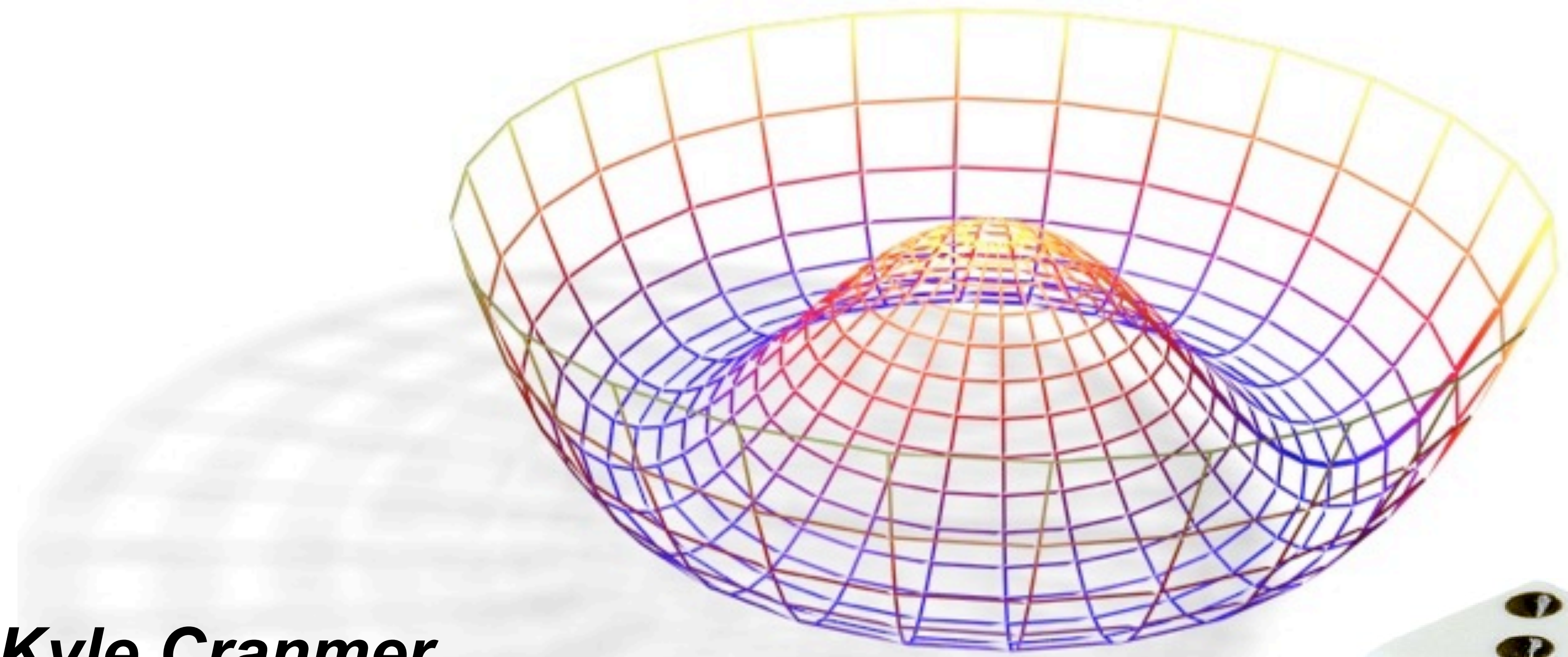
Center for
Cosmology and
Particle Physics

Update on the frequentist limit recommendation



Kyle Cranmer,
New York University

Outline

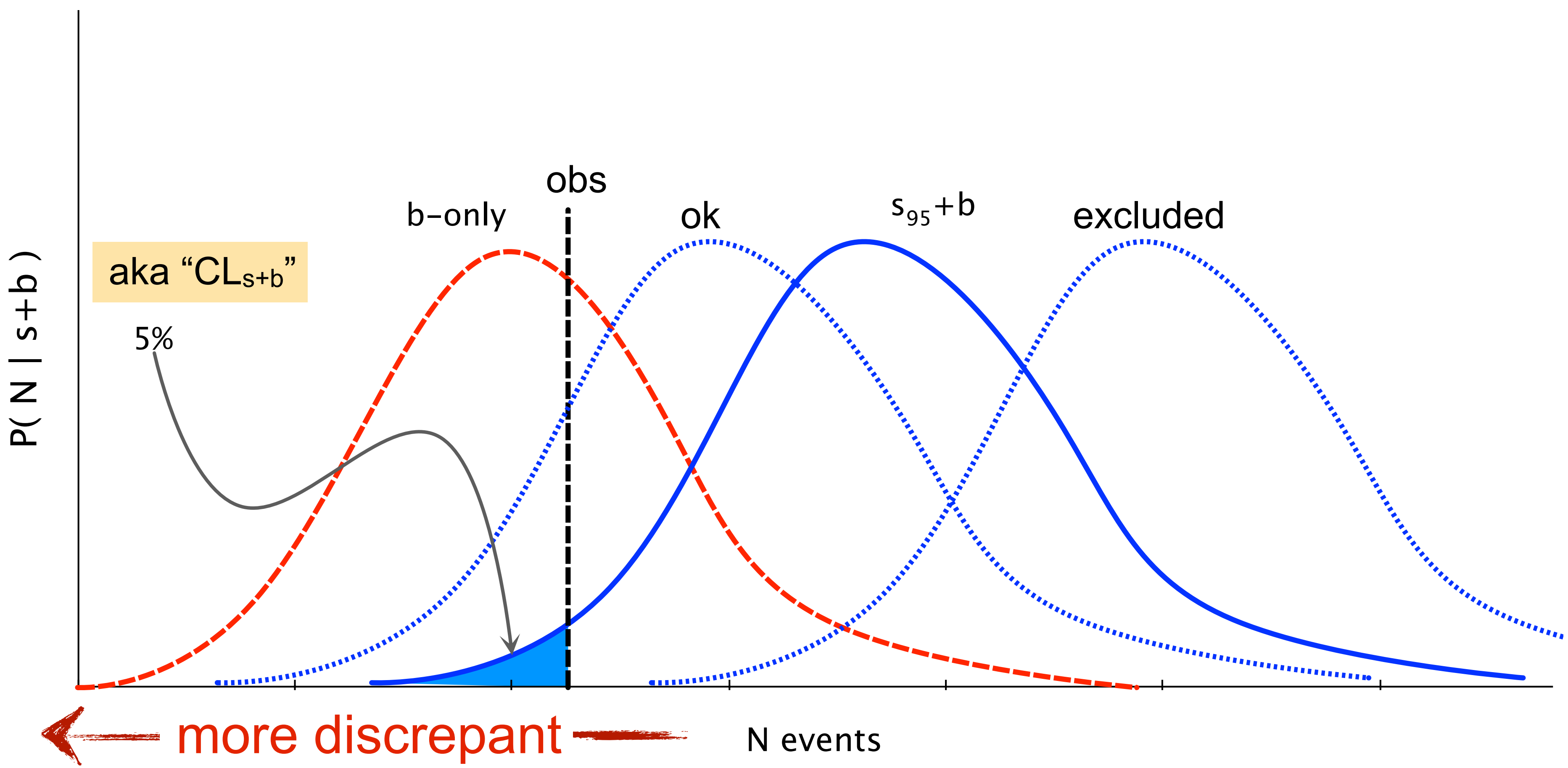
Brief reminder on the recommendation

A subtlety in the bands for small numbers of events

Some observations about the power-constraint

Recall, what we mean by 95% upper-limit

- increase s until tail probability is 5%



CL_s

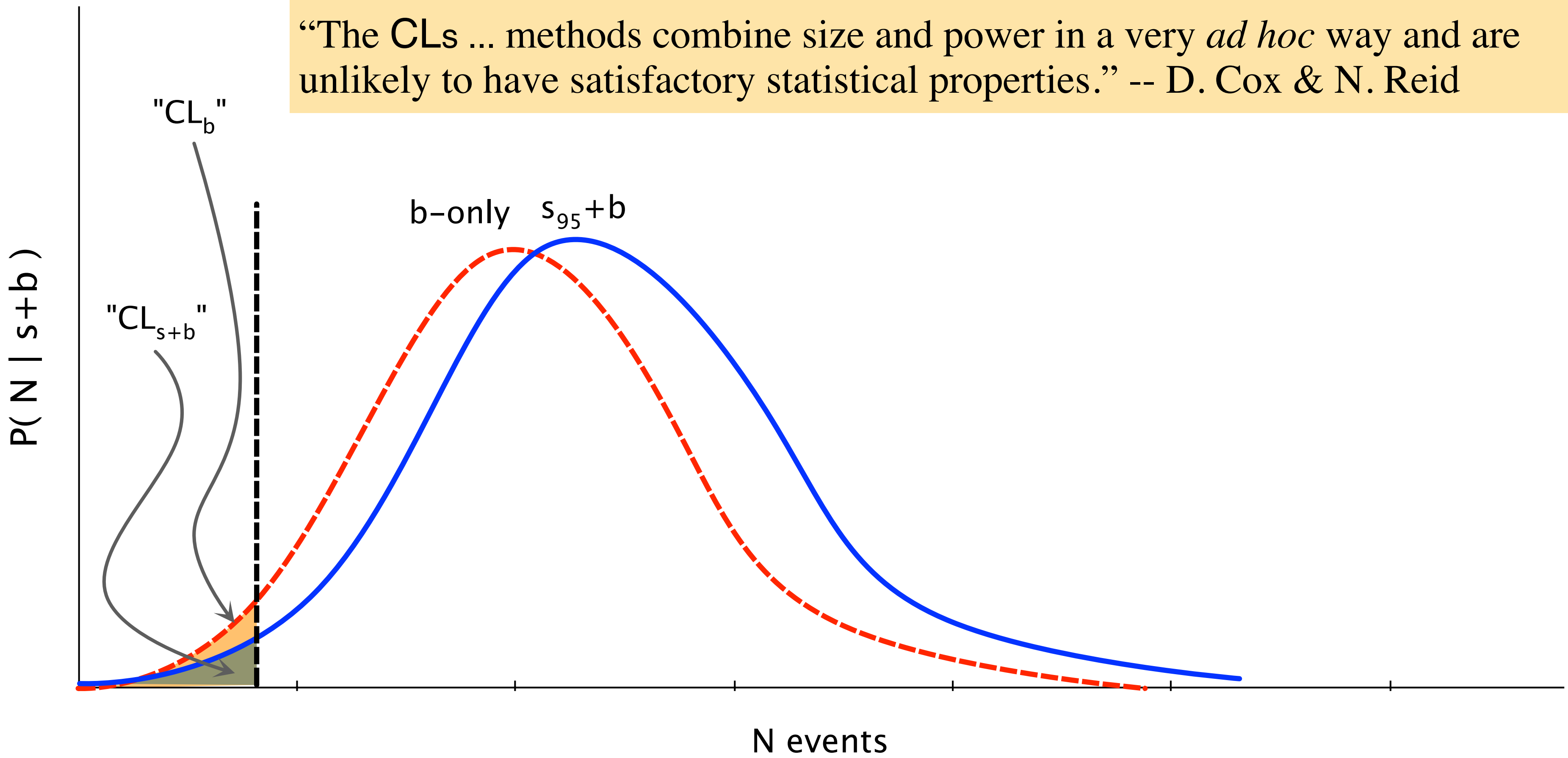
To address the sensitivity problem, CLs was introduced

- common (misused) nomenclature: $CL_s = CL_{s+b}/CL_b$
- idea: only exclude if CL_s<5% (if CL_b is small, CL_s gets bigger)

CL_s is known to be "conservative" (over-cover): expected limit covers with 97.5%

- amount by which CLs over-covers is not transparent to the reader

"The CLs ... methods combine size and power in a very *ad hoc* way and are unlikely to have satisfactory statistical properties." -- D. Cox & N. Reid



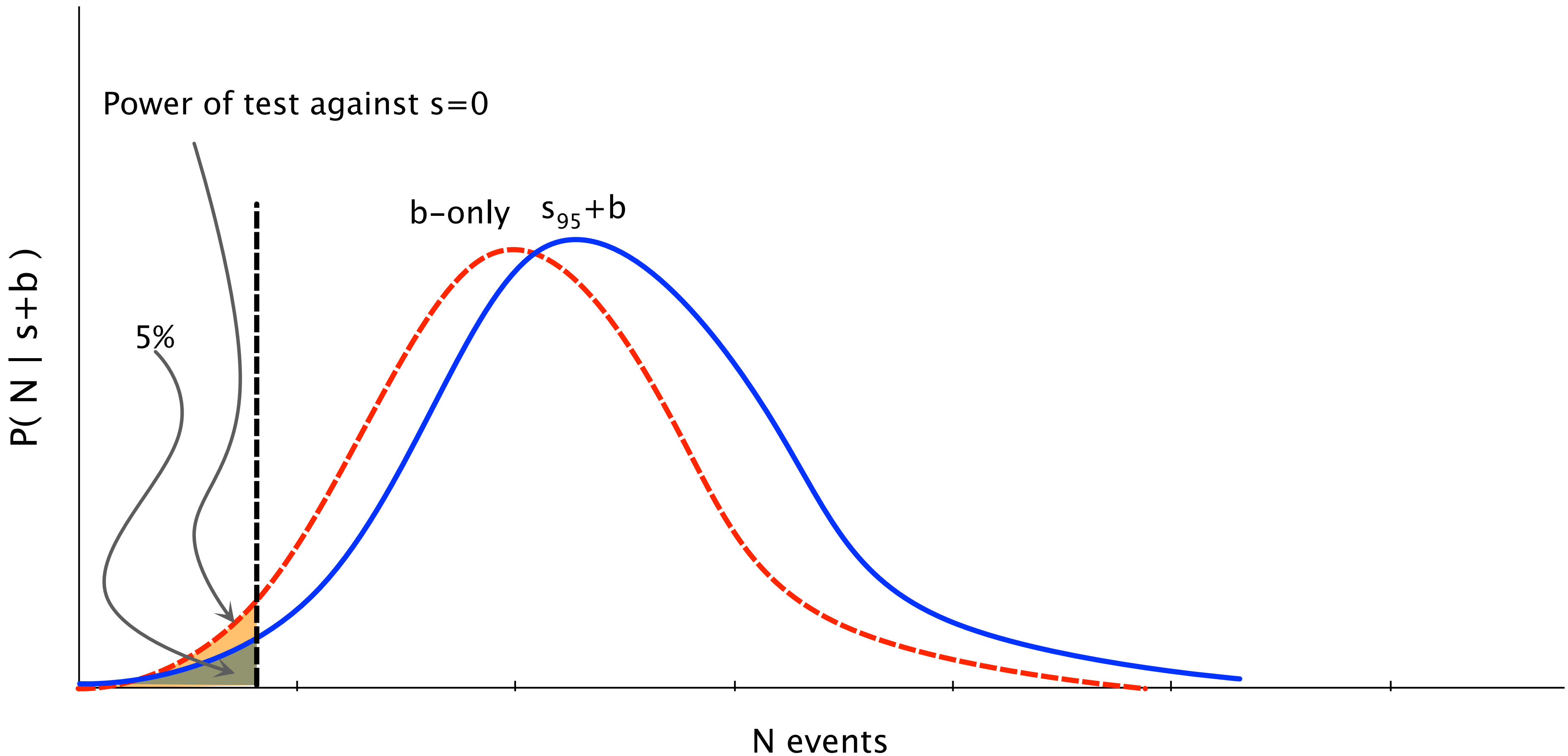
Power in the context of limits

Center for Cosmology and Particle Physics

The power-constraint approach uses the same information as CLs, but keeps the two pieces of information separate

- CL_{s+b} is used for the limit
- CL_b is used to define a “sensitivity”

Two pieces of information with well-defined properties (instead of one without)



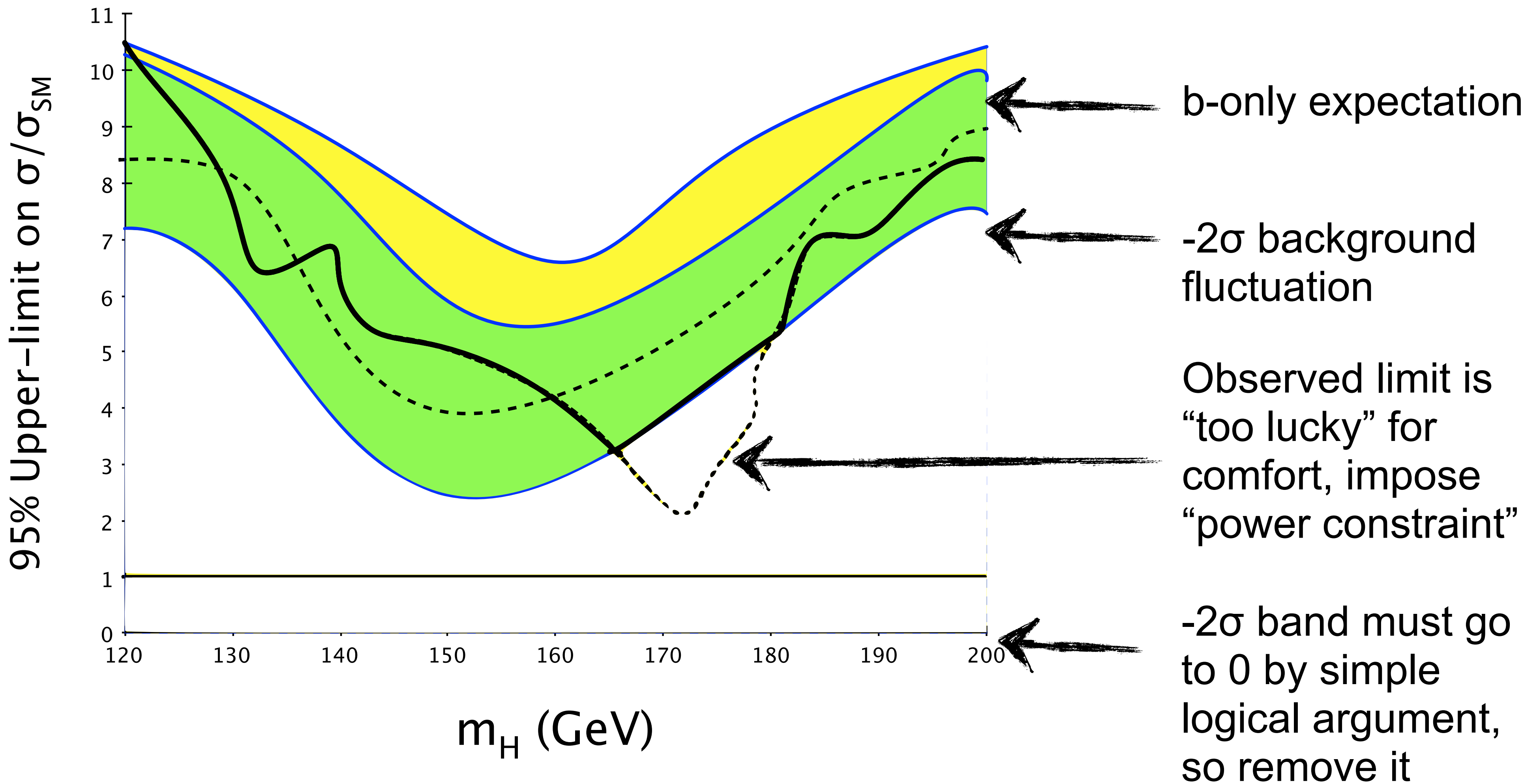
PCL and the bands

Center for
Cosmology and
Particle Physics

The recommended plot looks like the one below

- We have been using the -1σ band as the power-constraint
  - yes, it's a 16% is a convention... just like 95% is a convention

Focus here is on the importance of the bands

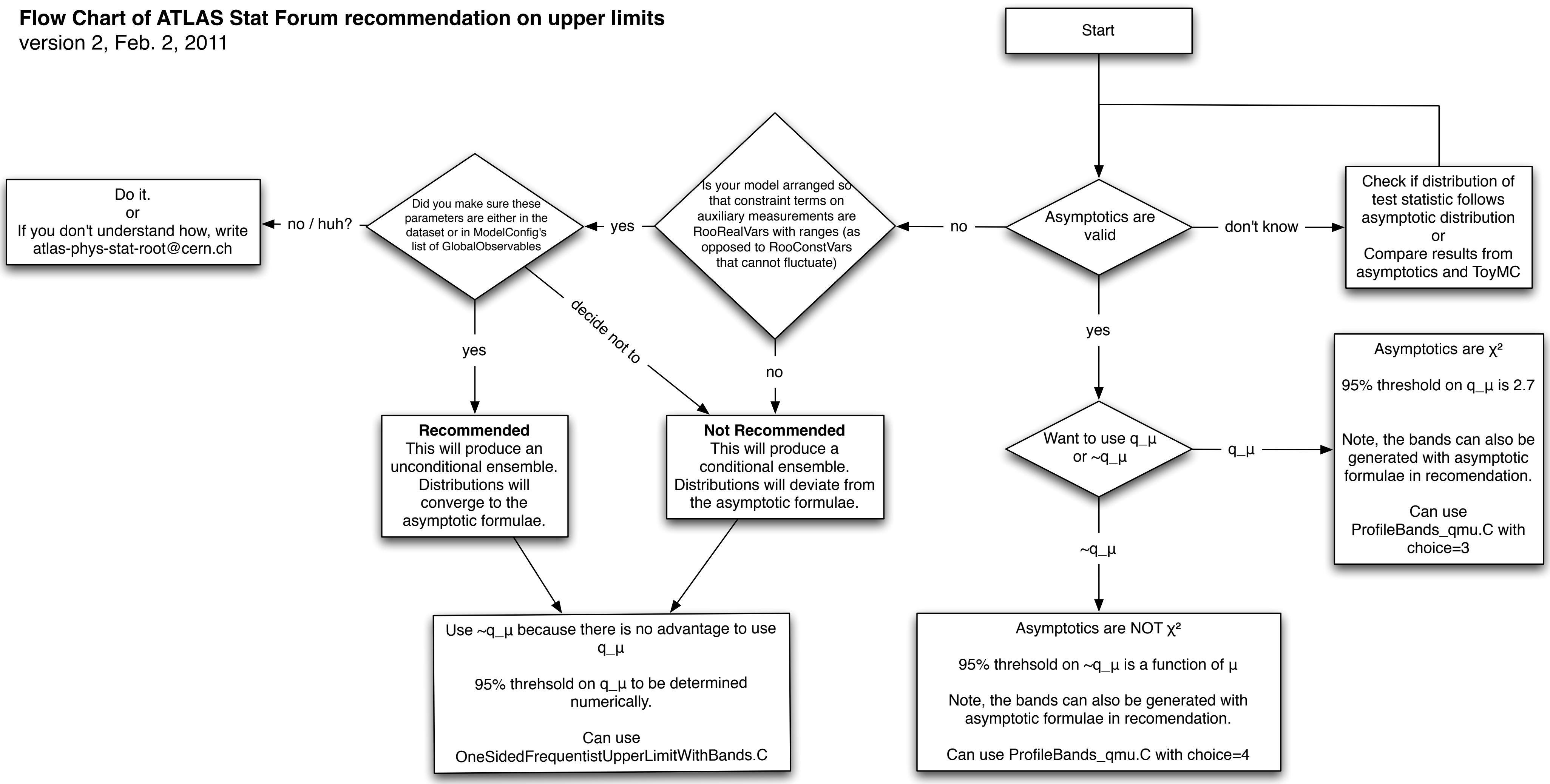


Recommendations: Flow-chart

Center for Cosmology and Particle Physics

Flow chart outlines recommendations. Specific scripts are available that implement the recommendations with RooStats tools.

Same workspace can be used with other statistical methods.



Some properties to keep in mind

These asymptotic properties are basis for much of the logic:

1. the value of the test statistic q_μ for some given data is **independent** of the value of the nuisance parameter θ
2. the distribution $f(q_\mu \mid \mu, \theta)$ is **independent** of the value of the nuisance parameter θ and has an analytic form
3. the distribution of $f(q_\mu \mid 0, \theta)$ **depends** on the value of the nuisance parameter θ

Thus:

- In the **asymptotic regime**, the distributions have a known form
- In an **intermediate regime**, we need to use toy MC to calibrate the distributions, but we can assume they are still roughly independent of θ
- In the **low-count regime**, we can't rely on this assumption
  - this is where we will update the recommendation

Note, this 3. means that even asymptotically, CL_s depends on the treatment of the nuisance parameters, while CL_{s+b} does not.

How we find the upper-limit

Center for Cosmology and Particle Physics

The confidence interval (upper-limit) is based on a Neyman-Construction.

- can't deal with space of all nuisance parameters, so we only perform construction along profiled path (called "Hybrid resampling" by statisticians)
- For each value of μ, we find threshold T(μ) that holds 95%.
- Exclude when $q_\mu > T(\mu)$,

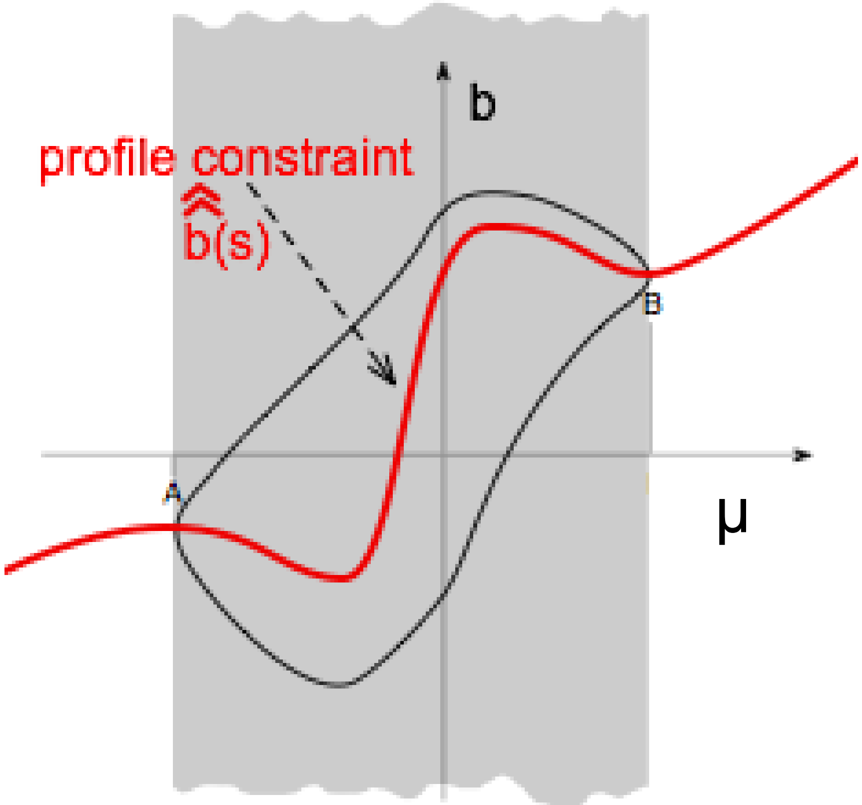
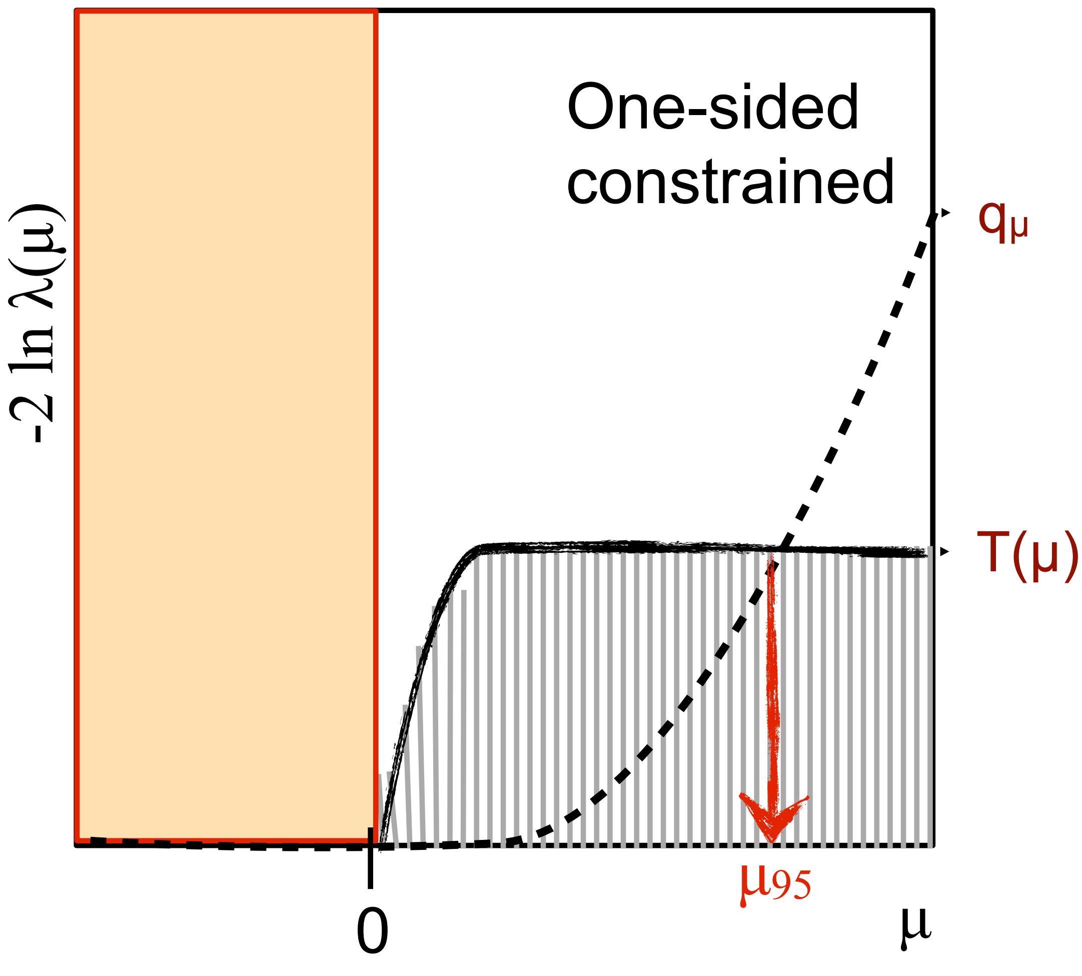


Figure 7.2: MINOS error confidence region for parameter 1

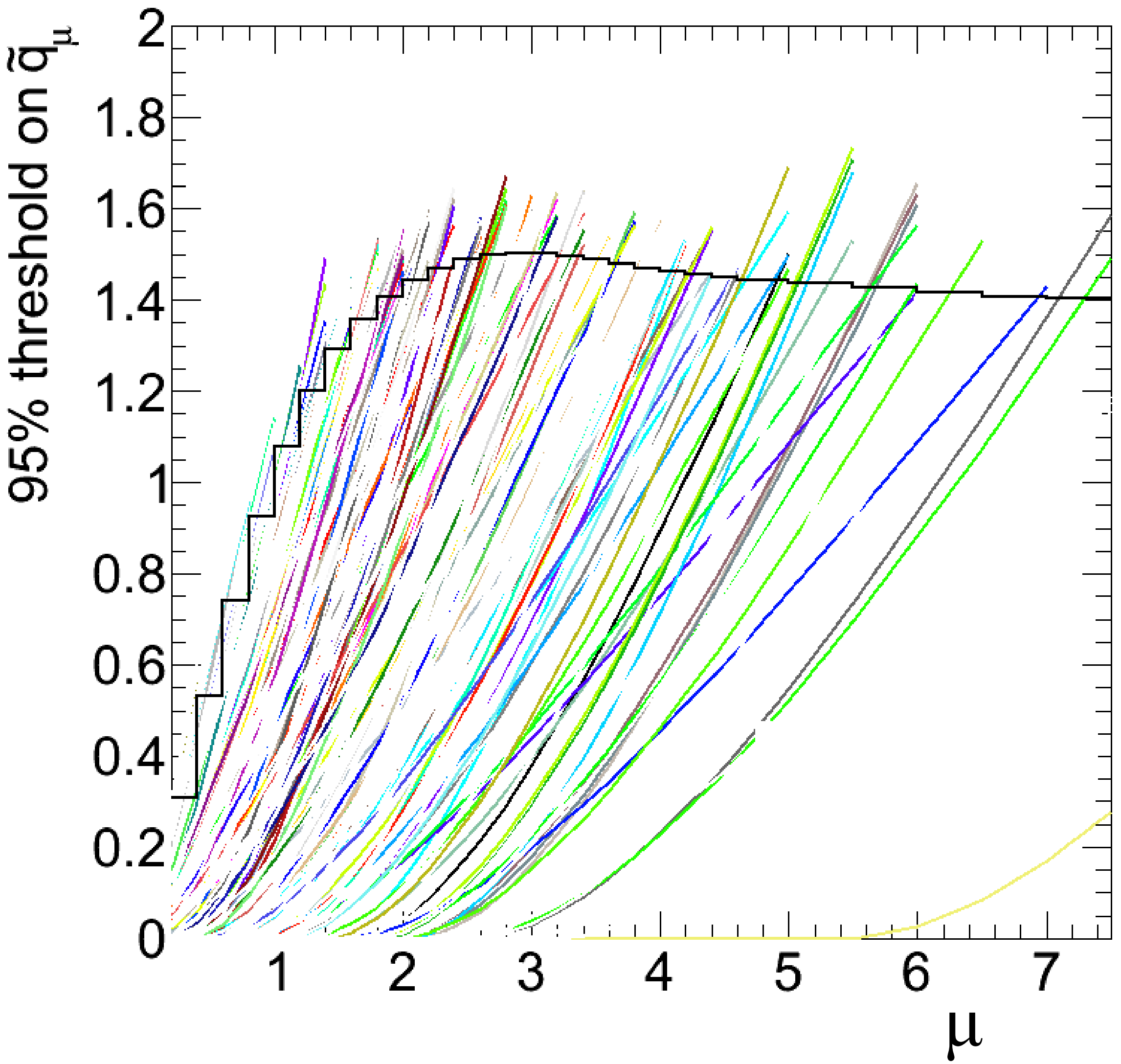


How we find the bands

Center for
Cosmology and
Particle Physics

Each colored curve represents q_μ for a single b-only pseudo-experiment

- Find upper-limit for each, build distribution of upper-limits
- use this to define bands, power-constraint



The subtlety we found with few events

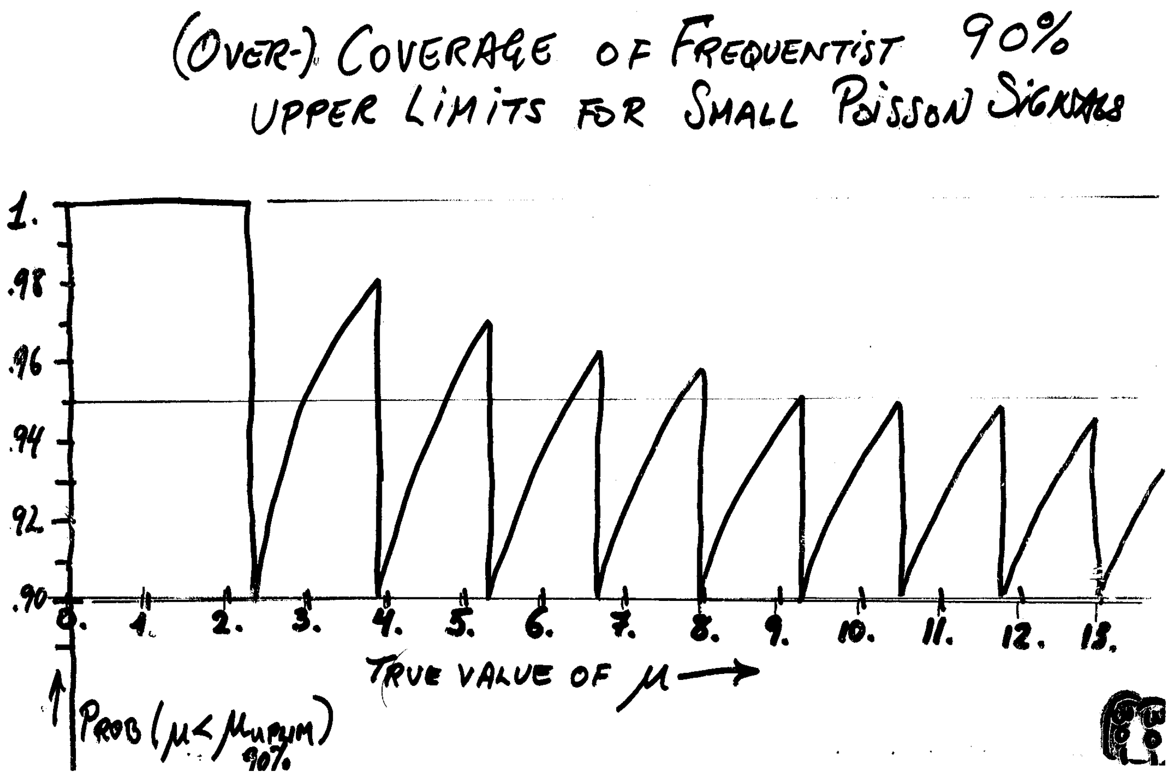
Reminder on Discrete Problems

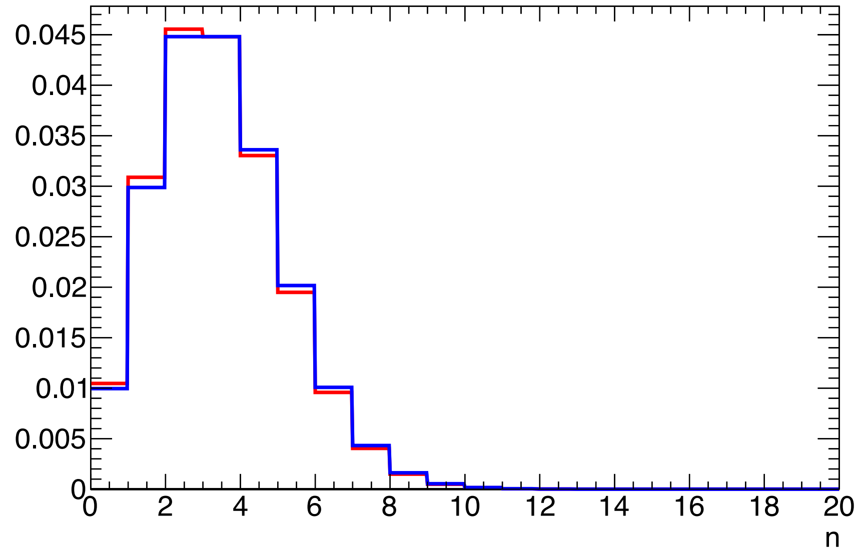
In discrete problems (eg. number counting analysis with counts described by a Poisson) one sees:

- discontinuities in the coverage (as a function of parameter)
- over-coverage (in some regions)

When there are systematics, the Poisson discreteness is broken

- For N=0 and b«1, the familiar limit of s_{95}=3 changes to s_{95}=2.3
- In some cases this 2.3 has exact coverage for all values, worst case is 90%

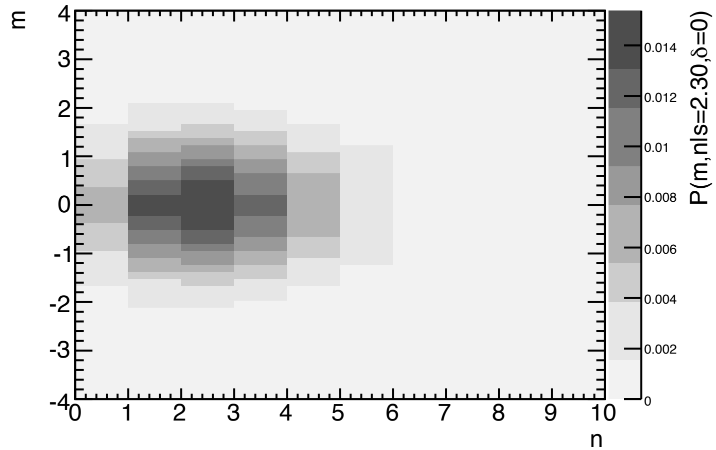


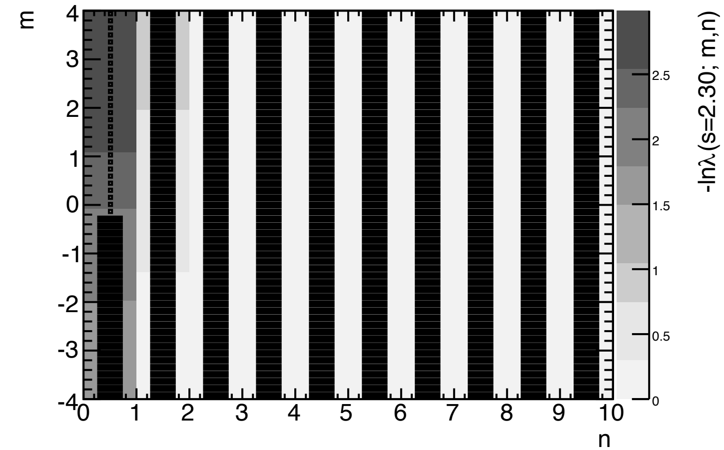


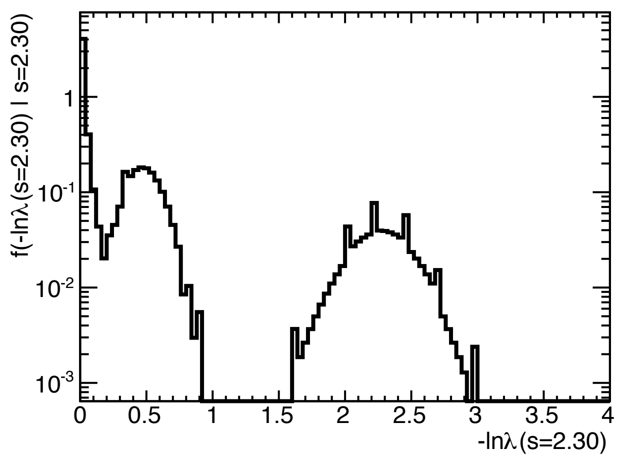
Toy Problem

In order to study the low-count situation with systematics, consider a simple extension to Pois(n | s+b) with systematic δ on signal and background rate, constrained by auxiliary measurement m

$$P(n,m|s,\delta) = Pois(n|(1+\eta_s\delta)s + (1+\eta_b\delta)b)\, Gaus(m|\delta,1).$$







Where one would have previously had delta functions at N=0,1,2,...

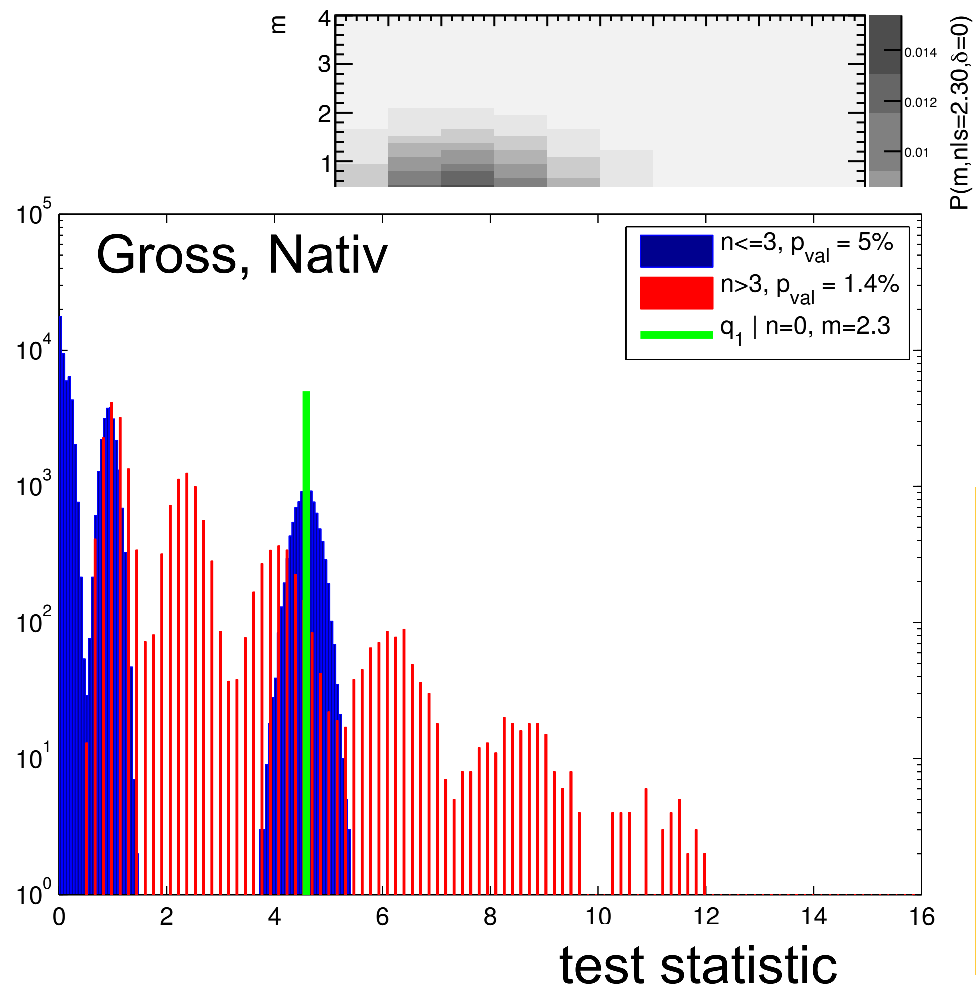
Now we get small mountains corresponding to fluctuations in the auxiliary measurement m

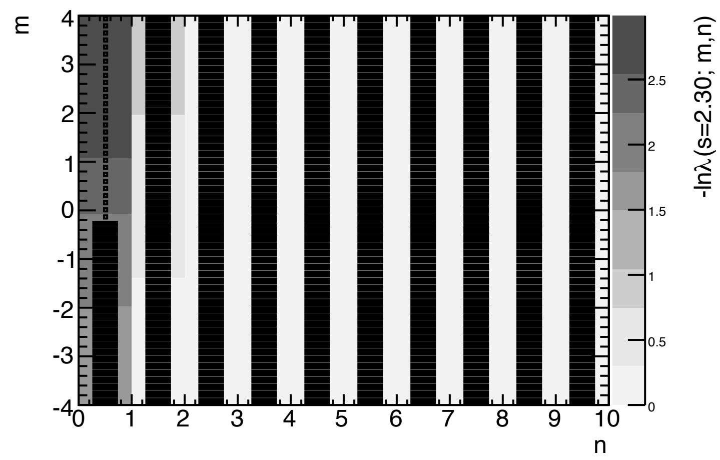
Toy Problem

Center for Cosmology and Particle Physics

In order to study the low-count situation with systematics, consider a simple extension to Pois(n | s+b) with systematic δ on signal and background rate, constrained by auxiliary measurement m

$$P(n, m|s, \delta) = Pois(n|(1+\eta_s\delta)s + (1+\eta_b\delta)b)\, Gaus(m|\delta, 1).$$





Where one would have previously had delta functions at N=0,1,2,...

Now we get small mountains corresponding to fluctuations in the auxiliary measurement m

The -1σ band for few events

In a recent analysis with N=0 and b≪1, the script that implements the recommendation was returning s_{95}~2.3 as expected, but the -1σ band was about 1.2 events.

- much discussion with Henri, Haichen, Ofer, Glen, myself
- In these cases, we expect N=0 background-only

Simply put, what type of fluctuation could lead to a limit that is almost twice as strong?

- If you repeat the argument of why one can get a limit of s_{95}~2.3 events for several b-only toys, you would expect the distribution of upper-limits from b-only to be very narrow around s_{95}~2.3

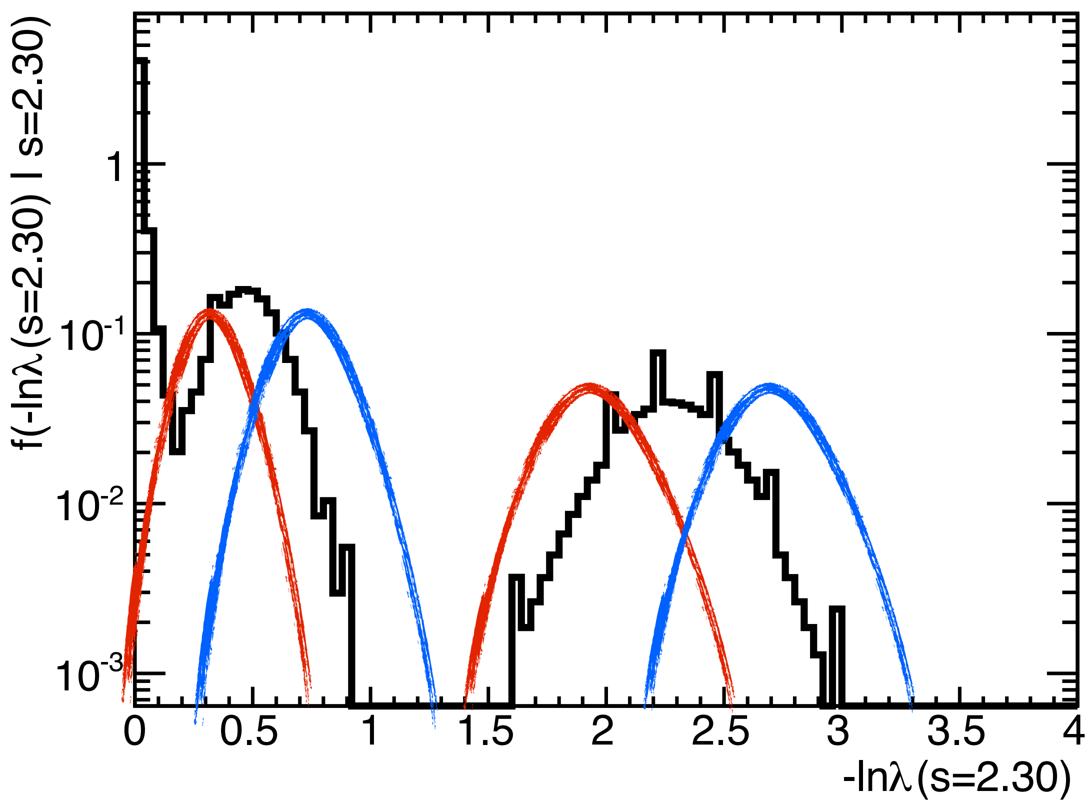
The problem

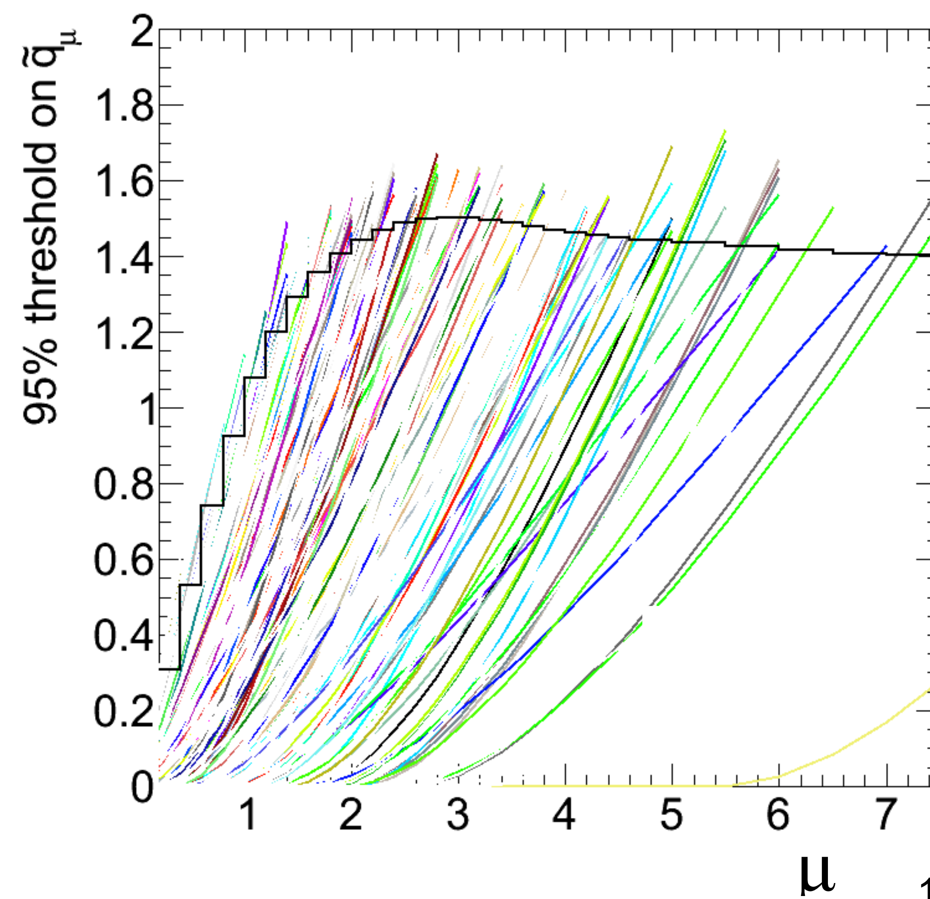
Center for Cosmology and Particle Physics

It’s a bit difficult to explain this, but essentially the point is that a fluctuations in the auxiliary measurement lead to small changes to the value of the test statistic.

- the problem is that we are re-using the T(μ) thresholds built from profiling on the observed data, not this particular b-only toy

$$P(n,m|s,\delta) = Pois(n|(1+\eta_s\delta)s + (1+\eta_b\delta)b)\,Gaus(m|\delta,1).$$





Solution

In short the solution is that in the low-count regime we need to repeat the entire procedure for each b-only toy

- this means a new profile construction for each b-only toy
- this will put the nuisance parameters so that the auxiliary measurement is near the median

Consequences: While this sounds like it would be computationally impractical, it's not as bad as it sounds

- Currently we use N toys for each of the M μ points we test to find T(μ). Then we run B toys and observed data to find limits. So we have ~NxM+B+1 toy runs
- If we only wanted the observed limit, we can do cleaver tricks so that we only need ~2N toys near μ_{95}
- So with about 2N*(B+1) toys we can get observed and build bands
- In very-low count, bands are narrow, so we may be able to use a smaller B

Practical: updated scripts in progress, another area we could use help

Some observations

Some observations about 16 vs. 50%

The median would actually have been stable to this problem that we observed.

Some have pointed out that over-estimating systematics might widen the band, thus reducing the power constraint... “being optimistic by being conservative”. But this not the case with the median.

Computational: It requires more b-only toys to estimate the 16% quantile than the median

Remember that CLs continues to have a sensitivity to the nuisance parameters even in the asymptotic regime

Comparison of 50% PCL & CLs

The CLs procedure purposefully over-covers (“conservative”)

- and it is not possible for the reader to determine by how much

The power-constrained approach has the specified coverage until the constraint is applied, at which point the coverage is 100%

- limits are not ‘aggressive’ in the sense that they under-cover
- arbitrary sensitivity estimate is explicit, coverage is explicit