Practical Issues In Producing Limits, Evidence, and Discoveries in Experimental High-Energy Physics

Tom Junk
Fermilab
BIRS Statistics in HEP Workshop
July 2010

You Can Make a Discovery with Just One Event

H_{null}= Bear rate=0. H_{test}= Bear rate > 0. p-value is *almost* zero.

Some contributions to the expected background rate:

- People dressing as grizzly bears (good selection requirements can reduce this background)
- Cardboard cutout pictures of grizzly bears
- Digital photograph manipulation

Each background source needs some kind of prior, or auxiliary measurement if possible. There is also not much skepticism about the discovery claim.

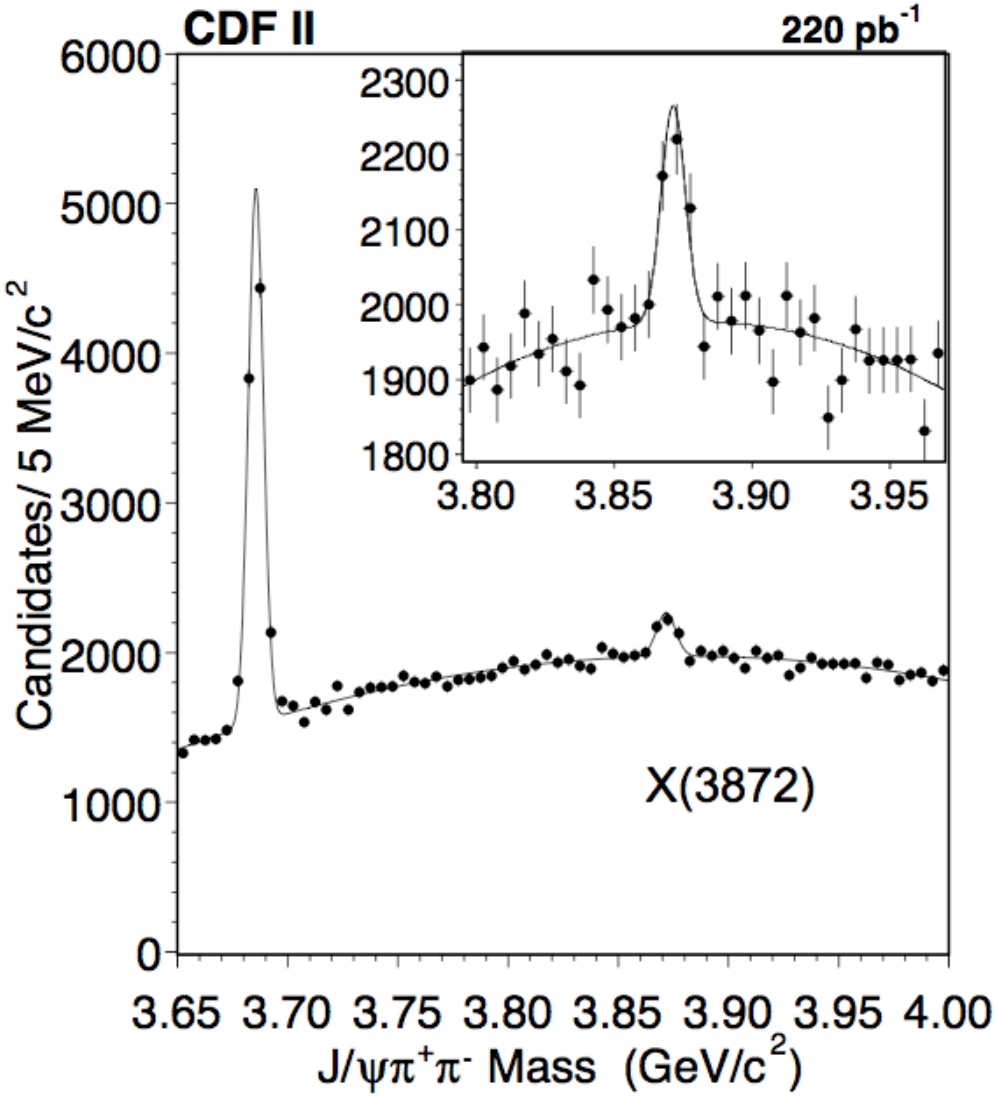
Extensions of Banff Challenge 1

$n_{off} * \tau$ as an estimate of b
n_{on} is the measurement in the signal region, with an estimated signal acceptance of ε. Given n_{off}, τ, ε, n_{on}, set a limit on the signal rate s (where $s\varepsilon$ is the expected signal yield and b is the background yield)

1) Usually there are multiple background sources b_1 ... b_n
2) Often there's more than one kind of signal, too. And they don't have to scale together (multidimensional signal parameter space). Grizzlies, brown bears, black bears, sun bears, ...
3) Usually there's more than one signal region ($n_{on\_1=}$, ... n_{on_n}), each with its own sets of ε's and τ's. Direct sightings of bears, observation of disturbed garbage cans, eyewitness accounts, auditory-only incidents, etc.
4) The ε's are uncertain. Sometimes they are just ratios of Poisson distributed numbers, but often there are more sources of uncertainty than just that. Same with the τ's. How to convert grizzlies/day to an expected number of pictures of grizzlies/day?
5) Often we have two or more "off-signal" auxiliary experiments used to evaluate b, each with its τ. What to do when they disagree?

Banff Challenge 2 samples 1, 3, and 4 above. 2 isn't so important as long as we can understand how to deal with the 1-signal problem, although problems occur in high-dimensional models that are not present in 1D models.

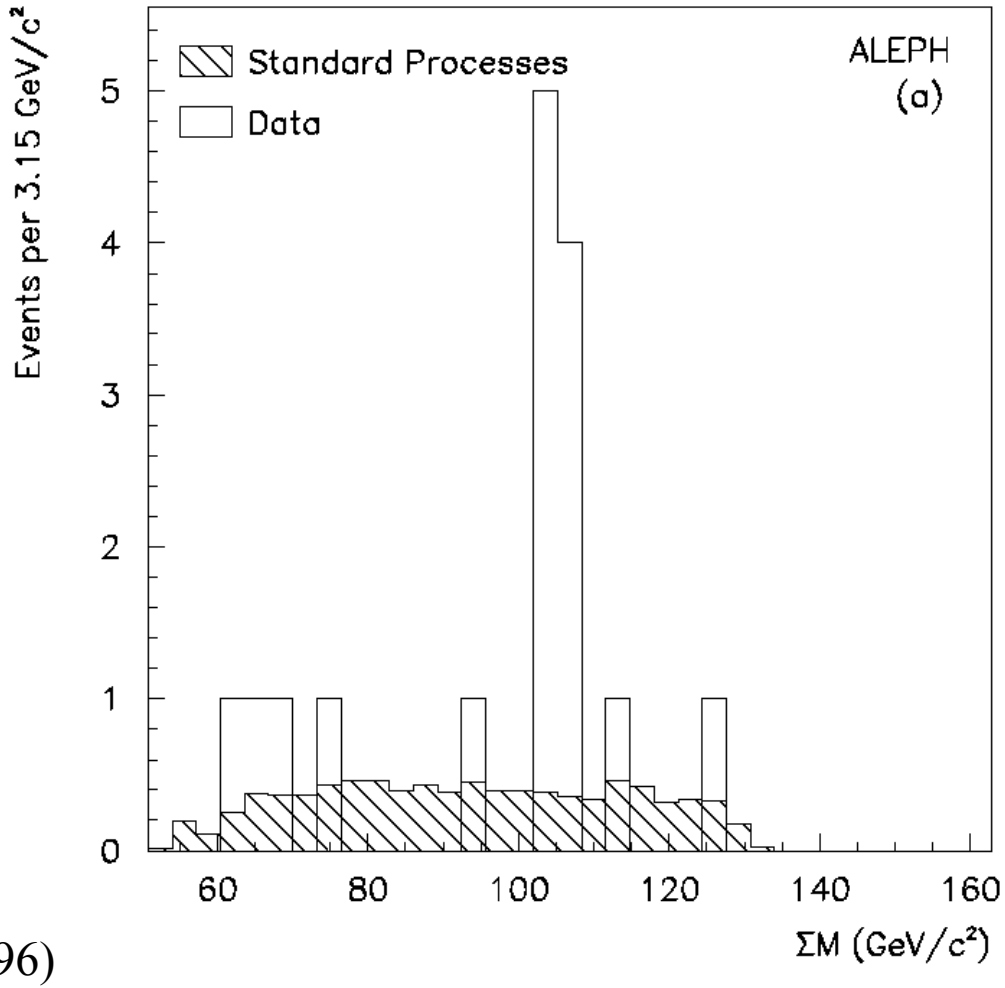
It would be nice if all discoveries were this clear (even with low s/b)



Guess a shape that fits the backgrounds, and fit it with a signal.

At Least they Explained what They Did

"the width of the bins is designed to correspond to twice the expected resolution ... and their origin is deliberately chosen to maximize the number of events found in any two consecutive bins"



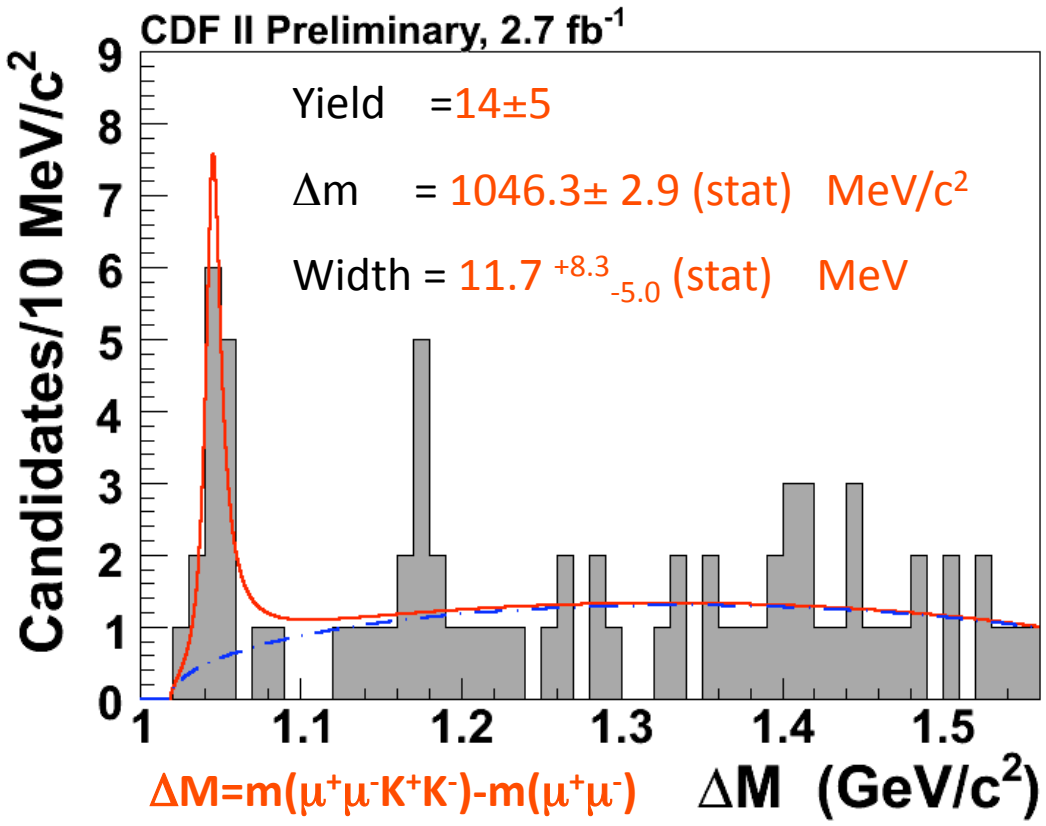
Dijet mass sum in $e^+e^- \rightarrow jjjj$

ALEPH Collaboration, Z. Phys. C71, 179 (1996)

Search for structures in $J/\psi\phi$ mass--Data

- We model the Signal (S) and Background (B) as:

  S: S-wave relativistic Breit-Wigner B: Three-body decay Phase Space



Convoluted with resolution (1.7 MeV)

Slide from K. Yi, Fermilab Joint Experimental/Theoretical Physics Seminar, March 17, 2009

How many bumps do you see?

$\sqrt{(-2\log(L_{max}/L_0))}=5.3$, need Toy MC to determine significance for low statistics

What if we don't have a signal model, and we're just on a hunting expedition? What's LEE now?

A Comment on low s and low b

Bins with tiny s and tiny b can have large s/b (Louis: large s/sqrt(b) is suspicious)

Naturally occurring in HEP and others seeking discovery:

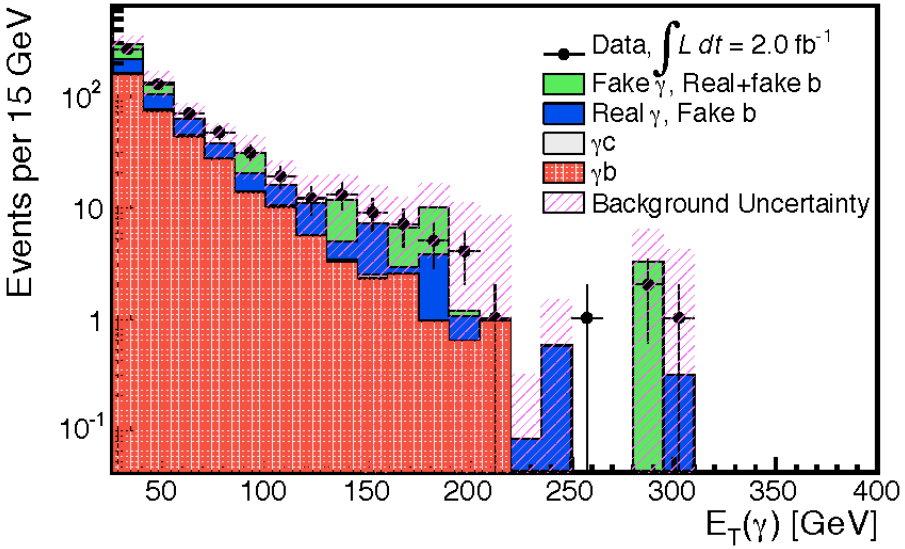
1) Each beam crossing has very small s and b but has the same s/b as neighboring beam crossings. Can make a histogram of the search for new physics separately for each beam crossing. Same s and b predictions, just scaled down very small.

   Adding is the same as a more elaborate combination if the histograms were accumulated under identical conditions (all rates, shapes, and systematics are the same)

2) Surveillance video catching a bear – each frame has a small s, b, but still worthwhile to collect each frame (and analyze them separately)

MC Statistics and "Broken" Bins

Overbinning is like overtraining a NN. s, b, and d can all be in different bins. A histogram can be partially overbinned, like this one here:

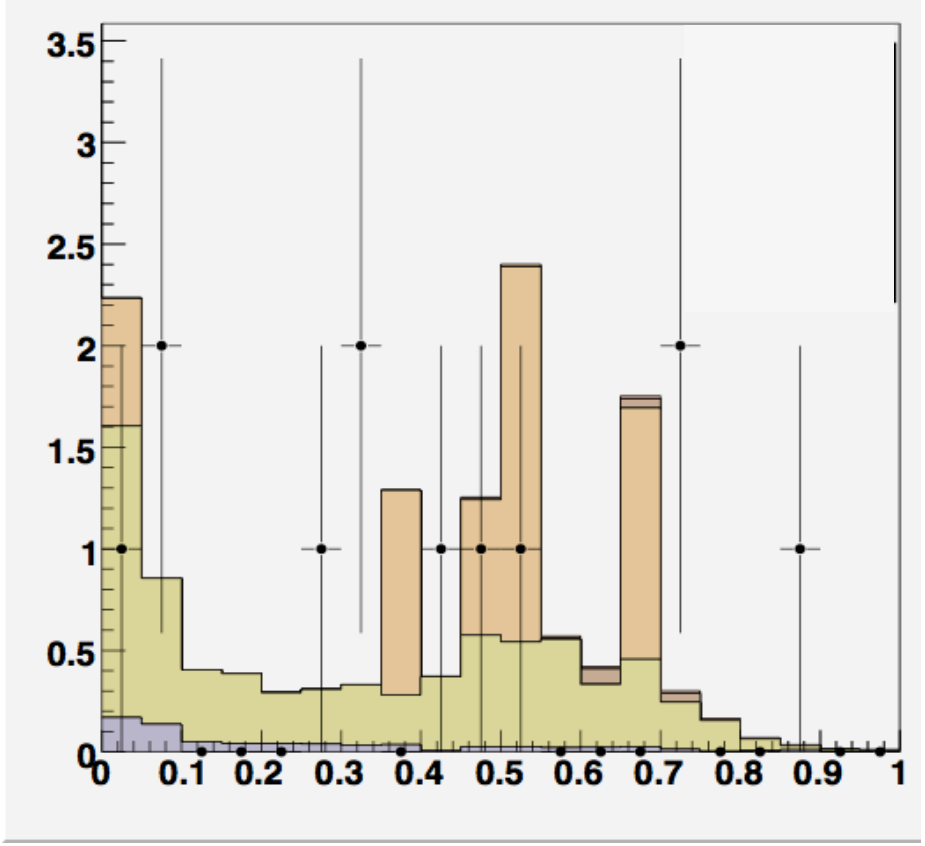


NDOF=?

- Limit calculators/discovery tools cannot tell if the background expectation is really zero or just a downward MC fluctuation.
- Real background estimations are sums of predictions with very different weights in each MC event (or data event)
- Rebinning or just collecting the last few bins together often helps.
- Problem compounded by requiring shape uncertainties to be evaluated! Alternate shape MC samples are often even more thinly populated than the nominal samples. Validation of adequate preparation of results is necessary? (but what are the criteria?)

A Pitfall -- Not Enough MC (data) To Make Adequate Predictions

An Extreme Example (names removed)



Questions: What's the shape we are trying to estimate? What is the uncertainty on that shape?

Cousins, Tucker and Linnemann tell us prior predictive p-values undercover with 0±0 events are predicted in a control sample.

CTL Propose a flat prior in true rate, use joint LF in control and signal samples. Problem is, the mean expected event rate in the control sample is n_{obs}+1 in control sample. Fine binning → bias in background prediction.

Overcovers for discovery, undercovers for limits?

An Extreme Example from Georgios Choudalakis

Ten MC events, used to estimate a background b, but with different weights.

τ_1=0.1
τ_2=0.2
τ_3=0.3
τ_4=0.4
τ_5=0.5
τ_6=0.6
τ_7=0.7
τ_8=0.8
τ_9=0.9
τ_{10}=1.0

The sum is 5.5 = b

But what to use for the prior on b?

Are there any possible (and possibly large) weights which are not represented here? Could we have gotten a MC event with weight=100?

Very little information about the distribution of the weights is present here.

Need acceptance as a function of weight.

General limit/discovery tools – do we need a histogram of weights for each bin of each signal and background contribution? What if this is insufficient anyway (as it is in this case).

Commonly Used Tools for Setting Limits and Discovering New Processes in use at the Tevatron

- Bayesian limits -- common at CDF
  - genlimit code by Joel Heinrich, added to mclimit code by Tom Junk. New MCMC calculations are more robust on big problems.
  - Implements posterior integrated over systematic uncertainties with a flat prior in cross section in 1D
  - Method described in PDG statistics reveiw
  - Extra feature -- “correlated prior”

- CL_s limits -- common at D0, but used at CDF as well.
  - Collie code by Wade Fisher in use at D0
  - Method described in PDG statistics review
  - mclimit was originally designed to do CL_s and still does.
  - Guaranteed to cover within the ensemble chosen
  - Often more optimal than Bayesian limits

Mini-Review: Bayesian Limits

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

Where r is an overall signal scale factor, and θ represents all nuisance parameters.

$$P_{Poiss}(data \mid r,\theta) = \frac{(rs_i(\theta) + b_i(\theta))^{n_i}\, e^{-(rs_i(\theta)+b_i(\theta))}}{n_i!}$$

where n_i is observed in each bin i, s_i is the predicted signal for a fiducial model (SM), and b_i is the predicted background. Dependence of s_i and b_i on θ includes rate, shape, and bin-by-bin independent uncertainties.

Mini-Review: Bayesian Limits

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

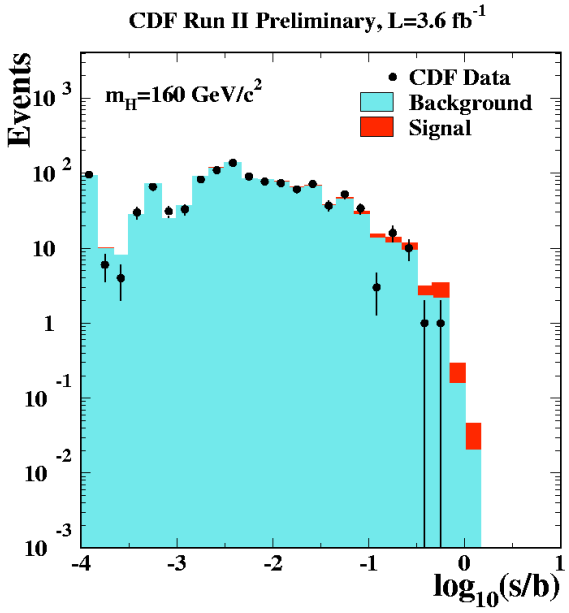
Typically $\pi(r)$ is constant
Other options possible.
Sensitivity to priors a concern.

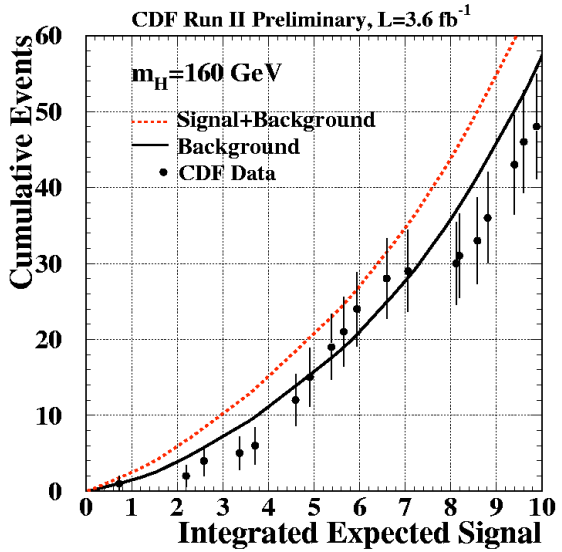
Measure r: $$0.68 = \int_{r_{low}}^{r_{high}} L'(data \mid r)\pi(r)dr$$

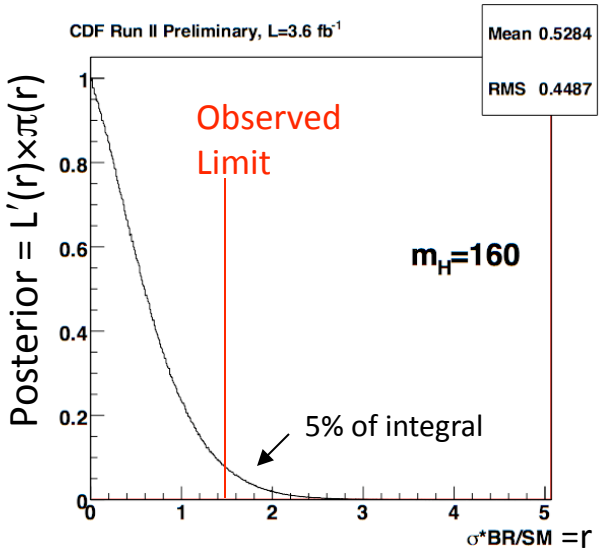
$$r = r_{\rm max}{}^{+(r_{high}-r_{\rm max})}_{-(r_{\rm max}-r_{low})}$$

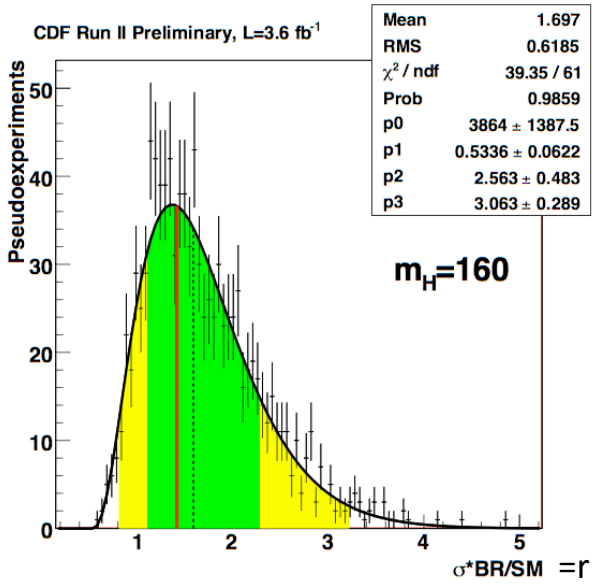
Usually: shortest interval containing 68% of the posterior (other choices possible)

Bayesian Example: CDF Higgs Search at m_H=160 GeV



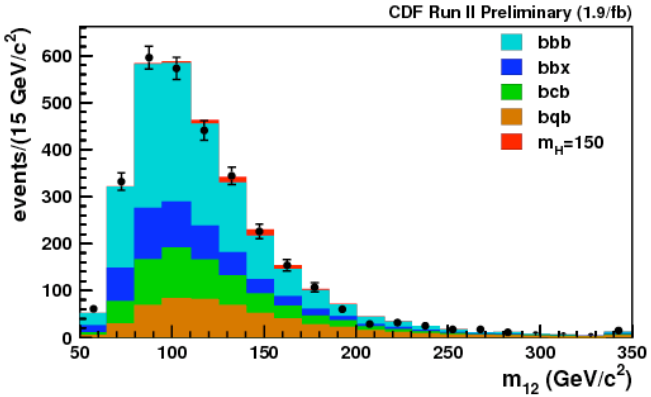


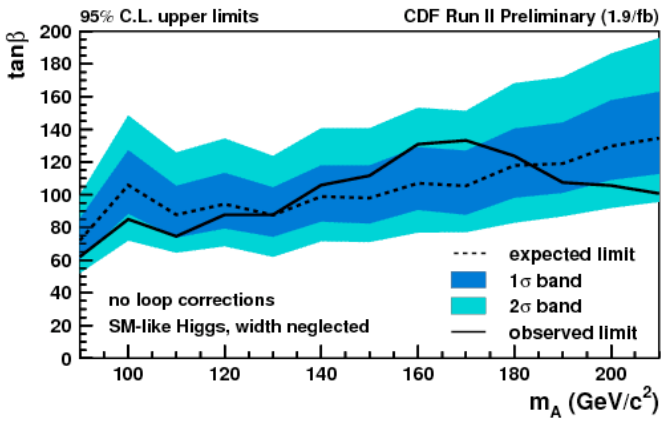


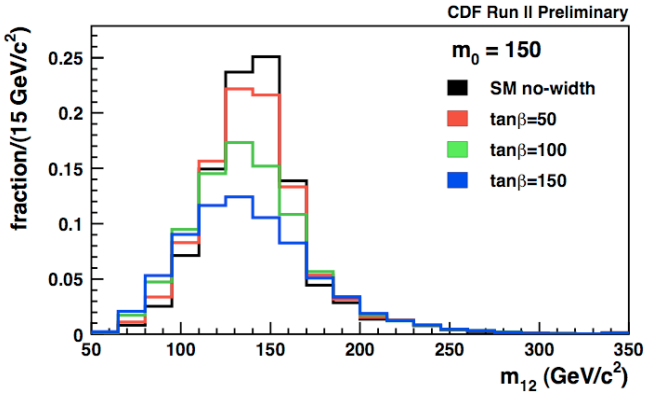


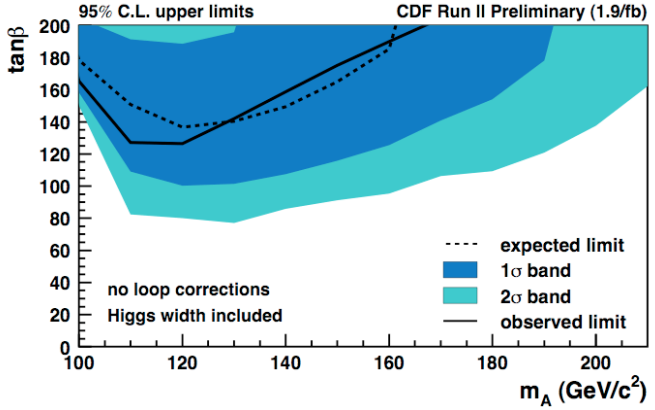
An Example Where Usual Bayesian Software Doesn't Work

- Typical Bayesian code assumes fixed background, signal shapes (with systematics) -- scale signal with a scale factor and set the limit on the scale factor
- But what if the kinematics of the signal depend on the cross section? Example -- MSSM Higgs boson decay width scales with $\tan^2\beta$, as does the production cross section.
- Solution -- do a 2D scan and a two-hypothesis test at each m_A,$\tan\beta$ point



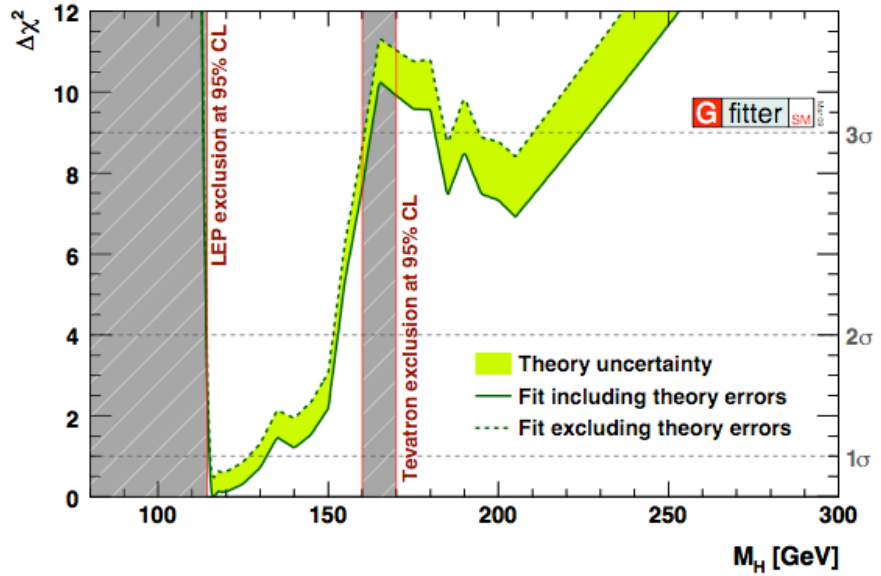




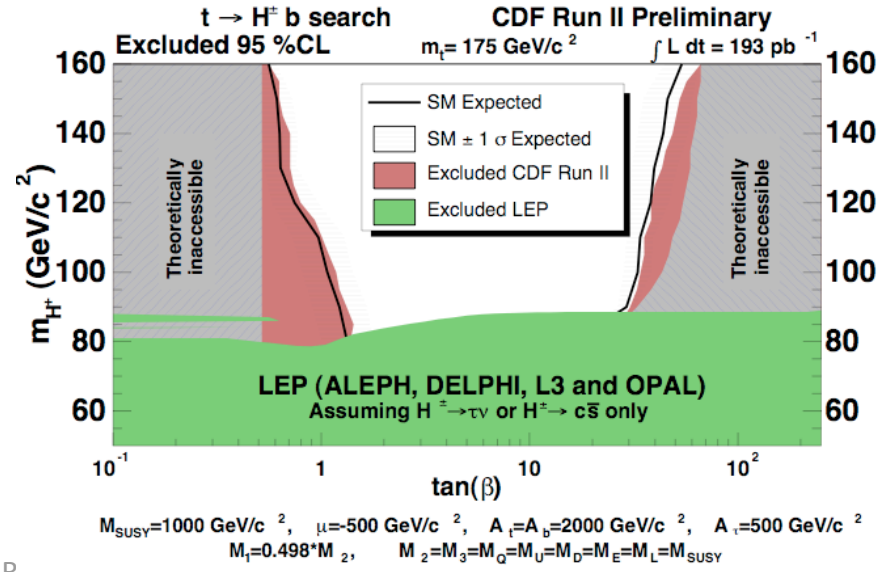


Priors in Non-Cross-Section Parameters of Interest

Example: take a flat prior in m_H; can we discover the Higgs boson by process of elimination? (assumes exactly one Higgs boson exists, and other SM assumptions)



Example: Flat prior in log(tanβ) -- even with no sensitivity, can set non-trivial limits..



Nuisance Parameter Priors

- An endless discussion for every measurement/limit/evidence/observation
- Central to the process -- A result is not ready until the collaboration is satisfied that all systematic uncertainties are estimated sufficiently and included in the result.

- Usually we do not have complete estimations:
  - ±1σ variations, but what does ±2σ look like? ±5σ?
  - alternate histogram shapes – is it safe to extrapolate?
  - What if the variation is say between two arbitrary models (Pythia vs. Herwig, or from a data –MC comparison?) Is it fair to extrapolate these? How can you be more like the data than the data?
  - What if a nuisance parameter variation makes a prediction go negative, say for the background or the signal? If a parameter is truncated for one prediction, should we apply that truncation everywhere?
  - Truncated Gaussians have biased means and medians (and sometimes modes)
  - Other priors move mean, median, or mode away from the central prediction, or have "corners" in their distributions.
  - Symmetric and asymmetric impacts for the same nuisance parameter on different predictions.

Predictions like b=1.0±0.1 are easy. What does 1.0±1.0 mean? What does $1.0^{+0.2}_{-0.1}$ mean? What does $1.0^{+5.0}_{-1.0}$ mean?

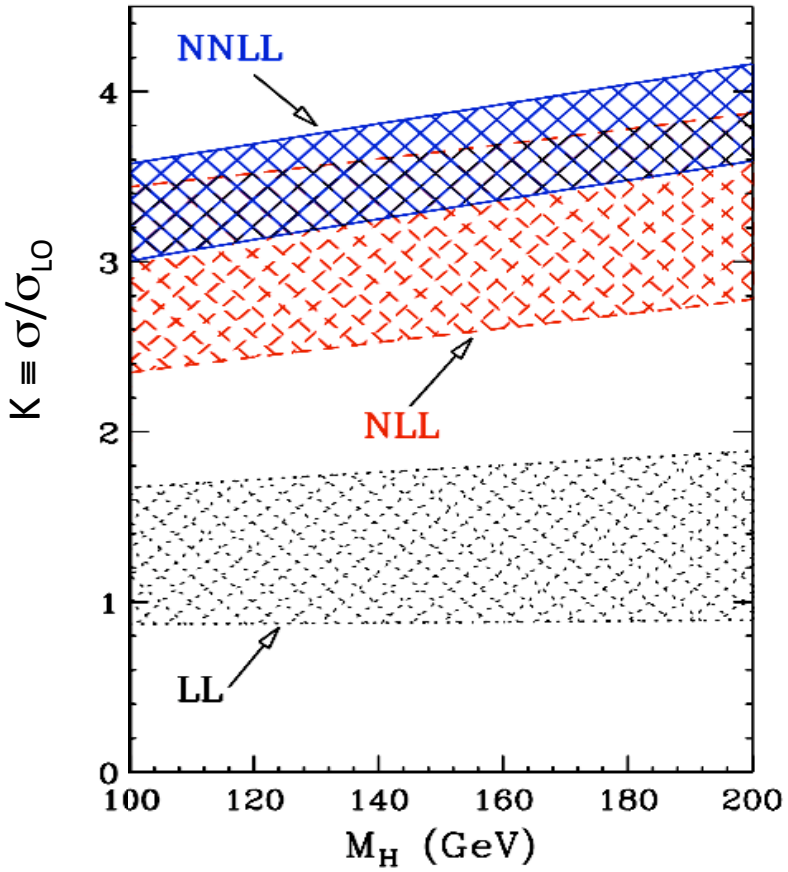
Our Theorists are not Statisticians

- NLO corrections -- ~80% (almost double the cross section)!
- NNLO QCD corrections -- An additional 40% on top of that! Residual uncertainty ~10%. Catani, de Florian, Grazzini, Nason JHEP **0307**, 028 (2003) hep-ph/0306211

Also resummed QCD corrections at NNLL

NLL,NNLL bands: $0.5m_H<\mu_F,\mu_R<2M_H$.
Bands on LO and LL unreliable.

We take a ±12% uncertainty on $\sigma_{gg\to H}$ for scale and PDF



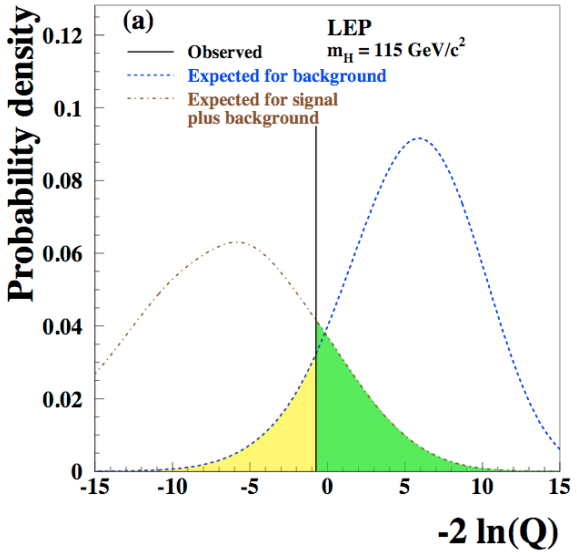
Mini-Review: CL_s Limits

- Based on p-values using the log likelihood ratio as the test statistic. Neyman-Pearson lemma says LLR is the uniformly most powerful test statistic, although the Neyman-Pearson one fits for the parameter of interest, not just the nuisance parameters, making the null hypothesis a subset of the test hypothesis

$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

Glen Cowan's LLR also fits for s (actually $r \times s$) in the numerator, while $r = 0$ in the denominator

Mini-Review: CL_s Limits



p-values:

Yellow area = $1\text{-}CL_b = 1\text{-}P(-2\ln Q > -2\ln Q_{obs} \mid b \text{ only})$

Green area = $CL_{s+b} = P(-2\ln Q > -2\ln Q_{obs} \mid s+b)$

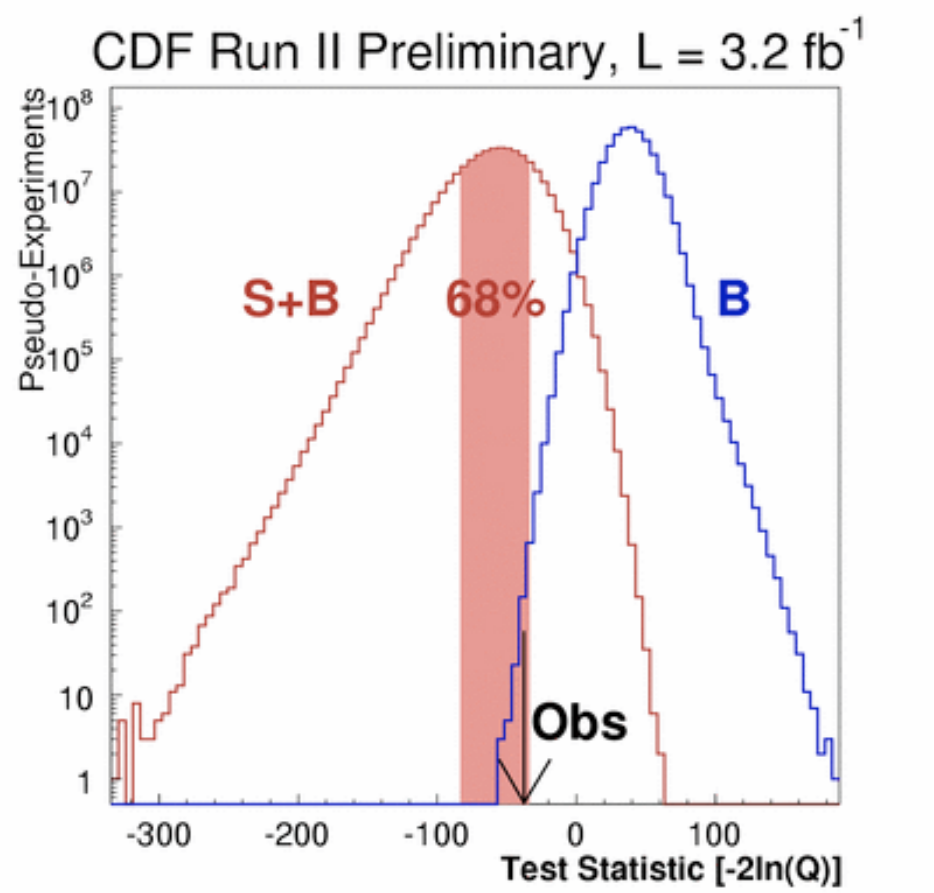
$CL_s \equiv CL_{s+b}/CL_b \geq CL_{s+b}$

Exclude if $CL_s < 0.05$

Vary r until $CL_s = 0.05$ to get r_{lim}

- Advantages:
  - Exclusion and Discovery p-values are consistent.
    Example -- a 2σ upward fluctuation of the data with respect to the background prediciton appears both in the limit and the p-value as such
  - Does not exclude where there is no sensitivity
    (big enough search region with small enough resolution and you get a 5% dusting of random exclusions with CL_{s+b})

Discovery with p-Values



Example: CDF single top.

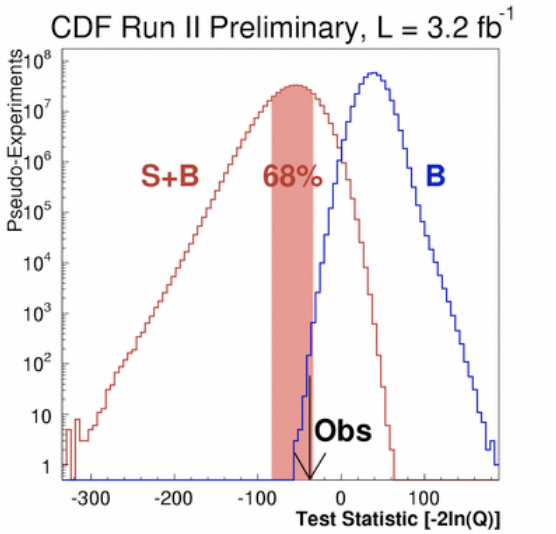
$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

100 M s+b and b-only pseudoexperiments, each with fluctuated nuisance parameters, and fit twice.

5σ: p-value of 2.77×10^{-7} or less.
3σ: p-value of 1.35×10^{-3} or less
2σ: p-value of 2.28% or less

Buzzword: "Prior Predictive ensemble"

Fitting and Fluctuating



$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

- Monte Carlo pseudoexperiments are used to get p-values.
- Test statistic -2lnQ is not uncertain for the data.
- Distribution from which -2lnQ is drawn is uncertain!

- Nuisance parameter fits in numerator and denominator of -2lnQ **do not incorporate systematics into the result**.
  Example -- 1-bin search; all test statistics are equivalent to the event count, fit or no fit.
- Instead, we fluctuate the probabilities of getting each outcome since those are what we do not know. Each pseudoexperiment gets random values of nuisance parameters.
- Can also try values of nuisance parameters that maximize the p-value, but that's very conservative (called the supremum p-value, still needs choices of parameter ranges).
- Why fit at all? It's an optimization. Fitting reduces sensitivity to the uncertain true values and the fluctuated values. For stability and speed, you can choose to fit a subset of nuisance parameters (the ones that are constrained by the data). Or do constrained or unconstrained fits, it's your choice.
- If not using pseudoexperiments but using Wilk's theorem, then the fits are important for correctness, not just optimality.

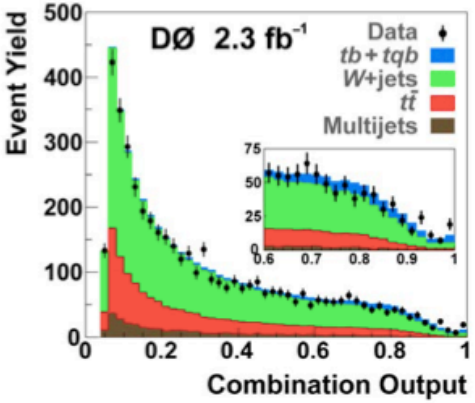
Using Bayesian Techniques as an ingredient for Discovery

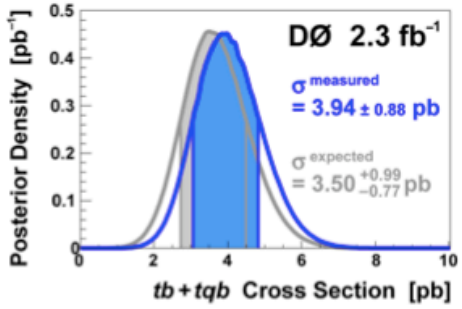
- D0 measures the single top cross section with a Bayesian technique
- The measured cross section is used as a test statistic for the p-value for significance. Pseudoexperiments fluctuate systematics.

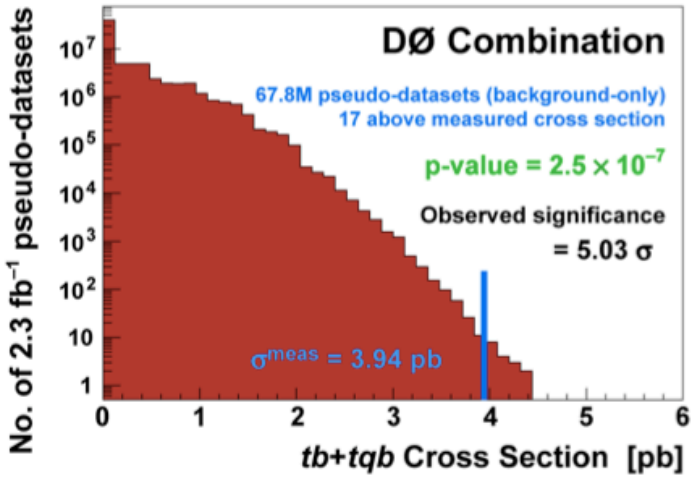
Combined Results

$$\sigma(p\bar{p} \to tb + X, tqb + X) = 3.94 \pm 0.88 \text{ pb}$$

(m_t=170GeV)







$$p - value = 2.5 \times 10^{-7}$$

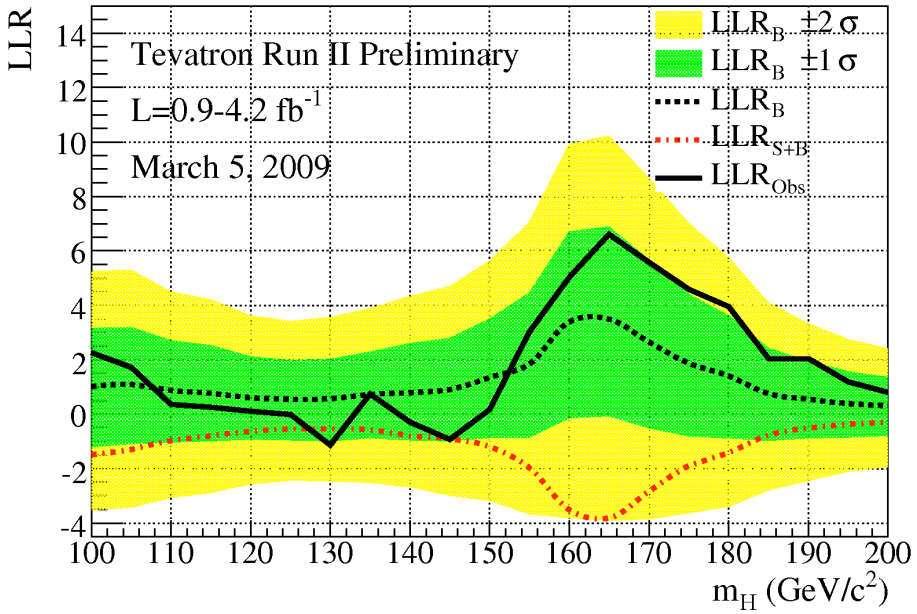
Measured Significance = 5.03σ

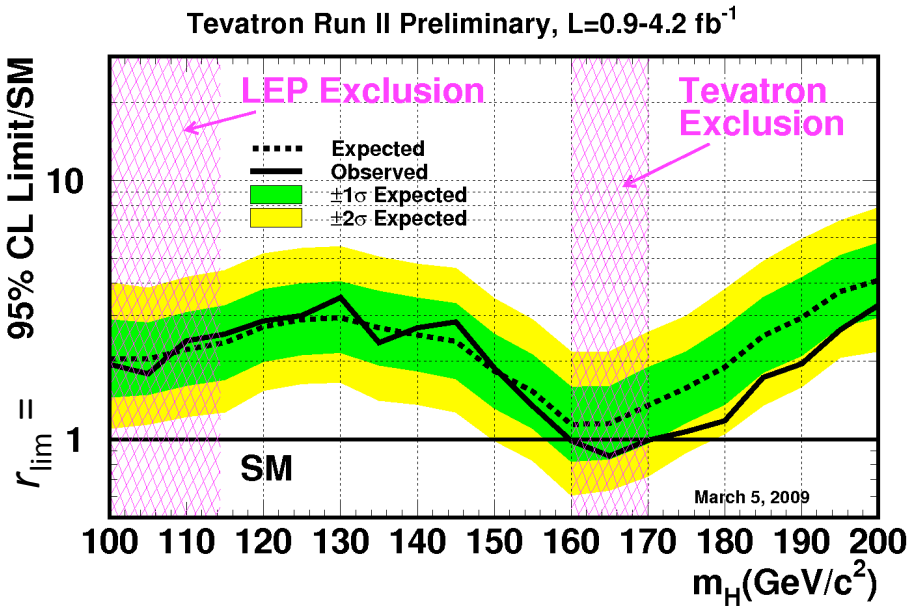
35

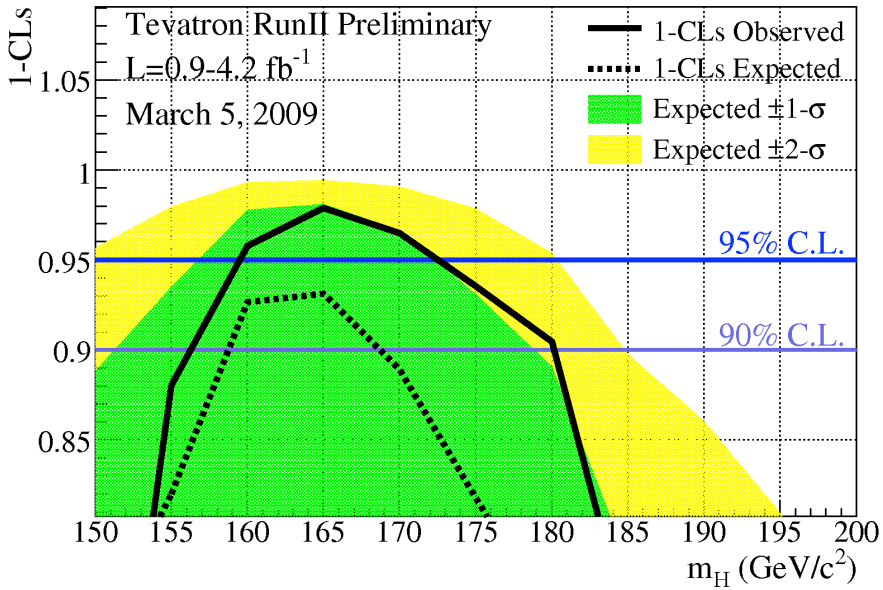
C. Gerber,
D0 single Top
Fermilab seminar
March 10, 2009

23

Tevatron Higgs Combination Cross-Checked Two Ways







Very similar results --

- Comparable exclusion regions
- Same pattern of excess/deficit relative to expectation

n.b. Using CL_{s+b} limits instead of CL_s or Bayesian limits would extend the bottom of the yellow band to zero in the above plot, and the observed limit would fluctuate accordingly. We'd have to explain the 5% of m_H values we randomly excluded without sufficient sensitivity.

Testing Two Non-Nested Hypotheses

Lambda = P(data|Intelligent Design)/P(data|Evolution)

- Two models share some parameters but not all (e.g. some nuisance parameters relating to the conversion of direct observation to interpretable statements)
- P(data|Intelligent Design) = 1? P(data|evolution) ~ 0?
  But what are the nuisance parameters?
- Maybe you can nest these, but is that scientific?

Sociological Issues

- Discovery is conventionally 5σ. In a Gaussian asymptotic case, that would correspond to a ±20% measurement.

- Less precise measurements are called "measurements" all the time

- We are used to measuring undiscovered particles and processes. In the case of a background-dominated search, it can take years to climb up the sensitivity curve and get an observation, while evidence, measurements, etc. proceed.

- Referees can be confused.