

Connections between statistical practice in elementary particle physics and the severity concept as discussed in Mayo's *Statistical Inference as Severe Testing*

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

February 22, 2020

Abstract

For many years, philosopher-of-statistics Deborah Mayo has been advocating the concept of *severe testing* as a key part of hypothesis testing. Her recent book, *Statistical Inference as Severe Testing*, is a comprehensive exposition of her arguments in the context of a historical study of many threads of statistical inference, both frequentist and Bayesian. Her foundational point of view is called *error statistics*, emphasizing frequentist evaluation of the errors called Type I and Type II in the Neyman-Pearson theory of frequentist hypothesis testing. Since the field of elementary particle physics (also known as high energy physics) has strong traditions in frequentist inference, one might expect that something like the severity concept was independently developed in the field. Indeed, I find that, at least operationally (numerically), we high-energy physicists have long interpreted data in ways that map directly onto severity. Whether or not we subscribe to Mayo's philosophical interpretations of severity is a more complicated story that I do not address here.

1 Introduction

In elementary particle physics (also known as high energy physics, HEP), the frequentist sampling properties of nearly every diagnostic tool are routinely studied. This is rooted in long tradition, including influential books, as well as in quantum mechanics, which is viewed by almost everyone in the field as providing the means for obtaining samples of fundamental physics processes that are perfectly independent and identically distributed (even if viewed through the imperfect and not identical capabilities of detectors and transducers). Thus,

*cousins@physics.ucla.edu

frequentist hypothesis tests in the tradition of Neyman and Pearson, as well as post-data p -values and confidence intervals, are routinely reported. Deborah Mayo’s 2018 monograph [1], *Statistical Inference as Severe Testing* (henceforth SIST), is a comprehensive exposition of her arguments (old and new) in favor of frequentist statistics, with the concept of severity supplementing the more talked-about frequentist methods. The book also contains detailed historical studies of many threads of statistical inference, both frequentist and Bayesian. As a high-energy physicist with long-standing interest in both the history and philosophical foundations of statistical inference, I found the book to be a fascinating read, even if I did not linger long enough over some subtle passages in order to fully absorb them. This note is a belated response to a kind invitation to participate in the discussion of SIST in Ref. [2].

The level of frequentist data analysis in HEP is highly developed, and has been improved with a lot of input from statisticians of all flavors in the last twenty years. Thus my main thought in approaching the severity concept was, to be frank, “If it is really useful, we are probably already doing something equivalent.” I have found that to be the case, with the key connection being confidence intervals that have an adjustable post-data critical confidence level, in the same manner that the p -value is a post-data critical value of the significance level α .

As with SIST, I consider separately the two cases of null hypothesis accepted and null hypothesis rejected. For the former, the usual practice in HEP of quoting upper limits (at least implicitly at more than one confidence level) maps directly onto severity concepts. For the latter, it maps as well, with the modification that we typically perform two-tailed tests if the point null hypothesis of the initial one-tailed test is rejected.

In order to make this note readable to a broad audience, I have erred on the side of too much pedagogical introduction, which can largely be skipped over by experts, except to see the notation defined. In order to keep the length at least somewhat acceptable, I focus on HEP’s (very) common case of testing a point null hypothesis versus a continuous alternative, and I emphasize what is typically reported in our publications. In this *descriptive* approach of *frequentist* methods in HEP, I avoid getting very deep into philosophy, and barely mention Bayesian methods.

2 Background about our field and the HEP prototype test

In HEP, the frequentist sampling properties of essentially everything we do are studied using traditional Monte Carlo (MC) simulation methods (less often Markov Chain). We perform simulations to obtain the distributions of each test statistic under the null and alternative hypotheses, and to evaluate frequentist coverage of interval estimates (whether obtained by frequentist or Bayesian-inspired recipes). Asymptotic formulas such as Wilks’s Theorem and its generalizations are also routinely used, though users are admonished to check by MC simulation that the asymptotic approximations are valid.

A lot of effort has gone into understanding how best to deal with nuisance parameters, which are present in all our statistical models, at a minimum to represent residual uncertainties in the calibrations of responses of our instruments. At the LHC, profile likelihood

functions are a common tool for dealing with nuisance parameters, though Bayesian-style marginalization (while treating the parameter of interest by frequentist methods) is also performed. For us, the bottom line is which of these approximate methods gives the best frequentist performance for parameter(s) of interest, in which situations. In this note, I suppress reference to nuisance parameters, but one can assume that they are present (and a nuisance!).

Furthermore, we appreciate well the point emphasized early and often in SIST (pp. 20, 94, 122, and others listed in the index under “auditing”) that the experimental procedure and the statistical model be scrutinized, and systematic effects in particular identified and quantified. As I wrote in Ref. [3], “It is certainly worth trying to understand if some physical parameter in the alternative core physics model is zero (corresponding to the SM), even if it is necessary to do so through the smoke of imperfect detector descriptions with many uninteresting and imperfectly known nuisance parameters. Indeed, much of what distinguishes the capabilities of experimenters is how well they can do precisely that by determining the detector response through careful calibration and cross-checks.” Looking back through history at (in)famous wrong results in HEP, I think that most of them were caused by failure to specify the measurement model (both the measurement apparatus and the procedure) correctly. However, in this note, the focus is on idealized situations where the model is correctly known (and for which there is still plenty of food for thought).

At the discovery frontiers of the field, where one is searching for new elementary particles and/or new forces of nature, the prototype test (idealized for this discussion) contains a parameter μ that characterizes the strength of a hypothetical new phenomenon. This can be the strength of a new force, the production rate of a new particle, or decay rates of known particles to final states that are forbidden for the known forces. In such cases, $\mu = 0$ means the absence of the new phenomenon, while $\mu > 0$ means “new physics”. Typically, there is no firm prediction for a specific value of $\mu > 0$.

It is sometimes useful in this discussion to add the complication that μ is a vector with two components of fundamental physics interest; one component could be the strength just described, while the other component could be a particle’s mass, or some other parameter in the speculative theory, such as a “mixing angle” that represents misalignment of sets of basis vectors in some Hilbert space of quantum field theory. However, unless explicitly mentioned, in this note μ is a single (scalar) parameter.

As usual I let \mathbf{x} (boldface following SIST) denote observable(s); more generally, \mathbf{x} is any convenient or useful function of the observable(s), i.e., a “statistic” or “test statistic”. Following usage in HEP as well as in SIST (p. 34, etc.), we let \mathbf{x}_0 denote a particular observed value (rather than using the statistician’s notation \mathbf{X} and observed \mathbf{x}). As with μ , I sometimes include the case where \mathbf{x} is vector with two or more materially different components. E.g., one component could be an energy and another component could be a position (or a direction). The “the statistical model”, or simply “the model” is then

$$p(\mathbf{x}|\mu), \tag{1}$$

the probability or probability density (from context) characterizing everything that determines the probabilities/densities of the observations, from laws of physics to experimental setup and protocol. (I use the “conditional” vertical line even in a frequentist context where

there is no distribution in μ , as in Ref. [4], rather than a semi-colon. While disfavored in SIST (p. 205), I do not find that it ever causes confusion in practice.) This definition of the model serves our needs, as long as one understands the physicist’s convention that $p(\mathbf{x}|\mu)$ and $p(\mathbf{y}|\mu)$ for some $\mathbf{y} \neq \mathbf{x}$ are *different functions* that would more properly be called $p_X(\mathbf{x}|\mu)$ and $p_Y(\mathbf{y}|\mu)$; in HEP the subscript is nearly always suppressed. I do not try to use the model notation attributed to Fisher (SIST p. 132).

Thus, a large fraction of what we do in HEP can be characterized as testing the “point” null hypothesis

$$H_0: \mu = 0 \tag{2}$$

versus the continuous alternative

$$H_1: \mu > 0. \tag{3}$$

In this note, I refer to this as the *HEP prototype test*.

The more general cases where $\mu < 0$, or where the point null is non-zero (as in Section 9.1.2), also exist. However, I think that most of the discussion can take place with the simpler special case. A key point is that frequently *the model is only defined for $\mu \geq 0$* . For example, if μ is the mean of a Poisson distribution, or a particle mass, or a reaction rate, then $p(\mathbf{x}|\mu)$ *does not exist* for $\mu < 0$: one would not know how to write a Monte Carlo simulation of the experimental setup! On the other hand, there is typically *nothing* wrong with $\mathbf{x} < 0$; it is just a sampled value from the model. This has led to a lot of confusion over the years, when \mathbf{x} is unfortunately called the “measured value of μ ”. We tried to dispel this in Ref. [5], not completely successfully. In any case, the situation in which ones observes $\mathbf{x} < 0$ for $\mu \geq 0$ remains a topic of much discussion [6, 7, 8], with multiple methods employed.

For those who might ask if a zero-width “point” null ever truly exists, I address this question in Ref. [3]. As has been noted in the statistics literature, all that matters is that any small width of the set of μ in H_0 be much narrower than the width of the likelihood function (i.e., the resolution of the measurement apparatus). Examples are plentiful in HEP.

The duality between confidence intervals for μ and frequentist hypothesis tests in the HEP prototype test was implicit in work in our field for many decades, and has been more explicitly advertised by myself and others for some 20 years. That there is *not* a duality between Bayesian credible intervals and the Bayesian treatment of this hypothesis test leads to the Jeffreys–Lindley paradox [3]. Thus, the HEP prototype test of Eqns. 2–3 is a great forum for discussion, with controversies both within the frequentist approach and in comparisons to the usual Bayesian approach.

As this duality is important for my understanding of severity, I belabor my low-level discussion of it here, essentially repeating how I teach it to HEP audiences [9]. This is partly to emphasize the choice of *ordering* that is used to determine which observations \mathbf{x} are *extreme*, and partly to make a point (Section 6.2) about a missing vocabulary word in the duality, relevant to severity. Perhaps my somewhat tedious discussion also lessons the chance for a miscommunication between the language that I use as a physicist and that which statisticians use.

Another point, discussed in Ref. [3] and mentioned in SIST (p. 212), is that many published results in HEP (and Ph.D. theses, and successful faculty promotion cases, including mine) do *not* reject the null hypothesis! (I am only beginning to try to understand how it is possible that this seems not always to be true in some other fields that are called “science”.)

Thus, we have highly developed methods for interpreting such “null results” as quantitative constraints on speculative “new physics”, i.e., on what can be said about $\mu > 0$ when $\mu = 0$ is not rejected. Since this is also what severity is about, it seems not surprising that we have concepts corresponding to it.

3 The basic idea of confidence intervals in two sentences

Given the model $p(\mathbf{x}|\mu)$ and the observed value \mathbf{x}_0 , we ask: For what values of μ is \mathbf{x}_0 an *extreme* value of \mathbf{x} ? Then we include in the confidence interval $[\mu_L, \mu_U]$ those values of μ for which \mathbf{x}_0 is *not* extreme.

That’s the concept! To specify the concept in full detail, we need to define “extreme”. We rank the possible values of \mathbf{x} and choose a particular fraction of highest-ranked values of \mathbf{x} that are *not* considered as extreme. This fraction is of course the *confidence level* (CL), say 68% or 95%. (I ignore any distinction between confidence coefficient and confidence level.) We also define $\alpha = 1 - \text{CL}$, i.e., the fraction that is the lower-ranked, extreme set of values. (This description here also corresponds to the construction of graphical *confidence belts* in the tradition of Neyman; for a given μ , the set of \mathbf{x} values that are not considered extreme for that μ is the *acceptance interval* in \mathbf{x} . I do not need either expression for this discussion, however.)

In order to rank the possible values of \mathbf{x} applicable to each μ , one needs to choose an *ordering principle*. By convention, *high rank means not extreme*. In SIST, a signed ordering $d(\mathbf{x})$ is defined on p. 129 as “a measure of accordance or discordance, fit, or misfit, $d(\mathbf{x})$ between possible answers (hypotheses) and data”. That is fine, but I prefer the more abstract notion of ordering \mathbf{x} values, since depending on the context, discordance may have a crucial arithmetic sign, or might be mapped monotonically to some other function, etc., as in the following examples. *I use the notation $r(\mathbf{x})$ (r for rank) for the ordering function.*

In approaching a discussion of severity for the HEP prototype test, it is useful to consider the *four* different orderings in use in HEP, as follows.

When \mathbf{x} and μ are 1D, and when $p(\mathbf{x}|\mu)$ is such that higher μ implies higher average \mathbf{x} , then we have the three traditional orderings:

Upper limit ordering Order \mathbf{x} from largest to smallest: the smallest values of \mathbf{x} are the most extreme (and so have the smallest $r(\mathbf{x})$). Given \mathbf{x}_0 , the confidence interval that contains μ for which \mathbf{x}_0 is not extreme will typically not contain the largest values of μ . This leads to confidence intervals known in HEP as *upper limits* on μ , which I write here as $\mu < \mu_{UL}$ (for values of μ that exist in the model). (I believe that some statisticians call them upper confidence bounds.)

Lower limit ordering Order \mathbf{x} from smallest to largest, so that the largest values of \mathbf{x} are most extreme (and so have the smallest $r(\mathbf{x})$). This leads to *lower limits* μ_{LL} : $\mu > \mu_{LL}$.

Central ordering Order \mathbf{x} using *central* quantiles of $p(\mathbf{x}|\mu)$, with the quantiles shorter in \mathbf{x} (least integrated probability of \mathbf{x}) containing higher-ranked \mathbf{x} (higher $r(\mathbf{x})$); lower-

ranked \mathbf{x} is added as the central quantile gets longer (contains more integrated probability of \mathbf{x}). This leads to *central* confidence intervals for μ , a special case of more general $[\mu_L, \mu_U]$.

For any CL, the endpoints of central confidence intervals with CL are, as is well-known, the same as upper/lower limits with confidence level given by $1 - (1 - \text{CL})/2 = (1 + \text{CL})/2$.

Both upper limits and central confidence intervals are ubiquitous in HEP. (Lower limits are typically only implicitly considered via hypothesis testing.) There are however long-standing controversies surrounding these three traditional orderings in the HEP prototype test, since the resulting confidence interval can be the empty set (also described by some as a non-empty interval that is “nonphysical”). Historically, this has resulted in various alternatives [8] to the traditional upper limit ordering, including a popular “modification” of the frequentist CL [10] and a Bayesian-inspired approach [11] (Section 39.4.1). Furthermore, since these orderings are defined only when \mathbf{x} is 1D, a more general ordering principle is needed for multi-D \mathbf{x} . In 1998, fellow physicist Gary Feldman and I advocated a *unified* ordering based on a likelihood ratio (LR) statistic, which we belatedly learned was the standard LR test statistic (!), described in “Kendall and Stuart” and successors [4] for some decades before and since:

Likelihood ratio ordering Order \mathbf{x} from high to low values of the likelihood ratio $\Lambda = \mathcal{L}(\mathbf{x}|\mu)/\mathcal{L}(\mathbf{x}|\hat{\mu})$, where $\hat{\mu}$ is that value of μ that maximizes \mathcal{L} . So $r(\mathbf{x})$ is any monotonic order-preserving function of Λ .

Of course, as SIST discusses on pp. 133 ff, this ratio was featured in the classic Neyman-Pearson papers, and has been found to be useful even when the strict criteria of the Neyman-Pearson Lemma (simple versus simple) are not satisfied.

As noted above, a key aspect of the HEP prototype test is that the model is only defined for $\mu \geq 0$. As the maximization of \mathcal{L} must respect this, it is usual that $\hat{\mu} = 0$ for various \mathbf{x} . This can change the unified LR ordering materially from all of the first three orderings.

In the HEP prototype test, this LR ordering “unifies” the previously disjoint sets of upper limits and central confidence intervals into one set of two-sided intervals that are no longer central, and *never the empty set*. The lower endpoint of the two-sided interval can be zero, in which case one can view the upper endpoint as an alternative to the traditional upper limit. The intervals with LR ordering also resolve another issue where one chooses between upper limits and central intervals based on the observed data (thus invalidating frequentist coverage). (SIST mentions a preference for two-sided intervals on page 358, but in a different context, the so-called two-sided test, where both signs of μ exist in the model.)

In summary, after observing \mathbf{x}_0 sampled from the model $p(\mathbf{x}|\mu)$, the confidence interval $[\mu_L, \mu_U]$ contains those values of μ for which \mathbf{x}_0 is *not* “extreme” at the chosen CL, *given the chosen ordering of \mathbf{x}* . E.g., at 68% CL, $[\mu_L, \mu_U]$ contains those μ for which \mathbf{x}_0 is in the highest-ranked (least extreme) 68% values of \mathbf{x} for each respective μ , according to probabilities obtained from the model $p(\mathbf{x}|\mu)$.

4 Frequentist hypothesis tests

As in my lectures in HEP [9], we set aside intervals for the moment and consider testing from first principles, essentially following Neyman and Pearson. For the null hypothesis H_0 , we order possible observations \mathbf{x} from least extreme to most extreme, using an ordering principle (which can depend on H_1 as well, as in the LR ordering). We choose a cutoff α (smallish number) called the *size* or *significance level* of the test.

We then “reject” H_0 if the observed \mathbf{x}_0 is in the *most extreme* fraction α of observations \mathbf{x} (generated under H_0). By construction:

α = probability (with \mathbf{x} generated according to H_0) of rejecting H_0 when it is true, i.e., false discovery claim (Type I error).

To quantify the performance of this test if H_1 is true, we further define:

β = probability (with \mathbf{x} generated according to H_1) of not rejecting (“accepting”) H_0 when it is false, i.e., not claiming a discovery when there is one (Type II error). The *power* of the test is defined as $1 - \beta$. The power depends on which value of μ in H_1 is considered.

In HEP, the tradeoff between α and β is studied a lot (essentially ROC curves) in intermediate steps of the analysis, for example in algorithms that identify the particle types discussed in Section 9.1.2. For pre-data characterization and comparison of the performance of high-level hypothesis-testing algorithms, typically $\beta = 0.5$ is used as a benchmark for evaluating α . I do not recall explicit use of power concepts being used in post-data analysis of our high-level hypothesis tests. What we use is more closely identified with the severity concept, as discussed below. (SIST distinguishes severity from various pre-data and post-data uses of power-related concepts (e.g., p. 343), but I have not studied these concepts.)

5 Nested hypothesis testing: Duality with intervals

As SIST emphasizes, in the frequentist formalism (but not the Bayesian formalism), the theory of these tests maps to that of confidence intervals! The way that I teach it to students is as follows. In this section, I do not require that the point null hypothesis be $\mu_0 = 0$, but rather a more general value μ_0 , so we have $H_0: \mu = \mu_0$.

1. Having observed data \mathbf{x}_0 , suppose the 90% CL confidence interval for μ is $[\mu_L, \mu_U]$. This contains all values of μ for which the observed \mathbf{x}_0 is ranked in the *least* extreme 90% of possible outcomes \mathbf{x} according to $p(\mathbf{x}|\mu)$ and the ordering principle in use.
2. With the same data \mathbf{x}_0 , suppose that we wish to test H_0 versus H_1 at Type I error probability $\alpha = 10\%$. We reject H_0 if \mathbf{x}_0 is ranked in the *most* extreme 10% of \mathbf{x} according to $p(\mathbf{x}|\mu)$ and the ordering principle in use.

Comparing the two procedures, we see that we reject H_0 at $\alpha = 10\%$ if and only if μ_0 is outside the 90% CL confidence interval $[\mu_L, \mu_U]$.

(In this verbal description, I am implicitly assuming that \mathbf{x} is continuous and that $p(\mathbf{x}|\mu)$ is a pdf that puts zero probability on a point \mathbf{x} with measure zero. Thus, I ignore any issues concerning endpoints of intervals.)

We conclude: *Given an ordering:* a test of H_0 versus H_1 at significance level α is equivalent to asking: Is μ_0 outside the confidence interval for μ with $CL = 1 - \alpha$?

As Kendall and Stuart put it, “There is thus no need to derive optimum properties separately for tests and for intervals; there is a one-to-one correspondence between the problems as in the dictionary in Table 20.1” [4] (p. 175). The table mentioned maps the terminology that historically developed separately for intervals and for testing, e.g.,

- $\alpha \leftrightarrow 1 - CL$
- Most powerful \leftrightarrow Uniformly most accurate
- Equal-tailed tests \leftrightarrow central confidence intervals

SIST (pp. 191-193) refers to a similar discussion in Lehmann and Romano’s treatise, with an example. Statisticians refer to this duality as “inverting a test” to obtain confidence intervals, and vice versa. Here I refer to the two sides of the duality as the *interval picture* and the *testing picture*. It is perhaps interesting that, while the duality is always there, in some contexts in HEP we tend to focus almost exclusively on (think in terms of) the interval picture, and in some other contexts the testing picture.

6 Post-data p -values and Z -values

The above N-P theory is all a *pre-data* characterization of the hypothesis test. A deep issue is how to apply it after \mathbf{x}_0 is known, i.e., *post-data*.

In N-P theory, α is *specified in advance*. Suppose after obtaining data, you notice that with $\alpha = 0.05$ previously specified, you reject H_0 , but with $\alpha = 0.01$ previously specified, you accept H_0 . In fact, you determine that with the data set in hand, H_0 would be rejected for $\alpha \geq 0.023$.

This interesting value has a name: *After* data are obtained, the *p-value* is the smallest value of α for which H_0 would be rejected, *had that value been specified in advance*. (This is also discussed on SIST p. 175.)

This is numerically (if not philosophically) the same as the definition used e.g. by Fisher and often taught: “The p -value is the probability under H_0 of obtaining \mathbf{x} as extreme *or more extreme* than the observed \mathbf{x}_0 .” I find the first definition to be more helpful in discussing severity.

In HEP, a p -value is typically converted to a Z -value (unfortunately commonly called “the significance S ” in HEP), which is the equivalent number of Gaussian (normal) standard deviations. E.g., for a one-tailed test in a search for an *excess*, $p\text{-value} = 2.87 \times 10^{-7}$ corresponds to $Z = 5$. Note that Gaussianity (normality) of the test statistic is typically *not* assumed when the p -value is computed; this conversion to equivalent Gaussian “number of sigma” is just for perceived ease of communication. This needs to be emphasized when communicating outside HEP, as I hear too often statisticians wondering about assumptions of normality, ironically indicating that our conversion is counter-productive in terms of clarity.

6.1 Interpreting p -values and Z -values

In the example above, it is crucial to realize that that value of α equal to the p -value (0.023 in the example) was typically *not* specified in advance. So p -values do *not* correspond to Type I error probabilities of experiments reporting them. Thus, the interpretation of p -values is a long, contentious story.

Whatever they are, p -values are *not* the probability that H_0 is true! This misinterpretation of p -values is unfortunately so common as to be used as an argument against frequentist statistics. I ask everyone in HEP to please keep in mind:

- That p -values are calculated *assuming that H_0 is true*, so they can hardly tell you the probability that H_0 is true!
- That the calculation of the “probability that H_0 is true” requires prior(s) to invert the conditional probabilities.

Regarding loose language about the interpretation of p -values, I am at least an associate member of the “ p -value police” (SIST p. 204) in that I think that one should always be very clear about what p -values *are* (easy to say, but not so easy to interpret), and what they *are not* (also easy to say correctly, so why be loose and risk encouraging confusion?).

6.2 A missing vocabulary word in the statistics literature?!

What seems to be *lacking* in the statistics literature is a phrase in the confidence interval picture that is dual to “ p -value” in the testing picture! I.e., we need a name for that *critical value of the confidence level* for which the point-null hypothesis value μ_0 is just on the edge of the confidence interval or region. For this note, I decided to call it the “ p -CL”. Then clearly, since $CL = 1 - \alpha$, we have

$$\text{“}p\text{-CL”} = 1 - \text{“}p\text{-value”}. \quad (4)$$

For me, as trivial as this definition is, having a name for it makes it easier to think about severity.

The old concept of *confidence distribution* is useful as a *technical device* for reading off the p -CL. (I have not tried to understand the potential of confidence distributions as claimed by Singh and others, as mentioned on p. 391.) In 1D, the (post-data) confidence distribution [12] is the set of all confidence intervals at different CL. (See also SIST p. 195.) I.e., one can make the post-data graph of the endpoints of the confidence interval as a function of CL (or provide the equivalent information, as in the likelihood scan in the third example of Section 9.1.2). Given μ_0 , one can then simply look at the confidence distribution to find p -CL, i.e., that CL for which μ_0 is an endpoint. (This is purely a technical comment, quite apart from any philosophical interpretations that confidence distributions may have.)

As Cox noted [12], “In applications it will often be enough to specify the confidence distribution, by for example a pair of intervals, and this corresponds to the common practice of quoting say both the 95 per cent and the 99 per cent confidence intervals.” Indeed, we see papers providing confidence intervals for a couple CL’s in examples from HEP in Section 9.1. We can have more complicated examples where the confidence interval or region is not simply

connected (as in Section 9.1.1), due to non-linearities (such as sine functions) in the model. However, the concept of confidence distribution as a technical device for determining p -CL is still workable, even if awkward to display.

6.3 Example of p -CL from HEP

Most papers in HEP use conventional values of CL, so that one needs to use the information given to infer a precise value of p -CL for some μ of particular interest. An exemplary case that I remember is from early attempts to measure a quantity known as $\sin 2\beta$, deeply related to differences between matter and antimatter in the equations of physics. The statistical model $p(\mathbf{x}|\sin 2\beta)$ is only defined for $-1 \leq \sin 2\beta \leq 1$ (but for *all* \mathbf{x}). The bounded parameter space made this a natural case for the LR ordering of \mathbf{x} for obtaining two-sided confidence intervals in $\sin 2\beta$. For the CDF collaboration’s first attempt at Fermilab [13], \mathbf{x} was a sample from a normal distribution centered on the unknown true value of $\sin 2\beta$ with $\sigma = 0.42$. The obtained value was $\mathbf{x}_0 = 0.79$. As $\sin 2\beta = 0$ is a special value in the physics equations (matter–antimatter symmetry), they calculated that value of CL for which 0 was an endpoint of the confidence interval, and obtained 93% for what I call p -CL. This of course corresponds to a p -value of 7% in the LR ordering. (Later experiments were much more precise.)

7 Sketch of top-level data analysis procedure in HEP

Finally, my long tutorial-level introduction is over, so we can look at long-standing practice in HEP in this section, and connect it to the severity concept in Sections 8 and 9.

In a search corresponding to the HEP prototype test, the p -value for the null hypothesis will typically be reported (often via the corresponding Z -value). In the interval picture, this is (as in Section 6.2), $1 - p$ -CL for the *lower limit* for μ (!), though I think that it is rare for practitioners to think about it that way: for testing the null hypothesis, the testing picture mentality is typical.

If the p -value is small enough to give some credence to the possibility of a non-zero value of μ in nature, then generally a *central* confidence interval is also reported. To statisticians, this is the so-called *effect size* in original units [3], much discussed in social science and medical literature. As described in SIST (p. 211), this can then be used to test a theoretical prediction, if one exists (as was the case for the Higgs boson).

If the p -value is not small (and sometimes even if it is), *upper* limit(s) for μ under the alternative hypothesis H_1 are also reported, for some conventional CL(s) (90% in some subfields of HEP, 95% at the LHC). Many papers also convey enough information so that the upper limit at other CL’s can be approximately inferred. (As mentioned in Section 3, it is common to replace the strict frequentist upper limits with one of the alternatives mentioned. But this does not affect the conceptual connection to severity.)

Thus, when reporting these three results of the analysis, the *ordering* (ranking) of \mathbf{x} in each is different (in fact opposite for upper and lower limits). Sometimes papers (and too many internal drafts) will say (paraphrasing) “We have no evidence for a signal [for $\mu > 0$], and so we calculate an upper limit”. As mentioned above, Ref. [5] criticized such a procedure, which we dubbed “flip-flopping” (using the *observed data* to choose the ordering

for testing/intervals). We advocated the *unified* LR ordering as mentioned, a single ordering replacing all three orderings. The LR ordering can of course be consistently used for both intervals and testing, per the duality, for a variety of CL’s.

It is easy to show that the p -value obtained for testing H_0 is the same whether one uses LR ordering or the traditional ordering. I.e., in the interval picture, given $\mu_0 = 0$, the p -CL for lower limits is the same as that for unified LR intervals. For tests of other μ , the p -CL’s and upper limits from the two orderings are different. In certain subfields of HEP (for example neutrino physics), results using the unified LR ordering are routinely included in reported experimental results, even if the traditional orderings are quoted as well.

From an operational point of view, in the first three traditional orderings of Section 3, the identification of the null and alternative hypotheses changes from ordering to ordering. Physicists at the LHC had to understand this in order to write general-purpose software with a guide how to use it. In the HEP prototype test, for computing the p -value the software treats $\mu = 0$ as the null hypothesis, as discussed. For computing the upper limit, the software in effect steps through the various positive values of μ one at a time, and for each value considers it to be a null hypothesis for a one-sided test in the opposite direction as that used to test $\mu = 0$.

Thus, I was interested to see that SIST (p. 346) cites Senn as saying what I think is the same thing: “Senn gives another way to view the severity assessment of $\delta > \delta'$, namely ‘adopt [$\delta = \delta'$] as a null hypothesis and then turn the significance test machinery on it’ ”. I did not quite follow why SIST then says that, “the error statistician is loath to advocate modifying the null hypothesis”, and in the next section, “We are not changing the original null and alternative hypotheses!”. I think that these may be philosophical statements, but operationally the calculation (computer code) is as if that change in null hypothesis is being made.

A related point is the discussion of Hoenig and Heisey (p. 357). The argument seems to hinge on whether one is “allowed” to switch from lower limits to upper limits. In HEP, the usual procedure certainly does so. (Lower limits are nearly always considered implicitly via the dual testing picture, while upper limits are conveyed in the interval picture.)

Whether speaking of two-sided intervals or upper limits, in HEP we tend to talk about “exclusion” of values of μ that are outside the confidence interval, at the quoted CL, e.g., one might say, “Values of μ above μ_{UL} are excluded at 95% C.L.”

To prepare a bit more for the comparison to severity, let us focus on the interval picture, and let $r(\mathbf{x})$ represent one of the four *orderings* for \mathbf{x} described in Section 3. Post-data, after observing \mathbf{x}_0 and not rejecting H_0 , the statement that “ $\mu < \mu_{UL}$ at 95% CL” means that for $\mu > \mu_{UL}$, \mathbf{x}_0 is *not* in the highest-ranked 95% values of \mathbf{x} using the ordering corresponding to upper limits. That is,

$$\begin{aligned} 0.95 &= \Pr(r(\mathbf{X}) > r(\mathbf{x}_0); \mu = \mu_{UL}) \\ &\leq \Pr(r(\mathbf{X}) > r(\mathbf{x}_0); \mu > \mu_{UL}). \end{aligned} \tag{5}$$

For some μ_1 of scientific interest that is not equal to μ_{UL} , one can find the its p -CL for this ordering by finding μ_1 (perhaps iteratively) such that,

$$p\text{-CL} = \Pr(r(\mathbf{X}) > r(\mathbf{x}_0); \mu = \mu_1). \tag{6}$$

If H_0 is rejected (and even if not), in HEP one will typically quote, at one or more conventional CL, a two-sided confidence interval $[\mu_L, \mu_U]$ obtained from either central ordering or LR ordering. This means, for that ordering, that \mathbf{x}_0 is in the CL top-ranked values of \mathbf{x} for μ in the confidence interval, and that \mathbf{x}_0 is in lower-ranked values of \mathbf{x} for μ outside the confidence interval:

$$\text{CL} \geq \Pr(r(\mathbf{X}) \leq r(\mathbf{x}_0); \mu \in [\mu_L, \mu_U]) \text{ and} \quad (7)$$

$$\text{CL} < \Pr(r(\mathbf{X}) > r(\mathbf{x}_0); \mu \notin [\mu_L, \mu_U]). \quad (8)$$

For some μ_1 of scientific interest that is not one of the endpoints of a quoted conventional confidence interval, one can find its p -CL, that value of CL for which μ_1 is just on the edge of the interval, either $\mu_1 = \mu_L$ or $\mu_1 = \mu_U$, with

$$p\text{-CL} = \Pr(r(\mathbf{X}) > r(\mathbf{x}_0); \mu \in [\mu_L, \mu_U]). \quad (9)$$

8 Severity definitions in SIST for comparison to HEP

As I understand it, the core concept of severity analysis is one that we appreciate well in HEP. Regardless of whether or not the null hypothesis is rejected, there is additional data-dependent information to be conveyed regarding the compatibility of any chosen value of μ (or range of μ) with the observed data. Confidence intervals with CL fixed in advance can convey some of this information, but confidence intervals with multiple values of CL (in the extreme, the whole confidence distribution) can convey more. The particular p -CL relevant to chosen μ of interest conveys this additional information economically.

I think that severity analysis accomplishes the same thing. Severity SEV is approached from a multitude of angles in SIST, including worked examples. Formally (pp. 143, 347) there are three arguments (test T, outcome \mathbf{x} , claim C), where “claim C” (p. 143) is also written as “inference H” (p. 347). Typically, the first two arguments are suppressed when the context is clear.

8.1 The SIST prototype test T+

The prototype test used throughout SIST (e.g., pp. 141, 323, 342, 351) is called T+ and defined by:

$$H_0: \mu \leq \mu_0 \quad (10)$$

versus

$$H_1: \mu > \mu_0. \quad (11)$$

(Sometimes the focus is on the special case $\mu_0 = 0$.) In the statistics literature (including SIST p. 141), this test is called the “the one-sided test”. To me the HEP prototype test could be called one-sided, but I follow the statistics nomenclature here.

There are important qualitative differences between the HEP and SIST prototype tests. In particular, the absence of a point null hypothesis in the latter allows amelioration of the more disturbing aspects of the Jeffreys-Lindley paradox [14]. This is not so important for the basic idea of severity, but the differences in H_0 do affect the specific considerations

when H_0 is rejected. In the discussion below, when H_0 is rejected in the SIST prototype test, the information to be conveyed via severity analysis is still in the form of a one-sided inequality. However, when H_0 is rejected in the HEP prototype test, scientifically we are typically interested in conveying information via two-sided intervals (Section 9.1).

In any case, in SIST the usual Neyman-Pearson test is applied (with α pre-defined, thus implying β as a function of μ), using observed data \mathbf{x}_0 . In the rest of this section, for reference I include nearly verbatim the key passages of SIST that define severity mathematically and contain interpretive remarks.

8.2 Severity when H_0 is not rejected: Severity Interpretation of Negative Results (SIN)

This case is described on p. 347 and summarized essentially verbatim on p. 351. Suppose that the null hypothesis of the SIST prototype test, Eqn. 10, is *not* rejected in the N-P test. Then in the severity application, the “inference H ” is not just Eqn. 10, but rather the generalization to $\mu \leq \mu_1$ for $\mu_1 > \mu_0$, sometimes written as $\mu_1 = (\mu_0 + \gamma)$, for $\gamma \geq 0$. Then SEV (test T+, outcome \mathbf{x} , inference H) gives “the severity with which $\mu \leq \mu_1$ passes test T+, with data \mathbf{x}_0 ”, written as

$$\text{SEV}(\text{T+}, d(\mathbf{x}_0), \mu \leq \mu_1), \quad (12)$$

contracted to

$$\text{SEV}(\mu \leq \mu_1), \quad (13)$$

and defined by

$$\begin{aligned} \text{SEV}(\mu \leq \mu_1) &= \Pr(d(\mathbf{X}) > d(\mathbf{x}_0); \mu \leq \mu_1 \text{ false}) \\ &= \Pr(d(\mathbf{X}) > d(\mathbf{x}_0); \mu > \mu_1) \\ &> \Pr(d(\mathbf{X}) > d(\mathbf{x}_0); \mu = \mu_1). \end{aligned} \quad (14)$$

The interpretation is:

- (a) Low severity: If there is a very *low* probability that $d(\mathbf{x}_0)$ would have been larger than it is, even if $\mu > \mu_1$, then $\mu \leq \mu_1$ passes with *low* severity: $\text{SEV}(\mu \leq \mu_1)$ is low. I.e., your test wasn’t very capable of detecting discrepancy μ_1 even if it existed, so when it’s not detected, it’s poor evidence of its absence.
- (b) High severity: If there is a very *high* probability that $d(\mathbf{x}_0)$ would have been larger than it is, were $\mu > \mu_1$, then $\mu \leq \mu_1$ passes the test with *high* severity: $\text{SEV}(\mu \leq \mu_1)$ is high. I.e., your test was highly capable of detecting discrepancy μ_1 if it existed, so when it’s not detected, it’s a good indication of its absence.

Another explanation that I also found useful for application to the HEP prototype test is (p. 343):

Severity analysis: If $\Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0); \mu_1) = \text{high}$ and the result is not significant, then it’s an indication or evidence that $\mu \leq \mu_1$.

In my rephrasing and notation, that would be:

Severity analysis: If H_0 is not rejected, and if the tail probability $p(d(\mathbf{X}) \geq d(\mathbf{x}_0)|\mu_1)$ is high, then $\mu \leq \mu_1$ passes the test with high severity (equal to the tail probability). [In HEP, we would often say that say that “ $\mu > \mu_1$ is excluded at high confidence level.”]

8.3 Severity when H_0 is rejected (small p -value): Severity Interpretation of Rejection (SIR)

Suppose that the null hypothesis of the SIST prototype test, Eqn. 10, *is* rejected in the N-P test. This case is described on p. 265 and summarized, with slightly different phrasing, on p. 351. If the significance level α is small, the rejection is indicative of some discrepancy from H_0 , and we’re concerned about the magnitude.

The severity concept is evidently applied to the “inference H ” (or claim C) that $\mu > \mu_1$ (pp. 143, 265). While I did not see a general definition of SEV corresponding to Eqn. 14 for this (SIR) case, from the worked example on p. 143, we can infer:

$$\begin{aligned} \text{SEV}(\mu > \mu_1) &= \Pr(d(\mathbf{X}) \leq d(\mathbf{x}_0); \mu > \mu_1 \text{ false}) \\ &= \Pr(d(\mathbf{X}) \leq d(\mathbf{x}_0); \mu \leq \mu_1) \\ &> \Pr(d(\mathbf{X}) \leq d(\mathbf{x}_0); \mu = \mu_1), \end{aligned} \tag{15}$$

which has the given interpretations (p. 265) for high and low severity (noting that $1 - \text{SEV} < \Pr(d(\mathbf{X}) > d(\mathbf{x}_0); \mu = \mu_1)$):

- (i) [*Some* discrepancy is indicated]: $d(\mathbf{x}_0)$ is a good indication of $\mu > \mu_1 = \mu_0 + \gamma$ if there is a high probability of observing a *less* statistically significant difference than $d(\mathbf{x}_0)$ if $\mu = \mu_0 + \gamma$.
- (ii) [I’m not *that* impressed]: $d(\mathbf{x}_0)$ is a poor indication of $\mu > \mu_1 = \mu_0 + \gamma$ if there is a high probability of an even more statistically significant difference than $d(\mathbf{x}_0)$ even if $\mu = \mu_0 + \gamma$.

These are restated on p. 351 for low and high severity (letting $d_0 = d(\mathbf{x}_0)$):

- (a) *low*: if there is a fairly high probability that d_0 would have been larger than it is, even if $\mu = \mu_1$, then d_0 is not a good indication $\mu > \mu_1$: $\text{SEV}(\mu > \mu_1)$ is low.
- (b) *high*: Here are two ways, choose your preferred:
 - (b-1) If there is a very high probability that d_0 would have been smaller than it is, if $\mu \leq \mu_1$, then when you observe so large a d_0 , it indicates $\mu > \mu_1$: $\text{SEV}(\mu > \mu_1)$ is high.
 - (b-2) If there’s a very low probability that so large a d_0 would have resulted, if μ were no greater than μ_1 , then d_0 indicates $\mu > \mu_1$: $\text{SEV}(\mu > \mu_1)$ is high.

9 Connection of practice in HEP to severity

In the case where H_0 is *not* rejected, then in the HEP prototype test, it seems that severity for the claim $\mu < \mu_1$ is the p -CL for an upper limit equal to μ_1 (!). For example, we typically report μ_{UL} at 95% CL. For $\mu_1 = \mu_{\text{UL}}$, $\text{SEV}(\mu < \mu_1)$ is 95%. For $\mu_1 < \mu_{\text{UL}}$, the severity will be lower, and for $\mu_1 > \mu_{\text{UL}}$, the severity will be higher (closer to unity). To get the exact $\text{SEV}(\mu < \mu_1)$ for some desired μ_1 , one just calculates (or reads off from a confidence distribution) the CL for which μ_1 is the upper limit. This seems evident from comparing HEP practice in Eqns. 5–6 with severity Eqn. 14. While SIST mentions the possible usefulness of a connection between confidence distributions and severity on page 195, I think that (at least from a technical point of view), it is very direct and simple.

In the case where H_0 is rejected, the standard severity analysis of Section 8.3 considers the severity of one-sided claims $\text{SEV}(\mu > \mu_1)$. In HEP, once the point null of Eqn. 2 is rejected, we are nearly always interested in a two-tailed test (or the dual confidence interval) for μ . The confidence interval can be either that of the traditional central ordering or that from the unified LR ordering, avoiding the flip-flopping issue of using a different ordering based on whether or not H_0 is rejected. The unified ordering gives two-sided (non-central) intervals for which the lower endpoint can be used to test $\mu = 0$, while both lower and upper endpoints can be used to test some $\mu_1 > 0$.

The connection of this practice to the severity concept again seems evident from comparing comparing HEP practice in Eqns. 7–9 with severity Eqn. 15. The equations are not identical because the severity equations are one-sided (which defines a sign for $d(\mathbf{X})$ that flips the inequality), and the HEP equations are a bit more awkward to write. But the concept appears to be the same: p -CL provides the data-dependent information about the CL that corresponds to severity. The claim that “ μ_1 is within the confidence interval at p -CL” passes the test with severity equal to the p -CL.

For me the situation is thus clear regarding the statements on pp. 192-193 about the “intimate relationship between severity and confidence limits”, and how “severity will break out of the fixed $(1 - \alpha)$ level...” The post-data p -CL does just that.

Regarding the discussion on page 216-217 about how this gets us to a new philosophy of inference, at this point I do not see how this maps onto the way we think in HEP. I have written some thoughts about our philosophy (for example the choice of α , and what effect sizes are meaningful) in Ref. [3].

9.1 Examples from HEP

For plots of upper limits or the multi-D equivalent in HEP, typically only one CL is chosen; the rest of the confidence distribution (and thus severity for values of μ_1 not exactly at the upper limit) must be inferred by the reader, as I think experienced readers do. Some authors do however consider it of interest to plot the upper limit for at least one other value of CL, as mentioned in the quote by Cox in Section 6.2. I give an example plot from neutrino physics here, which shows curves for experiments that reject H_0 as well as those that do not reject H_0 . I also give three examples from Higgs boson physics, in which the null hypothesis is not $\mu = 0$, but rather the (fairly precise) prediction of the Standard Model (SM) of elementary particles.

9.1.1 Severity in a search for sterile neutrinos

The first example is from a search for evidence of a so-called “sterile neutrino”. There are three known neutrino species in the SM; all interact with other particles via the “weak force”, one of the four known forces in nature. There is long-standing speculation about the existence of (at least) a fourth neutrino, called “sterile” because it does not interact with the known weak force. One way to detect the presence of a sterile neutrino is if one of the ordinary neutrinos transmutes into another neutrino via the interaction with a sterile neutrino mediated by a new, even weaker force of nature.

In a simple case, the probability of transmutation can be expressed in terms of two unknown parameters in the speculative models, namely the difference in mass-squared, Δm^2 , between the sterile neutrino and another neutrino; and an effective angle θ in a quantum-mechanical Hilbert space. Experiments typically report 2D confidence regions in a plane conveying the two parameters, usually plotted as Δm^2 versus $\sin^2 2\theta$. So for this example, we consider μ to be the two-component vector with these two parameters, and the null hypothesis of Eqn. 2 means $\Delta m^2 = \sin^2 2\theta = 0$. (In fact, both parameters must be non-zero for the transmutation to take place, but the resulting experimental degeneracy along two axes is not relevant here.)

Experiments named LSND and MiniBooNE have claimed evidence for such transmutations, and have hence reported confidence regions that exclude H_0 . Numerous other experiments have searched for such transmutations and not seen them, i.e., they have reported observations compatible with H_0 . The plot I chose as a case study is one from a 2013 paper by the ICARUS collaboration [15], which obtained results compatible with no transmutation, but the results were *not* stringent enough to rule out the evidence from the previous LSND and MiniBooNE experiments.

In Figure 1, the 2D confidence regions from ICARUS data are to the left of the red vertical curves, shown for 90% and 99% CL. They are compatible with the null hypothesis, $\Delta m^2 = \sin^2 2\theta = 0$. So parameter values to the left of each red curve pass the test with severity equal to the labeled CL.

The plot also shows the confidence regions from LSND and MiniBooNE, which do not include the null hypothesis. The MiniBooNE regions are given at more than one CL. The conclusion from the plot is that ICARUS data was compatible with the null hypothesis, but it did not have the sensitivity to exclude the parameter values favored by LSND and MiniBooNE. This was also the case for other experimental curves in the upper left of the figure, and for additional ICARUS data. I chose an obsolete plot to make the statistics point, but this subfield of neutrino physics is an exciting area that continues in earnest.

9.1.2 Severity in searches for non-standard couplings of the Higgs boson

After the discovery of a particle dubbed a “Higgs-like boson” in July 2012, a comprehensive program was undertaken to determine if all of its properties really are compatible with the Higgs boson of the SM. In less than a year, enough was learned to remove the qualifier “-like” from public discourse, but a decades-long campaign is still considered crucial for testing alternative possibilities. This campaign includes measuring, as precisely as possible, the strength of the interaction of each other particle in the SM with the Higgs boson (which

can be loosely thought of as the carrier of its own special force). For each elementary particle in the SM, there is a parameter called a “coupling” that determines the interaction strength with the Higgs boson. These couplings are predicted with small uncertainty by the SM, incorporating previous precise measurements of other quantities, in particular the particles’ masses. Any confirmed departure of couplings from the SM values would be an exciting indication of “new physics”.

The results of numerous analyses of experimental data are summarized by displaying measurements of interaction strengths (closely related to the couplings) denoted by μ , normalized so that $\mu = 1$ is the SM prediction. As discussed in SIST (p. 211) and in Ref. [3] (Section 5.4), historically the first thing to establish was that $\mu > 0$ for at least some of the couplings, i.e., that there was a new boson that coupled to known particles. Then attention quickly turned away from testing $\mu = 0$, and toward testing the SM value $\mu = 1$ for as many couplings as possible. Thus, in this subsection, we depart from the HEP prototype test of Eqns. 2 and 3, and instead perform a two-sided test of the point null hypothesis of the SM,

$$H_0: \mu = 1, \tag{16}$$

versus the continuous alternative

$$H_1: \mu \neq 1. \tag{17}$$

(In speculative physics models beyond the SM, $\mu < 1$ and $\mu > 1$ are possible.)

The three plots shown here are taken from a paper [16] written jointly by the ATLAS and CMS collaborations, who combined their data samples after first each performing the measurements separately. (The combination was a huge effort taking over a year, going a lot deeper than simply computing a weighted average of the separate measurements, and carefully evaluating the level of correlation in all the nuisance parameters.)

Figure 2 presents the measurements of the interaction strength as measured in the decay of the Higgs boson to *five* different decay modes, each to a pair of known particles: two photons ($\gamma\gamma$); two Z bosons (electrically neutral bosons that are carriers of the weak force); two W bosons (electrically charged bosons that are carriers of the weak force); two tau leptons (particles similar to the electron, but with nearly twice the mass of a proton); and two b quarks (the fifth quark, called “bottom” or “beauty”). The two confidence intervals plotted for each measurement (obtained via Wilks’s theorem after appropriate cross-checks) are approximately 68% and 95% CL central intervals, and are part of the confidence distribution of intervals at all CL.

These measurements from these early data sets were compatible with $\mu = 1$, with interval lengths that were quite impressive experimental achievements at that moment, so soon after the discovery. Translating our nearly every-day discussions in HEP into the language of severity, one could say that μ values within 10-20% of unity typically pass the test of compatibility with $\mu = 1$ with severity approximately indicated by the given CL values; a more precise value of the severity can be inferred as desired.

As wonderful as these early results were, it turns out that many speculative new physics models would alter $\mu = 1$ by only a few per cent or less (especially after being constrained by existing LHC results). Thus, these data cannot distinguish the SM from much of the speculation. The confidence intervals will shrink at the LHC over the next ten years and

more, but there is also a big push for new accelerator(s) to shrink them to be yet smaller, and hence test the speculative models as stringently as is feasible.

Figure 3 displays the results of the same suite of analyses (including Higgs boson production measurements) in a complementary way, motivated by the fact that the SM physics equations governing the coupling of the force carriers (γ , Z , W , known as vector bosons) to the Higgs boson are very different from the physics equations governing the couplings of leptons and quarks (collectively known as fermions) to the Higgs boson. One inserts an ad-hoc scale factor κ_V in the couplings to vector bosons in the equations, and a factor κ_F in the equations for fermion couplings, and measures those factors. Figure 3 displays 2D confidence *regions* in the plane of κ_F versus κ_V , for two different CL. Again, these are two of the confidence regions in a confidence distribution of regions at all CL. And again, the results are compatible with the null hypothesis of the SM prediction of $\kappa_F = \kappa_V = 1$. And again, the severity of claims regarding alternative values of the factors is evident. It was a fantastic beginning to the campaign of measuring couplings, but also motivates the effort (and expense) of obtaining far better precision in the years (decades) to come.

With so many quantities being measured and various ratios of quantities being measured as well (with the advantage that certain systematic effects can cancel in a ratio), it was expected that there might be some outliers, if only because of random fluctuations in the data. These were scrutinized and reported with additional information that I believe also corresponds to the severity concept. In an interesting case, μ is the ratio of Higgs boson decays *to two b quarks* to Higgs boson decays *to two Z bosons*, denoted as the ratio B^{bb}/B^{ZZ} , as usual normalized so that $\mu = 1$ is the SM prediction. The measured μ departed somewhat from $\mu = 1$. A scan of the profile likelihood ratio Λ as a function of μ was computed and displayed as $-2\Delta \ln \Lambda$, where $\Delta \ln \Lambda$ is the difference $\ln \Lambda(\mu) - \ln \Lambda(\hat{\mu})$. As seen in Fig. 4, the likelihood ratio scan is asymmetric, with the best-fit value being approximately 2.5σ below the SM prediction.

One can read off the confidence distribution for μ (p -CL for any CL), and hence the severity of claims regarding any desired μ_1 , from Fig. 4 as follows. Since Wilks’s Theorem is reasonably valid for these data, and as there is just one degree of freedom (μ) remaining after removing all other parameters via profiling, the quantity plotted has a χ^2 distribution with one degree of freedom. Thus, reading off the curve at any μ and taking the square root gives the “number of σ ” in an equivalent equal-tailed test of a normal distribution, and thus the p -CL of the dual interval. The horizontal lines intercept the curve at the endpoints of the 1σ (68%) and 2σ (95%) confidence intervals. The confidence interval with endpoint on the SM prediction corresponds to about $\sqrt{6.3}\sigma = 2.5\sigma$, i.e., p -CL $\approx 99\%$. One can similarly calculate p -CL for any desired μ_1 . Given the “multiple trials” of all the measurements considered (not yet included in the analysis of p -CL), as well as the high threshold in HEP for claiming “new physics”, this result was put in the category of something to keep an eye on as more data were accumulated.

10 Summary and final thoughts

In HEP, regardless of whether or not the null hypothesis is formally rejected according to some rigid application of N-P testing, there is additional data-dependent information

to be conveyed regarding the compatibility of any chosen value of μ (or range of μ) with the observed data. Confidence intervals with CL fixed in advance can convey some of this information, but confidence intervals with multiple values of CL (in the extreme, the whole confidence distribution) can convey more. The confidence interval having critical confidence level p -CL that just barely contains some μ_1 of interest is a quantity of interest that can often be inferred from typical plots in HEP, or computed as scientific interest merits.

The quantity p -CL in the interval picture is dual to the p -value in the testing picture, just as the CL is dual to α . So this whole note could be written in terms of α and p -value (*testing picture*). However, in the *interval picture*, the connection to severity is more apparent to me. Aside from different conventions in HEP and SIST about defining H_0 , and which of the four orderings to use after the null hypothesis is rejected, the operational methods of conveying information about compatibility of the data with ranges of μ seem to be the same in HEP and in SIST.

The underlying philosophy of what we are doing is talked about much less in HEP and rarely articulated in depth. In my opinion the tensions between different philosophies are unresolved in HEP, just as they are in the statistics profession, particularly since our prototype test has a point null hypothesis and a continuous alternative, the situation of the Jeffrey-Lindley paradox. Thus, I have intentionally kept this note largely descriptive, rather than prescriptive, regarding practice in HEP.

Acknowledgments

I thank Deborah Mayo, Andre David, and members of the CMS Statistics Committee for useful discussions and comments on earlier drafts. This work was partially supported by the U.S. Department of Energy under Award Number DE-SC0009937.

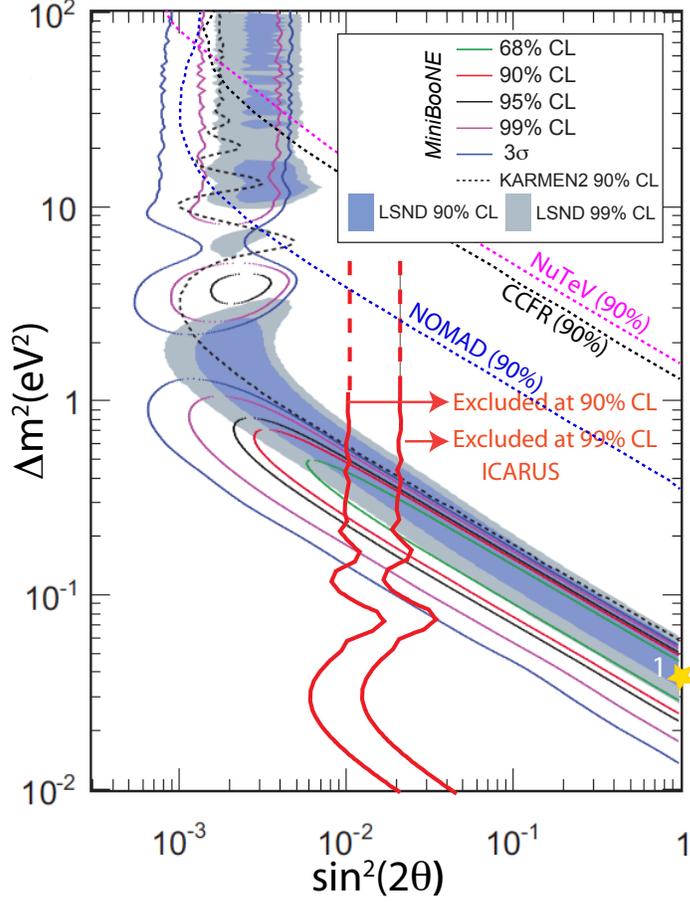


Figure 1: Plot of confidence regions in the plane of Δm^2 versus $\sin^2 2\theta$ for sterile neutrino searches. The ICARUS 2013 results (regions to the left of the red curves), do not reject the null hypothesis (the origin); the CL labeled on the “exclusion” curves corresponds to the severity of the claim that the true parameters are to the left of the respective curve. Also shown are the confidence regions of LSND and MiniBooNE that reject the null hypothesis (!); the values of CL given in the legend correspond to the interior of the regions outlined by the various curves, and hence to the severities of the claims that the true values are within the respective curves. There are also curves from some other less sensitive experiments. From Ref. [15], which gives references for all curves.

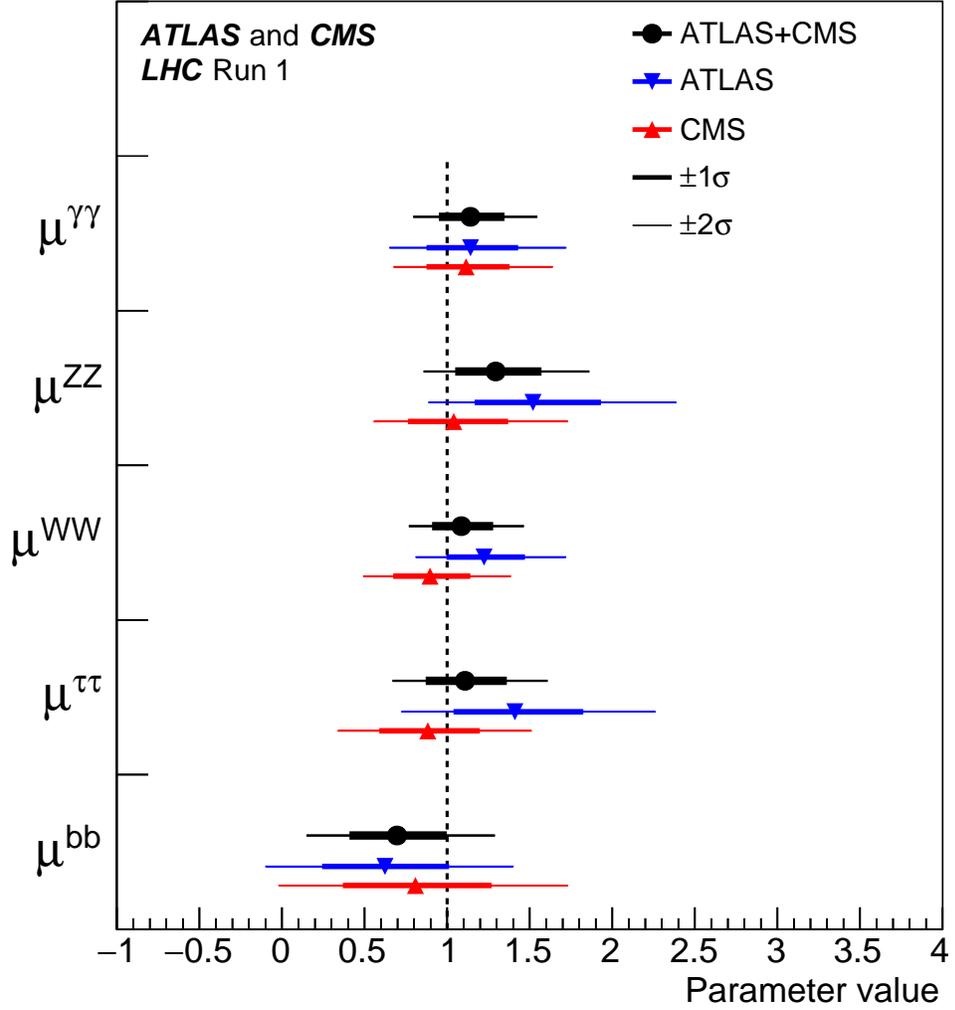


Figure 2: Point and interval estimates for the Higgs boson coupling strengths from the combination of ATLAS and CMS data. Also shown are the results from each experiment. The error bars indicate the 1σ (thick lines) and 2σ (thin lines) confidence intervals. They can thus be used to obtain the severity of the claims that μ is within each interval, and by extrapolation, other intervals. Figure 13 of Ref. [16].

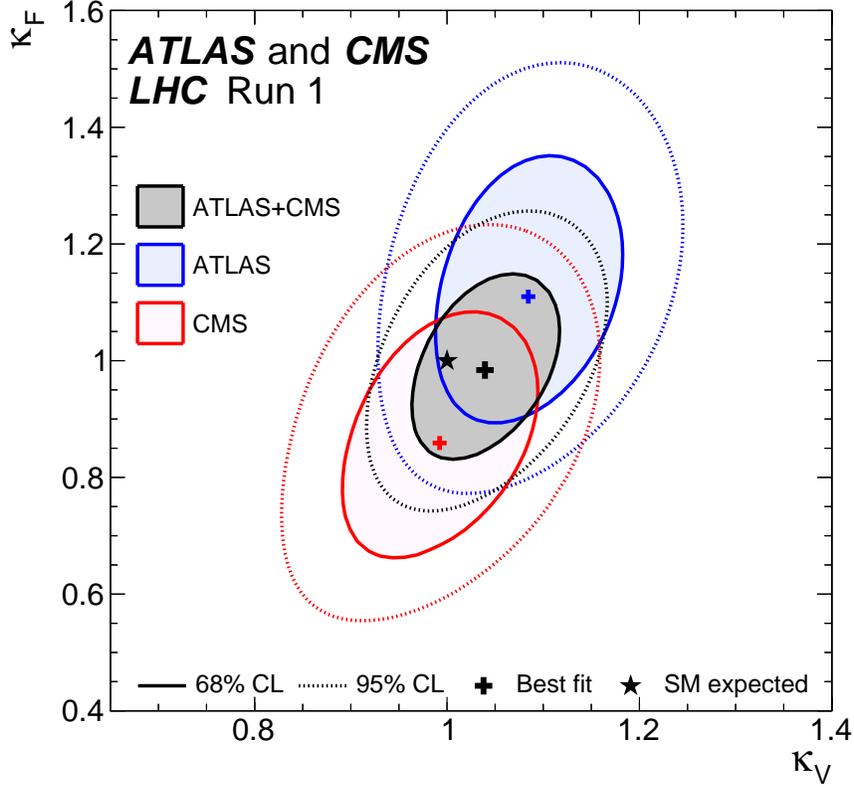


Figure 3: Negative log-likelihood contours at 68% and 95% CL in the (κ_V, κ_F) plane from the Higgs boson analysis of the combined ATLAS and CMS data sets. Also shown are the contours obtained for each experiment separately. The CL corresponds to the severity of the claim that the true values are within the respective region. In HEP, we say that values outside the regions are “excluded” at the respective CL. There is a long-term program to shrink the regions for the given CL, i.e., to test severely the claim that the true values lie within smaller regions. Of course, it would be of great interest to observe regions at very high CL that do not contain the SM values. Figure 26 (upper) of Ref. [16].

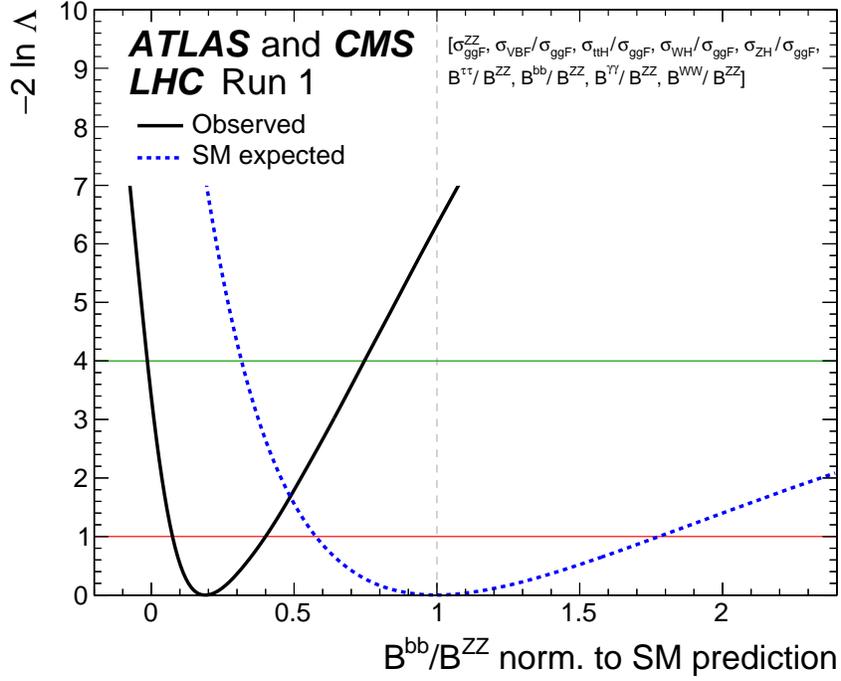


Figure 4: Observed (solid line) negative log-likelihood scan of the B^{bb}/B^{ZZ} parameter normalized to the corresponding SM prediction. All the other parameters are also varied in the minimization procedure (profiling). The dashed line is the result expected for data conforming exactly to the SM prediction. The red (green) horizontal line at the $-2\Delta \ln \Lambda$ value of 1 (4) indicates the value of the profile likelihood ratio corresponding to a 1σ (2σ) CL interval for the parameter of interest, assuming the asymptotic χ^2 distribution of the test statistic. The vertical dashed line indicates the SM prediction. Figure 9 of Ref. [16].

References

- [1] D. G. Mayo, “Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars”. Cambridge University Press, Cambridge, 2018.
- [2] A. Gelman et al., “Many perspectives on Deborah Mayo’s ‘Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars’ ”, (2019).
[arXiv:1905.08876](https://arxiv.org/abs/1905.08876) [stat.OT].
- [3] R. D. Cousins, “The Jeffreys–Lindley paradox and discovery criteria in high energy physics”, *Synthese* **194** (2017) 395, [arXiv:1310.3791](https://arxiv.org/abs/1310.3791).
[doi:10.1007/s11229-014-0525-z](https://doi.org/10.1007/s11229-014-0525-z), [10.1007/s11229-015-0687-3](https://doi.org/10.1007/s11229-015-0687-3).
- [4] A. Stuart, K. Ord, and S. Arnold, “Kendall’s Advanced Theory of Statistics”, volume 2A. Arnold, London, 6th edition, 1999. See also earlier editions by Kendall and Stuart. The hypothesis test that is dual to the confidence intervals of Ref. [5] is detailed (including profile likelihood for nuisance parameters) in the chapter on “Likelihood Ratio Tests and Test Efficiency,” pp. 238–239.
- [5] G. J. Feldman and R. D. Cousins, “Unified Approach to the Classical Statistical Analysis of Small Signals”, *Phys. Rev. D* **57** (1998) 3873, [arXiv:physics/9711021](https://arxiv.org/abs/physics/9711021).
[doi:10.1103/PhysRevD.57.3873](https://doi.org/10.1103/PhysRevD.57.3873).
- [6] M. Mandelkern, “Setting Confidence Intervals for Bounded Parameters”, *Statistical Science* **17** (2002) 149. With Comments by five statisticians and Rejoinder
<http://www.jstor.org/stable/3182816> .
- [7] R. D. Cousins, “Negatively Biased Relevant Subsets Induced by the Most-Powerful One-Sided Upper Confidence Limits for a Bounded Physical Parameter”, (2011).
[arXiv:1109.2023](https://arxiv.org/abs/1109.2023) [physics.data-an].
- [8] R. D. Cousins, “Bayes, Fisher, Neyman, Neutrino Masses, and the LHC”, (2011).
http://www.physics.ucla.edu/~cousins/stats/cousins_bounded_gaussian_virtual_talk_12sep2011.pdf .
- [9] R. Cousins, “Statistics in Theory: Prelude to Statistics in Practice”, (2018).
[arXiv:1807.05996](https://arxiv.org/abs/1807.05996). Writeup of Lectures at 2018 Hadron collider Physics summer School at Fermilab. Slides at
<http://www.physics.ucla.edu/~cousins/stats/theoryslides> .
- [10] A. L. Read, “Presentation of search results: The CL_s technique”, *J. Phys. G* **28** (2002) 2693. [doi:10.1088/0954-3899/28/10/313](https://doi.org/10.1088/0954-3899/28/10/313).
- [11] Particle Data Group, M. Tanabashi, et al., “Review of Particle Physics”, *Phys. Rev. D* **98** (2018) 030001. <http://pdg.lbl.gov/> .
- [12] D. R. Cox, “Some Problems Connected with Statistical Inference”, *Annals Math. Stat.* **29** (1958) 357. <http://www.jstor.org/stable/2237334> .

- [13] CDF Collaboration, “A measurement of $\sin(2\beta)$ from $B \rightarrow J/\psi K_S^0$ with the CDF detector”, *Phys. Rev. D* **61** (2000) 072005, [arXiv:hep-ex/9909003](https://arxiv.org/abs/hep-ex/9909003).
[doi:10.1103/PhysRevD.61.072005](https://doi.org/10.1103/PhysRevD.61.072005).
- [14] G. Casella and R. L. Berger, “Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem”, *J. Amer. Stat. Assoc.* **82** (1987) 106.
<http://www.jstor.org/stable/2289130> .
- [15] E. Segreto for the ICARUS Collaboration, “Experimental search for the LSND anomaly with the ICARUS LAr-TPC detector in the CNGS beam”, *J. Phys. Conf. Ser.* **447** (2013) 012064. [doi:10.1088/1742-6596/447/1/012064](https://doi.org/10.1088/1742-6596/447/1/012064).
- [16] ATLAS and CMS Collaborations, “Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV”, *JHEP* **08** (2016) 045, [arXiv:1606.02266](https://arxiv.org/abs/1606.02266). [doi:10.1007/JHEP08\(2016\)045](https://doi.org/10.1007/JHEP08(2016)045).