Statistical Methods in Particle Physics
Lecture 2: Limits and Discovery

SUSSP65

St Andrews

16–29 August 2009

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

Outline

Lecture #1: An introduction to Bayesian statistical methods

- Role of probability in data analysis (Frequentist, Bayesian)
- A simple fitting problem : Frequentist vs. Bayesian solution
- Bayesian computation, Markov Chain Monte Carlo

Lecture #2: Setting limits, making a discovery

- Frequentist vs Bayesian approach,
- treatment of systematic uncertainties

Lecture #3: Multivariate methods for HEP

- Event selection as a statistical test
- Neyman-Pearson lemma and likelihood ratio test
- Some multivariate classifiers:
  - NN, BDT, SVM, ...

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

Suppose the number of events found is roughly equal to the expected number of background events, e.g., $b = 4.6$ and we observe $n_{\text{obs}} = 5$ events.

The evidence for the presence of signal events is not statistically significant,

→ set upper limit on the parameter s, taking into consideration any uncertainty in b.

Setting limits

Frequentist intervals (limits) for a parameter s can be found by defining a test of the hypothesized value s (do this for all s):

> Specify values of the data n that are 'disfavoured' by s (critical region) such that $P(n \text{ in critical region}) \leq \gamma$ for a prespecified γ, e.g., 0.05 or 0.1.
>
> If n is observed in the critical region, reject the value s.

Now invert the test to define a confidence interval as:

> set of s values that would not be rejected in a test of size γ (confidence level is $1 - \gamma$).

The interval will cover the true value of s with probability $\geq 1 - \gamma$.

Frequentist upper limit for Poisson parameter

First suppose that the expected background b is known.

Find the hypothetical value of s such that there is a given small probability, say, $\gamma = 0.05$, to find as few events as we did or less:

$$\gamma = P(n \leq n_{\text{obs}}; s, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Solve numerically for $s = s_{\text{up}}$, this gives an upper limit on s at a confidence level of $1-\gamma$.

Example: suppose $b = 0$ and we find $n_{\text{obs}} = 0$. For $1-\gamma = 0.95$,

$$\gamma = P(n = 0; s, b = 0) = e^{-s} \quad \rightarrow \quad s_{\text{up}} = -\ln \gamma \approx 3.00$$

$[0, s_{\text{up}}]$ is an example of a confidence interval. It is designed to include the true value of s with probability at least $1-\gamma$ for any s.

Calculating Poisson parameter limits

Analogous procedure for lower limit $s_{\rm lo}$.

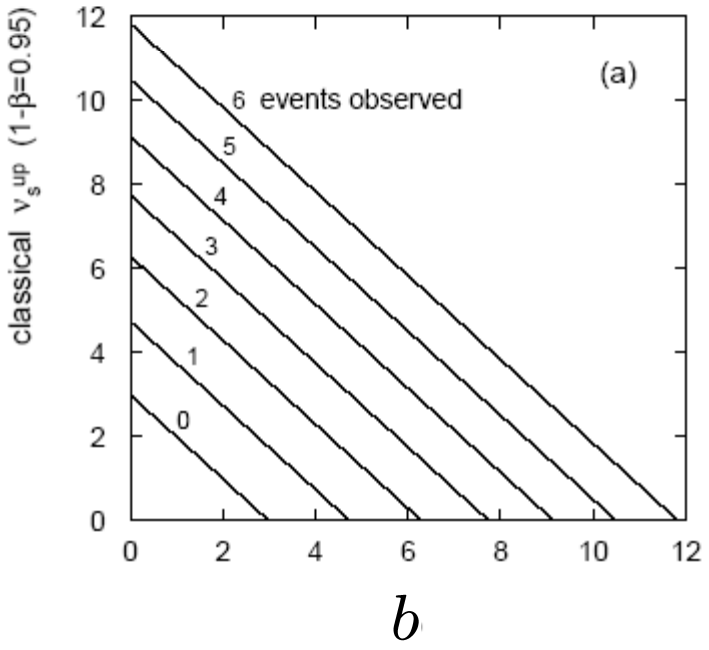
To solve for $s_{\rm lo}$, $s_{\rm up}$, can exploit relation to χ^2 distribution:

$$s_{\rm lo} = \frac{1}{2} F^{-1}_{\chi^2}(\alpha; 2n) - b$$

Quantile of χ^2 distribution

$$s_{\rm up} = \frac{1}{2} F^{-1}_{\chi^2}(1-\beta; 2(n+1)) - b$$

For low fluctuation of n this can give negative result for $s_{\rm up}$; i.e. confidence interval is empty.



Limits near a physical boundary

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL = 0.9, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (\text{CL} = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when limit of parameter is close to a physical boundary, cf. m_ν estimated using $E^2 - p^2$.

Expected limit for on s if $s = 0$

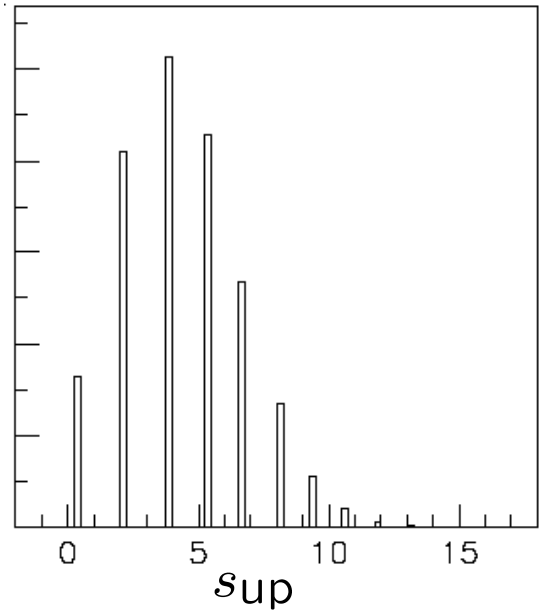
Physicist: I should have used CL = 0.95 — then $s_{up} = 0.496$

Even better: for CL = 0.917923 we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



Likelihood ratio limits (Feldman-Cousins)

Define likelihood ratio for hypothesized parameter value s:

$$l(s) = \frac{L(n|s,b)}{L(n|\hat{s},b)} \quad \text{where} \quad \hat{s} = \begin{cases} n-b & n \geq b, \\ 0 & \text{otherwise} \end{cases}$$

Here $\hat{s}$ is the ML estimator, note $0 \leq l(s) \leq 1$.

Define a statistical test for a hypothetical value of s:

Rejection region defined by low values of likelihood ratio.

Reject s if p-value = $P(l(s) \leq l_{\text{obs}} \mid s)$ is less than γ (e.g. $\gamma = 0.05$).

Confidence interval at CL = $1-\gamma$ is the set of s values not rejected.

Resulting intervals can be one- or two-sided (depending on n).

(Re)discovered for HEP by Feldman and Cousins,
Phys. Rev. D 57 (1998) 3873.

More on intervals from LR test (Feldman-Cousins)

Caveat with coverage: suppose we find $n >> b$.
Usually one then quotes a measurement: $\hat{s} = n - b\,, \quad \hat{\sigma}_{\hat{s}} = \sqrt{n}$

If, however, n isn't large enough to claim discovery, one sets a limit on s.

FC pointed out that if this decision is made based on n, then the actual coverage probability of the interval can be less than the stated confidence level ('flip-flopping').

FC intervals remove this, providing a smooth transition from 1- to 2-sided intervals, depending on n.

But, suppose FC gives e.g. $0.1 < s < 5$ at 90% CL,
p-value of s=0 still substantial. Part of upper-limit 'wasted'?

Nuisance parameters and limits

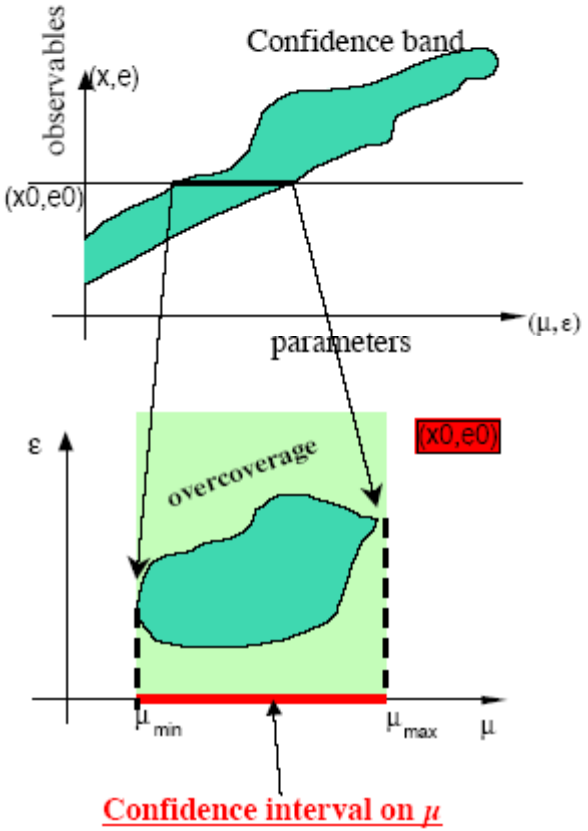
In general we don't know the background b perfectly.

Suppose we have a measurement of b, e.g., $b_{\text{meas}} \sim N(b, \sigma_b)$

So the data are really: n events and the value b_{meas}.

In principle the confidence interval recipe can be generalized to two measurements and two parameters.

Difficult and rarely attempted, but see e.g. talks by K. Cranmer at PHYSTAT03 and by G. Punzi at PHYSTAT05.



G. Punzi, PHYSTAT05

Nuisance parameters and profile likelihood

Suppose model has likelihood function

$$L(\mu, \nu) = P(\vec{x}|\mu, \nu)$$

Parameters of interest Nuisance parameters

Define the profile likelihood ratio as

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\nu}})}{L(\hat{\mu}, \hat{\nu})}$$

Maximizes L for given value of μ

Maximizes L

$\lambda(\mu)$ reflects level of agreement between data and μ $(0 \leq \lambda(\mu) \leq 1)$

Equivalently use $q_\mu = -2 \ln \lambda(\mu)$

p-value from profile likelihood ratio

Large q_μ means worse agreement between data and μ

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

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu \approx 1 - F_{\chi^2_n}(q_{\mu,\mathrm{obs}})$$

rapidly approaches chi-square pdf (Wilks' theorem)

chi-square cumulative distribution, degrees of freedom = dimension of μ

Reject μ if $p_\mu < \gamma = 1 - \mathrm{CL}$

(Approx.) confidence interval for μ = set of μ values not rejected.

Coverage not exact for all ν but very good if $\nu \approx \hat{\nu}$.

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

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

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

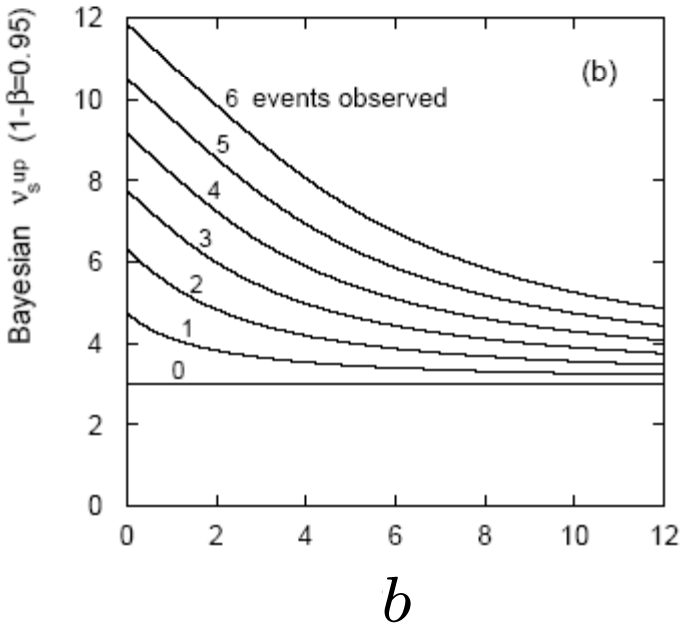
Solve numerically to find limit s_{up}.

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 $n = 0$.



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).

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)

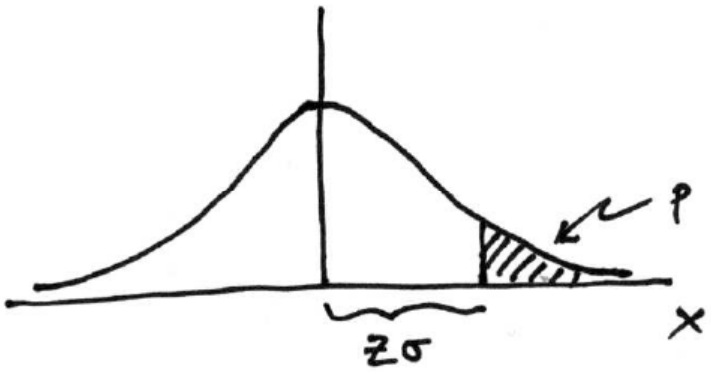
For example, if high values of t mean less compatibility,

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

If p-value comes out small, then this is evidence against the background-only hypothesis $\rightarrow$ 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

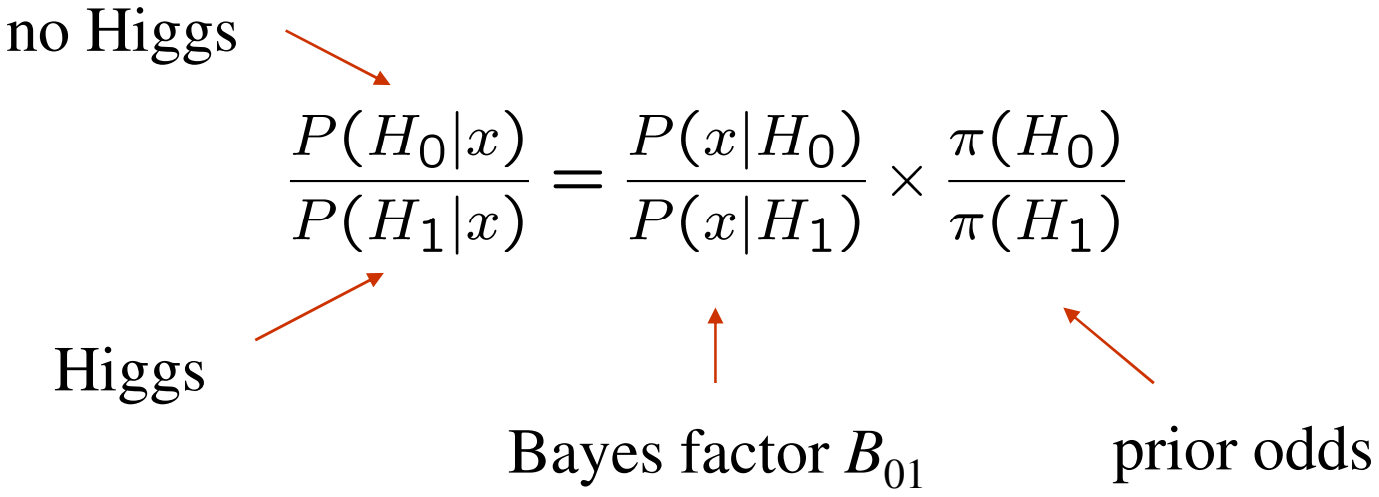
HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = 5$.

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}

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.

Will this be adopted in HEP?

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

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

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)
Nested sampling
...

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

Example of systematics in a search

Combination of Higgs boson search channels (ATLAS)

Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics, arXiv:0901.0512, CERN-OPEN-2008-20.

Standard Model Higgs channels considered (more to be used later):

$H \to \gamma\gamma$

$H \to WW^{(*)} \to e\nu\mu\nu$

$H \to ZZ^{(*)} \to 4l$ (l = e, μ)

$H \to \tau^+\tau^- \to ll, lh$

Used profile likelihood method for systematic uncertainties:

background rates, signal & background shapes.

Statistical model for Higgs search

Bin i of a given channel has n_i events, expectation value is

$$E[n_i] = \mu L \varepsilon_i \sigma_i \mathcal{B} + b_i \equiv \mu s_i + b_i$$

μ is global strength parameter, common to all channels.
$\mu = 0$ means background only, $\mu = 1$ is SM hypothesis.

Expected signal and background are:

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta}_s)\, dx \ ,$$

$$b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta}_b)\, dx$$

b_{tot}, $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$ are nuisance parameters

The likelihood function

The single-channel likelihood function uses Poisson model for events in signal and control histograms:

data in signal histogram

data in control histogram

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

here signal rate is only parameter of interest

$\boldsymbol{\theta}$ represents all nuisance parameters, e.g., background rate, shapes

There is a likelihood $L_i(\mu, \boldsymbol{\theta}_i)$ for each channel, $i = 1, \ldots, N$.

The full likelihood function is $L(\mu, \boldsymbol{\theta}) = \prod_i L_i(\mu, \boldsymbol{\theta}_i)$

Profile likelihood ratio

To test hypothesized value of μ, construct profile likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

$\hat{\hat{\boldsymbol{\theta}}}$: Maximized L for given μ

$\hat{\mu}, \hat{\boldsymbol{\theta}}$: Maximized L

Equivalently use $q_\mu = -2 \ln \lambda(\mu)$:

data agree well with hypothesized $\mu \rightarrow q_\mu$ small

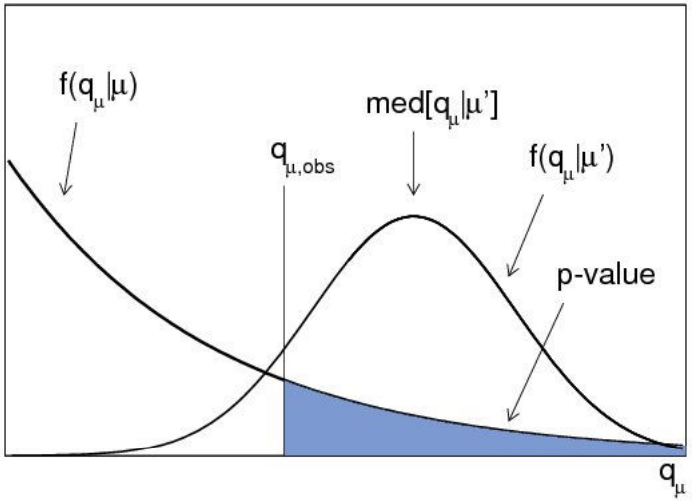
data disagree with hypothesized $\mu \rightarrow q_\mu$ large

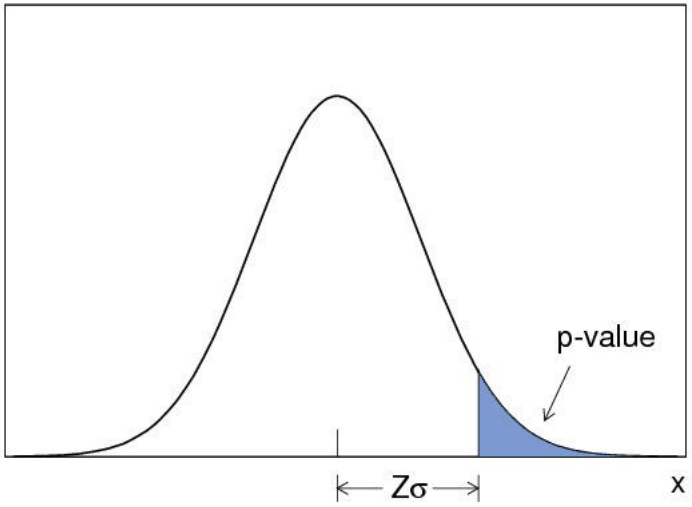
Distribution of q_μ under assumption of μ related to chi-square (Wilks' theorem, here approximation valid for roughly $L > 2$ fb^{-1}):

$$f(q_\mu|\mu) \approx \frac{1}{2} f_{\chi^2_1}(q_\mu) + \frac{1}{2}\delta(q_\mu)$$

p-value / significance of hypothesized μ

Test hypothesized μ by giving p-value, probability to see data with $\leq$ compatibility with μ compared to data observed:





Equivalently use significance, Z, defined as equivalent number of sigmas for a Gaussian fluctuation in one direction:

$$Z = \Phi^{-1}(1-p)$$

Sensitivity

Discovery:

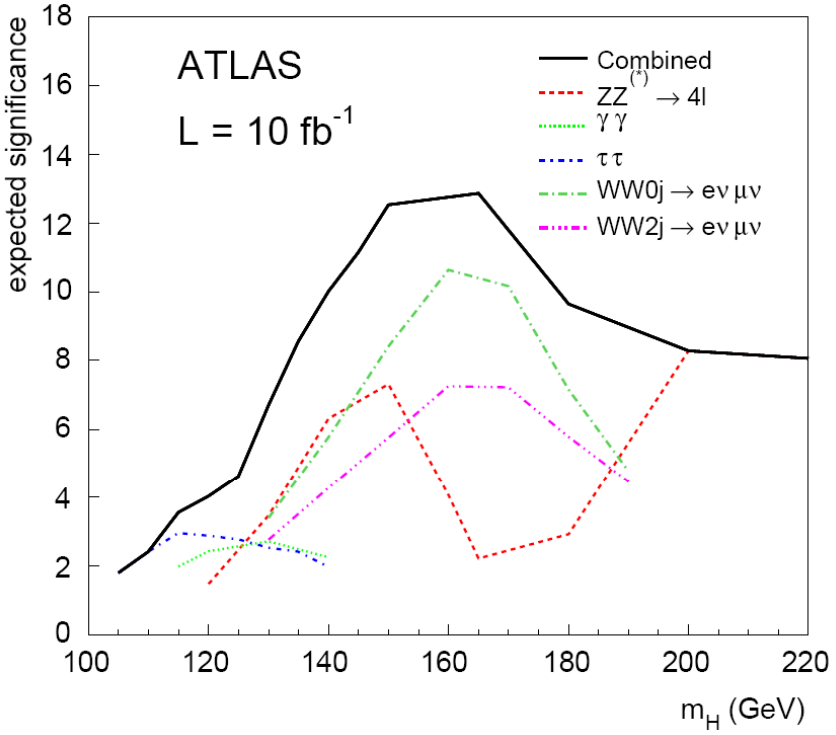
Generate data under s+b ($\mu = 1$) hypothesis;
Test hypothesis $\mu = 0 \rightarrow p$-value $\rightarrow Z$.

Exclusion:

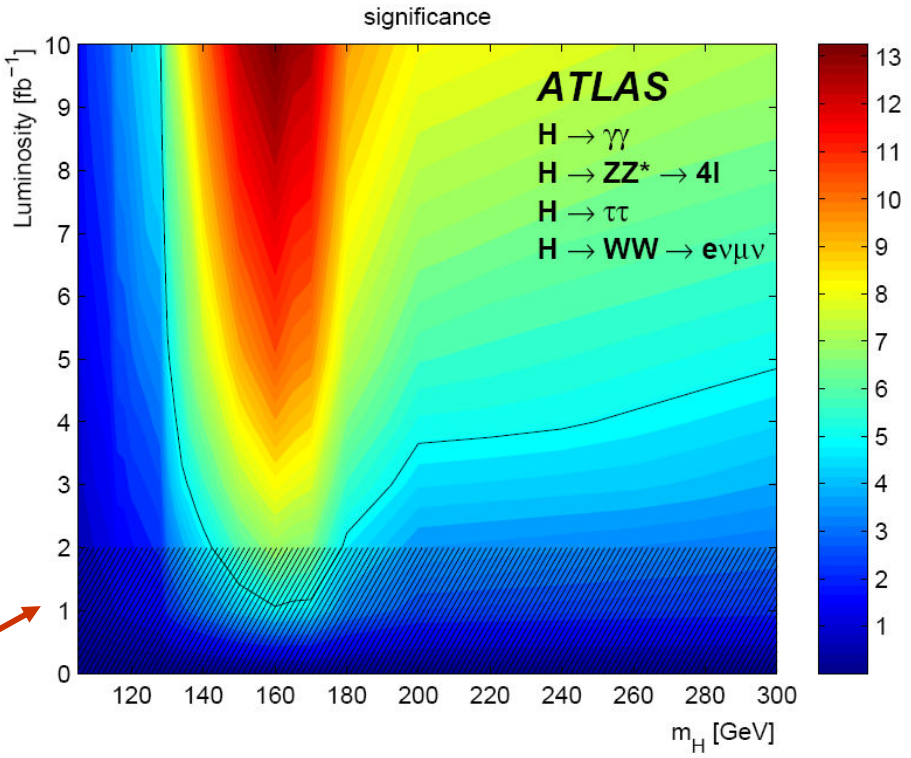
Generate data under background-only ($\mu = 0$) hypothesis;
Test hypothesis $\mu = 1$.
If $\mu = 1$ has p-value < 0.05 exclude m_{H} at 95% CL.

Presence of nuisance parameters leads to broadening of the profile likelihood, reflecting the loss of information, and gives appropriately reduced discovery significance, weaker limits.

Combined discovery significance



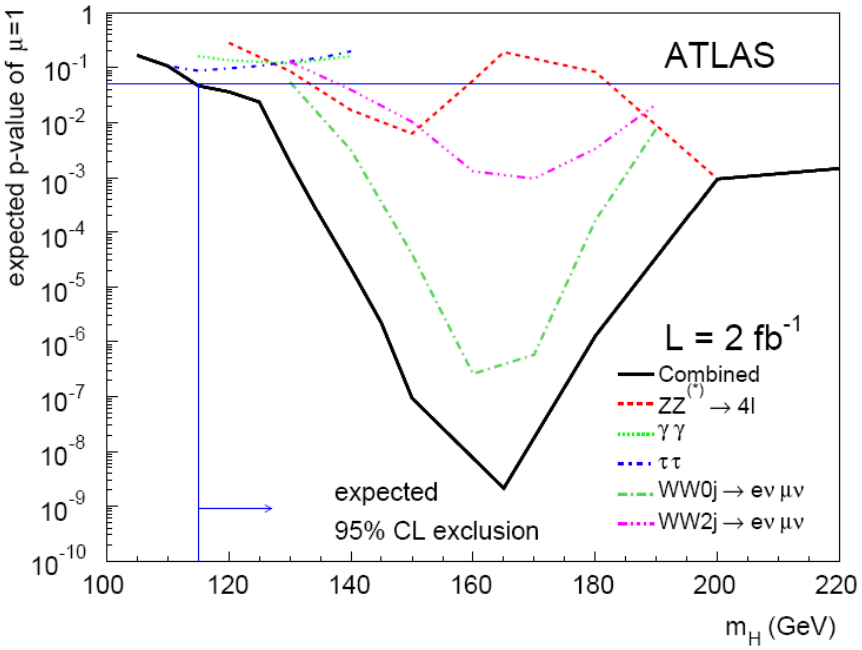
Discovery signficance (in colour) vs. L, m_H:

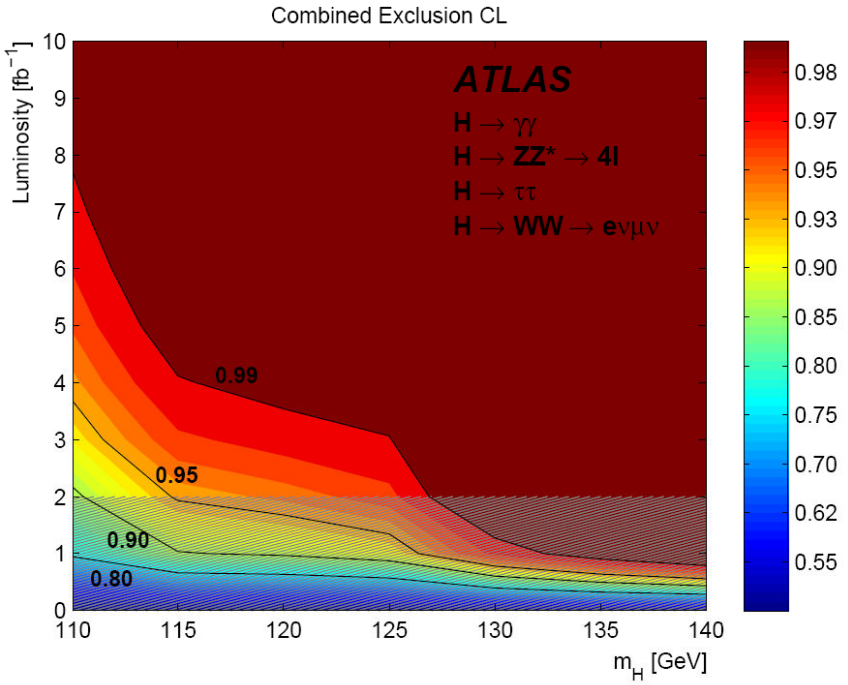


Approximations used here not always accurate for $L < 2$ fb^{-1} but in most cases conservative.

Combined 95% CL exclusion limits

1 – p-value of m_H (in colour) vs. L, m_H:





Summary 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?"

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

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

Summary on discovery

Current convention: p-value of background-only $< 2.9 \times 10^{-7}$ (5σ)

This should really depend also on other factors:

Plausibility of signal
Confidence in modeling of background

Can also use Bayes factor

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

Should hopefully point to same conclusion as p-value.

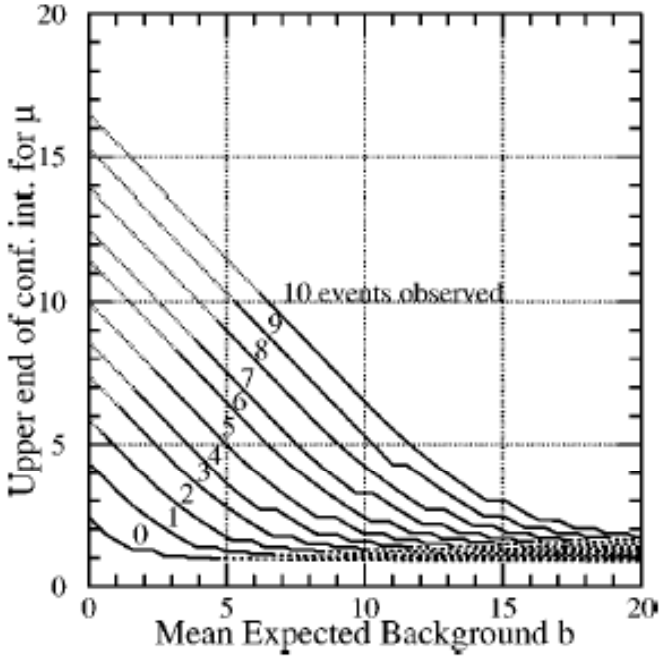
If not, need to understand why!

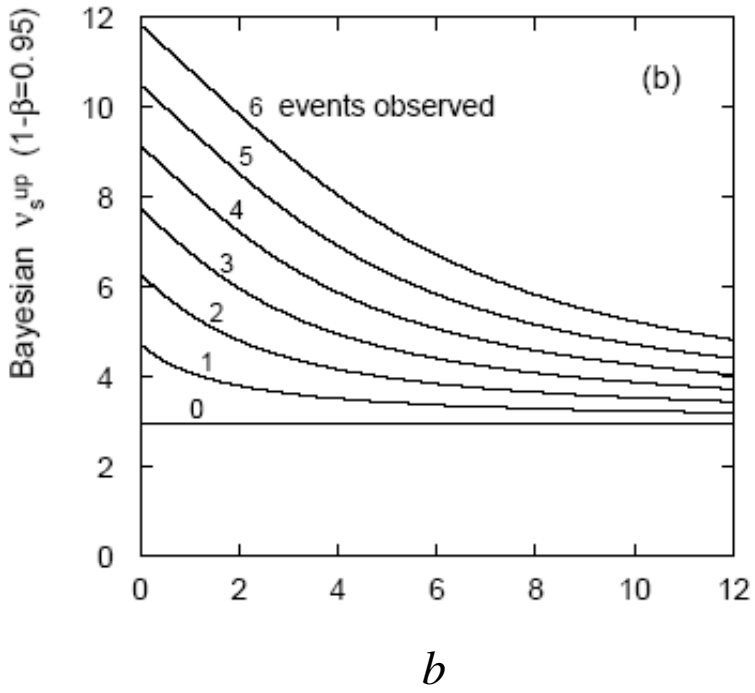
As yet not widely used in HEP, numerical issues not easy.

Extra slides

Upper limit versus *b*

Feldman & Cousins, PRD 57 (1998) 3873





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

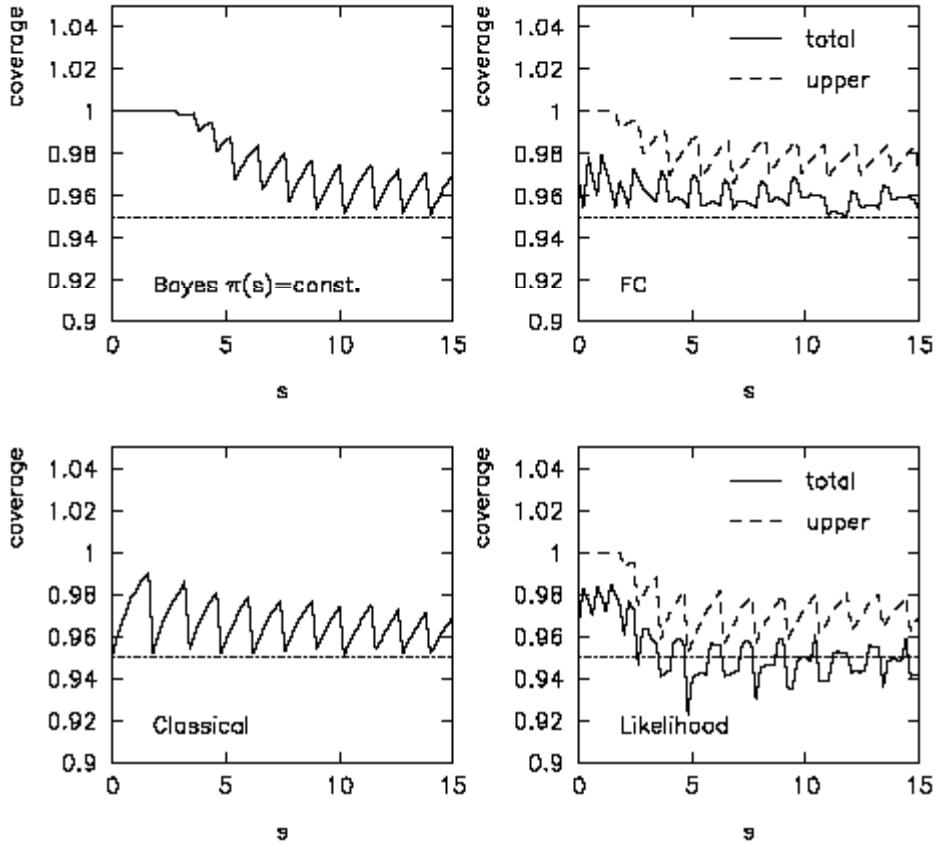
Classical: yes

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')



Cousins-Highland method

Regard b as 'random', characterized by pdf $\pi(b)$.

Makes sense in Bayesian approach, but in frequentist model b is constant (although unknown).

A measurement b_{meas} is random but this is not the mean number of background events, rather, b is.

Compute anyway $P(n;s) = \int P(n;s,b)\pi_b(b)\,db$

This would be the probability for n if Nature were to generate a new value of b upon repetition of the experiment with $\pi_b(b)$.

Now e.g. use this $P(n;s)$ in the classical recipe for upper limit at CL = $1 - \beta$: $\beta = P(n \leq n_{\text{obs}}; s_{\text{up}})$

Result has hybrid Bayesian/frequentist character.

'Integrated likelihoods'

Consider again signal s and background b, suppose we have uncertainty in b characterized by a prior pdf $\pi_b(b)$.

Define integrated likelihood as $L'(s) = \int L(s,b)\pi_b(b)\,db$,
also called modified profile likelihood, in any case not a real likelihood.

Now use this to construct likelihood ratio test and invert to obtain confidence intervals.

Feldman-Cousins & Cousins-Highland (FHC^2), see e.g. J. Conrad et al., Phys. Rev. D67 (2003) 012002 and Conrad/Tegenfeldt PHYSTAT05 talk.

Calculators available (Conrad, Tegenfeldt, Barlow).

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).

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,

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

posterior expectation

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).