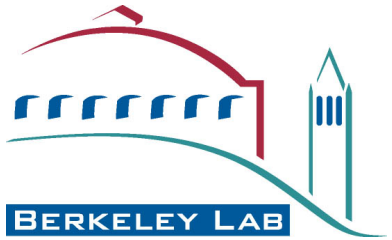
Power-Constrained Limits, etc.



LBNL-ATLAS Meeting
Berkeley, 17 June, 2011

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

* with Kyle Cranmer, Eilam Gross, Ofer Vitells

Outline

Confidence intervals from inversion of a test (review)

The problem of spurious exclusion and previous solutions (CLs)

Power Constrained Limits (PCL) (see CCGV arXiv:1105.3166)

PCL for upper limit based on a Gaussian measurement

Distribution of upper limit and choice of minimum power

Treatment of nuisance parameters

Issues concerning negatively biased relevant subsets

Summary and conclusions

Reminder about statistical tests

Consider test of a parameter μ, e.g., proportional to cross section.

Result of measurement is a set of numbers $\boldsymbol{x}$.

To define test of μ, specify *critical region* w_μ, such that probability to find $\boldsymbol{x} \in w_\mu$ is not greater than α (the *size* or *significance level*):

$$P(\mathbf{x} \in w_\mu | \mu) \leq \alpha$$

(Must use inequality since $\boldsymbol{x}$ may be discrete, so there may not exist a subset of the data space with probability of exactly α.)

Often use, e.g., $\alpha = 0.05$.

If observe $\boldsymbol{x} \in w_\mu$, reject μ.

Test statistics and p-values

Often construct a test statistic, q_μ, which reflects the level of agreement between the data and the hypothesized value μ.

For examples of statistics based on the profile likelihood ratio, see, e.g., CCGV arXiv:1007.1727 (the "Asimov" paper).

Usually define q_μ such that higher values represent increasing incompatibility with the data, so that the p-value of μ is:

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

observed value of q_μ

pdf of q_μ assuming μ

Equivalent formulation of test: reject μ if $p_\mu < \alpha$.

Confidence interval from inversion of a test

Carry out a test of size α for all values of μ.

The values that are not rejected constitute a *confidence interval* for μ at confidence level CL = $1 - \alpha$.

The confidence interval will by construction contain the true value of μ with probability of at least $1 - \alpha$.

Can give upper limit μ_{up}, i.e., the largest value of μ not rejected, i.e., the upper edge of the confidence interval.

The interval (and limit) depend on the choice of the test, which is often based on considerations of power.

Power of a statistical test

But where to define critical region? Usually put this where the test has a high *power* with respect to an alternative hypothesis μ'.

The *power* of the test of μ with respect to the alternative μ' is the probability to reject μ if μ' is true:

(M = Mächtigkeit, мощность)

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

E.g., for an upper limit, maximize the power with respect to the alternative consisting of $\mu' < \mu$.

Other types of tests not based directly on power (e.g., likelihood ratio).

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: $$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

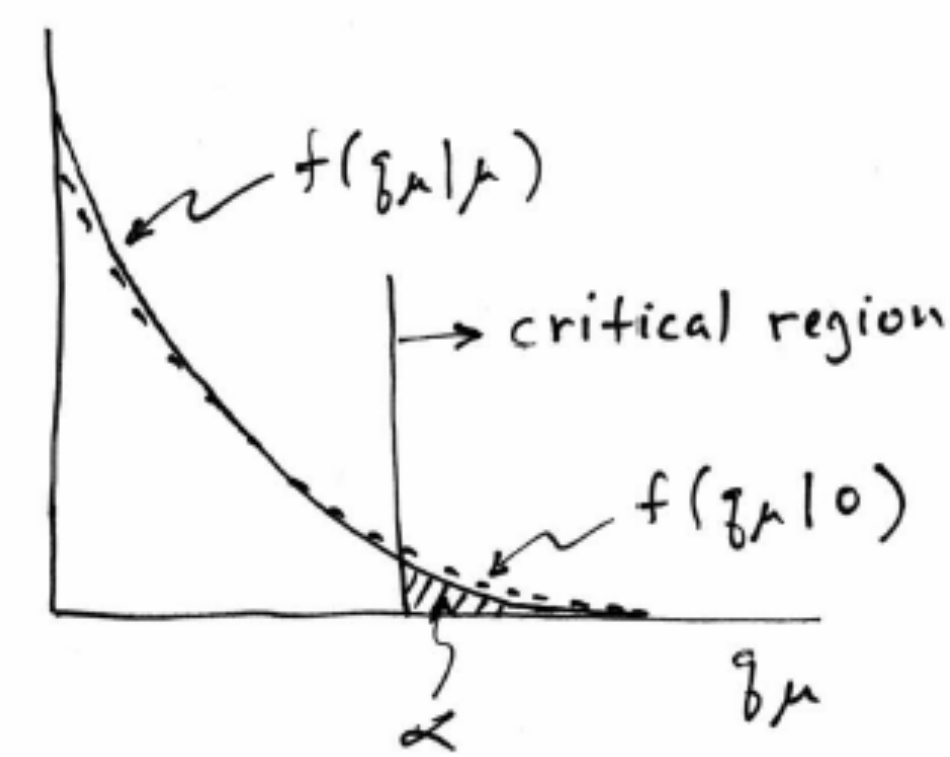
Large sample approximation: $$p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

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

Low sensitivity to μ

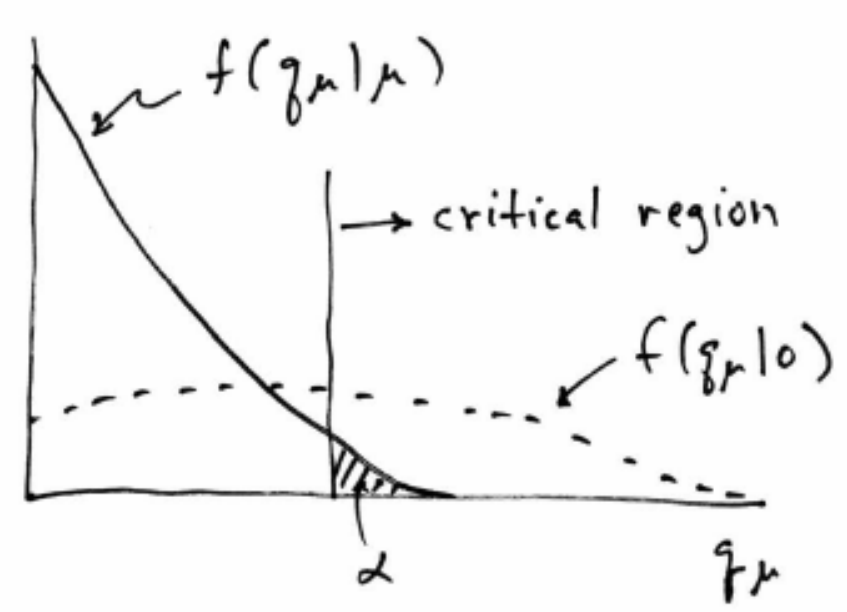
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

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:

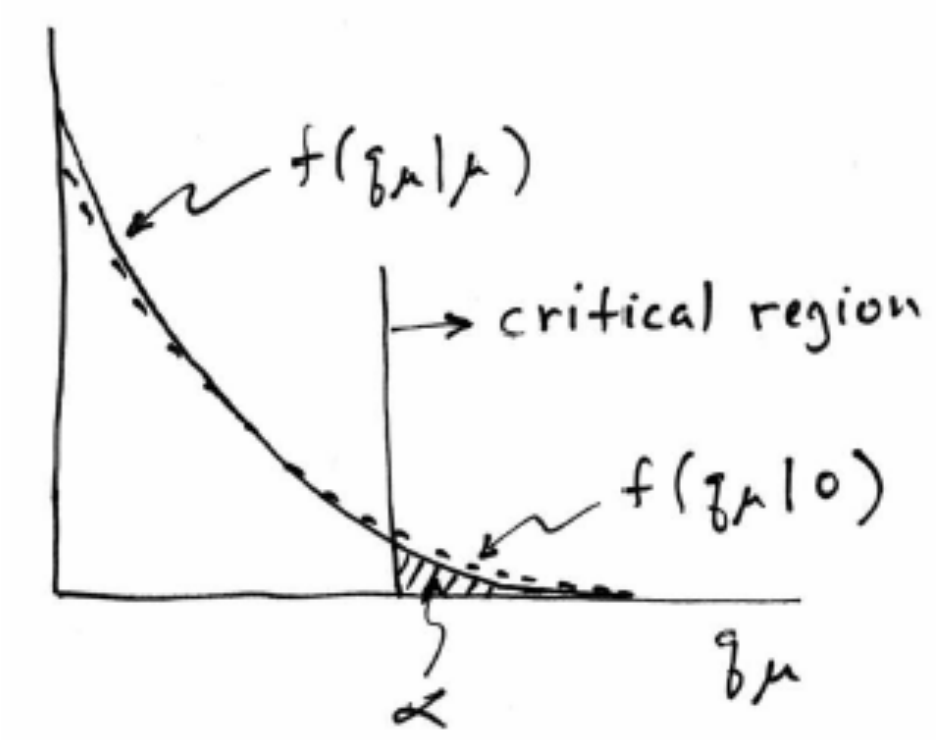


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α. We use this power as a measure of the sensitivity.

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"

Previous 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.

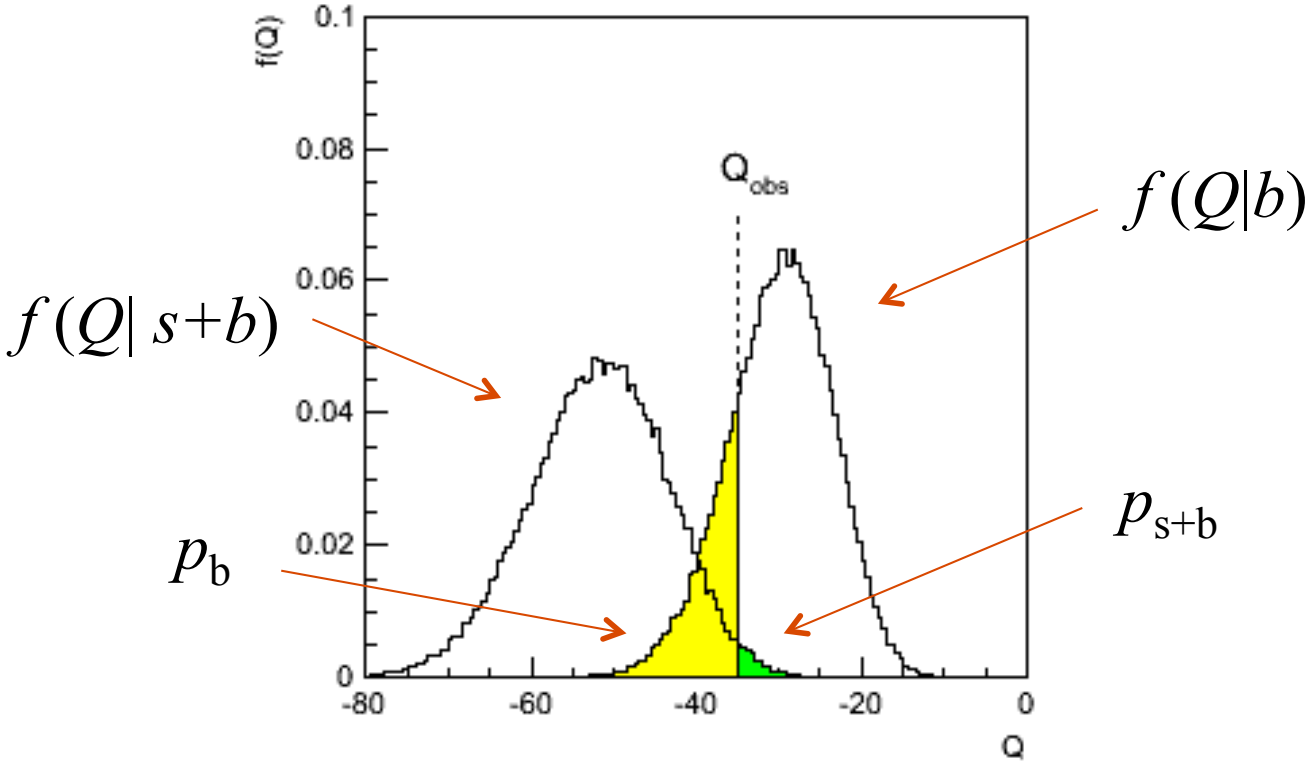
In the 1990s this was re-examined for the LEP Higgs search by Alex Read and others

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.

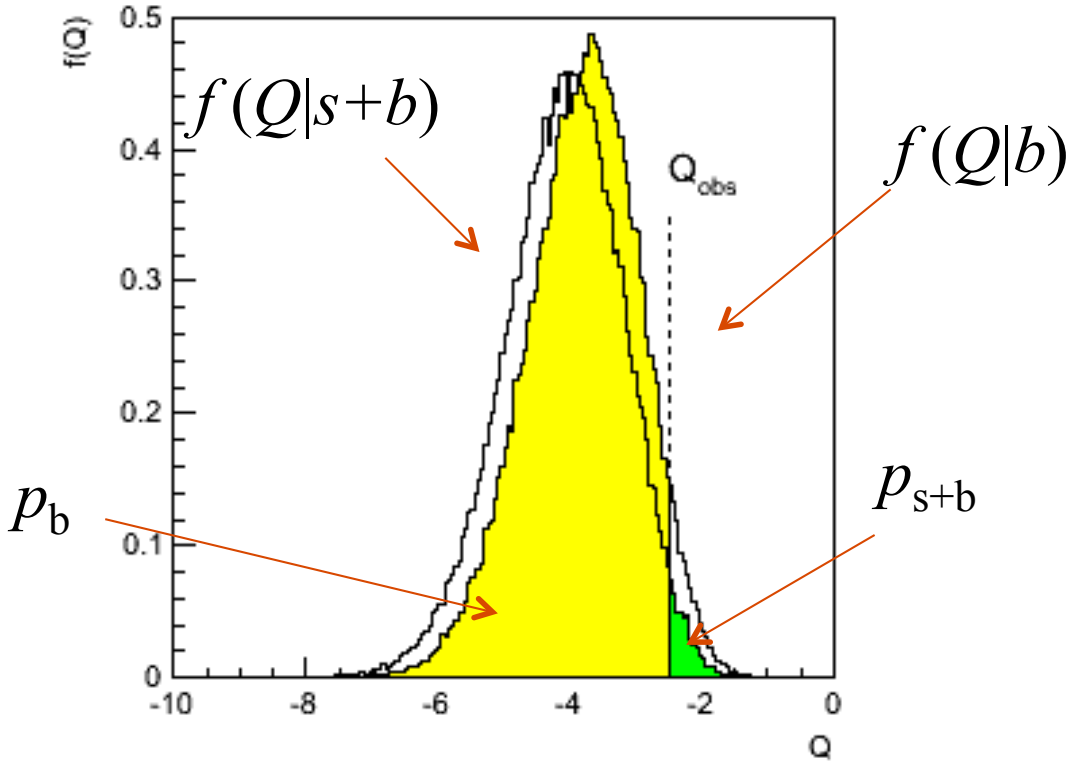
The CL_s procedure

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.,

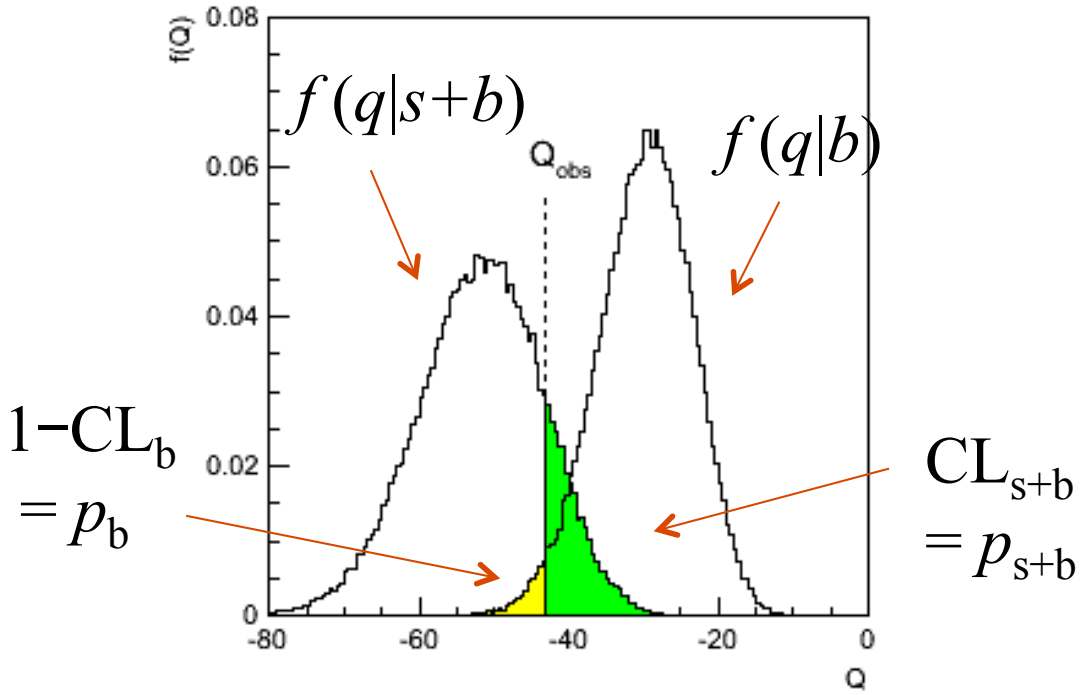
Define:

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

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

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).

Feldman-Cousins unified intervals

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., $M_0(\mu) \geq M_{\min}$.

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 $\quad \hat{\mu} < \mu - \sigma\Phi^{-1}(1-\alpha)$

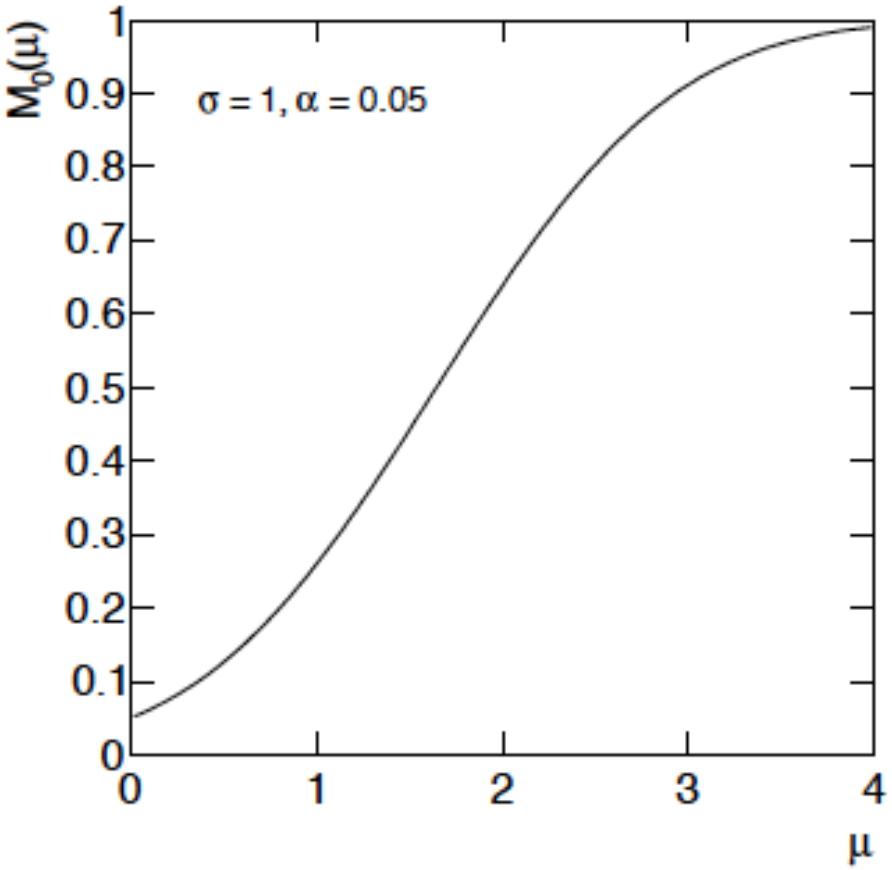
inverse of standard Gaussian cumulative distribution

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

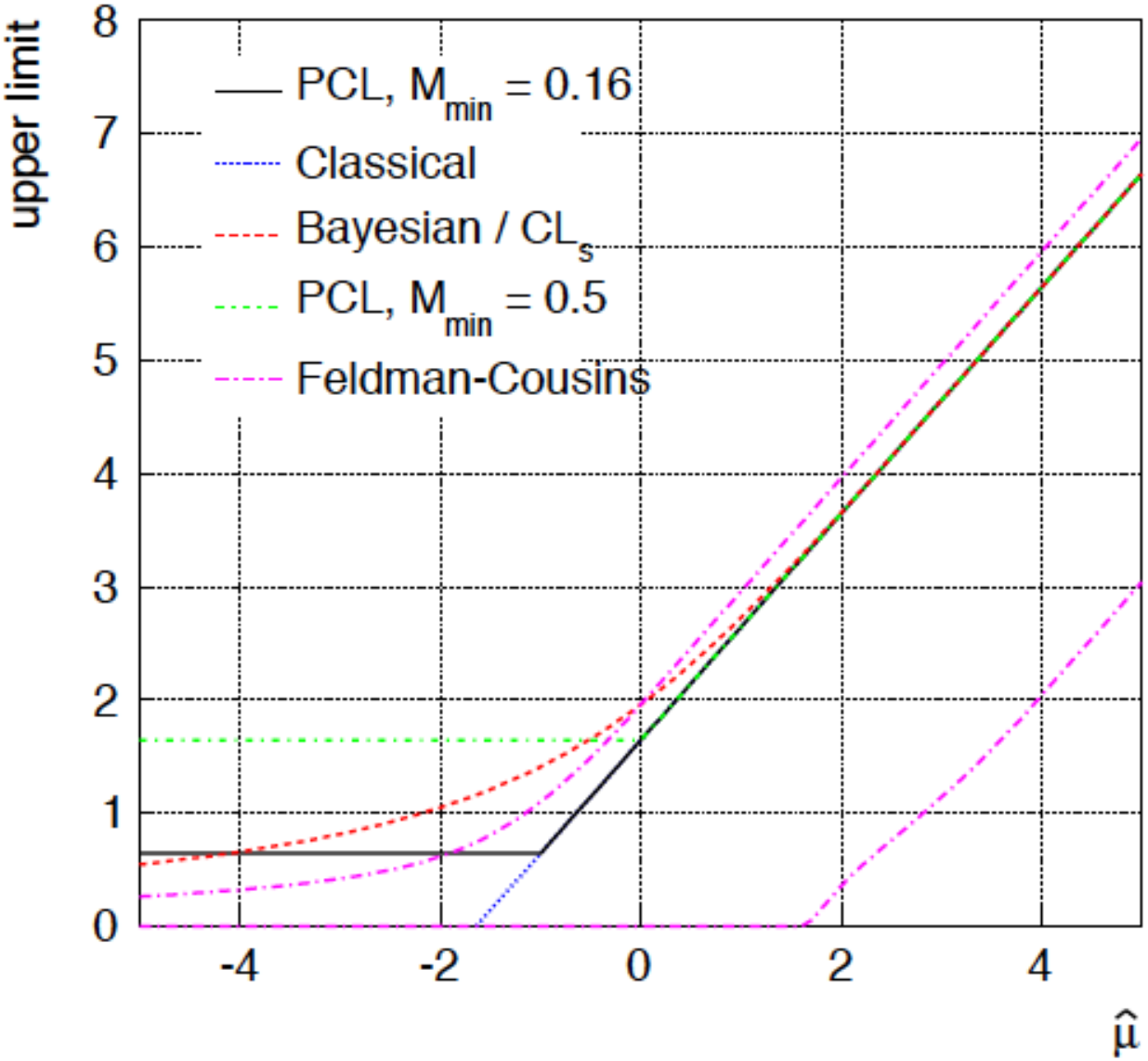
$$M_{\text{min}} = \Phi(-1) = 0.1587$$

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

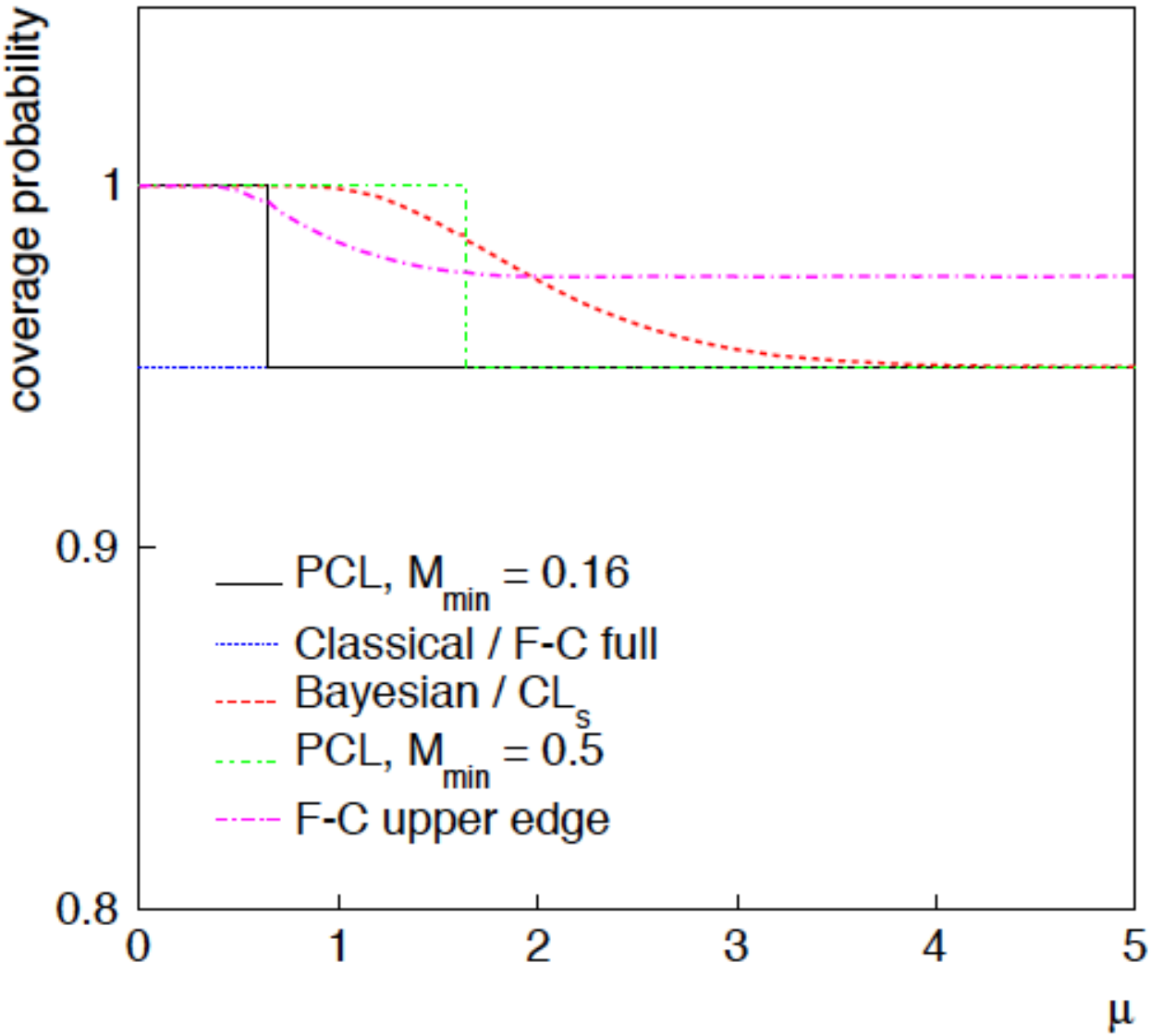
In fact the distribution of μ_{up} is often roughly Gaussian, so we call this a "1σ" (downward) fluctuation and use $M_{\text{min}} = 0.16$ regardless of the exact distribution of μ_{up}.

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).

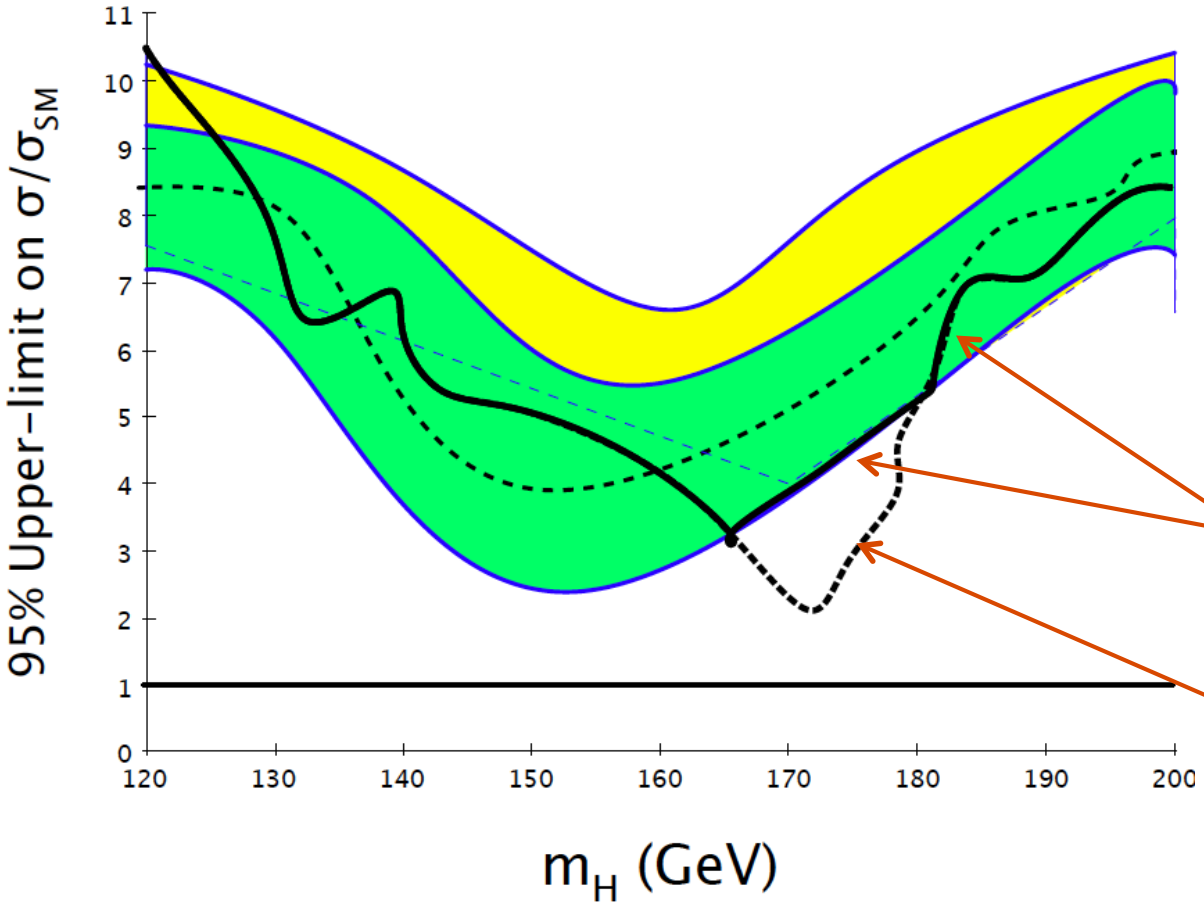
Upper limits for Gaussian problem



Coverage probability for Gaussian problem



PCL as a function of, e.g., m_H



PCL

Here power below threshold; do not exclude.

Some reasons to consider increasing M_{min}

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

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

PCL with $M_{\text{min}} = 16\%$ 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 $\mu \geq \mu_{\text{min}}$; 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

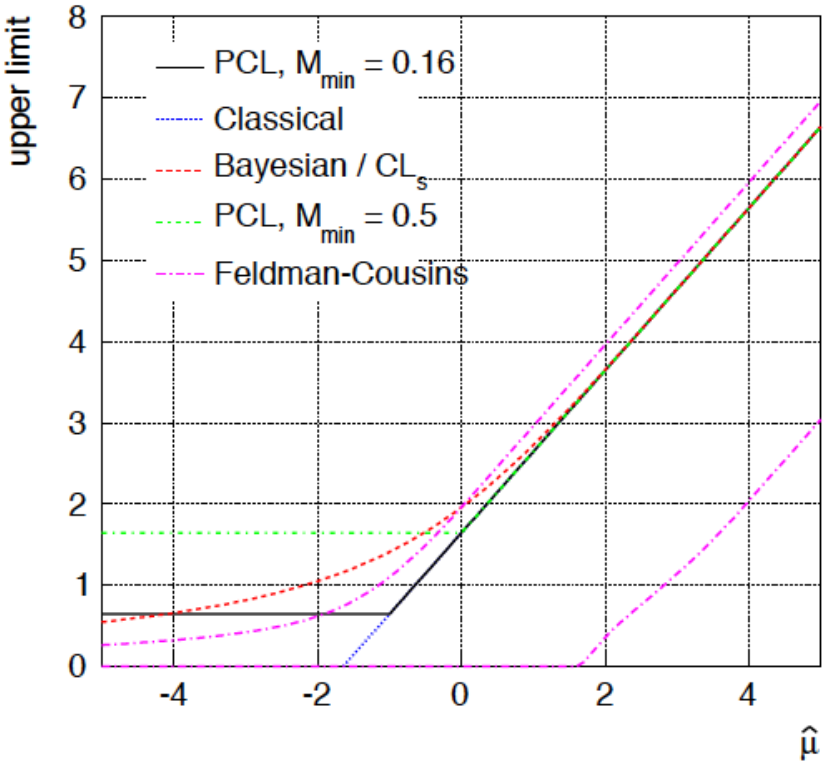
Obtaining PCL requires the distribution of unconstrained limits, from which one finds the M_{min} (16%, 50%) percentile.

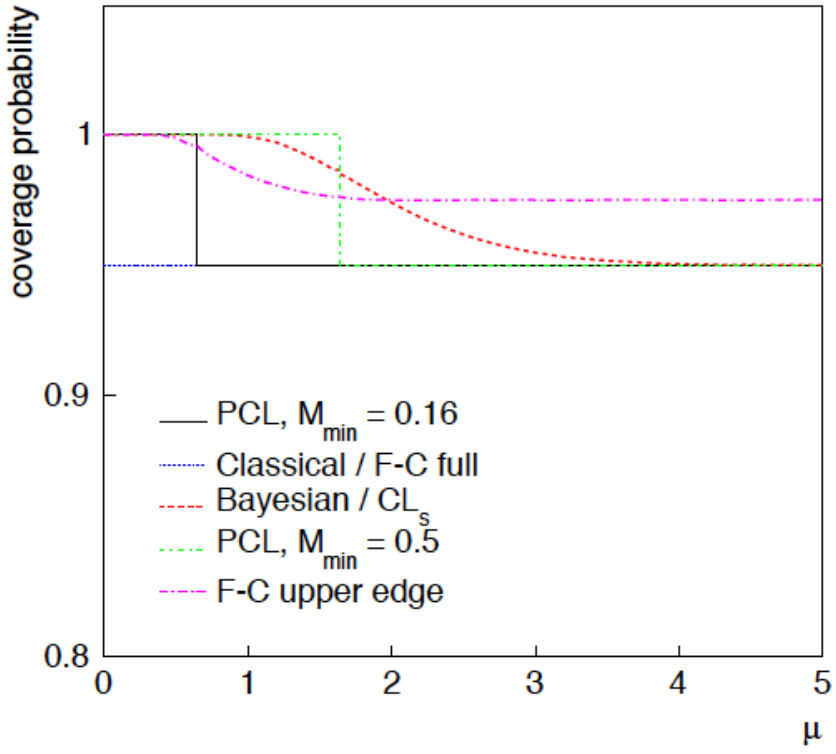
In some analyses this can entail calculational issues that are expected to be less problematic for $M_{\text{min}} = 50\%$ than for 16%.

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.

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.)

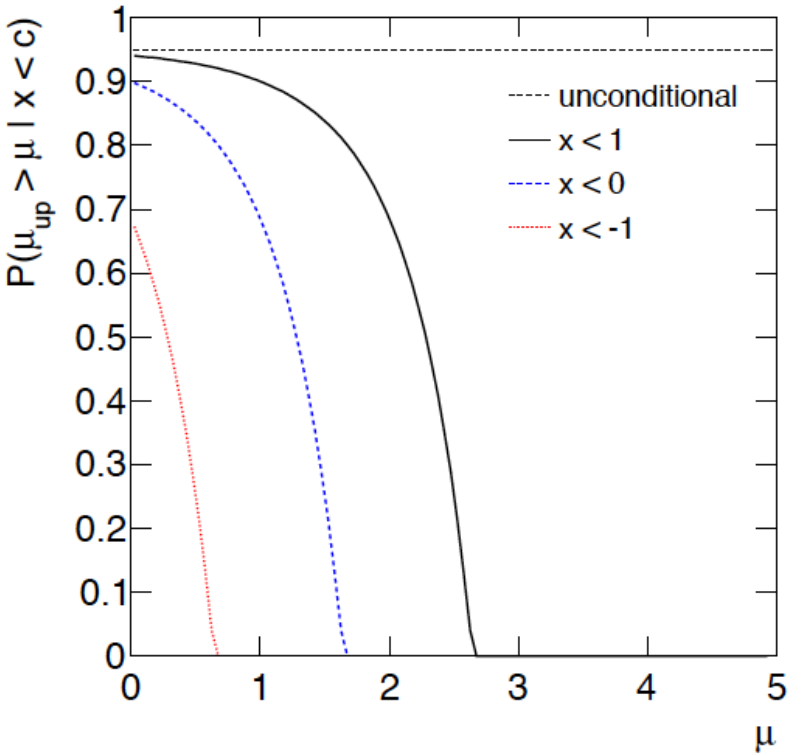
NBRS for unconstrained upper limit

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.



← $1 - \alpha$

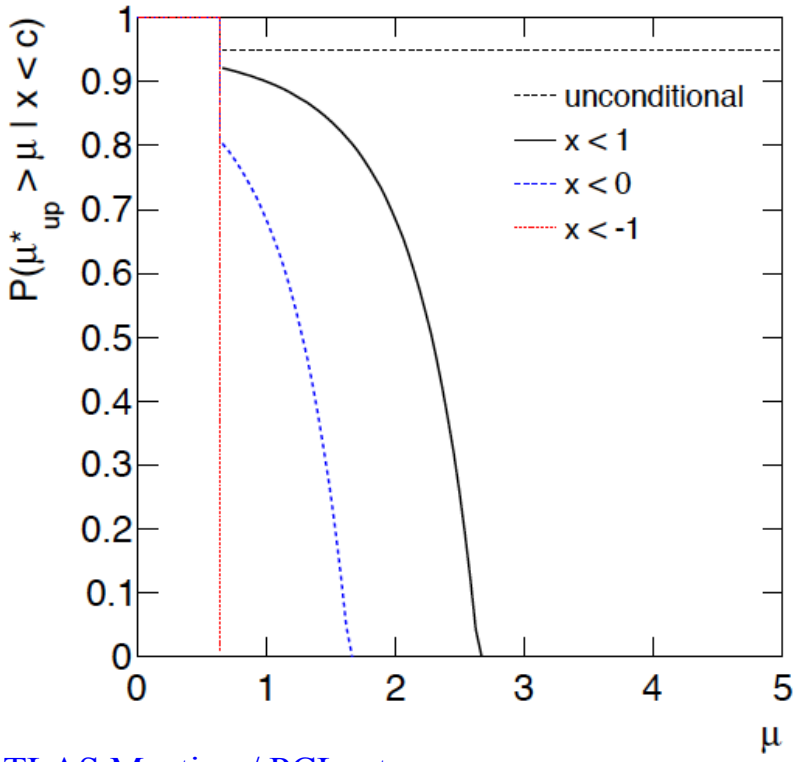
(Adapted) NBRS for PCL

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

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

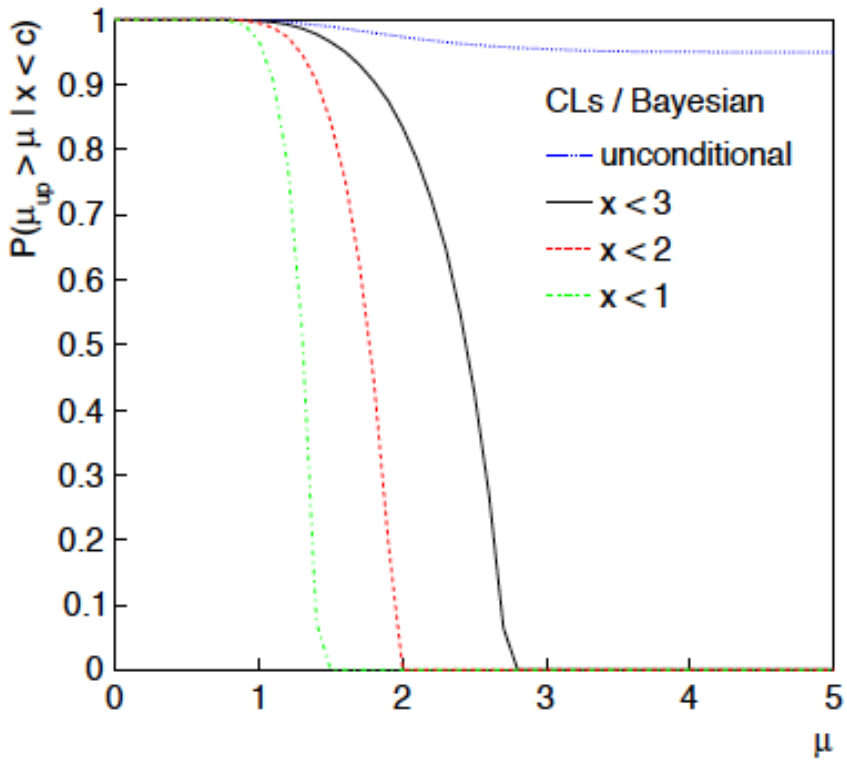
Coverage goes to 100% for $\mu < \mu_{\rm min}$, therefore no NBRS.

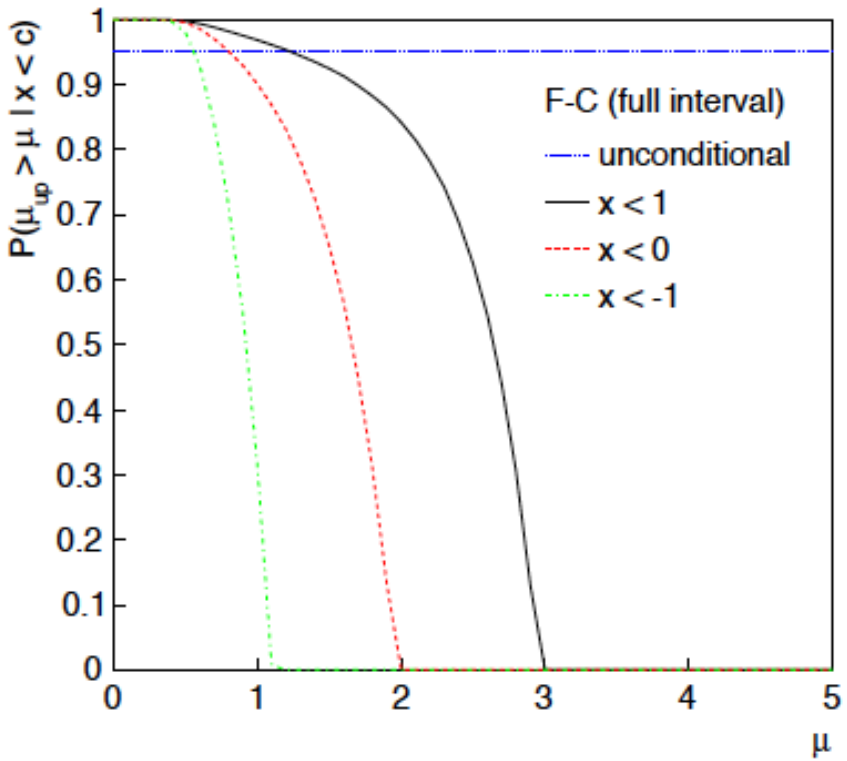
Note one does not have max conditional coverage $\geq 1-\alpha$ for all $\mu > \mu_{\rm min}$ ("adapted conditional coverage"). But if one conditions on μ, no limit would satisfy this.



← $1-\alpha$

Conditional coverage for CLs, F-C





Summary and conclusions

With a "usual" confidence 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).

PCL solves this problem by separating the parameter space into regions within which one has/hasn't sufficient sensitivity as given by the probability to reject μ if background-only model is true.

Recommendation for ATLAS: power $M_0(\mu) \geq 0.5$.

Within region with sufficient sensitivity, an upper limit can be set with a one-sided test (highest power) and exact $1 - \alpha$ coverage.

It is important to report both the constrained and unconstrained limits, so one can see where the power constraint comes into play.

Procedure easily adapted to problems with nuisance parameters (quote power at estimated values of nuisance parameters for $\mu = 0$).