2012

Statistical evaluation of experimental determinations of neutrino mass hierarchy

X. Qian

A. Tan

W. Wang
William & Mary

R. D. McKeown
William & Mary, bmck@jlab.org

J. J. Ling

Recommended Citation

Qian, X., Tan, A., Wang, W., Ling, J. J., McKeown, R. D., & Zhang, C. (2012). Statistical evaluation of experimental determinations of neutrino mass hierarchy. Physical Review D, 86(11), 113011.

PHYSICAL REVIEW D **86,** 113011 (2012)

Statistical evaluation of experimental determinations of neutrino mass hierarchy

X. Qian,[1,*] A. Tan,[2,†] W. Wang,[3,‡] J. J. Ling,[4] R. D. McKeown,[3,5] and C. Zhang[4]

[1]*Kellogg Radiation Laboratory, California Institute of Technology, Pasadena, California 91125, USA*
[2]*Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa 52242, USA*
[3]*Physics Department, College of William and Mary, Williamsburg, Virginia 23187, USA*
[4]*Brookhaven National Laboratory, Upton, New York 11973, USA*
[5]*Thomas Jefferson National Accelerator Facility, Newport News, Virginia 23606, USA*

(Received 12 October 2012; published 21 December 2012)

Statistical methods of presenting experimental results in constraining the neutrino mass hierarchy (MH) are discussed. Two problems are considered and are related to each other: how to report the findings for observed experimental data and how to evaluate the ability of a future experiment to determine the neutrino mass hierarchy, namely, the sensitivity of the experiment. For the first problem where experimental data have already been observed, the classical statistical analysis involves constructing confidence intervals for the parameter Δm^2_{32}. These intervals are deduced from the parent distribution of the estimation of Δm^2_{32} based on experimental data. Because of existing experimental constraints on $|\Delta m^2_{32}|$, the estimation of Δm^2_{32} is better approximated by a Bernoulli distribution (a binomial distribution with one trial) rather than a Gaussian distribution. Therefore, the Feldman-Cousins approach needs to be used instead of the Gaussian approximation in constructing confidence intervals. Furthermore, as a result of the definition of confidence intervals, even if it is correctly constructed, its confidence level does not directly reflect how much one hypothesis of the MH is supported by the data rather than the other hypothesis. We thus describe a Bayesian approach that quantifies the evidence provided by the observed experimental data through the (posterior) probability that either hypothesis of MH is true. This Bayesian presentation of observed experimental results is then used to develop several metrics to assess the sensitivity of future experiments. Illustrations are made by using a simple example with a confined parameter space, which approximates the MH determination problem with experimental constraints on the $|\Delta m^2_{32}|$.

DOI: 10.1103/PhysRevD.86.113011 PACS numbers: 14.60.Pq

I. INTRODUCTION

Neutrino mass hierarchy (MH), i.e., whether the mass of the third-generation neutrino (ν_3 mass eigenstate) is greater or less than the masses of the first- and the second-generation neutrinos (ν_1 and ν_2), is one of the main questions to be answered in the standard model. Besides its fundamental importance to neutrino oscillation physics, the resolution of the neutrino MH plays an important role for the search of neutrinoless double-beta decay, which would determine whether a neutrino is a Dirac or Majorana fermion. With the recent discovery of a large value of $\sin^2 2\theta_{13}$ from Daya Bay [1–4], T2K [5], MINOS [6], Double Chooz [7], and RENO [8], the stage for addressing the neutrino MH has been set. It became one of the major goals of current and next-generation long baseline neutrino experiments (T2K [9], NOνA [10], and LBNE [11]) and atmospheric neutrino experiments (Super-K [12], MINOS [13], PINGU [14], and INO [15]). Meanwhile, the idea of utilizing a reactor neutrino experiment to determine the MH is also intensively discussed [16–20].

The objective of this paper is to present appropriate ways to do statistical analysis that will help determine the neutrino mass hierarchy. We start by introducing a few symbols and state the physics problem in terms of a pair of statistical hypotheses. Let m_1, m_2, and m_3 denote the masses of the ν_1, ν_2, and ν_3 mass eigenstate neutrinos, respectively, and let $\Delta m^2_{ij} \equiv m_i^2 - m_j^2$ for $i, j = 1, 2, 3$. As reviewed in Ref. [21], it is known that $\Delta m^2_{21} > 0$ from measurements of solar neutrinos given the definition of mixing angle θ_{12}, whereas the sign of Δm^2_{32} is so far unknown, and it is common to use NH and IH to denote the two hypotheses, the normal hierarchy and the inverted hierarchy, respectively:

$$\begin{cases} \text{NH:} & \Delta m^2_{32} > 0; \\ \text{IH:} & \Delta m^2_{32} < 0. \end{cases} \tag{1}$$

A unique feature to the above hypothesis testing problem is that there is additional, rather strong information regarding the parameter Δm^2_{32} that needs to be taken into account properly. Actually, based on previous experiments, a 68% confidence interval of $M^2_{32} \equiv |\Delta m^2_{32}|$ is given by $(2.43 \pm 0.13) \times 10^{-3}$ eV^2 [22].

*Corresponding authors.
xqian@caltech.edu
†aixin-tan@uiowa.edu
‡wswang@wm.edu

We will mainly address two aspects of the hypothesis testing problem. The first one concerns conducting a test after data have been collected. We discuss a classical testing procedure based on a $\Delta\chi^2$ statistic [Eq. (3)] or, equivalently, the procedure of constructing confidence intervals by inverting the test. As a matter of fact, the classical procedure is derived upon the assumption that the best estimator of Δm^2_{32} based on experimental data would follow a distribution that is approximately Gaussian. But due to existing constraints on M^2_{32}, this assumption is far from being satisfied. Consequently, actual levels of the resulting confidence intervals may deviate substantially from their nominal levels, as we demonstrate in Sec. II. Instead, a general way to construct confidence intervals that are true to their nominal levels is the Feldman-Cousins approach [23], which we also illustrate in detail in Sec. II.

Still, there is a fundamental limitation to the use of confidence intervals. Note that in the MH determination problem, one of the most crucial questions is, what is the chance that the MH is indeed NH (or IH) given the observed experimental data? Classical confidence intervals are not meant to answer this question directly, whereas credible intervals reported by a Bayesian procedure are. In Sec. III, we present a Bayesian approach, which effortlessly incorporates prior information on M^2_{32} and outputs the easy-to-understand (posterior) probability of NH and IH to conclude the test. We will emphasize the importance to differentiate the Bayesian credible interval from the classical confidence interval.

The second aspect of the hypothesis testing problem that we address concerns assessment of experiments in their planning stage. It is critical to evaluate the "sensitivity" of a proposed experiment, i.e., its capability to distinguish NH and IH. Since this evaluation is performed before data collection, it has to be based on potential data from the experiment. An existing evaluation method (such as employed in Refs. [11,24,25]) assumes that the most typical data set under one hypothesis, say, NH, happens to have been observed. Such a data set is referred to as the *Asimov data set* [26]. The method then calculates $\overline{\Delta\chi^2}$, which stands for the statistic $\Delta\chi^2$ in Eq. (3), with the extra bar indicating its dependence on the Asimov data set. It can be seen that $\overline{\Delta\chi^2}$ reflects how much the Asimov data set under NH disagrees with the alternative model, IH. It is then common practice to quantify the amount of disagreement by finding the p-value corresponding to $\overline{\Delta\chi^2}$ after comparing it to the quantiles of a chi-square distribution with 1 degree of freedom (choice of MH). Finally, 1 minus this p-value is sometimes reported as a quantitative assessment of the experiment. We will show in Sec. II that the comparison of the value of $\overline{\Delta\chi^2}$ to the quantiles of a chi-square distribution is not justified, when previous knowledge imposes constraints on the range of possible values of the parameter Δm^2_{32}.

As an alternative solution, we adopt a Bayesian framework and develop a set of new metrics for sensitivity to evaluate the potential of experiments to identify the correct hypothesis.

The paper is organized as follows. In Sec. II, we review the steps to construct classical confidence intervals for the parameter Δm^2_{32}. In Sec. III, we describe a Bayesian approach that reports the probability of each hypothesis of MH given an observed data set. We further extend this Bayesian method to help assess the sensitivity for future experiments. In Sec. IV, we illustrate the Bayesian approach for a simplified version of the MH problem. In particular, analytical formulas of the approximations for the probability of the hypotheses and those for the sensitivity metrics are provided. Also, a numerical comparison is made between the $\overline{\Delta\chi^2}$ based on the Asimov data set and the sensitivity metrics based on the Bayesian approach. Finally, discussions and a summary are presented in Secs. V and VI, respectively.

II. ESTIMATION IN CONSTRAINED VERSUS UNCONSTRAINED PARAMETER SPACES

In this section, we review a classical statistical procedure of forming confidence intervals. For the problem of determining the neutrino mass hierarchy, we demonstrate that the procedure is valid in one scenario but fails in another where known constraints on M^2_{32} are taken into consideration. In the latter case, the Feldman-Cousins method [23] based on Monte Carlo (MC) simulation is recommended to obtain valid confidence intervals.

Consider a spectrum that consists of n energy bins. Assume that the expected number of counts in each bin is a function of Δm^2_{32} and a nuisance parameter η. For simplicity, we denote Δm^2_{32} by θ. Then for the ith bin, let $\mu_i(\theta,\eta)$ and N_i represent the expected and the observed counts of neutrino-induced reactions, respectively. When μ_i is large enough, the distribution of N_i can be well approximated by a Gaussian distribution with mean μ_i and standard deviation $\sqrt{\mu_i}$.

Once the data $x = \{N_i, i = 1,\dots,n\}$ are observed, the deviations from the expected values $\{\mu_i(\theta,\eta), i = 1,\dots,n\}$ are often calculated to help measure the implausibility of the parameter (θ,η). Specifically, when the systematic uncertainties are omitted, and that certain available knowledge concerning the parameters θ and η is taken into consideration, one useful definition of the deviation is given by

$$\chi^2(\theta,\eta) = \chi^2_{\mathrm{stat}}(\theta,\eta) + \chi^2_p(|\theta|) + \chi^2_p(\eta) = \sum_i \frac{(N_i-\mu_i(\theta,\eta))^2}{(\delta N_i)^2} + \frac{(|\theta|-|\theta_0|)^2}{(\delta|\theta|)^2} + \frac{(\eta-\eta_0)^2}{(\delta\eta)^2}. \tag{2}$$

Here, the general notation δw represents the standard deviation of a variable w. So $\delta N_i = \sqrt{\mu_i}$, and the corresponding χ^2_{stat} term is called Pearson's chi-square. Also, note that $|\theta| = M_{32}^2$, and it is taken from [22] that $|\theta_0| = 2.43 \times 10^{-3}$ eV^2 and $\delta|\theta| = 0.13 \times 10^{-3}$ eV^2.

Based on Eq. (2) and a standard procedure discussed in Ref. [22], confidence intervals can be obtained for the parameter of interest $\theta(\Delta m_{32}^2)$, the sign of which is an indicator of the neutrino MH. First, define $\theta_{\min}$ to be the best fit to the data in the sense that $(\theta_{\min}, \eta_{\min}) = \mathrm{argmin}_{\theta,\eta}\chi^2(\theta, \eta)$, where the minimum is taken over $\Theta \times H$, the space of all possible values of (θ, η). Here, the general notation $\mathrm{argmin}_w h(w)$ denotes the value of w which corresponds to the minimum of the given function h. Note that $\theta_{\min}$ suggested by the observed data set will not be exactly the true value of the parameter θ, and a repetition of the experiment would yield a data set that corresponds to a different $\theta_{\min}$. So instead of reporting only $\theta_{\min}$, it is more rational to report a set of probable values of θ that fit the observed data not too much worse than that of the best fit and state how trustworthy the set is. Indeed, for any given θ, let $\eta_{\min}(\theta) = \mathrm{argmin}_\eta \chi^2(\theta, \eta)$, and define

$$\Delta\chi_{\min}^2(\theta) \equiv \chi^2(\theta, \eta_{\min}(\theta)) - \chi^2(\theta_{\min}, \eta_{\min}), \quad (3)$$

then a *level-a confidence interval* based on Eq. (3) is defined to be

$$C_a = \{\theta \in \Theta: \Delta\chi_{\min}^2(\theta) \le t_a\}, \quad (4)$$

where we use the standard set-builder notation $\{h(w)\text{: restriction } w\}$ to denote a set that is made up of all the points $h(w)$ such that w satisfies the restriction to the right of the colon. The key in constructing Eq. (4) is to specify the correct threshold value t_a for a given confidence level a. (See the final paragraph of this section for a more detailed description of what confidence level means.) Most commonly examined confidence levels use $a = 68.27\%(1\sigma)$, $95.45\%(2\sigma)$, and $99.73\%(3\sigma)$, which are often linked to threshold values $t_a = 1$, 4, and 9, respectively [22]. Note that these three values are the 68.27%, 95.45%, and 99.73% quantiles of the chi-square distribution with 1 degree of freedom, respectively. They are used as threshold values because the parameter space Θ is of dimension one and that, under certain regularity conditions, $\Delta\chi_{\min}^2(\theta)$ would follow approximately a chi-square distribution with 1 degree of freedom when θ is the true parameter value. This procedure and its extensions to cases where θ is of higher dimension have been successfully applied in many studies [11,24,25,27–33] in order to constrain various parameters in the neutrino physics.

Although this procedure has been widely used in analyzing experimental data, note that it is not universally applicable. Its limitations have been addressed by Feldman and Cousins [23]. Below, we illustrate this point through a simple MC simulation study. It will be shown that, in a situation that is similar in nature to the MH determination problem in Eq. (1) where there exist special constraints on the possible values of θ, the aforementioned threshold values based on chi-square approximation could result in bad confidence intervals. That is, the actual coverage probabilities of the intervals strongly disagree with their nominal levels.

In the simulation, we set $n = 10$, $\mu_i(\theta) = 1000 + 15 \cdot \theta$ for $i = 1, \ldots, n$. (Here, no nuisance parameter η is introduced, and all the expected bin counts are assumed equal for simplicity. Nevertheless, these assumptions are not essential to the purpose of our simulation.) The following two cases are investigated:

(i) Case I: $\Theta = (-\infty, \infty)$,

(ii) Case II: $\Theta = \{-1, 1\}$.

Case I is a typical situation where nothing was known about θ before the current experiment, whereas case II is designed to imitate the situation where existing measurements of $|\theta| = M_{32}^2$ are very accurate at around 2.43×10^{-3} eV^2, and we simply denoted this value to 1 for clarity of presentation. Further, the definition of deviation analogue to Eq. (2) is taken to be $\chi^2(\theta) = \sum_i \frac{(N_i - \mu_i(\theta))^2}{\mu_i(\theta)}$ for case I. For case II, the chi-square definition is $\chi^2(\theta) = \sum_i \frac{(N_i - \mu_i(\theta))^2}{\mu_i(\theta)} + \frac{(|\theta| - |\theta_0|)^2}{(\delta|\theta|)^2}$ with experimental constraints on $|\theta|$. It is then reduced to $\chi^2(\theta) = \sum_i \frac{(N_i - \mu_i(\theta))^2}{\mu_i(\theta)}$ with θ being only 1 or -1.

Under each case, we set the true value of θ to be $\theta_0 = 1$, based on which 100 000 MC samples are simulated, denoted by $\{N_1^{(j)}, \ldots, N_{10}^{(j)}\}$ for $j = 1, \ldots, 100\,000$. Then for the jth sample, confidence intervals of levels $a = 68.27\%$, 95.45%, and 99.73% are constructed according to Eq. (4) by using threshold values 1, 4, and 9, respectively. Finally, at each of the three levels, we record the proportion of confidence intervals out of the 100 000 that include the truth $\theta_0 = 1$. The results are reported in the last three columns of Table I. It can be seen that, in case I, the actual coverage probabilities closely match the nominal levels. However, in case II, the actual coverage probabilities are always higher.

Without too much technical detail, we try to explain the reason why the chi-square procedure produced valid confidence intervals for case I but not for case II. In general, having observed data x from a parametric model $P(x|\theta)$, a sensible test for a pair of hypotheses, H_0: $\theta \in \Theta_0$ and H_1: $\theta \in \Theta - \Theta_0$ (the counterpart of H_0), is the likelihood ratio test that is based on the test statistic

$$\Delta\chi_{\min}^2 \equiv -2\log\left(\frac{P(x|\theta_{0,\min})}{P(x|\theta_{\min})}\right), \quad (5)$$

where $\theta_{0,\min} = \mathrm{argmin}_{\{\theta\in\Theta_0\}} P(x|\theta)$ and $\theta_{\min} = \mathrm{argmin}\{\theta \in \Theta_0\} P(x|\theta)$ are the best fit over the null parameter set Θ_0 and the full parameter set Θ, respectively. If the observed data x yields a large $\Delta\chi_{\min}^2$, it means that Θ_0 is implausible, which further leads to the rejection of H_0. Note that the statistic $\Delta\chi_{\min}^2(\theta)$ in Eq. (3) is a special case of Eq. (5) with Θ_0 consisting of a single point, θ.

TABLE I. Confidence levels for various of $\Delta\chi^2_{\min}$ region for the Gaussian and the Bernoulli distribution from MC calculations. In case I, the mean and the standard deviation of the Gaussian distribution are found to be about 1 and $\sigma = 0.67$, respectively. In case II, the parameter p of the Bernoulli distribution (e.g., percentage of $\theta_{\min} < 0$) is found to be about 6.8%.

Case	$\Delta\chi^2_{\min}(\theta_{\text{true}})$ distribution	$\theta_{\min}$ distribution	Distribution parameter within this example	$\Delta\chi^2_{\min} \le 1$ confidence level	$\Delta\chi^2_{\min} \le 4$ confidence level	$\Delta\chi^2_{\min} \le 9$ confidence level
I	Chi-square	Gaussian	mean = 1 and $\sigma = 0.67$	68.27%	95.48%	99.73%
II	$\cdots$	Bernoulli	$p = 0.0679$	95.12%	98.48%	99.86%

In order to determine the correct threshold values in rejecting or, equivalently, in constructing confidence intervals defined by Eq. (4), the distribution and quantiles of $\Delta\chi^2_{\min}(\theta_0)$ considering all possible data set need to be known, when the true parameter value is some $\theta_0 \in \Theta_0$. An important result in statistics, the Wilks theorem [34,35] states that, under certain regularity conditions, $\Delta\chi^2_{\min}(\theta_0)$ follows approximately a chi-square distribution with a degree of freedom equal to the difference between the dimension of Θ and that of Θ_0, when the data size is large. (In our problem, the data size is simply $\sum_i N_i$.) The main regularity conditions are, as we quote [35], "the model is differentiable in θ and that Θ_0 and Θ are (locally) equal to linear spaces." Essentially, such conditions imply that $\theta_{\min}$ follows an approximately Gaussian distribution centered at the true θ value, which eventually implies an approximate chi-square distribution for $\Delta\chi^2_{\min}(\theta_0)$.

In case I of our simulation, the best estimation of θ can be calculated directly from the number of events in each bin: $\theta_{\min} = [\sqrt{\sum_{i=1}^n N_i^2/n} - 1000]/15$.[1] The aforementioned regularity conditions are satisfied in this case, and the distribution of $\theta_{\min}$ and that of $\Delta\chi^2_{\min}(\theta_0)$ follow approximately the Gaussian and the chi-square distribution, respectively, as the Wilks theorem predicts. In the top two panels of Fig. 1, we reconfirm this fact by comparing their histograms based on the 100 000 MC samples (black shaded area) to the probability density function of the Gaussian and the chi-square distribution (blue long dash-dotted line). On the other hand, the full parameter space in case II consists of two isolated points and clearly violates the conditions required by the Wilks theorem. Indeed, in case II, the best estimation of θ is given by

$$\theta_{\min} = \begin{cases} 1 & \text{if } \chi^2(\theta = 1) < \chi^2(\theta = -1), \\ -1 & \text{otherwise} \end{cases}$$

and follows a Bernoulli distribution, and $\Delta\chi^2_{\min}(\theta_0)$ follows a distribution quite different from a canonical chi-square distribution. Approximations to the actual distributions of $\Delta\chi^2_{\min}(\theta_0)$ and $\theta_{\min}$ can be obtained from the 100 000 MC samples and are shown (black shaded area) in the bottom two panels of Fig. 1. Further, an analytical approximation (red dash-dot-dotted line) to the distribution is derived in Appendix A. The analytical calculation implies that, independent of whether the truth θ_0 is 1 or -1, the p-value[2] corresponding to an observed value of $\Delta\chi^2_{\min}(\theta_0)$, say, t, is approximately given by

$$p\text{-value}(t) = P(\Delta\chi^2_{\min}(\theta_0) \ge t) \approx \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left(\frac{t + \overline{\Delta\chi^2}}{\sqrt{8\overline{\Delta\chi^2}}}\right) \tag{6}$$

for any $t > 0$; and the p-value is 1 for any $t \le 0$. Here erf is the Gaussian error function: $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$. [We use the general notation $P(A)$ to denote the probability of an event A].

The discussions above suggest that, when constructing confidence intervals in special cases where conditions of the Wilks theorem do not hold (or that the user cannot be sure if the conditions hold), the regular threshold values (such as $t_\alpha = 1, 4, 9$ mentioned earlier) should not be taken for granted. Instead, alternative thresholds based on MC or case-specific analytical approximations are needed. We recommend using the MC method with a large MC sample size whenever possible, because, unlike other methods, it is guaranteed to produce a valid confidence interval for θ. We hereby review how to produce a valid 1σ (68.27%) confidence interval for θ using the MC method [23]. This method can easily be generalized to build confidence intervals of any level.

(1) Having observed data $x = \{N_1, \ldots, N_n\}$, apply the following procedure to every θ in the parameter space Θ (fix one θ at a time):
 (a) Calculate $\Delta\chi^2_{\min}(\theta)^x$ with Eq. (3) based on the observed data.
 (b) Simulate a large number of MC samples, say, $\{x^{(j)}\}_{j=1}^T$, where $x^{(j)} = \{N_1^{(j)}, \ldots, N_n^{(j)}\}$ is generated from the model with true parameter value θ. For $j = 1, \ldots, T$, calculate $\Delta\chi^2_{\min}(\theta)^{(j)}$, that is, Eq. (3) based on the jth MC sample $x^{(j)}$.

[1]Note that the above $\theta_{\min}$ can be closely approximated by $[(\sum_{i=1}^n N_i)/n - 1000]/15$, which is indeed the exact maximum likelihood estimator for θ had we assumed that each count N_i follows a Poisson distribution with mean $\mu_i(\theta) = 1000 + 15\cdot\theta$.

[2]The p-value at t is defined to be the percentage of potential measurements that result in the same or a more extreme value of the test statistic, say, $\Delta\chi^2_{\min}$, than t.

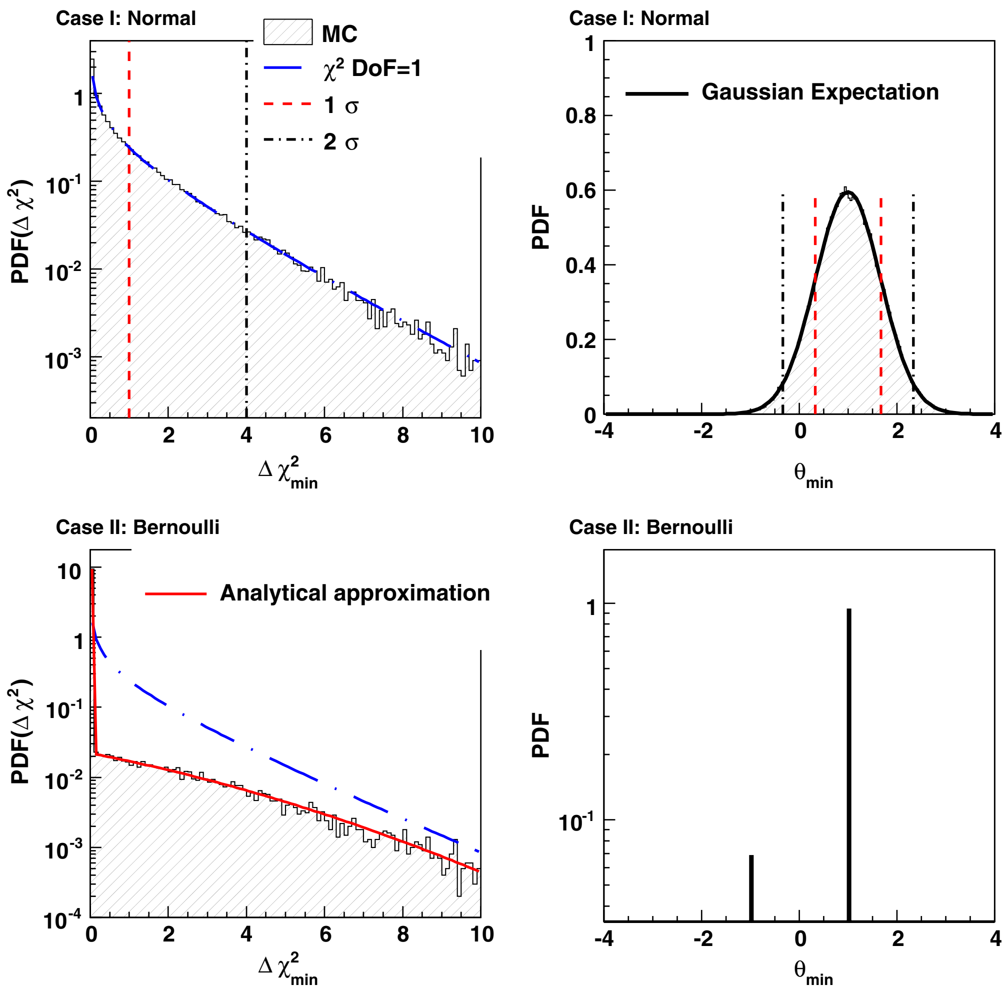

FIG. 1 (color online). Distributions of $\Delta\chi^2_{\rm min}(\theta_0)$ and $\theta_{\rm min}$ for case I and case II with 100 000 MC samples. The $\theta_{\rm min}$ distribution of case I (top right) and case II (bottom right) are a Gaussian and a Bernoulli distribution, respectively. The $\Delta\chi^2_{\rm min}(\theta_0)$ distribution of case I (top left) is consistent with the chi-square distribution with degree of freedom 1. The commonly used 1σ, (68.27% confidence level) and 2σ, (95.45% confidence level) regions are labeled with red dashed and black dash-dotted lines, respectively, for case I. The $\Delta\chi^2_{\rm min}(\theta_0)$ distribution of case II (bottom left) strongly deviates from the chi-square distribution. In case II, we also show the analytical approximation (derived in Appendix A) of the distribution of $\Delta\chi^2_{\rm min}$. We should emphasize that, while the chi-square distribution does not depend on any additional parameter (other than $\Delta\chi^2_{\rm min}$), the analytical approximation depends on $\overline{\Delta\chi^2}$.

This produces an empirical distribution of the statistic $\Delta\chi^2_{\rm min}(\theta)$.

(c) Calculate the percentage of MC samples such that $\Delta\chi^2_{\rm min}(\theta)^{(j)} < \Delta\chi^2_{\rm min}(\theta)^x$. Then θ is included in the 1σ confidence interval if and only if the percentage is smaller than 68.27%.

One can easily check that p-values analytically obtained from Eq. (6) for case II (basically the MC method) are consistent with the simulation results listed in Table I.

On a separate issue that was also emphasized in Ref. [23], classical confidence intervals should not be confused with Bayesian credible intervals. On one hand, the *confidence level* of a confidence interval, say, a, is an evaluation of this interval estimation procedure based on many *potential* repetitions of the experiment. More specifically, had the experiment been independently repeated 100 times, applying the estimation procedure to each would result in 100 intervals, and a represents the proportion of these intervals that we expect to contain the true value of the unknown parameter θ. The level-a confidence interval reported in practice is the result of applying such a procedure to the data observed in the current experiment. On the other hand, a Bayesian credible interval, say, of *credible level* b, is a region in the parameter space such that, given the *observed* data, it contains the true value of the unknown parameter with probability b. In general, an a-level confidence interval does not coincide with an a-level Bayesian credible interval. In other words, if C_a is an a-level confidence interval built from the observed data x, then it is generally inappropriate to give the interpretation that $P(\theta \in C_a|x)$ (the probability of true θ inside C_a given data x) is α. Nevertheless, in Appendix B, we discuss when confidence intervals approximately match Bayesian credible intervals. In the next section, we present a Bayesian approach to the problem of determining neutrino mass hierarchy.

III. A BAYESIAN APPROACH TO DETERMINE NEUTRINO MASS HIERARCHY

A. Bayesian inference based on observed data

The MH determination problem is concerned with comparing two competing models, NH and IH, having observed data x. The Bayesian approach to the problem is based on the probabilities that each model is true given x, namely, $P(\mathrm{NH}|x)$ and $P(\mathrm{IH}|x) = 1 - P(\mathrm{NH}|x)$. [In general, we adopt the notation $P(A|B_1, \ldots, B_n)$ to represent the probability of event A given events $B_1, \ldots, B_n$. Also, we use capital letters such as $S_1, \ldots, S_n$ and T to denote random variables and use small letters such as $s_1, \ldots, s_n$ and t to denote numbers inside the range of possible values of the random variables. Then $P_{T|S_1,\ldots,S_n}(t|s_1, \ldots, s_n)$ denotes for the conditional probability density function (pdf) or the conditional probability mass function given events $S_1, \ldots, S_n = s_1, \ldots, s_n$. The subscript to P is often omitted when it is clear what random variable is being considered.] Model NH will be preferred over IH if the odds $r(x) = P(\mathrm{IH}|x)/P(\mathrm{NH}|x) < 1$. Moreover, the size of r serves as an easy-to-understand measure for the amount of certainty of this preference. Alternatively, some people may feel more comfortable in interpreting $P(\mathrm{NH}|x) = 1/(1 + r(x))$ directly.

One can determine $P(\mathrm{NH}|x)$ and $P(\mathrm{IH}|x)$ within a Bayesian framework as follows. Let the true value of MH be either NH or IH, and let the counts N_i follow a Gaussian distribution with mean $\mu_i^{\mathrm{MH}}(\theta, \eta_{\mathrm{MH}})$ and standard deviation $\sqrt{\mu_i^{\mathrm{MH}}(\theta, \eta_{\mathrm{MH}})}$ for $i = 1, \ldots, n$. Here, θ is the parameter of interest, and η_{MH} denotes other unknown nuisance parameter(s). Here, a subscript accompanies η to emphasize that the nuisance parameter is allowed to have different interpretations and behavior under the two hypotheses. (We will omit this subscript whenever there is no possibility of confusion.) If prior knowledge is available for θ and η, then they should be elicited to form prior distributions, $P(\theta, \eta|\mathrm{MH})$ for MH = IH, NH. Sometimes, it is reasonable to assume that the parameters θ and η are independent conditional on MH and, hence, $P(\theta, \eta|\mathrm{MH}) = P(\eta|\mathrm{MH})P(\theta|\eta, \mathrm{MH}) = P(\eta|\mathrm{MH})P(\theta|\mathrm{MH})$.

Specific to the MH problem at hand, under NH (IH), previous knowledge (e.g., from Ref. [22]) suggests that a sensible prior for θ would be a Gaussian with mean 2.43×10^{-3} eV^2 (-2.43×10^{-3} eV^2) and standard deviation 0.13×10^{-3} eV^2. Since the hypotheses being tested are NH: $\theta \in \Theta_{\mathrm{NH}} = (0, \infty)$ versus IH: $\theta \in \Theta_{\mathrm{IH}} = (-\infty, 0)$, $P(\theta|\mathrm{NH})$ and $P(\theta|\mathrm{IH})$ are specified to be the truncated version of the above Gaussian distributions supported within Θ_{NH} and Θ_{IH}, respectively. Nevertheless, in our Bayesian model, $P(\theta \in \Theta_{\mathrm{IH}}|\mathrm{NH})$ and $P(\theta \in \Theta_{\mathrm{NH}}|\mathrm{IH})$ based on the Gaussian prior are so tiny that they will yield the same numerical results as the truncated version. A similar choice can be made for $P(\eta|\mathrm{MH})$.

According to Bayes' theorem, we have

$$P(\mathrm{NH}|x) = \frac{P(x|\mathrm{NH}) \cdot P(\mathrm{NH})}{P(x)} = \frac{P(x|\mathrm{NH}) \cdot P(\mathrm{NH})}{P(x|\mathrm{NH}) \cdot P(\mathrm{NH}) + P(x|\mathrm{IH}) \cdot P(\mathrm{IH})}. \quad (7)$$

Here, $P(\mathrm{NH})$ and $P(\mathrm{IH}) = 1 - P(\mathrm{NH})$ should reflect one's knowledge in NH and IH prior to the experiment. In the MH problem, it is reasonable to assume that NH and IH are equally likely, that is, $P(\mathrm{NH}) = P(\mathrm{IH}) = 50\%$. We will make this assumption throughout the paper. Consequently, Eq. (7) reduces to

$$P(\mathrm{NH}|x) = \frac{P(x|\mathrm{NH})}{P(x|\mathrm{NH}) + P(x|\mathrm{IH})}. \quad (8)$$

Based on probability theory, $P(x|\mathrm{MH})$, i.e., the likelihood of model MH, is a "weighted average" of $P(x|\theta, \eta, \mathrm{MH})$ over all possible values of (θ, η):

$$P(x|\mathrm{MH}) = \int_{H_{\mathrm{MH}}} \int_{\Theta_{\mathrm{MH}}} P(\eta|\mathrm{MH})P(\theta|\eta, \mathrm{MH}) \times P(x|\theta, \eta, \mathrm{MH})d\theta d\eta, \quad (9)$$

in which H_{MH} represents the phase space of nuisance parameter η given the choice of MH. Furthermore, under the assumption that θ and η are independent, Eq. (9) is reduced to

$$P(x|\mathrm{MH}) = \int_{H_{\mathrm{MH}}} \int_{\Theta_{\mathrm{MH}}} P(\eta|\mathrm{MH})P(\theta|\mathrm{MH}) \times P(x|\theta, \eta, \mathrm{MH})d\theta d\eta. \quad (10)$$

In practice, the integral in Eq. (9) is often analytically intractable but can be approximated by using MC methods. Using a basic MC scheme, first, a large number of samples $\{(\theta^{(j)}, \eta^{(j)}), j = 1, \ldots, T\}$ are randomly generated from the prior distribution $P(\theta, \eta|\mathrm{MH})$. Then for the observed data x, obtain $\hat{P}_T(x|\mathrm{MH}) := T^{-1}\sum_{j=1}^{T} P(x|\theta^{(j)}, \eta^{(j)}, \mathrm{MH})$. As the MC size T increases, the estimator $\hat{P}_T(x|\mathrm{MH})$ will have a probability approaching 1 of being arbitrarily close to the true $P(x|\mathrm{MH})$. Note that there exist much more efficient MC algorithms, such as importance sampling algorithms, that require smaller, more affordable T for the resulting estimators to achieve the same amount of accuracy as that of the basic MC scheme. Interested readers are pointed to Ref. [36] for further details and references.

There also exist (relatively crude) approximations to $P(x|\mathrm{MH})$ in Eq. (9) that avoid the intense computation in the MC approach. A most commonly used one is the one on which a popular model selection criteria, the Bayesian information criterion is based. This approximation is often presented in terms of an approximation to a one-to-one transformation of $P(x|\mathrm{NH})$, namely,

$$\Delta\chi^2(x) \equiv -2\log r(x) = -2\log(P(\mathrm{IH}|x)/P(\mathrm{NH}|x)). \quad (11)$$

Denote

$$\mathcal{T}_{\mathrm{MH}}(x) \equiv -2\log\{\max_{\theta,\eta} P(x|\theta,\eta,\mathrm{MH})P(\eta|\mathrm{MH})P(\theta|\mathrm{MH})\},$$

where the maximum is taken over $(\theta, \eta) \in \Theta_{\mathrm{MH}} \times H_{\mathrm{MH}}$ and

$$\Delta\mathcal{T}(x) \equiv \mathcal{T}_{\mathrm{IH}}(x) - \mathcal{T}_{\mathrm{NH}}(x). \quad (12)$$

Then if the sample size $\sum_i N_i$ is large, and η_{NH} and η_{IH} are of the same dimension,

$$\Delta\chi^2(x) = 2\log P(x|\mathrm{NH}) - 2\log P(x|\mathrm{IH}) \approx \Delta\mathcal{T}(x). \quad (13)$$

Here, the equality follows from Eq. (8), and the approximation is supported by a crude Taylor expansion around the maximum likelihood estimator for the parameters. There are other approximations that follow the same line that are more accurate but also computationally more demanding. See Ref. [36] for details.

One remark should be made regarding $\Delta\mathcal{T}$, as it is closely related to a commonly used test statistic in the classical testing procedure. Indeed, if the truncated Gaussian priors mentioned earlier are assigned for θ and a Gaussian prior with mean η_0 and standard deviation $\delta\eta$ are assigned for η under both NH and IH, then according to the definition of χ^2 in Eq. (2), we have

$$\delta\chi^2 \equiv \chi^2(\hat{\theta}', \hat{\eta}') - \chi^2(\hat{\theta}, \hat{\eta}) = \Delta\mathcal{T} - \sum_{i=1}^{n} \log\frac{\mu_i(\hat{\theta}', \hat{\eta}')}{\mu_i(\hat{\theta}, \hat{\eta})}, \quad (14)$$

where $(\hat{\theta}, \hat{\eta})$ and $(\hat{\theta}', \hat{\eta}')$ denote maximizers of

$$P(x|\theta, \eta, \mathrm{MH})P(\eta|\mathrm{MH})P(\theta|\mathrm{MH}), \quad \mathrm{MH} = \mathrm{NH},\ \mathrm{IH},$$

within their respective range. [Note that $\delta\chi^2$ is essentially an alternative version of $\Delta\chi^2_{\min}$ in Eq. (3), bearing some technical difference only.] Here, the term $\sum_{i=1}^{n} \log\frac{\mu_i(\hat{\theta}', \hat{\eta}')}{\mu_i(\hat{\theta}, \hat{\eta})}$ is the result of the normalization factor [e.g., $(2\pi\sigma^2)^{-\frac{1}{2}}$] of the Gaussian pdf and is in general small compared to $\Delta\mathcal{T}$. In the classical testing procedure, the observed value of $\delta\chi^2$ will be compared to its parent distribution to get a p-value, whereas the Bayesian approach described in this section directly interprets the value of $\Delta\mathcal{T}$, by transforming it to either the odds ratio between NH and IH, $r(x) = e^{-\Delta\chi^2(x)/2} \approx e^{-\Delta\mathcal{T}(x)/2}$, or the probability of NH,

$$P(\mathrm{NH}|x) = \frac{1}{1 + r(x)} = \frac{1}{1 + e^{-\Delta\chi^2(x)/2}} \approx \frac{1}{1 + e^{-\Delta\mathcal{T}(x)/2}}, \quad (15)$$

and similarly, the probability of IH,

$$P(\mathrm{IH}|x) = \frac{r(x)}{1 + r(x)} = \frac{e^{-\Delta\chi^2(x)/2}}{1 + e^{-\Delta\chi^2(x)/2}} \approx \frac{e^{-\Delta\mathcal{T}(x)/2}}{1 + e^{-\Delta\mathcal{T}(x)/2}}. \quad (16)$$

B. Sensitivity of experiments

So far, we described the Bayesian procedure for testing the two hypotheses, NH and IH, given observed data $x = \{N_1, \ldots, N_n\}$. Reasoning backwards, foreseeing what analysis will be done after data collection allows us to address the question that, before data are collected from a proposed experiment, how confidently do we expect it to be able to distinguish the two hypotheses NH and IH. We loosely refer to such an ability as the sensitivity of the experiment. There could be many ways to define sensitivity, and we list a few below. In practice, evaluating a proposed experiment by using one or several of these sensitivity criteria provides views from different angles of the potential return from the experiment.

Note that sensitivity depends on the underlying true model as well as future experimental results generated from this model. For example, if NH is true, then we have a population of potential experimental results $x \sim P(x|\mathrm{NH}) = \iint P(x|\theta, \eta, \mathrm{NH})P(\theta, \eta|\mathrm{NH})d\theta d\eta$. And each potential x is associated with a posterior probability $P(\mathrm{NH}|x)$. Then one could evaluate the ability of an experiment to confirm NH when it is truly the underlying model by looking at the distribution of $P(\mathrm{NH}|x)$. The most typical numerical summaries of this distribution include its mean, quantiles, and tail probabilities, all of which can be used to address sensitivity.

Below, we officially develop metrics for sensitivity under the assumption that NH is true. Note that these metrics can be similarly defined when IH is true.

(1) The average posterior probability of NH is given by

$$\begin{aligned}\bar{P}^{\mathrm{NH}}_{T=\mathrm{NH}} &= \int P(\mathrm{NH}|x)P(x|\mathrm{NH})dx \\ &= \int_{-\infty}^{\infty} \frac{1}{1 + e^{-\Delta\chi^2/2}} P(x|\mathrm{NH})dx \\ &= \int_{-\infty}^{\infty} \frac{1}{1 + e^{-\Delta\chi^2/2}} P(\Delta\chi^2|\mathrm{NH})d\Delta\chi^2. \quad (17)\end{aligned}$$

Note that the first integral above involves calculating an N-dim integral, and the last one is of 1-dim only. The latter is much easier to obtain, an example of which will be presented in the next section.

(2) The fraction of measurements x that favor NH, i.e., the fraction of x such that $P(\mathrm{NH}|x) > 0.5$, is given by

TABLE II. Tabulated results of $\Delta\chi^2_{\alpha\sigma}$. For a given α, the one-sided p-value is $p_\alpha = P(Z \geq \alpha)$ (probability of $Z \geq \alpha$), where Z stands for a standard Gaussian random variable. The corresponding $\Delta\chi^2$ value is given by $\Delta\chi^2_{\alpha\sigma} = -2\log(p_\alpha/(1-p_\alpha))$.

α	0.475	1	1.281	1.645	2	3	4	5
One-sided p-value: p_α	31.74%	15.87%	10%	5%	2.28%	0.13%	3.2×10^{-5}	3.0×10^{-7}
$\Delta\chi^2_{\alpha\sigma}$	1.53	3.33	4.39	5.89	7.52	13.29	20.70	30.04

$$F_{T=\mathrm{NH}} = \int_{\{x:\ P(\mathrm{NH}|x)>0.5\}} P(x|\mathrm{NH})dx = \int_0^\infty P(\Delta\chi^2|\mathrm{NH})d\Delta\chi^2. \tag{18}$$

Here, "F" and the subscript "$T = \mathrm{NH}$" stand for fraction and the NH assumption, respectively.

If NH is the correct hypothesis, then a good experiment should have a high probability of producing data that not only favors NH but indeed provides substantial evidence for NH. Hence, it is useful to generalize the term in Eq. (18) to gauge the chance of $P(\mathrm{NH}|x) > 1-p$ for any threshold value $1-p$ of interest. In particular, physicists are familiar with thresholds associated with the so-called $\alpha\sigma$ level, with one-sided $\alpha\sigma$ corresponding to $1 - p_\alpha = 1 - P(Z \geq \alpha)$ for a standard Gaussian random variable Z.[3] Accordingly, define

$$F^{\alpha\sigma}_{T=\mathrm{NH}} = \int_{\{x:\ P(\mathrm{NH}|x)>1-p_\alpha\}} P(x|\mathrm{NH})dx = \int_{\Delta\chi^2_{\alpha\sigma}}^\infty P(\Delta\chi^2|\mathrm{NH})d\Delta\chi^2. \tag{19}$$

A list of common $\alpha\sigma$ values and the corresponding p_α, as well as $\Delta\chi^2_{\alpha\sigma} = -2\log(p_\alpha/(1-p_\alpha))$, are listed in Table II.

(3) In addition, probability intervals (PIs) for $P(\mathrm{NH}|x)$ also provide useful information. For example, a 90% PI is denoted by $(P^{90\%}_{T=\mathrm{NH}}, 1)$, where $P^{90\%}_{T=\mathrm{NH}}$ is the $100 - 90 = 10$th percentile of $P(\mathrm{NH}|x)$. That is, had NH been the truth, 90% of the potential data would yield $P(\mathrm{NH}|x)$ larger than $P^{90\%}_{T=\mathrm{NH}}$.

All the above criteria reflect the capability of the experiment to distinguish the two competing hypotheses, and they convey different messages.

Finally, to get a complete picture of the sensitivity of an experiment, one should also obtain the above metrics under the assumption that IH is the underlying true model. The sensitivity scores under metrics 2 and 3 can be shown to depend on the underlying true model. For example, we experimented with simple examples (not shown) and observed that in general $F_{T=\mathrm{NH}} \neq F_{T=\mathrm{IH}}$, whereas for metric 1, we have $\bar{P}^{\mathrm{NH}}_{T=\mathrm{NH}} = \bar{P}^{\mathrm{IH}}_{T=\mathrm{IH}}$ as long as equal prior probabilities[4] $P(\mathrm{NH}) = P(\mathrm{IH})$ were assigned to the two models. This is because

$$\begin{aligned}\bar{P}^{\mathrm{NH}}_{T=\mathrm{NH}} - \bar{P}^{\mathrm{IH}}_{T=\mathrm{IH}} &= \int P(\mathrm{NH}|x)P(x|\mathrm{NH}) - P(\mathrm{IH}|x)P(x|\mathrm{IH})dx \\ &= \int \frac{P^2(x|\mathrm{NH})P(\mathrm{NH}) - P^2(x|\mathrm{IH})P(\mathrm{IH})}{P(x|\mathrm{NH})P(\mathrm{NH}) + P(x|\mathrm{IH})P(\mathrm{IH})}dx \\ &= \int P(x|\mathrm{NH}) - P(x|\mathrm{IH})dx = 0.\end{aligned}$$

In the next section, we use an example to show how one can easily calculate the posterior probability and the sensitivity measurements introduced above. We also contrast the resulting sensitivity measurements to a commonly used quantity that is known as "$\overline{\Delta\chi^2}$ of the Asimov data set."

IV. ILLUSTRATION OF THE BAYESIAN APPROACH IN A CONSTRAINED PARAMETER SPACE

In this section, we consider a situation where θ can take on only two possible values, 1 and -1, which correspond to the hypotheses NH and IH, respectively. This simplified setting is motivated by the fact that existing measurements of $|\theta| = M^2_{32}$ are very accurate at around 2.43×10^{-3} eV^2, and we simply denote this value to 1 for clarity of presentation. It is a special case of the Bayesian treatment in the previous section, where $P(\theta|\mathrm{NH})$ and $P(\theta|\mathrm{IH})$ are assigned degenerate distributions at 1 and -1, respectively. That is, $P(\theta = 1|\mathrm{NH}) = P(\theta = -1|\mathrm{IH}) = 1$. Furthermore, there is no nuisance parameter η. As a result, the expected bin counts will be denoted by $\mu_i^{\mathrm{NH}} = \mu(1)$ and $\mu_i^{\mathrm{IH}} = \mu(-1)$, respectively.

Below, we showcase numerical calculations of various sensitivity criteria for this example. In particular, we introduce approximations that are simple functions of a term commonly known as $\overline{\Delta\chi^2}$ of the Asimov data set in the physics literature. According to the definition in Ref. [26], the Asimov data set under hypothesis MH is given by $x^{\mathrm{MH}} = (\mu_1^{\mathrm{MH}}, \ldots, \mu_N^{\mathrm{MH}})$, where $\mu_i^{\mathrm{MH}} = \mu_i(\theta_0^{\mathrm{MH}}, \eta_0^{\mathrm{MH}})$ and $(\theta_0^{\mathrm{MH}}, \eta_0^{\mathrm{MH}}) = \mathrm{argmax}_{(\theta,\eta)} P(\theta, \eta|\mathrm{MH})$ is the prior

[3] Another commonly used term is two-sided $\alpha\sigma$, which corresponds to $1 - P(|Z| \geq \alpha)$.

[4] We acknowledge the referee for pointing out this important relation.

mode under MH. In words, the Asimov data set is the most typical data set under the most likely parameter values based on prior knowledge subject to the given model.

Interestingly, $\overline{\Delta\chi^2}$ is itself often used as a measure of sensitivity. Here, we will contrast the typical usage of $\overline{\Delta\chi^2}$ to that of the sensitivity criteria developed in the previous section. More accurate evaluations of these sensitivity criteria are also attainable via MC methods.

Suppose that the proposed experiment will collect enough data such that the expected counts under NH and IH are much larger than the difference between them: $\mu_i^{\rm NH} \sim \mu_i^{\rm iH} \gg |\mu_i^{\rm NH} - \mu_i^{\rm iH}|$. Using the notations introduced in Sec. II, if the nature is NH, then the observed counts N_i can be represented as

$$N_i = \mu_i^{\rm NH} + \sqrt{\mu_i^{\rm NH}} \cdot g_i, \tag{20}$$

where $g_1, \ldots, g_n$ are mutually independent standard Gaussian random variables. Then, the statistic $\Delta\chi^2$ of Eq. (11) becomes

$$\Delta\chi^2_{T=\rm NH} = \sum_{i=1}^n \frac{(\mu_i^{\rm NH} - \mu_i^{\rm IH})^2}{\mu_i^{\rm IH}} + \sum_{i=1}^n 2\frac{(\mu_i^{\rm NH} - \mu_i^{\rm IH})\sqrt{\mu_i^{\rm NH}} g_i}{\mu_i^{\rm IH}} + \sum_{i=1}^n \frac{\mu_i^{\rm NH} - \mu_i^{\rm IH}}{\mu_i^{\rm IH}} g_i^2 - \sum_{i=1}^n \log\left(1 + \frac{\mu_i^{\rm NH} - \mu_i^{\rm IH}}{\mu_i^{\rm IH}}\right). \tag{21}$$

Here, the subscript $T = \rm NH$ indicates that the nature is NH. Since $\mu_i^{i\rm H} \gg |\mu_i^{\rm NH} - \mu_i^{i\rm H}|$, the summation of the last two terms in Eq. (21) is negligible as it is approximately $\sum_{i=1}^n \frac{\mu_i^{\rm NH} - \mu_i^{\rm IH}}{\mu_i^{\rm IH}} \cdot (g_i^2 - 1)$ by a Taylor expansion of the last term. Therefore, $\Delta\chi^2_{T=\rm NH}$ follows a Gaussian distribution, with mean and standard deviation, respectively,

$$\begin{cases} \overline{\Delta\chi^2} & \equiv \sum_{i=1}^n \frac{(\mu_i^{\rm NH} - \mu_i^{\rm IH})^2}{\mu_i^{\rm IH}}, \\ \sigma_{\Delta\chi^2} & \equiv 2\sqrt{\sum_{i=1}^n \frac{(\mu_i^{\rm NH} - \mu_i^{\rm IH})^2 \cdot \mu_i^{\rm NH}}{(\mu_i^{\rm IH})^2}} \\ & = 2\sqrt{\sum_{i=1}^n \left(\frac{(\mu_i^{\rm NH} - \mu_i^{\rm IH})^2}{\mu_i^{\rm IH}} + \frac{(\mu_i^{\rm NH} - \mu_i^{\rm IH})^3}{(\mu_i^{\rm IH})^2}\right)} \\ & \approx 2\sqrt{\overline{\Delta\chi^2}}. \end{cases} \tag{22}$$

In the last step, since $\mu_i^{\rm NH} - \mu_i^{\rm IH} \ll \mu_i^{\rm NH} \sim \mu_i^{\rm IH}$, we further neglect the term $\frac{(\mu_i^{\rm NH} - \mu_i^{\rm IH})^3}{(\mu_i^{\rm IH})^2}$. Similarly, it is straightforward to show that when the nature is IH, $\Delta\chi^2_{T=\rm IH}$ would follow an approximate Gaussian distribution with mean $= -\overline{\Delta\chi^2}$ and standard deviation $\sigma_{\Delta\chi^2}$. In fact, when IH is true, $\overline{\Delta\chi^2_{\rm IH}} = -\sum_{i=1}^n \frac{(\mu_i^{\rm NH} - \mu_i^{\rm IH})^2}{\mu_i^{\rm NH}} \approx -\overline{\Delta\chi^2}$.

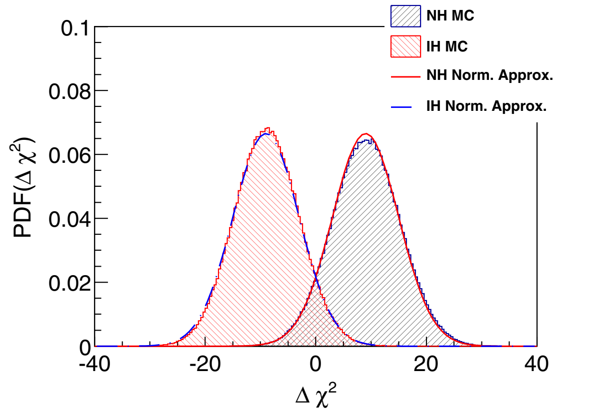

FIG. 2 (color online). The probability density functions $P(\Delta\chi^2|\rm NH)$ and $P(\Delta\chi^2|\rm IH)$ in the Bernoulli model are shown as the solid and dotted lines, respectively. The $|\overline{\Delta\chi^2}|$ is assumed to be 9.

To see how the above approximation works, we look at the example in Sec. II, where $\overline{\Delta\chi^2} \approx 9$. Figure 2 shows histograms (shaded area) based on large MC samples of $\Delta\chi^2$ under NH and IH, respectively. They agree very well with the analytical approximation (dashed lines) in Eq. (22).

Now, we are ready to calculate (i) the probability of a hypothesis post data collection and (ii) various measurements of sensitivity for an experiment concerning potential data generated from it.

First, given observed data $x = (N_1, \ldots, N_n)$, the probability $P(\rm NH|x)$ can be directly calculated from Eq. (7). Let $G(t; m, \sigma) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(t-m)^2}{2\sigma^2}}$ denote the pdf of a Gaussian random variable with mean m and standard deviation σ, evaluated at t, and then

$$\begin{aligned} P({\rm NH}|x) &= \frac{P(x|{\rm NH}) \cdot P({\rm NH})}{P(x|{\rm NH}) \cdot P({\rm NH}) + P(x|{\rm IH}) \cdot P({\rm IH})} \\ &= \frac{\Pi_i G\left(N_i; \mu_i^{\rm NH}, \sqrt{\mu_i^{\rm NH}}\right)}{\Pi_i G\left(N_i; \mu_i^{\rm NH}, \sqrt{\mu_i^{\rm NH}}\right) + \Pi_i G\left(N_i; \mu_i^{\rm IH}, \sqrt{\mu_i^{\rm IH}}\right)} \\ &= \frac{1}{1 + e^{-\Delta\chi^2(x)/2}}, \end{aligned}$$

where

$$\Delta\chi^2(x) = \sum_{i=1}^n \left[\log\frac{\mu_i^{\rm IH}}{\mu_i^{\rm NH}} + \frac{(N_i - \mu_i^{\rm IH})^2}{\mu_i^{\rm IH}} - \frac{(N_i - \mu_i^{\rm NH})^2}{\mu_i^{\rm NH}}\right].$$

We mention that, if one reduces the full data x to its function $\Delta\chi^2(x)$, then calculating $P(\rm NH|\Delta\chi^2)$ based on our approximation in Eq. (22) will recover $P(\rm NH|x)$:

$$P(\mathrm{NH}|\Delta\chi^2) = \frac{P(\Delta\chi^2|\mathrm{NH})\cdot P(\mathrm{NH})}{P(\Delta\chi^2)} = \frac{P(\Delta\chi^2|\mathrm{NH})}{P(\Delta\chi^2|\mathrm{NH}) + P(\Delta\chi^2|\mathrm{IH})} = \frac{G\left(\Delta\chi^2; \overline{\Delta\chi^2}, 2\sqrt{\overline{\Delta\chi^2}}\right)}{G\left(\Delta\chi^2; \overline{\Delta\chi^2}, 2\sqrt{\overline{\Delta\chi^2}}\right) + G\left(\Delta\chi^2; -\overline{\Delta\chi^2}, 2\sqrt{\overline{\Delta\chi^2}}\right)} = \frac{1}{1+e^{-\Delta\chi^2/2}}. \quad (23)$$

Next, we evaluate various sensitivity metrics of a future experiment, using again the Gaussian distribution for $\Delta\chi^2$ in Eq. (22):

$$\bar{P}^{\mathrm{NH}}_{T=\mathrm{NH}} \approx \int_{-\infty}^{\infty} \frac{1}{1+e^{-t/2}} G\left(t, \overline{\Delta\chi^2}, 2\sqrt{\overline{\Delta\chi^2}}\right) dt \equiv \bar{P}(\overline{\Delta\chi^2}), \quad (24)$$

$$F_{T=\mathrm{NH}} \approx \int_0^{\infty} G\left(t; \overline{\Delta\chi^2}, 2\sqrt{\overline{\Delta\chi^2}}\right) dt = \frac{1}{2}\left(1 + \mathrm{erf}\left(\sqrt{\frac{\overline{\Delta\chi^2}}{8}}\right)\right), \quad (25)$$

$$F^{\alpha\sigma}_{T=\mathrm{NH}} \approx \int_{\Delta\chi^2_{\alpha\sigma}}^{\infty} G\left(t; \overline{\Delta\chi^2}, 2\sqrt{\overline{\Delta\chi^2}}\right) dt = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\overline{\Delta\chi^2} - \Delta\chi^2_{\alpha\sigma}}{\sqrt{8\overline{\Delta\chi^2}}}\right)\right), \quad (26)$$

$$P^{A\%}_{T=\mathrm{NH}} \approx 1/(1 + e^{-\frac{1}{2}(\overline{\Delta\chi^2} - 2z^*_A\sqrt{\overline{\Delta\chi^2}})}). \quad (27)$$

In Eq. (24) above, $\bar{P}^{\mathrm{NH}}_{T=\mathrm{NH}}$ was approximated by $\bar{P}(\overline{\Delta\chi^2})$, which is a function of $\overline{\Delta\chi^2}$ only. In Eq. (27), z^*_A represents the Ath percentile of a standard Gaussian distribution; hence, $\overline{\Delta\chi^2} - 2z^*_A\sqrt{\overline{\Delta\chi^2}}$ is the $(100-A)$th percentile of $\Delta\chi^2$ according to the Gaussian approximation in Eq. (22). Since $P(\mathrm{NH}|\Delta\chi^2) = 1/(1+e^{-\Delta\chi^2/2})$ is increasing in $\Delta\chi^2$, this means that the right-hand side of Eq. (27) is the $(100-A)$th percentile of $P(\mathrm{NH}|\Delta\chi^2)$, which serves as the lower bound of the $A\%$ PI proposed in the previous section. In Table III, we list z^*_A for a few typical choices of probability intervals, assuming that the nature is NH.

TABLE III. Tabulated $\Delta\chi^2_{\mathrm{PI}}$ values for a few typical choice of probability intervals, assuming that the nature is NH.

$A\%$	68%	90%	95%	99%
Gaussian percentile z^*_A	0.468	1.282	1.645	2.326

For the example experiment used in the simulation of Sec. II, its $\overline{\Delta\chi^2} = 9$. Had one followed the common practice that directly compares $\sqrt{\overline{\Delta\chi^2}}$ to the quantiles of a Gaussian distribution, one would report the "specificity" of the experiment to be 99.87% (1 "one-sided p-value"). In contrast, we obtained various sensitivity metrics for the experiment according to Eqs. (24)–(27) and listed them in Table IV. First, assuming the Asimov data set is observed, we have $P(\mathrm{NH}|x^{\mathrm{NH}}) \approx P(\mathrm{IH}|x^{\mathrm{IH}}) \approx P(\mathrm{NH}|\Delta\chi^2 = 9) = 98.90\%$. Second, we calculated $\bar{P}^{\mathrm{NH}}_{T=\mathrm{NH}} = \bar{P}^{\mathrm{IH}}_{T=\mathrm{IH}} \approx \bar{P}(\overline{\Delta\chi^2} = 9) = 90.14\%$. That is, the average posterior probability for NH (or IH) when it is indeed the correct hypothesis is only about 90%, which is much lower than its Asimov counterpart of $P(\mathrm{NH}|\Delta\chi^2 = 9) = 98.90\%$. Third, the fraction $F_{T=\mathrm{NH}} = 93.32\%$ of potential data sets would yield a $\Delta\chi^2$ that favors NH. And to contrast with the Gaussian interpretation, we calculated that only $F^{3\sigma}_{T=\mathrm{NH}} = 23.73\%$ of potential data sets would yield a $\Delta\chi^2$ above 9 or, say, yield evidence as strong as $P(\mathrm{NH}|x) \geq 99.87\%$. Furthermore, the left panel of Fig. 3 displays the distribution (vertical axis in log scale) of $P(\mathrm{NH}|x) = P(\mathrm{NH}|\Delta\chi^2)$. The two vertical dashed lines show that 68% of potential data sets will result in $P(\mathrm{NH}|x) > 95.67\%$, whereas 90% of potential data sets will result in $P(\mathrm{NH}|x) > 65.79\%$.

Moving forward from a fixed $\overline{\Delta\chi^2}$ value, we next study how the various sensitivity metrics compare to each other for experiments with different $\overline{\Delta\chi^2}$ values. The right panel of Fig. 3 displays the lower bound of the 90% probability interval $P^{90\%}_{T=\mathrm{NH}}$, the average probability $\bar{P}^{\mathrm{NH}}_{T=\mathrm{NH}}$, and the Gaussian interpretation based on the one-sided p-value as functions of $\overline{\Delta\chi^2}$. Note that we plotted 1 minus the aforementioned metrics in order to zoom in the high probability regions. Interestingly, the line of average probability $\bar{P}$ yields a higher value than the lower bound of 90% P.I. for $\overline{\Delta\chi^2} < \sim 18$ and yields a lower value than the lower bound of 90% P.I. for $\overline{\Delta\chi^2} > \sim 18$. Such behavior is natural given the definition of each curve. Nevertheless, both curves are much higher than the Gaussian interpretation,

TABLE IV. Sensitivity metrics for an experiment with $\overline{\Delta\chi^2} = 9$.

Symbol Description	$\bar{P}$ Average	Gaussian interpretation	$P(\mathrm{NH}\|x)$ Asimov data set	$F_{T=\mathrm{NH}}$ $\Delta\chi^2 > 0$	$F^{3\sigma}_{T=\mathrm{NH}}$ $P > 99.87\%$	$P^{68\%}_{T=\mathrm{NH}}$ 68% P.I.	$P^{90\%}_{T=\mathrm{NH}}$ 90% P.I.
Sensitivity metric	90.14%	99.87%	98.90%	93.32%	23.73%	95.67%	65.79%

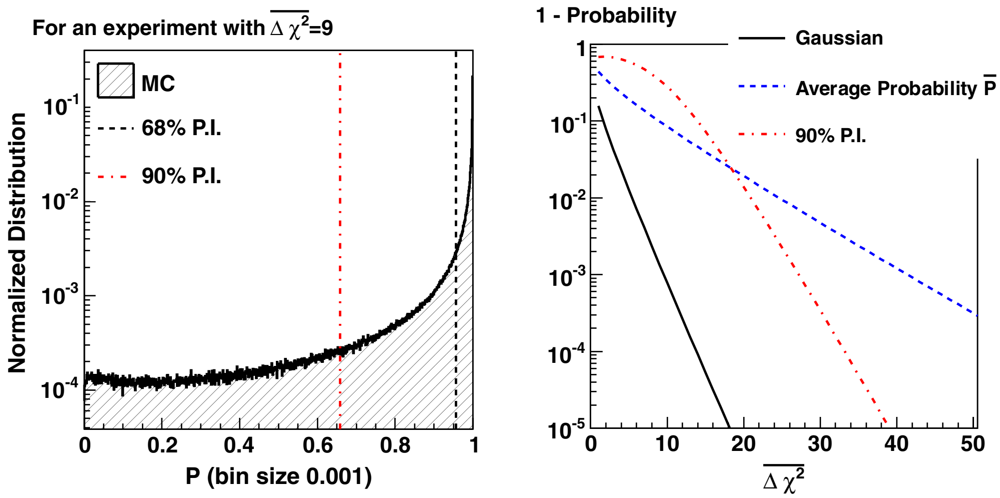

FIG. 3 (color online). The left panel shows the distribution of $P(\mathrm{NH}|x) = P(\mathrm{NH}|\Delta\chi^2)$ over the population of potential data x that arises from an experiment with $\overline{\Delta\chi^2} = 9$ where the truth is NH. The mean of this distribution is 90.14%. The lower bounds of the 68% and 90% probability intervals are plotted. That is, 68% (90%) of the data x would yield a $P(\mathrm{NH}|x)$ that falls to the right of the dash-dotted (dashed) line. These two lines are also commonly referred as the 32nd and the 10th percentile, respectively. The right panel plots several sensitivity metrics (subtracted from 1 for clarity), against $\overline{\Delta\chi^2}$ that ranges from 1 to 50. Note that all the lines are decreasing because higher values of $\overline{\Delta\chi^2}$ correspond to more sensitive experiments. This is done for three different criteria: the Gaussian interpretation (derived from the one-sided p-value with 1 degree of freedom), $\bar{P}$, and $P^{90\%}_{T=\mathrm{NH}}$. The Gaussian interpretation is seen to be overoptimistic in describing the ability of the experiment to differentiate the two hypotheses.

suggesting that the Gaussian interpretation is overoptimistic in describing the ability of an experiment to differentiate NH and IH.

V. DISCUSSIONS

A couple of comments should be made regarding the $\sqrt{\overline{\Delta\chi^2}}$ representation for sensitivity in determining the MH.

(1) We have seen that the distribution of the best estimator of $\theta = \Delta m^2_{32}$ is closer to a Bernoulli distribution than to a Gaussian distribution. Therefore, Wilks' theorem is not applicable, and direct interpretation of $\sqrt{\Delta\chi^2_{\min}}$ as the number of σ in the Gaussian approximation leads to incorrect confidence intervals. We provided an analytical formula [Eq. (6)] for the confidence interval in an ideal Bernoulli case, which can be used to generate approximate confidence intervals for similar cases. For more general cases, a full MC simulation is needed to construct confidence intervals, as advocated in Ref. [23].

(2) Even if a confidence interval for Δm^2_{32} is constructed correctly, its confidence level cannot be directly interpreted as how much the current measurement would favor the NH (IH) against the other. Despite possible agreement between confidence intervals and Bayesian credible intervals under certain circumstances as discussed in Appendix. B, such agreement does not apply to the current MH problem where there are strong constraints imposed on M^2_{32}.

Additional comments should be made regarding the Bayesian approach.

(1) In principle, results from different experiments can be combined within the Bayesian framework. One example can be found in Ref. [37], in which a Bayesian method was applied to constrain θ_{13} and CP phase δ with existing experimental data. Regarding the MH, results from different experiments can be combined through the integral in Eq. (9). Specifically, one can integrate over the nuisance parameters regarding experimental systematic uncertainties while leaving nuisance parameters regarding the relevant neutrino masses and mixing parameters unintegrated. For example, suppose there are two independently conducted experiments, labeled by $j = 1, 2$, and that their respective observed data x_j corresponds to the model $P(x_j|\theta, \eta^*, \eta_j, \mathrm{MH})$ under MH = NH or IH. Here the vector of nuisance parameter η in experiment j is separated into two pieces η^* and η_j, where η_j is unique to the experiment and η^* is common to both experiments. Of course, θ is the parameter of interest and hence always common to both. Then, it would be useful for the different experiments not only to present $\Delta\chi^2$ [Eq. (11)] but to also present

$$P(x_j|\theta, \eta^*, \mathrm{MH}) = \int P(x_j|\theta, \eta^*, \eta_j, \mathrm{MH}) \times P(\eta_j|\theta, \eta^*, \mathrm{MH}) d\eta_j,$$

in order that one can calculate the overall likelihood $P(x_1, x_2|\theta, \eta^*, \mathrm{NH}) = \Pi_{j=1}^{2} P(x_j|\theta, \eta^*, \mathrm{NH})$ for further inferences.

(2) We have listed a few different metrics to represent sensitivity of future experiments in Sec. III. Each of them conveys different information. In the case that one has to choose a single number to summarize the experiment sensitivity, one convenient choice would be $\bar{P} \equiv \bar{P}_{T=\mathrm{NH}}^{\mathrm{NH}} = \bar{P}_{T=\mathrm{IH}}^{\mathrm{IH}}$, the average probability reported for the true underlying model. For all other metrics that were introduced, the sensitivity scores need to be calculated separately by assuming NH or IH is the true model.

(3) For general models where nuisance parameters are present, it is possible to measure the specificity of an experiment conditional on different possible values of the nuisance parameters. For instance, suppose NH and that a particular value of the nuisance parameter, say, $\eta = \eta_0$, is true. Then the relevant population of potential experimental results consists of x generated from $P(x|\mathrm{NH}, \eta_0) = \int P(x|\theta, \eta_0, \mathrm{NH}) P(\theta, \eta_0|\mathrm{NH}) d\theta$. Accordingly, $P(\mathrm{NH}|x)$ can be obtained for each x in this population with Eq. (8),[5] and for, e.g., their mean $\bar{P}_{T=\mathrm{NH}}^{\mathrm{NH}}(\eta_0)$ and quantiles $P_{T=\mathrm{NH}}^{A}(\eta_0)$ serve as more refined sensitivity metrics for the experiment, and can be plotted against a range of possible η_0 values. Such an application is particularly useful when the separation of MH strongly depends on the value of η. One such example is long baseline ν_e or $\bar{\nu}_e$ appearance measurements (from the ν_μ or $\bar{\nu}_\mu$ beam), in which the sensitivity of MH strongly depends on the value of the *CP* phase of lepton section δ_{CP} and neutrino mixing angle θ_{23}.

(4) The Gaussian approximation in Eq. (22) allows analytical calculation of various sensitivity metrics. Be aware that such calculations are valid under the assumption that the possible range of θ under either hypothesis is narrow enough that it can be reasonably represented by a single point and that $\mu_i^{\mathrm{NH}} - \mu_i^{\mathrm{IH}} \ll \mu_i^{\mathrm{NH}} \sim \mu_i^{\mathrm{IH}}$. For more general cases, numerical such as MC methods are needed.

(5) Finally, we emphasize that sensitivity metrics are designed to evaluate an experiment in its planning stage. It can be used to see if an experiment with a proposed sample size, i.e., the expected bin counts $\{\mu_i, i = 1, \ldots, n\}$, will be large enough to have a high probability of generating desired strength of evidence to support the true hypothesis. But once the data are observed, the calculation of sensitivity metrics is no longer relevant. One should clearly differentiate results deduced from data from that from the sensitivity calculations.

[5]One should not take into account the information of $\eta = \eta_0$ in calculating the probability, since one does not know the true value of η when analyzing experimental data.

VI. SUMMARY

In this paper, we perform a statistical analysis for the problem of determining the neutrino mass hierarchy. A classical method of presenting experimental results is examined. Such a method produces confidence intervals through the parameter estimation of Δm_{32}^2 based on approximating the distribution of $\sqrt{\Delta\chi^2}$ as the standard Gaussian distribution. However, due to strong existing experimental constraints of $M_{32}^2 \equiv |\Delta m_{32}^2|$, the parent distribution of the best estimation of Δm_{32}^2 is better approximated as a Bernoulli distribution rather than a Gaussian distribution, which leads to a very different estimation of the confidence level. The importance of using the Feldman-Cousins approach to determine the confidence interval is emphasized.

In addition, the classical method is shown to be inadequate to convey the message of how much results from an experiment favor one hypothesis over the other, as the agreement between the confidence interval and the Bayesian credible interval also breaks down due to the constraints on M_{32}^2.

We therefore introduce the Bayesian approach to quantify the probability of MH. We further extend the discussion to quantify experimental sensitivities of future measurements.

ACKNOWLEDGMENTS

We thank Petr Vogel, Haiyan Gao, Jianguo Liu, Alan Gelfand, and Laurence Littenberg for fruitful discussions and careful reading. This work was supported in part by Caltech, the National Science Foundation, and the Department of Energy under Contracts No. DE-AC05-06OR23177, under which Jefferson Science Associates, LLC, operates the Thomas Jefferson National Accelerator Facility, and No. DE-AC02-98CH10886.

APPENDIX A: DERIVATION OF $P(\Delta\chi^2_{\mathrm{min}})$ FOR CASE II: $\Theta = \{-1, 1\}$

Let θ_0 denote the true parameter value from which the data are generated. Under case II, when $\theta_0 = 1$, the statistic $\Delta\chi^2_{\mathrm{min}}(\theta_0)$ in Eq. (3) is directly related to $\Delta\chi^2$ in Eq. (21) (recall that the notation $\theta = 1, -1$ refers to NH and IH, respectively) as $\Delta\chi^2_{\mathrm{min}}(1) = \max\{0, -\Delta\chi^2\}$. The result in Sec. IV implies that, under $\theta_0 = 1$, $-\Delta\chi^2$ follows an approximately Gaussian distribution with mean $-\overline{\Delta\chi^2}$ and standard deviation $2\sqrt{\overline{\Delta\chi^2}}$. Similarly, when $\theta_0 = -1$, the statistic $\Delta\chi^2_{\mathrm{min}}(\theta_0) = \max\{0, \Delta\chi^2\}$, where $\Delta\chi^2$ follows approximately Gaussian distribution with mean $\overline{\Delta\chi^2}$ and standard deviation $2\sqrt{\overline{\Delta\chi^2}}$. Therefore, whether the truth is θ_0 is 1 or -1, the distribution of $\Delta\chi^2_{\min}(\theta_0)$ is such that $P(\Delta\chi^2_{\min}(\theta_0) \geq t) = 1$ for $t \leq 0$ and that $P(\Delta\chi^2_{\min}(\theta_0) \geq t) \approx \frac{1}{2} - \frac{1}{2}\operatorname{erf}(\frac{t+\overline{\Delta\chi^2}}{\sqrt{8\overline{\Delta\chi^2}}})$ for $t > 0$.

APPENDIX B: CONFIDENCE INTERVAL VERSUS BAYESIAN CREDIBLE INTERVAL

As emphasized in Ref. [23], the classical confidence interval should not be confused with the Bayesian credible interval. However, it is rather common that physicists approximate the confidence interval as the Bayesian credible interval, especially in MC simulations, where previous measurements of some physics quantities are used as inputs. Such approximations turn out to be acceptable under the following condition.

Consider the condition that the pdf (or probability mass function) of the best estimation of the unknown parameter $\theta_{\min}$ only depends on its relative location with respect to the true parameter value, that is,

$$P_{\Theta_{\min}|\Theta_{\text{true}}}(\theta_{\min}|\theta_{\text{true}}) = h(\theta_{\min} - \theta_{\text{true}}), \tag{B1}$$

for some non-negative function h such that $\int_{-\infty}^{\infty} h(t)dt = 1$. Models that satisfy Eq. (B1) are said to belong to a location family, where θ_{true} is called the location parameter. When there is a lack of strong prior information for θ_{true}, it is usually reasonable to assign a uniform prior for it, that is, to assign $P_{\Theta_{\text{true}}}(\theta_{\text{true}}) \propto 1$. If so, we have

$$\begin{aligned}
&P_{\Theta_{\text{true}}|\Theta_{\min}}(\theta_{\text{true}}|\theta_{\min})\\
&\quad = P_{\Theta_{\min}|\Theta_{\text{true}}}(\theta_{\min}|\theta_{\text{true}})P_{\Theta_{\text{true}}}(\theta_{\text{true}})/P_{\Theta_{\min}}(\theta_{\min})\\
&\quad \propto P_{\Theta_{\min}|\Theta_{\text{true}}}(\theta_{\min}|\theta_{\text{true}})P_{\Theta_{\text{true}}}(\theta_{\text{true}}) \quad (\text{as a function of } \theta)\\
&\quad \propto P_{\Theta_{\min}|\Theta_{\text{true}}}(\theta_{\min}|\theta_{\text{true}}) = h(\theta_{\min} - \theta_{\text{true}}).
\end{aligned}$$

In the above, the first step follows from Bayes' theorem, and the third step incorporates the uniform prior on θ_{true}. Since for any fixed θ_{true}, $\int_{-\infty}^{\infty} h(\theta_{\min} - \theta_{\text{true}})d\theta_{\min} = 1$, the above indeed implies that

$$P_{\Theta_{\text{true}}|\Theta_{\min}}(\theta_{\min}|\theta_{\text{true}}) = h(\theta_{\min} - \theta_{\text{true}}). \tag{B2}$$

For any threshold level c and the observed value of $\theta_{\min}$, define a plausible region for θ_{true} by $A(\theta_{\min}, c) = \{\theta: P_{\Theta_{\min}|\Theta_{\text{true}}}(\theta_{\min}|\theta) > c\}$, and then

$$A(\theta_{\min}, c) = \{\theta: h(\theta_{\min} - \theta) > c\} = \{\theta_{\min} + t: h(0 - t) > c\} = \theta_{\min} + A(0, c), \tag{B3}$$

where the transformation $t = \theta - \theta_{\min}$ is used in step 2 and, in general, the notation $\alpha + A$ for a point α and a set A represents the set that consists of points $\alpha + a$ for all $a \in A$. In words, Eq. (B3) says that the plausible regions based on different $\theta_{\min}$ with a fixed threshold c are simply shifts in location of each other. First, under the Bayes framework, $A(\theta_{\min}, c)$ can be considered as a credible region (most often an interval). The probability that θ falls in $A(\theta_{\min}, c)$ is called the level of the credible region and is given by

$$\begin{aligned}
P_{\Theta_{\text{true}}|\Theta_{\min}}(\theta \in A(\theta_{\min}, c)|\theta_{\min}) &= \int_{A(\theta_{\min},c)} P_{\Theta_{\text{true}}|\Theta_{\min}}(\theta|\theta_{\min})d\theta\\
[\text{by Eqs. (B2)} \quad \text{and} \quad (\text{B3})] &= \int_{\theta_{\min}+A(0,c)} h(\theta - \theta_{\min})d\theta\\
(\text{letting } t = \theta - \theta_{\min}) &= \int_{A(0,c)} h(t)dt.
\end{aligned}$$

On the other hand, under the classical framework, $A(\theta_{\min}, c)$ serves as a confidence interval, the level of which is given by

$$\begin{aligned}
&P_{\Theta_{\min}|\Theta_{\text{true}}}(\theta \in A(\theta_{\min}, c)|\theta)\\
[\text{by Eq. (B3)}] &= P_{\Theta_{\min}|\Theta_{\text{true}}}(\theta \in \theta_{\min} + A(0, c)|\theta)\\
&= P_{\Theta_{\min}|\Theta_{\text{true}}}(\theta_{\min} \in \theta - A(0, c)|\theta)\\
[\text{by Eq. (B1)}] &= \int_{\theta - A(0,c)} h(\theta_{\min} - \theta)d\theta_{\min}\\
(\text{letting } t = \theta_{\min} - \theta) &= \int_{A(0,c)} h(t)dt.
\end{aligned}$$

In summary, the region $A(\theta_{\min}, c)$ can be interpreted as both a confidence interval and a credible region of the same level.

A most useful special case where Eq. (B1) is satisfied is the case where $\theta_{\rm min}$ strictly follows a Gaussian distribution with mean $\theta_{\rm true}$ (such as case I of Sec. II) and that the standard deviation of the Gaussian distribution did not depend on $\theta_{\rm true}$. As we mentioned in Sec. II, it is shown by Wilks [34] that, based on a large data sample size, the statistic $\theta_{\rm min}$ does approximately follow a Gaussian distribution with mean at $\theta_{\rm true}$ under certain regular conditions. Hence, it is not unacceptable to construct an α level confidence interval and interpret it as an α level credible interval, as long as the standard deviation of the Gaussian distribution has weak or no dependence on $\theta_{\rm true}$.

However, in the MH determination problem, the regularity conditions are violated due to the existing experimental constraints on $|\theta| = M^2_{32}$. As a result, condition Eq. (B1) is far from being satisfied, and there is no longer a correspondence between confidence intervals and Bayesian credible intervals. Indeed, strong inconsistency between implications of the two types of intervals can be seen from the following specific example belonging to case II of Sec. II. It is easy to come up with an observed data x that results in $\Delta\chi^2 = 1$ and $\overline{\Delta\chi^2} = 9$ [defined in Eqs. (11) and (22), respectively]. Then, according to the Bayesian approach, the probability is about 62.2% that NH is the correct hypothesis, or an odds of 5:3 of NH against IH. Most people would consider this a fairly weak preference for NH. On the other hand, the classical estimation procedure turns out to exclude the point IH from the 95% confidence interval according to (the correct) Table I. Had one attempted to interpret this 95% confidence interval as a Bayesian credible interval, one would conclude that the odds of NH against IH is at least 19:1. This conclusion is overconfident in the MH determination compared to the odds of 5:3 suggested by the well-founded Bayesian approach.

[1] F. P. An *et al.*, Phys. Rev. Lett. **108**, 171803 (2012).
[2] F. P. An *et al.*, Chin. Phys. C **37**, 011001 (2013).
[3] X. Qian (on behalf of the Daya Bay Collaboration), arXiv:1211.0570.
[4] F. P. An, Nucl. Instrum. Methods Phys. Res., Sect. A **685**, 78 (2012).
[5] K. Abe *et al.*, Phys. Rev. Lett. **107**, 041801 (2011).
[6] P. Adamson *et al.*, Phys. Rev. Lett. **107**, 181802 (2011).
[7] Y. Abe *et al.*, Phys. Rev. Lett. **108**, 131801 (2012).
[8] J. K. Ahn *et al.*, Phys. Rev. Lett. **108**, 191802 (2012).
[9] K. Abe *et al.*, Nucl. Instrum. Methods Phys. Res., Sect. A **659**, 106 (2011).
[10] NOνA experiment: http://www-nova.fnal.gov.
[11] T. Akiri *et al.*, arXiv:1110.6249.
[12] Super Kamiokande experiment: http://www-sk.icrr.u-tokyo.ac.jp/sk/index-e.html.
[13] MINOS experiment: http://www-numi.fnal.gov/Minos/.
[14] E. K. Akhmedov, S. Razzaque, and A. Y. Smirnov, arXiv:1205.7071.
[15] INO experiment: http://www.ino.tifr.res.in/ino.
[16] S. Choubey, S. T. Petcov, and M. Piai, Phys. Rev. D **68**, 113006 (2003).
[17] J. G. Learned, S. T. Dye, S. Pakvasa, and R. C. Svoboda, Phys. Rev. D **78**, 071302(R) (2008).
[18] L. Zhan, Y. F. Wang, J. Cao, and L. J. Wen, Phys. Rev. D **78**, 111103(R) (2008).
[19] X. Qian *et al.*, arXiv:1208.1551.
[20] E. Ciuffoli, J. Evslin, and X. M. Zhang, arXiv:1209.2227.
[21] R. D. McKeown and P. Vogel, Phys. Rep. **394**, 315 (2004).
[22] K. Nakamura *et al.*, J. Phys. G **37**, 075021 (2010).
[23] G. J. Feldman and R. D. Cousins, Phys. Rev. D **57**, 3873 (1998).
[24] J. Bernabeu *et al.*, arXiv:1005.3146.
[25] M. Blennow and T. Schwetz, J. High Energy Phys. 08 (2012) 058; 11 (2012) 098(E).
[26] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Eur. Phys. J. C **71**, 1554 (2011).
[27] X. H. Guo *et al.*, arXiv:hep-ex/0701029.
[28] F. Ardellier *et al.*, arXiv:hep-ex/0606025.
[29] J. K. Ahn *et al.*, arXiv:1003.1391.
[30] P. Huber and T. Schwetz, Phys. Rev. D **70**, 053011 (2004).
[31] P. Huber, M. Lindner, and W. Winter, Comput. Phys. Commun. **167**, 195 (2005).
[32] P. Huber, J. Kopp, M. Lindner, M. Rolinec, and W. Winter, Comput. Phys. Commun. **177**, 432 (2007).
[33] T. Schwetz, Phys. Lett. B **648**, 54 (2007).
[34] S. S. Wilks, Ann. Math. Stat. **9**, 60 (1938).
[35] A. W. Van der Vaart, *Asymptotic Statistics* (University of Cambridge, Cambridge, England, 1998).
[36] R. E. Kass and A. E. Raftery, J. Am. Stat. Assoc. **90**, 773 (1995).
[37] J. Bergstrom, J. High Energy Phys. 08 (2012) 163.