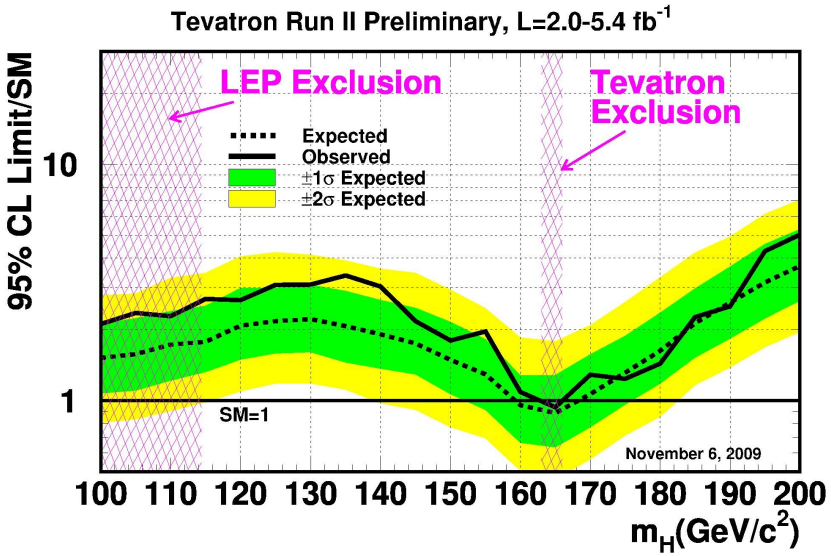
Limits, Discovery & Systematics



Wade Fisher

Michigan State University

May 17 2010

Nuts and Bolts of Hypothesis Testing

- Tools of the trade
  - Nuisance parameters & hypothesis testing
  - Semi-Frequentist Limits
  - Bayesian Limits
  - Discovery tests
- Nuances of systematic uncertainties
  - Priors!
  - Source: theory vs experimental
- Impact & presentation
  - How much do systematic uncertainties matter?
  - Alternative presentations

Models & Parameters

- Compound vs Simple hypotheses
  - Simple: The model probability distribution is fully specified. Eg, a Gaussian with width μ=3, σ=1.2
  - Compound: Model parameters are unknown. Eg, a Gaussian with width μ=3, σ=?
- Nuisance parameter
  - A parameter intrinsic to your model, but not the parameter you wish to test.
  - Example: We wish to test for the presence or absence of a new particle, but all event rate predictions depend on the total integrated luminosity being analyzed.
  - Operationally speaking, to test outcomes of a compound model, you must tranform to a simple hypothesis by assigning the nuisance parameter a prior (*ie*, some pre-specified probability distribution).

Hypothesis Testing

- The goal: Use data to distinguish two hypotheses
  - **H0** ⇨ null hypothesis: background-only model, eg Standard Model
  - **H1** ⇨ a test hypothesis: presence of a new particle, coupling, etc.

H0 is a compound hypothesis, with some set of nuisance parameters

H1 has the same form, maybe with some extra model parameters and nuisance parameters

Simple Example: **H0** describes the Standard Model background expectation for the result of an analysis. Nuisance parameters can be luminosity, acceptance, t-tbar cross section, etc...

H1 is the same as **H0**, but add a new physics signal. The model can be parametrized by mass / cross section / etc, and extra nuisance parameters come from signal acceptance, model parameters, etc...

Complicating Factors

- The hypotheses, **H1** & **H0**, are often subdivided according to final states with unique signatures
  - Orthogonal search channels defined to maximize acceptance, isolate high S/B regions, etc.
- The null hypothesis, **H0**, is the sum of several contributing Standard Model processes
  - Nuisance parameters are generally correlated amongst backgrounds (and signals!), but not always
- Discriminant distributions for expected and observed events are binned into histograms
  - Discriminating variable can be an observable (eg, angular distribution) or multi-variable function (Neural Nets, etc).
  - Uncertainties may affect rates, shapes of discriminating variable, or both. Many are asymmetric and are not necessarily Gaussian.

Semi-Frequentist Limits

- Given hypotheses, **H1** & **H0**, we must simulate outcomes of repeated experiments

Assume data is drawn randomly from a Poisson parent distribution

⇒ Generate pseudo-data via random Poisson with mean value from expected backgrounds (**H0**) or signal-plus-background (**H1**)

Systematics are a tricky Frequentist problem, so use a Bayesian model

⇒ Model uncertainties on nuisance parameters as Gaussian-distributed, sample randomly for each pseudo-experiment

⇒ Vary nominal background prediction according to smeared values of nuisance parameters, change mean of random Poisson each time

Each pseudo-experiment can be evaluated via some test statistic sensitive to the presence of signal: Neyman & Pearson say the LLR is best choice.

Pseudo-Experiments

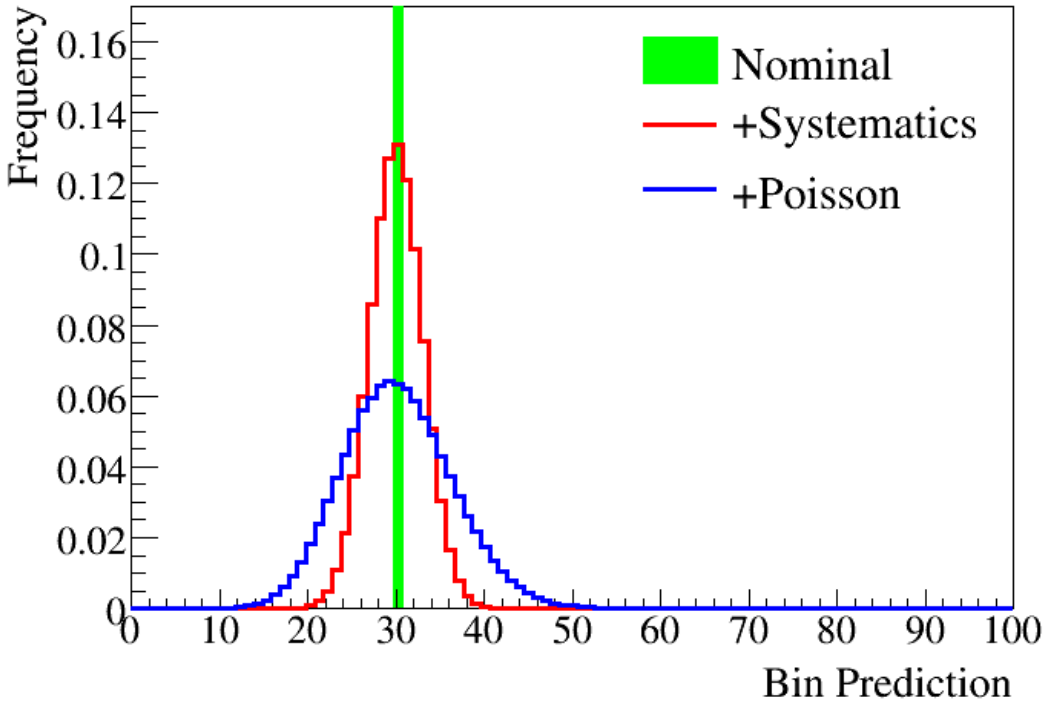
- The Semi-Frequentist model depends largely on how we generate pseudo-experiments and what we do with them.

For each bin of your histogrammed discriminant, can predict the number of observed events based on our construction

- For example:

  Monte Carlo predicts 30 events

  10% systematic uncertainty



Ordering Outcomes

- Each pseudo-experiment is evaluated against **H0** and **H1**

Construct LLR from the ratio of Poisson likelihoods:

$$Q(\vec{s},\vec{b},\vec{d}) = \prod_{i=0}^{N_C}\prod_{j=0}^{N_{bins}} \frac{(s+b)_{ij}^{d_{ij}} e^{-(s+b)_{ij}}}{d_{ij}!} \Big/ \frac{b_{ij}^{d_{ij}} e^{-b_{ij}}}{d_{ij}!}$$

$$LLR(\vec{s},\vec{b},\vec{d}) = -2\mathrm{Log}(Q) = \sum_{i=0}^{N_C}\sum_{j=0}^{N_{bins}} s_{ij} - d_{ij}\ln\left(1+\frac{s_{ij}}{b_{ij}}\right)$$

The Profile Likelihood

- To counteract the degrading effects of uncertainties on nuisance parameters, we begin by defining the Profile Likelihood

Likelihood becomes a function of signal, bkgd, data, and nuisance parameters

Maximizing the Profile Likelihood to a set of data points defines our "best fit" for that data in a given hypothesis

$$Q = \frac{L(x|\theta_{R1}, \hat{\theta}_S)}{L(x|\theta_{R0}, \hat{\hat{\theta}}_S)}$$

Two independent likelihood maximizations are performed over nuisance parameters parameters, one for each pseudo-experiment

θ_{R1} θ_{R0} : Physics parameters in ***H1*** and ***H0***, respectively

$\hat{\theta}_S$: Nuisance parameters which maximize L for ***H1***

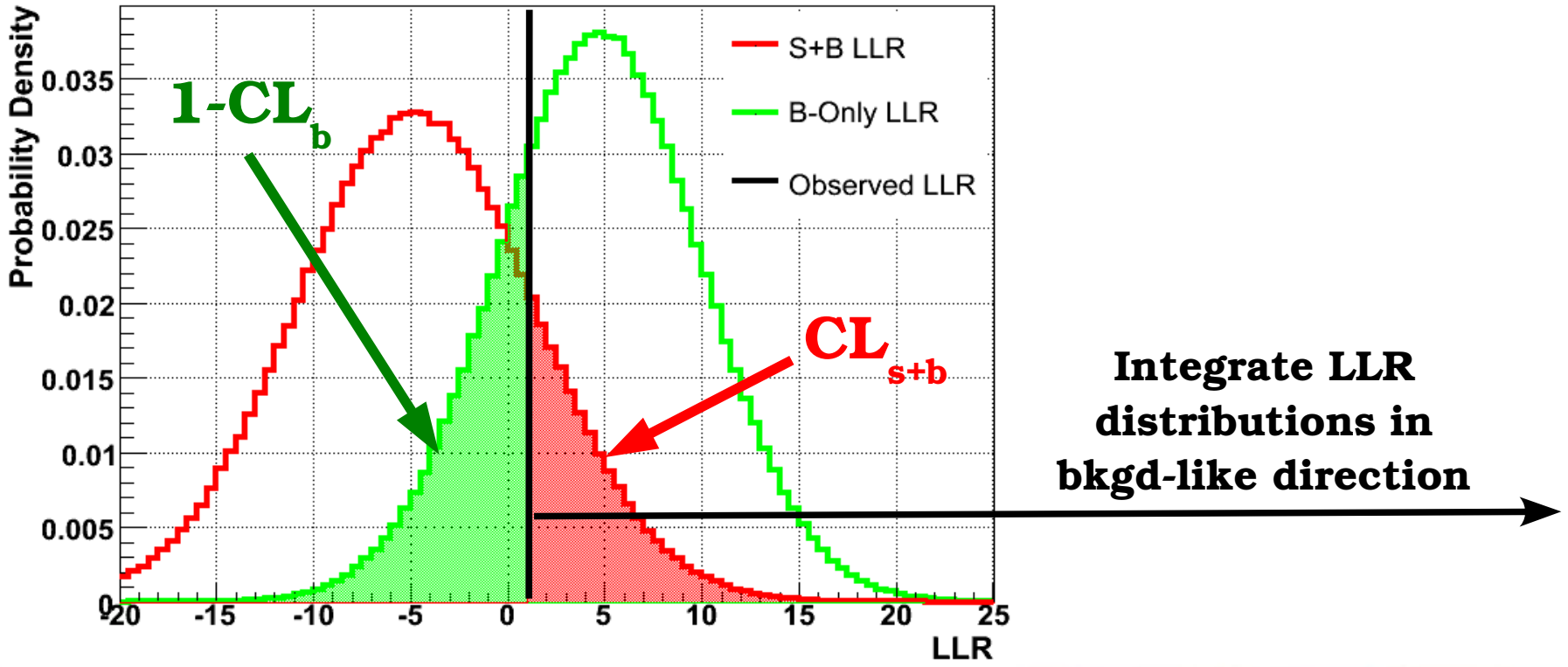
$\hat{\hat{\theta}}_S$: Nuisance parameters which maximize L for ***H0***

Ordering Outcomes

- Generate pseudo-experiments for both **H0** and **H1** and order each according to LLR test statistic.
  - Confidence levels (p-Values) based on frequency of outcomes

CL_b = fraction of **H0** pseudo-experiments less signal-like than data

CL_{sb} = fraction of **H1** pseudo-experiments less signal-like than data

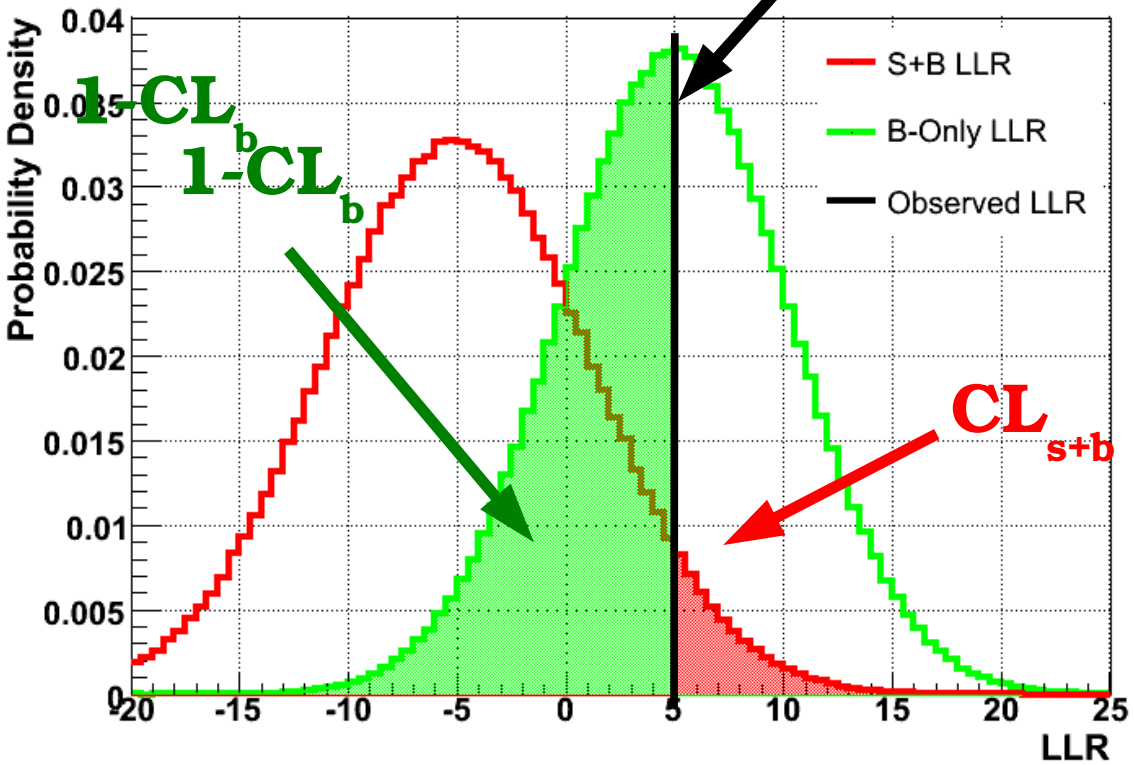


Expected limits

- Expected limits assume data = median expected background outcome

CL_b = 50% if background is well-modeled

"Expected" Data = Median Bkgd Outcome



The CLs Statistic

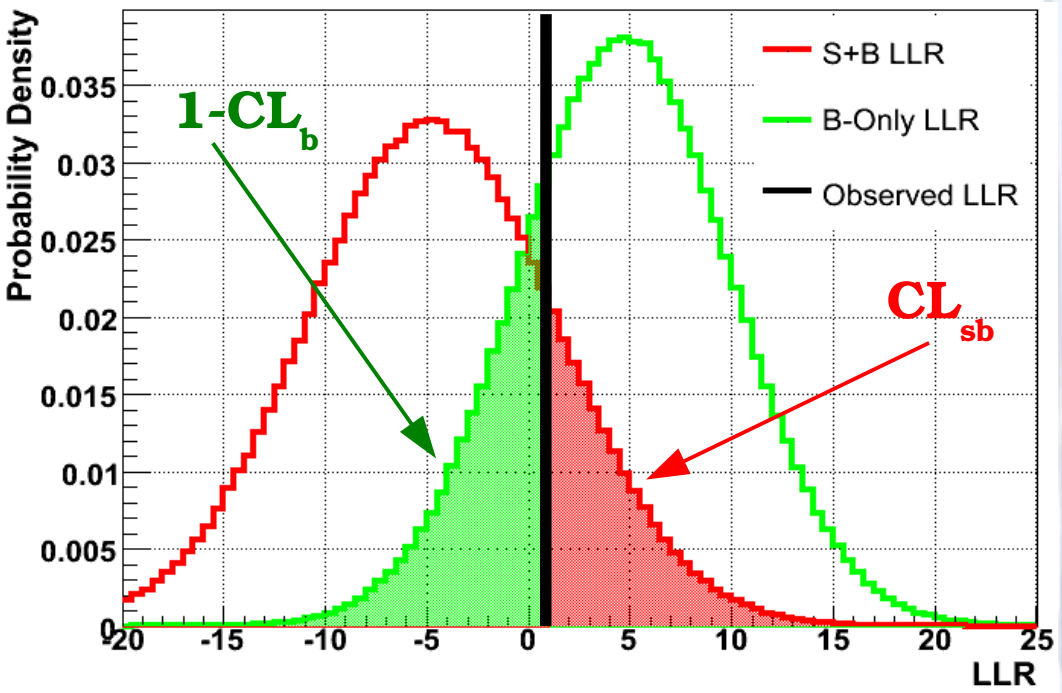
- Given our confidence levels for our two hypotheses ($\mathbf{CL_{sb}}$ and $\mathbf{CL_b}$), we want to describe confidence intervals relative to specific outcomes

A strict Frequentist definition would use $\mathbf{CL_{sb}}$ to define confidence interval

- The CLs prescription introduces an inherit dependence on the background model description

$$CL_s = \frac{CL_{s+b}}{CL_b}$$

New interval defined for 1-CLs = 1-α



The CLs Statistic

- Given our confidence levels for our two hypotheses ($\mathbf{CL_{sb}}$ and $\mathbf{CL_b}$), we want to describe confidence intervals relative to specific outcomes

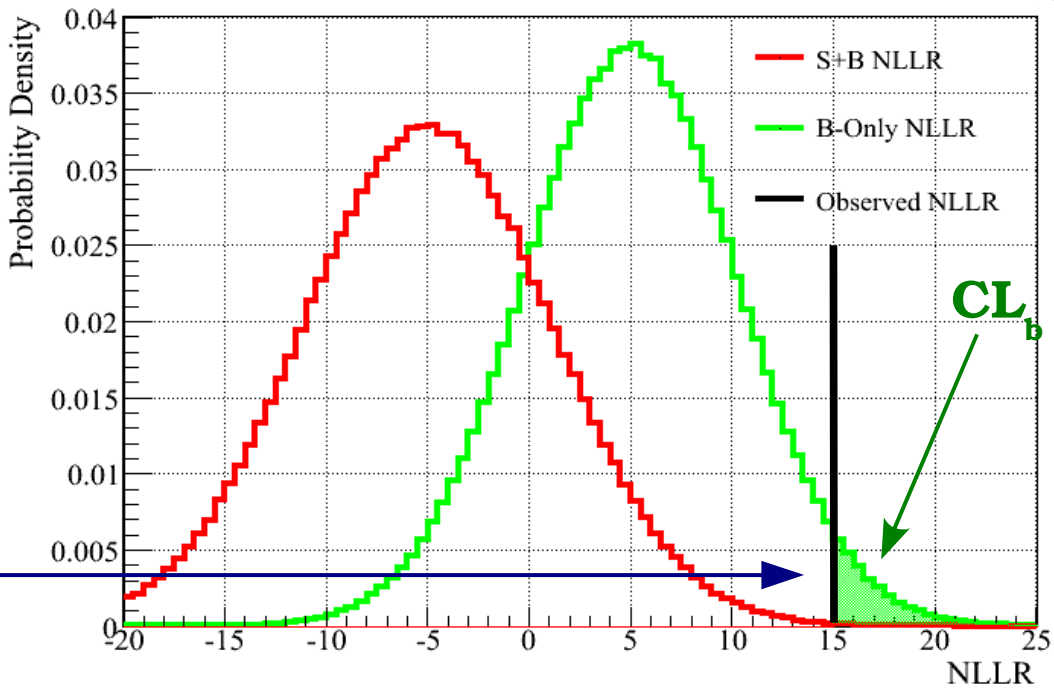
A strict Frequentist definition would use $\mathbf{CL_{sb}}$ to define confidence interval

- The CLs prescription introduces an inherit dependence on the background model description

$$CL_s = \frac{CL_{s+b}}{CL_b}$$

New interval defined for 1-CLs = 1-α

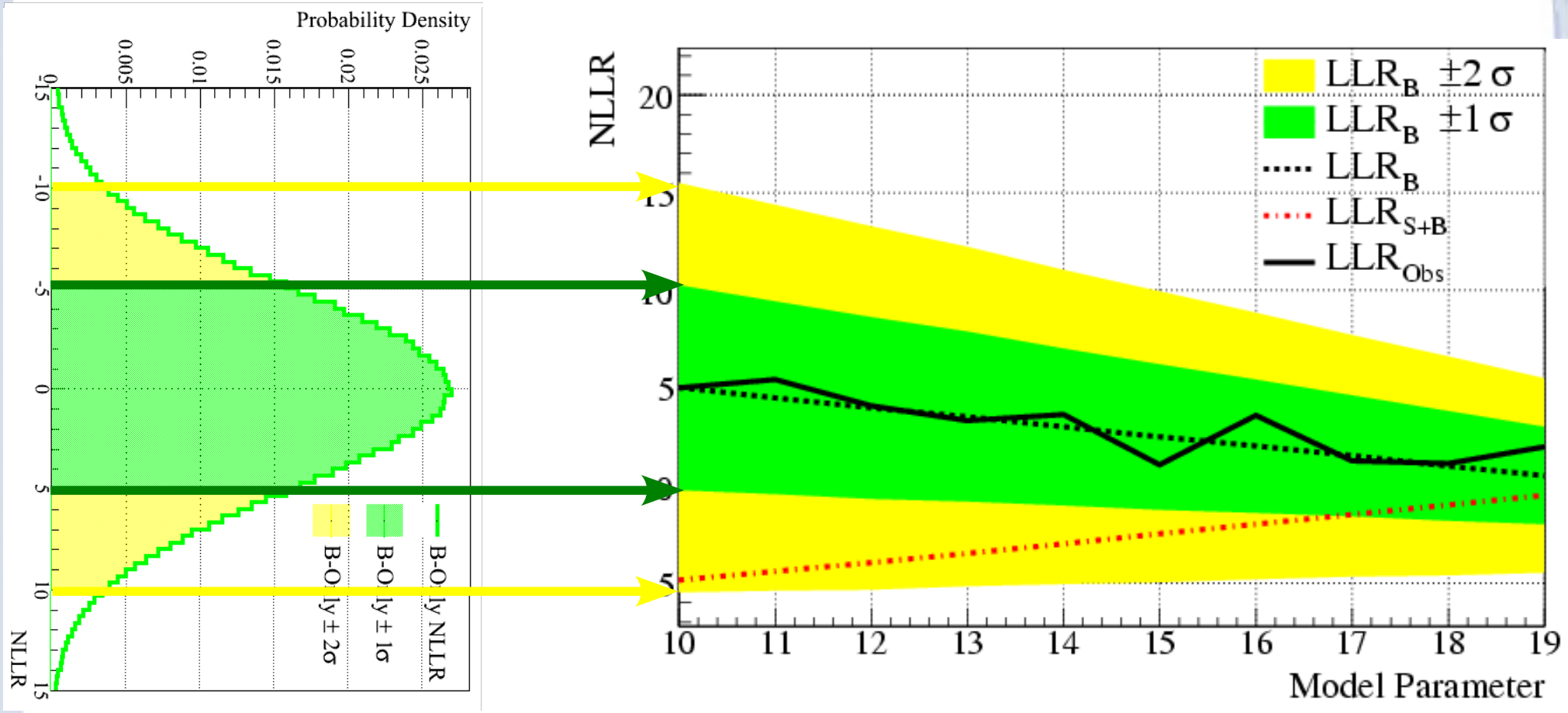
Provides protection against poor background modeling



LLR Distributions

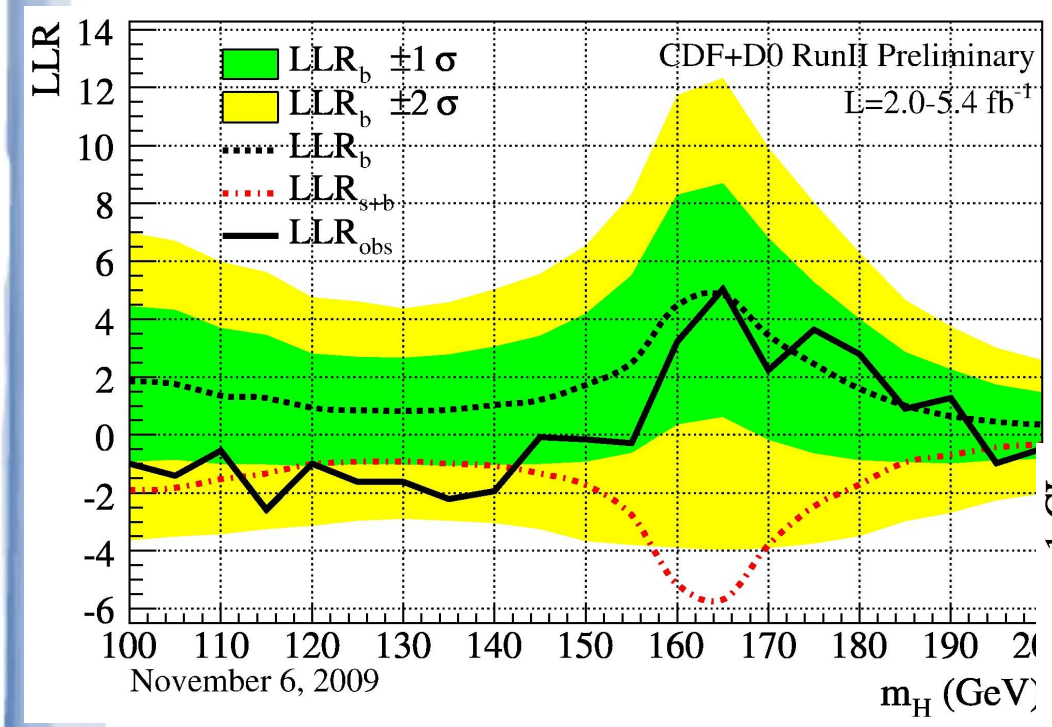
- We commonly plot LLR values as a function of a model variable
  Eg, Higgs mass

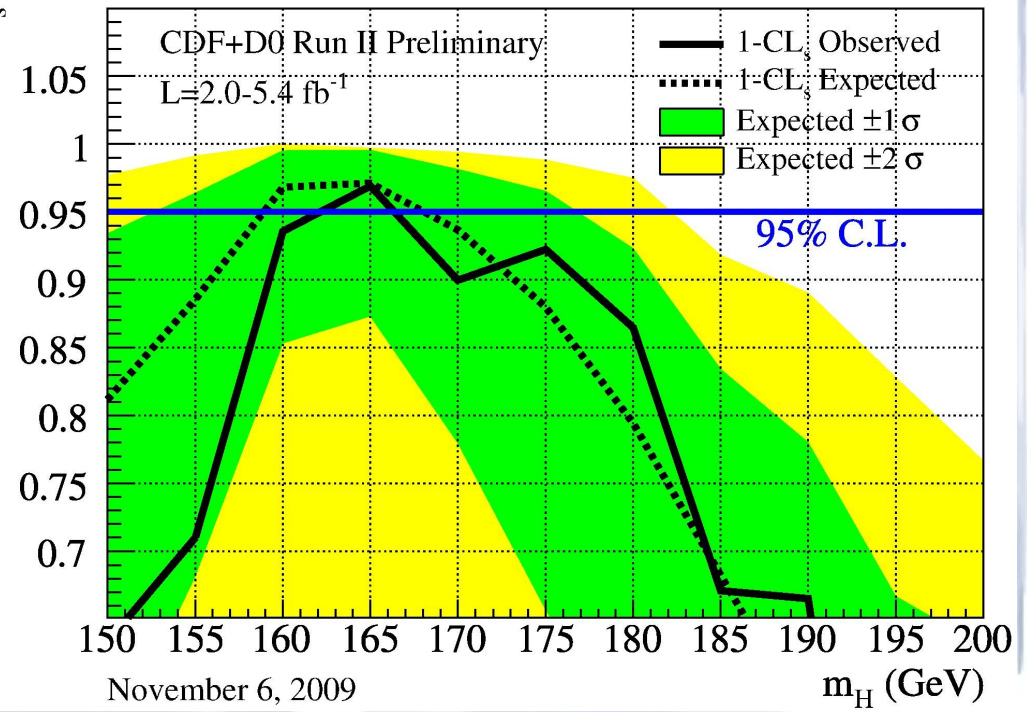
These are just "overhead" views of LLR distributions at each Higgs mass



Tevatron Plots

- Examples from the Tevatron Higgs search





Bayesian Limits

- Begin by defining a joint likelihood over input channels (searches) and histogram bins:

$$L(r,\theta) = \prod_{channels} \prod_{bins} P_{Poisson}(data|r,\theta)$$

  - here **r** is a signal rate scaling factor and **θ** is the set of all nuisance parameters for the model

$$P_{Poisson}(data|r,\theta) = \frac{(r \times s_i(\theta) + b_i(\theta))^{n_i} e^{-(r \times s_i(\theta) + b_i(\theta))}}{n_i!}$$

  - Where n_i, s_i, b_i represent the observed data, expected signal and expected background in **bin i** of your histogram.
  - The signal and background predictions may have any arbitrary dependence on the nuisance parameters **θ**

Bayesian Limits

From this morning

Including uncertainties on nuisance parameters θ

$$L'(data \mid r) = \int L(data \mid r, \theta)\pi(\theta)d\theta$$

where $\pi(\theta)$ encodes our prior belief in the values of the uncertain parameters. Usually Gaussian centered on the best estimate and with a width given by the systematic. The integral is high-dimensional. Markov Chain MC integration is quite useful!

Useful for a variety of results:

Limits: $$0.95 = \int_0^{r_{\lim}} L'(data \mid r)\pi(r)dr$$

Typically $\pi(r)$ is constant
Other options possible.
Sensitivity to priors a concern.

Measure r: $$0.68 = \int_{r_{low}}^{r_{high}} L'(data \mid r)\pi(r)dr$$

$$r = r_{\max}{}^{+(r_{high} - r_{\max})}_{-(r_{\max} - r_{low})}$$

Usually: shortest interval containing 68% of the posterior (other choices possible). Use the word "credibility" in place of "confidence"

Bayesian Limits

From this morning

Including uncertainties on nuisance parameters θ

$$L'(data \mid r) = \int L(data \mid r, \theta)\pi(\theta)d\theta$$

where $\pi(\theta)$ encodes our prior belief in the values of the uncertain parameters. Usually Gaussian centered on the best estimate and with a width given by the systematic. The integral is high-dimensional. Markov Chain MC integration is quite useful!

Useful for a variety of results:

Limits: $$0.95 = \int_0^{r_{lim}} L'(data \mid r)\pi(r)dr$$

Typically $\pi(r)$ is constant
Other options possible.
Sensitivity to priors a concern.

Measure r: $$0.68 = \int_{r_{low}}^{r_{high}} L'(data \mid r)\pi(r)dr$$

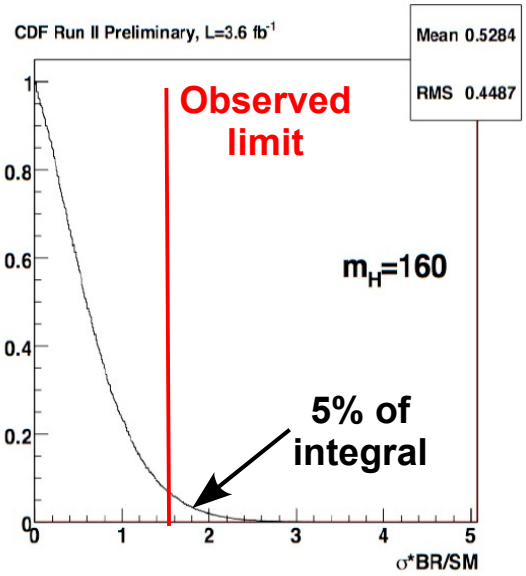
$$r = r_{\max}{}^{+(r_{high} - r_{\max})}_{-(r_{\max} - r_{low})}$$

Usually: shortest interval containing 68% of the posterior (other choices possible). Use the word "credibility" in place of "confidence"

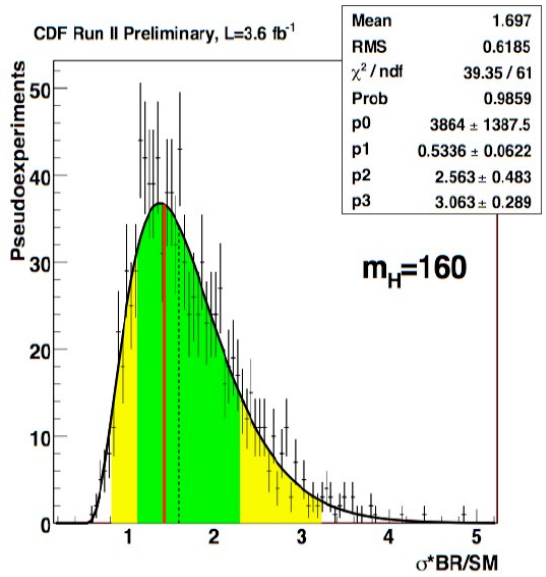
Bayesian Limits

- What's happening?
  - Given the likelihood previously defined, we can define a posterior density for our signal rate, given the observation

$$0.95 = \int_0^{r_{\lim}} L'(data \mid r)\pi(r)dr$$



- What about expected limits?
  - Bayes: information from the extra data you didn't take is meaningless!
  - So we mix in a little of the Frequentist approach by generating pseudo-experiments. Test each and study resulting distribution of outcomes.



Common Standards of Evidence

Physicists like to talk about how many "sigma" a result corresponds to and generally have less feel for p-values.

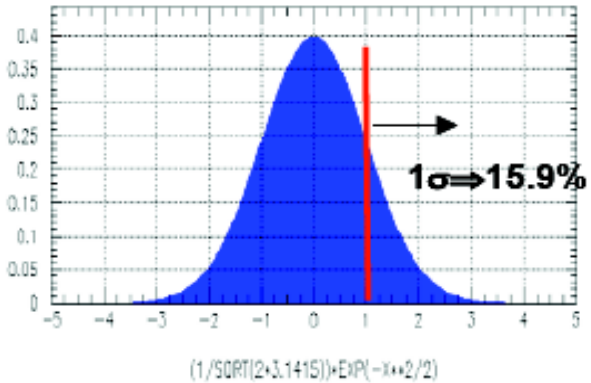
The number of "sigma" is called a "z-value" and is just a translation of a p-value using the integral of one tail of a Gaussian

Double_t zvalue = - TMath::NormQuantile(Double_t pvalue)

Folklore:
95% CL -- good for exclusion
3σ: "evidence"
5σ: "observation"
Some argue for a more subjective scale.

z-value (σ)	p-value
1.0	0.159
2.0	0.0228
3.0	0.00135
4.0	3.17E-5
5.0	2.87E-7

$$pvalue = \frac{\left(1 - erf\left(zvalue/\sqrt{2}\right)\right)}{2}$$

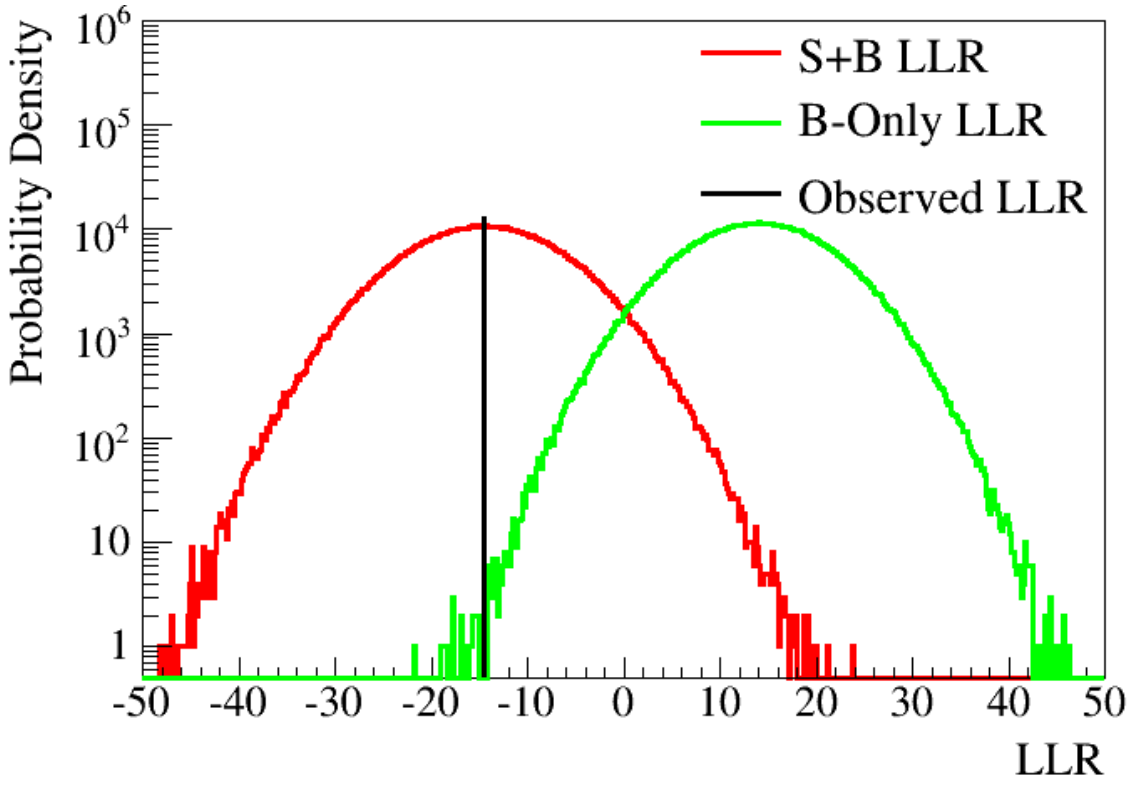


Tip: most physicists talk about p-values now but hardly use the term z-value

Discovery Criteria

- When we realize we cannot exclude, we begin to think about discovering something new
  - Discovery criteria can be thought of as excluding the background-only hypothesis (**H0**).

- Switch statistical tests to **1-CL$_b$**: values nearing zero indicate small probability for background-only hypothesis

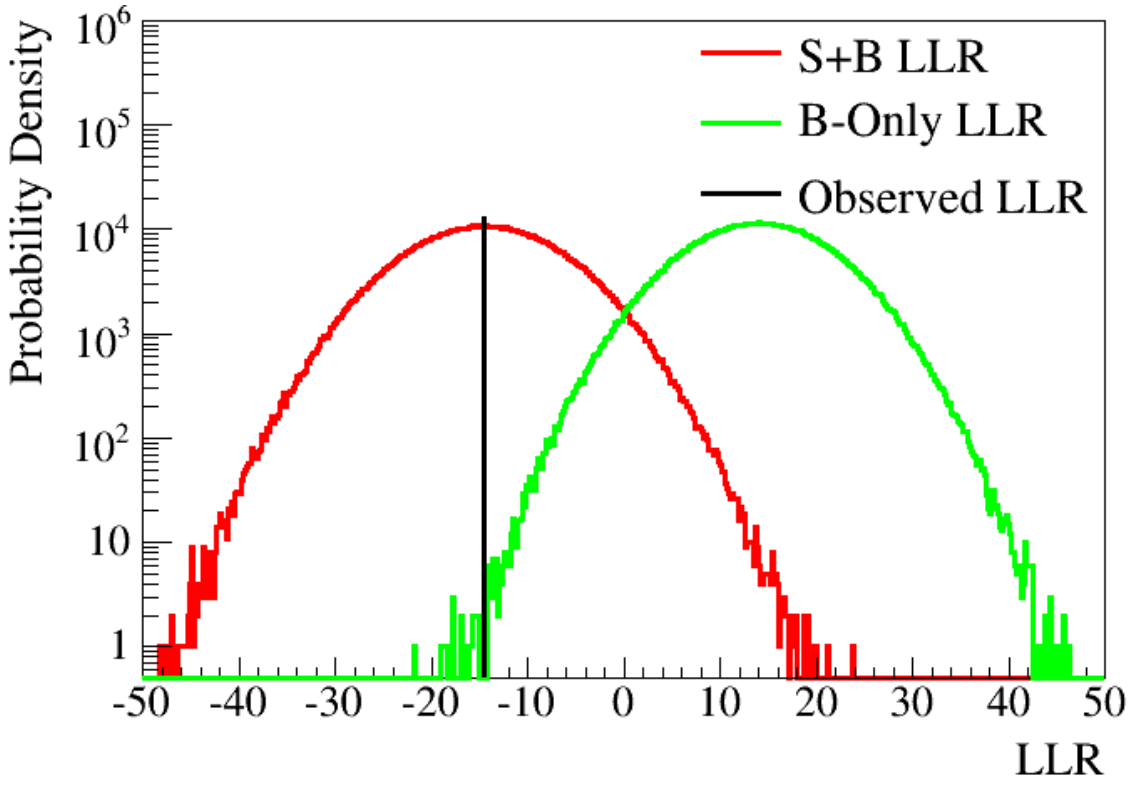


Discovery Criteria

- When we realize we cannot exclude, we begin to think about discovering something new
  - Discovery criteria can be thought of as excluding the background-only hypothesis (**H0**).

- Reduced dependence on signal model

  Used only in test-statistic (LLR): loose dependence on background p-Value.



Bayesian Discovery?

Bayes Factor

$$B = L'(data \mid r_{\max}) / L'(data \mid r = 0)$$

Similar definition to the profile likelihood ratio, but instead of maximizing L, it is averaged overr nuisance parameters in the numerator and denominator.

Similar criteria for evidence, discovery as profile likelihood.

Physicists would like to check the false discovery rate,
and then we're back to p-values.

But -- odd behavior of *B* compared with p-value for even a simple case

J. Heinrich, CDF 9678
http://newton.hep.upenn.edu/~heinrich/bfexample.pdf

Discovery vs Measurement

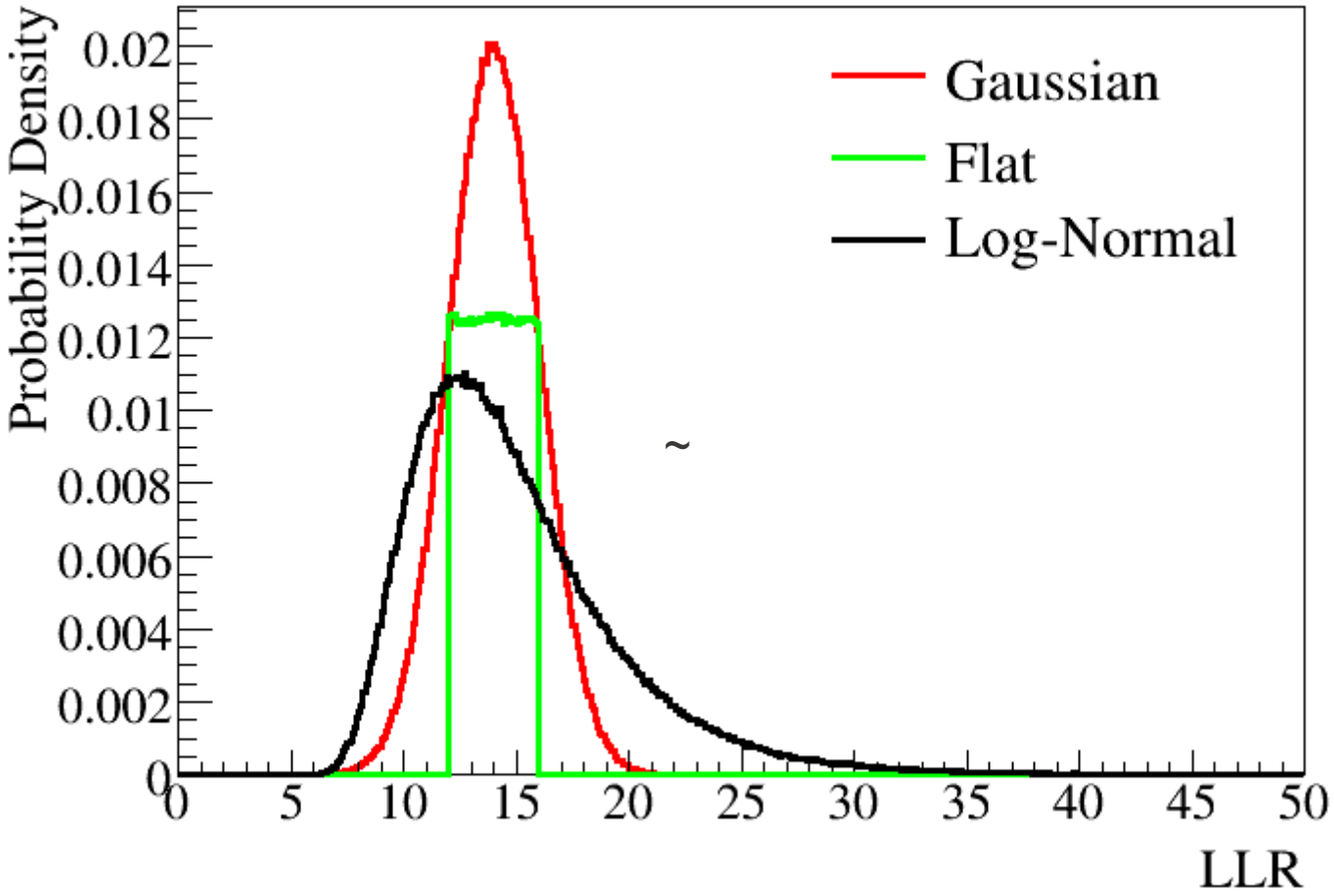
- Discovery is not always coincident with a good measurement!
  - Consider a search with nearly zero background predicted (~2E-7 events), low systematic uncertainty.
  - One observed event can be interpreted as a discovery!
    - Background-only is insufficient to describe observation.
  - Measurement is not so great: ±100%

Priors on Systematics

- Choosing a prior is not always a simple process
  - Even for experimental sources it's not always clear:
    - Many multiplicative factors are Gaussian
    - Acceptance errors often should be binomial
    - Large uncertainties force one to deal with probabilities nearing zero.
  - For theory uncertainties, it's far from clear
    - Uncertainty on PDF is generally prescribed (eg, 68% or 90%)
    - Scale uncertainty is tougher!
      - Flat? Gaussian? Other?

Priors on Systematics

- For each: 14±4 events, ~29%

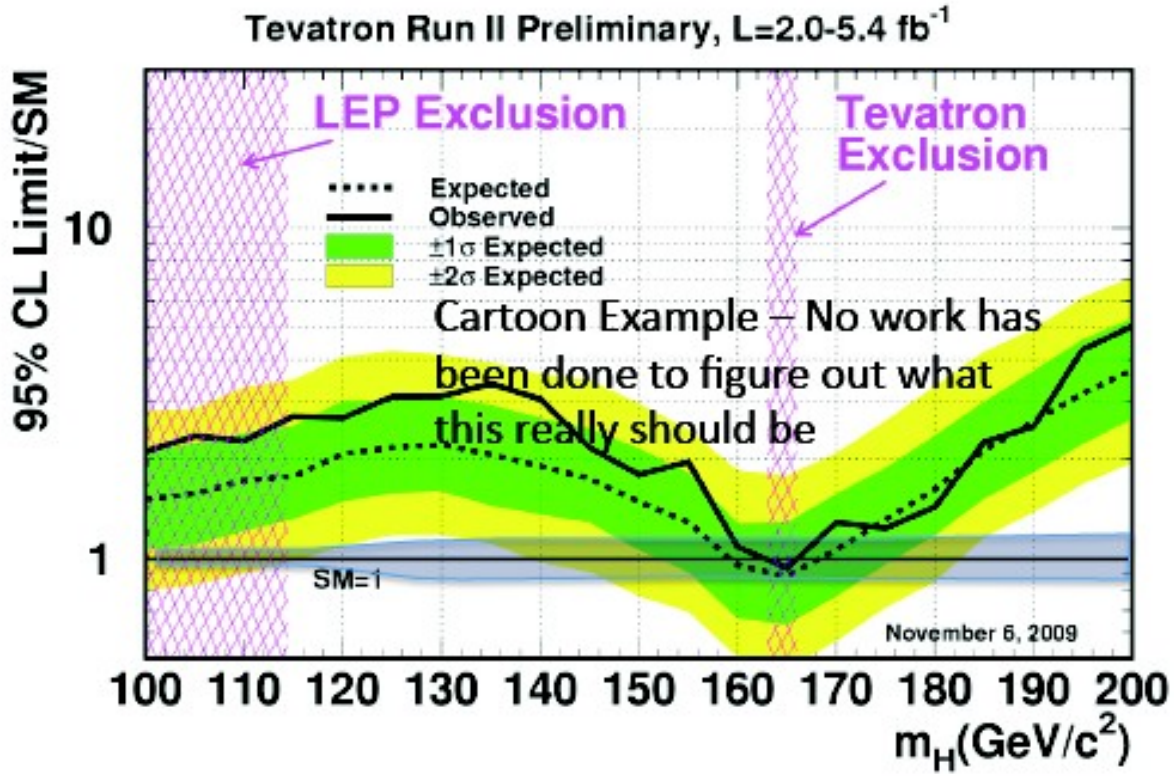


Priors on Systematics

- What should we do with asymmetric theory uncertainties?
  - Hard to assign a prior that doesn't change the mean!
    - Should the mean be equal to the nominal calculation?
    - Or should the uncertainty be allowed to change the rate?
    - *Or should it be left out of the calculation entirely?*
- Scale and PDF are often lumped together
  - PDF is an experimental issue that we know how to deal with (and potentially constrain!)
    - We probably should be separating these. Helps if theory calculations do too.

Uncertainty Presentation?

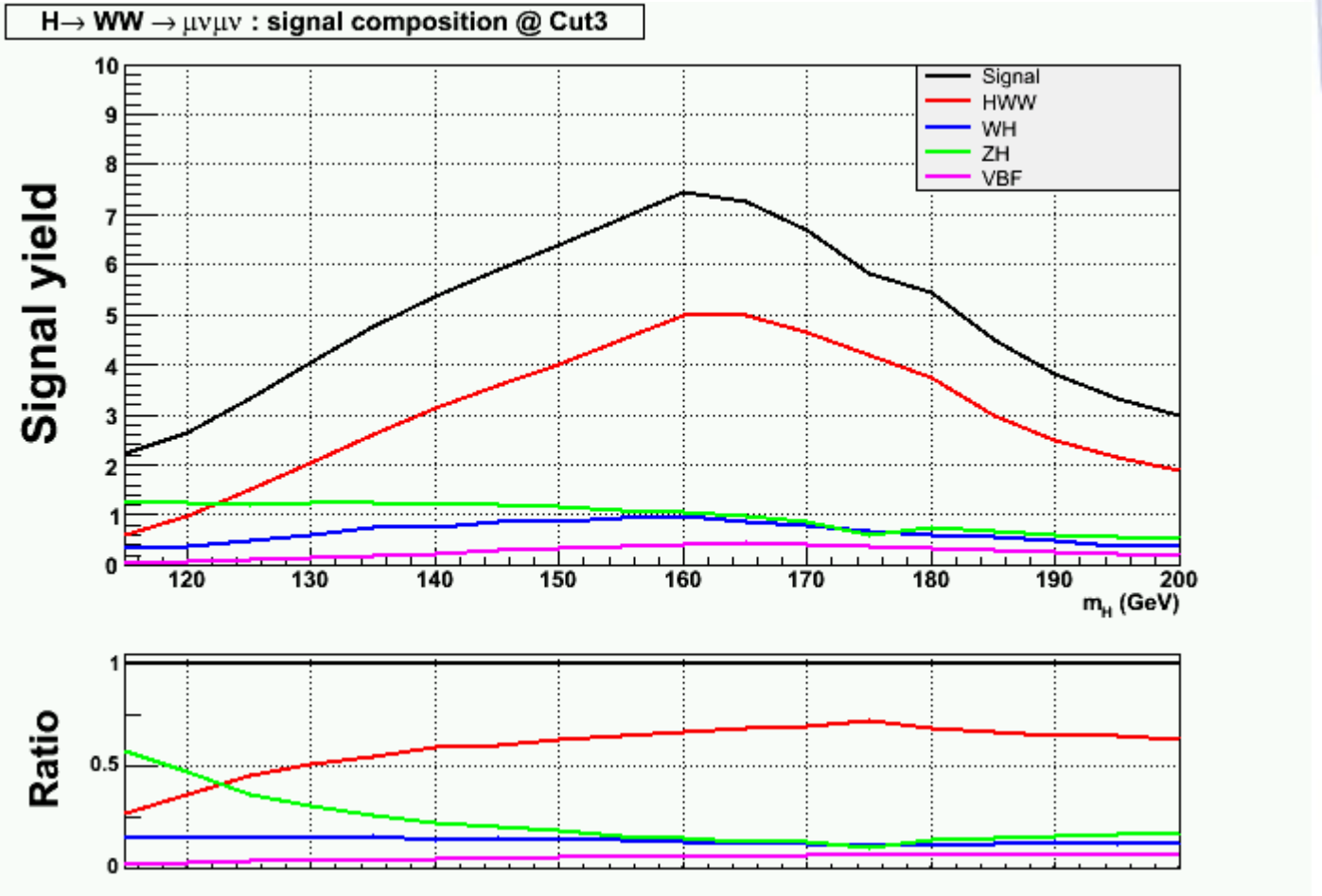
- Example from Tom's talk this morning
  - We present our limits in units of the ratio to the SM prediction.
  - Related to the fact that we are combining several different Higgs production mechanisms: Eg, ratio of ggH to VBF uncertain.



gg→H Uncertainty in H→WW

- What is the effective uncertainty size, considering a mix of processes?

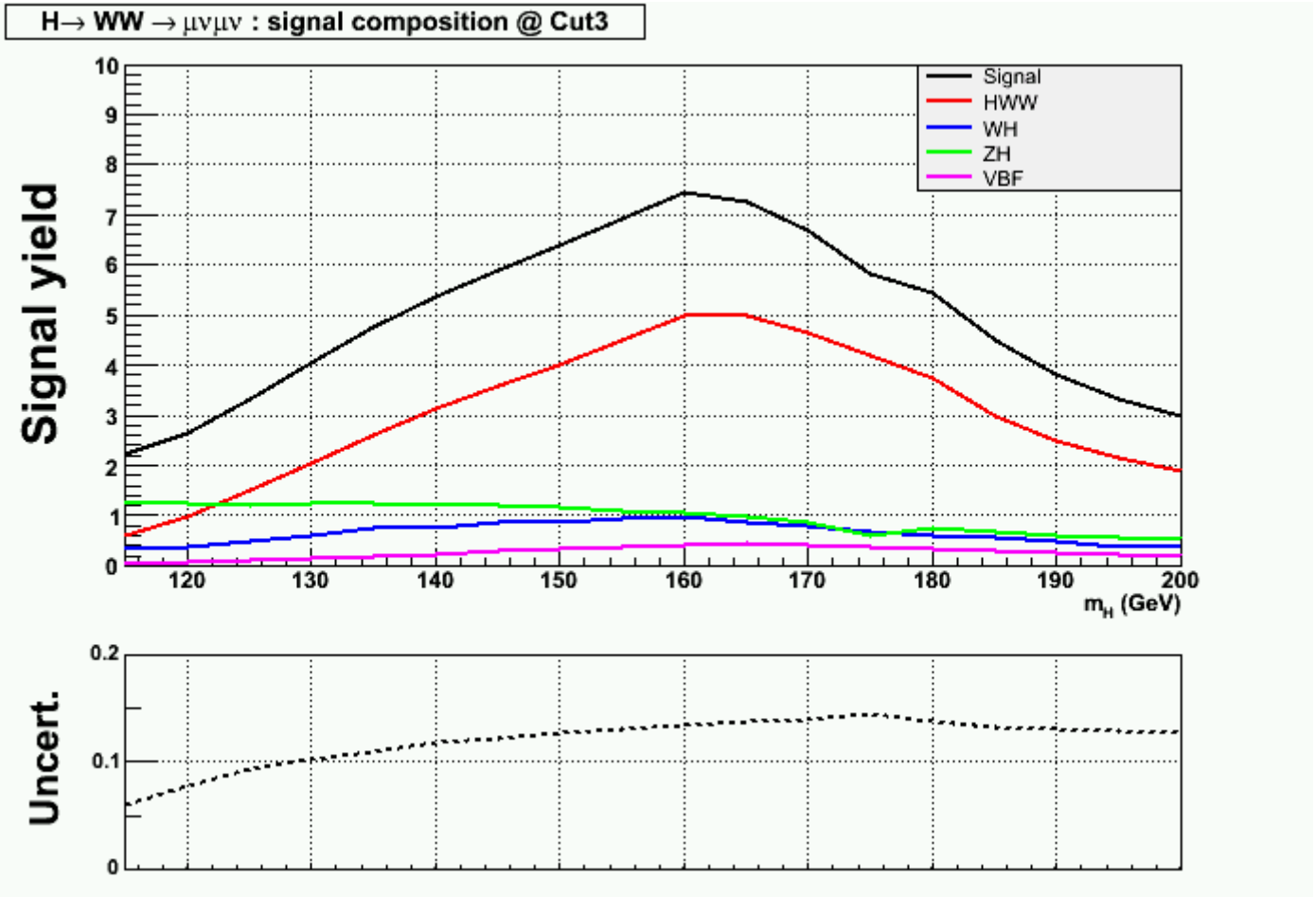
- GgH is 25-75% of total ll+MET signal yield



gg→H Uncertainty in H→WW

- What is the effective uncertainty size, considering a mix of processes?

- ggH is 25-75% of total ll+MET signal yield
- Assume uncertainties of:
  - 20% on ggH
  - 5% on the rest



gg→H Uncertainty in H→WW

- But does this even make sense?

- What one may care about is what limit you'd get for an alternative model for gg→H. IE, 3 limits:
  - gg→H scale up
  - gg→H nominal
  - gg→H scale down

- Not guaranteed to give the same answer, even if gg→H was the only source.