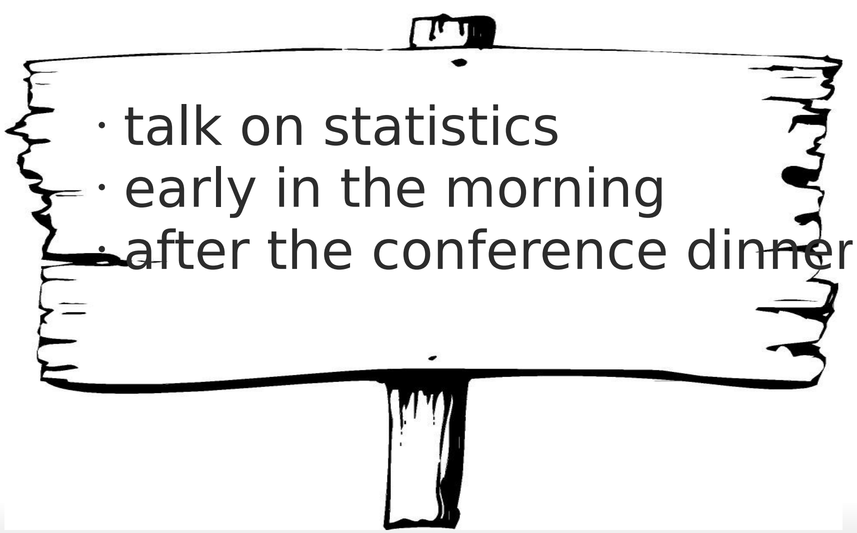
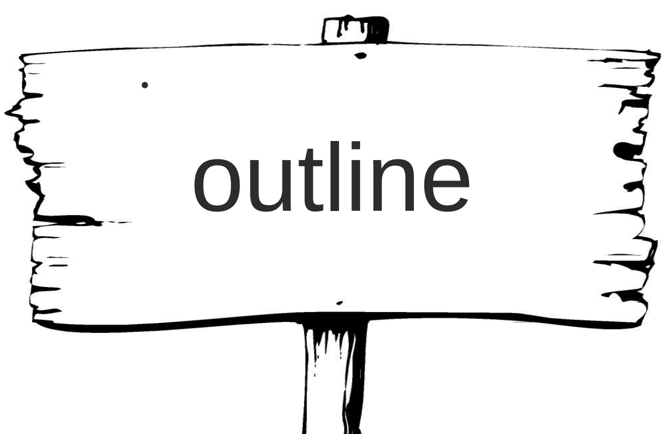
selected statistics topics
in neutrino telescopes

Aart Heijboer, Nikhef Amsterdam



selected statistics topics in neutrino telescopes

Aart Heijboer, Nikhef Amsterdam



- Searches
  - common aspects → observable
- Discovery
- Limits
  - from counting to continuous
  - problems with 'Neyman' limits
  - alternatives (PC, CLs, FC)
- Nuisance parameters and external constraints

not much about: reconstruction, event classification, Bayesian methods, measuring parameters, multi-variate methods, unfolding, ...

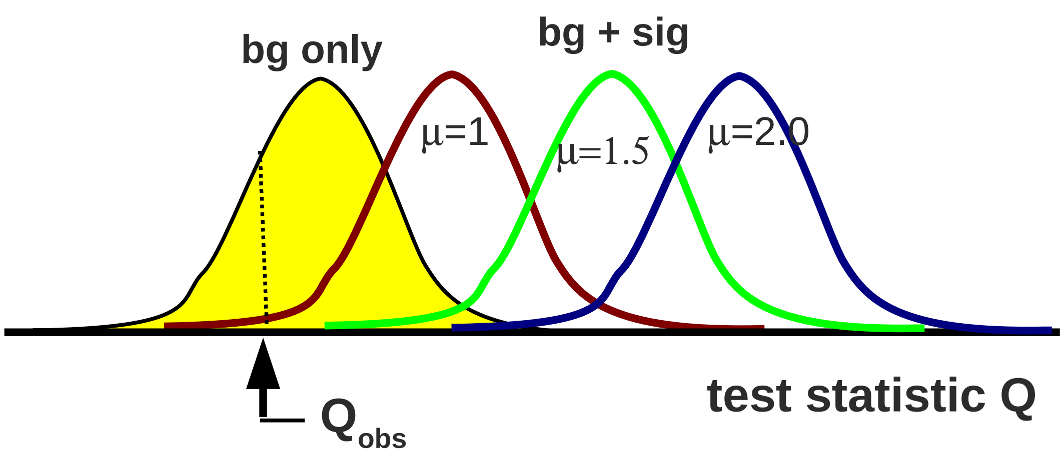
Introduction: searches

Common element: some observable which distinguishes signal from background

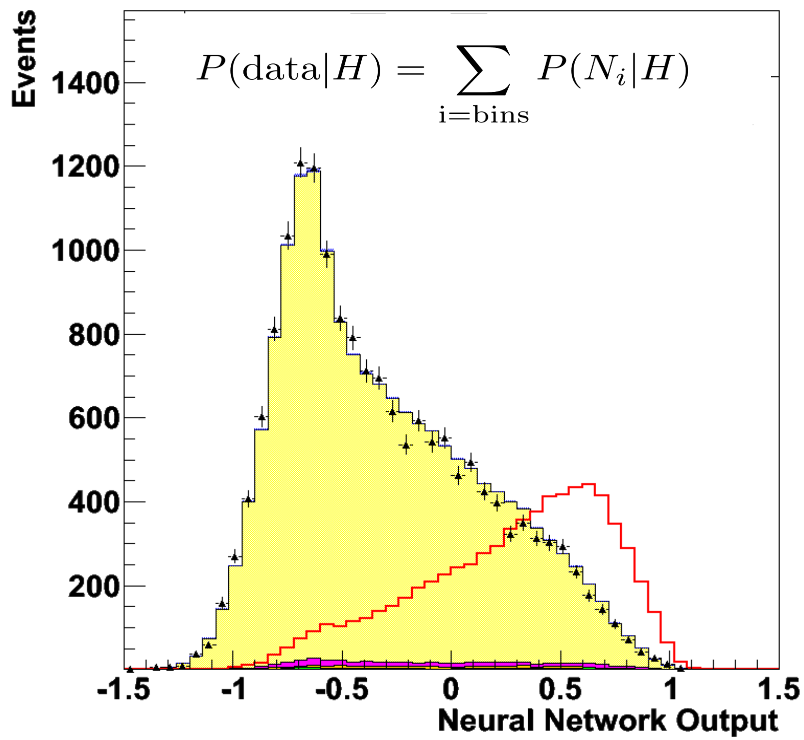
- number of events after cuts
- result of some multivariate method. BDT, NN, etc, or likelihood ratio Q

$$Q = \frac{\mathcal{L}^{s+b}}{\mathcal{L}^{b}} = \frac{P(\mathrm{data}|\mathrm{s+b})}{P(\mathrm{data}|\mathrm{b})} = \frac{P(\mathrm{data}|H_1)}{P(\mathrm{data}|H_0)}$$

end up with distributions of the test statistic (from MC) and one observed value (from data)



BDT, NN etc produce per-event output → cut and count or likelihood ratio from output



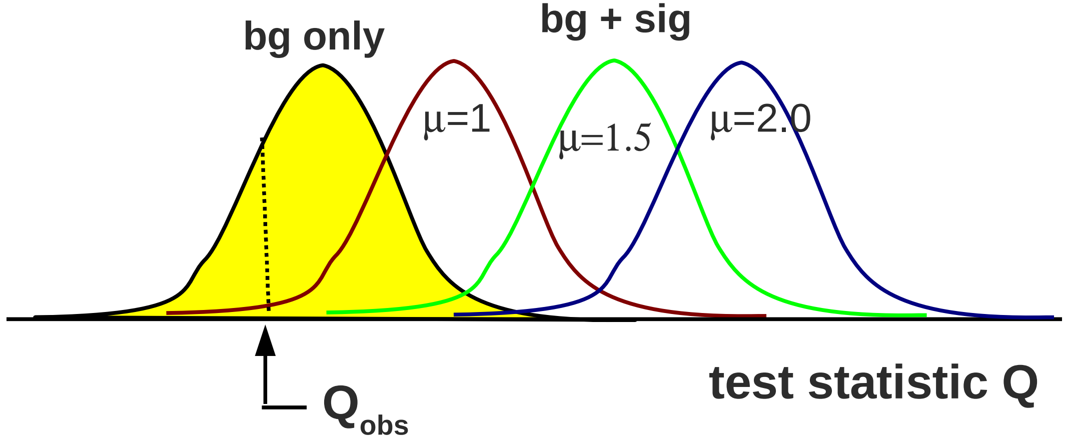
Introduction: searches

Common element: some observable (a.ka. test-statistic) which distinguishes signal from background

- number of events after cuts
- result of some multivariate method. BDT, NN, etc, or likelihood ratio Q

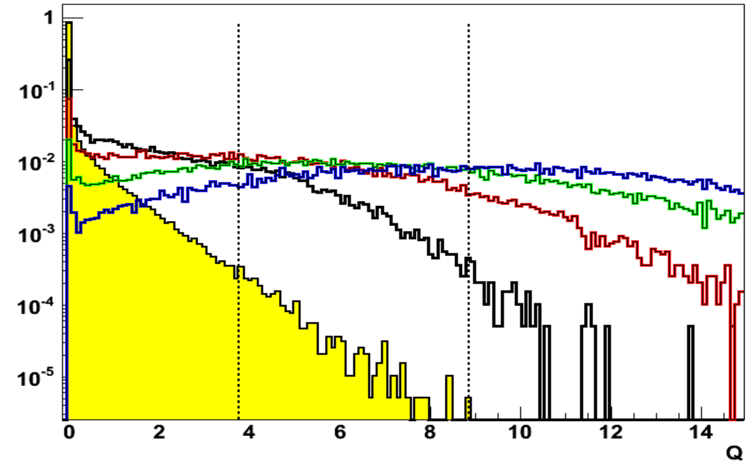
$$Q = \frac{\mathcal{L}^{s+b}}{\mathcal{L}^{b}} = \frac{P(\mathrm{data}|\mathrm{s+b})}{P(\mathrm{data}|\mathrm{b})} = \frac{P(\mathrm{data}|H_1)}{P(\mathrm{data}|H_0)}$$

end up with distributions of the test statistic (from MC) and one observed value (from data)



μ = expectation value of the signal size, here expressed in number of events.

μ can be predicted by some theory



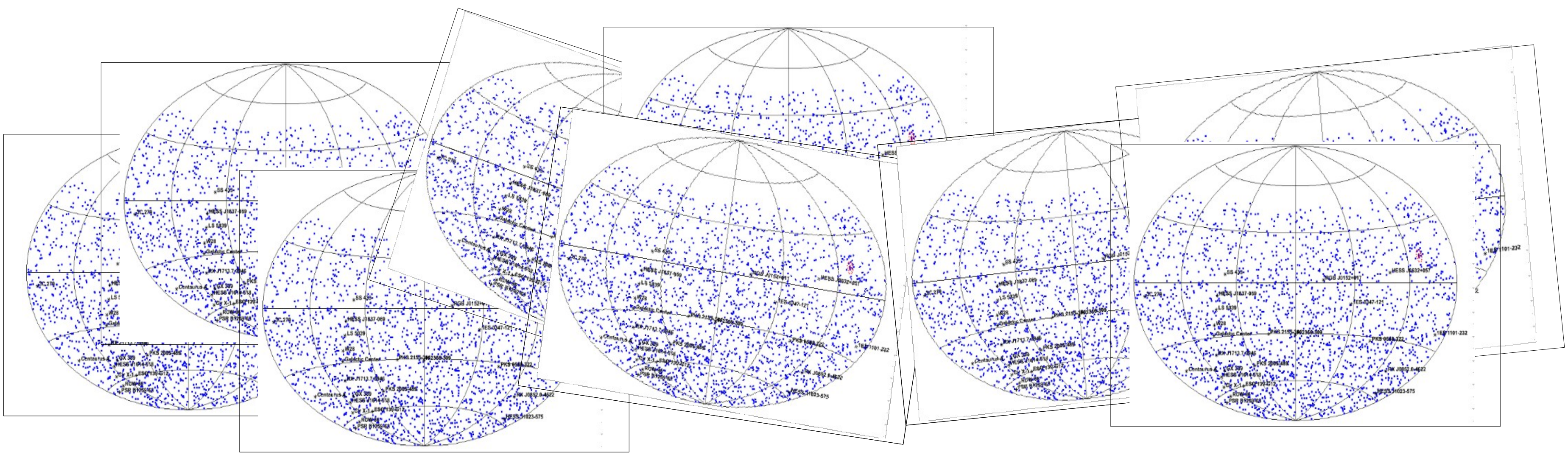
realistic version, produced by doing pseudo-experiments

pseudo-experiments

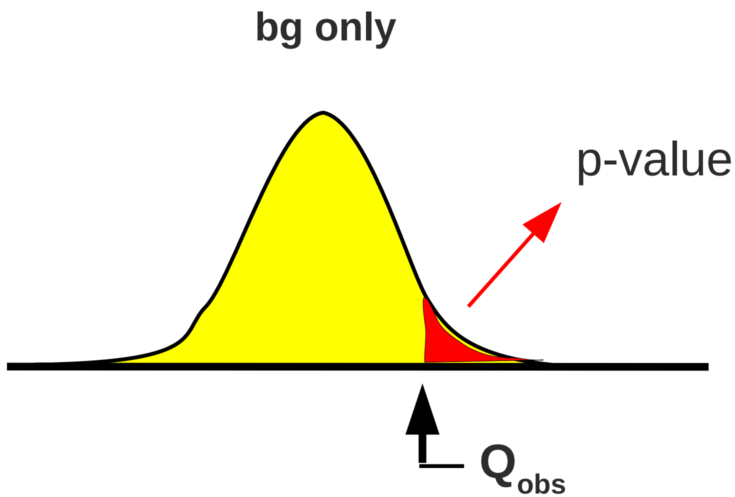
- To compute distribution of the observable (for bg-only and bg+sig hypotheses) do the full analysis, on 'toy' simulation on the dataset
  - not needed for counting experiment, but pretty much only solution for complex observables.

- Can sometimes find clever way to make independent toys (out of data), by randomizing or scrambling some key variable (eg.: randomize ra in pnt source search)

- Easy to conclude systematics : just vary the toys withing the syst. uncertainty.



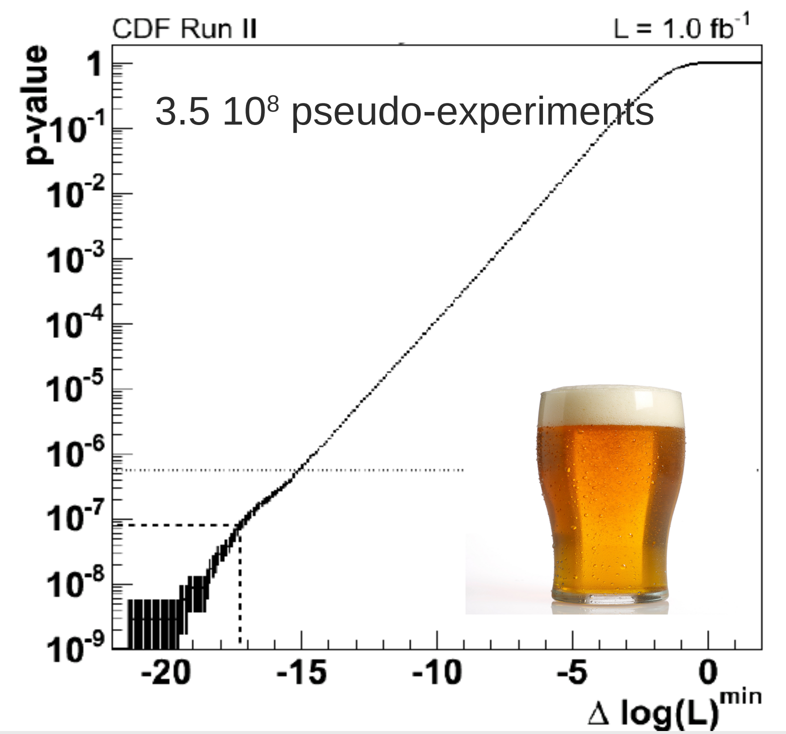
making a discovery



- significance quantified by p-value
- translates into "number of sigma's" (single or double sided convention)
- need to compute $p(Q_{obs})$

only minor problems:

- deal with trial-factors / look-elsewhere effect
  - can get philosophical ← just describe what was done.
  - once you decide what you want, it's easy with pseudo-experiments
- 5σ means running > 10^8 pseudo-experiments
  - usually not possible → extrapolate
  - math available: Wilk's theorem
  - would love to have that problem!



limit setting

- surprisingly hard:
- choices involved that matter for the numbers
  different limit setting method can change result by factor 2
- possibility of nonsense results
- statisticians do not agree which method to use (let alone physicists)

- not the end of the world, but good to be aware of some of the issues, especially when comparing experiments or using external data as input.

limit setting : coverage

the probability to get the data we got, or even
less signal-like data, is very small if the signal would be so-and-so large.

this is what we would like (coverage):

$$P(\text{ limit(data)} \geq \mu\ ) = 0.10$$

limit = random number
(because function of the data)

μ = a non-random, fixed number
(of which we don't know the value)

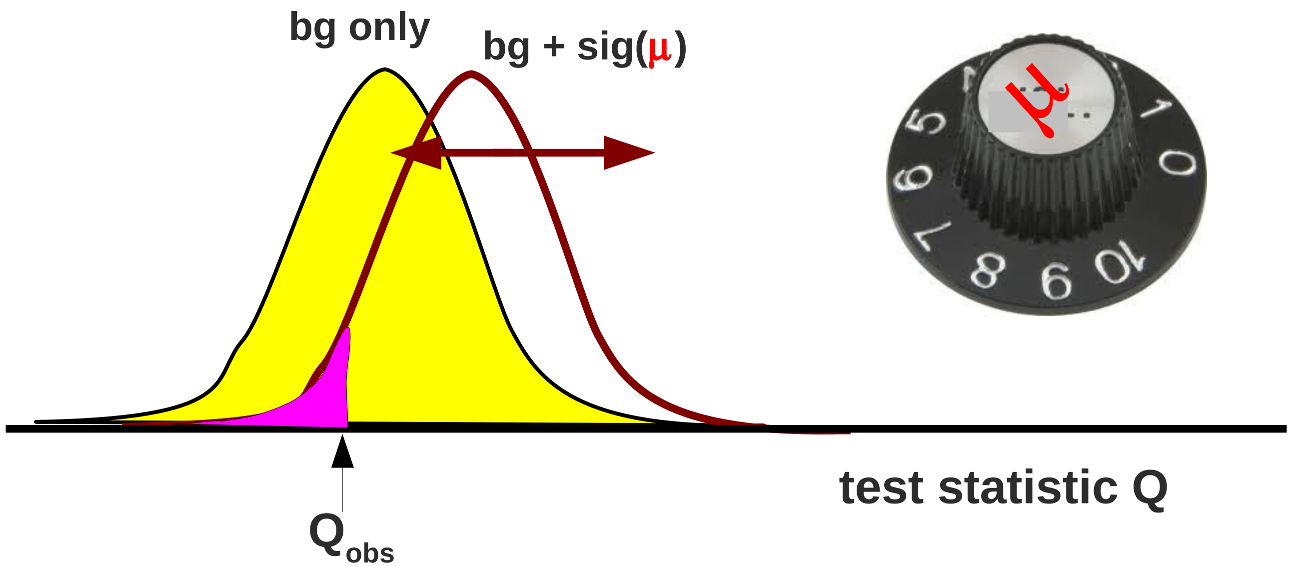
- This does not tell you what to do; i.e. how to define the function limit(data)

- Even with perfect coverage, one can still get 'undesired' results (examples follow)

- Talk only about frequentist limits
  Bayesian: compute PDF(μ | data, prior) and integrate to 10%
  → free of all the problems I will discuss next, still not prevalent (have to choose prior)

'Neyman' limits

* assume 90% CL everywhere



- Find the signal strength (μ) for which $P(Q_{obs}|\mu)$ = 10*%

- Note the bg-only distribution is not used (!)
- $P(Q_{obs}|\mu)$ is also called CL_{s+b}

- 'Neyman limits' is not the prefered nomenclature, since this is only one example of a Neyman construction. can also call them CL_{s+b}-limits.

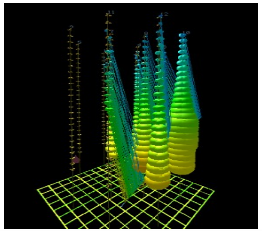
Example: counting experiment

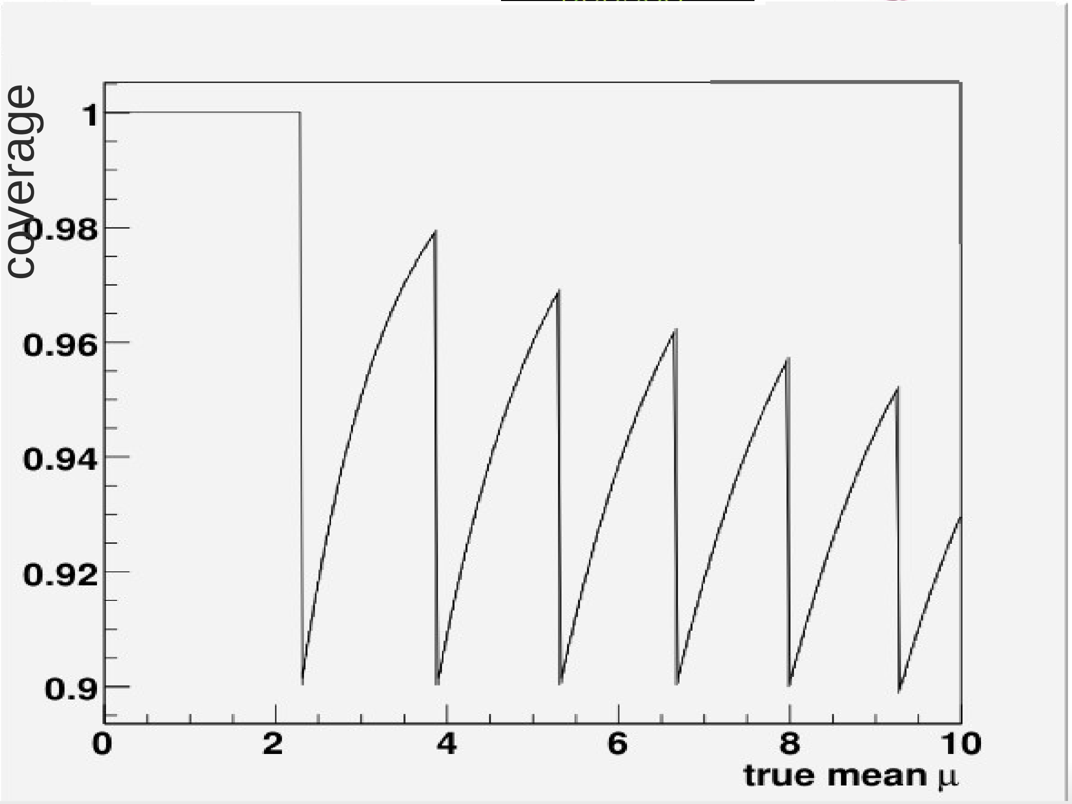
$$\mathrm{P}\ (\ \mathrm{N}(\mu) \leq N_{\text{obs}} | \mu) = 10\%$$

- Outcome of the experiment is discrete
- All experiments with a given N_{obs} must produce the same limit
  → exact coverage not possible and forced to be conservative (≤ sign)
- In low-background regime, the lowest possible limit is 2.3 signal events.
- (Severe) over-coverage
  - can live with that, but keep in mind if competing analysis does a lot better

N_{obs}	limit
0	2.30
1	3.89
2	5.32
3	6.68

For small expected background (pink elephant search)





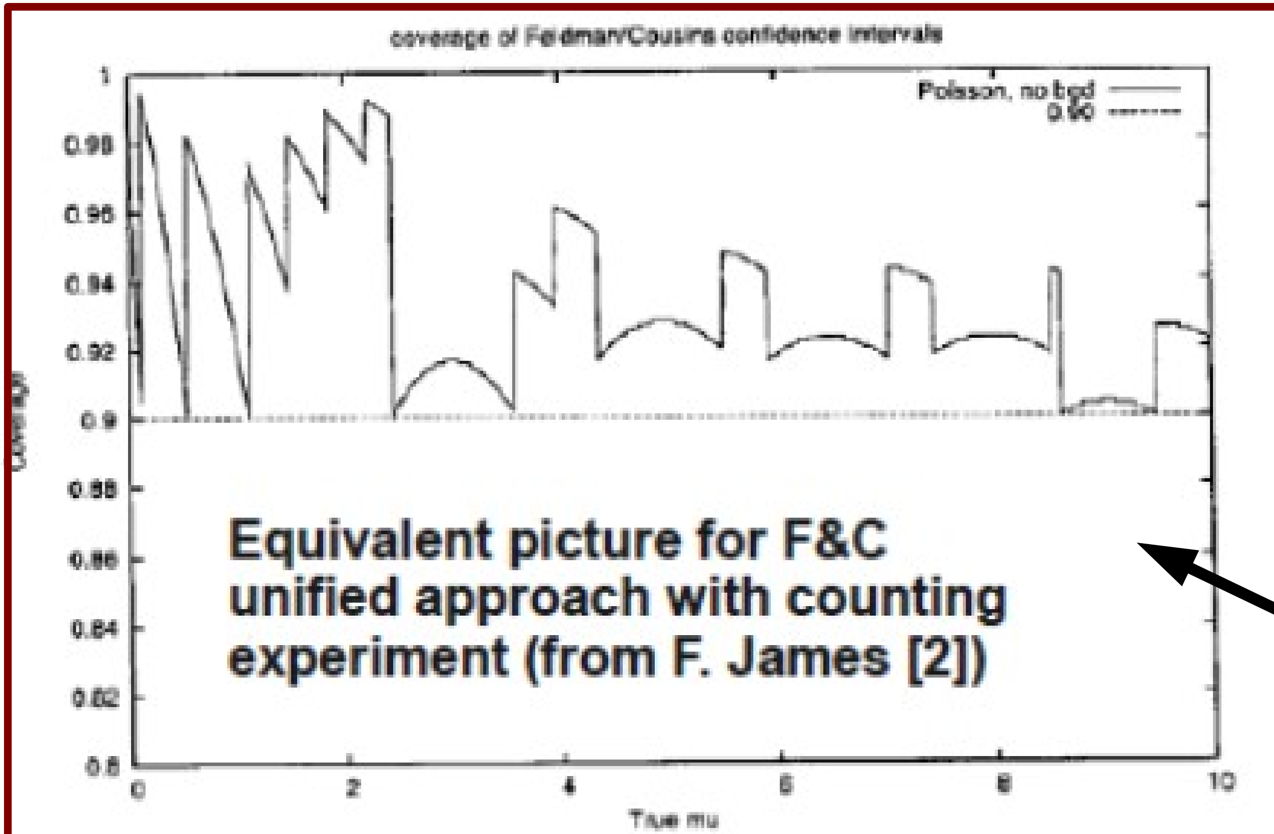
picture changes dramatically when using continuous variable

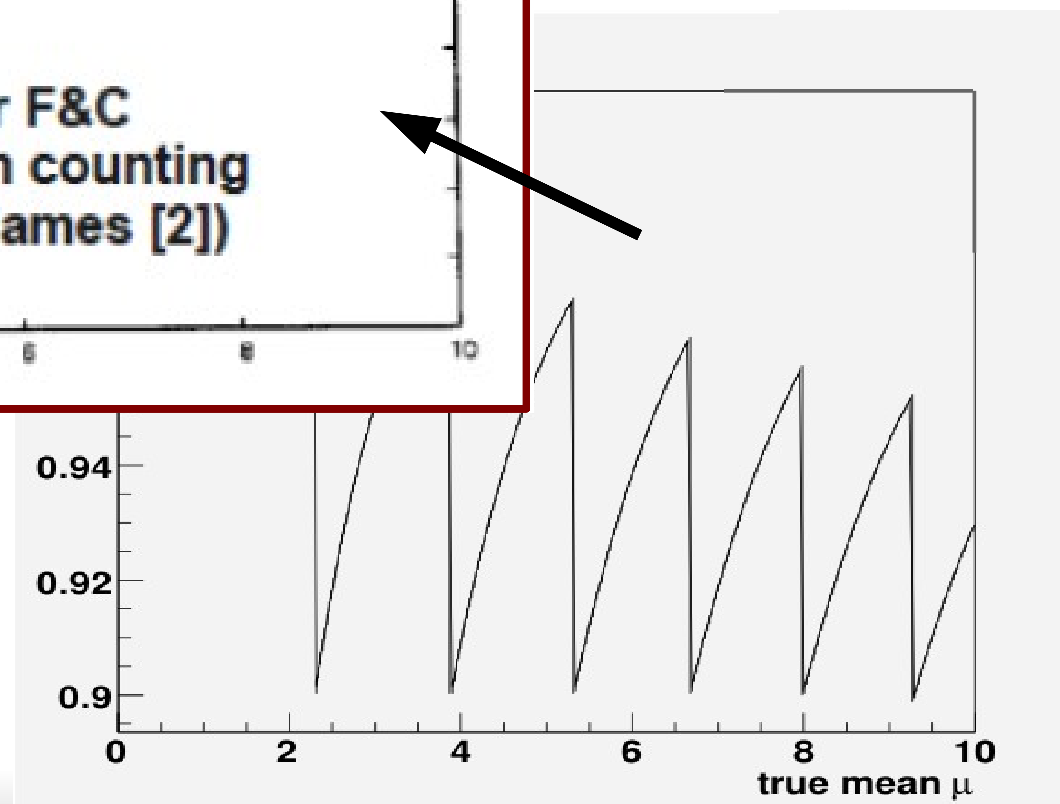
Example: counting experiment

P (

nall expected background
lephant search)

- Outc
- All e
  prod
  → e:
    fo
- In lo
  poss
- (Severe) over-coverage
  - can live with that, but keep in mind
    if competing analysis does a lot better



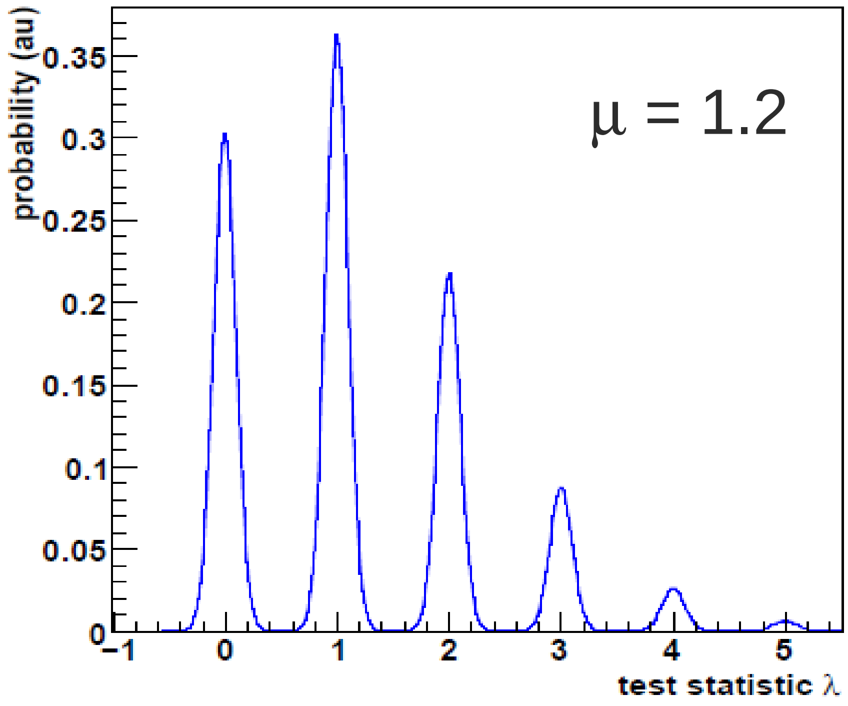


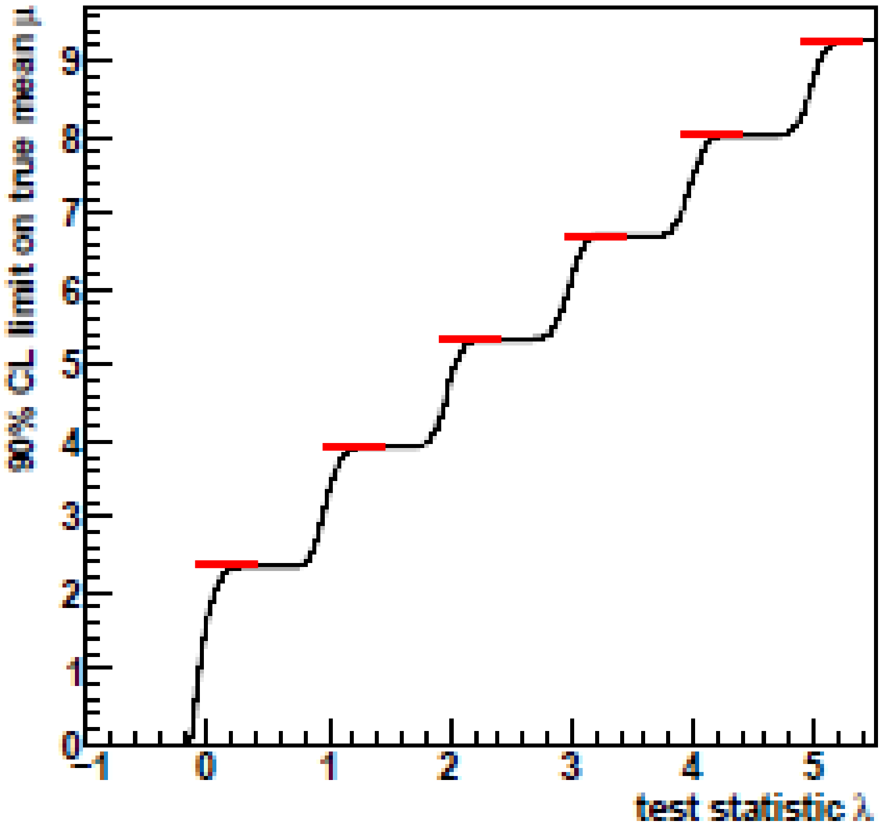
picture changes dramatically when using continuous variable

Illustration: 'Smeared counting experiment'

ANTARES-PHYS-2009-008

$$\lambda = N_{\text{obs}} + R_{\text{andom number}}$$





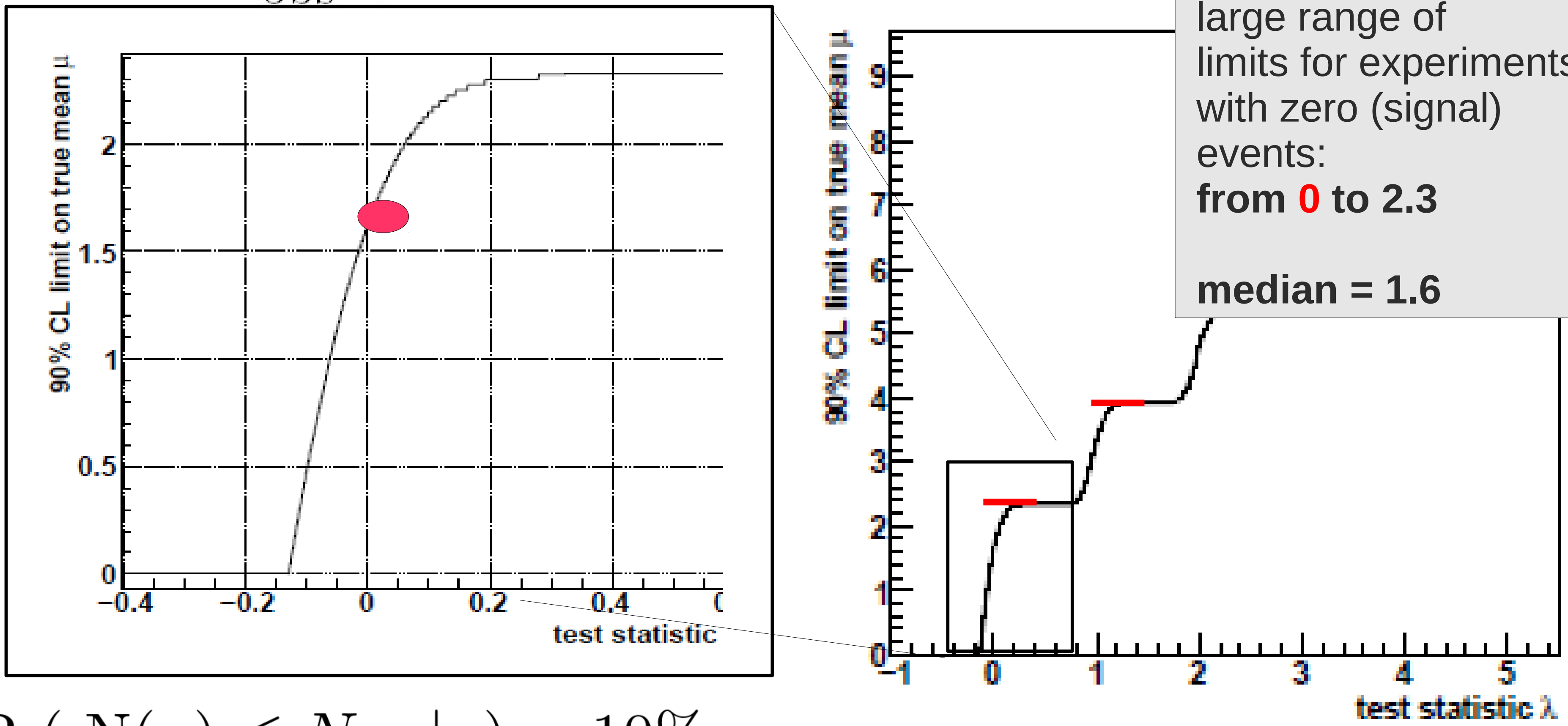
$$\mathrm{P}\ (\ \mathrm{N}(\mu) \leq N_{\text{obs}} | \mu) = 10\%$$

$$\mathrm{P}\ (\ \lambda(\mu) < \lambda_{\text{obs}} | \mu) = 10\%$$

Illustration: 'Smeared counting experiment'

ANTARES-PHYS-2009-008

$$\lambda = N_{\mathrm{obs}} + R$$ andom number



large range of limits for experiments with zero (signal) events:

from 0 to 2.3

median = 1.6

$$\mathrm{P}\ (\ \mathrm{N}(\mu) \leq N_{\mathrm{obs}} | \mu) = 10\%$$

$$\mathrm{P}\ (\ \lambda(\mu) < \lambda_{\mathrm{obs}} | \mu) = 10\%$$

Illustration: 'Smeared counting experiment'

- sensitivity (=expected limit) gets better by *just* using a continuous observable.
  - up to **40%** better, without adding information
- Gain comes from eliminating over-coverage of limits in case of discrete observable
- This can be (partially) why unbinned methods give better (expected) limits than binned
- Coverage is now exactly the stated 90% (for all μ)
- However: “Neyman” limits for a continuous observable, in the small background-regime, have a serious defect: sometimes the excluded value of μ is zero!
  - Fine for hardcore frequentist: it only happens in <10 % of the cases and so the limit still exceeds the true value at 90% CL
  - However, not considered a satisfactory answer in a search

Excluding a flux of zero

from CLs paper

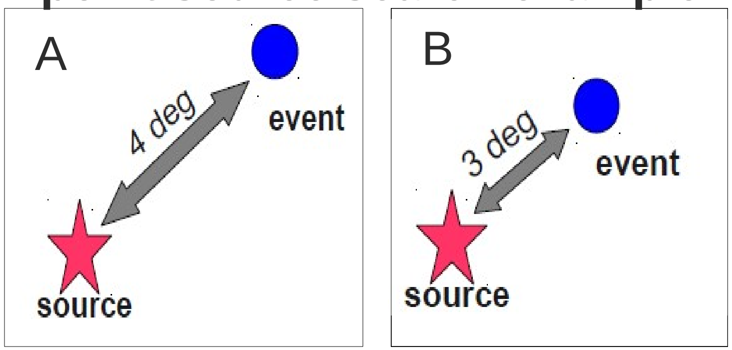
bounded. When an experimental result appears consistent with little or no signal together with a downward fluctuation of the background, the exclusion may be so strong that even zero signal is excluded at confidence levels higher than 95%. Although a perfectly valid result from a statistical point of view, it tends to say more about the probability of observing a similar or stronger exclusion in future experiments with the same expected signal and background than about the non-existence of the signal itself, and it is the latter which is of more interest to the physicist. Presumably a great deal of effort has already gone

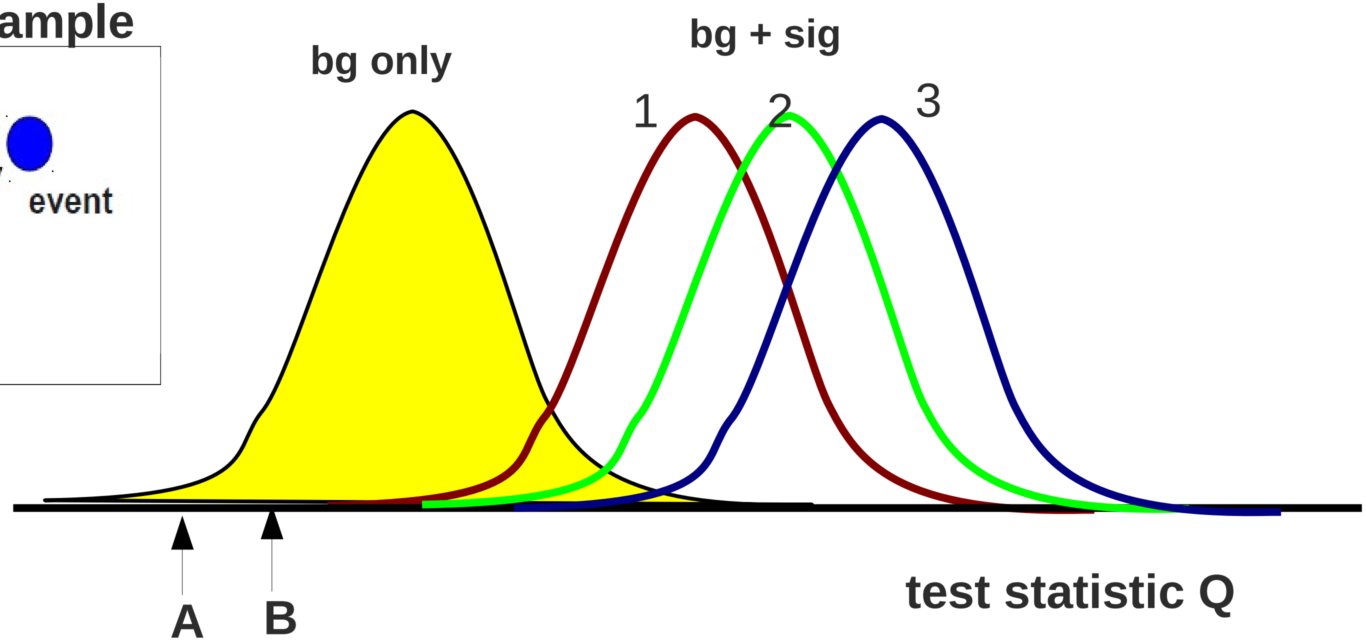
from PDG

probability to obtain a lower $\mathrm{CL_s}$ value) is less than α. This prevents exclusion of a parameter value that could result from a statistical fluctuation in situations where one has no sensitivity, *e.g.*, at very high Higgs masses. The procedure results in a coverage

(what to do with) BG-like experiments

point source search example





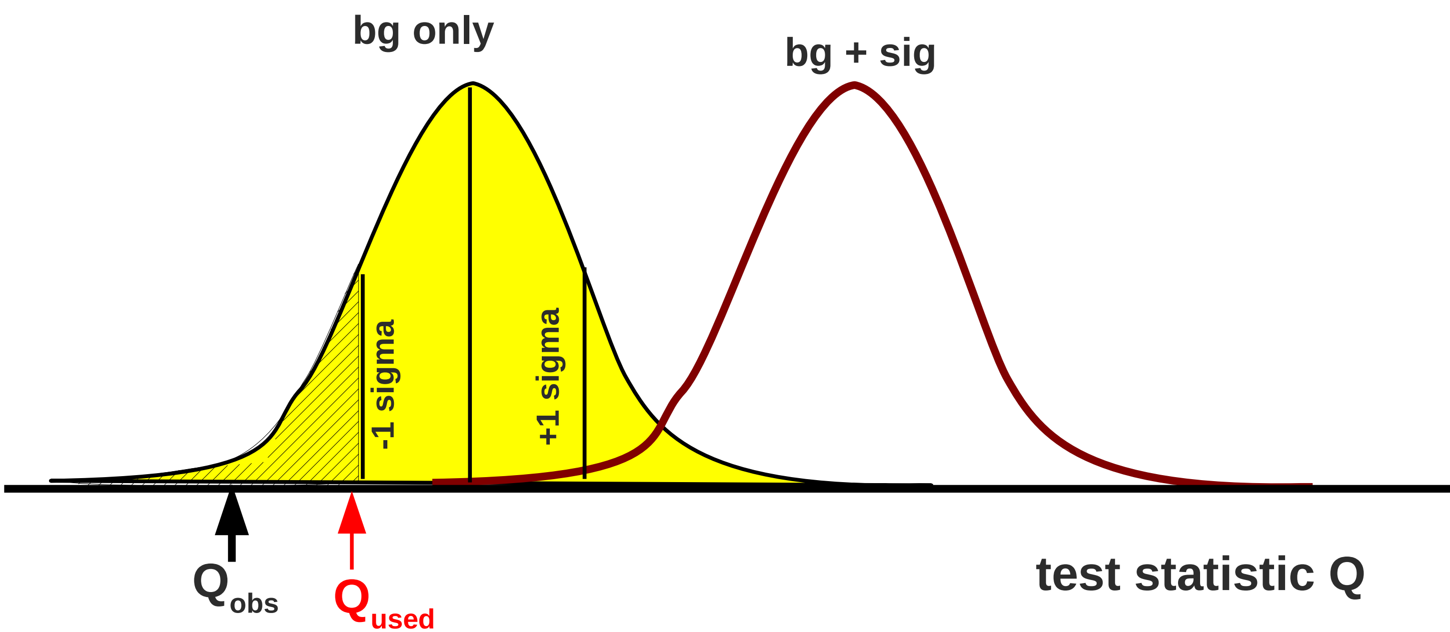
two schools of thought:

- experiment A is still more signal-like that experiment
  → B should have a more stringent limit
  (in that case, one must use a method that at least gives 'reasonable' limits)

- both experiments are ~equally compatible with any signal being present
  and the difference is just due to background fluctuation
  → They should yield the same limit
  - CL_s and power-constrained limits are an implementation of this

Power constrained "Neyman" / CL_{s+b}

- If the observed limit is lower than some threshold, the actual limit is reported for the threshold value.
- The threshold is determined from the bg-only distribution



nb: one can easy do something like this by accident.
… e.g by binning of Q

arXiv:1105.3166

arXiv:1105.3166v1 [physics.data-an] 16 May 2011

Power-Constrained Limits

Glen Cowan[1], Kyle Cranmer[2], Eilam Gross[3], Ofer Vitells[3]

[1] Physics Department, Royal Holloway, University of London, Egham, TW20 0EX, U.K.
[2] Physics Department, New York University, New York, NY 10003, U.S.A.
[3] Weizmann Institute of Science, Rehovot 76100, Israel

Abstract

We propose a method for setting limits that avoids excluding parameter values for which the sensitivity falls below a specified threshold. These "power-constrained" limits (PCL) address the issue that motivated the widely used CL_s procedure [1], but do so in a way that makes more transparent the properties of the statistical test to which each value of the parameter is subjected. A case of particular interest is for upper limits on parameters that are proportional to the cross section of a process whose existence is not yet established. The basic idea of the power constraint can easily be applied, however, to other types of limits.

arXiv:1006.4334

arXiv:1006.4334v1 [astro-ph.IM] 22 Jun 2010

accepted for publication in ApJ

On Computing Upper Limits to Source Intensities

Vinay L. Kashyap[1], David A. van Dyk[2], Alanna Connors[3],
Peter E. Freeman[4], Aneta Siemiginowska[1], Jin Xu[2], and Andreas Zezas[5]

[1] Smithsonian Astrophysical Observatory,
60 Garden Street, Cambridge, MA 02138

vkashyap@cfa.harvard.edu
asiemiginowska@cfa.harvard.edu

[2] Department of Statistics, University of California,
Irvine, CA 92697-1250

dvd@ics.uci.edu
jinx@ics.uci.edu

[3] Eureka Scientific
2452 Delmer Street Suite 100 Oakland, CA 94602-3017

aconnors@eurekabayes.com

[4] Department of Statistics, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

pfreeman@cmu.edu

[5] Physics Department, University of Crete,
P.O. Box 2208, GR-710 03, Heraklion, Crete, Greece

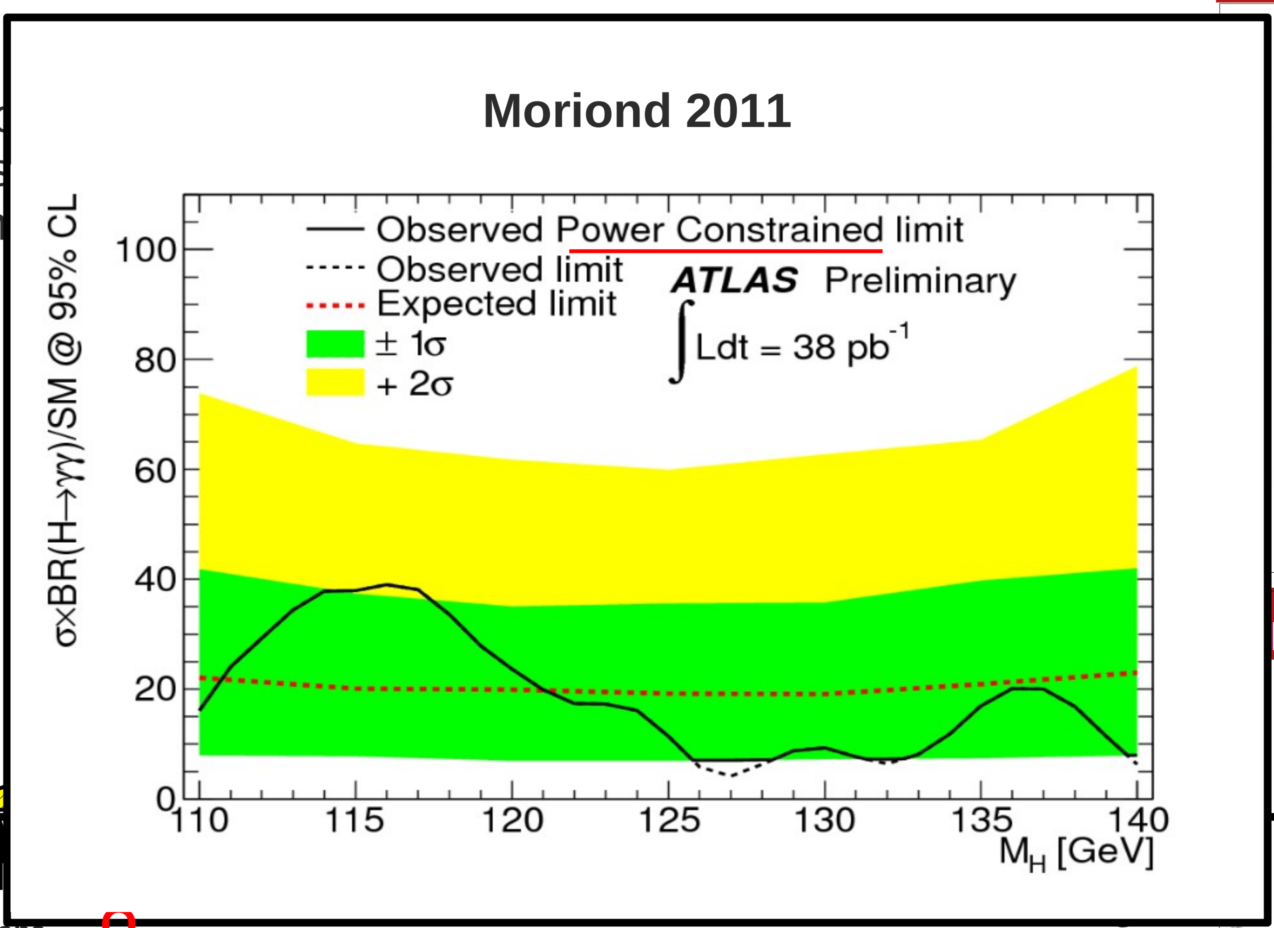
azezas@cfa.harvard.edu

ABSTRACT

A common problem in astrophysics is determining how bright a source could be and still not be detected in an observation. Despite the simplicity with which the problem can be stated, the solution involves complicated statistical issues that require careful analysis. In contrast to the more familiar confidence bound, this concept has never been formally analyzed, leading to a great variety of often ad hoc solutions. Here we formulate and describe the problem in a self-consistent manner. Detection significance is usually defined by the acceptable proportion of false positives (background fluctuations that are claimed as detections, or the Type I error), and we invoke the complementary concept of false negatives (real sources that go undetected, or the Type II error), based on the statistical power of a test, to compute an upper limit to the detectable source intensity. To determine the minimum intensity that a source must have for it to be detected, we first define a detection threshold, and then compute the probabilities of detecting sources of various intensities at the given threshold. The intensity that corresponds to the specified Type II error probability defines that minimum intensity, and is identified as the upper limit. Thus, an upper limit is a characteristic of the detection procedure rather than the strength of any particular source. It should not be confused with confidence intervals or

Power constrained "Neyman" / CL_{s+b}

- If the c… limit is…
- The th…



Q_{obs} Q_{used}


arXiv:1105.3166

Power-Constrained Limits

Glen Cowan[1], Kyle Cranmer[2], Eilam Gross[3], Ofer Vitells[3]

[1] Physics Department, Royal Holloway, University of London, Egham, TW20 0EX, U.K.
[2] Physics Department, New York University, New York, NY 10003, U.S.A.
[3] Weizmann Institute of Science, Rehovot 76100, Israel

Abstract

We propose a method for setting limits that avoids excluding parameter values for which the sensitivity falls below a specified threshold. These "power-constrained" limits (PCL) address the issue that motivated the widely used CL_s procedure [1], but do so in a way that makes more transparent the properties of the statistical test to which each value of the parameter is subjected. A case of particular interest is for upper limits on parameters that are proportional to the cross section of a process whose existence is not yet established. The basic idea of the power constraint can easily be applied, however, to other types of limits.


arXiv:1006.4334

accepted for publication in ApJ

On Computing Upper Limits to Source Intensities

Vinay L. Kashyap[1], David A. van Dyk[2], Alanna Connors[3], Peter E. Freeman[4], Aneta Siemiginowska[1], Jin Xu[2], and Andreas Zezas[5]

[1] Smithsonian Astrophysical Observatory, 60 Garden Street, Cambridge, MA 02138
[2] Department of Statistics, University of California, Irvine, CA 92697-1250
[3] Eureka Scientific, 2452 Delmer Street Suite 100 Oakland, CA 94602-3017
[4] Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213
[5] Physics Department, University of Crete, P.O. Box 2208, GR-710 03, Heraklion, Crete, Greece

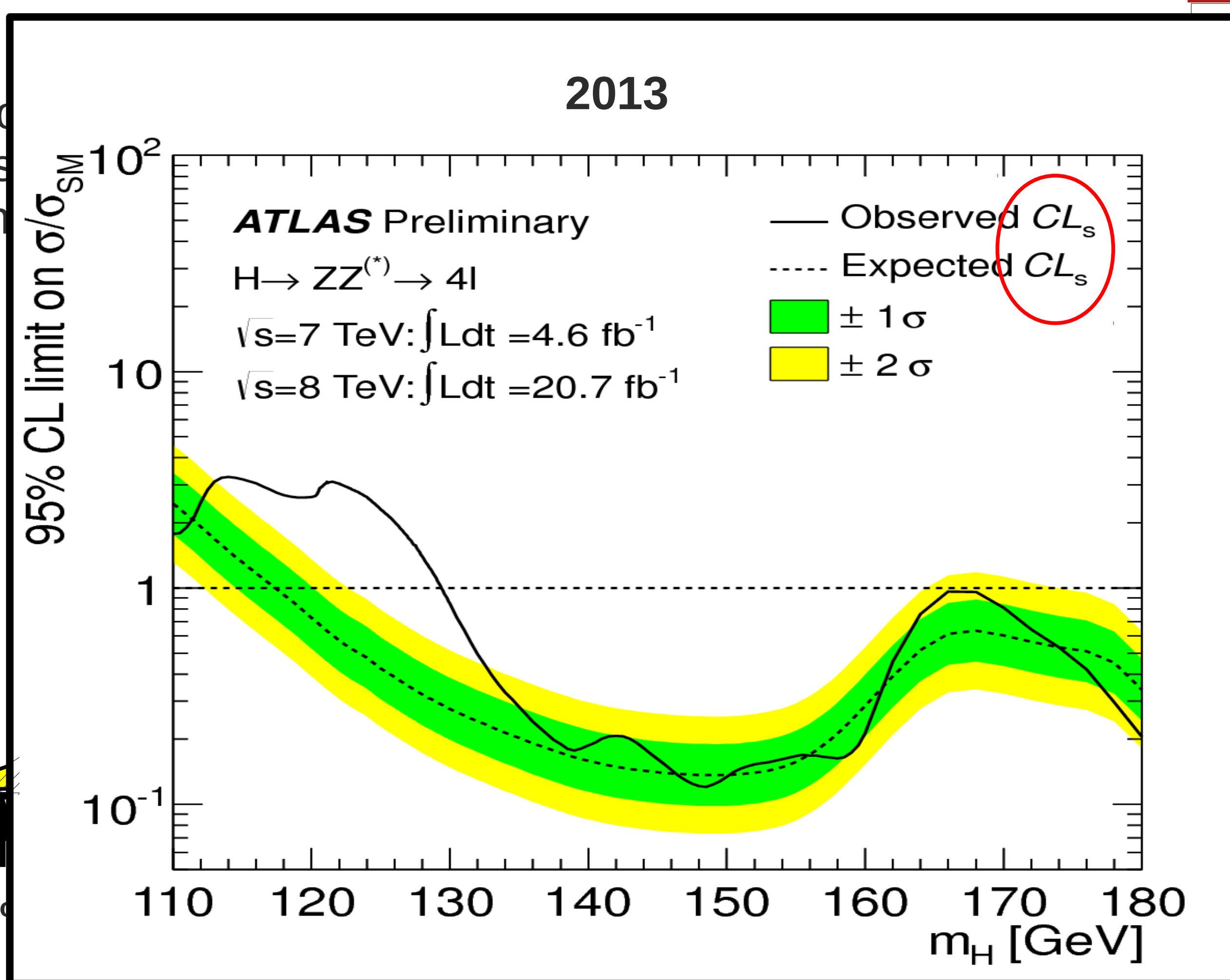
ABSTRACT

A common problem in astrophysics is determining how bright a source could be and still not be detected in an observation. Despite the simplicity with which the problem can be stated, the solution involves complicated statistical issues that require careful analysis. In contrast to the more familiar confidence bound, this concept has never been formally analyzed, leading to a great variety of often ad hoc solutions. Here we formulate and describe the problem in a self-consistent manner. Detection significance is usually defined by the acceptable proportion of false positives (background fluctuations that are claimed as detections, or the Type I error), and we invoke the complementary concept of false negatives (real sources that go undetected, or the Type II error), based on the statistical power of a test, to compute an upper limit to the detectable source intensity. To determine the minimum intensity that a source must have for it to be detected, we first define a detection threshold, and then compute the probabilities of detecting sources of various intensities at the given threshold. The intensity that corresponds to the specified Type II error probability defines that minimum intensity, and is identified as the upper limit. Thus, an upper limit is a characteristic of the detection procedure rather than the strength of any particular source. It should not be confused with confidence intervals or

nb: one can easy do something like this by accident.
… e.g by binning

Power constrained "Neyman" / CL_{s+b}

- If the c
  limit is
- The th



arXiv:1105.3166

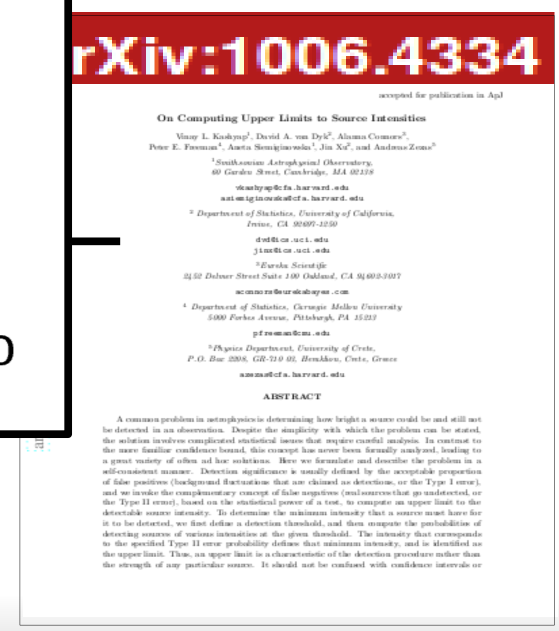
Power-Constrained Limits

Glen Cowan[1], Kyle Cranmer[2], Eilam Gross[3], Ofer Vitells[3]

[1] Physics Department, Royal Holloway, University of London, Egham, TW20 0EX, U.K.
[2] Physics Department, New York University, New York, NY 10003, U.S.A.
[3] Weizmann Institute of Science, Rehovot 76100, Israel

Abstract

We propose a method for setting limits that avoids excluding parameter values for which the sensitivity falls below a specified threshold. These "power-constrained" limits (PCL) address the issue that motivated the widely used CL_s procedure [1], but do so in a way that makes more transparent the properties of the statistical test to which each value of the parameter is subjected. A case of particular interest is for upper limits on parameters that are proportional to the cross section of a process whose existence is not yet established. The basic idea of the power constraint can easily be applied, however, to other types of limits.



rXiv:1006.4334

accepted for publication in ApJ

On Computing Upper Limits to Source Intensities

Vinay L. Kashyap[1], David A. van Dyk[2], Alanna Connors[3], Peter E. Freeman[4], Aneta Siemiginowska[1], Jin Xu[2], and Andreas Zezas[5]

[1] *Smithsonian Astrophysical Observatory, 60 Garden Street, Cambridge, MA 02138*

vkashyap@cfa.harvard.edu
asiemiginowska@cfa.harvard.edu

[2] *Department of Statistics, University of California, Irvine, CA 92697-1250*

dvd@ics.uci.edu
jinx@ics.uci.edu

[3] *Eureka Scientific 2452 Delmer Street Suite 100 Oakland, CA 94602-3017*

aconnors@eurekabayes.com

[4] *Department of Statistics, Carnegie Mellon University 5000 Forbes Avenue, Pittsburgh, PA 15213*

pfreeman@cmu.edu

[5] *Physics Department, University of Crete, P.O. Box 2208, GR-710 03, Heraklion, Crete, Greece*

azezas@cfa.harvard.edu

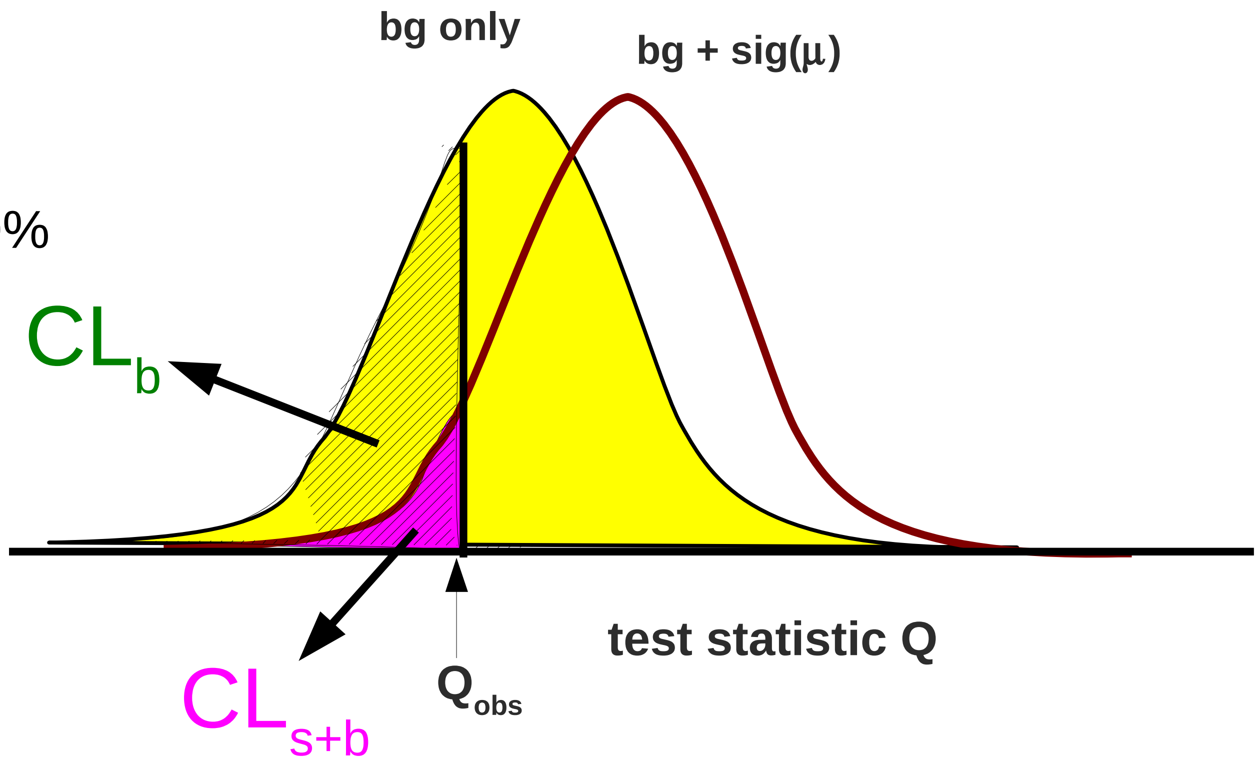
ABSTRACT

A common problem in astrophysics is determining how bright a source could be and still not be detected in an observation. Despite the simplicity with which the problem can be stated, the solution involves complicated statistical issues that require careful analysis. In contrast to the more familiar confidence bound, this concept has never been formally analyzed, leading to a great variety of often ad hoc solutions. Here we formulate and describe the problem in a self-consistent manner. Detection significance is usually defined by the acceptable proportion of false positives (background fluctuations that are claimed as detections, or the Type I error), and we invoke the complementary concept of false negatives (real sources that go undetected, or the Type II error), based on the statistical power of a test, to compute an upper limit to the detectable source intensity. To determine the minimum intensity that a source must have for it to be detected, we first define a detection threshold, and then compute the probabilities of detecting sources of various intensities at the given threshold. The intensity that corresponds to the specified Type II error probability defines that minimum intensity, and is identified as the upper limit. Thus, an upper limit is a characteristic of the detection procedure rather than the strength of any particular source. It should not be confused with confidence intervals or

nb: one can easy do *something like this* by accident.
… e.g by binning

CLs Method (a.k.a. Modified Frequentist)

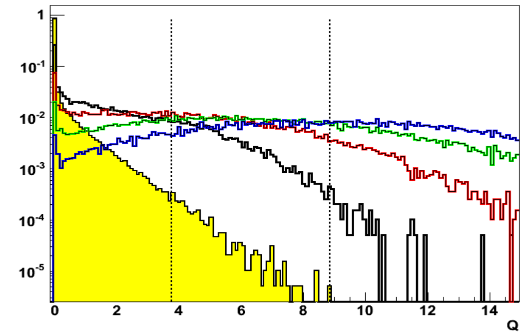
define:
$CL_s = CL_{s+b} / CL_b$
and require $CL_s(\mu^{limit}) = 10\%$
for a 90% 'CL' limit

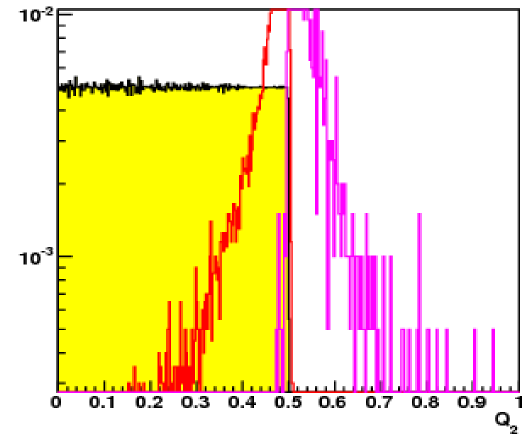


- Only exclude values for which there is some ability to observe them
- If $\mu = 0$, CLs = 1 → never exlude this
- in fact, for most bg-like outcomes give μ^{limit}=2.3 (same as counting experiment)
- Over-coverage : limits are 'worse'
- nevertheless quite widely used: LEP, Tevatron, **LHC**...
- easy to implement
- unpopular with statisticians : CLs is *not* a confidence level

CERN-OPEN-2000-205

Feldman-Cousins



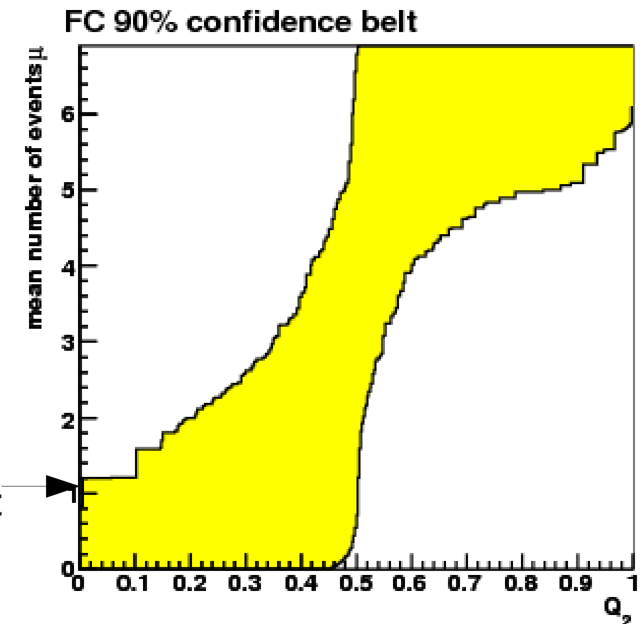


- Prevents excluding zero
  (by spending coverage on lower limit)
- produces double sided interval (we don't really care)
- Can be difficult to implement:
  - likelihood ordering requires many pseudo-experiments to work well..
  - a transformation of the test statistic can help, but still

for Antares point sources, we chose it because:

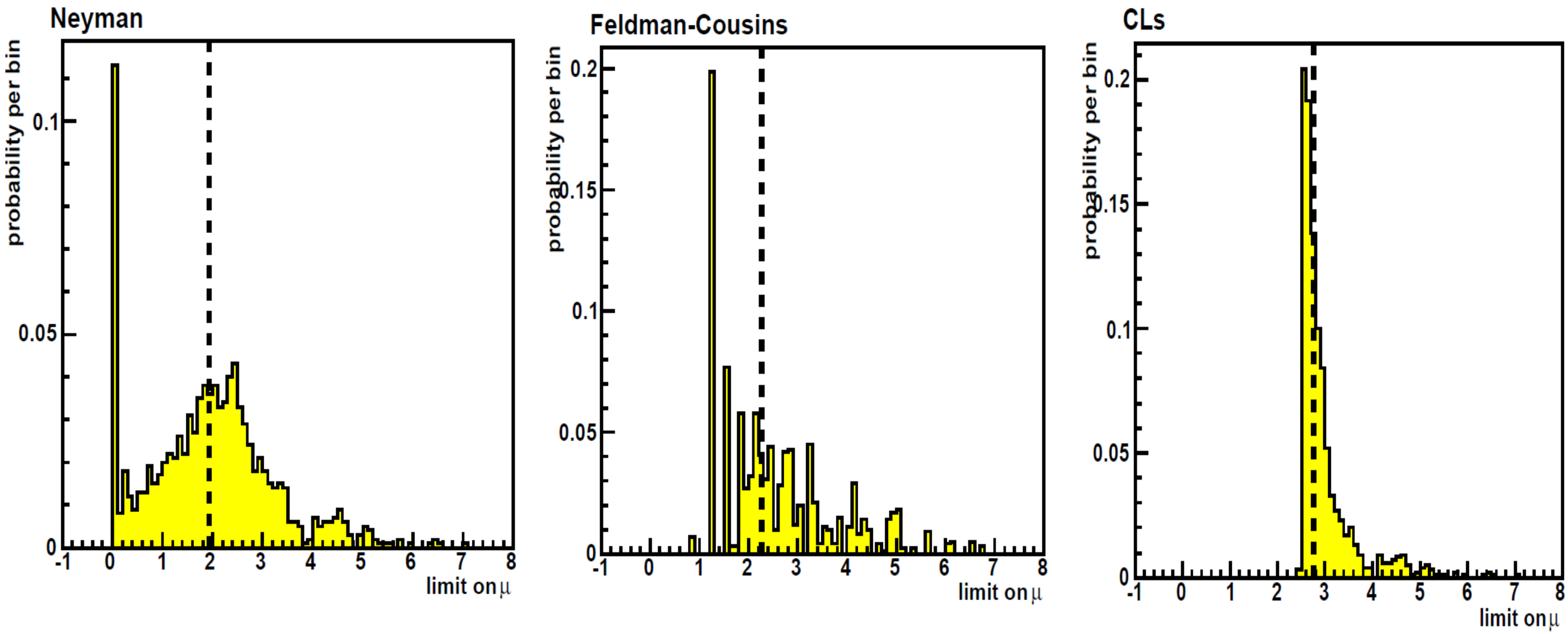
- IceCube was using it
- allows use of full range of continuous variable without the need for additional measures (like power-constraining or something that depends on the binning)
- better coverage (lower limits) than CL_s

- seems FC Is not really catching on at LHC, and many people in our community prefer something simpler.



lowest possible limit around 1 event (not unreasonable?)

comparing all three

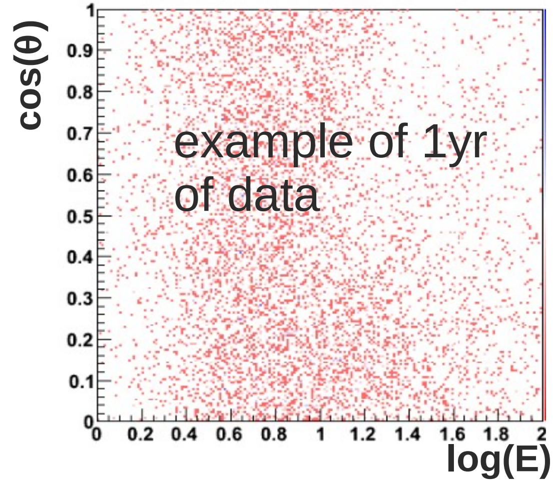


- plots from 1st antares point source analysis
- Neyman has best sensitivity (dashed line), but excludes a flux of zero

Likelihood ratio with nuisance parameters

What if the hypotheses under test have unknown (nuisance) parameters ?
e.g. for Hierarchy determination:

$$Q = \frac{P(\mathrm{data}|NH, \Delta m^2_{\mathrm{large}}, \Delta m^2_{\mathrm{small}}, \theta_{12}, \theta_{13}, \theta_{23})}{P(\mathrm{data}|IH, \Delta m^2_{\mathrm{large}}, \Delta m^2_{\mathrm{small}}, \theta_{12}, \theta_{13}, \theta_{23})}$$

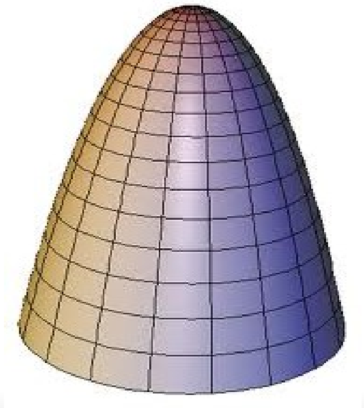


common recipe: plug in maximum likelihood values for the nuisance parameters → i.e. first fit them to the data.

What if we want to include external information?:

$$\log(\mathcal{L}) = \log P(\mathrm{data}^{\mathrm{us}}|H, \vec{\theta}) + \log P(\mathrm{data}^{\mathrm{others}}|H, \vec{\theta})$$

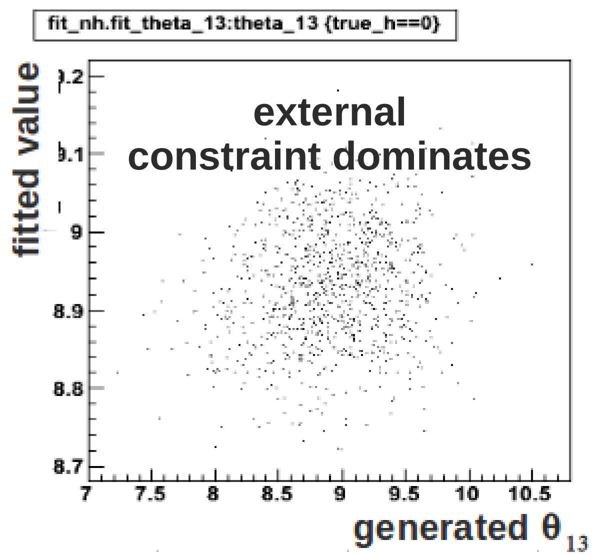
- adding constraints is equivalent to combining datasets
- ideally add full likelihood-grid of constraining measurement(s), alternatively, assume log(P) is paraboloid according to published central values and uncertainties

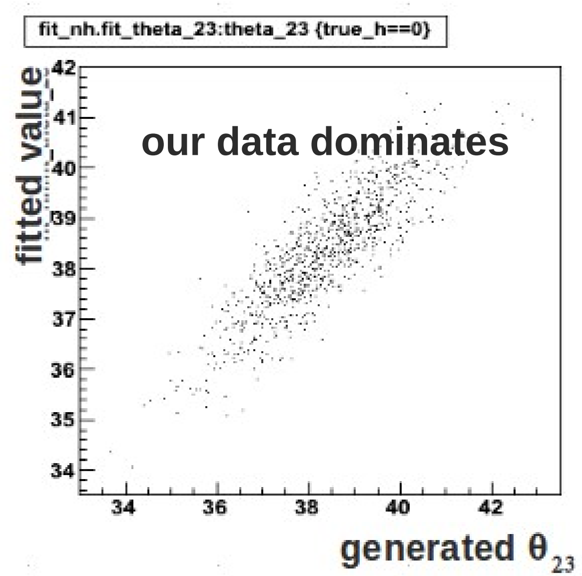


Likelihood ratio with nuisance parameters

TABLE I: Results of the global 3ν oscillation analysis, in terms of best-fit values and allowed 1, 2 and 3σ ranges for the 3ν mass-mixing parameters. We remind that Δm^2 is defined herein as $m_3^2 - (m_1^2 + m_2^2)/2$, with $+\Delta m^2$ for NH and $-\Delta m^2$ for IH.

Parameter	Best fit	1σ range	2σ range	3σ range
$\delta m^2/10^{-5}$ eV^2 (NH or IH)	7.54	7.32 – 7.80	7.15 – 8.00	6.99 – 8.18
$\sin^2\theta_{12}/10^{-1}$ (NH or IH)	3.07	2.91 – 3.25	2.75 – 3.42	2.59 – 3.59
$\Delta m^2/10^{-3}$ eV^2 (NH)	2.43	2.33 – 2.49	2.27 – 2.55	2.19 – 2.62
$\Delta m^2/10^{-3}$ eV^2 (IH)	2.42	2.31 – 2.49	2.26 – 2.53	2.17 – 2.61
$\sin^2\theta_{13}/10^{-2}$ (NH)	2.41	2.16 – 2.66	1.93 – 2.90	1.69 – 3.13
$\sin^2\theta_{13}/10^{-2}$ (IH)	2.44	2.19 – 2.67	1.94 – 2.91	1.71 – 3.15
$\sin^2\theta_{23}/10^{-1}$ (NH)	3.86	3.65 – 4.10	3.48 – 4.48	3.31 – 6.37
$\sin^2\theta_{23}/10^{-1}$ (IH)	3.92	3.70 – 4.31	3.53 – 4.84 $\oplus$ 5.43 – 6.41	3.35 – 6.63
δ/π (NH)	1.08	0.77 – 1.36	—	—
δ/π (IH)	1.09	0.83 – 1.47	—	—



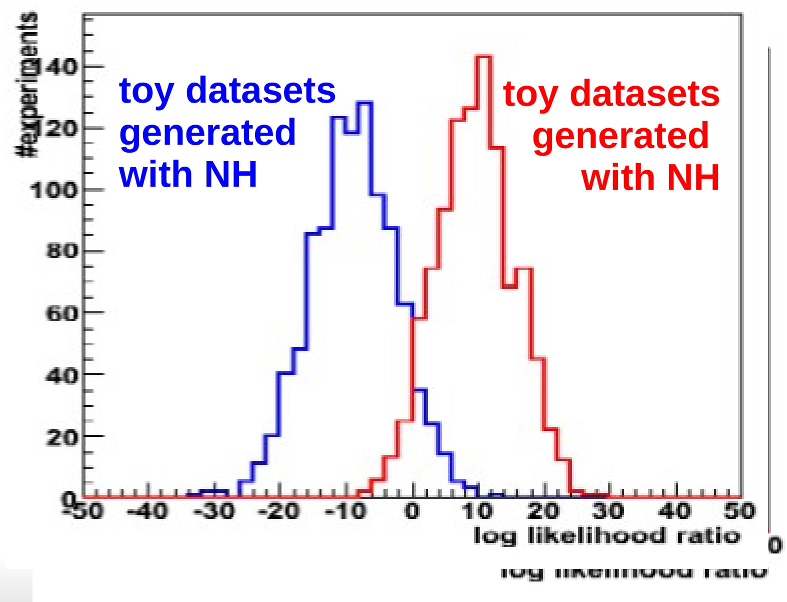


pseudo experiments are generated with parameters varied according to current uncertainties. (1)

in each PE, the nuisance parameters are fit to the data, constraint by current uncertainties (2)

(2) is done for the two hypotheses : NH and IH. Finally compute

$$Q = \frac{P(\mathrm{data}|NH, \vec{\theta}^{\mathrm{fit}}_{NH})}{P(\mathrm{data}|IH, \vec{\theta}^{\mathrm{fit}}_{IH})}$$



Conclusions

- Every search based on some observable, who's distribution can be computed e.g. by pseudo-experiments.

- Making discoveries is easy

- Setting limits is hard
  - Be careful comparing limits based on discrete and continuus variables
    - improvement seen may have nothing to do with s/b separation power of the analysis
  - Neyman / CLs+b limits
    - Over-cover in counting experiment (FC improves that a bit)
    - Severe problems for continuous variables (exclude zero)
  - Several alternatives : power constrain, CLs, FC, Bayesian
    - offer different trade-off between desired properties and 'lowness' of the limits

- Nuisance parameters (a.k.a. degeneracies)
  - fit to the data
  - external constraints can help and are easy to implement