

Statistical Tests and Limits

Lecture 3

IN2P3 School of Statistics

Autrans, France

17—21 May, 2010



Glen Cowan

Physics Department

Royal Holloway, University of London

g.cowan@rhul.ac.uk

www.pp.rhul.ac.uk/~cowan

Outline

Lecture 1: General formalism

Definition and properties of a statistical test

Significance tests (and goodness-of-fit) , p -values

Lecture 2: Setting limits

Confidence intervals

Bayesian Credible intervals



Lecture 3: Further topics for tests and limits

More on intervals/limits, CLs

Systematics / nuisance parameters

Bayesian model selection

Likelihood ratio limits (Feldman-Cousins)

Define likelihood ratio for hypothesized parameter, e.g., for expected number of signal events 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 $0 \leq l(s) \leq 1$.

Critical region defined by low values of likelihood ratio.

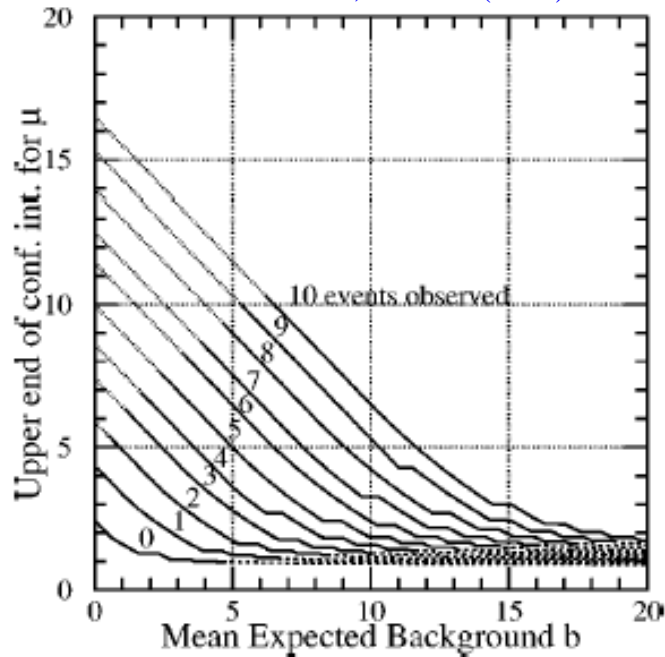
Resulting intervals can be one- or two-sided (depending on n).

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

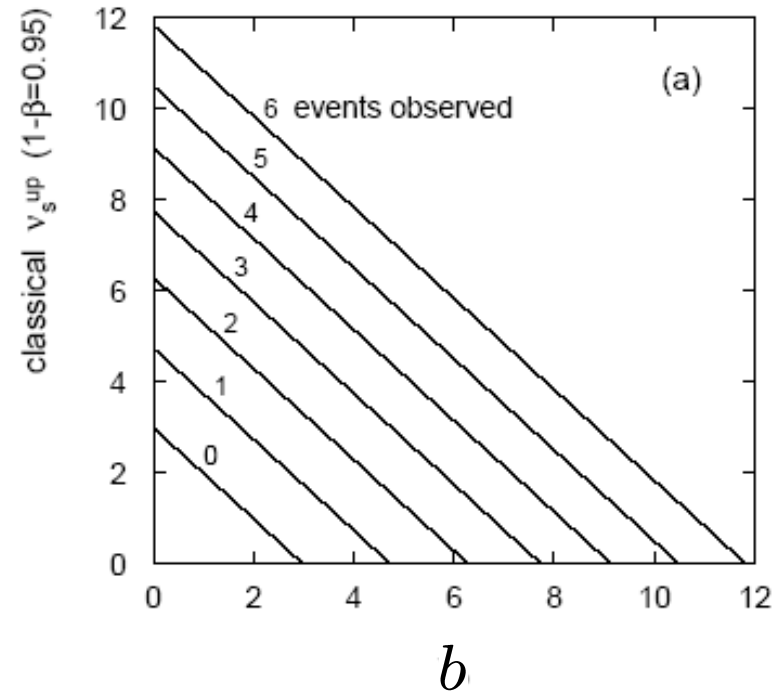
Upper limit versus b

Feldman-Cousins

Feldman & Cousins, PRD 57 (1998) 3873



“Classical”



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

Classical: yes

FC: yes, but less so

More on intervals from LR test (Feldman-Cousins)

Caveat with coverage: suppose we find $n \gg b$.

Usually one then quotes a measurement: $\hat{s} = n - b$, $\hat{\sigma}_{\hat{s}} = \sqrt{n}$

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'? For this reason, one-sided intervals for limits still popular.

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

$$0.95 = \int_{-\infty}^{\text{sup}} p(s|n) ds$$

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

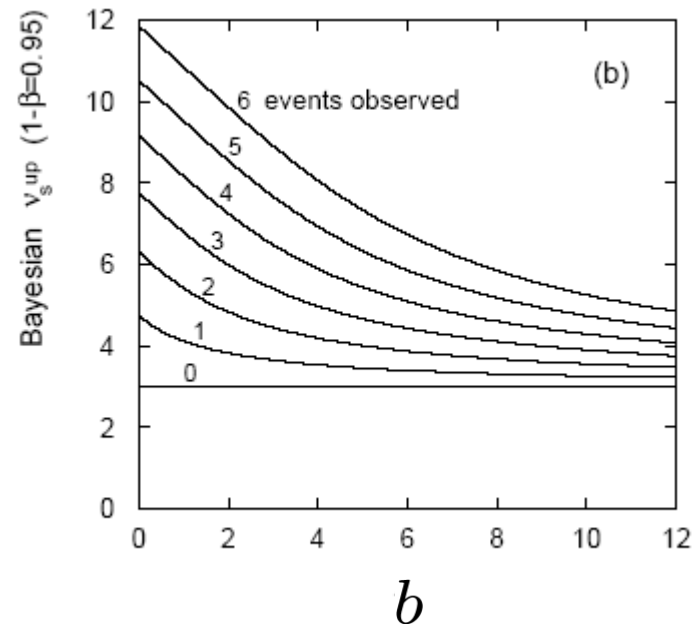
Solve numerically to find limit s_{up} .

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 $n = 0$.



Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called “objective priors”

Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In a Subjective Bayesian analysis, using objective priors can be an important part of the sensitivity analysis.

Priors from formal rules (cont.)

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes' theorem as a recipe to produce an interval with certain coverage properties. For a review see:

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in HEP, but there is recent interest in this direction; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, arxiv:1002.1111 (Feb 2010)

Jeffreys' prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E \left[\frac{\partial^2 \ln L(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] = - \int \frac{\partial^2 \ln L(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} L(\mathbf{x}|\boldsymbol{\theta}) dx$$

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters.

For a Gaussian mean, the Jeffreys' prior is constant; for a Poisson mean μ it is proportional to $1/\sqrt{\mu}$.

Jeffreys' prior for Poisson mean

Suppose $n \sim \text{Poisson}(\mu)$. To find the Jeffreys' prior for μ ,

$$L(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \qquad \frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\mu}$$

$$I = -E \left[\frac{\partial^2 \ln L}{\partial \mu^2} \right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

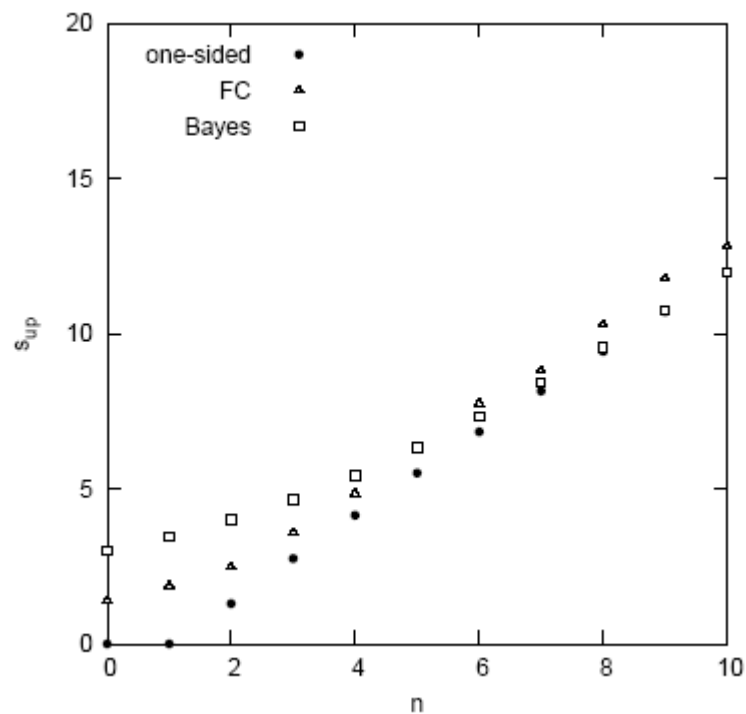
$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for $\mu = s + b$, this means the prior $\pi(s) \sim 1/\sqrt{s + b}$, which depends on b . But this is not designed as a degree of belief about s .

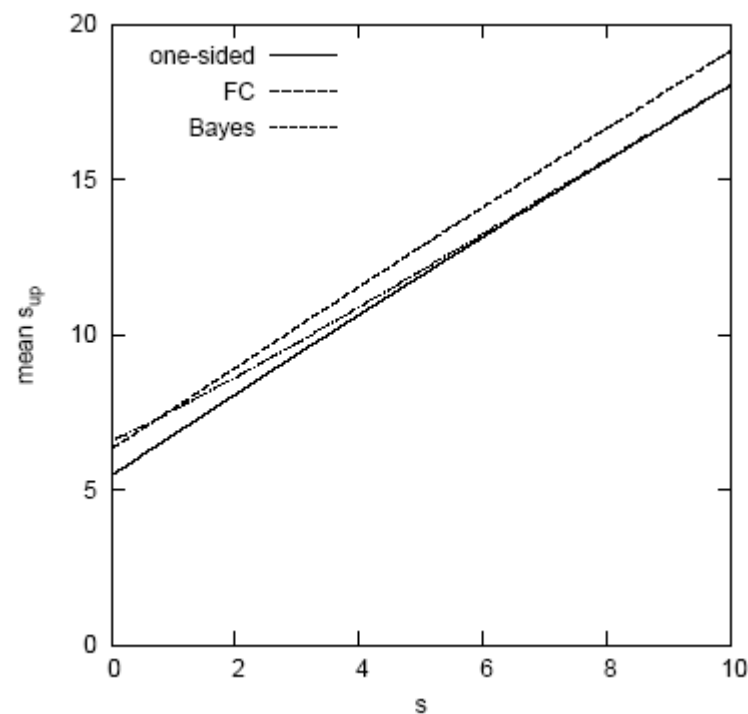
Properties of upper limits

Example: take $b = 5.0$, $1 - \gamma = 0.95$

Upper limit s_{up} vs. n

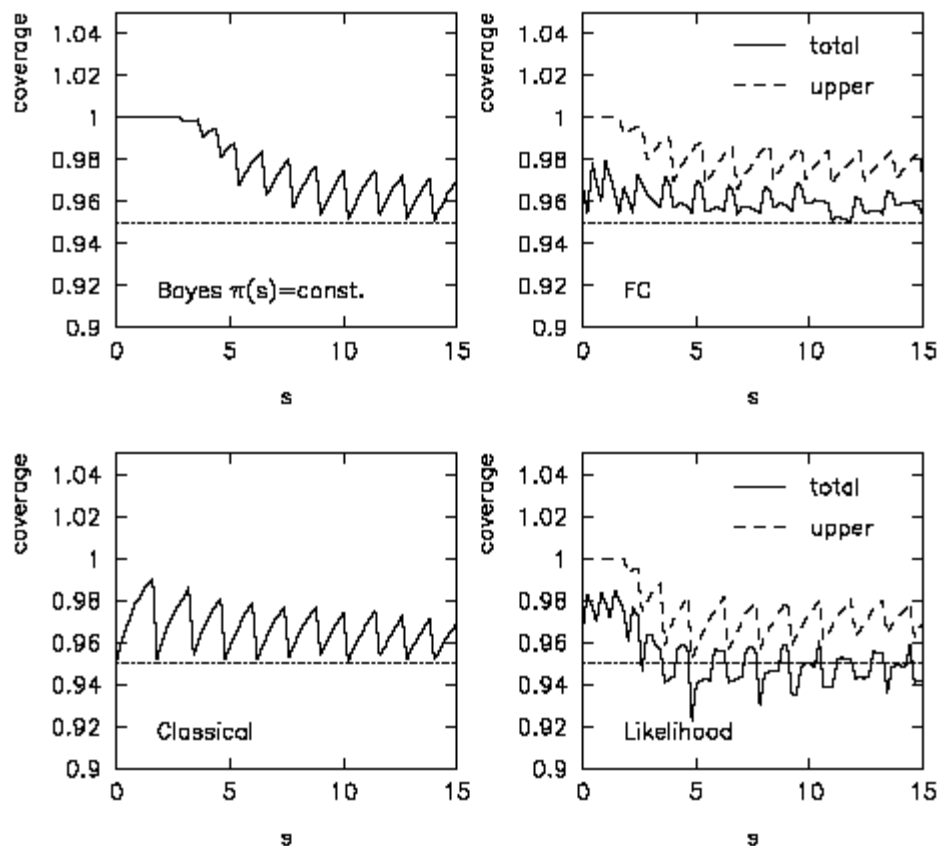


Mean upper limit vs. s



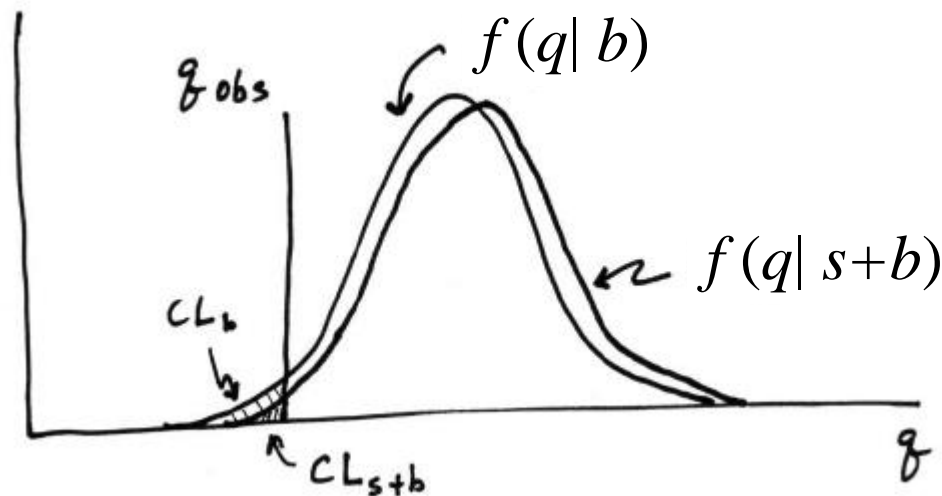
Coverage probability of intervals

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



The “ CL_s ” issue

When the cross section for the signal process becomes small (e.g., large Higgs mass), the distribution of the test variable used in a search becomes the same under both the b and $s+b$ hypotheses:



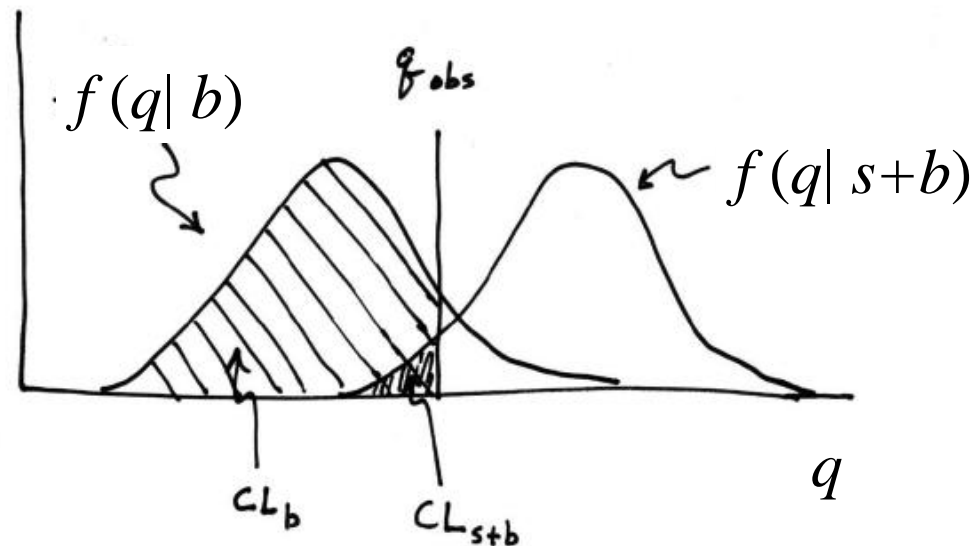
In such a case we will reject the signal hypothesis with a probability approaching $\alpha = 1 - CL$ (i.e. 5%) assuming no signal.

The CL_s solution

The CL_s solution (A. Read et al.) is to base the test not on the usual p -value (CL_{s+b}), but rather to divide this by CL_b (one minus the background of the b -only hypothesis, i.e.,

Define:

$$CL_s = \frac{CL_{s+b}}{CL_b} \\ = \frac{p_{s+b}}{1 - p_b}$$



Reject signal hypothesis if:

$$CL_s \leq \alpha$$

Reduces “effective” p -value when the two distributions become close (prevents exclusion if sensitivity is low).

CL_s discussion

In the CLs method the p-value is reduced according to the recipe

$$p_{\mu} \rightarrow \frac{P_{\mu}}{1 - p_b}$$

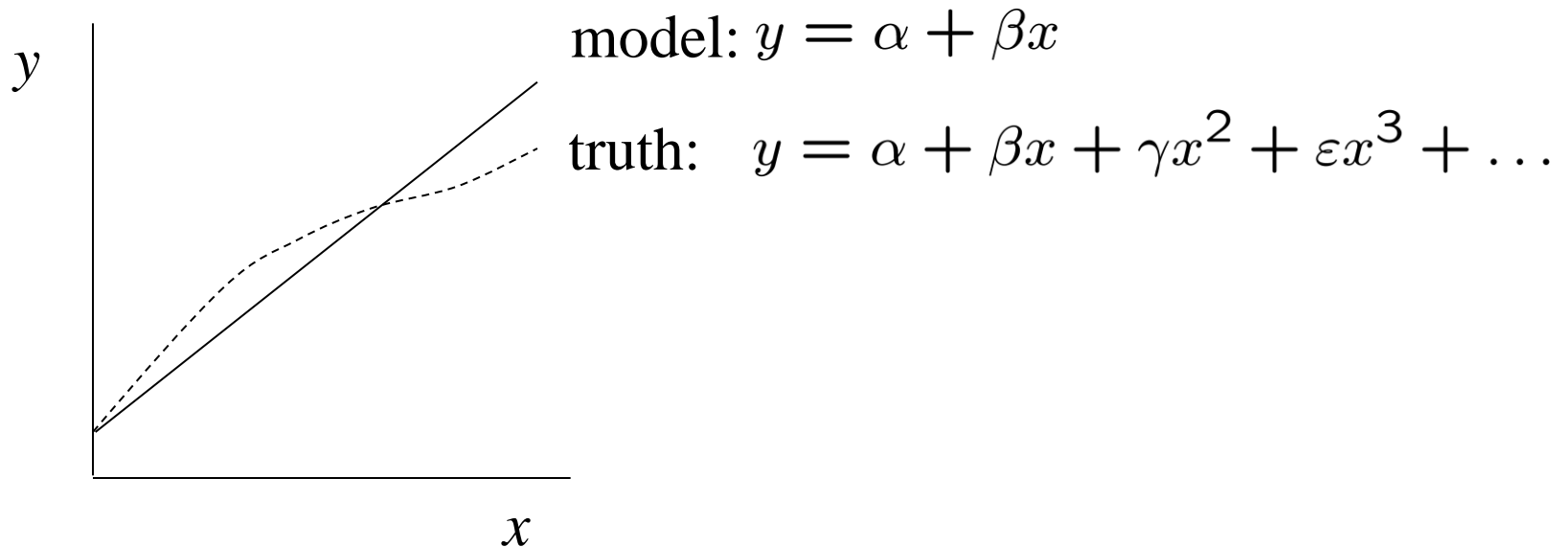
Statistics community does not smile upon ratio of p-values; would prefer to regard parameter μ as excluded if:

- (a) p-value of $\mu < 0.05$
- (b) power of test of μ with respect to background-only > some threshold (0.5?)

Needs study. In any case should produce CLs result for purposes of comparison with other experiments.

Systematic errors and nuisance parameters

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



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

systematic uncertainty \leftrightarrow nuisance parameters

Nuisance parameters and limits

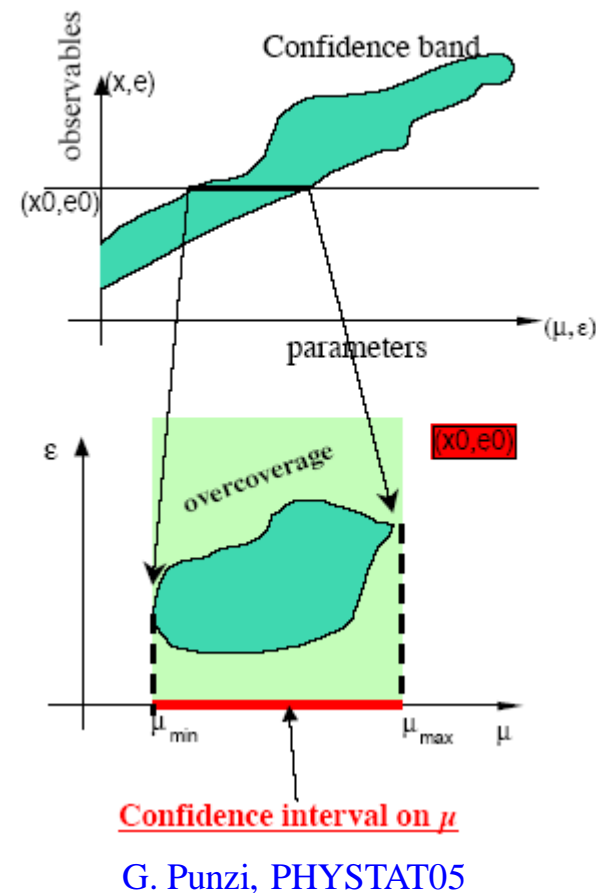
In general we don't know the background b perfectly.

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 exact confidence interval recipe can be generalized to multiple parameters, minimum coverage guaranteed.

Difficult because of overcoverage; see e.g. talks by K. Cranmer at PHYSTAT03 and by G. Punzi at PHYSTAT05.



Nuisance parameters in limits (2)

Connect systematic to nuisance parameters ν . Then form e.g.

Profile likelihood:
$$L_p(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \hat{\nu})$$

Marginal likelihood:
$$L_m(\boldsymbol{\theta}) = \int L(\boldsymbol{\theta}, \nu) \pi(\nu) d\nu$$

and use these to construct e.g. likelihood ratios for tests.

Coverage not guaranteed for all values of the nuisance params.

Results of both approaches above often similar, but some care is needed in writing down prior; this should truly reflect one's degree of belief about the parameters.

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

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

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)

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


rapidly approaches chi-square pdf
(Wilks' theorem)

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

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

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

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

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 $\text{CL} = 1 - \beta$: $\beta = P(n \leq n_{\text{obs}}; s_{\text{up}})$

Result has hybrid Bayesian/frequentist character.

Marginal likelihood in LR tests

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

Define marginal 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 (FHC²), see e.g. J. Conrad et al., Phys. Rev. D67 (2003) 012002 and Conrad/Tegenfeldt PHYSTAT05 talk.

Calculators available (Conrad, Tegenfeldt, Barlow).

Comment on profile likelihood

Suppose originally we measure x , likelihood is $L(x|\theta)$.

To cover a systematic, we enlarge model to include a nuisance parameter ν , new model is $L(x|\theta, \nu)$.

To use profile likelihood, data must constrain the nuisance parameters, otherwise suffer loss of accuracy in parameters of interest.

Can e.g. use a separate measurement to constrain ν , e.g., with likelihood $L(y|\nu)$. This becomes part of the full likelihood, i.e.,

$$L(x, y|\theta, \nu) = L(x|\theta, \nu)L(y|\nu)$$

Comment on marginal likelihood

When using a prior to reflect knowledge of ν , often one treats this as coming from the measurement y , i.e.,

$$\pi(\nu) \propto L(y|\nu)\pi_0(\nu)$$

 original prior,

Then the marginal likelihood is

$$L_m(\theta) = \int L(x|\theta, \nu)\pi(\nu) d\nu$$

So here L in the integrand does not include the information from the measurement y ; this is included in the prior.

Bayesian limits with uncertainty on b

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

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

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

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

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.

Framework for treatment of nuisance parameters well defined; choice of prior can still be problematic, but often less so than finding a “non-informative” prior for a parameter of interest.

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.

Google for ‘MCMC’, ‘Metropolis’, ‘Bayesian computation’, ...


MCMC generates **correlated** sequence of random numbers:
cannot use for many applications, e.g., detector MC;
effective stat. error greater than \sqrt{n} .


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


Bayesian model selection ('discovery')


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

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

no Higgs 

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.

Will this be adopted in HEP?

Rewriting the Bayes factor

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

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

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

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:

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

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

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

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.

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 $L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

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 $L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

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

p-values versus Bayes factors

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.

Summary

Bayesian approach to setting limits is straightforward; all information about the parameter is in the posterior probability, integrate this to get intervals with given probability.

Difficult to find appropriate “non-informative” prior.
Often use Bayesian approach as a recipe for producing interval, then study it in a frequentist way (e.g. coverage)

The key to treating systematic uncertainties is to include in the model enough parameters so that it is correct (or very close).

But too many parameters degrades information on parameters of interest

Bayesian model selection

Bayes factor = posterior odds if prior odds = 1.
Only requires priors for internal parameters of models.
Can be very difficult to compute numerically.