Why p-Values Make Sense, but Maybe Not in the Way You Think

Jonas Groschwitz

October 2024

Summary: p-Values are a valuable statistical measure. But they don't tell us directly whether a hypothesis is true or false. Rather, they can tell us how much to *shift* our *already existing (=prior) beliefs*. In this context, the arbitrary cutoff for "significant" p-values at e.g. 0.05 makes sense as a cutoff for "this experiment provides a meaningful shift to our beliefs", as opposed to a "cutoff for the truth".

1 Basics of Significance Testing

The goal of significance testing is to mathematically determine, from a set of data, how likely a certain claim is to be true. Or more simply put, to answer the question, "should I believe this claim?". Somewhat counterintuitively, but mathematically very usefully, significance testing does not look at the likelihood of the claim directly, but looks at the likelihood of a *null hypothesis*, usually denoted as H_0. A null hypothesis describes essentially the opposite of the claim, the absence of an effect.[1]

Take for example the classic lady tasting tea scenario described by Ronald Fisher. In this scenario, a woman claims to be able to taste whether the milk or the tea was added first to a cup of tea. The null hypothesis to this claim is that the woman is guessing at random.

To reject the null hypothesis, one can conduct an experiment, for example, giving the woman some cups of tea and writing down how often she is right, and how often she is wrong. Say, for example, the woman tastes 8 cups and is right 7 times. Let O then describe the set of experiment outcomes that are *at least as extreme* as the observed outcome. The idea of significance testing is then to look at the *p-value*, which is the probability that an outcome in O occurs, assuming that the null hypothesis H_0 is true. In mathematical notation the p-value is the conditional probability

$$P(O|H_0)$$

(read "P of O given H_0"). By convention, the null hypothesis is rejected if the p-value is at most some fixed value, usually 0.05 (i.e. 5%), that is, if

$$P(O|H_0) \leq 0.05$$

.

To summarize, if it is sufficiently unlikely that the null hypothesis could produce the observed outcome by random chance, then the experiment provides *significant* evidence against the null hypothesis and in favor of the claim.

However, this view on statistical significance always leaves me with two questions. First, we are primarily interested in the likelihood of the null hypothesis, not in the likelihood of the experiment outcomes. That is, we really care about $P(H_0|O)$, called the *posterior* or *posterior belief*, which is how much we should believe in the null hypothesis after having seen the experiment outcomes. We don't actually primarily care about the p-value. So how do the p-value and the posterior belief relate? And second, what's up with this seemingly arbitrary p-value cutoff at 0.05? A cutoff point of 5% seems like an odd choice if we really care about the truth; every 20-th bit of knowledge (on average) would just be wrong!

I found quite satisfying answers to these questions when looking at p-values through the lense of Bayes' theorem, and I'd like to share these answers with you below.

[1] It should be noted though that the null hypothesis is not always exactly the negation of the claim, but often it is close enough to think about it this way. To keep this essay simple, I'll restraing the argument to the cases where rejecting the null hypothesis is essentially the same as accepting the claim.

2 The Baysian View on Significance Testing

The central tool on our quest is Bayes' Theorem, a central theorem in probability theory that connects two conditional probabilities. Bayes' Theorem states that for two events A and B,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

This provides us with an answer to the first question: how to relate $P(H_0|O)$ and $P(O|H_0)$. Namely:

$$P(H_0|O) = \frac{P(O|H_0)P(H_0)}{P(O)}.$$

This is great news: we can now compute $P(H_0|O)$, which is what we really care about, from $P(O|H_0)$, which is what the statistical significance tests give us! The problem is only that there are two other pesky terms in the equation, $P(H_0)$ and $P(O)$.

$P(H_0)$ is called the *prior* or *prior belief* (we'll look at $P(O)$ in the next section). The prior belief is how much we believe that the null hypothesis is true *before* we conduct the experiment. The problem with the prior, in terms of objective reasoning, is that the prior belief is often subjective. For example, you can imagine that people will differ widely in how much they believe the lady that she can distinguish the tea, in the absence of any evidence. And one cannot compute the posterior without knowing the prior! This is the main reason why scientific analysis tends to report the p-value $P(O|H_0)$ rather than the posterior $P(H_0|O)$, because being derived from the subjective prior, the posterior is also subjective.

To summarize, it is not possible to compute the posterior fully objectively. The only thing we can really objectively get is $P(O|H_0)$. But the Baysian perspective is still very useful, as we will see in the next section.

3 Updating One's Beliefs with p-Values

In this section, we will first work out how exactly to update our beliefs using the p-value. That is, how to compute the posterior from the p-value and the prior. To shape our intuition, we'll then look at some concrete numbers of p-values and priors, and the posteriors they yield. Finally, we'll get back to our earlier two questions about significance testing, and see how this perspective resolves them quite nicely.

We left off above with the Baysian equation

$$P(H_0|O) = \frac{P(O|H_0)P(H_0)}{P(O)}.$$

To describe the posterior purely in terms of p-value $P(O|H_0)$ and prior $P(H_0)$, let us spell out $P(O)$, the likelihood of the evidence (without assuming the null-hypothesis). By the law of total probability, we have

$$P(O) = P(O|H_0)P(H_0) + P(O|\neg H_0)P(\neg H_0),$$

where $\neg H_0$ means "the null hypothesis does not hold". The first summand are our old acquaintances p-value and prior. In the second summand, we can write $P(\neg H_0)$ as $1 - P(H_0)$.

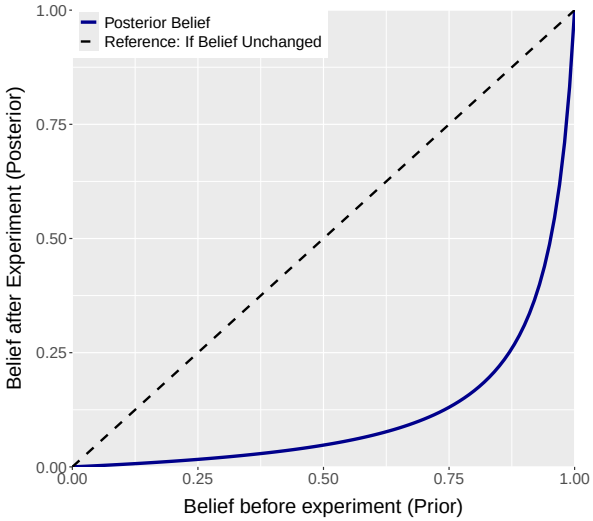
The remaining term, $P(O|\neg H_0)$, is the likelihood of the observation (or a more extreme result) under the condition that the null hypothesis does *not* hold. This number can often be assumed to be so close to one that we can just round it, $P(O|\neg H_0) \approx 1$. For example, if the lady can indeed taste the difference between putting the milk or the tea first into the cup, then the likelihood of her getting at least 7 out of 8 cups right is very high. Note that this assumption does not always hold, and what to do when it doesn't is a whole topic on its own. But here, we'll just focus on the simple cases where this approximation $P(O|\neg H_0) \approx 1$ is appropriate.

Under this assumption, we then get

$$P(O) = P(O|H_0)P(H_0) + 1 - P(H_0),$$

and thus our final formula for the posterior, expressed purely in terms of the p-value and the prior:

$$P(H_0|O) = \frac{P(O|H_0)P(H_0)}{P(O|H_0)P(H_0) + 1 - P(H_0)}. \tag{1}$$



Prior	0.001	0.05	0.2	0.5	0.8	0.95	0.999
Posterior	0.00005	0.003	0.012	0.048	0.167	0.487	0.980

Figure 1: Posterior belief, depending on the prior belief, for a p-value of 0.05 (assuming $P(O|\neg H_0) \approx 1$).

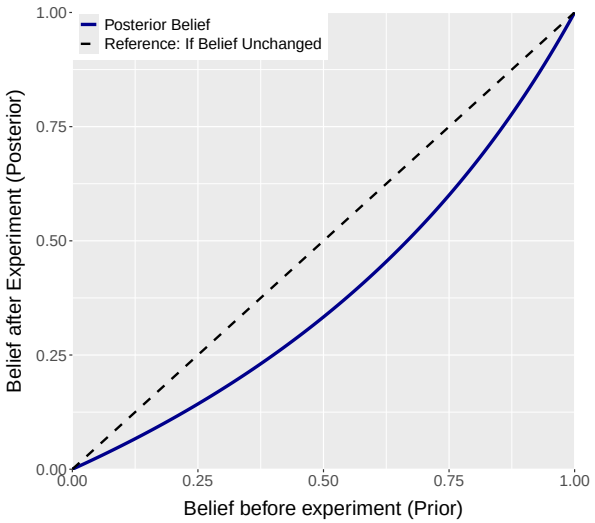
What a mouthful! To better understand what this means, let's plug in some numbers.

Figure 1 shows how different prior beliefs in the null hypothesis get updated to new posterior beliefs if the p-value is 0.05. We can see that the belief in the null hypothesis drops dramatically. For example, if the prior belief is 0.5 (50%), it drops below 5% after the experiment. If the prior belief is very small, the posterior is essentially the prior multiplied with the p-value (this can also be deduced from Equation (1): if the prior is very small, the denominator becomes essentially 1).

However, even an experiment with a small p-value of 0.05 – that is, under the null hypothesis, the likelihood of obtaining results as extreme as the ones observed is only 5% – does not necessarily mean that we should throw the null hypothesis out of the window. If before the experiment, we were quite sure that the null hypothesis is true – say, 95% sure – then we should still treat the null hypothesis as plausible after the experiment. And if we are extremely certain of the null hypothesis (99.9%), then afterwards we should still believe in it, but maybe have a bit more doubt (98% certainty).

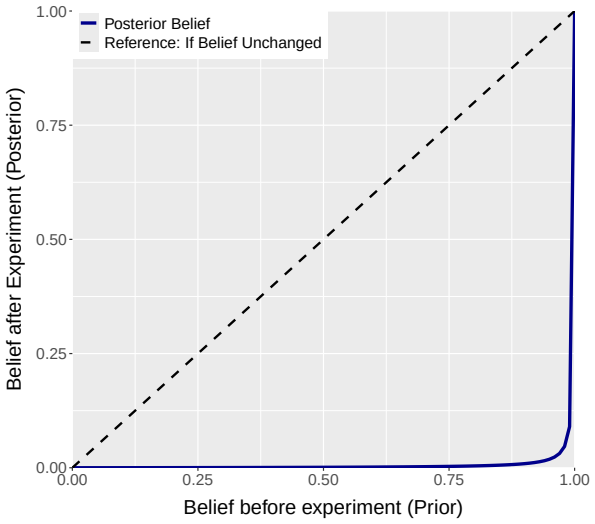
Figure 2 shows how to update our beliefs when the p-value is much larger, 0.5. Our belief in the null hypothesis should still decrease, but the effect is much smaller. By contrast, Figure 3 shows that an experiment with p-value 0.001 – under the null hypothesis, results as extreme as the observed ones would occur only once in a thousand experiment runs – is much more powerful in terms of shifting our beliefs, and can give us a lot of confidence in rejecting the null hypthesis.

An crucial observation here is that even though the prior belief is subjective, experiments with sufficiently small p-values can help normalize our beliefs. Have another look at Figure 3 (p-value 0.001) and see how, for people with a wide range of prior beliefs, the posterior belief has a near-complete consensus to reject the null-hypothesis. The exact degree of that rejection may vary, as you can see in the numbers in the table, but the disbelief in the null-hypothesis itself is a consensus that only people with an extreme prior will reject. And this consensus now is no longer based on subjective prior beliefs, but in objective observations. That is the power of science, and the power of proper application of p-values: they can form an objective consensus about the thruth. A single experiment with p-value 0.05 does not quite have such strong an effect, but the same principle is there, and each such experiment is at least a step towards a solid consensus.



Prior	0.001	0.05	0.2	0.5	0.8	0.95	0.999
Posterior	0.0005	0.026	0.11	0.33	0.67	0.9	0.998

Figure 2: Posterior belief, depending on the prior belief, for a p-value of 0.5 (assuming $P(O|\neg H_0) \approx 1$).



Prior	0.001	0.05	0.2	0.5	0.8	0.95	0.999
Posterior	0.000001	0.000053	0.00025	0.001	0.004	0.019	0.5

Figure 3: Posterior belief, depending on the prior belief, for a p-value of 0.001 (assuming $P(O|\neg H_0) \approx 1$).

4 Conclusion

Let us now come back to our original questions. First, if we care about the posterior $P(H_0|O)$, why do we compute the p-value $P(O|H_0)$ instead? And how do they relate? The answer is that only the p-value can be computed objectively, but our subjective beliefs can be updated with the methodology we just discussed. In other words, the p-value cannot tell us directly whether we should believe in the null-hypothesis (or disbelieve it, and correspondingly believe in the claim), but it can tell us how to shift our existing prior belief. This allows us to make our beliefs more and more based on evidence, bringing us as close to an objective consensus as we can get.

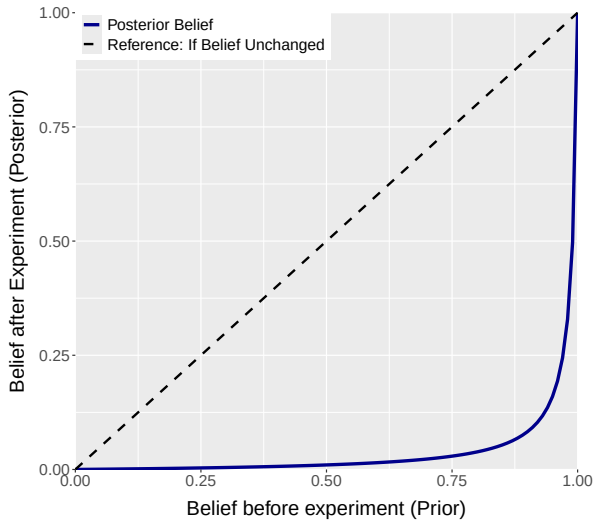
Second, why the p-value cutoff at 0.05? From a Baysian perspective, this arbitrary cutoff can be interpreted as stating that experiments of at least this p-value provide a shift in belief meaningful enough ("significant") that they are worth taking note of (and publishing a paper about them etc.). While the value 0.05 is still arbitrary, to me at least it actually makes intuitive sense as this sort of "update threshold" – much more so than a "truth threshold".

Taking on this point of view does not solve all issues with p-values and the current scientific landscape. For example, the multiple comparisons problem and publication bias are serious issues.

But nonetheless, I hope that I was able to give you a new appreciation for why p-values are useful, and maybe a bit more precise an intuition for how to interpret them.

Appendix: Prior and Posterior Values for p-Value 0.01

Since a p-value cutoff of 0.01 is also often used in science, here are the numbers and the plot for p-value 0.01.



Prior	0.001	0.05	0.2	0.5	0.8	0.95	0.999
Posterior	0.00001	0.00053	0.0025	0.0099	0.038	0.16	0.91

Figure 4: Posterior belief, depending on the prior belief, for a p-value of 0.001 (assuming $P(O|\neg H_0) \approx 1$).