Comments on “The Jeffreys-Lindley Paradox ...” by Bob Cousins

Harrison B. Prosper

Florida State University, Tallahassee, FL 32306 USA

(Dated: October 28, 2013)

I. INTRODUCTION

The discovery in 2012 of a new boson by the ATLAS [1] and CMS [2] collaborations, rapidly accepted as "a" Higgs boson consistent with that predicted by the Standard Model (SM), brings to a close decades of research guided in large measure by this thoroughly well-tested physical theory. However, going forward, no such well-tested theory exists to guide us, though there are any number of speculative alternatives to the SM. Consequently, the new era will be one in which "extraordinary claims will require extraordinary evidence" if we are not to be led astray. In a recent paper, Bob Cousins [3] provides an accessible review of the Jeffreys-Lindley paradox (JLP) that invites a re-examination of hypothesis testing and model selection. I believe that a re-examination is definitely warranted given the dogmatic manner in which high energy physicists make use of notions such as the "5σ" discovery threshold. Bob Cousins argues persuasively that we ought to do better than this. I agree.

What is less clear are the conclusions to be drawn from the Jeffreys-Lindley paradox and its relevance to what we high energy physicists do. One of my problems with the JLP is that it is *not* a model of what happens in practice. As we acquire more and more data, what we do with them changes. As the exciting saga of the Higgs boson discovery illustrates, we typically do not continue to make *exactly* the same hypothesis test. Almost as soon as we were convinced that the observation was real, we discarded the artificial null hypothesis – the SM without a Higgs boson, strictly speaking a mathematically incoherent theory – and switched to the characterization of the new particle assuming that the Higgs boson hypothesis, the alternative, is true. And, almost immediately, that alternative was "downgraded" to the null hypothesis, albeit this time a sensible one, namely, the SM with a Higgs boson, which was dutifully used in tests of a variety of alternative hypotheses. Thought experiments can be useful, and this is true of the thought experiment that underpins the JLP. But one should proceed cautiously because thought experiments are typically unrealistic and their relevance to what *is* realistic may not be absolutely clear.

The following are some remarks prompted, in part, by Bob Cousins' excellent review.

II. COMMENTS

One would hope that all sensible decision procedures would yield the same decisions given sufficient data. The Jeffreys-Lindley paradox [3] is the mathematical fact that Bayesian and p-value based decision procedures can lead to contradictory decisions even when one has access to arbitrarily large data sets.

I have never found the Jeffreys-Lindley "paradox" particularly puzzling. There are times when I am puzzled by my lack of puzzlement given the voluminous literature on the subject nicely reviewed in Bob Cousins' paper. Perhaps, I have simply failed to understand the JLP. I leave you to judge. I do not see what is so puzzling about arriving at contradictory decisions (accept or reject an hypothesis) when the decisions are based on different decision criteria. As I see it, the JLP is just a forceful demonstration of a truism: decisions, as currently formalized, are not based on data alone, but also on how these data are used and on the assumptions that are made. To be sure, it would be nice if we could all agree on a set of decision procedures that would more often than not yield the same decisions, especially when we have mountains of data. But, we have all encountered situations in which one test statistic will yield one decision – say that a fit is good, while another will yield its negation – that the same fit is bad. I do not find this at all surprising: unlike physics, absent mathematical error there is no impartial arbiter in statistical inference of what constitutes the "right" answer. Different people will find different principles more or less intellectually compelling.

In physics, by contrast, whatever our intellectual prejudices, the real world has a habit of ignoring them and setting us straight. So accustomed are we of getting an answer from Nature on which we all eventually agree that we make the categorical error of supposing the same is true of statistical inference. In fact, when it comes to statistical inference there are many mathematically right answers. Which of these we choose is ultimately a matter of intellectual taste, convention, tradition, or brute clout. The published results on the Higgs boson would have been different had we made a myriad of different decisions. This is merely to say that every inferential statement we make is necessarily *conditional* and so there is no single "right" answer.

That being said, I am willing to concede that it is disturbing that mathematically sound decision procedures can yield opposite conclusions even when we have arbitrarily large data sets. However, my inclination is to join the "trivial" resolution of the "paradox" camp: each decision rule does what it is designed to do. Equation (1) in Ref. [3] defines $\sigma_{\text{tot}} = \sigma/\sqrt{n}$, which for fixed σ goes to zero as the sample size $n \to \infty$. Therefore, *any* fixed value of $z = (\hat{\theta} - \theta_0)/\sigma_{\text{tot}}$, however crazily large z may be so long as $z < \infty$, implies that $\hat{\theta}$ converges in probability to θ_0. Admittedly, this convergence occurs in a decidedly improbable manner: the error in every estimate $\epsilon \equiv \hat{\theta} - \theta_0$ is always of the same size, namely $\epsilon = z\sigma/\sqrt{n}$. This is akin to tossing an unbiased coin forever and always getting heads. This can happen, but it would be a tad surprising.

A more realistic way [4] to conceptualize the JLP is to imagine we have an infinite ensemble of possibly disparate experiments with differing values of σ_{tot}, but which all report the same value of z. We then order the experiments in decreasing values of σ_{tot} thereby arriving at the same mathematical consequence as the single experiment collecting ever increasing amounts of data. The point is that however one chooses to conceptualize the JLP, the limit of the sequence of experiments is characterized by $\epsilon \to 0$.

It seems to me perfectly reasonable that if $\epsilon \to 0$, where H_0 corresponds to the hypothesis $\epsilon = 0$, there ought to exist at least one decision procedure that concludes the following: $\Pr(H_0|\hat{\theta}) \to 1$ as $n \to \infty$. This is the import of Eq. (3) in Ref. [3]. I fail to see what is puzzling about this. However, there is another far more important point to be made, which is exhibited starkly in Eq. (9) of Ref. [3]: the Bayes factor $\text{BF} = p(\hat{\theta}|H_0)/p(\hat{\theta}|H_1)$ is directly proportional to the measure τ of the support of the prior $g(\theta)$ that is placed on the alternative (compound) hypothesis H_1. Consequently, as $\tau \to \infty$, that is, as one becomes ever more vague about the alternative hypothesis, the Bayes factor is driven inexorably to infinity, regardless of the value of z, a behavior that many consider problematic.

But, as I see it, this merely shows that in Bayesian reasoning alternative hypotheses matter. If one cannot, or will not, characterize them accurately then one should *not* use Bayesian reasoning. In the search for the Higgs boson, a perfectly well-defined (local) Bayes factor $\text{BF}(m_H)$ could have been calculated at each hypothesized Higgs boson mass[1], m_H, because the then alternative hypothesis, namely the SM with a Higgs boson, predicts the Higgs boson signal θ given the hypothesized mass, which prediction could have been modeled by a prior $g(\theta)$ with support determined by the assigned theoretical uncertainty in the signal

[1] In fact, Bayes factors were calculated by my former student Joe Bochenek for the $H \to ZZ \to 4\ell$ channel. But turning a ship as ponderous as the CMS Higgs group proved to be too difficult a task and p-values prevailed.

prediction at each mass. We simply chose not to.

On the other hand, if the rule is that every experiment reporting the same value of z, say $z = 50$, is to reject H_0 regardless of the size of its data set then necessarily that is what will happen, even for those experiments with arbitrarily large data sets for which ϵ is arbitrarily close to zero. The p-value decision rule is an infinitely sharp scalpel that continues to distinguish the hypothesis $\theta = \theta_0$ and its closest alternatives $\theta = \theta_0 \pm \epsilon$ even when they are infinitely close to each other, while the Bayesian rule decides that these three hypotheses become indistinguishable in this limit. The rules do what they are supposed to do. What is relevant is which decision is of greater practical import, a question, I believe, that cannot be answered in the abstract but only in a context-dependent manner. Is it of practical importance to be able to distinguish infinitely close hypotheses or are we happy to consider such hypotheses as the same for all practical purposes?

III. CONCLUDING REMARKS

Bob Cousins makes the important point that the Bayes factor BF depends on the measure τ of the support of the prior $g(\theta)$. True. But what are we to conclude from this? Simply, that decision-making is a less robust activity than that of establishing the details of a model already known to be viable. However much we may wish to make decisions unsullied by subjective and arbitrary elements, we should entertain the possibility that in searching for rules to make such decisions we may be chasing after shadows.

Much progress in our intellectual development has come about by discarding deeply held ideas. The idea of science as an "objective" intellectual pursuit is one such idea. But, lest I be grossly misunderstood, let me quickly qualify this statement. Science is the best way we know of arriving at durable, consistent, knowledge that allows us to manipulate the world as we see fit. That part of science, its results, is by any sensible definition of the word objective. But, how we get to these results is anything but; the scientific method is a complex interplay of objective, subjective, and arbitrary elements. We can no more avoid this than we can avoid enjoying something we enjoy.

We seem unable to let go of the idea of "letting the data speak for themselves" in spite of overwhelming evidence that data alone are not endowed with such magical power. Data must be augmented with rules in order to make them useful, rules that *we* invent. The review by Bob Cousins vividly demonstrates that we are not even close to a consensus about which rules are preferred, if any, when it comes to testing hypotheses. However, it seems to me that all attempts to divorce decision rules from the decision maker are bound to be found wanting, which is why I tend to favor the Bayesian approaches in spite of their imperfections. The debate re-opened by Bob Cousins is very useful. But, in the final analysis, when it comes to decisions there is no substitute for the exercise of skeptical judgement and adhering to the maxim “extraordinary claims require extraordinary evidence”.

[1] G. Aad et al. [ATLAS Collaboration], “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC,” Phys. Lett. B **716**, 1-29 (2012).

[2] S. Chatrchyan et al. [CMS Collaboration], “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC,” Phys. Lett. B **716**, 30-61 (2012).

[3] R.D. Cousins, “The Jeffreys-Lindley Paradox and Discovery Criteria in High Energy Physics,” arXiv:1310.3791v1 [stat.ME].

[4] R.D. Cousins, private communication.