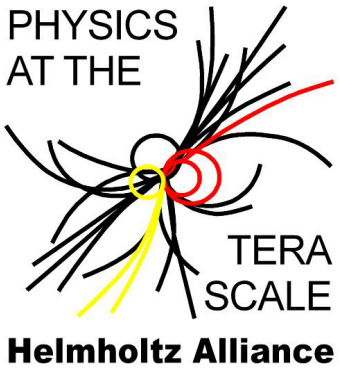
Statistical Methods for Discovery and Limits

Lecture 4: More on discovery and limits

http://www.pp.rhul.ac.uk/~cowan/stat_desy.html
https://indico.desy.de/conferenceDisplay.py?confId=4489



School on Data Combination and Limit Setting
DESY, 4-7 October, 2011

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

Outline

Lecture 1: Introduction and basic formalism
Probability, statistical tests, confidence intervals.

Lecture 2: Tests based on likelihood ratios
Systematic uncertainties (nuisance parameters)

Lecture 3: Limits for Poisson mean
Bayesian and frequentist approaches

→ Lecture 4: More on discovery and limits
Upper vs. unified limits (F-C)
Spurious exclusion, CLs, PCL
Look-elsewhere effect
Why 5σ for discovery?

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.)

Equivalently define a p-value p_μ such that the critical region corresponds to $p_\mu < \alpha$.

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

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

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 $1 - \alpha$.

The interval depends on the choice of the test, which is often based on considerations of power.

Power of a statistical test

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)

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

p-value of hypothesized μ

Choice of test for limits

Suppose 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.

The interesting alternative in this context is $\mu = 0$.

The critical region giving the highest power for the test of μ relative to the alternative of $\mu = 0$ thus contains low values of the data.

Test based on likelihood-ratio with respect to one-sided alternative → upper limit.

Choice of test for limits (2)

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.

For example, the process may be known to exist, and thus $\mu = 0$ is no longer an interesting alternative.

If the measure of incompatibility is taken to be the likelihood ratio with respect to a two-sided alternative, then the critical region can contain both high and low data values.

→ unified intervals, G. Feldman, R. Cousins, Phys. Rev. D 57, 3873–3889 (1998)

The Big Debate is whether to use one-sided or unified intervals in cases where the relevant alternative is at small (or zero) values of the parameter. Professional statisticians have voiced support on both sides of the debate.

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

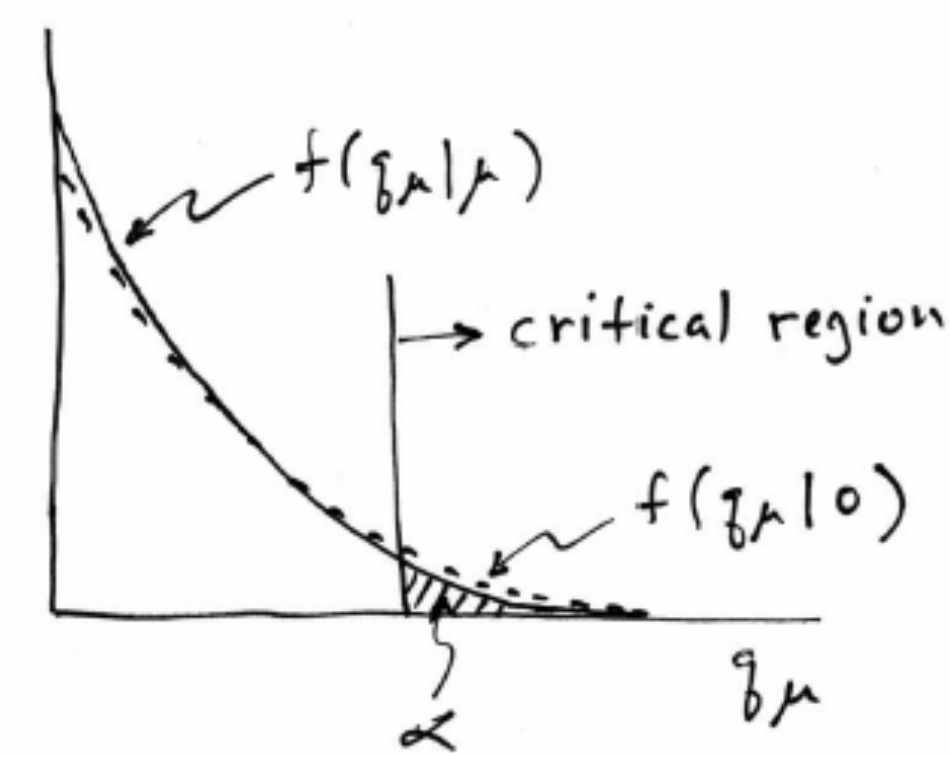
Large sample approximation: $$p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

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

Low sensitivity to μ

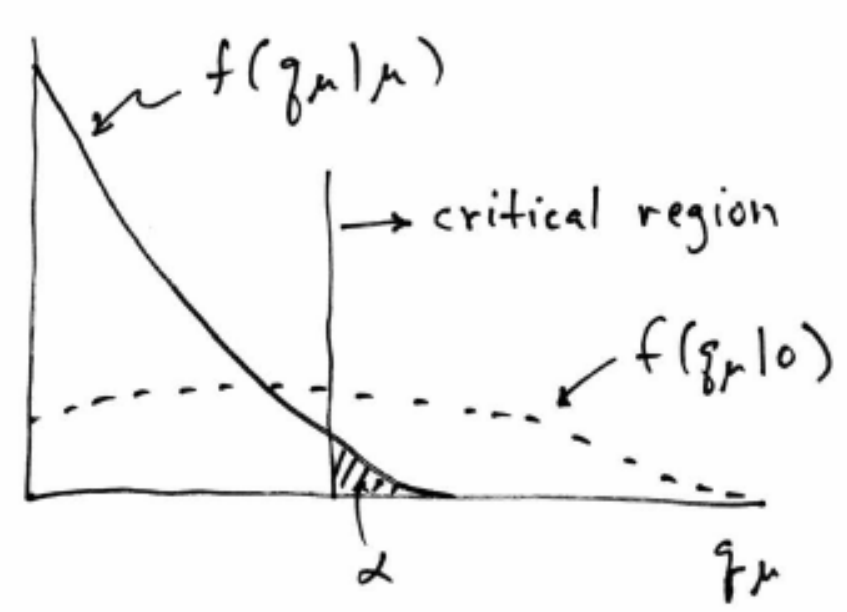
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

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:

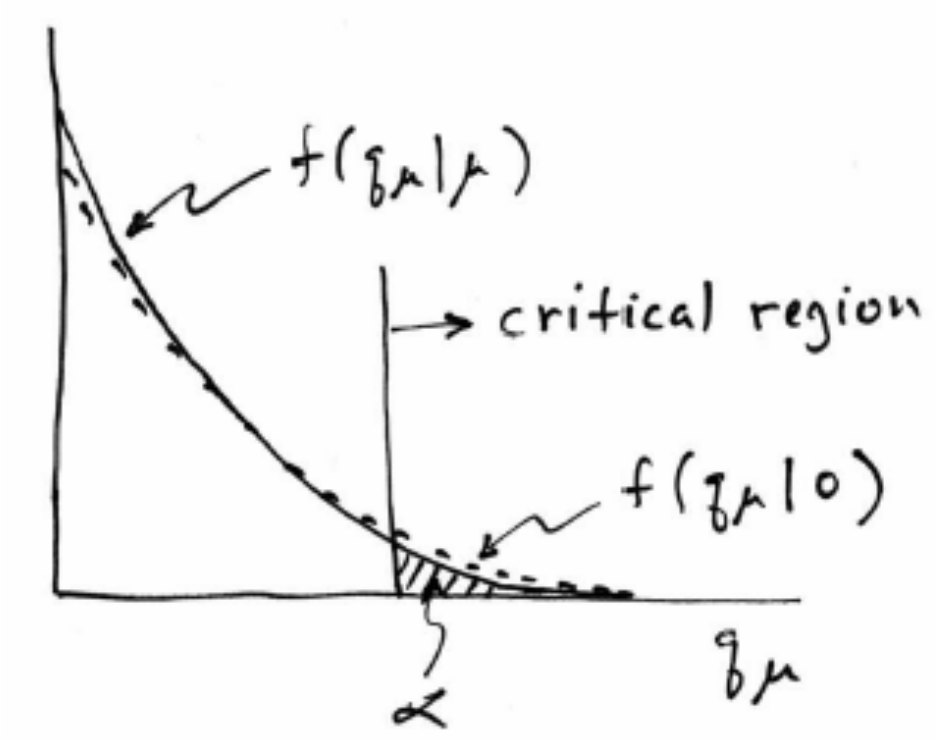


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α. Use this power as a measure of the sensitivity.

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

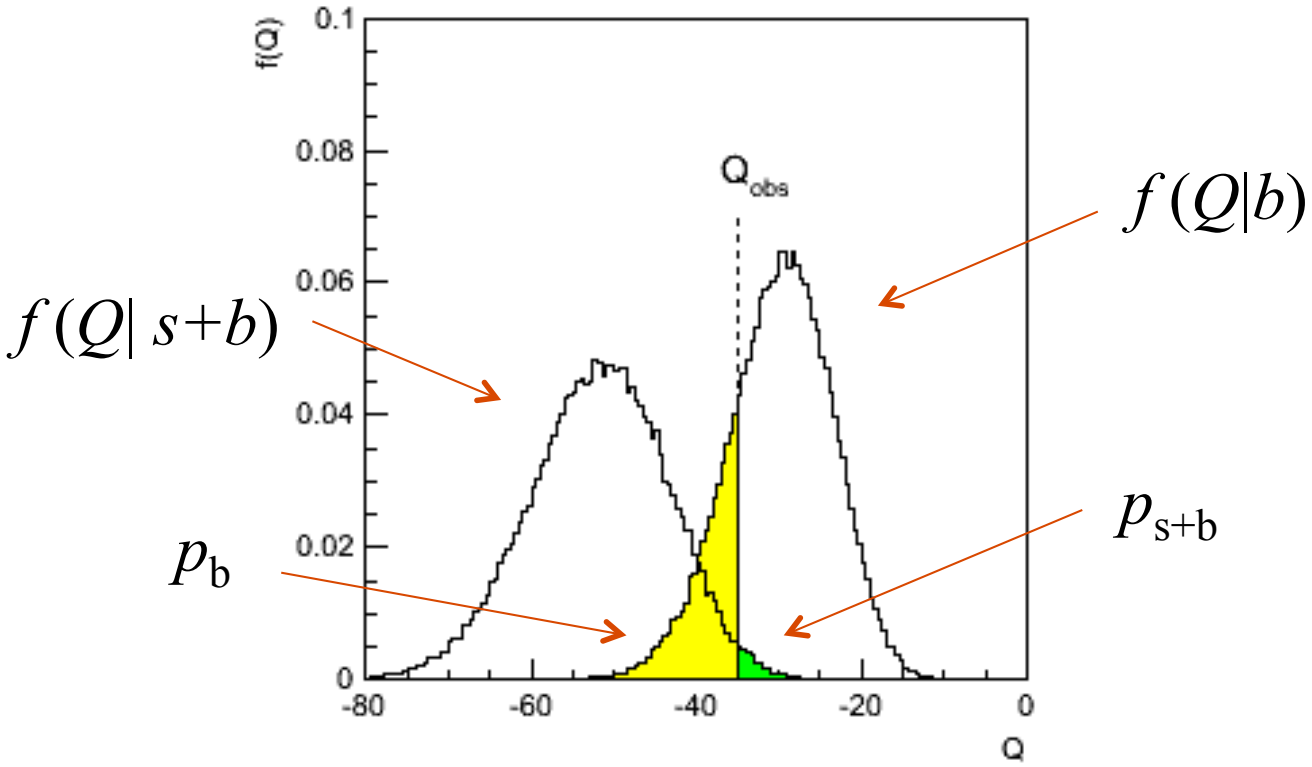
T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

and led to the "CL_s" procedure for upper limits.

Unified intervals also effectively reduce spurious exclusion by the particular choice of critical region.

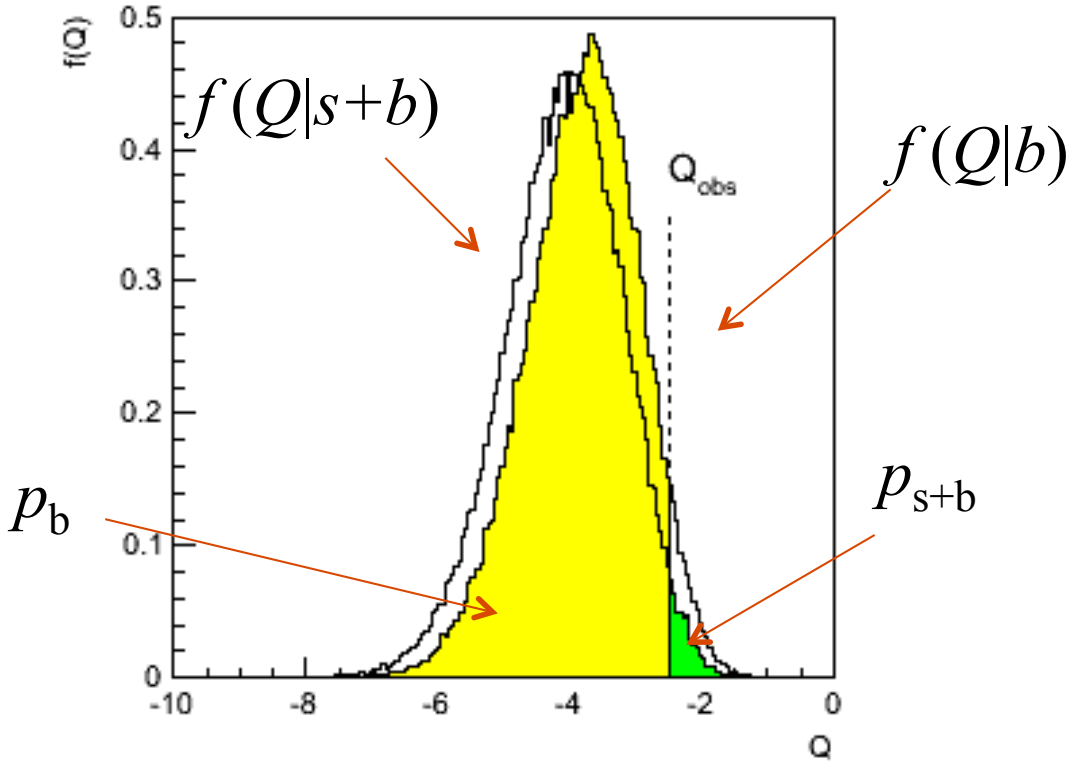
The CL_s procedure

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.,

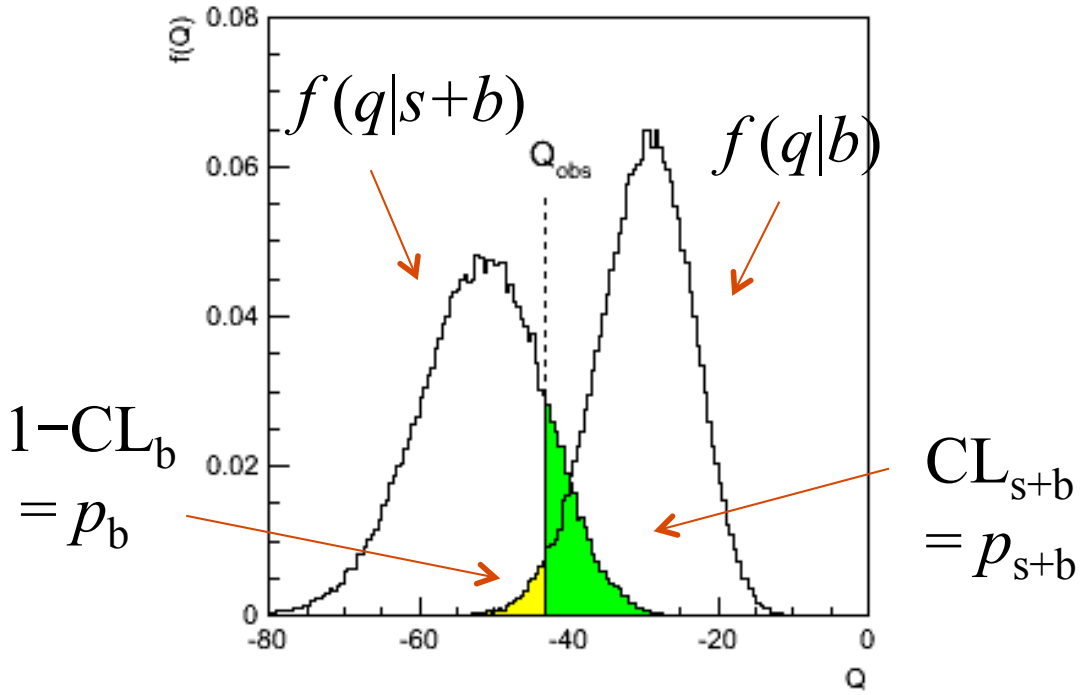
Define:

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

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

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).

Power Constrained Limits (PCL)

Cowan, Cranmer, Gross, Vitells, arXiv:1105.3166

CL_s has been criticized because the exclusion is based on a ratio of p-values, which did not appear to have a solid foundation.

The coverage probability of the CL_s upper limit is greater than the nominal CL = 1 − α by an amount that is generally not reported.

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:

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

Choice of minimum power

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

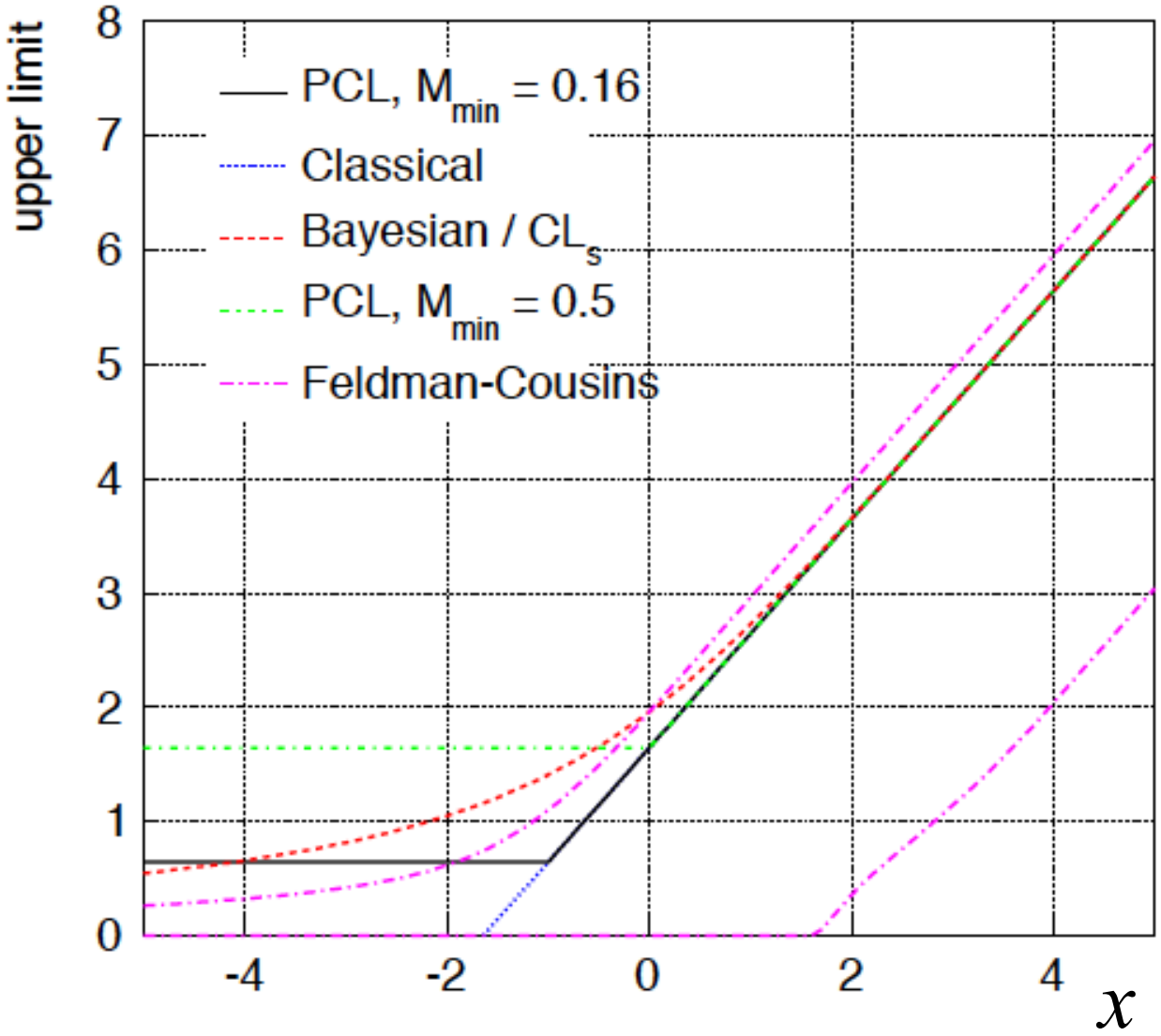
$$M_{\text{min}} = \Phi(-1) = 0.1587$$

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

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).

More recently for several reasons we have proposed $M_{\text{min}} = 0.5$, (which gives $\mu_{\text{min}} = 1.64\sigma$), i.e., one imposes the power constraint if the unconstrained limit fluctuations below its median under the background-only hypothesis.

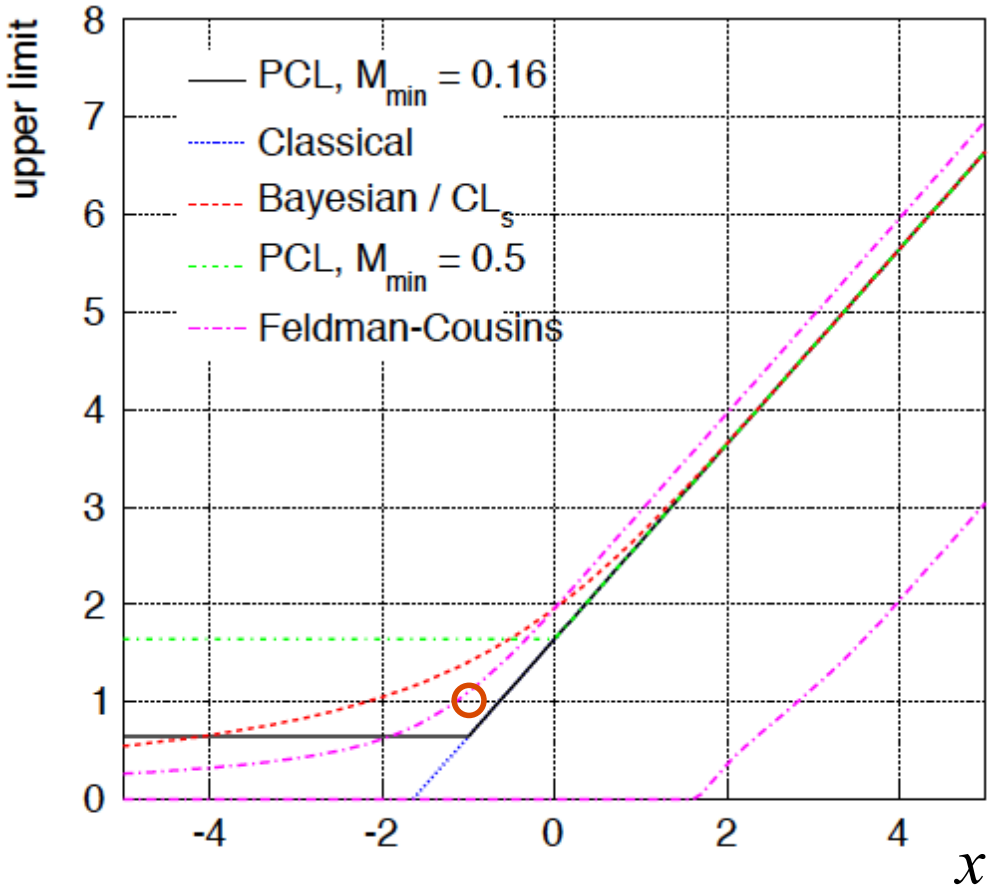
Upper limit on μ for $x \sim \text{Gauss}(\mu,\sigma)$ with $\mu \geq 0$



Comparison of reasons for (non)-exclusion

Suppose we observe $x = -1$.

$\mu = 1$ excluded by diag. line, why not by other methods?

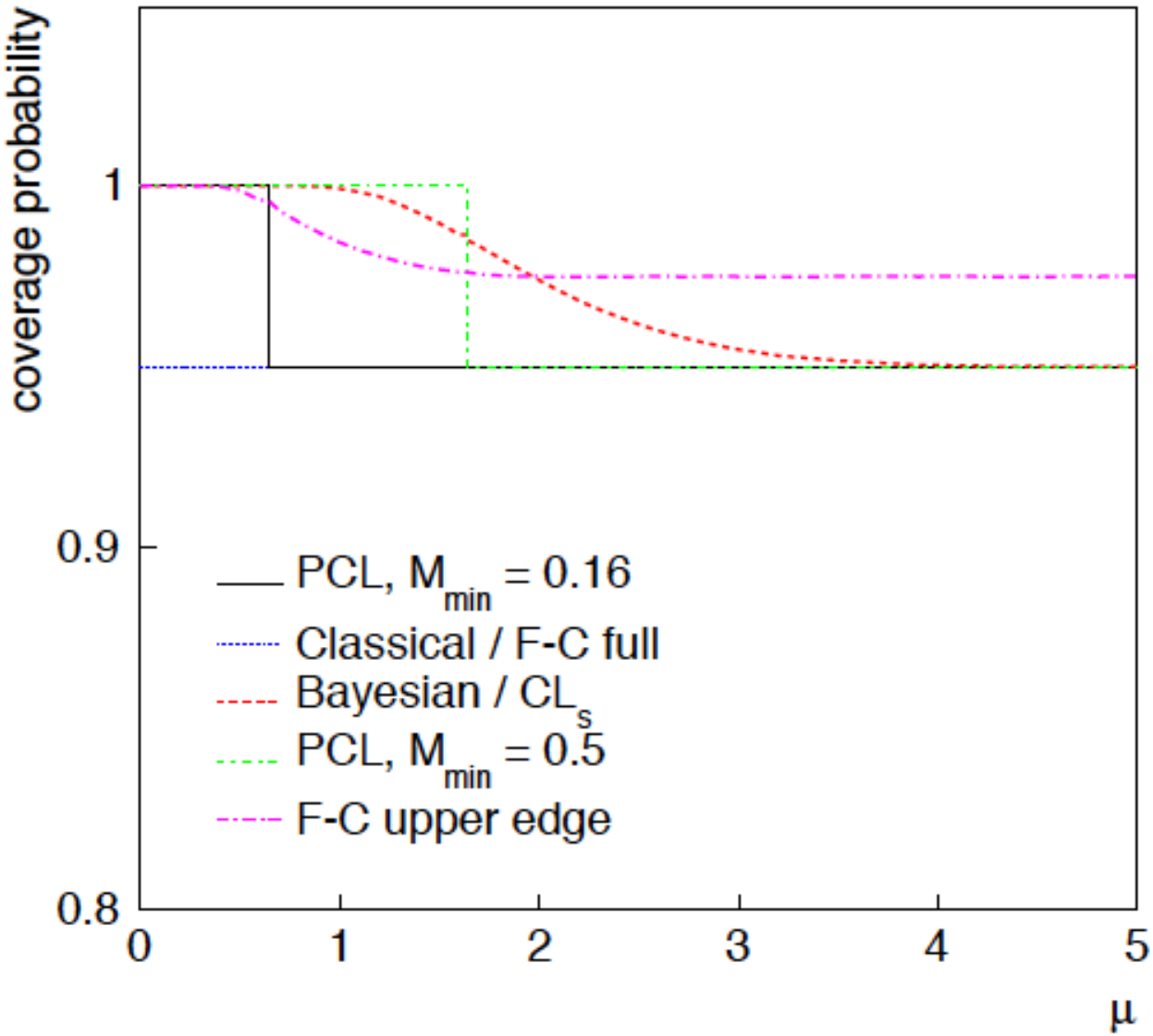


PCL (M_{min}=0.5): Because the power of a test of $\mu = 1$ was below threshold.

CLs: Because the lack of sensitivity to $\mu = 1$ led to reduced $1 - p_b$, hence CL_s not less than α.

F-C: Because $\mu = 1$ was not rejected in a test of size α (hence coverage correct). But the critical region corresponding to more than half of α is at high x.

Coverage probability for Gaussian problem



More thoughts on power*

**thanks to Ofer Vitells*

ALLAN BIRNBAUM Synthese 36 (1):5 - 13.

THE NEYMAN–PEARSON THEORY AS DECISION THEORY, AND AS INFERENCE THEORY; WITH A CRITICISM OF THE LINDLEY–SAVAGE ARGUMENT FOR BAYESIAN THEORY

Birnbaum formulates a concept of statistical evidence in which he states:

> A concept of statistical evidence is not plausible unless it finds 'strong evidence for H_2 as against H_1' with small probability (α) when H_1 is true, and with much larger probability ($1-\beta$) when H_2 is true.

More thoughts on power (2)*

**thanks to Ofer Vitells*

ON THE FOUNDATIONS OF STATISTICAL INFERENCE: BINARY EXPERIMENTS[1]

By Allan Birnbaum

The Annals of Mathematical Statistics,
Vol. 32, No. 2 (Jun., 1961), pp. 414-435

Thus in binary experiments *a small value of* α *does not in general imply high evidential strength in the outcome "reject* H_1*"*, and the determination of the evidential strength of such an outcome depends upon β as well as α, through the function $L_2 = (1 - \beta)/\alpha$.

This ratio is closely related to the exclusion criterion for CLs.

Birnbaum arrives at the conclusion above from the likelihood principle, which must be related to why CLs for the Gaussian and Poisson problems agree with the Bayesian result.

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.
See R. D. Cousins, arXiv:1109.2023

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.)

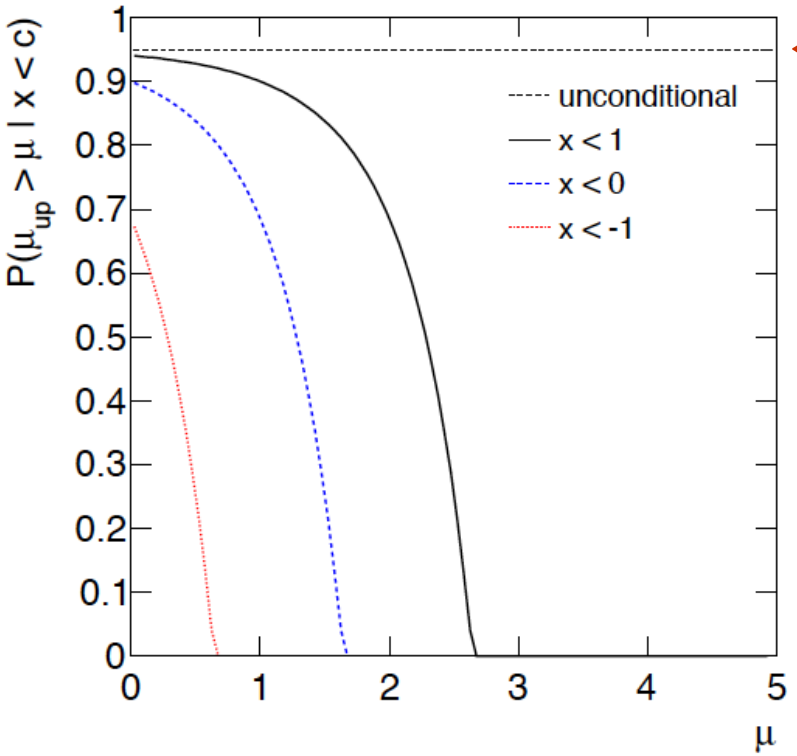
NBRS for unconstrained upper limit

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.



← $1 - \alpha$

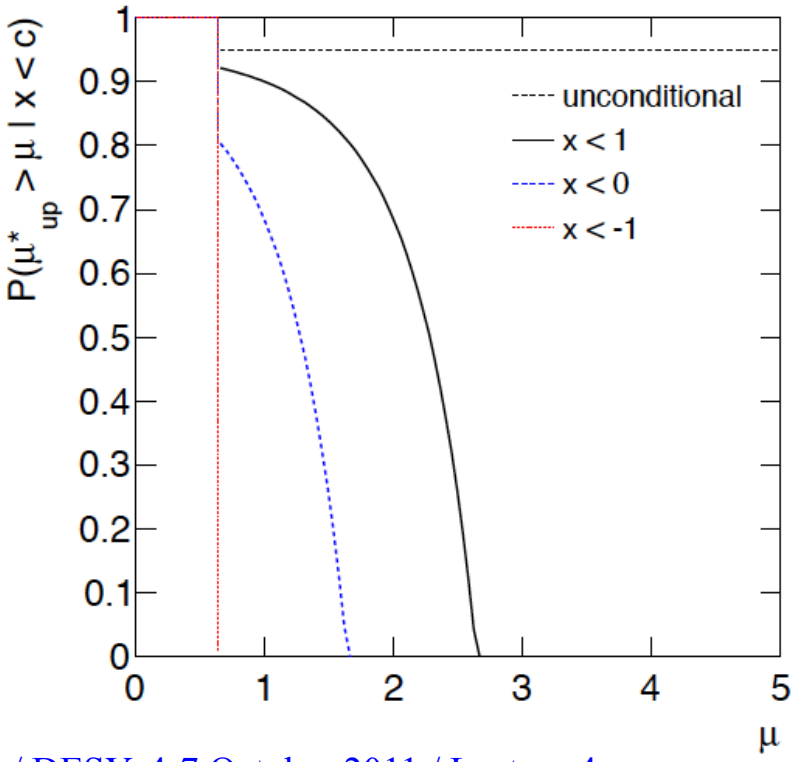
(Adapted) NBRS for PCL

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

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

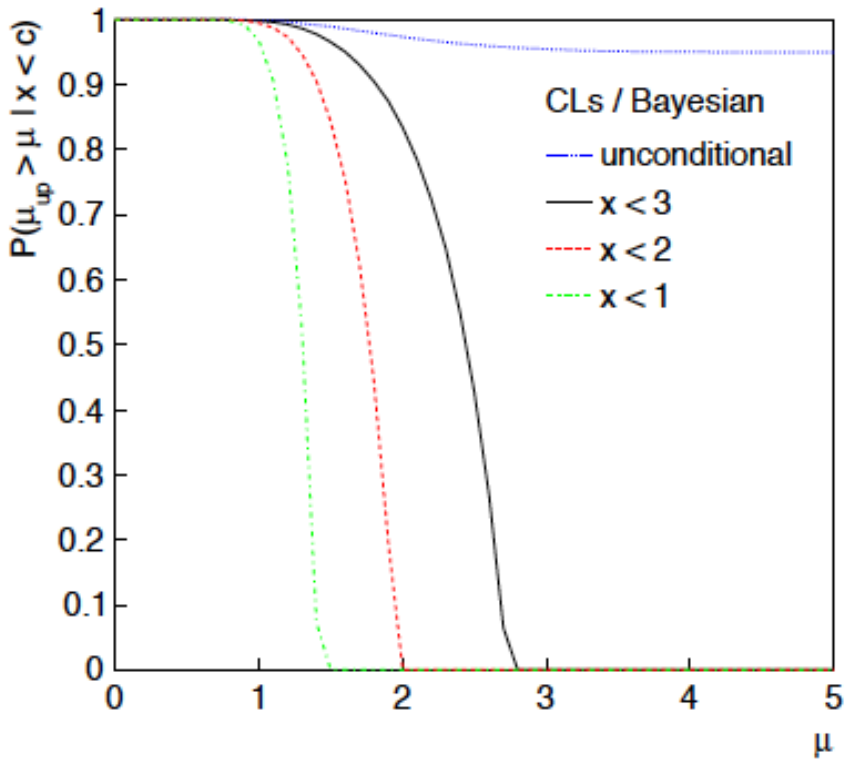
Coverage goes to 100% for $\mu < \mu_{\rm min}$, therefore no NBRS.

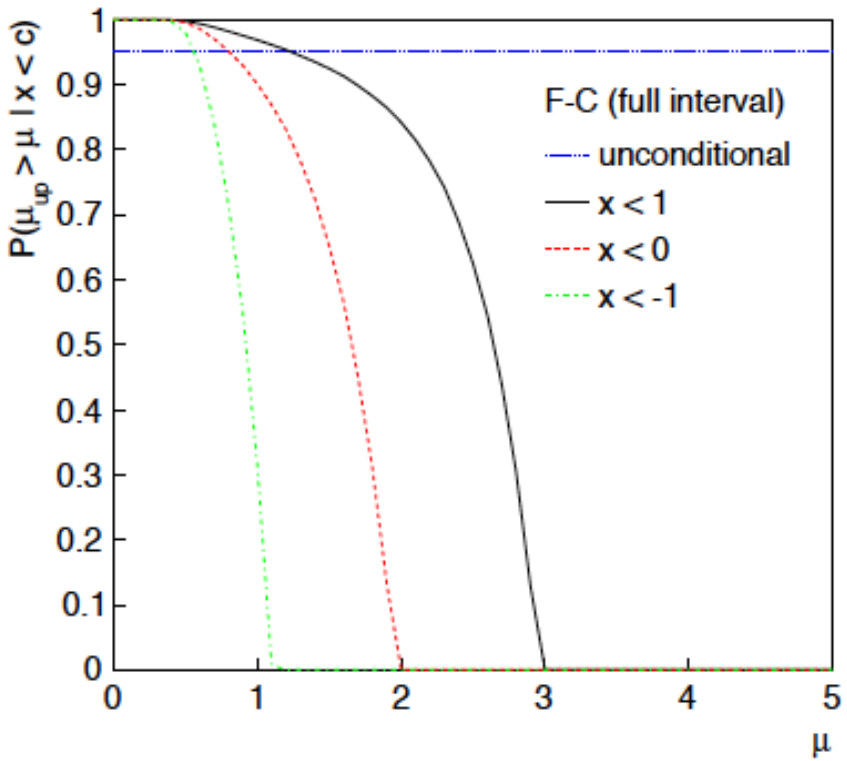
Note one does not have max conditional coverage $\geq 1-\alpha$ for all $\mu > \mu_{\rm min}$ ("adapted conditional coverage"). But if one conditions on μ, no limit would satisfy this.



← $1-\alpha$

Conditional coverage for CLs, F-C



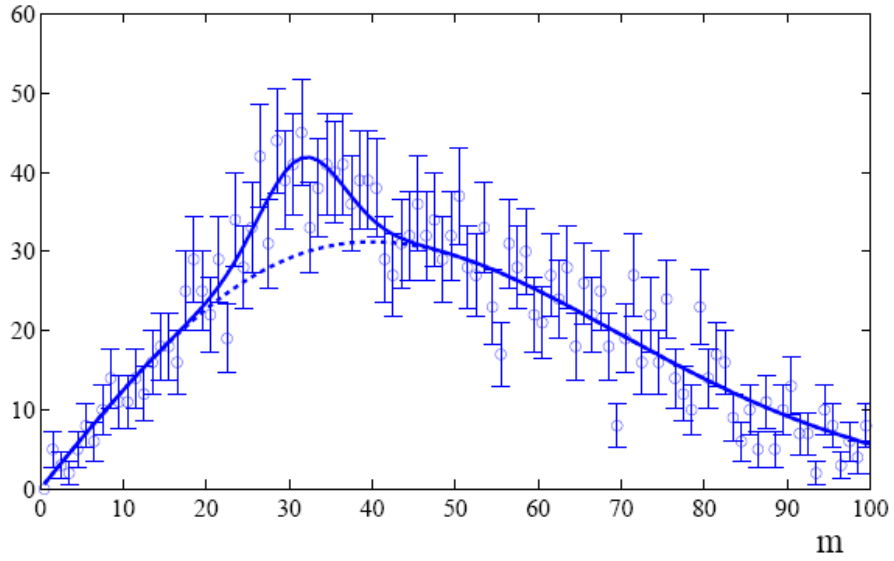


The Look-Elsewhere Effect

Gross and Vitells, EPJC 70:525-530,2010, arXiv:1005.1891

Suppose a model for a mass distribution allows for a peak at a mass m with amplitude μ.

The data show a bump at a mass m_0.



How consistent is this with the no-bump ($\mu = 0$) hypothesis?

p-value for fixed mass

First, suppose the mass m_0 of the peak was specified a priori.

Test consistency of bump with the no-signal ($\mu = 0$) hypothesis with e.g. likelihood ratio

$$t_{\text{fix}} = -2 \ln \frac{L(0, m_0)}{L(\hat{\mu}, m_0)}$$

where "fix" indicates that the mass of the peak is fixed to m_0.

The resulting p-value

$$p_{\text{fix}} = \int_{t_{\text{fix,obs}}}^{\infty} f(t_{\text{fix}}|0) dt_{\text{fix}}$$

gives the probability to find a value of t_{fix} at least as great as observed at the specific mass m_0.

p-value for floating mass

But suppose we did not know where in the distribution to expect a peak.

What we want is the probability to find a peak at least as significant as the one observed anywhere in the distribution.

Include the mass as an adjustable parameter in the fit, test significance of peak using

$$t_{\text{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{m})}$$

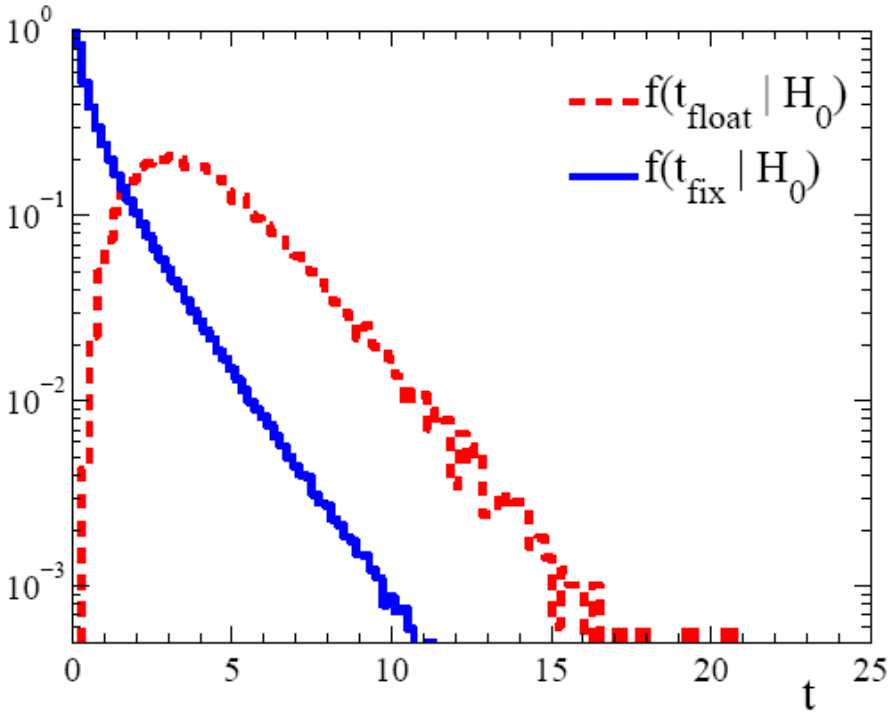
(Note m does not appear in the $\mu = 0$ model.)

$$p_{\text{float}} = \int_{t_{\text{float,obs}}}^{\infty} f(t_{\text{float}}|0) dt_{\text{float}}$$

Distributions of t_{fix}, t_{float}

For a sufficiently large data sample, t_{fix} ~chi-square for 1 degree of freedom (Wilks' theorem).

For t_{float} there are two adjustable parameters, μ and m, and naively Wilks theorem says t_{float} ~ chi-square for 2 d.o.f.



In fact Wilks' theorem does not hold in the floating mass case because on of the parameters (m) is not-defined in the $\mu = 0$ model.

So getting t_{float} distribution is more difficult.

Trials factor

We would like to be able to relate the *p*-values for the fixed and floating mass analyses (at least approximately).

Gross and Vitells show that the "trials factor" can be approximated by

$$F_{\text{trials}} \equiv \frac{p_{\text{float}}}{p_{\text{fix}}} \approx 1 + \sqrt{\frac{\pi}{2}}\langle \mathcal{N} \rangle Z_{\text{fix}}$$

where ‹*N*› = average number of "upcrossings" of −2ln*L* in fit range and

$$Z_{\text{fix}} = \Phi^{-1}(1 - p_{\text{fix}}) = \sqrt{t_{\text{fix}}}$$

is the significance for the fixed mass case.

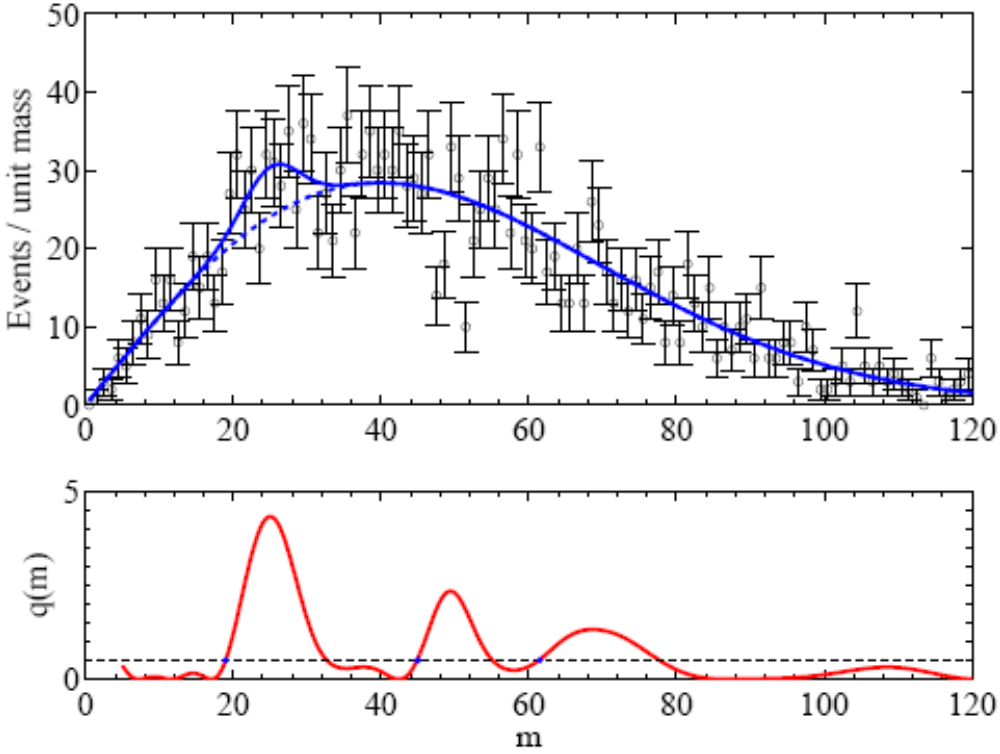
So we can either carry out the full floating-mass analysis (e.g. use MC to get *p*-value), or do fixed mass analysis and apply a correction factor (much faster than MC).

Upcrossings of $-2\ln L$

The Gross-Vitells formula for the trials factor requires the mean number "upcrossings" of $-2\ln L$ in the fit range based on fixed threshold.

$$P(q_0 > u)$$
$$\leq E[N_u] + P(q_0(0) > u)$$
$$= \mathcal{N}_1 e^{-u/2} + \frac{1}{2} P(\chi_1^2 > u)$$

estimate with MC at low reference level



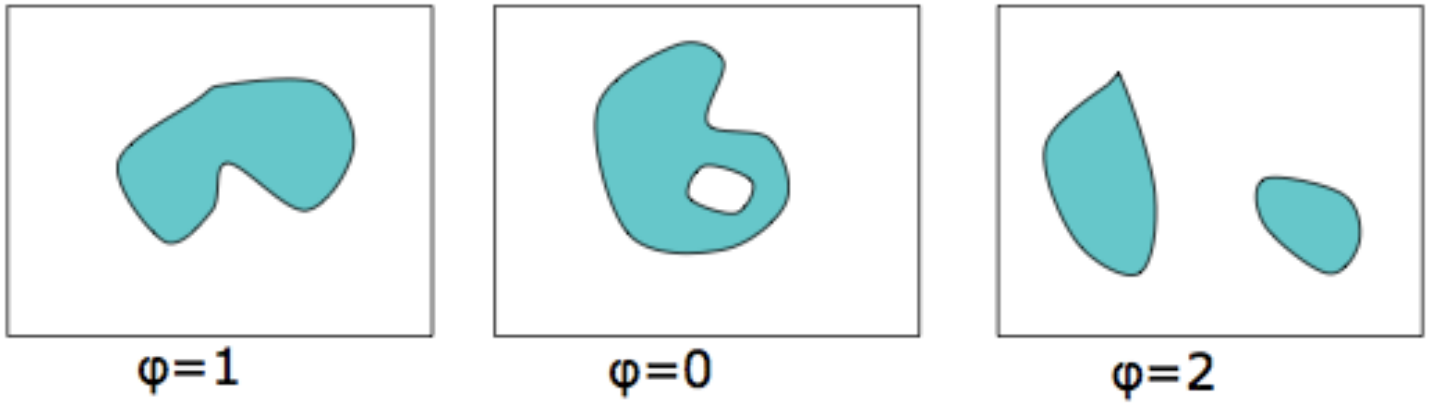
Vitells and Gross, arXiv:1105.4355

Multidimensional look-elsewhere effect

Generalization to multiple dimensions: number of upcrossings replaced by expectation of Euler characteristic:

$$\mathrm{E}[\varphi(A_u)] = \sum_{d=0}^{n} \mathcal{N}_d \rho_d(u)$$

- Number of disconnected components minus number of `holes'



Applications: astrophysics (coordinates on sky), search for resonance of unknown mass and width, ...

Summary on Look-Elsewhere Effect

Remember the Look-Elsewhere Effect is when we test a single model (e.g., SM) with multiple observations, i..e, in mulitple places.

Note there is no look-elsewhere effect when considering exclusion limits. There we test specific signal models (typically once) and say whether each is excluded.

With exclusion there is, however, the analogous issue of testing many signal models (or parameter values) and thus excluding some even in the absence of signal ("spurious exclusion")

Approximate correction for LEE should be sufficient, and one should also report the uncorrected significance.

> "There's no sense in being precise when you don't even know what you're talking about." — John von Neumann

Why 5 sigma?

Common practice in HEP has been to claim a discovery if the p-value of the no-signal hypothesis is below 2.9×10^{-7}, corresponding to a significance $Z = \Phi^{-1}(1 - p) = 5$ (a 5σ effect).

There a number of reasons why one may want to require such a high threshold for discovery:

The "cost" of announcing a false discovery is high.

Unsure about systematics.

Unsure about look-elsewhere effect.

The implied signal may be a priori highly improbable (e.g., violation of Lorentz invariance).

Why 5 sigma (cont.)?

But the primary role of the *p*-value is to quantify the probability that the background-only model gives a statistical fluctuation as big as the one seen or bigger.

It is not intended as a means to protect against hidden systematics or the high standard required for a claim of an important discovery.

In the processes of establishing a discovery there comes a point where it is clear that the observation is not simply a fluctuation, but an "effect", and the focus shifts to whether this is new physics or a systematic.

Providing LEE is dealt with, that threshold is probably closer to 3σ than 5σ.

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.

Thanks

Many thanks to Bob, Eilam, Ofer, Kyle, Alex.

Vielen Dank an die Organisatoren und Teilnehmer.

Extra slides

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

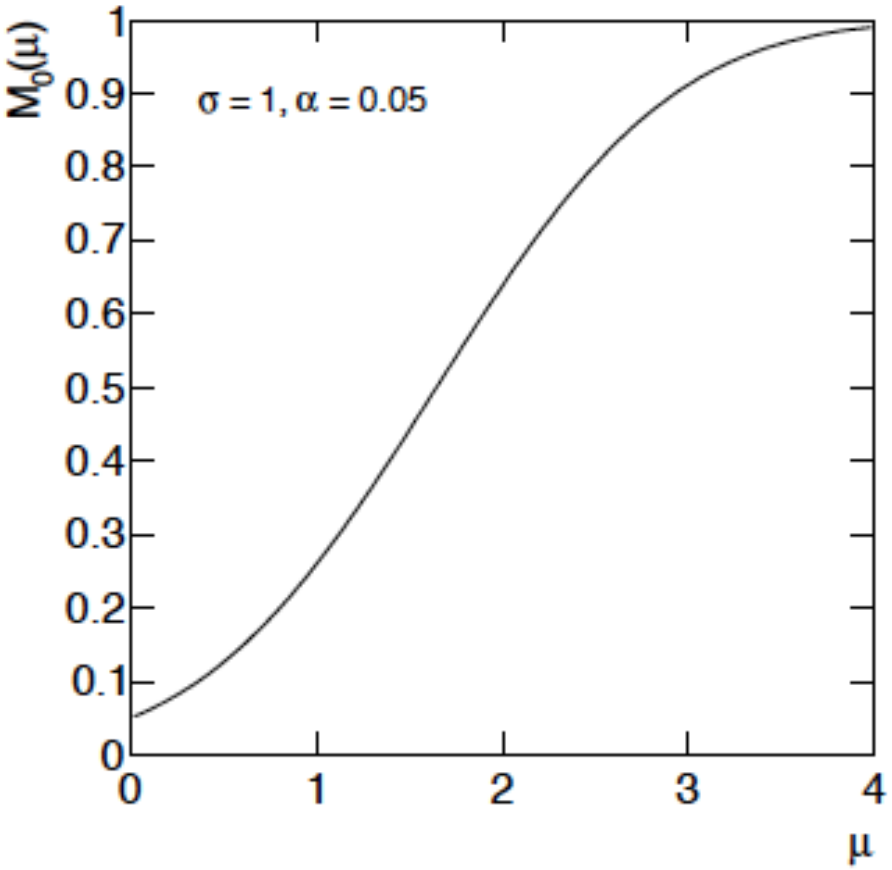
inverse of standard Gaussian cumulative distribution

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.

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 $\mu_{\rm up}$ 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.

Flip-flopping

F-C pointed out that if one decides, based on the data, whether to report a one- or two-sided limit, then the stated coverage probability no longer holds.

The problem (flip-flopping) is avoided in unified intervals.

Whether the interval covers correctly or not depends on how one defines repetition of the experiment (the ensemble).

Need to distinguish between:

(1) an idealized ensemble;

(2) a recipe one follows in real life that resembles (1).

Flip-flopping

One could take, e.g.:

Ideal: always quote upper limit (∞ # of experiments).

Real: quote upper limit for as long as it is of any interest, i.e., until the existence of the effect is well established.

The coverage for the idealized ensemble is correct.

The question is whether the real ensemble departs from this during the period when the limit is of any interest as a guide in the search for the signal.

Here the real and ideal only come into serious conflict if you think the effect is well established (e.g. at the 5 sigma level) but then subsequently you find it not to be well established, so you need to go back to quoting upper limits.

Flip-flopping

In an idealized ensemble, this situation could arise if, e.g., we take $x \sim \text{Gauss}(\mu, \sigma)$, and the true μ is one sigma below what we regard as the threshold needed to discover that μ is nonzero.

Here flip-flopping gives undercoverage because one continually bounces above and below the discovery threshold. The effect keeps going in and out of a state of being established.

But this idealized ensemble does not resemble what happens in reality, where the discovery sensitivity continues to improve as more data are acquired.