# Highlights from PHYSTAT 2011

*Glen Cowan*
Physics Department, Royal Holloway, University of London, Egham TW20 0EX, UK

**Abstract**

The PHYSTAT 2011 Workshop held at CERN from 17-20 January, 2011, brought together particle physicists working on statistical data analysis together with astrophysicists and statisticians to exchange ideas and report on recent developments. Highlights from the first three days of the meeting are summarized here. The fourth day, devoted to the problem of unfolding (deconvolution) is covered elsewhere in these proceedings.

## 1 Introduction

The highlights from PHYSTAT 2011 fall into several broad categories: frequentist methods, Bayesian methods, and tools and applications, reviewed below in Sections 2, 3 and 4, respectively. Apologies are extended in advance for any personal bias or imbalance in the emphasis of topics covered.

## 2 Frequentist methods

The frequentist methods discussed at PHYSTAT 2011 include use of order statistics for discovery [1], issues related to setting limits [2, 4], and treatment of systematic uncertainties using profile or marginal likelihoods [5, 6]. Additional contributions in this area covered the multiple testing problem or "look-elsewhere effect" as well as methods for combining results [7, 8].

### 2.1 Order statistics for discovery

Statistical tests for the discovery of new physics have often focused on specific signal models, and the test is thus optimal if the new model is correct. One runs the risk, however, of only discovering the types of phenomena that have been thought of in advance. It is therefore important to carry out some tests that probe the data in a more general way and explore departures from the Standard Model expectations that are not motivated by specific alternatives.

An example of this was presented by Cox [1], who proposes using order statistics as a basis for various tests. Suppose one carries out a test at $n$ positions (these could be, e.g., the $n$ bins of a histogram), and obtains as a result a set of $n$ $p$-values $P_1, \ldots, P_n$. These can be transformed using $Z = -\ln P$ and then ordered, i.e., $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)} = \mathrm{max} Z_j$.

A plot of the ordered $Z$ can then be used descriptively and forms the basis for formal tests. Under the null hypothesis, the $Z$ values form a straight line of unit slope. A single outlying point thus indicates an easily identifiable alternative. An incorrectly modeled shape for the histogram would lead to a smooth curve, and if the bins are correlated then one obtains a straight line but with a slope different from unity.

Such a method has a clear application in Particle Physics in a search for a bump in a histogram. Here, however, one would need to use a modified version of the test where the departure from the null (i.e., the new signal) is smeared out over several adjacent bins.

### 2.2 Frequentist limits

Before the existence of a given signal process is well established, it is often of interest to test the signal model using different values of its parameters and to see which values can be excluded. Specifically, one is often interested in testing parameters related to the overall rate of the signal process, and seeing which values can be excluded on the grounds that the predicted rate is too high relative to what is observed in the data. This corresponds to using a one-sided test to obtain an upper limit on the rate.

A long recognized difficulty with such one-sided tests is that effectively all physically allowed values of the signal rate may be excluded. For any unbiased test the probability to reject a given signal rate under the assumption of the background-only hypothesis is at least equal to the size of the test, e.g., $\alpha = 0.05$. This holds true even for very small signal rates, that is, for signal models to which one effectively has no sensitivity. If the number of events in data fluctuates low relative to the expected background, then one may end up rejecting even a very low signal rate and thus setting a limit that is substantially smaller than the intrinsic resolution of the measurement.

Already in the era of the LEP Higgs searches, the CLs procedure [9] was developed to prevent data fluctuations from leading to unrealistically strong limits. Here, the $p$-value of the hypothesized signal model is divided by one minus the $p$-value of the background-only hypothesis, and the signal model is only excluded if this ratio is found below a small threshold $\alpha$. The threshold thus plays the role of the significance level of the test, but it is not quite the same because the ratio of $p$-values is necessarily greater than the $p$-value of the signal model (the numerator of the ratio). Thus one is less likely to exclude a given signal rate and the CLs upper limit is in general higher than the corresponding limit based on a simple one-sided test.

More recently, as mentioned by Demortier [2], it has been suggested to only regard a parameter value as excluded if its $p$-value is below the test size $\alpha$ and also if one has sufficient sensitivity to that value. As a measure of sensitivity, one can require a certain minimum probability of discovering the signal (i.e., rejecting the background-only model) if it is true. This is essentially the power of a test of the background-only hypothesis with respect to the signal alternative, hence the name “Power-Constrained Limits” or PCL.

Alternatively one may take as the measure of sensitivity the power of a test of the signal model with respect to the background-only alternative. This is the approach recently used by the ATLAS Collaboration [3].

A similar approach is taken in the method described by van Dyk [4, 10] for reporting the results of a search for an astrophysical source. The proposed procedure is to give: (1) whether the source was detected, (2) a confidence interval for the source intensity and (3) the sensitivity of the observation, quantified as the minimum source strength for which one would have a given detection probability. Van Dyk has emphasized the importance of communicating both the usual upper limit (called an upper bound in the astrophysics community) as well as the sensitivity, rather than only the maximum of these two numbers.

### 2.3 Systematic uncertainties in likelihood-based tests

An important element of any analysis, frequentist or Bayesian, is ensuring that the model adequately describes the data. This means that one must have an accurate representation for the probability of an outcome, say, $x$, as a function of the model parameters $\theta$. If this model is not sufficiently accurate, then the situation can be improved by including additional nuisance parameters into the model. These provide an added degree of flexibility so that for some point in the enlarged parameter space, the model will be closer to the truth. Of course as more nuisance parameters are included, one’s sensitivity to the parameters of interest is diminished.

Conway [5] and Röver [8] brought up an important issue concerning two distinct methods for dealing with nuisance parameters: profiling and marginalization. Suppose originally one measures $x$, and the experiment is modeled with the likelihood $L(x|\theta)$, where $\theta$ is a parameter of interest. Now it may be that the model is found to be inadequate, and so it is enlarged by inserting a nuisance parameter $\nu$, so one now has the likelihood $L(x|\theta, \nu)$.

To constrain the nuisance parameter, one may set up a control measurement $y$ with likelihood

$L(y|\nu)$. This now becomes part of the full likelihood. If $x$ and $y$ are independent, this is simply

$$L(x, y|\theta, \nu) = L(x|\theta, \nu)L(y|\nu) \; . \tag{1}$$

To eliminate the nuisance parameter $\nu$, one can form the profile likelihood

$$L_{\mathrm{p}}(x, y|\theta) = L(x, y|\theta, \hat{\hat{\nu}}(\theta)) \; , \tag{2}$$

where the value $\hat{\hat{\nu}}(\theta))$ is the value of $\nu$ that maximizes the likelihood for the specified $\theta$.

Alternatively, in the Bayesian approach one can regard the measurement $y$ as supplying the prior information about the nuisance parameter $\nu$. This prior can be written

$$\pi(\nu) \propto L(y|\nu)\pi_0(\nu) \; , \tag{3}$$

where $\pi_0(\nu)$ reflects prior knowledge about $\nu$ even before the control measurement $y$. (It could be called the ur-prior, using the German prefix for original or primordial.) In the Bayesian approach one eliminates the nuisance parameter by integrating to find the marginal likelihood,

$$L_{\mathrm{m}}(\theta) = \int L(x|\theta, \nu)\pi(\nu) \, d\nu \; . \tag{4}$$

The point to notice here is that the observed value of $y$ is taken once to determine the prior $\pi(\nu)$, but is then not viewed as a quantity that varies upon repetition of the experiment. Thus if one simulates measurements based on the model (4), it is only $x$ that is generated. In contrast, the profile likelihood (2) models both the measurements $x$ and $y$. One must therefore simulate both of these values to determine the distribution of a statistic based on $L_{\mathrm{p}}(\theta)$.

It is easy to show that in simple cases, e.g., Gaussian measurements and a constant ur-prior, the two approaches (profiling and marginalization) are equivalent. And in the examples shown by Conway [5] and Röver [8], essentially no difference between the two methods can be seen. Trotta and Cranmer [11], however, discussed cases where this is not true.

A further important example where the two methods are not equivalent is when the main measurement $x$ is discrete, and the control measurement $y$ is continuous. For example, $x$ could represent a Poisson-distributed number of events, and $y$ could be a correction factor related to the efficiency, modeled as following a Gaussian distribution. The distribution of a test statistic based on the marginal likelihood (4), will be discrete, since the dependence on the continuous value $y$ has disappeared after integration over the nuisance parameter. Therefore confidence intervals based on the marginal likelihood will suffer from the over-coverage that is well known in discrete problems. A statistic based on the profile likelihood (2), however, follows a continuous distribution, because it retains a dependence on the continuous variable $y$. Thus the over-coverage due to discreteness is absent, which should be regarded as an advantage of this approach.

A further important aspect of tests based on the profile likelihood ratio is that one can use analytic formulae to approximate the distributions needed to carry out statistical tests. The formulae are exact only in the large sample limit, but for practical examples the approximations were shown to be reasonably accurate for surprisingly small data samples [6].

### 2.4 The look-elsewhere effect

Important progress was reported by Vitells [7] and Ranucci [12] on the problem of multiple testing, usually called in particle physics the 'look-elsewhere effect'. 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 $2.9 \times 10^{-7}$, corresponding to a $5\sigma$ 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., [6]). But this 'fixed-position' $p$-value would not be correct in general, as it assumes the position of the peak was known in advance.

Vitells 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 small Monte Carlo calculation. Suppose a test statistic $q_0$ is observed to have a value $u$, and the model contains a nuisance parameter $\theta$ (such as the peak position) which is only defined under the signal model (there is no peak in the background-only model). Then Vitells and Gross [13] find that the desired $p$-value can be written

$$P(q_0 > u) \approx N_1 e^{-u/2} + \frac{1}{2} P(q_0(0) > u) , \tag{5}$$

where $P(q_0(0) > u)$ is the 'fixed-position' $p$-value, and $N_1$ is the mean number of 'upcrossings' of the the statistic $q_0$ above the level $u$. The value of $N_1$ can be estimated by finding the number of upcrossings above some much lower value, $u_0$, from a relation due to Davis [15],

$$N_1 \approx \langle N_{u_0} \rangle e^{u_0/2} . \tag{6}$$

By choosing $u_0$ sufficiently low, the value of $N_1$ can be estimated by simulating, say, only 100 experiments, rather than the $10^8$ needed for a $5\sigma$ discovery.

Gross and Vitells 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 [14]. 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 fixed-position $p$-value can be reported along with an approximate correction to account for the range of parameter space in which the effect could have appeared.

Ranucci [12] also reported on the analogous problem in a time-series analysis. That is, if one examines any time series long enough, a feature that appears significant will eventually appear. The methods proposed for dealing with this problem are similar to those reported by Vitells.

## 3 Bayesian methods

In Bayesian statistics, probabilities are assigned to hypotheses (e.g., parameter values), in contrast to the frequentist approach where one only speaks of the probability of (repeatable) data outcomes. Given, say, a parameter $\theta$ and data $x$, Bayes' theorem is used to find the posterior probability of $\theta$ given $x$,

$$p(\theta|x) \propto L(x|\theta)\pi(\theta) \ , \tag{7}$$

where $L(x|\theta)$ is the likelihood and $\pi(\theta)$ is the prior probability. An important difficulty in the Bayesian approach stems from the requirement to supply priors. Although one may wish in cases to have these reflect a complete lack of prior information, it has long been realized that this is not a uniquely defined concept.

For example, Bayesian methods have a long tradition in particle physics for the problem of a Poisson distributed value $n$ used to make inference about the mean $\mu$. A widely used prior pdf for $\mu$ is the (improper) constant prior for $\mu \geq 0$, which is often thought of as reflecting a complete lack of prior knowledge about $\mu$. This is not really true, as it is not invariant under a change of parameter (e.g., it is not flat in $\ln \mu$), and furthermore it cannot possibly represent a meaningful degree of belief. Nevertheless it provides a simple benchmark and has been widely used.

A. Caldwell [16] proposed elicitation of prior probabilities through consensus of the particle physics community. As impossible as this task may sound, it could prove to be an interesting exercise and may well result in useful benchmarks.

In the absence of meaningfully usable prior information one may try to determine priors from formal rules, as described in the review by Kass and Wasserman [19]. A pioneering element of this approach is the reference prior due to J. Bernardo and J. Berger [20]. These were addressed by several speakers at this meeting and some important points are summarized in Sec. 3.1.

A further important element of Bayesian statistics that has not yet found wide application in particle physics is Bayesian model selection, which was addressed by Berger [18] and reviewed below in Sec. 3.2. Implicit Bayesian methods were described by Demortier [2] and the example of Approximate Bayesian Computation (ABC) is summarized in Sec. 3.3.

### 3.1 Reference priors

The particle physics community has been reluctant to assign subjective prior probabilities to important model parameters, no doubt driven by the desire to remain 'objective' and also because of the difficulties in reaching any sort of consensus on what these probabilities should be. A general prescription for prior probabilities from formal rules is thus very attractive to the community, and so the method of reference priors due to Bernardo and Berger [20] can perhaps provide a way forward.

As described by Bernardo [17], Demortier [2] and Pierini [21], to find the reference prior for a given problem one begins by considering the Kullback-Leibler divergence of the posterior $p(\theta|x)$ relative to a prior $\pi(\theta)$, obtained from the data $\vec{x} = (x_1, \ldots, x_n)$, which are assumed to consist of $n$ independent and identically distributed values of a random variable $x$:

$$D_n[\pi, p] = \int p(\theta|\vec{x}) \ln \frac{p(\theta|\vec{x})}{\pi(\theta)} \, d\theta \ . \tag{8}$$

This is effectively a measure of the gain in information provided by the data. The reference prior is chosen so that the expectation value of this information gain is maximized for the limiting case of $n \to \infty$, where the expectation is computed with respect to the marginal distribution of the data,

$$p(\vec{x}) = \int L(\vec{x}|\theta)\pi(\theta) \, d\theta \ . \tag{9}$$

The techniques for finding reference priors are relatively straightforward for the case of a single parameter, where it turns out to be the same as the well-known Jeffreys prior. Finding a general algorithm suitable for the multiparameter case, however, proves to be problematic. In particular the multiparameter reference prior can in general depend on the ordering of the parameters. Further discussion and applications to particle physics problems can be found in Ref. [22].

The interpretation of the posterior probabilities derived from reference priors is the subject of some debate. One may, for example, derive the result, then disregard its Bayesian origins and simply exploit its frequentist properties. For example, one can use the posterior pdf to derive an interval for the parameter, which will than have a certain probability to cover to the true parameter value in the same sense as a frequentist confidence interval. One may also use a reference prior as part of a sensitivity analysis, i.e., a study of how the posterior probabilities change under variation of the prior. These questions will no doubt receive further study should reference priors find wider application in particle physics.

### 3.2 Bayesian model selection

In the particle physics community, the usual frequentist measure of significance for establishing a discovery has been based on the $p$-value of the background-only hypothesis. That is, one gives the probability, under assumption of no new signal, to see data as signal-like as what was actually observed or more so. This is of course not exactly what one wants, which would more naturally be the probabilities of the background-only model or various signal models. The fact that the $p$-value is often confused for the probability of the no-signal model only makes matters worse.

The natural substitute for a $p$-value in the Bayesian paradigm is the posterior probability that signal is present or absent given the data. But this requires the prior probability for the hypotheses, and this is where constructing a result that is of value to the broader scientific community becomes difficult. What, after all, is the prior probability for the existence of the Higgs boson? Or of supersymmetry, or some other perhaps very speculative extension to the Standard Model? These things are highly subjective, and mixing them into the reporting of an experimental result cannot help matters.

As discussed by Berger [18], however, one can summarize an experimental result by use of a Bayes factor, $B_{01}$, which quantifies the degree to which one of two hypotheses, $H_0$ or $H_1$, is preferred by the data. This requires no overall prior probabilities for $H_0$ or $H_1$, but priors must be given for all of the internal parameters of the two models.

For a pair of hypotheses $H_0$ and $H_1$ the Bayes factor is defined as the posterior odds divided by the prior odds,

$$B_{01} = \frac{P(H_0|x)}{P(H_1|x)} \frac{\pi_1}{\pi_0} = \frac{P(x|H_0)}{P(x|H_1)} \; . \tag{10}$$

Here $x$ refers to the data and $\pi_i$ $(i = 0, 1)$ are the prior probabilities. That is, $B_{01}$ is the same as the posterior odds if one were to assume equal prior probabilities, and it is thus an indicator of which model is preferred by the data. The second equality in (10) follows from Bayes' theorem, and therefore the Bayes factor is also equal to the ratio of likelihoods.

If a model contains any internal parameters, then to obtain the likelihood these must be characterized by a meaningful prior pdf and marginalized, i.e.,

$$P(x|H_i) = \int P(x|H_i, \theta_i) \pi_i(\theta_i) \, d\theta_i \; , \tag{11}$$

where $\theta_i$ are the internal parameters for model $H_i$ $(i = 0, 1)$ and $\pi_i(\theta_i)$ is the corresponding prior pdf. It is important to note that in this case the prior pdf cannot be improper, as this would only be defined up to an arbitrary constant and the Bayes factor would not be well defined. Furthermore, if an improper prior is made proper by imposing a cut-off, then the Bayes factor will retain a dependence on this cut-off. Thus all internal parameters of the models must be characterized by meaningful, proper priors.

As an example, Berger examined the problem of a number of events $N$, assumed to be Poisson distributed with a mean $s+b$, where $s$ and $b$ are the contributions from signal and background processes, respectively. The Bayes factor for an observed value $N$ is

$$B_{01}(N) = \frac{\text{Poisson}(N|0+b)}{\text{Poisson}(N|s+b)\pi(s)\,ds} = \frac{b^N e^{-b}}{\int_0^\infty (s+b)^N e^{-(s+b)}\pi(s)\,ds} \tag{12}$$

In both numerator and denominator, the probabilities must be integrated over all internal parameters of the two models. In this example, this is only relevant in the denominator, where one has the signal parameter $s$, characterized by a prior pdf $\pi(s)$.

The prior $\pi(s)$ could be chosen subjectively, but for most problems there would be no consensus for what to use. One could show the result for a variety of subjective choices, which would convey some feel for how important prior information is in the given problem. Alternatively one could use what is called the 'intrinsic prior', which for this problem is $\pi^I(s) = b(s+b)^{-2}$.

Finally, one would use the Bayes factor for discovery in a manner analogous to how a frequentist would use the $p$-value for the background-only hypothesis. That is, for a sufficiently small value of $B_{01}$ or the $p$-value, one would reject the background-only model. The numerical values of cannot be directly compared, however, as illustrated by Berger in the example below.

Taking $N = 7$ and $b = 1.2$ gives a $p$-value of 0.00025, and as this number is very small one naturally thinks the probability of $s = 0$ must therefore also be small, and there can be a great temptation to identify the two numbers at least symbolically. But using the intrinsic prior in this case gives a Bayes factor $B_{01} = 0.0075$, i.e., a factor of 30 greater than the $p$-value.

That $B_{01}$ is substantially greater than the $p$-value for this problem cannot be pinned on the prior used. One can place a lower bound on the Bayes factor by making the prior $\pi(s)$ a delta function centred on the ML estimator, $\hat{s} = \max(0, N-b)$. In this case one finds $B_{01} = 0.0014$, still quite a bit larger than the $p$-value of 0.00025. So the lesson of the exercise is that the smallness of the $p$-value cannot be mentally transferred so easily onto a similarly small probability for the hypothesis.

In fact in many problems it may be more useful to report the Bayes factor as a function of a parameter rather than integrating over it. For example, rather than $B_{01}$, one could show $B_{0s}$ as a function of $s$. Of course in problems with a larger number of parameters this may become impractical.

An important impediment to the use of Bayes factors, however, is related to numerical challenges in computing the required marginal likelihoods represented by Eq. (11). One approach mentioned at the meeting [11, 23] involves a tool developed in the astrophysics community called nested sampling [24]. The key is to reparametrize the problem by defining

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)\,d\theta\ , \tag{13}$$

so that the desired integral can be written

$$\int L(\theta)\pi(\theta)\,d\theta = \int_0^1 X(\lambda)\,d\lambda\ . \tag{14}$$

An implementation of the nested-sampling algorithm is available in the `MultiNest` package [25].

Finally one should note that the Bayes factor is in many ways more intuitive than the $p$-value, as it addresses more directly the question of which model one believes to be true. For many years the particle physics community has used $p = 2.9 \times 10^{-7}$ for the $p$-value of the background-only model as a discovery threshold (a $5\sigma$ effect). But one's readiness to announce a discovery should surely depend on factors such as the degree to which the data are better described by an alternative model, and this is directly found in the Bayes factor.

### 3.3 Approximate Bayesian Computation (ABC)

The methods for dealing with systematic uncertainties often rely on having a parametric model containing corresponding nuisance parameters. In the frequentist framework, tests based on the profile likelihood can be used to eliminate the nuisance parameters, or in the Bayesian approach one assigns appropriate priors and marginalizes.

In many cases, however, a parameter $\mu$ may appear in a Monte Carlo model for a given process, but one does not have a parametric function for the probability of the data $x$ given $\mu$. Demortier described an approach known as Approximate Bayesian Computation (ABC), in which one can approximate the posterior probability $p(\mu|x)$ without requiring direct access to $p(x|\mu)$ [2].

First the prior pdf for $\mu$, $\pi(\mu)$, is sampled to obtain a value $\mu$, and then this is used in the Monte Carlo model to generate a data set $x^*$. One then computes a distance measure that quantifies the separation between $x^*$ and the data actually observed. If this distance is less than a given threshold the simulated $\mu$ is accepted, otherwise it is rejected. The distribution of accepted $\mu$ values is then used as an approximation for the posterior probability $p(\mu|x)$.

In principle this method could be used when combining measurements from two different experiments, although in that case the generation of Monte Carlo events would have to use the same parameter values and therefore some coordination would be necessary. ABC methods represent in any case an interesting approach that can address an important need in particle physics.

## 4 Applications and tools

The many contributions at PHYSTAT 2011 on statistical tools and applications clearly showed the community's increase in sophistication since the first PHYSTAT meeting more than ten years ago.

Cranmer [23] described the preparation for combination of the searches for the Higgs boson by ATLAS and CMS. This will exploit the full likelihood representing the joint outcomes of both experiments, with proper treatment of common nuisance parameters. The resulting model can be used in a variety of ways, such as in tests based on the profile likelihood or in a Bayesian analysis. The software for this task is being developed as part of the `RooStats` package [26, 27]. Studies on the combination of different decay channels shown by Zhukov [28] further illustrated the power of this software.

The Bayesian Analysis Toolkit (BAT) described by Pashapour [29] is a package designed for Bayesian computation, specifically, Markov Chain Monte Carlo integration for marginalization of posterior probabilities, and includes automated handling of tasks such as convergence diagnostics. As the user base for this package grows it will be important to include other aspects of Bayesian analyses, such as computation of marginal likelihoods (needed for Bayes factors) as well as support for various types of priors, particularly the reference priors mentioned in Sec. 3.1.

Prosper summarized lessons from the Tevatron [30]. The developments include multiparameter Bayesian analyses resulting in posterior densities for measured cross sections as well as searches for single top-quark production based on sophisticated multivariate classifiers. It will be interesting to see what role such classifiers will play in searches at the LHC, since their increased sensitivity can come with a loss of transparency. A $5\sigma$ signal from a Boosted Decision Tree may initially be met with some skepticism unless it is backed up by $4\sigma$ evidence from a cut-based analysis, and so the team that pursues both approaches may win the competition.

Among the most encouraging reports at the meeting were those on statistical practice by the LHC experiments, from which a rapid flow of publications is now emerging. The talks by Harel (CMS) [31], Casadei (ATLAS) [32] and Morata (LHCb) [33] show that different analysis groups are following different routes, and a movement towards some level of uniformity may take some time to achieve.

S. Forte reported on quantifying uncertainties related to parton densities [34], demonstrating that theorists as well as experimentalists are involved in application of statistical methods. Forte presented an analysis of the NNPDF Collaboration in which neural networks are used to parametrize parametrize parton densities. The increased flexibility of the neural network relative to the parametric functions used by other groups is found to provide a more satisfactory assessment of parton uncertainties.

As in past meetings, the PHYSTAT workshop provided an important opportunity to learn from colleagues in other fields about their statistical practice. Röver [8] and Sardy [35] reported on searches for gravitational waves. The analyses employ regularization methods to suppress noise that involve a bias-variance trade-off that is similar to what particle physicists encounter when unfolding a distribution for effects of limited resolution.

Lahav [36] summarized astrophysical applications, where one finds a far greater use of Bayesian methods than currently seen in particle physics. HEP can learn from the astrophysics community about tools such as nested sampling for computing the marginal likelihoods needed in Bayesian model selection. An exoplanet search provided an outstanding example for particle physicists of how to present clearly a result obtained from a variety of Bayesian priors.

Finally, congratulations to the winners, and thanks to the organizers, of the Banff Challenge 2a, which was reported by Junk [37]. This addressed a number of tricky issues, including the look-elsewhere effect and poorly constrained nuisance parameters. We look forward to the next round.

## 5 Outlook and conclusions

It is clear that great progress has been made in the methods and software used in HEP since the first 'Confidence Limits' workshop at CERN in 2000. Use of sophisticated multivariate classifiers has become an industry, both frequentist and Bayesian approaches for dealing with systematic uncertainties have made important advances, and new software tools allow experiments to combine results in a way that fully exploits the available information and correctly accounts for common systematics. The 'look-elsewhere' effect has long been a serious problem and the contributions seen at this meeting go a long way towards solving it.

There are also many areas where progress is ongoing, such as in finding Bayesian reference priors for important HEP problems, and developing new, physically motivated ways to improve models so as to account for systematic effects. The issue of how best to report limits and intervals has still not found a fully satisfactory solution, but at least the tools are becoming available that allow a variety of approaches to be pursued easily.

The HEP community continues to cling to a $5\sigma$ discovery threshold, that is, a $p$-value of $2.9\times10^{-7}$. As pointed out by van Dyk, this can be viewed as sweeping problems such as the look-elsewhere effect and poorly understood systematics under the rug. This is one of many issues that one hopes will be revisited as real discoveries from the LHC begin to arrive.

## Acknowledgements

I thank the speakers of PHYSTAT 2011 for their highly interesting contributions. I am grateful to the non-particle physicists and the statisticians for sharing their insights and expertise. And finally I thank the organisers, especially Louis Lyons, Albert de Roeck and Michelangelo Mangano, for arranging such a stimulating and productive meeting.

## References

[1] D. Cox, these proceedings.

[2] L. Demortier, these proceedings.

[3] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Power-Constrained Limits*, arXiv:1105.3166 (2011).

[4] D. van Dyk, these proceedings.

[5] J. Conway, these proceedings.

[6] G. Cowan, these proceedings; see also G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C (2011) 71:1554; arXiv:1007.1727.

[7] O. Vitells, these proceedings.

[8] C. Röver, these proceedings.

[9] T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

[10] Vinay L. Kashyap et al., *On Computing Upper Limits to Source Intensities*, Astrophysical Journal, 719, 900-914 (2010); arXiv:1006.4334.

[11] R. Trotta, these proceedings.

[12] G. Ranucci, these proceedings.

[13] E. Gross and O. Vitells, Eur. Phys. J C70 (2010) 525-530; arXiv:1005.1891.

[14] E. Gross and O. Vitells, *Estimating the significance of a signal in a multi-dimensional search*, arXiv:1105.4355.

[15] R.B. Davis, Biometrika 74, (1987) 33-43.

[16] A. Caldwell, these proceedings.

[17] J. Bernardo, these proceedings.

[18] J. Berger, these proceedings.

[19] R.E. Kass and L. Wasserman, J. Amer. Statist. Assoc. 91 (1996) 1343.

[20] J.M. Bernardo, J. Roy. Statist. B 41 (1979) 113–147; J.M. Bernardo and J.O. Berger, J. Amer. Statist. Assoc. 84 (1989) 200–207. See also J.M. Bernardo, *Reference Analysis*, in *Handbook of Statistics*, 25 (D.K. Dey and C.R. Rao, eds.), 17–90, Elsevier, 2005, and references therein.

[21] M. Pierini, these proceedings.

[22] L. Demortier, S. Jain and H. Prosper, Phys. Rev. D Phys. Rev. D 82, 034002 (2010); arXiv:1002.1111.

[23] K. Cranmer, these proceedings.

[24] J. Skilling, Bayesian Analysis (2006) 1, Number 4, pp. 833–860.

[25] F. Feroz, M.P. Hobson and M. Bridges, Mon. Not. Roy. Astron. Soc., 398, 4, 1601-1614 (2009); arXiv:0809.3437.

[26] L. Moneta, K. Belasco, K. Cranmer *et al.*, "The RooStats Project," proceedings of ACAT, 2010, Jaipur, India [arXiv:1009.1003 [physics.data-an]]. [https://twiki.cern.ch/twiki/bin/view/RooStats/]

[27] G. Schott, these proceedings.

[28] V. Zhukov, these proceedings.

[29] S. Pashapour, these proceedings.

[30] H. Prosper, these proceedings.

[31] A. Harel, these proceedings.

[32] D. Casadei, these proceedings.

[33] J. Morata, these proceedings.

[34] S. Forte, these proceedings.

[35] S. Sardy, these proceedings.

[36] O. Lahav, these proceedings.

[37] T. Junk, these proceedings.