Getting Started with Bayesian Statistics

Thomas J. Faulkenberry

October 28, 2020

Ten years have passed since Daryl Bem began circulating a preprint of his now infamous paper in which he claimed evidence for precognition – a phenomenon whereby future events can implicitly affect a person's present behaviors, even without a person's conscious awareness. This unbelievable result – which Bem operationalized as his subjects' ability to correctly predict the position of pictures on a computer screen at rates significantly greater than chance – was ultimately responsible for a host of methodological revolutions in our field. Since that time, considerable energy has gone into arguing for the merits of "open science", preregistration of studies, and a closer look at the statistical training of students and researchers, including calls for abandoning the use of *p*-values for hypothesis testing.

By now, one of the more familiar alternatives to traditional null hypothesis testing is *Bayesian hypothesis testing*. Despite increasing familiarity, Bayesian methods are still relatively under-utilized in psychology and are virtually absent from most courses in psychological statistics. I think this is quite unfortunate, as the core tenets of Bayesian inference are actually easier to understand than the traditional null-hypothesis-testing alchemy that we all grew up with. My goal in this column is to convince everyone that this statement is indeed true.

So where to start? Let's start with the familiar. Most readers will have some experience with hypothesis testing, and particularly the notion of a *p*-value, which we are taught to use as a magical dowsing rod for ascertaining whether something is "significant." Let's explore this a bit deeper. When we want to back up some quantitative statement – say, that the difference between two group means is significant – we formally proceed by defining two hypotheses: a null hypothesis H_0 which states that there is no difference between the means, and an alternative hypothesis H_1 which states that there is some difference. Then we collect some data, because we want to test how well these hypotheses hold up as models of our observed data. One way to do this is to *assume the null is true*, then calculate the probability of observing our data under H_0. This probability is the *p*-value, and traditional practice dictates that we determine whether this probability is small (i.e., less than 5%). If so, we say that our data is rare under H_0, and so we reject H_0 in favor of the alternative H_1. In short, we look for a difference in the group means by assuming that there is no difference, then showing that our (actually observed) data is implausible under such a hypothesis, rendering the null hypothesis itself implausible.

Two issues arise immediately. First, our evidence for H_1 (the model we actually care about in this situation) is indirect. We've shown that the null is not a good fit for our observed data – that's what the *p*-value shows – but nowhere have we actually assessed how well the *alternative* fits our data. Second, suppose we fail to reject H_0 (presumably because $p > 0.05$). As anyone who has taught undergraduate statistics knows by heart (because he or she has explained this countless times), simply failing to reject the null does not permit one to conclude *support* for the null. In these ways, the traditional null hypothesis testing procedure is lacking.

Bayesian hypothesis testing, on the other hand, takes care of these problems easily. At its simplest, Bayesian hypothesis testing replaces the *p*-value with the *Bayes factor*. To understand what a Bayes factor is, let's compare it to the *p*-value. Whereas the *p*-value tells us the likelihood of the data under the null hypothesis alone, the Bayes factor tells us the *relative* likelihood of the data under both H_0 and H_1. As the Bayes factor simultaneously compares two hypotheses, we must have some way to specify which hypothesis is being supported. We do this by specifying the "direction" of the Bayes factor with a subscript. For example, BF_{10} represents the Bayes factor for the alternative H_1 over the null H_0. On the other hand, BF_{01} represents the Bayes factor for the null H_0 over the alternative H_1.

Modern software packages such as JASP (which is freely downloadable at https://www.jasp-stats.org) make it simple for anyone to compute Bayes factors. Most of the common statistical tests you're already familiar with have Bayesian versions in JASP. The key to getting started is knowing how to interpret the Bayes factor. So, let's suppose we did a Bayesian independent samples *t*-test, and our data produced a Bayes factor of $BF_{10} = 20$. This means that the observed data are 20 times more likely under the alternative hypothesis H_1 than the null hypothesis H_0. Notice that instead of simply using a small *p*-value to reject the null as ill-suited to explain our observed data, we are reporting a relative degree of fit for *both* hypotheses. This Bayes factor tells us directly that the alternative hypothesis fits our data 20 times better than the null hypothesis. I think such statements are much easier to interpret, and in my experience, students have an easy time getting this too. Also, and perhaps most importantly, it is entirely possible that our data could instead be evidential for the null – for example, if we got $BF_{01} = 20$, that would mean that the data were 20 times more likely under the null than the alternative. Traditional null hypothesis testing simply cannot do this.

Once you started computing and interpreting Bayes factors for your own data, one natural question is "How big does the Bayes factor need to be?" After all, we are taught guidelines for how small a *p*-value must be in order to separate signal from noise. What sizes should we expect for Bayes factors? Remember, the Bayes factor is a ratio, so the "smallest" value we should get is $BF_{10} = 1$, which would mean that the data were equally likely under both the alternative and the null. In this case, we don't have any compelling evidence in favor of either model. A common recommendation is to consider a Bayes factor greater than 3 as representing positive evidence in favor of one hypothesis over the other. This is because having at least 3-to-1 odds in favor of a specific hypothesis equates to a posterior probability of at least 75% in favor of that hypothesis. My recommendation is to not worry about specific thresholds beyond this – simply report the value of the Bayes factor and tell the reader exactly what it means.

At this point, I invite you to start doing some Bayesian analyses of your own! Here, I've only scratched the surface of Bayesian hypothesis testing, but I've hopefully piqued your interest. There are plenty of resources to help you learn more, including this special issue of Psychonomic Bulletin & Review and my own recently published tutorial paper. Further, I usually offer workshops on Bayesian statistics at SWPA every year, so I will look forward to meeting you in one of those workshops. Feel free to send your Bayesian questions my way any time, or even invite me to give a workshop at your own university! You can email me at faulkenberry@tarleton.edu.