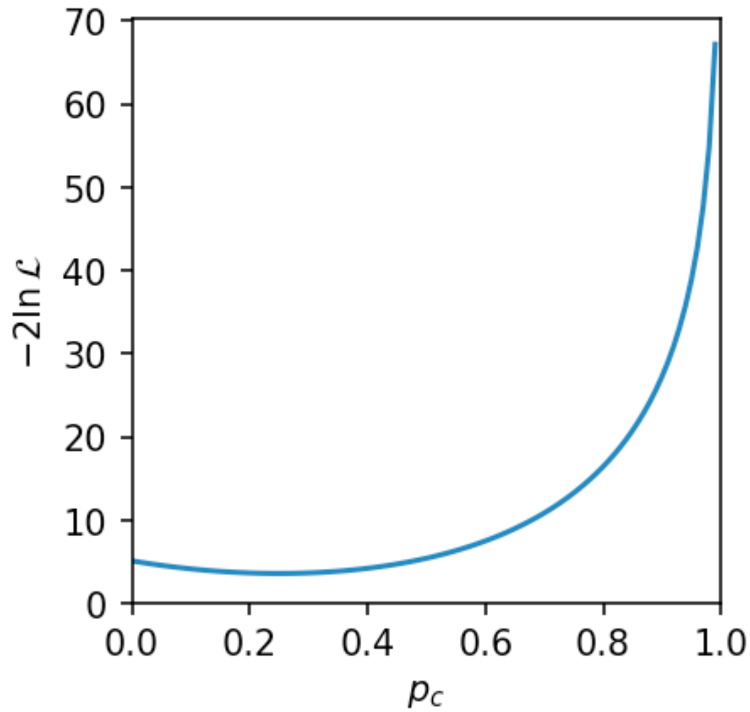
Intervals

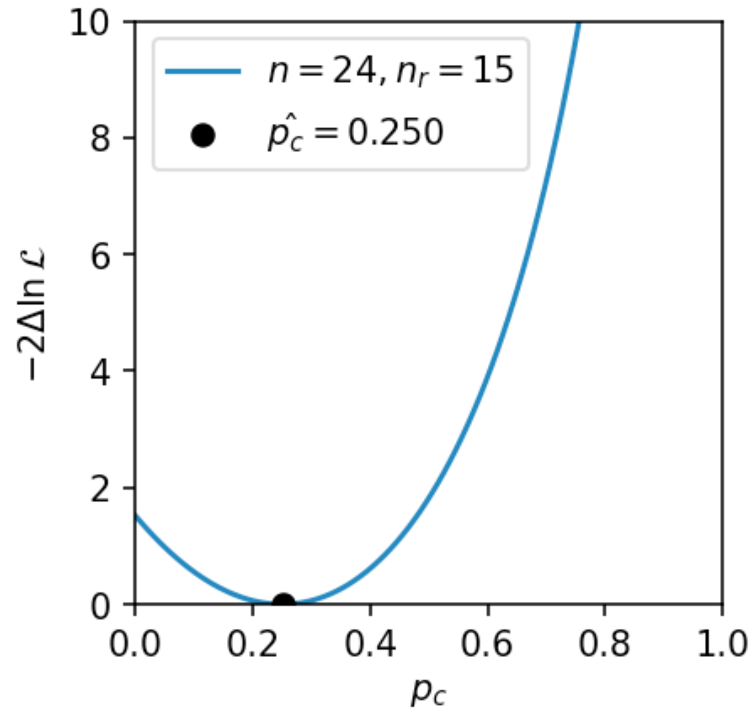
Topics: intervals

- Hypothesis tests
- Neyman intervals
- Likelihood ratio test statistic
  - Neyman-Pearson lemma
- Under-fluctuation and CLs
- Asymptotic behavior - Wilk's theorem

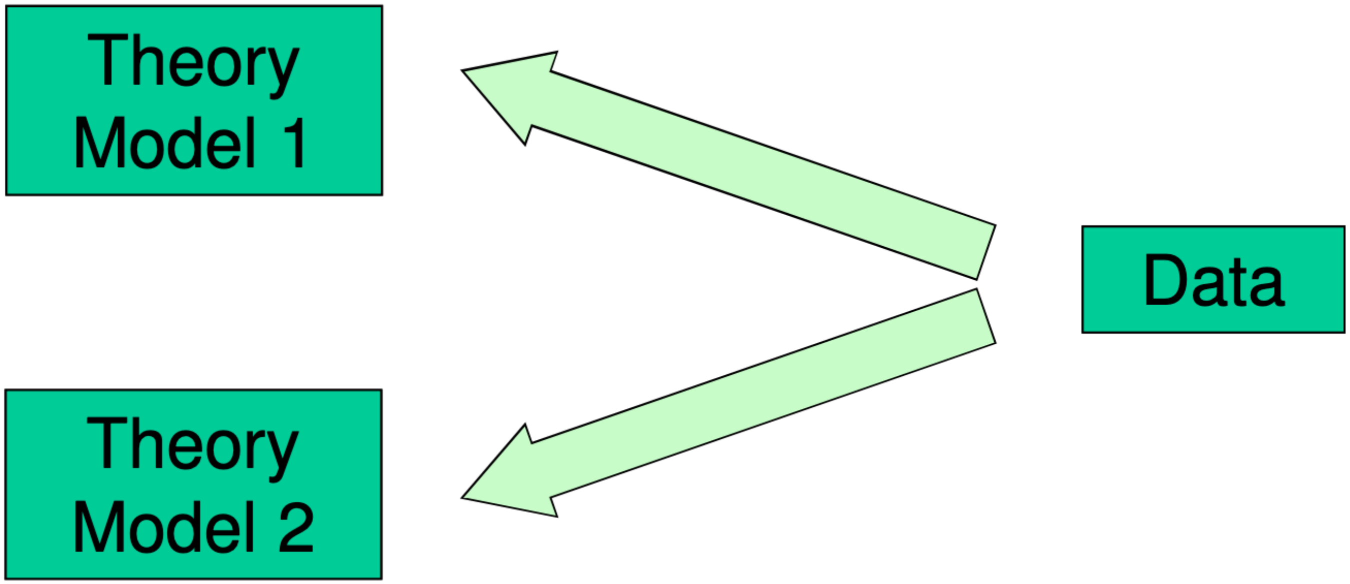
Inferring P(cheat) as a frequentist

- Scan $-2\ln\mathscr{L}(p_c) = -2\ln f(n_r; n, p_c)$
- Find minimum
- ???
- Profit





Hypothesis tests



Which hypothesis is the most consistent with the experimental data?

Hypothesis tests

- Simple test parameterized by two probabilities: α, β
  - β with respect to an alternate hypothesis

<table>
<tr><th colspan="2" rowspan="2">Table of error types</th><th colspan="2">Null hypothesis (H_0) is</th></tr>
<tr><th>TRUE</th><th>FALSE</th></tr>
<tr><td rowspan="2">Decision about null hypothesis (H_0)</td><td>Don't reject</td><td>Correct inference
(true negative)

(probability = $1-\alpha$)</td><td>Type II error
(false negative)

(probability = β)</td></tr>
<tr><td>Reject</td><td>Type I error
(false positive)

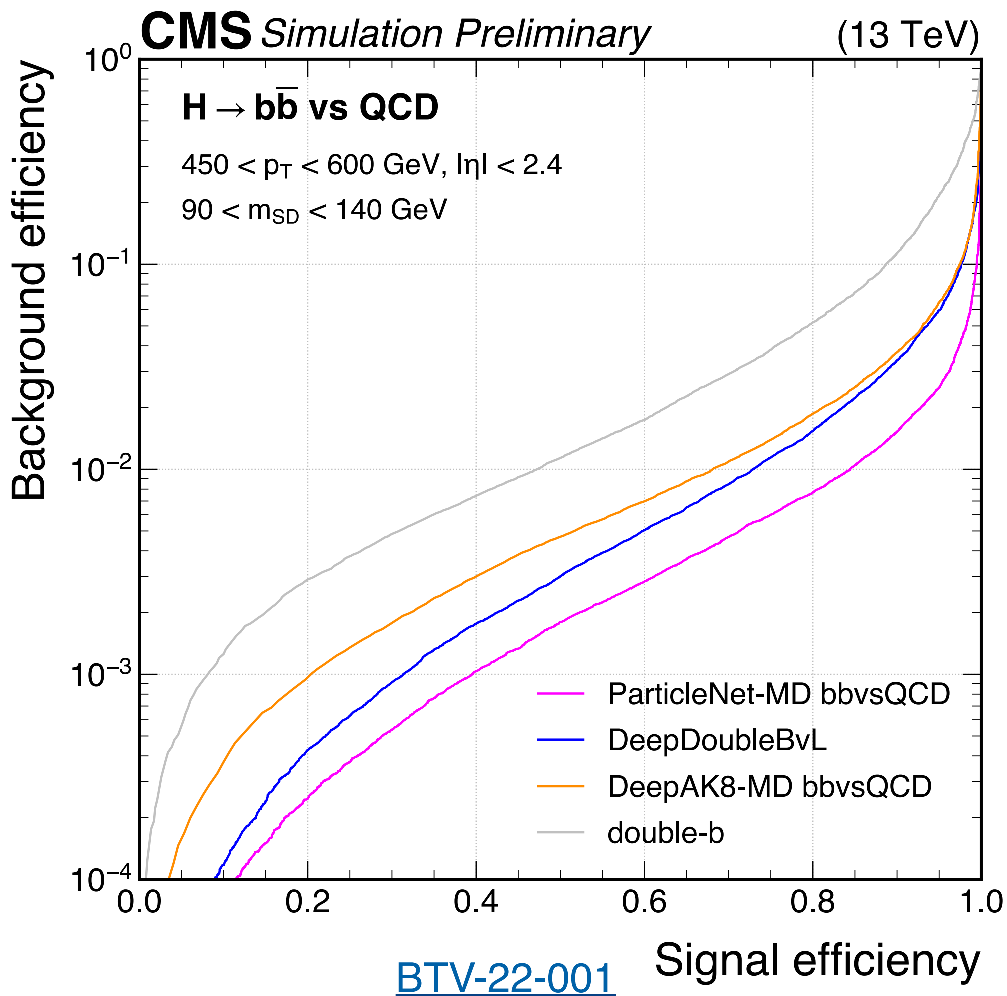
(probability = α)</td><td>Correct inference
(true positive)

(probability = $1-\beta$)</td></tr>
</table>

Hypothesis test example

- ROC plot: let H_0 be QCD jet
  - DNN tagger score is the test statistic
    - Lower values more consistent with H_0
- Background efficiency
  - True negative: accept H_0 when true
  - $1 - \alpha$ (where α is the *size*)
- Signal efficiency
  - True positive: reject H_0 when false
  - *Power* $1 - \beta$
- Prefer tests with higher *power* for a given *size*
  - Here, test size is large
  - Usually test size is *very small*

Comparison of the performance of the X→bb identification algorithms … in the 2018 data-taking conditions



BTV-22-001

Fermilab

Coverage (confidence) sets

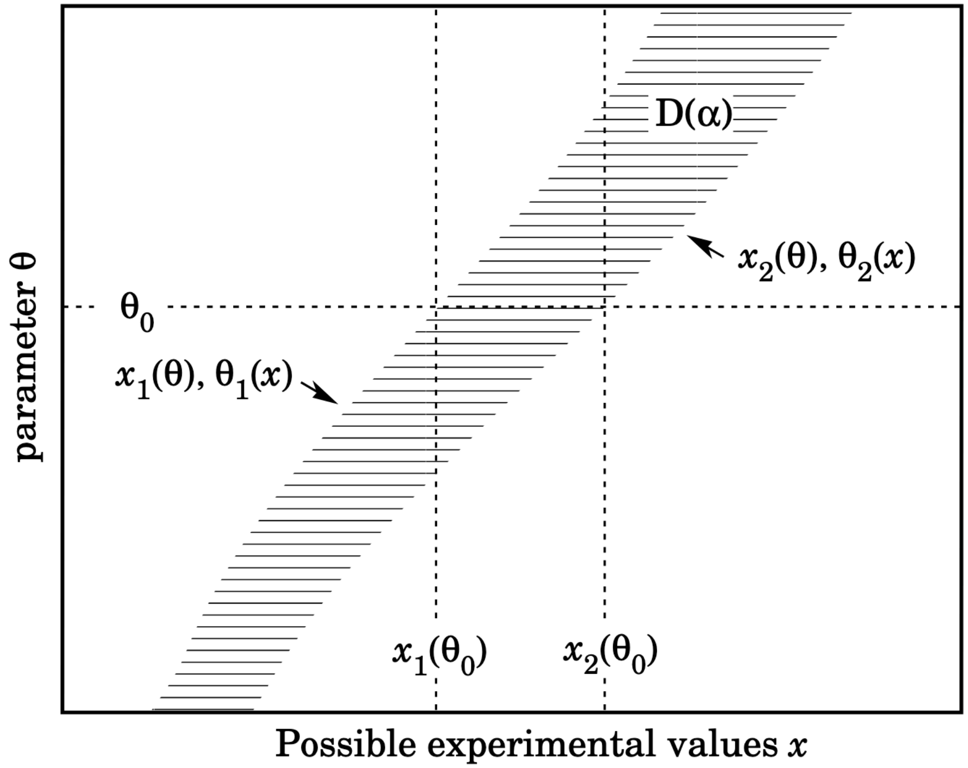
- One can ask, “assuming a value of θ, would x_{obs} be a likely outcome?”
  - This is a hypothesis test (sufficiently likely or not)
- We have P(x;θ), so we can answer this if we:
  - Choose a *size* (significance level) α of the test (e.g. 0.05)
  - Define a test statistic (ordering) of possible outcomes
  - Run pseudo-experiments (toys) for each θ to determine distribution of test statistic
  - Perform experiment, report set of θ where test statistic is below the 1-α quantile
- A good ordering will lead to
  - Good *coverage*: in repeated experiments the (unknown) θ_{true} will be in the set with probability at least 1-α, though it may over-cover
  - High *power* (1-β): the set does not contain θ_{alt} for some specified alternative hypothesis

Neyman interval

- Neyman construction:
  - For each θ, find range $[x_1, x_2]$ s.t.

$$P(x_1 < x < x_2; \theta) = \int_{x_1}^{x_2} P(x; \theta)\, dx \geq 1 - \alpha$$

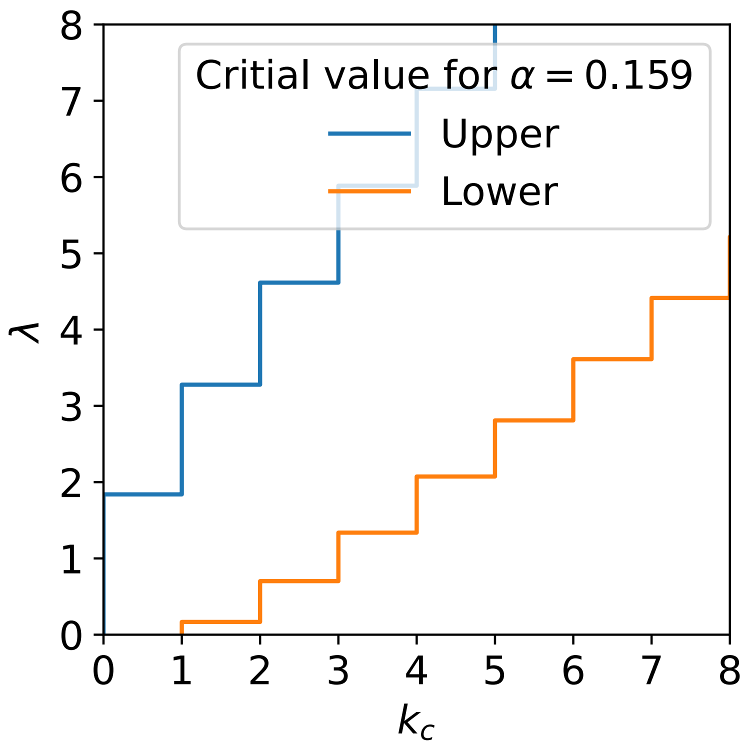
  - Perform experiment
  - Report confidence interval: $[\theta_1, \theta_2]$ where $x_{obs} \in [x_1(\theta), x_2(\theta)]$ for all $\theta \in [\theta_1, \theta_2]$

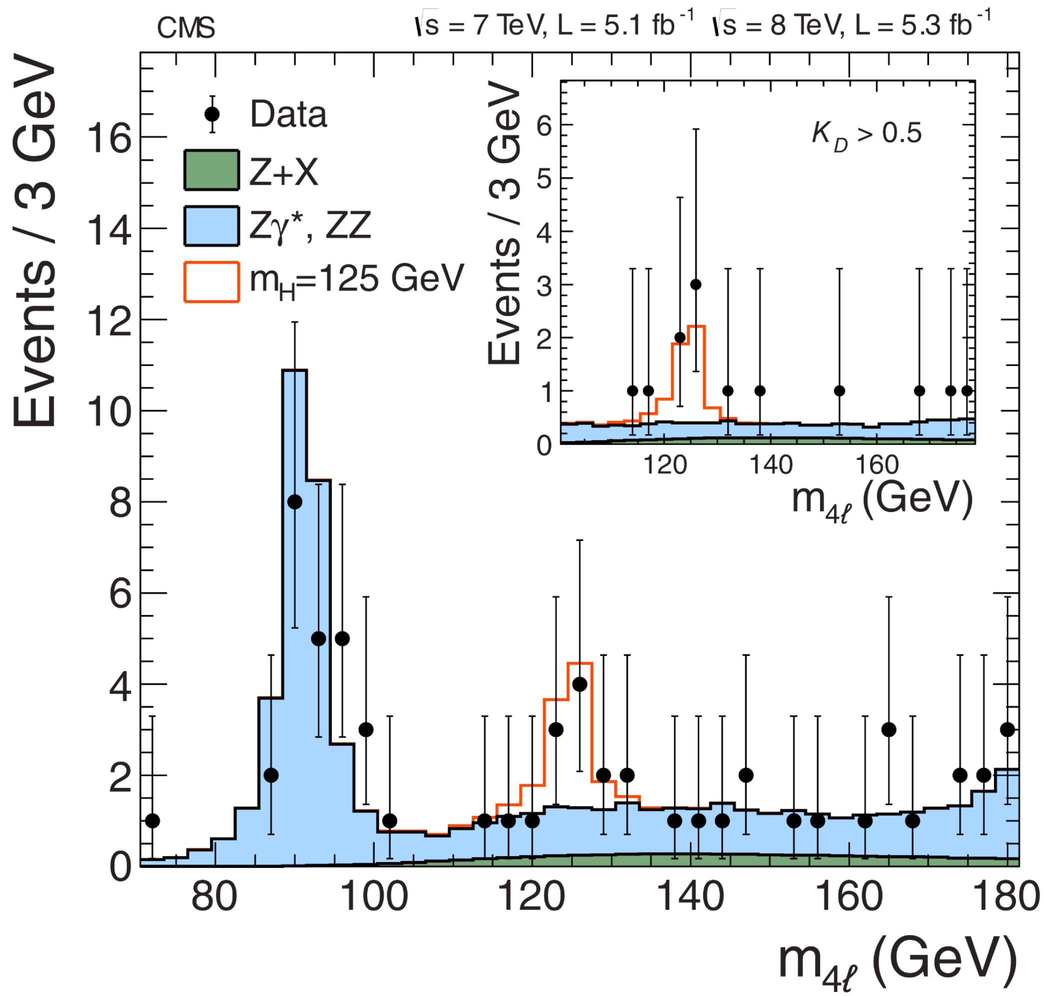


- Interval has coverage 1-α.
  - For an ensemble of experiments, the interval $[\theta_1, \theta_2]$ will contain (unknown) θ_{true} with probability 1-α. This is a statement about the distribution of θ_1 and θ_2, NOT θ_{true}.

Neyman interval

- You have all seen these: error bars on data points are the Neyman intervals for a Poisson distribution
  - With $\alpha = 0.159\ldots$
  - Also referred to as Garwood intervals





Likelihood ratio test statistic

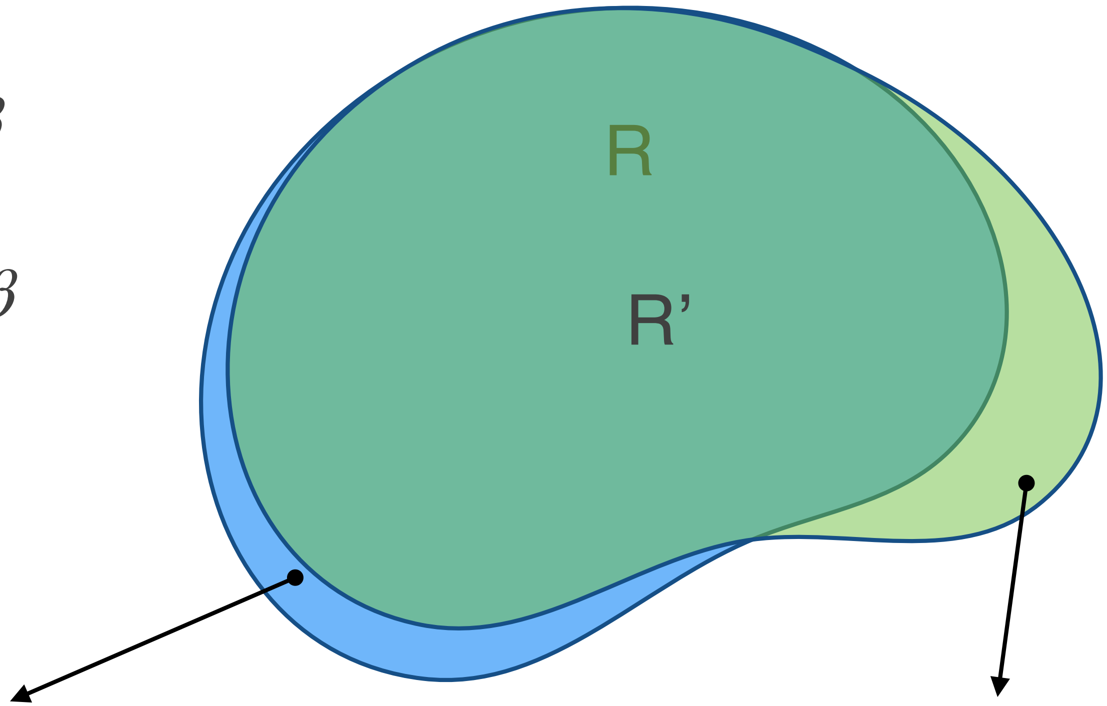
1. Define $t_\theta(x) = -2\ln\dfrac{\mathscr{L}(\theta)}{\mathscr{L}(\hat{\theta})} = -2\ln\dfrac{f(x;\theta)}{f(x;\hat{\theta})} \geq 0$
2. Compute associated pdf (change of variables)
$$f(t_\theta;\theta') = \int \delta(t_\theta - t_\theta(x))\, dP(x;\theta')$$
3. For each θ, find the critical value $t_{\theta,c}$ that covers
$$P(t_\theta < t_{\theta,c}) = \int_0^{t_{\theta,c}} f(t_\theta;\theta'=\theta)\, dt_\theta \geq 1-\alpha$$
4. Perform experiment, get x_{obs}, report confidence set $\{\theta \,|\, t_\theta(x_{obs}) < t_{\theta,c}\}$

Again, the set is the random variate, and will contain (unknown) θ_{true} with probability at least 1-α. Dimension of θ and x are arbitrary.

If θ is 1-d and t_θ is monotone, can make a Feldman-Cousins interval.

Neyman-Pearson lemma

- $R = \{x \,|\, t_\theta(x) \geq t_{\theta,c}\}$, or $\{x \,|\, f(x;\hat{\theta}) \geq e^{t_{\theta,c}/2} f(x;\theta) = cf(x;\theta)\}$
- $P(x \in R; \theta) = \alpha$
- $P(x \in R; \hat{\theta}) = 1 - \beta$
- $P(x \in R'; \theta) = \alpha$
- $P(x \in R'; \hat{\theta}) < 1 - \beta$
  - We reject θ less often



$$f(x;\hat{\theta}) \geq cf(x;\theta)$$

$$P(x \in R \backslash R'; \hat{\theta}) \geq \boxed{cP(x \in R \backslash R'; \theta)}$$

$$f(x;\hat{\theta}) < cf(x;\theta)$$

$$P(x \in R' \backslash R; \hat{\theta}) < \boxed{cP(x \in R' \backslash R; \theta)}$$

$$P(x \in R \backslash R'; \hat{\theta}) > P(x \in R' \backslash R; \hat{\theta})$$

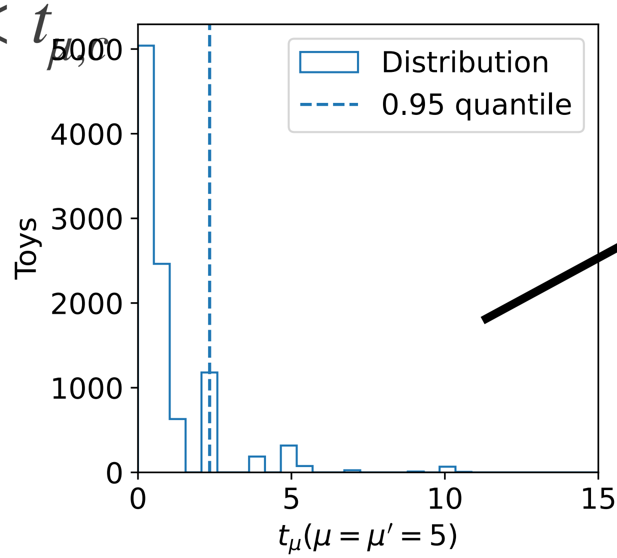
The (log-)likelihood ratio test is the most powerful test for a given size

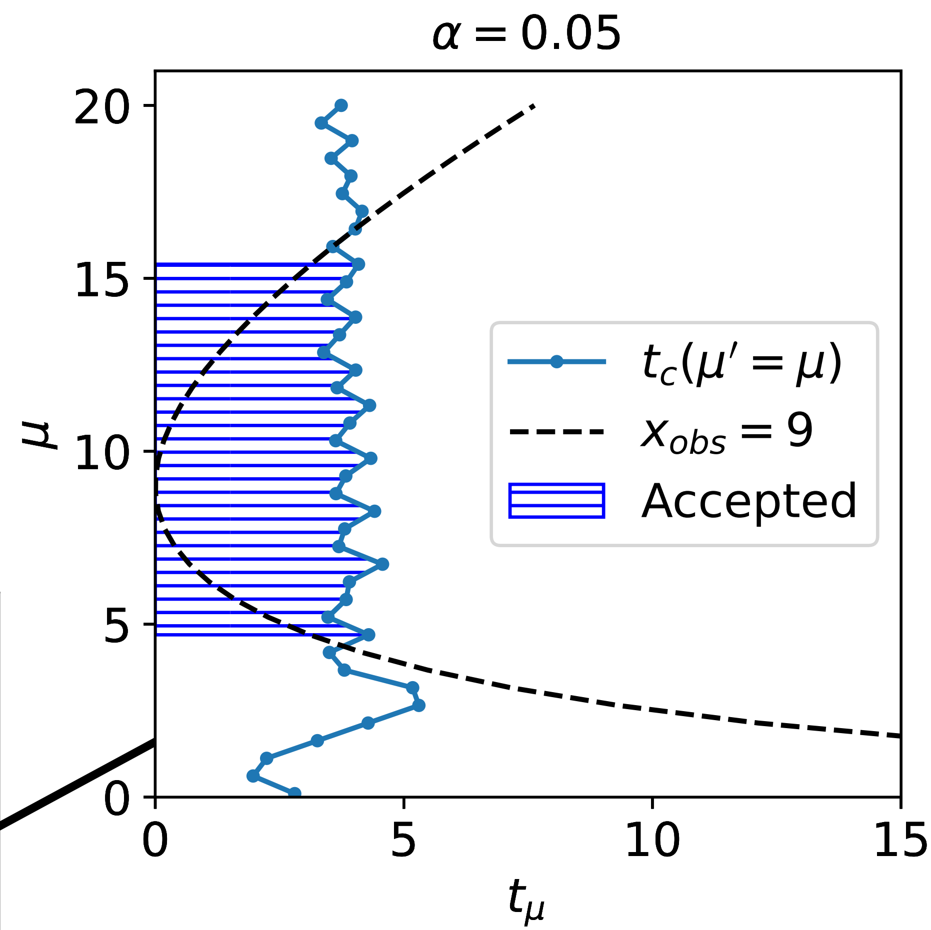
Frequentists…

Likelihood ratio examples

- Poisson example
  - $P(x;\mu) = \dfrac{\mu^x e^{-\mu}}{x!}$
- For each μ:
  - Throw 10k toys
  - Compute $\hat{\mu}$, t_μ
  - Find 0.95 quantile in distribution
- Draw $t_\mu(x_{obs})$ contour
- Accept $t_\mu(x_{obs}) < t_{c}$

Note: jagged behavior is due to discrete nature of t, not limited toy statistics

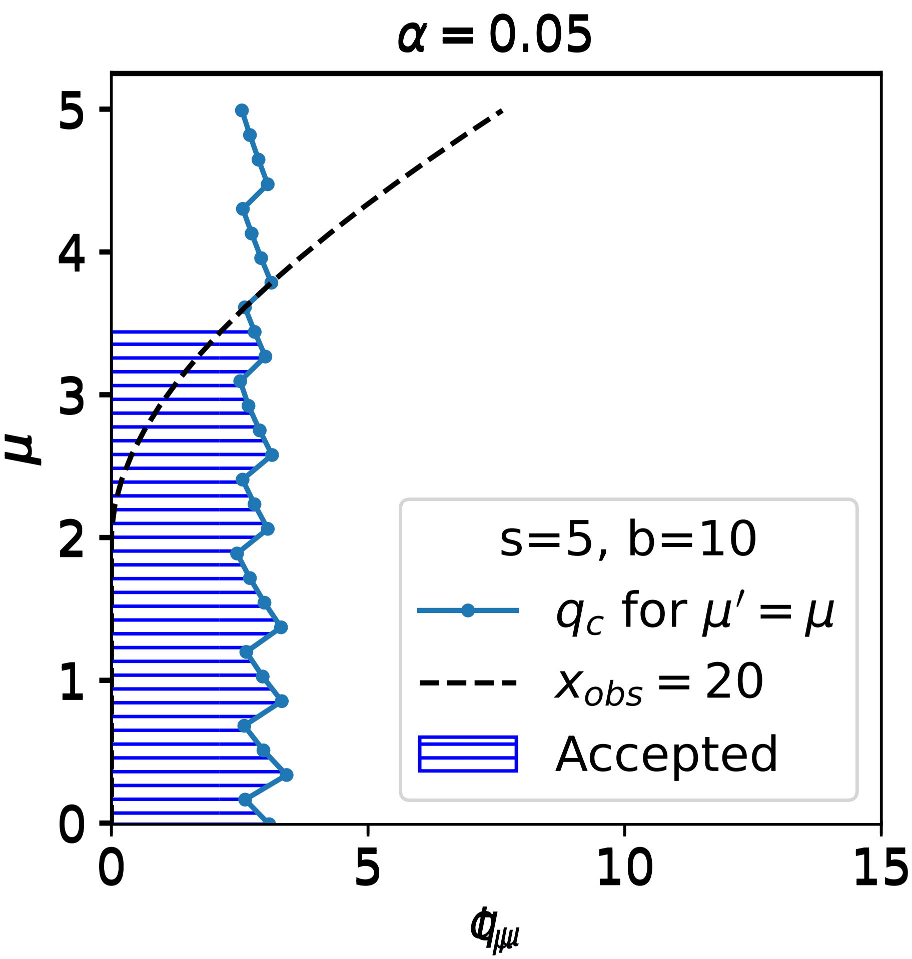




$$t_\mu = -2\ln\frac{\mathscr{L}(\mu)}{\mathscr{L}(\hat{\mu})}$$

Likelihood ratio examples

- Poisson with background example
  - $P(x; \mu s + b) = \dfrac{(\mu s + b)^x e^{-(\mu s + b)}}{x!}$
  - s=5, b=10 fixed, x=20
- Plan: set upper limit on μ
- Problem: two-sided region
  - We should not consider $\hat{\mu} > \mu$ to indicate less compatibility with a model that assumes a rate μ.
- Solution: modify test statistic
  - Define $q_\mu = -2 \ln \dfrac{\mathscr{L}(\mu)}{\mathscr{L}(\min(\mu, \hat{\mu}))}$
  - i.e. over-fluctuations are "not extreme"

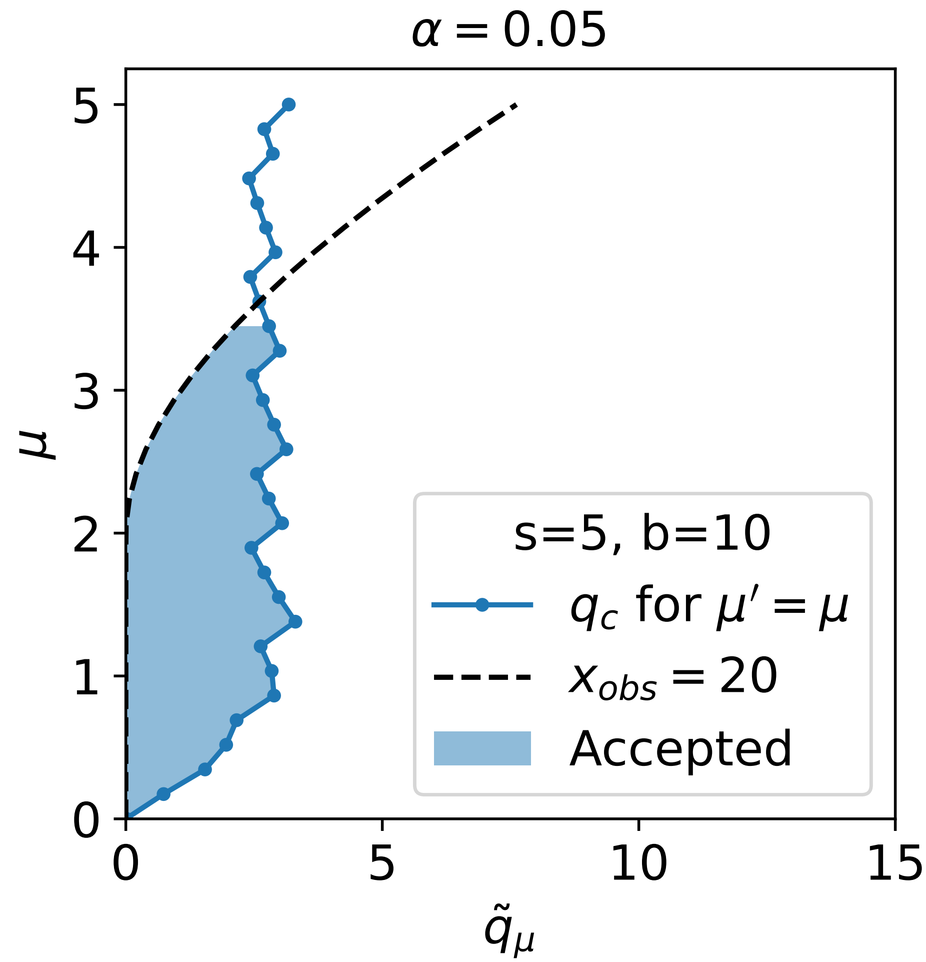


$$t_\mu = -2 \ln \frac{\mathscr{L}(\mu)}{\mathscr{L}(\hat{\mu})}$$

Likelihood ratio examples

- Poisson with background example
  - $P(x \mid \mu s + b) = \dfrac{(\mu s + b)^x e^{-(\mu s + b)}}{x!}$
  - s=5, b=10 fixed, x=20

- Problem: negative $\hat{\mu}$
  - Test stat distribution at 0 should collapse
    - Define $\tilde{q}_\mu = -2 \ln \dfrac{\mathscr{L}(\mu)}{\mathscr{L}(\max(0, \min(\mu, \hat{\mu})))}$
    - i.e. restrict $\hat{\mu}$ to be positive

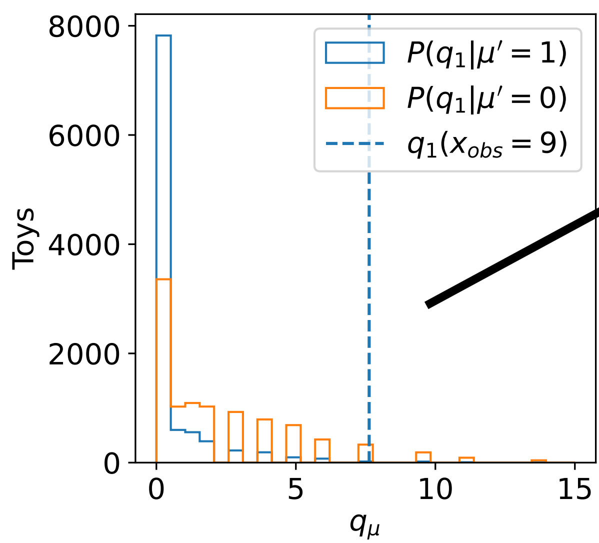


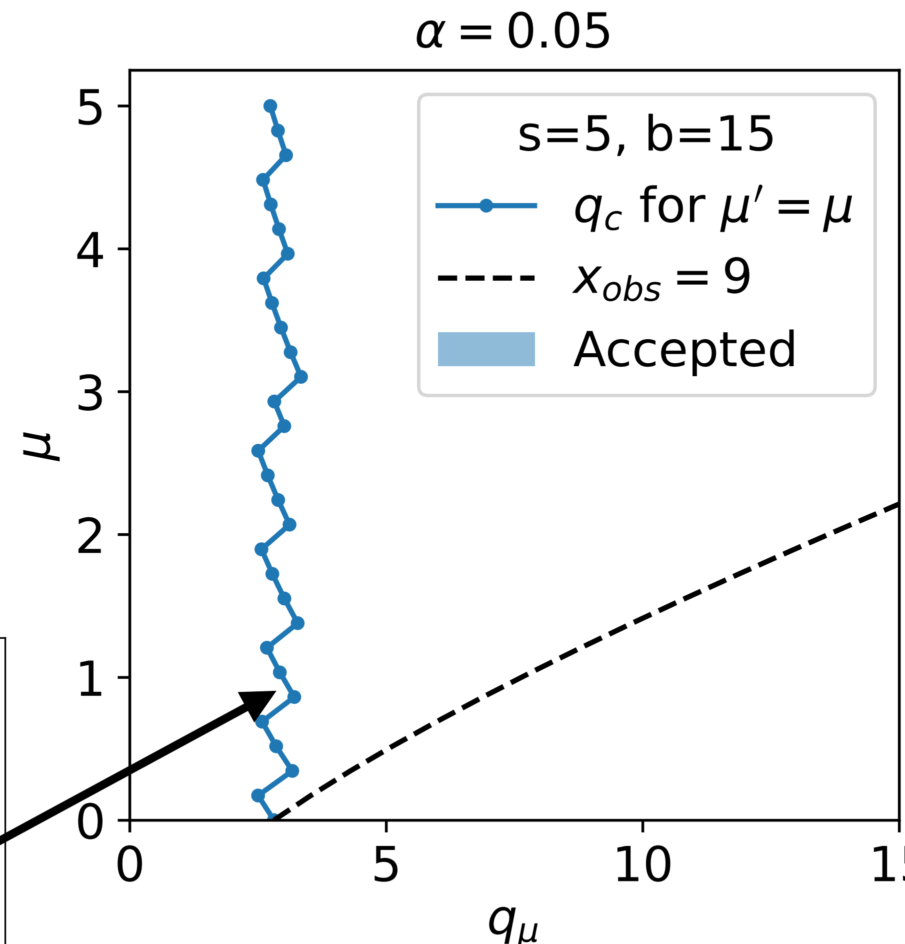
Fermilab

Likelihood ratio examples

- Same example as before, but b=15, x=9
- Problem: under fluctuation
  - No values accepted!
  - Possible but unsatisfying outcome of frequentist test

The result x=9 is rare for both S+B and B-only hypotheses.



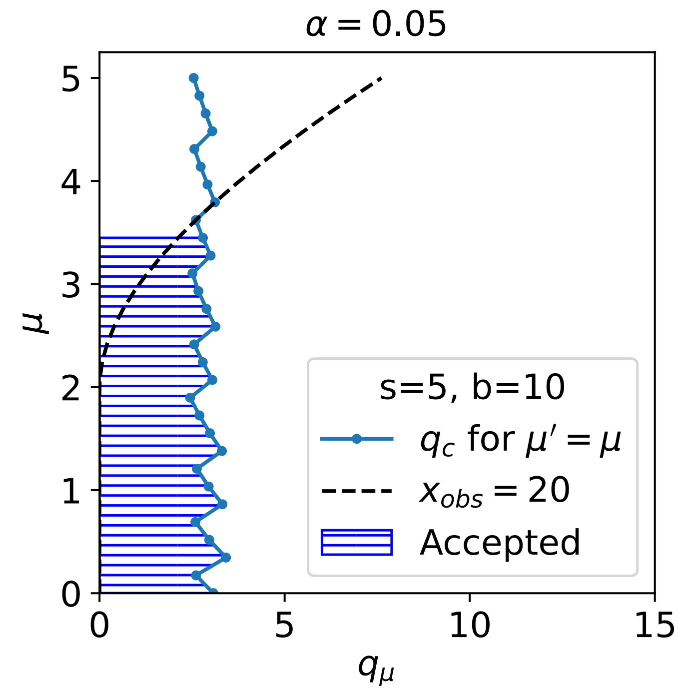


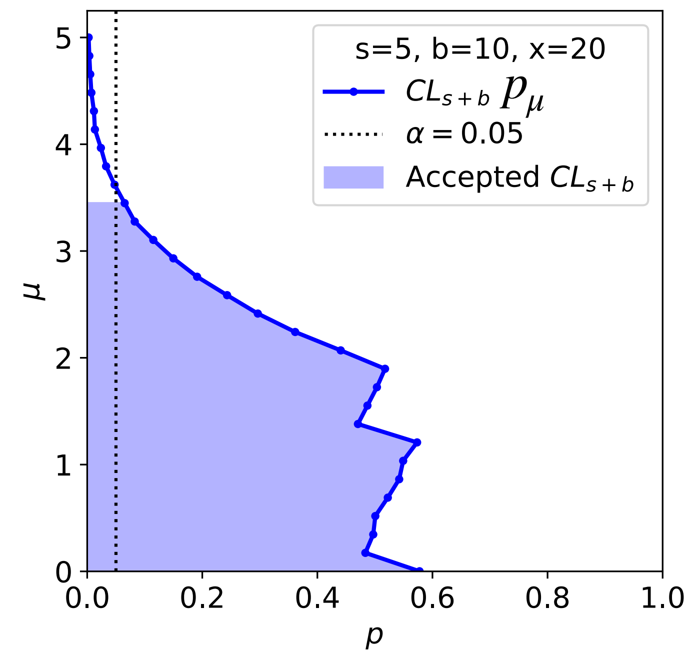
$$q_\mu = -2\ln\frac{\mathscr{L}(\mu)}{\mathscr{L}(\min(\mu,\hat{\mu}))}$$

CLs

- CL_s criterion departs from purely frequentist CL to ameliorate the null set problem (among others)
  - Original expositions by A. Read, T. Junk
  - See also PDG 40.4.2.4

- First we reformulate our old test:
  - Define $p_\mu = \int_{t_\mu(x_{obs})}^{\infty} P(t_\mu \,|\, \mu' = \mu)\, dt_\mu$
    - This is a p-value
  - Then we accept the region $p_\mu > \alpha$
    - Right: initial S+B example reformulated

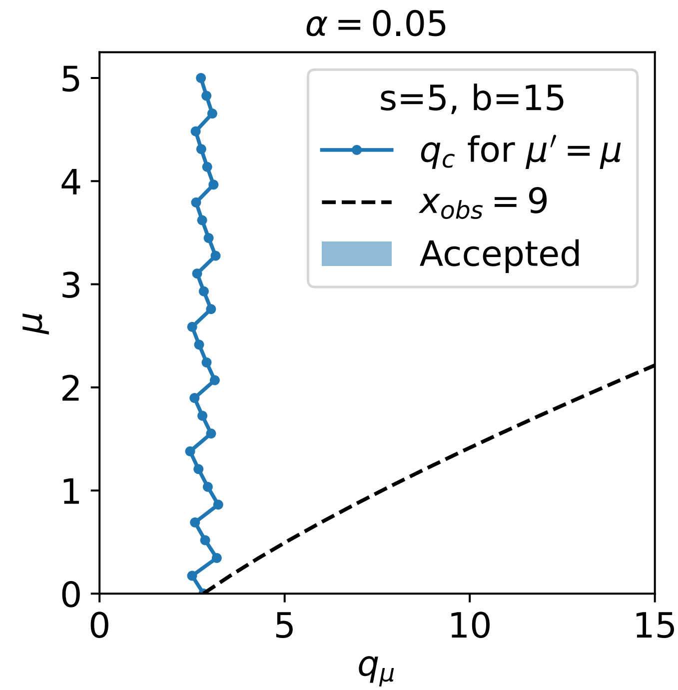


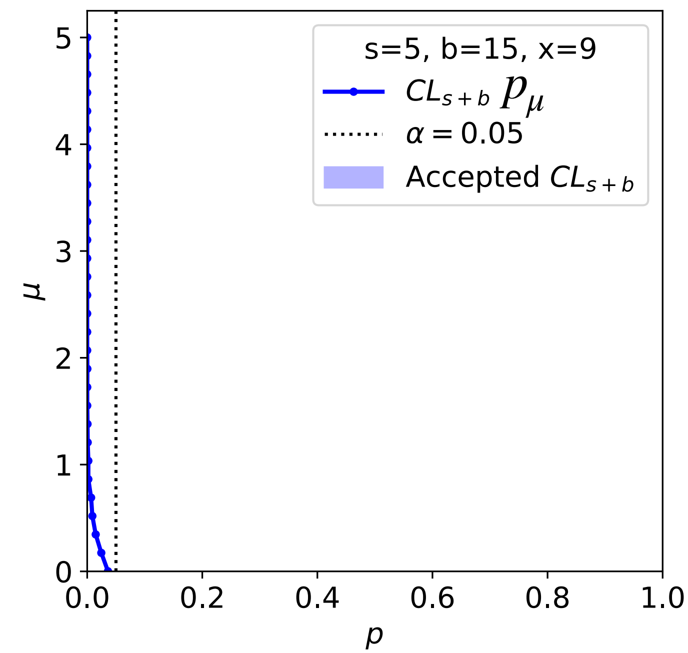


CLs

- CL_s criterion departs from purely frequentist CL to ameliorate the null set problem (among others)
  - Original expositions by A. Read, T. Junk
  - See also PDG 40.4.2.4

- First we reformulate our old test:
  - Define $p_\mu = \int_{t_\mu(x_{obs})}^{\infty} P(t_\mu \,|\, \mu' = \mu)\, dt_\mu$
    - This is a p-value
  - Then we accept the region $p_\mu > \alpha$
    - Right: under-fluctuation S+B example reformulated





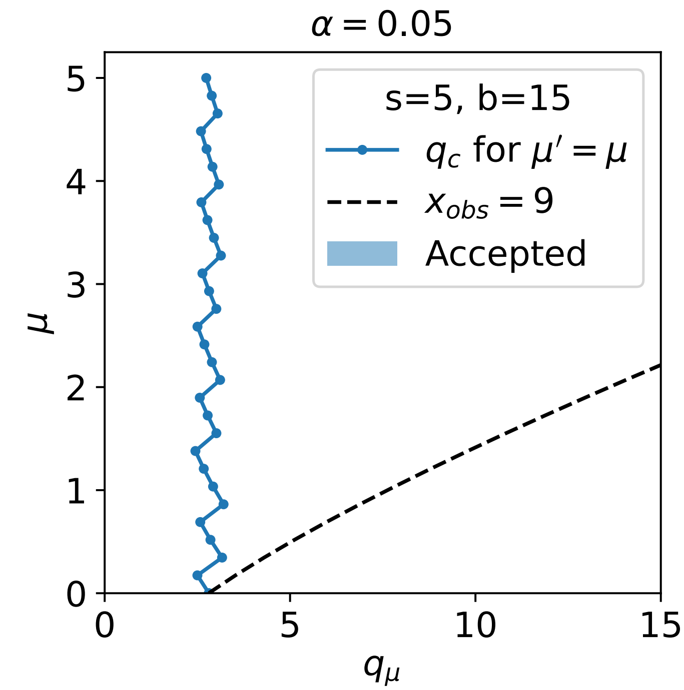
CLs

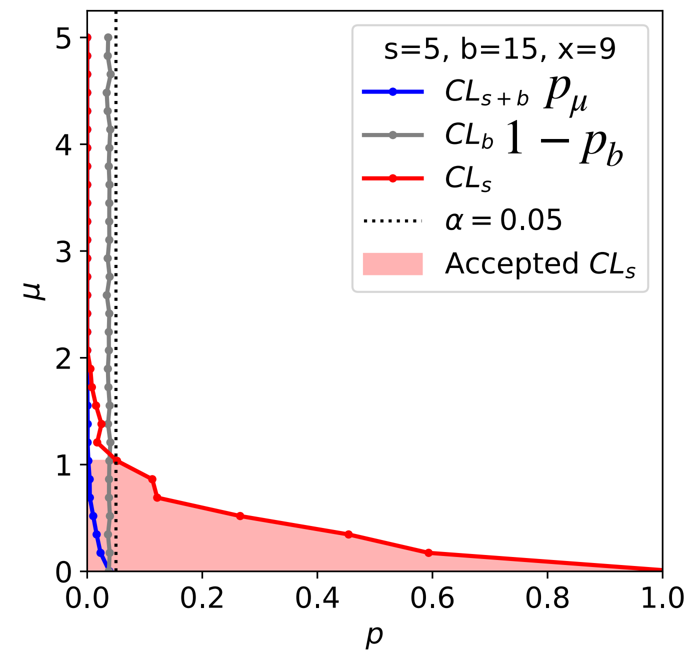
- CL_s criterion departs from purely frequentist CL to ameliorate the null set problem (among others)
  - Original expositions by A. Read, T. Junk
  - See also PDG 40.4.2.4

- First we reformulate our old test:
  - Define $p_\mu = \int_{t_\mu(x_{obs})}^{\infty} P(t_\mu \,|\, \mu' = \mu)\, dt_\mu$

- Now define background-only p-value
  - $1 - p_b = \int_{t_\mu(x_{obs})}^{\infty} P(t_\mu \,|\, \mu' = 0)\, dt_\mu$

- Accept instead $CL_s = \dfrac{p_\mu}{1 - p_b} > \alpha$





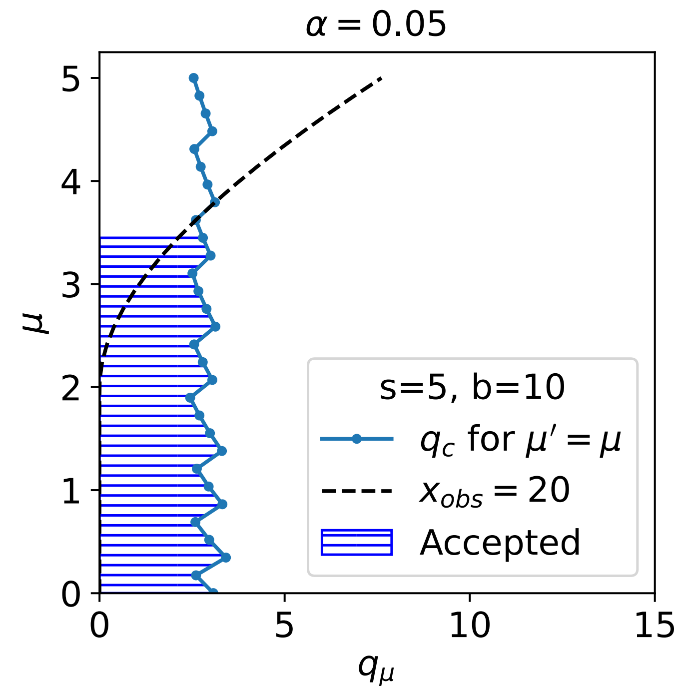
CLs

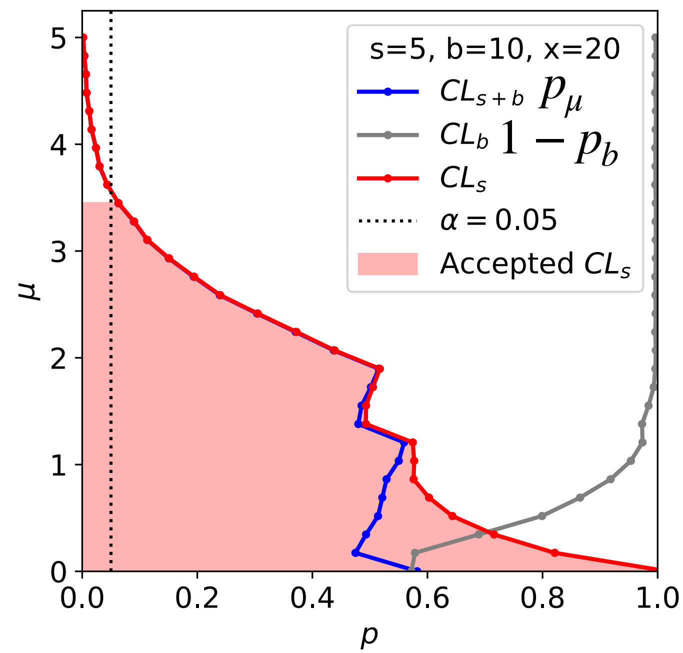
- CL_s criterion departs from purely frequentist CL to ameliorate the null set problem (among others)
  - Original expositions by A. Read, T. Junk
  - See also PDG 40.4.2.4

- First we reformulate our old test:
  - Define $p_\mu = \int_{t_\mu(x_{obs})}^{\infty} P(t_\mu \,|\, \mu' = \mu)\, dt_\mu$

- Now define background-only p-value
  - $1 - p_b = \int_{t_\mu(x_{obs})}^{\infty} P(t_\mu \,|\, \mu' = 0)\, dt_\mu$

- Accept instead $CL_s = \dfrac{p_\mu}{1 - p_b} > \alpha$
  - No effect in the first example



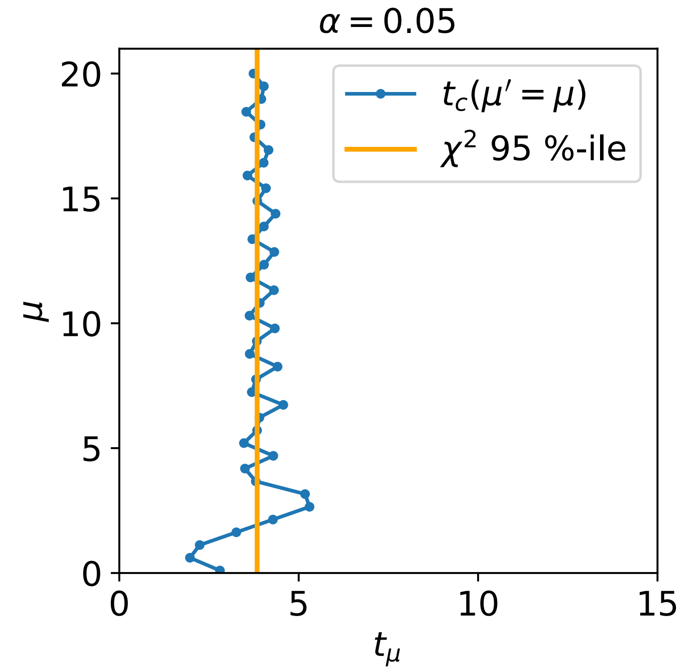


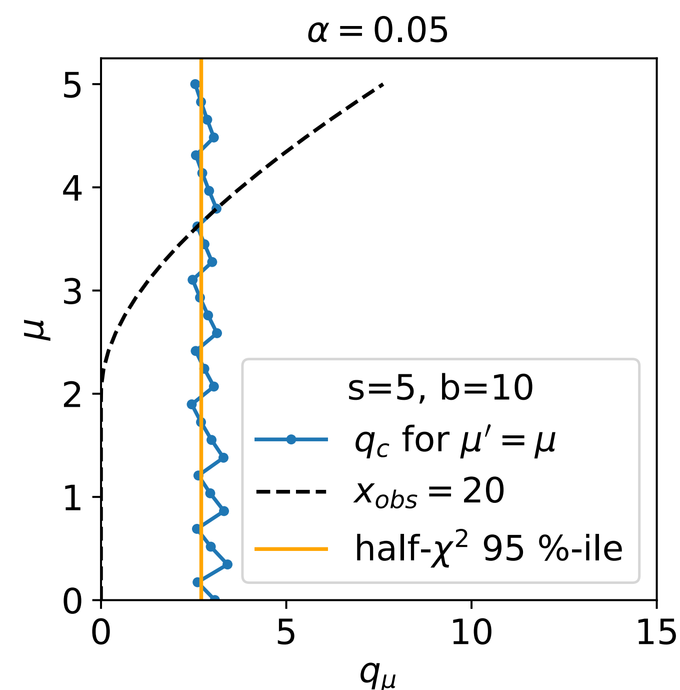
Asymptotic behavior

- Notice how $t_{\mu,c}$ and $q_{\mu,c}$ tend towards a constant?

- This is Wilk's theorem in action
  - Statement: as sample size grows, the distribution of the likelihood ratio P(t_θ|θ') approaches a χ^2 distribution
    - With df = dim(θ)
  - Hence we can approximate by just evaluating $t_\theta(x_{obs})$!

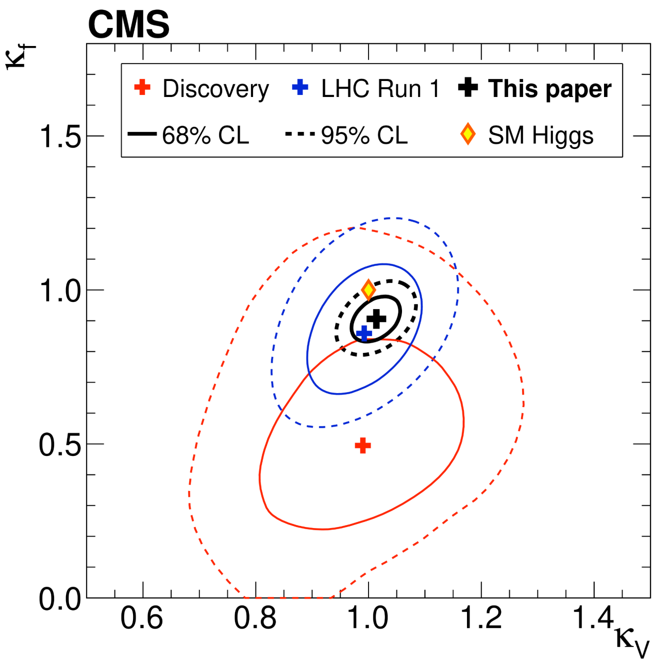
- For q, formulas slightly more complex
  - CCGV provide the recipe: non-central half-χ^2
  - The non-centrality is found using the *Asimov* dataset
    - A special x_μ for a given μ such that $\hat{\mu}(x) = \mu$
    - Note for Poisson data, it may be non-integral!
    - This dataset produces the median expected limit

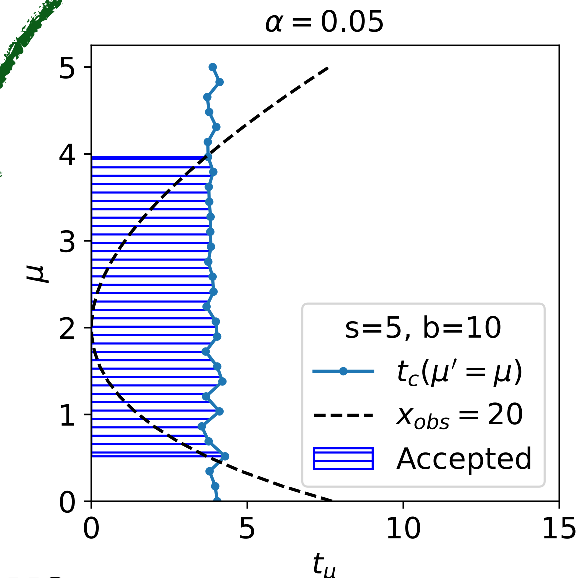


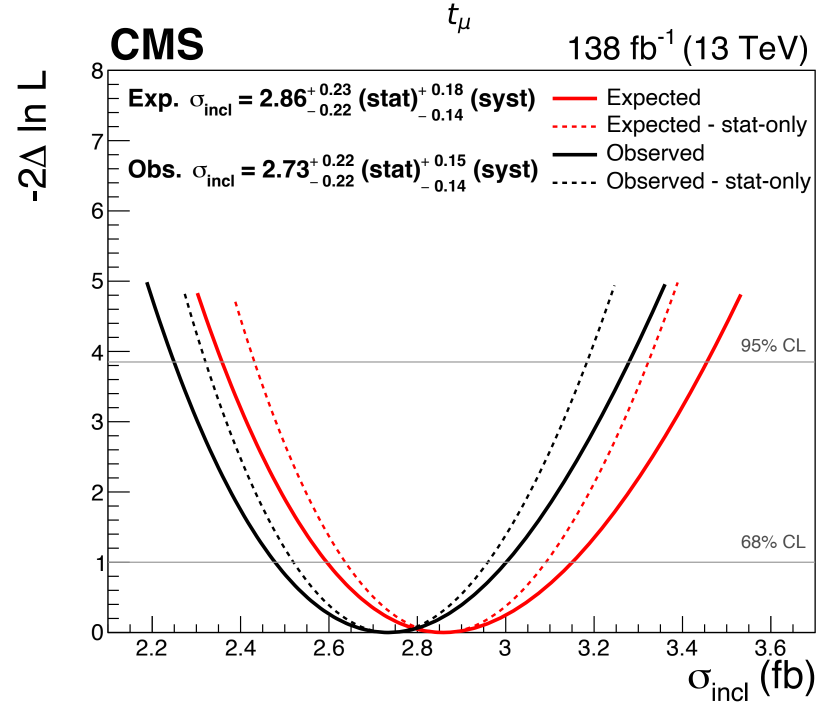


Asymptotic behavior

- This is how we make deltaNLL contours



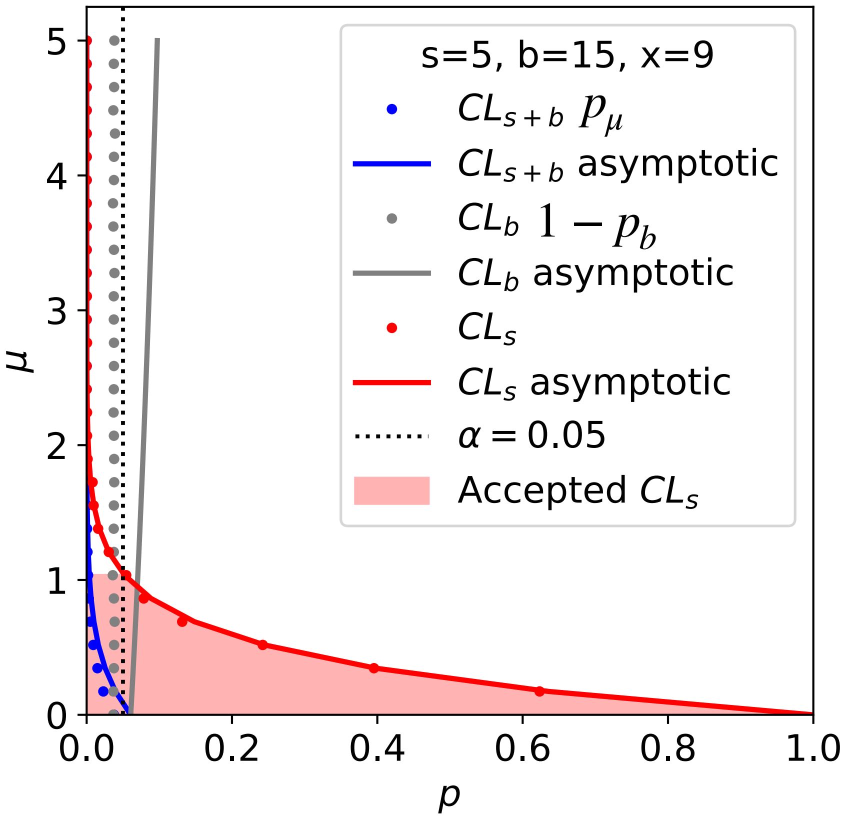


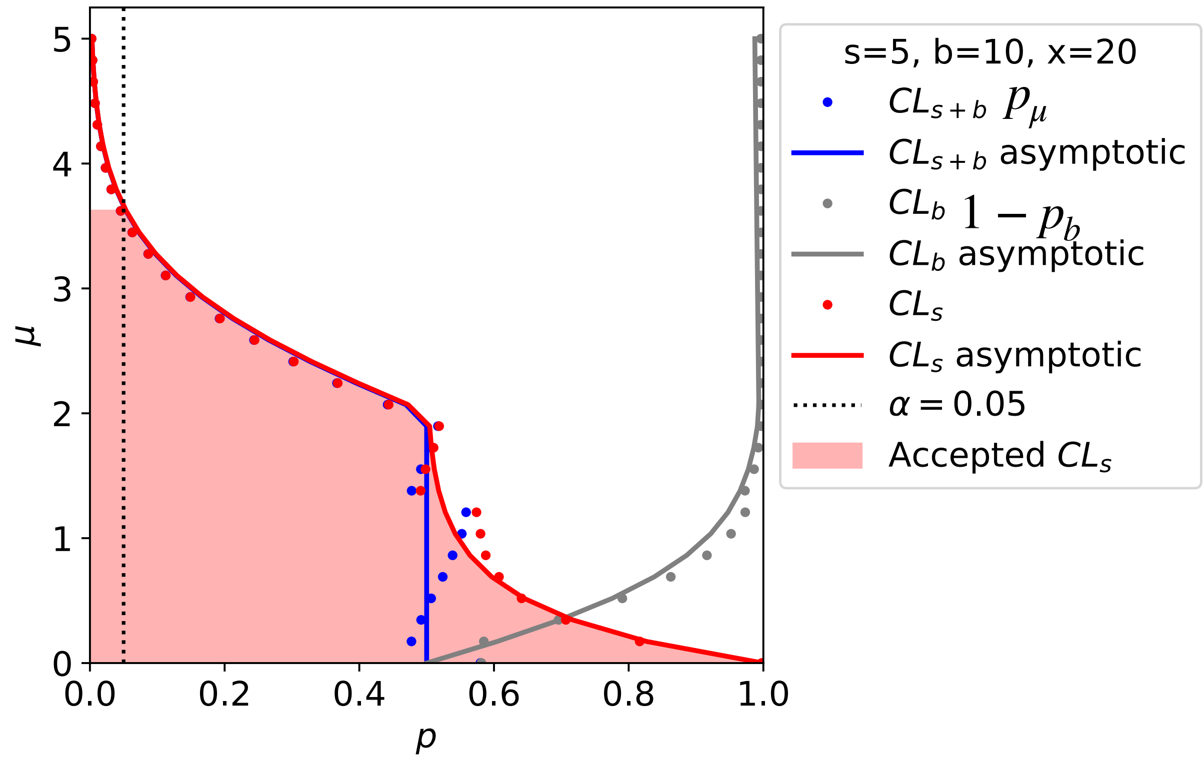


`scipy.stats.chi2.ppf(q, df)`

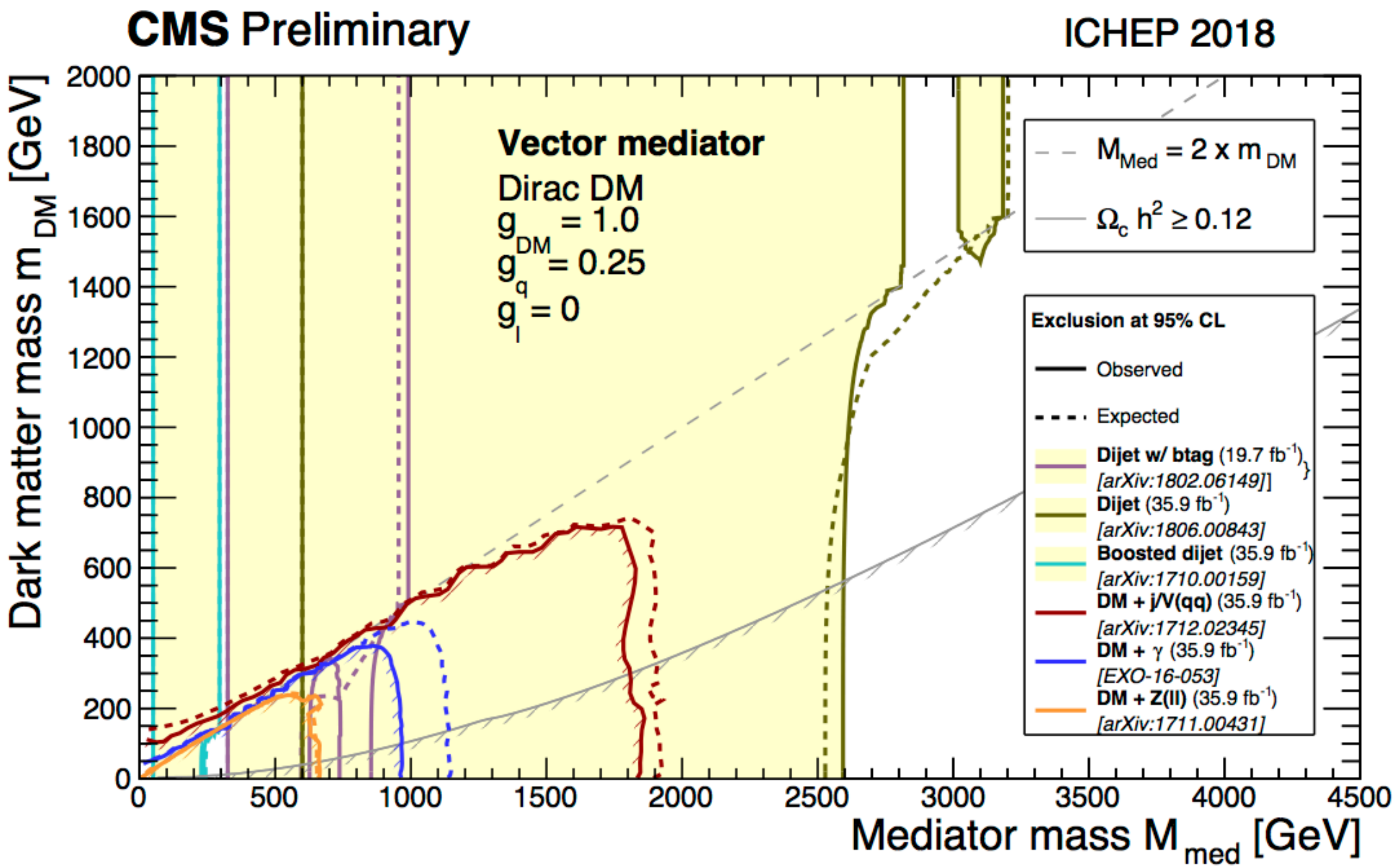
Quantile	t_c	
	df=1	***df=2***
0.68	***0.989***	***2.279***
1σ (0.6827…)	***1***	***2.296***
0.95	***3.841***	***5.991***
2σ (0.9545…)	***4***	***6.180***

Asymptotic CLs





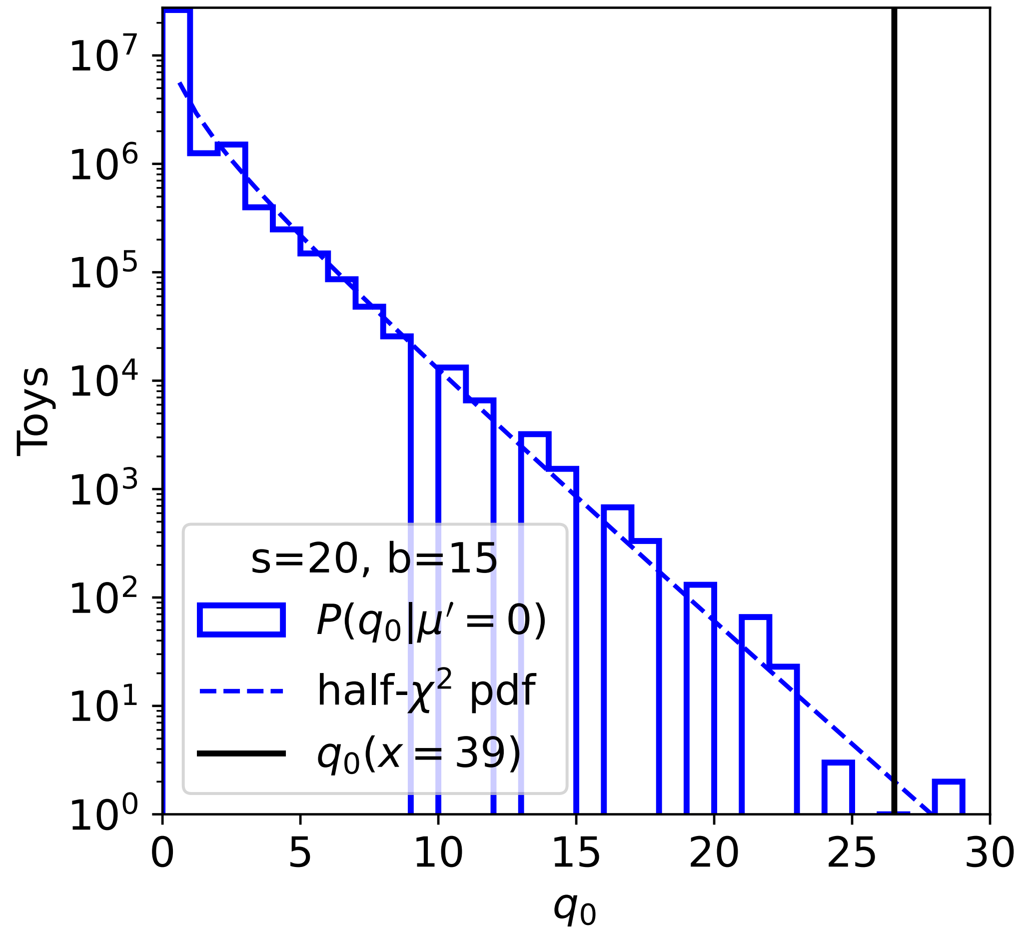
Examples



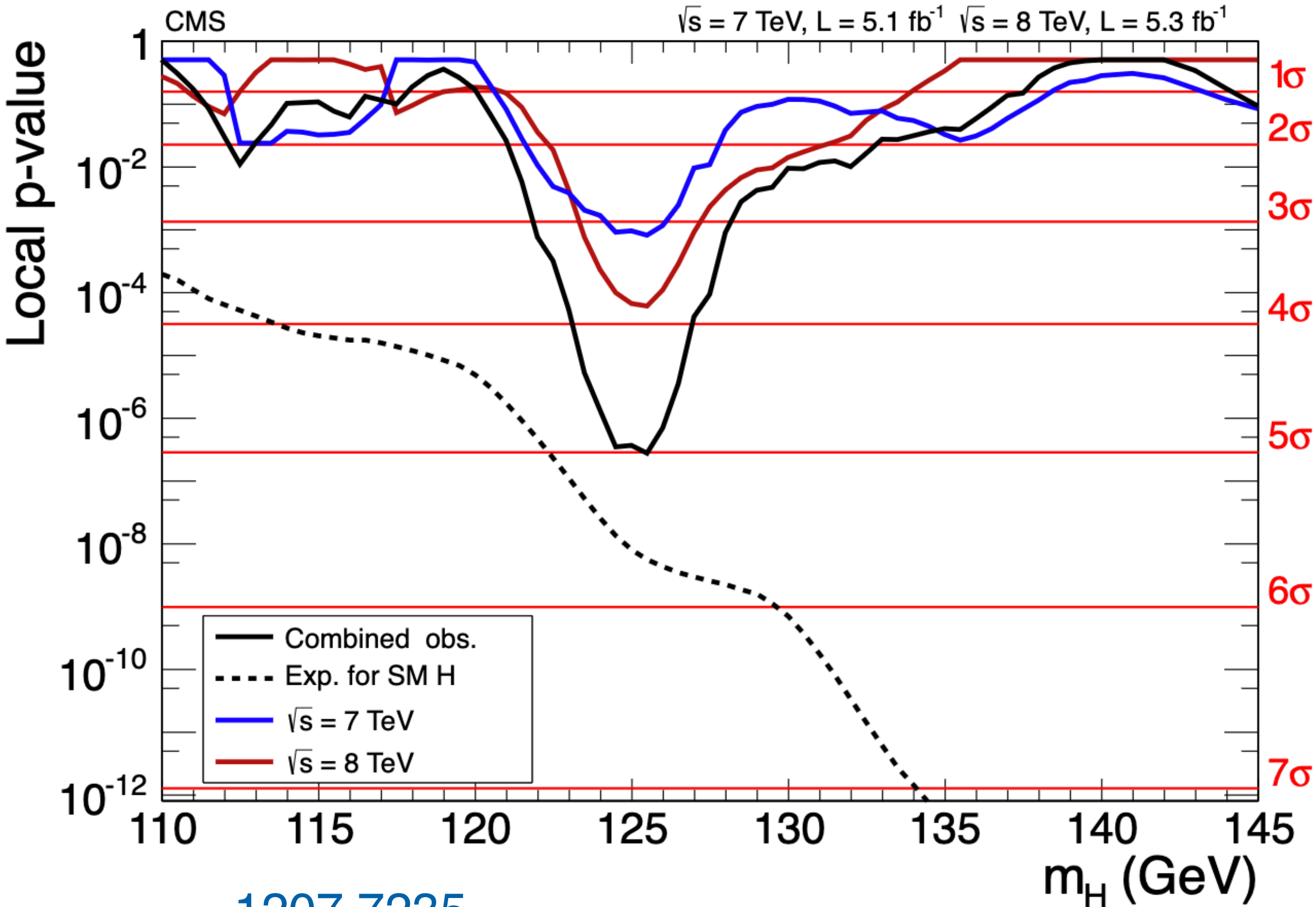
CMS Preliminary
ICHEP 2018
Vector mediator
Dirac DM
$g_{DM} = 1.0$
$g_q = 0.25$
$g_l = 0$
$M_{Med} = 2 \times m_{DM}$
$\Omega_c h^2 \geq 0.12$
Exclusion at 95% CL
Observed
Expected
Dijet w/ btag (19.7 fb^{-1})
[arXiv:1802.06149]]
Dijet (35.9 fb^{-1})
[arXiv:1806.00843]
Boosted dijet (35.9 fb^{-1})
[arXiv:1710.00159]
DM + j/V(qq) (35.9 fb^{-1})
[arXiv:1712.02345]
DM + γ (35.9 fb^{-1})
[EXO-16-053]
DM + Z(ll) (35.9 fb^{-1})
[arXiv:1711.00431]
Dark matter mass m_{DM} [GeV]
Mediator mass M_{med} [GeV]

Test statistic for discovery

- Poisson with background example
  - $f_P(x; \mu s + b) = \dfrac{(\mu s + b)^x e^{-(\mu s + b)}}{x!}$
  - s=20, b=15 fixed, x=39
- Cannot use t_μ:
  - Severe under-fluctuation would count as discovery! Certainly something was discovered, but not an excess over background. Disallow in test statistic:
    - Define $q_0 = -2 \ln \dfrac{\mathscr{L}(0)}{\mathscr{L}(\max(0,\hat{\mu}))}$
    - i.e. under-fluctuations are “not extreme”
  - Deceptively simple result: $Z = \sqrt{q_0(x_{obs})}$
    - Only true if one POI



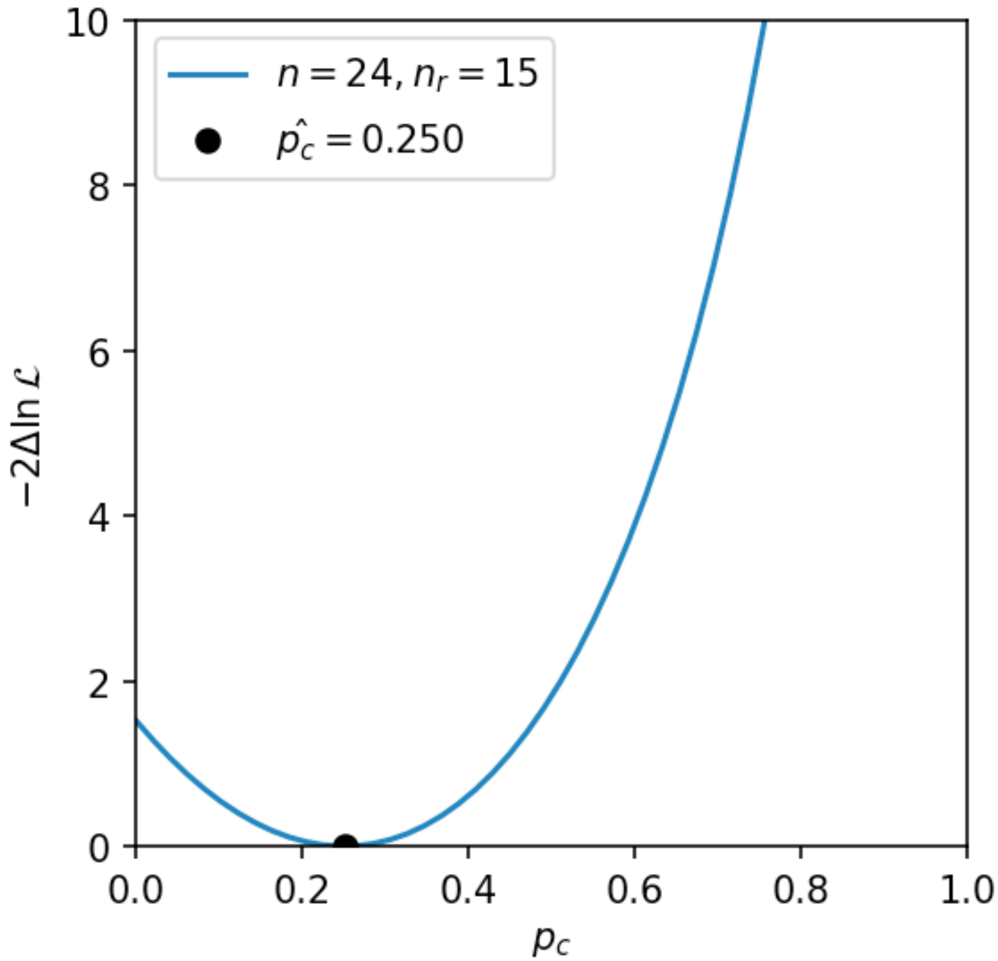
Discovery example



[1207.7235](1207.7235)

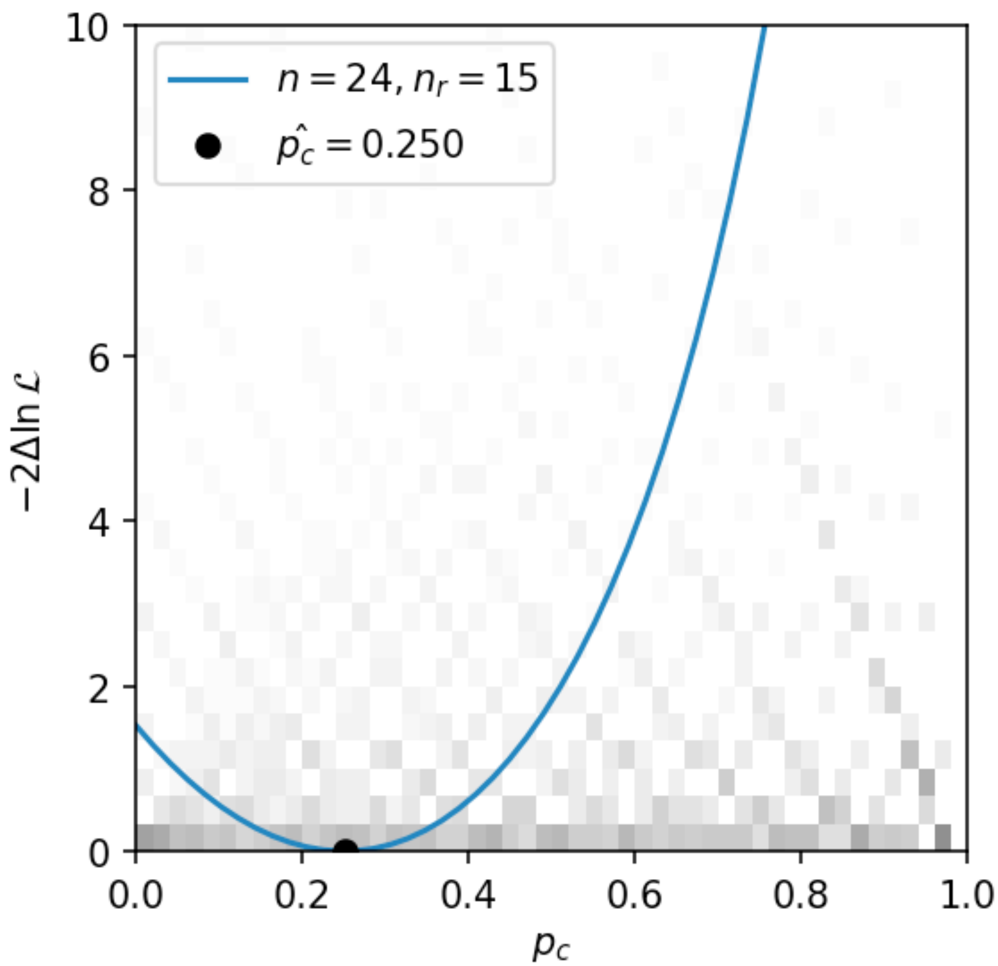
Intervals for P(cheat)

- Coming back to our favorite problem



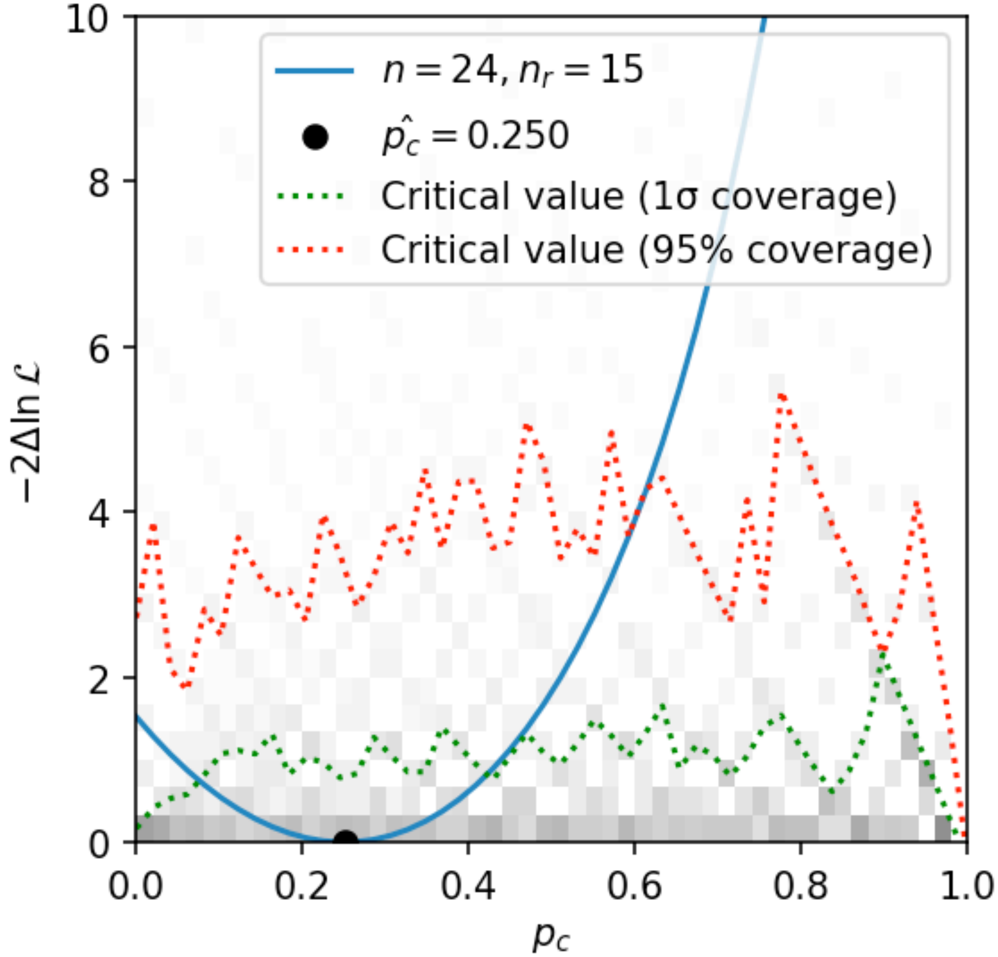
Intervals for P(cheat)

- Coming back to our favorite problem
  - For no-cheat null hypothesis, $p_0 \approx 0.2$, Z score: 0.88σ



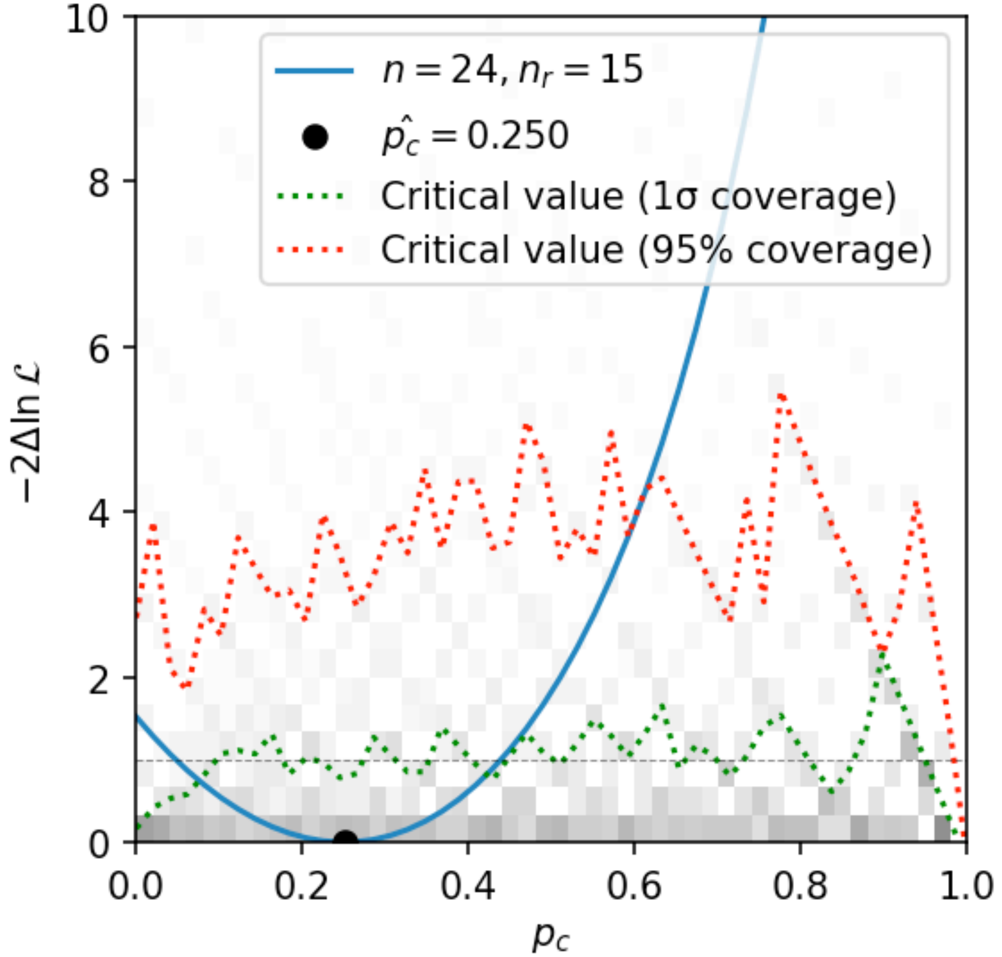
Intervals for P(cheat)

- Coming back to our favorite problem
  - For no-cheat null hypothesis, $p_0 \approx 0.2$, Z score: 0.88σ
  - 1σ central interval: $0.08 < p_c < 0.41$



Intervals for P(cheat)

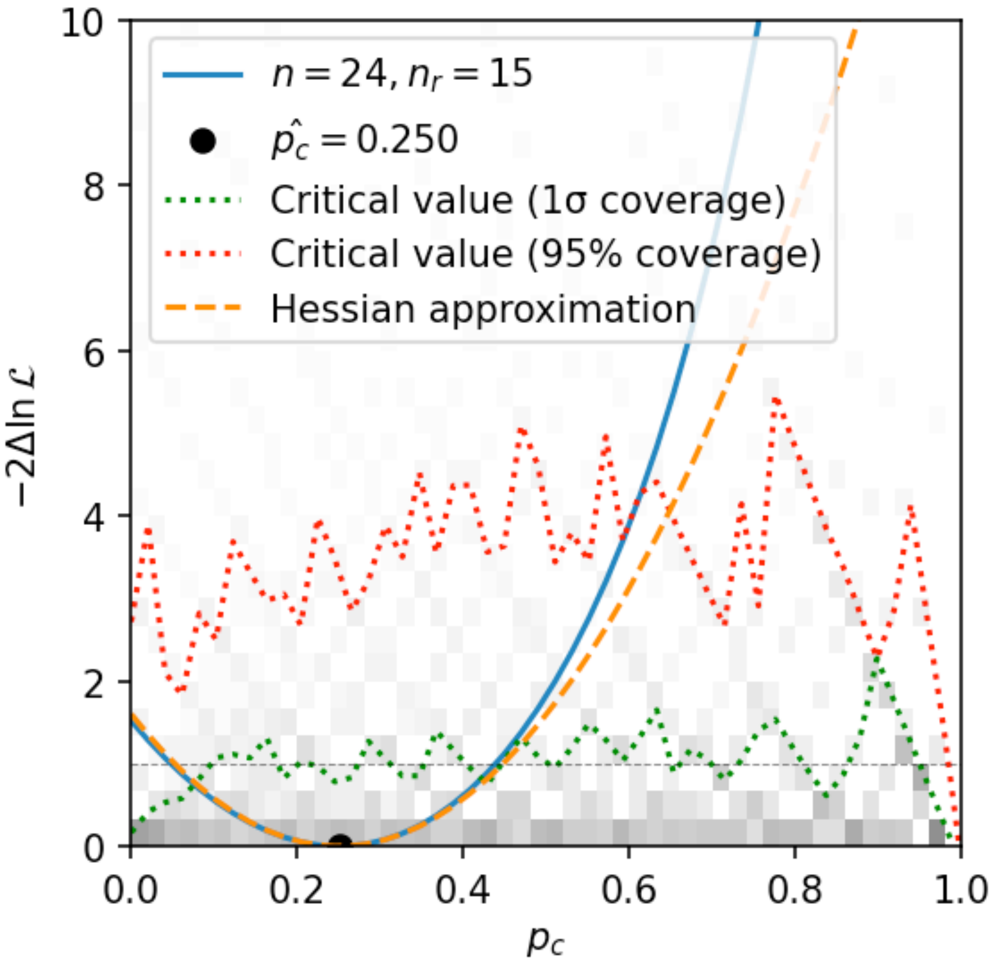
- Coming back to our favorite problem
  - For no-cheat null hypothesis, $p_0 \approx 0.2$, Z score: 0.88σ
  - 1σ central interval: $0.08 < p_c < 0.41$



Fermilab

Intervals for P(cheat)

- Coming back to our favorite problem
  - For no-cheat null hypothesis, $p_0 \approx 0.2$, Z score: 0.88σ
  - 1σ central interval: $0.08 < p_c < 0.41$



Fermilab