The posterior probability of a null hypothesis given a statistically significant result

Daniel J. Schad
University of Potsdam, Potsdam, Germany.

Shravan Vasishth
University of Potsdam, Potsdam, Germany.

Abstract

Some researchers informally assume that, when they carry out a null hypothesis significance test, a statistically significant result lowers the probability of the null hypothesis being true. Although technically wrong (the null hypothesis does not have a probability associated with it), it is possible under certain assumptions to compute the posterior probability of the null hypothesis being true. We show that this intuitively appealing belief, that the probability of the null being true falls after a significant effect, is in general incorrect and only holds when statistical power is high and when, as suggested by Benjamin et al., 2018, a type I error level is defined that is lower than the conventional one (e.g., $\alpha = 0.005$). We provide a Shiny app (https://danielschad.shinyapps.io/probnull/) that allows the reader to visualize the different possible scenarios.

Keywords: Null hypothesis significance testing, Bayesian inference, statistical power

Null hypothesis significance testing (NHST), as it is practised in all areas of science, involves a fairly straightforward procedure. We begin by positing a null hypothesis H_0, usually a point null hypothesis that a parameter μ has a specific value: $H_0 : \mu = \mu_0$. Then we collect data, compute the sample mean $\bar{x}$, and estimate the standard error SE from the sample standard deviation s and sample size n by computing $SE = s/\sqrt{n}$. Next, we compute some statistic such as the observed t-statistic, $t_{observed} = \frac{\bar{x}-\mu_0}{SE}$. If the absolute value of the observed t-statistic is larger than the absolute value of some critical t-value, we reject the null hypothesis. Usually we also compute the p-value, which is the probability of obtaining the observed t-statistic, or a value more extreme, assuming that the null hypothesis is true. Conventionally, when the p-value is less than 0.05, we reject the null hypothesis. A common phrasing is to say that we have a "statistically significant" result, and that the

Please send correspondence to {schad,vasishth}@uni-potsdam.de.

effect of interest is "reliable." As is well known, an issue that is of great importance here is false and true discovery rates (Betancourt, 2018).

The false discovery rate is the probability of incorrectly rejecting the null when the null is in fact true; this is referred to as Type I error. It is conventionally fixed at 0.05. The true discovery rate is the probability of correctly rejecting the null when μ has some specific point value that is not the null value μ_0; this is usually called power. The quantity (1-power) is called Type II error, often written as β; it is the probability of incorrectly accepting the null when it is false with some specific value for μ, i.e., when the true μ is some specific point value other than μ_0.

It is well-known that power needs to be high in order to draw inferences correctly from a statistically significant result; when power is low, statistically significant results are *guaranteed* to be overestimates or even to have the wrong sign (Gelman & Carlin, 2014), and they are likely to be unreplicable (Vasishth, Mertzen, Jäger, & Gelman, 2018). In other words, a significant result under low power is *never* "reliable" in any sense of the word.

Researchers sometimes assume that a significant result changes *the probability of the null hypothesis being true.* For example, a survey by Tam et al. (2018) reports this widespread misunderstanding of the p-value by medical doctors. As they put it:

> Many respondents conceptualised the P value as numerically indicating the natural probability of some phenomenon — for instance, a 95% or 5% chance of the truth or falsity of a hypothesis in the real world.

A second example comes from Doherty, Benson, and Higham (2002). On page 376, Table 2, they state that "[the p-value] is the probability that the null hypothesis is true." A third example comes from a textbook written for medical researchers (Harris & Taylor, 2003). On their page 24, they write: "The P value is used when we wish to see how likely it is that a hypothesis is true"; and on page 26, they write: "The P value gives the probability that the null hypothesis is true."

Here, we investigate the posterior probability of the null hypothesis being true under different possible assumptions. Specifically, when power is low, medium, or high, and when Type I error is 0.05, 0.01, or 0.005.

Intuitively, it does seem obvious that rejecting the null hypothesis after finding a significant result leads us to change our belief about the probability of the null hypothesis being true. We will show in this paper that this intuitive belief is in general wrong, except in two extreme situations that are rarely or never realized: when statistical power is high (greater than 0.80) and Type I error is low (0.005).

Note that, technically, talking about "the probability that the null hypothesis is true" is meaningless in the NHST framework. Probability mass functions can only be associated with discrete outcomes that constitute a random variable. An example from everyday life would be the probability of catching a train when running late: there are two possible outcomes, either one gets the train or not, and each outcome has a probability associated with it. By contrast, the null hypothesis is not a random variable and therefore cannot have a probability associated with the two possible outcomes of being true or false. The null hypothesis is either true or it is false.

However, in order to talk about the probability of the null hypothesis being true or false, we can assume for the moment that the null hypothesis is a random variable. Suppose that before running the experiment, we begin with the assumption that the null hypothesis is believed to be true with some probability θ_{prior}. Once we get a significant result, the probability of the null being true should (intuitively) fall to some lower value $\theta_{posterior}$.

But these point values θ_{prior} and $\theta_{posterior}$ only partly characterize our beliefs. Before running the experiment, we surely have some uncertainty about the probability θ_{prior} that the null is true. For example, the null may be true at the outset with probability somewhere between 85 and 95%. Thus, the prior probability of the null being true has a probability *distribution* associated with it, it cannot be just a *point* value. Still relying on intuition, we might say that a statistically significant result should shift this probability distribution to some lower range, say 10-20%. Such a hypothetical situation is visualized in Figure 1a.

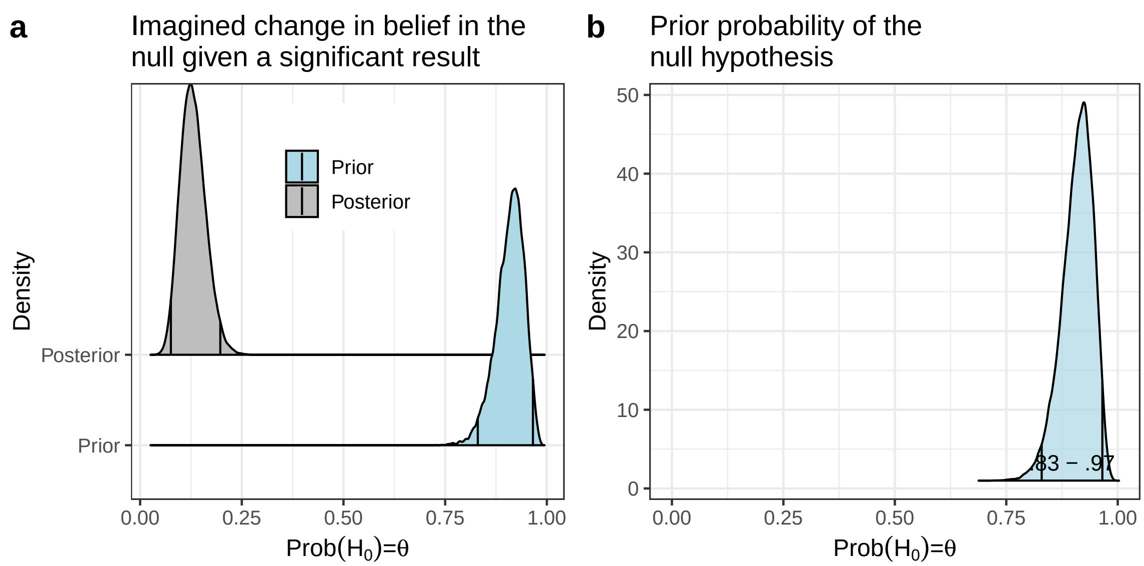

Figure 1. Prior and posterior probability of the null hypothesis. a) An illustration of how our belief in the null hypothesis—expressed as a probability distribution— might hypothetically shift once we see a statistically significant result. The vertical lines show the 95% credible intervals. b) Prior probability of the null hypothesis being true, expressed as a Beta(60,6) distribution.

To summarize, strictly speaking, the null hypothesis is either true or false, it has no probability distribution associated with it. So, one cannot even talk about the probability of the null hypothesis being true. Nevertheless, it is possible to relax this strict stipulation and ask ourselves: how strongly do we believe that the null is true before and after we do a significance test? A domain expert working in a particular field should be able to state, as a probability distribution, his or her a priori confidence level in a particular null hypothesis. In practice, some expert elicitation might be required (Oakley & O'Hagan, 2010; O'Hagan et al., 2006).

Bayes' rule allows us to calculate this posterior probability of the null hypothesis being true. Bayes' rule states that, given a vector of data y, we can derive the probability density function of a parameter or a vector of parameters θ given data, $f(\theta \mid y)$, by multiplying the likelihood function of the data, $f(y \mid \theta)$, with the prior probability of the parameter(s),

$f(\theta)$, and dividing by the marginal likelihood of the data, $f(y)$:

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)} \tag{1}$$

The marginal likelihood of the data can be computed by integrating out the parameter(s) θ:

$$f(y) = \int f(x, \theta)\, d\theta = \int f(y \mid \theta) f(\theta)\, d\theta \tag{2}$$

As McElreath (2016) mentions, Bayes' rule can be used to work out the posterior probability of the null being true given a significant effect. This is what we turn to next. Before we can carry out this computation, we have to decide on the prior probability of the null hypothesis being true; this is not too difficult to determine for specific research questions. We could start by eliciting from a researcher their prior belief about the probability of some particular null hypothesis being true: $Prob(H_0\ true)$. Given such a prior probability for the null hypothesis, we then stipulate a Type I error, $Prob(sig|H_0\ true) = \alpha$ and a Type II error, $Prob(not\ sig|H_0\ false) = \beta$. (When we write $H_0\ false$, we mean that the null is false with some specific value for the parameter μ). Once we have these numbers, Bayes' rule allows us to compute $Prob(H_0\ true|sig)$, the posterior probability of the null being true given a significant result. In other words, Bayes' rule helps us quantify the extent to which our prior belief should shift in the light of the long-run probability of obtaining a statistically significant result:

$$Prob(H_0\ true|sig) = \frac{Prob(sig|H_0\ true) \times Prob(H_0\ true)}{Prob(sig)} \tag{3}$$

The denominator $Prob(sig)$ is the marginal probability of obtaining a significant effect under repeated sampling, and is straightforward to compute using the law of total probability (Kolmogorov, 2018). This law states that the probability of a random variable Z, $Prob(Z)$, given another random variable A, is $Prob(Z|A)Prob(A) + Prob(Z|\neg A)Prob(\neg A)$. Translating this to our particular question, the event Z is the significant effect we obtained under repeated sampling, and the event A is the null hypothesis being true or false.

$$\begin{aligned} Prob(sig) =& Prob(sig|H_0\ true)Prob(H_0\ true) + Prob(sig|H_0\ false)Prob(H_0\ false) \\ =& \alpha \times Prob(H_0\ true) + (1-\beta) \times (1 - Prob(H_0\ true)) \\ =& \alpha \times \theta + (1-\beta) \times (1-\theta) \end{aligned} \tag{4}$$

The last line above arises because $Prob(sig|H_0\ true) = \alpha$ (Type I error), $Prob(sig|H_0 false) = 1-\beta$ (power), and $Prob(H_0\ true) = \theta$ (the prior probability of the null being true).

Thus, $Prob(H_0\ true|sig)$ is really a function of three quantities:

1. The false discovery rate, or Type I error α.

2. The true discovery rate, or power $(1-\beta)$, where β is Type II error.

3. The prior probability of null being true (θ).

We will now compute, under different assumptions, the posterior probability of the null being true given a significant effect. Before we can do this, we have to decide on a prior probability of the null hypothesis being true. What is a reasonable prior distribution to start with? In a recent paper, Benjamin et al. (2018) write the following: "Prediction markets and analyses of replication results both suggest that for psychology experiments, the prior odds of H1 relative to H0 may be only about 1:10. A similar number has been suggested in cancer clinical trials, and the number is likely to be much lower in preclinical biomedical research." A prior odds of 1:10 of the alternative being true relative to the null means that the probability of the null being true is about 90%. We take this estimate as a starting point; below we will also consider alternative scenarios where the probability of the null being true is lower.

For now, we will assume that the null hypothesis H_0 has the high prior probability of 90% of being true. Just as a coin has heads and tails as possible outcomes, the null hypothesis can have two possible outcomes, true or false, each with some probability. Thus, we can now talk about the probability θ of the null hypothesis being true. We can model this by assuming that a success or failure is generated from a Bernoulli process that has probability of success θ:

$$H_0 \sim Bernoulli(\theta) \tag{5}$$

Because our prior belief that the null is true will come with some uncertainty (it is not merely a point value), we can model this prior belief through a Beta distribution. For example, a prior $Beta(60, 6)$ on θ expresses the assumption that the prior probability of the null being true is between 0.83 and 0.97 with probability 95% (approximately), with mean probability approximately 0.90. The lower and upper bounds of the 95% credible interval can be computed using the inverse cumulative distribution function of the Beta distribution. We simply solve the integrals for the lower and upper bounds:

$$\int_{-\infty}^{lower} f(x)dx = \int_{-\infty}^{lower} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\, dx = 0.025 \tag{6}$$

and

$$\int_{-\infty}^{upper} f(x)dx = \int_{-\infty}^{upper} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\, dx = 0.975 \tag{7}$$

The mean can be computed from the fact that a random variable X that is generated from a Beta distribution with parameters a and b has mean:

$$E[X] = \frac{a}{a+b} \tag{8}$$

Given the a and b parameters, the mean is:

$$E[X] = \frac{a}{a+b} = \frac{60}{60+6} = 0.9 \tag{9}$$

Figure 1b visualizes this prior probability of the null hypothesis being true.

It is important to note that this prior probability does not state that the probability of the true mean μ being exactly 0 is this high. Rather, the claim is that our prior belief is about the null hypothesis distribution being $Normal(0, \sigma)$ (just using the normal distribution as an example). For example, in reaction time or reading data, we could assume that the difference between two conditions is $Normal(0, 5)$ on the milliseconds scale. This assumption entails that it is 95% probable that the true mean (which is usually a difference in means between two conditions being compared) lies between -10 and 10 ms, with mean 0. Such a null hypothesis could easily have a high prior probability of being true in many cases: Jäger, Engelmann, and Vasishth (2017) report in a meta-analysis involving about 100 reading studies that certain classes of working memory effects in sentence processing may be effectively 0 ms, with some uncertainty around this estimate (for example, 0 ± 6 ms in one case, see Table 4, p. 327 of Jäger et al., 2017). Such a meta-analytic estimate could be used as a starting point for defining a null hypothesis distribution when planning a future study.

Given this prior probability density function Beta(60,6) for θ, we are now in a position to investigate how the posterior probability of the null being true changes under different assumptions. We can use Monte Carlo sampling to compute the posterior probability of H_0 being true given significant results under repeated sampling:

1. Fix α (Type I error) and β (Type II error).
2. Do 100,000 times:
   (a) Sample one value θ from the Beta(60,6) distribution.
   (b) Compute posterior probability of θ given α, β, and the sampled value of θ from the prior distribution Beta(60,6):

$$\alpha \times \frac{\theta}{(\alpha \times \theta + (1 - \beta) \times (1 - \theta))} \tag{10}$$

   (c) Store this posterior probability of the null hypothesis being true.
3. Plot the distribution of the stored probabilities, or display summary statistics such as the mean and the 95% credible interval.

There are two interesting cases. The first is when statistical power is low (mean: 10%); we will show that in this case, it simply doesn't matter much if we get a significant result. The posterior probability of the null being true will not change substantially; this is regardless of whether Type I error is 0.05 or some lower value like 0.005, as recommended by Benjamin et al. (2018). The other case is where statistical power is high (mean: 90%); here, the posterior probability of the null being true will change considerably once we have a significant result under repeated sampling, especially if we follow the recommendation of Benjamin et al. (2018) to lower Type I error to 0.005.

Anticipating our main conclusion, when the prior probability of the null being true is low, the only situation where a statistically significant effect under repeated sampling can shift our belief substantially against the null hypothesis being true is when statistical power is high. When power is low, it simply doesn't matter whether you lower the Type I error to 0.005, as suggested by Benjamin et al. (2018) and others. Null hypothesis significance testing only makes sense if power is high; in all other situations, the researcher is wasting their time computing p-values. When power is low, the intuition that a significant result will lower the posterior probability of the null hypothesis being true is an illusion.

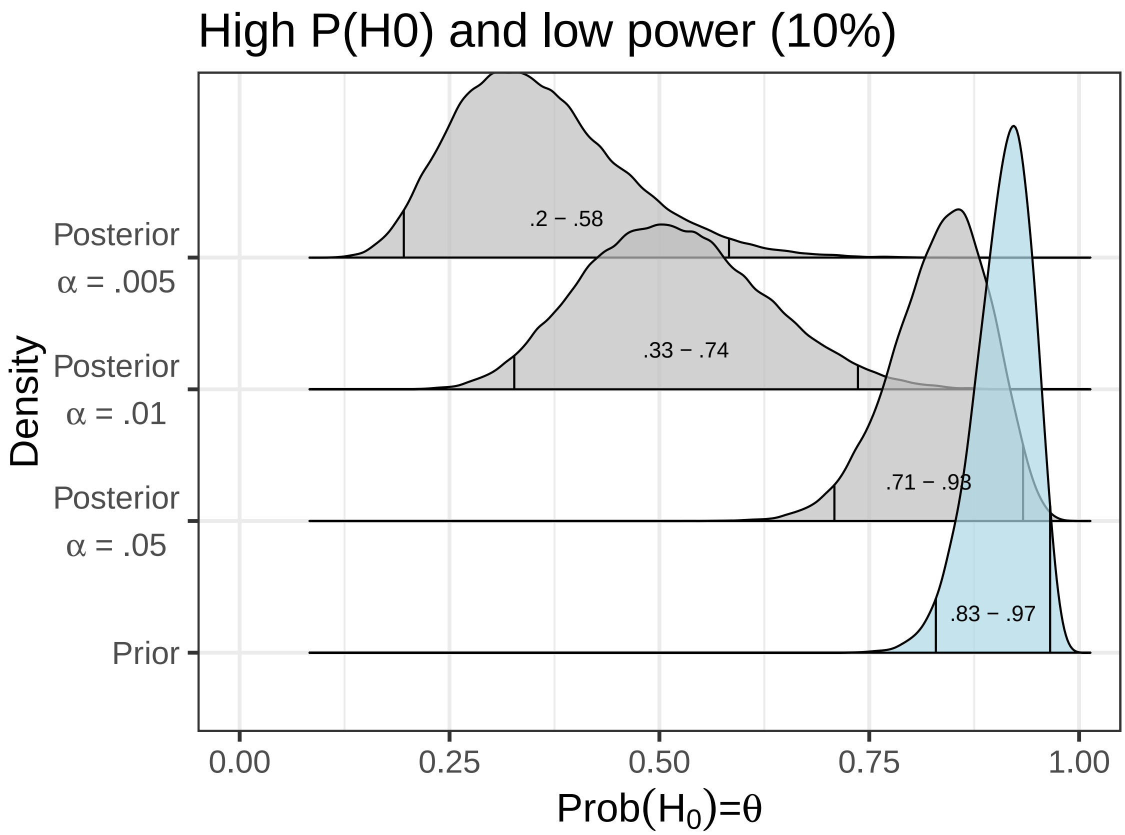

Figure 2. Probabilities for the Null-hypothesis, $P(H_0) = \theta$. Prior probability (blue) and posterior probabilities given a significant effect (grey) at a Type I error threshold α of 0.05 and 0.01. This is shown for a situation of low statistical Power of 10% (Type II error rate $\beta = 90\%$) and with a high prior probability for the null hypothesis ($\theta \sim Beta(60, 6)$; blue).

We next look at the posterior probability for different situations: As a first case, we investigate the posterior probability of the null hypothesis being true when the prior probability of the null hypothesis is high (Mean Prob(H0)=.90) and when power is low (Type II error 0.90). We investigate several scenarios by using different Type I errors ($\alpha = 0.05, 0.01$ and 0.005).

Scenario 1: Low power (0.10), Type I error 0.05. Let Type I error be $\alpha = 0.05$ and Type II error be $\beta = 0.90$. So, we have power at $1 - \beta = 0.10$. Such low power is by no means an uncommon situation in areas like psychology; examples are discussed in Jäger et al. (2017); Nicenboim, Roettger, and Vasishth (2018); Vasishth et al. (2018).

Figure 2 shows the prior and posterior distributions. The prior distribution is plotted in blue. Figure 2 shows that getting a significant result hardly shifts our belief regarding the null. This should be very surprising to researchers who believe that a significant result shifts their belief about the null hypothesis being true. Next, consider what happens when we reduce Type I error to 0.01, which is lower than the traditional 0.05.

Scenario 2: Low power (0.10), Type I error 0.01. Many researchers (Benjamin et al., 2018) have suggested that lowering Type I error will resolve many of the problems with NHST. Let's start by investigating what changes when we decrease Type I error to 0.01 (researchers like Benjamin et al., 2018 have proposed 0.005 as a threshold for Type I error; we turn to this proposal below). Type II error is held constant at 0.90.

Figure 2 shows that lowering Type I error does shift our posterior probability of the null being true a bit more but not enough to have any substantial effect on our beliefs. It seems unreasonable to discard a null hypothesis if the posterior probability of it being true lies between 30 and 70%.

Scenario 3: Low power, Type I error 0.05, incorporating uncertainty about Type II error. So far, we have been assuming a point value as representing power. However, power is really a function that depends (inter alia) on the magnitude of the true (unknown) effect. Power therefore also has some uncertainty associated with it, because we do not know the magnitude of the true effect, and we do not know the true standard deviation. We can introduce uncertainty about power (or equivalently, uncertainty about Type II error) into the picture by setting our prior on $\beta \sim Beta(10, 4)$, so that the Type II error is around 70%. Different levels of power (1-Type II error) are visualized in Figure 3, and the low power situation of 30% is shown in the bottom row of the figure.

Incorporating the uncertainty about Type II error (equivalently, power) increases the uncertainty about the posterior probability of the null quite a bit. Compare Figure 2 ($\alpha = 0.05$) and Figure 4a (low power). Figure 4a shows that the posterior of the null being true now lies between 40 and 90% (as opposed to 70 and 90% in Figure 2).

Scenario 4: Type I error 0.01, incorporating uncertainty in Type II error. Having incorporated uncertainty into Type II error, consider now what happens if we lower Type I error to 0.01, from 0.05. Figure 4d shows (cf. low power) that now the posterior distribution for the null hypothesis shifts to the left quite a bit more, but with wide uncertainty (10-60%). Even with a low Type I error of 0.01, we should be quite unhappy rejecting the null if the posterior probability of the null being true is between the wide range of 10 and 60%.

Scenario 5: Type I error 0.005, incorporating uncertainty in Type II error. Next, consider what happens if we lower Type I error to 0.005. This is the suggestion from Benjamin et al. (2018). Perhaps surprisingly, Figure 4g shows that now the posterior distribution for the null hypothesis does not shift much compared to Scenario 4 (see Fig. 4d): the range is 6 to 45% (compare with the range 10-60% in scenario 4). Thus, when power is low, there is simply no point in engaging in null hypothesis significance testing. Simply lowering the threshold of Type I error to 0.005 will not change much regarding our belief in the null hypothesis.

As a second case, we investigate the posterior probability of the null hypothesis being true when power is high. We consider the case where power is around 90%. We will still assume a high prior probability for the null (Mean Prob(H0) = .90). We will consider three scenarios: Type I error at 0.05, 0.01, and 0.005.

Scenario 6: High power (0.90), Type I error 0.05. First consider a situation where we have high power and Type I error is at the conventional 0.05 value. The question

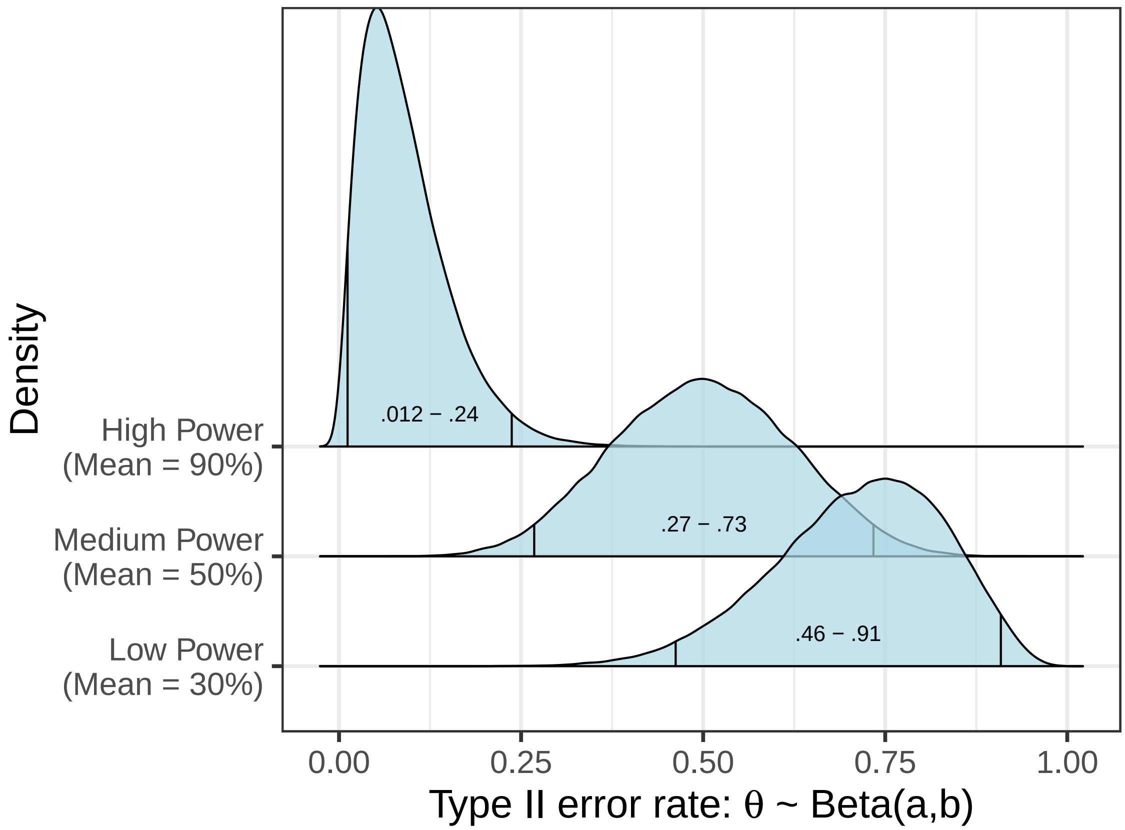

Figure 3. Visualization of the probability distribution associated with Type II error β corresponding to low power ($\beta \sim Beta(10, 4)$, mean power $E[1-\theta] = 30\%$), medium power ($\beta \sim Beta(8, 8)$, mean power $E[1-\theta] = 50\%$), and high power ($\beta \sim Beta(2, 20)$, mean power $E[1-\theta] = 90\%$) . Recall that Power is 1-Type II error.

here is: in high power situations, does a significant effect shift our belief considerably away from the null, with Type I error at the conventional value of 0.05? The prior on Type II error is shown in Figure 3. The mean Type II error is 10%, implying a mean for the power distribution to be 90%. Perhaps surprisingly, Figure 4a shows that even under high power, our posterior probability of the null being true does not shift dramatically: the probability lies between 20 and 60%.

Simulation 7: High power (mean 0.90), Type I error 0.01. Next, we reduce Type I error to 0.01. Figure 4d shows that when power is high and Type I error is set at 0.01, we get a big shift in posterior probability of the null being true: the range is 5-25%.

Simulation 8: High power (mean 0.90), Type I error 0.005. Next, in this high-power situation, we reduce Type I error to 0.005. Figure 4g shows that when power is high and Type I error is set at 0.005, we get a decisive shift in posterior probability of the null being true: the range is now 2-13%.

Finally, we consider cases where the prior probability for the null is medium or low.

Low prior probability for the null: Mean Prob(H0)=.10. One possible objection to the above analyses, however, is that the prior probability for the null hypothesis

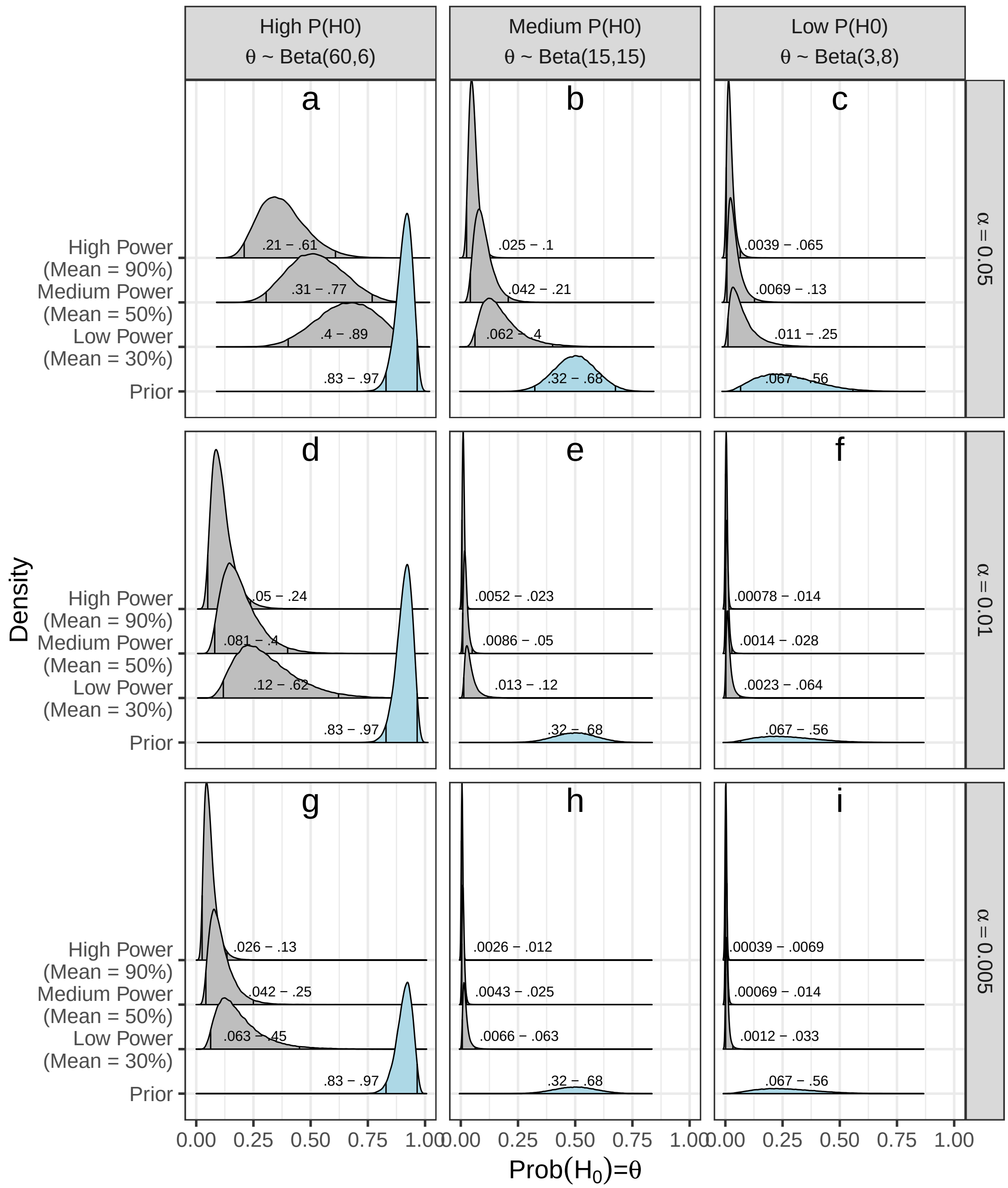

Figure 4. Probabilities for the null hypothesis, $P(H_0) = \theta$, considering uncertainty about power. Prior probability (blue) and posterior probabilities given a significant effect (grey) at a Type I error α of 0.05 (upper panels), 0.01 (middle panels), and 0.005 (lower panels). This is shown for situations of low statistical Power, $\beta \sim Beta(10, 4)$ (mean Type II error rate of about $\beta = 70\%$, mean Power of about 30%), medium statistical Power, $\beta \sim Beta(8, 8)$ (mean Type II error rate of $\beta = 50\%$, mean Power of 50%), and high statistical Power, $\beta \sim Beta(2, 20)$ (mean Type II error rate of about $\beta = 10\%$, mean Power of about 90%), and for situations with a high (left panels), medium (middle panels), and low (right panels) prior probability for the null hypothesis.

could often be much smaller than an average of 90%. Indeed, in some situations, the null hypothesis may be very unlikely. We here simulate a situation where the prior probability for the null is an average of 10% ($\theta \sim Beta(3, 8)$). For this situation, Figures 4c,f, and i show that the posterior probability for the null is always decisively low. Even for a conventional Type I error of $\alpha = 0.05$ in a low-powered study, the posterior probability for the null ranges from 1 to 25%, which is quite low, and when turning to smaller α levels or higher power, the effect is decisive.

However, of course this is not very informative, as we started out assuming that the null hypothesis was unlikely to be correct in the first place. Thus, we haven't learned much; a statistical significance test would just confirm what we already believed with high certainty before we carried out the test.

Medium prior probability for the null: Mean Prob(H0)=.50. Now consider the case where the prior probability for the null being true lies at an average of 50% (e.g., $\theta \sim Beta(15, 15)$). Here, we don't know whether the null or the alternative hypothesis is true a priori, and both outcomes seem similarly likely. In this situation, when we use a conventional Type I error level of $\alpha = 0.05$ in a low-powered study, a significant effect will bring our posterior probability for the null only to a range of 6-40%, and will thus leave us with much uncertainty after obtaining a significant effect.

However, either using a stricter Type I error level (e.g., $\alpha = 0.005$) or running a high-powered study each suffices to yield informative results: For a high-powered study and $\alpha = 0.05$, a significant result will (under our assumptions) bring the posterior probability to 2-10% (Figure 4b), which is quite informative. And for a Type I error level of $\alpha = 0.005$ a significant effect brings decisive evidence against the null for all the levels of power that we investigated (Figure 4h), with a posterior probabilty of 0.7-6% even for low-powered studies. This suggests that when the prior probabilities of the null and the alternative hypotheses are each at 50%, then either high power or a strict Type I error of $\alpha = 0.005$ will yield informative outcomes once a significant effect is observed.

Taken together, we analyzed the posterior probability for the null given a significant effect. We provide a shiny app (https://danielschad.shinyapps.io/probnull/) that allows computing the posterior distribution for different choices of prior, Type I and Type II error. For psychology and preclinical biomedical research, the prior odds of H1 relative to H0 are estimated to be about 1:10 (Benjamin et al., 2018), reflecting a high prior probability of the null of 90%. For this common and standard situation in psychology and other areas, when power is low, the posterior probability of the null being true doesn't change in any meaningful way after seeing a significant result, even if we change Type I error to 0.005. What shifts our belief in a meaningful way is reducing Type I error to say 0.005 (as suggested by Benjamin et al., 2018 and others), *as well as* running a high powered study. Only this combination of high power and small Type I error rate yields informative results.

One might object here that we set the prior probability of the null hypothesis being true at an unreasonably high value. This objection has some merit; although typically the prior probability for the null may lie at 10%, there may well be some situations where the null is unlikely to be true a priori. In this situation, our results show that a significant effect does indicate a very low posterior probability of the null. This is the case across a range of Type I error levels (α of 0.05, 0.01, 0.005) and for different levels of power (Figure 4c+f+i). Even for low power studies with $\alpha = 0.05$ the posterior probability is between

1 and 25%, which is quite low. So yes, if the prior probability of the null being true is already low, even with relatively low power and the standard Type I error level of 0.05, we are entitled to changing our belief quite strongly against the null once we have a significant effect. An obvious issue here is that if we already don't believe in the null before we do the statistical test, why bother to try to reject the null hypothesis? Even if we were satisfied with rejecting a null hypothesis we don't believe in in the first place, running low power studies is always a bad idea because of Type M and S error. As Gelman and Carlin (2014) and many others before them have pointed out, significant effects from low power studies will have exaggerated estimates of effects and could have the wrong sign of the effect. The probability of the null hypothesis being true is not the only important issue in a statistical test; accurate estimation of the parameter of interest is equally important.

In summary, we investigated the intuitive belief held by some researchers that finding a significant effect reduces the posterior probability of the null hypothesis. We show that this intuition is not true in general. The common situation in psychology and other areas is that the null hypothesis is a priori quite likely to be true. In such a situation, contrary to intuition, finding a significant effect leaves us with much posterior uncertainty about the null hypothesis being true. Obtaining a reasonable reduction in uncertainty is thus another reason to adopt the recent recommendation by Benjamin et al. (2018) to change Type I error to $\alpha = 0.005$. Furthermore, conducting high power studies is an obvious but neglected remedy. Otherwise, the results will be indecisive.

Our key result is that the posterior probability for the null given a significant effect varies widely across settings involving different Type I and Type II errors and different prior probabilities for the null. The intuition that frequentist p-values may provide a shortcut to this information is in general misleading.

Acknowledgements

Thanks to Valentin Amrhein and Sander Greenland for comments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 317633480, SFB 1287, Project Q, with principal investigators Shravan Vasishth and Ralf Engbert.

Author contribution

SV had the idea for the paper. SV and DJS performed analyses. SV and DJS generated the shiny app. SV and DJS wrote the paper.

Availability of simulations and computer code

All the computer code used for the simulations reported in the present manuscript, and the code for generating all Figures, will be freely available online at https://osf.io/9g5bp/. Moreover, we make a shiny app available at https://danielschad.shinyapps.io/probnull/ that allows computing the posterior probability for the null given a significant effect for many different settings of Type I and II error and for different prior probabilities for the null.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... others (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Betancourt, M. (2018). Calibrating model-based inferences and decisions. *arXiv preprint arXiv:1803.08393*.

Doherty, U. B., Benson, P. E., & Higham, S. M. (2002). Fluoride-releasing elastomeric ligatures assessed with the in situ caries model. *The European Journal of Orthodontics*, *24*(4), 371–378.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Harris, M., & Taylor, G. (2003). *Medical statistics made easy*. Scion Banbury, England.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316-339. doi: https://doi.org/10.1016/j.jml.2017.01.004

Kolmogorov, A. N. (2018). *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Vol. 122). CRC Press.

Nicenboim, B., Roettger, T. B., & Vasishth, S. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics*, *70*, 39-55. Retrieved from `https://osf.io/g5ndw/` doi: https://doi.org/10.1016/j.wocn.2018.06.001

Oakley, J. E., & O'Hagan, A. (2010). SHELF: The Sheffield Elicitation Framework (version 2.0) [Computer software manual]. University of Sheffield, UK. Retrieved from `http://tonyohagan.co.uk/shelf`

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons.

Tam, C. W. M., Khan, A. H., Knight, A., Rhee, J., Price, K., McLean, K., et al. (2018). How doctors conceptualise'p'values: A mixed methods study. *Australian Journal of General Practice*, *47*(10), 705.

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151-175. Retrieved from `https://osf.io/eyphj/` doi: https://doi.org/10.1016/j.jml.2018.07.004