Hypothesis Testing In HEP with Uncertain Nuisance Parameters

and observations on

Odd p-value Behavior

Tom Junk

University of Illinois at Urbana-Champaign

- The general problem – Rejecting a null hypothesis or rejecting a test hypothesis when neither of them is particularly well specified
- Combinations of search analyses
- Some software and examples
- Digression: Odd p-value behavior when channels are combined

The Problem: Searching for New Physics in a HEP experiment

Just about the only thing common is that the observations are all
integer event counts.

Two hypotheses:

H0=null hypothesis, usually the “Standard Model”
H1=a test hypothesis, which includes some new interaction or particle

H0 is a compound hypothesis, with nuisance parameters
H1 is the same, with physics parameters,
and (usually, but not necessarily) the same nuisance parameters as H0,
plus a few.

Example: H0 describes predicts a Standard Model background rate in each bin of each histogram for a search analysis. Nuisance parameters are things like luminosity, acceptance, SM cross section predictions, etc…

H1 is the SM with something new. Physics parameters are the mass and cross section of the new thing, and additional nuisance parameters are its acceptance and reconstruction resolution.

Analyzers are Smart and Careful and Want Everything Included

- Events are selected based on a model of new physics to be tested
    - Sometimes this model isn't very specific (searching for "anything" – another topic)
    - Usually searches are optimized for specific processes
- Most backgrounds are rejected (detector trigger, simple selection cuts)

- Sophisticated discriminants are formed (Neural Nets, Likelihoods, Matrix Element-based functions) out of event observables to separate signal from background. In the past, simple discriminants were reconstructed masses ("bump hunting").

- Histograms are formed using these discriminants, and data are compared against H0 and H1 predictions (usually estimated with Monte Carlo, but often with data). Bin contents are sums of different components!

- Uncertainties are evaluated on rates, histogram shapes, and MC (data) statististical uncertainties. Some nuisance parameters affect rates and shapes of signals and backgrounds. Some uncertainties are asymmetric.

- Often data samples are used to calibrate nuisance parameters

- Usually more than one histogram (channel, or even analysis team) is testing the same H1 in different ways, and we'd like to combine them all. Sometimes events are shared – it's easier when they aren't.
- H1 can predict FEWER events in some bins than H0. Be ready for it!

The Goals

- We would like to be the first to discover something new.
- We would like to set the strongest limits on models that do not predict nature.
- We'd like to have error rates specified by our choice of Confidence Level (95% for exclusion, 5σ for discovery, usually) ("do what you like but check the coverage – OK for correctness, but often not optimal).

These goals are usually not in conflict, unless an analysis is poorly designed.

Analyses are not often optimized with these in mind, but usually something easier to calculate.

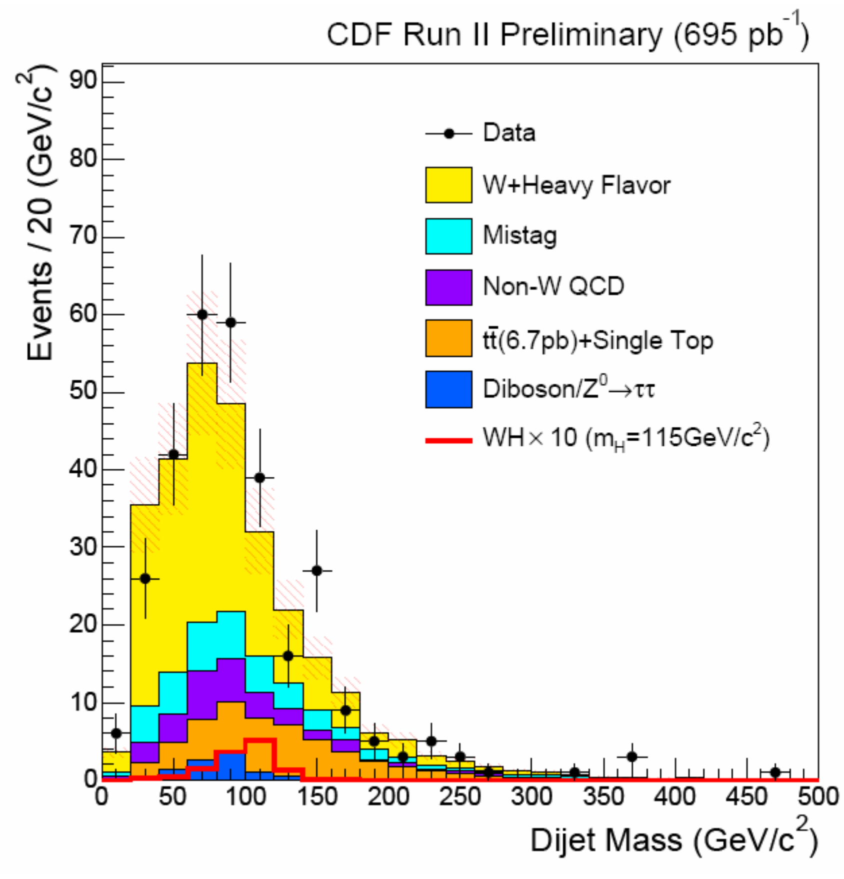
Examples: optimize

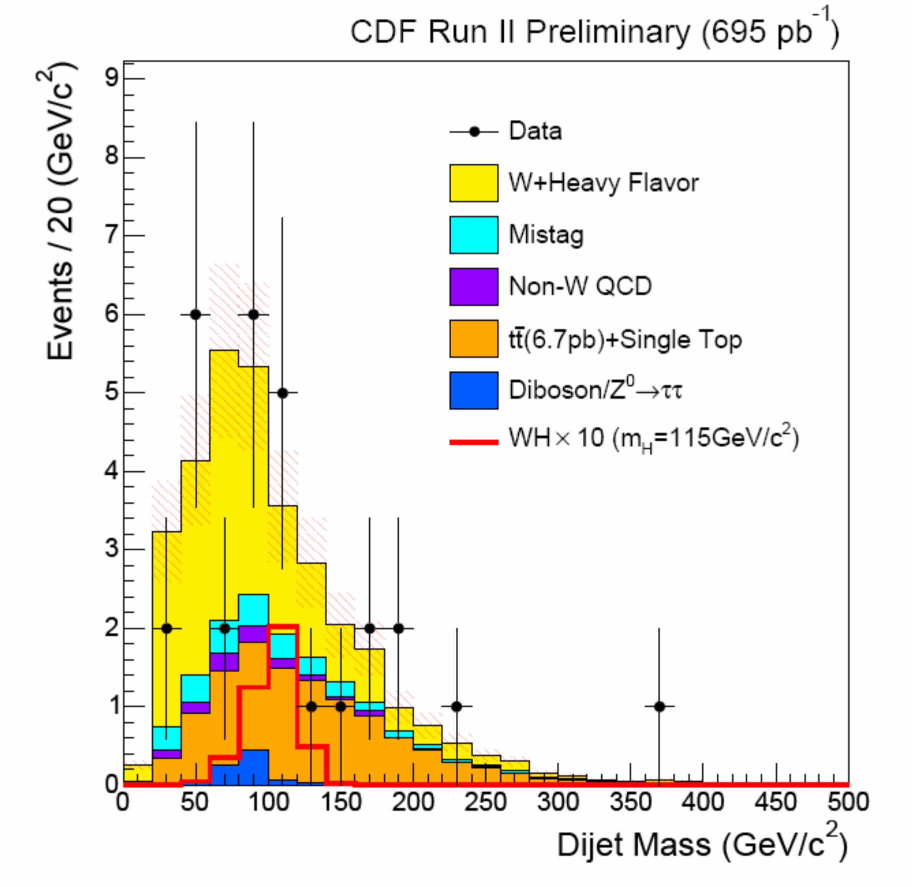
- s/sqrt(b)
- s/sqrt(s+b)
- Neural-Net Training Error sum. (Why do we care about this one? NN training programs are built around it).

In a multichannel (multi-bin) analysis, these can be ambiguous.

A Concrete Example: Standard Model Higgs Boson Search with CDF

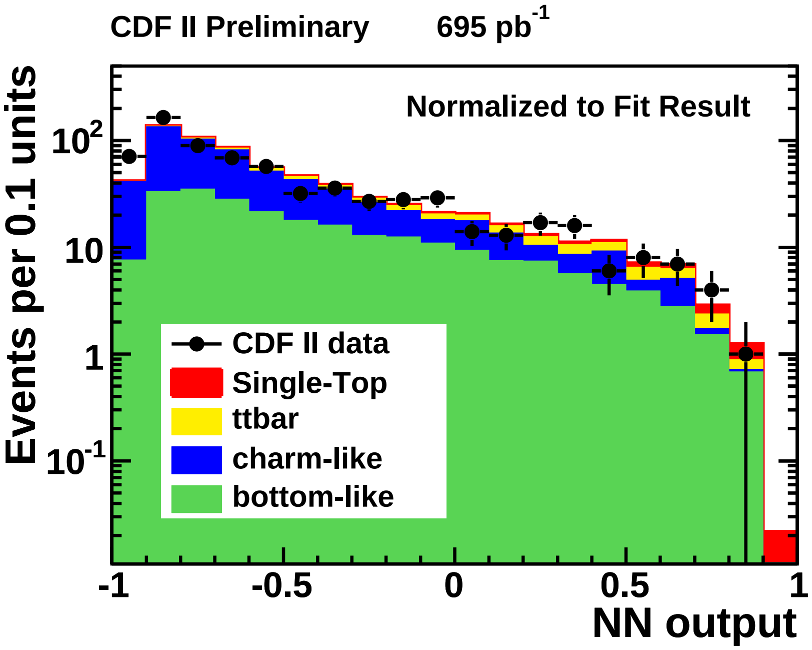
shown at APS Dallas, 2006



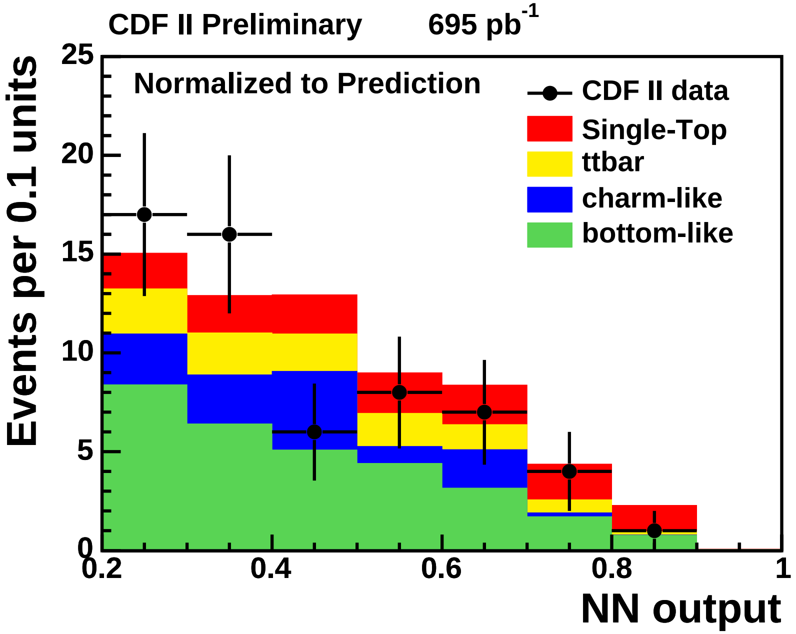


Two disjoint event samples, “single-tagged” and “double-tagged”
Nuisance parameters affect both in a correlated way:
Luminosity, background rates, Jet energy scale and resolution, signal detection efficiency. New physics prediction adds incoherently to the H0 prediction

Another Interesting Example: A Gaussian Problem on One Side and a Poisson Problem on the Other.



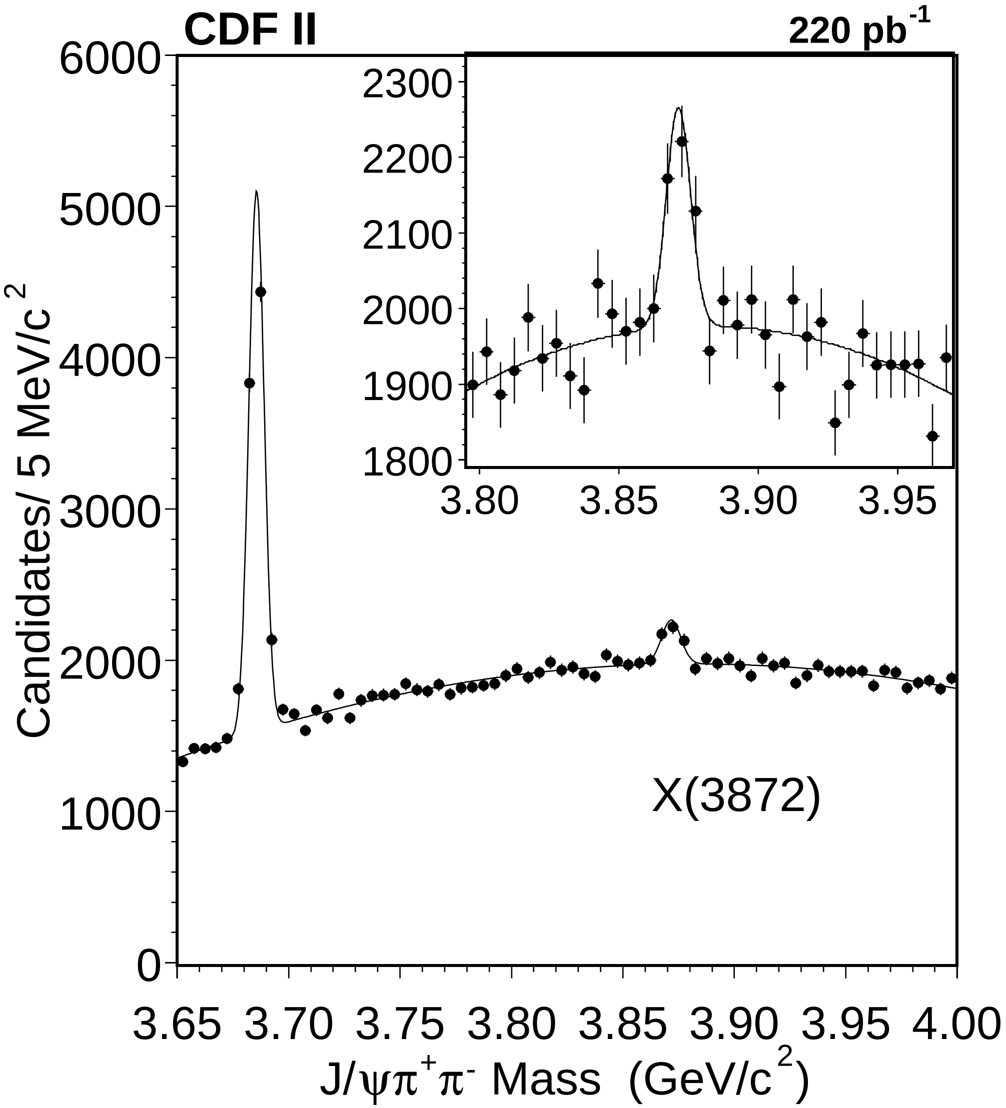
Histogram of a NN variable to search for single top quarks. Distributions in each bin ~Gaussian for most bins.



Zoomed in on last few bins
Poisson!

SM predictions can be highly uncertain (QCD is difficult to use for predictions), but data can constrain background rates in bins with little signal.

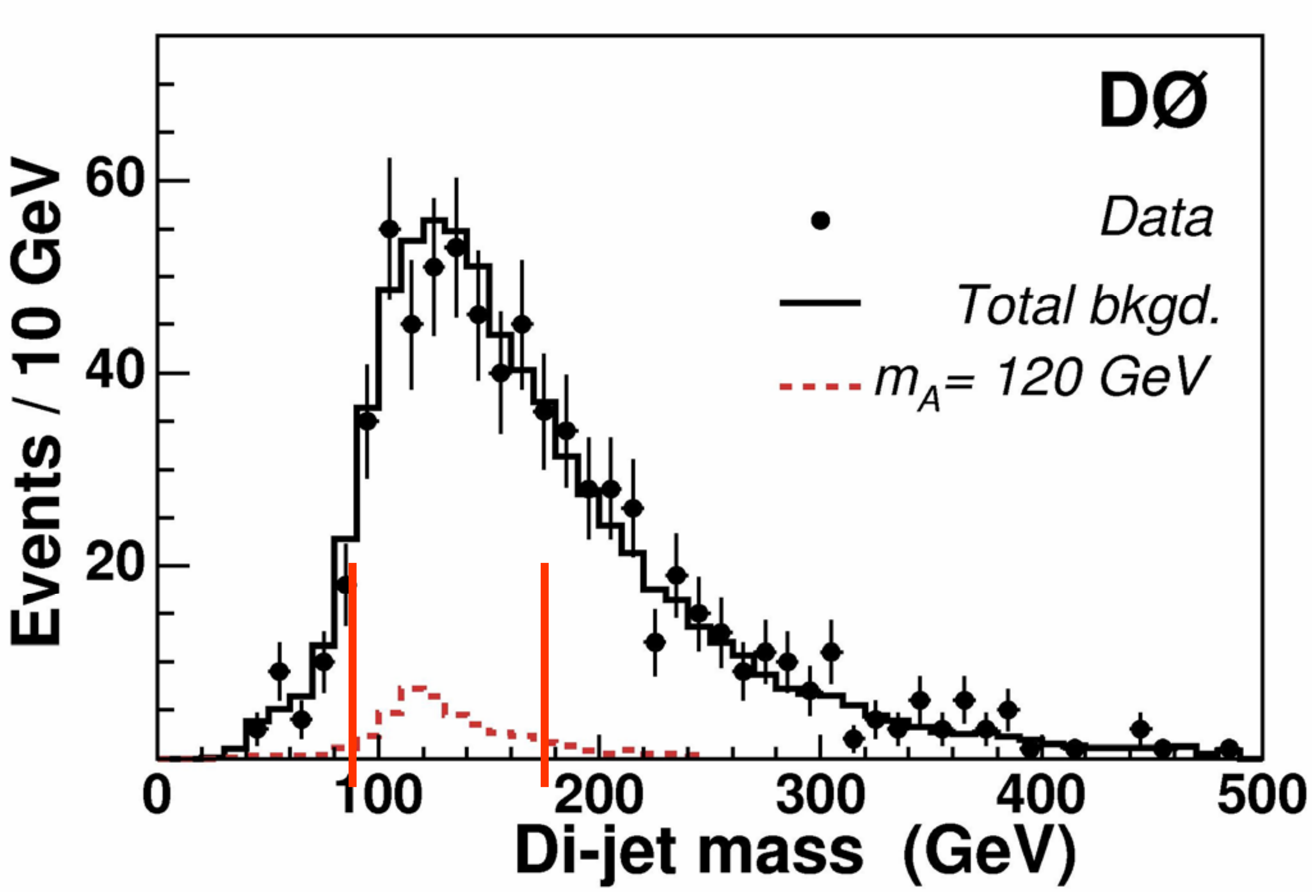
The Traditional Solution to Large, Uncertain Backgrounds: Sideband Fits



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

Sometimes Signal and Background Look Just a Little Too Alike

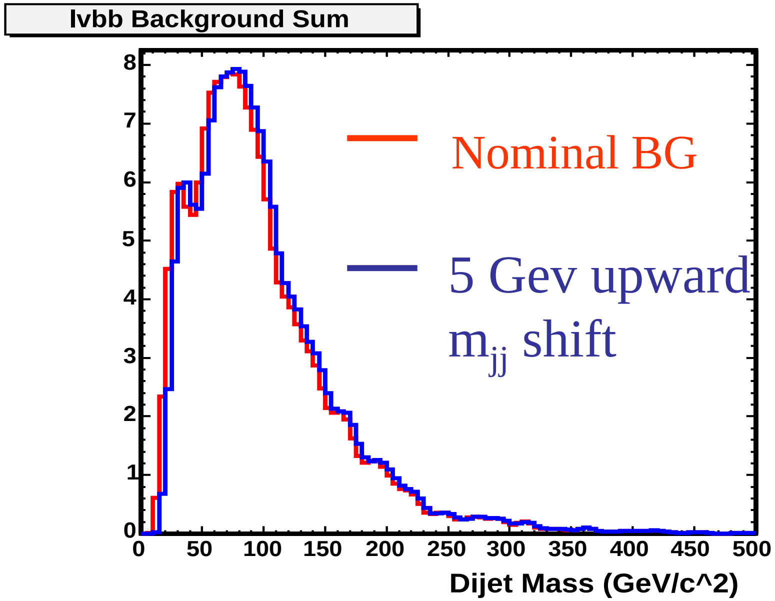
Sideband fit attempted anyway. 2-step process: background rate
in signal window determined from sideband fit. But there could be
signal in sideband. With increasing tanβ, signal gets wider. Made it into PRL.
This tecnique works well when signal is confined to a definite subset of bins.

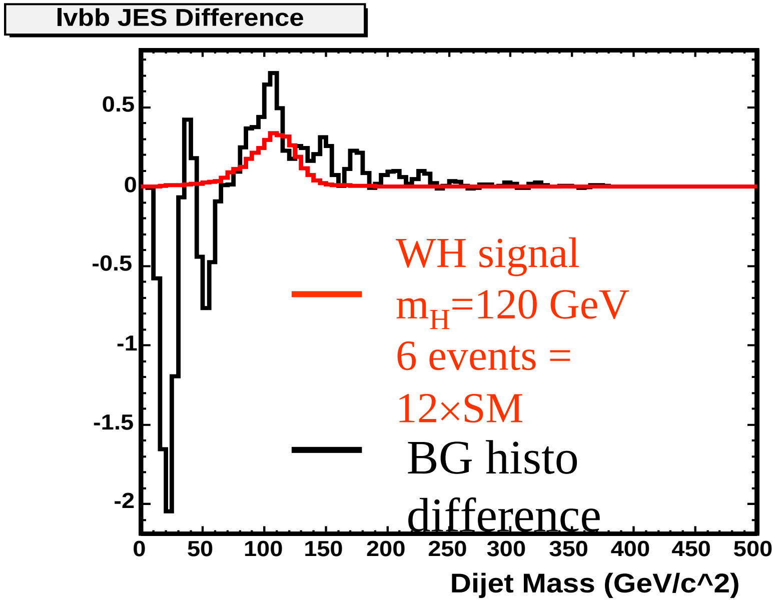


Shape Errors Can Be Show-Stoppers

Sinervo’s Type II errors

$$WH \to \ell\nu b\bar{b}$$





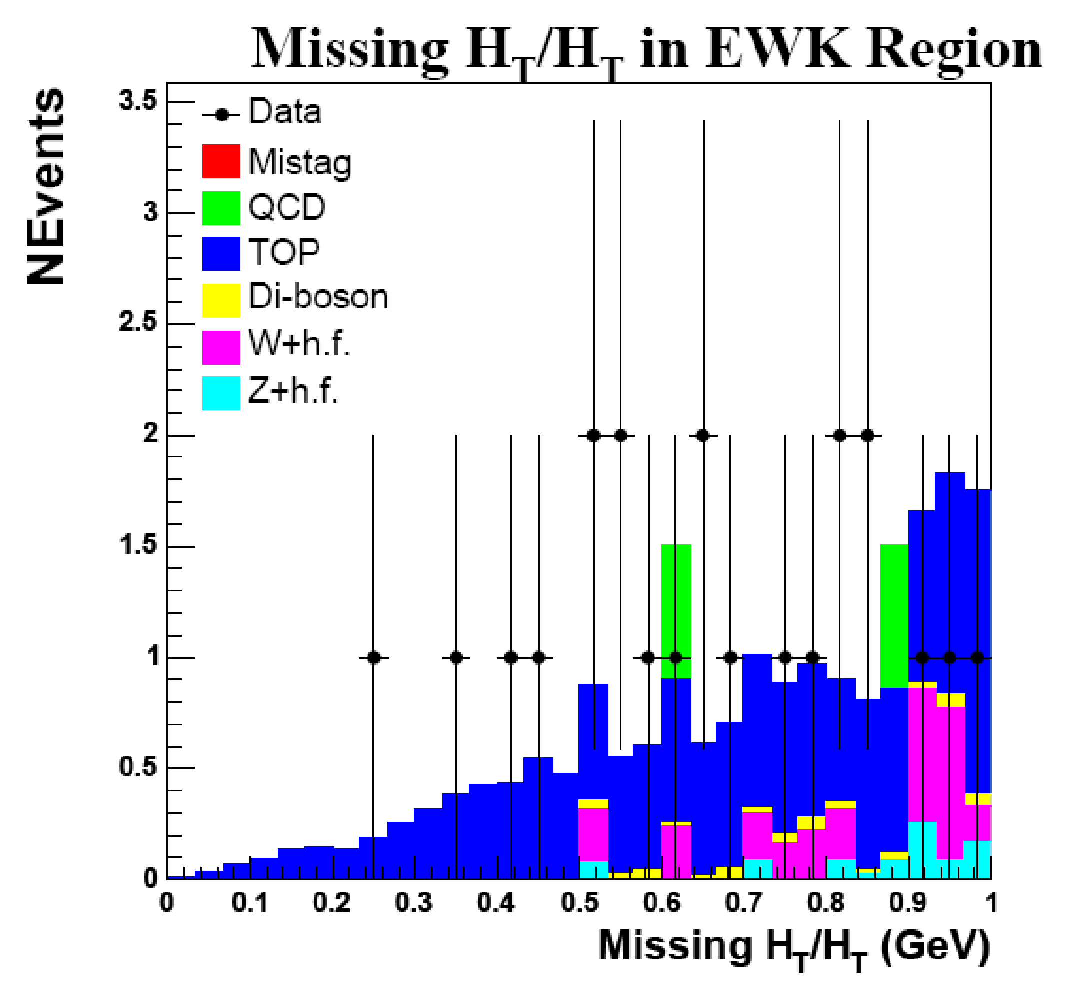
Example of a shifted histogram shape – Compare the difference between nominal and shifted against a signal – easy to fake a signal!

More accurately, if the uncertainty is included, we should be insensitive to such a signal – unable to discover because we are unable to be sure such a bump is a signal or just a background shape that looks a bit different from the central prediction. Need to put them in! Use them to evaluate sensitivity.

Monte Carlo Statistical Errors in Each Bin

SM prediction is a sum of Poisson components with different weights.

“QCD” component is poorly estimated: Just two MC events make it into the histogram.



Often MC statistical uncertainties are simply ignored, or they are treated as overall normalization uncertainties.

Proposed Solution(s) to Complicated Problems

Try both Bayesian and as-Frequentist-as-Possible approaches

Uncertainty on nuisance parameters usually only has a Bayesian interpretation.

- Sometimes we have access to the subsidiary measurement’s data and we can treat it in a frequentist manner, but
- Usually one or more nuisance parameters is a theoretical calculation, or some other number for which we assign some non-frequentist belief distribution. Example: comparing differing Monte Carlo predictions, neither of which can be ruled out by the data.

A mostly-frequentist approach: Use the Likelihood ratio as a test statistic, and simulate pseudoexperiments to find its distribution in H0 and H1. Vary the nuisance parameters according to their Bayesian credibility distributions in the pseudoexperiments to get the “Prior predictive ensemble”

“Cousins and Highland” – See Kyle Cranmer’s talk at Phystat03

Likelihood Ratios with Maximized Likelihoods

- Very similar to Kyle Cranmer's Proposal (PHYSTAT2003). Notation from Kendall and Stuart (almost)

$$Q = \frac{L(x|\theta_r, \hat{\theta}_s)}{L(x|\theta_{r0}, \hat{\hat{\theta}}_s)}$$

x = observation

No maximization is done over θ_r results are independent of the model space considered. Just two hypotheses at a time!

θ_r : Physics parameters in H1

θ_{r0} : Physics parameters in H0

θ_s : Nuisance Parameters (take a superset of all)

$\hat{\theta}_s$ Nuisance parameters which maximize L for the test hypothesis H1

$\hat{\hat{\theta}}_s$ Nuisance parameters which maximize L for the null hypothesis H0

In the Gaussian limit, $-2\ln Q \approx \Delta\chi^2$(H1,H0)

Monte Carlo Statistical Errors in Each Bin

- For each signal and background contribution in each bin, there may be a Monte Carlo statistical uncertainty which is uncorrelated with other bins' MC statistical uncertainty. It could be data too, not just MC statistics.

- Lots of extra nuisance parameters – the true values of each of the rates of each component in each bin. Would like to maximize L over all of these.

- An almost identical problem is sovled by Barlow and Beeston Comp. Phys. Commun 77 219 (1993), for unconstrained fraction fitting. (see TFractionFitter in ROOT). Here we have constraints, and possibly some sources of signal or background not evaluated with MC but with a smooth function, for example.

- Maximizing L with respect to all of these parameters amounts to solving a system of coupled quadratic equations in each bin – tractable numerically.

Asymmetrical Uncertainties

- Common in HEP. Example: $b_1 = 5.0^{+2.1}_{-1.8}$

Explored in detail by Barlow,
ArXiv:physics/0406120

But a question of how to combine a search with b_1 background events and the uncertainty coming from a nuisance parameter *s*, with a second search with $b_2 = 6.1 \pm 1.5$

where the uncertainty arise from the same nuisance parameter *s*.
One's symmetric and the other's asymmetric.

Abstraction: b_1 and b_2 aren't the nuisance parameters, but rather are just functions that depend on a common nuisance parameter *s*. Nuisance parameters are then related to the sources of uncertainty, not the things affected by these uncertainties (which may be affected by many sources of uncertainty).

Treat the nuisance parameter s with a prior that's a unit Gaussian centered on zero.

Parameterize the effect of s on b_1 and b_2 as quadratic functions functions (one of Barlow's options)
or something more complicated (if you don't like the truncated Gaussian prior).

Alex Read's PVMORPH
shape errors:
histogram interpolation

Alternative Nuisance Parameter Priors and Ambiguities in Combination

- Sometimes it’s desirable to have a Gamma prior for some uncertain parameter, such as an acceptance

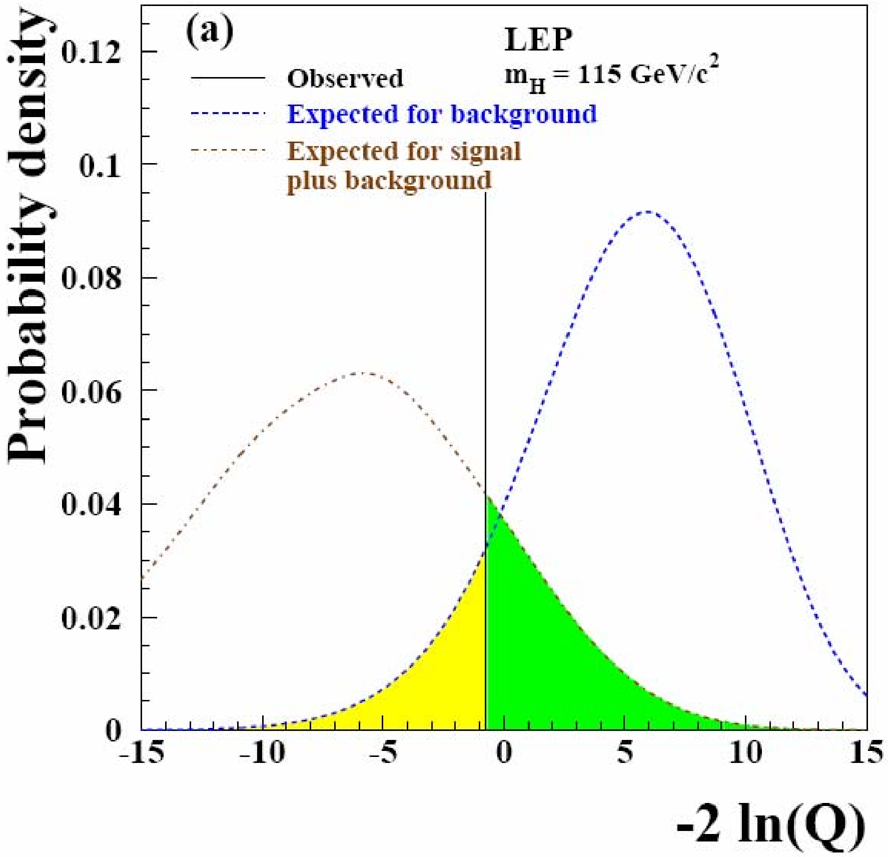
- But sometimes you may have another selection whose acceptance is anticorrelated with the first (common when dividing selections up into disjoint pieces to spread work across different teams of people).

$$\epsilon_1 = \epsilon_1^0 \pm \sigma_\epsilon$$

$$\epsilon_2 = \epsilon_2^0 \mp \sigma_\epsilon$$

They can’t both have a Gamma Prior!

Two-hypothesis test – Form test statistic out of all input histograms, and use Pseudoexperiments to find the distributions. Plot observation.

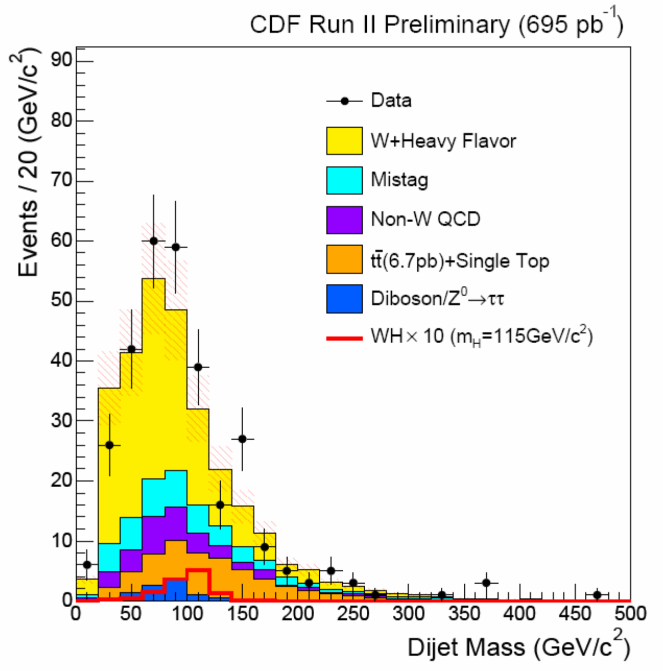


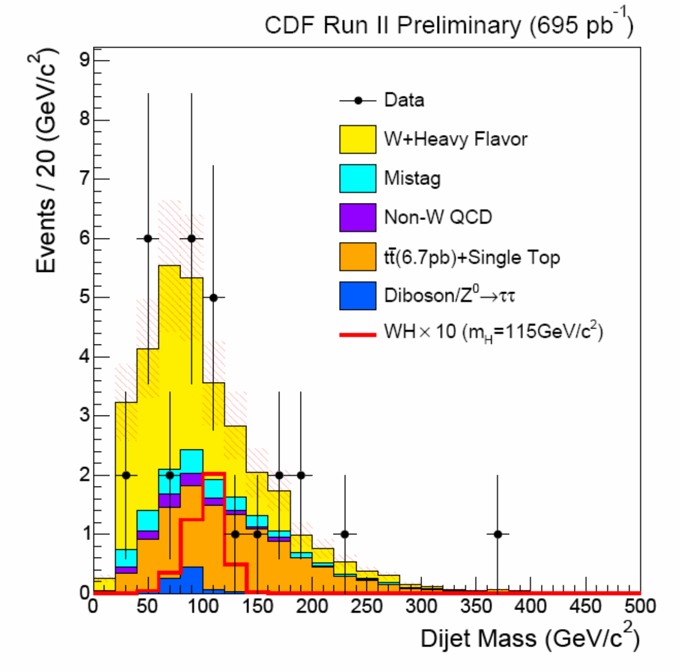
Yellow: 1-CL_b

Green: CL_{s+b}

At LEP, life was easy with large s/b ratios and small backgrounds. At the Tevatron and LHC, we are in trouble: low s/b, large backgrounds.

Back to the CDF WH Search





- 20% uncertainty on backgrounds, signals are very small (shown x10 for visibility).

- Background rates, shapes, signal acceptances fluctuate wildly in the prior-predictive ensemble – correalate all errors.
  Unknown true values in the data.

How to cope?

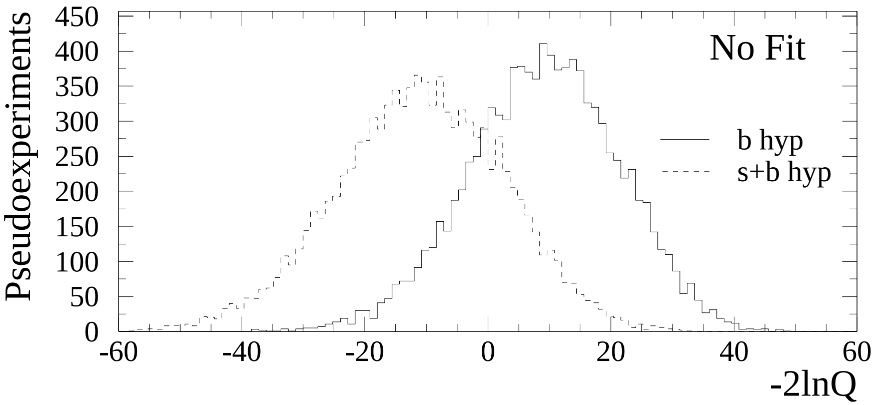
Answer – Maximizing L with respect to Nuisance Parameters improves sensitivity.

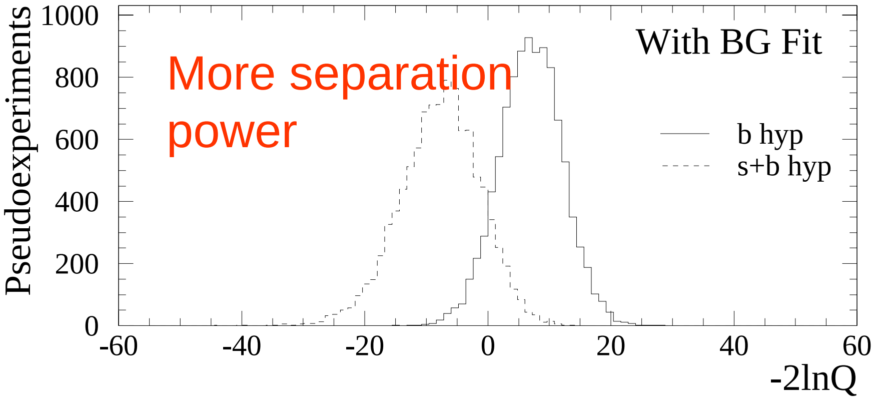
- Sometimes we think of maximizing L in a fit as a way of including the effects of nuisance parameters. But that's already done with the prior predictive ensemble when finding the PDF's of -2lnQ

- Fitting for the best values of the nuisance parameters makes both H1 and H0 fit the data (pseudodata) better.

- But it can improve the separation in PDF(-2lnQ) in the two hypotheses due to constraints on uncertain parameters:

Two sideband fits (really whole-spectrum fits) on each pseudoexperiment. In HEP we use MINUIT.




More separation power

Computing 5σ Significances

Hard to do MC integration to measure a fraction of 2.85×10^{-7} – need $\mathcal{O}$(billion) pseudoexperiments (x4xMINUIT) to discover something the brute force way.

A technique used at LEP (A. Read told me about this):

P(-2lnQ|H1)/P(-2lnQ|H0) = Q

You can reweight the outcome of the pseudoexperiments by Q to get the H1 PDF from H0's and vice-versa.

Only problem – a data outcome that's not represented well either by H0 or H1's distribution.

With systematic variations on the pseudoexperiments, the weight is a systematically varied Q (but keep the systematically unvaried Q as the test statistic! – see later).

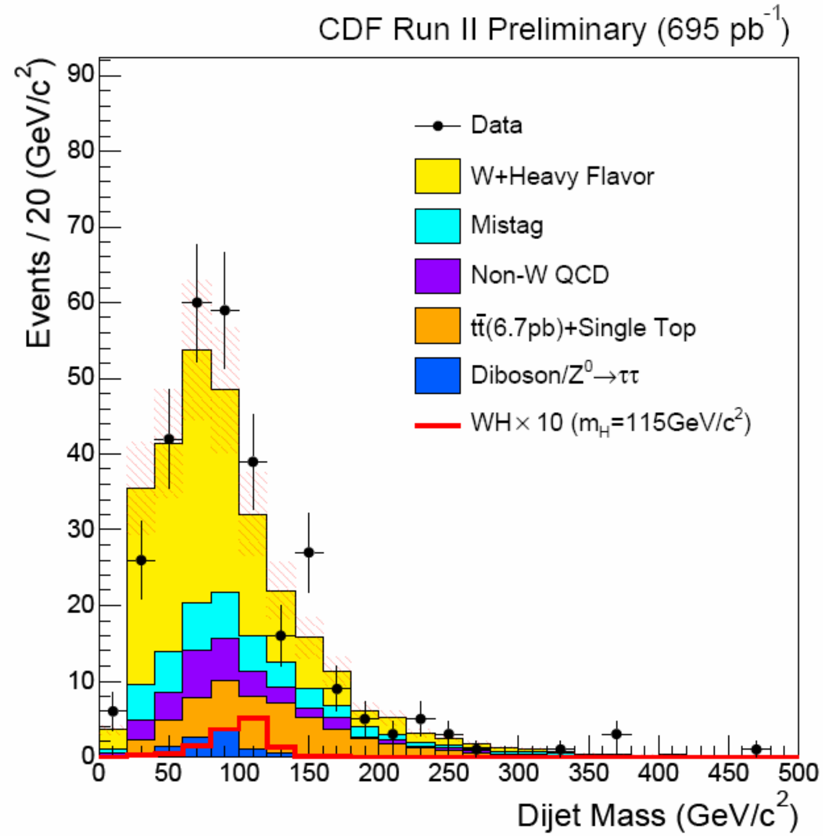
But – maximization of L over nuisance parameters ruins this.
P(-2lnQ|H1)/P(-2lnQ|H0)=αQ observed in at least one example for some α. General?

The Concrete Example: CDF's WH →lvbb, single-tag exclusive channel, L=695 pb^{-1}

CL$_s$ limit = 59.4*SM
expected limit= 49.0*SM

CL$_s$ limit with fits = 40.3*SM,
expected limit = 26.7*SM

Bayesian limit with marginalization:
Obs limit= 40.4*SM,
expected limit= 26.7*SM



Varying Likelihood according to C+H?

- Q (and the likelihoods) are just test statistics. They are not integrated over with a Cousins and Highland variation of nuisance parameters.

  - Q is just a function of the observed data (which is not uncertain) and functions which we choose (which are also not uncertain)

  - The distributions of Q in H0 and H1 are uncertain
  - Varying Q in a Cousins and Highland way “splits outcomes”

  -- addition of even the smallest uncertainty in a nuisance parameter can make a limit jump by a huge amount.
  (but see comment on addition of even the smallest amount of extra experimental information, in the form of a new channel).

  - We hope MINUIT gives us the same answer every time an identical problem is posed to it.

Need to Integrate over Nuisance Parameters even When Maximizing L

$$Q = \frac{L(x|\theta_r, \widehat{\theta}_s)}{L(x|\theta_{r0}, \widehat{\widehat{\theta}}_s)}$$

is just an ordering rule – we still do not know its distribution in H0 and H1 because of the nuisance parameters

Simplest example: one-bin counting experiment.

- All (sensible) ordering rules are equivalent to the event count.
- Q has a distribution of a sum of delta functions at fixed locations.
- Ignoring the variation in the nuisance parameters when making the distribution of Q is the same as ignoring systematic errors.

Look-Elsewhere for Exclusion too?

- Well-known effect produces false discoveries at any desired significance level if enough independent experiments are done.
    - Called “look-elsewhere”, “trials factor”, “greedy bump bias”, other names.
    - Usual prescription – dilution of significance based on how many independent experiments are conducted.

      Best practice – simulate actual procedure in Monte Carlo pseudoexperiments and see what the PDF of the lowest p-value is in the null hypothesis.

- In absence of conditioning, you get 5% false exclusions at 95% CL. Do we need to dilute the exclusion significance too when performing many independent tests? How about tests where we know we have no sensitivity (we can construct infinitely many of these).

Bayesian Approach to the General Problem

The marginalization approach:

$$0.95 = \frac{\int \int_0^{\theta_{r95}} L(x|\theta_r, \theta_s)\pi(\theta_r, \theta_s) d\theta_r d\theta_s}{\int \int_0^{\theta_{cut}} L(x|\theta_r, \theta_s)\pi(\theta_r, \theta_s) d\theta_r d\theta_s}$$

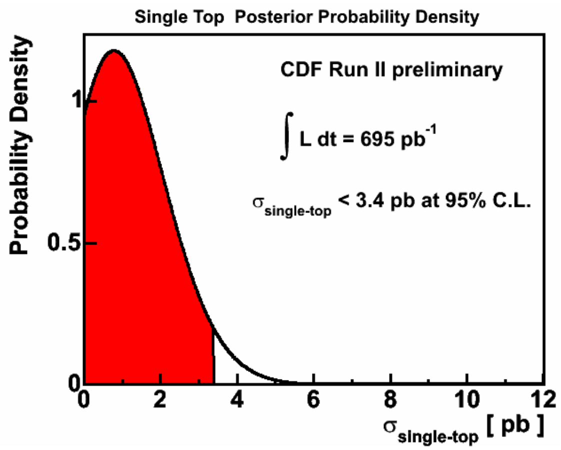
With a carefully chosen prior π, θ_{cut} can be infinity.

Typically the physics parameter integrated over is a cross section.
Other physics parameters, like branching fractions or masses
is typically not used and have pitfalls.

Profiling: instead of integrating over nuisance parameters, maximize L with respect to them.

Benefiting from the Sidebands with a Marginalized Bayesian Limit

$$0.95 = \frac{\int \int_0^{\theta_{r95}} L(x|\theta_r, \theta_s)\pi(\theta_r, \theta_s)d\theta_r d\theta_s}{\int \int_0^{\theta_{cut}} L(x|\theta_r, \theta_s)\pi(\theta_r, \theta_s)d\theta_r d\theta_s}$$



An Example. Similar shape, but much sharper when you plot

L×π as a function of the background rate.

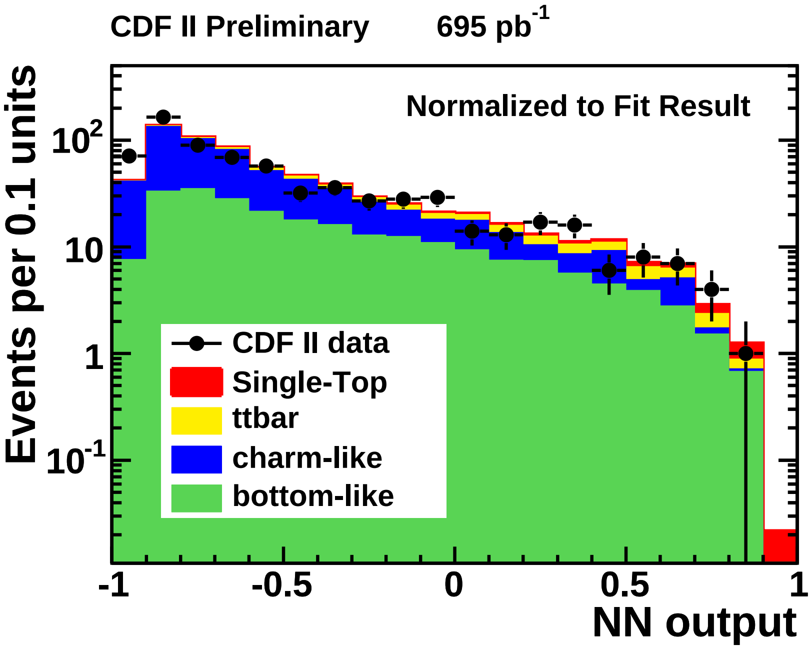
Integral emphasizes those values of θ_s most consistent with the data, even in bins with no signal expectation.

CDF WH search's Bayesian limit was nearly independent of prior background uncertainty

"We, along with other immigrant groups, have been the targets of detentions, deportations, marginalization, profiling, and now increased criminalization."

H. Soliveres, The Filipino Express Online (found with a Google search for "marginalization" and "profiling")

A Pitfall To Avoid with Marginalized Bayesian Limits



If you use the observed data to normalize the expected background rates and to define π for the background distribution,

Using the entire histogram again in a marginalized Bayesian limit calculation now uses the same data constraint twice, for a factor of sqrt(2) in constraint.

Suggestion: get π from another subsidiary experiment or other estimation.

Some Software and a Suggestion

Software and documentation provided at

http://www.hep.uiuc.edu/home/trj/cdfstats/mclimit_csm1/index.html

- Detailed note provided which documents the interfaces.
- Uses ROOT histogram classes (1D and 2D) and ROOT tools.

- User provides histograms of signal and background templates, (with unit-weight entries in order to do Poisson statistical errors) normalization factors, rate uncertainties, shape variation templates.

- Bayesian and CLs limits calculated. P-values for discovery (1-CLb) calculated. Expected limits and expected p-values computed.

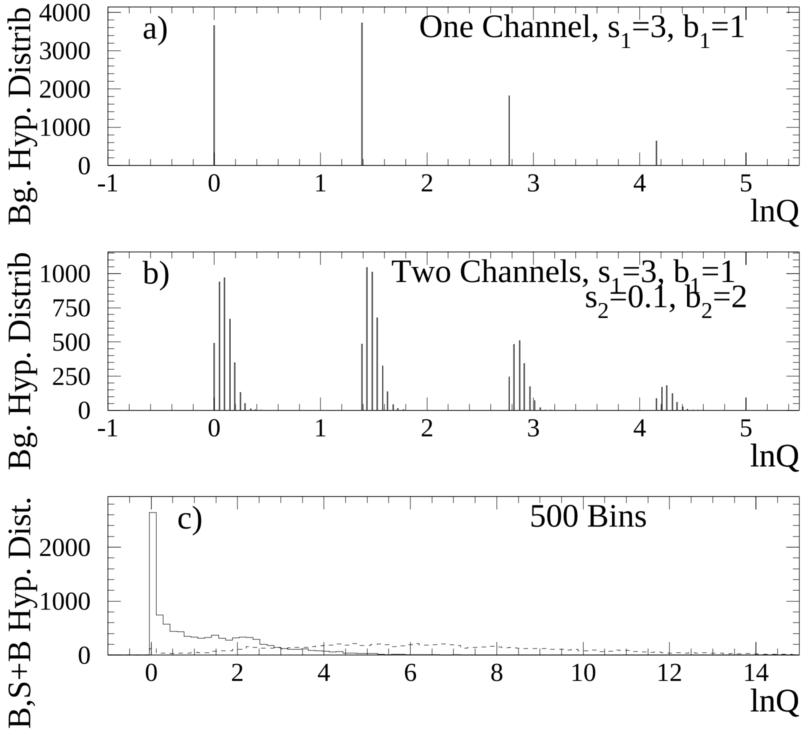
Interface is rather cumbersome due to the large number of possible systematic errors which need to be included. This is all much easier with Gaussian statistics, but some problems are mixed!

Suggsetion: Pick the method that has correct coverage, but maximizes the discovery potential while including all errors. I am a pragmatist and do not concern myself with philosophical purity of method.

An Odd Feature of P-values when Combining Two Bins

- Poisson Statistics regime – large probabilities of getting specific outcomes

- A weak channel is combined with a strong channel. Your limit can jump discontinuously when a weak channel is combined in. A sign of non-optimality.



One Channel: s_1=3, b_1=1

Two channels: add in
s_2=0.1, b_2=2

“split outcomes”

Continuous case:
$ds/dx = Ae^{-Bx}$,
$db/dx = 1$
Bin1: 0<x<1, Bin2: 1<x<3
A=10.6, B=3.4. Same s and b as two-bin case above, but you can choose more bins than that.