

Statistical issues relevant to significance of discovery claims

Richard Lockhart (Simon Fraser University),
Louis Lyons (University of Oxford), James Linnemann (Michigan State University)

July 11 – July 16, 2010

1 Motivation for the meeting

In 2006 a workshop was held at BIRS titled “Statistical inference Problems in High Energy Physics and Astronomy”. The outcome of the 2006 Workshop was so encouraging that we proposed another BIRS workshop for 2010. New facilities in Particle Physics and Astrophysics (e.g. the Large Hadron Collider and the GLAST telescope for gamma rays) are beginning to produce a large amount of data. There is a strong hope that these will result in exciting new discoveries. There are interesting statistical issues relating to discovery claims, and it is important to be able to give reliable, widely accepted statistical assessments of the evidence that the result is not due just to a statistical fluctuation.

A potentially disturbing example arises from an experimental Particle Physics collaboration who analyzed their data in 2003, and found that, at greater than a 5 sigma level, their data were inconsistent with the null hypothesis, and instead gave evidence for a new type of particle, the penta-quark. However a subsequent calculation of the Bayes factor comparing the null hypothesis with the alternative of a new particle was said to favour mildly the null hypothesis. This apparent sensitivity of an important conclusion to the statistical technique employed is worrying, and needs to be understood. The conflicting papers from the same authors analyzing the same data can be seen at: <http://arxiv.org/abs/hep-ex/0307018> and <http://arxiv.org/abs/0709.3154>.

This workshop therefore brought together particle physicists, astronomers and statisticians to discuss:

1. Why Particle Physicists like 5 sigma as a discovery criterion; for Statisticians, requiring a 5 standard error deviation from the null, which corresponds to a significance level on the order of 1 in a million, is extraordinarily stringent.
2. Allowing for multiple tests; research groups carry out many tests on the same data.
3. Goodness of fit tests for comparing sparse multi-dimensional data with theory.
4. Comparison of different techniques for comparing 2 hypotheses, for example:
 - (a) p -values (including methods for combining p -values for different tests);
 - (b) The so-called CL_s (ratio of p -values for null hypothesis and alternative), an approach to setting upper confidence limits which is little known in the statistical community;
 - (c) Likelihood ratio tests, even when null and alternative hypotheses are composite;
 - (d) Difference in chi-squared of 2 separate fits to the same data;
 - (e) Model selection techniques such as AIC or BIC;

- (f) Bayesian techniques such as posterior odds or Bayes factors (including the issue of choice of prior).
- 5. Adjusting for nuisance parameters in p-value and likelihood calculations.
- 6. Definitions of sensitivity of searches for new phenomena.

2 Presentation Highlights

The workshop format was very informal with the schedule being rejigged each evening in light of what happened that day. We wanted to focus on getting conversations and joint research projects going and we think we succeeded. In this section of the report we touch on main themes. Details of some presentations are provided in the next section.

The workshop opened with talks from a particle physicist (Louis Lyons) and an astronomer (Tom Laredo) setting the stage for the discussions and a response talk (Richard Lockhart) which tried to begin the process of translation from one technical language to another. We followed up by spending much of the first afternoon letting each participant say why he was there and what he hoped to get out of the meeting.

The second day began with talks from Jon Pumplin and Robert Thorne on parton distribution functions highlighting the following general problem. (Parton distribution functions describe the random partitioning of momentum among the quark and gluon constituents of a hadron such as a proton. When two hadrons collide it is really one of these constituents from each hadron which interact and these parton distribution functions then make it possible to describe the distribution of the momenta of the colliding constituents.) Several groups fit parton distribution function models to data arising from a variety of experiments. It is found that the fitted standard errors arising from standard chi-square approximations to log-likelihood ratio drops are unrealistic and that several experiments differ from the fitted values by more than is reasonable. It seems that differences of 50 need to be considered rather than differences of 1 or 2. This set of talks prompted much discussion and the talks had to be continued later in the meeting. See the commentary from Jon Pumplin, Robert Thorne and Steffen Lauritzen below.

Jim Linnemann talked about the on-off problem and Kyle Cranmer followed up with extensions therefrom. In particular he introduced the idea of “Asimov data sets”, new to statisticians and most physicist; for binned data an Asimov data set has bin counts equal to their expected value (even if that is not an integer). This sparked considerable discussion; see Kyle Cranmer’s remarks below. Also see Glen Cowan’s Wednesday talk on profile likelihood.

Elliott Bloom discussed the failure of Wilk’s theorem in an astrophysical example testing for source extension. Appropriate large sample approximations to the behaviour of likelihood ratio tests played an important role in the conference. See Eilam Gross and Ofer Vitells below. A talk by Glen Cowan on Wednesday touched on Wilk’s theorem

An important aspect of the workshop was the development of “Banff Challenge 2”. This challenge, running over fall 2010, is aimed at getting groups of statisticians and physicists to analyze data simulated data sets trying to detect signals either in analytically specified backgrounds or in situations where both the background and signals are described only by Monte Carlo data. An important component of the challenge, as with Banff Challenge 1, is the effect of systematic errors which are handled by either by specifying a prior of some sort or by auxiliary measurements.

The Challenge was discussed Wednesday morning in a series of presentations by workshop participants who had worked on a preliminary version of the problem. We returned to the discussion of the Challenge on Friday and a team led by Tom Junk and Wade Fisher has been working through the late summer to develop, distribute and publicize the challenge. Results are sought by early December in time for Phystat 2011 at CERN in January 2011.

One topic of intense discussion over the time period was the ‘look elsewhere effect’ or ‘multiple testing’ or ‘multiple comparisons’. Louis introduced the issue on Monday. A number of other speakers touched on aspects of the question and Eilam Gross spoke on Wednesday about the issue and about ‘trial factors’ – the ratio of a P -values appropriate for a single hypothesis to that for testing several hypotheses. Thus a trial factor is a number which corresponds to the number of hypotheses examined in a simple Bonferroni correction.

In a search for a peak on a background, a canonical problem in discovery, tests can be run looking for a peak at each fixed point in the spectrum and then we can think of scanning over those points and picking the smallest P -value. This generates a look elsewhere effect; it could be corrected for if we had effective large sample theory for the overall likelihood ratio statistic. This is a problem, however, because Wilk's theorem does not apply in this context. This issue is at the heart of Eilam's presentation. See his remarks below.

Thursday had many talks covering a variety of topics before we turned to summary presentations. David van Dyke summarized from the point of view of a statistician, Roberto Trotta from the point of view of an astrophysicist and Luc Demortier from the point of view of a particle physicist.

Friday morning had a talk from Wolfgang Rolke about nearest neighbour methods for doing multivariate goodness-of-fit – an important interest in the area which the workshop was not able to focus on enough. Friday morning also considered the future of the Banff Challenge 2.

3 Individual Reports on Progress Resulting from the meeting.

In this section we have gathered commentary from individual participants hoping to highlight the diversity of benefits we each took from the workshop. What follows are direct quotes, slightly edited by the conference organizers; they are presented in alphabetical order.

Henrique Araujo & Alistair Currie: Statistical analyses of WIMP search results are coming under close scrutiny as direct search experiments begin to probe the 'hot regions' of favoured parameter space. The problem of assessing the presence of a signal on the tails of poorly characterized backgrounds is a recurrent one in rare event physics. No general solution exists; several ideas were discussed with Glen Cowan, Bob Cousins, Bodhi Sen, Wolfgang Rolke and others on topics such as blind analyses, Feldman-Cousins with uncertain backgrounds, profile likelihood analyses, Bayesian methods and other topics.

Banff Challenge 2 proved useful in preparing for the workshop, helping us to identify similarities and differences between direct dark matter searches and typical collider scenarios.

Our Imperial colleague Roberto Trotta persuaded us that a Bayesian analysis of dark matter experiments should address uncertainties on the technology together with those on the Astrophysics and we could end up collaborating on that. The Banff meeting was the perfect setting for these discussions and for our ideas to mature during the week.

Tests presented to quantify the similarity of multivariate datasets also suggest applications to dark matter searches, where data-to-data and MC-to-data comparisons figure in background estimation and signal modelling. Overall it was a most useful workshop in the most wonderful of settings!

Jim Berger: Bayesian hypothesis testing and model selection were reviewed and illustrated with a very recent example of Vaccine Trials, indicating the importance of looking at matters from a Bayesian as well as a frequentist perspective. The initial analyses — which were based solely on p -values — could have erroneously caused a major shift in scientific efforts towards a likely incorrect avenue of research.

The reasons for the differences between p -values and Bayesian assessments of evidence were then discussed, with special focus on the issue of systematics or bias in the model for analysis. There is an intriguing indication that Bayes factors may be more resistant to such bias than are p -values.

The major difficulty with the Bayesian approach to hypothesis testing and model selection is the choice of prior distributions — this was highlighted in the talk by Bob Cousins. The "solutions" to this problem that exist in the statistics literature were discussed, although it was acknowledged that only a robustness or sensitivity analysis can definitively settle the issue.

Finally, the Bayesian approach to the "look elsewhere effect" ("multiple testing" or, simply, the problem of "multiplicity" in the statistics literature) was also discussed. Of particular interest was that Bayesians and frequentists approach multiplicity from very different directions, and that it is crucial to understand — and utilize — the relevant strengths of each.

Elliott Bloom: I found the discovery statistics meeting to be an instructive and very interesting meeting. For a number of months I have been working with a group of staff, post docs and graduate students at KIPAC-SLAC trying to understand the asymptotic behavior of the Test Statistic resulting from our likelihood fitting routine by comparing MC simulations to the predictions of Chernoff's theorem (I used to think that I was comparing to Wilks' Theorem, but the discovery statistics workshop educated me on this point).

In my talk to the conference I showed that the asymptotic behavior expected from Chernoff's theorem was badly violated in our MC simulations. Our results were received with interest by many people at the workshop and they made a number of valuable suggestions for further study by our group. This issue was also discussed in detail in the summary talk of David van Dyk, in which he also alluded to the result presented in my talk. Since returning we have tried a number of these suggestions and found that except for the absolute simplest case, fitting a peak with no background in a simple MC, Chernoff's theorem is not satisfied by our MC simulations, and as we approach more realistic simulations of our actual Fermi data analysis this disagreement becomes more and more severe. We are still trying to pin down root cause for these disturbing trends. One suggestion made by Roberto Trotta, was to try a Bayesian approach, we have been using frequentist theory. We are seriously considering implementing his suggestion at this time.

Jim Chiang: I found the BIRS workshop to have been very useful in many respects. Among the talks, I found the ones by Tom Loredo, Jim Berger, Michael Woodroffe, Chad Shafer, and David van Dyk to be particularly useful. Tom's discussions of the merits of the marginal likelihood vs the profile likelihood and the Neyman-Scott and related problems may be relevant for a bias we are finding in the analysis of Fermi data. David van Dyk's slides on his team's procedures for accounting for systematic uncertainties in the Chandra effective area described a framework in which similar sorts of calculations may be performed for Fermi data. He and his colleagues were kind enough to provide a draft of their paper in advance of publication, and we plan to implement some of their procedures. We are considering using the techniques Michael Woodroffe described for computing error probabilities via importance sampling for our own assessments of p-values. I'm happy to hear that Kyle and his student already have an implementation, and I have contacted Kyle about obtaining a copy of their code. As Kyle has already noted, he and I discussed RooStats at length, and I plan to implement some Fermi analysis using that framework and will provide feedback to Kyle and his colleagues that should help make that toolkit useful across disciplines. I had very useful discussions with Louis Lyons and Richard Lockhart on goodness-of-fit that helped clarify the relevant issues for me. My participation in this workshop inspired me to propose that the Fermi LAT team have a statistics board similar to those that exist for experiments such as CDF and ATLAS. A discussion that Elliott and I had with Joel Heinrich was extremely useful for defining the scope of the board functionality in this proposal, and the proposal was accepted by the LAT PI and analysis coordinators. Finally, I enjoyed the meals, hikes, BIRS lounge bull sessions, and after hours trips to the pub that helped make the whole event a very collegial experience.

Bob Cousins: This workshop was well worth my dedicated trip from CERN in Geneva, as it brought together a terrific mix of experts in a great environment. Thanks to excellent advance planning, the workshop attracted a nearly complete "Who's Who" group of high energy physicists who have an impact on statistical techniques in our field. The equivalent among astronomers was well represented as well, and the always-insightful statisticians included those which have learned about our problems in previous workshops (such as Jim Berger, Steffen Lauritzen and Michael Woodroffe) as well as some new faces (at least to me) who were impressively adept at understanding our specific issues and helping us with them.

My own interest was particularly sparked in several discovery-oriented areas, notably the Jeffrey-Lindley paradox, the Look-Elsewhere Effect, and comparisons of different ways of dealing with nuisance parameters (e.g., marginalization and profiling). In all these cases, I came away from the workshop with important insights that will directly affect my work in high energy physics. In other areas, such as uncertainties in parton distribution functions, I believe that I and the others at the workshop materially helped those struggling with their problems, offering suggestions and establishing contacts that will be mutually beneficial in the future.

My talk was on the Jeffreys-Lindley paradox, about which I had only superficial knowledge before preparing for the workshop. In this example, a Bayesian model selection calculation and a Frequentist hypothesis test for the same problem have different scaling behavior with the sample size n , so that in the limit of large sample size they can reach opposite conclusions, each with overwhelming significance. The workshop provided motivation for me to read some 25 papers and books on this topic, and to try to relate it to the way we approach analogous cases in high energy physics. I will almost certainly try to find the time to write up what I have learned and to follow up on a conjecture or two that grew out of this work. The Banff Centre was a perfect location for such a workshop. In additions to animated conversations over the three meals, we had a number of late-evening discussions in the Corbett Hall lounge that calcified some issues each day. I look forward to returning some day to the Banff Centre.

Glen Cowan: The BIRS meeting on Statistical Methods for LHC Physics provided an outstanding oppor-

tunity to finalize and report on recent work carried out by myself and three other workshop participants (E. Gross, O. Vitells and K. Cranmer) on use of profile likelihood methods for discovery significance and for setting limits. A draft of our paper had been finished just prior to the meeting (e-print: arxiv:1007.1727), and this was the basis of the talk that I presented. We apply asymptotic distributions based on the approximations of Wald and Wilks to find p -values for either the background-only hypothesis or a hypothesized signal and also to find the expected (median) discovery or exclusion significance.

Feedback received during the meeting was positive, and a few important points emerged that we are now incorporating into the paper's final draft. For example, our paper now addresses a criticism concerning zero-length confidence intervals. We also benefited from discussion with the statisticians present on the relation of the so-called Asimov data set to the expected Fisher information.

Beyond the progress related to our paper, I found the entire programme of the meeting interesting and useful, especially the work on the "look-elsewhere effect".

Kyle Cranmer: The BIRS workshop was very useful and very enjoyable. A number of my projects and collaborations either got a boost or were formed at the workshop.

Through conversations with Richard Lockhart and Earl Lawrence, we were able to precisely show the relationship of the "Asimov" dataset and the Fisher information matrix, which was only a conjecture before coming to BIRS. This result is being included in the second version of our paper on the arxiv, which Eilam, Ofer, Glen, and I will submit for publication shortly. We have thanked Richard, Earl, and BIRS in the acknowledgements. The result is also relevant for speeding up the calculation of Jeffreys's prior, which may also impact the work on reference priors being done by Luc Demortier and Harrison Prosper.

During the workshop we discussed a number of ideas which I hope to see implemented in RooStats, which may have impact on the entire field. In particular, my graduate student, Sven Kreiss, has implemented the importance sampling techniques described by Michael Woodroffe, which can bring huge gains for the computationally expensive LHC Higgs combinations. That development should go into the next release of ROOT. Similarly, Luc, Harrison, and I were able to develop a plan for how their `refpriors` package can be interfaced and incorporated into RooStats. Steffen Lauritzen graciously sat with me to work through the graphical models corresponding to our HEP problems. Particularly interesting was the "max propagation" and "random propagation" algorithms, which may provide important speedups for our most common HEP problems. We hope to employ these new techniques in the context of the Banff challenge 2, which I hope will be as successful as the first Banff challenge.

I was happy to get back in touch with Bodhi Sen, who gave me some important insight into the relationship of the bootstrap and the algorithm we currently use to generalize of the Feldman-Cousins technique with nuisance parameters. Of course, I enjoyed lunch and dinner conversations with all of the participants, most memorably those with Jim Berger, Bob Cousins, Gary Feldman, and Tom Lored.

Roberto Trotta and I were able to bring back to life a stalled project to estimate the coverage properties of current techniques that are used to infer regions of SUSY parameter space that compatible with a variety of experimental results. We hope to show some preliminary results at a conference in Stockholm in September.

Lastly, I had a long and pleasant conversation with Jim Chiang on our hike about the possibility of using RooStats in the analysis of data from the Fermi Gamma Ray Telescope. This development may have important consequences in our understanding of dark matter and a plausible combined analysis of LHC and Fermi data.

Luc Demortier: The most impressive aspect of the workshop was the high quality of all the talks. I learned something from each of them, but was particularly interested in some of the ideas presented by statisticians: a new importance sampling technique (M. Woodroffe), a goodness-of-fit test with Bayesian prior on alternatives (R. Lockhart), D. van Dyk's solution to the sensitivity problem in the calculation of upper limits, C. Schafer's decision-theoretic approach to parton densities and the Banff challenge, and Steffen Lauritzen's random effects model to determine the parton densities. Regarding the latter, the talks by J. Pumplin and R. Thorne were very useful and enlightening. It may be that one of the greatest successes of the workshop was the decision by these speakers to attempt a closure test on their procedure to determine the parton densities. On another front, there was a lot of discussion about the so-called Asimov data, but I remain somewhat skeptical of the validity of this method in more complicated settings than the usual illustrative examples. I enjoyed T. Lored's talk on profile versus marginal likelihood. Finally, I should also mention several useful discussions I had with J. Berger about Bayes/frequentist points of contact.

With the help of a summer student I have done some work on the Banff challenge, and plan to write up our results for discussion with other interested parties. Many workshop participants showed interest in the look-elsewhere effect, and this has inspired me to try and write up a review of the extensive and still evolving statistics literature on the subject.

I gave one of the summary talks at the workshop, and the above is more or less a summary of that summary.

Eilam Gross: The Banff workshop was the most beneficial workshop I have been to in my life. Besides learning a great deal about statistics and meeting remarkable people, it is thanks to the Banff workshop that together with Ofer Vitells, we have managed to fully complete and understand our own research on the “Look Elsewhere Effect”.

In his summary talk in the 2010 BANFF workshop Luc Demortier drew our attention to the work of Davies from 1977 which became the leading thread of our revised work. Michael Woodroffe whom we also met at BANFF, spent his valuable time to explain to us how to adopt the statistical language of Davies to the High Energy Physics jargon. He also spent valuable time in writing to us his impressions on the Look Elsewhere Effect. The revised version of our paper on the “look elsewhere effect” would have never been possible without the Banff workshop; for us this paper is a major scientific achievement.

On top of all this the magnificent atmosphere in Banff with the amazing hospitality set up the right ground for scientific developments. We have no words to express our thanks to the organizers and the team.

David Hand: The meeting was an eye-opener in revealing to me the breadth of statistical issues in which the particle physics and astronomy/cosmology communities had an interest. I knew of their interest in coping with massive data sets, but I had not realized that they also had a matching interest in the more philosophical subtleties of statistical inference. It served to reinforce my belief that those areas of physics are ones to which statisticians can make useful contributions. The recent surge of interest in these areas amongst statisticians (e.g. the establishment of the ISI group on astrostatistics) is very timely. I look forward to following up the various discussions I had outside the formal talks, and on the plans I made with various participants to collaborate on exploring some of their problems in detail.

Chris Hans: I found the Banff meeting to be very interesting. Most of the conferences I attend are organized by and for statisticians, and it was a pleasure to hear about statistical issues in astronomy and physics at this meeting directly from the source. While I can't say that I have developed any collaborations based on the meeting, I do feel that I gained a better understanding of which particular statistical areas are of interest – and importance – to researchers in these fields and will keep this in mind as I develop my research in these areas over the next few years. In terms of interactions at the meeting, I particularly enjoyed a few conversations I had with Tom Loredo about some of my work on Bayesian regularization priors and its connections to questions of model uncertainty. I also very much enjoyed meeting several statisticians who I had not yet met beyond earlier brief introductions (in particular Chad Schafer, Earl Lawrence, Nicolai Meinshausen and Bodhi Sen). In this sense, the meeting was successful in not only bringing together scientists across disciplines but also in bringing together statisticians across sub-disciplines who might not otherwise have an opportunity to interact and share ideas in such a small and productive setting.

Joel Heinrich: As a gathering of people from the HEP, Astrophysics, and Statistics worlds, the Banff meeting was helpful to me in several ways. I became informed on current trends in the HEP-statistics community, and the views of the statisticians regarding those trends. Learning about statistics practice in astrophysics provided a useful contrast to the practice in HEP which is familiar to me. Since the meeting, I have become involved in the design phase of Banff Challenge 2, which is intended to provide an additional forum for new methodology to be applied to the typical discovery problems in HEP.

Thomas Junk: This workshop was very productive. I met with other particle physicists, astrophysicists, and statisticians from July 11 to July 16. We discussed the issues related to how to make discoveries in particle physics and astrophysics; issues relating to the false discovery rate, such as Why do we like 5 sigma? What happened during historical non-discoveries like the Pentaquark and the 40-GeV top quark? Statisticians bring a unique point of view to the subject, and work is ongoing at CERN in the ATLAS and CMS statistics committees to work out their details of setting limits and discovery procedures. They were impressed with our care and rigor. I made three presentations, one on practical experimental details of interpreting search results, one on the challenge problems (homework for participants), filling in for Wade Fisher who could not attend, and one presentation on my solutions to the challenge problems. We will continue to work on

these challenge problems to generalize them and make them more useful in the near future. I also learned about “power-limited Feldman-Cousins” and the simpler “power-limited CLs+b” techniques and will give them some thought. Techniques also for reducing our need to run computationally expensive exclusion and evidence/observation calculations were also brought up that I am interested in testing in the future. I learned that the look-elsewhere effect depends on the data sample size, an effect that in hindsight makes a lot of sense, but I was unaware of it before this meeting. I am also much happier and more comfortable with our treatment of the look-elsewhere effect, which has a degree of arbitrariness in defining “elsewhere”; There was more agreement on that subject than I was expecting, and now we can proceed with confidence.

Steffen Lauritzen: A very interesting meeting. I am in contact with Jon Pumplin and Robert Thorne [see the discussion from Robert Thorne below] to follow up on my remarks about parton distribution functions. I hope something comes out of that.

Jim Linnemann: I found the Banff workshop useful in a number of ways. First, I found many of the talks informative and stimulating. I was also able to use the occasion to communicate directly with colleagues on matters of interest. In particular, I discussed with Robert Thorne (global parton distribution function fitting) and with Lorenzo Moneta some things I’ve learned in the last year on nonlinear fitting; I hope that Lorenzo can move some of this information into root where it will be accessible to a broad user community, and he seems interested in that path. I also found the discussion with colleagues on state ordering in the Look Elsewhere Effect to be clarifying, and expect this will be reflected in two physics papers in progress. Michael Woodroffe in particular has suggested some interesting ways to think about this problem, and I intend to follow up with him in this area. I have been involved in the setting of the initial Banff Challenge for this workshop, and also in the followup effort, which we expect will stimulate more effort and be reported on at the Phystat 2011 conference in Geneva. In addition, conversations with several statisticians (David Hand, Richard Lockhart, and Earl Lawrence) have identified common areas of interest which could lead to collaborations.

I also had interesting conversations with: Wolfgang Rolke on his proposed multidimensional goodness-of-fit (and pointed him to a paper by Friedman at an earlier phystat); Tom Loredo on 2-d angular difference measures in astrophysics; and with Gross et al’s whose paper shed light on issues we’d seen in effective number of trials in an astrophysics experiment.

Nicolai Meinshausen: The very stimulating meeting in Banff was interesting in many ways for me. The treatment of systematic errors in the particle physics simulation models is very much related to similar problems in climate models and I hope to be able to transfer some of the ideas between the fields in the future. It was also very fruitful to me meet some astronomers, notably Tom Loredo and people working on the Fermi experiment, and discuss the detection of periodic signals, a problem I have been working on and published about previously and which I intend to take up again, using partially the very useful input I got out of informal discussions at the meeting.

Lorenzo Moneta: This has been my first workshop at Banff and I have found it extremely useful for my work. It has been one of the most interesting and productive workshop I have participated. I have learned a lot about statistics at both theoretical and practical level from attending the lectures and participating in the discussions. For example, I have now a much clearer picture on what are the problemats in using the likelihood function to establish a discovery significance. This is very useful for my job to manage and develop software statistical tools for the data analysis of the LHC experiments.

I enjoyed very much the discussions with my colleagues from HEP, with the astro-particle physicists and the statisticians. The workshop provided a great opportunity to discuss together our statistical problems and to learn from each other. We have been discussing possible improvements for the RooStats package, like including the reference prior in the Bayesian analyses.

From listening to the lectures, I developed ideas for implementing new tools in the ROOT software package, such as automatic binning from histograms using Bayesian methods or new method for goodness of fit of multidimensional data. From discussing with Jim Linnemann, I will start investigating the possibility to improve the current minimization algorithm we are using in HEP (Minuit), to deal better with non linearity and with the problem of converging to a local minimum instead of the global minimum. This algorithm is the most common used algorithm in HEP for solving non linear fits, like those presented at the workshop for finding the parametrization of the parton structure function or for evaluating the discovery significance using the likelihood function. Furthermore, I enjoyed very much the pleasant atmosphere and the wonderful

location. Thank you very much to the organizers and to BIRS.

Chad Schafer: The main point of my talk was to present an approach to constructing confidence regions/hypothesis tests which are optimal with respect to a clearly defined, yet user-specified, notion of performance. In particular, using standard decision theoretic ideas, one can construct decision procedures that possess frequentist coverage, but have maximal power against alternatives considered physically feasible. It is common that one seeks procedures with such properties, and standard approaches do exist (e.g. Wilks' Theorem) for well-behaved situations. For situations in which one has a complex model (likelihood function), care must be exercised. Although I did not describe it in any detail, there is a Monte Carlo procedure for approximating the aforementioned optimal procedure; it is designed to work in (indeed, it was motivated by) these cases where one has a complex likelihood function, or for some other reason cannot rely upon the asymptotic approximations of Wilks' Theorem. My hope is that this approach could be of value in addressing both the Second Banff Challenge, and the quantifying of the amount of uncertainty in the estimates of the Parton distribution function. I found the Workshop to be an ideal setting to explore the challenging inference problems in particle physics, and look forward to pursuing these further in direct collaboration with the physicists.

Jeffrey D. Scargle: I found the Banff workshop was very useful for me on both practical and theoretical levels. One of my main interests is in the use of modern statistical techniques to astrophysical data. As you know, the lines between astrophysics and particle physics are blurring – hence the field of astroparticle physics. There were many presentations extraordinarily well focused on the corresponding issues. I enjoyed very much working on the Banff Challenge Data; one always learns a lot by coming to grips with actual data – be they synthetic, experimental, or observational.

Bodhisattva Sen: The Banff workshop was very exciting. It was mostly an educational experience for me, as I am still trying to understand the major statistical issues in HEP. I discussed some related statistical concepts to some of the physicists in one-to-one conversations. I hope that some of these synergistic activities will lead to real collaborations in the future. I also plan to take a closer look at the Banff Challenge data, in the near future. The organization of the workshop was exemplary. I very much enjoyed the visit and most of the talks, although I think that a few of the talks could have highlighted the statistical aspects of the problem more clearly.

Paul Sommers: Thank you very much for inviting me to participate in the Banff workshop on “Statistical issues relevant to significance of discovery claims.” This was a particularly valuable experience for me. As co-spokesperson for the Pierre Auger Collaboration, I am facing numerous problems that relate directly to statistical methods for assessing the significance of intriguing anisotropy correlations seen in our data, and also the problem of reporting sensible upper limits for point sources of neutral particles (neutrons and gamma rays). This was a great opportunity to learn from experts, and I was pleased to be able to present some results from the Auger Cosmic Ray Observatory.

The workshop was an opportunity to meet numerous distinguished persons whose work I know from the literature, and also to become acquainted with some outstanding scientists that I did not know about previously. It was an intellectually enriching experience in a delightful setting. Thanks again.

Robert Thorne: The Banff workshop was both useful and enjoyable, and from my viewpoint was unusual in the breadth of subject area expertise covered by the (relatively small number of) participants. However, this meant that my talk, which became talks, spent rather a longer time covering the basics than expected and did not really get to the precise details of how the procedures used by different groups differ in detail. However, it was gratifying that most (perhaps all) of the audience were happy to accept that this is a difficult problem, and also that the need to inflate the textbook determination for uncertainties of parton parameters was not found to be surprising.

It terms of determining more precise reasons for understanding why this inflation is necessary, the proposal to generate a set of data from the theory, but then to obtain the uncertainties by scattering according to the true experimental uncertainties in order to obtain a global set of data which is both consistent with itself and the theory will certainly be performed, and should be straightforward. Also generating a set from e.g. NNLO theory and attempting to fit with NLO, i.e. having self-consistent data set which does not match the theory perfectly is the obvious next step. Results will be interesting and I will keep people in touch.

I am also intrigued by various of the the proposals to solve the assumed problems of incompatibility of different data sets and/or of data and theory in a more statistically robust manner than used by the various

groups at present. In particular that of Steffen Lauritzen to modify the χ^2 definition to account for different data sets preferring different values of the parameters using Random Effect models. The general principles behind this do indeed seem to match the problem and I hope to pursue this further, though it will require more new work than the simpler checks above.

Roberto Trotta: I found the meeting highly interesting and enjoyable. I valued in particular the opportunity to interact with Paul Summers, Tom Loredo, Bob Cousins, Jim Berger, Chad Shafer, with whom I had several interesting discussions regarding various aspects of my research. Kyle Cranmer and I took the opportunity of the workshop to restart a project we had been working on together, with the aim of publishing the results after the Summer. I also had the opportunity to give one of the summary talks of the meeting, in which I tried to describe synergies and differences between the problems and approaches discussed during the workshop and some of the currently ongoing research in cosmology.

Ofar Vitells: I found the BIRS workshop very useful and educational. Both the lectures and discussions provided many important insights into the statistical problem that were addressed. In particular we had useful discussion and feedback on our “Asimov” paper which is about to be submitted for publication soon (with Kyle, Eilam and Glen) as already mentioned by Kyle. In addition we got very helpful comments and references related to our work on the “look elsewhere effect”. Michael Woodroffe and Luc Demortier pointed us to some related work that might help in placing some of our conjectures on a more solid ground. We are currently working on this in collaboration with Michael Woodroffe who has kindly agreed to help us with the mathematical formulation. I had also very interesting discussions with Henrique Araujo and Alastair Currie on their views on the statistical challenges of experiments that search for dark matter, and that will certainly contribute to our future work with the Xenon100 collaboration.

Michael Woodroffe: I got a better understanding of the physics and some new problems to pursue. I am following up with Eilam Gross and Jim Linnemann. With Eilam, I am working out the details of how Davies results apply to his problem. Jim’s nested multiple hypotheses remind me a bit of a problem that arose in sequential analysis circa 1980. I think that a similar formulation might capture the effect that he wants. I spoke on how importance sampling was used in sequential analysis and how I think it can be used in the discovery problem.

4 Summary

This workshop started many useful collaborations and introduced many of us to new ideas. The follow up over the fall of the Banff Challenge 2 should be very productive. Finally, this workshop will set the stage for much useful discussion at PHYSTAT 2011 at CERN in January 2011.