CONFERENCE SUMMARY TALK

Robert Cousins
Dept. of Physics and Astronomy, University of California, Los Angeles, CA 90095, USA

Abstract
This conference contained a wealth of illuminating talks, including a number of informative overviews of analysis methods, practical "reports from the trenches", and proposals for new ways to report results. The discussions were enriched by the participation of a Bayesian statistician.

1 INTRODUCTION

In my conference summary talk, my audience was of course composed of people who had heard every talk in the previous few days, and it was (almost) feasible to refer to most of the talks while emphasizing certain themes. In this written version, I attempt again to give an overview, but I am forced to reduce the number of topics in order to make it more comprehensible to those who might read this without having digested the rest of the conference. The conference attracted a number of talks of high quality which I discuss below in several broad categories:

- The "Other" PDF's: parton distribution functions.
- Reports from the Trenches: lessons from specific data analyses by physicists.
- Group Dynamics: how statistical analysis is performed in large collaborations and meta-collaborations.
- Tutorials and Overviews: surveys of tools use in data analysis.
- Studies of Intervals and Limits: confidence intervals, credible intervals, and alternatives.
- "Why Be A Bayesian?" A professional statistician's perspective.

In choosing what to speak about, I tried imagine which talks will have an impact beyond this conference. I think that talks on the following sorts of topics are particularly helpful.

- Well-founded methods from elsewhere in academe which are introduced in HEP and found to be practical and useful.
- HEP-specific extensions of standard methods when the extensions are understood and found to be practical and useful.
- Lucid explanations of subtle issues that arise in the context of a particular experiment.

With a higher threshold, we can add to the list completely new inventions by professional physicists/amateur statisticians. The threshold is higher because one must understand the foundations of any method and how broadly applicable it is.

Even sorting with these criteria, there is far too much to mention in this summary, so I am guided additionally by the desire to highlight methods that might be less well known.

2 THE "OTHER" PDF'S

At a statistical conference, a pdf is a probability density function, but for a subgroup here (which had its own sessions, which I won't be able to summarize), a PDF is a parton distribution function. In addition to their intrinsic interest, these PDF's matter in frontier experiments: as an example, we need only recall the difficulty in interpreting high-p_T spectra at experiments such as CDF. As was apparent in the introductory talk by **Robert Thorne ("Uncertainties in parton related quantities")**, those who calculate PDF's face an enormously difficult task: fitting vast data sets from diverse experiments, in order to extract several functions *and the errors on these functions.* This task is made even more difficult by the fact that the functional forms have systematic uncertainties due to finite-order theoretical calculations.

The walk-through by **Amanda Cooper-Sarkar** of **"Zeus NLO QCD fits to extract PDFs, $\alpha_s(M_Z^2)$ and their uncertainties"** also revealed points of controversy, and stimulated a discussion on what value of $\Delta\chi^2$ is appropriate, a discussion which I believe will extend well beyond this conference. (There were related talks in the parallel session on PDF's and in the main session by M.J. Wang.)

In spite of the difficulties, the progress has been impressive. For example, in predicting Higgs production cross sections, what matters is not the error at every point on every function, but rather the final answer which integrates over products of these functions. The errors on the Higgs cross section from various approaches are less than 5%.

3 REPORTS FROM THE TRENCHES

Here were heard a number of talks reporting on real-world applications, with all the difficulties that imperfect data and uncertain modeling bring. I mention a few:

3.1 Gary Hill and Tyce De Young: Application of Bayesian statistics to muon track reconstruction in Amanda

This was a fascinating talk in which Bayes's Theorem was applied to calculate the probability that a cosmic-ray muon was neutrino-induced or the direct result of meson decay. The technique is powerful and applies to other experiments which have a signal-to-background ratio which is dependent on position or some other variable. My only comment is a semantic one: while the authors called this technique "Bayesian", it would appear to be perfectly valid with the frequentist definition of probability.

Bayes's 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. Using the frequentist definition of P, it is difficult or impossible to define the "probability that supersymmetric particles have masses below a fixed value, say 1 TeV" but when there is a relevant ensemble it is quite possible to define "the probability that a randomly selected event with given characteristics will be a neutrino-induced muon." What is required is that the input P's to Bayes's theorem are themselves frequentist P's, which appears to be the case here: the fractions of muons, as a function of angle, that are neutrino-induced.

Thus, this application of Bayes's Theorem would appear to be as free of controversy as the common introductory example of Bayes's Theorem using a medical test. (See, e.g., Ref. [1] and talks by F. James and M. Goldstein at this conference). One needs to ensure, of course, that the input priors are determined independently in order to avoid a circularity to the measurement, but that is the same issue as in any other experiment in which calibration data of various sorts is used.

3.2 Volker Blobel and Claus Kleinwort: A New method for the high-precision alignment of track detectors

This beautiful talk is required reading for anyone faced with the task of fitting to large numbers of parameters.

3.3 Nigel Smith and Dan Tovey: Statistical Issues in Dark Matter Searches

This talk described important work by astro-particle physicists seeking an astounding observation: direct detection of the non-luminous matter which is known (through its gravitational interaction) to account for most of the mass of galaxies such as ours. There are a number of difficult statistics issues which are not yet completely solved, and which I am sorry we did not have more time to explore at this conference.

3.4 Rudy Bock and Wolfgang Wittek: Gamma/Hadron separation in atmospheric Cherenkov telescopes

Here is more required reading: a comparative study of signal and background separation in the context of an astro-particle physics experiment.

3.5 Sergei Redin: Advanced Statistical Techniques in the muon g–2 experiment at BNL

The beauty of this talk, aside from the specific issue addressed, was the reminder of how much can be learned with pencil and paper about how errors from various sources contribute to the overall error. A modern student can easily get the impression that multiple GEANT jobs are the way to approach any problem, but it may take a lot of simulations and log-log plots to gain the insight which analytic calculations can provide. In the case at hand, the experimenters at BNL perform a 5-parameter fit to extract the one parameter if interest, the precession frequency of the muons. The parts-per-million accuracy desired is far from the usual experience in HEP.

3.6 Fabrizio Parodi et al: How to use the Δm_S information in CKM fits

This was another interesting talk in which more work might be useful; I would have to study it more in order to understand the chosen method better. The speaker noted that an ad-hoc "modified" χ^2 proposed by someone else did not perform well. This is not too surprising: there is a large burden of proof placed on those who would invent chi-squares which are not based on those studied by statisticians.

3.7 Kay Kinoshita: Evaluating quality of fit in Unbinned Maximum Likelihood Fitting

This talk helped explain why a general-purpose goodness-of-fit (g.o.f.) test is not possible using only the value of the unbinned likelihood function at its maximum. (A recent internal CDF Note by Joel Heinrich also examines this issue.) I can add the reminder that the chi-square g.o.f. tests that we use (Gaussian and binned Poisson) can be derived starting from the likelihood ratio theorem [8]. It's the likelihood *ratio* which produces a g.o.f. test statistic which asymptotically obeys the chi-square distribution. With just one L, one has only the numerator, which is not dimensionless and which is metric-dependent. (As Fred James has emphasized in the past, for a given best-fit parameter one can therefore find a metric in which L is unity for all data sets which yield that best-fit parameter.) With some class of alternative models in mind, one can effectively obtain a relevant ratio for g.o.f. by comparing unbinned likelihoods, but this is more restricted than the usual classical g.o.f. tests in which no alternative is specified. The speaker points to some future work which will help clarify matters.

4 GROUP DYNAMICS

4.1 Bruce Yabsley: Statistical practice at the Belle experiment, and some questions

Not only is Belle a large collaboration spread around the world, but it is in head-to-head competition with BaBar on a broad menu of physics measurements requiring diverse statistical techniques. One has to accept that there may not be time to perform the ultimate analysis in the first instance. But two points can be made which both point toward education in statistical methods. First, while the initial results may be published hastily, the huge investment in these accelerators and experiments demands that an appropriate effort eventually be made to extract the most precise results from the data for physics of interest. Second, when compromises must indeed be made in order to have a timely announcement, the choice should be one informed by an understanding of how "dirty" the "quick" method is.

An interesting problem that this speaker highlighted relates to the $B^0\overline{B^0} \to \pi^+\pi^-$ analysis. The parameters extracted from the data are coefficients S and C of sine and cosine terms in the time dependence, respectively. This is an interesting variation on the physical-boundary situation because of the constraint $S^2 + C^2 \leq 1$. (Simple estimates of S and C may violate this inequality.) The collaboration

has constructed confidence intervals using the Unified Approach advocated by Feldman and Cousins [3], but issues remain. Notably, there is physics interest in the compatibility (or not) of the data with the special point C=0 and S=0 (no asymmetry) and the line segment C=0 (no direct CP violation). In the context of the Unified Approach, a p-value can be constructed: it is that value of the Significance Level such that the confidence interval/region would just exclude the point(s) of interest. If a Bayesian analysis is performed, I would argue that we are in the situation where a subjective prior with delta-functions at the interesting points are appropriate; one can do a sensitivity analysis with respect to the amount of probability put into the delta functions and elsewhere. (This sort of prior does not exist in the usual so-called "objective" Bayesian priors.) I return to this point below.

4.2 Chris Parkes: Practicalities of combining analyses: W physics results at LEP

The usual collaboration problem of agreeing on a common analysis method was multiplied by four when the LEP experiments formed a meta-collaboration in order to combine results on the most important measurements. The problem was exacerbated because each experiment had already published official results. Chris Parkes took us through the complex process by which they made compromises, understood correlated uncertainties, and achieved the goal of combined results.

LEP experiments contained a sizable fraction of world HEP community. As evident from this and other talks at this conference, they reached very mature state of analysis, having thought through many issues. We have much to learn from them, both theoretical and practical.

5 TUTORIALS AND OVERVIEWS

There were a number of excellent tutorials and reviews at the conference, including the following. (The talk by R. Bock above is also in this category.)

- **Fred James: Overview of Bayesian and Frequentist Principles**
- **Sherry Towers: 1) Overview of non-parametric Probability Density Estimation methods, and 2) Benefits of minimizing the number of discriminators used in a multivariate analysis.**
- **Berkan Aslan & Günter Zech: Comparison of different goodness of fit tests**
- **Roger Barlow: Systematic errors: Facts and fictions**
- **Niels Kjaer: Monte Carlo unifying Frequentist and Bayesian inference**
- **Paul Harrison: Blind Analyses**
- **Harrison Prosper: Multidimensional methods in data analysis: a unified perspective**
- **Tony Vaiciulis: Support Vector Machines in Analysis of Top Quark Production**
- **Pekka Sinervo: The significance of HEP observations**
- **Glen Cowan: A survey of unfolding methods for Particle Physics (See also Volker Blobel: An unfolding method for high energy physics experiments)**

I refer the reader to the writeups in these proceedings, which I believe will be a most valuable reference in the future. I mention here only two general points. Fred James emphasized the importance of knowing the statistical question one is asking, since confusing the question can quickly lead to a confusing answer. For example, avoid confusing confidence intervals (statements about parameters) with goodness of fit (statements about the model itself).

Harrison Prosper emphasized that there is a unifying theme to many of the above efforts, namely to classify events (or the equivalent) as one type or another. For definiteness, he considers signal S vs. background B. The Neyman-Pearson lemma tells us that the most powerful classical test of a simple hypothesis against a simple alternative is based on the likelihood ratio $L(S)/L(B)$. Where there is prior information favoring one hypothesis over the other, this generalizes to the Bayes discriminator, $D(x) = P(S|x)/P(B|x) = (L(S)/L(B))(P(S)/P(B))$, where x is the data. Frequently the likelihood

function L is not known a priori, and even its functional form may not be known. However it can be estimated by non-parametric methods, in particular by taking some kind of average of Monte Carlo events in a multi-dimensional space.

Prosper listed many of the methods discussed – Fisher Linear Discriminant, Principal Components Analysis, Independent Component Analysis, Self-Organizing Map, Grid Search, Probability Density Estimation (Kernal methods), Neural Networks, Support Vector Machines – and viewed them as attempts to solve the single classification problem whose solution is the Bayes discriminant. (As a direct application of the Bayes discriminant, see the paper by Gary Hill and Tyce De Young on AMANDA above. Papers using neural nets included those by F. Hakl et al. and by M. Wolter) He emphasized that multivariate analysis is hard and that it appears that there is no single optimal approximation – hence the proliferation of methods. He noted that it is important to use all the information contained in the full $D(x)$ (which might be lost, e.g., by marginalization).

6 STUDIES OF INTERVALS AND LIMITS

There were a number of talks about confidence intervals and confidence limits, which were the main topic of two previous workshops at CERN [4] and Fermilab [5]. Here I mention only a few of the interesting talks, primarily those with more recent results.

6.1 Alex Read: CL_s – Reporting Search Results

This was a beautiful talk which lucidly explained comparisons between the Unified Approach and CL_s. The CL_s method, which was developed for LEP experiments, is applied to neutrino oscillations. Alex Read now advocates CL_s only for limits and in case of signal, he now would use the Unified Approach (without the "unity" between limits and intervals that it provides).

6.2 Byron Roe & Michael Woodroofe: BooNE Neutrino Oscillations

For setting limits, Roe and Woodroofe advocate the approach of Bayesian calculations with approximate frequentist coverage, described at the FNAL workshop [5, 6]. In the case of a signal, they also would use the Unified Approach.

6.3 Dean Karlen: Credibility of Confidence Intervals

The speaker advocated that experiments report, along with a confidence interval, its credibility, namely the degree-of-belief that the true parameter is contained in the stated interval. This of course requires a prior pdf, which he suggests be taken as uniform in the region of interest. While there was generally interested-to-favorable reaction to this suggestion, I raised my usual objection [4] that uniform priors typically fail to capture degree of belief in any quantified manner, in addition to being ill-defined (since one must choose the metric in which the prior is uniform). We evaluate Bayesian intervals with serious frequentist methods. Why not evaluate confidence intervals with serious Bayesian methods? That includes subjective Bayesian methods, which truly deserve the description "degree of belief". In any case, this talk opened up a fruitful discussion which I am sure will continue.

6.4 Wolfgang Rolke & Angel Lopez: Bootstrap-corrected limits for rare signals

Even in a blind analysis in which the signal region is hidden while analysis criteria (cuts) are chosen, there can be a bias introduced by tuning cuts specifically to reduce events near the signal region, if these same events are used to estimate the background level. Some experiments therefore blind themselves to the events used to estimate the background level, which requires having even more events to tune the cuts. This speaker instead applied a generalization of the bootstrap method to show how to correct for the bias, so that all available events can be used to estimate the background level.

Resampling (such as the bootstrap) is common in other fields, but not in experimental HEP, as far as I am aware [2]. Such methods were also mentioned in Kjaer's talk above. It would be interesting to hear about more experience with them.

6.5 Rajendran Raja: Confidence Limits and their Errors

This talk dealt with the fact that the mean of a measurement gets quoted along with an error, but one does not quote an "error" on an upper limit. I believe there was some semantic confusion at the time, but the point many of us would agree with is that the sampling distribution of upper limits is of interest [7]. (This distribution depends, of course, on what is assumed about the signal strength, for example no signal events [3].) An upper limit is the end-point of a confidence interval, so it does not have an error or uncertainty in the usual sense, but the speaker raises a useful issue that I believe will be pursued in the future.

7 MICHAEL GOLDSTEIN: WHY BE A BAYESIAN?

It was a real pleasure to have Michael Goldstein, from the Univ. of Durham's Dept. of Mathematical Sciences, participate actively in the entire conference. He is a strong advocate of Bayesian methods using priors which really do represent beliefs – those priors which we often refer to as subjective priors. From this point of view there is no support for the existence of so-called "objective" or "non-informative" priors; such priors can be useful for illustrative purposes, but that is all. He insists on the Likelihood Principle, and hence finds that (frequentist) confidence intervals (which do not obey it) are fundamentally flawed. He notes difficulties with constructing the likelihood function – it may require subjective input – and additional pitfalls in high dimensions.

He also said that Bayesian methods are hard to do right, but they are the only way to attack certain hard problems. In his research, he and co-workers have been developing a Bayes Linear Methodology which addresses expectations rather than whole pdf's, in order to make some hard problems more tractable.

One point which I think Bayesians advocates in HEP should take seriously is his statement that "Sensitivity Analysis is at the heart of scientific Bayesianism". A sensitivity analysis looks at the posterior beliefs as a function of the prior beliefs. How skeptical would the community as a whole have to be in order not to be convinced that a discovery was made? What prior gives P(hypothesis) > 0.5 ? What prior gives P(hypothesis) > 0.99, etc?

Another point worth noting is his description of Bayesian methods as "hard", because he advocates doing the hard work to get a prior which really represents belief, in contrast to using a mathematical rule to obtain a prior which he finds "arbitrary".

Michael Goldstein represents only one school of Bayesian statistics. There are advocates of objective or non-informative priors, but I don't know of a school advocating a uniform prior for a Poisson mean, for example. I think that answers obtained with uniform priors are without much content unless they are evaluated by frequentist standards, as is indeed often the case in HEP. (In this case the importance of the Bayesian origin is not so much defining P as degree of belief, but rather that the Likelihood Principle is built in). Uniform priors can also of course provide examples in a sensitivity study.

On the other hand, subjective priors expand the sorts of sensitivity studies one can perform, since one can put part of the probability in a delta function at any point of particular interest, for example the point (0,0) in the CP violation study referred to above (Bruce Yabsley's talk). The result can be displayed as a function of the fraction of prior belief located at this special point (and of course, as a function of other subjective pieces of the prior).

8 CONCLUSION

Nearly all of the participants in this conference are physicists who have studied statistical techniques in order to have better tools available for their main passion, elementary particle physics. Thus it has been particularly impressive to listen to such cogent talks on a wide variety of statistical topics. I'm sure that we still have a lot to learn from the career statisticians, but the overall level of discussion has risen tremendously in recent years, while maintaining a congenial atmosphere. As our experiments get more expensive and take longer to perform, extracting the most information from the data by using advanced techniques translates into savings of running time and resources. We can look forward to the progress to be reported at the next such conference in a year or two.

ACKNOWLEDGEMENTS

I'm sure that everyone would join me in thanking the organizers of the conference. The local organizers performed a truly outstanding job: James Stirling, Mike Whalley, and Linda Wilkinson. This is the third workshop or conference in which Louis Lyons has been a driving force, having previously convened the CERN Workshop on Confidence Limits [4] (with Fred James) and the subsequent workshop at Fermilab [5]. (At the present conference no less than five authors of "statistics for physicists" books have been present: R. Barlow, G. Cowan, L. Lyons, F. James, and B. Roe.) We are grateful for Louis's service, and in fact he is already planning the next one. Finally I thank all the participants for their assistance in helping me to prepare this summary talk.

References

[1] G. Cowan, *Statistical Data Analysis* (Oxford University Press, Oxford, 1998)

[2] At the conference, I mentioned one paper I remembered, co-authored by Efron himself: K.G. Hayes, M. L. Perl, B. Efron, "Application of the bootstrap statistical method to the tau decay mode problem", Phys. Rev. **D39** 274 (1989). I thank Jim Linnemann for pointing me to another early application, A.L. Angelis et al, Nuclear Physics B327 (1989), and the talk by J. Linnemann, "Use of the Bootstrap in a High Energy Physics Experiment", Proceedings of the 1990 Symposium on the Interface between Computing Science and Statistics, Springer 1992, ed C. Page, R. Lepage, p 439-441.

[3] G.J. Feldman and R.D. Cousins, Phys. Rev. **D57** 3873 (1998). As noted in the Note Added in Proof, the Likelihood-Ratio ordering advocated for confidence intervals is a standard statistical method. In the present summary I refer to this approach as the "Unified Approach" for the reasons given in our article. It is not unique in this unified aspect, however, so a more precise description is the Likelihood-Ratio ordering used to construct the intervals.

[4] Proceedings of the Workshop on Confidence Limits (17-18 January 2000, Edited by F. James, L. Lyons, and Y. Perrin), CERN Yellow Report 2000-005 (2000). Available at http://user.web.cern.ch/user/Index/library.html

[5] Fermilab Workshop on Confidence Limits 27-28 March, 2000, http://conferences.fnal.gov/cl2k/

[6] B.P. Roe and M.B. Woodroofe, Phys. Rev. **D63** 013009 (2001).

[7] Sampling distributions are easy to study; for example Ref. [3] uses the mean of upper limits (which they call "sensitivity"), and the variance of upper limits is emphasized by C. Giunti and M. Laveder, hep-ex/0002020. Both are metric-dependent; in some cases the median and percentiles provide a useful metric-free alternative.

[8] For a review of the Poisson case, see S. Baker and R.D. Cousins, Nucl. Inst. Meth **A221** 437, 1984. As I learned during a discussion with Saul Perlmutter, the Gaussian case is analogous. In both these cases, the alternative hypothesis is the function that goes exactly through the data points.