The Statistical Philosophy of High Energy Physics: Pragmatism

Kent W. Staley

Saint Louis University

June 19, 2014

Contents

Abstract

The recent discovery of a Higgs boson prompted increased attention of statisticians and philosophers of science to the statistical methodology of High Energy Physics (HEP). Amidst long-standing debates within the field, HEP has adopted a mixed statistical methodology drawing upon both frequentist and Bayesian methods, but with standard frequentist techniques such as significance testing and confidence interval estimation playing a primary role. Physicists within HEP typically deny that their methodological decisions are guided by philosophical convictions, but are instead based on "pragmatic" considerations, thus distancing themselves from what they perhaps perceive as an ongoing pitched ideological battle between frequentists and Bayesians. Here I argue that there is a philosophical orientation to HEP that is neither exclusively frequentist nor Bayesian, but that lies squarely in the tradition of *philosophical pragmatism*. I further argue that understanding the statistical methodology of HEP through the perspective of pragmatism clarifies the role of and rationale for significance testing in the search for new phenomena such as the Higgs boson.

1 Introduction

On July 4, 2012 the CMS and ATLAS collaborations held a joint press conference at CERN to announce their latest findings from the search for the Higgs boson. ATLAS spokesperson Fabiola Gianotti declared that they had observed "clear signs of a new particle, at the level of 5 sigma, in the mass region around 126 GeV" (ATLAS 2012). The CMS statement reported the observation of an "excess of events at a mass of approximately 125 GeV with a statistical significance of five standard deviations above background expectations .... We interpret this to be due to the production of a previously unobserved particle with a mass of around 125 GeV" (CMS 2012). CMS and ATLAS considered the evidence insufficient to declare that the new particle was the Higgs boson itself, but only stated that the evidence was consistent with the expectations from decays of a Higgs boson.

Press coverage emphasized the appeal in these declarations to a standard of discovery: In order to announce the discovery of a new particle, the physicists needed to show that they had found an excess of candidate events beyond the expectations from background alone that would constitute a departure of at least five standard deviations ("5σ"). The associated probability statement was reported variously. In the New York Times, it was stated that "Both groups said that the likelihood that their signal was a result of a chance fluctuation was less than one chance in 3.5 million, 'five sigma,' which is the gold standard in physics for a discovery" (Overbye 2012). Reuters noted that "Five sigma, a measure of

probability reflecting a less than one in a million chance of a fluke in the data, is a widely accepted standard for scientists to agree the particle exists" (Wickham & Evans 2012).

Meanwhile, on a discussion forum on the website of the International Society for Bayesian Analysis, statisticians noted not only the appeal to the 5σ standard, but more generally the reliance of both groups on the significance testing methodology that yielded the relevant calculation: a p-value that was then converted into a distance from the background expectation, expressed in terms of a number of standard deviations. Tony O'Hagan, prompted by "[a] question from Dennis Lindley" posted a series of queries about the Higgs search results, including the following two:

1. Why such an extreme evidence requirement? We know from a Bayesian perspective that this only makes sense if (a) the existence of the Higgs boson (or some other particle sharing some of its properties) has extremely small prior probability and/or (b) the consequences of erroneously announcing its discovery are dire in the extreme. Neither seems to be the case, so why 5-sigma?

2. Rather than ad hoc justification of a p-value, it is of course better to do a proper Bayesian analysis. Are the particle physics community completely wedded to frequentist analysis? If so, has anyone tried to explain what bad science that is? (O'Hagan 2012a)[1]

These questions put into play two distinct issues regarding the statistical methodology of HEP. One is the use of the methodology of significance testing. The other is the apparent reliance on the 5σ standard for discovery claims. In this paper, I will focus on the question of the warrant for HEP's reliance on significance testing, though this will lead naturally to consideration of the 5σ standard. The second of O'Hagan's questions explicitly assumes that "it is better to do a proper Bayesian analysis." Were this the case, then the use of significance testing in HEP would indeed be puzzling, and one would want to investigate the reasons for the persistent failure of presumably well-trained and mathematically competent scientists to take advantage of the availability of a better method of analysis than that which they use. I will argue that O'Hagan's presupposition is in fact incorrect: the use of significance testing in such contexts as the Higgs search is well-warranted, and a Bayesian analysis is not "of course better."

My intention in this paper is not, however, to trudge along the well-worn paths of the debates between frequentists and Bayesians (though my course might intersect these at some points). Rather, I will ask what warrants the application of significance testing to the specific tasks for which HEP employs such tests? Answering that question requires attention both to the roles played by significance testing in the context of the experimental arguments of HEP and to the epistemological aims of experimental HEP. It also requires, I will argue, a framework for understanding how those aims inform the determination of a statistical methodology. For purposes of this paper, that framework is supplied by

pragmatism, a philosophical orientation first articulated in detail by C. S. Peirce. By considering the example of the search for the Higgs boson and the role of significance testing in the arguments emerging from that search, I will thus show how philosophical pragmatism can illuminate the statistical practices of HEP, revealing the warrant for those practices. The resulting perspective also holds promise, I will argue, for a better understanding of the second issue raised above: the warrant for the use of a 5σ standard for discovery in HEP.

I begin my argument in §2 with a quick summary of the methodology of significance testing, particularly as it is applied in the search for new phenomena in HEP. Then, in §3, I summarize some of the prominent criticisms of significance testing. To better understand the specific and limited role of significance testing, I turn in §4 to an overview of the argument offered by ATLAS in their July 2012 paper declaring "Observation of a New Particle." I sketch the basic tenets of pragmatism, roughly as espoused by Peirce, and extended by Churchman, in §5. Stage-setting thus completed, I will show in §6 how, by adopting the perspective of pragmatism, we can understand the warrant for HEP's use of significance testing, thus addressing the previously discussed criticisms of that practice. In §7 I offer some brief comments on the 5σ standard in light of the argument previously given, and then summarize the paper's conclusions in §8.

2 Significance testing in HEP

By characterizing their evidence in terms of an estimate of the statistical significance of their findings, ATLAS and CMS adopted the language and methodology of significance testing. Here I briefly summarize in a rough and informal way the main points of this widely used methodology.[2]

A significance test is a device for answering a question. To attempt an answer, one formulates a substantive hypothesis that is a possibly correct and testable answer to that question. This is the *null hypothesis* H_0. The investigator must devise a means of generating data for such a test, and then define some quantity, called a *test statistic*, that is a function of the data and has a known probability distribution supposing that hypothesis is true. The test statistic should be defined such that larger values indicate stronger evidence of departure from what is expected if the null hypothesis is true. The probability distribution of the test statistic under the null hypothesis is the *null distribution*. The null distribution thus serves as a mathematical *model* of the null hypothesis, and the direct target of the test is the *statistical hypothesis* that the data are generated by a process characterized by the null distribution. Knowing both the null distribution and the observed value of the test statistic, one may then ask: how probable is it that one would get a value as great or greater than that observed value, assuming the *statistical* null hypothesis is true? To the extent that the null distribution is a good model of the substantive null hypothesis, the answer to that question will

serve as a good estimate of the corresponding probability with regard to the null hypothesis itself.

In the context of a search for a new phenomenon in HEP, the purpose of the significance test is to quantify the extent to which indications of the sought-after phenomenon exceed what one would expect to see in the data in the absence of that phenomenon. The substantive null hypothesis is thus the hypothesis that only *background* processes contribute to the data. Typically one will conceptualize this in terms of a theoretical parameter μ, defined over a parameter space (set of possible values of μ) M, such that if the substantive null "background only" hypothesis is true, then $\mu = \mu_0; \mu_0 \in \mathrm{M}$, with the value of μ_0 conventionally set at zero. The *alternative hypothesis* H_1 can then be understood as asserting that $\mu = \mu_1; \mu_1 \in \mathrm{M}$, for some $\mu_1 \neq \mu_0$, so that the distinction between H_0 and H_1 amounts to a partition of M. Devising a test of the null hypothesis begins with the determination of a physical signature of the phenomenon sought after, based on its hypothetical features. In the case of a search for a particle, for example, such a signature might come in the form of the decay of the hypothetical particle into other, already established particles that are identifiable via their measurable properties. Experimenters must then operationalize that physical signature in terms of data selection criteria (*cuts*) that define *candidates* for the phenomenon in question. For a given set of cuts, they must then estimate the rate at which background processes will yield events satisfying those cuts. This amounts to the determination of the null distribution. However, HEP experiments commonly do

not use simply the number of candidate events as their test statistic, but instead the quantity $d(\mathbf{X}) = -2\ln\frac{\lambda(\mu_0|\mathbf{X})}{\sup\{\lambda(\mu_1|\mathbf{X})\}}$. This statistic uses the *likelihood function* $\lambda(\mu|\mathbf{X})$, a function that assumes different values for various values of $\mu \in \mathrm{M}$ for fixed data $\mathbf{X}$, such that $\lambda(\mu|\mathbf{X}) \equiv Pr(\mathbf{X};\mu)$. The quantity $d(\mathbf{X})$ thus takes greater values to the extent that the supremum (least upper bound) of the likelihoods of the alternative values of μ exceed the likelihood of the null value μ_0.

Once the data $\mathbf{X}$ are in hand, the observed value of the test statistic $d_0(\mathbf{X})$ can be recorded and the probability $Pr(d(\mathbf{X}) \geq d_0(\mathbf{X});\mu_0)$ can be calculated. This is the p-value of the experimental result. It has become standard practice in HEP to convert this probability number into a number of σ's by referring to the Standard Normal distribution (even when it does not describe the actual distribution of $d(\mathbf{X})$) and determining what number of standard deviations from the mean of that distribution would correspond to the p-value in question.

That, at any rate, is the basic outline of how significance testing is used in searches for new phenomena in HEP. Particular applications often involve additional complications, as we will see in the case of the Higgs search.

3 Problems with p-values?

O'Hagan's comment about HEP's reliance on frequentist statistical analysis (such as significance testing) being "bad science" calls to mind first of all the longstanding disputes between frequentist and Bayesian views about probability and statistical analysis. Significance testing in particular has come in for a great

deal of criticism, to the extent that calls for the reform or elimination of significance testing constitute a genre within the statistical literature.

In a classic of the genre Edwards, Lindman, and Savage (1963) note important respects in which inferences based on significance testing conflict with Bayesian inferences. They emphasize in particular the circumstances in which a small p-value will lead to rejection of (or conclusion of strong evidence against) a null hypothesis while a Bayesian inference would assign substantial probability to the null. This situation arises as a consequence of a general result, known as the *Jeffreys-Lindley paradox* (Jeffreys 1961; Lindley 1957). We need not here dwell on the derivation of this result, but can avail ourselves of the basic result. In Lindley's original paper, this is stated as follows:

> [I]f H is a simple hypothesis and x the result of an experiment, the following two phenomena can occur simultaneously:
>
> i. a significance test for H reveals that x is significant at, say, the 5% level;
>
> ii. the posterior probability of H, given x, is, for quite small prior probabilities of H, as high as 95%. (Lindley 1957, 187)

The intuitive explanation of this apparent conflict is that it comes about when the width of the sampling distribution is narrow in comparison to the range of possible parameter values. This can result, for example, from the accumulation of a large amount of data. Results that, *compared to the overall parameter space*, lie rather close to what one expects from the null distribution, may nonetheless register as

improbable under the null. Moreover, such an apparent mismatch between p-value and posterior probability can be made arbitrarily great for sufficiently large sample size n.

The condemnation of p-values that has followed from this result and related issues constitutes a veritable chorus calling either for the elimination of p-values or their modification through various ancillary devices (Berger & Sellke 1987; Diamond & Forrester 1983; Schervish 1996; Sellke, Bayarri, & Berger 2001; Sprenger 2013; Ziliak & McCloskey 2008).

Considering the problem a little more carefully, it becomes a little unclear exactly what basis the Jeffreys-Lindley paradox provides for objecting to the use of significance testing (see Robert 2014; Spanos 2013; Sprenger 2013 for some recent philosophical discussion). There is no direct contradiction, of course, between saying that the probability of getting $d(\mathbf{X}) \geq d(\mathbf{x_o})$, assuming H_0 is true, is less than 0.05 and saying that the probability of H_0 being true is 0.95. The probabilities in the two statements have entirely different meanings and entirely different foundations of assessment. Everyone acknowledges this point, but critics of significance testing go on to argue that the 'paradox' remains a problem for significance testing because it shows that taking p-values as a *measure of the evidence against the null hypothesis* is a bad idea. (This argument might be pressed under the assumption that some specific Bayesian measure of evidence (posterior probabilities, Bayes factors) *is* a good measure of evidence.)

Here again, though, the objection seems to miss the mark, since the

appropriate use of significance testing does not support the use of p-values as a measure of evidence against the null hypothesis. To object to the use of significance testing because it is sometimes misused is analogous to objecting to the use of electric guitars in general because you don't like 1980's "hair metal" bands. This point, too, is officially acknowledged by at least some critics, but they press the objection anyway by appealing to the ubiquity of the misuse. Hair metal bands are everywhere, and the only remedy is to ban electric guitars altogether. Berger and Sellke provide a good example of the dialectic. Having objected to the use of p-values as a measure of evidence against the null, they write:

> At this point, there might be cries of outrage to the effect that $p = .05$ was never meant to provide an absolute measure of evidence against H_0 and any such interpretation is erroneous. The trouble with this view is that, like it or not, people do hypothesis testing to obtain evidence as to whether or not the hypotheses are true, and it is hard to fault the vast majority of nonspecialists for assuming that, if $p = .05$, then H_0 is very likely wrong. (Berger & Sellke 1987, 114)

Here Berger and Sellke play off an ambiguity in the expression "people do hypothesis testing to obtain evidence." Their argument depends on understanding the use of hypothesis testing "to obtain evidence" as the performance of hypothesis testing *and* subsequent interpretation of the results of such tests as a measure of evidence. But another interpretation, and one that is compatible with the

appropriate use of significance testing, is that people regard the results of significance tests as *relevant evidence* regarding the correctness of the hypothesis under test, without the resulting p-values constituting a measure of the evidence against the tested hypothesis. Such would be the case, for example, if one thought that p-values should be considered in judging the correctness of the tested hypothesis, but that the specific bearing that any particular p-value has requires further interpretation in light of other pieces of information (which could include confidence intervals, effect size, severity analysis, likelihood functions, etc.) (Senn 2001).

I will not attempt here to assess whether it is true that the misuse or misinterpretation of p-values is widespread among scientists. I will not even directly address that question for the case of HEP in general. Instead I propose to look at the particular case of the Higgs boson discovery and examine how p-values are used in that case and whether the physicists of CMS and ATLAS are guilty of this allegedly widespread practice. To the extent that statistical methodology used in the arguments for the discovery of the Higgs is representative of statistical practice in HEP more generally, we can then at least formulate a hypothesis about whether significance testing in HEP is guilty of the charges raised by critics.

4 Context: the Higgs search at ATLAS

The announcements of the "observation" of the Higgs boson by CMS and ATLAS appeared in press releases made at the time of the July 4, 2012 press conference,

but the papers making the more detailed argument were released in preprint shortly thereafter (Aad et al. 2012b; Chatrchyan et al. 2012b). As an informed reading of those papers makes evident, it would be a grave misunderstanding to suppose that the arguments supporting their discovery claims depend on little more than the establishment of a p-value meeting the 5σ standard.

A detailed explanation of the arguments that CMS and ATLAS offer in support of their discovery claims would in fact far surpass what may reasonably be purveyed in the present context. Here I will restrict myself to a sketch, with just enough detail to establish the limited, though important, role of the significance calculation in the overall argument. I will also restrict myself to a discussion of the paper published by ATLAS. The CMS paper differs in some details, but the overall structure of their argument is the same in its essential features, and the general point could just as readily be made by reference to it.

The data on which ATLAS bases its discovery claim come from two distinct periods of data-collection. The 2011 dataset was collected with the LHC operating at a center-of-mass energy of $\sqrt{s} = 7$ TeV, while the 2012 dataset came from a $\sqrt{s} = 8$ TeV run. The introduction to ATLAS's paper notes that both ATLAS and CMS had already found excesses (Aad et al. 2012a; Chatrchyan et al. 2012a) beyond background expectations in the 2011 data "compatible with SM Higgs boson production and decay in the mass region 124–126 GeV, with significances of 2.9 and 3.1 standard deviations, respectively" (Aad et al. 2012b, 1). To further contextualize their present finding, they note the "broad excess in the mass region

120–135 GeV" reported by the CDF and D0 experiments at Fermilab's Tevatron collider.

Crucial to the argumentative strategy of the ATLAS paper is the identification of distinctive decay modes of the Higgs boson that lead to distinct search strategies. A Standard Model (SM) Higgs boson has seven distinct decay modes: $\gamma\gamma$, WW, ZZ, $\tau\tau$, bb, $Z\gamma$, and $\mu\mu$. All but the last two are mentioned in the ATLAS paper, but it is searches for the first three that provide the data on which the discovery claim rests. These different decay modes add evidential weight to the statistical significance argument insofar as they help to fix the theoretical interpretation of the excess indicated: the excesses show up in multiple channels in a manner that is predictable in light of knowledge about the rates at which the Higgs should decay in those channels and the size of the backgrounds in each of them. They also enable further probing of the import of the data insofar as most of the excess contributing to the statistical significance was found in just two of the five decay channels with relatively high signal-to-background ratios expected for a 125 GeV Higgs. The theoretical interpretation of the finding in terms of a Standard Model Higgs boson will require that excesses be found as well for the remaining channels, and the lack of data in these channels was one of the reasons why the ATLAS paper claims only the observation of "a new particle" rather than the Higgs boson specifically.

The guidance of theory is important to the validation of ATLAS's evidence claim in another way. ATLAS relies on statistical models of both the background

and the signal for a SM Higgs boson. Neither of these statistical models can be calculated directly from theory. Both require the use of simulation. For the signal, this simulation does depend, however, on a theoretical characterization of the processes by which Higgs bosons are produced (see, e.g., Harlander & Kilgore 2002). Understanding the signal is important both for developing and optimizing the analytic procedures to be applied to data and for the comparison of the observed excess with that expected for a SM Higgs with a mass near that reported. For present purposes, the latter consideration is particularly salient.

The theoretical description of the Higgs is crucial for ATLAS in guiding the simulation of the signal that is the target of their search. This is another respect in which an exclusive focus on the significance calculation obscures the overall character of their argument: in addition to demonstrating that they have achieved a statistical significance in excess of 5σ, ATLAS presents a comparison between the excess that they observe and what one would expect from SM Higgs decays near a mass of 125 GeV (see figure 1(b)). Moreover, they do this not only for the combined results, but also separately for the results from each of the three decay channels that figure in their discovery. In each case, it is important that the results fit, at least at a qualitative level, with the expectations from a SM Higgs boson with $m_H \sim 125$ GeV, and not so well with the expectations for a Higgs with mass far from that value. Such an agreement amongst different search modes would not be likely for data generated by background processes alone.

Another important aspect of the ATLAS argument is their characterization of

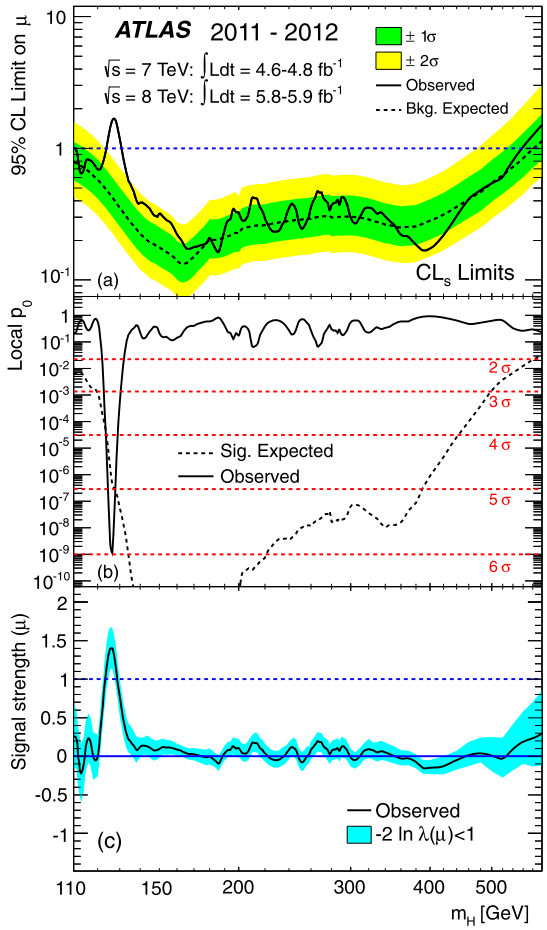

Figure 1: Three important ways of evaluating the ATLAS results. In (a) the solid line indicates 95% confidence limits on the value of μ established by the observed data, while the dotted line indicates the CL's that would be expected for background only, with bands showing the $\pm 1\sigma$ and $\pm 2\sigma$ uncertainties on the background expectation. In (b) the solid line gives the local p-value as a function of m_H, while the dotted line indicates the expected p-value based on simulation of the signal, also as a function of m_H. The best-fit estimate $\hat{\mu}$ of the signal strength as a function of m_H is given in (c).

the excess that they find. In particular, they estimate the mass of the new particle that they have observed using the profile likelihood ratio $\lambda(\mu, m_H)$.[3] On a plot of μ versus m_H, a confidence interval will, in the presence of a strong signal, yield a contour. ATLAS presents a plot that shows the 68% and 95% confidence intervals in the (μ, m_H) plane for each of the $H \to \gamma\gamma$, $H \to ZZ$, and $H \to WW$ channels. The first two form distinct but overlapping contours, while the latter yields no lower bound on m_H (figure 2). Regarding the separation between the contours based on the $H \to ZZ$, and $H \to WW$ channels, ATLAS notes that the "probability for a single Higgs boson-like particle to produce resonant mass peaks [in those two channels] separated by more than the observed mass difference, allowing the signal strengths to vary independently, is about 8%" (Aad et al. 2012b)

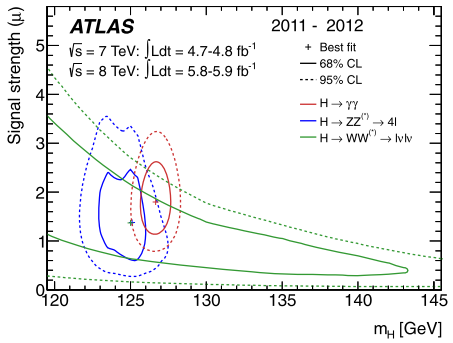

Figure 2: Confidence intervals in the (μ, m_H) plane for the $H \to \gamma\gamma$, $H \to ZZ$, and $H \to WW$ channels. Maximum likelihood estimates $(\hat{\mu}, \hat{m_H})$ are marked with '+'.

In particle physics experiments past, the problem of potential bias from

knowing too much about the data before finalizing the cuts has created difficulties. By "tuning the cuts," one might be able to enhance or create the appearance of a statistically significant signal. This problem had generated controversy in the early analyses of data from the search for the top quark at Fermilab (Staley 2002, 2004), and both ATLAS and CMS, well stocked with veterans of the CDF and D0 top searches, were eager to avoid a repetition of such problems. Consequently, ATLAS makes a point of establishing early in the paper that the new 8 TeV data, except for subsets used as control samples, were kept blinded while the analyses used for identifying candidate events in different categories were re-optimized on control and simulated data (Aad et al. 2012b, 1). This point supports the validity of that aspect of the argument that relies on the significance calculation: a biased procedure for selecting the cuts will invalidate the probability model on which the significance calculation depends. More precisely, the probability of a given excess of candidate events when only background processes are present is greater if the definition of 'candidate event' was chosen in a way that is tailored towards increasing the number of candidate events among the data in hand, compared to the number that would be found for cuts defined independently of the data in hand (Staley 2002). By keeping the data blinded while the analyses were developed, ATLAS protected themselves from being misled regarding the statistical characterization of the excess in their data.

At the same time, ATLAS implicitly acknowledges that the p-value itself is less than perfectly well-defined. The reason for this is related to the fact that the

distribution of the test statistic under the null hypothesis itself is not uniquely defined. The sensitivity of the experiment to the presence of decays of Higgs bosons depends in part on an unknown parameter: the mass of the Higgs boson m_H. Assuming that Higgs bosons do exist, the rate at which they are produced is a decreasing function of m_H. This bears on the definition of the test statistic, in which the likelihood function for the alternative hypothesis H_1 appears in the denominator. Put differently, m_H is a *nuisance parameter* in the Higgs search, a parameter on which the sampling distribution for $d(\mathbf{X})$ depends, but that has an unknown value. Both CMS and ATLAS faced this difficulty, and dealt with it using somewhat different implementations of the same strategy, which is to begin by regarding the p-value as a function of the parameter m_H, report that function (see Figure 1b),[4] and then find the minimum value of that function p_{min}. The latter quantity is then reported as the *local* p-value. However, the value of m_H, being unknown at the outset, is not in fact set in advance, and the local p-value is thus a fiction of sorts. The probability (call it "p_{real}" for now) that the experiment would report an excess as great as that observed *for some value or other* of the Higgs mass, assuming the null hypothesis is true, is greater than the local p-value. This is known as the "Look Elsewhere Effect" (LEE).

How much greater is p_{real} than p_{min}? That depends on the range of values of m_H that one considers, and just what that range should be is not uniquely defined. For this reason my suggested label p_{real} must be discarded as strictly aspirational at best, along with any illusions we might have held to this point about the p-value

for the Higgs search results having a uniquely well-defined value. Instead, both ATLAS and CMS reported, along with their local p-values, *global* p-values. Global p-values are defined relative to specified ranges of values for m_H, and both groups, in order to emphasize that these ranges are "arbitrary or subjective" (Cousins 2013, 33), reported both "wide" and "narrow" ranges, based on different criteria. Cousins notes, "Some possibilities were the range of masses for which the SM Higgs boson [had] not previously been ruled out at high confidence; the range of masses for which the experiment is capable of observing the SM Higgs boson; or the range of masses for which sufficient data had been acquired to search for any new boson. The experiments made different choices" (ibid.). They certainly did. The narrow range reported by ATLAS runs from 110 to 150 GeV (with a significance of 5.3σ), while the wide range reported by CMS is 110–145 GeV (with a significance of 4.5σ). Reporting these global significance values thus serves as a kind of check on the sensitivity of the statistical significance to the LEE.

To an outsider, the attitudes of particle physicists towards p-values evidenced here may seem conflicted, or even inconsistent. On the one hand, the concern over bias arising from tuning the cuts might suggest an *objectivist* attitude towards p-values, according to which the significance calculation is aimed at estimating the value of a quantity that has a well-defined objective value, a successful estimate of which will be close to that objective value. On the other hand, the discussion of the LEE acknowledges a crucial ambiguity regarding the p-value, reflecting what we might call a *fictionalist* attitude, which denies that the p-value has a uniquely

correct value, but persists in regarding it as a useful quantity to discuss. The objectivist attitude would not by itself commit particle physicists to regarding the p-value as a measure of the evidence that a body of data provides against the null hypothesis, but the fictionalist attitude is clearly incompatible with treating the p-value in this way.

I propose that ATLAS's concern about bias arising from tuning the cuts can be given a natural interpretation that is coherent with a fictionalist attitude toward their p-values while also making sense of their treatment of the LEE. According to this alternative interpretation, although there is no well-defined, uniquely correct p-value, some ways of estimating the statistical significance of the results of the Higgs search have a greater tendency than others to provide misleading guidance regarding the severity[5] with which potential errors have been ruled out, particularly the error of mistaking a stochastic variation in the size of the excess for an effect of a genuinely new physical phenomenon.

On this view, the relevance of the p-value of the Higgs search results for ATLAS's claim to have discovered a new particle is that it quantifies one dimension of a multi-dimensional evaluation of the evidence supporting that claim. It is important, for the purpose of cogently arguing for their claim, that ATLAS be able to establish that they have taken sufficient care to rule out, on a reasonable basis, the possibility of being misled by a stochastic fluctuation in the background. The calculation of a p-value addresses that need. As David Cox has noted, "significance tests ... address the question of whether the data are reasonably consistent with a

null hypothesis in the respect tested. This is in many contexts an interesting but limited question" (Cox 2006, 42). Other dimensions of ATLAS's assessment of the evidence that are essential to their experimental argument include the distribution of the candidate events across different decay modes, the comparison of the data with theoretical expectations for a Higgs boson with a mass in the range indicated by the data, the ability to arrive at an estimated mass for the candidate decay events, the comparison of the estimated mass for different decay channels, and the satisfaction of methodological constraints on the treatment of data.

5 Pragmatism

My contention in this paper is that attending to the role that significance calculations play in ATLAS's experimental argument will help us to understand the warrant for their reliance on significance testing, and that such understanding is facilitated by the philosophical perspective provided by pragmatism. In this section, I briefly characterize the pragmatic philosophical perspective, with particular attention to its application in the context of choosing statistical methods. A thorough discussion of pragmatism itself would naturally exceed the scope of the present paper, and I will seek only to sketch enough of the aims and approach of pragmatism to illuminate how it contributes to understanding the warrant for the use of significance testing in the Higgs search.

When philosophical issues are mentioned at all in particle physicists' discussions of statistical methodology, they are typically disavowed. For example,

in discussing the statistical methodology developed by the Higgs working group for establishing exclusion limits at a 2000 CERN workshop, A. L. Read notes that

> It has not been an explicit goal of the Higgs working group to choose a frequentist(-like) analysis rather than a Bayesian analysis on philosophical grounds. Our attitude is rather practical, we want to do the best we can with the data that we have, where the best we can means excluding the Higgs as strongly as possible in its absence . . . and confirming its existence as strongly as possible in its presence. . . while holding the probabilities of falsely excluding a true signal or falsely discovering a non-existent signal at or below specified levels. (Read 2000, 82)

Disavowals notwithstanding, I claim that Read's practical attitude expresses an implicit, underlying, pragmatic philosophical outlook. Pragmatism constitutes a philosophical orientation that grows out the writings of Charles Sanders Peirce, William James, and John Dewey, among others. Although 'pragmatism' is a capacious term covering a broad range of philosophical points of view, the writings of these "classical" pragmatists are of special importance in establishing the central features of pragmatic philosophies, and Peirce in particular has important insights into scientific reasoning.

I do not, of course, contend that physicists in HEP have any particular grounding in the writings of these authors. (I suspect most HEP physicists have never read them or given them a thought.) Neither do I suppose that there is a

uniform statistical philosophy shared by all members of the HEP community. Indeed, some particle physicists have advocated strongly for the abandonment of the frequentist methods that are routinely employed in HEP, in favor of Bayesian methods (D'Agostini 2003; D'Agostini 2011; Bhat, Prosper, & Snyder 1997; Prosper 2006). Rather, I maintain that HEP physicists have, in response to the inferential problems that face them, *collectively* developed an approach on their own that looks very much like applied pragmatism – i.e., they are doing just what one would, on the basis of philosophical pragmatism, recommend that they do. That pragmatism begins with a disavowal of commitments to a fixed statistical ontology, but it does not end there.

In an 1878 paper titled "How to Make Our Ideas Clear," Peirce sought to articulate a methodology for achieving a higher grade of clarity than previously articulated. Whereas Descartes's standard for "clear and distinct" ideas suffered from its reliance on the efficacy of a power of introspection, on which he had previously cast doubt in his 1868 essay "Questions Concerning Certain Faculties Claimed for Man" (Peirce 1992[1868]), Peirce now proposed "a method of reaching a clearness of thought of a far higher grade" (Peirce 1992[1878], 127). That method Peirce articulates in the form of a "rule for attaining the third grade of clearness of apprehension":

> Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these

effects is the whole of our conception of the object. (ibid., 132)

Let us call this rule *the pragmatic maxim*. The term "practical" here must be understood in its broad philosophical sense, in which it refers to intentional actions of any sort. Effects that might conceivably have practical bearings are thus those that we can consider, for some possible circumstances, to be relevant to some choice of action. Importantly, that action might be one undertaken in the course of pursuing some inquiry. Such is precisely the context in which scientists must choose a statistical inference procedure, and a pragmatic orientation towards statistics would require the application of the pragmatic maxim to the potential inferences to be drawn in a given inquiry.

The practical orientation of the pragmatic maxim requires such inferences to be regarded as a species of action, undertaken for a certain aim, with a range of potential consequences. This point was emphasized clearly in the work of C. West Churchman, who developed an account of "pragmatic inference" that incorporated the theories of statistical inference of Neyman and Pearson and Wald (which were in turn anticipated by Peirce's own "theory of probable inference" (Peirce 1883)) into a broader pragmatic framework drawing on the work of Peirce and Dewey (Churchman 1948). For an inference regarded pragmatically, three kinds of question must be addressed to adequately apply the pragmatic maxim: (1) What is your aim? (2) How might you go wrong? and (3) What is at stake? Tailoring these questions a little more closely to the problematic of a scientific investigator seeking

a statistical methodology, these become: (1) What are the learning goals of the experiment? (2) What are the possible errors that must be confronted? and (3) What are the foreseeable practical consequences of those errors or their absence, including those that bear on further and related inquiries?

6 Warrant for significance testing

It is these three questions that hold the key to the pragmatic warranting of ATLAS's statistical practices, so it is worth taking them in turn.

Beginning with the aims of the Higgs search at ATLAS, it goes without saying that among their aims was to answer the question "Does the Higgs boson exist?" But leaving it at that underspecifies their aim egregiously. A clearer statement can be found in the previously quoted words of A. L. Read: "we want to do the best we can with the data that we have, where the best we can means excluding the Higgs as strongly as possible in its absence ... and confirming its existence as strongly as possible in its presence... while holding the probabilities of falsely excluding a true signal or falsely discovering a non-existent signal at or below specified levels." It is the last-mentioned aspect of this aim that renders the reliance on significance testing appropriate: the aim of limiting the probability of declaring a discovery in the presence of a mere fluctuation in the background calls for precisely the kind of probabilistic assessment that issues from a calculation of a p-value.

Turning to the second question, we can see that the articulation of the aims of the Higgs search already incorporate an important part of the answer to the

question of the possible errors. Thinking of the possible errors simply as accepting the existence of the Higgs when only background processes are present and failing to accept the Higgs hypothesis when the Higgs is present provides us with only the coarsest-grained description of the landscape of errors surveyed by the ATLAS physicists. The complexity of that landscape reflects the complexity of the analytic procedures ATLAS brought to bear on their data. For the purpose of simply assessing the warrant for ATLAS's reliance on significance testing it will have to suffice that the two primary ways of going wrong – accepting the background only hypothesis when it is false, rejecting it when it is true – are the ultimate source of concern, and all of the more fine-grained possibilities of error become relevant precisely because of their potential to lead to one or the other of these two main errors.

ATLAS's choice of a methodology of significance testing is warranted in light of their aims and in light of the kinds of errors they sought to avoid. There is no evidence that they sought to provide a quantitative measure of the evidence against the background hypothesis or in favor of the Higgs (or even Higgs-like) hypothesis. They sought instead to be able to make a clear and compelling case either for or against the existence of the Higgs, while limiting the probability of doing so erroneously. Significance testing alone is not sufficient for this aim, but it can contribute to the pursuit of it, by providing evidence regarding the compatibility of the data with the background-only hypothesis. The Jeffreys-Lindley paradox entails that such evidence must be interpreted in light of

further information, but not that significance calculations are not relevant evidence regarding the acceptability of the null.

The third of the pragmatic questions receives no explicit treatment in ATLAS's published Higgs results. This, however, does not mean that consideration of it played no identifiable role in their deliberations over the statistical assessment of their data. On the contrary, the consequences of erroneously announcing a discovery of the Higgs played an important role in their reliance on what many regarded as an extremely strict standard of significance: the "5σ" rule previously mentioned. As this may constitute the clearest instance of pragmatic thinking in this episode, this point deserves its own discussion.

7 The 5σ standard

Although the requirement that discovery claims in HEP be premised on statistical excesses with p-values that equate to at least five standard deviations for a Gaussian distribution has assumed the status of tradition within the HEP community,[6] it has no official institutional codification and physicists will deny that its normative force is absolute. According to Joe Incandela, who was spokesperson for CMS at the time of the July 2012 announcements, "the 5 sigma standard is generally misunderstood outside the field. We do not take 5 sigma as absolutely necessary nor do we assume all 5 sigma results to be correct" (personal communication). Similarly, CMS member Robert Cousins comments, "I do not believe that experienced physicists have such an automatic response to a p-value,

but it may be that some people in the field may take the fixed threshold more seriously than is warranted" (Cousins 2013, 30). Meanwhile, some physicists have called for reform of the 5σ standard. Louis Lyons, for example, has called for a "more nuanced criterion" that would be more or less demanding for a variety of possible future discoveries, based on four criteria: the presence of an LEE, the magnitude of systematic uncertainties, the impact of the discovery, and the "degree of surprise" (also called the "subconscious Bayes' factor") (Lyons 2013).

Lyons' criteria cohere well with responses that Tony O'Hagan received from physicists to his query regarding the rationale for the 5σ criterion, mentioned in §1. Acknowledging the statistical (and pragmatic) inappropriateness of an ironclad significance threshold for discovery claims, these responses (apart from a minority of Bayesians calling for the abandonment of significance testing altogether) indicated an acceptance of 5σ as an appropriate standard for the Higgs search itself. (In Lyons' enumeration of varying significance standards from 3 to >8 standard deviations for fourteen different HEP searches, the standard for the Higgs search remains at 5σ.) Prominent among the considerations cited are the LEE and systematic uncertainty, or more generically, to quote O'Hagan's summary, the fact that "so much can go wrong that it makes sense to guard against false positives caused by errors in underlying assumptions, pre-processing, experimental controls, etc." (O'Hagan 2012b, 5).

The problems of the LEE and systematic uncertainties constitute obstacles toward taking the calculated local significance seriously as a measure of what it

purports to be: the probability of observing an excess as great as or greater than that observed, assuming that only background processes are present. They leave unaddressed the further questions of why the standard for discovery should be a very demanding one in the first place (why is 3σ not good enough?) and why it should not be even more demanding (why would 8σ not be even better?). These questions can be addressed only through the criterion of the impact of the discovery, which is to say, the consideration of the third pragmatic question: What are the foreseeable practical consequences of the possible errors or their absence? Although answers to this question do not determine univocally a precise standard that must be applied (as with any convention, some element of arbitrariness in the choice will always remain), they will illuminate the reasons that shaped the terrain in which the decision was made.

The pragmatic perspective requires us to acknowledge that the outcome of an inference is not only an event in an abstract realm of ideas, but is a decision with practical consequences. As noted by C. West Churchman,

> In pragmatic methodology, every scientific hypothesis is considered to be a possible course of action for accomplishing a certain end, or set of ends. Pragmatically speaking, an inability to say what one intends to do as a result of accepting one out of a set of alternative hypotheses, is an inability to state the hypotheses themselves in adequate terms. (Churchman 1948, 259)

We can place the consequences of the decisions of the ATLAS and CMS groups to

announce discoveries into two categories: those that pertain directly to the logical argumentation of future physics inquiries, and those that pertain indirectly to the aims of ATLAS, CMS, and the HEP community more broadly.

Regarding the first category, accepting the existence of a new boson amounts to a commitment to adopt statements entailing the existence of such a particle as premises in the pursuit of further inquiries. This commitment has its most obvious salience for the continued work of ATLAS and CMS themselves, as their analytic tasks turn from the aim of producing exclusion plots towards the aim of measuring the properties of the newly discovered particle and probing further implications of the Higgs hypothesis to fix more securely the theoretical interpretation of their finding. For other physicists working on SM and Beyond-SM problems, the announcement by ATLAS and CMS has the consequence of changing the logical terrain. Although each investigator must decide (whether as an individual or as a member of a working group) whether the evidence offered by the two CERN groups suffices to warrant agreement with their discovery claims, it seems likely that the burden now lies on those who would decline those claims to explain their dissent. These considerations contribute to our understanding of the 5σ standard for the Higgs search by highlighting the importance, for the pursuit of physics inquiries within ATLAS and CMS as well as beyond, of guarding against an erroneous discovery claim, while also pointing towards the tremendous value of that discovery claim, as it enables the pursuit of new inquiries that, prior to discovery, had to wait offstage.

The second category of consequences must be regarded as somewhat more speculative, but various statements of physicists involved in the Higgs search at least suggest some relevant considerations. CMS's published paper declares in its introduction that "The discovery or exclusion of the SM Higgs is one of the primary scientific goals of the Large Hadron Collider" (Chatrchyan et al. 2012b, 30). Given the great expense of building the LHC and operating the CMS and ATLAS experimental programs, it is not surprising that success at achieving this goal was highly valued. The much-anticipated discovery claims themselves were not merely attended by submitting papers for publication, but by a kind of scientific showmanship featuring a press conference that was broadcast via the internet worldwide and featured prominently among the news of the day. To get things wrong would have been tremendously embarrassing. Although one cannot be certain of the consequences of such an error, it is not unreasonable to imagine them including even a political dimension with negative consequences for the funding of HEP.

One respondent to O'Hagan's query communicates vividly the personal nature of such considerations: "In fact, we do have high standards because in our view we are trying to arrive at 'true' statements about the world in the pragmatic sense that these statements yield predictions that turn out to be correct. Given that the search for the Higgs took some 45 years, tens of thousands of scientists and engineers, billions of dollars, not to mention numerous divorces, huge amounts of sleep deprivation, tens of thousands of bad airline meals, etc., etc., we want to

be sure as is humanly possible that this is real" (O'Hagan 2012b, 5).

In addition to concerns about the amount of effort and expense that had gone into the search for the Higgs and its importance to the scientific project of the LHC, a broader sense of responsibility toward the public perception of science in general may have played a role in the cautious attitude toward any discovery announcement. According to CMS member Robert Cousins, the intense public spotlight that the LHC had felt since 2008 made it clear that there was an opportunity to try to show science of very high quality to the general public, in an environment where there was public skepticism about some scientific claims. Certainly making a discovery announcement that subsequently turned out to be erroneous carried a very high cost, and could only contribute to such skepticism (personal communication).

Taking the pragmatic perspective allows us to see that such considerations regarding the consequences of an inference are not extraneous to the scientific process, but rather help to clarify it. A clear articulation of the meaning of an inference will bring to light its practical dimension, thus helping us to understand the evidential standards that have been brought to bear on it – standards that might otherwise seem entirely arbitrary or mysterious.

8 Conclusion

This paper has argued that pragmatism helps us to understand how the statistical methodology of HEP is warranted. The argument focused on the particular case of

the use of p-values in the argument for the discovery of a new Higgs-like boson based on the Higgs search results at ATLAS and CMS. In spite of the oft-repeated objections to p-values, the LHC physicists' use of them was warranted because they employed significance testing for the specific purpose of providing evidence relevant to the multi-dimensional assessment of the hypothesis that their excess of Higgs candidates was due to a stochastic fluctuation of non-Higgs background processes, and not for the purpose of providing a quantitative measure of the evidence against the background-only hypothesis. Their use of significance testing was tailored to specific inferential aims, in light of explicit consideration of the possible errors that could be made in drawing that inference, and with at least implicit attention to the consequences of such errors, both for immediate matters of related scientific inquiries and for broader matters related to the place of HEP and science in society.

In spite of this defense, I do not wish to be understood as stating that all is well regarding the statistical methodology in HEP and we can be satisfied with the status quo. There are three dimensions along which improvements would be desirable, and to some extent are already being pursued by physicists themselves.

The first of these consists of the use of better statistical methods within HEP. Statisticians persist in the pursuit of ever more powerful methodologies, tailored to the demands of specific kinds of inferential problems. Particle physicists themselves have in the past contributed to the development of new statistical methods (Feldman & Cousins 1998; Prosper 1988). and there is no reason to think that further improvements cannot or should not be pursued. Acknowledging the

limitations on significance calculations as they are employed in HEP is itself a reason, not only to supplement p-values with other arguments as described here, but also to seek better tools. This task is ongoing and appropriately undertaken by physicists and statisticians in consultation, as well as by statistically sophisticated physicists.

A second route to doing better would be for more members of the HEP community to achieve the level of understanding of the results of applying statistical methods that has already been achieved by many physicists who are well-educated in statistics and reflective on the philosophical nuances of statistical methodology. Some members of the HEP community have already noted the connection between statistical education and clear communication of the meaning of scientific results (D'Agostini 2011). To improve in this respect, it is not necessary that physicists become philosophers of statistics, but only that the statistical education of physicists include an ongoing emphasis on the kind of conceptual and interpretive issues that are all too easily dismissed as "just philosophy."

The final route to doing better vis a vis statistics in HEP is that pointed out by attention to Peirce's pragmatic maxim and the three questions that derive from its application to problems of statistical inference, particularly the third question: "What is at stake?" I have argued here that considerations of the consequences of possible errors of inference played an important (though not exclusive) role in the determination of standards of evidence for purposes of announcing a discovery based on the Higgs search results at LHC. Discussions of the practical

consequences of accepting a hypothesis are part of the pragmatic clarification of an inference. Yet current norms governing scientific communication tend to force such discussions into informal, background contexts, so that the resulting decisions appear to the public as they were reported in the press following the Higgs announcement of July 2012: as a "gold standard" or as a "strict notion of scientific certainty" the status of which is simply to be taken for granted. Although I would not propose that every positive scientific claim must be accompanied by a detailed discussion of the deliberations that guided the choice of evidential standard that was applied to that claim, I do think that a more complete execution of the program of pragmatic clarification should include a more systematic expectation that scientists in fields such as HEP should address explicitly and thoroughly the considerations – including those regarding potential consequence of errors – that guide such decisions.

References

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., ... Zwalinski, L. (2012a). Combined search for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Physical Review D*, *86*, 032003. doi: 10.1103/PhysRevD.86.032003

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., ... Zwalinski, L. (2012b). Observation of a new particle in the search

for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters*, *B716*, 1–29. doi: 10.1016/j.physletb.2012.08.020

Abachi, S., Abbott, B., Abolins, M., Acharya, B. S., Adam, I., Adams, D. L., ... Zylberstejn, A. (1995). Observation of the top quark. *Physical Review Letters*, *74*, 2632–2637. Retrieved from `http://link.aps.org/doi/10.1103/PhysRevLett.74.2632` doi: 10.1103/PhysRevLett.74.2632

Abe, F., Akimoto, H., Akopian, A., Albrow, M. G., Amendolia, S. R., Amidei, D., ... Zucchelli, S. (1995). Observation of top quark production in $\bar{p}p$ collisions with the collider detector at Fermilab. *Physical Review Letters*, *74*, 2626–2631. doi: 10.1103/PhysRevLett.74.2626

Abe, F., Albrow, M. G., Amendolia, S. R., Amidei, D., Antos, J., Anway-Wiese, C., ... Zucchelli, S. (1994). Evidence for top quark production in $\bar{p}p$ collisions at $\sqrt{s}$ =1.8 TeV. *Physical Review D*, *50*, 2966–3026. doi: 10.1103/PhysRevD.50.2966

ATLAS. (2012). *Latest results from ATLAS Higgs search.* (Press release)

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*(397), pp. 112-122.

Bhat, P. C., Prosper, H. B., & Snyder, S. S. (1997). Bayesian analysis of multi-source data. *Physics Letters B*, *407*(1), 73–78. doi: http://dx.doi.org/10.1016/S0370-2693(97)00723-5

Chatrchyan, S., Khachatryan, V., Sirunyan, A., Tumasyan, A., Adam, W., Bergauer, T., ... Swanson, J. (2012a). Combined results of searches for the standard model higgs boson in pp collisions at $\sqrt{s} = 7$ TeV. *Physics Letters B*, *710*(1), 26 - 48. doi: http://dx.doi.org/10.1016/j.physletb.2012.02.064

Chatrchyan, S., Khachatryan, V., Sirunyan, A., Tumasyan, A., Adam, W., Bergauer, T., ... Swanson, J. (2012b). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters*, *B716*, 30–61. doi: 10.1016/j.physletb.2012.08.021

Churchman, C. W. (1948). Statistics, pragmatics, induction. *Philosophy of Science*, *15*(3), 249–268.

CMS. (2012). *Observation of a new particle with a mass of 125 GeV.* (Press release)

Cousins, R. D. (2013, October). *The Jeffreys-Lindley paradox and discovery criteria in high energy physics.* (arXiv:1310.3791)

Cox, D. R. (1970). *Analysis of binary data*. London: Methuen.

Cox, D. R. (2006). *Principles of statistical inference*. New York: Cambridge University Press.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman and Hall.

D'Agostini, G. (2003). *Bayesian reasoning in data analysis: A critical introduction*. Singapore: World Scientific.

D'Agostini, G. (2011). *Probably a discovery: Bad mathematics means rough*

scientific communication. (arXiv:1112.3620)

Diamond, G. A., & Forrester, J. S. (1983). Clinical trials and statistical verdicts: Probable grounds for appeal. *Annals of Internal Medicine*, *98*, 385–394.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242.

Feldman, G. J., & Cousins, R. D. (1998). Unified approach to the classical statistical analysis of small signals. *Physical Review D*, *57*, 3873–3889. doi: 10.1103/PhysRevD.57.3873

Franklin, A. (2013). *Shifting standards: Experiments in particle physics in the twentieth century.* Pittsburgh, PA: University of Pittsburgh Press.

Harlander, R. V., & Kilgore, W. B. (2002). Next-to-next-to-leading order Higgs production at hadron colliders. *Physical Review Letters*, *88*, 201801. doi: 10.1103/PhysRevLett.88.201801

Jeffreys, H. (1961). *Theory of probability* (Third ed.). Oxford: Oxford University Press.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1/2), 187–192.

Lyons, L. (2013). *Discovering the significance of* 5σ*.* (arXiv:1310.128)

Mayo, D. G. (1996). *Error and the growth of experimental knowledge.* Chicago: University of Chicago Press.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *The British Journal for the Philosophy of Science*, *57*(2), 323–357.

Mayo, D. G., & Spanos, A. (Eds.). (2009). *Error and inference: Recent exchanges on experimental reasoning, reliability, objectivity, and rationality.* New York: Cambridge University Press.

O'Hagan, T. (2012a). *Higgs boson.* Retrieved March 17, 2014, from `http://bayesian.org/forums/news/3648`

O'Hagan, T. (2012b). *Higgs boson digest and discussion.* Retrieved March 17, 2014, from `http://bayesian.org/forums/news/3830`

Overbye, D. (2012, July 4). Physicists find elusive particle seen as key to universe. *New York Times*.

Peirce, C. S. (1883). A theory of probable inference. In C. S. Peirce (Ed.), *Studies in logic: by members of the Johns Hopkins University* (pp. 126–181). Boston: Little, Brown, and Company.

Peirce, C. S. (1992[1868]). Questions concerning certain faculties claimed for man. In N. Houser & C. Kloesel (Eds.), *The essential peirce* (Vol. 1, pp. 11–27). Bloomington, IN: Indiana University Press.

Peirce, C. S. (1992[1878]). How to make our ideas clear. In N. Houser & C. Kloesel (Eds.), *The essential peirce* (Vol. 1, pp. 124–141). Bloomington, IN: Indiana University Press.

Prosper, H. B. (1988). Small-signal analysis in high-energy physics: A Bayesian approach. *Physical Review D*, *37*, 1153–1160. doi: 10.1103/PhysRevD.37.1153

Prosper, H. B. (2006). *Probability and statistical inference.* (arXiv:0606179)

Read, A. L. (2000). Modified frequentist analysis of search results (the CL_s method). In F. James, L. Lyons, & Y. Perrin (Eds.), *Workshop on confidence limits* (pp. 81–102). Geneva: CERN.

Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, *81*, 216–232.

Schervish, M. J. (1996). p values: What they are and what they are not. *The American Statistician*, *50*(3), 203–206.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62–71.

Senn, S. (2001). Two cheers for p-values? *Journal of Epidemiology and Biostatistics*, *6*, 193-204.

Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science*, *80*(1), 73–93.

Sprenger, J. (2013). Testing a precise null hypothesis: The case of Lindley's paradox. *Philosophy of Science*, *80*(5), 733–744.

Staley, K. W. (2002). What experiment did we just do? counterfactual error statistics and uncertainties about the reference class. *Philosophy of Science*, *69*(2), 279–299.

Staley, K. W. (2004). *The evidence for the top quark: Objectivity and bias in collaborative experimentation*. New York: Cambridge University Press.

Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical*

Society. Series C (Applied Statistics), *37*(1), 87–94.

Wickham, C., & Evans, R. (2012, July 4). "It's a boson:" Higgs quest bears new particle. *Reuters*.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

Notes

[1]O'Hagan collected and summarized the many replies he received to his post. In this digest, he noted that he had intentionally used somewhat inflammatory language to "provoke discussion" (O'Hagan 2012b).

[2]A more systematic treatment is given in numerous statistical texts; see, e.g., Cox & Hinkley 1974.

[3]When confronted with a statistical model with multiple parameters, all but one of which are considered 'nuisance' parameters, the profile likelihood for the parameter of interest is obtained by maximizing over the likelihoods of those parameters (Cox 1970; Venzon & Moolgavkar 1988).

[4]As Cousins states, "for each mass $[m_H]$ there is a p-value for the departure from H_0, *as if that mass had been fixed in advance*" (Cousins 2013, 33, emphasis in original).

[5]I have in mind here the notion of severity discussed by Mayo (e.g., 1996; 2006, 2009), though perhaps nothing in my argument depends essentially on this.

[6]Allan Franklin has documented the emergence of the 5σ standard in HEP (Franklin 2013). According to Franklin's narrative, the standard has only

assumed the weight that it does carry rather recently, around the time of the discovery of the top quark, for which an initial paper by CDF in 1994 (Abe et al. 1994) claimed only "evidence" (with a significance corresponding to 2.8σ for a Gaussian distribution), while later papers by CDF (Abe et al. 1995) and D0 (Abachi et al. 1995) claimed the top's "observation" on the basis of 5.0σ and 4.6σ, respectively (Staley 2004).