DISHONEST STOPPING, OR: BAYESIANS AND BAD EXPERIMENTS

WENBO GAO

ABSTRACT. A discussion of Bayesian inference in the context of deliberately biased experiments.

David Speyer [2] describes a seemingly paradoxical way of fooling Bayesian inference. Suppose that a pharmaceutical company creates a new drug, which either cures the patient with probability θ, or fails with probability $1-\theta$. Two experiments are carried out to test the efficacy of the drug:

- **E_1:** Scientist 1 carries out a random trial by giving the drug to 10^4 patients. She reports the number of cured patients.
- **E_2:** Scientist 2 is paid by the company to produce evidence that the drug works with probability greater than $1/2$. She gives the drug to patients in succession until n patients have taken the drug and at least $n/2+\sqrt{n}$ were cured.

Suppose that both scientists announce their experimental procedures, and then carry out their trials. Scientist 1 reports that 5100 of the patients were cured. Scientist 2 reports that $n = 10^4$ and that 5100 of the patients were cured. Both experiments suggest that $\theta > \frac{1}{2}$. Now, we ask: does E_1 provide *stronger evidence* for $\theta > \frac{1}{2}$ than E_2?

Let us take care to distinguish different questions:

(1) Does our posterior probability satisfy $p(\theta > 1/2|D, E_1) > p(\theta > 1/2|D, E_2)$? That is, given that the same data D was observed in both experiments, is E_1 more compelling than E_2?
(2) Will E_1 allow us to make a better inference for θ than E_2?

Perhaps surprisingly, (1) is false. In fact, for any θ',

$$p(\theta'|D, E_1) = p(\theta'|D, E_2) \tag{1}$$

And indeed, this seems intuitive as well. The data D was produced from a physical process, and the *intentions* of the scientist should have no impact on our interpretation of the data. However, our intuition paradoxically suggests that (2) might be true, since E_2 is clearly biased towards certain outcomes. And indeed, we generally have a strong preference for unbiased experiments in real life.

This feature (eq. (1)) of Bayesian inference is often used as an argument for favoring Bayesian methods over frequentist ones. Eliezer Yudkowsky [3] considers a similar setup, where one experimenter has a group of 100 subjects, and the other experimenter continues until at least 70% of the subjects are cured. It then happens that both experimenters test 100 subjects, and find 70 to be cured. Yudkowsky notes that a frequentist, doing hypothesis tests, might find one experiment to be statistically significant and the other insignificant. He dismisses this reasoning as absurd, asking rhetorically: "The evidential impact of a fixed experimental method, producing the same data, depends on the researcher's private thoughts?"

Even earlier, MacKay ([1], §37.2) makes a similar argument. In his example, 'Dr. Bloggs' tosses a coin 12 times and observes 9 heads. His statistician friend performs a hypothesis test and finds no evidence of bias in favor of heads at a significance level of 0.05. Dr. Bloggs then responds, "No, my procedure was to toss the coin until 3 tails were observed!" Performing another test, the friend finds that $p = 0.03$ and there *is* significant evidence of bias.

First, let's see why eq. (1) is true. The key point is that, given data $D = \{n, c\}$ (for trials, number cured) which is a possible outcome for both E_1 and E_2, and for any θ, θ', we have

$$\frac{p(D|\theta, E_1)}{p(D|\theta', E_1)} = \frac{p(D|\theta, E_2)}{p(D|\theta', E_2)}$$

Date: Feb. 14, 2017.

This can be shown by verifying that $p(D|\theta, E_1)$ and $p(D|\theta, E_2)$ have the same form. Clearly $p(D|\theta, E_1) = \binom{n}{c}\theta^c(1-\theta)^{n-c}$. For E_2, let $\mathcal{L}$ denote the set of outcomes $(Y_1, \ldots, Y_n)$ such that c of the Y_i are 1, and for all $k < n$, we have $\sum_{i=1}^{k} Y_i < k/2 + \sqrt{k}$. Every such path in $\mathcal{L}$ occurs with probability $\theta^c(1-\theta)^{n-c}$, so we have $p(D|\theta, E_2) = |\mathcal{L}|\theta^c(1-\theta)^{n-c}$. Thus, we see that the ratios are equal.

By applying Bayes Theorem,

$$\begin{aligned} p(\theta'|D, E_1) &= \frac{p(D|\theta', E_1)p(\theta')}{p(D|E_1)} \\ &= \frac{p(D, \theta'|E_1)p(\theta')}{\int p(D|\theta, E_1)p(\theta)d\theta} \\ &= \frac{p(\theta')}{\int \frac{p(D|\theta, E_1)}{p(D|\theta', E_1)}p(\theta)d\theta} \\ &= \frac{p(D|\theta', E_2)p(\theta')}{\int p(D|\theta, E_2)p(\theta)d\theta} \\ &= p(\theta'|D, E_2) \end{aligned}$$

This is a particular case of an idea known as the *likelihood principle*.

How then can we reconcile this understanding of (1) with our intuition about (2)? Though we make the same inference from E_2 as E_1 given the same outcome D, the distribution of possible D differs between the experiments. In particular, the probability of making a poor inference *based on a single experiment* is much higher with E_2 than E_1. Consider the case where E_1 involves a large number of trials. Hoeffding's inequality implies that the observed number of successes concentrates narrowly around θn, and the fact that n is large strengthens our belief. In contrast, E_2 has a high probability of stopping at a small value of n, which does not allow us to make any strong assertions (observing 7 successes out of 8 is not particularly convincing, and can occur frequently even if $\theta < 1/2$).

Moreover, if the true value is $\theta < 1/2$, then with positive probability, the experiment E_2 will never terminate. Observing such an outcome in E_2 suggests that $\theta < 1/2$, but we would obtain stronger evidence from a small number of successes in E_1.

That said, simulations done by Reginald Reagan [2] indicate that Bayesian inference is somewhat robust against this kind of deception. Even when the Bayesian is tricked after observing one experiment into believing that $\theta > 1/2$, the posterior distribution shows a high level of uncertainty about the true value of θ.

References

[1] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, United Kingdom, 2003.

[2] David Speyer. The dishonest stopping paradox. `https://sbseminar.wordpress.com/2015/05/10/the-dishonest-stopping-paradox/`, 2015.

[3] Eliezer Yudkowsky. Beautiful probability. `http://lesswrong.com/lw/mt/beautiful_probability/`, 2008.