How to Quantify Support For and Against the Null Hypothesis: A Flexible WinBUGS Implementation of a Default Bayesian t–Test

Ruud Wetzels, Jeroen G. W. Raaijmakers, Emőke Jakab, and Eric-Jan Wagenmakers
University of Amsterdam

Word count: ± 3734 words

Correspondence concerning this article should be addressed to:
Ruud Wetzels
University of Amsterdam, Department of Psychology
Roetersstraat 15
1018 WB Amsterdam, The Netherlands
Ph: (+31) 20–525–6420
Fax: (+31) 20–639–0279
E-mail may be sent to wetzels.ruud@gmail.com.

Abstract

We propose a sampling based Bayesian t–test that allows researchers to quantify the statistical evidence in favor of the null hypothesis. This WinBUGS Savage–Dickey t–test (WSD) is inspired by the Jeffreys–Zellner–Siow (JZS) test recently proposed by Rouder, Speckman, Sun, Morey, and Iverson (in press). The WSD test retains the key concepts of the JZS test but is applicable to a wider range of statistical problems. The WSD test allows researchers to test order–restrictions and applies two–sample situations in which the different groups do not share the same variance.

Keywords: Bayesian inference, t–test, statistical evidence, order–restrictions, Behrens–Fisher problem.

Never use the unfortunate expression "accept the null–hypothesis" – Wilkinson and the Task Force on Statistical Inference (1999, p. 599).

This research was supported by a Vidi grant from the Dutch Organization for Scientific Research (NWO). The WSD computer code and the data of Dr. Smith's hypothetical experiment can be found on the first author's website, `http://www.ruudwetzels.com`. Correspondence concerning this article may be addressed to Ruud Wetzels or Eric–Jan Wagenmakers, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands.

Introduction

Popular theories are difficult to overthrow. Consider, for instance, the following hypothetical sequence of events. First, Dr. John proposes a seasonal memory model (SMM). The model is intuitively attractive and quickly gains in popularity. Dr. Smith, however, remains unconvinced and decides to put one of SMMs predictions to the test. Specifically, SMM predicts that the increase in recall performance due to the intake of glucose is more pronounced in summer than in winter. Dr. Smith conducts the relevant experiment and finds the exact opposite, although the result is not significant. More specifically, Dr. Smith finds that with $n = 41$ the t–value equals 0.79, which corresponds to a two–sided p–value of .44 (see Table 1).

Clearly, Dr. Smith's data do not support SMMs prediction that the glucose–driven increase in performance is larger in summer than in winter. Instead, the data seem to suggest that the null hypothesis is plausible, and that no difference between summer and winter is evident. Dr. Smith submits his findings to the *Journal of Experimental Psychology: Learning, Memory, and the Seasons*. Three months later, Dr. Smith receives the reviews, and one of them is from Dr. John. This review includes the following comment:

> "From a null result, we cannot conclude that no difference exists, merely that we cannot reject the null hypothesis. Although some have argued that with enough data we can argue for the null hypothesis, most agree that this is only a reasonable thing to do in the face of a sizeable amount of data [which] has been collected over many experiments that control for all concerns. These conditions are not met here. Thus, the empirical contribution here does not enable readers to conclude very much, and so is quite weak (...)."

Table 1: Increase in recall performance due to intake of glucose in summer and winter, $t = 0.79$, $p = 0.44$ (NB: hypothetical example)

Season	N	Mean	SD
Winter	41	0.11	0.15
Summer	41	0.07	0.23

In this article, we outline a statistical procedure that allows Dr. Smith to quantify the evidence for the null hypothesis versus the SMM hypothesis. Our work is inspired by the automatic Jeffreys–Zellner–Siow (JZS) Bayesian t–test that was recently proposed by Rouder et al. (in press). Although the JZS test is able to quantify support in favor of the null hypothesis, it does not help Dr. Smith. This is because the prediction of SMM (i.e., the alternative hypothesis) is directional, one–sided, or order–restricted. In other words, SMM does not merely predict that the increase in recall performance differs from summer to winter, but it makes the more specific prediction that the increase in recall performance is *larger* in summer than it is in winter. The JZS test does not directly apply to this scenario. In addition, the JZS two–sample test assumes that both groups share the same variance. When this assumption is violated, the test may no longer be reliable, a phenomenon that statisticians have studied extensively (i.e., the Behrens–Fisher problem,

Kim & Cohen, 1998). To address these limitations, we have developed a flexible sampling based alternative to the JZS test. This alternative procedure, which we name the WinBUGS Savage–Dickey (WSD) test, retains the key concepts of the JZS test but applies to a wider range of statistical problems.

The outline of this article is as follows. First we provide the necessary Bayesian background, and then we discuss the statistical details of Rouder et al.'s JZS test. Next we explain our own procedure, the WSD test, and demonstrate by simulation that it mimics the JZS test—both for the one–sample and two–sample case. Subsequently, we outline two ways in which the WSD test is able to extend the JZS test. First, the WSD test enables researchers such as Dr. Smith to test order– restricted hypotheses (i.e., one–sided t–test). Second, the WSD test can deal with two–sample situations in which the different groups do not share the same variance.

Bayesian Hypothesis testing

In order to keep this article self–contained, we briefly recapitulate the basic principles of Bayesian hypothesis testing (for details see O'Hagan & Forster, 2004; Kass & Raftery, 1995; Myung & Pitt, 1997; Wasserman, 2000). First we explain the concept of *Bayes factors* and then we discuss Rouder et al.'s JZS test on which our method is based.

Bayes factors

In Bayesian inference, competing hypotheses (i.e., statistical models) are assigned probabilities. For instance, assume that you entertain two hypothesis, a null hypothesis H_0 and an alternative hypothesis H_1. Before seeing the data D, these hypotheses have *prior* probabilities $p(H_0)$ and $p(H_1)$. The ratio of these two probabilities defines the *prior odds*. When the data D come in, the prior odds are updated to *posterior odds*, which is defined as the ratio of posterior probabilities $p(H_0|D)$ and $p(H_1|D)$:

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(D|H_0)}{p(D|H_1)} \times \frac{p(H_0)}{p(H_1)}. \tag{1}$$

Equation 1 shows that the change from prior odds to posterior odds is quantified by $p(D|H_0)/p(D|H_1)$, the so–called *Bayes factor*. Thus, Equation 1 reads:

$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds}. \tag{2}$$

When the Bayes factor is, say, 14, this indicates that the data are 14 times more likely to have occurred under H_0 than under H_1, irrespective of the prior probabilities that you may assign to H_0 and H_1. When H_0 and H_1 are equally likely *a priori*, however, a Bayes factor of 14 translates directly to posterior probability—here this means that after seeing the data, H_0 is 14 times more likely than is H_1. Alternatively, one may state that the posterior probability in favor of H_0 equals $14/15 \approx 0.93$, and the posterior probability in favor of H_1 is its complement, that is, $p(H_1|D) = 1 - p(H_0|D) \approx 0.07$.

One of the attractions of the Bayes factor is that it follows the principle of parsimony: when two models fit the data equally well, the Bayes factor prefers the simple model over the more complex one (Berger & Jefferys, 1992; Myung & Pitt, 1997). This fact can be appreciated by considering how the components of the Bayes factor are calculated.

Specifically, both $p(D|H_0)$ and $p(D|H_1)$ are derived by averaging the likelihood over the prior:

$$p(D|H) = \int_{\theta \in \Theta_H} f_H(\theta|D) p_H(\theta) d\theta, \tag{3}$$

where Θ_H denotes the parameter space under the hypothesis of interest H, f_H is the likelihood, and p_H denotes the prior distribution on the model parameters θ. Because complex models have many parameters, and hence a large parameter space, such models have to spread out their prior probability quite thinly over the parameter space. As a result, the occurrence of any particular event will not greatly add to that model's credibility. A prior that is very spread out will occupy a relatively large part of the parameter space in which the likelihood for the observed data is almost zero, and this decreases the average likelihood $p(D|H)$ (Myung & Pitt, 1997).

Rouder et al.'s default Bayesian JZS t*–test*

Consider the one–sample t–test. We assume that the data are normally distributed with unknown mean μ and unknown variance σ^2. The null hypothesis states that the mean is equal to zero, that is, $H_0 : \mu = 0$. The alternative hypothesis states that the mean is not equal to zero, that is, $H_1 : \mu \neq 0$. Denote by BF_{01} the Bayes factor in favor of H_0 over H_1. From Equation 3, the separate components of BF_{01} are given by:

$$p(H_0|D) = \int_0^\infty f_0(D|\mu = 0, \sigma^2) p_0(\mu = 0, \sigma^2) d\sigma^2 \tag{4a}$$

$$p(H_1|D) = \int_{-\infty}^\infty \int_0^\infty f_1(D|\mu, \sigma^2) p_1(\mu, \sigma^2) d\sigma^2 d\mu \tag{4b}$$

These equations feature priors on the model parameters (i.e., p_0 and p_1). Rouder et al. (in press) proposed to define a prior on effect size $\delta = \mu/\sigma$, instead of on the mean μ. Specifically, Rouder et al. (in press) defined a Cauchy prior on δ (i.e., a t–distribution with one degree of freedom), and a Jeffrey's prior (Jeffreys, 1961) on the variance:

$$\delta \sim \text{Cauchy}, \tag{5}$$

$$p(\sigma^2) = 1/\sigma^2. \tag{6}$$

This completes the specification of H_0 and H_1. Rouder et al. (in press) then derive the following equation for the JZS Bayes factor:

$$BF_{01} = \frac{(1 + \frac{t^2}{\nu})^{-(\nu+1)/2}}{\int_0^\infty (1 + Ng)^{-1/2} (1 + \frac{t^2}{(1+Ng)\nu})^{-(\nu+1)/2} (2\pi)^{-1/2} g^{-3/2} \exp -1/(2g) dg}, \tag{7}$$

where t is the t–statistic for the one-sided t–test, N is the number of observations, and $\nu = N - 1$ equals the degrees of freedom.

In order to apply this Bayesian t–test to two–sample designs, Equation 7 needs to be adjusted in three ways: (1) replace the one–sample t– value with the two–sample t–value; (2) calculate N as $N_X N_Y/(N_X + N_Y)$, where X and Y denote the separate groups; and (3) calculate ν as $N_X + N_Y - 2$.

Now recall the data collected by Dr. Smith (see Table 1). Dr. Smith used a within–subject design, and hence a one–sample t–test on the difference scores is appropriate. From the Bayes factor calculator provided on Dr. Rouder's website[1] we obtain a Bayes factor of 6.08—this means that the data are about 6 times more likely under the null hypothesis than under the alternative hypothesis. When we assume that both hypotheses are equally likely *a priori*, we can compute $p(H_0|D)$, the posterior probability for the null hypothesis, as $6.08/7.08 \approx .86$.

Unfortunately, the test developed by Rouder and colleagues does not apply to the problem that confronts Dr. Smith. As mentioned earlier, the SMM predicts that the effect will go in a specific direction—a direction other than the one that is observed in Dr. Smith's experiment. In order to calculate the Bayes factors that are appropriate for a one–sided test, we have developed a sampling based alternative test.

WSD: An MCMC Sampling Based t–test

Calculation of the WinBUGS Savage–Dickey (WSD) t–test involves four steps. The associated computer programs can be found on the first author's website.

Step 1. Rescaling the Data

Prior to the analyses, we rescale the data such that one group has mean 0 and standard deviation 1. This scaling does not affect the test–statistic. For the data from Dr. Smith, for instance, the "summer mean" of 0.07 is subtracted from all observations, both in the winter condition and in the summer condition. Next, all observations are divided by the "summer standard deviation". The main advantage of this rescaling procedure is that the prior distributions for the parameters hold regardless of the scale of measurement: for our Bayesian WSD test, it does not matter whether, say, response times are measured in seconds or in milliseconds.

Step 2. Defining Prior Distributions

We follow Rouder et al. and use a Cauchy prior for effect size δ. For the standard deviation σ we use a half–Cauchy (Gelman & Hill, 2007), that is, a Cauchy distribution that is defined only for positive numbers. This choice for σ is reasonably uninformative, but—in contrast to Jeffrey's prior in Equation 6—the distribution is still "proper" (i.e., the area under the distribution is finite). In the two–sample t–test, we also use a Cauchy prior for the grand mean μ.

[1] `http://pcl.missouri.edu/bayesfactor`.

Step 3. Obtaining Posteriors using WinBUGS

The WinBUGS program[2] (Lunn, Thomas, Best, & Spiegelhalter, 2000; Spiegelhalter, Thomas, Best, & Lunn, 2003) uses built–in Markov chain Monte Carlo techniques (MCMC; Gamerman & Lopes, 2006) to obtain samples from posterior distributions. After specifying the WSD model in WinBUGS, the posterior distribution for effect size δ can be approximated to any desired degree of accuracy by increasing the number of samples. Because the WSD model is relatively simple, we can draw as many as one million samples in a matter of minutes.

Step 4. Calculating Bayes factors using the Savage–Dickey Density Ratio

To obtain the Bayes factor we compute the so–called Savage–Dickey density ratio (e.g., Dickey & Lientz, 1970; O'Hagan & Forster, 2004; Verdinelli & Wasserman, 1995). The Savage–Dickey method applies to nested models, and requires an estimate of the height of the prior distribution and the posterior distribution at the point of interest (i.e., $\delta = 0$), that is:

$$BF_{01} = \frac{p(D|H_0)}{p(D|H_1)} = \frac{p(\delta = 0|H_1, D)}{p(\delta = 0|H_1)}. \tag{8}$$

The height of the prior distribution for δ can be immediately derived from the Cauchy distribution. The height of the posterior distribution can be estimated from the MCMC samples. One method is to use a nonparametric density estimator, but in the work reported here we used a normal approximation to the posterior. As shown in the next two sections, the WSD test closely mimics Rouder et al.'s JZS test.

The One–Sample WSD t–Test: Comparison to Rouder et al.

The one–sample t–test is used to test whether the population mean of one particular sample of observations is equal to zero or not. In experimental psychology, the one–sample t–test is often used for within–subjects designs, in which the scores for two conditions can be reduced to a single difference score.

In order to clarify the structure of one–sample t–test we use graphical model notation (e.g., Gilks, Thomas, & Spiegelhalter, 1994; Lauritzen, 1996; Lee, 2008; Spiegelhalter, 1998). In this notation, nodes represent variables of interest, and the graph structure is used to indicate dependencies between the variables, with children depending on their parents. Double borders indicate that the variable under consideration is deterministic (i.e., they are calculated without noise from other variables) rather than stochastic. Finally, observed variables are shaded and unobserved variables are not shaded. The graphical model for the one–sample t–test is shown in Figure 1.

The graphical model shows that X represents the observed data, that are distributed according to a Normal distribution with mean μ_X and a variance σ_X^2. Because $\delta = \mu_X/\sigma_X$, μ_X is given by $\mu_X = \delta \times \sigma_X$. The null hypothesis puts all prior mass for δ on a single point, that is, $H_0 : \delta = 0$, whereas the alternative hypothesis assumes that δ is Cauchy distributed, $H_1 : \delta \sim$ Cauchy. It is relatively straightforward to implement this graphical

[2]WinBUGS is easy to learn and is supported by a large community of active researchers, see `http://www.mrc-bsu.cam.ac.uk/bugs/`.

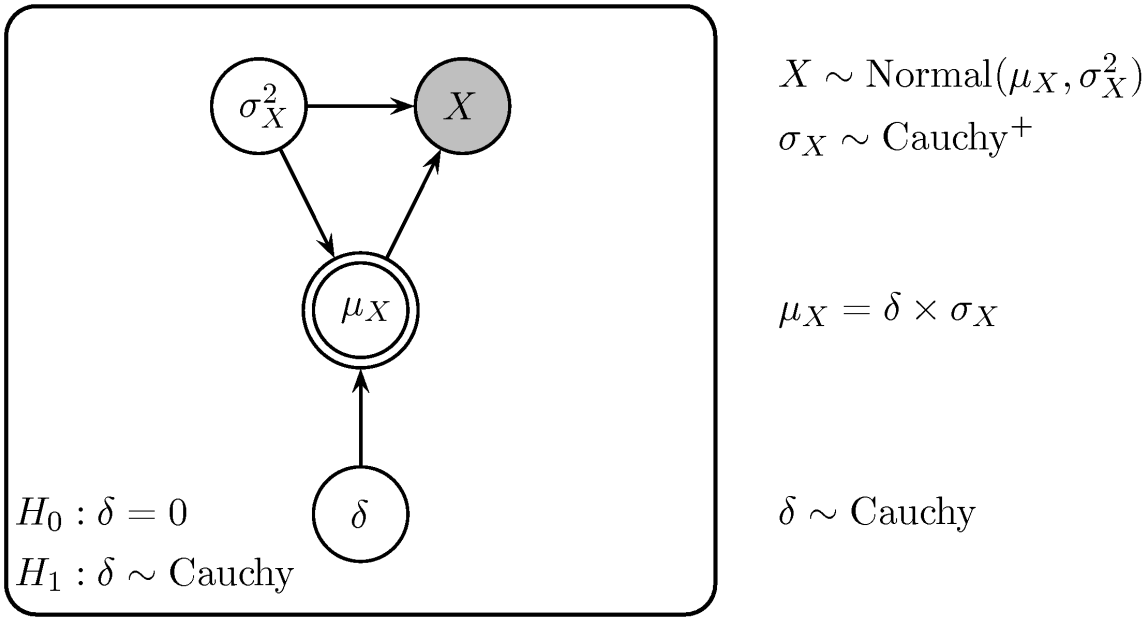


Figure 1. Graphical model for the WSD one–sample t– test. Cauchy$^+$ denotes the half–Cauchy defined for positive numbers only.

model in WinBUGS, obtain samples from the posterior distribution for δ, and carry out the Savage–Dickey test.

To compare this WSD test to Rouder et al.'s JZS test, we conducted a simulation study, the results of which are shown in Figure 2. For each panel, the x–axis gives the t–statistic, and the y–axis gives the associated posterior probability for the null hypothesis, $p(H_0|D)$, derived from the Bayes factor under the assumption that H_0 and H_1 are equally likely *a priori*. Each panel shows the overlap between the JZS test and the WSD test for a specific sample size (i.e., $N \in (20, 40, 80, 160)$), based on 100 simulated data sets. The results demonstrate that for the one–sample scenario, the WSD test closely mimics the JZS test.

The Two–Sample WSD t–Test: Comparison to Rouder et al.

The two–sample t–test is used to test whether the population means of two independent samples of observations are equal to each other or not. In experimental psychology, the two–sample t–test is often used for between–subjects designs.

The graphical model for the two–sample t–test is shown in Figure 3. The graphical model shows that X and Y represent the two groups of observed data. Both X and Y are distributed according to a Normal distribution with shared variance σ^2. The mean of X is given by $\mu + \alpha/2$, and the mean of Y is given by $\mu - \alpha/2$.

Because $\delta = \alpha/\sigma$, α is given by $\alpha = \delta \times \sigma$. As for the one–sample scenario, the null hypothesis puts all prior mass for δ on a single point, that is, $H_0 : \delta = 0$, whereas the alternative hypothesis assumes that δ is Cauchy distributed, $H_1 : \delta \sim$ Cauchy.

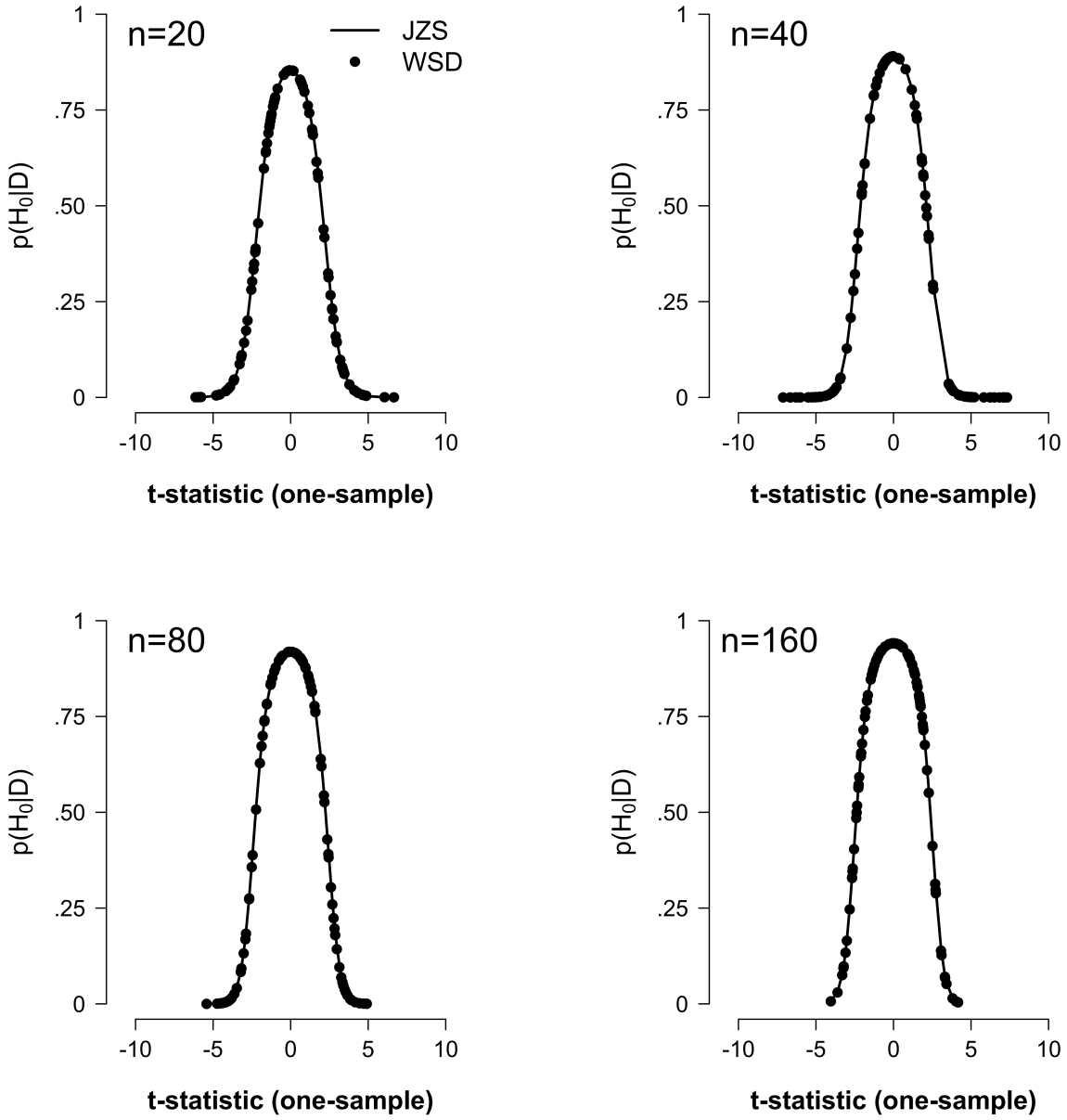


Figure 2. Comparison between the one–sample WSD values and JZS values, for various sample sizes. The black dots represent the WSD values and the solid line represents the JZS values.

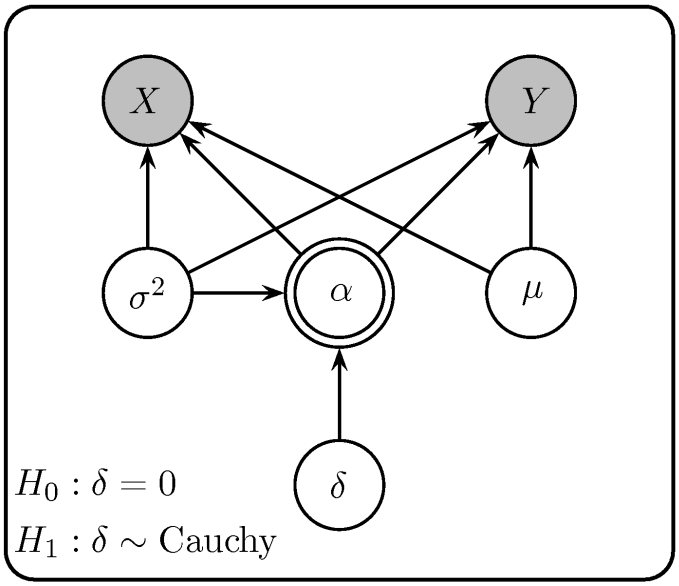


Figure 3. Graphical model for the WSD two–sample t– test. Cauchy$^+$ denotes the half–Cauchy defined for positive numbers only.

To compare this WSD test to Rouder et al.'s JZS test we conducted a simulation study, identical to the one–sample scenario in all respects except for the number of groups. The results of this simulation study are shown in Figure 4. The results demonstrate that for the two–sample scenario, the WSD test closely mimics the JZS test.

Extension 1: Order–Restrictions

Recall once again the experiment by Dr. Smith (see Table 1). The SMM predicted that the effect of glucose would be larger in summer than in winter. We now show how the WSD test can be used to test such order–restricted hypotheses, allowing Dr. Smith to quantify exactly the extent to which the data support the null hypothesis versus the alternative SMM hypothesis.

There exist several Bayesian methods to take into account order–restrictions. Here we use a method that is based on our normal approximation to the unrestricted posterior for effect size parameter δ. The top panel of Figure 5 shows the unrestricted prior and posterior distributions for δ for the data from Dr. Smith. Negative values of δ indicate that the effect of glucose is larger in summer than in winter. From the Savage–Dickey method we can compute the Bayes factor in favor of $H_0 : \delta = 0$ versus the unrestricted alternative $H_1 : \delta \neq 0$, instantiated as $\delta \sim$ Cauchy. Note that the result—$BF_{01} = 6.08$—is identical to the Bayes factor that is obtained from the JZS test: the data are about six times more likely under H_0 than under H_1.

The middle panel of Figure 5 shows the WSD test that applies to the prediction that Dr. Smith seeks to test, that is, $H_0 : \delta = 0$ versus the order–restricted hypothesis $H_1 : \delta < 0$, instantiated as $\delta \sim$ Cauchy$^-$, a half–Cauchy distribution that is defined only

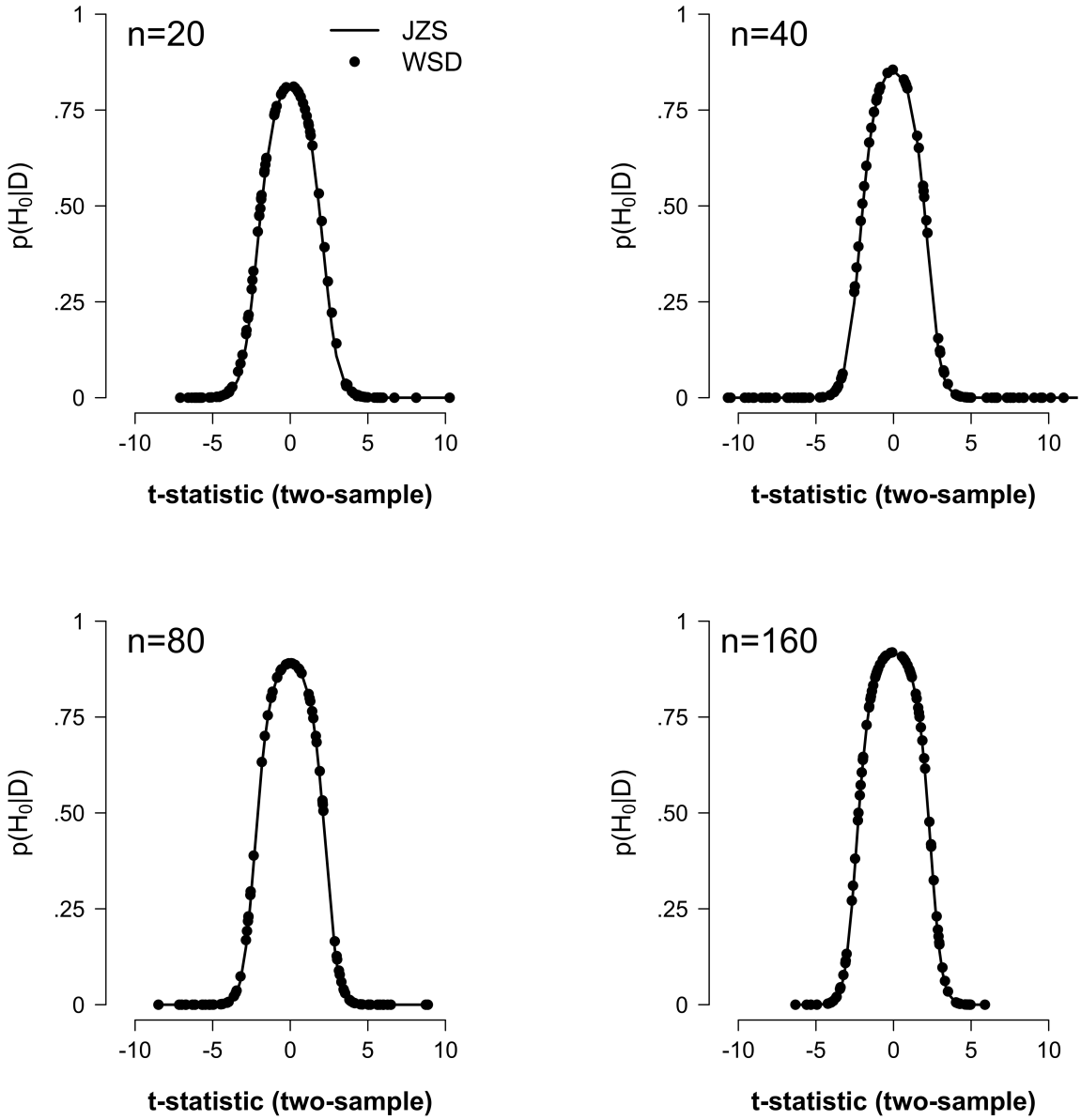


Figure 4. Comparison between the two–sample WSD values and JZS values, for various sample sizes. The black dots represent the WSD values and the solid line represents the JZS values.

for negative numbers. In order to calculate the height of the order–restricted posterior distribution at $\delta = 0$, we focus solely on that part of the unrestricted posterior for which $\delta < 0$. After renormalizing, we obtain a truncated normal posterior distribution that ranges from $\delta = -\infty$ to $\delta = 0$. Figure 5 shows both the half–Cauchy prior (solid line) and the truncated normal posterior (dashed line). The Savage–Dickey ratio at $\delta = 0$ yields a Bayes factor of $BF_{01} = 13.75$. This means that the data are almost 14 times more likely under H_0 than under the order–restricted H_1 that is associated with SMM. When H_0 and H_1 are equally likely *a priori*, the posterior probability in favor of the null hypothesis is about $13.75/14.75 \approx .93$, which is considered "positive evidence" for the null hypothesis (Raftery, 1995; Wagenmakers, 2007).

For completeness, the bottom panel of Figure 5 shows the WSD test for the alternative order–restriction. In this case, we seek to test $H_0 : \delta = 0$ versus $H_1 : \delta > 0$, instantiated as $\delta \sim \text{Cauchy}^+$, a half–Cauchy distribution that is defined only for positive numbers. The Savage–Dickey density ratio yields a Bayes factor of $BF_{01} = 3.91$, which indicates that the data are almost 4 times more likely under H_0 than under H_1.

Extension 2: Variances Free to Vary in the Two–Sample t–Test

For the two–sample scenario, the JZS test assumes that the separate samples share a common unknown variance. When this assumption is false, and both groups have unequal numbers of observations, the JZS t–test may not work as advertised.

This complication (i.e., testing for the difference of two normal means with unequal variances) is known as the Behrens–Fisher problem, and it is one of the oldest problems in statistics. Within the paradigm of p–value hypothesis testing, several solutions to the Behrens–Fisher problem have been proposed (Kim & Cohen, 1998). These solutions (i.e., corrections for unequal variances) have been implemented in popular statistical software packages such as SPSS and R.

In order to address the Behrens–Fisher problem, we adjusted the WSD test in two ways. First, as illustrated in Figure 6, each of the two groups now has its own variance. Second, the previous relation $\alpha = \delta \times \sigma$ no longer holds, as we now have two σ's. We use a standard solution and calculate the pooled standard deviation (Hedges, 1981):

$$\alpha = \delta \times \sqrt{\frac{(\sigma_1^2 \times (n_1 - 1)) + (\sigma_2^2 \times (n_2 - 1))}{n_1 + n_2 - 2}}. \tag{9}$$

After implementing these changes, calculation of the Bayes factor proceeds in the same fashion as before.

To illustrate the behavior of the separate variance WSD Bayes factors, we follow Moreno, Bertolino, and Racugno (1999) and apply the tests to hypothetical data from Box and Tiao (1973, p. 107). These data have the following properties: $n_1 = 20$, $\text{var}_1 = 12$, $n_2 = 12$, and $\text{var}_2 = 40$. As can be seen from Table 2, the support for the null hypothesis decreases as the difference in group means increases. The separate variance WSD test tends to favor the null hypothesis more than does the shared variance WSD test, although the difference is small. The intrinsic Bayes factors (Moreno et al., 1999) supports the null hypothesis the most. A more detailed treatment of the Behrens–Fisher problem is beyond

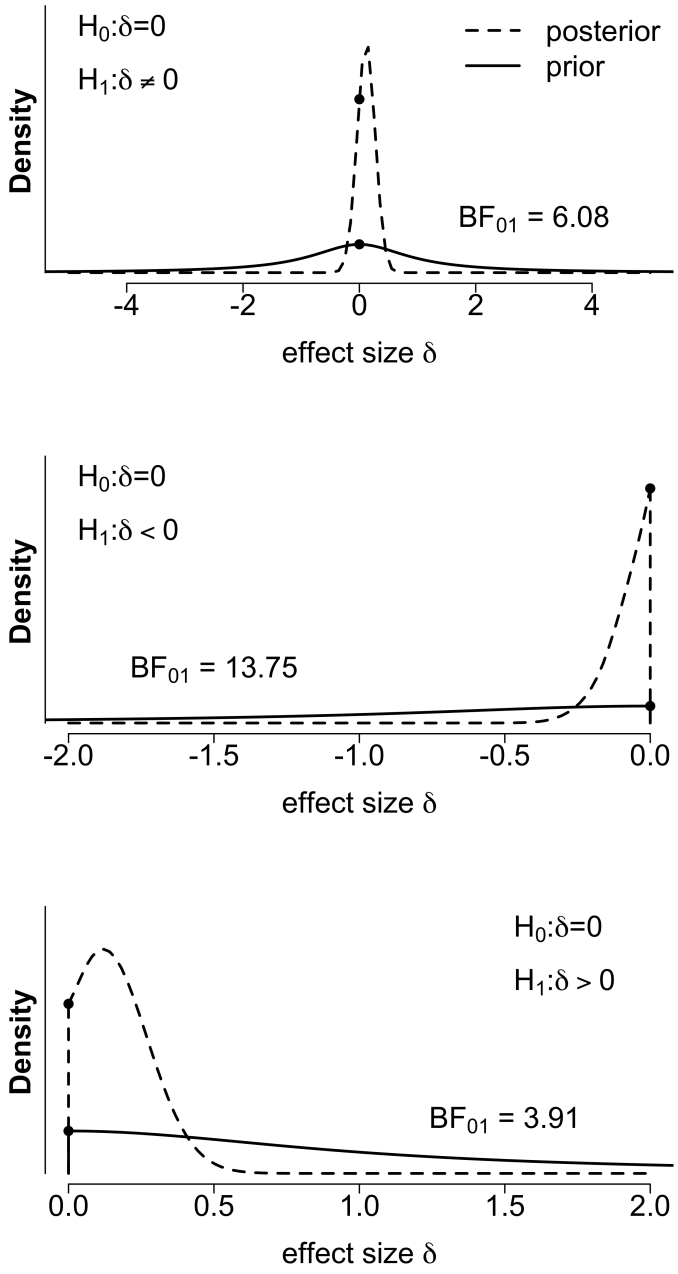


Figure 5. The prior and posterior distributions of effect size δ, based on the data from Dr. Smith (Table. 1). The top panel illustrates the unrestricted WSD test, the middle panel illustrates the order–restricted test associated with the SMM, and the bottom panel illustrates the WSD test for the alternative order–restriction. The dots mark the height of the prior and posterior distributions at δ=0.

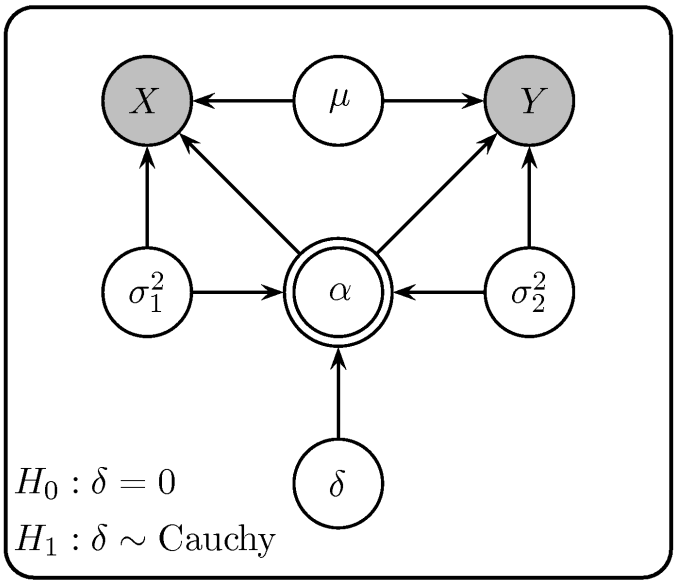


Figure 6. Graphical model for Rouder's default Bayesian two–sided t–test with unequal variances.

the scope of the present article—we include it here only to highlight the flexibility of the WSD test.

Table 2: Comparison of WSD Bayes factors to the intrinsic Bayes factor for hypothetical data reported in Box and Tiao (1973, p. 107) and analyzed in Moreno et al. (1999). Note. $BF_{01}^{\text{WSD1}\sigma}$ denotes the WSD Bayes factor using a shared variance, $BF_{01}^{\text{WSD2}\sigma}$ denotes the WSD Bayes factor using two separate variances, and BF_{01}^{I} denotes the intrinsic Bayes factor reported by Moreno et al.

$\bar{X}$-$\bar{Y}$	$BF_{01}^{\text{WSD1}\sigma}$	$BF_{01}^{\text{WSD2}\sigma}$	BF_{01}^{I}
0.00	3.93	3.36	5.00
2.20	2.08	2.16	2.86
4.22	0.45	0.81	0.76
5.00	0.21	0.51	0.40
10.0	<0.02	<0.02	<0.02

Summary and Conclusion

In this paper we developed a "WSD" Bayesian t–test that extends the Bayesian JZS t–test recently proposed by Rouder et al. (in press). Our sampling based WSD test can handle order–restrictions and addresses the situation in which two groups have unequal variance.

One of the advantages of the WSD test is its flexibility—for instance, it would be trivial to replace the default priors with priors that are informed by previous experiments

or detailed expert knowledge about the problem at hand. We chose to use the Cauchy prior for effect size δ, as proposed by Rouder et al., but many more prior distributions are possible. For example, Killeen (2007) argues that, based on extensive research in social psychology (Richard, Bond, & Stokes-Zoota, 2003), the distribution of effect sizes is Normally distributed with variance 0.3.

Another advantage of the WSD test, and Bayesian methods in general, is that they allow for *sequential inference*. As stated by Edwards, Lindman, and Savage (1963, p. 193), "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience". More concretely, this means that one can apply the WSD t–test and monitor the resulting Bayes factor after every new participant, stopping data collection whenever the evidence is sufficiently compelling. Note that within the paradigm of p-value hypothesis testing, such practice amounts to cheating; with enough time, money, and patience, "optional stopping" is guaranteed to yield a significant result (for a discussion see Wagenmakers, 2007).

In closing, we agree with Rouder et al. (in press) that many scientific hypotheses are formulated in terms of invariances, and that invariances can be formulated in terms of statistical null hypotheses (Wagenmakers, Lee, Lodewyckx, & Iverson, in press). To quantify the statistical evidence in favor of such substantive null hypotheses, we need to move away from p–value hypothesis testing (with which one can only "fail to reject" a null hypothesis) and move toward Bayesian hypothesis testing. In this paper, we have discussed a related problem of considerable scientific importance: a substantive hypothesis (i.e., the SMM) makes a specific prediction, and falsification of the theory requires that one is able to quantify the support in favor of the null hypothesis.

We believe that Bayesian hypothesis testing not only provides a coherent framework to quantify knowledge and uncertainty, but that it also addresses the kinds of questions that experimental psychologists would like to see answered. Bayesian t–tests such as Rouder et al.'s JZS test and our WSD test are the first steps towards a more rational and informative method for testing statistical hypotheses in psychology.

References

Berger, J., & Jefferys, W. (1992). The application of robust Bayesian analysis to hypothesis testing and Occam's Razor. *Statistical Methods and Applications*, *1*, 17–32.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading: Addison–Wesley.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference.* Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, *43*, 169–177.

Hedges, L. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational and Behavioral Statistics*, *6*, 107.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.

Killeen, P. R. (2007). Replication statistics as a replacement for significance testing: Best practices in scientific decision–making. In J. W. Osborne (Ed.), *Best practices in quantitative methods.* Thousand Oaks, CA: Sage Publications.

Kim, S., & Cohen, A. (1998). On the Behrens-Fisher Problem: A Review. *Journal of Educational and Behavioral Statistics*, *23*, 356–377.

Lauritzen, S. L. (1996). *Graphical models.* Oxford: Clarendon.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1–15.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Moreno, E., Bertolino, F., & Racugno, W. (1999). Default Bayesian analysis of the Behrens–Fisher problem. *Journal of Statistical Planning and Inference*, *81*, 323–333.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.).* London: Arnold.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwells.

Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

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

Spiegelhalter, D. J. (1998). Bayesian graphical modelling: A case–study in monitoring health outcomes. *Applied Statistics*, *47*, 115–133.

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS version 1.4 user manual.* Cambridge, UK: Medical Research Council Biostatistics Unit.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, *90*, 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., Lee, M. D., Lodewyckx, T., & Iverson, G. (in press). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Practical Bayesian approaches to testing behavioral and social science hypotheses.* New York: Springer Verlag.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.