The Jeffreys-Lindley Paradox and Discovery Criteria in High Energy Physics

Robert D. Cousins*
Department of Physics and Astronomy
University of California, Los Angeles, California 90095, USA

October 14, 2013

Abstract

The Jeffreys-Lindley paradox displays how the use of a p-value (or number of standard deviations z) in a frequentist hypothesis test can lead to inferences that are radically different from those of a Bayesian hypothesis test in the form advocated by Harold Jeffreys in the 1930's and common today. The setting is the test of a point null (such as the Standard Model of elementary particle physics) versus a composite alternative (such as the Standard Model plus a new force of nature with unknown strength). The p-value, as well as the ratio of the likelihood under the null to the maximized likelihood under the alternative, can both strongly disfavor the null, while the Bayesian posterior probability for the null can be arbitrarily large. The professional statistics literature has many impassioned comments on the paradox, yet there is no consensus either on its relevance to scientific communication or on the correct resolution. I believe that the paradox is quite relevant to frontier research in high energy physics, where the model assumptions can evidently be quite different from those in other sciences. This paper is an attempt to explain the situation to both physicists and statisticians, in hopes that further progress can be made.

*cousins@physics.ucla.edu

Contents

1 Introduction

On July 4, 2012, the leaders of two huge collaborations (CMS and ATLAS) presented results at a joint seminar at CERN on the French-Swiss border outside Geneva, each describing the observation of a "new boson", suspected to be the long-sought Higgs boson (Incandela and Gianotti, July 4, 2012). The statistical significances of the results were expressed in terms of "σ": carefully calculated p-values (not assuming normality) were mapped onto the equivalent number of standard deviations in a one-tailed test of the mean of a normal distribution. ATLAS observed 5σ significance by combining the two most powerful detection modes (different sets of particles into which the boson decayed) in 2012 data with full results from earlier data. With independent data taken with a different apparatus and only partially correlated analysis assumptions, CMS observed 5σ significance in an similar combination, and when combining with some other modes as CMS had planned for that data set, 4.9σ.

With both ATLAS and CMS also making similar observations about the size of the effect (magnitude of the signal strength), the new boson was immediately interpreted as the most anticipated and publicized discovery in high energy physics (HEP) since the Web was born (also at CERN). Journalists went scurrying for explanations for what a "σ" is, and why "high energy physicists require 5σ for a discovery". Meanwhile, those who knew a little or a lot about Bayesian model selection began to wonder out loud why high energy physicists were still using frequentist p-values to test a point null hypothesis against a composite alternative.

In this paper, I try to explain some of the traditions of discovery in high energy physics, which have a decidedly frequentist flavor, drawing in a pragmatic way on both Fisher and Neyman-Pearson, even where these giants disagreed over philosophical foundations. Of course, a number of us have been aware for many years of the criticisms of this approach, having had the real pleasure of interacting with some of the top Bayesian statisticians (both subjective and objective in flavor) who attended HEP workshops on statistics. These issues lead directly to a famous "paradox", as Lindley called it, when testing the hypothesis of a specific value of a parameter, θ_0, against a continuous set of alternatives θ. The different sample-size scaling of p-values and Bayes factors, described by Jeffreys and emphasized by Lindley, can lead the frequentist and the Bayesian to opposite inferences.

However, as this paper describes, it is an understatement to say that the community of Bayesian statisticians has not reached full consensus on what should replace p-values in scientific communication. For example, two of the most prominent voices of "objective" Bayesianism (Jim Berger and José Bernardo) advocate fundamentally different approaches to hypothesis testing for scientific communication, as discussed below. Furthermore, in surveying the Bayesian literature, it is striking to me how different the assumptions about "models" are in high energy physics compared to the social sciences, for example.

Thus, my goal in this paper is to describe today's rather unsatisfactory situation as I see it. Progress in high energy physics meanwhile continues, but I think it would be potentially quite useful if more statisticians became aware of our special circumstances, and reflected on what the Jeffreys-Lindley paradox means to high

energy physics, and vice versa.

In "high energy physics", also known as "elementary particle physics", the objects of study are the smallest building blocks of matter and the forces between them. The 2004 Nobel Lecture by prize-winner Frank Wilczek (2004) provides an introduction with insights and descriptions well beyond the topic for which he shared the prize (the equation for one of the four forces of nature). For a variety of reasons, the experimental techniques often make use of the highest-energy accelerated beams attainable. But due to the magic of quantum mechanics, we can probe even higher energies by carefully observing decays in intense lower-energy beams, with the ultimate reach potentially attainable by watching a large vat of liquid (hoping to see protons decay); and since the early universe was hotter than our most energetic beams, and still has powerful cosmic accelerators and extreme conditions, astronomical observations are a crucial source of information on "high energy physics". Historically, some discoveries in high energy physics have been in the category known to statisticians as "the interocular traumatic test; you know what the data mean when the conclusion hits you between the eyes." (Edwards et al, 1963, p. 217, citing J. Berkson). In other cases, the evidence accumulated slowly, and it was considered essential to quantify the evidence in a fashion that relates directly to this review.

A wide range of views on the Jeffreys-Lindley paradox can be found in reviews with commentary by many distinguished statisticians, in particular those of Shafer (1982), Berger and Sellke (1987), Berger and Delampady (1987a), and Robert, Chopin, and Rousseau (2009). The review of Bayes factors by Kass and Raftery (1995) is also a useful resource, and the earlier book by economist Leamer (1978) offers many interesting insights. Some of the views these authors express about the nature of typical statistical issues in data analysis are rather different than what we find in HEP, the greatest being that we *do* often have non-negligible belief that our null hypotheses are valid to a precision much greater than our measurement capability. Regarding the search by ATLAS and CMS leading to the discovery of "a Higgs boson", statistician David van Dyk (2014) has prepared an informative summary of the procedures that we used.

In Sections 2 and 3, I review the paradox, and explain how there may exist three different scales in θ, and that the paradox arises if they have a certain hierarchy that is common in high energy physics. In Section 4, I address the notions common among statisticians that all models are wrong, and that scientists are typically biased against the point null, so that the paradox is irrelevant. In passing, I briefly describe the likelihood-ratio commonly used in HEP as the test statistic. In Section 5, I discuss the difficult issue of how to choose the prior for θ, and in particular the scale τ of the plausible values of θ; to me the so-called objective methods that attempt to use the measuring apparatus to set this scale for hypothesis testing are not yet enlightening for scientific communication. Section 6 briefly describes the completely different approach to simple-vs-composite testing advocated by José Bernardo, which stands apart from the rest of the Bayesian literature. In Section 7, I discuss effect size and the historical usage of confidence intervals in high energy physics to augment the quoted p-value, and how tiny effect sizes can be a window into very high energy physics. Section 8 discusses the choice of Type I error α when adopting the approach of Neyman-Pearson

hypothesis testing, with some comments on the "5σ myth". Finally, in Section 9, I discuss the seemingly universal agreement that a single p-value is (at best) a woefully incomplete summary of the data, and how confidence intervals at various confidence levels can (and do) help the consumer in HEP. I conclude in Section 10.

2 The original "paradox" of Lindley, as corrected by Bartlett

Lindley (1957), with a crucial correction by Bartlett (1957), lays out the paradox in a form that is useful as our starting point. This section also draws on Section 5.0 of Jeffreys (1961) and on Berger and Delampady (1987a). I use (mostly) the notation of the latter, using the statistician's convention of upper case for the random variable and lower case for observed values.

Suppose X having density $f(x|\theta)$ is observed, with θ being an unknown element of the parameter space Θ. It is desired to test H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$. Following the Bayesian approach to hypothesis testing pioneered by Jeffreys (also referred to as Bayesian model selection), we assign prior probabilities π_0 and $\pi_1 = 1 - \pi_0$ to the respective hypotheses. Conditional on H_1 being true, one also has a continuous prior probability density $g(\theta)$ for the unknown parameter. If f is normal with mean θ and known variance σ^2, then with a random sample $\{x_1, x_2, \ldots x_n\}$, we have $\overline{X} \sim N(\theta, \sigma^2/n)$. For conciseness (and eventually to make the point that "n" can be obscure), I define

$$\sigma_{\rm tot} \equiv \sigma/\sqrt{n}. \tag{1}$$

The likelihood is then

$$\mathcal{L}(\theta) = \frac{1}{\sqrt{2\pi}\sigma_{\rm tot}} \exp\left\{-(\overline{x} - \theta)^2/2\sigma_{\rm tot}^2\right\}, \tag{2}$$

with maximum likelihood estimate (MLE) $\hat{\theta} = \overline{x}$, so that the posterior probabilities of the hypotheses are easily calculated:

$$P(H_0|\hat{\theta}) = \frac{1}{A}\,\pi_0\,\frac{1}{\sqrt{2\pi}\sigma_{\rm tot}} \exp\left\{-(\hat{\theta} - \theta_0)^2/2\sigma_{\rm tot}^2\right\} \tag{3}$$

and

$$P(H_1|\hat{\theta}) = \frac{1}{A}\,\pi_1 \int g(\theta)\frac{1}{\sqrt{2\pi}\sigma_{\rm tot}} \exp\left\{-(\hat{\theta} - \theta)^2/2\sigma_{\rm tot}^2\right\} d\theta. \tag{4}$$

Here A is a normalization constant to make the total probability unity, and the integral is over the support of $g(\theta)$.

There will typically be a scale τ that indicates the range of values of θ over which $g(\theta)$ is relatively large. One considers the case

$$\sigma_{\rm tot} \ll \tau, \tag{5}$$

so that $g(\theta)$ varies slowly where the rest of the integrand is non-negligible, and thus the integral is approximately $g(\hat{\theta})$. Then the ratio of posterior odds to prior odds for

H_0 with respect to H_1, i.e., the Bayes factor (BF), is independent of A and π_0, and given by

$$\begin{aligned} \mathrm{BF} &= \frac{P(H_0|\hat{\theta})}{P(H_1|\hat{\theta})} \Big/ \frac{\pi_0}{\pi_1} = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{tot}} g(\hat{\theta})} \exp\left\{-(\hat{\theta}-\theta_0)^2/2\sigma_{\mathrm{tot}}^2\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{tot}} g(\hat{\theta})} \exp(-z^2/2), \end{aligned} \tag{6}$$

where

$$z = (\hat{\theta}-\theta_0)/\sigma_{\mathrm{tot}} = \sqrt{n}(\hat{\theta}-\theta_0)/\sigma \tag{7}$$

is the usual test statistic indicating the departure from the null in units of σ_{tot}. Then the p-value for the two-tailed test considered here is $p = 2(1-\Phi(z))$, where Φ is the cumulative distribution function for the unit normal density. (As discussed below, in HEP typically θ is physically non-negative, and hence we use a one-tailed test, i.e., $p = 1-\Phi(z)$.)

Jeffreys (1961, p. 248) notes that $g(\hat{\theta})$ is independent of n and σ_{tot} goes as $1/\sqrt{n}$, and thus a given cutoff value of the BF does *not* correspond to a fixed value of z. This discrepancy in sample-size scaling of z and p-values compared to Bayes factors (already noted in the first edition in 1939, using a constant g on p. 194) is the core of the "paradox".

In Appendix B, Jeffreys (1961, p. 435) curiously downplays the discrepancy in a sentence that begins by summarizing concisely his objections to testing based on p-values (almost verbatim with p. 360 of the 1939 edition): "In spite of the difference in principle between my tests and those based on the [p-values], and the omission of the latter to give the increase in the critical values for large n, dictated essentially by the fact that in testing a small departure found from a large number of observations we are selecting a value out of a long range and should allow for selection, it appears that there is not much difference in the practical recommendations." He does say, "At large numbers of observations there is a difference", but suggests that this will be rare and that one should suspect and test for internal correlations as the cause.

In contrast, Lindley (1957) emphasized how bad the discrepancy could be, with the example where $g(\theta)$ was taken to be constant over an interval that contains $\hat{\theta}$ as well as the range of θ in which the integrand is non-negligible. For any arbitrarily small p-value (arbitrarily large z) that is traditionally interpreted as evidence *against the null*, there will always exist n for which the BF can be *arbitrarily large in favor of the null*.

Bartlett (1957) quickly noted that Lindley had neglected the length of the interval over which $g(\theta)$ is constant, which should appear in the numerator of the Bayes factor, and which makes the posterior probability of H_0 "much more arbitrary". More generally, the normalization of g always has a scale τ that characterizes the extent in θ of non-negligible g, so that $g(\hat{\theta}) \propto 1/\tau$. Thus there is a factor of τ the numerator of the BF. For example, Berger and Delampady (1987a)) and others consider $g(\theta) \sim N(\theta_0, \tau^2)$, which in the limit of Eqn. 5 leads to

$$\mathrm{BF} = \frac{\tau}{\sigma_{\mathrm{tot}}} \exp(-z^2/2). \tag{8}$$

One reaches the same proportionality in the Lindley/Bartlett example if the length of their interval is τ. The crucial observation is thus that the scaling,

$$\text{BF} \propto \frac{\tau}{\sigma_{\text{tot}}} \exp(-z^2/2), \tag{9}$$

is generic. Of course the details of g will in general lead to a different proportionality constant depending on $g(\hat{\theta})$.

Meanwhile, from Eqn. 2, the ratio λ of the likelihood of θ_0 under H_0 and the maximum likelihood under H_1 is

$$\lambda = \mathcal{L}(\theta_0)/\mathcal{L}(\hat{\theta}) \tag{10}$$

$$= \exp\left\{(\hat{\theta} - \theta_0)^2/2\sigma_{\text{tot}}^2\right\} \Big/ \exp\left\{(\hat{\theta} - \hat{\theta})^2/2\sigma_{\text{tot}}^2\right\} \tag{11}$$

$$= \exp(-z^2/2) \tag{12}$$

$$\propto \left(\frac{\sigma_{\text{tot}}}{\tau}\right) \text{BF}. \tag{13}$$

Thus, *unlike* the case of simple-versus-simple hypotheses discussed below in Section 2.2, this maximum likelihood ratio takes the side of the p-value in disfavoring the null for large z, independent of σ_{tot}/τ, and thus independent of sample size n. This difference between maximizing $\mathcal{L}(\theta)$ under H_1, and averaging it under H_1 weighted by the prior $g(\theta)$, can be dramatic.

From the derivation and the scaling equations, we also see that the paradox does not depend on the value of π_0 chosen; in particular it does not depend on the commonly suggested choice of $\pi_0 = 1/2$. The paradox follows from assigning any non-zero probability mass to the point of measure zero, $\theta = \theta_0$.

The factor σ_{tot}/τ, arising from the average of $\mathcal{L}$ weighted by g in Eqn. 4, is often touted as the "Ockham factor" that provides a desirable "Ockham's razor" effect (Jaynes, 2003, Chapter 20) by penalizing H_1 for lack of precision in specification of θ. But the fact that, even asymptotically, the Bayes factor is itself directly dependent on the scale τ of the prior $g(\theta)$ (and more precisely on $g(\hat{\theta})$) can come as a surprise to those deeply steeped in Bayesian estimation, where typically the dependence on all priors diminishes asymptotically. The surprise is perhaps enhanced since Bayes factors are often introduced as the factor by which (even subjective) prior odds are modified in light of the observed data, giving the initial impression that the subjective part has been factorized out from the BF.

The situation clearly invites robustness studies, and various authors, beginning with Edwards et al (1963), have explored in detail the effect of varying $g(\theta)$, making numerical comparisons of p-values to Bayes factors in various contexts such as testing a point null for a binomial parameter. Generally they give examples where the p-values are always numerically smaller than the Bayes factors, even when the prior for θ "gives the utmost generosity to the alternative hypothesis".

2.1 Is there really a "paradox"?

The trivial "resolution" of Jeffreys-Lindley paradox is to point out that there is no reason to expect the numerical results of frequentist and Bayesian hypothesis testing

to agree, since they calculate different quantities. Still, it is a bit unnerving to many that "hypothesis tests" that are nominally both communicating the same scientific result can have such a large discrepancy. So is it a *paradox*?

I prefer to use the word "paradox" with the meaning I first learned in my schoolboy days, namely a *seeming* contradiction that upon closer inspection is *not a contradiction*. This is the meaning of the word, for example, in the celebrated "paradoxes" of Special Relativity, such as the Twin Paradox and the Pole-in-Barn Paradox. (The "resolution" of a paradox is then a careful explanation of why it is not a contradiction.) I thus do *not* use the word paradox as a synonym for contradiction – that takes a word with (I think) a very useful meaning and wastes it on a redundant meaning of another word. It can however be confusing that what is paradoxical by my preferred definition depends on whether or not something "seems" contradictory, which depends on the person. Thus, if someone says, "What Lindley called a paradox is not a paradox", then typically they either define paradox as a synonym for contradiction, or it was always so obvious to them that the paradox is not a contradiction that they think it is not paradoxical. (They could also mean that it *is* a contradiction that cannot be resolved, with my preferred definition of paradox, but I have not seen that used as an argument for why it is not a paradox.) Although it may still be questionable as to whether there is a resolution satisfactory to everyone, for now I think that the word paradox is quite apt, just as it is for the Twin Paradox in Special Relativity. As the deep issue is the scaling of the BF with sample size (for fixed p-value) as pointed out by Jeffreys already in 1939, I follow some others in calling it the Jeffreys-Lindley paradox.

2.2 This paradox is *not* about testing simple vs simple

Testing simple H_0: $\theta = \theta_0$ versus *simple* H_1: $\theta = \theta_1$ provides another interesting contrast between Bayesian and frequentist testing, but this is *not* the case of the Jeffreys-Lindley paradox. *In contrast to the Jeffreys-Lindley paradox*, in the simple-versus-simple case, the Bayes factor and the likelihood ratio are the *same* (in the absence of nuisance parameters), and hence in agreement as to which hypothesis the data favor.

In the Jeffreys-Lindley paradox situation, there is a value of θ under H_1 that is *equal* to the MLE $\hat{\theta}$, and which hence has a likelihood no lower than that of θ_0. The extent to which $\hat{\theta}$ was not favored by the prior is encoded in the Ockham factor in Eqn. 13, and thus the BF and the likelihood ratio can disagree on both the magnitude and the direction of the evidence.

Simple-vs-simple tests are far less common in HEP than simple-versus-composite, but have in fact arisen in the last year as the CERN experiments have been performing tests of quantum mechanical properties of the new boson, namely quantum numbers known as spin and parity. Again supposing X having density $f(x|\theta)$ is observed, now one can form *two* well-defined p-values, namely p_0 indicating departures from H_0 in the direction of H_1, and also p_1 indicating departures from H_1 in the direction of H_0. Any physicist will examine both p-values in considering what inference to draw.

That the set of the *two* p-values is "the evidence" has been argued for example

by Thompson (2007, p. 108), and many in HEP may agree. If $\theta_0 < \hat{\theta} < \theta_1$ and $\sigma_{\rm tot} \ll \theta_1 - \theta_0$, then it is conceivable, for example, that H_0 is rejected at 5σ, while if H_1 is taken as the null, it would be rejected at 7σ. A physicist would be well aware of this circumstance and hardly fall into the straw-man trap of implicitly accepting H_1 by "rejecting" H_0. The natural reaction would be to question both hypotheses, i.e., the two-simple-hypothesis model would be questioned.

Senn (2001, pp. 200-201) has further criticism and references regarding the issue of sample-size dependence of p-values in the simple-vs-simple context.

3 Do point nulls exist in principle? In practice?

In the Bayesian literature, there are notably differing attitudes expressed regarding the point null hypothesis $\theta = \theta_0$ assumed above. There is disagreement on both its relevance to one's typical scientific work and on how to view its prior probability.

First, we recall that in the traditional frequentist paradigm, a point null value θ_0 is treated like any other θ. For example, "Kendall and Stuart" and successors (Stuart et al, 1999, p. 175), describe the duality between hypothesis testing and interval estimation via confidence intervals. The hypothesis test for $\theta = \theta_0$ at significance level ("size") α is entirely equivalent to whether or not θ_0 is contained in a confidence interval for θ with confidence level (CL) of one minus α. "Thus there is no need to derive optimal properties separately for tests and intervals: there is a one-to-one correspondence between the problems..."

This direct connection between estimation and testing is decried in much of the Bayesian literature starting with Jeffreys, and the fact that Bayesian hypothesis testing can treat a point null (also called "sharp hypothesis") in a special way is often touted as an advantage. The test is often phrased in the language of model selection: the "smaller" model H_0 is nested in the "larger" model H_1. From this point of view, it seems natural to have one's prior probabilities π_0 and π_1 for the two models. However, viewed from the point of view of putting a prior on the entire space Θ in the larger model, this corresponds to a non-regular prior that has counting measure ("probability mass") on θ_0 and Lebesgue measure (probability *density*) on $\theta \neq \theta_0$. At least one prominent advocate of "objective" priors (Bernardo, quoted below) argues against this feature in "objective" analyses.

As discussed by Casella and Berger (1987a), some of the more disturbing features of the Jeffreys-Lindley paradox are ameliorated (or even "reconciled") if there is no point null and the test is the so-called "one-sided test", namely H_0: $\theta \leq \theta_0$ versus H_1: $\theta > \theta_0$. This is also true for the completely different decision-based approach of Bernardo (Section 6). Given the importance of the issue of probability mass on a point null, I first cite some of the opinions expressed in the statistics literature, and describe our attitudes in HEP in Section 4.

We start with the position of Dennis Lindley (2009), who lauds the "triumph" of Jeffreys's "general method of significance tests, putting a concentration of prior probability on the null—no ignorance here—and evaluating the posterior probability using what we now call Bayes factors." As a strong advocate of the use of subjective

priors representing personal degree of belief, Lindley views the probability mass on the point null as subjective. (In the same Comment, Lindley criticizes Jeffrey's "error" of integrating over the sample space of unobserved data in formulating his eponymous priors used in estimation).

At the other end of the spectrum of Bayesian theorists, we find Lindley's student, José Bernardo (2009), commenting on Robert et al (2009): "Jeffreys intends to obtain a posterior probability for a precise null hypothesis, and, to do this, he is forced to use a mixed prior which puts a lump of probability $p = Pr(H_0)$ on the null, say $H_0 \equiv \theta = \theta_0$ and distributes the rest with a *proper* prior $p(\theta)$ (he mostly chooses $p = 1/2$). This has a very upsetting consequence, usually known as Lindley's paradox: for any fixed prior probability p independent of the sample size n, the procedure will wrongly accept H_0 whenever the likelihood is concentrated around a true parameter value which lies $O(n^{-1/2})$ from H_0. I find it difficult to accept a procedure which is *known* to produce the wrong answer under specific, but not controllable, circumstances." Bernardo goes on to advocate discrepancy measures such as those in his "Reference analysis" approach (Section 6), which "has the nontrivial merit of being able to use for both estimation and hypothesis testing problems a single, unified theory for the derivation of objective 'reference' priors."

When pressed by Commenters on his own proposals (Section 6 below), Bernardo (2011b) does say that "I am sure that there are situations where the scientist is willing to use a prior distribution highly concentrated at a particular region and explore the consequences of this assumption... What I claim is that, even in precise hypothesis testing situations, the scientist is often interested in an analysis which does *not* assume this type of sharp prior knowledge, and that standard reference priors may be used to give an objective Bayesian answer to the question of whether or not a particular parameter value is compatible with the data, without making such an important assumption."

A number of statisticians find point nulls irrelevant to their own work. In the context of an unenthusiastic comment on the Bayesian information criterion (BIC), Gelman and Rubin (1995) say "More generally, realistic prior distributions in social science do not have a mass of probability at zero..." Raftery (1995b) responds that "social scientists are prepared to act *as if* they had prior distributions with point masses at zero... social scientists often entertain the possibility that an effect is *small*".

In commenting on Bernardo (2011b), Christian Robert and Judith Rousseau say, "*Down with point masses!* The requirement that one uses a point mass as a prior when testing for point null hypotheses is always an embarrassment and often a cause of misunderstanding in our classrooms. Rephrasing the decision to pick the simpler model as the result of a larger advantage is thus much more likely to convince our students. What matters in pointwise hypothesis testing is not whether or not $\theta = \theta_0$ holds but what the consequences of a wrong decision are."

A number of comments on the point null are related to another claim, that all models and all point nulls are at best approximations that are wrong at some level. I discuss this point in more detail in Section 4, but include a few quotes here. Edwards et al (1963) say, "... in typical applications, one of the hypotheses—the null

hypothesis—is known by all concerned to be false from the outset," citing others including Berkson (1938). Vardeman (1987) claims, "Competent scientists do not believe their own models or theories, but rather treat them as convenient fictions. A small (or even 0) prior probability that the current theory is true is not just a device to make posterior probabilities as small as p values, it is the way good scientists think!"

Casella and Berger (1987b) object specifically to Jeffreys's use of $\pi_0 = \pi_1 = 1/2$, used in modern papers as well: "Most researchers would not put 50% prior probability on H_0. The purpose of an experiment is often to disprove H_0 and researchers are not performing experiments that they believe, *a priori*, will fail half the time!" Kadane (1987) expresses a similar sentiment: "For the last 15 years or so I have been looking seriously for special cases in which I might have some serious belief in a null hypothesis. I have found only one [testing astrologer]... I do not expect to test a precise hypothesis as a serious statistical calculation."

As discussed below, such statisticians have evidently not been socializing with too many HEP physicists. In fact, in the literature I consulted, I encountered very few statisticians who granted, as did Zellner (2009), that physical laws such as $E = mc^2$ are point null, and "Many other examples of sharp or precise hypotheses can be given and it is incorrect to exclude such hypotheses *a priori* or term them 'unrealistic'..."

As an opinion from a physicist outside HEP, over twenty years ago condensed matter physicist and Nobel Laureate Philip Anderson (1992) argued in our professional magazine, *Physics Today*, for Jeffreys-style hypothesis testing with respect to a claim for evidence for a fifth force of nature. "Let us take the 'fifth force'. If we assume from the outset that there *is* a fifth force, and we need only measure its magnitude, we are assigning the bin with zero range and zero magnitude an infinitesimal probability to begin with. Actually, we should be assigning this bin, which is the null hypothesis we want to test, some *finite a priori* probability—like 1/2—and sharing out the remaining 1/2 among all the other strengths and ranges."

Already in Edwards et al (1963, p. 235) there was a key point related to our situation in HEP: "Bayesians... must remember that the null hypothesis is a hazily defined small region rather than a point." They also emphasized the subjective nature of singling out a point null hypothesis: "At least for Bayesian statisticians, however, no procedure for testing a sharp null hypothesis is likely to be appropriate unless the null hypothesis deserves special initial credence."

That the "point" null can really be a "hazily defined small region" is clear from the derivation in Section 2. The general scaling conclusion of Eqn. 9 remains valid if "small region" means that the region of θ included in H_0 has a scale ϵ_0 such that $\epsilon_0 \ll \sigma_{\rm tot}$. Some modern authors, such as Berger and Delampady (1987a) have explored quantitatively the approximation induced in the BF by non-zero ϵ_0.

3.1 Three scales yield a paradox

We can conclude that the Jeffreys-Lindley paradox will arise if there exist *three scales* in the parameter space Θ:

1. ϵ_0, the scale under H_0,

2. $\sigma_{\rm tot}$, the scale for the total measurement uncertainty, and

3. τ, the scale under H_1;

and if there they have the hierarchy

$$\epsilon_0 \ll \sigma_{\rm tot} \ll \tau. \tag{14}$$

This situation is in fact common in frontier experiments in HEP. We even have cases, for example the mass of the photon, where $\epsilon_0 = 0$, i.e., most of the subjective prior probability is on $\theta = 0$.

As noted for example by Shafer (1982), the source of the precision of $\sigma_{\rm tot}$ does not matter as long as condition in Eqn. 14 is satisfied. The statistics literature tends to focus on the case where $\sigma_{\rm tot}$ arises from a sample size n via Eqn. 1. Then there can be pedantic discussions about whether or not n can really be arbitrarily large, etc. In my view the existence of a regime where the BF goes as $\tau/\sigma_{\rm tot}$ for fixed z is the fundamental characteristic that can lead to the Jeffreys-Lindley paradox, even if this regime does not extend to $\sigma_{\rm tot} \to 0$. As I discuss in Section 4.1, such regimes exist in HEP, and there is not always a well-defined n underlying $\sigma_{\rm tot}$, a point I return to in Sections 4.2 and 5 below in discussing τ. But first we consider the model itself.

4 Are all models wrong? Do we believe our models?

At the heart of our measurement model is typically what is commonly known as a "law of physics". By some historical quirks, the current "laws" of elementary particle physics, which have survived several decades of intense scrutiny with only a few well-specified modifications, are collectively called a "model", namely the Standard Model. In this paper I refer to such a physics law/model, or an alternative potential replacement for it, as a "core physics model". The currently accepted core physics models have parameters, such as masses of the quarks, which with few exceptions have all been measured reasonably precisely (even if tricky to define). In going from the core physics model to the full measurement model describing the probability densities for data (for example momentum spectra of particles emerging from proton-proton collisions at the Large Hadron Collider (LHC) at CERN), there are multiple complications.

Theoretical calculations based on the core physics model can be quite difficult and involve approximations due to truncation of power series, imperfect understanding of the internal structure of colliding protons, and imperfect understanding of the manner in which quarks emerging from the collision recombine into sprays of particles that we measure. The results of the theoretical calculations, with attendant uncertainties, must then be propagated through a simulation of the responses of the huge detectors, which have extremely complex descriptions involving multitudes of calibration constants, adjustments for inefficient detection, mis-identification of particles, etc. Much

of the work in data analysis in HEP involves subsidiary calibration measurements to measure detector responses, to perform checks against data for the theoretical calculations (in regimes where no departures are expected), and to confirm the accuracy of the simulations.

The aphorism "all models are wrong" (Box, 1976) can certainly apply to the detector simulation, where common assumptions of normality or log-normality are at best good approximations. But the key point is that the pure core physics models still exist as testable hypotheses about nature in which it makes sense to talk about point null hypotheses. Typically the alternative to the Standard Model is a more generalized model in which the Standard Model is nested, corresponding to parameters in the alternative model being set to zero, unity, or infinity. It is perfectly sensible to try to understand if some parameter in the core physics model is zero or not, even if one must do so through the smoke of imperfect detector descriptions with many nuisance parameters. Indeed much of what distinguishes the capabilities of experiments (and experimenters) is how well they can do precisely that by understanding the detector response through careful calibration and cross-checks. I believe that this distinction is over-looked in the notion that usually one cannot test a point null hypothesis in a core physics model because the rest of measurement model is imperfectly specified (as suggested in Berger and Delampady (1987a)).

There is a deeper point I would like to make about our core physics models, and that is the difference between the notion of a model being a good "approximation" in the ordinary sense of the word, and the concept of a mathematical limit. The equations of Newtonian physics have been superseded by those of Special and General Relativity, but we can now observe that they are not just approximations that did a pretty good job in predicting (most) planetary orbits; they are the correct *mathematical limits* in a precise sense. The kinematic expressions for momentum, kinetic energy, etc., are the limits of the Special Relativity equations in the limit as the speed goes to zero. That is, if you tell me how much error you are willing to tolerate in the approximation of Newtonian mechanics, I can tell you a speed below which they will be correct within that tolerance. Similarly, Newton's Universal Law of Gravity is the correct mathematical limit of General Relativity in the limit of small gravitational fields and low speeds (conditions that famously were not satisfied to observational precision for the orbit of the planet Mercury).

This limiting behavior can often be viewed from an appropriate power series. For example, we can expand the expression for kinetic energy from Special Relativity, $\sqrt{p^2+m^2}-m$, in powers of p^2/m^2 in the non-relativistic limit where momentum p is much less than mass m. The Newtonian physics expression $p^2/2m$ is the first term in the series, followed by the lowest order relativistic correction term, $p^4/8m^3$. (I use the usual HEP units in which the speed of light is 1, dimensionless.) An analogous, deeper concept arises in the context of effective field theories. An effective field theory in a sense consists of the correct first terms in a power series of inverse powers of some energy scale much higher than the applicable scale of the effective theory.

It is in this sense that physicists believe that the Standard Model, both its parts and the collective whole, is "true". (I am aware of course that there are deep philosophical questions about reality, and that this point of view can be considered "naive",

but this is a point of view that is common among high energy physicists.) We fully expect that the Standard Model is incomplete, in that more forces and quanta need to be added to it, and that the current mathematical entities will become embedded into larger mathematical entities; indeed much of our theoretical and experimental research programs are aimed at uncovering these extensions (while a significant effort is also spent on understanding further the consequences of the known equations). But whatever new physics is added, we also expect that the Standard Model will remain a correct mathematical limit, or a correct effective field theory with a well-defined relationship to the more correct theory.

It may be that on deep inspection the distinction between an ordinary "approximation" and a mathematical limit is not so great, as even crude approximations can perhaps be considered as some kind of limit. Also, the usefulness of our usual power series breaks down in certain important "non-perturbative" regimes. Nonetheless I think that the concepts of limits and effective field theories are helpful in explaining what we mean when we say that we believe our core physics models. It has occurred to me in reading the opinions of some statisticians that an important distinction is the absence of core "laws" in their models. In that case, one would naturally be averse to obsession about exact values of model parameters when the uncertainty in the model itself is already dominant. In high energy physics, we are typically in a different situation.

4.1 Examples of three scales in HEP

Many searches at the frontier of HEP have three scales with a hierarchy as in Eqn. 14. Here I mention a few of my personal favorites.

In the 1980's, I worked on an experiment searching for a particular decay of a particle called the long-lived neutral kaon, the K_L^0. This decay, to a muon and electron, had been previously credibly ruled out down to a branching fraction (probability per kaon decay) of 10^{-8} or so. With newer technology and beams, we proposed to search down to a level of 10^{-12}. The decay was forbidden at this level in the Standard Model, but there was a possibility that the decay occurred (via a diagram where neutrinos change type that was still within an expanded version of the Standard Model) at a much lower level, say 10^{-17} or less; since it was out of reach for us, this was included in our "null".

Thus our search was a "fishing expedition" for beyond-the-Standard-Model (BSM) physics (in this case a new force of nature) with, roughly speaking, $\epsilon_0 < 10^{-17}$ and $\sigma_{\rm tot} \approx 10^{-12}$. Both the scale τ of prior belief and $g(\theta)$ would be hard to define, as our motivation for performing the experiment was the capability to explore the unknown with a potentially huge discovery of a new force. For me personally, π_1 was small (say 1%), and the scale τ was probably close to that of the range we were exploring, 10^{-8}. (We were able to reach $\sigma_{\rm tot} \approx 10^{-11}$ in the first incarnation of the experiment, with unfortunately a null result.)

As perhaps the most extreme example, it is currently of great interest to determine whether or not protons decay, i.e., whether or not the decay rate is exactly zero, as so far seems to be the case experimentally. The experiments are currently probing

values of the average decay rate per proton of 1 decay per 10^{31} to 10^{33} years. This is part of the range of values predicted by certain unified field theories that extend the Standard Model (Wilczek, 2004). As the age of the universe is order 10^{10} years, this is a very small effect size indeed. Thanks to the exponential nature of decays, these experiments are feasible by observing nearly 10^{34} protons (many kilotons of water) for several years, rather than by observing several protons for 10^{34} years. Assigning the three scales is rather arbitrary, but I would say that $\sigma_{\rm tot} \approx 10^{-32}$ and τ initially was perhaps 10^{-28}. Historically the null under the Standard Model was considered to be a point at exactly zero decay rate, until 1976 when 't Hooft pointed out an exotic non-perturbative mechanism in the Standard Model that can cause proton decay on an immeasurably small scale. Hence again, Eqn. 14 applies.

Finally, among the multitude of current searches for BSM physics at CERN to which Eqn. 14 applies, I mention the search for a heavy version of the Z boson (Section 7), a so-called Z′ (pronounced Z-prime). This would be the quantum of a new force of nature that appears somewhat generically in many speculative BSM models; but there is no reliable prediction as to whether the mass or production rate is accessible at the LHC, or many orders of magnitude beyond our capability. For the Z′, ϵ_0 in the current Standard Model is zero; $\sigma_{\rm tot}$ is determined by the LHC beam energies, intensities, and our general-purpose detector's measuring capabilities; the scale τ is again rather arbitrary (as are π_0 and g), but much larger than $\sigma_{\rm tot}$.

A key point from these examples is that *the three scales are to a large extent independent*. There can be a loose connection in that an experiment may be designed with a particular subjective value of τ in mind, which then influences how resources are allocated to obtain a $\sigma_{\rm tot}$ that has a good chance of settling a particular scientific issue, if feasible. But this connection is so tenuous, and often absent in HEP (when an existing general-purpose apparatus is applied to a new measurement), that I think it unwise to attempt to assert a rule of thumb connecting τ to $\sigma_{\rm tot}$. When a subjective value of τ exists, one would seem to be better off declaring it and using it directly. I return to this below in criticizing a common notion among statisticians that somehow one can tie τ to $\sigma_{\rm tot}$ in some "objective" way.

Furthermore, even where I have indicated some sense of scale τ, there is still the arbitrariness in the form of g. Many in HEP think in terms of "orders of magnitude", with an implicit metric that is uniform in the log of the decay rate. (E.g., "the experiment is worth doing if it extends the reach by a factor of 10", or "it is worth taking data for another year if we double the data set".) But it is not at all clear such phrasing really corresponds to belief uniform in that metric.

4.2 Test statistics for computing *p*-values in HEP

There is a long tradition in HEP for using likelihood ratios, for both binned and unbinned data. This was no doubt inspired by frequentist theory such as the Neyman-Pearson Lemma and Wilks's Theorem, sometimes described in the jargon of HEP (James, 1980), and sometimes with more extensive sourcing (Eadie et al, 1971; Baker and Cousins, 1984; James, 2006). When merited, quite detailed likelihood functions (both binned and unbinned) are attempted, often based on Poisson models endemic

to HEP.

Thus, when it comes to computing confidence intervals and regions, the work horse test statistic is usually a likelihood-ratio λ that is either used to find approximate confidence intervals bounded by $-2\Delta \ln \lambda$, or in some irregular cases used in a carefully constructed Neyman-inspired hypothesis test inversion as advocated by Feldman and Cousins (1998). In discovery analyses, typically the distribution of the test statistic under H_0: $\theta = \theta_0$ is determined by simulation of samples using the Monte Carlo method. In many cases, θ is a physically non-negative quantity (such as a Poisson mean or mass) that vanishes under the null, so that $\theta_0 = 0$ and the alternative is H_1: $\theta > 0$. Thus, the observed test statistic and the null distribution are used in a one-tailed test to obtain a p-value, which is then converted to z, the equivalent number of "σ" for a one-tailed Gaussian test,

$$z = \Phi^{-1}(1-p) = \sqrt{2}\,\mathrm{erf}^{-1}(1-2p). \tag{15}$$

For example, $z = 3$ corresponds to a p-value of 1.35×10^{-3}, and $z = 5$ corresponds to a p-value of 2.9×10^{-7}. (For upper confidence limits, p-values are commonly modified to avoid some issues caused by downward fluctuations, but this does not affect the procedure for discovery.)

Nuisance parameters from detector calibration, estimates of background rates, etc., are abundant in these analyses. A large part of the data analysis effort is devoted to understanding and validating the (often complicated) descriptions of the responses of the experimental apparatus that are included in λ. For nuisance parameters, the uncertainties are typically listed as "systematic" in nature, the name that elementary statistics book use for uncertainties that are not reduced with more sampling. However, our systematic uncertainties typically *are* reduced as we take more data, since the subsidiary analyses that calibrate them also benefit from more data.

A typical example is the calibration of the response of the detector to a high-energy photon hitting it (crucial for one of the Higgs boson detection modes). The raw detector response (an optical flash converted to an analog electrical pulse that is digitized) must be converted to energy units. The resulting energy "measurement" has both a smearing due to resolution as well as an offset due to a scale error. We use calibration data and computer simulations to measure both the width and shape of the smearing, as well as to try to set an unbiased scale that has still residual uncertainty. Thus in terms of the simple $N(\theta, \sigma^2_{\mathrm{tot}})$ model discussed throughout this paper, we have the complications that the response shape may not be normal but is measured, the bias on θ may not be zero but is measured, and σ is also measured, with an uncertainty as well. All of the calibrations may change with temperature, position in the detector, radiation damage, etc., and many resources are put into the effort.

Such calibration takes place for all the different types of subdetectors in a HEP experiment, for all the basic types of particles detected (electrons, muons, pions, etc.). Ultimately, with enough data, some systematic uncertainties do often approach some constant value that limits the usefulness (for certain measurements) of further data-taking. In any case, after all this, it may not be clear at all what can be identified as

n if one thinks in term of the "unit measurement" (Section 5) with $\sigma = \sqrt{n}\sigma_{\rm tot}$ that is the basis for some "objective" methods of setting the scale τ.

Once the models for the nuisance parameters are selected, various approaches are used in HEP to "eliminate" them from the likelihood ratio λ (Cousins, 2005). Profiling the nuisances parameters (i.e., re-optimizing the MLEs of the nuisance parameters for each trial value of the parameter of interest) has been in our basic software tools (though not by those names) for decades (James, 1980). The Higgs results at the LHC used profiling, partly because asymptotic formulas for profile likelihoods were generalized (Cowan et al, 2011) and found to be useful. It is also common to integrate out nuisance parameters in λ in a Bayesian fashion (typically using evidence-based priors), usually by simple Monte Carlo integration (while still treating the parameter of interest in a frequentist manner).

In many of our analyses, the result is fairly robust to the treatment of nuisance parameters in the definition λ. For the separate step of obtaining the distribution of λ under the null hypothesis, we can sometimes use asymptotic theory, but when feasible we also perform Monte Carlo simulations of ensembles of experiments. These simulations require sampling using the nuisance parameters that is performed in various frequentist and Bayesian-inspired ways, again typically (though not always) fairly robust to the choice.

To the extent that we integrate out the nuisance parameters, or to the extent that profiling obtains the same results, then our use of λ as a test statistic for a frequentist p-value recalls Bayesian-frequentist hybrids in the statistics literature (Good, 1992, Section 1), including the prior-predictive p-value of Box (1980). Within HEP this mix of paradigms has been both advocated (Cousins and Highland, 1992) and questioned, as it has been in the statistics literature, but found to give reasonable results in a variety of circumstances.

4.3 Are we biased against the null in HEP?

Practitioners in other disciplines are sometimes accused of being biased against accepting nulls, to the point that experiments are set up with an artificial null just to be able to "reject the null". Allegedly the researchers might feel that they need to reject the null in order to publish their results, etc. I do not know to what extent these characterizations might be valid in other fields. But, in HEP it is often the case that we *do* have significant prior belief in both the model and the point null (within ϵ_0), notwithstanding some statisticians' opinions about competent scientists, etc., that I quote in Section 3. In many searches in HEP there is certainly a hope to reject the Standard-Model point null and make a major discovery of beyond-the-Standard-Model (BSM) physics in which the Standard Model is nested. But there is still high (or certainly non-negligible) prior belief in the null. This is especially the case since such a vast number of precise observations have turned out to be so compatible with predictions of the Standard Model.. There have been hundreds of experimental searches for BSM physics that have not rejected the null of the Standard Model.

Fortunately for the careers of practitioners of HEP, in our field we encourage

publishing results that advance exploration of the frontiers even if they do not reject the null. Our literature, including the most prestigious journals, has numerous papers beginning with "Search for..." that report that no significant evidence for the searched-for BSM physics was found. Many of these papers have some use in quantitatively constraining theoretical speculation, and providing some guidance for future searches.

Some statisticians have wrongly assumed that scientists would not conduct so many experiments in which the null is not rejected. On a related point, sometimes θ is a quantity that is physically non-negative (for example a mass) with $\theta_0 = 0$, and we find that about half the experiments result in an unbiased estimate of θ that is in the unphysical negative region. Some statisticians have thus suggested that our measurement model must be wrong. But our explanation is that our null hypotheses have tended to be true, or almost so, in which case an unbiased estimator would indeed have estimates in the unphysical region about half the time. As we have not recently found BSM physics in HEP, one could perhaps question our choices of experiments, but they are largely constrained by resources and by what nature has put there (or not) for us to discover. The huge experiments such as CMS and ATLAS are multiple-purpose experiments that, for any given process, may or may not be close to the ideal apparatus. Within resource constraints and loosely prioritized by speculation about where the BSM physics may be (not always pointing in fruitful directions, of course), we try to look wherever we have some capability, wherever that may be.

The main case in which we place *little* prior belief on the null is an artificial case in which the null hypothesis is the Standard Model with a missing piece! This is the situation in which one is looking for the "first observation" of a phenomenon that *is* predicted by the Standard Model, but hitherto not observed. In that case, we invent a null hypothesis that is everything in the Standard Model *except* the as-yet-unobserved phenomenon. Then the alternative hypothesis is equal to the complete Standard Model (including the searched-for phenomenon), but usually with a twist. One could naturally imagine the alternative as the complete Standard Model along with its precise (non-zero) prediction, $\theta = \theta_1$, for the new observation. Then, for the core physics, this would be a test of simple-vs-simple as in Section 2.2. Instead, the results are usually reported in two pieces. The simple-vs-composite test is performed, reporting the p-value under the null. In addition, one or more confidence intervals for θ are also reported, which can be then compared to θ_1. This allows for more flexibility in interpretation, including rejection of the null but with a surprising value of $\hat{\theta}$ that points to an alternative other than θ_1.

An example a few years ago at Fermilab was the search for production of single top quarks via the weak force in proton-antiproton collisions, a search made well after the weak force was well characterized, and well after pairs of top quarks had been discovered via their production by the strong force. The search for single top production was experimentally difficult, and the precise effect size could have been different than expectation, even indicating BSM physics. But I do not think that anyone gave much credence to the technical null hypothesis that was used to analyze the data and eventually rejected at more than 5σ, namely that single top production

did not exist at all.

Another, more recent example, is the search for a particular decay of certain particles containing bottom and strange quarks. The Standard Model prediction is that a few out of 10^9 decays yield two muons (heavy versions of electrons) as decay products. This difficult, delicate measurement had significant potential for discovering BSM physics that might enhance (or even reduce) the probability for this decay. The experimental search used the null hypothesis that the decay to two muons had zero probability, a null that was only recently rejected at the 5σ level. As with single top, the real physics interest was in the measured confidence interval(s), as there was negligible prior belief in the artificial null of exactly zero probability for the decay. Of course, a prerequisite for measuring the effect sizes was high confidence that these processes exist, so the observation at high significance by each of two experiments was one of the highlights of this year's results from the LHC.

As the Higgs boson is an integral part of the Standard Model, the null hypothesis used in the search for it was similarly taken to be an artificial model that had all of the Standard Model except the Higgs boson, with no BSM physics to take the place of the Higgs boson. Analogous to the previous two examples, the alternative was the complete Standard Model with a composite θ for the strength of the Higgs boson signal. However, the mass of the Higgs boson is a free parameter in the Standard Model that had been only partially constrained by various prior measurements and theoretical arguments. This complicated the search significantly, as the probabilities of various decay modes of the Higgs boson vary dramatically as a function of the mass. Unlike the examples of single top production and the rare decay to two muons just described, the null hypothesis of "no Higgs boson" probably carried some prior belief in physicists' minds, not so much in the artificial way it was posed, but in the sense that it was certainly possible that some BSM physics would be found, rather than the Standard Model's minimalist Higgs boson. (In fact, this was the hope of many.)

By July 4, 2012, this null was definitively rejected, so that observation of a new boson was announced by both ATLAS and CMS. The confidence intervals for θ (in various sub-classes) were in fairly encouraging agreement with predictions for the Standard Model Higgs boson, but not with great precision. Thus, a lot of the focus shifted to effect sizes of all sorts describing different production and decay mechanisms. For these measurements of effect sizes, the null has reverted back to the Standard Model Higgs boson and the tests use the frequentist duality between interval estimation and testing: one constructs confidence intervals and regions for parameters controlling various distributions, and checks whether or not the predicted values for the Standard Model Higgs boson are within the confidence regions.

5 What sets the scale τ?

The source of the scale τ (the range of values of θ over which the prior $g(\theta)$ is relatively large) is a significant issue, as discussed by Jeffreys (1961, p. 251) and re-emphasized by Bartlett (1957). Fundamentally it would seem to be personal and subjective, as

is the more detailed specification of $g(\theta)$. Berger and Delampady (1987a) state that "the precise null testing situation is a prime example in which objective procedures do not exist." They note that τ has a "dramatic effect" on the BF and posterior probability of H_0, and "furthermore, letting $\tau^2 \to \infty$ so that g becomes 'noninformative' is ridiculous, since then $P(H_0|x) \to 1$. Thus, a Bayesian must, at a minimum, subjectively specify τ^2, and there is no default value that 'lets the data speak for itself' ". In their Rejoinder to comments, they emphasize again, "Testing a precise hypothesis is a situation in which there is *clearly* no objective Bayesian analysis and, by implication, no sensible objective analysis whatsoever." (Berger and Delampady, 1987b).

These comments notwithstanding, Berger and others have attempted to formulate principles for specifying some default value of τ for scientific communication. The notion seems to be that, even if τ is fundamentally subjective, maybe there is some value that is useful for communicating scientific results (and "vastly superior to automatic use of p-values"), even if it should not be used for real decision-making.

I was rather startled to see that Bartlett (1957) suggests that τ might scale as $1/\sqrt{n}$, thus canceling the sample-size scaling in $\sigma_{\rm tot}$ and making the Bayes factor independent of n. David Cox (2006, p. 106) suggests this as well, on the grounds that "... in most, if not all, specific applications in which a test of such a hypothesis $[\theta = \theta_0]$ is thought worth doing, the only serious possibilities needing consideration are that either the null hypothesis is (very nearly) true or that some alternative within a range fairly close to θ_0 is true." This avoids the situation that he finds unrealistic, in which "the corresponding answer depends explicitly on n because, typically unrealistically, large portions of prior probability are in regions remote from the null hypothesis relative to the information in the data."

Although Andrews (1994) also explores the consequences of τ shrinking with sample size, I am not aware of a trend to follow this approach, even though part of Cox's argument was already given by Jeffreys (1961, p. 251), "... the mere fact that it has been suggested that $[\theta]$ is zero corresponds to some presumption that $[\theta]$ is small." Leamer (1978, p. 114) makes a similar point, "... a prior that allocates positive probability to subspaces of the parameter space but is otherwise diffuse represents a peculiar and unlikely blend of knowledge and ignorance". (As Section 4.1 discusses, this "peculiar and unlikely blend" is common in HEP.) Robert (1993) considered π_1 that increased with τ, but this seems not to have been pursued further.

Most of the attempts at a default τ that I have seen in the Bayesian literature lead to a scale τ (and certainly π_0) that does *not* depend on n, and hence does not remove the sample-size dependence of the Ockham factor. In the desperate search for any non-subjective sample-size-independent scale that even exists, the only option readily at hand is $\sigma = \sqrt{n}\sigma_{\rm tot}$, i.e., the scale of the measurement uncertainty when $n = 1$. This was suggested by Jeffreys (1961, p. 268), on the grounds that there is nothing else in the problem to set the scale, and followed for example in generalizations by Zellner and Siow (1980).

Kass and Wasserman (1995) do the same, which "has the interpretation of 'the amount of information in the prior on ψ is equal to the amount of information about ψ contained in one observation' ". They refer to this as "unit information priors".

citing Smith and Spiegelhalter (1980) as also using this "appealing interpretation of the prior."

Raftery (1995a, pp. 132, 135) also takes a prior for which, "roughly speaking, the prior distribution contains the same amount of information as would, on average, one observation". He is among the few to note the obvious problem in practice: the "important ambiguity... the definition of $[n]$, the sample size." He gives several examples for which he has a recommendation.

Thus far, I do not understand why this "unit information" approach is "appealing", or how it could lead to useful, universally cross-calibrated Bayes factors in HEP. As discussed in Section 4.2 our detector may also have some intrinsic $\sigma_{\rm tot}$ for which there is no obviously sensible n to consider.

Berger and Pericchi (2001, with commentary) review more general possibilities based on use of the information in a small subset of the data, in particular various versions of "intrinsic Bayes factors" (IBF) that use priors generated in a bootstrap fashion from either subsets of the data or simulated data. They claim, "... the IBF can also be thought of as the long sought device for generation of good conventional priors for model selection in nested scenarios". They recommend the median intrinsic Bayes factor (MIBF) "for those who desire at least *one* simple default model selection tool," but later say that for nested models, the arithmetic intrinsic Bayes factor (AIBF) is preferred. "Note that this is the first general approach to the construction of conventional priors in nested models." Berger (2008, 2011) applied an intrinsic prior to a pedagogical example and its generalization from high energy physics. Unfortunately, I am not aware of anyone in HEP who has pursued these suggestions. Meanwhile, recently Bayarri, Berger, Forte, and Garca-Donato (2012) have re-considered the issue and formulated principles resulting "... in a new model selection objective prior with a number of compelling properties."

5.1 Comments on non-subjective priors for estimation and model selection

For *estimation*, Jeffreys (1961) has two very different approaches for obtaining a prior for a physically non-negative quantity such as the magnitude of the charge q of the electron. Both involve invariance concepts. The *first* approach (pp. 120-123) involves *thinking about the parameter* being measured. In this example, one person might think that the charge is the fundamental parameter, while another might think that the charge-squared (or some other power) is the fundamental quantity. Faced with arbitrariness in the power m of q, everyone will arrive at the same posterior density if they each take the prior to be $1/q^m$ with their personal choice for m, since all expressions $d(q^m)/q^m)$ differ only by a proportionality constant. (Equivalently, they all take the prior to be uniform in $\ln q^m$,i.e., uniform in $\ln q$).

Jeffreys's *second* approach, much better known, and leading to his eponymous Rule and "Jeffreys's priors", is based on the likelihood function and some averages over the sample space. Statisticians say it is based on "the model". But as an experimenter, one day I realized that what they meant is that "Jeffreys's prior" is

derived *not* by thinking about the parameter being measured, but rather by *thinking about the measuring apparatus.* At first (or even second) sight this might appear strange. The Jeffreys prior for a Gaussian (normal) measurement apparatus is uniform in the measured value. So taking this approach, if the measuring apparatus has Gaussian response in q, the prior is uniform in q. If the measuring apparatus has Gaussian response in q^2, then the prior is uniform in q^2. If the physical parameter is measured with Gaussian resolution and is physically non-negative, as in this case, then the functional form of the prior remains the same (uniform) and is set to zero in the unphysical region (Berger, 1985, p. 89).

Berger and Bernardo refer to the "non-subjective" priors such as Jeffreys's prior as "objective" priors. To me, this is rather like referring to "non-cubical" volumes as "spherical" volumes, which is to say, one is giving a new meaning to the word. Bernardo (2011b) defends the use of the word as follows. "No statistical analysis is really objective, since both the experimental design and the model assumed have very strong subjective inputs. However, frequentist procedures are often branded as 'objective' just because their conclusions are only conditional on the model assumed and the data obtained. Bayesian methods where the prior function is directly derived from the assumed model are objective is this limited, but precise sense."

Whether or not one accepts this explanation, there are many claims for the practical usefulness *for estimation* of so-called "objective" priors. And there seems to be a deep (frequentist) reason for their potential appeal: Because the priors are derived by using knowledge of the properties of the *measuring apparatus*, it is at least conceivable that Bayesian credible intervals based on them might have better-than-typical frequentist coverage properties when they are interpreted as approximate frequentist confidence intervals. As Welch and Peers (1963) showed, for Jeffreys's priors this is indeed the case for one parameter; under suitable regularity conditions, the approximate coverage of the resulting Bayesian credible intervals is uniquely good to order $1/n$, compared to the slower convergence, good only to order $1/\sqrt{n}$, for other priors. So except at very small n, by using "objective" priors, one can (at least approximately) obey the Likelihood Principle and get decent frequentist coverage, which for some is a preferred "compromise". Reasonable coverage is also claimed to be the experience for Reference Priors with more than one parameter. This all works, in spite of the fact that the objective priors are improper for many prototype problems, because the ill-defined normalization constant cancels out in the calculation of the posterior. (Or equivalently, if a cut-off parameter is introduced to make the prior proper, the dependence on the cut-off parameter vanishes as the cut-off increases without bound.)

For model selection, Jeffreys proposed a *third* approach to priors. As discussed above, from the point of view of the larger model, the prior is irregular, having probability mass (a Dirac delta function to physicists) on the null value θ_0 that has measure zero. For $g(\theta)$ on the rest of Θ, for the Gaussian measurement model Jeffreys argued for a Cauchy density ("Lorentzian" to atomic physicists and "Breit-Wigner" to nuclear and high energy physicists).

Apart from the many subtleties that led Jeffreys to choose the Cauchy form for g, there is the major issue of the scale τ of g, as discussed in Section 5. Here again, the

assumption of the objective Bayesians is that, basically by definition, the only "objective" τ is one that is derived from the measuring apparatus. And then, *under the assumption that $\sigma^2_{\rm tot}$ comes from n measurements with an apparatus having variance σ^2, as in Eqn. 1*, they invoke σ as the scale of the prior g.

At this point, I have some sympathy with Lindley's repeated criticisms (e.g., in commenting on Bernardo (2011b)) that objective Bayesians can get lost in the Greek letters and lose contact with the actual context. After arguing that the Ockham's factor is a crucial feature of Bayesian logic that is absent from frequentist reasoning, I find it remarkable that this factor would be chosen based on the measurement apparatus, and on a concept of sample size n that can be very difficult to define. The textbook by Lee (2004, p. 130) appears to agree that this is without compelling foundation: "Although it seems reasonable that $[\tau]$ should be chosen proportional to $[\sigma]$, there does not seem to be any convincing argument for choosing this to have any particular value...".

In order for the concept of "objective" choice of τ to be useful in scientific communication, it seems to me that some features need to be demonstrated as to how it truly provides for useful cross-calibration across experiments with different $\sigma_{\rm tot}$ when n is not well-defined. Otherwise I would agree with Jim Berger's first instinct (or at least the first half of it): "Testing a precise hypothesis is a situation in which there is *clearly* no objective Bayesian analysis and, by implication, no sensible objective analysis whatsoever" (Berger and Delampady, 1987b).

Another voice emphasizing the practical problem is Robert Kass (2009), saying that Bayes factors for hypothesis testing "remain sensitive—to first order—to the choice of the prior on the parameter being tested." The results are thus "contaminated by a constant that does not go away asymptotically." Thus he says that this approach is "essentially nonexistent" in neuroscience.

6 The reference analysis approach of José Bernardo

Among the well-known Bayesian statisticians whose papers I have tried to understand, José Bernardo has a singularly different point of view. A proper discussion of his approach would require (at least) a full paper devoted to it, so here I just try to capture the flavor and the contrast with other Bayesians. Bernardo (1999) (with critical discussion by Lindley and others) defines hypothesis testing in terms very different from calculating the posterior probability of H_0: $\theta = \theta_0$. Rather, he proposes to judge whether or not H_0 is *compatible* (his italics) with the data.

"Any Bayesian solution to the problem posed will obviously require a prior distribution $p(\theta)$ over Θ, and the result may well be very sensitive to the particular choice of such prior; note that, in principle, there is no reason to assume that the prior should necessarily be concentrated around a particular θ_0; indeed, for a judgement on the compatibility of a particular parameter value with the observed data to be useful for scientific communication, this should only depend on the assumed model and the observed data, and this requires some form of non-subjective prior specification for θ which could be argued to be 'neutral'; a sharply concentrated prior around a par-

ticular θ_0 would hardly qualify." He later continues, "In this paper, it is argued that nested hypothesis testing problems are better described as specific decision problems about the choice of a useful model and that, when formulated within the framework of decision theory, they do have a natural, fully Bayesian, coherent solution."

Unlike Jeffreys, Bernardo advocates using the *same* non-subjective priors (even when improper) for testing as for estimation. He defines a discrepancy measure d whose scaling properties can be complicated for small n, but which asymptotically can be much more akin to those of p-values than to those of Bayes factors. In fact, if the posterior becomes asymptotically normal, then d approaches $(1+z^2)/2$ (Bernardo, 2011a,b). Thus, a fixed cutoff for his d (which he considers to be the objective approach), like a fixed cutoff for z, is inconsistent in that it does not accept H_0 when it is true with probability 1 as the sample size increases without bound.

Bernardo and Rueda (2002) elaborate this approach further, emphasizing that the Bayes factor approach, when viewed from Bernardo's decision theory framework, corresponds to a "zero-one" loss-difference function, which they refer to as "simplistic". They prefer continuous loss functions (such as quadratic loss) that do not require the use of non-regular priors. A sharply spiked prior on θ_0 "*assumes* important prior knowledge ... *very strong* prior beliefs," and hence "Bayes factors should *not* be used to test the *compatibility* of the data with H_0, for they inextricably combine what the data have to say with (typically subjective) *strong* beliefs about the value of θ." This contrasts strongly with the common notion, following Jeffreys (1961, p. 246) that, "To say that we have no information initially as to whether the new parameter is needed or not we must take" $\pi_0 = \pi_1 = 1/2$. Bernardo and Rueda reiterate Bernardo's recommendation, mentioned above, for applying the discrepancy measure, (expressed in "natural" units of information) according to "an *absolute* scale which is independent of the problem considered".

Bernardo (2011b) gives a major review, also with extensive commentary, referring unapprovingly to point nulls in an "objective" framework: "However, since the pioneering book by Jeffreys (1961), Bayesian methods have often made use of two *radically different* types of priors, some for estimation and some for hypothesis testing. We argue that this is certainly not necessary, and probably not convenient, and describe a particular form of doing this within the framework of Bayesian decision theory." He clarifies his view of testing, that it is a decision whether or not "to act *as if* H_0 were true", based on the expected posterior loss of using the simpler model rather than the alternative (full model) in which it is nested, which "is true by assumption". (There are a number of subtleties that I did not quite follow, but I think that these quotes capture the flavor.)

In his rejoinder, he states that the Jeffreys-Lindley paradox "clearly poses a very serious problem to Bayes factors, in that, under certain conditions, they may lead to misleading answers. Whether you call this a paradox or a disagreement, the fact that the Bayes factor for the null may be arbitrarily large for sufficiently large n, *however relatively unlikely the data may be under* H_0 is, to say the least, deeply disturbing... the Bayes factor analysis may be completely misleading, in that it would suggest *accepting* the null, even if the likelihood ratio for the m.l.e. *against* the null is very large." This is quite a notable statement on the Ockham factor in Eqn. 13.

At a recent PhyStat workshop, Bernardo (2011a) summarized this approach for an HEP audience. High energy physicist Luc Demortier (2011) also discussed this approach to testing, considering it appropriate when point null is just a convenient simplification if the loss in using it is low, rather than a point having significant prior probability. He noted (as did Bernardo) that the formalism can in fact allow for point nulls if the analyst so desires.

7 Effect size in HEP

Practitioners in other disciplines are sometimes criticized for focusing on p-values to the neglect of effect size, i.e., the point estimate $\hat{\theta}$ (and associated interval estimates). High energy physicists are practically always quite mindful of effect sizes, though we use other nomenclature for them. It is intrinsic to HEP and its precursors (atomic and nuclear physics) to compare quantitatively the predictions of theory (based on quantum mechanics and field theory) to experimental results. Point estimates and confidence intervals for parameters in the models are the basic results of most experiments. For experiments in which one beam scatters off of (or interacts with) a fixed target or other beam, the parametric meeting point for comparison of theory and experiment is frequently a probability of interaction normalized in a conventional way and called a "cross section". (The name is based on an analogy that can be made between geometrical sizes of objects and quantum-mechanical probabilities of interacting. Cross sections are measured in units of area; a low cross section expresses a low probability of interacting, as if the particle were small and thus hard to hit.)

For particles that are produced in interactions and then later observed to decay, the parametric meeting point for comparison of theory and experiment is typically the decay rate $\Gamma = -(dN(t)/dt)/N(t)$, where $N(t)$ is the number of particles not yet decayed at time t after creation. For many processes, Γ is a constant unique to that process (given by "Fermi's Golden Rule" of quantum mechanics), so that $N(t)$ decays exponentially with mean lifetime $1/\Gamma$, i.e., $N(t) = N(t{=}0)\exp(-\Gamma t)$. Both cross section measurements and lifetime measurements are subdivided into various subprocesses, as functions of both continuous parameters (such as angles) and discrete parameters (such as the probabilities known as "branching fractions" for decay into differing sets of decay products). Other type of parameters in the theory are also measured, for example masses of particles.

In the example of the Higgs-like boson discovery, the effect size was quantified with confidence intervals on the production cross section times branching fraction for several sets of decay products. These confidence intervals were exciting indications that the new boson was indeed Higgs-like, as described in the highly publicized discovery talks (Incandela and Gianotti, July 4, 2012) and the subsequent ATLAS and CMS publications (Aad et al, 2012; Chatrchyan et al, 2012). However, the precision at that time was rather limited, with both ATLAS and CMS concluding that more data were needed to determine more precisely the nature of the new boson. By spring 2013, more data had been analyzed and it seemed clear to both collaborations that the boson was at least "a" Higgs boson. Some of the key figures are reproduced and

described in the information accompanying the announcement of the recent Nobel Prize in Physics (Swedish Academy, 2013, e.g., Figures 6 and 7).

7.1 No effect size is too small in core models of HEP

Some of the literature in other disciplines makes the point that one must distinguish between mathematical statistical significance and practical significance: if there is 5σ evidence for an extremely small departure from the null, then that may have little practical significance. Furthermore, since "all models are wrong", a tiny effect on a parameter in a Gaussian model in psychology, which is conditional on the model being true, is likely to be properly disregarded as uninteresting. In contrast, our core models in physics are what are colloquially known as "laws of physics". It is big news if they can be shown to be wrong at any level.

In HEP, tests of our core physics models also benefit from what we believe to be the world's most perfect random-sampling mechanism: quantum mechanics. In each of many repetitions of a given initial state, nature randomly picks out a final state according to the weights given by the (true, not completely known) laws of physics and quantum mechanics. Furthermore, the most perfect incarnation of "identical" is achieved through the fundamental quantum mechanical property that elementary particles of the same type are literally *indistinguishable*. Thus the underlying model is nearly always akin to Bernoulli trials and their generalizations and approximations, quite frequently the Poisson distribution.

The nature of frontier physics research is to make inferences about the laws of physics from observations in certain domains (speeds, temperatures, densities, strength of fields, numbers of particles, etc.) that are limited by our capabilities at a given time; and then to extrapolate and test these laws outside the domains in which they are first formulated. The most direct tests of the extrapolations are, of course, to extend the experimentally accessible domains. However, the first hints that the extrapolations might fail (signaling "new physics") can come from very precise measurements in the original limited domains. As described in the next section, in high energy physics, we can push this approach further, and use the nature of quantum mechanics to glimpse the effect of very massive particles well before the technology exists to make them.

7.2 In HEP, smaller effect size can correspond to new particles at higher energy

In the modern view, for every force there is a quantum field that permeates all space. As suggested in 1905 by Einstein for the electromagnetic (EM) field, associated with every such field is an "energy quantum" (called the photon for the EM field) that is absorbed or emitted ("exchanged") by other particles interacting via that field. While (as noted above) the mass of the photon is presumed to be exactly zero, the masses of the quanta of some other fields are non-zero. The nominal mass m, the energy E, and momentum p, are related by Einstein's classical equation, $m^2 = E^2 - p^2$.

(For unstable particles, exactly what I mean by "nominal mass" becomes somewhat technical, but there are agreed-on conventions.)

Much of high energy physics is possible because energy quanta can be exchanged even when the energy ΔE and momentum Δp being transferred in the interaction do not correspond to the nominal mass of the quantum being exchanged. With quantity q^2 (unrelated to symbol for the charge q of the electron) defined in a process by $q^2 = (\Delta E)^2 - (\Delta p)^2$, quantum mechanics reduces the probability of the reaction as q^2 departs from m^2 of the exchanged particle. In many processes, the reduction factor is at leading order proportional to

$$\frac{1}{(m^2 - q^2)^2}. \tag{16}$$

(As q^2 can be negative, the relative sign of q^2 and m^2 depends on details of the process. The singularity of $m^2 = q^2$ is softened to be finite by higher-order corrections.) What q^2 is accessible depends on the technology available; in general, higher q^2 requires higher-energy particle beams and thus more costly accelerators.

For the photon, $m = 0$ and the interaction probability goes as $1/q^4$. On the other hand, if the mass m of the quantum of a force is very much higher than the q^2 attainable with existing technology, the probability for an interaction to occur due to the exchange of such a quantum is not zero, but proportional to $1/m^4$. Thus, by looking for interactions or decays having *very* low probability, we are probing the existence of *very* high-mass quanta, beyond the energies directly attainable with concurrent technology.

An illustrative example, studied by the historian of science Peter Galison (1983), is the accumulation of evidence for the existence of the Z boson (with mass m_Z), an electrically neutral quantum of the weak force hypothesized in the 1960's. Difficult experiments were performed in the late 1960's and early 1970's using intense beams of neutrinos scattering off targets of ordinary matter. The available q^2 was much less than m_Z^2, resulting in a small reaction probability in the presence of other processes obscuring the signal. CERN staked the initial claim (Hasert et al, 1973), and after a period of confusion, both experimental teams agreed that they had observed interactions mediated by Z bosons even though no Z bosons were directly detected, as the energies involved (and hence $|q|$) were well below m_Z).

In a second type of experiment probing the Z boson, conducted at SLAC in the late 1970's (Prescott et al, 1978), specially prepared electrons were scattered off nuclei to look for a very subtle left-right asymmetry in the scattered electrons due to the combined action of the electromagnetic and weak forces. In an exquisite experiment widely praised both for its conception and its execution, an asymmetry of about 1 part in 10^4 was measured to about 10% statistical precision with an estimated systematic uncertainty also about 10%. The experiment was essentially Bernoulli trials with the ability to measure departures from unity of twice the binomial parameter with an uncertainty of about 10^{-5}. I.e., the sample size of scattered electrons was of order 10^{10}. This is a precision in a binomial parameter finer than that in an ESP example that has already generated a lively discussion in the statistics literature on the Jeffreys-Lindley paradox Bernardo (2011b, pp. 19, 26, and cited references, and comments

and rejoinder). More recent experiments measure this scattering asymmetry even more precisely. The results of Prescott et al. confirmed predictions of the model of electroweak interactions put forward by Glashow, Weinberg, and Salam, clearing the way for these three to receive the Nobel Prize in 1979.

Finally, in 1982, the technology for creating interactions with $q^2 = m_Z^2$ was attained at CERN (and later at Fermilab). And in 1989, "Z factories" turned on at SLAC and CERN, colliding electrons and positrons at beam energies tuned to $q^2 = M_Z^2$. At this q^2, the singularity in Eqn. 16 causes the *tiny* effect size in the previous experiments to become a *huge* bump in a plot of cross section, a factor of 1000 increase in scattering cross section compared to the null hypothesis of "no Z boson". (The instability of the Z boson to decay leads to a finite peak height.)

This sequence of events in the experimental pursuit of the Z boson is somewhat of a prototype for what many of us hope will happen multiple times in the future of high energy physics. A given process (scattering or decay) has probability zero (or immeasurably small ϵ_0) according to the Standard Model. If, however, there is a new boson X with mass m_X much higher than accessible with current technology, then the boson may give a non-zero probability, proportional to $1/m_X^4$, for the given process. The null hypothesis is that X does not exist and the probability for the process is immeasurably small. As m_X is unknown, the possible probabilities for the process if X *does* exist are a continuum, including probabilities arbitrarily close to zero. But these tiny numbers in the continuum map onto possibilities for real, discrete, modifications to the laws of nature – new forces!

The searches for rare decays described in Section 4.1 are examples of this approach. In the example of rare decays of the K_L^0, a non-zero effect at the 10^{-11} level would have indicated a new mass scale about a 1000 times greater than the mass of the Z boson, more than a factor of 10 above currently accessible q^2 even with the LHC. The observation of proton decay with a decay rate at the level probed by current experiments would spectacularly indicate a new mass scale over 10^{13} times greater than the mass of the Z boson, i.e., new force-carrying particles having a mass over 10^{15} times greater than the mass of the proton.

Alas, none of these latter searches has observed the searched-for decays that would constitute BSM physics. In the intervening years, there have however been major discoveries in neutrino physics that redefined and extended the Standard Model. These discoveries established that the mass of the neutrino, while tiny, is not equal to zero. A reasonable inference from these discoveries is that there is a new mass scale, very high and perhaps approaching the scale probed by proton decay (Wilczek, 2004).

As another example of the incredible precision that is sometimes possible in HEP, I mention that the difference in the mass of the K_L^0 and the mass of a closely related particle, the short-lived neutral kaon (K_S^0) has been measured. In the units we use for these masses (MeV), the mass of one of these particles has been measured to be 497.614 ± 0.024 MeV (about half the mass of a proton), already an impressively precise absolute measurement for particles that have a mean lifetime of a few billionths of a second. But incredibly, the *difference* in mass between the K_L^0 and K_S^0 has been measured to be $(3.484 \pm 0.006) \times 10^{-12}$ MeV (Particle Data Group et al, 2012), i.e., a

part in 10^{14} relative to their masses. This mass difference is due to high-order terms in the weak interaction, and is in agreement with the Standard Model prediction. It is extremely sensitive to some classes of BSM physics speculation, which it thus tightly constrains.

Finally, I mention one measurement currently published with a tantalizing discrepancy from the Standard Model. A muon (unstable heavy version of an electron) is a tiny magnet whose strength is predicted by the Standard Model, but can be modified in various BSM physics scenarios. In the natural metric, the strength is about a part per mil larger than 1 unit, more precisely 1.001659. But the theoretical prediction and experimental results don't stop at that precision. Currently there is a discrepancy between theory and experiment at the level of 2.87×10^{-9}, with an estimated uncertainty of 0.8×10^{-9} (Miller et al, 2012). This discrepancy, greater than 3σ, is much discussed; only time will tell if it holds up as a potential major discovery of BSM physics.

8 Neyman-Pearson testing and the choice of Type I error probability α

In HEP, our confidence intervals typically have conventional confidence levels (68%, 95%, etc.), so by the duality with hypothesis tests mentioned in Section 3, whether or not θ_0 is in a confidence interval corresponds to the dual test. Thus if experimenters report a p-value, consumers can each invoke the Neyman-Pearson (N-P) accept/reject paradigm by comparing the p-value to one's own unique (pre-test) value of α. From a mathematical point of view, one can *define* the post-data p-value as the smallest significance level α at which the null hypothesis would be rejected, had that α been specified in advance. (Rice, 2007, p. 335). This offends some adherents of Fisher who point out that Fisher did not think about it this way when he introduced the term, but these protests do not negate the trivially true mathematical identity with Fisher's p-value, even though the differing interpretations should be kept distinct.

In any case, regardless of the steps by which one learns whether or not the test statistic λ is in the rejection region for a particular value of θ, much has been written about how to choose its size α, the Type I error probability of rejecting H_0 when it is true. N-P introduced the alternate hypothesis H_1 and the Type II error β, the probability under H_1 that H_0 is not rejected when it is false. As they noted (Neyman and Pearson, 1933a, p. 296) "These two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second. ... The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator."

Lehmann and Romero (2005, p. 57, and earlier editions by Lehmann) echo this point in terms of the *power* of the test, defined as $1-\beta$: "The choice of a level of significance α is usually somewhat arbitrary... the choice should also take in consideration the power that the test will achieve against the alternatives of interest..."

For the case of simple vs simple discussed in Section 2.2, such considerations can be

well-defined since the power $1-\beta$ is well-defined. Neyman and Pearson (1933b, p. 497) discuss how one might balance the two types of errors, for example by considering their total. It is well-known today that such an approach, including minimizing a weighted sum, can go a long way toward removing some of the most unpleasant aspects of fixed-α testing, such as inconsistency. But Neyman and Pearson (1933b, p. 496) realized of course that this solution becomes ill-defined for a test of simple vs composite when the composite hypothesis has values of θ arbitrarily close to θ_0, since the limiting value of β is 0.5, independent of α. Robert (2013) echoes these concerns: "In the Neyman–Pearson referential, there is a fundamental difficulty in finding a proper balance (or imbalance) between Type I and Type II errors, since such balance is not provided by the theory, which settles for the sub-optimal selection of a *fixed* Type I error. In addition, the whole notion of *power*, while central to this referential, has arguable foundations in that this is a *function* that inevitably depends on the unknown parameter θ. In particular, the power decreases to the Type I error at the boundary between the null and the alternative hypotheses in the parameter set."

Unless one picks out a value of θ among those in the composite hypothesis as being of special enough interest to use it for power considerations, there is no well-defined procedure. A Bayesian, of course, will in effect perform the optimization by weighting the values of θ under H_1 by the prior $g(\theta)$. As Raftery (1995a, p. 142) put it, "Bayes factors can be viewed as a precise way of implementing the advice of [Neyman and Pearson (1933a)] that power and significance be balanced when setting the significance level... there is a conflict between Bayes factors and significance testing at predetermined levels such as .05 or .01." Remarkably, Neyman and Pearson (1933b, p. 502) suggest this possibility if multiple θ_i under the alternative are genuinely sampled from probabilities Φ_i: "...if the Φ_i's were known, a test of greater resultant power could almost certainly be found."

"Kendall and Stuart" and successors (Stuart, Ord, and Arnold, 1999, Section 20.29) view the choice of α in terms of costs: "...unless we have supplemental information in the form of the *costs* (in money or other common terms) of the two types of error, and costs of observations, we cannot obtain an optimal combination of α, β, and n for any given problem." But of course a Bayesian (or good scientist) will also insist that prior belief must play a role, and Lehmann and Romero (2005, p. 58) (and earlier editions by Lehmann) agree: "Another consideration that may enter into the specification of a significance level is the attitude toward the hypothesis before the experiment is performed. If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low."

Of course, none of these vague statements about choosing α comes close to formal decision theory, which is however not visibly practiced in HEP. For the case of simple vs composite relevant to the Jeffreys-Lindley paradox, I think that HEP physicists do informally take into account prior belief, the effect size (confidence interval), and relative costs of errors, contrary to myths about 5σ. But as I discuss in Section 4.2, the concept of sample size can be ill-defined in some of our measurements where we essentially just have the net $\sigma_{\rm tot}$.

8.1 5σ mythology

Nowadays one often reads that 5σ is the criterion for a discovery in high energy physics. Notwithstanding that a fixed one-size-fits-all level of significance violates one of the most basic tenets of science — that the more extraordinary the claim, the more extraordinary must be the evidence — I suspect that some people in the field may take the fixed threshold more seriously than it merits.

The (quite sensible) historical roots of the 5σ were in a specific context, namely searches in the 1960's for new "elementary particles", now known to be composite particles with different configurations of quarks. A plethora of histograms were made, and presumed new particles known as "resonances" showed up as localized excesses ("bumps") spanning a few histogram bins. Upon finding an excess and defining those bins to be the "signal region", one could estimate what is now called the "local p-value". First one uses nearby bins in the histogram ("sidebands") to formulate the null hypothesis corresponding to the expected number of events in the signal region in the absence of a new particle. Then one could calculate the probability of seeing a bump as large as that seen, or larger, under the null hypothesis, and express the result in terms of "σ" by analogy to a one-sided test of a normal model. The problem was that the location of the resonance was typically not known in advance, so that the local significance did not include the fact that "pure chance" has lots of opportunities (lots of histograms and their bins) to have an unlikely occurrence.

Over time many of the alleged new resonances were not confirmed in repeat experiments. In the group led by Luis Alvarez at Berkeley, "5σ" became a useful threshold for predicting which resonances would be confirmed. The story is mentioned in Alvarez's Nobel Prize acceptance speech in 1968. An article by Arthur Rosenfeld (1968, p. 465) describes computer simulations and a rough hand calculation of the number of trials, and concludes, "To the theorist or phenomenologist the moral is simple: wait for nearly 5σ effects. For the experimental group who have spent a year of their time and perhaps a million dollars, the problem is harder... go ahead and publish... but they should realize that any bump less than about 5σ calls only for a repeat of the experiment."

Thus the original concept of "5σ" in HEP was mainly motivated as a (fairly crude) way to account for a multiple trials factor, known these days in HEP as the "Look Elsewhere Effect". It had at least one other motivation, however, namely that spurious claimed discoveries are sometimes in retrospect attributed to mistakes in modeling the detector or other so-called "systematic effects" that were either unknown or not properly taken into account. Thus "5σ" also builds in a crude robustness against such mistakes.

Unfortunately, as time goes by many practitioners in HEP are unaware of the original motivations for "5σ", and some may apply it without much thought. For example, it is sometimes used as a threshold when a trials-factor correction has already been applied (as in Section 9.1 below), or when there is no trials factor from multiple bins or histograms because one is performing a one-off measurement. (In this case, there is still the ill-posed question of whether to account for the trials in all the other experiments in HEP, or for that matter in all of science.)

Further thoughts on 5σ are in a recent note by Louis Lyons (2013).

9 Can *p*-values be calibrated as a data summary? If augmented by confidence interval(s) for θ?

Generally in HEP we believe that the primary goal is to communicate the method and results of the experiment in a manner that allows the reader to draw his or her own conclusions, supplemented by an interpretation section in the paper. Confidence intervals and p-values are often intended to perform this function. While in some cases providing a more extensive description of the likelihood function, the writer is often implicitly assuming that confidence intervals (sometimes given for more than one confidence level) and p-values (often expressed as equivalent z) are sufficient input into inferences or even decisions to be made by readers.

Thus one can ask (as in the statistics literature) what is the result of taking the p-value as the "observed data" that is then the input for a full (subjective) Bayesian calculation of the posterior probability of H_0. One could even attempt to go further and formulate a decision on whether or not to claim publicly that H_0 is false, using a (highly subjective) loss function describing one's personal costs of falsely declaring a discovery versus waiting and getting scooped in a real discovery. In fact, I think that high energy physicists frequently base decisions on informal attempts to combine observed p-values, prior belief, and the cost/benefit of doing more work before presenting their work.

From Eqn. 9, clearly z alone is not sufficient to recover the Bayes factor and proceed as a Bayesian. This point is repeatedly emphasized in articles already cited. (Even worse is to try to recover the BF using only the binary inputs as to whether or not the p-value was above or below pre-data fixed thresholds (Dickey, 1977; Berger and Mortera, 1991; Johnstone and Lindley, 1995).) The oft-repeated argument (e.g., Raftery (1995a, p. 143)) is that there is no justification for the step in the derivation of the p-value where one replaces "probability density for data as extreme as that observed" with "probability for data as extreme, *or more extreme*". Jeffreys (1961, p. 385) still seems to be unsurpassed in his ironic way of saying this (italics in original), "*What the use of [the p-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.*" Berger and Delampady (1987a) conclude, "...it becomes ridiculous to argue that we can intuitively learn to properly calibrate P-values... First and foremost, when testing precise hypotheses, formal use of P-values should be abandoned," a conclusion of course not unanimously concurred with in the Comments. Hinkley (1997) says it more mildly, "Unfortunately, p-values are not generally comparable from one experiment to another... there is no universal inferential scale to which p-values can be judged... the usual p-value cannot be interpreted fully without reference to the relevant information available..."

Good (1992) found that, "The real objection to [p-values] is not that they usually are utter nonsense, but rather that they can be highly misleading, especially if the

value of $[n]$ is not also taken into account and is large." He suggested a rule of thumb for taking n into account by standardizing the p-value to an effective size of $n = 100$, but this seems not to have attracted a following.

Meanwhile, a confidence interval for θ (as invariably reported in HEP publications, at a minimum for 68% CL and sometimes for additional CLs) *does* give one a good sense of the magnitude of $\sigma_{\rm tot}$ (though this might be misleading in some special cases). And one has one's subjective prior and hence τ. *Thus, at least crudely, it seems that one has the required inputs to recover the result from something like Eqn. 9.* It is perhaps doubtful that a typical physicist would use them to arrive at the same Ockham factor as one who consciously calculated a BF from the original likelihood. On the other hand, a BF based on an arbitrary ("objective") τ does not seem to me an obviously better way to communicate.

While the "5σ" criterion in HEP gets a lot of press (Section 8.1), when a decision needs to be made, I think that physicists intuitively and informally adjust their decision-making based on the confidence interval, their prior belief in H_0 and $g(\theta)$, and on how high the stakes are in getting it right.

Mayo and Spanos (2006) argue that confidence intervals do not solve the problem, and that Mayo's concept of "severe testing" is the key to scientific inference. Spanos (2013) argues this specifically in the context of the Jeffreys-Lindley paradox. I am not aware of widespread application of this approach, and do not yet understand it well enough to see how it would improve scientific communication in HEP if adopted as the standard.

9.1 Trials factors for nuisance parameters *not* eliminated

In HEP, the situation may arise where there is a nuisance parameter ψ that we choose not to eliminate by profiling, marginalization, or other means. Rather, we communicate the results (p-value and confidence interval for θ) as a function of ψ. The search for the Higgs boson was a typical example, in which ψ is the mass of the boson, while θ is the Poisson mean (relative to that expected for a Higgs boson) of any putative excess of events at mass ψ. That is, for each mass, we reported a p-value for the departure from H_0 *as if that mass had been fixed in advance*, as well as a confidence interval for θ, given that ψ. We refer to this p-value as the "local" p-value, the probability for a deviation as extreme as that seen, or larger, at that *particular* mass. (Local p-values are correlated with those of nearby masses within the experimental uncertainties on the mass measurement.)

We then scan all masses in a specified range and find the smallest local p-value, $p_{\rm min}$. Obviously the probability of having a local p-value as small or smaller than $p_{\rm min}$ *anywhere in a specified mass range* is greater than $p_{\rm min}$, by a factor that we refer to as the "Look Elsewhere Effect" (LEE), essentially a multiple trials effect. When feasible, we use Monte Carlo simulations to calculate the p-value that takes the LEE into account, which we refer to as a "global" p-value for the specified mass range. When this is too computationally demanding, we estimate the effect using the method advocated by high energy physicists Gross and Vitells (2010), which is based on that of statistician Davies (1987).

To emphasize that the range of masses used for the LEE is arbitrary or subjective, and to indicate the sensitivity to the range, we try to give the global p-value for at least two ranges of mass. Some obvious possibilities are the range of masses for which the Standard Model Higgs boson has not previously been ruled out at high confidence; or the range of masses for which the experiment is able to search with some reasonable sensitivity for the Standard Model Higgs; or the range of masses for which we have data and could search for any new boson.

10 Conclusions

More than a half century after Lindley drew attention to the difference in sample size scaling between p-values and Bayes factors (already described two decades earlier by Jeffreys), there is still no consensus in the statistics literature on how best to communicate scientific results; and the argument continues internally within the broader Bayesian community on a number of points. While there is a large and always-growing literature criticizing p-values and praising the "logical" approach of Bayes factors, in my opinion much of this literature (especially the secondary literature by scientists) has still not come to terms with the fact that the Ockham factor $\sigma_{\rm tot}/\tau$ is either arbitrary or personal, even (especially) asymptotically.

It has always been important in estimation problems for the analyst to describe the sensitivity of results to choices of prior probability, especially as the dimensionality grows. In testing, sensitivity analysis is clearly mandatory. I am not so concerned with the difference in numerical value of p-values and Bayes factors (or posterior probabilities), as one must commit the error of probability inversion ("error of the third kind") to equate the two. Rather, the issue is whether a summary of the data, with say two or three numbers, can (even in principle) be interpreted by consumers in a manner cross-calibrated across different experiments. The difference in sample-size scaling (or more generally, the difference in scaling with $\sigma_{\rm tot}/\tau$) between the BF and λ is already apparent in Eqn. 13 and hence cannot be entirely blamed on the additional issue of tail probabilities pithily derided by Jeffreys.

For us high energy physicists, I think that it is important to gain a lot more experience with Bayes factors, and also with Bernardo's proposals (which I find quite intriguing). For statisticians, I hope that this discussion of the issues in high energy physics provides "existence proofs" of situations where one cannot ignore the Jeffreys-Lindley paradox, and renews some attempts to improve methods of scientific communication.

As for the Higgs boson discovery, I think that our message was very well discussed and understood internally, and in general well-communicated externally, both to fellow high energy physicists and to the general public. Probably we could have done with less talk about a fixed 5σ threshold and more discussion about why something around that level was useful for the Higgs boson, not so much because the prior on "no Higgs boson" was high, but because there was a potentially large (and avoidable) cost to a premature announcement before we approached that level. We met the ideal scientific standard of the two largely independent "observations" of CMS

and ATLAS (which both inflicted interocular trauma on many of us) before CERN declared a "discovery". That was the right decision in my opinion.

Given our detailed presentations of effect sizes, I am not sure how Bayes factors would have helped, especially given the lack of tradition of Bayes factors in HEP and thus context for interpreting them. (The latter is of course a circular justification that can be repaired in the future. In that respect, a retrospective attempt at Bayes factors would be illuminating and provide a calibrated example.) All of us at the LHC look forward to the next data sets beginning in 2015, when the energy and intensity of the beams will both be increased, and we resume the search for physics beyond the Standard Model. That is one relevant time scale during which new statistical tools can certainly be considered.

Acknowledgments I thank my numerous colleagues in high energy physics and in the CMS experiment in particular for many useful discussions, and members of the CMS Statistics Committee for comments on drafts. The PhyStat series of workshops organized by Louis Lyons has been helpful to many of these discussions and has also brought us enlightening contact with a number of eminent statisticians, including several cited herein. This material is based upon work partially supported by the U.S. Department of Energy under Award Number de-sc0009937.

References

Aad G, et al (2012) Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Physics Letters B 716(1):1–29, DOI 10.1016/j.physletb.2012.08.020

Anderson PW (1992) The Reverend Thomas Bayes, needles in haystacks, and the fifth force. Physics Today 45(1):9–11, DOI 10.1063/1.2809482, URL `http://link.aip.org/link/?PTO/45/9/1`

Andrews DWK (1994) The large sample correspondence between classical hypothesis tests and Bayesian posterior odds tests. Econometrica 62(5):1207–1232, URL `http://www.jstor.org/stable/2951513`

Baker S, Cousins RD (1984) Clarification of the use of chi-square and likelihood functions in fits to histograms. Nucl Instrum Meth 221:437–442, DOI 10.1016/0167-5087(84)90016-4

Bartlett MS (1957) A comment on D. V. Lindley's statistical paradox. Biometrika 44(3/4):533–534, URL `http://www.jstor.org/stable/2332888`

Bayarri MJ, Berger JO, Forte A, Garca-Donato G (2012) Criteria for Bayesian model choice with application to variable selection. The Annals of Statistics 40(3):1550–1577, URL `http://www.jstor.org/stable/41713685`

Berger J (2008) A comparison of testing methodologies. In: Prosper H, Lyons L, Roeck AD (eds) Proceedings of PHYSTAT LHC Workshop on Statistical Issues for

LHC Physics, CERN, Geneva, Switzerland, 27-29 June 2007, CERN, CERN-2008-001, pp 8–19, URL http://cds.cern.ch/record/1021125

Berger J (2011) The Bayesian approach to discovery. In: Prosper HB, Lyons L (eds) Proceedings of PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland, 17-20 January 2011, CERN, CERN-2011-006, pp 17–26, URL http://cdsweb.cern.ch/record/1306523

Berger JO (1985) Statistical Decision Theory and Bayesian Analysis, 2nd edn. Springer Series in Statistics, Springer, New York

Berger JO, Delampady M (1987a) Testing precise hypotheses. Statistical Science 2(3):317–335, URL http://www.jstor.org/stable/2245772

Berger JO, Delampady M (1987b) [Testing precise hypotheses]: Rejoinder. Statistical Science 2(3):348–352, URL http://www.jstor.org/stable/2245779

Berger JO, Mortera J (1991) Interpreting the stars in precise hypothesis testing. International Statistical Review / Revue Internationale de Statistique 59(3):337–353, URL http://www.jstor.org/stable/1403691

Berger JO, Pericchi LR (2001) Objective Bayesian methods for model selection: Introduction and comparison. Lecture Notes-Monograph Series 38:135–207, URL http://www.jstor.org/stable/4356165

Berger JO, Sellke T (1987) Testing a point null hypothesis: The irreconcilability of p values and evidence. Journal of the American Statistical Association 82(397):112–122, URL http://www.jstor.org/stable/2289131

Berkson J (1938) Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association 33(203):526–536, URL http://www.jstor.org/stable/2279690

Bernardo JM (1999) Nested hypothesis testing: The Bayesian reference criterion. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian Statistics 6. Proceedings of the Sixth Valencia International Meeting, Oxford U. Press, Oxford, U.K., pp 101–130

Bernardo JM (2009) [Harold Jeffreys's theory of probability revisited]: Comment. Statistical Science 24(2):173–175, URL http://www.jstor.org/stable/25681292

Bernardo JM (2011a) Bayes and discovery: Objective Bayesian hypothesis testing. In: Prosper HB, Lyons L (eds) Proceedings of PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland, 17-20 January 2011, CERN, CERN-2011-006, pp 27–49, URL http://cdsweb.cern.ch/record/1306523

Bernardo JM (2011b) Integrated objective Bayesian estimation and hypothesis testing. In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) Bayesian Statistics 9. Proceedings of the Ninth Valencia International Meeting, Oxford U. Press, Oxford, U.K., pp 1–68, URL http://www.uv.es/bernardo/

Bernardo JM, Rueda R (2002) Bayesian hypothesis testing: A reference approach. International Statistical Review / Revue Internationale de Statistique 70(3):351–372, URL http://www.jstor.org/stable/1403862

Box GEP (1976) Science and statistics. Journal of the American Statistical Association 71(356):791–799, URL http://www.jstor.org/stable/2286841

Box GEP (1980) Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society Series A (General) 143(4):pp. 383–430, URL http://www.jstor.org/stable/2982063

Casella G, Berger RL (1987a) Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association 82(397):106–111, URL http://www.jstor.org/stable/2289130

Casella G, Berger RL (1987b) [Testing precise hypotheses]: Comment. Statistical Science 2(3):344–347, URL http://www.jstor.org/stable/2245777

Chatrchyan S, et al (2012) Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. Physics Letters B 716(1):30 – 61, DOI 10.1016/j.physletb.2012.08.021

Cousins RD (2005) Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature. In: Lyons L, Unel MK (eds) Proceedings of PHYSTAT 05 Statistical Problems in Particle Physics, Astrophysics and Cosmology, Oxford, U.K, September 12-15, 2005, Imperial College Press, pp 75–85, URL http://www.physics.ox.ac.uk/phystat05/proceedings/

Cousins RD, Highland VL (1992) Incorporating systematic uncertainties into an upper limit. Nuclear Instruments and Methods A 320:331–335, DOI 10.1016/0168-9002(92)90794-5

Cowan G, Cranmer K, Gross E, Vitells O (2011) Asymptotic formulae for likelihood-based tests of new physics. Eur Phys J C 71:1554, DOI 10.1140/epjc/s10052-011-1554-0

Cox DR (2006) Principles of Statistical Inference. Cambridge University Press, Cambridge

Davies RB (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 74(1):33–43

Demortier L (2011) Open issues in the wake of Banff 2010. In: Prosper HB, Lyons L (eds) Proceedings of PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland, 17-20 January 2011, CERN, CERN-2011-006, pp 1–11, URL http://cdsweb.cern.ch/record/1306523

Dickey JM (1977) Is the tail area useful as an approximate Bayes factor? Journal of the American Statistical Association 72(357):138–142, DOI 10.1080/01621459.1977.10479922, URL http://www.jstor.org/stable/2286921

Eadie W, et al (1971) Statistical Methods in Experimental Physics, 1st edn. North Holland, Amsterdam

Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. Psychological Review 70(3):193–242

Feldman GJ, Cousins RD (1998) Unified approach to the classical statistical analysis of small signals. Phys Rev D 57:3873–3889, DOI 10.1103/PhysRevD.57.3873, physics/9711021

Galison P (1983) How the first neutral-current experiments ended. Rev Mod Phys 55:477–509, DOI 10.1103/RevModPhys.55.477, URL http://link.aps.org/doi/10.1103/RevModPhys.55.477

Gelman A, Rubin DB (1995) Avoiding model selection in Bayesian social research. Sociological Methodology 25:165–173, URL http://www.jstor.org/stable/271064

Good IJ (1992) The Bayes/non-Bayes compromise: A brief review. Journal of the American Statistical Association 87(419):597–606, URL http://www.jstor.org/stable/2290192

Gross E, Vitells O (2010) Trial factors or the look elsewhere effect in high energy physics. Eur Phys J C 70:525–530, DOI 10.1140/epjc/s10052-010-1470-8

Hasert F, et al (1973) Observation of neutrino-like interactions without muon or electron in the Gargamelle neutrino experiment. Physics Letters B 46(1):138 – 140, DOI 10.1016/0370-2693(73)90499-1

Hinkley D (1997) [Unified frequentist and Bayesian testing of a precise hypothesis]: Comment. Statistical Science 12(3):155–156, URL http://www.jstor.org/stable/2246364

Incandela J, Gianotti F (July 4, 2012) Latest update in the search for the Higgs boson, public seminar at CERN. Video: http://cds.cern.ch/record/1459565; slides: http://indico.cern.ch/conferenceDisplay.py?confId=197461

James F (1980) Interpretation of the shape of the likelihood function around its minimum. Comput Phys Commun 20:29–35, DOI 10.1016/0010-4655(80)90103-4

James F (2006) Statistical Methods in Experimental Physics, 2nd edn. World Scientific, Singapore, sec. 11.6

Jaynes E (2003) Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, U.K.

Jeffreys H (1961) Theory of Probability, 3rd edn. Oxford University Press, Oxford

Johnstone D, Lindley D (1995) Bayesian inference given data 'significant at α': Tests of point hypotheses. Theory and Decision 38(1):51–60, DOI 10.1007/BF01083168

Kadane JB (1987) [Testing precise hypotheses]: Comment. Statistical Science 2(3):347–348, URL http://www.jstor.org/stable/2245778

Kass R (2009) Comment: The importance of Jeffreys's legacy. Statistical Science 24(2):179–182, URL http://www.jstor.org/stable/25681294

Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90(430):773–795, URL http://www.jstor.org/stable/2291091

Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association 90(431):928–934, URL http://www.jstor.org/stable/2291327

Leamer EE (1978) Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley series in probability and mathematical statistics, Wiley, New York

Lee PM (2004) Bayesian Statistics: An Introduction, 3rd edn. Wiley, Chichester U.K.

Lehmann E, Romero JP (2005) Testing Statistical Hypotheses, 3rd edn. Springer, New York

Lindley D (2009) [Harold Jeffreys's theory of probability revisited]: Comment. Statistical Science 24(2):183–184, URL http://www.jstor.org/stable/25681295

Lindley DV (1957) A statistical paradox. Biometrika 44(1/2):187–192, URL http://www.jstor.org/stable/2333251

Lyons L (2013) Discovering the significance of 5 sigma, arXiv:1310.128

Mayo DG, 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):pp. 323–357, URL http://www.jstor.org/stable/3873470

Miller JP, Eduardo de R, Roberts BL, Stöckinger D (2012) Muon $(g-2)$: Experiment and theory. Annual Review of Nuclear and Particle Science 62(1):237–264, DOI 10.1146/annurev-nucl-031312-120340

Neyman J, Pearson ES (1933a) On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character 231:289–337, URL http://www.jstor.org/stable/91247

Neyman J, Pearson ES (1933b) The testing of statistical hypotheses in relation to probabilities a priori. Mathematical Proceedings of the Cambridge Philosophical Society 29:492–510, DOI 10.1017/S030500410001152X

Particle Data Group, Beringer J, et al (2012) Review of particle physics. Phys Rev D 86:010,001, DOI 10.1103/PhysRevD.86.010001, URL http://pdg.lbl.gov/

Prescott C, et al (1978) Parity non-conservation in inelastic electron scattering. Physics Letters B 77(3):347 – 352, DOI 10.1016/0370-2693(78)90722-0

Raftery AE (1995a) Bayesian model selection in social research. Sociological Methodology 25:111–163, URL http://www.jstor.org/stable/271063

Raftery AE (1995b) Rejoinder: Model selection is unavoidable in social research. Sociological Methodology 25:185–195, URL http://www.jstor.org/stable/271066

Rice JA (2007) Mathematical Statistics and Data Analysis, 3rd edn. Thomson, Belmont, CA

Robert CP (1993) A note on Jeffreys-Lindley paradox. Statistica Sinica 3(2):601–608

Robert CP (2013) On the Jeffreys-Lindley paradox, arXiv:1303.5973v2[stat.ME]

Robert CP, Chopin N, Rousseau J (2009) Harold Jeffreys's theory of probability revisited. Statistical Science 24(2):141–172, URL http://www.jstor.org/stable/25681291

Rosenfeld AH (1968) Are there any far-out mesons or baryons? In: Baltay C, Rosenfeld AH (eds) Meson spectroscopy: A collection of articles, W.A. Benjamin, New York, pp 455–483, From the preface: based on reviews presented at the Conference on Meson Spectroscopy, April 26-27, 1968, Philadelphia, PA USA. "...not, however, intended to be the proceedings..."

Senn S (2001) Two cheers for p-values? Journal of Epidemiology and Biostatistics 6(2):193–204

Shafer G (1982) Lindley's paradox. Journal of the American Statistical Association 77(378):325–334, URL http://www.jstor.org/stable/2287244

Smith AFM, Spiegelhalter DJ (1980) Bayes factors and choice criteria for linear models. Journal of the Royal Statistical Society Series B (Methodological) 42(2):213–220, URL http://www.jstor.org/stable/2984964

Spanos A (2013) Who should be afraid of the Jeffreys-Lindley paradox? Philosophy of Science 80(1):73–93, URL http://www.jstor.org/stable/10.1086/668875

Stuart A, Ord K, Arnold S (1999) Kendall's Advanced Theory of Statistics, vol 2A, 6th edn. Arnold, London, and earlier editions by Kendall and Stuart

Swedish Academy (2013) Advanced information: Scientific background: The BEH-mechanism, interactions with short range forces and scalar particles. URL http://www.nobelprize.org/nobel_prizes/physics/laureates/2013/advanced.html

Thompson B (2007) The Nature of Statistical Evidence. Lecture Notes in Statistics, Springer, New York

van Dyk DA (2014) The role of statistics in the discovery of a Higgs boson. Annual Review of Statistics and Its Applications, to appear

Vardeman SB (1987) [Testing a point null hypothesis: The irreconcilability of p values and evidence]: Comment. Journal of the American Statistical Association 82(397):130–131, URL http://www.jstor.org/stable/2289136

Welch BL, Peers HW (1963) On formulae for confidence points based on integrals of weighted likelihoods. Journal of the Royal Statistical Society Series B (Methodological) 25(2):pp. 318–329, URL http://www.jstor.org/stable/2984298

Wilczek F (2004) Nobel lecture: Asymptotic freedom: From paradox to paradigm. URL http://www.nobelprize.org/nobel_prizes/physics/laureates/2004/wilczek-lecture.html, This web page has video, slides, and pdf writeup.

Zellner A (2009) [Harold Jeffreys's theory of probability revisited]: Comment. Statistical Science 24(2):187–190, URL http://www.jstor.org/stable/25681297

Zellner A, Siow A (1980) Posterior odds ratios for selected regression hypotheses. Trabajos de Estadistica Y de Investigacion Operativa 31(1):585–603, DOI 10.1007/BF02888369