

Statistical Methods for Experimental Particle Physics

Tom Junk



Pauli Lectures on Physics

ETH Zürich

30 January — 3 February 2012

Day 3:

Bayesian Inference

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”

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

$\{x\}$ = set of observations

$\{v\}$ = 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: $p(\{v\} | data)$

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

Bayes' Theorem

$p(\{\nu\} | data)$ 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

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

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

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

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

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

Typically $\pi(r)$ is constant
Other options possible.

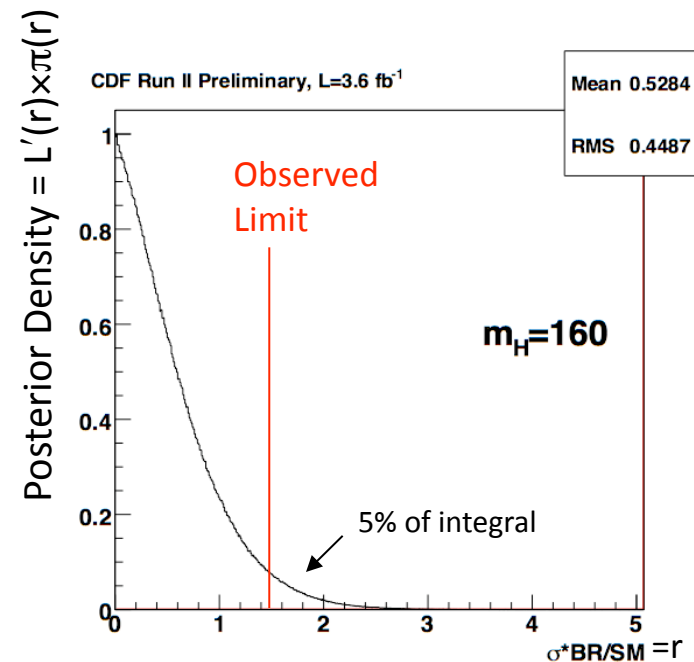
Sensitivity to priors a concern.

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:

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



Bayesian Cross Section Extraction

Same handling of nuisance parameters as for limits

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

$$0.68 = \frac{\int_{r_{low}}^{r_{high}} L'(data | r) \pi(r) dr}{\int_0^{\infty} L'(data | r) \pi(r) dr}$$

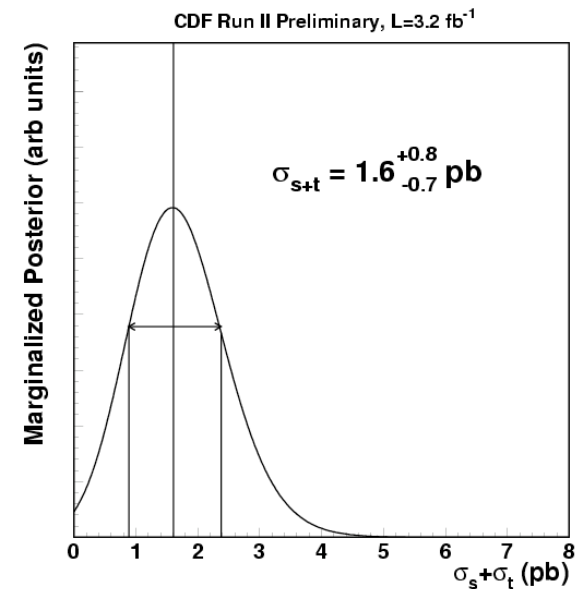
The measured cross section and its uncertainty

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

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

If the 68% CL interval does not contain zero, then the posterior at the top and bottom are equal in magnitude.

The interval can also break up into smaller pieces! (example: WW TGC@LEP2)



Extending Our Useful Tip About Limits

It takes almost exactly 3 expected signal events to exclude a model.

If you have zero events observed, zero expected background, and no systematic uncertainties, then the limit will be 3 signal events.

Call s =expected signal, b =expected background. $r=s+b$ is the total prediction.

$$L(n = 0, r) = \frac{r^0 e^{-r}}{0!} = e^{-r} = e^{-(s+b)}$$

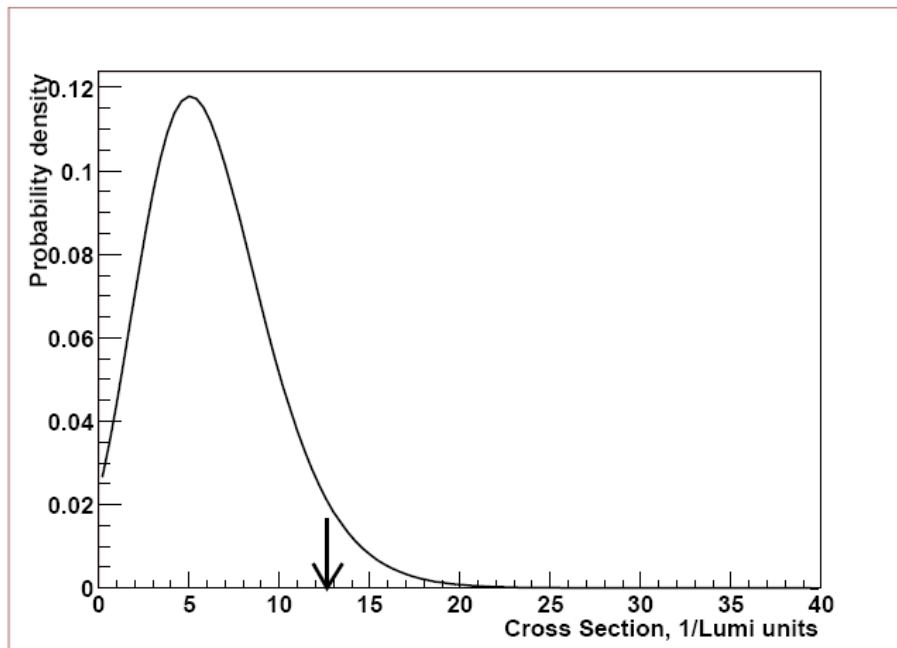
$$0.95 = \frac{\int_0^{r_{\text{lim}}} L'(data | r) \pi(r) dr}{\int_0^{\infty} L'(data | r) \pi(r) dr} = \frac{-e^{-(s+b)} \Big|_0^{r_{\text{lim}}}}{-e^{-(s+b)} \Big|_0^{\infty}} = e^{-r_{\text{lim}}}$$

The background rate cancels! For 0 observed events, the signal limit does not depend on the predicted background (or its uncertainty). This is also true for CL_s limits, but not PCL limits (which get stronger with more background)

If $p=0.05$, then $r=-\ln(0.05)=2.99573$

A Handy Limit Calculator

D0 (http://www-d0.fnal.gov/Run2Physics/limit_calc/limit_calc.html) has a web-based, menu-driven Bayesian limit calculator for a single counting experiment, with uncorrelated uncertainties on the acceptance, background, and luminosity. Assumes a uniform prior on the signal strength. Computes 95% CL limits (“Credibility Level”)



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

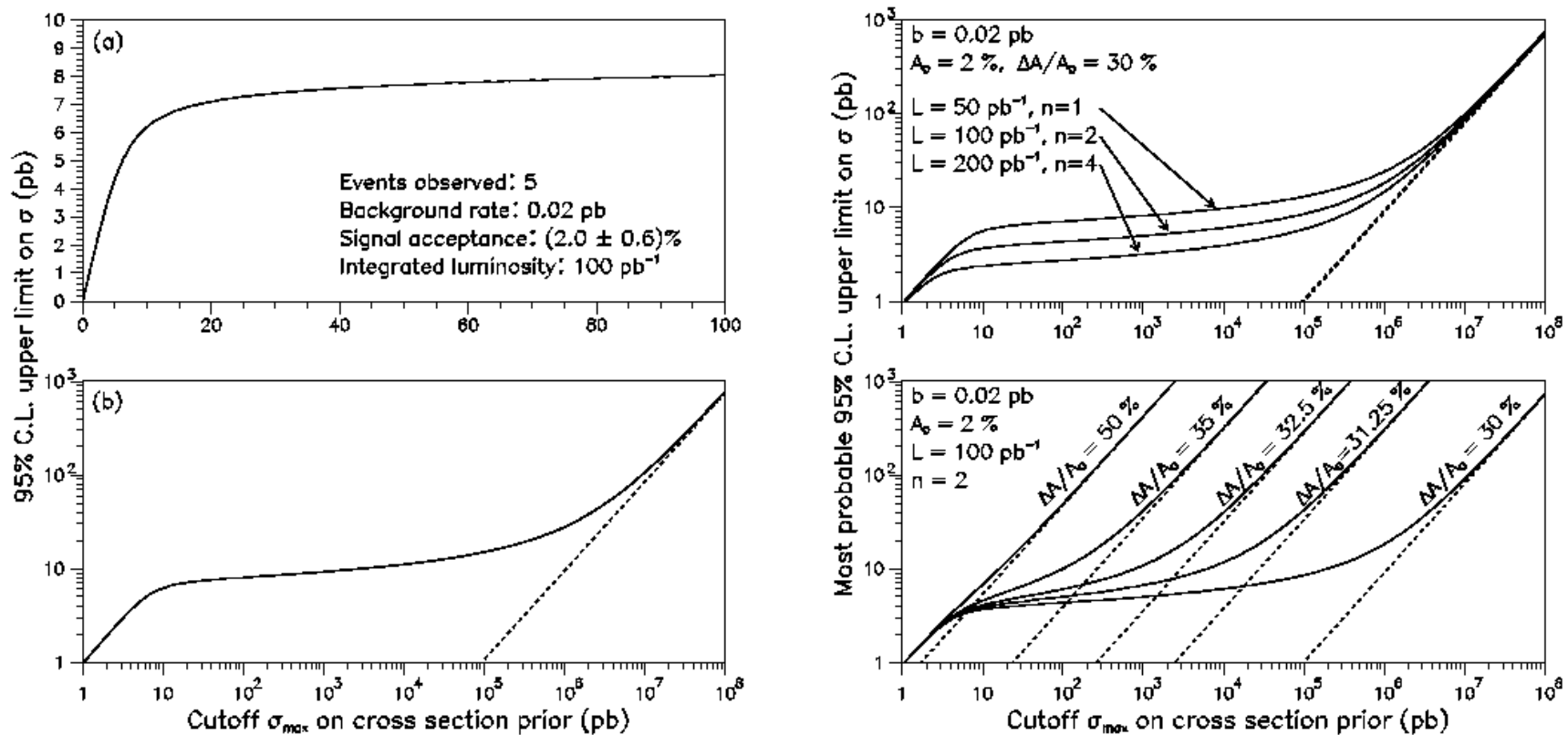


Figure 1: Bayesian upper limits at the 95% credibility level on a hypothetical cross section σ , as a function of the cutoff σ_{\max} on the flat prior for σ .

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? (answer: 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.

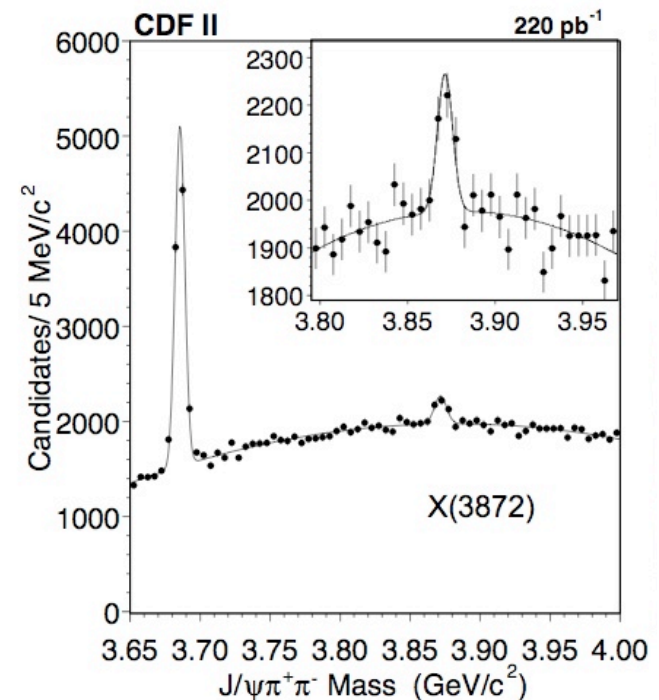
Integrating over Systematic Uncertainties Helps Constrain their Values with Data

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

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.

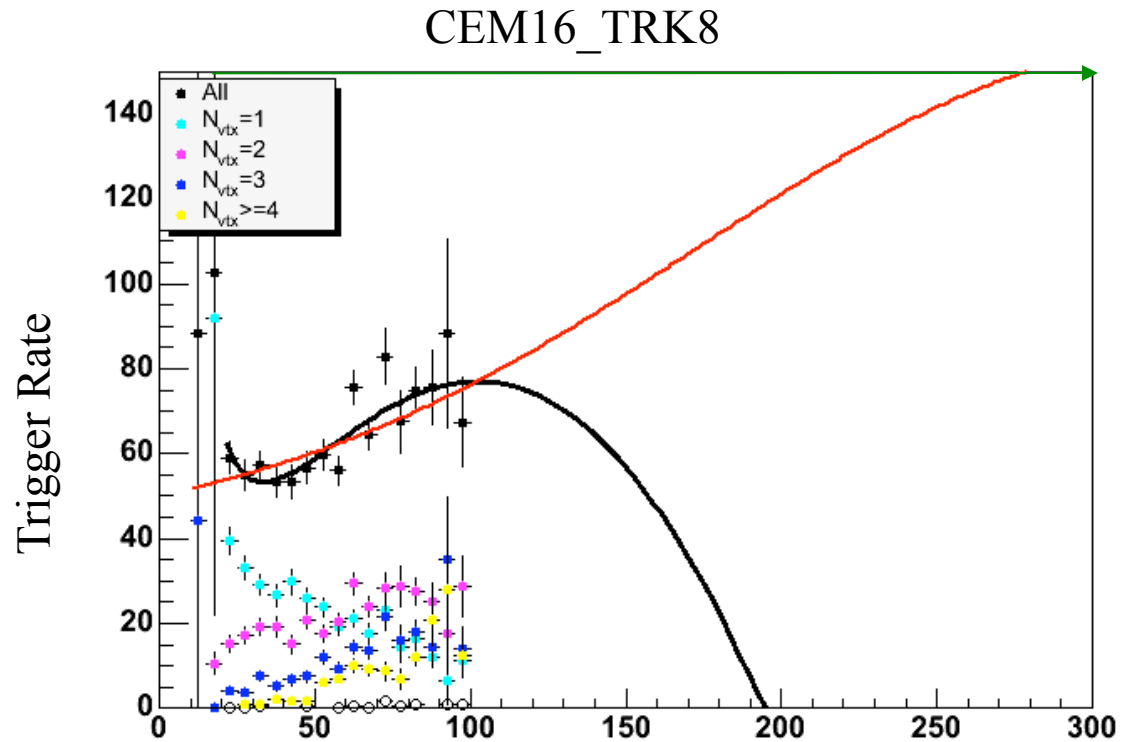
Coping with Systematic Uncertainty

- “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

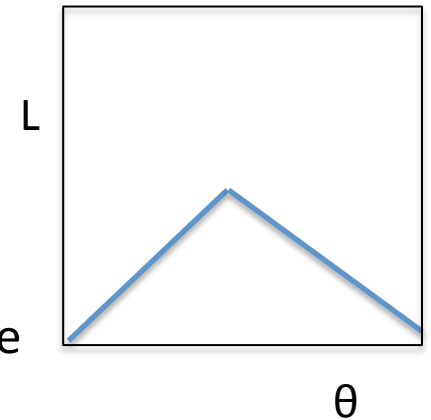


Other Pitfalls of Fitting

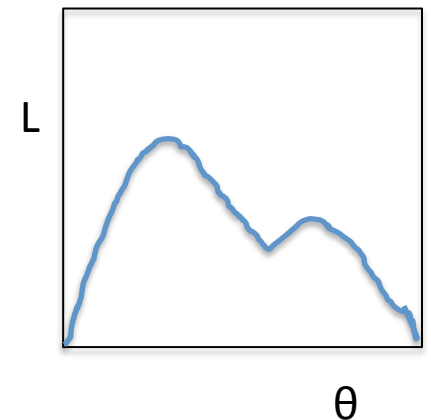
Usually methods relying on profiling and marginalizing provide numerically similar results, but there are exceptions.

Sometimes a likelihood function has multiple maxima.

Prediction = 10^{+2}_{+3} . Observe data=12. What's the best fit nuisance parameter? Its Uncertainty? Integrating over the whole shape provides the most information.



Sometimes a likelihood function has a discontinuous first derivative (care should be taken to avoid this, but sometimes it happens. e.g. using Barlow and Beeston's TFractionFitter in a likelihood function).



MINUIT gets stuck in corners. Uncertainty in fit value is also ill-defined.

Asymmetric Uncertainties and Priors

Measurements, and even theoretical calculations, frequently are assigned asymmetric uncertainties:

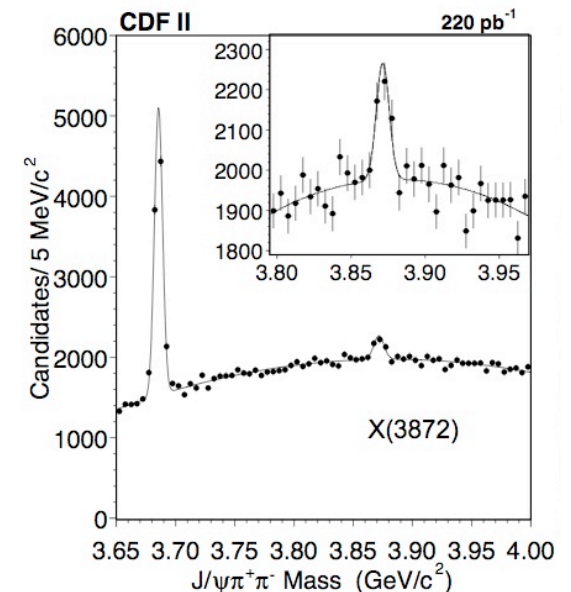
Value = 10^{+2}_{-1} , or more extremely, 10^{+2}_{+2} (ouch). When the uncertainties have the same sign on both sides, it is worthwhile to check and see why this is the case.

Example – we seek a bump in a mass distribution by counting events in a small window around where the bump is sought.

The detector calibration has an energy uncertainty (magnetic field or chamber alignment for tracks, or much larger effect, calorimeter energy scales for jets).

Shift the calibration scale up – predicted peak shifts out of the window → downward shift in expected signal prediction.

Shift the calibration down – predicted peak shifts out of the other side of the window → downward shift in expected signal prediction



Treatment of Asymmetric Uncertainties

These cases are pretty clear – the underlying parameter, the energy scale, has a (Gaussian? Your choice) distribution, while it has a nonlinear, possibly non-monotonic *impact* on the model prediction.

The same parameter may have a linear, symmetrical impact on another model prediction, and we will have to treat them as correlated in statistical analysis tools.

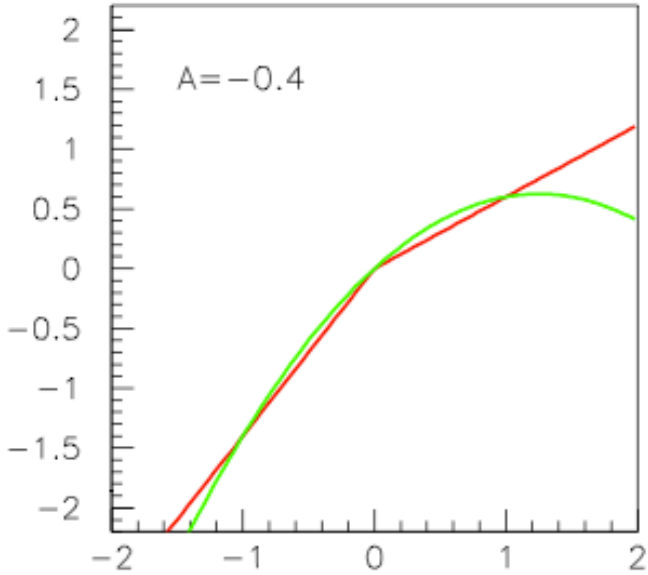
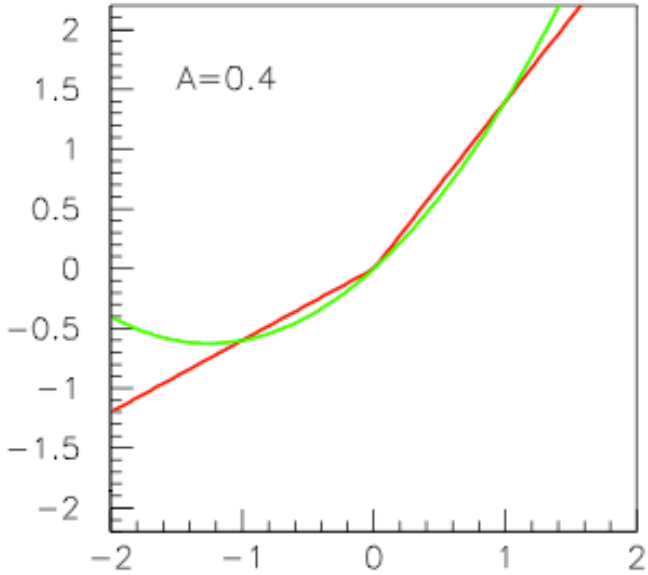
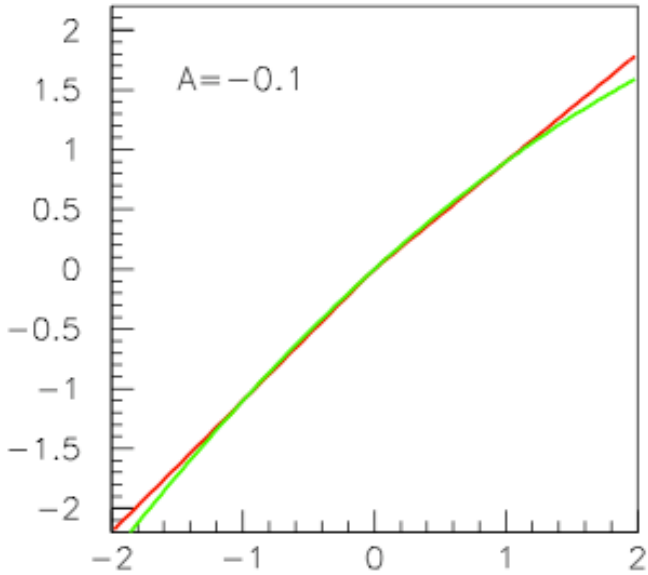
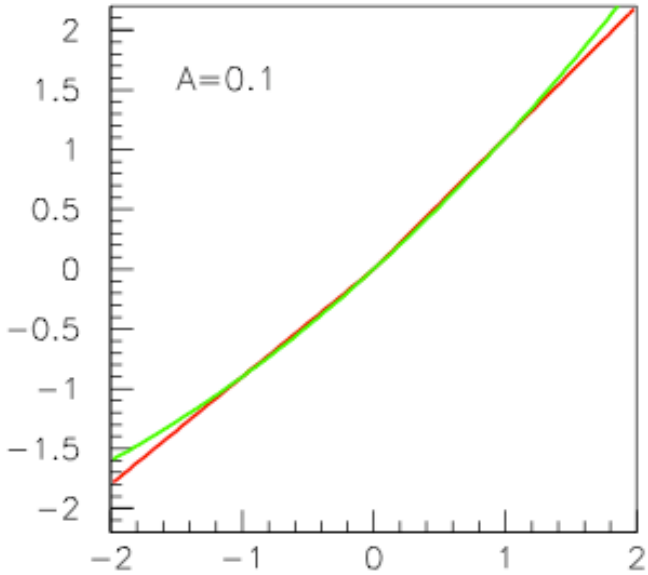
Treatment is ambiguous when little is known why the uncertainties are asymmetric, or it is not clear how to extrapolate/interpolate them.

See R. Barlow,

“Asymmetric Systematic Errors”, arXiv:physics/0306138

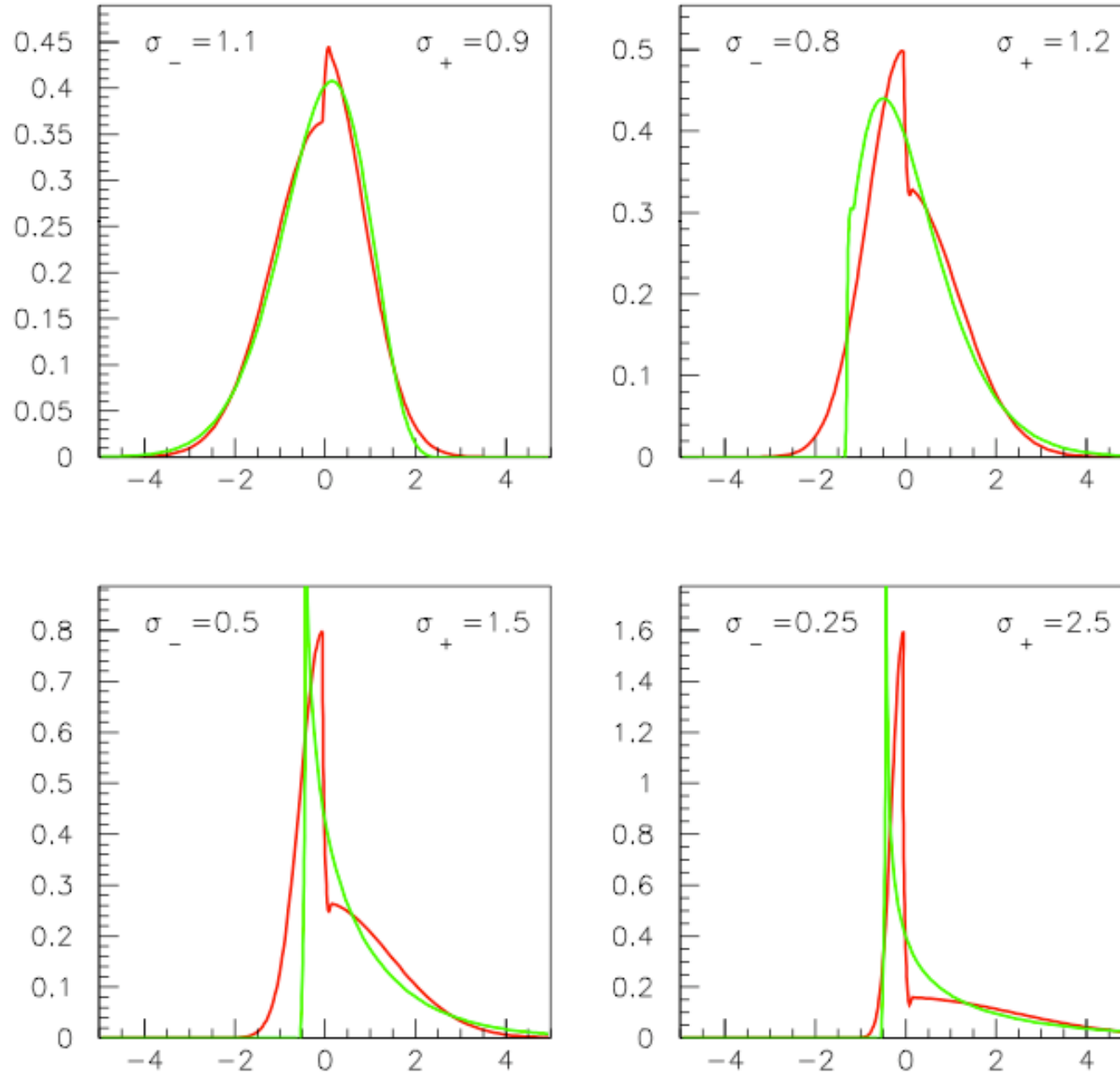
“Asymmetric Statistical Errors”, arXiv:physics/0406120

Quadratic Impacts of Asymmetric Uncertainties



R. Barlow

Resulting Prior Distributions for alternative handling of Asymmetric Impacts



R. Barlow

Other Ideas on Handling Asymmetric Uncertainties

- Quadratic interpolation for small values of the uncertain parameter
 - avoids the kink at zero
- Gradual switchover (use an exponential or other asymptotic function) to linear for large values of the nuisance parameter
 - avoids large quadratic divergence from more sensible linear extrapolation

Arbitrary! But this one's nice. What are our criteria for what's "nice"?

Preserve the mean of the prior distribution to be the central value.
Otherwise people will complain of bias.

Preserve the median of the prior distribution to be the central value.
Otherwise an up-variation in the parameter will produce a down-variation in the impacted prediction.

Preserve the mode of the prior distribution
The best fit value should be the central prediction.

We may be asking too much! What does 1^{+10}_{-1} mean, anyway?

Statistical Uncertainties on Systematic Uncertainties?

Answer: No. But some systematic uncertainties are difficult to evaluate properly.

See Roger Barlow's "Systematic errors: Facts and Fictions",
arXiv: hep-ex/0207026

The idea: If a systematic uncertainty is estimated by comparing two data samples or two MC samples, or data vs. MC, then if one or both of them have a limited size, then the magnitude of the systematic can be poorly constrained.

Ideally, work harder (run more MC) to get a better prediction of the expected signal and background, under all assumptions of systematic variation.

Monte Carlo Statistical Uncertainty is a Systematic Uncertainty

but don't double-count it for each separate MC variation of each nuisance parameter. Easy to do by comparing central vs. varied MC samples.

Statistically weak tests should be handed as cross checks. If they are consistent, consider the test to have passed, but do not add systematic uncertainty.

If they fail, however, and a discrepancy between two MC's or data and MC cannot be understood and fixed, then a systematic uncertainty is called for.

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”

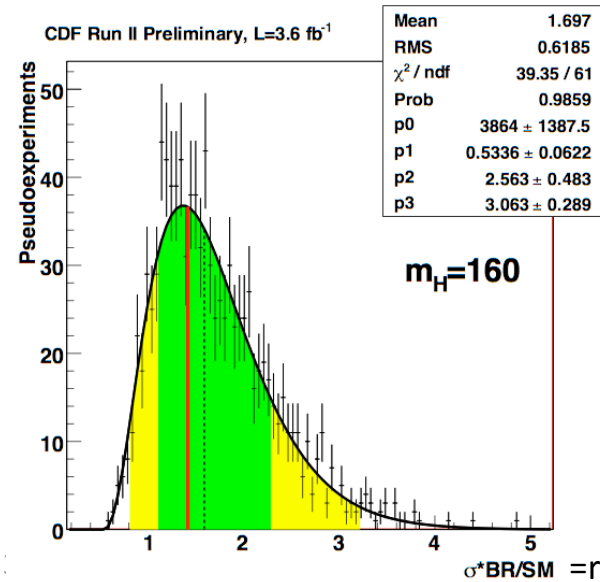
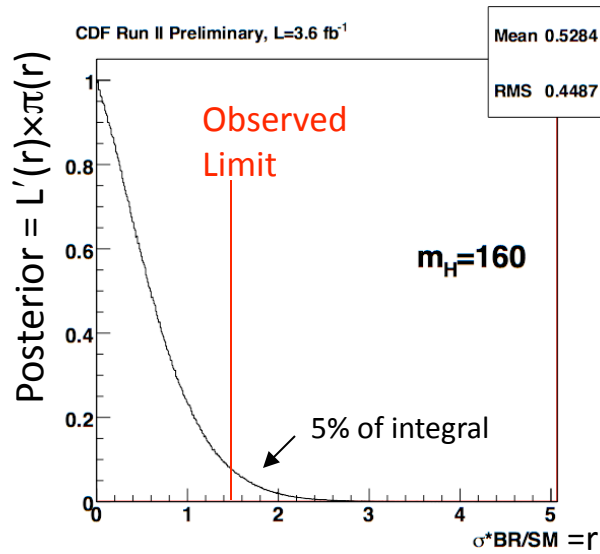
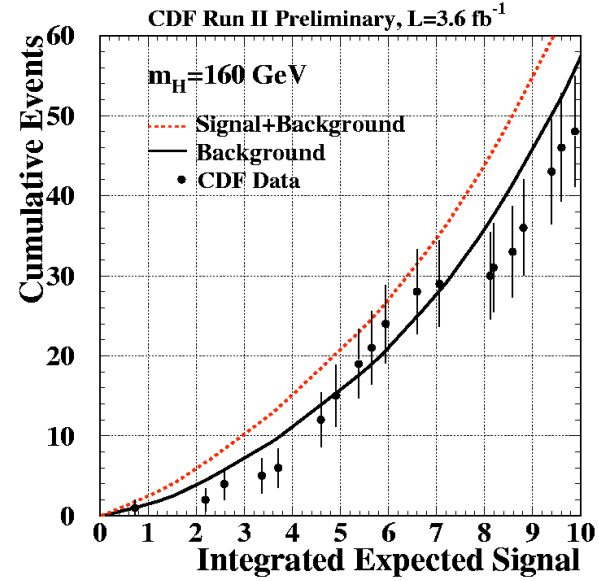
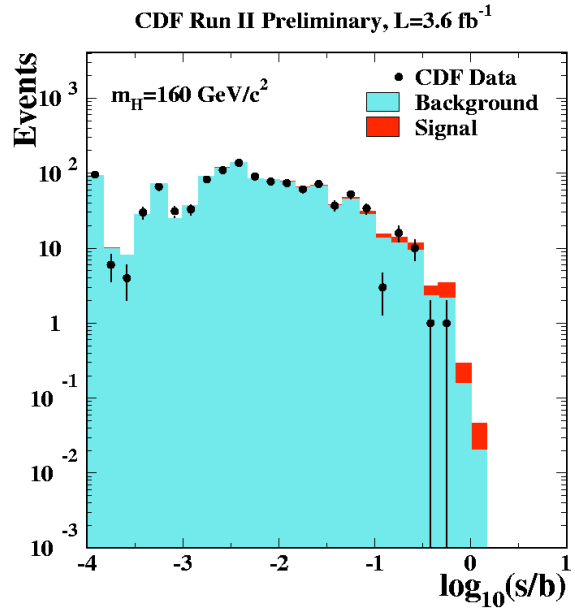
- But we still want the sensitivity estimated! An experiment can get a strong upper limit not because it was well designed, but because it was lucky.

How to optimize an analysis before data are observed?

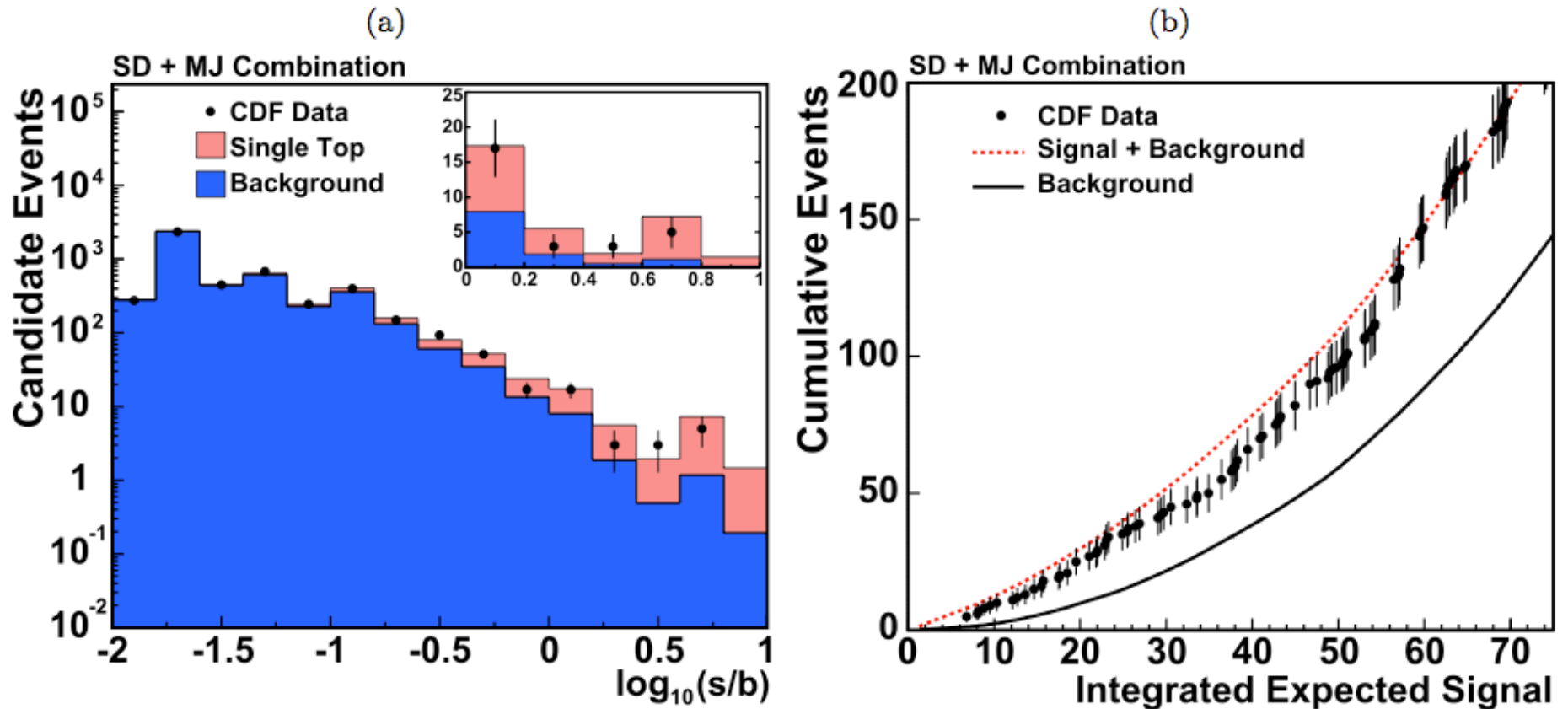
So -- run Monte Carlo simulated experiments and compute a Frequentist distribution of possible limits. Take the **median**-- metric independent and less pulled by tails.

But even Bayesian/Frequentists have to be Bayesian: use the Prior-Predictive method -- vary the systematics on each 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 older one)



What These Look Like for a 5.0σ Observation



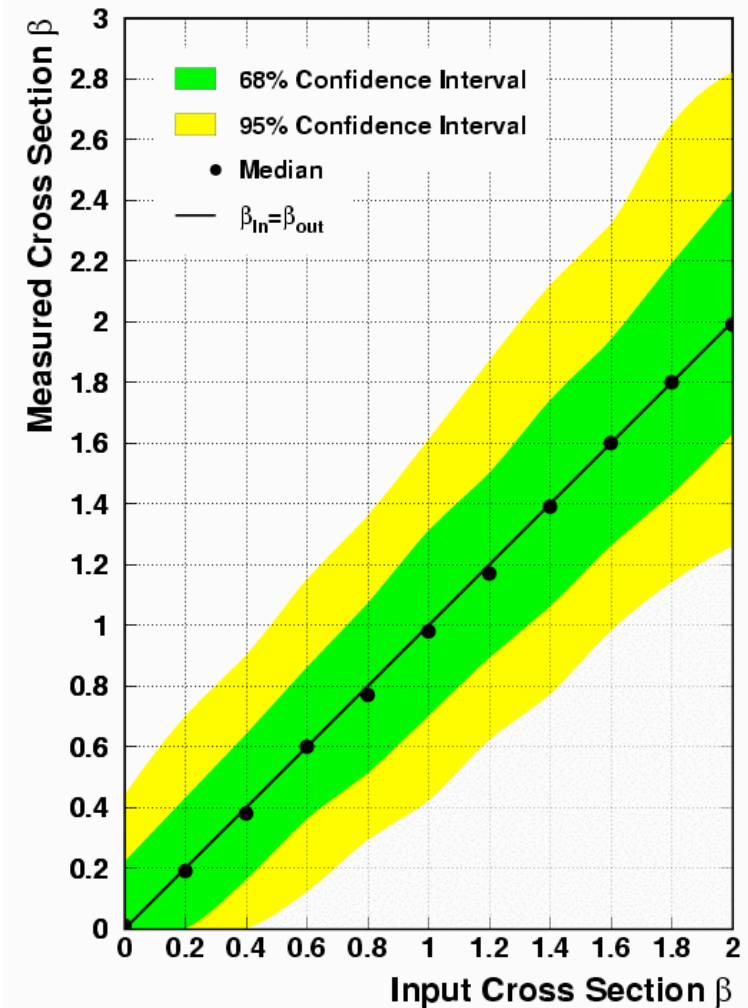
CDF Single Top, 3.2 fb^{-1}

Even Bayesians have to be a little Frequentist

We would like to know how the cross section calculations behave in an ensemble of possible experimental outcomes.

Procedure:

- Inject a signal.
- Vary systematics on each pseudoexperiment (which integrates over them in the ensemble)
- Calculate Bayesian cross section for each outcome and plot distribution.
- Black line is the median, not the mean
- Check the width of the distribution against the quoted uncertainties. Specifically, the distribution of (meas-inject)/uncertainty should be a Unit-width Gaussian (when not up against zero).



This is in fact a Neyman construction!
Can do Feldman-Cousins with this
(correct for fit biases, if any).

Some Features of the Linearity Plot

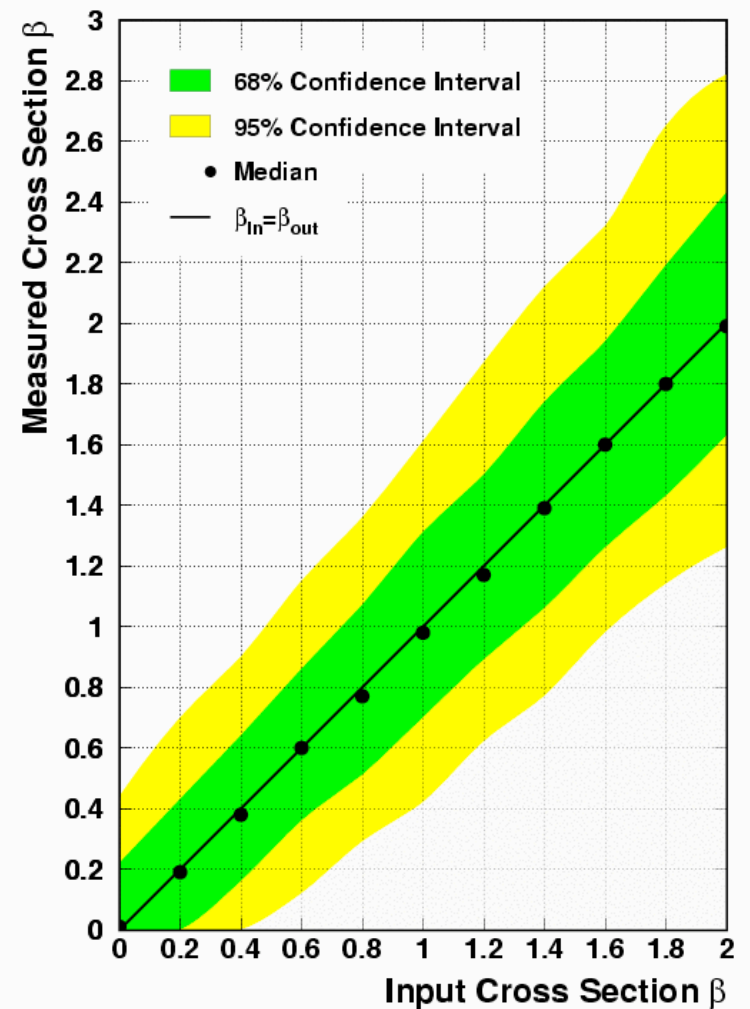
The distribution of fit outcomes at an injected signal of 0 is a delta function at zero with 50% of the total amount. The other 50% of the distribution has a width from the measurement resolution.

When computing pulls, use the up error if the measured value is below the injected rate, and the down error if it is above.

For a fully systematics-dominated measurement, the band edges should be straight lines pointing at the origin. (e.g, if the only uncertainty were acceptance). Also largely the case for high s/b statistics-limited measurements.

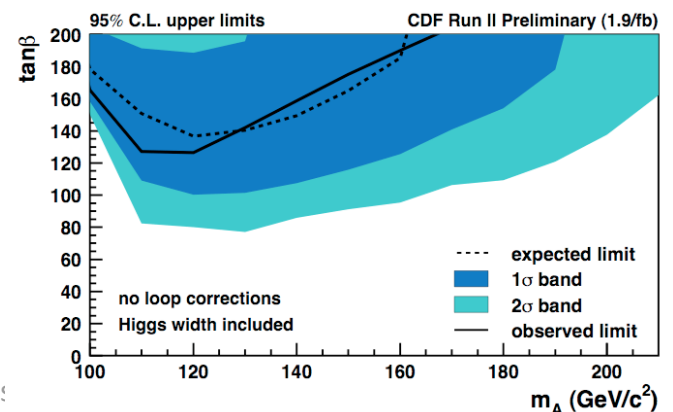
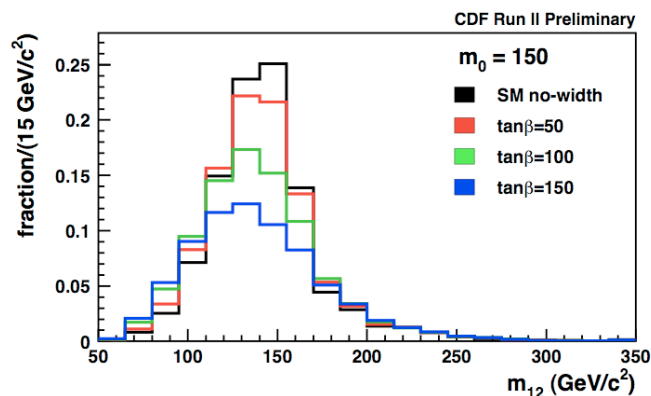
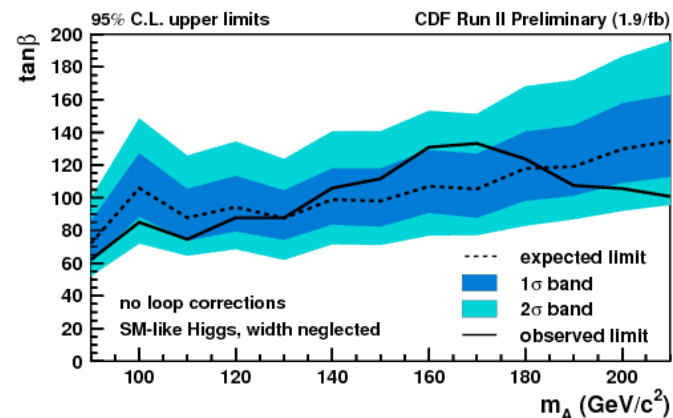
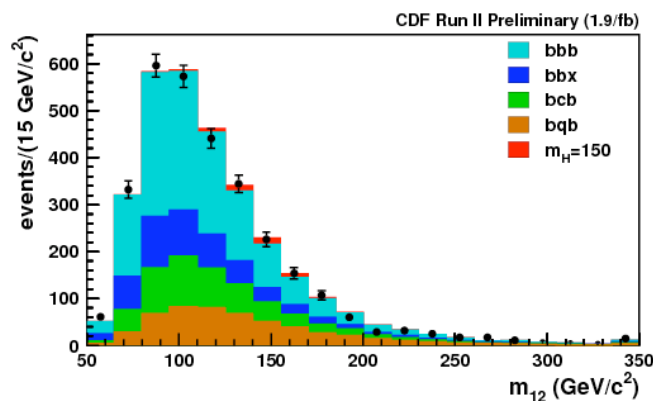
For this measurement, there was a small signal and a large, uncertain background. The total uncertainty on the signal is less dependent on the value of the signal.

Using the fit value of the uncertainty can be biasing – also quote expected fit uncertainty



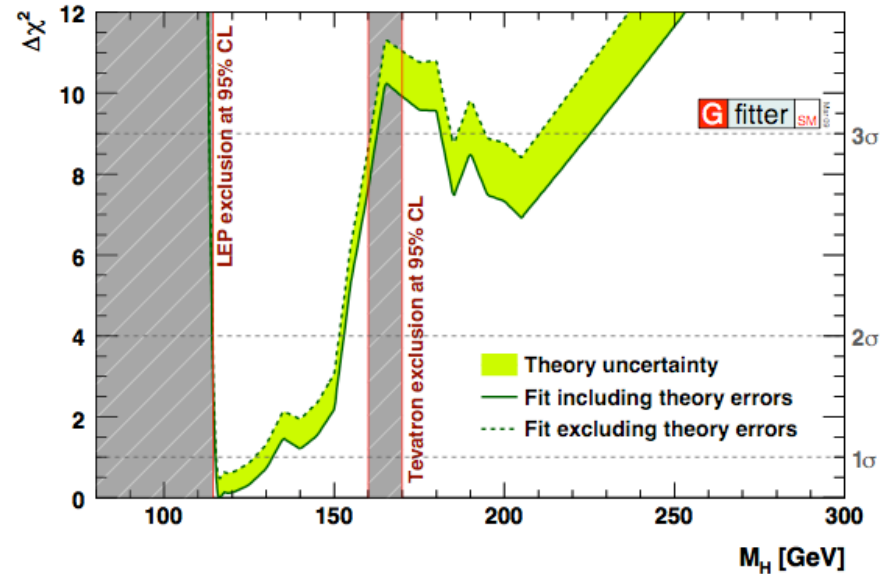
An Example Where Usual Bayesian Software Doesn't Work

- 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

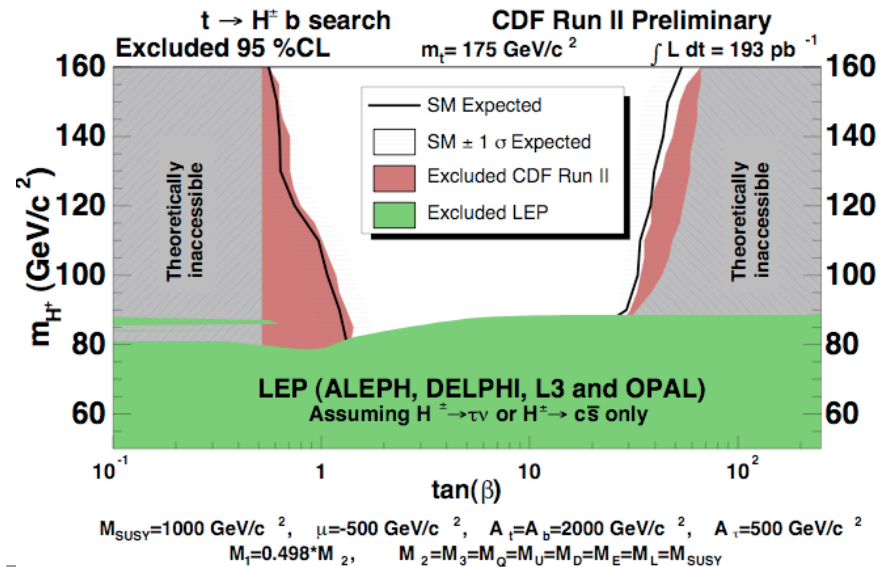


Priors in Non-Cross-Section Parameters

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 $\log(\tan\beta)$ -- even with no
 sensitivity, can set non-trivial
 limits..



Bayesian Discovery?

Bayes Factor

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

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.

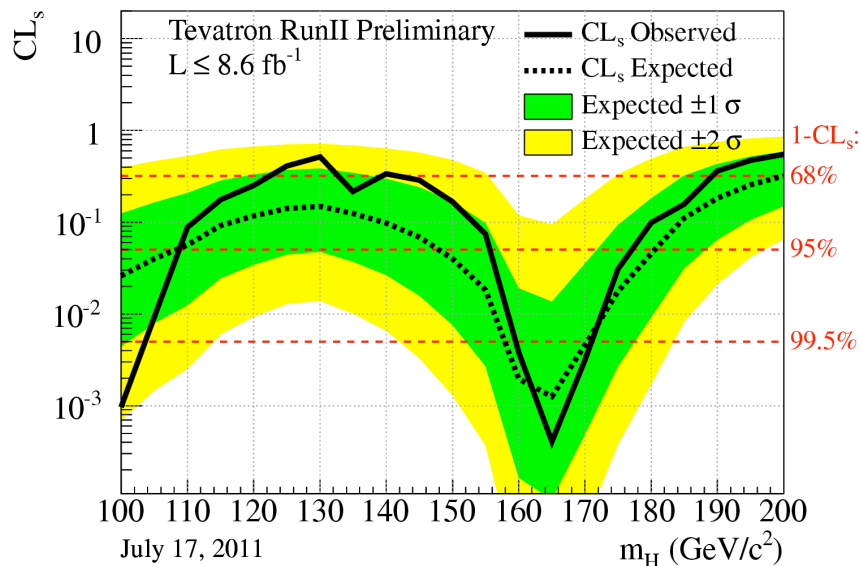
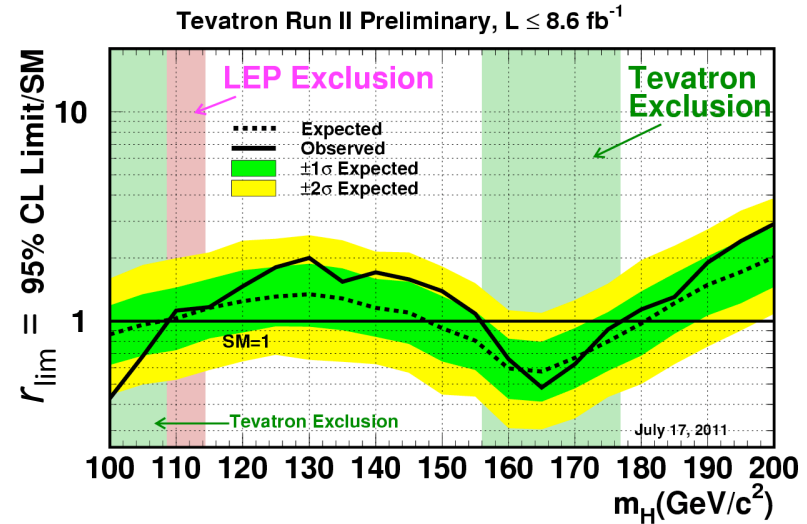
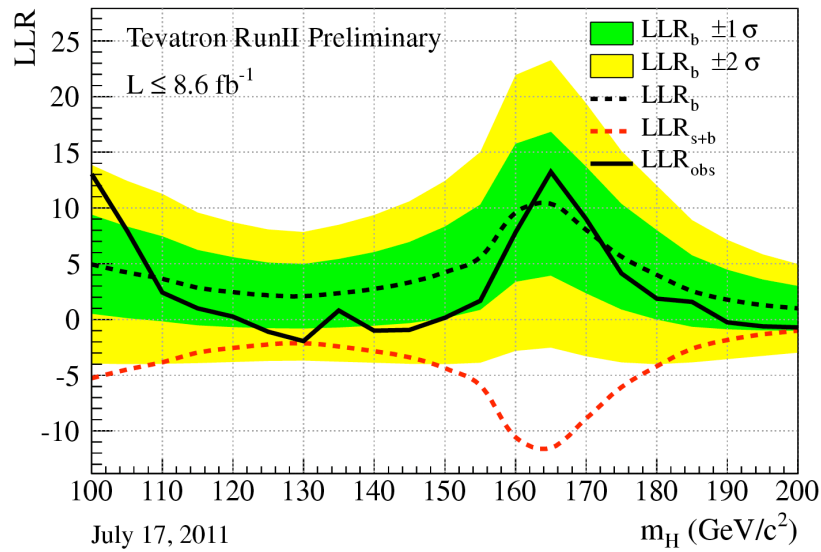
Physicists would like to check the false discovery rate, and then we're back to p-values.

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 $b=2 \times 10^{-7}$ events, expected signal=1 event, observe 1 event, no systematics.

p-value $\sim 2 \times 10^{-7}$ is a discovery! (hard to explain that event with just the background model). But have $\pm 100\%$ 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.

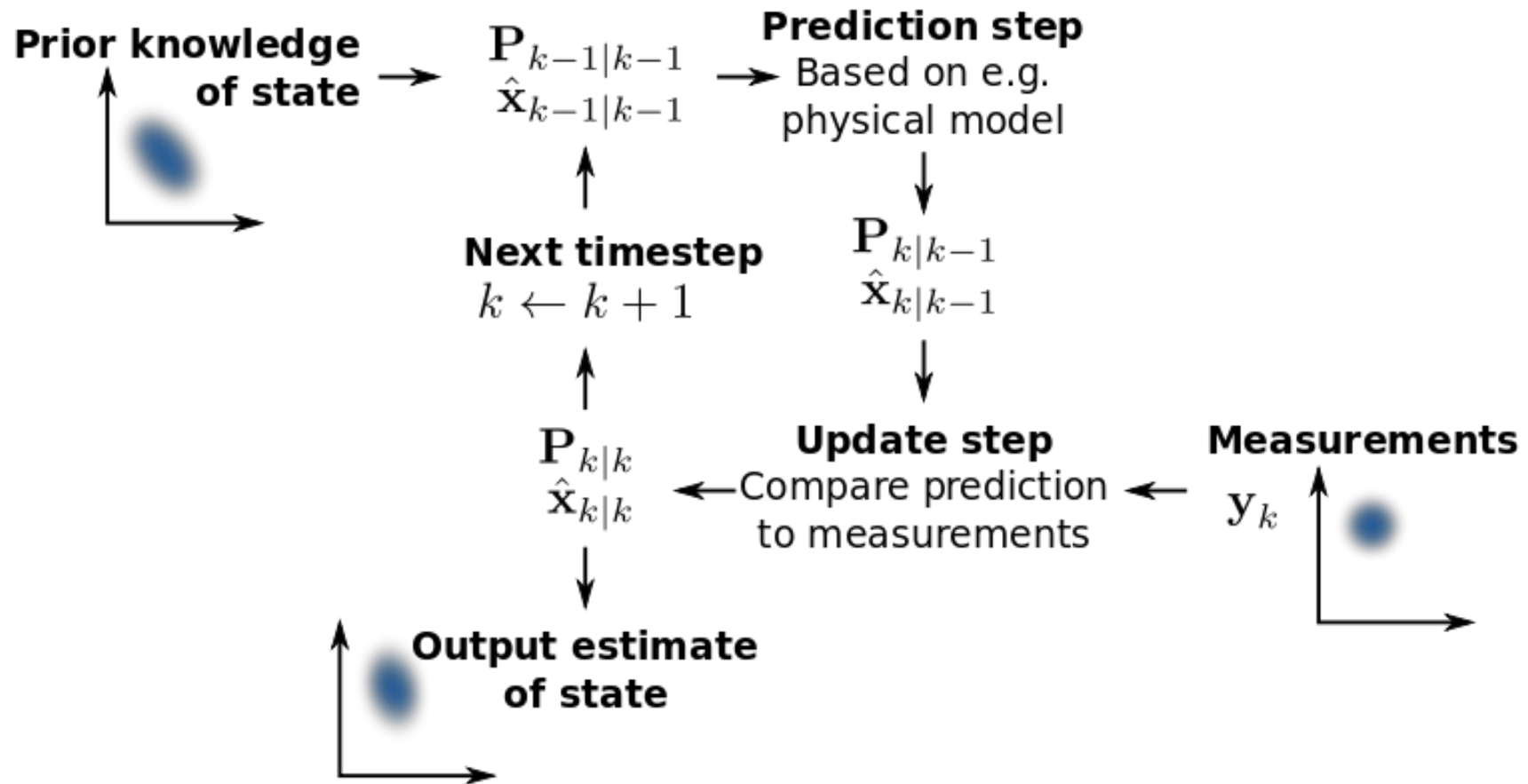
In all practicality, discriminant distributions have a wide spectrum of s/b , even in the same histogram. But some good bins with $b < 1$ event

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)

Another Application of Bayesian Reasoning: The Kalman Filter

Used in HEP to fit tracks in a particle detector

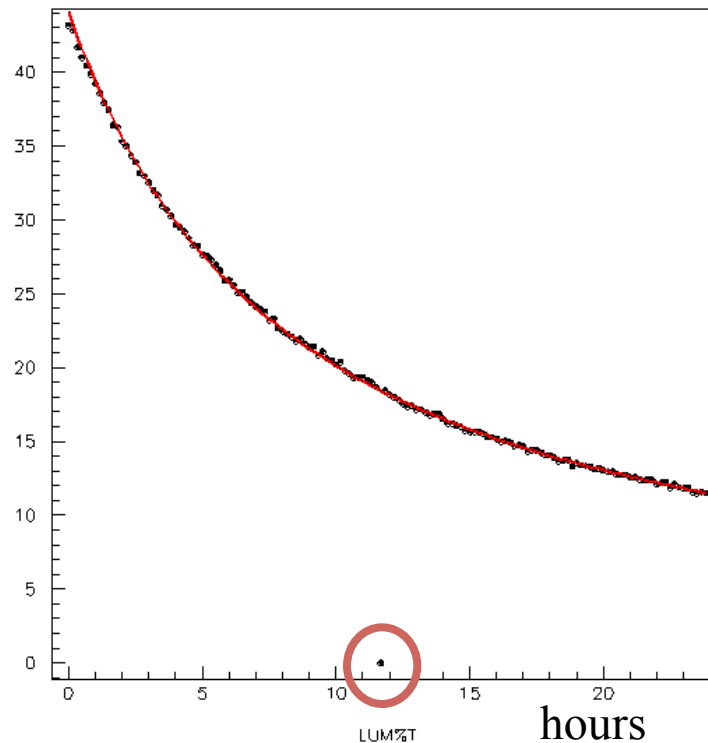


From the Wikipedia article

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 at CDF vs. time
(a Tevatron 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.

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 $\pi(s) = 1$ and find the limit by integrating:

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

Not too tricky; easy to explain.

- But where did $\pi(s)$ 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 hypothesis 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:

$$s/\sqrt{s+b}$$

Approximation to expected discovery significance

$$s/\sqrt{b}$$

Systematic Uncertainties in Frequentist Approaches

- Can construct multi-dimensional Confidence intervals, with each nuisance parameter (=source of uncertainty) constrained by some measurement.
- Not all nuisance parameters 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., $p(\text{data}|\text{Standard Model})$ is not completely well determined due to nuisance parameters
- One approach -- “ensemble of ensembles” -- include in the ensemble variations of the nuisance parameters.

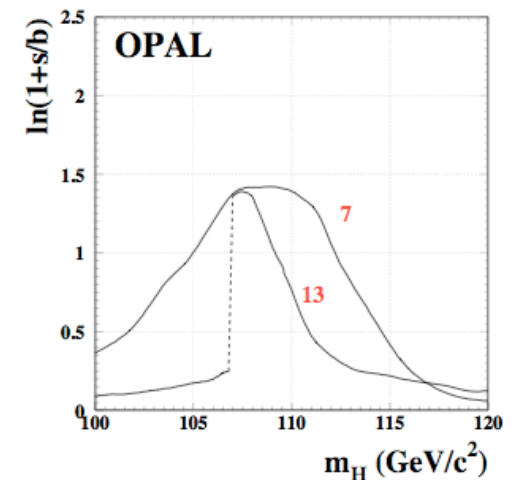
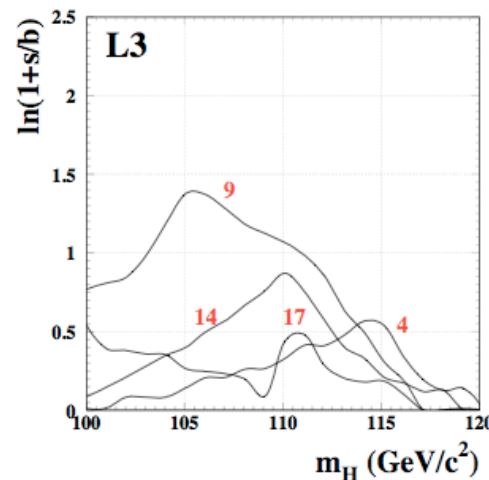
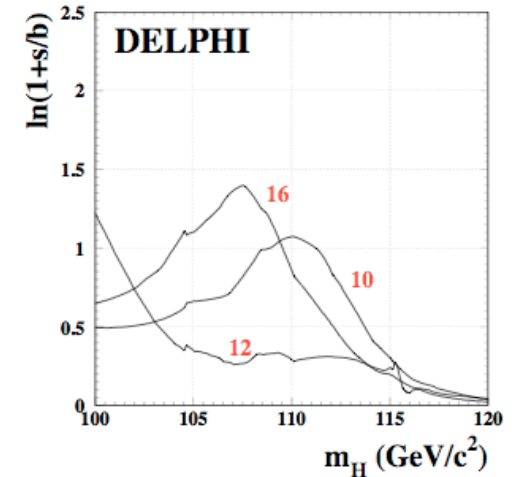
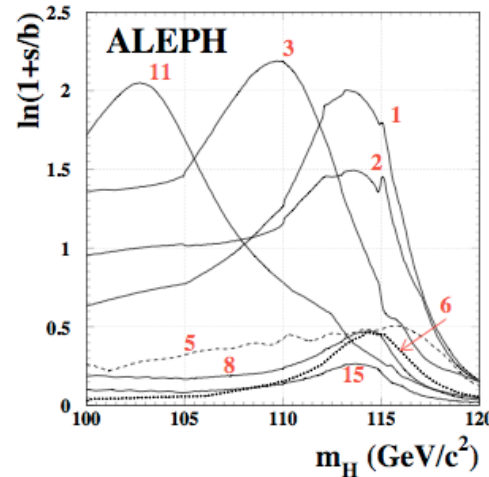
(even Frequentists have to be a little Bayesian sometimes)

Individual Candidates Can Make a Big Difference

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

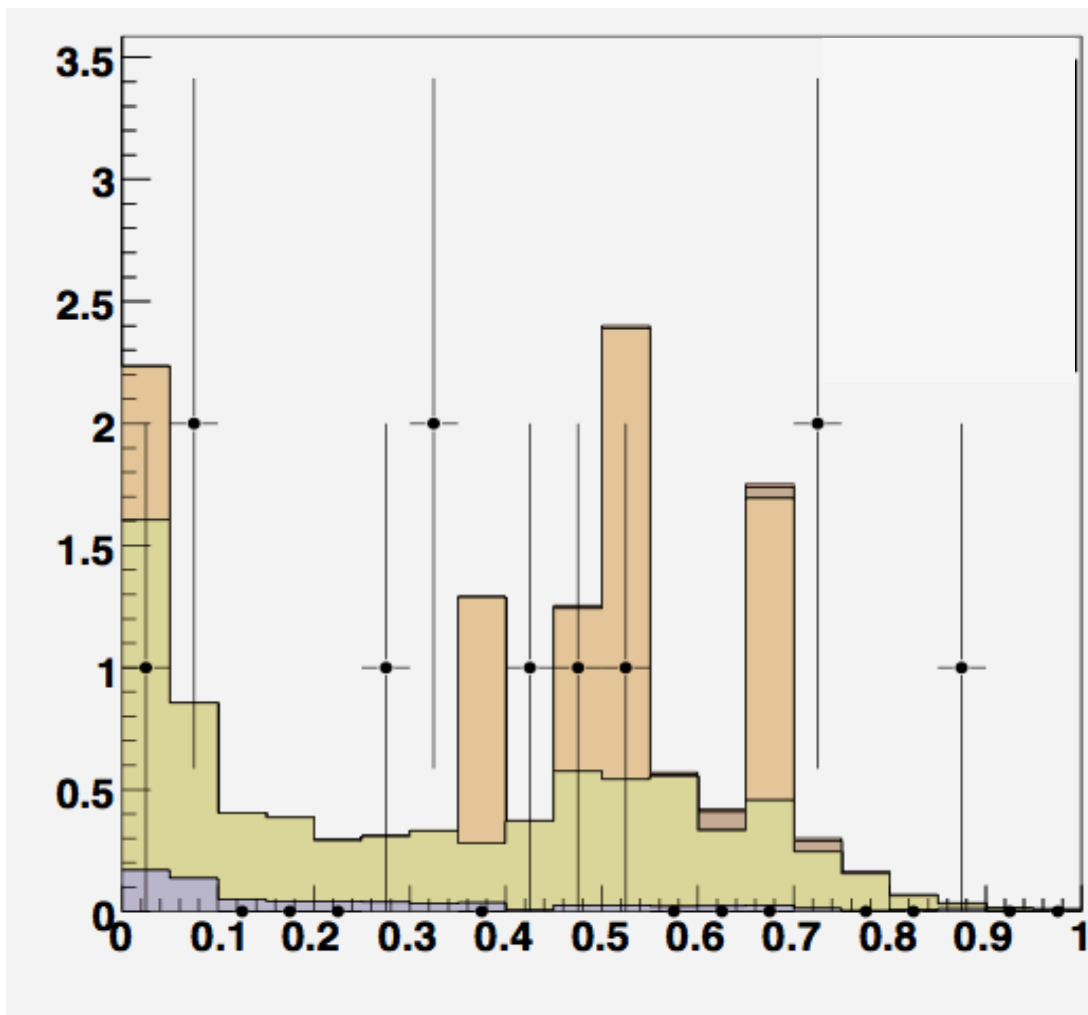
if s/b is high enough near each one.

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



A Pitfall -- Not Enough MC (or data in sideband regions) 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 $n_{\text{obs}} + 1$ in control sample. Fine binning \rightarrow bias in background prediction.

Overcovers for discovery, undercovers for limits?