Why isn't every physicist a Bayesian?

Robert D. Cousins
Department of Physics, University of California, Los Angeles, California 90024-1547

(Received 1 June 1994; accepted 3 November 1994)

Physicists embarking on seemingly routine error analyses are finding themselves grappling with major conceptual issues which have divided the statistics community for years. While the philosophical aspects of the debate may be endless, a practicing experimenter must choose a way to report results. The results can depend on which of the two major frameworks, classical or Bayesian, one adopts. This article reviews reasons why most data analysis in particle physics has traditionally been carried out within the classical framework, and why this will probably continue to be the case. However, Bayesian reasoning has recently made significant inroads in some published work in this field, and many other particle physicists may frequently think in a Bayesian manner without realizing it. I illustrate the issues involved with a few simple, commonly encountered examples which reveal how each framework can sometimes lead to unsatisfying results. © *1995 American Association of Physics Teachers.*

I. INTRODUCTION

In recent years, many particle physicists[1] have become increasingly aware of the deep conceptual conflicts in statistical inference, and of why these conflicts cannot always be dismissed as philosophical arcana: published experimental results can depend on which of the two statistical frameworks, classical or Bayesian,[2] one adopts. The lurking controversy can come as a shock to a graduate student who encounters a statistical problem at some late stage in writing up the Ph.D. dissertation. Upon asking colleagues what statistical method to use, he or she will frequently get a classical answer having its roots in a few influential reference texts[3] or unpublished lecture notes.[4,5] However, the persistent student may discover that a "standard" classical technique has an undesirable feature, and that a Bayesian method can be more attractive in certain contexts. In fact, the statistical methods section of the Particle Data Group's desk reference[6] has recognized some Bayesian reasoning over the last few years.

Many particle physicists have therefore been motivated to learn more about the logical and philosophical foundations of statistical methods. Engineers[7] and physicists[8] from other specialties have been vocal as well, e.g., a Nobel Laureate in condensed matter theory who decreed matter-of-factly that Bayesian statistics "are the correct way to do inductive reasoning from necessarily imperfect experimental data."[9] Still, pure unabashed Bayesians are an almost invisible minority in particle physics publications.[10,11] In the spirit of the 1986 article, "Why isn't everyone a Bayesian?," by statistics researcher B. Efron,[12] one can ask this question of physicists. In this article I try to answer this question from the point of view of particle physicists, amidst whom most of my experience and discussions have taken place.

I restrict this article primarily to the discussion of confidence intervals, i.e., what physicists call either the "error" on a measurement, or an "upper limit" in the case of a null result. I follow common practice in referring to these intervals generically as "confidence intervals" regardless of whether the construction is classical or Bayesian, even though the phrase was originally coined to distinguish the classical construction. (Bayesian alternative appellations such as "credible intervals" are not used in particle physics.) A more complete discussion would include the *decision*-making process, but I do not attempt to cover that here. Confidence intervals are the way particle physicists normally report results, and there is reasonable consensus in the field regarding appropriateness of various methods. Decision theory remains much less formally defined in particle physics, even though a crucial part of one's work consists of deciding which experimental results to believe.

In Sec. II, I use the simple Gaussian example to outline the dichotomy between classical and Bayesian methods. I note the differing roles of the likelihood function: classically, it provides a handy means for computing *approximate* confidence intervals, while for Bayesian intervals, it has a more fundamental role. Since the concept of frequentist *coverage* is so central to classical confidence intervals, I devote Sec. III to its explication. In Sec. IV, I discuss the case in which a measured quantity is known in advance to be constrained to a physical region, and how unhappiness with classical results in this case has led to the acceptance of a Bayesian technique. Section V uses the common particle physics case of Poisson statistics to emphasize the difference between Bayesian and classical methods. Stark contrasts follow from the asymmetric nature of the distribution and the fact that the parameter is continuous while the observed values are discrete. This section also highlights the thorny problem of specifying an uninformative prior density in Bayesian statistics. I then describe in Sec. VI a common case where purely classical statistics leads to "unacceptable" behavior when combining the Gaussian and Poisson statistics of the earlier sections. I conclude in Sec. VII with a discussion of the merits and prospects of classical and Bayesian methods in particle physics.

II. DEFINITIONS AND SIMPLEST EXAMPLE

I begin with a simple familiar problem in order to define terms.[13] Suppose a measurement of the mass m of an elementary particle yields the value m_0, and it is known that the measuring apparatus yields values normally distributed about the unknown true mass m_t, with a known rms deviation σ_m. Also assume in this section that m_0 is many rms deviations above zero. The probability density $P(m|m_t)$ for obtaining the value m given the true mass m_t is[14]

$$P(m|m_t)=N(m_t,\sigma_m)=\frac{1}{\sqrt{2\pi\sigma_m^2}}\,e^{-(m-m_t)^2/2\sigma_m^2}. \qquad (1)$$

We wish to construct a confidence interval (m_1,m_2) at a

 © 1995 American Association of Physics Teachers

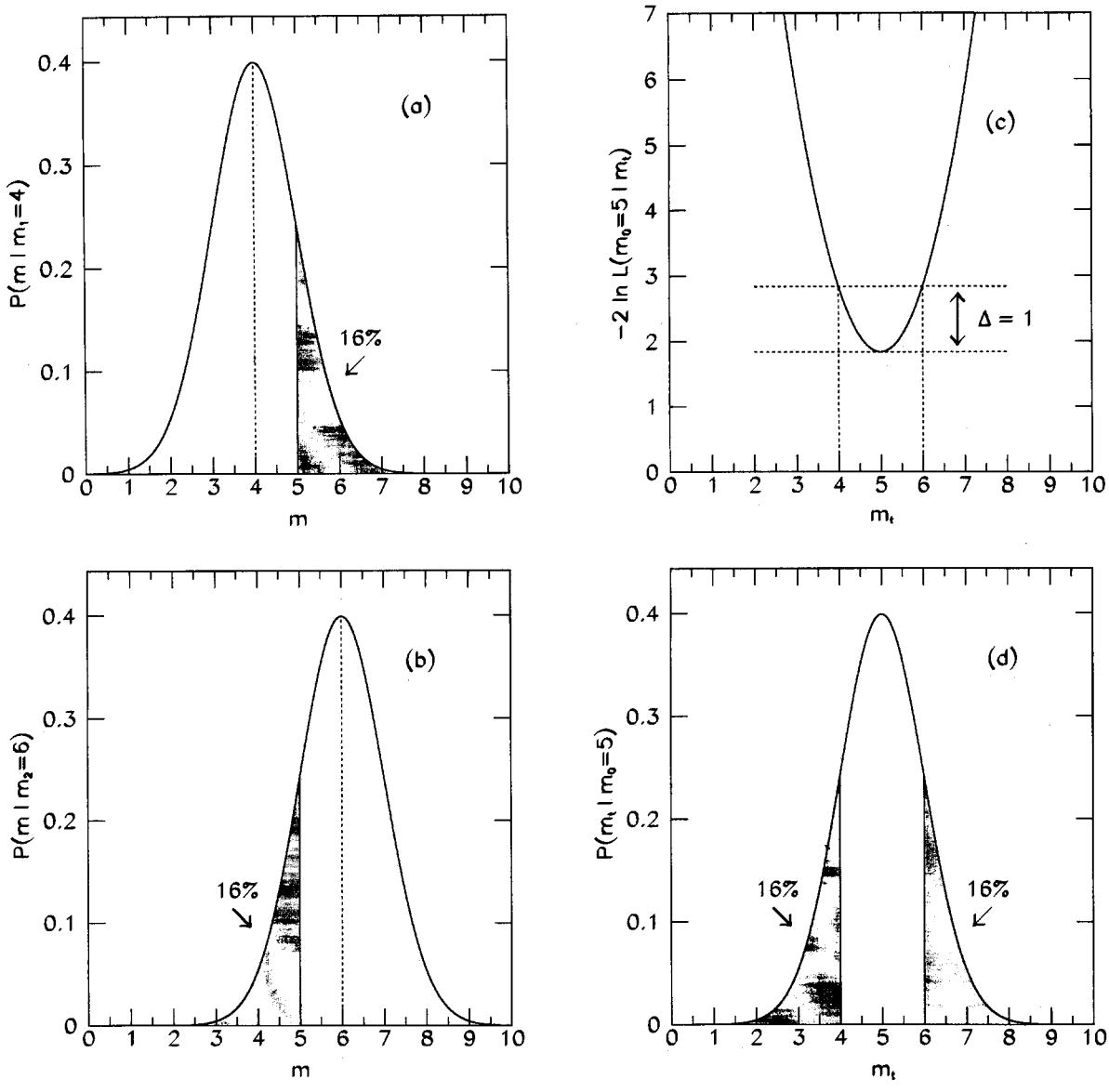

Fig. 1. Methods for 68% C.L. confidence interval construction in case of Gaussian pdf with known $\sigma_m=1.0$, after single measurement yields $m_0=5.0$. (a) and (b) Classical construction of left and right interval endpoints; (c) increase of $-2\ln\mathscr{L}$ by one unit (d) Bayesian with uniform prior.

specified confidence level (C.L.), which I take to be 68% in this section.[15]

Classical confidence intervals are those based on the method outlined in the famous 1937 paper by Neyman.[16] They have the property that no matter what m_t is, 68% of the intervals calculated by an ensemble of experiments will contain m_t. This defining property of "frequentist coverage" is emphasized in Sec. III. One classical construction of a (central) 68% C.L. confidence interval (m_1,m_2) proceeds by finding $m_1<m_0$ such that 16% of the area under $N(m_1,\sigma_m)$ is at values of m greater than m_0 [Fig. 1(a)], and by finding $m_2>m_0$ such that 16% of the area under $N(m_2,\sigma_m)$ is at values of m smaller than m_0 [Fig. 1(b)]. One obtains $m_1=m_0-\sigma_m$ and $m_2=m_0+\sigma_m$, and by common convention one states that the measured mass is "$m_0\pm\sigma_m$."

Many people do not think about the problem this way. Instead, they mentally construct a normal curve *centered on the measured value* m_0. This is usually justified in terms of likelihood functions, which provide the uneasy common ground (and hence one source of confusion) between the classical and Bayesian methods.

The likelihood $\mathscr{L}(m_0|m_t)$ is given by the same expression as Eq. (1) with the important change in point of view: by writing $\mathscr{L}$ instead of P we draw attention to the fact that we are considering its behavior for different values of m_t, given the particular datum $m=m_0$ obtained in this experiment:[17]

$$\mathscr{L}(m_0|m_t)=\frac{1}{\sqrt{2\pi\sigma_m^2}}\,e^{-(m_0-m_t)^2/2\sigma_m^2}. \tag{2}$$

The "measured value" (i.e., point estimate) of a physical quantity is frequently taken as that value which maximizes the likelihood $\mathscr{L}$ or equivalently the log-likelihood $\ln\mathscr{L}$. The rub comes in choosing how to extract "errors" (confidence intervals) from $\ln\mathscr{L}$, particularly when it is not parabolic. Most particle physicists follow the recommendation of the classical texts[3] and use *differences* of $\ln\mathscr{L}$ from its maxi-

mum value (or equivalently *likelihood* ratios) to map out confidence intervals (or regions in the multiparameter case). This is conveniently done with an available FORTRAN routine.[18,19]

From this classical point of view, it is legitimate to draw a likelihood function centered on the measured value m_0, but it is illegitimate to consider *areas* under the curve, i.e., to integrate the likelihood function as if it were a probability density. The usual classical method for obtaining a 68% C.L. confidence interval from $\mathscr{L}$ is to let m_1 and m_2 be those points at which $\ln \mathscr{L}$ is down from its maximum by a difference of 1/2 unit. This is often expressed equivalently as choosing those points at which $-2 \ln \mathscr{L}$ increases by one unit [Fig. 1(c)]. [For the Gaussian case this is same as the chi-square method, since $-2 \ln \mathscr{L}=(m_0-m_t)^2/\sigma_m^2$.] In non-Gaussian situations, this and other classical techniques[20] for extracting intervals from $\mathscr{L}$ yield only *approximate* confidence intervals in the sense of Neyman, as discussed further in Sec. V A below.

In contrast, in the Bayesian framework one does use $\mathscr{L}$ as one of two ingredients for constructing the *probability density function* (pdf) for the unknown true value of m_t, and in many cases that pdf can be $\mathscr{L}$ itself. Those who view $\mathscr{L}$ as a pdf will obtain 68% C.L. (central) confidence intervals by finding those values of m_1 and m_2 which yield tails which each contain 16% of the area under the curve [Fig. 1(d)].

Philosophically, there is a whole arena of controversy regarding whether or not it makes sense to have a pdf for the true value of a physical quantity, etc. That debate goes on, but meanwhile, a pragmatist can consider the utility of equations generated by the two approaches while skirting the issue of buying into a whole philosophy of science. Operationally, the line of demarcation between classical and Bayesian methods essentially lies in whether or not $\mathscr{L}$ is integrated. Thus, we will identify a method as Bayesian if the likelihood function is used in forming a probability density from which intervals are computed by integration.[2]

Bayesian methods proceed by invoking an interpretation of Bayes's Theorem[3,21,22] in which one deems it sensible to consider a pdf for the unknown true value m_t. We let $P(m_t)$ be the "prior" pdf for m_t, which reflects our beliefs before doing the experiment. We let $P(m_t|m_0)$ be the "posterior" pdf for m_t (given the data m_0), which reflects our modified beliefs after incorporating the results of the experiment. Then it is precisely the likelihood function $\mathscr{L}(m_0|m_t)$ which relates the two:

$$P(m_t|m_0)=\mathscr{L}(m_0|m_t) \times P(m_t) \Big/ \int_{\text{all } m_t} \mathscr{L}(m_0|m_t)P(m_t)dm_t . \tag{3}$$

The method illustrated in Fig. 1(d) can thus be identified as Bayesian with uniform prior [i.e., $P(m_t)=\text{const}$]. In general, Bayesian confidence intervals are constructed using the posterior pdf, as illustrated more extensively in Sec. V B. In particle physics, the prior is almost always taken to be uniform (where nonzero), although this assumption often goes unemphasized by those who merely report that they "integrated the likelihood function."

In the example of Fig. 1, the 68% confidence interval obtained is of course independent of the method chosen. However, as demonstrated in Secs. IV–VI, some of the most common analysis problems in particle physics go straight to the core of the classical Bayesian debate in a way that cannot be avoided.

III. THE FREQUENTIST CONCEPT OF "COVERAGE"

Although particle physicists may use the words "confidence interval" loosely, the most common meaning is still in terms of original classical concept of coverage which follows from the method of construction suggested in Figs. 1(a) and 1(b). This concept is usually stated (too narrowly, as noted below) in terms of a hypothetical ensemble of similar experiments, each of which measures m and computes a confidence interval for m_t with say, 68% C.L. Then the classical construction guarantees that in the limit of a large ensemble, 68% of the confidence intervals contain the unknown true value m_t, i.e., they "cover" m_t. This property, called coverage in the frequentist sense, is the defining property of classical confidence intervals. It is important to see this property as what it is: it reflects the relative frequency with which the statement "m_t is in the interval (m_1,m_2)" is a true statement. The probabilistic variables in this statement are m_1 and m_2; m_t is fixed and unknown. It is equally important to see what frequentist coverage is *not*: it is not a statement about the degree of belief that m_t lies within the confidence interval of a particular experiment. The whole concept of "degree of belief" does not exist with respect to classical confidence intervals, which are cleverly (some would say devilishly) defined by a construction which keeps strictly to statements about $P(m|m_t)$ and *never* uses a probability density in the variable m_t.

This strict classical approach can be considered to be either a virtue or a flaw, but I think that both critics and adherents commonly make a mistake in describing coverage from the narrow point of view that I described in the preceeding paragraph. As Neyman himself pointed out from the beginning,[16] the concept of coverage is not restricted to the idea of an ensemble of hypothetical nearly identical experiments. Classical confidence intervals have a much more powerful property: if, in an ensemble of *real, different experiments*, each experiment measures whatever observable it likes, and constructs a 68% C.L. confidence interval, then in the long run 68% of the confidence intervals cover the true value of their respective observables. This is directly applicable to real life, and is the real beauty of classical confidence intervals.

Sometimes one intentionally constructs confidence intervals using a method which gives *greater* coverage than claimed by the stated confidence level. Such intervals are referred to as "conservative." I believe that normally, if one wants to ensure greater coverage, a better way is to construct an interval with a higher stated C.L. (Though with discrete distributions such as Poisson, some level of conservatism can be unavoidable.) However, this notion of conservatism can come in handy for someone advocating a Bayesian method. If the advocate can show that the Bayesian method gives conservative intervals, then he or she will likely encounter less resistance or even acceptance (as in the case of physical constraints discussed in Sec. IV). On the other hand, if a Bayesian method is known to yield intervals with frequentist coverage appreciably less than the stated C.L. for some pos-

sible value of the unknown parameters (i.e., they *undercover*), then it seems to have no chance of gaining consensus acceptance in particle physics.

Both classical and Bayesian methods of interval construction rely critically on knowing $P(m|m_t)$ correctly. (In the Bayesian case it is used to construct the likelihood function.) Much of the skill in experimental physics involves designing an experiment for which $P(m|m_t)$ can be known reliably. This requires subsidiary calibration measurements that can exceed the main measurement in terms of work and perseverance. If the 68% C.L. confidence interval for a particular experiment fails to cover m_t, then it can be difficult to know if the experiment was in the "unlucky" 32%, or if a mistake was made in calculating $P(m|m_t)$.

In real physics research, mistakes are of course made, and much of the scrutiny given to a surprising result consists of trying to find mistakes in the calculation of $P(m|m_t)$. Scientists digesting reported confidence intervals may in effect modify $P(m|m_t)$ by adding allowance for "unknown errors" depending on the reputation of the experimenter, difficulty and novelty of the experimental technique, etc. How one reacts to reported confidence intervals brings us into the important subject of decision theory, which is however beyond the scope of this paper.

IV. INCORPORATING CONSTRAINTS INTO CONFIDENCE INTERVALS

The strongest challenge to the dominance of classical confidence intervals in particle physics comes when there is "objective" prior information, in particular constraints on the possible physical values of the quantity measured. This issue was long anticipated in the statistics literature[23] and has been highly visible in the particle and nuclear physics communities in connection with upper limits on the mass of the electron neutrino.[24] The result was that a Bayesian method gained recognition by the Particle Data Group (PDG).[6]

Masses[25] are physically constrained to be non-negative, and in fact the quantity more directly measured in these experiments is the square of the neutrino mass m^2. In the standard model, $m^2=0$, but there is widespread speculation (and some would say evidence) that neutrinos actually have a small, nonzero mass. Direct measurements of m^2 for the electron neutrino emitted in tritium beta decay have been attempted by many groups.

Interestingly, a number of the experiments have reported negative values for their best estimate of m^2, typically obtained by maximizing a likelihood function. In fact, the PDG's weighted average[6] over all experiments gives a measured value and central 68% classical confidence interval as $m^2=(-54\pm30)$ eV^2. The whole interval is in the unphysical region! The preponderance of experiments reporting negative values for m^2 is disconcerting and there is continued speculation about unknown sources of error. However, there is nothing unusual or alarming about a particular experiment obtaining an unphysical value. If neutrinos have zero mass and the resolution function is an unbiased Gaussian, then even in the *absence* of unknown biases, half of an ensemble of experiments will obtain negative values, and 16% of experiments will have 68% classical confidence intervals completely in the unphysical region. Independent of one's attitude toward possible unknown errors in the electron neutrino mass measurements, the question they raise regarding how to incorporate physical constraints goes right to the heart of a prominent Bayesian classical conflict.

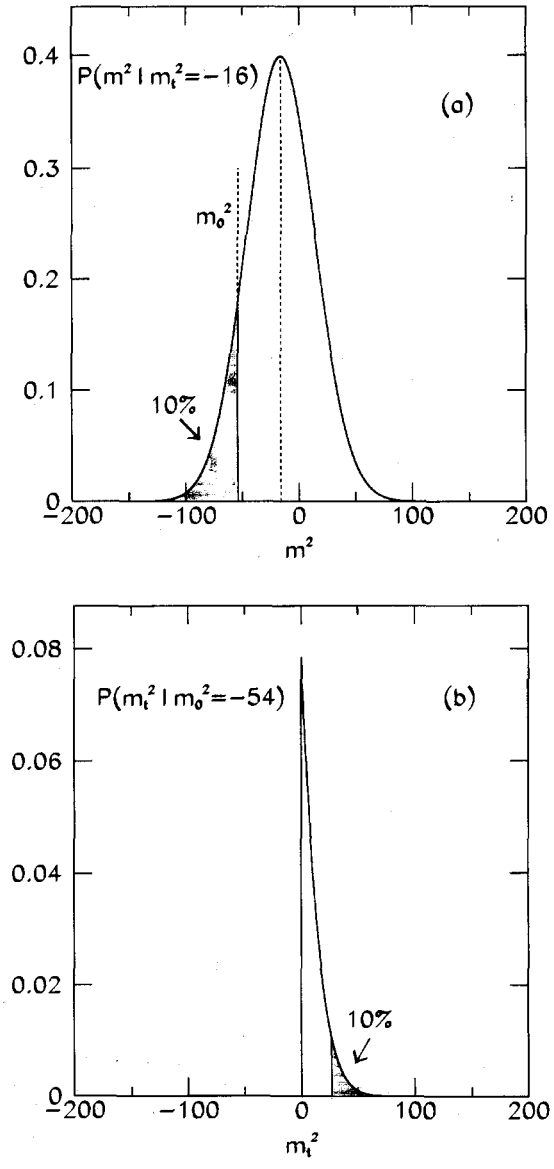

Fig. 2. For mass-squared of electron neutrino, construction of 90% C.L. upper limit on m_t^2 given measured value $m_0^2=(-54\pm30)$ eV^2. (a) Classical construction, and (b) Bayesian construction with prior $P(m_t^2)$ vanishing for $m_t^2<0$ and uniform for $m_t^2\geq0$.

In the absence of strong evidence for a nonzero mass, one typically quotes an *upper limit* on the mass at a specified confidence level. (Of course,[26] one should also report the measured central value and error to facilitate combining results from different experiments, and to allow the reader to construct limits as he or she desires.) The 90% C.L. upper limit is the right endpoint of the highly noncentral 90% C.L. confidence interval whose left endpoint is $-\infty$. The classical construction, illustrated in Fig. 2(a) using the PDG world average, results in an unphysical 90% C.L. upper limit of -16 eV^2. (This is the mean of the Gaussian, with rms deviation -30 eV^2, which has 10% of its area to the left of -54 eV^2.) Such results are allowed in the pure classical method; they are simply included in the 10% of all "90% C.L. upper

limits" which are false statements. However, when the upper limit is negative, one *knows* that the result is in the false 10%, which is of course disturbing.

In the face of such unphysical limits, it has become common[6] to use the Bayesian Eq. (3), only with m^2 substituted for m everywhere. The Bayesian approach naturally accommodates the physical constraint simply by using a prior pdf $P(m_t^2)$ which is zero if $m_t^2<0$ and uniform otherwise. The resulting posterior pdf is just the tail of $\mathscr{L}(m_0^2|m_t^2)$ which lies at $m_t^2\geq 0$ [Fig. 2(b)]. Integrating yields $m_t^2<26.6$ eV^2 at 90% C.L., naively implying $m_t<5.2$ eV, though the PDG urges caution in interpreting a limit such as this. A major factor in the acceptance of this method is that the resulting upper limits are less restrictive than the classical intervals and hence *conservative* in the absence of mistakes.

Though Bayesians may enjoy the imprimatur of the PDG, the inherent difficulty in the Bayesian method is also immediately apparent: in which quantity (m, m^2, ln m, etc.) should the prior be uniform? The consensus view settled on m^2, but the fact that the upper limit depends on this choice remains extremely unsettling to many. I come back to this important point in Sec. V B below.

As a test of where one stands on the issue of physical constraints, one can consider the upper limit on the number of light neutrinos, as reported by the Mark II collaboration[27] in 1989. At the time of their measurement, the existence of at least three types of neutrinos (associated with electron, muon, and tau) was well established, and the outstanding question was whether or not there was a fourth. At the SLAC Linear Collider, they measured Z-boson resonance parameters, with precision which was to be improved immensely within months by competition from the LEP experiments at CERN.[6] In an analysis which assumed standard model couplings but which fit for the Z mass and the number of neutrinos N_ν, the Mark II collaboration obtained $N_\nu=2.8\pm0.6$. Apparently using the classical method, they also reported, "the 95% C.L. limit, $N_\nu<3.9$, excludes to this level the presence of a fourth massless neutrino species within the standard model framework." This was perfectly respectable, but it was also of interest that the Mark II's ability to exclude $N_\nu=4$ at this level was partly due to a mild (1/3 standard deviation) downward fluctuation in their central value, from 3.0 to 2.8. A Bayesian analysis with (uniform) prior vanishing for $N_\nu<3$ would have yielded a slightly weaker statement on this crucial issue.

In the above, the measured quantity N_ν is treated as a continuous variable, which is sensible in broader contexts because new physics (such as supersymmetry) can manifest itself as an *apparent* fractional increase in N_ν. However, for the narrow question of how many neutrino types there are if the rest of the standard model is not extended, the true unknown value of N_ν is an integer. This discreteness suggests a likelihood ratio analysis in which one compares the relative likelihoods of the various integer values of N_ν. A Bayesian may then additionally incorporate his or her prior probabilities for each N_ν (in particular zero for $N_\nu<3$), and compute posterior probabilities. I leave this as an exercise for the interested reader.

Thus, although the physics issue of the number of neutrino types quickly became moot because of the LEP data, the Mark II result remains a superb illustration of a general statistics problem, especially since the Mark II result was not *so* unphysical as to generate doubts about the experiment.

V. CONFIDENCE INTERVALS USING POISSON STATISTICS

Poisson statistics naturally arise in counting experiments, e.g., as the limit of binomial statistics when one is counting the number of times a rare process is observed ("successes"), which is a small fraction of all processes taking place. In the last decade, searches for new processes (new particle production, reactions, or particle decay modes) have become commonplace as part of a large effort to discover new laws of physics "beyond the standard model." Nearly all such searches have resulted in no claims of new physics, but rather in upper limits on their rates. In discussing the statistics of such limits, I restrict most of my discussion to the simplest cases, where it is already apparent that Bayesian intervals do not have frequentist coverage except in special cases. However, an innocent-looking complication in Sec. VI introduces a bizarre twist for the classicists.

Consider an experiment which searches for a possible rare process (a rare decay such as $\mu\to e\gamma$ or a rare interaction such as that producing the top quark) by observing with a detector the number of events n which appear to be from the signal. Then n is drawn from a Poisson distribution with unknown true mean μ_t [and rms deviation equal to $(\mu_t)^{1/2}$]:

$$P(n|\mu_t)=\frac{\mu_t^n e^{-\mu_t}}{n!}. \tag{4}$$

Let R_t be the unknown true value of the relevant physical quantity which is to be measured (branching ratio, cross section, etc.). Then we can always write $\mu_t=R_tS_t$, where S_t is the true value of the experiment's sensitivity factor: a combination of the number of interactions or decaying particles, observing live time, detection efficiency, etc. In order to use n to make an inference about the value of R_t, one needs information about the value of S_t. This information is typically obtained from subsidiary measurements that give an estimate $\hat{S}$ for S_t, and its uncertainty σ_S.

We assume until Sec. VI that σ_S is negligibly small, so that we know the sensitivity S_t exactly. Then all inferences about the branching ratio R_t directly correspond to inferences about μ_t. Furthermore, throughout this article, we assume that there are no background events, i.e., we assume that all events attributed to the rare process are in fact real. Then suppose that n_0 events are observed. One way to report the result of the experiment (the point estimate $\hat{R}$ of the true value R_t) is

$$\hat{R}=\left(\frac{n_0}{S_t}\right)\pm\left(\frac{\sqrt{n_0}}{S_t}\right), \tag{5}$$

using the estimate $\hat{\mu}=n_0\pm(n_0)^{1/2}$ of the true mean μ_t. Here the quantity following the "$\pm$" is an estimate of the rms deviation from Poisson statistics. However, a more common convention is for $\pm$ to indicate a 68% confidence interval, which for small n_0 can differ significantly from Eq. (5).

In analogy with Fig. 1, there is a variety of ways to construct 68% C.L. intervals for μ_t (and hence R_t). All methods in common use retain n_0 as the best estimate of μ_t, but the variously computed confidence intervals (which typically do not have n_0 as midpoint) are generally different. The fact that n is discrete while μ_t is continuous means that the classical and Bayesian construction methods bear no pictorial resemblance to each other.

For purposes of illustration, we examine in detail the case where $n_0=3$. Such a small number of observed events is

Table I. 68% C.L. confidence intervals (μ_1,μ_2) for the mean of a Poisson distribution, based on the single observation $n_0=3$, calculated by various methods.

Method	Prior	Defining equation(s)	Interval	Length	Coverage?
Root-mean-square deviation	···	$n_0 \pm \sqrt{n_0}$	(1.27, 4.73)	3.46	no
Classical central	···	Eqs. (6) and (7)	(1.37, 5.92)	4.55	yes
Classical shortest	···	Method of Crow and Gardner[a]	(1.29, 5.25)	3.96	yes
Likelihood ratio	···	Eq. (9)	(1.58, 5.08)	3.50	no
Bayesian central	1	Eqs. (16) and (17)	(2.09, 5.92)	3.83	no
Bayesian shortest	1	Eq. (16); minimum $\mu_2-\mu_1$	(1.55, 5.15)	3.60	no
Bayesian equal $\pm$	1	Eq. (16); $\hat{\mu}-\mu_1=\mu_2-\hat{\mu}$	(1.15, 4.85)	3.70	no
Bayesian central	$1/\mu_t$	Eqs. (16) and (17)	(1.37, 4.64)	3.27	no
Bayesian shortest	$1/\mu_t$	Eq. (16); minimum $\mu_2-\mu_1$	(0.86, 3.85)	2.99	no
Bayesian equal $\pm$	$1/\mu_t$	Eq. (16); $\hat{\mu}-\mu_1=\mu_2-\hat{\mu}$	(1.36, 4.63)	3.27	no

[a]Reference 31.

typical of a pioneering frontier experiment. Different ways of calculating 68% C.L.[15] confidence intervals are discussed below, and summarized in Table I. However, constructing a 68% C.L. interval for μ_t is only one option. An experimenter unsure that the background is negligible may prefer to quote an upper confidence limit, as discussed in Sec. V C.

A. Classical confidence intervals in the Poisson case

Starting from Eq. (4), the most common construction of a classical 68% C.L. confidence interval (μ_1, μ_2) proceeds[28,29] by finding the value μ_1 such that $P(n\geq n_0|\mu_1)=16\%$,

$$\sum_{n=n_0}^{\infty} P(n|\mu_1)=0.16 \tag{6}$$

[Fig. 3(a)]; and by finding the value μ_2 such that $P(n\leq n_0|\mu_2)=16\%$,

$$\sum_{n=0}^{n_0} P(n|\mu_2)=0.16 \tag{7}$$

[Fig. 3(b)]. The criterion of $(1-0.68)/2=16\%$ on each side represents a choice (central intervals[30]) dictated by taste or convention. An interesting alternative is to seek intervals of minimum length, i.e., those which minimize $\mu_2-\mu_1$. Crow and Gardner[31] gave a recipe for their construction, which for $n_0=3$ results in the 68% C.L. interval (1.29, 5.25) with length 3.96, compared to the usual interval (1.37, 5.92) with length 4.55.

A classical approach might use the likelihood function to obtain approximate confidence intervals. The likelihood $\mathscr{L}(n_0|\mu_t)$ is given by the same expression as Eq. (4), with the important change in point of view: we now consider its variation with μ_t, given the particular data n_0 obtained in this experiment:

$$\mathscr{L}(n_0|\mu_t)=\frac{\mu_t^{n_0}e^{-\mu_t}}{n_0!}. \tag{8}$$

$\mathscr{L}$ is maximized for $\mu_t=\hat{\mu}=n_0$. In analogy with the Gaussian case, an approximate 68% classical confidence interval is obtained by the *likelihood ratio* method, often implemented using differences of negative log likelihoods. One finds $\mu_1<n_0$ and $\mu_2>n_0$ such that

$$\begin{aligned}-2\ \ln \mathscr{L}(n_0|\mu_1)&=-2\ \ln \mathscr{L}(n_0|\mu_2)\\&=-2\ \ln \mathscr{L}(n_0|\hat{\mu})+1.\end{aligned} \tag{9}$$

Though not exact, this method and is easy to implement on a computer[18] and to generalize to higher-dimensional problems.[19] It is also generally more accurate than the parabolic approximation using a symmetric interval of half-length:

$$\delta\mu=-\frac{1}{2}\frac{\partial^2}{\partial\mu_t^2}\ln \mathscr{L}(n_0|\mu_t)\Big|_{\mu_t=\hat{\mu}}, \tag{10}$$

which in this case gives $\delta\mu=(n_0)^{1/2}$. More sophisticated methods for extracting approximate confidence intervals from the likelihood function exist,[20] but they all seem to undercover when the true mean μ_t is small.

As noted in Sec. II it is important to realize that $\mathscr{L}(n_0|\mu_t)$ by itself is *not* a pdf in μ_t. This is particularly clear from the mathematics in this Poisson case since $\mathscr{L}$ is constructed from the expression for probabilities of a discrete variable n, not from a probability *density*. Nothing in the construction of $\mathscr{L}$ allows one to multiply it by $d\mu_t$ and integrate, or to consider areas under it. Classically, one strictly looks only at *ratios* of $\mathscr{L}$ for different values of μ_t.

B. Bayesian confidence intervals in the Poisson case

In order to make mathematical sense out of putting $\mathscr{L}(n_0|\mu_t)$ inside an integral over μ_t, one must multiply it by a pdf in μ_t, and this is precisely the approach taken in Bayesian statistics. In analogy with Eq. (3), the posterior pdf for μ_t is proportional to the product of the likelihood function and the prior pdf for μ_t:

$$P(\mu_t|n_0)d\mu_t=\mathscr{L}(n_0|\mu_t)[P(\mu_t)d\mu_t]\Big/\int_{\text{all }\mu_t}\mathscr{L}(n_0|\mu_t)P(\mu_t)d\mu_t. \tag{11}$$

I have explicitly written the differential element $d\mu_t$ on both sides to emphasize that mathematically it goes hand-in-hand with a probability density, which on the right-hand side is $P(\mu_t)$. The denominator must be carefully justified,[3,22] but here it just serves to normalize the pdf, and it is suppressed below. $P(\mu_t|n_0)$ is a density in μ_t which can be integrated, and it makes sense (at least to a Bayesian) to use areas under $P(\mu_t|n_0)$ to construct confidence intervals.

The problem lies in what to use for the prior pdf $P(\mu_t)$. For a large sample of data or a sharply peaked likelihood function, one is not too sensitive to the choice of $P(\mu_t)$. But in the case typified by our single experiment with small n_0,

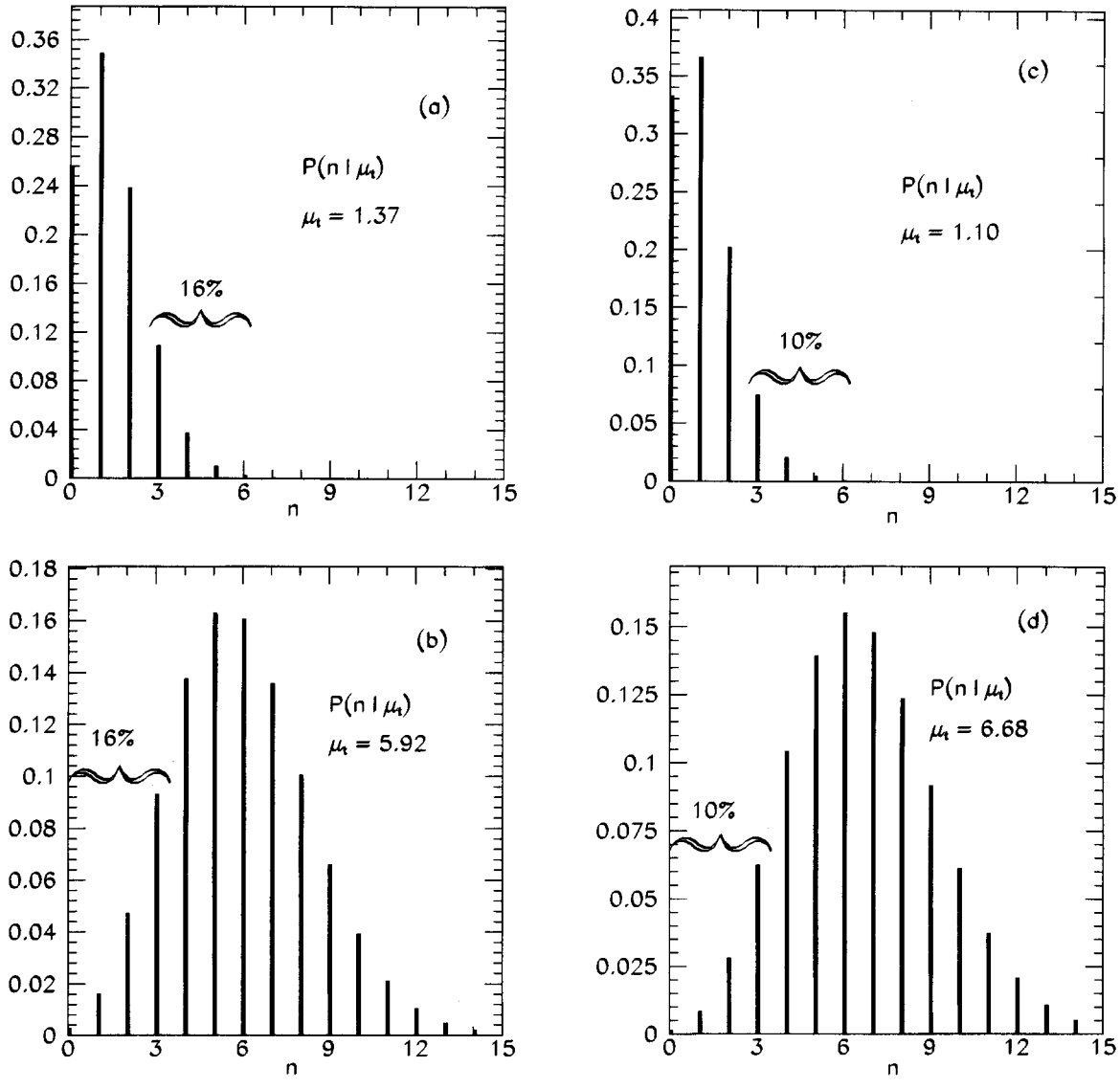

Fig. 3. Classical solutions in Poisson case with $n_0=3$. (a) and (b) Left and right 68% C.L. interval endpoints according to Eqs. (6) and (7); (c) and (d) same construction for 80% C.L. interval, yielding endpoints of 90% C.L. lower and upper limits, respectively.

the choice of $P(\mu_t)$ can be critical. For reporting results, one would like to use a prior density which is objective in a vaguely defined sense, corresponding to what statisticians call "uninformative" priors,[22,32] which attempt to represent an absence of prior knowledge about μ_t.

By far the most common practice in particle physics is to use the uniform prior $P(\mu_t)\equiv 1$, wherever μ_t is defined (i.e., $\mu_t>0$). Thus one has the equation

$$P(\mu_t|n_0)d\mu_t \propto \mathscr{L}(n_0|\mu_t)d\mu_t, \tag{12}$$

with its hidden "1" for the prior. This represents a rather naive choice, but I think it survives for Poisson statistics partly because of a mathematical curiosity in one special case: if Bayesian *upper* limits are calculated with this prior, then one obtains *precisely the same upper limits* as with the classical construction! (See Sec. V C below.)

The naivety of the choice of uniform prior is exposed by two considerations. First, we note that the physical process giving rise to Poisson statistics is often exponential decay, which is equivalently described by either the mean lifetime or its inverse, the decay rate. Accordingly, one experimenter may prefer to think in terms of μ_t and hence use a prior which is uniform in μ_t, while another experimenter may prefer to think in terms of $1/\mu_t$ and hence use a prior which is uniform in $1/\mu_t$: $P(1/\mu_t)\equiv 1$. The experimenter who analyzes the data in terms of $1/\mu_t$ will write

$$P(1/\mu_t|n_0)d(1/\mu_t) \propto \mathscr{L}(n_0|1/\mu_t)\times P(1/\mu_t)d(1/\mu_t). \tag{13}$$

The likelihood function has the well-known property that it is invariant with respect to transformations of the variable μ_t. That is, $\mathscr{L}(n_0|1/\mu_t)=\mathscr{L}(n_0|\mu_t)$, since both likelihood functions are actually constructed from Eq. (4). (This underscores the fact that a likelihood function is not a pdf, and does not behave like a pdf.) Equation (13) becomes [using $d(1/\mu_t)=-d\mu_t/\mu_t^2$]

$$P(1/\mu_t|n_0)d(1/\mu_t) \propto -\mathscr{L}(n_0|\mu_t)d\mu_t/\mu_t^2. \tag{14}$$

But the posterior densities must for consistency obey $-P(1/\mu_t|n)d(1/\mu_t)=P(\mu_t|n)d(\mu_t)$, so that Eq. (14) becomes

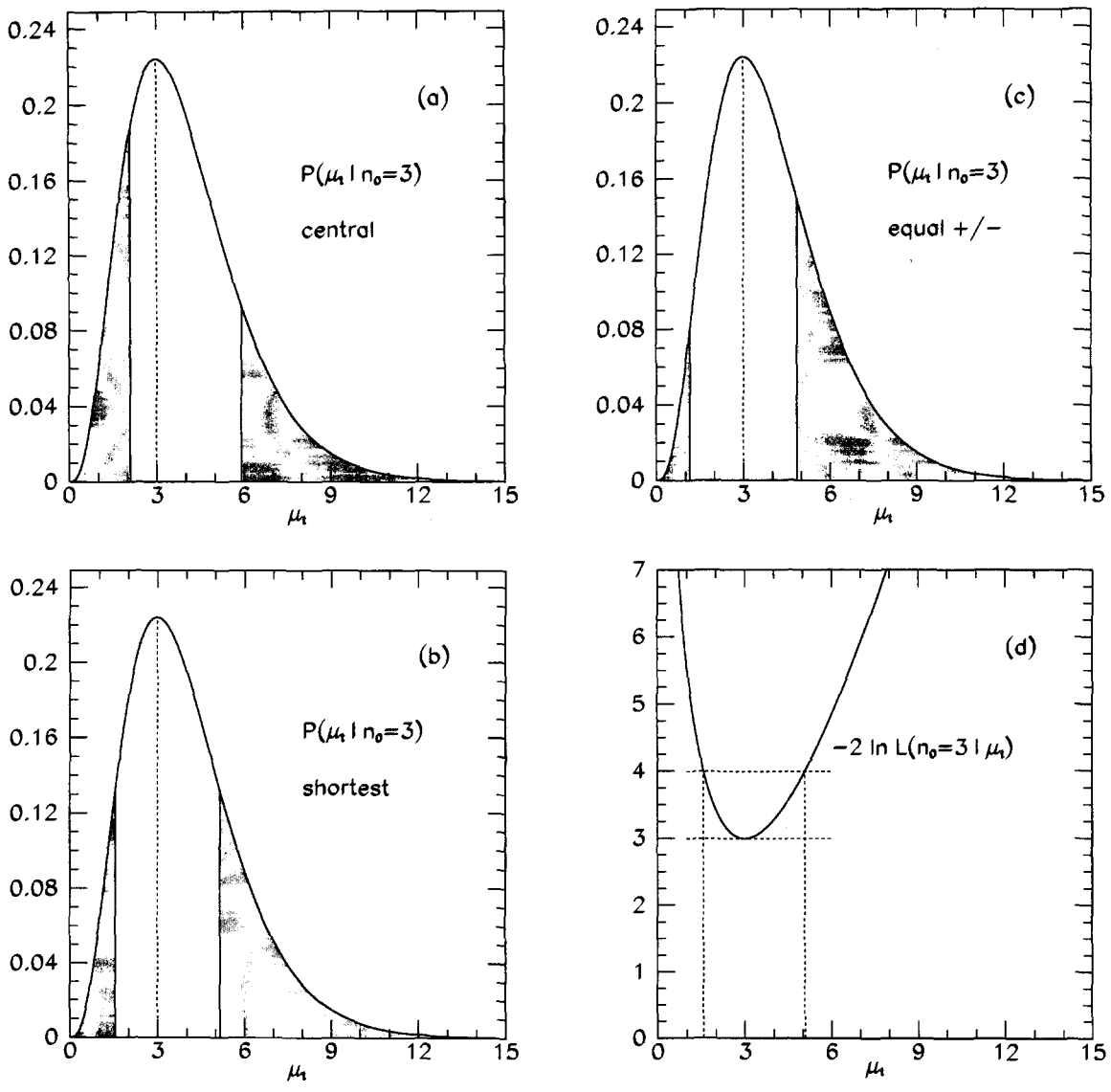

Fig. 4. More 68% C.L. intervals in Poisson case with $n_0=3$. Bayesian (uniform prior) using Eq. (16) with subsidiary conditions (a) [Eq. (17)], (b) minimum width, and (c) $\hat{\mu}-\mu_1=\mu_2-\hat{\mu}$. (d) Likelihood ratio method, Eq. (9).

$$P(\mu_t|n_0)d\mu_t \propto (1/\mu_t^2)\mathscr{L}(n_0|\mu_t)d\mu_t, \tag{15}$$

in contradiction with Eq. (12). The contradiction of course arises because the assumptions of uniform prior for both μ_t and $1/\mu_t$ are inconsistent. The dilemma of what to use for an uninformative prior can be stated as: in what metric, i.e., for what function of μ_t, should the prior be uniform? The answer is certainly not "*obviously* μ_t!"

The choice of uniform prior is naive for another reason, namely that statisticians known for advocating Bayesian statistics most strongly include those who have argued that for Poisson statistics, the uninformative prior $P(\mu_t)$ should be either $(\mu_t)^{-1/2}$ or $1/\mu_t$. The monograph by Harold Jeffreys,[33] which reflects a career devoted to the exposition of a Bayesian theory of probability, discusses both $(\mu_t)^{-1/2}$ and $1/\mu_t$, and seems to come down on the side of $1/\mu_t$ if absolutely nothing is known about μ_t.

Jeffreys's original argument[33] for $P(\mu_t)=1/\mu_t$ was that it *is* invariant under changes of power of the parameter being estimated, such as the change from μ_t to $1/\mu_t$ above. That is, the priors $P(\mu_t)d\mu_t=d\mu_t/\mu_t$ and $P(\mu_t^k)d\mu_t^k=d\mu_t^k/\mu_t^k$ are consistent for any power k. Both are proportional to $d(\ln \mu_t)$, and we see that $\ln \mu_t$ is the metric in which this prior is uniform. This may amuse those physicists who sometimes view "orders of magnitude" as the natural metric!

A related argument[34] for $P(\mu_t)=1/\mu_t$ is based on requiring consistency for two experimenters, each measuring the rate of the same decay process, but each using his or her own absolute standard for the passage of time.

The arguments[23,35] for $P(\mu_t)=(\mu_t)^{-1/2}$ involve a choice of metric which is deemed "natural," either because of a criterion of shape invariance under data translation, or because it is the *information* metric familiar to statisticians.

Thus, among these Bayesians, there are two main candidates for the uninformative prior for the mean μ_t of a Poisson process, and neither of them is the uniform prior! I compare the uniform and $1/\mu_t$ priors in illustrations below.

Once the prior is specified, Bayesian 68% C.L. confidence intervals (μ_1, μ_2) can be easily calculated to obey the defining criterion

$$\int_{\mu_1}^{\mu_2} P(\mu_t|n_0)d\mu_t=0.68. \tag{16}$$

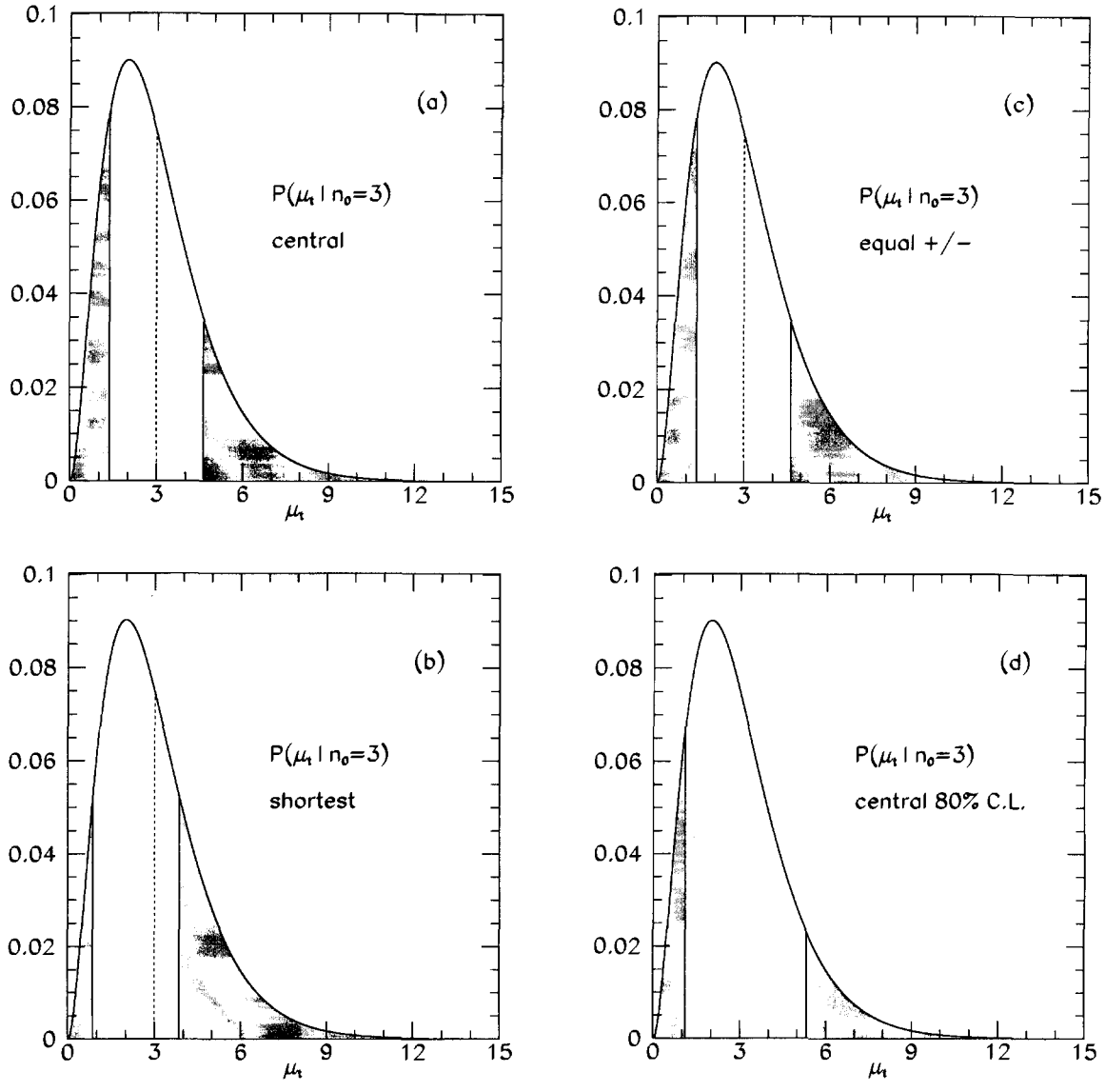

Fig. 5. Bayesian intervals for $1/\mu_t$ prior with $n_0=3$. 68% C.L. intervals with (a) central area [Eq. (17)], (b) minimum width, and (c) $\hat{\mu}-\mu_1=\mu_2-\hat{\mu}$. (d) 80% C.L. central interval, giving 90% C.L. lower and upper limits.

As with classical intervals, they are not uniquely determined without further choice dictated by taste or convention. Various possibilities are illustrated in Figs. 4 and 5 for both the uniform and $1/\mu_t$ prior. Bayesian analogs of *central* intervals are obtained from the subsidiary requirement

$$\int_0^{\mu_1} P(\mu_t|n_0)d\mu_t = \int_{\mu_2}^{\infty} P(\mu_t|n_0)d\mu_t, \tag{17}$$

which puts 16% of the posterior probability on each side of the interval. [See Figs. 4(a) and 5(a).]

Alternatively, one can find the *shortest* Bayesian intervals. There is even an independent argument for choosing shortest intervals, because it would seem very reasonable to require intervals to have the property that the posterior $P(\mu_t|n_0)$ for any μ_t outside the confidence interval be less than $P(\mu_t|n_0)$ for all μ_t inside the confidence interval. This latter criterion is sufficient to obtain the shortest intervals. [See Figs. 4(b) and 5(b).]

Except in the special case of upper limits, Bayesian confidence intervals in the Poisson case typically *fail* the criterion of the frequentist coverage. One can see why by comparing the central 68% C.L. intervals for $n_0=3$: (1.37, 5.92) for the classical construction, (2.08, 5.92) for the Bayesian construction with uniform prior, and (1.37, 4.63) for the Bayesian construction with $1/\mu_t$ prior. The Bayesian intervals each have one endpoint identical to the classical interval, with the other endpoint *inside* the classical interval, thereby shortening it!

This coincidence of endpoints results from a wonderful property of the Poisson distribution which connects the classical and Bayesian prescriptions:

$$\int_{\mu_2}^{\infty} P(n_0|\mu_t)d\mu_t = \sum_{n=0}^{n_0} P(n|\mu_2), \tag{18}$$

so that also

$$\int_0^{\mu_1} P(n_0|\mu_t)d\mu_t = \sum_{n_0+1}^{\infty} P(n|\mu_1). \tag{19}$$

In the case of uniform prior, where $P(\mu_t|n_0) \propto \mathscr{L}(n_0|\mu_t) = P(n_0|\mu_t)$, these equations allow us to compare the classical equations (6) and (7) with the Bayesian analog, Eqs. (16) and (17). They imply that the Bayesian μ_2 is identical with the corresponding classical μ_2 inferred from n_0, while the Bayesian μ_1 is identical with the corresponding classical μ_1 inferred from data n_0+1, not (!) from data n_0.

With the $1/\mu_t$ prior, the power of μ_t coming from $\mathscr{L}$ is reduced by one, with the result that the Bayesian μ_1 is then identical with classical μ_1 inferred from n_0, while the Bayesian μ_2 is identical with the classical μ_2 inferred from data n_0-1, not from data n_0.

C. Upper and lower limits in Poisson case

These results have immediate implications for upper and lower limits. In both classical and Bayesian constructions, an upper limit μ_2 is merely a special case of a confidence interval in which the subsidiary choice $\mu_1=0$ is made in order to determine the interval uniquely. Similarly, a lower limit is a special case determined by requiring $\mu_2=\infty$. It follows that the endpoints of a *central* 68% C.L. confidence interval (μ_1, μ_2) have the property that μ_2 is a 16% C.L. upper limit on μ_t, and that μ_1 is a 16% C.L. lower limit. More typical C.L. values for upper and lower limits, such as 90% and 95%, have a corresponding relationship to the endpoints of 80% and 90% C.L. central confidence intervals, respectively. Construction of 90% C.L. upper and lower limits is illustrated in Figs. 3(c), 3(d), 5(d), and 6, and summarized in Tables II and III.

We see that Bayesian upper limits on μ_t derived with uniform prior are identical with classical upper limits, while those obtained with the $1/\mu_t$ prior (or any prior of the form μ^k with $k<0$) fail frequentist coverage. The $1/\mu_t$ prior has an even bigger problem when $n_0=0$, for then the posterior probability is not normalizable. Since the computation of an upper limit when $n_0=0$ is commonly encountered, this is a serious defect.

In contrast, Bayesian *lower* limits using the $1/\mu_t$ prior are identical to the classical lower limits, while those based on the uniform prior fail to cover!

Table I summarizes all the above results for 68% C.L. confidence intervals for classical, Bayesian, and likelihood-based methods, along with other possibilities including the rms deviation estimate $3\pm\sqrt{3}$. Tables II and III summarize the results for 90% C.L. lower and upper limits. The entry in the last column is "yes" if the method of construction yields frequentist coverage for all values of the unknown μ_t. One who insists on frequentist coverage is not tempted by anything other than the classical constructions, except in the peculiar cases where Bayesian upper or lower limits coincide with the classical ones. Hard-core Bayesians reject the criterion of frequentist coverage, and hence claim not to be bothered when it is not met. But I believe that the particle physics community will never knowingly accept a method which does not provide (at least approximate) frequentist coverage.

Before classical hubris becomes acute, it is worthwhile to consider a seemingly innocuous complication which is unavoidable in a real experiment.

VI. COMBINING THE GAUSSIAN AND POISSON CASES

One of the most common statistical problems in particle physics is to calculate an upper limit on the rate of a Poisson process *in the presence of an approximately Gaussian error on the sensitivity of the experiment.* Since the classical and (uniform prior) Bayesian methods give the *same* intervals for Gaussian densities (Sec. II), and the *same* upper limits of Poisson processes (Sec. V C), one would hope that combining the Gaussian and Poisson cases would be a routine matter, again with classical/Bayesian agreement. Thus it can be startling to find a total departure of classical and Bayesian results when one simply combines the two cases. The purely classical result for upper limits is so counterintuitive that it is not used in particle physics!

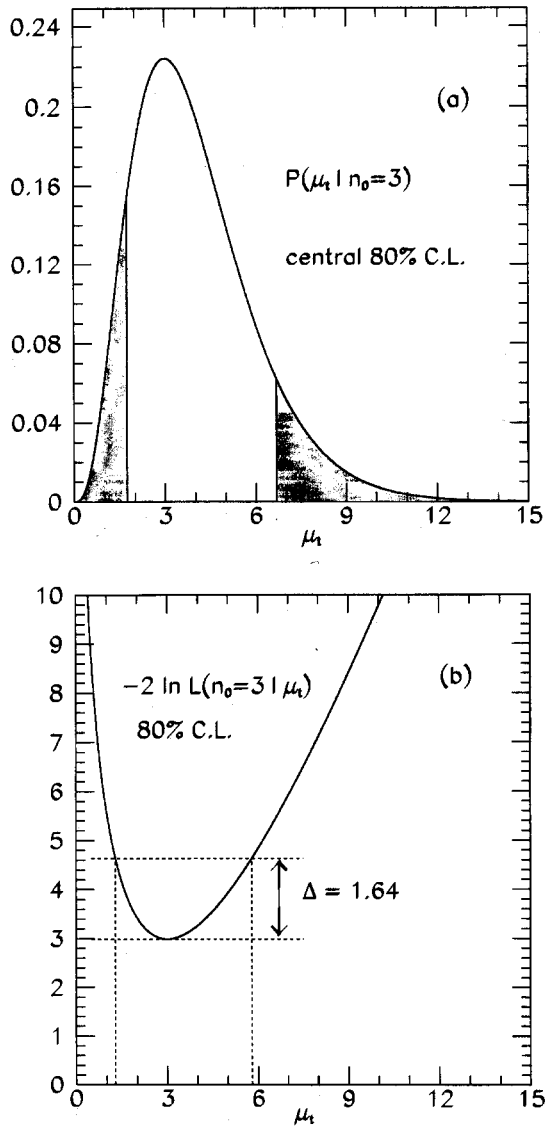

Fig. 6. (a) Bayesian (with uniform prior) and (b) likelihood ratio constructions, for 80% C.L. intervals in Poisson case with single measurement $n_0=3$. The endpoints are 90% C.L. lower and upper limits.

Recently Highland and I have described a sort of classical Bayesian hybrid method[36] which gives simply calculable, intuitive results. Others had previously and independently applied the same sort of reasoning in computer calculations of limits, but very little seems to have been written on this problem.

The problem occurs when computing an upper limit on the rate $R_t=\mu_t/S_t$ of a Poisson process, as described in Sec. V, except now we consider the more realistic case in which the

Table II. 90% C.L. *lower* limit μ_1 for the mean of a Poisson distribution, based on the single observation $n_0=3$, calculated by various methods. The denominator in the Bayesian equations is the normalization coming from integrating over $\mu_t \in (0,\infty)$.

Method	Prior	Defining equation	Limit	Coverage?
Classical	···	$\Sigma_{n=n_0}^{\infty} P(n\|\mu_1)=0.10$	1.10	yes
Likelihood ratio	···	$-2\ln\mathscr{L}(n_0\|\mu_1)=-2\ln\mathscr{L}(n_0\|\hat{\mu})+1.64$	1.29	no
Bayesian	1	$0.1=\int_0^{\mu_1}\mathscr{L}(n_0\|\mu_t)d\mu_t/(\text{denom})$	1.74	no
Bayesian	$1/\mu_t$	$0.1=\int_0^{\mu_1}(1/\mu_t)\mathscr{L}(n_0\|\mu_t)d\mu_t/(\text{denom})$	1.10	yes

true sensitivity factor S_t is unknown and is estimated by $\hat{S}\pm\sigma_S$ from subsidiary measurements. The uncertainty in determining $\hat{S}$ typically arises from the additive effect of many things, so that one can consider $\hat{S}$ to be sampled from an approximately Gaussian probability density $P(S|S_t)=N(S_t,\sigma_S)$ whose rms deviation σ_S is presumed known (after a lot of calibration work).[37]

One frequently refers to the expected Poisson fluctuation in the number of events as the "statistical error" and the uncertainty in $\hat{S}$ as the "systematic error." In reporting a significantly nonzero result, it is common to report the statistical and systematic errors separately, and then combine them in quadrature if a single overall error is desired. The generalization of Eq. (5) is then $\hat{R}=(n_0/\hat{S})\pm\sigma_R$, where $\sigma_R=\hat{R}(1/n_0+\sigma_{\text{rel}}^2)^{1/2}$, and $\sigma_{\text{rel}}=\sigma_S/\hat{S}$ is the *relative* uncertainty in $\hat{S}$.

The case of interest here is when n_0 is small and one wishes to place an upper limit on the branching ratio $R_t=\mu_t/S_t$. One desires a way to introduce the systematic error σ_{rel} into the pure Poisson upper limit from Sec. V. The problem can be illustrated with the simplest case, $n_0=0$ (no signal events are observed). The classical 90% C.L. upper limit on μ_t [from Eq. (7) with 0.10 on the right-hand side] is well-known to be $\ln 10\approx 2.3$: if the mean of a Poisson distribution is larger than 2.3, there is less than 10% chance of obtaining $n_0=0$. If there is no systematic error ($\sigma_{\text{rel}}=0$ so $\hat{S}=S_t$), then the 90% C.L. upper limit on R_t is $2.3/S_t$. The sticky problem arises when $\sigma_{\text{rel}}\neq 0$; e.g., $\sigma_{\text{rel}}=0.1$, i.e., the sensitivity is known to 10%.

Two of the most common methods for dealing with this case have been (1) to ignore σ_{rel} and report the upper limit $2.3/\hat{S}$, or (2) to add 10% to $2.3/\hat{S}$ and report $2.53/\hat{S}$. No justification is usually attempted for either method (though the second can claim to be conservative). One thing is "clear" to all however: *An acceptable method for incorporating σ_{rel} into the upper limit must not result in an upper limit less than* $(2.3/\hat{S})$! Otherwise, if two experiments each find $n_0=0$ and have the same $\hat{S}$, the poorly calibrated one will report a more restrictive limit than the superbly calibrated one.

Thus it may come as a great shock that a purely classical construction of the 90% C.L. upper limit *does* yield a value less than $(2.3/\hat{S})$! How this happens can be understood by first examining the coverage of the upper limit ignoring σ_{rel} in the case where μ_t is near 2.3 compared to σ_{rel}: say $\mu_t=2.28$ or $\mu_t=2.32$ if $\sigma_{\text{rel}}=0.1$. For either $\mu_t=2.28$ or $\mu_t=2.32$, one observes $n\geq 1$ approximately 90% of the time, and in these cases a computed upper limit which ignores σ_{rel} will be $3.9/\hat{S}$ or greater; these limits cover $R_t=\mu_t/S_t$ virtually always. In the remaining cases where the observed n_0 is zero (about 10% of the time), the computed upper limit ignoring σ_{rel} is $2.3/\hat{S}$, which will cover R_t whenever $\hat{S}<(2.3/\mu_t)S_t$, which is about half of these remaining cases! Thus, if $\mu_t=2.28$ or $\mu_t=2.32$, a classical construction which assumes $\sigma_{\text{rel}}=0$ covers R_t about 95% of the time. Since this conclusion is independent of the sign of $(2.3-\mu_t)$, we see that in general ignoring σ_{rel} *over*covers the true value by a finite amount.

Therefore, a classical construction which incorporates $\sigma_{\text{rel}}\neq 0$ and computes an upper limit with *actual* 90% coverage will result in a more restrictive limit, i.e., less than $2.3/\hat{S}$. If $\sigma_{\text{rel}}=0.1$, then 90% coverage can amazingly be obtained by stating upper limits of $2.0/\hat{S}$ when $n_0=0$ and similarly lower-then-usual upper limits when $n_0\neq 0$. As with all claims about frequentist coverage, this is easily verified by Monte Carlo simulation.[38]

This peculiar effect is a consequence of the discrete nature of the observations in a Poisson process. Discrete distributions (the binomial distribution is another) lead to upper limits which can cover the true value more than claimed by the confidence level. As soon as a continuously varying observable is introduced into the problem, the upper limits can sometimes be relaxed in a paradoxical way.[39]

I cannot imagine these perfectly valid classical upper limits being generally accepted. The paper by Highland and me[36] treated the Poisson mean classically, reflecting the bias in the field against Bayesian statistics, and avoiding the issue of prior density for μ_t. But we took a Bayesian approach to the sensitivity in that we considered the subsidiary measure-

Table III. 90% C.L. *upper* limit μ_2 for the mean of a Poisson distribution, based on the single observation $n_0=3$, calculated by various methods. The denominator in the Bayesian equations is the normalization coming from integrating over $\mu_t \in (0,\infty)$.

Method	Prior	Defining equation	Limit	Coverage?
Classical	···	$\Sigma_{n=0}^{n_0} P(n\|\mu_2)=0.10$	6.68	yes
Likelihood ratio	···	$-2\ln\mathscr{L}(n_0\|\mu_2)=-2\ln\mathscr{L}(n_0\|\hat{\mu})+1.64$	5.80	no
Bayesian	1	$0.1=\int_{\mu_2}^{\infty}\mathscr{L}(n_0\|\mu_t)d\mu_t/(\text{denom})$	6.68	yes
Bayesian	$1/\mu_t$	$0.1=\int_{\mu_2}^{\infty}(1/\mu_t)\mathscr{L}(n_0\|\mu_t)d\mu_t/(\text{denom})$	5.32	no

ments to yield a posterior pdf $P(S|\hat{S},\sigma_S)$ for the true sensitivity S_t. We then calculated the upper limits such that the true value is covered by 90% of an ensemble of experiments with sensitivities sampled from $P(S|\hat{S},\sigma_S)$. The effect is to yield a reported upper limit *greater* than $2.3/\hat{S}$ in the case $n_0=0$, as intuitively demanded. For small $\sigma_{\rm rel}$ [independent of whether or not $P(S|\hat{S},\sigma_S)$ is Gaussian], the approximate formula is

$$R_t<2.30(1+2.30\sigma_{\rm rel}^2/2)/\hat{S} \quad (90\% \text{ C.L.}). \tag{20}$$

Generalizations are given in our paper.

A fully Bayesian treatment with essentially uniform prior gives similar results, if the prior for S_t is used to rule out S_t unreasonably close to zero. Thus, in a deceptively simple situation, the classical construction gives an unacceptable result, and a touch of Bayesianism changes the sign of the effect and gives it a reasonable value!

VII. CONCLUDING REMARKS

The most superficial answer to the question posed in the title is that people have generally been taught classical methods rather than Bayesian methods. Thus we may ask why the (relatively few) influential teachers adopted the classical point of view. Much of the answer lies in the fact that the entrance of one's prior knowledge (or beliefs) can be postponed in classical statistics until the constructed confidence intervals are used as input to a decision. This neatly separates the reporting of confidence intervals from their interpretation by individual readers. In Bayesian statistics, prior knowledge is incorporated from the beginning. This may be viewed as a virtue when one is looking at the logical consistency of a statistical paradigm,[12] but it continues to be viewed as a defect by scientists who seek to report their results in the most objective manner.

Both classical and Bayesian statisticians can agree on the importance of the likelihood function, in particular on its value as a concise *summary* of the experimental data.[40] Indeed, one often publishes a graph of $\mathscr{L}$ when it is asymmetric or useful for showing correlations among parameters. This is a particularly good practice, since it allows a reader to construct either Bayesian or (approximate) classical intervals, at any desired confidence level.

What are the biggest obstacles to widespread use of Bayesian confidence intervals in particle physics? First in my mind is our nearly universal insistence on frequentist coverage of the unknown true value. Bayesian methods with prior densities for Poisson mean suggested in the statistics literature can lead to intervals which severely undercover, and hence are unacceptable in the consensus of particle physicists, though occasionally studied.[10,11] However, a Bayesian method which yields intervals with the requisite frequentist coverage encounters much less resistance and, as discussed in Sec. IV, can even earn the supposed accolade "conservative."

The other big obstacle, related to the first, is the need to specify the prior density. Once one realizes that the naive objective choice of uniform prior is in fact an arbitrary choice (since one has to choose the metric in which the prior is uniform), the situation is very difficult. As illustrated in the case of Poisson statistics, particle physicists favor prior densities which lead to good frequentist coverage (or better yet, the same upper limits as the classical method), even when they are not the prior densities which make the most sense to a pure Bayesian. For this reason, I am not sure that continuing developments[41] in the research statistics literature will be of much influence in particle physics.

As the examples have shown, both classical and Bayesian methods can lead to results which would be unacceptable to most particle physicists. Thus, there is an uneasy equilibrium, in which typically classical intervals are used, unless they give unacceptable results. In that case, one typically turns to a Bayesian result, as long as it provides frequentist coverage or overcoverage! This approach can be charitably described as pragmatic,[42] and is enhanced if a graph of the likelihood function is provided whenever its shape cannot be easily inferred.

The pragmatic approach works well enough so that most of the time, we do not concern ourselves with the philosophical issues of statistical inference. However, all physicists should at least be aware that computation of confidence intervals can lead to some real mind-benders. Niels Bohr supposedly said[43] that if quantum mechanics did not make you dizzy then you did not really understand it. I think that the same can be said about statistical inference!

ACKNOWLEDGMENTS

Much of this work was inspired or carried out while teaching occasional graduate seminars on data analysis in particle and nuclear physics, and I thank the students for their interest. I thank Virgil Highland and Frederick James for helping me to understand the issues discussed here, and for their encouragement and advice on revising an earlier draft. The referees provided a number of useful suggestions and several additional references. Portions of this work related to my particle physics research were supported by the U. S. Department of Energy.

[1]As is common, I use "particle physics" as a shorthand for "elementary particle physics."

[2]I adopt this simplistic dichotomy between classical and Bayesian methods (using the terms "classical" and "frequentist" synonymously) at the risk of offending statisticians of some persuasions. Nowadays classical statistics is amalgamation of the methods pioneered by the rivals R. A. Fisher and J. Neyman, who often disagreed. There are also raging debates splintering Bayesians into various factions.

[3]Among the more influential are H. Cramér, *Mathematical Methods of Statistics* (Princeton University, Princeton, NJ, 1958); W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics* (North Holland, Amsterdam, 1971); A. G. Frodeson, O. Skjeggestad, and H. Tøfte, *Probability and Statistics in Particle Physics* (Columbia University, New York, 1979). More recent books include L. Lyons, *Statistics for Nuclear and Particle Physicists* (Cambridge University, Cambridge, 1986); B. P. Roe, *Probability and Statistics in Experimental Physics* (Springer, New York, 1992).

[4]Jay Orear, "Notes on statistics for physicists, revised" [Cornell preprint CLNS 82/511 (1982), unpublished]. This primarily espouses classical likelihood ratios, but on one page describes the Bayesian construction.

[5]F. James, "Determining the statistical significance of experimental results" [CERN preprint DD/81/02 (1982), unpublished].

[6]Particle Data Group, "Review of particle properties," Phys. Rev. D **50**, 1173–1826 (1994). Confidence interval discussion is on pp. 1278–1282. Electron neutrino mass limits are on p. 1390. Discussion of the current combined LEP result for the number of neutrino types ($2.985\pm0.023\pm0.004$) is on pp. 1416–1417.

[7]Bayesian methods are popular in various branches of applied science; see for example, A. F. M. Smith, "An overview of the Bayesian approach," in *Bayesian Methods in Reliability*, edited by P. Sander and R. Badoux (Kluwer Academic, Dordrecht, 1991), pp. 15–79; K. Weise and W. Wöger, "A Bayesian theory of measurement uncertainty," Sci. Technol. **3**, 1–11 (1992). See also the workshop series of which the most recent published occurrence is *Proceedings of the Twelfth International Workshop on Maxi-*

mum Entropy and Bayesian Methods (1992, Paris, France), edited by Ali Mohammad-Djafari and G. Demoment (Kluwer Academic, Dordrecht, 1993).

[8]For advocacy of Bayesian statistics in this journal, see D. E. Raeside "A physicist's introduction to Bayesian statistics," Am. J. Phys. **40**, 688–694 (1972); "A physicist's introduction to Bayesian statistics. II," *ibid.* **40**, 1130–1133 (1972); H. W. Lewis, "What is an experiment?," *ibid.* **50**, 1164–1165 (1982); Stig Steenstrup, "Experiments, prior probabilities, and experimental prior probabilities," *ibid.* **52**, 1146–1147 (1984).

[9]Philip W. Anderson, "The Reverend Thomas Bayes, needles in haystacks, and the fifth force," Physics Today, **45**(1), 9–11 (1992). The context is actually hypothesis testing rather than confidence intervals. A more extensive related discussion is by W. H. Jefferys and J. O. Berger, "Ockham's razor and Bayesian analysis," Am. Sci. **80**, 64–72 (1992).

[10]Among the few articles in the particle physics literature advocating Bayesian methods for Poisson processes are O. Helene, "Upper limit of peak area," Nucl. Instrum. Methods **212**, 319–322 (1983); O. Helene, "Errors in experiments with small numbers of events," *ibid.* **228**, 120–128 (1984), with criticism by F. James, Nucl. Instrum Methods A **240**, 203–204 (1985); O. Helene, "Determination of the upper limit of a peak area," *ibid.* A **300**, 132–136 (1991); and H. B. Prosper, "A Bayesian analysis of experiments with small numbers of events," *ibid.* A **241**, 236–240 (1985). Helene's 1983 method for incorporating subtraction of background events has been adopted by the Particle Data Group; a curious claim for a classical derivation has been made by G. Zech, *ibid.* A **277**, 608–610 (1989).

[11]H. B. Prosper, "Small signal analysis in high-energy physics: A Bayesian approach," Phys. Rev. D **37**, 1153–1160 (1988); and comment by D. A. Williams, *ibid.* **38**, 3582–3583 (1988); and reply by H. B. Prosper, *ibid.* **38**, 3584–3585 (1988).

[12]B. Efron, "Why isn't everyone a Bayesian?" Am. Stat. **40**, 1–11 (1986); the discussion following indicates that the argument still rages among statisticians.

[13]As similarly noted by Eadie *et al.* (Ref. 3), physicists say "estimate," "measure," and "error," when a statistician says "guess," "point estimate," and "confidence interval," respectively. Furthermore, particle physicists say "Gaussian" and "Breit–Wigner" when a statistician says "normal" and "Cauchy," respectively.

[14]I use the statistician's notation $N(\mu,\sigma)$ to denote a Gaussian (normal) curve centered at μ with rms deviation σ.

[15]Confidence level of "68%" is more precisely 68.27%, corresponding to plus-or-minus one standard deviation for a Gaussian. Statisticians define the variable α equal to 1 minus the C. L. Although this article uses 68% for definiteness, the generalization to other α is obvious, for example, $\alpha/2$ instead of 0.16 in Eq. (6).

[16]J. Neyman, Philos. Trans. R. Soc. London Sect. A **236**, 333–380 (1937). Reprinted in *A Selection of Early Statistical Papers on J. Neyman* (University of California, Berkeley, 1967), pp. 250–289. See in particular pp. 250–252, 261–268, and 285–286 of the reprint. Unfortunately, the quantity which Neyman calls α is precisely what is called $(1-\alpha)$ in modern references.

[17]I follow usual practice in using the vertical line to mean "given," except in this common notation for $\mathscr{L}$, which has a schizophrenic bidirectional nature. Since m_0 can be considered to be what is actually given, some prefer to write $\mathscr{L}(m_t|m_0)$ rather than $\mathscr{L}(m_0|m_t)$.

[18]F. James and M. Roos, "MINUIT—A system for function minimization and analysis of the parameter errors and correlations," Comput. Phys. Commun. **10**, 343–367 (1975). An updated version of the routine is maintained in the CERN Program Library.

[19]F. James, "Interpretation of the shape of the likelihood function around its minimum," Comput. Phys. Commun. **20**, 29–35 (1980).

[20]D. A. S. Fraser, "Statistical inference: Likelihood to significance," J. Am. Stat. Assoc. **86**, 258–265 (1991), and references therein.

[21]As an elementary statement of two ways to compute the probability that an event is in both sets A and B, Bayes's theorem is unassailable: $P(A|B)P(B)=P(B|A)P(A)$. The controversy arises when either A or B is taken to be a hypothesis or an unknown true value.

[22]A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics, Vol. 1, Distribution Theory*, 5th ed. (Oxford University, New York, 1987); see also earlier editions by Kendall and Stuart. Uninformative priors are discussed on pp. 283–289.

[23]A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics, Vol. 2, Classical Inference and Relationship*, 5th ed. (Oxford University, New York, 1991); see also earlier editions by Kendall and Stuart. Constraints are discussed on pp. 1227–1228.

[24]A variety of *ad hoc* methods to deal with the constraint were critically discussed by Virgil L. Highland, "Estimation of upper limits from experimental data," COO-3539-38, Temple University Physics Dept. (1986–7, unpublished); see also R. G. H. Robertson and D. A. Knapp, "Direct measurements of neutrino mass," Annu. Rev. Nucl. Part. Sci. **38**, 185–215 (1988); D. Denegri, B. Sadoulet, and M. Spiro, "The number of neutrino species," Rev. Mod. Phys. **62**, 1–42 (1990), especially the Appendix.

[25]In particle physics, "mass" m always means "rest mass." Masses are commonly quoted in terms of their equivalent energy, in electron volts (eV).

[26]F. James and M. Roos, "Statistical notes on the problem of experimental observations near an unphysical region," Phys. Rev. D **44**, 299–301 (1991).

[27]G. S. Abrams *et al.*, "Measurement of Z-boson resonance parameters in e^+e^- annihilation," Phys. Rev. Lett. **63**, 2173 (1989).

[28]F. Garwood, "Fiducial limits for the Poisson distribution," Biometrika **28**, 437–442 (1936).

[29]W. E. Ricker, "The concept of confidence or fiducial limits applied to the Poisson frequency distribution," J. Am. Stat. Assoc. **32**, 349–356 (1937).

[30]More complicated criteria can be imposed using Neymans's original construction, which in the case of one random variable is called the method of *confidence belts*, described in Ref. 3.

[31]E. L. Crow and R. S. Gardner, "Confidence intervals for the expectation of a Poisson variable," Biometrika **46**, 441–453 (1959).

[32]The literature also uses the terms vague priors and noninformative priors.

[33]Harold Jeffreys, *Theory of Probability*, 3rd ed. (Clarendon, Oxford, 1983). See in particular pp. 135–138. The several editions of this book were important in the development of Bayesian statistics in this century.

[34]E. T. Jaynes, "Prior probabilities," IEEE Trans. Syst. Sci. Cybernet. **SSC-4**, 227–241 (1968). This article is also widely known for its advocacy of the maximum entropy method when the unknown parameter takes on only discrete values. A related exchange in the particle physics literature is in Ref. 11.

[35]H. Jeffreys, "An invariant form for the prior probability in estimation problems," J. R. Statist. Soc. Ser. A **186**, 453–461 (1946); G. E. P. Box and G. C. Tiao, *Bayesian Inference and Statistical Analysis* (Addison-Wesley, Reading, MA, 1973), see in particular pp. 25–46; R. E. Kass, "Data-translated likelihood and Jeffreys's rules," Biometrika **77**, 107–114 (1990); J. G. Ibrahim and P. W. Laud, "On Bayesian analysis of generalized linear models using Jeffreys's prior," J. Am. Stat. Assoc. **86**, 981–986 (1991), and references therein.

[36]R. D. Cousins and V. Highland, Nucl. Instrum. Methods A **320**, 331–335 (1992), and references therein.

[37]The uncertainly often comes from the *multiplicative* effect of many things, so that S is more properly sampled from a log-normal distribution, but the Gaussian approximation is nearly always used.

[38]In general, integrals may be evaluated by Monte Carlo simulations, and vice versa. A Monte Carlo simulation of a Bayesian method includes sampling of "true" values from a pdf in the true value. In contrast, a Monte Carlo simulation of frequentist coverage of any method selects trial true values, and for a given true value samples observed values, computing intervals and testing for coverage.

[39]For another situation where introducing a continuous parameter has an interesting effect, see R. D. Cousins, "A method which eliminates the discreteness in Poisson confidence limits and lessens the effect of moving cuts specifically to eliminate candidate events," Nucl. Instrum. Methods A **337**, 557–565 (1994).

[40]B. Efron, "Maximum likelihood and decision theory," Ann. Stat. **10**, 340–356 (1982).

[41]For recent discussion in the statistics literature, see *Bayesian Statistics 4, Proceedings of the Fourth Valencia International Meeting*, edited by J. M. Bernardo *et al.* (Clarendon, Oxford, 1992); in particular, articles by D. V. Lindley; J. O. Berger and J. M. Bernardo; and J. K. Ghosh and R. Mukerjee; and the discussion following them.

[42]There has long been interest in the statistics literature in studying frequentist coverage yielded by various Bayesian priors, for example: B. L. Welch and H. W. Peers, "On formulas for confidence points based on integrals of weighted likelihoods," J. R. Stat. Soc. B **25**, 318–329 (1963); R. Tibshirani, "Noninformative priors for one parameter of many," Biometrika **76**, 604–608 (1989).

[43]N. D. Mermin, "What's wrong with this pillow?," Physics Today, **42**(4), 9 (1989).