Bayesian Statistics at the LHC (and elsewhere)

Cambridge HEP Seminar

7 March, 2008

UNIVERSITY OF CAMBRIDGE

Royal Holloway
University of London

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

Outline

0 Why worry about this?

1 The Bayesian method

2 Bayesian assessment of uncertainties

3 Bayesian model selection ("discovery")

4 Outlook for Bayesian methods in HEP

5 Bayesian limits

Extra slides

Statistical data analysis at the LHC

High stakes

"5 sigma"

"4 sigma"

and expensive experiments, so we should make sure the data analysis doesn't waste information.

Specific challenges for LHC analyses include

Huge data volume

Generally cannot trust MC prediction of backgrounds; need to use data (control samples, sidebands...)

Lots of theory uncertainties, e.g., parton densities

People looking in many places ("look-elsewhere effect")

Dealing with uncertainty

In particle physics there are various elements of uncertainty:

theory is not deterministic

quantum mechanics

random measurement errors

present even without quantum effects

things we could know in principle but don’t

e.g. from limitations of cost, time, ...

We can quantify the uncertainty using PROBABILITY

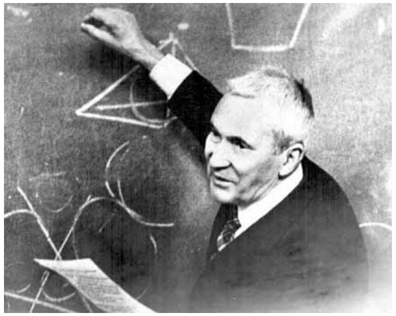
A definition of probability

Consider a set *S* with subsets *A*, *B*, ...

$$\text{For all } A \subset S, P(A) \geq 0$$

$$P(S) = 1$$

$$\text{If } A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$$



Kolmogorov axioms (1933)

Also define conditional probability:

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

Interpretation of probability

I. Relative frequency

A, B, ... are outcomes of a repeatable experiment

$$P(A) = \lim_{n \to \infty} \frac{\text{times outcome is } A}{n}$$

cf. quantum mechanics, particle scattering, radioactive decay...

II. Subjective probability

A, B, ... are hypotheses (statements that are true or false)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

- Both interpretations consistent with Kolmogorov axioms.
- In particle physics frequency interpretation often most useful, but subjective probability can provide more natural treatment of non-repeatable phenomena:

  systematic uncertainties, probability that Higgs boson exists,...

Bayes' theorem

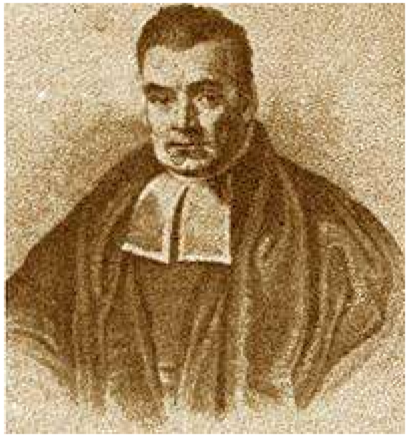
From the definition of conditional probability we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

but $P(A \cap B) = P(B \cap A)$, so

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

Bayes' theorem



First published (posthumously) by the Reverend Thomas Bayes (1702−1761)

An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Soc. **53** (1763) 370; reprinted in Biometrika, **45** (1958) 293.

Frequentist Statistics − general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
$P\,(0.117 < \alpha_s < 0.121)$,

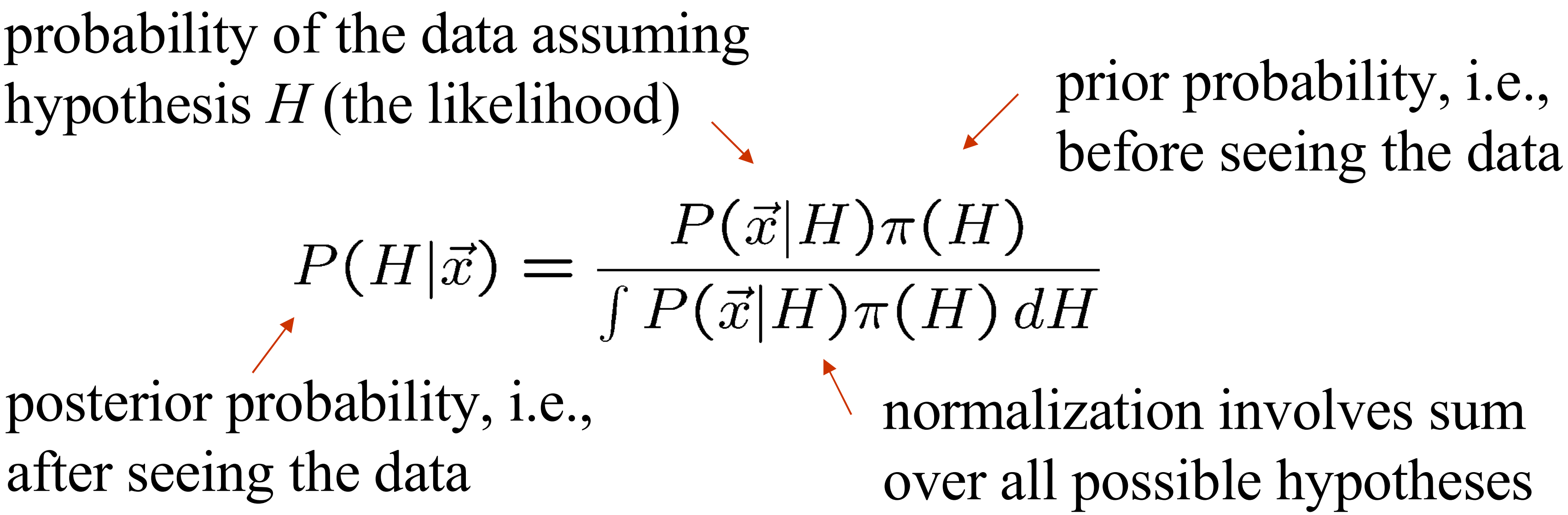
etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered 'usual'.

Bayesian Statistics – general philosophy

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability). Use this for hypotheses:



probability of the data assuming hypothesis *H* (the likelihood)

prior probability, i.e., before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayesian methods can provide more natural treatment of non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists,...

No golden rule for priors (“if-then” character of Bayes’ thm.)

Outline

0 Why worry about this?

1 The Bayesian method

→ 2 Bayesian assessment of uncertainties

3 Bayesian model selection ("discovery")

4 Outlook for Bayesian methods in HEP

5 Bayesian limits

Extra slides

Statistical vs. systematic errors

Statistical errors:

How much would the result fluctuate upon repetition of the measurement?

Implies some set of assumptions to define probability of outcome of the measurement.

Systematic errors:

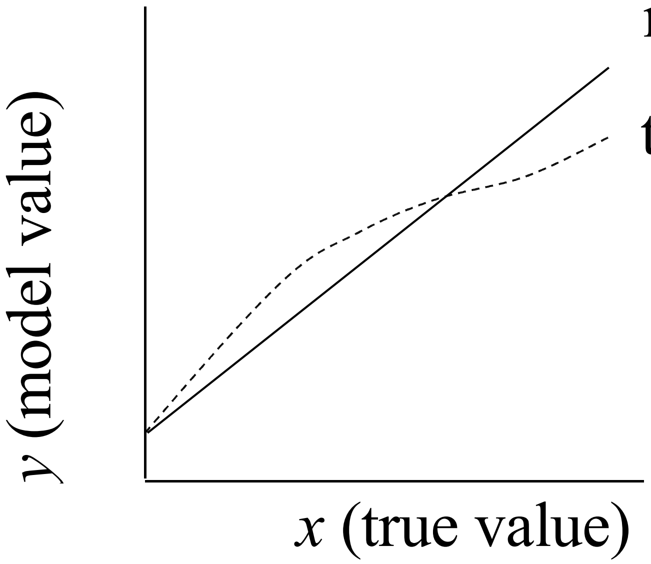
What is the uncertainty in my result due to uncertainty in my assumptions, e.g.,

model (theoretical) uncertainty;
modelling of measurement apparatus.

Usually taken to mean the sources of error do not vary upon repetition of the measurement. Often result from uncertain value of, e.g., calibration constants, efficiencies, etc.

Systematic errors and nuisance parameters

Model prediction (including e.g. detector effects) never same as "true prediction" of the theory:



model: $y = \alpha + \beta x$

truth: $y = \alpha + \beta x + \gamma x^2 + \varepsilon x^3 + \ldots$

Model can be made to approximate better the truth by including more free parameters.

systematic uncertainty ↔ nuisance parameters

Example: fitting a straight line

Data: $(x_i, y_i, \sigma_i)\ , i = 1, \ldots, n\ .$

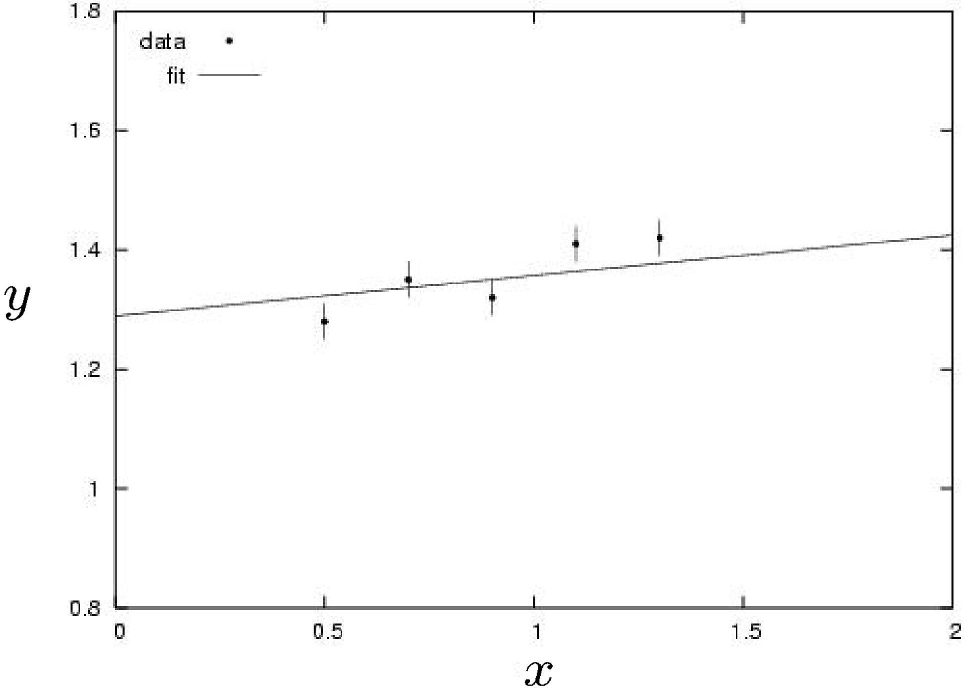
Model: measured y_i independent, Gaussian: $y_i \sim N(\mu(x_i), \sigma_i^2)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x\ ,$$

assume x_i and σ_i known.

Goal: estimate θ_0

(don't care about θ_1).



Frequentist approach

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right] ,$$

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from

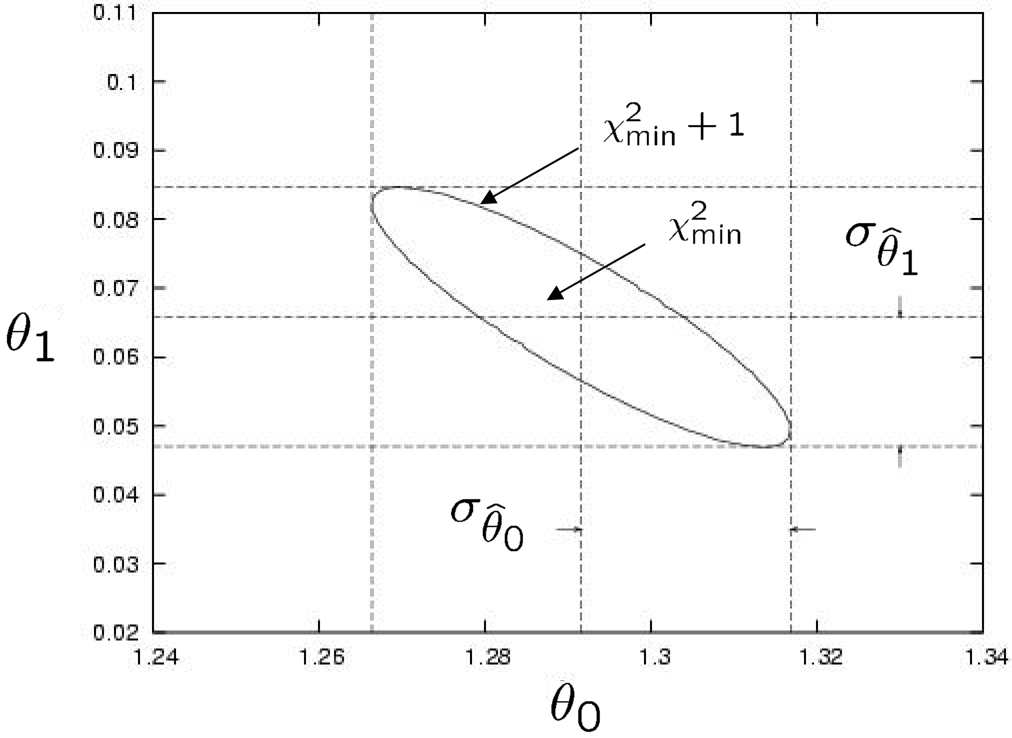
tangent lines to contour

$$\chi^2 = \chi^2_{\text{min}} + 1 .$$

Correlation between

$\hat{\theta}_0, \ \hat{\theta}_1$ causes errors

to increase.

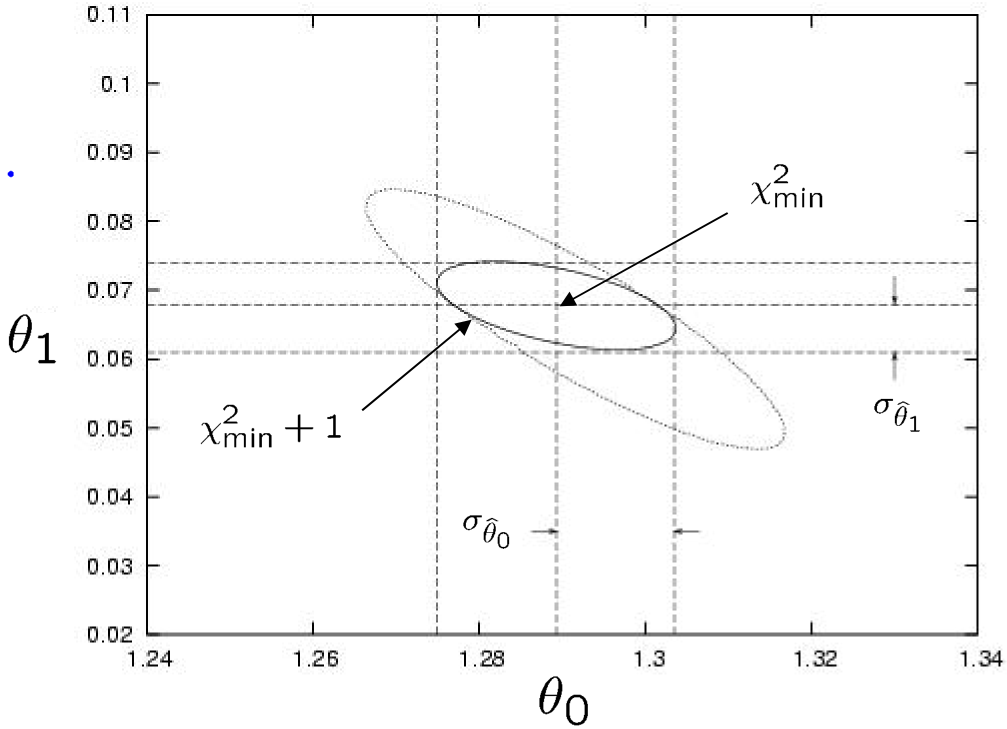


Frequentist case with a measurement t_1 of θ_1

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2} \, .$$

The information on θ_1

improves accuracy of $\hat{\theta}_0$.



Bayesian method

We need to associate prior probabilities with θ_0 and θ_1, e.g.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\, \pi_1(\theta_1)$$

$$\pi_0(\theta_0) = \text{const.}$$

reflects ‘prior ignorance’, in any case much broader than $L(\theta_0)$

$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

← based on previous measurement

Putting this into Bayes’ theorem gives:

$$p(\theta_0, \theta_1 | \vec{y}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i;\theta_0,\theta_1))^2/2\sigma_i^2}\ \pi_0\ \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

posterior Θ likelihood × prior

Bayesian method (continued)

We then integrate (marginalize) $p(\theta_0, \theta_1 \,|\, x)$ to find $p(\theta_0 \,|\, x)$:

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x) \, d\theta_1 \; .$$

In this example we can do the integral (rare). We find

$$\begin{aligned} p(\theta_0|x) &= \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma_{\theta_0}^2} \quad \text{with} \\ \hat{\theta}_0 &= \text{same as ML estimator} \\ \sigma_{\theta_0} &= \sigma_{\hat{\theta}_0} \text{ (same as before)} \end{aligned}$$

Ability to marginalize over nuisance parameters is an important feature of Bayesian statistics.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

often high dimensionality and impossible in closed form, also impossible with 'normal' acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

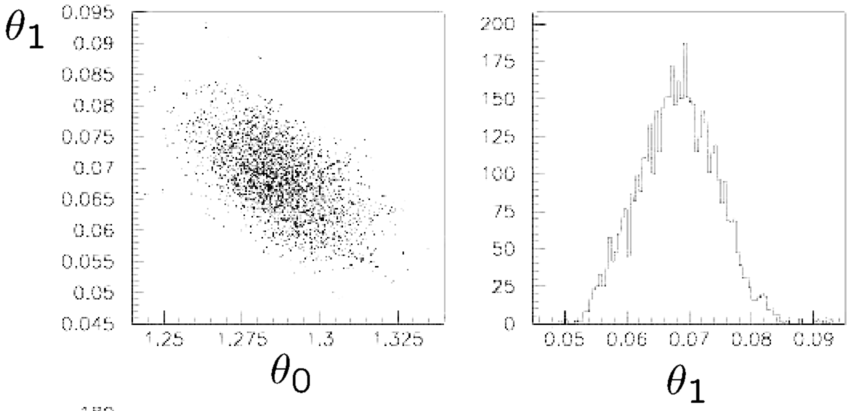
MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

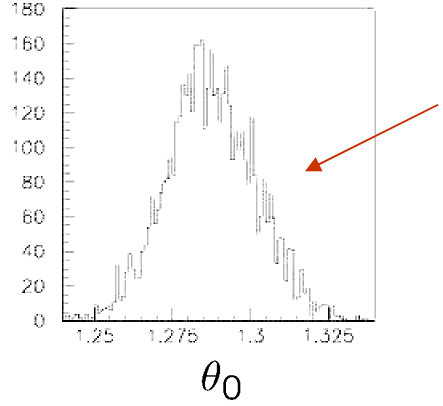
cannot use for many applications, e.g., detector MC;
effective stat. error greater than naive $\sqrt{n}$.

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:





Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Bayesian method with vague prior

Suppose we don’t have a previous measurement of θ_1 but rather some vague information, e.g., a theorist tells us:

$\theta_1 \geq 0$ (essentially certain);

θ_1 should have order of magnitude less than 0.1 ‘or so’.

Under pressure, the theorist sketches the following prior:

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau} \,, \quad \theta_1 \geq 0 \,, \quad \tau = 0.1 \,.$$

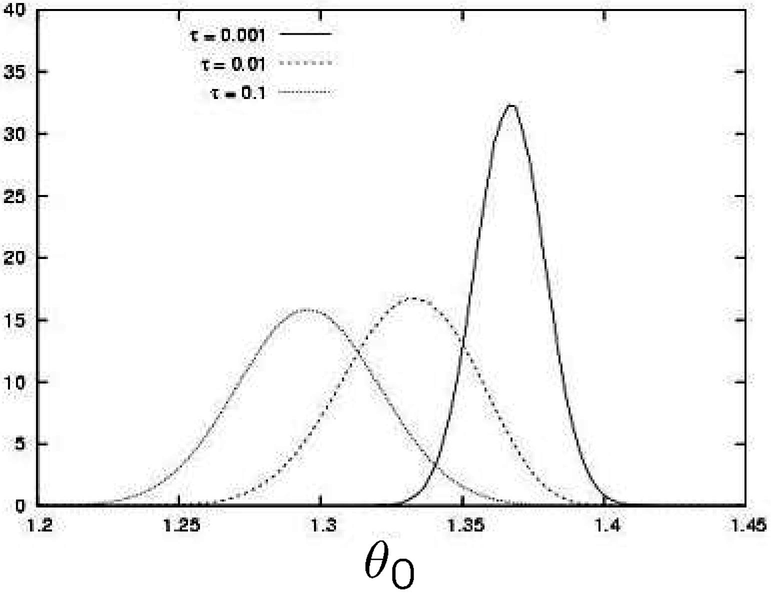
From this we will obtain posterior probabilities for θ_0 (next slide).

We do not need to get the theorist to ‘commit’ to this prior; final result has ‘if-then’ character.

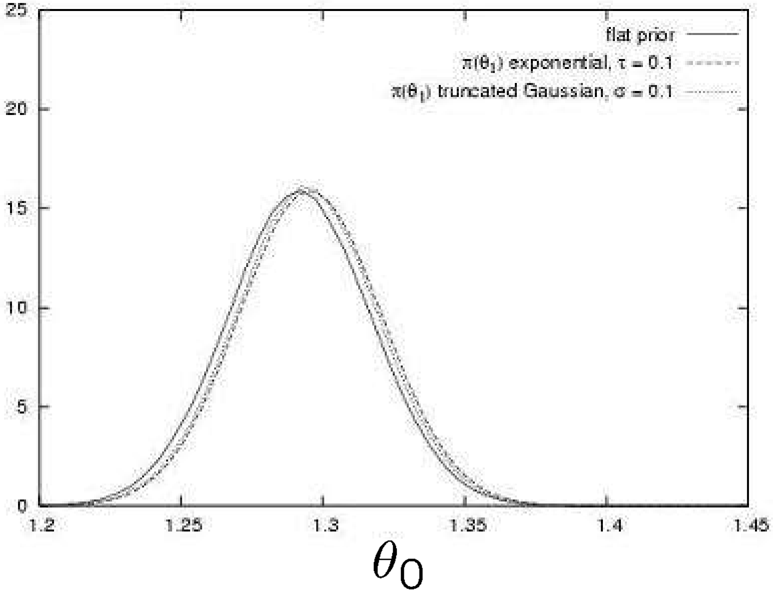
Sensitivity to prior

Vary $\pi(\theta)$ to explore how extreme your prior beliefs would have to be to justify various conclusions (sensitivity analysis).

Try exponential with different mean values...



Try different functional forms...



A more general fit (symbolic)

Given measurements: $y_i \pm \sigma_i^{\mathrm{stat}} \pm \sigma_i^{\mathrm{sys}}, \quad i = 1, \ldots, n\,,$

and (usually) covariances: $V_{ij}^{\mathrm{stat}},\ V_{ij}^{\mathrm{sys}}\,.$

Predicted value: $\mu(x_i;\theta)\,,$ expectation value $E[y_i] = \mu(x_i;\theta) + b_i$

control variable

parameters

bias

Often take: $V_{ij} = V_{ij}^{\mathrm{stat}} + V_{ij}^{\mathrm{sys}}$

Minimize $\chi^2(\theta) = (\vec{y} - \vec{\mu}(\theta))^T V^{-1} (\vec{y} - \vec{\mu}(\theta))$

Equivalent to maximizing $L(\theta)$ » $e^{-\chi^2/2}$, i.e., least squares same as maximum likelihood using a Gaussian likelihood function.

Its Bayesian equivalent

Take $$L(\vec{y}|\vec{\theta},\vec{b}) \sim \exp\left[-\frac{1}{2}(\vec{y}-\vec{\mu}(\theta)-\vec{b})^T\, V_{\mathrm{stat}}^{-1}\,(\vec{y}-\vec{\mu}(\theta)-\vec{b})\right]$$

$$\pi_b(\vec{b}) \sim \exp\left[-\frac{1}{2}\,\vec{b}^T\, V_{\mathrm{sys}}^{-1}\,\vec{b}\right]$$

$$\pi_\theta(\theta) \sim \mathrm{const.}$$

Joint probability for all parameters

and use Bayes' theorem: $$p(\theta,\vec{b}|\vec{y}) \propto L(\vec{y}|\theta,\vec{b})\pi_\theta(\theta)\pi_b(\vec{b})$$

To get desired probability for θ, integrate (marginalize) over b:

$$p(\theta|\vec{y}) = \int p(\theta,\vec{b}|\vec{y})\, d\vec{b}$$

→ Posterior is Gaussian with mode same as least squares estimator, σ_θ same as from $\chi^2 = \chi^2_{\mathrm{min}} + 1$. (Back where we started!)

The error on the error

Some systematic errors are well determined

Error from finite Monte Carlo sample

Some are less obvious

Do analysis in *n* ‘equally valid’ ways and extract systematic error from ‘spread’ in results.

Some are educated guesses

Guess possible size of missing terms in perturbation series;

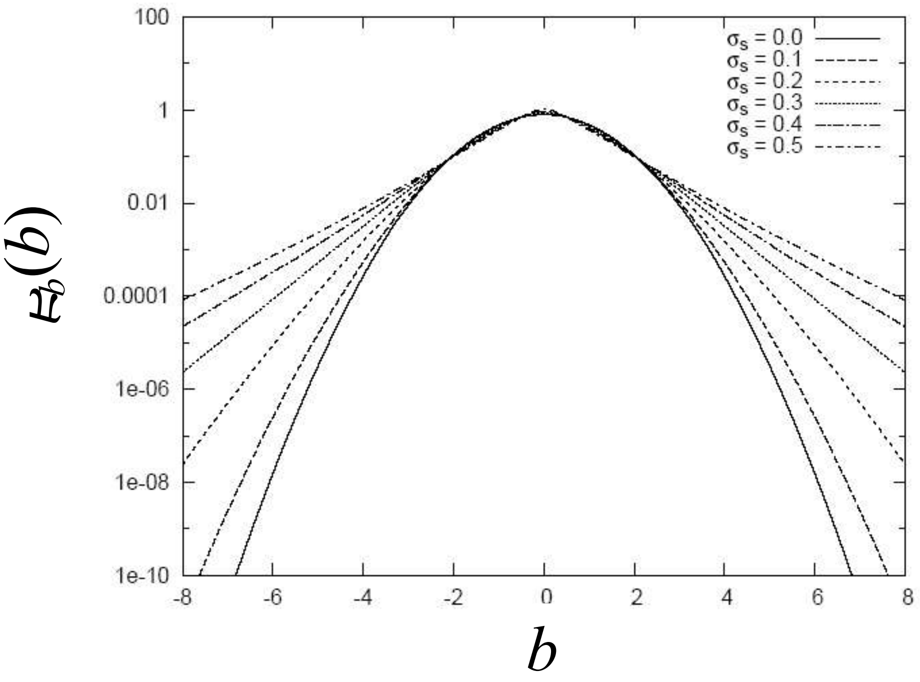
vary renormalization scale $(\mu/2 < Q < 2\mu\ ?)$

Can we incorporate the ‘error on the error’?

(cf. G. D’Agostini 1999; Dose & von der Linden 1999)

A prior for bias $\pi_b(b)$ with longer tails

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi} s_i \sigma_i^{\rm sys}} \exp\left[-\frac{1}{2}\frac{b_i^2}{(s_i\sigma_i^{\rm sys})^2}\right] \pi_s(s_i)\, ds_i$$



Represents 'error on the error';

standard deviation of $\pi_s(s)$ is σ_s.

Gaussian ($\sigma_s = 0$) $P(|b| > 4\sigma_{\rm sys}) = 6.3 \times 10^{-5}$

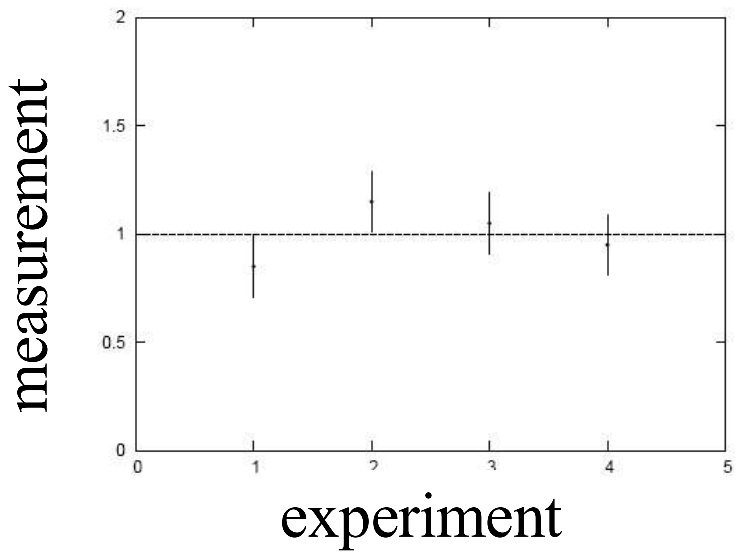
$\sigma_s = 0.5$ $P(|b| > 4\sigma_{\rm sys}) = 6.5 \times 10^{-3}$

A simple test

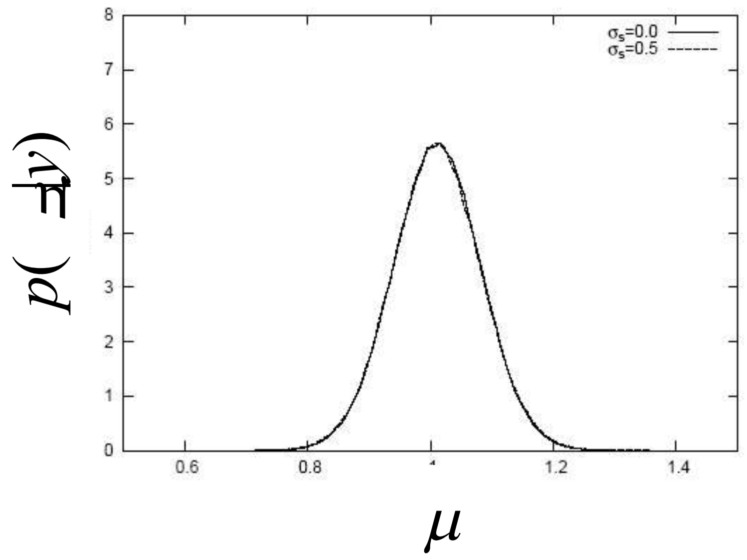
Suppose fit effectively averages four measurements.

Take $\sigma_{\mathrm{sys}} = \sigma_{\mathrm{stat}} = 0.1$, uncorrelated.

Case #1: data appear compatible



Posterior $p(\mu|y)$:



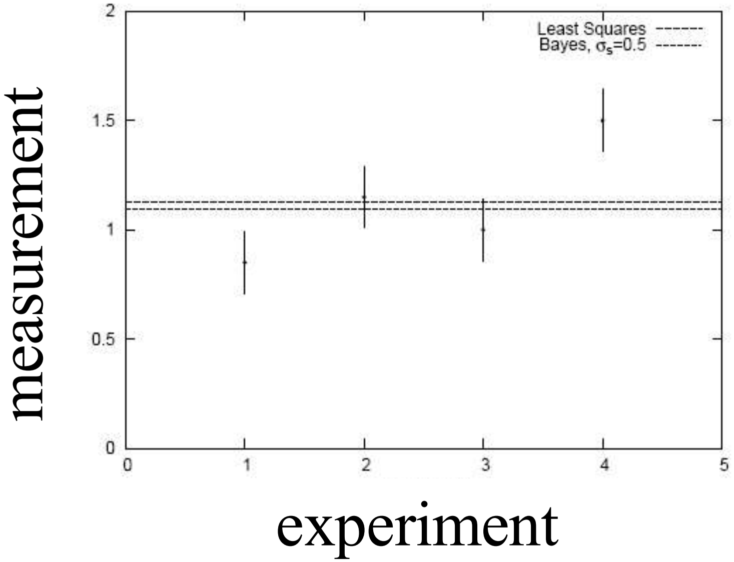
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$$\sigma_s = 0.0 : \quad \hat{\mu} = 1.000 \pm 0.071$$

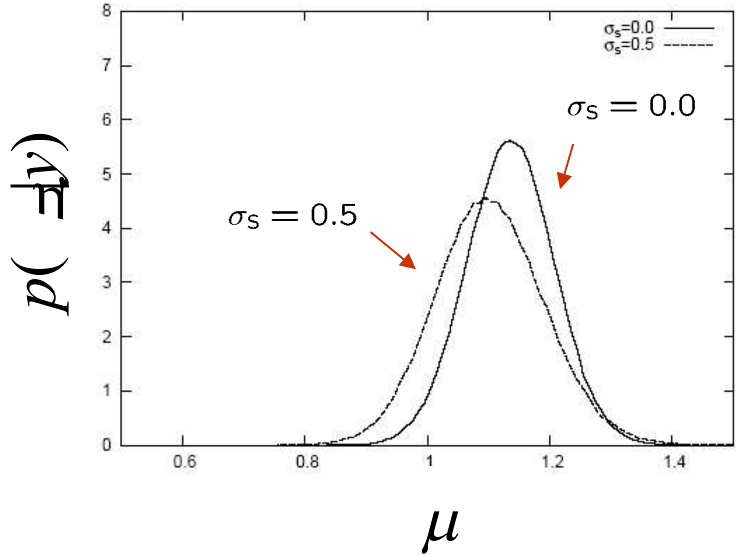
$$\sigma_s = 0.5 : \quad \hat{\mu} = 1.000 \pm 0.072$$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



$\sigma_s = 0.0: \quad \hat{\mu} = 1.125 \pm 0.071$

$\sigma_s = 0.5: \quad \hat{\mu} = 1.093 \pm 0.089$

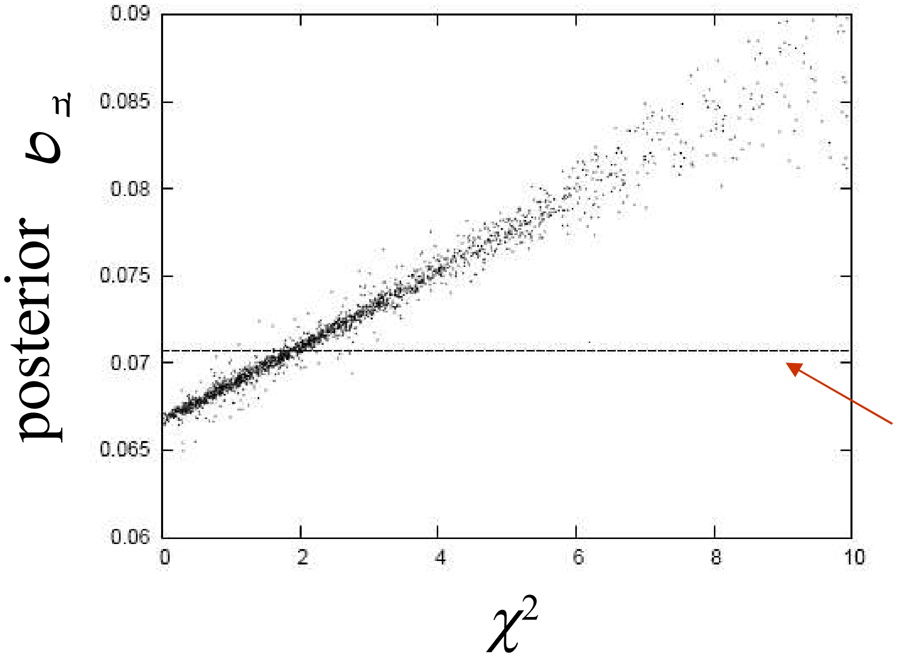
→ Bayesian fit less sensitive to outlier.

→ Error now connected to goodness-of-fit.

Goodness-of-fit vs. size of error

In LS fit, value of minimized χ^2 does not affect size of error on fitted parameter.

In Bayesian analysis with non-Gaussian prior for systematics, a high χ^2 corresponds to a larger error (and vice versa).



2000 repetitions of experiment, $\sigma_s = 0.5$, here no actual bias.

σ_μ from least squares

Uncertainty from parametrization of PDFs

Try e.g. $xf(x) = ax^b(1-x)^c(1+d\sqrt{x}+ex)$ (MRST)

or $xf(x) = ax^b(1-x)^c e^{d \cdot x}(1+e^e x)^f$ (CTEQ)

The form should be flexible enough to describe the data; frequentist analysis has to decide how many parameters are justified.

In a Bayesian analysis we can insert as many parameters as we want, but constrain them with priors.

Suppose e.g. based on a theoretical bias for things not too bumpy, that a certain parametrization 'should hold to 2%'.

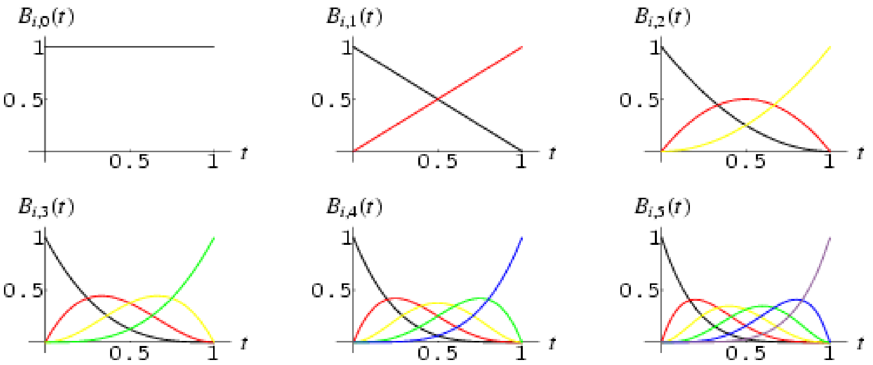
How to translate this into a set of prior probabilites?

Residual function

Try e.g. $xf(x) = ax^b(1-x)^c(1+\ldots) + r(x)$ ← 'residual function'

where $r(x)$ is something very flexible, e.g., superposition of

Bernstein polynomials, coefficients ν_i: $r(x) = \sum_i \nu_i B_i(x)$



mathworld.wolfram.com

$$B_{i,n} = \binom{n}{i} x^i (1-x)^{n-i}$$

Assign priors for the ν_i centred around 0, width chosen to reflect the uncertainty in $xf(x)$ (e.g. a couple of percent).

→ Ongoing effort.

Outline

0 Why worry about this?

1 The Bayesian method

2 Bayesian assessment of uncertainties

→ 3 Bayesian model selection ("discovery")

4 Outlook for Bayesian methods in HEP

5 Bayesian limits

Extra slides

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 ≤ compatibility with H_0 when compared to the data we got | H_0)

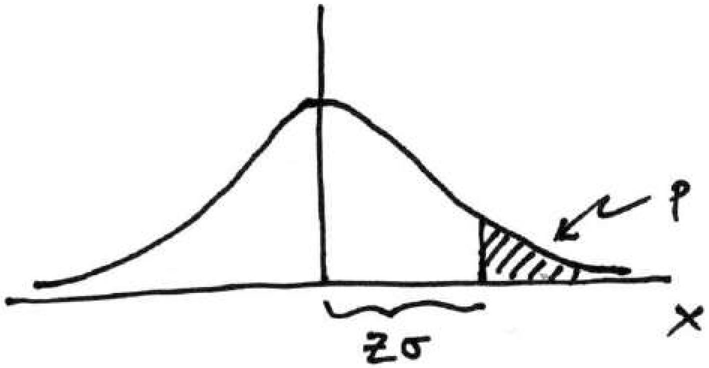
For example, if high values of t mean less compatibility,

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

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^0D^0 mixing	~ 0.05
Higgs	$\sim 10^{-7}$ (?)
Life on Mars	$\sim 10^{-10}$
Astrology	$\sim 10^{-20}$

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:

no Higgs

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)}{P(x|H_1)} \times \frac{\pi(H_0)}{\pi(H_1)}$$

Higgs

Bayes factor B_{01}

prior 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 $\pi(H_i, \vec{\theta}_i) = p_i \pi_j(\vec{\theta}_j)$

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:

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

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

Use Bayes theorem

So therefore the Bayes factor is

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

Ratio of marginal likelihoods

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

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})\, d\vec{\theta} \quad \longleftarrow \quad \text{‘marginal likelihood’}$$

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

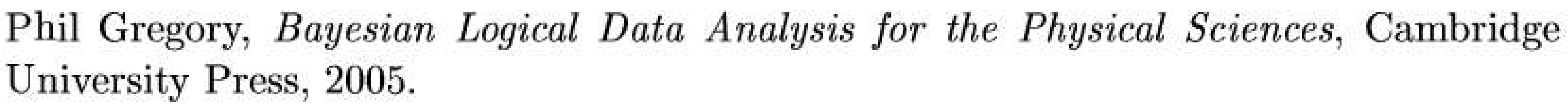
$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E_f\left[\frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})}\right]$$

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.



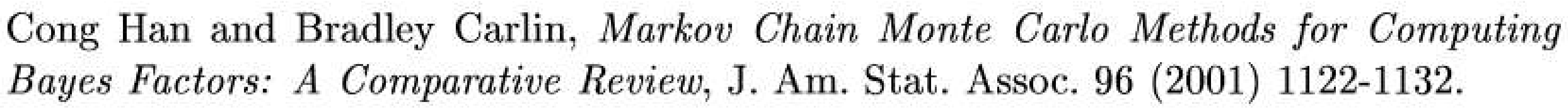
Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press, 2005.

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:

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

(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 $\mu_{\rm up}$; consider $m_{\rm H}$ excluded if upper limit $\mu_{\rm up} < 1.0$.

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:

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

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

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

Bayes factors for Higgs analysis

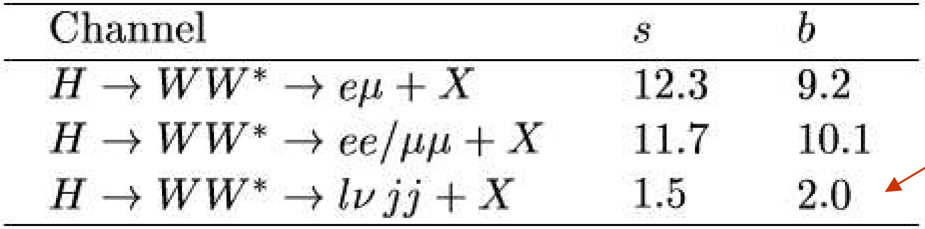
The Bayes factor B_{10} is

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

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}, $m_{\mathrm{H}} = 130$ 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

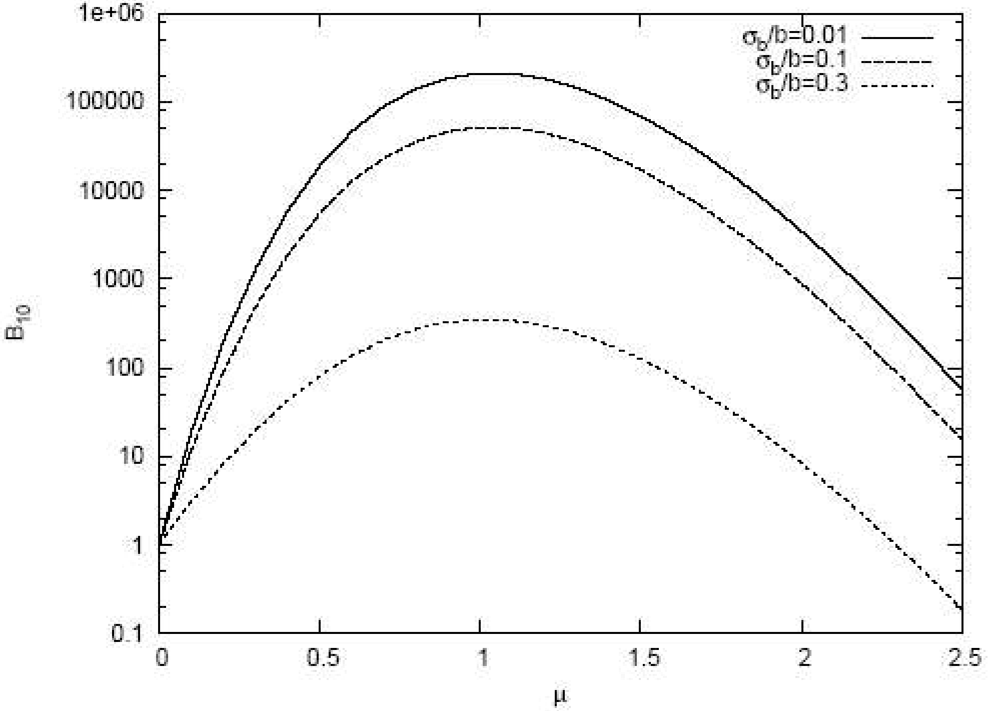
$l\nu jj$ was for 30 fb^{-1}, in paper; divided by 3

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)

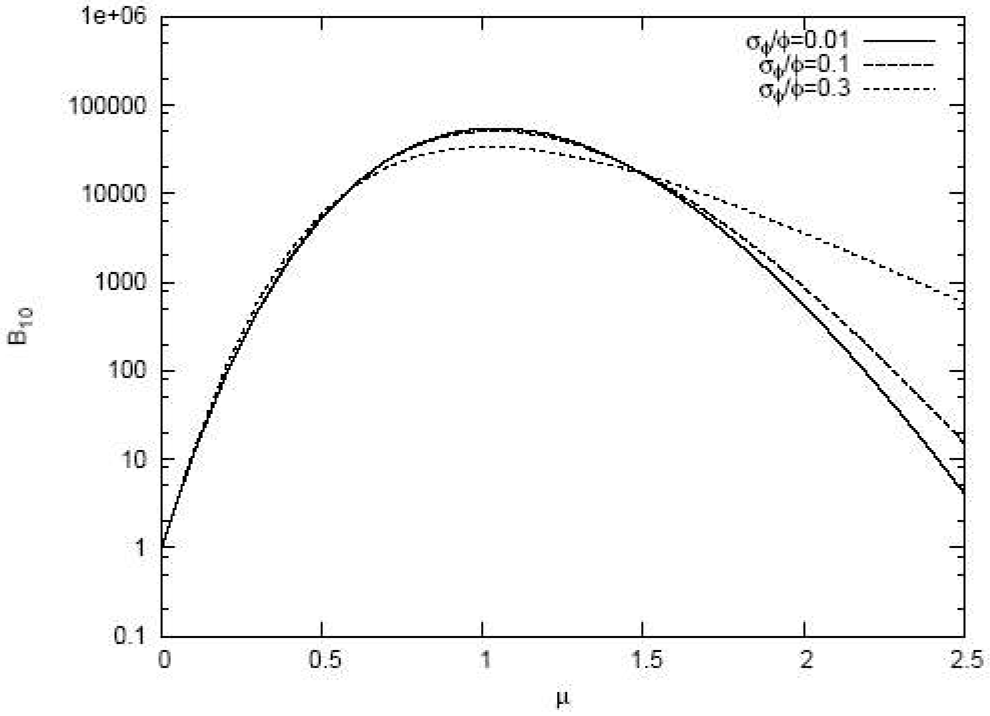
Create data set by hand with $n_i \sim$ nearest integer $(\phi_i + b_i)$, i.e., $\mu = 1$:
$n_1 = 22$, $n_2 = 22$, $n_3 = 4$.

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)



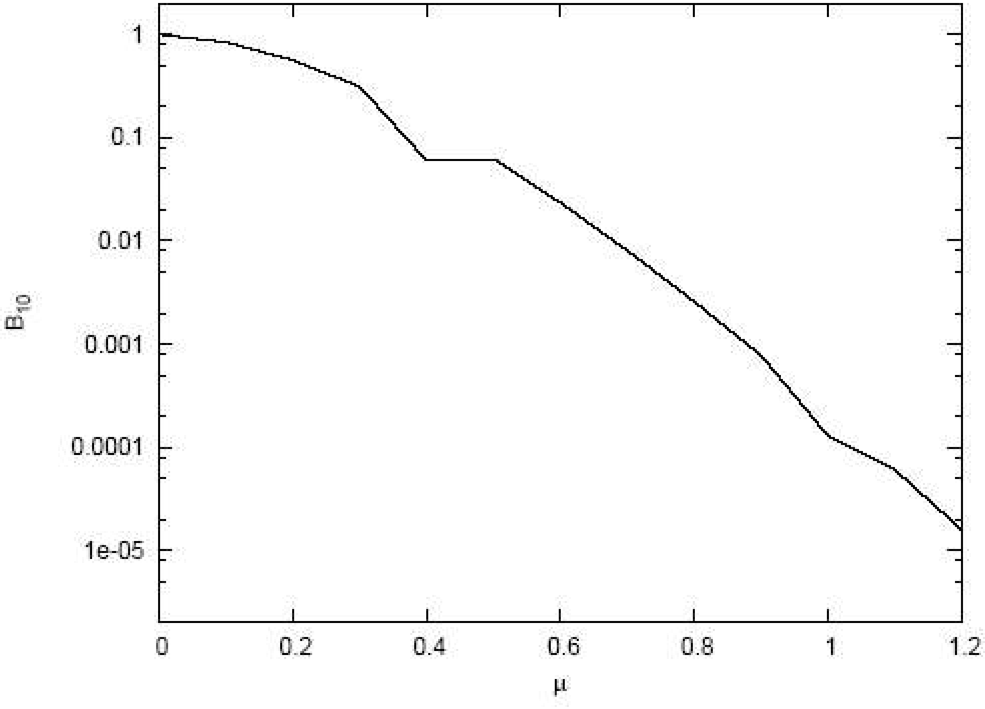
B_{10} for $\sigma_b/b = 0.1$, different values of σ_ϕ/ϕ, as a function of μ.

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 $\sigma_b/b = 0.1$, $\sigma_\phi/\phi_i = 0.1$.



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.)

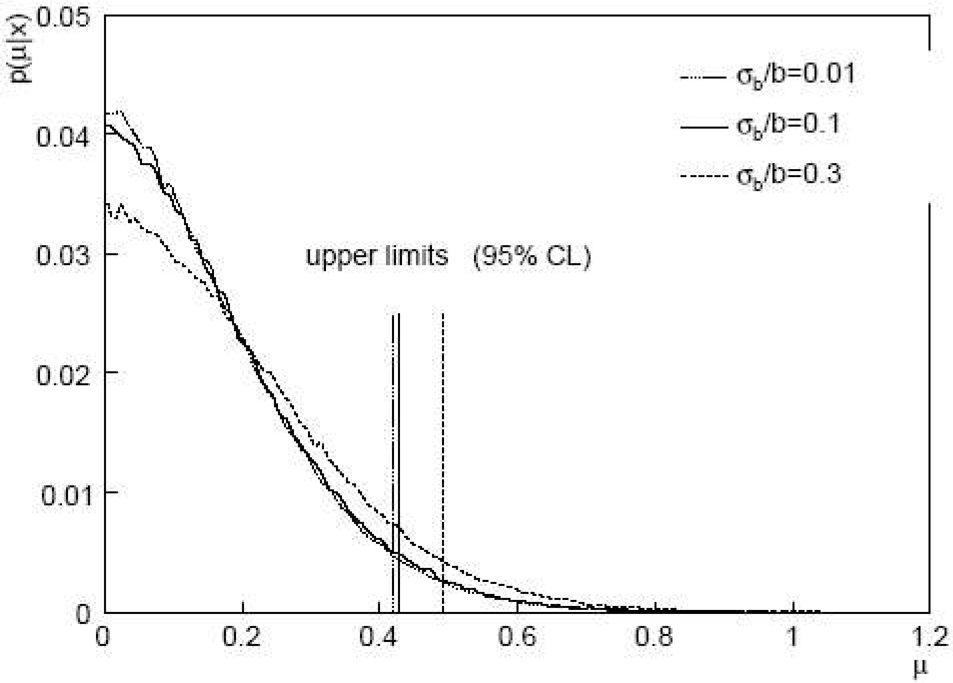


Figure 4: The posterior distribution of μ with a data set compatible with background only ($n_1 = 9$, $n_2 = 10$, $n_3 = 2$) for $\sigma_\varphi/\varphi = 0.1$ and several values of σ_b/b.

Posterior pdf for μ , upper limits (2)

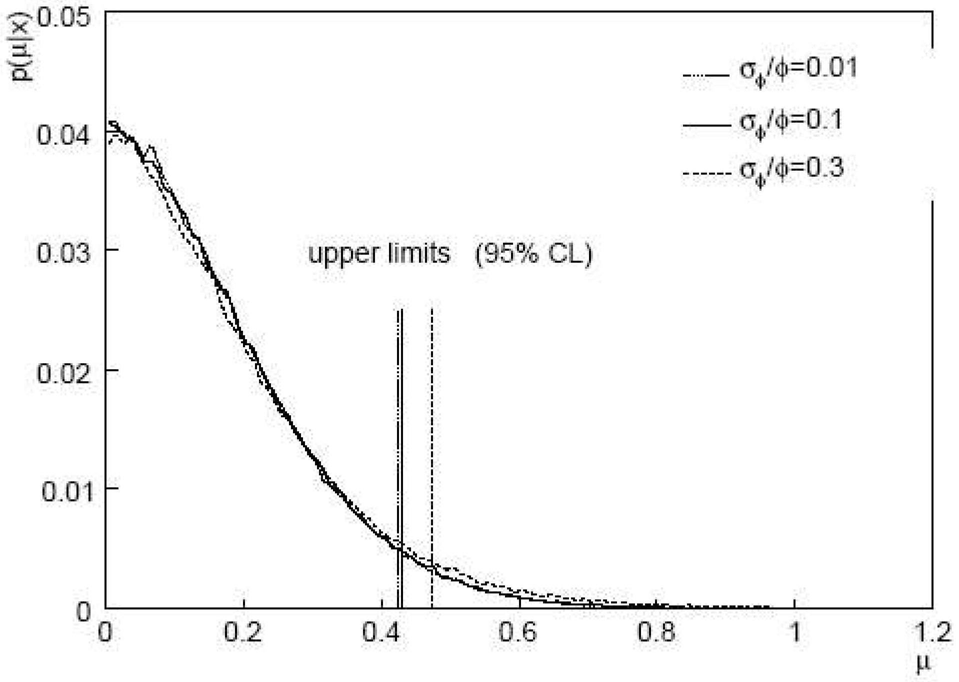


Figure 5: The posterior distribution of μ with a data set compatible with background only ($n_1 = 9$, $n_2 = 10$, $n_3 = 2$) for $\sigma_b/b = 0.1$ and several values of σ_φ/φ.

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

Extra slides

Some Bayesian references

P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, CUP, 2005

D. Sivia, *Data Analysis: a Bayesian Tutorial*, OUP, 2006

S. Press, *Subjective and Objective Bayesian Statistics: Principles, Models and Applications*, 2nd ed., Wiley, 2003

A. O'Hagan, Kendall's, *Advanced Theory of Statistics, Vol. 2B, Bayesian Inference*, Arnold Publishers, 1994

A. Gelman et al., *Bayesian Data Analysis*, 2nd ed., CRC, 2004

W. Bolstad, *Introduction to Bayesian Statistics*, Wiley, 2004

E.T. Jaynes, *Probability Theory: the Logic of Science*, CUP, 2003

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

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

Analytic formulae for limits

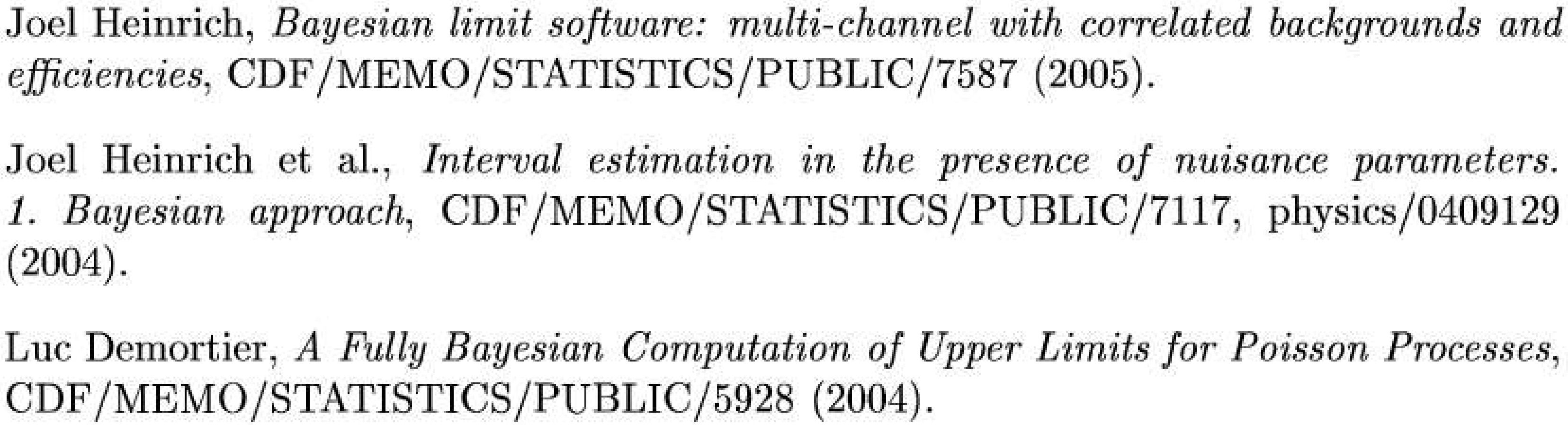
There are a number of papers describing Bayesian limits for a variety of standard scenarios

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*

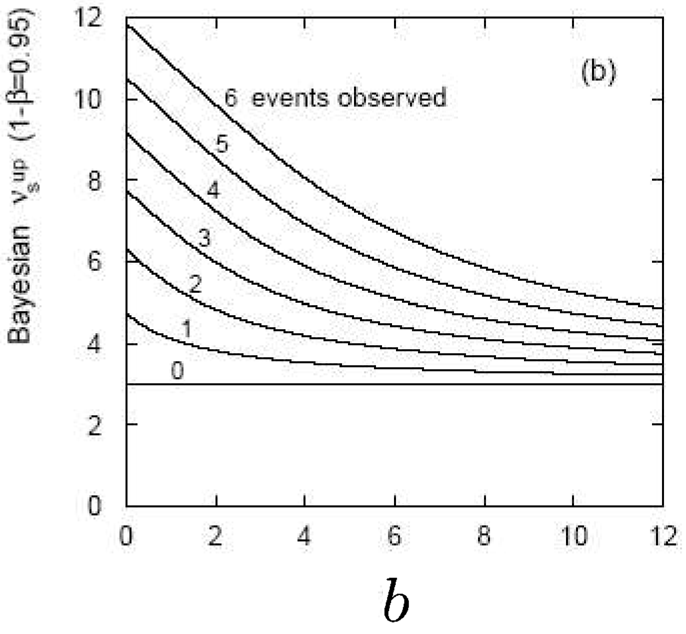
Solve numerically to find limit s_{up}.

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').

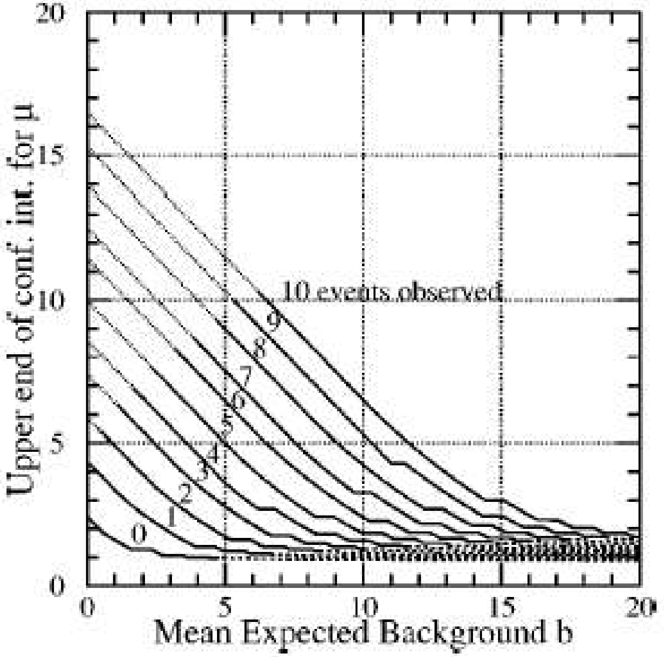
Never goes negative.

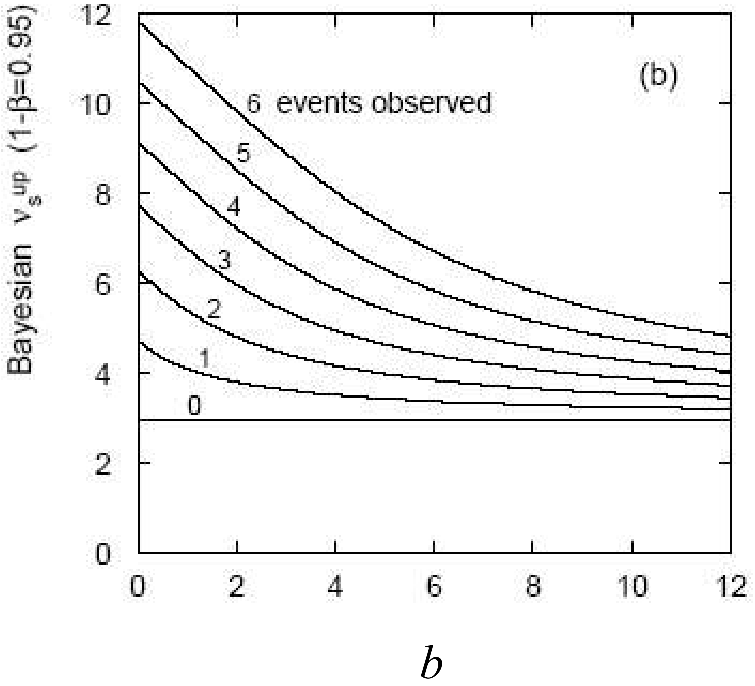
Doesn't depend on b if $n = 0$.



Upper limit versus *b*

Feldman & Cousins, PRD 57 (1998) 3873





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

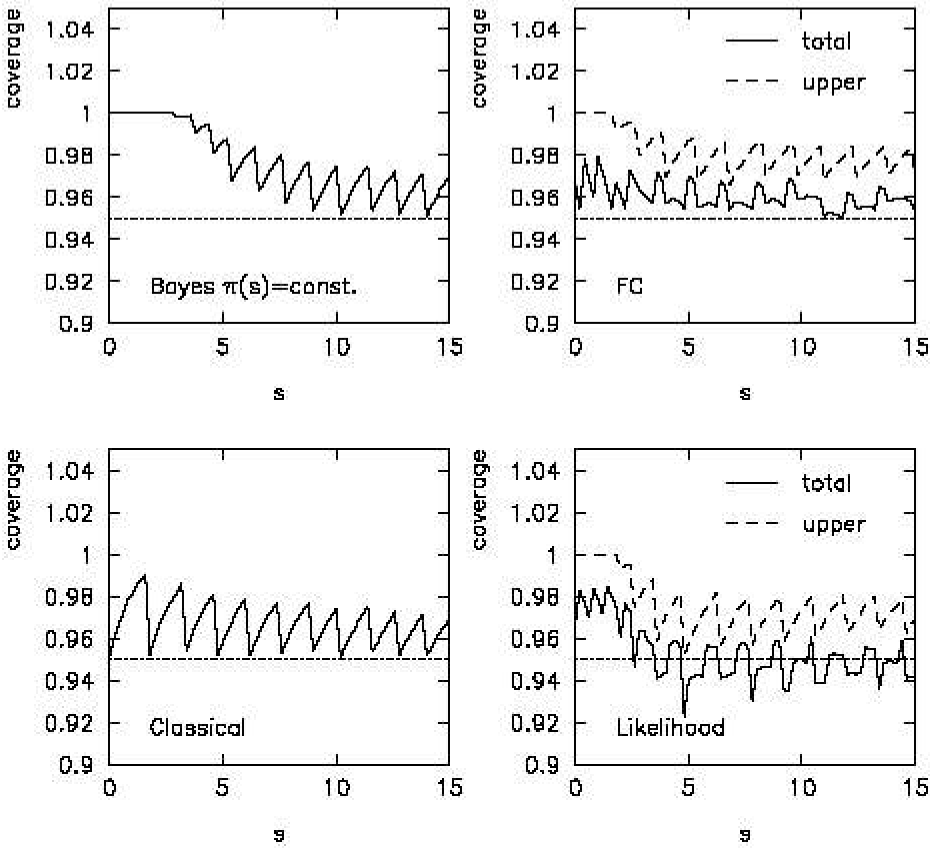
Classical: yes

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.,

$$
\begin{aligned}
\pi(s,b) &= \pi_s(s)\pi_b(b) \quad \text{(or include correlations as appropriate)} \\
\pi_s(s) &= \text{const}, \quad \sim 1/s, \ldots \quad ? \ \text{(see R. Barlow talk)} \\
\pi_b(b) &= \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad \text{(or whatever)}
\end{aligned}
$$

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., $E[s_{\mathrm{up}} \mid s = 0]$.

Consider also:

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

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

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

Need likelihood (or summary thereof).

consumer’s prior

MCMC basics: Metropolis-Hastings algorithm

Goal: given an *n*-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$ e.g. Gaussian centred about $\vec{\theta}_0$

3) Form Hastings test ratio $\alpha = \min\left[1, \frac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \mathrm{Uniform}[0, 1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than naive $\sqrt{n}$.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

Test ratio is (*Metropolis*-Hastings): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.
If proposed step rejected, hop in place.

Metropolis-Hastings caveats

Actually one can only prove that the sequence of points follows the desired pdf in the limit where it runs forever.

There may be a "burn-in" period where the sequence does not initially follow $p(\vec{\theta})$.

Unfortunately there are few useful theorems to tell us when the sequence has converged.

Look at trace plots, autocorrelation.

Check result with different proposal density.

If you think it's converged, try it again with 10 times more points.