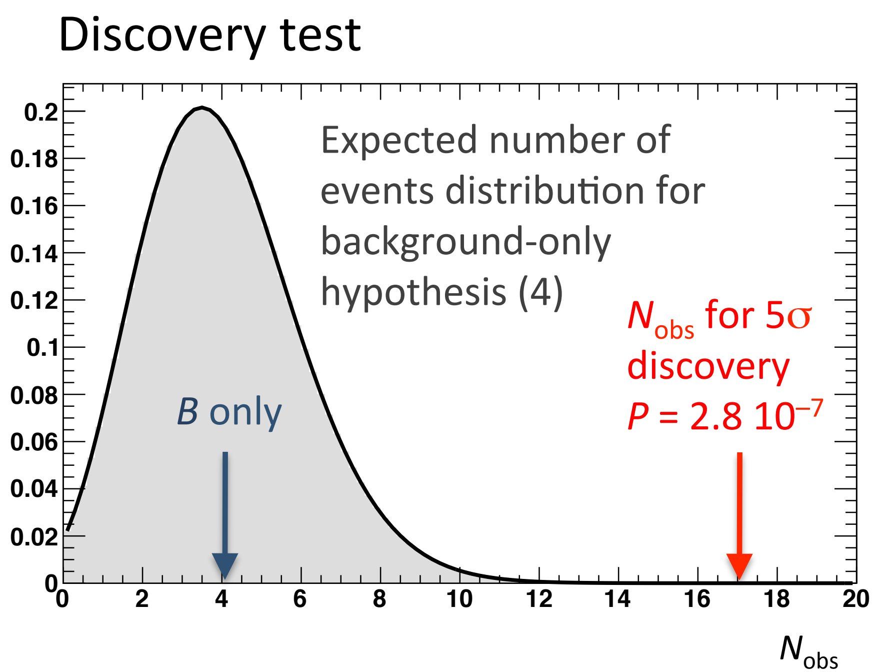
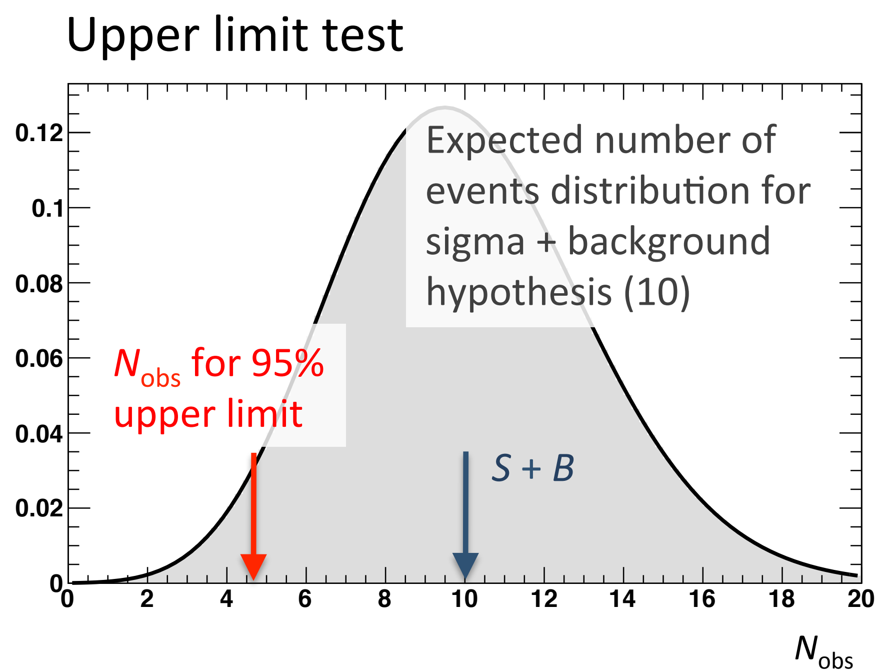
Easy Limit Statistics

Andreas Hoecker

CAT Physics, Mar 25, 2011

The Goals





- In a *discovery test* one wants to measure the probability of an upward fluctuation for background only
- In an *upper limit test*, one wants to measure the probability of a downward fluctuation of signal + background

Funny Parameters

L	$\rightarrow$	Maximum of likelihood
$L(\mu S+B)$	$\rightarrow$	μ is "signal strength" parameter
$L(\mu S+B,\theta)$	$\rightarrow$	θ is "nuisance parameter" (e.g., errors on S and B)
$L(\mu S+B,\hat{\theta})$	$\rightarrow$	"hat" (or "hat+hat") means parameter floats for max. L
$L(\mu,\hat{\hat{\theta}})$	$\rightarrow$	Short-hand version: only $\hat{\hat{\theta}}$ floats
$L(\hat{\mu},\hat{\theta})$	$\rightarrow$	$\hat{\mu}$ and $\hat{\theta}$ float ($\hat{\theta}$ usually different from $\hat{\hat{\theta}}$)
$\dfrac{L(\mu,\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})}$	$\rightarrow$	Nominator and denominator independently maximised

Likelihood Function

- L can be very simple, eg, for a counting experiment:
  - Number counting:

$$L(N_{\text{obs}},\mu) = e^{-\mu S - B}(\mu S + B)^{N_{\text{obs}}} \frac{1}{N_{\text{obs}}\,!}$$

  - Number counting with background uncertainty (nuisance parameter)

$$L(N_{\text{obs}},\mu,\theta) = e^{-\mu S - B}(\mu S + \theta B)^{N_{\text{obs}}} \frac{1}{N_{\text{obs}}\,!} \times \text{Gauss}(\theta - 1, \sigma_\theta)$$

  - Signal prediction (expected numbers of events) usually also has nuisance parameters: cross section, selection efficiency, luminosity uncertainties, etc.

Likelihood Function

- *L* can be very simple, eg, for a counting experiment:
- *L* can also be complex
  - Several distinct signal and background contributions
  - Several discriminating variables (use product of PDFs)
  - Some variables may have event-by-event scaling factors
  - Signal, background and PDF shape parameters may be floating
  - Physical parameters may be number of events but also signal properties
  - Likelihood may be split into categories with different subpopulations of events with common and non-common parameters
- Most ATLAS search analyses so far dealt with counting likelihoods in presence of signal cross section and efficiency uncertainties, as well as background abundance uncertainties

One-sided Test Statistics

$$\tilde{\lambda}(\mu) = \begin{Bmatrix} \dfrac{L(\mu,\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})}, & 0 \le \hat{\mu} \le \mu \\ \dfrac{L(\mu,\hat{\hat{\theta}})}{L(0,\hat{\theta})}, & \hat{\mu} < 0 \end{Bmatrix} = \frac{L(\mu,\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})}, \qquad 0 \le \hat{\mu} \le \mu$$

$$\tilde{q}_\mu = \begin{cases} -2\ln\tilde{\lambda}(\mu), & 0 \le \hat{\mu} \le \mu \\ 0, & \hat{\mu} > \mu \end{cases}$$

- Large values of $\tilde{q}_\mu$ correspond to increasing disagreement between data and hypothesis μ
- This test statistics behaves asymptotically similar to a χ^2 for large data samples and Gaussian nuisance parameters

One-sided Test Statistics

$$\tilde{\lambda}(\mu) = \left\{ \begin{array}{ll} \dfrac{L(\mu,\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})}, & 0 \le \hat{\mu} \le \mu \\ \dfrac{L(\mu,\hat{\hat{\theta}})}{L(0,\hat{\theta})}, & \hat{\mu} < 0 \end{array} \right\} = \dfrac{L(\mu,\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})}, \quad 0 \le \hat{\mu} \le \mu$$

"ratio of likelihoods", why ?

Why not simply using $L(\mu,\theta)$ as test statistics ?

- The number of degrees of freedom of the fit would be $N_\theta + 1_\mu$
- However, we are **not** interested in the values of θ (→ they are *nuisance* !)
- Additional degrees of freedom dilute interesting information on μ
- The "profile likelihood" (= ratio of maximum likelihoods) concentrates the information on what we are interested in
- It is just as we usually do for chi-squared: $\Delta\chi^2(m) = \chi^2(m, \theta_{best'}) - \chi^2(m_{best}, \theta_{best})$
- $N_{d.o.f.}$ of $\Delta\chi^2(m)$ is 1, and value of $\chi^2(m_{best}, \theta_{best})$ measures "Goodness-of-fit"

One-sided Test Statistics

$$\tilde{\lambda}(\mu) = \left\{ \begin{array}{ll} \dfrac{L(\mu,\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})}, & 0 \le \hat{\mu} \le \mu \\ \dfrac{L(\mu,\hat{\hat{\theta}})}{L(0,\hat{\theta})}, & \hat{\mu} < 0 \end{array} \right\} = \frac{L(\mu,\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})}, \quad 0 \le \hat{\mu} \le \mu$$

"one-sided" upper limit condition, why ?

Consider Discovery Case

- Want to test significance of signal excess

➠ Test p-value of background-only hypothesis

$$\tilde{\lambda}(0) = \frac{L(\hat{\mu}S + \hat{\hat{\theta}}B)}{L(\hat{\theta}B)} = \frac{L(\hat{\mu},\hat{\hat{\theta}})}{L(0,\hat{\theta})}, \qquad 0 \leq \hat{\mu}$$

- Produce toy experiments with $\mu = 0$ (fluctuate N_{obs} around B, and fluctuate θ), maximise both likelihoods, determine PDF($\tilde{q}_0 | B$) and compute:

$$\text{p-value [SM correct]} = \int_{\tilde{q}_0 \leq \tilde{q}_{0,\text{obs}}} \text{PDF}(\tilde{q}_0 \,|\, B)\ \mathrm{d}\tilde{q}_0$$

Consider Discovery Case

- Want to test significance of signal excess

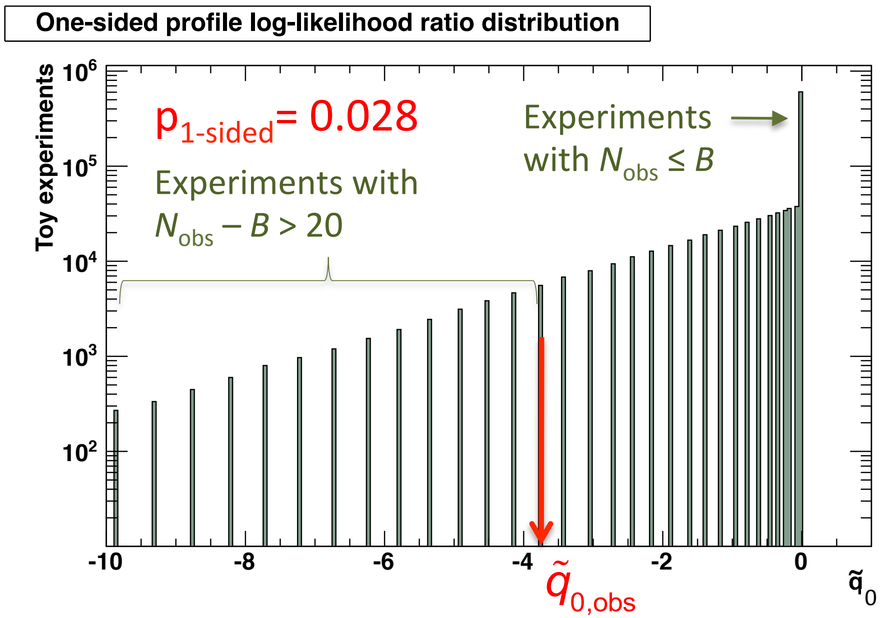
➠ Test p-value of background-only hypothesis

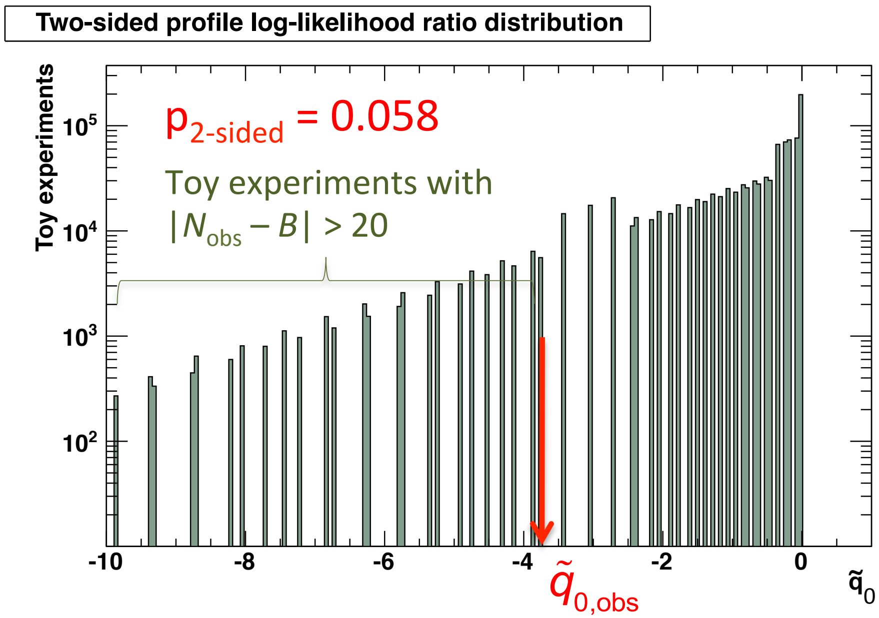
$$\tilde{\lambda}(0) = \frac{L(\hat{\mu}S + \hat{\hat{\theta}}B)}{L(\hat{\theta}B)} = \frac{L(\hat{\mu},\hat{\hat{\theta}})}{L(0,\hat{\theta})}, \qquad 0 \leq \hat{\mu}$$

If new physics cannot destructively interfere with SM (background), can inject that $S \geq 0$

Consider Discovery Case

- Example: N_{obs} = 120, B = 100 no uncertainty on B

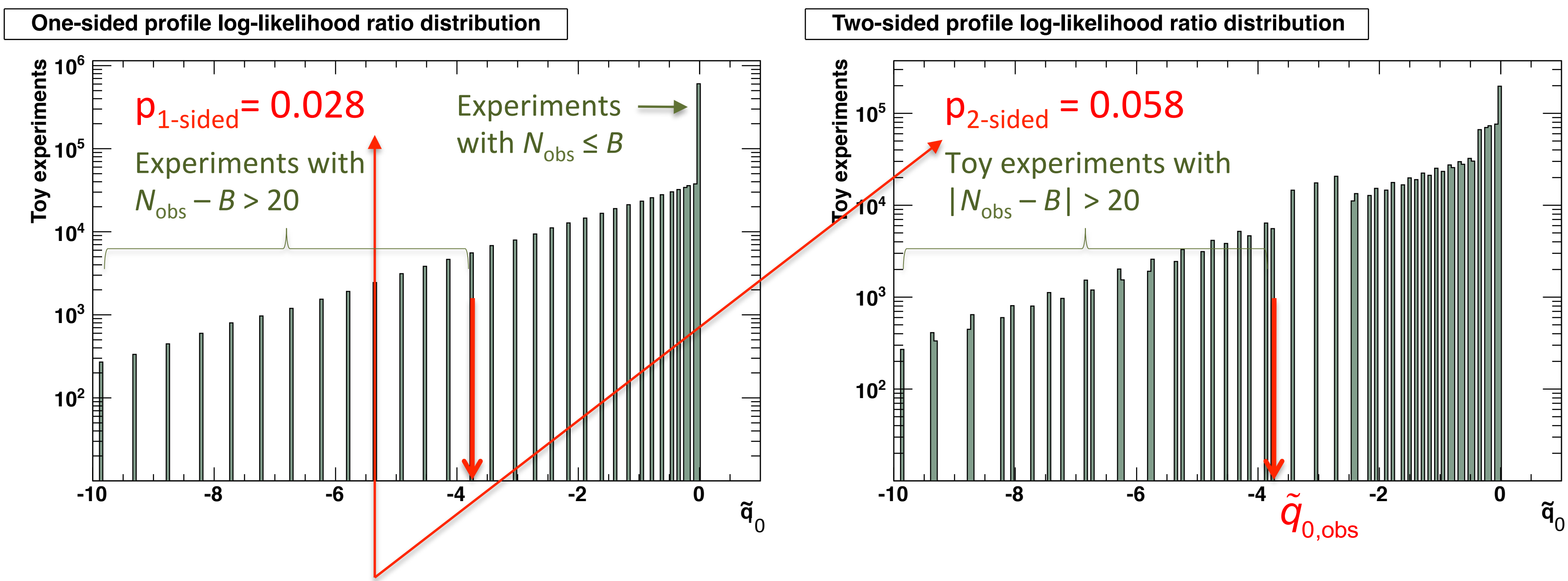




- Injecting $S \geq 0$ information has reduced p-value by factor of ≈2 and thus enhanced discovery reach
- $S < 0$ solution represents a dilution of the statistical information in the data

Consider Discovery Case

- Example: N_{obs} = 120, B = 100 no uncertainty on B



- Difference between factor = 2 due to asymmetric Poisson statistics
- Compare: N_{obs} = 1062, B = 1000 → $p_{1\text{-sided}}$ = 0.027, $p_{2\text{-sided}}$ = 0.054

  N_{obs} = 15, B = 9 → $p_{1\text{-sided}}$ = 0.041, $p_{2\text{-sided}}$ = 0.096

Upper Limit Case

- No signal excess, want to obtain upper limit

➠ Test p-value of signal + background hypothesis

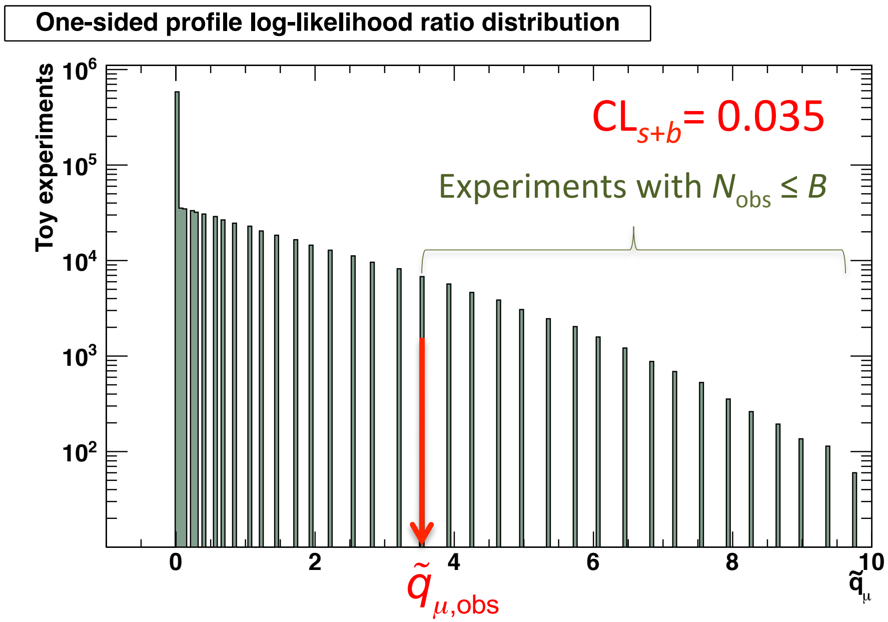
$$\tilde{\lambda}(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}, \qquad 0 \leq \hat{\mu} \leq \mu$$

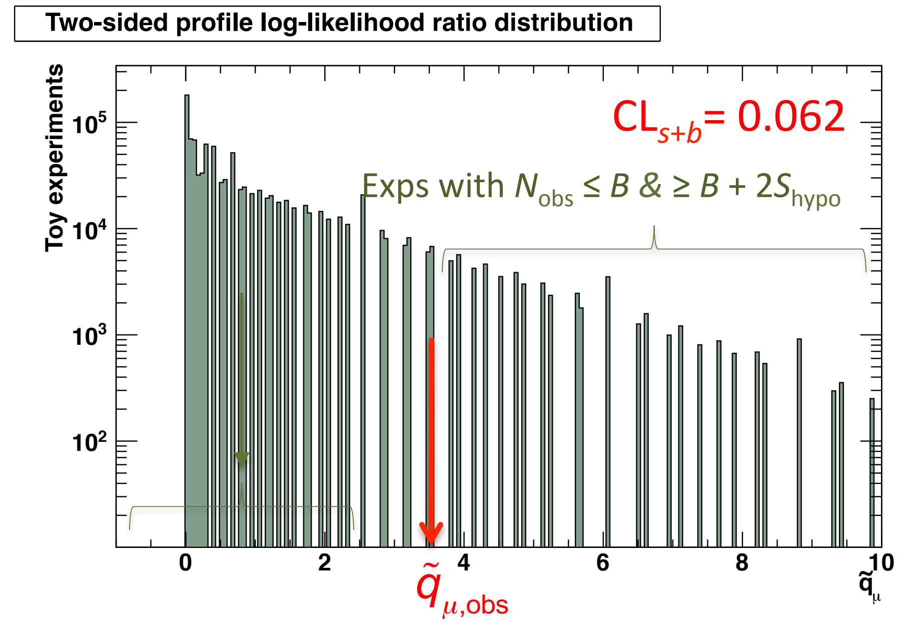
- Produce toy experiments with $\mu = \mu_{\mathrm{hypo}}$ (fluctuate N_{obs} around $\mu S + B$, and fluctuate θ), maximise likelihoods, determine PDF($\tilde{q}_\mu \mid \mu S + B$) and compute:

$$\mathrm{CL}_{s+b}(\mu) = \int_{\tilde{q}_\mu \geq \tilde{q}_{\mu,\mathrm{obs}}} \mathrm{PDF}(\tilde{q}_\mu \mid \mu S + B)\ \mathrm{d}\tilde{q}_\mu$$

Upper Limit Case

- Example: $N_{\text{obs}} = 100$, $B = 100$ (no error), $S_{\text{hypo}} = 20$





- Again, injecting $S \ge 0$ information has improved sensitivity of analysis (95% CL limits of $18.1_{\text{1-sided}}$ vs. $21.3_{\text{2-sided}}$)

Nuisance Parameters

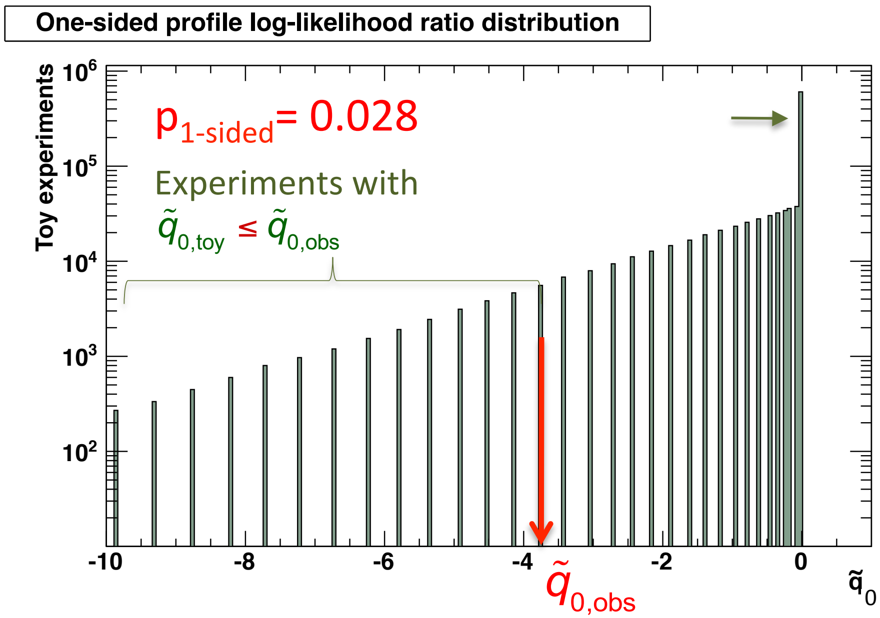
- So far only discrete cases considered: nothing else than Poisson probability summation
- The problems come when maximising likelihoods with respect to nuisance parameters

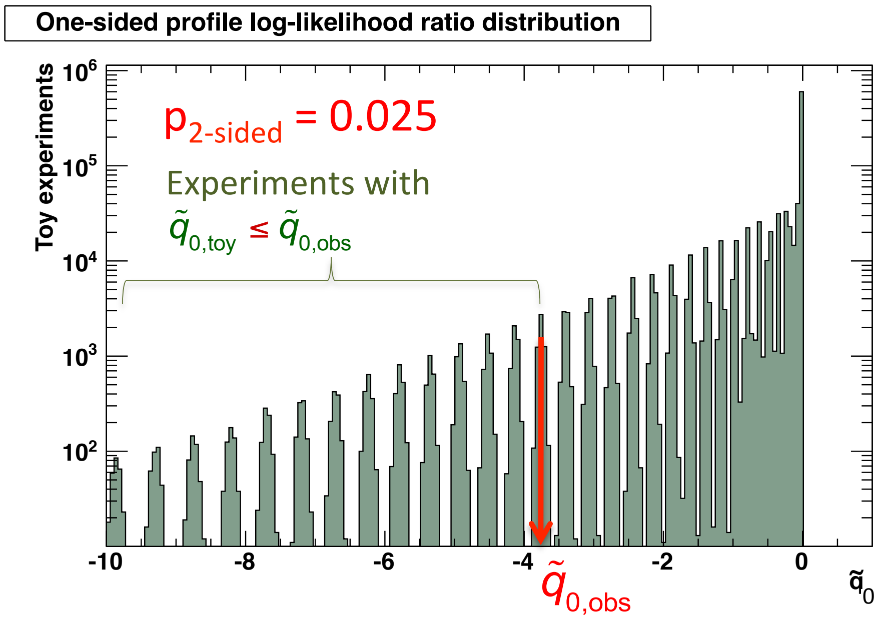
$$\tilde{\lambda}(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}, \qquad 0 \le \hat{\mu} \le \mu$$

- Additional Gaussian terms make *L* continuous

Discovery Case with Error on *B*

- Example: N_{obs} = 120, B = 100 ± 0.1

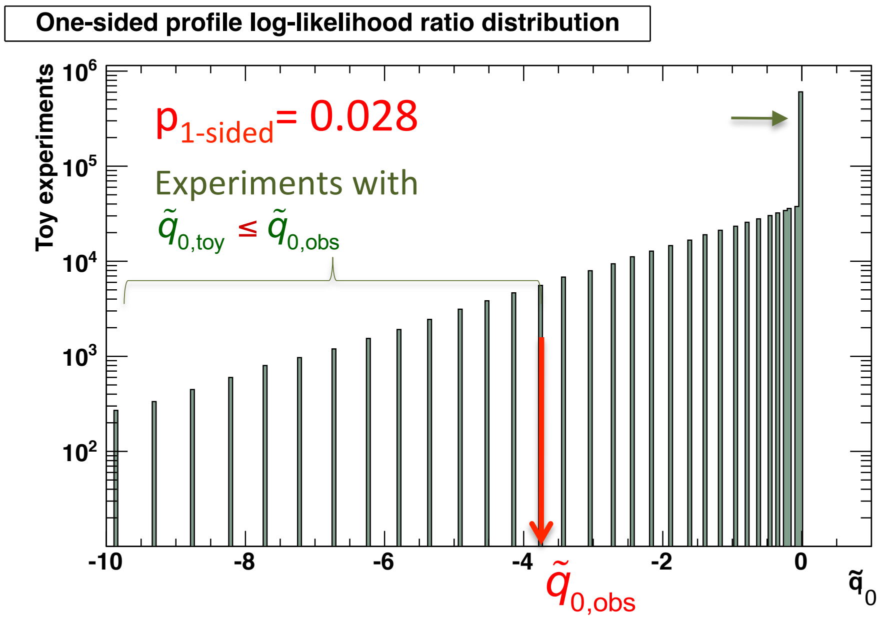


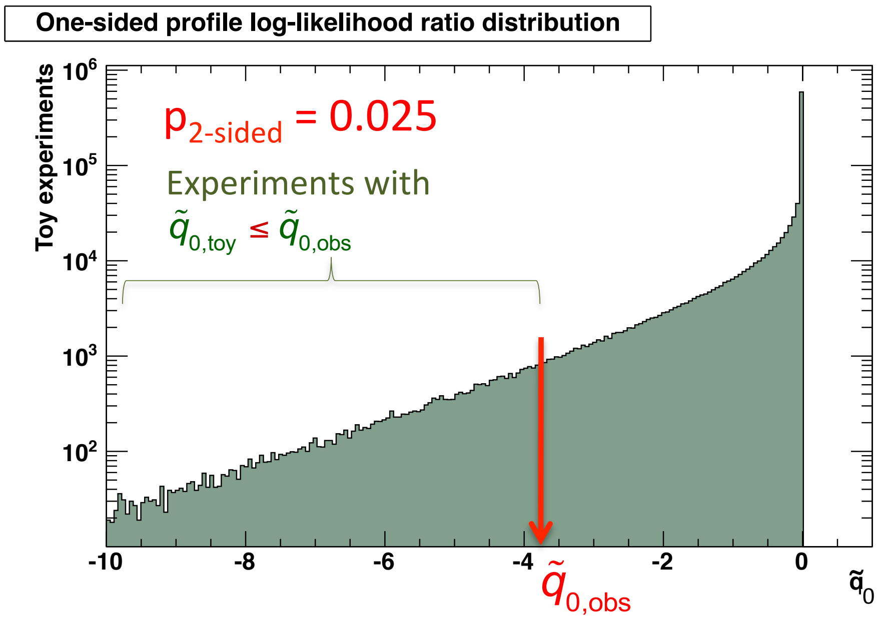


- Half of experiments with $q_{0,toy} \approx q_{0,obs}$ unaccounted in continuous case
- This gives a better (!) discovery reach, and also a more stringent upper limit

Discovery Case with Error on *B*

- Example: $N_{obs} = 120$, $B = 100 \pm 1$

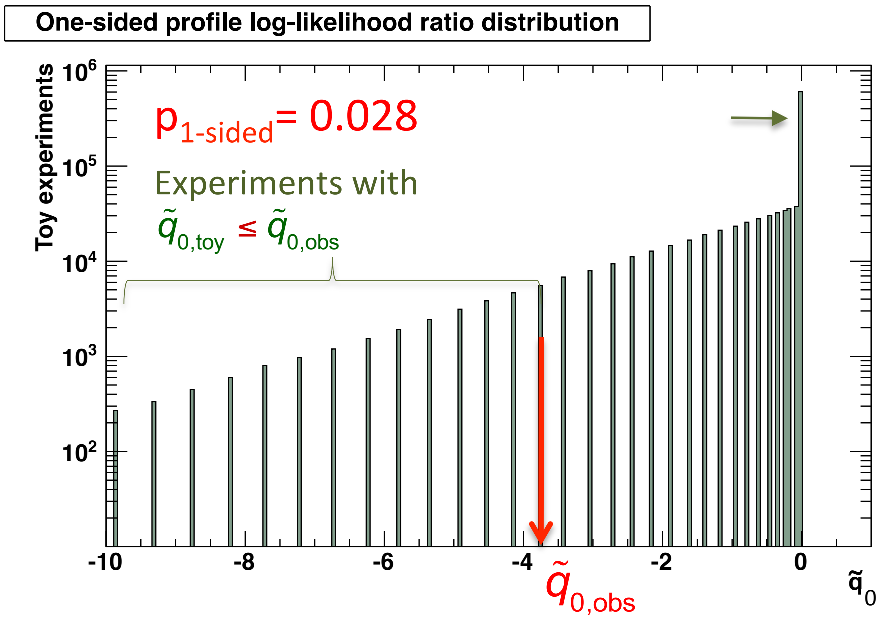


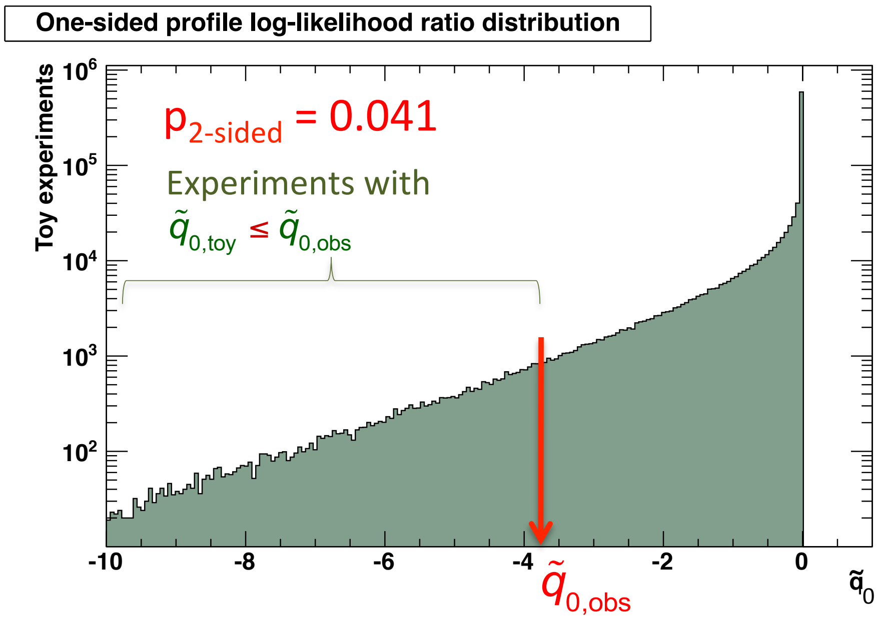


- Half of experiments with $q_{0,toy} \approx q_{0,obs}$ unaccounted in continuous case
- With increasing background uncertainty the p-value gets larger again

Discovery Case with Error on *B*

- Example: N_{obs} = 120, B = 100 ± 5

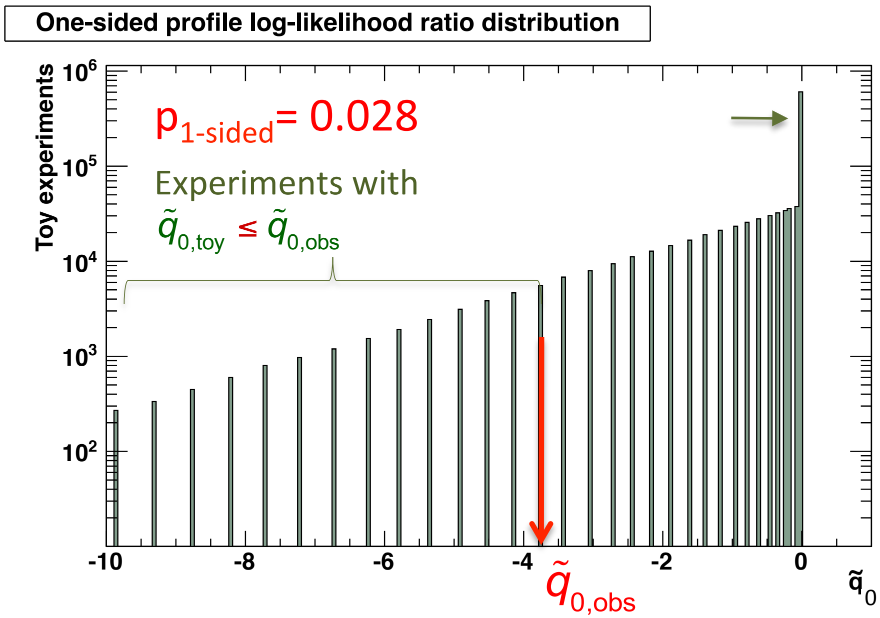


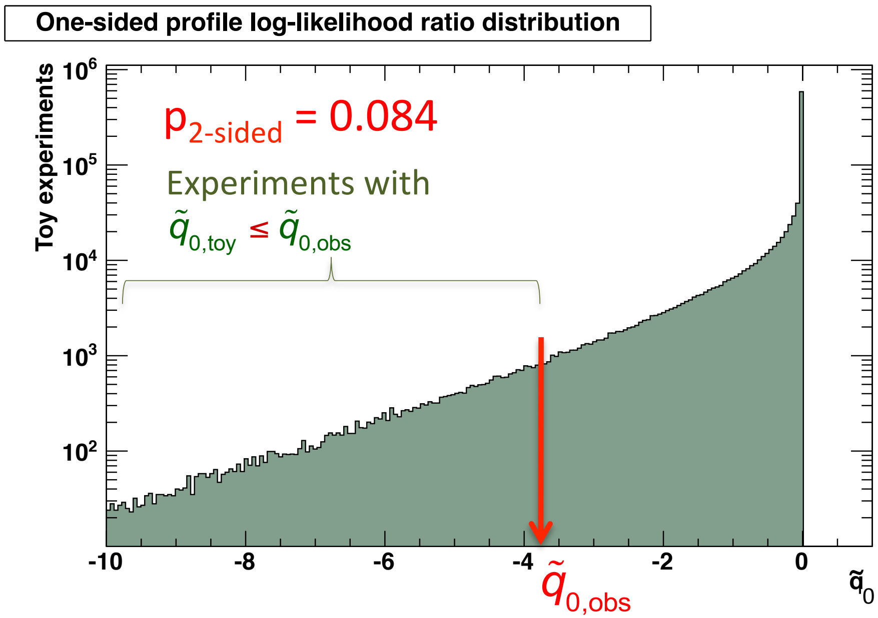


- Half of experiments with $q_{0,toy} \approx q_{0,obs}$ unaccounted in continuous case
- With increasing background uncertainty the p-value gets larger again

Discovery Case with Error on B

- Example: N_{obs} = 120, B = 100 ± 10





- Half of experiments with $q_{\mu,toy} \approx q_{\mu,obs}$ unaccounted in continuous case
- Eventually, the discovery reach becomes worse than in the discrete case

Discrete vs. Continues Test Statistics

- To bring discrete and continuous case together for negligible error on *B*, compute p-value as follows:

$$\text{p-value}\ (N_{\text{obs}}, B) = \frac{1}{2} \cdot P(N_{\text{obs}}, B) + \sum_{n=N_{\text{obs}}+1}^{\infty} P(n, B)$$

P is Poisson probability

- In that case, p-value of previous example decreases from 0.028 to 0.025 (= continuous case with small σ_B)
- Justification: discrete case "overcovers"
- Will get back to coverage later…
- See: [document on discreteness problem (Glen + Eilam)](https://twiki.cern.ch/twiki/pub/AtlasProtected/ATLASStatisticsFAQ/PLvsInt.pdf)

https://twiki.cern.ch/twiki/pub/AtlasProtected/ATLASStatisticsFAQ/PLvsInt.pdf

Upper Limit with Null Observation

- Naïve solution:

$$0.05 = P(N_{obs} = 0, S_{95} + B)$$
$$\Rightarrow S_{95} = 2.996 - B$$

- With new prescription:

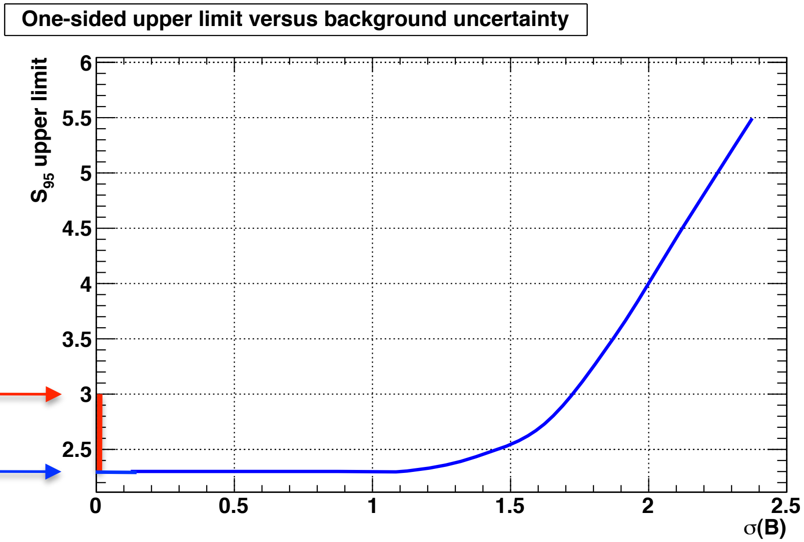
$$0.05 = \frac{1}{2} \cdot P(N_{obs} = 0, S_{95} + B)$$
$$\Rightarrow S_{95} = 2.303 - B$$

- Example for

$N_{obs} = 0,\ B = 0 \pm \sigma(B)$

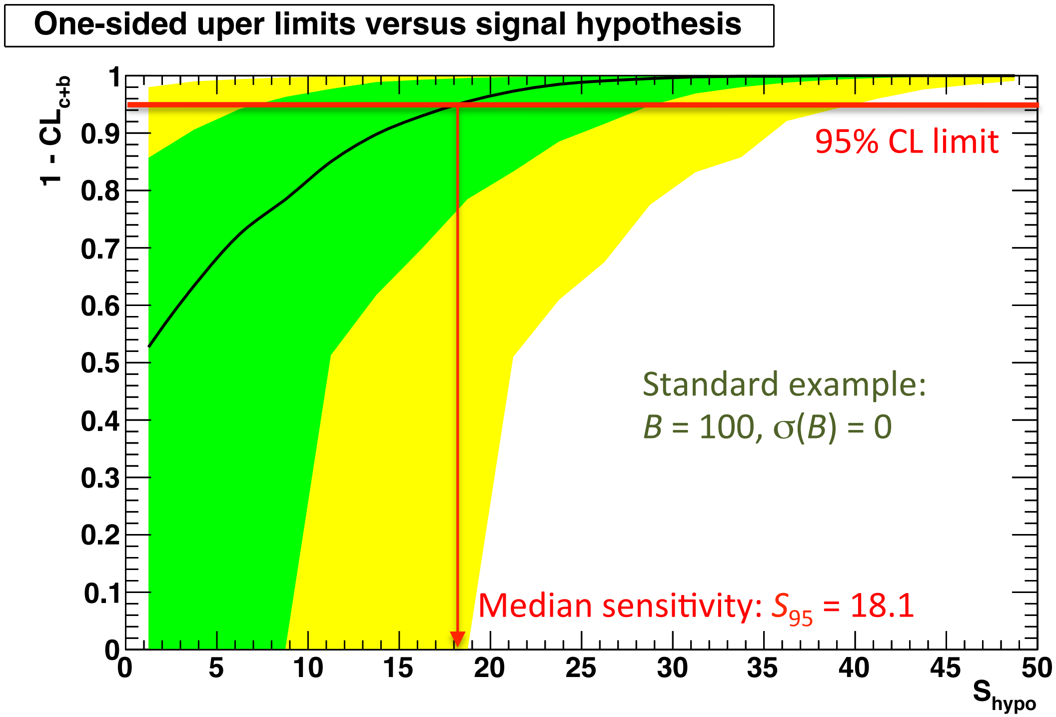
Discrete limit without background uncertainty

Discrete limit with new prescription



Expected Limits – Median Sensitivity

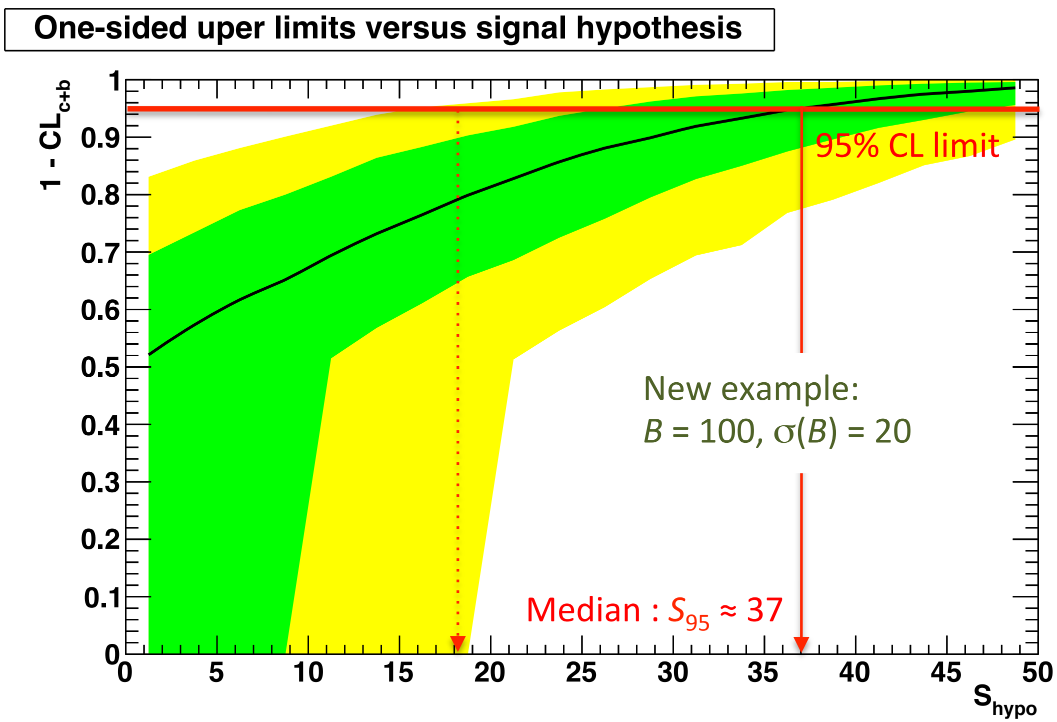
- Prescription to compute "green & yellow bands"
  - Median sensitivity is based on background only hypothesis



1. Create toy experiments where N_{obs} fluctuates around B only
2. Scan through S_{hypo}
3. For each toy experiment compute $CL_{s+b}(S_{\text{hypo}})$ [from another toy !]
4. Determine median and 68%, 95% error bands for $CL_{s+b}(S_{\text{hypo}})$
5. Plot bands and publish yet another limit

Expected Limits – Median Sensitivity

- Prescription to compute "green & yellow bands"
  - Median sensitivity is based on background only hypothesis



1. Create toy experiments where N_{obs} fluctuates around B only
2. Scan through S_{hypo}
3. For each toy experiment compute $CL_{s+b}(S_{hypo})$ [from another toy !]
4. Determine median and 68%, 95% error bands for $CL_{s+b}(S_{hypo})$
5. Plot bands and publish yet another limit

Being a Good Citizen

- Our CL_{s+b} UL takes benefit from upwards fluctuations in background (remember the N_{obs} = 0 case: S_{95} = 2.3 – B)

  [Would not be the case for: $\tilde{\lambda}(0) = L(\hat{\mu},\hat{\hat{\theta}}) \big/ L(0,\hat{\theta})$ → null observation limit increases with B !]

- With some luck, limits (far) better than sensitivity could be obtained

- Discuss two remedies here: CL_s and PCL

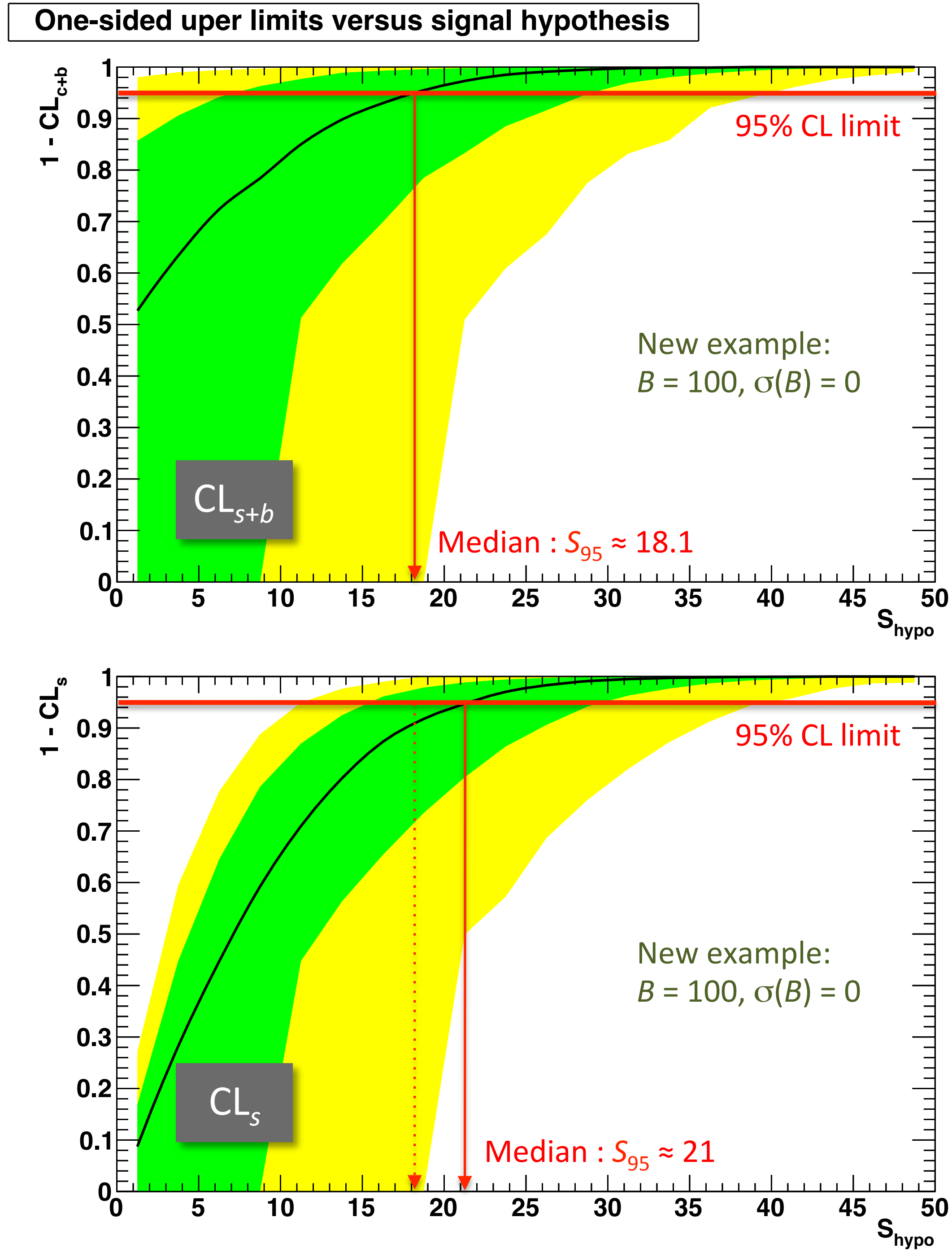
Modified Frequentist Method

- LEP (A. Read) & Tevatron: $CL_s = CL_{s+b} / CL_b$, where:

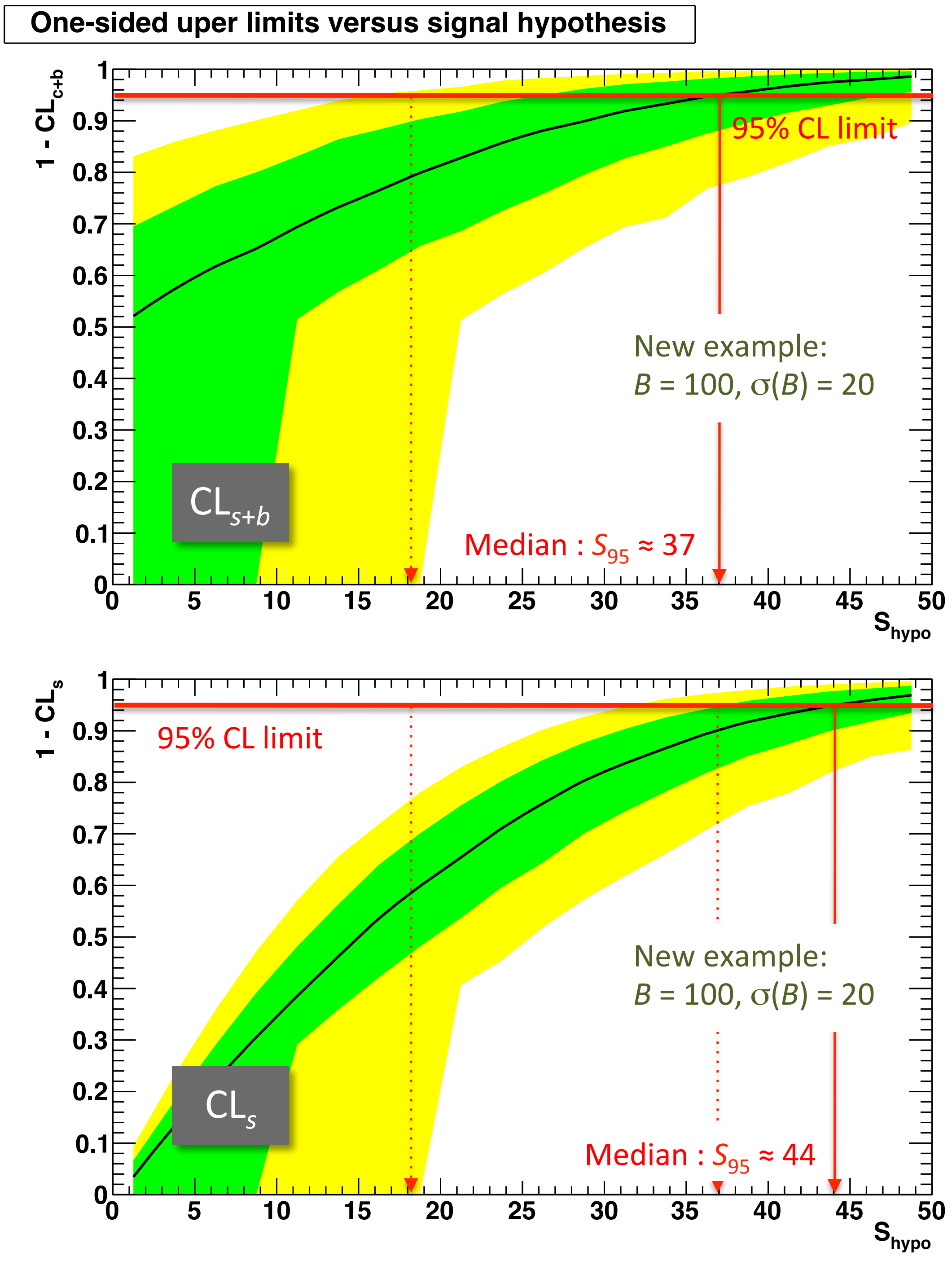
$$CL_b(\mu) \quad = \int_{\tilde{q}_\mu \geq \tilde{q}_{\mu,\text{obs}}} \text{PDF}(\tilde{q}_\mu \mid B) \;\, d\tilde{q}_\mu$$

  - This is not a statistical method in the proper sense: the ratio of two probabilities is not a probability
  - $CL_s(S_{95,\text{obs}}) = 0.05$ determines 95% CL upper limit $S_{95,\text{obs}}$
  - Dividing by CL_b is a penalty: in case of a fluctuation away from expected B, both CL_{s+b} and CL_b will be small, but not CL_s
  - CL_s has overcoverage in general

Reuse previous example to illustrate CL_s



Reuse previous example to illustrate CL_s



Power-Constrained Limit (PCL)

- Keep CL_{s+b} and solve problem of over-exclusion by introducing a "power constraint"
  - $\mathrm{CL}_{s+b}(S_{95}) = 0.05$ determines 95% CL upper limit $S_{95,\mathrm{obs}}$
  - However, use constraint: $S_{95,\mathrm{obs}} = \mathrm{Max}(S_{95,\mathrm{obs}}, S_{95,\mathrm{median}} - 1\sigma)$
  - Choice of power constraint is arbitrary, but fixed
  - PCL has advantage of proper coverage, and protects against excluding non-testable hypotheses
  - CL_s is also arbitrary and overcovers, but has advantage of being smooth → may appear less *ad hoc* to non-experts (at conferences)

Remark on Coverage

- "CL_{s+b}, if obtained from toy experiments has correct coverage". Correct ? No !
  - It only has proper coverage if the nuisance parameters used to create the toys correspond to the truth
  - This assumption can only be wrong
  - Limits obtained will depend on θ_{truth} values used
  - Custom but not unique choice is to use best fit values θ_{fit}
  - A conservative limit should include θ_{truth} variations, but full Neyman construction impossible because θ_{truth} unbound
  - Try *ad hoc* variation $\theta_{\text{truth}} = \theta_{\text{fit}} \pm 1\sigma$ and redetermine limits

➠ Effect on standard example very small (N_{obs} = 100, B = 100 ± 20): ΔS_{95} = 1.3%

How to Generate Toy Experiments

- The way how toy experiments are generated matters
- To obtain upper limit for given signal hypothesis:

1. Compute observed test statistics in data
2. Generate for toy $\{i\}$ $N_{\mathrm{obs},i}$ around expected background + signal hypothesis using best fit values for nuisance parameters (unsmeared!)
3. Generate Gaussian-smeared nuisance parameters θ_i around best fit values for hypothesis ("unconditional ensemble")
4. Compute test statistics using $N_{\mathrm{obs},i}$ and smeared θ_i, representing the measurements of that toy experiment
5. Count how often toy test statistics is larger or equal than data test statistics and compute CL_{s+b}

Short Cuts – Asymptotic Behaviour

- One could not want to bother with toys and use "Wilk's theorem" instead, ie, postulate: $\Delta\chi^2(\mu) = q_\mu - q_{\min}$, and compute $CL_{s+b}(\mu)$ = TMath::Prob($\Delta\chi^2(\mu)$, 1)
  - Usually not good in presence of small numbers
  - Should preferably not be used for the observed limit or small evidence p-value
  - For 5σ discovery, one would need at least 10M toys to see a few events, impractical
  - Could be used to derive median sensitivity and error bands, which may be necessary in case of very complex, CPU-intensive fits

Short Cuts – Asymptotic Behaviour

- The test statistics $\tilde{q}_{\mu}$ has well defined asymptotic behaviour for sufficiently large data samples
  - Asymptotic PDF for given μ hypothesis known analytically
  - PDF requires standard deviation of floating signal strength parameter, which can be obtained for given μ
  - Very useful for expected limit ("yellow & green band") computation
  - This is nicely described in G. Cowan et al. arXiv:1007.1727

http://arXiv.org/abs/arXiv:1007.1727

References

- ATLAS SCs Frequentist Limit Recommendation
  https://twiki.cern.ch/twiki/pub/AtlasProtected/StatisticsTools/Frequentist_Limit_Recommendation.pdf
- Document on discreteness problem (Glen + Eilam)
  https://twiki.cern.ch/twiki/pub/AtlasProtected/ATLASStatisticsFAQ/PLvsInt.pdf
- Paper on asymptotic formulas (G. Cowan et al)
  http://arXiv.org/abs/arXiv:1007.1727
- 1st ATLAS Physics & Statistics meeting, Mar 15, 2011
  https://indico.cern.ch/conferenceDisplay.py?confId=131204
- ATLAS Physics & Statistics workshop, April 15, 2011
  https://indico.cern.ch/conferenceDisplay.py?confId=132499
- Nicolas Berger's asymptotic behaviour study for $H \rightarrow \gamma\gamma$
  https://indico.cern.ch/getFile.py/access?contribId=1&resId=1&materialId=slides&confId=130102
- Most recent CDF + D0 Higgs combination paper
  http://www-d0.fnal.gov/Run2Physics/WWW/results/prelim/HIGGS/H106/H106.pdf