Simulation Studies as a Tool to Understand Bayes Factors

Don van Ravenzwaaij[1] and Alexander Etz[2]
[1]Department of Psychology, University of Groningen
[2]Department of Psychology, University of California, Irvine

Correspondence concerning this article should be addressed to:
Don van Ravenzwaaij
University of Groningen, Department of Psychology
Grote Kruisstraat 2/1, Heymans Building, room 169
9712 TS Groningen, The Netherlands
Ph: (+31) 50 363 7021
E–mail should be sent to d.van.ravenzwaaij@rug.nl.

Abstract

When social scientists wish to learn about an empirical phenomenon, they perform an experiment. When they wish to learn about a complex numerical phenomenon, they can perform a simulation study. The goal of this paper is twofold. Firstly, this paper introduces how to set up a simulation study using the relatively simple example of simulating from the prior. Secondly, this paper demonstrates how simulation can be used to learn about the Jeffreys-Zellner-Siow (JZS) Bayes factor: a currently popular implementation of the Bayes factor employed in the BayesFactor R-package and freeware program JASP. Many technical expositions exist on JZS Bayes factors, but these may be somewhat inaccessible to researchers that are not specialized in statistics. This paper aims to show in a step-by-step approach how a simple simulation script can be used to approximate the calculation of the JZS Bayes factor. We explain how a researcher can write such a sampler to approximate JZS Bayes factors in a few lines of code, what the logic is behind the Savage Dickey method used to visualize JZS Bayes factors, and what the practical differences are for different choices of the prior distribution for calculating Bayes factors.

Keywords: Jeffreys-Zellner-Siow (JZS) Bayes factor, Savage Dickey method, prior distributions, statistical inference.

Research in the social sciences hinges on the existence of tools for conducting statistical testing. For the last one hundred years or so, arguably the golden standard has been the Null Hypothesis Significance Test, or NHST. Not without its protests though, the last twenty years in particular have seen an enormous amount of papers either questioning, or seeking to improve upon, the typical way statistical testing is (or was) conducted (Benjamin et al., 2018; Cumming, 2014; Gigerenzer, 2004; Harlow, Mulaik, & Steiger, 1997; Johnson, 2013; Wagenmakers, 2007; van Ravenzwaaij & Ioannidis, 2017, 2019).

Suggested alternatives to traditional ways of conducting statistical testing are not infrequently a variety of Bayesian hypothesis testing (see e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009; Dienes, 2011; Lee & Wagenmakers, 2013; Kruschke, 2014; van Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2019; van Ravenzwaaij & Wagenmakers, 2020, see van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017 for a general review of the use of Bayesian statistics in psychology). Perhaps the most popular method of the Bayesian hypothesis test quantifies statistical evidence through a vehicle known as the *Bayes factor*. The Bayes factor is a flexible tool for model comparison, allowing one to evaluate the evidence for and against any theories we care to specify through clever specification of the prior distribution, or prior (Etz, Haaf, Rouder, & Vandekerckhove, 2018). In practice, however, it is perhaps most common to see some sort of convenient default specification used for priors in Bayesian analyses (Kass & Wasserman, 1996; Gelman, Jakulin, Pittau, & Su, 2008). In scenarios calling for one of the most basic and often-used statistical tests, the *t*–test, a popular default specification uses the so–called Jeffreys-Zellner-Siow (JZS) class of priors. Developments of Bayes factors using these priors, often referred to as default Bayes factors, are inspired by the work of Jeffreys (1961) and Zellner and Siow (1980).

The goal of this paper is not to rehash statistical debates about *p*-values and Bayes factors. Nor will we give an exhaustive introduction to default Bayes factors. Many technical expositions already exist on default Bayes factors (e.g., Gönen, Johnson, Lu, & Westfall, 2005; Rouder et al., 2009; Morey & Rouder, 2011) and their extensions (Gronau, Ly, & Wagenmakers, 2018), but these papers are not always easily accessible to those researchers who are not statistical experts. This is unfortunate because the existence of easy-to-use tools for calculating the default Bayes factor, such as the point-and-click program JASP (The JASP Team, 2018) and the script-based BayesFactor package in R (Morey et al., 2018), make it imperative that researchers understand what these tools do.

The present paper is aimed at researchers that lack the time or confidence to delve into the advanced mathematics necessary to understand what is being calculated when software produces a default Bayes factor. Specifically, this paper will contain the bare minimum of equations and focus instead on a conceptual and intuitive understanding of the specific choices that underlie the default Bayes factor approach to the *t*–test.

The way to facilitate this improvement in intuition of Bayes factors is through the lens of *simulation*. We find that a useful analogue to simulation is experimentation. In an experiment, drawing samples can be used to learn about a population of interest. In a simulation, drawing samples can be used to learn about a complex numerical phenomenon. The 'population' of interest in a simulation can be anything from a known distribution to a quantity for which no analytical expression exists. Just like in an experiment, one draws a representative sample of this population. A visual display or numerical summary of the results can be used to learn something about this population. Throughout this paper, we use simple simulations with annotated code to show the reader how these can be used to learn about priors, Bayes factors, and posterior distributions (posteriors).

While not strictly necessary to understanding this tutorial, the reader may benefit from some conceptual knowledge of Bayesian statistical inference and Markov Chain Monte Carlo (MCMC) sampling. For those who would like to brush up on these topics we recommend our recent introductions in Etz and Vandekerckhove (2018, at least the first half) and van Ravenzwaaij, Cassey, and Brown (2018). Both of these papers are geared towards being accessible to researchers that are not statistical experts.

The current manuscript is organized as follows: In the first part, we introduce simulation studies and use them to explore a prior. In the second part, we provide a brief introduction on the model specifications that are used to calculate default Bayes factors in the context of a one-sample *t*–test. After this introduction, we use the simulation approach to generate data from the prior under the null hypothesis and from the prior under the alternative hypothesis to approximate the Bayes factor for hypothetical data that has not yet been observed.

In the third part, we provide sample code in JAGS (an acronym for software program Just Another Gibbs Sampler; Plummer, 2003) to approximate posterior distributions based on the default Bayes factor approach. This allows readers to obtain the output provided by either the BayesFactor package or the JASP software themselves while seeing exactly what choices are made for the priors and likelihood functions.

In the fourth part of this paper, we progress from posterior distributions to a second way to represent the Bayes factor: the Savage Dickey method (see e.g., Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Using the basic JAGS code provided in the previous section, we show the intuition behind the method, and a way to approximate the default Bayes factor by using the samples from JAGS.

In the fifth part of this paper, we use simulations to explore the effect of using different priors on the resulting Bayes factor. The aim is to show the reader how progressively more extreme priors change the conclusions compared to the priors employed by the default Bayes factor approach.

Motivating Simulation Studies

Social scientists typically use observation to learn about a certain population of interest (often, the population consists of humans). It is usually not possible to study the entire population, so social scientists draw a representative sample from this population. For instance, when we wish to learn if the consumption of alcohol affects perceptual discrimination, we may set up an experiment. In this experiment, a group of people randomly drawn from the population performs a perceptual discrimination task after having consumed different doses of alcohol (van Ravenzwaaij, Dutilh, & Wagenmakers, 2012). We might look at the data obtained in this random sample to learn something about the original question and consider this procedure of random sampling pretty straightforward.

Yet, those same social scientists may be daunted when they read a technical exposition on Bayes factors.[1] It can be hard to intuitively grasp how a statistical method works from looking at a complicated equation. Perhaps surprisingly, these scientists have a very similar tool at their disposal as the experimental procedure that is so helpful for empirical questions. This tool is the *simulation study.* In what follows below, we will illustrate how one can use simulations to learn about two key concepts in Bayesian inference: the prior and the Bayes factor. The goal of the first example is to show how one can explore aspects of a prior and the information it represents. The goal of the second example is to demonstrate how priors

[1] Both authors of the present manuscript can attest to having been there at one point in their career.

can be used to generate predictions about an experiment, and how these predictions form a key component for computation of the Bayes factor.

Exploring a Prior Distribution

One of the earliest roadblocks for researchers who want to adopt Bayesian methods in their research is the choice of prior for their analysis. Many methods exist to elicit priors from subject-matter experts (who are often the researchers themselves), but often a default prior is chosen for convenience. Regardless of how the prior is chosen, it remains an abstract mathematical object that can be nebulous to even experienced Bayesian analysts. Fortunately, we can use simulation studies to gain some intuition about the prior distribution and what it implies about our knowledge of the parameter of interest.

A common choice of prior in Bayesian analyses is the normal distribution. The normal distribution is typically one of the first things that is taught to social scientists in their introductory methodology or statistics undergraduate course. Students are typically taught that a standard Normal distribution (with mean 0 and standard deviation 1) has a distinctive bell shape similar to that depicted in Figure 1. Although most students learn to recognize these bell-shaped curves as being Normal distributions, we hazard that fewer students (or indeed, graduated social scientists) would know much about their properties beyond the 68-95-99.7 rule or the fact that the mean, median, and mode are equal.

Now consider the case of a one-sample *t*-test, where the parameter of interest is the standardized effect size, $\delta = \mu/\sigma$. If we choose a $N(0,1)$ prior distribution for δ, we hazard few readers would know off the top of their head what the probability is that the value for δ lies between -.5 and .5 (i.e., smaller in magnitude than a ‘medium’ effect). The authors of the present paper certainly do not. We could try writing out the equation for the standard normal distribution,

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \tag{1}$$

but, frankly, this may do more to intimidate than illuminate. Here is where we can turn to simulation. We encourage the reader to perform these operations alongside with us, at this stage all that is required is a working copy of the freely available program R (R Development Core Team, 2004). The Rmarkdown document underlying this manuscript, which includes all code, is available at `https://osf.io/9kwz4/`.

We can sample a value of δ from the prior many times, say a thousand times, and draw a histogram of the sampled values using the following line of code:

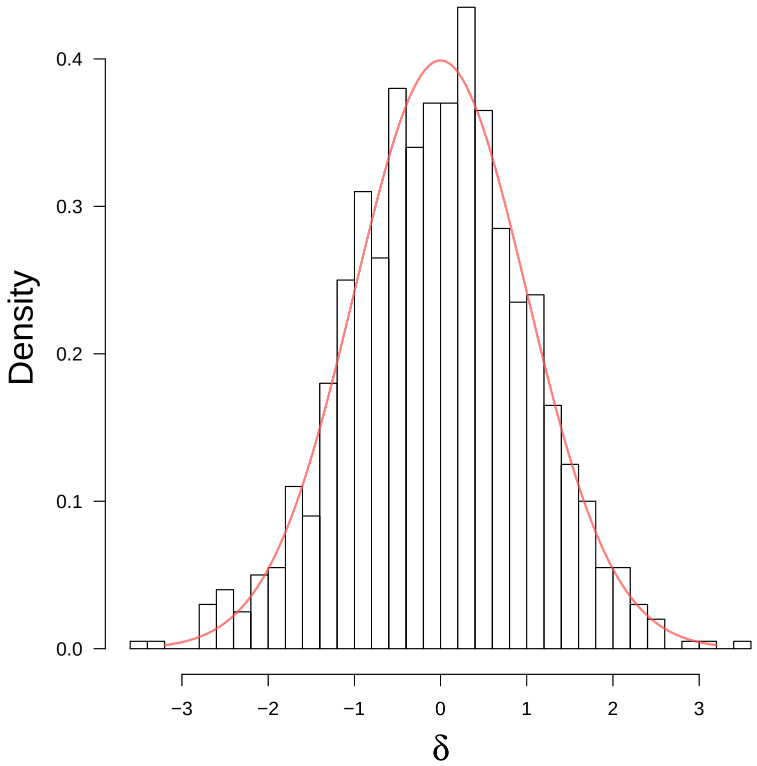

Figure 1. The simulated prior distribution for δ, which is specified as a Normal distribution with mean 0 and standard deviation 1. Note the close correspondence between the histogram of the samples and the analytical distributions overlaid as a red line.

```
# Sets seed that creates same pseudo-random sequence every time
# this code is run
set.seed (8675309)

# Create a vector of 1000 random numbers drawn from standard normal
# distribution (mean=0, sd=1)
delta = rnorm (1000)

# Plots a histogram of the sampled values
hist (delta, freq = F, breaks = 30)
```

The set.seed() command sets the starting number used to generate a sequence of pseudo-random numbers. It ensures that even if you do not save your output, you will obtain the exact same results next time you run your script. The histogram of the sampled values is shown in Figure 1, and gives us a sense of how the standard Normal distribution is built up, without having to decompose the equation.[2] Note the close correspondence with the analytical distribution overlaid as a red line. From here we can answer our question about the probability that the value of δ lies between -.5 and .5 by computing the proportion of simulated values of δ that fall within those limits. To do this we use the following code:

```
# Label the samples as 1 if they fall in the limits, 0 otherwise
deltasInTheLimits = (delta > -.5) & (delta < .5)

# Compute the proportion of samples of delta that are in the limits
proportionInTheLimits = mean (deltasInTheLimits)

# What is the proportion of delta samples in the limits?
# (Approximates the probability that delta is between those limits)
print (proportionInTheLimits)

# [1] .385
```

Thus, with this simple simulation we have found that, for the given prior, the probability that the value of δ lies between -.5 and .5 is approximately .385.[3] With the tool of simulation at our disposal we can look at any probability statement we wish. For instance, the probability that the value of δ lies between .5 and .8, that is, a 'medium' to 'large' positive effect, is found by changing the logical check in the first line in the above code to `deltasInTheLimits = (delta > .5) & (delta < .8)`; the resulting probability is approximately .089.

We have used simulation to explore our prior distribution for δ and have so far come away with two insights. First, this prior distribution corresponds to the a priori expectation that the true effect size is probably not smaller than 'medium' in magnitude; the probability that $|\delta| < .5$ is approximately .385, meaning the probability that $|\delta| > .5$ is approximately $1 - .385 = .615$. Second, according to this prior distribution it is unlikely a priori that the effect size is both positive and between 'medium' and 'large' in magnitude.

In the next two sections, we will use simulation to approximate Bayes factors. In the upcoming section, we first introduce some theory behind Bayes factors. Next, we use simulation to generate data using samples from two priors, each belonging to a different hypothesis. These predictions from the prior can be used to approximate Bayes factors for different values of the data, should they be observed. In other words, we can examine what the Bayes factor would be for data that is not yet observed. The section after that, we take the opposite approach and use simulation to approximate the posterior and calculate a Bayes factor for data that were actually observed.

[2] Using a different seed would give only slightly different results.

[3] The exact probability is found using the `pnorm` command as follows: `pnorm(.5) - pnorm(-.5)`. The result is .383.

Exploring the Bayes Factor

Theory of Bayes Factors

Before we go into the specifics of the default Bayes factor approach, it is worthwhile to provide a brief reminder of Bayes' rule in the context of two contrasting hypotheses:

$$\underbrace{\frac{P(H_0|\text{data})}{P(H_A|\text{data})}}_{\text{Posterior odds}} = \underbrace{\frac{P(H_0)}{P(H_A)}}_{\text{Prior odds}} \times \underbrace{\frac{P(\text{data}|H_0)}{P(\text{data}|H_A)}}_{\text{Bayes factor}} \tag{2}$$

The quantity on the left is the *posterior odds*, or the probability of the null hypothesis H_0 given the data relative to the probability of the alternative hypothesis H_A given the data. The quantity in the middle is the *prior odds*, or the probability of the null hypothesis H_0 before having seen the data relative to the probability of the alternative hypothesis H_A before having seen the data. The quantity to the right is the *Bayes factor*, or the probability of the data given the null hypothesis H_0 relative to the probability of the data given the alternative hypothesis H_A.

If one wants to use statistical inference to test hypotheses, one must first make some choices regardless of whether one employs the traditional NHST method or Bayesian testing. First, one must decide on the form of H_0 and H_A. A convenient way to specify these hypotheses is to relate them to an effect size δ parameter. In this context, H_0 usually specifies that the effect size is exactly zero, whereas H_A can be one-sided/directional (e.g., the effect size is larger than zero) or two-sided/non-directional (e.g., the effect size is different from zero).

Furthermore, both NHST and Bayesian testing require making an assumption about the way the data is distributed, as that will affect the choice of your likelihood functions (Etz, 2018). For example, in the case of a t-test both NHST and Bayesian testing assume that data are normally distributed. When conducting Bayesian inference, one might choose a normal distribution for the likelihood function to reflect this assumption.

For Bayesian statistical inference, two more choices need to be made. The first choice is about the prior odds, or the ratio of prior model probabilities. Does one believe H_0 and H_A to be equally plausible before having seen any data? This degree of belief can be informed by prior study results, or by a researcher's intuition, but will likely contain a certain degree of subjectivity. Fortunately, the prior odds have no effect on the Bayes factor, so every reader of a study is welcome to combine the reported Bayes factor with their own prior odds to arrive at their own posterior odds. In the context of hypothesis testing, textbooks often follow a convention set by Jeffreys (1961) and assume a-priori that both hypotheses are equally likely (cf. top row of Figure 2) by setting the prior odds to 1 (but see Kruschke & Liddell, 2018, for a discussion of scenarios in which you have more specific information on prior odds). When this is the case the Bayes factor and posterior odds are equal.

The second choice a Bayesian needs to make concerns the prior distribution of the effect size parameter under each hypothesis. Contrary to the prior odds described in the last paragraph, the prior distributions of the effect size parameter *do* affect the resulting Bayes factor, as we will soon see in the section "Prior Influence on the Bayes Factor". The prior distribution is quite simple for an H_0 that specifies the effect size is exactly zero (there

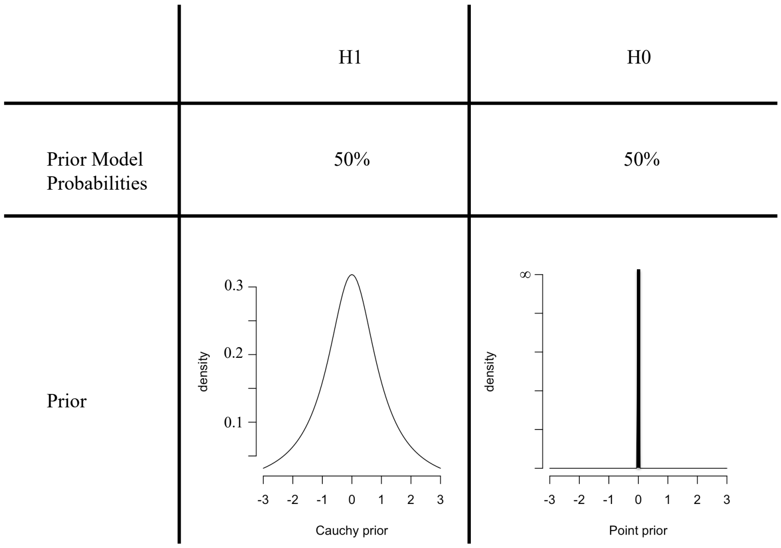

Figure 2. Prior odds and prior distributions. See text for details

is only one permissible value for the effect size, so the distribution consists of a "spike" at zero, see the bottom-right panel of Figure 2), but for an H_A that specifies the effect size is different from zero, a probability distribution is needed to specify which values of the effect size δ are more likely than others.

So far, nothing of the above is specific to the default Bayes factor approach. What distinguishes this approach from any other Bayesian hypothesis test approach is the choice of the prior distribution for the effect size parameter δ under H_A (i.e., the bottom-left panel of Figure 2). The chosen distribution is a Cauchy distribution centered on zero, usually with a scale parameter equal to $\sqrt{2}/2$ (Morey et al., 2018). One may think of this prior as a standard normal distribution with fatter tails. The scale parameter is the upper and lower bound that encompasses 50% of the distribution. So, a Cauchy distribution centered on zero with a scale parameter of $\sqrt{2}/2$ has 50% of the distribution between $-\sqrt{2}/2$ and $\sqrt{2}/2$ (or approximately -.71 to .71).

The Cauchy prior has some desirable mathematical properties (see e.g., Bayarri, Berger, Forte, & García-Donato, 2012; Consonni, Fouskakis, Liseo, & Ntzoufras, 2018), such as model selection consistency (for data generated under a model, the corresponding Bayes factor should go to infinity as sample size goes to infinity), predictive matching (a minimum sample size should exist for which the Bayes factor is 1, such that models are indistinguishable), and information consistency (a minimum sample size should exist for which data that result in test statistics that go to infinity should have corresponding Bayes factors that also go to infinity). Other priors may share some of these desirable properties, but the Cauchy prior has caught on as the go-to choice because it satisfies them all and is relatively easy to specify and interpret.

As a perhaps more intuitive way to grasp why such a prior makes sense, we consider why it should be so that the prior density is relatively high for values closer to zero and why the distribution should be symmetrical. For most studies, it should be the case that an effect size of say $\delta = 10$ is substantially less likely to be found than an effect size of say $\delta = 5$. Similarly, $\delta = 5$ should be less likely than $\delta = 2$, which in turn should be less likely than $\delta = 0.8$. Moreover, in the specific context of testing the null hypothesis that δ is zero, "the mere fact that we are seriously considering the possibility that it is zero may be associated with a presumption that if it is not zero it is probably small" (Jeffreys, 1961, p. 332). This accounts for the fact that the distribution is peaked instead of flat.

The second thing to bear in mind is that in the context of two-sided testing, negative effect sizes should be just as plausible as positive effect sizes, as every parameter can be flipped around such that the sign of the effect size can switch (e.g., happiness can be relabeled unhappiness, Group 1 can be relabeled Group 2, etc.). This accounts for the fact that the distribution is symmetrical around zero.

The reader might wonder if it would not make more sense to have a distribution with two peaks, one above say $\delta = 0.2$ and one above $\delta = -0.2$. Such a distribution would still be non-flat and symmetrical, but would incorporate the fact that researchers probably have some intuition about the phenomena they investigate, such that small effects are more likely to be studied than null effects. The beauty of the Bayesian approach is that everyone is at liberty to pick their own prior distribution, the one they think best reflects the a-priori knowledge of the field. The Cauchy prior described above is considered by many to be a sensible default prior. It is relatively diffuse, reflecting the fact that the researcher is not willing to commit to very specific values of the effect size parameter a-priori. Such a prior will have a comparatively small influence on the posterior distribution, such that most of the diagnosticity comes from the likelihood of the data. We will see examples of the effect of choosing different kinds of priors in section "Prior Influence on the Bayes Factor" below.

In the next sub-section, we will use simulation to generate data from the priors under the null and alternative hypothesis to gain insight into the mechanics of the Bayes factor.

Simulation of Bayes Factors

In the previous section, we have learned that the Bayes factor is computed by taking the ratio of two probabilities: The probability of the data given the null hypothesis, $P(D|H_0)$, and the probability of the data given the alternative hypothesis, $P(D|H_A)$. The section before that, we used simulation to draw samples from a prior distribution. In this section we will combine these two ideas to obtain Bayes factors for data that have not yet been observed (see also Etz et al., 2018). We will illustrate this idea using a running example of Kim the educational psychologist.

Kim is interested in examining whether a new program focused on more systematic rehearsal of learned topics leads to lasting increases in IQ scores among high-school students. She randomly selects 50 students from high-schools in the Netherlands and has them enroll in her program (with permission from their teachers and parents, of course). Kim administers an IQ test to the 50 students directly before the program and half a year after the program. She is interested in whether there is a gain in the IQ score that lasts until half a year after the program.

Kim does not have any data yet, but we are going to use simulation to examine what data she might observe if the null hypothesis is true and what data she might observe if the alternative hypothesis is true. The first component of the Bayes factor that we will focus on is $P(D|H_0)$, the probability of the data given the null hypothesis. As indicated in the bottom-right panel of Figure 2, the null hypothesis is a point null. This means that under the null hypothesis, the *population* effect size δ can only be zero. Using simulation, we can examine the sampling distribution of the *sample* effect size, Cohen's d, when the null hypothesis is true and the sample size is 50. We use the following code to generate 10,000 sample effect sizes under the null hypothesis:

```
# Our sample size for each experiment
n = 50

# Number of simulated experiments to generate
nSims = 10000

# Create an empty (for now) vector in which to store sample effect sizes
Sample0 = c()

# Repeat nSims times: create data set -> compute effect size
for (i in 1:nSims)
{
  # Generate data set from N(0,1)
  data = rnorm (n, 0, 1)

  # Compute one sample effect size and store in position i of the vector
  Sample0[i] = mean(data) / sd(data)
}
```

The resulting sample effect sizes are represented by the green histograms and density in Figure 3. All four panels show the same data with different granularity.

Simulating data for the second component of the Bayes factor, $P(D|H_A)$, is a little more involved. The complication comes from the fact that for the alternative hypothesis we do not merely set δ to some fixed value. We instead specify a prior distribution for δ, reflecting the fact that we do not yet know its true value. As indicated in the previous section, this prior is Cauchy centered on zero with scale $\sqrt{2}/2$.

The way to incorporate the fact that we do not know the true value for δ is to add an additional step to the simulation. First, we sample a population effect size from the distribution of potential population effect sizes. In this case, the distribution of potential population effect sizes is dictated by the Cauchy prior. This extra step is implicit in the previous simulation for the null hypothesis, because the population effect size is always the same (i.e., zero). Once we have generated this population effect size, we draw a data set of size 50 from this population effect size, just as in the simulation for the null hypothesis presented previously. We use the following code to generate 10,000 sample effect sizes under the alternative hypothesis:

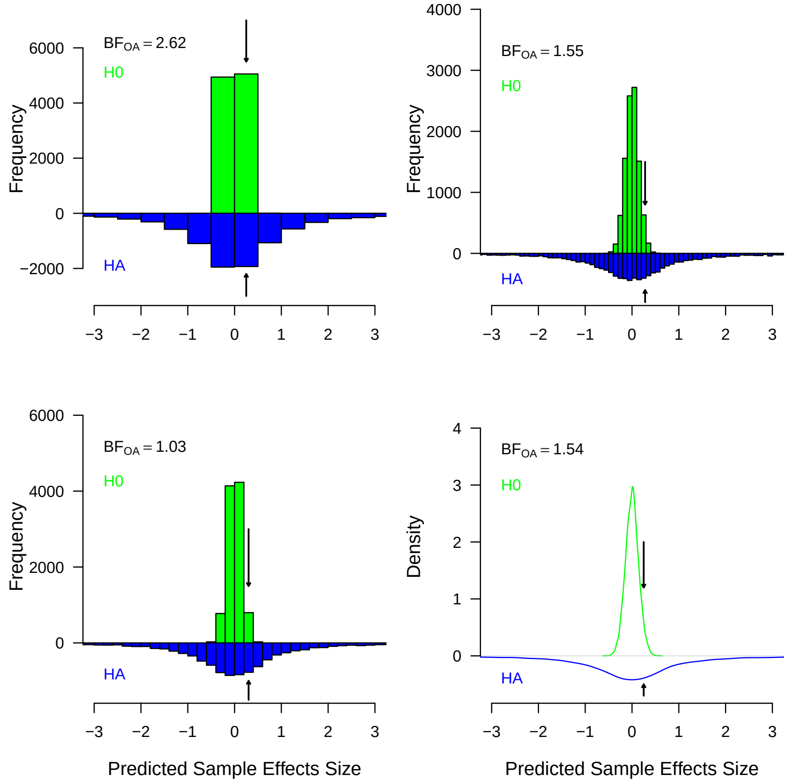

Figure 3. Sample effect sizes simulated from the prior distribution under H_0 and from the prior distribution under H_A. Note that we are plotting the negative of the frequency and density for the alternative hypothesis, resulting in a reflection across the x-axis for easier comparison.

```
# Create an empty (for now) vector to store sample effect sizes
SampleA = c()

# Repeat nSims times: sample a parameter -> create data set ->
# compute effect size
for (i in 1:nSims)
{
  # Generate a delta parameter (true effect size) from the Cauchy dist.
  delta = rcauchy (1,0,scale=sqrt(2)/2)

  # Generate data set from N(delta, 1)
  data = rnorm (n, delta, 1)

  # Compute one sample effect size and store in position i of the vector
  SampleA[i] = mean (data) / sd (data)
}
```

The resulting sample effect sizes are represented by the blue histograms and density in Figure 3. The distributions of sample effect sizes under both hypotheses are called *prior predictives* (Ntzoufras, 2009). Now that we have distributions of hypothetical data under the null and under the alternative hypothesis, the next step is to turn these into Bayes factors. Recall that a Bayes factor is nothing more than the ratio of the probability of the data under one hypothesis over the probability of the data under the other hypothesis. In other words, we can compare the green histograms to the blue histograms for a specific part of the data, and the ratio of these two will be our Bayes factor.

As an example, Kim has not collected any data yet, but let us consider the scenario in which Kim has collected some data with a sample effect size d of 0.25. Under which hypothesis is this Cohen's d more likely? In our simulated data, an exact value of 0.25 will not have occurred (at least not prior to rounding), but we can approximate the Bayes factor by binning the data. For instance, the top-left panel of Figure 3 shows the data binned with a bin-width of 0.5. We can now get a very rough approximation of the Bayes factor for sample effect size d of 0.25 by dividing the number of times a sample effect size d between 0 and 0.5 occurred under the null hypothesis and under the alternative hypothesis:

```
# Set the limits of our bin that is .5 wide
Bin = c(0, 0.5)

# Proportion of null-generated sample effect sizes within the bin limits
Nulls = mean (Sample0<Bin[1] & Sample0<Bin[2])

# Proportion of alternative-generated sample effect sizes in bin limits
Alts =  mean (SampleA<Bin[1] & SampleA<Bin[2])

 # Approximate BF given by ratio of the two proportions
BF0A = Nulls / Alts
```

Essentially what we are doing here is dividing the height of the green bar marked with an arrow by the height of the blue bar marked with an arrow. Running this script, we obtain a Bayes factor of 2.62. This means that *if* Kim will observe a sample effect size somewhere in between 0 and 0.5, that data will have been slightly more likely under H_0 than under H_A. Put differently, the null hypothesis predicts a sample effect size d in the range of 0 and 0.5 more strongly than the alternative hypothesis.

What happens if we make the bins more narrow? The reader can experiment with this themselves by changing the values of the bin. The bottom-left panel examines the approximate Bayes factor for bin $0.2-0.4$ and the top-right panel examines the approximate Bayes factor for bin $0.2-0.3$. For our specific samples, resulting approximate Bayes factors are 1.03 and 1.55, respectively. For reference, the exact Bayes factor corresponding to a sample effect size of 0.25 and $n = 50$ is 1.54.[4] We see that even a bin with a width of 0.1 comes pretty close already.

[4] This result can for instance be obtained from the BayesFactor R package using ‘1/exp(ttest.tstat(t=.25*sqrt(50), n1=50, rscale = sqrt(2)/2)[[’bf’]])’

What if instead of 0.25, Kim had collected data with a sample effect size d of 0.75? Inspection of the histograms shows us that around value 0.75 on the x-axis, the blue bars are actually much larger than the green bars, indicating that this data is (much) more likely to occur under H_A than under H_0: the corresponding Bayes factor is over 6,000.

In the next two sections, we turn our attention to using simulation to approximate Bayes factors for data that is actually observed. In doing so, we substantially change the nature of our simulations. Before we generated many instances of data that were consistent with two different priors. In what follows, we zoom in on one specific data set that was actually observed, and use simulation to draw samples from the posterior distribution. The samples from the prior and posterior distributions, in turn, can be used to obtain a Bayes factor.

Simulation of Posterior Distribution

In the previous section, we simulated a wide range of data from the prior. We used the resulting simulated data sets to approximate Bayes factors for specific hypothetical data, should it have been observed. In this section, we assume one specific data set has actually been observed, and we are going to use simulation to explore the relationship between the posterior, the prior, and the Bayes factor (see section 8.1 of Lee & Wagenmakers, 2013, for a similar demonstration). Put differently: in the previous section we used simulation to generate multiple data sets that could be observed for a (set of) true parameter value(s). In the next two sections we use simulation to explore for a range of parameter values how likely it is to have generated a single data set. So in essence, we turn it around: first data that could have been observed from one parameter, now parameters that could have generated one data set.

Before we get to our example, a quick refresher of Bayes' rule for estimation may be useful. Bayes' rule states that the posterior density for an individual parameter value θ after seeing the data is given by

$$\underbrace{p(\theta|\text{data}, H)}_{\text{Posterior density}} = \underbrace{p(\theta|H)}_{\text{Prior density}} \times \underbrace{\frac{p(\text{data}|\theta, H)}{p(\text{data}|H)}}_{\text{Updating factor}} \tag{3}$$

for a given prior density $p(\theta|H)$, a likelihood for the data given a specific value for θ $p(\text{data}|\theta, H)$, and a likelihood for the data that is a weighted average across each possible value of θ $p(\text{data}|H)$. The latter is typically called marginal likelihood, as it is a likelihood in which one variable is collapsed over 'marginalized out.' For example, say the probability of rain on any given day in January in the Netherlands is 30%: $p(\text{rain}|January, Netherlands) = .3$. Furthermore, say that the probability of rain on any given day in July in the Netherlands is 40%: $p(\text{rain}|July, Netherlands) = .4$. Assuming, for this simple example, that these are the only two months of the year, our marginal likelihood $p(\text{rain}|Netherlands)$, in which the variable 'month' is marginalized out, now becomes $(.3 + .4)/2 = .35$; the number of days in both months are identical, so that the weighted average of the probabilities is simply the mean. The likelihood we work with in the rest of these sections is $p(\text{data}|\delta, H)$, or the probability of the data given a specified population effect size δ, as dictated by the t-test model. The marginal likelihood we work with in the rest of these sections is $p(\text{data}|H)$, or the probability of the data given any population effect size, obtained by marginalization over the possible values of δ (weighted by the prior distribution).

The grouping of terms in Equation 3 makes it clear that the posterior density for a given parameter value θ is merely the prior density for that point multiplied by an updating factor. As we learned through simulation in the previous section, the likelihood and marginal likelihood (the terms comprising the updating factor) are given by the heights of the prior predictive distributions at the point corresponding to the data, for a given parameter value θ (in our previous example, the null hypothesis, with value zero) and a weighted-average over all θ values (in our previous example, the alternative hypothesis) respectively. Essentially, Bayes' rule says that values of θ whose predictive distribution assign relatively high probability to the observed data get a bump in density, and those that assign relatively low probability decrease in density.

In the following example, we are going to use an overly simplistic data set. Our fictional data set consists of seven values: -2, -1, 0, 1, 2, 3, and 4. We are interested in testing if the population mean differs from zero. We can run this analysis simply in JASP by creating a csv-file with a column of these seven values and running a Bayesian one-sample *t*-test (see Wagenmakers et al., 2018, for some JASP examples). The output is shown in the top panel of Figure 4.

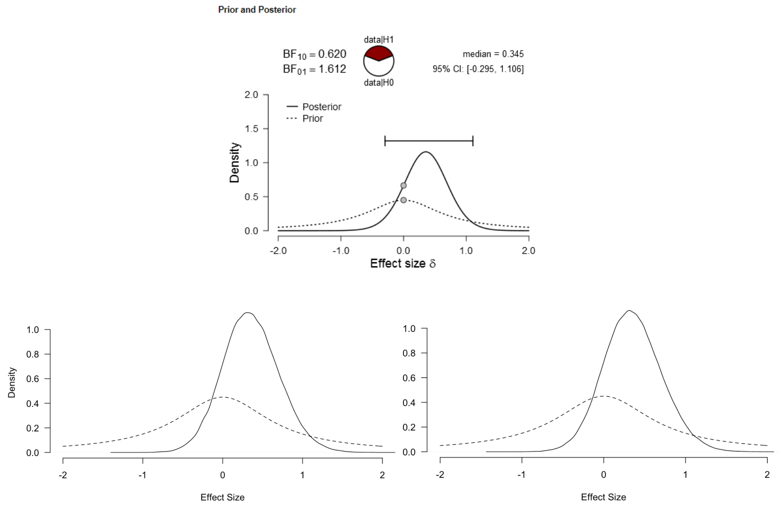

Figure 4. JASP output (top), JAGS output based on raw data (bottom-left), and JAGS output based on summary statistics (bottom-right) for the simple {-2, -1, 0, 1, 2, 3, 4} data set.

In the remainder of this section, we are going to approximate what JASP computes directly (and, as a result, somewhat obscurely) with what a sampler can do more intuitively. For this to work, we need a working copy of JAGS (Plummer, 2003) in addition to the program R we have been using so far. We also need to install R-packages 'R2jags' and (for later) ‘polspline’. With this in order, the following lines of code approximate the posterior distribution that JASP produced. The JAGS model itself is contained in the object ‘JZSfulldata’.

```
# Load the package R2jags, and interface between R and JAGS
library (R2jags)

# The data used for our sample
dat = -2:4

# Number of data points in the sample
n   = length (dat)

# The following JZSfulldata object is a JAGS model string
# It will subsequently be used to specify the model in the jags() function below
JZSfulldata <- "model{
  # This for loop specifies the likelihood for the data
  # ("How were the data from the sample generated?")
  for (i in 1:n)
  {
    # Data point i is normally distributed with mean mu and precision invsigma2
    dat[i] ~ dnorm (mu, invsigma2)
  }

  # Next come prior distributions for delta and invsigma2 parameters

  # Cauchy prior on delta (using the t-dist. with 1 df)
  delta ~ dt (0, 2, 1)

  # Improper prior for sigma2 (approximating the Jeffreys’s prior)
  invsigma2 ~ dgamma (.00001, .00001)

  # Finally, transform back to the variables mu and sigma
  sigma <- sqrt (1/invsigma2)
  mu <- delta * sigma
}"

# List of variables to be passed to JAGS (data and sample size)
Fulldata  = list (dat = dat, n = n)

# This tells JAGS which parameters’ samples we want to see when it finishes
JAGSparam = c("mu", "sigma", "delta")

# Finally, the jags() function calls JAGS to run the simulations as we specified
FitFulldata = jags (data = Fulldata, parameters.to.save = JAGSparam,
  n.thin = 1, n.iter = 20000, n.burnin = 10000, n.chains = 1,
  model.file = textConnection(JZSfulldata))
```

The posterior distribution is ultimately obtained via compromise between the prior distribution and the information provided by the data through the likelihood (for an illustration see example 3 of Etz & Vandekerckhove, 2018). So now we need to provide prior distributions and a likelihood for the data. Starting with the likelihood, we assume that each data point comes from a normal distribution with unknown population mean μ and variance σ^2 (as we would with a traditional t-test). Note that JAGS is a bit unorthodox when it comes to statistical software, because it works with mean and "precision" parameters, or the inverse of the variance. So rather than specifying a model using μ and σ or σ^2, we instead need to specify the model using μ and $1/\sigma^2$. Thus, we specify a normal distribution with mean mu and precision $invsigma2$ in the code.

To complete our model, we need prior distributions for the parameters that are specified in our likelihood function. However, in the default Bayesian t-test we test hypotheses using the standardized effect $\delta = \mu/\sigma$, so instead of specifying a prior for the mean μ directly we specify priors for δ and $1/\sigma^2$ and then convert back to μ using $\mu = \delta \times \sigma$ (conversions take place in the "# Computed variables" model section). Recall from earlier that the prior for δ is a Cauchy distribution centered on zero with a scale parameter of $\sqrt{2}/2$ (equivalent to a t-distribution with a scale of $\sqrt{2}/2$ and one degree of freedom, see bottom-left panel of Figure 2). The parameters required by JAGS for the t distribution are mean, precision (inverse of scale, squared), and degrees of freedom. As such, the precision parameter of the δ prior we need to provide is the squared inverse of the scale: 2.

Finally, we need a prior for the precision $1/\sigma^2$. The formal prior used in the default Bayesian t-test is known as the *Jeffreys prior* (Rouder et al., 2009), an improper prior because the area under the curve does not add up to 1.[5] Unfortunately, improper priors are not allowed in JAGS. For our purposes, this prior is approximated near-perfectly by an inverse gamma distribution with shape and scale parameters of 0.00001 (which is equivalent to a gamma distribution on σ^2 with shape and scale parameters of 0.00001). A visualization of a gamma distribution with shape and scale parameters of 0.00001 is provided in Figure 5.

The remaining bit of code specifies the data and runs the JAGS sampler. Specifically, we are going to draw 10,000 values from the posterior distribution (n.iter-n.burnin). More would be possible, either by increasing the number of iterations, or by running the simulation multiple times (i.e., increasing the number of 'chains'), but for purposes of this example, 10,000 values offer more than enough precision to approximate the posterior distribution. We have chosen to run a large number of samples, because that will allow us to approximate the Bayes factor that analytical methods provide more accurately. To plot the output of the sampler, we can use the following lines of code:

[5] This statistical nomenclature can be somewhat misleading. "Improper" here only refers to the technical aspect of the area not adding up to 1; it does not refer to the validity of this type of prior for modeling purposes. Improper priors are quite suitable for many modeling endeavors.

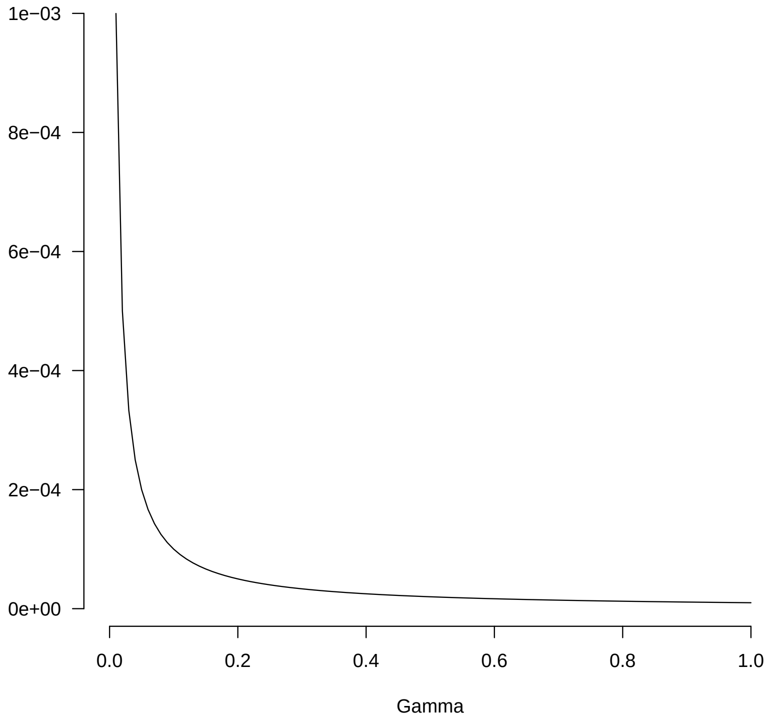

Figure 5. Gamma distribution with shape and scale parameters of 0.00001.

```
# Extract the mcmc samples JAGS generated
Fulldatamcmc = as.mcmc (FitFulldata)

# Pull out the mcmc samples for the delta parameter specifically
Fulldelta = Fulldatamcmc[[1]][,"delta"]

# Create a density plot of the posterior samples of delta
plot (density (Fulldelta, n = 4096), xlim=c(-2,2), bty = ’n’, axes = F,
  xlab = "Effect Size", ylab = "Density", main = "")

# Create the axes
axis (1); axis (2, las = 1)

# Add the prior distribution of delta for comparison
curve (dcauchy (x, 0, sqrt(2)/2), from = -2, to = 2, lty = 2, add = T)
```

The resulting output is shown in the bottom-left panel of Figure 4. The posterior shows us the probability density of different values of unknown population parameter δ, given the observed data set of {-2, -1, 0, 1, 2, 3, 4} under the alternative hypothesis. We see that at least at first glance, the posterior distribution (the solid black ones in both plots) looks very similar to the one produced by JASP in the top panel. In the next section, we will explore whether the Bayes factors for the analytical JASP approach and the approximate JAGS approach agree.

In the example, we specified a likelihood for each data point separately. It is entirely possible to summarize all the relevant characteristics of the data set using the sample test statistic t and put our likelihood on that instead. Such a specification is provided in section 'Simulation of Posterior Distribution Using Test Statistic' in the Appendix.

With an approximation of the posterior distribution for our effect size parameter δ under the alternative hypothesis in hand, we now turn to the next section, in which we use this information to obtain the default Bayes factor.

Bayes Factor Visualization: The Savage Dickey Method

We have used JAGS to great effect to obtain the posterior distribution for the effect size parameter δ. We can now calculate a Bayes factor by taking the ratio of the prior and posterior density of δ evaluated at zero, a technique known as the Savage Dickey method (see e.g., Wagenmakers et al., 2010). Conveniently, the Bayes factor is nothing more than an updating factor (see Equation 3), as it quantifies whether a parameter or hypothesis is more plausible after having seen the data (quantified by the posterior) than before having seen the data (quantified by the prior).

Although it can be shown mathematically why the Bayes factor can be represented as the ratio of prior and posterior densities (see Box 1), in our opinion, understanding why the Savage Dickey method works is not intuitive. In what follows below, we will use the simulation results from the previous section to explain the rationale behind the Savage Dickey method.

Box 1: The Savage-Dickey Density Ratio.
The Savage-Dickey density ratio (often shortened to just the Savage-Dickey ratio or Savage-Dickey method), is the ratio of posterior density to prior density for a parameter value (Dickey, 1971; Dickey & Lientz, 1970). The Savage-Dickey method is useful because it connects Bayes' rule for hypothesis testing (Equation 2) with Bayes' rule for estimation (Equation 3) and gives us a way to "see" how large the Bayes factor is.

Consider the hypothesis testing case when hypothesis H_0 is nested within hypothesis H_A, meaning H_0 sets a parameter present in H_A to equal some pre-determined value. The t-test is one such example, where H_0 restricts δ to take the value zero. In this scenario, $p(\text{data}|\delta = 0, H_A)$ will equal $p(\text{data}|H_0)$ because H_0 is just H_A with the restriction $\delta = 0$. Moreover, the same $p(\text{data}|H_A)$ shows up in both the Bayes factor and the estimation updating factor. Thus, we have that the Bayes factor testing whether $\delta = 0$ equals the estimation updating factor at $\delta = 0$. Now if we divide each side of Equation 3 by the prior density we see the following result:

$$\underbrace{\frac{p(\delta = 0|\text{data}, H_A)}{p(\delta = 0|H_A)}}_{\text{Savage-Dickey ratio}} = \underbrace{\frac{p(\text{data}|\delta = 0, H_A)}{p(\text{data}|H_A)}}_{\text{Updating factor}} = \underbrace{\frac{p(\text{data}|H_0)}{p(\text{data}|H_A)}}_{\text{Bayes factor}}. \tag{4}$$

Hence, the Savage-Dickey ratio, the updating factor, and the Bayes factor are all equal and we can conveniently visualize the Bayes factor as a comparison of the prior and posterior densities. When the posterior density is larger (smaller) than the prior density, the Bayes factor will show evidence in favor of (against) H_0.

However, it is important to note that this simple relationship between the Bayes factor and the Savage-Dickey density ratio can become more complicated in hypothesis testing scenarios involving models with many interdependent parameters. In such cases, it is possible that the Savage-Dickey ratio and Bayes factor diverge by a positive scalar factor; that is, the ratio of posterior to prior density equals k but the Bayes factor equals αk for some positive α. For technical explanations and examples of this phenomenon see Heck (2019), Verdinelli and Wasserman (1995), and Wagenmakers, Gronau, Dablander, and Etz (2020, Section 6).

In order to gain some intuition with respect to the Savage Dickey method that is typically used to visualize the Bayes factor, we are going to briefly move away from our original null hypothesis H_0 and alternative hypothesis H_A. Specifically, we are going to change our point null hypothesis ($\delta = 0$) into an interval null hypothesis (and change our alternative hypothesis to all values outside the null interval). Say, for instance, that our null hypothesis is given by $-0.5 < \delta < 0.5$ and our alternative hypothesis is given by $|\delta| > 0.5$ (i.e., the remaining possible values of δ). This scenario is visualized in the top panel of Figure 6, with the null hypothesis corresponding to the area between the vertical dashed lines.

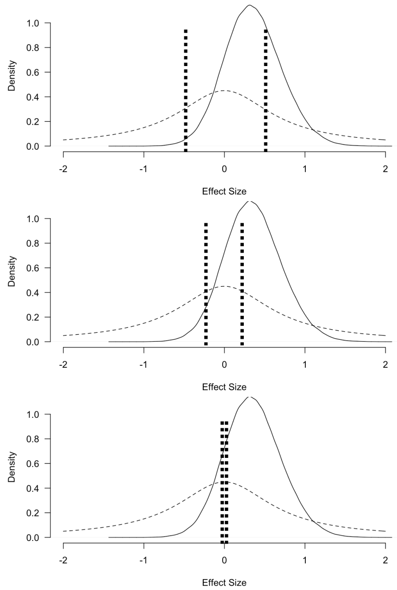

Figure 6. Three different interval null hypotheses. Top panel: $-0.5 < \delta < 0.5$. Middle panel: $-0.25 < \delta < 0.25$. Bottom panel: $-0.01 < \delta < 0.01$.

Recall that we can use Equation 2 to obtain the Bayes factor. Practically speaking, we need the area of the posterior (i.e., the solid curve) between the vertical lines and outside of the vertical lines, and the area of the prior (i.e., the dashed curve) between the vertical lines and outside of the vertical lines. With our JAGS samples in hand, we are going to approximate the area of the posterior within the vertical lines by calculating the proportion of samples that fell within the vertical lines. The area outside of the dashed lines is approximated by subtracting the previous proportion from 1 (recall that probability distributions sum to 1).

Because our prior distribution is an exact distribution, we can calculate the area of the prior within the vertical lines exactly. Code to calculate these four quantities is shown below:

```
# The max absolute value of delta under the null hypothesis
Margin  = .5

# Area under the posterior for the null hypothesis
PostH0  = mean (Fulldelta>-Margin & Fulldelta<Margin)

# Area under the posterior for the alternative hypothesis
PostHA  = 1 - PostH0

# Area under the prior for the null hypothesis
PriorH0 = pcauchy (Margin, 0, sqrt(2)/2) - pcauchy (-Margin, 0, sqrt(2)/2)

# Area under the prior for the alternative hypothesis
PriorHA = 1 - PriorH0

# Dividing the posterior odds by the prior odds gives the Bayes factor
BF0A    = (PostH0/PostHA) / (PriorH0/PriorHA)
```

For our samples, we get a posterior area between the vertical lines of .67, but because they are samples the reader might get slightly different results. The area of the prior between the vertical lines is approximately .39. The resulting Bayes factor would be $BF_{0A} = \frac{.67}{1-.67} / \frac{.39}{1-.39} \approx 3.20$, so the relative support for the null hypothesis provided by the data is approximately 3.

Let us repeat this procedure, but now choosing a more narrow band around $\delta = 0$ as our null hypothesis: $-0.25 < \delta < 0.25$ (see middle panel of Figure 6). The reader can perform the calculation with us, all that is required is to change the 'Margin' variable in the previous bit of code to reflect our new band of {-0.25, 0.25}. For our samples, we get a posterior area between the dashed lines of .36. The area of the prior between the dashed lines is approximately .22. The reader will note that both areas are smaller than for the previous band as they should be. The resulting Bayes factor is $BF_{0A} = \frac{.36}{1-.36} / \frac{.22}{1-.22} \approx 2.04$.

We are going to repeat this procedure one last time, now choosing a very narrow band around $\delta = 0$ as our null hypothesis: $-0.01 < \delta < 0.01$ (see bottom panel of Figure 6). For our samples, we get a posterior area between the dashed lines of .015. The area of the prior between the dashed lines is approximately .009. The reader might note that the areas of both the prior and the posterior for $H0$ are now very small. The consequence of this is that the areas of both the prior and the posterior for $H1$ are close to one.

We can exploit this to simplify our expression for the Bayes factor as follows: $BF_{0A} = \frac{.015}{1-.015}/\frac{.009}{1-.009} \approx \frac{.015}{.009} = 1.64$. In words, we can approximate the Bayes factor by dividing the area of the posterior for H_0 by the area of the prior for H_0. We do not need to explicitly take into account the areas outside of the vertical lines anymore! This approximation becomes better and better the more narrow we choose our band around $\delta = 0$, providing we drew enough samples. In the limit where we use the point null hypothesis $\delta = 0$, dividing the height of the posterior for $\delta = 0$ by the height of the prior for $\delta = 0$ provides the exact Bayes factor. This method of obtaining the Bayes factor is called the Savage Dickey method (see e.g., Wagenmakers et al., 2010), we can calculate the Bayes factor using the Savage Dickey method for our samples using the following bit of code:

```
# An R package that ’guestimates’ smooth densities based on data
library (polspline)

# Posterior density for null value under alt hypothesis
Post0underHA = dlogspline (0, logspline (Fulldelta))

# Prior density for null value under alt hypothesis
Prior0underHA = dcauchy (0, 0, sqrt(2)/2)

# The Savage Dickey ratio gives the Bayes factor
BFOA = Post0underHA / Prior0underHA
```

The logspline function essentially treats the sampled posterior distribution as if it were a proper density function and evaluates the density at $\delta = 0$ *as if* the curve were smooth. Now that we are no longer working with interval hypotheses, the prior and posterior for the null hypothesis are no longer smooth continuations of the prior and posterior for the alternative hypothesis. Thus, our notation now refers to the prior and posterior density for the null value under the alternative hypothesis (i.e., the value above zero in the bottom-left panel of Figure 2). The resulting Bayes factor is 1.64 for our sample, which corresponds fairly well with the JASP output (see top panel of Figure 4, value behind BF_{01}).

The last two sections have attempted to visualize an alternative way of obtaining Bayes factors from what happens behind the scenes when JASP calculates Bayes factors for a simple one-sample t test design. One advantage of such a hands-on approximation is that it becomes quite simple to examine how strong the influence is of the choice of prior employed in the default Bayes factor approach. We provide a few basic examples of how such an examination might be conducted in the next section.

Prior Influence on the Bayes Factor

We have seen in the previous sections the distinguishing feature for the JZS Bayes factor: the Cauchy prior. We have also seen how the Bayes factor can be calculated by (1) generating prior predictives under the null hypothesis and under the alternative hypothesis, or by (2) evaluating the prior and posterior distribution under the alternative hypothesis, evaluated at $\delta = 0$, for an observed data set.

One reasonable question to ask is: how much does the choice of prior affect the Bayes factor? After all, sometimes multiple defensible non-default priors can be specified in addition to or instead of a default prior (e.g., Dienes, 2019; Gronau, Ly, & Wagenmakers, 2019; Jones & Johnson, 2014; Saunders, Milyavskaya, Etz, Randles, & Inzlicht, 2018). The sensitivity of Bayes factors to priors has been discussed previously by Liu and Aitkin (2008) and Vanpaemel (2010). In what follows, we demonstate how to examine the effect of prior sensitivity oneself by calculating the Bayes factor for three alternative choices of prior that differ in increasing amounts from the Cauchy prior. The three other priors we examine are:

- A normal prior with mean 0 and variance 1.
- A uniform prior with range -2 to 2.
- A bi-modal normal prior, essentially a mixture of two normal priors with means of -2 and 2 respectively and standard deviations of 1 each.

The Cauchy, Normal, Uniform, and bi-modal priors are visualized in the four columns of Figure 7.

We run the sampler for our original data set {-2, -1, 0, 1, 2, 3, 4} and for a second, more substantial data set in which we drew 40 random samples from a Normal distribution with mean 1 and standard deviation 1. Code for running these models is provided below. Note that we use the sample test statistic coding alluded to at the end of section ‘Simulation of Posterior Distribution’ and explained in the Appendix to keep the length of the code block manageable.

```
# Loads a module in JAGS that can deal with a mixture of distributions
load.module("mix")

# We create four different JAGS models
JZS <- list ("model{
  # The same likelihood is used for each of the four models. We chose
  # a more efficient specification of the likelihood (see appendix)
  tstat ~ dnt(delta * sqrt(n), 1, n-1)
  # Prior 1, the Cauchy prior
  delta ~ dt (0, 2, 1)
}",
  "model{
  tstat ~ dnt(delta * sqrt(n), 1, n-1)
  # Prior 2, the Normal prior
  delta ~ dnorm (0, 1)
}",
  "model{
  tstat ~ dnt(delta * sqrt(n), 1, n-1)
  # Prior 3, the Uniform prior
  delta ~ dunif (-2, 2)
}",
  "model{
  tstat ~ dnt(delta * sqrt(n), 1, n-1)
  # Prior 4, the bi-modal Normal prior
  delta ~ dnormmix (c(-2,2), c(1,1), c(1,1))
}")

# The two data sets used for our sample, put into a list
Dum = list (-2:4, rnorm (40, 1, 1))

# We create a list for our four models, each of which will
# be applied to two data sets
FitTstat = list ()

# A counter to put each result in a consecutive slot in the list
Count = 0

# Repeating variable specification for both data sets
for (i in 1:length(Dum))
{
  dat = Dum[[i]]
  m = mean (dat)                          # Sample mean
  n = length (dat)                        # Number of data points
  s = sd (dat)                            # Sample sd
  tstat = (m/s)*sqrt(n)                   # t statistic
  Tstatdata = list (tstat = tstat, n = n) # JAGS variables
  JAGStparam = c("delta")                 # JAGS parameter to be returned

  # Repeating running the jags() function call for each model
  for (j in 1:length(JZS))
  {
    Count = Count + 1

    # Runs the simulations for each data set and each model
    FitTstat[[Count]] = jags (data = Tstatdata,
      parameters.to.save = JAGStparam, n.thin = 1, n.iter = 20000,
      n.burnin = 10000, n.chains = 1,
      model.file = textConnection(JZS[[j]]))
  }
}
```

The resulting eight posterior distributions for the four different priors and two different data sets can be found in Figure 7.

For the remainder of this section, we present BF_{A0} instead of BF_{0A}, indicating the probability of the data under the alternative relative to the null hypothesis. These can be

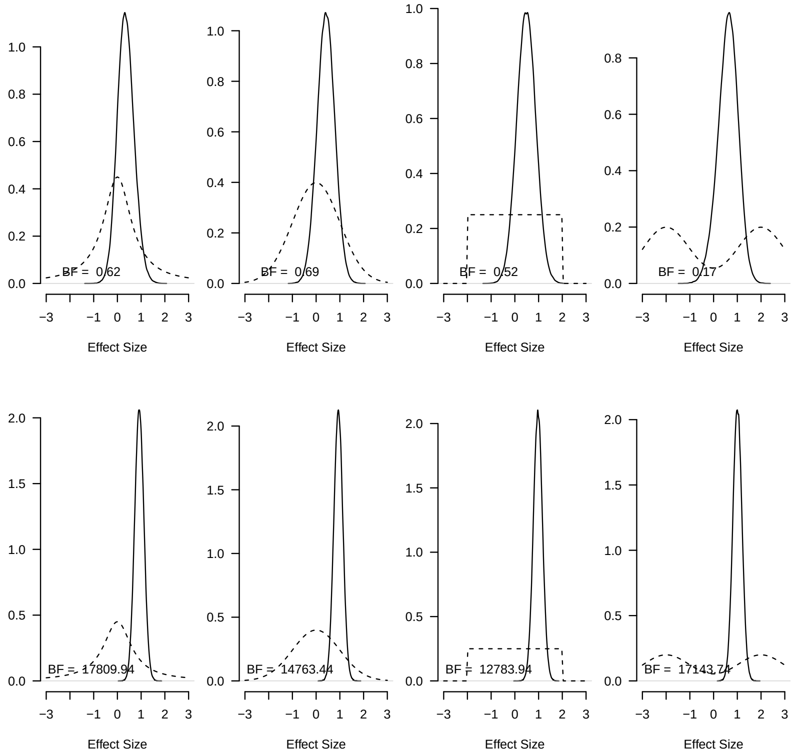

Figure 7. Bayes factors for four different priors (dashed lines), from left to right: Cauchy, Normal, Uniform, bi-modal. Small data set (top), larger data set (bottom). Posterior distributions are indicated by solid lines, Bayes factors are displayed in the bottom-left of each panel. See text for details.

converted to BF_{0A} by calculating $1/BF_{A0}$. Looking first at the smaller overly simplistic data set, we see that the difference in the Bayes factor between the Cauchy and Normal prior is negligible (from 0.62 to 0.69). When we substantially change the prior to a Uniform, the Bayes factor increases a bit to 0.92. Finally, when we radically change the prior to a bi-modal distribution, we get a completely different Bayes factor (well over 0.17).

For the larger data set, we see that the influence of the prior diminishes. The difference in the Bayes factor between the Cauchy and Normal prior is small in relative terms (from about $18,000$ to about $15,000$). When we substantially change the prior to a Uniform, the Bayes factor decreases a bit further to about $13,000$. Finally, when we radically change the prior to a bi-modal distribution, our Bayes factor is about $17,000$.

The small demonstration above is not meant to be an exhaustive simulation, but it should provide the insight that prior distributions that are fairly diffuse (in the sense that they are spread over a wide range of values) and have somewhat similar shapes provide robust outcomes, even for small and strangely distributed data sets. As priors get more extreme, their effect on the Bayes factor becomes more pronounced, especially when data is sparse. Specifically, when priors are highly concentrated in regions of the parameter space that are largely inconsistent with the data, Bayes factors with extreme values can be expected. We would recommend against using such extreme priors unless one has especially strong reasons to do so, such as copious and reliable previous data on the topic.

Conclusion

In this paper, we have attempted to show in an intuitive matter how simulation studies can be employed to learn about statistical or mathematical concepts. We started out by providing code to simulate a normal prior distribution by sampling random values from this distribution.

Next, we transitioned to a description of Bayes factors for quantifying evidence in Bayesian testing. Our exposition was mainly conceptual, eschewing equations for intuition. We discussed a popular implementation, the Jeffreys Zelner Siow (JZS) Bayes factor, which is the Bayes factor used in the BayesFactor package in R and the statistical freeware package JASP. We then used the idea of simulating from a prior distribution to show how generating data from priors under two hypotheses can be used to approximate Bayes factors for hypothetical, unobserved, data sets.

In the subsequent section, we took a different approach to approximating Bayes factors through simulation. It is not always easy to see for social scientists what happens in the black box of programs like the BayesFactor R package and the JASP software program. We approximated their operations by using simulations to approximate areas under posterior distributions. By effectively replacing the point null by a small interval around null, simulations were used to approximate the exact Bayes factor provided by the Savage Dickey method.

Finally, we harnessed the power of simulation to assess the effect of the choice of prior distribution on the Bayes factor for a modest selection of priors and two specific data sets. Such an approach can be implemented by researchers for their specific data set, but we caution against the use of extreme priors without proper a-priori justification.

It is worth pointing out that the JZS Bayes factor is not the only popular implementation of the Bayes factor out there. Other notable candidates include versions by Gönen et al. (2005), the online tool by Dienes (2008) (see `https://medstats.github.io/bayesfactor.html`), and the minimum Bayes factor available in the pCalibrate R package (Ott & Held, 2017).

Note that we do not advocate using simulations for the calculation of Bayes factors to report in scientific articles. Whenever analytical methods exist, they should be preferred over approximating methods such as simulations, much like experiments on a representative sample are unnecessary when the relevant information about the population is known.

With all of this said, in a world where the use of programs like R is increasingly more common among researchers and students, we hope that viewing the JZS Bayes factor through the lens of a simulation approach increases understanding of this ever more popular vehicle for reporting the results of statistical testing.

Author Contributions

DvR and AE jointly generated the idea for the study. DvR wrote the R-code used in the manuscript. AE verified the accuracy of the R-code. DvR and AE drafted the manuscript, and both authors critically edited it. Both authors approved the final submitted version of the manuscript.

Conflicts of Interest

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

Acknowledgements

This research was supported by a Dutch scientific organization VIDI fellowship grant to DvR, which supported the writing of the manuscript (016.Vidi.188.001). AE was supported by a National Science Foundation Graduate Research Fellowship Program (DGE1321846). We thank Quentin Gronau for helpful discussions about the manuscript.

References

Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, *40*, 1550–1577.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6.

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, *13*, 627–679.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*, 7–29.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 204–223.

Dickey, J. M., & Lientz, B. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a markov chain. *The Annals of Mathematical Statistics*, 214–226.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* Macmillan International Higher Education.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Dienes, Z. (2019). How do i know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377.

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, *1*(1), 60–69.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, 2515245918773087.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio–Economics*, *33*, 587–606.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample t test. *The American Statistician*, *59*, 252–257.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2018). Informed Bayesian T-tests. Manuscript submitted for publication.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2019). Informed bayesian t-tests. *The American Statistician.*

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Lawrence Erlbaum.

Heck, D. W. (2019). A caveat on the Savage–Dickey density ratio: The case of computing Bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 316–333.

Jeffreys, H. (1961). *Theory of probability.* Oxford, UK: Oxford University Press.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*, 19313–19317.

Jones, G., & Johnson, W. O. (2014). Prior elicitation: Interactive spreadsheet graphics with sliders can be fun, and informative. *The American Statistician*, *68*(1), 42–51.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*, 1343–1370.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Elsevier Science.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*, 155–177.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406.

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). *Bayes Factor (Version 0.9.12-4.1) [computer software].* Retrieved from `https://CRAN.R-project.org/package=BayesFactor`

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS.* Hoboken: Wiley.

Ott, M., & Held, L. (2017). *pCalibrate (Version 0.1-1) [computer software].* Retrieved from `https://cran.r-project.org/web/packages/pCalibrate`

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 20–22.

R Development Core Team. (2004). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3–900051–00–3)

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t–tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Saunders, B., Milyavskaya, M., Etz, A., Randles, D., & Inzlicht, M. (2018). Reported self-control is not meaningfully associated with inhibition-related executive function: A bayesian analysis. *Collabra: Psychology*, *4*(1), 39.

The JASP Team. (2018). *JASP (Version 0.8.6) [computer software].* Retrieved from `https://jasp-stats.org/`

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*, 217-239.

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo. *Psychonomic Bulletin & Review*, *25*, 143–154.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2012). A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology*, *219*, 1017–2025.

van Ravenzwaaij, D., & Ioannidis, J. P. A. (2017). A simulation study of the strength of evidence in the recommendation of medications based on two trials with statistically significant results. *PLoS One*, *12*, e0173184.

van Ravenzwaaij, D., & Ioannidis, J. P. A. (2019). True and false positive rates for different criteria of evaluating statistical evidence from clinical trials. *BMC: Medical Research Methodology*, *19*, 218.

van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC: Medical Research Methodology*, *19*, 71.

van Ravenzwaaij, D., & Wagenmakers, E.-J. (2020). Advantages masquerading as 'issues' in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019). Manuscript submitted for publication.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, *90*(430), 614–618.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p*-values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E.-J., Gronau, Q. F., Dablander, F., & Etz, A. (2020). The support interval. *Erkenntnis*, 1–13.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. P. P. P. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–159.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... van Doorn, J. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585–603). Valencia: University Press.

Appendix

Simulation of Posterior Distribution Using Test Statistic

Code for obtaining the posterior using the *t*–test statistic is provided below:

```
library (R2jags)

JZStstat <- "model{
  tstat ~ dnt(delta * sqrt(n), 1, n-1)  # Likelihood based on test stat
  delta ~ dt (0, 2, 1)                  # Cauchy prior on delta
}"

dat  = -2:4                             # The data
m    = mean (dat)                       # Mean of data
n    = length (dat)                     # Number of data points
s    = sd (dat)                         # Sd of data
tstat= (m/s)*sqrt(n)                    # Test statistic of data

Tstatdata = list (tstat = tstat, n = n) # Variables to be passed to JAGS
JAGStparam = c("delta")                 # JAGS parameters to be returned

FitTstat = jags (data = Tstatdata, parameters.to.save = JAGStparam,
  n.thin = 1, n.iter = 20000, n.burnin = 10000, n.chains = 1,
  model.file = textConnection(JZStstat))
```

The main changes compared to the example in section 'Simulation of Posterior Distribution' are the following:

- We provide JAGS with the sample test statistic instead of the raw data. For the one sample case, the test statistic is calculated as the sample mean divided by the sample standard deviation times the square root of the sample size, or the sample mean divided by the standard error.
- We need a likelihood for the sample test statistic. In other words, a distribution that characterizes the probability of obtaining specific sample test statistics given an underlying population effect size δ and sample size n. This likelihood is given by a non-central t distribution with non-centrality parameter $\delta \times \sqrt{n}$, precision 1, and $n - 1$ degrees of freedom (see Figure A1, for a visualization).
- We no longer explicitly specify a prior distribution for σ^2. However, this setup implicitly uses the same improper prior for σ^2 as before (see Gronau et al., 2018).

The non-central t distribution may sound fancy, but in practice it boils down to the same thing as the likelihood on each individual data point, because the procedure behind computing a t test statistic is based on the assumption that the data is normally distributed, which is made explicit in the likelihood for individual data points in the first

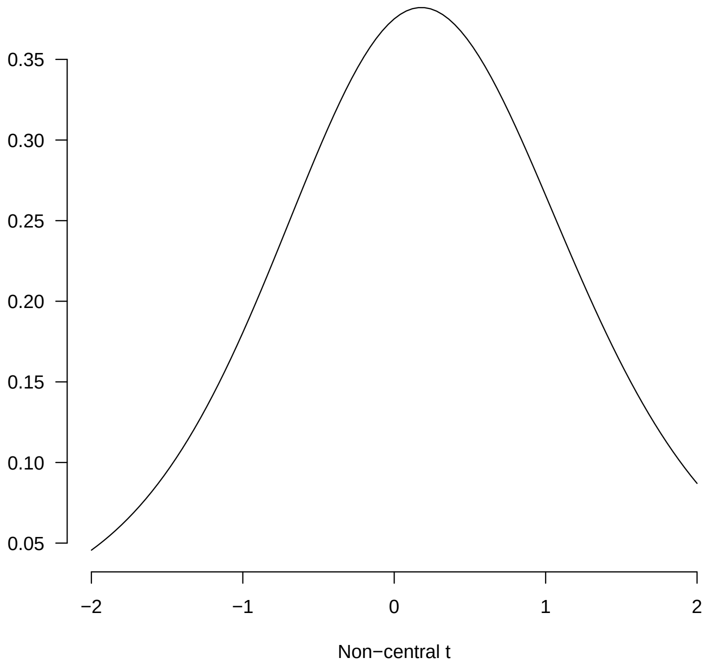

Figure A1. Non-central t distribution non-centrality parameter 0.2, precision 1 and df = 6.

JAGS model. We can use this fact to our advantage when we wish to reanalyze published study results with a default Bayes factor approach in cases where we have access to the sample test statistics, but not the raw data. The procedure for a two-sample t test is very similar, in the JAGS model the reader need only modify the likelihood to factor in the new non-centrality parameter $\delta \times \sqrt(n1 \times n2/(n1 + n2))$ and degrees of freedom $n1 + n2 - 2$.

Similar to our example that used the raw data, we plot the output of the sampler with:

```
Tstatmcmc = as.mcmc (FitTstat)
Tstatdelta = Tstatmcmc[[1]][,"delta"]
plot (density (Tstatdelta, n = 4096), xlim=c(-2,2), bty = ’n’, axes = F,
  xlab = "Effect Size", ylab = "Density", main = "")
axis (1); axis (2, las = 1)
curve (dcauchy (x, 0, sqrt(2)/2), from = -2, to = 2, lty = 2, add = T)
```

To confirm that the output visually matches the output of the previous JAGS script based on the raw data, we use similar plotting code to obtain the output presented in the bottom-right panel of Figure 4.