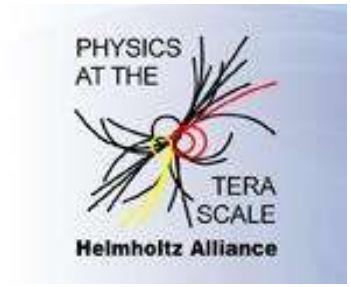
Bayesian statistical methods for HEP

Terascale Statistics School

DESY, Hamburg

1 October, 2008



Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

Tuesday:

The Bayesian method

Bayesian assessment of uncertainties

Bayesian computation: MCMC

Wednesday:

Bayesian limits

Bayesian model selection ("discovery")

Outlook for Bayesian methods in HEP

The Bayesian approach to limits

In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta')\,d\theta'} \propto L(x|\theta)\pi(\theta)$$

Integrate posterior pdf $p(\theta\,|\,x)$ to give interval with any desired probability content.

For e.g. Poisson parameter 95% CL upper limit from

$$0.95 = \int_{-\infty}^{s_{\mathrm{up}}} p(s|n)\,ds$$

Analytic formulae for limits

There are a number of papers describing Bayesian limits for a variety of standard scenarios

- Several conventional priors
- Systematics on efficiency, background
- Combination of channels

and (semi-)analytic formulae and software are provided.

Joel Heinrich, *Bayesian limit software: multi-channel with correlated backgrounds and efficiencies*, CDF/MEMO/STATISTICS/PUBLIC/7587 (2005).

Joel Heinrich et al., *Interval estimation in the presence of nuisance parameters. 1. Bayesian approach*, CDF/MEMO/STATISTICS/PUBLIC/7117, physics/0409129 (2004).

Luc Demortier, *A Fully Bayesian Computation of Upper Limits for Poisson Processes*, CDF/MEMO/STATISTICS/PUBLIC/5928 (2004).

But for more general cases we need to use numerical methods (e.g. L.D. uses importance sampling).

Example: Poisson data with background

Count n events, e.g., in fixed time or integrated luminosity.

s = expected number of signal events

b = expected number of background events

$n \sim$ Poisson(s+b): $$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Sometimes b known, other times it is in some way uncertain.

Goal: measure or place limits on s, taking into consideration the uncertainty in b.

Widely discussed in HEP community, see e.g. proceedings of PHYSTAT meetings, Durham, Fermilab, CERN workshops...

Bayesian prior for Poisson parameter

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s<0$.

Often try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as $L(s)$ dies off for large s.

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true s).

Bayesian interval with flat prior for s

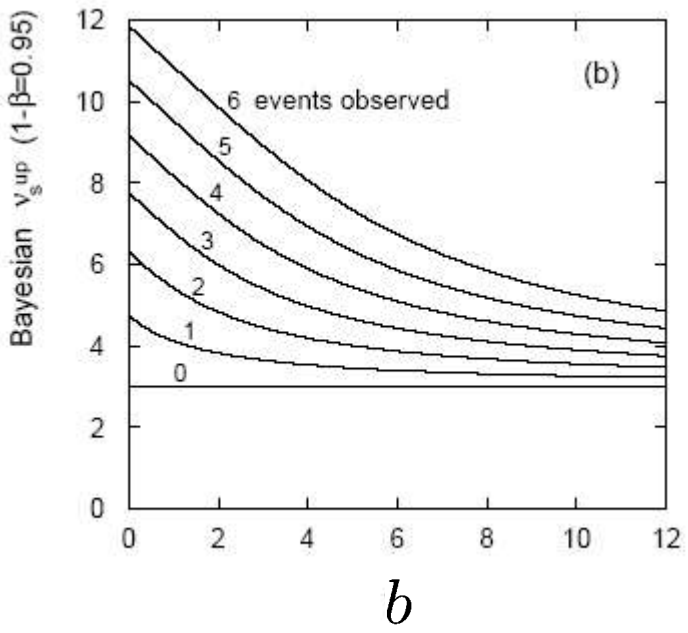
Solve numerically to find limit s_{up}.

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as classical case ('coincidence').

Otherwise Bayesian limit is everywhere greater than classical ('conservative').

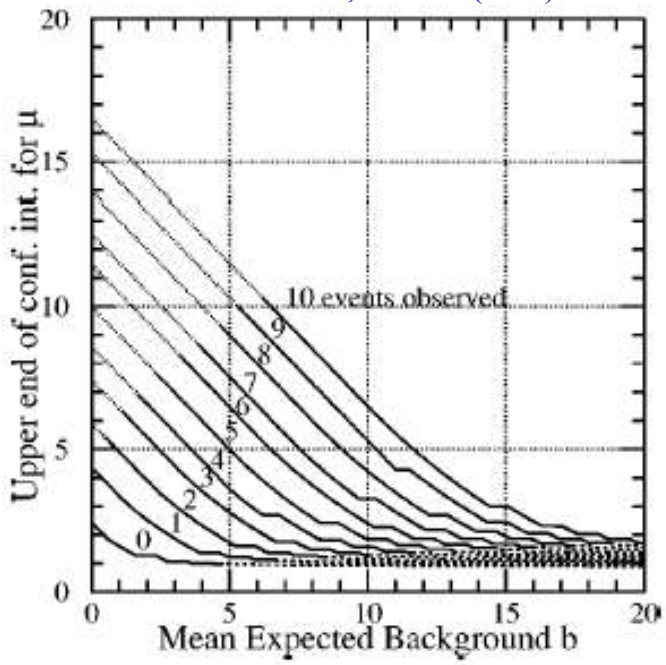
Never goes negative.

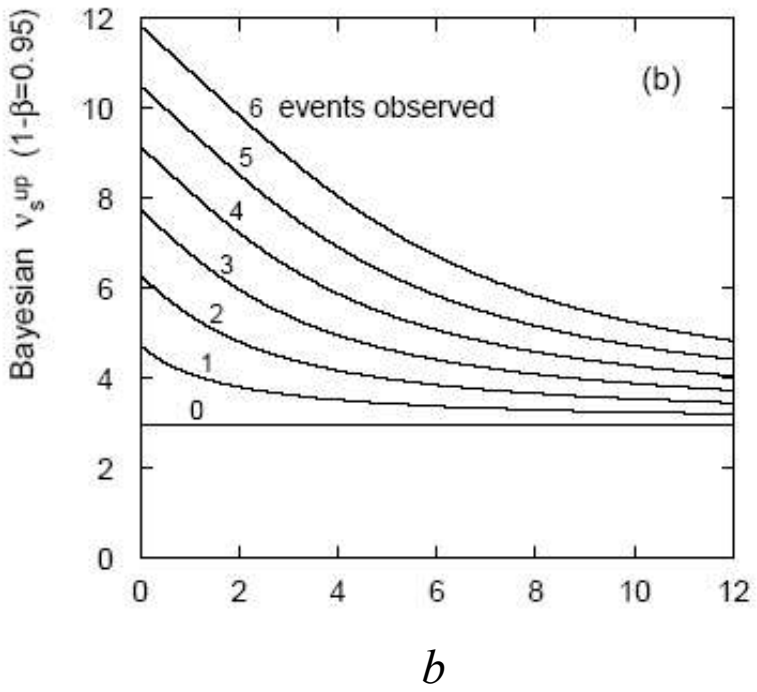
Doesn't depend on b if $n = 0$.



Upper limit versus *b*

Feldman & Cousins, PRD 57 (1998) 3873





If $n = 0$ observed, should upper limit depend on b?

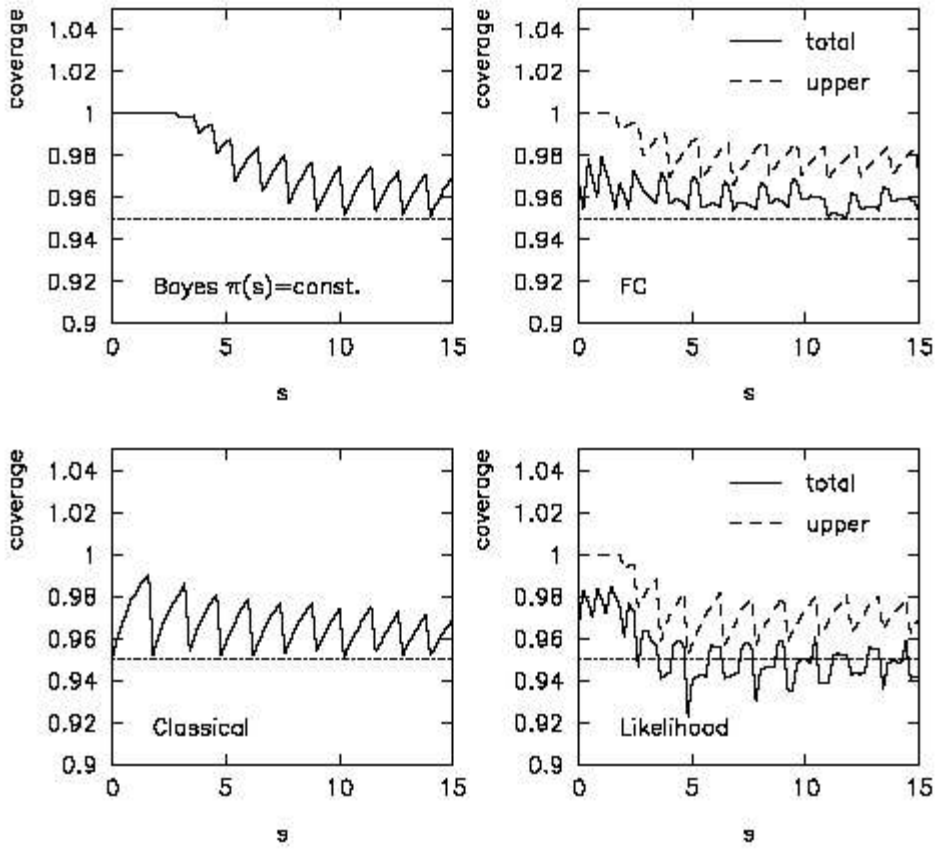
Classical: yes

Bayesian: no

FC: yes

Coverage probability of confidence intervals

Because of discreteness of Poisson data, probability for interval to include true value in general > confidence level ('over-coverage')



Bayesian limits with uncertainty on b

Uncertainty on b goes into the prior, e.g.,

$$\pi(s,b) = \pi_s(s)\pi_b(b) \quad \text{(or include correlations as appropriate)}$$

$$\pi_s(s) = \text{const}, \quad \sim 1/s, \ldots$$

$$\pi_b(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad \text{(or whatever)}$$

Put this into Bayes' theorem,

$$p(s,b|n) \propto L(n|s,b)\pi(s,b)$$

Marginalize over b, then use $p(s|n)$ to find intervals for s with any desired probability content.

Controversial part here is prior for signal $\pi_s(s)$
(treatment of nuisance parameters is easy).

Discussion on limits

Different sorts of limits answer different questions.

A frequentist confidence interval does not (necessarily) answer, "What do we believe the parameter's value is?"

Coverage — nice, but crucial?

Look at sensitivity, e.g., $E[s_{\text{up}} \mid s = 0]$.

Consider also:

politics, need for consensus/conventions;
convenience and ability to combine results, ...

For any result, consumer will compute (mentally or otherwise):

$$p(\theta|\text{result}) \propto L(\text{result}|\theta)\pi(\theta)$$

Need likelihood (or summary thereof).

consumer's prior

Frequentist discovery, *p*-values

To discover e.g. the Higgs, try to reject the background-only (null) hypothesis (H_0).

Define a statistic t whose value reflects compatibility of data with H_0.

p-value = Prob(data with $\leq$ compatibility with H_0 when compared to the data we got | H_0)

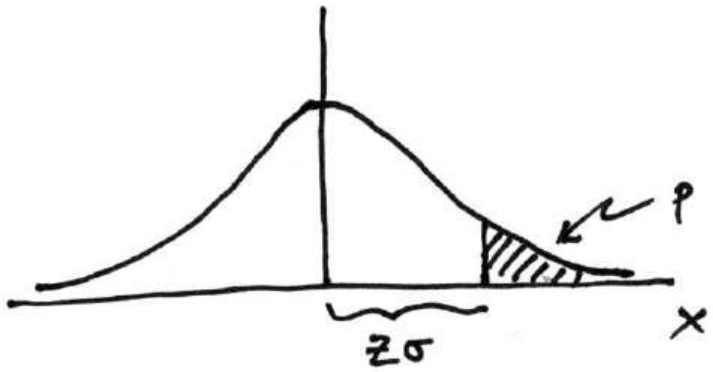
For example, if high values of t mean less compatibility,

$$p = \int_t^\infty f(t'|H_0)\, dt' \ .$$

If p-value comes out small, then this is evidence against the background-only hypothesis → discovery made!

Significance from p-value

Define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ `TMath::Prob`

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

When to publish

HEP folklore is to claim discovery when $p = 2.85 \times 10^{-7}$, corresponding to a significance $Z = 5$.

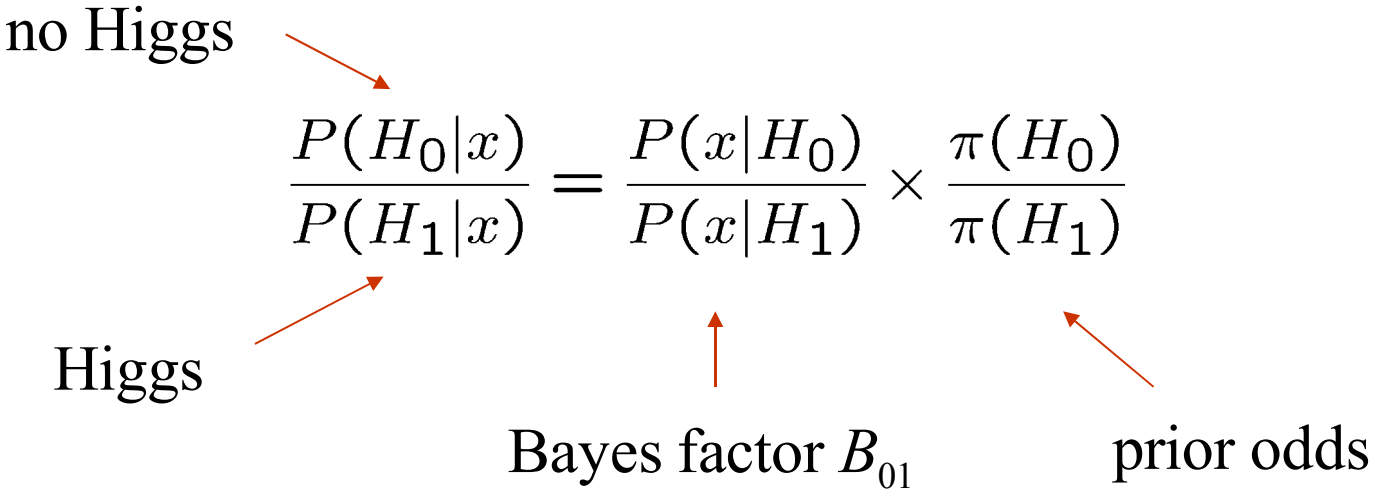
This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

phenomenon	reasonable p-value for discovery
$D^0\bar{D}^0$ mixing	~0.05
Higgs	~ 10^{-7} (?)
Life on Mars	~10^{-10}
Astrology	~10^{-20}

Note some groups have defined 5σ to refer to a two-sided fluctuation, i.e., $p = 5.7 \times 10^{-7}$

Bayesian model selection ('discovery')

The probability of hypothesis H_0 relative to its complementary alternative H_1 is often given by the posterior odds:



The Bayes factor is regarded as measuring the weight of evidence of the data in support of H_0 over H_1.

Interchangeably use $B_{10} = 1/B_{01}$

Assessing Bayes factors

One can use the Bayes factor much like a p-value (or Z value).

There is an "established" scale, analogous to our 5σ rule:

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Kass and Raftery, *Bayes Factors*, J. Am Stat. Assoc 90 (1995) 773.

11 May 07: Not clear how useful this scale is for HEP.
3 Sept 07: Upon reflection & PHYSTAT07 discussion, seems like an intuitively useful complement to p-value.

Rewriting the Bayes factor

Suppose we have models H_i, $i = 0, 1, \ldots$,

each with a likelihood $p(x|H_i, \vec{\theta}_i)$

and a prior pdf for its internal parameters $\pi_i(\vec{\theta}_i)$

so that the full prior is $\pi(H_i, \vec{\theta}_i) = p_i \pi_i(\vec{\theta}_i)$

where $p_i = P(H_i)$ is the overall prior probability for H_i.

The Bayes factor comparing H_i and H_j can be written

$$B_{ij} = \frac{P(H_i|\vec{x})}{P(H_i)} \Big/ \frac{P(H_j|\vec{x})}{P(H_j)}$$

Bayes factors independent of $P(H_i)$

For B_{ij} we need the posterior probabilities marginalized over all of the internal parameters of the models:

$$P(H_i|\vec{x}) = \int P(H_i, \vec{\theta}_i|\vec{x})\, d\vec{\theta}_i$$

Use Bayes theorem

$$= \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) p_i \pi_i(\vec{\theta}_i)\, d\vec{\theta}_i}{P(x)}$$

So therefore the Bayes factor is

Ratio of marginal likelihoods

$$B_{ij} = \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) \pi_i(\vec{\theta}_i)\, d\vec{\theta}_i}{\int L(\vec{x}|H_j, \vec{\theta}_j) \pi_j(\vec{\theta}_j)\, d\vec{\theta}_j}$$

The prior probabilities $p_i = P(H_i)$ cancel.

Numerical determination of Bayes factors

Both numerator and denominator of B_{ij} are of the form

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})\, d\vec{\theta}$$ ← 'marginal likelihood'

Various ways to compute these, e.g., using sampling of the posterior pdf (which we can do with MCMC).

Harmonic Mean (and improvements)
Importance sampling
Parallel tempering (~thermodynamic integration)
...

See e.g. Kass and Raftery, *Bayes Factors*, J. Am. Stat. Assoc. 90 (1995) 773-795.

Harmonic mean estimator

E.g., consider only one model and write Bayes theorem as:

$$\frac{\pi(\boldsymbol{\theta})}{m} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})}$$

$\pi(\boldsymbol{\theta})$ is normalized to unity so integrate both sides,

posterior expectation

$$m^{-1} = \int \frac{1}{L(\mathbf{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p[1/L]$$

Therefore sample $\boldsymbol{\theta}$ from the posterior via MCMC and estimate m with one over the average of $1/L$ (the harmonic mean of L).

M.A. Newton and A.E. Raftery, *Approximate Bayesian Inference by the Weighted Likelihood Bootstrap*, Journal of the Royal Statistical Society B 56 (1994) 3-48.

Improvements to harmonic mean estimator

The harmonic mean estimator is numerically very unstable; formally infinite variance (!). Gelfand & Dey propose variant:

Rearrange Bayes thm; multiply both sides by arbitrary pdf $f(\boldsymbol{\theta})$:

$$\frac{f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{f(\boldsymbol{\theta})}{m}$$

Integrate over $\boldsymbol{\theta}$:

$$m^{-1} = \int \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x}) = E_p\left[\frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right]$$

Improved convergence if tails of $f(\boldsymbol{\theta})$ fall off faster than $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

Note harmonic mean estimator is special case $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$.

A.E. Gelfand and D.K. Dey, *Bayesian model choice: asymptotics and exact calculations*, Journal of the Royal Statistical Society B 56 (1994) 501-514.

Importance sampling

Need pdf $f(\boldsymbol{\theta})$ which we can evaluate at arbitrary $\boldsymbol{\theta}$ and also sample with MC.

The marginal likelihood can be written

$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E_f\left[\frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})}\right]$$

Best convergence when $f(\boldsymbol{\theta})$ approximates shape of $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Use for $f(\boldsymbol{\theta})$ e.g. multivariate Gaussian with mean and covariance estimated from posterior (e.g. with MINUIT).

Bayes factor computation discussion

Also can use method of parallel tempering; see e.g.

Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press, 2005.

Harmonic mean OK for very rough estimate.

I had trouble with all of the methods based on posterior sampling.

Importance sampling worked best, but may not scale well to higher dimensions.

Lots of discussion of this problem in the literature, e.g.,

Cong Han and Bradley Carlin, *Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review*, J. Am. Stat. Assoc. 96 (2001) 1122-1132.

Bayesian Higgs analysis

N independent channels, count n_i events in search regions:

$$P(\mathbf{n}|\mathbf{s},\mathbf{b}) = \prod_{i=1}^{N} \frac{(s_i+b_i)^{n_i}}{n_i!} e^{-(s_i+b_i)}$$

Constrain expected background b_i with sideband measurements:

$$P(\mathbf{m}|\mathbf{b},\boldsymbol{\tau}) = \prod_{i=1}^{N} \frac{(\tau_i b_i)^{m_i}}{m_i!} e^{-\tau_i b_i}$$

Expected number of signal events: $s_i = \mu\sigma_{\mathrm{SM}}\mathcal{B}_i\varepsilon_{s,i}L_i \equiv \mu\varphi_i$

(μ is global parameter, $\mu = 1$ for SM).

Consider a fixed Higgs mass and assume SM branching ratios B_i.

Suggested method: constrain μ with limit μ_{up}; consider m_{H} excluded if upper limit $\mu_{\mathrm{up}} < 1.0$.

For discovery, compute Bayes factor for $H_0 : \mu = 0$ vs. $H_1 : \mu = 1$

Parameters of Higgs analysis

E.g. combine cross section, branching ratio, luminosity, efficiency into a single factor ϕ:

$$s_i = \mu\sigma_{\rm SM}\mathcal{B}_i\varepsilon_{s,i}L_i \equiv \mu\varphi_i$$

Systematics in any of the factors can be described by a prior for ϕ, use e.g. Gamma distribution. For now ignore correlations, but these would be present e.g. for luminosity error:

$$\pi_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}) = \prod_{i=1}^{N} \frac{a_i(a_i\varphi_i)^{b_i-1}e^{-a_i\varphi_i}}{\Gamma(b_i)}$$

a_i, b_i from nominal value $\phi_{i,0}$ and relative error $r_i = \sigma_{\phi,i} / \phi_{i,0}$:

$$a = \frac{1}{\varphi_0 r_\varphi^2}, \quad b = \frac{1}{r_\varphi^2}.$$

Bayes factors for Higgs analysis

The Bayes factor B_{10} is

$$B_{10} = \frac{\int\int\int L(\mathbf{n},\mathbf{m}|\mu,\mathbf{b},\boldsymbol{\varphi})\,\pi_\mu(\mu)\pi_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\pi_{\mathbf{b}}(\mathbf{b})\,d\mu\,d\boldsymbol{\varphi}\,d\mathbf{b}}{\int\int L(\mathbf{n},\mathbf{m}|\mu=0,\mathbf{b},\boldsymbol{\varphi})\,\pi_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\pi_{\mathbf{b}}(\mathbf{b})\,d\boldsymbol{\varphi}\,d\mathbf{b}}$$

Compute this using a fixed μ for H_1, i.e., $\pi_\mu(\mu) = \delta(\mu - \mu')$,

then do this as a function of μ'. Look in particular at $\mu = 1$.

Take numbers from VBF paper for 10 fb^{-1}, m_H = 130 GeV:

Channel	s	b
$H \to WW^* \to e\mu + X$	12.3	9.2
$H \to WW^* \to ee/\mu\mu + X$	11.7	10.1
$H \to WW^* \to l\nu jj + X$	1.5	2.0

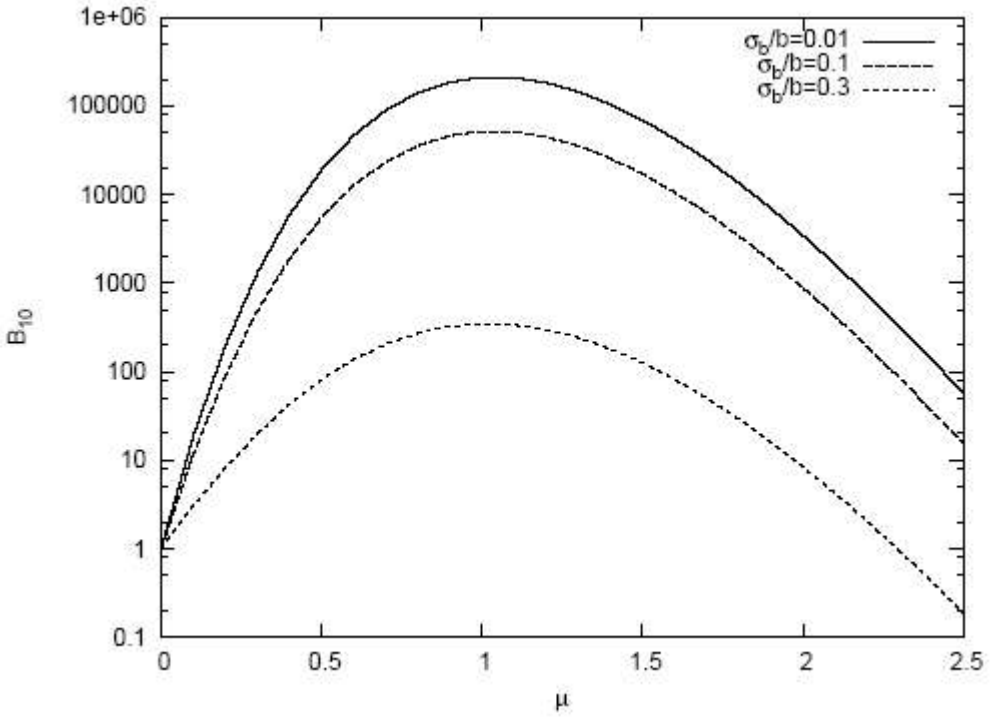
$l\nu jj$ was for 30 fb^{-1}, in paper; divided by 3

S. Asai et al., *Prospects for the Search for a Standard Model Higgs Boson in ATLAS using Vector Boson Fusion*, Eur. Phys. J. C32S2 (2004) 19-54; hep-ph/0402254.

Bayes factors for Higgs analysis: results (1)

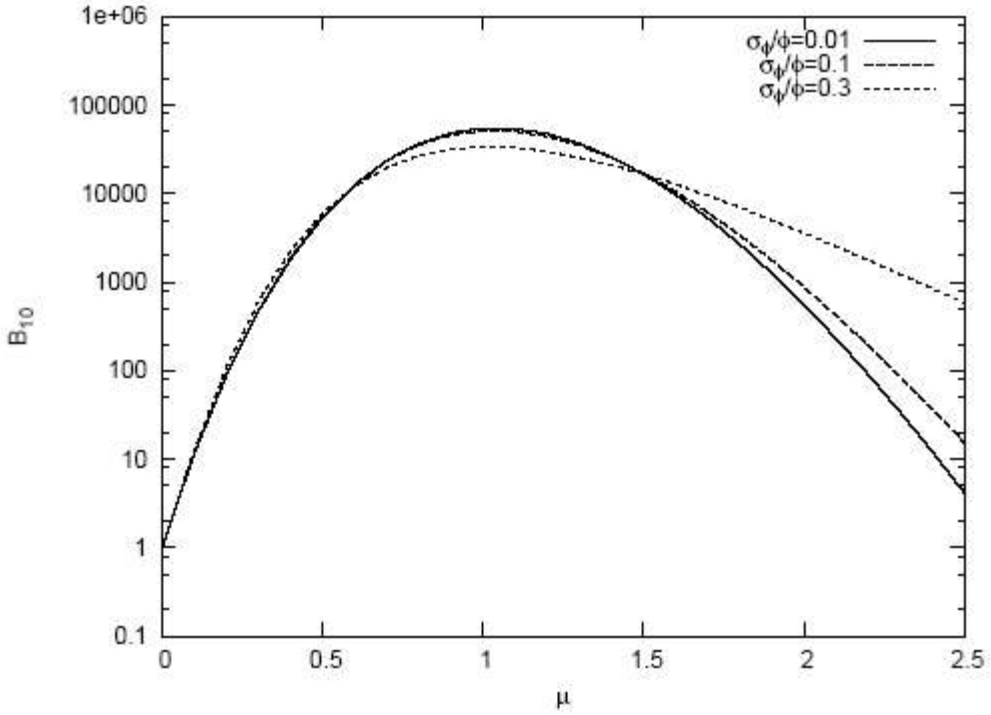
Create data set by hand with $n_i \sim$ nearest integer $(\phi_i + b_i)$, i.e., $\mu = 1$: $n_1 = 22$, $n_2 = 22$, $n_3 = 4$.

For the sideband measurements m_i, choose desired σ_b/b, use this to set size of sideband (i.e. $\sigma_b/b = 0.1 \rightarrow m = 100$).



B_{10} for $\sigma_\phi/\phi = 0.1$, different values of σ_b/b., as a function of μ.

Bayes factors for Higgs analysis: results (2)



B_{10} for $\sigma_b/b = 0.1$,
different values of σ_ϕ/ϕ,
as a function of μ.

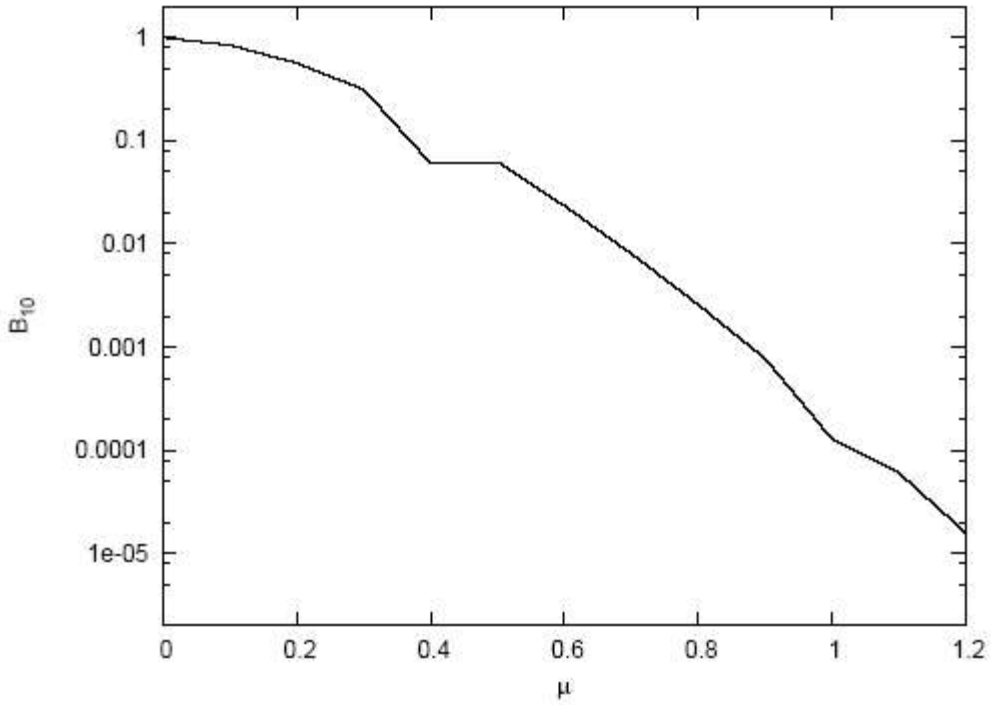
Effect of uncertainty in ϕ_i (e.g., in the efficiency):
$\mu = 1$ no longer gives a fixed s_i, but a smeared out distribution.

→ lower peak value of B_{10}.

Bayes factors for Higgs analysis: results (3)

Or try data set with $n_i \sim$ nearest integer b_i, i.e., $\mu = 0$:

$n_1 = 9$, $n_2 = 10$, $n_3 = 2$. Used $\sigma_b/b = 0.1$, $\sigma_\phi/\phi, = 0.1$.



Here the SM $\mu = 1$
is clearly disfavoured,
so we set a limit on μ.

Posterior pdf for μ , upper limits (1)

Here done with (improper) uniform prior, $\mu > 0$.
(Can/should also vary prior.)

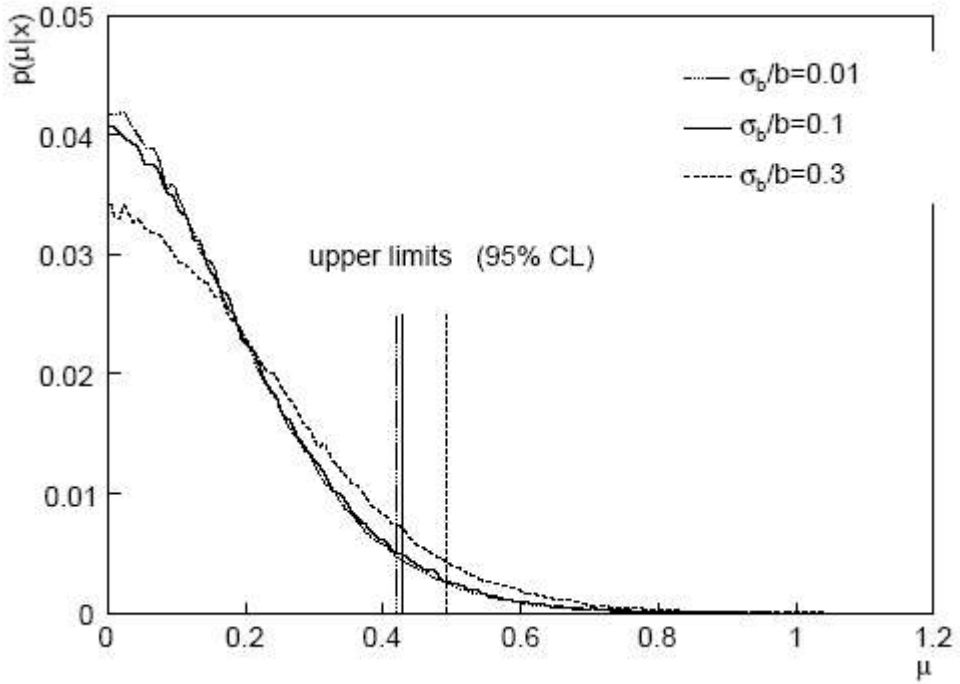


Figure 4: The posterior distribution of μ with a data set compatible with background only ($n_1 = 9$, $n_2 = 10$, $n_3 = 2$) for $\sigma_\varphi/\varphi = 0.1$ and several values of σ_b/b.

Posterior pdf for μ , upper limits (2)

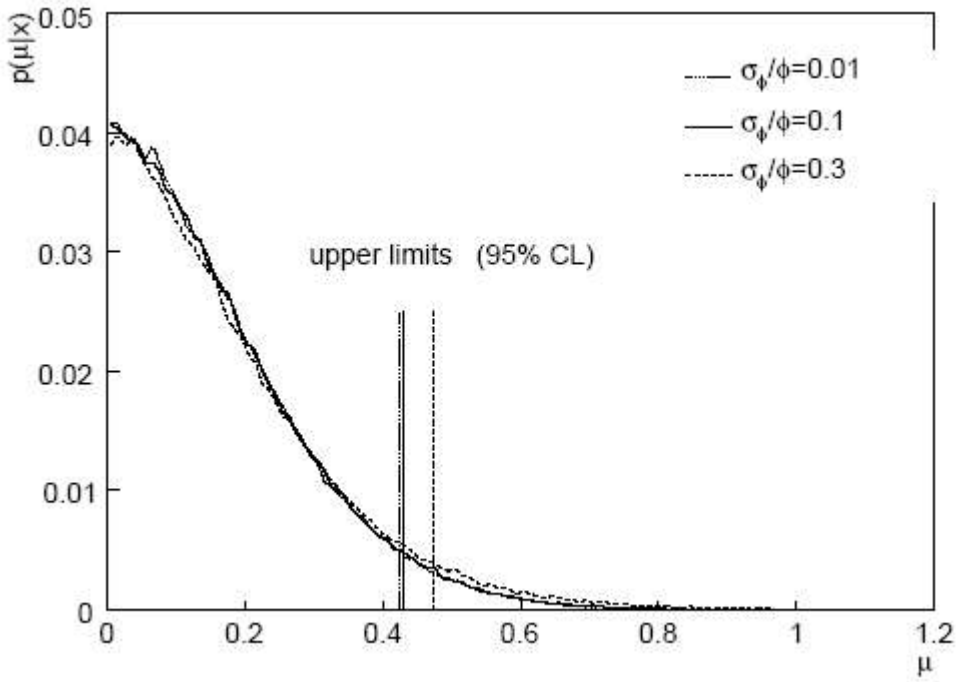


Figure 5: The posterior distribution of μ with a data set compatible with background only ($n_1 = 9$, $n_2 = 10$, $n_3 = 2$) for $\sigma_b/b = 0.1$ and several values of σ_φ/φ.

Outlook for Bayesian methods in HEP

Bayesian methods allow (indeed require) prior information about the parameters being fitted.

This type of prior information can be difficult to incorporate into a frequentist analysis

This will be particularly relevant when estimating uncertainties on predictions of LHC observables that may stem from theoretical uncertainties, parton densities based on inconsistent data, etc.

Prior ignorance is not well defined. If that's what you've got, don't expect Bayesian methods to provide a unique solution.

Try a reasonable variation of priors -- if that yields large variations in the posterior, you don't have much information coming in from the data.

You do not have to be exclusively a Bayesian or a Frequentist

Use the right tool for the right job