There is Nothing Magical About Bayesian Statistics: An Introduction to Epistemic Probabilities in Data Analysis for Psychology Starters

Please cite as: Świątkowski, W., & Carrier, A. (2020). There is Nothing Magical about Bayesian Statistics: An Introduction to Epistemic Probabilities in Data Analysis for Psychology Starters. *Basic and Applied Social Psychology*, *42*(6), 387–412. https://doi.org/10.1080/01973533.2020.1792297

Wojciech ŚWIĄTKOWSKI, PhD*

Department of Social Psychology, University of Lausanne, Switzerland

wojciech.swiatkowski@unil.ch

Antonin CARRIER, PhD

Department of Psychology, University of Bordeaux, France

antonin.carrier@gmail.com

*Corresponding author information: University of Lausanne, Department of Social Psychology UNILaPS, Quartier UNIL-Mouline, Bâtiment Géopolis, Lausanne 1015, Switzerland.

Abstract

This paper is a reader-friendly introduction to Bayesian inference applied to psychological science. We begin by explaining the difference between frequentist and epistemic interpretations of probability that underpin respectively frequentist and Bayesian statistics. We use a concrete example – a student wondering whether s/he carries the virus *statisticus malignum* – to explain how both approaches are different one from another. We illustrate Bayesian inference with intuitive examples, before introducing the mathematical framework. Different schools of thoughts and recommendations are discussed to illustrate how to use priors in Bayes Factor testing. We discuss how psychology could benefit from a greater reliance on Bayesian methods. Finally, we illustrate how to compute Bayes Factors analyses with real data and provide the R code.

Key-words: Bayesian statistics, probability, Bayes Factor, statistical inference

There is Nothing Magical About Bayesian Statistics:

An Introduction to Epistemic Probabilities in Data Analysis for Psychology Starters

For the past few years, matters such as the improvement of research practices, reproducibility of published findings and data analysis have been among the hottest topics in the mainstream psychological literature. In the ongoing debates, Bayesian statistics have often been advocated as a viable tool to supplement the existing practices (e.g., Benjamin et al., 2017; Johnson, 2013; Lindsay, 2015; Wasserstein & Lazar, 2016). Bayesian statistics were introduced in psychology for the first time already decades ago (Edwards et al., 1963), yet it is only quite recently that they have inspired a wide interest across the academia. The recent proliferation of introductory texts (e.g., Etz & Vandekerckhove, 2018; Kruschke & Liddell, 2018a; Wagenmakers, Marsman et al., 2018) and special contributions highlighting the most recent developments (e.g., Hoijtink, & Chow, 2017; Mulder & Wagenmakers, 2016) testifies of a global trend toward an increase in interest in Bayesian methods across psychological science (van de Schoot et al., 2017).

The present paper is intended for those who wish to begin their experience with Bayesian statistics and get a clear understanding of what they are about. By means of concrete examples and intuitive explanations, it offers a non-technical, reader-friendly introduction that guides the reader through the essential topics in the following, incremental way. We start with discussing the notion of probability (Section 1) from which the rationale of Bayesian inference derives (Section 2), and we illustrate it with several examples (Section 3). We then outline a formal presentation of Bayes' Theorem (Section 4). After having discussed the issue of priors and current practical recommendations on how to specify them (Section 5), we discuss hypothesis testing with Bayes Factors (Section 6). The paper also seeks to provide some food for thought on how Bayesian statistics could improve analytical practices in psychology (Section 7). Finally, we present an example of Bayesian hypothesis testing based on real data from a social psychology research and include the corresponding R code (Section 8). For those who are already familiar with Bayesian

inference, this paper also offers a more in-depth introduction to Bayesian concepts and formulas, as well as an up-to-date literature review on the ongoing recommendations and practices.

We also want to make clear what *is not* the purpose of this paper. Firstly, we do not intend to give an exhaustive account of Bayesian analyses. The article covers some approaches instead of others and we take care to mention complementary references whenever relevant. Secondly, we do not provide an extensive training on how to compute the analyses in a statistical software. We mostly focus here on laying down the conceptual bases and we mention useful statistical programs and references to available manuals. That being said, we still show how to conduct an actual Bayesian analysis in the last section of the article.

Finally, we also wish to emphasize that the present paper does not have the vocation to criticize the classical, frequentist framework of *p*-values for the benefit of Bayesian statistics. Though some authors do espouse such an approach and advocate for a complete switch to Bayesian statistics (e.g., Kruschke & Liddell, 2018b; Wagenmakers, Marsman et al., 2018), here we focus on discovery rather than persuasion.[1] With this respect, the present paper adopts an approach of comparative pedagogics: We first lay down the fundamentals of the classical, frequentist approach to data analysis so that readers can develop useful intuitions on how Bayesian inference differs from what psychologists are mostly used to with the *p*-values.

1. Frequentist versus Epistemic Probabilities

There are at least two ways to interpret what "probability" means. The key-step in understanding Bayesian statistics first requires making explicit their underlying *epistemic*[2] interpretation to probability, which is fundamentally different from the *frequentist* interpretation underpinning classical statistics.

Frequentist probability

The frequentist interpretation defines probability as a long-term frequency of occurrence of an event in a specific set of events (von Mises, 1957). Frequentist probabilities are said to be *in the world*. They refer to intrinsic characteristics of nature that can be assessed more or less accurately

via observation. For instance, let us assume that *statisticus malignum* is a rare virus that triggers a serious disease known as *statisticophilia gravita*. The symptoms of the disease, in its severest form, include compulsive reading of statistical textbooks and pathologic feelings of fruition when performing regression analyses in R. Let us further assume that some serious scholars estimated that in the population of psychology students across the globe, the virus is present in 1 out of 1000 individuals. The frequentist probability of having the virus is then 1/1000 = .001. It can be obtained by computing the ratio between the number of actual occurrences of the event (i.e. carrying the virus) and the total number of possible occurrences of the event (i.e. being a psychology student), which *in the long run* would converge to .001. From the frequentist perspective, this probability is the *frequency* of carrying the virus in this specific *reference class*, namely the population of all psychology students. This means that if we randomly sample *over and over again* from this population, it would be expected that 1 out of 1000 students would have the virus, on average.

Epistemic Probability

The epistemic interpretation defines probability as a measure of degree of confidence (Jeffreys, 1961). Epistemic probabilities are said to be *in the mind* because they relate to states of knowledge and represent subjective degrees of belief. In other words, epistemic probabilities can be used to quantify the degree of uncertainty regarding the occurrence of an event or the truthfulness of a statement. Imagine yourself as a psychology student wondering whether you carry the virus *statisticus malignum*. Having read the epidemiological reports written by the serious scholars about the prevalence of the virus in your group of reference – that is, .001 – you feel relieved, at first. Then, you realize that you have always enjoyed your statistics classes and among various statistical software you came across, R is your favorite one. You believe that your odds of having the virus are somewhat greater than you could expect based on the epidemiological reports. You are still pretty confident that most likely you do not carry the virus. You actually believe that there is a 5% chance of you being infected. Notice that here, the 5% is nothing but a belief. It quantifies your degree of uncertainty about having the virus, which is conceptually different from some objective frequency

of virus contamination within a reference class (here, the population of psychology students; see also Dienes, 2008, 2011). For the sake of illustration, the epistemic probability of 5% in this example is also numerically different from the frequentist probability of .001 mentioned above. Note however that modeling epistemic probabilities is not inconsistent with using frequentist probabilities as a starting point: you could have perfectly assumed that there was a .1% chance of you being infected (in line with the epidemiological reports), had you felt it represented accurately your state of knowledge. Using epistemic probabilities is fundamental to Bayesian statistics and, as we will later see, some proponents of Bayesian methods in fact stipulate that epistemic probabilities should be tuned to frequentist probabilities, whenever the latter are known.

2. Frequentist versus Bayesian Statistics

This section extends the previous one and discusses how each of the two interpretations of probability is determinant for each school of inferential statistics. The aim here is to compare both schools of inference, so that hopefully readers can grasp some useful intuitions about what Bayesian inference is and how it differs from the standard, frequentist approach to statistical inference (see *Table 1* for a summary). For a more detailed account, we invite readers to read Dienes (2008, 2011), Kruschke and Liddell (2018b), Wagenmakers (2007) and Wagenmakers et al. (2008).

Frequentist Statistics

The classical statistics that psychologists use when they perform ordinary *t*-tests, ANOVAs, regression analyses, etc. rely on the frequentist interpretation of probability. As explained above, for frequentist probabilities to be meaningful, a specific set of events (i.e. a reference class) must be pre-specified. When one performs a statistical test that yields the usual *p*-value, the reference class refers to the sampling distribution under H_0. For instance, when one obtains a *p*-value = .025 based on some data, then it means that such a set of data occurs with a maximal frequency of .025 under H_0 *in the long-run*. Stated otherwise, if H_0 were true, such data would only appear with a frequency of .025 if we were to repeat the same study over and over again, all other things being equal. The classical approach to statistical inference relies on such statements about frequencies – or

frequentist probabilities – of *some data* given *some hypothesis*. Specifically, in the Neyman-Pearson (NP) approach, one first specifies two statistical hypotheses about the investigated effect (H_0 and H_1), a sample size, and α and β levels that indicate rejection regions for each hypothesis based on cost-benefit trade-offs, before conducting the study[3] (Gigerenzer et al., 2004; Lakens et al., 2018). After conducting the study, one then calculates the *p*-value: the probability of observing some sets of data at least as extreme as the actual set of data obtained in the study *if it was run over and over again* assuming H_0 were true[4]. If $p < \alpha$ (conventionally $\alpha = .05$, but see Lakens et al., 2018), then one rejects H_0 and accepts H_1. Thus, the level of significance α can be used as the error rate control in making wrong decisions of rejecting H_0. Specifically, α refers to the relative frequency of Type-I errors (i.e. accepting H_1 while H_0 is true) in the long run of using this procedure. Here, the *p*-value is used only to take a yes-no decision: either to accept or reject a hypothesis. Since the *p*-value pertains to some possible sets of data that would be – *but were not* – obtained in the long run, frequentist inference does not depend *stricto sensu* on the observed data from a study. Instead, the frequentist inference uses probabilities to simulate possible outcomes based on the observed data in order to draw a generalization beyond the observed data, upon which a decision to accept or to reject a hypothesis is made.

We illustrate the frequentist logic of the NP approach with the following example. Again, imagine that you are the psychology student and that you want to assess whether you carry the virus *statisticus malignum* or not. You decide to consult with Dr. Nikolaevich Golmorokov, a highly regarded Russian scholar in research on the virus. Dr. Golmorokov is also a sharp-minded statisticologist: he specializes in detecting the virus and has developed a highly accurate test to detect the presence of this virus in psychology students. Dr. Golmorokov developed the test as follows. He considered two hypotheses: H_0: "The student does not carry the virus" and H_1: "The student does carry the virus". He assumed accepting H_0 while H_1 is true (Type II error) to be the costliest error, because leaving the virus unnoticed could lead to the contamination of the general population of students. He thought however that accepting H_1 while H_0 is true (Type I error) is not

so problematic, because treating the virus does not involve any substantial costs for patients. Thus, he designed his test to ensure high statistical power set at .99 ($\beta = .01$), and with an error rate of $\alpha = .10$. In his office, the doctor explains the calibration of the test: the latter will miss the presence of the virus on average in one among a hundred of students who really carry the virus (Type II error) and that among a hundred of students who do not carry the virus, his test will falsely yield its presence in ten individuals, on average (Type I error). Having performed the test, the doctor gives you the results. It is positive, which indicates that you carry the virus. Without any time to waste, the doctor gives you the cure. He takes some red pill out of his pocket, asks you to swallow it, then accepts your payment of 99 CHF for the treatment, and finally sets you free to go, supposedly cured.

Notice that in the example above, it was completely irrelevant whether the doctor believed or not if the student carried the virus. He only *acted* accordingly with the test result. The idea is that with α and β probability levels under control, the doctor can monitor the long-term frequency of wrong decisions. This example is meant to illustrate that the underlying frequentist approach to statistical inference answers the question "What should I do?". The aim is to guide one's behavior in dichotomous decision-making in rejection and acceptance of hypotheses while controlling the error rates committed by the repetition of this procedure *in the long run*. Consequently, everything that might alter this error rate in the long run – such as multiple comparisons, sampling intentions, post-hoc testing – must be taken into account if one seeks to keep α at the desired, fixed level (Wagenmakers, 2007; Wagenmakers et al., 2008). Notice that taking a decision means only to act accordingly, with no consequence whatsoever on believing whether the decision is right or wrong (Dienes, 2011).

Bayesian Statistics

In contrast, Bayesian statistics rely on the epistemic interpretation of probability. In essence, epistemic probability is about quantifying uncertainty, and Bayesian statistics are about using uncertainty for inference. In this approach, one starts by formulating an initial belief about a

hypothesis of interest. This initial belief translates one's uncertainty about the truthfulness of a hypothesis and is called the *prior*. Later sections will further describe Bayesian priors and discuss their choice in greater depth. At this stage, it suffices to say that the prior reflects one's subjective state of knowledge – more or less precise – about some effect of interest that is available before conducting the study and collecting the data. As we will illustrate with the example below, the aim of Bayesian statistics is to use data in order to update the prior in the *posterior.* In other words, Bayesian statistics basically use data to update beliefs about the phenomenon of interest.

For the last time, think yourself again as the psychology student wondering whether you carry the virus *statisticus malignum*. Even though the test by Dr. Golmorokov turned out to be positive, you would like to get an idea of the probability that *you* carry the virus. As you recall, your initial belief was that there was a 5% chance of you carrying the virus. You assign to the hypothesis H: "I carry the virus" a prior $p(H) = .05$, and this is your initial belief of this hypothesis being true.

However, you cannot ignore the fact that you consulted with Dr. Golmorokov and his test indicated that you carry the virus. You decide to use the knowledge of this data D (i.e. the positive test result) to learn something more about your chances of having the virus. In other words, you intend to use D in order to update your prior. You recall that the calibration of the test was as follows: (1) if you carry the virus, there is a 99% chance of the test being positive; (2) if you do not carry the virus, there is only a 10% chance of the test being positive. With this information in mind, you use Bayesian inference (which is illustrated more formally in Section 4), to calculate your probability of carrying the virus, *given* the fact that your result was positive. *Abracadabra*: you now know that there is a 0.3425606 probability of you carrying the virus, given that the test result was positive. Somewhat confused, you realize that you might have intuitively overestimated the probability of you carrying the virus because you tended to neglect how low your prior was in the first place ($p(H) = .05$) and relied only on the result given by the test. With this in mind, you now understand that, although the test diagnosed you as positive, there is still around a one-in-three chance of you actually having the virus[5].

Notice that the probability 0.3425606 is the new epistemic probability of the hypothesis *H* that you actually carry the virus. This probability has been obtained based on the prior probability of the hypothesis $p(H) = .05$, which has been then updated by the knowledge of data *D*. This is a conditional probability – $p(H|D) = 0.3425606$ – because now it refers to the probability of *H* being true given the fact that the test result was positive (see also Appendix A and Section 4 for the detailed calculation). This probability is typically referred to as the posterior, since it is established *a posteriori* to knowing the data. Getting from prior to posterior probabilities of hypotheses by the means of data is the very kind of statistical inference one seeks to perform within the Bayesian framework.

The example described above aimed to illustrate that Bayesian statistics answer the question *"*How should I adjust my beliefs?*"*. The aim of Bayesian inference is adjusting one's confidence – or uncertainty – in statistical hypotheses. Importantly, the rules of probability used in Bayesian inference that will be explained later imply that once the prior is set and data are known, there is only one way in which these two pieces of information combine together to get the posterior probability of a hypothesis. Importantly, this implies that Bayesian inference requires no adjustments for factors such as multiple testing, sampling intentions or post-hoc testing (Dienes, 2011, 2016; Kruschke, 2014; Rouder, 2014). This is because, contrary to frequentist statistics, there is no reason to address the question of "what would happen in the long run if… *"*. Therefore, the posterior information remains unchanged regardless of, for example, how many other tests involving different hypotheses/priors were – *or could have been* – conducted, or regardless of how data was obtained (e.g., sequentially or not). However, this is also the reason why using Bayesian inference in decision making for research purposes will not allow controlling for the amount of wrong decisions made in the long run. The notion of long-term error rate is irrelevant in Bayesian statistics (Dienes, 2016) and therefore cannot be controlled for. This does not mean, however, that one cannot strategically use frequentist long-term error rates resulting from using Bayesian

inference in decisions to accept or reject hypotheses (Dienes, 2016; Schönbrodt & Wagenmakers, 2017).

[Insert Table 1 near here]

3. Some Intuitive Examples of Bayesian Analyses

Before we get to the formal introduction to Bayesian statistics, we first present some easy but conceptually meaningful examples of Bayesian analyses. These are meant to help readers develop further intuitions about their underlying mechanics (see also van de Schoot et al., 2014). Unlike in the previous section where the underlying parameter of interest was dichotomous (i.e. carrying or not carrying the virus), here we consider a more typical scenario where the parameter of interest is continuous. In the following examples, we illustrate how different priors can combine with the same data to result in different posteriors.

Imagine a group of three undergraduate students who, as part of their research project, are interested in the link between the *statisticophilia gravita* disease and the introversion personality trait. They want to estimate the mean level of introversion in the population of psychology students diagnosed with the disease. They run the study together: they ask participants diagnosed with the disease to fill in a short questionnaire measuring introversion. In statistical terms, their parameter of interest – which we refer to as μ – is the mean score of introversion in this population, with $\mu = 0$ denoting the scale midpoint. They try to estimate the value of μ. Though they ran the study together, each of the three students had formulated a different prior before conducting the study and thus analyzed the data separately. *Figure 1* displays their analyses (see Appendix B, for calculations). The data are portrayed by the likelihood function. The likelihood gives the conditional probability of obtaining the data given each considered value of μ. *Figure 1* shows that the value of μ maximizing the probability of observing this set of data is $\mu = 4.5$.

[Insert Figure 1 near here]

Student 1 assumed μ to be positive: he hypothesized that psychology students with the disease would be highly introverted and would take more joy in socializing with statistical manuals rather than with their fellow students. *Figure 1A* illustrates his relatively *informed* prior. It uses a bell-shaped curve typical to a normal distribution centered on $\mu = 2$. This means that the parameter is believed to be most likely positive, with values close to the mean being the most plausible *a priori*. Once the prior is updated in the posterior, the most probable parameter value given the data is $\mu = 3.4$. Note that because a range of μ values were both probable under the prior and supported by the data, the belief in the truthfulness of these parameter values has strengthened in the posterior distribution of μ. *Figure 1A* shows that the Bayesian analysis maximized the posterior probability of those μ values that were congruent both with the specified prior and the likelihood function.

Unlike Student 1, Student 2 did not hold any strong expectations regarding μ values before the study. Often researchers dispose of only little – if any – knowledge about the effect before running a study, and therefore cannot favour a specific hypothesis. From the Bayesian perspective, this also refers to a kind of knowledge to be integrated into the analysis. In such instances, *weakly informed*, *vague* or *uninformed* priors can be used. *Figure 1B* illustrates the choice of such a prior: it is an almost flat-shaped normal distribution centered on $\mu = 0$ with a substantially greater variance than the prior used by Student 1. The prior from *Figure 1B* has considerably less *precision* than the prior from *Figure 1A*, meaning that the uncertainty of Student 2 is much higher than that of Student 1. The prior's precision is the inverse of its variance and illustrates the strength of the belief: the stronger the belief, the higher the precision and the lower the variance. With such a weakly informed prior as the one depicted in *Figure 1B*, the posterior is almost exclusively determined by the data. While the prior favored parameter values close to $\mu = 0$ to a minor extent, the posterior now favors values centered on $\mu = 4.1$. Once again, the intuitive nature of information integration under uncertainty is captured with Bayesian statistics: when there is only a glimpse of prior

knowledge available about a phenomenon, the resulting posterior is essentially determined by the likelihood function (that is, empirically).

Finally, Student 3 assumed μ to be most likely null. Since the student held a very strong assumption about the value of this parameter, he used a *highly informed* prior (see *Figure 1C*). With such a strong allocation of prior probability, the data has only little impact on the posterior. In realistic situations, a prior that strong would have little reason to be used. If choosing such a highly informed prior was acceptable in a scientific community, using scarce research resources to conduct a study would be hardly justified. Otherwise, a prior that subjective could be simply deemed of limited interest. This example is merely presented for illustrative purposes to show another property of Bayesian analyses: with a strong prior allocation of probability in a hypothesis, much data is needed to exert a genuine impact and change the belief.

These examples are meant to outline some basic principles of Bayesian inference. The take-home message is that the posterior probability of a hypothesis is a coherent trade-off between the prior and the data, weighted by their respective precision. Bayesian inference is intuitively rational (Kruschke, 2014): strongly informed priors require much novel data to be changed, which is more easily achieved with weakly informed priors. We encourage readers to check Kristoffer Magnusson's interactive app (http://rpsychologist.com/d3/bayes/) that illustrates the principles described here.

For this pedagogical purpose, the choice of the prior was systematically arbitrary. It is only but a legitimate question to ask how one should proceed in real-life situations when choosing a prior. Specifying priors is certainly the most challenging, conceptually difficult and also frequently criticized part of Bayesian statistics. In Section 5, we introduce different schools of thoughts on how to choose the prior and on discussing what role it should play in data analysis.

4. The Bayesian Approach to Statistical Inference

Bayesian statistics rely on the assumption that it is possible to assign degrees of epistemic probability to scientific hypotheses (Dienes, 2008, 2011). Unlike in classical, frequentist statistics

where we assume hypotheses about statistical parameter values to be either true or false (e.g., "H_0: $\mu = 0$" is false and "H_1: $\mu \neq 0$" is true), here we consider that hypotheses can be described as being more or less plausible or credible possibilities (Kruschke, 2014). Each possible hypothesis can be assigned to a certain level of epistemic probability that reflects the degree of confidence or belief one puts in the truthfulness of the given hypothesis. Bayesian statistics are all about updating this prior allocation of uncertainty across considered hypotheses by the means of data into a posterior allocation of uncertainty (Dienes, 2011; Kruschke, 2014; Lee & Wagenmakers, 2014).

The way the prior relates with the posterior is given by Bayes' Theorem (Bayes & Price, 1763; Laplace, 1814; see Appendix C, for derivation), which states that, for a hypothesis *H* and a set of data *D*:

$$p(H|D) = \frac{p(D|H) \times p(H)}{p(D)}. \quad (1)$$

This formula indicates how the prior probability of a hypothesis must be updated after getting to know data in order to get the posterior probability of the hypothesis being true. Specifically, the posterior probability of the hypothesis being true given data $p(H|D)$ is given by *likelihood* $p(D|H)$ times prior $p(H)$, divided by the probability of data $p(D)$ referred to as *evidence*, which is the literal transcription of Eq. (1).

The likelihood $p(D|H)$ refers to the conditional probability of the observed data, given the hypothesis. It is a mathematical function that indicates under which parameter values specified by the hypothesis the data are most likely to occur. All of the relevant information for the inference and the support that is provided by the data for a given hypothesis is captured by the likelihood function (Birnbaum, 1962, as cited in Dienes, 2011). Importantly, a hypothesis with the highest likelihood value is not necessarily the one with the highest probability of being true given the data: the likelihood $p(D|H)$ should not be confused with the posterior probability $p(H|D)$. The latter depends both on the likelihood and the prior.

To illustrate the use of Bayes' Theorem, we return to the example of the psychology student assessing the chances of carrying the virus given that the test result was positive (see *Figure 2*). We assumed the prior for H_1: "I carry the virus" such that $p(H_1) = .05$. Based on the information available on the test calibration, we use the rule of marginalization (see Appendix D) to calculate the probability of the evidence $p(D)$ necessary for calculations. Then, by applying Bayes' Theorem, we determine that the probability of carrying the virus, given that the test was positive, is 0.3425606 (see *Figure 2*).

[Insert Figure 2 near here]

Notice that in order to obtain the posterior $p(H_1|D)$, we needed to calculate the probability of data $p(D)$. For real life situations involving continuous and multiple parameters, this element of the equation will be impossible to calculate.[6] Moreover, performing Bayesian analyses will consist most of the time in comparing posterior probabilities of several hypotheses, rather than computing the exact value of a posterior for a given statistical hypothesis (as we did here merely for pedagogical purposes). Because all of these comparisons will rely on the same data, the value of $p(D)$ stays unchanged: it is only a normalizing constant that can be safely omitted from analyses.[7]

To sum up, the rationale of Bayesian statistics is quite straightforward: you start with an initial belief about a hypothesis on the effect at hand and use the data from your study to update this belief. Remarkably, this reasoning may seem quite natural as it mirrors some aspects of human reasoning in situations of uncertainty (Ajzen & Fishbein, 1975). Imagine you drive a car on a road you know well and find yourself unexpectedly in a traffic jam. You obviously do not know for sure why there is unusual traffic (uncertainty), yet you suspect that a car accident seems quite likely to have occurred (prior). Suddenly, you hear an alarm siren and see an upcoming ambulance (data). Because you know that ambulances are especially likely to appear when car accidents happen and you rarely see them otherwise (likelihood), the fact that you saw an ambulance has just strengthened

your prior belief that the traffic comes from a car accident (posterior). From the Bayesian perspective, updating personal beliefs in scientific hypotheses in the light of data is precisely the point of scientific inference (Dienes, 2008). The Bayesian framework can be relevant in research methodology because these "personal" prior beliefs may actually refer to previously accumulated scientific knowledge or reflect theoretically driven predictions about studied effects. The next section elaborates on this point.

5. Specifying Bayesian Priors: On Some Intricacies of Subjectivity in Statistics

A natural question that arises when discovering Bayesian statistics is: *"*How do I choose my prior?*"*. Specifying priors is undoubtedly one of the most challenging parts of Bayesian statistics, and might seem difficult at first glance. There are no unanimously shared rules on how to choose one's prior. At least two schools of thoughts can be distinguished in this regard, namely *subjective* and *objective* Bayesian statistics (Berger, 2000, 2006; Goldstein, 2006; Sprenger, 2018). We espouse this distinction here because of its heuristic and pedagogical value, though we invite readers to keep in mind that it does not fully embody the complexity of disagreements and nuances between different Bayesian stances (e.g., Bandyopadhyay & Brittan, 2010; Berger, 2000, 2006; Gelman, 2008; 2011; Gelman & Shalizi, 2013; Goldstein, 2006; Williamson, 2010). After laying down the basic presumptions of each of the two schools, we will address the issue of subjectivity in Bayesian statistics. The section will end up with an articulation of the strengths of both objective and subjective approaches.

Subjective Bayesian Statistics

The subjective school uses priors to reflect personal beliefs about the studied phenomenon, and this subjectivity has a key role in data interpretation. Priors serve a descriptive function to model one's initial assumptions about the problem, and Bayes' Theorem is used to adjust these in the light of evidence (Goldstein, 2006). Subjective priors can be based on purely personal judgments and intuition (e.g., Howard et al., 2000). In this tradition, priors can also be calibrated based on frequentist probabilities if the latter are known (see Williamson, 2010), as we already

mentioned earlier. Priors can be also subjective in the sense that they are *informed* in a scientifically relevant way. Firstly, priors can be constructed with reference to past research that addressed similar questions, drawing from individual studies, literature reviews or meta-analyses. "Today's posterior is tomorrow's prior" (Lindley, 2000, p. 301). The rationale of subjective Bayesian statistics is consistent with the cumulative nature of science, where new research is always introduced with reference to past research. Such available knowledge can be accounted for in the Bayesian framework (Kruschke, 2014). Secondly, subjective priors can also represent predictions one makes from a theory (Dienes, 2008, 2011, 2014, 2016, 2019; Dienes & Mclatchie, 2018). Here, a prior is said to be "subjective" in the sense that it is a rational translation of a theoretical prediction that is only possible via an *informed* scientific judgment. Dienes (2011, 2014, 2019; Dienes & Mclatchie, 2018) provided some useful guidelines on how to derive such theory-based priors, using uniform, normal and semi-normal distributions. Notice that for both these approaches, subjectively informed priors are not "capricious and idiosyncratic and covert, but instead […] based on publicly agreed facts or theories […] [and] must be admissible by a skeptical scientific audience" (Kruschke, 2014, pp. 114–115).

Objective Bayesian Statistics

It is considered in objective Bayesian statistics that priors should reflect as little assumptions as possible so as to limit their influence in data interpretation and remain as uncontroversial as possible (Berger, 2006). Instead of committing to beliefs or informed judgments, the objective approach to Bayesian inference often involves using pre-determined rules of thumbs when formulating priors that can be used in common situations. Objective Bayesian analyses thus refer to a set of conventional and default procedures to be used whenever subjectivity is not integrated into the analyses (Berger, 2006). In parameter estimation, it involves using *vague* priors that diffuse prior probability density over a wide range of parameter values. These priors exert virtually no impact on the posterior, which is essentially determined by data (Kruschke, 2014; Kruschke & Liddell, 2018a, 2018b; *Figure 1B* shows an example). In hypothesis testing, commonly used

objective priors are known as *default* priors. These classes of priors are meant to be either non-informative in the sense that they limit the assumptions about the range of effect sizes (e.g., Rouder et al., 2009), or *weakly-informed* in the sense that some initial assumptions are necessary but still compatible with a large range of effect sizes and thus represent only diffuse prior knowledge (e.g., Rouder & Morey, 2012). A popular default prior is the JZS prior formulated by Rouder and colleagues (2009). It uses a Cauchy distribution[8] centered on 0, which favors effect sizes closest to 0, but still allocates reasonable amounts of prior probability to stronger effect sizes in each direction. This prior also uses a scale factor parameter r that specifies the range of expected effect sizes to be observed in data. The default value is set on $r = .707$, which means that 50% of expected standardized effect sizes falls within the [-.707, .707] interval. Recent research in Bayesian statistics has been marked by substantial efforts in the formulation of such priors that can serve as defaults in commonly used designs such as *t*-tests (Rouder et al., 2009; Wetzels et al., 2009), ANOVAs (Rouder et al., 2012; Wetzels et al., 2012), correlations (Wetzels & Wagenmakers, 2012) or multiple regressions (Liang et al., 2008; Rouder & Morey, 2012). The developments of such priors greatly mitigated the concerns raised by the use of priors based on idiosyncratic assumptions and made the Bayesian methods accessible for a variety of complex models. One may ask, however, what is the point of performing Bayesian analyses and struggle with epistemic probabilities if the goal is to limit their influence in data interpretation. Arguably, such methods offer an interesting trade-off: They allow making statements about probabilities of hypotheses and provide statistical evidence, while still mostly capitalizing on data. Finally, default priors are less time- and effort-consuming to use, more widely accessible to a non-expert audience, easier to communicate, and may represent more compelling evidence to other researchers than subjectively informed priors (Berger, 2006; Wagenmakers, Marsman et al., 2018).

A Brief Commentary on Subjectivity in Statistics

We suspect that some of the readers raised in the frequentist tradition may find the objective approach to Bayesian analyses more appealing than the subjective one, supposedly because the

former falls closer to the standard of objectivity inherent to science. This issue points right at the heart of a major criticism of Bayesian statistics that has fueled a long-standing debate in the history of epistemology of science, namely its subjectivity (Gelman, 2008). In particular, it is often argued that Bayesian statistics – and especially the subjective approach we discussed above – lack concordant objectivity (e.g., Sprenger, 2018; Wagenmakers, Marsman et al., 2018). The objection essentially unfolds as follows: scientists using different priors to analyze their data may reach diverging conclusions while using the same dataset. Obviously, discussing the issue of subjectivity and objectivity in data analysis in depth is complex and falls outside the scope of the present paper (see Gelman & Hennig, 2017, for a recent account). Still, it is worth mentioning two arguments to address this recurrent criticism. Firstly, we will argue that both frequentist and Bayesian approaches actually involve some degrees of subjectivity – each of a different kind – to lay down a general argument that the notion of "objectivity" in data analysis is more of an illusion than a reality. Secondly, we will argue that subjectivity is a vital component in Bayesian methods without which a proper progress in psychological sciences could hardly take place. Before we begin, we emphasize that we do not wish to give a false impression that these are definite answers to the problem. This complex issue has no unique solution over which statisticians and philosophers alike would come to an agreement. Instead, approaching it ultimately leads to endorsing a particular epistemological tradition on how to do science.

Bayesian school of inference has a well-established reputation of being a controversial approach to statistics, partly because of their subjectivity (e.g., Gelman, 2008, 2011; Sprenger, 2018; Wagenmakers, Marsman, et al., 2018). The objection is generally grounded in a conviction that statistical analyses should be objective in order to warrant the objectivity of scientific conclusions and should be based on data and nothing else. Such a perception goes hand in hand with a belief that objectivity in data analysis is attainable through the frequentist approach with the *p*-values (e.g., Berger & Berry, 1988; Gelman, 2011), paired with an instinct to escape from subjectivity at any cost. After all, *p*-values can be obtained straightforwardly based on observed

data and the statistical model assuming the null hypothesis, irrespective of anyone's idiosyncratic judgments and beliefs about the research question at hand. However, several authors have thoroughly illustrated that even the *p*-values are not exempt from subjective choices, such as researcher's sampling and testing intentions (Berger & Berry, 1988, Dienes, 2008; Kruschke, 2014; Kruschke & Liddell, 2018b; Wagenmakers, 2007; Wagenmakers et al., 2008; Sprenger, 2018). For instance, assume that a particular data set of 100 participants yields a significant effect, say a $p < .03$. Most likely, you implicitly assumed that the data set was collected with the following sampling and testing intentions: "1) collect 100 participants, then 2) compute the *p*-value". However, the same data set could have been obtained with different sampling and testing intentions, such as "1) collect 50 participants, then 2) compute the *p*-value and if not significant, then 3) collect 50 other participants, then 4) compute the *p*-value. The latter scenario corresponds to a sequential analysis and would require dividing the α threshold by the total number of performed tests, in which case the $p < .03$ would no longer remain statistically significant at $\alpha/2 = .05/2 = .025$. As a matter of fact, the same data set could have been obtained with a whole variety of different sampling and testing intentions that may eventually lead to opposite conclusions (see Dienes, 2008, Chapter 3; Kruschke & Liddell, 2018b; Wagenmakers, 2007, for a more detailed account of the issue; also see Wagenmakers et al., 2008, Section 2.3, for a humorous illustration). In the same example, the significance of the *p*-value may also be contingent on whether the test was one- or two-tailed, or whether it was predicted before or after the data was collected (e.g., Dienes, 2008; Sprenger, 2018). Such a dependency on *subjective* sampling and testing intentions thus leaves the door open for situations in which a single data set leads to different statistical conclusions, which is precisely the main criticism against Bayesian inference. Now, we hasten to emphasize that the word "subjective" must not be regarded as ungrounded; rather, it refers to an *informed judgment* that must be justified in the light of other available options. Making informed judgments in frequentist statistics is part of the routine and there is nothing unusual about it. When one uses the *p*-values, one is expected to conduct power analyses that require committing to an appropriate effect size, which must be

justified. Indeed, the choice of a suitable alternative to the null hypothesis in power analyses also requires to make an informed judgment. Finally, the choice of α thresholds in frequentist inference is also left to the researchers' discretion. Researchers can use custom α thresholds in their work based on subjective cost-benefit considerations and are expected to justify their α (Lakens et al., 2018). Let alone, the usual α threshold fixed at .05 used in most of psychology journals is arbitrary and can be more or less stringent depending on the discipline. Likewise, Bayesian inference can also involve thresholds – as we are about to discover in the next section – that are no less arbitrary and can vary across journals and scientific communities, without a straightforward theoretical justification.

We insist that the point made here is not meant to undermine the validity of frequentist inference. Our point is that both Bayesian and frequentist approaches involve various degrees of subjectivity. While the latter is bound up with the choices such as the researcher's intentions to collect and test data or effect size specification in power analyses, the former is bound up with the choice of priors. This leads us to the more general statement put forward by many acknowledged authors, that objectivity in data analysis is illusory (Berger & Berry, 1988; Gelman & Hennig, 2017). The seemingly virtuous attitude of "letting the data speak" is simply incompatible with the wide variety of subjective choices that one needs to make in data analysis (Morey et al., 2016; Steegen et al., 2016; Silberzahn et al., 2018), such as model specification and building, measurement methods, data processing and construction (i.e. "cleaning") or yet outlier management. Accordingly, Gelman and Hennig (2017) argued that the potentially misleading terms of "objective" and "subjective" should be abandoned and replaced by more meaningful attributes such as "transparent", "consensual" or "context-dependent". *Suma sumarum*, subjectivity – as long as it is understood as scientifically informed and publicly disclosed judgments rather than personal or idiosyncratic opinions – can be argued to be widely present across statistical practices and in science in general.

We now turn to our second argument: subjectivity may be regarded as a desirable feature and actually one of the strengths of Bayesian statistics. As we outlined earlier, there are good reasons to appreciate the pragmatical advantages of using the "objective" default priors. Default priors are particularly praised for their lack of subjective informativeness, minimal and neutral assumptions and for their wide adaptability as a default strategy to address various research questions (e.g., Berger, 2000). One of the reasons for their popularity is that they are designed to allow inferences to capitalize on data. However, learning from data is not equal to testing theory-driven predictions, which is vital to the development of empirical sciences (e.g., Chalmers, 1982; Dienes, 2008). To bare on scientific progress, priors must reflect predictions that are theoretically meaningful. Theory-driven priors cannot be narrowed to mere default models since the latter will not yield substantive theoretical predictions (e.g., Dienes, 2014, 2019; Dienes & Mclatchie, 2018) nor can they simply summarize previously accumulated evidence (Krefeld-Schwalb et al., 2018). However, as Dienes (2016) points it out, formulating theory-driven priors is not a self-evident process: a theory will usually yield not one but several models that operationalize it. Firstly, because most psychological theories are not expressed with formal models but are stated in ordinary language (e.g., Fiske, 2004), which does not prompt unequivocal numerical predictions (Meehl, 1967). Secondly, because a thought must be given to auxiliary assumptions under which we assume the theory to work (Earp & Trafimow, 2015; Świątkowski & Dompnier, 2017). All in all, translating a theory into specific priors requires ascertaining what the theory actually predicts but also many other aspects, such as quality and relevance of past research related to the theory at hand, reflecting upon the range of plausible effect sizes, the auxiliary assumptions, etc. All of these involve a great deal of informed scientific judgments in the process. This leads us to the core of our argument: the subjectivity inherent to the Bayesian framework is transparent and can be constructively discussed and debated within a research community (Dienes, 2019; Kruschke, 2014; Morey et al., 2016, Sprenger, 2018). For example, Sprenger (2018) argued that the subjective approach to Bayesian inference fosters constructive criticism and illustrated it in the case of psi

research. All controversy put aside, conflicting parties could work on their disagreements because they held different assumptions that were transparently translated into different priors. In this sense, the inherent subjectivity of Bayesian inference can be considered as its strength as it encourages to outline the assumptions laying behind the scientific debate as precisely and transparently as possible.

Beyond Objective and Subjective: Specifying Informed Priors

The arguments discussed above lead us to the take-home message of this subsection: the distinction between "subjective" and "objective" Bayesian statistics may be somewhat spurious with respect to scientific inference (Dienes, 2014; Morey, 2017). On the one hand, scientifically meaningful priors should be formulated in full transparency, in a way that can be defended against a skeptical research audience and that invites for constructive criticism (e.g., Kruschke, 2014; Sprenger, 2018). On the other hand, they should convey theoretically meaningful and concomitant predictions about hypothesized effects (e.g., Dienes, 2014, 2016, 2019; Morey et al., 2016). It seems to us that current recommendations broadly available in the psychological literature align well with these two constraints. For instance, those authors who advocate for defaults priors clearly underscore the importance of customizing such priors depending on the researchers' background knowledge and research peculiarities (e.g., Morey, 2017; Rouder et al., 2009; Rouder et al., 2016; Wagenmakers, Marsman et al., 2018). Likewise, Dienes (2014) also stressed that his approach to Bayesian inference cannot be reduced to either of the two approaches – subjective or objective – which he put in an insightful way:

> "The approach is objective in that [it involves] rules of thumb that can act as (contextually relevant) defaults, where the probability distributions are specified in simple objective ways by reference to data or logical or mathematical relations inherent in the design. No example relied on anyone saying, "according to my intuition the mean should be two because that's how I feel" (cf. Howard et al., 2000, for such priors). But the approach is

> subjective in that […] only scientific judgment can determine the right representation of the theory's predictions given the theory and existing background knowledge; and that scientific judgment entails that all defaults are defeasible – because science is subject to the frame problem" (p.13)

A first practical consequence that follows is that default priors scaled at arbitrary parameter values should not be used mechanistically in every possible situation, irrespective of the research context. Such an attitude would be unwarranted for sound inference (see Gigerenzer, 2004, for "statistical rituals"), neither is it actually advised by the tenants of default priors. Arguably, the degree to which researchers can appeal to their judgments in constructing their priors should depend on their advancement in their research program when studying a particular effect. This point holds to the fact that researchers accumulate knowledge as they conduct successive studies within a broader research program, and hence do not possess the same knowledge at its beginning than at more advanced stages of it. Therefore, an interesting possibility could be to start with a default prior at the beginning of a research and tune it up as the research advances.

As such, using the default JZS prior as a starting point to formulate one's priors may seem appealing (see Dienes, 2014, 2019 and Dienes & Mclatchie, 2018, for alternative priors and useful strategies to construct them for hypothesis testing). First, the default JZS prior uses a Cauchy distribution and is centered on 0, so it assumes effect sizes close to 0 as the most plausible *a priori*. It is a conservative assumption for an alternative hypothesis that describes one's expectations about the world when H_0 is false. Also, recall that the default JZS is scaled at $r = .707$, so that 50% of the expected effect sizes falls within the [-.707, .707] range. Importantly, note that this default effect size range is arbitrary: it has no straightforward justification (Morey, 2017). While it may be well-suited in those psychology subfields that are interested in substantial effect sizes, it is less so in domains that have interest in smaller effects, such as social psychology. With this respect, Morey (2017) urged for customizing the JZS prior instead of relying on the $r = .707$ default value. Ideally,

the JZS prior's scale factor parameter should be adjusted to match the expected range of effect sizes typical to one's research field, or otherwise to a theoretically justified range. For instance, considering the mean effect size in social psychological research of .21 in terms of correlation coefficient (Richard et al., 2003), equivalent to .43 in terms of Cohen's *d*, researchers from this field may want to adjust the JZS prior's *r* scale to .43 for standardized mean difference when performing a *t*-test (see also Williams et al., 2017). This would be a default starting point when no relevant knowledge is available, and no predictions can be made. With clear hypotheses in mind and at more advanced stages of research programs, a finer tailoring of the JZS prior becomes desirable. Strong hypotheses will usually specify the direction of an effect, and so such directional hypotheses can be tested with semi-Cauchy prior distributions (Morey, 2017; Morey & Rouder, 2011). With enough expert knowledge at hand, it is also possible to center Cauchy prior distributions around more theoretically meaningful mean parameter values than the null (see Gronau et al., 2019, for the informed Bayesian *t*-test). After all, if the theory at hand is not utterly wrong, a non-zero parameter value is more likely to be true. Keep in mind that such customizations of default priors warrant adequate justifications. Finally, it is also recommended to use several priors – ranging from the default to the very informed – to assess whether conclusions vary substantially across different choices in prior specification (i.e. robustness checks; see Dienes, 2019 and Wagenmakers, Love et al., 2018).

Now that we reviewed some ways of choosing priors, we turn to discussing practical methods of the analyses. The next section introduces Bayesian hypothesis testing, discussing both theoretical underpinnings and practical ways of computation.

6. Bayesian Hypothesis Testing: The Bayes Factor

Bayesian hypothesis testing involves computing Bayes Factors (Dienes, 2011, 2016; Jeffreys, 1961; Kass & Raftery, 1995; Kruschke, 2011; Lee & Wagenmakers, 2014; Morey et al., 2016; Mulder & Wagenmakers, 2016; Rouder et al., 2009), which is the equivalent of frequentist

hypothesis testing. The Bayes Factor (BF) tests two hypotheses by comparing how much data are likely to have occurred under each (see Equation 3). It derives from Bayes' Theorem as follows:

$$\frac{p(H_1|D)}{p(H_0|D)} = BF_{10} \times \frac{p(H_1)}{p(H_0)}, \quad (2)$$

$$where\ BF_{10} = \frac{p(D|H_1)}{p(D|H_0)} \quad (3)$$

Equation (2) reads "Posterior odds equals Bayes Factor times Prior odds". Prior odds indicate to what extent one favors one hypothesis over another before getting data. Posterior odds are an updated version of prior odds: they indicate how much more probable is hypothesis H_1 than H_0, given data at hand. Eq. (2) makes it clear that the BF gives the amount by which prior beliefs about hypotheses must be adjusted once data are known, to get the ratio of posterior probabilities of these hypotheses. Interpreting the BF value is very straightforward: for instance, a $BF_{10} = 8$ means that the observed data are 8 times more likely to have occurred under H_1 (e.g., experimental hypothesis) than under H_0 (e.g., null hypothesis). A $BF_{10} = 1/5$ is equivalent to $BF_{01} = 5$, meaning that data is 5 times better predicted by H_0 than H_1. For more computational details on how the BF is calculated, see Appendix E.

Importantly, note that the notion of "prior" has two distinct and independent meanings in Bayesian hypothesis testing (see *Figure 3*; see also Stefan & Schönbrodt, 2017). The first meaning refers to prior probability of a hypothesis, which simply quantifies the degree to which one believes the hypothesis to be true before seeing any data. A prior in this meaning, for instance the $p(H_1)$ in Eq. (2), is used to calculate prior odds. The second meaning refers to prior probability of a model parameter $p_H(\theta)$: it is the distribution of probability over the range of values of a parameter θ as specified by the hypothesis H. In this sense, a prior can be understood as a model (Dienes, 2016, 2019; Dienes & Mclatchie, 2018): it gives a statistical operationalization of what is actually predicted by a hypothesis and is used to calculate the evidence that data provide for the hypothesis

(Appendix E shows exactly how the prior intervenes in the computation of the likelihood term $p(D|H)$). The prior in this meaning is, for instance, the Cauchy distribution in the default JZS we discussed earlier. With its scale parameter set at r = .707, this prior distributes the probability of the parameter of interest – the standardized effect size δ – so that one half of the expected effect sizes falls within the [-.707, .707] interval and the other half falls outside (see Rouder et al., 2009). In essence, the two meanings of "prior" refer to two independent questions in data analysis (Dienes, 2014, 2016): the first meaning relates to the question "how much does one believe in a hypothesis before seeing the data", while the second relates to the question "what does the hypothesis actually predict".

[Insert Figure 3 near here]

Furthermore, notice that in order to come up with posterior odds that indicate how probable is one hypothesis compared to another given data, one must necessarily commit to a ratio of prior odds and multiply it by the BF value. Although setting prior odds is entirely up to one's personal beliefs (Dienes, 2008), it is suggested to conveniently set this ratio at 1 (e.g., Rouder et al., 2009), so that both hypotheses are equally probable before data are collected. Thus, with prior odds set at 1, a BF_{10} = 8 is equivalent with saying that H_1 is 8 times more likely than H_0, given the data at hand. Although some proponents of the objective approach to Bayesian statistics would advocate for this practice, it still may be regarded as questionable. Specifically, setting the same prior probability to a point-hypothesis (e.g., the null hypothesis) and to an interval-hypothesis (e.g., the alternative) may be dubious. Keep in mind however that the BF can be interpreted on its own without any reference to prior odds. One can quantify the degree to which data support one hypothesis over another without having to state how much one believed in these hypotheses before seeing the data, as we discussed it in the paragraph above. The only prior information that must be specified then is the prior probability $p_H(\theta)$ that is necessary to compute the BF. Therefore, the BF can be interpreted as

a likelihood ratio that gives a measure of strength of evidence without having to specify the prior odds (see Dienes, 2014, 2016).

The BF is advocated as being a conceptually simple, theoretically sound and coherent measure of evidence (Dienes, 2014, 2016; Lodewyckx et al., 2011; Morey et al., 2016). The BF compares the predictive accuracy of two competing hypotheses to explain data, hence it is a direct expression of relative strength of evidence for one hypothesis over another (Lodewyckx et al., 2011; Morey et al., 2016). The BF is a continuous measure of evidence from which three possible conclusions follow: (1) there is evidence for the alternative (H_1) relative to the null (H_0) when $BF_{10} > 1$; (2) there is evidence for the null (H_0) relative to the alternative (H_1) when $BF_{10} < 1$; the data is insensitive to discriminate between the two hypotheses when $BF_{10} \sim 1$ (Dienes, 2014). According to a commonly used convention (see *Table 2*), the BF value must favor at least 3 times more one hypothesis over another for the evidence to be said "decisive". For instance, a $BF_{10} = 5$ would be regarded as "substantial" evidence in favor of (H_1) compared to (H_0) according to this convention, whereas the evidence of $BF_{10} = 2$ would be merely regarded as "anecdotal". Importantly however, some authors have raised concerns that such a categorization of continuous BF values into discrete and arbitrary labels is neither useful nor desirable (Morey, 2015, 2017). Morey (2015) argued for instance that labels such as "anecdotal" or "substantial" to interpret BF values may inspire different interpretations across researchers, and "impose an arbitrary, unjustified homogeneity on judgments of evidential strength". On top of that, relying too heavily on these cutoffs entails the same risk as academia already experiences with the $p < .05$ threshold. For instance, a BF = 2.9 could be deemed insufficient to give grounds to publication, whereas a BF = 3.1 could be deemed high enough, though their evidential weights are quantitatively similar.

[Insert Table 2 near here]

Since the BF tests two hypotheses simultaneously one against another, it is a symmetrical measure of evidence (Dienes, 2014, 2016; Dienes & Mclatchie, 2018). The BF can provide support either in favor of H_1 against H_0, or in favor of H_0 against H_1 (or indicate data insensitivity as mentioned above). When H_1 is true, BF_{10} will converge to plus infinity as the sample size increases, and it will converge to 0 when H_0 is true (Morey & Rouder, 2011). What follows is that the BF can be used to corroborate the null hypothesis (e.g., Dienes, 2014, 2016; Gallistel, 2009; Rouder et al., 2009). Some authors raised concerns however that taken literally, the null hypothesis is always false (e.g., Cohen, 1994; Meehl, 1978), limiting the interest of providing evidence for such a point-null hypothesis. The BF can also provide evidence for interval null hypotheses, asserting approximate rather than exact invariances (Morey & Rouder, 2011). In any case, remember that the BF is a relative measure of evidence: it supports a hypothesis only relative to another. In other words, the BF cannot prove a hypothesis to be true in an absolute sense, but only indicate that this hypothesis accounts for data better than a competing one. For instance, imagine that your data indicate a very small correlation, say r = .03. With a reasonable sample size, a BF comparing the probability of such data under a model assuming the effect to be non-zero relative to a model assuming the effect to be zero would most likely yield support for the latter, say $BF_{01} = 50$. It is important to notice that such a BF value does not mean the null hypothesis is true, but rather that the data are more likely to occur under the null compared to a supposedly reasonably specified alternative (see also Rouder et al., 2009, Figure 5).

For simple designs such as mean comparisons and when priors follow a (semi)normal distribution, BFs can be calculated analytically with online calculators (e.g., https://medstats.github.io/bayesfactor.html; also see Dienes, 2014). In more complex designs such as ANOVAs and multiple regression, computing BFs involve substantially greater computational difficulties, but they can be efficiently estimated by Savage-Dickey ratio using MCMC iterative methods (see Morey et al., 2011 and Wetzels et al., 2009, for more details). Among statistical programs computing BFs (Mulder & Wagenmakers, 2016), JASP (jasp-stats.org; JASP Team,

2016) and several R packages, including the *BayesFactor* (Morey & Rouder, 2015) or *bayesmeta* (Röver, 2020) may seem the most appealing for psychologists in daily needs, both multiplatform and freely available. JASP is a user-friendly software providing both frequentist and Bayesian features for most common models in psychology (e.g., ANOVA, linear regression, repeated measures). R users may use the *BayesFactor* package, covering mostly the same applications and allowing extra-flexibility typical for the software or *bayesmeta* for meta-analyses. Both *JASP* and *BayesFactor* perform omnibus and covariate tests (see Rouder & Morey, 2012). Importantly, remember that both *JASP* and *BayesFactor* are tuned to use the default JZS prior with default scales (e.g., $r = .707$ for the *t*-test). Though one can obtain BF values without having to explicitly specify the JZS *r* scale in JASP or *BayesFactor*, it is however of *utmost* importance not to rely on such default *r* scale values automatically, but rather to use priors that are meaningful with respect to the research context (see Section 5 and Rouder et al., 2009 for heuristics on how to specify the JZS scale). For detailed explanations on how to use JASP or *BayesFactor* package to compute BFs, we encourage readers to read the manuals (Morey & Rouder, 2015; Wagenmakers, Marsman et al., 2018; Wagenmakers, Love et al., 2018). For some examples on how to report BF analyses, see Chatard et al. (2017) and Rouder and Morey (2012).

Some limitations of the BF must be mentioned. Remember that the BF measures the strength of evidence for one hypothesis with reference to another, thus the evidence is relational. Incautious hypothesis testing may yield an apparent support for a hypothesis (e.g., $BF_{10} = 100$), which in reality conceals the fact that both hypotheses are not well supported by data, only one more strongly than the other (e.g., $BF_{10} = 0.0001 / 0.000001 = 100$). Importantly, BFs are also limited with respect to their sensitivity to prior specification (e.g., Kruschke, 2014; Liu & Aitkin, 2008; Rouder & Morey, 2012). Even BFs based on non-informative or weakly informative default priors do not fully address the concern about subjectivity, as BFs' values may strongly depend on the choice of priors, even with large sample sizes. Robustness checks should therefore be regularly performed by checking whether the BF value changes substantially if alternative priors are used.

Lastly, computing BFs for covariate testing in multiple regressions may be limited when predictors are highly correlated (Rouder & Morey, 2012), just as in frequentist statistics.

At this point, readers should have enough understanding of the BF to bring us to the closing point. We turn back to the conceptual discussion on priors to emphasize the practical consequences of customizing the JZS prior. Here we present a series of simulations that illustrate how the choice of priors can impact the accuracy of decisions based on a particular BF analysis. We consider a common setup involving an independent two-sample Student *t*-test. For each simulation, we specified the following: the magnitude of population effect size, the scale factor parameter of the prior (i.e. JZS' *r)* and the sample size. Effect sizes varied from 0 to 1 by steps of .10 in terms of Cohen's *d*. We used two JZS priors: 1) a JZS prior scaled at $r = .707$, assuming a median effect size of .707 in absolute value used by the default in the software 2) a JZS prior scaled at $r = .43$, assuming a median effect size of .43 in absolute value and corresponding to the mean effect size in social psychology. For the sake of simplicity, we refer to the former as the "default prior" and to the latter as the "informed prior". Finally, the simulations used several sample sizes, ranging from n=25 to n=100.

Each simulation unfolded as follows. First, we drew a random sample of size *n* from two normally distributed populations, each parameterized according to the desired effect size. We then computed a BF_{10} with a Bayesian two-samples *t*-test on that data, using one of the two priors as the alternative to test against H_0. If $BF_{10} < 1/3$, we decided to accept H_0; If $1/3 \leq BF_{10} < 3$, we considered the test as inconclusive; if $BF_{10} > 3$, we decided to reject H_0. Table 3 summarizes the results of repeating this procedure for an arbitrary high number of times. For comparative purposes and in line with our pedagogical approach, we also displayed statistical power proper to each scenario. The R code used to run the simulation is available online (https://osf.io/8x6kg).

[Insert Table 3 near here]

What we are most interested in here is the accuracy reached with small effect sizes depending on the prior. In domains like social psychology, small effect sizes are more usual than large effect sizes. The default prior assumed a medium/large median effect size. It was unreasonably large with respect to usual expectations of a social psychologist. Contrariwise, the informed prior assumed small effects to be more likely and it was meant to be more suitable. Note that overall, both priors performed roughly equally well for rejecting accurately H_0 for non-zero effects. However, stronger expectations for medium/large effect sizes under the default prior came at a cost in accuracy with respect to making wrong decisions. For instance, for effects ranging from .20 to .40 Cohen's *d*, the rate of wrong decisions to accept H_0 oscillated between 48% to 16% for $n = 50$, when using the default prior. In this case, the informed prior performed much better: for the same range of effect sizes and sample size, the rate of wrong decisions to accept H_0 was comprised between 17% and 4%. Now, where does this difference come from? Remember that a prior is an allocation of probability over the range of the possible values of a parameter. When the default prior assumes that that the median expected effect size is $d = .707$, it allocates 50% of its probability mass to effect sizes comprised between [-.707, .707]; the informed prior, on the other hand, allocates 50% of probability mass to effect sizes comprised between [-.43, .43]. In other words, the default prior spreads the probability mass to a greater extent for more extreme values and to a lesser extent for smaller values than the informed prior. If small effects are especially likely to be observed, then the corresponding likelihood will be more compatible with probability mass allocation under a narrower prior than under a wider prior. A narrower (informed) prior will be thus more competitive against the null hypothesis than a wider (default) prior. Consequently, BF values obtained with the former will be higher than those obtained with the latter. This in turn explains why the decisions based on the informed prior more often led to data insensitivity compared with the default prior.

Finally, notice the difference in accuracy obtained with this procedure when H_0 was true. In this case, decisions based on the default prior led to higher accuracy relatively to the informed prior. For instance, with a small sample size ($n = 25$), using the informed prior literally never led to

correctly accepting the null hypothesis, whereas this rate was around 47% under the same conditions when the default prior was used. This illustrates another trade-off in prior customization: narrower priors will constitute more severe tests against the null hypothesis than wider priors. Stated differently, a narrower prior will be more competitive to account for random variations around the null than will be a wider prior.

7. Bayes in Psychology: What Could Change?

In this section, we review how a greater reliance on Bayesian statistics could have non-trivial consequences on virtually each step of the process of scientific inquiry, from theory construction to publication and replication. Rather than providing an exhaustive account of each increment, this section aims at highlighting some important consequences of using Bayesian inference. We thus encourage readers to check the references we mention below for further readings.

Theory Development

Formulating the Alternative Hypothesis.

Arguably, among substantial consequences of a wider use of Bayesian statistics could be stronger theorizing in our field. Given that one must necessarily commit to a certain prior when performing Bayesian analyses, a thought must be given to what one actually expects (i.e. predicts) to find in data. In other words, Bayesian statistics force to explicitly state one's hypotheses with respect to studied effects (see Dienes, 2016). Common criticisms toward NHST users object that the latter never commit to any alternative hypothesis and the only one being actually tested is H_0, which often leads to rejecting it with only weak evidence (Rouder et al., 2016; Wetzels et al., 2011). As we mentioned earlier, this is partly because most psychological theories are not mathematically precise enough to allow specifying an *a priori* alternative hypothesis as expected in the NP approach (Meehl, 1967). We hope to have illustrated throughout the paper that the Bayesian framework provides flexible ways in which priors can be formulated to reflect researcher's expectations and hypotheses. On their road from exploration of effects to corroboration of theories,

the strength and the precision of researchers' expectations and hypotheses vary substantially. To begin with, a possible application of Bayesian statistics could be in exploratory research. In this case, one ignores the exact number of statistical tests to be performed, hence testing intentions are not known *a priori* (De Groot, 2014). Relying on frequentist statistics may be dubious, both because no specific alternative hypothesis is specified, and because one does not control error rates. Here, a $p < .05$ hardly means what it is intended to mean, and chances are high that it might be a false positive especially when extensive multiple testing is applied, which is typical to exploration. Bayesian analyses seem more adequate because they preserve their meaning regardless of testing intentions. Moreover, researchers can use defaults priors (such as the JZS) and customize them so that they match the range of effect sizes of interest for exploration. On the other hand, closer to the "confirmatory" end of the spectrum, more subjective or informed versions of priors can be used. As researchers' get experience with each study they conduct, their default priors can be flexibly shaped into more specific predictions, just as we discussed in Section 5. Building priors can be also based on the available literature (e.g., Dienes, 2019). This should be done with extra caution because of possible context dependency that could constitute some "hidden assumptions" of a theory (see Świątkowski & Dompnier, 2017). Stated differently, past research related to some effect of interest may exist, but it may turn out to be misleading, especially if it is produced within a specific culture or context.

Theorizing and Assessing Invariances.

Researchers can use Bayesian methods to corroborate H_0 (Dienes, 2016; Gallistel, 2009; Rouder et al., 2009), which is not possible with conventional null hypothesis significance testing. Although one does not need to use Bayesian statistics *per se* to corroborate the null hypothesis (see Lakens, 2017, for frequentist equivalence testing), it is reasonable to expect that with the increased popularity of Bayesian statistics, an increased number of researchers would become less "aversive" against corroborating the null hypothesis. Establishing invariances can represent theoretical interests (e.g., Gallistel, 2009; Morey et al., 2018). Moreover, researchers often want to ensure that

variables such as sex or age exert no influence on dependent variables, or that such variables do not interact with experimental treatments. In such cases, corroborating H_0 can be properly addressed with BFs instead of incorrectly with non-significant *p*-values.

Methodology

Sequential Analyses.

Arguably, using Sequential Bayes Factors analyses (SBF; Schönbrodt & Wagenmakers, 2017; Schönbrodt et al., 2017) could represent a major change for psychologists in the way they conduct their studies. Since Bayesian inference does not depend on sampling and testing intentions (i.e., multiple testing, collecting more data after initial testing; early data collection stopping), they allow a greater flexibility in research design than traditional methods. Here, one can monitor evidence throughout the process of data collection and stop it whenever it is deemed compelling enough. One can also collect more data once the planned sample size has been reached if the evidence is not informative enough. Both of these situations do not require to make any adjustments, as they would if one were to use frequentist statistics. Thus, SBF can be appealing because genuine *a priori* power analyses cannot always be performed, or they also may rely or inaccurate, overestimated effect sizes. SBF therefore guarantees that researchers do not end up with "non-significant" studies, and eventually be able to provide support either for H_0 or for H_1. Remember however that Bayesian statistics do not control for long term error rates. Using SBF for decisions to stop or continue data collection and testing should be performed very carefully if one still wishes to control error rates, which is possible with simulations (see Dienes, 2016 and Schönbrodt & Wagenmakers, 2017).

Data analysis

Intellectual Hygiene.

According to some authors, many researchers are interested in knowing the probabilities of their hypotheses being true (e.g., Cohen, 1994; Gigerenzer et al., 2004). In this regard, Bayesian statistics seem naturally appealing, because they actually give what researchers look for when they

interpret *p*-values as a measure of falsehood of H_0. Hence using Bayesian analyses would be a matter of intellectual hygiene. However, we also think that if the Bayesian framework were to become widely recognized in the field next to traditional methods, researchers would be more prone to acknowledge the important difference between the two conditional probabilities that are often conflated (e.g., Badenes-Ribera et al., 2016), namely the probability of some data given some hypothesis (e.g., a *p*-value) and the probability of some hypothesis given some data (e.g., a posterior). This would contribute to better statistical knowledge and education.

Measuring Statistical Evidence.

As we hope to have explained it in Section 2, Bayesian statistics answer a radically different kind of statistical question than frequentist statistics. Instead of using *p*-values to know what decision about hypothesis acceptance or rejection to make based on some data, Bayesian statistics and specifically BFs can be used to assess the degree to which some data support one hypothesis over another. This approach allows to measure statistical evidence data provide for a hypothesis, because we can measure the degree to which a conviction in a hypothesis must be changed relative to another based on the data. This approach allows assessing the evidential value of data for a theory because the data are interpreted in a strict relationship with the theory (Dienes, 2016; Morey et al., 2016). In practice, using BFs for decision making to accept hypotheses can sometimes lead to different conclusions than one would get using traditional statistics (see Dienes, 2011; Johnson, 2013; Wetzels et al., 2011).

No More "p > .05" Ambiguity.

Non-significant *p*-values are part of psychologists' daily routine, yet they are hard to interpret. They occur for two reasons: (1) either because H_0 is true; (2) or because data present too much variance and are insensitive to discriminate between H_0 and H_1. The *p*-value does not allow differentiating between these two situations, whereas the BF does. When BF_{10} is close to 1 this point toward data insensitivity, and when it is close to 0, it gives support for H_0 over H_1 (Dienes, 2014). Remember that when H_0 is true, the *p*-value does not converge to any specific value as the

sample size increases, whereas the BF_{10} converges to 0 (Morey & Rouder, 2011). Irrespective of one's preferences for frequentist or Bayesian statistics, Dienes (2014; 2017) suggests computing BFs for non-significant *p*-values in order to disambiguate their meaning.

Publication

More Balanced Literature.

Dienes (2016) argued that since Bayesian statistics can provide evidence in favor of the null hypothesis over an alternative, there is no reason not to publish such evidence to track when data goes against a theory. This could likely contribute to a more balanced record of published literature between the "positive" and "negative" findings, hence possibly reducing the publication bias issue.

Reduced QRPs.

Let there be no illusion about it: as with every tool, it is possible to engage in questionable research practices (QRPs) with Bayesian statistics (e.g., using observed data to shape priors *a posteriori*; post-hoc covariate inclusion in a model to reach evidence, etc.). However, according to Dienes (2016), a wider reliance on Bayesian analyses could contribute to reducing the problem of QRPs. Since Bayesian statistics do not depend on factors such as sampling intentions (e.g., optional stopping, collecting more data after initial testing…) and multiple comparisons, these are ruled out from possible options to "hack" data.

Replication

Assessing Replication Research.

Imagine you conduct a replication study of an effect with statistical power set at .95, and you get a $p = .21$. Did you replicate or not? For the reasons mentioned above, a non-significant *p*-value is hard to interpret in such cases, despite high statistical power (Dienes, 2016). As Dienes stated: "[…] knowing power alone is not enough; once the data are in, the obtained evidence needs to be assessed for how sensitively H_0 is distinguished from H_1" (2016, p. 80). In such cases, one could again rely on BF's capacity to make the three-way distinction to assess whether data support H_1, H_0 or are just insensitive to make the difference. This view on data could provide a more

informative view on replication studies than a simple "replicated/not-replicated" dichotomy (see also Etz, & Vandekerckhove, 2016).

8. Bayes Factors in Practice: A Commented Example with R Code

This paper would not have been complete without a practical illustration. In the last section, we draw from a published research in social psychology to illustrate how to conduct an actual Bayesian analysis with Bayes Factors. We use the data from Chatard et al.'s (2017) research and show how to compute Bayes Factors using the *BayesFactor* R package. The annotated R code needed to reproduce the analyses is also available through the OSF website (https://osf.io/8x6kg).

Chatard and colleagues' research (Chatard et al., 2017) addresses the question of automaticity in social comparison. The authors provide evidence that subliminal exposure (20 ms) to pictures of ultra-thin women is sufficient to increase body appearance anxiety. In the first part of their second experiment involving a within-subject design, female participants were subliminally exposed (vs. not exposed) to pictures of thin women while doing an unrelated task. Right after, they had to say how anxious they felt "right now" about their body and their weight. This constituted the main dependent variable of the experiment (coded "AnxietyPrime" and "AnxietyControl" in the R script). Then, participants were exposed a second time to a similar task and they were asked to guess whether a subliminal image was embedded in the task or not. The comparison between the "hits" (i.e. the picture was present and they saw it) and the "false alarms" (i.e. the picture was absent but they said it was present) served to check whether the exposure to the picture was indeed outside awareness. These two variables are coded as "Hits" and "FalseAlarms" in the R script. The authors tested two hypotheses.

First, Chatard et al. sought to test whether the presentation of stimuli was outside participant's consciousness. To this end, the authors compared participants' false alarms with hits and predicted that the latter should not be higher than the former. A frequentist paired-samples *t*-test yielded a $t(50) = -0.43$, $p = .66$, indicating that the two were not statistically significantly different.

Using the formula below, we computed a Bayesian paired-sample *t*-test that pits the null hypothesis against a default alternative hypothesis assuming a Cauchy prior scaled at $r = .707$:

```
1 / ttestBF (Hits, FalseAlarms, paired = TRUE ,rscale  = .707)
```

This analysis yielded a $BF_{01} = 5.99$ and means that the data were almost 6 times more likely under the null hypothesis than under a default alternative hypothesis. However, Chatard et al. used a more informed version of the Cauchy prior. They were specifically interested in testing an alternative directional hypothesis assuming that participants would have greater Hits than False Alarms. They thus used a semi-Cauchy prior scaled at the default value. We obtain the corresponding analysis with the following formula, by adding the "nullInterval" argument to the script and specifying that the Cauchy must be constrained to the positive side of the distribution:

```
1 / ttestBF (Hits, FalseAlarms, paired = TRUE, nullInterval = c(0,+Inf), rscale  = .707)
```

This analysis yielded a $BF_{01} = 8.92$ and means that their data were roughly 9 times more likely to occur under the null hypothesis than under the alternative hypothesis. It provides support to the claim that participants could not detect the presence of stimuli above chance level. Finally, Chatard et al. complemented their analyses with a robustness check to ensure that the same conclusion would follow if a narrower prior was used. Specifically, they further customized the JZS prior and used a one-side Cauchy scaled at $r = .20$:

```
1 / ttestBF (Hits, FalseAlarms, paired = TRUE, nullInterval = c(0,+Inf), rscale  = .20)
```

The robustness check yields a $BF_{01} = 3.03$ that is consistent with the analyses obtained above. Using conventional cut-offs for decision-making (Jeffreys, 1961) would lead to accept the null hypothesis in spite of the alternatives for every prior that was used.

Turning to the main analyses, Chatard et al. tested their experimental hypothesis assuming that participants would feel greater anxiety after being exposed to pictures with ultra-thin women (i.e. AnxietyPrime) than in the control condition (i.e. AnxietyControl). A frequentist paired-samples *t*-test, $t(50) = 3.36$, $p = .001$, indicated that the two variables were statistically significantly different. To start the Bayesian analysis, we can begin by computing a Bayesian paired-sample *t*-test that tests an alternative hypothesis assuming a default Cauchy prior scaled at $r = .43$ against the null hypothesis:

```
ttestBF (AnxietyPrime, AnxietyControl, paired = TRUE, rscale = .43)
```

The output of this analysis is a $BF_{10} = 22.59$. If we used the default $r = .707$ scale, the corresponding analysis would yield a value of $BF_{10} = 19.98$. Yet again, to the extent that Chatard et al.'s experimental hypothesis assumed a specific direction of their effect, it is more relevant to custom the prior and use a semi-Cauchy distribution as displayed below:

```
ttestBF (AnxietyPrime, AnxietyControl, paired=TRUE, nullInterval = c(0,+Inf), rscale = .707)
```

With this formula, we obtain a $BF_{10} = 39.93$ in line with Chatard et al.'s results. We could also run additional robustness checks to see to what extent the results are sensible to prior specification. Scaling the JZS prior to semi-Cauchy distributions scaled at $r = .43$ and at $r = .20$ yielded $BF_{10} = 45.12$ and $BF_{10} = 38.07$, respectively. Overall, the data provide substantial support to their hypothesis that a subliminal exposure to the thin ideal increases body appearance anxiety.

Conclusion

The primary purpose of this article was to get across to the reader the understanding what Bayesian statistics are and how they can be applied in a relevant way for psychological research. We hope to have achieved this goal and to have demonstrated that there is really nothing magical about Bayesian statistics. They simply rely on a different interpretation of probability and answer different kinds of questions than those addressed in frequentist statistics. We further hope that the present content will successfully challenge some frequent controversies and clichés surrounding the use of Bayesian statistics, such that they would be too complex, subjective, or only to be used to publish "null" findings. We believe the provided examples illustrated how interpreting Bayesian statistics is straightforward and intuitive. Importantly, modern statistical programs have also made Bayesian statistics easy to compute. Furthermore, though the subjectivist approach to understand probability distributions as states of beliefs can be indeed debatable (Gelman, 2008), we have discussed several ways in which current recommendations for specifying Bayesian priors – such as the use of default JZS priors – mitigate usual concerns about how researcher's subjectivity is involved in the analysis. Finally, we also hope that the article illustrated relevant ways of making psychological research benefit from using Bayesian statistics, beyond the sole possibility to corroborate the null hypothesis.

Finally, it is also important to stress that our intention here was not to provide a universal tool for statistical inference (see Gigerenzer & Marewski, 2015). We believe that both frequentist and Bayesian statistics have their utility in psychological science and should be viewed as complementary, rather than mutually exclusive statistical tools (e.g., Bayarri & Berger, 2004; Krefeld-Schwalb et al., 2018; Witte & Zenker, 2017). With this respect, the APA guidelines state that "Researchers in the field of psychology use numerous approaches to the analysis of data, and no one approach is uniformly preferred as long as the method is appropriate to the research questions being asked and the nature of the data collected" (APA, 2010, p.33).

Footnotes

[1] None of the authors advocates for an exclusive use of frequentist or Bayesian statistics. Both advocate for sound statistical practices.

[2] Elsewhere in the psychological literature, epistemic probabilities are also referred to as subjective probabilities (e.g., Dienes, 2011)

[3] Though partially inspired by the NP approach, current practices for justifying scientific claims in psychology rely almost exclusively on the *null hypothesis significance testing* (NHST). Here, one rejects the statistical hypothesis of nonexistence of the effect under investigation (H_0), if the probability of observing under this hypothesis one's set of data or a more extreme set of data (the *p*-value) is smaller than the conventional threshold of .05. Despite the undisputed popularity in academia, this procedure has no straightforward theoretical justification (Gigerenzer, 2004; Gigerenzer et al., 2004). NHST in its most common use consists both in rejecting and accepting statistical hypotheses on the one hand, and providing evidence against the null hypothesis on the other. This reflects a hybrid logic of competing statistical theories of Neyman and Pearson (1933) and Fisher (1956), respectively (see Gigerenzer et al., 2004). A commonly endorsed practice in NHST, inspired by Fisher's work, consists in computing *p*-values to reflect the falsehood of H_0, where lower *p*-values indicate stronger evidence against H_0 and allow being more confident that it is false, than higher *p*-values (see Wagenmakers et al., 2008, for why this is a less than an ideal way to quantify evidence against H_0). This is a quasi-Bayesian approach to NHST (Gigerenzer, 1993) that conflicts with a strict frequentist approach to inference, because the level of significance serves, in practice, the same role as epistemic probabilities. This is not warranted within the framework that considers statistical hypotheses only as *true* or *false*, hence mentions like *probable* or *improbable* do not apply.

[4] Assuming the data are being collected with the same sampling intentions as the original study (see Wagenmakers et al., 2008).

[5] It may seem counterintuitive that with a 99% chance of obtaining a positive test result when one does really carry the virus, the fact that the test was positive only makes the probability of carrying the virus still relatively as low as less than 35%. This is due to a relatively low prior probability for carrying the virus, namely 5% (see Bayes' Theorem application in Section 4).

[6] In many realistic applications, the exact value of $p(D)$ can be impossible to calculate due to computational difficulties (see Kruschke, 2014). This explains why Bayesian methods have been historically only seldom used. Scholars have known this approach to inference for centuries, yet only a limited subset of relatively simple models were mathematically solvable under Bayesian analyses (Kruschke, 2014). Without knowing $p(D)$, deriving the exact posterior probability distribution $p(H|D)$ directly from Bayes' Theorem is not possible. This is not a problem anymore, since recent developments in computational techniques known as Markov Chains Monte Carlo (MCMC; Gamerman & Lopes, 2006; Kruschke, 2014; Lunn et al., 2012; Rouder & Lu, 2005) extended the applicability of Bayesian statistics to a whole range of useful models for daily applications. Their usefulness relies on the fact that they allow to approximate posterior probability distributions without having to calculate $p(D)$.

[7] A simpler version of Bayes' Theorem states that the posterior is proportional to likelihood times prior.

[8] A Cauchy distribution is a Student *t* distribution with 1 degree of freedom.

Acknowledgements

We express our sincere gratitude to Julien Diard, Benoît Dompnier, Fabrizio Butera, Jean-Philippe Antonietti and Mathias Schmitz for providing valuable feedbacks on the paper. We wish to thank the two reviewers whose comments helped to improve this paper. We thank Victoria Davoine and Ocyna Rudmann for reading and commenting on the earlier versions of this paper. Finally, we would like to thank Alexandre Charpentier Poncelet for his help in proofreading.

Declaration of interests

The authors have no conflict of interests associated with the publication of this article.

References

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin*, *82*(2), 261–277. https://doi.org/10.1037/h0076477

American Psychological Association (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Badenes-Ribera, L., Frias-Navarro D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian Academic Psychologists. *Frontiers in Psychology, 7:*1247. https://doi.org/10.3389/fpsyg.2016.01247

Bandyopadhyay, P. S., & Brittan, G. (2010). Two dogmas of strong objective Bayesianism. *International Studies in the Philosophy of Science, 24*(1), 45-65. https://doi.org/10.1080/02698590903467119

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science, 1*, 58-80. https://doi.org/10.1214/088342304000000116

Bayes, T., & Price, M. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. Philosophical Transactions (1683-1775), 370-418.

Berger, J. O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statistical Association, 95*(452), 1269–1276. https://doi.org/10.1080/01621459.2000.10474328

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 1*(3), 385–402. https://doi.org/10.1214/06-BA115

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76*(2), 159-165. https://www.jstor.org/stable/27855070

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., … Johnson, V. (2017). Redefine statistical significance. *Nature Human Behavior, 2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Chalmers, A. F. (1982). *What is this thing called science?: An assessment of the nature and status of science and its methods. 2nd ed*. St. Lucia: University of Queensland Press.

Chatard, A., Bocage-Barthélémy, Y., Selimbegović, L., & Guimond ,S. (2017).The woman who wasn't there: Converging evidence that subliminal social comparison affects self-evaluation. *Journal of Experimental Social Psychology, 73*, 1–13. http://dx.doi.org/10.1016/j.jesp.2017.05.005

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

De Groot, A. D. (2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta Psychologica, 148*, 188–194. https://doi.org/10.1016/j.actpsy.2014.02.001

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 78*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Dienes, Z. (2017, March 23–25). *Conventions for using Bayes Factors* [Paper presentation]. International Convention of Psychological Science, Vienna, Austria.

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science, 2*(4), 364–377. https://doi.org/10.1177/2515245919876960

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychomonic Bulletin & Review*, 25, 207-218.

https://doi.org/10.3758/s13423-017-1266-z

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology, 6*, 621. https://doi.org/10.3389/fpsyg.2015.00621

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193–242. https://doi.org/10.1037/h0044139

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLOS ONE 11*(2): e0149794. https://doi.org/10.1371/journal.pone.0149794

Etz, A., & Vandekerckhove, J. (2018). Bayesian inference for psychology. *Psychonomic Bulletin & Review, 25*, 5–34. https://doi.org/10.3758/s13423-017-1262-3

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh, UK: Oliver & Boyd.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439–453. https://doi.org/10.1037/a0015251

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis, 3*(3), 445–449. https://doi.org/10.1214/08-BA318

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals, 2*, 67-78.

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180*(4), 967-1033. https://doi.org/10.1111/rssa.12276

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*(1), 8-38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis, *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G., Krauss, S., & Vitouch. (2004). The Null Ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41*(2), 421– 440. https://doi.org/10.1177/0149206314547522

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian analysis, 1*(3), 403-420. https://doi.org/10.1214/06-BA116

Gronau, Q. F., Ly, A., & Wagenmakers, E. J. (2019). Informed Bayesian t-tests. *The American Statistician, 74*(2), 137-143. https://doi.org/10.1080/00031305.2018.1562983

Hoijtink, H., & Chow, S.-M. (2017). Bayesian hypothesis testing: Editorial to the Special Issue on Bayesian data analysis. *Psychological Methods, 22*(2), 211-216. http://dx.doi.org/10.1037/met0000143

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5*(3), 315. https://doi.org/10.1037/1082-989X.5.3.315

JASP Team (2017). JASP (Version 0.8.2) [Computer software]. https://jasp-stats.org/

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist, 80*(1), 64–72. http://www.jstor.org/stable/29774559

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press: Clarendon Press.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences, 110*(48), 19313–19317. https://doi.org/10.1073/pnas.1313476110

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Krefeld-Schwalb, A., Witte, E. H., & Zenker, F. (2018). Hypothesis-testing demands trustworthy data—a simulation approach to inferential statistics advocating the research program strategy. *Frontiers in psychology, 9*, 460. https://doi.org/10.3389/fpsyg.2018.00460

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312. https://doi.org/10.1177/1745691611406925

Kruschke, J. K. (2014). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan* (2nd ed.). Boston: Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*(1), 155-177. https://doi.org/10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2018b) The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206. https://doi.org/10.3758/s13423-016-1221-4

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour, 2*, 168-171. https://doi.org/10.1038/s41562-018-0311-x

Laplace, P. S. (1814). *Essai Philosophique sur les Probabilities [A philosophical essay on probabilities]*. New York, NY: Courcier.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association, 103,* 410–423. https://doi.org/10.1198/016214507000001337

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician), 49*(3), 293–337. https://doi.org/10.1111/1467-9884.00238

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52*(6), 362–375. https://doi.org/10.1016/j.jmp.2008.03.002

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E. J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology, 55*(5), 331–347. https://doi.org/10.1016/j.jmp.2011.06.001

Lunn, D. J., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Morey, R. D. (2015, January 30). On verbal categories for the interpretation of Bayes Factor. *BayesFactor: An R package on Bayesian data analysis.* http://bayesfactor.blogspot.com/2015/01/on-verbal-categories-for-interpretation.html

Morey, R. D. (2017, March 23–25). *Bayes factors, BayesFactor and 'default' priors* [Paper presentation]. International Convention of Psychological Science, Vienna, Austria.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406–419. https://doi.org/10.1037/a0024377

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor R package* (Version 0.9.12-2) [Computer Software].

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology, 55*(5), 368–378. https://doi.org/10.1016/j.jmp.2011.06.004

Morey, R. D., Homer, S., Proulx, T. (2018). Beyond statistics: Accepting the null hypothesis in mature sciences. *Advances in Methods and Practices in Psychological Science, 1*(2), 245-258. https://doi.org/10.1177/2515245918776023

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments." *Journal of Mathematical Psychology, 78*, 1–5. https://doi.org/10.1016/j.jmp.2016.01.002

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*(1), 79–95. https://doi.org/10.3758/BF03210778

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, 231*, 289–337. https://doi.org/10.1098/rsta.1933.0009

R Core Team. (2015). R: A language and environment for statistical computing. http://www.R-project.org/.

Richard, F. D., Bond, C. F. Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573–604. https://doi.org/10.3758/BF03196750

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for model selection in regression. *Multivariate Behavioral Research, 47*(6), 877–903. https://doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8*, 520–547. https://doi.org/10.1111/tops.12214

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Röver, C. (2020). Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software, 93*(6),1-51. https://doi.org/10.18637/jss.v093.i06

Schönbrodt, F. D., & Wagenmakers, E.-J. (2017). Bayes Factor Design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*.

https://doi.org/10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes Factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322-339. https://doi.org/10.1037/met0000061

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... & Carlsson, R. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337-356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sprenger, J. (2018). The objectivity of subjective Bayesianism. *European Journal for Philosophy of Science, 8*(3), 539-558. https://doi.org/10.1007/s13194-018-0200-1

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702-712. https://doi.org/10.1177/1745691616658637

Stefan, A., & Schönbrodt, F. D. (2017, January 17). Two meanings of priors, part II: Quantifying uncertainty about model parameters. http://www.nicebread.de/two-meanings-of-priors-2/

Świątkowski, W., & Dompnier, B. (2017). Replicability Crisis in Social Psychology: Looking at the Past to Find New Pathways for the Future. *International Review of Social Psychology, 30*, 111–124. http://doi.org/10.5334/irsp.66

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*(3), 842–860. https://doi.org/10.1111/cdev.12169

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217-239. http://dx.doi.org/10.1037/met0000100

von Mises, R. V. (1957*). Probability, Statistics, and Truth.* New York: Dover Publications, Inc.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, NY: Springer.

Wagenmakers, E. J, Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432. https://doi.org/10.1037/a0022790

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, L., … Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25,* 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E. J., Love, L., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25,* 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wetzels, R., Grasman, R. P., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for ANOVA designs. *The American Statistician, 66*(2), 104–111. https://doi.org/10.1080/00031305.2012.695956

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology an empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E. J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review, 16*(4), 752–760. https://doi.org/10.3758/PBR.16.4.752

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review, 19*(6), 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Williams, M., Bååth, R. A., & Philipp, M. C. (2017). Using Bayes Factors to test hypotheses in developmental research. *Research in Human Development, 14*, 1–17. https://doi.org/10.1080/15427609.2017.1370964

Williamson, J. (2010). *In defence of objective Bayesianism*. Oxford: Oxford University Press.

Witte, E. H., & Zenker, F. (2017). From discovery to justification: outline of an ideal research program in empirical psychology. *Frontiers in psychology, 8*, 1847. https://doi.org/10.3389/fpsyg.2017.01847

Appendices

Appendix A. Definition of conditional probability

For two events *a* and *b*, with *p(b)* > 0,

$$p(a|b) = \frac{p(a,b)}{p(b)}$$

The probability that will *a* happen given that *b* has occurred, is the probability of *a* and *b* happening together (i.e. joint probability), divided by the probability of *b*.

Appendix B. Formulas for calculating posterior normal distributions.

The normal posterior distributions for Figures 1A, 1B and 1C can be conveniently computed with the following formulas for normal distributions (from Dienes, 2008, p.94), based on information about prior distributions and the likelihood provided below.

Mean of prior = M_0

Mean of likelihood function = M_d

Standard-deviation of prior = SD_0

Standard-deviation of likelihood (standard-error of sample) = SE

Precision or prior = $c_0 = 1/SD_0^2$

Precision of likelihood = $c_d = 1/SE^2$

Posterior precision = $c_1 = c_0 + c_d$

Posterior mean = $M_1 = (c_0/c_1) * M0 + (c_d/c_1) * M_d$

Posterior standard-deviation = SD_1 = square root$(1/c_1)$

For all figures: Mean of likelihood = 4.5, Standard-deviation of likelihood = 2.3

Figure 1A: Mean of prior = 2, standard-deviation of prior = 2.7

Figure 1B: Mean of prior = 0, standard-deviation of prior = 7

Figure 1C: Mean of prior = 0, standard-deviation of prior = 1.08

Appendix C. Deriving the Bayes' Theorem

For two events *a* and *b*, we know that :

$$p(a|b) = \frac{p(a,b)}{p(b)} \rightarrow p(a|b)p(b) = p(a,b) \text{ , for } p(b) > 0$$

and alternatively,

$$p(b|a) = \frac{p(b,a)}{p(a)} \rightarrow p(b|a)p(a) = p(b,a) \text{ , for } p(a) > 0$$

Since $p(a,b) = p(b,a)$,

$$p(a|b)p(b) = p\ (b|a)p(a)$$

By dividing the last equation by $p(b)$, we get Bayes' Theorem:

$$\boldsymbol{p(a|b) = \frac{p\ (b|a)p(a)}{p(b)}}$$

Appendix D. Rule of marginalization

For a set of data D and a parameter θ we can deduce from axioms of probability that:

$$p(D,\theta) = p(D|\theta)p(\theta)$$

which means that the joint probability of D and θ – $p(D,\theta)$ – equals the conditional probability of D given θ – $p(D|H)$– times the probability of θ, $p(\theta)$.

Then:

- For a set of data *D* and a discrete-valued parameter θ, we know that:

$$p(D) = \sum_{\theta^*} p(D,\theta^*) = \sum_{\theta^*} p(D|\theta^*)p(\theta^*)$$

where θ* means that θ takes all possible values. This equation means that the probability of data can be expressed as function of (i.e. marginalized) the discrete values of θ. Specifically, the probability of data – $p(D)$ – is the sum for all values of θ of the products of the conditional probability of D given θ – $p(D|H)$ – and the probability of θ, $p(\theta)$.

- For a set of data *D* and a continuous-valued parameter θ, we know that

$$p(D) = \int p(D,\theta^*)d\theta^* = \int p(D|\theta^*)p(\theta^*)d\theta^*$$

where θ* means that θ takes all possible values. This equation means that the probability of data can be expressed as function of the continuous values of θ. Specifically, the probability of data – $p(D)$ – is the continuous sum (i.e. the integral $\int$) for all values of θ of the products of the conditional probability of D given θ –

$p(D|H)$ – and the probability of θ, $p(\theta)$. The symbol $d\theta^*$ simply means that the integration (the summing) is being performed over the parameter θ.

Appendix E. Bayes Factor as a ratio of two marginal likelihoods.

To get further insights on how the BF is calculated, it is necessary to consider what is actually predicted by each of the two tested hypotheses (see also Morey & Rouder, 2011, Rouder et al., 2009 and Stefan & Schönbrodt, 2017). Specifically, it is important to understand that each hypothesis can be given a concrete, statistical specification of what it actually predicts. In other words, a hypothesis can be operationalized by a model, which comprises a set of statistical parameters that appropriately describe the effect of interest. A model must specify the probability of observing some particular data given the model's structure and parameter values – $f_H(D|\theta_H)$ – along with prior probabilities of these parameter values $p_H(\theta)$ (Kruschke, 2014). For instance, the hypothesis used by Student 1 assuming a non-null mean score on introversion scale in the population of students with the disease can be modeled with two parameters, μ and σ^2, using a normal distribution to describe the data such that: $y \sim N(\mu, \sigma^2)$, where y refers to individual scores on the scale. Also, Student 1 used a normal prior on the parameter μ, such that $\mu \sim N(\delta_\mu, \sigma^2_\mu)$, and used arbitrary values for these priors, such that $\delta_\mu = 2$ and $\sigma^2_\mu = 2.7^2$. Since the prior $p_H(\theta)$ intervenes in the calculation of the likelihood $p(D|H)$, the specification of this prior will have an important consequence on the likelihood value, and thus on the BF value. Thus, a proper specification of prior probability on parameter values $p_H(\theta)$ is essential for a BF analysis to be meaningful. This property is apparent if we use the rule of marginalization (see Appendix D), to re-express a Bayes Factor in the following way:

$$BF_{10} = \frac{p(D|H_1)}{p(D|H_0)} = \frac{\int_{\theta_1 \in \Theta_{H_1}} f_{H1}(D|\theta_1) p_{H_1}(\theta_1) d\theta_1}{\int_{\theta_0 \in \Theta_{H_0}} f_{H0}(D|\theta_0) p_{H_0}(\theta_0) d\theta_0}. \quad (4)$$

Equation (4) shows that the BF can be expressed as a ratio of marginal likelihoods $p(D|H_1)$ and $p(D|H_0)$, which simply indicate how much the observed data D are likely to occur under each hypothesis, H_1 and H_0, respectively. Eq. (4) further shows how these marginal likelihoods are obtained (Morey et al., 2011; Rouder et al., 2009; Wetzels et al., 2009). The term "marginal likelihood" implies specifically that the likelihoods $f_H(D|\theta_H)$ will be averaged across all possible

values of θ (hence the integral $\int$ to sum across the parameter space Θ) according to the density function f_H of the data specified by the hypothesis *H*, weighted by their respective prior probability $p_H(\theta)$ specified by the hypothesis *H*. For instance, the literal transcription of the right part of Eq. (4) for H_1 could be then: "Across all values of parameters θ from the set of parameters defined by H_1, compute the continuous sum of the conditional probabilities of the data given the value of θ according to the density function f_H that are weighed by the prior probability accorded to the value of θ". In sum, a marginal likelihood refers to the probability of some data given some hypothesis, and can be understood as the continuous average of likelihoods over all constituent point parameters, where priors serve as weights (Rouder et al., 2009). Importantly, Eq. (4) shows that the choice of prior $p_H(\theta)$ is determinant for the computation of marginal likelihood $p(D|H)$. The fact that the likelihoods $f_H(D|\theta)$ are weighted by priors $p_H(\theta)$ across all parameter values procures desirable properties to BF (see Lodewyckx et al., 2011; Rouder et al., 2009; Wetzels et al., 2009, for more details). First, it ensures that BF operates according to the principle of parsimony in hypothesis testing (Jefferys & Berger, 1992; Myung & Pitt, 1997; Wetzels et al., 2009): when two statistical models specified by each hypothesis fit the data equally well, the simpler model is favored over the more complex one. This is because for the $p(D|H)$ to be high, the hypothesis *H* must both be able to predict accurately the observed data and not predict another, different set of data. Crucially then, an unreasonably high allocation of prior probability to large values of θ will tend to lower the value of $p(D|H)$, thus decreasing the degree to which the observed data *D* supports the hypothesis *H* (Rouder et al., 2009).

Tables

Table 1.

Summary of some important differences between frequentist (Neyman-Pearson approach) and Bayesian statistics.

	Frequentists statistics	Bayesian statistics
What is the aim in inference?	Decision-making about accepting/rejecting statistical hypotheses	Adjusting one's confidence in statistical hypotheses
What is the question answered?	"What should I do?"	"How should I adjust my beliefs?"
Do condition on the actual data?	No	Yes
Do provide with evidence?	No	Yes
Do control error rates?	Yes	No
Sensitive to multiple testing?	Yes	No
Sensitive to sampling rules?	Yes	No
Sensitive to the time of prediction?	Yes	No

Table 2.

Jeffrey's (1961; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011) classification to interpret the values of Bayes Factor.

Bayes Factor, BF_{10}			Interpretation
	>	100	Extreme evidence for H_1*
30	–	100	Very strong evidence for H_1*
10	–	30	Strong evidence for H_1*
3	–	10	Substantial evidence for H_1*
1	–	3	Anecdotal evidence for H_1*
	1		No evidence
1/3	–	1	Anecdotal evidence for H_0**
1/10	–	1/3	Substantial evidence for H_0**
1/30	–	1/10	Strong evidence for H_0**
1/100	–	1/30	Very strong evidence for H_0**
	<	1/100	Extreme evidence for H_0**

Note. "*": in comparison with H_0; "**": in comparison with H_1.

Table 3.

Summary of decision rates (percentages) for accepting and rejecting H_0 and declaring data inconclusiveness based on simulated Bayes Factors values using a default (r = .707) and informed (r = .43) JZS prior through a Bayesian independent-samples t-test.

			True population effect size (Cohen's d)										
			0	**.1**	**.2**	**.3**	**.4**	**.5**	**.6**	**.7**	**.8**	**.9**	**1**
N = 25/cell		Power	0,05	0,06	0,11	0,18	0,28	0,41	0,55	0,68	0,79	0,88	0,93
	Default prior (r = .707)	Accept H_0	0,47	0,45	0,39	0,29	0,20	0,12	0,07	0,03	0,01	0,01	0,00
		Inconclusive	0,52	0,52	0,57	0,62	0,65	0,62	0,56	0,45	0,33	0,21	0,14
		Reject H_0	0,02	0,03	0,05	0,09	0,15	0,27	0,38	0,52	0,65	0,78	0,86
	Informed prior (r=.43)	Accept H_0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
		Inconclusive	0,98	0,97	0,95	0,90	0,84	0,73	0,61	0,48	0,33	0,23	0,13
		Reject H_0	0,02	0,03	0,05	0,10	0,16	0,27	0,39	0,52	0,67	0,77	0,87
N = 50/cell		Power	0,05	0,08	0,17	0,32	0,51	0,70	0,84	0,93	0,98	0,99	1,00
	Default prior (r = .707)	Accept H_0	0,69	0,63	0,48	0,31	0,16	0,07	0,02	0,01	0,00	0,00	0,00
		Inconclusive	0,29	0,34	0,44	0,51	0,52	0,41	0,27	0,15	0,06	0,02	0,00
		Reject H_0	0,02	0,03	0,08	0,17	0,32	0,52	0,70	0,85	0,94	0,98	0,99
	Informed prior (r=.43)	Accept H_0	0,27	0,24	0,17	0,09	0,04	0,01	0,00	0,00	0,00	0,00	0,00
		Inconclusive	0,71	0,73	0,75	0,72	0,61	0,45	0,28	0,14	0,06	0,02	0,00
		Reject H_0	0,02	0,04	0,09	0,19	0,35	0,54	0,72	0,86	0,94	0,98	1,00
N = 75/cell		Power	0,05	0,09	0,23	0,45	0,68	0,86	0,95	0,99	1,00	1,00	1,00
	Default prior (r = .707)	Accept H_0	0,76	0,68	0,47	0,25	0,10	0,03	0,01	0,00	0,00	0,00	0,00
		Inconclusive	0,23	0,29	0,42	0,49	0,41	0,25	0,12	0,04	0,01	0,00	0,00
		Reject H_0	0,01	0,03	0,10	0,26	0,48	0,71	0,88	0,96	0,99	1,00	1,00
	Informed prior (r=.43)	Accept H_0	0,52	0,44	0,28	0,12	0,04	0,01	0,00	0,00	0,00	0,00	0,00
		Inconclusive	0,46	0,52	0,60	0,59	0,44	0,25	0,10	0,03	0,01	0,00	0,00
		Reject H_0	0,02	0,04	0,12	0,29	0,52	0,75	0,90	0,97	0,99	1,00	1,00
N = 100/cell		Power	0,05	0,11	0,29	0,56	0,80	0,94	0,99	1,00	1,00	1,00	1,00
	Default prior (r = .707)	Accept H_0	0,80	0,69	0,44	0,20	0,06	0,01	0,00	0,00	0,00	0,00	0,00
		Inconclusive	0,19	0,27	0,42	0,45	0,32	0,15	0,04	0,01	0,00	0,00	0,00
		Reject H_0	0,01	0,04	0,14	0,34	0,62	0,84	0,96	0,99	1,00	1,00	1,00
	Informed prior (r=.43)	Accept H_0	0,62	0,51	0,28	0,10	0,02	0,00	0,00	0,00	0,00	0,00	0,00
		Inconclusive	0,37	0,44	0,56	0,52	0,31	0,13	0,04	0,01	0,00	0,00	0,00
		Reject H_0	0,02	0,05	0,16	0,38	0,66	0,86	0,96	0,99	1,00	1,00	1,00

Note. Each scenario was simulated 5000 times. Power levels correspond to a two-tailed test for alpha = .05

Figures

Figure 1.

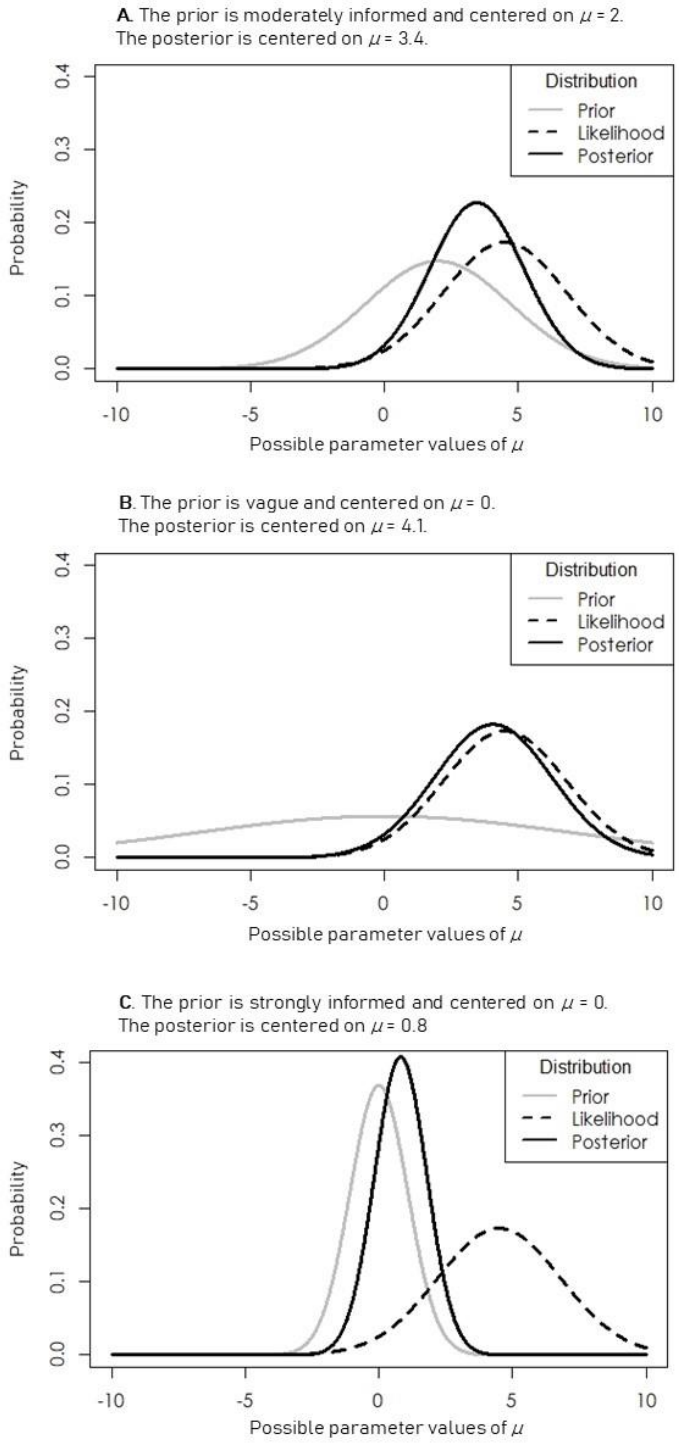

Figure 2.

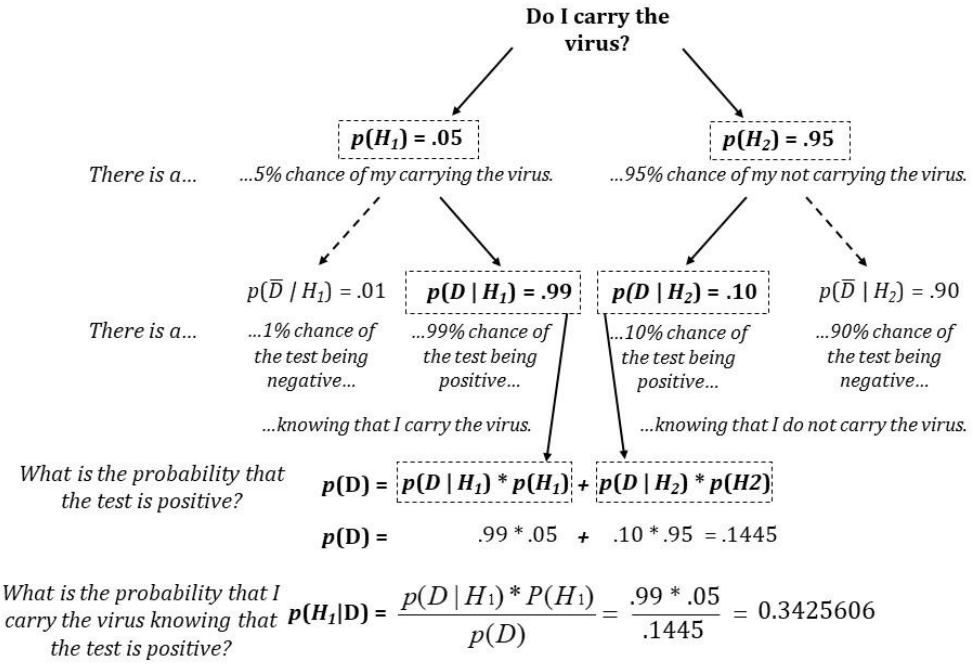
Do I carry the virus?
p(H1) = .05
p(H2) = .95
There is a...
...5% chance of my carrying the virus.
...95% chance of my not carrying the virus.
p(D̄ | H1) = .01
p(D | H1) = .99
p(D | H2) = .10
p(D̄ | H2) = .90
There is a...
...1% chance of the test being negative...
...99% chance of the test being positive...
...10% chance of the test being positive...
...90% chance of the test being negative...
...knowing that I carry the virus.
...knowing that I do not carry the virus.
What is the probability that the test is positive?
p(D) = p(D | H1) * p(H1) + p(D | H2) * p(H2)
p(D) = .99 * .05 + .10 * .95 = .1445
What is the probability that I carry the virus knowing that the test is positive?
p(H1|D) = p(D | H1) * P(H1) / p(D) = .99 * .05 / .1445 = 0.3425606

Figure 3.

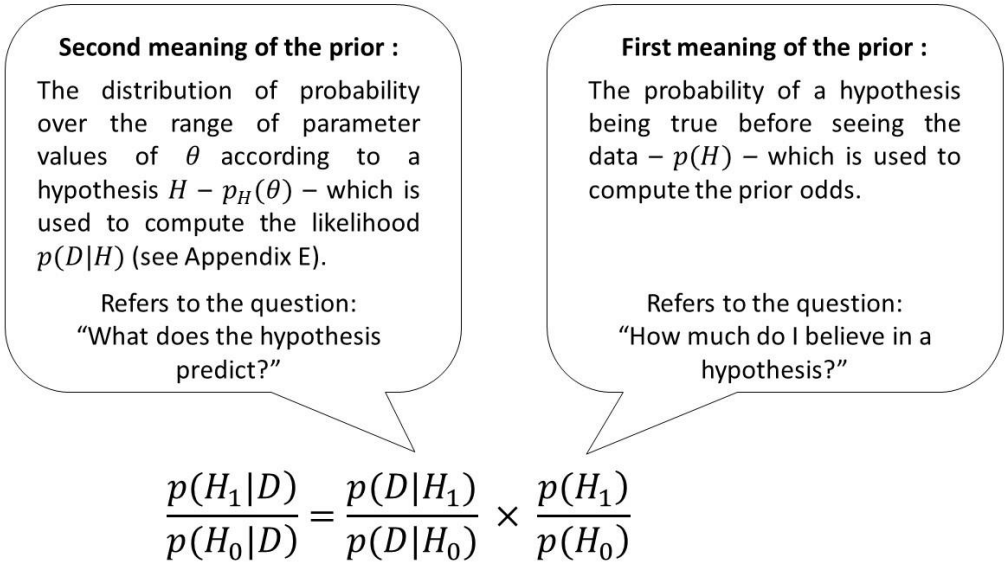
Second meaning of the prior :
The distribution of probability over the range of parameter values of θ according to a hypothesis $H - p_H(\theta)$ – which is used to compute the likelihood $p(D|H)$ (see Appendix E).
Refers to the question:
"What does the hypothesis predict?"
First meaning of the prior :
The probability of a hypothesis being true before seeing the data – $p(H)$ – which is used to compute the prior odds.
Refers to the question:
"How much do I believe in a hypothesis?"
$\frac{p(H_1|D)}{p(H_0|D)} = \frac{p(D|H_1)}{p(D|H_0)} \times \frac{p(H_1)}{p(H_0)}$

Figure captions

Figure 1. *Three Bayesian analyses based on each student's prior.*

Figure 2. *Conditional probabilities for the test result given the presence of the virus and application of Bayes' Theorem.*

Figure 3. *The two meanings of the prior in Bayes Factor hypothesis testing.*