Systematic Errors

Joel Heinrich[1] and Louis Lyons[2]

[1] Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104; email: heinrich@hep.upenn.edu

[2] Department of Physics, University of Oxford, Oxford OX1 3RH, United Kingdom; email: l.lyons@physics.ox.ac.uk

Annu. Rev. Nucl. Part. Sci. 2007. 57:145–69

First published online as a Review in Advance on May 2, 2007

The *Annual Review of Nuclear and Particle Science* is online at http://nucl.annualreviews.org

This article's doi:
10.1146/annurev.nucl.57.090506.123052

0163-8998/07/1123-0145$20.00

Key Words

systematics, nuisance parameters, confidence interval, *p*-values, Bayes, frequentism

Abstract

To introduce the ideas of statistical and systematic errors, this review first describes a simple pendulum experiment. We follow with a brief discussion of the Bayesian and frequentist approaches. Two widely used applications of statistical techniques in particle physics data include extracting ranges for parameters of interest (e.g., mass of the *W* boson, cross section for top production, neutrino mixing angles, etc.) and assessing the significance of possible signals (e.g., is there evidence for Higgs boson production?). These two topics are first discussed in the absence of systematics, and then methods of incorporating systematic effects are described. We give a detailed discussion of a Bayesian approach to setting upper limits on a Poisson process in the presence of background and/or acceptance uncertainties. The relevance of the choice of priors and how this affects the coverage properties of the method are described.

Contents

שגיאות מי יבין מנסתרות נקני

Who can understand errors? Save me from hidden effects.
Psalm 19:13 ± 1

1. INTRODUCTION

1.1. A Trivial Example: The Simple Pendulum

The first experiment that many physics students perform at school is to measure g, the acceleration due to gravity, using a simple pendulum of length l. The time T for n swings is measured to determine the period $\tau = T/n$, and then g is $4\pi^2 l/\tau^2$. We use this example to introduce the basic ideas of statistical and, especially, systematic errors, the latter being the focus of this review. [See References 1–6 for previous reviews and notes on systematics, and the many research articles in the PHYSTAT series of conferences and workshops (7–12).]

Contributions to the statistical error on g arise from the limited accuracy with which we can measure T and l. These can be assessed from the measuring instruments, the quantities to be measured, and the experimentalists. Because the result will vary for repetitions of the measurement, the estimated statistical error can be checked with the standard deviation of a large set of independent measurements. Furthermore, because the statistical errors on a series of independent measurements are uncorrelated, their effect can be reduced by combining the results of such measurements. The usefulness of this reduction in statistical error may be limited by the fact that the systematic error does not automatically go down.

The measured quantities may also have systematic uncertainties, which could arise from uncertainties in the calibration of the clock and ruler used to measure T and l, respectively. These quantities can be determined by performing our own calibration in some subsidiary experiment. For example, we could use the ruler to estimate the length of an object, whose size is precisely known. In doing so we hope to have reduced the systematic error and replaced it by the smaller uncertainty of how well we know the new calibration of our ruler. We return to a more detailed consideration of subsidiary experiments in Sections 4.1.4, 4.1.5, and, especially, 5.

There are, however, other sources of possible systematics:

- The formula quoted for τ applies to *undamped* oscillations of a *small amplitude* of a simple pendulum, which consists of a *point mass* suspended by a *massless*, *extensionless string* from a *rigid support*. All italicized items are only approximated in a real experiment, and so corrections must be estimated for them or the result will be biased. The uncertainties in these corrections are systematics.
- We may be interested in the value of g at sea level, rather than at the elevation at which we performed the measurement. This requires a correction that depends not only on our elevation, but also on the material below us. We may have access to other people's estimate of this correction, but its uncertainty is a systematic. It is also possible that more than one such estimate exists, in which case we must decide how to combine them and how to deal with any discrepancies among the estimates.

The nature of systematic effects is such that they may not cause different answers when the experiment is repeated. Thus, a consistent set of results does not imply the absence of systematics. Furthermore, as is already apparent from the pendulum example, systematic effects occur not only on directly measured quantities. Thus, in general, the reliable assessment of systematics requires much more thought and work than for the corresponding statistical error.

Some errors are clearly statistical (e.g., those associated with the reading errors on T and l), and others are clearly systematic (e.g., the correction of the measured g to its sea level value). Others could be regarded as either statistical or systematic (e.g., the uncertainty in the recalibration of the ruler). Our attitude is that the type assigned to a particular error is not crucial. What is important is that possible correlations with other measurements are clearly understood.

The result of the experiment may be quoted as $g \pm \sigma_{\text{stat}} \pm \sigma_{\text{syst}}$, where the statistical and systematic errors are shown separately. If a single error is required, then typically σ_{stat} and σ_{syst} are combined in quadrature, on the grounds that they are uncorrelated. At the other extreme, some or all of the systematic errors can be shown individually. This would be useful in combining different measurements, for which some of the systematic effects may be correlated between the different measurements. Also, it is possible that in the future there may be an improvement in some relevant external information. (In the pendulum example, this could be the correction to sea level.) The way this reduces the systematic error can readily be assessed if this particular contribution is quoted separately.

The relevance of correlations can be seen by extending the above example to consider measuring the ratio of the g-values at two different locations. Then, for example, calibration errors in the T measurements at the two locations could cancel.

The examples discussed below are basically extensions of this simple case, but include the effects of non-Gaussian errors. Also the Poisson distribution, with its discrete observations, usually plays an important role in particle physics analyses.

1.2. Structure of this Review

The differences in the Bayesian and frequentist philosophies are described in Section 2, as is the construction of credible or confidence intervals for parameters. Section 3 deals with p-values for assessing the significance of any potential discoveries. These two sections largely avoid discussion of systematics. The various ways of including these effects for intervals and for p-values are considered in Section 4. A detailed example of calculating upper limits can be found in Section 5, while the last two sections contain a few miscellaneous items and the conclusions. The Supplemental Material (follow the Supplemental Material link from the Annual Reviews home page at **http://www.annualreviews.org**) contains a glossary.

2. BAYES AND FREQUENTISM

2.1. Probability

To a Bayesian, probability is interpreted as the degree of belief in a statement. It can vary from person to person because they can have different information about a situation. It is quantified by the fair bet: A Bayesian should be prepared to accept bets in either direction, with odds determined by the numerical value he assigns for the probability.

In contrast, frequentists define probability via a repeated series of almost identical trials; it is the limit of the fraction of successes as the number of trials tends to infinity Thus, frequentists will not assign a probability to a one-off event (e.g., will the first astronaut to Mars return to Earth alive?) or to the value of a physical constant (e.g., is the value of the strong coupling constant between 0.110 and 0.115?).

2.2. Bayes' Theorem

Bayes' theorem is derived from $P(A$ and $B)$, the probability that two events A and B both happen. Then

$$P(A \quad \text{and} \quad B) = P(A\,|\,B)P(B) = P(B\,|\,A)P(A), \qquad 1.$$

where $P(B)$ is the probability of B happening, while $P(A|B)$ is the conditional probability of A happening given that B has occurred. For example, A could be the probability that a high-energy proton-proton collision contains a top quark, and B the probability that it contains a W boson. Then Bayes' theorem

$$P(A\,|\,B) = P(B\,|\,A)P(A)/P(B) \qquad 2.$$

relates $P(A|B)$ to $P(B|A)$. Frequentists regard Bayes' theorem as completely noncontroversial, provided that the probabilities occurring in it are acceptable frequentist probabilities.

The dispute occurs when Bayesians choose B as data, and A as one or more parameter values. This then gives

$$p(\mu|x) \sim p(x|\mu)\pi(\mu), \qquad 3.$$

where the implied constant of proportionality is simply the normalization constant $1/\int p(x|\mu)\pi(\mu)\,d\mu$; the likelihood $p(x|\mu)$ is derived from the probability-density function (*pdf*) for obtaining measurements x, given the parameter μ; and the Bayes prior $\pi(\mu)$ specifies the assumed probability density for μ, before the experiment was performed. Here we use P to denote probabilities of discrete variables, and p or π for probability densities of continuous variables. In contrast $p(\mu|x)$ is the Bayesian posterior for μ and gives the probability density for μ after the data are obtained; it is determined by both the likelihood function and the prior. Bayes' theorem thus provides a way of using the data from our experiment to update our prior knowledge about a parameter.

All this is unacceptable for frequentists because they would object to assigning a probability distribution to a physical parameter μ. A Bayesian would counter this by explaining that the frequentist view of probability is too narrow and that Bayes' theorem should be interpreted in terms of degrees of belief.

2.3. Bayesian Priors

A practical problem arises in the Bayesian approach because of the need to choose the prior $\pi(\mu)$. If the value of the parameter had been well determined in some previous measurement as $a \pm b$, $\pi(\mu)$ could perhaps be a Gaussian distribution centered on a and with variance b^2. However, if we are looking for some previously unobserved process, it is more likely that very little is known prior to our experiment. Then we need to choose $\pi(\mu)$ to express our relative ignorance about μ.

An unreasonably popular choice is the uniform distribution

$$\pi(\mu) = \text{constant}, \qquad 4.$$

as it favors no particular value of μ. However, this prior implies that the range from 0 to 1 for μ is only as likely as that from 176,391.3 to 176,392.3. Another feature of a uniform prior over an infinite range is that it cannot be normalized.

Furthermore, although there does seem to be something natural in using a uniform distribution as an uninformative prior, the choice is in fact far from obvious. Priors uniform in μ^2, $\ln(\mu)$, or $1/\sqrt{\mu}$, and so on may be equally plausible and are different from our initial choice. There is thus arbitrariness in how ignorance is parametrized, and this will affect our posterior probability distribution.

The problem is exacerbated in several dimensions, with uniform priors having more undesirable properties. For example, in an analysis involving several different final states, each of which has its own acceptance A_i, the total acceptance $A_{\text{tot}} = \sum A_i$. Then priors that are uniform in each A_i (for positive A_i) correspond to a prior that grows with A_{tot}, and it is not obvious that this is what was intended (see Section 5).

The variety of Bayesians includes those who prefer subjective priors or objective ones. The former are happy that a different prior could be chosen by each scientist, as this encapsulates their varying degrees of knowledge and beliefs. Objective Bayesians try to find priors with desirable theoretical properties. Examples include the Jeffreys priors $1/\mu$ and $1/\sqrt{\mu}$ for specific cases.

A positive feature of the Bayesian approach is that it is simple to incorporate the fact that part of the infinite range for μ may be unphysical (e.g., a reaction rate or the mass of a particle should not be negative). Whatever functional form for $\pi(\mu)$ is used in the allowed region, it is set equal to zero where μ is unphysical. This ensures that any Bayesian interval for a parameter will always be completely physical.

2.4. Bayesian Intervals

The output of using Bayes' theorem is $p(\mu|x)$, the posterior probability distribution for the parameter of interest. This is supposed to contain a complete summary of what is known about μ. From it, various intervals at a given credible level α can be extracted. These could include, for example, upper or lower limits, central intervals [with probability $(1-\alpha)/2$ on each side], or those defined by the largest probability density or some other ordering rule. Of all intervals at level α, those selected according to highest probability density have the shortest length in μ. However, they are not invariant with respect to nonlinear transformation of the parameter μ, for example, to μ^2 or to $1/\mu$. If there are several parameters of interest, the posterior $p(\mu_1, \mu_2, \ldots)$ can similarly be used to define a region in the space of the parameters at a credible level α.

2.5. Frequentist Approach

2.5.1. The Neyman construction.

The frequentist or classical approach to parameter determination is very different from the Bayesian one. It uses only $p(x|\mu)$ and never considers a prior $\pi(\mu)$ or posterior $p(\mu|x)$.

Figure 1 illustrates the construction of frequentist confidence intervals for a single non-negative parameter μ and a measurement x that depends on μ. For example, μ could be the temperature at the center of the Sun, and x the measured flux of solar

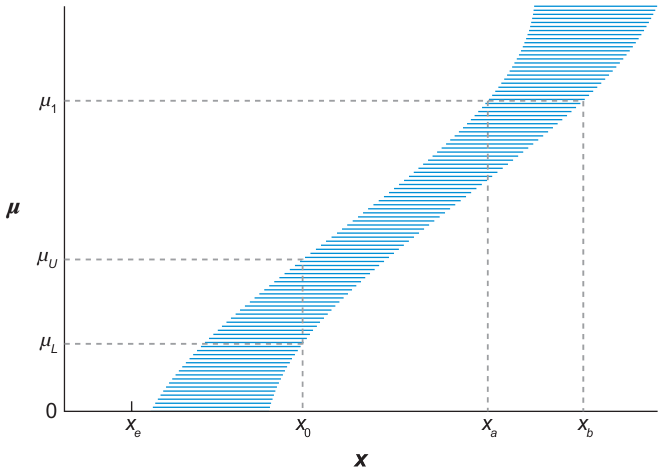

Figure 1

The Neyman construction for a confidence band at probability level α. It makes use of the *pdf* $p(x|\mu)$, which specifies the probability density for a measurement x for any value of the non-negative parameter μ. For the parameter value μ_1, there is a probability α of the measurement being in the region from x_a to x_b. This procedure is repeated for all possible μ, to build up the confidence belt (*shaded area*). For a particular measurement x_0, the values μ_L and μ_U at the edges of the confidence belt are found. The confidence interval from μ_L to μ_U thus specifies the values of μ for which our measurement x_0 is likely (at level α). For the shown confidence belt, a smaller measurement x_e would result in an empty confidence interval; there are no values of μ for which x_e is likely.

neutrinos, as determined by running a large neutrino detector for a month. We assume we know the *pdf* $p(x|\mu)$ for all physical μ. For a given μ, such as μ_1 in **Figure 1**, we can use the *pdf* to construct an interval from x_a to x_b such that the probability of x falling in this range is α, for example, 0.9. When x is continuous, it is possible to choose an x range such that the included probability is exactly α. This will not be so in general when x is discrete, in which case the conservative choice of having the probability at least equal to α is made.

The choice of a range in x is then repeated for every possible μ until the shaded area in **Figure 1** is built up. This shows likely measured results for any μ. Now assume we perform a measurement and obtain x_0. By finding where a vertical line at x_0 cuts the shaded region, limits μ_L and μ_U are established for μ. These then specify the range of values for μ for which x_0 is a likely result, as favored by the ordering rule (see Sections 2.5.2, 2.5.3, and 4.1.5).

Thus, the frequentist statement that $\mu_L \leq \mu \leq \mu_U$ at the 90% confidence level is a probability statement about μ_L and μ_U; if the measurement was repeated very many times and the range extracted for each of them, (at least) 90% of them should contain the true value of μ. This is known as coverage. This contrasts with the similar-looking statement by Bayesians. For them, μ_L and μ_U are determined in the particular experiment and are fixed. Then the statement specifies the percentage of the posterior probability distribution for μ within this range.

2.5.2. Feldman and Cousins' method.

The prescription for constructing a range in x for each μ is not unique in that there are infinitely many ranges that contain a fraction α of the *pdf*, and so a choice must be made among them; this is usually specified by an ordering rule. Thus, a central interval could have $(1-\alpha)/2$ of the *pdf* on either side of the range, a range from x_L to $+\infty$ would be suitable for upper limits on μ, and so on. Feldman & Cousins (13) have exploited this freedom of choice, already noted by Neyman, to interesting effect.

In their unified approach to parameters with a physical bound (e.g., $\mu \geq 0$), Feldman and Cousins choose a likelihood-ratio ordering rule (see the Supplemental Material and Section 4.1.5), which results in favoring smaller values of the data x in the construction of the confidence interval. This has the consequence of greatly reducing the probability of having confidence intervals that contain only $\mu = 0$. Another feature of the Feldman-Cousins approach is that there is no longer the arbitrariness of choosing central intervals, upper limits, and so on (hence the name unified).

2.5.3. Several parameters.

In principle, the Neyman construction of **Figure 1** for one dimension of data and one physical parameter can be extended to cases where there are more data dimensions and/or parameters. For each possible combination of parameters, the region in data space at confidence level α is determined according to the chosen ordering rule. The actual data then serve to define a confidence region in the multiparameter space. This can then be projected down onto any given axis to produce a confidence interval for that parameter. This almost always results in overcoverage, which becomes more serious as the number of parameters increases. It thus becomes important to choose a good ordering rule so as to produce a region in parameter space that is narrow for the parameter of interest but elongated in the directions of the others. Because the ordering rule applies to the data at fixed parameter values, but the desired shape is in parameter space, this is nontrivial (14, 15). (See Section 4.1.5 for a fuller description.)

2.6. Comparison of Bayes and Frequentist Approaches

The main advantage of the frequentist method is that it does not require a prior and hence tends to be favored by particle physicists who want the data to speak for themselves and who prefer not to incorporate personal beliefs into their result. Another positive feature is that coverage is guaranteed. However, it becomes increasingly difficult computationally to perform the frequentist Neyman construction as the number of parameters and measurements increases.

Bayesians could argue that frequentists address the boring question about which values of the parameter are such that the data are likely, rather than the more interesting one about what parameter values we favor (after having performed the experiment and in light of all we previously knew about it). Bayesians also point out that frequentists need to define an ensemble of which their measurement is a member. Unfortunately, the choice of ensemble is often not unique (see the Supplemental Material and Reference 16).

Bayesian intervals are always physical, whereas frequentist ones can be empty. Frequentists have the freedom (or ambiguity) to choose the ordering rule, which results in different possible confidence ranges. This in some way mirrors the different credible intervals that can be extracted from a Bayesian posterior. Bayesian intervals can also be sensitive to the choice of the prior, especially for the case of upper limits.

2.7. Other Methods

Not all statistical procedures are Bayesian or frequentist. What distinguishes these two techniques is that they claim some self-consistent justification for their approach, whereas other methods are more ad hoc. Hence, the other possibilities do not usually achieve either coverage or Bayesian credibility.

Alternatives for parameter determination include the following:

- χ^2: This usually is used for a histogram of data. For large enough expected numbers of observations for each bin, the Gaussian approximation for the bin contents may be adequate. Then S, the weighted sum over all bins of squared deviations, behaves like χ^2 for the appropriate number of degrees of freedom. The best values of the parameters are determined by minimizing S, and regions in parameter space are defined by letting S increase from its minimum by the appropriate amount, depending on the effective number of free parameters and the required confidence level.
- Maximum likelihood: A confidence region for the parameters $p_1, p_2, \ldots, p_n$ involves determining a region in parameter space within a contour $\ln L(p_1, p_2, \ldots, p_n) = \ln L_{\max}(p_1, p_2, \ldots, p_n) - k$, where k is an appropriately chosen constant. Alternatively, the likelihood could be profiled over several uninteresting parameters. Then if, for example, only one parameter p_1 remains, the profile likelihood is defined by $L_{\text{prof}}(p_1) = L(p_1, q_2, \ldots, q_n)$, where q_j are the values of p_j that maximize L for each particular p_1. Then the range corresponding to ±1 error on p_1 is taken conventionally by finding where $\ln(L_{\text{prof}}(p_1))$ decreases by 0.5 compared with $\ln(L_{\max})$. Adjusted likelihood methods (17) have been developed because the standard likelihood approach may not result in good coverage properties.
- Mixed methods: Because of the desire to use a frequentist method for the parameter of interest, and because of the practical difficulties of using a fully frequentist method for more than one or two nuisance parameters, Cousins & Highland (18) advocate using a mixed method, where a Bayesian approach is used for the systematics but the Neyman construction is used for the physics parameter. Their paper describes this technique for upper limits, but the mixed approach can be used in other contexts too. This is discussed further in Section 4.1.6.

3. *p*-VALUES

3.1. What *p*-Values Are

Both significance (e.g., claiming discovery of a new resonance) and goodness of fit (e.g., does the new resonance have a Breit-Wigner shape?) are traditionally quantified by p-values. A statistic t (a function of the observed data) is chosen, which becomes larger as the disagreement between the data and an assumed model increases. The p-value is the probability (when the assumed model is actually correct, the experiment is unbiased, etc.) that the value of t, over many repetitions of the experiment, will be greater than or equal to that observed in the data. (Because the calculation of a p-value involves an ensemble of experiments, the use of p-values is inherently non-Bayesian.) A small p-value is evidence that the assumed model does not match the data, casting doubt on any parameters derived from a fit using the model.

For example, if data x are expected to have a Gaussian distribution centered on zero and with unit variance, the p-value is simply the fractional area in the tail(s) beyond the measured x_0. (Whether the one-sided or two-sided tail area is considered depends on whether deviations in either direction are considered significant. Thus, in searching for oscillations of solar neutrinos, a reduced flux is expected, so only the lower tail is relevant.) The one-sided tail probability corresponding to $x_0 = 5.0$ is 3×10^{-7}. Often, in non-Gaussian situations, p-values are converted into the equivalent number of standard deviations for a Gaussian distribution, thereby providing a number that is easier to remember.

A second example is a data histogram with n_i events in the i-th bin, when μ_i is the expectation from our model. Pearson's χ^2 statistic

$$\chi^2_P(n_1, n_2 \ldots n_N) = \sum_{i=1}^{N} \frac{(n_i - \mu_i)^2}{\mu_i}$$

is often used to quantify goodness of fit. Here, μ_i is the variance of a Poisson distribution with mean μ_i. (Soon we are going to let μ_i have an uncertainty.) The p-value associated with the goodness of fit of the data to our model is then calculated using Poisson probabilities:

$$P = \sum_{\mathrm{K}} \left[\prod_{i=1}^{N} \frac{e^{-\mu_i} \mu_i^{k_i}}{k_i!} \right],$$

where K is the set of all $(k_1, k_2 \ldots k_N)$ such that $\chi^2_P(k_1, k_2 \ldots k_N) \geq \chi^2_P(n_1, n_2 \ldots n_N)$. When μ_i are sufficiently large and assuming the model is correct, the distribution of χ^2_P is approximately independent of the actual values of μ_i, and the p-value can be approximated by the tail area beyond $t_0 = \chi^2_P(n_1, n_2 \ldots n_N)$ of a mathematical χ^2 distribution for N degrees of freedom (assuming no free parameters). A small p-value leads to rejection of the model. When the model under which the p-value is calculated has nuisance parameters (i.e., systematic uncertainties), the proper computation of the p-value is more complicated (see Section 4.2).

3.2. Discrete Data

When x is a continuous variable, the ideal p-value distribution is uniform between zero and unity. This implies that Prob $(p \leq p_0) = p_0$ for any p_0 between zero and unity. When x is discrete (e.g., for the data of a Poisson distribution), so are the possible p-values, and thus their distribution cannot be completely uniform. The discrete distribution is most uniform when the above equation is true for all achievable p-values.

3.3. What p-Values Are Not

It is vital to remember that a p-value is not the probability that the relevant hypothesis is true. Thus, statements such as “our data show that the probability that the Standard Model is true is below 1%” are incorrect interpretations of p-values. Similarly, a p-value of 10% or larger is not evidence that the null hypothesis is true; it is merely that the data are not inconsistent with it.

4. METHODS OF INCLUDING SYSTEMATICS

4.1. Parameter Estimation

We describe methods for incorporating systematics into measurements and limits.

4.1.1. Shift method. The shift method, based on linear propagation of errors, is simple but not always applicable. Given N nuisance parameters μ_i with uncorrelated Gaussian uncertainties σ_i, and an estimator of the parameter of interest $f(\mu_1, \mu_2, \ldots, \mu_N)$, the linear approximation yields

$$\sigma_f^2 \simeq \sum_{i=1}^{N} \left(\frac{\partial f}{\partial \mu_i} \right)^2 \sigma_i^2,$$

where σ_f is the combined systematic uncertainty of the measurement. If f is approximately linear over the region $\mu_i \pm \sigma_i$, the partial derivatives can be approximated by finite differences as

$$\frac{\partial f}{\partial \mu_i} \simeq \frac{f(\mu_1, \mu_2, \ldots, \mu_i + \sigma_i, \ldots, \mu_N) - f(\mu_1, \mu_2, \ldots, \mu_i, \ldots, \mu_N)}{\sigma_i} \equiv \frac{\Delta_i}{\sigma_i},$$

and one obtains $\sigma_f^2 \simeq \sum_{i=1}^{N} \Delta_i^2$. That is, one adds the 1σ shifts in quadrature. Given the full error matrix V_{ij} including correlations, this becomes

$$\sigma_f^2 \simeq \sum_{i,j} \Delta_i \Delta_j \left(\frac{V_{ij}}{\sigma_i \sigma_j} \right),$$

where the matrix in parentheses, the correlation matrix, has by construction ones along its main diagonal. (Contrast the multisim approach of Section 6.3.3.)

For example, m_H, the mass of the Higgs boson, could be determined from a fit to a mass spectrum, consisting of a peak above a background, some of which is caused by $t\bar{t}$ events. To estimate the contribution from the uncertainty of the mass of the top quark on a measurement of m_H, one simply reanalyzes the data with the top quark mass shifted by its 1σ uncertainty. Typically, this would also necessitate producing new Monte Carlo events at the shifted top mass, a straightforward, albeit time-consuming, task. The statistical error on the estimate of this contribution can be reduced by varying the nuisance parameter by more than one error. Then the resulting change is divided by the appropriate factor. This requires the dependence of the answer on the nuisance parameter to be linear over a wider range.

When one is setting a limit, rather than obtaining a point estimate with associated uncertainty, the shift method is not applicable. If the function f is significantly nonlinear in the nuisance parameters, the shift method is not reliable.

4.1.2. Likelihood method.

Here we need the full likelihood as a function of both the parameters of interest and the nuisance parameters. (Although the likelihood also depends on the observed data, only its dependence on the parameters is explicit below.) Ideally, one uses the product of the actual likelihood functions from the main and subsidiary measurements. [Punzi (15) has discussed the case where only a range is defined for the nuisance parameter.] Otherwise, should only estimates $\hat{\mu}_i$ with an error matrix V_{ij} be available for a set of nuisance parameters μ_i, one typically approximates the likelihood associated with those nuisance parameters by a multidimensional Gaussian:

$$L \simeq \exp\left(-\frac{1}{2}\sum_{i,j}(\mu_i - \hat{\mu}_i)V_{ij}^{-1}(\mu_j - \hat{\mu}_j)\right).$$

We define

$$\ell(\vec{\mu}) = -2\sum_k \ln(L_k),$$

where L_k represent the component likelihoods (from independent parts of the overall *pdf*) for the parameters of interest and nuisance parameters, and the vector $\vec{\mu}$ contains both parameters of interest and nuisance parameters. [It is numerically more convenient to deal with −2 times the logarithm of the likelihood, as the value of the likelihood is often outside the range of floating point representation. The factor −2 is introduced by convention; with this factor, $\ell(\vec{\mu})$ is a χ^2 in the Gaussian case.] Given a starting point $\vec{\mu}_0$ close to the minimum of $\ell(\vec{\mu})$, the minimization package MINUIT (19) will numerically find the values $\hat{\vec{\mu}}$ that minimize $\ell(\vec{\mu})$, which represent maximum likelihood estimates for the parameters of interest and nuisance parameters. MINUIT (via its MIGRAD routine) will also produce a complete error matrix V_{ij}, which derives from the shape of the likelihood, for the parameters. By using $L = L_{\text{main}}L_{\text{subsid}}$, the statistical and systematic uncertainties are fully incorporated into the errors on the parameters of interest.

This method uses a numerically more precise approximation of the likelihood than the shift method. However, the method is not recommended for limits and assumes that the shape of the likelihood near its maximum can be approximated by a multidimensional Gaussian.

4.1.3. Profile likelihood.

Here one obtains estimates of the parameters by maximizing the likelihood as above, but the uncertainties are handled differently. To define the profile likelihood, we divide the parameters into a single parameter of interest μ and the rest of the parameters $\vec{\phi}$, writing $L(\mu, \vec{\phi})$. The profile likelihood with respect to μ is then

$$L_{\mathrm{P}}(\mu) = \text{maximum with respect to } \vec{\phi} \text{ of } L(\mu, \vec{\phi}).$$

Here also it is convenient to work with $\ell_{\mathrm{P}}(\mu) = -2\ln(L_{\mathrm{P}}(\mu))$. The uncertainties are given by

$$\ell_{\mathrm{P}}(\hat{\mu} \pm \sigma_{\pm}) - \ell_{\mathrm{P}}(\hat{\mu}) = 1,$$

where $\hat{\mu}$ is the value of μ that maximizes $L_{\mathrm{P}}(\mu)$ and can be calculated using the MINUIT MINOS routine (see Reference 20 for further discussion). Because in general $\sigma_+ \neq \sigma_-$, these errors can be asymmetric. When $\mu + \sigma_+$ or $\mu - \sigma_-$ is outside the allowed region for μ, the method can be unreliable. Otherwise, the method does a better job of handling both the correlations between the parameters and any non-Gaussian behavior of the likelihood. Asymmetric errors give more information about the shape of the profile likelihood, but their proper interpretation is not always clear (21).

4.1.4. Fully Bayesian.

The Bayesian approach requires a prior for the nuisance parameters. Because there may be correlations between the nuisance parameters, we write this as a joint nuisance prior $\pi(\vec{\phi})$. In cases where some groups of nuisance parameters are unrelated, the joint prior may be the product of several individual priors. Ideally, $\pi(\vec{\phi})$ would be derived from Bayesian posteriors provided by subsidiary measurements. Sometimes some portion of $\pi(\vec{\phi})$ is based, partially or wholly, on the physicist's judgment (or personal belief). It is important to state the source of the nuisance priors.

In some cases, the combined prior for μ and $\vec{\phi}$ is not factorizable into separate priors $\pi(\mu)$ and $\pi(\vec{\phi})$. For example, in a Poisson counting experiment with rate $s + b$, $1/\sqrt{s+b}$ is often suggested as a prior for the parameter of interest s, where the background rate b is the nuisance parameter. The posterior is then calculated as

$$p(\mu) = \frac{\int L(\mu, \vec{\phi})\,\pi(\mu, \vec{\phi})d\vec{\phi}}{\int\int L(\mu, \vec{\phi})\,\pi(\mu, \vec{\phi})\,d\vec{\phi}d\mu},$$

where $\pi(\mu, \vec{\phi})$ is the joint prior for all parameters.

When the prior can be factorized as $\pi(\mu, \vec{\phi}) = \pi(\mu)\pi(\vec{\phi})$, one may equivalently calculate the marginalized likelihood (or marginalized *pdf*)

$$L(\mu) = \int L(\mu, \vec{\phi})\,\pi(\vec{\phi})\,d\vec{\phi} \qquad 5.$$

first. Then one proceeds to treat the marginalized likelihood as one would a simple likelihood: multiply by the prior for μ to obtain the normalized posterior *pdf*:

$$p(\mu) = \frac{L(\mu)\pi(\mu)}{\int L(\mu)\pi(\mu)\,d\mu}.$$

As usual, the posterior is integrated to define an interval or limit. The Bayesian procedure works for all cases, for both one- and two-sided intervals, with no assumption of linearity or Gaussian shape. The marginalization integral may be difficult to calculate in some cases. One should check that the posterior is normalizable, in addition to checking for numerical accuracy, when integrating numerically.

The methods of Sections 4.1.1 through 4.1.3 are approximations of the fully Bayesian method, with a flat prior for μ. In the fully Bayesian approach, however, other priors are permitted, and the lack of a unique prior is viewed by frequentists as a drawback. This freedom of choice of prior extends to the nuisance parameters as well, as a nuisance prior for the main experiment is either based directly on personal belief or is the posterior of a subsidiary measurement, for which some choice of prior was also made. (A prior for the subsidiary measurement combined with the likelihood for the subsidiary measurement yields the subsidiary posterior; the subsidiary posterior becomes the nuisance prior in the main measurement.) Concerns about the answer's dependence on the choice of prior(s) may be alleviated by investigating the frequentist coverage of the method.

4.1.5. Fully frequentist. The fully frequentist method is based on the Neyman construction of confidence intervals. This requires the probability distribution of the data $\vec{x}$ given the parameter of interest μ and the nuisance parameters $\vec{\phi}$. In this case, $\vec{x}$ includes the data from the main measurement, and from the subsidiary measurements for the nuisance parameters.

To calculate intervals at confidence level α, one proceeds as follows: For each allowed value of $(\mu, \vec{\phi})$, one selects a region $\mathrm{R}(\mu, \vec{\phi})$ in $\vec{x}$ space that would contain the result of the experiment (both the main and subsidiary experiments) with probability = α (discrete cases may require probability $\geq\alpha$). For observed data $\vec{x}$, the confidence interval is then the set $\mathrm{I}(\vec{x})$ of all $(\mu, \vec{\phi})$ such that $\mathrm{R}(\mu, \vec{\phi})$ contains $\vec{x}$.

For a given $\vec{x}$, the interval $\mathrm{I}(\vec{x})$ will have some complicated shape in $(\mu, \vec{\phi})$ space. For any possible true value $(\mu, \vec{\phi})$, the distribution of resulting intervals $\mathrm{I}(\vec{x})$ has the property that the true value $(\mu, \vec{\phi}) \in \mathrm{I}(\vec{x})$ with probability α ($\geq\alpha$ for discrete cases). To produce an interval $[\mu_L, \mu_U]$ for μ alone (given $\vec{x}$), one takes the projection of $\mathrm{I}(\vec{x})$ onto the μ-axis; $\mu_L \leq \mu$ for any $(\mu, \vec{\phi}) \in \mathrm{I}(\vec{x})$, and $\mu_U \geq \mu$ for any $(\mu, \vec{\phi}) \in \mathrm{I}(\vec{x})$. Then for any possible true value of $(\mu, \vec{\phi})$, the distribution of resulting intervals $[\mu_L, \mu_U]$ will also have the property that $\mu_L \leq \mu \leq \mu_U$ with probability $\geq\alpha$.

Projection induces overcoverage, especially with many nuisance parameters. Thus, one would like to define $\mathrm{R}(\mu, \vec{\phi})$ so that the intervals $\mathrm{I}(\vec{x})$ are already wide in the $\vec{\phi}$ variables. The regions $\mathrm{R}(\mu, \vec{\phi})$ are often defined via an ordering rule

$$\mathrm{R}(\mu, \vec{\phi}) = \{\vec{x} | \rho(\vec{x}, \mu, \vec{\phi}) \leq \rho_0(\alpha, \mu, \vec{\phi})\}.$$

Because of computational difficulty, there are few fully frequentist examples (14–16, 22, 23). Coverage is guaranteed by construction, but one should check for other pathologies, some of which are listed in Reference 24.

4.1.6. Mixed frequentist-Bayesian. Here one applies the Neyman construction to obtain an interval from $L(\mu)$ of Equation 5. Because only the parameter of interest remains after marginalization, the Neyman construction is not numerically difficult at this stage. The method is therefore easier to implement than the fully frequentist approach, while still avoiding the necessity of choosing a prior for the parameter of interest. References 18 and 25 describe the mixed approach.

Despite the Bayesian methodology for the nuisance parameters, the overall method may still have frequentist coverage. However, neither frequentist coverage nor Bayesian credibility for the parameter of interest is guaranteed by the method. Philosophically, the approach can be regarded as a black box whose properties must be determined. Investigation of coverage is therefore in order (if coverage is desired).

4.2. *p*-Values

Section 3 discussed *p*-values in the absence of systematic uncertainties. We now mention several ways of incorporating them into *p*-values. For each method described below, we assume that an appropriate statistic *t*, which may depend on data from both the main and subsidiary measurements, has been selected for the task at hand, and the *p*-values are the upper-tail probability of the distribution of the statistic *t*. Without any systematic uncertainties, the *p*-value corresponding to an observed value of the statistic t_0 is then

$$P(t_0) = \int_{t_0}^{\infty} f(t)\, dt,$$

where $f(t)$ is the probability density of the distribution of the statistic *t*.

Systematic uncertainties give the distribution of *t* a dependence on nuisance parameters $\vec{\phi}$: $f(t|\vec{\phi})$. In some cases, especially for goodness of fit, a parameter of interest can be a nuisance parameter for the purposes of calculating a *p*-value. (For example, in assessing whether a mass spectrum shows evidence for the production of the Higgs boson, the physically interesting mass of the Higgs is merely a nuisance parameter.) It is desirable to construct a goodness-of-fit statistic whose distribution is almost independent of unknown parameters, but exact independence is usually not possible. In the binned goodness-of-fit example, the Poisson likelihood ratio λ, reviewed in Reference 26, is less dependent than Pearson's χ^2 on the true values of the expected bin contents, but not completely independent. The distribution of either statistic may then depend on the parameter of interest, which becomes a nuisance parameter for the associated *p*-value.

The following sections describe methods for incorporating uncertainties on nuisance parameters into a *p*-value; they are discussed more fully in Reference 27. Inherent is the assumption that a small *p*-value leads to the rejection of the hypothesis in question, whereas a large *p*-value only means that the hypothesis in question is not rejected—the normal approach in questions of significance. A conservative approach is then to consider the largest possible *p*-value. In some goodness-of-fit applications, for example, performing checks that data match an expected distribution, the emphasis is somewhat different: A reasonably large *p*-value is interpreted to mean that no serious problems are present and it is safe to proceed with the analysis. In such applications, a method that tends to overestimate the *p*-value is not conservative; it tends to hide problems. Here, one may prefer a range, or distribution, of *p*-values that are likely to be correct, as a large goodness-of-fit *p*-value is intended to lend

Table 1 Errors in using p-values for goodness of fit and for significance

	Errors of first kind	Errors of second kind
Definition	Reject true H_0	Accept false H_0
p-value	Small	Larger
Goodness of fit	Loss of efficiency	Source of background
Significance	False discovery	Failure to discover
Effect of conservatism (larger p-value)	Fewer errors	More errors

p-values are used to quantify the degree of consistency between data and a null hypothesis H_0. In goodness of fit, we look for reasonable p-values (say $p > 1\%$) in order to not reject H_0 and, for example, to accept the values of some fitted parameters or to select a sample of a given type of events. For significance, we hope for very small p-values (say 10^{-7}) in order to reject H_0 and perhaps claim a discovery. The usual effect of nuisance parameters is an increase in p-values, which thus reduces the chance of false discovery but makes it more likely that H_0 may be incorrectly accepted.

confidence that the data are well modeled. For simplicity, the methods are illustrated for the significance case. How to properly incorporate uncertainties on nuisance parameters into a goodness-of-fit p-value is less well established. **Table 1** summarizes some differences in how systematics associated with p-values influence significance and goodness of fit.

4.2.1. Prior predictive. This is the method most commonly used in high-energy physics. One marginalizes the *pdf* over the priors for the nuisance parameters before calculating the p-value:

$$P(t_0) = \int_{t_0}^{\infty} \int f(t|\vec{\phi})\,\pi(\vec{\phi})\,d\vec{\phi}\,dt.$$

As with parameter determination, the prior $\pi(\vec{\phi})$ may be the posterior of a subsidiary measurement. This method may not be appropriate for goodness-of-fit p-values when the distribution of t depends strongly on the parameter(s) of interest, as the prior for the parameter of interest is typically chosen to be noninformative.

4.2.2. Posterior predictive. In a fully Bayesian method, the main experiment often has some information about the nuisance parameters. One can obtain the posterior $p(\vec{\phi})$ for the nuisance parameters by marginalizing the joint posterior over the parameter of interest μ,

$$p(\vec{\phi}) = \frac{\int L(\mu, \vec{\phi})\,\pi(\mu, \vec{\phi})\,d\mu}{\int\int L(\mu, \vec{\phi})\,\pi(\mu, \vec{\phi})\,d\mu d\vec{\phi}}.$$

The posterior-predictive p-value is then calculated as

$$P(t_0) = \int_{t_0}^{\infty} \int f(t|\vec{\phi})p(\vec{\phi})\,d\vec{\phi}\,dt.$$

Here also it is unclear that this method is appropriate for goodness-of-fit p-values when the distribution of t depends strongly on the parameter(s) of interest. The fact that the data are used twice, once to help determine $p(\vec{\phi})$ and again in the definition of $P(t_0)$, may represent a logical defect of the posterior-predictive method.

4.2.3. The plug-in method. The plug-in p-value is calculated as

$$P(t_0) = \int_{t_0}^{\infty} f(t|\vec{\phi}_0)\,dt,$$

where $\vec{\phi}_0$ is some estimate of the nuisance parameters. In some cases, as when $\vec{\phi}_0$ is estimated under the null hypothesis, the plug-in p-value is conservative. For example, with a null hypothesis of no signal, the background estimate must include all bins in a histogram, including those normally considered signal, to qualify as conservative. This to some extent compensates for the fact that the plug-in method does not account for the uncertainties associated with $\vec{\phi}$. Reference 28 shows how using a plug-in p-value for goodness of fit can lead to false confidence.

4.2.4. The supremum method. The supremum p-value is the largest possible p-value obtainable from any allowed values of the nuisance parameters:

$$P(t_0) \;=\; \text{maximum with respect to } \vec{\phi} \text{ of } \int_{t_0}^{\infty} f(t|\vec{\phi})\,dt.$$

It is useful when the statistic t can be chosen so that $f(t|\vec{\phi})$ is at least approximately independent of $\vec{\phi}$. Otherwise the supremum p-value is too conservative, destroying any effect.

As no prior is used, the supremum p-value is a completely frequentist construction. For example, suppose that there are n signal region events and m events in a background-only region, where, for simplicity, equal background contributions are expected in each region. Here the background-only region represents a subsidiary measurement, yielding the expected background in the signal region (with an uncertainty). As the statistic for a significance calculation, a frequentist approach may use the likelihood ratio (for equal expected rates in the two regions, compared with completely free rates), defined as

$$\lambda(n,\; m) \;=\; \frac{\text{max w.r.t. } \mu \;=\; v \text{ of } \exp(-\mu)\mu^n/n! \; \exp(-v)v^m/m!}{\text{max w.r.t. } \mu,\; v \text{ of } \exp(-\mu)\mu^n/n! \; \exp(-v)v^m/m!} \;=\; \frac{\left(\frac{n+m}{2}\right)^{n+m}}{n^n m^m}.$$

It is convenient to define the significance statistic as $t(n, m) \;=\; -2\ln[\lambda(n, m)]$. The p-value

$$P(n_0, m_0|\nu) = \sum_{t(n,m)\geq t(n_0,m_0)} \frac{e^{-\nu}\nu^n}{n!}\,\frac{e^{-\nu}\nu^m}{m!}$$

is evaluated under the no-signal hypothesis and depends on the value of the nuisance parameter v (the background rate in each bin). In evaluating the supremum p-value, we maximize $P(n_0, m_0|\nu)$ over the range $\nu = 0$ to ∞. This can be done numerically, and $P(n_0, m_0|\nu)$ does not depend strongly on v, so the method can be considered reasonable. For other choices of statistic, the p-value may be strongly dependent on the assumed value of a nuisance parameter. In such cases, the supremum p-value is useless.

The value of the nuisance parameter at which the maximum p-value occurs is not considered an estimate of that parameter; it is common for the p-value to have its maximum at a value of the nuisance parameter that is 10σ or more from its best estimate.

For goodness of fit, it may be more useful to quote both the minimum and maximum p-value, not just the maximum. For example, a range of 0.2–0.8 would be a clear indication of an acceptable fit, whereas a range of 10^{-4}–0.8 would be ambiguous.

4.2.5. Supremum over confidence interval. In some cases it may be computationally difficult to maximize the p-value over an infinite range, or the p-value may be maximized at a value of a nuisance parameter excluded at very high confidence levels. A variation of the supremum method is to maximize the p-value over a confidence interval for $\vec{\phi}$. Suppose W is a region that contains the true value of $\vec{\phi}$ at confidence level $1 - \beta$. Then the p-value defined as

$$P(t_0) = \beta + \text{maximum with respect to } \vec{\phi} \in \text{W of} \int_{t_0}^{\infty} f(t|\vec{\phi})\, dt$$

has valid frequentist coverage properties (29). No prior is necessary, but obtaining a confidence interval for the nuisance parameters requires that the *pdf* for the subsidiary measurements be available.

As defined, this p-value can never be less than β. One should choose a β much smaller than the smallest p-value to which one wishes to be sensitive. For example, to retain the possibility of obtaining p-values of order 10^{-10}, $\beta = 10^{-12}$ would be a reasonable choice. It is not valid to select β based on what is observed in the data; β must be chosen on other grounds.

The method allows one to ignore values of nuisance parameters rejectable at high confidence levels, at the cost of introducing an artificial floor below which the reported p-value may never descend. However, it is useful mainly in the same situation that the supremum p-value is useful: when $f(t|\vec{\phi})$ is at least approximately independent of $\vec{\phi}$.

5. FULLY BAYESIAN CROSS-SECTION UPPER LIMITS

Heinrich et al. (24) describe a fully Bayesian approach to upper limits on cross sections in the presence of a single nuisance parameter. In that example, the main measurement observes n events from a Poisson distribution with mean $s\varepsilon + b$, where the cross section s is the parameter of interest; the acceptance ε is a nuisance parameter; and, to keep the example simple, the background b is taken to be a known constant.

A subsidiary measurement to determine ε is specified in which m events are observed from a Poisson process with mean $\kappa\varepsilon$, where κ is a known constant. In the subsidiary measurement, a prior for ε of the form ε^{q-1} yields a posterior for ε that becomes the prior:

$$\pi(\varepsilon) = \frac{(\kappa\varepsilon)^{\mu} e^{-\kappa\varepsilon}}{\varepsilon\Gamma(\mu)}$$

for ε in the main measurement, where $\mu = m + q$. This is a gamma distribution, with mean μ/κ and variance μ/κ^2.

A prior for the cross section s, which in Reference 24 is chosen to have the form s^{r-1}, is also necessary. The posterior for the cross section s, which in this simple case can be obtained analytically, is then

$$p(s) = \frac{\Gamma(\mu + n)}{\Gamma(\mu - r)\Gamma(r + n)} \frac{s^{r+n-1}\kappa^{\mu-r}}{(s + \kappa)^{\mu+n}} \frac{M(-n, 1 - n - \mu, b(s + \kappa)/s)}{M(-n, 1 - n - r, b)}.$$

The confluent hypergeometric function (30) $M(-n, a, x)$ with integer $n \geq 0$ is a polynomial of degree n in x. As $M(-n, 1 - n - r, b)$ is a polynomial in b whose highest-order term contains the factor b^n/r, with $b > 0$, the posterior approaches a δ function at $s = 0$ in the limit as $r \to 0$. In the presence of background, a $1/s$ prior has too much weight at $s = 0$.

Other popular choices of priors for s are $r = 1$ (a flat prior) and $r = 1/2$ (a $1/\sqrt{s}$ prior). A flat prior for s is conservative, as it results in mild overcoverage for upper limits on s. Coverage drops as r decreases. Consequently, for priors of the form s^{r-1} in this problem, optimal coverage properties for upper limits are achieved by an r somewhere in the range of $0 < r \leq 1$.

Reference 31 advocates the use of coverage as a diagnostic for the selection of priors—one approach within a larger objective Bayesian methodology—and seems a practical way to reduce the ambiguity associated with the choice of prior. Objective Bayesians attempt to minimize the influence of priors on the posterior, rather than derive them from personal belief, as in the subjective Bayesian approach.

When not derived directly from the posterior of a subsidiary measurement, it is common to assume a Gaussian form for the prior for ε (set to zero for $\varepsilon < 0$). When combined with a flat prior for s, this leads to a posterior density for s that is not normalizable, behaving like $1/s$ at large s. Some blame the Gaussian ε prior, others the flat prior for s, but certainly the combination of the two is pathological. (Even here, the flat prior is still conservative; upper limits become infinite.)

However, in this example it is only for s that flat is conservative. A flat subsidiary prior for ε is less conservative than a $1/\sqrt{\varepsilon}$ prior. An ε^{q-1} subsidiary prior yields larger upper limits for s as q decreases because underestimating the acceptance means overestimating the cross section. This reversal becomes important in the multichannel generalization of the problem; Reference 32 shows that a flat prior for ε can lead to significant undercoverage when there are more than two channels with Poisson subsidiary measurements. For N channels with acceptances ε_i, assigning a flat prior to each ε_i results in an effective prior for the total acceptance $\varepsilon = \sum \varepsilon_i$ proportional to ε^{N-1}. For large N and for small numbers of events in the subsidiary experiments, this can lead to significant undercoverage.

6. MISCELLANEOUS

In this section, we discuss several different issues of relevance to systematics. As with many statistical situations but especially with systematics, although textbook problems have unique solutions, in real life the issues are such that decisions often involving personal judgment must be made.

6.1. Blind Analyses

Corrections for potential systematic biases can be significant in magnitude and, as already mentioned, can involve personal choices. There is thus the opportunity/danger that a physicist can decide which corrections to apply in order, subconsciously, to adjust the corrections until a desired result is obtained. (Similar remarks apply to the choice of statistical technique.) This can be avoided if a blind analysis is employed. Various techniques for blind analyses are described in References 33–36.

6.2. Separating Signal from Background

Almost every analysis in particle physics involves the separation of the wanted signal from various sources of background. This is often achieved by using a classifier that has been trained on simulated samples of signal and backgrounds, which have been checked to describe the actual data with reasonable accuracy.

The traditional way of handling systematic effects is to train the classifier with simulated samples for which the nuisance parameters are fixed at their optimal values, and then to investigate how the classifier's performance is affected when these nuisance parameters are changed. Better performance of the classifier may be achieved if the generated training samples already contain the systematic uncertainties. That is, the nuisance parameters should be varied randomly event by event over their expected probability distributions (37, 38).

6.3. Simulation Issues

6.3.1. Reweighting. One method for estimating a contribution to systematics is to determine δa, the change in the answer, when the nuisance parameter v is changed by its uncertainty σ_ν (see Section 4.1.1). This usually involves generating Monte Carlo samples with v equal to its best value ν_0 and to $\nu_0 + \sigma_\nu$ (and maybe $\nu_0 - \sigma_\nu$ too).

If the Monte Carlo samples at ν_0 and at $\nu_0 + \sigma_\nu$ are generated independently, the uncertainty on δa due to the limited Monte Carlo statistics will be $\sqrt{2}\sigma_{MC}$, where σ_{MC} is the statistical error on each. If instead the sample at $\nu_0 + \sigma_\nu$ is obtained by a suitable reweighting of the events at ν_0, the error on the difference can be even smaller than σ_{MC} because of the correlations between the samples, provided the events' reweighting factors are not too different from each other. Thus, for example, a set of events generated from an exponential decay distribution with a lifetime $\tau = \tau_1$ can be converted to a distribution with lifetime $\tau = \tau_2$ by reweighting each event with decay time t by a factor of $\tau_1/\tau_2 \times \exp(-[1/\tau_2 - 1/\tau_1]t)$.

It is better to vary v in the direction such that the parts of the distribution with smaller numbers of events are reweighted downward rather than upward. For example, it is better to reweight a Gaussian to a narrower width rather than to a larger one.

6.3.2. Allowance for statistical errors of simulation. With a contribution to the systematic quantities estimated as $c_i \pm d_i$ (where d_i is the statistical error on c_i arising from the limited simulation statistics), the question is, what should be used for the actual contribution to our error estimate? Suggestions have included $c_i, c_i + d_i, \sqrt{c_i^2 + d_i^2}, \sqrt{c_i^2 - d_i^2}$, and c_i if c_i is larger than d_i but otherwise zero, and so on. We suggest an aggressive approach in which the contribution to the variance is taken as $c_i^2 - d_i^2$, even if this is negative. Assuming that the separate contributions are independent, the total variance for the systematic is then $\sum(c_i^2 - d_i^2)$, except that if this sum were negative, we would set it to zero. Our logic is that, if a potential systematic were in fact negligible, we would expect c_i to be Gaussian distributed about zero with variance d_i^2. If each contribution to the variance were taken as the larger of $c_i^2 - d_i^2$ and zero, we would overestimate our systematic. A problem this suggestion shares with most others is that, if one of the contributions to the systematics is poorly determined, then so is the total systematic.

6.3.3. Unisim or multisim? The contribution from a possible systematic can be estimated by seeing the change in the answer a when the nuisance parameter is varied by its uncertainty. The various contributions from each systematic are calculated separately and then combined (see Section 4.1.1). It essentially makes use of the first-order Taylor expansion

$$a = a_0 + \sum \frac{\partial a}{\partial v_i} \delta v_i. \qquad 6.$$

The simulations to estimate the derivatives are termed unisims by the MiniBooNE Collaboration.

An alternative, even for just one source of systematic, is to generate separate simulated samples, in each of which the values for each systematic have been chosen randomly from their expected distributions and then an answer a is extracted for each sample. The width of the distribution in a values then is used to estimate the total effect of all the systematics. The advantage is that it can handle non-Gaussian distributions and also nonlinearities in the dependence of a on the nuisance parameters. In some cases there can also be savings in the total amount of simulation needed by this multisim approach, to achieve the same accuracy as for the corresponding unisim. Roe (39) has compared the relative accuracies of the two methods.

6.4. Theory Uncertainties

In some cases, the extraction of a physical result requires some theoretical formulation, and there can be more than one variant of how this is implemented. (For example, several different sets of parton distribution functions exist.) The question arises of how the different results contribute to the systematic error. Again, this is a situation without a unique answer.

If there are only two possible versions of the theory, it is probably best to quote separately the results for each. This deals most simply with the situation in which one version is subsequently ruled out by other data. However, if there are multiple choices, quoting the result for each may not be practical. It can also depend on whether the different theoretical variants are regarded as equally plausible. If so, it is common to quote the result as the average, and in the case of two possibilities, half the difference is usually taken as the error. If the different results are regarded as samples of possible results (which from a frequentist viewpoint is hard to maintain), it may not be unreasonable to calculate their standard deviation; for two values, this would give 0.7 times their difference.

A similar situation arises when different functional forms are used to parametrize a distribution. For example, a peak may have more events in its tails than is appropriate for the Gaussian distribution used to fit the data. This is simpler to deal with if the alternative can be formulated in such a way that the original distribution is nested within it. Thus, a sum of Gaussians or a Student's *t* distribution with *N* degrees of freedom may be used to describe the heavier tails. Such considerations can be especially important for nuisance parameters, where the approximation of (truncated) Gaussian distributions may be overly optimistic, especially in the tails. This can cause problems when high significances (e.g., larger than 5σ) are demanded and have been calculated using an incorrect probability density distribution.

6.5. Fits with Correlated Systematics

In comparing a theoretical distribution with a data histogram where systematic effects are involved, the relevant χ^2 expression is

$$S = \sum \left(y_i^{\text{data}} - y_i^{\text{pred}}\right) H_{ij} \left(y_j^{\text{data}} - y_j^{\text{pred}}\right), \tag{7.}$$

where H_{ij} is the inverse error matrix on the data y_i^{data}, and the correlations are likely due to the systematics. Demortier (40) showed that this is equivalent to minimizing

$$S' = \sum \frac{\left(y_i^{\text{data}} - z_i^{\text{pred}}(c_1, c_2, \ldots)\right)^2}{(\sigma_{\text{stat}})_i^2} + \sum c_s^2, \tag{8.}$$

where $z_i^{\text{pred}}(c_1, c_2, \ldots)$ is the sum of the predictions y_i^{pred} in the absence of systematics plus the possible systematic effects, and c_s are the coefficients of the different systematics. Each systematic is assumed to have a defined magnitude in each of the n bins of the experimental distribution. With enough data, more than n different coefficients c_i for the magnitudes of the systematic sources can be extracted from the data. For low statistics, Equation 8 can be modified by using the likelihood based on the Poisson distribution of the small number of events in each bin of a histogram.

6.6. Asymmetric Errors

Sometimes separate values are quoted for positive and for negative errors, for example, $1.0^{+0.4}_{-0.2}$ ps for a lifetime. These can arise from statistical errors where the likelihood function has a non-Gaussian shape, or for systematic errors when there are asymmetric changes as a nuisance parameter is changed upward and downward by one error from its central value.

When various contributions with uncorrelated but asymmetric errors are combined, a popular method is to combine the upper errors in quadrature, and similarly for the lower errors, but there is no basis for this. The quadrature rule is applicable when the separate errors can be in the same direction or can tend to cancel. But upper errors are always in the same sense, so the justification disappears. Barlow (21) has discussed ways of dealing with asymmetric errors for calculating χ^2 or for combining different measurements of the same quantity.

6.7. The Best Linear Unbiased Estimate

A well-known technique for combining several measurements $a_i \pm \sigma_i$ of a single quantity a is to use the best linear unbiased estimate (41). This is

$$\hat{a} = \sum \beta_i a_i, \tag{9.}$$

where the coefficients satisfy $\sum \beta_i = 1$ and are chosen such that the error on $\hat{a}$ is minimized. For uncorrelated errors, this is equivalent to minimizing $\sum (a_i - \hat{a})^2/\sigma_i^2$. When each error is expressed as $\pm(\sigma_{\text{stat}})_i \pm(\sigma_{\text{syst}})_i$, the method can still be used, with the possibility of including any correlations between the various errors. The method is not recommended for highly correlated results, where $\hat{a}$ can be outside the range of the individual a_i, and its value depends sensitively on the exact values of the correlation coefficients. An advantage of extracting the coefficients β_i is that this enables the separate contributions of the statistical and systematic errors to $\hat{a}$ to be calculated.

7. CONCLUSIONS

Various methods for incorporating systematic effects in parameter estimation have been discussed. These range from the fully Bayesian approach to a complete frequentist Neyman construction. There are also many other recipes that are neither fully Bayesian nor fully frequentist, for example, profile likelihood, mixed Bayes-frequentist, and so on. Although many physicists would regard the fully frequentist method as desirable (in that it avoids the necessity to choose priors and is guaranteed not to undercover), practicalities generally require the use of some other technique. The Bayesian calculation of upper limits on a Poisson production process was discussed in Section 5. It resulted in a usable algorithm, with reasonable coverage properties, provided priors were appropriately chosen.

In searching for new physics, possible discrepancies between the data and the null hypothesis (currently the Standard Model for particle physics) are expressed as *p*-values. Possible ways of incorporating nuisance parameters were listed in Section 4.2. Even though it is possible to specify the properties of various procedures for incorporating nuisance parameters in calculations of intervals or *p*-values, by their very nature systematic errors are such that their evaluation will often depend on the experience and judgment of the experimentalist.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We wish to thank our experimental colleagues, too numerous to list individually, for many stimulating and vigorous discussions over the years, and especially the CDF Statistics Committee, who also provided valuable feedback on this review. We are also grateful to those statisticians who have participated in the PHYSTAT series of conferences and had the patience to explain relevant aspects of statistics to us. The Leverhulme Foundation kindly provided L.L. with a grant, which partially supported this work.

LITERATURE CITED

1. Barlow R. See Ref. 9, p. 134 (2002)
2. Sinervo P. See Ref. 10, p. 122 (2003)
3. Cousins RD. See Ref. 11, p. 75 (2006)
4. Punzi G. CDF Note 7975. **http://www-cdf.fnal.gov/physics/statistics/notes/punzi-systdef.ps** (2001)
5. Blocker C. CDF Note 6506. **http://www-cdf.fnal.gov/publications/cdf6506_systematics.ps** (2001)
6. Barlow R, et al. (BaBar Stat. Work. Group) *Chapter* 7. **http://www.slac.stanford.edu/BFROOT/www/Statistics/** (2002)
7. James F, Lyons L, Perrin Y, eds. *CERN Yellow Report*. CERN 2000–005. Geneva, Switz.: CERN (2000)
8. Lyons L, et al. *Proc. Fermilab Confid. Limits Workshop*, Batavia, Illinois, 2000. **http://conferences.fnal.gov/cl2k/** (2000)
9. Whalley MR, Lyons L, eds. *Proc. Conf. Adv. Stat. Tech. Part. Phys.* IPPP/02/39. Durham, UK: Univ. Durham. 333 pp. (2002)
10. Lyons L, Mount R, Reitmeyer R, eds. *Proc. PHYSTAT 2003 Stat. Probl. Part. Phys. Astrophys. Cosmol.* SLAC-R-703. Menlo Park, CA: Stanford Linear Accel. Cent. 334 pp. (2003)
11. Lyons L, Karagoz Unel M, eds. *Proc. PHYSTAT05 Stat. Probl. Part. Phys. Astrophys. Cosmol.* London: Imp. Coll. Press. 310 pp. (2006)
12. Linnemann J, Lyons L, Reid N. *BIRS PHYSTAT Workshop Stat. Inference Probl. High Energy Phys. Astron.*, Banff, 2006. **http://www.pims.math.ca/birs/birspages.php?task=displayevent&event_id=06w5054** (2006)

13. Feldman GJ, Cousins RD. *Phys. Rev. D* 57:3873 (1998)
14. Cousins RD. *Nucl. Instrum. Methods A* 417:391 (1998)
15. Punzi G. See Ref. 11, p. 88 (2006)
16. Demortier L. physics/0312100 v2 (2003)
17. Reid N. See Ref. 10, p. 265 (2003)
18. Cousins RD, Highland VL. *Nucl. Instrum. Methods A* 320:331 (1992)
19. James F. Function minimization and error analysis reference manual. *CERN Program Libr. Long Writeup D506*. Geneva, Switz.: CERN (1998)
20. Rolke WA, Lopez AM, Conrad J. See Ref. 11, p. 97 (2006)
21. Barlow R. See Ref. 10, p. 250 (2003)
22. Nicolo D, Signorelli G. See Ref. 9, p. 152 (2002)
23. Sen B, Walker M, Woodroofe M. *Statistica Sin*. 17:In press
24. Heinrich J, et al. physics/0409129 (2004)
25. Conrad J, Tegenfeldt F. See Ref. 11, p. 93 (2006)
26. Baker S, Cousins R. *Nucl. Instrum. Methods A* 221:437 (1984)
27. Demortier L. CDF Note 8662. **http://www-cdf.fnal.gov/publications/cdf8662_p_values_in_detail.pdf** (2007); Cranmer K. See Ref. 11, p. 112 (2006); Linnemann J. See Ref. 10, p. 35 (2003)
28. Heinrich J. See Ref. 10, p. 52 (2003)
29. Berger RL, Boos DD. *J. Am. Stat. Assoc.* 89:1012 (1994)
30. Slater LJ. In *Handbook of Mathematical Functions*, ed. M Abramowitz, IA Stegun, p. 503. New York: Dover (1968)
31. Bayarri MJ, Berger JO. *Stat. Sci.* 19:58 (2004)
32. Heinrich J. See Ref. 11, p. 98 (2006)
33. Harrison P. See Ref. 9, p. 278 (2002)
34. Roodman A. See Ref. 10, p. 166 (2003)
35. Heinrich J. CDF Note 6576. **http://www-cdf.fnal.gov/publications/cdf6576_blind.pdf** (2003)
36. Lyons L. *Physics perspective*. Presented at Stat. Chall. Mod. Astron., 4th, State College, 2006. CDF Note 8514. **http://www-cdf.fnal.gov/publications/cdf8514_HEP_Astro_Stats.ps** (2006)
37. Lyons L. *Nucl. Instrum. Methods A* 324:565 (1993)
38. Neal R. BIRS PHYSTAT Workshop Stat. Inference Probl. High Energy Phys. Astron., Banff, 2006. **http://www.pims.math.ca/birs/birspages.php?task=displayevent&event_id=06w5054** (2006)
39. Roe B. *Nucl. Instrum. Methods A* 570:159 (2007)
40. Demortier L. CDF Note 8661. **http://www.cdf.fnal.gov/publications/cdf8661_chi2fit_w_corr_syst.pdf** (1999)
41. Lyons L, Gibaut D, Clifford P. *Nucl. Instrum. Methods A* 270:110 (1988)