Open Peer Review on Qeios

The Bayesian Range: Simple Methods for Incorporating Bayesian Reasoning into Psychological Research for Psychologists and Psychology Students Unfamiliar with Bayes' Theorem

Eric Prichard[1]

1 University of Arkansas - Monticello

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

Bayes' theorem is used to calculate posterior probabilities of an event occurring given that a second event occurs. Advanced Bayesian statistics are now commonly used in psychology, but simpler versions can be adopted by social psychologists and psychology students who have little or no prior experience with Bayes' Theorem using commonly obtained values. Using the post-hoc statistical power of a study, the selected alpha level for determining significance, and a subjective personal probability for the probability that a null hypothesis is false, it is possible to create a version of Bayes' Theorem that calculates the probability that a null hypothesis is false given the null hypothesis is rejected. Because the subjective personal probability of the null being false is an estimate, the paper recommends that for novel empirical findings, researchers take one of two approaches to incorporating Bayes' Theorem. The first is to calculate the posterior probability for a range of probabilities that the null is false. The second is to treat a novel finding as though it is unlikely to be true and give the probability that the null is false a low starting probability. With each replication, the posterior probability can be updated. Such a practice would encourage multiple replications of a novel finding before it can be reported with confidence. The emphasis of the paper is on the ready application of Bayes' Theorem for non-experts in Bayesian statistics who are more comfortable with traditional significance testing.

Eric C. Prichard, Ph.D.[*]

School of Social and Behavioral Sciences

University of Arkansas at Monticello

[*]**Corresponding Author Contact Information:**

University of Arkansas at Monticello
School of Social and Behavioral Sciences
UAM Box 3619
562 University Drive
Monticello, AR 71656
Email: prichard@uamont.edu

Keywords: Bayes' Theorem, Statistics, Significance Testing, Effect Size, Statistical Power.

A classic problem those of us who teach statistics to psychology students face is trying to explain the meaning of the p-value. For example, it can be an effort to get students to understand that a p of .01 does not mean that there is only a 1% chance the null hypothesis is true. It only means that an obtained statistical value of a given absolute magnitude or greater would be expected to occur in around 1/100 random samples of a given size if the null hypothesis were assumed to be true. However, even for trained psychologists, it is hard not to fall into the trap of seeing p-values as the probability the null is true versus false. Is there a way to estimate how likely it is that the null hypothesis is false given that one rejected the null hypothesis? Would doing so help encourage replication at an earlier stage of research? This brief paper is meant to provide a very simple application of Bayes' Theorem as a way of calculating a range of probabilities to give one an idea of how likely it actually is that a null hypothesis is false given the decision to reject the hypothesis.

A few words about the proposed technique. Mathematically speaking, nothing in this paper is groundbreaking. A lot of authors recommend using Bayesian statistical methods in psychological research as a supplement to or replacement for Null Hypothesis Statistical Testing, hereafter referred to as NHST (e.g., van de Schoot et al., 2017; Yalch, 2016). However, to the knowledge of the author, the particular application presented herein is not widespread and was derived from a very basic version of Bayes Theorem and general definitions taken from common practices in power analysis. In other words, while the author does not know of others applying Bayesian statistics in the way presented, the calculations are so simple that it would not be surprising if it has been done. More accurately, the author would be surprised if someone has NOT used similar calculations somewhere at some point. This paper makes no claim to be the first to calculate the posterior probabilities of the null hypothesis being false using the methods described below. Rather, the purpose of the paper is to show how psychologists and psychology students with little familiarity with Bayes' Theorem can apply a simple version of it, which is widely taught in upper-level high school and undergraduate level math classes, to put probabilistic boundaries around their confidence in novel empirical findings. Finally, as will be discussed further, the goal of this paper is not to convince researchers to replace NHST. Instead, it is to offer an accessible application of Bayes Theorem to the everyday researcher who may not be versed in Bayesian techniques, but who might benefit from incorporating Bayesian thinking into their research program. Those who are already well versed in Bayesian methods may find the recommendations rudimentary. The target audience is not the expert Bayesian, but the excitable social or personality psychology graduate student who gets a low p-value with an underpowered sample and (understandably) wants to shout their "significant" finding to the world. Bayes' Theorem can help temper those impulses and encourage replication.

The Basic Calculation

As stated in the introduction, this paper centers on the simple version of Bayes' Theorem that can be found in many textbooks at the high school and undergraduate level. Further, no special computer program is needed to apply the theorem in the way recommended. A calculator would do, provided the researcher already has some means of doing a post-hoc power analysis (e.g., G*Power).

Bayes' Theorem is a formula that deals with what are called inverse probabilities and uses known base rates or estimated base rates to calculate the probability of one event given that another event also occurs (Berkson, 1930). It is often used to answer questions like, "how likely is it that I have cancer given a positive test result after a cancer screening?" Using the test's sensitivity rate, its false positive rate, and the base rate of the cancer in question, one can come up with a posterior probability. To the surprise of many who are exposed to Bayes' Theorem for the first time, the posterior probability is often considerably smaller than a test's sensitivity would suggest. A test that provides a correct diagnosis 95% of the time cancer is present will still produce more false positives than hits if the base rate of the cancer is small. The common version of the formula taught in schools looks like this:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)}$$

P(A|B) means the probability of A given B. P(B|A) means the probability of B given A. The ~ symbol refers to negation, so P(B|~A) means the probability of B given A is false. Using the basic concepts of power analysis (Cohen, 1992), we can construct a Bayesian formula for estimating the posterior probability that a null hypothesis is false (A) given that one rejects the null hypothesis (B).

Later in the paper, methods for choosing a starting P(A) will be discussed. Typically, this is an unknown value, but researchers have an intuition or hypothesis about whether it will be likely that a hypothesis is true or false. For now, let us assume a researcher obtains a statistically significant test value and, prior to the study, was completely agnostic as to whether the null hypothesis was true or not. Their personal probability that the null is false can be set at 0.5. In Cohen's analysis, power is a measure of how likely one is to reject the null hypothesis if the null is false. Often power analyses are done to estimate effect sizes, but programs like G*Power can do post-hoc estimates of power based on the effect size obtained and the sample size used in a study. This post-hoc measure of power can stand in for P(B|A). P(~A) will simply be the inverse of P(A). To get P(~A), subtract P(A) from one. P(B|~A) is the probability the null is rejected given that the null is true. This value can be filled in using the α-level for a statistical test. If one chooses, say, 0.05, then one way to interpret that is, “there is a probability of approximately 0.05 of rejecting the null if the null is true.” Now, it is possible to rewrite Bayes’ Theorem for hypothesis tests as:

$$\text{P(Null is False given a decision to reject the null)} = \frac{\text{Power } \times P(\text{ Null is False })}{\text{Power } \times P(\text{ Null is False }) + \alpha \times P(\text{ Null is True })}$$

Cohen (1992) suggests that researchers try to obtain power of at least 0.8 for studies. Power is a function of both sample size and effect size. The smaller the effect, the larger the sample needed for adequate power. Let us assume that the researcher in question calculated a post-hoc power for his study of the recommended 0.8 and selected an α of 0.05. The hypothetical researcher can now calculate the posterior probability that the null is false given the decision to reject the null:

$$\text{P(Null is False given a decision to reject the null)} = \frac{0.8 \times 0.5}{(0.8 \times 0.5) + (0.05 \times 0.50)} = 0.941$$

At first glance, this would seem like a promising result for the researcher in question. However, this actually yields a probability that the null is true which is higher than traditional p-values would suggest. For example, this yields a posterior probability that the null is true of around 0.06. This is higher than the traditional α-level of 0.05. Further, the p-value is dependent on sample size. In order to get adequate power of 0.8, the researcher may have had to use a large sample that would have yielded a low p-value for an effect of a modest or small effect. For example, the researcher may have obtained a p-value of 0.01, which is 6 times smaller than the posterior probability that the null is true. Under these idealized conditions, the researcher can still have some confidence in the obtained effect, but the confidence should be more tempered than the initial p-value might indicate.

Where such an analysis becomes really useful is when conditions are not ideal. For example, it has long been recognized that psychological studies are frequently underpowered (Cohen, 1992; Maxwell, 2004). Suppose that the researcher in the study above did a post-hoc power analysis and found that their power was 0.5 instead of 0.8. The posterior probability that the null is false given the null hypothesis was rejected falls to approximately 0.91. As the obtained power falls, the posterior probability will also fall. As power is a joint function of effect size and sample size, this indicates that small effects and small samples reduce one’s confidence in a finding relative to larger samples and effects.

In addition, the lower the perceived personal probability that the null is false, the less confident a researcher should be in a finding. Imagine a scientist uses a two-tailed statistical test but makes a prediction based on previous experiments she has done and on previous published literature. Prior to the study, she assigns a probability that the null will be rejected in direction 1 (e.g., “the experimental group will have a higher mean than the control group”) of 0.7, a probability of a null result of 0.2, and a probability the null will be rejected in direction 2 (e.g., “the control group will have a mean higher than the mean obtained from the experimental group") of 0.1. She unexpectedly obtains a result in the direction opposite of her prediction, and the control group does indeed have a significantly higher mean. Post-hoc power analysis yields power of 0.8, and her criterion for significance was a p-value of less than .05. Since the result flies in the face of previous research, one should be more skeptical of it. Using 0.1 as P(A), the posterior probability that the null is false given the null was rejected would be:

$$\text{P(Null is False given a decision to reject the null)} = \frac{0.8 \times 0.1}{(0.8 \times 0.1) + (0.05 \times 0.90)} = 0.64$$

At this point, given the researcher's personal probability of her result prior to the study, there is over a 1/3 chance she has a Type I error. Now, the researcher does not have to throw this data out. Instead, the nice thing about Bayesian statistics is that they can be used in an iterative fashion. She can run a replication and use her posterior probability from the first study as P(A) in the second study. Suppose she replicates the result in study two and has power of 0.8 and uses α = 0.05 again. After replicating the result, the posterior probability the null is false given the null was rejected rises to 0.966. It should be evident that the subjective personal probability that the null is false has a major impact on the final posterior probability. This raises the fair question, how should one choose P(Null is False) to begin with? The next section addresses this question.

The Bayesian Range and Assuming You Are Wrong

Suppose one is interested in incorporating posterior probabilities into their statistical reporting. On the one hand, post-hoc power incorporates sample size and the obtained effect size. It is based on data. On the other hand, the probability that the null is false is a subjective personal probability. What is the best way to proceed? This paper will recommend two approaches.

One approach this paper shall call "The Bayesian Range for Intermediate Levels of Confidence in a Theory" or "The Bayesian Range," for short. Just as Cohen popularized the idea of reporting and interpreting effect sizes, one could always report posterior probabilities after rejecting a null hypothesis. For a novel hypothesis that is being tested, one approach would be to report posterior probabilities for several intermediate levels of confidence that the null is false. Suppose a researcher obtains a novel result, but it is woefully underpowered at 0.43 with alpha set to 0.05. Using calculations similar to the ones above, the author could report a range of posterior probabilities for three levels of intermediate confidence. Let it be supposed the researcher chooses 0.25, 0.5, and 0.75 as the intermediate levels of confidence and plugs each one into the formula. They could put their posterior probabilities into a table such as Table 1.

Table 1. *Bayesian Range for Hypothetical Study with Power of.43 and α-level of 0.05*

P(Null is False)	Posterior Probability Null is False Given Null Rejected
0.25	0.741
0.50	0.910
0.75	0.963

What such a range tells us is that at the lower bounds of the range of confidence, the probability the null is false given that the null was rejected could be as low as 0.74, which means there would be as high as a ¼ chance that the null is true. At the upper range of the intermediate levels of confidence, the probability that the null is false given that the null was rejected is as high as 0.96, but that still leaves almost a 1/25 chance the null is true. What such an analysis suggests is that type I errors should be more common than would be implied by p-values alone and that the more unexpected the result, the greater the initial risk of a type I error. Because Bayesian statistics are iterative, a replication of the original study could use one of the posterior probabilities from the table as the probability the null is false in a Bayesian calculation done for the results of the replication. The safest way to do this would be to choose the lowest probability from the range as the new prior probability that the null is false. This more conservative approach will slow down the rate at which the posterior probabilities blow up with replications.

This latter point is important, because if a study's result is replicated several times, it will not take long for the probability that the null is false, given that the null was rejected, to grow very quickly. That is why a second, and even more conservative, approach is to set the initial probability that the null hypothesis is false to an arbitrarily low number such as 0.01 and to iteratively replicate the study, making each posterior probability the null is false the new prior probability the null is false in the next study. A researcher may choose to not consider a set of studies ready to present until the posterior probability that the null is false surpasses some predetermined value. This could be particularly important for situations in which findings are novel, counterintuitive, or contradict previous studies. For the sake of argument, suppose this value is set to be 0.90. Suppose the hypothetical underpowered study mentioned in Table 1 is a test of a novel hypothesis. Because it is novel and underpowered, the researchers set the probability the null is false to 0.01. With a power of 0.43, the probability the null is false given the null was rejected would only be about 0.08. However, this probability becomes the new probability the null is false in the second study. Also, imagine that the researchers increased their sample size to bring power up to an adequate 0.8. If the same result is obtained in study two, the posterior probability that the null is false given the null was rejected jumps to 0.58. A third replication with adequate power would then bring the probability over 0.90. After two replications with the same results and adequate power, the researchers would be justified in arguing there is strong evidence the null is false.

Advantages and Limitations of the Approach

The approach outlined above has several advantages. First, it is easy to calculate, and it uses a variation of Bayes' Theorem that psychologists are most likely to have encountered. It does not require any special expertise, and professors teaching statistics can introduce it tomorrow without having to completely scrap NHST. Even if some researchers and statisticians would prefer NHST to be dropped eventually, the reality is that it is still common and still widely taught. The proposed procedure is a way to introduce Bayesian thinking without forcing an immediate overhaul of psychological education and practice.

Second, it raises an important question about the novelty of findings. When the Open Science Collaboration (2015) published its multi-lab attempt to replicate 100 studies from top-tier psychology journals, the overall replication rate was dismal. Social psychology studies fared particularly poorly, with only around 30% of social psychology studies replicating. However, the journals that were chosen were ones with high impact factors and readership and ones that put an emphasis on novel findings because of the competition for space. However, as scientists, we should actually expect that most hypotheses and ideas are wrong to some degree or the other. Novel findings should be the ones that are most likely to be false positives instead of being the ones that are most likely to be given maximum exposure and endorsement from top-tier journals. A Bayesian approach should result in initial skepticism of a novel finding, but through the iterative replication process, will allow scientists to present new findings with more confidence so long as they replicate. In our publish-or-perish world, this might mean the inconvenience of doing 3 or 4 versions of a study instead of 1 or 2 before publication. On the flip side, it might mean fewer failed replications later.

One of the limitations of the approach suggested in this paper is that it only supplements, not replaces, NHST. Further, there is a level of subjectivity involved. Judgements must be made about the initial probability that the null is false. As such, it would be a stretch to call the probabilities exact estimates of the likelihood that the null is false given the null was rejected. Instead, the approach offers more of a way for researchers to put boundaries around their uncertainty, to express it numerically, and to provide a counterweight to the siren song of p-values. That is why the safest way to use Bayesian analysis might be to take the most conservative approach, especially when a result is novel or unexpected. That is to say, give the probability that the null is false a low value and replicate findings until that value is acceptable.

Another drawback is that the method described in this paper does not have a great defense against the file drawer effect. Successful replications may quickly lead to large posterior probabilities, but a failed replication should have the effect of dramatically reducing the probability that the null is false. If only studies that reject the null are considered, then the Bayesian analysis provides no additional benefits. But that raises an additional question: what to do with a failed replication? If a scientist believes a failed replication is a type II error, it still ought to be reported, and the result ought to be reflected in a reduced probability that the null is false when future Bayesian analyses are done. How much that reduction should be is up for debate. However, because two successful replications can take an initial probability the null is false of 0.01 and ultimately yield a posterior probability it is false of over 0.9, there is no reason not to go conservative with reconsidering a result that has failed to replicate. One recommendation is to reduce the probability that the null is false by a factor of 10 for every failed replication. Thus, just as two successful replications would yield a posterior probability that the null is false of over 0.9, two failed replications would reduce it to below 0.01. With such a procedure, a set of experiments with as many or more failed replications than successful ones would yield a posterior probability the null is false of less than 0.5.

In conclusion, incorporating Bayesian thinking into everyday statistical practice need not be daunting or even that computationally intensive. A simple approach, such as the one presented here, could be readily adopted by most psychologists and students of psychology and could be done as a supplement to NHST. It is understood that an expert in Bayesian statistics might see the approach outlined in this paper and feel it is crude and rudimentary. However, the goal of the paper is accessibility. How can Bayes' Theorem help psychologists think a little more carefully about their findings? Further, how can Bayes' Theorem be introduced in a way that makes an immediate impact for people who have not typically used it or thought about it?

References

- Berkson, J. (1930). Bayes' theorem. *The Annals of Mathematical Statistics*, *1*(1), 42-56.
- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98-101.
- Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods, 9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 1–8.
- 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. https://doi.org/10.1037/met0000100
- Yalch, M. M. (2016). Applying Bayesian statistics to the study of psychological trauma: A suggestion for future research. *Psychological Trauma: Theory, Research, Practice, and Policy, 8*(2), 249–257. https://doi.org/10.1037/tra0000096