Open issues in the wake of Banff 2010

Luc Demortier
The Rockefeller University, New York, NY 10065, USA

Abstract
We review, in some cases very succinctly, statistical issues in the formulation of discovery procedures for high energy physics. This includes alternatives to p-value tests, the look-elsewhere effect, measurement sensitivity, implicit statistical models, parton density uncertainties, reference priors, profile likelihood methods, and extreme value theory.

1 Introduction

From 11 to 16 July 2010, a group of statisticians and physicists met at the Banff International Research Station in the Canadian Rockies to debate statistical issues related to the significance of discovery claims. Although these discussions did not lead to a miraculous consensus on how to claim or not claim discoveries, progress was made in understanding some questions and in learning about potentially useful statistical techniques that are not yet known in the high energy physics community. Section 2 starts with a critical look at the way we quantify evidence against a given hypothesis, and why the almost exclusive use of p-values in our field may not be optimal. Section 3 discusses how to report a failure to discover, and the importance of measurement sensitivity for this. Difficulties arising from likelihood functions that cannot be written down analytically are explored in section 4, and the question of parton density uncertainties is summarized in section 5. Finally, some technical advances in profile likelihood techniques and reference priors are briefly described in sections 6 and 7 respectively, and the potential usefulness of extreme value theory is mentioned in section 8.

2 Discovery claims

Discovery claims in high energy physics are almost universally based on p-value calculations, regardless of the type of hypothesis that is being tested. Equally universal is the discovery threshold, which is set at five standard deviations, corresponding to a Type-I error rate of 2.87×10^{-7}. This threshold was chosen a long time ago [1], based on a back-of-the-envelope estimate of the probability of a false discovery claim in the vast number of histograms examined by all high energy physicists in the course of one year. Since then, statisticians have given ample warning that the evidence contained in a dataset for or against a given hypothesis depends strongly on the type of hypothesis being tested, on the formulation of alternatives, on sample size, on the dimensionality of the problem, and on the stopping rule. Thus it may be time to question the universality of high energy physics procedures in this regard, or at least to explore alternatives. As it turns out, these alternatives can produce results that are quite different from those obtained with p-values.

Another reason to explore alternatives is that p-values are easily misinterpreted, if not by the physicists who produce them, then almost certainly by the public at large. The two most common misinterpretations are that a p-value represents the posterior probability of a hypothesis, or the odds against it, in light of the data. Since these concepts of posterior probability and odds actually belong to the Bayesian paradigm, it is natural to turn to the latter in a search for alternative testing methods. To contrast p-values with Bayesian measures of evidence we start with a well-known paradox formulated by Lindley in 1957 [2].

2.1 Lindley's paradox

Suppose first that we have n measurements $X_1, X_2, \ldots, X_n$ distributed according to a Gaussian with unknown mean μ and known variance σ^2. The likelihood can be reduced to:

$$\mathcal{L}(\mu) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{\bar{x}_o-\mu}{\sigma/\sqrt{n}}\right)^2}}{\sqrt{2\pi}\,\sigma/\sqrt{n}}, \tag{1}$$

where $\bar{x}_o$ is the observed average of all the measurements. We are interested in testing

$$H_0:\ \mu=\mu_0 \quad \text{versus} \quad H_1:\ \mu=\mu_1, \tag{2}$$

with $\mu_1 > \mu_0$. A sufficient test statistic is the average $\bar{X}$, and large values of $\bar{X}$ indicate deviation from H_0 in the direction of H_1. The p-value is therefore:

$$p_0 \;=\; \mathbb{P}\big[\bar{X} \geq \bar{x}_o \big| H_0\big] \;=\; 1-\Phi(z_o), \tag{3}$$

where $\Phi(z)$ is the cumulative standard normal distribution and $z_o \equiv (\bar{x}_o-\mu_0)/(\sigma/\sqrt{n})$ is the number of standard deviations away from H_0. For a Bayesian analysis we must first assign prior probabilities π_0 to H_0 and π_1 to H_1, with $\pi_0+\pi_1=1$. A typical non-informative choice is $\pi_0=\pi_1=1/2$, but the argument works for any value $\pi_0>0$. The posterior probability of H_0 is:

$$p(H_0\,|\,\vec{x}) \;=\; \frac{\pi_0\,\mathcal{L}(\mu_0)}{\pi_0\,\mathcal{L}(\mu_0)+\pi_1\,\mathcal{L}(\mu_1)} \;=\; \left[1+\frac{\pi_1}{\pi_0}\,e^{\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)\left(\frac{\bar{x}_o-(\mu_0+\mu_1)/2}{\sigma/\sqrt{n}}\right)}\right]^{-1}. \tag{4}$$

Note how this posterior couples the measurement sensitivity, $(\mu_1-\mu_0)/(\sigma/\sqrt{n})$, with the evidence contained in the data, $z_{\rm Bayes} \equiv [\bar{x}_o-(\mu_1+\mu_0)/2]/(\sigma/\sqrt{n})$. If either quantity is zero, the posterior probability of H_0 reduces to π_0. Rewriting the posterior in terms of z_o,

$$p(H_0\,|\,\vec{x}) \;=\; \left[1+\frac{\pi_1}{\pi_0}\,e^{-\frac{1}{2}\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)^2+\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)z_o}\right]^{-1}, \tag{5}$$

shows that for a fixed p-value p_0 (or equivalently, a fixed z_o value), the posterior probability of H_0 goes to 1 as the sample size n increases. With α the Type-I error rate, it could happen that a frequentist finds $p_0<\alpha$ and rejects H_0, whereas a Bayesian concludes that the evidence in the data supports H_0. The reason for this discrepancy is clear: evidence in the p-value sense is measured by z_o, which only takes H_0 into account, whereas evidence in the Bayes sense is measured by $z_{\rm Bayes}$, which takes both H_0 and H_1 into account. As the measurement resolution $\sigma/\sqrt{n}$ improves, the only way to keep z_o fixed is to increase the number of standard deviations between the data $\bar{x}_o$ and H_1. Eventually the Bayesian evidence will favor H_0.

Within the Neyman-Pearson theory of testing, the alternative hypothesis H_1 influences the test via the Type-II error rate β, the probability of incorrectly rejecting H_1. As the sample size increases, keeping α fixed allows β to become arbitrarily small, thereby shifting the emphasis from protecting H_0 (the usual goal of an experimenter) to protecting H_1. This can be avoided by letting α decrease as the sample size increases.

What happens if we remove the advantage the Bayesian approach draws from looking at a *precise* alternative hypothesis? Suppose we replace test (2) by:

$$H_0:\ \mu=\mu_0 \quad \text{versus} \quad H_1':\ \mu>\mu_0. \tag{6}$$

Here the alternative hypothesis is completely vague: the lack of focus on H_1 that characterizes the p-value has been incorporated in H_1'. This is the situation examined by Lindley in his famous paradox. A p-value analysis of this test yields the same result as before, namely Eq. (3). On the other hand,

a Bayesian analysis requires that, in addition to the hypothesis priors π_0 and π_1, we specify a prior distribution $g(\mu)$ for μ under H_1'. The actual form of $g(\mu)$ does not matter much because we will be taking the limit $n \to \infty$. We only assume that $g(\mu)$ is continuous and integrates to 1 over $\mu > \mu_0$. At large values of n and for positive z_o the posterior probability of H_0 is then given by:

$$p(H_0 \,|\, \vec{x}) \;=\; \frac{\pi_0 \,\mathcal{L}(\mu_0)}{\pi_0 \,\mathcal{L}(\mu_0) + \pi_1 \int_{\mu>\mu_0} \mathcal{L}(\mu)\, g(\mu)\, d\mu} \;\simeq\; \left[1 + \frac{\pi_1}{\pi_0}\, \frac{\sqrt{2\pi}\,\sigma}{\sqrt{n}}\, e^{\frac{z_o^2}{2}}\, g(\mu_0)\, \Phi(z_o)\right]^{-1}. \qquad (7)$$

Thus we find again that for a fixed p-value evidence z_o against H_0, the posterior probability of H_0 goes to 1 at large n, hence the paradox. A striking aspect of this paradox is that it arises in the large-sample limit, where the Bayesian and frequentist paradigms often agree in problems of point and interval estimation.

2.2 Resolution

The statistics literature on Lindley's paradox is extensive, and many resolutions have been proposed [3]. A recurring theme in this literature is that the choice of test procedure should depend on one's *prior* beliefs in the hypotheses being tested. For test (6) one can imagine three possibilities:

1. due to past experience or compelling theoretical arguments, there is a concentration of prior belief on H_0;
2. H_0 is not particularly believable, but represents a valuable simplification of our description of the physics process under study;
3. H_0 is not particularly believable, but is stated for convenience (e.g. the hypothesis we are really interested in is $\mu \leq \mu_0$, but $\mu = \mu_0$ is easier to analyze).

When searching for new physics, test (6) is rather common, with μ representing, for example, the production rate of a new particle. Our prior beliefs regarding H_0 and H_1 can then be characterized as follows:

- Even though the physical theory underlying H_0 (the standard model of particle physics) describes a vast body of previous observations extremely well, we know that it is incomplete, and that somewhere it predicts something that will not be observed. Fundamentally the theory is wrong.
- However, we do not know where the breakdown will occur. There are many predictions that can be tested. Furthermore, if the test at hand should be the one to detect a breakdown, there may be more than one physics explanation that could incorporate the alternative hypothesis. It is also possible that the correct physics explanation hasn't been formulated yet.

Which of the three prior belief structures does this situation correspond to? If we leave aside the third case (misspecification of H_0), it could be argued that we have strong belief in the (limited) validity of the standard model (case 1), or that we only view the standard model as a useful simplification of a more fundamental theory (case 2). Each of these views receives its own treatment within the Bayesian paradigm and leads to further insights into Lindley's paradox. It is also possible to accomodate both views in a single treatment.

2.2.1 Case 1: the null hypothesis enjoys strong prior belief

An important insight here is that it is very rare that one tests a true point null hypothesis. Even if the theoretical hypothesis is a point (e.g. the production rate of the Higgs boson is exactly zero because the Higgs boson does not exist), there are always unknown measurement biases that cause the actually tested hypothesis to be "fuzzy". Without arguing this point in detail, it is relatively easy to see how it leads to a resolution of Lindley's paradox [4].

Suppose that by $H_0 : \mu = \mu_0$ we really mean to approximate the hypothesis $H_0' : \mu_0 \leq \mu \leq \mu_0 + \epsilon$ for some small positive ϵ that describes the unknown biases. The test is therefore:

$$H_0' : \mu_0 \leq \mu \leq \mu_0 + \epsilon \quad \text{versus} \quad H_1'' : \mu > \mu_0 + \epsilon, \tag{8}$$

and the p-value is:

$$p_0' \;=\; \sup_{\mu_0 \leq \mu \leq \mu_0 + \epsilon} \mathbb{P}\big[\bar{X} \geq \bar{x}_o \big| H_0'\big] \;=\; \sup_{\mu_0 \leq \mu \leq \mu_0 + \epsilon} \left[1 - \Phi\Big(\frac{\bar{x}_o - \mu}{\sigma/\sqrt{n}}\Big)\right] \;=\; 1 - \Phi\Big(\frac{\bar{x}_o - \mu_0 - \epsilon}{\sigma/\sqrt{n}}\Big). \tag{9}$$

For the Bayesian analysis we suppose that there is a continuous, proper prior $\pi(\mu)$ that peaks inside H_0', such that $\pi_0 = \int_{H_0'} \pi(\mu)\, d\mu$. The posterior probability of H_0' is:

$$p(H_0' \,|\, \vec{x}_o) \;=\; \frac{\int_{\mu_0}^{\mu_0 + \epsilon} \mathcal{L}(\mu)\, \pi(\mu)\, d\mu}{\int_{-\infty}^{+\infty} \mathcal{L}(\mu)\, \pi(\mu)\, d\mu}. \tag{10}$$

At large enough n the likelihood $\mathcal{L}(\mu)$ concentrates around $\bar{x}_o$. Solving equation (9) for $\bar{x}_o$ yields:

$$\bar{x}_o \;=\; \mu_0 + \epsilon + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - p_0'). \tag{11}$$

Hence for fixed p_0' the likelihood concentrates at the edge of H_0' as n becomes large. For a smooth prior $\pi(\mu)$ the numerator of posterior (10) can therefore be approximated by:

$$\int_{\mu_0}^{\mu_0 + \epsilon} \mathcal{L}(\mu)\, \pi(\mu)\, d\mu \;\simeq\; \pi(\mu_0 + \epsilon) \left[\Phi\Big(\frac{\mu_0 + \epsilon - \bar{x}_o}{\sigma/\sqrt{n}}\Big) - \Phi\Big(\frac{\mu_0 - \bar{x}_o}{\sigma/\sqrt{n}}\Big)\right] \;\simeq\; \pi(\mu_0 + \epsilon)\, p_0', \tag{12}$$

where the approximation is valid in the limit where n goes to infinity while p_0' remains constant. A similar calculation for the denominator of (10) yields $\pi(\mu_0 + \epsilon)$. Taking the ratio, we find that $p(H_0' \,|\, \vec{x}_o) \to p_0'$, thus resolving the paradox.

Both Lindley's paradox and the above resolution are formulated in the large-sample limit. However, in problems of practical interest it is rare that one is able to specify ϵ, and in finite samples it is not possible to determine how close the p-value will be to the posterior probability of the null hypothesis H_0 without knowing the prior $\pi(\mu)$. Unfortunately it is notoriously difficult to construct objective priors for testing a precise hypothesis against a vague one (as in equation 6). The problem is that objective priors often tend to be improper. To circumvent this problem, ref. [4] studies lower bounds on Bayes factors and posterior probabilities over wide classes of proper priors. The surprising result is that even these lower bounds are significantly larger than the corresponding p-values, indicating that the latter overestimate the evidence against the null hypothesis. Furthermore, p-values cannot be "recalibrated" for a variety of reasons: the calibration would depend on the sample size, on the postulated probability density of the observations, on the stopping rule of the experiment, and on the type of null hypothesis being tested.

2.2.2 Case 2: the null hypothesis provides a useful simplification

Again we consider test (6), but this time we assume that, although belief in H_0 is not particularly high, this hypothesis embodies a *useful* simplification of the theory that describes the observations [5]. In other words, μ_0 is special in terms of a utility function rather than in terms of prior belief. Let $u(d_i, \mu)$ be the utility of choosing d_i when μ is the value of the parameter of interest, where d_i represents the decision to accept H_i. It seems reasonable to require that the gain in the utility of accepting H_1 be an increasing function of the distance $\delta(\mu, \mu_0)$ between μ and μ_0. For simplicity we set:

$$u(d_1, \mu) - u(d_0, \mu) \;=\; \delta(\mu, \mu_0) - \delta_0, \tag{13}$$

where δ_0 is a constant, which can be interpreted as a penalty for using the more complicated model implied by H_1 when the simpler H_0 would suffice (since $u(d_1, \mu_0) = u(d_0, \mu_0) - \delta_0$). Rejecting H_0 is the optimal decision when it leads to an expected gain in utility:

$$\mathbb{E}\big[u(d_1, \mu) - u(d_0, \mu) \,\big|\, \vec{x}\big] \;>\; 0 \quad \text{or} \quad U(\vec{x}) \;\equiv\; \mathbb{E}\big[\delta(\mu, \mu_0) \,\big|\, \vec{x}\big] \;>\; \delta_0, \tag{14}$$

where the expectation is taken with respect to the posterior distribution of μ. For the Gaussian model (1) used in Lindley's paradox, an appropriate choice of δ is the Mahalanobis distance

$$\delta(\mu, \mu_0) \;=\; \Big(\frac{\mu - \mu_0}{\sigma}\Big)^2, \tag{15}$$

and an appropriate prior for μ is the reference prior, which in this case is the indicator function of the set $\mu \geq \mu_0$. The posterior expected utility can then be written as:

$$U(\vec{x}) \;=\; \int_{\mu_0}^{+\infty} \frac{e^{\frac{1}{2}\left(\frac{\bar{x}_o - \mu}{\sigma/\sqrt{n}}\right)^2}}{\sqrt{2\pi}\,(\sigma/\sqrt{n})\,\Phi(z_0/\sqrt{2})} \Big(\frac{\mu - \mu_0}{\sigma}\Big)^2 d\mu \;=\; \frac{1}{n}\bigg[1 + z_o^2 + \frac{z_o\, e^{-z_o^2/2}}{\sqrt{2\pi}\,\Phi(z_o)}\bigg]. \tag{16}$$

Since one rejects H_0 whenever $U(\vec{x}) > \delta_0$, and since $U(\vec{x})$ is a one-to-one function of z_o, it is possible to choose δ_0 so as to make this procedure identical to the p-value test $p_0 < \alpha$. However, if we consider the situation in Lindley's paradox, where n is increased while z_o stays constant, agreement between the two procedures for a fixed penalty δ_0 can only be achieved if α decreases with n. Thus we are led to the same conclusion that we obtained by considering the Type-II error rate β of test (2).

Note that in this utility based approach it is perfectly possible to use objective priors, even if they are improper. It is also possible to put a finite prior weight on the null hypothesis, thereby obtaining a treatment that mixes the first two cases in our description of possible belief structures for test (6). Further details on this methodology can be found in ref. [6].

2.3 Application to the look-elsewhere effect

To illustrate the discrepancy between p-value and Bayesian measures of evidence, we briefly consider the problem of searching for a resonance peak somewhere in a spectrum of finite width. Since the location of the peak is not known a priori, the significance of an interesting local excess must be corrected for the fact that a background fluctuation like the observation could have occurred *anywhere* in the spectrum. This is the look-elsewhere effect (LEE).

The statistician R.B. Davies computed the LEE correction to p-values in 1987 [7]. Suppose that for each value of the resonance location $\theta \in [A, B]$, the test statistic $S(\theta)$ is (asymptotically) chisquared with s degrees of freedom. Davies derived the following formula for the LEE-corrected tail probability:

$$\mathbb{P}\left[\sup_{A \leq \theta \leq B} S(\theta) \;>\; u\right] \leq \mathbb{P}(\chi^2_s > u) \;+\; \langle N(u) \rangle, \tag{17}$$

where $\langle N(u) \rangle$ is the expected number of upcrossings of the level u by the process $S(\theta)$. LEE-corrected p-values are typically obtained via Monte Carlo simulation, which can be very time consuming for large values of u. Ref. [8] solves this problem by providing an analytical formula for the scaling of $\langle N(u) \rangle$ with u. The computation can then be considerably shortened by performing it at some low value of u and using the formula to extrapolate to the observed value.

For a simple example that doesn't require the full generality of Davies's result, consider the spectrum of observed Poisson counts shown in the left panel of Fig. 1. We assume that the background noise is the same in all bins, and that any signal can only appear in one bin. The p-value in any given bin i is

$$p(n_{o,i}) \;=\; \sum_{n=n_{o,i}}^{\infty} \frac{\mu^n}{n!}\, e^{-\mu}, \tag{18}$$

where μ is the background level and $n_{o,i}$ is the observed count in bin i. We are interested in the most significant effect, as identified by the smallest p-value in the spectrum, say $p_{\min}$. If the total number of bins examined is B, the LEE-corrected significance is:

$$p_{LEE} \;=\; \mathbb{P}\Big[\min_{1\le i\le B} p(n_{o,i}) \le p_{\min} \,\Big|\, H_0\Big] \;=\; 1-(1-p_{\min})^B. \tag{19}$$

Note that p_{LEE} is larger than $p_{\min}$.

The Bayesian calculation starts with the likelihood function:

$$\mathcal{L}(\eta,\ell) \;=\; \prod_{i=1}^{B} \frac{[\mu+\eta\,\delta_{i\ell}]^{n_{o,i}}}{n_{o,i}!}\, e^{-\mu-\eta\,\delta_{i\ell}}, \tag{20}$$

where η is the signal magnitude and ℓ its bin number. We wish to test

$$H_0 : \eta = 0 \quad \text{versus} \quad H_1 : \eta > 0. \tag{21}$$

The problem has one nuisance parameter, the signal location ℓ. In the absence of any information about ℓ, we take its prior to be uniform: $\pi(\ell) = 1/B$. If the value of η was specified under H_1, the posterior probability of H_0 would be:

$$\begin{aligned} p(H_0\,|\,\vec{n}_o)_{LEE} \;&=\; \frac{\pi_0 \sum_\ell \mathcal{L}(0,\ell)\,\pi(\ell)}{\pi_0 \sum_\ell \mathcal{L}(0,\ell)\,\pi(\ell) + (1-\pi_0)\sum_\ell \mathcal{L}(\eta,\ell)\,\pi(\ell)} \\ &=\; \left[1+\frac{1-\pi_0}{\pi_0}\frac{1}{B}\sum_{l=1}^{B}\Big(1+\frac{\eta}{\mu}\Big)^{n_{o,\ell}}\, e^{-\eta}\right]^{-1}, \end{aligned} \tag{22}$$

where π_0 is the prior probability of H_0. How can we handle the fact that η is actually *not* specified under H_1? The preferred option is a subjective Bayesian analysis: introduce a proper prior for η under H_1 and integrate it out. A second option is to do an objective Bayesian analysis by constructing a "neutral" prior for η; however this prior needs to be proper, otherwise the posterior probability of H_0 will be undefined. Methods for doing this are described in ref. [9]. A third option is the utility-based approach of section 2.2.2. Finally, one could simply plot $p(H_0\,|\,\vec{n}_o)_{LEE}$ from equation (22) as a function of η to get a sense of the variation of the Bayesian evidence regarding H_0. This is shown in the right panel of Fig. 1 for $\pi_0 = 1/2$. It is quite remarkable that, *even at its minimum,* the posterior probability of H_0 is still about an order of magnitude higher than the p-value. Of course one could reduce this discrepancy by lowering π_0, but this would mean that a substantial fraction of the evidence against H_0 is due to one's prior opinion about H_0 rather than to the data.

The plot also shows the effect of the LEE correction on both the p-value and the posterior probability. For this particular example, the effect is about the same on both quantities.

3 Measurement sensitivity

So far we have concentrated our attention on the interpretation of evidence supporting discovery claims. For tests such as (6), where the alternative hypothesis specifies a *range* of values for the parameter of interest, an equally important and difficult issue is what to report when no discovery can be claimed. If the p-value (3) is greater than the significance threshold α, we accept H_0. However, this does not mean that all values of μ under H_1' are now rejected: there are values of μ that our experiment is not sensitive to, and others that the data won't allow us to exclude. One way to investigate this is to test individual values of μ under H_1':

$$H_1'[\mu_1] :\; \mu = \mu_1 \quad \text{versus} \quad H_0 :\; \mu = \mu_0, \tag{23}$$

where, as before, $\mu_1 > \mu_0$. A $(1-\gamma)$ C.L. upper limit μ_u can then be defined as the largest value of μ_1 that is not rejected by the test at some significance level γ.

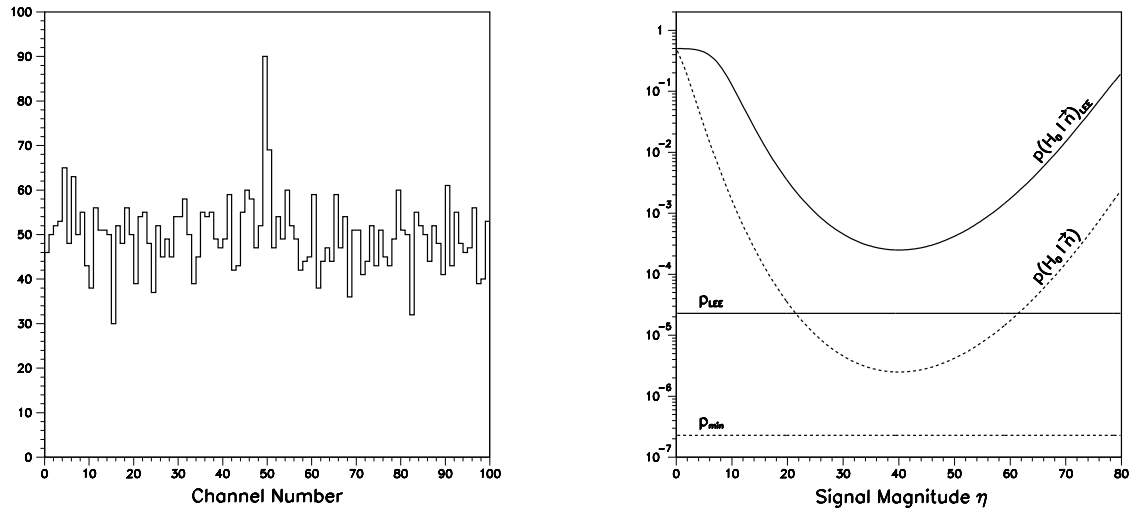

Fig. 1: Left: spectrum of Poisson counts used to illustrate the look-elsewhere effect on p values and posterior probabilities. Right: posterior probability of the background-only hypothesis as a function of the tested signal magnitude η, with and without LEE correction, compared with the corresponding p values.

In the frequentist approach to testing, μ_u can be obtained by solving the p-value equation $p_1(\mu_u) = \gamma$, where

$$p_1(\mu_1) \;=\; \mathbb{P}\big[\bar{X} < \bar{x}_o \,\big|\, H_1'[\mu_1]\big] \tag{24}$$

is the p-value for testing $H_1'[\mu_1]$. For the Gaussian likelihood (1), the upper limit derived this way is given by:

$$\mu_u \;=\; \bar{x}_o + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\gamma). \tag{25}$$

Due to measurement resolution effects it may happen that $\bar{x}_o$ is such that the upper limit μ_u falls below the lower boundary μ_0 of the physical parameter space. In this case the upper limit is unphysical and the corresponding interval is empty: all values of μ are excluded, regardless of the actual measurement sensitivity.

This problem has been known for at least twenty-five years [10]. As recently emphasized by Bob Cousins, the underlying issue is lack of conditioning in the standard frequentist approach, which, in the presence of physical boundaries, yields what is known in the statistics literature as "relevant subsets" [11]. These are subsets of sample space with respect to which the conditional coverage of a confidence interval procedure is consistently above or consistently below the nominal coverage for *all* parameter values.

Many solutions have been proposed over the years. Here we only mention those that are based on solid statistical principles. The first one is to calculate a Bayesian upper limit: the resulting intervals are never empty, but they require the choice of a prior and typically do not achieve exact frequentist coverage. The second solution is to do a frequentist construction with a so-called "unified" ordering rule, such as the likelihood-ratio ordering rule of ref. [12]. This procedure has coverage and never yields empty intervals, but there are cases where the behaviour of interval length as a function of the observations is unsatisfactory. In addition, it only accomodates one confidence level where high energy physicists typically require three: one for the discovery significance ($1.0 - 2.87 \times 10^{-7}$), one for the upper limit (95%) reported in the absence of a discovery claim, and one for the two-sided interval (68%) reported with a discovery claim. A third possibility is to modify the statistical model of the measurement, in particular its error structure [13]. For the Gaussian example, one typically assumes that the standard

deviation is known exactly and is independent of the mean, neither of which may be true. Finally, some astrophysicists have recently proposed to keep reporting the standard frequentist upper limit, but to complement it with a minimum sensitivity bound, defined as the smallest parameter value that one would have a pre-specified probability of detecting at a pre-specified level of significance if it was the true value [14]. As indicated by its definition, the construction of this sensitivity bound requires two pre-specified numbers; in addition, the handling of nuisance parameters is not trivial.

There is at present no consensus on the optimal method.

4 Implicit statistical models

High energy physics measurements are complex in the sense that we typically do not know the exact analytical dependence of the likelihood function on some parameters of the model. All we have is the underlying stochastic mechanism, which we can simulate with a Monte Carlo algorithm. This difficulty occurs for both nuisance and interest parameters.

As illustration, consider the measurement of the mass μ of a new particle. The data sample consists of a signal component (events containing the new particle) and an irreducible background component. If we have an event by event estimator X of μ, the likelihood has the form:

$$\mathcal{L}(\mu) \;=\; \prod_{i=1}^{N} \Big[(1-\epsilon_b)\, f_s(x_i\,;\,\mu) \;+\; \epsilon_b\, f_b(x_i) \Big] \;\times \ldots, \tag{26}$$

where ϵ_b is the background contamination of the sample, and f_s and f_b are the signal and background distributions of X. These distributions are usually approximated by histograms from Monte Carlo simulations, which may be smoothed or fitted with parametric representations. In addition, the f_s distribution must be constructed on a grid of μ values supplemented with interpolation. This is inefficient since a lot of time is wasted modeling $f_s(x\,;\,\mu)$ at μ values far from the maximum-likelihood estimate (MLE). Finally, f_s and f_b also depend on nuisance parameters such as energy scales, initial and final state radiation, parton densities, etc. Generalizing the above approach to multiple parameters quickly becomes impractical [15].

Over the years, a number of ingenious but somewhat dubious shortcuts were invented by high energy physicists to take nuisance parameters into account. An example shortcut is to evaluate the shift $\Delta\mu$ in the MLE of μ induced by a one-sigma variation of a given nuisance parameter, and then to replace the likelihood by its convolution with a Gaussian with standard deviation $\Delta\mu$:

$$\mathcal{L}(\mu) \;\to\; \tilde{\mathcal{L}}(\mu) \;\equiv\; \int \mathcal{L}(\mu') \, \frac{e^{-\frac{1}{2}\left(\frac{\mu-\mu'}{\Delta\mu}\right)^2}}{\sqrt{2\pi}\,\Delta\mu} \, d\mu' \tag{27}$$

When there is more than one nuisance parameter, $\Delta\mu$ is replaced by the sum in quadrature of the individual shifts. The validity of this method has never been studied in detail.

In the next two subsections we examine approaches, one Bayesian and the other frequentist, that may be useful for handling implicit models.

4.1 Approximate Bayesian computation methods

In the Bayesian paradigm, the likelihood is integrated over the nuisance parameters, a feature that lends itself well to Monte Carlo computations. Implicit statistical models can be analyzed with the help of so-called ABC methods (Approximate Bayesian Computation). The goal is to *approximate* the posterior distribution $\pi(\mu \,|\, x) \propto p(x \,|\, \mu)\, \pi(\mu)$. All we need is a suitable distance function $d(x_a, x_b)$ between two datasets x_a and x_b. Let x_{obs} be the observed dataset. The simplest ABC algorithm is the ABC rejection sampler:

1. Sample $\mu^\star$ from $\pi(\mu)$.
2. Simulate a dataset $x^\star$ from $p(x \mid \mu^\star)$.
3. If $d(x_{obs}, x^\star) \leq \epsilon$, accept $\mu^\star$, otherwise reject.
4. Return to step 1.

The output of an ABC algorithm is a sample of parameters $\mu^\star$ from a distribution $\pi(\mu \mid d(x_{obs}, x^\star) \leq \epsilon)$. If ϵ is sufficiently small, this distribution will be a good approximation to the posterior $\pi(\mu \mid x_{obs})$. A delicate issue is the choice of distance function $d(x_a, x_b)$. There is no general theory for this, and the choice must be made on a case-by-case basis.

There exist other ABC algorithms, which are more efficient than the rejection sampler and even work with improper priors [16].

When combining the results from different experiments, common uncertainties and the resulting correlations must be taken into account. This seems doable with ABC methods, although the generation of Monte Carlo samples (an industry in itself) will have to be carefully coordinated between experiments.

4.2 Decision-Theoretic Methods

In the frequentist paradigm one is interested in procedures that have coverage for all values of the interest and nuisance parameters. Other requirements besides coverage are needed to specify unique procedures.

For the construction of confidence intervals, one approach, based on decision-theoretic ideas, is known as minimax expected size (MES): it minimizes the maximum expected size of the confidence set over parameter space. In a Monte Carlo implementation of MES, parameter values are drawn at random from the parameter space, and a dataset is simulated for each parameter value. Each simulated dataset is compared to the observed dataset using a likelihood ratio test. Inverting the likelihood ratio test minimizes the probability of including false values in the confidence region, which in turn minimizes the expected size of the confidence region. This Monte Carlo algorithm does not require explicit knowledge of the likelihood function, only of the data generating mechanism [17]. In addition, it is well suited for handling physical boundaries in parameter space.

At present the Bayesian approach via ABC methods seems a lot more flexible than the above frequentist method, since ABC methods produce an approximation to the posterior itself. The decision-theoretic procedure only produces confidence intervals, and only of the MES type (no choice of ordering rule).

5 Parton Density Function Uncertainties

Currently the parton densities are determined by a fit to ~ 35 datasets with a total of ~ 3000 data points. The standard parametrization uses ~ 25 parameters, and the fit quality is characterized by a χ^2 value. Uncertainties on the parton densities are derived from a $\Delta\chi^2$ procedure, but the standard $\Delta\chi^2 = 1$ rule yields clearly unrealistic uncertainties. Instead, 90% C.L. uncertainties are obtained via $\Delta\chi^2 = 100$ or 50, depending on the group doing the fit.

These uncertainties are not yet understood from a statistical point of view. Some suggestions were made at Banff to improve this situation:

- A decision-theoretic approach such as MES.
  This may be of value for quantifying the uncertainty in the pdf estimates.
- A random effects model.
  Assume that the theory does not quite fit each experiment, resulting in underestimated prediction errors. Propose as solution that the theory parameter is slightly different in each experiment, and all these individual parameters are constrained to the formal parameter of the theory via some distributional assumptions (such as a multivariate-t prior).

– A closure test.
First verify that for data generated from the theoretical distributions, the $\Delta\chi^2 = 1$ criterion yields reasonable uncertainties. Then study how inferences are affected by biasses in theory and/or data.

6 Profile Likelihood Methods

Using results due to Wilks and Wald, ref. [18] derives a comprehensive set of asymptotic formulae, based on the profile likelihood, for use in searches for new physics.

An interesting technique introduced in that paper is the so-called Asimov dataset, which is in a sense the most representative dataset of an ensemble: when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values. Asimov datasets can be used to simplify the estimation of measurement sensitivities and to compute Jeffreys' prior.

7 Reference Priors

Uniform priors have been the norm in high energy physics for a long time, partly because they *seem* reasonable (by the principle of indifference), and partly because the corresponding posterior intervals sometimes exhibit reasonable frequentist behaviour. However, they are also known to suffer from two major drawbacks: they give inconsistent results if the parametrization of the problem is changed, and they are not guaranteed to yield proper posteriors.

Reference priors have been developed over the past thirty years with the aim of providing a "standard" for presenting and comparing measurements of quantities about which little or no prior knowledge is available. Similarly to other standards (e.g. lengths and weights), the reference prior standard was designed with some rational considerations in mind: the algorithm is based on information theory and is very generally applicable; reference posteriors are invariant under one-to-one transformations of the parameter of interest, have good frequentist coverage properties, and avoid the so-called marginalization paradoxes that plague other non-informative constructions. In high energy physics, reference priors are now available for cross section measurements, when partial information is available for acceptances and background sources [19].

There are still some important issues however:

1. Can reference posterior inferences be reported by themselves, or should they be reported only as part of a sensitivity analysis? If the latter, how should one choose alternative priors?
2. The general definition of reference priors involves the taking of limits, and this must be done carefully in order to avoid infinities. The standard approach is to use sequences of nested compact sets that converge to the whole parameter space. Unfortunately there is no unique way of choosing these compact sets, and there is no guarantee that different choices lead to the same result, or even that all choices lead to a proper posterior. This ambiguity prevents us from designing a completely general numerical algorithm.
3. How should we handle implicit statistical models? Can we combine ABC methods with numerical algorithms for computing reference posteriors?

8 Extreme value theory

Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed random variables. Whereas central limit theory is concerned with the behavior of the partial sums $X_1 + X_2 + \ldots + X_n$ as $n \to \infty$, extreme value theory studies the behavior of the sample extremes $\max\{X_1, X_2, \ldots, X_n\}$ as $n \to \infty$. This theory has many applications, for example to the question of how high dikes should be built in the Netherlands to protect land below sea level from storm surges that drive the seawater level up along the coast.

In high energy physics we are often interested in extreme events, that is, collision events in which some measurable quantity takes on a very large value. Extreme value theory may help here, by providing a solid basis for extrapolating from measurements at lower values of the quantity of interest [20].

9 Acknowledgements

I wish to thank the organizers and participants of both the 2010 Banff meeting and the 2011 Phystat workshop for many interesting talks and productive discussions.

References

[1] 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.

[2] D. V. Lindley, "A statistical paradox," Biometrika **44**, 187 (1957).

[3] See for example G. Shafer, "Lindley's paradox," (with discussion) J. Amer. Statist. Assoc. **77**, 325 (1982).

[4] J. O. Berger and M. Delampady, "Testing precise hypotheses," Statist. Sci. **2**, 317 (1987); `http://www.stat.duke.edu/~berger/papers/p-values.pdf`.

[5] M. J. Bayarri, "Comment," Statist. Sci. **2**, 342 (1987).

[6] J. Bernardo, these proceedings. Note that Bernardo uses a loss function instead of our utility function; one is simply the negative of the other.

[7] R. B. Davies, "Hypothesis testing when a nuisance parameter is present only under the alternative," Biometrika **74**, 33 (1987).

[8] E. Gross and O. Vitells, "Trial factors for the look-elsewhere effect in high energy physics," Eur. Phys. J. C**70**, 525 (2010); arXiv:1005.1891v3 [physics.data-an].

[9] J. O. Berger, "A comparison of testing methodologies," *PHYSTAT LHC Proceedings on "Statistical issues for LHC physics,"* CERN Yellow Report CERN-2008-001 (7 March 2008), pg. 8.

[10] V. L. Highland, "Estimation of upper limits from experimental data," Temple University preprint C00-3539-38 (1987).

[11] R. J. Buehler, "Some validity criteria for statistical inferences," Ann. Math. Statist. **30**, 845 (1959).

[12] G. J. Feldman and R. D. Cousins, "Unified approach to the classical statistical analysis of small signals," Phys. Rev. D **57**, 3873 (1998).

[13] G. Casella, "Comment," Statist. Sci. **17**, 159 (2002).

[14] V. L. Kashyap *et al.*, "On computing upper limits to source intensities," Astrophysical J. **719**, 900 (2010); arXiv:1006.4334v1 [astro-ph.IM]. Caveat: the terminology of this paper interchanges the concepts of *upper limit* and *upper bound* as understood in high energy physics.

[15] P.J. Diggle and R.J. Gratton, "Monte Carlo methods of inference for implicit statistical models," J. R. Statist. Soc. B**46**, 193 (1984).

[16] T. Toni *et al.*, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," J. R. Soc. Interface **6**, 187 (2009).

[17] C. M. Schafer and P. B. Stark, "Constructing confidence regions of optimal expected size," J. Amer. Statist. Assoc. **104**, 1080 (2009); `http://www.stat.cmu.edu/~cschafer/cmspbs.pdf`.

[18] G. Cowan *et al.*, "Asymptotic formulae for likelihood-based tests of new physics," arXiv:1007.1727v2 [physics.data-an] (2010).

[19] L. Demortier, H. B. Prosper, and S. Jain, "Reference priors for high energy physics," Phys. Rev. D **82**, 034002 (2010).

[20] L. de Haan and A. Ferreira, "Extreme value theory: an introduction," Springer (2006).