Limit setting

(in point source searches)

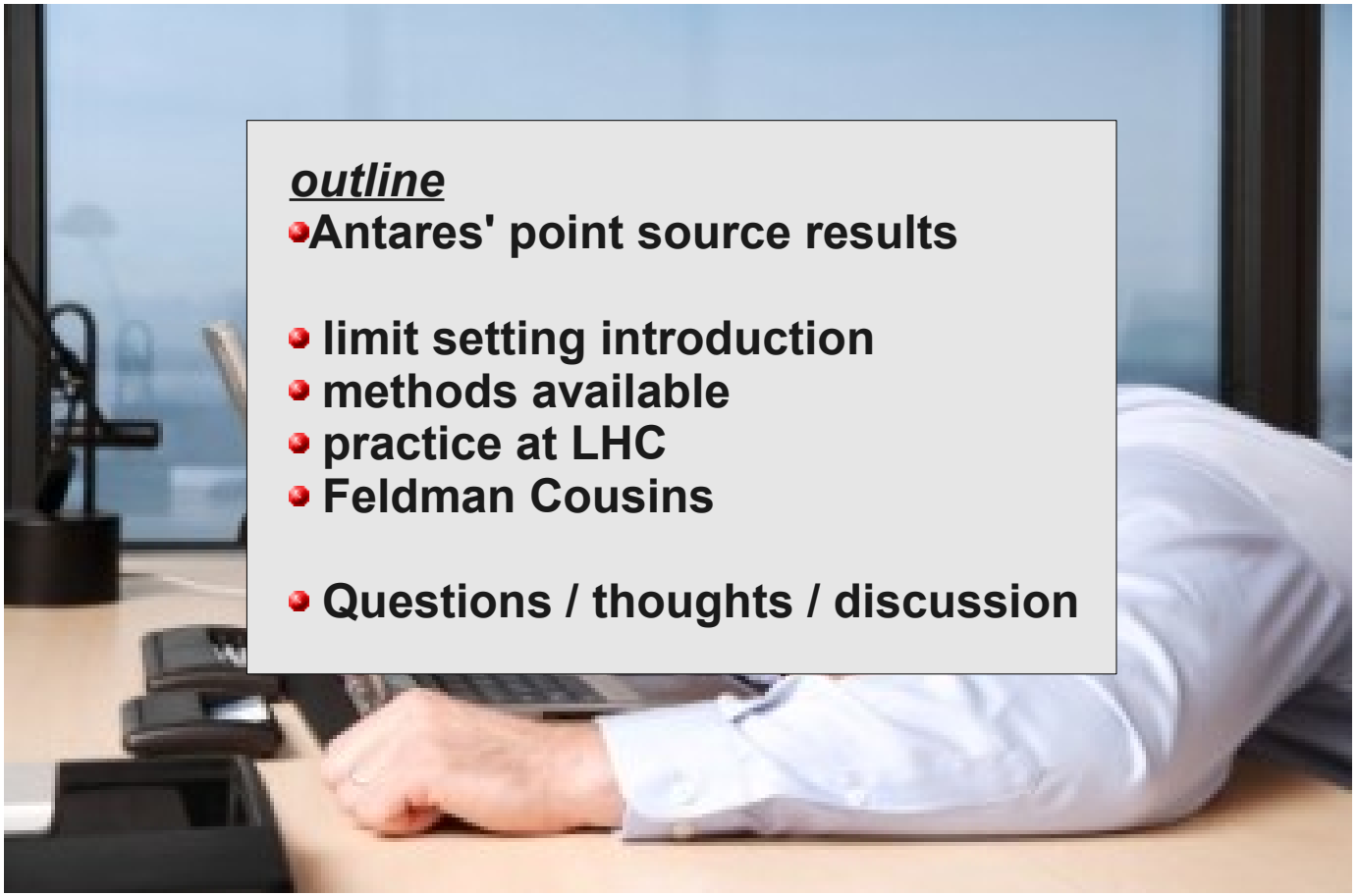
Aart Heijboer,
Nikhef

since you're here early Sunday morning, I'll assume you're interested.....

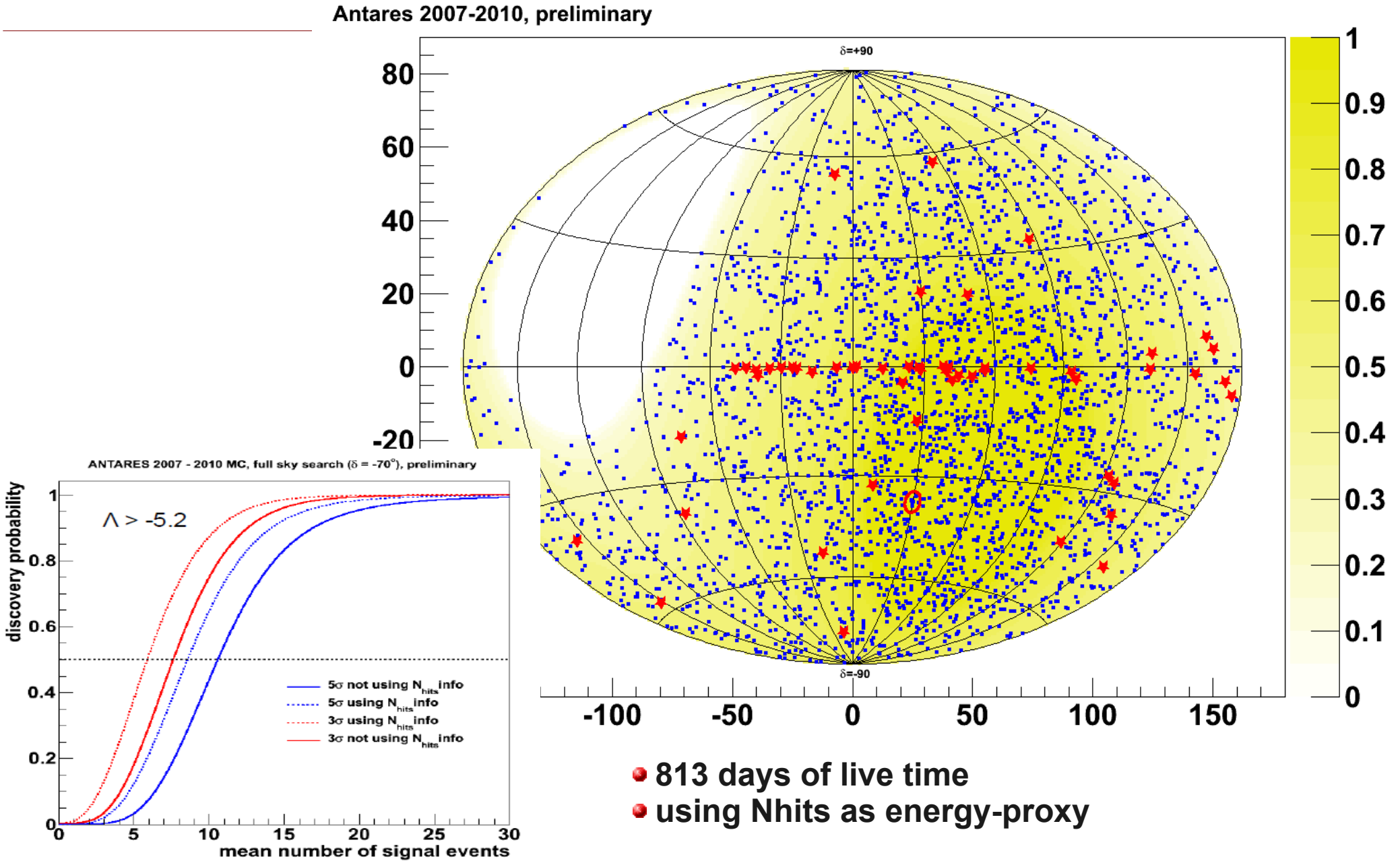
Limit setting

(in point source searches)

Aart Heijboer,
Nikhef



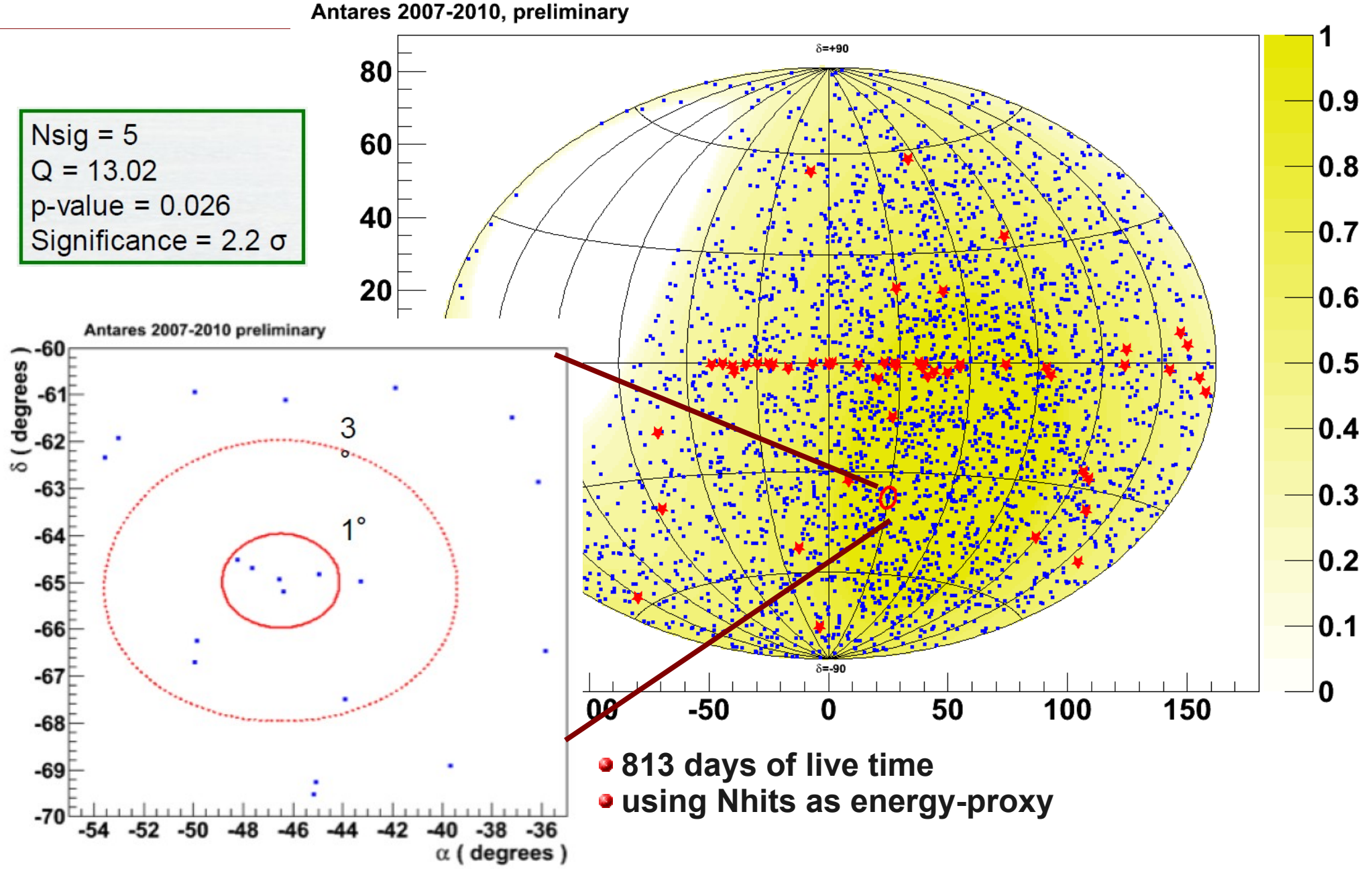
Antares' latest point source results



- 813 days of live time
- using Nhits as energy-proxy

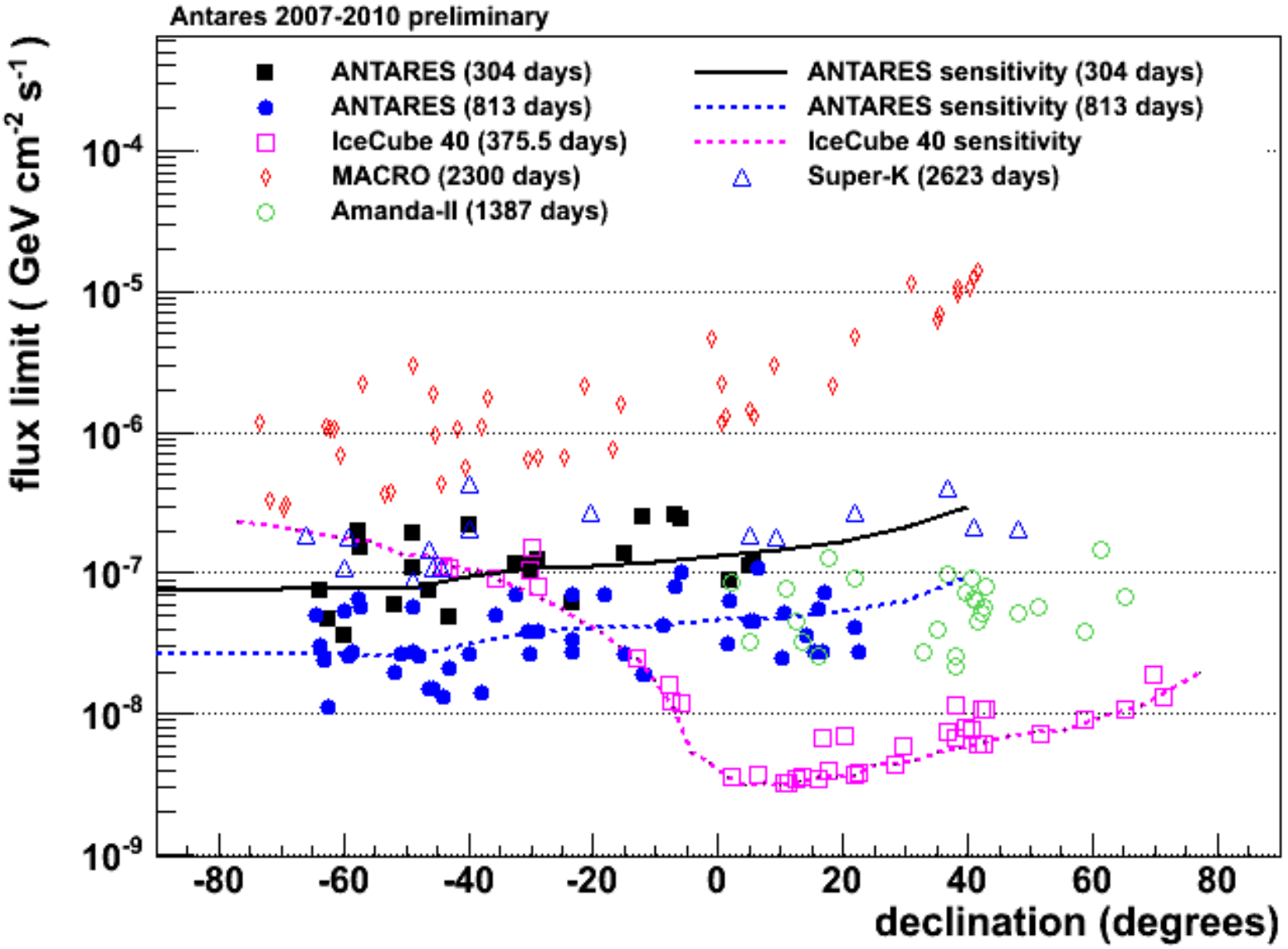
Antares' latest point source results

Nsig = 5
Q = 13.02
p-value = 0.026
Significance = 2.2 σ



- 813 days of live time
- using Nhits as energy-proxy

Limits (F&C)


Antares 2007-2010 preliminary
ANTARES (304 days)
ANTARES (813 days)
IceCube 40 (375.5 days)
MACRO (2300 days)
Amanda-II (1387 days)
ANTARES sensitivity (304 days)
ANTARES sensitivity (813 days)
IceCube 40 sensitivity
Super-K (2623 days)
flux limit (GeV cm^{-2} s^{-1})
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
-80
-60
-40
-20
0
20
40
60
80
declination (degrees)

Limit setting : overview of methods and issues

Introduction

All searches use some likelihood ratio test statistic. We call it Q:

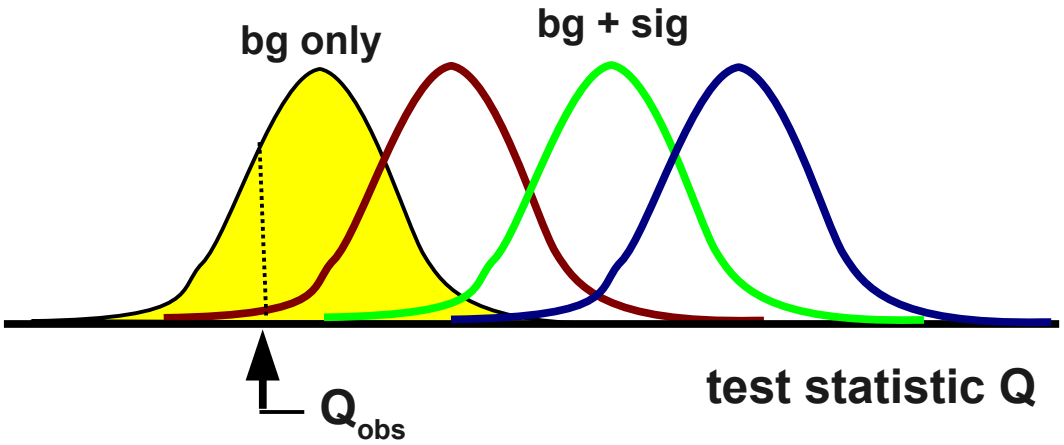
$$\log \mathcal{L}_{\mathrm{s+b}} = \sum \log[\mu_{\mathrm{sig}} \times \mathcal{F}(\beta_i(\delta_s, \alpha_s)) \times \mathcal{N}(N_{hits}^{i,sig}) + \mathcal{B}_i \times \mathcal{N}(N_{hits}^{i,bkg})] + \mu_{\mathrm{tot}}$$

$$Q = \log \mathcal{L}_{\mathrm{s+b}}^{\max} - \log \mathcal{L}_{\mathrm{b}}$$

- making discoveries
  - easy!
  - p-values easy to compute
  - ~no systematics

- setting limits
  - surprisingly hard:
  - choices involved that matter for the numbers
    different limit setting method can change result by 40%
  - possibility of nonsense-results
  - statisticians do not agree

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

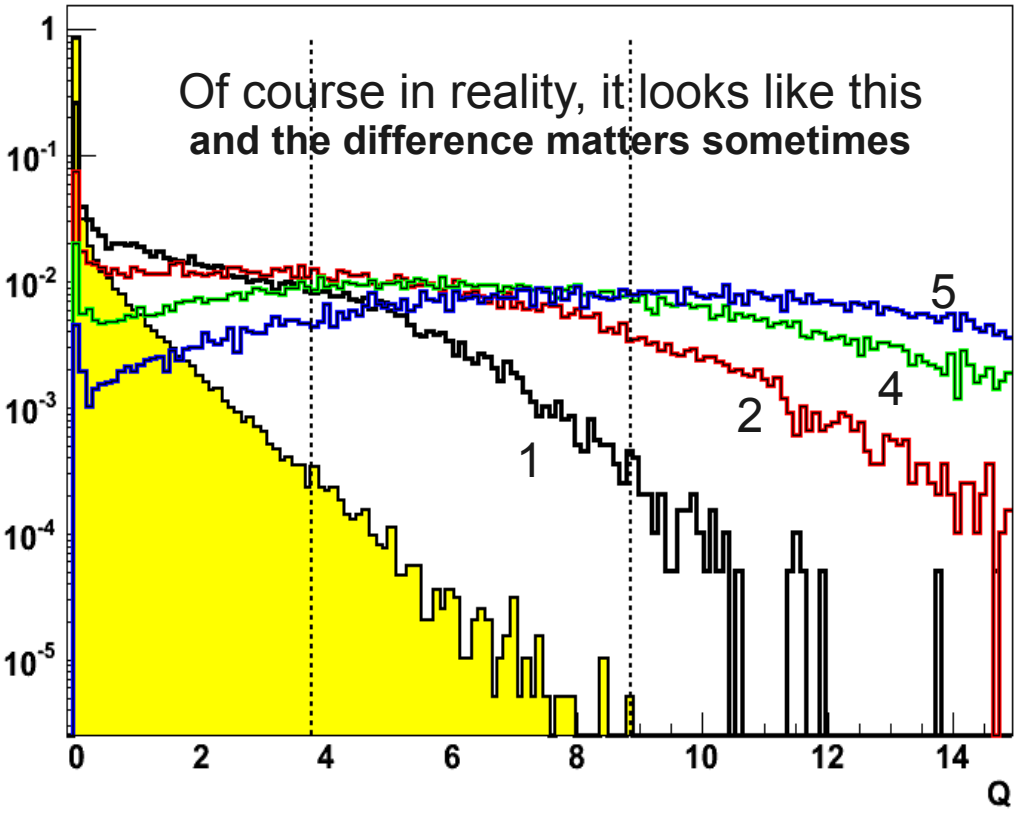


Introduction

$$\log \mathcal{L}_{\mathrm{s+b}} = \sum \log[\mu_{\mathrm{sig}} \times \mathcal{F}(\beta_i(\delta_s, \alpha_s)) \times \mathcal{N}(N_{hits}^{i,sig}) + \mathcal{B}_i \times \mathcal{N}(N_{hits}^{i,bkg})] + \mu_{\mathrm{tot}}$$

$$Q = \log \mathcal{L}_{\mathrm{s+b}}^{\max} - \log \mathcal{L}_{\mathrm{b}}$$

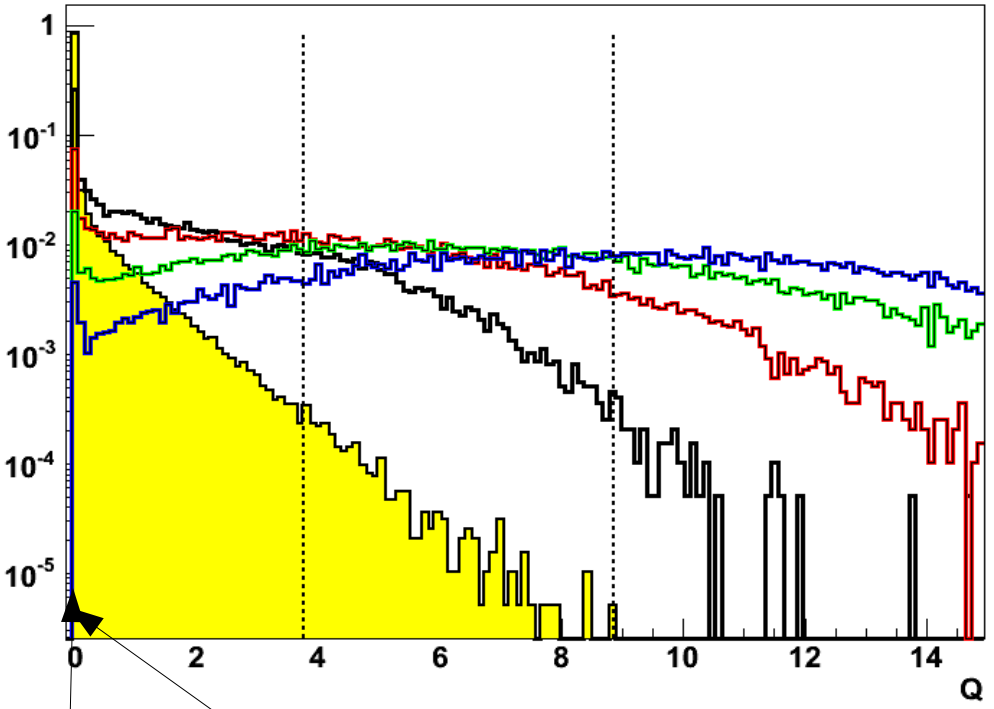
- making discoveries
  - easy!
  - p-values easy to compute
  - no systematics
- setting limits
  - surprisingly hard:
  - choices involved that matter for the result
  - possibility of nonsense-results
  - statisticians do not agree

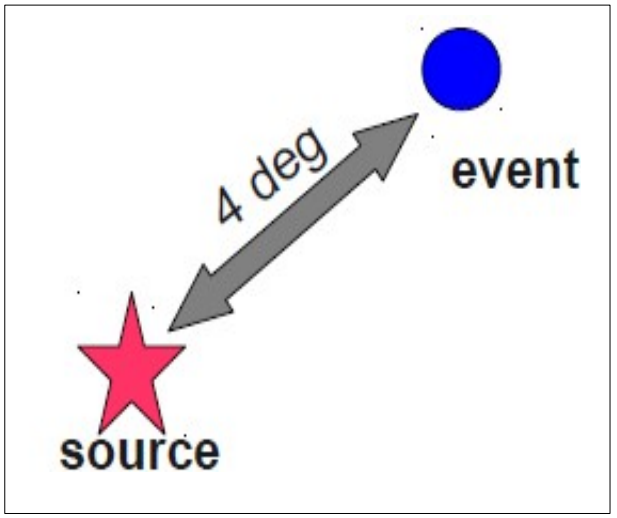


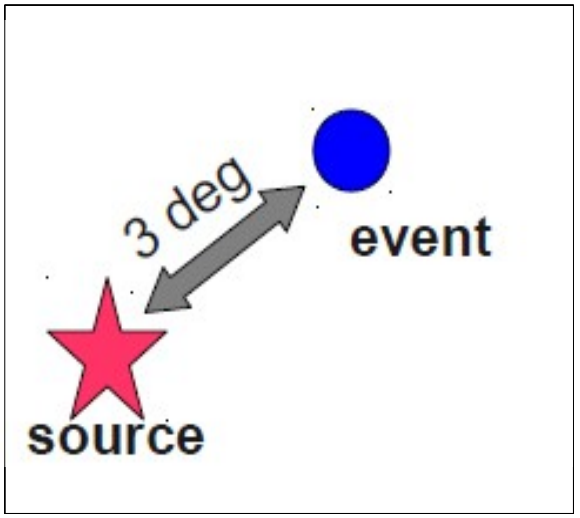
Q distributions from running analysis on pseudo-experiments. PE generation can include all the systematics.

BG-like experiments

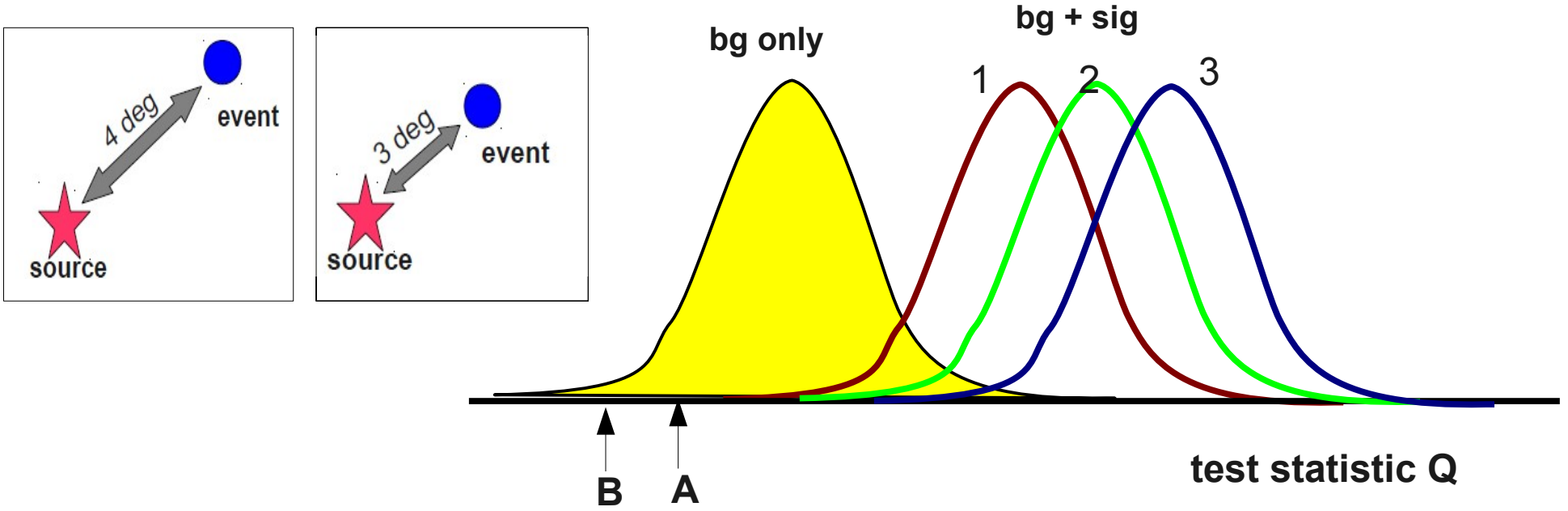
How to treat this peak?







BG-like experiments

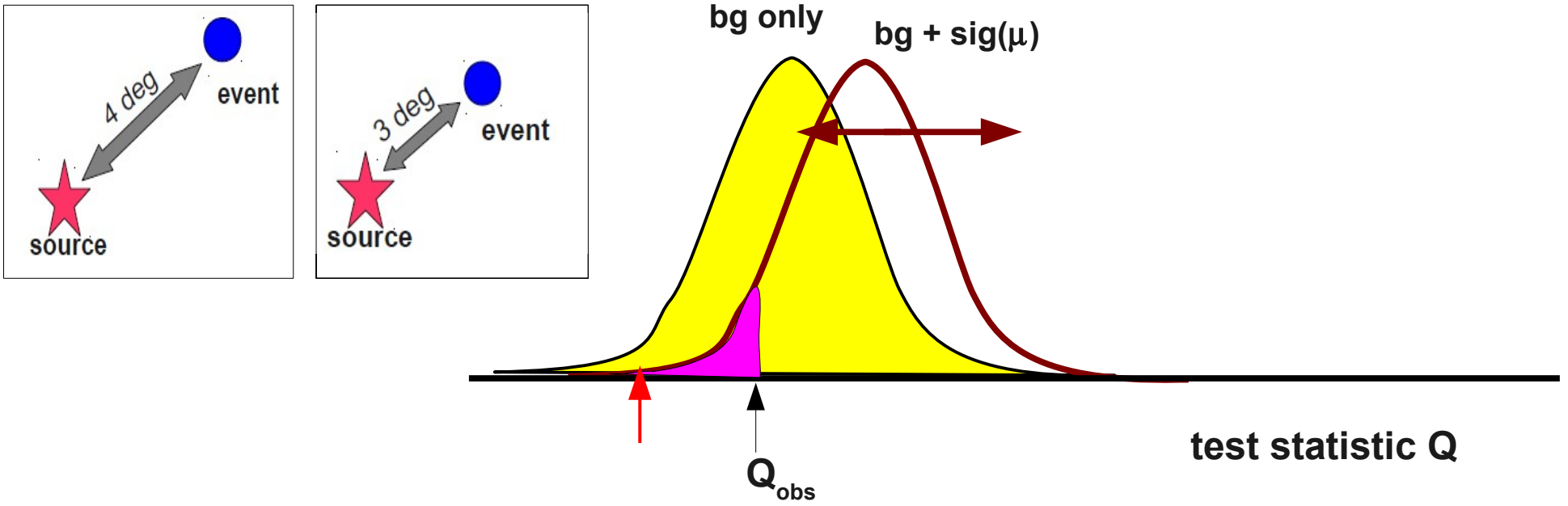


two schools of thought:

- experiment A is more signal-like that experiment
  → B should have a more stringent limit
- 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

'Neyman' limits

(a.k.a CL_{s+b} limits)



'neyman limits' or CL_{s+b} : find the signal strength m so that

$P(Q<Q_{obs} \mid \mu) = 10\ \%$

- produces very different limits for different background fluctuations typically in the region <~1 signal event.
- If Q_{obs} is very bg-like (in the 10% tail) → exclude even $\mu=0$

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

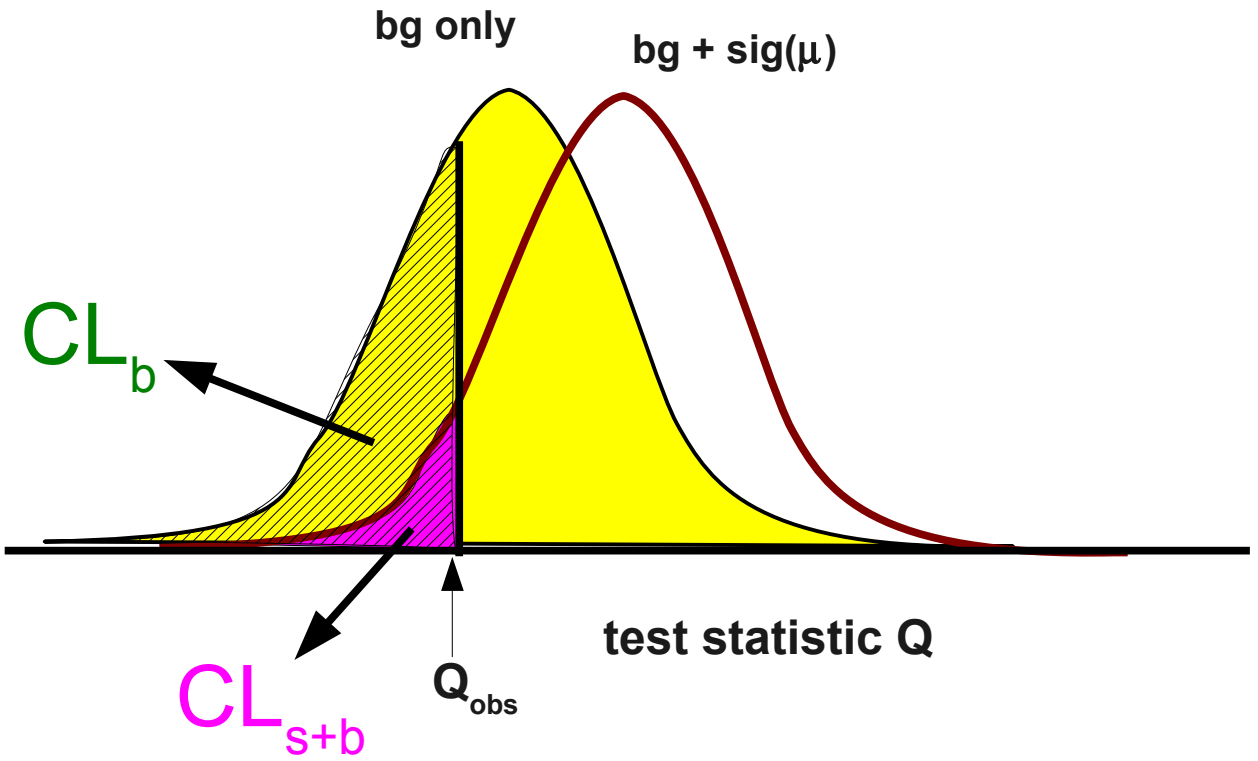
from PDG

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.

happens in 10% of the cases. i.e. ~sure to happen in a candidate sourc search

Modified Frequentist (a.k.a. CL_s) method

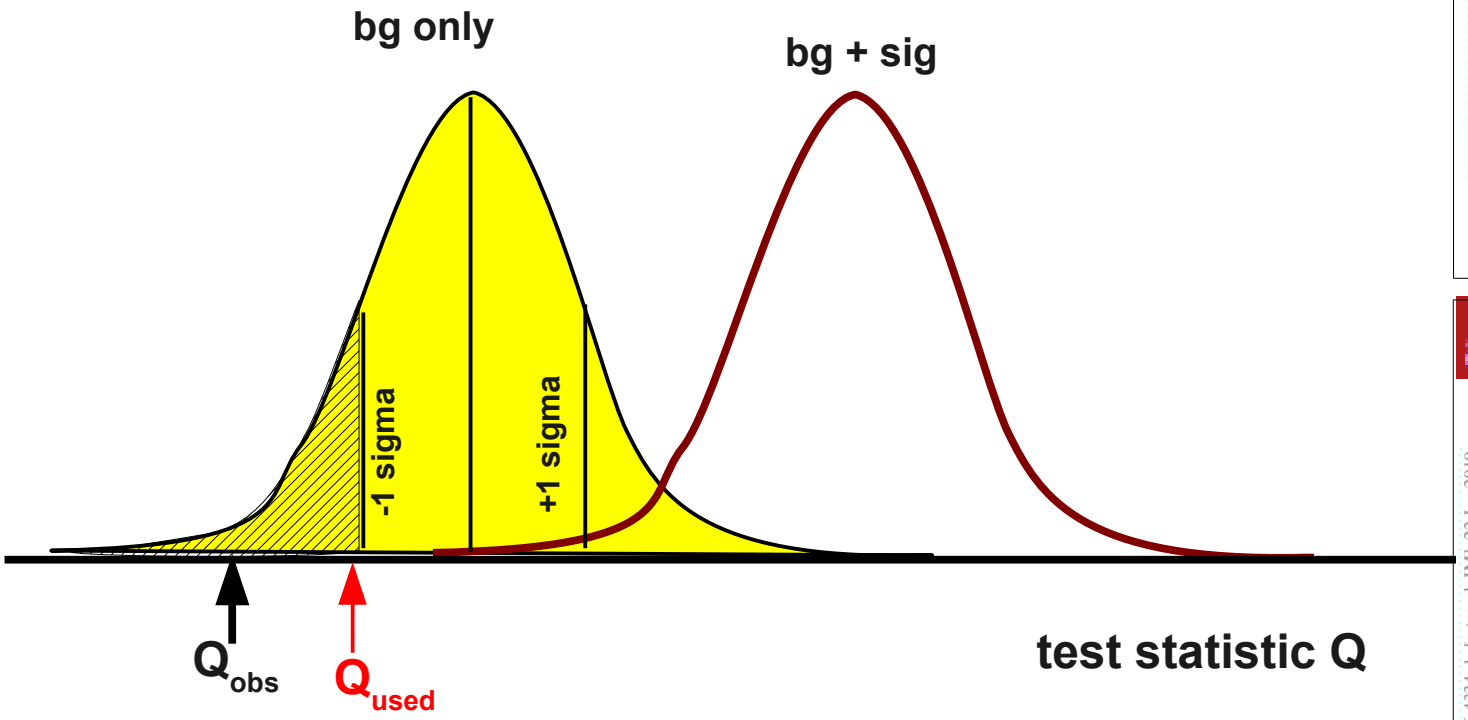
define:
$CL_s = CL_{s+b} / CL_b$
and require $CL_s(\mu) = 10\%$
for a 90% 'CL' limit



- If m = 0, CLs = 1 → never exlude this
- Only exclude values for which there is some ability to observe them
- Overcoverage : limits are 'worse'
  - nevertheless quite widely used: LEP, Tevatron, LHC...
- easy to implement
- unpopular with statisticians :
  - CLs is not a confidence level

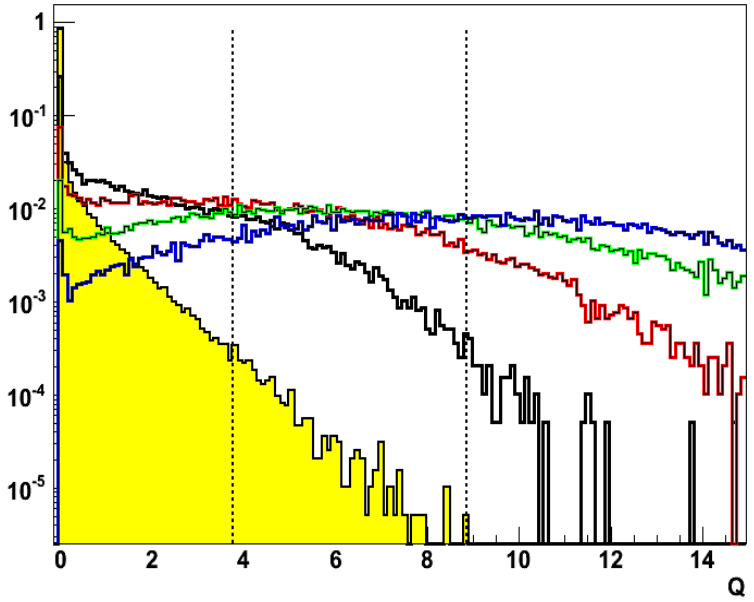
Power constrained limits

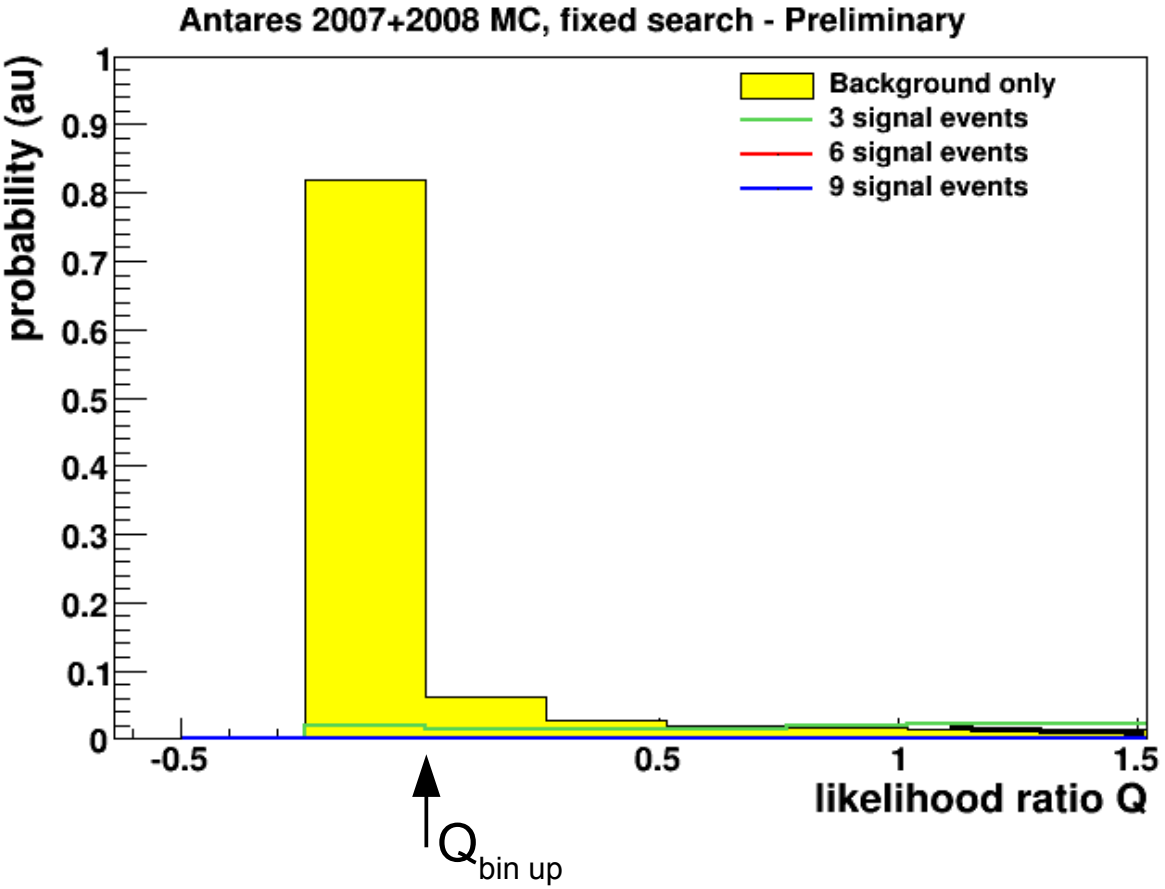
- If the observed limit is lower than some threshold, the actual limit is reported for the threshold value.
- The threshold is determined from the bg-only distribution



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

'Power constrained' by accident?





doing this:

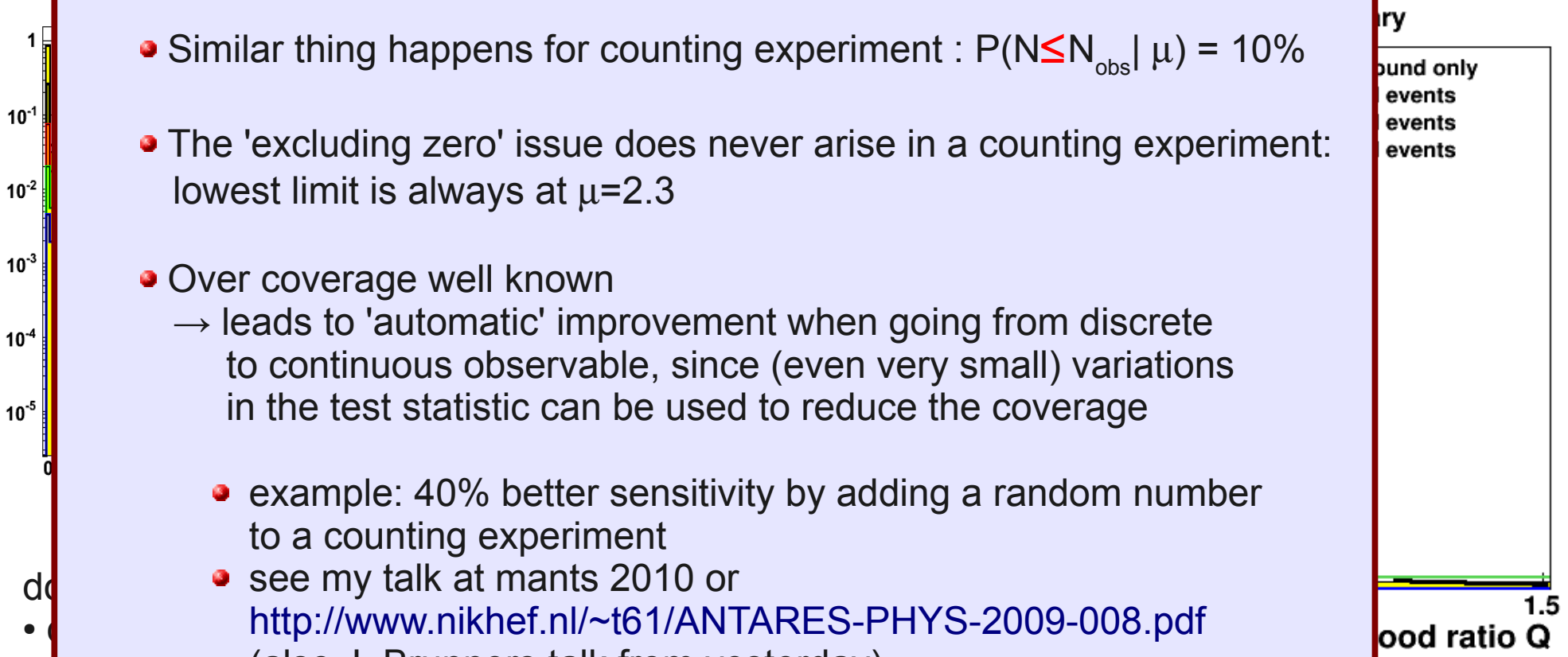
- over-covers (badly)
  → higher limits that needed for coverage
- can solve/hide the problem of excluding zero
- result depends on binning chosen (probably not desirable)

what happens depends on details of the code, but for events in 1st bin likely to amount to:

$$P(bin \leq bin_{obs}) = P(Q < Q_{bin\ up} \mid \mu) = 10\ \%$$

'Power constrained' by accident?

- Similar thing happens for counting experiment : $P(N \leq N_{obs} | \mu) = 10\%$
- The 'excluding zero' issue does never arise in a counting experiment: lowest limit is always at μ=2.3
- Over coverage well known
  → leads to 'automatic' improvement when going from discrete to continuous observable, since (even very small) variations in the test statistic can be used to reduce the coverage
  - example: 40% better sensitivity by adding a random number to a counting experiment
  - see my talk at mants 2010 or http://www.nikhef.nl/~t61/ANTARES-PHYS-2009-008.pdf (also J. Brunners talk from yesterday)



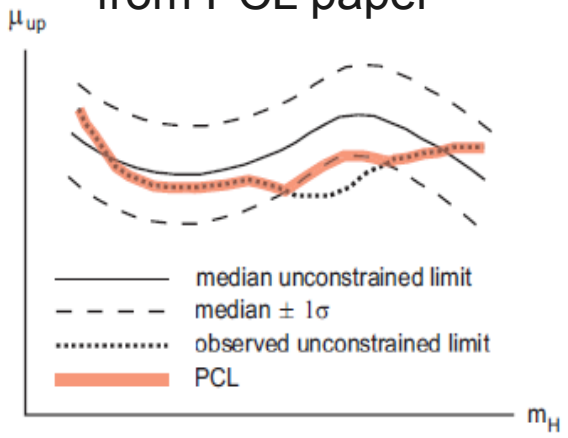
- can solve/hide the problem of excluding zero
- result depends on binning chosen (probably not desirable)

what happens depends on details of the code, but for events in 1st bin likely to amount to:

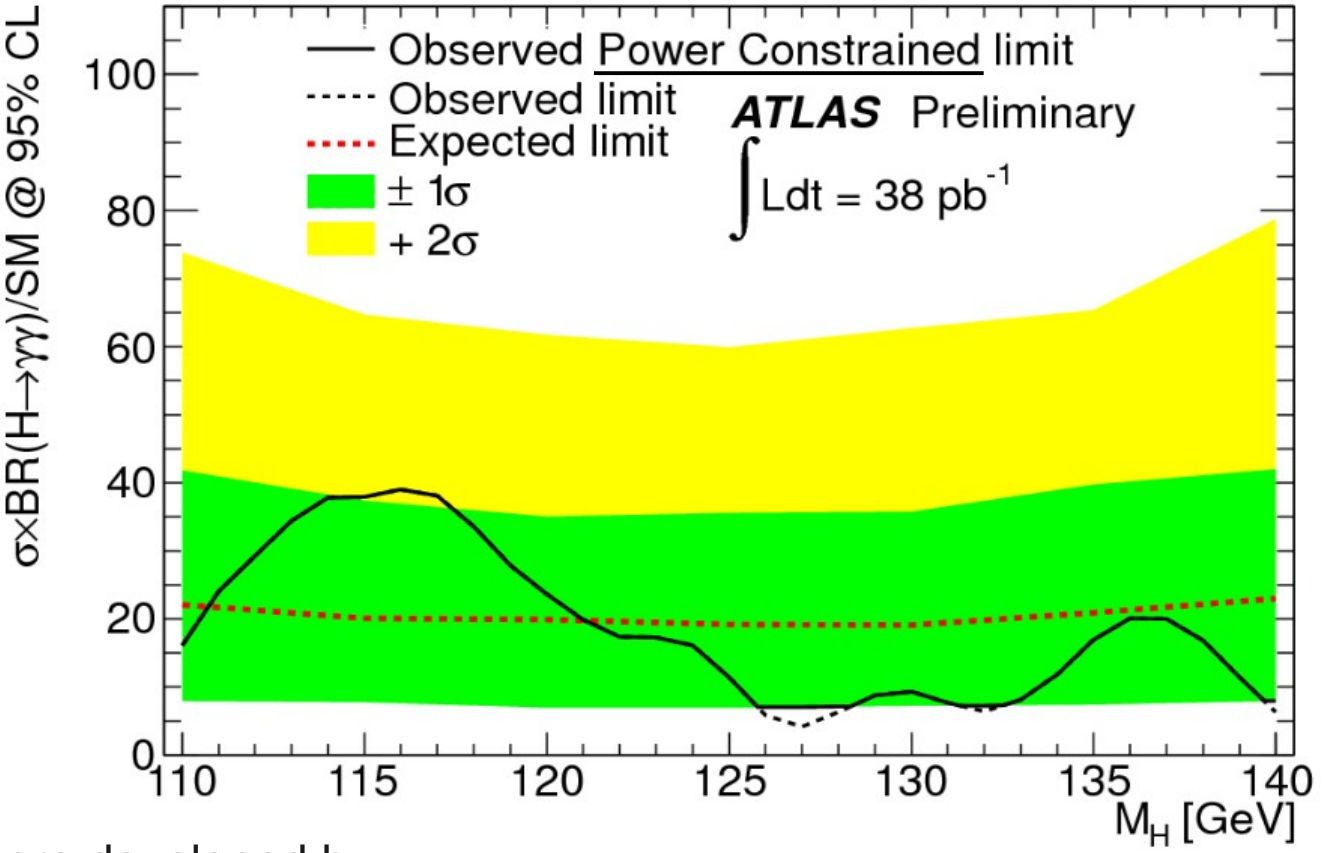
$$P(\text{bin} \leq \text{bin}_{obs}) = P(Q < Q_{\text{bin up}} | \mu) = 10\ \%$$

Meanwhile at the LHC...

from PCL paper



from Moriond



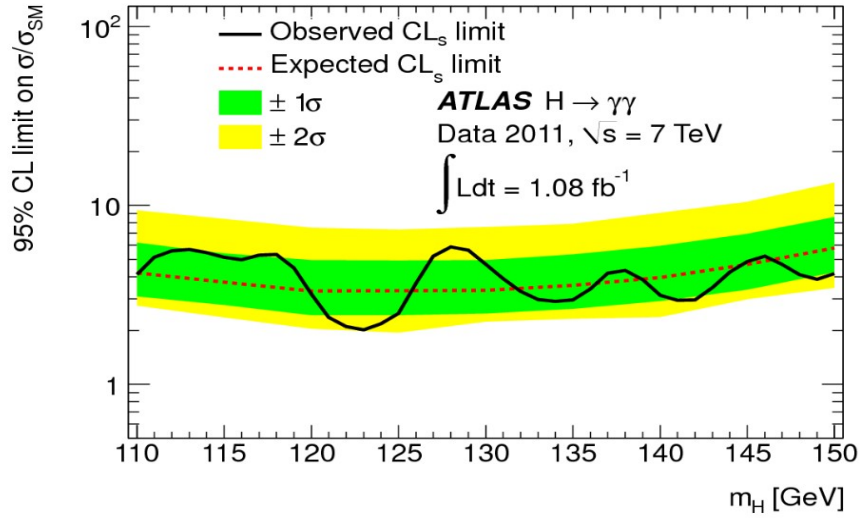
- Power-constrained limits were developed by Atlas member and addopted as 'official'
- Used for several Atlas analyses (Moriond 2011)

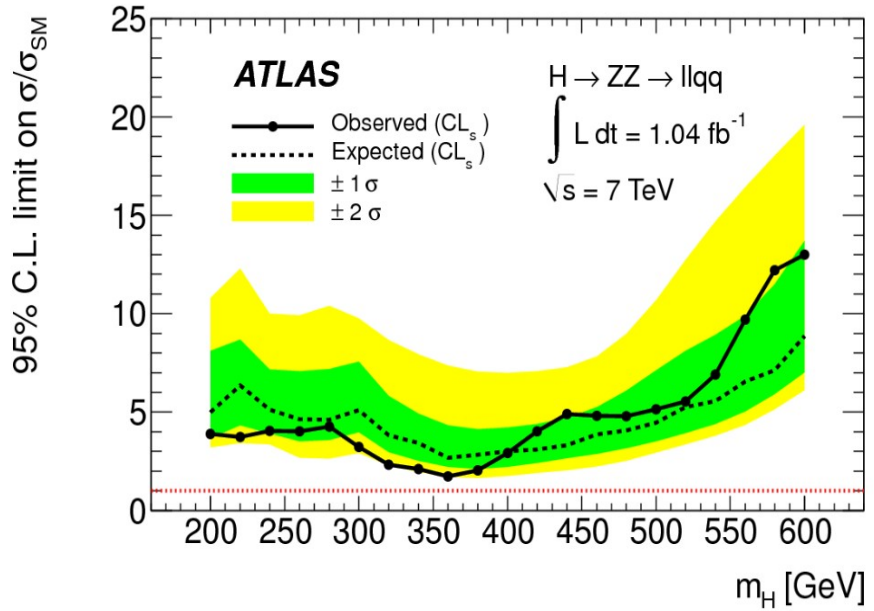
- note: they use threshold = median -1σ
  could also use: threshold = median (Juergen would like that...)

however..

Meanwhile at the LHC...

lepton-photon





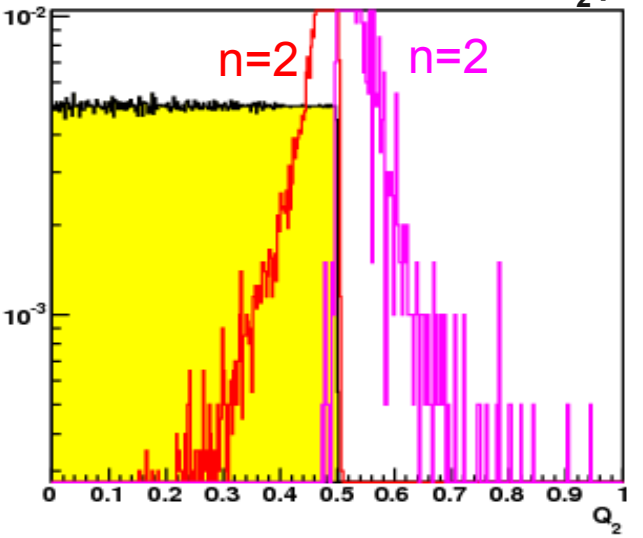
Atlas has now decided that it will produce CLs -type limits for its results.. (as a temporary solution).

- **after discussion with CMS → allows to compare directly**
- **No power constrained limits shown for recent (lepton-photon) results.**
- **Bayesian methods also still allowed (I have not talked about them .. ..they're especially popular in CMS)**
- **seems CLs is not going away easily (but plan is still to use PCL in the future)**
- **Feldman & Cousins seems not to be on their radar**

Feldman & Cousins

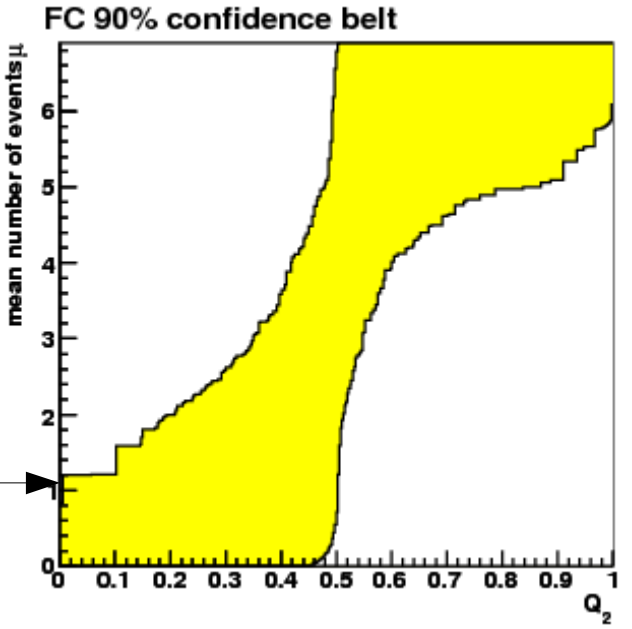
- Prevents excluding zero (by spending coverage on lower limit)
- produces double sided interval (we don't really care)
- Can be difficult to implement:
  - likelihood ordering requires many pseudo-experiments to work well..
  - a transformation of the test statistic can help, but still

transformed test statistic $Q_2(Q)$



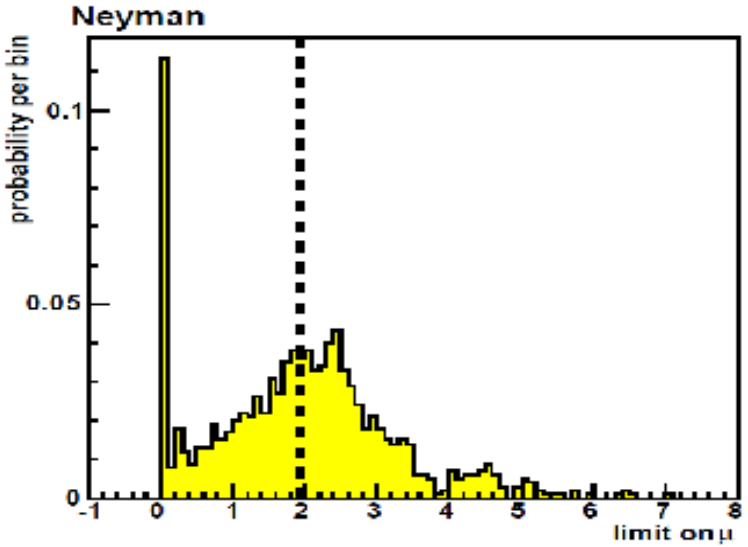
we chose it because:

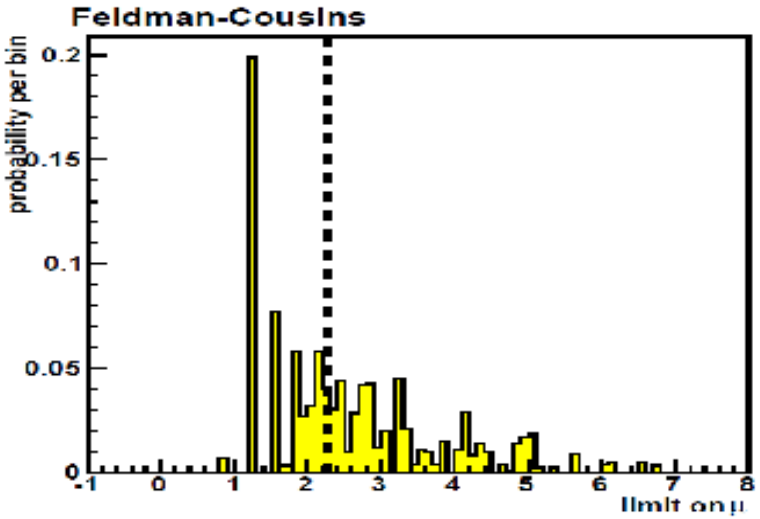
- IceCube uses it
- allows use of full range of continuous variable without the need for additional measures (like power-constraining or something that depends on the binning)
- better coverage (lower limits) than CL**s**

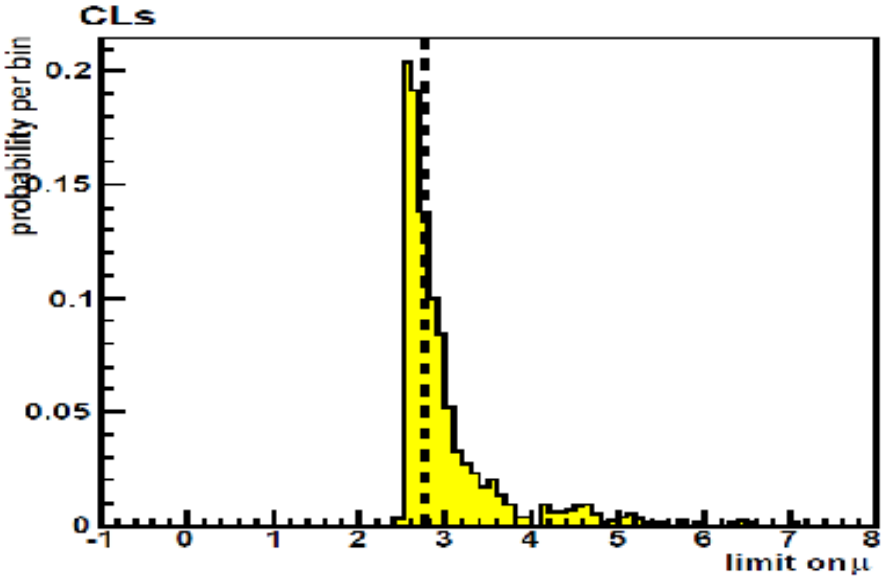


lowest possible limit around 1 event (not unreasonable)

Comparison of methods







- Neyman:
  - Easy to implement, exact coverage = lowest possible limits
  - non-physical limits (undesirable)
- Feldman-Cousins
  - tedious to implement (for continuous variable)
  - modest overcoverage
  - no unphysical limits
- CLs / Modified Frequentist (CERN-OPEN-2000-205)
  - easy to implement
  - limit does not depend on bg-only fluctuations that do not look like signal
  - severe overcoverage -> high values for limit
- **Power constrained limits**
  - easy modification of 'Neyman'
  - not yet widely accepted (but maybe soon)
  - threshold is somewhat arbitrary

Questions and thoughts

- Do we desire to use a single limit setting method
  - across experiments (Antares/IceCube/others?)
  - different measurement (e.g. do we care if the point sources use another type of limit than the diffuse flux analysis.. this is currently the case)
- Do we treat the very bg-like events in the same way?
  - limit distribution suggests that we do not
    (ic40 result looks like there are very few points below the sensitivity)
- For point sources: do we want to change from F&C to..
  - Power constraint limits (fine, but perhaps a bit too new for some readers)
  - CLs (used very widely still in HEP despite that statisticians don't like it)
  - something else?