Subjective Bayesian Hypothesis Testing

David (Daiwei) Zhang

March 7, 2017

Hypothesis testing is an important procedure in scientific inquiry. The classical way of testing hypothesis is the frequentist method. However, the emergence of Bayesian inference provides an alternative to this traditional methodology and has been proved to have multiple advantages. Within the Bayesian school of thoughts, there are two subdivisions: subjective Bayesianism and objective Bayesianism, with the difference lying on the type of prior distributions used in the analysis. In this literature review, I will show the basic framework of Bayesian hypothesis testing, the difference between the frequentist and the Bayesian methods, and finally the advantages of subjective Bayesianism.

The frequentists' classical way of testing hypothesis goes back to Fisher. His idea of forming and rejecting the null hypothesis with the data was highly influenced by the philosopher of his time Karl Popper. Popper advocates a science of falsification rather than verification. He argues that a scientific hypothesis can never verified by any data, since we can only collect finitely many pieces of evidence and will never know if there are some evidence yet to be discovered that will contradict the hypothesis. However, although it takes infinitely many data points to verify a hypothesis, it only needs one data point as a counterexample to falsify a hypothesis. Thus Popper believes that science is a process of rejecting false hypotheses and the theory we believe to be true is technically the theory that has yet to be proven to be false, or rather, the theory that has the least data against it (Thornton (2016)). We can clearly see the shadow of Popper in Fisher's way of testing hypothesis. Rather than verifying the hypothesis that we think might be true, we form a strawman hypothesis and falsify it with data. This is the essence of frequentist hypothesis testing (Press (2003)).

In contrary, Bayesian hypothesis testing provides a drastically different philosophy and methodology. In Bayesian inference, probability is regarded not as the limiting relative frequency of a repeatable event but rather as a measure of a person's degree of belief. This makes it legitemate to assign probability to hypotheses, a taboo for the frequentists. Rather than fixing the parameter and calculate the probability for it to produce the observed data, Bayesian statisticians fix the observed data and calculate the probability for it to be produced by the hypothesis. This framework of thinking makes hypothesis testing less frequently in need in Bayesian inference, or at least it makes hypothesis testing as easy as drawing any other conclusions from the posterior distribution. All we need is to look at the posterior distribution and find the probability for the parameter to be in the range that we are interested in. Notice that point hypothesis testing (e.g. $\theta = 0$), which is frequently used by frequentists, has no meaning in Bayesian inference, since a point has zero mass in the posterior distribution. But this shows another advantage of Bayesian hypothesis testing: rather than answering whether θ is zero, it tells us how θ is distributed, which gives us much more information (Gelman et al. (2004) Ch. 8).

In addition to the infrequent need to test hypotheses and the ability to assign probabilities to hypotheses, Bayesian hypothesis testing is advantageous in the weakness area of the frequentist method. First of all, the significance level in the frequentist hypothesis testing is arbitrary. The number 0.05 is not a magic number and can be well replaced by 0.10, 0.01234, or $1/\pi$. On the other hand, for the Bayesian, all that is needed is whether the posterior distribution favors one hypothesis or the other, that is, whether the odds is greater or less than one:

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)} \frac{P(x|H_0)}{P(x|H_1)}$$

This is naturally extended to testing more than two hypotheses. In the case when the parameter is continuous and the alternative hypothesis is composite, we can simply take the average for the alternative hypothesis in the posterior distribution:

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)} \frac{P(x|H_0,\theta)}{\int P(x|H_1,\theta)g(\theta)d\theta},$$

where g is θ's prior distribution under H_1.

Moreover, frequentists often test point hypotheses of the form $\theta = \theta_0$, but in reality, θ is almost never equal to zero but rather often in some ϵ neighborhood of zero. This means that if the testing procedure is consistent, we can reject any hypothesis we want simply by making the size of the sample large enough. This provides a loophole to show the significance that does not exist. Finally, frequentist hypothesis testing involves data that is never observed. When we set a significance level for our p-value, we are dividing the sample space into two regions. However, either of these two regions contains data points that have never been observed. This violates the likelihood principle, which states that all the information in a sample that is relevant to the parameters of the model is contained in the likelihood function. In addition, studies (Berger & Selke (1987) and Casella & Berger (1987)) have shown that data against the null hypothesis is generally not as strong as what is reflected in the p-value. Thus Bayesian hypothesis is often more conservative than its frequentist counterpart (Press (2003)).

Although Bayesian inference has several advantages over frequentist inference, it also has its own challenges. One component in Bayesian hypothesis testing that statisticians often disagree on is the choice of the prior distribution. On this issue, Bayesians are divided into two categories, the so-called objective Bayesians and subjective Bayesians. Subjective Bayesians suggest quantifying experts' prior experience on the issue into the prior distribution. In contrast, objective Bayesians do not want subjective elements in statistical analysis and propose to use "non-informative" prior distributions. There are different ways to make a distribution non-informative, and these distributions are often relatively flat. A commonly used method for this task is Jeffrey's invariance principle. Jeffrey argues that a non-informative prior should be invariant to transformation of variables. If $p(\theta)$ is the non-informative prior for θ, and $\phi = h(\theta)$ is the transformed variable, then we should have

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| .$$

Jeffrey's principle implies that the non-informative prior should be

$$p(\theta) \propto J(\theta)^{1/2},$$

where $J(\theta)$ is Fisher's information for θ. Although Jeffrey's principle provides reasonable non-informative priors for single-parameter models, it becomes more controversial when multiple parameters need to be evaluated (Gelman et al. (2004) Ch. 2).

The objective Bayesianism's use of non-informative priors can be useful when quantifying experts' prior experience is not worth the effort. However, non-informative priors have at least three major drawbacks. First, the search for non-informative priors can be misguided. If we truly want the likelihood to dominate the posterior, then the choice of the prior does not matter much. Looking for "the" prior distribution will tempt us to use it inappropriately. Second, there is often no clear choice for non-informative prior distributions. A density that is flat under on parameter may no longer be so under another. Finally, non-informative priors can be improper, and this can cause problems in the analysis of the posterior (Gelman et al. (2004) Ch. 2).

Subjective Bayesianism provides a reasonable alternative to the difficulties in objective Bayesianism. However, there are at least two reasons that scares people away from this solution. First of all, quantifying experts' experience is not an easy task and consumes money and time. But more importantly, many scientists believe that subjective elements have no place in scientific inquiry and therefore should be avoided in statistical analyses. Goldstein (2006) responded to both of these issues. He first points out that the pain-to-gain ratio is indeed too high for converting experience into prior distribution in some cases. However, when the data is too limited or when the system is too complex, such as software testing, quantifying experts' experience is an investment that will save the overall financial and time cost.

As for the presence of subjectivity in scientific inquiry, he argues that there are always subjective elements in science, just that they are better-hidden in the so-called "objective" methods. Even objective Bayesian inference measures the degree of belief of the observer rather than the physical properties of the observed. Subjective Bayesian inference provides a way to quantify subjectivity so that scientists disagreeing on a conclusion can trace their disagreement in a clear manner. Moreover, when the choice of prior can significantly change the conclusion, it means that the issue is currently controversial and therefore impossible to find a universally acceptable answer. Finally, labeling a statistical procedure as "objective" can be misleading to non-statisticians and tempts them to use the result of the analysis beyond what it promises. Therefore, subjectivity is unavoidable in any scientific analysis and subjective Bayesianism provides a reasonable way to dissect and quantify it.

In conclusion, Bayesian hypothesis testing differs from frequentist hypothesis testing in both philosophy and practice. Furthermore, subjective Bayesianism and objective Bayesianism disagrees on the role of statistical analysis in scientific inquiry, which leads to their difference in the treatment of the prior distribution. It is interesting to see these fundamental disagreements in issues even as basic as hypothesis testing in statistics. Thus it requires further studies and investigation to find a unifying solution or a better understanding of the differences in various schools of thoughts in statistics.

References

Berger, J. O. and Selke, T.. "Testing A Point Null Hypothesis: The Irreconcilability of p-Values and Evidence", Jr. Am. Statist. Assoc., **82**(397), 112-122.

Casella, G. and Berger, R. L.. "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem", Jr. Am. Statist. Assoc., **82**(397), 106-111.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B.. *Bayesian Data Analysis*, 2nd edition, Chapman & Hall.

Goldstein, M.. "Subjective Bayesian Analysis: Principles and Practice", International Society for Bayesian Analysis, **1**(3), 403-420.

Press, S. J.. *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, John Wiley and Sons.

Thornton, Stephen, "Karl Popper", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.).