Statistical Methods for Discovery and Limits in HEP Experiments

Day 3: Exclusion Limits

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

Vorlesungen des GK *Physik an Hadron-Beschleunigern*, Freiburg, 27-29 June, 2011

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

Outline

Day 1: Introduction and basic formalism
Probability, statistical tests, parameter estimation.

Day 2: Discovery
Quantifying discovery significance and sensitivity
Systematic uncertainties (nuisance parameters)

→ Day 3: Exclusion limits
Frequentist and Bayesian intervals/limits

Interval estimation — introduction

In addition to a 'point estimate' of a parameter we should report an interval reflecting its statistical uncertainty.

Desirable properties of such an interval may include:

communicate objectively the result of the experiment;
have a given probability of containing the true parameter;
provide information needed to draw conclusions about the parameter possibly incorporating stated prior beliefs.

Often use +/− the estimated standard deviation of the estimator.
In some cases, however, this is not adequate:

estimate near a physical boundary,
e.g., an observed event rate consistent with zero.

We will look briefly at Frequentist and Bayesian intervals.

Frequentist confidence intervals

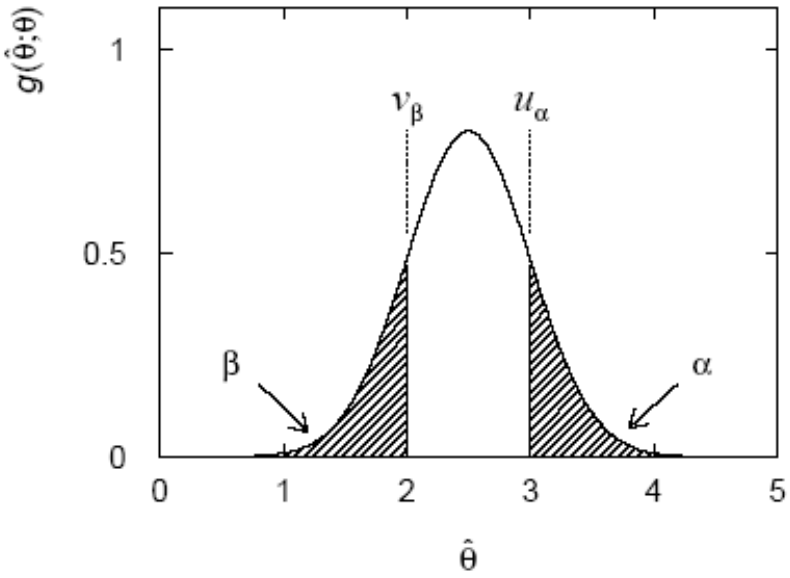
Consider an estimator $\hat{\theta}$ for a parameter θ and an estimate $\hat{\theta}_{\text{obs}}$.

We also need for all possible θ its sampling distribution $g(\hat{\theta};\theta)$.

Specify upper and lower tail probabilities, e.g., $\alpha = 0.05$, $\beta = 0.05$, then find functions $u_\alpha(\theta)$ and $v_\beta(\theta)$ such that:

$$\begin{aligned}\alpha &= P(\hat{\theta} \geq u_\alpha(\theta)) \\ &= \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\, d\hat{\theta}\end{aligned}$$

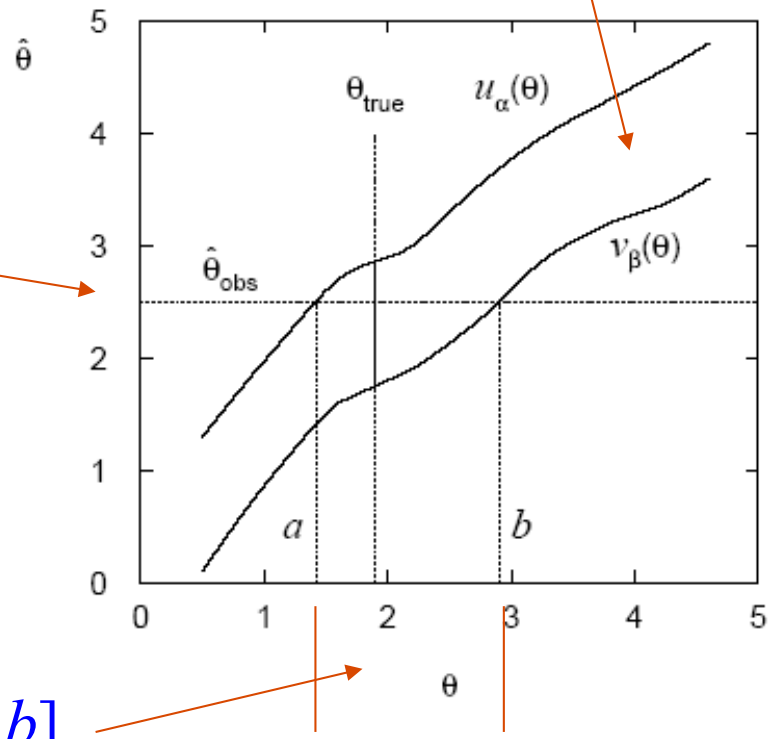
$$\begin{aligned}\beta &= P(\hat{\theta} \leq v_\beta(\theta)) \\ &= \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\, d\hat{\theta}\end{aligned}$$



Confidence interval from the confidence belt

The region between $u_\alpha(\theta)$ and $v_\beta(\theta)$ is called the confidence belt.

Find points where observed estimate intersects the confidence belt.



This gives the confidence interval $[a, b]$

Confidence level = $1 - \alpha - \beta$ = probability for the interval to cover true value of the parameter (holds for any possible true θ).

Confidence intervals by inverting a test

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

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

Now invert the test to define a confidence interval as:

> set of θ 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 θ with probability $\geq 1 - \gamma$.

Equivalent to confidence belt construction; confidence belt is acceptance region of a test.

Relation between confidence interval and p-value

Equivalently we can consider a significance test for each hypothesized value of θ, resulting in a p-value, p_θ..

If $p_\theta < \gamma$, then we reject θ.

The confidence interval at CL = $1 - \gamma$ consists of those values of θ that are not rejected.

E.g. an upper limit on θ is the greatest value for which $p_\theta \geq \gamma$.

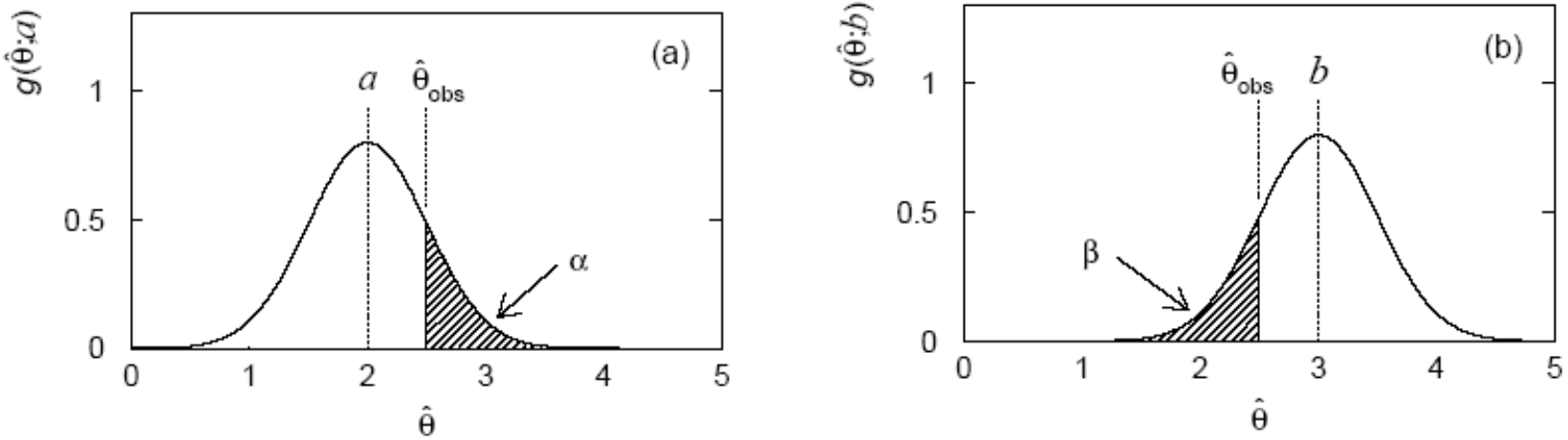
In practice find by setting $p_\theta = \gamma$ and solve for θ.

Confidence intervals in practice

The recipe to find the interval $[a, b]$ boils down to solving

$$\alpha = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\, d\hat{\theta} = \int_{\hat{\theta}_{\rm obs}}^{\infty} g(\hat{\theta};a)\, d\hat{\theta}\,,$$

$$\beta = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\, d\hat{\theta} = \int_{-\infty}^{\hat{\theta}_{\rm obs}} g(\hat{\theta};b)\, d\hat{\theta}\,.$$



→ a is hypothetical value of θ such that $P(\hat{\theta} > \hat{\theta}_{\rm obs}) = \alpha$.

→ b is hypothetical value of θ such that $P(\hat{\theta} < \hat{\theta}_{\rm obs}) = \beta$.

Meaning of a confidence interval

N.B. the interval is random, the true θ is an unknown constant.

Often report interval $[a, b]$ as $\hat{\theta}^{+d}_{-c}$, i.e. $c = \hat{\theta} - a$, $d = b - \hat{\theta}$.

So what does $\hat{\theta} = 80.25^{+0.31}_{-0.25}$ mean? It does not mean:

$P(80.00 < \theta < 80.56) = 1 - \alpha - \beta$, but rather:

repeat the experiment many times with same sample size,

construct interval according to same prescription each time,

in $1 - \alpha - \beta$ of experiments, interval will cover θ.

Setting limits on Poisson parameter

Consider again the case of finding $n = n_s + n_b$ events where

n_b events from known processes (background)
n_s events from a new process (signal)

are Poisson r.v.s with means s, b, and thus $n = n_s + n_b$ is also Poisson with mean = $s + b$. Assume b is known.

Suppose we are searching for evidence of the signal process, but the number of events found is roughly equal to the expected number of background events, e.g., $b = 4.6$ and we observe $n_{obs} = 5$ events.

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

→ set upper limit on the parameter s.

Upper limit for Poisson parameter

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:

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

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

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

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

Calculating Poisson parameter limits

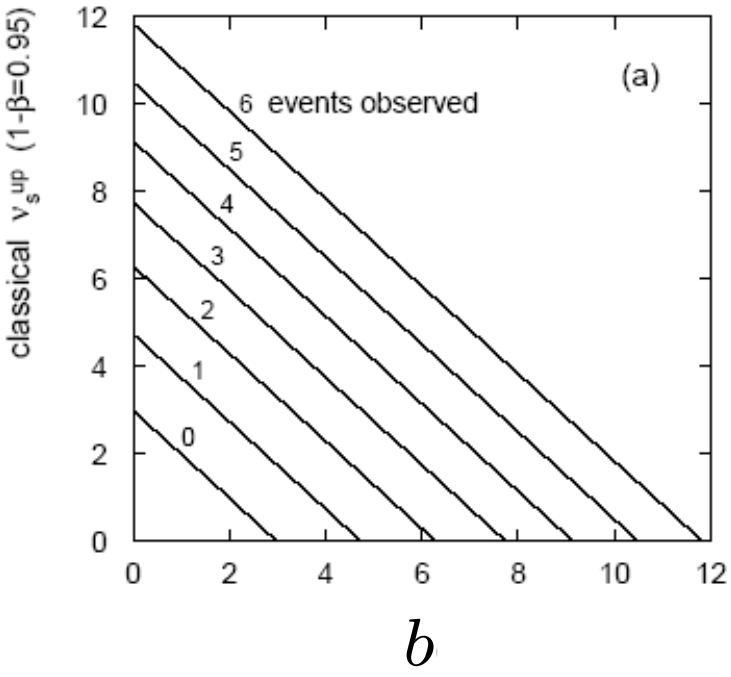
To solve for $s_{\rm lo}$, $s_{\rm up}$, can exploit relation to χ^2 distribution:

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

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.

Expected limit for $s = 0$

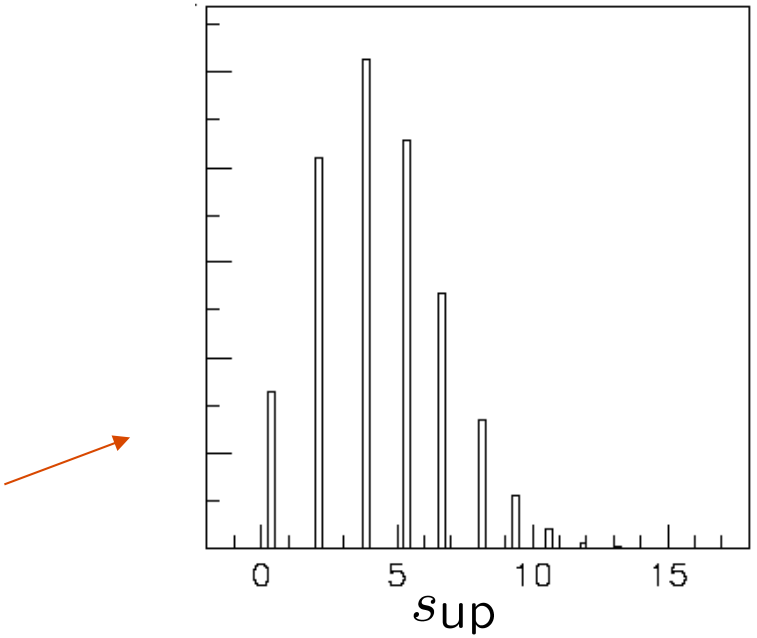
Physicist: I should have used CL = 0.95 — then $s_{up} = 0.496$

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



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. $n \sim$ Poisson(s+b), 95% CL upper limit on s from

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

Bayesian prior for Poisson parameter

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

Could 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

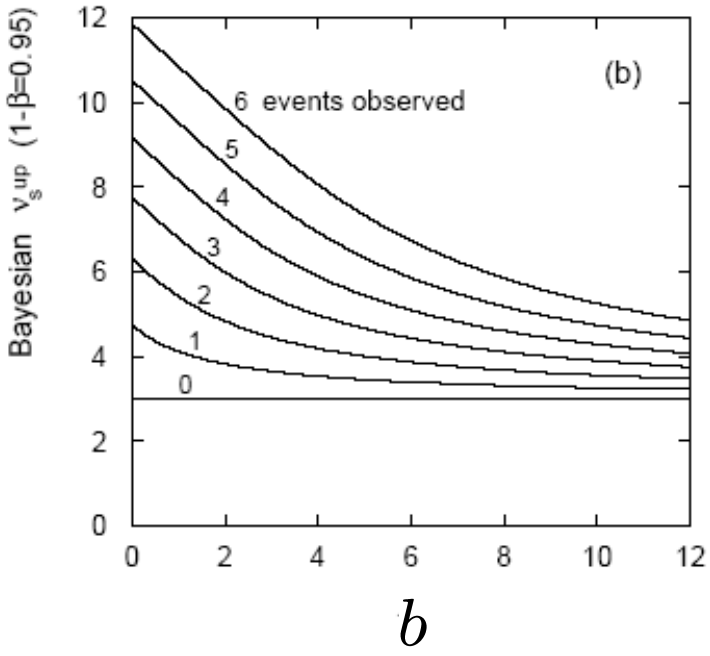
Solve numerically to find limit s_{up}.

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*, Phys. Rev. D 82 (2010) 034002, 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

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

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.

Bayesian limits with uncertainty on b

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

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

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

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

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.

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.

Comment on priors

Suppose we measure $n \sim \text{Poisson}(s+b)$, goal is to make inference about s.

Suppose b is not known exactly but we have an estimate $\hat{b}$ with uncertainty σ_b.

For Bayesian analysis, first reflex may be to write down a Gaussian prior for b,

$$\pi(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-\hat{b})^2/\sigma_b^2}$$

But a Gaussian could be problematic because e.g.

$b \geq 0$, so need to truncate and renormalize;
tails fall off very quickly, may not reflect true uncertainty.

Gamma prior for b

What is in fact our prior information about b? It may be that we estimated b using a separate measurement (e.g., background control sample) with

$m \sim \text{Poisson}(\tau b)$ (τ = scale factor, here assume known)

Having made the control measurement we can use Bayes' theorem to get the probability for b given m,

$$\pi(b|m) \propto P(m|b)\pi_0(b) \propto \frac{(\tau b)^m}{m!} e^{-\tau b} \pi_0(b)$$

If we take the "original" prior $\pi_0(b)$ to be to be constant for $b \geq 0$, then the posterior $\pi(b|m)$, which becomes the subsequent prior when we measure n and infer s, is a Gamma distribution with:

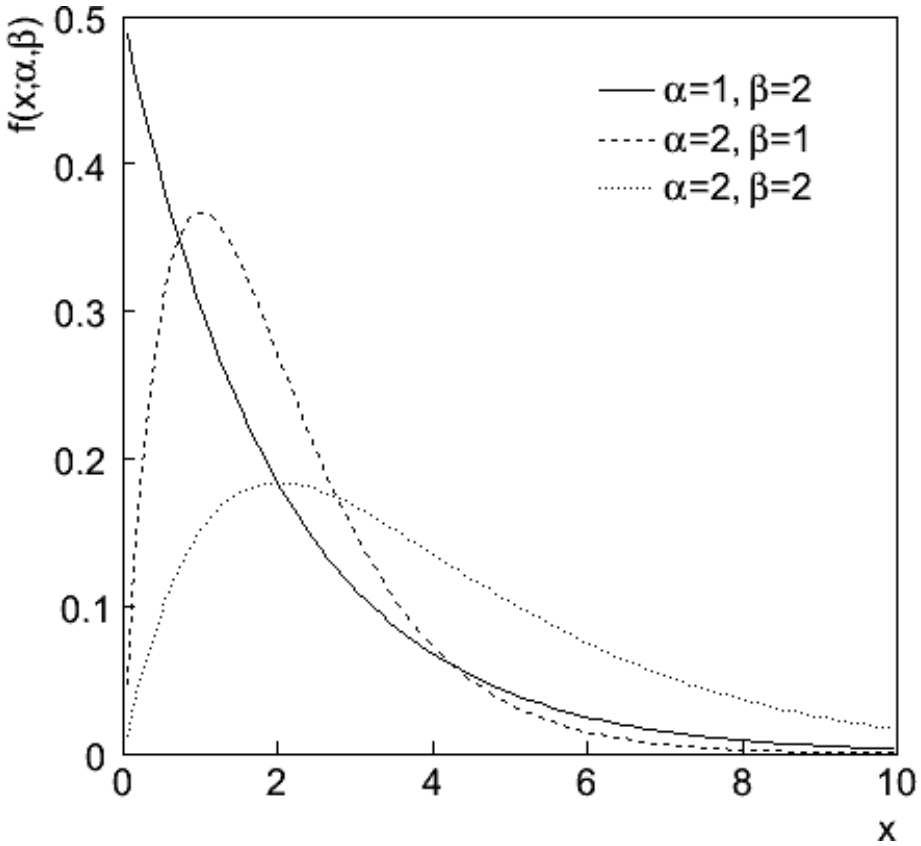
mean = $(m + 1)/\tau$
standard dev. = $\sqrt{(m + 1)}/\tau$

Gamma distribution

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\, x^{\alpha-1}\, e^{-x/\beta}$$

$$E[x] = \alpha\beta$$

$$V[x] = \alpha\beta^2$$



Frequentist approach to same problem

In the frequentist approach we would regard both variables

$n \sim \text{Poisson}(s+b)$
$m \sim \text{Poisson}(\tau b)$

as constituting the data, and thus the full likelihood function is

$$L(s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Use this to construct test of s with e.g. profile likelihood ratio

$$\lambda(s) = \frac{L(s, \hat{\hat{b}})}{L(\hat{s}, \hat{b})}$$

Note here that the likelihood refers to both n and m, whereas the likelihood used in the Bayesian calculation only modeled n.

Choice of test for limits

Often we want to ask what values of μ can be excluded on the grounds that the implied rate is too high relative to what is observed in the data.

To do this take the alternative to correspond to lower values of μ.

The critical region to test μ thus contains low values of the data.

→ One-sided (e.g., upper) limit.

In other cases we want to exclude μ on the grounds that some other measure of incompatibility between it and the data exceeds some threshold (e.g., likelihood ratio wrt two-sided alternative).

The critical region can contain both high and low data values.

→ Two-sided or unified (Feldman-Cousins) intervals.

Test statistic for upper limits

For purposes of setting an upper limit on μ use

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \end{cases} \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

I.e. for purposes of setting an upper limit, one does not regard an upwards fluctuation of the data as representing incompatibility with the hypothesized μ.

From observed q_μ find p-value: $$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

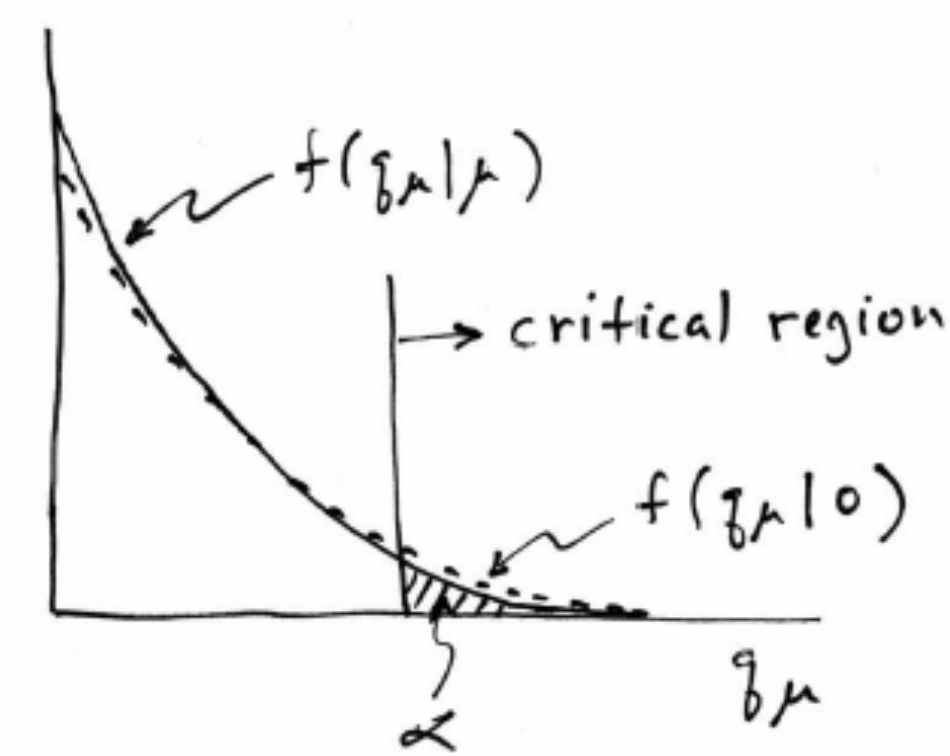
Large sample approximation: $$p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

95% CL upper limit on μ is highest value for which p-value is not less than 0.05.

Low sensitivity to μ

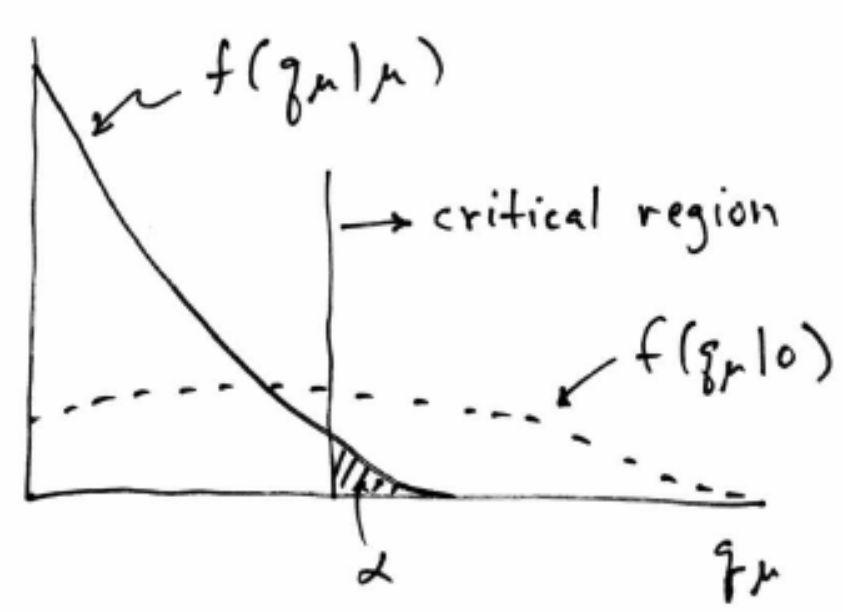
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

This means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ will be almost the same:



Having sufficient sensitivity

In contrast, having sensitivity to μ means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ are more separated:

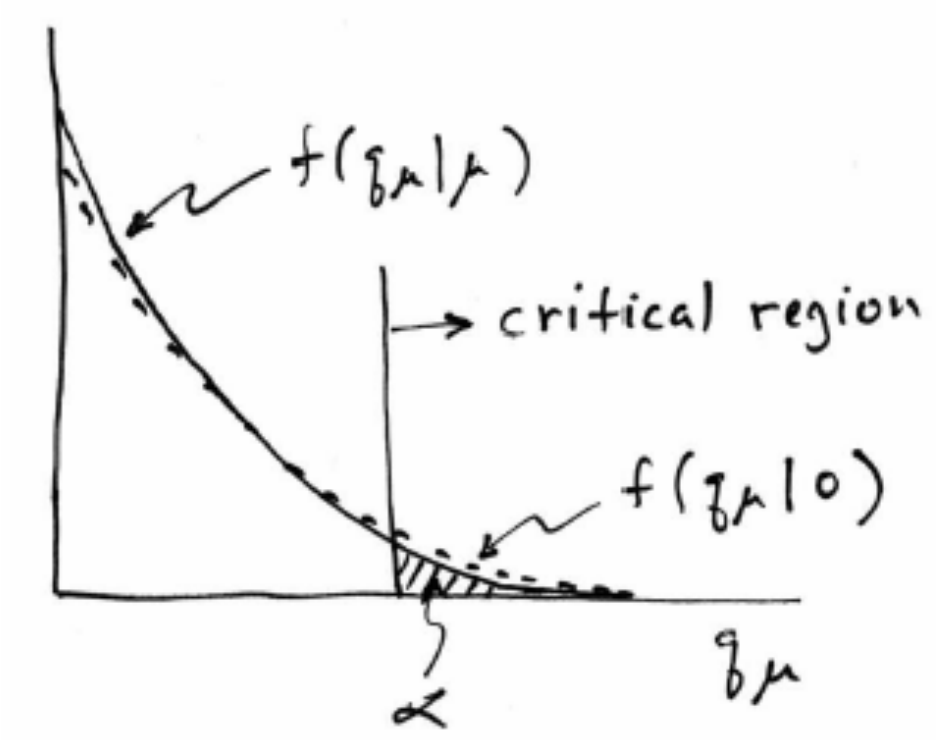


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α. We use this power as a measure of the sensitivity.

Spurious exclusion

Consider again the case of low sensitivity. By construction the probability to reject μ if μ is true is α (e.g., 5%).

And the probability to reject μ if $\mu = 0$ (the power) is only slightly greater than α.



This means that with probability of around $\alpha = 5\%$ (slightly higher), one excludes hypotheses to which one has essentially no sensitivity (e.g., $m_{\mathbf{H}} = 1000$ TeV).

"Spurious exclusion"

Ways of addressing spurious exclusion

The problem of excluding parameter values to which one has no sensitivity known for a long time; see e.g.,

Virgil L. Highland, *Estimation of Upper Limits from Experimental Data*, July 1986, Revised February 1987, Temple University Report C00-3539-38.

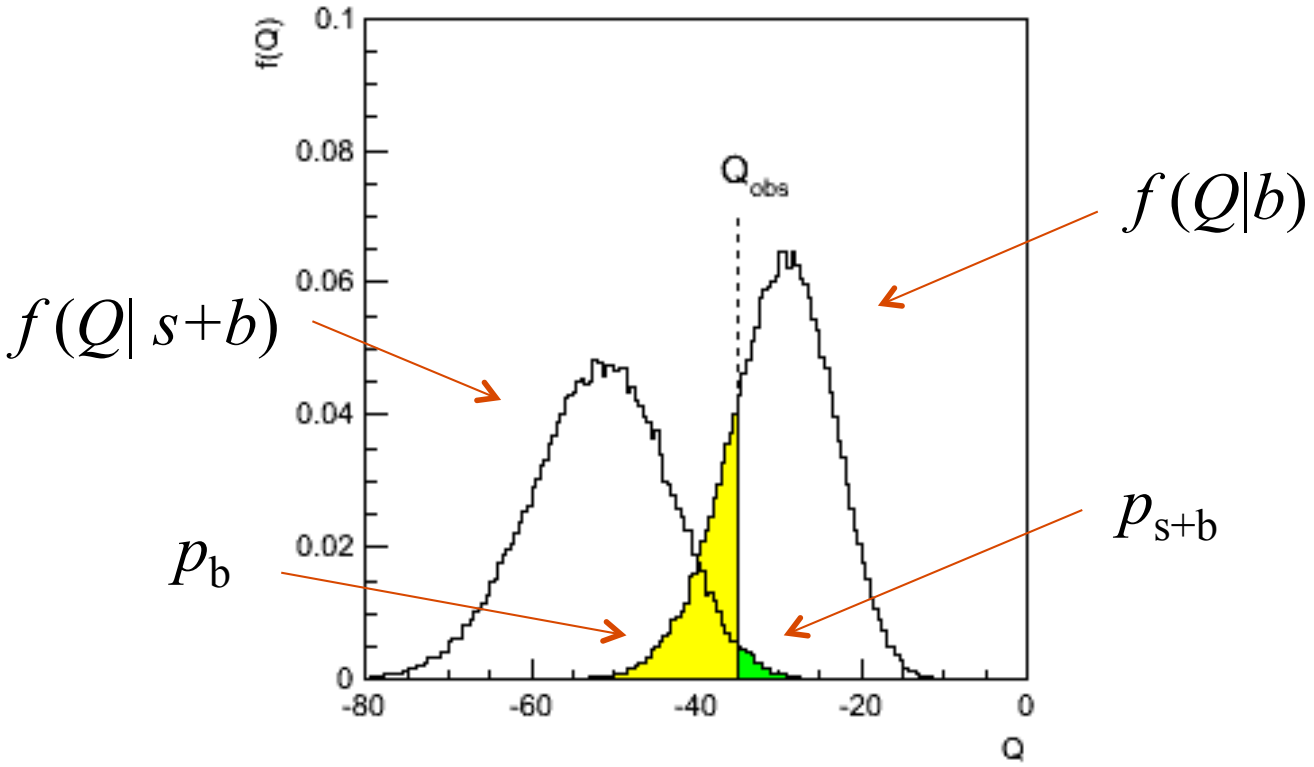
In the 1990s this was re-examined for the LEP Higgs search by Alex Read and others

T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

and led to the "CL_s" procedure.

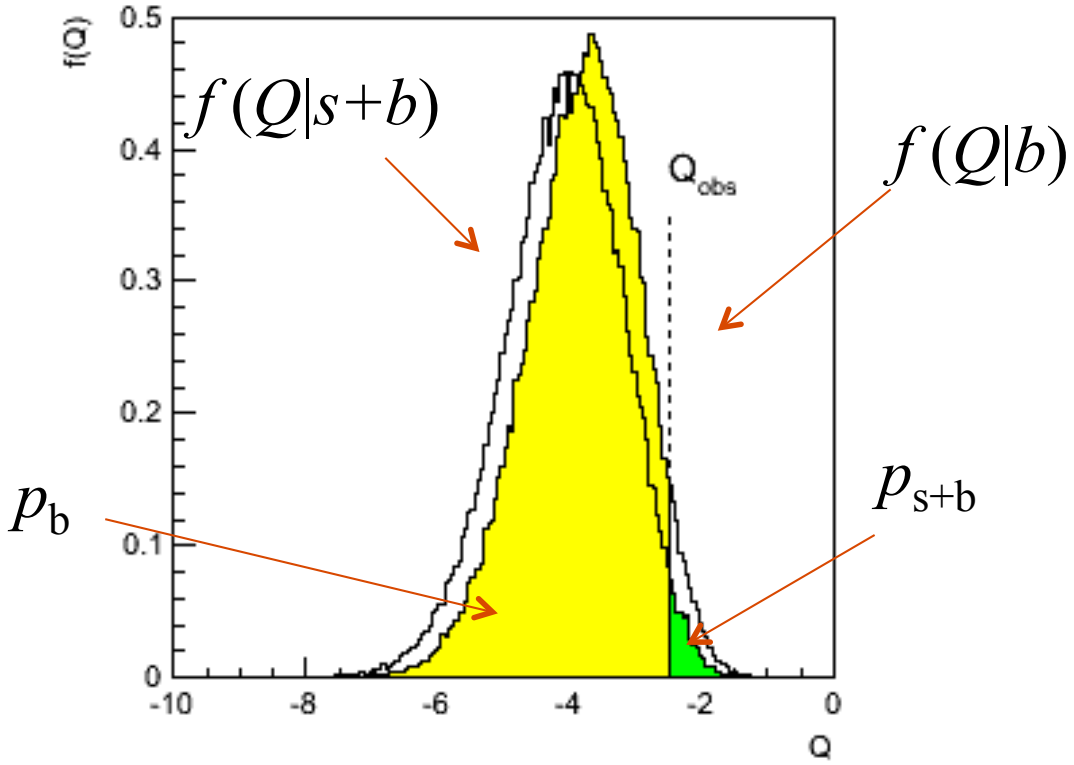
The CL_s procedure

In the usual formulation of CL_s, one tests both the $\mu = 0$ (b) and $\mu = 1$ (s+b) hypotheses with the same statistic $Q = -2\ln L_{s+b}/L_b$:



The CL_s procedure (2)

As before, "low sensitivity" means the distributions of Q under b and $s+b$ are very close:



The CL_s procedure (3)

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 *p*-value of the *b*-only hypothesis, i.e.,

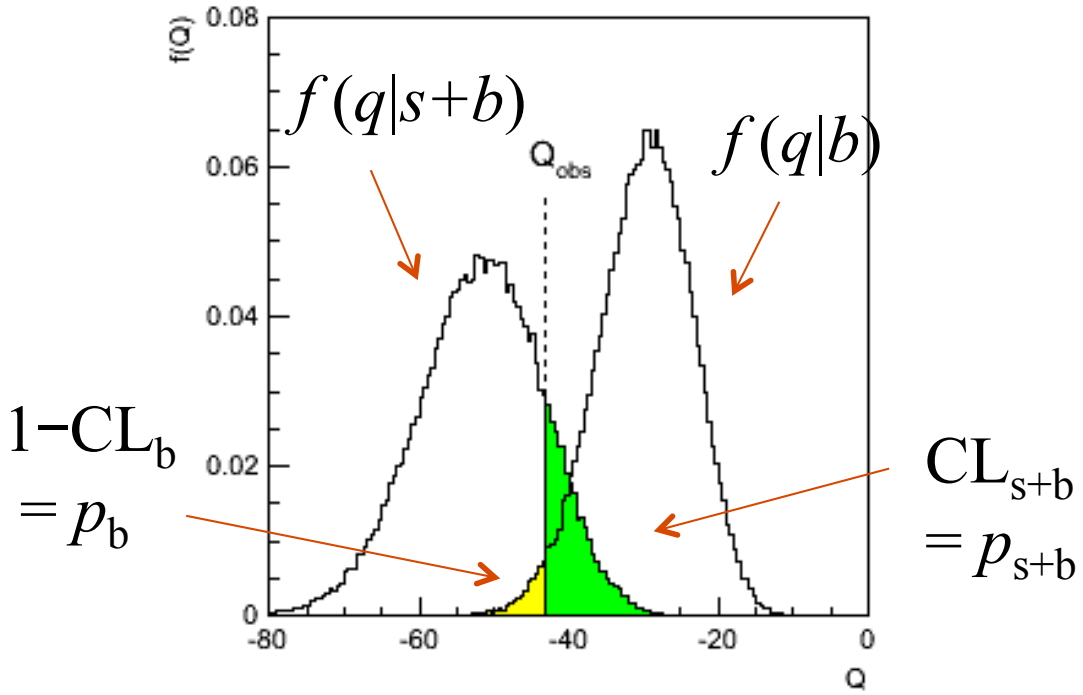
Define:

$$CL_s = \frac{CL_{s+b}}{CL_b}$$

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

Reject s+b hypothesis if:

$$CL_s \leq \alpha$$



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

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 $0 \leq l(s) \leq 1$.

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.

See also Cowan, Cranmer, Gross & Vitells, arXiv:1007.1727
for details on including systematic errors and on asymptotic
sampling distribution of likelihood ratio statistic.

Profile likelihood ratio for unified interval

We can also use directly

$$t_\mu = -2 \ln \lambda(\mu) \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

as a test statistic for a hypothesized μ.

Large discrepancy between data and hypothesis can correspond either to the estimate for μ being observed high or low relative to μ.

This is essentially the statistic used for Feldman-Cousins intervals (here also treats nuisance parameters).

Distribution of t_μ

Using Wald approximation, $f(t_\mu|\mu')$ is noncentral chi-square for one degree of freedom:

$$f(t_\mu|\mu') = \frac{1}{2\sqrt{t_\mu}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}+\frac{\mu-\mu'}{\sigma}\right)^2\right)+\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}-\frac{\mu-\mu'}{\sigma}\right)^2\right)\right]$$

Special case of $\mu = \mu'$ is chi-square for one d.o.f. (Wilks).

The p-value for an observed value of t_μ is

$$p_\mu = 1 - F(t_\mu|\mu) = 2\left(1-\Phi\left(\sqrt{t_\mu}\right)\right)$$

and the corresponding significance is

$$Z_\mu = \Phi^{-1}(1-p_\mu) = \Phi^{-1}\left(2\Phi\left(\sqrt{t_\mu}\right)-1\right)$$

Feldman-Cousins discussion

The initial motivation for Feldman-Cousins (unified) confidence intervals was to eliminate null intervals.

The F-C limits are based on a likelihood ratio for a test of μ with respect to the alternative consisting of all other allowed values of μ (not just, say, lower values).

The interval's upper edge is higher than the limit from the one-sided test, and lower values of μ may be excluded as well. A substantial downward fluctuation in the data gives a low (but nonzero) limit.

This means that when a value of μ is excluded, it is because there is a probability α for the data to fluctuate either high or low in a manner corresponding to less compatibility as measured by the likelihood ratio.

Power Constrained Limits (PCL)

CL_s has been criticized because the coverage probability of the upper limit is greater than the nominal CL = 1 − α by an amount that is not readily apparent (but can be computed).

Therefore we have proposed an alternative method for protecting against exclusion with little/no sensitivity, by regarding a value of μ to be excluded if:

(a) the value μ is rejected by the test, i.e., $\mathbf{x} \in w_\mu$ or equivalently $p_\mu < \alpha$, and

(b) one has sufficient sensitivity to μ, i.e., $M_0(\mu) \geq M_{\text{min}}$.

Here the measure of sensitivity is the power of the test of μ with respect to the alternative $\mu = 0$:

$$M_0(\mu) = P(\mathbf{x} \in w_\mu | 0) = P(p_\mu < \alpha | 0)$$

Constructing PCL

First compute the distribution under assumption of the background-only ($\mu = 0$) hypothesis of the "usual" upper limit $\mu_{\rm up}$ with no power constraint.

The power of a test of μ with respect to $\mu = 0$ is the fraction of times that μ is excluded ($\mu_{\rm up} < \mu$):

$$M_0(\mu) = P(\mu_{\rm up} < \mu|0)$$

Find the smallest value of μ ($\mu_{\rm min}$), such that the power is at least equal to the threshold $M_{\rm min}$.

The Power-Constrained Limit is:

$$\mu^*_{\rm up} = \max(\mu_{\rm up}, \mu_{\rm min})$$

PCL for upper limit with Gaussian measurement

Suppose $\hat{\mu} \sim \text{Gauss}(\mu, \sigma)$, goal is to set upper limit on μ.

Define critical region for test of μ as $\quad \hat{\mu} < \mu - \sigma\Phi^{-1}(1-\alpha)$

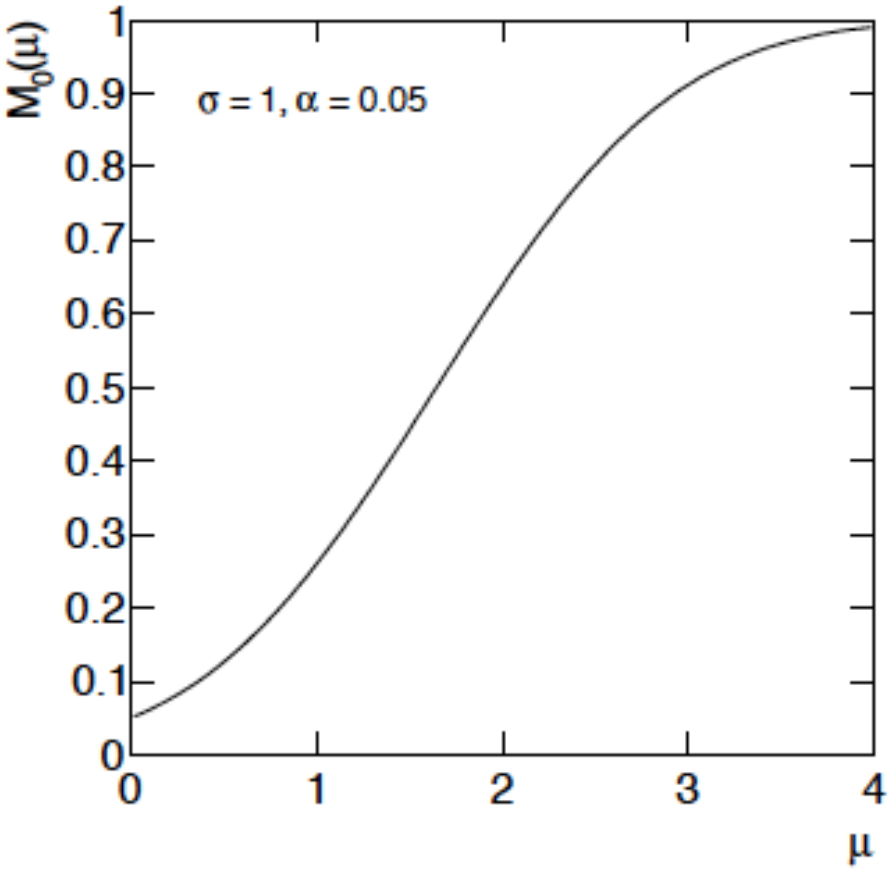
inverse of standard Gaussian cumulative distribution

This gives (unconstrained) upper limit: $\quad \mu_{\text{up}} = \hat{\mu} + \sigma\Phi^{-1}(1-\alpha)$

Power $M_0(\mu)$ for Gaussian measurement

The power of the test of μ with respect to the alternative $\mu' = 0$ is:

$$M_0(\mu) = P\left(\hat{\mu} < \mu - \sigma\Phi^{-1}(1-\alpha)|0\right) = \Phi\left(\frac{\mu}{\sigma} - \Phi^{-1}(1-\alpha)\right)$$



standard Gaussian cumulative distribution

Spurious exclusion when $\hat{\mu}$ fluctuates down

Requiring the power be at least M_{min}

$$\Phi\left(\frac{\mu}{\sigma} - \Phi^{-1}(1-\alpha)\right) \geq M_{\text{min}}$$

implies that the smallest μ to which one is sensitive is

$$\mu_{\text{min}} = \sigma\left(\Phi^{-1}(M_{\text{min}}) + \Phi^{-1}(1-\alpha)\right)$$

If one were to use the unconstrained limit, values of μ at or below μ_{min} would be excluded if

$$\hat{\mu} < \sigma\Phi^{-1}(M_{\text{min}})$$

That is, one excludes $\mu < \mu_{\text{min}}$ when the unconstrained limit fluctuates too far downward.

Choice of minimum power

Choice of M_{min} is convention. Formally it should be large relative to α (5%). Earlier we have proposed

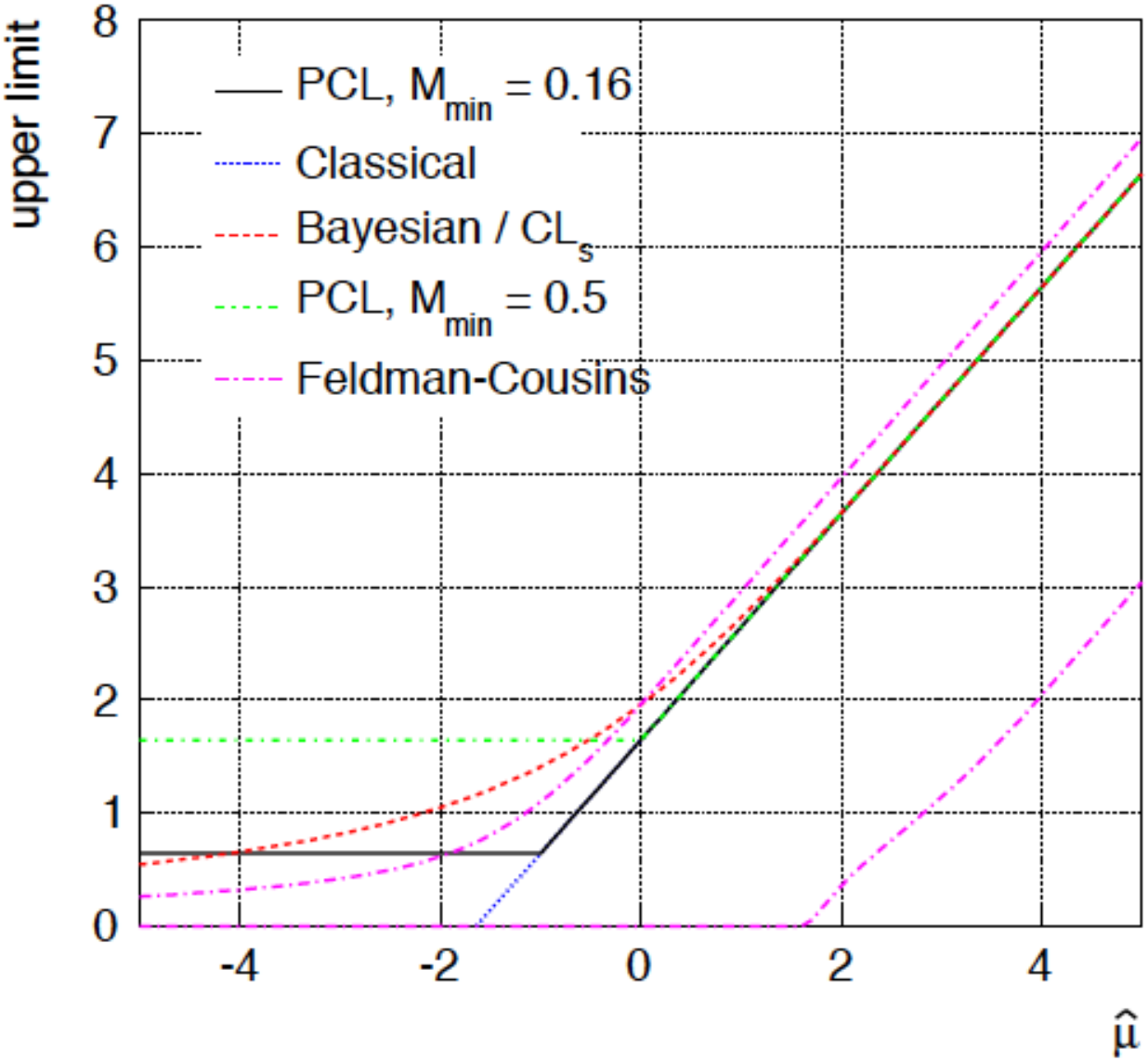
$$M_{\text{min}} = \Phi(-1) = 0.1587$$

because in Gaussian example this means that one applies the power constraint if the observed limit fluctuates down by one standard deviation.

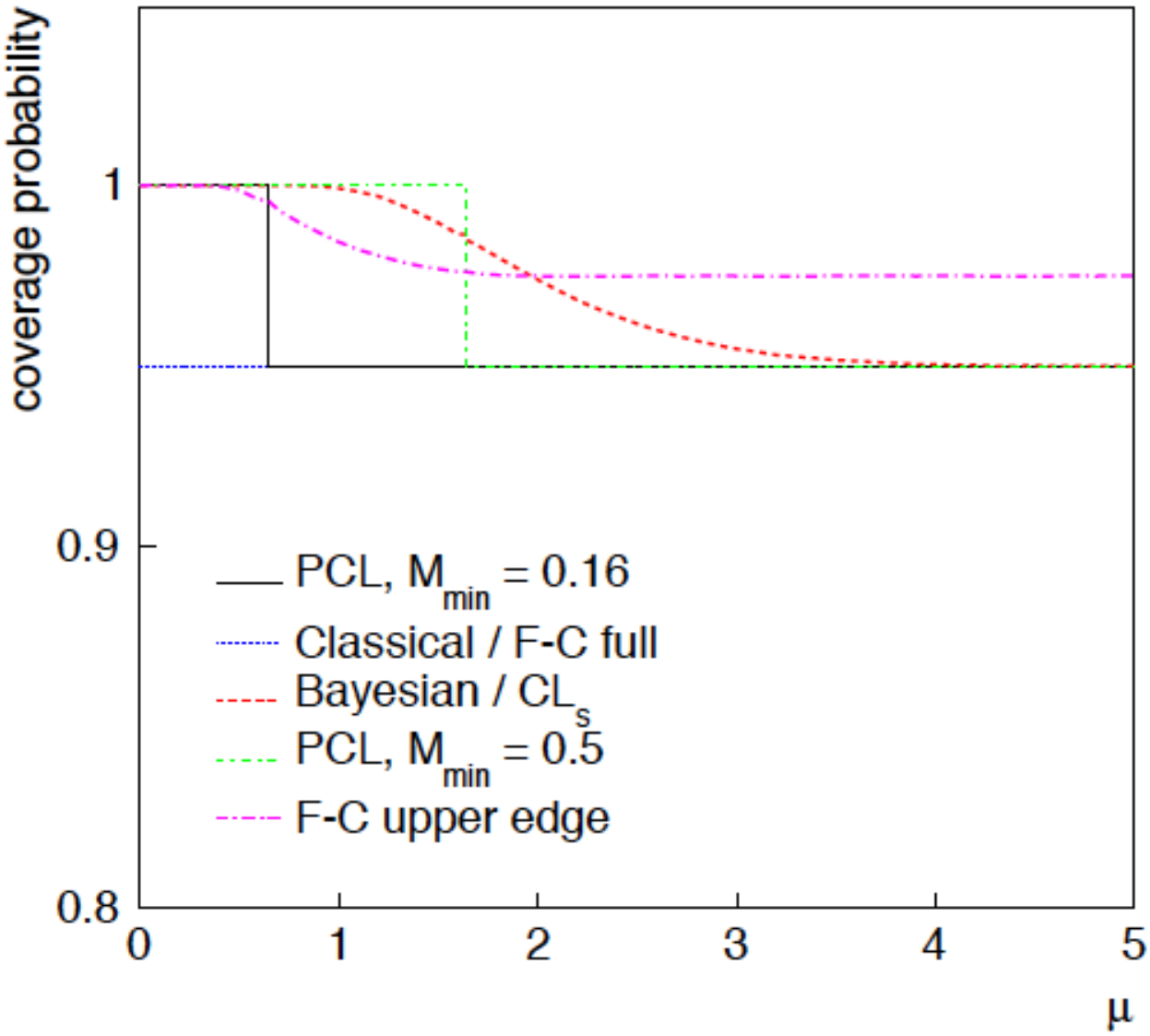
In fact the distribution of μ_{up} is often roughly Gaussian, so we call this a "1σ" (downward) fluctuation and use $M_{\text{min}} = 0.16$ regardless of the exact distribution of μ_{up}.

For the Gaussian example, this gives $\mu_{\text{min}} = 0.64\sigma$, i.e., the lowest limit is similar to the intrinsic resolution of the measurement (σ).

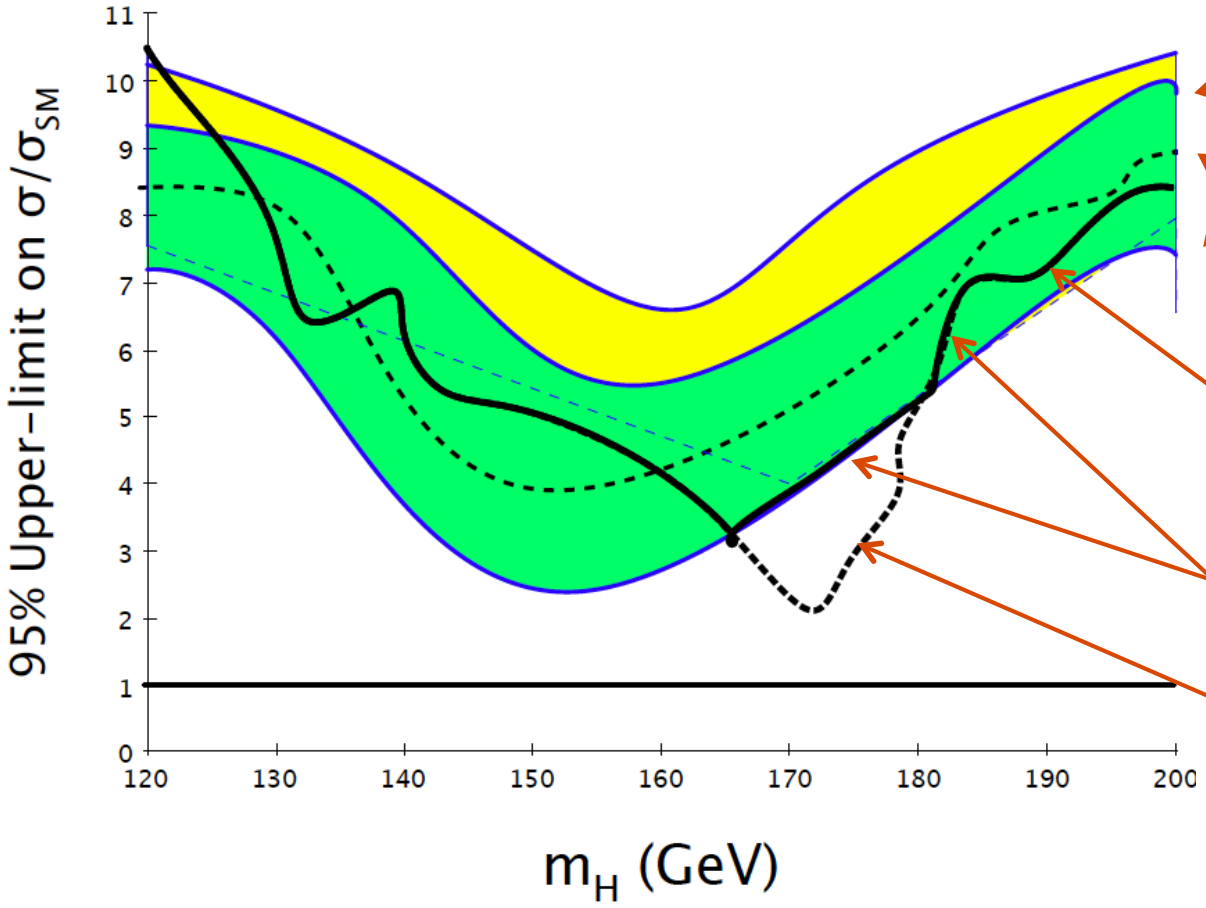
Upper limits for Gaussian problem



Coverage probability for Gaussian problem



PCL in practice



+/− 1σ band of limit dist. assuming $\mu = 0$.

median limit (unconstrained)

observed limit

PCL with $M_{\text{min}} = 0.16$

Here power below threshold; do not exclude.

Important to report both the constrained and unconstrained limits.

Some reasons to consider increasing $M_{\min}$

$M_{\min}$ is supposed to be "substantially" greater than α (5%).

So $M_{\min} = 16\%$ is fine for $1 - \alpha = 95\%$, but if we ever want $1 - \alpha = 90\%$, then16% is not "large" compared to 10%; $\mu_{\min} = 0.28\sigma$ starts to look small relative to the intrinsic resolution of the measurement. Not an issue if we stick to 95% CL.

PCL with $M_{\min} = 16\%$ is often substantially lower than CLs. This is because of the conservatism of CLs (see coverage).

But goal is not to get a lower limit per se, rather

- to use a test with higher power in those regions where one feels there is enough sensitivity to justify exclusion and
- to allow for easy communication of coverage (95% for $\mu \geq \mu_{\min}$; 100% otherwise).

Aggressive conservatism

It could be that owing to practical constraints, certain systematic uncertainties are over-estimated in an analysis; this could be justified by wanting to be conservative.

The consequence of this will be that the +/−1 sigma bands of the unconstrained limit are broader than they otherwise would be.

If the unconstrained limit fluctuates low, it could be that the PCL limit, constrained at the −1sigma band, is lower than it would be had the systematics been estimated correctly.

conservative = aggressive

If the power constraint M_{min} is at 50%, then by inflating the systematics the median of the unconstrained limit is expected to move less, and in any case upwards, i.e., it will lead to a less strong limit (as one would expect from “conservatism”).

A few further considerations

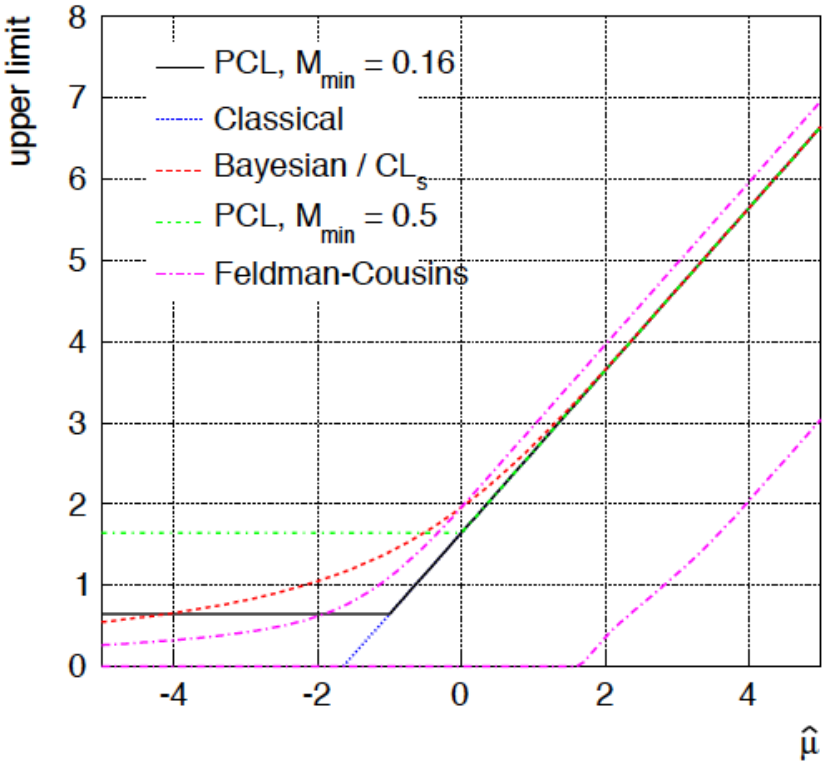
Obtaining PCL requires the distribution of unconstrained limits, from which one finds the M_{min} (16%, 50%) percentile.

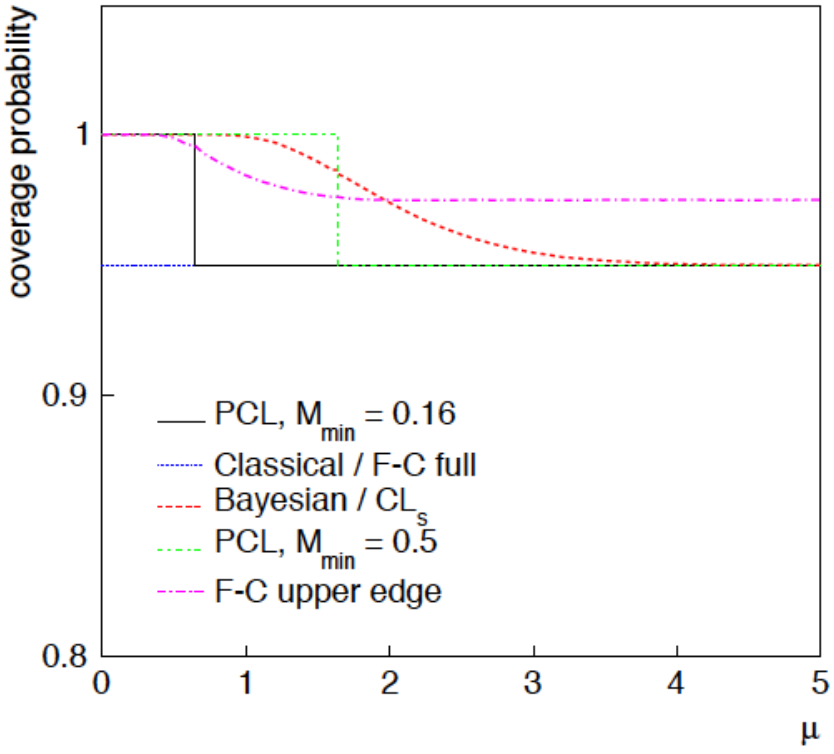
In some analyses this can entail calculational issues that are expected to be less problematic for $M_{\text{min}} = 50\%$ than for 16%.

Analysts produce anyway the median limit, even in absence of the error bands, so with $M_{\text{min}} = 50\%$ the burden on the analyst is reduced somewhat (but one would still want the error bands).

We therefore recently proposed moving M_{min} to 50%.

PCL with $M_{\min} = 0.16$, 0.50 (and other limits)





With $M_{\min} = 50\%$, power constraint is applied half the time.

This is somewhat contrary to the original spirit of preventing a "lucky" fluctuation from leading to a limit that is small compared to the intrinsic resolution of the measurement.

But PCL still lower than CLs most of the time (e.g., $x > -0.4$).

Treatment of nuisance parameters

In most problems, the data distribution is not uniquely specified by μ but contains nuisance parameters θ.

This makes it more difficult to construct an (unconstrained) interval with correct coverage probability for all values of θ, so sometimes approximate methods used ("profile construction").

More importantly for PCL, the power $M_0(\mu)$ can depend on θ. So which value of θ to use to define the power?

Since the power represents the probability to reject μ if the true value is $\mu = 0$, to find the distribution of μ_{up} we take the values of θ that best agree with the data for $\mu = 0$: $\hat{\hat{\boldsymbol{\theta}}}(0)$

May seem counterintuitive, since the measure of sensitivity now depends on the data. We are simply using the data to choose the most appropriate value of θ where we quote the power.

Summary and conclusions

Exclusion limits effectively tell one what parameter values are (in)compatible with the data.

Frequentist: exclude range where p-value of param < 5%.
Bayesian: low prob. to find parameter in excluded region.

In both cases one must choose the grounds on which the parameter is excluded (estimator too high, low? low likelihood ratio?) .

With a "usual" upper limit, a large downward fluctuation can lead to exclusion of parameter values to which one has little or no sensitivity (will happen 5% of the time).

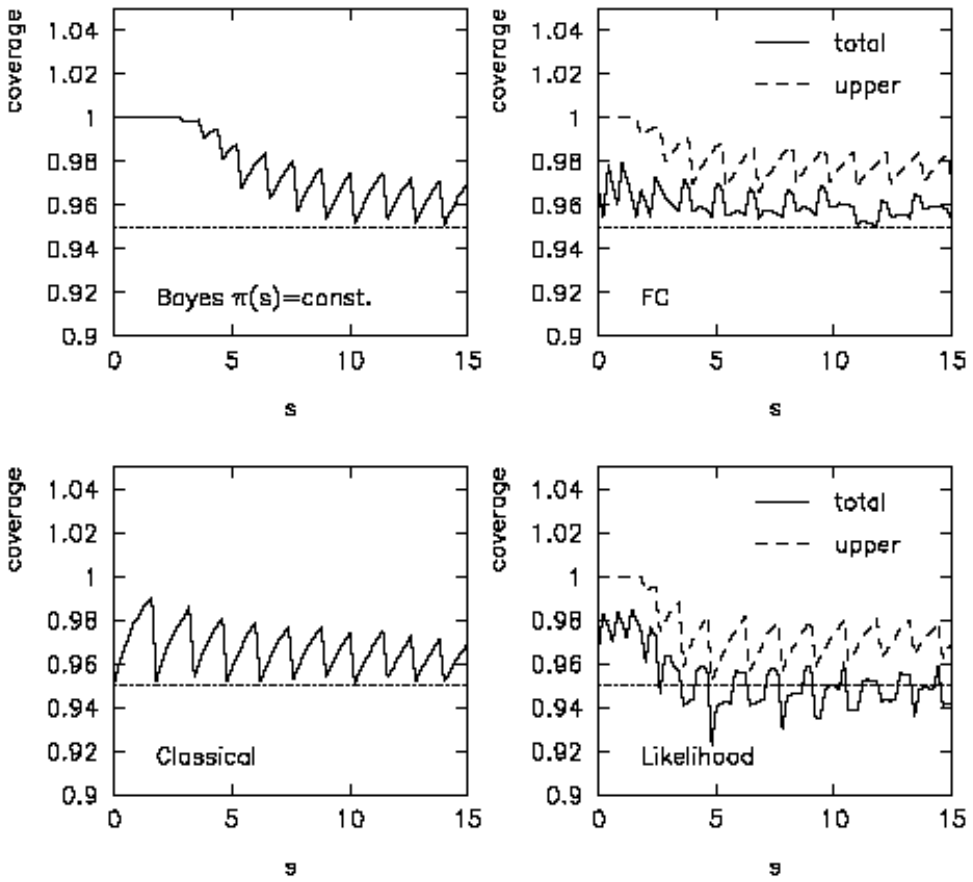
"Solutions": CLs, PCL, F-C

All of the solutions have well-defined properties, to which there may be some subjective assignment of importance.

Extra slides

Coverage probability of intervals

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



Intervals from the likelihood function

In the large sample limit it can be shown for ML estimators:

$$\hat{\vec{\theta}} \sim N(\vec{\theta}, V) \qquad (n\text{-dimensional Gaussian, covariance } V)$$

$$L(\vec{\theta}) = L_{\max} \exp\left[-\frac{1}{2} Q(\hat{\vec{\theta}}, \vec{\theta})\right], \quad Q(\hat{\vec{\theta}}, \vec{\theta}) = (\hat{\vec{\theta}} - \vec{\theta})^T V^{-1} (\hat{\vec{\theta}} - \vec{\theta})$$

$Q(\hat{\vec{\theta}}, \vec{\theta}) = Q_\gamma$ defines a hyper-ellipsoidal confidence region,

$$P(\text{ellipsoid covers true } \vec{\theta}) = P(Q(\hat{\vec{\theta}}, \vec{\theta}) \leq Q_\gamma)$$

If $\hat{\vec{\theta}} \sim N(\vec{\theta}, V)$ then $Q(\hat{\vec{\theta}}, \vec{\theta}) \sim \text{Chi-square}(n)$

$$\text{coverage probability} \equiv 1 - \gamma = \int_0^{Q_\gamma} f_{\chi^2}(z; n)\, dz = F_{\chi^2}(Q_\gamma; n)$$

Approximate confidence regions from $L(\theta)$

So the recipe to find the confidence region with CL = $1-\gamma$ is:

$$\ln L(\vec{\theta}) = \ln L_{\max} - \frac{Q_\gamma}{2} \quad \text{or} \quad \chi^2(\vec{\theta}) = \chi^2_{\min} + Q_\gamma$$

where $Q_\gamma = F^{-1}_{\chi^2}(1-\gamma; n)$

Q_γ	$1-\gamma$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

For finite samples, these are approximate confidence regions.

Coverage probability not guaranteed to be equal to $1-\gamma$;

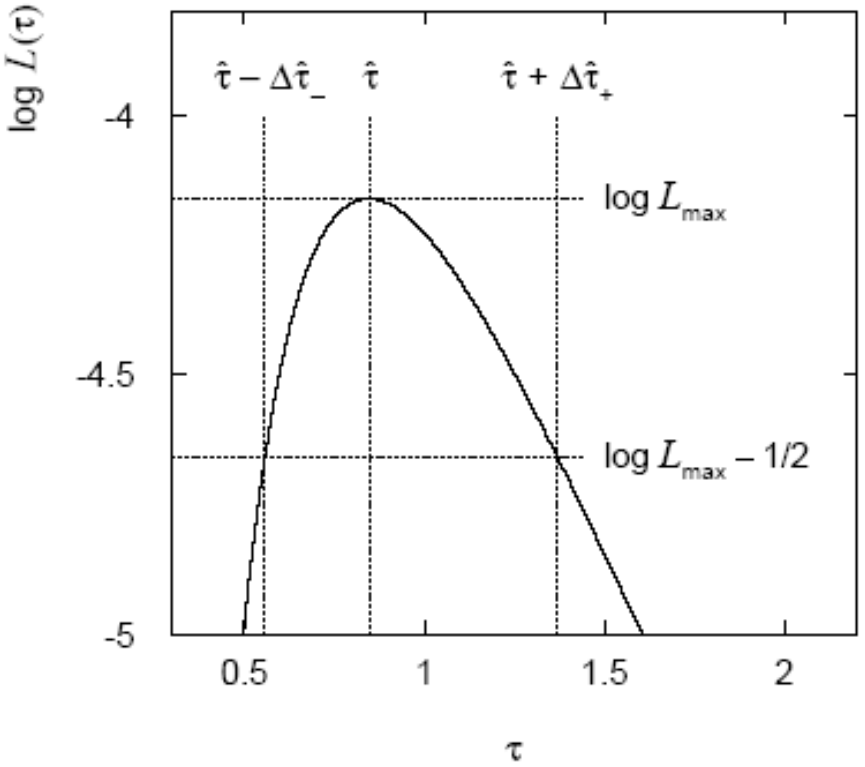
no simple theorem to say by how far off it will be (use MC).

Remember here the interval is random, not the parameter.

Example of interval from ln $L(\theta)$

For n=1 parameter, CL = 0.683, $Q_\gamma = 1$.

Our exponential example, now with $n = 5$ observations:



$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

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 $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

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

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): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

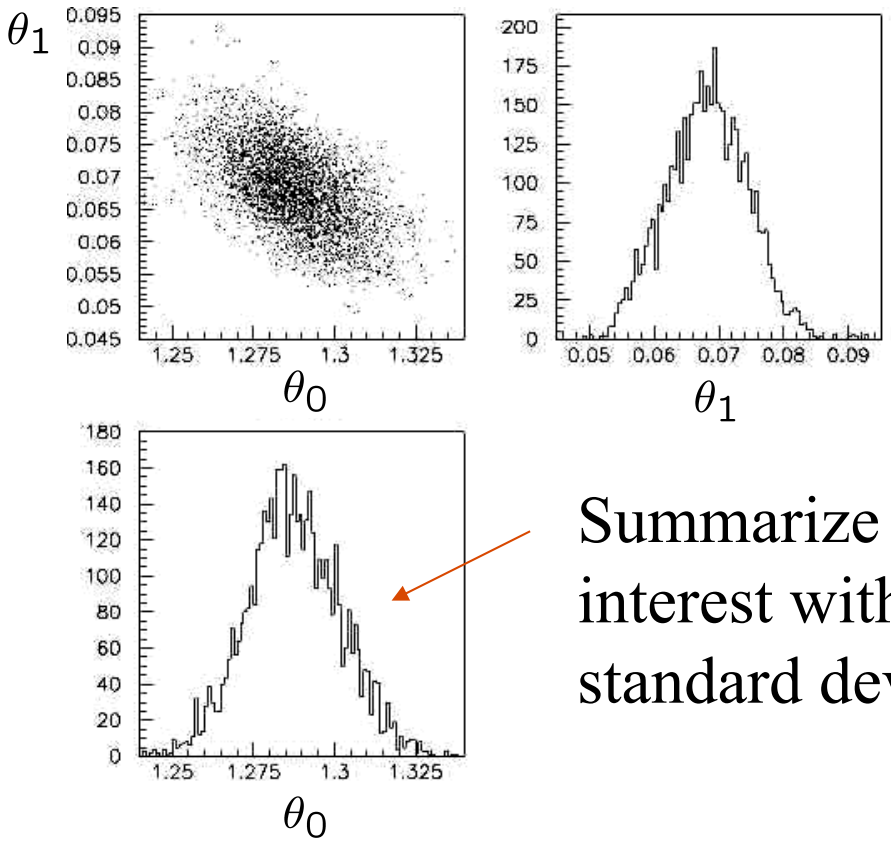
I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it;

if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.

If proposed step rejected, hop in place.

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?)

Negatively Biased Relevant Subsets

Consider again $x \sim \text{Gauss}(\mu, \sigma)$ and use this to find limit for μ.

We can find the conditional probability for the limit to cover μ given x in some restricted range, e.g., $x < c$ for some constant c.

This conditional coverage probability may be greater or less than $1 - \alpha$ for different values of μ (the value of which is unkown).

But suppose that the conditional coverage is less than $1 - \alpha$ for *all* values of μ. The region of x where this is true is a *Negatively Biased Relevant Subset*.

Recent studies by Bob Cousins (CMS) and
Ofer Vitells (ATLAS) related to earlier publications,
especially, R. Buehler, Ann. Math. Sci., 30 (4) (1959) 845.

Betting Games

So what's wrong if the limit procedure has NBRS?

Suppose you observe x, construct the confidence interval and assert that an interval thus constructed covers the true value of the parameter with probability $1 - \alpha$.

This means you should be willing to accept a bet at odds $\alpha : 1 - \alpha$ that the interval covers the true parameter value.

Suppose your opponent accepts the bet if x is in the NBRS, and declines the bet otherwise. On average, you lose, regardless of the true (and unknown) value of μ.

With the "naive" unconstrained limit, if your opponent only accepts the bet when $x < -1.64\sigma$, (all values of μ excluded) you always lose!

(Recall the unconstrained limit based on the likelihood ratio never excludes $\mu = 0$, so if that value is true, you do not lose.)

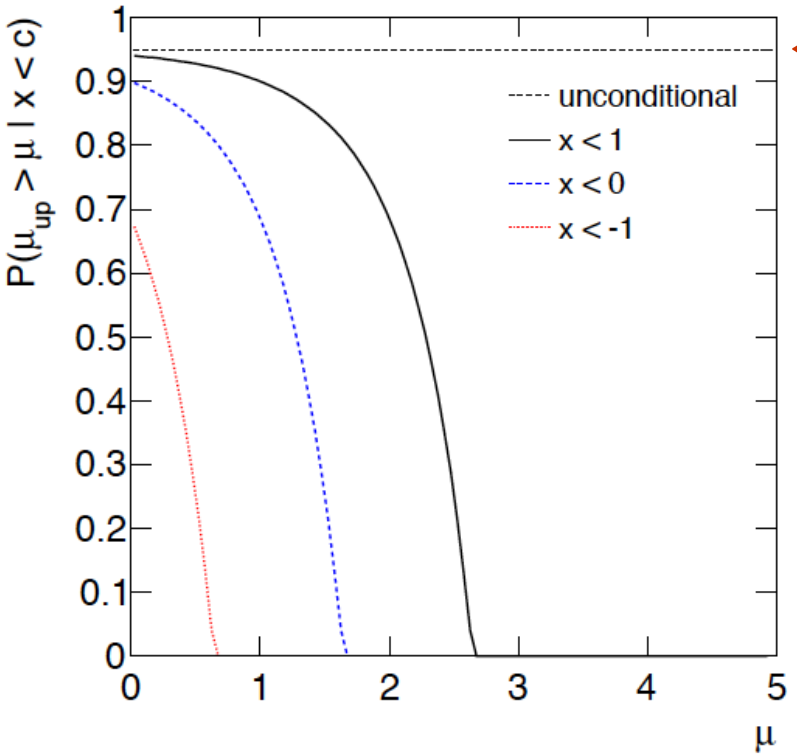
NBRS for unconstrained upper limit

For the unconstrained upper limit (i.e., CL_{s+b}) the conditional probability for the limit to cover μ given $x < c$ is:

$$P(\mu_{\rm up} > \mu | x < c) = \frac{1 - \alpha - \Phi\left(\frac{\mu - c}{\sigma}\right)}{1 - \Phi\left(\frac{\mu - c}{\sigma}\right)}$$

Maximum wrt μ is less than $1-\alpha$ → Negatively biased relevant subsets.

N.B. $\mu = 0$ is never excluded for unconstrained limit based on likelihood-ratio test, so at that point coverage = 100%, hence no NBRS.



← $1 - \alpha$

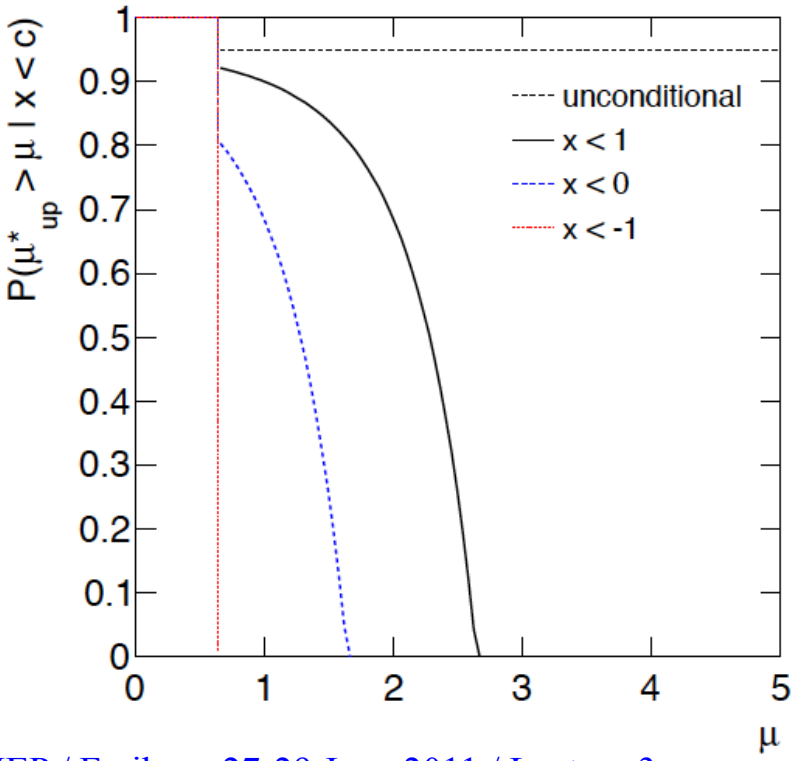
(Adapted) NBRS for PCL

For PCL, the conditional probability to cover μ given $x < c$ is:

$$P(\mu^*_{\rm up} > \mu | x < c) = \begin{cases} 1 & \mu < \mu_{\rm min}, \\ \frac{1-\alpha-\Phi\left(\frac{\mu-c}{\sigma}\right)}{1-\Phi\left(\frac{\mu-c}{\sigma}\right)} & \text{otherwise.} \end{cases}$$

Coverage goes to 100% for $\mu < \mu_{\rm min}$, therefore no NBRS.

Note one does not have max conditional coverage $\geq 1-\alpha$ for all $\mu > \mu_{\rm min}$ ("adapted conditional coverage"). But if one conditions on μ, no limit would satisfy this.



← $1-\alpha$

Conditional coverage for CLs, F-C