Published in final edited form as:
Nat Neurosci. 2020 July 01; 23(7): 788–799. doi:10.1038/s41593-020-0660-4.

Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence

Christian Keysers[1,2], **Valeria Gazzola**[1,2], **Eric-Jan Wagenmakers**[2]

[1]Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands [2]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

Abstract

Most neuroscientists would agree that for brain research to progress, we have to know which experimental manipulations have no effect as much as identify those that do have an effect. The dominant statistical approaches used in neuroscience rely on *p*-values and can establish the latter but not the former. This makes non-significant findings difficult to interpret – do they support the null hypothesis or are they simply not informative? Here we show how Bayesian hypothesis testing can be used in neuroscience studies to establish both whether there is evidence of absence and whether there is absence of evidence. Through simple tutorial-style examples of Bayesian t-tests and ANOVAs using the open source project JASP, this article aims to empower neuroscientists to use this approach to provide compelling and rigorous evidence for the absence of an effect.

Neuroscientists would need to know and publish whether a manipulation does not have an effect as much as whether it does. One may use drugs to block a candidate pathway. If the drug has an effect, that pathway is involved; if it doesn't, one would like to conclude the pathway is not involved. Or one may alter activity in a brain region X and measure behaviour B. If de-activating X changes B, X is involved in B; if B remains unchanged, one would like to conclude X is not involved in B.

Currently, the field of neuroscience is characterized by advanced measurement techniques and sophisticated experimental designs; in contrast, the resulting data analyses almost always employ the standard framework of frequentist statistics, prominently featuring *p*-value Null-Hypothesis Significance Testing (NHST) and confidence intervals. This exclusive reliance on NHST shackles us. NHST is arguably appropriate when the interest lies in quantifying evidence *against* the null hypothesis (H_0: there is no effect), and therefore for the presence of an effect (but see [1]); however, NHST is problematic when quantifying evidence

Correspondence to: Christian Keysers.

Correspondence should be addressed to Christian Keysers, c.keysers@nin.knaw.nl; Meibergdreef 47, 1105BA, Amsterdam, The Netherlands.

Competing Interests: EJ Wagenmakers declares that he coordinates the development of the open-source software package JASP (jasp-stats.org), a non-commercial, publicly-funded effort to make Bayesian statistics accessible to a broader group of researchers and students. CK and VG declare not having any competing interests.

Europe PMC Funders Author Manuscripts

for the null hypothesis. Within the frequentist framework, it is notoriously difficult to establish whether non-significant results support the null hypothesis (i.e., yield evidence for absence) or are simply not informative (i.e., show absence of evidence; [2–4]). NHST biases us to glorify positive effects because those are the effects it equips us to quantify, but to dishonour null findings because we are ill-equipped to convince readers that we have strong evidence for such an absence of an effect. If we agree that the absence of an effect is important information, isn‘t this bias unacceptable? Here we aim to highlight how an alternative statistical framework known as Bayesian inference can resolve this imbalance in neuroscience practice.

We will first illustrate why it is problematic to quantify evidence for the null hypothesis based on the dominant frequentist approaches. We will then show how Bayesian statistics provides a way out of this predicament through simple tutorial-style examples of Bayesian *t*-tests and ANOVAs using the open source project JASP[5].

1 The P-value Predicament

When we conduct a *t*-test to compare two conditions A and B, a resulting *p*-value below a critical threshold α shows that one is unlikely to encounter differences this extreme or more, if the experimental manipulation had no effect (H_0: $\mu_A=\mu_B$). For a fixed sample size, the smaller the *p*, the more evidence we have *against* H_0. Sir Ronald Fisher, one of the first *p*-value proponents, argued that a low *p*-value signals that “either the null hypothesis is false, or an exceptionally rare event has occurred”. (i.e., *Fisher’s disjunction* [6]). In Fisher’s view, the *p*-value is a metric of evidence against the null; and when the null is deemed false, it appears that the alternative hypothesis (H_1: $\mu_A\neq\mu_B$) must then be true[a]. But what if we find no significant effect (e.g., $p=0.3$)? Is that evidence that we have no effect? Apart from sampling variability (i.e., "bad luck"), there are two fundamentally different causal explanations for a non-significant *p*-value: (1) the manipulation had a non-zero effect, but the sample size was too small to detect it (i.e., there was insufficient power); or (2) the manipulation had no effect (i.e., the true effect is zero). When sample size is small, either explanation is plausible. As sample size grows, a non-significant *p*-value increasingly suggests the manipulation really did not have an effect, or at least only one too small to matter. The problem is that the relationship between sample size, power, *p*-value, and evidence for H_0 is complex enough that we are rightly reticent to draw strong conclusions from a non-significant *p*-value. This has been famously and elegantly phrased in the antimetabole: ‘absence of evidence [read: the data are not informative, the design was underpowered] is not evidence of absence [read: the data provide support in favour of the null]’[7].

Intuitively, one may believe that, if lower *p*-values provide more evidence against H_0, higher *p*-values should provide more evidence in favour of H_0. We would thus expect that if we simulate truly random data with no effect, high *p*-values should be relatively frequent, especially with large sample sizes. This, however, is not the case. When we draw random samples from two identical distributions (i.e., where H_0 is true, Figure 1A left-most column),

[a]Fisher warned against this intuitive conclusion, arguing that it requires that H_1 be specified exactly.

p<0.05 is rare, as it should be, but all *p*-values are equally likely. What is 'worse', as sample size increases, and we thus intuitively have more evidence that the two distributions have the same mean, high *p*-values do not become more frequent (Figure 1A, left-most column comparing top and bottom row). Higher *p*-values are thus not a reliable metric for more evidence for H_0.

Hence, NHST leaves the neuroscientist in a peculiar predicament: significant *p*-values indicate evidence against H_0 (but see [1,8]), but non-significant *p*-values are inherently ambiguous and do not allow us to conclude that the data support H_0. This inherent limitation of *p*-values impedes our ability to draw the important conclusion that a manipulation has no effect – and hence that a particular pathway or brain circuitry is not involved, or that a particular stimulus dimension does not matter for brain activity (e.g., [3]). In a way, knowledge becomes like the moon: the side that faces us is familiar, but the other one remains in the dark.

2 A Bayesian Solution

Frequentist NHST focuses exclusively on the null hypothesis (H_0), and *p* simply reflects the likelihood of any data at least as extreme as the observed data under the null hypothesis (orange area in Figure 2D). The alternative hypothesis (H_1) is never used in the *p*-value calculations. In contrast, the Bayesian approach continually updates beliefs based on incoming data, and Bayesian hypothesis testing in particular aims to quantify the relative plausibility of alternative hypotheses (Box 1). This Bayesian approach requires that H_1 be specified with enough precision to also predict the data.

Figure 2 shows an example of how evidence is computed for the case of a *t*-test, where the question of interest is whether or not an experimental manipulation has a positive effect. This translates into two rival hypotheses: the manipulation had no effect versus the manipulation increased the dependent variable. Rather than expressing hypotheses in raw values specific to a given experiment, they are expressed using the population standardized effect size δ (with $\delta=(\mu_A-\mu_B)/\sigma$), which is more comparable across experiments. The sceptic's hypothesis, H_0: δ=0, then states that the effect is absent, whereas the alternative hypothesis, H_+: δ>0, states that the effect is positive (Figure 2A). Note that a "one-tailed" H_1 is denoted as H_+ to indicate the direction of the hypothesized effect. To quantify which hypothesis best predicts the data, we then quantify the observed effect size d ($d=(m_A-m_B)/s$) in the data, and transform it into a t-value ($t=d*\sqrt{n}$) because the distribution of t-values expected for any δ is well known. Next we transform the qualitative hypotheses H_0 and H_+ into quantitative predictions about the probability of encountering every *t*-value using this well-known t-distributions. This is achieved by assigning prior probability distributions to δ (Figure 2B), and then computing the probability of each observable *t* based on these δ-value distributions (Figure 2C). For the sceptic's H_0: δ=0, the distribution of effect sizes is simply a spike at δ=0 (red in Figure 2B), and this makes predictions about the likelihood of each observable *t*-value using the same distribution that is used in a frequentist *t*-test with *n* participants: the student-*t* distribution with *n*-2 degrees of freedom (red in Figure 2C). For H_+:δ>0 we need to be specific about the probability of each possible positive δ to become specific about *t*. The one-tailed nature of our hypothesis is reflected in a truncated

distribution, with negative values having zero probability under H_+ (ref [9] p. 283; note that two-tailed hypotheses are usually implemented by means of symmetrical distributions, e.g. the dotted line in Figure 6B). We also know that most neuroscience papers report effect sizes of $\delta<1$[10] with smaller effect sizes more common than larger effect sizes, which is reflected in a peak for small positive δ, and low probability for $\delta>1$. Indeed, that we feel that we need to perform a test in the first place corresponds to this presumption that the effect size must be fairly small (ref [9] p. 251). Together, this generates the prior distribution shown in blue in Figure 2B (see also Section 4). For each of the hypothesized δ values, we can make predictions about *t* using the noncentral *t* distribution with μ=δ. The mixture of these non-central *t*-distributions associated with each d, weighted by the prior plausibility of that δ, predicts the probability of each possible *t*-value under H_+ (blue in Figure 2C). When the data arrive (Figure 2D), we first calculate the *t*-value for our data, which we will call t_1, and then see where t_1 falls on the *t*-distribution expected under H_0 (red) and under H_+ (blue). The traditional frequentist *p*-value corresponds to the area to the right of t_1 on the red distribution -- note that the predictions from H_+, indicated by the blue distribution, are entirely ignored in the frequentist approach. In contrast, for the Bayesian approach, we take the ordinates $p(t_1 \mid H_0)$ and $p(t_1 \mid H_+)$, and calculate the evidence that the data provide in favour of H_+ over H_0 as $p(t_1 \mid H_+) / p(t_1 \mid H_0)$ (Figure 2E). At that specific t_1 value, the ratio equals 4, indicating that our data was predicted 4 times better by H_+ than H_0; we may conclude that our data supports H_+, albeit only moderately (see Box 1). The evidence --the relative predictive performance of H_0 versus H_+-- is known as the *Bayes factor*[9,11,12]. We abbreviate it as 'BF', and use subscripts to denote which model is in the numerator versus the denominator; thus, $BF_{+0} = p(t_1 \mid H_+) / p(t_1 \mid H_0)$ and $BF_{0+} = p(t_1 \mid H_0) / p(t_1 \mid H_+)$.

Consider the scenario where the *t*-value from our data had been closer to 0, as exemplified by another hypothetical *t*-value, t_2 (Figure 2E). At this new *t*-value the ordinates of the red and blue distributions are about equally high, indicating that the observed t_2 is about equally likely to occur under H_0 and H_+; hence the predictive performance of H_0 and H_+ is about equal, the Bayes factor is near 1, and consequently we have *absence of evidence*. On the other hand, if the *t*-value were to fall at t_3 (Figure 2E), this value would be 4 times more likely to occur under H_0 than under H_+; consequently, $BF_{+0} = ¼$, that is, $BF_{0+} = 4$, and we may conclude that our data support H_0, albeit only moderately -- in other words, we have some *evidence of absence*.

While the *p*-value of a frequentist approach has two logical states – significant or not – which translate into evidence for H_1 ("great, I found the effect") or a state of suspended disbelief ("I did not find an effect, but it could be because I was unlucky, or because the effect does not exist, or because my sample size was too small"), the BF has three qualitatively different logical states: $BF_{10}>x$ ("great, I have compelling evidence for the effect"), $1/x<BF_{10}<x$ ("oops, my data are not sufficiently diagnostic"), $BF_{10}<1/x$ ("great, I have compelling evidence for the absence of the effect"). Here *x* is the researcher-defined target level of evidence. Although the BF should primarily be seen as a continuous measure of evidence, with larger deviations from 1 providing stronger evidence, in Appendix B[9] of his book, Jeffreys proposed reference values to guide the interpretation of the strength of the evidence. These values were spaced out in exponential half steps of 10, $10^{0.5}\approx3$, $10^1=10$, $10^{1.5}\approx30$, to be equidistant on a log scale. He then compared these values with critical

values in frequentist *t*-tests (see Extended Data Figure 1a for a modern equivalent) and χ^2, and declared "Users of these tests speak of the 5 per cent point [p=0.05] in much the same way as I should speak of the K = $10^{1/2}$ [i.e. BF_{10}=3] point,and of the 1 per cent [p=0.01], point as I should speak of the K = 10^1 point [i.e. BF_{10}=10]; and for moderate numbers of observations the points are not very different" (p435 ref [9]). These reference values remain in use, with a BF>3 considered moderate evidence for the hypothesis in the numerator (i.e., H_1 if BF_{10}>3), roughly similar to p<0.05, and BF>10 considered strong evidence (roughly similar to p<0.01)[13]. Because BF_{10}=1/BF_{01}, this automatically also defines the bounds for evidence for the hypothesis in the denominator: BF<1/3 is moderate and BF<1/10 is strong evidence. BF values between 1/3 and 3 indicate that there is insufficient evidence to draw much of a conclusion for or against either hypothesis. While these guidelines allow us to reach somewhat discrete conclusions, the magnitude of the BF should be considered as a continuous quantity, and the strength of the conclusions expressed in the discussion section of a paper should reflect in tone the magnitude of the BF. For new discoveries, Jeffreys suggested that x=10 is more appropriate than x=3, but each scientist and field will need to decide whether to privilege the sensitivity of the test for small samples or effects, by using smaller critical values such as 3, or to avoid false decisions by using higher x values such as 10. Whichever decision a scientist will take, readers can judge the strength of the evidence directly from the continuous numerical value of BF, with a BF twice as high providing evidence twice as strong. In contrast, with p-values, it can be difficult to interpret the actual p-value as strength of evidence, with p=0.01 certainly not five times as much evidence than p=0.05.

Crucially, the three-state system of the Bayes factor allows us to differentiate between evidence of absence and absence of evidence. This represents a fundamental conceptual step forward in the way we interpret the data we so painstakingly collect. Instead of one outcome (i.e., p<α) that generates knowledge, we now have two (i.e., BF_{10}>x and BF_{01}>x). Both sides of the moon are now in the light.

Figure 1B shows how a Bayesian *t*-test performs under the same scenario in which Figure 1A was analysed. In this figure we used a target level of evidence equal to x=3, as this was proposed to roughly correspond to the α-level of .05 we use for the frequentist *t*-test in panel A[9]. The figure shows that when an effect is absent (μ=0), as in the frequentist approach, the Bayesian test will seldom come to the erroneous conclusion that an effect is present (less than 4% BF_{+0}>3). However, unlike the frequentist approach --in which larger p-values do not become more frequent when larger sample sizes intuitively provide more evidence of absence -- the Bayesian *t*-test provides increasing evidence for the absence of an effect (see green percentages in Figure 1B), as sample size grows. The same is true when an effect is present: evidence for an effect increases as sample size or effect size increases (Extended Data Figure 1b). Hence, unlike the frequentist p-value, the BF has the symmetric property to quantify evidence for the presence or the absence of an effect that scale with evidence in either direction -- be it due to increased sample size or effect size. In each case, inconclusive cases (i.e., absence of evidence, defined here as 1/3<BF<3) become increasingly rare as sample size increases.

Another aspect that can be appreciated from Figure 1B, is the statistical power to provide evidence for or against an effect. When an effect is absent, evidence of absence ($BF_{+0}<1/3$) in the presence of noise is limited when sample size is very small (40% at n=5), but reasonable in sample sizes often used in neuroscience (n=20-100). When an effect is present, evidence for the presence of an effect ($BF_{+0}>3$) is slightly less frequent than that of the frequentist approach ($p<0.05$), but not dramatically different. However, as sample sizes become very large, the Bayes factor and p values diverge more dramatically, with p values becoming significant even for arguably irrelevantly small effect sizes (e.g. at n=1000, d=0.05, $t(999)=d\sqrt{1000}$, p=0.05), while the BF continues to require more relevant effect sizes (Extended Data Figure 1b). It should be noted that for two-tailed tests, evidence for the null hypothesis becomes substantially harder to provide and requires larger sample-sizes because the predictions of the null are directly flanked by the high likelihood of finding small effect sizes in either directions under H_1.

If Bayesian inference is so simple and rewarding, why isn't it used more often in the neurosciences? We speculate that one of the main reasons for the slow adoption of Bayesian inference is pragmatic: until relatively recently it was difficult to conduct Bayesian analyses for standard statistical scenarios; even a Bayesian t-test required an understanding of a monograph such as Jeffreys's *Theory of Probability* [9], which is accessible only to those with considerable statistical training. In order to remove this hurdle, a number of packages have been developed to make Jeffreys-style Bayesian hypothesis tests easier to perform. Here, to make Bayesian testing available to the widest community of neuroscientists we concentrate on the multi-platform open-source program JASP (Jeffreys's Amazing Statistics Program; jasp-stats.org) that uses an accessible graphical user interface, but the R-package BayesFactor[14] is a powerful alternative. The next section illustrates how JASP allows neuroscientists to conduct comprehensive Bayesian analyses with ease.

3 JASP - A Convenient Tool for Bayesian Inference

The main goal of JASP is to popularize Bayesian inference by removing the pragmatic hurdles that have so far frustrated its broader adoption. In the JASP graphical user interface, data files are uploaded in a variety of formats (including .csv formats easily exported from excel and the .jasp format), analyses are selected from simple drop-down menus, variables are dragged and dropped into windows, and output is generated on the fly (Movie 1 at https://osf.io/md9kp/). Increasingly detailed analyses can be executed by ticking checkboxes. As a result, for many statistical scenarios, a comprehensive Bayesian (re)analysis can be executed in a matter of seconds. The examples below showcase the ways in which the output from such Bayesian analyses should be interpreted and how they allow researchers to go beyond the conclusions from the classical frequentist p-values. The supplementary materials contain a number of csv files associated with the examples presented below, as well as a commented code in the language R to replicate the BF values we obtain with JASP for power users that favor command line coding in order to apply an analysis to a large number of units (e.g., to classify hundreds of neurons recorded using calcium imaging into those responding and those not responding to a particular stimulus).

Europe PMC Funders Author Manuscripts

3.1 Example of a two-sample *t*-test

To illustrate the Bayesian *t*-test we use an example inspired by Carrillo et al., 2019. In Carrillo et al., we hypothesized that the rat anterior cingulate cortex (ACC) is critical for emotional contagion and that deactivating the ACC by locally injecting muscimol should thus reduce emotional contagion compared to injecting saline. The injected animal observed a conspecific receive electroshocks (ShockObs), and its freezing was measured as an index of emotional contagion. To explore specificity, the same animals underwent a non-social control condition in which they were exposed to a tone previously associated with shocks (CS playback). To illustrate how to analyze this kind of design using Bayesian statistics, we generated two synthetic data sets (Muscimol1.csv and Muscimol2.csv that can be found at https://osf.io/md9kp/) that illustrate two slightly different scenarios. We use simulated data rather than the actual data from the paper to guide the reader though alternative scenarios and to allow the reader to modify the data and test the effect this has on the analysis (using the r-script GenerateMuscimolData.R, that can be found at https://osf.io/md9kp/).

Movie 1 can be found at https://osf.io/md9kp/ and shows how to setup the analyses in JASP to examine the data of Muscimol1.csv. Our main analyses of interest are two independent sample *t*-tests on the freezing measures that compare H_+:Saline>Muscimol against H_0:Saline=Muscimol separately for the ShockObs and CS. To establish specificity, we will later use an ANOVA. We use a one-tailed alternative hypothesis because deactivating the ACC should reduce (not enhance) freezing in the muscimol condition and hence lead to higher freezing in the saline condition. The frequentist approach can also be performed in JASP using 'Independent Samples T-Test' to encourage scientists to combine frequentist and Bayesian approaches on the same data set using a single package.

The frequentist approach confirms that for ShockObs, muscimol reduced freezing significantly ($t_{(38)}$=3.961, p<0.001), i.e., the observed (or larger) difference in freezing are highly unlikely under H_0. For CS, the result is nonsignificant ($t_{(38)}$=-0.519, p=0.7), which could signal evidence for absence or absence of evidence. To adjudicate between these alternative interpretations, we now perform the 'Bayesian Independent Samples T-Test'. Similarly, to the frequentist approach we selected ShockObs and CS as dependent variables, Group as Grouping Variable, and the one-tailed Group1>Group2 analysis (after selecting Saline as group1 and Muscimol as group2 in the data viewer as shown in Movie1). The results are shown in Figure 4.

In the input panel on the left-hand side, we selected BF_{10} as the output, i.e., p(data | H_+) / p(data | H_0), with a one-tailed hypothesis of Group1[saline]>Group2[muscimol]. The results table in the panel on the right summarizes the main outcomes. For ShockObs, BF_{+0}=162.282, signalling that the data are 162 times more likely under H_+ than under H_0. The data thus provides what is considered extremely strong evidence for our hypothesized reduction in socially triggered freezing following ACC deactivation. For CS, BF_{+0}=0.223. This value is below 1/3, and according to the classification scheme by Jeffreys [9,15] our data thus provide moderate evidence for H_0, i.e., that ACC deactivation does not lead to a reduction of non-socially triggered freezing. Switching to option BF_{01} in the left-hand panel

inverts the Bayes factor; now BF_{0+} for CS equals 4.494 (1/0.223), meaning that the data are 4.5 times more likely under H_0 than under H_+.

For the muscimol2 data, the frequentist *t*-test again reveals a significant reduction in ShockObs ($t_{(38)}$=3.8, p<0.001) and a non-significant result for CS ($t_{(38)}$=1.2, p=0.11). The Bayesian analysis confirms that the data provide extremely strong evidence for a reduction of freezing for ShockObs (BF_{+0}=120). However, this time, for CS, BF_{+0}=0.97. This result indicates an absence of evidence (in contrast to Muscimol1, which showed some evidence of absence), with the data being about equally likely under H_+ and H_0.

3.2 Example of an ANOVA

We can also examine whether muscimol had a greater effect on ShockObs than on CS, a question that is different from asking whether one has and the other does not show an effect, by examining if we have evidence for an interaction between group (saline vs muscimol) and condition (ShockObs vs CS) [16,17]. In a frequentist approach, we can conduct this analysis using the JASP 'Repeated Measures ANOVA' (rmANOVA) menu option. The results show a significant main effect of Condition ($F_{(1,38)}$=14.6, p<0.001), of Group ($F_{(1,38)}$=5.4, p=0.026), and importantly a significant Condition x Group interaction ($F_{(1,38)}$=14.3, p<0.001). We can also perform this analysis using the 'Bayesian Repeated Measures ANOVA' menu option (Figure 5), the functionality of which is based on the 'BayesFactor' R package [18].

The Bayesian approach to the rmANOVA is to compare the predictive performance of models with and without each of the factors and interactions. Conceptually, it starts from a null model that predicts data based on a constant for each subject without considering any experimental factors. It computes the likelihood L_{null} of that null model, i.e. the probability of the observed data D under this null model. It then also calculates the likelihood L_{Group} of a model additionally including an effect of Group. If the Bayes factor calculated as L_{Group} / L_{null} is >1there is evidence for the factor Group. If BF<1, i.e. the null model outperforms the more complex Group model, there is evidence for the absence of an effect of Group. If BF≈1 we have absence of evidence. This Bayes factor can be interpreted using the same bounds discussed in Figure 2 and Extended Data Figure 1.

Complex models always *fit* data at least as well as simpler models. How can a simpler model thus ever outperform a more complex model in the Bayesian sense? The answer is simple: a Bayes factor model comparison does not compare the *fit* of models for a specific parameter value (i.e., the maximum likelihood) but the predictive performance of models across all plausible parameter values (i.e. average likelihood) [19–21]. If we consider the models D = Subject + β*Group (i.e. the Group model) and D = Subject (i.e., the null model), the average likelihood of the data under the models is the weighted average of the probability of the data D under the full range of plausible values assigned to β in the parameter prior: $L = \int P(D|\beta)P(\beta)d\beta$. Hence, the null model's L is calculated entirely atβ=0, whereas the Group model's L considers β=0, but averaged with the predictions from all other plausible β values. The effect of this integration over β can be appreciated in Extended Data Figure 2.

Essentially, because the null model concentrates all its predictions on β=0, small differences across the two groups are more likely under this null model, providing evidence for absence.

Figure 5 applies this logic to our data. The top table in the output panel of indicates all the models that are being considered and compared. This includes the abovementioned Null model with subject constants only, a model that adds only the effect of Condition, one with only Group, one with both main effects, and one also including the interaction. The *P*(M) column indicates the prior probability of these various rival models, which is set to be equal not to influence the outcome of the test. Note that this *model* prior probability reflects how likely you believe each model to be true and is different from the *parameter* prior distribution that characterizes each model. Next, we see how likely each model is after having seen our data, *P*(M|data). Here we see that the full model with the interaction (Condition + Group + Condition*Group) is by far the most likely (*P*(M|data)=0.983). The following columns indicate the relative likelihood of each model compared with the average of all other models (BF_M), or compared with the best or worst model (BF_{10}). Specifically, BF_M for the null model for instance is calculated as P(M|data) for the null model divided by the average of the P(M|data) over all other models. For BF_{10}, the calculation depends on what is chosen in the menu 'Order'. If one selects "Compare to best model", as we did in Figure 5, the best model is shown first, and all other BF_{10} express likelihood relative to that best model. If you select "Compare to null model", the models are shown with the Null model on top, and all other BF_{10} represent the P(M|data) of that model divided by that of the null model, and can be read as describing how much more likely that model is than the null model. Switching to BF_{01} then inverts the BF and expresses how much better the null model is than that particular alternative model. The error column finally estimates the margin of error in the BF computation.

Now that we know that amongst the tested models the full model is the most likely in the light of our data, we still wonder which of its components improved its predictive performance. The simplest way to explore this question systematically is to select the 'Effects' option, which generates the second, 'Analysis of Effects' table (Figure 5). This analysis uses the P(M|data) column of the model comparison above to quantify the contribution of each component. When selecting the default option 'across all models', for each component, the BF_{incl} (last column) is calculated as *p*(models with that factor|data)/*p*(models without that factor|data). For Condition for instance, BF_{incl} is the average P(M|data) for all models with Condition (i.e.,Condition, Condition + Group and Condition + Group + Condition*Group) divided by that of all models without Condition (i.e., Null model and Group). Selecting "across matched models", restricts the comparison to models that only differ in the presence/absence of a particular component, and for Condition BF_{incl} is then the average P(M|data) for Condition and Condition + Group divided by the average P(M|data) of their matched models, i.e., models identical except for the absence of Condition, namely the Null and Group models. In this calculation, the interaction model is not included in the nominator, because it lacks a matched model Group + Condition*Group. We recommend the 'matched model option', as it provides a more conservative estimate of each factor's contribution.

This effects table then allows us to draw inferences about the contribution of each factor and interaction in the spirit of a traditional ANOVA. BF_{incl} for Condition (similar to the main effect of condition) is 37.5, indicating that the models including the factor Conditions are much (37.5 times) more likely than those not including it. The BF_{incl} for Group (main effect of Group) is 1.7, showing that models with Group are marginally more likely than those without that main effect, but the evidence is too weak to be conclusive. BF_{incl} for the interaction is 96, meaning that the full model with the interaction is 96 times more likely than that without. This effect of interaction provides extremely strong evidence for the fact that deactivating the ACC has a much stronger impact on ShockObs than on the CS condition. Performing the same analysis on Muscimol2, in which we had inconclusive evidence with regard to a reduction of freezing in the CS muscimol condition (BF_{+0}=.97) however provides no evidence for an interaction (BF_{incl}=1.16, i.e., absence of evidence), showing that in Muscimol2, we remain uncertain whether deactivating the ACC impairs freezing in the CS condition (because the *t*-test BF_{+0} is inconclusive), and about whether deactivating the ACC has a stronger effect on ShockObs than CS. Had we found a BF_{incl}<1/3, we would have had evidence of absence: namely that muscimol has the same effect on ShockObs and CS.

4 Default Priors Provide an Objective Anchor

As shown in Figure 2, to calculate a Bayes factor we have to specify H_1 such that its predictive adequacy can be assessed. We are generally uncertain about the true value of the parameters (such as effect size), and this uncertainty is reflected in a prior distribution across the parameter values. Defining this prior distribution introduces the dreaded element of "subjectivity", one that scientists fear jeopardizes the objectivity and generalizability of their inferences (e.g.,[22] but see [23]). There is however a simple two-step solution: first, use a default prior that is designed to fulfil general statistical desiderata[24]; then, check how much your inference is robust against motivated changes in the prior.

For the *t*-test and ANOVA, there is broad consensus on certain parameter priors being appropriate under most circumstances. We recommend using these parameter priors to increase the objectivity of the analyses, and provide a common frame of reference that ensures the direct comparability of Bayes factors from different experiments. Indeed, these defaults are implemented in JASP (and in the BayesFactor package in R for those that prefer a command line environment), and in section 3 we performed all our inferences so far without even thinking about prior distributions. However, it is informative to consider these parameter priors in more detail.

For the t-test exemplified in section 3.1., the default prior is the Cauchy distribution with a scale parameter of $r = \sqrt{2}/2 \approx 0.707$ as shown in Figure 2 and 4. A Cauchy distribution resembles a Gaussian distribution but has fatter tails. The prior specifies the a priori plausibility of each effect size, and the default specifies that half the effect sizes are within the scale parameter, i.e., ±0.707, with smaller effect sizes more likely than larger effect sizes. For ANOVAs, the parameters are also assumed to follow a Cauchy prior distribution, but their scale depends on the type of factor one explores (fixed effects r=0.5, random effects r=1, and covariates r=0.354, see [19] for details).

To examine the effect of changing the width of that prior distribution in our *t*-test example, it suffices to select the option 'Bayes factor robustness check' to generate the plots of Figure 6A. The default width of the prior distribution for *t*-tests is the above mentioned Cauchy with scale 0.707 [18], and which prior is used can be seen (and changed) by pulling down the 'Prior' option on the bottom left panel (Figure 4). The robustness graph on the top of Figure 6A then shows how BF_{+0} changes as a function of the prior scale or width, with the one set in the menu 'Prior' shown as the 'user prior' at the gray circle. Wider priors (wide, black circle and ultrawide, empty circles), assume that larger effects are more likely than the default prior and then show how the BF_{+0} changes if one uses these less informative priors. We term wider priors less informative because if one knows nothing about what effect size to expect, all effect sizes would be equally likely a priori, and the prior would be infinitely wide). For ShockObs (Figure 6A left), evidence for H_+ is extremely high for all but the narrowest prior distributions, and our conclusion that deactivating the ACC reduces freezing is thus robust against reasonable changes in the prior. For CS (Figure 6A right), evidence favours H_0 also robustly across all but the narrowest prior distributions. In both cases, such robustness is reassuring, and warrants confident conclusions. In contrast, when conclusions vary dramatically across a range of reasonable prior distributions, then caution may be in order. Note that when the scale parameter is zero, H_+ reduces to H_0, and the Bayes factor equals 1 regardless of the data, explaining why all robustness lines will converge to 1 for the narrowest prior distributions.

Selecting the option 'Prior and posterior and additional info' outputs the results shown in Figure 6B for our one-tailed hypothesis. Under H_+, the half-Cauchy prior distribution is shown as a dotted line, and the posterior distribution after taking the data into account is shown as the black line. This posterior shows the effect size distribution after updating the prior based on the data (Box 1 and 2). The posterior median and credible interval summarize the Bayesian estimate of the effect if H_+ holds (median δ=1.109, 95% credibility interval: [0.406, 1.810]). This effect size estimate is not simply the Cohen's *d* observed in the sample (which equals 1.24), but it reflects the posterior distribution that is a combination of prior distribution and data (Box 1). The Cauchy prior distribution assumes that small effect sizes are more likely than large effect sizes, and this knowledge exerts a small pull towards zero on the sample estimates, a reasonable and conservative approach, leading to the Bayesian point estimate of δ=1.1 (using the median, and assuming H_+ is true). For small sample sizes, the estimate will be more influenced by the prior, whereas for larger sample sizes, the estimate will approach the sample value *d*. This property is desirable in the way it counteracts the systematic overestimation of effect sizes in frequentist approaches with low power[25]. For CS (right), we see that the posterior is folded at zero because of our one-tailed hypothesis which implies that negative effect sizes are impossible. For parameter estimation of d, we recommend to adopt a two-tailed hypothesis by clicking on "Group1≠Group2", leading to estimates that are more suitable to report as effect size estimates (second row). Note that for the Muscimol1 column, the posterior distribution for effect size is mostly unaffected by whether a two-sided or a one-sided prior distribution is used; in contrast, the Bayes factor against the null hypothesis is about twice as high for the one-sided analysis as for the two-sided analysis (i.e., BF_{+0} = 162 and BF_{10} = 81).

We recommend to report the median and 95% credible interval (abbreviated as 95% CI) in addition to the BF because it provides complementary information. For instance, for ShockObs, the BF_{+0} reveals strong evidence for the presence of an effect, but it leaves unaddressed how strong the effect is. This is because the same effect size δ will lead to different BF values at different sample sizes (Extended Data Figure 1b). The 95% CI provides us with information about this effect size – namely that the effect for ShockObs is probably very large (as suggested by the median δ=1.1) and that we can be quite confident that it exceeds δ=0.4 (lower bound of the 95% CI). If one looks for effects of clinical relevance, for instance, knowing that a manipulation has an effect in a group of 1000 patients (as revealed by the BF) is often less interesting than knowing how strong the effect is likely to be (as revealed by the CI). That the 95% CI does not include δ = 0, is further indication for the presence of an effect. For CS, the BF_{+0} provides evidence for the absence of an effect. In such cases, considering the 95% CI is perhaps less relevant, because the CI only makes sense under H_1. However, the bounds of the CI specify that even if H_+ were true (despite the observed data being 4 times more likely under H_0) the effect size is unlikely to exceed 0.4 (upper bound of the CI), and is likely to be very small (median=-0.12). This then informs the kind of group size one would need to systematically study such an effect. That the CI includes δ = 0 is in line with the notion that the data reflect the absence of an effect; however unlike the BF, the CI alone cannot help us distinguish absence of evidence from evidence for absence. If scientists prefer to see the CI in the original units of measurement, the bounds should be multiplied by the population standard deviation σ.

For the ANOVA, extracting credible intervals of effect sizes in JASP is work in progress[26]. In the meantime, post-hoc Bayesian *t*-tests could be performed to obtain Bayesian CI for specific contrasts of interest, or the effect size (e.g., $\eta 2$) of the corresponding frequentist ANOVA could be reported.

The effect of the directionality of H_1 on the BF and posterior distribution is important. In frequentist statistics, one-tailed hypothesis testing is sometimes frowned upon - if one focuses on the risk of false positives, a more conservative two-tailed statistics is arguably preferable, and the only difference is typically that *p* values double. With Bayesian statistics, the focus shifts to giving H_1 and H_0 a more balanced 'chance', and the ability to provide evidence for H_0 becomes an important consideration. In that context, if we hypothesize a specific direction of effect (e.g., that injecting muscimol into the ACC should reduce freezing in response to ShockObs but not CS), we strongly recommend to test this directional hypothesis with the appropriate directional H_+ effect size prior distribution. The reason is particularly apparent in small group sizes: with n=8, under a two-tailed Bayesian one-sample *t*-test, *t*-values> 2.8 (corresponding to δ~0.8) can provide evidence for H_1 (BF_{10}>3), but even t=0 (the datum with the highest evidence for H_0) falls short of providing modest evidence for H_0 (BF_{01}=2.97). Using the theoretically appropriate H_+ resolves this imbalance, as even small negative *t*-values can provide evidence for H_0 over H_+ (e.g.,t=-0.3, BF_{0+}=3.62). One-tailed testing is thus typically a fairer balance between the ability to provide evidence for H_0 and H_1.

Finally, it is important to consider that some scenarios do call for using user-defined priors (see [23] for a more extensive discussion of how to create informed priors). For instance, to

test a claim that a candidate drug has an effect size δ>0.8, default priors are unhelpful. Custom priors with H_0: δ<0.8 vs H_1: δ>0.8 would need to be specified, and their likelihood compared.

5 Accumulation of Evidence

While conducting experiments, a difficult question is often how many participants to acquire. By selecting the option 'Sequential Analysis' we can see how the BF changes as one considers an increasing number of data points in our Bayesian *t*-test examples (Figure 3 and Figure 4C). For Muscimol1, we observe a clear upward trend to ShockObs in favour of H_+ and a downward trend to CS playback in favour of H_0. Such consistent trends provide confidence in the effect *a posteriori*. Importantly, we could have performed this analysis during data collection to avoid predefining our sample size (see Supplementary Materials Sequential Testing). We could have used a data collection plan in which we collect a minimum of n=20 animals (10 per group) at first, and then we would have kept adding new animals to the saline and muscimol group until the BF_{+0} crosses a predetermined critical value (e.g., BF_{+0}>6 or BF_{+0}<⅙) or until a preset maximum of animals (e.g., n=40) is depleted. In our example, we would then have stopped at n=20 animals in the ShockObs condition and continued until n=40 animals in the CS condition. This would have saved n=20 animals while warranting the same conclusions. As discussed in the supplementary materials, such an approach is unacceptable in NHST. One reason for this difference is that Bayesian statistics can provide evidence for H_0 and H_1, while NHST can only provide evidence against H_0. Hence, testing until a significant result is found in NHST will per definition always find evidence against H_0 -- such "testing to a foregone conclusion" does not hold for Bayesian testing.

For Muscimol2, evidence for CS playback shows no steep up- or downwards trend. This is diagnostic of small effect sizes. The second half of the curve suggests a mild up-wards trend, and extending this trend shows that hundreds of animals would probably have to be added to reach a conclusion. This is in line with how our data has been generated (see Figure 3), with a traditional power analysis suggesting that hundreds of animals would be necessary to detect such small effect sizes.

6 Reporting Both Frequentist and Bayesian Results -- A Plea for Statistical Inclusivity

One concern for aspiring Bayesian neuroscientists is that reviewers in neuroscience journals will be relatively unfamiliar with Bayes factors; for instance, reviewers may be more impressed by p<0.01 than by BF_{10}=10.3. This argument is well taken and our pragmatic recommendation is therefore to consistently report the frequentist and Bayesian statistics jointly (e.g., $t_{(38)}$=3.961, p<0.001, BF_{+0}=162, with median posterior δ =1.1, 95%CI=[0.4, 1.8]). Where evidence for H_1 is to be presented, using a p-value with a standard frequentist test is perhaps the pragmatic thing to do and adding the BF_{10} provides an additional quantification. Finally, the effect size estimate helps the reader estimate how strong an effect the manipulation caused. Where there is no evidence for H_1, a nonsignificant p-value cannot indicate whether the data is evidence of absence (i.e., evidence for H_0) or the absence of

evidence (i.e., inconclusive); then the use of BF_{01}presents an attractive way to adjudicate between these two conceptually different alternative interpretations. This hybrid approach is a powerful opportunity to reap the best of both statistical approaches. Borderline cases in which frequentist and Bayesian approaches do not quite concur, e.g., p<0.04, suggesting a significant effect but BF_{10}=2.3, suggesting only anecdotal evidence, are cases in which reporting both is fair and invites us to refrain from discussing the effect in too strong a tone, and also reporting the CI on the effect size is informative, as at high *n,p* values and BF dissociate for small effect sizes.

In terms of details of reporting, we recommend to report the results as follows. In the statistical analysis section of the methods, write: “Differences across the muscimol and saline groups were analysed using the Bayesian Independent Samples T-Test as implemented in JASP vXXX using default effect size priors (Cauchy scale 0.707). Results are reported using the one-tailed Bayes factor BF_{+0} that represents p(data|H_+:Saline>Muscimol) / p(data|H_0:Saline=Muscimol). Effect size estimates are reported as median posterior Cohen’s δ with 95% credibility interval using a two-tailed H_1 in order not to bias estimates in the expected direction. In the results section for Muscimol1, we would then write: "We found extreme evidence for a reduction of freezing in the Muscimol compared to the Saline group for ShockObs ($t_{(38)}$=3.961, p<0.001, BF_{+0}=162.282, with median posterior δ = 1.11, 95%CI=[0.42, 1.795]) and moderate evidence for the absence of a reduction for the CS ($t_{(38)}$=-0.519, p=0.7, BF_{+0}=0.223, with median posterior δ =-0.133, 95%CI=[-0.712, 0.414]”. In the discussion, one can then say “Our data supports the notion that the ACC is involved in socially triggered freezing and that the ACC is not involved in freezing triggered by a CS”.

In the results of Muscimol2 instead we would write: “We found extreme evidence for a reduction of freezing in the Muscimol compared to the Saline group for ShockObs ($t_{(38)}$=3.8, p<0.001, BF_{+0}=120, with median posterior δ =1.07, 95%CI=[0.427, 1.765]). For the CS condition, results were inconclusive ($t_{(38)}$=1.2, p=0.11, BF_{+0}=0.98, with median posterior δ =0.31, 95%CI=[-0.23, 0.90]), which suggests that the data are equally likely under H_0 and H_1”. In the discussion: “Although it remains unclear whether or not muscimol injection to the ACC reduces freezing following CS playback, if there is an effect, it is relatively small. In contrast, our data strongly support our hypothesis that muscimol reduces freezing during the ShockObs condition, and the effect size appears to be large.”

For the ANOVA, in the Methods section we recommend writing: “Bayesian ANOVAs were conducted using JASP with default priors, and effects are reported as the Bayes factor for the inclusion of a particular effect, calculated as the ratio between the likelihood of the data given the model with vs the next simpler model without that effect”. In the results, for Muscimol1, we would write: “A repeated measures ANOVA revealed strong evidence for the presence of an interaction of Group*Cond ($F_{(1,38)}$=14.3, p<0.001, BF_{incl}=239)”. In the discussion, we could then add "our data further provides strong evidence for the specificity of the involvement of the ACC, in that the effect of Muscimol injection in the ACC was substantially stronger during ShockObs than CS playback". For Muscimol2, we would write “… revealed inconclusive evidence regarding the presence of an interaction of Group*Cond ($F_{(1,38)}$=3.4, p=0.072, BF_{incl}=1.16)”. In the discussion, in the case of Muscimol2, in which we could not draw conclusions about CS using the *t*-test, we now also conclude “it remains

unclear whether muscimol injection to the ACC reduces freezing following CS playback and it remains unclear whether muscimol reduces freezing less in the CS than the ShockObs condition".

7 Concluding Comments

Bayesian inference offers unique practical advantages for neuroscience. Above all, Bayes factors provide a continuous and symmetric measure of statistical evidence. The Bayes factor can support H_0 as much as it can support H_1. At a time in which we have become acutely aware of the devastating impact of publication biases and *p*-value hacking, it offers us the tools to become less biased by making evidence for absence as solid as evidence for presence. With software like JASP and the BayesFactor package for R, calculating Bayes factors for the most used tests in neuroscience has become as easy as calculating *p*-values, encouraging us to harvest the added information afforded by Bayesian analyses with little additional effort.

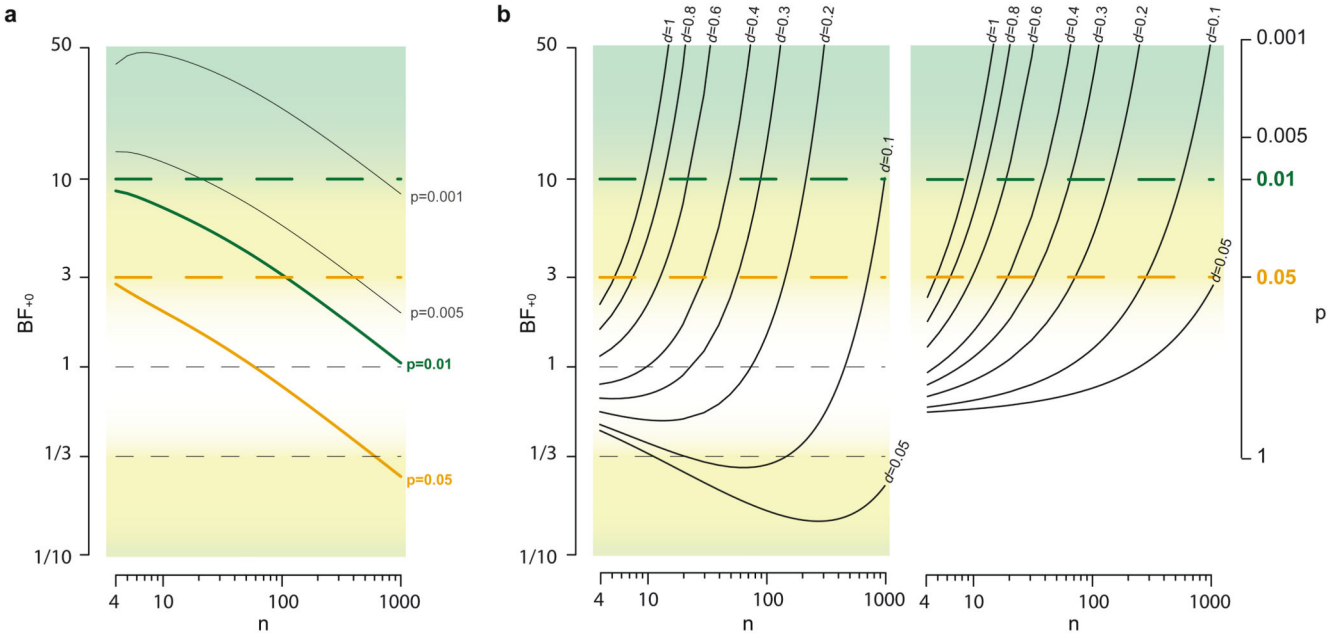
We have presented some examples of scenarios in which Bayesian statistics have become turn-key and simple to adopt in neuroscience. Some applications will however require more development. Applications like neuroimaging often require statistical testing over hundreds of thousands of voxels, and frameworks for the correction for multiple comparisons are still in their infancy for the Bayes factor. Other datasets are highly non-normal in their distributions. Just like parametric frequentist approaches, the Bayesian t-test and ANOVA we leveraged here assume normality. Non-parametric Bayesian tests so far only exist for certain applications (e.g., some t-tests and regressions have a tick-mark for non-parametric approaches in JASP, and R code exists for a number of additional cases[27]), but remain in development for others (e.g., ANOVAs).

As a field, we have been slow to take up Bayesian statistics out of a perception that Bayesian hypothesis testing is difficult to perform and to interpret. With the emergence of new software and accessible packages, performing Bayesian equivalents of the most prevalent tests has become easy. By supplementing our frequentist approaches with Bayesian analyses, we come to richer interpretations that allow more informative conclusions. Null findings become interpretable and publishable. We have a chance to shed light on the hitherto dark side of our scientific enterprise.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

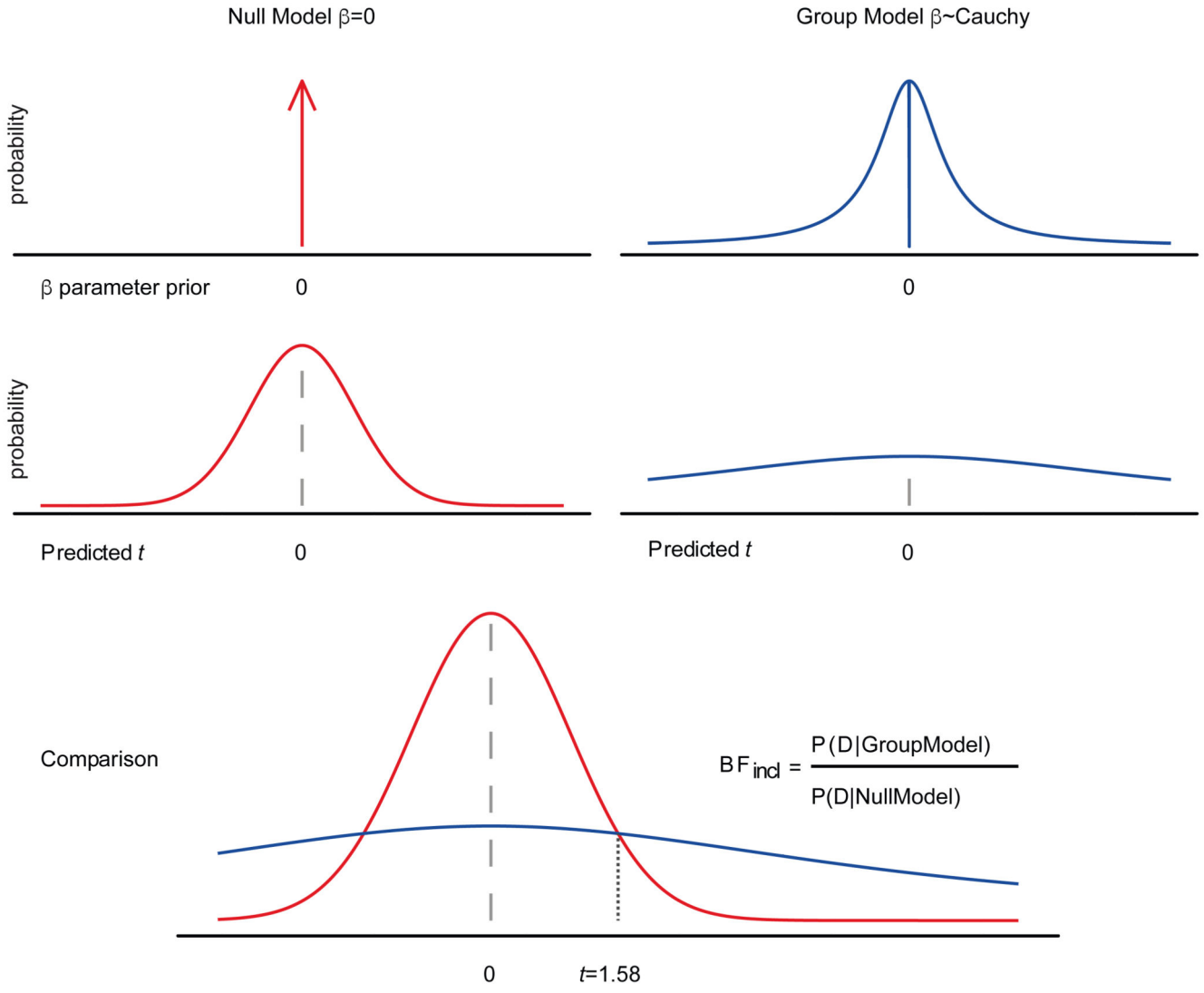
Extended Data



Extended Data Figure 1. The relationship between BF, p, and effect sizes values. (a)

This log-log plot shows the BF_{+0} values corresponding to familiar critical *p* values for a one-tailed one-sample *t*-test at different sample sizes (*n*). The curves show the BF_{+0} values obtained in a Bayesian *t*-test based on the critical *t*-value that provides p=0.05 (yellow), p=0.01 (green), p=0.005 (black) and p=0.001 (black). The yellow dashed horizontal line indicates the BF_{+0}=3 bound for moderate evidence considered by Jeffreys[9] to be similar to p=0.05, the green one the BF_{+0}=10 for strong evidence considered similar to p=0.01. The two black dashed lines mark BF_{+0}=1, i.e. the line of no evidence, and BF_{+0}=1/3, the bound for moderate evidence of absence. The background gradient reminds the reader that the BF reference values of 3 and 10 should not be considered hard bounds. Instead the BF should be interpreted as a continuous value, with values diverging more from 1 supporting stronger conclusions. This panel makes two points. First, there is no simple equivalence between *p* and BF that holds over all sample sizes. This is because in a frequentist *t*-test, the observed effect size (*d*) sufficient to generate a specific *p* value decreases with √n more rapidly than for the BF. As a result, at large *n*, very small effect sizes generate 'significant' *t*-test: at n=1000, the critical *t*-value for a one-tailed p=0.05 is 1.65, corresponding to d=1.65/√n =0.05. For the BF, such a minuscule effect is 4 times more likely under H_0 than H_+ (BF_{+0}=0.26). Hence, for small sample sizes *p* and BF support similar conclusions (e.g., p=0.05 at n=4 corresponds to BF_{+0}>3, supporting the same conclusion of evidence for an effect), but for large sample sizes the frequentist and Bayesian conclusions can diverge in the presence of very small effect sizes (e.g., p=0.05 at n=1000 corresponds to $BF_{+0}<1/3)_{p207,\ 38}$. Considering confidence or credible intervals of the effect size in addition to *p* or BF values helps interpret such cases. Second, the fact that the dashed lines are above the curve of the same colour for all n>4 shows that BF_{+0}=3 and BF_{+0}=10 indeed protect against Type I errors in a frequentist sense at least at p=0.05 or p=0.01, respectively. In other words, if BF_{10}>3,

p<0.05, and if BF_{10}>10, *p*<0.01, but how much lower than 0.05 or 0.01 the exact *p*-value is, depends on *n.* **(b) BF** $_{+0}$ (left) and *p* (right) values as a function of measured effect- and sample-sizes. These panels illustrate the measured effect sizes necessary to provide evidence for an effect at different sample sizes in a one-sample one-tailed t-test using the BF vs. traditional p values. Each curve connects the results at different sample sizes for the specified value of *d*. The logarithmic BF and p scales are aligned so as to place BF=3 next to *p*=0.05, and BF=10 next to *p*=0.01.



Extended Data Figure 2. Evidence for or against a factor in a Bayesian ANOVA

A Bayesian ANOVA is a form of model comparison. This figure illustrates how the Bayes factor can provide evidence for a simpler model by concentrating its predictions on a single parameter value. This example ANOVA determines whether or not the data D depend on the value of the factor Group by comparing the Null Model D=0*Group (left) against the Group Model D=β*Group, with a Cauchy prior on β (right). The top row illustrates the prior probability attributed to the different values of β under the two competing models. Note how

both models include β = = 0 as a possibility, but given that the probability values must integrate to 1 over the entire β space, for the Null Model $p(\beta = 0) = 1$ while for the Group Model, the probability is distributed across all plausible alternative values. The middle row shows the predicted t-values based on these priors, where t represents the difference between the data from the two groups as in Figure 2. Note how these predictions are more peaked for the Null compared to the Group model. The bottom row compares the predicted probability of finding particular *t*-values under the two models, and shows how values close to zero (i.e., small or no difference between the groups) are predicted more often by the Null compared to the Group Model, while the opposite is true for large *t*-values. If conducting the experiment reveals a measured *t*-values close to zero, the Bayes Factor for including the factor Group would be substantially below 1, providing evidence for the absence of an effect of Group, while the inverse would be true for high *t*-values

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

CK is funded by NWO VICI grant 453-15-009, VG by ERC grant 758703 and NWO VIDI grant 452-14-015. We thank Frantisek Bartos for help with Figure 2.

Data Availability

All data, code and figures can be downloaded at: https://osf.io/md9kp/.

References

1. Benjamin DJ, et al. Redefine Statistical Significance. Nat Hum Behav. 2018; 2:6–10. [PubMed: 30980045]
2. Dienes Z. Using B ayes to Get the Most out of Non-Significant Results. Front Psycholology. 2014; 5:781.
3. Gallistel CR. The Importance of Proving the Null. Psychol Rev. 2009; 116:439–453. [PubMed: 19348549]
4. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t Tests for Accepting and Rejecting the Null Hypothesis. Psychon Bull Rev. 2009; 16:225–237. [PubMed: 19293088]
5. Love J, et al. JASP: Graphical statistical software for common statistical designs. J Stat Softw. 2019; 88:1–17.
6. Wagenmakers, E--J, , et al. >The Need for Bayesian Hypothesis Testing in Psychological SciencePsychological Science Under Scrutiny: Recent Challenges and Proposed Solutions. Lilienfeld, SO, Waldman, I, editors. John Wiley and Sons; 2017. 123–138.
7. Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. BMJ. 1995; 311:485. [PubMed: 7647644]
8. Edwards W, Lindman H, Savage LJ. Bayesian Statistical Inference for Psychological Research. Psychol Rev. 1963; 70:193–242.
9. Jeffreys, H. Theory of Probability. Oxford University Press; 1961.
10. Szucs D, Ioannidis JPA. Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature. PLOS Biol.
11. Etz A, Wagenmakers E--JJBS. Haldane's Contribution to the Bayes Factor Hypothesis Test. Stat Sci. 2017; 32:313–329.

12. Kass RE, Raftery AE. Bayes Factors. J Am Stat Assoc. 1995; 90:773–795.
13. Lee, MD, Wagenmakers, E--J. Bayesian Cognitive Modeling: {A} Practical Course. Cambridge University Press; 2013.
14. Morey RD, Rouder JN. BayesFactor. 2018; 0.9:12–4.2.
15. Jeffreys H. Theory of Probability. Oxford University Press. 1939
16. Nieuwenhuis S, Forstmann BU, Wagenmakers E--J. Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance. Nat Neurosci. 2011; 14:1105–1107. [PubMed: 21878926]
17. Gelman A, Stern H. The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. Am Stat. 2006; 60:328–331.
18. Morey RD, Rouder JN. Bayes Factor Approaches for Testing Interval Null Hypotheses. Psychol Methods. 2011; 16:406–419. [PubMed: 21787084]
19. Rouder JN, Morey RD, Speckman PL, Province JM. Default {B}ayes Factors for {ANOVA} Designs. J Math Psychol. 2012; 56:356–374.
20. Rouder JN, Engelhardt CR, McCabe S, Morey RD. Model comparison in ANOVA. Psychon Bull Rev. 2016; 23:1779–1786. [PubMed: 27068543]
21. Myung IJ, Pitt MA. Applying {O}ccam’ Razor in Modeling Cognition: {A} {B} yesian Approach. Psychon Bull Rev. 1997; 4:79–95.
22. Efron B. Why Isn’t Everyone a Bayesian? Am Stat. 1986; 40:1–5.
23. Lee MD, Vanpaemel W. Determining Informative Priors for Cognitive Models. Psychon Bull Rev. 2018; 25:114–127. [PubMed: 28194721]
24. Bayarri MJ, Berger J, Forte A, Garcia-Donato G. Criteria for Bayesian Model Choice With Application to Variable Selection. Ann Stat. 2012; 40:1550–1577.
25. Cremers HR, Wager TD, Yarkoni T. The relation between statistical power and inference in fMRI. PLoS One. 2017; 12:e0184923. [PubMed: 29155843]
26. Marsman M, Waldorp L, Dablander F, Wagenmakers EJ. Bayesian estimation of explained variance in ANOVA designs. Stat Neerl. 2019; 73:351–372. [PubMed: 31341338]
27. van Doorn J, Marsman M, Ly A, Wagenmakers E--J. Bayesian Latent Normal Inference for the Rank Sum Test, the Signed Rank Test, and Spearman’s ρ. Manuscr Submitt Publ. 2017
28. Wagenmakers E--J, Morey RD, Lee MD. Bayesian Benefits for the Pragmatic Researcher. Curr Dir Psychol Sci. 2016; 25:169–176.
29. Sutton, RS, Barto, AG. Reinforcement Learning: An Introduction. The MIT Press; 1998.
30. Wrinch D, Jeffreys H. On Certain Fundamental Principles of Scientific Inquiry. Philos Mag. 1921; 42:369–390.
31. Rozeboom WW. The Fallacy of the Null--Hypothesis Significance Test. Psychol Bull. 1960; 57:416–428. [PubMed: 13744252]
32. Stefan AM, Gronau QF, Schönbrodt FD, Wagenmakers E--J. A Tutorial on Bayes Factor Design Analysis using an Informed Prior. Behav Res Methods. 2019; 51:1042–1058. [PubMed: 30719688]
33. Wagenmakers E--J, et al. Bayesian Inference for Psychology. Part II: Example Applications with JASP. Psychon Bull Rev. 2018; 25:58–76. [PubMed: 28685272]
34. Rouder JN. Optional Stopping: No Problem for Bayesians. Psychon Bull Rev. 2014; 21:301–308. [PubMed: 24659049]
35. Schönbrodt FD, Wagenmakers E--J. Bayes Factor Design Analysis: Planning for Compelling Evidence. Psychon Bull Rev. 2018; 25:128–142. [PubMed: 28251595]
36. Consonni G, Fouskakis D, Liseo B, Ntzoufras I. Prior Distributions for Objective Bayesian Analysis. Bayesian Anal. 2018; 13:627–679.
37. Gronau QF, Ly A, Wagenmakers E--J. Informed Bayesian t-tests. Am Stat. 2019
38. Jeffreys H. Some Tests of Significance, Treated by the Theory of Probability. Proc Cambridge Philos Soc. 1935; 31:203–222.

Box 1

Bayesian Updating

The Bayesian formalism describes how an optimal observer updates beliefs in response to data. In the context of hypothesis testing, at the start, observers entertain a set of two or more rival accounts. In the context of a t-test, they would be called hypotheses H_0 and H_1, in the case of an ANOVA, they would be called models. Each is specified via parameters we can call θ, e.g. the effect size δ in a t-test hypothesis or a regression parameter β in an ANOVA. Prior to looking at the data, the rival accounts have prior probabilities and the parameter values within each account have prior probabilities. At the level of the accounts, we may assume them to be equally believable a priori (e.g. prior hypothesis probabilities $P(H_0)=P(H_1)=0.5$). At the level of the parameters within each account, they are associated with prior parameter distributions (e.g. H_0: δ=0, H_1: d~Cauchy, see Figure 2). When data become available, the probabilities are reallocated: Accounts and parameters-within-accounts that predict the data relatively well receive a boost in credibility, whereas those that predict the data poorly suffer a decline[28]. Note the similarity to models of reinforcement learning[29]. Mathematically, this updating is done using Bayes' rule, as described below, separately for parameters and accounts.

Updating parameter estimates

$$\underbrace{p(\theta|\text{data})}_{\substack{\text{Posterior beliefs}\\ \text{about } \theta}} = \underbrace{p(\theta)}_{\substack{\text{Prior beliefs}\\ \text{about } \theta}} \times \underbrace{\frac{p(\text{data}|\theta)}{P(data)}}_{\substack{\text{Predictive}\\ \text{updating factor}}}.$$

Here, the probability of each possible value of θ within an account after seeing the data (i.e. posterior parameter beliefs) are calculated as the product of the prior probability of that value (i.e. parameter prior beliefs) times the predictive updating factor. The latter reflects how likely the observed data is according to that particular parameter value divided by the average predictive performance across all values of θ weighted by their prior probability, i.e. $p(\text{data})= \int p(data \mid \theta) \cdot p(\theta)\, d\theta$. The result is the basis for the Credible Intervals that the Bayesian analysis provides for the parameters conditional on a given model

Updating the plausibility of the rival accounts

For two rival accounts of the data (e.g., H_0 vs H_1), Bayes' rule can best be written in the form of odds[30]:

$$\underbrace{\frac{p(\mathcal{H}_0|data)}{p(\mathcal{H}_1|data)}}_{\text{Posteriors odds for } \mathcal{H}_0 \text{ vs. } \mathcal{H}_1} = \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\text{Prior odds for } \mathcal{H}_0 \text{ vs. } \mathcal{H}_1} \times \underbrace{\frac{p(data|\mathcal{H}_0)}{p(data|\mathcal{H}_1)}}_{\text{Predictive updating factor}}.$$

This equation shows that the change from prior hypothesis odds to posterior hypothesis odds is brought about by the predictive updating factor -- commonly known as the *Bayes factor*[12]. For instance, assume the rival hypotheses are equally plausible a priori (i.e.,

$P(H_0)=P(H_1)=0.5$). The prior hypothesis odds are then equal to one. If the predictive updating factor is 10 (i.e., the observed data is 10 times more likely under H_0 than under H_1), this means that the posterior odds are then also 10. Given that for mutually exclusive hypotheses $P(H_0)+P(H_1)=1$, these odds mean that the data have increased the probability of H0 from 0.5 (the prior hypothesis probability) to $10/11 \approx 0.91$ (the posterior H_0 probability).

The Bayes factor quantifies the degree to which the data warrant a change in beliefs, and it therefore represents the strength of evidence that the data provide for H_0 vs. H_1. Note that this strength measure is symmetric: evidence may support H_0 just as it may support H_1 -- neither of the rival hypotheses enjoys a special status.

For a neuroscientist deciding whether or not their manipulation had an effect, the posterior odds might *seem* like the most obvious metric, as they reflect the plausibilityof one hypothesis over another after considering the data. However, these posterior odds depend both on the evidence provided by the data (i.e. the Bayes factor) and the prior odds. The latter capture subjective believes prior to the experiment and introduce an often-undesirable element of subjectivity that could bias the conclusions drawn from the posterior believes. Scientists who embrace a certain theoretical standpoint and those who do not, might fiercely disagree on these prior odds while agreeing on the evidence, that is, the extent to which the data should change their beliefs. With beliefs considered less valuable for scientific reporting than evidence, the data-informed Bayes factor makes is the less controversial and thus favoured metric to report.

There are three broad qualitative categories of Bayes factors. First, the Bayes factor may support H_1; second, the Bayes factor may support H_0; third, the Bayes factor is near 1 and supports neither of the two rival hypotheses. In the second case we have “evidence of absence”, and in the third care we have “absence of evidence” (see also [2]). More fine-grained classification schemes have been proposed[15].

In order to develop an intuition for the continuous strength of evidence that a Bayes factor provides one may use a probability wheel. One example is shown in Box 1 Figure. In order to construct the wheel, we have assumed that H_0 and H_1 are equally likely; the red part in the wheel is then the posterior probability for H_1, and the white part is the complementary probability for H_0. Now pretend that the wheel is a pizza with the red area covered with pepperoni and the white area covered with mozzarella. Imagine that you poke your finger blindly onto the pizza and that it comes back covered in the non-dominant topping -- in this case, pepperoni. How surprised are you? Your level of imagined surprise is an indication for the strength of evidence that a Bayes factor provides. We additionally compare the BF with traditional *P*-values in Extended Data Figure 1.

Europe PMC Funders Author Manuscripts

Box 1 Figure.
A probability wheel representation of a Bayes factor of 10 in favor of H_0. The circle has area 1.

Europe PMC Funders Author Manuscripts

Box 2

Six Advantages of a Bayesian Analysis for Pragmatic Neuroscientists

Pragmatic neuroscientists may be convinced to start conducting Bayesian analyses -- and Bayes factor hypothesis tests in particular-- only when the practical advantages of doing so are sufficiently evident. Below is a select overview of such practical advantages:

1. **Bayesian hypothesis testing allows researchers to discriminate evidence of absence from absence of evidence.**

   Non-significant p-values are notoriously ambiguous. Indeed, a *P*-value of 0.25 may indicate that the experiment was underpowered ("absence of evidence") or that the data support the null-hypothesis ("evidence of absence").

2. **Bayesian results are relatively straightforward to interpret and communicate.**

   Compared to frequentist conclusions, Bayesian conclusions are remarkably intuitive. While $p<0.01$ is not 5 times as convincing as $p<0.05$, $BF_{10}=6$ really does mean twice the evidence compared to $BF_{10}=3$. When neuroscientists make positive claims (e.g., The ACC is necessary for vicarious freezing), reviewers and readers may find it convincing if these claims are accompanied by an assessment of the statistical evidence, that is, an assessment of the extent to which H_1outpredicted H_0.

3. **Bayes factor hypothesis testing encourages researchers to quantify evidence on a continuous scale.**

   The advantage of retaining a continuous representation of evidence was stressed by Rozeboom ([31] pp. 422-423): "The null-hypothesis significance test treats 'acceptance' or 'rejection' of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true."

4. **For most statistical scenarios, Bayes factor hypothesis testing is now relatively easy.**

   Until recently, carrying out a Bayesian analysis for a standard statistical test required mathematical expertise and knowledge of probabilistic programming. This alone would be enough to deter most pragmatic neuroscientists who just wish to conduct a quick Bayesian *t*-test. However, recent R packages[14], Shiny apps[32], and GUI-based software packages such as JASP[33] now allow comprehensive Bayesian analyses with a minimum of effort.

Europe PMC Funders Author Manuscripts

5. **Bayesian inference allows researchers to monitor the results as the data accumulate.**

   As illustrated in Box 1, and Figure S1, the Bayesian predict-update cycle of learning continues indefinitely. In an experimental setting, neuroscientists may decide to terminate data collection when the result is deemed compelling or when they have ran out of time, money, or patience[8,34]. This means that experiments can be flexibly shortened or lengthened according to the evidence that has already been collected. If error control guarantees are put in place such flexibility can reduce the required sample size by as much as 50%[32,35].

6. **Bayes factor hypothesis testing allows researchers to include prior knowledge for a more diagnostic test.**

   Although the default prior parameter distributions allow for a robust reference analysis[36], these distributions can be adjusted in light of relevant background information. This background information acts to sharpen the predictions from the models, making them easier to discriminate. For instance, prior distributions for effect size may respect the direction of the prediction, or even its location[37].

Europe PMC Funders Author Manuscripts

Figure 1. *P*-value of a *t*-test and BF_{+0} as a function of effect size and sample size.

(A) Each histogram shows the distribution of *p*-values obtained from 1000 one-tailed one-sample *t*-tests based on *n* random numbers drawn from a normal distribution with mean μ and sd=1. To differentiate levels of significance, the first bin was split in multiple bins based on standard critical values. Note how, when there is an effect in the data (i.e., μ>0, all but leftmost column), increasing sample size (downwards) or effect size (rightwards) leads to a leftwards shift of the distribution: more evidence for an effect leads to lower *p*-values. In this case, *p*-values<0.05 are considered hits, and are shown in green, while *p*>0.05 are

considered misses and shown in red. However, somewhat counterintuitively, the converse does not hold true: in the absence of an effect, (μ=0, leftwards column), increasing sample size does *not* lead to a rightward shift (increase) of the *p*-values. Instead the distribution is completely flat, with all *p*-values equally likely (note that the distribution seems to thin out below 0.05, but this is because we subdivided the last leftmost bin into several bins to resolve levels of significance). In this case, $p<0.05$ are false alarms, shown in red, and $p>0.05$ are correct rejections, shown in green. *P*-values are thus not a symmetrical instrument: cases with much evidence for H_1 (high effect size and sample size) give us quasi certainty to find a very low *p*-value, whereas cases with much evidence for H_0 (e.g., μ=0 with n=100) do not make *p*-values close to 1 highly likely -- instead, any *p*-value remains as likely as any other. (**B**) Distribution of BF_{+0} (using $r = \sqrt{2}/2$ for the effect size prior Cauchy width) values obtained from 1000 *t*-test based on *n* random numbers drawn from a N(μ,1) normal distribution with mean μ and sd=1. Each histogram has the same bounds specified below the graphs, representing conventional limits for moderate and strong evidence. When an effect is absent (μ=0, leftmost column), evidence of absence (green bars and percentages, $BF_{+0}<1/3$) increases with increasing sample size, and false alarm rate is well controlled. When an effect is present (μ>0), evidence for a positive effect ($BF_{+0}>3$, green bars and green percentages) increases with sample size and effect size, and misses ($BF_{+0}<3$, red bars and red percentages) are rare (μ=0.5) or absent (μ=1.2 or 2). When percentages are not shown, they are 0% (red) or 100% (green). Data can be found at https://osf.io/md9kp/.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

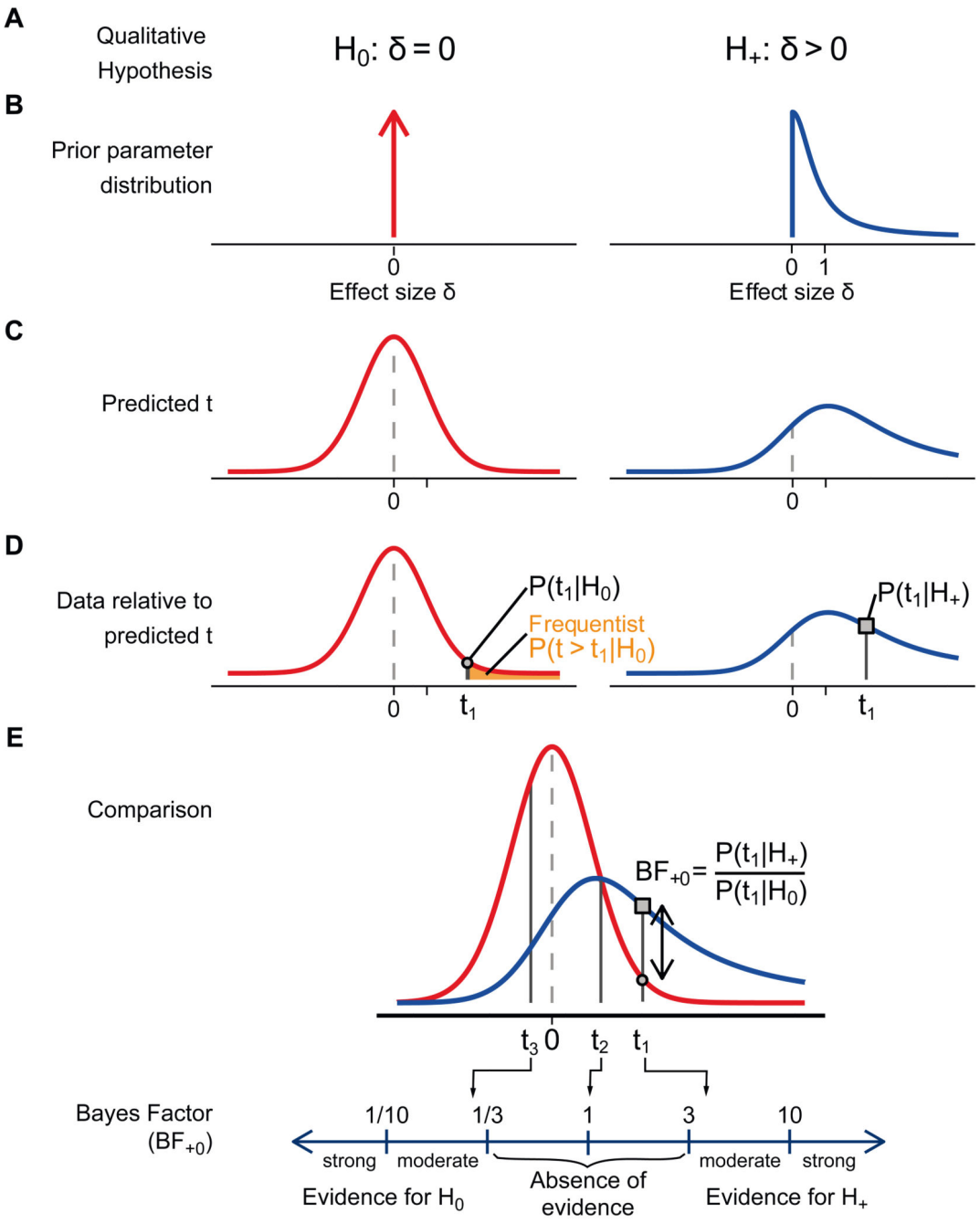


Figure 2. Hypothesis testing under the Bayesian Framework

(A) Two competing qualitative hypotheses are expressed in terms of a test parameter such as the population effect size δ. H_+ represents a directional hypothesis of a positive effect size. (B) The two rival hypotheses are formulated in terms of specific probability distributions expressing the plausibility or probability of each effect size value. (C) Each effect size distribution is transformed into expected *t*-values. For H_0, this is simply the standard *t*-distribution used in frequentist *t*-tests. For H_+, for each hypothesized effect size, a non-central *t*-distribution with that effect size is multiplied with the hypothesized probability of

that effect size in B. All of these weighted non-central *t*-distributions are then summed together to get the distribution in C. (D) After the data is obtained, the observed *t*-value (t_1) can be interrogated in each distribution. Note that, in frequentist statistics, the *p*-value is derived from the H_0 distribution alone, as the area where $t > t_1$. (E) The likelihood of t_1 under H_0 and H_+ is then compared to calculate the BF_{+0}. Here we illustrate three examples of observed *t*-values. At an observed value of t_1, the blue distribution is 4 times higher than the red; hence BF_{+0}=4, and we have (moderate) evidence for H_+. At an observed value of t_2, where the two distributions are equal, BF_{+0} = 1 and we have absence of evidence. At an observed value of t_3, the red distribution is 4 times higher than the blue; hence BF_{0+} = 4 and we have moderate evidence for H_0. Here we illustrated one-tailed hypotheses, as these respect the directional nature of the underlying theory and yield more diagnostic predictions. More agnostic two-tailed hypotheses are calculated using the same principles, but the truncated blue distribution in B is then replaced with a untruncated, symmetric distribution, as shown in the dotted line in Figure 6B. Data can be found at https://osf.io/md9kp/

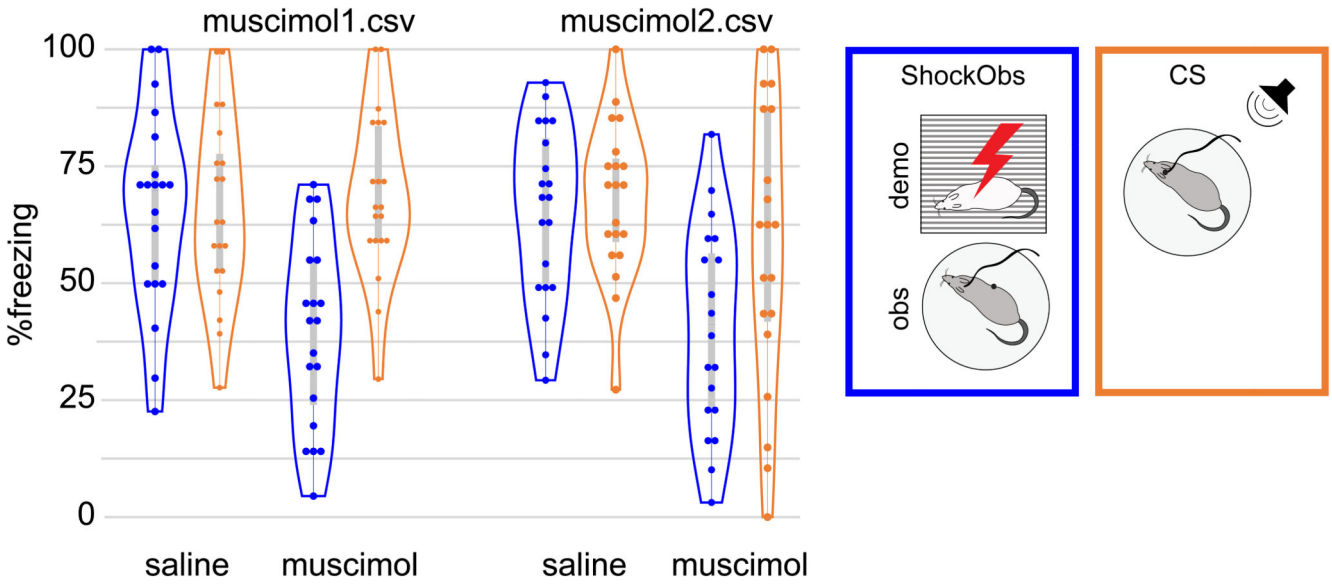


Figure 3. Illustration of the data for the two simulated scenarios

Muscimol1 data were simulated using μ=70 and σ=20 for all conditions (imposing a floor of 0 and a ceiling of 100), except ShockObs (in blue) under Muscimol which was simulated using μ=40. Muscimol2 data was simulated using the same parameters except for CS (in orange) under Muscimol, which had μ=65 and σ=40. Based on these data we should find evidence for H_+:Saline>Muscimol in all cases for ShockObs. For CS (orange), Muscimol1 should reveal evidence for H_0 (evidence of absence) given that data were drawn from the same μ=70, σ=20 distributions. For Muscimol2, CS was drawn from different distributions for saline and muscimol, but with n=20, it might be hard to adjudicate the difference, and we might thus expect absence of evidence. Data can be found at Data can be found at https://osf.io/md9kp/. Plots are violin plots, with the gray bar showing the middle two quartiles.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

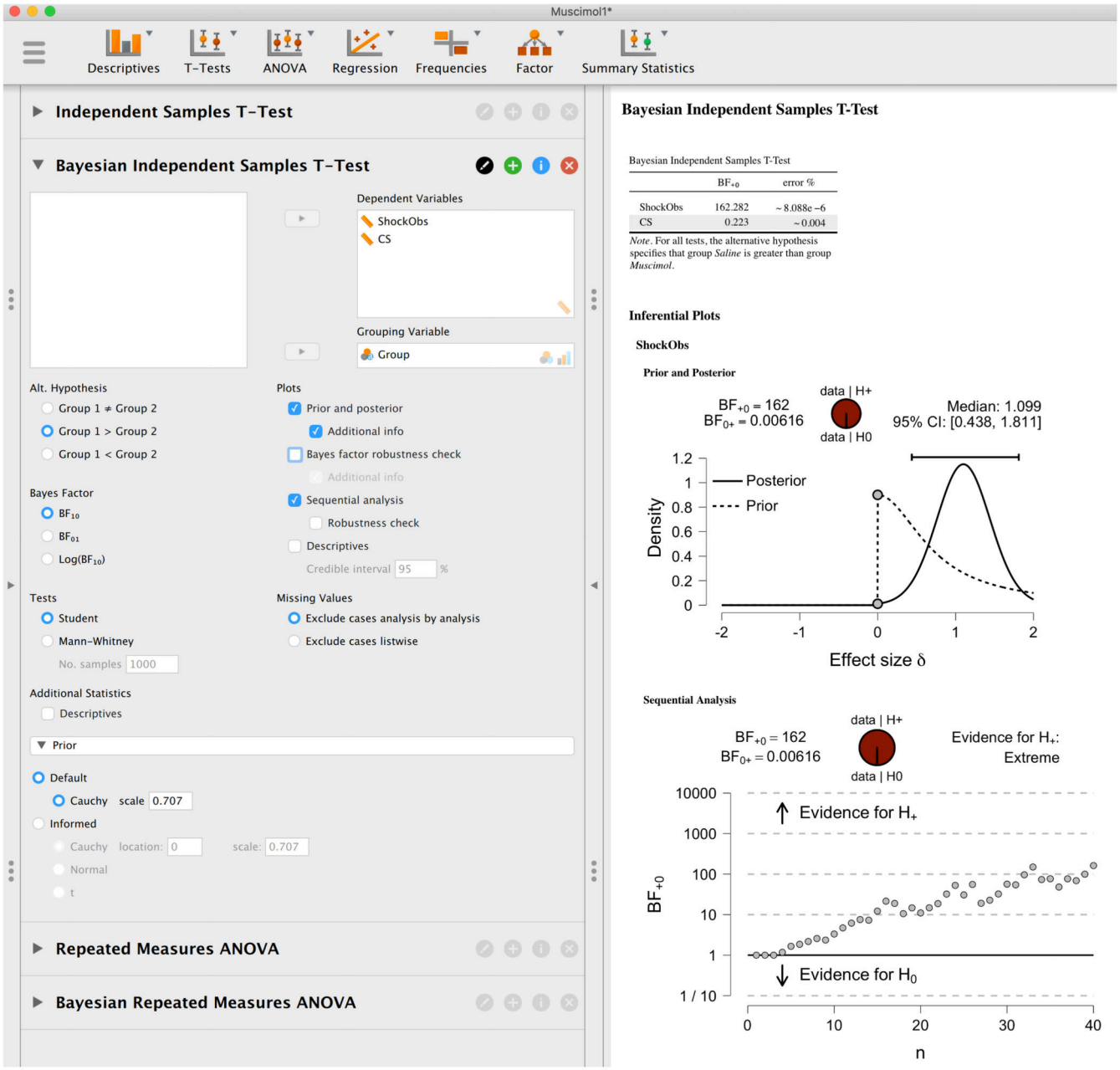


Figure 4. Screenshot from the Bayesian Independent Samples T-Test in JASP

The top right shows the Bayes factor for the two variables, followed by the inferential plot showing the credible interval of the effect size and the sequential analysis. The inferential plots shown on the right will be discussed in sections 4 and 5. Data can be found at https://osf.io/md9kp/

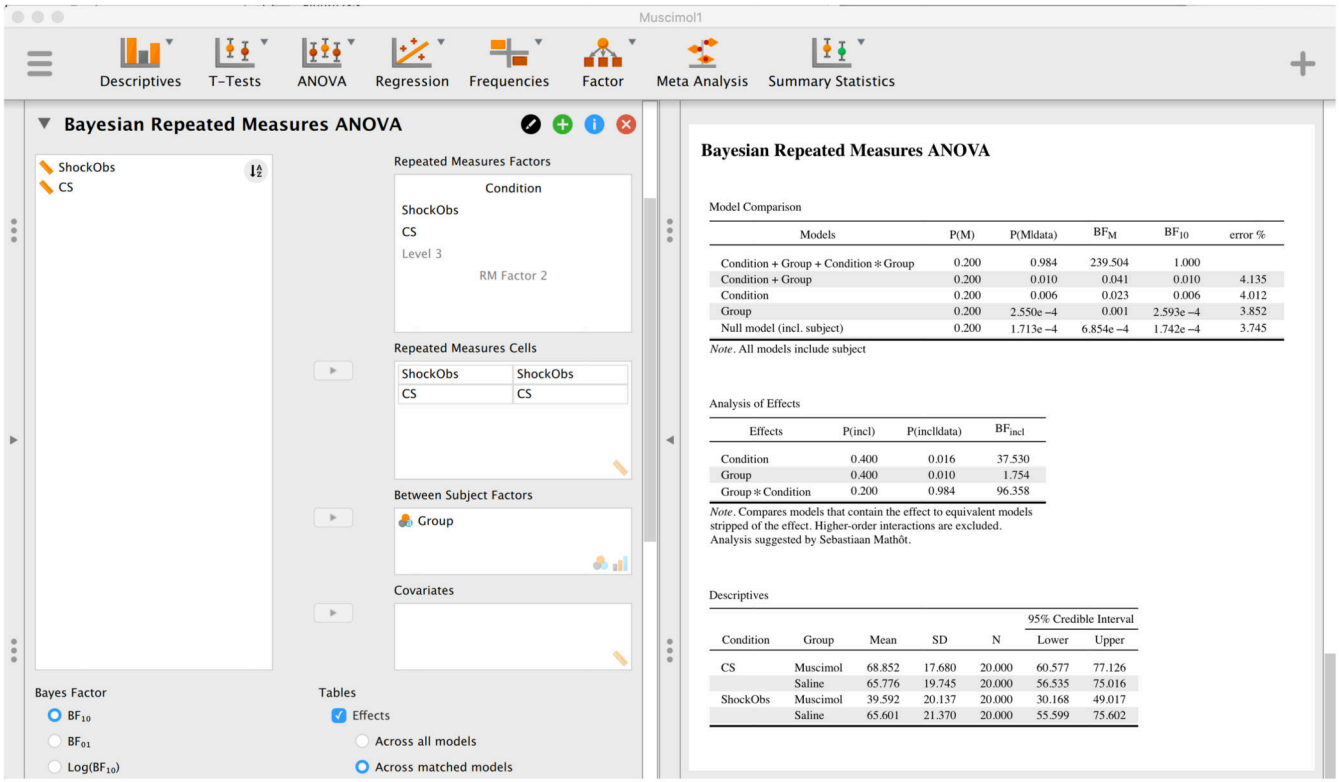


Figure 5. Screenshot of the Bayesian repeated measures ANOVA of Muscimol1

The muscimol1.jasp analysis file can be downloaded at https://osf.io/md9kp/

Europe PMC Funders Author Manuscripts

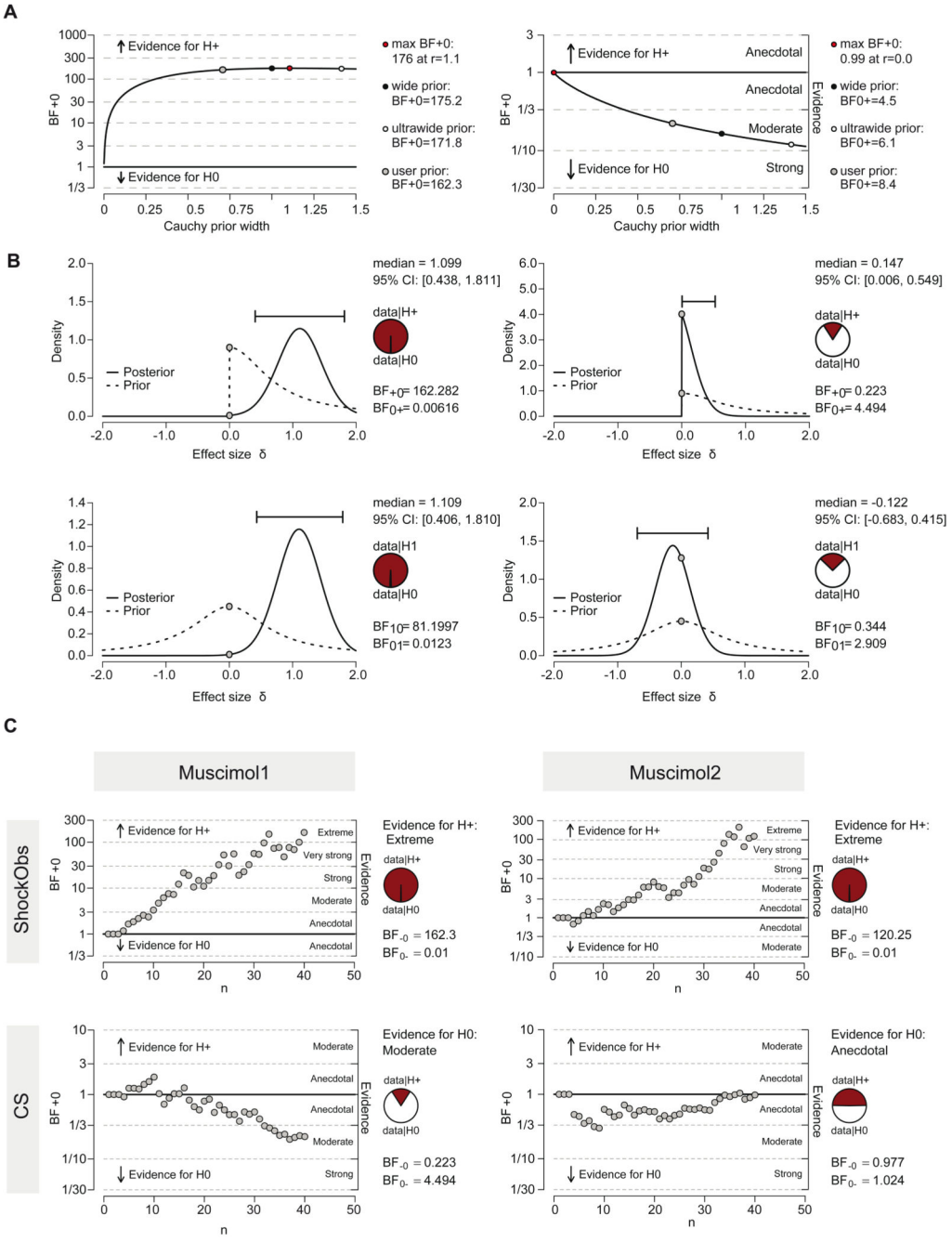


Figure 6. Further outputs for the Bayesian *t*-test on Muscimol1.csv

(A) Clicking the option Bayes Factor Robustness Check will plot for each variable (ShockObs on the left and CS on the right) the BF as a function of the effect size prior. The user prior (gray) is by default set at Cauchy scale 0.707 as recommended in[18]. The wide and ultrawide prior are flatter priors that are sometimes used, especially when the goal is parameter estimation. As can be seen, there is extreme evidence for H_1 in ShockObs, across all but the smallest priors (i.e., the gray, black and white dots all have BF_{+0}>160), and there is moderate evidence for H_0 for all but the smallest priors for CS (most BF_{0+}>4.5). The

Europe PMC Funders Author Manuscripts

interpretation of the data does thus not depend on the choice of prior scale within a reasonable range. (B) Priors and Posteriors for ShockObs and CS together with median and CI of the effect size. Results are shown for a one-tailed prior (top row) often more suited for hypothesis testing and two-tailed prior (bottom row) more suited for parameter estimation. (C) Accumulation of evidence with increasing sample size using the ‘Sequential analysis’ option. Data can be found at https://osf.io/md9kp/

Europe PMC Funders Author Manuscripts