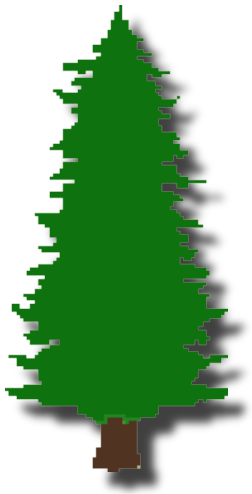
"The Hunting of the Sigma":
Statistical Interpretations in ATLAS

Jason Nielsen

Santa Cruz Institute for Particle Physics

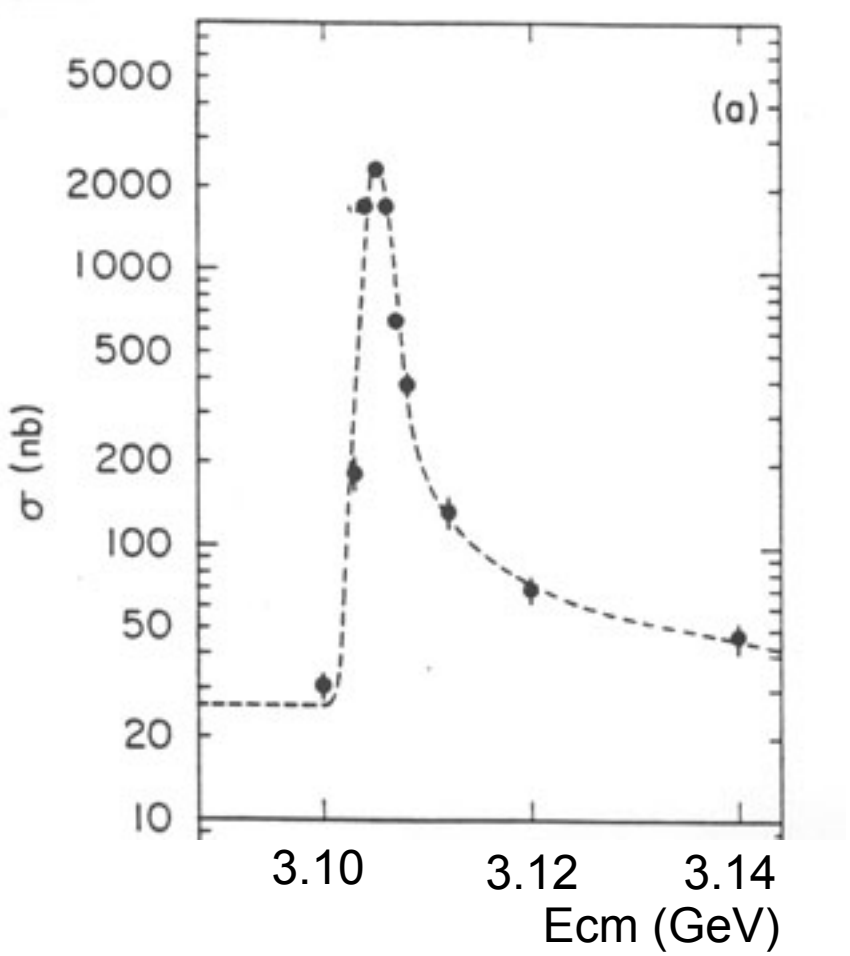
University of California, Santa Cruz



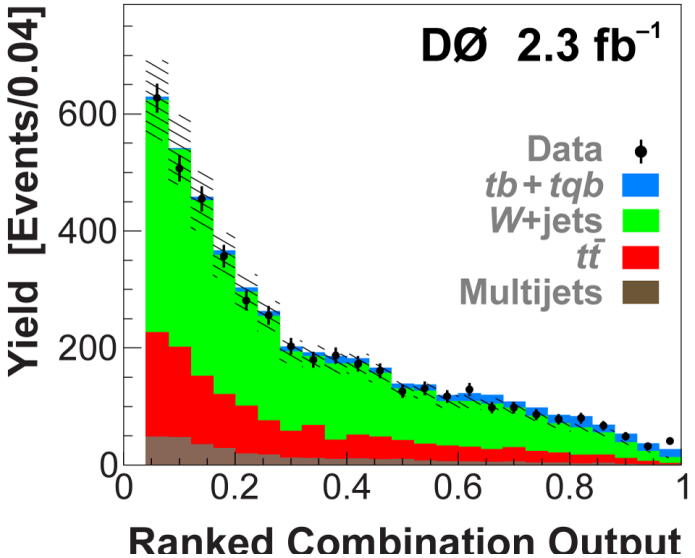
SLAC Theory Seminar

July 14, 2011

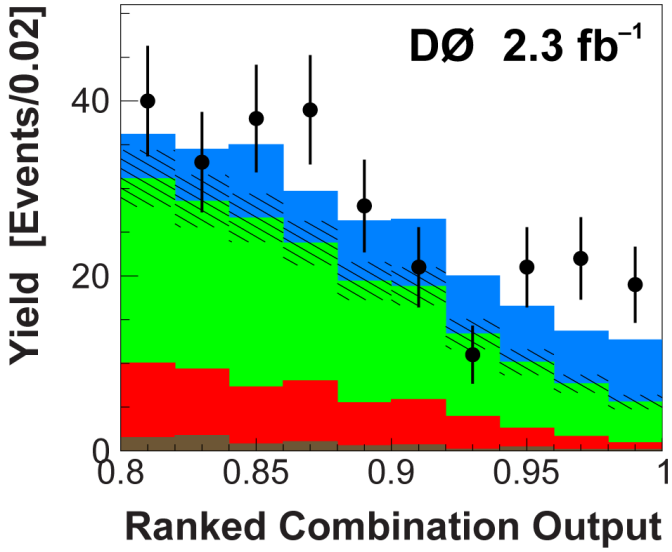
Two Examples of Discovery



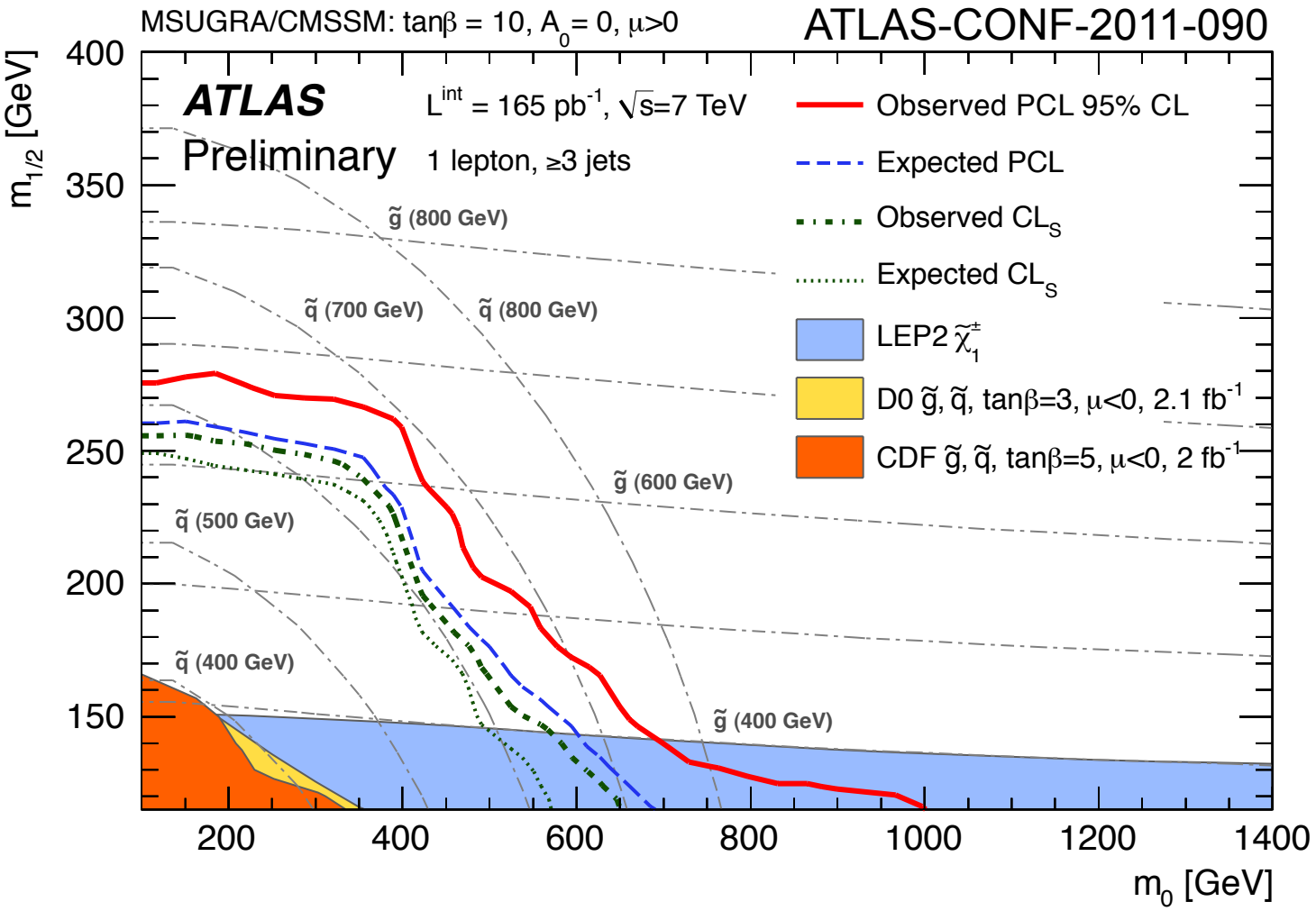
Final Discriminant



Signal Region



Comparison of Limit-Setting Methods



ATLAS Limit Recommendations

I focus on frequentist methods, as opposed to Bayesian

- Adopt profile likelihood ratio to deal with "nuisance parameters" (uncertainties and background estimates)
- Reduce sensitivity to out-of-reach signals by calculating "Power-Constrained Limits" and "CLs method"
- Use asymptotic behavior of profile likelihood and upper limits when number of events is large
- New: agree to produce results with CLs method to facilitate comparison with CMS/CDF/D0

Historo-Pedagogical Outline of Topics

- Rise of frequentist methods in LEP Higgs analysis
- "Semi-Bayesian" results at LEP and Tevatron
- Rediscovery of the profile likelihood ratio
- Unnatural sensitivity: CLs vs. PCL methods
- Agreement between ATLAS and CMS
- Asymptotic behavior for large numbers of events

- Introduction of RooStats tools intended to provide uniform validated implementation of techniques

A Frequentist Checks an Unfair Coin

Example experiment: 99 heads out of 100 trials

- We will not make any statement about whether this coin is fair or not, nor any estimate of "how fair"
- We will say "this is a very unlikely outcome from a fair coin hypothesis, based on our knowledge of fair coins"
- **Frequentist methods** give probabilities under a certain hypotheses; the **confidence interval** is a way of finding hypotheses for which the outcome is not too unlikely

Frequentist Interpretations at LEP

- Used for Higgs boson searches; basis of current schemes
- Simplified prescription of hypothesis testing:
  - Define a test statistic that increases in value when the data sample becomes more "signal-like" (could be number of events or something more sophisticated like likelihood)
  - Generate a test statistic distribution for an ensemble of pseudo-data samples containing
    - signal+background
    - background-only
  - Find where the observed data's test statistic value lies in the distribution of all potential experiments
  - Calculate a p-value, giving the probability for the observed experimental result

Likelihood Ratio as Test Estimator

For testing two hypotheses, the likelihood ratio is the most powerful test statistic

$$Q = \frac{L(s+b)}{L(b)}$$

In a counting experiment, calculate Poisson probabilities

$$= \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{n!}{b^n} e^b$$

We can calculate ln Q by hand for the example case with s=8, b=8, n=16

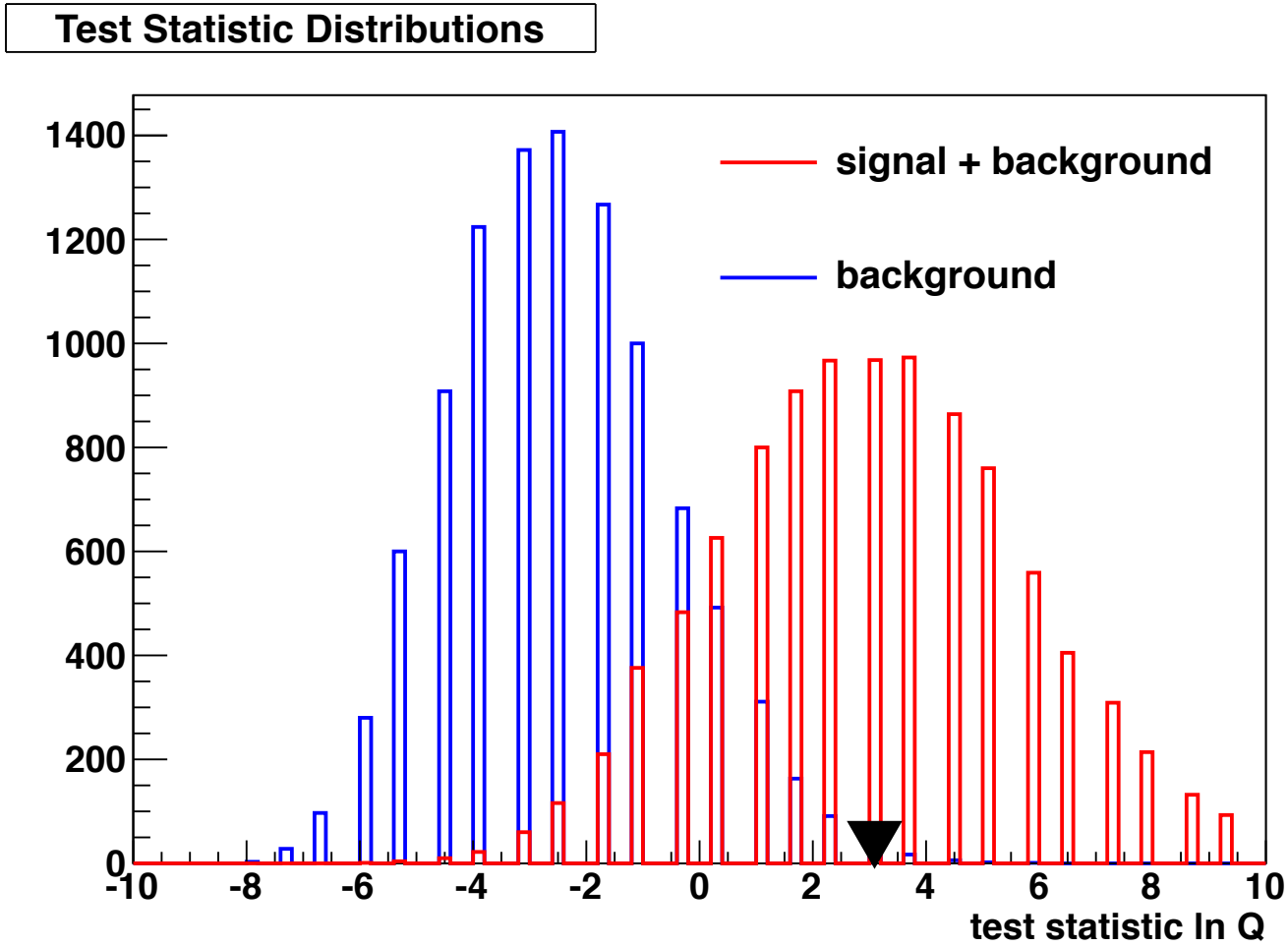
$$\ln Q = -s + n \ln\left(1 + \frac{s}{b}\right)$$

$$= -8 + 16 \ln\left(1 + \frac{8}{8}\right)$$

$$= 3.090$$

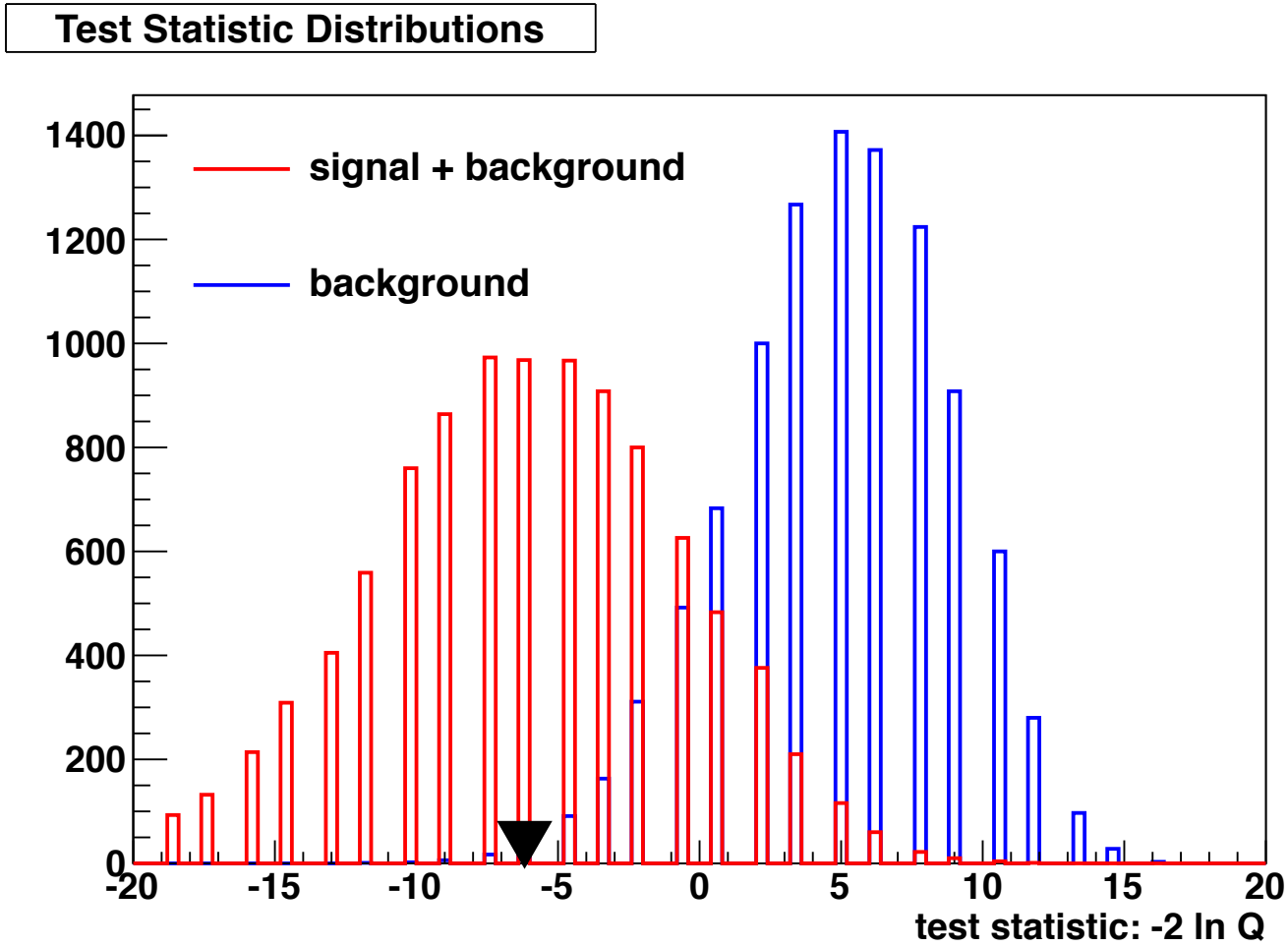
But of course we do not expect to collect 16 events each time!

Discrete Distribution with No Uncertainties



The frequentist asks: "Is the observed outcome likely, given the background-only hypothesis?"

Conventional Definition of Test Statistic



With this definition, we are closer to the $\Delta\chi^2$ treatment

Calculating CLsb and CLb

We see that the "CL" definitions are related to the P-values (integrals of probability distributions)

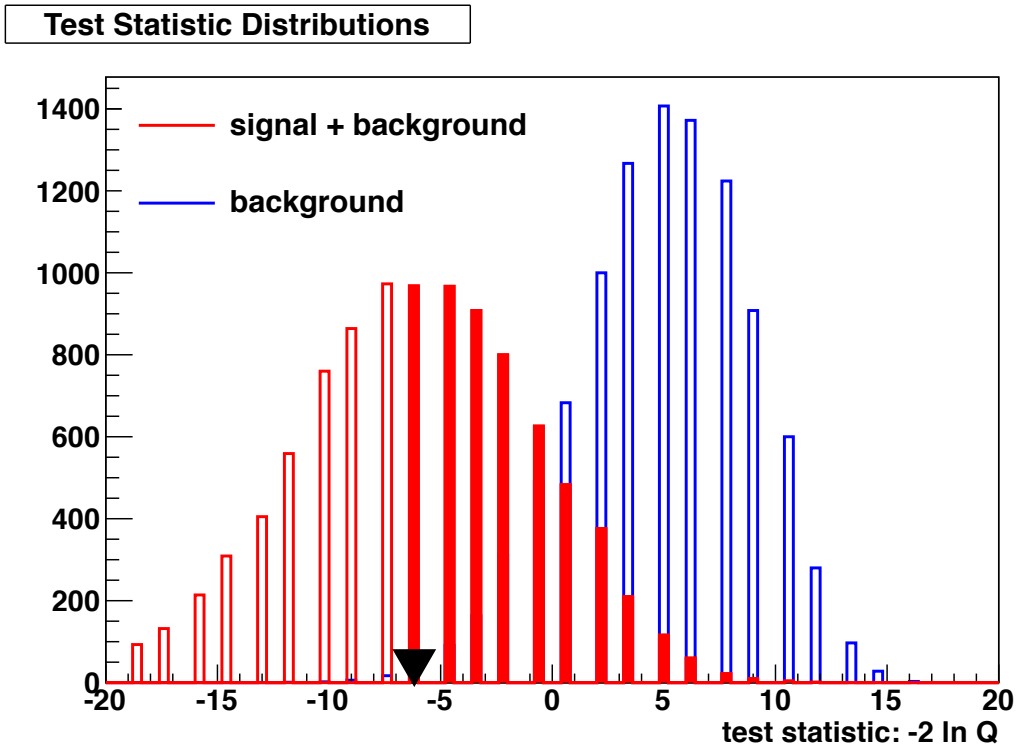
$$CL_{s+b} = P_{s+b}(Q \leq Q_{\rm obs})$$

Or, with a definition q = -2 ln Q,

$$CL_{s+b} = P_{s+b}(q \geq q_{\rm obs})$$

$$1 - CL_b = P_b(Q \geq Q_{\rm obs})$$

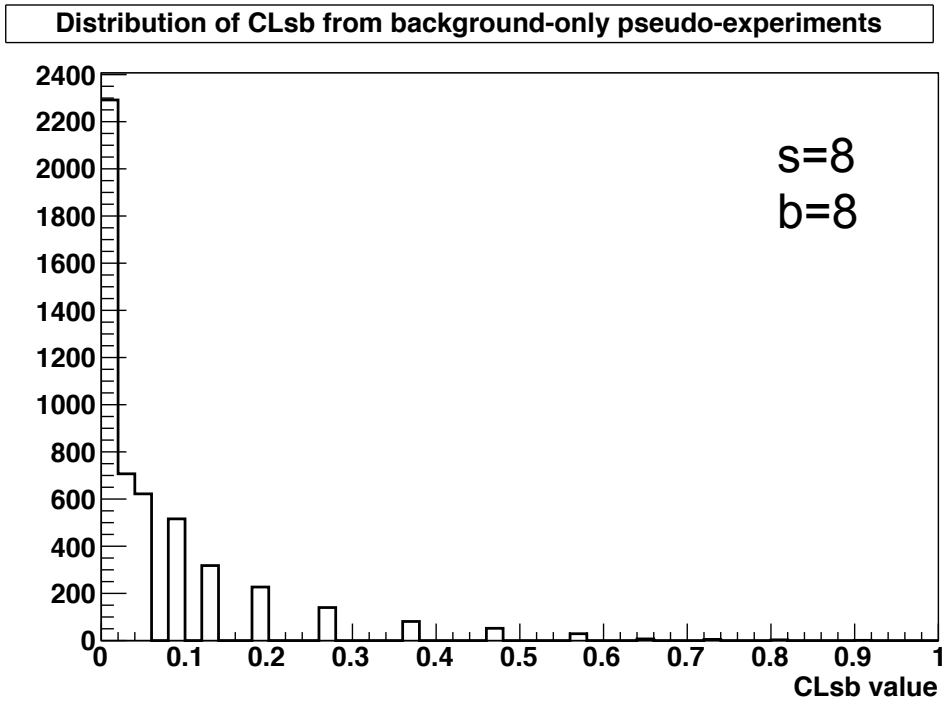
Note different comparison operator!

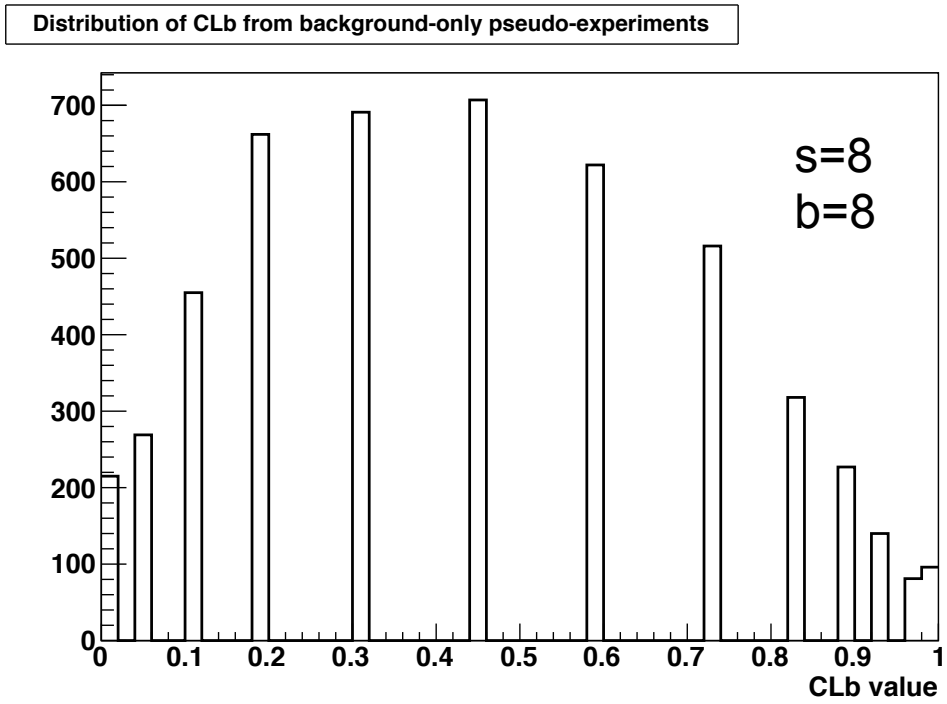


For this example (s=8, b=8, nobs=16), the integral is a simple sum:
Observed CL_{s+b} = 0.5551, and observed CL_b = 0.9926

Calculating <CLsb> and <CLb>

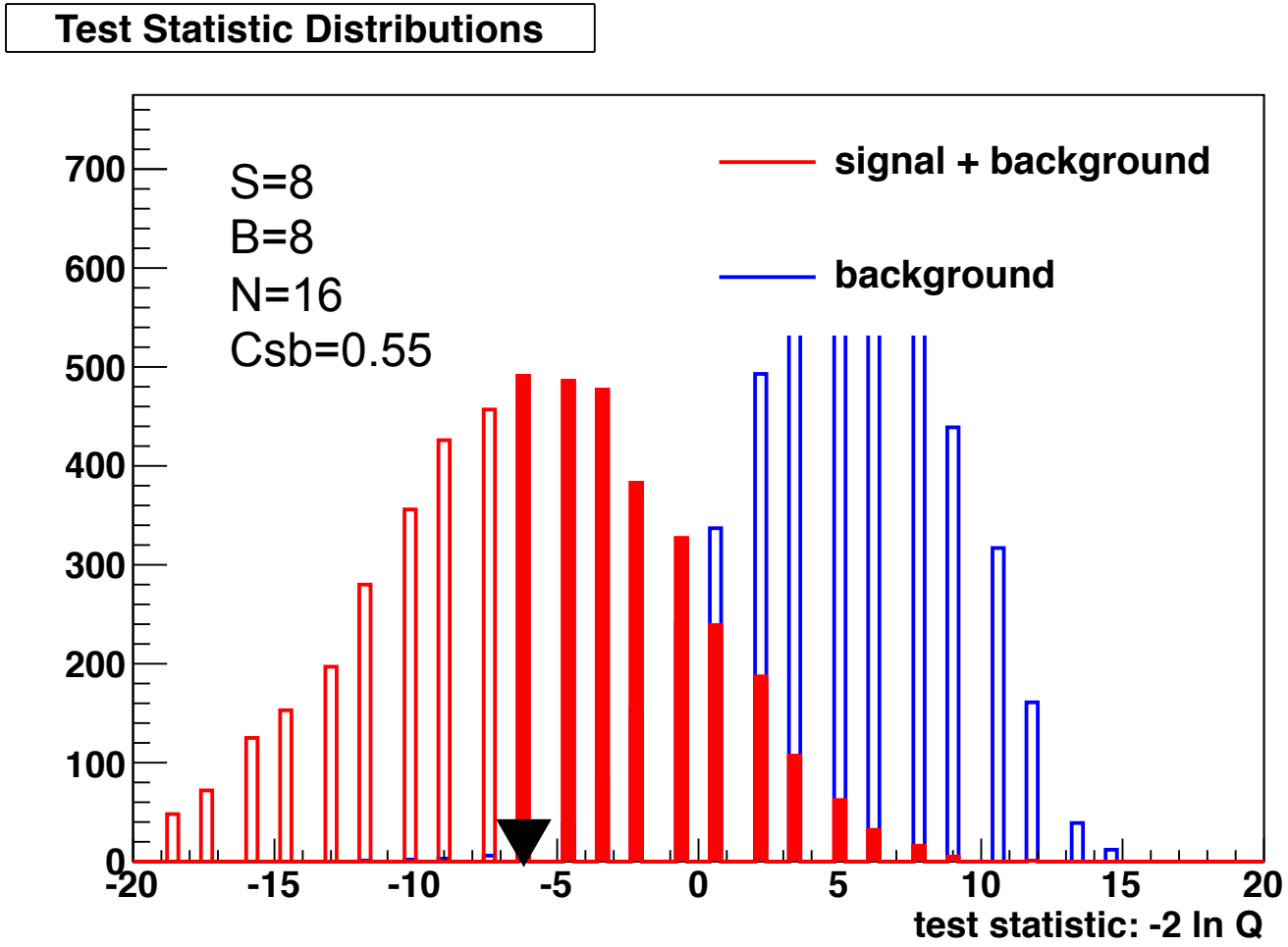
Expectation values of the p-values use CL_{s+b} and CL_b distributions from background-only pseudo-experiments ("expected limits")



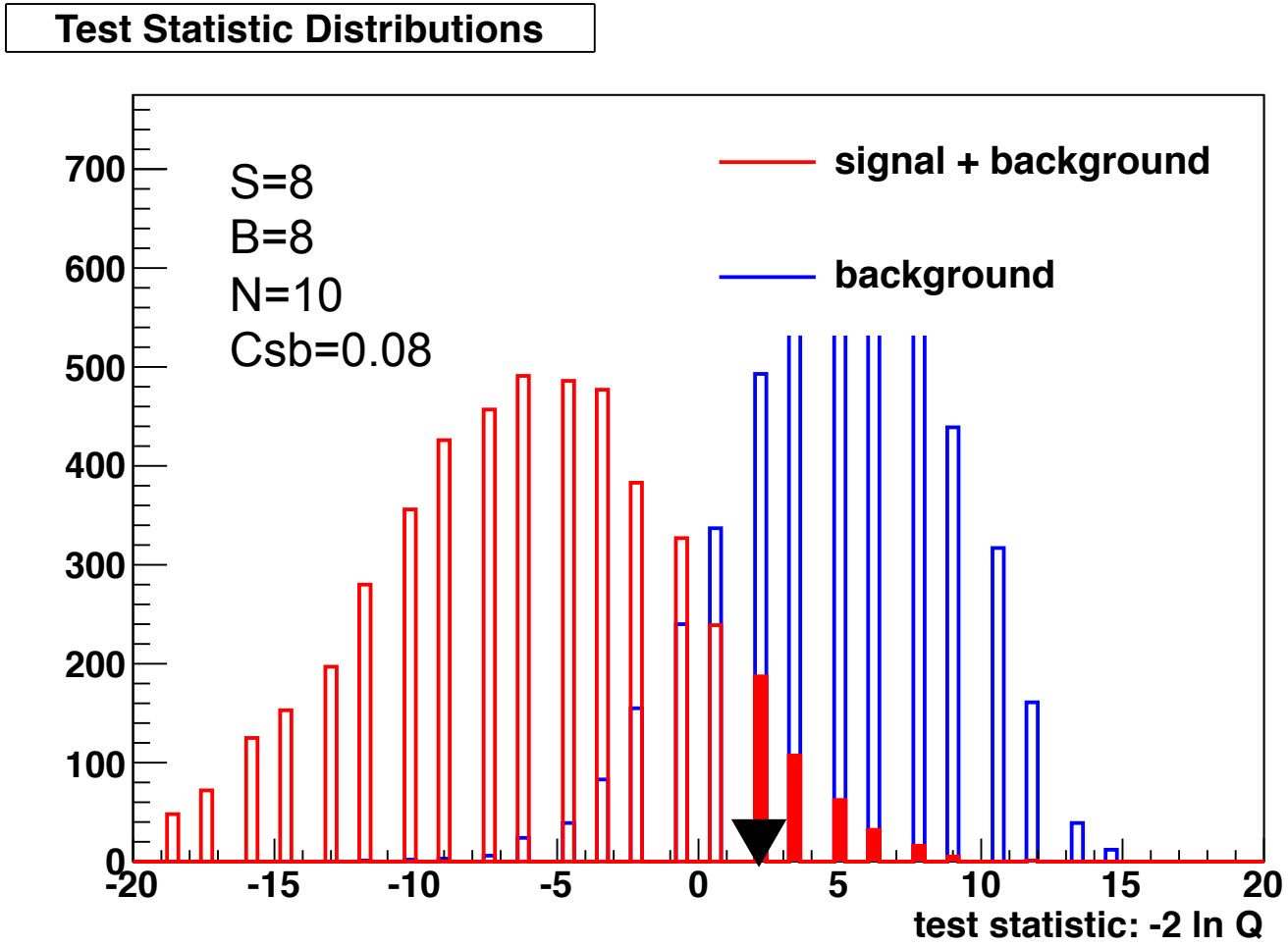


The 50th percentile of CL_b occurs in the bin with n_{obs}=8, the 16th at nobs=11, and the 84th at nobs=5. This similarity to our "sigma intuition" for cases when distributions are roughly Gaussian.

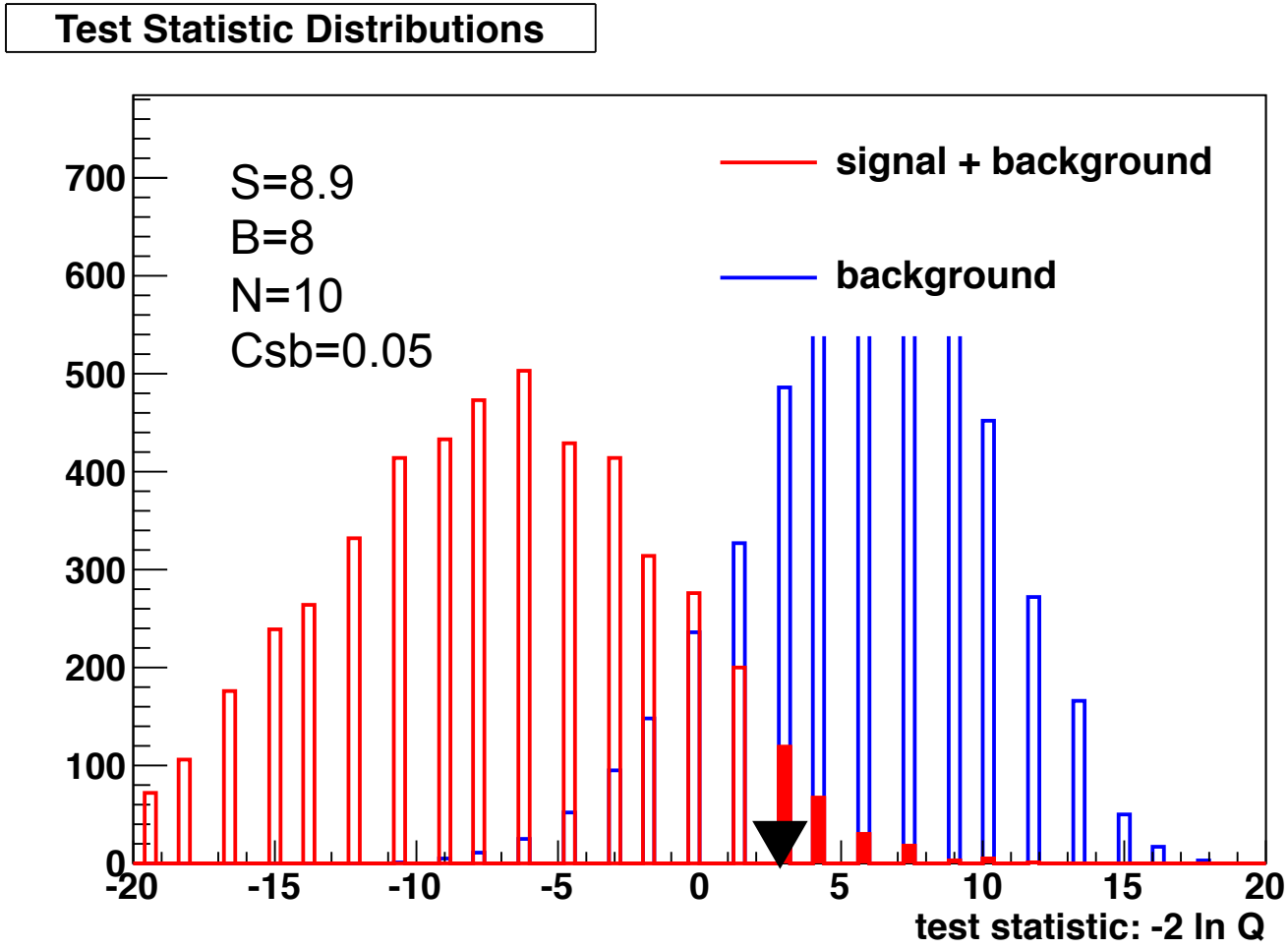
Example: Different Observations



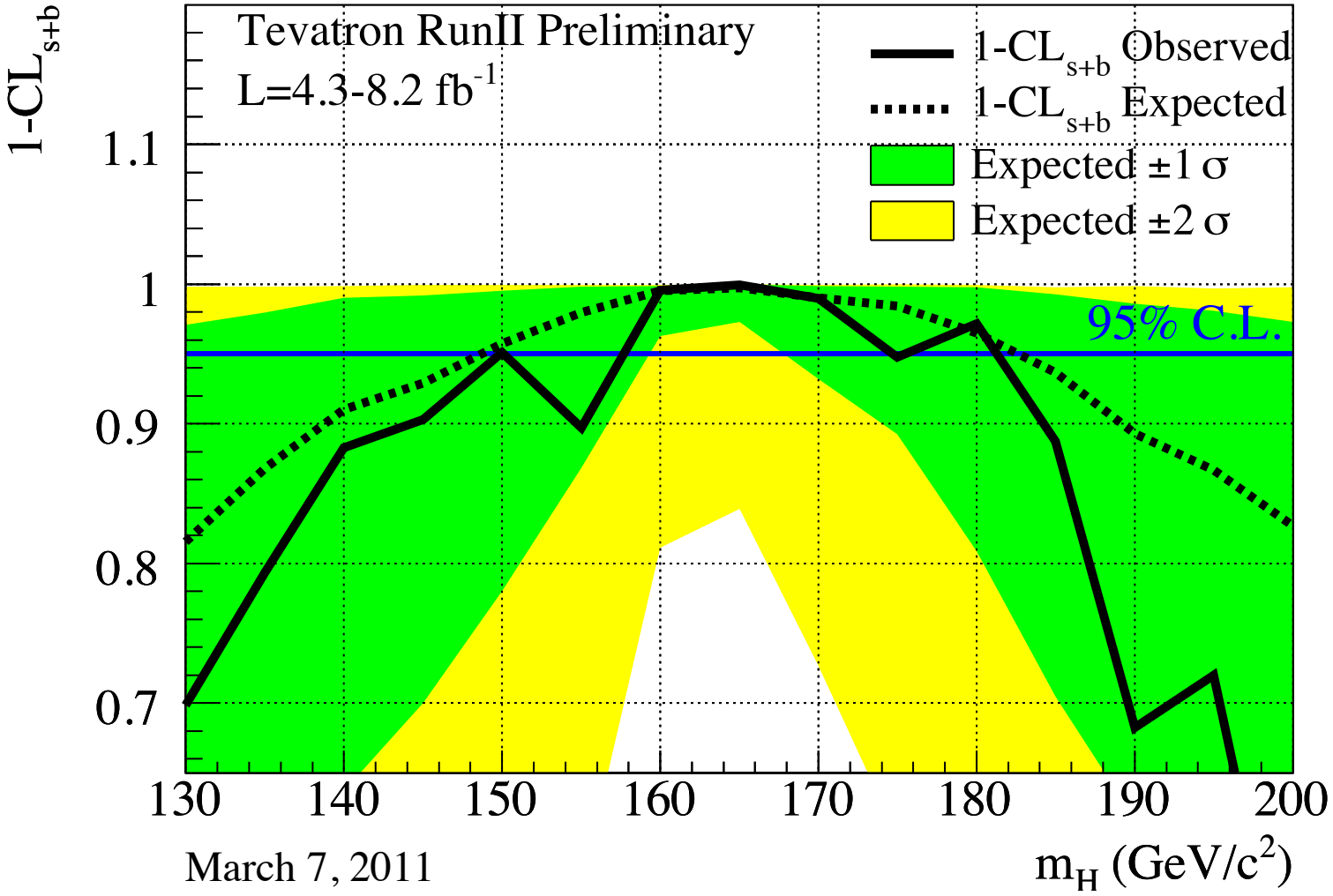
Example: Different Observations



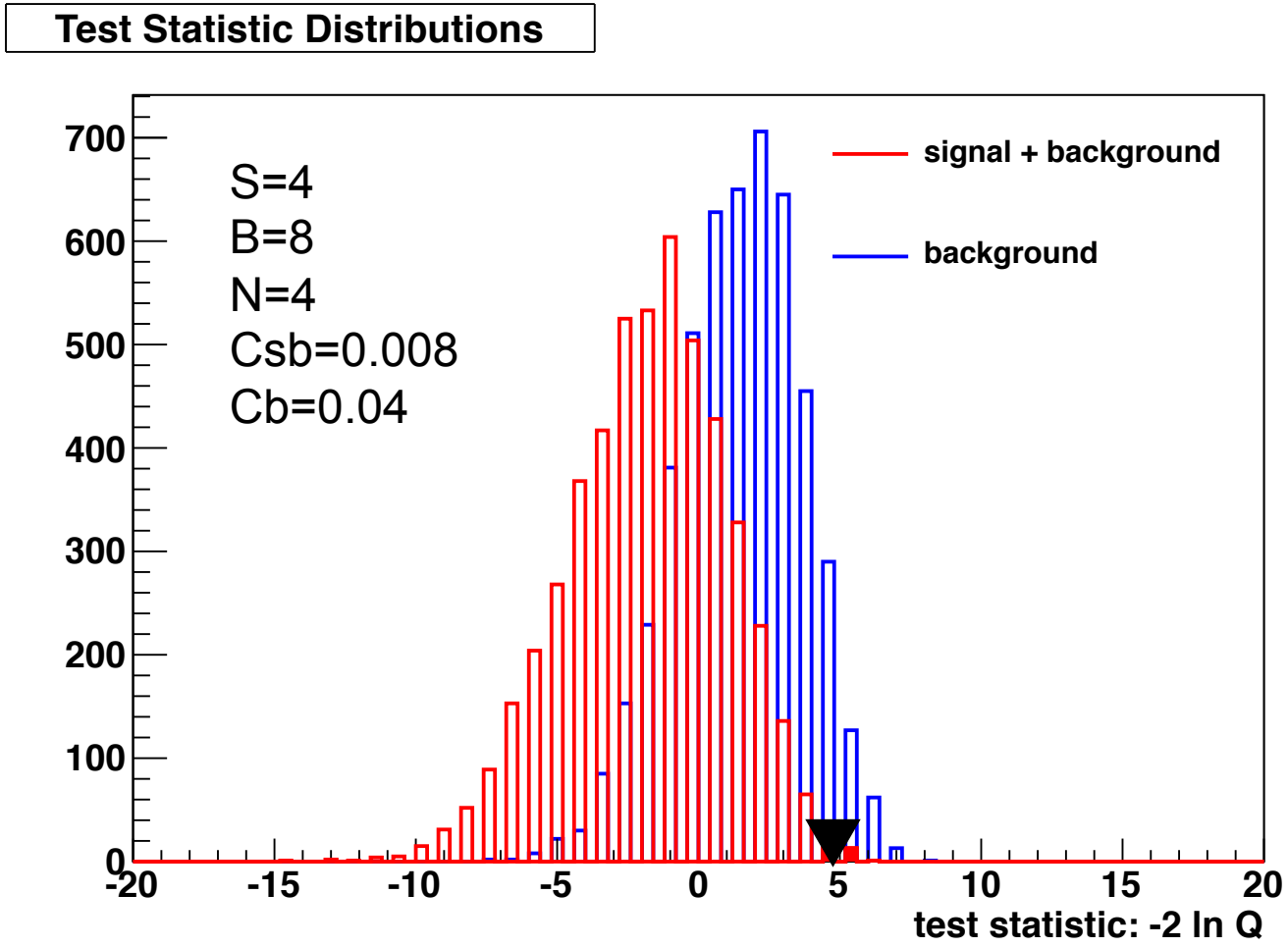
Example: Different Observations



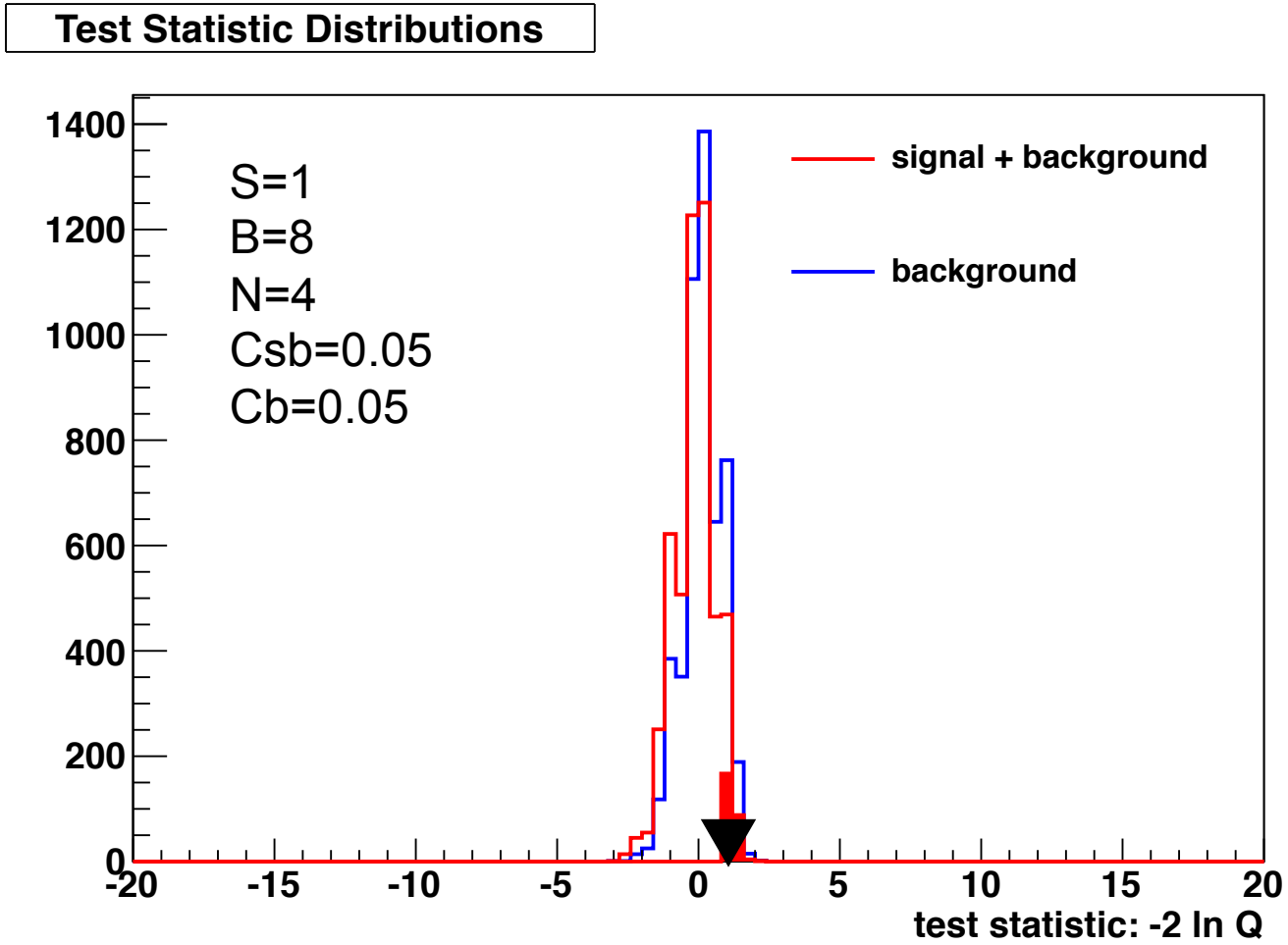
CLsb with varying s but fixed n



Example: Different Observations



Example: Different Observations



"Unnatural" exclusion due to the fact that we insist on the 95th percentile

This result is exceedingly unlikely, even for the background-only hypothesis!

Calculating CLs and <CLs>

- Problem: CL_{s+b} procedure excludes, with probability close to 5%, signal hypotheses to which we have no sensitivity
  - Can see this clearly when distributions are nearly degenerate and data fluctuate below background expectation
- When s is very small, we would like the prob to be 0%
- CLs method corrects for downward fluctuations by penalizing CLsb:

(LEP?/Tevatron version)

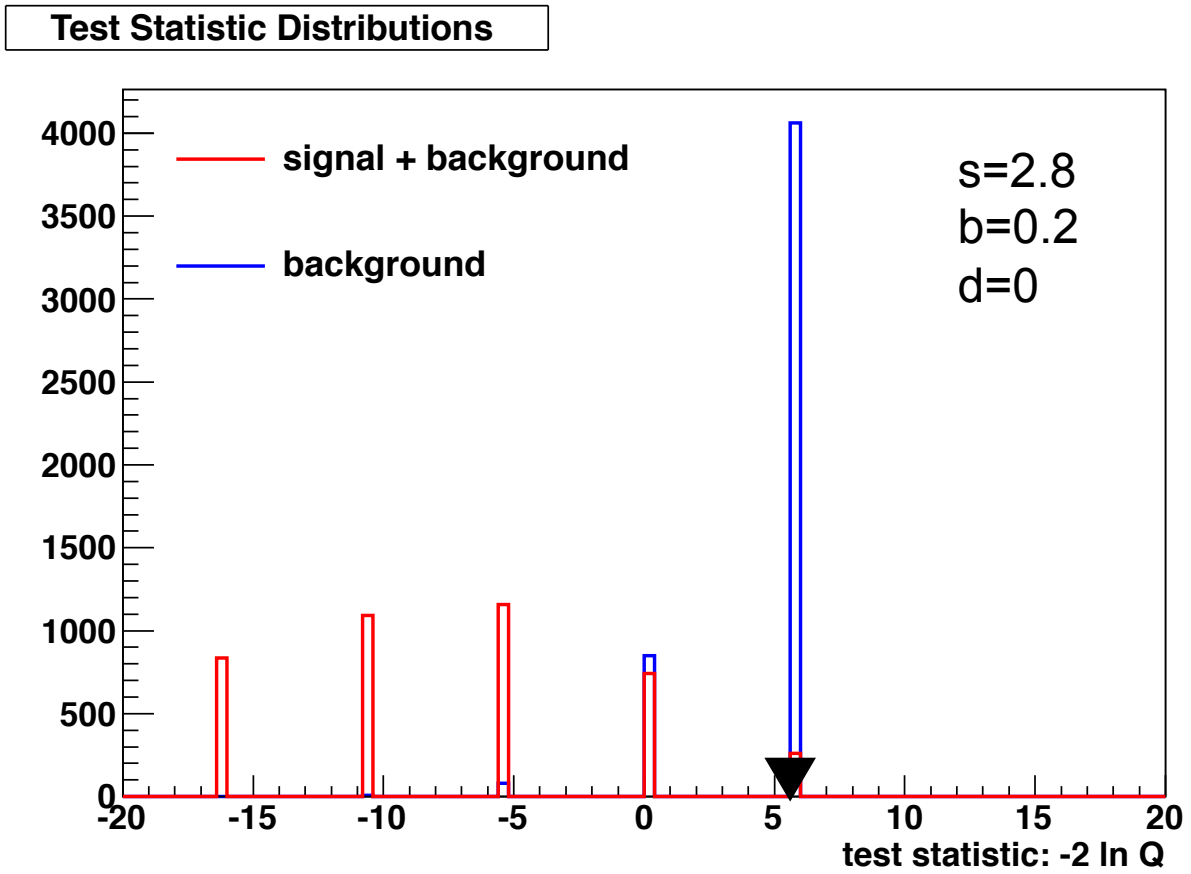
$$CL_s = \frac{P_{s+b}(q \geq q_{\rm obs})}{P_b(q \geq q_{\rm obs})}$$

- Has been described as an *ad hoc* method, but it accomplishes the goal of avoiding hyper-sensitivity
  - Technically it ensures overcoverage in a certain regime

Example: n_{obs}=0

Typical case with early searches – it even has specific rule of thumb

For n=0, the upper limit on s+b is 3.0 events, at 95% CL



ROOT's TLimit Class

Implementation of Junk's algorithm for frequentist CL calculation

In a form with no uncertainties, it is extremely straightforward to use

```
TLimit* theLimit = new TLimit();
TConfidenceLevel* theCL = theLimit->ComputeLimit(s, b, data);
cout << theCL->CLsb() << endl
     << theCL->CLb() << endl
     << theCL->GetExpectedCLsb_b() << endl
     << theCL->GetExpectedCLb_b() << endl;
```

```
For s=2.8, b=0.2, data=0:
Observed CLsb=0.0497463
Observed CLb=0
Expected CLsb_b=0.0497463
Expected CLb_b=0.81806
```

Systematic Uncertainties

- The spread in the p-values so far has been due to statistical distributions (Poisson distribution about mean)
- What effect do additional systematic uncertainties have on the result?

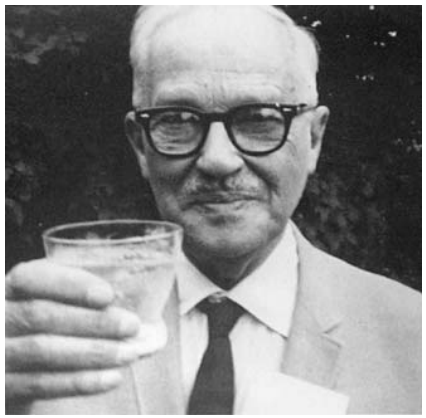
- First attempt is to integrate probability over the presumed values (usually Gaussian distribution) when deriving the test statistic distribution
  - Even if we use Monte Carlo, it is still an integration!

"Semi-Bayesian" Criticism

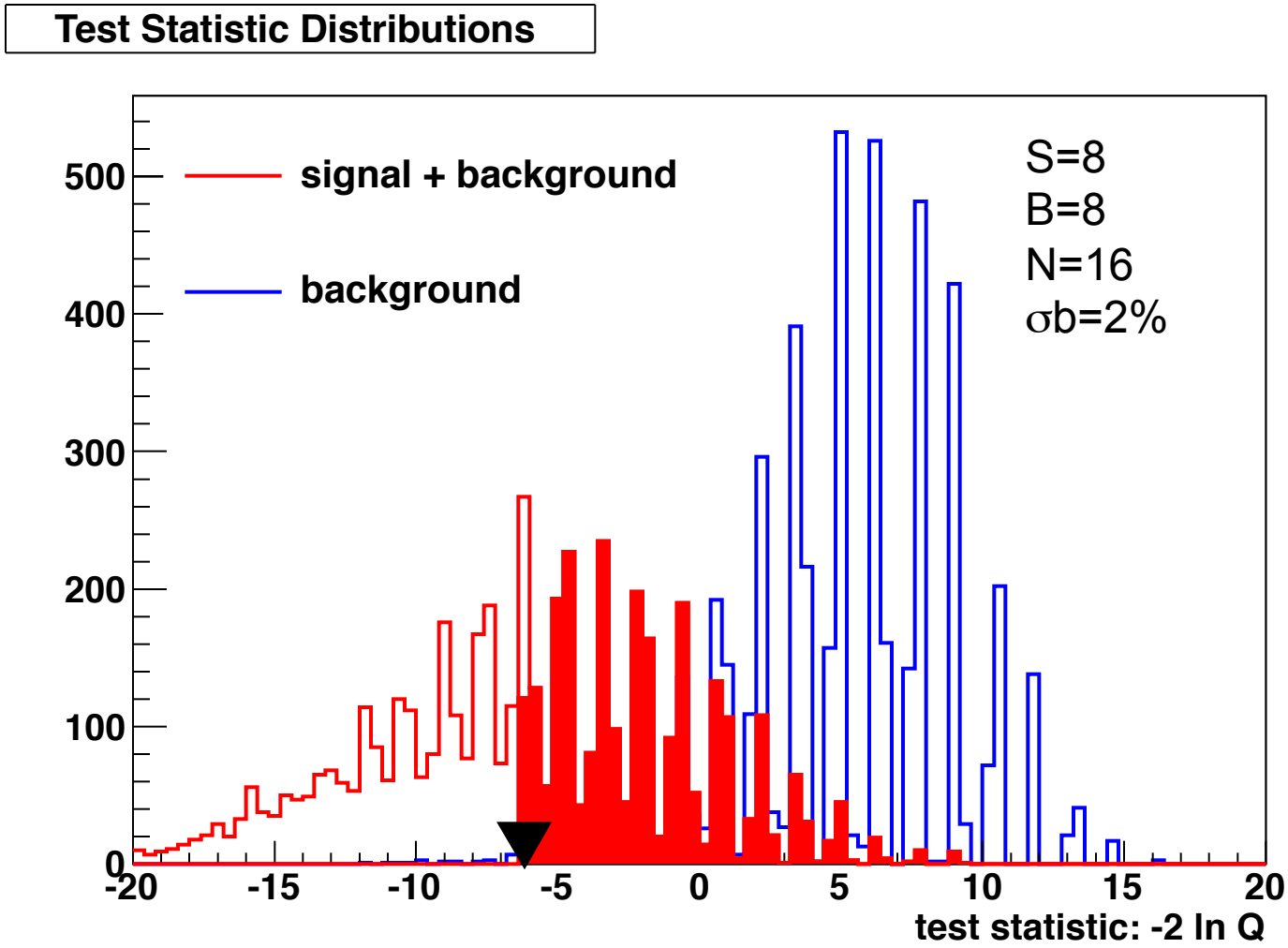
- This approach is effectively a Bayesian approach, since the presumed distribution is a Bayesian prior (Gaussian)
- Known as "marginalization" of the likelihood
- At LEP, all uncertainties were treated in this way to give a hybrid Bayesian-frequentist result

- This mixed approach causes some angst if you are a card-carrying frequentist

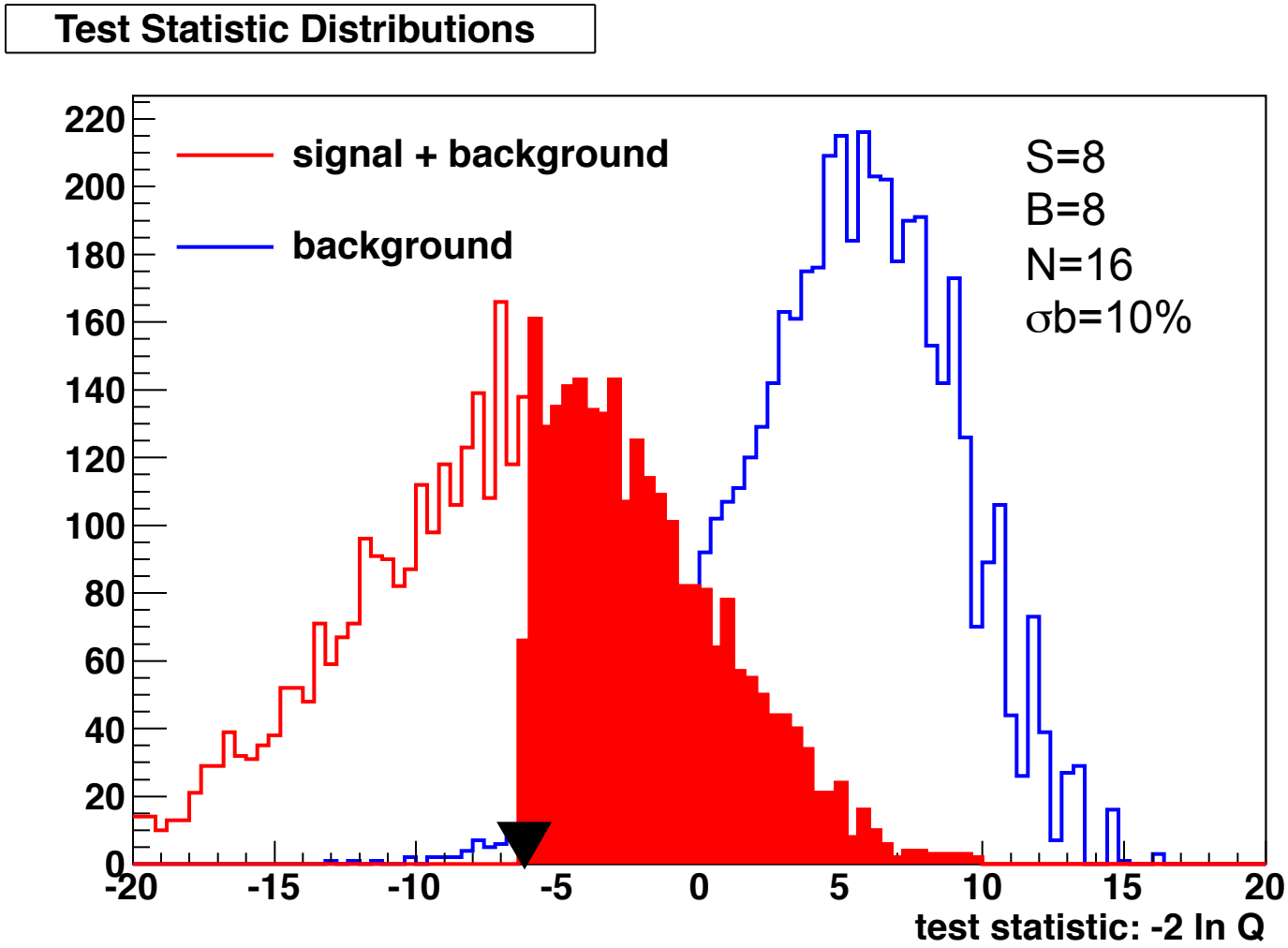
Jerzy Neyman
Frequentist
UC Berkeley
1894-1981



Distributions with Uncertainty on b

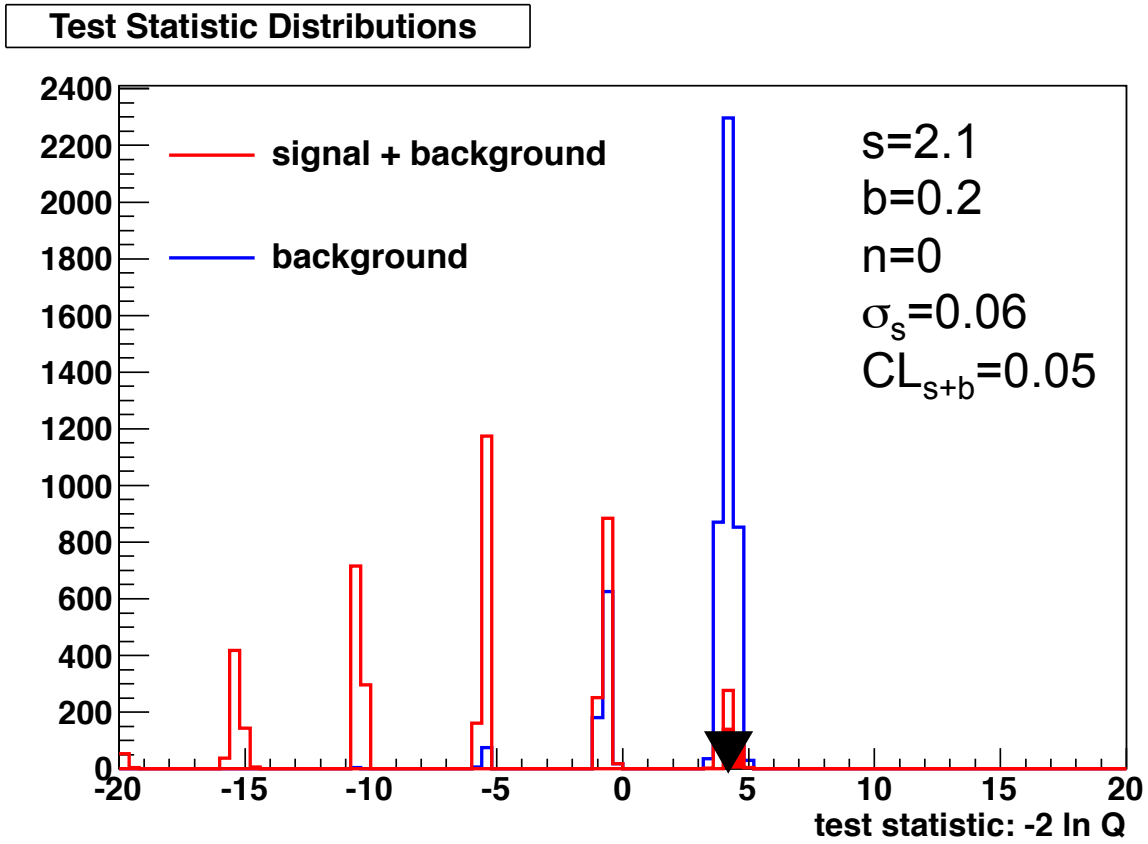


Distributions with Uncertainty on b



Example: n_{obs}=0

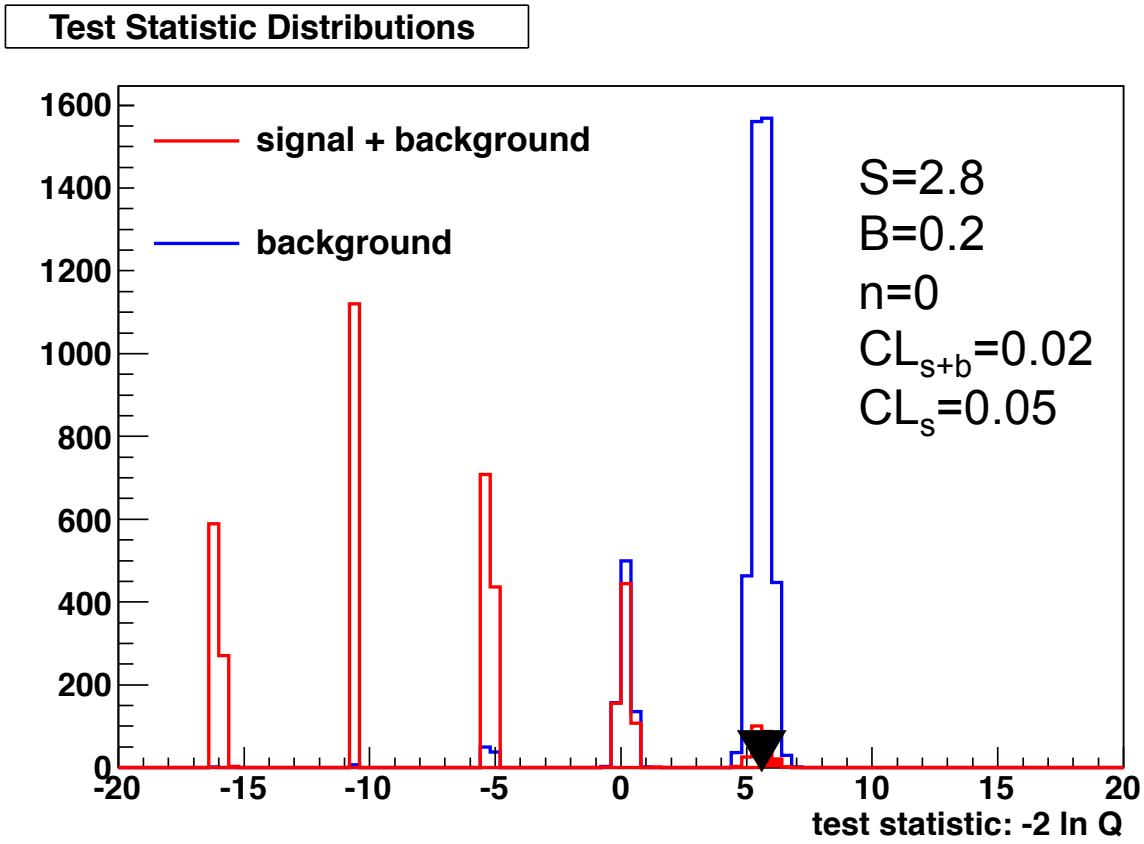
Effect of signal uncertainty is to smear individual peaks in distribution, so that only half of the last peak contributes to CL_{s+b} and CL_b



Uncertainty has improved the limit s_{up} from 2.8 to 2.1 events!

Example: n_{obs}=0

But the CLs method fixes the problem, restores limit to 3.0



We will see that this is one motivation for using CLs in final results

mclimit.f Program at Tevatron

- Used for most CDF results, following Junk's algorithm
- Attempts to evade the hybrid Bayesian treatment of some nuisance parameters by going back to source of the uncertainty and treating in frequentist way
  - Background uncertainties resulting from limited MC statistics

- Tevatron results use CLs method to set limits
- Test statistic has changed to the profile likelihood

Introducing Profile Likelihood

- Goal: incorporate nuisance parameters directly in the likelihood function, so that marginalization is no longer necessary for making the test statistic distributions
- Define signal strength μ and nuisance parameter vector θ
  - These are free parameters (dependent variables) for likelihood
  - Signal strength μ can even be negative, if total signal = $\mu s < 0$

$L(\mu, \theta)$	the likelihood function (from Poisson probability)
$L(\mu, \hat{\hat{\theta}}(\mu))$	the likelihood function, using most likely value of nuisance parameters θ, given that value of μ
$L(\hat{\mu}, \hat{\theta}(\mu))$	the likelihood function at its extremum μ and θ are values that are most likely, given experiment

Profile Likelihood Implies Fitting

- Since these functions are evaluated using $\hat{\hat{\theta}}$ that maximizes the likelihood, the mathematics is the same as for ML fit of those parameters in an experiment!
  - In fact, Rolke introduced this (2005) by noting similarities to venerable CERNLIB fitting program called MINOS
- ROOT's dedicated RooFit library is designed for this case, and is used as the basis for the RooStats profile likelihood
- Presence of nuisance parameters introduces loss of information about μ, broadening likelihood function

- Distinction: μ is the signal strength parameter under test, while $\hat{\mu}$ is the best fit to data

Simple Definition of Profile Likelihood

Most general definition of profile likelihood ratio (no restrictions on μ)

$$\lambda(\mu|\theta) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}$$

Unlike Tevatron, both L are L_{s+b}

Test statistic follows same definition as the LEP/Tevatron framework (note: no functional dependence on θ)

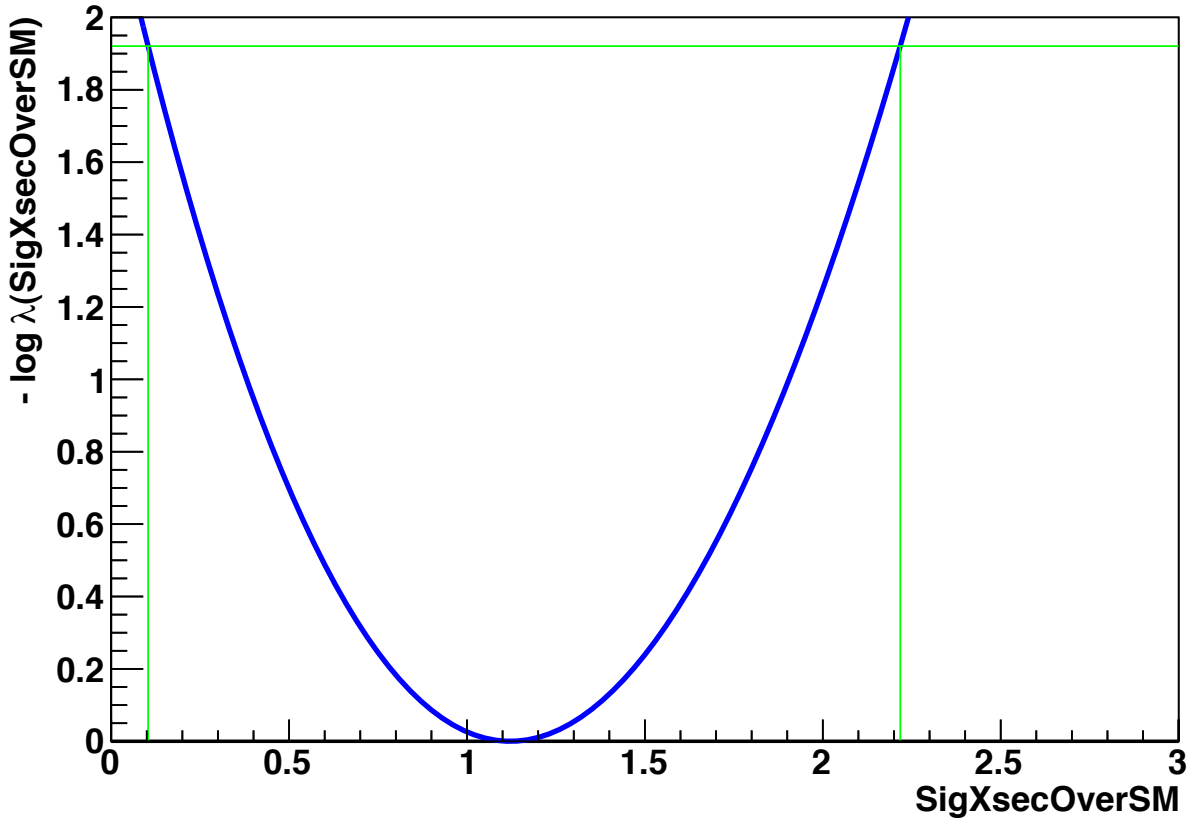
$$t_\mu = -2 \ln \lambda(\mu)$$

If we are interested in correct upper limits, then $\hat{\mu} > \mu$ is not relevant, and we do not count that region

$$q_\mu = \begin{cases} -2 \ln \lambda(\mu), & \hat{\mu} \leq \mu \\ 0, & \hat{\mu} > \mu \end{cases}$$

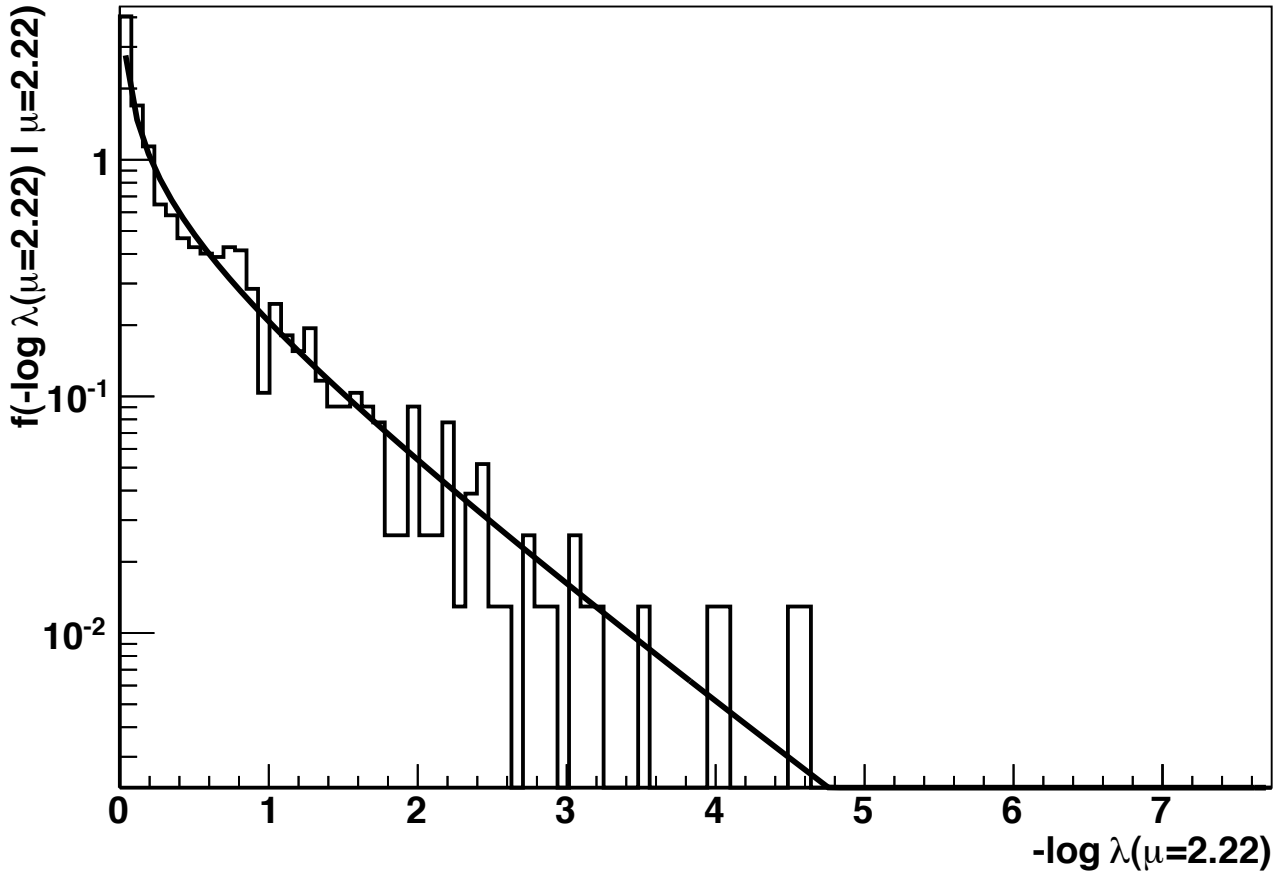
Profile Likelihood Ratio

Sample of 230 data events (30 signal) gives smooth symmetric function



Constructing Test Statistic Distributions

Shape is different from LEP/Tevatron – there is no balancing point at 0

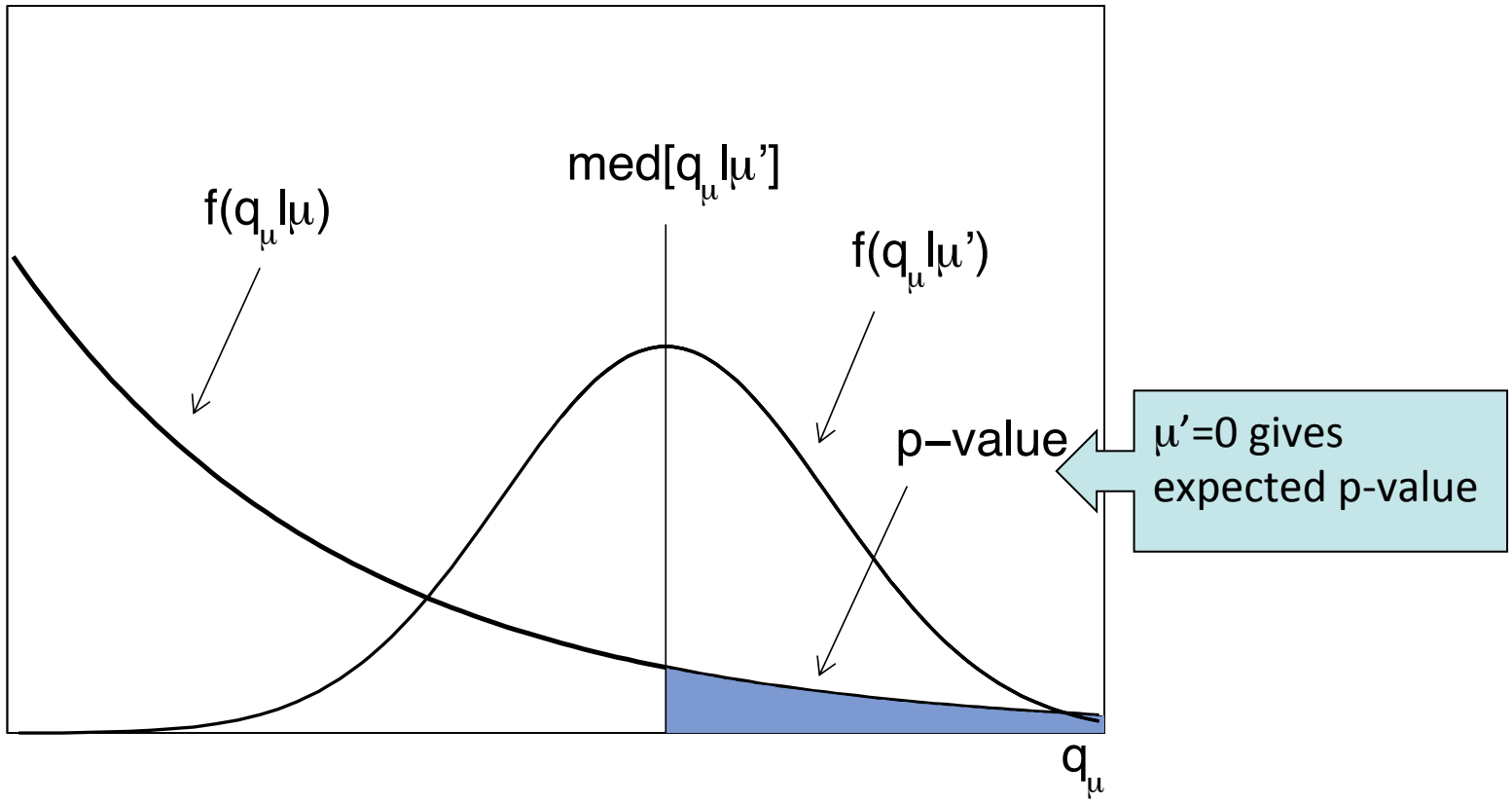


Observed p-value (and limits) still calculated by integrating the tail

Calculating expected p-values

Expected p-values require an ensemble of events with a given μ'

CCGV found a shortcut: use a single representative dataset ("Asimov" dataset) corresponding to mean number of events expected



Solution #1: CLs Method

- Designed overcoverage so limit does not benefit from downward fluctuations

(ATLAS/LHC version)

$$CL_s = \frac{P_{s+b}(q \geq q_{\rm obs})}{1 - P_b(q \leq q_{\rm obs})}$$

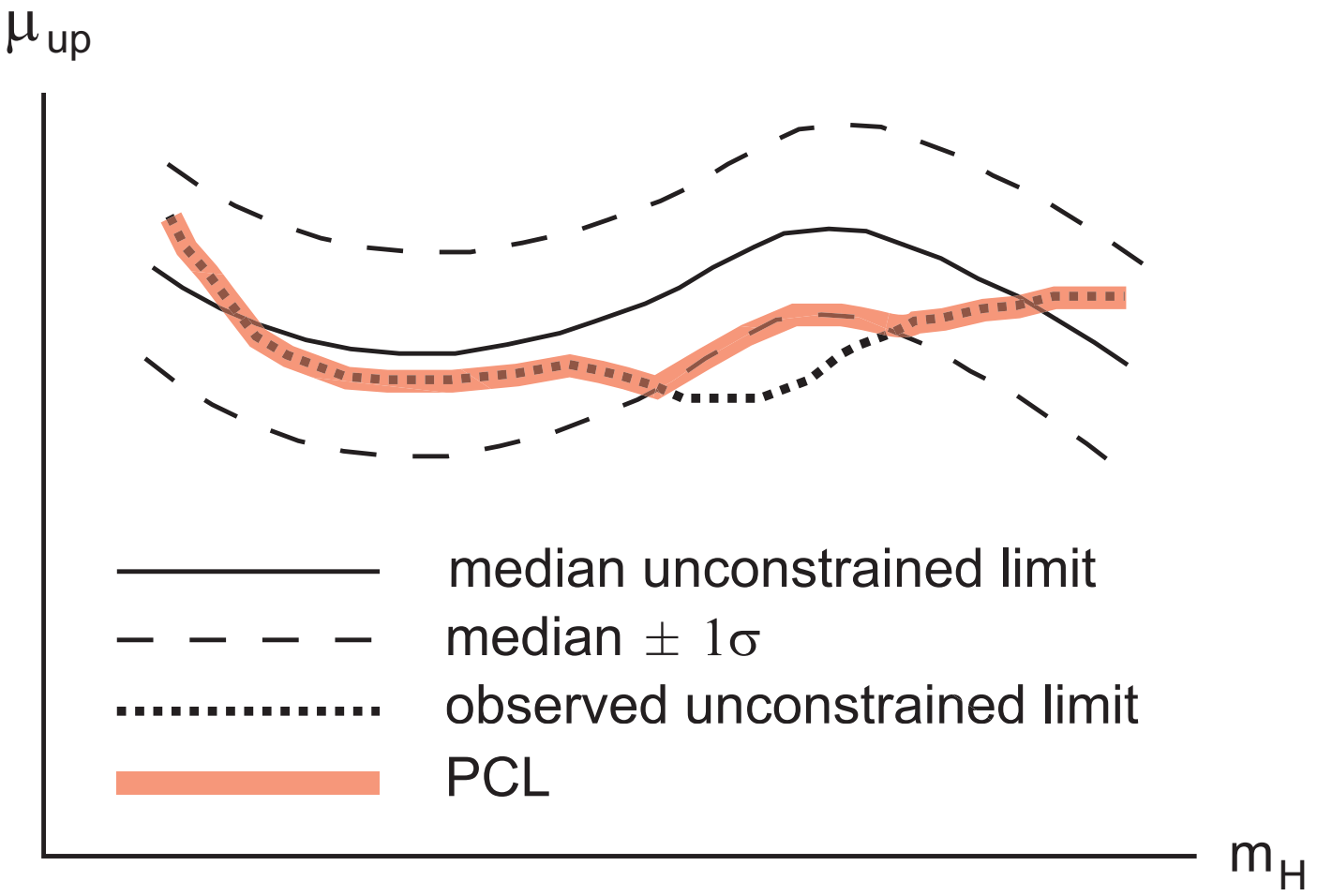
- The CLb penalty smoothly decreases with decreasing fluctuations, but in a non-intuitive way
- Main objection is varying (inconsistent) coverage, a key concern for frequentist methods

Solution #2: Power-Constrained Limits

- Proposed by Cowan, Cranmer, Gross, Vitells as a way to make this suppression more concrete and controlled

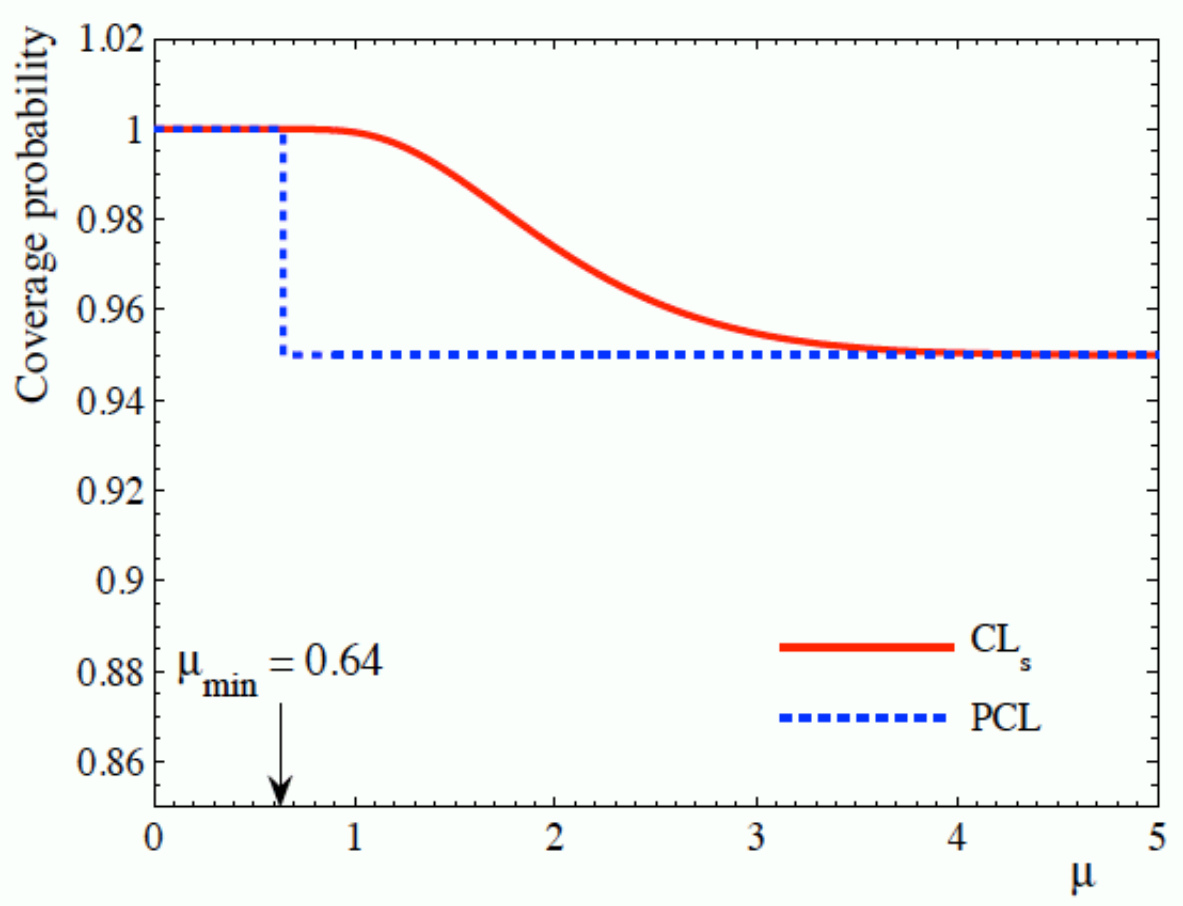
- Define the power (sensitivity) of a search to a certain value of signal strength μ
  - Note: the power is always greater than the significance level 5%
- If this power is less than some cut-off value for a value m, do not allow that m to be excluded
- Practically speaking, if we take a cut-off value of 16% (1σ), then we never allow the limit line to pass below that well-defined (but still arbitrary!) value

Upper Limit Results with PCL (16 % cutoff)



Coverage Comparison

Remember: the frequentist goal is 95% coverage probability

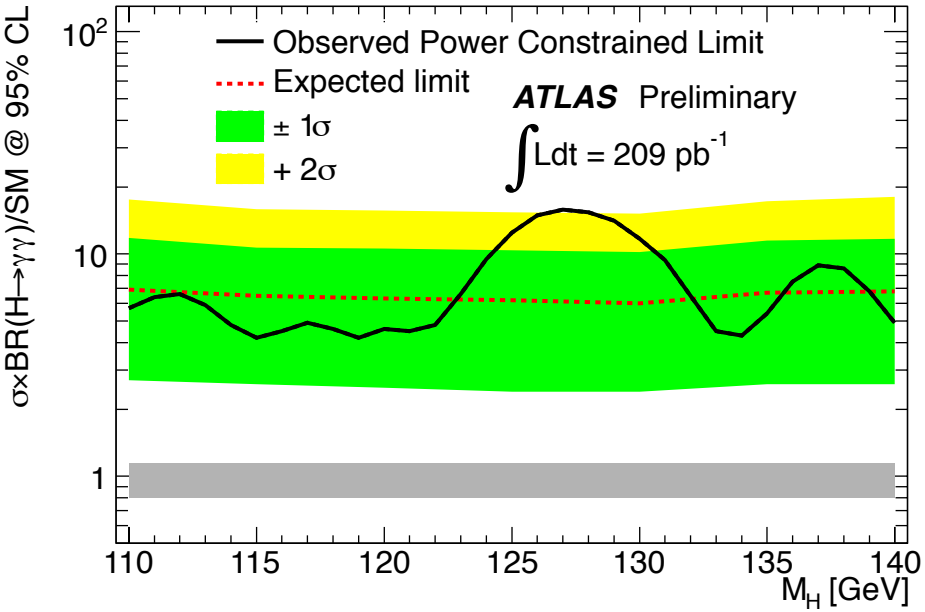


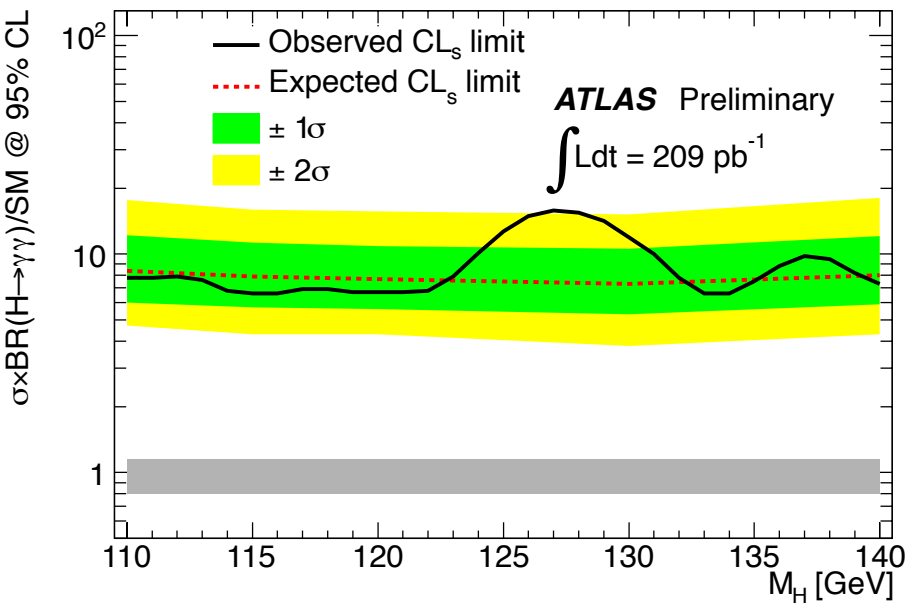
PCL over-covers by construction when we hit the 16% cutoff value

Comparison of Techniques

ATLAS SM Higgs limits in the diphoton channel give both results for a direct comparison of the CLs and PCL methods

ATLAS-CONF-2011-085

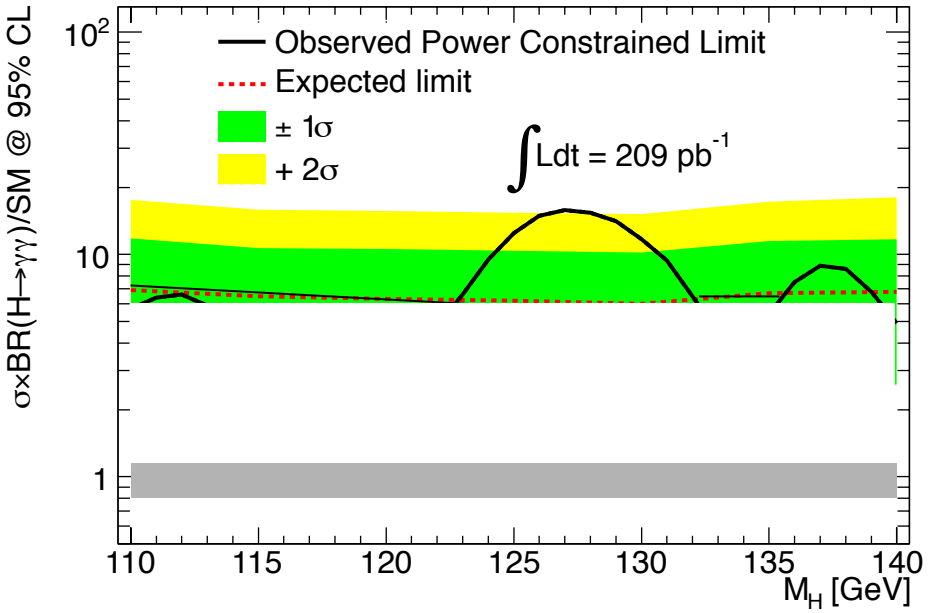




In this case, the PCL cutoff does not kick in because there is no 2σ deficit anywhere in the analysis region of interest

Upper Limit Results with PCL (50% cutoff)

ATLAS has recently decided to weaken the PCL arbitrarily by using a power of 50% instead of 16% (despite arguments in CCGV paper)



Example of new approach: not an ATLAS approved result!

Practically speaking, this means ATLAS results will never benefit from any downward fluctuation in data, below expectation

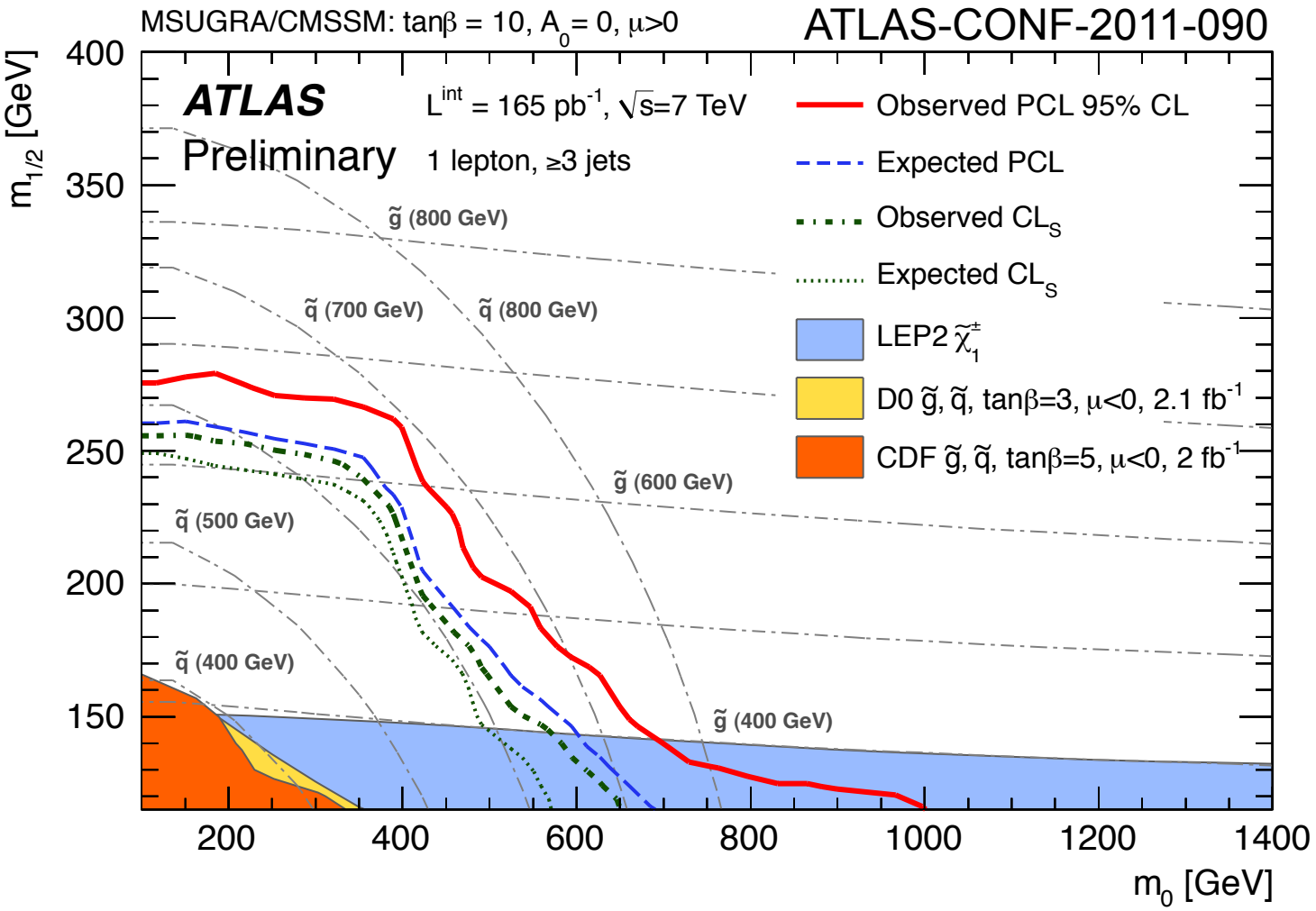
How is this similar to the CLs overcoverage on p.37?

Implementation in RooStats

- RooStats routines grow out of research into these methods, based on the RooFit classes and math libs
- Tutorials in $ROOTSYS/tutorials/roostats:
  - StandardProfileLikelihoodDemo
  - StandardTestStatDistributionDemo
  - rs101_limitexample
  - OneSidedFrequentistUpperLimitWithBands (if you are brave…)

- Still checking some of these results against code developed from scratch by many users

Comparison of Techniques



Common Agreement with CMS

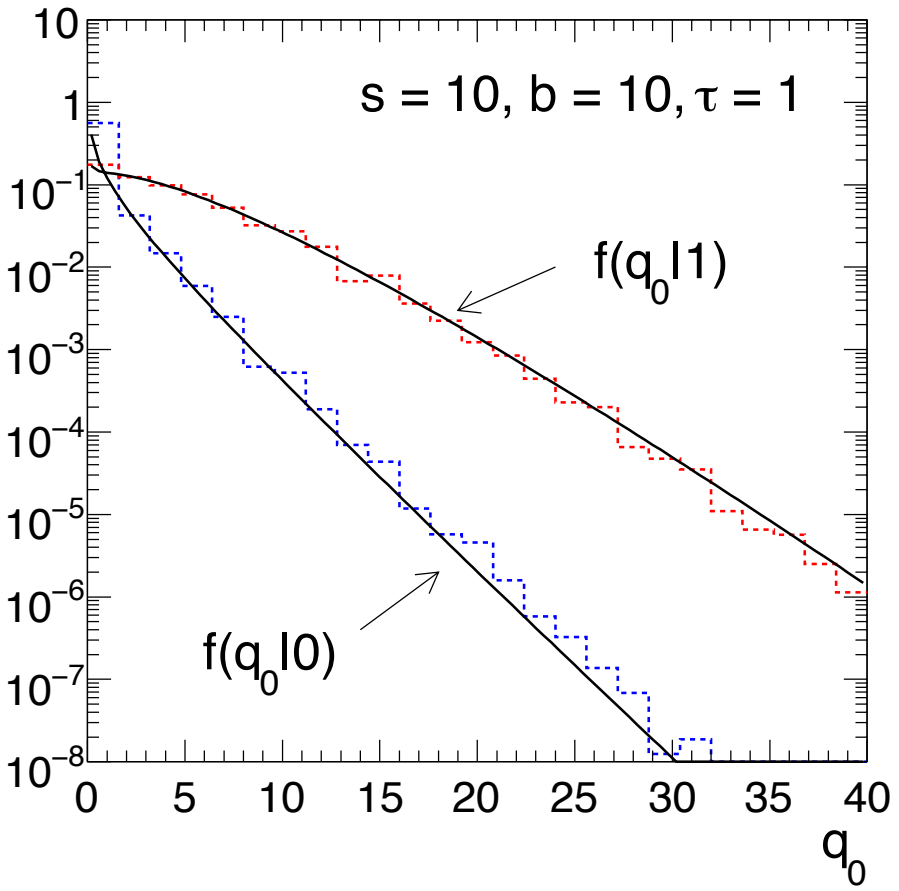
- CMS choose to use different methods than ATLAS
  - CLs method instead of Power-Constrained Limits
  - Greater use of Bayesian methods
- ATLAS agrees to show CLs limits as a common point of comparison with CMS and Tevatron (and LEP)

- Debate continues in dedicated workshops, with professional statisticians appalled by CLs method

Asymptotic Behavior of Distributions

- Asymptotic techniques work with χ^2 distributions by evaluating the integral functions of the tails

- Fails for small number of events (early searches), but we are rapidly passing out of that regime for many ATLAS searches
  - Fall back on toy Monte Carlo when this fails



Asymptotic Formulas for q_μ

- Distribution of f is a half-chi-square distribution

$$f(q_\mu|\mu) = \frac{1}{2}\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}e^{-q_\mu/2}$$

- With cumulative distribution given by

$$F(q_\mu|\mu) = \Phi\left(\sqrt{q_\mu}\right)$$

- The p-value (tail integral) can be read off from F

$$p_\mu = 1 - F(q_\mu|\mu) = 1 - \Phi(\sqrt{q_\mu})$$

- By evaluating Φ, we can find the value of q_μ such that

$$p_{\mu 95} = 0.05$$
$$q_{\mu 95} = 1.64^2$$

- The problem has been reduced to calculating q for various strength parameters (much less CPU time)

Summary

- ATLAS frequentist standard method uses a profile likelihood ratio to deal with uncertainties
  - Some analyses will continue with Bayesian results for updates
- ATLAS Power-Constrained Limits take "sensitivity power" as input, and this was 16% (1σ), now 50% (all downward)
- ATLAS/CMS now agree to show results of CLs method for the sake of comparison between experiments
  - ATLAS standard tends to be more aggressive (less designed overcoverage) than the CLs method used by CMS/Tevatron/LEP

- Recommend RooStats tools, which have implemented these techniques (some parts still being commissioned)

Public References

- *CLs method*: A.L. Read, J. Phys. G **28**, 2693 (2002)
- *Semi-Bayesian:* T. Junk, Nucl. Instrum. Methods A **434** (1999) 435
- *Tevatron:* http://www-cdf.fnal.gov/~trj/mclimit/mclimit_csm.pdf
- *Profile Likelihood and asymptotic behavior:* Rolke et al., Nucl. Instrum. Methods A 551 (2005) 493; Cowan et al., Eur. Phys. J. C (2011) **71 :** 1554
- *Power-Constrained Limits:* Cowan et al., **arXiv:1105.3166v1**