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].