CERN–2008–001
7 March 2008

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

PHYSTAT LHC Workshop
on
Statistical Issues for LHC Physics

CERN, Geneva, Switzerland, 27–29 June 2007

Proceedings

Editors: H. B. Prosper
L. Lyons
A. De Roeck

GENEVA
2008

CERN–520 copies printed–March 2008

Abstract

A PHYSTAT workshop on the topic of *Statistical issues for LHC physics* was held at CERN. The workshop focused on issues related to discovery that we hope will be relevant to the LHC. These proceedings contain written versions of nearly all the talks, several of which were given by professional statisticians. The talks varied from general overviews, to those describing searches for specific particles. The treatment of background uncertainties figured prominently. Many of the talks describing search strategies for new effects should be of interest not only to particle physicists but also to scientists in other fields.

Preface

The PHYSTAT LHC Workshop took place at CERN in June 2007. It brought this series of meetings back to its birthplace — the first one was at CERN in January 2000 on Confidence Limits. Subsequent meetings were held at Fermilab (March 2000), Durham, UK (April 2002), SLAC (March 2003), Oxford (September 2005), and Banff (July 2006). The Fermilab workshop was also on confidence limits, but the Durham, SLAC, and Oxford meetings were on a wide range of statistical issues in particle physics as well as in astrophysics and cosmology.

The Programme Committee at the Oxford conference in 2005 considered what the future of these meetings should be. Rather than have another one which dealt with any statistical issues in the relevant field, it was felt that it would be better to change in two ways. The first was to have general educational sessions on practical statistics, mainly for graduate students. It was felt that a whole summer school devoted to statistical problems would not be very attractive, and it would be better to try to persuade existing summer schools to include a few lectures on statistics. The second approach was to have workshops devoted to more specific statistical issues. The Banff meeting thus addressed just three topics: upper limits, separating signal from background, and discovery issues. In a similar spirit, the CERN LHC workshop concentrated on statistical issues to do with discovery claims. It is hoped that this will be relevant for the data that will be accumulated when the Large Hadron Collider starts running in 2008; and when astrophysics facilites such as GLAST commence operation. It was somewhat unfortunate that, because of the narrow focus of the meeting, the Programme Committee had to turn down some potentially interesting talks of a more general nature.

Over 200 people attended the meeting, the majority being experimental physicists. Because the meeting concentrated on a single topic, it was felt inappropriate to have parallel sessions. There were vigorous discussions, prompted by various statistical approaches to the same problem, and by the hope that the real data could contain really exciting results. We hope that these proceedings give a flavour of this excitement.

There are several people who deserve warm thanks for the success of PHYSTAT LHC. First and foremost is my co-Chairperson Albert De Roeck. He organized everything — yes, everything — at CERN. His activities included battling with the CERN administration for some funding for this meeting; being the Workshop's photographer; organizing the Dinner; being responsible for the web site; buying and serving the drinks for the Welcome event, and then clearing up afterwards; etc. And this was all done while looking after CMS issues that arose during the Workshop. A really big thank you, Albert.

Thanks are also due to Dorothee Denise and Kate Ross, who were the Conference secretaries. We are also grateful to Yves Perrin for the design of the poster and logo.

Having statisticians at PHYSTAT LHC added enormously to the usefulness of the meeting. We are very grateful to you all for coming, and giving us the benefit of your expertise in a field in which we are just amateurs. We appreciated your invited talks, the informed comments after other presentations, and the fact that you were around to explain statistical issues to us during the breaks before, between, and after sessions. We hope you continue to attend these meetings, and perhaps we can even induce some of you to join in our analyses in an even more direct manner.

Our invited particle physics speakers clearly went to a lot of trouble to bring into focus the statistical issues involved. It was particularly pleasing to see that ATLAS and CMS could come up with a joint talk. This bodes well for future collaboration between these large experiments, which will make it easier to compare their results and eventually to combine them.

Harrison Prosper deserves special mention for his willingness to take on the very large job of editing these proceedings. He and I want to thank all the contributers who sent us their articles, and the Programme Committee who reviewed them, in some cases with enthusiasm. The session Chairpersons all played an important role in keeping a tight programme running smoothly and on time.

Albert and I also wish to thank all participants who helped to make this a productive meeting. We wish you well with your analyses, and in your searches for exciting new discoveries.

We would also like to acknowledge CERN for providing financial and logistical support. Without that, the workshop would not have been possible.

Louis Lyons

Contents

Upper limits

Analysis methods

Statistical software

Concluding remarks

Appendix

* Invited Talk

Statistical viewpoints

The Last Fifty Years of Statistical Research and their Implications for Particle Physics

D.R. Cox
Nuffield College, Oxford, England

1 Introduction

The title may be interpreted in at least two ways. The last fifty years, although drawing heavily on earlier work, have seen both the development of a large number of particular statistical techniques and also their general availability achieved through relatively painless software packages. These procedures are very widely used in a great range of scientific and technological fields. In so far as these methods are based on probabilistic models of the data, the models tend to be broad descriptions of commonly occurring patterns of haphazard variability; the models relatively rarely contain an important specific subject-matter basis. One possible interpretation of the title is that within this vast mass of material lies the answer to at least some of the demanding statistical issues facing particle physics. Such a source is not the route taken in this paper. While the provision of such an armoury of methods can be claimed to be a massive contribution to science, it seems more fruitful to regard the problems of particle physics as ones to be tackled largely from first principles. See, however, the comments in this volume of Cox and Reid (2007) on some of the so-called statistical wishlist.

The emphasis in this paper is therefore largely on broad principles. I have recently (Cox, 2007) reviewed the nature of statistical considerations in scientific research, strongly emphasizing the need for unity of statistical and subject-matter thinking. In many fields a statistician involved in discussions of a research study will be the most quantitatively-minded member of the group. Clearly this is not the case in discussions with physicists with their very strong and highly successful tradition of independent mathematical thought, so that many of the points in my review hardly apply. The objective of the present paper is therefore partly to outline some general principles and partly, in order to be more specific, to describe in outline issues connected with discovery against a large essentially random background, including an account of false discovery rates.

2 Some methodological themes

A very broad classification of statistical concepts is as follows:

- ideas not specifically based on probability models, for example
  - clustering algorithms
  - visualization of multivariate data
- design of experiments, including simulation studies, and of sampling (data capture) procedures
- construction of probability models
  - families of empirical models of haphazard variation
  - substantive stochastic models
- statistical methods for analysis and interpretation of data in the light of a probability model
  - model checking
- formal theory of inference

Reid (2007) below discusses design of experiments, noting that the experiments discussed in the statistical literature are comparative studies, for example of a modified condition with a control, and are

thus unlike experiments in the present context, so that the ideas are of most relevance in particle physics in the context of computer experiments.

The construction of special stochastic models, especially for dynamic problems, is now virtually a special area of its own. In some countries, for example UK but not in US, it used to be firmly part of the statistical field. It is unclear what the role of such work may be in the context of particle physics.

In the next section a few aspects of more formal theory are discussed before proceeding to more specific matters.

3 An aspect of formal theory

The notion of probability used in formulating a statistical model for data is based on long-run frequency under real or hypothetical repetition. It aims to capture the essence of the data-generating procedure. Probability is used also in defining and interpreting the most appropriate method of analysis. There are various ways in which this can be done, for example by significance tests, confidence limits and so on or by posterior probabilities. This last usually involves a different notion of probability, namely as in some sense a degree of belief. These two approaches are broadly called frequentist and Bayesian respectively. Both ideas occur often in statistical discussions in particle physics and the following brief note is intended to clarify the distinctions between them and, very importantly, the distinctions between different interpretations of Bayesian analyses.

Key ideas of the frequentist approach were set out in very major papers by R.A. Fisher (1922, 1925). Neyman and Pearson, in a series of papers, reformulated Fishers ideas aiming, as they said, for greater clarity. Some recent authors consider that while Neyman and Pearson certainly did achieve mathematical clarity and introduced important new ideas, there was some loss of scientific relevance and there has been some move towards Fisher′s original formulations. For many purposes these differences are relatively minor; the key point is that statistical procedures are calibrated by their performance under hypothetical repetition.

By contrast there are at least five quite different interpretations of Bayesian theory which makes the use of the word Bayesian decidedly ambiguous if not actually confusing. The common link is the mathematical one of using the basic laws of probability to pass from the probability of data given explanation to the probability of the explanation given the data, accounting for the older and, in some ways preferable, term inverse probability. This calculation requires the existence of and knowledge of probabilities for the various explanations in the absence of the specific data under analysis, the so-called prior distribution.

The five interpretations are in outline as follows:

- the prior distribution is a frequency distribution known or estimable from appropriate data. The use of inverse probability is then uncontroversial, and all probabilities are frequentist. This may be called empirical Bayes.
- the prior is intended to be neutral in the sense of introducing no or very little information about the issue under study, leaving the data to supply that information. The idea goes back at least to Laplace and has been developed in detail by Jeffreys and later Jaynes. Demortier (2007) has described the latest thinking on this in a particle physics context, the contributions of Bernardo (2005) to the notion of a reference prior being central; another major contributor is Berger as in Berger (2007). Important contributions to a counting problem with noise have shown that somewhat casual choice of a flat prior can have very bad consequences. This approach may be called objective Bayes.
- the prior distribution may be indirectly data-based, or at least evidence-based in some sense, and provides a way of introducing into an analysis important additional information.
- the prior may encapsulate the opinions of a particular individual, usually called You, and as such

be the basis of Your decision making. There is no necessary implication for any other individual. This may be called the personalistic view.

– Bayesian calculations with a suitably standardized prior may be regarded purely as a convenient algorithm for finding procedures with good frequentist properties, no special notion of probability being needed.

It is important that these, while united by the mathematical techniques used, represent very different views. The first is entirely uncontroversial. The second pursues the same objectives as those of frequentist inference. The third offers the important possibility of incorporating additional information; note, however, that if this is directly based on empirical data, techniques for the combination of information may be used and these should include looking at the mutual consistency of the two sets of data. The personalistic approach is strongly focussed on personal decision making and, while it may be helpful to clarify personal thinking on an issue, it is not of clear relevance to the public discussions of scientific research and to the presentation of evidence for public discussion. The position over Bayesian calculations as algorithmic procedures with a frequency justification is not entirely clear. With a single unknown parameter very close matching between Bayesian and frequentist solutions is achieved with the Jeffreys prior. Such close matching is achievable with more than one parameter only in exceptional cases, although with modest numbers of parameters reasonable results may often be achieved.

Major issues arise when the number of parameters fitted is large relative to the amount of information available. Naive use of objective priors leads to very bad answers. Direct use of maximum likelihood may be misleading. Frequentist techniques for overcoming such difficulties stem from Bartlett (1937) and have been the subject of much recent work; for an account with many examples, see Brazzale, Davison and Reid (2007). Issues associated with models with many parameters are a challenge for all approaches to statistical inference.

4 Many hypotheses

4.1 General formulation

The remainder of the paper is concerned with much more specific issues. Suppose that data are available to test a large number n of null hypotheses, each hypothesis may or may not be true. If only the smallest p-value is reported we are likely to be misled if in fact all null hypotheses are true, for example if the data are totally noise. This is a well-understood selection effect.

There are two distinct problems:

– some small but almost certainly nonzero number of the hypotheses are false and it is required to assess which those are

– it is quite possible that all the null hypotheses are true: how strong is the evidence against this on the basis of $m = \min(p_j)$

These two questions require quite different answers.

4.2 Selection of real effects

There are two broad approaches to this issue, one involving notional error rates and the other, probably the preferable one, empirical Bayesian in formulation. The former approach was first studied systematically by Schweder and Spjotvoll (1982). Suppose that R of the null hypotheses are rejected of which F are in fact rejected in error. Then the false discovery rate is defined as

$$E(F/R \mid R > 0).$$

Procedures are required to ensure that the false discovery rate does not exceed some specified limit (Benjamini and Hochberg, 1995; Storey, 2002). A very minor modification is to define the false rejection

rate F/R to be zero if $R = 0$. In a more elaborate version, F/R is regarded as a random variable which is required to be less than some specified limit with suitably high probability (Genovese and Wasserman, 2006).

For the second approach in its simplest formulation suppose that the n test statistics $T_1, \ldots, T_n$ have under the respective null hypotheses the densities $f_0(t)$, whereas under the alternative hypothesis in each case the density is $f_1(t)$. Suppose further that a proportion θ of null hypotheses are false.

In an important special case the two distributions are Gaussian distributions with unit variance and means zero and μ_1 for the null and for the alternative hypotheses respectively. The unit variance under the null hypothesis is achieved virtually without loss of generality by definition of the test statistic. The unit variance under the alternative is a simplifying assumption which could be tested given sufficient data.

Then for any given t the posterior odds that the value comes from the alternative rather than from the null distribution are

$$\log \frac{P(f_1 \mid t)}{P(f_0 \mid t)} = \log \frac{\theta}{1-\theta} + \mu_1(t - \mu_1/2).$$

Thus, provided θ and μ_1 can be reasonably estimated, the posterior odds corresponding to any given t can be found. Numerical work suggests that at least for a preliminary analysis estimation of the two parameters from the first two moments of the test statistics gives good results (Cox and Wong, 2004). A threshold in t could be set to achieve a preassigned false recovery rate; a main advantage of the method, however, is that it attaches a measure of uncertainty to each value of t rather than merely giving a dichotomy.

A more elaborate approach (Efron et al, 2001) assumes that both densities $f_0(t)$ and $f_1(t)$ are unknown and need to be estimated nonparametrically. The emphasis of the discussion is rather different depending on whether the immediate interpretation of the analysis is important or whether a multi-step selection procedure is involved. In the latter those hypotheses chosen in step one are then tested more searchingly in a second or further stages.

4.3 Global null hypothesis

Suppose now that we consider $m = \min(p_j)$ as the test statistic for the global null hypothesis that all individual null hypotheses are simultaneously satisfied. If the individual tests are independent the p-value allowing for selection is

$$1 - (1-m)^n$$

and without the independence assumption

$$mn$$

is an upper bound, often sharp.

If n is large, for example of the order of 10^3, achievement of an interesting level of significance requires m to be extremely small. In particular this involves sensitivity to the assumptions involved in calculating the individual p_j to a degree that will often be quite unreasonable.

A possible solution, which has been used independently by Professor David Clayton in a genetical context, is as follows.

The first step is to produce a graphical summary of the $\{p_j\}$ that will emphasize the p_j of most interest, namely the small values. For this write $y_j = -\log p_j$ so that under the null hypothesis these have an exponential distribution with unit mean. Equivalently the $2y_j$ have a chi-squared distribution with two degrees of freedom.

Order the values in the form

$$y_{(1)} \geq y_{(2)} \geq \cdots y_{(n)}.$$

Under the global null hypothesis these have expected values

$$1 + 1/2 + \ldots + 1/n, 1/2 + \ldots + 1/n, \ldots, 1/n.$$

These are close to the quantiles of the unit exponential distribution.

Under the global null hypothesis and, assuming that the formal distribution theory is totally appropriate, a plot of the ordered y against their expected values should show a straight line of unit slope. Failure of one or a small number of the null hypotheses is shown by the final points being well above the line. On the other hand, provided that it is unlikely that many of the null hypotheses are false, failure in the tails of the underlying distribution theory is shown by the plot producing a smooth curve; outlying points can then be assessed as departures from this smooth curve.

Detailed discussion of the assessment of significance requires some further work. Under simple assumptions the ordered $y_{(j)}$ considered as functions of j for $j = n, n-1, \ldots, 1$ form a simple Markov process, in fact a nonstationary random walk.

References

[1] Bartlett, M.S. (1937). *Proc. Roy. Soc.* A, **160**, 268-282.

[2] Benjamini, Y. and Hochberg, Y. (1995). *J.R. Statist. Soc.* B **57**, 289-300.

[3] Berger, J. (2007). This volume.

[4] Bernardo, J.M. (2005). Reference analysis. In *Handbook of statistics,* vol. **35**. Amsterdam: Elsevier.

[5] Brazzale, A.R., Davison, A.C. and Reid, N. (2007). *Applied asymptotics.* Cambridge University Press.

[6] Cox, D.R. (2007). *Ann. Appl. Statist.* **1**, 1-18.

[7] Cox, D.R. and Reid, N. (2007). This volume

[8] Cox, D.R. and Wong, M.Y. (2004). *J. R. Statist. Soc.* B **66**, 395-400.

[9] Demortier, L. (2007). This volume.

[10] Efron, B., Tibshirani, R. and Storey, J. (2001). *J. Amer. Statist. Assoc.* **96**, 1151-1160.

[11] Fisher, R.A. (1922). *Phil. Trans. Roy. Soc.* A, **222**, 309-368.

[12] Fisher, R.A. (1925). *Proc. Camb. Phil. Soc.* **22**, 700-725.

[13] Genovese, C.R. and Wasserman, L. (2006). *J. Amer. Statist. Assoc.* **101**, 1408-1417.

[14] Reid, N. (2007). This volume.

[15] Schweder, T. and Spjotvoll, E.J. (1982). *Biometrika* **69**, 493-502.

[16] Storey, J.D, (2002). *J. R. Statist. Soc.* B **64**, 479-498.

A Comparison of Testing Methodologies

James Berger
Duke University, Durham NC, USA

Abstract
This is a mostly philosophical discussion of approaches to statistical hypothesis testing, including p-values, classical frequentist testing, Bayesian testing, and conditional frequentist testing. We also briefly discuss the issue of multiplicity, an issue of increasing concern in discovery. The article concludes with some musings concerning what it means to be a frequentist.

1 Introduction

Because of the tradition in high-energy physics of requiring overwhelming evidence before stating a discovery, there has been limited attention paid to formal statistical testing. With the increasing cost of data, and issues involving simultaneous performance of a multitude of tests, there is likely to be an increasing interest in more formal testing. The main purpose of this article is to review the major approaches to testing, utilizing the basic high-energy physics problem as the vehicle for the discussion.

The following are some of the conclusions that will be argued:

- Tests are very different creatures than confidence intervals or confidence bounds, and it is often not correct to conclude an hypothesis is wrong because it lies outside a confidence interval.
- p-values are typically much smaller than actual error probabilities.
- Objective Bayesian and (good) frequentist error probabilities can agree, providing simultaneous frequentist performance with conditional Bayesian guarantees.

There will also be a brief discussion of multiplicity in testing in Section 3, highlighting the Bayesian approach to dealing with the problem. Section 4 contains some musings about the meaning of frequentism, motivated by presentations and discussions at the Phystat 07 conference.

2 Hypothesis testing

We review, and critically examine, p-values, classical frequentist testing, Bayesian testing and conditional frequentist testing. An ongoing example used in the discussion is a high-energy physics example described in the next section. For pedagogical reasons, a very stylized version of the problem will be considered here, ignoring most of the real physics.

2.1 The pedagogical testing problem and statistical model

Suppose the data, X, is the number of events observed in time T that are characteristic of Higgs boson production in an LHC particle collision experiment. The probabilistic model for the data is that X has density

$$\text{Poisson}(x \mid \theta + b) = \frac{(\theta + b)^x e^{-(\theta+b)}}{x!},$$

where θ is the mean rate of production of Higgs events in time T in the experiment and b is the (assumed known) mean rate of production of events from background sources in time T. Two specific values of X and b that we will follow through various analyses are

$$\textit{Case 1: } x = 7 \text{ and } b = 1.2; \qquad \textit{Case 2: } x = 6 \text{ and } b = 2.2.$$

The main purpose of the experiment is supposedly to determine whether or not the Higgs boson exists which, in terms of the probability model for the data, is typically phrased as testing $H_0 : \theta = 0$

versus $H_1 : \theta > 0$. Thus H_0 corresponds to 'no Higgs.' (Later we will discuss the issue of whether this statistical test is the correct representation of the desired scientific test.) There are many secondary issues that are of interest, such as "What is a lower confidence bound for the mass of the Higgs?" We will not discuss this issue in depth (noting that it has been the focus of many of the Phystat conferences), but will contrast the statistical analysis of the issue with the basic existence issue answered by the test.

2.2 Classical statistical analysis

There are two types of classical analysis: use of p-values, as recommended by Fisher [1], and use of fixed error probability tests, as recommended by Neyman [2].

2.2.1 p-values

The p-**value** in this example, corresponding to observed data x, is

$$p = P(X \geq x \mid b, \theta = 0) = \sum_{m=x}^{\infty} \text{Poisson}(m \mid 0 + b)\,.$$

This is the probability, under the null hypothesis, of observing data as or more extreme than the actual experimental data, and the tradition is to reject the null hypothesis if p is small enough. The part of the definition that may seem odd is the inclusion of *more extreme* data in the probability computation. Indeed, the oddity of doing so led to Jeffreys's [3] famous criticism of p-values "... a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred." (It is worth spending the time to understand that sentence.) For the two cases,

Case 1: $p = 0.00025$ if $x = 7$ and $b = 1.2$; *Case 2:* $p = 0.025$ if $x = 6$ and $b = 2.2$.

There is general agreement that a small p-value indicates that something unusual has happened, but that the p-value does not have a direct quantitative interpretation as evidence against the null hypothesis. Thus Luc Demortier observed in his talk at the Phystat 07 conference:

> In any search for new physics, a small p-value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery.

2.2.2 Fixed α-level testing

Under this approach, one pre-specifies the set of data for which one would reject the hypothesis – the *rejection region* – selecting the set so that the probability of rejection under the null hypothesis is the desired error probability α. Often, as in our example, one can formally state the rejection region in terms of the p-value, namely "reject if $p \leq \alpha$." Because X has a discrete distribution in our example, α should be limited to the possible values allowed by this discreteness; otherwise, one would have to artificially introduce some randomization which is unappealing. (That this rejection region indeed has probability α at the allowed values, follows from an easy computation.)

There are two major concerns with using fixed error probability testing. The first is that it does not properly seem to reflect the evidence in the data. For instance, suppose one pre-selected $\alpha = 0.001$. This then is the error one must report whether $p = 0.001$ or $p = 0.000001$, in spite of the fact that the latter would seem to provide much stronger evidence against the null hypothesis.

The second concern, as it applies to typical high-energy physics experiments, is more subtle: data naturally arrives, and is analyzed, sequentially and typical frequentist computations of fixed error probabilities must take this into account. For instance, suppose the experimental plan is to review the accumulated data at the end of each month, with there being a possibility of claiming a discovery at each

review. The rejection region is then a complicated set involving possible rejection at each of the time points (together with a lack of previous rejection); the frequentist error probability is the probability of this complicated rejection region and is typically much larger than the probability of the rejection region at a particular time. To achieve an error probability of $\alpha = 0.001$ for instance, the rejection region might have to be something such as "reject at each review if $p \leq 0.0001$", so that the frequent looks at the data require a higher standard of evidence to achieve the desired error probability. Note that p-values are affected by this same issue and in roughly the same way: much smaller p-values are needed in a sequential experiment to convey the same evidence as in a fixed sample size experiment.

Louis Lyons raised the interesting point that, with the LHC, declaration of a discovery would not stop the data gathering process, as is common in sequential experimentation in, say, clinical trials. (In clinical trials, claim of a discovery would ethically necessitate stopping the trial, in an attempt to save lives while, as Louis points out, no one we know of really cares if a few more particles are smashed.) So, in principle, a mistake made by this 'sequential look-elsewhere effect' could be corrected with later data.

In practice, however, declaration of a discovery often does have other effects – e.g., people stop research along lines that are incompatible with the discovery – so there is a serious cost to erroneous claims of discovery (in addition to having to return the Nobel prizes), even if there is a possibility of later correction. Also, we shall see that there are readily available reports (both Bayesian and frequentist) that can be made on an interim basis and which do not have difficulty with this sequential look-elsewhere effect, so the entire philosophical conundrum can be avoided.

2.3 Bayesian testing

2.3.1 Bayes factor

The **Bayes factor** of H_0 to H_1 in our ongoing example is given by

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0 + b)}{\int_0^\infty \text{Poisson}(x \mid \theta + b)\pi(\theta)\, d\theta} = \frac{b^x\, e^{-b}}{\int_0^\infty (\theta + b)^x\, e^{-(\theta+b)}\pi(\theta)\, d\theta}\,;$$

in the *subjective Bayesian approach*, the prior density, $\pi(\theta)$, is chosen to reflect the beliefs of the investigators (e.g., it could reflect the standard model predictions pertaining to the Higgs) while, in the *objective Bayesian approach*, it is chosen conventionally and nominally reflects a lack of knowledge concerning θ.

A reasonable objective prior here (to be justified later, but note that it is a proper prior) is $\pi^I(\theta) = b(\theta + b)^{-2}$. For this prior, the Bayes factor is given by

$$B_{01} = \frac{b^x\, e^{-b}}{\int_0^\infty (\theta + b)^x\, e^{-(\theta+b)} b(\theta + b)^{-2}\, d\theta} = \frac{b^{(x-1)}\, e^{-b}}{\Gamma(x - 1, b)}\,,$$

where Γ is the incomplete gamma function. The result for the two cases is

Case 1: $B_{01} = 0.0075$ (recall $p = 0.00025$); *Case 2:* $B_{01} = 0.26$ (recall $p = 0.025$)

2.3.2 Objective posterior probabilities of the hypotheses

The objective choice of prior probabilities of the hypotheses is $\Pr(H_0) = \Pr(H_1) = 0.5$, in which case

$$\Pr(H_0 \mid x) = \frac{B_{01}}{1 + B_{01}}\,.$$

For the two cases in the example,

Case 1: $\Pr(H_0 \mid x) = 0.0075$ (recall $p = 0.00025$); *Case 2:* $\Pr(H_0 \mid x) = 0.21$ (recall $p = 0.025$).

Of course, one can specify subjective prior probabilities of each hypothesis and determine the resulting posterior probabilities, but scientific communication is usually done through objective posterior probabilities or Bayes factors, since any individual can take either and easily convert it into the individual's personal subjective answer.

2.3.3 Complete posterior distribution

In addition to the uncertainty in the hypotheses, there is also uncertainty in θ, given that H_1 were true. The complete posterior distribution is thus determined by

- $\Pr(H_0 \mid x)$, the posterior probability of the null hypothesis;
- $\pi(\theta \mid x, H_1)$, the posterior distribution of θ under H_1.

For Case 1 in the example, Figure 1 presents these two parts of the full posterior distribution. One way of thinking of this is that the vertical bar gives the probability that one has just observed noise, while the density part says where θ is likely to be if there is a discovery.

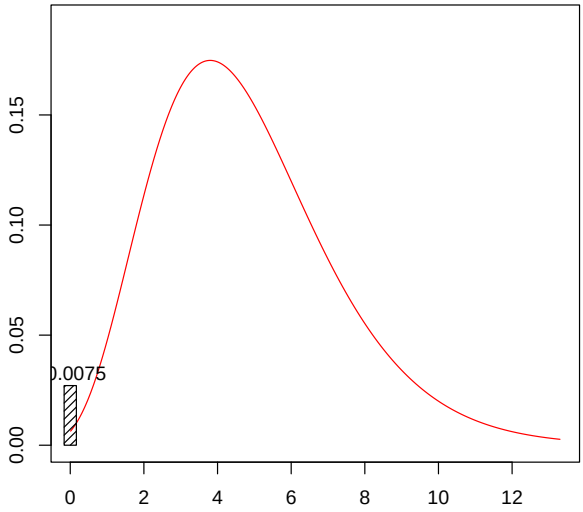

Fig. 1: For Case 1, $\Pr(H_0 \mid x)$ (the vertical bar), and the posterior density for θ given $x = 7$ and H_1.

A useful summary of the complete posterior is $\Pr(H_0 \mid x)$ and C, a (say) 95% posterior confidence interval for θ under H_1. For the two cases, and with C chosen to be an equal-tailed 95% posterior confidence interval (i.e., omitting 2.5% of the posterior mass on the left and the right)
Case 1: $\Pr(H_0 \mid x) = 0.0075$ and $C = (1.0, 10.5)$; *Case 2:* $\Pr(H_0 \mid x) = 0.21$ and $C = (0.2, 8.2)$.
C could, alternatively, be chosen to be a one-sided confidence bound, if desired.

Note that confidence intervals alone are *not* a satisfactory inferential summary. In Case 2, for instance, the 95% confidence interval does not include 0, and so many mistakenly believe that one can accordingly reject $H_0 : \theta = 0$. But, the full posterior distribution also has a probability of 0.21 that $\theta = 0$, which would hardly imply a confident rejection.

A Brief Aside: A precise null hypothesis, such as $H_0 : \theta = 0$, is typically never true *exactly*; rather, it is used as a surrogate for a 'real null' $H_0^\epsilon : \theta < \epsilon, \quad \epsilon$ small. In the Higgs example for instance, while the scientific null is real (i.e., the Higgs might not exist), the statistical null is based on the experimental measurements, and there is undoubtedly some small bias ϵ in the experiment. Berger and Delampady [4] show that, under reasonable conditions, if $\epsilon < \frac{1}{4}\sigma_{\hat{\theta}}$, where $\sigma_{\hat{\theta}}$ is the standard error of the estimate of θ, then $Pr(H_0^\epsilon|\mathbf{x}) \approx Pr(H_0|\mathbf{x})$, so that the point null is then a reasonable approximation to the real null.

2.4 The discrepancy between p-values and posterior probabilities

The Bayesian error probabilities given in the previous section differed from the corresponding p-values by factors of 30 and 10 in the two cases, respectively. What explains this?

It might be tempting to say that there is something wrong with the Bayesian analysis, but even a pure likelihood analysis (favored by many Fisherians) reveals the same effect. In particular (following Edwards, Lindeman and Savage [10]), note that a lower bound on the Bayes factor over all possible priors can be found by choosing $\pi(\theta)$ to be a point mass at $\hat{\theta}$ (the maximum likelihood estimate), yielding

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0+b)}{\int_0^\infty \text{Poisson}(x \mid \theta+b)\pi(\theta)\, d\theta} \geq \frac{\text{Poisson}(x \mid 0+b)}{\text{Poisson}(x \mid \hat{\theta}+b)} = \min\{1, \left(\frac{b}{x}\right)^x e^{x-b}\}\,. \qquad (1)$$

In 'likelihood language,' this says that, for the given data, the likelihood of H_0 relative to the likelihood of H_1 is at least the bound on the right hand side of (1). For the two cases, this bound is

Case 1: $B_{01} \geq 0.0014$ (recall $p = 0.00025$); *Case 2:* $B_{01} \geq 0.11$ (recall $p = 0.025$),

so that a serious discrepancy remains even when the prior is eliminated. This can be traced to the fact that the p-value is based on the probability of the tail area of the distribution, rather than the probability of the actual observed data.

It is well known that Bayesian analysis utilizing suitable proper priors will automatically penalize more complex models (i.e., has an Ockham's razor effect – cf. Jefferys and Berger [5]), and it is useful to separate this effect from that observed above in explaining the difference between p-values and posterior probabilities or Bayes factors. Thus in Case 1, where the p-value ($\approx .00025$) and the objective posterior probability of the null (≈ 0.0075) differ by a factor of 30,

- a factor of $.0014/.00025 \approx 5.6$ is due to the difference between a tail area $\{X : X \geq 7\}$ and the actual observation $X = 7$ (as reflected through the likelihood ratio for the observation);
- the remaining factor of roughly 5.4 in favor of the null results from the Ockham's razor penalty resulting from the conventional proper prior that was used.

An Aside – Robust Bayesian Analysis: Robust Bayesian theory (cf. Berger [6] for references) takes a more sophisticated look at the type of bounding over priors that is done in (1). For instance, it might be deemed scientifically reasonable to restrict attention to priors $\pi(\theta)$ that are nonincreasing, in which case it is easy to see that

$$B_{01}(x) \geq \frac{b^x\, e^{-b}}{\sup_c \int_0^c (\theta+b)^x\, e^{-(\theta+b)} c^{-1}\, d\theta}\,.$$

For the two cases, this bound is

Case 1: $B_{01} \geq 0.0024$ (recall $p = 0.00025$); *Case 2:* $B_{01} \geq 0.15$ (recall $p = 0.025$).

2.5 Conditional frequentist testing

There is a powerful (but, alas, largely overlooked) frequentist school called *conditional frequentist analysis*. This school was formalized by Kiefer [7] and Brown [8], and proceeds as follows:

- find a statistic S that reflects the "strength of evidence" in the data;
- compute the frequentist measure of error conditional on S.

Artificial example (from Berger and Wolpert [9]): Observe X_1 and X_2 where

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2. \end{cases}$$

A classical (unconditional) 75% confidence set (here a point) for the unknown θ is

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2\,; \end{cases}$$

it is easy to compute that $P_\theta(C(X_1, X_2) \text{ contains } \theta) = 0.75$. It is, however, clearly silly to report this; when $X_1 \neq X_2$, it is a certainty that the confidence set equals θ while, if $X_1 = X_2$, it is intuitively 50-50 as to whether the confidence set equals θ. The issue here is typically phrased in statistics as that of desiring good conditional performance (for relevant subsets of the actual data); in the Phystat literature it is more commonly phrased as desiring *Bayesian credibility:* for a reasonable prior, the Bayesian coverage of the confidence set should be reasonable. In this example, for instance, if one uses the objective prior $\pi(\theta) = 1$, then $C(X_1, X_2)$ has Bayesian credibility of 100% if $x_1 \neq x_2$ and 50% if $x_1 = x_2$, so that the report of 75% confidence in all circumstances would be seriously deficient from the viewpoint of Bayesian credibility.

The conditional frequentist approach here would

- measure the strength of evidence in the data by, say, $S = |X_1 - X_2|$ (either 0 or 2)
- compute the conditional coverage

$$P_\theta(C(X_1, X_2) \text{ contains } \theta \mid S) = \begin{cases} 0.5 & \text{if } S = 0 \\ 1.0 & \text{if } S = 2\,, \end{cases}$$

which is clearly the right answer.

Returning to the testing problem, Berger, Brown and Wolpert [11] for continuous data, and Dass [12] for discrete data, proposed the following conditional frequentist testing procedure for testing a simple hypothesis versus a simple alternative:

- Develop S, the measure of strength of evidence in the data, as follows:
  - let $p_i(x)$ be the p-value from testing H_i against the other hypothesis;
  - define $S = \max\{p_0(x), p_1(x)\}$; its use is based on deciding that data (in either the rejection or acceptance regions) with the same p-value has the same 'strength of evidence.'
- Accept H_0 when $p_0 > p_1$, and reject otherwise.
- Compute Type I and Type II conditional error probabilities as

$$\alpha(s) = P_0(\text{rejecting } H_0 \mid S = s) \equiv P_0(p_0 \leq p_1 \mid S(X) = s)$$
$$\beta(s) = P_1(\text{accepting } H_0 \mid S = s) \equiv P_1(p_0 > p_1 \mid S(X) = s),$$

  where P_i refers to probability under H_i.

The surprising feature of this conditional test is stated in the following theorem from those papers.

Theorem 1 *The conditional frequentist error probabilities, $\alpha(s)$ and $\beta(s)$, exactly equal the (objective) posterior probabilities of H_0 and H_1, so conditional frequentists and Bayesians report the same error probabilities.*

In our ongoing example, the conditional Type I error is thus $\alpha(s) = \Pr(H_0 \mid x) = B_{01}/(1 + B_{01})$ (=0.0075 in Case 1; =0.21 in Case 2). Some features of this:

- The conditional test can be viewed as a way to convert p-values into real frequentist error probabilities when there is an alternative hypothesis.
- The conditional error probabilities $\alpha(s)$ and $\beta(s)$ are fully data-dependent (being smaller when p is smaller, in contrast to the fixed α-level tests), yet are fully frequentist.
- The conditional test also applies without any change in sequential settings; since Bayesian error probabilities are known to ignore the stopping rule, so must the conditional frequentist test (Berger, Boukai and Wang [13]).

The conditional frequentist test thus overcomes all of the difficulties with the fixed α-level test that were discussed earlier, and so can be used happily by frequentists. Of course, one need not go through the formal conditional frequentist computation, since the theorem guarantees that the answer which would be obtained is the same as the objective Bayesian answer (which can be obtained much more directly).

There is the caveat that the above discussion was given only for the testing of two simple hypotheses. In our ongoing example, on the other hand, H_1 was a composite hypothesis (involving an unknown θ). The papers mentioned above do cover the extension of the theory to the composite alternative case, with the only modification being that the conditional Type II error that is obtained is a certain average Type II error over θ; the conditional Type I error is unaffected. Extensions to composite null hypotheses are considered in Dass and Berger [14] for composite null hypotheses that have an invariance structure to group operations; this class of composite null hypotheses includes most classical situations of testing. The nice feature of this class of composite null hypotheses is that the conditional Type I error is constant over the null hypothesis, and so no averaging over Type I error needs to be done. (There are other technical caveats to the conditional frequentist testing paradigm that are discussed in the mentioned papers, but they have essentially no practical impact.)

2.6 Implementing Bayesian testing

To implement objective Bayesian estimation (and confidence procedures) there are, in principle, excellent objective priors available, such as *reference priors* (see Bernardo [15] for a review and references). In practice, determination of such objective priors can be challenging but the goal is, at least, clear.

In Bayesian hypothesis testing and model selection, however, determination of suitable prior distributions is considerably more challenging, in part because it is typically the case that improper prior distributions cannot be used (or at least have to be used very carefully). Use of 'vague proper priors' (another staple of many Bayesians in estimation problems) is even worse, and will typically give nonsensical answers in testing and model selection. There has thus been a huge effort in statistics to derive objective (or at least conventional) priors for use in hypothesis testing and model selection. These issues and this literature can be accessed through Berger and Pericchi [16].

For our ongoing example, an appealing methodology for default prior construction is the *intrinsic* or *expected posterior* prior construction. For the situation where the data consists of i.i.d. observations from a density $f(x \mid \theta)$, and for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, the construction is as follows:

- let $\pi^O(\theta)$ be a good estimation objective prior, so that $\pi^O(\theta \mid \boldsymbol{x}) = [\prod_{i=1}^n f(x_i \mid \theta)]\pi^O(\theta)/m^O(\boldsymbol{x})$ is the resulting posterior, where $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $m^O(\boldsymbol{x}) = \int [\prod_{i=1}^n f(x_i \mid \theta)]\pi^O(\theta)\, d\theta$;
- then the intrinsic prior is $\pi^I(\theta) = \int \pi^O(\theta \mid \boldsymbol{x}^*)[\prod_{i=1}^q f(x_i \mid \theta_0)]\, d\boldsymbol{x}^*$, with $\boldsymbol{x}^* = (x_1, \ldots, x_q)$ being (unobserved) data of the minimal sample size q such that $m^O(\boldsymbol{x}^*) < \infty$.

Note that this will be a proper (not vague proper) prior.

The idea behind this prior is that, if one were handed the data $\boldsymbol{x}^*$ but allowed to use it only for prior construction, one would happily compute $\pi^O(\theta \mid \boldsymbol{x}^*)$ and use this proper prior to conduct the test. We don't have $\boldsymbol{x}^*$ available, but we could simulate $\boldsymbol{x}^*$ from the null model, and compute the resulting 'average' prior. There are many other justifications of this prior; see Pérez and Berger [17] for discussion and references. Note, however, that use of such conventional proper priors is inherently more contentious than use of objective priors for estimation problems. Indeed, it would be better to determine $\pi(\theta)$ from consensus scientific knowledge, providing the knowledge is relatively precise and quantifiable.

For our ongoing example, suppose we choose $\pi^O(\theta) = 1/(\theta + b)$. (Jeffreys prior, the square root of π^O, would probably be better, but leads to a much more difficult computation.) Following the ideas in Berger and Pericchi [18], we represent the Poisson observation, X, over the time period T from the distribution in the example as a sum of i.i.d observations from an exponential inter-arrival time process.

Indeed, for $i = 1, \ldots$, consider $Y_i \sim f(y_i \mid (\theta + b)/T) = (\theta + b)T^{-1} \exp\{-(\theta + b)y_i/T\}$; then $X \equiv \{\text{first } j \text{ such that } S_j = \sum_{i=1}^{j} Y_i > T\} - 1$. A minimal sample size for this exponential distribution can easily be seen to be $q = 1$. Computation then yields $\pi^I(\theta) = \int \pi^O(\theta \mid y_1) f(y_1 \mid 0) dy_1 = b/(\theta+b)^2$, which was the conventional proper prior used for Bayesian testing in the example.

3 Multiplicities

The issue of dealing with multiplicities in discovery is increasingly being recognized to be important. One type of multiple testing has already been discussed, namely sequential experimentation in which one periodically evaluates the incoming data to see if a discovery can be claimed. It is interesting that frequentist analyses often need to be adjusted to account for these 'looks at the data,' while Bayesian analyses (and optimal conditional frequentist analyses) do not. That Bayesian analysis claims no need to adjust for this 'look elsewhere' effect – called the *stopping rule principle* – has long been a controversial and difficult issue in statistics, as admirably expressed by Savage [19]: "I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that people resist an idea so patently right." See Berger and Berry [20] for discussion of this controversy, and note that the controversy is no longer a frequentist versus Bayesian issue, because of the fact that optimal conditional frequentist tests also obey the stopping rule principle.

Another common situation of multiple testing is when one is scanning many possible data sets for a discovery. For instance, suppose 1000 energy channels are searched for a signal expected from a non-standard theory. It is well known that one cannot proceed with separate testing of each data set, but the classical solution – the Bonferonni adjustment – is often viewed as being too harsh. The Bonferonni adjustment assumes each test is independent, in which case one divides the desired error probability α by the number of tests to determine the significance level that an individual test must achieve to declare a discovery. Thus if $\alpha = 0.001$ is desired for 1000 independent tests, the per-test significance level should be set at 0.000001 for declaring a discovery.

I have been told that the assumption of (at least approximate) independence of test statistics does hold for many high-energy physics experiments, in which case use of the Bonferonni correction is fine. When the various test statistics are dependent, however (as happens in most non-physics examples I know of), the Bonferroni correction can be much too conservative, so it's use would incur a dramatic loss of power for discovery. Finding appropriate correction for multiple testing under dependence is, unfortunately, quite difficult from the frequentist viewpoint. Note, also, that there are no shortcuts here; simple alternative methods such as the 'false discovery rate' are fine for screening purposes, but are not useful for claiming a discovery.

One of the highly attractive features of the Bayesian approach to multiple testing or model selection is that (if done properly) it will automatically adjust for multiplicities, and do so in a way that preserves as much discriminatory power as possible. The Bayesian adjustment for multiplicity occurs, somewhat curiously, directly through the prior probabilities assigned to the tests or models. Consider two illustrative cases:

Mutually exclusive hypotheses: Suppose one is testing mutually exclusive hypotheses $H_i,\ i = 1, \ldots, m$, where it is known that one is true. An objective Bayesian would choose $p_i = \Pr(H_i) = 1/m$. Suppose, for instance, that a signal is known to exist, but it is not known in which of 1000 energy channels it will manifest. Then each channel would be assigned prior probability 0.001 of containing the signal, an automatic penalization of each hypothesis.

Suppose instead that 1000 channels are searched for a signal expected from a non-standard theory that could manifest in only one channel. Then one should assign some prior mass – e.g. 1/2 – to 'no signal,' giving prior probability of 0.0005 to each channel. Note that these simple adjustments apply no matter what the dependence is between the test statistics, indicating why it is much easier to approach

multiplicity adjustment from the Bayesian perspective.

Independently occurring hypotheses: Consider, instead, the situation in which there are multiple possible discoveries, and that the signal from each would appear in a separate channel. If we knew nothing about these possible signals, we might choose to assign prior probabilities by first defining p as the probability that any given channel will manifest a signal. This would typically be unknown, and hence would need to be assigned a prior distribution $\pi(p)$. This could be chosen according to scientific knowledge, or set equal to a default prior such as the uniform distribution. That an assignment of prior probabilities such as this automatically deals with multiplicity is demonstrated in Scott and Berger [21].

There is a large and increasing literature on discovery techniques in the face of multiplicity. Two recent references are Storey, Dai and Leek [22] and Guindani, Zhang and Mueller [23].

4 Musings on the meaning of frequentism

4.1 Introduction and example

During the Phystat meeting, there were a number of interesting problems discussed that caused me to reflect on the meaning of frequentism. To facilitate the discussion here, consider the following version of the basic HEP problem, but now focusing on confidence bounds (see, e.g., Heinrich [24] for background).

Suppose $X_{s+b} \sim \text{Poisson}(X_{s+b} \mid s+b)$, where s is the unknown signal mean and b now an unknown background mean. The goal is to find an upper confidence limit for s. There is also information available about the nuisance parameter b, arising from either

- *Case 1:* independent sideband data $X_b \sim \text{Poisson}(X_b \mid b)$,
- *Case 2:* randomness in b from experiment to experiment arising from a known random mechanism,
- *Case 3:* agreed scientific beliefs.

4.2 Bayesian analysis

Suppose we have an agreed upon objective prior density $\pi^O(s \mid b)$ for s given b (the best objective priors will typically depend on nuisance parameters such as b here). The information about b would be encoded in a prior density $\pi(b)$. This density would be derived differently in each case:

- *Case 1:* With the sideband data X_b, a standard approach would be to chose an initial objective prior $\pi^O(b)$, and then choose the final $\pi(b)$ to be the posterior $\pi^O(b \mid X_b) \propto \text{Poisson}(X_b \mid b)\pi^O(b)$.
- *Case 2:* $\pi(b)$ describes the physical randomness of the (otherwise unmeasured) background from experiment to experiment.
- *Case 3:* $\pi(b)$ is chosen to encode accepted scientific beliefs.

In all three cases, Bayesian analysis would proceed in the same way, constructing a $100(1-\alpha)\%$ upper confidence limit U for s as the solution to

$$1-\alpha = \int_0^U \pi(s \mid X_{s+b})\, ds\,,$$

where $\pi(s \mid X_{s+b})$ is the posterior distribution

$$\pi(s \mid X_{s+b}) = \frac{\int \text{Poisson}(X_{s+b} \mid s+b)\pi^O(s \mid b)\pi(b)\, db}{\int\int \text{Poisson}(X_{s+b} \mid s+b)\pi^O(s \mid b)\pi(b)\, db\, ds}\,.$$

The point is that Bayesian analysis does not care about the nature of the randomness in the modeling of the information about b.

4.3 Frequentist analysis

Frequentist analysis can be quite different in the three cases.

4.3.1 Frequentist analysis in Case 1.

The natural frequentist goal: Frequentist coverage with respect to the joint distribution of X_{s+b} and X_b, i.e. control of

$$P(s \leq U(X_{s+b}, X_b) \mid s, b) = \sum_{X_{s+b}=0}^{\infty} \sum_{X_b=0}^{\infty} 1_{\{s \leq U(X_{s+b}, X_b)\}} \text{Poisson}(X_{s+b} \mid s+b) \text{Poisson}(X_b \mid b)\,,$$

where $1_{\{s \leq U(X_{s+b}, X_b)\}}$ is 1 if $s \leq U(X_{s+b}, X_b)$ and 0 otherwise.

This problem has been extensively studied in the Phystat literature. It is interesting that there is, as of yet, no solution which is agreed by all to be adequate in terms of both frequentist coverage and Bayesian credibility (conditional performance). The objective Bayesian holy grail in this problem would be to find the reference prior for (s, b), with s being the parameter of interest; the hope is that the upper confidence bound arising from such a prior would do an excellent job of balancing frequentist coverage and Bayesian credibility. Finding the reference prior is very challenging, however, as was discussed in the Phystat talk of Luc Demortier (and see Demortier [25]).

4.3.2 Frequentist analysis in Case 2.

The natural frequentist goal: Frequentist coverage with respect to the marginal density of X_{s+b}, given by $f(X_{s+b} \mid s) = \int \text{Poisson}(X_{s+b} \mid s+b)\pi(b)db$. The coverage target is then

$$P(s \leq U(X_{s+b}) \mid s) = \sum_{X_{s+b}=0}^{\infty} 1_{\{s \leq U(X_{s+b})\}} f(X_{s+b} \mid s)\,.$$

The reason this is the natural frequentist goal is because b changes from experiment to experiment according to $\pi(b)$, and *real* frequentism is about performance of a statistical procedure in actual repeated use of the procedure in differing experiments, as discussed in Neyman [2]. (The textbook definition of frequentism – in which one considers *imaginary* repetition of the same experiment – makes no sense in terms of reality; the standard definition has mathematical relevance, but the philosophical appeal of frequentism to scientists is presumably its relevance to real experimentation over time.)

Attaining this frequentist goal while achieving good Bayesian credibility is potentially rather straightforward, since the problem has been reduced to a one-parameter problem. Indeed, one simply computes the reference (Jeffreys) prior corresponding to $f(X_{s+b} \mid s)$, namely

$$\pi^J(s) = \sqrt{I(s)}\,, \qquad I(s) = -\sum_{X_{s+b}=0}^{\infty} f(X_{s+b} \mid s) \frac{d^2}{ds^2} \log f(X_{s+b} \mid s)\,.$$

The resulting Bayesian confidence bound will automatically have good Bayesian credibility (conditional performance), and the Jeffreys prior for one-parameter problems typically results in Bayes procedures with excellent frequentist coverage properties (except possibly at the boundary $s = 0$; see Bayarri and Berger [26] for discussion).

4.3.3 Frequentist analysis in Case 3.

The natural frequentist goal: Here the situation is quite murky. Since $\pi(b)$ is not physical randomness, but simply scientific opinion, a classical frequentist could insist that, for every given s and b, we control

$$P(s \leq U(X_{s+b}) \mid s, b) = \sum_{X_{s+b}=0}^{\infty} 1_{\{s \leq U(X_{s+b})\}} \text{Poisson}(X_{s+b} \mid s+b)\,.$$

This is actually not possible to control unless there is a known bound on b, but a classical frequentist would philosophically wish to control this coverage.

Alternatively, one could argue that, since $\pi(b)$ arises from consensus scientific opinion, it should be treated the same as when it arises from physical randomness, and so one should seek to control coverage as in Case 2, i.e.

$$\begin{aligned} P(s \leq U(X_{s+b}) \mid s) &= \sum_{X_{s+b}=0}^{\infty} 1_{\{s \leq U(X_{s+b})\}} f(X_{s+b} \mid s) \\ &= \int P(s \leq U(X_{s+b}) \mid s, b)\pi(b)db\,. \end{aligned}$$

The second expression for this coverage shows that the criterion can be interpreted as an average of the coverage for given s and b, averaged over the consensus prior distribution for b.

There are many situations in which it has been argued that a frequentist should use an average coverage criterion; see Bayarri and Berger [26] for examples and references. Here it seems clearly right because of necessity; what else can be done given the available information? The point worth pondering is – if average coverage is fine here, why should it be philosophically problematical in other cases?

Acknowledgements

This work was supported by NSF Grants AST-0507481 and DMS-0103265. Also very helpful was the extensive discussion by participants in the working group on Particle Physics (lead by Louis Lyons) during the program on Astrostatistics at the Statistical and Applied Mathematical Sciences Institute in March of 2006. Finally, thanks to Luc Demortier and Louis Lyons for their very helpful comments and discussions concerning this paper.

References

[1] Fisher, R.A. (1973). *Statistical Methods and Scientific Inference (3rd ed.)*. Macmillan, London.

[2] Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthèse*, **36**, 97–131.

[3] Jeffreys, H. (1961). *Theory of Probability*, London: Oxford University Press.

[4] Berger, J. and Delampady, M. (1987). Testing precise hypotheses (with Discussion). *Statist. Science* **2**, 317–352.

[5] Jefferys, W. and Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64–72.

[6] Berger, J. (1994). An overview of robust Bayesian analysis. *Test*, **3**, 5–124.

[7] Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with Discussion). *J. Amer. Statist. Assoc.* **72**, 789–827.

[8] Brown, L. D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist.* **6**, 59-71.

[9] Berger, J., and Wolpert, R. L. (1988). *The Likelihood Principle: A Review, Generalizations, and Statistical Implications* (second edition, with Discussion). Hayward, CA: Institute of Mathematical Statistics.

[10] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

[11] Berger, J., Brown, L.D. and Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics*, **22**, 1787–1807.

[12] Dass, S. C. (2001). Unified Bayesian and conditional frequentist testing procedures for discrete distributions. *Sankhya Ser. B*, **63**, 251–269.

[13] Berger, J., Boukai, B. and Wang, Y. (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika*, **86**, 79–92.

[14] Dass, S. and Berger, J. (2003). Unified Bayesian and conditional frequentist testing of composite hypotheses. *Scandinavian Journal of Statistics*, **30**, 193–210.

[15] Bernardo, J. M. (2005). Reference analysis. Handbook of Statistics 25 (D. K. Dey and C. R. Rao eds.). Amsterdam: Elsevier, 17–90.

[16] Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison (with Discussion). In *Model Selection*, P. Lahiri, ed., Institute of Mathematical Statistics Lecture Notes – Monograph Series, volume 38, Beachwood Ohio, 135–207.

[17] Pérez, J.M. and Berger, J. (2002). Expected posterior prior distributions for model selection. *Biometrika*, **89**, 491–512.

[18] Berger, J. and Pericchi, L. (2004). Training samples in objective Bayesian model selection. *Ann. Statist.*, **32**, 841–869.

[19] Savage, L.J. (1962). *The Foundations of Statistical Inference.* London: Methuen.

[20] Berger, J. and Berry, D. (1988). The relevance of stopping rules in statistical inference (with Discussion). In *Statistical Decision Theory and Related Topics IV.* Springer–Verlag, New York.

[21] Scott, J. and Berger, J. (2006) An exploration of aspects of Bayesian multiple testing, Journal of Statistical Planning and Inference, Vol. 136, No. 7. (1 July 2006), pp. 2144–2162.

[22] Storey, J.D., Dai, J.Y and Leek, J.T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, **8***(2)*, 414–432.

[23] Guindani, M., Zhang, S. and Mueller, P.M. (2007). A Bayesian discovery procedure. Technical Report, MD Anderson Medical Center.

[24] Heinrich, J. (2008). Review of the Banff challenge on upper limits. These Proceedings.

[25] Demortier, L. (2005). Bayesian reference analysis for particle physics. *PHYSTAT05 Proceedings on "Statistical Problems in Particle Physics, Astrophysics and Cosmolgy"*. Oxford University Press.

[26] Bayarri, M.J. and Berger, J. (2004). The interplay between Bayesian and frequentist analysis. *Statistical Science*, **19**, 58–80.

Discovery

P Values and Nuisance Parameters

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

Abstract
We review the definition and interpretation of p values, describe methods to incorporate systematic uncertainties in their calculation, and briefly discuss a non-regular but common problem caused by nuisance parameters that are unidentified under the null hypothesis.

1 Introduction

Statistical theory offers three main paradigms for testing hypotheses: the Bayesian construction of hypothesis probabilities, Neyman-Pearson procedures for controlling frequentist error rates, and Fisher's evidential interpretation of tail probabilities. Since practitioners often attempt to combine elements from different approaches, especially the last two, it is useful to illustrate with a couple of examples how they actually address different questions [1].

The first example concerns the selection of a sample of $p\bar{p}$ collision events for measuring the mass of the top quark. For each event one must decide between two hypotheses, H_0 : *The event is background*, versus H_1 : *The event contains a top quark*. Since the same testing procedure is sequentially repeated on a large number of events, the decision must be made in such a way that the rate of wrong decisions is fully controlled in the long run. Traditionally, this problem is solved with the help of Neyman-Pearson theory, with a terminology adapted to physics goals: the Type-I error rate α translates to background contamination, and the power of the test to selection efficiency. In principle either hypothesis can be assigned the role of H_0 in this procedure. Since only the Type-I error rate is directly adjustable by the investigator (via the size of the critical region), a potentially useful criterion is to define as H_0 the hypothesis for which it is deemed more important to control the incorrect rejection probability than the incorrect acceptance probability.

The second example is encountered in searches for new phenomena, when one has observed an enhancement in a background spectrum and one wishes to characterize and quantify the evidence this provides against the background-only null hypothesis. When a well-defined alternative hypothesis can be formulated, a coherent characterization is best done with the help of likelihood ratios [2] or Bayes factors [3]. Often however, the alternative is not unique, and there is a desire to quantify the evidence against the null hypothesis in a way that does not depend on the alternative. Although the idea that this can be done meaningfully is rejected by some statisticians, it has a long history in the scientific and statistics literature. The method of solution is based on p values, the focus of this contribution.

Section 2 reviews the definition and interpretation of p values. A major obstacle to their calculation is assessing the effect of systematic uncertainties. As is standard in high energy physics, we assume that the latter can be modelled by so-called nuisance parameters [4], so that the task of incorporating a systematic uncertainty (physics terminology) reduces to that of eliminating the corresponding nuisance parameter (statistics terminology). Methods for solving this problem are described in section 3. A non-regular form of this problem, known among physicists as the "look-elsewhere" effect, is briefly discussed in section 4. Our conclusions are contained in section 5.

2 Definition and Interpretation of p Values

Suppose we collect some data X and wish to test a hypothesis H_0 about the distribution $f(x\,|\,\theta)$ of the underlying population. The first step is to find a test statistic $T(X)$ such that large realizations of its observed value, $t_0 \equiv T(x_0)$, are evidence against H_0. One way to calibrate this evidence is to compute

the probability for observing $T = t_0$ or a larger value under H_0; this tail probability is known as the p value of the test:

$$p \;=\; \mathbb{P}\mathrm{r}(T \geq t_0 \,|\, H_0). \tag{1}$$

Hence, small p values are evidence against H_0. The usefulness of this calibration is that the distribution of p under H_0 is in principle uniform, and therefore known to the experimenter and the same in all testing problems to which the procedure is applied. Unfortunately, in practice it is often difficult to obtain uniform p values, either because the test statistic is discrete or because of the presence of nuisance parameters. The following terminology characterizes the null distribution of p values:

$$\begin{array}{lcl} p \text{ exact or uniform} & \Leftrightarrow & \mathbb{P}\mathrm{r}(p \leq \alpha \,|\, H_0) = \alpha, \\ p \text{ conservative or overcovering} & \Leftrightarrow & \mathbb{P}\mathrm{r}(p \leq \alpha \,|\, H_0) < \alpha, \\ p \text{ liberal or undercovering} & \Leftrightarrow & \mathbb{P}\mathrm{r}(p \leq \alpha \,|\, H_0) > \alpha, \end{array}$$

where α is a number between 0 and 1. Compared to an exact p value, a conservative one tends to understate the evidence against H_0, whereas a liberal one tends to overstate it. It is of course possible for a p value to be conservative for some values of α and liberal for others.

Even though the definition of p values is straightforward, their interpretation is notoriously subtle and has been the subject of numerous papers in the statistics literature. Here we limit ourselves to a few important caveats. The first one is that p values are not frequentist error rates or confidence levels. Indeed, the latter are performance criteria that must be chosen *before* the experiment is done, whereas p values are post-data measures of evidence. Secondly, p values should not be confused with posterior hypothesis probabilities. Compared to the latter, p values often tend to exaggerate the evidence against the null hypothesis. Finally, the notion that equal p values represent equal amounts of evidence should be regarded with a healthy dose of scepticism. Arguments can be formulated to show that the evidence provided by a p value depends on sample size as well as on the type of hypothesis being tested.

Because of these and other caveats, it is better to treat p values as nothing more than useful exploratory tools or measures of surprise. In any search for new physics, a small p value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery. A detailed discussion of the role of p value tests can be found in Refs.[5, 6].

3 Incorporating Systematic Uncertainties

In order to evaluate the various methods that are available to incorporate systematic uncertainties in p value calculations, it is useful to discuss some properties one would like these methods to enjoy:

1. Uniformity: An important aspect of p values is their uniformity under H_0, since this is how the evidence provided by a test statistic is calibrated. If exact uniformity is not achievable in finite samples, then asymptotic uniformity may still provide a useful criterion.
2. Monotonicity: For a fixed data sample, increases in systematic uncertainty should devalue the evidence against H_0, i.e. increase the p value.
3. Generality: The method should not depend on the testing problem having a special structure, but should be applicable to as wide a range of situations as possible.
4. Power: Although p values are generally not constructed with a specific alternative in mind, it may sometimes be useful to compare their power against a whole class of physically relevant alternatives.

To compare methods we consider the following benchmark problem. A measurement $N = n_0$ is made of a Poisson variate N whose mean is the sum of a background strength ν and a signal strength μ:

$$\mathbb{P}\mathrm{r}(N = n_0) \;=\; \frac{(\nu+\mu)^{n_0}}{n_0!}\, e^{-\nu-\mu}. \tag{2}$$

We wish to test

$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu > 0. \tag{3}$$

Since large values of n_0 are evidence against H_0 in the direction of H_1, the p value is simply:

$$p \;=\; \sum_{n=n_0}^{+\infty} \frac{\nu^n}{n!}\, e^{-\nu}, \tag{4}$$

and requires knowledge of ν to be computed. A frequent situation is that only partial information is available about ν, either from an auxiliary measurement or from a Bayesian prior. In the next subsections we examine six methods for incorporating such information in the calculation of p: conditioning, supremum, confidence set, bootstrap, prior-predictive, and posterior-predictive. In principle all of these methods can be applied to the case where information about ν comes from an actual measurement, but only the last two can handle information in the form of a prior.

3.1 Conditioning Method

Suppose we make an independent, Poisson distributed measurement M of the quantity $\tau\nu$, with τ a known constant. The conditional distribution of N, given a fixed value s_0 of the sum $S \equiv N + M$, is binomial:

$$\begin{aligned} \mathbb{P}\mathrm{r}(N = n_0 \,|\, S = s_0) \;&=\; \frac{\mathbb{P}\mathrm{r}(N = n_0 \text{ and } S = s_0)}{\mathbb{P}\mathrm{r}(S = s_0)} \;=\; \frac{\mathbb{P}\mathrm{r}(N = n_0)\,\mathbb{P}\mathrm{r}(M = s_0 - n_0)}{\mathbb{P}\mathrm{r}(S = s_0)} \\ &=\; \binom{s_0}{n_0} \left(\frac{1 + \mu/\nu}{1 + \mu/\nu + \tau}\right)^{n_0} \left(1 - \frac{1 + \mu/\nu}{1 + \mu/\nu + \tau}\right)^{s_0 - n_0}. \end{aligned} \tag{5}$$

Under the null hypothesis that $\mu = 0$ this distribution is independent of the nuisance parameter ν and can therefore be used to compute a p value:

$$p_{cond} \;=\; \sum_{n=n_0}^{s_0} \binom{s_0}{n} \left(\frac{1}{1 + \tau}\right)^{n} \left(1 - \frac{1}{1 + \tau}\right)^{s_0 - n}. \tag{6}$$

Because of the discreteness of the measurements N and M, p_{cond} is by construction a conservative p value. For continuous measurements it would be exact.

Note that tail probabilities of the distribution (5) cannot be used to construct confidence intervals for μ under H_1, since the dependence on ν is only eliminated under H_0. Such a limitation does not exist when the mean of the Poisson variate N is the product rather than the sum of μ and ν. The product case leads to a well-known technique for calculating confidence intervals on the ratio of two Poisson means.

As illustrated above, the conditioning method requires the existence of a statistic S that is sufficient for the nuisance parameter under the null hypothesis. This special structure is not present in most problems encountered in high energy physics. Although other special structures are sometimes available, it is clear that a more universal approach is needed for routine applications.

3.2 Supremum Method

A very general technique consists in maximizing the p value with respect to the nuisance parameter(s):

$$p_{\sup} \;=\; \sup_{\nu} p(\nu). \tag{7}$$

It may happen that the supremum is reached at some value $\nu_{\max}$ within the interior of the ν region allowed by the null hypothesis. In this case $p_{\sup} = p(\nu_{\max})$. Clearly, $\nu_{\max}$ is in no sense a valid estimate

of the true value of ν. Hence, the supremum method should not be confused with "profiling", which consists in substituting the maximum likelihood estimate of ν in $p(\nu)$, and which will be discussed as one of the bootstrap methods in section 3.4.

In contrast with p_{cond}, $p_{\sup}$ is not a tail probability. It is conservative by construction, and may yield the trivial result $p_{\sup} = 1$ if one is not careful in the choice of test statistic. In general the likelihood ratio is a good choice. Suppose for example that in our benchmark problem information about ν is available in the form of a Gaussian measurement $X = x_0$ with mean ν and known standard deviation $\Delta\nu$ (in this form the problem cannot be solved by the conditioning method). The likelihood function is then:

$$\mathcal{L}(\nu, \mu \,|\, n_0, x_0) \;=\; \frac{(\nu+\mu)^{n_0}\, e^{-\nu-\mu}}{n_0!}\; \frac{e^{-\frac{1}{2}\left(\frac{x_0-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu}. \tag{8}$$

We assume that x_0 can take on negative as well as positive values due to resolution effects. The likelihood ratio statistic is:

$$\lambda(n_0, x_0) \;=\; \frac{\sup_{[\,\nu\geq 0\ \&\ \mu=0\,]}\ \mathcal{L}(\nu, \mu \,|\, n_0, x_0)}{\sup_{[\,\nu\geq 0\ \&\ \mu\geq 0\,]}\ \mathcal{L}(\nu, \mu \,|\, n_0, x_0)} \;=\; \frac{\mathcal{L}\big(\hat{\hat{\nu}}, 0 \,\big|\, n_0, x_0\big)}{\mathcal{L}\big(\hat{\mu}, \hat{\nu} \,\big|\, n_0, x_0\big)}, \tag{9}$$

where $\hat{\hat{\nu}}$ is the maximum likelihood estimate of ν under the constraint of the null hypothesis:

$$\hat{\hat{\nu}} \;=\; \frac{x_0 - \Delta\nu^2}{2} \;+\; \sqrt{\left(\frac{x_0 - \Delta\nu^2}{2}\right)^2 \;+\; n_0\,\Delta\nu^2}, \tag{10}$$

and $(\hat{\mu}, \hat{\nu})$ the unconditional maximum likelihood estimate of (μ, ν):

$$(\hat{\mu}\,,\,\hat{\nu}) = \begin{cases} (n_0\,,\,0) & \text{if} \quad x_0 < 0, \\ (n_0 - x_0\,,\,x_0) & \text{if} \quad 0 \leq x_0 \leq n_0, \\ (0\,,\,\hat{\hat{\nu}}) & \text{if} \quad x_0 > n_0. \end{cases} \tag{11}$$

Plugging $\hat{\hat{\nu}}$, $\hat{\nu}$, and $\hat{\mu}$ into equation (9) and taking twice the negative logarithm yields finally:

$$-2\ln\lambda(n_0, x_0) = \begin{cases} 2n_0 \ln(n_0/\hat{\hat{\nu}}) - \hat{\hat{\nu}}^2/\Delta\nu^2 & \text{if} \quad x_0 < 0, \\ 2n_0 \ln(n_0/\hat{\hat{\nu}}) - (\hat{\hat{\nu}}^2 - x_0^2)/\Delta\nu^2 & \text{if} \quad 0 \leq x_0 \leq n_0, \\ 0 & \text{if} \quad x_0 > n_0. \end{cases} \tag{12}$$

Tail probabilities of the distribution of $-2\ln\lambda$ under the null hypothesis are easily calculable by numerical methods. Setting $q_0 \equiv -2\ln\lambda(n_0, x_0)$, the observed value of $-2\ln\lambda$, we have:

$$\mathbb{P}\mathrm{r}\Big[-2\ln\lambda(N, X) \geq q_0 \,\Big|\, \mu = 0, \nu \geq 0\Big] \;=\; \sum_{\substack{n \\ -2\ln\lambda(n,x)\,\geq\, q_0}} \int dx \quad \frac{\nu^n\, e^{-\nu}}{n!}\; \frac{e^{-\frac{1}{2}\left(\frac{x-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu}. \tag{13}$$

The x derivative of $-2\ln\lambda(n, x)$ is strictly negative in the region $x < n$; one can therefore implicitly define a function $\tilde{x}(n, q_0)$ by the equation

$$-2\ln\lambda(n, \tilde{x}(n, q_0)) \;=\; q_0 \quad \text{for } q_0 > 0, \tag{14}$$

which can be solved numerically. The integration region $-2\ln\lambda(n, x) \geq q_0$ is equivalent with $x \leq \tilde{x}(n, q_0)$, so that the expression for the tail probability simplifies to:

$$\mathbb{P}\mathrm{r}(-2\ln\lambda(N, X) \geq q_0 \,|\, \mu = 0, \nu) \;=\; \begin{cases} \displaystyle\sum_{n=1}^{+\infty} \frac{\nu^n\, e^{-\nu}}{n!}\; \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\tilde{x}(n, q_0) - \nu}{\sqrt{2}\,\Delta\nu}\right)\right] & \text{if } q_0 > 0, \\ 1 & \text{if } q_0 = 0. \end{cases} \tag{15}$$

According to equation (7), the dependence of this tail probability on ν is eliminated by taking the supremum with respect to ν.

For $\Delta\nu$ values of order 1 or larger, a graphical examination of eq. (15) shows that $-2\ln\lambda$ is stochastically increasing with ν, so that the supremum (7) equals the limiting p value:

$$p_\infty \;=\; \lim_{\nu\to+\infty} \mathbb{P}\mathrm{r}(-2\ln\lambda(N,X) \geq q_0 \,|\, \mu=0,\nu). \tag{16}$$

The distribution of $-2\ln\lambda$ in the large ν limit is described by asymptotic theory. In the present case, care must be taken of the fact that the null hypothesis, $\mu = 0$, lies on the boundary of the physical parameter space, $\mu \geq 0$. The correct asymptotic result is that, under H_0, half a unit of probability is carried by the singleton $\{-2\ln\lambda = 0\}$, and the other half is distributed as a chisquared with one degree of freedom over $0 < -2\ln\lambda < +\infty$; this distribution is sometimes written as $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. Thus, for $q_0 > 0$, p_∞ equals half the tail probability to the right of q_0 in a χ^2_1 distribution.

For small values of $\Delta\nu$, the discreteness of n causes the tail probabilities of $-2\ln\lambda$ to oscillate as a function of ν, and their supremum to slightly exceed the asymptotic value. In this case the correct supremum is much more difficult to find, although p_∞ is often a useful approximation.

3.3 Confidence Set Method

The supremum method has two significant drawbacks. The first one is computational, in that it is often difficult to locate the global maximum of the relevant tail probability over the entire range of the nuisance parameter ν. Secondly, the very data one is analyzing often contain information about the true value of ν, so that it makes little sense conceptually to maximize over all values of ν. A simple way around these drawbacks is to maximize over a $1-\beta$ confidence set C_β for ν, and then correct the p value for the fact that β is not zero [7, 8]:

$$p_{cset} \;=\; \sup_{\nu\in C_\beta} p(\nu) \;+\; \beta. \tag{17}$$

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

$$\mathbb{P}\mathrm{r}(p_{cset} \leq \alpha) \;\leq\; \alpha \quad \text{for all } \alpha \in [0,1]. \tag{18}$$

We emphasize that this inequality is only true if β is chosen before looking at the data. Since p_{cset} is never smaller than β, the latter should be chosen suitably small. If one is using a 5σ discovery threshold for example ($\alpha = 5.7\times10^{-7}$), then it would be reasonable to take a 6σ confidence interval for ν, i.e. $\beta = 1.97\times10^{-9}$. Constructing an interval of such high confidence level may be difficult however, as one rarely has reliable knowledge of the relevant distributions so far out in the tails.

3.4 Bootstrap Methods

Conceptually the simplest method for handling the nuisance parameter ν in the p value (4) is to substitute an estimate for it. This is known as a parametric bootstrap, or plug-in method. Estimation of ν should be done under the null hypothesis, to maintain consistency with the general definition of p values. For example, in the case where information about ν comes from an auxiliary Gaussian measurement, one should use the $\hat{\hat{\nu}}$ estimate of equation (10). The plug-in p value is thus:

$$p_{plug} \;=\; \sum_{n=n_0}^{+\infty} \frac{\hat{\hat{\nu}}(n_0,x_0)^n}{n!}\, e^{-\hat{\hat{\nu}}(n_0,x_0)}. \tag{19}$$

Two criticisms can be levelled at this method. First, it makes double use of the data, once to estimate the nuisance parameter under H_0, and then again to calculate the tail probability. This tends to favor H_0.

Second, it does not take into account the uncertainty on the parameter estimate. This tends to exaggerate the significance and hence works against H_0. There are several ways for correcting these deficiencies.

One option is to base the plug-in estimate of ν on the auxiliary measurement x_0 only, in order to avoid potential signal contamination from the observation n_0. This is equivalent to extracting the estimate of ν from the conditional distribution of the data given the test statistic n_0, and the resulting p value is therefore referred to as a conditional plug-in p value. Although this method avoids double use of the data, it still ignores the uncertainty on the estimate of ν, which can lead to significant undercoverage.

A better way is to adjust the plug-in p value by the following procedure. Let $F_{plug}(p_{plug} \,|\, \nu)$ be the cumulative distribution function of p_{plug}. It depends on the nuisance parameter ν, whose value we don't know. However, we can estimate it, and substitute that estimate in F_{plug}. This yields the so-called adjusted plug-in p value:

$$p_{plug,adj} \;=\; F_{plug}(p_{plug} \,|\, \hat{\hat{\nu}}). \tag{20}$$

This adjustment algorithm is known as a double parametric bootstrap and can be implemented by a Monte Carlo calculation [9]. Since double bootstrap calculations tend to require large amounts of computing resources, methods have been developed to speed them up [10, 11, 12].

Another way to correct the plug-in p value is to work with a different test statistic. Ideally one would like to use a test statistic that is pivotal, i.e. whose distribution under the null hypothesis does not depend on any unknown parameters. Often this is not possible, but an *asymptotically* pivotal statistic can be found; this is then still better than a non-pivotal statistic. The test statistic used above for p_{plug}, namely n, is clearly not pivotal, not even asymptotically. However, twice the negative log-likelihood ratio, equation (12), is asymptotically pivotal, having a $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ distribution in the large-sample limit. The parametric bootstrap evaluation of the likelihood ratio p value consists in substituting $\hat{\hat{\nu}}$ for ν in equation (15). We will write $p_{plug,\lambda}$ for this plug-in p value, to distinguish it from the one based on n. In finite samples, $p_{plug,\lambda}$ is usually more accurate than the asymptotic p value p_∞ (eq. 16).

Figure 1 compares the relative coverage error, $R \equiv (\alpha - \mathbb{P}\mathrm{r}(p \leq \alpha))/\alpha$, of p_{plug}, $p_{plug,adj}$, $p_{plug,\lambda}$, and p_∞ for our Poisson benchmark problem with a Gaussian uncertainty $\Delta\nu$ on ν. Positive values of R indicate overcoverage, negative ones undercoverage. Exactly uniform p values have $R = 0$. In terms of uniformity, $p_{plug,\lambda}$ performs best, followed by $p_{plug,adj}$, p_∞, and p_{plug}, in that order. The first two exhibit some minor undercoverage, which varies with the value of $\Delta\nu$.

An interesting alternative to the bootstrap, known as Bartlett adjustment, can be applied to any log-likelihood ratio statistic T whose asymptotic distribution under the null hypothesis is chisquared with k degrees of freedom. In finite samples one assumes that T is distributed as a *scaled* chisquared variate with expectation value $\langle T \rangle = k\,(1 + B)$, where B goes to zero with increasing sample size. An estimate of B can be extracted from a Monte Carlo calculation of $\langle T \rangle$, in which unknown parameters are replaced by their maximum likelihood estimates under H_0. For continuous data it turns out that the Bartlett-adjusted statistic $T/(1 + B)$ is a better approximation than T to a chisquared statistic with k degrees of freedom. For discrete data the improvement is less consistent.

3.5 Prior-predictive Method

The last two nuisance parameter elimination methods we examine are inspired by a Bayesian approach to model selection. It is assumed that information about the nuisance parameter ν is available in the form of a prior distribution $\pi(\nu)$. The prior-predictive method consists in averaging the p value $p(\nu)$ over this prior:

$$p_{prior} \;=\; \int d\nu\; \pi(\nu)\, p(\nu). \tag{21}$$

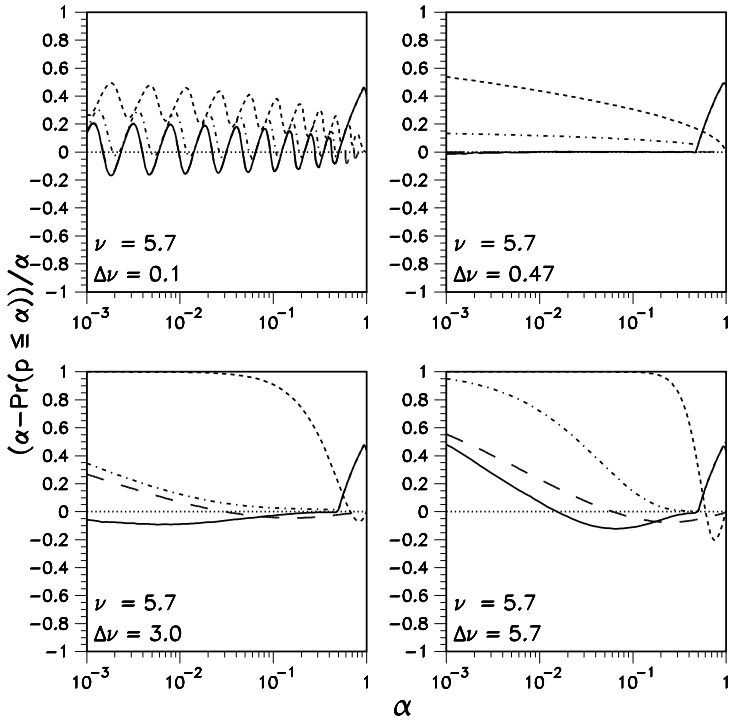

Fig. 1: Relative coverage error, $R \equiv (\alpha - \mathbb{P}\mathrm{r}(p \leq \alpha))/\alpha$ versus significance threshold α, for p_{plug} (short dashes), $p_{plug,adj}$ (long dashes), $p_{plug,\lambda}$ (solid), and p_∞ (dot-dashes). The dotted lines represent zero relative error. The values of ν and $\Delta\nu$ used to generate the reference ensemble are indicated in the lower-left corner of each plot. R values for $p_{plug,adj}$ and $p_{plug,\lambda}$ are almost indistinguishable for $\Delta\nu = 0.1$ and 0.47.

Substituting expression (4) from our benchmark example into the above integral, and interchanging the order of integration and summation, yields:

$$p_{prior} \;=\; \sum_{n=n_0}^{+\infty} m_{prior}(n), \quad \text{where} \quad m_{prior}(n) \;=\; \int d\nu\; \pi(\nu)\, \frac{\nu^n\, e^{-\nu}}{n!}, \tag{22}$$

showing that p_{prior} is itself the tail probability of a distribution, namely the prior-predictive distribution $m_{prior}(n)$ [13]. The latter characterizes the ensemble of all experimental results that one could obtain, taking into account prior uncertainties about the model parameters. A small value of p_{prior} is therefore evidence against the overall model used to describe the data, and could in principle be caused by a badly elicited prior as well as by an invalid likelihood (or unlikely data).

Despite its Bayesian motivation, the prior-predictive p value can be used to analyze frequentist problems. If prior information about ν comes from a bona fide auxiliary measurement with likelihood $\mathcal{L}_{aux}(\nu \,|\, x_0)$, the prior $\pi(\nu)$ can be derived as the posterior for that measurement:

$$\pi(\nu) \;\equiv\; \pi_{aux}(\nu \,|\, x_0) \;=\; \frac{\mathcal{L}_{aux}(\nu \,|\, x_0)\, \pi_{aux}(\nu)}{\int d\nu\; \mathcal{L}_{aux}(\nu \,|\, x_0)\, \pi_{aux}(\nu)}, \tag{23}$$

where the auxiliary measurement prior $\pi_{aux}(\nu)$ is in some sense noninformative or neutral. For example, the testing problem discussed in section 3.2 can be analyzed this way, with $\mathcal{L}_{aux}(\nu \,|\, x_0)$ a Gaussian likelihood. Choosing a flat prior for $\pi_{aux}(\nu)$, truncated to positive values of ν, leads to:

$$\pi(\nu) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{\nu - x_0}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu\,\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x_0}{\sqrt{2}\,\Delta\nu}\right)\right]}. \tag{24}$$

Inserting this prior and the Poisson p value (4) in equation (21) yields, after some simple algebra:

$$p_{prior} \;=\; \begin{cases} \displaystyle\int_0^{+\infty} du \;\; \frac{1+\text{erf}\left(\frac{x_0-u}{\sqrt{2}\Delta\nu}\right)}{1+\text{erf}\left(\frac{x_0}{\sqrt{2}\Delta\nu}\right)} \; \frac{u^{n_0-1}\, e^{-u}}{(n_0-1)!} & \text{if } n_0 > 0, \\ 1 & \text{if } n_0 = 0. \end{cases} \tag{25}$$

In this type of application it is interesting to study the characteristics of p_{prior} with respect to a purely frequentist ensemble, in which both n_0 and x_0 fluctuate according to their uncertainties. This is to be contrasted with the prior-predictive ensemble, where n_0 and ν fluctuate, and with respect to which p_{prior} is exactly uniform by construction. Figure 2 shows the behaviour of the prior-predictive p value (25) with respect to the frequentist ensemble. It appears to be everywhere conservative.

3.6 Posterior-predictive Method

The prior-predictive p value is undefined when $\pi(\nu)$ is improper. A possible way to overcome this problem is to average the p value over the posterior $\pi(\nu \,|\, n_0)$ instead of the prior $\pi(\nu)$ [14]:

$$p_{post} \;=\; \int d\nu \; \pi(\nu \,|\, n_0)\, p(\nu), \quad \text{with} \quad \pi(\nu \,|\, n_0) \equiv \frac{\mathcal{L}(\nu \,|\, n_0)\, \pi(\nu)}{m_{prior}(n_0)}. \tag{26}$$

This posterior-predictive p value is also a tail probability, as can be seen by the same manipulations that led from eq. (21) to eq. (22) in analyzing our Poisson benchmark problem:

$$p_{post} \;=\; \sum_{n=n_0}^{+\infty} m_{post}(n), \quad \text{where} \quad m_{post}(n) \;=\; \int d\nu \; \pi(\nu \,|\, n_0)\, \frac{\nu^n\, e^{-\nu}}{n!}. \tag{27}$$

The posterior-predictive distribution $m_{post}(n)$ is the predicted distribution of n ***after*** having observed n_0. Therefore, p_{post} estimates the probability that a ***future*** observation will be at least as extreme as the current observation if the null hypothesis is true.

The posterior-predictive p value uses the data n_0 twice, first to calculate m_{post} and then again when evaluating p_{post}. As was the case for p_{plug}, this makes p_{post} conservative, increasing the risk of accepting a bad model. The behaviour of p_{post} with respect to the frequentist ensemble for our benchmark problem is compared to that of p_{prior} in Fig. 2. Note that for small values of $\Delta\nu$, inferences about ν are dominated by the prior (24), so that p_{prior} and p_{post} become indistinguishable.

An advantage of p_{post} over p_{prior} is that the former can be used to calibrate discrepancy variables in addition to test statistics. In contrast with statistics, discrepancy variables depend on both data and parameters. A typical example is a sum $D(\vec{x}, \vec{\theta})$ of squared residuals between data $\vec{x}$ and model predictions that depend on unknown parameters $\vec{\theta}$. Whereas a frequentist approach consists in minimizing $D(\vec{x}, \vec{\theta})$ with respect to $\vec{\theta}$, the posterior-predictive approach integrates the joint distribution of $\vec{x}$ and $\vec{\theta}$, given the observed value $\vec{x}_0$ of $\vec{x}$, over all values of $\vec{x}$ and $\vec{\theta}$ that satisfy $D(\vec{x}, \vec{\theta}) \geq D(\vec{x}_0, \vec{\theta})$ [14].

In spite of its advantages, the extreme conservativeness of p_{post} remains troubling and has led some statisticians to propose a recalibration [15] or modified constructions [16].

4 Nuisance parameters that are present only under the alternative

Even though p values are designed to test a single hypothesis (the "null"), they often depend on the general type of alternative envisioned. The benchmark example of section 3 illustrates this, since only positive excursions of the background level are of interest, and negative excursions, no matter how large, are never considered part of the alternative. This clearly affects the calculation of the p value, which depends on one's definition of "more extreme than the observation". As another example, consider

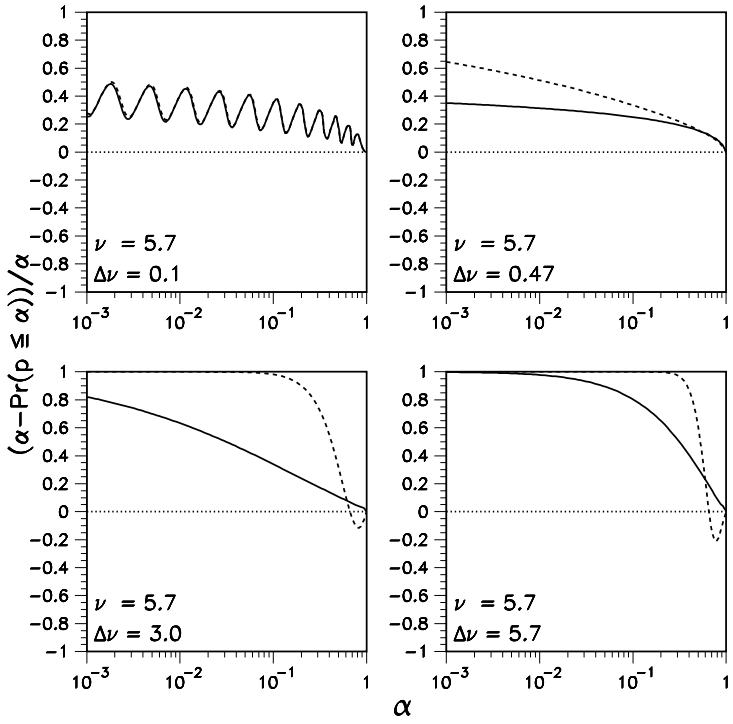

Fig. 2: Relative coverage error $(\alpha - \mathbb{P}\mathrm{r}(p \leq \alpha))/\alpha$ versus significance threshold α, for prior-predictive p values (solid lines) and posterior-predictive p values (dashes). At $\Delta\nu = 0.1$ the two curves are indistinguishable. The dotted lines represent zero relative error.

a positive finding from a search for a signal peak or trough on top of a one-dimensional background spectrum. In this case the alternative hypothesis includes excursions from the background level at any location on the spectrum, not just where the observation was made. The significance of the latter will be degraded due to what physicists call the look-elsewhere effect. Statisticians on the other hand, blame the location parameter of the signal, which they characterize as "a nuisance parameter that is present under the alternative but not under the null" [17, 18]. Since the nuisance parameter is not present under the null, none of the methods described in section 3 can be applied here, and a separate treatment is needed.

As usual, the first step towards a solution consists in choosing an appropriate test statistic. If the signal location θ is known beforehand, the optimal test statistic is simply the likelihood ratio λ. Otherwise λ is a function of θ, and Ref. [19] discusses several ways to eliminate this dependence:

$$\text{SupLR} \equiv \sup_{L \leq \theta \leq U} [-2 \ln \lambda(\theta)], \tag{28a}$$

$$\text{AveLR} \equiv \int_L^U d\theta \; w(\theta) \; [-2 \ln \lambda(\theta)], \tag{28b}$$

$$\text{ExpLR} \equiv \int_L^U d\theta \; w(\theta) \; \exp\left[\frac{1}{2}\,[-2 \ln \lambda(\theta)]\right], \tag{28c}$$

where L and U are the spectrum boundaries and $w(\theta)$ is a weight function. These are two-sided statistics; one-sided versions also exist and do not add any particular difficulty. If there is no prior information about the location of the signal ($w(\theta)$ uniform between L and U), then SupLR and ExpLR appear to be equally good choices, whereas AveLR is significantly less powerful. Note that SupLR is the likelihood ratio statistic when θ is unknown. However, because θ is unidentified under H_0, SupLR does not have the usual asymptotic null distribution, nor does it enjoy the usual asymptotic optimality properties [19].

In general, the null distribution of the selected test statistic has to be obtained by a Monte Carlo

simulation, or a parametric bootstrap if there are unknown parameters under the null. This is often a very complex calculation, in which each simulated dataset must undergo many fits, one under the null hypothesis and several under the alternative, in order to obtain the likelihood ratio as a function of θ. Such a procedure is not easy to automate over the millions of simulated datasets required to prove a 5σ effect, the standard of discovery in high-energy physics.

This computational burden may be somewhat alleviated by using asymptotic approximations when the sample size allows it. For simplicity we illustrate this technique in the case of a binned spectrum with N bins. Background and signal shapes can then be represented by N-vectors whose components are expected bin contents. Suppose that the background spectrum is a linear combination of k independent N-vectors $\vec{b}_i$, whose coefficients are unknown parameters, and that the signal shape is described by N-vector $\vec{s}(\theta)$. We can introduce a metric in the space of N-vectors by defining $\langle \vec{a} \,|\, \vec{b} \rangle \equiv \sum_{i=1}^{N} a_i b_i / \sigma_i^2$, where a_i, b_i are components of the N-vectors $\vec{a}$ and $\vec{b}$, and σ_i is the standard deviation of bin i under the null hypothesis. Let $\vec{v}(\theta)$ be that linear combination of the $\vec{b}_i$ and $\vec{s}(\theta)$ that is orthogonal to each $\vec{b}_i$ and normalized to 1. It can be shown that, asymptotically:

$$-2 \ln \lambda(\theta) \;\sim\; \left[\sum_{i=1}^{N} \frac{v_i(\theta)}{\sigma_i} \, Z_i \right]^2 , \tag{29}$$

where the Z_i are independent standard normal random variables, and the symbol '$\sim$' stands for equality in distribution. For known θ, eq. (29) reduces to $-2 \ln \lambda(\theta) \;\sim\; \chi^2_1$, as expected. For unknown θ, it properly accounts for correlations between values of $-2 \ln \lambda(\theta)$ at different θ locations, an essential requirement for correctly evaluating the statistics (28). That this result simplifies significance calculations is easy to see, since it gives the likelihood ratio without having to fit a spectrum. The vector $\vec{v}(\theta)$ should be constructed over a fine grid of θ values before starting the simulation. Then, to simulate a dataset, one generates N normal deviates Z_i, computes $-2 \ln \lambda(\theta)$, and plugs the result into the desired test statistic, SupLR, ExpLR, or AveLR.

Reference [20] generalizes equation (29) to unbinned likelihoods and non-regular problems other than the one discussed here.

5 Summary

General methods for handling nuisance parameters in p value calculations fall in three categories: worst-case evaluation (supremum or confidence set), bootstrapping, and Bayesian predictive (prior or posterior). The performance of these methods depends strongly on the choice of test statistic, and the likelihood ratio is usually optimal. Of all the methods considered, bootstrapping the likelihood ratio seems the most successful at preserving the uniformity of p values with respect to frequentist ensembles.

Significance problems in high energy physics typically involve many nuisance parameters, not all of which can be handled in the same way. Our understanding of detector energy scales for example, is usually far too complex to be modelled by a likelihood function. A sensible solution is to construct a prior representing this understanding, and assume a prior-predictive approach to incorporate it into a significance. This suggests a hybrid treatment of systematics, where the main effects are handled by bootstrapping a likelihood ratio, whereas auxiliary effects are accounted for with a supremum or predictive method. Such a treatment is well suited to the Monte Carlo approach often necessitated by the complexity of physics analyses.

References

[1] R. Christensen, *Testing Fisher, Neyman, Pearson, and Bayes*, Amer. Statist. **59**, 121 (2005).

[2] R. Royall, *On the probability of observing misleading statistical evidence [with discussion]*, J. Amer. Statist. Assoc. **95**, 760 (2000).

[3] R. E. Kass and A. E. Raftery, *Bayes factors*, J. Amer. Statist. Assoc. **90**, 773 (1995).

[4] P. K. Sinervo, *Definition and treatment of systematic uncertainties in high energy physics and astrophysics*, in *Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics, and Cosmology (PhyStat2003)*, SLAC, Stanford, California, September 8–11, 2003.

[5] D. R. Cox, *The role of significance tests*, Scand. J. Statist. **4**, 49 (1977).

[6] D. R. Cox, *Statistical significance tests*, Br. J. clin. Pharmac. **14**, 325 (1982).

[7] R. L. Berger and D. D. Boos, *P values maximized over a confidence set for the nuisance parameter*, J. Amer. Statist. Assoc. **89**, 1012 (1994).

[8] M. J. Silvapulle, *A test in the presence of nuisance parameters*, J. Amer. Statist. Assoc. **91**, 1690 (1996); correction, ibid. **92**, 801 (1997).

[9] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*, Cambridge University Press, 1997 (582pp.).

[10] M. A. Newton and C. J. Geyer, *Bootstrap recycling: a Monte Carlo alternative to the nested bootstrap*, J. Amer. Statist. Assoc. **89**, 905 (1994).

[11] J. C. Nankervis, *Stopping rules for double bootstrap tests*, Working Paper 03/14, University of Essex Department of Accounting, Finance and Management (2003).

[12] R. Davidson and J. G. MacKinnon, *Improving the reliability of bootstrap tests with the fast double bootstrap*, Computational Statistics and Data Analysis **51**, 3259 (2007).

[13] George E. P. Box, *Sampling and Bayes' inference in scientific modelling and robustness [with discussion]*, J. R. Statist. Soc. A **143**, 383 (1980).

[14] X. L. Meng, *Posterior predictive p-values*, Ann. Statist. **22**, 1142 (1994).

[15] N. L. Hjort, F. A. Dahl, and G. H. Steinbakk, *Post-processing posterior predictive p values*, J. Amer. Statist. Assoc. **101**, 1157 (2006).

[16] M. J. Bayarri and J. O. Berger, *P-values for composite null models [with discussion]*, J. Amer. Statist. Assoc. **95**, 1127 (2000).

[17] R. B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, Biometrika **64**, 247 (1977).

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

[19] D. W. K. Andrews and W. Ploberger, *Optimal tests when a nuisance parameter is present only under the alternative*, Econometrica **62**, 1383 (1994).

[20] D. W. K. Andrews, *Testing when a parameter is on the boundary of the maintained hypothesis*, Econometrica **69**, 683 (2001).

Testing for a Signal

Wolfgang A. Rolke and Angel M. López
University of Puerto Rico - Mayaguez

Abstract
We describe a statistical hypothesis test for the presence of a signal based on the likelihood ratio statistic. We derive the test for one case of interest and also show that for that case the test works very well, even far out in the tails of the distribution. We also study extensions of the test to cases where there are multiple channels.

1 Introduction

In recent years much work has been done on the problem of setting limits, beginning with the seminal paper by Feldman and Cousins [1]. A fairly comprehensive solution for limits on the signal rate in the presence of background and efficiency, both measured with some uncertainty, was given in Rolke, López and Conrad [2]. In this paper we will study a related problem, namely that of claiming a new discovery, say of a new particle or decay mode. Statistically this falls under the heading of hypothesis testing. We will describe a test derived in a fairly standard way called the likelihood ratio test. The main contribution of this paper is the study of the performance of this test. This is essential for two reasons. First, discoveries in high energy physics require a very small false-positive, that is the probability of falsely claiming a discovery has to be very small. This probability, in statistics called the type I error probability α, is sometimes required to be as low as $2.87 \cdot 10^{-7}$, equivalent to a 5σ event. The likelihood ratio test is an approximate test, and whether the approximation works this far out in the tails is a question that needs to be investigated. Secondly, in high energy physics we can often make use of multiple channels, which means we have problems with as many as 30 parameters, 20 of which are nuisance parameters. The sizes of the samples needed to insure that the likelihood ratio test works need to be determined.

2 Likelihood Ratio Test

We will consider the following general problem: we have data $\mathbf{X}$ from a distribution with density $f(\mathbf{x};\theta)$ where θ is a vector of parameters with $\theta \in \Theta$ and Θ is the entire parameter space. We wish to test the null hypothesis $H_0 : \theta \in \Theta_0$ (no signal) vs the alternative hypothesis. $H_a : \theta \in \Theta_0^c$ (some signal), where Θ_0 is some subset of Θ. The likelihood function is defined by

$$L(\theta|\mathbf{x}) = f(\mathbf{x};\theta)$$

and the likelihood ratio test statistic is defined by

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})}$$

Intuitively we can understand the statistic in the case of a discrete random variable. In this case the numerator is the maximum probability of the observed sample if the maximum is taken over all parameters allowed under the null hypothesis. In the denominator we take the maximum over all possible values of the parameter. The ratio of these is small if there are parameter points in the alternative hypothesis for which the observed sample is much more likely than for any parameter point in the null hypothesis. In that case we should reject the null hypothesis. Therefore we define the likelihood ratio test to be: reject the null hypothesis if $\lambda(\mathbf{x}) \leq c$, for some suitably chosen c, which in turn depends on the type I error probability α.

How do we find c? For this we will use the following theorem: under some mild regularity conditions if $\theta \in \Theta_0$ then $-2\log\lambda(\mathbf{x})$ has a chi-square distribution as the sample size $n \to \infty$. The degrees of freedom of the chi-square distribution is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\theta \in \Theta$.

A proof of this theorem is given in Stuart, Ord and Arnold [3] and a nice discussion with examples can be found in Casella and Berger [4].

3 A Specific Example: A Counting Experiment with Background and Efficiency

We begin with a very common type of situation in high energy physics experiments. After suitably chosen cuts we find n events in the signal region, some of which may be signal events. We can model n as a random variable N with a Poisson distribution with rate $es + b$ where b is the background rate, s the signal rate and e the efficiency on the signal. We also have an independent measurement y of the background rate, either from data sidebands or from Monte Carlo and we can model y as a Poisson with rate τb, where τ is the relative size of the sidebands to the signal region or the relative size of the Monte Carlo sample to the data sample, so that y/τ is the point estimate of the background rate in the signal region. Finally we have an independent measurement of the efficiency z, usually from Monte Carlo, and we will model z as a Gaussian with mean e and standard deviation σ_e. So we have the following probability model:

$$N \sim Pois(es + b) \qquad Y \sim Pois(\tau b) \qquad Z \sim N(e, \sigma_e)$$

In this model s is the parameter of interest, e and b are nuisance parameters and τ and σ_e are assumed known. Now the joint density of N, Y and Z is given by

$$f(n, y, z; e, s, b) = \frac{(es+b)^n}{n!} e^{-(es+b)} \frac{(\tau b)^y}{y!} e^{-\tau b} \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{-\frac{1}{2}\frac{(z-e)^2}{\sigma_e^2}}$$

Finding the denominator of the likelihood ratio test statistic λ means finding the maximum likelihood estimators of e, s, b. They are given by $\widehat{s} = n - y/\tau$, $\widehat{b} = y/\tau$ and $\widehat{e} = z$.

We wish to test $H_0 : s = 0$ vs $H_a : s > 0$, so under the null hypothesis we have

$$\begin{aligned} &\log f(n, y, z; 0, b, e) = n\log(b) - \log(n!) - b + \\ &y\log(\tau b) - \log(y!) - (\tau b) - \tfrac{1}{2}\log(2\pi\sigma_e^2) - \tfrac{1}{2}\tfrac{(z-e)^2}{\sigma_e^2} \end{aligned}$$

and we find that this is maximized for $\widetilde{b} = \frac{n+y}{1+\tau}$ and $\widetilde{e} = z$. Now

$$\lambda(n, y, z) = \frac{\sup\ L(0,b,e|n,y,z)}{\sup\ L(s,b,e|n,y,z)} = \frac{f(n,y,z|0,\widetilde{b}\,,\widetilde{e}\,)}{f(n,y,z|\widehat{s}\,,\widehat{b}\,,\widehat{e}\,)} =$$

$$\frac{\left(\frac{n+y}{1+\tau}\right)/n!\,\exp\left(-\frac{n+y}{1+\tau}\right)\left(\tau\frac{n+y}{1+\tau}\right)^{y}/y!\,\exp\left(-\tau\frac{n+y}{1+\tau}\right)\frac{1}{\sqrt{2\pi\sigma_e^2}}\,e^{-\frac{1}{2}\frac{(z-z)^2}{\sigma_e^2}}}{n^n/n!\exp(-n)\,y^y/y!\exp(-y)\frac{1}{\sqrt{2\pi\sigma_e^2}}\,e^{-\frac{1}{2}\frac{(z-z)^2}{\sigma_e^2}}} =$$

$$\frac{\left(\frac{n+y}{1+\tau}\right)^{n+y}\tau^{y}}{n^n\,y^y}$$

One special case of this needs to be studied separately, namely the case $y = 0$. In this case we can not take the logarithm and the maxima above have to be found in a different way. It turns out that the maximum likelihood estimators are $\widehat{s} = n$, $\widehat{b} = 0$, $\widehat{e} = z$, and under the null hypothesis we find $\widetilde{b} = \frac{n}{1+\tau}$ and $\widetilde{e} = z$. With this we find $\lambda(n, 0, z) = (1 + \tau)^{-n}$.

First we note that the test statistic does no involve z, the estimate of the efficiency. This is actually clear: the efficiency is for the detection of signal events, but under the null hypothesis there are none. Of

course the efficiency will affect the power curve: if e is small the observed n will be small and it will be much harder to reject the null hypothesis.

Now from the general theory we know that $-2\log\lambda(N,Y,Z)$ has a chi-square distribution with 1 degree of freedom because in the general model there are 3 free parameters and under the null hypothesis there are 2. So if we denote the test statistic by $L(n,y)$ we get

$$L(n,y) = -2\log\lambda(n,y,z) = \begin{cases} 2\left[n\log(n) + y\log(y) - (n+y)\log\left(\frac{n+y}{1+\tau}\right) - y\log(\tau)\right] & \text{if } y > 0 \\ 2n\log(1+\tau) & \text{if } y = 0 \end{cases}$$

and we have $L(N,Y) \sim \chi^2_1$, approximately.

Obviously we will only claim a disovery if there is an excess of events in the signal region, and so the test becomes: reject H_0 if $n > y/\tau$ and $L(n,y) > c$. Now it can be shown that c is the $(1-2\alpha)$ quantile of a chi-squared distribution with one degree of freedom.

The situation described here has previously been studied in Rolke, López and Conrad [2] in the context of setting limits. They proposed a solution based on the profile likelihood. This solution is closely related to the test described here. In fact it is the confidence interval one finds when inverting the test described above.

4 Multiple Channels

In high energy physics we can sometimes make use of multiple channels. There are a number of possible extensions from one channel. We will consider the following model: there are k channels and we have $N_i \sim Pois(e_i s_i + b_i)$, $Y_i \sim Pois(\tau_i b_i)$, $i = 1,..,k$, all independent. We will again find that the efficiencies do not affect the type I error probability. We will discuss two ways to extend the methods above to multiple channels, both with certain advantages and disadvantages.

4.1 Method 1: (Full LRT)

We can calculate the likelihood ratio statistic for the full model. It turns out that the test statistic L_k is given by

$$L_k(\mathbf{n},\mathbf{y}) = \sum_{i=1}^{k} L(n_i,y_i)I(n_i > y_i/\tau_i)$$

where I is the indicator function, that is $I(n > y/\tau) = 1$ if $n > y/\tau$, and 0 otherwise. In other words the test statistic is simply the sum of the test statistics for each channel separately. The test is then as follows: we reject H_0 if $L_k(\mathbf{n},\mathbf{y}) > c$. It can be shown that the distribution of the test statistic under the null hypothesis is a linear combination of chi-square distributions. Tables of critical values as well as a routine for calculating them are available from the authors.

4.2 Method 2: (Max LRT)

Here we will use the following test: reject H_0 if $M = \max_i\{L(n_i,y_i)I(n_i > y_i/\tau_i\} > c$, that is, we claim a discovery if there is a significant excess of events in any one channel. For this method the critical value c is found using Bonferroni's method, see for example Casella and Berger [4]. We therefore reject H_0 if $M > c$, where c is the $(1-2(1-\sqrt[k]{1-\alpha}))$ quantile of a chi-square distribution with one degree of freedom.

As we shall see soon, which of these two methods performs better depends on the experiment.

5 Performance

How do the above tests perform? In order to be a proper test they first of all have to achieve the nominal type I error probability α. If they do we can then further study their performance by considering their power function $\beta(s)$ given by

$$\beta(s) = P(\text{reject } H_0 | \text{ true signal rate is } s)$$

Of course we have $\alpha = \beta(0)$. $\beta(s)$ gives us the discovery potential, that is the probability of correctly claiming a discovery if the true signal rate is $s > 0$.

In simple cases the true type I error probability α and the power $\beta(s)$ can be calculated explicitly, in more difficult cases we generally need to use Monte Carlo. Moreover, if Monte Carlo is used a technique called importance sampling makes it possible to find the true type I error probability even out at 5σ.

First we will study the true type I error probability as a function of the background rate. In figure 1 we calculate α (expressed in sigma's) for background rates ranging from $b = 5$ to $b = 50$. Here we have used $\tau = 1$ and α corresponding to 3σ, 4σ and 5σ.

It is clear that even for moderate background rates (say $b > 20$) the true type I error is basically the same as the nominal one. For smaller background rates, the method is conservative, that is, the true significance of a signal is actually even higher than the one claimed, and it is therefore safe to use the method even for small b.

In figure 2 we have the power curves for $b = 50$, $\tau = 1$, $e = 1$, s from 0 to 100 and α corresponding to 3σ, 4σ and 5σ. This clearly shows the "penalty" of requiring a discovery threshold of 5σ: at that level the true signal rate has to be 83 for a 90% chance of making a discovery. If 3σ is used a rate of 52 is sufficient, and for 4σ it is 67.

Let us now consider the case of multiple channels. In figure 3 we have the results of the following simulation: There are 5 channels, all with the same background, going from 10 to 100, and the same $\tau = 1$. Again we see that the test achieves the nominal α even for small background rates.

For the last study we will compare the two methods for multiple channels. In figure 4 we have the power curves for the following situations: we have 5 channels with $b = 50$, $e = 1$, and $\tau = 1$ for all channels. In case 1 the signal rate s goes from 0 to 75 and is the same in all channels. In case 2 we have s_1 going from 0 to 100 and $s_2 = .. = s_5 = 0$. All simulations are done using $\alpha = 5\sigma$. Clearly in case 1 Full LRT does better whereas in case 2 it is Max LRT.

This is not surprising because the maximum makes this method more sensitive to the "strongest" channel whereas the sum makes Full LRT more sensitive to a "balance" of the channels. In practice, of course, a decision on which method to use has to be made before any data is seen. A discussion of the optimum strategy for making such a decision is beyond the scope of this paper.

6 Further Extensions

Our extension to multiple channels assumes possibly different signal rates in each channel. The most common situation involves different decay channels of a particle whose existence is being tested. In that case, the different signal rates are due to different branching ratios such that $si = ris$ with a common s. A detailed discussion of this case along with the inclusion of information on certain variables in each event (a technique generally known as marked Poisson) will be found in an upcoming paper.

7 Summary

We have discussed a hypothesis test for the presence of a signal. For the case of a Poisson distributed signal with a background that has either a Poisson or a Gaussian distribution we have carried out the calculations and done an extensive performance study. We have shown that the test achieves the nominal

type I error probability α, even at a 5σ level. We extended the test to the case of multiple channels with two possible tests and showed that both achieve the nominal α. Either one or the other has better performance depending on the specific experiment.

References

[1] G.J. Feldman and R.D. Cousins "A Unified Approach to the Classical Statistical Analysis of Small Signals", *Phys. Rev*, **D57**, (1998) 3873.

[2] W.A. Rolke, A. López and J. Conrad, "Limits and Confidence Intervals in the Presence of Nuisance Parameters", Nuclear Instruments and Methods A, 551/2-3, 2005, pp. 493-503, physics/0403059

[3] A. Stuart, J.K.Ord and S. Arnold, "*Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*", 6^{th} Ed., London Oxford University Press (1999)

[4] G. Casella and R.L. Berger, "*Statistical Inference*", 2^{nd} Ed., Duxbury Press, (2002)

Figure 1

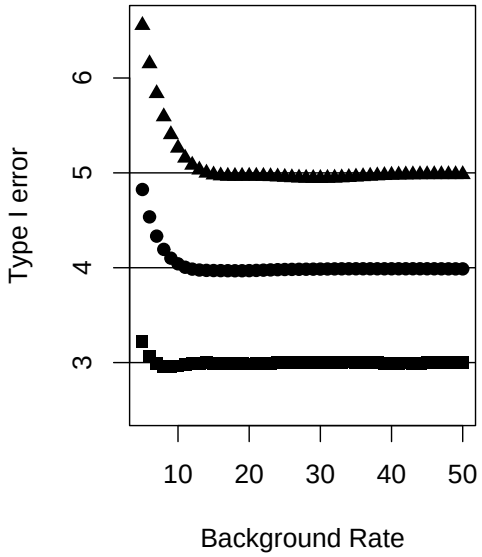

Figure 2

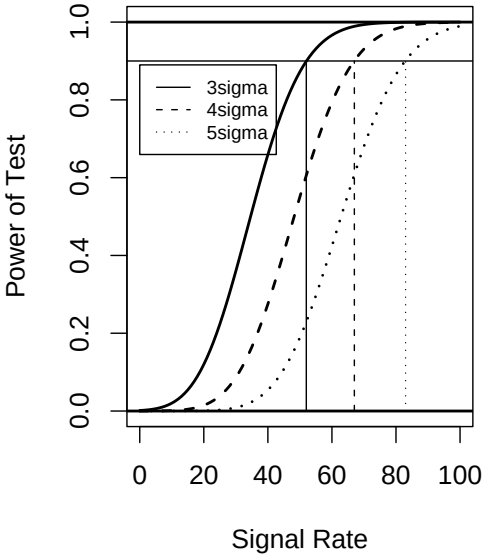

Figure 3

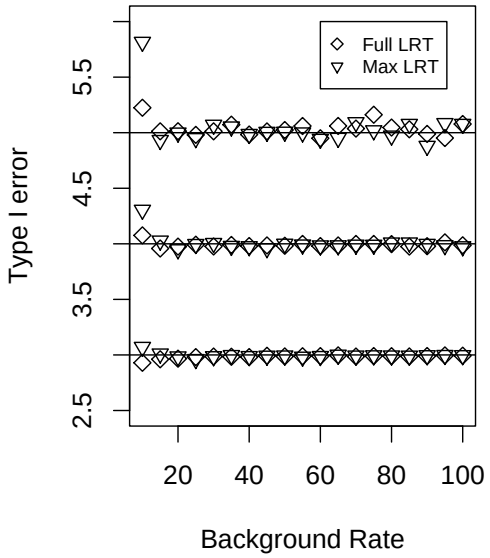

Figure 4

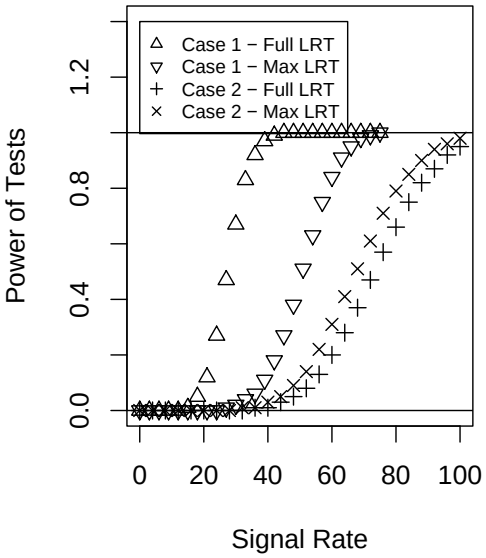

Fig. 1: Type I error probability α for different values of the background rate b

Fig. 2: Power of Test for $b = 50, \tau = 1$

Fig. 3: Type I error probability α for different values of the background rate b for the 5 channel case.

Fig. 4: Power of two methods with 5 channels. Case 1 has equal signal in all channels, case 2 has signal in channel one and no signals in the others.

Evaluation of Two Methods for Incorporating a Systematic Uncertainty into a Test of the Background-only Hypothesis for a Poisson Process

Jordan Tucker
Dept. of Physics and Astronomy, University of California, Los Angeles, California, USA

Abstract
Hypothesis tests for the presence of new sources of Poisson counts amidst background processes are frequently performed in high energy physics, gamma ray astronomy, and other branches of science. This talk briefly summarizes work in which we evaluate two classes of algorithms for dealing with uncertainty in the mean background in such tests.

This talk briefly summarizes studies, performed with Robert Cousins and described in Ref. [1], of two methods for incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process. In a situation common in both gamma-ray astronomy (GRA) and high-energy physics (HEP), $n_{\rm on}$ events are observed from a Poisson process with mean $\mu_{\rm s} + \mu_{\rm b}$; the signal mean $\mu_{\rm s}$ is of interest, while the background mean $\mu_{\rm b}$ is a nuisance parameter. In this work, we study tests of the background-only null hypothesis ($\mu_{\rm s} = 0$) in two prototypical problems in GRA and HEP as follows.

The "on/off" problem. In GRA, $n_{\rm on}$ photons are detected with a telescope pointed on-source, i.e. with some putative source in the field of view; and $n_{\rm off}$ photons are detected with the telescope pointed off-source. The ratio τ of observing time $t_{\rm off}/t_{\rm on}$ is assumed known exactly. In the analogous example from HEP, one counts $n_{\rm on}$ events in a signal region where one is looking for an excess above background. One observes $n_{\rm off}$ events in a background control (sideband) region where no excess is expected. The ratio τ of sideband to signal region events under the background-only null hypothesis is again assumed known.

The "Gaussian-mean background" problem. In another scenario, there is a subsidiary measurement which determines $\mu_{\rm b}$ with normal (Gaussian) uncertainty with rms deviation $\sigma_{\rm b}$. We assume $\sigma_{\rm b}$ to be precisely known, either absolutely, or as a set fraction of $\mu_{\rm b}$.

In either problem, for a data set one can then proceed to calculate the tail probability (p-value) under the null hypothesis. In HEP, one typically quotes the significance S (known in the statistics literature as the Z-value) of the data set, namely the p-value converted to equivalent normal standard deviations.

As detailed by Linnemann [2] at PhyStat 2003, there is an approximate correspondence between the two problems. For the on/off problem, an estimate of the mean background in the signal region is

$$\hat{\mu}_{\rm b} = n_{\rm off}/\tau, \tag{1}$$

and the (rough) uncertainty on this estimate is then

$$\sigma_{\rm b} = \sqrt{n_{\rm off}}/\tau. \tag{2}$$

Combining the two equations and eliminating $n_{\rm off}$, we have

$$\tau = \hat{\mu}_{\rm b}/\sigma_{\rm b}^2. \tag{3}$$

This suggests that a recipe to estimate the significance for one of the prototypical problems can be applied to the other; then the performance of the recipe can be studied. Here, we quantify performance in terms of coverage.

There is a frequentist solution to the on/off problem, discussed by Linnemann [2] at PhyStat 2003 and by a very few references in that work. The key idea is to reformulate the null hypothesis: if the

signal mean $\mu_{\rm s}$ is zero, then the ratio of Poisson means in the sideband region and the signal region is exactly τ (i.e. it is the ratio expected given background alone, and no signal). Then, one can use the standard frequentist solution for the hypothesis test for the ratio of Poisson means, expressed in terms of binomial probabilities. This recipe ($Z_{\rm Bi}$) is then easily carried out, for example in ROOT [3]. In ROOT, one function call returns a p-value, and another calculates the equivalent number of standard deviations, Z. By the properties of the frequentist construction, $Z_{\rm Bi}$ never under-covers, but it over-covers due to the discreteness of n, especially for small counts.

It is common in HEP to integrate out the nuisance parameter (here, the unknown background mean $\mu_{\rm b}$) in an otherwise frequentist calculation (Cousins and Highland [4] integrated out an unknown luminosity). For the Gaussian-mean background problem, starting from the Poisson probability to obtain $n_{\rm on}$ or more background events:

$$p_{\rm P} = \sum_{j=n_{\rm on}}^{\infty} e^{-\mu_{\rm b}} \mu_{\rm b}^{j}/j!, \tag{4}$$

one can calculate the weighted average over a given pdf for the background mean $\mu_{\rm b}$ to obtain the p-value:

$$p = \int p_{\rm P} p(\mu_{\rm b}) {\rm d}\mu_{\rm b}. \tag{5}$$

Then, depending on the pdf one chooses for $\mu_{\rm b}$, one has different recipes to calculate Z-values.

Choosing a gamma function pdf for $\mu_{\rm b}$ (the result of a flat prior times the likelihood function from the Poisson sideband observation of $n_{\rm off}$), one has the recipe Z_Γ. Amazingly, this yields an answer which is identical [2] to that of the frequentist-constructed $Z_{\rm Bi}$!

Letting the pdf for $\mu_{\rm b}$ be a Gaussian with rms deviation $\sigma_{\rm b}$ as above, one obtains the recipe $Z_{\rm N}$ (with the subscript denoting normal). This method was presented in a poster at PhyStat 2005 by Bityukov [5] et al. and was the recommendation out of the CMS Higgs group, adopted by CMS. But, the frequentist coverage of $Z_{\rm N}$ is not guaranteed, and Cranmer [6] gave examples where it was poor.

We check the coverage of the two recipes, $Z_{\rm Bi}$ and $Z_{\rm N}$, scanning over the true background mean $\mu_{\rm b}$ and the other experimental setup parameter (τ for the on/off problem, or $f = \sigma_{\rm b}/\mu_{\rm b}$ for the Gaussian-mean background problem). Choosing a "claimed" Z-value, $Z_{\rm claim}$, from the common choices 1.28 (corresponding to a p-value of 0.1), 3, or 5, we calculate the frequency, in the absence of a signal, that the claimed Z- value is exceeded for an ensemble of experiments with the chosen $\mu_{\rm b}$ and τ or f. This is then converted to the "true" Z-value, $Z_{\rm true}$. We then plot what we call $\Delta Z = Z_{\rm true} - Z_{\rm claim}$; then the coverage is easily checked by looking for deviations above or below $\Delta Z = 0$, corresponding to over or under-coverage, respectively.

We present here just four sample plots showing the results of these scans for a claimed Z-value of 5 (i.e. a claimed p-value of 2.87×10^{-7}). One pair of plots applying $Z_{\rm Bi}$ and $Z_{\rm N}$ to the on/off problem is shown in Fig. 1, and another pair applying the two recipes to the Gaussian-mean background problem for absolute $\sigma_{\rm b}$ is shown in Fig. 2. Plots for other combinations of problems, recipes, claimed Z-values, and for larger values of $\mu_{\rm b}$ and τ or f are in Ref. [1].

For the on/off experiments analyzed using the $Z_{\rm Bi}$ recipe (Fig. 1 (left)), $Z_{\rm true} \geq Z_{\rm claim}$ everywhere, as expected. Using the $Z_{\rm N}$ recipe (Fig. 1 (right)), one gets under-coverage as severe as two units of ΔZ for some regions in the plot. This agrees with the result of Cranmer [6], who (using a Monte Carlo coverage calculation method) finds that for the ensemble of experiments with $\mu_{\rm b} = 100$ and $\tau = 1$ using $Z_{\rm N}$ for the on/off problem under-covers for $Z_{\rm claim} = 5$, obtaining $Z_{\rm true} = 4.2$.

There is significant over-coverage for small values of n, as seen in the lower left corners of the plots; there the discreteness issues come into play as mentioned above. We choose to leave blank those regions in the plot where a p-value less than $\sim 10^{-15}$ is obtained and the precise calculation of Z breaks down due to numerical precision limitations.

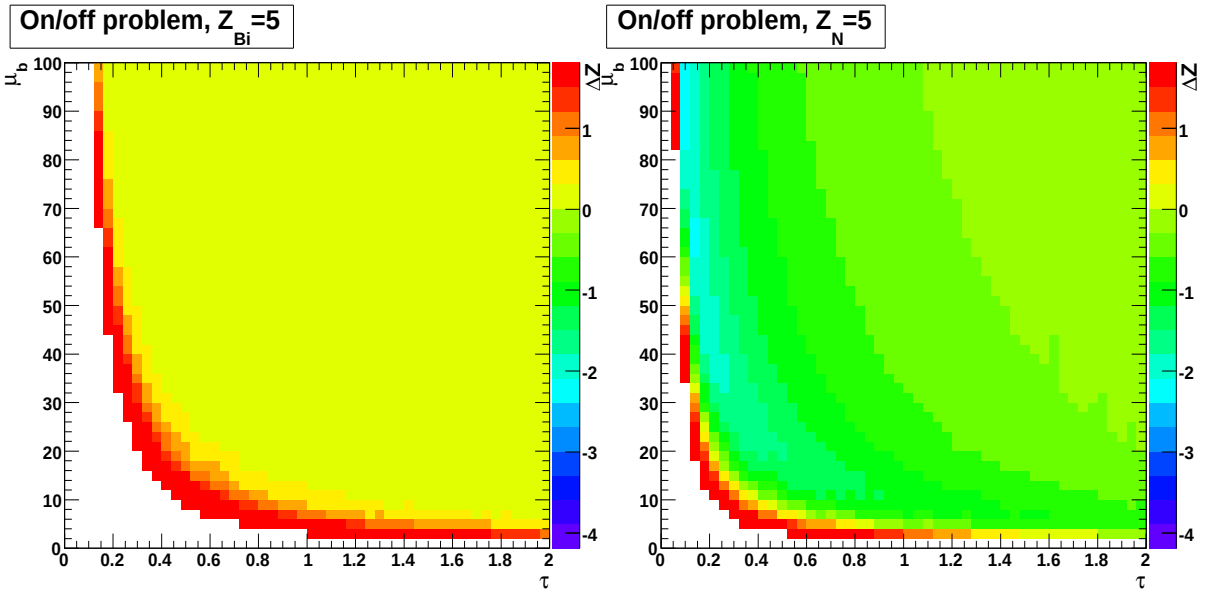

Fig. 1: For the on/off problem analyzed using the $Z_{\rm Bi}$ (left) and $Z_{\rm N}$ (right) recipes, for each fixed value of τ and $\mu_{\rm b}$, the plot indicates the calculated $Z_{\rm true} - Z_{\rm claim}$ for the ensemble of experiments quoting $Z_{\rm claim} \geq 5$. The lower left corner is devoid of entries due to machine round-off, as described in in Ref. [1].

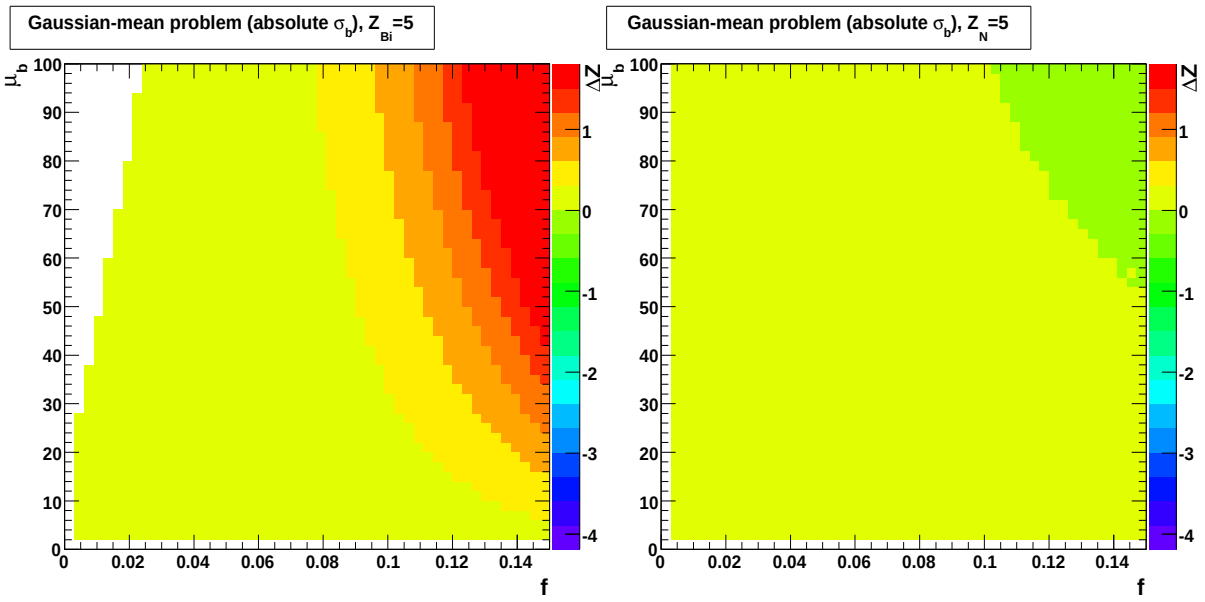

Fig. 2: For the Gaussian-mean background problem with exactly known $\sigma_{\rm b}$, analyzed using the $Z_{\rm Bi}$ (left) and $Z_{\rm N}$ (right) recipes, for each fixed value of $f = \sigma_{\rm b}/\mu_{\rm b}$ and $\mu_{\rm b}$, the plot indicates the calculated $Z_{\rm true} - Z_{\rm claim}$ for the ensemble of experiments quoting $Z_{\rm claim} \geq 5$. The upper left corner of the left plot is again devoid of entries due to machine round-off, as described in the in Ref. [1].

For the Gaussian-mean background experiments with exactly known $\sigma_{\rm b}$ analyzed with $Z_{\rm Bi}$ (Fig. 2 (left)), there is over-coverage everywhere, and by a large amount for increasing values of $f = \sigma_{\rm b}/\mu_{\rm b}$. Using $Z_{\rm N}$, one runs into under-coverage for increasing f and $\mu_{\rm b}$. This under-coverage can be even more severe for other choices of $Z_{\rm claim}$ and values of $\mu_{\rm b}$ and f, as seen in the full set of plots in Ref. [1].

For small values of f and larger values of $\mu_{\rm b}$, using the correspondence Eqn. 3 to approximate the Gaussian-mean background problem as Poisson for calculating $Z_{\rm Bi}$ leads to numerical difficulties as explained in Appendix B of Ref. [1]; therefore we leave the upper left region of the left plot in Fig. 2 blank.

Recommendations. For the on/off problem $Z_{\rm Bi} = Z_\Gamma$ avoids under-coverage by construction, but can be quite conservative for small numbers of events; we recommend $Z_{\rm Bi}$ for general use in this problem. One may wish to use less conservative tests, either ones constructed directly for the ratio of Poisson means and never under-cover, or (as our referee Nancy Reid suggested) less conservative approximate methods for the binomial problem, such as mid p-values.

For the Gaussian-mean background problem, $Z_{\rm Bi}$ works as well as or better than $Z_{\rm N}$ in much of the space, but in this implementation there are numerical issues for very small uncertainties on a large mean background. Since neither $Z_{\rm Bi}$ nor $Z_{\rm N}$ has coverage built in by construction for the Gaussian-mean background problem, one should check the coverage where used.

Of course, it is of interest to extend the studies to other recipes and more complex problems, as previously begun by Tegenfeldt and Conrad [7], and by Rolke, Lopez, and Conrad [8]. For example, we have not yet considered the uncertainty on τ.

We thank Kyle Cranmer and James Linnemann for numerous enlightening discussions. This work was partially supported by the U.S. Department of Energy.

References

[1] Robert D. Cousins and Jordan Tucker, "Evaluation of two methods for incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process", [arXiv:physics/0702156].

[2] James T. Linnemann, "Measures of significance in HEP and astrophysics," Proceedings of PhyStat 2003: Statistical Problems in Particle Physics, Astrophysics, and Cosmology (SLAC, Stanford, California USA Sept. 8-11, 2003) [arXiv:physics/0312059]. `http://www.slac.stanford.edu/econf/C030908/papers/MOBT001.pdf`

[3] ROOT, An Object-Oriented Data Analysis Framework. `http://root.cern.ch`.

[4] R.D. Cousins and V.L. Highland, "Incorporating systematic uncertainties into an upper limit," Nucl. Instrum. Meth. A **320** (1992) 331.

[5] S.I. Bityukov, S.E. Erofeeva, N.V. Krasnikov, A.N. Nikitenko, "Program for Evaluation of Significance, Confidence Intervals and Limits by Direct Calculation of Probabilities", Proceedings of PhyStat 05: Statistical Problems in Particle Physics, Astrophysics and Cosmology (Oxford, Sept. 12-15, 2005), `http://www.physics.ox.ac.uk/phystat05/proceedings/`.

[6] Kyle Cranmer, "Statistical challenges for searches for new physics at the LHC," Proceedings of PhyStat 05: Statistical Problems in Particle Physics, Astrophysics and Cosmology (Oxford, Sept. 12-15, 2005) [arXiv:physics/0511028].

[7] Fredrik Tegenfeldt and Jan Conrad, "On Bayesian treatment of systematic uncertainties in confidence interval calculations," Nucl. Instrum. Meth. A **539** (2005) 407. [arXiv:physics/0408039].

[8] W. A. Rolke, A. M. Lopez and J. Conrad, "Confidence Intervals with Frequentist Treatment of Statistical and Systematic Uncertainties," Nucl. Inst. Meth. A **551** (2005) 493. [arXiv:physics/0403059].

Experimental issues

Statistics for the LHC: Progress, Challenges, and Future

Kyle S. Cranmer
New York University

Abstract

The Large Hadron Collider offers tremendous potential for the discovery of new physics and poses many challenges to the statistical techniques used within High Energy Physics. I will review some of the significant progress that has been made since the PhyStat 2005 conference in Oxford and highlight some of the most important outstanding issues. I will also present some ideas for future developments within the field and advocate a progressive form of publication.

1 Introduction

There are several direct and indirect indications that some type of new physics will show up at the TeV scale – the energy scale being explored by the Large Hadron Collider (LHC) and the multi-purpose detectors ATLAS and CMS. There are a plethora of theoretical models that have been proposed for this new physics; some with few parameters that make specific predictions, some with many parameters and diverse phenomenology[1], and some that are quite vague. For the models that make sharp predictions, I will summarize the substantial progress that has been made recently regarding the statistical procedures used to establish a discovery, and indicate some of the challenges and open issues that remain. In the case of high-dimensional models or models with vague predictions, the statistical challenge is more strategical in nature. I will outline some of the approaches that have been proposed and discuss some new directions that may bear fruit. In addition to the work being done by experimentalists, I will review some of the work being done by the growing community of theorists using sophisticated statistical techniques. I will conclude with some discussion of how the theoretical and experimental communities can improve their communication and speed the iteration cycle needed to interpret signs of new physics.

This paper is largely a continuation of my contribution to PhyStat 2005, and I urge readers to consult those proceedings for a more thorough statement of the problem and introduction to notation [1]. For completeness, it should be said that within High Energy Physics (HEP) we use a theoretical formalism called Quantum Field Theory that allows us to predict the "cross-section" of any particular interaction, which is proportional to the probability that it will occur in a given collision. The number of observed events, n, is Poisson distributed. Distributions of discriminating variables (angles, energies, masses, etc.) of the particles produced in a collision are described as a convolution of fundamental distributions predicted by theory and complicated detector effects that can only be modeled with Monte Carlo techniques. The resulting distributions are generically called "shapes" and are denoted $f(m)$ (where m may have many components). The "Standard Model" is a specific theory that has survived all our tests so far, thus it is our Null, or "background-only", hypothesis. Uncertainties in the detector performance, deficiencies in our theoretical modeling, and finite computational resources lead to uncertainties in our prediction of the background and often force us to resort to an effective description using a parametric model $f(m|\nu)$ instead of Monte Carlo. The parameters ν are nuisance parameters and reflect our uncertainty in the background. Incorporating nuisance parameters into (or eliminating them from) the statistical techniques used to claim discovery is the focus of the next section. The situation is complicated when the signal hypothesis is composite (has additional free parameters). When there are relatively few parameters in the signal model (eg. $\{m_H\}$ or $\{m_A, \tan\beta\}$) we refer to the problem as the "look elsewhere effect". A more severe form of this problem occurs when the model space has many parameters (e.g. $\{m_0, M_{1/2}, A_0, \tan\beta\}$, the 105 parameters of the MSSM, or the even larger set of models in hep-ph), calling for a more radical approach.

[1] Phenomenology in this context refers to the expected signature of new physics.

2 Searches for Specific Signatures

The claim of discovery of new physics is a statement that the data are inconsistent with our current Standard Model to a high degree. Often, the signature of new physics is evidence of an excess in some distribution above the background. Compared to recent experiments, the LHC experiments have a combination of large background uncertainties and an enormous discovery potential. The large background uncertainties are largely due to the fact that the machine collides protons, which are not fundamental objects, and because we will probe new kinematic regimes. There has been a noble effort by the theoretical community to model these effects and improve Monte Carlo tools; however, it is expected that there will still be significant uncertainties in the rate and shape of the various backgrounds to new physics searches [2, 3].

The expected number of events from background processes is typically denoted b, and b is used as a subscript when needed.[2] Hence the model for our null hypothesis has the form of a "marked Poisson" and can be written

$$L(\mathbf{m}|H_0) = \text{Pois}(n|b) \prod_j^n f_b(m_j;\nu), \tag{1}$$

where ν represents the nuisance parameters used to incorporate uncertainty in our background model and the boldface $\mathbf{m}$ is used to indicate we have a measurement of m for each of the n events.[3] Similarly, the signal is often manifest as an excess above the background, with an expected rate and shape, denoted s and $f_s(m)$. Thus when the signal is purely additive, the model for the alternate hypothesis can be written

$$L(\mathbf{m}|H_1) = \text{Pois}(n|s+b) \prod_j^n f_{s+b}(m_j;\nu) = \text{Pois}(n|s+b) \prod_j^n \frac{s f_s(m_j) + b f_b(m_j;\nu)}{s+b}. \tag{2}$$

When the signal is not additivie (eg. in cases like the Z' where interference effects lead to a deficit) the shape for the alternate, $f_{s+b}(m_j)$, is not a simple mixture model. Often the signal model also has free parameters, but that complication is deferred to Section 2.4.

For quite some time, High Energy Physics has been aware of the Neyman-Pearson lemma and heavily utilized the event-wise likelihood ratio $L(m_j|H_1)/L(m_j|H_0)$ for the selection of signal candidates or the experiment-wise likelihood ratio $L(\mathbf{m}|H_1)/L(\mathbf{m}|H_0)$ as a test statistic in hypothesis testing [4]. The main area of development in the last few years has been the treatment of the nuisance parameters ν and uncertainty in the background rate b [1, 5, 6, 7, 8, 9].

In the LEP Higgs searches, background shapes were known quite well, and shape uncertainties were essentially neglected – or, more accurately, were treated as a systematic error in a way that was decoupled from the rest of the statistical formalism. Normalization uncertainties were included into an otherwise frequentist calculation by "smearing" the background rate according to some (posterior) distribution. This technique of smearing is fundamentally Bayesian (via integrating or marginalizing the nuisance parameter b), and is referred to by several names including Prior Predictive, Cousins-Highland, Z_N, S_{cP}, etc [10, 11, 12]. Searches at the Tevatron have had to deal with shape uncertainties, and the `MCLimit` program developed by Tom Junk employs a mixture of integration and maximization to eliminate nuisance parameters. The techniques being used by the Tevatron currently appear to be adequate for relatively low significance statements (eg. 2σ limits), but may not have good coverage properties at high significance (eg. the 5σ customary for discovery).

2.1 Number Counting Experiments

Analyses that do not take advantage of shape information are called number-counting analyses, and rely purely on the Poisson nature of the counts. Because there are no other discriminating variables, the

[2] With the exception of s and b, Roman characters are reserved for observable quantities and Latin characters are used for model parameters

[3] L will be used interchangeably for a probability density function and a likelihood function

role of background uncertainty is of utmost importance. There has been considerable attention paid to a prototype problem in which a subsidiary measurement y is used to constrain the background rate and the main measurement x is used to test the presence of signal [11].

$$L(x, y|s, b) = \text{Pois}(x|s + b)\text{Pois}(y|\tau b). \tag{3}$$

In my contribution to PhyStat2005, I compared the coverage of several methods in High Energy Physics for calculating significance. The surprising result from that study was that the Bayesian smearing technique, one of the most common techniques in HEP, significantly undercovered for $b = 100$ and $\tau = 1$, an important regime for the LHC. This result was generalized by Cousins and Tucker [12].

An encouraging result of the study presented in my PhyStat2005 contribution was that the use of the Profile Likelihood Ratio[4]

$$\lambda(s = 0) = \frac{L(x, y|s = 0, \hat{\hat{b}})}{L(x, y|\hat{s}, \hat{b})} \tag{4}$$

together with the assumption that $-2 \log \lambda$ is distributed as χ^2_1 (under the null) had good coverage out to 5σ.[5] It is somewhat surprising that the asymptotic result worked so well even with a single observation (x, y) and modest values of the background rate (the Poisson parameter b). This result has spurred significant interest in use of the Profile Likelihood Ratio for LHC searches since it is capable of dealing with many nuisance parameters, has good coverage properties, and can be implemented with one of our field's most throughly debugged tools: `MINUIT/MINOS` [13].

2.2 Coverage Studies With Shapes

In order to explore the coverage properties of the Profile Likelihood Ratio in the presence of shapes and nuisance parameters associated with the shapes, Jan Conrad and I performed a massive Monte Carlo coverage study. We considered a simple extension of the prototype problem:

$$L(x, y, \mathbf{m}|s, b, \nu) = \text{Pois}(x|s + b)\text{Pois}(y|\tau b) \prod_{j=1}^{x} \frac{s f_s(m_j|\nu) + b f_b(m_j|\nu)}{s + b} \tag{5}$$

where

$$f_s(m|\nu) = \frac{1 - e^{-\nu}}{\nu} e^{\nu(m-1)} \qquad \text{and} \qquad f_b(m|\nu) = \frac{1 - e^{-\nu}}{\nu} e^{-\nu m}.$$

We generated $O(10^8)$ pseudo experiments for several values of s, b, and ν. For each we used `MINUIT` to fit $L(x, y, \mathbf{m}|s = 0, \hat{\hat{b}}, \hat{\hat{\nu}})$ and $L(x, y, \mathbf{m}|\hat{s}, \hat{b}, \hat{\nu})$, and used the asymptotic distribution $-2 \log \lambda \sim \chi^2_1$. We tested coverage for background-only scenarios and signal-plus-background scenarios when the shape parameter was assumed to be known and when it was a nuisance parameter. In each of the cases we studied, the coverage from assuming the asymptotic distribution of $-2 \log \lambda$ was very good. There was some indication that for strong discrimination in the shape (large values of $|\nu|$) that there was some undercoverage. See Tab. 1 for the results of that study. While we had planned to compare the power of the profile likelihood ratio with the Bayesian marginalization technique, that study has not been concluded, partially due to the fact that the Bayesian calculation is much more computationally intensive.

This scenario is somewhat artificial because the nuisance parameter ν is shared between the signal and background contributions. Often, the signal shape has its own nuisance parameters, ν', and those nuisance parameters have no effect on the likelihood when $s = 0$. This results in a non-χ^2 distribution for $-2 \log \lambda$ and requires special care. The look-elsewhere effect is an example of this situation, and one must either rely on another asymptotic distribution, use Monte Carlo to estimate the distribution, or recalibrate the p-value obtained.

[4] The use of a single $\hat{}$ denotes the unconditional maximum likelihood estimate, while the double $\hat{}$ denotes the conditional maximum likelihood estimate under the constraint $s = 0$.

[5] If one constrains $s > 0$, then one expects $-2 \log \lambda \sim 1/2\delta(0) + 1/2\chi^2_1$. The factor of 2 in the p-value has a small influence in the significance expressed in σ when one is testing at the 5σ level.

Table 1: Performance of a 5σ test using the profile likelihood ratio in a simple model including shapes with nuisance parameters. Coverage is expressed in σ and quantifies the probability that the true s was included in the 5σ confidence interval. Power is the probability to reject the $s = 0$ hypothesis at the 5σ level.

s	b	τ	ν	coverage [σ]	power [%]
0	20	1	-1	5.1	-
0	40	4	-4	5.1	-
25	100	1	-1	5.1	1.4
50	100	1	-1	5.0	12
50	100	1	-3	4.8	99

2.3 Combining Search Channels

It is common that a new physics signature is manifest in multiple different particle interaction processes. For instance, if the Higgs boson exists, it is expected to decay into different combinations of final state particles with different rates. When the final state particles are different, a different analysis is required: these are commonly referred to as "channels". Obviously, we have more sensitivity to the new physics signature if we combine the different channels. Combining multiple channels by considering the likelihood ratio for the multiple-channel experiment as a test statistic was used by the LEP Higgs searches and is a widely accepted technique.

The LEP Higgs group combined multiple channels by using the multi-channel likelihood ratio

$$\frac{L(\mathbf{m}|H_1)}{L(\mathbf{m}|H_0)} = \prod_{i\in\text{channels}} \left[\frac{\text{Pois}(n_i|s_i+b_i)\ \prod_j^{n_i} \frac{s_i f_{s,i}(m_{j,i})+b_i f_{b,i}(m_{j,i})}{s_i+b_i}}{\text{Pois}(n_i|b_i)\ \prod_j^{n_i} f_{b,i}(m_{j,i})} \right]. \tag{6}$$

It should be noted that here the alternate is a simple model where each of the s_i are known. Moreover, this implies that the relationship of the s_i's is known and is incorporated in the discrimination with the null hypothesis.

As previously mentioned, uncertainty in the background rate, b_i, was included by marginalizing it with respect to some distribution $P(b_i)$, which was taken as a truncated Gaussian and can be considered as a posterior distribution for b_i. The ATLAS experiment used the same formalism to calculate its sensitivity to a low-mass Higgs boson with a pure number counting analysis (eg. no use of f_s or f_b) [14]. Given the undercoverage that was found in this technique of incorporating background uncertainty [1, 12], the combination was repeated using the profile likelihood ratio. In that case, each channel's likelihood function was extended to include an auxiliary, or sideband, measurement y_i that constrains the background via $\text{Pois}(y_i|\tau_i b_i)$. In order to maintain the structure between the different channels (eg. keeping constant the ratios $s_i/s_{i'}$) the s_i were considered to be fixed and an overall signal strength, μ, (related to the production cross-section of the particle) was introduced. Thus, the multi-channel model

$$L(\mathbf{x},\mathbf{y}|\mu,\mathbf{s},\mathbf{b}) = \prod_{i\in\text{channels}} \text{Pois}(x_i|\mu s_i + b_i)\,\text{Pois}(y_i|\tau_i b_i), \tag{7}$$

and the profile likelihood ratio

$$\lambda(\mu=0) = \frac{L(\mathbf{x},\mathbf{y}|\mu=0,\mathbf{s},\hat{\hat{\mathbf{b}}})}{L(\mathbf{x},\mathbf{y}|\hat{\mu},\mathbf{s},\hat{\mathbf{b}})} \tag{8}$$

was used as the test statistic for the combination. See Tab. 2 for a comparison of profile likelihood ratio combination and the marginalization performed with `LEPStats4LHC` [17].

Rolke and López considered the same combination, but used two different approaches [18]. In the technique they called MaxLRT, they considered a discovery to be determined not by the combined likelihood ratio, but solely by the most significant channel. In order to re-calibrate the p-value they

Table 2: Comparison of expected significance of ATLAS Higgs searches with 5 fb^{-1} of data calculated with Prior-Predictive and Profile-Likelihood Ratio. Note, this table is based on previously published ATLAS estimates, but is not itself a result of the ATLAS collaboration. The table is intended to draw attention to the relative difference of the methods rather than the expected significance.

m_H (GeV)	Smearing [σ]	Profile [σ]
110	2.11	1.83
120	3.45	2.43
130	4.76	3.83
140	6.78	5.21
150	8.78	7.45
160	10.43	9.92
170	10.19	9.65
180	8.57	8.02
190	5.77	5.57

made a Bonferroni-type correction. The technique they called FullLRT considered a likelihood ratio as a product of each of the individual channels, but without the notion of an overall signal strength μ. Instead, they took

$$L_{Full}(\mathbf{x},\mathbf{y}|\epsilon,\mathbf{s},\mathbf{b}) = \prod_i \mathrm{Pois}(x_i|\epsilon_i s_i + b_i)\,\mathrm{Pois}(y_i|\tau_i b_i).$$

The p-value for their method is based on the distribution of $-2\log\lambda$ being a linear combination of χ^2 distributions. The main difference in this approach to what was considered in Eq. 8 is that the ratio of unconstrained maximum likelihood estimators $\hat{\epsilon}_i\hat{s}_i$ are not constrained to have the same structure as $\hat{\mu}s_i$. It seems intuitively obvious to me (as a consequence of the Neyman-Pearson lemma) that imposing the additional structure assumed by the alternate hypothesis will translate to additional power, but this has not been confirmed explicitly. Thus, it remains an **open question** if Eq. 8 is more powerful than L_{Full}.

2.4 The Look Elsewhere Effect

So far we have considered the scenario in which the signal model is well specified, and focused on the incorporation of the nuisance parameters ν in the background model. The coverage property that we want our to satisfy is that the rate of Type I error is less than or equal to α *for all* values of the nuisance parameter (eg. $\forall\nu\ \alpha(\nu) < \alpha$). Geometrically, the discovery region corresponds to the *union* of the acceptance regions at every ν, and this union may cause over-coverage for any particular ν.

Now consider the case in which the signal is composite. Let us separately consider signal parameters with physical significance, γ, and those which are more akin to background nuisance parameters, ν_s. If one is interested in γ, then it is not a nuisance parameter and we should expect to represent our results in the $s-\gamma$ (or $\mu-\gamma$) plane. Consider for a moment that γ corresponds to the true mass of the Higgs boson, then our results would be reported in terms of contours in the Higgs cross-section and mass plane. In that case, for every point in the plane, we are asking if the data are consistent with that particular point in the plane. If we restrict ourselves to questions of this form, there is no problem and the relationship between Frequentist confidence intervals and inverted hypothesis tests is clear.

A problem does arrise, however, if one makes a claim of discovery if there is an excess *for any* value of the signal parameter γ (eg. for any mass of the Higgs boson). Clearly, we have a much larger chance to find an excess in a narrow window if we scan the window across a large spectrum. This is often called the "look elsewhere effect", and is typically corrected by scaling the p-value by the "number of places that we looked" or a "trials factor" (often the mass range divided by the mass resolution). This approach of scaling the observed p-value by the trials factor is often called a Bonferroni-type correction.

Formally, one might write "discovery!" $\iff \exists\gamma \ni \forall\nu\ p(\nu,\gamma) < \alpha$. Geometrically, the discovery region now corresponds to *intersection* of the acceptance regions across γ, and this intersection may cause under-coverage for all γ.

In the context of the profile likelihood ratio, one does not explicitly scan γ, but implicitly scans when finding the maximum likelihood estimate $\hat{\gamma}$. The look elsewhere effect is manifest by a non-χ^2 distribution for $-2\log\lambda$. This is known to happen in cases where the alternate model has parameters that do not belong to the null hypothesis (eg. the mass of a new particle sitting on a smooth background). At this conference, Luc Demortier presented various modified asymptotic distributions for $-2\log\lambda$ in these cases [15]. Furthermore, Bill Quayle demonstrated the non-χ^2 distributions via Monte Carlo simulation and proposed an insightful technique to estimate the distribution [16]. It is worth mentioning that the conditions that lead to non-χ^2 distributions for $-2\log\lambda$ seem to only be relevant for discovery, and that all the conditions for a χ^2 distribution are satisfied for measurements of γ or even setting limits on s.

For simple cases (eg. when γ is 1-dimensional), the Bonferroni-type correction is quite straightfoward. In Section 4, I will consider cases in which γ is high-dimensional and the trials factor is either difficult to estimate or leads to a significant loss of power. It is an **open question** whether in simple cases the look elsewhere effect "factorizes" in the sense that a simple re-calibration of the "local" p-value has the same power compared to a method that incorporates the look elsewhere effect in the distribution of the test statistic. A counter example would be equally helpful. Another **open question** is what effect other nuisance parameters in the signal ν_s have on the asymptotic distributions of $-2\log\lambda$.

2.5 Coverage as Calibration & Comparing Multiple Methods

In the last six months, both ATLAS and CMS have created their own statistics committees, and we have already convened joint ATLAS-CMS statistics sessions. One of the outcomes from those discussions was that we plan on using multiple methods for computing the significance of a (hopefully) future observation, and for incorporating systematic errors. As was shown in Ref. [1], the true rate of Type-I error from the different methods may deviate significantly from the nominal value (eg. over- or under-coverage). While one may argue that the accuracy of the coverage at 5σ is irrelevant in absolute terms, it is quite important in relative terms. In particular, we do not want to be in a situation where one experiment requires substantially less data to make a claim of discovery if it is purely due to convention and not because it is actually more powerful. This has furthered the notion that one can think of coverage as a way to "calibrate our statistical apparatus".

Developments such as the RooFit/RooStats framework are being developed to allow us to easily compare different techniques (eg. methods based on the Neyman-Construction, the "profile" construction, profile likelihood ratio, and various Bayesian methods) within the same framework [43].

3 Some Comments on Multivariate Methods

As mentioned in Sec. 2, our field has has been aware of the Neyman-Pearson lemma and heavily utilized the event-wise likelihood ratio $L(m_j|H_1)/L(m_j|H_0)$ for the selection of signal candidates. Here we remind the reader that m_j may be a multi-component discriminating variable and introduce the index k for those d components. To avoid clutter, the event index j will be suppressed. The most basic multivariate analysis, often called "naive Bayes", ignores correlations among the components and builds the event-wise likelihood as a simple (naive) product, viz. $L(m|H_0) = \prod_{k=1}^{d} L(m_k|H_0)$. This technique is very common within HEP, but is rapidly being displaced by other multivariate classifiers like neural networks, decision trees, etc. that can incorporate and leverage non-trivial correlations. It is not surprising that those multivariate classifiers have better performance; however, there are often objections to their "black-box" nature. Furthermore, it is less clear how to incorporate the systematic uncertainties of the Monte Carlo procedures that produced the training data used to train these classifiers. Finally, many of the classifiers have been borrowed from computer science and are optimized with respect to classification

accuracy, a GINI index, or some other heuristic that may not be the most appropriate for the needs of HEP. The next two subsections consider two aspects to multivariate analysis that I hope will complement the other contributions in these proceedings.

3.1 Optimization

Within the context of a search for new physics, one wants to optimize the power of an experiment. In an earlier PhyStat contribution, I introduced the notion of direct and indirect multivariate methods [19, 20]. Essentially, direct methods, such as the genetic programming approach introduced to HEP in Refs. [21, 22], attempt to directly optimize a user-defined performance measure – in this case the power of a 5σ search including background uncertainty. It was shown that many of the common heuristics lead to a function that is at least approximately one-to-one with the likelihood ratio. When neglecting background uncertainty the background hypothesis is no longer composite, both the null and alternate hypotheses are simple, and the Neyman-Pearson lemma holds; thus, in those cases optimization with respect to the heuristic coincides with optimization of power. However, when background uncertainty is taken into account, it is no longer obvious if the heuristic is actually optimizing the power of the search.

A similar point was made by Whiteson and Whiteson when they compared neural networks optimized for classification accuracy to neural networks that were directly optimizing the uncertainty in a top mass measurement [23]. Intuitively, they realized that the top mass measurement is more sensitive to some backgrounds than others, so classification accuracy missed an essential aspect of the problem they were trying to solve. In that case, they found that the uncertainty on the top mass measurement for the classifier that was directly optimizing the mass measurement was $\sim 29\%$ smaller than the networks optimizing classification accuracy. While genetic (a.k.a. evolutionary) strategies easily incorporate user-defined performance measures and direct optimization, many of the "off the shelf" multivariate classifiers from computer science do not. I can only encourage our field to be more aware of this distinction.

The Bayesian Neural Networks that have been advocated by Radford Neal [24] and used in Ref. [26] preserve a clear connection between the statistical goals of the experiment and the optimization of the multivariate classifier. An **open question** is whether the formalism that he uses provides a practical way to incorporate rate and shape uncertainties in the optimization procedure.

3.2 Matrix-Element Methods

Often, the components of the discriminating variable m are kinematic in nature, eg. masses, momenta, angles, or functions of those quantities. These variables are often strongly and non-trivially correlated, which is why multivariate techniques are so powerful. The kinematic quantities and their correlations are well modeled by a theory. By using Feynman diagrams (a perturbative expansion of Quantum Field Theory) we can readily calculate a complex number called the "matrix element" for an arrangement of initial- and final-state particles. The square of the modulus of the matrix element $|\mathcal{M}|^2$ together with phase space dPS and the parton densities D_{parton} predicts the differential cross section $d\sigma/d\vec{r}$ for the kinematic quantities $\vec{r}$, which is proportional to the probability density function $f(\vec{r})$.

In practice, we use particle-level Monte Carlo programs to sample the distribution $f(\vec{r})$. Since we do not measure the kinematic quantities $\vec{r}$ perfectly, we must also simulate the impact of detector effects, which gives us measured quantities $\vec{r}_m$. The final discriminating quantities m are then calculated from the kinematic quantities $\vec{r}_m$. As previously mentioned, the simulation of the detector can be very complicated and we rely on Monte Carlo techniques for the probabilistic mapping $\vec{r} \rightarrow m$. Standard practice has been to use pseudo-data for m as training data for multivariate classifiers, and, as previously mentioned, the resulting classifiers are often regarded as a "black box".

While the detector simulation is very detailed, the probability that a true $\vec{r}$ results in a measured m can often be approximated with a "transfer function" denoted $W(\vec{r}, m)$. Clearly, a more transparent multivariate approach would be to construct a multivariate classifier by numerically confronting the

convolution of the differential cross-section with the transfer function $W(\vec{r}, m)$.

$$\frac{L(m_j|H_1)}{L(m_j|H_0)} = \frac{\int dPS(\vec{r})\, |\mathcal{M}_{s+b}(\vec{r})|^2\, W(\vec{r}, m_j) D_{parton}(\vec{r})}{\int dPS(\vec{r})\, |\mathcal{M}_{b}(\vec{r})|^2\, W(\vec{r}, m_j) D_{parton}(\vec{r})} \tag{9}$$

The success of this method is limited by the accuracy of the transfer function $W(\vec{r}, m)$ and the computational complexity of the convolution. Modern computers now make the numerical convolution tractable, and experience at the Tevatron shows that these "matrix element techniques" are competitive with other multivariate techniques. These matrix element techniques have been used for the most precise measurements of the top mass [25] and in the context of searches for single top [6] are competitive with boosted decision trees and Bayesian neural networks [26, 27].

In addition to experimental applications of the "matrix element technique", the method has substantial capacity to influence phenomenological studies. A large class of phenomenological studies are sensitivity studies, which ask "what is the sensitivity of a given experiment to a given signature predicted by a new theory?". Traditionally, this question is addressed by generating particle-level Monte Carlo for the kinematic quantities $\vec{r}$, smearing those quantities with a parametrized detector response $W(\vec{r}, \vec{r}_m)$, using creativity and insight to find good discriminating variables $m(\vec{r}_m)$, designing a simple cut-analysis to select signal-like events, and then estimating sensitivity with $s/\sqrt{b}$. In cases where the estimated significance is large, then this theory should be taken seriously and studied in more detail by the experimentalists. However, if the estimated significance is low, it is not clear if the experiment is truly not sensitive to this signature for new physics or if the choice of the discriminating variables and the simple cut analyses were just sub-optimal. To avoid this situation, Tilman Plehn and I considered the use of the matrix element technique as in a phenomenological context [28]. Instead of calculating $L(m_j|H_1)/L(m_j|H_0)$ for an observed event's discriminating variables m_j, one can integrate over the joint distribution $f_{s+b}(\vec{r}_m)$ under the alternate hypothesis and calculate an expected significance. Moreover, since the kinematic variables $\vec{r}_m$ encode all the kinematic information, this expected significance provides an upper-bound. If the upper-bound is low, then one can be sure that the experiment truly is not sensitive to the signature for new physics.

It is worth mentioning that the typical procedure in the matrix element method is to integrate over the "true" particles' kinematics $\vec{r}$. This is comfortable for physicists because that is what we do when we calculate cross-sections and we know the phase-space factors associated with $\vec{r}$. Since we are integrating over "true" quantities, this has a Bayesian feel – but it is a use of Bayes theorem that a Frequentist would not mind because we can consider a frequency distribution for $\vec{r}$. Another point of view might be that for this particular event, the particles had some particular true value, and that it doesn't make sense to talk about a sampling distribution for $\vec{r}$. In that vein, one could imagine $\vec{r}$ to be a vector of nuisance parameters for this event, and that one should choose maximization over integration (marginalization). Such techniques have been considered in the context of supersymmetric mass determination [29]. This point of view brings up several **open questions**: a) which method is more powerful? b) how is $-2\log\lambda$ distributed if new nuisance parameters are added to the problem for each of the x events (which is itself a random variable)? and c) is the maximization approach simpler computationally?

4 Challenges of Searches for Beyond the Standard Model Physics

Sections 2 and 3 considered the scenario in which the signal model was well specified, and focused on the incorporation of the nuisance parameters ν in the background model. Section 2.4 considered the case in which the signal is composite, but the dimensionality of physically significant parameters, γ, in the alternate hypothesis was small enough that the "trials factor" can be readily estimated and no severe loss of power is expected by recalibrating the p-value. In this section, we consider the case where the signal model is composite and γ has many parameters and the case in which the signal is quite vague.

[6] The matrix element techniques used in the D0 searches did not perform as well as the other multivariate techniques; however, it is known that they neglected the matrix element for some of the background processes.

Before continuing, it is worth considering a few specific examples. Perhaps the most studied scenario for "beyond the standard model" physics is supersymmetry (SUSY). If supersymmetry exists, it is a broken symmetry in nature. There is no established mechanism for supersymmetry breaking, so the minimal supersymmetric extension to the Standard Model (MSSM) parametrizes all the soft breaking terms with ~ 105 parameters. Thus, one might look at the unconstrained MSSM as more of a theoretical framework than a theory per se. Within this 105 dimensional space, are a few well-motivated subspaces corresponding to particular scenarios for SUSY breaking, eg. mSUGRA parametrized by four real-valued constants and a sign. Even in this restricted parameter space, the signatures for new physics are quite diverse. In general, one is faced with a generic tradeoff between more powerful searches for specific model points and less-powerful but more robust searches. Despite its complexity, supersymmetric models that conserve something called R-parity, in which there is a generic signature of large missing energy in the detector. This allows for a search strategy that is both powerful (enough) and robust (enough); however, other models do not necessarily have an equivalent generic signature.

There are a host of models in addition to supersymmetry that have been proposed to be relevant to the "terra scale" and accessible to the LHC. To give a feeling for the activity, there have been 32,000 papers in hep-ph since 2000. Clearly, even 4000 physicists cannot give due consideration to all of the proposed models in the landscape. As a result, it is interesting to consider more radical approaches and formalisms that can be applied more generically.

4.1 False Discovery Rate

In 1995, Benjamini and Hochberg introduced a technique called False Discovery Rate (FDR) to confront the challenge of multiple hypothesis tests [30]. In contrast to the Bonferroni-type corrections (eg. trials factor), which seeks to control the chance of even a single false discovery among all the tests performed, the FDR method controls the proportion of errors among those tests whose null hypotheses were rejected. Since that time, FDR has become quite popular in astrophysical data analysis [31]. The properties which make FDR popular are

- It has a higher probability of correctly detecting real deviations between model and data.
- It controls a scientifically relevant quantity: the average fraction of false discoveries over the total number of discoveries.
- Only a trivial adjustment to the basic method is required to handle correlated data.

The definition of the false discovery rate is given by

$$FDR = \frac{N^{\text{reject}}_{\text{null true}}}{N^{\text{reject}}} \tag{10}$$

While it seems like a vacuous re-casting of Type-I and Type-II error, it is fundamentally different since it controls a property of a set of discoveries. In brief the FDR technique is implemented by performing N tests, ordering the tests according to the observed p-values (ascending), specifying an acceptable false discovery rate q, and then rejecting the null hypothesis for the first r tests, where r is the largest j that satisfies $p_j < jq/C_N N$. The quantity C_N is only needed if the tests are correlated. This definition means that the value of the threshold on the largest p-value is not known a priori, but is adaptive to the data set. Furthermore, one does not need to specify an alternate hypothesis – though p-values determined from test statistics that are designed with an alternate in mind can be expected to be more powerful.

FDR has not been widely used within High Energy Physics, but it seems to have a natural place in the context of searches for exotica. While our experiments do not have enough resources to study every model that will be proposed, we do have several clever collaborators and an enormous amount of computing power, so we need to address the multiple testing problem. I do not see FDR as being particularly relevant for searches when we expect to claim only a single discovery (eg. the Higgs); however, I do see that it might have a role in the global analysis of LHC data.

4.2 Interpreting New Physics: The Inverse Problem

If we are lucky enough to find evidence of physics beyond the standard model, the next order of business will be to interpret what we have observed and measure the relevant parameters of the new standard model (sometimes called the "inverse problem" [34] historically just called "physics"). Perhaps we will observe new physics that is consistent with an already well-developed theory; perhaps we will observe something consistent with several different theories and we will need to discriminate between them; or perhaps we will observe something more unexpected and be stumped for some time to provide a concise description or even identify the fundamental parameters to be measured.

4.3 Parameter Scans

For cases such as supersymmetry, in which we have a well-developed theory with known fundamental parameters, it is common to simply scan the parameters on a naive grid. At each parameter point in the scan, one might consider an optimized analysis for that point using an automated procedure [32], or test the consistency of the model with data with the aim to measure the fundamental quantities [33]. Naive parameter scans face two problems, one practical and one conceptual. The practical problems are that grid-scans don't scale to high-dimensions and that simple Monte Carlo sampling is not very efficient. Markov Chain Monte Carlo techniques address the practical problems and are addressed below. The conceptual problem is that the space of the parameters does not have a natural metric – why do we take equal steps in $\tan\beta$ and not in β? Information Geometry provides an elegant, though computationally challenging, solution by equipping the space of the parameters with an experimentally relevant metric. Information Geometry may also provide a useful tool for theorists to formalize the "cliffs" and "valleys" in the landscape.

4.3.1 Markov Chain Monte Carlo & Hierarchical Bayes

Markov Chain Monte Carlo (MCMC) has been used successfully (mainly by the collaborations of theorists and experimentalists) to map out the regions of the constrained MSSM (CMSSM) preferred by existing Standard Model and astrophysical measurements [35, 36, 37, 38]. The CMSSM is described by four parameters and a sign. Typically, the MCMC scans provide a Bayesian posterior distribution for the parameters. As Cousins critiqued in his proceedings to this conference, the groups often use flat priors in relatively high dimensions.

Recently, a similar analysis was performed with a more theoretically driven choice for the prior [39]. There, the authors considered the prior probability density for a given SUSY breaking scale M_S:

$$\begin{aligned} p(m_0, M_{1/2}, A_0, \mu, B, sm|M_S) &= p(m_0|M_S)\, p(M_{1/2}|M_S)\, p(A_0|M_S) \\ &\quad p(\mu|M_S)\, p(B|M_S)\, p(sm), \end{aligned} \tag{11}$$

assuming that the SM experimental inputs sm do not depend upon M_S. A particular choice was made relating the SUSY breaking scale M_S and the parameters m_0, $M_{1/2}$, A_0, by relating them to M_S at the "order of magnitude" level:

$$p(m_0|M_S) = \frac{1}{\sqrt{2\pi w^2} m_0} \exp\left(-\frac{1}{2w^2}\log^2(\frac{m_0}{M_S})\right). \tag{12}$$

The parameters A_0 and B are allowed to have positive or negative signs and values may pass through zero, so a prior of a different form was used:

$$p(A_0|M_S) = \frac{1}{\sqrt{2\pi e^{2w}} M_S} \exp\left(-\frac{1}{2(e^{2w})}\frac{A_0^2}{M_S^2}\right). \tag{13}$$

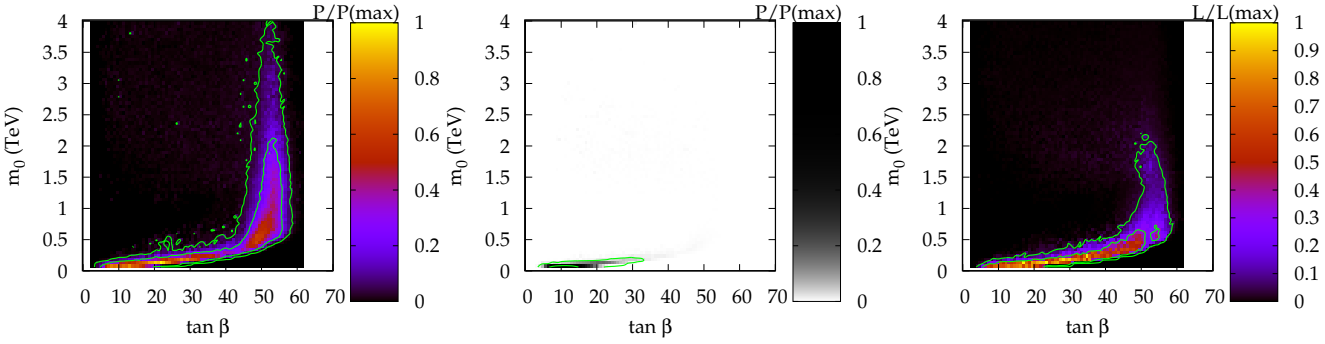

Fig. 1: CMSSM fits marginalised in the unseen dimensions for (a) flat $\tan\beta$ priors, (b) the hierarchical prior with $w = 1$. Figure (c) shows the result of the profile likelihood ratio, in which the unseen dimensions are evaluated at their conditional maximum likelihood values. Contours showing the 68% and 95% regions are shown in each case. The posterior probability in each bin of (a) and (b), normalized to the probability of the maximum bin, is displayed by reference to the color bar on the right hand side of each plot.

Finally, since one does not know M_S a priori, it was treated as a "hyper-parameter" and marginalized giving

$$p(m_0, M_{1/2}, A_0, \mu, B) = \int_0^\infty dM_S\, p(m_0, M_{1/2}, A_0, \mu, B|M_S)\, p(M_S) \tag{14}$$

$$= \frac{1}{(2\pi)^{5/2} w^5 m_0 |\mu| M_{1/2}} \int_0^\infty \frac{dM_S}{M_S^2} \exp\left[-\frac{1}{2w^2}\left(\log^2(\frac{m_0}{M_S}) + \log^2(\frac{|\mu|}{M_S}) + \right.\right.$$
$$\left.\left. \log^2(\frac{M_{1/2}}{M_S}) + \frac{w^2 A_0^2}{e^{2w} M_S^2} + \frac{w^2 B^2}{M_S^2 e^{2w}}\right)\right] p(M_S),$$

where $p(M_S)$ is the prior for M_S itself, which was taken to be flat in the logarithm of M_S. The marginalisation over M_S amounts to a marginalisation over a family of prior distributions, and as such constitutes a hierarchical Bayesian approach. Fig. 1 shows a comparison between the results obtained with flat priors (a) and those obtained with the hierarchical approach (b). As far as I am aware, Ref. [39] is the first example of the use of hierarchical Bayesian techniques in particle physics.

4.3.2 Frequentist Approach

It is clear from Fig. 1 that the choice of prior has a large effect on the results obtained. In the sense of "forecasting" what the LHC might see, the hierarchical approach is playing an important role by injecting our physical insight and sharpening our focus. However, in terms of an experimental result the dependence on a prior is often seen as undesirable, thus it is interesting to consider frequentist approaches.

The MCMC scans were performed in a four-dimensional parameter space, but the figures show two-dimensional projections. In the Bayesian approach, one marginalizes the unseen dimensions with respect to the prior. A frequentist analysis would eliminate the unseen dimensions by maximization instead of marginalization – eg. use the profile likelihood ratio. Fig. 1(c) shows the result of the same analysis with the profile likelihood ratio. We see similar constraints, except that the tail at high $\tan\beta$ up to larger values of $m_0 > 2$ TeV has been suppressed in the profile. From the difference we learn the following facts: in this high $\tan\beta$-high m_0 tail, the fit to data is less good than in other regions of parameter space. However, it has a relatively large volume in unseen dimensions of parameter space, which enhances the posterior probability in Fig. 1(a). The difference between the two plots is therefore a good measure of the so-called "volume effect". While one may argue that flat priors distort the inference by pushing all the probability away from the origin, it is clear that the hierarchical priors had much more of an effect on the inference (reflecting the fact that the data are not dominating the Bayesian inference).

Other groups have performed frequentist analyses of essentially the same problem, though without the use of MCMC to scan the parameter space [40, 41]. In both cases the asymptotic distribution of the profile likelihood ratio was used in constructing confidence intervals. Given the complexity of the likelihood function, it is an **open question** if the asymptotic χ^2 distributions provide good coverage properties for these studies.

4.4 Information Geometry

Information Geometry is a synthesis of statistics and differential geometry. In essence information geometry equips model space with a "natural" metric that is invariant to reparametrization of observables, m, and covariant to reparametrization of theoretical parameters, γ [42].

$$g_{ij}(\gamma) = \int dm f(m;\gamma) \left[\frac{\partial \log f(m;\gamma)}{\partial \gamma_i}\right] \left[\frac{\partial \log f(m;\gamma)}{\partial \gamma_j}\right] \tag{15}$$

By equipping the space of the models with a metric, one can do many powerful things. It has been shown in the context of machine learning that learning algorithms that take equal steps in this natural geometry can converge exponentially faster than one that takes equal steps in the naive parameters of the learning machine. In the context of experimental high energy physics, one can imagine that Information Geometry could make parameter scans significantly more efficient.

Information Geometry may play an even more useful role in theoretical analyses. For instance, the authors of Ref. [34] considered a 15-dimensional supersymmetric model and an exhaustive list of relevant observables. The authors sought to analyze the structure of this space by finding degeneracies (ie. points γ_a and γ_b where the observables are essentially unchanged) and "cliffs" (ie. regions where a small change in γ gives rise to a large change in the observables). These questions could be addressed formally if one had access to the metric $g_{ij}(\gamma)$. Instead, their analysis used a rather ad hoc $\Delta\chi^2$-like discriminant for the observables and a non-invariant Euclidean-like distance for the parameter space γ. While their results seemed quite reasonable, and the degeneracies they found correspond to physically reasonable scenarios, it would be a significant advance if such studies could be formalized.

4.5 Interpretation and The Theory-Experiment Interface

Another challenge of beyond the standard model searches is how to represent the result of an observation and communicate sufficient information to the field. Because of the complexity of some of the models it is not possible to represent the results as a simple one-dimensional likelihood curve or a two-dimensional contour without substantial loss of information. For instance, Fig. 1 only shows two of four interesting dimensions in the theory's parameter space. Ideally, experiments would publish a likelihood map in the full dimensionality of relevant quantities – this is technically possible in many cases [39] and a new feature of the RooFit/RooStats framework [43].

Another issue for publication is model-dependence; it is common for a single experimental signature to be described by several models with different fundamental parameters. It is not feasible for the experiments to report the results tailored for each conceivable model. Instead, experiments prefer to report their results in a model-independent way. In some cases (eg. different models for a Z') there is a model-relevant and model-neutral set of parameters that can be measured, which encompass several different theoretical models, while still providing enough information to distinguish among them. This is an ideal case, but it is not always obvious which measurements are sufficient to distinguish between competing models.

Recently an old theoretical tool (on-shell effective theory) was given a new spin, in the form of a toolset called MARMOSET [44]. MARMOSET is meant to quickly provide a simplified description of new physics, especially in cases where the data are not described by an already well-developed theoretical model. Despite its simplifications, the authors of MARMOSET argue that it does maintain the

essential features of many scenarios for new physics. It is an **open question** if full likelihood maps of the parameters of the best fitting on-shell effective theories provide a general purpose solution for model-neutral and model-relevant publications for the LHC.

5 Conclusion

We are entering a very exciting time for particle physics. The LHC will be probing the "tera-scale", which may reveal the mechanism for electroweak symmetry breaking, new symmetries of nature, and evidence for additional space-time dimensions. The rich landscape of theoretical possibilities and the particularly challenging experimental environment of the LHC place particular emphasis on our statistical techniques. Searches for specific signatures, like the Higgs boson, must address large background uncertainties and consistently combine several search channels. Substantial progress has been made in terms of incorporating systematics in our statistical machinery. Searches for beyond standard model physics have additional challenges, which are more strategical in nature. In particular, how should we approach the search when the signal model is vague or the model space is very large and the phenomenology is diverse? We still have not fully addressed the multiple testing problem for the LHC, but perhaps methods like False Discovery Rate have a role to play in the global analysis of the LHC data. If we are fortunate enough to discover new physics at the LHC, we will begin the process of interpreting what we saw. Perhaps we will see something expected and the process will be fairly straightforward; however, we must be prepared for something more unexpected. Ideally, the experiments will publish their results in terms of a full likelihood scan of a model-neutral and model-relevant parameter space. The technical challenge of reporting a full likelihood map has been addressed by the RooFit framework, the remaining challenge is choosing how to represent the data. In some cases the field has already converged on an appropriate set of parameters to measure for a given signature, but we do not have an adequate solution in the case of something more unexpected. A recent proposal is to use on-shell effective theories as a concise summary of LHC data, which could provide a general purpose solution for publishing model-neutral and model-relevant results. While there remain many open questions to address, the PhyStat conference series has been very effective in preparing our field for the statistical challenges of the LHC.

References

[1] K. Cranmer, *proceedings of PhyStat05, Oxford, England, United Kingdom, (2005)*

[2] S. Frixione and B. R. Webber, JHEP **0206** (2002) 029 [arXiv:hep-ph/0204244].

[3] Andreas Schalicke and Frank Krauss. *JHEP*, 07:018, (2005).

[4] LEP Higgs Working Group. *Phys. Lett.*, B565:61–75, (2003).

[5] Nancy Reid. proceedings of PhyStat2003, (2003).

[6] W. A. Rolke and A. M. Lopez. *Nucl. Instrum. Meth.*, A458:745–758, (2001).

[7] W. A. Rolke, A. M. Lopez, and J. Conrad. *Nucl. Instrum. Meth.*, A551:493–503, (2005).

[8] K. Cranmer. *proceedings of PhyStat2003* (2003) [physics/0310108].

[9] G. Punzi. *proceedings of PhyStat2005* (2005) [physics/0511202].

[10] R.D. Cousins and V.L. Highland. *Nucl. Instrum. Meth.*, A320:331–335, (1992).

[11] J. Linnemann. *proceedings of PhyStat2003* [physics/0312059], (2003).

[12] R.D. Cousins and J. Tucker, [arXiv:physics/0702156] (2007)

[13] F. James and M. Roos. *Comput.Phys.Commun.*, 10:343–367, (1975).

[14] S. Asai et al. *Eur. Phys. J.*, C3252:19–54, (2004).

[15] L. Demortier p-values *these proceedings*, PhyStat2007 (2007).

[16] W. Quayle *talk presented at*, PhyStat2007 (2007).

[17] K. Cranmer, "LEPStats4LHC", software available at *phystat.org* repository. [packages/0703002].

[18] W. A. Rolke and A. M. Lopez, [arXiv:physics/0606006].

[19] K. Cranmer, *In the Proceedings of PHYSTAT2003: Statistical Problems in Particle Physics, Astrophysics, and Cosmology, WEJT002 (2003),* [arXiv:physics/0310110].

[20] K. Cranmer. *Acta Phys. Polon.*, B34:6049–6068, (2003).

[21] K. Cranmer and R. S. Bowman. *Comp. Phys. Commun.*, 167(3):165–176, (2005).

[22] J. M. Link *et al.* [FOCUS Collaboration], Phys. Lett. B **624** (2005) 166 [arXiv:hep-ex/0507103].

[23] S. Whiteson and D. Whiteson, [arXiv:hep-ex/0607012].

[24] R.M. Neal, *Bayesian Learning of Neural Networks*. Springer-Verlag, New York, (1996)

[25] V. M. Abazov *et al.* [D0 Collaboration], Nature **429** (2004) 638 [arXiv:hep-ex/0406031].

[26] V. M. Abazov *et al.* [D0 Collaboration], Phys. Rev. Lett. **98** (2007) 181802 [arXiv:hep-ex/0612052].

[27] W. Wagner, [hep-ex/0610074]; The CDF Collaboration, public conference note 8185, April 2006; M. Bühler, Diplomarbeit Universität Karlsruhe, FERMILAB-MASTERS-2006-02, (2006).

[28] K. Cranmer and T. Plehn, Eur. Phys. J. C **51**, 415 (2007). [arXiv:hep-ph/0605268].

[29] C.G. Lester, "Part X:" , *Les Houches 'Physics at TeV Colliders 2003' Beyond the Standard Model Working Group: Summary report,* [arXiv:hep-ph/0402295].

[30] Y. Benjamini and Y. Hochberg *J. R. Stat. Soc.*-B 57:289-300 (1995)

[31] C. J. Miller *et al.*, Astro.Jour., 122.6,3492-3505 (2001) [arXiv:astro-ph/0107034].

[32] V. M. Abazov *et al.* [D0 Collaboration], Phys. Rev. Lett. **87**, 231801 (2001)

[33] I. Hinchliffe, F. E. Paige, M. D. Shapiro, J. Soderqvist, and W. Yao. *Phys. Rev.*, D55, (1997).

[34] N. Arkani-Hamed, G. L. Kane, J. Thaler and L. T. Wang, JHEP **0608** (2006) 070

[35] E. A. Baltz and P. Gondolo, JHEP **0410** (2004) 052 [arXiv:hep-ph/0407039].

[36] B. C. Allanach and C. G. Lester, Phys. Rev. D **73** (2006) 015013 [arXiv:hep-ph/0507283].

[37] R. R. de Austri, R. Trotta and L. Roszkowski, JHEP **0605** (2006) 002 [arXiv:hep-ph/0602028].

[38] M. Rauch, R. Lafaye, T. Plehn and D. Zerwas, arXiv:0710.2822 [hep-ph].

[39] B. C. Allanach, K. Cranmer, C. G. Lester and A. M. Weber, JHEP **08**, 023 (2007). Likelihood maps published at *http://users.hepforge.org/∼allanach/benchmarks/kismet.html*

[40] W. de Boer, M. Huber, C. Sander and D. I. Kazakov, Phys. Lett. B **515** (2001) 283.

[41] J. R. Ellis, K. A. Olive, Y. Santoso and V. C. Spanos, Phys. Rev. D **69** (2004) 095004

[42] S. Amari, *Differential-geometrical methods in statistics*, Springer-Verlag, Berlin, 1985

[43] W. Verkerke, *http://roofit.sourceforge.net/* and *these proceedings*

[44] N. Arkani-Hamed, P. Schuster, N. Toro, J. Thaler, L. T. Wang, B. Knuteson and S. Mrenna, arXiv:hep-ph/0703088.

Experiences from Tevatron Searches

Wade Fisher
Fermilab, Illinois, United States

Abstract
In preparation for the possibility of new physics at the Large Hadron Collider at CERN, experiences gained at the Fermilab Tevatron collider experiments can be a useful guide for potential problems. This paper presents a review of recent applied statistical techniques and the problems for which they were required.

1 Introduction

The two Fermilab Tevatron collider experiments, CDF and DØ , both probe a broad range of elementary particle physics. Of particular interest are searches for new particles or evidence for new physics. It is anticipated that the new energy frontier at the Large Hadron Collider (LHC) at CERN will create opportunities for observing new physics beyond the electroweak scale. Experiences from searches for new physics at the Tevatron should provide insight into the problems that LHC experiments will likely face. This paper presents a brief review of the problems addressed by several modern Tevatron searches.

Many searches at the Tevatron result in a relatively unambiguous statistical significance. Two excellent examples of this are the direct observation of the $\Xi_b^{\pm}$ baryon at DØ [1] and the observation of orbitally excited B_s^{**} mesons [2]. Figure 1 shows the reconstructed mass of the $\Xi_b^{\pm}$ baryon and Fig. 2 shows the mass difference in candidates for orbitally excited B_s^{**} decays. Although both are exciting examples of discovery, these searches do not represent challenges or ambiguity in the estimation of search significance. Most of the problems for Tevatron searches arise from the convergence of small signal rates, large background rates, and large relative uncertainties on nuisance parameters inherent to the search. It is these cases which will be the focus of this paper.

2 Systematic Uncertainties

In many searches for small signals, a significant limiting factor is the relative size and nature of systematic uncertainties on the measurement of background processes. In the large-statistics limit of a search, a precise knowledge of systematic uncertainties is required to reliably determine search significance. In

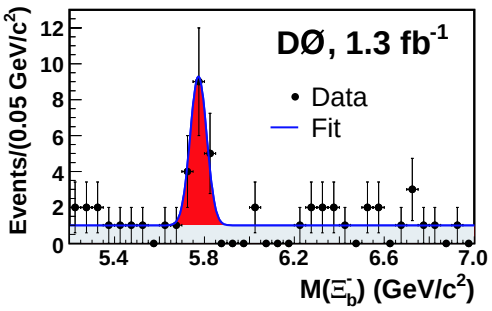

Fig. 1: The background-subtracted invariant mass of $\Xi_b^{\pm}$ baryons at the DØ experiment.

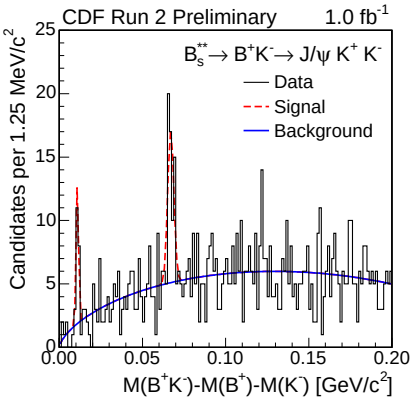

Fig. 2: The mass difference in candidates for orbitally excited B_s^{**} decays at the CDF experiment.

searches for small signals at the Tevatron, the uncertainty on the background rate is often over an order of magnitude larger than the signal rate itself, and is thus a dominant factor in signal sensitivity. There are two broad classes of such uncertainties:

- Type I: Uncertainties related to gross normalizations or rates of acceptance. These are generally constrainable by control samples and the relative size scales according to the statistics of the control samples.
- Type II: Uncertainties related to the understanding of the features of data measurements and experimental resolutions. These often manifest as uncertainties on the shapes of differential distributions involved in event selection procedures.

Type I uncertainties are generally assumed to arise from parent distributions which are Gaussian in nature. Errors in the estimation of background process rates occur when these uncertainties are actually non-Gaussian or asymmetric. Estimates of Type I uncertainties must also be sensitive to regions of truncation which occur, for example, when efficiencies reach either 0% or 100%. Type II uncertainties can present a significant challenge when their impact to an event selection is difficult to properly measure. The size of these uncertainties are often inflated to partially accommodate this difficulty, which in turn degrades search sensitivity. A worrying scenario arises when Type II uncertainties are not propagated through high-dimension multivariate analyses, thus incorrectly overestimating signal significance. A careful understanding of systematic uncertainties is considered a prerequisite to performing a detailed statistical analysis for a search.

The finite statistics of simulated samples used to predict the rates of different classes of events represents a particular challenge in searches for small signals. In many cases, the uncertainty on the shape of a differential distribution or multivariate analysis discriminant is dominated by the statistical uncertainty on the prediction. A common solution is to use a smoothing algorithm to make an estimate of the true parent distribution. There are two smoothing techniques used frequently at the Tevatron: the 353QH algorithm [3] which is implemented in the ROOT software package [4] and Gaussian kernel estimation [5]. A more complete comparison of the two algorithms is presented in [5]. As an alternative to smoothing, the true shape of the parent distribution of a statistics-limited sample can be estimated from the shape of the same variable at a less restricted point in the selection process. For example, an analysis that selects two quark-jets and requires that both jets be tagged by a b-quark identification algorithm could model the shape of the double b-tagged distribution by the shape of the less restrictive single b-tagged selection. After a proper normalization, the remaining biases of the b-tagging algorithm are often smaller and better understood than the uncertainty on the statistics-limited shape of the double-tagged sample. This method is often used in conjunction with a smoothing algorithm, but is sensitive to the nature of the intermediate selection used.

3 The Tevatron Higgs Search

The search for a standard model (SM) Higgs boson at the Tevatron provides a good example of the implementation of several statistical techniques used in searches for small signals. The statistical analysis begins with an assumption of two hypotheses which are to be tested:

- Null hypothesis (**H0**): A compound hypothesis with an associated set of nuisance parameters each with its own uncertainty. This hypothesis can be considered to be the background-only hypothesis.
- Test hypothesis (**H1**): The test hypothesis is the same as **H0**, but a signal for new physics is added. Thus, **H1** adopts the necessary signal model parameters and possibly additional nuisance parameters.

In this example, **H0** describes the SM background expectation for the results of a Higgs search and is the sum of several contributing physical processes. The nuisance parameters are the luminosity

normalization, acceptance, background cross sections, etc. The test hypothesis adds the SM Higgs signal and is parameterized by Higgs mass, production cross section, and decay branching fractions. Both **H0** and **H1** are subdivided according to final states with unique signatures. These final states are orthogonal search channels defined to maximize acceptance and to isolate regions with high signal significance. At the Tevatron, two different statistical analysis treatments are utilized: a Bayesian integration [6] and the semi-Frequentist CL_S method [7]. More detailed references for the Tevatron SM Higgs searches can be found here [8, 9].

3.1 The Bayesian Treatment

The Bayesian approach utilized in the Tevatron Higgs search begins by assuming a Poisson-distributed probability distribution for the observed numbers of events selected. The posterior probability density function (PDF) for a set of signal and background model parameters θ_R is given by:

$$p(\theta_R|\vec{x}) = \frac{\int L(\vec{x}|\theta_R,\theta_S)\,\pi(\theta_R,\theta_S)\,d\theta_S}{\int\int_{-\infty}^{\theta_{R_{cut}}} L(\vec{x}|\theta_R,\theta_S)\,\pi(\theta_R,\theta_S)\,d\theta_S\,d\theta_R} \tag{1}$$

where $\pi(\theta_R,\theta_S)$ is the prior probability density for θ_R and the set of nuisance parameters θ_S. The likelihood $L(\vec{x}|\theta_R,\theta_S)$ is the joint probability density over all analysis channels and bins of the final variables and the observed data $\vec{x}$. The parameter $\theta_{R_{cut}}$ is chosen to ensure unitarity, and for an appropriately chosen prior can generally be infinity. The most common choices for prior probability density for the signal is the Heaviside unit step function, and the prior for all nuisance parameters is taken as Gaussian. The limit on the rate of signal events is then determined by integrating the posterior density function to the desired fraction of the total integral, β:

$$\beta = \int_{-\infty}^{\theta_{R_\beta}} p(\theta_R|x)\,d\theta_R \tag{2}$$

which defines θ_{R_β} as the limit on the model parameter at a Bayesian confidence level of β.

3.2 The CL_S Treatment

The semi-Frequentist CL_S approach also assumes Poisson-distributed sources of events and begins by constructing a joint likelihood ratio test statistic:

$$Q = -2\,Log\,\frac{L(\vec{x}|\theta_{R1},\hat{\theta_S})}{L(\vec{x}|\theta_{R0},\hat{\hat{\theta}_S})} \tag{3}$$

where the profile likelihood $L(\vec{x}|\theta_{R1},\hat{\theta_S}))$ is the Poisson likelihood for the physics parameters of **H1** and the set of nuisance parameters which maximize the likelihood for **H1** ($\hat{\theta_S}$). Likewise, $L(\vec{x}|\theta_{R0},\hat{\hat{\theta}_S}))$ represents a maximization for the physics parameters of **H0**. This test statistic is used to describe the outcomes of multiple repetitions of the experiment for which pseudo-experiment trials are drawn from Poisson-distributed outcomes of the **H0** and **H1** hypotheses. The uncertainties on the N nuisance parameters for each source of events are assumed to have a Gaussian PDF. For each pseudo-experiment, the mean value for the expected number of signal and background events is randomly drawn from an N-dimensional Gaussian distribution, described by the N nuisance parameters and their uncertainties.

The PDFs for the test statistics of **H0** and **H1** are then used to evaluate confidence intervals of the following definition:

$$CL_S = 1 - \frac{CL_{S+B}}{CL_B} \tag{4}$$

where the confidence levels CL_{S+B} and CL_B are defined by integrating the corresponding PDFs from the observed test statistic to infinity. An exclusion of a signal model parameter (parameterized as F) at a confidence level of α is achieved when the model parameter satisfies $\alpha \leq 1 - \frac{CL_{S+B}(F)}{CL_B(F)}$.

3.3 Drawbacks of Methodology

Practically speaking, statistical treatments with discrete numbers of events lead to imperfect coverage for any method. This unavoidable consequence demands a conservative treatment to ensure well-understood coverage. The Bayesian treatment suffers from both larger issues with coverage but also the choice of prior. While priors are somewhat unpopular in the field of particle physics due to potential biases, a careful choice of prior can generally contruct a test with the desired properties. It is generally accepted that the CL_S method is a less rigorous means of communicating exclusions of signal hypotheses. However, the properties of the CL_S test appear to be robust for this purpose. The formulation of the test has the agreeable feature of giving a conservative response for exclusion in insensitive signal regions. The approach chosen at the Tevatron is to maintain both treatments in as many cases as possible. This serves both as a cross check of results, but can also lead to insight in the fundamental behavior of small signal searches.

4 The Tevatron Search for Single Top Quark Production

In 2007, the DØ collaboration announced it had observed first evidence for the electroweak production of single top quarks [10]. At the same time, the CDF collaboration announced that with analyses of similar sensitivity it did not find the same results [11, 12, 13]. This scenario may occur between the LHC experiments and the treatment at the Tevatron therefore has pedagogical value. Adopting the nomenclature from the previous section, the descriptions of the **H0** and **H1** hypotheses are identical for the Tevatron search for single top quarks aside from the parameterization of the signal.

4.1 The DØ Single Top Search

The DØ measurement was constructed using a Bayesian calculation similar that defined in Eqn. 1. Three semi-independent analyses observed an excess of events consistent with single top quark production near the expected SM rate of 2.9 pb. The measurement of the observed cross section was based on a binned likelihood derived from the analysis discriminants:

$$y = a\sigma + \sum_{s=1}^{N_{bkgd}} b_s \tag{5}$$

$$P(D|y) = P(D|\sigma, a, b) = \prod_{i=1}^{N_{bins}} P(D_i|y_i) \tag{6}$$

where the index s runs over the number of background sources in **H0**. The Bayesian posterior PDF was calculated as a function of cross section, assuming a Heaviside unit step function prior density probability. As demonstrated in Fig. 3, the measured signal cross section is determined from the peak of the posterior PDF and the uncertainty is taken from the width near the peak. Figure 4 shows the actual measurement derived from one of the three analysis techniques.

The estimates of search sensitivity and the significance of the observed result were described using p-values derived from pseudo-experiments. For each pseudo-experiment, the values of the nuisance parameters were sampled from Gaussian PDFs. The p-values were reported with the following definitions:

– Expected p-value: The fraction of **H0** pseudo-experiments measuring at least the SM single top quark cross section of 2.9 pb.

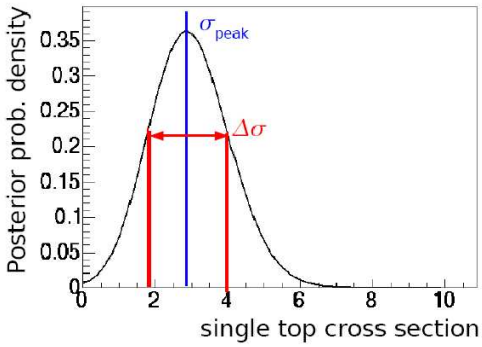

Fig. 3: An example of the Bayesian posterior density distribution for pseudo-experiments with a single top quark signal with the SM rate of 2.9 pb for the DØ experiment.

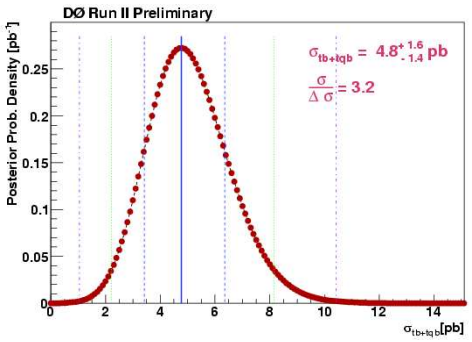

Fig. 4: The Bayesian posterior density distribution for pseudo-experiments derived from the observed data distribution for the DØ experiment.

Table 1: DZero ST Data

	Expected p-value	Observed p-value	SM p-value	Observed Cross Section
Analysis 1	0.019 (2.1σ)	0.00035 (3.4σ)	0.11 (1.2σ)	4.9 pb
Analysis 2	0.037 (1.8σ)	0.0021 (2.9σ)	0.21 (0.8σ)	4.6 pb
Analysis 3	0.097 (1.3σ)	0.0089 (2.4σ)	0.175 (0.9σ)	5.0 pb

Table 2: DZ ST Cov Matrix

Analysis 1	1.0	0.57	0.51
Analysis 3	-	1.0	0.45
Analysis 2	-	-	1.0
Weight	0.401	0.452	0.146

– Observed p-value: The fraction of **H0** pseudo-experiments measuring at least the observed cross section.

– SM p-value: The fraction of **H1** pseudo-experiments measuring at least the observed cross section.

The results of these measurements along with the observed cross section for each analysis are given in Table 1. The results seem to indicate an upward fluctuation in the data rate, but are demonstrably more compatible with the **H1** hypothesis.

Given three highly correlated analyses, the DØ researchers employed the Best Linear Unbiased Estimate (BLUE) technique [14] to combine the measurements and determine a more sensitive estimate of the observed signal significance. The BLUE technique essentially describes a linear function of weighted measurements. The weights are determined by inverting the correlation matrix for the system of measurements. The covariance matrix obtained via pseudo-experiments and the corresponding linear weights for the DØ analyses are shown in Table 2. Using this linear combination, a new set of pseudo-experiments was generated to determine the observed signal significance. Via this technique, the original best value of 4.9 ± 1.4 pb was refined to 4.8 ± 1.3 pb, reported as 3.5 standard deviations, and more recently as 4.7 ± 1.3 pb [15].

4.2 The CDF Single Top Search

The CDF collaboration performed a similar search for single top quark production in an equal-sized data sample. Using a similar analysis approach, CDF researchers applied three different multivariate analyses

Table 3: CDF ST Data

	Exp. p-value	Obs. p-value	Exp. CL_S	Obs. CL_S	Exp. Limit	Obs. Limit
Analysis 1	0.005 (2.6σ)	0.525 (-)	-	-	2.9 pb	2.6 pb
Analysis 2	0.025 (2.0σ)	0.585 (-)	0.05	0.039	2.9 pb	2.7 pb
Analysis 3	0.006 (2.5σ)	0.01 (2.3σ)	-	-	-	-

Table 4: CDF ST Cov Matrix

Analysis 1	1.0	0.59	0.70
Analysis 3	-	1.0	0.65
Analysis 2	-	-	1.0

to estimate significance. However, the expected and observed significance of each analysis was reported in a slightly different manner. As in the case of the DØ statistical analyses, the values of the nuisance parameters used in pseudo-experiments are drawn from Gaussian PDFs:

- Analysis 1:
  - Expected p-value: The fraction of **H0** pseudo-experiments measuring at least the SM single top quark cross section of 2.9 pb.
  - Observed p-value: The fraction of **H0** pseudo-experiments measuring at least the observed cross section.
- Analysis 2: The same p-values reported by Analysis 1 were given, and the CL_S confidence level (Eqn. 4) was also reported.
- Analysis 3: The same p-values reported by Analysis 1 were given, and the Bayesian calculation described in the Sec. 4.1 was used to measure an observed cross section.

The analysis results were mixed, with two analyses excluding single top quark production above 2.6 pb and 2.7 pb respectively, while the third analysis observed a 2.3σ effect at 2.7 ± 1.2 pb. All three analyses had a similar expected sensitivity and used the same data, reconstruction, and Monte Carlo simulation. To understand the compatibility of the three measurements, the CDF researchers also utilized the BLUE technique. The covariance matrix for the CDF analyses is given in Table 4. With the linearly-combined estimator, a combined measurement was evaluated and a χ^2 value for each pseudo-experiment was determined. An estimate of analysis compatibility was determined by measuring the fraction of pseudo-experiments whose χ^2 value exceed the value observed data. This fraction was found to be 0.65%.

4.3 Comparison of Results

Both Tevatron experiments devised multiple analyses, all with similar search sensitivities in the same size data sample. The agreement of results amongst the three DØ analyses is not unexpected considering the design of the analyses, and the additional search sensitivity gained via the linear combination of results is indeed small. The conflicting results from the CDF analyses is perhaps more interesting. It is conceivable that the results reflect internal biases within the analyses. It is also possible that the analyzers were unlucky, to use the particle physics jargon, and the data reflected a downward fluctuation of stochastic processes. At the time of this publication, the DØ researchers have updated two of their analysis techniques and found similar results [15]. At this time, there is no public update of the CDF single top quark search.

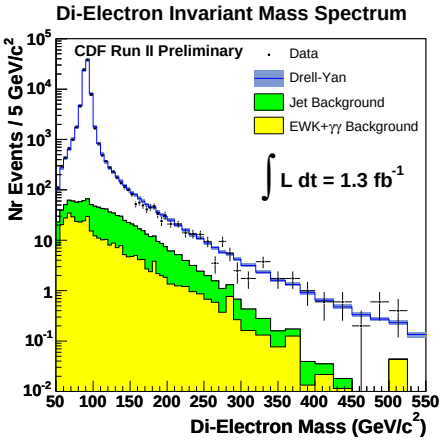

Fig. 5: The di-electron invariant mass spectrum from the CDF experiment.

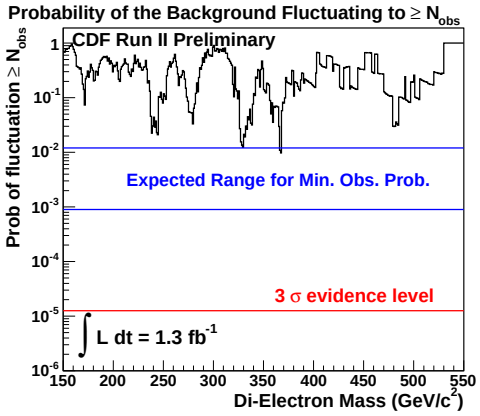

Fig. 6: The Poisson probability for data as a function of di-electron mass.

5 Model Independent Searches

The previous two examples of the standard model Higgs boson search and the single top quark search exemplified directed searches for new physics at the Tevatron. Each uses very modern approaches to analysis and statistical interpretation, but are inherently linked to assumptions which impact the interpretation of the results. As an alternative, a second class of searches removes all model parameters for new physics from the analysis design and allows them to be introduced after a statistical interpretation of the results has been performed. There are two general categories for such searches: bump hunts and broad spectrum searches.

5.1 Bump Hunts

Given a well-understood differential distribution, researchers can search for deviations from nominal predictions of the shape and rate of that distribution. As an example, we will consider the model-independent search for a high-mass resonance in the di-electron mass spectrum at the CDF experiment. The di-electron mass spectrum is well-studied at the Tevatron as the global energy-scale calibration procedure for both experiments relies upon accurate knowledge of Drell-Yan Z boson production and high-resolution detection of electrons. Figure 5 contains an example of the CDF di-electron mass spectrum corresponding to 1.3 fb^{-1} of data [16].

The analysis of the di-electron mass spectrum proceeded by using a variable-sized sliding window of the form $W(M_{ee}) = 4.8 + 0.044 \times M_{ee}$ with a step size of 1 GeV$/c^2$. For each step, the significance of any excess above the SM prediction was evaluated via a Frequentist p-value. Assuming Poisson-distributed background rates and Gaussian uncertainties on the rates of backgrounds, the p-value is defined as the fraction of background-only pseudo-experiments which equal or exceed the number of events observed in the window. The results of this test are shown in Figure 6 as a function of the di-electron mass. The authors define an expected range for minimum observation probability of 5%-0.27% region as the expected range to find the minimum Frequentist p-value over the tested mass range [16]. The minimum p-value for the spectrum is 9.7×10^{-3} with the sliding window centered at $M_{ee} = 367$ GeV$/c^2$.

Such a search is certainly hampered the lack of a signal model, but presents a broad application for theoretical model interpretation. As such, this technique is a valuable approach to studying the gross features of a data sample. The analyzers went on to interpret the search using both a set of Z' models and Randall-Sundrum graviton models, each resulting in a more sensitive probe to the specified new physics model than the model-independent search.

5.2 Broad Spectrum Searches

The range of physics processes at the Tevatron and LHC contains a rich phenomenology of search possibilities. Despite being an exciting opportunity for the observation of new physics, there are several challenges which must be faced. First, limited human and computing resources force the prioritization of search design and implementation. Second, a comparison and correlation of search results in a broad range of final states is often made opaque by differing search techniques and interpretations. The Tevatron has seen the development of a few broad spectrum search techniques which attempt to address these problems, amongst others. As an example, we present the two most modern versions implemented at the CDF experiment: Vista and Sleuth.

The Vista program [17] searches for large cross-section physics in final states with high-p_T (high transverse momentum) physics objects. The basic algorithm proceeds as follows:

1. Select high-p_T ($p_T > 17\text{GeV}/c$) electrons, photons, muons, tau-leptons, hadronic jets, and neutrinos (manifested as missing transverse energy) which pass the detector's physics triggers.
2. Events are passed through an offline filter to isolate interesting final states.
3. Standard model background simulations are generated and the detector response is simulated.
4. Orthogonal subsets of events are formed and kinematic distributions are populated.

The current implementation of Vista identifies a total of 344 final states and 16486 kinematic distributions. The program forms a global χ^2 for a comparison of the simulation to data and minimizes it over 44 total nuisance parameters, 26 of which are externally constrained. Following this minimization, the total numbers of events are compared for each final state and each kinematic distribution is evaluated using the Kolmogorov-Smirnov statistic. An example of a discrepant distribution identified by the Vista program is shown in Fig. 7.

The Sleuth program [18] is a quasi-model-independent tool used to search for new physics in high-p_T final states. To be sensitive to electroweak-scale new physics, the program analyses the tails of the summed transverse momentum ($\sum p_T$). The program interfaces with the Vista program by adopting its orthogonal set of final states and its comprehensive correction model. The statistical test for each final state in the Sleuth program is as follows:

1. Identify D regions in D data points defined by the semi-infinite integral of the $\sum p_T$ kinematic distribution.
2. Define the interestingness ($\mathcal{P}_N$) of a region containing N data points as the Poisson probability the SM prediction would fluctuate up to or above N.
3. The most interesting region ($\mathcal{R}$) is found by minimizing $\mathcal{P}_N$ for the final state.
4. Pseudo-experiments of the SM $\sum p_T$ distribution are used to generate a population of $\mathcal{P}_N$ associated with the value $\mathcal{R}$ for each final state.
5. The fraction of regions more interesting than $\mathcal{R}_{obs}$ quantifies each final state.

Considering all final states, the program determines the most interesting region $\mathcal{R}_{max}$. Using this region, the statistic $\tilde{\mathcal{P}}$ is defined as the fraction of pseudo-experiments that would generate a region in any final state more interesting than $\mathcal{R}_{max}$, including a proper accounting for the number of final states considered. Assuming that the simulation and correction model are accurate, the distribution of $\tilde{\mathcal{P}}$ in all final states should be uniform in the range $0 \rightarrow 1$. In the presence of an unmodeled source of physics, the value of $\tilde{\mathcal{P}}$ should be small.

In 927 pb^{-1} of data at the CDF experiment, the Sleuth program found $\tilde{\mathcal{P}} = 0.46$. The Sleuth program would interpret this as no indication for new physics in the distributions it probes, while the threshold for the pursuit of potential discovery is chosen by the authors at $\tilde{\mathcal{P}} < 0.001$. The most interesting final state in this data sample as identified by the Sleuth program is the two b-quark final state, as

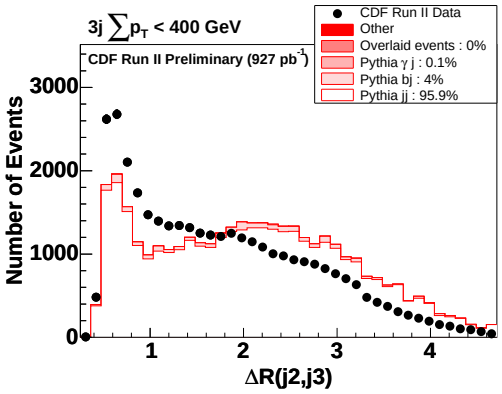

Fig. 7: One of the most discrepant kinematic distributions identified by the Vista program at the CDF experiment.

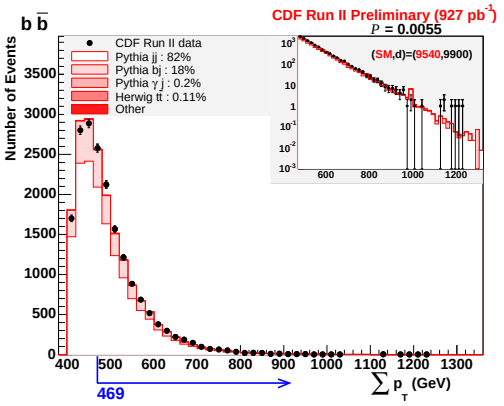

Fig. 8: The most discrepant summed p_T distribution found by the Sleuth program at the CDF experiment.

shown in Fig. 8. It should be noted that this tool is not intended to be a bypass for directed searches, but rather an effective means for evaluating an experiment's data in a systematic manner.

6 Summary

The statistical techniques employed in recent Tevatron searches encompass a broad range of interpretation and utility. The transition to the new energy frontier at the CERN LHC will indeed be exciting and is eagerly anticipated by many. It is expected that the LHC experiments will face similar challenges in searches for new physics as those seen at the Tevatron. The experiences from Tevatron searches will hopefully be both useful and instructive for probing the data at the LHC.

References

[1] The DØ collaboration, Phys. Rev. Lett. **99**, 1052001 (2007)

[2] The CDF Collaboration, Phys. Rev. D **64**, 072002 (2001)

[3] J.W. Tukey, "Exploratory Data Analysis", Vol. I, Chapter 7, Addison-Wesley, Reading, Mass., 1970

[4] Rene Brun and Fons Rademakers, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A **389** (1997) 81-86. See also `http://root.cern.ch/`.

[5] K.S. Cranmer, Comput. Phys. Commun., **136** (2001) 198, arXiv:0011057 [hep-ex]

[6] Joel Heinrich *et. al.*, "Interval estimation in the presence of nuisance parameters. 1. Bayesian approach.", CDF Note 1771, (2004). `http://www\-cdf.fnal.gov/publications/cdf7117\_bayesianlimit.pdf`

[7] A.L. Read, "Optimal statistical analysis of search results based on the likelihood ratio and its application to the search for the MSM Higgs boson at sqrt(s) =161 and 172", DELPHI collaboration note 97-158 PHYS 737 (1997); T. Junk, Nucl. Intrum. Meth. A **434**, 435 (1999);

[8] The CDF collaboration, "Combined Upper Limit on Standard Model Higgs Boson Production at CDF for Summer 2007", CDF Public Note 8941. `cdfhiggs`

[9] The DØ collaboration, "Combined Upper Limits on Standard Model Higgs Boson Production from the DØ Experiment in 0.9-1.7 fb^{-1}", DØ Note 5487-CONF. `http://www-d0.fnal.gov/Run2Physics/WWW/results/prelim/HIGGS/H41/`

[10] The DØ collaboration, Phys. Rev. Lett. **98**, 181802 (2007)

[11] The CDF collaboration, "Search for Electroweak Single-Top-Quark Production using Neural Networks with 955 pb^{-1} of CDF II data", CDF Public Note 8677. http://www-cdf.fnal.gov/physics/new/top/confNotes/cdf8677_public_NN1FB.pdf

[12] The CDF collaboration, "Search for Single Top Quark Production in 955 pb^{-1} using the Matrix Element Technique", CDF Public Note 8588. http://www-cdf.fnal.gov/physics/new/top/confNotes/cdf8588_me_singletop_1fb.pdf

[13] The CDF collaboration, "Multivariate Likelihood Search for Single-Top-Quark Production with 1 fb^{-1}", CDF Public Note 8585. http://www-cdf.fnal.gov/physics/new/top/confNotes/cdf8585_lf_singletop_1fb.pdf

[14] L. Lyons, D. Gibaut and P. Clifford, NIM A **270**, 110 (1988). L. Lyons, A. Martin and D. Saxon, Phys. Rev. D **41**, 3 (1990). A. Valassi, NIM A **500**, 391 (2003)

[15] The DØ Collaboration, "Updated Combination Results from Three Single Top Quark Cross Section Measurements using the BLUE Method", DØ Note 5396-CONF. http://www-d0.fnal.gov/Run2Physics/WWW/results/prelim/TOP/T49/T49.pdf

[16] The CDF Collaboration, arXiv:0707.2524 [hep-ex].

[17] The CDF Collaboration, arXiv:0710.2372 [hep-ex].

[18] The CDF Collaboration, arXiv:0710.2378 [hep-ex]. The DØ Collaboration, Phys. Rev. Lett. **86**, 3712 (2001) [arXiv:hep-ex/0011071].

ATLAS and CMS Statistics Wish-List

Eilam Gross
Weizmann Institute, Rehovot, Israel

Abstract
A wish-list of statistics related issues, which were regarded by ATLAS and CMS as requiring a deeper understanding and perhaps the response of a professional statistician, is given.

1 Introduction

The first Phystat meeting was a workshop at CERN on Confidence Limits followed by a similar workshop at Fermilab. Fred James who organized the meeting with Louis Lyons presented then his personal wish list titled: "What I would like to see" (see Figure 1). Fred wishes that physicists learn the vocabulary of statistics, all searches use Feldman and Cousins unified method [1] to derive confidence intervals and Bayesian methods are used only in policy decisions. When accepting upon myself to collect a wish list from Atlas and CMS my only experience was with LEP statistics [2] and the so called CL_s method for deriving limits which was used at LEP. In the two months of preparation of this lecture I had a steep learning curve during which I partially fulfilled the first item in Fred's wish list but found myself in mild debate with his other two points. However when enquiring around I discovered that most physicists are mainly concerned with old fashioned systematics issues and the majority have only a vague idea of the meaning of the term "Nuisance parameters" and the meaning and difference between the Feldman and Cousins vs the profile likelihood methods etc. It became clear to me that my mission in this lecture is not only to communicate to the statisticians our unsolved difficulties but also to make sure that when Atlas and CMS publish a combined limit or discovery significance not only the few statisticians amongst the Physicists (which I propose to call "Phystatisticians") will understand but also the majority of the HEP experimentalists. Therefore I decided to expand the contents of my talk to include also a pedestrian guide to LHC statistics. This guide which also provides the separate title to these proceedings [3]also provides the reader with the statistics background required to understand the wish list. The statisticians and phystatisticians are exempted from reading it.

2 A Wish List of ATLAS and CMS

The wish-list is made of statistics related issues which were regarded by ATLAS and CMS as requiring a deeper understanding and perhaps a professional statistician response.

2.1 Modeling of the underlying process

The raison d'etre of our meeting here are 10^9 Protons that will collide with 10^9 protons per second in the 27 km long LHC tunnel. The Proton is made of partons which are Quarks and Gluons. The underlying process is a collision between two partons. To understand the process we need to know the Parton Distribution Function $f_i(x; Q^2, \alpha_s(Q^2))$ of parton i in a Proton. x is the fraction of momentum carried by the parton and Q^2 is the energy scale squared. To obtain the Parton Distribution Functions a global fit analysis is performed[4], mainly the CTEQ and MRST, [5]. However some strange phenomena occur with these fits. For example, for most of the individual experiments from which the data is taken the χ^2/dof is around 1. However, in the global fit, the CTEQ group set $\Delta\chi^2 \sim 100$ in order to get reasonable errors [6]. Stump [7] argues: "What we have are estimates on the uncertainties, not the true ones. The increase of χ^2 if the estimators are biased or wrong might be bigger than 1. We find that alternate pdfs that would be unacceptable differ in χ^2 by an amount of order 100". This is a very vague statement. Robert Thorne, the "S" in MRST concludes with a wish: "It would be nice to have a more

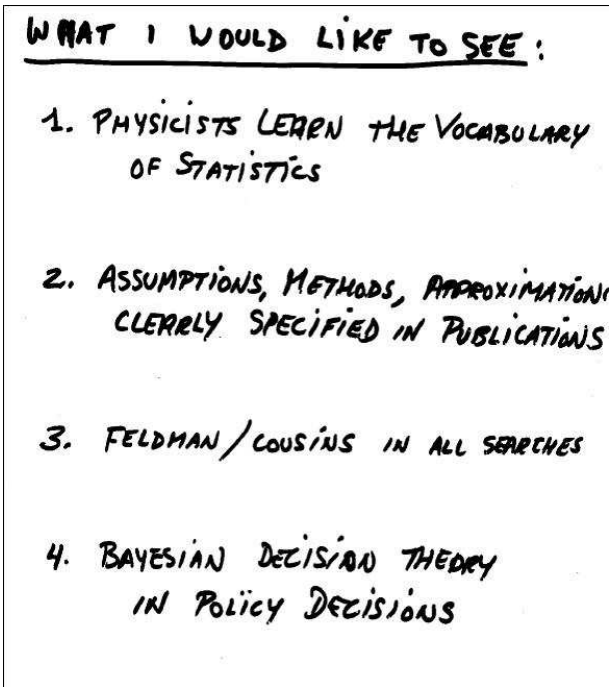

Fig. 1: Fred James personal wish-list presented in the first Phystat meeting, year 2000.

systematic way of accounting for this, e.g. a modified definition of goodness of fit to account for non-Gaussian nature of errors, a quantitative way of accounting for theoretical errors, etc..." [8]. Since Throne is elaborating the issue in this conference [6] I decided not to dwell anymore with it, though it is in the heart of proton-proton collisions and must be sorted out in order to reduce the systematics involved with all processes involved.

2.2 Why 5σ?

The null hypothesis is usually taken to be the background only hypothesis. The alternate hypothesis is the signal+background hypothesis. When analyzing results in HEP (High Energy Physics) it became a habit either to reject the signal hypothesis at the 95% Confidence Level or announce an observation at the 3σ level or a discovery at the 5σ level. Statistically speaking a 5σ discovery corresponds to a fluctuation of a Gaussian distributed background expectation at the level of $5.4 \cdot 10^{-7}$ (here we adopt the 2-sided interpretation). At that level one is probing tails of distributions. In order to make such statements one needs to understand the data and the detector response to that level. Nobody really knows where this habit of 5σ was born. A back of the envelope calculation reveals a possible explanation which has to do with the "look elsewhere effect". Suppose when searching for a new phenomena (Higgs boson...) one is combining 100 search channels each with a discriminating variable distributed within 100 resolution bins. The false discovery rate of 5σ will be $10^4 \cdot 2.7 \cdot 10^{-7} = 0.27\%$. This degrades the discovery sensitivity to 3σ. More examples and insights into this problem can be found in [9, 10, 11].

So is there a way to clarify how many σs are needed for discovery? Is there a problem in seeking for an effect at at tail of a pdf? How can we take a fluctuation into account?

2.3 Look Elsewhere Effect in Time

An ongoing discussion in the LHC collaborations is the need and possibility to perform a blind analysis. Even if from the scientific integrity point of view the pros are clear, with each collaboration having over 2000 physicists it is hard to believe such a habit can be adopted. Moreover, it will take years till the

detectors are understood, and understanding the detector necessitates looking at the data as it comes. However, that raises another issue, the issue of sequential analysis. Should a statistical stopping rule be established? Is it possible that by adopting a stopping rule we would achieve a discovery with less luminosity than needed with a blind analysis? Understanding stopping rules is a complicated issue. It is probably more relevant in medical experiments where one would like to minimize the damage that might be referred to as patients trying new drugs. No doubt the HEP community must adopt very strict rules for looking at the data and publishing results to minimize human bias. The HEP Physicist might be bothered by another related issue, which might be refereed to as a "look elsewhere in time" effect: How much the p-value is increased as a result of the fact that we have already looked at the data a few times before and got no satisfactory significance?

2.4 Estimating Systematics

There are two issues related to systematics. Classifying and estimating them and implementing them in the analysis. Implementation of the systematics in the statistical hypothesis tests is discussed at length in [3]. Knowing the type of error one is dealing with is very important and make its estimation clearer. Sinervo [12] classified the sytematics into three types: Class I: Statistics-like uncertainties that are reduced with increasing statistics. Example: Calibration constants for a detector whose precision of (auxiliary) measurement is statistics limited. Class II: Systematic uncertainties that arise from one's limited knowledge of some data features and cannot be constrained by auxiliary measurements. One has to make some assumptions. Example: Background uncertainties due to fakes, isolation criteria in QCD events, shape uncertainties. These uncertainties do not normally scale down with increasing statistics. Class III: The "Bayesian" kind. The theoretically motivated ones, uncertainties in the model, Parton Distribution Functions, Hadronization Models. *The most accurate way to communicate the systematic error is to separate one type from another and quote them separately*. Some bad habits should cease. For example adding all sorts of systematics in quadrature and quote only the final result.

Another unfortunate habit in estimating systematics is when Physicists do not differentiate between cross-checks and identifying the sources of the systematic uncertainty. For example we shift cuts around and measure the effect on the observable. Very often the observed variation is dominated simply by the statistical uncertainty in the measurement [13].

2.5 Reference Priors in Demand of a Code

Analytical derivation of reference priors might be technically complicated. Bernardo [14] proposes an algorithm (pseudo code) to obtain a numerical approximation to the reference prior in the simple case of a one parameter model. The pseudo code should work for any number of parameters (of interest and nuisance) provided you make non informative priors for ALL! If the code could be extended to multiple parameters, including some with informative priors, it would be more useful for the HEP community. Another complication is that the order of the parameters matter. This should be further investigated and clarified. The wish is to have a generalized routine (REAL CODE) to numerically calculate reference priors for parameters $\{\theta\}$ given the Likelihood $L(\{\theta\})$ as an input.

2.6 Subsequent Inference

Often the background distribution is fitted with a polynomial, $\sum_{i=0}^{n} a_i x^i$ with the degree n determined with a stepwise test. However, the fitted coefficients a_i were obtained as if we know a priori the degree of the polynomial. How does one take the prior test into account? Perhaps the degree was wrong to start with?

Table 1: Hypothesis Test Methods. The columns indicate if the method obeys the Likelihood Principle (LP), if it has a coverage and if it uses priors.

		LP	Coverage	Priors	Comments
1	F&C	No	Yes	No	Pioneering in HEP
2	F&H&C[2]	No	No	Yes	
3	Profile Likelihood (PL)	Yes	Asymptotic	No	`Best Value for Money`
4	PL F&C Construction	No	Satisfactory	No	
5	PL Full Construction	No	Yes	No	Cumbersome
6	Bayesian	Yes	No	Yes	Choose priors with care
7	CL_s with C&H	Yes	Partial	Yes	For upper limits only

2.7 Multivariate Analysis

The number of Physicists objecting to MultiVariate (MV) analyses (like ANN, Decision Trees) is getting smaller as the average year of birth of the active physicists go up. Evaluating the systematics with MV analyzes is very unclear. Many physicists have the habit of changing the input parameter by what they believe is a standard deviation, do it one at a time or randomly with all of them together. There must be a better way to do it. Can the community come up with good figures of merit for the robustness optimization of a MV analysis (and not only for the significance)?

Note that the articles by Linnemann ('A pitfall in estimating systematic errors') and by Reid ('Some aspects of experiment design') in these proceedings also deal with this issue of problems with changing one variable at a time.

2.8 Telling Between Multi Hypotheses

Is the scalar particle we have just discovered a Standard Model one, a CP-odd SUSY one or a CP-even SUSY one [15]? Here are three hypotheses regarding the nature of the Higgs Boson. The Neyman Pearson lemma tell us the best test statistic to tell between two simple hypotheses. In case of more than one equivalent alternate hypotheses, what is the best test statistics to use besides testing them one against the other? Is there anyway to do it without a Bayesian assumption that all hypotheses have an a priori degree of belief?

2.9 Hypothesis Test

Testing a preferred hypothesis includes the estimation and incorporation of systematic errors. The result is then interpreted in terms of exclusion, measurement or discovery. Hypothesis testing is a science by itself. The LEP collaboration has chosen the CL_s method integrating out the systematics using the C&H method. This was one possibility out of many. In the years since LEP we have grown up to understand many more methods. The frequently used ones in HEP are introduced in [3] and compared in Table 1. Since each method has its pros and cons the ATLAS and CMS combined statistics forum has expressed its wish that the analyses be interpreted in a few of them. The frequentist based Profile Likelihood and a Bayesian method are highly recommended. CL_s will certainly be practised for exclusion (the power of a habit...). It is also recommended to try one of the Neyman construction methods. If all methods agree a trust in the result will be established, however, if one method gives completely different inference from others, this should be further investigated.

3 Conclusion and a Personal Wish

Some of the statistical issues raised in this paper are quite picky. The level of the discussions on hypothesis tests is quite advanced, which indicates that the HEP community's understanding of statistics has matured since the days of LEP (thanks in large measure to the Phystat meetings). It is therefore my

personal wish that when the real data analysis phase arrives (one hopes soon) every physicist will make the effort to become a Phystatistician to some degree, so he or she understands what is a p-value, what is Profile Likelihood, a Confidence Limit, a Confidence Interval etc.

4 Acknowledgements

This work would have never been possible without the patience and amazing support of Bob Cousins, Kyle Cranmer and Glen Cowan. They helped me to become as close as possible to a Phystatistician in two months.

I would also like to thank Alex Read, Bill Quayle, Luc Demortier and my student Ofer Vitells for always being there to answer my questions and requests. Last but not least, applause to Louis Lyons for keeping the Phystat meetings going. These meetings are the ultimate source of statistics for the High Energy Physicists. Long may these meetings live! Toda Louis. I am obliged to the Benoziyo center for High Energy Physics for their support of this work. This work was also supported by the Israeli Science Foundation(ISF), by the Minerva Gesellschaft and by the Federal Ministry of Education, Science, Research and Technology (BMBF) within the framework of the German-Israeli Project Cooperation in Future-Oriented Topics(DIP).

References

[1] Gary J. Feldman and Robert D. Cousins. A unified approach to the classical statistical analysis of small signals. Phys. Rev., D57:3873-3889, 1998.

[2] Eilam Gross and Amit Klier, Higgs Statistics for Pedestrians, hep-ex/0211058.

[3] Eilam Gross, LHC Statistics for pedestrians, these proceedings.

[4] A nice introduction can be found in John Pumplin's talk: Light Front Physics and Parton Distributions (Talk at LC2006 Minneapolis May 17, 2006); `http://www.pa.msu.edu/people/pumplin/talks/lc.pdf`

[5] A CTEQ and MRST Parton Distribution Functions generator can be found in `http://durpdg.dur.ac.uk/hepdata/pdf3.html`

[6] Robert Throne in these proceedings.

[7] Daniel Stump, Uncertainties of Parton Distribution Functions, Phystat2003.

[8] Private Communication with Robert Throne.

[9] See talk by Kyle Cranmer in these proceedings.

[10] Gary Feldman, Concluding Remarks: Phystat 2005

[11] 95% CL limits and 5σ discoveries in CMS TDR Appendix A. See also talk by Alexey Drozdetskiy, this conference.

[12] P. Sinervo, Definition and treatment of systematic uncertainties in high energy physics and astrophysics PhyStat2003.

[13] See an excellent discussion on estimating sysyematics in the 2000 SLUO Lectures on Statistics and Numerical Methods in HEP, lecture number 5, Roger Barlow, `http://www-group.slac.stanford.edu/sluo/Lectures/stat_lecture_files/sluolec5.pdf`

[14] Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. J. American Statistical Association 84, 200-207; Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. Bayesian Statistics 4 (J. M. Bernardo, J. O. Berger, D. V. Lindley and A. F. M. Smith, eds). Oxford: Oxford University Press, 61-77 (with discussion). See also Luc Demortier, Bayesian Reference Analysis for Particle Physics, Phystat05.

[15] D. Rainwater, D. Zeppenfeld and K. Hagiwara, Phys . Rev. D 59, 014037 (1999); T. Plehn, D. Rainwater and D. Zeppenfeld, Phys. Lett. B 454, 297 (1999); Phys. Rev. D 61, 093005 (2000).

Some Statistical Issues in the LHCb Experiment

Yuehong Xie
The University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK

Abstract

This paper describes statistical issues that are of particular importance and interest to the LHCb experiment in probing new physics beyond the Standard Model through study of CP violation and rare phenomena in B decays. A wish list for statistical methods and tools that will help LHCb to exploit its full physics potential is given at the end.

1 Introduction

The LHCb experiment is a dedicated B physics experiment at the Large Hadron Collider (LHC). Its physics aim is to study CP violation and rare phenomena in B decays with very high precision in order to test the Standard Model (SM) in the quark flavour sector and to look for physics beyond the SM. Different from the ATLAS and CMS experiments, which will explore the high energy frontier to search for new physics particles directly produced in proton-proton collisions at the LHC, the LHCb experiment will pursue precision measurements to understand the quantum effects of possible virtual new particles appearing in loop diagrams. LHCb will need to put enormous efforts to understand how to deal with background, control systematic uncertainties and incorporate theoretical errors. Improving signal significance is also very important, especially for measurements of very rare decay processes.

This paper discusses the major issues in LHCb physics analysis that require special statistical treatments. These are illustrated using examples of analysis. In addition a list of statistical methods and tools that LHCb wishes to develop, improve or understand better is given.

2 The physics of the LHCb experiment

In the SM, quark-flavour mixing and CP violation are fully described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix with four independent parameters. The task of flavour physics is to determine these parameters and more importantly to check the validity of the CKM mechanism. LHCb will take two routes to achieve this task. LHCb will make many measurements to over-constrain the CKM matrix. These will provide stringent tests of the SM. Any inconsistency will mean that some new source of flavour mixing and CP violation must exist. LHCb will also study Flavour-Changing-Neutral-Current (FCNC) loop decays. The FCNC decays are forbidden at tree level in the SM, therefore new physics may have significant effects in these processes. Comparing asymmetries or rates in these decays with their SM expectations will be a sensitive test of the SM. Any established discrepancy will indicate new FCNC couplings beyond CKM mixing.

3 A statistical view of the LHCb experiment

Statistics plays an important role in quantifying the level of consistency/inconsistency between physics models and experimental data. In the language of statistics, LHCb will perform hypothesis testing. The null hypothesis is that "the SM is valid at the energy scale relevant to B meson decays". No alternative hypothesis is explicitly given, but rejecting the null hypothesis implies new physics is needed to explain the data. What LHCb needs to do for a hypothesis test of the SM is

– Identify a test statistic, i.e. an observable, in flavour physics which has high power to separate the SM and potential new physics models;

- Measure the test statistic from data;
- Evaluate the tail probability of the null hypothesis, that is, the p-value;
- If the p-value is judged too small, reject the null hypothesis and look for a possible alternative;
- Otherwise go for another observable and repeat the test.

4 Application of statistics in LHCb physics analysis

Statistical methods and concepts are used in almost every aspect of B physics experiments, ranging from pattern recognition to averaging measurements. Since many of these issues have been widely discussed in the B physics community, this paper focuses on those aspects of LHCb data analysis which can potentially benefit a lot from improvement of statistical methods.

4.1 B flavour tagging

For most CP measurements with neutral B decays it is necessary to know the flavour of the B meson at production. This can be inferred from the information carried by the following tagging categories:

- Same side tagger: charge of the particle accompanying the signal B at production;
- Opposite side tagger: charge of muon, electron or kaon from the decay of the opposite side B hadron;
- Vertex charge tagger: the weighted sum of the charges of all particles found to be compatible with being from the opposite side B decay.

The tagging result is a decision made on a statistical basis combining all available taggers. The figures of merit of tagging is the effective tagging power $\epsilon(1-2\omega)^2$, where ϵ is the tagging efficiency and ω is the mistag probability. The current estimates using simulated data are $\epsilon \sim 50-60\%$, $\omega \sim 30-35\%$ and $\epsilon(1-2\omega)^2 \sim 4-10\%$. Since the statistical errors of the CP asymmetries in neutral B decays decrease linearly with the square root of the tagging power, it is crucial to maximize the tagging power using appropriate statistical methods.

In LHCb neural network methods are employed to get event-by-event mistag probability of each tagger, the performance of which depends on the way the neural networks are constructed and trained. We expect some room for improvement here. Combining different tagging categories is non-trivial when these are correlated. For example, if the opposite side tagger and the vertex charge tagger use the same particle, correct handling of this correlation requires splitting the data sample into sub-samples depending on whether there is a particle used by the opposite side tagger and the vertex charge tagger or not. LHCb is developing new techniques to optimize the procedure of combining taggers. A possibility for tagging improvement is to investigate using better methods to assign particles to vertices. The tagging algorithm needs determine if a particle originates from a primary vertex or from a tagging B vertex. It may lead to a wrong tagging decision if for example a charged lepton from a primary vertex is mistakenly regarded as being from the tagging B hadron or loss of tagging efficiency if a charged lepton from the tagging B hadron is mistakenly treated as being from a primary vertex. In both cases, the effective tagging power is compromised. We have already investigated various methods to minimize this loss of tagging power, but room for further improvement is still possible.

4.2 Separating signal and background events

Separating signal and background events is a demanding task in LHCb for two reasons: after trigger the ratio of inclusive $b\bar{b}$ background events to signal events is at the order of one million to one in a typical decay channel and can be even larger for a very rare decay channel such as $B_s \to \mu^+\mu^-$; in each $b\bar{b}$ event there are not only the two B hadron decays but also about 50 tracks from proton-proton interactions.

The following information can be used for signal and background separation

- Particle identification;
- Kinematic information such as particle momenta and invariant masses of particle combinations;
- Geometrical information such as the secondary vertex χ^2 and event topology.

There are typically 10-20 variables to look at in an analysis, each alone with limited discrimination power. Therefore, a cut-based analysis method is usually not optimum in terms of statistical precision. Multivariate analysis methods can be more powerful for signal and background separation but this involves more complexity to understand the systematic issues. In a real analysis one needs to find a trade-off between better statistical precision and smaller systematic uncertainty.

A multivariate analysis entails constructing a best test statistic from multiple input variables for a hypothesis test. In principle the Neyman-Pearson lemma tells us the likelihood ratio is the best choice (for simple hypotheses). A straightforward application is to represent the probability density functions (PDFs) of signal and background by multi-dimensional histograms which can be obtained from Monte Carlo simulation. The likelihood ratio then can be computed as the ratio of the two PDFs. However, this procedure becomes impractical when the dimensions of the PDFs are too large.

There are alternative methods which construct estimators to approach the likelihood ratio under certain conditions. Examples include decorrelated likelihood classifier, linear estimators such as Fisher's discriminants, nonlinear estimator such as neural networks and Boosted decision trees. The Toolkit for MultiVariate data Analysis (TMVA) [1], an integrated part of the Root framework, hosts a variety of these multivariate algorithms and provides many techniques that are useful in LHCb data analysis.

The TMVA package has been applied in a simulation study of the decay channel $B_s \to e^{\pm}\mu^{\mp}$ [3], which is forbidden in the SM and therefore requires high selection efficiency and low background level. In this study, variables showing very clear separation between signal and background are directly cut on first. TMVA is used to deal with the less powerful variables. This effectively reduces the complexity in understanding systematics. In this particular case no non-linear correlations between these are expected. A sample of simulated data is used to train several classifiers and an independent data sample is used to evaluate their performances. The results are shown in Fig. 1. Just as the Neyman-Pearson lemma implies, the decorrelated likelihood method, denoted as LikelihoodD in Fig. 1, gives the highest signal efficiency for the same number of background events. This is not necessarily the case in other more complicated analyses with non-linear correlations between the input variables. There are indications that the current TMVA mechanism to monitor over-training of classifiers using independent samples for training and testing may not be sufficient. A way to control over-training in the training phase is desirable.

4.3 Setting confidence limits in case of a small signal

As in all HEP experiments, we need to quote confidence intervals/limits for experimental measurements. This issue is especially important when working on very rare decays with small signals and large background. Here we use the analysis of $B_s \to \mu^+\mu^-$ [3] as an example to illustrate the statistical procedure which LHCb adopts for determination of the experimental sensitivity in very rare decay channels. We know that $B_s \to \mu^+\mu^-$ is highly suppressed in the SM and its branching ratio is expected to be around 3.4×10^{-9}. This can be greatly enhanced in new physics models. We use the average exclusion limit as a measure of the experimental sensitivity, which is defined as the average upper limit that would be obtained from an ensemble of experiments with the expected background and no true signal [4]. We evaluate the average exclusion limit by generating toy experiments with only background events and using the "N-counting" method described below to set an upper limit for each toy experiment. The "N-counting" method includes the following steps:

- Construct geometrical, muon-ID and $\mu^+\mu^-$ invariant mass likelihood ratios between signal and

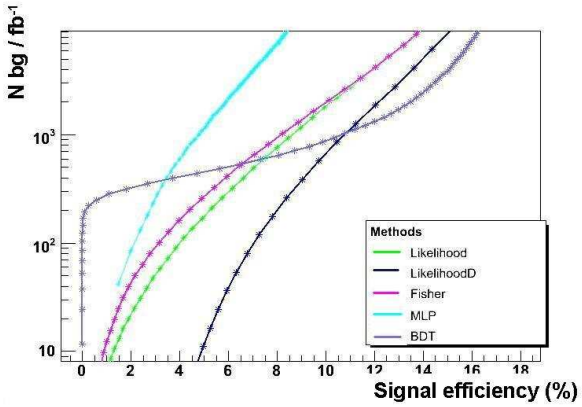

Fig. 1: Number of retained background events per fb^{-1} as a function of signal efficiency (%) for various multivariate methods.

background hypotheses for each event, where the decorrelated likelihood method is used for the geometrical likelihood ratio;

- Divide the 3-dimensional space of the three likelihood ratios into a number of bins and count the number of events in each bin, denoted as d_i;
- Estimate the number of expected background events b_i and signal events s_i for each examined branching ratio in each bin;
- Construct a total likelihood ratio between the signal+background and background-only hypotheses for the whole experiment

$$X = \prod_i \frac{P(d_i, < d_i >= s_i + b_i)}{P(d_i, < d_i >= b_i)}, \tag{1}$$

where $P(x, < x >)$ denotes the Poisson probability of a variable x with the average value $< x >$;
- Evaluate the p-value of the signal+background hypothesis: $probability(X < X_{observed}; S + B)$ and that of the background-only hypothesis: $probability(X > X_{observed}; B)$;
- Form a statistic called CL_s [5] from the ratio

$$CL_s = \frac{p\text{ -value of signal plus background hypothesis}}{1 - (p\text{ -value of background hypothesis})}; \tag{2}$$

- Make a statistical statement: if $CL_s(BR) < \alpha$, then the branching ratio BR is excluded at $1 - \alpha$ confidence level.

The comparison of the N-counting method and a simple counting method is shown in Fig. 2. It can be clearly seen that the N-counting method requires less data to reach the same average exclusion limit at 10% confidence level. While the CL_s test statistic has some advantages over the likelihood ratio X and the CL_s limit is easy to compute, the way the confidence level is set is known to be conservative [6]. The normal procedure requires the p-value of the signal+background hypothesis, not the CL_s, to be smaller than α in order to exclude the signal+background hypothesis at $1 - \alpha$ confidence level.

It should be noted that the significance of a measured result is given by the p-value of the background-only hypothesis, which should not be confused with the p-value of the signal+background hypothesis or the CL_s value.

4.4 Analysis tools for data modelling and fitting

The maximum likelihood fit method is generally used in B physics experiments. This is largely facilitated by the data modelling and fitting package RooFit [7]. The RooFit package has been widely used in LHCb

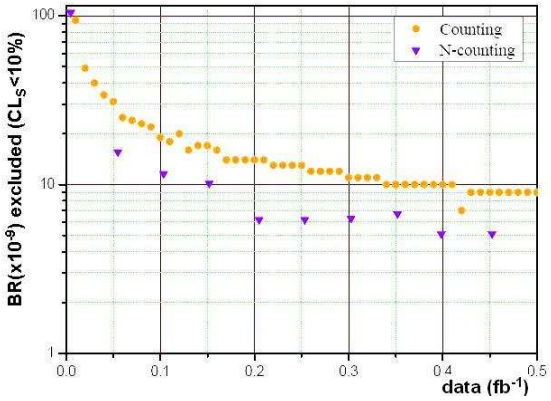

Fig. 2: $B_s \to \mu^+\mu^-$ average exclusion limit (10^{-9}) at 90% confidence level as a function of the integrated luminosity (fb^{-1}) for the N-counting method and the simple counting method as a comparison.

sensitivity studies. The experience shows that LHCb can better benefit from this package if:

- The event generation for complicated PDFs can be made faster;
- We learn how to make fits converge that employ non-factorizable multi-dimensional PDFs that have no analytical normalization and can only be numerically integrated.

4.5 Controlling systematic errors

Systematic errors arise from incorrect modelling of the detector and/or background effects. Delicate statistical methods are needed to acquire knowledge of these effects from real data and to model them. An example is the efficiency as a function of the proper decay time t and phase space position Ω, denoted as $\varepsilon(t,\Omega)$. Correct modelling of $\varepsilon(t,\Omega)$ is very important for time-dependent and/or an angular analysis. Here we discuss a technique [8] to absorb the effect of $\varepsilon(t,\Omega)$ into a normalization factor. Generally the PDF describing a signal decay has the form

$$p(t,\Omega;A) = \frac{\sum_j h_j(A) f_j(t) g_j(\Omega)\varepsilon(t,\Omega)}{\sum_j h_j(A) \int_t \int_\Omega f_j(t) g_j(\Omega)\varepsilon(t,\Omega) dt d\Omega}, \tag{3}$$

where $h_j(A)$, $f_j(t)$ and $g_j(\Omega)$ are functions that depend only on the physical parameters A, decay time t or phase space position Ω respectively. The likelihood of all signal events is

$$L = \prod_i l_i = \prod_i p(t_i,\Omega_i;A). \tag{4}$$

Varying parameters A to maximize L requires evaluating

$$\frac{d\ln l_i}{dA} = \frac{d}{dA}\ln\frac{\sum_j h_j(A) f_j(t) g_j(\Omega)\varepsilon(t,\Omega)}{\sum_j h_j(A) \int_t \int_\Omega f_j(t) g_j(\Omega)\varepsilon(t,\Omega) dt d\Omega} \equiv \frac{d}{dA}\ln\frac{\sum_j h_j(A) f_j(t) g_j(\Omega)}{\sum_j h_j(A)\Phi_j}, \tag{5}$$

where we have defined $\Phi_j \equiv \int_t \int_\Omega f_j(t) g_j(\Omega)\varepsilon(t,\Omega) dt d\Omega$. These factors Φ_j are independent of the physical parameters A and therefore can be obtained from Monte Carlo simulation before fitting. Note the acceptance function $\varepsilon(t,\Omega)$ in the numerator of Eq. 3 drops out in the last step of Eq. 5 because it only contributes a constant to the log-likelihood function. There is no need to know the explicit form of $\varepsilon(t,\Omega)$ in the maximum likelihood fitting. In addition to the general problem of a lack of goodness-of-fit in a unbinned likelihood fit, a lack of knowledge of $\varepsilon(t,\Omega)$ makes it difficult to check the fit quality by

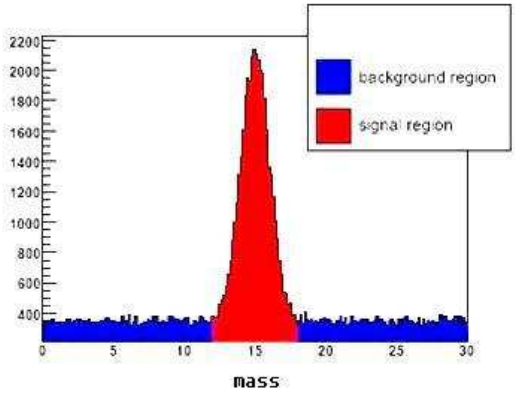

Fig. 3: Schematic view of signal and sideband regions defined with invariant mass.

comparing measured distributions and fitted projections. A solution must be found to ensure the fit result is reliable. This is still under investigation.

When background events are taken into account the total PDF becomes

$$p_{tot} = f \cdot p_{sig} + (1 - f) \cdot p_{bkg}, \tag{6}$$

where f is the fraction of signal events in the data sample and p_{sig}, p_{bkg} denotes signal and background PDF respectively. The total PDF no longer has the form of Eq. (3). Therefore the normalization trick of Eq. (5) cannot be employed.

One solution to this problem is to use a pseudo-log-likelihood method [9], which avoids the use of any background PDF. Instead of maximizing the usual likelihood defined using the total PDF in Eq. (6), one maximizes the pseudo-log-likelihood defined using only the signal PDF

$$\ln L_{pseudo} = \sum_{i=1}^{N_{sig}} \ln(p_{sig}(t_i, \Omega_i; A)) - \frac{N_{sb}}{N_b} \sum_{j=1}^{N_b} \ln(p_{sig}(t_j, \Omega_j; A)), \tag{7}$$

where $N_{sig/b}$ is the total number of events in the signal/sideband region and N_{sb} is the number of expected background events in the signal region. The signal and sideband regions can be defined in terms of invariant mass as shown in Fig. 3. While the minimization of the negative pseudo-log-likelihood leads to a unbiased estimate of the physical parameters A, the errors returned by Minuit at the end of the minimization are generally under-estimated. There are efforts underway to derive a formalism to give the correct estimates of parameter errors [10]. Useful discussions on the topic of background-subtraction during this workshop can be found in [11]. Every effort needs to be made to model the detector and background effect correctly in order to minimize systematic errors. In addition, a proper statistical procedure should be established and strictly followed when estimating systematic errors so that the arbitariness in assigning systematic uncertainties can be minimized.

4.6 CKM fit and other global fits

An area in B physics where statistical analysis plays a major role is the CKM fit. The goal of the CKM fit is to test if different measurements of the sides and angles of the Unitarity Triangle are consistent or not. Currently there are two groups working on this using the B factory measurements: the UTFit group [12] which uses a Bayesian method and the CKMFitter group [13] which employs a frequentist approach. While the two methods are very different from each other, the basic conclusions made by the two groups are the same: no inconsistency between measurements is found so far. Once LHCb starts to produce precision measurements, a more stringent test of the CKM mechanism can be done. At that time it might

be necessary to consider improving the statistical treatments of these existing tools to make best use of the high precision measurements at LHCb. For example, we may want to know how better to incorporate theoretical uncertainties and how to tell if an inconsistency is due to new physics or to under-estimated systematic uncertainties or under-estimated theoretical uncertainties.

LHCb will also measure many rare decay channels. Individually they test the SM and probe new physics in one way or another. We may want to perform a global fit in the SM in order to achieve better sensitivity to new physics. This is not an easy job as the SM relations between the measured quantities in the rare decay channels are not precisely known and the SM predictions of these quantities are usually subject to sizable uncertainties. A lot of analysis efforts are first needed to understand the SM and make better predictions. In terms of statistical methods, some thinking is required to understand how to construct a test statistic using the measurements in rare decay channels and their SM predictions and taking into account the uncertainties of these predictions and the correlation between the predicted errors due to common theoretical sources.

5 LHCb's statistical wish list

Having discussed the aspects of the LHCb experiment that will need a careful statistical analysis, we can give a specific list of topics that LHCb wishes to develop or to improve:

- A well supported tool for data modelling and fitting that can handle general multi-dimensional problems;
- A multivariate analysis tool that is capable of dealing with multiple discriminating variables with non-linear correlations and has a reliable mechanism to monitor and control over-training of classifiers;
- Better understanding of how to treat systematic and theoretical uncertainties;
- New statistical methods to improve flavour tagging;
- Better understanding of how to set confidence limits in case of insignificant signals;
- Recommendation on statistical procedures in data analysis.

References

[1] A. Hocker *et al.*, CERN-OPEN-2007-007.

[2] W. Bonivento and N. Serra, CERN-LHCb-2007-028.

[3] D. Martinez *et al.*, CERN-LHCb-2007-033.

[4] G. J. Feldman and R.D. Cousins, Phys. Rev. D57, 3873 (1998).

[5] T. Junk, CERN-EP/99-041.

[6] W.-M. Yao *et al.*, *Review of Particle Physics*, Journal of Physics G 33, 1 (2006).

[7] W. Verkerkear and D. Kirkby, Xiv:physics/0306116.

[8] S. T. Jampens's thesis, available at https://oraweb.slac.stanford.edu/pls/slacquery /BABAR_DOCUMENTS.DetailedIndex?P_BP_ID=3629.

[9] B. Aubert, Phys. Rev. D71, 032005 (2005).

[10] Private communication with Joe Boudreau at CDF.

[11] J. Linnemann and A. J. Smith, these proceedings.

[12] M. Bona *et al.*, UTfit Collaboration, J. High Energy Phys. 0610, 081 (2006), updated results available at http://www.utfit.org/.

[13] J. Charles *et al.*, CKMfitter Group, Eur. Phys. J. C41, 1 (2005), and updated results available at http://www.slac.stanford.edu/xorg/ckmfitter/.

ALICE Statistical Wish-list

Iouri Belikov for the ALICE collaboration
CERN, Geneva, Switzerland

Abstract
A few statistical problems faced by the event reconstruction in ALICE experiment at CERN are discussed in this paper. We outline several ad-hoc extensions of traditional Kalman-filter track finding which seem to increase the quality of tracks reconstructed in high multiplicity events anticipated for Pb–Pb collisions at LHC. These extensions, however, need a stricter formulation and justification from the theoretical side. The particle identification in ALICE is done by combining the information from different detecting systems using a Bayesian method. Having many clear advantages, this approach introduces into the analysis additional complications which are also discussed here.

1 Introduction

A Large Ion Collider Experiment (ALICE) [1] at CERN is a general-purpose heavy-ion experiment designed to study the physics of strongly interacting matter and the Quark-Gluon Plasma in nucleus-nucleus collisions at the LHC. In addition to heavy systems, the ALICE Collaboration will study collisions of lower-mass ions, in order to vary the energy density, and protons (both pp and pA), which primarily provide reference data for the nucleus–nucleus collisions. The pp data will also allow for a number of genuine pp physics studies.

The detector consists of a central part (see Fig. 1), which, event-by-event, measures hadrons, electrons and photons, and of a forward spectrometer to measure muons. The central part, which covers polar angles from 45° to 135° over the full azimuth, is embedded in the large L3 solenoidal magnet. It consists of an Inner Tracking System (ITS) of high-resolution silicon detectors; a cylindrical Time-Projection Chamber (TPC); three particle identification arrays, a Time-Of-Flight (TOF) detector, a Transition-Radiation Detector (TRD) and a single-arm ring imaging Cherenkov detector (HMPID) and a single-arm electromagnetic calorimeter (PHOS). The forward muon spectrometer (covering polar angles $180^\circ - \theta = 2^\circ - 9^\circ$) consists of a complex arrangement of absorbers, a large dipole magnet, and fourteen planes of tracking and triggering chambers. Several smaller detectors for global event characterization and triggering are located at forward angles.

The detector is optimized for charged-particle density $\mathrm{d}N_{\mathrm{ch}}/\mathrm{d}y = 4000$ and its performance is checked in detailed simulations up to $\mathrm{d}N_{\mathrm{ch}}/\mathrm{d}y = 8000$. The track reconstruction efficiency in the acceptance of the TPC is about 80% down to transverse momentum of $p_{\mathrm{t}} \sim 0.2$ GeV/c and about 90% for tracks with $p_{\mathrm{t}} > 1$ GeV/c. It is limited only by the particle decays and small dead zones between the TPC sectors. Typical momentum resolution obtained with the magnetic field of 0.5 T is $\sim 1\%$ at $p_{\mathrm{t}} \sim 1$ GeV/c and $\sim 4\%$ at $p_{\mathrm{t}} \sim 100$ GeV/c. The secondary vertices can be reconstructed with the precision better then 100 μm.

The detector has excellent particle identification (PID) capabilities. From $p \sim 0.1$ GeV/c to a few GeV/c the charged particles are identified by combining the PID information provided by ITS, TPC, TRD, TOF and HMPID. Statistically, the charged particles can be identified up to a few tens GeV/c using the relativistic rise of $\mathrm{d}E/\mathrm{d}x$ in the TPC. Electrons above 1 GeV/c are identified by the TRD, and muons are registered by the muon spectrometer.

To achieve the benchmarks described above, the ALICE reconstruction has to cope with a few statistical problems. We will discuss some of them in this paper (the details can be found in Chapter 5 of the ALICE Physics Performance Report [2]).

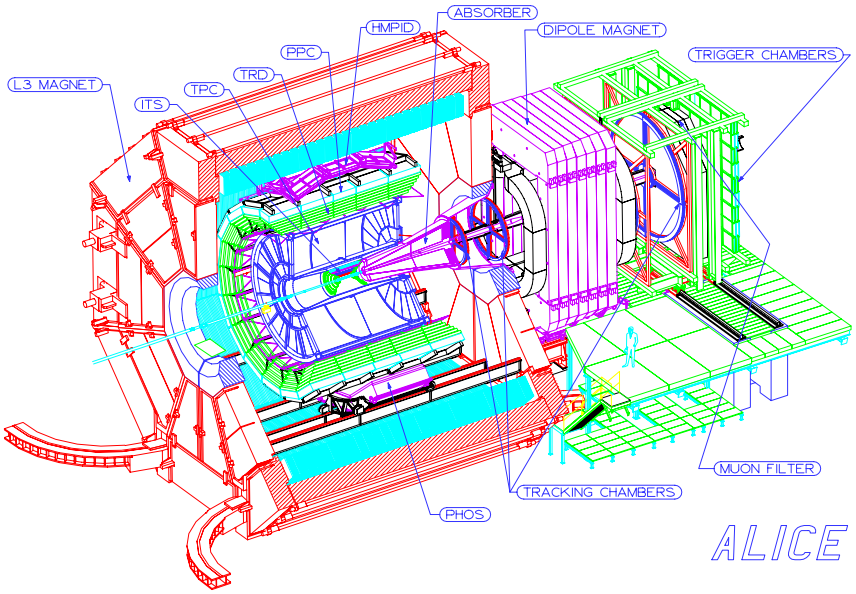

Fig. 1: Schematic layout of the ALICE detector.

2 Statistical problems with track finding in ITS

The track reconstruction in ALICE starts in the TPC, and then the tracks have to be prolonged in the ITS. This is difficult, because the distance between the inner wall of the TPC and the outer layer of the ITS is rather large and the track density inside the ITS is so high that there are always many ITS clusters found within the prolongation 'window' defined, mainly, by the multiple scattering in the material. The same often happens between the ITS layers as well (see Fig. 2). All this leads to a non-negligible probability of assigning to tracks many wrong clusters, if we use just the criterion of minimal χ^2 at each layer. Therefore we have to find the ways to improve the classical Kalman filter track-finding procedure [3].

For each event, we do two reconstruction passes over the set of clusters in the ITS: first, with a 'primary vertex constraint' (see below) and then without the constraint. In the both cases, we try to assign to a track, one by one, all the hits within the predicted window that have a χ^2 below a given limit, and not only the one with minimal χ^2. This way, for each track from TPC, we build a whole 'tree' of all possible prolongations in ITS. To speed up building the tree, the branches are sorted after each layer according to χ^2 and only a restricted number of acceptable branches are propagated further. Finally, we choose the most probable track candidate (i.e. the path along the tree) taking into account the quality of the whole path (sum of χ^2s at the layers, total number of assigned clusters and a few other criteria).

Because most of tracks are expected to be primary, the first reconstruction pass is done applying an ad-hoc 'primary vertex constraint' (see Fig. 3). Since the primary vertex in ALICE can be reconstructed in advance sufficiently well, the idea is to use this additional information during the track finding. When going over the clusters within the 'window', we take into account not only the positions of clusters and the track intersection point with the layer, but also the direction towards the primary vertex. Technically, this is done by extending the vector of measurement m

$$m^T = \{y, z\} \rightarrow \{y, z, \sin(\phi), \tan(\lambda)\},$$

where $\{y, z\}$ are the coordinates of the cluster position and the angles $\{\phi, \lambda\}$ define the direction to the primary vertex and are calculated using the current value of the track curvature. The elements of the covariance matrix of the extended measurement vector that correspond to the two angles are evaluated considering the material which this track would cross on its way to the primary vertex. The subsequent evaluation of the χ^2 and update of the track parameters become thus 4-dimensional problem.

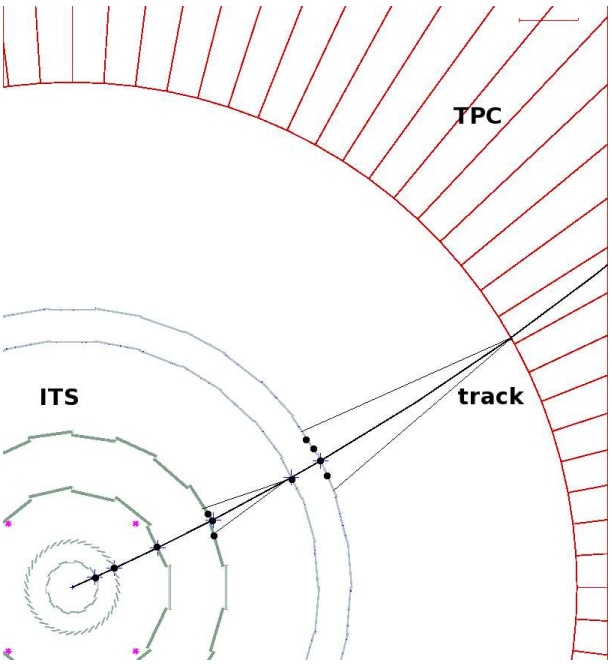

Fig. 2: The problem of track finding in ITS in high multiplicity events: Several clusters are found within the prolongation 'window' from one layer to another.

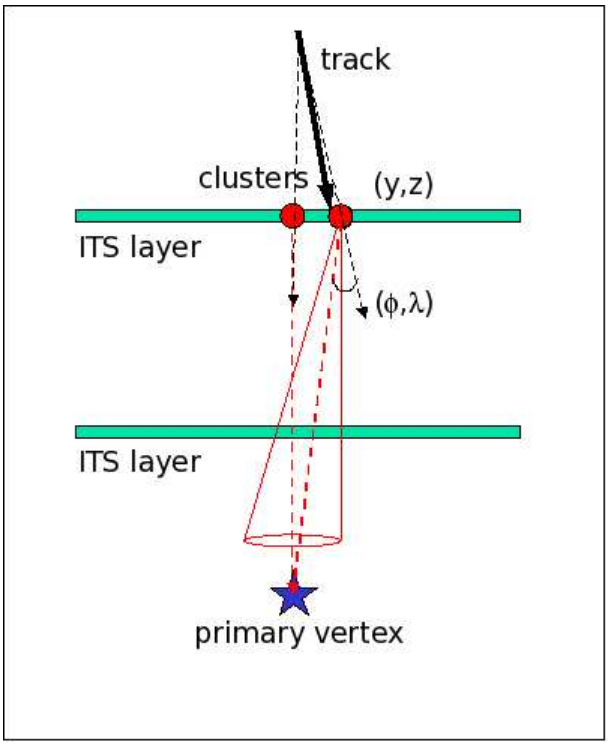

Fig. 3: Taking into account the information about the primary vertex position by applying a 'vertex constraint' (see the text).

Detailed Monte-Carlo studies performed with ALICE offline simulation and reconstruction framework AliRoot [1] show that the outlined ad-hoc 'vertex constraint' significantly reduces the probability of wrong cluster assignment, and so the quality of reconstructed tracks improves. Unfortunately, the procedure is not free of flaws. For example, for each of the tracks, it uses several times (even though with different 'weights') the same information about the primary vertex position. This is done as many times per a track as there are detector layers. Consequently, one of the undesirable features is that the resulting covariance matrix of the track parameters becomes underestimated (which can be overcome by an additional refitting step, however).

In future, we would like to incorporate the vertex constraint into the Kalman-filter track finding in a stricter way. A possible solution can probably be found by introducing the information about the primary vertex position in the form of Bayesian priors.

3 Statistical problems with particle identification

The ALICE experiment is able to identify particles with momenta from 0.1 GeV/c and up-to a few tens GeV/c (statistically, on the relativistic rise of $\mathrm{d}E/\mathrm{d}x$ in TPC). This can be achieved by combining several detecting systems that are efficient in some narrower and complementary momentum sub-ranges. The situation is complicated by the amount of data to be processed (about 10^7 events with about 10^4 tracks in each). Thus, the particle identification (PID) procedure should satisfy the following requirements:

1. It should be as much as possible automatic.
2. It should be able to combine PID signals of different nature (*e.g.* $\mathrm{d}E/\mathrm{d}x$ and time-of-flight measurements).
3. When several detectors contribute to the PID, as it is shown in Fig. 4, the procedure must profit from this situation by providing an improved PID.
4. When some of the detectors can not separate the particle species, the signals from the other detectors must not affect the combined PID.
5. It should take into account the fact that, due to different event and track selection, the PID depends on the kind of analysis.

The method described here is similar to that in Ref. [4]. Let $r(s|i)$ be a conditional probability density function to observe in some detector a PID signal s if a particle of type i ($i = \mathrm{e}, \mu, \pi, \mathrm{K}, \mathrm{p}, ...$) is detected. The probability to be a particle of type i if the signal s is observed, $w(i|s)$, depends not only on $r(s|i)$, but also on how often this type of particles is registered in the experiment (*a priori* probabilities C_i to find a particle of i-type in the detector). The corresponding relation is given by Bayes's formula:

$$w(i|s) = \frac{r(s|i)C_i}{\sum_{k=e,\mu,\pi,...} r(s|k)C_k}. \quad (1)$$

If C_i and $r(s|i)$ are not strongly correlated we can rely on the following approximation:

– The functions $r(s|i)$ reflect only properties of the detector ('detector response functions') and do not depend on other external conditions like event and track selections.
– On the contrary, the quantities C_i ('relative concentrations' of particles of type i) do not depend on the detector properties, but reflect the external conditions, selections etc.

In the case of several detectors, the signal s is replaced by a vector of the PID measurements $\bar{s}$ in the detectors. The response function $r(s|i)$ becomes some 'combined response function' $R(\bar{s}|i)$ of the whole system of the detectors involved (in the simplest case, this is the product of the single-detector PID response functions). The PID procedure is then done in steps:

– First, the detector response functions are obtained (theoretically, or in beam tests). This can be done 'once and forever' before the reconstruction starts as a part of detector calibration.
– Second, for each track, a value $R(\bar{s}|i)$ is calculated using the PID signals measured for this track. This is done during the event reconstruction.
– Third, the relative concentrations of particle species C_i are estimated for the subset of events and tracks selected for a specific physics analysis. For obtaining better results, the particle concentrations C_i can be considered as functions of momentum.
– Finally, for each track within the selected subset, the array of probabilities $w(i|\bar{s})$ is calculated using the formula (1). This steps, as well as the previous one, can be done only during the physics analysis of the data.

Doing the particle identification in this way, we naturally satisfy all the requirements mentioned at the beginning of this paragraph. However there are two problems which we are still working on.

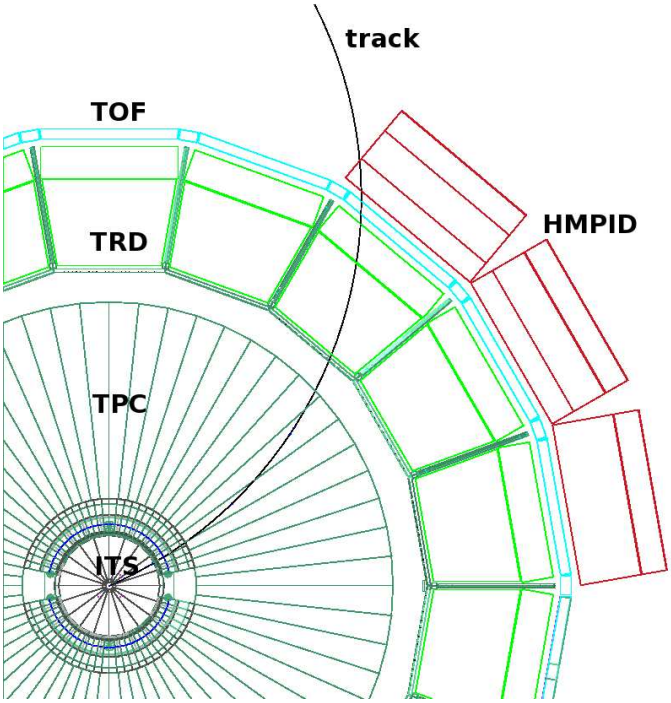

Fig. 4: The particle identification for the shown track is done by combining PID signals from five detectors: ITS, TPC, TRD, TOF and HMPID.

Since the results of such a PID procedure depend explicitly on the choice of *a priori* probabilities C_i (and, in fact, this kind of dependence is unavoidable in any approach), the question of stability of the results with respect to the choice of C_i becomes important. This problem seems to be related to the 'Punzi effect' discussed in Ref. [5]. At lower momenta, there is always some momentum region where the single-detector response functions for different particle types of at least one of the detectors do not significantly overlap, and so the stability is guaranteed. The final PID weights $w(i|\bar{s})$ are defined by the detector response functions. The more detectors enter the combined PID procedure, the wider this momentum region becomes and the results are more stable. But, finally, as the momentum goes up, all the detectors lose separation power, and the PID decision is given by the bare priors C_i (which we can not in this case estimate independently). The question is: can we somehow quantify the 'contribution of priors' to the final PID weights so that if they become dominant, relative to the 'contribution of detector responses', we know when to stop trying to identify particles of higher momenta?

The second problem is of a different nature. The formula (1) fundamentally assumes that all the components of vector $\bar{s}$ are the results of PID measurements done for *the same* particle. In other words, the procedure of assigning clusters to tracks has to be ideal, which is not the case in reality. The consequences are seen, for example, in Fig. 5, where the PID efficiency and contamination in ALICE TOF detector are shown. In spite of the fact that separation of the particle species by the time of flight method improves greatly with decreasing momentum (Fig. 5, upper pad), the actual situation with the PID becomes worse, especially for the particles below 0.5 GeV/c (Fig. 5, lower pad). This is because the low-momentum particles decay, suffer from scattering in material and so have a higher probability of being assigned to the wrong cluster in the TOF detector. This mismatching effect is not taken into account by the formula (1), and so the combined PID result becomes biased at low momenta.

The effect of mismatching can be corrected by excluding from the vector $\bar{s}$ the components that deviate too much from 'reasonable expectation'. This is possible, for example, in the case of the ALICE TOF detector, because we calculate the expected time of flight during the track finding in ITS, TPC and TRD. However, in general case, we may not know what the 'reasonable expectation' is. Also, applying sharp cuts in an otherwise smooth procedure based on formula (1) may cause additional difficulties with finding the best values for the cuts. Thus, a better solution for the problem of dealing with the mismatching is still to be found.

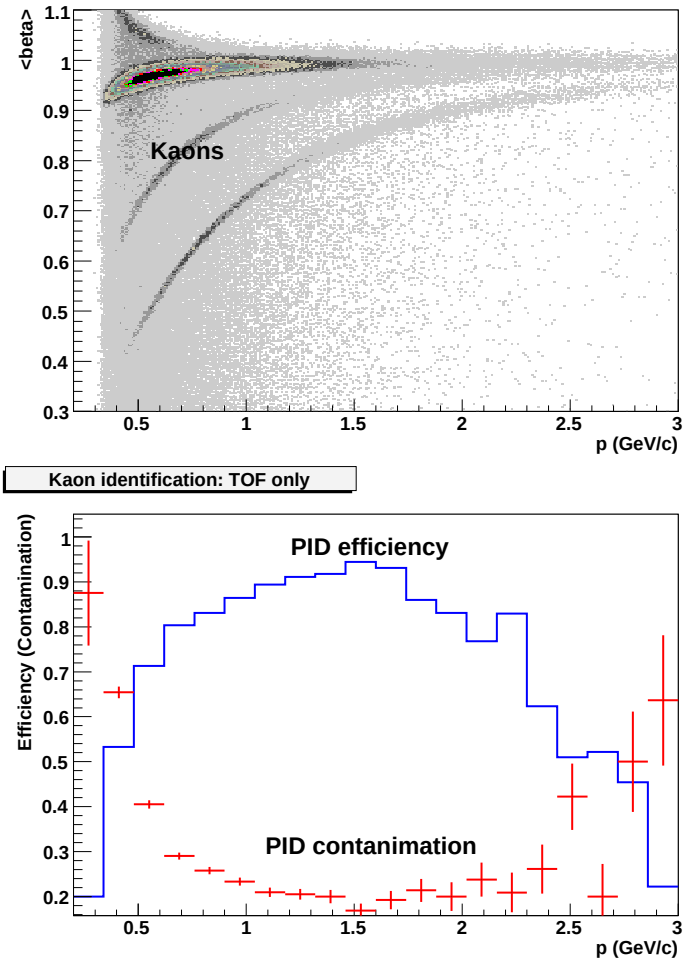

Fig. 5: PID efficiency and contamination for charged kaon identification with ALICE TOF detector. The deterioration of PID at low momenta is the consequence of the mismatching effect (see the text).

4 The wish-list

The statistical problems arising in event reconstruction in ALICE and discussed above represent the ALICE reconstruction statistical wish-list. In short, we would like to find better, theory supported, ways for

1. introducing constraints in the standard Kalman filter (needed for improving the track finding in high-multiplicity events in the ALICE ITS);
2. quantifying the relative importance of prior information and the results of actual measurements when making Bayesian decision (needed to define the highest momentum up-to which the Bayesian particle identification still makes sense);
3. taking into account mismeasurements in Bayesian combination of information (needed for improving the low-momentum particle identification in ALICE).

References

[1] F. Carminati *et al.* [ALICE Collaboration], J. Phys. G **30**, (2004) 1517.

[2] B. Alessandro *et al.* [ALICE Collaboration], J. Phys. G **32** (2006) 1295.

[3] P. Billoir, Nucl. Instrum. Meth. A **225** (1984) 352.

[4] M. Aguilar-Benitez *et al.*, Z. Phys. C **50** (1991) 405.

[5] P. Catastini and G. Punzi, *PHYSTATO5: Statistical Problems in Particle Physics, Astrophysics and Cosmology, Oxford, England, United Kingdom, 12-15 Sep 2005*.

Dilution of a Statistical Significance of a Signal in the Higgs Boson Searches in the $\mathrm{H} \to \mathrm{ZZ}^{(*)} \to 4\mu$ Channel at LHC

A. Drozdetskiy, A. Korytov, G. Mitselmakher
University of Florida, Gainesville, FL, USA

Abstract
Should an event excess compatible with the $\mathrm{H} \to \mathrm{ZZ}^{(*)} \to 4\mu$ decay channel be observed at LHC, the statistical significance of the access must be properly scaled down to account for the systematic errors and the fact that the search is performed in a wide-open range of possible Higgs boson masses. We present results of studies addressing both of the two contributions and show that the required corrections in Higgs boson search in this particular channel are by far not negligible.

1 Introduction

The $\mathrm{H} \to \mathrm{ZZ}^{(*)} \to 4\mu$ process is one of the cleanest channels for discovering the Standard Model Higgs boson at the LHC. Signal events have a relatively narrow resonance peak in the four-muon invariant mass $m_{4\mu}$ distribution, which allows for effective background suppression. This channel is the frontrunner for discovering the Higgs boson in a broad range of its possible masses [1]. Figure 1 shows distributions of $m_{4\mu}$ for signal and background events after all analysis cuts [2] for the CMS detector. Figure 2 shows integrated luminosity needed for discovering the Higgs boson in its four-muon decay channel with a 5σ significance.

If an excess of events is indeed observed in such a search, to evaluate its true significance, one needs to address two issues [2]: background uncertainties (cross sections, efficiencies, etc.) and the fact that the search is actually carried in a broad range of masses (see e.g. Refs. [3]–[5]). Both will result in a dilution of an observed event excess significance. This paper quantifies these two effects in the context of a counting experiment approach [1]. The first effect is evaluated by folding a background probability density function $f(b)$ into significance calculations, while the second effect is addressed by a brute force of generating a large number of pseudo-experiments.

In these studies we use S_{cL} as a significance estimator:

$$S_{cL} = sign(n_o - b)\sqrt{2n_o \ln(n_o/b) - 2(n_o - b)}, \tag{1}$$

where n_o is the number of observed events and b is the expected background; the signal is then defined as $s = n_o - b$. The signed S_{cL} takes negative values for the cases when a lack of events is observed. It is worthwhile pointing out that this definition of significance follows very closely the significance defined as the probability P to observe $n \geq n_o$ background events:

$$P = \int_S^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx. \tag{2}$$

2 Including $\mathrm{ZZ}^{(*)} \to 4\mu$ background uncertainties in significance calculations

After applying analysis cuts, ZZ production is the dominant irreducible background, with all other processes giving much smaller contributions. This reduces the analysis of systematic errors to the $ZZ \to$

[1] Although we use the counting experiment approach as the case study, the problem is of a very general nature. For example, an analysis based on Log-Likelihood Ratio built for the entire $m_{4\mu}$ spectrum would suffer from the same problems and with a similar scale of the effect.

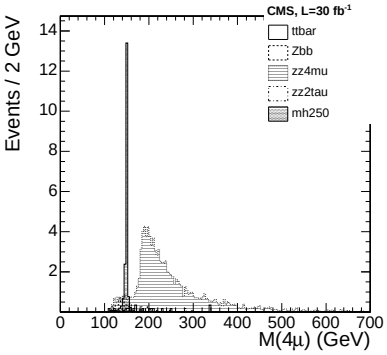

Fig. 1: Four-muon invariant mass distribution for the three background subprocesses and a Higgs-boson signal at $M_H = 150\,\mathrm{GeV}/c^2$, after applying cuts on muon isolation and p_T.

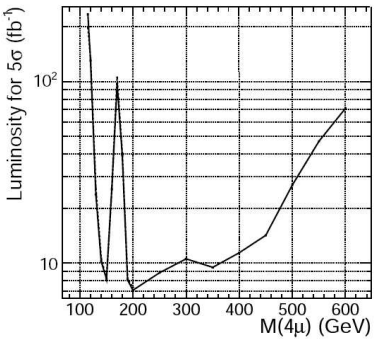

Fig. 2: Integrated luminosity needed for discovering the Higgs boson in its four-muon decay channel with a 5σ significance. Systematic errors are not included; neither is included the effect of significance dilution due to the fact that the Higgs boson mass is not known a priori.

4μ process. The main uncertainties are as follows: parton distribution function (PDF) and QCD scale uncertainties, differences between the predictions of leading order (LO) Monte Carlo models and the observed data, uncertainty in integrated luminosity, trigger efficiency, muon reconstruction efficiency, muon isolation cut efficiency, four-muon invariant mass scale and resolution. To minimize systematic errors due to these uncertainties, we developed methods for evaluating muon reconstruction and isolation cut efficiencies directly from data. Furthermore, we propose to estimate the ZZ-background around a particular $m_{4\mu}$ point (signal region) via a reference to a *measured* control sample. We explored two options for a control sample: an inclusive $Z \to 2\mu$ process and sidebands in the $m_{4\mu}$-distribution itself. Use of appropriate control samples completely eliminates uncertainties associated with measuring the luminosity and reduces significantly the sensitivity to theoretical uncertainties in PDF and QCD-scales. See Ref. [2] for details on all of the above.

Figure 3 shows systematic errors in estimation of the $ZZ \to 4\mu$ background via a measured number of $Z \to 2\mu$ events. We further assume a log-normal form for a probability density function $f(b)$ with the expected number of background events b_0 and a combined relative uncertainty δ:

$$f(b) = \frac{1}{\sqrt{2\pi}\ln(k)} \exp\left(-\frac{\ln^2(b/b_0)}{2\ln^2(k)}\right)\frac{1}{b}, \tag{3}$$

where $k = 1 + \delta$. For relatively small errors, this form of equation gives a Gaussian distribution with average b_0 and $\sigma = \delta \cdot b_0$. However, unlike the Gaussian distribution, the log-normal distribution is always positively defined, gives an intuitively correct representation for very large uncertainties ("a factor of two uncertainty" would mean $k = 2$), and has more conservative tails than the Gaussian distribution.

When the background b is estimated from a number of events B in sidebands via $b = \rho B$, one needs to take into account systematic errors in the factor ρ and statistical fluctuations in the number B. The latter is dominant in the case of a Higgs search at LHC. The background *pdf* $f(b)$ associated with statistical fluctuations in sidebands can be obtained using Bayes' theorem and a flat prior:

$$f(b) = \frac{1}{\rho}\frac{(b/\rho)^B e^{-b/\rho}}{\Gamma(B+1)}. \tag{4}$$

Then, the probability of observing at least n_o events becomes

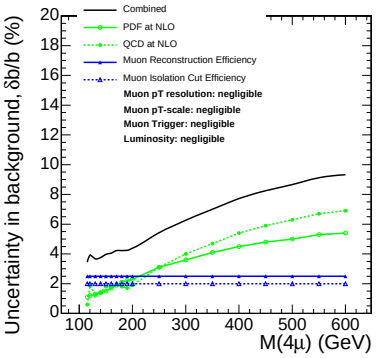

Fig. 3: Uncertainties in the number of $ZZ \to 4\mu$ background events in the signal region window at different $m4\mu$ referenced to the number of $Z \to 2\mu$ events.

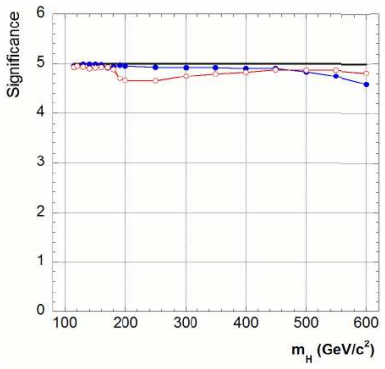

Fig. 4: Dilution of significance due to uncertainties on the backgrounds at luminosity corresponding to S=5: solid line – no uncertainties, curve with filled circles – normalization to Z, curve with empty circles – normalization to ZZ (syst+stat uncertainties).

$$P = p(n \geq n_o|b) \otimes f(b) = \int_0^{+\infty} p(n \geq n_o|b)\, f(b)\, db, \tag{5}$$

which can be easily converted into significance as defined in Eq. 2. Dilution of significance due to the background uncertainties at luminosities at which significance would be 5, if not for these uncertainties (Fig. 2), is shown as a function of the Higgs boson mass in Fig. 4.

3 Dilution of significance due to a search being carried out in a broad ranges of possible Higgs boson masses

The analysis is based on performing $\approx 10^8$ pseudo-experiments. Each pseudo-experiment is an ensemble of N randomly generated events with the $m_{4\mu}$ probability density function as given by the expected background. The number of events N per pseudo-experiment is sampled according to the Poisson distribution. For each pseudo-experiment, we conduct a pseudo-search for a Higgs signal. Within a priori defined range of search (110-600 GeV$/c^2$), we slide a $m_{4\mu}$-dependent mass window $\Delta m(m_{4\mu})$, whose width is driven by the expected signal peak width in such a way that it would give the best significance S_{cL}, should the Higgs boson be indeed present at that mass. The scanning step of 0.4 GeV$/c^2$ is much smaller than the mass window width. The mass point M_{max} at which the observed S_{cL} is maximum (S_{max}) is a Higgs candidate with a naive significance S_{max}. Figures 5 and 6 give an example of a pseudo-experiment where a plain statistical fluctuation results in a pseudo-discovery with $S > 5$.

After performing 10^8 pseudo-experiments, we obtain a cumulative probability function $P(S_{max} > S)$, Fig. 7. The dashed (lower) line in Figure 7 shows the integrated probability that one associates with the true significance via Eq. 2. One can see that the observed probability $P(S_{max} > S)$ is substantially higher. If desired, the real probability $P(S_{max} > S)$ as obtained in such studies can be used to derate the observed S_{max} to its true significance S_{true} via Eq. 2. The result of such significance derating is presented in Fig. 8. We also checked (Ref. [2]) that the obtained results depend only on a $m_{4\mu}$-range of search and signal width; they do not depend on an integrated luminosity and/or the actual background shape.

4 Summary

Should an event excess compatible with the H $\to$ ZZ$^{(*)}$ $\to 4\mu$ decay channel be observed at LHC, statistical significance of the access must be properly scaled down to account for uncertainties in background

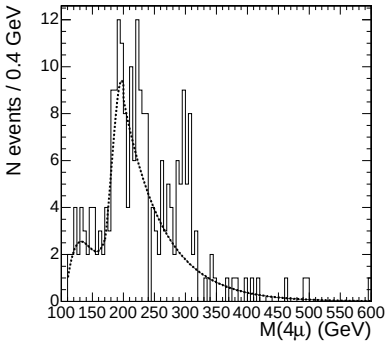

Fig. 5: A pseudo-experiment example (histogram). The dashed line is the four-muon mass *pdf* for the ZZ-background.

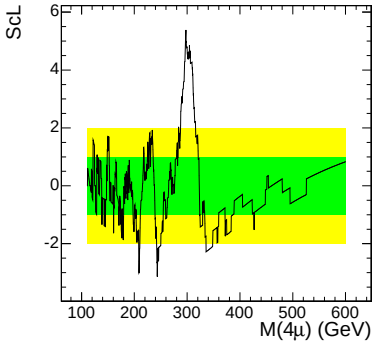

Fig. 6: S_{cL} profile for the pseudo-experiment example shown on the left. Green (inner) and yellow (outer) bands denote $\pm 1\sigma$ and $\pm 2\sigma$ intervals. Steps one the plot are due to events coming in or dropping off from the Δm-window as we scan it along x-axis.

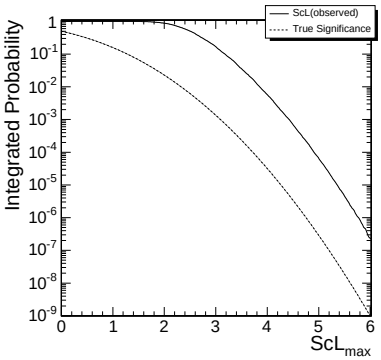

Fig. 7: S_{cL} cumulative probability density function.

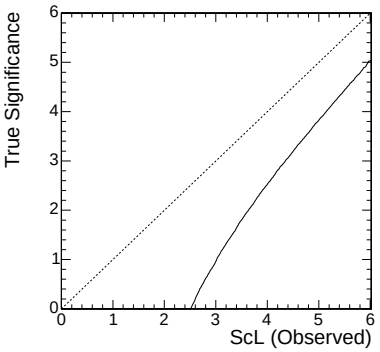

Fig. 8: Local significance derating (solid line).

evaluation and the fact that the search is performed in a broad range of possible Higgs boson masses. We present methodologies for taking into account both effects. We show that systematic/statistical errors in background estimations may result in significance derating by as much as $\sim$0.4 units. Moreover, since the search will be conducted in the Higgs boson mass range as broad as 100-600 GeV/c^2, the significance will have to be further derated by as much as one unit.

Acknowledgments

The authors would like to thank R. Cousins, S. Nikitenko, and G. Quast for very fruitful discussions of the analysis methodology and the obtained results.

References

[1] M.Della Negra, A. Petrilli, A. Ball, L. Foa et al., J. Phys. G: Nucl. Part. Phys., vol34 (2007), 995-1579.

[2] S. Abdullin et al., CMS Note 2006/122.

[3] E. L. Lehmann, The Annals of Mathematical Statistics, Vol. 28 1, 1, (1957).

[4] R. O'Neill and G. B. Wetherill, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 33 2, 218, (1971).

[5] B. Knuteson, thesis, 2000.

A Pitfall in Evaluating Systematic Errors

James T. Linnemann
Michigan State University, 3245 BPS Building, E. Lansing, Michigan 48823; linnemann@pa.msu.edu

Abstract
A common practice in evaluating the contribution of systematic uncertainties is to change the parameter by $\pm 1\sigma$ of systematic uncertainty, then to add the induced changes in the result in quadrature. This is typically justified by arguing that the individual systematic effects are statistically independent. However, if the response to one parameter depends on the value of another parameter ("interaction" in the statistical Design of Experiments jargon), a significant portion of the actual uncertainty may be missed by applying the usual formula. In particular, any terms such as xy are completely missed unless more than one systematic parameter is varied in a single evaluation.

1 Introduction: Ideal Evaluation of Systematic Errors?

We have a result f and we wish to express our uncertainty in f due to our imperfect knowledge of parameters the results depends upon (the systematic uncertainties). The result f might be a single top cross section, or a Higgs mass upper limit.

This often involves an implicitly Bayesian point of view in which we have a prior distribution π of possible values for the unknown parameters. Here I'll use a 2-dimensional example $\pi(x, y)$, where the imperfectly known parameters might be a jet energy scale factor, or a luminosity calibration constant.

For algebraic simplicity in what follows, consider the coordinates for (x, y) to be centred about their nominal values, and define $d = f - f_o = f(x, y) - f(0, 0)$ to be the change in the response in moving away from the nominal values.

Then the ideal evaluation of the variance V in the output d induced by our imperfect knowledge of the systematic parameters x, y can be written as

$$V[f] = \int dx\, dy\, d^2(x, y)\, \pi(x, y) \tag{1}$$

This formulation takes f_o as the reference point for the variance calculation; it implicitly assumes that f_o is the expected value of f–a symmetry assumption about f and π. Unfortunately, to carry out this integral requires two things we don't have: an analytical form for $d(x, y)$, and a reliable probabilistic description of our knowledge of the parameters, $\pi(x, y)$.

2 Standard Evaluation of Systematics

In contrast to this ideal procedure, typical practice is to instead run simulations to evaluate f at $(0, 0)$, and then at $+1\sigma_i$ for each systematic. Then defining the differences $d_i = d(0 + 1\sigma_i)$, one combines in quadrature:

$$S^2 = \sum d_i^2 \tag{2}$$

and reports the result with systematic error

$$f_0 \pm S \tag{3}$$

How do we justify this procedure? I believe we appeal to the first order covariance formula [1]:

$$V_1[f] \approx \sum_i \sum_j \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} Cov(x_i, x_j) \tag{4}$$

Table 1: OFAT vs. ideal variance estimation vs. DOE

f	S^2	V	2^2 DOE	comment
$x+y$	a^2+b^2	a^2+b^2	a^2+b^2	truly linear
x^2+y^2	a^4+b^4	$3a^4+2a^2b^2+3b^4$	0	quadratic; S^2 not so hot, DOE worse
xy	0	a^2b^2	a^2b^2	bilinear; S^2 fails completely, DOE fine

where the partials are evaluated at the origin. Three features of this expression are worth noting. First, the distribution π has vanished completely, having been replaced by its first two moments. Second, we have again implicitly assumed that $f(0) = f_0$ is the expected value of f over π. Third, if the systematic parameters x_i are uncorrelated, the covariance matrix may be replaced by its diagonal elements:

$$V_d[f] \approx \sum_i \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2. \tag{5}$$

On approximating the partial derivatives by finite differences $\frac{\partial f}{\partial x_i} \to \frac{\Delta f}{\Delta x_i} = d_i/\sigma_i$ we have arrived at Eq. 2. We justified this standard procedure by appealing to the lack of correlation between our systematic parameters in $\pi(x, y)$, and by approximations appropriate to measuring a first order effect.

3 One Factor At A Time

I learned at my thesis advisor's knee: "A good physicist should be able to diagnose and fix almost any single problem. It should require two things both wrong simultaneously to seriously confuse a physicist." This leads to an immediate corollary: "Changing more than one thing at a time is *asking* for trouble." This is another, practical, justification for the procedure in Eq. 2. The statisticians have a name for this procedure: One Factor at a Time (OFAT).

Let's see how well OFAT works, compared to the ideal evaluation. Assume an uncorrelated normal distribution for the systematics

$$\pi(x, y) = N(0, a) \otimes N(0, b) \tag{6}$$

and consider its effects on several functional forms for f. Table 1 gives the results; for now concentrate on the S^2 and V columns.

If all you are interested in is the linear terms, OFAT does fine. But the whole reason we perform a MC evaluation is that we don't know the functional form f, which may not be politely linear in all parameters.

What went wrong with the nonlinear terms? First, the quadratic terms are underestimated substantially: finite differences are not enough to fully account for the nonlinearities. Second, the diagonal covariance matrix does *not* protect us from the xy terms. This is because xy and its derivatives are all zero on the axes, as if f were actually independent of x, y. But $f = xy$ has a twisted surface: the x derivatives depend on y and vice versa. To be sensitive to such behavior one *must* consider points off the axes: that is, change more than one thing at a time. If you are considering *any* nonlinearities of f to evaluate V, you need at least all the quadratic terms in the Taylor series for V, which puts (by rotation) xy terms on the same footing as the "curvature" terms involving x^2, y^2. Another way of stating the matter is that accounting for the nonlinear effects in f means estimating the higher moments from the lower ones–assuming a fair bit about the shape $\pi(x, y)$ describing our knowledge of the parameters. But to know which higher moments are relevant also requires knowing f, and we typically don't.

I have slightly oversimplified the standard procedure. Very often, one runs MC at 0 and both of $\pm 1\sigma$. In part this is intended to make us sensitive to curvature terms, where a nonlinearity changes our response differently in the $\pm$ directions, though there are issues as to how to treat the resulting

Table 2: Top: values of x, y for Runs of 2^2 design, and results for various f. Bottom:Analysis of Runs.

Run	x/σ_x	y/σ_y	Sgn(xy)	$f = x + y$	$f = x^2 + y^2$	$f = xy$
f_1	+1	+1	+	$a + b$	$a^2 + b^2$	ab
f_2	+1	-1	-	$a - b$	$a^2 + b^2$	$-ab$
f_3	-1	+1	-	$-a + b$	$a^2 + b^2$	ab
f_4	-1	-1	+	$-a - b$	$a^2 + b^2$	$-ab$
$A = [(f_1 - f_3) + (f_2 - f_4)]/4 =$				a	0	0
$B = [(f_1 - f_2) + (f_3 - f_4)]/4 =$				b	0	0
$C = [(f_1 - f_2) + (f_4 - f_3)]/4 =$				0	0	ab
$f_0 = [(f_1 + f_4) + (f_2 + f_3)]/4 =$				0	$a^2 + b^2$	0

asymmetric errors [2]. However, even this improved sensitivity to the quadratic on-axis terms still leaves us *completely blind* to the bilinear xy terms.

4 Design of Experiments (DOE): not a funding agency

DOE aims to choose good patterns ("designs") for exploring the x, y space. In an OFAT design, each MC "run" (evaluation of f) changes only one parameter from nominal. DOE typically suggests designs (sets of MC runs) in which *every* parameter is varied from its nominal value, in every run.

OFAT is not a term of endearment. Statisticians wish you had talked to *their* thesis advisor, who told them to *always* change more than one thing at a time [3, 4]. I believe we should take such exhortations seriously, even though Roger Barlow, whose advice I normally try to follow assiduously, has commented [5] that in particle physics, "The whole experimental design field–Latin squares and similar techniques used to minimize uncontrollable effects–is not needed as such effects are not a problem." But missing xy effects while looking for x^2 effects does seem like a problem to me.

4.1 A 2-D Example of DOE

Many treatises [6, 7] are devoted to DOE methods. In Table 2 gives a 2 variable example. Though it is too simple to exhibit all the features and the advantages of more sophisticated designs, it's enough to glimpse how things work. There are four MC runs, each of which vary both parameters.

The top part of the table gives the (x, y) values to use in evaluating f, and the resulting f values, assuming $\sigma_x = a,\ \sigma_y = b$, for the same functional forms of f as used in Table 1. The bottom part of the table shows how the evaluations can be linearly combined to "fit" the data to the model $f_0 + Ax + By + Cxy$ appropriate for this design. There are two evaluations of each term which can be cross-checked; the redundancy could be used to assess statistical errors if these were not already available.

How well does this simple design perform? The DOE variance deduced from the model parameters is just $A^2 + B^2 + C^2$. Referring back to the DOE column on Table 1, as advertised it correctly extracts the linear and interaction terms. To recover the curvature term, a more complex " composite" design with more runs is needed, to fit coefficients missing from our oversimplified model.

4.2 OFAT vs. Design

DOE procedures assume each run has sufficient statistical power to measure effects of interesting size. Then for the same number of runs, you can learn (by averaging) more than you could in OFAT (with a single run for each coefficient). The difference becomes more dramatic, the larger number of parameters you have to explore. DOE methods intend to allow you to search for effects in order of likely importance–linear ("main") effects first, then 2-variable bilinear ("interaction"), then 3-variable effects. Typically, a few effects dominate and one expects small interaction effects if the corresponding main effects are

small; of course a pure $f = xy$ would violate that hope. "Interaction" is the DOE term for a twisting in the response surface f, i.e. a slope wrt. a variable depending on the value of another variable.

OFAT is not without advantages. It is simpler to set up (change only one thing per run), and works well if the main (linear) effects dominate; and it's easier to analyze without specialized software. One bad run loses less information, and curvature can be recognized with help of the zero point. Designed experiments can estimate interactions or show them to be negligible, give savings (especially for larger numbers of parameters), result in more accurate determinations for fixed effort (since multiple runs contribute to measurement of each significant effect), and with the use of the nominal point can also identify curvature effects.

4.3 A DOE glossary

The DOE literature requires translation for physicists. Understanding the terminology, and motivations behind standard designs, may help when reading (or applying to non-analysis tasks we do).

Response surface f

Factor x_i a systematic parameter; from Analysis of Variance, also a linear combination of variables

Level a value used in a design. $\pm\sigma$ is two levels; adding 0 makes three

Additive (effects or models) f linear in x_i's; also called main effects

Active factors main effects which are significant

Interaction multilinear terms in f: $x_i x_j$ or trilinear or higher

Curvature quadratic terms x_i^2

Twisting of Response Surface $\partial_x f(x, y) \neq \partial_x f(x, 0)$

Factorial Design (or simply Design) Nothing to do with a gamma function. A plan for sampling from the x_i space. The name comes from Factor Analysis, related to Analysis of Variance.

Full Factorial Design (L^k) All L^k combinations of L levels of k factors; our example was 2^2.

Fractional Factorial Design (L^{k-m}) Not all L^k combinations; can't distinguish ("confounds") m kinds of interactions; if you were fitting a k-dim polynomial model to the data you wouldn't have enough data points to separately distinguish all the coefficients that you could with a L^k design. "Screening" designs are examples, chosen to focus directly on main effects.

Confounding A fractional design can't distinguish all interactions; it can detect whether one class is active, but may confound higher order interactions with lower order ones or main effects.

4.4 Motivation for typical Designs

Typical goals of DOE are optimization, or robustification, which are somewhat different than our goal, which is sensitivity analysis. In optimization (finding a local minimum or maximum), one seeks the best pattern of points for searching for the best yield from epoxy curing time and temperature for building a tracker system, or finding a minimum variance set of cuts for a mass determination. One picks points which are good for calculating numerical derivatives, seeks an uphill direction or checks for a hilltop. Response surface methodology [8] is fully quadratic DOE; "composite" designs add more points (including the nominal 0 setting) to the classic designs to better characterize quadratic curves. In robustification [9] one looks instead for stationary points (maxima, or ridges, for which the outcome locally does not depend on parameters which are hard to control: say, the humidity when curing epoxy). The methods also treat compromise among multiple objectives (f's). Warning: this "Taguchi" design literature has a number of strangely named metrics.

5 Summary

Even if your systematics *are* independent, your measurement probably correlates them for you. If you worry about curvature (asymmetry of response to $\pm 1\ \sigma$), you should worry about xy terms too–they're at the same order in the Taylor expansion. But OFAT is blind to multi-linear (xy-like) terms; you *must* leave OFAT to see these terms. OFAT also does not get all the nonlinear effects you might have hoped for (even if you don't ignore the point at the nominal settings). Design of Experiments just might help.

Note Added: After the conference I was encouraged to coordinate a workshop on this subject. A conversation with Jim Berger led me to the SAMSI web site and the literature [10] in the closely related field of computer "experiments".

References

[1] F. James, "Statistical Methods in Experimental Physics", World Scientific, 2006, section 2.4.5

[2] R. Barlow, "Asymmetric Statistical Errors", in proceedings of PHYSTAT 2005, ed. L. Lyons and M. Ünel, Imperial College Press, 2006.

[3] N. Reid, "Experimental Design", PHYSTAT 2007, CERN, Geneva, Switzerland (these proceedings). Thanks to Nancy for agreeing to give this introduction.

[4] B. Gunter, "How statistical Design Concepts Can Improve Experimentation in the Physical Sciences", Computers in Physics, 7 May (1993).

[5] R. Barlow, "Introduction to Statistical Issues in Particle Physics", Proceedings of PHYSTAT2003, ed. L. Lyons, R. Mount, R. Richtmeyer.

[6] D. R. Cox, N. Reid, "The Theory of Design of Experiments", Chapman & Hall, 2000.

[7] "The NIST/SEMATECH e-Handbook of Statistical Methods", `http://www.itl.nist.gov/div898/handbook/`,2007

[8] G. Box, J. Hunter, W. Hunter, "Statistics for Experimenters", Wiley, 2005. Mount, R. Reitmeyer, SLAC, 2003, and arXiv:physics/0311105

[9] W. Fowlkes, C. Creveling, "Engineering Methods for Robust Product Design", Addison-Wesley, 1995

[10] T. Santner, B. Williams, W. Notz, "The Design and Analysis of Computer Experiments", Springer 2003; K. Fang, R. Li, A. Sudjianto, "Design and Modeling for Computer Experiments", Chapman & Hall, 2006; J. Kleijnen, "Design and Analysis of Simulation Experiments", Springer, 2007.

Some Aspects of Design of Experiments

Nancy Reid
University of Toronto, Toronto Canada

Abstract
This paper provides a brief introduction to some aspects of the theory of design of experiments that may be relevant for high energy physics experiments and associated Monte Carlo investigations.

1 Introduction

'Design of experiments' means something specific in the statistical literature, which is different from its more general use in science. The key notion is that there is an *intervention* applied to a number of *experimental units*; these interventions are conventionally called treatments. The treatments are usually assigned to experimental units using a randomization scheme, and randomization is taken to be a key element in the concept in the study of design of experiments. The goal is then to measure one or more responses of the units, usually with the goal of comparing the responses under the various treatments. Because the intervention is under the control of the experimenter, a designed experiment generally provides a stronger basis for making conclusions on how the treatment affects the response than an observational study.

The original area of application was agriculture, and the main ideas behind design of experiments, including the very important notion of randomization, were developed by Fisher at the Rothamsted Agricultural Station, in the early years of the twentieth century. A typical agricultural example has as experimental units some plots of land, as treatments some type of intervention, such as amount of or type of fertilizer, and as primary response yield of a certain crop. The theory of design of experiments is widely used in industrial and technological settings, where the experimental units may be, for example, manufactured objects of some type, such as silicon wafers, the treatments would be various manufacturing settings, such as temperature of an oven, concentration of an etching acid, and so on, and the response would be some measure of the quality of the resulting object. In so-called *computer experiments*, the experimental units are simulation runs, of, for example, a very complex system such as used for climate modelling or epidemic modelling; the 'treatments' are settings for various systematic or forcing parameters, and the response is the output of the climate model or epidemic model. Principles of experimental design are also widely used in clinical trials, where the experimental units are often patients, the treatments are medical interventions, and the response is some measure of efficacy of the treatment.

If the experimenter is able to ensure that the experimental units are homogeneous, and the treatments are assigned randomly, then there is some basis for attributing a difference in response under different treatments to the effect of the treatment; in some contexts the effect might be presumed then to be a causal effect. In most settings the randomization is subject to some constraints; for example experimental units might be subdivided into more homogeneous groups, conventionally called blocks, and treatments assigned to units at random within blocks. In clinical trials it is more or less impossible to ensure homogeneity of treatment groups, and several background variables will be recorded in order to attempt to assess whether an observed difference between two treatments might be ascribed to some other feature, such as, for example, a possibly small but important age difference between the groups. Randomization will on average balance out differences on all these so-called confounding variables, but with small groups of patients the balance may be far from perfect. In computer experiments such elaborate protections will not normally be needed, although it might be used if there could be, for example, some potential drift in conditions over time.

Table 1: A 2^4 factorial design of 16 runs, with the response labelled according to conventional notation for the factor levels.

run	A	B	C	D	response
1	-1	-1	-1	-1	$y_{(1)}$
2	-1	-1	-1	$+1$	y_d
3	-1	-1	$+1$	-1	y_c
4	-1	-1	$+1$	$+1$	y_{cd}
5	-1	$+1$	-1	-1	y_b
6	-1	$+1$	-1	$+1$	y_{bd}
7	-1	$+1$	$+1$	-1	y_{bc}
8	-1	$+1$	$+1$	$+1$	y_{bcd}
9	$+1$	-1	-1	-1	y_a
10	$+1$	-1	-1	$+1$	y_{ad}
11	$+1$	-1	$+1$	-1	y_{ac}
12	$+1$	-1	$+1$	$+1$	y_{acd}
13	$+1$	$+1$	-1	-1	y_{ab}
14	$+1$	$+1$	-1	$+1$	y_{abd}
15	$+1$	$+1$	$+1$	-1	y_{abc}
16	$+1$	$+1$	$+1$	$+1$	y_{abcd}

2 Factorial experiments

A very useful class of designed experiments are *factorial* experiments, in which the treatments are combinations of levels of several factors. These are used in many applications of experimental design, but especially in technological experiments, where the factors might be, for example, time, concentration, pressure, temperature, etc. It is very common to use a small number of levels for each of the factors, often just two levels, in which case a design with k treatment factors has 2^k treatments and is called a 2^k factorial design. As an example, in a computer experiment, if there were 10 systematic parameters then a full 2^{10} factorial might have each systematic parameter set at $\pm 1\sigma$; of course in this case it would be usual as well to have one or more runs at the central 'mean value' or 'best guess' of all the systematics.

A 2^k factorial design is to be contrasted with a one-factor-at-a-time, or OFAT, design, where, for example, a single simulation run would keep 9 of the 10 systematics at their mean values and use $+1\sigma$ for the 10th systematic; the next run would do the same but use -1σ for the 10th systematic, and subsequent runs would proceed through the other values. An OFAT design has the advantage that if a large change is observed in a single run, the change can be attributed to the systematic that was altered in that run, but it is a very inefficient way to extract this information. In fact the mean effects of each systematic can be estimated in a 2^k factorial design with considerable savings.

Table 1 gives the settings for a 2^4 factorial experiment; usually the order of the runs would be randomized, but the structure of the experiment is easier to see in the un-randomized form. The run called '1', for example, has all four factors set to their low level, whereas the run called '2', has factors A, B and C set to their low level and D set to its high level. Note that the estimated effect in going from the low level of A, say, to the high level of A, is based on comparing the averages of 8 observations taken at the low level with 8 observations taken at the high level. Each of these averages has a variance equal to $1/8$ the variance in a single observation, or in other words to get the same information from an OFAT design we would need 8 runs with A at 1σ and 8 runs with A at $+1\sigma$, all other factors held constant. Repeating this for each factor would require 64 runs, instead of 16. The balance of the 2^4 design ensures that we can estimate the effects for each of the four factors in turn from the average of 8 observations at

Table 2: The 2^4 factorial showing all of the interaction effects.

run	**A**	**B**	**C**	**D**	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
1	-1	-1	-1	-1	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	-1	-1	-1	-1	$+1$
2	-1	-1	-1	$+1$	$+1$	$+1$	-1	$+1$	-1	-1	-1	$+1$	$+1$	$+1$	-1
3	-1	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	$+1$	-1
4	-1	-1	$+1$	$+1$	$+1$	-1	-1	-1	-1	$+1$	$+1$	$+1$	-1	-1	$+1$
5	-1	$+1$	-1	-1	-1	$+1$	$+1$	-1	-1	$+1$	$+1$	$+1$	-1	$+1$	-1
6	-1	$+1$	-1	$+1$	-1	$+1$	-1	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$
7	-1	$+1$	$+1$	-1	-1	-1	$+1$	$+1$	-1	-1	-1	$+1$	$+1$	-1	$+1$
8	-1	$+1$	$+1$	$+1$	-1	-1	-1	$+1$	$+1$	$+1$	-1	-1	-1	$+1$	-1
9	$+1$	-1	-1	-1	-1	-1	-1	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	-1	-1
10	$+1$	-1	-1	$+1$	-1	-1	$+1$	$+1$	-1	-1	$+1$	-1	-1	$+1$	$+1$
11	$+1$	-1	$+1$	-1	-1	$+1$	-1	-1	$+1$	-1	-1	$+1$	-1	$+1$	$+1$
12	$+1$	-1	$+1$	$+1$	-1	$+1$	$+1$	-1	-1	$+1$	-1	-1	$+1$	-1	-1
13	$+1$	$+1$	-1	-1	$+1$	-1	-1	-1	-1	$+1$	-1	-1	$+1$	$+1$	$+1$
14	$+1$	$+1$	-1	$+1$	$+1$	-1	$+1$	-1	$+1$	-1	-1	$+1$	-1	-1	-1
15	$+1$	$+1$	$+1$	-1	$+1$	$+1$	-1	$+1$	-1	-1	$+1$	-1	-1	-1	-1
16	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$

the high level compared to 8 observations at the low level: for example the main effect of D is estimated by

$$(y_{abcd} - y_{abc} + y_{abd} - y_{ab} + y_{acd} - y_{ac} + y_{ad} - y_a + y_{bcd} - y_{bc} + y_{bd} - y_b + y_{cd} - y_c + y_d - y_{(1)})/8,$$

and similar estimates can be constructed for the effects of B and C.

Note that by constructing these four estimates we have used four linear combinations of our 16 observations. One linear combination, the simple average, is needed to set the overall level of response, leaving 11 linear combinations not yet used to estimate anything. These combinations are in fact used to estimate the *interactions* of various factors, and the full set of combinations is given by the set of signs in Table 2.

For example, the interaction of factors A and B is estimated by the contrast given by the fourth column of table 2:

$$\{y_{abcd} + y_{abc} + y_{abd} + y_{ab} - y_{bcd} - y_{bc} - y_{bd} - y_b - (y_{acd} + y_{ac} + y_{ad} + y_a - y_{cd} - y_c - y_d - y_{(1)})\}/8$$

which takes the difference of the difference between responses with A at high level and A at low level with the difference between responses with B at high level and B at low level. The column of signs in Table 2 for the interaction effect AB was obtained simply by multiplying the A column by the B column, and all the other columns are similarly constructed.

This illustrates two advantages of designed experiments: the analysis is very simple, based on linear contrasts of observations, and as well as efficiently estimating average effects of each factor, it is possible to estimate interaction effects with the same precision. Interaction effects can never be measured with OFAT designs, because two or more factors are never changed simultaneously.

The analysis, by focussing on averages, implicitly assumes that the responses are best compared by their mean and variance, which is typical of observations that follow a Gaussian distribution. However the models can be extended to more general settings, as will be briefly discussed in the next section.

Table 3: A screening design for 7 factors in 8 runs, built from a 2^3 factorial design.

run	A	B	C	D	E	F	G
1	-1	-1	-1	$+1$	$+1$	$+1$	-1
2	-1	-1	$+1$	-1	-1	$+1$	$+1$
3	-1	$+1$	-1	-1	$+1$	-1	$+1$
4	-1	$+1$	$+1$	$+1$	-1	-1	-1
5	$+1$	-1	-1	$+1$	-1	-1	$+1$
6	$+1$	-1	$+1$	-1	$+1$	-1	-1
7	$+1$	$+1$	-1	-1	-1	$+1$	-1
8	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$

In most applications the interpretation of 3- and 4- factor interactions would be rather difficult, and in fact these higher order interactions might be expected to be zero. If they are indeed zero, then 5 of the contrasts outlined in Table 2 are estimating zero, and their squares could then be pooled to provide an estimate of the variance of a single observation, with 4 degrees of freedom. This pooling of higher order interactions is often used in settings where the interactions are expected to be small, and no external estimate of variance is available. Sometimes even two-factor interactions are pooled and used to estimate the error.

Alternatively, we could assign new factors to the higher order interactions, leading to the class of fractional factorial designs. For example, we could use introduce a fifth factor, E, to the 2^4 factorial of Table 2, using the signs for the $ABCD$ interaction. That is, in the first run E would be set to its high level, in the second run to its low level, in the third run to its low level, and so on, following the pattern of $+1$ and -1 in the last column of Table 2. The resulting contrast $(y_{(1)} - y_a - y_b + y_{ab} \pm \dots)/8$ is estimating the main effect of factor E (i.e. the difference between responses on the high level of E to the low level of E), but it is also estimating the $ABCD$ interaction: these two effects are completely aliased. The working assumption is that the $ABCD$ interaction is likely to be very small, so any observed effect can be attributed to E. The main effects of A, B, C and D are estimated as before, and we now have information on 5 factors from a 16 run design. However all the main effects are aliased with 4 factor interactions: for example A is aliased with $BCDE$, B with $ACDE$, and so on. Further, all 2 factor interactions are aliased with 3 factor interactions. Again, the working assumption is typically that any observed effect is more likely to be due to a 2 factor interaction than a 3 factor interaction.

This process can be continued; we might for example assign a new factor F, say, to the ABC interaction (which is aliased with DE), giving a 2^{6-2} design, sometimes called a $1/4$ fraction of a 2^6. This allows us to assess the main effect of 6 factors in just 16 runs, instead of 64 runs, although now some 2 factor interactions will be aliased with each other.

There are very many variations on this idea; one is the notion of a screening design, in which only main effects can be estimated, and everything else is aliased. The goal is to quickly assess which of the factors is likely to be important, as a step in further experimentation involving these factors and their interactions. Table 3 shows an 8 run screening design for 7 factors. The basic design is the 2^3 factorial in factors A, B and C shown in the first 3 columns; then 4 new factors have been assigned to the columns that would normally correspond to the interactions BC, AC, AB and ABC.

There is a very large literature on fractional factorial designs; a good introduction aimed at physicists is given in [1] and much of this paper draws on those ideas. A detailed but quite accessible introduction is given in [2]. Some advantages of these fractional factorial designs is the ability to screen a large number of factors in a few runs, in settings where many factors are expected to be inactive. More

Table 4: Data and design for a 2^{5-1} factorial.

A	B	C	D	E	response
-1	-1	-1	-1	$+1$	29.17
-1	-1	-1	$+1$	-1	29.39
-1	-1	$+1$	-1	-1	22.13
-1	-1	$+1$	$+1$	$+1$	27.64
-1	$+1$	-1	-1	-1	11.53
-1	$+1$	-1	$+1$	$+1$	16.20
-1	$+1$	$+1$	-1	$+1$	14.99
-1	$+1$	$+1$	$+1$	-1	19.29
$+1$	-1	-1	-1	-1	16.30
$+1$	-1	-1	$+1$	$+1$	22.40
$+1$	-1	$+1$	-1	$+1$	19.42
$+1$	-1	$+1$	$+1$	-1	23.85
$+1$	$+1$	-1	-1	$+1$	6.70
$+1$	$+1$	-1	$+1$	-1	13.17
$+1$	$+1$	$+1$	-1	-1	8.53
$+1$	$+1$	$+1$	$+1$	$+1$	19.04

complete fractional factorials, such as $1/2$ fractions or $1/4$ fractions permit assessing a small number of main effects and two-factor interactions. Often a number of the inactive effects can be pooled to provide an internal estimate of variability.

These designs are more complicated to run than OFAT designs, as several factors settings need to be changed with each run. If some factor levels are difficult to change, for example temperature of an oven, in a manufacturing context, then a full factorial design will not be feasible. In such cases it is often possible to have an 'outer' factorial with the difficult-to-change factors, and an 'inner' factorial of the other factors; the analysis of these *split plot* designs is a little more complex. There is a lot of information in a single run of a factorial design, so if a run is lost, the associated balance is lost along with quite a bit of information. It is often necessary to block runs to ensure homogeneity; for example if all runs cannot be completed in a single day and there is concern about changes in conditions from one day to the next. This is relatively straightforward to implement but the analysis of the results is again a little more complicated.

3 Analysis of the data

Implicit in the discussion above is a linear model with Gaussian error

$$y = Z\beta + \epsilon$$

where y is an $n \times 1$ vector of responses, and Z is the so-called *design* matrix, with n rows and p, say, columns, and we assume ϵ follows a Gaussian distribution with mean 0 and covariance matrix σ^2 times the identity. This is exactly a linear regression formulation, but the design matrix Z has a particularly simple form. The first column is a column of $+1$, and the remaining columns have elements ± 1 according to the factorial structure. Table 2 gives an example: in a single run of a 2^4 factorial, y will have length 16, and the 16×16 matrix Z has columns 2 through 16 given by the columns of this table. Any standard regression package will fit this model, although there will be no degrees of freedom available to estimate the error. By specifying a simpler model with just main effects and 2-factor interactions, so that Z now

has dimension 16×11, we will have 5 degrees of freedom left to estimate σ^2. The design matrix Z is completely orthogonal, which makes the least squares fitting of the model particularly simple; in fact it can be computed by hand, and an early algorithm to do this computation invented by Yates is a pre-cursor to the fast Fourier transform.

Most statistical software can deduce from the specification of the model which effects are aliased, and in some packages, including `R` and `Splus` it is relatively easy to produce a graphical display of the estimated effects that allows one to assess which effects are non-zero, at least in part so that the 'nearly zero' effects can be pooled to estimate the error.

An example of the standard linear analysis for a 2^{5-1} factorial, carried out in `R` is given in Figure 1. Figure 2 gives the display of estimated effects described above. It is conventional to ignore the sign of the effects, so the ordered (absolute) values are then plotted against the expected values of ordered (absolute) standard Gaussian variables. The data and design are given in Table 4, following the 2^4 design of Table 1, but assigning factor E to the 4-factor interaction.

If the response is non-Gaussian, then the model will normally assume that some transformation of the mean of the response follows a linear structure of the form $Z\beta$; these models are often called generalized linear models in the statistical literature. For example if y follows a Poisson distribution with mean μ, we might assume $\log \mu = Z\beta$, and fit the model by maximum likelihood. If y is a proportion, then a version of logistic regression is often used, assuming that $\log\{p/(1-p)\} = Z\beta$, where p is the mean value of y. These models can be fit using the `glm` command of `R`. An example of a fractional factorial fit to Poisson data is given in [3], §5.4. An alternative is to transform the responses to something approximately Gaussian, and use the linear model formulation above. If the response is more complex, such as a histogram, then analysis might proceed by constructing one or more derived responses, at least as a first step.

4 Response surface designs

Very often, especially in manufacturing settings, the factors correspond to underlying quantitative variables, and the levels, denoted ± 1 in the previous section, are codes for particular numerical values: temperatures at 80 and 100 degrees, for example. In such cases the choice of factor levels involves both subject matter expertise, and at least in the early stages, considerable guesswork. As well, the goal of the experiment might not be to compare responses at different factor settings, but to find the combination of factor settings that leads to minimum or maximum response.

Factorial designs adapted to these conditions are called response surface designs. The basic idea is that the response y is a continuous function of some input variables x_1, x_2, and so on, and factorial designs are used sequentially to explore the shape of the response surface. Sequential experimentation in the relevant range of x-space usually begins with a screening design, to quickly assess which of several factors have the largest effect on the response. Then second stage is a factorial design at new values of the underlying variables, chosen in the direction of increasing (or decreasing) response. Near the maximum additional factor levels are added, to model curvature in the response surface. In the setting of simulation experiments, the goal might be to see which values of the systematics produce simulated data consistent with the observations; thus we would be seeking to minimize a derived response measuring discrepancy of the simulation with the data would be. Another goal might be simply to see which systematic parameters affect the simulation output, and whether they affect it linearly or in a more complex fashion.

Figure 3 is adapted from [4], although similar pictures can be found in, for example [2], and other treatments of response surface methods, such as [5]. The contours of a smooth response surface in two quantitative variables are indicated, along with an initial 2^2 factorial design. The $+$ symbols indicate the design points for the first experiment, and the results could lead to a second 2^2 factorial carried out at the circles indicated. The results would show that the maximum is 'surrounded', so to speak, at which

```
>  y <- c(29.17,29.39,22.13,27.64,11.53,16.20,14.99,19.29,16.30,
> +   22.40,19.42,28.85,6.70,13.17,8.53,19.04)
> A <- c(rep(-1,8),rep(1,8))
> B <- c(rep(-1,4),rep(1,4),rep(-1,4),rep(1,4))
> C <- rep(c(rep(-1,2),rep(1,2)),4)
> D <- rep(c(-1,1),8)
> E <- A*B*C*D
> A <- factor(A); B <- factor(B); C <- factor(C); D <- factor(D); E <- factor(E)
> # The factor() tells R to interpret the levels of A as qualitative instead of quantitative.
> # This enables use of abbreviated notation for the  model, as shown below.

> fact.lm <- lm(y~A*B*C*D*E) # lm is the general linear model fitting routine
# we avoided specifying the full matrix Z, although this could have been done instead

> fact.lm                 # there are many ways to summarize the output
Call:
lm(formula = y ~ A * B * C * D * E)

Coefficients:
   (Intercept)              A1              B1              C1
        25.686          -9.386         -14.156          -3.556
            D1              E1           A1:B1           A1:C1
         3.704           3.484           1.697           4.878
         B1:C1           A1:D1           B1:D1           C1:D1
         3.367           4.453           1.173           3.073
         A1:E1           B1:E1           C1:E1           D1:E1
        -1.238           0.612          -0.448          -4.303
      A1:B1:C1        A1:B1:D1        A1:C1:D1        B1:C1:D1
            NA              NA              NA              NA
  # ...[other NAs deleted]

> coef(fact.lm)[2:16]
     A1      B1      C1      D1      E1   A1:B1   A1:C1   B1:C1   A1:D1
 -9.386 -14.156  -3.556   3.704   3.484   1.697   4.878   3.367   4.453
  B1:D1   C1:D1   A1:E1   B1:E1   C1:E1   D1:E1
  1.173   3.073  -1.238   0.612  -0.448  -4.303

> library(faraway) # a package named "faraway" gives easy access to half normal plots
> halfnorm(coef(fact.lm)[2:16],labs=fact.names,main="Half-normal plot for identifying
 effects", ylab="effect estimates")

> # A simpler model with just factors A, B and C and their interactions
> anova(lm(y~A+B+C+A:B+A:C+B:C+A:B:C))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
A          1     81      81    3.70 0.0907 .
B          1    461     461   21.11 0.0018 **
C          1     14      14    0.65 0.4444
A:B        1      3       3    0.13 0.7257
A:C        1     24      24    1.09 0.3270
B:C        1     11      11    0.52 0.4915
A:B:C      1     19      19    0.85 0.3840
Residuals  8    175      22
```

Fig. 1: Some R code illustrating analysis of a factorial design

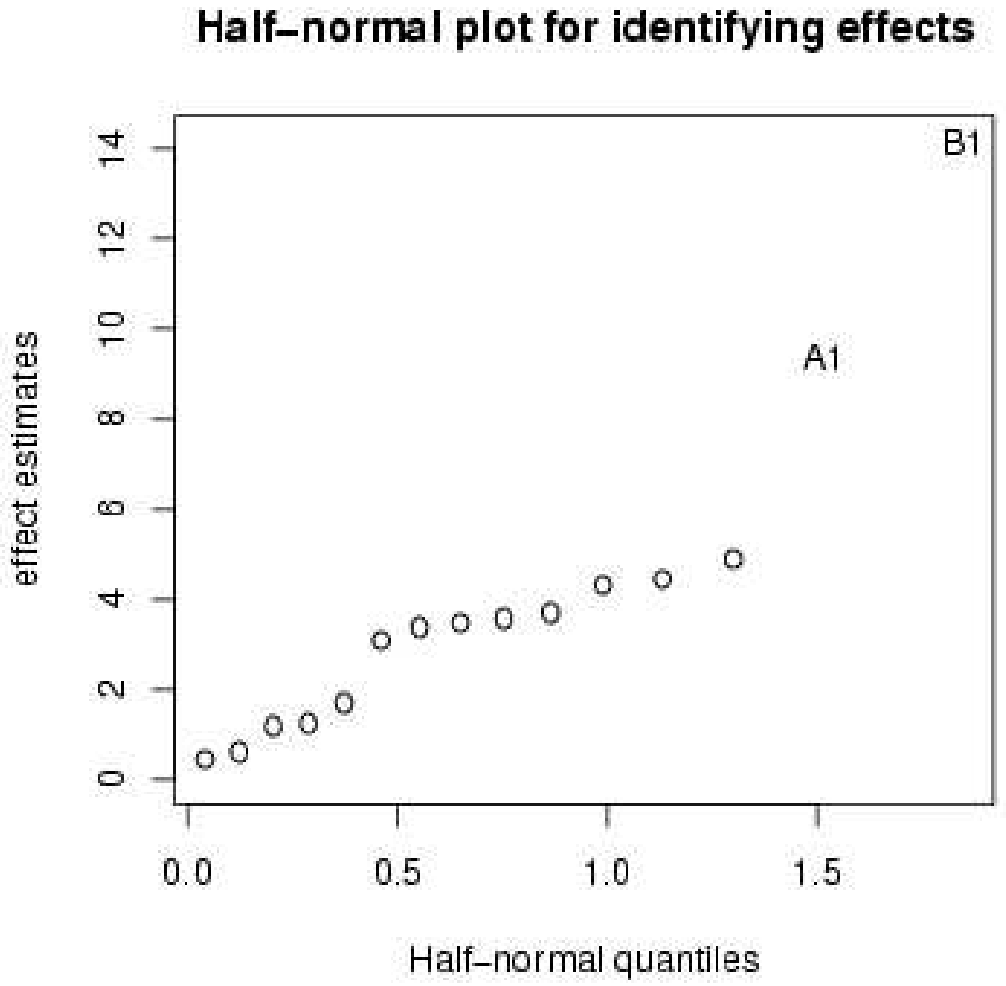

Fig. 2: A graphical display of the estimated effects for the data from Table 4; the two largest effects are labelled.

stage it would be usual to add center points and radial points to attempt to quantify the curvature at the maximum. Note that with an OFAT design, the sequential stages of experimentation could only proceed along lines parallel to the coordinate axes, which is less efficient unless the axes of the elliptical contours are aligned with the coordinate axes.

A two-level factorial design can only detect linear effects of x_1 and x_2, and their interaction, x_1x_2. The other quadratic effects, x_1^2 and x_2^2 need a minimum of three levels to be estimated. A very common approach to estimating a smooth, curved, response surface is to add center points at $(0, 0)$, often replicated, to give an internal estimate of error, and then to add further points on the radius of a circle. Such designs are called *central composite* designs. This is illustrated for two factors in Figure 4, but the idea is very general.

5 More specialized designs

The 8 run screening design illustrated in Table 3 is a 2^{7-4} fractional factorial, but is also an example of an *orthogonal array*, which is by definition an array of symbols, in this case ± 1, in which every symbol appears an equal number of times in each column, and any pair of symbols appears an equal number of times in any pair of columns. An orthogonal array of size $n \times (n-1)$ with two symbols in each column

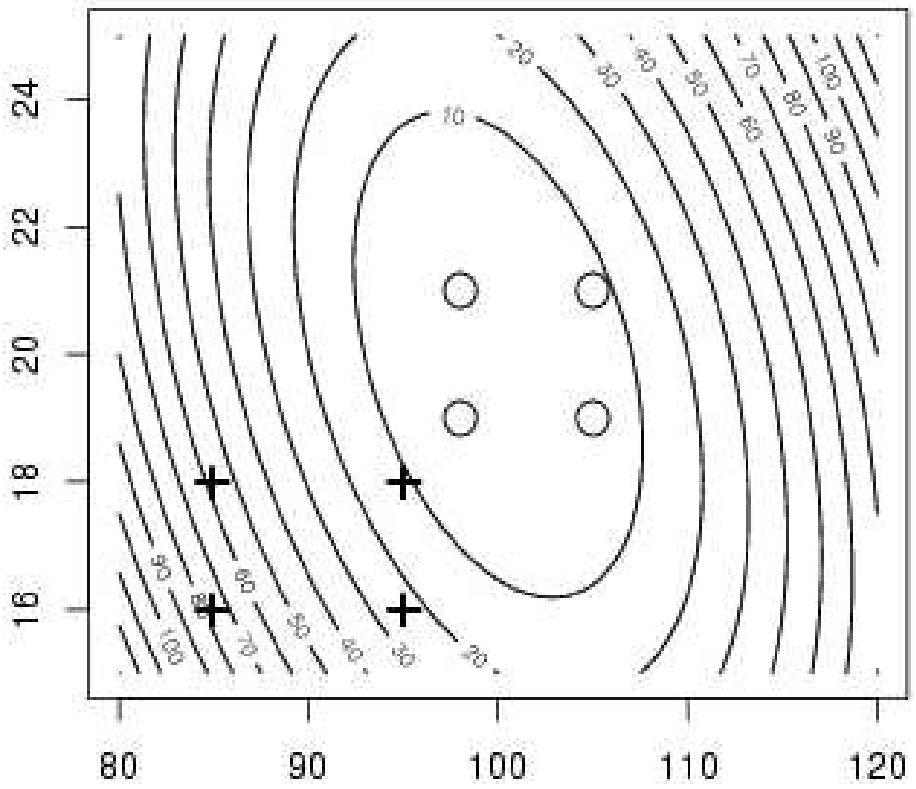

Fig. 3: Two 2^2 experiments to explore a response surface: + shows the design points for the first experiment, and o shows the design points for the second.

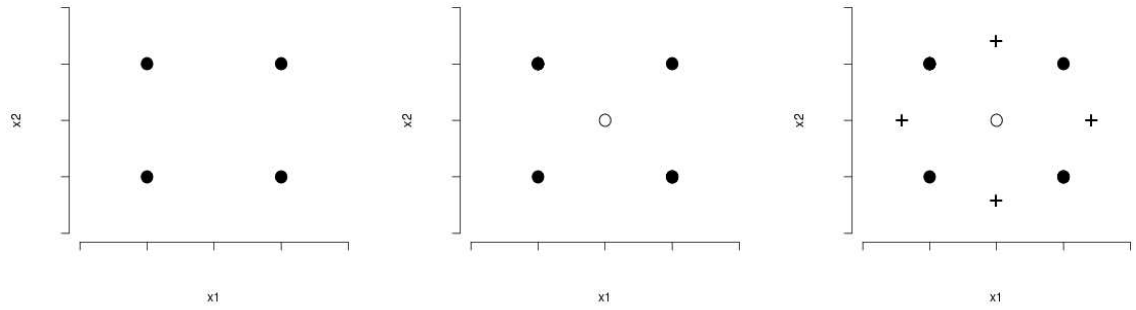

Fig. 4: A series of experiments adding points to capture (to first order) the nonlinearity in the response surface; the third is a central composite design.

Table 5: An orthogonal array for 6 factors each at 3 levels

run	A	B	C	D	E	F
1	-1	-1	-1	-1	-1	-1
2	-1	0	0	0	0	0
3	-1	$+1$	$+1$	$+1$	$+1$	$+1$
4	0	-1	-1	0	0	$+1$
5	0	0	0	$+1$	$+1$	-1
6	0	$+1$	$+1$	-1	-1	0
7	$+1$	-1	0	-1	$+1$	0
8	$+1$	$+1$	-1	$+1$	0	-1
9	$+1$	$+1$	-1	$+1$	0	$+1$
10	-1	-1	$+1$	$+1$	0	0
11	-1	0	-1	-1	$+1$	$+1$
12	-1	$+1$	0	0	-1	-1
13	0	-1	0	$+1$	-1	$+1$
14	0	0	$+1$	-1	0	-1
15	0	$+1$	-1	0	$+1$	0
16	$+1$	-1	$+1$	0	$+1$	-1
17	$+1$	0	-1	$+1$	-1	0
18	$+1$	$+1$	0	-1	0	$+1$

specifies an n-run screening design for $n-1$ factors. The designs with symbols ± 1 are called Plackett-Burman designs and Hadamard matrices defining them have been shown to exist for all multiples of four up to 424. A Plackett-Burman design is used for studying simulations in [9].

More generally, an $n \times k$ array with m_i symbols in the ith column is an *orthogonal array* of strength r if all possible combinations of symbols appear equally often in any r columns. The symbols correspond to levels of a factor. Table 5 gives an orthogonal array of 18 runs, for 6 factors with three levels each.

Orthogonal arrays are particularly popular in applications of so-called Taguchi methods in technological experiments and manufacturing. An extensive discussion is given in [6]. In the discussion of the talk at CERN, Jim Linneman pointed out that these applications could be relevant to HEP experiments at the stage at which the equipment is being designed, tested and manufactured.

A related approach to the exploration of possibly complex response surfaces is the use of *space filling* designs. These are especially popular in computer experiments and simulation experiments, as well as in numerical integration, where the method is known as quasi Monte Carlo. In the approximation of a multi-dimensional integral

$$\int_{R^k} f(x)dx \approx \frac{1}{n}\sum f(X_i)$$

the X_i are taken to be 'space-filling' points, whereas in simple Monte Carlo the X_i would be sampled randomly, possibly using a uniform distribution on each coordinate. The difficulty with the simple approach is that in high dimensions the compounding of the uniform points tends to concentrate on the outer shell, and interior points are too rarely sampled. Orthogonal arrays can be used as the basis of space-filling designs, and in this application are often called Latin hypercube designs. A good reference is [7]; however in the discussion of the talk at CERN Fred James said that in fact he had little success with space-filling designs in high dimensional integrations.

6 An example motivated by miniBoone

In this section I report on some preliminary work by Zi Jin, Radford Neal and myself. This does not in fact use the orthogonal constructions described in the preceding sections, but is some preliminary work to see if these methods could provide improvement in simulation experiments. The basic ideas were described in the context of the miniBoone experiments by Byron Roe at the Banff workshop in summer 2006.

Suppose that a simulation run generates M background events and mistakenly identifies y of these a signal, with some small probability p, say $p = 0.001$ or $p = 0.0001$. We use as a first approximation the Poisson model for y with mean p. This mean is assumed to depend on various settings for the systematics, presumably in a fairly complex way that could be explored using factorial designs and other concepts from the previous sections.

We bypass that by making the very rough approximation that p depends on these systematics via a Gamma distribution with parameters a and b. The mean and variance of the gamma distribution are a/b and a/b^2, respectively, so a and b would be chosen to allow p to vary between, say $\pm 3\sigma$ about its mean. We then explore how the Fisher information in a full set of N simulation runs depends on the trade-off between sampling K different values of these systematics or sampling M events at each value of the systematics, under the constraint $N = MK$.

For example, suppose p is approximately 0.0001. If we fix the total number of simulations N at $2,000,000$, then the optimal split is roughly $M = 160,000$ runs at each of $K = 13$ different parameter settings. This represents a 10-fold increase in precision (inverse variance) over the rather arbitrary choice of $M = 100,000$ and $K = 20$. The improvement indicated by these preliminary results suggests that it will be worthwhile to investigate the information in the more complex orthogonal array based designs. A more complete account of these results will be presented elsewhere.

7 A short guide to the references

Much of this paper was inspired by two articles by Gunter, [1] and [4], which are written for scientists. The book by Box, Hunter and Hunter, [2] has very detailed explanation of the sequential nature of experimentation in industrial applications, and the book by Montgomery [5] has a more concise discussion, in something of a cookbook style. A somewhat more theoretical approach is given in [3]. Ref. [6] has a lot of material on orthogonal arrays, but their use in numerical integration is probably best approached from [7]. Ref [8] is an introduction to the general field of computer experiments, but [9] describes applications that are probably closer to those needed for HEP simulations.

References

[1] B.H. Gunter, *Computers in Physics* **7**, 262–272.

[2] Box, G.E.P., Hunter, J.S. and Hunter, W. G. (1978) *Statistics for Experimenters.* Wiley & Sons, New York.

[3] Cox, D.R. and Reid, N. (2000) *The Theory of Design of Experiments.* Chapman & Hall/CRC, London.

[4] Gunter, B.H. (2007) "Sequential Experimentation", to appear in *Encyclopedia of Statistics in Quality and Reliability*, Wiley & Sons, New York.

[5] Montgomery, D. C. (2005). *Design and Analysis of Experiments (6th ed.)* Wiley & Sons, New York.

[6] Wu, C.F.J. and Hamada, M. *Experiments: Planning, Analysis, and Parameter Design Optimization.* Wiley & Sons, New York.

[7] Owen, A. (1992). Orthogonal arrays for computer experiments, integration and visualisation. *Statistica Sinica* **2**, 459–462.

[8] Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989). Design and analysis of computer experiments. *Statistical Science* **4** 409–423.

[9] Welsh, J.P., Koenig, G.G. and Bruce, D. (1997). Screening design for model sensitivity studies. *J. Geophys. Res.* **102** D14, 16,499–16,505.

Computing Likelihood Functions for High-Energy Physics Experiments when Distributions are Defined by Simulators with Nuisance Parameters

Radford M. Neal
Dept. of Statistics, University of Toronto

Abstract

When searching for new phenomena in high-energy physics, statistical analysis is complicated by the presence of nuisance parameters, representing uncertainty in the physics of interactions or in detector properties. Another complication, even with no nuisance parameters, is that the probability distributions of the models are specified only by simulation programs, with no way of evaluating their probability density functions. I advocate expressing the result of an experiment by means of the likelihood function, rather than by frequentist confidence intervals or p-values. A likelihood function for this problem is difficult to obtain, however, for both of the reasons given above. I discuss ways of circumventing these problems by reducing dimensionality using a classifier and employing simulations with multiple values for the nuisance parameters.

1 The Problem

I will discuss a class of problems that I hope at least resemble those encountered in high-energy physics experiments, such as searches for the Higgs Boson with the LHC. The solutions that I examine will have much in common with some present practice, though I will not attempt to provide comprehensive references. I hope that my discussion will clarify the role of existing techniques, such as the training of classifiers for 'signal' vesus 'background' events, and also point to possible new approaches.

In this paper, I deal with experiments where we will observe O events, indexed by $i = 1, \ldots, O$, that are described by variables, v_i, computed from the raw observational data. Events can either be from the 'background' or (if it exists) from the 'signal' — for instance, an event in which a previously-unobserved particle appears. I assume that simulation programs for background and signal events exist, which stochastically generate the variables from either a background distribution, which has probability density function $p_0(v)$, or a signal distribution, which has probability density function $p_1(v)$. The real events come from a *mixture* of signal and background distributions, with an unknown proportion, f, of signal. We may be most interested in whether or not f is zero — since $f > 0$ may, for instance, correspond to the existence of a previously-unknown particle.

Our first difficulty is that no explicit formulas for $p_0(v)$ and $p_1(v)$ exist. We may 'know' p_0 and p_1 in some sense, or we couldn't have written the simulator programs, but we have no way of translating this knowledge into a practical method for computing these density functions.

Our second difficulty is that, typically, we don't actually know p_0 and p_1 exactly. The simulators for generating from these distributions have some parameters — relating either to the physics or to the behaviour of the detector — whose values are not known precisely. Call these parameters ϕ. I'll assume that although ϕ is not known, we have a suitable prior distribution for ϕ, with density $p(\phi)$. Note that ϕ is a 'nuisance' parameter, since our only real interest is in f. The fact that ϕ is unknown is just an annoyance. (Though ϕ might be of interest to other people, such as the designers of the detector.)

2 The Role of the Likelihood Function

The *likelihood function* is the probability (or probability density) of the observed data, seen as a function of the model parameter(s). The likelihood function is defined only up to an arbitrary constant factor, and hence only ratios of likelihoods for different values of the parameters are meaningful.

When there are no nuisance parameters, the likelihood function for our problem (assuming independent observations) is a function of f alone:

$$L(f) \;=\; \prod_{i=1}^{O} \Big[f p_1(v_i) \;+\; (1-f) p_0(v_i) \Big] \tag{1}$$

Here, $f p_1(v_i) \;+\; (1-f) p_0(v_i)$ is simply the probability density for obtaining the observation v_i from either the signal distribution (with probability f) or the background distribution (with probability $1-f$). When there are nuisance parameters, ϕ, the likelihood is a function of both f and ϕ. I defer consideration of nuisance parameters to Section 4.

According to the *likelihood principle* (see, for example, the discussion by Cox and Hinkley [1], Section 2.3), the likelihood function contains all the information from the experiment that is relevant to inference for the parameters. So inference should not depend on aspects of the data that do not enter into the likelihood function. (An exception is that checks of the appropriateness of the model on which the likelihood function is based may utilize other aspects of the data.) Note that the likelihood function is itself a function of the data, and sometimes (not always!) depends only on some low-dimensional statistic computed from the data, such as the sample mean and/or the sample variance. A quantity computed from the data that can be used to compute the likelihood function is known as a 'sufficient statistic'. The likelihood function itself is a 'minimal sufficient statistic', containing no irrelevant information.

This 'weak' form of the likelihood principle is accepted by most statisticians. The 'strong' form, which is not universally accepted, says that the same conclusions should be drawn from two experiments (involving the same parameters) if they produced the same likelihood function. Bayesian inference obeys the strong likelihood principle, since it simply combines the likelihood with a prior distribution (which presumably does not vary with the choice of experiment). The strong likelihood principle is accepted by some non-Bayesian statisticians as well, however, partly because it follows from the weak likelihood principle together with a form of the principle that one should condition on an ancillary statistic (whose distribution does not depend on the parameters).

Classical (ie, non-Bayesian, frequentist) confidence intervals and p-values (other than those used for model checking) often violate the likelihood principle. For example, consider observations of n independent binary events, of which k turned out to be 1, with the remaining $n-k$ being 0. If we are interested in inferring the probability, θ, that an event is 1 (assumed the same for all events), the likelihood function will be $L(\theta) = \theta^k (1-\theta)^{n-k}$. In particular, if $n = 10$ and $k = 1$, the likelihood function is $L(\theta) = \theta(1-\theta)^9$. This likelihood function is the same regardless of whether we had decided to observe $n = 10$ events and found that $k = 1$ of them were 1, or we had decided to observe events until $k = 1$ of them were 1, and found that this was reached when $n = 10$. Hence, according to the likelihood principle, our conclusions should be the same in these two scenarios. However, the one-sided p-value for testing the null hypothesis that $\theta = 1/2$ versus the alternative that $\theta < 1/2$ is $P(k \leq 1) = (10+1)\,2^{-10}$ when the number of events is fixed at $n = 10$, but is $P(n \geq 10) = 2^{-9}$ when the number of 1 events is fixed at $k = 1$. Of the many arguments why such differing results should not be accepted, I will mention only consideration of an observer who knows everything that the experimenter does and sees, but doesn't know the experimenter's thoughts. Does this observer really need to ask the experimenter whether the stopping condition was $n = 10$ or $k = 1$ in order to draw an inference from the data? And would inference really be impossible if the experimenter had forgotten?

This issue arises also with Feldman and Cousins' [2] method for constructing confidence intervals from data, n, that is a sum of Poisson-distributed counts of 'signal' events (with unknown mean, μ) and 'background' events (with known mean, b). (We will see in Section 3 that this problem can arise as a much-reduced form of the problem discussed in this paper.) Their method (as well as some others) produces different confidence intervals for μ from an observed count of zero depending on the mean number of background events — the interval is tighter (with smaller upper limit) when the mean number of background events is higher. This violates the strong likelihood principle, since the likelihood functions for a

count of zero from experiments with different background means differ only by a constant factor (which is irrelevant for likelihoods):

$$L(\mu) \;=\; \exp(-(b+\mu)) \;=\; \exp(-b)\,\exp(-\mu) \;\propto\; \exp(-\mu) \tag{2}$$

It makes intuitive sense that the inference drawn when the count is zero should not depend on the mean background — with a count of zero, we *know* that no background events occurred, so how many would occur on average if we were to repeat the experiment is of no relevance.

Feldman and Cousins are aware of this issue, but in responding to it, appear to have lost track of the scientific purpose of a statistical analysis, as is not uncommon in such discussions. They say that "for making decisions, [Bayesian inference] is probably how many scientists do (and should) think", but that "[classical] confidence intervals provide the preferred option for publishing numerical results of an experiment in an objective way. However it is critical not to interpret them as Bayesian intervals, i.e., as statements about $P(\mu_t|x_0)$". They remark with respect to the dependence of their intervals on the expected background when the observed count is zero that "We find that objections to this behaviour are typically based on a misplaced Bayesian interpretation of classical intervals, namely the attempt to interpret them as statements about $P(\mu_t|n_0)$." In further discussion of this situation, and in particular their method's production of confidence intervals for a count of zero that are tighter when the experiment is more poorly designed (with higher mean background), they say "The origin of these concerns lies in the natural tendency to want to interpret these results as the probability ... of a hypothesis given data rather than what they really are related to... It is the former that a scientist may want to know in order to make a decision, but the latter which classical confidence intervals relate to." In their discussion, they say nothing about what actual scientific use their classical confidence intervals might have, leaving (at least to me) the impression that they believe classical confidence intervals should be computed and reported simply as a ritual activity.

Fortunately, Feldman and Cousins do say that "it is important to publish relevant ingredients to the calculation so that the reader... can (at least approximately) perform alternative calculations or combine the result with other experiments". The "relevant ingredient" is in fact the likelihood function. Nothing more is needed, and nothing less than the full likelihood function would allow (for example) any Bayesian with any prior to make inferences.

When there is only a single parameter, such as f, the result of an experiment can easily be communicated fully by a plot of $L(f)$ versus f. In general, such a plot contains more information than a classical confidence interval, or any other interval that attempts to summarize the result. However, in many situations, the likelihood function approaches the exponential of a quadratic function as the amount of data increases (see [1], Section 10.6), and for simple Gaussian models, the log likelihood may be a quadratic function even for small samples. In such situations it is possible to specify the likelihood function using only two numbers (recall that the likelihood is defined only up to a constant factor). The end-points of a classical confidence interval can sometimes serve this purpose. If the parameter space is also unbounded, it is possible to develop intuitions about the meaning of classical confidence intervals that reflect what really matters — the likelihood function — explaining (in my view at least) how their use has survived.

However, when the log likelihood is not approximately quadratic, or when the parameter space is bounded, as in the example above where $L(\mu) \propto \exp(-\mu)$, with $\mu \in (0, \infty)$, applying these intuitions about confidence intervals, developed in another contexts, is dangerous. One simply cannot represent the various forms that a likelihood function can in general take using only two numbers. Obtaining the full likelihood function, or a good approximation to it, is therefore a crucial objective of statistical inference.

3 First Difficulty: We Can't Compute the Likelihood

Consider again our model with no nuisance parameters, with likelihood function given by equation (1). For a model like this with only a single scalar parameter, the full result of the experiment can easily be

communicated by simply plotting the likelihood function. In typical problems, one can also easily find the maximum likelihood parameter estimate, as well as various Bayesian inferences, such as the posterior density obtained when some prior distribution is assumed.

But for our problem, we don't know how to compute the likelihood! So we can't easily produce a plot of $L(f)$ versus f. The likelihood involves p_0 and p_1, which are known only through simulation programs. If the v_i are low-dimensional (not more than around four dimensional), we could generate many points from p_0 and p_1, and use them to get good estimates for these density functions, but for high-energy physics experiments, it seems that it is more typical for each event to be described by dozens or hundreds of values. It might be possible to compute the p_0 and p_1 densities by using techniques similar to those used to compute free energies from Monte Carlo simulations, but these techniques would likely be too slow for this application, since the number (O) of observations for which these densities would need to be computed is typically quite large.

Fortunately, we only really need to compute the ratio $p_1(v_i)/p_0(v_i)$ for each observation. Since constant factors in the likelihood can be ignored, we can reduce the likelihood as follows:

$$
\begin{aligned}
L(f) &= \prod_{i=1}^{O} \Big[f p_1(v_i) \;+\; (1-f) p_0(v_i) \Big] && (3)\\
&= \prod_{i=1}^{O} p_0(v_i) \Big[f \, \frac{p_1(v_i)}{p_0(v_i)} \;+\; (1-f) \Big] && (4)\\
&= \Big[\prod_{i=1}^{O} p_0(v_i) \Big] \cdot \prod_{i=1}^{O} \Big[f \, \frac{p_1(v_i)}{p_0(v_i)} \;+\; (1-f) \Big] && (5)\\
&\propto \prod_{i=1}^{O} \Big[f \, \frac{p_1(v_i)}{p_0(v_i)} \;+\; (1-f) \Big] && (6)
\end{aligned}
$$

Here, we can ignore the product of the $p_0(v_i)$ since factors in the likelihood not depending on f can be ignored. So we can look for a way to compute $p_1(v)/p_0(v)$ without having to compute $p_0(v)$ and $p_1(v)$.

One way to compute $p_1(v)/p_0(v)$ is to produce a classifier to distinguish signal and background events, training it on many simulated signal and background events, drawn according to p_1 and p_0. This is commonly done, as illustrated, for example, by [3]. The classifier could be based on neural networks, decisions trees, or many other methods, though I will assume here that the classifier can produces probabilities for the two classes, not just a guess at the class, with no indication of how likely it is to be correct. Suppose that the fraction of simulated events used to train such a classifier that are from the signal distribution is s. If we manage to train an excellent classifier, the probability it outputs that an event described by variables v is a signal event (call this $c(v)$) will match the true probability that such a simulated event is signal, so that

$$
c(v) \;=\; \frac{s \, p_1(v)}{(1-s) \, p_0(v) \;+\; s \, p_1(v)} \qquad (7)
$$

Once we have this classifier, we can find the desired ratios as follows:

$$
\frac{p_1(v)}{p_0(v)} \;=\; \frac{c(v)}{1-c(v)} \, \frac{1-s}{s} \qquad (8)
$$

If we really trust our classifier, we can now compute the likelihood function for f, and present a plot of $L(f)$ as the result of the experiment.

If we don't totally trust our classifier, we can still use it to get good results. We just treat it as a way of reducing the dimensionality of the data — from the multidimensional measurements, v_i, to the

scalar $r_i = c(v_i)$ produced using the classifier. If the classifier were perfect, this reduction would not lose any useful information (since the r_i determine the likelihood function). If it's not perfect, it will throw away a bit of information, but the reduction to a scalar allows us to easily estimate $p_0(r)$ and $p_1(r)$ from simulation data, and use them to compute a likelihood function given the r_i. The results will be valid (ie, not systematically misleading), since this likelihood captures what can be learned from the experiment if one insists on reducing dimensionality in this way. However, the results may not be as precise (ie, as informative) as would have been obtained using a perfect classifier, which would have produced the true likelihood given all the information in the data.

One could reduce the data further by binning the r_i values, but this loses information. Using a fairly large number of bins might be OK, however, if it loses little information, and makes estimating the probabilities easier. If only two bins are used (ie, the output of the classifier is thresholded at some value), we get a Poisson count with background problem, of the sort discussed in Section 2 — assuming there are many events and signal events are rare, the number of events in the "signal" bin will be Poisson distributed, with some being real signal events and some being mis-classified background events. But such a drastic reduction of the data might throw away quite a bit of relevant information.

4 Second Difficulty: Nuisance Parameters for the Physics and the Detector Behaviour

In practice, we probably don't know p_0 and p_1 exactly. The simulators for generating from these distributions will have some parameters, ϕ, relating either to the physics or to the behaviour of the detector, which are not known precisely. (As a convenience, we can assume that ϕ is the same for simulating p_0 and p_1, since we can let some components of ϕ be used by only one of these simulators.)

We have to assume that these ϕ parameters are known to some degree, or there's no hope of solving the problem. I'll assume that based on theory or previous experiments, a prior distribution for ϕ is available, with density $p(\phi)$. It's unlikely that this prior will be perfect — eg, it might assume independence of components of ϕ when it really ought not to. We must hope that the results are not too sensitive to this — formally checking whether this is true is a difficult problem.

Once there are ϕ parameters, the likelihood is a function of both f and ϕ:

$$L(f,\phi) \;=\; \prod_{i=1}^{O} \Big[f p_1(v_i|\phi) \;+\; (1-f) p_0(v|\phi) \Big] \tag{9}$$

where $p_0(v|\phi)$ and $p_1(v|\phi)$ denote probability densities for generating v from the background and signal simulators with parameters set to ϕ.

This is a high dimensional function (since ϕ is typically high dimensional), and hence will be difficult to visualize. Just plotting $L(f,\phi)$ will *not* be a feasible way of presenting the results of the experiment. Many ways of dealing with this type of problem have been proposed. For instance, we might look at the "profile likelihood", a function of f alone defined as $\sup_\phi L(f,\phi)$. This ignores many aspects of the likelihood function, however. A Bayesian approach is to instead integrate $L(f,\phi)$ with respect to a prior distribution for ϕ, to obtain a *marginal likelihood function* for f alone:

$$\underline{L}(f) \;=\; \int L(f,\phi)\, p(\phi)\, d\phi \tag{10}$$

We often could compute this fairly easily by simple Monte Carlo (sampling from the prior for ϕ), if we could compute $L(f,\phi)$. Since the marginal likelihood is one-dimensional, we would then be able to present the result of the experiment by simplying plotting $\underline{L}(f)$, if we could compute it.

The role of the prior, $p(\phi)$, in producing a marginal likelihood is worth examining. When there are no nuisance parameters, the likelihood function $L(f)$ is an 'objective' presentation of the experimental result (if one ignores subjectivity in the choice of model). Inferences can then be drawn using this

likelihood in various, possibly 'subjective', ways. There is no need for the experimenters to draw such inferences (though they may of course do so if they wish), and hence no need for them to choose a prior for the parameter of interest, f. The situation is different for the nuisance parameters, ϕ, since many components of ϕ will relate to experimental details about which the experimenters are much more knowledgeable than anyone else. It therefore seems most sensible for the experimenters to decide on a suitable prior, $p(\phi)$, and use this to produce a marginal likelihood, $\underline{L}(f)$, that can be interpreted by others.

Unfortunately, actually computing $L(f,\phi)$, and from it $\underline{L}(f)$, is at least as difficult as computing $L(f)$ when there are no nuisance parameters. We might try, as in Section 3, to rewrite the likelihood in terms of ratios of probabilities:

$$L(f,\phi) \;=\; \prod_{i=1}^{O} \Big[f p_1(v_i|\phi) \;+\; (1-f) p_0(v_i|\phi) \Big] \tag{11}$$

$$\;=\; \Big[\prod_{i=1}^{O} p_0(v_i|\phi) \Big] \;\cdot\; \prod_{i=1}^{O} \Big[f \, \frac{p_1(v_i|\phi)}{p_0(v_i|\phi)} \;+\; (1-f) \Big] \tag{12}$$

However, unlike before, the first factor is now relevant, since it depends on the parameter ϕ. Properties of events that are irrelevant for classifying them as signal versus background may still be relevant for inferring ϕ, and hence indirectly for inferring f.

As we did in Section 3, we might try to avoid our difficulties by reducing the dimensionality of the data. If we can map the high-dimensional v_i to quantities r_i that are low-dimensional (and then possibly binned), it will be feasible to estimate $p_0(r_i|\phi)$ and $p_1(r_i|\phi)$ using a reasonable number of events generated by the simulators. We probably can't expect to reduce dimensionality in a way that preserves all relevant information, but we can hope to keep the loss of information small.

We can define a likelihood function based on this reduced data:

$$L_r(f,\phi) \;=\; \prod_{i=1}^{O} \Big[f p_1(r_i|\phi) \;+\; (1-f) p_0(r_i|\phi) \Big] \tag{13}$$

Since we have likely lost information by going from v_i to r_i, this is not the same function as $L(f,\phi)$. But it can be used to make valid (though less efficient) inferences, provided that the mapping from v to r was not chosen based on the observed data. We can again define a marginal likelihood, integrating over the prior for ϕ:

$$\underline{L}_r(f) \;=\; \int L_r(f,\phi)\, p(\phi)\, d\phi \tag{14}$$

If we can compute this, plotting it will display the results of the experiment, as well as possible given our computational limitations, which forced the reduction from v_i to r_i.

To compute $\underline{L}_r(f)$, we could choose K values for ϕ from the prior, labelled $\phi_1, \ldots, \phi_K$, either randomly or by some quasi-Monte Carlo scheme, and then average $L_r(f,\phi_k)$ over these K values to approximate the integral above. Computing $L_r(f,\phi_k)$ will require simulating many events from the background and signal distributions with parameters ϕ_k, and then using these to estimate the probability densities $p_0(r|\phi_k)$ and $p_1(r|\phi_k)$ (or the bin probabilities, if the r_i were binned).

Though not easy, this computation seems to be feasible, provided a value for K in the hundreds or thousands is adequate, and the dimensionality of the r_i is small enough that for each ϕ_k the densities $p_0(r|\phi_k)$ and $p_1(r|\phi_k)$ can be adequately modeled using a few thousand events generated by each of the simulators. (Alternatively, one might try to build one general model for the conditional densities $p_0(r|\phi)$ and $p_1(r|\phi)$ using data generated with all values ϕ_k.) The total number of simulated events required would then be no more than a few tens of millions.

The choice of K is not easy, however. If one is sure that $L_r(f,\phi)$ does not vary drastically with ϕ, a value for K of a few hundred would suffice. However, if drastic variation is conceivable, a much larger value of K might be needed in order to be confident that the results are valid. Suppose, for example, that the actual observations are such that for most ϕ, $L(0,\phi)$ is small (compared to $L(f,\phi)$ for some $f>0$), but that in some region of ϕ values with small but not negligible prior probability, $L(0,\phi)$ is very much larger, sufficiently so that this region dominates the integral defining $\underline{L}(0)$ — or to put it another way, if ϕ is in this region, the experiment will produce many more background events that look like signal events than for other values of ϕ. If none of the K values for ϕ_k that were chosen happen to lie in this region, the value for $\underline{L}(0)$ that is computed will be much smaller than the true value. If this situation is a possibility, one would need to use a value for K that is sufficiently large for the p-value, or other measure of confidence, that one aims to report for a discovery (certainly no smaller than the reciprocal of this p-value, and preferrably somewhat larger), in order to reduce to the required level the chance that such problematic values for ϕ might have been missed.

The most difficult problem is deciding how to reduce dimensionality. Training a classifier to distinguish signal from background still seems to be a useful way of isolating relevant information. This might be done in several ways, however, and we may also wish to preserve other information, relating to the first factor in equation (12).

We could train a classifier using background and signal events generated using a single value for the nuisance parameters (eg, the prior mean), reducing the data from v_i to $r_i = P(\text{signal}|v_i)$, as approximated by this classifier. This is much the same as we would do if there were no nuisance parameters (equivalently, if the correct ϕ were known). The predictions could of course be binned, and with only two bins, we would end up with Poisson-distributed counts in which the mean background is uncertain, but with a distribution that can be estimated from simulations, as described more generally above.

The danger of this approach is that the classifier that is trained may not work well for events generated with other values of ϕ, and in particular, may not work well for the true ϕ. If so, the number of background events misclassified as signal may be large, leading to a substantial loss of information (though not to misleading results, if the rest of the inference task is properly done).

Another simple approach is to generate events with many values of ϕ, drawn from the prior (a different ϕ_i for every v_i), and train a classifier with v_i as inputs on all of this data. The classifier might then learn how to distinguish signal from background in a robust way, that works for all ϕ. Of course, it could well be that there is no way to accurately classify without knowing ϕ, in which case this method will also lose much information.

When neither of these approaches work, it seems that one must rely on the data being informative about ϕ. I sketch here a scheme that may perhaps provide a feasible solution in this situation, based on reducing the dimensionality of both ϕ and v.

As before, we will train a classifier for signal versus background events, using data generated from the two simulation programs, with some fraction s of signal events, using values for ϕ drawn from its prior. This classifier will take both ϕ and v as inputs — ie, it learns to classify for any value of ϕ, provided that the correct ϕ is known. Furthermore, this classifier will contain "bottlenecks" for both ϕ and v, which force the classifier to learn to use reduced-dimension versions of these inputs. In detail, we specify some small dimensionality for ϕ^* and some small dimensionality for v^* (eg, perhaps both are two-dimensional), and then train a classifier that has the following functional form for the estimated probability that an event is from the signal distribution:

$$P(\text{signal}|v,\phi) \quad \approx \quad d(g(\phi),h(v)) \tag{15}$$

where $\phi^* = g(\phi)$ is the reduced form of ϕ and $v^* = h(v)$ is the reduced form of the variables describing the event. The functions d, g, and h are parameterized in some way, such as with a multilayer perceptron ('backprop') neural network. Training of the classifier is done by adjusting these parameters to match the

data as sell as possible (eg, by some maximum penalized likelihood criterion). Dimensionality reduction schemes similar to this have long been used with neural networks (eg, see [4]).

We cannot use this classifier on real data, since we don't know the correct value for ϕ — nor for ϕ^*, which is all we would need. The only reason to train this classifier is to obtain the mappings $\phi^* = g(\phi)$ and $v^* = h(v)$, which, if the classifier is successful, must preserve most of the relevant information from ϕ and v. Note that we *can* obtain $v_i^* = h(v_i)$ for all the real events, since (once training is finished) h does not depend on the unknown value of ϕ. To obtain information about ϕ^* from the real events, we use simulated data to train a regression model (probably non-linear, perhaps a neural network) that approximates the expectation of ϕ^* given a single observation, v, by $e(v)$, where e is another parameterized function, learned from the data. Using data simulated only from the background distribution, p_0, may be sufficient, since even when the fraction of signal events is non-zero, it is typically quite small. This regression model also is merely used for dimensionality reduction, and will not be directly used to predict ϕ^*.

We are now in a position to produce the reduced data we will use for inference, consisting of the pairs $r_i = (h(v_i), e(v_i))$. If the dimensionality of these pairs is quite small, we can hope to build good models of the conditional densities $p_0(r|\phi)$ and $p_1(r|\phi)$ — most crudely, just by binning values for r, and using many simulated events with each value for ϕ, though more sophisticated methods may work better. Inference for f is then based on the marginal likelihood defined in equation (14), along with equation (13), computed by Monte Carlo, using a sample of values for ϕ drawn from its prior.

Note that the validity of the inference will depend only on the accuracy with which these densities are estimated, not on the quality of the classification and regression models described above. However, if these models are poor, much information may be lost, so the inferences may be uninformative. If instead our models are good, the $e(v_i)$ component of the r_i values will carry information about ϕ^*, with the result that most of the weight in the integral of (14) will be on values of ϕ close to the true value (or at least for which ϕ^* is close to its true value). For these values of ϕ, the likelihood will tend to favour values for f near the true value, because of the information carried in the $h(v_i)$ portion of r_i.

This scheme is untested, and may prove in practice to either lose too much information or be computationally infeasible — it may, for instance, prove necessary to train something more than a simple regression model in order to learn about ϕ^*. We can hope that success will be achievable using some such strategy for reducing dimensionality so that estimates for probabilities or probability density functions become feasible, allowing an approximation to the marginal likelihood function to be computed. When the parameter of interest (f in this problem) is one-dimensional (or more generally, of low enough dimension that a plot is intelligible), such a likelihood function is the most complete, and most useful, report of the experimental result.

Acknowledgements

I thank Nancy Reid, Zi Jin, and Byron Roe for helpful discussions. This research was supported by the Natural Sciences and Engineering Research Council of Canada. The author holds a Canada Research Chair in Statistics and Machine Learning.

References

[1] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, Chapman & Hall/CRC, 1974.

[2] G. J. Feldman and R. D. Cousins, "A unified approach to the classical statistical analysis of small signals", `http://arxiv.org/abs/physics/9711021v2`, 1997.

[3] H.-J. Yang, B. P. Roe, and J. Zhu, "Studies of Boosted Decision Trees for MiniBooNE Particle Identification", `http://arxiv.org/abs/physics/0508045`, 2005.

[4] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *Science*, vol. 313. no. 5786, pp. 504 - 507, 2006.

The Wish-lists: Some Comments

D.R. Cox and N. Reid
Nuffield College, Oxford and University of Toronto, Canada

Abstract
We provide brief comments on some common threads arising from the 'wish-lists' set out in some of the other papers in this volume. The discussion is necessarily incomplete: in particular we have dealt only with points for which a reasonably compact answer seems possible.

1 Introduction

The wish-lists are wide-ranging and raise issues of varying difficulties, ranging up to the seemingly impossible. The following comments concern just some of the topics raised.

2 Combination of independent sets of data

In the simplest situation there are m independent sets of data from each of which a common parameter θ can be estimated, representing for example some constant of interest. Separate analyses of the individual sets give estimates $t_1, \ldots, t_m$ with uncorrelated estimates of the variances $s_1^2, \ldots, s_m^2$. For an initial discussion ignore errors in the s_j^2.

If there are no additional sources of variation and the studies do indeed estimate the same unknown, there is the implicit model

$$t_j = \theta + \epsilon_j,$$

where the ϵ_j are independent Gaussian errors of zero mean and variances estimated by s_j^2. The parameter θ is estimated by weighted least squares or equivalently by ordinary least squares applied to a modified version of the model, namely

$$t_j/s_j = \theta/s_j + \epsilon_j/s_j,$$

where now the errors have unit variance. The estimate is

$$\tilde{\theta} = \frac{\Sigma t_j/s_j^2}{\Sigma 1/s_j^2},$$

with

$$\mathrm{var}(\tilde{\theta}) = \frac{1}{\Sigma 1/s_j^2}.$$

Importantly also the residual sum of squares from the modified model, namely

$$\Sigma(t_j/s_j - \tilde{\theta}/s_j)^2 = \Sigma t_j^2/s_j^2 - \tilde{\theta}^2 \Sigma 1/s_j^2,$$

has under the model a chi-squared distribution with $m-1$ degrees of freedom.

The argument can be refined, essentially by an empirical Bayes approach, to allow for errors in estimating the variances. The main practical point is that there can be major drawbacks to giving relatively high weight to individual estimates that have very small values of s_j^2 arising by chance.

Suppose now that the chi-squared test indicates clear heterogeneity, that is, the t_j vary too much. There are a number of possible explanations:

– the internal estimates of variance are unrealistically small

– there may be a small number of anomalous values
– there may be characteristics of the different studies which if entered into a regression equation for the t_j account for the additional variation.

If none of these is applicable and provided m is not too small, for example is at least, say, 10 it may be reasonable to suppose that there is an additional source of random error producing inter-study variation and to replace the starting model by

$$t_j = \theta + \eta_j + \epsilon_j,$$

where the ϵ_j are as before and the η_j are independent random variables of zero mean and unknown variance σ_η^2.

If in fact that variance were known, the least squares estimate of θ is

$$\frac{\Sigma t_j/(s_j^2+\sigma_\eta^2)}{\Sigma 1/(s_j^2+\sigma_\eta^2)}$$

a value intermediate between the simple weighted mean $\tilde{\theta}$ and the unweighted mean $\Sigma t_j/m$. The component of variance σ_η^2 can be estimated by maximum likelihood or, slightly less efficiently, by equating a sum of squares to its expectation. The assumptions involved in this formulation are quite strong and estimation of σ_η^2 is fragile if m is small. When the estimates are based on Poisson-distributed counts, the variances are functions of the counts themselves and this involves some changes in any more refined formulae. The additive representation for additional variation may be better replaced by a multiplicative form.

These issues are treated in more depth in [2].

3 Comparison of fit of a small number of models

Suppose first there are just two models neither of which is nested within the other. Two broad approaches might be considered. There are formidable practical difficulties in most situations with a Bayesian discussion, not so much connected with specifying the prior probabilities of the two models as with the conditional densities of the (different) parameters within each model. Unless these priors can be specified at least approximately on external evidence there is difficulty in computing the posterior probabilities required for model assessment.

A frequentist approach is to test model 1 for departures in the direction of model 2 and compute a p-value. Then switch the roles of the two models. There results information about whether both models give an adequate fit in the respect tested, whether one but not the other fits or whether neither model is adequate. In the last case, further analysis to develop an improved model would normally be required. Note that such a possibility cannot be directly obtained from the formal Bayesian approach.

With three or more models the best procedure is usually to test model 1, say, in turn against model 2 and then model 3 and to take the smaller p-value adjusted for selection as an assessment of model 1, and so on.

4 Systematics

Most statistical analysis focuses on random errors, it being assumed that the impact of systematic errors has been eliminated by design, that is by arranging that the effects of interest are estimated by comparisons of groups of data equally affected by systematic errors. There is also a substantial literature on estimating sources of variability in complex measurement systems intended, in particular, to aid the standardization of measurement techniques.

In the present context these methods are largely not applicable and explicit consideration of systematic errors seems unavoidable. A common approach seems to be to estimate the effect of an estimated physical constant on the final result of interest by re-computing this final result with the physical constant changed by plus and minus one standard error. Half the difference between these two resulting values is then approximately the derivative of that quantity with respect to the physical constant. This could be combined with other estimated sources of error in a propogation of errors formula, but it is essential to note that the errors in estimating the constant must be independent of errors from other sources. A less formal approach would be to investigate the sensitivity to the result of interest to errors in the physical constant by re-computing the results over a range of plausible values for the constant.

If there are k sources of systematic error and these can be given bounds, taken without loss of generality as say $(-\Delta_j, \Delta_j)$ for $j = 1, \ldots, k$, a very cautious approach is to do 2^k possible analyses based on the set of extreme possibilities, each with its confidence limits for the effect of interest and to take the union of these intervals as the basis of inference. Assumptions that the Δ_j are random variables may be reasonable but the key issue will often concern the independence assumptions involved which may have very strong implications. Ref. [3] has given a careful account of these issues from a Bayesian perspective.

Systematic errors that are essentially nuisance parameters in a model that is fully specified, or even partially specified, can be eliminated from the full likelihood by either maximizing over them or by integrating over them, with respect to a weight function. The integration approach is emphasized in [4]. . The maximization approach results in a profile likelihood, discussed in [5], and is implemented as MINOS in MINUIT. The limiting distribution of statistics based on the profile likelihood is the same as that for a simple likelihood. However the approximations given by this limiting theory, such as the χ^2 approximation to twice the log-(profile) likelihood ratio, can be quite inaccurate, especially if there are large numbers of nuisance parameters. Several adjustments to profile likelihood have been suggested in the statistical literature (see for example [1], Chs. 2 and 3), to take account of the uncertainty in estimating the nuisance parameters. These adjustments are implicit in the weight function applied in the integration approach, although the weight function is best thought of as a prior distribution on the nuisance parameters. The choice of the prior is important, and a large body of evidence now indicates that flat priors on the nuisance parameters are not appropriate, and can lead to very poorly calibrated inference, especially if there are large numbers of nuisance parameters. In some applications it may be possible to construct an empirical prior distribution from previous observations or from simulations.

5 Comparison of alternative test statistics

Tests are conventionally assessed by the power curve. In the simplest case of testing a null hypothesis H_0 that a single parameter θ is equal to θ_0 against alternatives $\theta > \theta_0$, the power curve shows the probability that the test "rejects" H_0 at level α as a function of θ. Equivalently the power curve shows the probability of a p-value less than α versus θ. If correctly calibrated the curve should pass through (θ_0, α). It is often a good idea to plot $\Phi^{-1}(\text{power})$ against θ, where $\Phi(x) = \int_\infty^x \{1/\sqrt{(2\pi)}\} \exp(-y^2/2) dy$ is the standard Gaussian distribution function. This produces a series of roughly parallel curves, or even approximately lines, for different α. In comparing two tests the steeper the curves the better.

More mathematically for test statistics that are approximately normally distributed we may define the efficacy of a test T as

$$\{ \frac{\partial E(T;\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \}^2 / \text{var}(T;\theta_0).$$

This measures the sensitivity of the expectation of T near the null hypothesis relative to the variance.

For two test statistics T_1, T_2 of the same null hypothesis the ratio of their efficacies is the asymptotic relative efficiency (ARE) of T_1 relative to T_2. Because efficacy usually scales as sample size, the ARE compares the sample sizes needed to achieve the same power with the two tests. Thus for testing

the mean of a Gaussian distribution the ARE of the median relative to the mean is $2/\pi$ so that tests based on the median of n observations and on the mean of $0.63n$ observations have about the same power.

These ideas may be useful even if the properties of the tests are studied primarily by simulation.

6 p-values and limits

The CL_s or CL_{s+b} methods combine size and power in a very *ad hoc* way and are unlikely to have satisfactory statistical properties. As is emphasized in Neal [4], upper and lower one-sided confidence limits should replace confidence intervals, and a full plot of the log-likelihood function is better still. A related point is that the construction of a p-value for discoveries, i.e. for confirming the existence of a particular effect, should be treated as a separate problem from the establishment of limits on the magnitude of a well-established effect. When there are several parameters of interest, a decision is needed about whether they can be assessed separately, treating the other parameters as nuisance parameters for each of these assessments, or whether it is physically more relevant to consider two (or more) of the parameters as forming a single vector. In the latter case approximate p-values can be computed using the usual asymptotic theory of likelihood, or a more refined version, but the construction of confidence regions is considerably more difficult and often not very illuminating.

References

[1] A.R. Brazzale, A.C.Davison, N. Reid, *Applied Asymptotics.*, Cambridge University Press, 2007.

[2] D.R. Cox, *Encyclopedia of Statistical Science* N.L. Johnson and S.Kotz, eds. New York: Wiley **2**, 45-53, 1982.

[3] S. Greenland, *Journal of the Royal Statistical Society: Series A* **168**, 267-306, 2005.

[4] R. Neal, in this volume.

[5] N. Reid and D.A.S. Fraser, in *Proceedings of PHYSTAT2003*, L. Lyons, R. Mount, R. Reitmeyer, eds. SLAC e-Conf C030908, 265–271, 2003.

Upper limits

Review of the Banff Challenge on Upper Limits

Joel Heinrich
University of Pennsylvania

Abstract
We report the results of the Limits Challenge project, in which participants were asked to provide upper limits on a cross section s measured in a counting experiment with nuisance parameters.

1 Introduction

In July of 2006, 40 physicists and statisticians met at the Banff International Research Station (BIRS) for the *Statistical Inference Problems in High Energy Physics and Astronomy Workshop* [1] organized by James Linnemann, Louis Lyons, and Nancy Reid. Here we report on the resulting Limits Challenge project. The specification of the challenge was:

The main experiment observes events with a Poisson rate that derives from a signal of cross section s (with acceptance ϵ) and background b. Nuisance parameters (ϵ, not constrained to be ≤ 1, is actually acceptance times integrated luminosity) are measured via Poisson subsidiary measurements:

$$\begin{array}{ll} n_i \sim \mathrm{Pois}(\epsilon_i s + b_i) & \text{(main measurement)} \\ y_i \sim \mathrm{Pois}(t_i b_i) & \text{(subsidiary background measurement)} \\ z_i \sim \mathrm{Pois}(u_i \epsilon_i) & \text{(subsidiary acceptance measurement)} \end{array}$$

Channels $i = 1, 2 \ldots N$. Constants t_i and u_i are known. Upper limits (or 2-sided intervals if required by the method) on parameter of interest s to be calculated at 90% and 99% level. The $2N$ parameters ϵ_i and b_i are to be considered nuisance parameters.

It was decided that participants would provide intervals for two situations, single channel and 10 channels. The data to be used was as follows: $N = 1$: I provided a list of ~ 100000 $(n_1, y_1, z_1, t_1, u_1)$ cases for which the intervals were returned by the participants. I made coverage curves from these (using importance sampling) and calculated the Bayesian credibility of the returned intervals. $N = 10$: Same as for $N = 1$ (I provided 50 numbers for each case). Participants were warned of possible coverage problems for Bayesian methods in higher dimensions[2, 3]. The test cases I provided to the participants consisted of 3 files obtainable from [4], described as follows.

Single channel data sets: two files in ASCII text format. Each line of each file has an (n, y, z, t, u) instance for which the participants provided two upper limits: at the 90% and 99% level. (Some methods provide 2-sided intervals for some (n, y, z, t, u) combinations.) Set-1 (60229 lines) has nuisance parameters with uncertainties of about 10%, while in set-2 (39700 lines) this is increased to about 30%.

One 10 channel data set: a single file (70000 lines) in ASCII text format. Each line of each file has the (n, y, z, t, u) for each of the 10 channels (for a total of 50 numbers per line). Nuisance parameter uncertainties are about 30%. Upper limits to be provided as specified above.

2 The Submitted Methods

Eleven methods were submitted. The raw files submitted by the participants are available from [4]. Not all the participants have submitted results for all data sets. Some of the methods have built-in preferences for upper limits or 2-sided intervals. Table 1 summarizes the received entries. General reviews of strategies that have been applied to this problem are available in [5] and [6].

Table 1: Submitted methods: • for '90% and 99% intervals', ○ for '90% intervals only'.

		upper limits			2-sided intervals			
designation	type	set-1	set-2	set-3	set-1	set-2	set-3	Section
MINUIT	profile	•	•	•	•	•	•	2.1
RLC	profile′	•	•		•	•		2.1
Davison–Sartori	H-O likelihood	•	•	•				2.2
Demortier	Bayesian	•	•	•				2.3
FHC^2	mixed				•			2.4
MBT	mixed				•			2.4
Baines	Bayesian	•	•	•			•	2.5
Baines-2	Bayesian	•		•				2.5
Edlefsen	Bayesian	•	•	•				2.6
Yu	Bayesian			•				2.7
Punzi	frequentist	○	•					2.8

2.1 MINUIT and RLC

This is the profile likelihood method, submitted by Wolfgang Rolke, historically known as the MINUIT [7] method in high energy physics. Jan Conrad has written a ROOT class `TRolke` [8] that implements the scheme for Poisson upper limits in a convenient way for ROOT users. `TRolke` actually has two variations on the profile likelihood: the default 'unbounded likelihood' method (here designated 'MINUIT'), and the 'bounded likelihood' method (designated 'RLC'). The main reference for the methods is [9], which shows coverage curves that can be compared with the 1-channel coverage curves in this study. As MINUIT is based only on the likelihood, the likelihood principle is obeyed. That is, the resulting intervals depend only on the form of the likelihood, not the probability (as in the frequentist approach). Nevertheless, profile methods are neither frequentist nor Bayesian, so both the coverage and credibility are of interest in this study.

2.2 Davison–Sartori

This submission, a higher order likelihood method, is from statistics professors Anthony Davison and Nicola Sartori. The method is described in [10], which lists the following features: parameterization-invariant; computation almost as easy as first order asymptotics; more accurate than use of Bartlett correction; error $O(n^{-3/2})$ in continuous response models; gives continuous approximation to discrete response models, with error $O(n^{-1})$ at support points of the discrete distribution; relative (not absolute) error, so highly accurate in tails; see [11] for a recent review.

2.3 Demortier

This submission is from Luc Demortier. It is a Bayesian approach using reference priors for the main and subsidiary measurements considered separately, not the full reference prior, which would consider the three Poisson measurements simultaneously.

2.4 FHC^2 and MBT

Jan Conrad and Fredrik Tegenfeldt have implemented a mixed method that is Bayesian with respect to the nuisance parameters and frequentist with respect to the parameter of interest. The Poisson probability is multiplied by priors for the nuisance parameters and integrated (marginalization), leaving only a dependence on the parameter of interest. Then the unified method of Feldman and Cousins [12] is employed to extract intervals. This approach is analogous to the procedure of Cousins and Highland [13], hence the designation 'FHC^2'.

The MBT method ('modified Bayesian treatment') is a variation of the FHC^2 method described in Section 2.4, in which the ordering rule is modified. This modification is a suggestion of Gary Hill [14]. Conrad and Tegenfeldt implement MBT and compare it with FHC^2 in [15].

2.5 Baines and Baines-2

This submission is from Harvard PhD statistics student Paul Baines, who presented the matching prior approach at this conference [16]. He has provided the following brief description of the method:

The method uses a basic 'one-level' Bayesian approach (i.e. fixed hyperparameters, no hyperpriors). A limited 'grid search' was performed in a simulation study, using priors of the form:

$$p(s,b,e) \propto (s^{\alpha_s-1})(b^{\alpha_b-1})(e^{\alpha_\epsilon-1})$$

for numerous $(\alpha_s, \alpha_b, \alpha_\epsilon)$ triplets. From simulation studies, 'Pseudo-Jeffreys' priors $(1/\sqrt{\cdot})$ for the nuisance parameters and a flat prior for the interest parameter appear to perform better than most 'one-level' schemes, although slight undercoverage is expected. The approach is simple, fast to compute, and provides a benchmark for comparison with other schemes. Other 'one-level' Empirical Bayes schemes were tried with limited success. Indeed, fully Bayesian hierarchical models (e.g., as implemented by Yaming Yu) appear to offer more flexibility in accurately modelling the three-Poisson structure of the problem.

The ten-channel submission is an implementation of Jeffreys prior (i.e. $\sqrt{\det I}$) where I is the Fisher Information matrix). The one-channel entry is a minor modification of Jeffreys prior: $(1/\epsilon)*$Jeffreys. Jeffreys prior has excellent coverage properties in the absence of nuisance parameters (it is 'first order probability matching', see below). However, coverage properties are known to deteriorate in many cases when nuisance parameters are present. This implementation was used to measure the deterioration in this particular example.

His description of Baines-2 is:

This submission was another Bayesian implementation, this time using a class of priors from Tibshirani (Biometrika, 1989). I am actually giving a contributed talk at the conference about this class of priors, they are related to 'Probability Matching Priors' which give Bayesian posterior intervals with Frequentist validity. The actual submission is not of this form and is a (poor!) approximation to it. I've made some progress on this class of priors since the submission.

2.6 Edlefsen

This submission is from Harvard PhD statistics student Paul Edlefsen. He has provided the following explanation of the method:

I have produced one-sided intervals for the BIRS A1 Challenge using a numerical approximation to the Dempster-Shafer (DS) relative plausibility of singletons function. This approach results in a Bayesian posterior, but uses an intermediate calculus (DS) that is a superset of the Bayesian calculus. Unlike pure-Bayesian approaches, this does not necessitate the use of a prior. Simply put, we consider random intervals that contain the true unknown s. The intervals have distributions deduced logically from the model using the relationship between Poisson processes and exponential sums. The one-channel posterior probability distribution for s, $F(s)$, is proportional to the probability that the random interval contains s. The ten-channel distribution is proportional to the product of these one-channel distributions. The method is simpler than non-DS Bayesian methods, and requires less time to compute.

2.7 Yu

This submission is from statistics Professor Yaming Yu. He has provided the following description:

This approach treats the 10 channels as exchangeable and builds a fully Bayesian hierarchical model. We specify a common prior distribution for the nuisance parameters ϵ_i's, and vague but proper hyper-priors for the parameters of this distribution. (Likewise for the b_i's.) The hyper-priors as well as the prior on the parameter of interest (s, or source intensity) are chosen to have good frequency properties as evaluated by separate simulations. After the model is specified, posterior inference is done through Markov chain Monte Carlo. Though Monte Carlo error is present in the reported 90 and 99 percent upper bounds, it can be reduced by running a longer chain or by using more sophisticated methods to estimate quantiles from the Monte Carlo output.

2.8 Punzi

Giovanni Punzi has been developing a fully frequentist method for this problem[17, 18]. He has collaborated with Pierluigi Catastini to calculated the submitted intervals, and they provided the following summary:

The limits are obtained by implementing in a C program the method described in [17]. Limits can be produced in the same way from any desired ordering (you can have two-sided FC limits, central limits, or whatever you like), but for this challenge they were explicitly required to be upper limits. The limits are constructed to always have coverage for any value of the physical and nuisance parameters. The program takes about 1 day to run for each of the proposed files. The step size in s was 0.2, and the scan goes up to $s = 20$. This limitation has no effect on the standard coverage plots of the challenge, but causes an underestimate of the actual credibility of the intervals (we thought of this side effect only after the run).

3 Coverage

Frequentist coverage is the first criterion by which the submissions are compared. The coverage probability is defined in the single channel case as

$$C(s, b, \epsilon) = \sum_{}^{'} \frac{e^{-\mu}\mu^n}{n!} \frac{e^{-\nu}\nu^y}{y!} \frac{e^{-\rho}\rho^z}{z!}$$

where $\mu = \epsilon s + b$, $\nu = tb$, and $\rho = u\epsilon$, and ϵ and b are fixed representative values for the coverage calculation (t and u are fixed values specified without uncertainty). Here $\sum'$ means sum only over values of (n, y, z) that yielded an interval that includes s.

This is the classic definition of frequentist coverage probability. s, ϵ and b are thought of as the 'true' values of the parameters that are unknown in real life. One investigates how the method performs for (representative) fixed true values of the parameters.

The 'true' values somewhat arbitrarily selected for the nuisance parameters to produce $C(s)$ for $0 \leq s \leq 20$ are:

set-1: $b = 3$, $\epsilon = 1$

set-2: $b = 3$, $\epsilon = 1$

set-3: $b_i = 0.31$, $\epsilon_i = 0.1$

Because of the range of (n, y, z, t, u) values provided in the 3 sets, the b and ϵ values assumed for a plot of $C(s)$ can be varied somewhat, e.g., for set-1 $\sim 2.5 \leq b \leq\sim 3.5$ is doable, but not much further outside that range. But no significant changes were observed for other values in the allowable range, so just one representative set is shown here.

Figure 1 shows $C(s)$ for selected 90% intervals, and Fig. 2 shows 99% intervals. Coverage curves for all submitted sets are available at [4]. Briefly summarizing:

– MINUIT covers at $\sim$ nominal for sets 1 and 2; set-3 is a bit lower, but is still acceptable.

– RLC's coverage can oscillate in the $0 < s < 5$ region, but otherwise OK.
– Davison–Sartori often undercovers at small s.
– Demortier undercovers in set-3.
– FHC^2 and MBT overcover slightly.
– Baines undercovers set 3.
– Baines-2 shows slight undercoverage.
– Edlefsen covers $\sim$ nominal for sets 1 and 3; for set-2 overcovers.
– Yu shows slight undercoverage.
– Punzi shows moderate overcoverage.

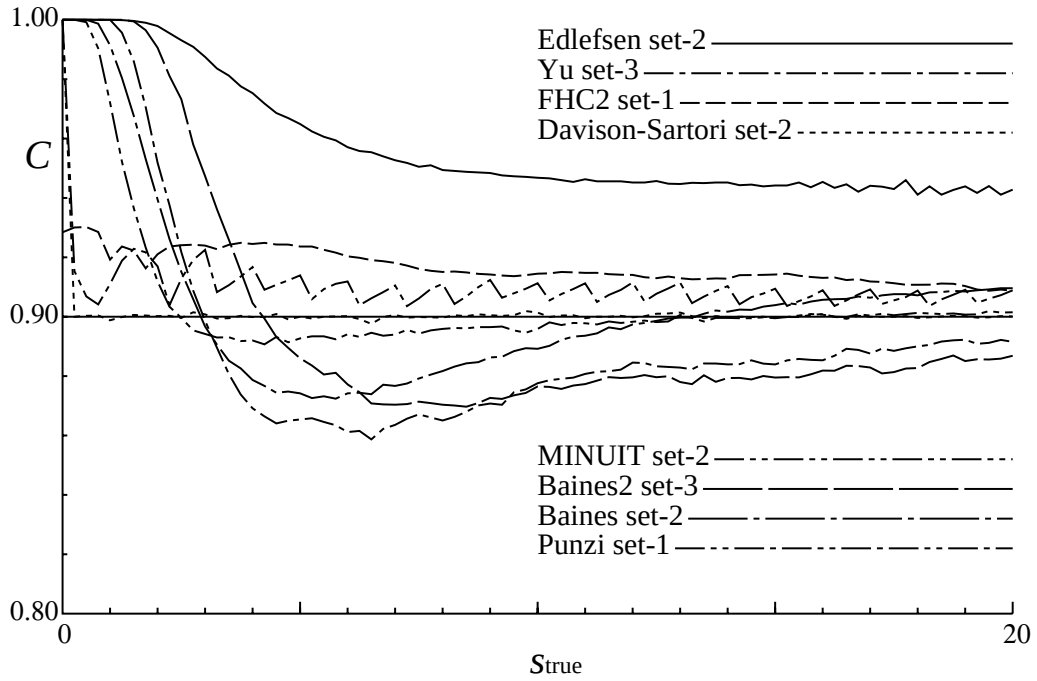

Fig. 1: Coverage of selected 90% intervals as a function of the true value of s.

All methods with submitted results on all 3 data sets show at least moderate deviations from nominal coverage (either overcoverage or undercoverage) for at least one of the sets, but most of these methods still seem usable. MINUIT, for example, achieves coverage properties similar to the more sophisticated Bayesian methods, but with much less computational cost.

4 Credibility

To further characterize the performance, one would like to have the Bayesian credibility for each of the supplied intervals. While the coverage calculation is completely specified by the definition, calculating the Bayesian credibility of the intervals supplied by the participants presents a bit of a problem, as one needs to select priors for the parameter of interest and the nuisance parameters. I have somewhat arbitrarily selected the following priors:

sets 1 and 2: flat for s, b and ϵ

set-3: flat for s. Priors for b and ϵ are $b_i^{-0.9}$ and $\epsilon_i^{-0.9}$

The priors for the nuisance parameters (applied to the likelihood for the auxiliary measurements) in the 10-channel case are chosen so that the effective priors for the total background $b' = \sum_i b_i$ and total acceptance $\epsilon' = \sum_i \epsilon_i$ are flat.

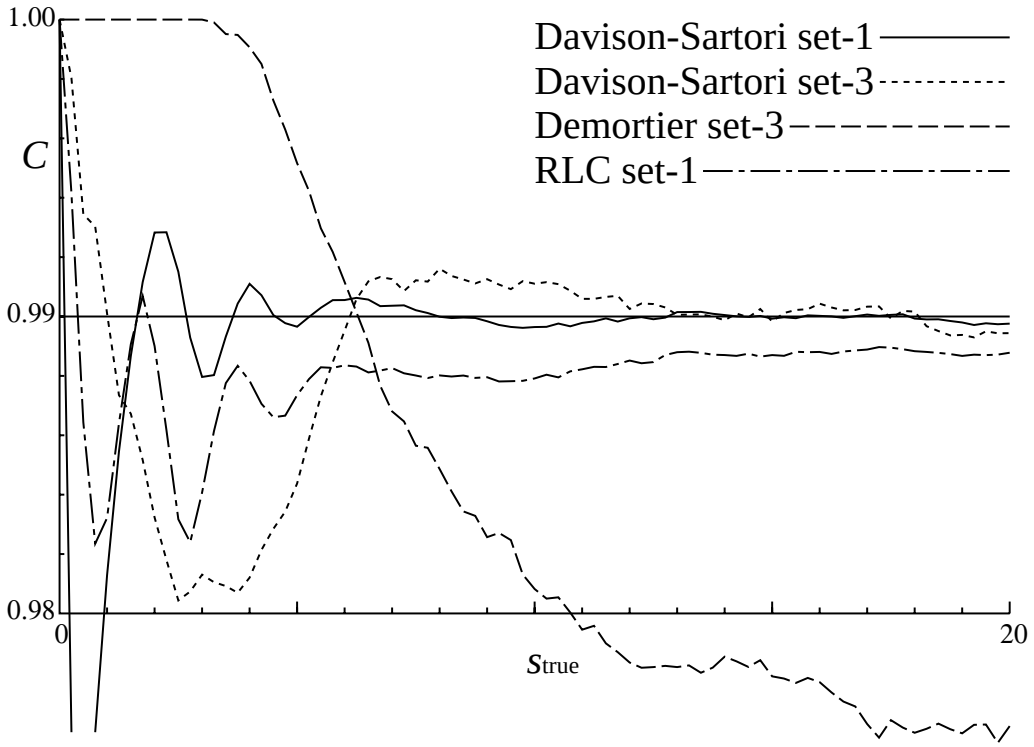

Fig. 2: Coverage of selected 99% intervals as a function of the true value of s.

Sample distributions of credibilities are shown in Figs. 3–5; see [4] for a complete set.

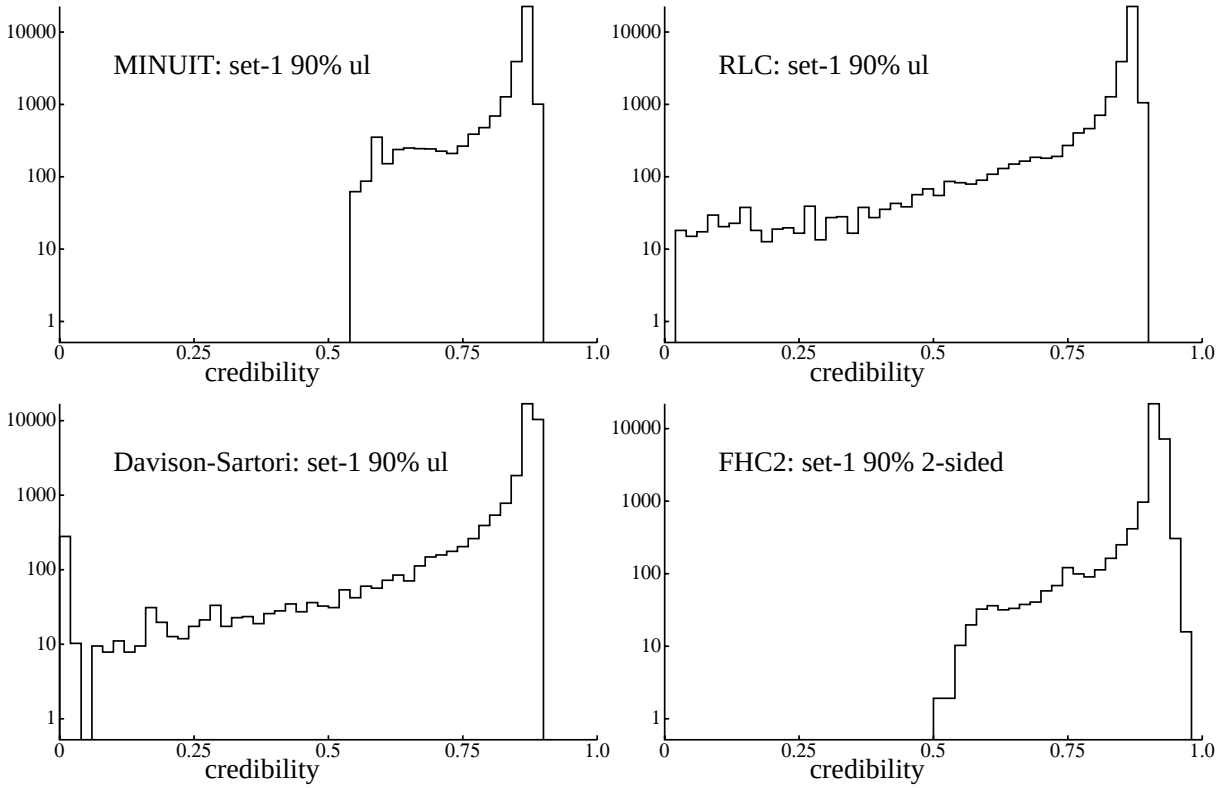

Fig. 3: Distribution of set-1 credibilities for selected methods.

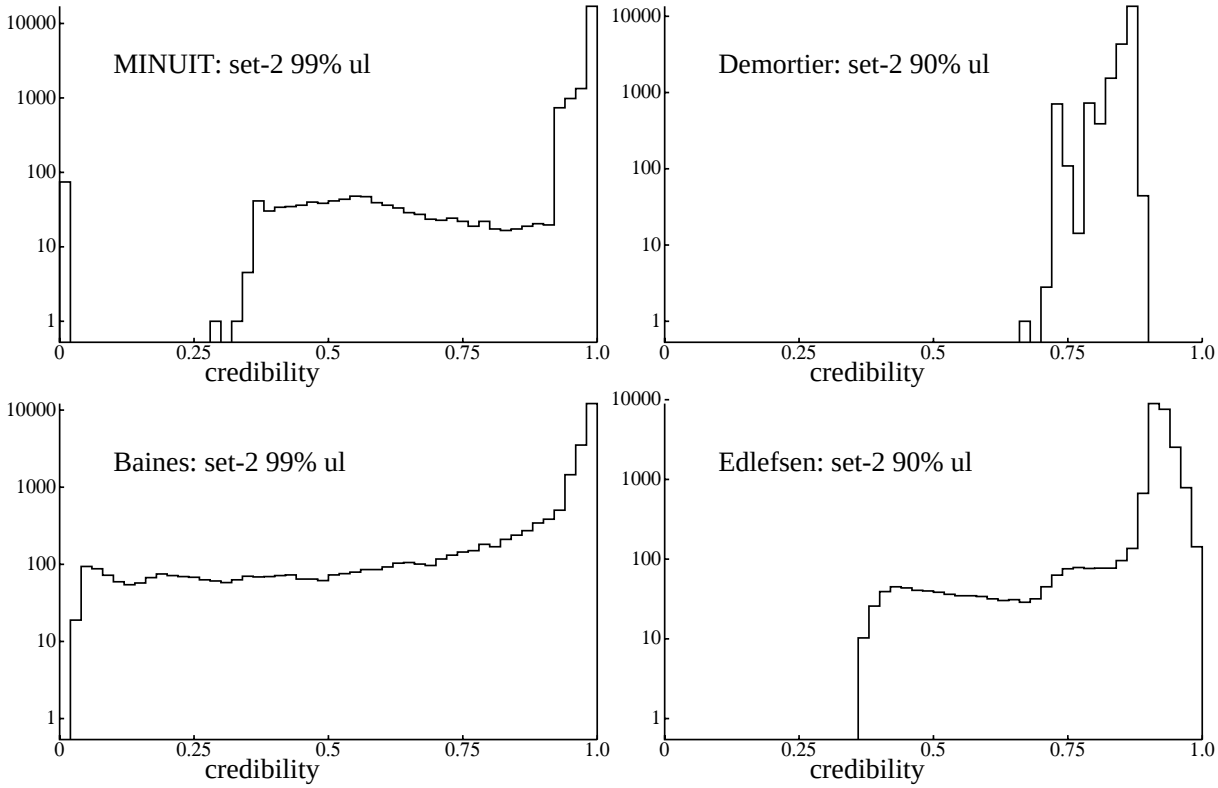

Fig. 4: Distribution of set-2 credibilities for selected methods.

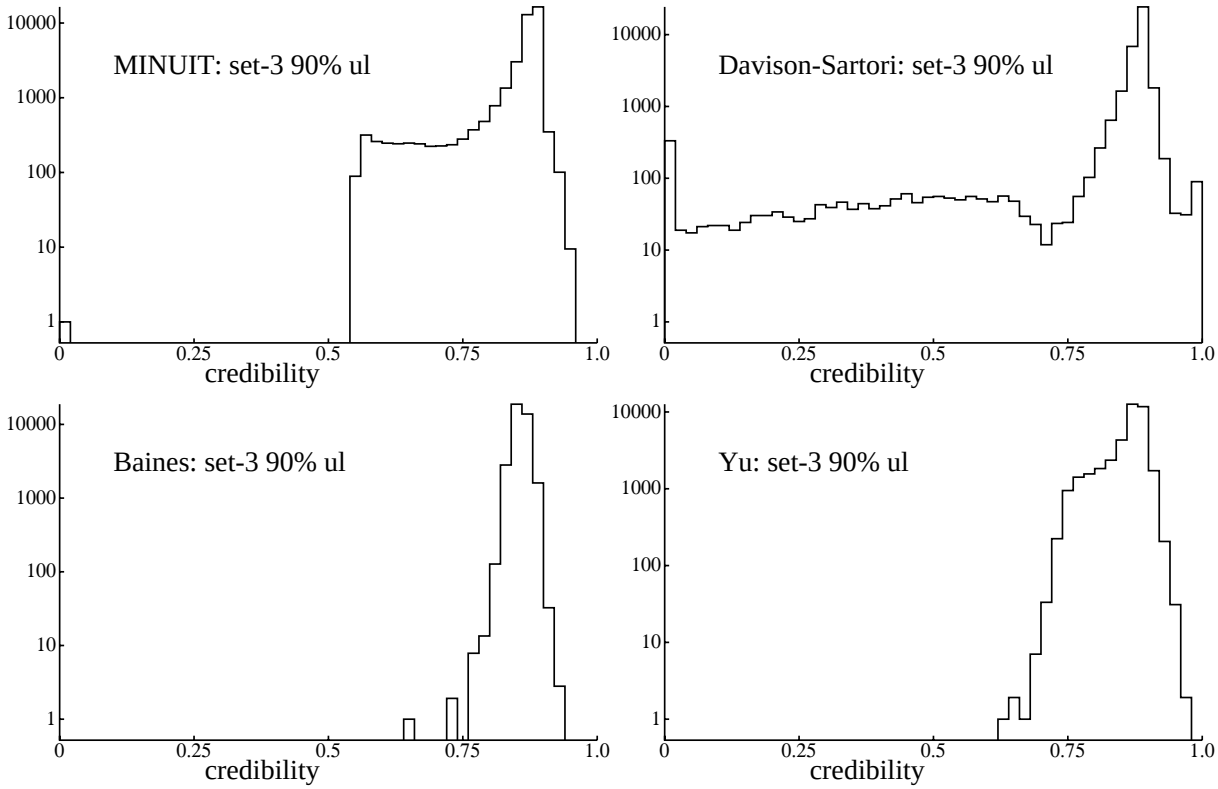

Fig. 5: Distribution of set-3 credibilities for selected methods.

Large deviations in credibility from nominal require investigation, but as the choice of prior is not unique, moderate deviations are not considered significant.

Table 2 shows some set-1 upper limits and calculated credibilities for comparison. The estimate of b (based on the observed y) increases as one moves down the table. RLC shows some intervals with rather low credibility. Focusing on the set-1 90% upper limits, one finds some intervals with credibilities as low as 2%. With $n = 1$, as y, the number of background events observed in the subsidiary background measurement, increases, the resulting upper limit drops rapidly to 0.02 at $y = 95$, then jumps discontinuously to 1.4 at $y = 96$:

Table 2: Selected 90% upper limits with credibilities for set-1 with $n = 1$, $z = 111$, $t = 33$, $u = 100$.

	RLC		D–S		Punzi		FHC^2		MINUIT	
y	ul	cred	ul	cred	ul	cred	ul	cred	ul	cred
79	0.458	0.307	0.727	0.445	1.0	0.560	2.013	0.819	1.13	0.606
84	0.322	0.229	0.591	0.383	0.8	0.483	1.861	0.796	1.10	0.601
89	0.187	0.141	0.455	0.313	0.6	0.392	1.664	0.760	1.08	0.598
90	0.160	0.122	0.428	0.297	0.6	0.393	1.664	0.760	1.08	0.599
91	0.133	0.103	0.401	0.282	0.6	0.393	1.664	0.761	1.07	0.599
92	0.106	0.083	0.374	0.266	0.6	0.394	1.664	0.761	1.07	0.597
95	0.025	0.020	0.292	0.216	0.4	0.284	1.664	0.763	1.06	0.595
96	1.366	0.692	0.265	0.198	0.4	0.284	1.664	0.764	1.05	0.592
98	1.312	0.678	0.211	0.162	0.4	0.285	1.512	0.731	1.05	0.594
101	1.230	0.656	0.130	0.103	0.4	0.287	1.512	0.732	1.04	0.592
103	1.175	0.640	0.076	0.061	0.2	0.155	1.512	0.734	1.03	0.590
107	1.066	0.605	0.000	0.000	0.2	0.156	1.375	0.701	1.02	0.589
114	0.876	0.537	0.000	0.000	0.0	0.000	1.375	0.704	1.00	0.586

Davison–Sartori also shows intervals with low credibility. As the background estimate increases, the upper limit drops gradually to zero, and stays there. Punzi, implementing a fully frequentist method, shows similar behaviour.

One of the benefits of the unified method of Feldman and Cousins is that it tends to avoid this behaviour. MINUIT also shows good performance with respect to this criterion; the credibility staying reasonably large:

4.1 Behaviour with zero observed events

With $n = 0$, the Poisson likelihood is $\exp[-(\epsilon s + b)]$. The shape of the likelihood with respect to the parameter of interest s is, in this special case, independent of the true value of b. Methods that obey the likelihood principle will consequently show no dependence of the upper limit for s on the background estimate or its uncertainty.

Alternatively: When zero events are observed in the main measurement, one knows that *zero signal* events were observed (and also zero background events). For the $n = 0$ special case, we have absolute separation between signal and background; consequently the uncertainty associated with not knowing if the events were signal or background is absent.

I check each submitted method to see whether the resulting intervals depend on the background estimate. Looking at set-1, for example, MINUIT, Demortier, Baines, Baines-2, and Edlefsen demonstrate background-independent $n = 0$ intervals.

For set-1, both Davison–Sartori and Punzi always produce zero-length 90% intervals whenever $n = 0$. As shown in Table 3, RLC and FHC^2 show a strange dependence of the limit on the background estimate when $n = 0$, and at 99%, Davison–Sartori shows a few rather narrow but finite intervals.

Table 3: Selected upper limits for set-1 with $n = 0$, $z = 110$, $t = 33$, $u = 100$.

y	RLC 90%	FHC^2 90%	D–S 99%
84	0.325	0.908	0.180
90	0.161	0.825	0.017
102	1.213	1.000	0.000
112	0.939	0.908	0.000
119	0.746	0.825	0.000
128	0.500	0.750	0.000

5 Conclusions

Comparison of the submitted results leads to the following conclusions about the performance of the methods. The conclusions are specific to the particular type of Poisson problem investigated in this project; they will not necessarily generalize to other applications (measurements of particle masses or lifetimes, for example) or to 5-σ confidence level. Comparing the methods:

- Overall, MINUIT (i.e. profile) is the easiest of the methods computationally, and its performance seems quite acceptable on the whole. The RLC variant performs less well, and was not provided for the 10-channel case.
- The fully Bayesian methods can perform excellently, but take more computational effort. One needs some care in selecting the priors, especially for the 10-channel case.
- The FHC^2 and MBT methods (mixed frequentist-Bayesian providing two sided intervals) behave well in general with respect to the coverage and credibility criteria, but it's not numerically clear what happens when $n = 0$ events are observed. (Of course, the frequentist component of FHC^2 and MBT does not necessarily satisfy the likelihood principle.)
- The fully frequentist method of Punzi and the higher order likelihood method of Davison–Sartori can produce zero-length or excessively narrow (i.e. low credibility) intervals. Punzi is not yet available for 10-channels. Davison–Sartori shows oscillations of coverage.

General conclusions are:

- Bugs are a ubiquitous problem; no software package is immune. Coverage and credibility checks were useful in uncovering some of these bugs. (Several of the entries were re-submitted after the initial coverage plots were viewed by the submitters.)
- Coverage is a well defined performance criterion. Bayesian credibility depends on the choice of prior(s), but intervals with very low credibility are worth investigating.
- Zero-length intervals are widely viewed as undesirable; very low credibility intervals seem undesirable for essentially the same reasons. Nevertheless, a document *Why Frequentists Should Care About Bayesian Credibility* may be necessary to convince hard core frequentists. (Does such a document already exist?)
- The companion document *Why Bayesians Should Care About Frequentist Coverage* would also be useful, and probably already exists.
- The Limits Challenge project has attracted significant interest, including both physicists and statisticians. It seems likely that after the PHYSTAT-LHC workshop more submissions will be sent to fill some of the gaps (or to fix some bugs) still present in the current submissions. These are certainly welcome.
- It would be useful to preserve the software that calculates the coverage and credibility, as well as the data sets and submitted files.

Acknowledgements

I would like to thank the organizers of the PHYSTATLHC and BIRS Conferences, and the many statisticians and physicists who submitted their methods to this project.

References

[1] http://www.pims.math.ca/birs/birspages.php?task=displayevent\&event_id=06w5054

[2] J. Heinrich, in *PHYSTAT05 Proceedings Statistical Problems in Particle Physics, Astrophysics and Cosmology* L. Lyons, U.M. Karagoz, eds. London: Imp. Coll. Press, 98 (2006)

[3] http://www.samsi.info/200506/astro/workinggroup/phy/jgh.pdf

[4] http://newton.hep.upenn.edu/~heinrich/birs/

[5] R. Cousins, in *PHYSTAT05 Proceedings Statistical Problems in Particle Physics, Astrophysics and Cosmology* L. Lyons, U.M. Karagoz, eds. London: Imp. Coll. Press, 75 (2006)

[6] J. Heinrich and L. Lyons, "Systematic Errors", in *Annual Review of Nuclear and Particle Science*, Vol. 57, Palo Alto, Annual Reviews, 145 (2007)

[7] F. James, *MINUIT—Function Minimization and Error Analysis*, Version 94.1, CERN Program Library Long Writeup D506, (1994)

[8] http://root.cern.ch/root/html/TRolke.html

[9] W. Rolke, A. Lopez, and J. Conrad, Nucl. Instrum. Methods Phys. Res. A 551/2-3, 493 (2005)

[10] http://www.imsv.unibe.ch/htm/sstats/NicolaSartori.pdf
http://newton.hep.upenn.edu/~heinrich/birs/davison-sartori/doc/

[11] A.R. Brazzale, A.C. Davison and N. Reid, *Applied Asymptotics: Case Studies in Small-Sample Statistics*, Cambridge, Cambridge University Press (2007)

[12] G.J. Feldman and R.D. Cousins, Physical Review D **57**, 3873 (1998)

[13] R.D. Cousins and V.L. Highland, Nucl. Instrum. Methods Phys. Res. A 331 (1992)

[14] Gary C. Hill, Physical Review **D67**, 118101 (2003)

[15] J. Conrad and F. Tegenfeldt, in *PHYSTAT05 Proceedings Statistical Problems in Particle Physics, Astrophysics and Cosmology* L. Lyons, U.M. Karagoz, eds. London: Imp. Coll. Press 93 (2006)

[16] Paul Baines, *Probability matching priors in LHC Physics—a pragmatic approach*, these proceedings

[17] G. Punzi, in *PHYSTAT05 Proceedings Statistical Problems in Particle Physics, Astrophysics and Cosmology* L. Lyons, U. M. Karagoz, eds. London: Imp. Coll. Press 88 (2006)

[18] http://www.samsi.info/200506/astro/workinggroup/phy/SAMSI_punzi.ppt.pdf

Probability Matching Priors in LHC Physics

P.D. Baines and X.-L. Meng
Department of Statistics, Harvard University

Abstract
Probability matching priors (PMPs) provide a bridge between Bayesian and frequentist inference by yielding Bayesian posterior intervals with frequentist validity. PMPs are, in general, challenging to implement as they are defined as solutions to a potentially high-dimensional and non-linear PDE. Outside the orthogonal case, no general framework exists for the implementation of PMPs. Recent work has made progress in this area, although no approach can yet be applied in generality. We consider PMPs for the three Poisson system arising in LHC experiments. Connections to reference and reverse reference priors are also considered. Theoretical and simulation results are presented, with comparison to other Bayesian techniques.

1 The Problem & Motivation

The problem of reliably estimating the intensity of a 'signal' in the presence of background and calibration uncertainties is a common one in LHC Physics and throughout the scientific world. Here we consider application of a class of Bayesian prior distributions to this problem, known as *probability matching priors* (PMPs). PMPs provide a bridge between the two main paradigms of statistical inference: frequentist and Bayes. Direct implementation of PMPs is, in general, extremely challenging as a result of possibly high-dimensional and non-linear partial differential equations (PDEs) that must be solved. This paper introduces both the rich rewards that may be reaped from applying PMPs in LHC Physics analyses, as well as the challenges that must first be overcome.

The primary criterion for the methods considered here will be their coverage properties. Other criteria such as credibility, length, bias and behavior in 'boundary' cases are also of importance and shall be addressed where space permits. PMPs are constructed to have (approximate) frequentist validity, will have good credibility over a range of prior distributions, and avoid many undesirable properties such as zero-length intervals. In this sense, where the desired coverage can be achieved, PMPs would appear to provide an 'optimal' solution likely to be accessible to both Bayesians and frequentists. However, existence of a PMP is not guaranteed. In the LHC example presented here, a large class of candidate priors are shown to not be PMPs.

While the theoretical properties of PMPs are well understood (see Ref. [1] for a review), their implementation remains an immense challenge. Recent papers by Levine & Casella [2] and Sweeting [3] have attempted to address this challenge, albeit not yet in full generality. In section 2 we provide a brief introduction to PMPs and orthogonality. Implementation is discussed in section 3, with an LHC application presented in section 4. Brief discussion is provided in section 5.

2 Introduction to Probability Matching Priors

2.1 Probability Matching Priors

The definition of a PMP for $\psi \in \mathbb{R}$, is that the posterior quantiles of ψ have (approximate) frequentist validity. See Ref. [1] for a formal definition. Peers [4] derived a PDE that a prior distribution must satisfy if it is to be first order probability matching (PM) (i.e., coverage of $\psi^{(1-\alpha)}$, the $100(1-\alpha)$ posterior percentile of ψ, is $1-\alpha+o(n^{-1/2})$ for all $0<\alpha<1$, where n is the sample size).

Theorem 1 ***First Order PMP Condition****: Let ψ be a univariate parameter of interest, with $\phi \in \mathbb{R}^{p-1}$ a nuisance parameter. The data are assumed to be generated from the family $f(\cdot;\psi,\phi)$. Let I_{ij} and I^{ij} denote the corresponding elements of the Fisher Information matrix and its inverse respectively. A prior $\pi(\cdot)$ is first order PM if and only if it satisfies the PDE:*

$$\frac{\partial}{\partial\psi}\left\{\pi(\psi,\phi)\cdot(I^{\psi\psi})^{1/2}\right\}+\sum_{j=1}^{p-1}\frac{\partial}{\partial\phi_j}\left\{\pi(\psi,\phi)I^{\phi_j\psi}(I^{\psi\psi})^{-1/2}\right\}=0. \tag{1}$$

Analytic solutions to this generally nonlinear $p-$dimensional PDE are rarely possible, and numerical solutions are often equally as elusive. However, in the case of an orthogonal parameterisation, that is, $I^{\psi,\phi_j}=0$ for all j, the solution is trivially given by:

$$\pi(\psi,\phi)=I_{\psi\psi}^{1/2}\cdot d(\phi) \tag{2}$$

where $d(\phi)$ is an arbitrary smooth function of the nuisance parameter (see Tibshirani, Ref. [5]). We, therefore, naturally attempt to extend the utility of (2) even when the parameterisation fails to be exactly orthogonal. The arbitrary function $d(\phi)$ can have a strong impact on finite-sample properties: the reverse reference prior [6] is a recommended tool for selecting within this class.

2.2 Orthogonality

The formal definition of orthogonality, from Cox & Reid [7], is that the partitioned Fisher Information (FI) is block diagonal, that is, $I^{\psi,\phi_j}=0$ for all j. Cox & Reid showed that for a scalar parameter of interest there always exists a transformation to achieve orthogonality with a $(p-1)-$dimensional nuisance parameter. However, the transformation is defined as the solution to a set of $(p-1)$ PDE's. These equations are in general not solvable by standard methods, and pose arguably a greater challenge than the PMP PDE (1). Hence, two obvious routes to finding probability matching priors, from the definition and via orthogonal parameterisation, are blocked by the obstacle of an intractable (set of) PDE(s). A third route is to derive either the reference prior of Berger and Bernardo [8], or reverse reference prior and check whether it is probability matching (frequently they are). However, outside the orthogonal case, their derivation can also become extremely challenging.

3 Existing Implementation Methods & Their Limitations

Levine & Casella [2] (LC) describe a Monte Carlo scheme to sample from the posterior distribution under a prior that is a solution to (1), when the nuisance parameter is univariate. The high run-time for the algorithm also makes it infeasible for large-scale simulation studies as considered here. Sweeting [3] proposes a more general approach that removes the restriction to univariate nuisance parameters, by seeking a *local probability matching prior*, using data-dependent approximations. The approach requires a non-trivial condition on the parameterisation, a condition that is not satisfied in the LHC application of section 4. In the general case it is unclear how to construct a parameterisation satisfying the condition if one is not immediately obvious. Indeed, in the LHC examples of section 4 the condition is not satisfied.

4 LHC Physics Example

The following problem is a common one in LHC Physics. The parameter of interest, s, represents the signal, monitored for M decay channels, with ϵ_i and b_i unknown channel-specific effective area and background parameters. Consider,

$$n_i|s,\epsilon_i,b_i\sim Pois\left(\epsilon_i s+b_i\right),\qquad y_i|b_i\sim Pois\left(t_i b_i\right),\qquad z_i|\epsilon_i\sim Pois\left(u_i\epsilon_i\right), \tag{3}$$

with $i=1,\ldots,M$, $\{t_1,\ldots,t_M,u_1,\ldots,u_M\}$ known constants and observations assumed to be independent. The goal is to find a PMP for s under this model. For simplicity we consider only the single channel

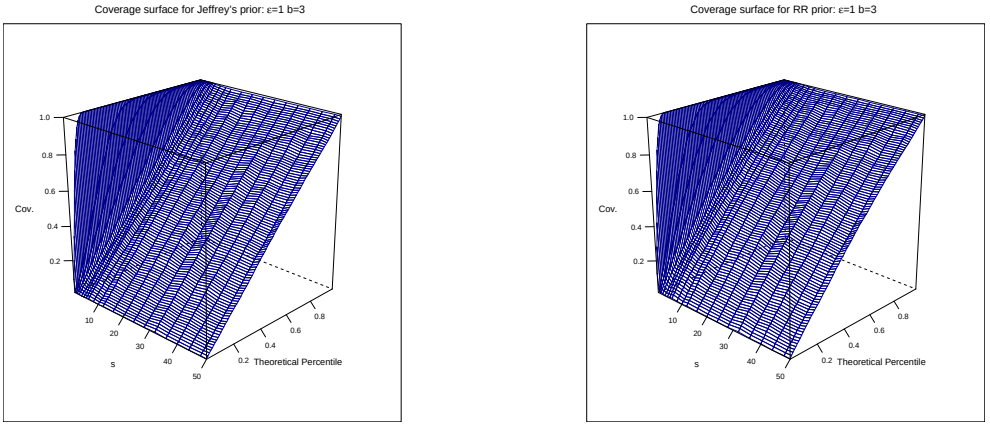

Fig. 1: Coverage surfaces for Jeffreys' [L] and the reverse reference [R] priors. The $z-$axis displays the coverage, x and y-axes indicate nominal coverage and the value of s. A 'perfect coverage method' would give a plane at 45°.

($M = 1$) case and drop the subscripts. The multi-channel setting is known to be more challenging, see Heinrich [9]. The first order PMP PDE can be shown to be:

$$\frac{\partial}{\partial s}\left\{\pi\sqrt{\frac{\epsilon st(u+s)+bu(1+t)}{\epsilon^2 tu}}\right\}-\frac{\partial}{\partial b}\left\{\pi\sqrt{\frac{b^2u}{\epsilon st^2(u+s)+but(1+t)}}\right\}-\frac{\partial}{\partial \epsilon}\left\{\pi\sqrt{\frac{s^2\epsilon t}{\epsilon stu(u+s)+bu^2(1+t)}}\right\}=0 \tag{4}$$

This cannot be directly solved by standard software, which may suggest that no solution exists. The prior from (2) here becomes $\pi(s,b,\epsilon) \propto d(b,\epsilon)/\sqrt{s\epsilon+b}$. Jeffreys prior is found to be a special case, where $d(b,\epsilon)=\sqrt{\epsilon tu/b}$. This is not the case for $M>1$. In all cases posterior propriety must be checked. The general $M-$channel reverse reference prior π_{rr} can be fully derived. The regular reference prior π_r for the ordered parameterisation $\psi = s, \phi = (\mathbf{b}, \epsilon)$, if it exists, is of the form:

$$\pi_{rr}(s,\mathbf{b},\epsilon) \quad \propto \quad \sqrt{\frac{\sum_{j=1}^{M}\epsilon_j u_j}{\prod_{j=1}^{M} b_j\epsilon_j}\cdot\sum_{j=1}^{M}\frac{\epsilon_j^2}{s\epsilon_j+b_j}\cdot\frac{1}{\sum_{j=1}^{M}\epsilon_j}}\,, \tag{5}$$

$$\pi_{r}(s,\mathbf{b},\epsilon) \quad \propto \quad g(s)\sqrt{\prod_{j=1}^{M}\frac{b_ju_j(1+t_j)+\epsilon_j s t_j(s+u_j)}{b_j\epsilon_j(b_j+\epsilon_j s)}}\,. \tag{6}$$

By plugging in the form of the prior distribution into (4), it can be proved that, for the single-channel case, neither the regular reference prior nor any priors within the Tibshirani class of priors from (2) can be a PMP. For example, plugging the reference prior into (1), we obtain an ODE for the function $g(s)$. However, this ODE can be shown to have no solution. An analogous proof holds for the Tibshirani class from (2), hence, also the reverse reference prior. These results, combined with the failure to directly solve (4), strongly suggest that there in fact may be no PMP in this example.

Instead, we considered three priors of the form (2):

$$d_J(b,\epsilon)=\sqrt{\epsilon/b} \qquad d(b,\epsilon)=1/\sqrt{b\epsilon} \qquad d(b,\epsilon)=1, \tag{7}$$

where d_J corresponds to Jeffreys prior and $d=1/\sqrt{b\epsilon}$ to a form of pseudo-Jeffreys' prior for b and ϵ. For comparison, priors of the form $\pi(s,b,\epsilon)\propto\frac{1}{\sqrt{s}}$ and $\frac{1}{s}$ are also considered. 110,000 datasets were simulated from (3) with $b=3$, $\epsilon=1$, $t=33.0$, $u=100.0$ and with s taking on 22 values in the range 0.1 to 48.0. Posterior intervals were obtained under all of the above prior distributions. Figure 1 displays the coverage surface for Jeffreys' prior. Numerical results are presented in Table 1. Both Jeffreys' and the $d=\frac{1}{\sqrt{b\epsilon}}$ prior have excellent coverage properties over a wide range of s. For $M\geq 8$, say, coverage properties often deteriorate. Overcoverage for small s is inevitable under the Bayesian methodology, and a necessary price to pay for any method that does not produce zero-length intervals.

Table 1: For $s = 20$, the actual coverage of nominal $5, 10, 25, 50, 75, 90, 95$ & 99^{th} percentiles produced by using each of the different priors discussed in section 4

$s^{(\alpha)}$	π_{rr}: $d = \frac{1}{\sqrt{b}}$	Jeffreys': $d = \sqrt{\frac{\epsilon}{b}}$	$d = \frac{1}{\sqrt{b\epsilon}}$	$d = 1$	$\pi \propto 1$	$\pi \propto \frac{1}{\sqrt{s}}$
$s^{(0.05)}$	0.06	0.05	0.05	0.05	0.06	0.06
$s^{(0.10)}$	0.12	0.10	0.11	0.11	0.12	0.12
$s^{(0.25)}$	0.29	0.25	0.27	0.28	0.29	0.29
$s^{(0.50)}$	0.54	0.49	0.51	0.52	0.54	0.54
$s^{(0.75)}$	0.78	0.74	0.75	0.76	0.78	0.78
$s^{(0.90)}$	0.91	0.89	0.90	0.91	0.91	0.91
$s^{(0.95)}$	0.96	0.95	0.95	0.96	0.96	0.96
$s^{(0.99)}$	0.99	0.99	0.99	0.99	0.99	0.99

5 Discussion

Since the primary goal is to produce intervals with frequentist validity: "why not just be a frequentist?" One benefit of PMPs is that they produce intervals accessible to frequentists and Bayesians alike. Moreover, this accessibility is independent of the criteria by which they are evaluated. Other criteria are discussed in Heinrich [9]. In many of these respects PMPs may provide a more satisfactory solution than other methods produced from frequentist principles alone. For example, as discussed in Heinrich [9], both frequentist and likelihood-based methods can produce undesirable zero-length intervals. This behaviour cannot occur under the Bayesian construction presented here, a side-effect of this is overcoverage for small signal s.

PMPs, where they exist, may provide an 'optimal' solution to coverage problems. In the LHC example considered here, no exact PMP has been found so far, but approximate PMPs seem to exist over restricted ranges of the parameter space, and may be all that is required for practical purposes. Reference and reverse priors are also recommended as an effective default prior for Bayesian inference, often satisfying the PMP property. Further progress on both computational issues and operational properties will help give practitioners another option for making reliable inference about important physical parameters arising in LHC experiments.

References

[1] Datta, G.S. & Mukerjee, R., (2004) Probability Matching Priors: Higher Order Asymptotics. *Lecture Notes in Statistics 178, Springer-Verlag*.

[2] Levine, R.A. & Casella, G. (2003) Implementing probability matching priors for frequentist inference. *Biometrika* **90**, 127-137.

[3] Sweeting, T.J. (2005) On the implementation of local probability matching priors for interest parameters. *Biometrika* **92**, 47-57.

[4] Peers, H.W. (1965) On Confidence sets and Bayesian probability points in the case of several parameters. *J. of the Royal Stat. Soc. B* **53**, 611-618.

[5] Tibshirani, R.J. (1989) Noninformative priors for one parameter of many. *Biometrika* **76**, 604-608.

[6] Berger, J.O. (1993) Discussion of "Non-informative Priors" by Ghosh, J.K, Mukerjee, R. *Bayesian Statistics 4*, 205-206

[7] Cox, D.R. & Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *J. of the Royal Stat. Soc. B* **49**, 1-39.

[8] Berger, J.O, & Bernardo, J.M. (1992) On the Development of Reference Priors. *Bayesian Statistics 4*, 35-60.

[9] Heinrich, J. (2007) Report on the Banff Limits Challenge. `http://newton.hep.upenn.edu/~heinrich/challenge.pdf`

Analysis methods

The Role of Uncertainties in Parton Distribution Functions

R.S. Thorne[1*]
Department of Physics and Astronomy, University College London, WC1E 6BT, UK

Abstract
I consider the uncertainties in parton distributions and the consequences for hadronic cross-sections. There is ever-increasing sophistication in the relationship between the uncertainties of the distributions and the errors on the experimental data used to extract them. However, I demonstrate that this uncertainty is frequently subsumed by that due to the choice of data used in fits, and more surprisingly by the precise details of the theoretical framework used. Variations in heavy flavour prescriptions provide striking examples.

1 Introduction

When calculating cross-sections for scattering processes involving hadronic particles one requires detailed knowledge of the input parton distributions. The uncertainties in the latter propagate into the uncertainties on the former, and are often significant and sometimes dominant. The parton distributions can be derived within QCD using the Factorization Theorem, i.e. the cross-section for a physical cross-section at the LHC can be written in the factorised form

$$\sigma(pp \to X_P) \propto \sum_i \sum_j C_{ij}^P(x_1, x_2, \alpha_s(M^2)) \otimes f_i(x_1, M^2) \otimes f_j(x_2, M^2), \tag{1}$$

up to small corrections, where P represents some arbitrary process with hard scale (e.g. particle mass, jet E_T, ...). The coefficient functions $C_{ij}^P(x_1, x_2, \alpha_s(M^2))$ describing the hard scattering process of the two incoming partons are process dependent but calculable as a power-series in $\alpha_S(M^2)$. The $f_i(x, M^2)$ are the parton distributions – heuristically the probability of finding a parton of type i carrying a fraction x of the momentum of the proton. The parton distributions are not calculable from first principles, but evolve with M^2 in a perturbative manner governed by the splitting functions $P_{ij}(x, \alpha_s(M^2))$ which are calculable order by order in perturbation theory. Hence, once measured at one scale the distributions can be predicted at other scales.

In this article I will briefly review the extraction of the parton distributions and the resulting uncertainties. This is an update of a previous article in this series of Workshops [1], so I will concentrate on new developments. A full discussion of fitting procedures and uncertainties due to experimental errors on the input data is found in [1], but I will very briefly restate the essentials, including some updates.

There are a variety of sets of parton distributions which are obtained by a comparison to all available data (so-called global fits) [2, 3] or to smaller subsets of mainly structure function data [4, 5, 6], sometimes only in the nonsinglet sector [7, 8]. All follow the same general principle. The fit usually proceeds by starting the parton evolution at a low scale Q_0^2 and evolving partons upwards (sometimes also downwards) using fixed order evolution equations. The default has long been next-to-leading order (NLO), but the next-to-next-to-leading order (NNLO) splitting functions were recently calculated [9], and sets of NNLO distributions are also available [11, 10]. In principle, there are 11 different parton distributions (assuming isospin symmetry and ignoring the top quark) – the 5 quarks, up, down, strange, charm, and bottom and their antiquarks, and the gluon distribution. Until recently these were not all considered independent, but there is now some evidence for asymmetry between strange quarks and antiquarks [13], and moreover all quarks evolve slightly differently from their antiquarks due to evolution effects which begin at NNLO. However, in practice $m_c, m_b \gg \Lambda_{\rm QCD}$, so the heavy parton distributions

[*]Royal Society University Research Fellow

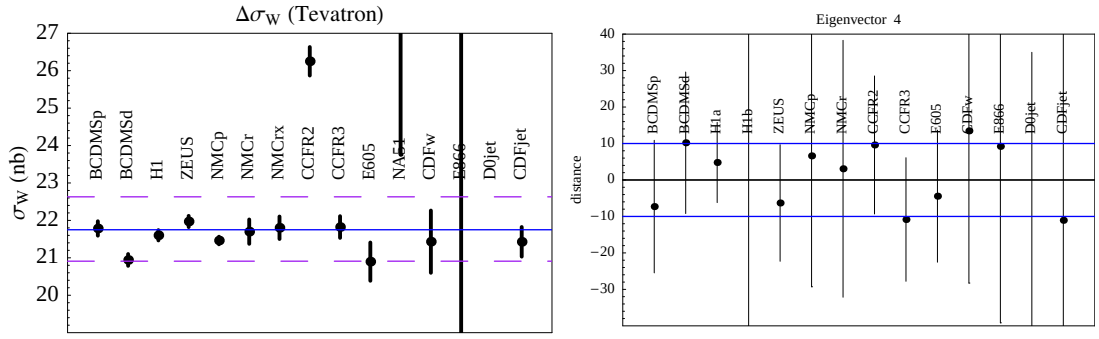

Fig. 1: The best value of σ_W and the uncertainty using $\Delta\chi^2 = 1$ for each data set in the CTEQ fit (left) and the 90% confidence limits for each data set as a function of $\sqrt{\Delta\chi^2}$ for one particular eigenvector (right)

are usually determined perturbatively and there are 7 independent input parton sets, each parameterised in a particular form, e.g.

$$xf(x,Q_0^2) = A(1-x)^\eta(1+\epsilon x^{0.5}+\gamma x)x^\delta. \tag{2}$$

The partons are constrained by a number of sum rules: i.e. conservation of the number of valence up and down quarks, zero number asymmetry for the other quarks and the conservation of the momentum carried by partons. The last is an important constraint on the form of the gluon, which is only probed indirectly. In determining partons one needs to consider that not only are there many different distributions, but there is also a wide distribution of x from 0.75 to 0.00003. One needs many different types of experiment for full determination, as discussed in [1]. For instance, the MRST (now MSTW [14]) group use 29 different types of data set.

The quality of the fit is determined by the χ^2 of the fit to data, which may be calculated in various ways. The simplest is to add statistical and systematic errors in quadrature, which ignores correlations between data points, but is sometimes quite effective. Also, the information on the data often means that only this method is available. More properly one uses the full covariance matrix which is constructed as

$$C_{ij} = \delta_{ij}\sigma^2_{i,stat} + \sum_{k=1}^{n}\rho^k_{ij}\sigma_{k,i}\sigma_{k,j}, \qquad \chi^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}(D_i - T_i(a))C^{-1}_{ij}(D_j - T_j(a)), \tag{3}$$

where k runs over each source of correlated systematic error, ρ^k_{ij} are the correlation coefficients, N is the number of data points, D_i is the measurement and $T_i(a)$ is the theoretical prediction depending on parton input parameters a. An alternative that produces identical results if the errors are small is to incorporate the correlated errors into the theory prediction

$$f_i(a,s) = T_i(a) + \sum_{k=1}^{n} s_k\Delta_{ik}, \qquad \chi^2 = \sum_{i=1}^{N}\left(\frac{D_i - f_i(a,s)}{\sigma_{i,unc}}\right)^2 + \sum_{k=1}^{n} s_k^2, \tag{4}$$

where Δ_{ik} is the one-sigma correlated error for point i. One can solve analytically for the s_k [15].

Having defined the fit quality there are a number of different approaches for obtaining parton uncertainties. The most common is the Hessian (Error Matrix) approach. One defines the Hessian matrix by

$$\chi^2 - \chi^2_{min} \equiv \Delta\chi^2 = \sum_{i,j} H_{ij}(a_i - a_i^{(0)})(a_j - a_j^{(0)}). \tag{5}$$

One can then use the standard formula for linear error propagation:

$$(\Delta F)^2 = \Delta\chi^2 \sum_{i,j}\frac{\partial F}{\partial a_i}(H)^{-1}_{ij}\frac{\partial F}{\partial a_j}. \tag{6}$$

This has been used to find partons with errors by $H1$ [6] and Alekhin [4]. In practice it is problematic due to extreme variations in $\Delta\chi^2$ in different directions in parameter space. This is improved by finding and

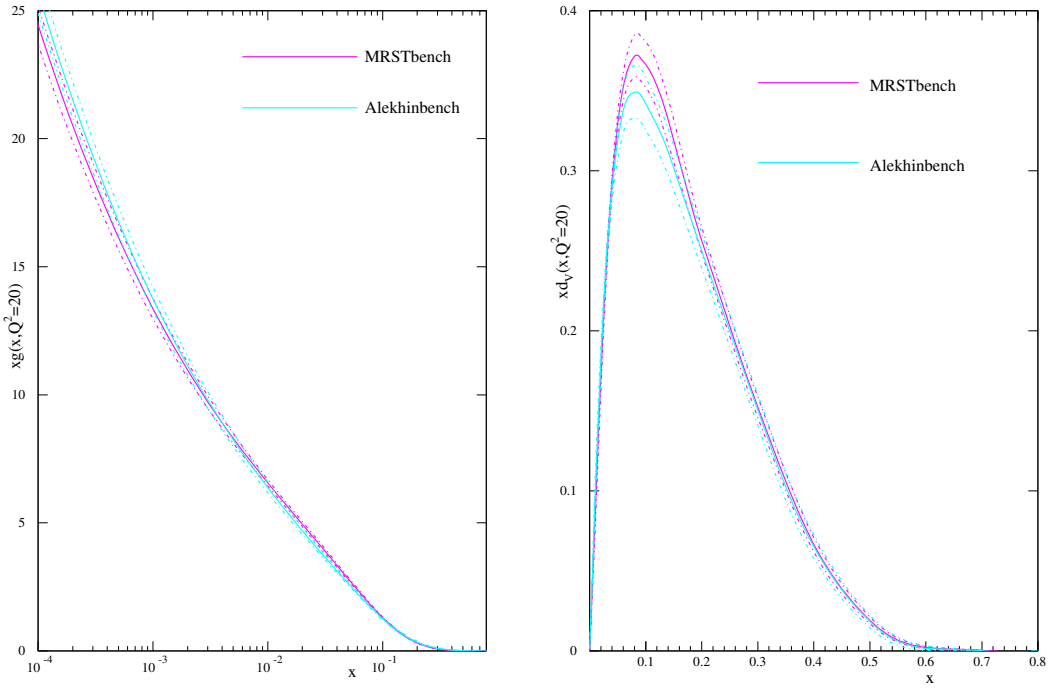

Fig. 2: Comparison of the benchmark gluon distributions and d_V distributions

rescaling the eigenvectors of H, a method developed by CTEQ [16, 17], and now used by most groups. The uncertainty on a physical quantity is

$$(\Delta F)^2 = \frac{1}{2}\sum_i \big(F(S_i^{(+)}) - F(S_i^{(-)})\big)^2, \tag{7}$$

where $S_i^{(+)}$ and $S_i^{(-)}$ are PDF sets displaced along eigenvector directions by the given $\Delta\chi^2$.

One can also investigate the uncertainty on a given physical quantity using the Lagrange Multiplier method, first suggested by CTEQ [15] and also used by MRST [18]. One performs the global fit while constraining the value of some physical quantity, i.e. minimise

$$\Psi(\lambda, a) = \chi^2_{global}(a) + \lambda F(a) \tag{8}$$

for various values of λ. This gives the set of best fits for particular values of the parameter $F(a)$ without relying on the quadratic approximation for $\Delta\chi^2$, but has to be done anew for each quantity.

In each approach there is uncertainty in choosing the "correct" $\Delta\chi^2$. In principle this should be one unit, but given the complications of a full global fit this gives unrealistically small uncertainties. This can be seen in the left of Fig. 1 where the variation in the predictions for σ_W using $\Delta\chi^2 = 1$ for each data set has an extremely wide scatter compared to the uncertainty. CTEQ choose $\Delta\chi^2 \sim 100$ [15]. The 90% confidence limits for the fits to the larger individual data sets when $\sqrt{\Delta\chi^2}$ in the CTEQ fit is increased by a given amount are shown in the right of Fig. 1. As one sees, a couple of sets may be some way beyond their 90% confidence limit for $\Delta\chi^2 = 100$. The MRST/MSTW group chooses $\Delta\chi^2 = 50$ to represent the 90% confidence limit for the fit. Other groups with much smaller data sets and fewer complications still use $\Delta\chi^2 = 1$.

There are other approaches to finding the uncertainties. In the offset method the best fit is obtained by minimising the χ^2 using only uncorrelated errors. The systematic errors on the parton parameters a_i are determined by letting each $s_k = \pm 1$ and adding the deviations in quadrature. This method was used in early H1 fits [19] and by early ZEUS fits [20], but is uncommon now. There is also the statistical approach used by Neural Network group [8]. Here one constructs a set of Monte Carlo replicas $\sigma^k(p_i)$ of the original data set $\sigma^{data}(p_i)$ which gives a representation of $P[\sigma(p_i)]$ at points p_i. Then one trains

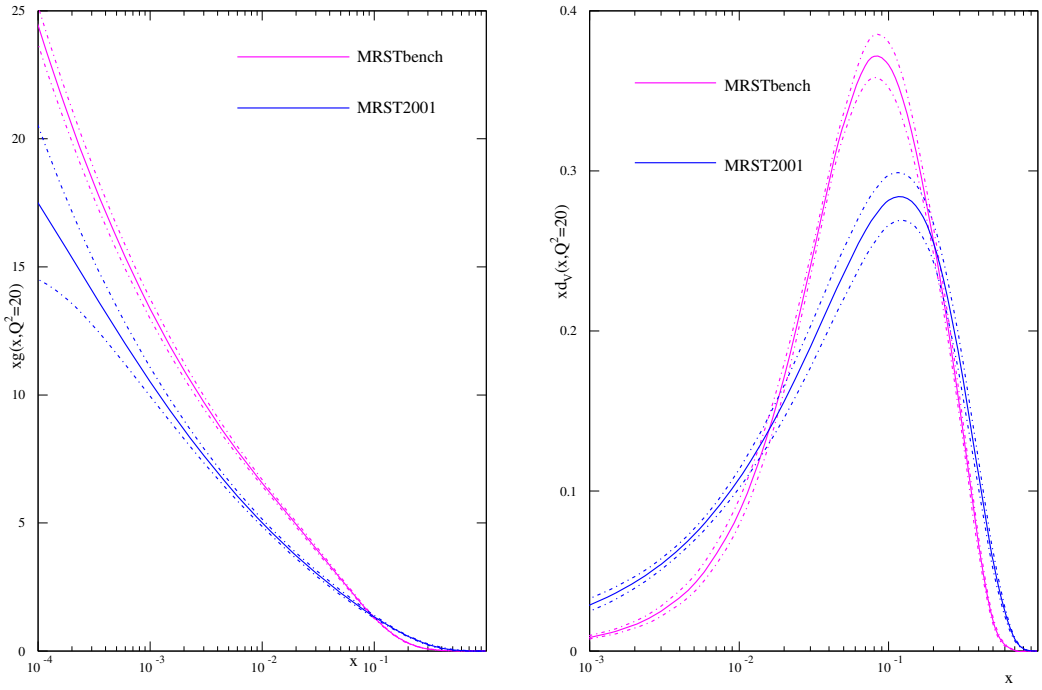

Fig. 3: Comparison of the benchmark gluon and d_V distribution with the corresponding MRST2001E partons

a neural network for the parton distribution function on each replica, obtaining a representation of the pdfs $q_i^{(net)(k)}$. The set of neural nets is a representation of the probability density – i.e. the mean μ_O and deviation σ_O of an observable O is given by

$$\mu_O = \frac{1}{N_{rep}} \sum_{1}^{N_{rep}} O[q_i^{(net)(k)}], \quad \sigma_O^2 = \frac{1}{N_{rep}} \sum_{1}^{N_{rep}} (O[q_i^{(net)(k)}] - \mu_O)^2. \tag{9}$$

One can incorporate full information about measurements and their error correlations in the distribution of $\sigma^{data}(p_i)$. This is does not rely on the approximation of linear propagation of errors but is more complicated and time intensive. It is currently done for the nonsinglet sector only.

2 Sources of Uncertainty

In recent years there has been a great deal of work on the correct and complete inclusion of the experimental errors on the data when extracting the partons and their uncertainties. However, to obtain a complete estimate of errors, one also needs to consider the effect of the decisions and assumptions made when performing the fit, e.g. cuts made on the data, data sets fit and parameterization for the input sets.

As an exercise for the HERA-LHC [21] workshop, partons were produced from fits to some sets of structure function data for $Q^2 > 9\text{GeV}^2$ using a common form of parton inputs at $Q_0^2 = 1\text{GeV}^2$. Partons were obtained using the rigorous treatment of all systematic errors (labelled Alekhin) and using the simple quadratures approach (labelled MRST), both using $\Delta\chi^2 = 1$ to define the limits of uncertainty. This benchmark test is clearly a very conservative approach to fitting that should give reasonable partons with bigger than normal uncertainties. As seen in Fig. 2 there are small differences in the central values and similar errors, i.e. the two sets are fairly consistent. It is more interesting to compare the HERA-LHC benchmark partons to partons obtained from a global fit [18], where the uncertainty is determined using $\Delta\chi^2 = 50$. There is an enormous difference in the central values, sometimes many σ, as seen in Fig. 3, although the uncertainties are similar using $\Delta\chi^2 = 1$ compared to $\Delta\chi^2 = 50$ with approximately twice the data. Moreover, $\alpha_S(M_Z^2) = 0.1110 \pm 0.0015$ from the benchmark fit compared to $\alpha_S(M_Z^2) = 0.119 \pm 0.002$. Something is clearly seriously wrong in one of these analyses, and indeed partons from

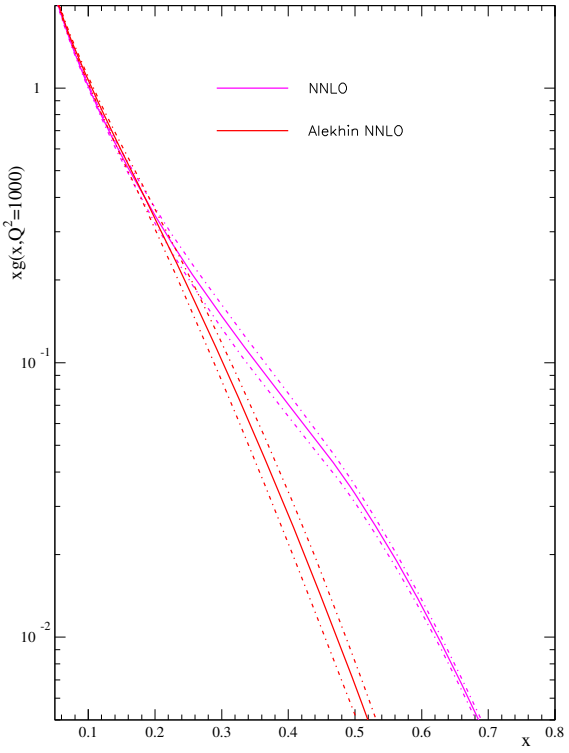

Fig. 4: Comparison of MRST and Alekhin NNLO gluon distributions at high x

the benchmark fit fail when compared to most data sets not included. This implies that partons should be constrained by all possible reliable data.

The benchmark partons above are not a realistic set of partons, but similar examples are found when comparing different sets of published parton distributions. For example, the valence quarks extracted from the nonsinglet analysis in [7] (see Figs. 9 and 10) are different from a variety of alternatives by much more than the uncertainties. Indeed, various gluon distributions, all obtained by fitting to small x HERA data [6, 22] are very different despite what is meant to be the main constraint on the data being the same in each case. It is particularly illustrative to look at the difference in the high-x gluons of MRST and Alekhin in Fig. 4. This is for NNLO, but is similar at NLO. Here the difference above $x = 0.2$ is a large factor, and very much bigger than each uncertainty (calculated using $\Delta\chi^2 = 1$ for Alekhin and $\Delta\chi^2 = 50$ for MRST.) It seems that the HERA data require a gluon distribution for the very best fit which is incompatible with the Tevatron jet data [23], and the standard error analysis does not accommodate this. As a further point, at NNLO one of the few hard cross-sections required in a global fit which is not fully known is that for the jet cross-section. It might be argued that one should leave the data out rather than rely on the NLO hard cross-section, as done by MRST. However, this correction is very likely to be $\sim 5\%$, whereas the change in the gluon distribution if the data are left out can be $> 100\%$. This implies, to the author at least, that it is better to include a data set relying on a slight approximation than to leave it out and obtain partons which are completely incompatible with it.

Even when similar data sets are fit, there can still be significant differences in parton distributions and their predictions.The prediction for σ_W at NLO at the LHC using CTEQ6.5 partons is 202 ± 9 nb and using MRST04 partons is 190 ± 5 nb. This is despite the rather similar data sets and procedures used in the two fits. The different predictions are easily explained by looking at the left of Fig. 5. The CTEQ gluon is much bigger than MRST at small x and drives quark evolution to be larger. This difference is not fully understood but is probably partially due to the fact the MRST have lower Q^2 cuts on the structure function data, and also due to the different input parameterisations for the gluon. MRST allow their gluon to be negative at small x at input ($Q_0^2 = 1\mathrm{GeV}^2$) while the CTEQ gluon is positive at small x input ($Q_0^2 = 1.69\mathrm{GeV}^2$), but is very small indeed. (Further analysis suggests a slightly negative input

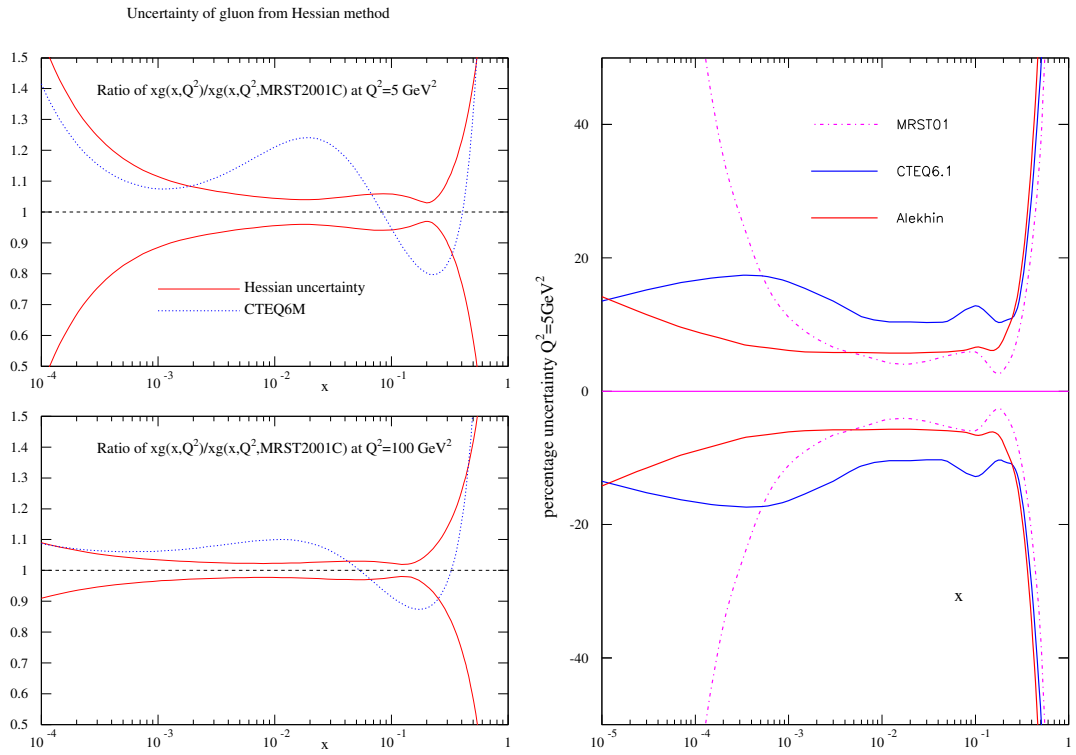

Fig. 5: The MRST gluon distribution with percentage uncertainties, and the central CTEQ distribution (left) and the uncertainties on the MRST, CTEQ and Alekhin gluon distributions at $Q^2 = 5\mathrm{GeV}^2$(right)

gluon is preferred, but only barely [12] so the freedom is not introduced.)

The parameterization can have an even more dramatic effect on the uncertainty than on the central value. The uncertainties on the gluon distributions for MRST, CTEQ and Alekhin are shown in the right of Fig. 5. One would expect the uncertainty to increase significantly at very small x as constraints die away. This happens for the MRST gluon. The Alekhin gluon does not have as much freedom, but is input at higher scales and behaves like $x^{-\lambda}$ at small x. The uncertainty is due to the uncertainty in λ (the situation is similar for ZEUS and H1 partons). The CTEQ input gluon behaves like x^{λ} at small x where λ is large and positive. The small-x input gluon is tiny and has a very small absolute error. At higher Q^2 all the uncertainty is due to evolution driven by the higher-x, well-determined gluon. The very small x gluon no more uncertain than at $x = 0.01 - 0.001$.

Another important source of uncertainty only now becoming clear is due to the strange distribution. Until recently this was taken to be a fixed and constant fraction of the total sea quark distribution. This did not allow any intrinsic uncertainty on the strange quark. It is now being fit more directly by comparison to dimuon data in neutrino scattering [13]. In the MSTW fits [14] this results in an increased uncertainty on all sea quarks since allowing the strange to vary independently gives the up and down quarks more freedom. CTEQ have produced specific parton sets with fits to the strange quark [24], and in Fig. 6 we see predictions from these for production of $W^+ + \bar{c}$. CTEQS0 represents the best fit when the strange is fit directly. Worryingly, this can be outside the uncertainty band for the default set.

3 Theoretical Uncertainties

Even if we had an unambiguous definition for the parameterization and the data sets and cuts used, there would still be additional uncertainties due to the limited accuracy of the theoretical calculations. The sources of theoretical error include higher twist at low scales and higher orders in α_S, and it now seems likely that there may be sizable corrections from higher order electroweak corrections at the LHC (see e.g. [25]), due to $\alpha_W \ln^2(E^2/M_W^2)$ terms in the expansion. The higher order QCD errors are due not only to fixed order corrections, but also to enhancements at large and small x because of terms of the form $\alpha_s^n \ln^{n-1}(1/x)$ and $\alpha_s^n \ln^{2n-1}(1-x)$ in the perturbative expansion. This means that renormalization

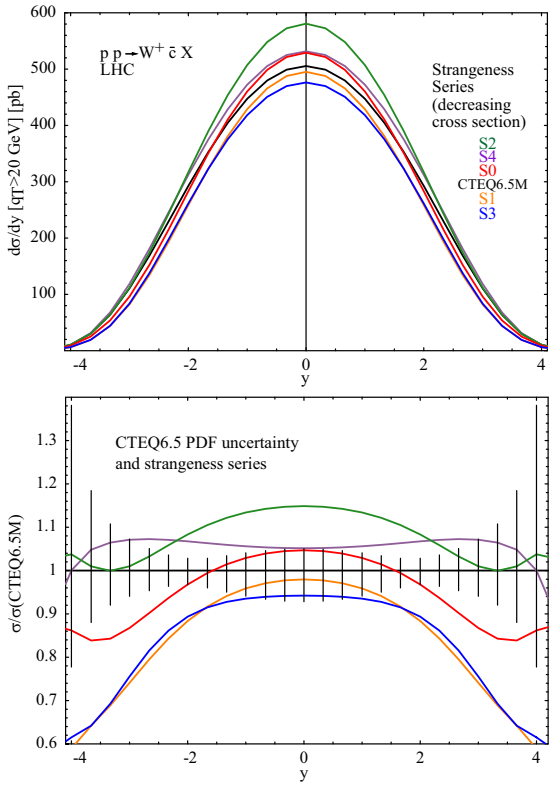

Fig. 6: Uncertainty of predictions for $W^+ + \bar{c}$

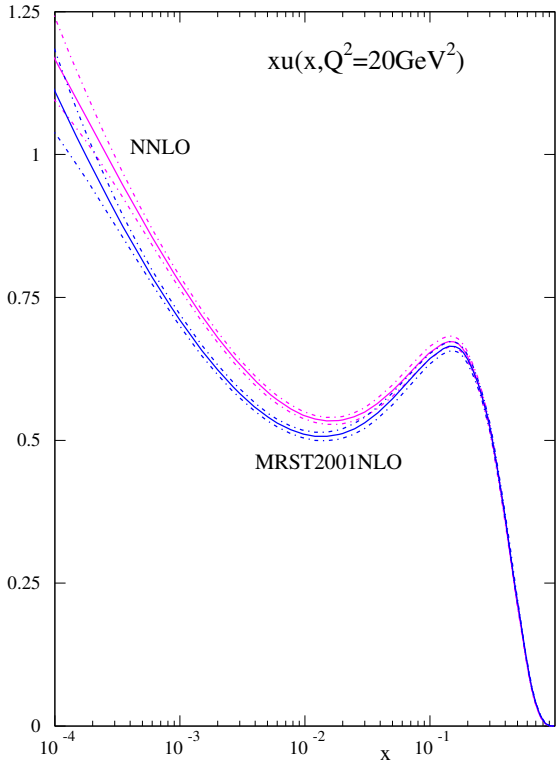

Fig. 7: Comparison between the NLO and NNLO up quark distribution

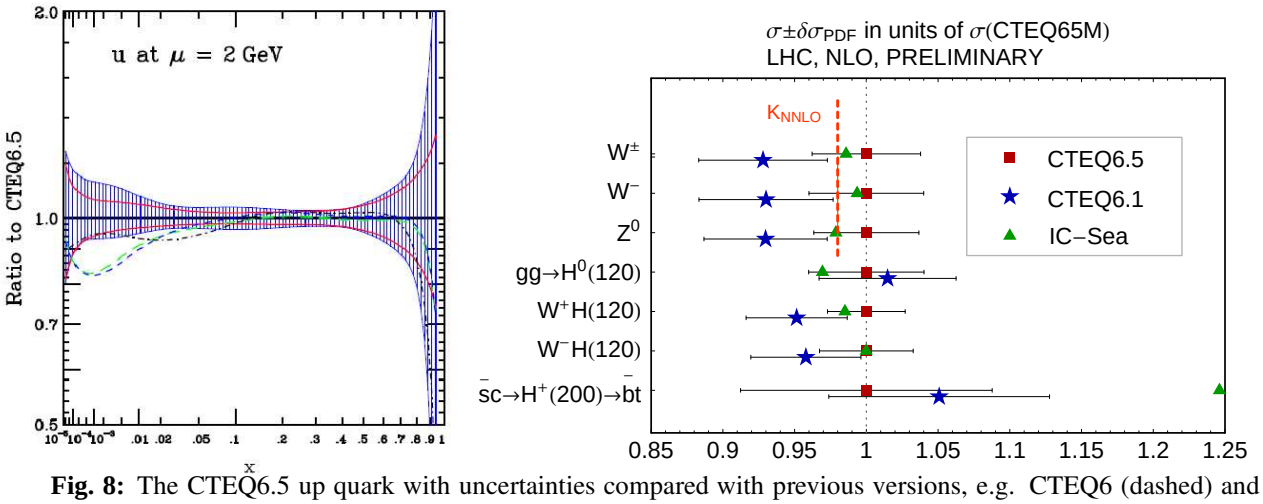

Fig. 8: The CTEQ6.5 up quark with uncertainties compared with previous versions, e.g. CTEQ6 (dashed) and MRST04 (dot-dashed) (left) and the change in predictions using CTEQ6.5 partons for LHC cross-sections as opposed to CTEQ6 (right)

and factorization scale variation are not a reliable way of estimating higher order effects because a scale variation at one order will not give any indication of an extra $\ln(1/x)$ or $\ln(1-x)$ at higher orders. Hence, in order to investigate the true theoretical error we must consider some way of performing correct large and small x resummations, and/or use what we already know about going to higher orders.

We are now able to look at the size of the corrections as we move from NLO to NNLO. The up quark distribution at the two orders is illustrated in Fig. 7. As one can see, the change in the central value is somewhat larger than the uncertainty due to the experimental errors. The predictions for various physical processes have been calculated. The change for quark-dominated processes, such as W and Z production, is not very large, i.e. 4% or less [26], but is sometimes bigger than the quoted uncertainty at each order. Changes in gluon dominated quantities, such as $F_L(x, Q^2)$, can be much larger [27]. Similarly there are implications that resummations may have significant effects on LHC predictions, particularly at high rapidity [28].

Very recently it has become clear that a less obvious source of theoretical errors can have surprisingly large effects, i.e. the precise treatment of heavy quark effects. For many years CTEQ have had a procedure for extrapolating from the limit where quarks are very heavy to the limit where they are effectively massless, i.e. a general-mass variable flavour number scheme (GM-VFNS) [29]. Nevertheless, they have chosen the scheme where the quark masses are zero as soon as the heavy quark evolution begins, i.e. zero-mass variable flavour number scheme (ZM-VFNS), to be the default parton set. In the most recent analysis [3] they have switched to the GM-VFNS definition as default and noticed that this has a very large effect on their small-x light quark distributions, mainly determined by fitting to HERA data, where mass corrections are important, and on LHC predictions. This is shown in Fig. 8, where one sees the prediction for σ_W increase by 8%.

Perhaps even more surprising is the change observed by MRST at NNLO. Because early approximate "NNLO" sets (e.g. [26]) were based on approximate splitting functions the MRST group used a (fully explained) approximate treatment of heavy quarks at NNLO, in particular not including the discontinuities at transition points that occur at this order [30]. The correction of this approximate NNLO VFNS between [2] and [10] using the scheme in [31] led to large corrections to the gluon distribution at small x and by evolution, also to the light quark distributions at higher scales, as seen in Fig. 9. This results in the corrections to LHC cross-sections shown in Table 1, i.e. up to 6%. In this case the change in procedure was less dramatic than that for the CTEQ6.5 result, where the original approximation was of massless quarks, and was also at one order lower. The size of the change was certainly unexpected. It is important to note that in both these cases the change is not really representative of an uncertainty, since each represents a correction of something that was known to be wrong. However, in each case the

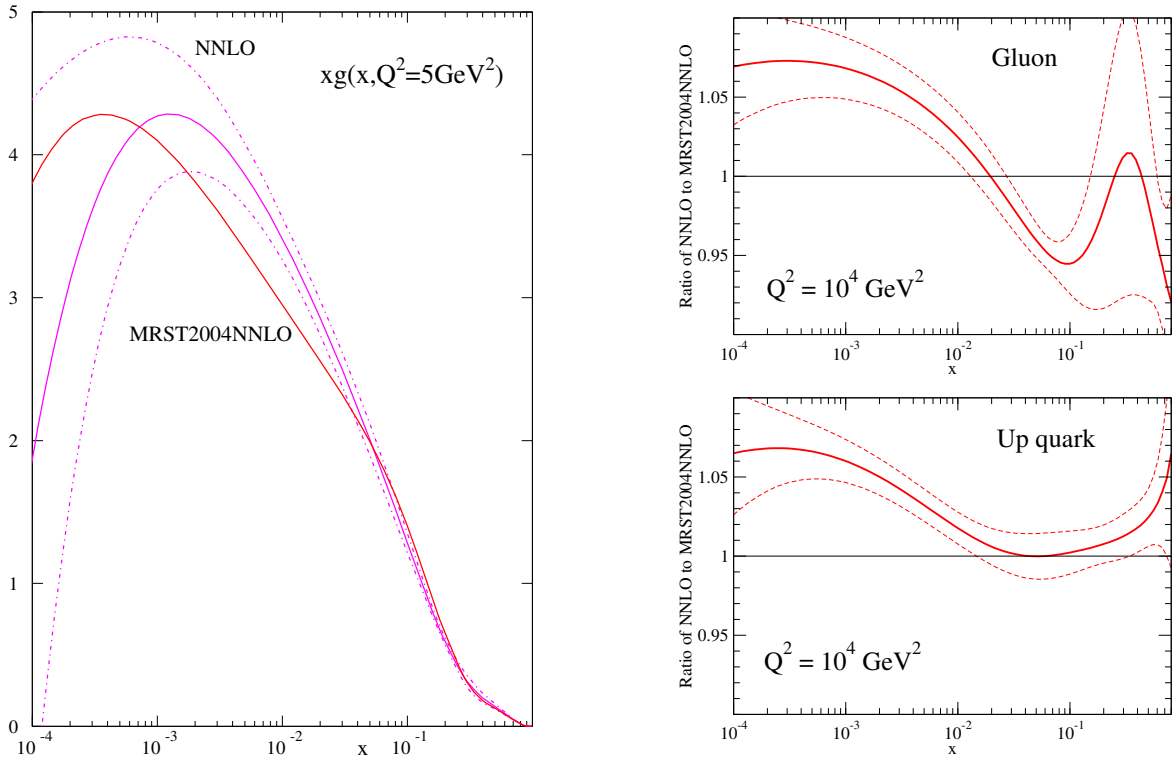

Fig. 9: Comparison of the NNLO gluon distribution (and its uncertainty) with the previous approximate NNLO distribution at $Q^2 = 5$ GeV^2 (left), and the ratio at $Q^2 = 10^4$ GeV^2 for both the gluon and the up quark (right).

Table 1: Total W and Z cross-sections multiplied by leptonic branching ratios at the Tevatron and the LHC, calculated at NNLO using the updated NNLO parton distributions. The predictions using the 2004 NNLO sets are shown in brackets.

	$B_{l\nu} \cdot \sigma_W$(nb)	$B_{l^+l^-} \cdot \sigma_Z$(nb)
Tevatron	2.727 (2.693)	0.2534 (0.2518)
LHC	21.42 (20.15)	2.044 (1.918)

"wrongness" was thought to be an approximation requiring only a small correction, an expectation that was optimistic. Some parton sets currently available are still extracted using similar (or worse) "approximations", and even in the best case the limited order of the calculation means that everything is to some extent an approximation, with the size of the correction being by definition uncertain.

4 Conclusions

One can determine the parton distributions from fits to existing data and predict cross-sections at the LHC. The fit quality using NLO or NNLO QCD is fairly good. There are various ways of looking at uncertainties due to the errors on data. For genuinely global fits, using $\Delta\chi^2 = 1$ is not a sensible option due to incompatibility between data sets and possibly between data and theory. Uncertainties due to parton distributions from experimental errors lead to rather small, $\sim 1-5\%$ uncertainties for most LHC quantities, and are fairly similar for all approaches. However, sometimes the central values using different sets differ by more than this. The uncertainties from input assumptions, e.g. cuts on data, sets used, parameterisations *etc.*, are comparable and sometimes larger than statistical uncertainties. In particular, the detail of uncertainties on the flavour decomposition of the quarks is still developing.

Uncertainties from higher orders/resummation in QCD are significant, and electroweak corrections are also potentially large at very high energies. At the LHC measurement at high rapidities, e.g. W, Z,

would be useful in testing our understanding of QCD. Our limited knowledge of the theory is often the dominant source of uncertainty. There has recently been much progress: more processes known at NLO, and some at NNLO; improved heavy flavours treatments; developments in resummations *etc.*. In particular, essentially full NNLO parton distribution determinations are now possible. But further theoretical improvements and complementary measurements are necessary for a full understanding of the best predictions and their uncertainties.

References

[1] R.S. Thorne, J. Phys. **G28** (2002) 2705 [arXiv:hep-ph/0205235].

[2] A.D. Martin, R.G. Roberts, W.J. Stirling and R.S. Thorne, Phys. Lett. **B604** 61 (2004).

[3] W.K. Tung, H.L. Lai, A. Belyaev, J. Pumplin, D. Stump and C.P. Yuan, JHEP **0702** (2007) 053.

[4] S. Alekhin, Phys. Rev. **D68** 014002 (2003).

[5] ZEUS Collaboration: S. Chekanov *et al.*, Eur. Phys. J. **C42** 1 (2005).

[6] H1 Collaboration: C. Adloff *et al.*, Eur. Phys. J. **C21** 33 (2001).

[7] J. Blümlein, H. Böttcher and A. Guffanti, Nucl. Phys. **B774** (2007) 182.

[8] L. Del Debbio, S. Forte, J.I. Latorre, A. Piccione and J. Rojo, JHEP **0703** (2007) 039.

[9] S. Moch, J.A.M. Vermaseren and A. Vogt, Nucl. Phys. **B688** (2004) 101; A. Vogt, S. Moch and J.A.M. Vermaseren, Nucl. Phys. **B691** (2004) 129.

[10] A.D. Martin, W.J. Stirling, R.S. Thorne and G. Watt, Phys. Lett. **B652** 292 (2007).

[11] S. Alekhin, K. Melnikov and F. Petriello, Phys. Rev. **D74** (2006) 054033.

[12] J. Huston, J. Pumplin, D. Stump and W.K. Tung, JHEP **0506** (2005) 080.

[13] NuTeV Collaboration: M. Goncharov *et al.*, Phys. Rev. **D64** (2001) 112006.

[14] R.S. Thorne, A.D. Martin, W.J. Stirling and G. Watt, arXiv:0706.0456 [hep-ph].

[15] D. Stump *et al.*, Phys. Rev. **D65** (2002) 014012.

[16] J. Pumplin, D.R. Stump and W.K. Tung , Phys. Rev. **D65** (2002) 014011.

[17] J. Pumplin *et al.*, Phys. Rev. **D65** (2002) 014013.

[18] A.D. Martin, R.G. Roberts, W.J. Stirling and R.S. Thorne, Eur. Phys. J. **C28** (2003) 455.

[19] C. Pascaud and F. Zomer 1995 *Preprint* LAL-95-05.

[20] M. Botje, Eur. Phys. J. **C14** (2000) 285.

[21] M. Dittmar *et al.*, "Parton distributions: Summary report for the HERA - LHC workshop," `hep-ph/0511119`.

[22] H1 Collaboration: C. Adloff *et al.*, Eur. Phys. J. **C13** 609 (2000); Eur. Phys. J. **C19** 269 (2001); S. Chekanov *et al.*, Eur. Phys. J. **C21** 443 (2001); Phys. Rev. **D70** 052001 (2004).

[23] D0 Collaboration: B. Abbott *et al.*, Phys. Rev. Lett. **86** (2001) 1707; CDF Collaboration; A. Abulencia *et al.*, Phys. Rev. **D75** 092006 (2007).

[24] H.L. Lai, P. Nadolsky, J. Pumplin, D. Stump, W.K. Tung and C.P. Yuan, JHEP **0704** (2007) 089.

[25] U. Baur, Phys. Rev. **D75** (2007) 013005, and references therein.

[26] A. D. Martin, R. G. Roberts, W. J. Stirling and R.S. Thorne, Phys. Lett. **B531** (2002) 216.

[27] A. D. Martin, W. J. Stirling and R. S. Thorne, Phys. Lett. **B635** (2006) 305.

[28] A. D. Martin, R. G. Roberts, W. J. Stirling and R. S. Thorne, Eur. Phys. J. **C35** (2004) 325.

[29] M. Aivazis *et al.*, Phys. Rev. **D50** 3102 (1994); W.K. Tung *et al.*, J. Phys. **G28** 983 (2002); S. Kretzer *et al.*, Phys. Rev. **D69** 114005 (2004).

[30] M. Buza *et al.*, Eur. Phys. J. **C1** (1998) 301.

[31] R. S. Thorne, Phys. Rev. **D73** (2006) 054019.

Weighting Background-Subtracted Events

James T. Linnemann[1] *(presenter)*
Andrew J. Smith[2]
[1]Michigan State University, 3245 BPS Building, E. Lansing, Michigan 48823
[2]Department of Physics, University of Maryland, College Park, MD 20742

Abstract
Often a full maximum likelihood (ML) estimate is inconvenient for computational reasons (e.g., iteration over large data sets). If a variable x is a discriminating variable ($s(x) \neq b(x)$), a weight function can be found which allows estimation of the number of signal events with a variance approaching that of a ML estimate of the same quantity. We derive a formula and discuss it in the context of more general results on event weighting from earlier papers by Barlow and Tkachov, which also find weighting out-performs cutting.

1 Introduction

The origin of this talk lies within the Milagro cosmic ray experiment [1]. However, the results apply just as well within LHC experiments, because both have to subtract backgrounds in order to see signals. Milagro's physics goal is to look at structure with TeV gamma rays, which are outnumbered by charged cosmic rays by 10^3. Thus our analysis procedures must calculate backgrounds correctly to 1 ppt. We make background-subtracted sky maps of measured photon excess $m = n - \hat{B}$. To enhance our statistical significance, we seek discriminator variables x whose probability density distributions differ for signal $s(x)$ and background $b(x)$. We then consider the *background subtracted* excess $m(x) = n(x) - \hat{B}b(x)$.

What is the best way to combine (say) bins of $m(x)$ for a best overall estimate of the excess? The naive solution is just to sum all the bins. My colleague Andy Smith argued you could do better by weighting the bins by the ratio of expected signal and background contributions to each bin

$$w(x) = E[S(x)]/E[B(x)] = K\ s(x)/b(x) = K\ r(x) \tag{1}$$

where K is a constant (independent of x) which can be ignored for calculating relative weights of bins of x. My first reaction was that this was cheating, since you've already used the expected background in the subtraction that led to the observed $m(x)$. But Andy was right!

2 Event Weighting

The underlying hypothesis is that the signal distribution across x bins is governed by an overall intensity M, and the signal distribution $s(x)$. The naive estimate (from x bins i) of M and its variance would be:

$$\hat{M_1} = \sum m_i;\ \ V[M_1] = \sum V(m_i) = \sum V_i \tag{2}$$

But this is not using the information about the expected relationship between bins, $s(x)$, which allows each bin to independently estimate M:

$$E[m_i] = Ms_i;\ \ \hat{M_i} = m_i/s_i;\ \ V[\hat{M_i}] = V_i/s_i^2 \tag{3}$$

How should these independent estimates $\hat{M_i}$ be best combined? The Best Linear Unbiased Estimate (BLUE) method invokes the Gauss-Markov theorem (James [2] §7.4.4) for the solution, which is to weight the estimates by the inverse of their variance, so that smaller-variance estimates are more heavily

weighted; the result has the minimum variance among the class of unbiased linear estimators (of which $\hat{M}_1$ is an inferior member).

$$\hat{M} = \sum \hat{M}_i w_i / \sum w_i = \sum (m_i s_i / V_i) / \sum (s_i^2 / V_i); \tag{4}$$

This solution holds for *any* distribution of uncorrelated, unbiased, variables. But it does require that the actual variance be used (and fluctuations to low estimated variance are particularly damaging [3]). $\hat{M}$ is identical to the minimum M found for

$$\chi^2 = \sum (m_i - M s_i)^2 / V_i \; . \tag{5}$$

$\hat{M}$ can also be regarded as applying for each event in bin i a weight u_i as defined below:

$$\hat{M} = k \sum (m_i s_i / V_i) = k \sum m_i u_i; \;\; u_i = s_i / V_i; \;\; 1/k = \sum (s_i^2 / V_i) \tag{6}$$

Finally, we arrive at Andy's weight by recognizing that for a large Poisson background, $V_i \approx B_i = B b_i$, so that $u_i = K' s_i / b_i$, as in Eq. 1. Milagro deals with billions of events, so it is a considerable advantage to sum event weights, rather than minimizing for each pixel of sky.

We can calculate the variance of the BLUE solution (using the definitions of k, u in Eq. 6):

$$V[\hat{M}] = k^2 \sum V[m_i] u_i^2 = k^2 \sum V_i u_i^2 = 1 / \sum (s_i^2 / V_i) = k \tag{7}$$

For sufficient data, the BLUE solution approaches the Cramer-Rao minimum variance bound (James §7.4.5). Because we can formulate the BLUE solution as event weighting (often referred to as the method of moments) we find that despite the "suboptimal" reputation of the method of moments (James §8.2.2), in this case it is competitive with ML (assuming Gaussian uncertainties).

3 Sensitivity to Assumptions

It is worth clarifying here how the solution depends on the assumptions made. Since we have independently normalised the shapes s, b, their absolute normalisation S, B does not matter. But we are sensitively dependent on the shapes $s(x),\ b(x)$. In Milagro we determine $b(x)$ from the data and can use it to check the simulations. But s comes from the simulation, and depends on the input shower physics, and (if the variable x is correlated with energy) on the assumed source energy spectrum. We test by comparing the MC $s(x)$ distribution with data.

4 Barlow's Event Weighting

Was the good performance of event weighting a fluke of this particular problem? Remarkably, no! In a 1987 paper, Barlow [4] found the *best* event weight function to count signal events. He first wrote the expected weight $E[w(x)]$ in terms of $s,\ b$ as

$$E[w_d] = (M\mu_s + B\mu_b)/N = (a\mu_s + \mu_b)/(a+1); \quad a = M/B = M/(N-M) \tag{8}$$

where μ_s, μ_b are the expected mean weights for signal and background. Substituting the observed data weight $\overline{w_d} = \sum w(x_j)/N$ for the expected and solving for M gives:

$$\widehat{M_B} = \sum (w_j - \mu_b)/(\mu_s - \mu_b) \tag{9}$$

These two equations are independent of the specifics of w, though $\mu_s,\ \mu_b$ depend on the form of w and its parameters. Eq. 9 makes it crystal clear that $\widehat{M_B}$ is unchanged by multiplicative *or* additive x-independent constants in $w(x) \to Cw(x) + D$.

The method of moments is quite general: calculate any function of the data, then solve for parameters, considering expected moments as functions $f(u)$ of the parameters of the true pdf. Eq. 9 is an example. Typically one chooses power moments $w' = x^n$ and hopes for the best.

4.1 Barlow's Optimal Weight

But Barlow did (much) better: he calculated a completely general equation for the variance $V[\widehat{M_B}]$ and used the calculus of variations to minimise $V[\widehat{M_B}]$ with respect to the function $w(x)$. He found the variance with the optimal weight function approached the ML variance (and Cramer-Rao bound), but unlike the ML solution, required no iteration through all the events! Further, the variance is *less* than the variance resulting from cutting on the optimal weight variable, though fitting to the distribution of $w(x)$ is also close to optimal. The optimum weight function Barlow found (after choosing a suitable normalization) was

$$w(x) = a_o s(x)/(a_o s(x) + b(x)) \text{ or} \tag{10}$$

$$w(x) = a_o r(x)/(1 + a_o r(x)); \qquad r(x) = s(x)/b(x) \tag{11}$$

Clearly $w \in [0, 1]$ (though Eq. 9 reminds us $\hat{M} \neq \sum w$). This optimal weight function should look remarkably familiar. The Neyman-Pearson lemma (James §10.3.1) tells us that the best variable for testing the hypothesis of whether an event is signal or background is $r(x) = s(x)/b(x)$; and as a result the best Bayesian discriminant (ideal neural net output) is the posterior signal probability $d(s|x) = as/(as + b)$, where $a = \pi_s/(1 - \pi_s)$ is the prior odds ratio.

There is a mild catch: one has to make an initial guess at M_o/B (why we wrote a_o instead of a). But the optimum is quadratic, so close ($a_o \approx a$) is quite good. Further, guessing is actually advantageous: it relieves you of iterating through the data. Since $E[w]$ already has a near-optimal dependence on the pdf parameters, and all you lose by the guess is a bit of variance increase, not a bias. However, wrong $s,\ b$ functions still give a biased $\hat{M}$ since you are fitting normalisation to an incorrect shape and $\mu_s,\ \mu_b$.

4.2 Comparison with BLUE Weight

Barlow notes that knowing the expected Poisson mean B reduces $V[\hat{M}]$, but finds the same $w(x)$ is optimal. When the fraction a_o of signal events is small, as it is in Milagro, then the Eq. 10 weight becomes $w \approx K''s/b$, showing the subtraction weight of Eq. 1 to be near-optimal for large backgrounds.

5 Tkachov Weights and the ML Solution

Barlow solved the specific problem of the best weight for separating signal and background. But Tkachov [5] later solved the more general problem of choosing the optimal $w(x)$ ("generalized moment") to estimate *any* pdf parameter u with minimum variance, again using the calculus of variations. His result is both more general, and simpler! Having *fixed* w, one estimates $\hat{u}$ through the dependence $E[w] = f(u)$ of the expected moment on the pdf parameters u, solving $\overline{w_d} = f(\hat{u})$. Functional differentiation relates the variance of $V[\hat{u}]$ in first order to the moment variance $V[w]$. More functional differentiation minimises $V[\hat{u}]$ wrt w, giving the optimum $w(x)$ choice, intimately related to the ML solution:

$$w_{opt}(x) = C(u)\ \frac{\partial Ln[p(x;u)]}{\partial u} + D(u) \tag{12}$$

Specializing to our case, $u = a$, and seeking $\hat{a}$ with $p = (as + b)/(1 + a),\ C = a,\ D = a/(1 + a)$, and $a \to a_o$ (our pre-data guess)

$$w = s/(as + b) - 1/(1 + a) \to w = a_o s/(a_o s + b)\ , \tag{13}$$

matching Barlow's Eq. 10 weight. The expected data weight is then

$$E[w_d(a_o, a)] = \int w(x; a_o)\ p(x; a)\ dx = \int \left(\frac{a_o s}{a_o s + b}\right)\left(\frac{as + b}{a + 1}\right)dx = \frac{a_o + (a - a_o)\mu_s}{a + 1} \tag{14}$$

which can be compared with the iterative ML solution written in terms of the weight function:

$$\sum_j \frac{\partial Ln[p(x_j; a)]}{\partial a} = 0 \Rightarrow \ \sum_j \frac{w(x_j, a_o \to a)}{Na} = \frac{1}{a + 1} \tag{15}$$

Tkachov shows that with the optimal weight function choice, no matter what the parameter, the method of moments gives a variance approaching the best possible.

One final comment: the optimal weight function of Eq. 10 is strongly reminiscent of the Wiener optimal frequency filter [6] with squared amplitudes (absolute power) instead of pdf's. The derivation minimises the reconstructed variance wrt the true signal, again using the calculus of variations.

6 Summary

A near-optimal weight can achieve near-ML accuracy. Weighting methods are powerful and simple. There is a rational scheme leading to choice of optimal weight (moment) functions. And both Barlow and Tkachov show that weighting (or fitting to a weight distribution) is more accurate (lower variance) than making cuts in even an optimal weight variable. A longer version of this paper is in preparation for submission to the Astrophysical Journal (and the arxiv server).

7 Acknowledgement

JTL thanks Harrison Prosper for having brought to my attention Refs. [4, 5] (years ago by now); it's a pleasure to tie them together. JTL also appreciated conversations with Sekhar Chivukula and Neil Christensen of MSU on some mathematical points.

References

[1] R. Atkins *et al.*, "Observation of TeV Gamma Rays from the Crab Nebula with Milagro Using a New Background Rejection Technique" Astrophysical Journal 595 (2003) 803-811; A. Abdo *et al.*, "TeV Gamma-Ray Sources from a Survey of the Galactic Plane with Milagro", Astrophysical Journal Letters 664 (2007) L91-L94.

[2] F. James, "Statistical Methods in Experimental Physics", World Scientific, 2006, §7.4.4; and J. Rice, "Mathematical Statistics and Data Analysis", 2nd ed, Duxbury 1995, §14.8.pr5; the latter generalizes the Gauss-Markov theorem to variables with unequal variance by dividing each variable by the square root of its variance, arriving at weighted least squares.

[3] L. Lyons, "Statistics for nuclear and particle physicists", Cambridge 1986, §1.6.(ii)

[4] R. Barlow, "Event Classification Using Weighting Methods", J. Comp. Phys 72 (1987) p202. Barlow uses N_S, $\bar{w}$, A where we use M, μ, $1/a$. He uses individual events j instead of bins i.

[5] F. Tkachov, "Approaching the Parameter Estimation Quality of Maximum Likelihood via Generalized Moments" Part. Nucl. Lett. 111(2002) 28 or arXiv:physics/0001019; arXiv:hep-ph/0210116; arXiv:physics/0604127. Tkachov uses f, π, P where we use w, p, a. Derivatives of moments wrt pdf parameters a receive no contribution from $w(x, a_o)$ because the choice of a_o has no functional dependence on a, though one finds the optimal *value* of a_o is a. Tkachov prefers $E[w] = 0$, e.g. $C = 1,\ D = 0$ in Eq. 13 rather than $w \in [0, 1]$. A caution: the 2nd reference derives Eq 10 but is cavalier about the normalisation of $p(x)$, which would give the wrong ML solution.

[6] W. Press *et al.*, Numerical Recipes in C++, Cambridge 2002, §13.3. JTL thanks Prof. Igor Volobouev for pointing out the resemblance.

Subtracting and Fitting Histograms using Profile Likelihood

F.M.L. de Almeida Jr. and A.A. Nepomuceno
Instituto de Física, Universidade Federal do Rio de Janeiro,RJ,Brazil

Abstract
It is known that many interesting signals expected at LHC are of unknown shape and strongly contaminated by background events. These signals will be difficult to detect during the first years of LHC operation due to the initial low luminosity. In this work, one presents a method of subtracting histograms based on the profile likelihood function when the background is previously estimated by Monte Carlo events and one has low statistics. Estimators for the signal in each bin of the histogram difference are calculated so as limits for the signals with 68.3% of Confidence Level in a low statistics case when one has a exponential background and a Gaussian signal. The method can also be used to fit histograms when the signal shape is known. Our results show a good performance and avoid the problem of negative values when subtracting histograms.

1 Introduction

The search for signals of low statistics has led to a strong development on statistics methods for high energy physics. Recently, methods based on profile likelihood has been widely used in problems related to setting limits to a signal and to test hypotheses. This approach shows very good performance in extracting signal information in the presence of nuisance parameters [1].

In this work one considers a χ^2-function obtained from the profile likelihood for subtracting histograms where the signal to backgrounds ration is small, and both distribution have unknown shape. One also shows that, when the signal distribution is known, one can use this χ^2-function to fit the signal without fitting the background. It is presented in the next Section the road map to this new χ^2-function. Section 3 presents the results for extracting signal information by subtracting histograms, and limits to signal are computed using the proposed χ^2-function. Section 4 shows an example on the fit method.

2 Likelihood and Profile Likelihood

Let us assume a counting experiment such that the signal and background events are completely independent and both obey to Poisson distributions. The background events are first estimated using the Monte Carlo method, running the experiment in "idle" mode or by any other technique. Suppose that during the experiment k data events are obtained and that m background events were previously estimated using Monte Carlo(MC) techniques. Since the number of previously estimated MC events depends on computational resources, it is possible to generate τ samples, such that

$$\tau = \mathcal{L}_{\mathcal{MC}}/\mathcal{L}_{\mathcal{EXP}}, \tag{1}$$

where $\mathcal{L}_{\mathcal{EXP}}$ and $\mathcal{L}_{\mathcal{MC}}$ are the experimental and MC luminosities, respectively, and $\tau > 0$. When one has limited computer resources, τ may be restricted τ to the range $0 < \tau < 1$. Any information about the background is helpful in order to extract as clean a signal as possible. The likelihood corresponding to the above discussion is

$$L(s, b; k, m, \tau) \propto (s+b)^k e^{-(s+b)} (\tau b)^m e^{-\tau b}, \tag{2}$$

where s and b are related to the signal and background distributions, respectively.

To obtain a b independent likelihood, one can find the maximum likelihood estimator of the background $\widehat{b}$ as a function of s and replace the true value b by $\widehat{b}$ in Eq. (2). Taking the derivative of that equation, and solving it for $b \geq 0$, one gets

$$\widehat{b}(s) = max\left(0, \frac{k+m-(1+\tau)s+\Delta(s)}{2(1+\tau)}\right), \tag{3}$$

where

$$\Delta(s) = \sqrt{[k+m-(1+\tau)s]^2 + 4m(1+\tau)s} \geq 0. \tag{4}$$

Replacing b by $\widehat{b}(s)$ in Eq. (2), one obtains the profile likelihood $L_P(s;k,m,\tau)$, which does not dependent on b [2].

$$L_P(s;k,m,\tau) \propto (s+\widehat{b}(s))^k e^{-(s+\widehat{b}(s))}(\tau\widehat{b}(s))^m e^{-\tau\widehat{b}(s)}. \tag{5}$$

The maximum value of L_P and the most probable value of s, $\widehat{s}$, are obtained by solving Eq (5). The simple analytical solution for $\widehat{s}$ is an unbiased value

$$\widehat{s} = max\left(0, k - \frac{m}{\tau}\right), \tag{6}$$

since $s \geq 0$ due to physical constrains. The parameter $\widehat{s}$ is just the maximum profile likelihood estimator of s.

Let us construct now an approximate χ^2-function using Eq (5). The maximum profile likelihood ratio is given by

$$\lambda_P = \frac{L_P(s,k,m,\tau)}{L_P(\widehat{s},k,m,\tau)}, \tag{7}$$

where the denominator is the maximum profile likelihood, which occurs when $s = \widehat{s}$. According to the maximum likelihood ratio theorem $\chi^2_P \approx -2\log\lambda_P$ and hence, the profile χ^2_P-function is written as

$$\chi^2_P = 2\left\{(s-\widehat{s}) + (\tau+1)\left(\widehat{b}(s) - \widehat{b}(\widehat{s})\right) + k\ln\left(\frac{\widehat{s}+\widehat{b}(\widehat{s})}{s+\widehat{b}(s)}\right) + m\ln\left(\frac{\widehat{b}(\widehat{s})}{\widehat{b}(s)}\right)\right\}, \tag{8}$$

where $\widehat{b}(s)$ and $\widehat{s}$ are given by Eqs (3,6), respectively, so as $\widehat{b}(\widehat{s})$.

3 Subtracting Histograms and Setting Limits

In order to show the applicability of the χ^2_P-function obtained we generated 500 Toy Monte Carlo events, such that 50 were signal and 450 were background, distributed in a histogram of 50 bins. The signal and background were generated according to Gaussian and Exponential functions, respectively,

$$S \sim Gauss(1.2, 0.2),\ B \sim Exp(-x). \tag{9}$$

The number of background events in each bin was previously estimated by generating 2250 background events, corresponding to $\tau = 5$. It is useful to mention at this point that there is no advantage in taking $\tau > 5$ when one estimates the background from MC, since there is no relevant change in the χ^2_P-function for $\tau > 5$. Figure 1 shows the 'data', the background previously estimated and the signal. To

extract the signal histogram from the 'data', one can use Eq (6), which give us the signal estimated for each bin. Its limits (s_{min}, s_{max}) are obtained by solving the system

$$\begin{cases} \int_{s_{min}}^{s_{max}} f_P(s;k,m,\tau)ds = 1-\alpha \\ \\ \chi^2_P(s_{min}) = \chi^2_P(s_{max}) \\ \\ 0 \leq s_{min} < s_{max} \end{cases} \tag{10}$$

where $f_P(s;k,m,\tau)$ is the normalized probability distribution of s given k, m and τ obtained normalizing $L_P(s;k.m,\tau)$ with respect to s, and α depends on the chosen confidence level.

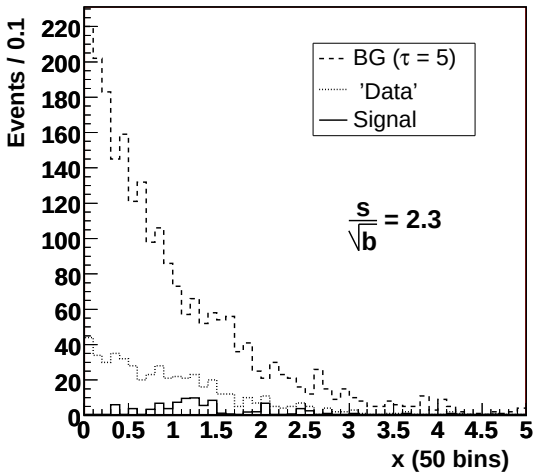

Fig. 1: Toy Monte Carlo Example. The full line represents the signal contained in the 'data'. The background was previously estimated with $\tau = 5$.

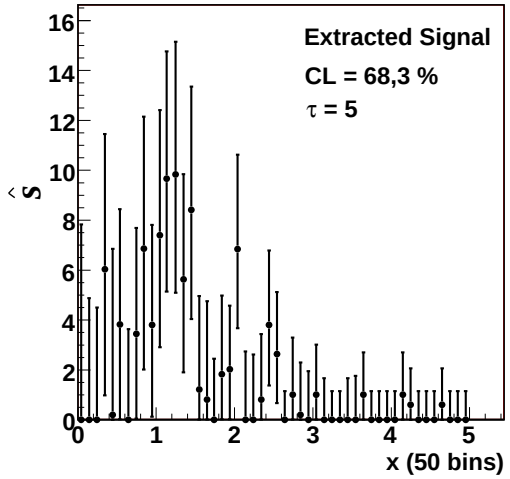

Fig. 2: Extracted signal. The signal limits were calculated for a confidence level of 68.3%. The constraining $\widehat{s} > 0$ avoid bins with negative values.

The subtracted histogram result is shown in Fig. 2. The points are the signal estimated for each bin and the error bars were calculated using Eq (10) for a confidence level of 68.3%. Notice that we have no bin with negative values due the constraint $\widehat{s} > 0$. It is important to mention also that one did not need to know the true background rate b in order to get signal limits, since the χ^2_P-function, given by Eq. (8), does not depend on that parameter since it has been replaced by an estimate.

The signal significance can be obtained by looking at the *P*-value under the hypothesis that one has no signal. Taking into account just the bins between $x = 0.85$ and $x = 2.5$, one gets a *P*-value of 0.022.

4 Fitting Histograms

When the signal shape is known, one can use Eq. (8) to fit histograms. In such case, the χ^2_P-function that will be minimized is given by the sum of all $\chi^2_{P_i}(s_i, k_i, m_i, \tau)$ which correspond to N bin contributions, where s_i must be substituted by the function $f(x_i, \boldsymbol{\theta})$ to be fit, x_i being the corresponding ordinate in the i^{th} bin and $\boldsymbol{\theta}$ the parameter vector to be fitted.

One can apply this approach to fit the signal in the Monte Carlo sample shown before, but now the events are distributed in a histogram of 100 bins, since one knows now the signal distribution shape, as shown in Fig. 3. The number of previously estimated background events in each bin m_i is given by the

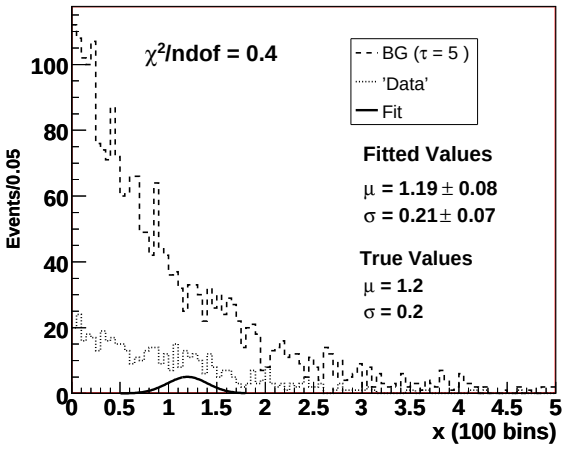

Fig. 3: Previously estimated background, 'data' and fitted curve.

histogram labeled BG in Fig. 3, and k_i is the number of 'data' events in each bin. The signal distribution s_i is substituted by a Gaussian function, and $\tau = 5$. By the minimization of the χ^2_P-function, one gets the fitted parameters $\mu = 1.19 \pm 0.08$ and $\sigma = 0.21 \pm 0.07$, which are in very good agreement with the "true" values 1.2 and 0.2, respectively. The full line in Fig. 3 shows the fitted curve.

Notice that as Eq. (8) depends just on $f(x_i, \boldsymbol{\theta})$, k_i, m_i and τ, one did not need to fit the background distribution, and the only necessary information from background was its number of events m_i estimated by MC. This is the great advantage of this method. The χ^2_P-function already incorporates the background statistical fluctuations. Besides reducing the numbers of fitted parameters, this method presents no problems when one has few or no events in one or more bins as can occur in data with long tails. Even the bins with $k_i = 0$ and/or $m_i = 0$ contributes to the χ^2_P-function. It is is only necessary to fit the signal function parameters which will allow us to obtain a much cleaner and less noisy analysis. This will affect in a positive way the parameter covariance matrix. A systematic study of this method was done for different τ values and different signal and background distributions, and in all cases the method showed very good performance.

5 Summary

The proposed χ^2_P-function can be used to extract signal information without need to know the background distribution shape. The fact that one just needs to fit the signal reduce the number of parameters to be fitted and avoid the uncertainties carrying by the lack of knowledge of the exact background parameters. The method works well even in situation where there is very low statistics.

Acknowledgments

This work was partially supported by the Brazilian Agencies CNPq and CAPES and the HELEN project.

References

[1] W. A. Rolke, A.M. Lopez, Nuclear Instruments and Methods **A551** (2005) 493.

[2] W. A. Rolke, A.M. Lopez, Nuclear Instruments and Methods **A458** (2001) 745.

SFitter: Determining Supersymmetric Parameters

Rémi Lafaye[1], *Michael Rauch*[2], *Tilman Plehn*[2] *and Dirk Zerwas*[3]
[1]LAPP Annecy
[2]University of Edinburgh
[3]LAL Orsay

Abstract
If supersymmetry (or a similar complex phenomenon) is found at the LHC, the goal for all colliders over the coming decades will be to extract the fundamental parameters of an underlying model from the measurements. Dedicated state–of–the–art tools will be necessary to link a wealth of measurements to an e.g. 20–dimensional MSSM parameter space. Starting from a general log–likelihood function of this high-dimensional parameter space we show how we can find the best–fit parameter values and determine their errors. Beyond a single best–fit point we illustrate how distinct secondary minima occur in complex parameter spaces. In cases where there are flat dimensions in the likelihood we comment on the benefits and limitations of marginalizing over additional dimensions.

1 Introduction

The LHC start is now a matter of months, and high energy physicists are eager to see the first sign of a Higgs boson or any alternative to such a fundamental particle. However, fundamental scalars naturally lead to the existence of an ultraviolet completion of the the Standard Model. Such an extension of the current Standard Model, might even at the same time solve the second main mystery of high–energy physics, the existence of cold dark matter.

Supersymmetry is one appealing extension of the Standard Model and is already constrained by previous experiments such as LEP and Tevatron. The LHC era might give many hints about new–physics scenarios and it will certainly rule out large classes of extensions to the Standard Model. However, it will not give us anything like a one–to–one map between a limited number of observables and a well–defined small set of parameters.

The SFitter[1] program aim at extracting parameters compatible with the available observables, while relying on as few assumptions as possible. For example, assuming an MSUGRA scenario, the LHC might provide sufficient measurements to give good uncertainties on the fundamental parameters, with the use of well–known fitting techniques. But to test the assumption of the GUT scale unification, one needs to scan a full 20 parameter space at the TeV scale. This task requires a subtle and careful scanning of the parameter space, and is better performed with the use of techniques such as Markov chains.

2 Fitting principles

The SFitter program uses as input high–energy physics experimental data and low–energy scale constraints and compare them to theoretical predictions to compute a likelihood value. Theoretical quantities are computed to the highest order (NLO in most cases) available thanks to SUSY spectrum calculators[2, 3]. Other quantities, such as cross–sections, branching ratios and dark matter relic density can then be derived using other available programs [4, 5, 6].

The likelihood can be computed using two different scheme, depending on the final use:

– The RFit scheme as defined by Höcker et al.[7] should be used for the sake of correctness in frequentist analysis. In this case, the theoretical errors are interpreted as a lack of knowledge on a

parameter. The combined likelihood (including both experimental uncertainty σ_{exp} and theoretical error σ_{th}) is defined as:

$$-2\ln\mathcal{L} = \begin{cases} 0, & \forall\, |x_{exp} - x_{th}| < \sigma_{th} \\ \left(\frac{|x_{exp}-x_{th}|-\sigma_{th}}{\sigma_{exp}}\right)^2 & \forall\, |x_{exp} - x_{th}| \geq \sigma_{th} \end{cases} \tag{1}$$

– The standard convolution of experimental and theoretical uncertainties. This is done assuming the theoretical error has a probability density function (either flat or Gaussian) following Bayesian statistics.

The region of the parameter space in which the likelihood is to be computed depends greatly on the fitting techniques used. Ideally, a scan covering the whole parameter space region of interest, is performed to find local likelihood maxima. Then, a gradient fit around each maximum allows to determine the parameter values and errors.

The parameter space region of interest definition and how the scan is performed relies on priors. There is no way around this. But hopefully if the likelihood shape is sufficiently smooth compared to the scan steps, the gradient fit will converge to the true minima. The main problem with a 20–dimensional parameter space (like with the phenomenological MSSM) is to perform an efficient scan. Efficient in terms of coverage and computing time. This is were the Markov chains come into play.

SFitter provides all relevant frequentist or Bayesian answers in three steps: first (1), we compute a likelihood map of the entire parameter space, using either a simple grid method or a Markov Chains approach (described later). This map is completely exclusive, i.e., it includes all dimensions in the parameter space. Then (2), we rank the best local likelihood maxima in the map according to their log–likelihood values. This way we identify the global maximum, and everybody can include their personal prior towards secondary maxima (i.e., SUSY breaking scenario), without mistaking such a prior for actual likelihood. Last (3), we compute likelihood or probability maps of lower dimensionality, down to one–dimensional distributions, by properly removing or marginalizing unwanted parameter dimensions.

3 Markov Chains

A Markovian process is defined as a stochastic process in which the conditional probability distribution of future states depends only on the present state and not on any past state.

For the purpose of fitting the state is defined as a point in the parameter space and its associated likelihood value. The future parameter values (*new*) are then chosen according to the current position (*cur*) and kept (as the next point of the chain) if it satisfies one of the two following conditions:

$$\begin{cases} \mathcal{L}_{new} > \mathcal{L}_{cur} \\ \texttt{Random}[0,1] < \frac{\ln\mathcal{L}_{cur}}{\ln\mathcal{L}_{new}} \end{cases} \tag{2}$$

If the new point is chosen randomly over the whole parameter space without any dependence on the current point, then the Markov chain is equivalent to a Monte–Carlo fitting method. The main drawback in this case is that high likelihood regions will not be favored with respect to low likelihood ones unless we a–priori know the likelihood shape and can generate new points accordingly (as in Monte–Carlo resonant process generation). Another way to improve the Markov chains efficiency is to generate the new point depending on the current point. In SFitter, this is done using a Landau distribution separately for each parameter. The Landau peak is taken as the current value of the parameter and the distribution extends to the parameter limits.

The main advantage of the Markov chains method compared to a crude scan is its convergence speed, which can go linearly with the number of parameters. Indeed, parameters which have no influence on the likelihood value do not slow down the convergence process. Also, it does not rely on the likelihood

χ^2	m_0	$m_{1/2}$	$\tan\beta$	A_0	μ	m_t
0.09	102.0	254.0	11.5	-95.2	$+$	172.4
1.50	104.8	242.1	12.9	-174.4	$-$	172.3
73.2	108.1	266.4	14.6	742.4	$+$	173.7
139.5	112.1	261.0	18.0	632.6	$-$	173.0
...						

Table 1: SFitter output for MSUGRA in SPS1a. List of the best log–likelihood values over the MSUGRA parameter space. All masses are given in GeV.

shape in the parameter space, and like any other scanning process, it has the ability to find secondary minima. However, it should be used cautiously as it is not meant to find the exact value of the minima, and a bad choice of priors can lead to scans on limited parameter space regions. Theoretically increasing the number of points in the chain can overcome these problems. Alternatively, one can try different priors to make sure the whole parameter space is correctly scanned and then use a gradient fit to find the exact minima.

4 MSUGRA as a toy model

Clearly, in the LHC era no model for supersymmetry breaking should be assumed for analysis. Instead, the breaking mechanism should be inferred from data. At the LHC (and certainly in combination with the ILC) there is little need for top–down analysis, which are known to reveal more about their author's imagination than about physics. As a matter of fact, supersymmetry should only be considered one possible interpretation of for example cascade decays. However, completely generalizing an intelligent analysis to a general new–physics model space seems not viable at the moment. We will therefore assume that supersymmetry, little–Higgs models or extra–dimensional models can be distinguished by simpler hypothesis testing.

Running the Markov chains algorithm on the MSUGRA SPS1a point[9] using 300 fb^{-1} LHC toy data[8], we obtain the results summarized in Table 1. As discussed before this ranked list of likelihood maxima has to be refined with a gradient fit. Building a profile likelihood map (i.e., looking for the maximum likelihood in all directions but the ones of interest) can be compared to the Bayesian approach (marginalizing over all other dimensions to obtain a probability density function). The two illustrations shown in Fig. 1 look similar. However, there are two differences in the details: first, the area around the true parameter point is less pronounced in the Bayesian pdf, compared to the profile likelihood. When we integrate over a direction in parameter space we largely collect noise from regions with small likelihood. This noise washes out the peaked structures. The second effect is the more pronounced branch structure for the Bayesian pdf, while in the profile likelihood the area between the two branches is filled by single good parameter points in the parameter projected away; the marginalization provides us with 'typical' likelihood values in this region which in general does not fit the data well.

The washing–out effect also smears considerably the one–dimensional Bayesian pdf distribution in m_0 as shown in Fig. 2. But here the marginalization over $m_{1/2}$ also creates a higher peak at $m_0 = 50$ GeV, which should only be interpreted as higher likelihood density. The only reliable source of information on likelihood value being the profile map.

5 Conclusions

It will be hard for the LHC to give a conclusive answer to the crucial question, namely what is the 'correct' ultraviolet completion of the Standard Model. The impact of the LHC on the vast model space will at best be locally conclusive. Be it from a frequentist or a Bayesian point of view, the use of improved fitting techniques, such as used in SFitter, will prove to be very useful. At least until we reach sufficient

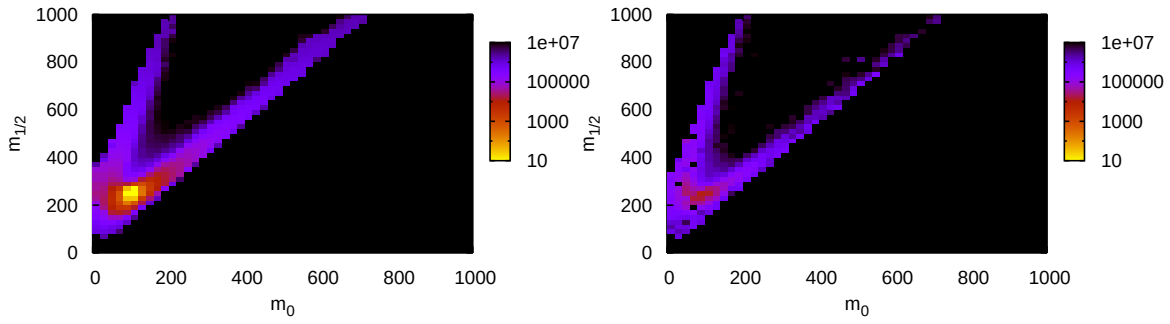

Fig. 1: SFitter output for MSUGRA in SPS1a. Left: two–dimensional profile likelihood χ^2 over the m_0–$m_{1/2}$ plane. Right: two–dimensional Bayesian pdf χ^2 over the m_0–$m_{1/2}$ plane marginalized over all other parameters. All masses are given in GeV.

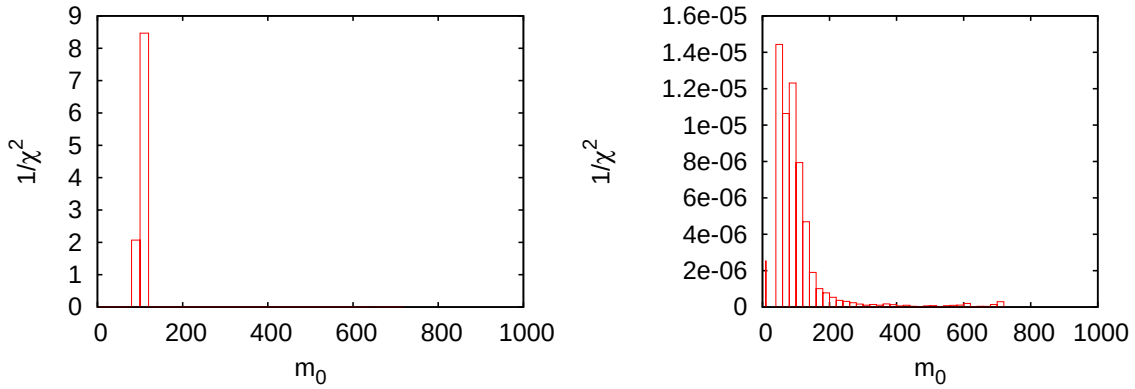

Fig. 2: SFitter output for MSUGRA in SPS1a. Left: one–dimensional profile likelihood $1/\chi^2$ for m_0. Right: one–dimensional Bayesian pdfs $1/\chi^2$ for m_0. All masses are given in GeV.

experimental and theoretical precision to narrow down the *Next Standard Model* parameters.

References

[1] R. Lafaye, T. Plehn, M. Rauch and D. Zerwas, arXiv:0709.3985 [hep-ph].

[2] A. Djouadi, J. L. Kneur and G. Moultaka, arXiv:hep-ph/0211331.

[3] B. C. Allanach, Comput. Phys. Commun. **143**, 305 (2002).

[4] W. Beenakker, R. Höpker, M. Krämer, T. Plehn, M. Spira, P. M. Zerwas, Nucl. Phys. B **492**, 51 (1997); Nucl. Phys. B **515**, 3 (1998); Phys. Rev. Lett. **83**, 3780 (1999); arXiv:hep-ph/9809319.

[5] A. Djouadi, M.M. Muhlleitner and M. Spira CERN-PH-TH/2006-200

[6] G. Bélanger, F. Boudjema, A. Pukhov and A. Semenov Comput.Phys.Commun.176:367-382,2007

[7] J. Charles, A. Höcker, H. Lacker, S. Laplace, F. Le Diberder, J. Malcles, J. Ocariz, M. Pivk, L. Roos Eur. Phys. J. C41, 1-131 (2005)

[8] G. Weiglein et al. [LHC/LC Study Group], arXiv:hep-ph/0410364.

[9] B. C. Allanach et al., in Proc. of the APS/DPF/DPB Summer Study on the Future of Particle Physics (Snowmass 2001) ed. N. Graf, Eur. Phys. J. C 25 (2002) 113

A Bayesian Approach to the Constrained MSSM

Leszek Roszkowski,[1] *Roberto Ruiz de Austri*[2] *and Roberto Trotta*[3]
[1]Department of Physics and Astronomy, University of Sheffield, Sheffield S3 7RH, UK
[2]Departamento de Física Teórica C-XI and Instituto de Física Teórica C-XVI, Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain
[3]Astrophysics Department, Oxford University, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK

Abstract
We present a new analysis of the Constrained MSSM in terms of Bayesian statistics. We illustrate our results with the light Higgs boson whose inferred mass range one should be able to exclude at the Tevatron with high confidence.

1 Introduction

Softly-broken low-energy supersymmetry (SUSY) offers a promising framework within which many questions challenging particle physics and cosmology, such as the hierarchy problem or the nature of dark matter, can be addressed. Despite many attractive features, without a reference to grand (or string) unification, SUSY models suffer from the lack of predictivity due to a large number of free parameters (e.g., over 120 in the Minimal Supersymmetric Standard Model (MSSM)). The MSSM with one particularly popular choice of universal boundary conditions at the unification scale is called the Constrained MSSM, or CMSSM [1]. The CMSSM is defined in terms of five free parameters: common scalar (m_0), gaugino ($m_{1/2}$) and tri-linear (A_0) mass parameters (all specified at the unification scale), plus the ratio of Higgs vacuum expectation values $\tan\beta$ and sign(μ), where μ is the Higgs/higgsino mass parameter. The economy of parameters makes the CMSSM a useful framework for exploring SUSY phenomenology.

Many studies have explored the CMSSM or other SUSY models, mostly by evaluating the goodness-of-fit of points scanned using fixed grids in parameter space. However, this approach has a number of severe limitations. Firstly, the number of points required scales as k^N, where N is the number of a model's parameters and k the number of points for each of them, making the approach highly inefficient for exploring with sufficient resolution parameter spaces of even modest dimensionality, say $N > 3$. Secondly, narrow "wedges" and similar features of parameter space can easily be missed by not setting a fine enough resolution (which, on the other hand, may be completely unnecessary outside such special regions). Thirdly, extra sources of uncertainties (e.g., those due to the lack of precise knowledge of SM parameter values) and relevant external information (e.g., about the parameter range) are difficult to accommodate in this scheme.

Here we present a different approach, encoded in the publicly available package SuperBayes [2]. It is based on Bayesian statistics and Markov Chain Monte Carlo scanning methods. After introducing our procedure we will present our results obtained in the framework of the CMSSM. In particular we focus on the lightest Higgs boson h^0. We also comment on prospects for superpartner searches at the LHC and on direct neutralino dark matter detection. We refer the reader to [3, 4, 5] for a detailed presentation. The Bayesian approach has several technical and statistical advantages over the more traditional fixed-grid scan technique, the most important being perhaps the ability to incorporate all relevant sources of uncertainties, e.g., the residual uncertainty in the value of SM parameters. This means that the inferred high probability regions of the CMSSM parameters (or resultant observables) take fully into account all sources of uncertainty relevant to the problem. For other recent works applying a similar approach to the CMSSM, see [6, 7, 8, 9].

2 Parameter space, priors and data used

We consider the 8 dimensional parameter space $m = (\theta, \psi)$, where $\theta = (m_0, m_{1/2}, A_0, \tan\beta)$ is a vector of CMSSM parameters, while $\psi = (M_t, m_b(m_b)^{\overline{MS}}, \alpha_{\rm em}(M_Z)^{\overline{MS}}, \alpha_s(M_Z)^{\overline{MS}})$ is a vector of relevant SM parameters, where M_t is the pole top quark mass, $m_b(m_b)^{\overline{MS}}$ is the bottom quark mass at m_b, and $\alpha_{\rm em}(M_Z)^{\overline{MS}}$ and $\alpha_s(M_Z)^{\overline{MS}}$ are the electromagnetic and the strong coupling constants at the Z pole mass M_Z. The last three quantities are evaluated in the $\overline{MS}$ scheme. Since we are only interested in the effect of the residual uncertainty in the experimental determination of the SM parameters on our observables (see below), we treat them as "nuisance parameters" and at the end we integrate them out from our probability distribution function (pdf). It turns out that including them has an important impact in widening high probability regions of the CMSSM parameters.

In Bayesian statistics the posterior probability distribution $p(m|d)$ is computed using the Bayes theorem, $p(m|d) = p(d|m, f(m))\pi(m)/p(d)$. The *likelihood* $p(d|m, f(m))$ supplies the information provided by the data, by comparing the base parameters m or any derived function $f(m)$ to the data d. The quantity $\pi(m)$ denotes a *prior probability density function* (hereafter called simply *a prior*) which encodes our state of knowledge about the values of the parameters before we see the data. Here we first take the prior to be flat (i.e., constant) in the variables m; below we specify their ranges. If the constraining power of the likelihood is strong enough to override the choice of the prior, than the latter does not matter in the final inference based on the posterior pdf. We have adopted a wide prior region of up to 4 TeV for $m_0, m_{1/2}$ (in order to include the so-called "focus point" (FP) region at large m_0), $|A_0| \leq 7$ TeV and $2 \leq \tan\beta \leq 62$. The prior range on the nuisance parameters does not influence the final results, since the SM parameters are rather tightly constrained by the data: $M_t = 171.4(2.1)$ GeV , $m_b(m_b)^{\overline{MS}} = 4.20(0.07)$ GeV, $\alpha_s(M_Z)^{\overline{MS}} = 0.1176(0.002)$ and $1/\alpha_{\rm em}(M_Z)^{\overline{MS}} = 127.955(0.018)$.

In our analysis, for each choice of m we compute a series of derived observable quantities $f(m)$. We list them here along with their experimental values and (estimated) theoretical errors, which are added in quadrature: the W gauge boson mass $M_W = 80.392(0.029)(0.015)$ GeV, the effective leptonic weak mixing angle $\sin^2\theta_{\rm eff} = 0.23153(0.00016)(0.00015)$, a SUSY contribution to the anomalous magnetic moment of the muon, $\delta a_\mu^{\rm SUSY} = a_\mu^{\rm expt} - a_\mu^{\rm SM} = 28(8.1)(1) \times 10^{-10}$, the branching ratio $BR(\bar{B} \to X_s\gamma) = 3.55(0.26)(0.21) \times 10^{-4}$ and the cosmological neutralino relic abundance $\Omega_\chi h^2 = 0.104(0.009)(0.1\,\Omega_\chi h^2)$. For existing limits we take: $BR(B_s \to \mu^+\mu^-) < 1.0 \times 10^{-7}$, the light Higgs mass $m_h > 114.4(3\text{ th. error only})$ GeV (91.0 GeV) and superpartner masses; see [5] for a complete list. The above data are included in the likelihood and used to constrain high posterior probability regions of the model. The likelihood is modified in such a way that it includes estimated theoretical errors in the mapping from CMSSM and SM parameters to derived quantities, another major advantage of employing a Bayesian approach (see [3, 5] for details).

3 Numerical results

First, in the left panel of Fig. 1 (from [5]) we present the 2-dim posterior pdf for $m_{1/2}$ and m_0, with all other parameters marginalized over. The 68% total probability region lies mostly at large $m_0 \gtrsim 1$ TeV and not as large $m_{1/2}$, predominantly in the FP region. This is caused mostly by a recent downwards shift of the SM value of $BR(\overline{B} \to X_s\gamma)$ [10], below the current experimental world average, as explained in [5]. Surprisingly enough, with this new value, SUSY predictions from our analysis fit the experimental distribution of $BR(\overline{B} \to X_s\gamma)$ better for $\mu < 0$ (the case which we have also explored) than for $\mu > 0$. (Despite this, the case of $\mu < 0$ shows a rather poor overall fit to the data - for details see [5].) Most other observables fit the data well (or even very well), except for the anomalous magnetic moment of the muon. The overall preference for large m_0 makes $\delta a_\mu^{\rm SUSY}$ rather small. As a result, for both signs of μ the peaks of the relative probability of $\delta a_\mu^{\rm SUSY}$ are far below the central experimental value (about 3.2σ for $\mu > 0$ versus about 3.7σ for $\mu < 0$), and close to each other.

Clearly, while the 95% total probability region lies well within the assumed prior of $m_{1/2}$ (as well

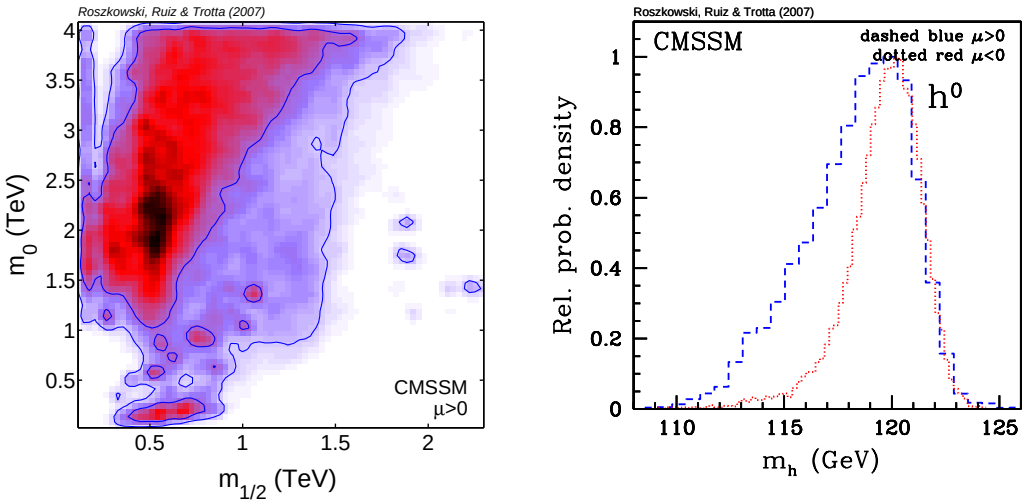

Fig. 1: Left panel: The 2-dimensional probability density in the $m_{1/2}$ and m_0 plane (with all other parameters marginalized), with the contours containing 68% and 95% probability also marked. Right panel: The 1-dim relative probability density for the light Higgs boson mass m_h for $\mu < 0$ (dotted red) and $\mu > 0$ (dashed blue).

as A_0 and $\tan\beta$, see [5]), this is not the case for m_0. This should be kept in mind in deriving conclusions from the pdfs of any observables that depends directly on m_0, such as sfermion masses. There are also sizable uncertainties associated with the FP region, in particular with the reliability of existing numerical codes in computing mass spectra. Also, the SM value of $BR(\overline{B} \to X_s\gamma)$ may still change somewhat after the NNLO calculation is completed. Thus this result should still be treated with a pinch of salt.

Note also that the above high-probability regions do not necessarily coincide with the best fitting points in parameter space if the pdf is strongly non-Gaussian, as in the present case. See [3, 5] for a detailed description of the discrepancy and a discussion of its meaning in terms of probabilistic inference.

Despite these outstanding issues, some results seem fairly robust. One is the properties of the lightest Higgs boson h^0. In the right panel of Fig. 1 (from [5]) we present, for each sign of μ, the 1-dim relative pdf of the h mass, obtained after marginalizing over all other parameters. (A previous plot, obtained in [4] with the previous value of $BR(\overline{B} \to X_s\gamma)$ is nearly identical, and also agrees rather well with ref. [8].) It is clearly well confined, with the ranges of posterior probability given by $115.4\,\text{GeV} < m_h < 120.4\,\text{GeV}$ (68%) and $112.5\,\text{GeV} < m_h < 121.9\,\text{GeV}$ (95%). A finite tail on the l.h.s. of the 1-dim pdf for m_h, below the final LEP-II lower bound of 114.4 GeV (95% CL) is a consequence of the fact that our likelihood function does not simply cut off points with m_h below some arbitrary CL, but instead it assigns to them a lower probability. On the other hand, a sharp drop-off on the r.h.s. of the relative probability density is mostly caused by the assumed upper bound on $m_0 < 4$ TeV. For instance, adopting a much more generous upper limit $m_0 < 8$ TeV would lead to changing the above ranges to $120.4\,\text{GeV} \lesssim m_h \lesssim 124.4\,\text{GeV}$ (68% CL) and $115.4\,\text{GeV} \lesssim m_h \lesssim 125.6\,\text{GeV}$ (95% CL). Other properties of the lightest Higgs boson, including its couplings to ZZ and WW pairs, for the most part closely resemble those of the SM Higgs with the same mass [4]. This means that ongoing SM Higgs searches at the Tevatron almost directly apply to h^0. According to ref. [11], with about $2\,\text{fb}^{-1}$ of integrated luminosity per experiment (with around $3\,\text{fb}^{-1}$ already on tape), a 95% CL exclusion limit can be set for the whole 95% posterior probability light Higgs mass range given derived for $m_0 < 4$ TeV ($\sim 2.5\,\text{fb}^{-1}$ for $m_0 < 8$ TeV). It is remarkable that negative Higgs searches at the Tevatron should allow one to make definitive conclusions about the ranges of CMSSM parameters, in particular m_0, which extend well beyond the reach of even the LHC in direct searches for superpartners.

We have also studied in detail prospects for dark matter detection, both direct [5] and indirect [12].

To give some highlights, for $\mu > 0$ the neutralino dark matter direct detection elastic scattering cross section σ_p^{SI} shows two main features (see Fig. 11 in [5]). Firstly, there is a strong probability peak around 10^{-8} pb, mostly due to a contribution from the FP region. Secondly, there is another high probability region of σ_p^{SI} which extends between about 10^{-10} pb and about 10^{-7} pb (which is roughly today's experimental sensitivity) and which shows a strong anticorrelation with m_χ. The largest values of σ_p^{SI} correspond, for the most part, to the FP region of large m_0. Thus this region will soon be tested in DM searches while remaining inaccessible to the LHC, except for smallest values of m_0 in the FP region.

4 Conclusions

We have presented a new method of exploring the CMSSM parameters using a state-of-the-art Bayesian method, encoded in the package SuperBayes [2]. The power and flexibility of the approach allows one to probe many previously unexplored choices of parameters and to fully incorporate the effects of remaining uncertainties in relevant SM parameters and other theoretical uncertainties in computing observables.

Using the method, we derived high probability ranges of the CMSSM parameters and showed that current data (most notably the SM value prediction for $BR(\overline{B} \to X_s\gamma)$) favour the focus point region. Despite some theoretical uncertainties in that region, we delineated high probability ranges of m_h which one should be able to rule out with high confidence on the basis of the data already collected at the Tevatron (although not yet fully analyzed). Prospects for dark matter detection in the CMSSM also look very promising. So far, as a starting point, we only assumed flat priors in the CMSSM and SM parameters - studies using different priors are in progress. Higgs properties are fairly robust with respect to changes in the *a priori* allowed range for the parameters or to the exclusion of the anomalous magnetic moment of the muon measurement from the analysis. The observables which depend on m_0 are not.

5 Acknowledgements

R.RdA is supported by the program "Juan de la Cierva" of the Ministerio de Educación y Ciencia of Spain. R.T. is supported by the Royal Astronomical Society through the Sir Norman Lockyer Fellowship and by St Anne's College, Oxford. The authors would like to thank the European Network of Theoretical Astroparticle Physics ILIAS/N6 under contract number RII3-CT-2004-506222 for financial support.

References

[1] G. L. Kane, C. F. Kolda, L. Roszkowski and J. D. Wells, Phys. Rev. **D49** (1994) 6173 [hep-ph/9312272].

[2] The code is available from `superbayes.org`.

[3] R. Ruiz de Austri, R. Trotta and L. Roszkowski, JHEP **0605** (2006) 002 [hep-ph/0602028].

[4] L. Roszkowski, R. Ruiz de Austri and R. Trotta, JHEP **0704** (2007) 084 [hep-ph/0611173].

[5] L. Roszkowski, R. Ruiz de Austri and R. Trotta, JHEP **0707** (2007) 075 [arXiv:0705.2012].

[6] B. C. Allanach and C. G. Lester, Phys. Rev. **D73** (2006) 015013 [hep-ph/0507283].

[7] B. C. Allanach, Phys. Lett. **B635** (2006) 123 [hep-ph/0601089].

[8] B. C. Allanach, C. G. Lester and A. M. Weber, JHEP**0605** (2006) 065 [hep-ph/0609295].

[9] B. C. Allanach, K. Cranmer, C. G. Lester and A. M. Weber, arXiv:0705.0487.

[10] M. Misiak *et al.*, Phys. Rev. Lett. **98** (2007) 022002 [hep-ph/0609232]; M. Misiak and M. Steinhauser, Nucl. Phys. **B764** (2007) 62 [hep-ph/0609241].

[11] M. Carena, J. S. Conway, H. E. Haber and J. D. Hobbs, hep-ph/0010338; L. Babukhadi *et al.* [CDF and D0 Working Group Members], FERMILAB-PUB-03-320-E, Oct. 2003.

[12] L. Roszkowski, R. Ruiz de Austri, J. Silk and R. Trotta, arXiv:0707.0622.

Statistical software

Statistical Software for the LHC

Wouter Verkerke
NIKHEF, Amsterdam, the Netherlands

Abstract
I give an overview of existing statistical software for high energy physics and present a proposal for the organization of new high level statistics tools for the LHC in a common framework that will simplify exchange of information and cross validation of techniques.

1 Introduction

The Large Hadron Collider (LHC) is expected to start taking data in 2008 and will produced unprecedented amounts of data to be processed and analyzed. I present an overview of software for the statistical analysis of this data that is currently available or being developed, focusing on analysis working environments, classification tools to separate signal and background, and in some more detail on software fitting, minimization, error analysis and tools for calculating significances and limits.

2 Analysis Working Environment

In the 2005 PHYSTAT conference several presentations were given describing and promoting the R analysis environment[1]. R is a language and environment for statistical computing and graphics, is related the commercial S environment, and is popular among statisticians. The R analysis environment has many features that are appealing to analysis of HEP data: R has many tools for data visualization, can read many data formats including ROOT `TTree`s and sessions can be saved to disk and recovered at a later time. Extensions of the R environment come in packages and the distributed package management is integrated in R. This structure allows users to download an install new features from a central repository on the fly in their R session (e.g. `install.packages("Rcmdr", dependencies=kTRUE)`). One of the points of concern of R for HEP is that the performance of R does not scale well to very large datasets of order million events and larger.

While R has a user community in HEP, the field has a strong tradition of home grown analysis environments starting with PAW in 1985 developed by R. Brun et al., and the ROOT[2] environment, developed by the same team in 1995 with a market share close to 100% for the physicists associated with one the LHC experiments. This market share has made ROOT the de facto standard analysis environment for High Energy Physics. ROOT components are integrated in the common software infrastructure of all LHC experiments and the data format of LHC experiments is based on ROOT. For physics analysis several LHC experiments are developing software libraries that allow to read centrally reconstructed detector data (the 'AOD') directly into the ROOT analysis environment. A key asset in ROOT data management is its performance in the handling Terabyte-sized data volumes. The ROOT command line interface is a C++ interpreter, which exposes the object-oriented structure of the underlying C++ code directly to the end user. The standard ROOT distribution provides several libraries with C++ classes that manage histogramming and graphics, define standard mathematical functions, random number generators, linear algebra tools, numerical algorithms and provide function minimization and error analysis tools.

Both the R and ROOT environment have their own repository function. While ROOT doesn't have the modular and automated package structure of R a growing number of external tools packages such as TMVA[3], RooFit[4] have found their way in the ROOT software repository, as well a growing number of smaller scale tools such as `TLimit` and `TFractionFitter` that are contained in a single C++ class. A number of classical software repositories are alive as well. Specifically for High Energy Physics tools there are `phystat.org`, which contains among others StatPatternRecognition[5],

TMVA and LepStats4LHC[6], `hepforge.net`, which provides mostly physics Monte Carlo generators and `freehep.net`. For non HEP-specific repositories the web page[7] compiled by Jim Linneman for PHYSTAT 2005 still an excellent resource.

3 Signal and Background, tools for Multivariate analysis

A large range of techniques is nowadays available to construct classifiers to separate signal from background, despite the fact HEP physicists have traditionally been reluctant in embracing the use of novel multivariate analysis. The software tool with the longest history in HEP is the multi-layer perceptron neural network, available as the MLPFit package in PAW since 1999 and still available in ROOT as the `TMultiLayerPerceptron` class. Since the PHYSTAT 2005 conference two software libraries, targeted primarily to an HEP audience, and with a much broader scope have been developed: TMVA (Tool for MultiVariate Analysis) and StatPatternRecognition. The existence of these new tools now make it possible for the average HEP physicist to get started with a relatively modern technique like boosted decision trees in less than a day of work. The easy availability of large number of new techniques will be crucial in the development of their acceptance and use in HEP.

3.1 Tool design and available classifiers

TMVA is an interactive-style tool strongly coupled to the ROOT environment. The kit implements the following classifiers: Optimized rectangular cuts, projective and multi-dimensional likelihood estimators, Fisher and H-matrix discriminants, artificial neural networks (three types of multi-layer perceptrons,), boosted/bagged decision tress with automatic node pruning, a rule fitter and a support vector machine. Plans exists for the addition of generalized non-linear discriminants and committee classifiers. TMVA provides a framework for testing, training, evaluation and application of multi-variate classifiers. Each classifier provides a ranking of the input variables. Input variables can be de-correlated or projected upon their principle components. The training results and full configuration are written to weight files and applied by a separate Reader class.

The StatPatternRecognition toolkit is a standalone C++ tool. It is designed to be a production tool rather than an interactive tool and care has been taken to ensure processing scalability to very large datasets and large number of dimensions while retaining reasonable CPU and memory consumption. At the moment of writing StatPatternRecognition implements the following classifiers: decision split/stump, decision trees (regular and top-down), bump hunting (PRIM, Friedman and Fisher), linear and quadratic discriminants, logistic regression, boosting (discrete and real AdaBoost, epsilon-Boost) of any sequence of classifiers, Arc-x4 (a variant of boosting from Breiman), bagging of any sequence of classifiers, random forests, back propagation neural networks with logistic activation function, multiclass learners (Allwein, Shapire and Singer), and an interface to SNNS neural net.

3.2 Training, analysis and validation tools

Both tools provide training tools for all classifiers that require training and include facilities to detect over training through comparison with statistically independent control samples and provide information on the discrimination power of input observables e.g. by ranking them.

TMVA developers has put a lot of emphasis on graphical presentation of results and classifier structures such as network topologies and decision trees and promote easiness of use. One of the toolkits strong points is the easy (graphical) comparison of the performance of the various discriminants. Several benchmarks are implemented: signal versus background efficiency as function of a cut on the classifier output, the separation, and the discrimination significance.

StatPatternRecognition pays in addition particular attention to cross validation and bootstrap. For cross validation data samples are split into N sub samples. Then one subsamples is removed, the discriminant is re-optimized on the N-1 subsamples the predicted error is estimated for the removed sample,

and the procedure is repeated for all N subsamples. For bootstrap, N events are randomly drawn with replacement out of the data sample. The discriminant is optimized on the bootstrap sample and seen how well it predicts the behavior of the original sample.

It should be noted that even though there is a fair amount of overlap in the classifiers provided by TMVA and SPR, the implementations of these are quite different and their performance is sometimes also different as this often depends on implementation details.

3.3 Distribution and availability

Both packages are available on the Open Source development platform SourceForge (`tmva.sourceforge.net` and `www.sf.net/projects/statpatrec`). TMVA is distributed with the ROOT in addition. Both packages are described in more detail in contributed proceedings of this conferences.

4 Fitting, Minimization and Error Analysis

From a software point of view a fit of a model to data involves three separate steps: first a model must be constructed. This is usually a (probability density) function. Next, a test statistic is constructed from the model and the data. For fitting likelihood and χ^2 are the most common test statistics. Finally, the test statistic is minimized with respect to its parameters and an error analysis is performed to determine the uncertainty on the parameters.

Historically there is no common language or software tool to construct models. In the early days models were implemented as FORTRAN subroutines, more recently C and C++ functions are used. While minimally sufficient, any support for common operations on models such as integration, toy Monte Carlo sampling and visualization are left as a problem to the physicist. Since 1995, the ROOT environment provides dedicated C++ classes that represent 1,2,3-dimensional functions that provide basic common functionality such as numeric integration and toy Monte Carlo sampling, though do not scale to very complex models. The RooFit toolkit, integrated in the ROOT environment, has recently taken the concept of data modeling further by providing a modular collection of classes that can be used to compose models of arbitrary complexity and provide extensive support for p.d.f normalization, fitting, toy Monte Carlo generation and visualization in a computationally optimized way.

High quality software tools for minimization and error analysis on the other hand have been around for several decades. While several commercial or public packages exist that contain tools to perform function minimization, notable the NAG software library and more recently the GNU Scientific Library, the industry standard for function minimization and error analysis in HEP is, and has been, MINUIT[8] for nearly 40 years. It is highly regarded for its robustness and ability to deal with difficult problems. Its multi decade track record started in the Fortran-based CERNLIB, which was later interface to the PAW analysis environment and is now available in ROOT as well. The default version in ROOT is a straight Fortran-to-C++ port of the original application by the ROOT team. A rewrite in C++ by the authors of Minuit, known as Minuit2, is available as well since a couple of years and has a cleaner interface. MINUIT contains aside from the MIGRAD minimization code, two algorithms to do error analysis: HESSE, based on the analysis of second derivatives and MINOS, a hill climbing algorithm that can also quantify asymmetric errors.

A number of alternative fitter/minimizers have been added to ROOT in recent years that deal with specific problems. These are `TFumili`, an implementation of the Fumili algorithm[9] that achieves faster convergence for certain types of problems, `TLinearFitter`, which implements an analytical solution for problems linear in their parameters and `TFractionFitter` and `TBinomialEfficiencyFitter` for fitting of Monte Carlo templates and efficiencies from ratio respectively.

5 Constructing models to describe the data

I focus in some more detail on the topic of ROOT and RooFit data modeling language as this will be an important aspect for the design of future statistical analysis tools and because it is not covered in much detail in other contributions of this conference. This section assumes a basic knowledge of the C++ programming language.

5.1 ROOT function classes

Basic ROOT function classes have an elegant design for uses cases of lower complexity. For example a 1-dimensional function named `fa1` with a single observable `x` is instantiated as

```
TF1* fa1 = new TF1("fa1","sin(x)/x",0,10) ;
```

Functions that cannot be easily expressed in a single line of code can be created as generic C++ functions bound to `TF1` objects.

The advantage of a common interface for all types of function implementations is easy access to common functionality. For example series of toy Monte Carlo events can be sampled from any `TF1` object through

```
TH1F* h1 = new TH1F("histo","sampling of myfunc",100,0,10) ;
h1->FillRandom("myfunc",20000) ;
```

or be fit to a TH1 histogram through

```
f1->SetParameters(800,1) ;
h1->Fit("myfunc",20000) ;
```

ROOT function classes provide a basic concept of modularity as all functions are named and registered in a central repository, and can reference each other by name. Issues that are not addressed by ROOT classes are: support for normalized probability density functions, calculation of an unbinned likelihood, automatic computational optimization of function expressions and advanced introspection and modification tools for complex composite objects. The ROOT authors are currently working on redesigning their function structure and fitting interface to address a number of these issues.

5.2 The RooFit toolkit for data modeling

The RooFit toolkit for data modeling was developed in the BaBar collaboration in 1999 to address all of these issues, as a spin-off from the flag-ship measurement of the CP violation parameter $\sin 2\beta$. This measurement required a fit to a five dimensional dataset with over 30 floating parameters. The initial implementation of the model was written in FORTRAN, but this solution turned out not to scale well to other analyses and presented problems with maintenance over time. While initially developed for BaBar, the toolkit is available in the general ROOT distribution since 2005.

5.2.1 Basic concepts

The key concept of RooFit is to take modularity in terms of C++ objects one step further than was done for the ROOT function classes by making each mathematical component of a model a separate software object, i.e. not only functions are represented by objects but also the variables that appear in them are represented by separate objects. Consequently, RooFit code to write a simple function like a Gaussian is somewhat more elaborate than when done through a ROOT `TF1`:

```
RooRealVar x("x","x",-10,10) ;          // declare variable x
RooRealVar m("m","mean",-10,10) ;
RooRealVar s("m","sigma",0.1,10) ;
RooGaussian g("g","g(x,m,s)",x,m,s) ;   // declare p.d.f in terms of x,m,s
```

Name	Math concept	RooFit class
variable	x, p	`RooRealVar`
function	$f(\vec{x})$	`RooAbsReal`
p.d.f.	$F(\vec{x}, \vec{p})$	`RooAbsPdf`
space point	$\vec{x}$	`RooArgSet`
integral	$\int_{x_{min}}^{x_{max}} f(x)dx$	`RooRealIntegral`
list of space points	$\vec{x}_k$	`RooAbsData`

Table 1: Mapping between mathematical concepts and RooFit classes

The upside of this approach is that a number of configuration and documentation issues is interfaced in a very transparent way. Each variable has a intrinsic name, an optional more descriptive title, and associated information concerning the allowed range of each variable that is stored with the object representing the variable. The value of all objects can be retrieved with the universal `getVal()` method and objects that represent variables or invertible transformations can also be assigned a value with the corresponding `setVal()` method

```
m.setVal(3.5) ; // sets value of m to 3.5
x.getVal() ;    // returns value of variable x
g.getVal(x) ;   // returns value of Gaussian p.d.f at present
                // values of x,m,s, normalized over observable x
```

5.2.2 Parameters, observables and normalization

The evaluation of the probability density function `g` takes an argument that specifies which of the variables of `g` should be interpreted as observables because RooFit p.d.f. objects have no intrinsic notion of observables and parameters. This may seem confusing at first, but has several important advantages. First, it allows a natural use of p.d.f.s in both Bayesian and Frequentist contexts. For example a call to `g.getVal(s)` ruturns a normalized probability density function in the observable `s`. Second, it allows for a clean extension and composition concept reusing existing p.d.f classes. One can for example substitute the variable `m` with a polynomial function $m = a_0 + a_1 \cdot y$ as follows:

```
RooRealVar y("y","y",-10,10) ;
RooRealVar a0("a0","a0",1
RooRealVar a1("a1","a1",-3) ;
RooPolyVar m("m","m(y,a0,a1)",y, RooArgList(a0,a1)) ; }
```

In this case the argument `m` of the Gaussian is neither observable nor parameter and is unproblematic because class `RooGaussian`, like any other RooFit function object, makes no assumption other than that its argument is a real-valued entity. This example effectively makes a p.d.f. that could be used as a two-dimensional probability function in `x` and `y`. Substitutions of this type are generically possible for any variable argument of any function or p.d.f.

5.2.3 Function class library, writing new functions

RooFit is shipped with library of about 20 p.d.f. classes that include basic shapes (Gaussian, polynomial etc), non-parametric shapes (histogram-based, kernel estimation [10]) and physics inspired shapes (Breit-Wigner, CrystalBall, several B physics decay models) that can all be adapted to a specific use case as described above. If a particular shape is missing a new class is easy to write. Code for a new p.d.f class can be generated semi-automatically with a class factory: e.g.

```
RooClassFactory::makePdf("RooWavingPdf","a,b,x","a+b*cos(x)") ;
```

generates a fully functional p.d.f class `RooWavingPdf` with three variables a, b, x and above definition that is ready to use in ROOT. A normalization term to make the above function expression a proper normalized probability density function is automatically inserted by RooFit, with the exact expression depending on which variable is considered observable in the use context. If users know how to analytically integrate or normalize their expression, this knowledge can be integrated into the p.d.f and will be used, otherwise numeric integration is used, for which a variety of techniques is available.

5.2.4 Composing complex models

Composition operations like addition, multiplication and convolution are expressed through dedicated operator p.d.f classes that allows the core code to recognize such operations and apply computational optimizations accordingly. Additions require a number of fraction parameters to be specified with the components to be added:

```
RooRealVar sigFrac("sigFrac","f(sig)",0.5,0.,1.) ;
RooAddPdf  sum("sum","sig+bkg",RooArgList(sig,bkg),sigFrac) ;
```

Multiplication of p.d.f.s can be done for terms without correlations or with correlations.

```
RooProdPdf    prod("prod","sigx*sigy",RooArgList(sigx,sigy)) ; // uncorrelated
RooProdPdf    fg("fg","f(x|y)*g(y)",g(y),Conditional(f,x)) ;   // f(x|y)*g(y)
```

Correlations expressed through conditional p.d.f.s, such as $f(x|y) \cdot g(y)$, are computationally most efficient, but other forms are also accepted. Similar classes exist for generic convolution of p.d.f.s either through brute-force calculation (`RooNumConvPdf`) or by using the convolution theorem (`RooFFTConvPdf`). Conversely, existing p.d.f can be reduced in dimensionality by integrating out observables. This example

```
RooAbsPdf* fx = fxy->createProjection(y) ;
```

constructes a one-dimensional p.d.f $f_x(x)$ through integration of the two-dimensional p.d.f $f_x y(x, y)$.

5.2.5 Fitting, likelihood calculations and toy MC event generation

Once a p.d.f. object has been created, it has universal functionality for fitting, toy Monte Carlo generation and visualization, regardless of its internal complexity. Basic operations are one-line operations:

```
model->fitTo(data) ;                          // fit model to data
RooDataSet* data = model.generate(x,10000) ; // sample data from model
```

While the `fitTo()` method takes care of both creating the (unbinned) likelihood object and its subsequent minimization and error analysis, these steps can also be separated for a greater degree of user control. The likelihood itself is a regular RooFit function that can be manipulated in any way any RooFit function can be. The minimization can be controlled interactively by creating a `RooMinuit` interface object and performing the various minimization steps one by one.

```
RooNLLVar nll("nll","-log(L)",pdf,data) ;
RooMinuit m(nll) ;
m.migrad() ;       // minimize likelihood
m.hesse() ;        // error analysis
p.setConstant() ;  // fix parameter p of likelihood
m.hesse() ;        // rerun error analysis
```

The result of the minimization by MINUIT is automatically propagated to the value of the parameters of the p.d.f. The result of the error analysis by HESSE is is stored in the error property of those parameters. The `RooMinuit` interface can minimize any real-valued function so that is easy to e.g. add a penalty term to a likelihood to be minimized

```
RooFormulaVar nllp("nllp","nll +0.5*pow((m-m0)/s,2)",RooArgSet(nll,m,m0,s)) ;
RooMinuit m(nllp) ;
```

5.2.6 Computational optimization

The base class for the calculation of test statistics like `RooNLLVar` contains several algorithms that optimize the performance of the test statistic calculation using generic optimization algorithms. Prior to each use it detects and pre-calculates all terms that depend exclusively on constant parameters. It caches integral values and only recalculates them if any of the parameters of the integrand has changed. In general multi-dimensional problems are factorized as much as possible to achieve maximum granularity and optimization performance. This automatic optimization keeps the users code clean, yet yields optimal performance for every use case. The typical speedup of a likelihood calculated of realistic complex fits ranges from a factor 3 to a factor 10. Test statistic calculations can be parallelized for use on multi-CPU and multi-core machines: it suffices to indicate the number of CPUs to be used in the `fitTo()` function call to activate this feature.

The toy Monte Carlo generator code is optimized is a similar way. Each p.d.f can optionally advertise an internal generator if a generation method exists that is more efficient that accept/reject sampling. Each generation request to a composite p.d.f will use the most efficient combination of internal and accept/reject methods available, generation components of summed p.d.f.s and multiplied p.d.f separately. The generation order of observables for expressions containing conditional p.d.f.s is automatically determined. If cyclical conditional dependencies are present the entire cyclical expression is generated with the accept/reject technique.

6 Calculating significances and confidence intervals

The calculation of a signal significance, a Frequentist confidence interval or limits of various types require complex interactions with both model and data. In many cases these calculations have been performed with code specifically tailored to the analysis in question. A number of smaller and larger software tools have been made available in the past years, often a cleaned and generalized version of the code that was once used for a specific physics analysis. Several of these tools are found in the ROOT distribution, others are published in external repositories. The following is a short survey of available tools.

ROOT class TRolke calculates the confidence intervals for the rate of a Poisson in the presence of background and an efficiency. It makes a fully Frequentist treatment of the uncertainties in the efficiency and and background estimate using the profile likelihood method. The signal is always assumed to be a Poisson. Seven different models are included with varying choices for background and efficiency models (Poisson, Gaussian, binomial or known). The class only handles the count of signal and background events, treatment of discriminating variables is not included.

ROOT class TFeldmanCousins calculates the fully Frequentist construction as described by Feldman and Cousins. It is not capable of treating uncertainties in nuisance parameters and is intended to be used for cases with no or negligible uncertainties. It only handles the count of signal and background events, treatment of discriminating variables is not included.

ROOT class TLimit contains an algorithm to compute 95% C.L. limits using the likelihood ratio semi-Bayesian method. It takes signal, background, data histograms as input. It runs a set of Monte Carlo experiments to compute the limits. If needed, inputs can be fluctuated according to their quoted systematic uncertainties. The ROOT class is a rewrite of the original `mclimit.f` FORTAN software. Class `TLimit` does work with discriminating variables. The TLimit code, like the original fortran code does not take systematics on the shapes of inputs into account. A newer version of that code `mclimit_csm` does have that capability

Standalone C++ tool `mclimit_csm` is designed to calculate limits for complex realistic analyses such as the CDF Standard Model Higgs Boson search in which two disjoint samples "single-tagged" and "double-tagged" are analyzed and in which nuisance parameters affect both in a correlated way. The `mclimit_csm` tool used binned input data, can handle multiple signal and background sources. The model predictions are sums of template histograms from different sources. Most importantly, each source of signal and background can have a rate and shape uncertainty from multiple sources of systematic uncertainty. Shape uncertainties are handle by a template morphing algorithm or by simple linear interpolation within each bin. Uncertain parameters can be fit for, in the data and in each pseudo-experiment. Finite MC statistical errors in each bin are included in the fits and the pseudo experiments. The output consists of the p values given the data and the model. In addition CLs and CLS-based cross section limits can be computed as well as Bayesian limits using a flat prior in the cross section. The code is available at [11].

Standalone C++ tool `LepStats4LHC` comprises a series of tools for Frequentist limit calculations that implement the LEP-style calculation of significances in C++. It uses the external Fast Fourier Transform package FFTW[12] to implement the required convolution calculations. The interface is a series of command line utilities. `PoissonSig` calculates the significance of a number counting analysis. `PoissonSig_syst` calculates the significance including a systematic error on the background expectation. `Likelihood` calculates the combined significance of several search channels or to calculate the significance of a search channel with a discriminating variables. Finally `Likelihood_syst` does that taking a systematic error associated with each channel into account.

ROOT class TSplot is in a separate category of statistics tools. The sPlot concept [13] provides a techniques for the analysis of multi-dimensional likelihood models and for the subtraction of background by exploiting information of the full correlation matrix in event weighting. The details of the sPlot technique are described in detail in PHYSTAT 2005 proceedings. A basic version of sPlots is implemented in the ROOT class `TSplot`.

7 Toward a common framework for statistical tools

Experiments at LEP, Tevatron and the B factories have created tools that combine multiple channels and included systematic uncertainties. These tools generally implement a specific technique and the construction of combined results requires significant manual intervention. The survey of the last section shows that the interfaces for these tools are very different from each other and that tools generally avoid dealing with analytical models. Tools either do not deal at all with discriminating variables or represent them through histograms.

7.1 Desired properties of a common framework for the LHC

For the LHC it would be desirable to have a more versatile, generic solution for statistics tools. In addition to providing tools for simple calculations, this framework should be able to combine the results of multiple measurements, be able to incorporate systematic uncertainties and facilitate the technical aspects of code sharing. Such a framework should implement all major classes of statistical techniques, i.e. likelihood based techniques, where all inference is made from likelihood curves; Bayesian techniques, where a prior on parameters is used to compute P(theory|data), and Frequentist techniques which are restricted to statements on P(data|theory). Within each of these classes there are several ways to approach the problem and the framework should support each of these types of techniques and provide some common abstractions. The usefulness of this approach has already been demonstrated: in the PHYSTAT2005 workshop Kyle Cranmer compared the coverage of several common methods which can incorporate systematic errors on an identical problem and found significant discrepancies in significance for these methods. It will be beneficial for the LHC if tools and methods can be cross calibrated before they are used for physics publications.

7.2 Toward common tools - the RooStats project.

An initiative has been started by Rene Brun and Kyle Cranmer to organize a suite of common statistics tools in ROOT. Essentially all of the three basic classes of statistical methods start with the probability density function or the likelihood function. Thus building a good model is the hard part as such a model should be reusable for multiple methods and be able to interface to common tools. Following a survey of existing software and feedback from the user community it has been proposed to build this tools suite on top of the RooFit toolkit for data modeling as that already provides solutions for most of the hard problems. Since RooFit has no static notion of parameters and observables in the core code it is naturally suitable to work with both Bayesian and Frequentist techniques. The idea of the initiative is to have a few core developers maintaining the framework and have a mechanism for users and collaborators to contribute concrete tools.

7.3 Sharing and publishing models

In the spring of 2007 Kyle Cranmer and I have worked out a basic design of the class structure of RooStats to identify which existing part of RooFit can be used and which parts are missing. The first concrete development from this design study is to enhance the RooFit data modeling language with persistence, i.e. the ability to store any models constructed in memory in a ROOT file. To this end a new concept has been introduced, the `RooWorkSpace`, which can contain all components of a RooFit model definition (p.d.f.s, variables, functions) as well as RooFit datasets.

A ROOT file with a workspace in essence a universal language to describe, store and share models. Effectively, the `RooWorkspace` can be the ultimate publication of a physics results, as it allows to share the *actual* likelihood function of a measurement in a form that can be read and manipulated by anyone in ROOT without requiring any experiment- or analysis-specific software. The language of these models is generic and universal, so that e.g. a theorist should be able to use a likelihood published by an experiment without the need for any experiment-specific software. Similarly, it will be possible to perform statistical analysis on combined results given an number of supplied workspaces and just a few lines of code.

A few code examples below illustrate the simplicity of using workspaces. Given a p.d.f `g` of arbitrary complexity and a dataset `d`, we can publish the actual likelihood of our measurement as follows:

```
RooNLLVar nll("nll","likelihood",*g,*d) ; // construct likelihood
RooWorkspace w("my workspace") ;          // create workspace
w.import(nll) ;                           // store likelihood in workspace

TFile f("myresult.root","RECREATE") ;     // open ROOT file
w.Write() ;                               // save workspace
```

The importing of the likelihood object, as shown above, recursively imports and stores all the necessary dependent objects: p.d.f.s function, variables and datasets, which can comprise hundreds of objects for complex realistic models. All these components will remain individually accessible in the workspace. At this point, any physicists with a working ROOT environment and can analyze and interpret these stored data. For example, to access the likelihood one does

```
TFile f("myresult.root") ;
RooAbsReal* nll = w.function("nll") ; // extract likelihood
RooArgSet* nll_pars = nll->getVariables() ; // obtain set of all parameters of L
```

The universal introspection functions of RooFit classes like the `getVariables()` method above allow to access all of its properties. The following continuation performs a profile likelihood analysis:

```
RooProfileNLL pnll("pnll","profile likelihood,nll,*p) ; // construct profile L
RooPlot* frame = p->frame() ;
pnll.plotOn(frame) ;
```

On other words, if e.g. an Atlas physicist were given a CMS workspace, *these few lines of code are all that are needed* to resurrect the actual CMS likelihood of the analysis in question without the need for any CMS specific code libraries. Along similar lines a combination analysis can be performed by opening multiple ROOT files with workspaces previously created by various authors or even various experiments. Generic RooFit utility classes allow to add likelihoods a posteriori and to perform a joint statistical analysis. The workspace class is equipped with utility functions that allow to renames variables and functions on the fly to solved naming conflicts and mismatches that may arise when combining workspaces from different sources.

7.4 Development plans for RooStats

The workspace concept, already available in ROOT 5.17, provides in addition to a vehicle for sharing and publishing models, a common interface for new statistics tools to be developed. The next step in the RooStats project, expected for ROOT release 5.18, is to implement new and to refactor existing statistics tools to form an initial set tools that interface to the workspace and perform various types of limit and confidence interval calculations, envisioning classes like `RooBayesianInterval`, `RooNeymanConstruction`, `RooProfileLikelihood` and `RooSPlot`. An important design goal for this next step is to aim for an easy-to-use interface for these tools, which sometimes require a substantial amount of input information, and to aim to abstract all common aspects in required inputs in a common interface to promote similarity and interoperability of these tools.

8 Summary

The software landscape for statistics tools for HEP is evolving rapidly. A decade ago most physicists were performing analyses in FORTRAN and writing their own (statistical) analysis tools. The enormous progress object-oriented programming, brought to the analysis environment in HEP by the C++ ROOT environment has promoted modularity and interoperability and offers unprecedented new possibilities. While most statistics tools for LEP and Tevatron were experiment specific and written during data taking it now looks technically feasible for the LHC to have a fully functional set of common statistics tools and electronic publishing and sharing of results before data taking has started.

References

[1] `http://www.r-project.org/`
[2] `http://root.cern.ch`
[3] `http://tmva.sf.net`
[4] `http://roofit.sf.net`
[5] `http://www.sf.net/projects/statpatrec`
[6] `http://phystat.org/`, select package `LepStats4LHC`
[7] `http://www.pa.msu.edu/people/linneman/stat_resources.html`
[8] MINUIT, Cern Program Library Long Writeup D506.
[9] `http://root.cern.ch/root/html/TFumili.html`
[10] K. Cranmer, Comput. phys. commun.vol.136 (2001),no3,pp.198-207
[11] `http://www.hep.uiuc.edu/home/trj/cdfstats/mclimit_csm1/index.html`
[12] `http://www.fftw.org`
[13] M. Pivk and F.R. Le Diberder, Nucl. Inst. Meth A 555, 356-369, 2005.

ROOT Statistical Software

L. Moneta, I. Antcheva, R. Brun, A. Kreshuk
CERN, Geneva, Switzerland

Abstract
Advanced mathematical and statistical computational methods are required by the LHC experiments for analyzing their data. Some of these methods are provided by the ROOT project, a C++ Object Oriented framework for large scale data handling applications. We review the current mathematical and statistical classes present in ROOT, emphasizing the recent developments.

1 ROOT Math Work Package

The ROOT MATH work package is responsible to provide and to support a coherent set of mathematical and statical libraries required for simulation, reconstruction and analysis of high energy physics data. Existing libraries provided by ROOT are in the process of being re-organized in a new set of mathematical libraries with the aim to avoid duplication, increase modularity and to facilitate support in the long term. The main library components are the followings and shown in figure 1.

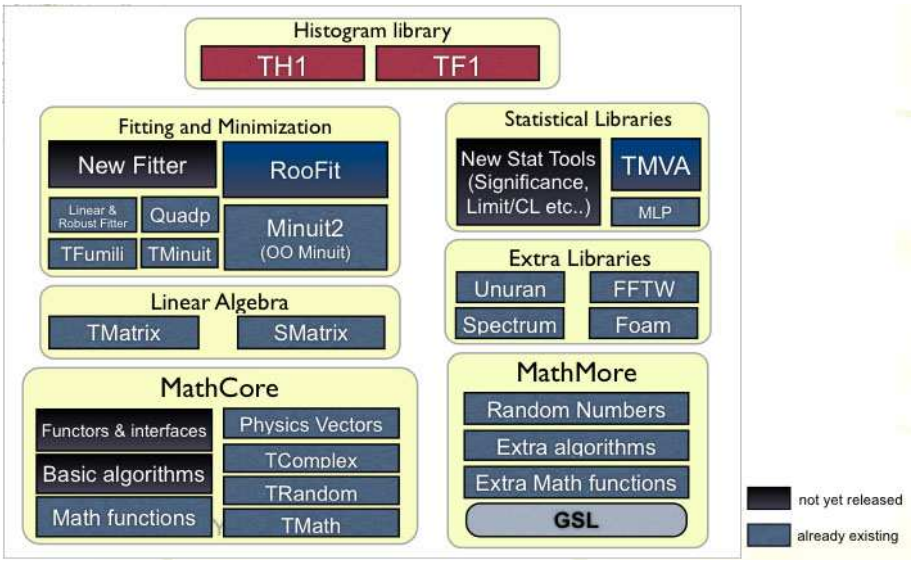

Fig. 1: New structure of the ROOT Mathematical Libraries. A different color code is used to distinguish components already existing from those which are in the process of being developed.

- **MathCore**: a self-consistent minimal set of mathematical functions and C++ classes for the basic needs of HEP numerical computing.
- **MathMore**: a package incorporating functionality which might be needed for an advanced user (as opposed to MathCore which addresses the primary needs of users) and dependent on external libraries like the GNU Scientific Library [1].
- **Linear Algebra**: vector and matrix classes and their related linear algebra functions. Two libraries exist: a general matrix package completed with a large variety of linear algebra algorithms and SMatrix, a dedicated package for small and fixed size matrices with optimal performances.
- **Fitting and minimization libraries**: classes and libraries implementing various types of fitting and function minimization methods, like Minuit and the new object-oriented version Minuit2.

– **Statistical libraries**: packages providing various algorithms for multi-variate analysis or classes for computing confidence levels and discovery significances using frequentist or Bayesian statistics.
– **Histogram libraries**: advanced classes for displaying and analyzing one, two and three dimensional data. It provides the histograms and profiles classes. Multi dimensional data sets are handled by the tree library.

In the following sections a detailed description is given for some of these components which have been recently developed and released. A brief description will be given also for those components that are planned to be introduced in ROOT.

2 Mathematical functions

New mathematical functions have been added recently in the MathCore and MathMore library to complement the functions existing in the namespace TMath and present in the ROOT core library. The new special functions are those proposed in the next extension of the C++ Standard Library [2] and follow the same naming scheme. These functions include all the major special functions, like the gamma, beta, error functions and also Bessel functions, hypergeometric functions, elliptic integrals, Legendre and Laguerre polynomials. Furthermore the MathCore and MathMore libraries provide all the major statistical distribution functions such as normal, χ^2, Cauchy, etc.., in a coherent naming scheme. For each statistical function, the probability density function, with suffix `_pdf`, (for example `normal_pdf` for the normal distribution), the cumulative distribution function with suffix `_cdf`, the complement of the cdf with suffix `_cdf_c`, the quantile function (inverse of the cdf), with suffix `_quantile`, and the inverse of the complement of the cdf, with suffix `_quantile_c` are provided.

Extensive tests of these functions have been performed [3] by comparing the numerical results obtained with the functions from other packages like Mathematica or Nag [4]. Often an accuracy at the level of 10^{-16} (double numerical accuracy) is reached for these functions.

3 Random Numbers

In ROOT pseudo-random numbers can be generated using the `TRandom` classes. A base class provides the methods for generating uniform and non-uniform numbers (according to specific distributions) while the derived classes, `TRandom1`, `TRandom2` and `TRandom3` implement pseudo-random number generators. These classes have been recently improved by replacing some obsolete generators. The following pseudo-random number generators are currently provided:

– Mersenne and Twister generator [5] implemented in the class `TRandom3`. This is the default generator in ROOT and the recommended one for the very good random propriety and its speed. It can also be seeded automatically using a 128 bit UUID number in order to generate independent streams of random numbers.
– RanLux generator [6] provided by the class `TRandom1`.
– Tausworthe generator [7] from L'Ecuyer provided by the class `TRandom2`. This generator has the advantage to use only 3 words of 32 bits for its state.

The CPU time results for generating a pseudo-random number using the ROOT generators are shown in table 3.

The base class `TRandom` provides also a Linear Congruential Generator. This generator has a state of only 32 bits and therefore a very short period and should not be used in any statistical application. `TRandom` implements as well methods for generating random numbers according to specific distributions. Recently a new faster algorithm for generating normal distributed random numbers, based on the acceptance-complement ratio method (ACR) [8], has been added to ROOT. This algorithm is much faster

Random Number Generator	Intel 32	Intel 64
MT (`TRandom3`)	22 ns	9 ns
Tausworthe (`TRandom2`)	17 ns	6 ns
RanLux (`TRandom1`)	120 ns	98 ns

Table 1: CPU time (in nanoseconds) for generating one pseudo-random number on a Linux box with the 32 or 64 bit architecture running CERN Scientific Linux 4 and using the GNU gcc version 3.4 compiler

than the traditional Box-Muller (polar) method used previously in ROOT which requires the evaluation of mathematical functions like `sqrt` or `log`. For example, on a 64 Intel Linux box running ROOT compiled with gcc 3.4, the time for generating one random gaussian number has been decreased from 183 to 42 ns.

The latest releases of ROOT contains in addition an interface to UNU.RAN [9], a software package for generating non-uniform pseudo-random numbers. It contains universal (also called automatic or black-box) algorithms that can generate random numbers from large classes of continuous (in one or multi-dimensions), discrete distributions, empirical distributions (like histograms) and also from practically all standard distributions. Efficient methods based on Markov-Chain Monte Carlo are as well provided for multi-dimensional distributions.

4 Numerical Algorithms

New numerical algorithms based on the GNU Scientific Library (GSL) [1] are provided by the MathMore library. Classes for numerical differentiation, various adaptive and non-adaptive integration, interpolation, minimization and root finding algorithms for one-dimensional functions are currently present. Algorithms for multi-dimensional functions like Monte Carlo integration and minimizations are in the process of being added. Fast Fourier Transforms are as well provided via an interface to the FFTW [10] package. The new algorithms are designed by presenting a single interface to the user for the various implementations. Alternative implementations which can be present in different libraries can then be loaded at run-time using the plug-in manager system.

5 Minimization and Fitting

Fitting in ROOT is possible directly via the `Fit(..)` methods of the various data object classes like histograms (classes `TH1`, `TH2`, `TH3`), graphs (classes `TGraph`, `TGraphErrors`, `TGraphAsymmErrors` and `TGraph2D`) and trees (class `TTree`). Methods like least-squares or binned and un-binned likelihood fits are supported. An interface class, `TVirtualFitter` exists to perform more sophisticated fits and to interface the minimization packages, like Minuit [11], Fumili [12] or Minuit2 [3]. In the case of linear fits, a dedicated class `TLinearFitter` exists to solve the resulting linear system. An extension to the linear fitter (robust fitter) for removing bad observations, outliers, based on the approximate Fast Least Trimmed Squares (LTS) regression algorithm for large data sets [13] exists as well. More complex fits can be performed by using the RooFit package [14], which is now distributed within ROOT.

A new object-oriented version of Minuit has been recently developed and it is now integrated inside ROOT as a new package, called Minuit2. It provides and enhances all the functionality of the original version. The profits from basing on an object oriented design are increased flexibility, easy maintainability in the long term and opening to extensions such as integration of new algorithms, new functionality and changes in user interfaces. For example, the Fumili algorithm has been integrated directly inside the minimization framework provided by Minuit2. Various extensive tests have been performed to study and validate the numerical quality, convergence power and computational performances of this new version. In the future it is expected to improve the functionality by adding the possibility of supplying constraints on the parameters.

A new GUI for fitting has been introduced in order to drive the fitting process. It is possible to select the fitting function, to set the initial parameter values, fitting and minimization options with possibility of choosing the minimization engine. It is foreseen to be improved soon by adding advance drawing functionality such as contour plots, residuals and confidence levels.

In the future it is planned to improve the existing ROOT fitting classes, by extending the functionality of the `TVirtualFitter` class, by providing support for parallel fits, various fitting and minimization methods and easier integration with RooFit.

6 Statistical tools

For multi-variate analysis and signal-background discrimination a new package, TMVA [15], has been integrated recently in ROOT. It provides various algorithms, like automatic cuts optimizations, likelihood estimators, neural networks and boosted decision trees with common interfaces to use them easily together. Neural networks can also be used directly via the class `TMultiLayerPerceptron`. `TMultiDimFit` is another multi-dimensional method present in ROOT, which provides the possibility to find the parametrization of multi-dimensional functions with polynomials, Chebyshev or Legendre functions. It is used for example to parametrize the LHCb magnetic field from the measured field map. The class `TPrincipal` gives the possibility to perform principal component analysis to reduce dimensionality of the data while keeping as much information as possible. The class `TRobustEstimator` implements the Minimum Covariance Determinant estimator, a robust technique [16] to find the location and scatter of multi-dimensional data.

For estimating confidence levels, the class `TFeldmanCousin` computes upper limits for Poisson processes in the presence of background using the Feldman-Cousin method [17]. The class `TRolke` computes again the confidence intervals for Poisson processes but including the treatment of uncertainties in the background and in the signal efficiency using a profile likelihood method [18]. The class `TLimit` computes instead the confidence intervals using the CL_s method used for LEP Higgs searches [19]. It is applied to histograms representing the data, the simulated signal and the background and it incorporates the systematic uncertainty using a Bayesian approach.

A new package is also currently being developed to extend and improve the functionality of estimating confidence levels to satisfy the LHC requirements and focusing in particular on estimating discovery significances. It will both include frequentists and Bayesian methods and it will be based on the RooFit data modeling framework [20]. Tools for easy statistical combinations of results will be as well provided by this new package.

7 Conclusions

ROOT contains already a large variety of mathematical and statistical functionality required for the analysis of LHC data. An effort is on-going to consolidate and improve the existing libraries by replacing obsolete algorithms, by making them easier to use and by improving their modularity to gain in long term maintainability. The needs and the feedback received from users working on data analysis and reconstruction of the experiment data are as well taken into account in this consolidation process. Many of the statistical tools currently present in ROOT have been developed by various contributors from the high energy physics community. It is therefore important to ensure a continuation of these user contributions and to provide as well an easy way for the users to plug-in their developed tools. This consolidation effort should as well aim to remove duplications and provide implementations which are considered standard by the community.

References

[1] M. Galassi et al, *The GNU Scientific Library Reference Manual* - Second Edition, ISBN = 0954161734 (paperback). See also the url `http://www.gnu.org/software/gsl`

[2] W. Brown and M. Paterno, *A proposal to Add Mathematical Special Functions to the C++ Standard Library*, WG21/N1422 = J16/03-0004, available at the url `http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2004/n1687.pdf`.

[3] M. Hatlo *et al.*, *IEEE Transactions on Nuclear Science* **52-6**, 2818 (2005)

[4] The Numerical Algorithm Group (Nag) C Library, see the url `http://www.nag.co.uk/numeric/cl/CLdescription.asp`

[5] M. Matsumoto and T. Nishimura, *Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generators*, ACM Trans. on Modeling and Computer Simulations, 8, 1, (1998), 3-20

[6] F. James, *RANLUX: A Fortran implementation of the high quality pseudo-random number generator of Lüscher*, Computer Physics Communication, 79 (1994) 111.

[7] P. L'Ecuyer, *Maximally Equidistributed Combined Tausworthe Generators*, Mathematics of Computation, 65, 213 (1996), 203-213

[8] W. Hoermann and G. Derflinger, *The ACR Method for generating normal random variables*, OR Spektrum 12 (1990), 181-185.

[9] see the url `http://statistik.wu-wien.ac.at/unuran`.

[10] see the url http://www.fftw.org.

[11] F. James, *MINUIT Reference Manual*, CERN Program Library Writeup D506.

[12] S. Yashchenko, *New method for minimizing regular functions with constraints on parameter region*, Proceedings of CHEP'97 (1997).

[13] P.J. Rousseeuw and K.Van Driessen, *Computing LTS Regression for Large Datasets*, *Estadistica* **54**, 163 (2002).

[14] see the url `http://roofit.sourceforge.net`.

[15] F. Tegenfeld *et al.*, *TMVA - Toolkit for multivariate data analysis with ROOT*, proceedings to this workshop, see also the url `http://tmva.sourceforge.net`.

[16] P.J. Rousseeuw and K.Van Driessen, *A fast algorithm for the minimum covariance determinant estimator*, *Technometrics* **41**, 212 (1999).

[17] G.J. Feldman and R.D. Cousins, *Unified approach to the classical statistical analysis of small signals*, *Phys.Rev.* **D57**, 3873 (1998).

[18] W. Rolke, A. Lopez, J. Conrad, *Nuclear Instruments and Methods* **A551**, 493-503 (2005).

[19] T. Junk, *Nuclear Instruments and Methods* **A434**, 435-443 (1999).

[20] W. Verkerke, *Statistical software tools for LHC analysis*, proceedings of this workshop, 2007.

TMVA, the Toolkit for Multivariate Data Analysis with ROOT

Andreas Höcker[1], *Peter Speckmayer*[1], *Jörg Stelzer*[1] *Fredrik Tegenfeldt*[2] *and Helge Voss*[3]
[1]CERN, Switzerland
[2]Iowa State U., USA
[3]Max-Planck-Institut für Kernphysik, Heidelberg

Abstract
Multivariate classification methods based on machine learning techniques have become a fundamental ingredient to most physics analyses. The classification techniques themselves have also significantly evolved in recent years. Statisticians have found new ways to tune and to combine classifiers to further gain in performance. Integrated into the analysis framework ROOT, TMVA is a toolkit offering a large variety of multivariate classification algorithms. TMVA manages the simultaneous training, testing and performance evaluation of all the classifiers with a user-friendly interface, and also steers the application of the trained classifiers to data.

1 Introduction

The Toolkit for Multivariate Data Analysis (TMVA) provides a ROOT-integrated framework for the processing and parallel evaluation of many different multivariate classification techniques. The classification is done in terms of two event categories, e.g. signal and background. The idea of TMVA is to integrate a large variety of powerful multivariate classifiers in one common environment with a single interface allowing the user to compare all classification techniques for any given problem. TMVA offers convenient preprocessing possibilities for the data prior to feeding them into any of the classifiers. Auxiliary information about the data is provided such as the correlations between the input variables, their separation power and ranking, various classifier specific validations and finally efficiency versus background rejection curves for all trained classifiers. These criteria allow the user to choose the optimal classifier for the given problem. The package currently includes implementations of:

- Multi-dimensional rectangular cut optimisation using a genetic algorithm or Monte Carlo sampling;
- Projective likelihood estimation;
- Multi-dimensional likelihood estimation (k-nearest neighbour (k-NN) and probability density estimator range-search (PDERS));
- Linear and nonlinear discriminant analysis (Fisher, H-Matrix, Functional Discriminant Analysis);
- Artificial neural networks (three different multilayer percpetron implementations);
- Support Vector Machine;
- Boosted/bagged decision trees with pruning;
- Predictive learning via rule ensembles.

A detailed description of the individual classifiers including the configuration parameters available for their tuning is given in the TMVA Users Guide [1]. TMVA provides training, testing and performance evaluation algorithms, visualisation scripts and auxiliary tools such as parameter fitting and variable transformations.

2 Data Preprocessing, Training and Testing

Training and testing of the classifiers is performed with user-supplied data sets with known event classification. This data is given in form of either ROOT trees or ASCII text files. The data sets are divided into

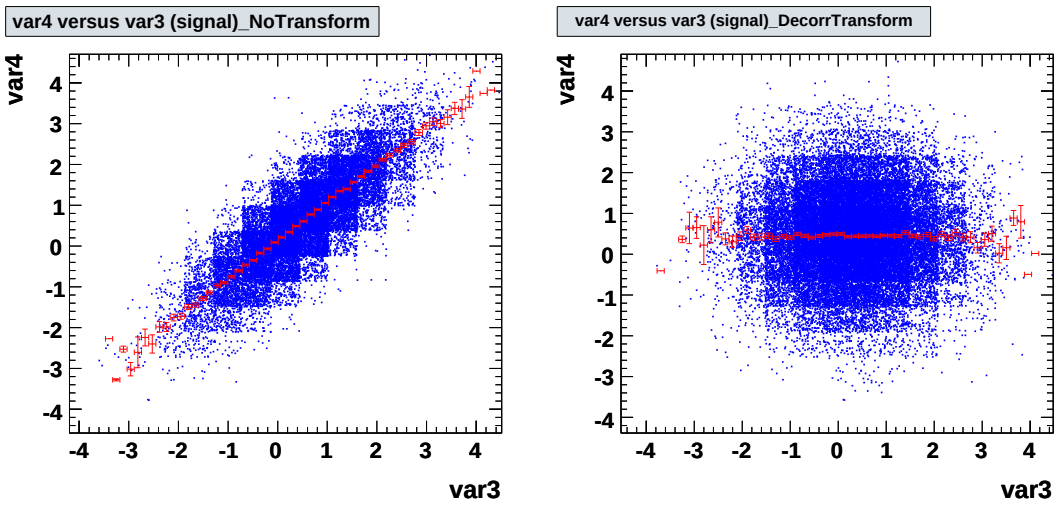

Fig. 1: Correlation between input variables. Left: correlations between var3 and var4 for signal training events from a Gaussian toy Monte Carlo. Right: the same after applying a linear decorrelation transformation.

statistically independent samples of training and testing data, omitting here an independent validation sample. Individual event weights may be attributed when specified in the data set. All classifiers see the same data sets and use the same prescription for the evaluation allowing for an objective comparison between them. A *Factory* class organises the interaction between the user and the TMVA analysis steps including preanalysis and preprocessing of the training data.

During the preanalysis, a preliminary ranking of the input variables is provided and their linear correlation coefficients are displayed. The variable ranking is later superseded by the ranking provided for each of the classifiers.

Preprocessing of the data set includes the application of conventional preselection cuts that are common for all classifiers. In addition one can apply two different variable transformations, decorrelation via the square-root of the covariance matrix and via a principal component decomposition. These transformations can be individually chosen for any particular classifier. Removing linear correlations from the data sample may be useful for classifiers that intrinsically do not take into account variable correlations as for example rectangular cuts or projective likelihood. A demonstration of the decorrelation procedure is shown in Fig. 1. It shows the decorrelation applied to a toy Monte Carlo with linearly correlated and Gaussian distributed variables that is supplied together with the TMVA package.

After the training, each classifier writes the entire information needed for its later application to weight files[1]. The classifiers are then tested and evaluated to assess their performance. The optimal classifier to be used for a specific analysis strongly depends on the problem at hand and no general recommendations can be given. To simplify the choice, TMVA computes and displays for each classifier a number of benchmark quantities such as:

- The *signal efficiency and background rejection* obtained from cuts on the classifier output. The area of the background rejection versus signal efficiency function is used for ranking the different classifiers.
- The *separation* $\langle S^2 \rangle$ of a classifier y, defined by the integral [2]

$$\langle S^2 \rangle = \frac{1}{2} \int \frac{\left(\hat{y}_S(y) - \hat{y}_B(y)\right)^2}{\hat{y}_S(y) + \hat{y}_B(y)} dy \,, \tag{1}$$

[1] A stand alone C++ code of the trained classifier which is independent of the TMVA libraries is also provided.

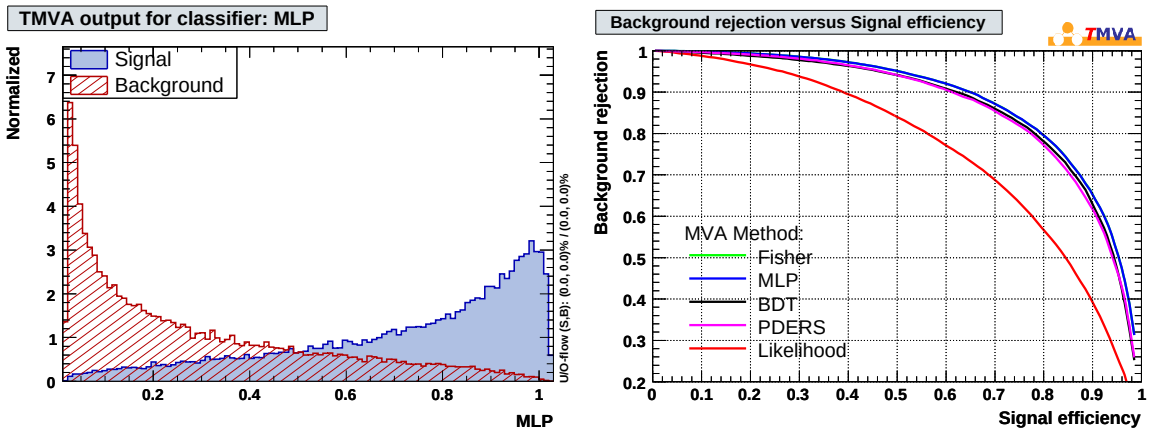

Fig. 2: The top left shows and example plot for classifier output distributions for signal and background events from the Neural Network (MLP) analysis on a toy Monte Carlo data sample. The background rejection versus signal efficiency obtained by cutting on the classifier output for the events of the test sample is shown at the top right.

where $\hat{y}_S$ and $\hat{y}_B$ are the signal and background PDFs of y, respectively. The separation is zero for identical signal and background shapes, and it is one for shapes with no overlap.

– The discrimination *significance* of a classifier, defined by the difference between the classifier means for signal and background divided by the quadratic sum of their root-mean-squares.

A cut placed on the classifier's output value y is typically used to classify an event as either signal or background. Upon user request TMVA also provides the classifier's signal and background PDFs, $\hat{y}_{S(B)}$. The PDFs can be used to derive classification probabilities for individual events. It is also used to compute the *Rarity* transformation.

– *Classification probability*: The probability for event i to be of signal type is given by,

$$P_S(i) = \frac{f_S \cdot \hat{y}_S(i)}{f_S \cdot \hat{y}_S(i) + (1 - f_S) \cdot \hat{y}_B(i)} \,, \tag{2}$$

where $f_S = N_S/(N_S + N_B)$ is the expected signal fraction, and $N_{S(B)}$ is the expected number of signal (background) events (default is $f_S = 0.5$).

– *Rarity*: The Rarity $\mathcal{R}(y)$ of a classifier y is given by the integral [3]

$$\mathcal{R}(y) = \int_{-\infty}^{y} \hat{y}_B(y')\, dy' \,, \tag{3}$$

which is defined such that $\mathcal{R}(y_B)$ for background events is uniformly distributed between 0 and 1, while signal events cluster towards 1. The signal distributions can thus be directly compared among the various classifiers. The stronger the peak towards 1, the better is the discrimination. Another useful aspect of the Rarity is the possibility to directly visualise deviations of a test background (which could be physics data) from the training sample, by exhibition of non-uniformness.

In addition, the variable distributions, correlation matrices and scatter plots, overtraining validation plots, as well as classifier specific information such as likelihood reference distributions, the neural network architecture and decision trees are conveniently plotted using ROOT macros executed via a graphical user interface that comes with TMVA. An example of the output is given in Fig. 2.

3 Classifier Application

The application of the trained classifiers to the selection of events from a data sample with unknown signal and background composition is handled via a light-weight *Reader* object. It reads and interprets the weight files of the chosen classifier and can be included in any C++ executable, ROOT macro or python analysis job.

For standalone use of the trained classifiers, TMVA also generates stand alone C++ response classes for most classifiers, which contain the encoded information from the weight files and the classifier's functionality. These classes do not depend on TMVA or ROOT, neither on any other external library.

4 Summary

TMVA is a toolkit that unifies highly customisable sophisticated multivariate classification algorithms in a single framework thus ensuring convenient use and an objective performance assessment since all classifiers see the same training and test data, and are evaluated following the same prescription.

Emphasis has been put on the clarity and functionality of the *Factory* and *Reader* interfaces to the user applications, which will hardly exceed a few lines of code. All classifiers run with reasonable default configurations and should have satisfying performance for average applications. It is stressed however that, to solve a concrete problem, all classifiers require at least some specific tuning to deploy their maximum classification capability. Individual optimisation and customisation of the classifiers is achieved via configuration strings that are detailed in [1].

TMVA is an open source project. The newest TMVA development version can be downloaded from Sourceforge.net at `http://tmva.sourceforge.net`. It is also part of the standard ROOT distribution kit (v5.14 and higher).

Acknowledgements

The fast growth of TMVA would not have been possible without the contributions from many developers listed as co-authors in the Users Guide [1]) and the crucial feedback from the user community. We thank in particular the CERN summer students Matt Jachowski (Stanford U.) for the implementation of TMVA's MLP neural network and Yair Mahalalel (Tel Aviv U.) for a significant improvement of PDERS. The Support Vector Machine has been contributed to TMVA by Andrzej Zemla and Marcin Wolter (IFJ PAN Krakow), and the k-NN classifier has been written by Rustem Ospanov (Texas U.). We also thank René Brun and the ROOT team for their support.

References

[1] A. Höcker, P. Speckmayer, J. Stelzer, F. Tegenfeldt, H. Voss, K. Voss, A. Christov, S. Henrot-Versillé, M. Jachowski, A. Krasznahorkay Jr., Y. Mahalalel, R. Ospanov, X. Prudent, M. Wolter, A. Zemla arXiv:physics/0703039 (2007).

[2] The BABAR Physics Book, BABAR Collaboration (P.F. Harrison and H. Quinn (editors) *et al.*), SLAC-R-0504 (1998); S. Versillé, PhD Thesis at LPNHE, `http://lpnhe-babar.in2p3.fr/theses/these_SophieVersille.ps.gz` (1998).

[3] To our information, the *Rarity* has been originally defined by F. Le Diberder in an unpublished Mark II internal note. In a single dimension, as defined in Eq. (3), it is equivalent to the μ-transform developed in: M. Pivk, "*Etude de la violation de CP dans la désintégration* $B^0 \to h^+h^-$ $(h = \pi, K)$ *auprès du détecteur BABAR à SLAC*", PhD thesis (in French), `http://tel.archives-ouvertes.fr/documents/archives0/00/00/29/91/index_fr.html` (2003).

StatPatternRecognition in Analysis of HEP and Astrophysics Data

I. Narsky
California Institute of Technology, Pasadena, CA, USA

Abstract

StatPatternRecognition (SPR) is a C++ package for supervised machine learning. Introduced in 2005, it has been used by several HEP and astrophysics collaborations, as well as non-academic researchers, for analysis of complex multivariate data. The package implements powerful classification algorithms such as boosting (three flavors), arc-x4, bagging, random forest, neural networks, decision trees (two flavors), bump hunter (PRIM), multi-class learner, logistic regression, linear and quadratic discriminant analysis, combiner of classifiers, and others. It also offers a suite of tools for data analysis: estimation of variable importance, bootstrap, cross-validation, computation of data moments, multivariate goodness-of-fit estimation, and others. SPR is a standalone package with an optional dependency on Root for data input/output. The user can access major SPR methods from an interactive Root session by loading the SPR shared library. The package is under active development and shows a growing number of users in the HEP community and elsewhere. The latest source release of the package can be obtained under General Public License from Sourceforge [1]. A full list of notes and talks about the package can be found on the author's web page [2].

1 Introduction

For several decades the HEP community has been using various classification methods to separate signal from background. Among these methods, only binary decision splits, also known as "cuts" in physics jargon, Fisher discriminant [3], and neural networks [4] have become truly popular. Stimulated by discussion at the Phystat workshops and related publications in physics journals and web archives, the community is now exploring new powerful classifiers recently introduced in the statistics literature. In particular, boosted decision trees [5] and random forest [6, 7] are becoming increasingly popular.

One cannot apply these advanced methods to physics data without software. In the past, physicists used to adapt packages from other communities or write their own implementations of desired algorithms. This practice is still ongoing. Both approaches require a substantial investment of manpower and often involve replication of effort. A package that can be used for physics analysis off the shelf should reduce this waste of effort to minimum.

What are the code requirements for such a package? First and foremost, it must be written in C++. It is by far the most popular choice among HEP researchers and the base for software frameworks maintained by large HEP collaborations. A package for multivariate classification must implement various methods and provide tools for comparison of their performance on the same dataset. Such a package should offer methods particularly useful for physics analysis, for example, optimization and monitoring of HEP-specific figures of merit (FOM's) such as the signal significance $S/\sqrt{(S+B)}$, a 90% upper limit and others. One of the distinctive features of HEP analysis is the enormous amount of experimental data available. Thus, the package should perform well on big datasets in many dimensions. This package needs to be interfaced to Root [8], a widely accepted framework for data storage and access within the HEP and astrophysics communities. Prior to SPR, such a package was not available.

2 Distribution

SPR is distributed as source code off Sourceforge [1] under General Public License. Installation instructions are included in INSTALL. Users reported successful builds on 32-bit (Scientific Linux 3 and 4, RedHat Enterprise 4, and Solaris 9) and 64-bit Unix platforms (Fedora and Debian). Enthusiasts have adapted the package to Windows and MacOS. SPR has been included as an extra package in Fedora Core 6 and later versions.

I am committed to supporting two versions of the package: a standalone version that uses ASCII text for input and output of data, and a Root-dependent version. The user can choose between the two versions during installation by setting an appropriate parameter of the configuration script. The ASCII version found consumers outside of the HEP community, while the Root-dependent version is more popular among physicists. No graphical tools are offered for the ASCII version of the package; however, one can go through the full analysis chain using ASCII output from SPR executables, as long as one can tolerate digesting information in the form of text tables instead of plots.

From day one it has been my goal to deliver a package ideal for CPU-intensive long-running job batch submissions. A typical HEP user trains boosted decision trees or random forest on datasets with up to millions of events in up to hundreds of dimensions. The package includes two dozen executables, one for each implemented classifier plus other analysis methods. For the most part, graphical tools have not been in the focus of development. However, recently I introduced SprRootAdapter — this class wraps SPR functionality in a shared library that can be loaded in Root. Now the user has access to major SPR methods from an interactive Root session. Scripts for making Root plots of various SPR quantities are also provided. For Root users, this should make the graphical analysis of SPR data much easier.

Documentation is supplied in the README file distributed with the package. All implemented methods and executables are described in sufficient detail.

3 Methods

It is impossible to fit a description of all SPR algorithms into a 4-page note. It is not necessary either because these algorithms have been described in many books on machine learning, advertised in several recent publications by physicists, and discussed at numerous seminars and workshops. Below is a brief summary of what is available.

The full analysis chain for a classifier of choice consists of training and testing. At the training stage, the user creates a trainable classifier and trains it for a specified number of cycles, either by supplying parameters for this classifier to one of the SPR executables or by using the SprRootAdapter interface from an interactive Root session. For a classifier that requires more than one training cycle, the user can monitor classification error computed for validation data. After the training is completed, the user can save the trained classifier configuration into a file. The saved configuration contains full information about the classifier. At any time the user can read the saved configuration from the file back into memory and either continue accumulating more training cycles or apply the trained classifier to test data. SPR allows to read configurations from several files and apply them to test data simultaneously turning the classifier comparison into an easy task.

SPR decision trees [9, 10] come in two flavors — a "regular" decision tree and a top-down tree. They use the same algorithm for training but store their configurations in different formats. A "regular" tree stores its terminal nodes as rectangular regions. If the number of nodes is small, this tree can be easily interpreted by a human. A top-down tree stores its configuration as a full path from the root of the tree. A top-down tree is faster because the lookup time grows as $\log(N)$ versus the number of nodes N while for the "regular" tree the lookup time grows linearly.

The bump hunter algorithm [11] finds one rectangular region in a multidimensional space by optimizing a chosen FOM. Both the bump hunter and the decision tree can optimize FOM's widely used in the machine-learning research such as Gini index or cross-entropy, as well as FOM's suited for physics

analysis such as $S/\sqrt{S+B}$, a 90% upper limit and others.

Boosting [5] works by adding many weak classifiers sequentially and increasing weights of misclassified events at each step. By focusing on events that are misclassified most of the time, boosting typically achieves a very good predictive power. SPR implements three flavors of boosting: Discrete AdaBoost, Real AdaBoost, and ϵ-Boost. Although boosted decision trees have lately become popular in HEP analysis, one can successfully boost other classifiers as well. Boosted binary splits and boosted neural networks are two other typical applications of boosting found in the machine-learning literature. SPR allows the user to boost an arbitrary sequence of classifiers.

The bagging (bootstrap aggregation) algorithm [6, 12] averages over many weak classifiers built on bootstrap replicas of a training set. SPR allows the user to bag an arbitrary sequence of classifiers. Bagged decision trees have been often used in machine-learning research and are now being applied to physics analysis, in particular, for particle identification at *BABAR*. Another popular method, bagged neural networks, will hopefully find its way into physics analysis as well.

Random forest [7], typically used in conjunction with bagging, represents a set of decision trees. Each tree is built using randomly selected input variables for each decision split. Random sampling of input variables reduces correlation among the decision trees and improves the overall classification power. This method has been applied with success to several *BABAR* physics analyses.

SPR implements a feedforward backpropagation neural net with a logistic activation function [4] well known to physicists.

SPR implements a tool for combining several powerful classifiers. One can train several classifiers on subsets of input variables and then train a global classifier in the space of their outputs. The user needs to specify how the classifiers are to be combined through a configuration file.

All algorithms described above can be only used for separation of two classes, signal and background. A multi-class method [13] reduces a problem with an arbitrary number of classes to a set of two-class problems and then converts the solutions to these binary problems into an overall multi-category classification label.

SPR offers various methods for estimation of variable importance. For decision trees, the importance of an input variable can be estimated by adding changes in the optimized FOM due to decision splits on this variable. For any classifier, the importance of an input variable can be estimated by randomly permuting class labels across this variable and estimating an increase in the overall classification error due to this permutation.

SPR implements other tools for data analysis that will not be described here due to limited space.

4 Examples of Use

In 2005 the SPR implementation of the random forest algorithm was applied to muon identification at *BABAR*, and a significant improvement over the traditional neural-net approach was achieved. This exercise, presented at the CHEP 2006 and ACAT 2007 conferences, instigated development of an SPR muon selector. At present the *BABAR* PID group is working on electron, proton and kaon SPR selectors as well. An SPR-based K_L selector is also available. These selectors outperform likelihood-, cut- and neural-net-based selectors that are still used in *BABAR*. An SPR-based electron selector has been recently introduced at CMS.

Several *BABAR* physics analyses use SPR methods for background suppression. A search for $B^+ \to K^+\nu\nu$ and a measurement of the branching fractions for the decays $B \to \rho\gamma$ and $B \to \omega\gamma$ exploit the SPR bump hunter algorithm to find the optimal combination of orthogonal cuts and apply the random forest method with dozens or hundreds of input variables to achieve the ultimate classification power. SPR boosted decision trees are used in a search for $B^+ \to \tau^+\nu$ and a measurement of exclusive $b \to s\gamma$ modes.

There are several published results as well. SPR boosted decision trees and random forest were applied to identify supernovae [14]. 32 input variables were used to separate signal modeled as fake supernovae inserted into real sky images from background. The decision tree methods reduce background by 1-2 orders of magnitude in a broad range of the true positive identification rate compared to the traditional threshold-cut approach and a support vector machine classifier.

Another study [15] used SPR boosted decision trees for tagging of b-jets. $W \rightarrow l\nu q\bar{q}$ events were generated by simulating the environment of the LHC collider with $p\bar{p}$ collisions at a center-of-mass energy of 14 TeV. The training sample consisted of 50k b-jet events and 50k u-jet events with 7 input variables. For moderate b-tagging efficiencies, the boosted decision trees improve the u-jet rejection by several dozen percent compared to the multi-layer perceptron implemented in Root.

This is by far an incomplete list of analyses using SPR. I have little or no knowledge of how the package is used by collaborations other than *BABAR* or CMS, and even less so by analysts outside the HEP and astrophysics communities.

5 Acknowledgments

Many people have contributed to SPR development by submitting code, offering advice, providing feedback about the package installation and use, and by presenting the package at conferences on my behalf. A full list of code contributors can be found in the AUTHORS file included in the package. I would like to thank the organizers of Phystat07 for giving me an opportunity to present this work at the conference.

References

[1] `http://sourceforge.net/projects/statpatrec/`.

[2] `http://www.hep.caltech.edu/~narsky/spr.html`.

[3] R.A. Fisher, *The use of multiple measurements in taxonomic problems*, Annals of Eugenics **7**, 179-188 (1936).

[4] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1999.

[5] Y. Freund and R.E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, J. of Computer and System Sciences **55**, 119-139 (1997); J. Friedman, T. Hastie and R. Tibshirani, *Additive Logistic Regression: a Statistical View of Boosting*, Annals of Statistics **28(2)**, 337-407 (2000).

[6] L. Breiman, *Bagging Predictors*, Machine Learning **26**, 123-140 (1996).

[7] L. Breiman, *Random Forests*, Machine Learning **45**, 5-32 (2001).

[8] `http://root.cern.ch/`.

[9] L. Breiman et al., *Classification and Regression Trees*, Waldsworth International, 1984.

[10] I. Narsky, *StatPatternRecognition: A C++ Package for Statistical Analysis of High Energy Physics Data*, physics/0507143 (2005).

[11] J. Friedman and N. Fisher, *Bump hunting in high dimensional data*, Statistics and Computing **9**, 123-143 (1999).

[12] I. Narsky, *Optimization of Signal Significance by Bagging Decision Trees*, physics/0507157 (2005).

[13] E.L. Allwein, R.E. Schapire and Y. Singer, *Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers*, J. of Machine Learning Research **1**, pp. 113–141 (2000).

[14] S. Bailey et al., *How to Find More Supernovae with Less Work*, The Astrophysical Journal **665**, 1246-1253 (2007); arXiv:0705.0493 [astro-ph].

[15] J. Bastos, *Tagging heavy flavours with boosted decision trees*, arXiv:physics/0702041; J. Bastos and Y. Liu, *A Multivariate approach to heavy flavour tagging with cascade training*, arXiv:0704.3706 [physics.data-an].

Concluding remarks

PhysStat-LHC Conference Summary

Robert D. Cousins
Dept. of Physics and Astronomy, University of California, Los Angeles, California, USA

Abstract
This timely conference in the PhyStat series brought together physicists and statisticians for talks and discussions having an emphasis on techniques for use at the Large Hadron Collider experiments. By building on the work of previous generations of experiments, and by developing common tools for comparing and combining results, we can be optimistic about our readiness for statistical analysis of the first LHC data.

1 Introduction

In attempting to summarize the content of such a dynamic conference, for my commentary I selected a subset of the many talks, based either on the significance, or simply on my somewhat arbitrary interest. I have also tried to make some connections to earlier PhyStat meetings.

Data analysis at the LHC will benefit from, and build on, the vast experience coming from LEP, B factories, the Tevatron, and earlier experiments. The task during the next year is to consolidate this experience into common tools that the LHC experiments can use, while continuing to add to them. Already there has been progress in this area, and we have good reason to expect that ATLAS and CMS will be better positioned to compare and combine results than were experiments at first turn-on of other accelerators. I begin by mentioning a few selected topics, then discussing some aspects of the Bayesian analyses presented, and then turning to the more global issues of statistics at the LHC.

2 Topical Talks

2.1 P values and Nuisance Parameters

Luc Demortier has written an extensive review (174 pages!) [1] on p values (roughly speaking, the probability of obtaining a value of a test statistic as extreme or more extreme than that observed), including many aspects of the inclusion of nuisance parameters [3]. I take the opportunity to mention as well my recently posted annotated bibliography [4] on combining p-values. Especially since p-values are easy to misinterpret, Demortier's work deserves to be widely read.

One of many issues we face in computing p-values is what is the best way to enumerate all the possibilities used in computing the probability entering into the p-value, while accounting for the effect of the many places one looks. This is coupled to the issue of what value of a p-value corresponds to a "discovery". A. Drozdetskiy (with A. Korytov and G. Mitselmakher) and W. Quayle each described ways to account for the multiple Higgs masses at which one looks for a signal; an interesting practical issue is to what extent this effect can be factorized out of the more complicated analysis.

I neglected to mention in my talk what I perceive to be an alarming propagation in HEP generally of the notion that there can be universal values of p-values which correspond to "evidence", "discovery", and other words in the scientific process. Without getting into even more fundamental objections to p-values, I hope that it is clear (to paraphrase Carl Sagan) that the more extraordinary the claim, the more extraordinary the evidence must be, so that there cannot be a universal p-value for all claims.

2.2 Weighting Background-Subtracted Events

Jim Linnemann (with Andrew Smith), also citing Roger Barlow and F. Tkachov, discussed optimal weighting. Historically, HEP seems to have under-utilized these "direct" calculational ways of well-

approximating a maximum-likelihood estimate. Is our computing now so advanced that we have less use for it? Even if so, it would seem that this should be part of our statistics toolkit.

2.3 Banff Challenge on Upper Limits, and other studies

Joel Heinrich reported on the performance of methods employed by many physicists and statisticians in a challenge in which participants were asked to provide upper limits or 2-sided intervals on cross sections measured in a counting experiment with nuisance parameters. Heinrich then evaluated the results by both frequentist and Bayesian criteria. This fascinating study has a lot of food for thought. In a related paper Tucker (with myself) evaluated the frequentist performance of several algorithms, in particular high-lighting the little-known application of the binomial test to a common problem [6]. Also, Rolke (with Lopez) presented studies using the likelihood ratio test statistic.

2.4 Design of Experiments

Statistician Nancy Reid reviewed some of the theory of experimental design. This is another example where most high energy physicists seemed to have missed a whole area of study that has some relevance to our work. In particular, as emphasized in a talk by Jim Linnemann, it seems that the usual way of checking the influence of variations in parameters has missed important cross terms. We should all take a look at these talks and the references.

2.5 The "Other PDF's"

As he did at the 2002 conference in Durham [8], Robert Thorne (the "T" of MRST PDFs) reviewed the status of calculating uncertainties on Parton Distribution Functions. This is *still* a very tough business, which is however important for our experiments. Any consumer of these uncertainties (especially one contemplating interpreting them literally out to several sigma) is well-advised to learn how hard this problem is. Since "For Global Fits, using $\Delta\chi^2 = 1$ is not a sensible option", CTEQ uses $\Delta\chi^2 \sim 100$, and MRST/MSTW use $\Delta\chi^2 \sim 50$, for 90% C.L. intervals (for which the book value of $\Delta\chi^2$ is 2.7). This is in the spirit of a (large!) PDG scale factor, and points to inconsistencies in the input data and/or the model.

2.6 Multivariate Methods

Multivariate methods using machine learning techniques have been a very common theme in the PhyStat workshops since the Durham workshop in 2002 (where Harrison Prosper provided a useful overall perspective on the several methods discussed there). At PhyStat 2003 (SLAC) and 2005 (Oxford), we were fortunate to have one of the world's experts, Jerome Friedman, actively involved. At the present conference, we heard talks on packages implementing many of these methods, TMVA (Fredrik Tegenfeldt and collaborators) and SPR (Ilya Nasky). General frameworks (ROOT, RooStats) for incorporating these and many other tools were described by Lorenzo Moneta and by Wouter Verkerke; I return to these in my discussion below.

2.7 A Statistician's view of Nuisance Parameters

Statistician Radford Neal emphasizes the likelihood principle, and therefore uses the likelihood function as input to a classifier. He urges us not to use frequentist confidence intervals, particularly not two-sided ones. He integrates out the nuisance parameters using a prior.

2.8 Use of Bayes' Theorem for Particle ID

Iouri Belikov, while presenting the statistical "wish-list" of the Alice collaboration, showed a nice application of Bayes' Theorem to particle identification. It reminded me of a similarly nice application at

the 2002 Durham Conference [9], and I have the same comment [10], namely a semantic one: while this technique was called "Bayesian", it would appear to be perfectly valid with the frequentist definition of probability.

Bayes' Theorem applies to any P which obeys the axioms of probability, including both the degree-of-belief P commonly referred to as "Bayesian" and the frequency definition of P more commonly used in HEP. The example of Belikov would seem to be perfectly consistent with the frequentist definition of P, and hence pleasing to frequentists and Bayesians alike. At PhyStat 2003, statistician Bradley Efron put it this way: "Bayes' rule is satisfying, convincing, and fun to use. But using Bayes' rule does not make one a Bayesian; *always* using it does, and that's where difficulties begin." [11]

3 Discussion of Bayesian Methods

3.1 Cox's Five faces of Bayesian statistics (and the sixth from HEP)

Renowned statistician David Cox, in a stimulating talk, compared a number of approaches to the problem of inference when there are many parameters or many hypotheses. I chose for the summary talk one slide in which he described five types of Bayesians among statisticians. What is notable is that typical HEP Bayesians do not fall into any of these categories, if they are using priors which are uniform in arbitrary variables (sometimes claimed to be preferred on the grounds that they are "fundamental" or "what is directly measured"). This has been tolerated, I think, partly because typically the likelihood overwhelms the prior (and frequentist coverage is in the end good), and partly because flat priors for the Poisson mean yield upper limits with conservative frequentist properties. However, it is unfortunate that there are workers in HEP who are using (or even advocating) Bayesian techniques but who are completely unfamiliar either with the subjective Bayes foundations of Savage and De Finetti, or with writings on non-subjective priors such as those of James Berger and the review of Kass and Wasserman [12]. The Jeffreys prior does not seem to be commonly used in HEP, and I am not aware of any examples in HEP of the use of the Reference Priors of Bernardo and collaborators, although Luc Demortier has advocated their use [13] (and in this conference put such software on our wish-list for statisticians).

Furthermore, the flat priors used in HEP can be susceptible to ill behavior in high dimensions, where one can easily add undesired "information" without realizing it. As Bradley Efron noted at PhyStat 2003, "Perhaps the most important general lesson is that the facile use of what appear to be uninformative priors is a dangerous practice in high dimensions."[11]. Joel Heinrich gave a specific example relevant to HEP at PhyStat 2005 at Oxford, noting a problem encountered with a multi-dimensional nuisance parameter and observing, "In hindsight, this should have led us to distrust a prior flat in multiple dimensions, since this is well known to lead to problems" [14].

Thus, I had a somewhat questioning reaction to the talk by Leszek Roskowski on "A Bayesian approach to Constrained MSSM", but I hope in a constructive way. The problem described is an important, difficult one, namely trying to synthesize particle and cosmological data in order to constrain supersymmetric models. Closely related work was presented by Lafaye (with Plehn, Rauch, Zerwas) regarding Sfitter. While I have not studied the physics inputs and models for these talks, it would appear that the latter talk explored more of the "space" of methods, as I would advocate. Once one has the likelihood function, one can obtain either approximate frequentist confidence regions via the profile likelihood (MINUIT MINOS in HEP), or Bayesian credible regions by adding priors and integrating. As has been much discussed at past PhyStat conferences, the first step (plotting contours of the likelihood) is always useful, if only to be used as comparison with other methods. For a Bayesian analysis, before multiplying the likelihood by the prior, it can be very instructive to take only the multi-dimensional priors, marginalize over the nuisance parameters, and see what one is left with, i.e., the posterior one would obtain if the likelihood function were constant.

It would seem that only after performing these two exercises is one truly prepared to multiply the likelihood and the prior, and start integrating. With a sensitivity analysis to compare various priors, and

with comparison to the profile likelihood answer, one can understand if one is faced with a pathological situation (for example where the likelihood has a spike that is so narrow that there is negligible area under it in any reasonable metric), or is in asymptopia (where all methods agree), or somewhere in between.

Statistician Paul Baines (with Xiao-Li Meng) gave an enticing, if somewhat sobering talk on Probability Matching Priors, i.e., priors which lead to posterior intervals with good frequentist coverage. A bottoms-up approach is extremely difficult. Meanwhile in HEP we have gained quite a bit of experience regarding specific cases when Bayesian calculations give reasonable coverage; it important to continue to do so.

3.2 James Berger on Bayesian analysis: objectivity, multiplicity and discovery

It was a pleasure to have statistician Jim Berger back, as he was first introduced to our community at the Fermilab Confidence Limits Workshop in 2000 [15]. He is a leading proponent of the "Objective Bayes" approach in which one uses Bayesian techniques (thereby building in the likelihood principle and consistent treatment of probabilities once the all-important priors are chosen) with priors which do not always represent personal belief, but rather are chosen by some formal rules.

One striking aspect of Berger's talk is that, for an unknown binomial parameter, it is obvious to him that the objective prior to use is the Jeffreys prior, from both the objective Bayesian (invariance) and frequentist (approximate coverage) points of view. And yet, in HEP, on several occasions I have seen people seeking a non-informative prior for a binomial parameter and without any thought taking the uniform prior. In higher dimensions, Berger advocates the use of Reference Priors, and we discussed with him how useful it would be to have some software tools for this.

Berger conveyed a key part of his message to us in both 2000 and 2007. In 2000 [15], Berger said, "What should be the view today: Objective Bayesian analysis is the best frequentist tool around." (This was after quoting M.G. Kendall, whom we in HEP know best via his book with A. Stuart, as giving the 'old' frequentist viewpoint of Bayesians: "...if they [Bayesians] would only do as he [Bayes] did and publish posthumously we should all be saved a lot of trouble." [16].) An important part of his message this year is that "Good versions [of Objective Bayes] are argued to yield better frequentist answers than asymptotic frequentist methods", concluding that "There is great appeal to simultaneously being objective Bayesian and frequentist."

In Durham in 2002, we had the pleasure of interacting with a Bayesian statistician of a different flavor, Michael Goldstein [17], who is solidly in the (personalistic) subjective camp. He understood of course that one cannot publish only a posterior probability based on one's personal subjective prior. The key point, which I believe our community has been rather feeble in undertaking, is to study the sensitivity of the result to changing the prior. In the 2002 Proceedings, Goldstein says, "Again, different individuals may react differently, and the sensitivity analysis for the effect of the prior on the posterior is the analysis of the scientific community, so that the answer should now be an interval of posterior values which may be reasonably held by individual scientists...In this view, a sensitivity analysis over the reasonable a priori judgments of the scientific community gives the full analysis." I copied from his transparencies at the time a slogan which I think Bayesians in HEP should take to heart: "Sensitivity Analysis is at the heart of scientific Bayesianism."

As we go forward in HEP, I hope that high energy physicists using Bayesian methods will look to both of these points of view for understanding.

4 Collider Physics

Complementary overview talks with lots of food for thought were presented by Wade Fisher (Tevatron methods), Kyle Cranmer (practical problems in LHC searches), and Eilam Gross (ATLAS+CMS wish-list), with related talks by Yuehong Xie (LHCb wish-list) and Iouri Belikov (ALICE wish-list). As to summarize these important talks would be tantamount to repeating them, I urge everyone to consult the

writeups in these proceedings.

In trying to synthesize all the experience from past experiments, we have the sociological lessons as well as the statistical ones. At the 2002 Durham workshop, Chris Parkes provided some fascinating insight into the combination of LEP results [18], and the Tevatron experience demonstrated similar issues. Meanwhile, as a result of the PhyStat conferences and our interaction with statisticians, we have learned a lot about the technical and foundational aspects of many of our methods. Like the statisticians before us, we seem to be getting over the hump of foundational wars and becoming pragmatists, and we are getting some residual skirmishes out of the way before we have data. And we understand the necessity to compare and combine results in order to maximize return on society's huge investment in us.

For me the ideal situation, which is indeed already underway, is for ATLAS and CMS to have a technical framework in which results can be compared and combined in a transparent way, while allowing for differences of opinion about which method is preferred. One of the key aspects is to make it "easy" for an experimenter to compute a result (statistical significance of an effect, a measured value, or an interval) by multiple techniques, so that the consuming physicist is not confined to the narrow preferences of one analyst.

In this respect, I am quite enthusiastic about all the work (by many people) described in talks by Lorenzo Moneta and Wouter Verkerke. From the ROOT environment, one will be able to perform analyses, share the results, and combine analyses, both for multiple channels within one experiment, and with other experiments. Since the workshop, progress has continued in this direction, with involvement of physicists from both ATLAS and CMS.

If we take interval estimation (including nuisance parameters) as an example, the three main classes of methods were discussed during the workshop:

- Profile likelihood, known in HEP as the MINUIT MINOS method, based on likelihood ratios (differences in log likelihood), without attaching a metric to the unknown parameters.
- Bayesian methods, based on the likelihood function, with metric attached via the prior pdf.
- Frequentist confidence intervals, either constructed a la Neyman, or by a technique meant to assure frequentist coverage.

It is common to mix aspects of the methods, for example integrating out some nuisance parameters in a profile likelihood or frequentist confidence interval treatment. The RooStats framework is gathering momentum as a forum where technical implementations of all of the above techniques (and popular variants thereof) can be implemented with a common interface. Our community could then effectively demand that a result derived from one technique also be derived from the other techniques, and that the sampling properties be studied.

This will help to educate students and veterans alike. When the methods agree, one is happily in asymptopia; when the methods disagree, one will be reminded that the methods answer different questions and have different definitions of probability. Bayesian answers depend on the prior and can have poor frequentist coverage properties, while frequentist confidence intervals typically violate the likelihood principle and the probability of containing the true value is a property of the set, not of any one interval. Furthermore, as more advanced methods continue to be developed and are "plugged in", in this environment one should be able to evaluate the new methods in a controlled way.

All this points to a future which I believe will be quite productive, as we eagerly await first data from the LHC. By the time of the next PhyStat, we expect to have to have *real* LHC data on which to demonstrate our techniques!

5 Thanks

On behalf of all participants of the meeting, I am pleased once again to thank Louis Lyons for his continuing efforts to organize this series of workshops, and to thank his co-organizer of this meeting,

Albert De Roeck. On behalf of the physicists, we thank again the statisticians who helped educate us and showed only good-natured tolerance as we sometimes abused their discipline's techniques and principles.

This work was partially supported by the U.S. Dept. of Energy and by the National Science Foundation.

References

[1] Luc Demortier, "P Values: What They Are and How to Use Them", CDF/MEMO/STATISTICS/PUBLIC/8662 (June 2007) http://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf.

[2] Proceedings of Phystat 2005 Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology, Oxford, England, 12-15 Sept 2005, Imperial College Press, http://www.physics.ox.ac.uk/phystat05/proceedings/default.htm.

[3] Robert D. Cousins, "Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature", in Ref. [2].

[4] Robert D. Cousins, "Annotated Bibliography of Some Papers on Combining Significances or p-values", arXiv:0705.2209 [physics.data-an].

[5] Proceedings of PHYSTAT2003 Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology, SLAC, 8-11 Sept 2003, http://www.slac.stanford.edu/econf/C030908/.

[6] James T. Linnemann, "Measures of significance in HEP and astrophysics," in Ref. [5]; [arXiv:physics/0312059].

[7] Proceedings of the Conference on Advanced Statistical Techniques in Particle Physics, Durham, England, 18-22 Mar 2002, Report number IPPP/02/39, http://www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings.shtml.

[8] R.S. Thorne, "Uncertainties in Parton Related Quantities", in Ref. [7].

[9] T. Deyoung and G.C. Hill, "Application of Bayes' Theorem to Muon Track Reconstruction in Amanda", in Ref. [7].

[10] Robert Cousins, "Conference summary Talk", in Ref. [7].

[11] Bradley Efron, "Bayesians, Frequentists, and Physicists", in Ref. [5].

[12] Robert E. Kass and Larry Wasserman, "The Selection of Prior Distributions by Formal Rules", J. Amer. Stat. Assoc. 91 1343 (1996). http://lib.stat.cmu.edu/~kass/papers/rules.pdf.

[13] Luc Demortier, "Bayesian Reference Analysis", in Ref. [2].

[14] Joel Heinrich, "The Bayesian Approach to Setting Limits: What to Avoid", in Ref. [2].

[15] Workshop on Confidence Limits, Fermilab, 27-28 March 2000, http://conferences.fnal.gov/cl2k/.

[16] M.G. Kendall, "On the Future of Statistics–A Second Look", J. Royal Stat. Soc. Series A 131 182 (1968). The quote is from p. 185. The context is an extended complaint about too many papers being published, for a variety of reasons that I think we physicists might recognize today as well.

[17] Michael Goldstein, "Why Be a Bayesian", in Ref. [7].

[18] C. Parkes, "Practicalities of Combining Analyses: W Physics Results at LEP", in Ref. [7].

The Early History of Bayesian Ideas

F. James
CERN, Geneva, Switzerland

Abstract
A brief after-dinner talk presented at the PhyStat LHC Workshop, in which I present some unexpected Bayesian thinking dating back to the Middle Ages.

As I am often accused of being an incorrigible frequentist, I thought I should do some more studying of the Bayesian methodology and in particular the early history and foundations of Bayesian thinking.

I found to my great surprise that many of the ideas and even the terminology I thought originated in the 18th, 19th and 20th centuries can actually be traced back long before Reverend Thomas Bayes' famous paper on the Doctrine of Chances.

The earliest traces I could find are from the 12^{th} and 13^{th} centuries. In those days in Europe, there were many ways in which one could lead a religious life. Probably the most devoted servants of the faith were the monks, who lived in monasteries (as their name implies), and the friars, who led humble lives in the outside world. The name of the latter group derives from the French *frère*, or brother.

Although we tend to have a romantic view of monasteries now, by all accounts life in the monasteries was not very comfortable. Everything was in stone or hard wood, and meals were taken while seated on long wooden or even stone benches. Although quite uncomfortable, these benches were very important, for it was here that one encountered the highest posterior densities.

The friars, on the other hand, were not concentrated in monasteries, so the friar density was much more spread out and uniform, as it should be. Not completely uniform, of course, because friars were believers, so the friar density reflected the degree of belief, or faith, for a given region.

In the beginning, friars were supposed to have no possessions and live from begging alone, but this was not an entirely workable arrangement, so most of them eventually took on regular jobs. Those who worked in the library were known as reference friars, and those who did ironing were called flat friars.

As it must happen with any social group, some friars, and indeed those most often encountered in public, were accused of improper behaviour. These improper friars soon became a source of scandal, starting with their provocative dress often characterized by considerable undercoverage. There was some discussion about how much coverage a friar should have, and it was decided that, at the very least, their posteriors should have adequate coverage.

Finally, one case was reported of a friar with no coverage at all! This friar was arrested, but was later released for lack of evidence.

There was a suggestion to organize the friars into groups, with leaders that would oversee the behaviour of the group members, but the idea of hierarchical friars was not well received. Finally, a physicist friar by the name of Jeffrey decided to form his own group of friars known as Jeffrey's friars, who would promise to behave themselves better. Most importantly, their behaviour was to be invariant.

But even if it was invariant, the behaviour of some of Jeffrey's friars was still improper. Moreover, they were accused of being unprincipled, since they refused to obey an obscure religious dogma known as the Principle of Likelihood.

Just at this time an additional problem arose with those friars who (like many monks) had taken vows of silence. These friars were called noninformative friars, and there was considerable discussion about just how noninformative they were. An influential author in Valencia even published a pamphlet with the provocative title "Noninformative Friars do not Exist!"

Further problems arose as an unexpected group, the Multidimensional Friars, exhibited a new form of unacceptable behaviour: inconsistency. But by this time, the Reformation was in full swing in much of Europe, and in the confusion that followed, the trail of early Bayes history has been lost.

Fred James,
with help from Bob Cousins, Kyle Cranmer and Jim Linnemann

Appendix

LHC Statistics for Pedestrians

Eilam Gross
Weizmann Institute, Rehovot, Israel

Abstract
A pedestrians guide aimed at the LHC laymen statisticians is presented. It is not meant to replace any text book but to help the confused physicist to understand the jargon and methods used by HEP Phystatisticians[1]

1 Introduction

The first Phystat meeting was a workshop at CERN on Confidence Limits followed by a similar workshop at Fermilab. Fred James who organized the meeting with Louis Lyons presented then his personal wish list titled: 'What I would like to see'. Fred wishes that physicists learn the vocabulary of statistics. This pedestrian guide is aimed at the Atlas and CMS physicists who wish to become Phystatisticians so that when ATLAS or CMS publish a combined limit or discovery significance they will know what it is all about.

When interpreting the result of the experiment, there are two alternate questions and one must not confuse between them. Question number one would be: Did I or did I not establish a discovery? Question number two would be: How well does my alternate model describe this discovery? The first question has to do with the goodness of the fit of the observed data to the good and old Standard Model while the second question has to do with hypotheses testing and the derivation of confidence intervals and upper limits. The LHC physics community is not only a mixture of physicists speaking all sorts of languages, from Hebrew and Chinese to English, German and French but who are also refugees of all sorts of experiments each with its preferred statistical method. Physicists educated at LEP advocate the CLs method while some Tevatron physicists prefer Bayesian methods with some of their friends from BaBar and Belle using pure frequentist methods. It seems that the only way out is to do it all... But, in a way, as we will show, conceptually one way leads to another.

But in order to introduce the different methods and compare them a basic lesson in the related statistics jargon is necessary.

2 Test Statistics

A test statistic is a quantity calculated from our sample of data. Its value can be used to estimate how probable is the result that we observe with respect to some null hypothesis. A physicist's intuition will attribute the null hypothesis to the 'background only' hypothesis. Normally it depends on the nature of the problem, but in this write up we will stick to this definition. In this context the value of the test statistic is used to decide whether or not the null hypothesis should be rejected in our hypothesis test.

It is important to note that the observed test statistics is based on our ONE experiment and could be a result of years of data collecting! Normally to conclude anything based on the observed test statistic one needs the pdf of the test statistic. This can sometimes be calculated analytically but can always be generated with toy Monte Carlo experiments.

A consequence of the Neyman-Pearson lema is that if H_0 is the null hypothesis (background only) and H_1 is the alternate hypothesis (say, a Higgs Boson with a mass m) then the most powerful test statistic one can construct (in absence of systematics) is the Likelihood ratio

$$Q(m) = \frac{L(H_1)}{L(H_0)} = \frac{L(s(m)+b)}{L(b)} \tag{1}$$

[1] A Phystatistician is a Physicist who knows his way in Statistics and knows how Kendall's advanced theory of statistics book looks like....

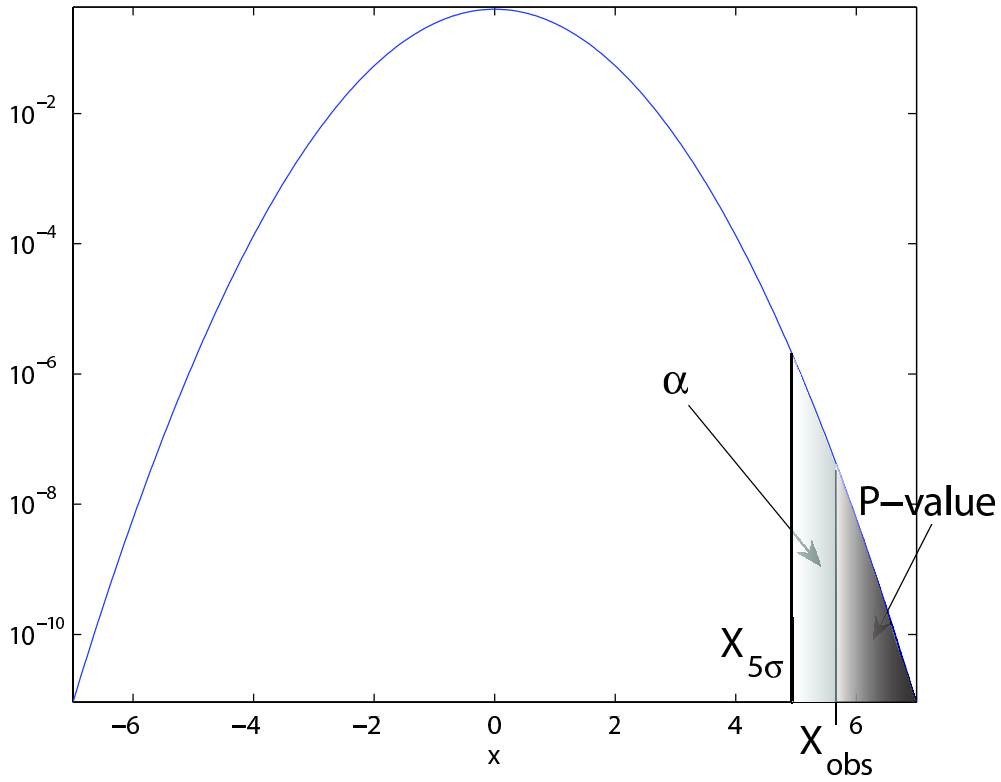

Fig. 1: An illustration showing the control area α and the $p-value$ of a Gaussian distribution. Note, in this example $X_{obs} > X_{5\sigma}$.

In a counting experiment s and b would be the average number of the expected signal and background events and the Likelihoods would be derived from the data using Poisson statistics.

3 p-value

At LEP, trying to discover the Higgs boson, people examined the distribution of the observed $1 - CL_b$ as a function of the hypothesized Higgs mass and looked for troughs.... That might have been the right thing to do but the wrong statistical jargon. A discovery by definition is a deviation from the Standard Model, i.e. the "background only" hypothesis (H_0). Given the pdf of the test statistic for background only experiments, it is common in HEP to announce a discovery if the result is at least 5 σ away from the expectation. Given a pdf $g(x|H_0)$ of the test statistic x, one can define a control area of size α at the tail of the pdf distribution (for this example let us assume that the less probable result is on one side of the distribution only), i.e. $\alpha = \int_{x_{5\sigma}}^{\infty} g(x|H_0)dx$ (Figure 1). If the observed result $x_{obs} > x_{5\sigma}$ then the probability to get a result which is as or less compatible with the background hypothesis is given by $p = \int_{x_{obs}}^{\infty} g(x|H_0)dx$ and it is smaller than α. This probability is called the $p-value$ and a discovery is considered when $p < \alpha$. This means also that the background-only hypothesis is rejected with a probability of $1 - p$.

Historically physicists have the tendency to mix confidence with p-value. In looking for the Higgs, the LEP experiments used the 'confidence level in the background, $1 - CL_b(m_H)$, where $CL_{b(m_H)}$ is defined as the tail area $CL_b = \int_{-\infty}^{x_{obs}} g(x|H_0)dx$, with the statistic x being the log likelihood ratio for the background plus signal model (i.e. Standard Model with Higgs of mass m_H) as compared with background only (i.e. H_0, the Standard Model with no Higgs in the observable mass range). This

implies that in an ensemble of backgound only experiments, a fraction $1 - CL_b$ would be expected to have a larger value than the observed value. The terminology is confusing since $1 - CL_b$ is in fact a p value. The LEP experiments were looking for tiny values of $1 - CL_b$, which would indicate a very large fluctuation of the background (or the presence of a signal), but none was found. The correspondence between the hypothesis test property $1 - p$ and the background confidence estimation, CL_b is further discussed in [1].

4 The Look Elsewhere Effect

The Standard Model predicts a Higgs Boson but not its mass. It can be anywhere up to a few hundreds of GeV. We can specify an hypothesis with a specific Higgs mass but had we observed some possible signal we should take into account that this signal could be a fluctuation which could be observed anywhere in our sensitivity range [2]. Here we change the signal hypothesis from a Higgs with a specific mass m_H to a Higgs with some mass in the observed region. It is not clear how to take these effects into account. One common way is to degrade the observed p-value by multiplying it by the size of the sensitivity region divided by the experimental resolution. A common claim is that the control region for discovery is so small that "who cares".... Another common belief is that the "look elsewhere effect" is the reason for the habit of defining a discovery as a 5σ and not for example 4σ, because even if you quote 5σ your effective significance is lower.

5 Confidence Intervals and Coverage

Assume you have a measurement m_{meas} of m with m_t being the true value of m and suppose you know the pdf $p(m_{meas}|m)$. You use some method to calculate a 90% confidence interval $[m_1, m_2]$. What does it mean?

Most physicists interpret it as if the probability that there is a Higgs Boson with $m_t \in [m_1, m_2]$ is 90%. However, this is totally wrong. If you run a bunch of toy Monte Carlo experiments, each one will yield a different interval. The correct statement is that if there is a Higgs with a mass, m_t, then, in an ensemble of experiments, 90% of the obtained confidence intervals will contain the true value of m, m_t. More on the source of this misconception in section 6.

Subsequent to the above definition of interval is the notion of coverage. The confidence interval is estimated using the physicist preferable method. If in an ensemble of Monte Carlo experiments the true value of m is covered within (e.g.) 90% of the estimated confidence intervals, we claim a coverage. If it occurs less than 90%, the method is claimed to undercover.

Some physicists doubt the importance of coverage. Their claim is that coverage answers the wrong question. What we really want to know, so they claim, is the probability that the Higgs Boson exists and is in the specified mass interval. So there are two possibilities here. Either educate the physicists about the correct meaning of coverage or try to answer the "right" question...

6 Subjective Bayesian

What is the "right" question? It must be: Is there a Higgs Boson? When pronouncing this question, I cannot escape from an immediate association to the question: Is there a God? Can one really answer this question based on the data (earth)? The answer is yes, but with many significant prior assumptions.... each weakens the credibility of the answer.

I believe that the source of the common misconception regarding the interpretation of a confidence interval is that our mind is sometimes acting in a Bayesian manner. We try to deduce something about the Higgs ("asking the right question"), we derive a confidence integral and translates it to our degree of belief that there is a Higgs given the data, i.e. $Prob(m_t \in [m_1, m_2]|data)$.

A model (A Higgs Boson with a mass m) can only be assigned a degree of belief, but not a probability in a frequentist manner (i.e. as a random variable in a repetitive set of experiments).

The relation between the degree of belief and the true probabilities is given by the Bayesian relation

$$Prob(Higgs|data) = \frac{L(data|Higgs)\pi(Higgs)}{Normalization}$$

where $\pi(Higgs)$ is the prior for a Higgs Boson which many times is taken to be uniform in the Higgs mass (or simply 1) without even noticing!

Last comment here; in this approach instead of talking about confidence intervals we talk about credible intervals, where $p(Higgs|data)$ is the credibility of the Higgs given the data.

7 The Likelihood Principle

Bayesian inference obeys by definition the Likelihood Principle (LP). According to this, the Likelihood function $L(\{\theta\})$ contains the full information from the experimental data. A consequence to the LP is that methods that provide different results for a measurement yet have proportional likelihood functions are inconsistent. A nice discussion about the LP can be found in [3].

8 Who is Afraid of Nuisance Parameters?

The answer to the question appearing in the title is nobody, yet everybody.... Nobody, because Nuisance parameters is just the term used by statisticians for what we physicists refer to as systematics. Everybody, because systematics can kill an experimental observation if not under control. The significance of an observation is given in the limit of large numbers as $S/\sqrt{B}$, however, this number is degraded in the presence of a systematic uncertainty Δ on the background and becomes $S/\sqrt{B(1+\Delta^2 \cdot B)}$, which in the limit of infinite luminosity (and large B) becomes $\frac{S}{B \cdot \Delta}$. So if there is 10% background uncertainty, one will never reach a 5σ significance if $S/B < 0.5$.

Physicists find difficulties in both classifying and estimating the systematic uncertainties and implementing them in the analysis interpretation. There are systematic errors that reduce with increasing statistics and therefore can be handled, and those that do not. In what follows, we will concentrate on the possible treatment of systematics in the interpretation phase of the analysis.

9 Integrating Out the Systematic Errors

When applied to Bayesian credibilities, integrating out the systematics via marginalization with a prior is a natural thing to do. If we denote by s the Higgs signal, by b the background which has some systematic uncertainty, the equation in section 6 becomes

$$p(s,b|data) = \frac{L(s,b|data)\pi(s,b)}{Normalization}$$

The prior is often assumed to factorize $\pi(s,b) = \pi(s)\pi(b)$ with the signal prior taken to be flat. Hence the background systematics is explicitly included in the background prior. We can then integrate the background systematics via $p(s|data) = \int p(s,b|data)db$.

Integrating the nuisance parameters is also used in the so called Cousins-Highland hybrid-frequentist technique [4]. Here the recipe is given by $p(data, data'|s) = \int p(data|s,b)p(b|data')db$ where the $data$ is used for the main measurement and the $data'$ for the auxiliary measurement of the background (e.g. via a side band). It is to be noted that one can fake an auxiliary measurement in order to apply for example a 5% systematics to the background. The Bayesian nature of this method is apparent by the use of the posterior $p(b|data')$.

10 Priors

A prior, e.g. $\pi(\lambda)$ is interpreted as a description of what we believe about a parameter λ preceding the current experiment. One can distinguish two kinds of priors. Informative priors which are based on some information one has on λ and uninformative priors. When the parameter is that of no-interest (nuisance) an auxiliary measurement might supply a legitimate basis for an informative prior. The Higgs signal, on the other hand, is a parameter of interest. Some would say that all priors of the parameters of interest should be uninformative. I would say that using the lower bound of 115 GeV on the Higgs mass as part of a prior, is hard to argue with..... But note also that choosing a prior is a science by itself. A prior flat in the coupling g is not flat in the cross section $\sigma \sim g^2$. That led to the development of reference priors [5]. Reference priors have a minimal effect (relative to the data) on our prospective final inference. In the simple one dimensional case, with one parameter, the reference prior is reduced to the Jeffry's prior which is metric invariant, i.e. $\int L(data|s,\lambda)\pi(\lambda)d\lambda = \int L(data|s,\lambda)\pi(f(\lambda))df$ and can be easily obtained in an analytic way.

11 Doing Justice with the CL_s Method

In section 3 we defined the confidence level CL_b. In a similar manner one can define the signal+background confidence level CL_{s+b}. But what is the meaning of a signal confidence level? Using the terminology of confidence levels CL_s was defined as $CL_s \equiv \frac{CL_{s+b}}{CL_b}$ [6].

The CL_s method is the most discredited method in HEP statistical inference.The reason is that it lacks a frequentist coverage. However, it lacks it in places where the experiment is insensitive to the expected signal! And this is not necessarily a disadvantage from some physicists point of view! Here is what happens:

One uses the Neyman-Pearson likelihood ratio as a test statistics (see section 2) and construct its pdf for background only and signal+background experiments. When the expected signal is very low the two pdf are almost overlapping (see Figure 2). When the number of observed events fluctuates far below the expected background, both hypotheses $s(m_H), s(m_H)+b$ are not favored, yet, given the low p-value of the $s+b$ hypothesis $p_{s+b} = 3\%$ for example, one might exclude the $s(m_H)+b$ hypothesis and the common physicist will interpret the result as if a Higgs with a mass m_h (e.g. 116 GeV in LEP case) is excluded at the 97% Confidence Level. But this is a false statement. To protect against such an inference one defines a new quantity with an unfortunate name $CL_s = \frac{p_{s+b}}{1-p_b}$. In the limit of a light Higgs mass $CL_s \overset{m_H\downarrow}{\longrightarrow} CL_{s+b}$. As a result the false exclusion rate is too low for heavy Higgs Bosons, i.e. the method undercovers where the experiment lacks sensitivity. However this is conservative because it avoids excluding when there is no sensitivity, while simple usage of the pure frequentist CL_{s+b} could result in an exclusion.

12 Neyman Construction

The Neyman construction is a method of parameter estimation that ensures coverage. One scans over all the possible true values of some parameter s and defines an acceptance interval for each s, based on the known pdf, $g(s_m|s)$, of the measured s_m given a possible true s (there is only ONE unknown true s though). The (e.g.) 68% acceptance interval $[s_l, s_h](s)$ is defined via the integration $[s_l, s_h](s) = \{s_m | \int_{s_l}^{s_h} g(s_m|s)ds_m = 68\%\}$ (Figure 3). Even in the simplest case where g is a Gaussian, there is an ambiguity in the choice of the integration limit, which will lead to two-sided intervals, or one-sided integral bounded from below or above. To sort out the integration limits one needs to specify an ordering rule. The construction of the acceptance intervals for all s turns out to be a belt from which one can easily get the corresponding (e.g.) 68% confidence interval $[s_d, s_u](s_o)$ (see section 5), given one measurement s_o via inversion (Figure 3). Due to space limitations there is no way I can describe here the Neyman construction in the necessary detail. Full descriptions can be found in [7].

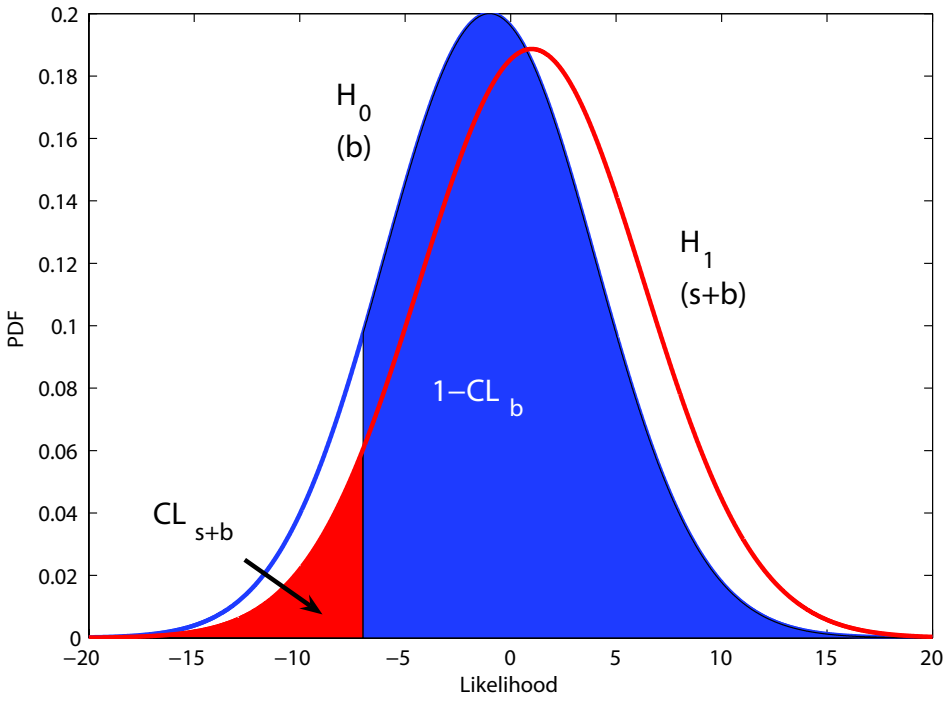

Fig. 2: An illustration showing the reasoning of the CL_s method. In this situation a signal+background hypothesis might be rejected though the experiment has no sensitivity to observe that particular signal.

13 The Feldman-Cousins Method

The full Neyman construction was introduced to HEP by Feldman and Cousins [8]. The test statistic is the likelihood ratio $Q(s) = \frac{L(s+b)}{L(\hat{s}+b)}$ where $\hat{s}$ is the physically allowed mean s that maximizes the Likelihood $L(\hat{s}+b)$. To construct an acceptance 68% interval in the number of observed events, $[n_1, n_2]$, one is using Q as an ordering rule, i.e. $\sum_{n_1}^{n_2} p(n|s,b) \geq 68\%$ where only terms with decreasing order of $Q(n)$ are included in the sum, till the sum exceeds the 68% confidence. When n_o events are observed, one is using this constructed Neyman belt to derive a confidence interval, which, depending on the observation, might be a one-sided or a two-sided interval. This method is therefore called the unified method, because it avoids a flip-flop of the inference (i.e. one decides to flip from a limit to an interval if the result is significant enough...).

The difficulty with this approach is that an experiment with higher expected background which observes no events might set a better upper limit than an experiment with lower or no expected background. This would never occur with the CL_s method.

Another difficulty is that this approach does not incorporate a treatment of nuisance parameters. However, it can either be plugged in "by hand", using the hybrid Cousins and Highland method [9] or a Neyman construction can be performed, as described below.

14 The Profile Likelihood Full Construction Method

Treating the background as a nuisance parameter, one can perform a full Neyman construction with the Feldman-Cousins test statistic used as an order $\ell(s) = \frac{L(s,\hat{\hat{b}})}{L(\hat{s},\hat{b})}$. This is a very cumbersome construction. In this relatively simple example, the construction is done in a 4-dimensional space, the two observables (n, b_m) and the two possible true values (s, b) . For each s the MLE of b is found, $\hat{\hat{b}}(s,n)$. So far only

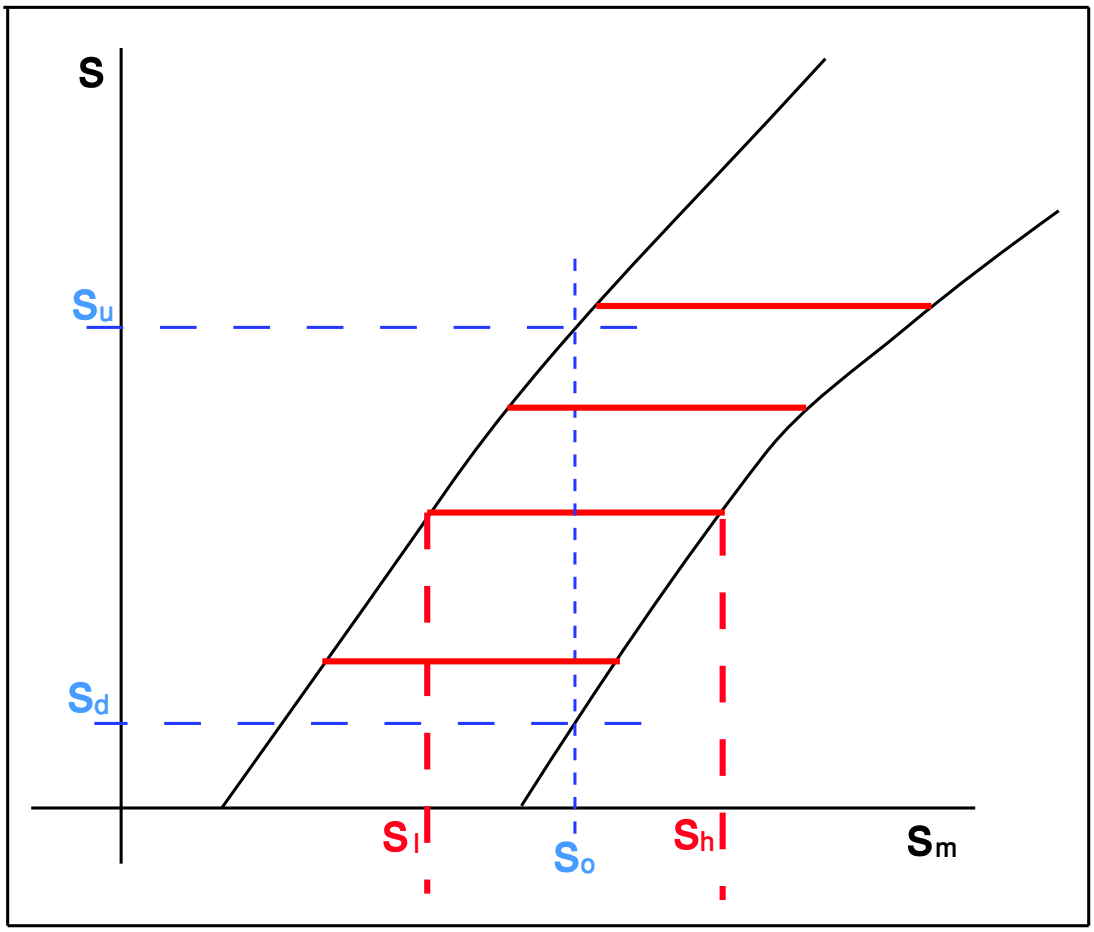

Fig. 3: An illustration showing the Neyman belt. The horizontal lines are the acceptance intervals in the measured parameter space s_m for a given possible true s, $[s_l, s_h](s)$. Given an observation s_o one can construct the confidence interval $[s_d, s_u]$.

low dimensional toy models were fully constructed [10]. To ease the procedure an approximate Neyman construction was suggested [11] by fixing $\hat{\hat{b}}$ to be $\hat{\hat{b}}(s, n_{obs})$. Gary Feldman does not recommend to try the full construction at home for many reasons [12]. One of them is that using a simple Profile Likelihood method works quite well.

15 The Profile Likelihood Method

The simplest way to incorporate systematics into hypothesis inference is the Profile Likelihood. High Energy Physicists are unaware of their familiarity with this method via its implementation in the MINOS process within MINUIT [13].

For simplicity let us define the Profile Likelihood for one channel as $\lambda(s) = \frac{L(s,\hat{\hat{b}}(s))}{L(\hat{s},\hat{b})}$. Here $\hat{\hat{b}}(s)$ is the MLE of b given s and $\hat{s}, \hat{b}$ are the MLE of s and b. When generating experiments, each with data distributed according to $Poisson(n, s+b)$ we find that the pdf of $-2ln\lambda(s)$ is distributed as a $\chi^2(1)$. This is not surprising since in the asymptotic limit the Likelihood function $L(s)$ becomes a Gaussian centered about the ML estimator $\hat{s}$, i.e. $lnL(\hat{s} \pm N\sigma_{\hat{s}}) = lnL_{max} - \frac{N^2}{2}$. The magic of the Profile Likelihood method is that the χ^2 approximation works very well and there is no need for toy Monte Carlo experiments... One can calculate the exclusion or discovery sensitivity or significance in a fraction

of a second.

16 Acknowledgements

This work would have never been possible without the patience and amazing support of Bob Cousins, Kyle Cranmer and Glen Cowan. They helped me to become as close as possible to a Phystatistician in two months...

I would also like to thank Alex Read, Bill Quayle, Luc Demortier and my student Ofer Vitells for always being there to answer my questions and requests.

Last but not least, applause to Louis Lyons for keeping the Phystat meetings going. These meetings are the ultimate source of statistics for the High Energy Physicists. Long may these meetings live! Toda Louis.

I am obliged to the Benoziyo center for High Energy Physics for their support of this work. This work was also supported by the Israeli Science Foundation(ISF), by the Minerva Gesellschaft and by the Federal Ministry of Education, Science, Research and Technology (BMBF) within the framework of the German-Israeli Project Cooperation in Future-Oriented Topics(DIP).

References

[1] A. Stuart and J.K. Ord, Kendallś Advanced Theory of Statistics, Vol. 2, Classical Inference and Relationship, 5th Ed. (Oxford University Press, New York, 1991), tests of hypotheses, Table 20, chapter 20.10.

[2] See talks by Alexey Drozdetskiy and Luc Demortier, these proceedings. See also the CMS TDR appendix A.

[3] G. Zech, Confronting classical and Bayesian confidence limits to examples, contribution to the first Phystat meeting (workshop on Confidence Limits) CERN, Jan 2000. arXiv:hep-ex/0004011. See also refrences therein.

[4] R.D. Cousins and V.L. Highland. Incorporating systematic uncertainties into an upper limit. Nucl. Instrum. Meth., A320:331, 1992.

[5] Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. J. American Statistical Association 84, 200-207; Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. Bayesian Statistics 4 (J. M. Bernardo, J. O. Berger, D. V. Lindley and A. F. M. Smith, eds). Oxford: Oxford University Press, 61-77 (with discussion). See also Luc Demortier, Bayesian Reference Analysis for Particle Physics, Phystat05.

[6] Presentation of search results: the CLs technique, A L Read 2002 J. Phys. G: Nucl. Part. Phys. 28 2693-2704, **doi:10.1088/0954-3899/28/10/313**

[7] I would recommend the Statistics books of Fredrick James and Glen Cowan for a full description of the Neyman construction.

[8] Gary J. Feldman and Robert D. Cousins. A unified approach to the classical statistical analysis of small signals. Phys. Rev., D57:38733889, 1998.

[9] J. Conrad et al.,Including systematic uncertainties in confidence interval construction for Poisson statistics, Phys. Rev. D67 (2003) 012002

[10] K. Cranmer, Frequentist hypothesis testing with background uncertainty. PhyStat2003 physics (2003) 0310108 .

[11] Gary Feldman,Multiple Measurements and Parameters in the Unified Approach,Workshop on Confidence Limits Fermilab March 28, 2000

[12] Gary Feldman, Concluding Remarks: Phystat 2005

[13] F.James and M. Roos, Comput.Phys.Commun. 10, 343 (1975);

International Committee

Jouri Belikov
Bob Cousins
Luc Demortier
Albert De Roeck
David Cox
Kyle Cranmer
Ulrik Egede
Ian Hinchliffe
Fred James
Patrick Janot
Jim Linnemann
Louis Lyons
Harrison Prosper
Nancy Reid

Local Committee

Dorothee Denise (Conference Secretary)
Kate Ross (Conference Secretary)
Albert De Roeck
Louis Lyons
Yves Perrin

List of Participants

Name	Institution
AHMAD, Ashfaq	SUNY Stony Brook, USA
ALBRECHT, Johannes	PI Heidelberg, GERMANY
ALISON, John	University of Pennsylvania, USA
ASK, Stefan	CERN, SWITZERLAND
ATRAMENTOV, Oleksiy	Florida State University, USA
BABB, John Michael	University of California, Riverside, USA
BABU, G. Jogesh	Penn State University, USA
BACHAS, Konstantinos	Aristotle University of Thessaloniki, GREECE
BAINES, Paul	Harvard University, USA
BALA, Suman	Punjab University, INDIA
BANSAL, Vikas	University of Pittsburgh, USA
BELIKOV, Iouri	CERN, SWITZERLAND
BELOTSKIY, Konstantin	MePhl, RUSSIA
BENJAMIN, Doug	Duke University, USA
BERGER, James	Duke University, USA
BHASIN, Anju	School of Physics and AstroPhysics, UK
BHAT, Pushpalatha	Fermilab, USA
BHATTACHARYA, satyaki	University of Delhi, INDIA
BIANCO, Michele	Univ. + INFN, ITALY
BITYUKOV, Sergey	IHEP, RUSSIA
BOCCI, Andrea	Duke University, USA
BOTJE, Michiel	NIKHEF, NETHERLANDS
BRANDT, Oleg	UK
BRUCKMAN DE RENSTROM, Pawel	University of Oxford, UK
BRUN, Rene	CERN, SWITZERLAND
BRUNELIERE, Renaud	Freiburg, GERMANY
CAKIR, Orhan	University of Ankara, TURKEY
CAMPANELLI, Mario	Michigan State University, USA
CARSON, Laurence	University of Glasgow, UK
CASADEI, Diego	New York University, USA
CAVANAUGH, Rick	University of Florida, USA
CHAO, Yuan	National Taiwan University, TAIWAN
CHAUHAN, Sushil Singh	CDRST, INDIA
CLARE, Robert	University of California, Riverside, USA
CLEMENTS, Daniel Robert	University of Glasgow, UK
COADOU, Yann	CERN, SWITZERLAND
CONWAY, John	University of California, Davis, USA
COPIC, Katherine	Columbia University, USA
COUSINS, Robert	University of California, Los Angeles, USA
COWAN, Glen	Royal Holloway, UK
COX, David	University of Oxford, UK
COX, Tim	University of California, Davis, USA
CRANMER, Kyle	BNL, USA
D'AURIA, Saverio	University of Glasgow, UK
D'HONDT, Jorgen	Vrije Universiteit Brussel, BELGIUM
DE LA CRUZ BURELO, Eduard	University of Michigan, USA

DE ROECK, Albert	CERN, SWITZERLAND
DEISSENROTH, Marc	Universitaet Heidelberg, GERMANY
DELMASTRO, Marco	CERN, SWITZERLAND
DELMEIRE, Evelyne	Universiteit Antwerpen, BELGIUM
DEMIN, Pavel	Universite Catholique de Louvain, BELGIUM
DEMORTIER, Luc	Rockefeller University, USA
DENISE, Dorothee	CERN, SWITZERLAND
DONINI, Julien	Padova University, INFN, ITALY
DROZDETSKIY, Alexey	University of Florida, USA
DUCKECK, Guenter	LMU Munich, GERMANY
DUNFORD, Monica	University of Chicago, USA
EIFERT, Till	University of Geneva, SWITZERLAND
ESTEVES RAMALHETE, Pedro Miguel	LIP, FRANCE
EZHELA, Vladimir	IHEP, RUSSIA
FABIEN, Tarrade	BNL, USA
FAYARD, Louis	LAL Orsay, FRANCE
FERNANDEZ TELLEZ, Arturo	High Energy Physics Group, MEXICO
FERREIRA PARRACHO, Pedro Guilherme	IST, PORTUGAL
FILTHAUT, Frank	Dept. of Experimental HEP, NETHERLANDS
FISHER, Wade	Fermilab, USA
FONSECA MARTIN, Teresa	CERN, SWITZERLAND
GARCIA-ABIA, Pablo	CIEMAT, SPAIN
GELE, Denis	IPHC, FRANCE
GENTILE, Simonetta	Universita di Roma I , ITALY
GIANOTTI, Fabiola	CERN, SWITZERLAND
GOLLUB, Nils	PH-ATA, CERN, SWITZERLAND
GORINI, Edoardo	Salento University, ITALY
GOVONI, Pietro	Universita’ ed INFN Milano-Bicocca, ITALY
GROSS, Eilam	Weizmann Institute, ISRAEL
GRUNEWALD, Martin	University College, IRELAND
GUTIERREZ, Phillip	University of Oklahoma, USA
HAMACHER, Klaus	Fachbereich C / Physik, GERMANY
HARVEY, Alex	Hampton University, USA
HEINRICH, Joel	University of Pennsylvania, USA
HINCHLIFFE, Ian	LBNL, USA
HINTZ, Wieland	Labor fur Hochenergiephysik, SWITZERLAND
HOECKER, Andreas	CERN, SWITZERLAND
HOOFT VAN HUYSDUYNEN, Loek	NIKHEF, NETHERLANDS
HORVATH, Dezso	Res. Inst. Particle Nucl. Phys., HUNGARY
IANNI, Aldo	LNGS INFN, ITALY
JACHOLKOWSKI, Adam	Universita di Catania, FRANCE
JAMES, Fred	CERN, SWITZERLAND
JANA, Dilip Kumar	University of Oklahoma, SWITZERLAND
JANOT, Patrick	CERN, SWITZERLAND
JEN-LA PLANTE, Imai	University of Chicago, USA
JOHNSON, Kurtis	Florida State University, USA
KANAYA, Naoko	University of Tokyo / ICEPP, JAPAN
KAO, Shih-Chuan	University of California, Riverside, USA
KARAGOZ UNEL, Muge	University of Oxford, UK
KARASU UYSAL, Ayben	Yildiz Technical University, TURKEY

KATSAS, Panagiotis	University of Athens, GREECE
KAYIS TOPAKSU, Aysel	Physics Department, TURKEY
KOESTNER, Stefan	CERN, SWITZERLAND
KONO, Takanori	CERN, SWITZERLAND
LAFAYE, Remi	LAPP, FRANCE
LAFORGE, Bertrand	LPNHE, FRANCE
LAGOURI, Theodota	Universidad Autonoma de Madrid, SPAIN
LAKTINEH, imad	IPN LYON, FRANCE
LAMPEN, Tapio	HIP, FINLAND
LARI, Tommaso	Milano University and INFN, ITALY
LE DIBERDER, Franois	CNRS/IN2P3, FRANCE
LETHUILLIER, Morgan	IPN Lyon / IN2P3 / CNRS, FRANCE
LEVEQUE, Jessica	CPPM, FRANCE
LIEBIG, Wolfgang	NIKHEF, NETHERLANDS
LINNEMANN, James	Michigan State University, USA
LIPNIACKA, Anna	University of Bergen, NORWAY
LOPEZ, Angel Mario	Univ. Puerto Rico, USA
LYONS, Louis	Oxford, UK
M, Saleem	University of Oklahoma, USA; CERN, SWITZERLAND
MAJUMDER, Gobinda	TIFR, INDIA
MARTINS, Pedro	LIP, PORTUGAL
MCGLONE, Helen	CERN, SWITZERLAND
MILENOVIC, Predrag	ETH zurich, SWITZERLAND
MONETA, Lorenzo	CERN, SWITZERLAND
MORSCH, Andreas	CERN, SWITZERLAND
MOSS, Joshua James	Ohio State University, UK
NARSKY, Ilya	Caltech, USA
NATION, Nigel	Boston University, SWITZERLAND
NEAL, Radford	University of Toronto, CANADA
NEPOMUCENO, Andre	Federal University of Rio de Janeiro, BRAZIL
NEWMAN, Harvey	Caltech, USA
NIKOLOPOULOS, Konstantinos	University of Athens, GREECE
NILSEN, Bjorn	Ohio State University, USA
NILSSON, Paul	University of Texas, Arlington, USA
OFER, Vitells	Weizmann Institute, ISRAEL
OHAD, Silbert	Weizmann Institute, ISRAEL
OLIVITO, Dominick	University of Pennsylvania, USA
OULD-SAADA, Farid	Oslo University, NORWAY
PAGAN GRISO, Simone	INFN and University of Padova, ITALY
PAKHOTIN, Yuriy	University of Florida, USA
PANARETOS, Victor	Ecole Polytechnique Fdrale de Lausanne, SWITZERLAND
PERRINO, Davide	INFN, ITALY
PORTELL BUESO, Xavier	Freiburg, GERMANY
PRAVAHAN, Rishiraj	uta, SWITZERLAND
PRICE, Darren	Lancaster University, UK
PROSPER, Harrison	Florida State University, USA
PUROHIT, Milind	University of South Carolina, USA
QUAST, Gunter	Institut fuer Experimentelle Kernphysik, GERMANY
QUAYLE, William	University of Wisconsin, USA
RAZZAK, Meera Lebbai	University Of Oklahoma, SWITZERLAND

REDIN, Sergei	Budker Institute, RUSSIA
REECE, Ryan David	University of Pennsylvania, USA
REID, Nancy	University of Toronto, CANADA
RESENDE, Bernardo	CPPM, FRANCE
RIU, Imma	UAB/IFAE, SPAIN
ROLKE, Wolfgang	UPR-RUM, PUERTO RICO
ROSZKOWSKI, Leszek	University of Sheffield, UK
ROY, Dipanjan	University of Texas, Arlington, USA
ROZEN, Yoram	Technion, ISRAEL
SALVATORE, Tupputi	Universita degli Studi di Bari, ITALY
SANNINO, Mario	INFN Genoa, ITALY
SANTONI, Claudio	Universite Blaise Pascal de Clermont-Ferrand II, FRANCE
SASCHA, Caron	University Freiburg, GERMANY
SCHOTT, Gregory Alfred	Universitaet Karlsruhe, GERMANY
SEIXAS, Jose	Univ. Federal do Rio de Janeiro, BRAZIL
SIOLI, Maximiliano	Universita degli Studi di Bologna, ITALY
SARDY, Sylvain	Geneva University, SWITZERLAND
SIRUNYAN, Albert	Yerevan Physics Institute, ARMENIA
SLABOSPITSKY, Sergey	IFVE, RUSSIA
SONNENSCHEIN, Lars	CERN, SWITZERLAND
SOTIRIS, Vlachos	National Technical Univ. of Athens, GREECE
T'JAMPENS, Stephane	LAPP - CNRS/IN2P3, FRANCE
TEGENFELDT, Fredrik	Iowa State University, USA
THERESE BERGE, Sjursen	University of Bergen, NORWAY
THORNE, Robert	UCL, UK
TIMCIUC, Vladlen	Caltech, USA
TONOYAN, Arshak	University of Bergen, NORWAY
TRACZYK, Piotr	Soltan Institute for Nuclear Studies, POLAND
TRIPIANA, Martin	UNLP, ARGENTINA
TRENTADUE, Raffaello	Bari University, ITALY
TUCKER, Jordan	University of California, Los Angeles, USA
UNAL, Guillaume	CERN, SWITZERLAND
USAI, Giulio	University of Chicago, USA
VASQUEZ SIERRA, Ricardo	University of California, Davis, USA
VASSILAKOPOULOS, Vassilis	Hampton University, USA
VERKERKE, Wouter	NIKHEF, NETHERLANDS
VOLOBOUEV, Igor	Texas Tech University, USA
VOSS, Helge	MPI-K Heidelberg, SWITZERLAND
WENG, Joanna	IPP, SWITZERLAND
WICKRAMAGE, Nadeesha Manohari	Univ. Ruhuna, SRI LANKA
WOLF, Gustavo	Louis Dreyfus Commodities, SWITZERLAND
WRIGHT, Catherine	University of Glasgow, UK
XIE, Yuehong	University of Edinburgh, UK