Open Issues in the Wake of Banff 2010

Luc Demortier

The Rockefeller University

PHYSTAT 2011 Workshop, CERN, 17–20 January 2011

Outline

1. Discovery claims
2. Measurement sensitivity
3. Implicit statistical models
4. Parton density function uncertainties
5. Profile likelihood methods
6. Reference priors
7. Extreme value theory

The website of the Banff 2010 workshop contains links to the presentations and other relevant material:
http://www.birs.ca/events/2010/5-day-workshops/10w5068.

1 Discovery Claims

Discovery Claims

Discovery claims can only be justified if one has a method for quantifying the evidence for or against a given hypothesis. . .

- The tradition in High Energy Physics is to compute a p-value (i.e. the probability of obtaining something at least as extreme as the actual observation), and claim discovery if $p \leq 2.87 \times 10^{-7}$.
- An alternative approach is to compute a Bayes factor (likelihood ratio) or a hypothesis probability.

These two approaches do not always agree, as demonstrated, for example, by Lindley's paradox:

Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\theta, \sigma^2)$, and suppose we wish to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Prior probability is π_0 on H_0, and uniformly distributed over a large interval $\mathcal{I}$ of θ values under H_1. Then the posterior odds in favor of H_0 are:

$$\frac{\pi_0}{1-\pi_0}\,\mathcal{I}\,\frac{e^{-Z^2/2}}{\sqrt{2\pi}\,\sigma/\sqrt{n}}, \qquad \text{where} \qquad Z \equiv \frac{\bar{X}-\theta_0}{\sigma/\sqrt{n}}.$$

Suppose now that Z indicates strong evidence against H_0; keeping Z fixed, it is always possible to find n large enough for the posterior odds to favor H_0.

Structure of Hypothesis Tests in HEP (1)

One of the lessons from Lindley's paradox is that, in order to correctly interpret the result of a hypothesis test, one needs to specify as clearly as possible one's prior belief structure regarding the hypotheses being tested. When searching for new physics, a common type of test is the following:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0,$$

where θ is, for example, the production rate of a new particle.

What is our prior belief structure regarding H_0 and H_1?

- Even though the physical theory underlying H_0 describes a vast body of previous observations extremely well, we know that it is incomplete, and that somewhere it predicts something that will not be observed. Fundamentally the theory is wrong.
- However, we do not know where the breakdown will occur. There are many possibilities for this, and many potential explanations. It is also possible that we haven't yet formulated the correct new physical theory.

Therefore, when it comes to testing H_0 for a particular type of breakdown, our prior belief in H_0 tends to be much stronger than that in H_1.

Structure of Hypothesis Tests in HEP (2)

For this type of test, a precise null hypothesis with a strong prior belief in it, it has been shown that *p*-values substantially overestimate the evidence against the null as measured by Bayes factors, posterior probabilities, and lower bounds on these over wide classes of priors.

Furthermore, *p*-values cannot be "recalibrated", because

1. the calibration depends on the sample size;
2. the interpretation of a *p*-value depends on the model pdf;
3. the interpretation of a *p*-value depends on the stopping rule;
4. the interpretation of a *p*-value depends on the type of null hypothesis being tested (precise versus diffuse).

For further details, see J. O. Berger and M. Delampady, "Testing precise hypotheses," Statist. Sci. **2**, 317 (1987).

For the above reasons it is unfortunate that in our field we have a rigid 5σ discovery threshold. . .

The Look-Elsewhere Effect: p-Value Corrections

When searching for a resonance somewhere in a spectrum of given width, the significance of an interesting local excess must be corrected for the fact that a background fluctuation like the observation could have occurred *anywhere* in the spectrum. This is the look-elsewhere effect (LEE).

The statistician R.B. Davies computed the LEE correction to p-values in 1987 (Biometrika **74**, 33). Suppose that for each value of the resonance location $\theta \in [A, B]$, the test statistic $S(\theta)$ is (asymptotically) chisquared with s degrees of freedom. Davies derived the following formula for the LEE-corrected tail probability:

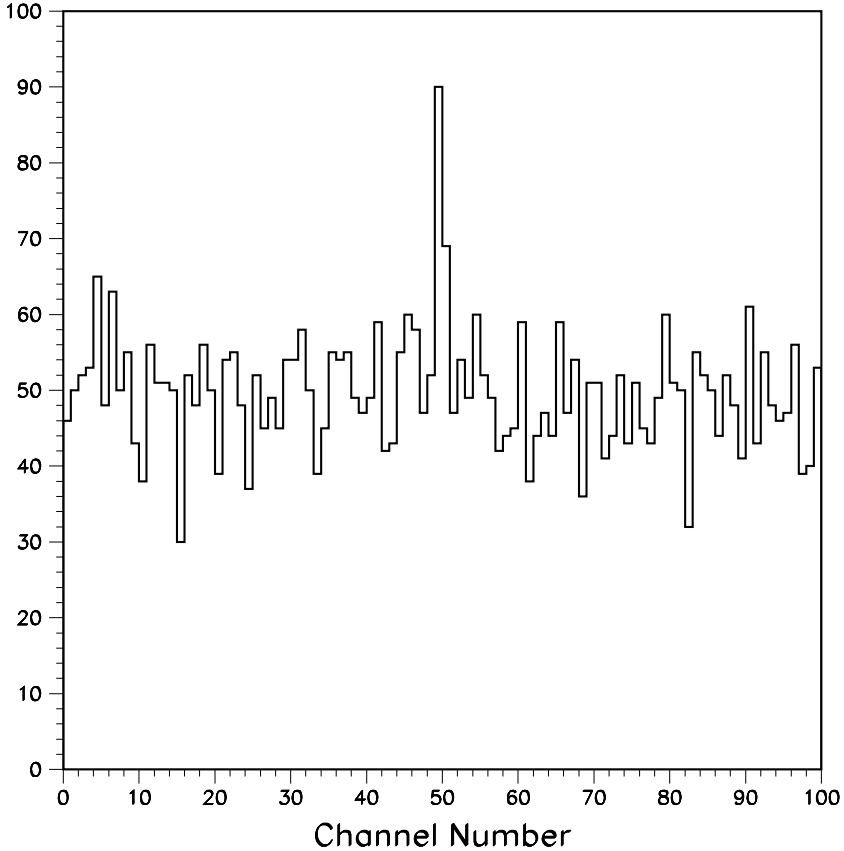
$$\mathbb{P}\left[\sup_{A\le\theta\le B} S(\theta) > u\right] \le \mathbb{P}(\chi^2_s > u) + \langle N(u)\rangle, \qquad (1)$$

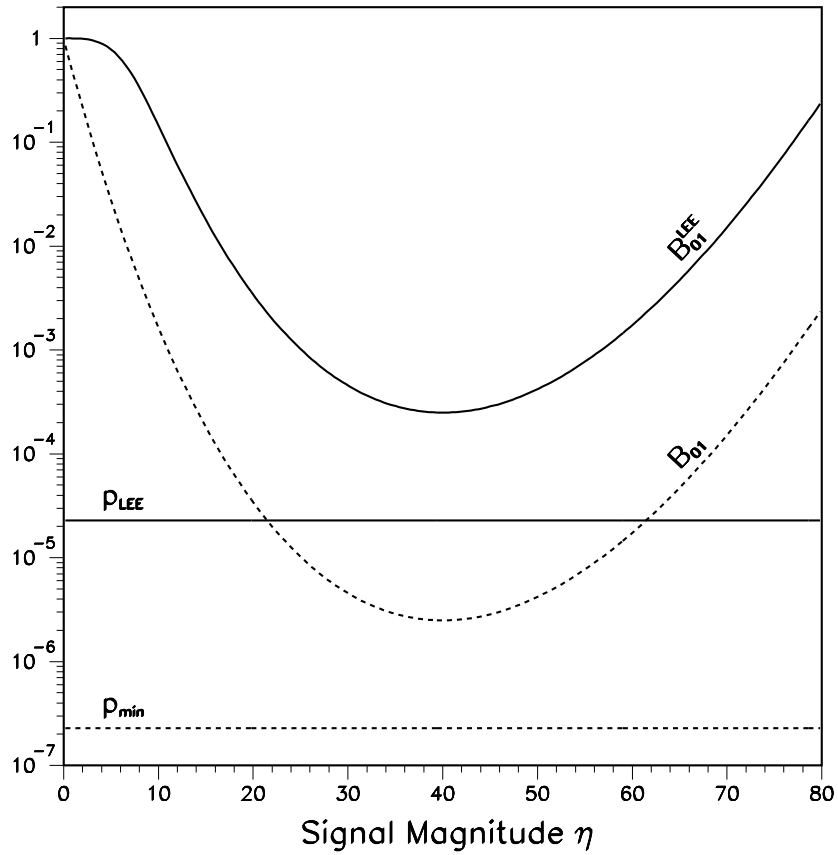
where $\langle N(u)\rangle$ is the expected number of upcrossings of the level u by the process $S(\theta)$.

E. Gross and O. Vitells recently described a simple procedure to compute $\langle N(u)\rangle$ and gave an interpretation of Davies's formula in the limit of large u. See http://arxiv.org/abs/1005.1891v3.

The Look-Elsewhere Effect: *p*-Values versus Bayes Factors

Search for a peak of unknown location and magnitude:





Left: spectrum of Poisson counts used to illustrate the look-elsewhere effect on *p* values and Bayes factors. Right: Bayes factor in favor of the background-only hypothesis, with and without LEE correction, compared with the corresponding *p* value.

2 Measurement Sensitivity

Measurement Sensitivity (1)

Let X be an observable whose pdf depends on a parameter θ, such that we wish to test:

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta > 0, \tag{2}$$

using the p-value:

$$p_0 = \mathbb{P}\left[X \geq x_0 \,\middle|\, H_0\right].$$

Suppose we find $p_0 > \alpha$, so that we accept H_0. This does not mean that all values of θ under H_1 are now rejected: there are values of θ that our experiment is not sensitive to, and others that the data won't allow us to exclude.
One way to investigate this is to test

$$H_1[\theta_1] : \theta = \theta_1 \quad \text{versus} \quad H_0 : \theta = 0, \tag{3}$$

for each θ_1 under H_1, using the p-value:

$$p_1(\theta_1) = \mathbb{P}\left[X < x_0 \,\middle|\, H_1[\theta_1]\right].$$

We then reject $H_1[\theta_1]$ if $p_1(\theta_1) \leq \gamma$ for some γ, and we can define a $(1-\gamma)$ C.L. upper limit θ_u as the largest value of θ_1 that is not rejected by test (3).

The problem with this procedure is that there is a finite probability of excluding all θ values (i.e., whenever $\theta_u \leq 0$), regardless of the measurement sensitivity.

Measurement Sensitivity: the CL_S Method

To prevent exclusion of parameter values to which the experiment is not sensitive, the CL_S method replaces the frequentist rejection criterion

$$p_1(\theta) < \gamma$$

by

$$\frac{p_1(\theta)}{p_1(0)} < \gamma.$$

This prevents rejection of small values of θ, i.e. values of θ for which $H_1[\theta]$ gets too close to H_0. The price to pay for this feature is an excess of frequentist coverage at low θ.

It is interesting to note that for the simplest problems, CL_S upper limits agree with Bayesian limits for some choice of prior (typically improper).

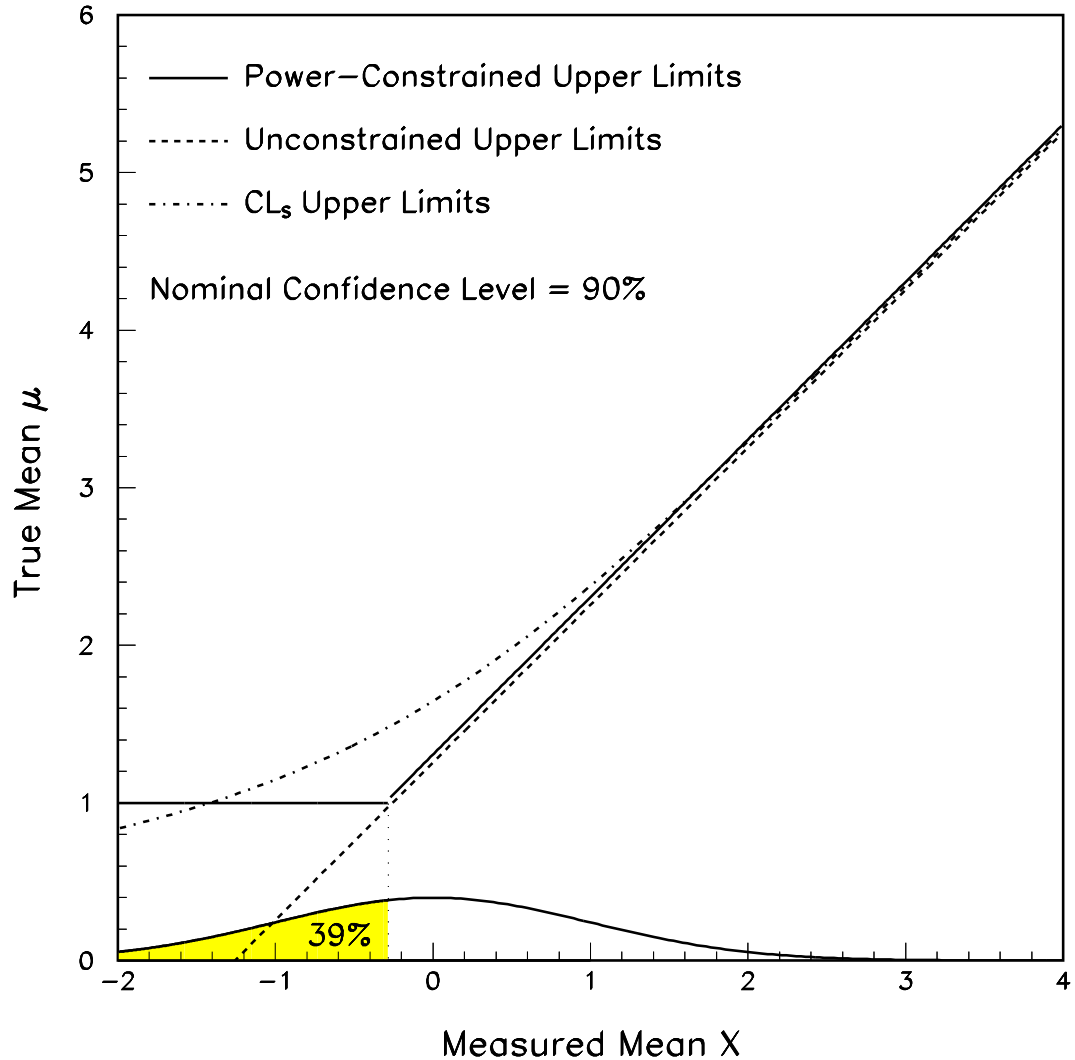
Measurement Sensitivity: Power-Constrained Limits

An alternative approach is to report the observed upper limit only if it is above a prespecified "sensitivity bound". If the observed limit is below the bound, only the bound itself is reported. This was proposed by V. Highland in an unpublished note in 1987. Some colleagues from ATLAS have motivated this method with a statistical power argument: you shouldn't reject a given parameter value unless you have a decent probability of detecting it when it is the true value. Hence the name "power-constrained limits" (PCL). A delicate issue here is the choice of sensitivity bound.

Some astrophysicists have proposed to always report both the observed upper limit *and* a minimum sensitivity bound (caveat: their terminology interchanges the concepts of *upper limit* and *upper bound* as understood in HEP). See V. L. Kashyap *et al.*, "On computing upper limits to source intensities," arXiv:1006.4334v1 [astro-ph.IM] (2010).

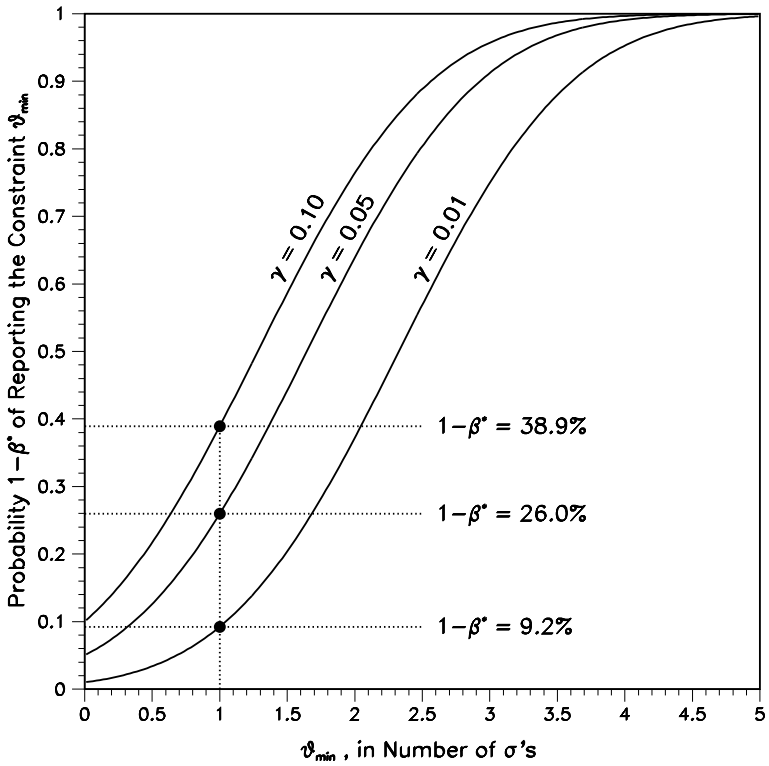
Measurement Sensitivity: Comparison of Methods

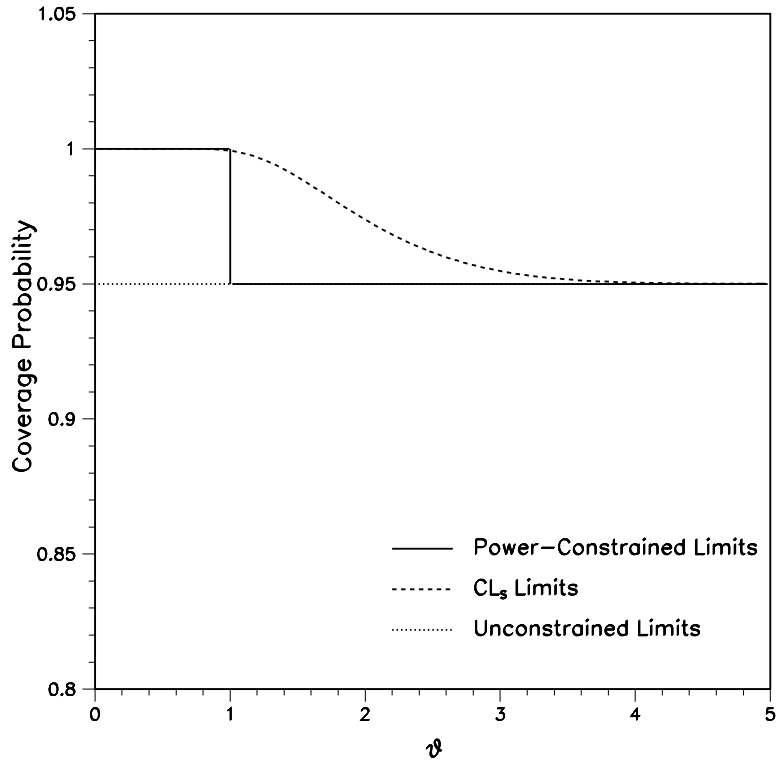
Example: measuring the mean of a Gaussian distribution with unit variance.



Measurement Sensitivity

Gaussian example, continued:

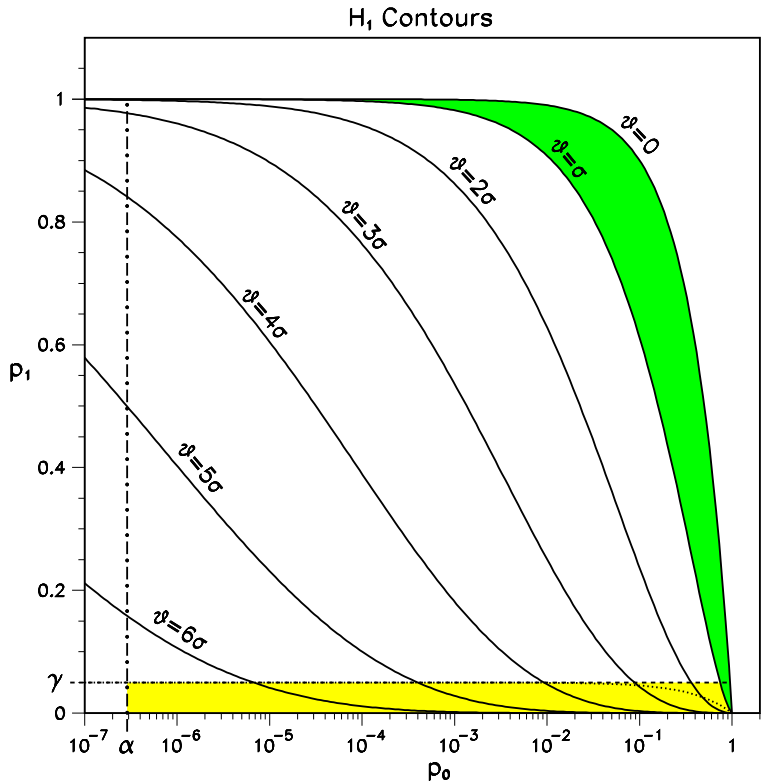


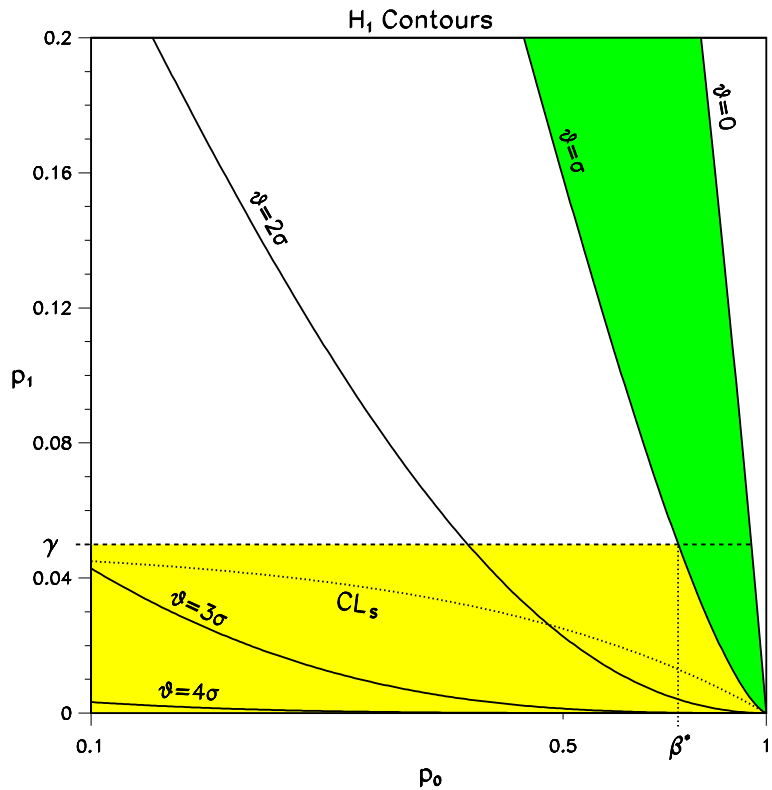


- Left: Probability of reporting the sensitivity bound θ_{min}, versus θ_{min}, for three choices of γ, where $1 - \gamma$ is the upper limit confidence level.
- Coverage of PCL, CL_S, and standard frequentist limits.

Measurement Sensitivity

Still for the Gaussian example:





- Left: p_1 versus p_0 plot. Every measuring apparatus is represented by a contour, and every measurement is a point on that contour.
- Right: Zoom-in on the lower right-hand corner of the left figure.

3 Implicit Statistical Models

Implicit Statistical Models (1)

Many HEP measurements are complex: we do not know the exact analytical dependence of the likelihood function on some parameters of the model; all we have is the underlying stochastic mechanism, which we can simulate with a Monte Carlo algorithm. This difficulty occurs for both nuisance and interest parameters.

Take for example the measurement of the mass μ of a new particle. The data sample consists of a signal component (events containing the new particle) and an irreducible background component. If we have an event by event estimator X for μ, the likelihood has the form:

$$\mathcal{L}(\mu) \;=\; \prod_{i=1}^{N} \Big[(1-\epsilon_b)\, p_s(x_i \,|\, \mu) \;+\; \epsilon_b\, p_b(x_i) \Big] \;\times \ldots$$

Typically the X distributions p_s and p_b are histograms obtained via Monte Carlo simulations. These histograms may be smoothed or fitted with parametric representations. In addition, the p_s distribution must be constructed on a grid of μ values supplemented with interpolation. This is inefficient since a lot of time is wasted modeling $p_s(x \,|\, \mu)$ at μ values far from the MLE.

Implicit Statistical Models (2)

Finally, p_s and p_b also depend on nuisance parameters such as energy scales, initial and final state radiation, parton densities, etc. Generalizing the above approach to multiple parameters quickly becomes impractical [see P.J. Diggle and R.J. Gratton, "Monte Carlo methods of inference for implicit statistical models," J. R. Statist. Soc. B**46**, 193 (1984).]

Over the years, a number of ingenious but somewhat dubious shortcuts were invented by high energy physicists to take nuisance parameters into account. An example shortcut is to evaluate the shift $\Delta\mu$ in the MLE of μ induced by a one-sigma variation of a given nuisance parameter, and then to replace the likelihood by its convolution with a Gaussian with standard deviation $\Delta\mu$:

$$\mathcal{L}(\mu) \;\rightarrow\; \tilde{\mathcal{L}}(\mu) \;\equiv\; \int \mathcal{L}(\mu') \, \frac{e^{-\frac{1}{2}\left(\frac{\mu-\mu'}{\Delta\mu}\right)^2}}{\sqrt{2\pi}\,\Delta\mu} \, d\mu'$$

When there is more than one nuisance parameter, $\Delta\mu$ is replaced by the sum in quadrature of the individual shifts.
The validity of this method has never been verified!

Better approaches may be available, both for frequentist and Bayesian inferences

Implicit Statistical Models: ABC Methods

In the Bayesian paradigm, the likelihood is integrated over the nuisance parameters, a feature that lends itself well to Monte Carlo computations. Implicit statistical models can be analyzed with the help of so-called ABC methods (Approximate Bayesian Computation). The goal is to *approximate* the posterior distribution $\pi(\mu \mid x) \propto p(x \mid \mu)\,\pi(\mu)$. All we need is a suitable distance function $d(x_a, x_b)$ between two datasets x_a and x_b. The simplest ABC algorithm is the ABC rejection sampler:

1. Sample $\mu^\star$ from $\pi(\mu)$.
2. Simulate a dataset $x^\star$ from $p(x \mid \mu^\star)$.
3. If $d(x_{obs}, x^\star) \leq \epsilon$, accept $\mu^\star$, otherwise reject.
4. Return to step 1.

The output of an ABC algorithm is a sample of parameters $\mu^\star$ from a distribution $\pi(\mu \mid d(x_{obs}, x^\star) \leq \epsilon)$. If ϵ is sufficiently small, this distribution will be a good approximation to the posterior $\pi(\mu \mid x_{obs})$.

There are other ABC algorithms, more efficient than the rejection sampler, and which work with improper priors; see T. Toni *et al.*, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," J. R. Soc. Interface **6**, 187 (2009).

Implicit Statistical Models: Decision-Theoretic Methods

In the frequentist paradigm, one is interested in procedures that have coverage for all values of the interest and nuisance parameters. Other requirements besides coverage are needed to specify unique procedures.

For the construction of confidence intervals, one approach, based on decision-theoretic ideas, is known as minimax expected size (MES): it minimizes the maximum expected size of the confidence set over parameter space.

In a Monte Carlo implementation of MES, parameter values are drawn at random from the parameter space, and a dataset is simulated for each parameter value. Each simulated dataset is compared to the observed dataset using a likelihood ratio test. Inverting the likelihood ratio test minimizes the probability of including false values in the confidence region, which in turn minimizes the expected size of the confidence region. This Monte Carlo algorithm does not require explicit knowledge of the likelihood function, only of the data generating mechanism.

See C. M. Schafer and P. B. Stark, "Constructing confidence regions of optimal expected size," J. Amer. Statist. Assoc. **104**, 1080 (2009).

Implicit Statistical Models: Two Final Comments

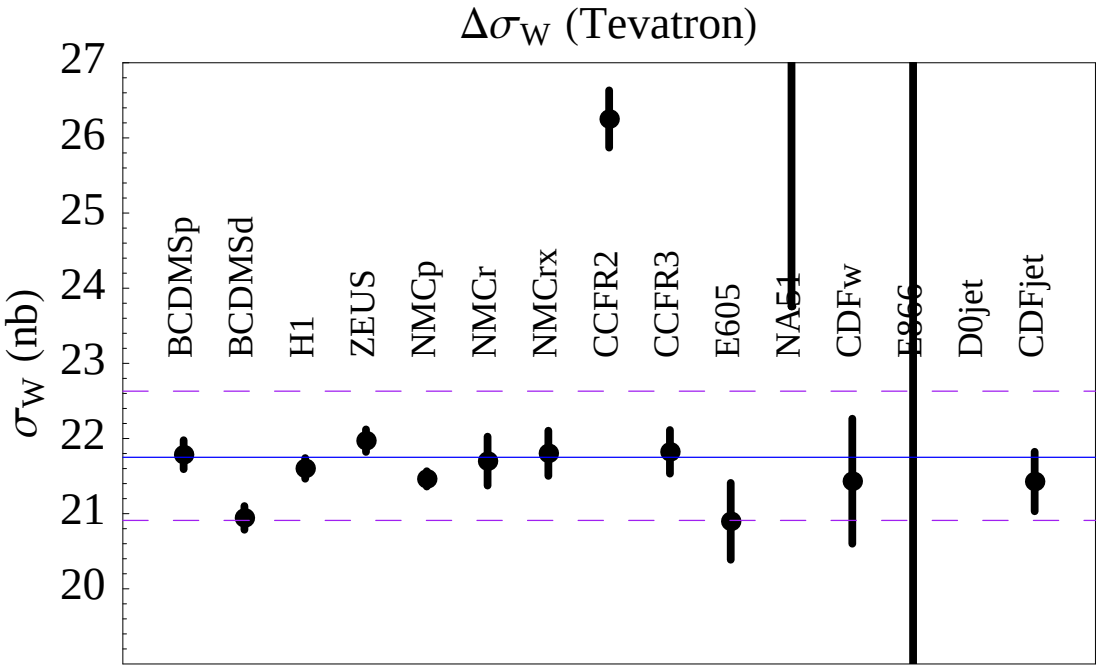
1. At present the Bayesian approach via ABC methods seems a lot more flexible than the frequentist approach, since ABC methods produce an approximation to the posterior itself. The decision-theoretic procedure only produces confidence intervals, and only of the MES type (no choice of ordering rule).

2. An important consideration in HEP is the combination of measurements from two or more experiments. Different experiments may have common uncertainties, and the resulting correlations must be taken into account. This seems doable with ABC methods, although the generation of Monte Carlo samples (an industry in itself) will have to be redesigned and much more carefully coordinated between experiments.

4 Parton Density Function Uncertainties

Parton Density Function Uncertainties (1)

PDF uncertainties affect our ability to make predictions and claim discoveries. Currently the PDFs are determined by a fit to $\sim$ 35 datasets with a total of $\sim$ 3000 data points. The standard parametrization of the PDFs uses $\sim$ 25 parameters. Fit quality is characterized by a χ^2 value. PDF uncertainties are derived from a $\Delta\chi^2$ procedure, but the standard $\Delta\chi^2 = 1$ rule yields clearly unrealistic uncertainties:



Instead, 90% CL uncertainties are obtained via $\Delta\chi^2 = 100$ or 50, depending on the group doing the fit.

Parton Density Function Uncertainties (2)

The pdf uncertainties are not yet understood from a statistical point of view. Some suggestions were made at Banff to improve this situation:

- A decision-theoretic approach such as MES.
  This may be of value for quantifying the uncertainty in the pdf estimates (Chad Schafer).
- A random effects model.
  Assume that the theory does not quite fit each experiment, resulting in underestimated prediction errors. Propose as solution that the theory parameter is slightly different in each experiment, and all these individual parameters are constrained to the formal parameter of the theory via some distributional assumptions (i.e. a prior, such as multivariate-t). (Steffen Lauritzen)
- A closure test.
  Harrison Prosper has teamed up with a theorist from LPSC (Grenoble) to do this: first verify that for data generated from the theoretical distributions, the $\Delta\chi^2 = 1$ criterion yields reasonable uncertainties. The next step is to study how inferences are affected by biasses in theory and/or data.

5 Profile Likelihood Methods

Profile Likelihood Methods

Using results due to Wilks and Wald, ATLAS colleagues G. Cowan, K. Cranmer, E. Gross, and O. Vitells derived asymptotic formulae for use in searches for new physics and based on the profile likelihood; see arXiv:1007.1727v2 [physics.data-an].

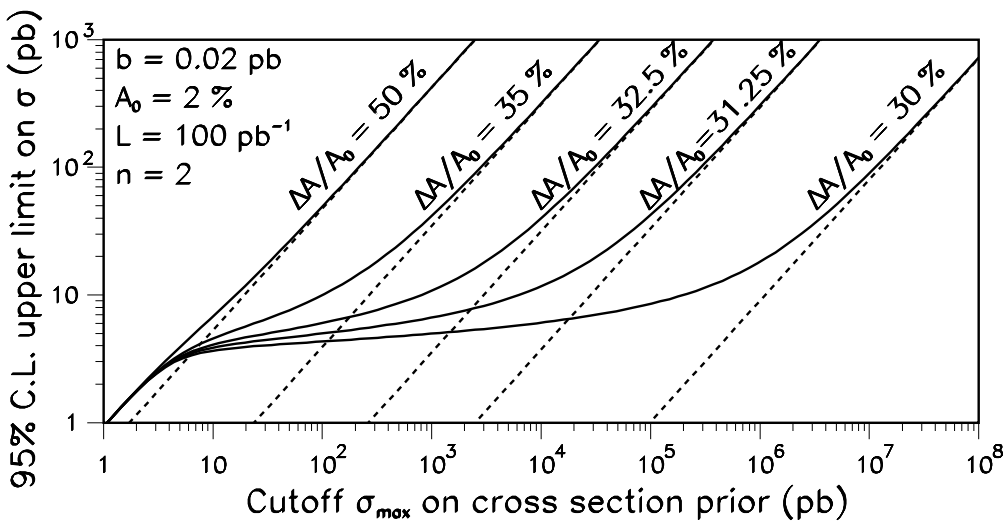
An interesting technique they developed is the so-called Asimov dataset, which is in a sense the most representative dataset of an ensemble: when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values. Asimov datasets can be used to simplify the estimation of measurement sensitivities and even Jeffreys' prior.

6 Reference Priors

Reference Priors (1)

In HEP uniform priors have been the norm for a long time, partly because they *seem* reasonable (principle of indifference), and partly because the corresponding posterior intervals sometimes exhibit reasonable frequentist behavior. However, they are also known to suffer from major drawbacks:

1. They give inconsistent results if the parametrization of the problem is changed;
2. They are not guaranteed to yield proper posteriors;
   the best known example is the measurement of a cross section σ from a Poisson variate n with mean $LA\sigma + b$, where L is a known integrated luminosity, b a known background source, and A an efficiency;
   the prior for σ is uniform, whereas that for A is truncated Gaussian with mean A_0 and standard deviation ΔA:



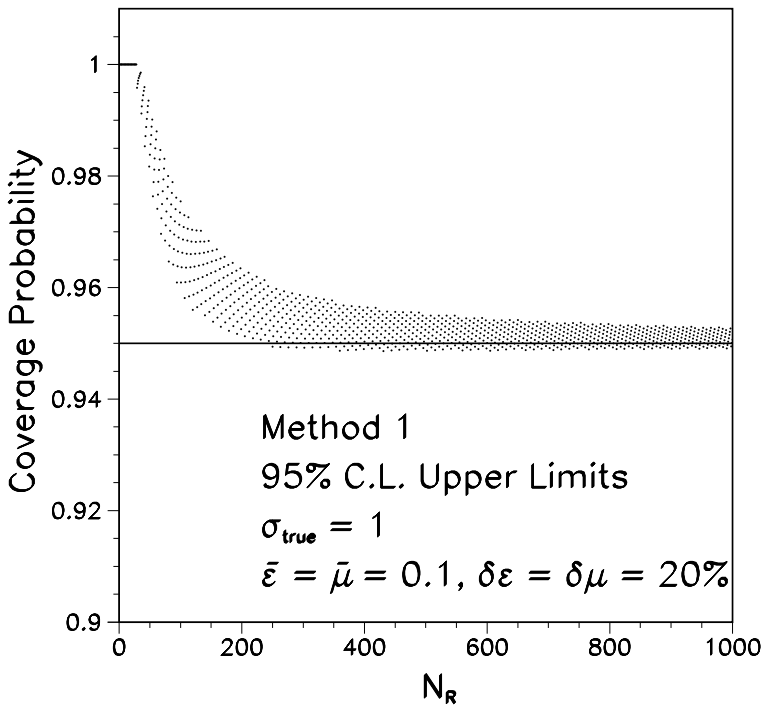
Reference Priors (2)

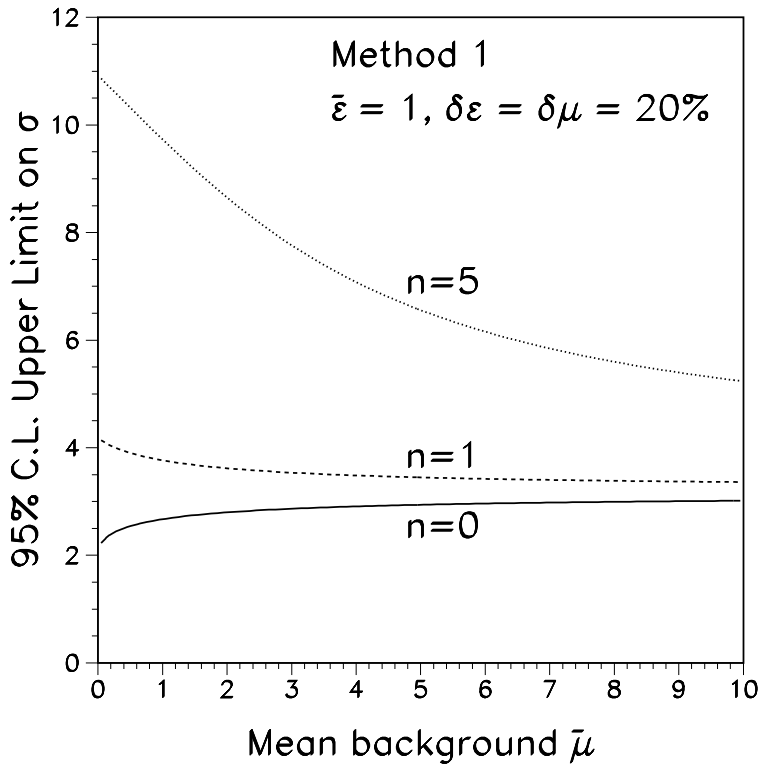
Reference priors have been developed over the past thirty years with the aim of providing a "standard" for presenting and comparing measurements of quantities about which little or no prior knowledge is available. Similarly to other standards (e.g. lengths and weights), the reference prior standard was designed with some rational considerations in mind:

- information-theory based: reference priors minimize the influence of the prior, relative to the model, on the final inference;
- generality: a well-defined algorithm exists to create a reference prior for almost any estimation problem, and the resulting posterior is proper;
- invariance: given a one-to-one map from a parameter θ to a parameter ϕ, applying the reference prior construction separately to θ and ϕ yields posteriors that are related by the correct transformation law, $\pi(\phi \mid x) = \pi(\theta \mid x)\,|\partial\theta/\partial\phi|$;
- sampling consistency: the posterior densities from an ensemble of experiments tend to cluster around the true values of the parameters;
- coherence: inferences derived from reference priors avoid marginalization paradoxes.

Reference Priors (3)

In HEP we now have reference priors for cross section measurements, when partial information is available for acceptances and background sources. The resulting inferences behave reasonably:





[L. D., H. Prosper, S. Jain, "Reference priors for high energy physics," Phys. Rev. D **82**, 034002 (2010).]

Reference Priors (4)

There still seem to be some important puzzles regarding reference priors:

1. What is the proper probabilistic interpretation of a reference posterior?
   - Reference posterior probabilities are not subjective probabilities! So what are they then?
   - Can reference posterior inferences be reported by themselves, or should they be reported only as part of a sensitivity analysis? If the latter, how should one choose alternative priors?
2. How should we deal with the compact set normalization procedure?
   - The general definition of reference priors involves the taking of limits, and this must be done carefully in order to avoid infinities; the standard approach is to use sequences of nested compact sets that converge to the whole parameter space.
   - Unfortunately there is no unique way of choosing these compact sets, and there is no guarantee that different choices lead to the same result, or even that all choices lead to a proper posterior.
   - This ambiguity prevents us from designing a completely general numerical algorithm.
3. How should we handle implicit statistical models?
   - Can we combine ABC methods with numerical algorithms for computing reference posteriors?

7 Extreme Value Theory

Extreme Value Theory

Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed random variables. Central limit theory is concerned with the behavior of the partial sums $X_1 + X_2 + \ldots + X_n$ as $n \to \infty$, whereas extreme value theory is concerned with the behavior of the sample extremes $\max\{X_1, X_2, \ldots, X_n\}$ as $n \to \infty$.

Extreme value theory has many applications, for example to the question of how high dikes should be built in the Netherlands to protect land below sea level from storm surges that drive the seawater level up along the coast.

In HEP we are often interested in extreme events, that is, collision events in which some measurable quantity takes on a very large value. Extreme value theory can help by providing a solid basis for extrapolating from measured distributions at lower values of the quantity of interest.

[See L. de Haan and A. Ferreira, "Extreme value theory: an introduction," Springer, 2006.]

8 Summary

Summary

- Progress has been made on several fronts: profile likelihood methods, Bayesian reference priors, look-elsewhere effect.
- Some issues seem too difficult to decide (maybe it's a cultural problem?): the 5σ threshold for discovery, accounting for measurement sensitivity.
- Other problems appear solvable but will require a lot of effort: parton density function uncertainties, implicit statistical models.
- We may not yet be exploiting everything statistics has to offer: extreme value theory.
- All in all, we do seem to be getting closer to the goal formulated by Bob Cousins, which is to be able to present the results of a measurement as analyzed via at least three paradigms: frequentist, likelihood, and Bayes.