Banff International Research Station

for Mathematical Innovation and Discovery

Statistical Issues Relevant to
Significance of Discovery Claims

Second Workshop, Third Summary

Luc Demortier

The Rockefeller University

Banff, July 16, 2010

1. In the Beginning There Were Questions. . .

The Quiz (1)

1. Frequentist testing: the standard test of H_0 versus H_1 has only two outcomes: Accept or Reject H_0. In HEP we typically have four outcomes: Reject H_0 and accept H_1, Accept H_0 and exclude H_1, Make no decision, and Reject both H_0 and H_1. What is the error structure of such a test?
2. Rejection threshold: traditionally set at 5σ, regardless of sample size, prior evidence, accounting for look-elsewhere effect (LEE), etc. Can we come up with a more rational, more flexible criterion?
3. Wilks's theorem: in cases where it does not apply (e.g. LEE), we have to run Monte Carlo toy experiments. What is the most efficient way of doing this?
4. Peak searches: how should we calculate the significance of a possible second peak in a spectrum?
5. Is there a LEE for exclusion? If at each Higgs mass we exclude Higgs production at the 95% C.L., what is the confidence level for excluding over the whole mass range?
6. Is there a look-elsewhen effect?

The Quiz (2)

7. In a counting experiment: as the observed count sinks below the expected background, what do we want the upper limit to do?
8. Parton Density Functions: they are hugely important in searches for new physics, and yet we still do not have a statistically grounded understanding of their uncertainties.
9. How can we convince high energy physicists to share their data?
10. How can two different experiments coherently combine their reference posteriors for the same parameter of interest?
11. Can statisticians help with prior selection?
12. What is the best way to work with likelihood functions that have a non-analytical dependence on some important nuisance parameters (e.g. the likelihood for the dijet invariant mass spectrum as a function of the jet energy scale nuisance parameter)?

2. . . . Then Came Some Answers.

The Jeffreys-Lindley Paradox

[Cousins]

Consider a model for some data X that depends on an unknown parameter θ, and suppose that we wish to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. It is entirely possible that for some observation x_{obs} the p-value calculated by a frequentist will lead to rejection of H_0, whereas the Bayes factor calculated by a Bayesian will strongly support H_0. Hence the paradox. Who can we blame?

- The frequentist, because she did not adjust her rejection criterion for the power of the test;
- The experimenter, because he chose to test a point-null hypothesis instead of an interval-null hypothesis;
- The objective Bayesian, because he used a vague prior for θ under H_1.

Only the subjective Bayesian is totally immune to this paradox! 

Bayes/non-Bayes: from Compromise to Cooperation?

[Berger, Cousins, Linnemann, Lockhart,. . .]

The statistician I.J. Good had an approach to statistics that he called "the Bayes/non-Bayes compromise". One example is his resolution of the Jeffreys-Lindley paradox by "renormalizing" p-values: $p \rightarrow \min\{p * \sqrt{N/100}, 1/2\}$. Over the years there have been remarkable examples where the "compromise" turns into "cooperation", in the sense that Bayesian and frequentist inferences agree exactly, thereby enriching the interpretation of measurement results.

Earliest example in HEP literature is probably the Cousins-Highland paper, where Bayesian marginalization is used to handle systematic uncertainties. Later, Giunti proposed a modification of Feldman-Cousins intervals in which the ordering rule incorporates a Bayesian posterior mean. This is still a fully frequentist method, but the behavior of intervals and upper limits as the data count sinks below background is improved.

This workshop offered several examples of this compromise/cooperation. . .

Bayes/non-Bayes: Hypothesis Testing with a Prior on the Alternatives

[Lockhart]

If one has a range of alternatives to test against, and there is prior information about “interesting targets”, then it makes sense to construct a prior density over the space of alternatives so as to optimize the power of the test. This does not affect the Type-I error rate of the test.

This can be used to design goodness-of-fit tests. There are also examples in the literature where this approach is used to develop *standard* tests for the look-elsewhere effect (i.e. tests with chisquared distributions under the null).

Bayes/non-Bayes: Prior-Predictive *p*-Values

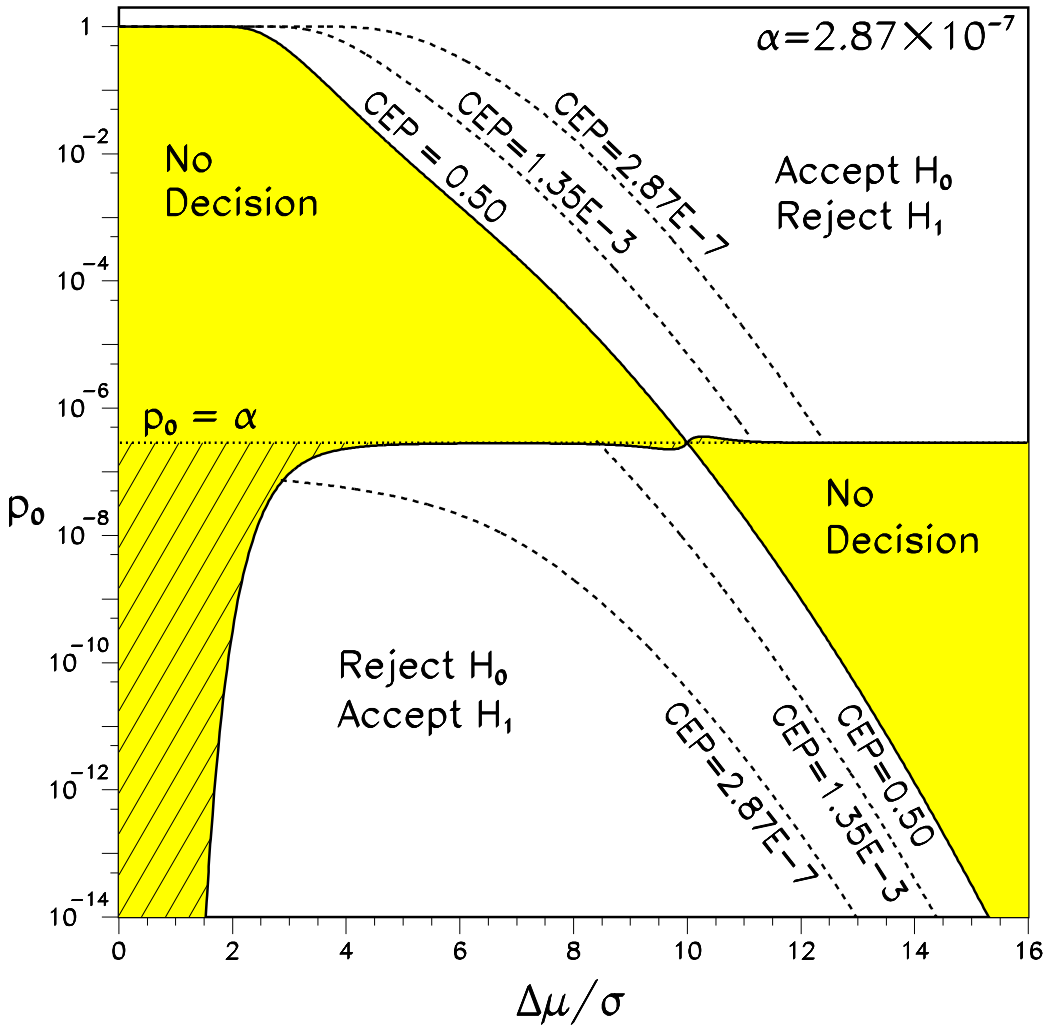
Linnemann discussed some prior-predictive *p*-values for the Poisson problem with unknown background:

- Z_Γ: uses a Γ prior on the Poisson mean. Numerically this gives exactly the same result as a purely frequentist calculation where information about the background comes from an auxiliary Poisson measurement. Therefore Z_Γ "covers". Perfect Bayes/non-Bayes cooperation!
- Z_N: uses a normal prior on the Poisson mean, and tends to undercover for very high significances.

The coverage calculations done here are with respect to an ensemble where only the data vary, and the unknown parameters (background magnitude) are fixed. Personally I have no problem in considering "extended" ensembles, where detector properties and other model parameters can vary. In the prior-predictive spirit, such extended ensembles represent our state of knowledge, not potentially true states of nature. The crucial point however, is to properly model the uncertainties on the nuisance parameters: Gaussian is usually wrong in HEP, Gamma is better, but sometimes we need something with even heavier tails, e.g. log-normal.

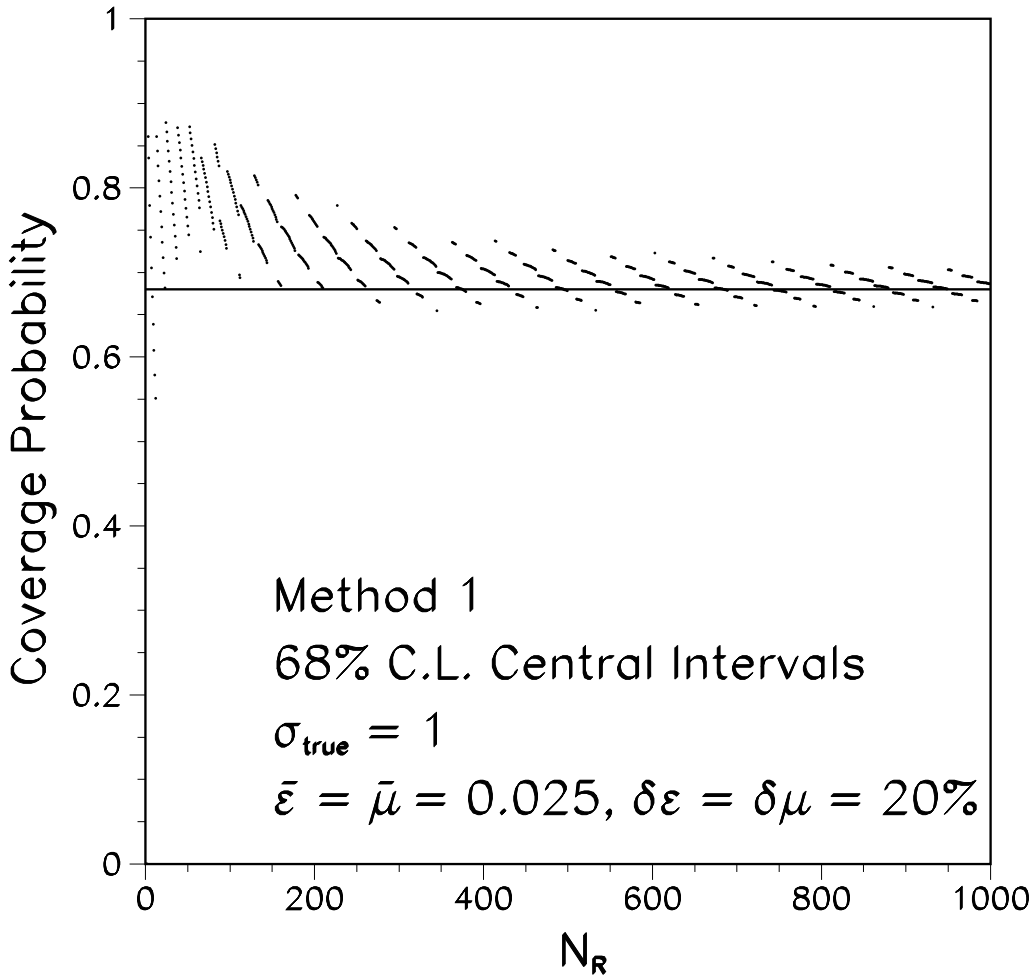
Bayes/non-Bayes: Hypothesis Testing

[Berger] Partition sample space according to the *evidential strength* Q of possible observations, where $Q \equiv \min\{(1-p_0)/(1-\alpha), (1-p_1)/(1-\beta)\}$. Then, frequentist error rates conditioned on Q agree exactly with Bayesian posterior probabilities of hypotheses.



Bayes/non-Bayes: Reference Priors

[Prosper] High energy physicists have now learned that indiscriminately used flat priors may lead to improper posteriors. A better alternative is provided by reference priors, which have reasonable frequency behavior:



The Sensitivity Problem

[Cowan, Cranmer, Murray, van Dyk. . .]

Bill M. gave a touching recantation of CL_S, but not quite for the best (imho) of reasons. . . He does not like the way the CL_S upper limit varies with the observation in a Gaussian measurement. Unfortunately the CL_S limit for this problem coincides exactly with the Bayesian limit (for a flat prior). You can't reject one without rejecting the other.

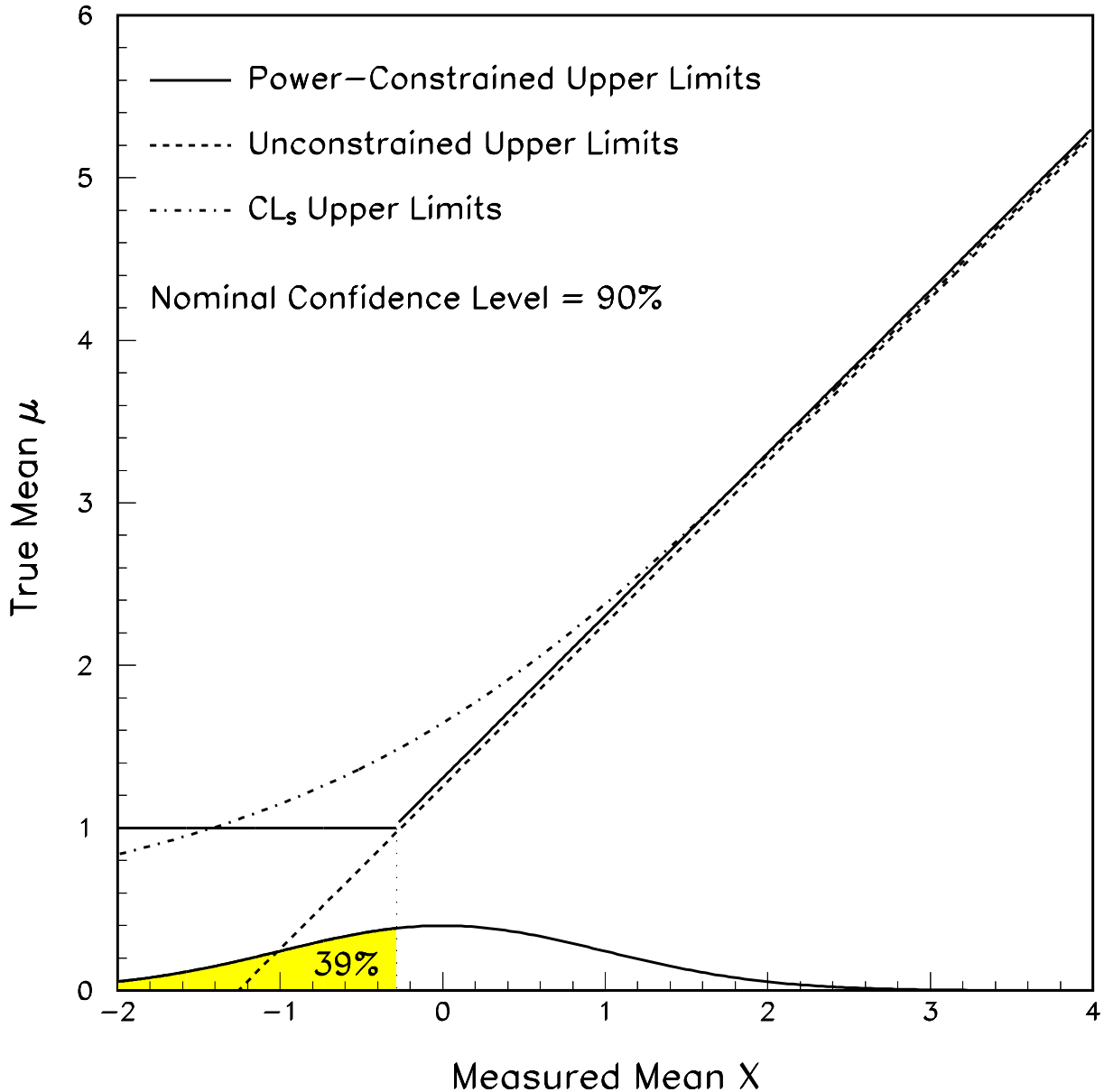
There are better reasons to abandon CL_S, foremost that it does not have a well-defined interpretation (not even as a conditional probability, as is sometimes claimed), and that it only guarantees a lower bound on the frequentist coverage.

Cowan, Cranmer, Gross proposed an alternative, called "power-constrained upper limits": can't reject a parameter value unless you have a decent probability of detecting that value when it is the true value. So if the observed upper limit is "too low", we reset it to a minimum.

Van Dyk proposes to always report both the observed upper limit *and* a minimum sensitivity bound (caveat: his terminology interchanges the concepts of upper limit and upper bound as understood in HEP).

The Sensitivity Problem (2)

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



The Look-Elsewhere Effect

[Berger, Bloom, Cranmer, Gross, Linnemann, Lyons, Vitells, Woodroofe,. . .]

In the statistics literature this effect shows up under various guises:

- Nuisance parameters that are “not identified under the null hypothesis”, or “only present under the alternative”;
- Multiple tests;
- Testing for the number of components in a mixture;
- Change-point detection;
- Level crossings of random fields;
- Singular information matrix.

Many solutions have been worked out over the years (lack-of-fit tests, Euler characteristics, Hotelling’s volume of tube formula, penalized likelihood, parameter transformations, directed graphs,. . .)

The Look-Elsewhere Effect (2)

Eilam showed an interesting result for the LEE-corrected tail probability in the case of a Gaussian signal with unknown mean, on top of a smooth background:

$$\mathbb{P}\left[t_{\text{float}} > t\right] \;\approx\; p_{\chi^2_2}\,\langle N\rangle, \qquad \text{when} \qquad p_{\chi^2_2}\,\langle N\rangle \ll 1. \tag{1}$$

The derivation is somewhat heuristic. In fact, this problem has been studied many times in the statistics literature, see for example R. B. Davies, "Hypothesis testing when a nuisance parameter is present only under the alternative," Biometrika **74**, 33 (1987). Suppose that for each value of $\theta \in [L, U]$, your test statistic is (asymptotically) chisquared with s degrees of freedom. Davies derives the following formula for the LEE-corrected tail:

$$\mathbb{P}\left[\sup_{L\le\theta\le U} S(\theta) \;>\; u\right] \le \mathbb{P}(\chi^2_s > u) \;+\; \int_L^U \psi(\theta)\, d\theta, \tag{2}$$

where

$$\psi(\theta) \;=\; K_s\, \frac{u^{(s-1)/2}\, e^{-u/2}}{2^{s/2}\,\sqrt{\pi}\,\Gamma((s+1)/2)}. \tag{3}$$

The Look-Elsewhere Effect (3)

For one degree of freedom ($s = 1$), $S(\theta) = Z^2(\theta)$, where Z is standard normal, and Davies's formula reduces to:

$$\mathbb{P}\left[\sup_{L\le\theta\le U} S(\theta) > u\right] \le \mathbb{P}(\chi^2_1 > u) + K_1\,\mathbb{P}(\chi^2_2 > u). \tag{4}$$

This is a linear combination of two χ^2 tails (only turns out this way for $s = 1$!). The coefficient of the χ^2_2 tail is

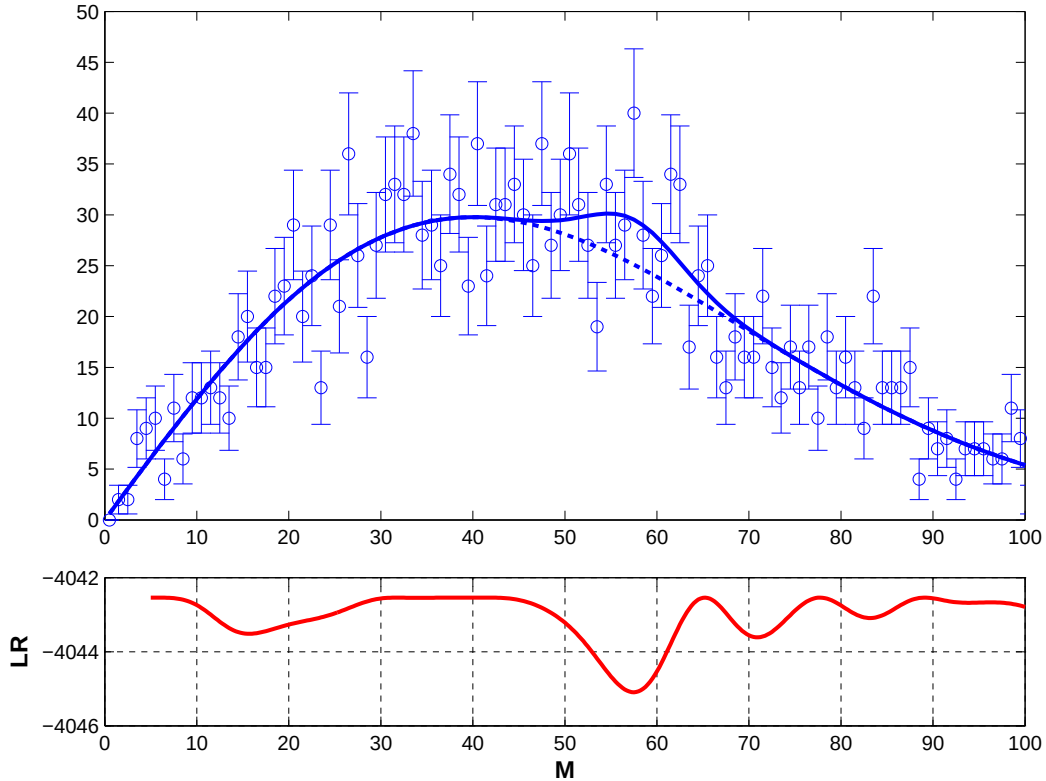
$$K_1 = \frac{1}{\pi}\int_L^U \sqrt{\mathrm{var}\left[\frac{\partial Z}{\partial\theta}\right]}\,d\theta. \tag{5}$$

Davies worked out an approximation for this:

$$\sqrt{2\pi}\,K_1 \approx \int_L^U \left|\frac{\partial S^{1/2}(\theta)}{\partial\theta}\right| d\theta = |S^{1/2}(\theta_1) - S^{1/2}(L)| + |S^{1/2}(\theta_2) - S^{1/2}(\theta_1)| + \ldots + |S^{1/2}(U) - S^{1/2}(\theta_m)|, \tag{6}$$

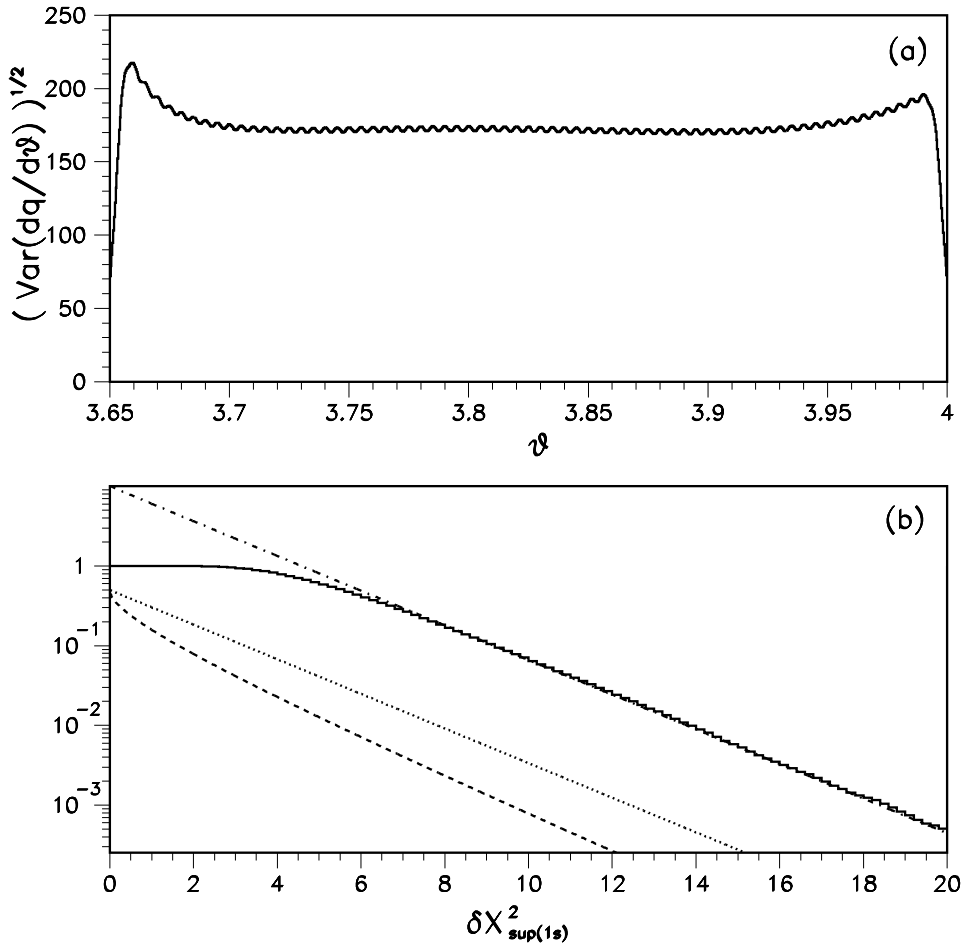
where $\theta_1, \ldots, \theta_m$ are the turning points of $S^{1/2}(\theta)$. So K_1 is proportional to the total variation of a Gaussian random field, which is related to the number of local maxima in Eilam and Ofer's formula.

The Look-Elsewhere Effect (3)



Background-only experiment example from arXiv:1005:1891v1 [physics.data-an]. Bottom panel shows local minima and maxima of inverse likelihood ratio.

The Look-Elsewhere Effect (4)



Top: integrand of the quantity $K(\theta)$ of equation (5), as a function of θ. Bottom: survivor function of sup S: toy experiments (solid), Davies formula (dot-dashes), $\chi^2_1/2$ (dashes), and $\chi^2_2/2$ (dots) [Demortier 2006].

The Look-Elsewhere Effect (5)

Note that in the Davies formula:

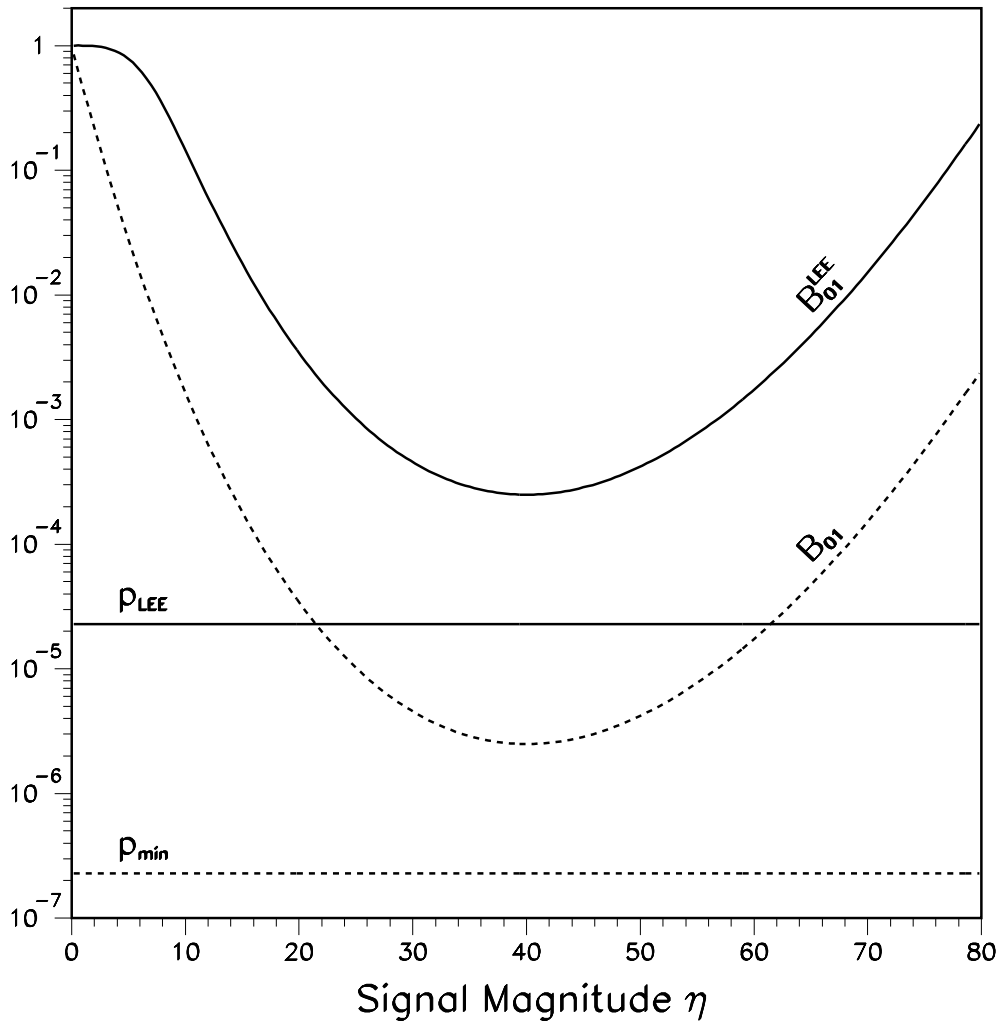
$$\mathbb{P}\left[\sup_{L\leq\theta\leq U} S(\theta) > u\right] \leq \mathbb{P}(\chi^2_1 > u) + K_1\,\mathbb{P}(\chi^2_2 > u), \tag{7}$$

the quantity $\mathbb{P}(\chi^2_1 > u)$ is the "fixed mass" significance, i.e. the significance without accounting for the LEE. For the ratio of the floating- to the fixed-mass significances, Eilam and Ofer obtain:

$$\text{trial\#}_{\text{observed}} = \langle N\rangle \sqrt{\frac{\pi}{2}}\, Z_{\text{fix}}. \tag{8}$$

However, this expression of the trial factor in terms of Z_{fix} is an artefact of mathematical identities that take place in the case $s = 1$; it does not generalize to $s \neq 1$ and should therefore not be taken too seriously.

The Look-Elsewhere Effect (6)



Solid line: Bayes factor for a peak search, as a function of peak height, when the peak location is unknown. This is the Bayes factor in favor of the background-only hypothesis. Its denominator is averaged over a uniform prior for the signal location.

The Look-Elsewhere Effect (7)

Some additional issues on this topic:

- Linnemann presented an interesting graphical method to prioritize subsets in the space of alternative hypotheses, in such a way that the LEE causes less significance dilution for the alternatives we care most about. However it is not clear how high energy physicists would set up such a graph for the large number of complex models they have to contend with.
- Lyons wondered about the "Look-ElseWhen Effect", and claimed it doesn't matter because "relevant data is (essentially) all data up to the present". Strictly speaking, if this is how we proceed, then we should adjust α for the sample size.

Other Intriguing Topics

1. Asimov data [Cowan, Cranmer]: can we avoid time-consuming expectation-value calculations by using only one, cleverly chosen datapoint? Although there is some evidence that this can be done, more understanding is needed. Cranmer suggested that we can use Asimov data to compute Jeffreys' prior. This may work for exponential models, but how general is it?
2. Profile likelihood versus marginal likelihood [Loredo]: helps us understand when the former is reliable and when it is not.
3. Banff Challenge 2 [Fisher, Junk, Scargle, Schafer, Vitells]: some interesting ideas to analyze these datasets were presented. However we will probably need ensembles in order to study the effectiveness of the methods proposed.

The End

This was a workshop rich in interesting and useful ideas, and I couldn't do justice to all the beautiful talks that were given. Thank you all for your contributions!