Statistics for Searches at the LHC

Glen Cowan

Abstract This chapter describes several topics in statistical data analysis as used in High Energy Physics. It focuses on areas most relevant to analyses at the LHC that search for new physical phenomena, including statistical tests for discovery and exclusion limits. Particular attention is payed to the treatment of systematic uncertainties through nuisance parameters.

1 Introduction

The primary goal of data analysis in High Energy Physics (HEP) is to test our current understanding of particle interactions and in doing so to search for phenomena that go beyond the existing framework of the Standard Model. These lectures describe some of the basic statistical tools that alone one to do this.

Despite efforts to make the lectures self contained, some familiarity with basic ideas of statistical data analysis is assumed. Introductions to the subject can be found, for example, in the reviews of the Particle Data Group [1] or in the texts [2, 3, 4, 5, 6].

Brief reviews of probability are given in Sec. 2 and frequentist hypothesis tests in Secs. 3 and 4. These are applied to establishing discovery and setting limits (Sec. 5) and are extended using the profile likelihood ratio (Sec. 6), from which one can construct unified intervals (Sec. 7). Bayesian limits are discussed in Sec. 8 and all of the methods for limits are illustrated using the example of a Poisson counting experiment in Sec. 9. Application of the standard tools for discovery and limits leads to a number of unexpected difficulties, such as exclusion of models to which one has no sensitivity (Sec. 10) and the look-elsewhere effect (Sec. 11). In Sec. 12 we examine why one traditionally requires five-sigma significance to claim a discovery and fi-

Glen Cowan
Royal Holloway, University of London, Physics Department, Egham, Surrey, TW20 0EX, UK, e-mail: g.cowan@rhul.ac.uk

nally some conclusions are drawn in Sec. 13. The lectures presented at SUSSP also included material on unfolding or deconvolution of measured distributions which is not included here but can be found in Ref. [7] and Chapter 11 of Ref. [2].

2 Review of probability

When analyzing data in particle physics one invariably encounters uncertainty, at the very least coming from the intrinsically random nature of quantum mechanics. These uncertainties can be quantified using probability, which was defined by Kolmogorov [8] using the language of set theory. Suppose a set S contains elements that can form subsets A, B, ... As an example, the elements may represent possible outcomes of a measurement but here we are being abstract and we do not need to insist at this stage on a particular meaning. The three axioms of Kolmogorov can be stated as

1. For all $A \subset S$, there is a real-valued function P such that $P(A) \geq 0$;
2. $P(S) = 1$;
3. If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

In addition we define the conditional probability of A given B (for $P(B) \neq 0$) as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \,. \tag{1}$$

From these statements we can derive all of the familiar properties of probability. They do not, however, provide a complete recipe for assigning numerical values to probabilities nor do tell us what these values mean.

Of the possible ways to interpret a probability, the one most commonly found in the physical sciences is as a limiting frequency. That is, we interpret the elements of the sample space as possible outcomes of a measurement, and we take $P(A)$ to mean the fraction of times that the outcome is in the subset A in the limit where we repeat the measurement an infinite number of times under "identical" conditions:

$$P(A) = \lim_{n \to \infty} \frac{\text{times outcome is in } A}{n} \,. \tag{2}$$

Use of probability in this way leads to what is called the *frequentist* approach to statistics. Probabilities are only associated with outcomes of repeatable observations, not to hypothetical statements such as "supersymmetry is true". Such a statement is either true or false, and this will not change upon repetition of any experiment.

Whether SUSY is true or false is nevertheless uncertain and we can quantify this using probability as well. To define what is called *subjective probability* one interprets the elements of the set S as *hypotheses*, i.e., statements that are either true or false, and one defines

$$P(A) = \text{degree of belief that } A \text{ is true.} \tag{3}$$

Use of subjective probability leads to what is called *Bayesian statistics*, owing to its important use of Bayes' theorem described below.

Regardless of its interpretation, any quantity that satisfies the axioms of probability must obey Bayes' theorem, which states

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \; . \tag{4}$$

This can be derived from the definition of conditional probability (1), which we can write as $P(A \cap B) = P(B)P(A|B)$, or equivalently by changing labels as $P(B \cap A) = P(A)P(A|B)$. These two probabilities are equal, however, because $A \cap B$ and $B \cap A$ refer to the same subset. Equating them leads directly to Eq (4).

In Bayes' theorem (4) the condition B represents a restriction imposed on the sample space S such that anything outside of B is not considered. If the sample space S can be expressed as the union of some disjoint subsets A_i, $i = 1, 2, \ldots$, then the factor $P(B)$ appearing in the denominator can be written $P(B) = \sum_i P(B|A_i)P(A_i)$ so that Bayes' theorem takes on the form

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} \; . \tag{5}$$

In *Bayesian* (as opposed to frequentist) statistics, one uses subjective probability to describe one's degree of belief in a given theory or hypothesis. The denominator in Eq. (5) can be regarded as a constant of proportionality and therefore Bayes' theorem can be written as

$$P(\text{theory}|\text{data}) \propto P(\text{data}|\text{theory})P(\text{theory}) \; , \tag{6}$$

where "theory" represents some hypothesis and "data" is the outcome of the experiment. Here $P(\text{theory})$ is the *prior* probability for the theory, which reflects the experimenter's degree of belief before carrying out the measurement, and $P(\text{data}|\text{theory})$ is the probability to have gotten the data actually obtained, given the theory, which is also called the *likelihood*.

Bayesian statistics provides no fundamental rule for obtaining the prior probability; in general this is subjective and may depend on previous measurements, theoretical prejudices, etc. Once this has been specified, however, Eq. (6) tells how the probability for the theory must be modified in the light of the new data to give the *posterior* probability, $P(\text{theory}|\text{data})$. As Eq. (6) is stated as a proportionality, the probability must be normalized by summing (or integrating) over all possible hypotheses.

3 Hypothesis tests

One of the fundamental tasks in a statistical analysis is to test whether the predictions of a given model are in agreement with the observed data. Here we will use $\mathbf{x}$ to denote the outcome of a measurement; it could represent a single quantity or a collection of values. A hypothesis H means a statement for the probability to find the data $\mathbf{x}$ (or if $\mathbf{x}$ includes continuous variables, H specifies a probability density function or pdf). We will write $P(\mathbf{x}|H)$ for the probability to find data $\mathbf{x}$ under assumption of the hypothesis H.

Consider a hypothesis H_0 that we want to test (we will often call this the "null" hypothesis) and an alternative hypothesis H_1. In frequentist statistics one defines a *test* of H_0 by specifying a subset of the data space called the *critical region* w, such that the probability to observe the data there satisfies

$$P(\mathbf{x} \in w|H_0) \leq \alpha \, . \tag{7}$$

Here α is a specified small constant, such as 5%. For continuous data, one takes the relation above as an equality. If the data are discrete, such as a number of events, then there may not exist any subset of the data values whose summed probability is exactly equal to α, so one takes the critical region to have a probability up to α. The critical region w defines the test. If the data are observed in w, one rejects the hypothesis H_0.

Up to this point the sole defining property of the test is Eq. (7), which states that the probability to find the data in the critical region is not more than α. But there are in general many if not an infinite number of possible subsets of the data space that satisfy this criterion, and it is not clear which should be taken as the critical region. This is where the alternative hypothesis H_1 comes into play. One would like the critical region to be chosen such that there is as high a probability as possible to find the data there if the alternative is true, while having only the fixed probability α assuming H_0, as illustrated schematically in Fig. 1.

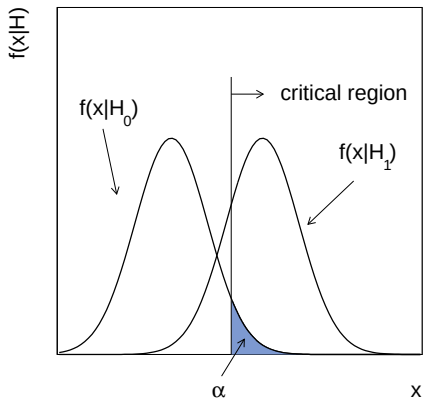

Fig. 1 Illustration of the critical region of a statistical test (see text).

Rejecting the hypothesis H_0 if it is in fact true is called a type I error. By construction the probability for this to occur is the size of the test, α. If on the other hand we do not reject H_0 but we should have, because the alternative H_1 was true, then this is called a type II error. The probability to reject the null hypothesis if the alternative H_1 is true is called the *power* of the test with respect to H_1, which is one minus the probability of a type II error.

A *significance test* of a hypothesis H is closely related to the tests described above. Suppose a measurement results in data $\mathbf{x}$ (a single number or a collection of many values) for which the hypothesis H predicts the probability $P(\mathbf{x}|H)$. We observe a single instance of $\mathbf{x}$, say, $\mathbf{x}_{\text{obs}}$, and we want to quantify the level of agreement between this outcome and the predictions of H.

To do this the analyst must specify what possible data values would constitute a level of incompatibility with H that is equal to are greater than that between H and the observed data $\mathbf{x}_{\text{obs}}$. Once this is given, then one computes the p-value of H as the probability, under assumption of H, to find data in this region of equal or greater incompatibility.

When computing the p-value there is clearly some ambiguity as to what data values constitute greater incompatibility with H than others. To give meaning to the statement that a given $\mathbf{x}$ has less compatibility with H, we must imply that it has more compatibility with some alternative hypothesis.

We should emphasize that the p-value is not the probability of H. To compute this, we would need to use Bayes' theorem (8), and this requires that we specify prior probabilities for all hypotheses, whereas the p-value does not depend on prior probabilities.

We can relate the two types of frequentist tests by specifying a critical region for a test of H_0 of size α as the set of data values that would have a p-value less than or equal to α. The resulting test will have a certain power with respect to any given alternative H_1, although these were not used explicitly in constructing the critical region.

In the language of a frequentist test, we *reject* H_0 if the data are found in the critical region, or equivalently, if the p-value of H_0 is found less or equal to α. Despite this language, it is not necessarily true that we would would then believe H_0 to be false. To make this assertion we should quantify our degree of belief about H_0 using subjective probability as described above, and it must be computed using Bayes' theorem:

$$P(H_0|\mathbf{x}) = \frac{P(\mathbf{x}|H_0)\pi(H_0)}{\sum_i P(\mathbf{x}|H_i)\pi(H_i)} \,. \tag{8}$$

As always, the posterior $P(H_0|\mathbf{x})$ is proportional to the prior $\pi(H_0)$, and this would need to be specified if we want to express our degree of belief that the hypothesis is true.

For most of these lectures we will stay within the frequentist framework. The result of our analysis will be a p-value for the different models considered. If this is less than some specified value α, we reject the model.

Often the p-value is translated into an equivalent quantity called the *significance* Z, defined by

$$Z = \Phi^{-1}(1-p)\,, \tag{9}$$

where Φ is the cumulative standard Gaussian distribution (zero mean, unit variance) and Φ^{-1} is its inverse function, also called the *quantile* of the standard Gaussian. The definition of significance illustrated in Fig. 2(a) and the significance versus p-value is shown in Fig. 2(b). Often a significance of $Z = 5$ is used as the threshold for claiming discovery of a new signal process. This corresponds to a very low p-value of 2.9×10^{-7}. The rationale for such an extreme threshold is discussed further in Sec. 12.

Although we can simply take Eq. (9) as our defining relation for Z, it is useful to compare to the case of measuring a quantity x that follows a Gaussian distribution with unknown mean μ. Suppose we want to test the hypothesis $\mu = 0$ against the alternative $\mu > 0$. In this case we would take the critical region of the test to contain values of x greater than a certain threshold, or equivalently, we would define the p-value to be the probability to find x as large as we found or larger. In this case the significance Z is simply the value of x observed, measured in units of its standard deviation σ. For this reason one often refers to finding a significance Z of, say, 2.0 as a *two*-sigma effect.

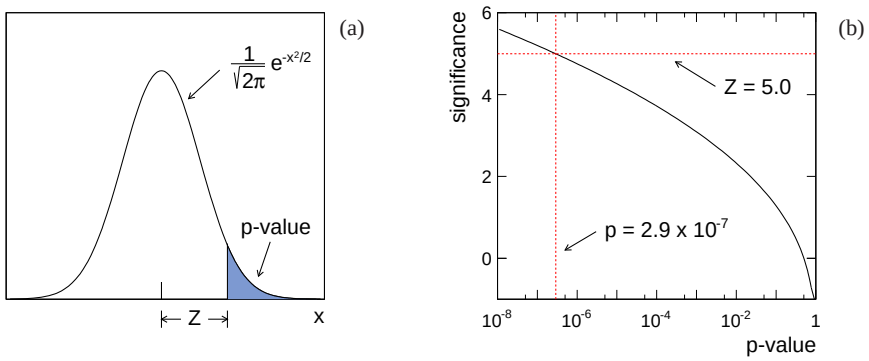

Fig. 2 (a) Illustration of the definition of significance Z and (b) the significance as function of the p-value.

4 Choice of critical region, test statistics

We now examine more closely the question of how best to define the critical region of a test and for this consider the example of selecting a sample of events of a desired

type (signal, denoted s) from others that we do not want (background, b). That is, for each event we will measure some set of quantities $\mathbf{x}$, which could represent different kinematic variables such as the missing energy, number of jets, number of muons, and so forth. Then for each event carry out a test of the background hypothesis, and if this is rejected it means we select the event as a candidate signal event.

Suppose that the s and b hypotheses imply probabilities for the data of $P(\mathbf{x}|s)$ and $P(\mathbf{x}|b)$, respectively. Figures (3) show these densities for two components of the data space along with possible boundaries for the critical region.

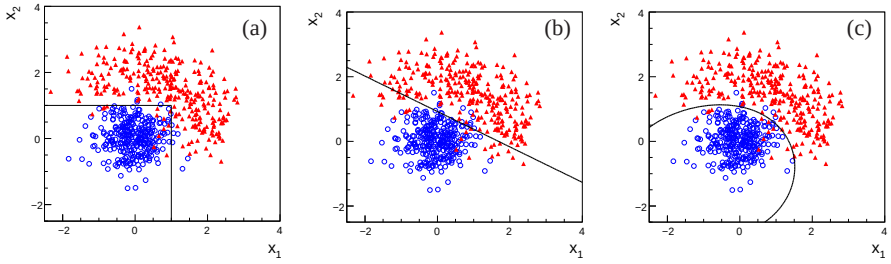

Fig. 3 Scatter plots of two variables corresponding to two hypotheses: background (H_0) and signal (H_1). The critical region for a test of H_0 could be based, e.g., on (a) cuts, (b) a linear boundary, (c) a nonlinear boundary.

Figure 3(a) represents what is commonly called the 'cut-based' approach. One selects signal events by requiring $x_1 < c_1$ and $x_2 < c_2$ for some suitably chosen cut values c_1 and c_2. If x_1 and x_2 represent quantities for which one has some intuitive understanding, then this can help guide one's choice of the cut values.

Another possible decision boundary is made with a diagonal cut as shown in Fig. 3(b). One can show that for certain problems a linear boundary has optimal properties, but in the example here, because of the curved nature of the distributions, neither the cut-based nor the linear solution is as good as the nonlinear boundary shown in Fig. 3(c).

The decision boundary is a surface in the n-dimensional space of input variables, which can be represented by an equation of the form $y(\mathbf{x}) = y_{\text{cut}}$, where y_{cut} is some constant. We accept events as corresponding to the signal hypothesis if they are on one side of the boundary, e.g., $y(\mathbf{x}) \leq y_{\text{cut}}$ could represent the acceptance region and $y(\mathbf{x}) > y_{\text{cut}}$ could be the rejection region.

Equivalently we can use the function $y(\mathbf{x})$ as a scalar *test statistic*. Once its functional form is specified, we can determine the pdfs of $y(\mathbf{x})$ under both the signal and background hypotheses, $p(y|s)$ and $p(y|b)$. The decision boundary is now effectively a single cut on the scalar variable y, as illustrated in Fig. 4.

We would like to design a test to have maximum reject a hypothesis if it is false, which is what we have called the power of the test. Unfortunately a test with maximum power with respect to one alternative will not be optimal with respect to others, so there is no such thing as an ideal "model-independent" test. Nevertheless, for a specific pair of signal and background hypotheses, it turns out that there is a well

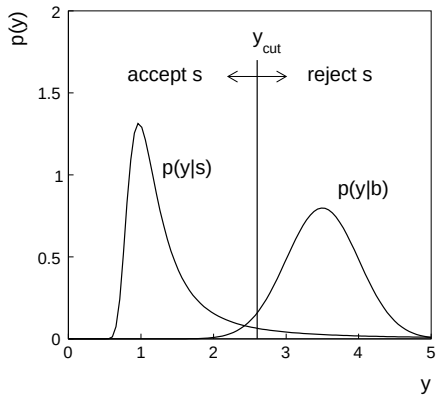

Fig. 4 Distributions of the scalar test statistic $y(\mathbf{x})$ under the signal and background hypotheses.

defined optimal solution to our problem. The *Neyman–Pearson* lemma states that one obtains the maximum power relative for the signal hypothesis for a given significance level (background efficiency) by defining the acceptance region such that, for $\mathbf{x}$ inside the region, the *likelihood ratio*, i.e., the ratio of pdfs for signal and background,

$$y(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)} \,, \tag{10}$$

is greater than or equal to a given constant, and it is less than this constant everywhere outside the acceptance region. This is equivalent to the statement that the ratio (10) represents the test statistic with which one obtains the highest signal efficiency for a given background efficiency, or equivalently, for a given signal purity.

In principle the signal and background theories should allow us to work out the required functions $f(\mathbf{x}|s)$ and $f(\mathbf{x}|b)$, but in practice the calculations are too difficult and we do not have explicit formulae for these. What we have instead of $f(\mathbf{x}|s)$ and $f(\mathbf{x}|b)$ are complicated Monte Carlo programs, that is, we can sample $\mathbf{x}$ to produce simulated signal and background events. Because of the multivariate nature of the data, where $\mathbf{x}$ may contain at least several or perhaps even hundreds of components, it is a nontrivial problem to construct a test with a power approaching that of the likelihood ratio.

In the usual case where the likelihood ratio (10) cannot be used explicitly, there exists a variety of other multivariate classifiers such as neural networks, boosted decision trees and support vector machines that effectively separate different types of events. Descriptions of these methods can be found, for example, in the textbooks [9, 10, 11, 12], lecture notes [13] and proceedings of the PHYSTAT conference series [14]. Software for HEP includes the `TMVA` [15] and `StatPatternRecognition` [16] packages.

5 Frequentist treatment of discovery and limits

The use of a statistical test to in an analysis involving different event types comes up in different ways. Sometimes both event classes are known to exist, and the goal is to select one class (signal) for further study. For example, proton–proton collisions leading to the production of top quarks are a well-established process. By selecting these events one can carry out precise measurements of the top quark's properties such as its mass. This was the basic picture in the previous section. The measured quantities referred to individual events, and we tested the the hypothesized event type for each.

In other cases, the signal process could represent an extension to the Standard Model, say, supersymmetry, whose existence is not yet established, and the goal of the analysis is to see if one can do this. Here we will imagine the "data" as representing not single events but a sample of events, i.e., an entire "experiment". If the signal process we are searching for does not exist, then our sample will consist entirely of background events, e.g., those due to Standard Model processes. If the signal does exist, then we will find both signal and background events. Thus the hypothesis we want to test is

$$H_0 \ : \ \text{only background processes exist}$$

versus the alternative

$$H_1 \ : \ \text{both signal and background exist.}$$

We will refer to the hypothesis H_0 as the background-only model (or simply "b") and the alternative H_1 as the signal-plus-background model, $s+b$. The Neyman-Pearson lemma still applies. In a test of H_0 of a given size, the highest power relative to H_1 is obtained when the critical region contains the highest values of the likelihood ratio $L(H_1)/L(H_0)$. Here, however, the likelihood is the probability for the entire set of data from the experiment, not just for individual events.

Rejecting H_0 means in effect discovering a new phenomenon. Of course before we believe that we have made a new discovery, a number of other concerns must be addressed, such as our confidence in the reliability of the statistical models used the plausibility of the new phenomenon and the degree to which it can describe the data. Here however we will simply focus on question of statistical significance and in effect equate "rejecting the background-only" hypothesis with "discovery". Often in HEP one claims discovery when the p-value of the background-only hypothesis is found below 2.9×10^{-7}, corresponding to a 5-sigma effect. We will revisit the rationale behind this threshold in Sec. 12.

Even if one fails to discover new physics by rejecting the background-only model, one can nevertheless test various signal models and see if these can be rejected. Signal models are usually characterized by some continuous parameters representing, e.g., the masses of new particles. If we carry out a test of size α for all possible values of the parameters, then those that are not rejected constitute what is

called a *confidence region* with a *confidence level* of $\mathrm{CL} = 1-\alpha$. By construction a parameter value will, if it is true, be rejected with probability α. Therefore the confidence region will contain the true value of the parameter with probability $1-\alpha$. For purposes of confidence limits one typically uses a test of size $\alpha = 0.05$, which is to say the regions have a confidence level of 95%.

If the problem has only one parameter, then the region is called a confidence interval. An important example is where a parameter μ is proportional to the cross section for the signal process being sought. Here one is interested in testing a hypothetical value relative to the alternative hypothesis that the signal does not exist, i.e., $\mu = 0$. The critical region of the test is then taken have higher probability for the lower values of the parameter.

For example, suppose the data consist of a value x that follows a Gaussian distribution with unknown mean μ and known standard deviation σ. If we test a value μ relative to the alternative of a smaller value, then the critical region will consist of values of $x < c$ for some constant c such that

$$\alpha = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx = \Phi\left(\frac{c-\mu}{\sigma}\right) , \tag{11}$$

or

$$c = \mu - \sigma\Phi^{-1}(1-\alpha) . \tag{12}$$

If we take, e.g., $\alpha = 0.05$, then the factor $\Phi^{-1}(1-\alpha) = 1.64$ says that the critical region starts at 1.64 standard deviations below the value of μ being tested. If x is observed any lower than this, then the corresponding μ is rejected.

Equivalently we can take the p-value of a hypothesized μ, p_μ, as the probability to observe x as low as we found or lower, and we then reject μ if we find $p_\mu < \alpha$. The highest value of μ that we do not reject is called the *upper limit* of μ at a confidence level of $1-\alpha$, and we will write this here as μ_{up} Lower limits μ_{lo} can of course be constructed using an analogous procedure. In practice these points are found by setting $p_\mu = \alpha$ and solving for μ. There are a number of subtle issues connected with limits derived in this way and we will return to these in Sec. 10.

5.1 A toy example

Consider an experiment in which we measure for each selected event two quantities, which we can write as a vector $\mathbf{x} = (x_1, x_2)$. Suppose that for background events $\mathbf{x}$ follows

$$f(\mathbf{x}|b) = \frac{1}{\xi_1} e^{-x/\xi_1} \frac{1}{\xi_2} e^{-x/\xi_2} , \tag{13}$$

and for a certain signal model they follow

$$f(\mathbf{x}|s) = C\frac{1}{\sqrt{2\pi}\sigma_1}e^{-(x_1-\mu_1)^2/2\sigma_1^2}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-(x_2-\mu_2)^2/2\sigma_2^2}\,. \tag{14}$$

where $x_1 \geq 0$, $x_2 \geq 0$ and C is a normalization constant. The distribution of events generated according to these hypotheses are shown in Fig. 5(a).

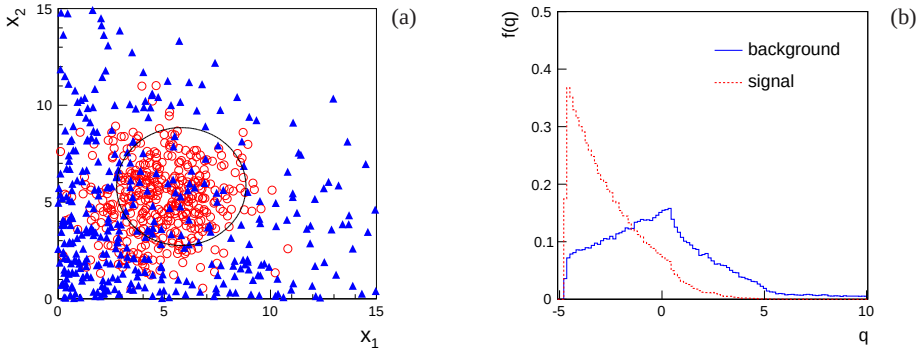

Fig. 5 (a) Distributions of $\mathbf{x} = (x_1, x_2)$ for events of type signal (red circles) and background (blue triangles) shown with a contour of constant likelihood ratio; (b) the distribution of the statistic q for signal and background events.

First, suppose that the signal and background both correspond to event types that are known to exist and the goal is simply to select signal. In this case we can exploit the Neyman-Pearson lemma and base the selection on the likelihood ratio

$$y(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}\,. \tag{15}$$

We can define the same critical region by using any monotonic function of the likelihood ratio, and in this case it is useful to take

$$q = \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - \frac{2x_1}{\xi_1} - 2\frac{2x_2}{\xi_2} = -2\ln y(\mathbf{x}) + \text{const.} \tag{16}$$

Distributions of the statistic q for the signal and background hypotheses (14) and (13) are shown in Fig. 5(b). This shows that a sample enhanced in signal events can be selected by selecting events with q less than a given threshold, say, q_{cut}.

Consider now a search for a signal process whose existence is not yet established. Suppose that the expected numbers events are b of background s for a given signal model. For now assume that the model's prediction for both of these quantities can be determined with negligible uncertainty. The the actual number of events n that we find can be modeled as a Poisson distributed quantity whose mean we can write

as $\mu s + b$, where μ is a parameter that specifies the strength of the signal process. That is, the probability to find n events is

$$P(n|\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \; . \tag{17}$$

The values of $\mathbf{x}$ follow a pdf that is a mixture of the two contributions from signal and background,

$$f(\mathbf{x}|\mu) = \frac{\mu s}{\mu s + b} f(\mathbf{x}|s) + \frac{b}{\mu s + b} f(\mathbf{x}|b) \; , \tag{18}$$

where the coefficients of each term give the fraction of events of each type.

The complete measurement thus consists of selecting n events and for each one measuring the two-dimensional vector quantity $\mathbf{x}$. The full likelihood is therefore

$$L(\mu) = P(n|\mu) \prod_{i=1}^{n} f(\mathbf{x}_i|\mu) = \frac{e^{-(\mu s + b)}}{n!} \prod_{i=1}^{n} [\mu s f(\mathbf{x}_i|s) + b f(\mathbf{x}_i|b)] \; . \tag{19}$$

We can now carry out tests of different hypothetical values of μ. To establish the existence of the signal process we try to reject the hypothesis of the background-only model, $\mu = 0$. Regardless of whether we claim discovery we can set limits on the signal strength μ, which we examine further in Sec. 10.

Let us first focus on the question of discovery, i.e., a test of $\mu = 0$. If the signal process exists, we would like to maximize the probability that we will discover it. This means that the test of the background-only ($\mu = 0$) hypothesis should have as high a power as possible relative to the alternative that includes signal ($\mu = 1$). According to the Neyman-Pearson lemma, the maximum power is achieved by basing the test on the likelihood ratio $L(1)/L(0)$, or equivalently on the statistic

$$Q = -2 \ln \frac{L(1)}{L(0)} = -s + \sum_{i=1}^{n} \ln \left(1 + \frac{s}{b} \frac{f(\mathbf{x}_i|s)}{f(\mathbf{x}_i|b)} \right) \; . \tag{20}$$

The term $-s$ in front of the sum is a constant and so only shifts the distribution of Q for both hypotheses equally; it can therefore be dropped.

Other than $-s$ on the right-hand side of Eq. (20) there is a sum of contributions from each event, and because the $\mathbf{x}$ values follow the same distribution for each event, each term in the sum follows the same distribution. To find the pdf of Q we can exploit the fact that the distribution of a sum of random variables is given by the convolution of their distributions. The full distribution can therefore be determined using Fourier transform techniques from the corresponding single-event distributions; details can be found in Ref. [17].

Following our toy example, suppose we take the expected numbers of events to be $b = 100$ for background and $s = 20$ for signal. The distribution of the statistic Q is found in this case simply by generating experiments according to the $\mu = 0$ and $\mu = 1$ hypotheses, computing for each Q according to Eq. (20) and recording the values in histograms. This results in the distributions shown in Fig. 6(a).

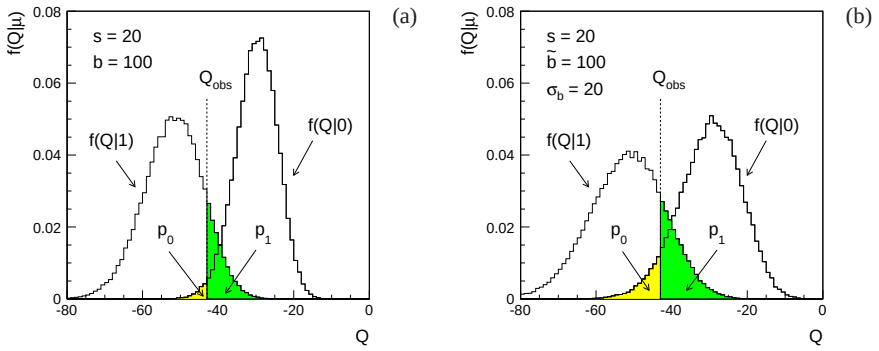

Fig. 6 (a) Distribution of the statistic Q assuming $s = 20$ and $b = 100$ under both the background-only ($\mu = 0$) signal-plus-background ($\mu = 1$) hypotheses; (b) same as in (a) but with b treated as having an uncertainty of $\sigma_b = 20$ (see text).

To establish discovery of the signal process, we use the statistic Q to test the hypothesis that $\mu = 0$. As the test statistic is a monotonic function of the likelihood ratio $L(1)/L(0)$, we obtain maximum power relative to the alternative of $\mu = 1$. The p-vale of $\mu = 0$ is computed as the area below $Q_{\rm obs}$ in Fig. 6, i.e. $p_0 = P(Q \leq Q_{\rm obs}|0)$, because here lower Q corresponds to data more consistent with a positive μ (e.g., $\mu = 1$). The p-value can be converted into a significance Z using Eq (9) and if this is greater than a specific threshold (e.g., 5.0) then one rejects the background-only hypothesis.

To set limits on μ we can use the statistic

$$Q_\mu = -2 \ln \frac{L(\mu)}{L(0)} \,, \tag{21}$$

defined such that the special case Q_1 is the same as the statistic Q used above for discovery. This will provide maximum power in a test of μ relative to the background-only alternative. The distribution of Q is also shown in Fig. 6) for $\mu = 1$. The p-value of the $\mu = 1$ hypothesis is given by the area above the observed value $Q_{\rm obs}$, since higher values of Q are more consistent with the alternative of $\mu = 0$. This is shown here for the special case of $\mu = 1$ but one can repeat the procedure using $f(Q_\mu|\mu)$ for any other value of μ and compute the p-value p_μ in the analogous manner. To find the upper limit one would carry out the analysis as described above for all values of μ and reject those that have $p_\mu < \alpha$ for, say $\alpha = 0.05$. The highest value of μ not rejected is then the upper limit at 95% C.L.

5.2 Systematic uncertainties and nuisance parameters

Until now we have treated the expected number of background events b as known with negligible uncertainty. In practice, of course, this may not be true and so we may need to regard b as an adjustable parameter of our model. That is, we regard Eq. (19) as giving $L(\mu, b)$, where μ is the parameter of interest and b is a *nuisance parameter*.

There are two main ways of eliminating the nuisance parameters from the problem. First we consider the method motivated by Bayesian statistics; an alternative approached using the profile likelihood is discussed in Sec. 6. We may have a best guess for b, say, $\tilde{b}$, and our degree of belief about the true value of the parameter may be described in a Bayesian sense by a prior pdf $\pi(b)$. As an example this could be a Gaussian distribution centred about $\tilde{b}$ with a standard deviation σ_b:

$$\pi(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-\tilde{b})^2/2\sigma_b^2} \,. \tag{22}$$

In fact a Gaussian pdf for b may not be the most appropriate model and but it we will use it here for illustrative purposes.

Using the pdf $\pi(b)$ we can construct what is called the marginal (or prior predictive) likelihood,

$$L_{\rm m}(n, \mathbf{x}_1, \ldots, \mathbf{x}_n|\mu) = \int L(n, \mathbf{x}_1, \ldots, \mathbf{x}_n|\mu, b)\pi(b)\, db \,, \tag{23}$$

where in the notation above we have emphasized that the likelihood of a model is the probability for the data under assumption of that model.

Notice that the marginal model does not represent the probability of data that would be generated if we were to really repeat the experiment. In that case we would not know the true value of b, but we could at least assume it would not change under repetition of the experiment. Rather, the marginal model represents a situation in which every repetition of the experiment is carried out with a new value of b randomly sampled from $\pi(b)$. It is in effect an average of models each with a given b, where the average is carried out with respect to the density $\pi(b)$.

For our tests we can use the same test statistic Q as before, but now we need to know its distribution under assumption of the prior predictive model. That is, if b is known exactly then we obtain distributions $f(Q|\mu, b)$ such as those shown in Fig. 6(a). What we want instead is the distribution based on data that follows the marginal model,

$$f_{\rm m}(Q|\mu) = \int f(Q|\mu, b)\pi(b)\, db \,. \tag{24}$$

Although it may not be obvious how to compute this integral, it can be done easily with Monte Carlo by generating a value of b according to $\pi(b)$, then using this value to generate the data n, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and with these we find a value of Q which is recorded in a histogram. By repeating the entire procedure a large number of times

we obtain distributions as shown in Fig. 6(b), which are generated with a Gaussian prior for b with $\tilde{b} = 100$ and $\sigma_b = 20$.

As can be seen in Fig. 6, the effect of the uncertainty on b broadens the distributions of Q such that the p-values for each hypothesis are increased. That is, one may be able to reject one or the other hypothesis in the case where b was known because the p-value may be found less than α. When the uncertainty in b is included, however, the p-values may no longer allow one to reject the model in question.

As a further step one could consider using the marginal likelihood as the basis of the likelihood ratio used in the test statistic, i.e., we take $Q = -2\ln(L_{\rm m}(1)/L_{\rm m}(0))$. Use of a different statistic simply changes the critical region of the test and thus alters the power relative to the alternative models considered. This step by itself, however, does not take into account the uncertainty in b and it will not result in a broadening of $f(Q|\mu)$ and an increase in p-values as illustrated above. This is achieved by generating the distribution of Q using the marginal model through Eq. (24). Test statistics based on the ratio of marginal likelihoods would be very difficult to compute and are not used in practice. In some cases, however, it may be beneficial to base Q on a ratio of profile likelihoods, which are described below.

6 Tests based on the profile likelihood

Suppose as before that the parameter of interest is μ and the problem may contain one or more nuisance parameters θ (such as the parameter b in the previous example). An alternative way to test hypothetical values of μ is to define the *profile likelihood*,

$$L_{\rm p}(\mu) = L(\mu, \hat{\hat{\theta}}(\mu)) \,, \tag{25}$$

where $\hat{\hat{\theta}}(\mu)$, called is the profiled value of the nuisance parameter θ, is the value that maximizes $L(\mu,\theta)$ for the specified value of μ. This is then used to construct the *profile likelihood ratio*

$$\lambda(\mu) = \frac{L_{\rm p}(\mu)}{L(\hat{\mu},\hat{\theta})} \,, \tag{26}$$

where $\hat{\mu}$ and $\hat{\theta}$ are the values of the parameters that maximize the likelihood. In some models it may be that μ can only take on values in a restricted range, e.g., $\mu \geq 0$ if this parameter is proportional to the cross section of the signal process. In this case we can, however, regard $\hat{\mu}$ as an effective estimator that is allowed to take on negative values. This will allow us to write down simple formulae for the distributions of the test statistics used that are valid in the limit where the data sample is very large.

The quantity $\lambda(\mu)$ is defined so that it lies between zero and one, with higher values indicating greater compatibility between the data and the hypothesized value of μ. We can therefore use $\lambda(\mu)$ to construct a statistic to test different values of μ.

Suppose as above that μ is proportional to the rate of the sought after signal process and we want to test the background-only ($\mu = 0$) hypothesis.

Often the signal process is such that only positive values of μ are regarded as relevant alternatives. In this case we would choose the critical region of our test of $\mu = 0$ to correspond to data outcomes characteristic of positive μ, that is, when $\hat{\mu} > 0$. It could happen that we find $\hat{\mu} < 0$, e.g., if the total observed number of events fluctuates below what is expected from background alone. Although a negative $\hat{\mu}$ indicates a level of incompatibility between the data and hypothesis of $\mu = 0$, this is not the type of disagreement that we want to exploit to declare discovery of a positive signal process.

Providing our signal models are of the type described above, we can take the statistic used to test $\mu = 0$ as

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0\,, \\ 0 & \hat{\mu} < 0\,, \end{cases} \tag{27}$$

where $\lambda(0)$ is the profile likelihood ratio for $\mu = 0$ as defined in Eq. (26). In this way, higher values of q_0 correspond to increasing disagreement between data and hypothesis, and so the p-value of $\mu = 0$ is the probability, assuming $\mu = 0$ to find q_0 at least high or higher than the observed value.

If we are interested in an upper limit for the parameter μ, then we want the critical region to correspond to data values characteristic of the alternative $\mu = 0$. This can be achieved by defining

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \leq \mu\,, \\ 0 & \hat{\mu} > \mu\,. \end{cases} \tag{28}$$

For both discovery and upper limits, therefore, the p-value for a hypothesized μ is then

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu,\theta)\,dq_\mu\,, \tag{29}$$

If we use the statistic q_μ then we find the upper limit μ_{up} at confidence level $1-\alpha$ by setting $p_\mu = \alpha$ and solving for μ. This will have the property $P(\mu_{\mathrm{up}} \geq \mu) \geq \alpha$.

To find the p-value we need the distribution of the test statistic under assumption of the same μ being tested. For sufficiently large data samples one can show that these distributions approach an asymptotic form related to the chi-square distribution, where the number of degrees of freedom is equal to the number of parameters of interest (in this example just one, i.e., μ). The asymptotic formulae are based on theorems due to Wilks [18] and Wald [19] and are described in further detail in Ref. [20].

An important advantage of using the profile likelihood ratio is that its asymptotic distribution is independent of the nuisance parameters, so we are not required to choose specific values for them to compute the p-value. In practice one has of course

a finite data sample and so the asymptotic formulae are not exact. Therefore the p-values will in general depend on the nuisance parameters to some extent.

Providing the conditions for the asymptotic approximations hold, one finds a very simple formula for the p-value,

$$p_\mu = \Phi\left(\sqrt{q_\mu}\right) \,, \tag{30}$$

where Φ is the cumulative distribution of the standard Gaussian. From Eq. (9) we find for the corresponding significance

$$Z_\mu = \sqrt{q_\mu} \,. \tag{31}$$

For discovery, we could require Z_0 greater than some threshold such as 5.0, which corresponds to $p_0 < 2.9 \times 10^{-7}$. When setting limits one usually excludes a parameter value if its p-value is less than, say, 0.05, corresponding to a confidence level of 95%, or a significance of 1.64. Although Eqs. (30) and (31) are only exact for an infinitely large data sample, the approach to the asymptotic limit is very fast and the approximations often turn out to be valid for moderate or even surprisingly small data samples. Examples can be found in Ref. [20].

For data samples not large enough to allow use of the asymptotic formulae, one must determine the distribution of the test statistics by other means, e.g., with Monte Carlo models that use specific values for the nuisance parameters. In the exact frequentist approach we would then only reject a value of μ if we find its p-value less than α for all possible value of the nuisance parameters. Therefore only a smaller set of μ values are rejected and the resulting confidence interval becomes larger, which is to say the limits on μ become less stringent. The confidence interval then *overcovers*, which is to say its probability to contain the true μ is greater than $1-\alpha$, at least for some values of the nuisance parameters.

It may seem unfortunate if we cannot reject values of μ that are retained only under assumption of nuisance parameter values that may be highly disfavoured, e.g., for theoretical reasons. A compromise solution is test μ using the p-value based only on the profiled values of the nuisance parameters, i.e., we take

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu, \hat{\hat{\theta}}(\mu))\, dq_\mu \,, \tag{32}$$

This procedure has been called *profile construction* [21] in HEP or *hybrid resampling* [22, 23] amongst statisticians. If the true values of of the nuisance parameters are equal to the profiled values, then the coverage probability of the resulting confidence interval for μ is exact. For other values of θ the interval for μ may over- or undercover. In cases where it this is crucial one may include a wider range of nuisance parameter values and study the coverage with Monte Carlo.

7 Unified intervals

The test of μ used for an upper limit assumes that the relevant alternative hypothesis is $\mu = 0$, and the critical region is chosen accordingly. In other cases one may regard values of μ both higher and lower than the one being tested as valid alternatives, and one would therefore like a test that has high power for both cases. One can show that in general there is no single test (i.e., no given critical region) that will have the highest power relative to all alternatives (see, e.g., Ref. [24], Chapter 22).

Nevertheless we can use the statistic

$$t_\mu = -2 \ln \lambda(\mu) \tag{33}$$

to construct a test for any value of μ. As before, higher values of the statistic correspond to increasing disagreement between the data and the hypothesized μ. Here, however, the critical region can include data corresponding to an estimated signal strength $\hat{\mu}$ greater or less than μ. The resulting test therefore has power relative to values of μ both higher and lower than the one being tested. If one carries out a test of all values of μ using this statistic, then both high and low values of μ may wind up being rejected.

Suppose the lowest and highest values not rejected are μ_1 and μ_2, respectively. One may be tempted to interpret the upper edge of such an interval as an upper limit in the same sense as the one derived above using q_μ from Eq. (28). The coverage probability, however, refers to the whole interval, i.e., one has $P(\mu_1 \leq \mu \leq \mu_2) \geq 1 - \alpha$. One cannot in general make a corresponding statement about the probability for the upper or lower edge of the interval alone to be above or below μ, analogous to the statement $P(\mu_{\text{up}} \geq \mu) \geq 1 - \alpha$ that holds for an upper limit.

The confidence intervals proposed by Feldman and Cousins [25], also called *unified intervals*, are based on a statistic similar to t_μ from Eq. (33) with the additional restriction that the estimator $\hat{\mu}$ that appears in the denominator of the likelihood ratio is restricted to physically allowed values of μ. Large-sample formulae for the distributions and corresponding p-values can be found in Ref. [20]. (In that reference the statistic for the case $\mu \geq 0$ is called $\tilde{t}_\mu$.)

8 Bayesian limits

Although these lectures focus mainly on frequentist statistical procedures we provide here a brief description of the Bayesian approach to setting limits. This is in fact conceptually much simpler than the frequentist procedure. Suppose we have a model that contains a parameter μ, which as before we imagine as being proportional to the rate of a sought-after signal process. In addition the model may contain some nuisance parameters θ. As in the frequentist case, we will have a likelihood $L(\mathbf{x}|\mu,\theta)$ which gives the probability for the data $\mathbf{x}$ given μ and θ. In a Bayesian analysis we are allowed to associate a probability with parameter values, and so we

assess our degree of belief in a given model (or set of parameter values) by giving the posterior probability $p(\mu, \theta|\mathbf{x})$. To find this we use Bayes' theorem (4), which we can write as a proportionality

$$p(\mu, \theta|\mathbf{x}) \propto L(\mathbf{x}|\mu, \theta)\pi(\mu, \theta) \,, \tag{34}$$

where the prior pdf $\pi(\mu, \theta)$ specifies our degree of belief in the parameters' values before carrying out the measurement.

The problematic ingredient in the procedure above is the prior pdf $\pi(\mu, \theta)$. For a nuisance parameter θ, one typically has some specific information that constrains one's degree of belief about its value. For example, a calibration constant or background event rate may be constrained by some control measurements, leading to a best estimate $\tilde{\theta}$ and some measure of its uncertainty σ_θ. Depending on the problem at hand one may from these subsidiary measurements as well as physical or theoretical constraints construct a prior pdf for θ. In many cases this will be independent of the value of the parameter of interest μ, in which case the prior will factorize, i.e., $\pi(\mu, \theta) = \pi_\mu(\mu)\pi_\theta(\theta)$. For the present discussion we will assume that this is the case.

The more controversial part of the procedure is the prior $\pi_\mu(\mu)$ for the parameter of interest. As one is carrying out the measurement in order to learn about μ, one usually does not have much information about it beforehand, at least not much relative to the amount one hopes to gain. Therefore one may like to write down a prior that is *non-informative*, i.e., it reflects a maximal degree of prior ignorance about μ, in the hopes that one will in this way avoid injecting any bias into the result. This turns out to be impossible, or at least there is no unique way of quantifying prior ignorance.

As a first attempt at a non-informative prior for μ we might choose to take it very broad relative to the likelihood. Suppose as before that μ represents the rate of signal so we have $\mu \geq 0$. As an extreme example of a broad prior we may try

$$\pi_\mu(\mu) = \begin{cases} 1 & \mu \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

This so-called flat prior is problematic for a number of reasons. First, it cannot be normalized to unit area, so it is not a proper pdf; it is said to be *improper*. Here this defect is not fatal because in Bayes' theorem the prior always appears multiplied by the likelihood, and if this falls off sufficiently rapidly as a function of μ, as is often the case in practice, then the posterior pdf for μ may indeed be normalizable.

A further difficulty with a flat prior is that our inference is not invariant under a change in parameter. For example, if we were to take as the parameter $\eta = \ln\mu$, then according to the rules for transformation of variables we find for the pdf of η

$$\pi_\eta(\eta) = \pi_\mu(\mu)\left|\frac{d\mu}{d\eta}\right| = e^\eta \pi_\mu(\mu(\eta)) \,, \tag{36}$$

so if $\pi_\mu(\mu)$ is constant then $\pi_\eta(\eta) \propto e^\eta$ which is not. So if we claim we know nothing about μ and hence use for it a constant prior, we are implicitly saying that we known something about η.

Finally we should note that the constant prior of Eq. (35) cannot in any realistic sense reflect a degree of belief, since it assigns a zero probability to the range between any two finite limits.

The difficult and subjective nature of encoding personal knowledge into priors has led to what is called *objective Bayesian statistics*, where prior probabilities are based not on an actual degree of belief but rather derived from formal rules. These give, for example, priors which are invariant under a transformation of parameters or which result in a maximum gain in information for a given set of measurements. For an extensive review see, for example, Ref. [36]; applications to HEP are discussed in Refs. [37, 38].

The constant prior of Eq. (35) has been used in HEP so widely that it serves a useful purpose as a benchmark, despite its shortcomings. Although interpretation of the posterior probability as a degree of belief is no longer strictly true, one can simply regard the resulting interval as a given function of the data, which will with some probability contain the true value of the parameter. Unlike the confidence interval obtained from the frequentist procedure, however, the coverage probability will depend in general on the true (and unknown) value of the parameter.

We now turn to the Bayesian treatment of nuisance parameters. What we get from Bayes' theorem is the joint distribution of all of the parameters in the problem, in this case both μ and θ. Because we are not interested in the nuisance parameter θ we simply integrate (or sum in the case of a discrete parameter) to find the marginal pdf for the parameter of interest, i.e.,

$$p(\mu|\mathbf{x}) = \int p(\mu,\theta|\mathbf{x})\, d\theta \ . \tag{37}$$

One typically has not one but many nuisance parameters and the integral required to marginalize over them cannot be carried out in closed form. Even Monte Carlo integration based on the acceptance-rejection method becomes impractical if the number of parameters is too large, since then the acceptance rate becomes very small. In such cases, Markov Chain Monte Carlo (MCMC) provides an effective means to calculate integrals of this type. Here one generates a correlated sequence of points in the full parameter space and records the distribution of the parameter of interest, in effect determining its marginal distribution. An MCMC method widely applicable to this sort of problem is the Metropolis-Hastings algorithm, which is described briefly in Ref. [13]. In-depth treatments of MCMC can be found, for example, in the texts by Robert and Casella [32], Liu [33], and the review by Neal [34].

9 Limits for a Poisson counting experiment

As a simple example, consider an experiment in which one counts a number of events n, modeled as following a Poisson distribution with a mean of $s+b$, where s and b are the contributions from signal and background processes, respectively. Suppose that b is known and we want to set an upper limit on s. Here we will do this with both frequentist and Bayesian methods.

To construct the frequentist upper limit we should test all hypothetical values of s against to the alternative of $s = 0$, so the critical region consists of low values of n. This means we take the p-value of a hypothesized s to be the probability to find n as small as observed or smaller, i.e.,

$$p_s = \sum_{m=0}^{n} \frac{(s+b)^m}{m!} e^{-(s+b)} \,. \tag{38}$$

The upper limit at $\mathrm{CL} = 1-\alpha$ is found from the value of s such that the p-value is equal to α, i.e.,

$$\alpha = \sum_{m=0}^{n} \frac{(s_{\mathrm{up}}+b)^m}{m!} e^{-(s_{\mathrm{up}}+b)} = 1 - F_{\chi^2}(2(s_{\mathrm{up}}+b), 2(n+1)) \,, \tag{39}$$

where in the second equality we used a trick that relates the sum of Poisson probabilities to the cumulative chi-square distribution for $2(n+1)$ degrees of freedom. This allows us to solve for the upper limit

$$s_{\mathrm{up}} = \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha, 2(n+1)) - b \,, \tag{40}$$

where $F^{-1}_{\chi^2}$ is the chi-square quantile (inverse of the cumulative distribution). The upper limit s_{up} is shown in Fig. 7(a) for $1-\alpha = 95\%$ as a function of b for different numbers of observed events n.

To find the corresponding upper limit in the Bayesian approach we need to assume a prior pdf for s. If we use the flat prior of Eq. (35), then by using Bayes' theorem we find the posterior pdf

$$p(s|n) \propto \frac{(s+b)^n}{n!} e^{-(s+b)} \tag{41}$$

for $s \geq 0$ and $p(\mu|n) = 0$ otherwise. This can be normalized to unit area, which gives

$$p(s|n) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(b, n+1)} \tag{42}$$

where $\Gamma(b, n+1) = \int_b^\infty x^n e^{-x}\, dx$ is the upper incomplete gamma function.

Since in the Bayesian approach we are assigning a probability to s, we can express an upper limit simply by integrating the posterior pdf from the minimum value $s = 0$ up to an upper limit s_{up} such that this contains a fixed probability, say, $1-\alpha$. That is, we require

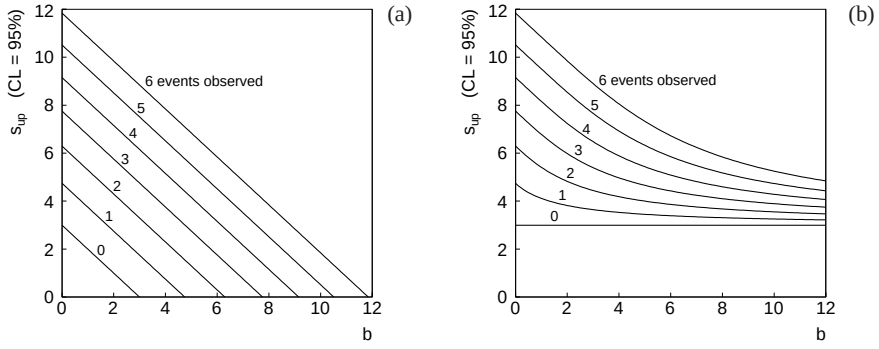

Fig. 7 Upper limits on the mean number of signal events s at 95% confidence level as a function of the expected background b for (a) the frequentist method and (b) Bayesian method with a flat prior.

$$1-\alpha = \int_0^{s_{\rm up}} p(s|n)\,ds\,. \tag{43}$$

To solve for $s_{\rm up}$ we can use the integral

$$\int_0^a x^n e^{-x}\,dx = \Gamma(n+1)F_{\chi^2}(2a,2(n+1))\,, \tag{44}$$

where again F_{χ^2} is the cumulative chi-square distribution for $2(n+1)$ degrees of freedom. Using this we find for the upper limit

$$s_{\rm up} = \frac{1}{2}F^{-1}_{\chi^2}(p,2(n+1)) - b\,, \tag{45}$$

where

$$p = 1-\alpha\left(1-F_{\chi^2}(2b,2(n+1))\right)\,. \tag{46}$$

This is shown in Fig. 7(b). Interestingly, the upper limits for the case of $b=0$ happen to coincide exactly with the values we found for the frequentist upper limit, and for nonzero b the Bayesian limits are everywhere higher. This means that the probability for the Bayesian interval to include the true value of s is higher than $1-\alpha$, so in this sense one can say that the Bayesian limit is conservative. The corresponding unified interval from the procedure of Feldman-Cousins is described in Ref. [25].

If the parameter b is not known, then this can be included in the limit using the methods discussed above. That is, one must treat b as a nuisance parameter, and in general one would have some control measurement that constrains its value. In the frequentist approach b is eliminated by profiling; in the Bayesian case one requires a prior pdf for b and simply marginalizes the joint pdf of s and b to find the posterior

$p(s|n)$. The problem of a Poisson counting experiment with additional nuisance parameters is discussed in detail in Refs. [26, 37].

10 Limits in cases of low sensitivity

An important issue arises when setting frequentist limits that is already apparent in the example from Sec. 9. In Fig. 7(a), which shows the frequentist upper limit on the parameter s as a function of b, one sees that $s_{\rm up}$ can be arbitrarily small. Naive application of Eq. (40) can in fact result in a negative upper limit for what should be an intrinsically positive quantity. What this really means that all values of s are rejected in a test of size α. This can happen if the number of observed events n fluctuates substantially below the expected background b. One is then faced with the prospect of not obtaining a useful upper limit as the outcome of one's expensive experiment. It might be hoped that such an occurrence would be rare but by construction it should happen with probability α, e.g., 5% of the time.

Essentially the same problem comes up whenever we test any hypothesis to which we have very low sensitivity. What "low sensitivity" means here is that the distributions of whatever statistic we are using is almost the same under assumption of the signal model being tested as it is under the background-only hypothesis. This type of situation is illustrated in Fig. 8(a), where here we have labeled the model including signal $s+b$ (in our previous notation, $\mu = 1$) and the background-only model b (i.e., $\mu = 0$)).

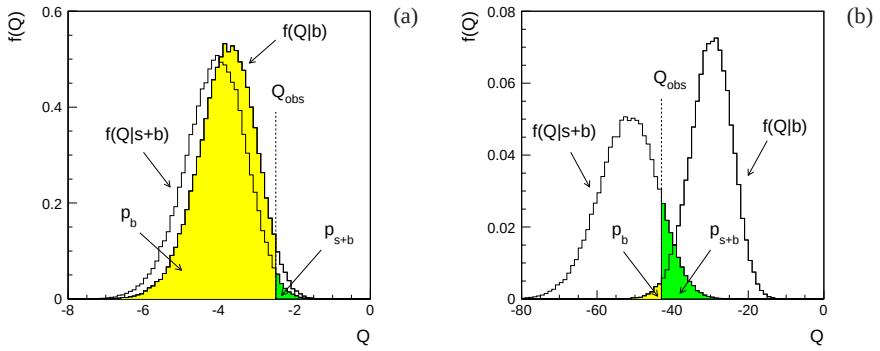

Fig. 8 (a) Distributions of the statistic Q indicating low sensitivity to the hypothesized signal model; (b) illustration of the ingredients for the CL_s limit.

A test of the $s+b$ hypothesis consists of high values of Q. Equivalently, the p-value is the probability $p_{s+b} = P(Q \geq Q_{\rm obs}|s+b)$. Because the distributions of Q under both hypotheses are very close, the power of the test of $s+b$ is only slightly

greater than the size of the test α, which is equivalent to the statement that the quantity $1 - p_b$ is only slightly greater than p_{s+b}.

If we have no sensitivity to a particular model, such as the hypothesis of a Higgs boson with a mass much greater than what we could produce in our experiment, then we do no want to not reject it, since our measurement can produce no evidence to justify such a claim. Unfortunately, the frequentist procedure that rejects the signal model if its p-value is found less then α will do just that with a probability of at least α. And this will happen even if the model is, from an experimental standpoint, virtually indistinguishable from the background-only hypothesis. Since we typically take $\alpha = 0.05$, we will exclude one model out of every twenty to which we have no sensitivity.

One solution to this problem is the CL_s procedure proposed by Alex Read [28, 29], whereby the threshold for rejecting a model is altered in a way that prevents one from rejecting a model in the limit that one has very little sensitivity, but reverts to the usual frequentist procedure when the sensitivity is high. This is achieved by defining

$$\mathrm{CL}_s = \frac{P(Q \geq Q_{\mathrm{obs}}|s+b)}{P(Q \geq Q_{\mathrm{obs}}|b)} = \frac{p_{s+b}}{1-p_b} \; . \tag{47}$$

The quantity CL_s then is then used in place of the p-value p_{s+b}, i.e., the $s+b$ model is rejected if one finds $\mathrm{CL}_s \leq \alpha$. The ingredients are illustrated in Fig. 8(b).

One can understand qualitatively how this achieves the desired goal by considering the case where the distributions of Q under the two hypotheses $s+b$ and b are close together. Suppose the observed value Q_{obs} is such that p_{s+b} is less than α, so that in the usual frequentist procedure we would reject the $s+b$ hypothesis. In the case of low sensitivity, however, the quantity $1 - p_b$ will also be small, as can be seen from Fig. 8(a). Therefore the quantity CL_s will be greater than p_{s+b} such that the $s+b$ model is not rejected by the criterion of Eq. (47).

If on the other hand the distributions are well separated, and Q_{obs} is such that the $p_{s+b} < \alpha$, then p_b will also be small and the term $1 - p_b$ that appears in the denominator of CL_s will be close to unity. Therefore in the case with high sensitivity, using CL_s is similar to what is obtained from the usual frequentist procedure based on the p-value p_{s+b}.

The largest value of s not rejected by the CL_s criterion gives the corresponding CL_s upper limit. Here to follow the traditional notation we have described it in terms of the mean number of signal events s rather than the strength parameter μ, but it is equivalent to using $\mathrm{CL}_\mu = p_\mu/(1-p_0)$ to find an interval for μ.

The CL_s procedure described above assumes that the test statistic Q is continuous. The recipe is slightly different if the data are discrete, such as a Poisson distributed number of events n with a mean $s+b$. In this case the quantity CL_s is defined as

$$\mathrm{CL}_s = \frac{P(n \leq n_{\mathrm{obs}}|s+b)}{P(n \leq n_{\mathrm{obs}}|b)} \; , \tag{48}$$

where $n_{\rm obs}$ is the number of events observed. Here the numerator is p_{s+b}, which the same as in Eq. (47). The p-value of the background-only hypothesis is $p_b = P(n \geq n_{\rm obs}|b)$, but the denominator in Eq. (48) requires n less than *or equal* to $n_{\rm obs}$, so this is not exactly the same as $1 - p_b$. Eq. (48) is the fundamental definition and it reduces to the ratio of p-values for the case of a continuous test statistic.

For a Poisson distributed number of events, the CL_s upper limit coincides exactly with the Bayesian upper limit based on the flat prior as shown in Fig. 7(b). It is thus also greater than or equal to the limit based on the p-value and is in this sense conservative. It also turns out that the CL_s and Bayesian limits (using a flat prior) agree for the important case of Gaussian distributed data [28].

11 The look-elsewhere effect

Recently there has been important progress made on the problem of multiple testing, usually called in particle physics the "look-elsewhere effect" [30, 31]. The problem often relates to finding a peak in a distribution when the peak's position is not predicted in advance. In the frequentist approach using a p-value, one must determine the probability, under the background-only hypothesis, to find a peak as significant as the one found more more so anywhere in the search region.

The "brute-force" solution to this problem involves generating data under the background-only hypothesis and for each data set, fitting a peak of unknown position and recording a measure of its significance. To establish a discovery one often requires a p-value less than 2.9×10^{-7}, corresponding to a 5σ effect. Thus determining this with Monte Carlo requires generating and fitting an enormous number of experiments, perhaps several times 10^7.

In contrast, if the position of the peak were known in advance, then the fit to the distribution would be much faster and easier, and furthermore one can in many cases use formulae valid for sufficiently large samples that bypass completely the need for Monte Carlo (see, e.g., [20]). But this "fixed-position" p-value would not be correct in general, as it assumes the position of the peak was known in advance.

Gross and Vitells [30] have described a method that allows one to modify the p-value computed under assumption of a fixed position to obtain the correct value by use of a relatively fast Monte Carlo calculation. Suppose a test statistic q_0, defined so that larger values indicate increasing disagreement with the data, is observed to have a value u. Furthermore suppose the model contains a nuisance parameter θ (such as the peak position) which is only defined under the signal model (there is no peak in the background-only model). An approximation for the desired "global" p-value is found to be

$$p_{\rm global} \approx p_{\rm local} + \langle N_u \rangle \ , \tag{49}$$

where $p_{\rm local}$ is the p-value assuming a fixed value of θ (e.g., fixed peak position), and $\langle N_u \rangle$ is the mean number of "upcrossings" of the the statistic q_0 above the level u in the range of the nuisance parameter considered (e.g., the mass range).

The value of $\langle N_u \rangle$ can be estimated from the number of upcrossings $\langle N_{u_0} \rangle$ above some much lower value, u_0, by using a relation due to Davis [35],

$$\langle N_u \rangle \approx \langle N_{u_0} \rangle e^{-(u-u_0)/2} \,. \tag{50}$$

By choosing u_0 sufficiently low, the value of $\langle N_u \rangle$ can be estimated by simulating only a very small number of experiments, rather than the 10^7 needed if one is dealing with a 5σ effect.

Vitells and Gross also indicate how to extend the correction to the case of more than one parameter, e.g., where one searches for a peak of both unknown position and width, or for searching for a peak in a two-dimensional space, such as an astrophysical measurement on the sky [31]. Here one may find some number of regions where signal appears to be present, but within those regions there may be islands or holes where the significance is lower. In the generalization to multiple dimensions, the number of upcrossings of the test statistic q_0 is replaced by the expectation of a quantity called the Euler characteristic, which is roughly speaking the number of disconnected regions with significant signal minus the number of 'holes'.

It should be emphasized that an exact accounting of the look-elsewhere effect requires that one specify where else one looked, e.g., the mass range in which a peak was sought. But this may be have been defined in a somewhat arbitrary manner, and one might have included not only the mass range but other variables that were also inspected for peaks but where none was found. Thus it perhaps not worth expending great effort on an exact treatment of the look-elsewhere effect, as one would do in the brute-force method mentioned above. Rather, the more easily obtained local p-value can be reported along with an approximate correction to account for the range of measurements in which the effect could have appeared.

12 Why 5 σ?

Common practice in HEP has been to regard an observed signal to be worthy of the word "discovery" when its significance exceeds $Z = 5$, corresponding to a p-value of the background-only hypothesis of 2.9×10^{-7}. This is in stark contrast to many other fields (e.g., medicine, psychology) in which a p-value of 5% ($Z = 1.64$) is considered significant.

First, it is not clear that the same significance threshold should be used in all cases. Whether one is convinced that a discovery is real should take into account the plausibility of the implied signal and how well it describes the data. If the discovered phenomenon is a priori very unlikely, then more evidence is required to produce a given degree of belief that the new phenomenon exists. As Carl Sagan said, "... extraordinary claims require extraordinary evidence" [39]. This follows directly

from Bayes' theorem (34), whereby the posterior probability of a hypothesis is proportional to its prior probability. If an experimental result can only be explained by phenomena that may not be impossible but nevertheless highly improbable (fifth force, superluminal neutrinos), then we should demand a higher level of statistical significance.

Some phenomena, on the other hand, are regarded by the community as quite likely before they are observed experimentally. Most particle physicists would have bet on the Higgs boson well in advance of the direct experimental evidence. As with the Higgs, however, when a discovery is announced in HEP it is usually something fairly important and the cost of a false claim is perceived to be quite high. Every time the community endures a retracted discovery there is a tendency to think that the threshold should be higher.

Another reason for the high five-sigma threshold is that the experimenter may be unsure of the statistical model on which the reported significance relies. To first approximation one can think of the significance Z as the estimated size of the signal divided by the standard deviation σ in the estimated background. Here σ characterizes the level of random fluctuation in the background, i.e., it is a statistical error. If we have a systematic uncertainty in the background as well, then roughly speaking these should get added in quadrature. If an underestimate of our systematic errors would result in our σ being wrong by a factor of several, then a mere three-sigma effect may be no real effect at all. The high threshold in this case thus compensates for modeling uncertainty.

Another important issue is the look-elsewhere effect, where as discussed in Sec. 11 it is difficult to define exactly where else one looked. That is, should one correct for the fact that the search histogram had 100 bins, or also for the fact that one looked at 100 different histograms, or perhaps account for the thousands of scientists all carrying out searches? Surely in such a scenario someone will see a bump in a histogram somewhere that appears significant. Since it is impossible to draw an unambiguous boundary around where one "looked", there always remains a nagging feeling that one's correction for this effect may have been inadequate, hence the desire for a greater margin of safety before announcing a discovery.

The p-value, however, really only addresses the issue of whether a fluctuation in the background-only model is likely to lead to data as dissimilar to background as what was actually obtained. It is not designed to compensate for systematic errors in the model, the cost of announcing a false discovery or the plausibility of the phenomena implied by the discovery. When a new phenomenon is discovered, it often emerges first as only marginally significant, then increases to the point where everyone is convinced. At first, everyone asks whether the apparent signal is just a fluctuation. After a while, people stop asking that question. It is obvious one has seen something, the question is whether that is "new physics" or an uncontrolled systematic effect. Provided that the look-elsewhere effect is taken into account in a reasonable way, this transition probably takes place closer to the three-sigma level, in any case well before $Z = 5$.

Nevertheless, the 5-sigma threshold continues to be used to decide when the word "discovery" is appropriate. In future the HEP community should perhaps think of

better ways of answering the different questions that arise when searching for new phenomena, since the statistical significance is really only designed to say whether the data, in the absence of a signal, is likely to have fluctuated in manner at least as extreme as what was observed. Lumping all of the issues mentioned above into the p-value simply makes them more difficult to disentangle.

13 Conclusions

To search for new physical phenomena we need to be able to demonstrate quantitatively that our data cannot be described using only known proceesses. In these lectures we have seen how statistical tests allow us to carry out this task. They provide a framework for rejecting hypotheses on the basis that the data we observed were uncharacteristic for them and more indicative of an alternate explanation. Frequentist statistical tests nevertheless prevent one from asking directly certain seemingly relevant questions, such as "what is the probability that my theory is true?". Bayesian statistics does allow one to quantify such a degree of belief, at the expense of having to supply subjective prior probabilities. The frequentist and Bayesian approaches answer different but related questions and both are valuable tools.

We did not have time to discuss in detail many other statistical issues such as Bayesian methods for establishing discovery, multivariate techniques and more sophisticated means for improving the accuracy of statistical models through carefully motivated nuisance parameters. These methods will no doubt play an important role when the LHC enters its next data-taking phase.

Acknowledgements I wish to convey my thanks to the students and organizers of the 69th SUSSP in St. Andrews for a highly stimulating environment. The friendly atmosphere and lively discussions created a truly enjoyable and productive school.

References

1. C. Amsler et al. (Particle Data Group), Physics Letters B667 (2008) 1; available at `pdg.lbl.gov`.
2. G.D. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.
3. L. Lyons, *Statistics for Nuclear and Particle Physicists*, Cambridge University Press, 1986.
4. R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989.
5. F. James, *Statistical Methods in Experimental Physics*, 2nd ed., World Scientific, 2006.
6. S. Brandt, *Data Analysis* 3rd ed., Springer, 1999.
7. G. Cowan *A Survey of Unfolding Methods for Particle Physics*, in Proc. Conf. on Advanced Statistical Techniques in Particle Physics, M.R. Whalley and L. Lyons (eds.), IPPP/02/39, Durham 2002.
8. A.N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, 1933; *Foundations of the Theory of Probability*, 2nd ed., Chelsea, 1956.
9. C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

10. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.
11. R. Duda, P. Hart, D. Stork, *Pattern Classification*, 2nd ed., Wiley, 2001.
12. A. Webb, *Statistical Pattern Recognition*, 2nd ed., Wiley, 2002.
13. G. Cowan, *Topics in Statistical Data Analysis for HEP*, in *LHC Physics*, T. Binoth, C. Buttar, P.J. Clark, E.W.N. Glover (eds.), Taylor and Francis, 2012.
14. Links to the Proceedings of the PHYSTAT conference series (Durham 2002, Stanford 2003, Oxford 2005, and Geneva 2007) can be found at `phystat.org`.
15. A. Höcker et al., *TMVA Users Guide*, `physics/0703039` (2007); software available from `tmva.sf.net`.
16. I. Narsky, *StatPatternRecognition: A C++ Package for Statistical Analysis of High Energy Physics Data*, `physics/0507143` (2005); software available from `sourceforge.net/projects/statpatrec`.
17. Hongbo Hu and Jason Nielsen, *Analytic Confidence Level Calculations using the Likelihood Ratio and Fourier Transform*, arXiv:physics/9906010 [physics.data-an] (1999).
18. S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.
19. A. Wald, *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*, Transactions of the American Mathematical Society, Vol. **54**, No. 3 (Nov., 1943), pp. 426-482.
20. G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C (2011) 71:1554; arXiv:1007.1727.
21. K. Cranmer, *Statistical Challenges for Searches for New Physics at the LHC*, in Proceedings of PHYSTAT 2005, Louis Lyons and Muge Karagoz Unel (eds.), Oxford, 2005; arXiv:physics/0511028.
22. C. Chuang and T.L. Lai., Statist. Sinica, 10:150, 2000.
23. M. Walker B. Sen and M. Woodroofe, Statist. Sinica, 19:301314, 2009.
24. A. Stuart, J.K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A: *Classical Inference and the Linear Model* 6th Ed., Oxford Univ. Press (1999), and earlier editions by Kendall and Stuart.
25. Robert D. Cousins and Gary J. Feldman, Phys. Rev. D 57, 3873 (1998).
26. Kyle Cranmer, *Frequentist Hypothesis Testing with Background Uncertainty* in L. Lyons et al. (eds.), Proceedings of PHYSTAT 2003, SLAC, Stanford California, September 8–11, 2003, 261–264.
27. Luc Demortier, *Objective Bayesian Upper Limits for Poisson Processes*, CDF Memo 5928, 2005.
28. A.L. Read, J. Phys. G **28**, 2693 (2002).
29. T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999).
30. E. Gross and O. Vitells, Eur. Phys. J C70 (2010) 525-530; arXiv:1005.1891.
31. E. Gross and O. Vitells, *Estimating the significance of a signal in a multi-dimensional search*, arXiv:1105.4355.
32. C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., Springer, 2004.
33. J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, 2001.
34. R.M. Neal, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, available from `www.cs.toronto.edu/~radford/res-mcmc.html`.
35. R.B. Davis, Biometrika 74, (1987) 33-43.
36. Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435 (1996) 1343–1370.
37. L. Demortier, S. Jain and H.B. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82, 034002 (2010); eprint: arXiv:1002.1111 [stat.AP].
38. Diego Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) P01012; eprint: arXiv:1108.4270 [physics.data-an].
39. Carl Sagan, *Cosmos*, Episode 12, PBS (December 14, 1980).