Statistical Methods for Experimental Particle Physics

Theory and Lots of Examples

Thomas R. Junk
Fermilab

TRIUMF Summer Institute
July 20 - 31, 2009

Day 3: Bayesian Inference
Miscellaneous Topics

Reasons for Another Kind of Probability

- So far, we've been (mostly) using the notion that probability is the limit of a fraction of trials that pass a certain criterion to total trials.

- Systematic uncertainties involve many harder issues. Experimentalists spend much of their time evaluating and reducing the effects of systematic uncertainty.

- We also want more from our interpretations -- we want to be able to make decisions about what to do next.
  - Which HEP project to fund next?
  - Which theories to work on?
  - Which analysis topics within an experiment are likely to be fruitful?

  These are all different kinds of bets that we are forced to make as scientists. They are fraught with uncertainty, subjectivity, and prejudice.

  Non-scientists confront uncertainty and the need to make decisions too!

Bayes' Theorem

Law of Joint Probability:

$$p(A \text{ and } B) = p(A|B)p(B) = p(B|A)p(A)$$

Events A and B interpreted to mean "data" and "hypothesis"

$$p(\{\nu\}\,|\,data) = \frac{L(data\,|\,\{\nu\})\pi(\nu)}{\int L(data\,|\,\{\nu'\})\pi(\{\nu'\})d\{\nu'\}}$$

$\{x\}$ = set of observations

$\{\nu\}$ = set of model parameters

A frequentist would say: Models have no "probability". One model's true, others are false. We just can't tell which ones (maybe the space of considered models does not contain a true one).

Better language: $p(\{\nu\}\,|\,data)$

describes our belief in the different models parameterized by $\{\nu\}$

Bayes' Theorem

$p(\{\nu\} \mid data)$ is called the "posterior probability" of the model parameters

$\pi(\{\nu\})$ is called the "prior density" of the model parameters

The Bayesian approach tells us how our existing knowledge before we do the experiment is "updated" by having run the experiment.

This is a natural way to aggregate knowledge -- each experiment updates what we know from prior experiments (or subjective prejudice or some things which are obviously true, like physical region bounds).

Be sure not to aggregate the same information multiple times! (groupthink)

We make decisions and bets based on all of our knowledge and prejudices

"Every animal, even a frequentist statistician, is an informal Bayesian." See R. Cousins, "Why Isn't Every Physicist a Bayesian", Am. J. P., Volume 63, Issue 5, pp. 398-410

How I remember Bayes's Theorem

$$p(\text{hypothesis}|\text{data}) = \frac{p(\text{data}|\text{hypothesis}) \times p(\text{hypothesis})}{p(\text{data})}$$

Posterior "PDF"
("Credibility")

"Likelihood Function"
("Bayesian Update")

"Prior belief
distribution"

Normalize this so that

$$\int p(\text{hypothesis}|\text{data}) d(\text{hypothesis}) = 1$$

for the observed data

Bayesian Application to HEP Data: Setting Limits on a new process with systematic uncertainties

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

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 in a realistic example.

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

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

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

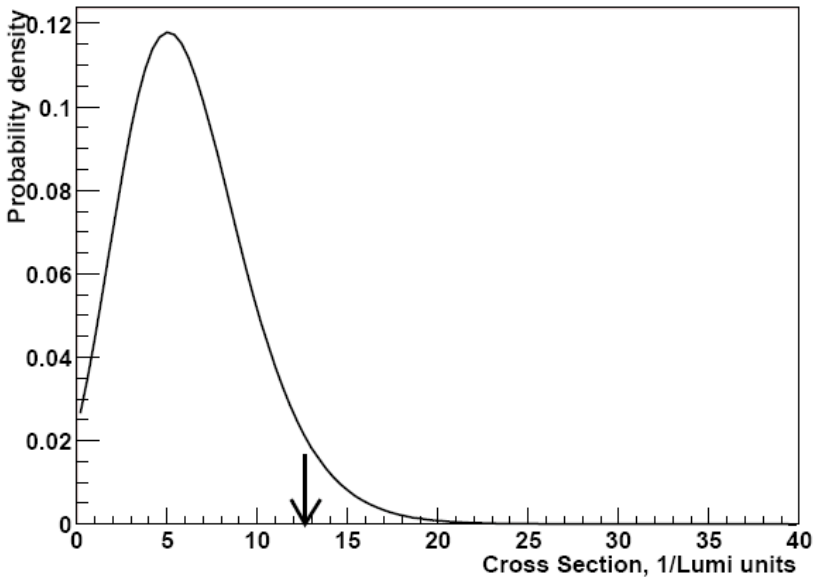
$$r = r_{\max}{}^{+(r_{high}-r_{\max})}_{-(r_{\max}-r_{low})}$$

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

Be Explicit About Introduction of Priors

- Typical example of a Bayesian calculation of a 95% CL upper limit

DØ (www-d0.fnal.gov/~hobbs/limit_calc.html) has a web based menu driven product which is interesting—similar to `bayes.f`, but produces limit (95% only) and posterior plot online:




```
Data:  10
Background:  5 +- 1
Efficiency:  1.0 +- 0.1
Luminosity:  1.0 +- 0.0
```

```
The cross section 95% CL upper limit is 12.666
```

Sensitivity of upper limit to Even a "flat" Prior

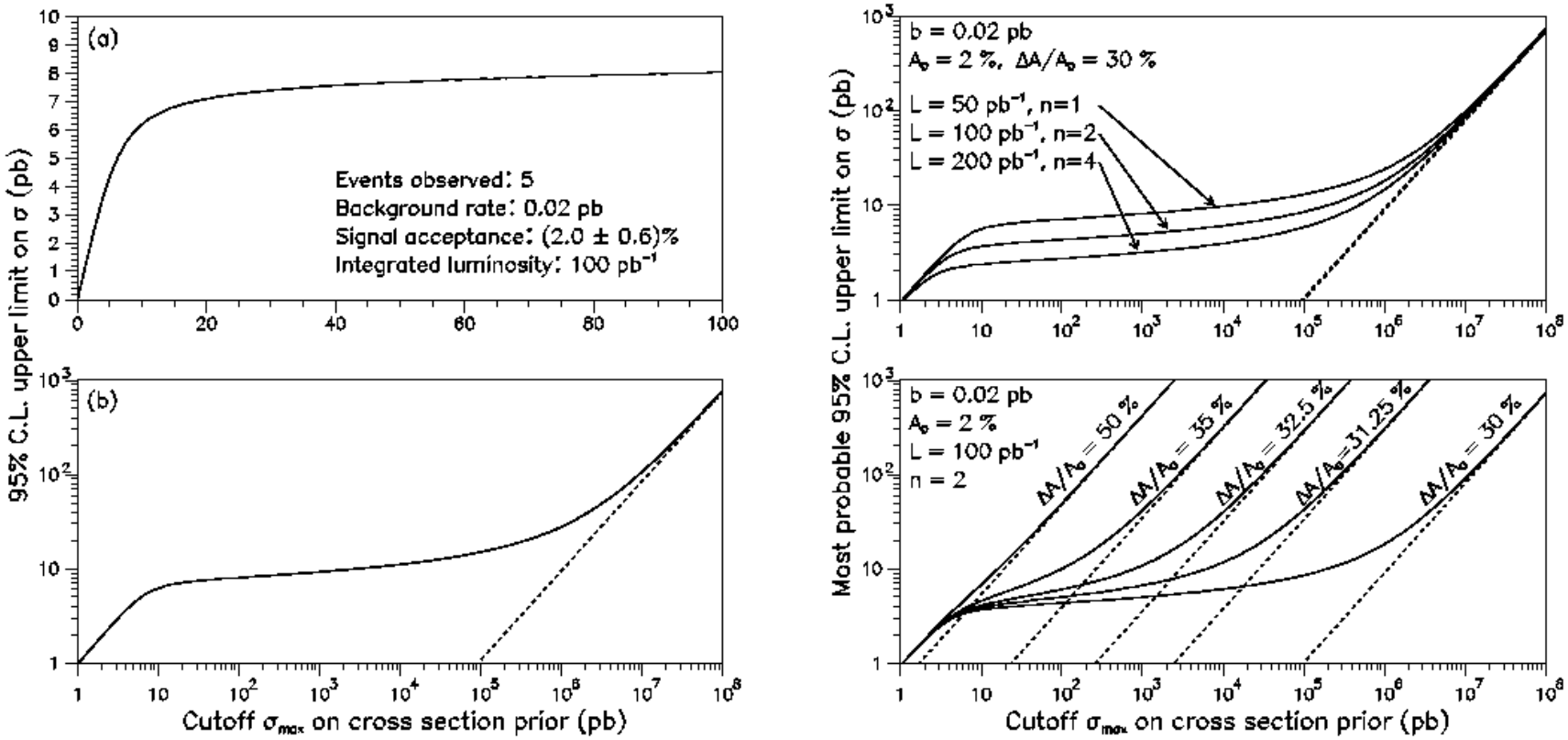


Figure 1: Bayesian upper limits at the 95% credibility level on a hypothetical cross section σ, as a function of the cutoff $\sigma_{\max}$ on the flat prior for σ.

L. Demortier, Feb. 4, 2005

Systematic Uncertainties

Encoded as priors on the nuisance parameters $\pi(\{\theta\})$.

Can be quite contentious -- injection of theory uncertainties and results from other experiments -- how much do we trust them?

Do not inject the same information twice.

Some uncertainties have statistical interpretations -- can be included in L as additional data. Others are purely about belief. Theory errors often do not have statistical interpretations.

Aside: Uncertainty on our Cut Values? (ans: no)

- Systematic uncertainty -- covers unknown differences between model predictions and the "truth"
- We know what values we set our cuts to.
- We aren't sure the distributions we're cutting on are properly modeled.
- Try to constrain modeling with control samples (extrapolation assumptions)
- Estimating systematic errors by "varying cuts" isn't optimal -- try to understand bounds of mismodeling instead.

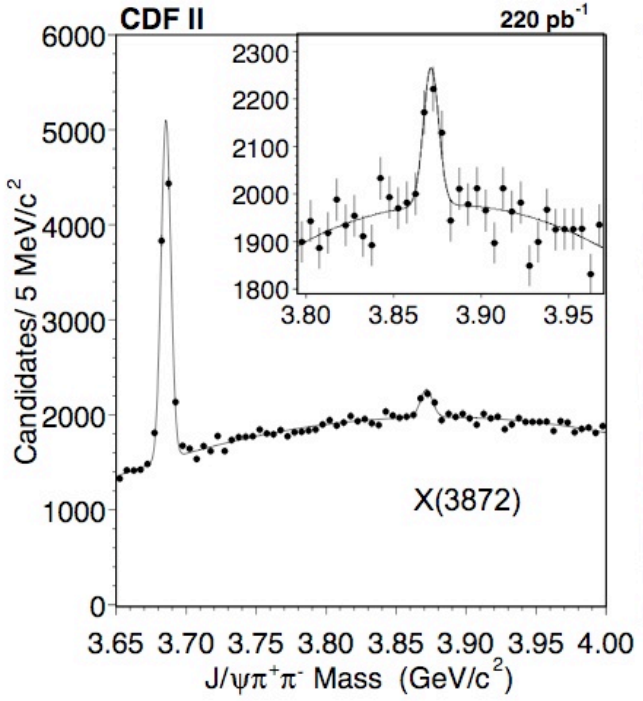
Integrating over Systematic Uncertainties Helps Constrain their Values with Data

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

Nuisance parameters: θ

Parameter of Interest: r

Example: suppose we have a background rate prediction that's 50% (fractionally) uncertain -- goes into $\pi(\theta)$. But only a narrow range of background rates contributes significantly to the integral. The kernel falls to zero rapidly outside of that range.



Can make a posterior probability distribution for the background too -- narrow belief distribution.

Coping with Systematic Uncertainty

- "Profile:"
  - Maximize *L* over possible values of nuisance parameters include prior belief densities as part of the χ^2 function (usually Gaussian constraints)

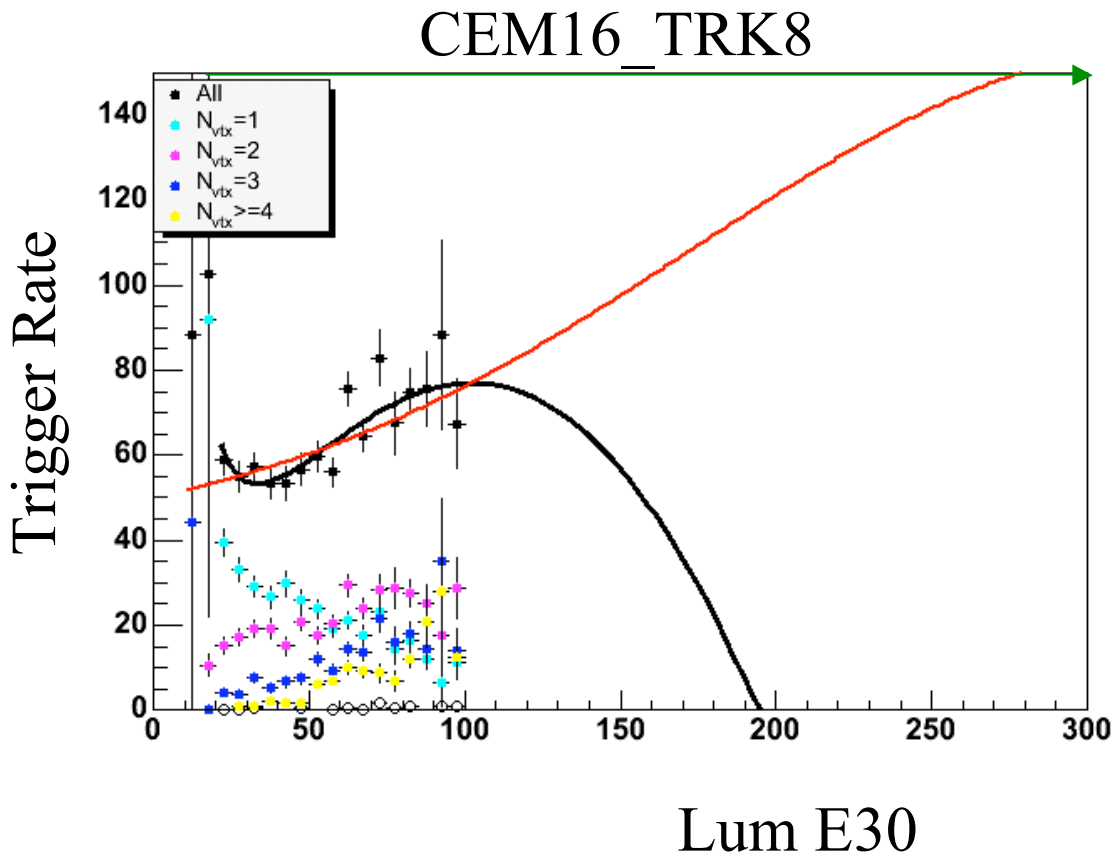
- "Marginalize:"
  - Integrate L over possible values of nuisance parameters (weighted by their prior belief functions -- Gaussian, gamma, others...)
  - Consistent Bayesian interpretation of uncertainty on nuisance parameters

- Aside: MC "statistical" uncertainties are systematic uncertainties

Example of a Pitfall in Fitting Models

- Fitting a polynomial with too high a degree
- Can extrapolations be trusted?

Trigger x-section extrapolation vs. luminosity



Even Bayesians have to be a little Frequentist

- A hard-core Bayesian would say that the results of an experiment should depend only on the data that are observed, and not on other possible data that were not observed.

  Also known as the “likelihood principle”

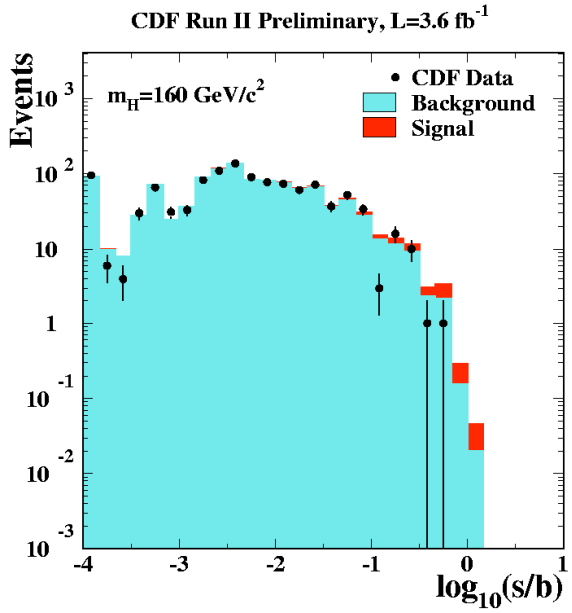
- But we still want the sensitivity estimated! An experiment can get a strong upper limit not because it was well designed, but because it was lucky.

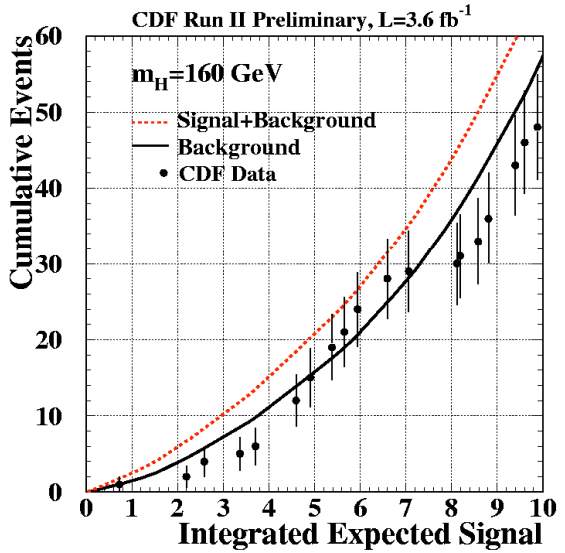
  How to optimize an analysis before data are observed?

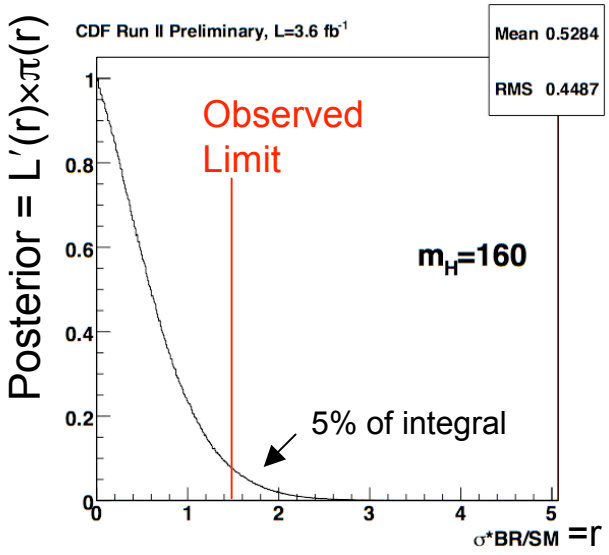
  So -- run Monte Carlo simulated experiments and compute a Frequentist distribution of possible limits. Take the median-- metric independent and less pulled by tails.

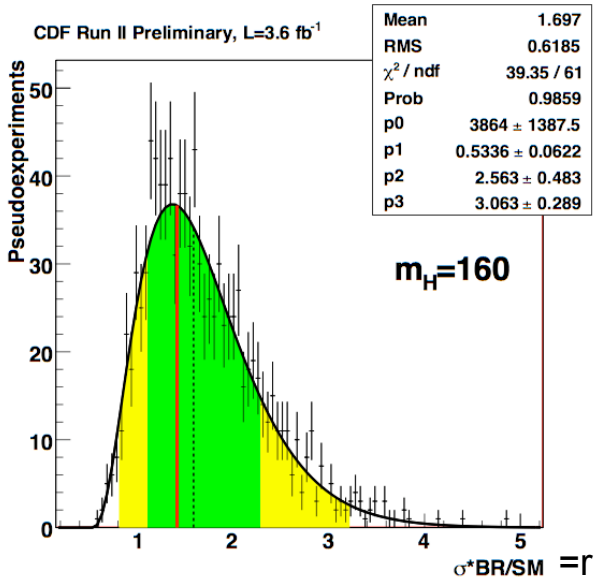
  But even Bayesian/Frequentists have to be Bayesian: use the Prior-Predictive method -- vary the systematics on eachc pseudoexperiment in calculating expected limits. To omit this step ignores an important part of their effects.

Bayesian Example: CDF Higgs Search at m_H=160 GeV



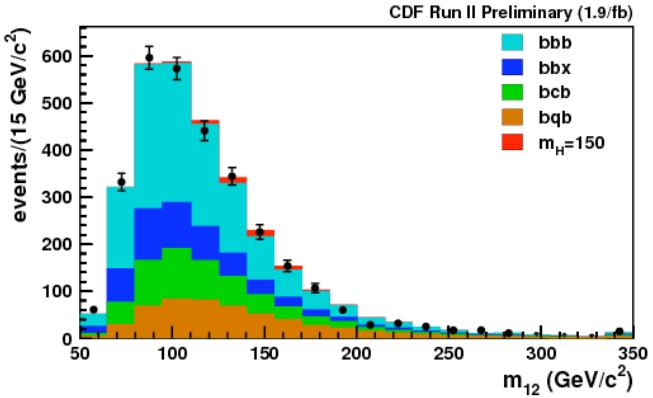


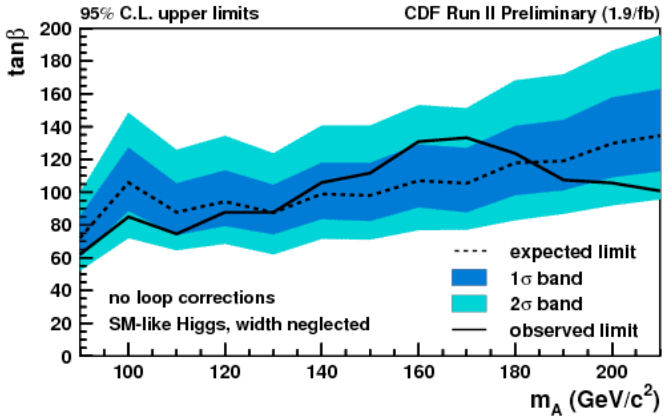


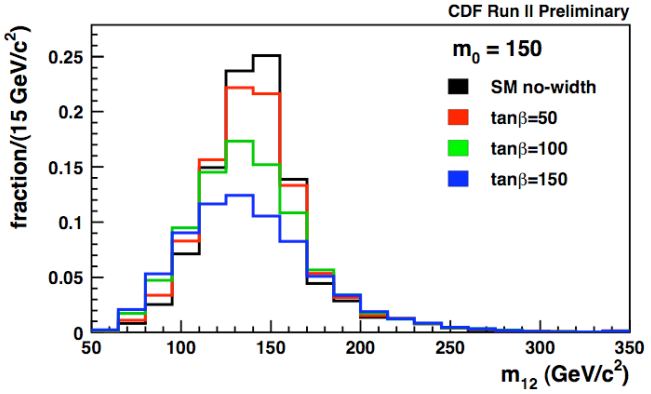


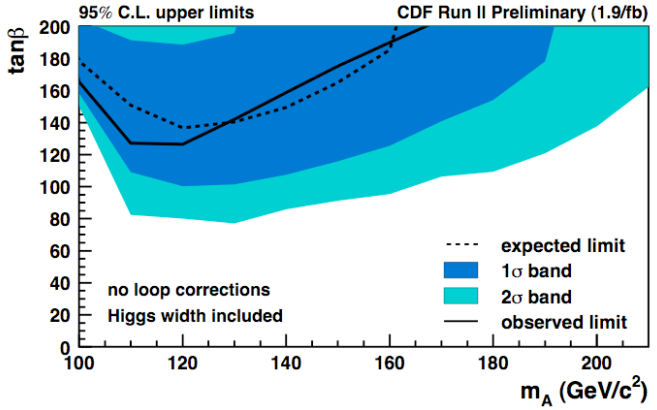
An Example Where Usual Bayesian Software Doesn't Work

- 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



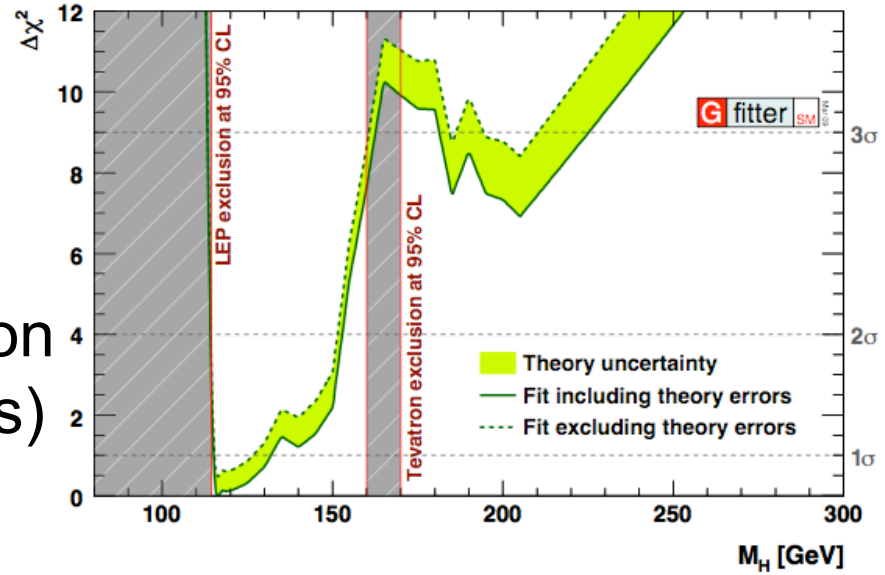




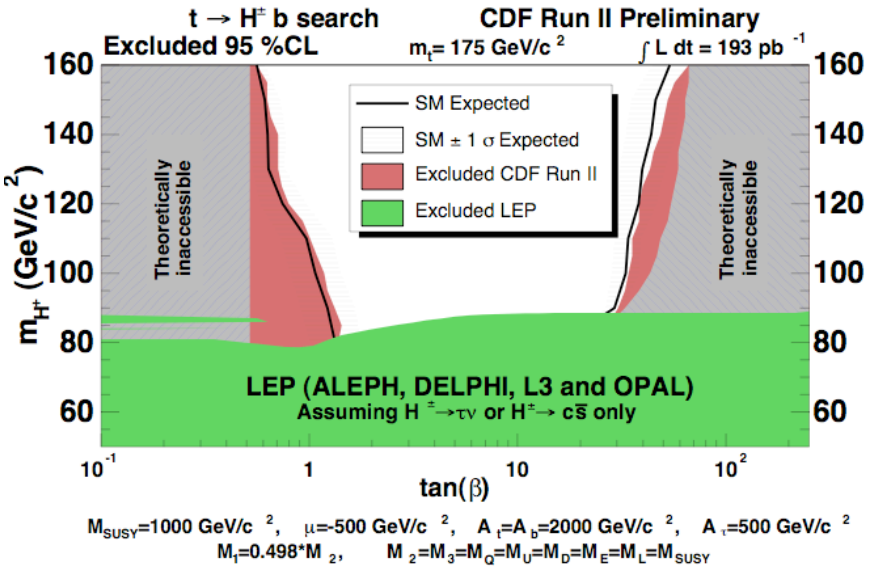


Priors in Non-Cross-Section Parameters

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..



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 over nuisance parameters in the numerator and denominator.

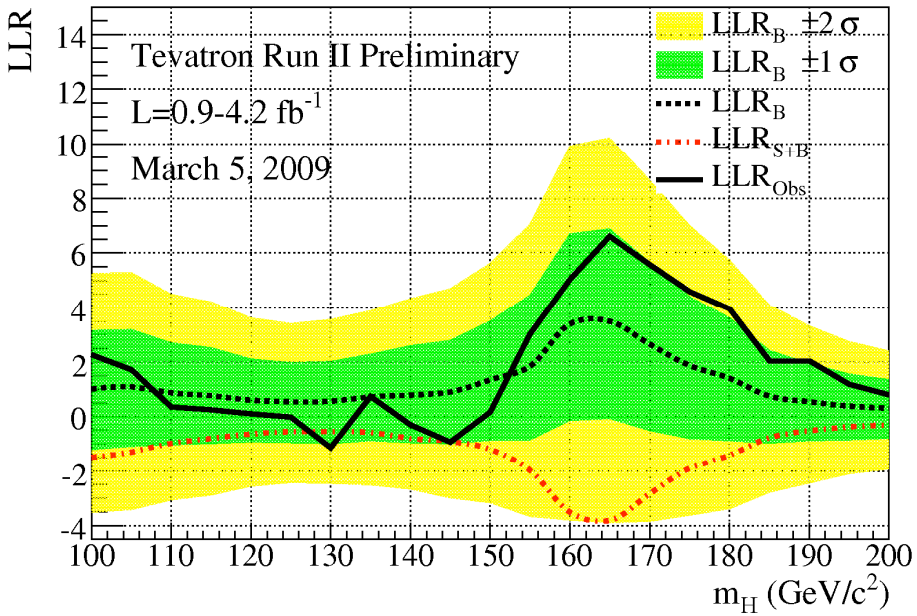
Similar criteria for evidence, discovery as profile likelihood.

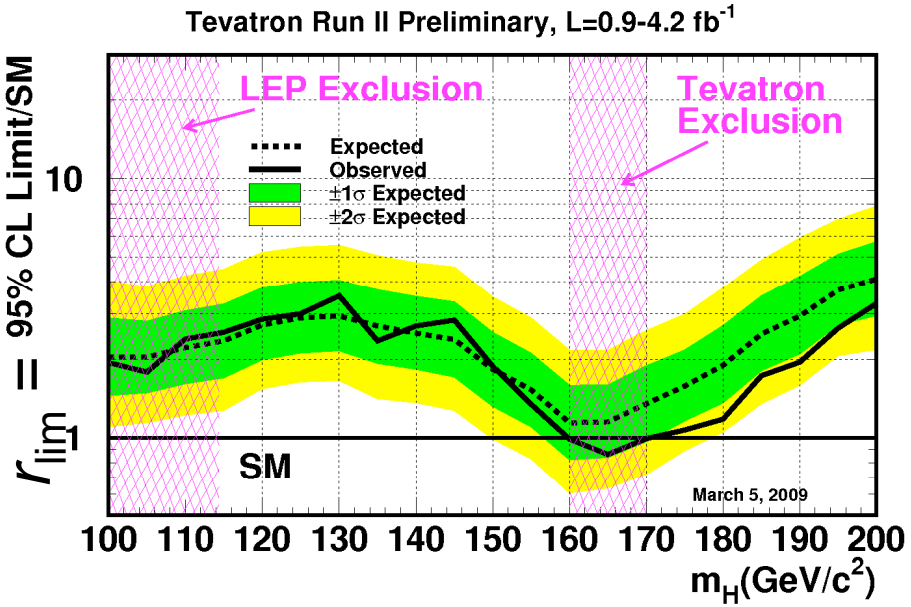
Physicists would like to check the false discovery rate,
and then we're back to p-values.

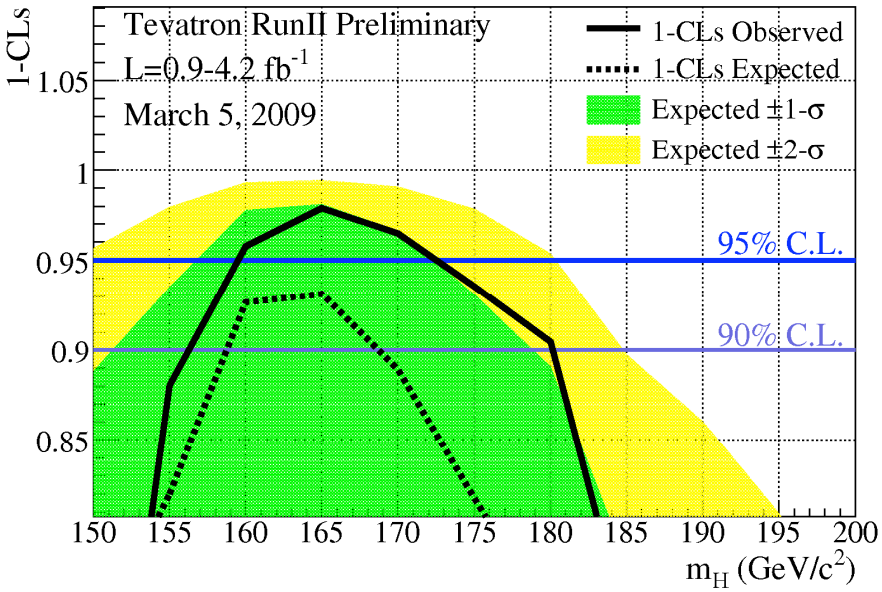
But -- odd behavior of *B* compared with p-value for even a simple case

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

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.

Measurement and Discovery are Very Different

Buzzwords:

- Measurement = “Point Estimation”
- Discovery = “Hypothesis Testing”

You can have a discovery and a poor measurement!
Example: Expected $b=2\times10^{-7}$ events, expected signal=1 event, observe 1 event, no systematics.

p-value ~2×10^{-7} is a discovery! (hard to explain that event with just the background model). But have ±100% uncertainty on the measured cross section!

In a one-bin search, all test statistics are equivalent. But add in a second bin, and the measured cross section becomes a poorer test statistic than the ratio of profile likelihoods.

In all practicality, discriminant distributions have a wide spectrum of s/b, even in the same histogram. But some good bins with b<1 event

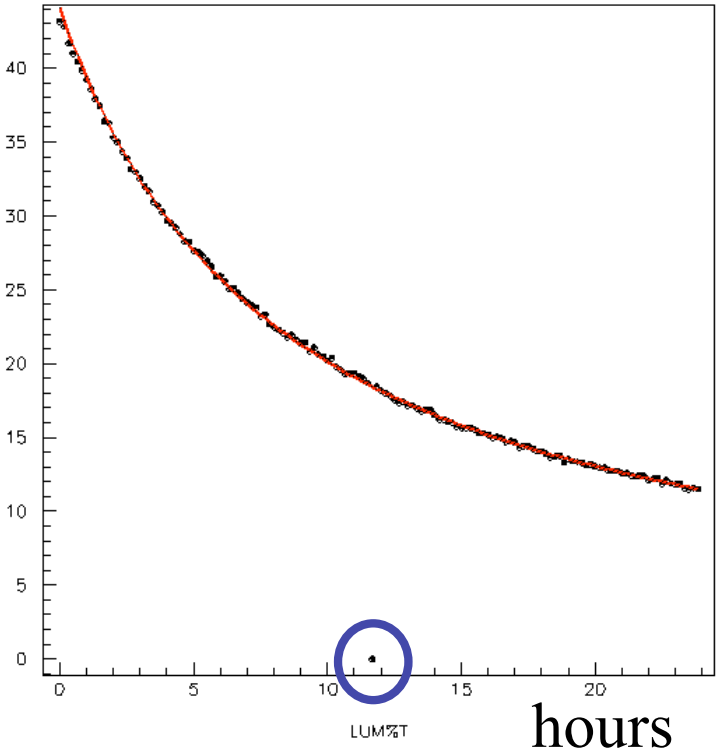
Advantages and Disadvantages of Bayesian Inference

- Advantages:
  - Allows input of *a priori* knowledge:
    - positive cross-sections
    - positive masses
  - Gives you "reasonable" confidence intervals which don't conflict with *a priori* knowledge
  - Easy to produce cross-section limits
  - Depends only on observed data and not other possible data
  - No other way to treat uncertainty in model-derived parameters
- Disadvantages:
  - Allows input of *a priori* knowledge (AKA "prejudice") (be sure not to put it in twice...)
  - Results are metric-dependent (limit on cross section or coupling constant? -- square it to get cross section).
  - Coverage not guaranteed
  - Arbitrary edges of credibility interval (see freq. explanation)

Outliers

- Sometimes they're obvious, often they are not.
- Best to make sure that the uncertainties on all points honestly include all known effects. Understand them!

Lum E30



L. Ristori,
Instantaneous
Luminosity vs. time
(a store in 2005)

Summary

Statistics, like physics, is a lot of fun!

It's central to our job as scientists, and about how human knowledge is obtained from observation.

Lots of ways to address the same problems.

Many questions do not have a single answer. Room for uncertainty. Probability and uncertainty are different but related.

Think about how your final result will be extracted from the data before you design your experiment/analysis -- keep thinking about it as you improve and optimize it.

Thanks

To You!

To the organizers, Isabel, Bernd, Rob, ...

Extra Material

Bayesian Upper Limit Calculation

$$L(s) = \frac{(s+b)^n e^{-(s+b)}}{n!}$$

data = n

b = background rate

s = signal rate (= cross section when luminosity=1)

Multiply by a flat prior π(s) = 1 and find the limit by integrating:

$$0.95 = \int_0^{s_{\mathrm{lim}}} L(s)\pi(s)ds$$

Not too tricky; easy to explain.

- But where did π(s) come from?
- What to do about systematic uncertainty on signal and background?

Frequentist Analysis of Significance of Data

- Most experiments yield outcomes with measure ~0

- A better question: Assuming the null hypothesis is true, what are the chances of observing something as much like the test hypothesis as we did (or more)?
  used to reject the null hypothesis if small

- Another question: If test hypothesis is true, what are the chances that we'd see something as much like the null hypothesis as we did (or more)?
  used to reject the test hypothesis if small

It is possible to reject **both** hypotheses! (but not with C+F or Bayesian techniques).

Frequentist Interpretation of Data

- Relies on an abstraction -- an infinite ensemble of repetitions of the experiment. Can speak of probabilities as fractions of experiments.

- Constructed to give proper coverage:

  95% CL intervals contain the true value 95% of the time, and do not contain the true value 5% of the time, if the experiment is repeated.

- Two kinds of errors:
  - Accepting test hypothesis if it is false
  - Excluding test hypothesis if it is true
- Two kinds of success
  - Accepting test hypothsis if it is true
  - Excluding test hypothesis if it is false

Difference between "power" and "coverage"

Undesirable Behavior of Limit-Setting Procedures

- Empty confidence intervals: we know with 100% certainty that an empty confidence interval doesn't contain the true value, even though the technique produces correct 95% coverage in an ensemble of possible experiments. Odd situation when we know we're in the "unlucky" 5%.

- Ability of an experiment to exclude a model to which there is no sensitivity.
  Classic example: fewer selected data events than predicted by SM background. Can sometimes rule out SM b.g. hypothesis at 95% CL and also any signal+background hypothesis, regardless of how small the signal is.

Annoying, but not actually flaws of a technique

- Experiments with less sensitivity (lower s, or higher b, or bigger errors) can set more stringent limits if they are lucky than more sensitive experiments
- Increasing systematic errors on b can result in more stringent limits (happens if an excess is observed in data).

Solution to Annoying Problems -- Expected Limits

- Sensitivity ought to be quoted as the median expected limit (or median discovery probability) or median expected error bar in a large ensemble of possible experiments, not the observed one. Called "a priori limits" in CDF Run 1 parlance.

- Systematic errors will always weaken the expected limits (observed limits may do anything!)

- Best way to compare which analysis is best among several choices -- optimize cuts based on expected limits is optimal

Approximations to expected limit: $s/\sqrt{b}$

$s/\sqrt{s+b}$

Systematic Uncertainties in Fequentist Approaches

- Can construct multi-dimensional Confidence intervals, with each nuisance parameter (=source of uncertainty) constrained by some measurement.

- Not all nuisance paramters can be constrained this way -- some are theoretical guesses with belief distributions instead of pure statistical experimental errors.

- Systematic uncertainty is uncertainty in the predictions of our model: e.g., p(data|Standard Model) is not completely well determined due to nuisance parameters

- One approach -- “ensemble of ensembles” -- include in the ensemble variations of the nuisance parameters.

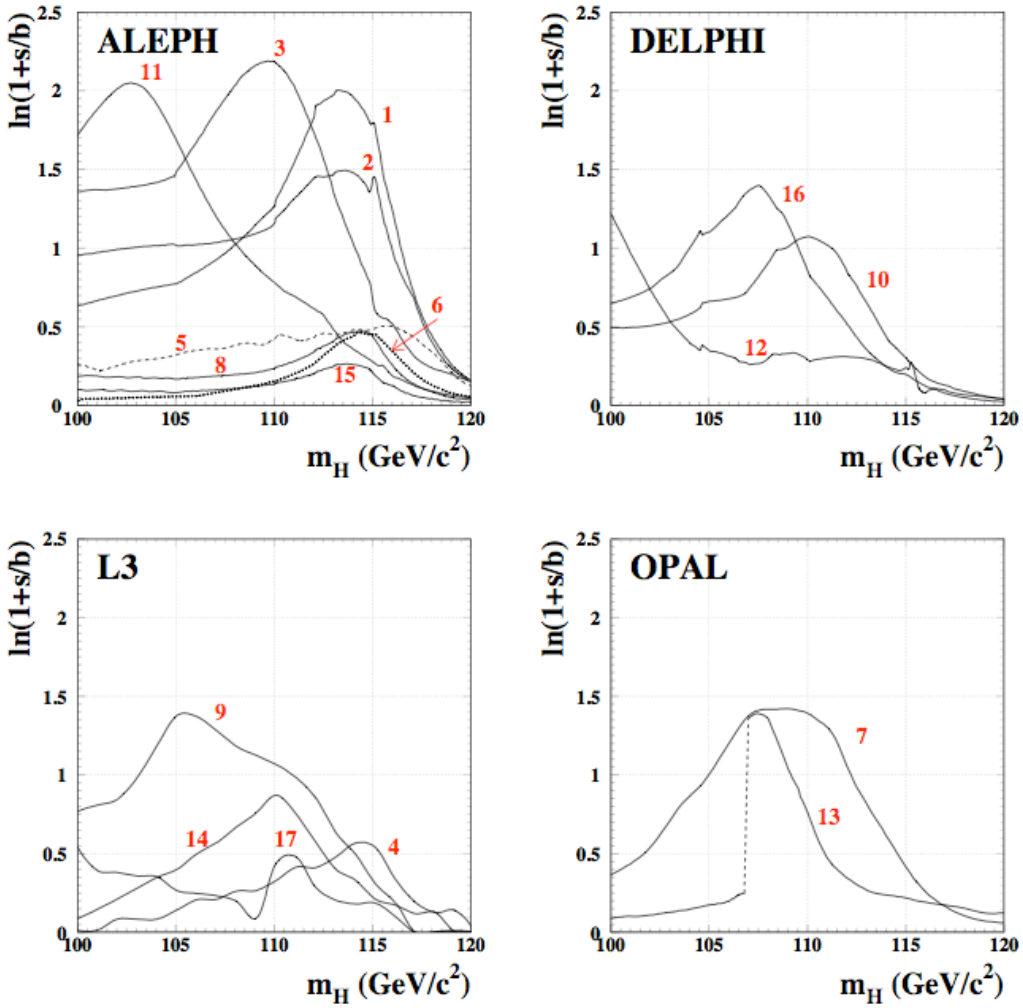
(even Frequentists have to be a little Bayesian sometimes)

Individual Candidates Can Make a Big Difference

At LEP -- can follow individual candidates' interpretations as functions of test mass

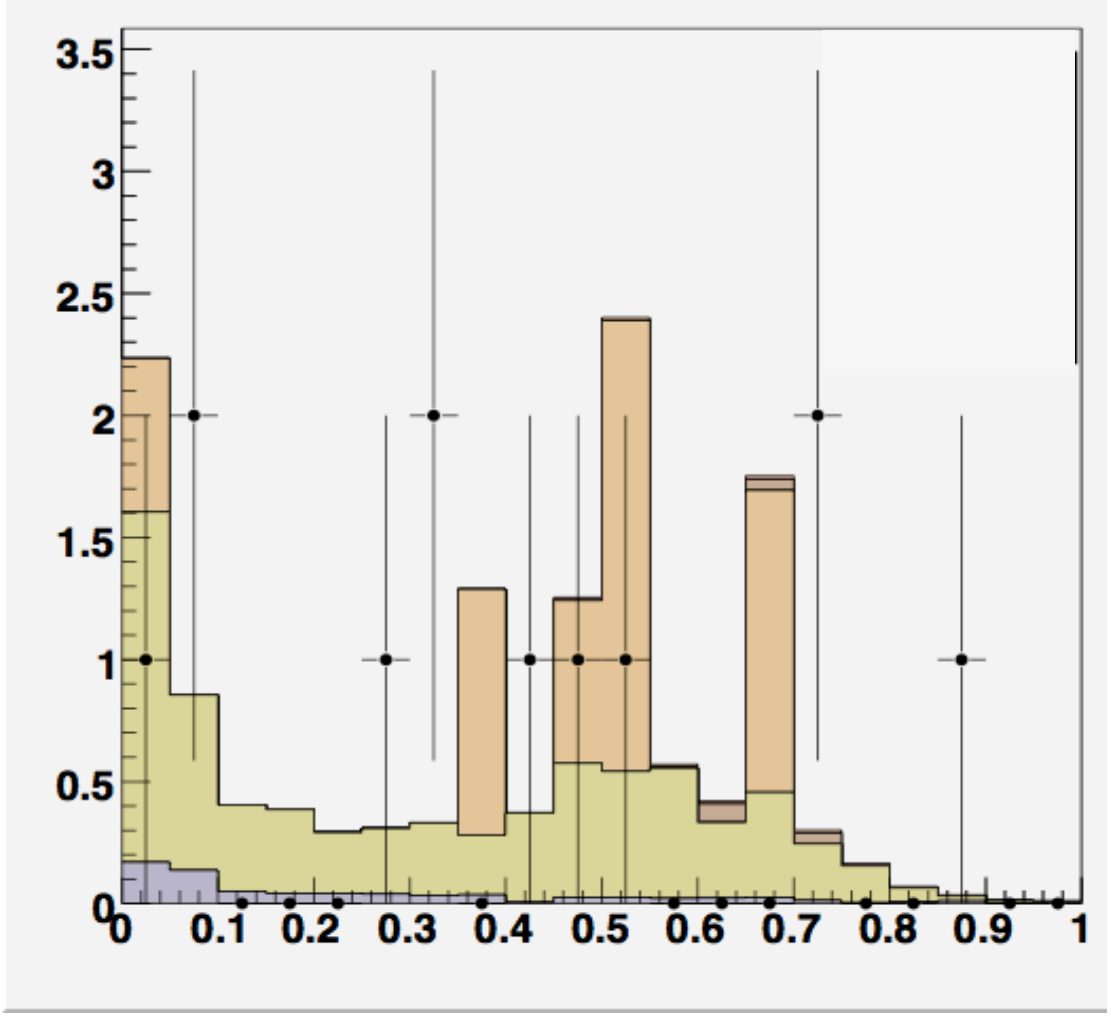
if s/b is high enough near each one.

Fine mass grid -- smooth interpolation of predictions -- some analysis switchovers at different m_H for optimization purposes



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

An Extreme Example (names removed)



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?