DEALING WITH DATA: SIGNALS, BACKGROUNDS, AND STATISTICS

L. DEMORTIER

Laboratory of Experimental High Energy Physics, Rockefeller University, New York, NY 10065, U.S.A.
E-mail: luc.demortier@rockefeller.edu
www.rockefeller.edu

We review the basic statistical tools used by experimental high energy physicists to analyze, interpret, and present data. After an introduction on the meaning of probability, we describe approaches to hypothesis testing, interval estimation, and search procedures.

Keywords: Bayes; Frequentism; Hypothesis testing; Interval estimation; Search procedures

1. Introduction

The primary goal of these lectures is to review the basic statistical concepts needed to understand and interpret experimental results in the high energy physics literature. These results are typically formulated in terms of point estimates, intervals, p values, likelihood functions, Bayes factors, and/or posterior probabilities. Knowing the meaning of these quantities, their limitations, and the rigorous methods needed to extract them from data, will help in evaluating the reliability of published results. A secondary goal is to provide some tools to phenomenologists who would like to estimate the sensitivity of a particular experiment to a model of new physics.

These goals are facilitated by the availability of many web resources. For example, several experimental collaborations have formed internal statistics committees whose purpose is to make recommendations on proper statistical methods, to act as consultants on specific data analyses, and to help with the comparison and combination of experimental results from different experiments; some of these committees have public web pages with plenty of useful information.[1–3] In addition, high energy physicists and astrophysicists regularly meet with professional statisticians to discuss problems and

methods. These so-called PhyStat meetings have their own webpages and proceedings.[4–9] Finally, there is a repository of statistics software and other resources at `http://phystat.org`, and professional statistics literature is available online through `http://www.jstor.org`.

We begin our review with a discussion of the frequentist and Bayesian concepts of probability in section 2. This is followed by sections on hypothesis testing and interval estimation. Section 5 combines these two methodologies in the design of search procedures, which are at the heart of everyone's hopes for the success of the LHC program. Finally, section 6 contains some remarks about systematic uncertainties.

2. What Is Probability?

There is a long-standing philosophical dispute on the appropriate definition of probability, between two contenders known as frequentism and Bayesianism. This dispute has interesting implications both for the interpretation of scientific measurements and for the determination of quantum states.

2.1. *Frequentism*

Frequentists attempt to define probabilities as relative frequencies in sequences of trials. This corresponds to the common-sense intuition that if, for example, we toss a coin a large number of times, we can use the fraction of times it falls heads up as an estimate of the probability of "heads up", and this estimate becomes more accurate as the total number of tosses increases. To physicists this is a very attractive aspect of the frequentist definition: probabilities are postulated to be real, objective quantities that exist "outside us" and can be measured just as the length of a table or the weight of a book. Unfortunately it is very difficult to formulate a rigorous, non-circular definition of probability in terms of sequences of trials.[10] One possibility is to define probability as the limiting relative frequency in an infinite sequence of trials, or as the limiting relative frequency which would be obtained if the sequence of trials were extended to infinity. However, unlike a table or a book, infinite sequences are unobservable to finite beings like us. Furthermore, they may not even be empirically relevant. If at some point very far into an infinite sequence, the probability of interest suddenly changes by a discrete amount, this will affect the "infinite-sequence" value of the probability, but why should we care if we do not get to live until that point? Thus from a practical point of view it would seem more sensible to define the probability of an event as the relative frequency of that event in

a *sufficiently long* sequence of trials. This is clearly a much weaker definition though. Indeed, given a finite number of trials, *every* sequence has a non-zero probability of occurring, and therefore also every probability value allowed by the discreteness of the measurement. The only way to resolve the difficulties in the frequentist definition of probability is to assume that the trials in the defining sequence are independent and equally probable. Hence the circularity: we need the concept of equal probability in order to be able to define probability.

Setting aside these foundational problems, the frequentist definition of probability seriously constrains the type of inferences that can be made. Indeed, according to frequentism, a random variable is a physical quantity that fluctuates from one observation to the next. Hence it is not possible to assign a meaningful probability value to a statement such as "the true mass M_H of the Higgs boson is between 150 and 160 GeV/c^2", since M_H is a fixed constant of nature. Frequentism therefore needs an additional, separate concept to describe the reliability of inferences: this is the concept of confidence. As applied to interval estimates of M_H, confidence represents the probability that the measurement *procedure* will yield an interval that contains the true value of M_H if the experiment is repeated a large number of times; it does *not* represent the probability that the numerical interval actually obtained from the data at hand contains that true value. Thus, even though confidence is defined in terms of probability, it should not be confused with the latter since it is applied to statements to which a (non-trivial) frequentist probability value cannot be assigned.

The objective of frequentist statistics is then to transform measurable probabilities of observations into confidence statements about physics parameters, models, and hypotheses. This transformation is not unique however. In the great variety of measurement situations, frequentism offers many "ad hoc" rules and procedures. In contrast with Bayesianism, to be described next, there is no unique frequentist principle that guides the process of drawing inferences.

2.2. *Bayesianism*

Bayesianism makes a strict distinction between propositions and probabilities.[11] Propositions include statements such as "the Higgs mass is between 150 and 160 GeV/c^2", and "it will rain tomorrow". These are either true or false. On the other hand, Bayesian probabilities are degrees of belief about the truth of some proposition. They are themselves not propositions and are therefore neither true nor false. In contrast with frequentist probability,

which claims to be a measurable physical reality, Bayesian probability is a logical construct.

It can be shown that *coherent* degrees of belief satisy the usual rules of probability theory. The Bayesian paradigm is therefore entirely based on the latter, viewed as a form of extended logic:[12] a process of reasoning by which one extracts uncertain conclusions from limited information. This process is guided by Bayes' theorem, which prescribes how degrees of belief about a parameter $\theta \in \Theta$ are to be updated when new data x become available:

$$\pi(\theta \,|\, x) \;=\; \frac{p(x \,|\, \theta)\; \pi(\theta)}{m_{prior}(x)}. \tag{1}$$

On the left-hand side, the quantity $\pi(\theta \,|\, x)$ represents the posterior probability density of θ, after having observed data value x. It is expressed as a function of the prior probability density $\pi(\theta)$ and the likelihood function $p(x \,|\, \theta)$, which is the probability density of the data x for a given value of θ, viewed as a function of θ; to emphasize this view, the likelihood is sometimes written as $\mathcal{L}(\theta \,|\, x)$. Finally, the denominator $m_{prior}(x)$ is the marginal distribution of x, also called prior-predictive distribution, or evidence, depending on the context:

$$m_{prior}(x) \;\equiv\; \int_\Theta p(x \,|\, \theta)\; \pi(\theta)\; d\theta. \tag{2}$$

All the basic tools of Bayesian statistics are direct applications of probability theory. A typical example is marginalization. Suppose we have a model for some data that depends on two parameters, θ and λ, but that we are only interested in θ. The posterior density of θ can then be obtained from the joint posterior of θ and λ by integration:

$$\pi(\theta \,|\, x) \;=\; \int_\Lambda \pi(\theta, \lambda \,|\, x)\; d\lambda. \tag{3}$$

Another useful example involves prediction. Suppose we observe data x and wish to predict the distribution of future data y. This can be done via the posterior-predictive distribution:

$$m_{post}(y \,|\, x) \;=\; \int_\Theta p(y \,|\, \theta)\; \pi(\theta \,|\, x)\; d\theta. \tag{4}$$

We emphasize that the output of a Bayesian analysis is always the *full* posterior distribution (1). The latter can be summarized in various ways, by providing point estimates, interval estimates, hypothesis probabilities, predictions for new data, etc., but the summary should not be substituted for the "whole story".

2.2.1. *Bayesian Priors: Evidence-Based Constructions*

The elicitation of prior probabilities on an unknown parameter or incompletely specified model is often difficult work, especially if the parameter or model is multidimensional and prior correlations are present. In particle physics we can usually construct so-called evidence-based priors for parameters such as the position of a detector element, an energy scale, a tracking efficiency, or a background level. Such priors are derived from subsidiary data measurements, Monte Carlo studies, and theoretical beliefs.

If for example the position of a detector is measured to be $x_0 \pm \Delta x$, and Δx is accurately known, it will be sensible to make the corresponding prior a Gaussian distribution with mean x_0 and standard deviation Δx. On the other hand, for an energy scale, which is usually a positive quantity, it will be more natural to use a gamma distribution, and for an efficiency bounded between 0 and 1 a beta distribution should be appropriate. In each of these cases, other functional forms should be tried to assess the robustness of the final analysis result to changes in prior shape. Note that evidence-based priors are always proper, that is, they integrate to 1.

2.2.2. *Bayesian Priors: Formal Constructions*

In physics data analysis we often need to extract information about a parameter θ about which very little is known a priori, or perhaps we would like to pretend that very little is known for reasons of objectivity. How do we apply Bayes' theorem in this case? How do we construct the prior $\pi(\theta)$?

Historically, this problem is the main reason for the development of alternative statistical paradigms: frequentism, likelihoodism, fiducial probability, and others. Even Bayesianism has come up with its own solution, known as objective Bayes. In general, results from these different methods tend to agree on large data samples, but not necessarily on small samples (discovery situations). For this reason, statistics committees in various experiments recommend data analysts to cross-check their results using alternative methods.

At its most optimistic, objective Bayesianism tries to find a completely coherent, objective Bayesian methodology for "letting the data speak for themselves". A much more modest goal is to provide a collection of useful methods to learn from the data as part of a robustness study. There are in fact several approaches to objective Bayesianism, all of which attempt to construct prior distributions that are minimally informative in some sense. Some approaches make use of concepts from information theory, others

exploit the group invariance properties of some problems, and still others try to produce posterior distributions for which Bayesian credibilities can be matched with frequentist confidence statements. Bayesian analyses in high energy physics tend to err on the side of simplicity by using flat priors for parameters about which nothing is known a priori. The naive justification for flat priors is that they give the same weight to all parameter values and therefore represent ignorance. However, flat priors are not invariant under parameter transformations and they sometimes lead to improper posterior distributions and other kinds of problems.

Objective priors are also known as neutral, formal, or conventional priors. Although they are often improper when the parameter space is unbounded, they must lead to proper posteriors in order to make sense. A very important example of objective Bayesian prior is due to Harold Jeffreys. Suppose the data X have a distribution $p(x \,|\, \theta)$ that depends on a single continuous parameter θ; Jeffreys' prior is then:

$$\pi_J(\theta) \;\equiv\; \left\{ -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ln p(x \,|\, \theta) \right] \right\}^{1/2}, \tag{5}$$

where the expectation is with respect to the data distribution $p(x \,|\, \theta)$. This prior illustrates how formal priors depend on the model assumed for the data; however, they do not depend on the data themselves. When θ is multidimensional, Jeffreys' prior tends to misbehave and must be replaced by the more general reference analysis prescription.[13]

2.3. *Quantum Probabilities*

An argument that is sometimes made is that frequentism must be the correct approach to data analysis because quantum mechanical probabilities are frequentist.[14] This argument is specious however, because the process by which we learn from our observations is logically distinct from the process that generates these observations. Furthermore, advances in quantum information science have shown that it is possible to interpret quantum mechanical probabilities as states of knowledge, i.e. as Bayesian.[15]

Part of the motivation for research into whether quantum probabilities are frequentist or Bayesian comes from EPR-style arguments. Suppose two systems A and B are prepared in some entangled quantum state and then spatially separated. By measuring one of two observables on A alone, one can immediately write down a new state for B. If one accepts that the "real, objective state of affairs" at B cannot depend on measurements made at

A, then the simplest interpretation of the new state for B is that it is a *state of knowledge*.

It is possible to develop this idea of quantum states as states of knowledge in a fully consistent way. There are many aspects to this:[15]

- Subjective probability assignments must follow the standard quantum rule for probabilities (Gleason's theorem).
- The connection between quantum probability and long-term frequency still holds, but is a non-trivial consequence of Gleason's theorem and the concept of maximal information in quantum theory.
- Even quantum certainty (probability-1 predictions for pure states) is always some agent's certainty. Any agent-independent certainty about a measurement outcome would correspond to a pre-existing system property and would be in conflict with locality.[11]

Aside from providing yet another interpretation of quantum mechanics, do Bayesian quantum probabilities have any practical consequence? This is very much an open question. It may be for example, that vacuum fluctuations represent a Bayesian uncertainty rather than a real, physical phenomenon. If so, we do not need to worry about their contribution to the cosmological constant. Arguments for the physical reality of vacuum fluctuations are usually based on the experimental observations of spontaneous emission, the Lamb shift, and the Casimir effect. However E.T. Jaynes showed that spontaneous emission and the Lamb shift can both be derived without the need for vacuum fluctuations,[16] and R. L. Jaffe proved this for the Casimir effect.[17]

2.4. *Data Analysis: Frequentist or Bayesian?*

With some reasonable care, frequentist and Bayesian inferences generally agree in large samples. Disagreements tend to appear in small samples, where prior assumptions play a more important role both for frequentists and Bayesians. For a small number of problems, the Bayesian and frequentist answers agree exactly, even in small samples.

An often fruitful approach is to start with a Bayesian method, and then verify if the solution has any attractive frequentist properties. For example, if a Bayesian interval is calculated, does the interval contain the true value of the parameter of interest sufficiently often when the measurement is repeated? This approach has been formally studied by professional statisticians and is quite valuable.

On the other hand, if one starts with a purely frequentist method, it is also important to check its Bayesian properties for a reasonable choice of prior.

In experimental HEP we often use a hybrid method: a frequentist method to handle the randomness of the primary observation, combined with Bayesian techniques to handle uncertainties in auxiliary parameters. This is not easy to justify from a foundational point of view, but if the auxiliary parameter uncertainties are small, the overall measurement result may exhibit acceptable frequentist coverage.

3. Testing a Hypothesis

Hypothesis testing in high energy physics comes up in two very different contexts. The first one is when we wish to decide between two hypotheses, in such a way that if we repeat the same testing procedure many times, the rate of wrong decisions will be fully controlled in the long run. For example, when selecting good electron candidates for a measurement of the mass of the W boson, we need to minimize background contamination and signal inefficiency. The second context is when we wish to characterize the evidence provided by the data against a given hypothesis. In searching for new phenomena for example, we need to establish that an observed enhancement of a given background spectrum is evidence against the background-only hypothesis, and we need to quantify that evidence.

Traditionally, the first problem is solved by Neyman-Pearson theory and the second one by the use of p values, likelihood ratios, or Bayes factors.

3.1. *The Neyman-Pearson Theory of Testing*

Suppose we wish to decide which of two hypotheses, H_0 (the "null") or H_1 (the "alternative"), is more likely to be true given some observation X. The frequentist strategy is to minimize the probability of making the wrong decision over the long run. However, that probability depends on which hypothesis is actually true. There are therefore two types of error that can be committed:

- Type-I error: Rejecting H_0 when H_0 is true;
- Type-II error: Accepting H_0 when H_1 is true.

To fix ideas, suppose that the hypotheses have the form:

$$H_0 : \; X \sim f_0(x) \quad \text{versus} \quad H_1 : \; X \sim f_1(x), \tag{6}$$

by which one means that the observation X has probability density $f_0(x)$ under H_0 and $f_1(x)$ under H_1. For the test to be meaningful, f_0 and f_1 must be distinguishable given the measurement resolution. In other words, there must be a region C in sample space (the space of all possible data X) where the observation is much more likely to fall if H_1 is true than if H_0 is true. This region is called the critical region of the test and is used as follows: if the observation X falls inside C, we decide to reject H_0, otherwise we decide to accept it. The Type-I error probability α and the Type-II error probability β are then given by:

$$\alpha \;=\; \int_C f_0(x)\,dx \quad \text{and} \quad \beta \;=\; 1 - \int_C f_1(x)\,dx. \tag{7}$$

The probability of correctly accepting the alternative hypothesis equals $1-\beta$ and is known as the power of the test.

In general the critical region C is constructed so as to achieve a suitably small Type-I error rate α, but there are many possible critical regions that will yield the same α. The idea of the Neyman-Pearson theory is to choose the C that minimizes β for a given α. In the above example, the distributions f_0 and f_1 are fully specified before the test (this is known as "simple vs. simple testing"). In this case it can be shown that, in order to minimize β for a given α, C must be of the form:

$$C \;=\; \{x:\; f_0(x)/f_1(x) \;<\; c_\alpha\}, \tag{8}$$

where c_α is a constant depending on α. This result is known as the Neyman-Pearson lemma, and the quantity $y \equiv f_0(x)/f_1(x)$ is known as the likelihood ratio statistic.

Unfortunately, f_0 and/or f_1 are often composite, meaning that they depend on an unknown, possibly multidimensional parameter $\theta \in \Theta$. This happens when the measurement is affected by systematic uncertainties (in which case θ or one of its components could be an imperfectly known detector energy scale or tracking efficiency) or when the alternative hypothesis does not fully specify the value of a parameter of interest (as when θ or one of its components represents the production cross section for a new physics process and one is testing whether that cross section is exactly zero or strictly positive). The likelihood ratio is then defined as:

$$\lambda \;\equiv\; \frac{\sup\limits_{\theta\in\Theta_0} f_0(x_{obs}\,|\,\theta)}{\sup\limits_{\theta\in\Theta} f_1(x_{obs}\,|\,\theta)}, \tag{9}$$

where $\Theta_0 \subset \Theta$ is the subspace of θ values allowed by the null hypothesis. Although the Neyman-Pearson lemma does not generalize to this composite

situation, the likelihood ratio remains an extremely useful test statistic. This is partly due to Wilks' theorem, which states that for large samples the distribution of $-2\ln\lambda$ under H_0 is that of a chisquared variate with number of degrees of freedom equal to the difference between the dimensionality of Θ and that of Θ_0. Under some rather general conditions, this theorem can be used to construct approximate critical regions for finite samples (however, see section 3.4).

As already stated, the Neyman-Pearson theory of testing is most useful in data quality control applications, when a given test has to be repeated on a large sample of identical items. In HEP we use this technique to select events of a given type. For example, if we want to select a sample of events to measure the mass of the top quark, we define H_0 to be the hypothesis that a given event contains a top quark, and try to minimize the background contamination β for a given signal efficiency $1-\alpha$.

On the other hand, this approach to testing is not very satisfactory when dealing with one-time testing situations, for example when testing a hypothesis about a new phenomenon such as the Higgs boson or SUSY. This is because the result of a Neyman-Pearson test is either "accept H_0" or "reject H_0", without consideration for the strength of evidence contained in the data. In fact, the level of confidence in the decision resulting from the test is already known *before* the test: it is either $1-\alpha$ or $1-\beta$. One would like a way to quantify evidence from observed data, *after* the test. The frequentist solution to this problem uses p values exclusively, whereas the Bayesian one works with p values, Bayes factors and posterior hypothesis probabilities.

3.2. *The p Value Method for Quantifying Evidence*

Suppose we collect some data $\mathbf{X}$ and wish to characterize the evidence contained in $\mathbf{X}$ against a hypothesis H_0 about the distribution $f(\mathbf{x}\,|\,\theta)$ of the population from which $\mathbf{X}$ was drawn. A general approach is to construct a test statistic $T(\mathbf{X})$ such that large observed values of T are evidence against H_0 in the direction of some alternative of interest H_1. Often a good choice for T is $1/\lambda$, where λ is the likelihood ratio statistic defined in eq. (9). In general, different testing problems require different test statistics, and the observed values of these test statistics cannot be directly compared across problems. We therefore need a method for *calibrating* the evidence provided by T. One way to do this is to calculate the probability for observing $T = t_{\mathrm{obs}}$ *or a larger value* under H_0; this tail probability is known as the p value

of the test:

$$p \;=\; \mathbb{P}(T \geq t_{\rm obs} \,|\, H_0). \tag{10}$$

Thus, small p values are evidence against H_0. Typically one will reject H_0 if $p \leq \alpha$, where α is some predefined, small error rate. This α has essentially the same interpretation as in the Neyman-Pearson theory of testing, but the emphasis here is radically different: with p values we wish to characterize *post-data* evidence, a concept which plays no role whatsoever in the Neyman-Pearson theory. Indeed, the only output of the latter is a report of acceptance or rejection of H_0, together with *pre-data* expectations of long-run error rates.

Clearly, the usefulness of p values for *calibrating* evidence against a null hypothesis H_0 depends on their null distribution being known to the experimenter and being the same in all problems considered. In principle, the very definition (10) of a p value guarantees that its distribution under H_0 is uniform. In practice however, this guarantee is rarely fulfilled exactly, either because the test statistic is discrete or because of the presence of nuisance parameters. The following terminology then characterizes the true null distribution of p values:

$$\begin{array}{ll} p \text{ exact} & \Leftrightarrow \mathbb{P}(p \leq \alpha \,|\, H_0) = \alpha, \\ p \text{ conservative} & \Leftrightarrow \mathbb{P}(p \leq \alpha \,|\, H_0) < \alpha, \\ p \text{ liberal} & \Leftrightarrow \mathbb{P}(p \leq \alpha \,|\, H_0) > \alpha. \end{array}$$

Compared to an exact p value, a conservative p value tends to understate the evidence against H_0, whereas a liberal p value tends to overstate it.

In spite of the apparent simplicity of the motivation and definition of p values, their correct interpretation in terms of evidence is notoriously subtle. In fact, p values themselves are controversial. Here is partial list of caveats:

(1) P values are neither frequentist error rates nor confidence levels.
(2) P values are not hypothesis probabilities.
(3) Equal p values do not necessarily represent equal amounts of evidence (for example, sample size also plays a role).

Because of these and other caveats, it is better to treat p values as nothing more than useful "exploratory tools," or "measures of surprise." In any search for new physics, a small p value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better

explanation of the observations than the null hypothesis, can one make a convincing case for discovery.

3.2.1. *The 5σ Discovery Threshold*

A small p value has little intuitive appeal, so it is conventional to map it into the number N_σ of standard deviations a normal variate is from zero when the probability outside $\pm N_\sigma$ equals $k \cdot p$, where $k = 1$ or 2:

$$p \;=\; \frac{2}{k} \int_{N_\sigma}^{+\infty} \frac{e^{-x^2/2}}{\sqrt{2\,\pi}}\, dx \;=\; \frac{1}{k} \left[1 \;-\; \mathrm{erf}(N_\sigma/\sqrt{2})\right]. \qquad (11)$$

Experiments at the LHC set $k = 2$. This choice is not universal however.

The threshold for discovery is typically set at $N_\sigma = 5$. This convention can be traced back to a 1968 paper by A. Rosenfeld,[18] where the author argued that, given the number of histograms examined by high energy physicists every year, one should expect several 4σ claims per year. He therefore recommended that experimental groups publish any tantalizing effect that passes the 3σ threshold, as a recompense for the time and funds invested in their experiments, but that they take additional data in the amount needed to confirm a real effect at the 5σ level. As for theorists, they should always wait for 5σ (or nearly 5σ) effects.

Rosenfeld's argument was based on what is known as the look-elsewhere effect, and according to which the probability of a significant background fluctuation scales with the number of places one looks in. This is a 40-year old calculation however, and it is legitimate to ask whether the discovery threshold should be adjusted for the increase in the number and scope of searches for new physics that have been performed every year since then. A purely empirical answer is that at the present time there is still no evidence that the rate of false 5σ claims is running out of control. Sure, there is the occasional false alarm,[19] but this is balanced by the increased sophistication of experimental methods, in particular a better understanding of particle interactions inside detectors, the investment of large amounts of computer power in the modeling of background processes and systematic uncertainties, and the use of "safer" statistical techniques such as blind analysis.[20] In any case, professional statisticians are usually surprised by the stringency of our discovery threshold, and few of them would trust our ability to model the tails of distributions beyond 5σ. Thus, raising the current discovery threshold could not be justified without first demonstrating our understanding of such extreme tails.

3.3. *The Problem of Nuisance Parameters in the Calculation of p Values*

Often the distribution of the test statistic, and therefore the p value (10), depends on parameters that model various uninteresting background processes and instrumental features such as calorimeter energy scales and tracking efficiencies. The values of these parameters usually have uncertainties on them, known as systematic uncertainties, and since this complicates the evaluation of p values the corresponding parameters are referred to as "nuisance parameters". There is obviously considerable interest in methods for calculating p values that eliminate the dependence on nuisance parameters while taking into account the corresponding systematic uncertainties. In fact there are many such methods, but before we discuss them, it is useful to list some desiderata that we might wish them to satisfy:

(1) *Uniformity:* the method should preserve the uniformity of the null distribution of p values. If exact uniformity is not achievable in finite samples, then asymptotic uniformity should be aimed for.
(2) *Monotonicity:* for a fixed value of the observation, systematic uncertainties should decrease the significance of null rejections.
(3) *Generality:* the method should not depend on the testing problem having a special structure, but should be applicable to as wide a range of problems as possible.
(4) *Power:* all other things being equal, more power is better.

Keeping these criteria in mind, in the following subsections we discuss four classes of methods for eliminating nuisance parameters: structural, supremum, bootstrap, and predictive. Only the first three of these methods are compatible with frequentism; the last one requires a Bayesian concept of probability.

3.3.1. *Structural Methods*

We label "structural" any purely frequentist method that requires the testing problem to have a special structure in order to eliminate nuisance parameters. A classical example is the pivotal method introduced by W. S. Gossett. Assume we have $n \geq 2$ observations X_i from a Gaussian distribution with mean μ and standard deviation σ, both unknown, and suppose we wish to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, for a given value μ_0. The obvious test statistic here is the average $\bar{X}$ of all observations, but it can't be used because its distribution depends on the unknown parameter

σ. However, Gosset discovered that the quantity

$$T \;\equiv\; \frac{\bar{X}-\mu_0}{S/\sqrt{n}}, \quad \text{where} \quad S \;\equiv\; \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}, \tag{12}$$

is a pivot, i.e. a function of both data and parameters whose distribution under H_0 is itself independent of unknown parameters:

$$T \;\sim\; \frac{\Gamma(n/2)}{\sqrt{(n-1)\,\pi}\,\Gamma((n-1)/2)}\Big(1+\frac{t^2}{n-1}\Big)^{-n/2}. \tag{13}$$

Thus, if we evaluate T for our observed data, we can use the above distribution to calculate a p value and perform the desired test.

Another interesting example is the conditioning method: suppose that we have some data X and that there exists a statistic $C = C(X)$ such that the distribution of X given C is independent of the nuisance parameter(s). Then we can use that conditional distribution to calculate p values. A simple illustration of this idea involves observing a number of events N from a Poisson distribution with mean $\mu + \nu$, where μ represents a signal rate of interest, whereas ν is a nuisance parameter representing the rate of a background process. Without further knowledge about ν it is not possible to extract information from N about μ and hence to test the null hypothesis that $\mu = 0$. Suppose however that we perform a subsidiary experiment in which we observe M events from a Poisson distribution with mean $\tau\nu$, where τ is a known calibration constant. We have then:

$$N \;\sim\; \text{Poisson}(\mu+\nu) \quad \text{and} \quad M \;\sim\; \text{Poisson}(\tau\nu). \tag{14}$$

It turns out that this problem has the required structure for applying the conditioning method, if we use as conditioning statistic $C \equiv N+M$. Indeed, the probability of observing $N = n$ given $C = n+m$ is binomial under H_0:

$$\begin{aligned}\mathbb{P}(N=n\,|\,C=n+m) \;&=\; \frac{\mathbb{P}(N=n\,\&\,C=n+m)}{\mathbb{P}(C=n+m)} \\ &=\; \frac{\mathbb{P}(N=n\,\&\,M=m)}{\mathbb{P}(C=n+m)} \;=\; \frac{[\nu^n e^{-\nu}/n!]\;[(\tau\nu)^m e^{-\tau\nu}/m!]}{(\nu+\tau\nu)^{n+m}e^{-\nu-\tau\nu}/(n+m)!} \\ &=\; \binom{n+m}{n}\Big(\frac{1}{1+\tau}\Big)^n\Big(1-\frac{1}{1+\tau}\Big)^m. \end{aligned} \tag{15}$$

The dependence on ν has disappeared in the final expression for this probability, allowing one to compute a conditional p value:

$$p_{cond} \;=\; \sum_{i=n}^{n+m}\binom{n+m}{i}\Big(\frac{1}{1+\tau}\Big)^i\Big(1-\frac{1}{1+\tau}\Big)^{n+m-i}. \tag{16}$$

Since m/τ is the maximum likelihood estimate of ν from the subsidiary measurement, this p value is based on defining as more extreme those observations that have a larger N value and simultaneously a lower background estimate than the actual experiment. This method is sometimes used to evaluate the significance of a bump on top of a background spectrum, where "sidebands" provide a subsidiary measurement of the background level in the signal window. Fluctuations in both the signal window and the sidebands are Poisson.

In Fig. 1 we study the uniformity of the conditional p value (16) under H_0, for several values of τ and the true background magnitude ν_{true}. In

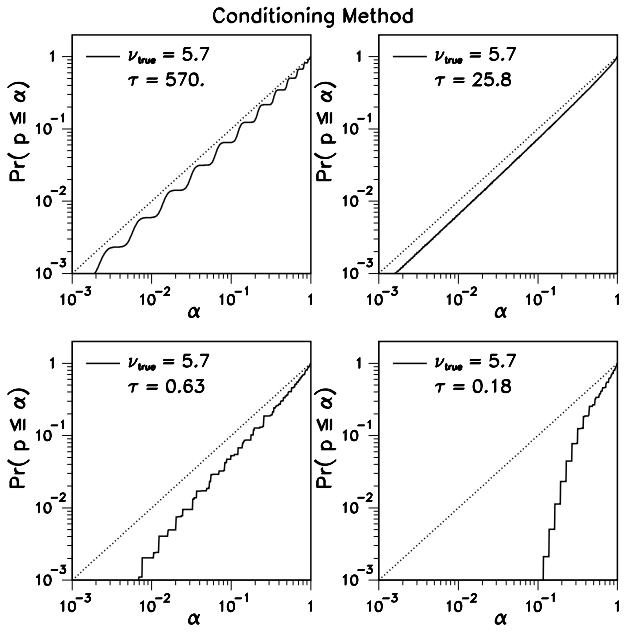

Fig. 1. Solid lines: cumulative probability distribution of conditional p values under the null hypothesis, $\mathbb{P}(p \leq \alpha \,|\, H_0)$ as a function of α. The dotted lines indicate a uniform distribution, $\mathbb{P}(p \leq \alpha \,|\, H_0) = \alpha$. Note the log-log scale.

all cases the p value turns out to be conservative, and the conservativeness increases as τ decreases, i.e. as the uncertainty on the background estimate increases. Note that if the problem only involved continuous statistics

instead of the discrete N and M, the conditional p value would be exact.

3.3.2. *Supremum Methods*

Structural methods have limited applicability due to their requirement that the testing problem have some kind of structure. A much more general technique consists in maximizing the p value with respect to the nuisance parameter(s):

$$p_{\sup} \;=\; \sup_{\nu} p(\nu). \tag{17}$$

This is a form of worst-case analysis: one reports the largest p value, or the smallest significance, over the whole parameter space. By construction $p_{\sup}$ is guaranteed to be conservative, but may yield the trivial result $p_{\sup} = 1$ if one is not careful in the choice of test statistic. In general the likelihood ratio is a good choice. For an example, we consider again the Poisson problem from the previous section, but this time with a Gaussian distribution with mean ν and standard deviation $\Delta\nu$ for the subsidiary measurement:

$$N \;\sim\; \text{Poisson}(\mu + \nu) \quad \text{and} \quad X \;\sim\; \text{Gauss}(\nu, \Delta\nu). \tag{18}$$

The joint likelihood is:

$$\mathcal{L}(\nu, \mu \,|\, n, x) \;=\; \frac{(\nu+\mu)^n \, e^{-\nu-\mu}}{n!} \; \frac{e^{-\frac{1}{2}\left(\frac{x-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu}, \tag{19}$$

and the likelihood ratio statistic is (compare eq. (9)):

$$\lambda \;=\; \frac{\sup\limits_{\nu\geq 0,\ \mu=0} \mathcal{L}(\nu, \mu \,|\, n, x)}{\sup\limits_{\nu\geq 0,\ \mu\geq 0} \mathcal{L}(\nu, \mu \,|\, n, x)}. \tag{20}$$

Small λ is evidence against H_0. It can be shown that for large values of ν, the quantity $-2\ln\lambda$ has the following distribution under H_0:

$$\begin{aligned} \mathbb{P}(-2\ln\lambda = 0) \;&=\; \frac{1}{2}, \\ \mathbb{P}(-2\ln\lambda > x) \;&=\; \frac{1}{2}\int_x^\infty \frac{e^{-t/2}}{\sqrt{2\pi x}}\, dx \;=\; \frac{1}{2}\left[1 - \operatorname{erf}\left(\sqrt{\frac{x}{2}}\right)\right]. \end{aligned} \tag{21}$$

For small ν however, the distribution of $-2\ln\lambda$ depends on ν and is a good candidate for the supremum method. Here the supremum p value can be rewritten as:

$$p_{\sup} \;=\; \sup_{\nu\geq 0} \mathbb{P}(\lambda \leq \lambda_0 \,|\, \mu = 0) \tag{22}$$

A great simplification occurs when $-2\ln\lambda$ is stochastically increasing [a] with ν, because then $p_{\rm sup} = p_\infty \equiv \lim_{\nu\to\infty} p(\nu)$ and we can still use (21). Unfortunately this is not generally true, and is often difficult to check. When $p_{\rm sup} \neq p_\infty$, p_∞ will tend to be liberal. Figure 2 shows the cumulative distribution of p_∞ under H_0, for problem (18) and several values of ν_{true} and $\Delta\nu$. It is seen that the p_∞ approximation to $p_{\rm sup}$ is generally conservative,

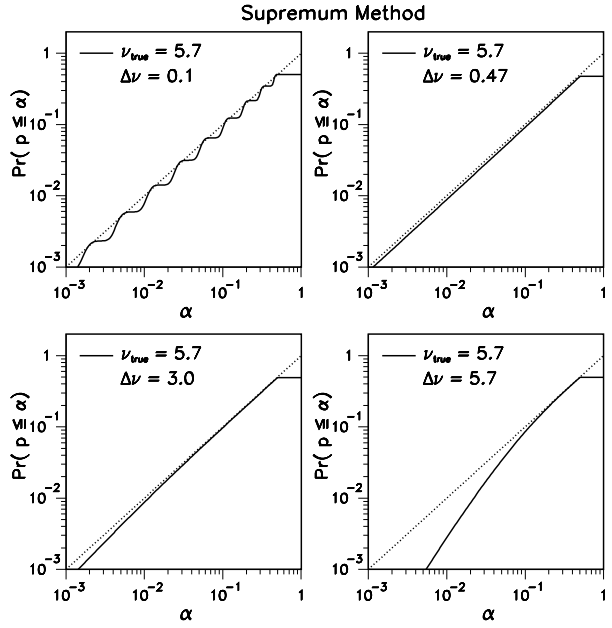

Fig. 2. Cumulative probability distribution, under the null hypothesis, of the asymptotic approximation to the supremum p value, for a Poisson event count with Gaussian measurement of the mean.

except at low $\Delta\nu$, where some minor, localized liberalism can be detected.

The supremum method has two important drawbacks. Computationally, it is often difficult to locate the global maximum of the relevant tail probability over the entire range of the nuisance parameter ν. Secondly, the

[a] A statistic X with cumulative distribution $F(x\,|\,\theta)$ is stochastically increasing with the parameter θ if $\theta_1 > \theta_2$ implies $F(x\,|\,\theta_1) \leq F(x\,|\,\theta_2)$ for all x and $F(x\,|\,\theta_1) < F(x\,|\,\theta_2)$ for some x. In other words, X tends to be larger for larger values of θ.

very data one is analyzing often contain information about the true value of ν, so that it makes little sense to maximize over *all* values of ν. A simple way around these drawbacks is to maximize over a $1-\gamma$ confidence set C_γ for ν (see section 4.1), and then to correct the p value for the fact that γ is not zero:

$$p_\gamma = \sup_{\nu \in C_\gamma} p(\nu) + \gamma. \tag{23}$$

This time the supremum is restricted to all values of ν that lie in the confidence set C_γ. It can be shown that p_γ, like $p_{\sup}$, is conservative:

$$\mathbb{P}(p_\gamma \leq \alpha) \;\leq\; \alpha \quad \text{for all } \alpha \in [0,1]. \tag{24}$$

Although there is a lot of flexibility in the choice of γ and C_γ, both should be chosen *before* looking at the data.

3.3.3. *Bootstrap Methods*

The first bootstrap method we consider is the plug-in. It gets rid of unknown parameters by estimating them, using for example a maximum-likelihood estimate, and then substituting the estimate in the calculation of the p value. For example (18) with likelihood function (19), the maximum-likelihood estimate of ν under H_0 is obtained by setting $\mu = 0$ and solving $\partial \ln \mathcal{L}/\partial\nu = 0$ for ν. This yields:

$$\hat{\nu}(x,n) \;=\; \frac{x - \Delta\nu^2}{2} \;+\; \sqrt{\left(\frac{x-\Delta\nu^2}{2}\right)^2 \;+\; n\,\Delta\nu^2}. \tag{25}$$

The plug-in p value is then:

$$p_{plug}(x,n) \;\equiv\; \sum_{k=n}^{+\infty} \frac{\hat{\nu}(x,n)^k\; e^{-\hat{\nu}(x,n)}}{k!}. \tag{26}$$

In principle two criticisms can be leveled at the plug-in method. Firstly, it makes double use of the data, once to estimate the nuisance parameters under H_0, and then again to calculate a p value. Secondly, it does not take into account the uncertainty on the parameter estimates. The net effect is that plug-in p values tend to be too conservative. The adjusted plug-in method attempts to overcome this.

If we knew the exact cumulative distribution function F_{plug} of plug-in p values under H_0, then the quantity $F_{plug}(p_{plug})$ would be an exact p value since its distribution is uniform by construction. In general however, F_{plug} depends on one or more unknown parameters and can therefore not be used

in this way. The next best thing we can try is to substitute estimates for the unknown parameters in F_{plug}. Accordingly, one defines the adjusted plug-in p value by:

$$p_{plug,adj} \equiv F_{plug}(p_{plug} \,|\, \hat{\theta}), \tag{27}$$

where $\hat{\theta}$ is an estimate for the unknown parameters collectively labeled by θ. This adjustment algorithm is known as a double parametric bootstrap and can also be implemented in Monte Carlo form.

Some cumulative distributions of the plug-in and adjusted plug-in p values are plotted in Fig. 3 for example (18). The adjusted plug-in p value provides a strikingly effective correction for the overconservativeness of the plug-in p value.

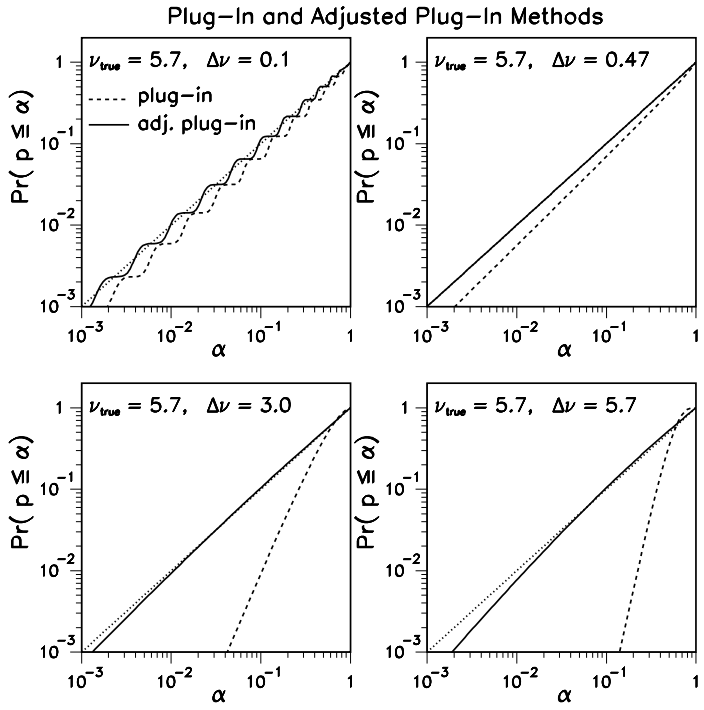

Fig. 3. Cumulative probability distribution of the plug-in (dashed lines) and adjusted plug-in (solid lines) p values under the null hypothesis for a Poisson event count with Gaussian measurement of the mean.

3.3.4. *Predictive Methods*

So far we have assumed that information about the nuisance parameter comes from a subsidiary measurement. This allows one to treat the problem of eliminating nuisance parameters in a purely frequentist way. The structural, supremum, and confidence interval methods are guaranteed to be conservative. The asymptotic approximation to the supremum method and the bootstrap methods do not provide this guarantee but are still frequentist. We now turn to the situation where information about the nuisance parameter comes in the form of a Bayesian prior. We discuss two approaches, known as prior-predictive and posterior-predictive.

The prior-predictive distribution of a test statistic T is the predicted distribution of T before the measurement:

$$m_{prior}(t) \;=\; \int p(t \,|\, \theta)\; \pi(\theta)\; d\theta, \tag{28}$$

where $\pi(\theta)$ is the prior probability density of θ. After having observed $T = t_0$ we can quantify how surprising this observation is by referring t_0 to m_{prior}, e.g. by calculating the prior-predictive p value:

$$\begin{aligned} p_{prior} \;=\; \mathbb{P}_{m_{prior}}(T \geq t_0 \,|\, H_0) \;&=\; \int_{t_0}^{\infty} m_{prior}(t)\; dt \\ &=\; \int \pi(\theta) \left[\int_{t_0}^{\infty} p(t \,|\, \theta)\; dt \right]\; d\theta, \end{aligned} \tag{29}$$

where the last equality follows from interchanging two integral signs. This last expression for p_{prior} shows that the prior-predictive p value can be interpreted as the average of the usual p value over the prior for the unknown parameter.

The posterior-predictive distribution of a test statistic T is the predicted distribution of T *after* measuring $T = t_0$:

$$m_{post}(t \,|\, t_0) \;=\; \int p(t \,|\, \theta)\; \pi(\theta \,|\, t_0)\; d\theta. \tag{30}$$

The posterior-predictive p value estimates the probability that a *future* observation will be at least as extreme as the current observation if the null hypothesis is true:

$$\begin{aligned} p_{post} \;=\; \mathbb{P}_{m_{post}}(T \geq t_0 \,|\, H_0) \;&=\; \int_{t_0}^{\infty} m_{post}(t \,|\, t_0)\; dt \\ &=\; \int \pi(\theta \,|\, t_0) \left[\int_{t_0}^{\infty} p(t \,|\, \theta)\; dt \right]\; d\theta. \end{aligned} \tag{31}$$

As the last expression on the right shows, the posterior-predictive p value can also be written as an average, this time over the posterior for the unknown parameter. Note the double use of the observation t_0 in p_{post}: first to compute the posterior for θ, and then again in the tail probability calculation. We encountered the same feature in the definition of the plug-in p value, and the same effect will be observed here, namely that the posterior-predictive p value is overly conservative.

What about the uniformity of p_{prior} and p_{post}? How well calibrated are these predictive p values? The answer depends on the distribution of the test statistic T under the null hypothesis. One can argue that this should be the prior-predictive distribution (28), since this distribution is fully specified and is available before observing the data. It is clear that, by construction, p_{prior} will be uniform with respect to the prior-predictive distribution. On the other hand, because of its double-use of the data, p_{post} will be conservative.

Frequentists will argue that the prior-predictive distribution is not frequentist and therefore does not provide a valid reference ensemble to check the uniformity of p_{prior} and p_{post}. If the testing problem of interest is purely frequentist, a different approach is in fact possible. Consider for example the Poisson+Gauss problem of eq. (18). One way to apply a predictive method to this problem is to construct a posterior for the subsidiary Gaussian measurement of ν, and then use this posterior as a prior for ν when calculating a predictive p value for the Poisson event count N. We still need a prior for the subsidiary measurement however, and in the absence of further information about ν, it is appropriate to use an objective rule such as Jeffreys'. For a Gaussian likelihood with unknown mean, the Jeffreys' prior (5) is a constant. Thus the subsidiary posterior is:

$$\pi_{\text{sub.}}(\nu \,|\, x) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{\nu - x}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu\,\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}\,\Delta\nu}\right)\right]}, \tag{32}$$

where the normalization comes from the requirement that ν, being a Poisson mean, is a positive parameter. We can use this posterior as a prior to construct p_{prior} and p_{post}. Furthermore, for every value of ν we now have a frequentist reference ensemble to check the uniformity of these p values, namely the set of all (X, N) pairs where X is a Gaussian variate with mean ν and standard deviation $\Delta\nu$, and N is an independent Poisson variate with mean ν. Contrast this with the reference ensemble represented by the prior-predictive distribution, which is defined for every value of x rather than every value of ν, and is the set of (ν, N) pairs where ν is a Gaussian

variate with mean x and standard deviation $\Delta\nu$, and N is a *dependent* Poisson variate whose mean is the ν value in the same pair. Because of the random nature of the parameter ν in this ensemble, it is clearly Bayesian. Figure 4 shows the cumulative distributions of p_{prior} and p_{post} with respect to the frequentist ensemble, for several values of $\Delta\nu$. Both p values appear to be (mostly) conservative, and p_{post} much more so than p_{prior}, especially at large $\Delta\nu$.

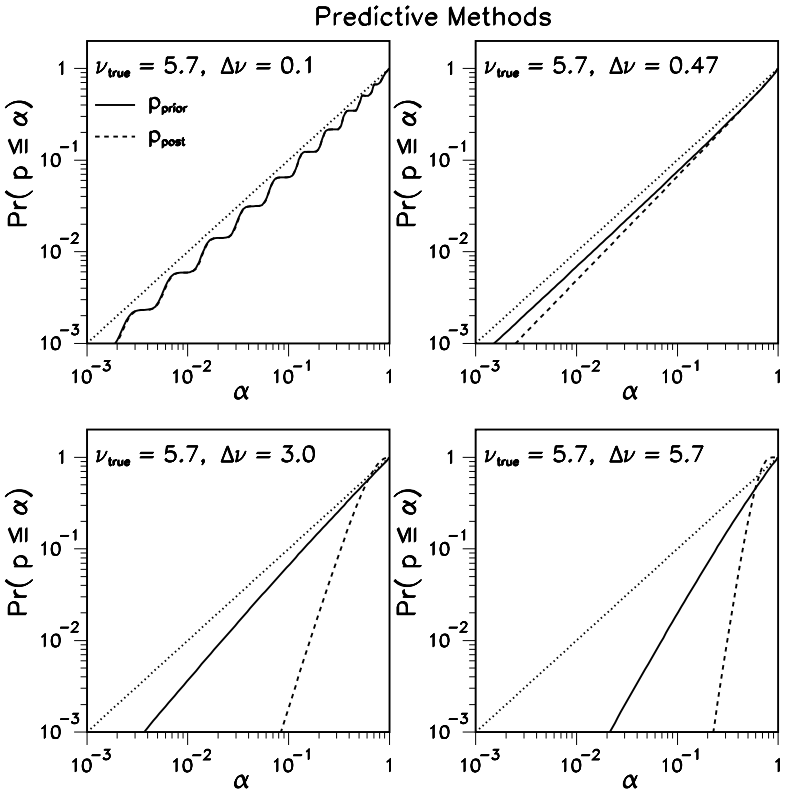

Fig. 4. Cumulative distributions of the prior-predictive (solid lines) and posterior-predictive (dashed lines) p values for a Poisson event count with a Gaussian uncertainty on the mean. The dotted lines correspond to exact p values.

We end this discussion of predictive p values with some general comments:

- Prior-predictive p values cannot be defined for improper priors; in this case, posterior-predictive p values often provide a solution.
- Posterior-predictive p values can be calculated for discrepancy variables (i.e. functions of data *and* parameters) in addition to test statistics.
- Rather than simply reporting a predictive p value, it may be more informative to plot the observed value of the test statistic against the appropriate predictive distribution.
- There are other types of predictive p values, which avoid some of the problems of the prior- and posterior-predictive p values.[21]

3.3.5. *Summary of p Value Methods*

To guide our summary of the various nuisance parameter elimination methods just described, we return to the desiderata listed at the beginning of section 3.3.

Figures 1 to 4 indicate quite a variation in uniformity, or rather lack thereof, of p value distributions under the null hypothesis. For the examples studied, the adjusted plug-in and supremum methods perform quite well, but this behavior depends strongly on the choice of test statistic. The likelihood ratio is generally a good choice. Our examples also show that uniformity tends to be violated on the conservative side, but this is only guaranteed for fully frequentist methods such as conditioning, supremum, and confidence interval. For other methods uniformity will have to be checked explicitly for the problem at hand. This is of course important if one wants to avoid overestimating the significance of a result.

An interesting point to note is that some p values tend to converge in the asymptotic limit. This is numerically illustrated for example (18) in Table 1, which shows that the supremum, adjusted plug-in, and prior-predictive p values give almost identical results on a data sample of thousands of events. Whenever possible, it is always instructive to compare the results of different methods.

Figure 5 compares the power functions of the supremum, adjusted plug-in, and prior-predictive p values for problem (18). There is not much difference between the curves, except perhaps at high $\Delta\nu$, where the prior-predictive p value seems somewhat less powerful. Note that as the signal strength goes to zero, the power function converges to α if the p value is exact.

Finally, we comment on the monotonicity property: for the examples and methods studied here, it is true that the p value increases with the magnitude of the systematic uncertainty. In other words, significance claims

Table 1. P values for a Poisson observation of $n_0 = 3893$ events over an estimated background of $x_0 = 3234 \pm \Delta\nu$ events, where $\Delta\nu = 10$ or 100. For the confidence interval p value a 6σ upper limit was constructed for the nuisance parameter ($\gamma = 9.87 \times 10^{-10}$).

Method	$\Delta\nu = 10$		$\Delta\nu = 100$	
	P value	N_σ	P value	N_σ
Supremum	1.16×10^{-28}	11.05	9.81×10^{-9}	5.62
Confidence Interval	9.87×10^{-10}	6.00	1.23×10^{-8}	5.58
Plug-In	8.92×10^{-28}	10.86	1.86×10^{-3}	2.90
Adjusted Plug-In	1.13×10^{-28}	11.05	9.90×10^{-9}	5.61
Prior-Predictive	1.23×10^{-28}	11.04	9.85×10^{-9}	5.61
Posterior-Predictive	5.27×10^{-27}	10.70	1.35×10^{-2}	2.21

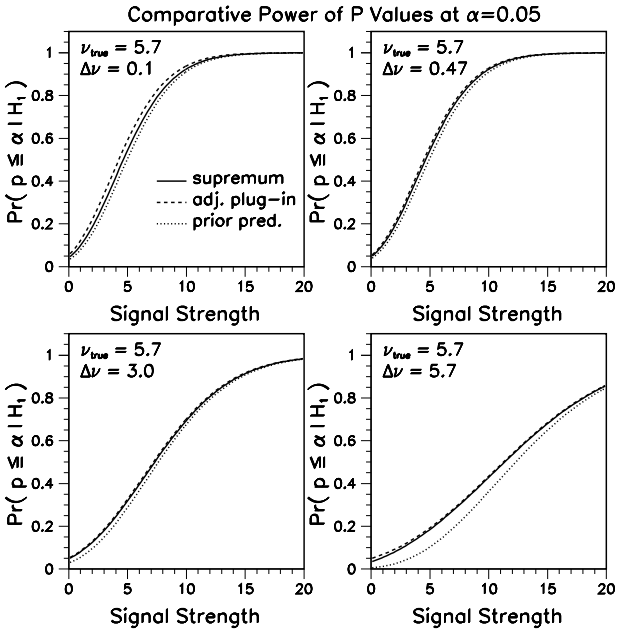

Fig. 5. Power functions of the supremum (solid), adjusted plug-in (dashed), and prior-predictive (dotted) p values for testing for the presence of a Poisson signal on top of a Poisson background whose mean ν_{true} has a Gaussian uncertainty $\Delta\nu$. The power is calculated for a test level of $\alpha = 0.05$ and is plotted as a function of true signal strength.

are degraded by the presence of systematics. However, in practical problems not covered by this review, monotonicity will have to be checked explicitly.

3.4. *Caveats about the Likelihood Ratio Statistic*

As mentioned previously, the likelihood ratio λ defined in (9) is often a good choice of test statistic, because it is intuitively sensible, and is even optimal in the special case of simple vs. simple testing. Although this optimality does not transfer to the testing of composite hypotheses,[22] λ remains popular in that case due to Wilks' theorem, which gives the asymptotic distribution of $-2\ln\lambda$ under the null hypothesis as that of a chisquared (see section 3.1). Unfortunately, the conditions for this theorem to be applicable do not always hold in high energy physics problems. What follows are some examples where these regularity conditions are violated.

- One of the regularity conditions is that the tested hypotheses must be nested, i.e. H_0 must be obtainable by imposing parameter restrictions on the model that describes H_1. A counter-example is a test that compares two new-physics models that belong to separate families of distributions.
- Another regularity condition is that H_0 should not be on the boundary of the model that describes H_1. A typical violation of this condition is when θ is a positive signal magnitude and one is testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$.
- A third condition is that there must not be any nuisance parameters that are defined under H_1 but not under H_0. Suppose for example that we are searching for a signal peak on top of a smooth background. The location, width, and amplitude of the peak are unknown. In this case the location and width of the peak are undefined under H_0, i.e. when the amplitude is zero. Hence $-2\ln\lambda$ will not have a chisquared distribution under H_0.

There does exist some analytical work on the distribution of the likelihood ratio when the above regularity conditions are violated; however, these results are not always easy to apply and still require some numerical calculations. Physicists aware of the limitations of Wilks' theorem usually prefer to estimate the distribution of $-2\ln\lambda$ with the help of a Monte Carlo calculation. The advantage of this approach is that it allows one to incorporate all the relevant details of the experimental data analysis; the disadvantage is that it sometimes requires enormous amounts of CPU time.

3.5. *Expected Significances*

Probably the most useful way to describe the sensitivity of a model of new physics, given specific instrumental conditions, is to calculate the integrated luminosity for which there is a 50% probability of claiming discovery at the 5σ level. The calculation can be done as follows:

(1) Compute (or simulate) the distribution of p values under the new physics model and assuming a fixed integrated luminosity.
(2) Find the median of the p value distribution from (1).
(3) Repeat steps (1) and (2) for several values of the integrated luminosity and interpolate to find the integrated luminosity at which the median p value is 2.7×10^{-7} (5σ).

To determine the most sensitive method, or the most sensitive test statistic for discovering new physics, another useful measure is the expected significance level (ESL), defined as the observed p value averaged over the new physics hypothesis. If the test statistic X has density $f_i(x)$ under H_i, and if $p = 1 - F_0(X) \equiv 1 - \int_{-\infty}^{X} f_0(t)\,dt$, then:

$$\text{ESL} \;\equiv\; \mathbb{E}(p \,|\, H_1) \;=\; \int [1 - F_0(x)]\, f_1(x)\, dx \;=\; \int F_1(x)\, f_0(x)\, dx. \qquad (33)$$

The integral on the right is easy to estimate by Monte Carlo, since it represents the probability that $X \geq Y$, where X and Y are independent random variables distributed according to F_0 and F_1, respectively.

3.6. *Combining Significances*

When searching for new physics in several different channels, or via different experiments, it is sometimes desired to summarize the search by calculating a combined significance. This is a difficult problem. The best approach is to combine the likelihood functions for all the channels and derive a p value from the combined likelihood ratio statistic. However, it may not always be possible or practical to do such a calculation. In this case, if the individual p values are independent, another possibility is to combine the p values directly.[23] Unfortunately there is no unique way of doing this. The general idea is to choose a rule $S(p_1, p_2, p_3, \ldots)$ for combining individual p values p_1, p_2, p_3,..., and then to construct a combined p value by calculating the tail probability corresponding to the observed value of S. Some plausible combination rules are:

(1) The product of p_1, p_2, p_3,... (Fisher's rule);

(2) The smallest of p_1, p_2, p_3,... (Tippett's rule);
(3) The average of p_1, p_2, p_3,...;
(4) The largest of p_1, p_2, p_3,....

This list is by no means exhaustive. To narrow down the options, there are some properties of the combined p value that one might consider desirable. For example:

(1) If there is strong evidence against the null hypothesis in at least one channel, then the combined p value should reflect that, by being small.
(2) If none of the individual p values shows any evidence against the null hypothesis, then the combined p value should not provide such evidence.
(3) Combining p values should be associative: the combinations $((p_1, p_2), p_3)$, $((p_1, p_3), p_2)$, $(p_1, (p_2, p_3))$, (p_1, p_2, p_3), should all give the same result.

Now, it turns out that property 1 eliminates rules 3 and 4; property 2 is satisfied by all four rules, and property 3, called evidential consistency, is satisfied by none. This leaves Tippett's and Fisher's rules as reasonable candidates. Actually, it appears that Fisher's rule has somewhat more uniform sensitivity to alternative hypotheses of interest in most problems. So Fisher's rule is quite popular.

Here is a simple mathematical trick to combine n p-values by Fisher's rule: take twice the negative logarithm of their product and treat it as a chisquared variate for $2n$ degrees of freedom (this is valid because the cumulative distribution of a chisquared variate for 2 d.o.f. is $1 - e^{-x/2}$, and chisquared variates are additive). The general result is that

$$p_{\rm comb} \equiv \Pi \sum_{j=0}^{n-1} \frac{(-\ln \Pi)^j}{j!}, \quad \text{where } \Pi \equiv \prod_{j=1}^{n} p_j, \tag{34}$$

will have a uniform distribution under H_0 if the individual p_i are uniform. One situation in which the p_i will not be uniform is if they are derived from discrete test statistics. In this case the formula will give a combined p value that is larger than the correct one, and therefore conservative.

The literature on combining p values is extensive; see Ref. 24 for an annotated bibliography.

3.7. *Bayesian Hypothesis Testing*

The Bayesian approach to hypothesis testing is to calculate posterior probabilities for all hypotheses in play. When testing H_0 versus H_1, Bayes'

theorem yields:

$$\pi(H_0 \,|\, x) \;=\; \frac{p(x \,|\, H_0)\,\pi_0}{p(x \,|\, H_0)\,\pi_0 \;+\; p(x \,|\, H_1)\,\pi_1}, \tag{35}$$

$$\pi(H_1 \,|\, x) \;=\; 1 \;-\; \pi(H_0 \,|\, x), \tag{36}$$

where π_i is the prior probability of H_i, $i = 0, 1$. If $\pi(H_0 \,|\, x) < \pi(H_1 \,|\, x)$, one rejects H_0 and the posterior probability of error is $\pi(H_0 \,|\, x)$. Otherwise H_0 is accepted and the posterior error probability is $\pi(H_1 \,|\, x)$.

In contrast with frequentist Type-I and Type-II errors, which are known *before* looking at the data, Bayesian error probabilities are fully conditioned on the observations. They do depend on the prior hypothesis probabilities however, and it is often interesting to look at the evidence against H_0 provided by the data alone. This can be done by computing the ratio of posterior odds to prior odds and is known as the Bayes factor:

$$B_{01}(x) \;=\; \frac{\pi(H_0 \,|\, x)/\pi(H_1 \,|\, x)}{\pi_0/\pi_1} \tag{37}$$

In the absence of unknown parameters, $B_{01}(x)$ is a likelihood ratio.

Often the distributions of X under H_0 and H_1 will depend on unknown parameters θ, so that posterior hypothesis probabilities and Bayes factors will involve marginalization integrals over θ:

$$\pi(H_0 \,|\, x) \;=\; \frac{\displaystyle\int p(x \,|\, \theta, H_0)\,\pi(\theta \,|\, H_0)\,\pi_0\,d\theta}{\displaystyle\int \Big[p(x \,|\, \theta, H_0)\,\pi(\theta \,|\, H_0)\,\pi_0 \;+\; p(x \,|\, \theta, H_1)\,\pi(\theta \,|\, H_1)\,\pi_1\Big]\,d\theta} \tag{38}$$

$$\text{and:}\quad B_{01}(x) \;=\; \frac{\displaystyle\int p(x \,|\, \theta, H_0)\,\pi(\theta \,|\, H_0)\,d\theta}{\displaystyle\int p(x \,|\, \theta, H_1)\,\pi(\theta \,|\, H_1)\,d\theta} \tag{39}$$

Suppose now that we are testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Then:

$$B_{01}(x) \;=\; \frac{p(x \,|\, \theta_0)}{\displaystyle\int p(x \,|\, \theta, H_1)\,\pi(\theta \,|\, H_1)\,d\theta} \;\geq\; \frac{p(x \,|\, \theta_0)}{p(x \,|\, \hat{\theta}_1)} \;=\; \lambda, \tag{40}$$

where $\hat{\theta}_1$ maximizes $p(x \,|\, \theta, H_1)$. Thus, the ratio between the Bayes factor and the corresponding likelihood ratio is larger than 1. It is sometimes called the Ockham's razor penalty factor: it penalizes the evidence against H_0 for the introduction of an additional degree of freedom under H_1, namely θ.[25]

The smaller B_{01}, or equivalently, the larger $B_{10} \equiv 1/B_{01}$, the stronger the evidence against H_0. A rough descriptive statement of standards of evidence provided by Bayes factors against a hypothesis is given in Table 2.[26] There is at present not much experience with Bayes factors in high energy physics.

Table 2. Verbal description of standards of evidence provided by Bayes factors.

$2 \ln B_{10}$	B_{10}	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

For a hypothesis of the form $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, a Bayesian test can be based directly on the posterior distribution of θ. First calculate an interval for θ, containing an integrated posterior probability β. Then, if θ_0 is outside that interval, reject H_0 at the $\alpha = 1 - \beta$ credibility level. An exact significance level can be obtained by finding the smallest α for which H_0 is rejected. There is a lot of freedom in the choice of posterior interval. A natural possibility is to construct a highest posterior density (HPD) interval. If the lack of parametrization invariance of HPD intervals is a problem, there are other choices (see section 4.4).

If the null hypothesis is $H_0 : \theta \leq \theta_0$, a valid approach is to calculate a lower limit θ_L on θ and exclude H_0 if $\theta_0 < \theta_L$. In this case the exact significance level is the posterior probability of $\theta \leq \theta_0$.

4. Interval Estimation

Suppose that we make an observation $X = x_{obs}$ from a distribution $f(x \,|\, \mu)$, where μ is a parameter of interest, and that we wish to make a statement about the location of the true value of μ, based on our observation x_{obs}. One possibility is to calculate a point estimate $\hat{\mu}$ of μ, for example via the maximum-likelihood method:

$$\hat{\mu} \;=\; \arg\max_{\mu} f(x_{obs} \,|\, \mu). \tag{41}$$

Although such a point estimate has its uses, it comes with no measure of how confident we can be that the true value of μ equals $\hat{\mu}$.

Bayesianism and Frequentism both address this problem by constructing an interval of μ values believed to contain the true value with some confidence. However, the interval construction method and the meaning of the associated confidence level are very different in the two paradigms.

On the one hand, frequentists construct an interval $[\mu_1, \mu_2]$ whose boundaries μ_1 and μ_2 are random variables that depend on X in such a way that if the measurement is repeated many times, a fraction γ of the produced intervals will cover the true μ; the fraction γ is called the confidence level or coverage of the interval construction.

On the other hand, Bayesians construct the posterior probability density of μ and choose two values μ_1 and μ_2 such that the integrated posterior probability between them equals a desired level γ, called credibility or Bayesian confidence level of the interval.

4.1. *Frequentist Intervals: the Neyman Construction*

The Neyman construction is the most general method available for constructing interval estimates that have a guaranteed frequentist interpretation. The principal steps of the construction are illustrated in Fig. 6 for the simplest case of a one-dimensional continuous observation X whose probability distribution depends on an unknown one-dimensional continuous parameter μ. The procedure can be described as follows:

Step 1: Make a graph of the parameter μ versus the data X, and plot the density distribution of X for several values of μ (plot a);

Step 2: For each value of μ, select an interval of X values that has a fixed integrated probability, for example 68% (plot b);

Step 3: Connect the interval boundaries across μ values (plot c);

Step 4: Drop the "scaffolding", keeping only the two lines drawn at step 3; these form a *confidence belt* that can be used to construct an interval $[\mu_1, \mu_2]$ for the true value of μ every time you make an observation x_{obs} of X (plot d).

To see why this procedure works, refer to Fig. 7. Suppose that $\mu^\star$ is the true value of μ. Then $\mathbb{P}(x_1 \leq X \leq x_2 \,|\, \mu^\star) = 68\%$ by construction. Furthermore, for every $X \in [x_1, x_2]$, the reported μ interval will contain $\mu^\star$ and for every $X \notin [x_1, x_2]$, the reported μ interval will *not* contain $\mu^\star$. Therefore, the probability of covering $\mu^\star$ is exactly 68%, and this holds regardless of the value of $\mu^\star$. For problems with discrete statistics (such as Poisson event counts), the construction yields intervals that are conservative, i.e. which cover above the nominal level for some parameter values.

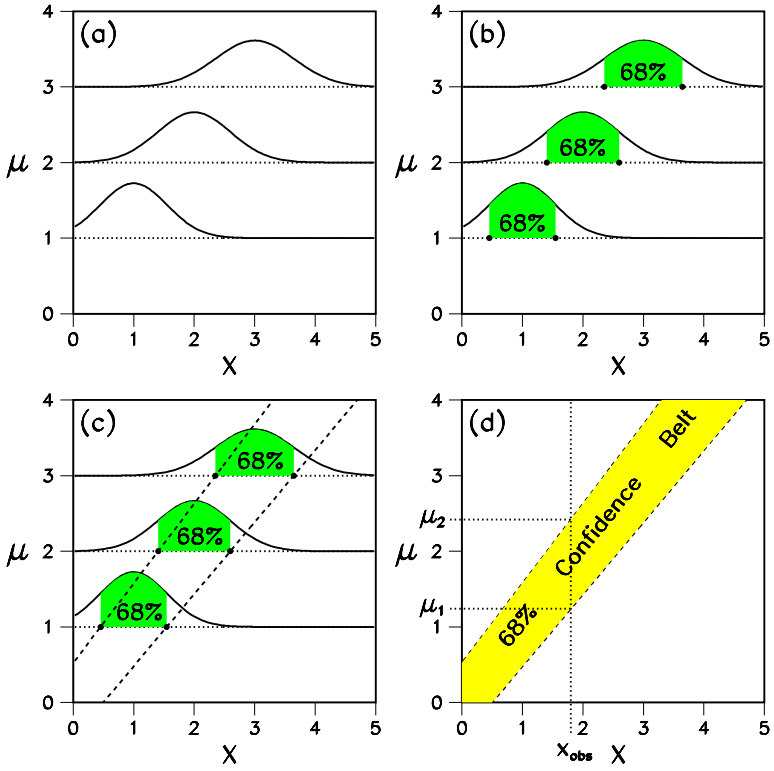

Fig. 6. Four steps in the Neyman construction of confidence intervals (see text).

There are four basic ingredients in the Neyman construction: an estimator $\hat{\mu}$ of the parameter of interest μ, an ordering rule, a reference ensemble, and a confidence level. We now take a look at each of these individually.

4.1.1. *Ingredient 1: the Estimator*

The estimator is the quantity plotted along the abscissae in the Neyman construction plot. Suppose for example that we collect n independent measurements x_i of the mean μ of a Gaussian distribution with known standard deviation. Then clearly we should use the average $\bar{x}$ of the x_i as an esti-

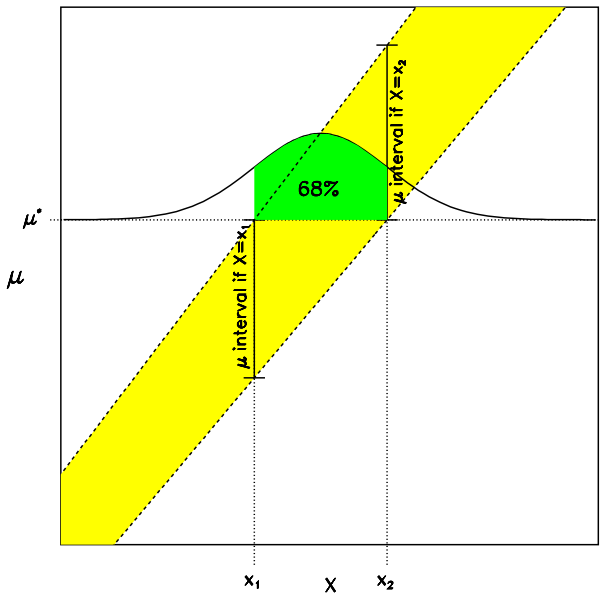

Fig. 7. Why the Neyman construction works (see text).

mate of μ, since $\bar{x}$ is a sufficient statistic [b] for μ. On the other hand, if μ is constrained to be positive, then it would make sense to use either $\hat{\mu} = \bar{x}$ or $\hat{\mu} = \max\{0, \bar{x}\}$. These two estimators lead to intervals with very different properties. We will come back to this example in section 4.5.

4.1.2. *Ingredient 2: the Ordering Rule*

The ordering rule is the rule we use to decide which X values to include in the interval at step 2 of the construction. The only constraint on that interval is that it must contain 68% of the X distribution (or whatever confidence level is desired for the overall construction). For example, we could start with the X value that has the largest probability density and then keep adding values with lower and lower probability density until we

[b] A statistic $T(X)$ is sufficient for μ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on μ. In a sense, $T(X)$ captures all the information about μ contained in the sample.

cover 68% of the distribution. Another possibility is to start with $X = -\infty$ and add increasing values of X, again until we reach 68%. Of course, in order to obtain a smooth confidence belt at the end, we should choose the ordering rule consistently from one μ value to the next. In this sense it is better to formulate the ordering rule in terms of μ rather than X. This emphasizes the inferential meaning of the resulting intervals: an ordering rule is a rule that orders parameter values according to their perceived compatibility with the observed data. Here are some examples, all assuming that we have observed data x and are interested in a 68% confidence interval $[\mu_1, \mu_2]$ for a parameter μ whose maximum likelihood estimate is $\hat{\mu}(x)$:

- Central ordering
  $[\mu_1, \mu_2]$ is the set of μ values for which the observed data falls between the 16^{th} and 84^{th} percentiles of its distribution.
- Probability density ordering
  $[\mu_1, \mu_2]$ is the set of μ values for which the observed data falls within the 68% most probable region of its distribution.
- Likelihood ratio ordering
  $[\mu_1, \mu_2]$ is the set of μ values for which the observed data falls within a 68% probability region R, such that any point x inside R has a larger likelihood ratio $\mathcal{L}(\mu \,|\, x)/\mathcal{L}(\hat{\mu}(x) \,|\, x)$ than any point outside R.
- Upper limit ordering
  $]-\infty, \mu_2]$ is the set of μ values for which the observed data is at least as large as the 32^{nd} percentile of its distribution.
- Minimal expected length
  This rule minimizes the average interval length $(\mu_2(X) - \mu_1(X))$ over the sample space.

4.1.3. *Ingredient 3: the Reference Ensemble*

This refers to the replications of a measurement that are used to calculate coverage. In order to specify these replications, one must decide which random and non-random aspects of the measurement are relevant to the inference of interest. When measuring the mass of a short-lived particle for example, it may be that its decay mode affects the measurement resolution. Should we then refer our measurement to an ensemble that includes all possible decay modes, or only the decay mode actually observed?

For simplicity assume that the estimator X of the mass μ is normal with mean μ and standard deviation σ, and that there is a $p = 50\%$ probability that the particle will decay hadronically, in which case $\sigma \equiv \sigma_h = 10$; other-

wise the particle decays leptonically and $\sigma \equiv \sigma_\ell = 1$. As interval ordering rule we'll use minimal expected length. Since the decay mode is observable, one can proceed in two ways:

- Unconditional minimization;
  The reference ensemble includes all decay modes. We report $x \pm \delta_h$ if the decay is hadronic and $x \pm \delta_\ell$ if it is leptonic, where δ_h and δ_ℓ are constants that minimize the expected interval length, $2[p\delta_h + (1-p)\delta_\ell]$, subject to the constraint of 68% coverage over the whole reference ensemble. Substituting the given numbers, this yields $\delta_h = 5.06$, $\delta_\ell = 2.20$, and an expected length of 7.26.
- Conditional minimization;
  The reference ensemble includes only the observed decay mode. We report $x \pm \sigma_h$ in the hadronic case and $x \pm \sigma_\ell$ in the leptonic one; the expected interval length is $2[p\sigma_h + (1-p)\sigma_\ell] \;=\; 11.0$.

The expected interval length is quite a bit larger for the conditional method than for the unconditional one. If one were to repeat the measurement a large number of times, one would find that in the conditional analysis the coverage of the interval is 68% both within the subensemble of hadronic decays and within the subensemble of leptonic decays. On the other hand, in the unconditional analysis the coverage is 39% for hadronic decays and 97% for leptonic decays, correctly averaging to 68% over all decays combined. Qualitatively, by shifting some coverage probability from the hadronic decays to the higher precision leptonic ones, the unconditional construction is able to reduce the average interval length.

The above problem is an adaptation to high-energy physics of a famous example in the statistics literature,[27,28] used to discuss the merits of conditioning versus power (or interval length).

4.1.4. *Ingredient 4: the Confidence Level*

The confidence level labels a family of intervals; some conventional values are 68%, 90%, and 95%. It is very important to remember that a confidence level does *not* characterize single intervals; it only characterizes families of intervals. The following example illustrates this.

Suppose we are interested in the mean μ of a Gaussian population with unit variance. We have two observations, x and y, so that the maximum likelihood estimate of μ is $\hat{\mu} = (x+y)/2$. Consider the following two intervals

for μ:

$$I_1 : \ \hat{\mu} \pm 1/\sqrt{2} \quad \text{and} \quad I_2 : \ \hat{\mu} \pm \sqrt{\max\{\, 0,\ 4.60 - (x-y)^2/4\}}$$

Both I_1 and I_2 are centered on the maximum likelihood estimate of μ. Interval I_1 uses likelihood ratio ordering, is never empty, and has 68% coverage. Interval I_2 uses probability density ordering, is empty whenever $|x - y| \geq 4.29$, and has 99% coverage. Suppose next that we observe $x = 10.00$ and $y = 14.05$. It is easy to verify that the corresponding I_1 and I_2 intervals are numerically identical and equal to 12.03 ± 0.71. Thus, the same numerical interval can have two very different coverages (confidence levels), depending on which ensemble it is considered to belong to.

4.2. *Handling of Nuisance Parameters in the Neyman Construction*

In principle the Neyman construction can be performed when there is more than one parameter; it simply becomes a multidimensional construction, and the confidence belt becomes a "hyperbelt". If some parameters are nuisances, they can be eliminated by projecting the final confidence region onto the parameter(s) of interest at the end of the construction. This is a difficult problem: the ordering rule has to be designed so as to minimize the amount of overcoverage introduced by projecting.

There are simpler solutions. A popular one is to eliminate the nuisance parameters ν from the data probability density function (pdf) first, by integrating them over proper prior distributions:

$$f(x \,|\, \mu, \nu) \ \rightarrow \ \tilde{f}(x \,|\, \mu) \ \equiv \ \int f(x \,|\, \mu, \nu)\, \pi(\nu)\, d\nu \tag{42}$$

This is a Bayesian step: the data pdf it yields depends only on the parameter(s) of interest and can then be used in a standard Neyman construction.

Another possibility is to eliminate the nuisance parameters by profiling the pdf. This is particularly useful if one has an independent measurement y of ν, with pdf $g(y \,|\, \nu)$:

$$f(x \,|\, \mu, \nu) \ \rightarrow \ \breve{f}(x \,|\, \mu) \ \propto \ \max_{\nu} \Big\{ f(x \,|\, \mu, \nu)\, g(y \,|\, \nu) \Big\} \tag{43}$$

The profiled pdf is then used in a Neyman construction.

Note that the coverage of the simpler solutions is not guaranteed! However, if necessary it is sometimes possible to "recalibrate" these methods in such a way that coverage *is* achieved. Recall the Neyman construction of a

γ-level confidence interval:

$$C_\gamma(x_{obs}) \;=\; \Big\{\mu :\; x_{obs} \in S_\gamma(\mu)\Big\}, \tag{44}$$

where $S_\gamma(\mu)$ is a subset of sample space that satisfies:

$$\mathbb{P}_{\mu,\nu}\Big[X \in S_\gamma(\mu)\Big] \;\geq\; \gamma \quad \text{for all } \mu \text{ and } \nu, \tag{45}$$

or equivalently:

$$\min_\nu \mathbb{P}_{\mu,\nu}\Big[X \in S_\gamma(\mu)\Big] \;\geq\; \gamma \quad \text{for all } \mu. \tag{46}$$

In the recalibrated profile likelihood method, one sets:

$$S_\gamma(\mu) \;=\; \Big\{x :\; \lambda_\mu(x) \equiv \frac{\mathcal{L}(\mu, \hat\nu_\mu(x)\,|\,x)}{\mathcal{L}(\hat\mu(x), \hat\nu(x)\,|\,x)} \;\geq\; c_\gamma(\mu)\Big\}, \tag{47}$$

where $\hat\nu_\mu(x)$ maximizes $\mathcal{L}(\mu,\nu\,|\,x)$ for given μ and x, and $(\hat\mu(x), \hat\nu(x))$ maximizes $\mathcal{L}(\mu,\nu\,|\,x)$ for given x. For each μ one adjusts $c_\gamma(\mu)$ to satisfy (46).

4.3. *Other Frequentist Interval Construction Methods*

In practice, a popular method for constructing intervals is via *test inversion*. Suppose we are interested in some parameter $\theta \in \Theta$, and that for each allowed value θ_0 of θ we can construct an exact p value to test $H_0 : \theta = \theta_0$. We then have a family $\{p_\theta\}$ of p values indexed by the θ value of the corresponding test, and we can use this family to construct one- and two-sided γ confidence-level intervals for θ:

$$C_{1\gamma} \;=\; \Big\{\theta :\; p_\theta \geq 1-\gamma\Big\} \quad \text{and} \quad C_{2\gamma} \;=\; \Big\{\theta :\; \frac{1-\gamma}{2} \leq p_\theta \leq \frac{1+\gamma}{2}\Big\}. \tag{48}$$

To describe the one-sided construction for example, one would say that a γ confidence limit for θ is obtained by collecting all the θ values that are not rejected at the $1-\gamma$ significance level by the p value test. Indeed:

$$\begin{aligned}\mathbb{P}[\theta_{\rm true} \in C_{1\gamma}] \;&=\; \mathbb{P}[p_{\theta_{\rm true}} \geq 1-\gamma] \;=\; 1 - \mathbb{P}[p_{\theta_{\rm true}} < 1-\gamma] \\ &=\; 1-(1-\gamma) \;=\; \gamma, \end{aligned} \tag{49}$$

where the first equality follows from the definition of $C_{1\gamma}$ and the third one from the uniformity of p values under the tested hypothesis.

In general one can expect the properties of a family of p values to be reflected in the properties of the resulting family of intervals. A conservative p value will lead to conservative intervals, and a powerful p value will result in short intervals.

Another popular interval construction method is based on the likelihood function.[29] In one dimension, an approximate 68% confidence interval can be obtained by collecting all the parameter values for which the log-likelihood is within half a unit from its maximum. The validity of this approximation tends to increase with sample size.

Finally, another method explored by statisticians is based on objective Bayesian ideas. Objective priors can be designed in such a way that the resulting posterior intervals have a frequentist coverage that matches their Bayesian credibility to some order in $1/\sqrt{n}$, n being the sample size. When there are no nuisance parameters and the parameter of interest is one-dimensional, the matching prior to $\mathcal{O}(1/n)$ for one-sided intervals is Jeffreys' prior (5). Results are harder to come by in higher dimensions, but it is believed that reference analysis offers the best hope.[13] A major advantage of this approach is that it automatically yields intervals with Bayesian credibility, meaning intervals that are relevant for the actually observed data.

4.4. *Bayesian Interval Constructions*

As emphasized in section 2.2, the output of a Bayesian analysis is *always* the complete posterior distribution for the parameter(s) of interest. However, it is often useful to summarize the posterior by quoting a region with a given probability content. Such a region can be an interval or a union of intervals. Several schemes are available:

- Highest probability density regions;
  Any parameter value inside such a region has a higher posterior probability density than any parameter value outside the region, guaranteeing that the region will have the smallest possible length (or volume). Unfortunately this construction is not invariant under reparametrizations, and there are examples where this lack of invariance results in zero coverage for a subset of parameter values (of course this would only be of concern to a frequentist or an objective Bayesian).
- Central intervals;
  These are intervals that are symmetric around the median of the posterior distribution. For example, a 68% central interval extends from the 16^{th} to the 84^{th} percentiles. Central intervals are parametrization invariant, but they can only be defined for one-dimensional parameters. Furthermore, if a parameter is constrained to be non-negative, a central interval will usually not include the value zero; this may be problematic

if zero is a value of special physical significance.
- Upper and lower limits;
  For one-dimensional posterior distributions, these one-sided intervals can be defined using percentiles.
- Likelihood regions;
  These are standard likelihood regions where the likelihood ratio between the region boundary and the likelihood maximum is adjusted to obtain the desired posterior credibility. Such regions are metric independent and robust with respect to the choice of prior. In one-dimensional problems with physical boundaries and unimodal likelihoods, this construction yields intervals that smoothly transition from one-sided to two-sided.
- Intrinsic credible regions;
  These are regions of parameter values with minimum reference posterior expected loss[30] (a concept from Bayesian reference analysis).

High energy physicists using Bayesian procedures are generally advised to check the sensitivity of their result to the choice of prior, and its behavior under repeated sampling (coverage).

4.5. *Examples of Interval Constructions*

The effect of a physical boundary on frequentist and Bayesian interval constructions is illustrated in Figures 8 and 9 for the measurement of the mean μ of a Gaussian with unit standard deviation. The mean μ is assumed to be positive. All intervals are based on a single observation x. In general intervals have many properties that are worth studying: here we only examine the Bayesian credibility of frequentist constructions and the frequentist coverage of Bayesian constructions.

Figure 8 shows only frequentist constructions; Feldman-Cousins intervals[31] use x as estimator of μ and are based on a likelihood ratio ordering rule, whereas Mandelkern-Schultz intervals[32] use $\max\{0, x\}$ as estimator of μ and are based on a central ordering rule. The central, Feldman-Cousins, and upper limit confidence sets have very low credibility when the observation X is a large negative number. Mandelkern-Schultz intervals avoid this problem by reporting the same result for any negative X as for zero X, resulting in excess credibility at negative X.

Figure 9 shows central, highest posterior density, intrinsic, and upper limit Bayesian constructions, using Jeffreys' rule as prior for μ. They generally have good frequentist coverage, except near $\mu = 0$, where the curves

for central and intrinsic intervals dip to zero.

Note how frequentist coverage and Bayesian credibility always agree with each other when one is far enough from the physical boundary.

5. Search Procedures

Search procedures combine techniques from hypothesis testing and interval construction. The basic idea is to test a hypothesis about a new physics model, and then characterize the result of the test by computing point and interval estimates. We discuss both the frequentist and Bayesian approaches to this problem.

5.1. *Frequentist Search Procedures*

The standard frequentist procedure to search for new physics processes is as follows:

(1) Calculate a p value to test the null hypothesis that the data were generated by standard model processes alone.
(2) If $p \leq \alpha_1$ claim discovery and calculate a two-sided, α_2 confidence level interval on the production cross section of the new process.
(3) If $p > \alpha_1$ calculate an α_3 confidence level upper limit on the production cross section of the new process.

Typical confidence levels are $\alpha_1 = 2.9 \times 10^{-7}$, $\alpha_2 = 0.68$, and $\alpha_3 = 0.95$.

There are a couple of issues regarding this procedure. The first one is coverage: since the procedure involves one p value and two confidence intervals, an immediate question concerns the proper frequentist reference ensemble for each of these objects. The second issue arises when one fails to claim a discovery and calculates an upper limit. The stated purpose of this limit is to exclude cross sections that the experiment is sensitive to and did not detect. How then does one avoid excluding cross sections that the experiment is *not* sensitive to? We take a closer look at these two issues in the following subsections.

5.1.1. *The Coverage Issue*

In a 1998 paper on frequentist interval constructions,[31] Feldman and Cousins characterize as *flip-flopping* the procedure by which some experimenters decide whether to report an upper limit or a two-sided interval on a physics parameter that is constrained to be non-negative (such as a

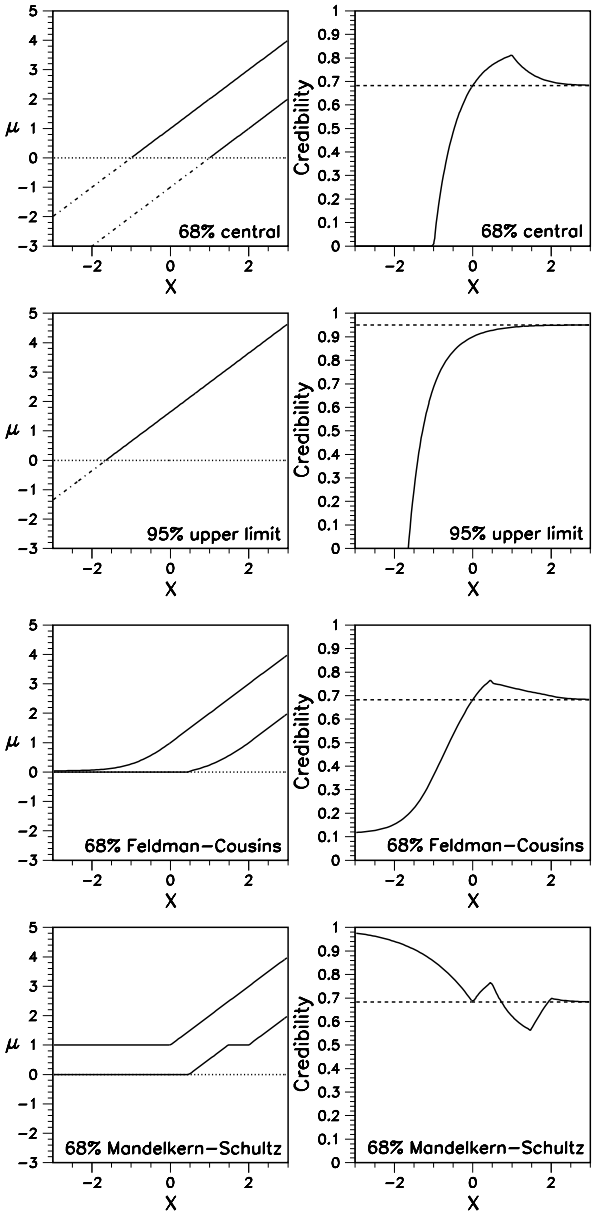

Fig. 8. Frequentist interval constructions. Left: graphs of μ versus X. Right: Bayesian credibility levels based on Jeffreys' prior; dashed lines indicate the frequentist coverage.

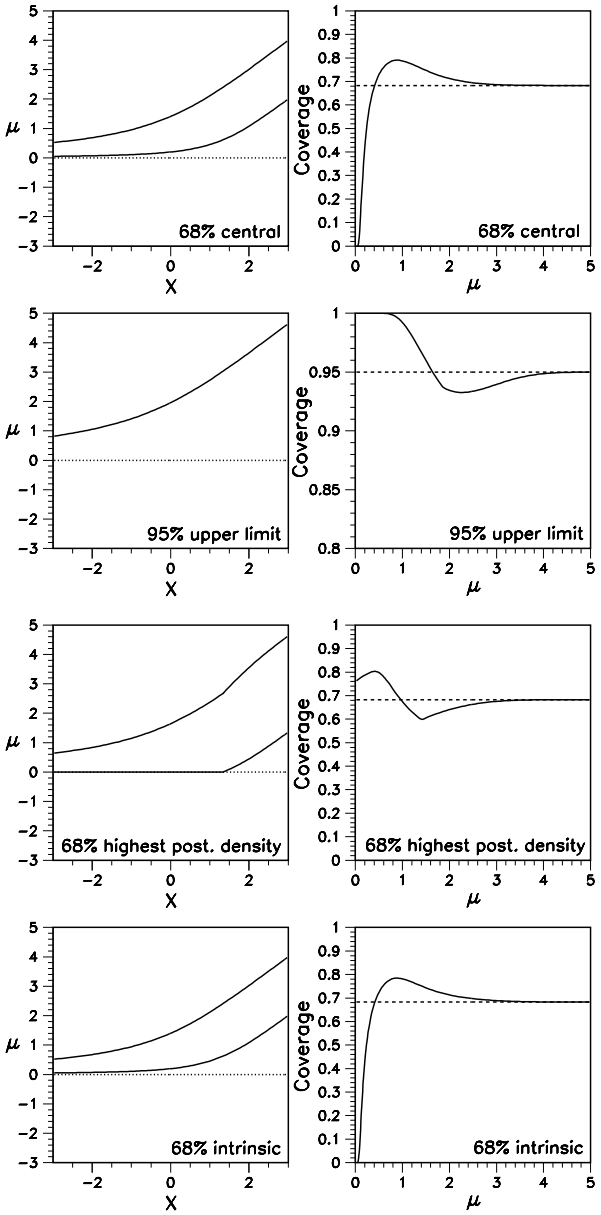

Fig. 9. Bayesian interval constructions. Left: graphs of μ versus X. Right: frequentist coverage levels; dashed lines indicate the Bayesian credibility.

mass or a mean event rate). In flip-flopping, this decision is based on first inspecting the data; an upper limit is then reported if the data is less than 3σ away from the physical boundary, and a two-sided interval otherwise. Because the initial data inspection is not taken into account in constructing the intervals, the flip-flopping procedure undercovers and is therefore invalid from a frequentist point of view.

In the frequentist search procedure just described, the decision to calculate a two-sided interval or an upper limit is based on the significance of the observed data with respect to the null hypothesis, a clear case of flip-flopping. One way to solve this problem would be to construct a Feldman-Cousins interval, since the latter transitions automatically from a one-sided to a two-sided interval as the data exhibits increasing evidence against the null hypothesis (see Fig. 8, top right). Unfortunately the Feldman-Cousins construction requires $\alpha_1 = \alpha_2 = \alpha_3$; this is unsatisfactory because it leads either to intervals that are too wide or test levels that are too low.

Another possibility is to construct *conditional* frequentist intervals: If $p \leq \alpha_1$, calculate a two-sided α_2 confidence level interval conditional on the observation that $p \leq \alpha_1$; otherwise, calculate an α_3 confidence level upper limit conditional on the observation that $p > \alpha_1$. What this means practically, in terms of the Neyman construction of each interval, is that the estimator X along the horizontal axis must be constrained to live within the region of sample space selected by the test, i.e. $p \leq \alpha_1$ or $p > \alpha_1$. The distribution of X must be appropriately truncated and renormalized in each case. An example of such a construction is shown in Fig. 10, for a simple search that involves testing whether the mean μ of a Gaussian distribution is zero (the null hypothesis) or greater than zero (the alternative). The data consists of a single sample X from that Gaussian, and can be negative or positive. The Gaussian width is assumed known. The plot shows that the conditional upper limit diverges as the discovery threshold is approached from the left, indicating that, so close to discovery, it becomes impossible to exclude *any* non-zero value of μ. On the other hand, as the threshold is approached from the right, the conditional two-sided interval turns into an upper limit, indicating that, so close to failing to make a discovery, it is possible that the true value of μ is zero and that the observed effect is just a background fluctuation. Note that a likelihood-ratio ordering rule was used here, in order to avoid creating a region of X values for which the reported μ interval is empty. For a central ordering rule for example, a small such region appears just above the discovery threshold.

In general physicists using the frequentist search procedure don't bother

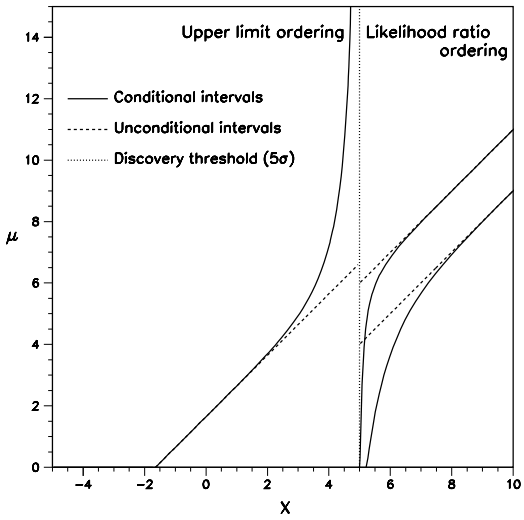

Fig. 10. Neyman construction of conditional intervals (solid lines) for the positive mean μ of a Gaussian, after having tested at the 5σ level whether $\mu = 0$. On the left of the discovery threshold, a 95% confidence level upper limit is shown, and on the right a 68% confidence level interval. Dashed lines indicate the corresponding unconditional intervals.

with the conditional construction. It is a more complicated calculation, and in any case its results coincide with those of the unconditional construction if one is far enough from the rejection threshold. Presumably one could argue that eventually, as more data are collected, one *will* be far enough.

So far our discussion of frequentist search procedures is based on a strict error-rate interpretation of measurement results. An alternative approach, not widely known in HEP, is to adopt an evidential interpretation.[33] This approach is centered around the p value, and the reported intervals serve to quantify the actual severity with which the hypothesis test has probed deviations from the null hypothesis. Suppose for example that we are testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$. If the p value against H_0 is not small, this is regarded as evidence that the true value of μ must be less than $\mu_0 + \delta$ for some δ. Thus one may examine the probability $\beta(\delta)$ of observing a worse fit of the data if the true value of μ is $\mu_0 + \delta$. If that probability is near one, the data are good evidence that $\mu < \mu_0 + \delta$. What physicists do in practice is to solve $\beta(\delta) = \alpha_3$ for δ, and report that all values of μ above $\mu_0 + \delta$ are excluded at the $1 - \alpha_3$ confidence level. A similar reasoning can

be followed to justify the reporting of a two-sided interval for μ when the p value against H_0 is small.

5.1.2. *The Sensitivity Issue*

Suppose the result of a test of H_0 is that it can't be rejected: we find $p_0 > \alpha_1$, where the subscript 0 on the p value emphasizes that it is calculated *under the null hypothesis*. A natural question is then: what values of the new physics cross section μ can we actually exclude? This is answered by calculating an α_3 C.L. upper limit on that cross section, and the easiest way to do this is by inverting a p value test: exclude all μ values for which $p_1(\mu) \leq 1 - \alpha_3$, where $p_1(\mu)$ is the p value under the alternative hypothesis that μ is the true value.

If our measurement has little or no sensitivity for a particular value of μ, this means that the distribution of the test statistic is (almost) the same under H_0 and H_1. In this case $p_0 \sim 1 - p_1$, and under H_0 we have:

$$\begin{aligned}\mathbb{P}_0(p_1 \leq 1 - \alpha_3) \;\sim\; \mathbb{P}_0(1 - p_0 \leq 1 - \alpha_3) \;=\; \mathbb{P}_0(p_0 \geq \alpha_3) \\ = \; 1 - \mathbb{P}_0(p_0 < \alpha_3) \;=\; 1 - \alpha_3. \quad (50)\end{aligned}$$

For example, if we calculate a 95% C.L. upper limit, there will be a $\sim 5\%$ probability that we will be able to exclude μ values for which we have no sensitivity. Some experimentalists consider that 5% is too much; to avoid this problem they only exclude μ values for which

$$\frac{p_1(\mu)}{1 - p_0} \leq 1 - \alpha_3. \quad (51)$$

For historical reasons, the ratio of p values on the left-hand side is known as CL_s. The resulting upper limit procedure *over*covers.

It is often useful to examine plots of p_1 versus p_0 for a given experimental resolution.[34] If $F_i(x)$ is the cumulative distribution function of the test statistic X under H_i, then we have $p_1 = F_1(x)$ and $p_0 = 1 - F_0(x)$ (assuming that large values of X are evidence against H_0). Hence, $p_1 = F_1[F_0^{-1}(1 - p_0)]$. This is illustrated in Fig. 11 for the simple case where $F_i(x)$ is Gaussian with mean μ_i and known width σ. The horizontal dashed line in the plot is the standard frequentist exclusion threshold: any μ value for which $p_1(\mu)$ is below that line will be excluded at the α_3 confidence level. In the lower right-hand corner of the plot, one sees that even for experiments with no resolution ($\Delta\mu/\sigma = 0$) p_1 can dip below the horizontal line, leading to the rejection of some values of μ. This is avoided by the CLs procedure (51), represented by the slanted line of dots. Interestingly, Bayesian upper limits

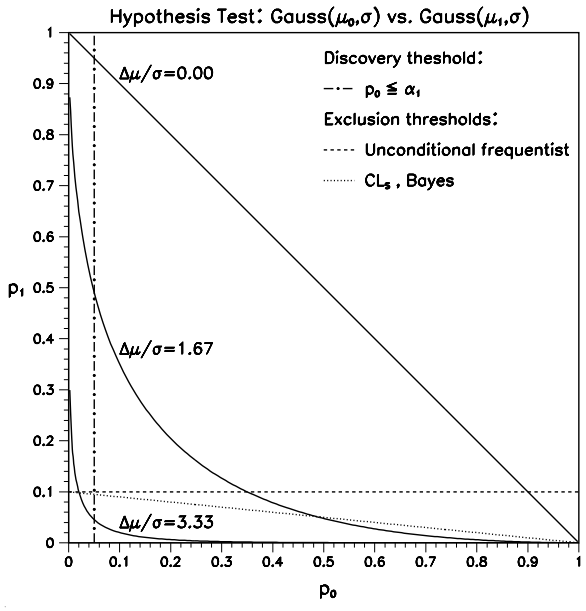

Fig. 11. Plot of p_1 versus p_0 in a test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, where μ is the mean of a Gaussian of known width σ. The experimental resolution is $\Delta\mu/\sigma$, with $\Delta\mu = |\mu_1 - \mu_0|$.

coincide with CLs limits for this problem. As the measurement resolution $\Delta\mu/\sigma$ increases, the corresponding p_1 versus p_0 contour approaches the lower left-hand corner of the plot, with the result that the probability of rejecting a false H_0 increases, and conversely, the probability of excluding a given μ value if H_0 is true also increases.

This last observation provides an interesting way to quantify *a priori* the sensitivity of a search procedure when the new physics model depends on a parameter μ, namely by reporting the set S of μ values for which

$$1 - \beta(\alpha_1, \mu) \;\geq\; \alpha_3, \tag{52}$$

where $\beta(\alpha_1, \mu)$ is the frequentist Type-II error rate corresponding to a discovery threshold α_1 and a value μ for the parameter under the alternative hypothesis. The set S has a couple of valuable interpretations:[35]

(1) If the true value of μ belongs to S, the probability of making a discovery is at least α_3, by definition of β.
(2) If the test does not result in discovery, it will be possible to exclude *at least* the entire sensitivity set with confidence α_3. Indeed, if we fail

to reject H_0 at the α_1 level, then we can reject any μ in H_1 at the $\beta(\alpha_1, \mu)$ level, so that $p_1(\mu) \leq \beta(\alpha_1, \mu)$; furthermore, if $\mu \in S$, then $\beta(\alpha_1, \mu) \leq 1 - \alpha_3$ and therefore $p_1(\mu) \leq 1 - \alpha_3$, meaning that μ is excluded with confidence α_3.

In general the sensitivity set depends on the event selection and the choice of test statistic. Maximizing the size of the sensitivity set provides a criterion for optimizing the event selection and choice of test statistic. The appeal of this criterion is that it optimizes the result regardless of the outcome of the test.

5.2. *Bayesian Search Procedures*

The starting point of a Bayesian search is the calculation of a Bayes factor. For a test of the form $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$, this can be written as:

$$B_{01}(x) \;=\; \frac{p(x \,|\, \theta_0)}{\displaystyle\int p(x \,|\, \theta, H_1)\, \pi(\theta \,|\, H_1)\, d\theta}, \tag{53}$$

and points to an immediate problem: what is an appropriate prior $\pi(\theta \,|\, H_1)$ for θ under the alternative hypothesis? Ideally one would be able to elicit some kind of proper "consensus" prior representing scientific knowledge prior to the experiment. If this is not possible, one might want to use an "off the rack" objective prior, but such priors are typically *improper*, and therefore only defined up to a multiplicative constant, rendering the Bayes factor totally useless.

A possible objective solution is to use the so-called *intrinsic* or *expected posterior* prior construction:[25]

- Let $\pi^O(\theta)$ be a good estimation objective prior (for example a reference prior), and $\pi^O(\theta \,|\, x)$ the corresponding posterior.
- Then the intrinsic prior is

$$\pi^I(\theta) \equiv \int \pi^O(\theta \,|\, y)\, p(y \,|\, \theta_0)\, dy, \tag{54}$$

  where $p(y \,|\, \theta_0)$ is the pdf of the data under H_0. The dimension of y (the sample size) should be the smallest one for which the posterior $\pi^O(\theta \,|\, y)$ is well defined.

The idea is that if we were given separate data y, we would compute the posterior $\pi^O(\theta \,|\, y)$ and use it as a proper prior for the test. Since we are

not given such data, we simply compute an average prior over all possible data.

In addition to the Bayes factor we need prior probabilities for the hypotheses themselves. An "objective" choice is the impartial $\pi(H_0) = \pi(H_1) = 1/2$. The posterior probability of H_0 is then

$$\pi(H_0 \,|\, x) \;=\; \frac{B_{01}}{1 + B_{01}}, \tag{55}$$

and the complete outcome of the search is this probability $\pi(H_0 \,|\, x)$, plus the posterior distribution of θ under the alternative hypothesis, $\pi(\theta \,|\, x, H_1)$. Often it will be useful to summarize the posterior distribution of θ under H_1 by calculating an upper limit or a two-sided interval.

6. Systematic Uncertainties

Although we have mentioned systematic uncertainties in our treatment of nuisance parameters in section 3.3, they deserve some additional remarks in a separate section. To begin, systematic uncertainties should be distinguished from statistical uncertainties, which are due to random fluctuations resulting from the finite size of the data sample. Systematic uncertainties are associated with the measuring apparatus, assumptions made by the experimenter, and the model used to draw inferences. Whereas statistical uncertainties from different samples are independent, this is not usually the case with systematics, which tend to be correlated across samples.

One can distinguish three types of systematic uncertainties:[36]

(1) Systematics that can be constrained by ancillary measurements and can therefore be treated as statistical uncertainties. As example, consider the measurement of the mass of the top quark in a $t\bar{t}$ channel where at least one top quark decays hadronically, i.e. $t \to Wb \to j_1 j_2 b$, where j_1 and j_2 are light-quark jets; since these come from the decay of the W, the known W mass can be used to constrain the jet energy scale.
(2) Systematics that cannot be constrained by existing data and are due to poorly understood features of the model used to draw inferences. Here, examples include background composition and shape, gluon radiation, higher-order corrections, and fragmentation parameters.
(3) Sources of uncertainty not easily modeled in a standard probabilistic setup, such as unknown experimenter bias.

In general a measurement result f is affected by several systematic uncertainties simultaneously. Assuming that these are all Type-2 systematics

and that we adopt a Bayesian framework, we can find a prior $\pi(\mu, \nu, \ldots)$ for the corresponding nuisance parameters. The variance of f due to these uncertainties is then:

$$V[f] \;=\; \int \Big[f(\mu, \nu, \ldots) - f(\mu_0, \nu_0, \ldots) \Big]^2 \, \pi(\mu, \nu, \ldots) \, d\mu \, d\nu \ldots, \qquad (56)$$

where μ_0, ν_0, $\ldots$, are the nominal values of the nuisance parameters. In HEP however, we usually quantify the effect of these systematics on f by summing independent variations in quadrature:

$$S^2 \;=\; \Big[f(\mu_0 + \sigma_\mu, \nu_0, \ldots) - f(\mu_0, \nu_0, \ldots) \Big]^2 + \Big[f(\mu_0, \nu_0 + \sigma_\nu, \ldots) - f(\mu_0, \nu_0, \ldots) \Big]^2 + \ldots \qquad (57)$$

This procedure is called OFAT, for "One Factor At a Time", and only takes into account linear terms in the dependence of f on the nuisance parameters. This may be a mistake, as there often are quadratic (μ^2, ν^2, $\ldots$), mixed ($\mu\nu$), and even higher order terms that should be included in the calculation of the variance of f.

Techniques exist to estimate these higher-order effects by order of importance — this is called DOE, for "Design Of Experiments". The idea is to vary several systematics simultaneously instead of just one by one. DOE techniques are not much used in current experimental high energy physics. However, it is believed that these are valuable ideas that should be kept in mind as the complexity of data analyses continues to increase.[37,38]

References

1. The BABAR statistics committee web page: `http://www.slac.stanford.edu/BFROOT/www/Statistics`
2. The CDF statistics committee web page: `http://www-cdf.fnal.gov/physics/statistics/statistics_home.html`
3. The CMS statistics committee web page: `https://twiki.cern.ch/twiki/bin/view/CMS/StatisticsCommittee`
4. Workshop on *Confidence Limits*, (CERN, Geneva, Switzerland, 2000); `http://doc.cern.ch/cernrep/2000/2000-005/2000-005.html`.
5. Workshop on *Confidence Limits*, (Fermilab, Batavia, Illinois, 2000); `http://conferences.fnal.gov/cl2k/`.
6. Conference on *Advanced Statistical Techniques in Particle Physics*, (University of Durham, UK, 2002); `http://www.ippp.dur.ac.uk/old/Workshops/02/statistics/`.
7. Conference on *Statistical Problems in Particle Physics, Astrophysics and Cosmology (PHYSTAT2003)*, (Stanford Linear Accelerator Center, Stanford, California, 2003); `http://www-conf.slac.stanford.edu/phystat2003/`.

8. Conference on *Statistical Problems in Particle Physics, Astrophysics and Cosmology (PHYSTAT05)*, (Oxford, UK, 2005); `http://www.physics.ox.ac.uk/phystat05/index.htm`.
9. PHYSTAT LHC Workshop on *Statistical Issues for LHC Physics*, (CERN, Geneva, Switzerland, 2007); `http://phystat-lhc.web.cern.ch/phystat-lhc/`.
10. D. M. Appleby, "Probabilities are single-case, or nothing," arXiv:quant-ph/0408058v1 (8 Aug 2004).
11. C. M. Caves, C. A. Fuchs, and R. Schack, "Subjective probability and quantum certainty," arXiv:quant-ph/0608190v2 (26 Jan 2007).
12. E. T. Jaynes, "Probability theory: the logic of science," (ed. by G. L. Bretthorst), Cambridge University Press, 2003 (727pp).
13. J. M. Bernardo and A. F. M. Smith, "Bayesian theory," John Wiley & Sons, 1994 (586pp).
14. F. James, "Introduction and statement of the problem," in *Proceedings of the 1st workshop on confidence limits, 17–18 January 2000, CERN, Geneva, Switzerland*, L. Lyons, Y. Perrin, F. James, eds., CERN Yellow Report cernrep/2000-005.
15. C. M. Caves, C. A. Fuchs, and R. Schack, "Quantum probabilities as Bayesian probabilities," Phys. Rev. A **65**, 022305 (2002).
16. E. T. Jaynes, "Probability in quantum theory," `http://bayes.wustl.edu/etj/articles/prob.in.qm.pdf` (1990).
17. R. L. Jaffe, "The Casimir effect and the quantum vacuum," Phys. Rev. D **72**, 021301 (2005).
18. A. H. Rosenfeld, "Are there any far-out mesons or baryons?," in *Meson Spectroscopy. A collection of articles*, C. Baltay and A. H. Rosenfeld, eds., W.A. Benjamin, Inc., New York, Amsterdam, 1968, pg. 455.
19. S. Stepanyan *et al.* (CLAS Collaboration), "Observation of an exotic $S = +1$ baryon in exclusive photoproduction from the deuteron," Phys. Rev. Lett. **91**, 252001 (2003); B. McKinnon *et al.* (CLAS Collaboration), "Search for the Θ^+ pentaquark in the reaction $\gamma d \to pK^-K^+n$," Phys. Rev. Lett. **96**, 212001 (2006).
20. A. Roodman, "Blind analysis in particle physics," in *Proceedings of the PhyStat2003 Conference*, SLAC, Stanford, California, September 8–11, 2003, pg. 166; also at arXiv:physics/0312102v1 (17 Dec 2003).
21. M. J. Bayarri and J. O. Berger, "P-values for composite null models [with discussion]," J. Amer. Statist. Assoc. **95**, 1127 (2000).
22. E. L. Lehmann, "On likelihood ratio tests," arXiv:math/0610835v1 [math.ST] (27 Oct 2006).
23. C. Goutis, G. Casella, and M. T. Wells, "Assessing evidence in multiple hypotheses," J. Amer. Statist. Assoc. **91**, 1268 (1996).
24. R. D. Cousins, "Annotated bibliography of some papers on combining significances or p-values," arXiv:0705.2209v1 [physics.data-an] (15 May 2007).
25. J. Berger, "A comparison of testing methodologies," CERN Yellow Report CERN-2008-001, pg. 8; see `http://phystat-lhc.web.cern.ch/phystat-lhc/proceedings.html`

26. R. E. Kass and A. E. Raftery, "Bayes factors," J. Amer. Statist. Assoc. **90**, 773 (1995).
27. D. R. Cox, "Some problems connected with statistical inference," Ann. Math. Statist. **29**, 357 (1958).
28. J. V. Bondar, Discussion of "Conditionally acceptable frequentist solutions," by G. Casella, in "Statistical decision theory and related topics IV," Vol. 1, S. S. Gupta and J. O. Berger, eds., Springer-Verlag 1988, pg. 91.
29. F. James, "Statistical methods in experimental physics," 2nd ed., World Scientific Publishing Co., 2006 (345pp).
30. J. Bernardo, "Intrinsic credible regions: an objective Bayesian approach to interval estimation," Test **14**, 317 (2005); see also `http://www.uv.es/~bernardo/2005Test.pdf`.
31. G. J. Feldman and R. D. Cousins, "Unified approach to the classical statistical analysis of small signals," Phys. Rev. D **57**, 3873 (1998).
32. M. Mandelkern and J. Schultz, "The statistical analysis of Gaussian and Poisson signals near physical boundaries," J. Math. Phys. **41**, 5701 (2000).
33. D. G. Mayo and D. R. Cox, "Frequentist statistics as a theory of inductive inference," IMS Lecture Notes — Monograph Series: 2nd Lehmann Symposium — Optimality, Vol. 49, pg. 77-97 (2006); arXiv:math/0610846v1 [math.ST] (27 Oct 2006); see also D. G. Mayo's comment on J. O. Berger, "Could Fisher, Jeffreys and Neyman have agreed on testing?," Statist. Science **18**, 1 (2003).
34. The suggestion to study plots of p_1 versus p_0 was made by Louis Lyons.
35. G. Punzi, "Sensitivity of searches for new signals and its optimization," in *Proceedings of the PHYSTAT2003 Conference*, SLAC, Stanford, California, September 8–11, 2003, pg. 79; also at arXiv:physics/0308063v2 (4 Dec 2003).
36. P. Sinervo, "Definition and treatment of systematic uncertainties in high energy physics and astrophysics," in *Proceedings of the PHYSTAT2003 Conference*, SLAC, Stanford, California, September 8–11, 2003, pg. 122.
37. J. T. Linnemann, "A pitfall in evaluating systematic errors," CERN Yellow Report CERN-2008-001, pg. 94.
38. N. Reid, "Some aspects of design of experiments," CERN Yellow Report CERN-2008-001, pg. 99.