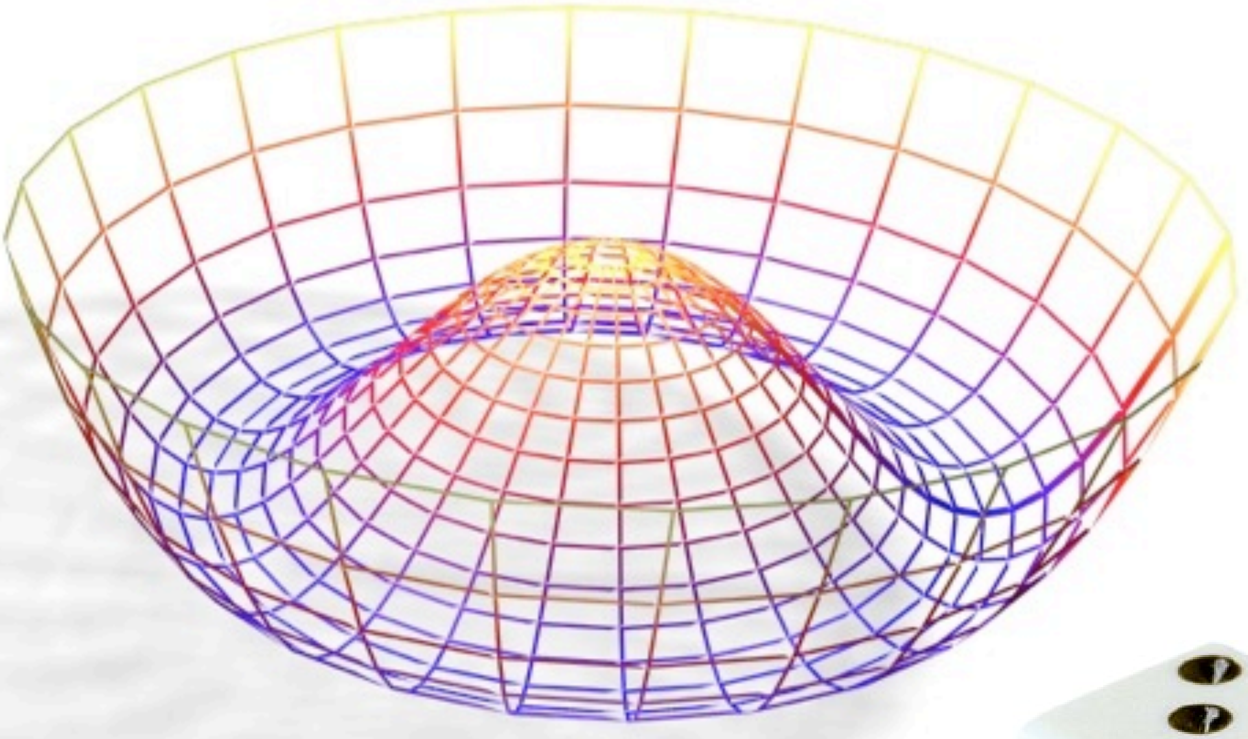
Center for
Cosmology and
Particle Physics

Statistics for the LHC: Quantifying our Scientific Narrative



Kyle Cranmer,
New York University

Introduction

Statistics plays a vital role in science, it is the way that we:

- quantify our knowledge and uncertainty
- communicate results of experiments

Big questions:

- make discoveries, test theories, measure or exclude parameters, etc.
- how do we get the most out of our data
- how do we incorporate uncertainties
- how do we make decisions

Statistics is a very big field, and it is not possible to cover everything in 4 hours. In these talks I will try to:

- **explain** some fundamental ideas & prove a few things
- **enrich** what you already know
- **expose** you to some new ideas

I will try to go slowly, because if you are not following the logic, then it is not very interesting.

- Please feel free to ask questions and interrupt at any time

Further Reading

By physicists, for physicists

G. Cowan, Statistical Data Analysis, Clarendon Press, Oxford, 1998.

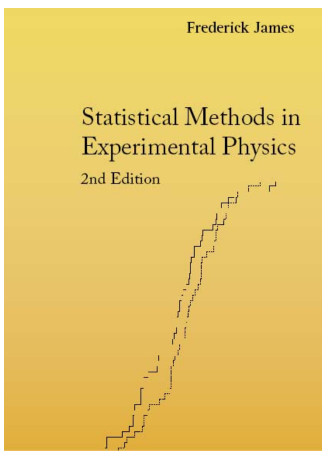
R.J.Barlow, A Guide to the Use of Statistical Methods in the Physical Sciences, John Wiley, 1989;

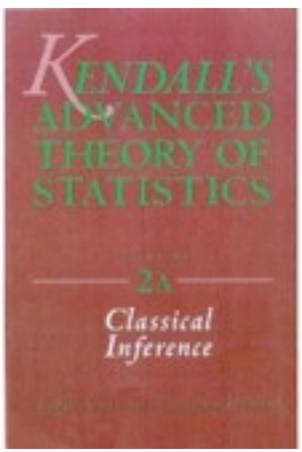
F. James, Statistical Methods in Experimental Physics, 2nd ed., World Scientific, 2006;

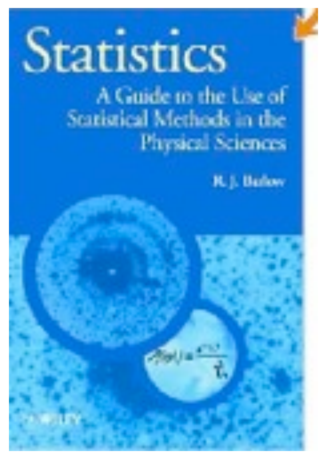
- W.T. Eadie et al., North-Holland, 1971 (1st ed., hard to find);

S.Brandt, Statistical and Computational Methods in Data Analysis, Springer, New York, 1998.

L.Lyons, Statistics for Nuclear and Particle Physics, CUP, 1986.







My favorite statistics book by a statistician:

Stuart, Ord, Arnold. "Kendall's Advanced Theory of Statistics" Vol. 2A *Classical Inference & the Linear Model.*

Other lectures

Fred James's lectures

http://preprints.cern.ch/cgi-bin/setlink?base=AT&categ=Academic_Training&id=AT00000799

http://www.desy.de/~acatrain/

Glen Cowan's lectures

http://www.pp.rhul.ac.uk/~cowan/stat_cern.html

Louis Lyons

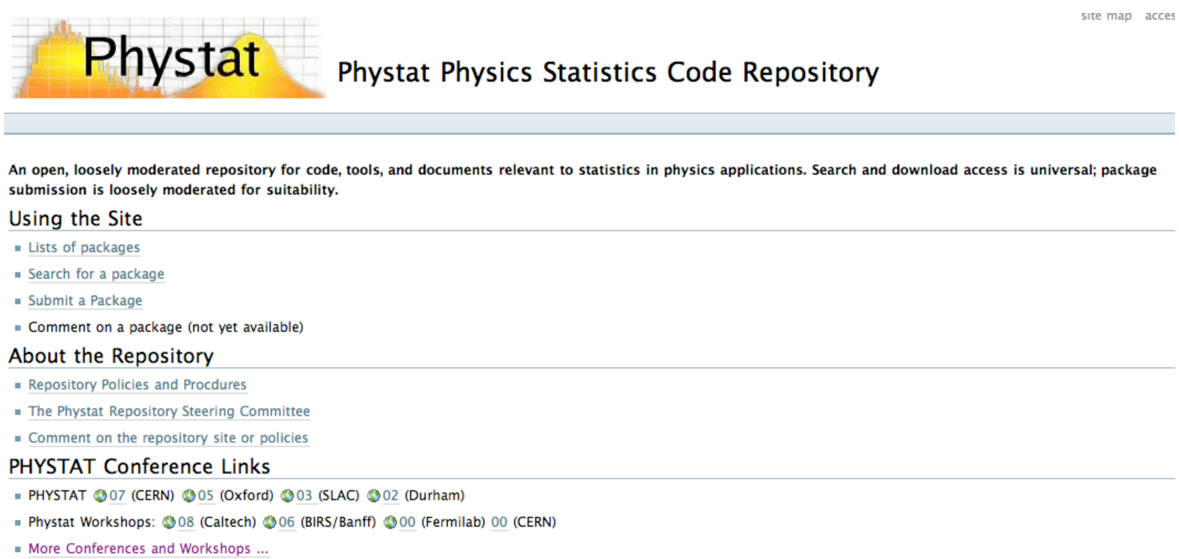
http://indico.cern.ch/conferenceDisplay.py?confId=a063350

Bob Cousins gave a CMS lecture, may give it more publicly

Gary Feldman "Journeys of an Accidental Statistician"

http://www.hepl.harvard.edu/~feldman/Journeys.pdf

The PhyStat conference series at PhyStat.org:



Comments on these lectures

I also gave "Statistics for LHC" academic training lectures in 2009

http://indico.cern.ch/conferenceDisplay.py?confId=48425

Now that we have data, I will put emphasis on realistic problems representative of current analyses

2009	2011
Foundations of Probability	Modeling & Scientific Narrative
Hypothesis Tests	Hypothesis Tests
Confidence Intervals	Confidence Intervals
Generalization for complex problems	Bayesian Methods
	Likelihood Methods

Lecture 1

Preliminaries

Probability Density Functions

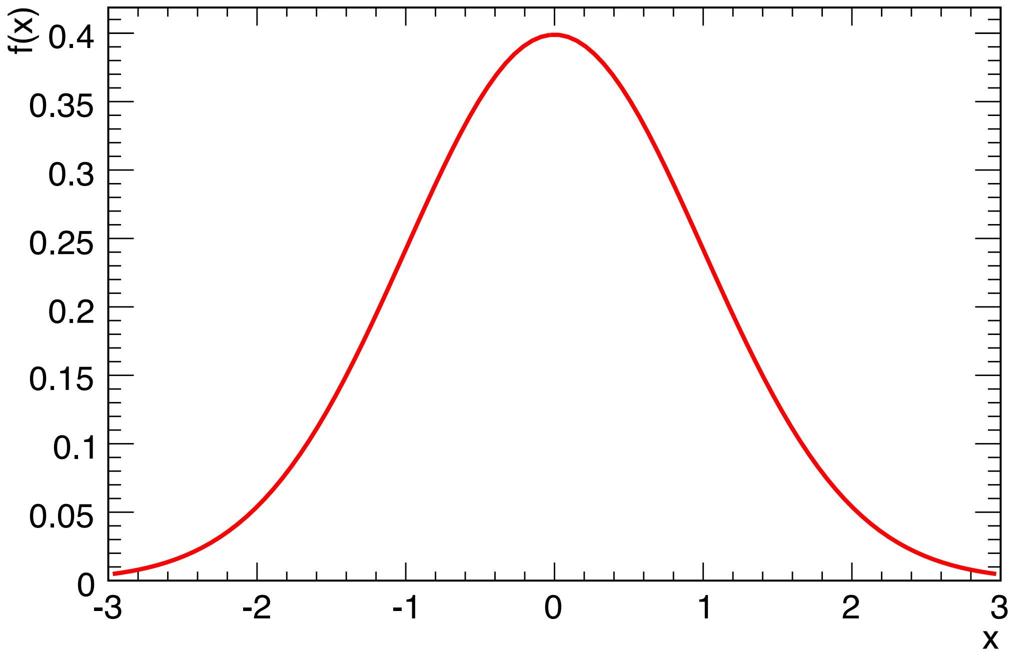
When dealing with continuous random variables, need to introduce the notion of a **Probability Density Function** (PDF... not parton distribution function)

$$P(x \in [x, x+dx]) = f(x)dx$$

Note, $f(x)$ is NOT a probability

PDFs are always normalized

$$\int_{-\infty}^{\infty} f(x)dx = 1$$



Probability Density Functions

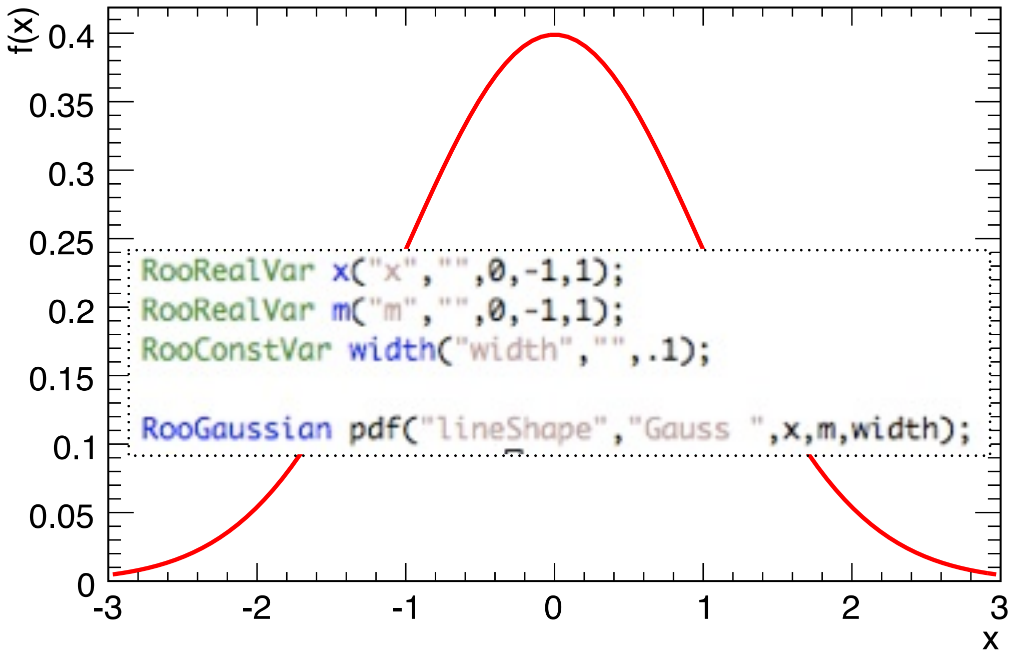
When dealing with continuous random variables, need to introduce the notion of a **Probability Density Function** (PDF... not parton distribution function)

$$P(x \in [x, x+dx]) = f(x)dx$$

Note, $f(x)$ is NOT a probability

PDFs are always normalized

$$\int_{-\infty}^{\infty} f(x)dx = 1$$



The Likelihood Function

A Poisson distribution describes a discrete event count n for a real-valued mean μ.

$$Pois(n|\mu) = \mu^n \frac{e^{-\mu}}{n!}$$

The likelihood of μ given n is the same equation evaluated as a function of μ

- Now it's a continuous function
- But it is not a pdf!

$$L(\mu) = Pois(n|\mu)$$

Common to plot the -2 ln *L*

- helps avoid thinking of it as a PDF
- connection to χ^2 distribution

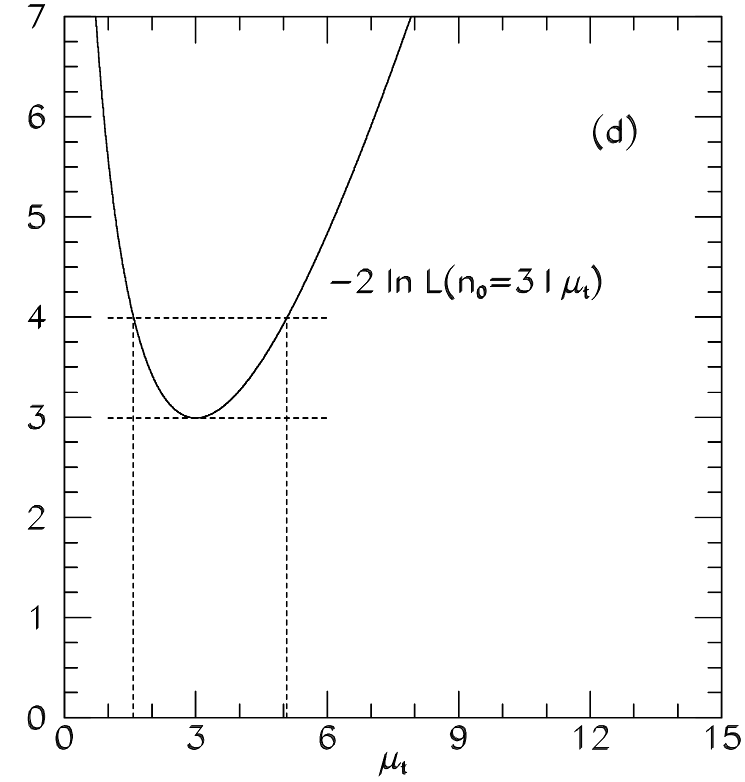


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

Parametric PDFs

CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS

Many familiar PDFs are considered **parametric**

- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- defines a family of distributions
- allows one to make inference about parameters

I will represent PDFs graphically as below (directed acyclic graph)

- every node is a real-valued function of the nodes below

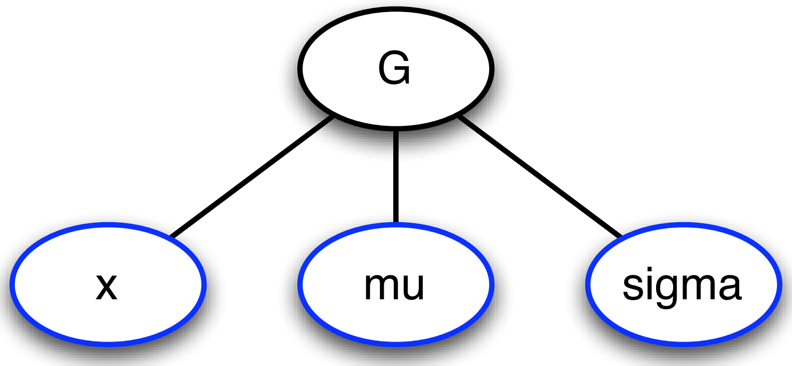
Parametric PDFs

Many familiar PDFs are considered **parametric**

- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- defines a family of distributions
- allows one to make inference about parameters

I will represent PDFs graphically as below (directed acyclic graph)

- every node is a real-valued function of the nodes below



Parametric PDFs

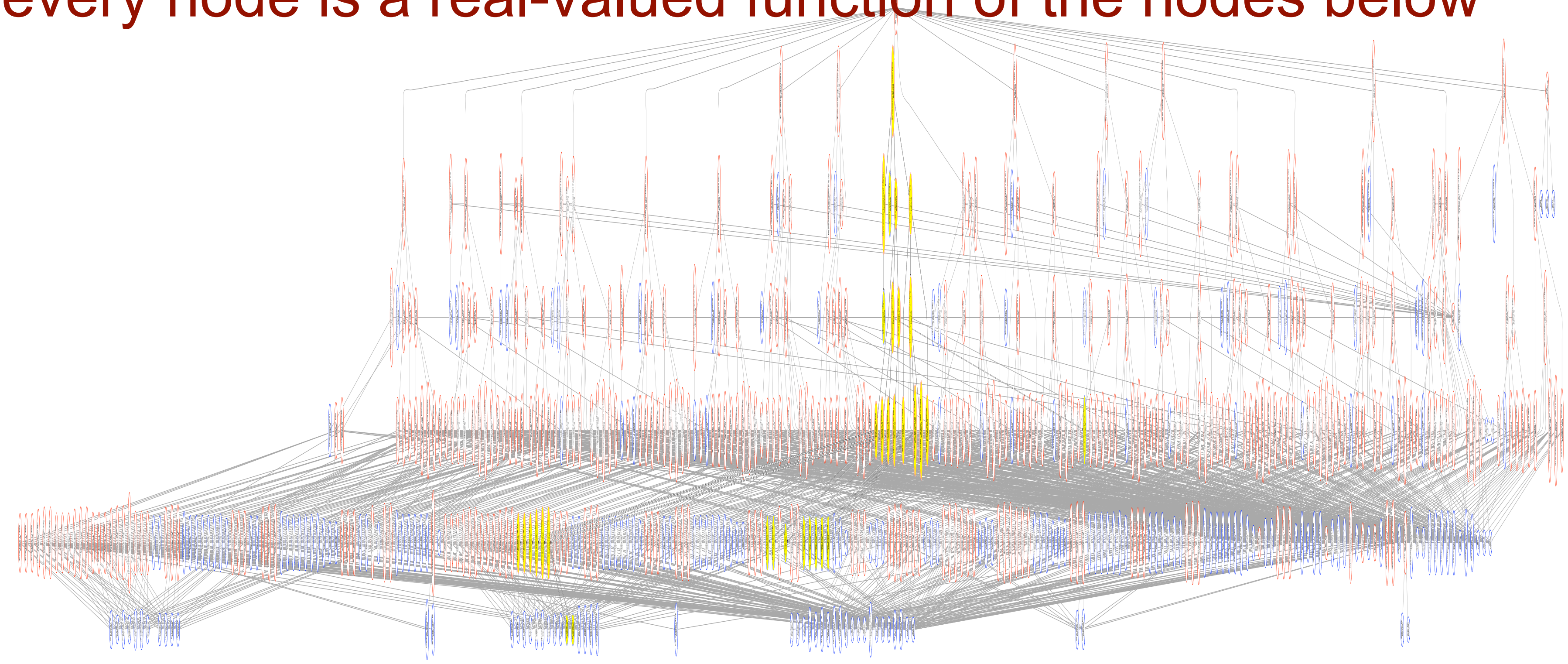
Center for Cosmology and Particle Physics

Many familiar PDFs are considered **parametric**

- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- defines a family of distributions
- allows one to make inference about parameters

I will represent PDFs graphically as below (directed acyclic graph)

- every node is a real-valued function of the nodes below



Modeling:
The Scientific Narrative

Building a model of the data

Before one can discuss statistical tests, one must have a "**model**" for the data.

- by "model", I mean the full structure of P(data | parameters)
  - holding parameters fixed gives a PDF for data
  - ability to evaluate generate pseudo-data (Toy Monte Carlo)
  - holding data fixed gives a **likelihood function** for parameters
    - note, likelihood function is not as general as the full model because it doesn't allow you to generate pseudo-data

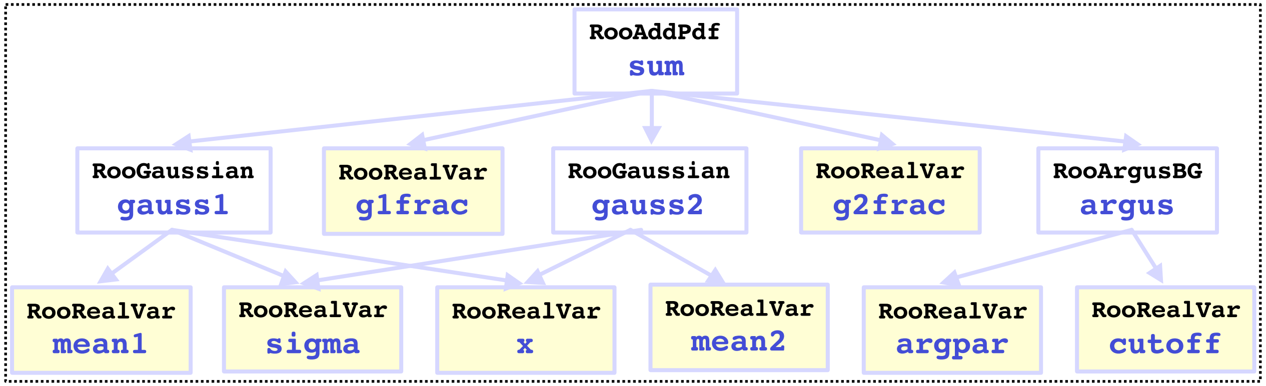
Both Bayesian and Frequentist methods start with the model

- it's the objective part that everyone can agree on
- it's the place where our physics knowledge, understanding, and intuiting comes in
- building a better model is the best way to improve your statistical procedure

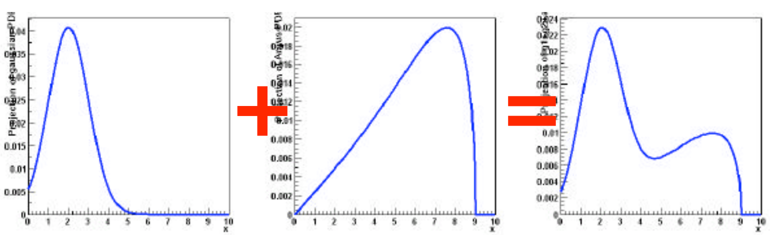
RooFit: A data modeling toolkit

Center for
Cosmology and
Particle Physics

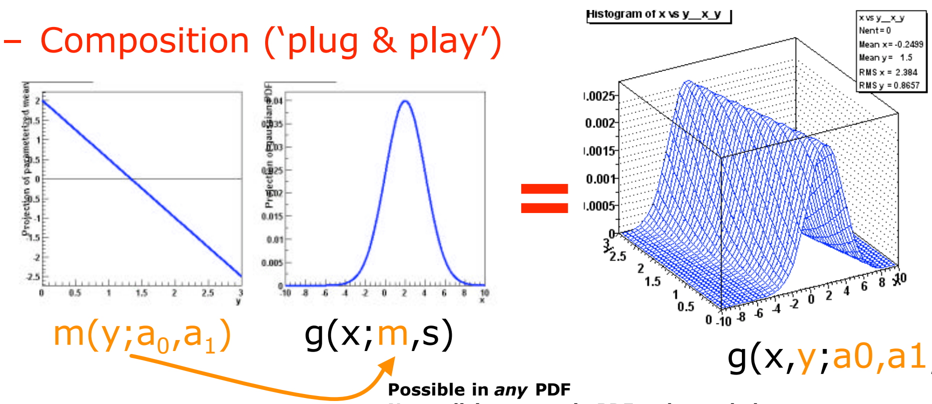
RooFit is a major tool developed at BaBar for data modeling.
RooStats provides higher-level statistical tools based on these PDFs.



– Addition

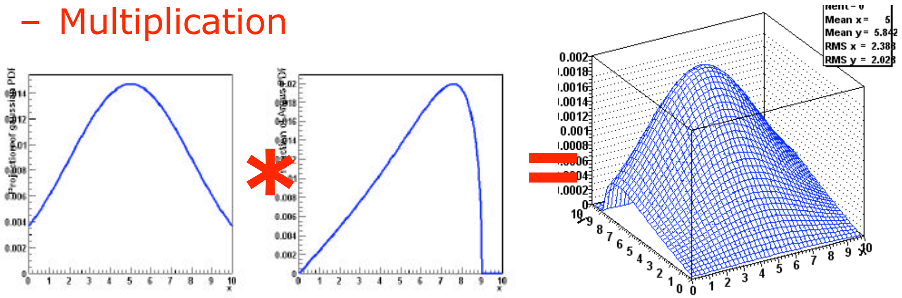


– Composition ('plug & play')



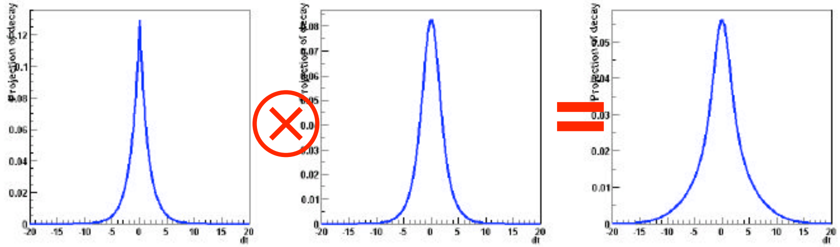
Possible in *any* PDF
No explicit support in PDF code needed

– Multiplication



Wouter Verkerke,

– Convolution



Wouter Verkerke, UCSB

The Scientific Narrative

The model can be seen as a quantitative summary of the analysis

- If you were asked to justify your modeling, you would tell a **story** about why you know what you know
  - based on previous results and studies performed along the way
- the quality of the result is largely tied to how convincing this story is and how tightly it is connected to model

I will describe a few "narrative styles"

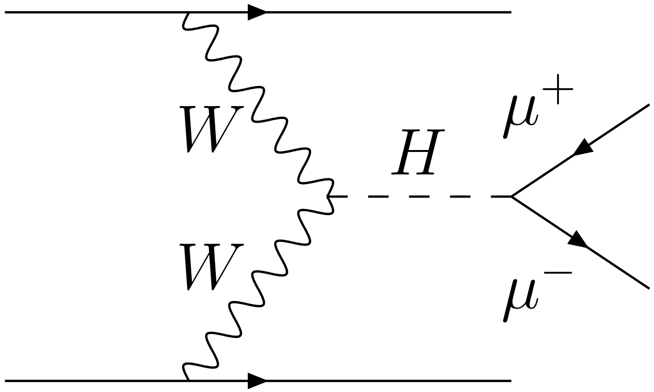
- The "Monte Carlo Simulation" narrative
- The "Data Driven" narrative
- The "Effective Modeling" narrative
- The "Parametrized Response" narrative

Real-life analyses often use a mixture of these

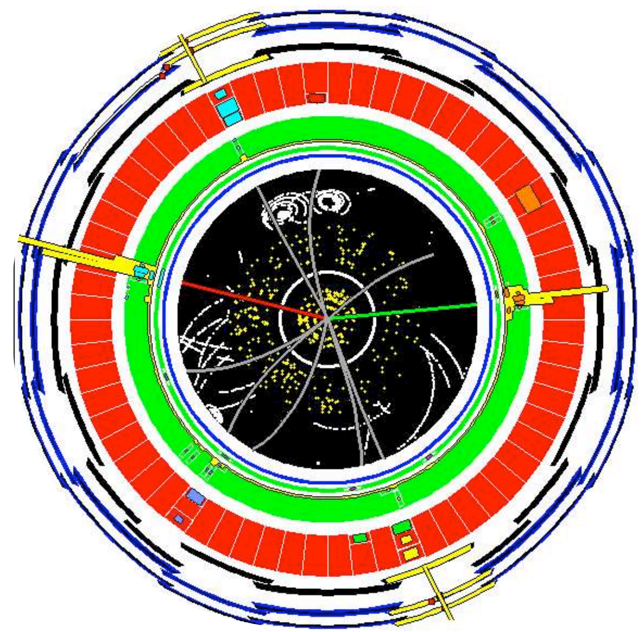
The Monte Carlo Simulation narrative

Let’s start with “the Monte Carlo simulation narrative”, which is probably the most familiar

$P =$



$\oplus$



Cross-sections and event rates

From the many, many collision events, we impose some criteria to select n candidate signal events. We hypothesize that it is composed of some number of signal and background events.

$$\mathrm{Pois}(n|s+b)$$

The number of events that we expect from a given interaction process is given as a product of

- L : a time-integrated luminosity (units 1/cm^2) that serves as a measure of the amount of data that we have collected or the number of trials we have had to produce signal events
- σ : "cross-section" (units cm^2) a quantity that can be calculated from theory
- ε : fraction of signal events selected by selection criteria

The simulation narrative

1) The language of the Standard Model is Quantum Field Theory
Phase space Ω defines initial measure, sampled via Monte Carlo

$$P = \frac{|\langle f|i\rangle|^2}{\langle f|f\rangle\langle i|i\rangle}$$

$$P \to L\sigma$$

$$d\sigma \to |\mathcal{M}|^2 d\Omega$$

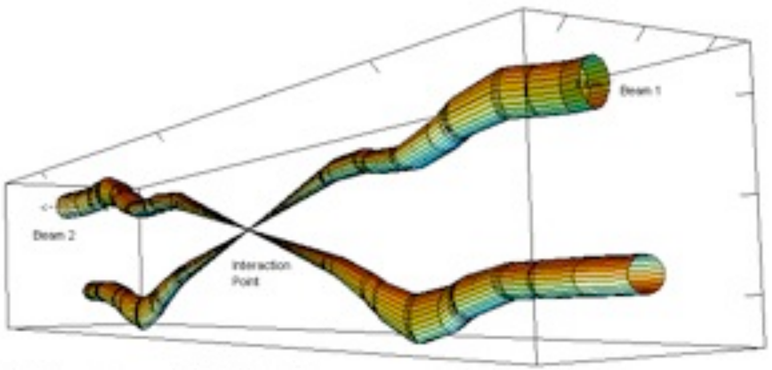
The simulation narrative

1) The language of the Standard Model is Quantum Field Theory
Phase space Ω defines initial measure, sampled via Monte Carlo

$$P = \frac{|\langle f|i\rangle|^2}{\langle f|f\rangle\langle i|i\rangle}$$

$$P \rightarrow L\sigma$$

$$d\sigma \rightarrow |\mathcal{M}|^2 d\Omega$$



Relative beam sizes around IP1 (Atlas) in collision

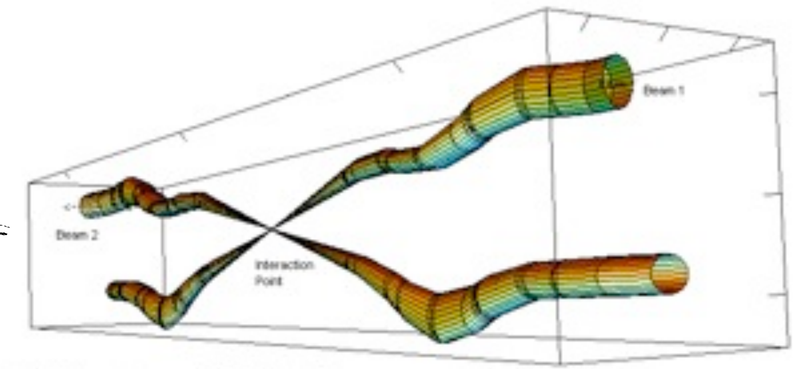
1) The language of the Standard Model is Quantum Field Theory

Phase space Ω defines initial measure, sampled via Monte Carlo

$$P = \frac{|\langle f|i\rangle|^2}{\langle f|f\rangle\langle i|i\rangle}$$

$$P \to L\sigma$$

$$d\sigma \to |\mathcal{M}|^2 d\Omega$$



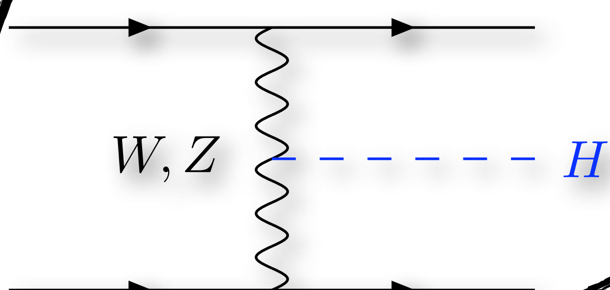
Relative beam sizes around IP1 (Atlas) in collision

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'YB_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)\,\phi\right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^\mu T_a q)\,G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{R}\phi_c L + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$

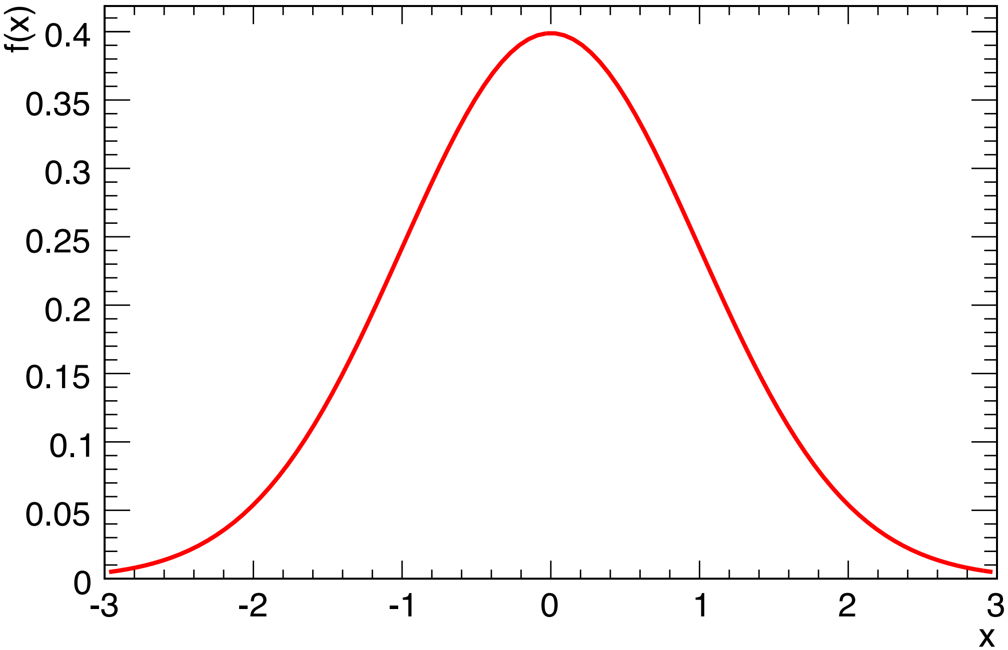


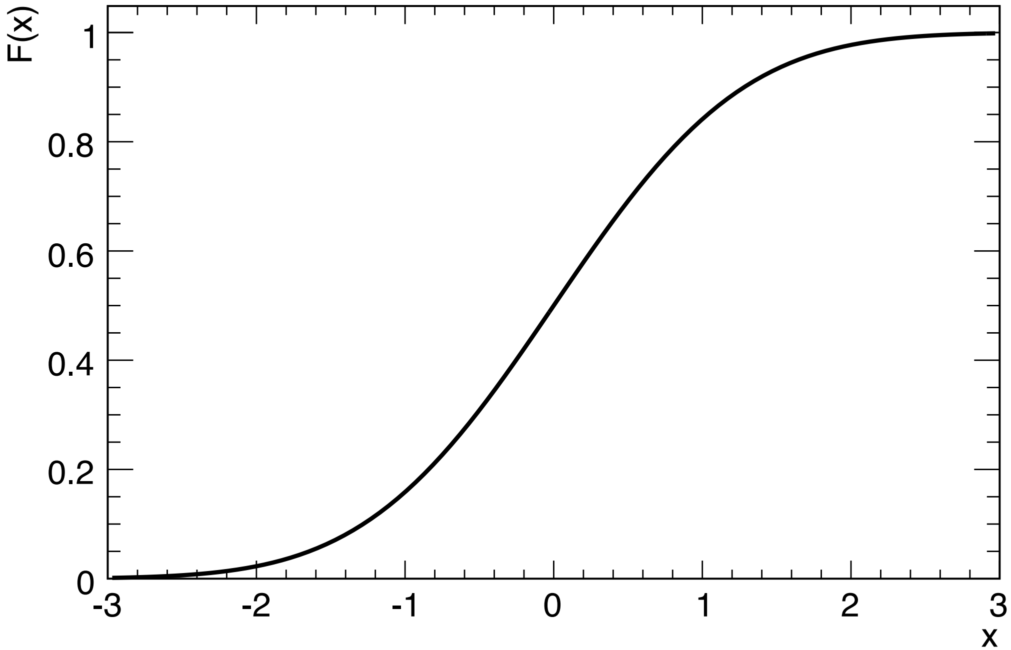
Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension:

$$\int_{-\infty}^{x} f(x')dx' = F(x)$$



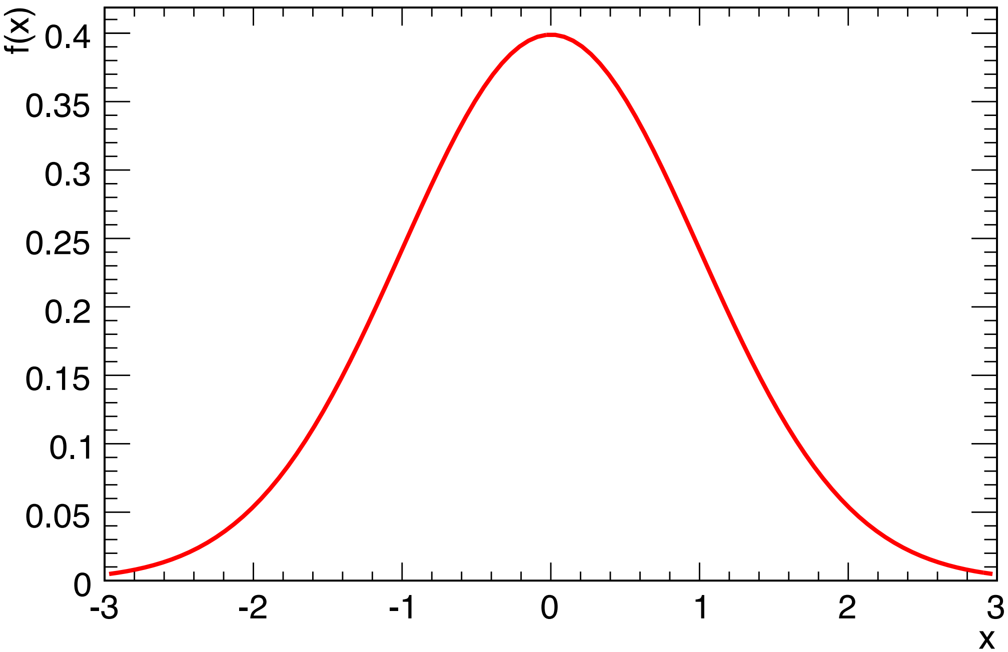


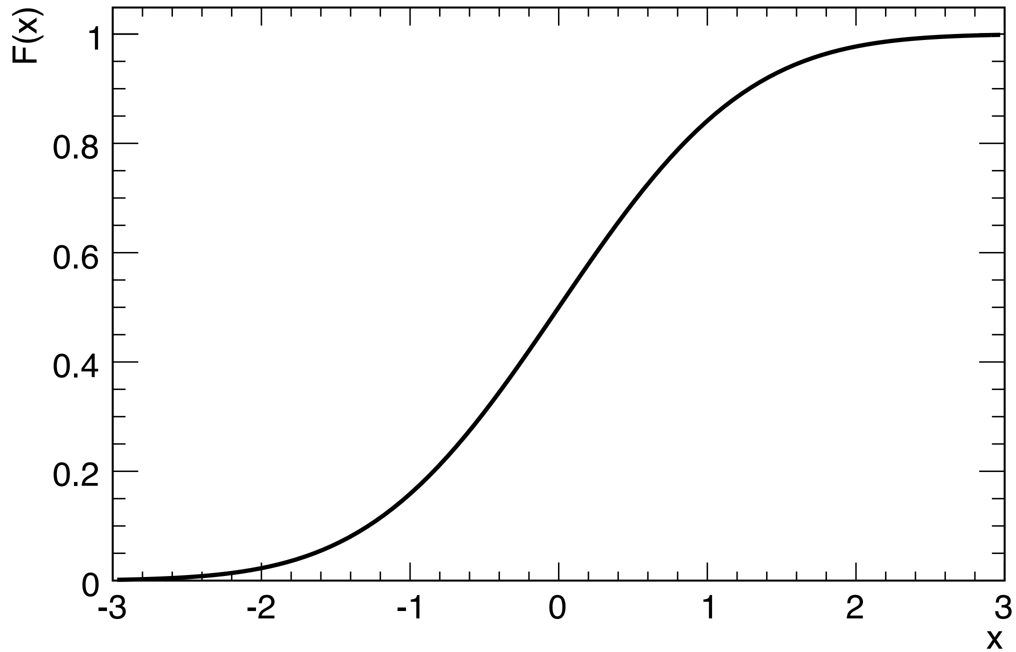
Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension:

$$\int_{-\infty}^{x} f(x')dx' = F(x)$$





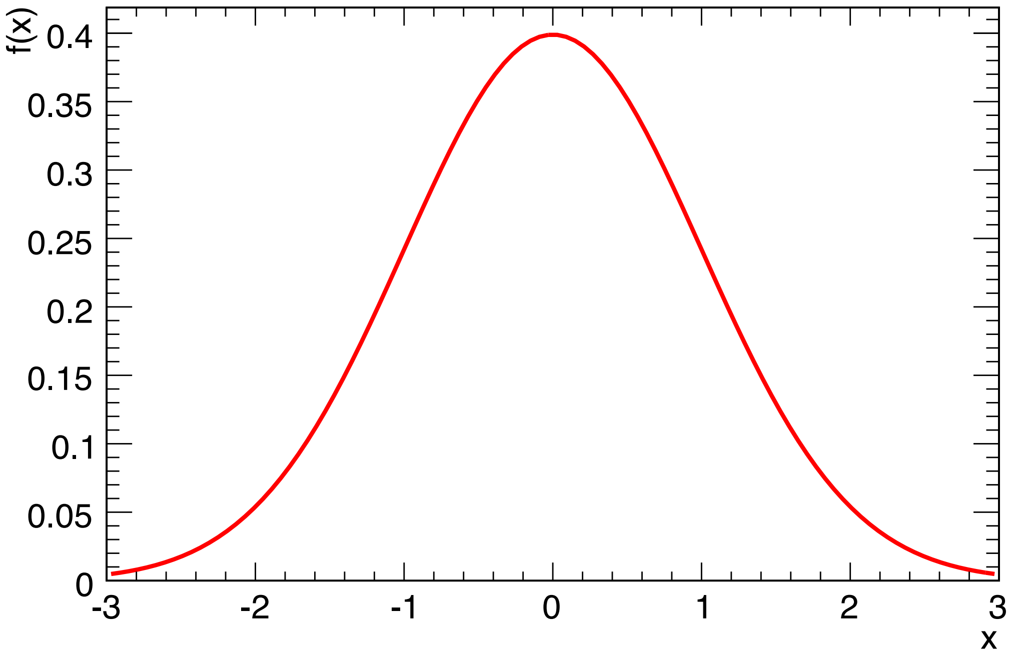
- alternatively, define density as partial of cumulative:

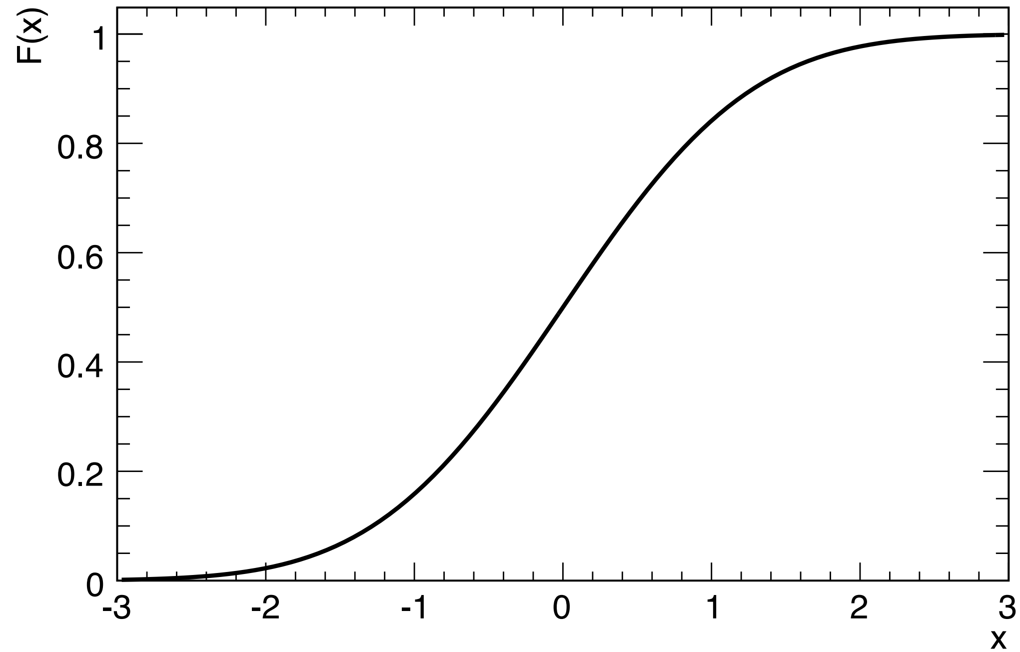
$$f(x) = \frac{\partial F(x)}{\partial x}$$

Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension: $$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

$$f(x) = \frac{\partial F(x)}{\partial x}$$

- same relationship as total and differential cross section:

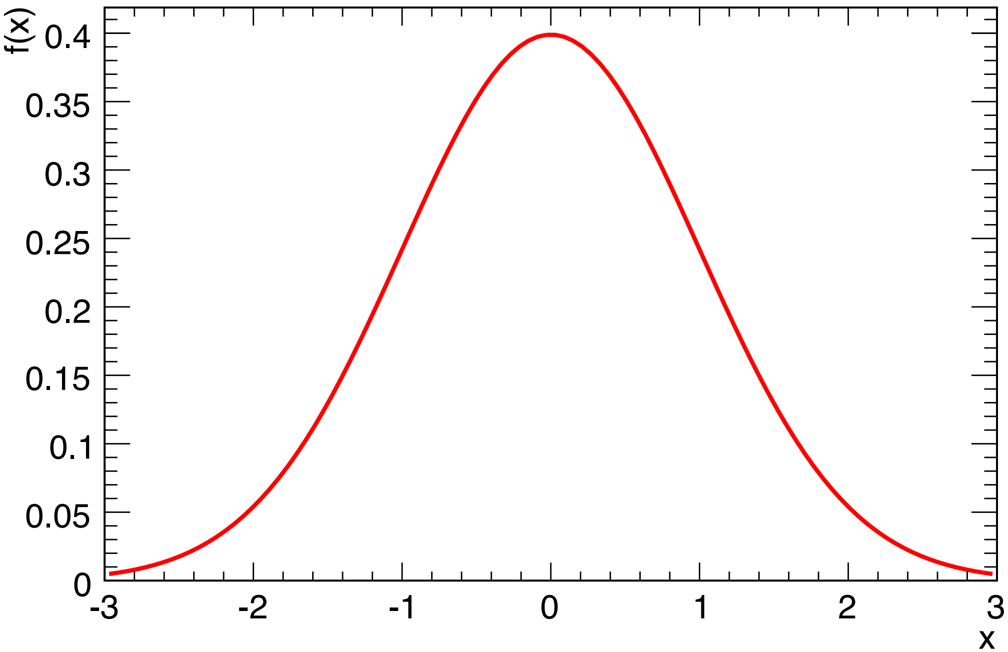
$$f(E) = \frac{1}{\sigma}\frac{\partial \sigma}{\partial E}$$

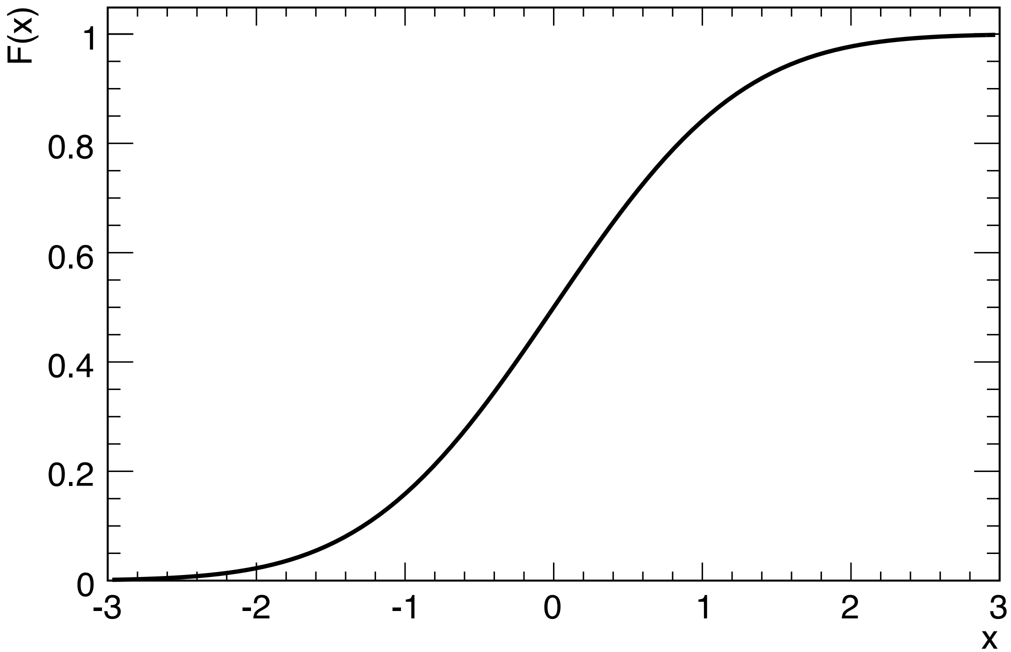
Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension:

$$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

$$f(x) = \frac{\partial F(x)}{\partial x}$$

- same relationship as total and differential cross section:

$$f(E, \eta) = \frac{1}{\sigma} \frac{\partial^2 \sigma}{\partial E \partial \eta}$$

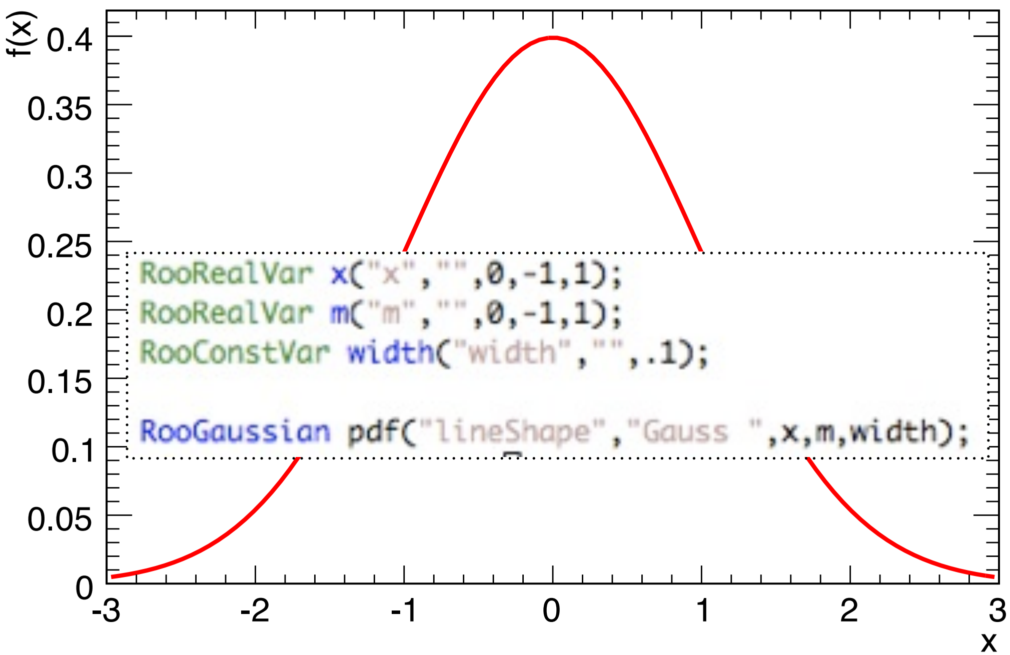
Cumulative Density Functions

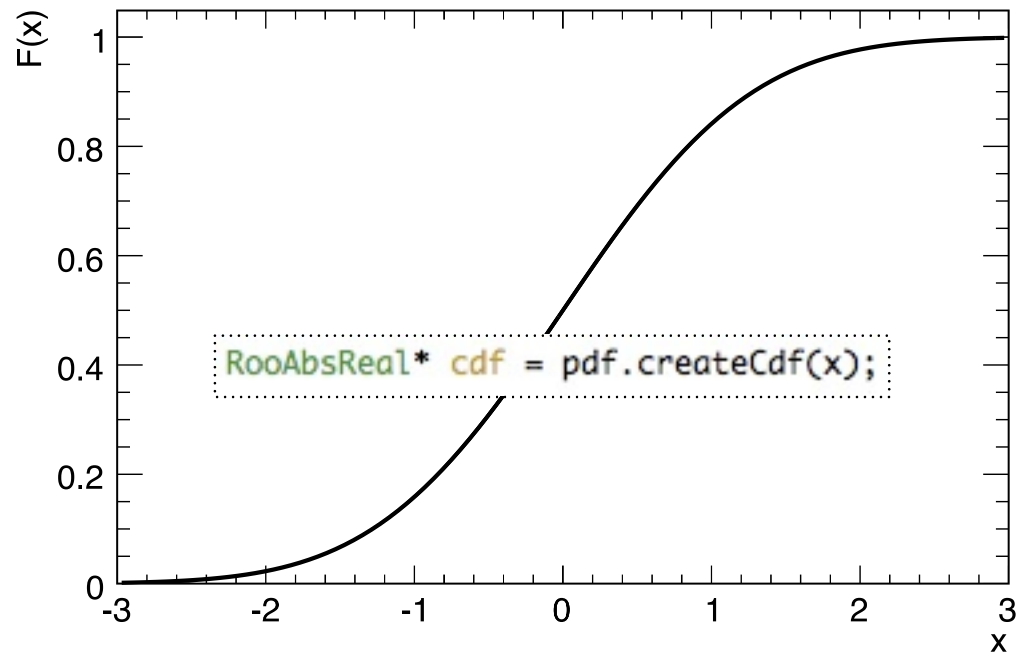
Center for Cosmology and Particle Physics

Often useful to use a cumulative distribution:

- in 1-dimension:

$$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

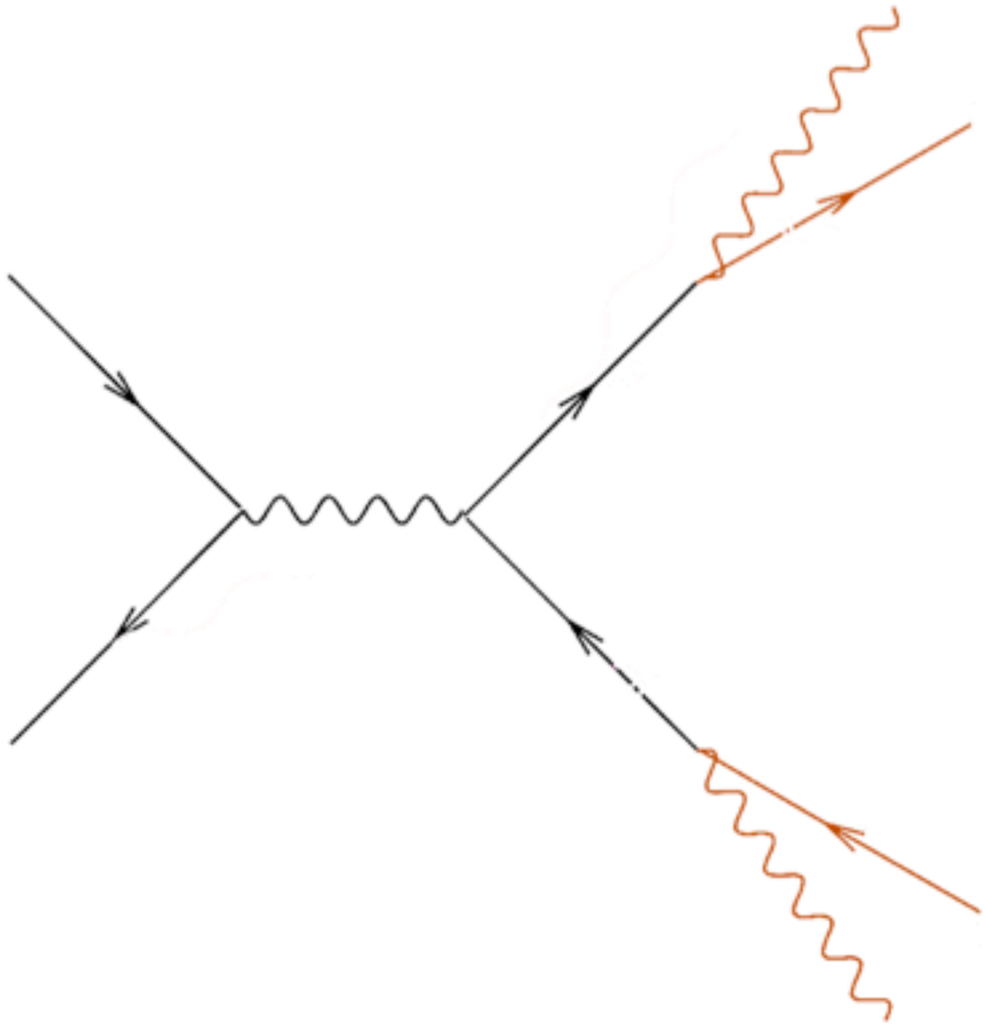
$$f(x) = \frac{\partial F(x)}{\partial x}$$

- same relationship as total and differential cross section:

$$f(E,\eta) = \frac{1}{\sigma}\frac{\partial^2 \sigma}{\partial E \partial \eta}$$

The simulation narrative

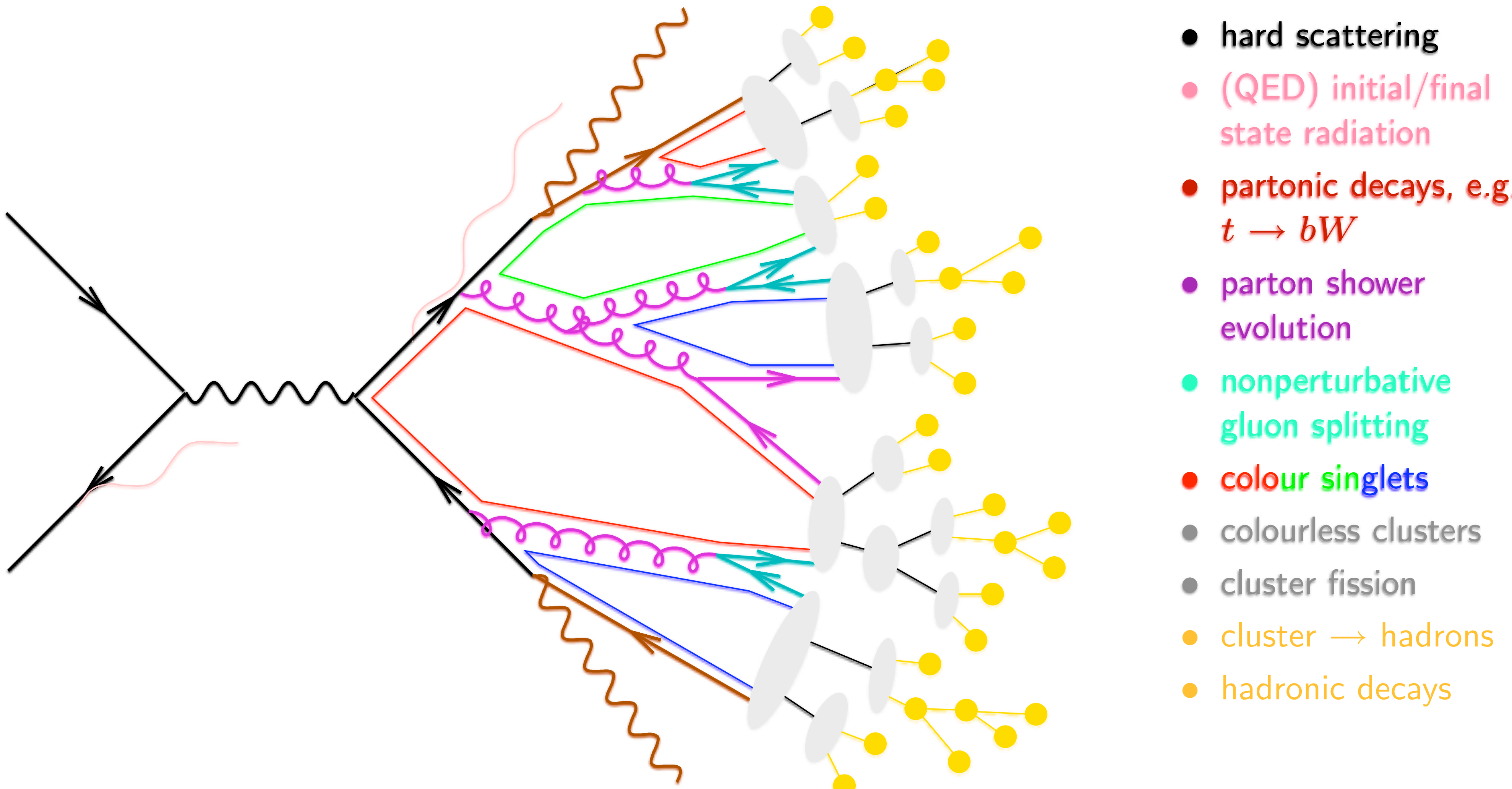
2) a) Perturbation theory used to systematically approximate the theory.
b) splitting functions, Sudokov form factors, and hadronization models
c) all sampled via accept/reject Monte Carlo **P(particles | partons)**



- hard scattering
- partonic decays, e.g. $t \to bW$

The simulation narrative

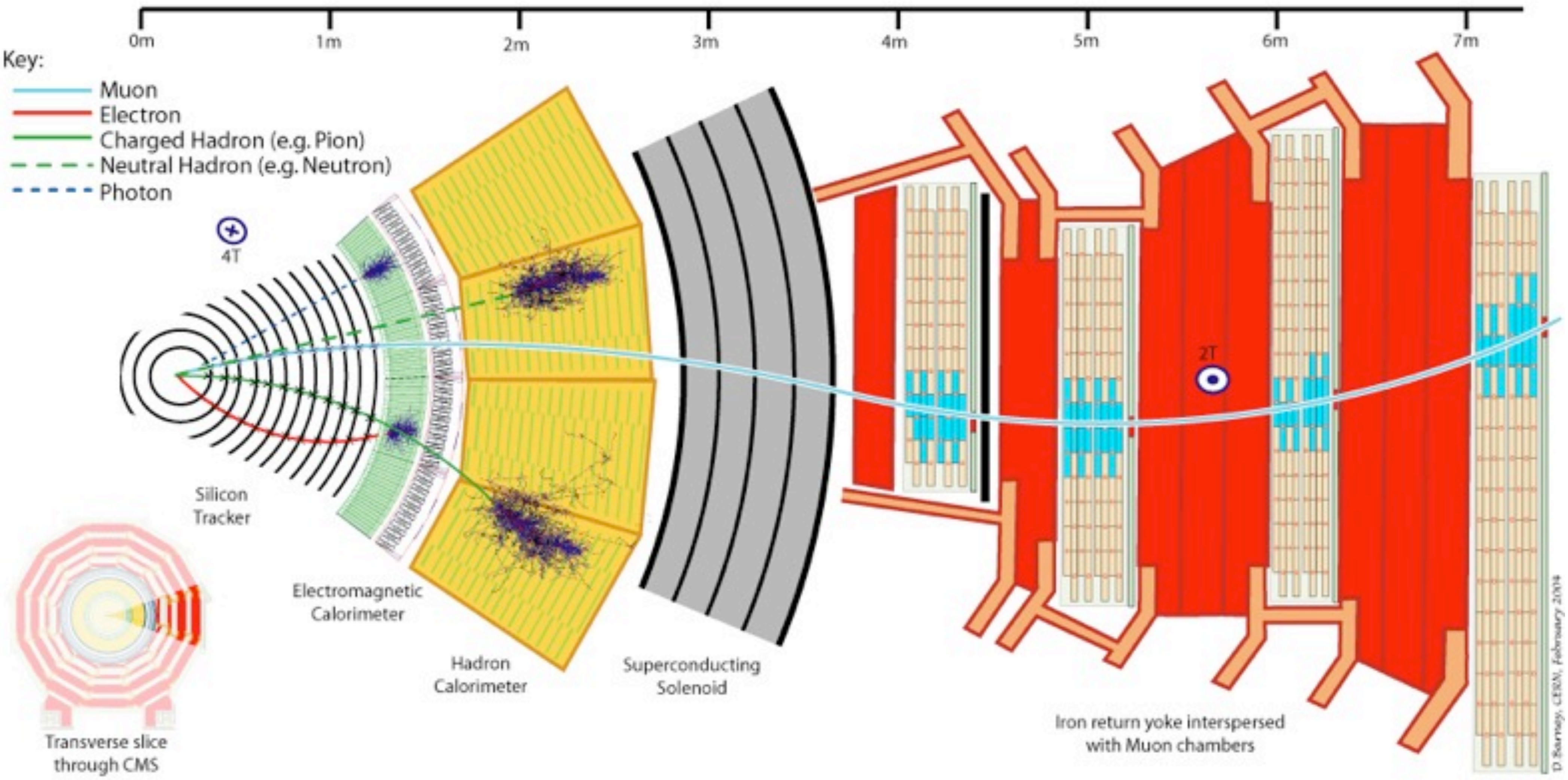
2) a) Perturbation theory used to systematically approximate the theory.
b) splitting functions, Sudokov form factors, and hadronization models
c) all sampled via accept/reject Monte Carlo **P(particles | partons)**



The simulation narrative

Center for
Cosmology and
Particle Physics

3) Next, the interaction of outgoing particles with the detector is simulated.
Detailed simulations of particle interactions with matter.
Accept/reject style Monte Carlo integration of very complicated function
P(detector readout | initial particles)



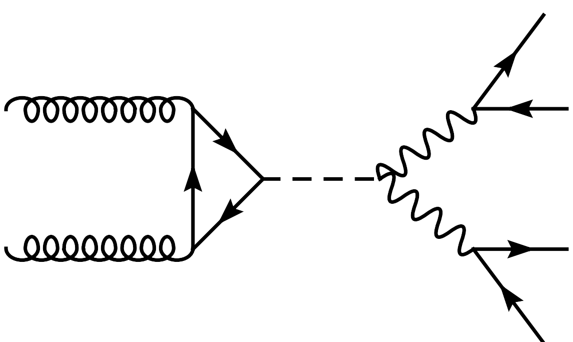
Theoretical Predictions

In addition to the rate of interactions, our theories predict the distributions of angles, energies, masses, etc. of particles produced

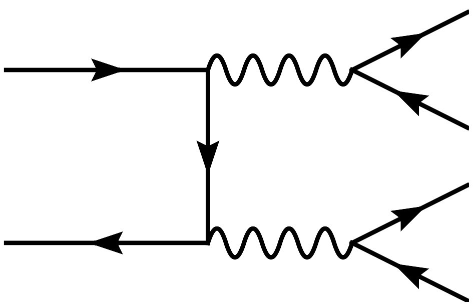
- we form functions of these called **discriminating variables** *m*,
- and use Monte Carlo techniques to estimate *f(m)*

In addition to the hypothesized signal process, there are known background processes.

- thus, the distribution of *f(m)* is a **mixture model**
- the full model is a **marked Poisson process**



signal process



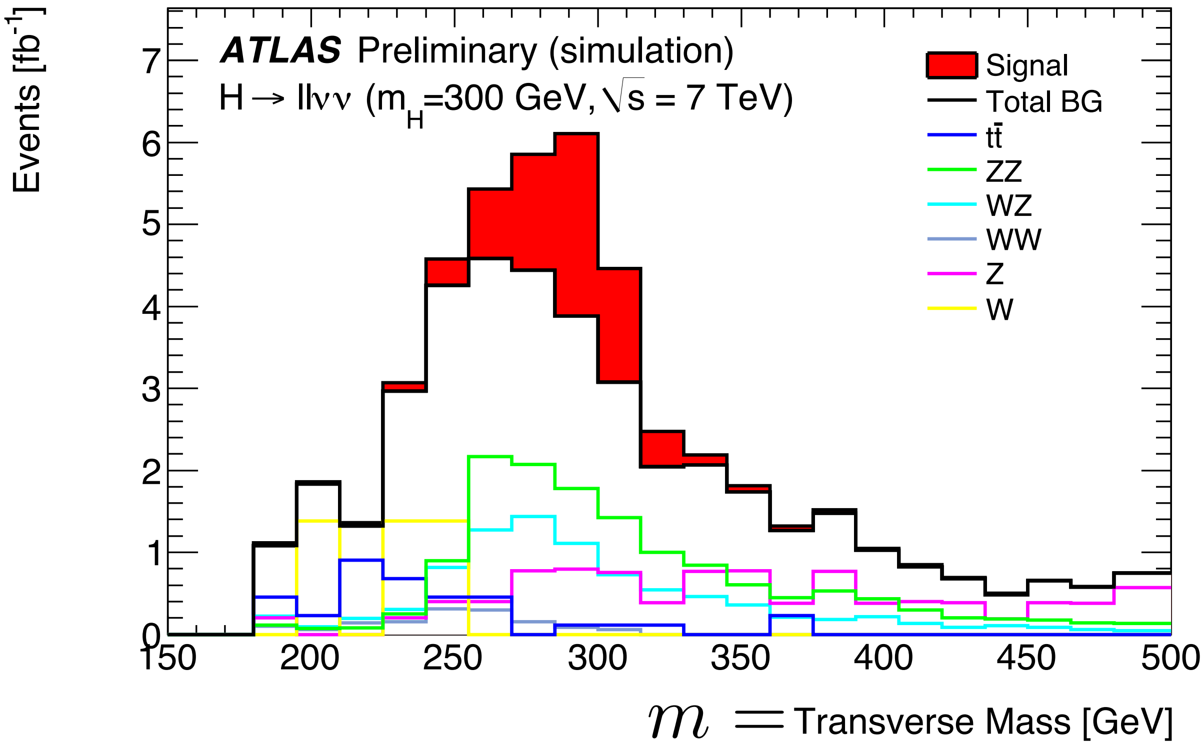
background process

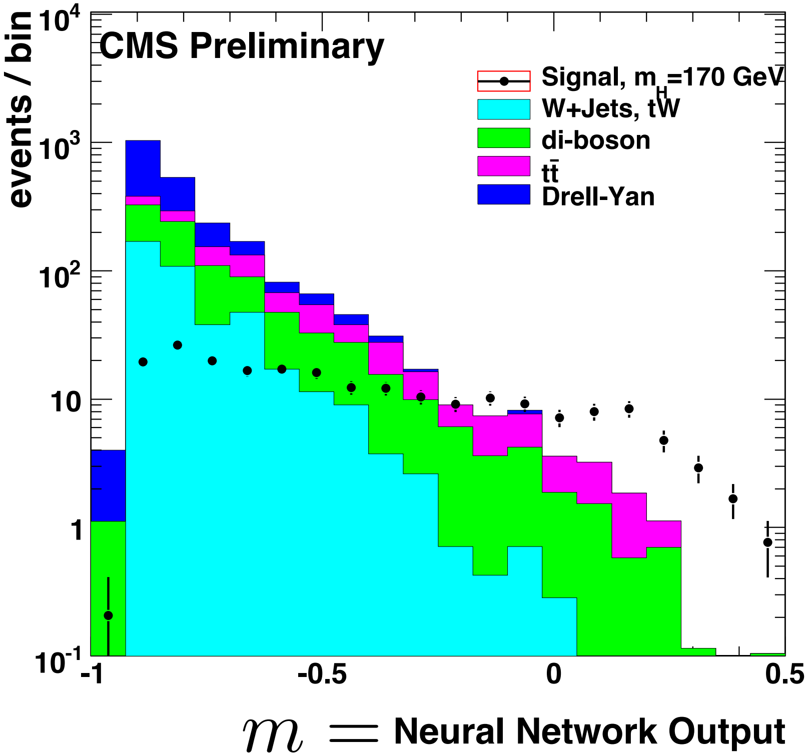
$$P(\mathbf{m}|s) = \text{Pois}(n|s+b)\prod_{j}^{n}\frac{sf_s(m_j)+bf_b(m_j)}{s+b}$$

Example model

Here is an example prediction from search for H→ZZ and H→WW

- sometimes multivariate techniques are used





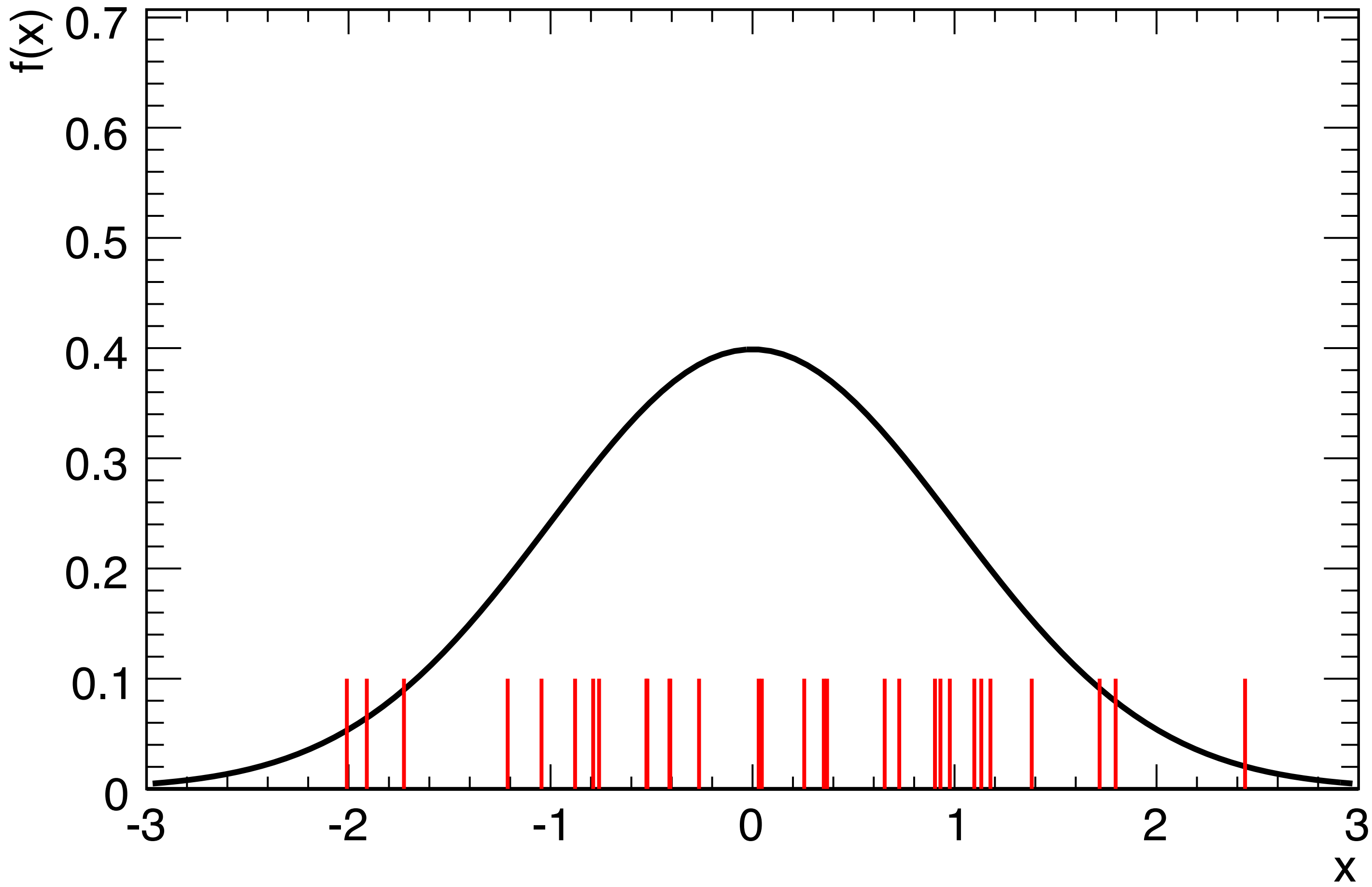
$$P(\mathbf{m}|s) = \mathrm{Pois}(n|s+b) \prod_j^n \frac{s f_s(m_j) + b f_b(m_j)}{s+b}$$

Parametric vs. Non-Parametric PDFs

No parametric form, need to construct **non-parametric** PDFs

From Monte Carlo samples, one has empirical PDF

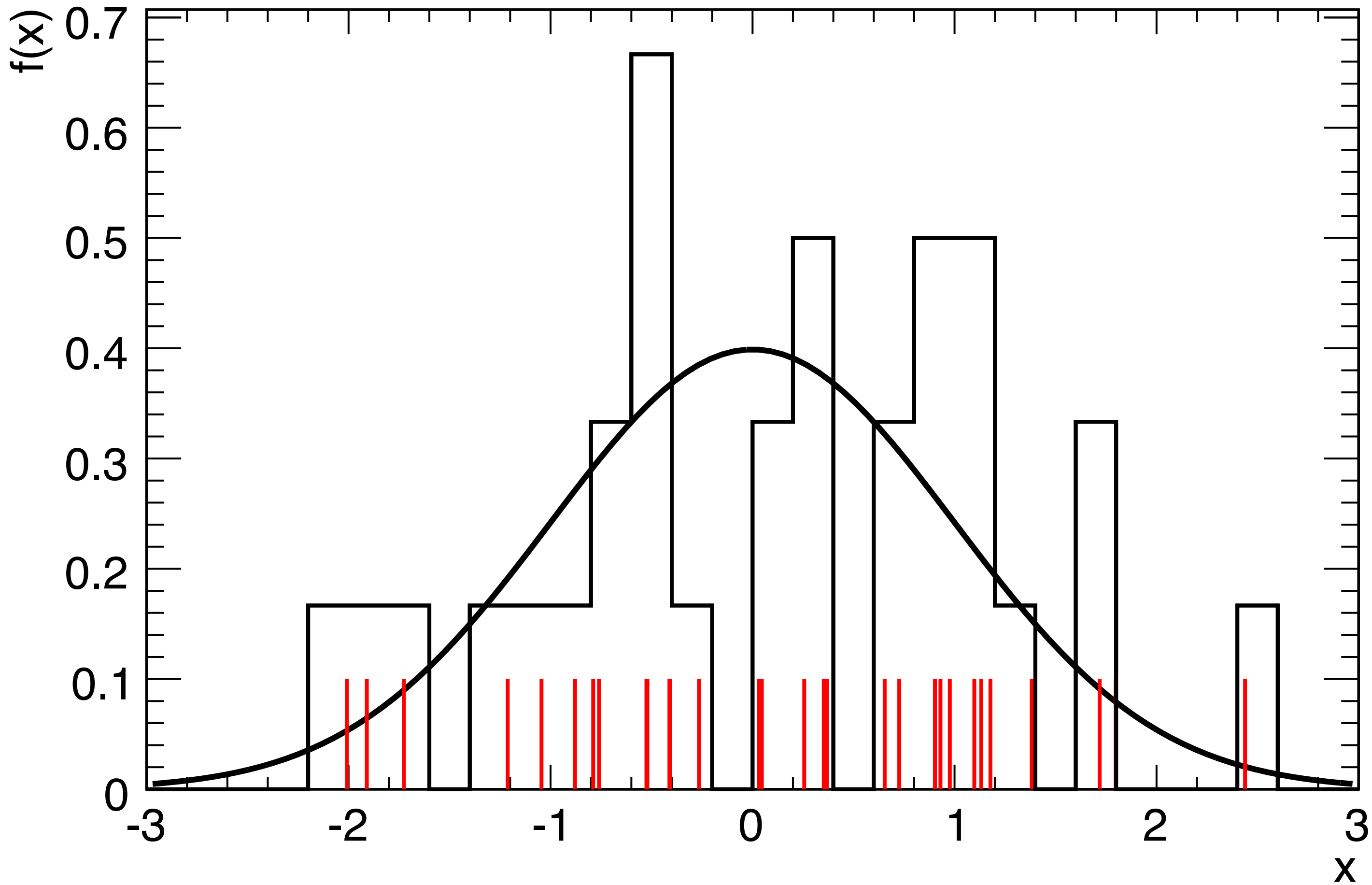
$$f_{emp} = \frac{1}{N}\sum_{i}^{N}\delta(x - x_i)$$



Parametric vs. Non-Parametric PDFs

Classic example of a **non-parametric** PDF is the histogram

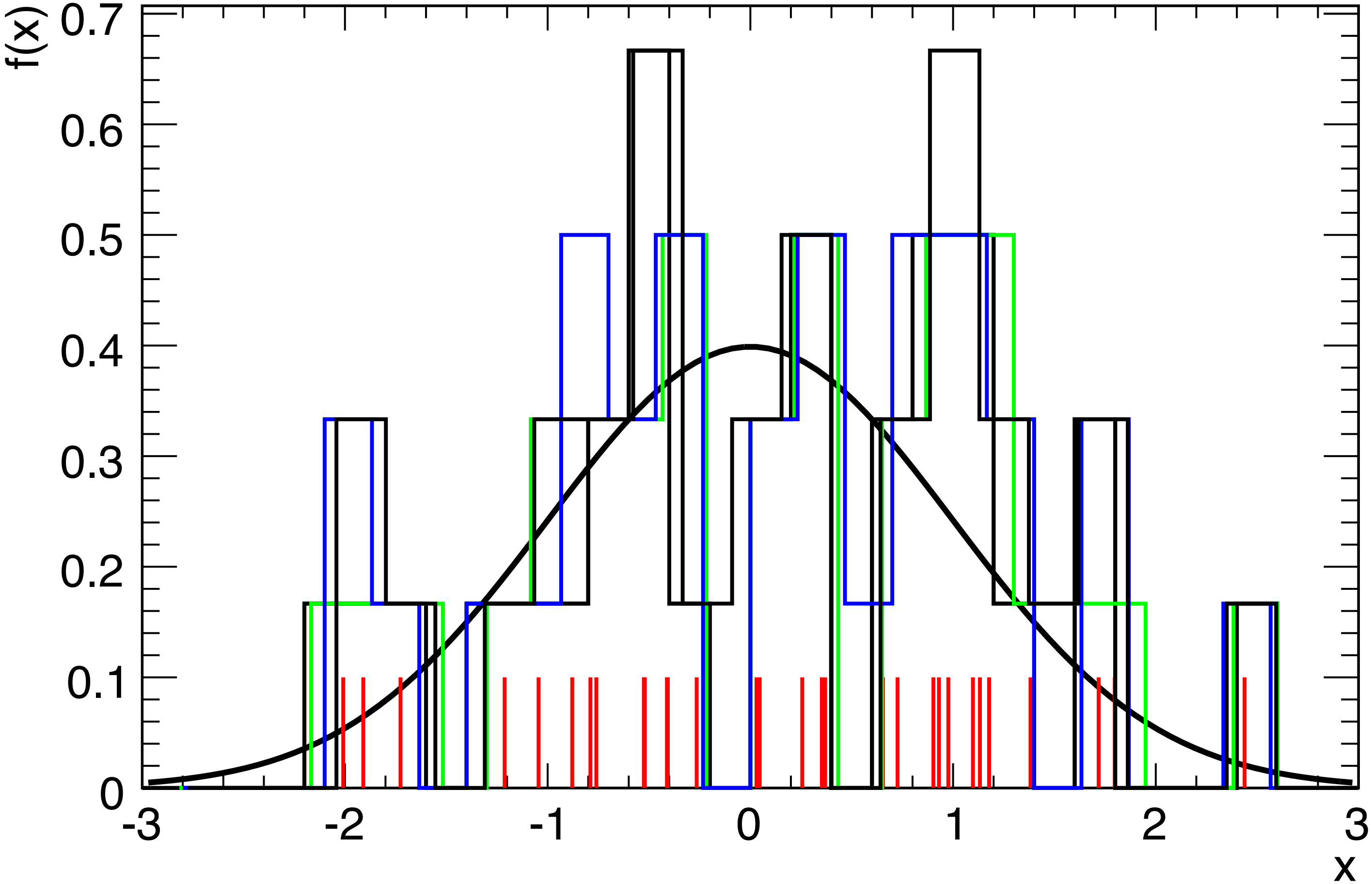
$$f^{w,s}_{hist}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$



Classic example of a **non-parametric** PDF is the histogram

but they depend on bin width and starting position

$$f_{hist}^{w,s}(x) = \frac{1}{N} \sum_i h_i^{w,s}$$

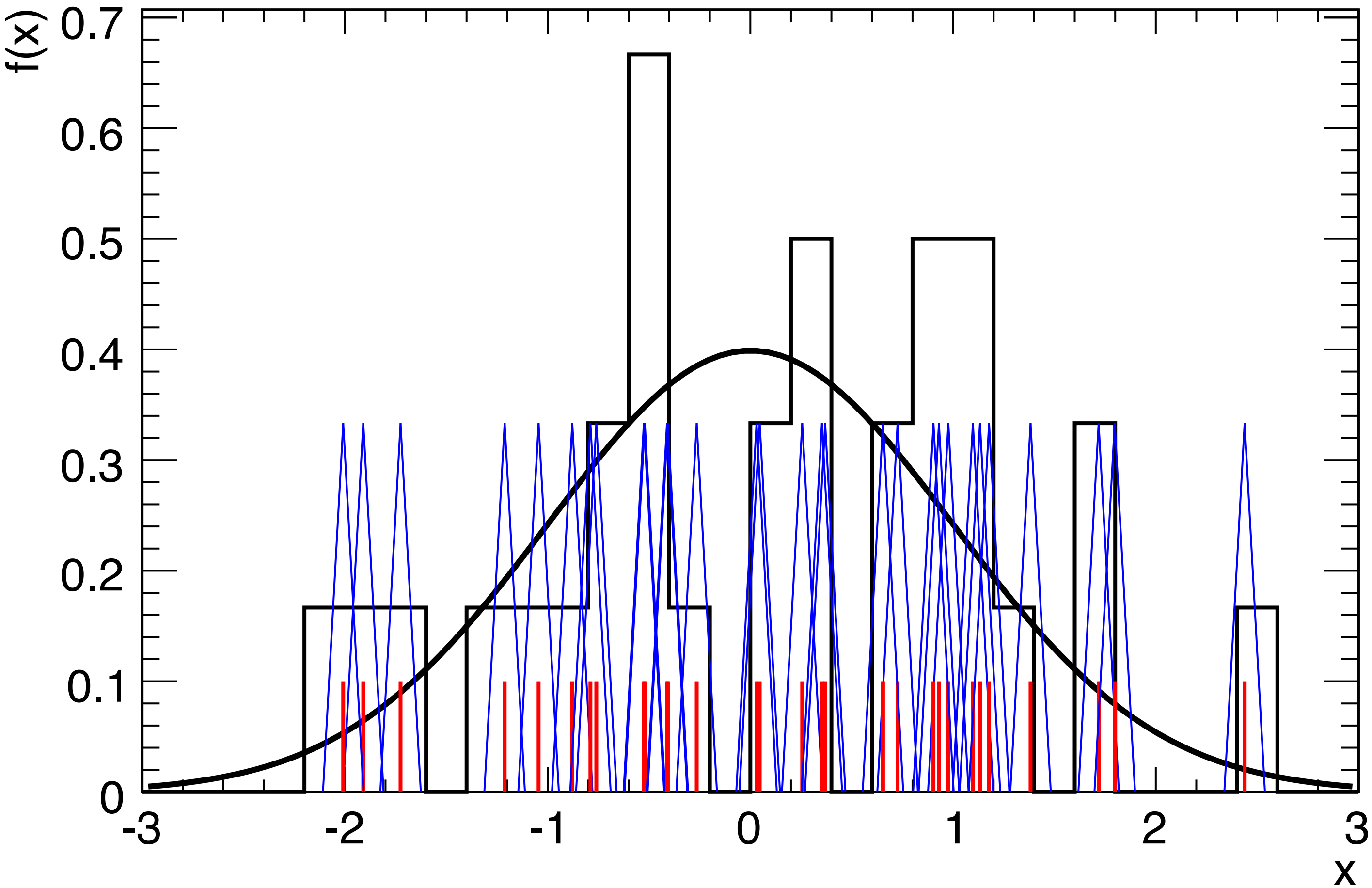


Parametric vs. Non-Parametric PDFs

Classic example of a **non-parametric** PDF is the histogram

"Average Shifted Histogram" minimizes effect of binning

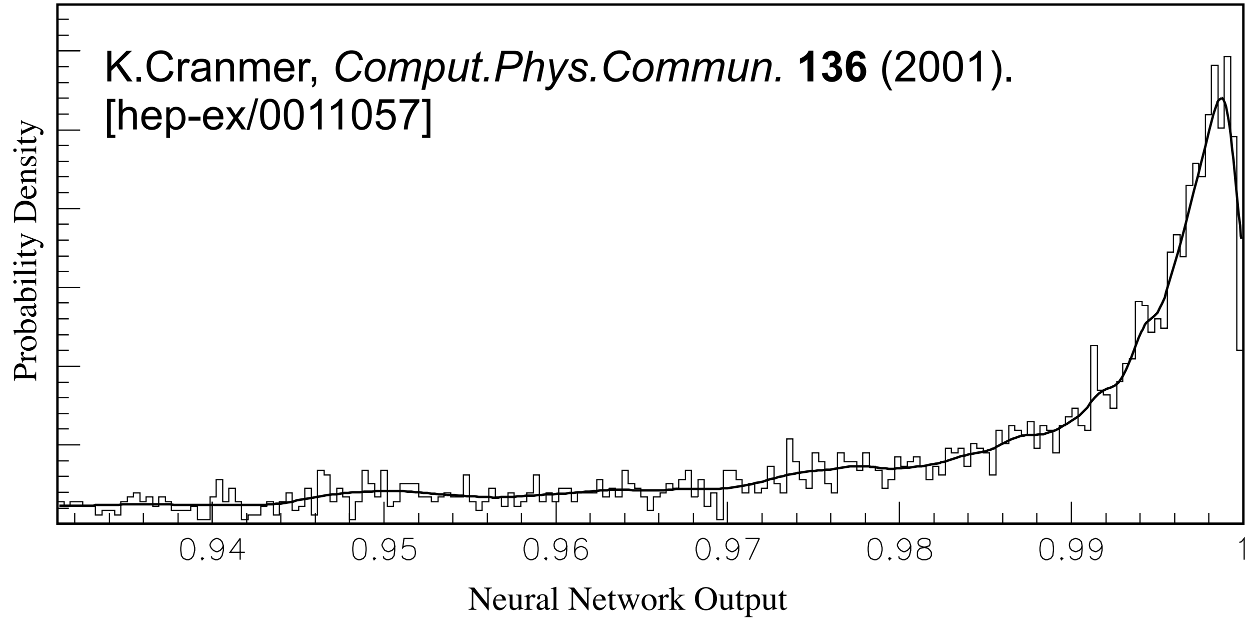
$$f^w_{ASH}(x) = \frac{1}{N}\sum_i^N K^w(x - x_i)$$



Kernel estimation is the generalization of Average Shifted Histograms

$$\hat{f}_1(x) = \sum_i^n \frac{1}{nh(x_i)} K\left(\frac{x - x_i}{h(x_i)}\right)$$

$$h(x_i) = \left(\frac{4}{3}\right)^{1/5} \sqrt{\frac{\sigma}{\hat{f}_0(x_i)}} n^{-1/5}$$



"the data is the model"

Adaptive Kernel estimation puts wider kernels in regions of low probability

Used at LEP for describing pdfs from Monte Carlo (KEYS)

Multivariate, non-parametric PDFs

Kernel Estimation has a nice generalizations to higher dimensions

- practical limit is about 5-d due to curse of dimensionality

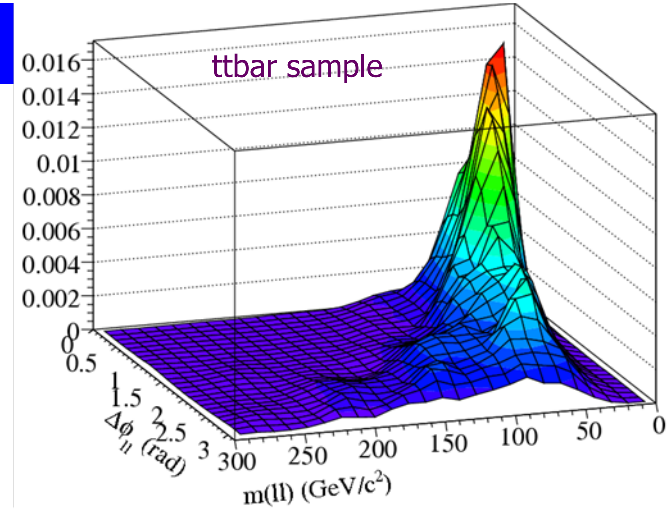
Max Baak has coded N-dim KEYS pdf described in Comput.Phys.Commun. **136** (2001) in RooFit.

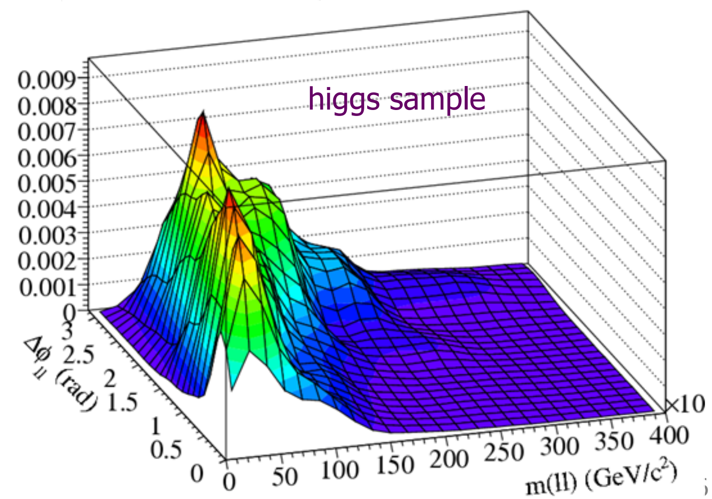
These pdfs have been used as the basis for a multivariate discrimination technique called "PDE"

$$D(\vec{x}) = \frac{f_s(\vec{x})}{f_s(\vec{x}) + f_b(\vec{x})}$$

Correlations

- 2-d projection of pdf from previous slide.
- RooNDKeys pdf automatically models (fine) correlations between observables ...



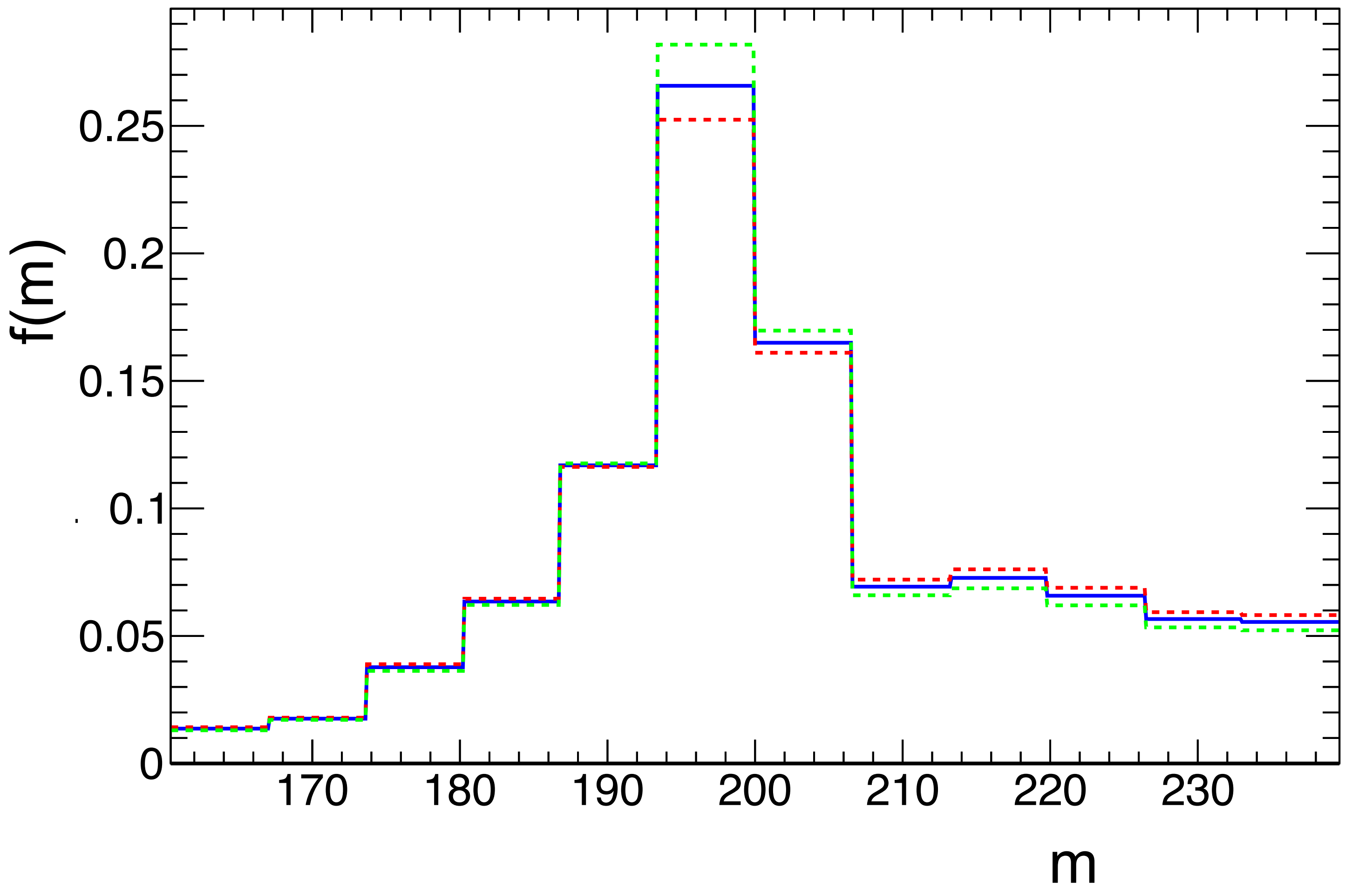


Max Baak

Incorporating Systematic Effects

Of course, the simulation has many adjustable parameters and imperfections that lead to systematic uncertainties.

- one can re-run simulation with different settings and produce **variational histograms** about the **nominal prediction**



Explicit parametrization

Important to distinguish between the **source** of the systematic uncertainty (eg. jet energy scale) and its **effect.**

- The same 5% jet energy scale uncertainty will have different effect on different signal and background processes
  - not necessarily with any obvious functional form
- Usually possible to decompose to independent "uncorrelated" sources

Imagine a table that **explicitly quantifies** the effect of each source of systematic.

- Entries are either normalization factors or variational histograms

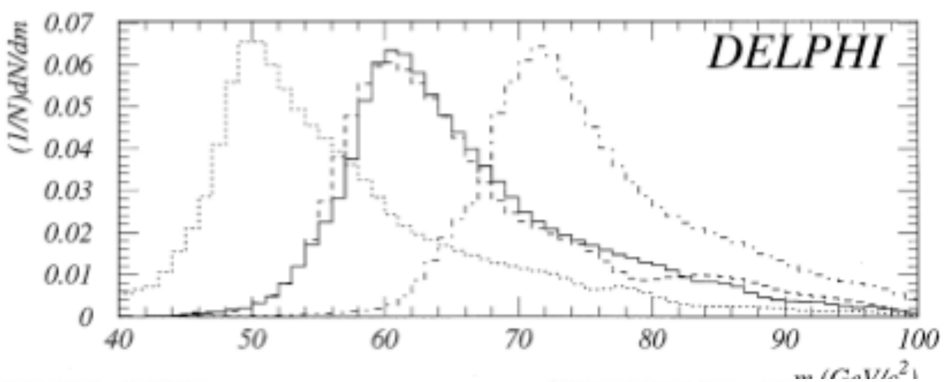
	sig	bkg 1	bkg 2	...
syst 1				
syst 2				
...				

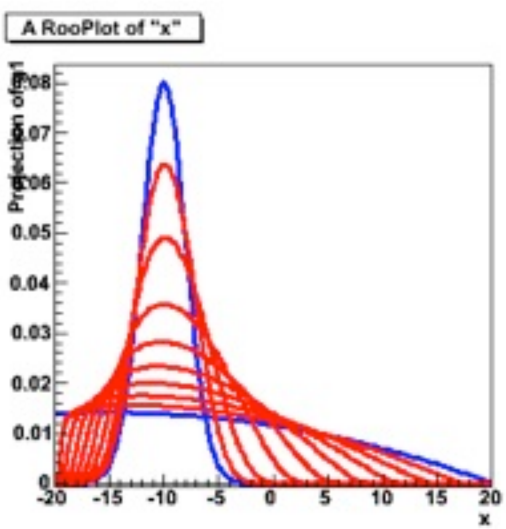
Histogram Interpolation

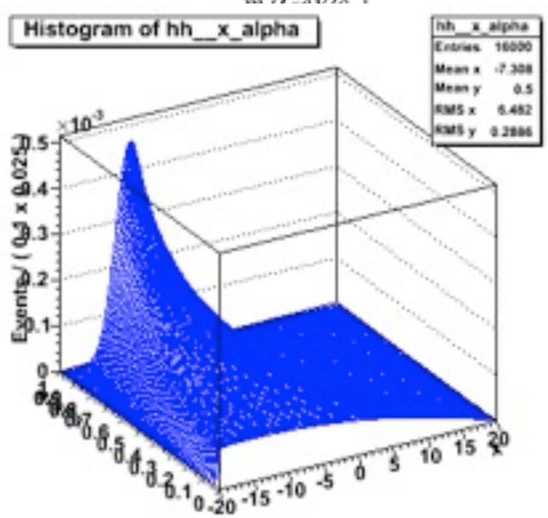
Several interpolation algorithms exist: eg. Alex Read's "horizontal" histogram interpolation algorithm (RooIntegralMorph in RooFit)

- take several PDFs, construct interpolated PDF with additional **nuisance parameter** α

A.L. Read / Nuclear Instruments and Methods in Physics Research A 425 (1999) 357–360







Simple "vertical" interpolation bin-by-bin.

Alternative "horizontal" interpolation algorithm by Max Baak called "RooMomentMorph" in RooFit (faster and numerically more stable)

Incorporating systematics

Let's consider a simplified problem that has been studied quite a bit to gain some insight into our more realistic and difficult problems

- number counting with background uncertainty
  - in our main measurement we observe n_{on} with $s+b$ expected

$$\mathrm{Pois}(n_{\mathrm{on}}|s+b)$$

- and the background has some uncertainty
  - but what is "background uncertainty"? Where did it come from?
  - maybe we would say background is known to 10% or that it has some pdf $\pi(b)$
    - then we often do a smearing of the background:

$$P(n_{\mathrm{on}}|s) = \int db \,\mathrm{Pois}(n_{\mathrm{on}}|s+b)\,\pi(b),$$

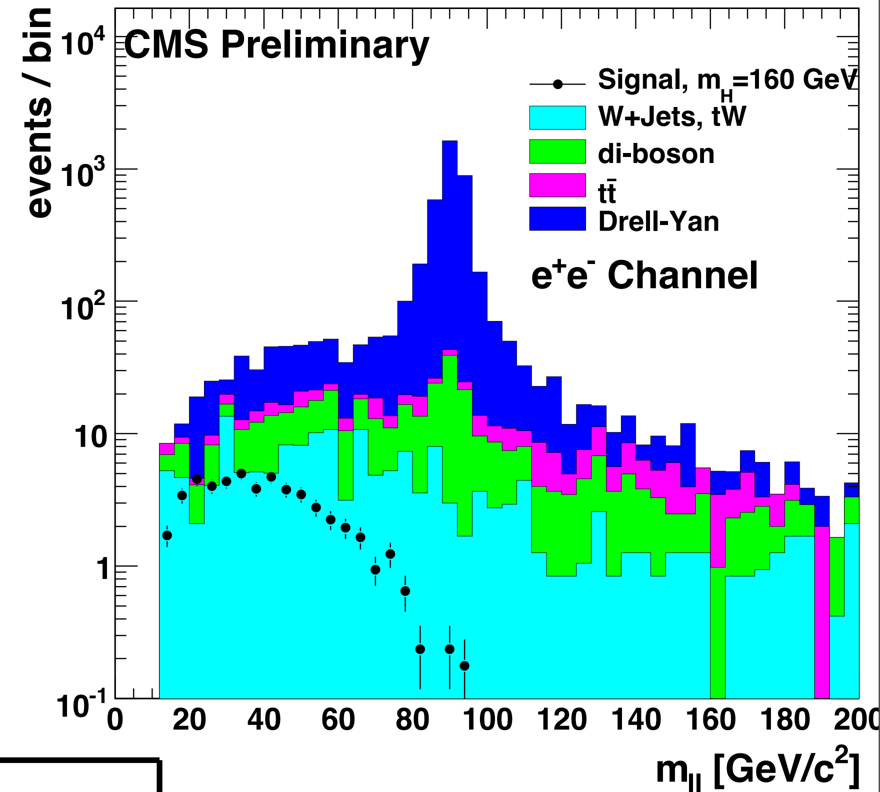
  - Where does $\pi(b)$ come from?
    - did you realize that this is a Bayesian procedure that depends on some prior assumption about what *b* is?

The Data-driven narrative

Center for Cosmology and Particle Physics

Regions in the data with negligible signal expected are used as control samples

- simulated events are used to estimate extrapolation coefficients
- extrapolation coefficients may have theoretical and experimental uncertainties



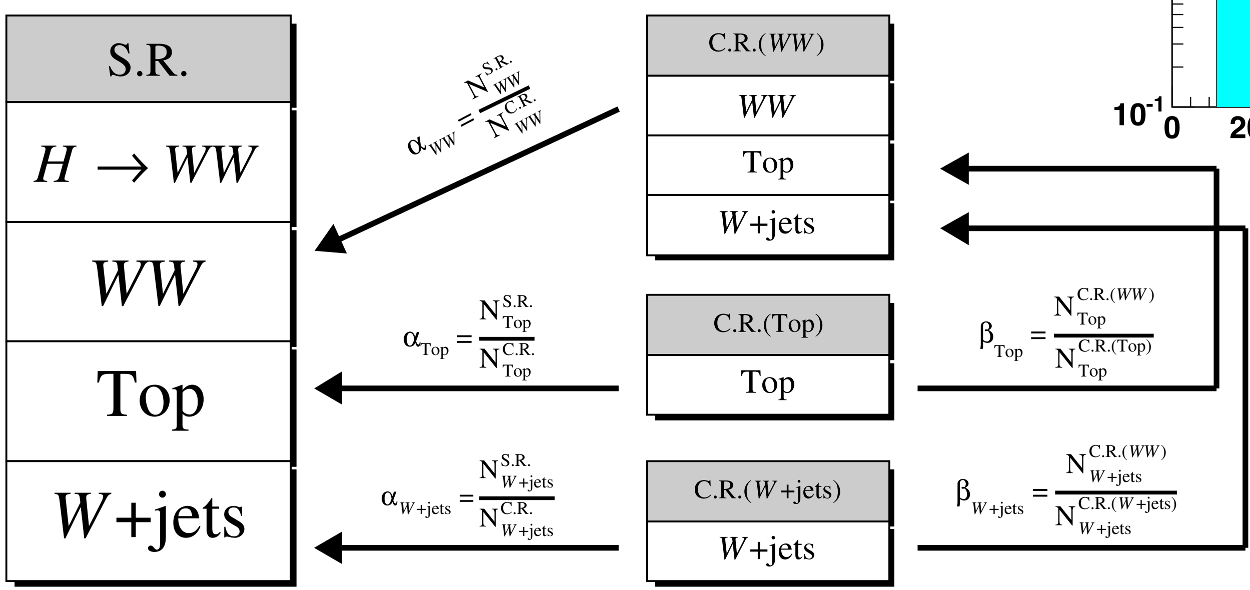


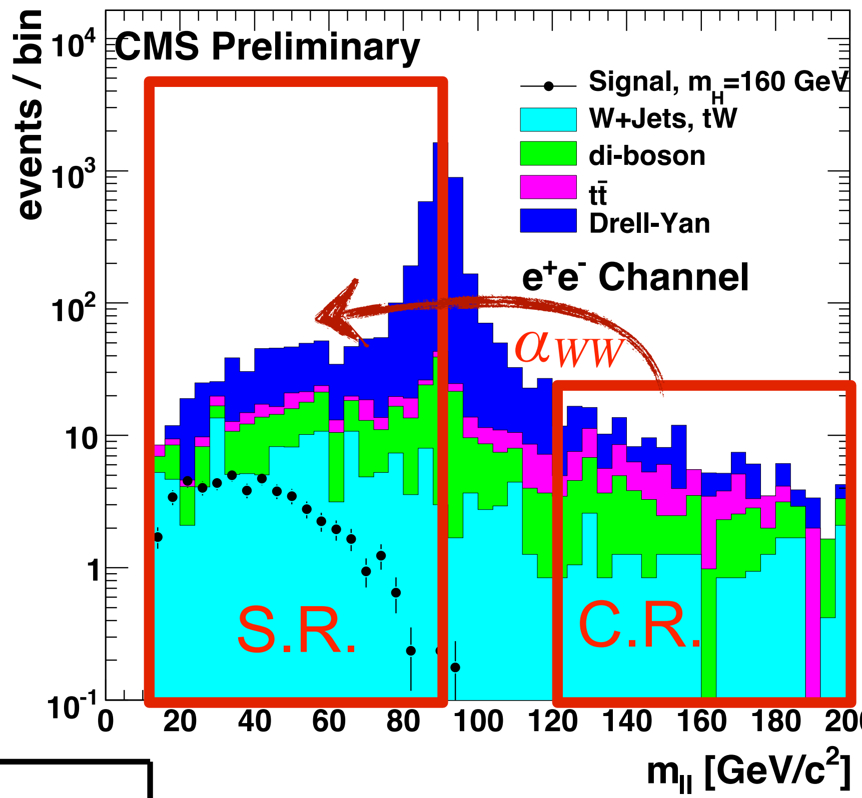
Figure 10: Flow chart describing the four data samples used in the $H \to WW^{(*)} \to \ell\nu\ell\nu$ analysis. S.R and C.R. stand for signal and control regions, respectively.

The Data-driven narrative

Center for Cosmology and Particle Physics

Regions in the data with negligible signal expected are used as control samples

- simulated events are used to estimate extrapolation coefficients
- extrapolation coefficients may have theoretical and experimental uncertainties



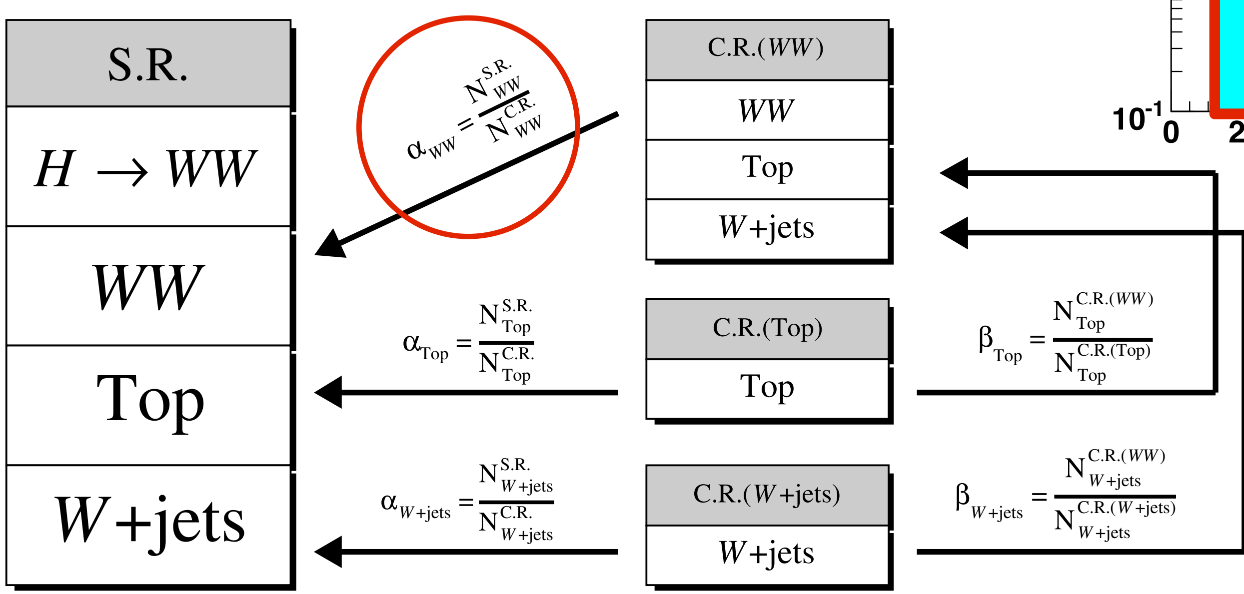


Figure 10: Flow chart describing the four data samples used in the $H \to WW^{(*)} \to \ell\nu\ell\nu$ analysis. S.R and C.R. stand for signal and control regions, respectively.

The “on/off” problem

Now let’s say that the background was estimated from some control region or sideband measurement.

- We can treat these two measurements simultaneously:
  - main measurement: observe n_{on} with $s+b$ expected
  - sideband measurement: observe n_{off} with τb expected

$$\underbrace{P(n_{\rm on}, n_{\rm off}|s,b)}_{\text{joint model}} = \underbrace{\text{Pois}(n_{\rm on}|s+b)}_{\text{main measurement}}\ \underbrace{\text{Pois}(n_{\rm off}|\tau b)}_{\text{sideband}}$$

  - In this approach “background uncertainty” is a statistical error
  - justification and accounting of background uncertainty is much more clear

How does this relate to the smearing approach?

$$P(n_{\rm on}|s) = \int db\, \text{Pois}(n_{\rm on}|s+b)\, \pi(b),$$

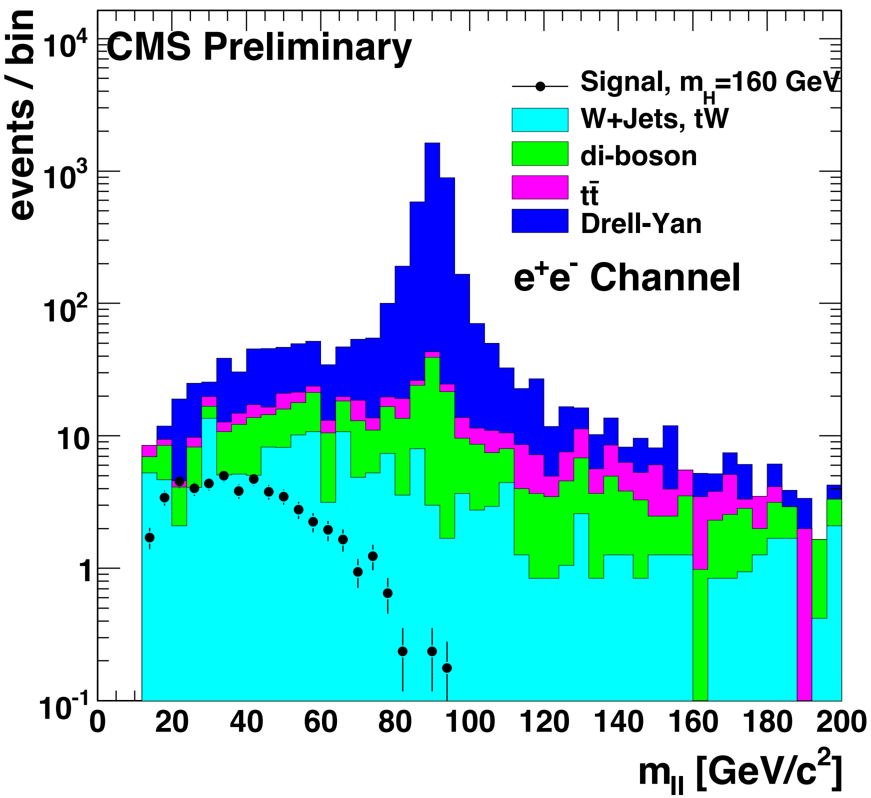
- while $\pi(b)$ is based on data, it still depends on a prior $\eta(b)$

$$\pi(b) = P(b|n_{\rm off}) = \frac{P(n_{\rm off}|b)\eta(b)}{\int db P(n_{\rm off}|b)\eta(b)}.$$

Going beyond on/off

Often the extrapolation parameter has uncertainty

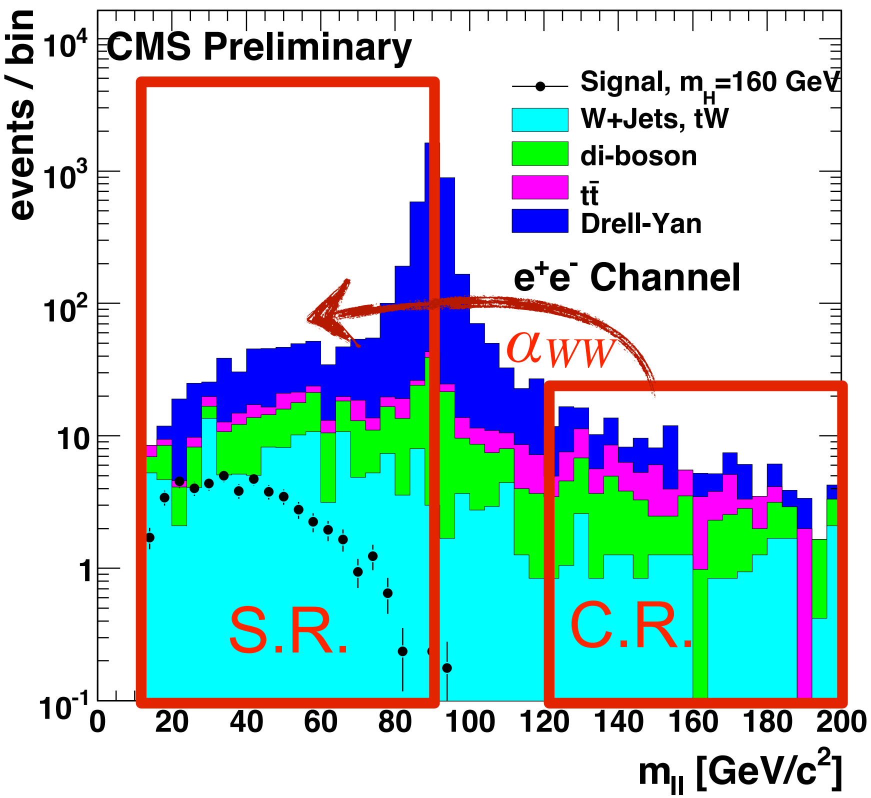
- introduce a new measurement to constrain it as in the ABCD method
- what if..., what if ..., what if..., what if ..., what if..., what if ...



Going beyond on/off

Often the extrapolation parameter has uncertainty

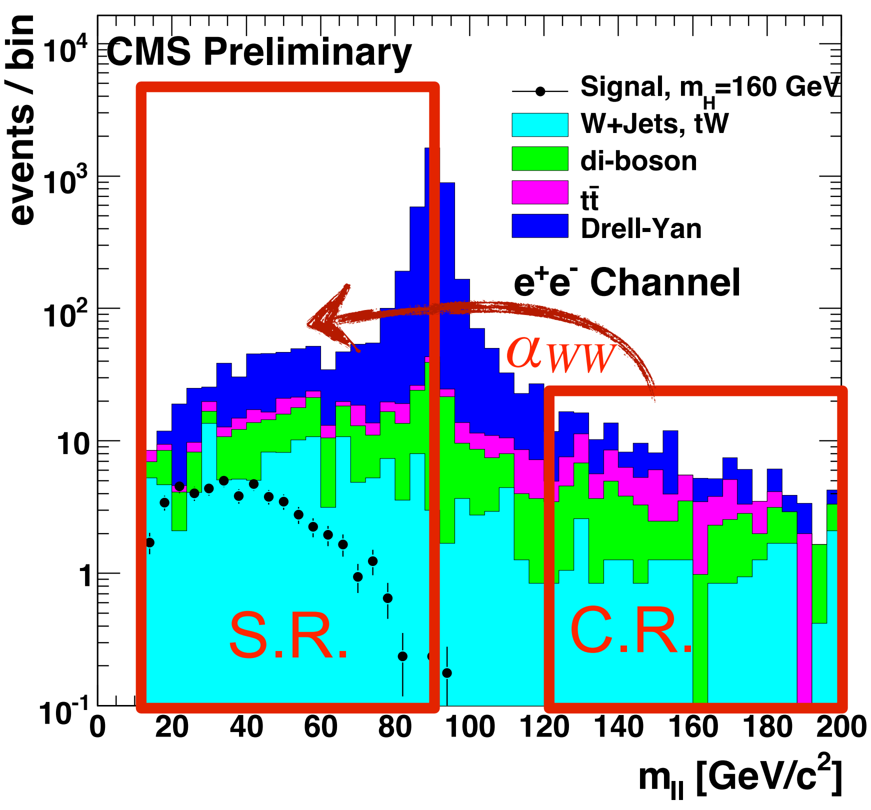
- introduce a new measurement to constrain it as in the ABCD method
- what if..., what if ..., what if..., what if ..., what if..., what if ...



Going beyond on/off

Often the extrapolation parameter has uncertainty

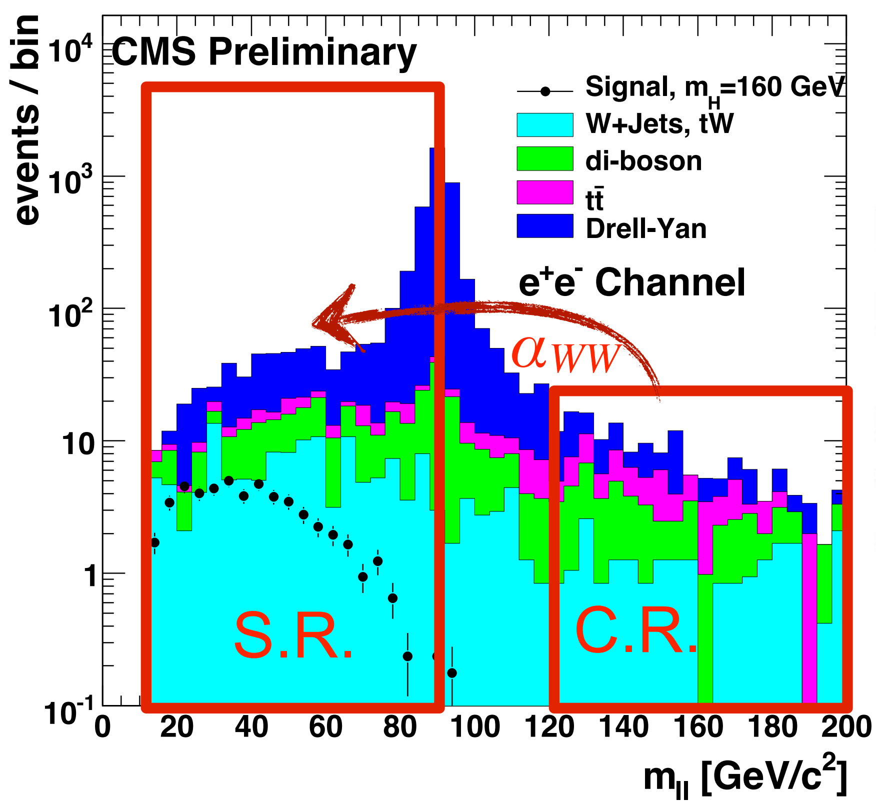
- introduce a new measurement to constrain it as in the ABCD method

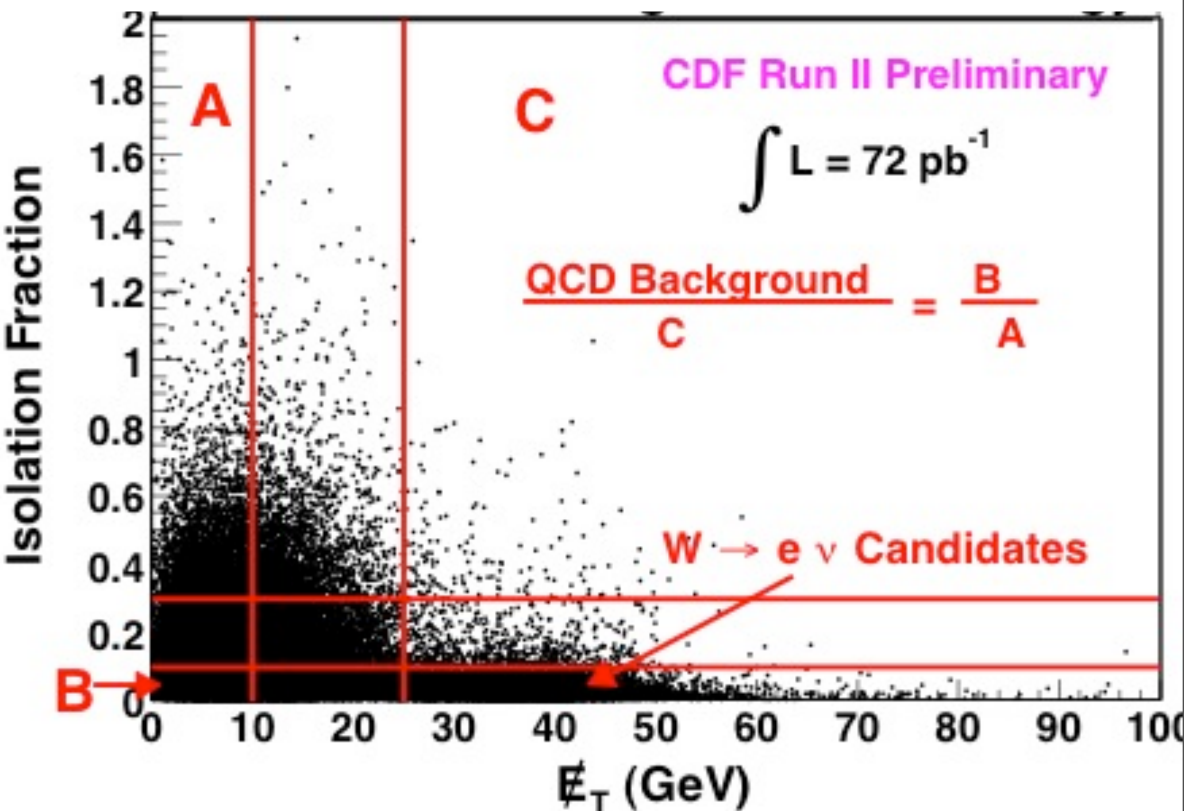


Going beyond on/off

Often the extrapolation parameter has uncertainty

- introduce a new measurement to constrain it as in the ABCD method

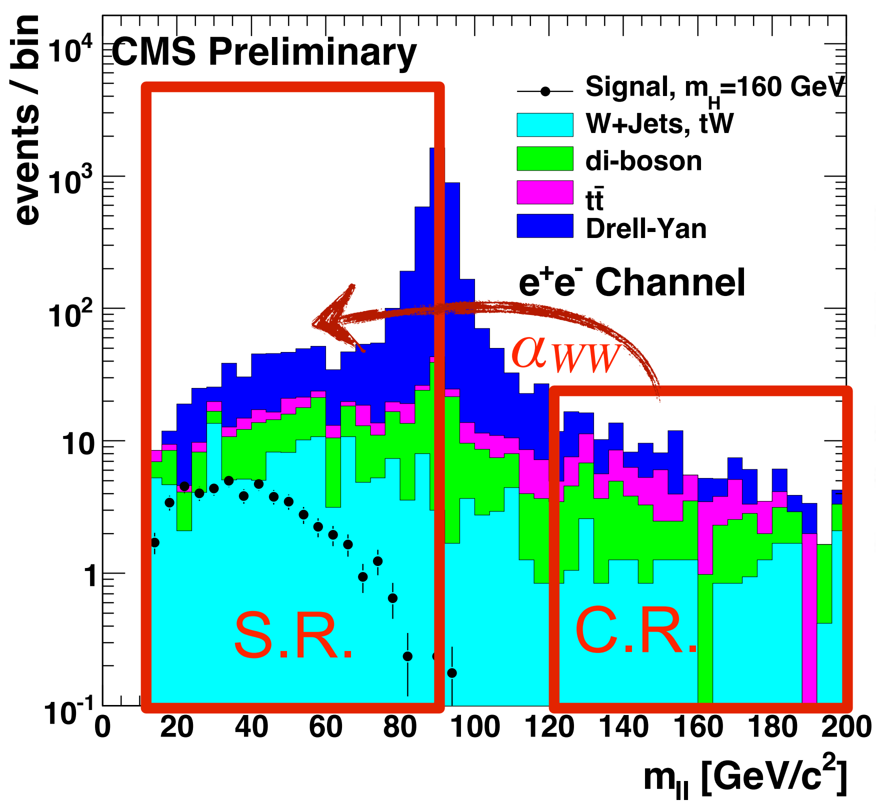


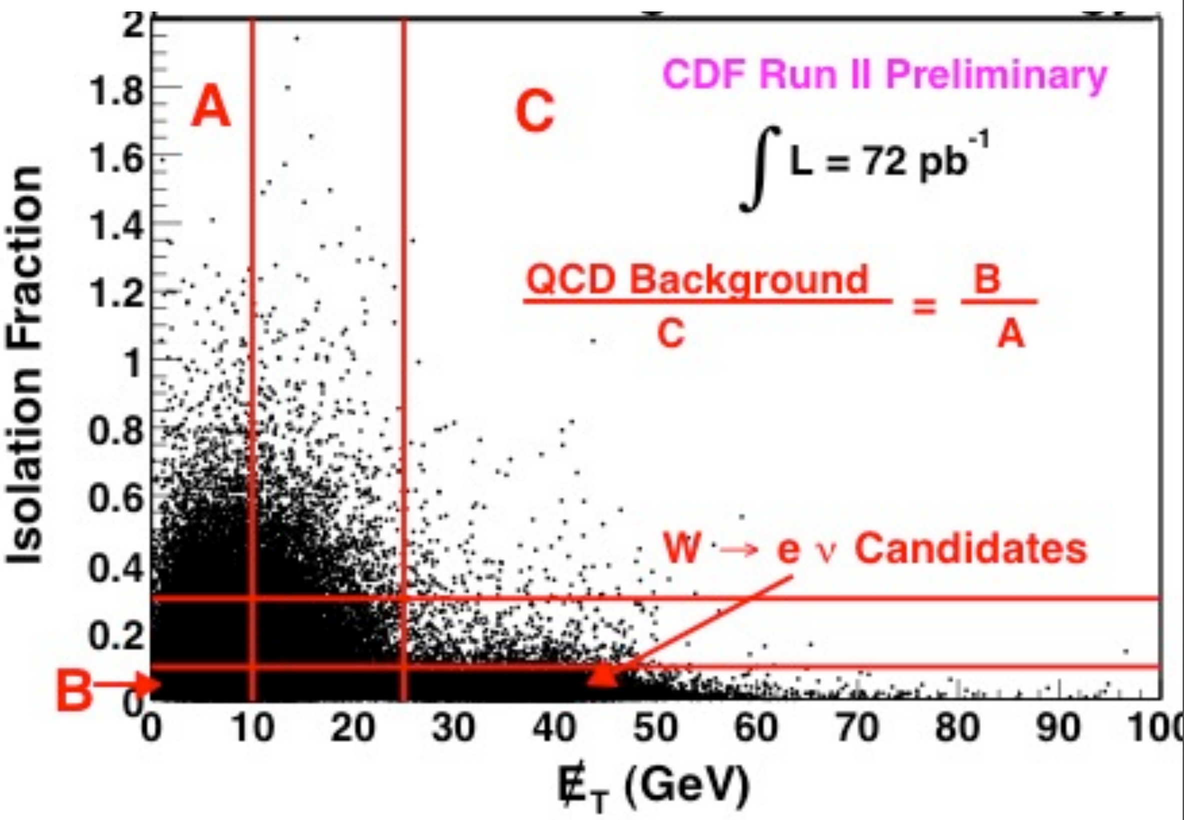


Going beyond on/off

Often the extrapolation parameter has uncertainty

- introduce a new measurement to constrain it as in the ABCD method
- what if..., what if ..., what if..., what if ..., what if..., what if ...





Going beyond on/off

Center for Cosmology and Particle Physics

Often the extrapolation parameter has uncertainty

- introduce a new measurement to constrain it as in the ABCD method
- what if..., what if ..., what if..., what if ..., what if..., what if ...

Classification of Systematic Uncertainties

Taken from Pekka Sinervo's PhyStat 2003 contribution

Type I - "The Good"

- can be constrained by other sideband/auxiliary/ancillary measurements and can be treated as statistical uncertainties
  - scale with luminosity

Classification of Systematic Uncertainties

Taken from Pekka Sinervo's PhyStat 2003 contribution

Type I - "The Good"

- can be constrained by other sideband/auxiliary/ ancillary measurements and can be treated as statistical uncertainties
  - scale with luminosity

Type II - "The Bad"

- arise from model assumptions in the measurement or from poorly understood features in data or analysis technique
  - don't necessarily scale with luminosity
  - eg: "shape" systematics

Classification of Systematic Uncertainties

Center for Cosmology and Particle Physics

Taken from Pekka Sinervo's PhyStat 2003 contribution

Type I - "The Good"

- can be constrained by other sideband/auxiliary/ ancillary measurements and can be treated as statistical uncertainties
  - scale with luminosity

Type II - "The Bad"

- arise from model assumptions in the measurement or from poorly understood features in data or analysis technique
  - don't necessarily scale with luminosity
  - eg: "shape" systematics

Type III - "The Ugly"

- arise from uncertainties in underlying theoretical paradigm used to make inference using the data
  - a somewhat philosophical issue

Separating the prior from the objective model

Recommendation: where possible, one should express uncertainty on a parameter as a statistical (random) process

- explicitly include terms that represent auxiliary measurements in the likelihood

Recommendation: when using a Bayesian technique, one should explicitly express and separate the prior from the objective part of the probability density function

Example:

- By writing $P(n_{\rm on}, n_{\rm off}|s,b) = {\rm Pois}(n_{\rm on}|s+b)\,{\rm Pois}(n_{\rm off}|\tau b).$
  - the objective statistical model is for the background uncertainty is clear
- One can then explicitly express a prior $\eta(b)$ and obtain:

$$\pi(b) = P(b|n_{\rm off}) = \frac{P(n_{\rm off}|b)\eta(b)}{\int db P(n_{\rm off}|b)\eta(b)}.$$

Constraints on Nuisance Parameters

Center for Cosmology and Particle Physics

Many uncertainties have no clear statistical description or it is impractical to provide

Traditionally, we use Gaussians, but for large uncertainties it is clearly a bad choice

- quickly falling tail, bad behavior near physical boundary, optimistic p-values, ...

For systematics constrained from control samples and dominated by statistical uncertainty, a Gamma distribution is a more natural choice [PDF is Poisson for the control sample]

- longer tail, good behavior near boundary, natural choice if auxiliary is based on counting

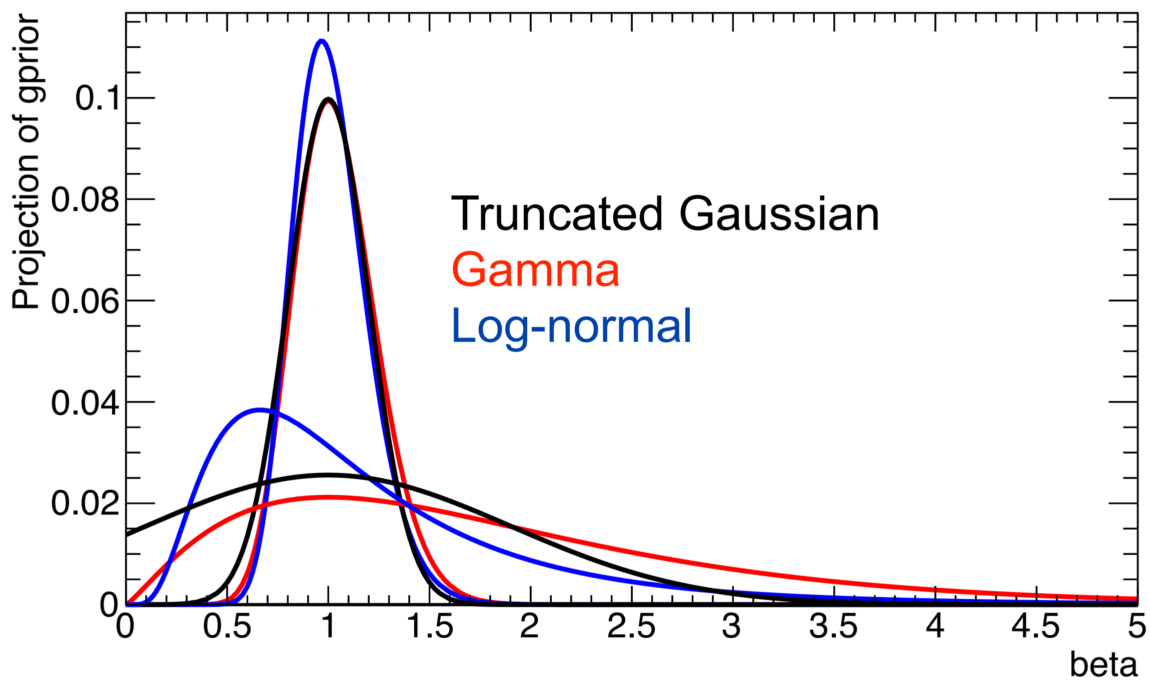
For "factor of 2" notions of uncertainty log-normal is a good choice

- can have a very long tail for large uncertainties

None of them are as good as an actual model for the auxiliary measurement, if available

To consistently switch between frequentist, Bayesian, and hybrid procedures, need to be clear about prior vs. likelihood function

PDF	Prior	Posterior
Gaussian	uniform	Gaussian
Poisson	uniform	Gamma
Log-normal	reference	Log-Normal



Building the model: HistFactory (RooStats)

Several analyses have used the tool called **hist2workspace** to build the model (PDF)

- command line: **hist2workspace myAnalysis.xml**
- construct likelihood function below via XML + histograms

$$\mathcal{L}(\mu, \alpha_i) = \prod_{m \in \text{bins}} \text{Pois}(n_m | \nu_m) \prod_{i = \in \text{Syst}} N(\alpha_i)$$

$$\nu_m = \mu L \eta_1(\alpha)\, \sigma_{1m}(\alpha) + \sum_{j \in \text{Bkg Samp}} L \eta_j(\alpha)\, \sigma_{jm}(\alpha),$$

interpolation convention

$$\eta_j(\alpha) = \prod_{i \in \text{Syst}} I(\alpha_i; \eta_{ij}^+, \eta_{ij}^-)$$

$$\sigma_{jm}(\alpha) = \sigma_{jm}^0 \prod_{i \in \text{Syst}} I(\alpha_i; \sigma_{ijm}^+/\sigma_{jm}^0, \sigma_{ijm}^-/\sigma_{jm}^0)$$

$$I(\alpha; I^+, I^-) = \begin{cases} 1 + \alpha(I^+ - 1) & \text{if } \alpha > 0 \\ 1 & \text{if } \alpha = 0 \\ 1 - \alpha(I^- - 1) & \text{if } \alpha < 0 \end{cases}$$

```
<!DOCTYPE Channel  SYSTEM 'Config.dtd'>

  <Channel Name="channel1" InputFile="./data/example.root" HistoName="" >
    <!--<Data Name="data" InputFile="" HistoPath="" HistoName=""/>-->
    <Sample Name="signal" HistoPath="" HistoName="signal">
      <OverallSys Name="syst1" High="1.05" Low="0.95"/>
      <NormFactor Name="SigXsecOverSM" Val="1" Low="0.5" High="1.8" Const="True" />
    </Sample>
    <Sample Name="background1" HistoPath="" NormalizeByTheory="True" HistoName="background1">
      <OverallSys Name="syst2" Low="0.95" High="1.05"/>
    </Sample>
    <Sample Name="background2" HistoPath="" NormalizeByTheory="True" HistoName="background2">
      <OverallSys Name="syst3" Low="0.95" High="1.05"/>
      <!-- <HistoSys Name="syst4" HistoPathHigh="" HistoPathLow="histForSyst4"/>-->
    </Sample>
  </Channel>
```

Constraint terms

Center for Cosmology and Particle Physics

For each systematic effect, we associated a nuisance parameter α

- for instance electron efficiency, JES, luminosity, etc.
- the background rates, signal acceptance, etc. are parametrized in terms of these nuisance parameters

These systematics are usually known ("constrained") within ± 1σ.

- but here we must be careful about Bayesian vs. frequentist
- Why is it constrained? Usually b/c we have an **auxiliary measurement** *m* and a relationship like:

$$G(m|\alpha, \sigma)$$

  - Saying that α has a Gaussian distribution is Bayesian.
    - has form "Probability of parameter"
  - The frequentist way is to say that that *m* fluctuates about α

While *m* is a measured quantity (or "observable"), there is only one measurement of *m* per experiment. Call it a "**Global observable**"

An example ModelConfig from HistFactory

The RooStats tools, use the RooFit PDF interface, but the tools need some additional meta information. The **ModelConfig** class encapsulates this meta information

- The PDF itself, the observables, the "global observables", the parameter of interest, and the nuisance parameters. Also the prior for Bayesian methods.

root [7] modelConfig->Print()
=== Using the following for ModelConfig ===
Observables: RooArgSet:: = (obs_h2e2nu_200)

Parameters of Interest: RooArgSet:: = (SigXsecOverSM)

Nuisance Parameters: RooArgSet:: = (Lumi,alpha_SysBtagEff,alpha_SysElecScale,alpha_SysElecSmear,alpha_SysJetScale,alpha_SysJetSmear,alpha_SysMETHadScale,alpha_SysMETHadSmear,alpha_SysMuonScale,alpha_SysMuonSmear,alpha_dieleceff,alpha_mjet2enorm,alpha_signorm,alpha_topnorm,alpha_wnorm,alpha_wwnorm,alpha_wznorm,alpha_znorm,alpha_zznorm)

Global Observables: RooArgSet:: = (nominalLumi,nom_alpha_dieleceff,nom_alpha_signorm,nom_SysMuonScale,nom_SysMETHadSmear,nom_SysElecSmear,nom_SysMuonSmear,nom_SysJetSmear,nom_SysBtagEff,nom_SysJetScale,nom_SysMETHadScale,nom_SysElecScale,nom_alpha_topnorm,nom_alpha_wwnorm,nom_alpha_wznorm,nom_alpha_zznorm,nom_alpha_wnorm,nom_alpha_znorm,nom_alpha_mjet2enorm)

PDF: RooProdPdf::model_h2e2nu_200[lumiConstraint * alpha_dieleceffConstraint * alpha_signormConstraint * alpha_SysMuonScaleConstraint * alpha_SysMETHadSmearConstraint * alpha_SysElecSmearConstraint * alpha_SysMuonSmearConstraint * alpha_SysJetSmearConstraint * alpha_SysBtagEffConstraint * alpha_SysJetScaleConstraint * alpha_SysMETHadScaleConstraint * alpha_SysElecScaleConstraint * alpha_topnormConstraint * alpha_wwnormConstraint * alpha_wznormConstraint * alpha_zznormConstraint * alpha_wnormConstraint * alpha_znormConstraint * alpha_mjet2enormConstraint * h2e2nu_200_model] = 0

CMS Higgs example

Center for Cosmology and Particle Physics

The CMS input:

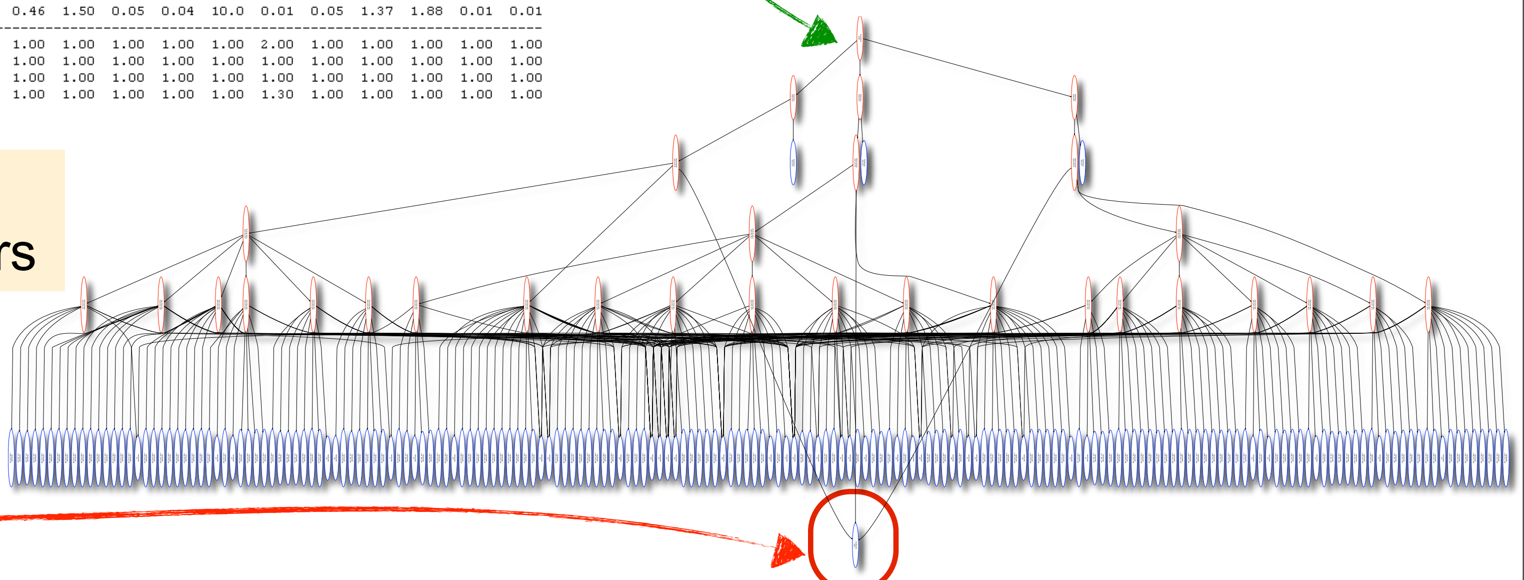
- cleanly tabulated effect on each background due to each source of systematic
- systematics broken down into uncorrelated subsets
- used lognormal distributions for all systematics, Poissons for observations

Started with a txt input, defined a mathematical representation, and then prepared the RooStats workspace

```
Date: June 22, 2010
Description: HWW-->2l2v, 0jets, cut-and-count for 3 channels: mumu, ee, emu; made-up numbers for a ATLAS+CMS combination exercise
mH     160  Higgs mass hypothesis
comE   7.0   center of mass energy
lumi   1  luminosity in fb-1
----------------------------------------------------------------------------------------------------------
imax    3  number of channels
jmax    6  number of backgrounds
kmax   37  number of nuisance parameters
----------------------------------------------------------------------------------------------------------
Observation   15  7  13
----------------------------------------------------------------------------------------------------------
bin          1     1     1     1     1     1     1     2     2     2     2     2     2     2     3     3     3     3     3     3     3
process      H    Wj    Zj    tX    WW    WZ    ZZ     H    Wj    Zj    tX    WW    WZ    ZZ     H    Wj    Zj    tX    WW    WZ    ZZ
process      0     1     2     3     4     5     6     0     1     2     3     4     5     6     0     1     2     3     4     5     6
----------------------------------------------------------------------------------------------------------
rate      10.5  0.01  0.05  0.94  3.39  0.01  0.01  5.39  0.01  0.05  0.46  1.50  0.05  0.04  10.0  0.01  0.05  1.37  1.88  0.01  0.01
----------------------------------------------------------------------------------------------------------
 1  lnN   1.00  2.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  2.00  1.00  1.00  1.00  1.00  1.00
 2  lnN   1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  2.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00
 3  lnN   1.00  1.30  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00
 4  lnN   1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.30  1.00  1.00  1.00  1.00  1.00  1.00  1.30  1.00  1.00  1.00  1.00  1.00
```

$$L_{b+rs} = \prod_i \left(\frac{\left(\sum_{j=0,1,..} \tilde{n}_{ij} \cdot \kappa_{ijk}^{\theta_k} \right)^{N_i}}{N_i!} \cdot \exp\left(-\sum_{j=0,1,..} \tilde{n}_{ij} \cdot \kappa_{ijk}^{\theta_k} \right) \right) \cdot \prod_k f(\theta_k)$$

3 observables and
37 nuisance parameters



$$n = \mu L \epsilon \sigma_{SM}$$

ATLAS Higgs Example

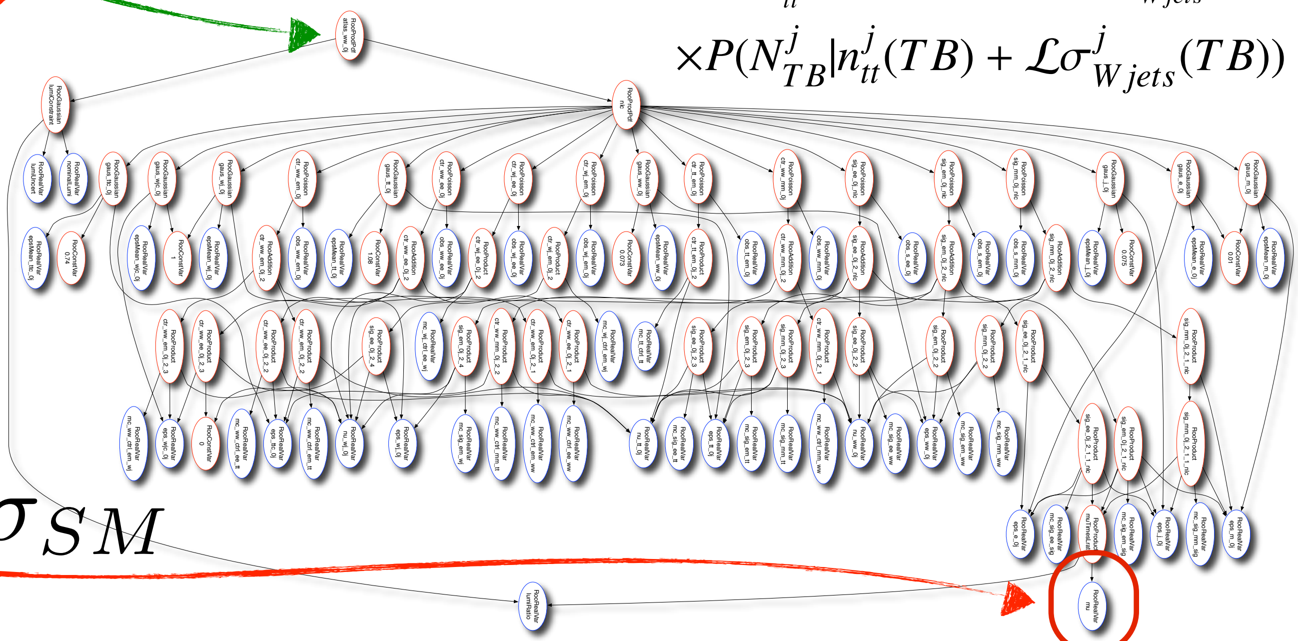
The ATLAS input:

- Poisson terms 3 signal regions and 6 control regions
- Initially uncertainties in extrapolation coefficients treated with one Gaussians and it wasn't possible to identify individual systematics effects
  - thus, unable to identify any correlated systematic (eg. theory uncertainty)
- Now individual uncertainties are explicitly parameterized

$$L^{j,e\mu}_{Pois} = P(N^j_{SR} | n^j_s(SR) + \alpha^j_{WW}\nu_{\alpha^j_{WW}} n^j_{WW}(CR) + \alpha^j_{t\bar{t}}\nu_{\alpha^j_{t\bar{t}}} n^j_{t\bar{t}}(TB) + \alpha^j_{Wjets}\nu_{\alpha^j_{Wjets}} n^j_{Wjets}(LL) + \mathcal{L}\sigma^j_{DY}(SR))$$

$$\times P(N^j_{CR} | n^j_s(CR) + n^j_{WW}(CR) + \beta^j_{t\bar{t}}\nu_{\beta^j_{t\bar{t}}} n^j_{t\bar{t}}(TB) + \beta^j_{Wjets}\nu_{\beta^j_{Wjets}} n^j_{Wjets}(LL) + \mathcal{L}\sigma^j_{DY}(CR))$$

$$\times P(N^j_{TB} | n^j_{tt}(TB) + \mathcal{L}\sigma^j_{Wjets}(TB)) \times P(N^j_{LL} | n^j_{Wjets}(LL))$$

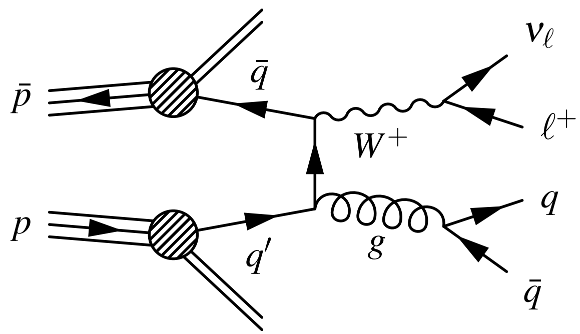


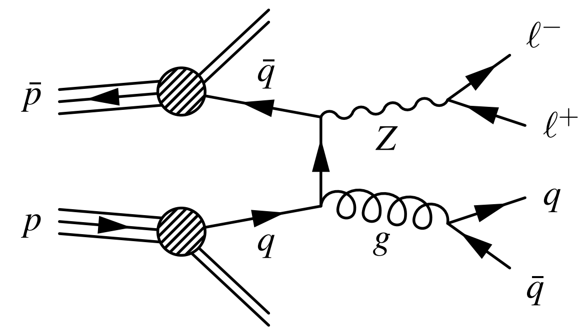
$$n = \mu L \epsilon \sigma_{SM}$$

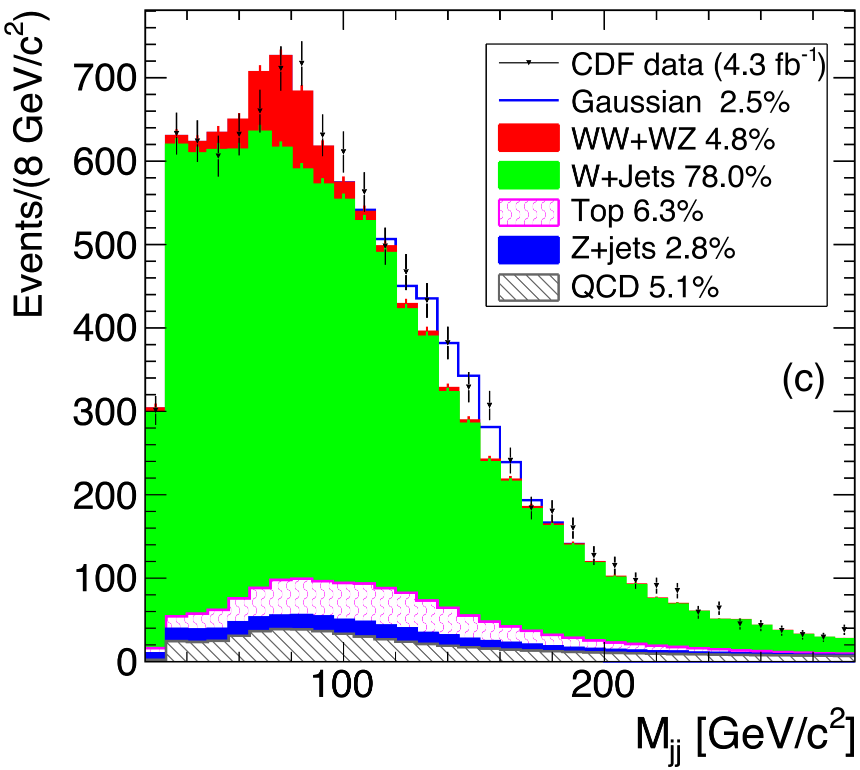
Data driven estimates

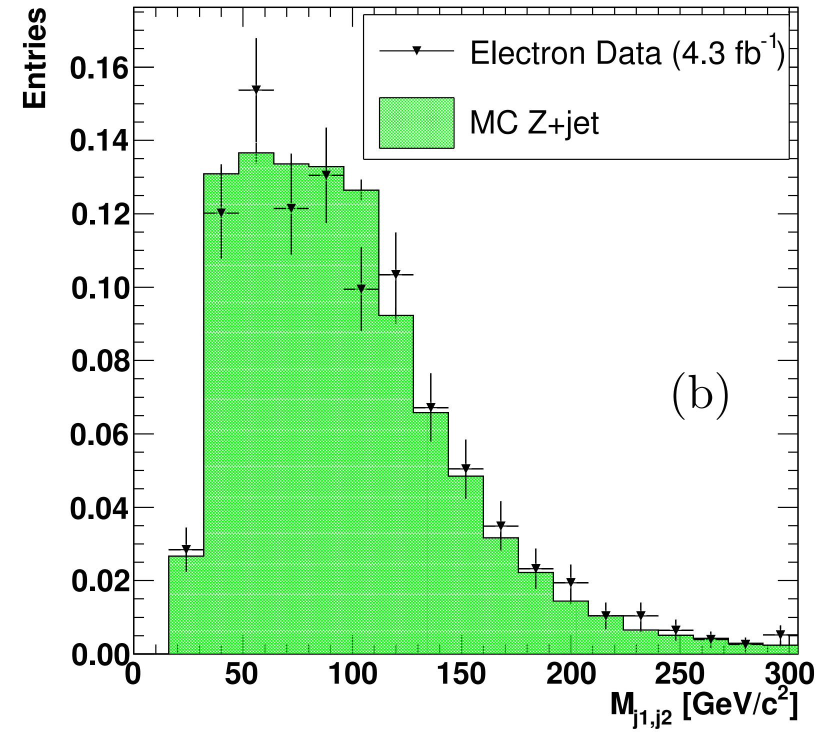
Center for Cosmology and Particle Physics

In the case of the CDF bump, the Z+jets control sample provides a data-driven estimate, but limited statistics. Using the simulation narrative over the data-driven is a **choice**. If you trust that narrative, it's a good choice.





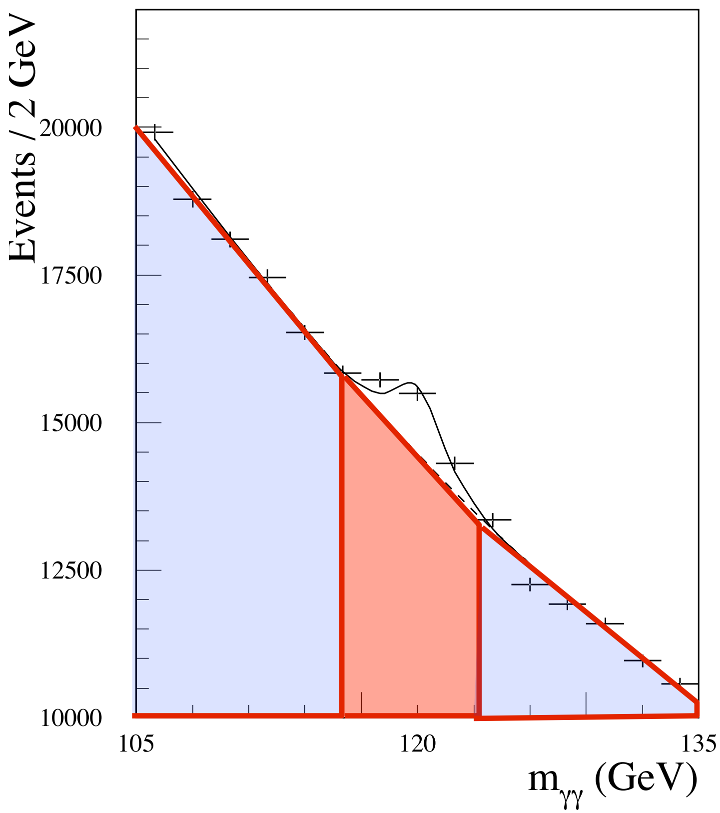




The Effective Model

It is common to describe a distribution with some parametric function

- "fit background to a polynomial", exponential, ...
- While this is convenient and the fit may be good, the narrative is weak



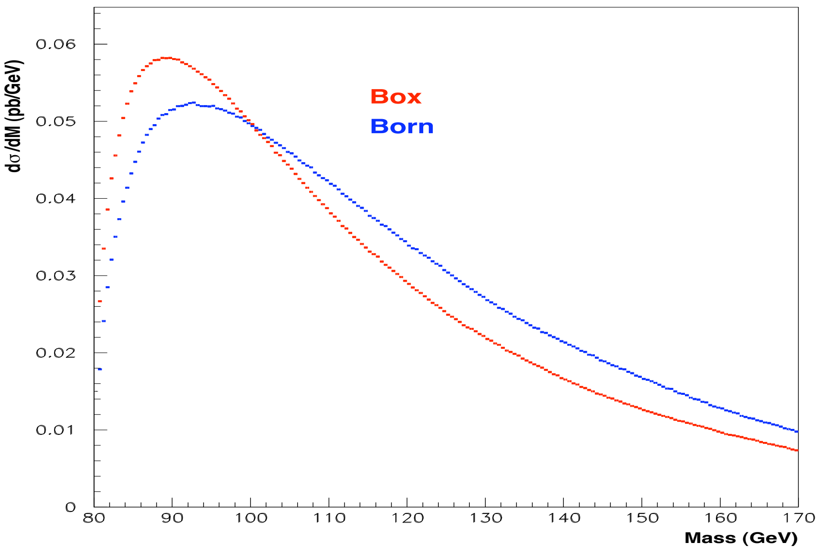
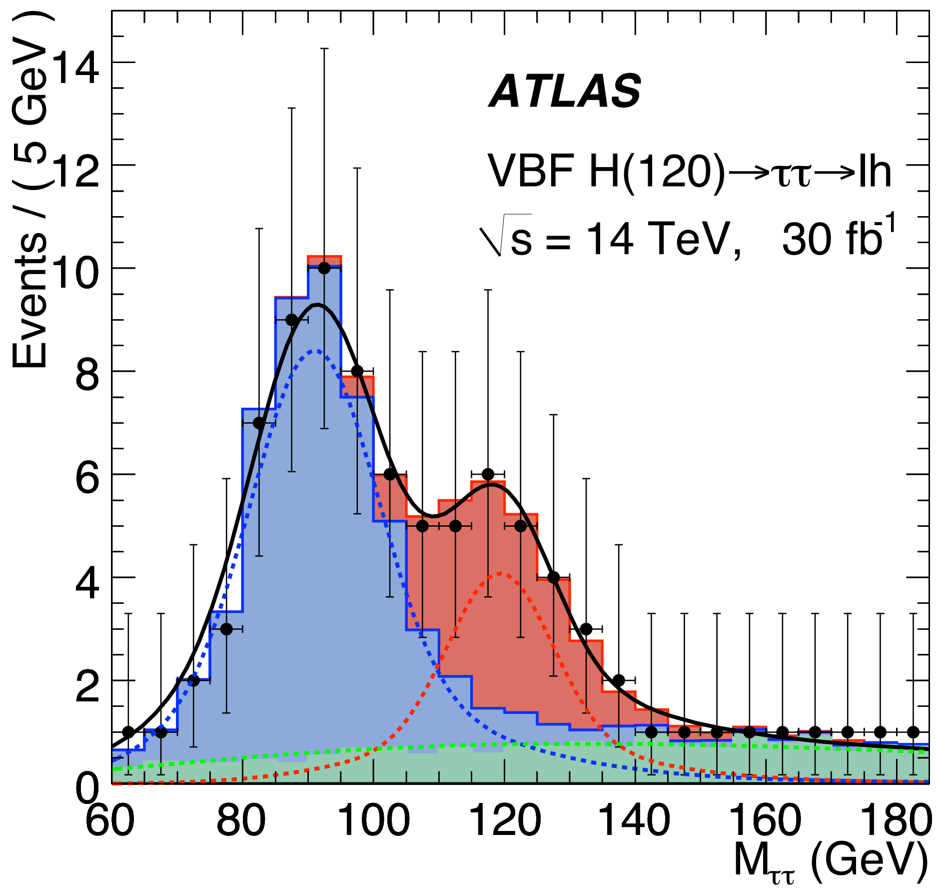


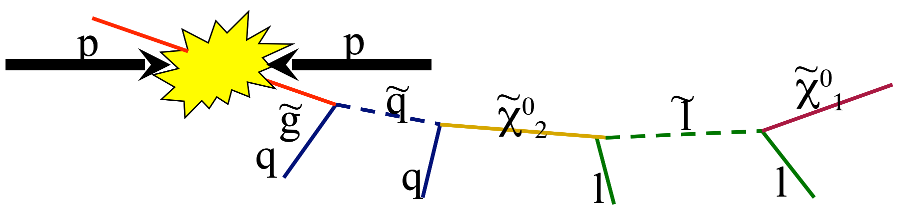
Figure 5. Two plausible shapes for the continuum $\gamma\gamma$ mass spectrum at the LHC.

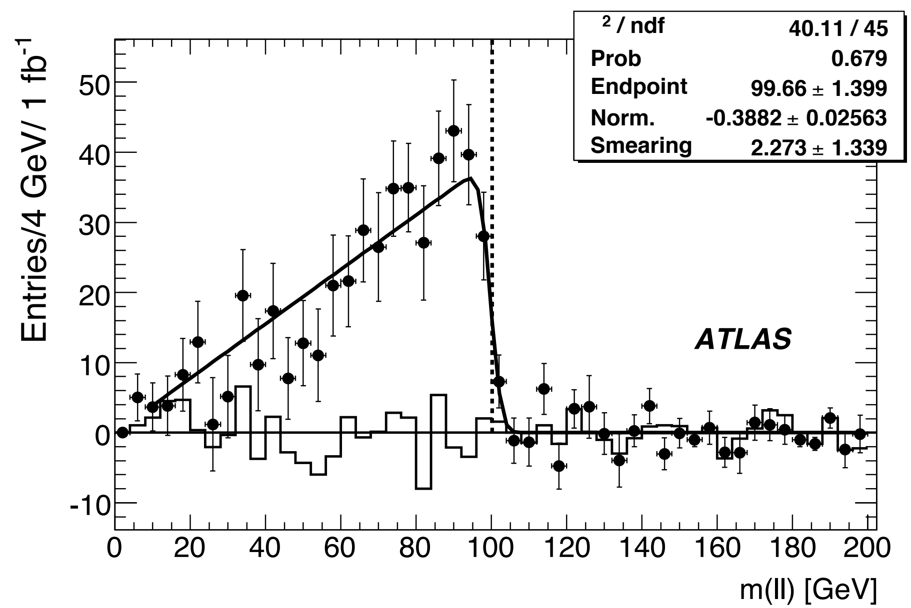
The Effective Model narrative

However, sometimes the effective model comes from a convincing narrative

- convolution of resolution with known distribution
- for example, the "invariant mass" of some final state particles







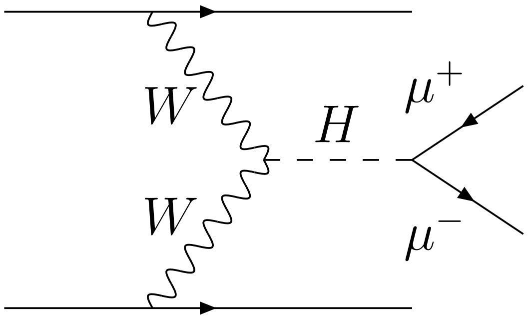
The parametrized response narrative

Center for Cosmology and Particle Physics

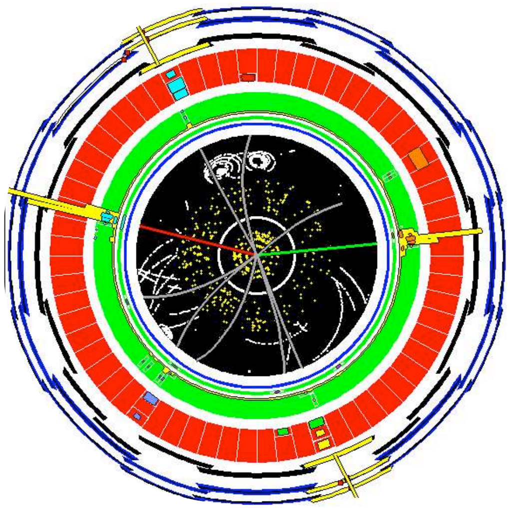
The Matrix-Element technique is conceptually similar to the simulation narrative, but the detector response is parametrized.

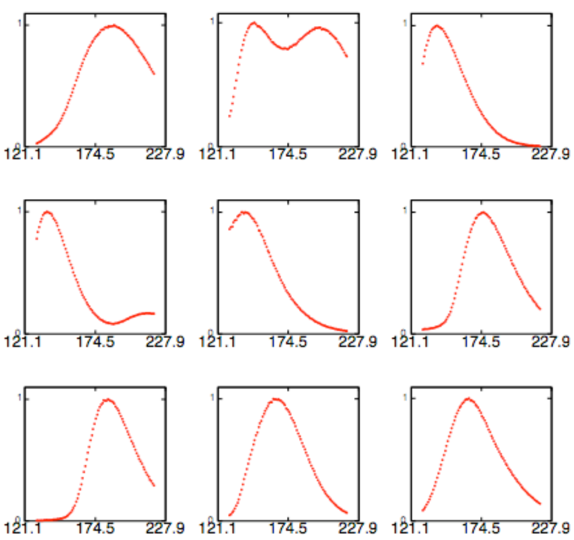
- Doesn't require building parametrized PDF by interpolating between non-parametric templates.

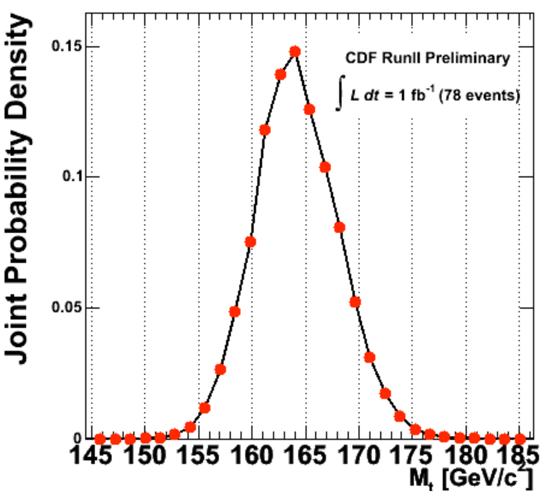
$$L(x|H_0) =$$



$\oplus$







The parametrized response narrative

Center for Cosmology and Particle Physics

The Matrix-Element technique is conceptually similar to the simulation narrative, but the detector response is parametrized.

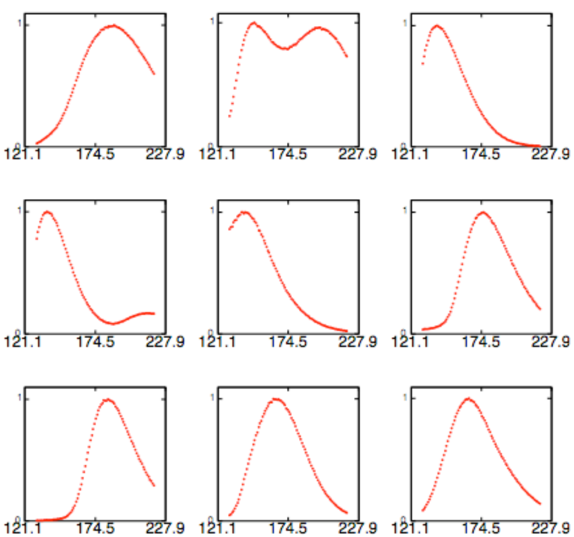
- Doesn't require building parametrized PDF by interpolating between non-parametric templates.

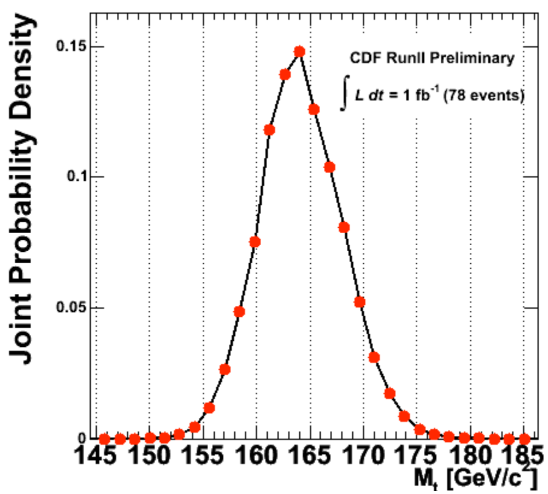
$$P(\mathbf{x}|M_t) = \frac{1}{N}\int d\Phi \; |\mathcal{M}_{t\bar{t}}(p; M_t)|^2 \prod_{jets} f(p_i, j_i) f_{PDF}(q_1) f_{PDF}(q_2)$$

Phase-space Integral

Matrix Element

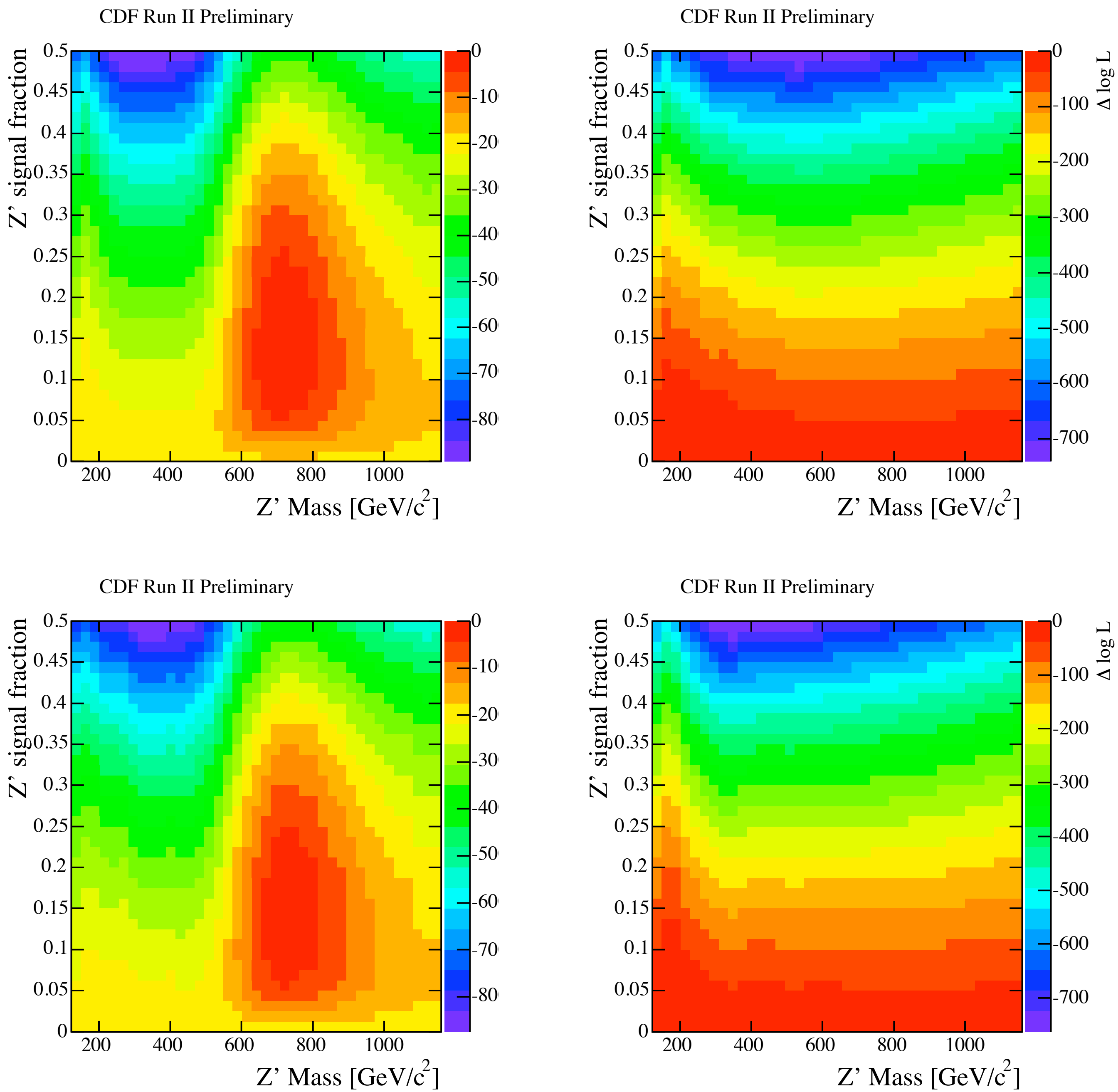
Transfer Functions





Example likelihoods from CDF Z’

Center for
Cosmology and
Particle Physics



CENTER FOR COSMOLOGY AND PARTICLE PHYSICS

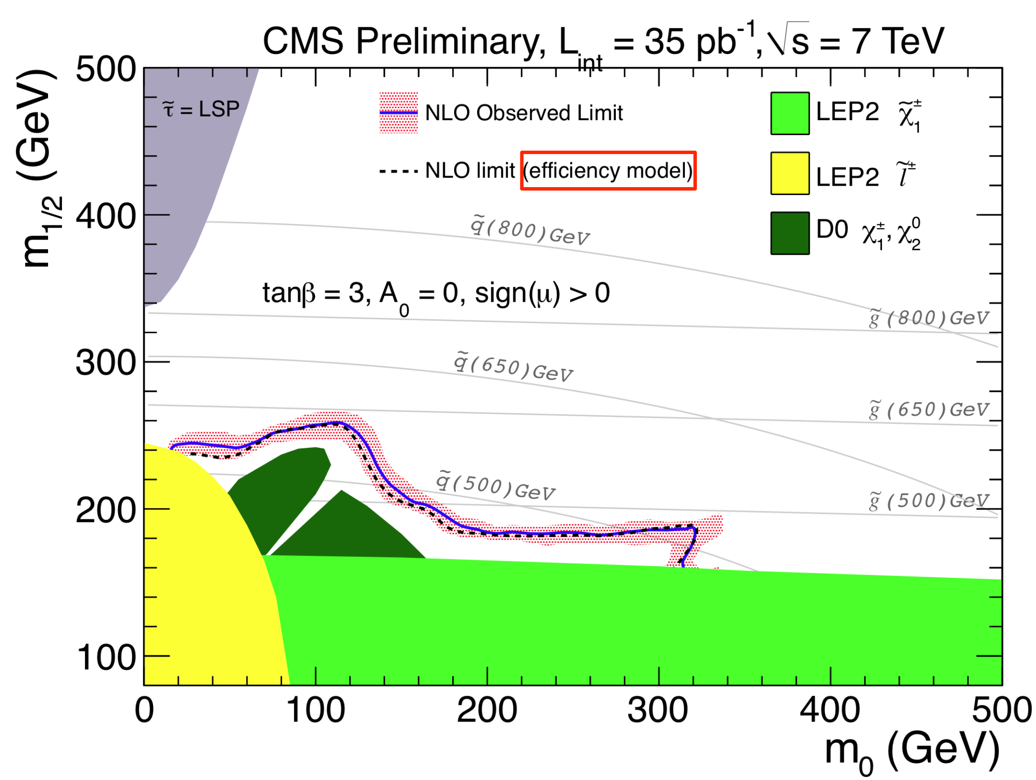
Fast simulations based on parametrized detector response are very useful and can often be tuned to perform quite well in a specific analysis context

- For example: tools like PGS, Delphis, ATLFAST, ...

But these tools still use accept/reject Monte Carlo.

- Would be much more useful if the parmaetrized detector response could be used as a transfer function in Matrix-Element approach

Same sign di-lepton + jets + MET search



Paper includes a simple efficiency model (i.e. for PGS calibrations) and compares full limit to limit with simple model.

CMS SUSY Results, D. Stuart, April 2011, SUSY Recast, UC Davis 36

Narrative styles

The Monte Carlo Simulation narrative (MC narrative)

- each stage is an accept/reject Monte Carlo based on P(out|in) of some microscopic process like parton shower, decay, scattering
- PDFs built from non-parametric estimator like histograms or kernel estimation
  - need to supplement with interpolation procedures to incorporate systematics
  - smearing approach fundamentally Bayesian
- **pros:** most detailed understanding of micro-physics
- **cons:** computationally demanding, loose analytic scaling properties, relies on accuracy of simulation
- **new ideas:** improved interpolation, Radford Neal's machine learning, "design of experiments"

The Data-driven narrative

- independent data sample that either acts as a proxy for some process or can be transformed to do so
- **pros**: nature includes "all orders", uses real detector
- **cons**: extrapolation from control region to signal region requires assumptions, introduces systematic effects. Appropriate transformation may depend on many variables, which becomes impractical

Narrative styles

Effective modeling narrative

- parametrized functional form: eg. Gaussian, falling exponential para polynomial fit to distribution, etc.
- **pros**: fast, has analytic scaling, parametric form may be well justified (eg. phase space, propagation of errors, convolution)
- **cons**: approximate, parametric form may be ad hoc (eg. polynomial from)
- new ideas: using non-parametric statistical methods

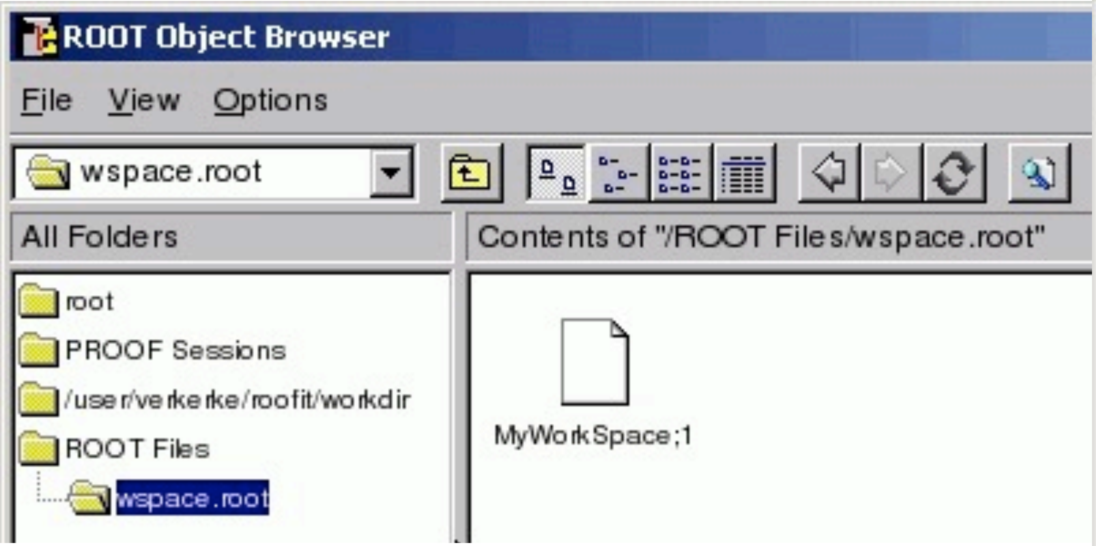
Parametrized detector response narrative (eg. kinematic fitting, Matrix-Element method, ~fast simulation)

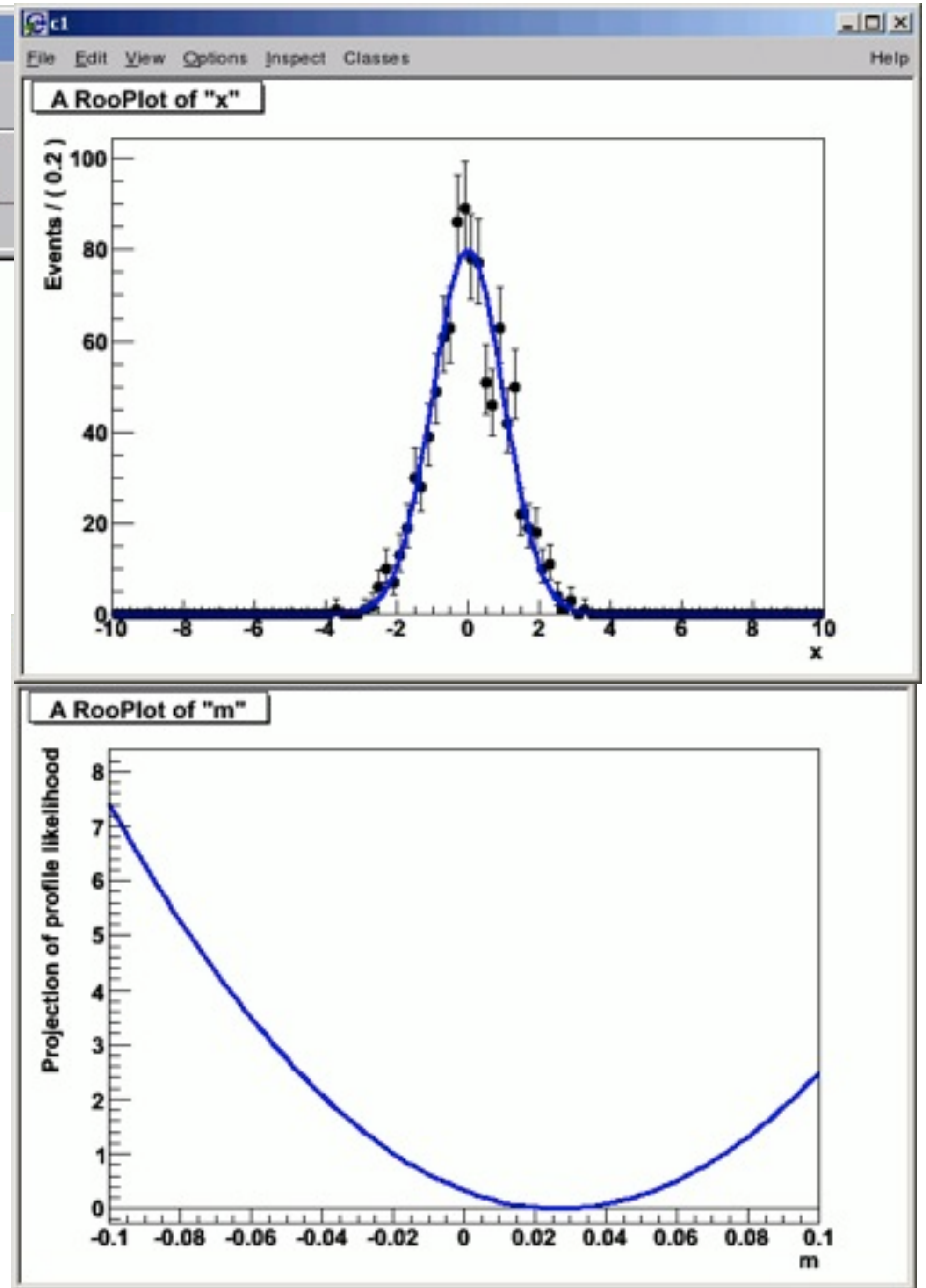
- **pros**: fast, maintains analytic scaling, response usually based on good understanding of the detector, possible to incorporate some types of uncertainty in the response analytically, can evaluate P(out|in) for arbitrary out,in.
- **cons**: approximate, best parametrized detector response is often not available in convenient form
- new ideas: fast simulation is typically parametrized, but we use it in an accept/reject framework (see Geant5)

Combinations, Rich Modeling, and Publishing

Example of Digital Publishing

CENTER FOR COSMOLOGY AND PARTICLE PHYSICS



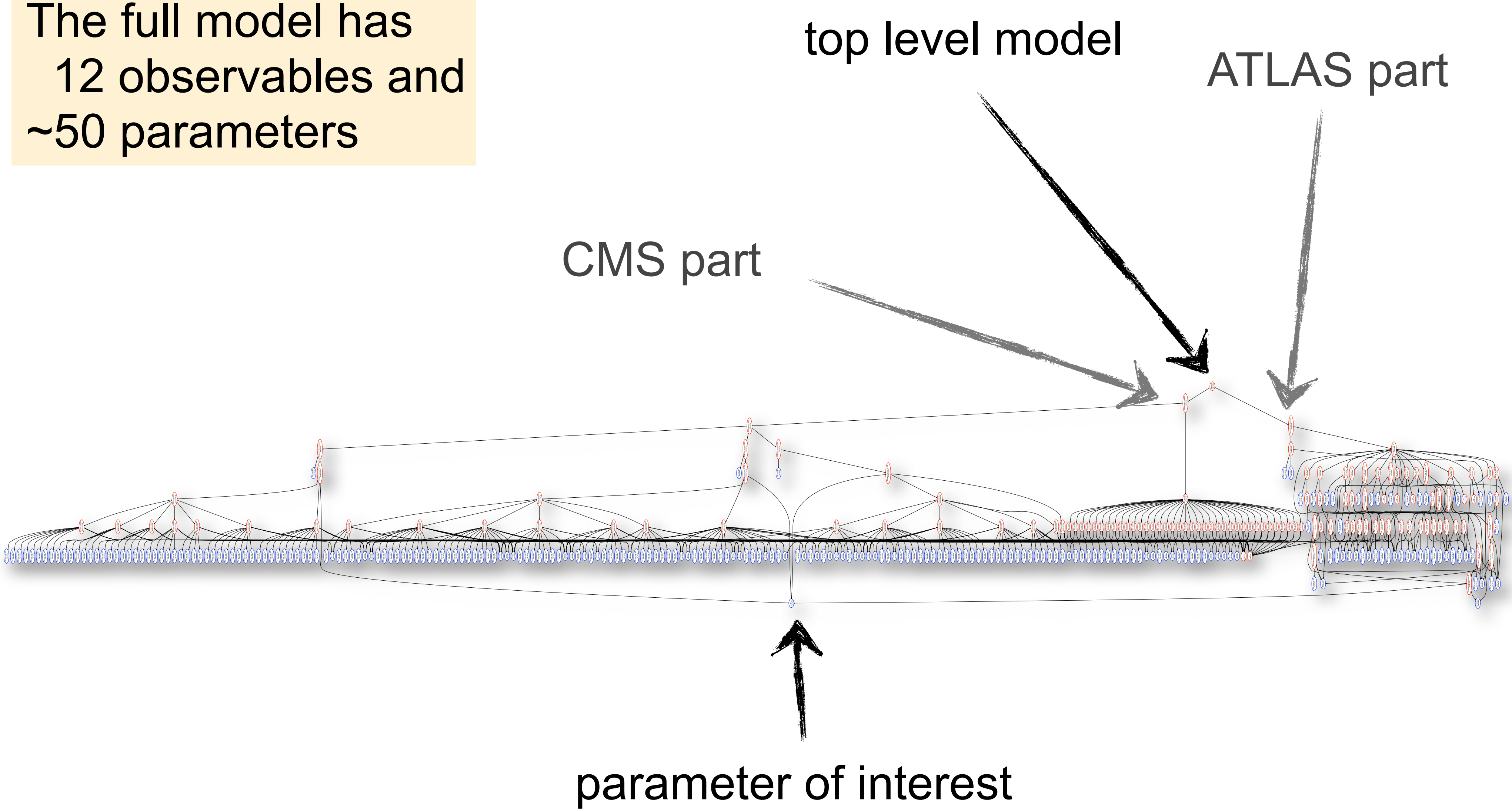


RooFit's Workspace now provides the ability to save in a ROOT file the full likelihood model, any priors you might want, and the minimal data necessary to reproduce likelihood function.

Need this for combinations, as p-value is not sufficient information for a proper combination.

Visualization of the ATLAS+CMS Workspace

The full model has 12 observables and ~50 parameters



$$\mu = \frac{\sigma BR}{\sigma_{SM} BR_{SM}}$$

Combinations & Rich Modeling

As we saw, constraint terms for nuisance parameters can often be related to auxiliary measurements

- we only considered very simple auxiliary measurements, like number of events in a sideband, but even in that case there are likely to be common systematics
- idea can be generalized to more sophisticated measurements
  - for example, γ-jet or Z-jet balance measurements to constrain the Jet Energy Scale uncertainty

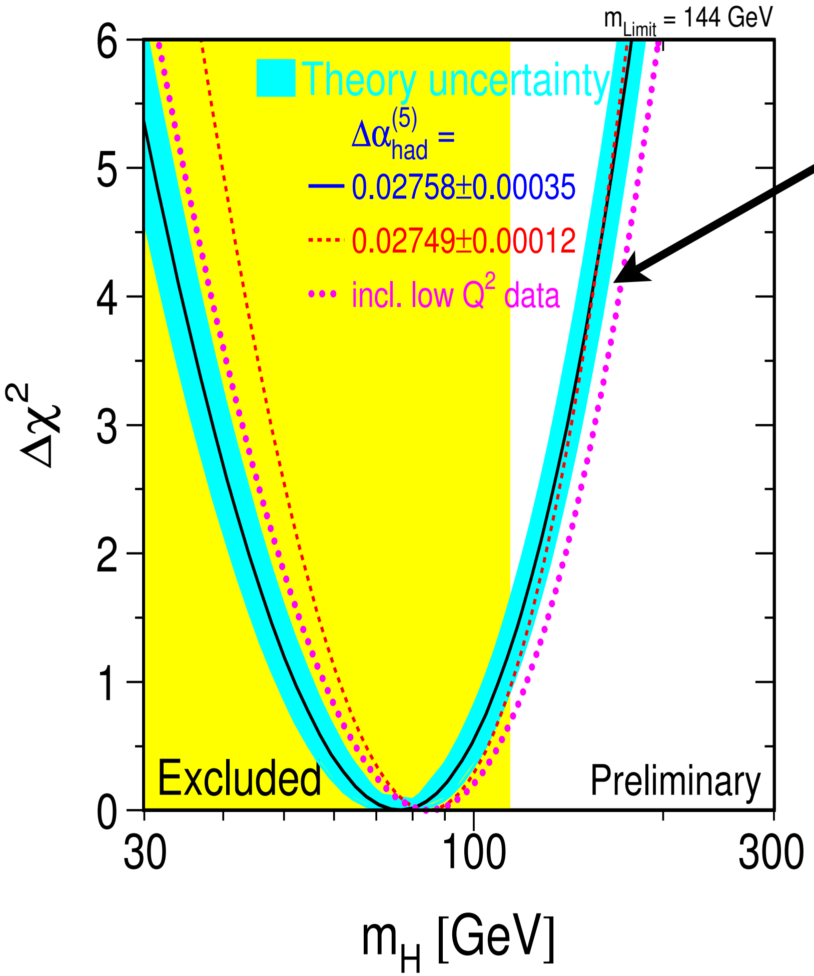
The point is that combining these models leads to a qualitiative change in how we represent what we know: **rich modeling**

Now the distinction has been blurred between a Higgs combination and a sophisticated modeling of systematics

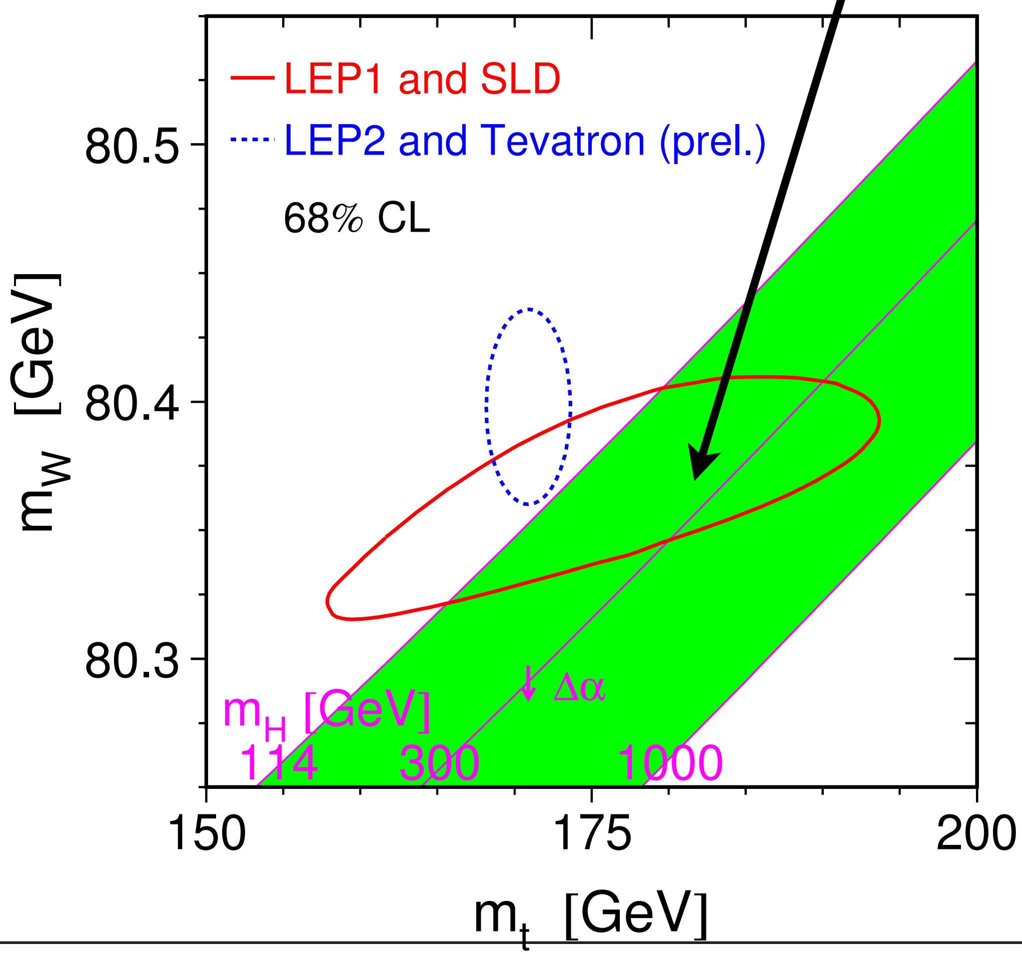
Examples of Published Likelihoods

At previous PhyStats, we agreed to publish likelihood functions

You can find examples of published likelihoods in 1D



In 2-D you just get the contours



Surely we can do better!

Origins I: The First "Statistics in HEP" conference

WORKSHOP ON CONFIDENCE LIMITS

CERN, Geneva, Switzerland
17–18 January 2000

CERN 2000–005

Massimo Corradi

Does everybody agree on this statement, to publish likelihoods?

Louis Lyons

Any disagreement ? Carried unanimously. That's actually quite an achievement for this Workshop.

...[Fred James wants to be able to calculate coverage, Don Groom wants to able to calculate goodness of fit]...

Cousins

I thought the point of unanimity was that publishing the likelihood function was a *necessary* condition, not a sufficient condition.

But a practical problem remained: *How* to communicate multi-D likelihood?

http://indico.cern.ch/conferenceDisplay.py?confId=100458

Current scenario

Taken from the GFitter paper

[23] This procedure only uses the M_H value under consideration, where Higgs-mass hypothesis and measurement are compared. It thus neglects that in the SM a given signal hypothesis entails background hypotheses for all M_H values other than the one considered. An analysis accounting for this should provide a statistical comparison of a given hypothesis with all available measurements. This however would require to know the correlations among all the measurement points (or better: the full experimental likelihood as a function of the Higgs-mass hypothesis), which are not provided by the experiments to date. The difference to the hypothesis-only test employed here is expected to be small at present, but may become important once an experimental Higgs signal appears, which however has insufficient significance yet

Combining Results: An Example

A combination example

- Combining 'ATLAS' and 'CMS' result from persisted workspaces

Read ATLAS workspace

```
TFile* f = new TFile("atlas.root") ;
RooWorkspace *atlas = f->Get("atlas") ;
```

Read CMS workspace

```
TFile* f = new TFile("cms.root") ;
RooWorkspace *cms = f->Get("cms") ;
```

Construct combined LH

```
RooAddition nllCombi("nllCombi","nll CMS&ATLAS",
                     RooArgSet(*cms->function("nll"),*atlas->function("nll"))) ;
```

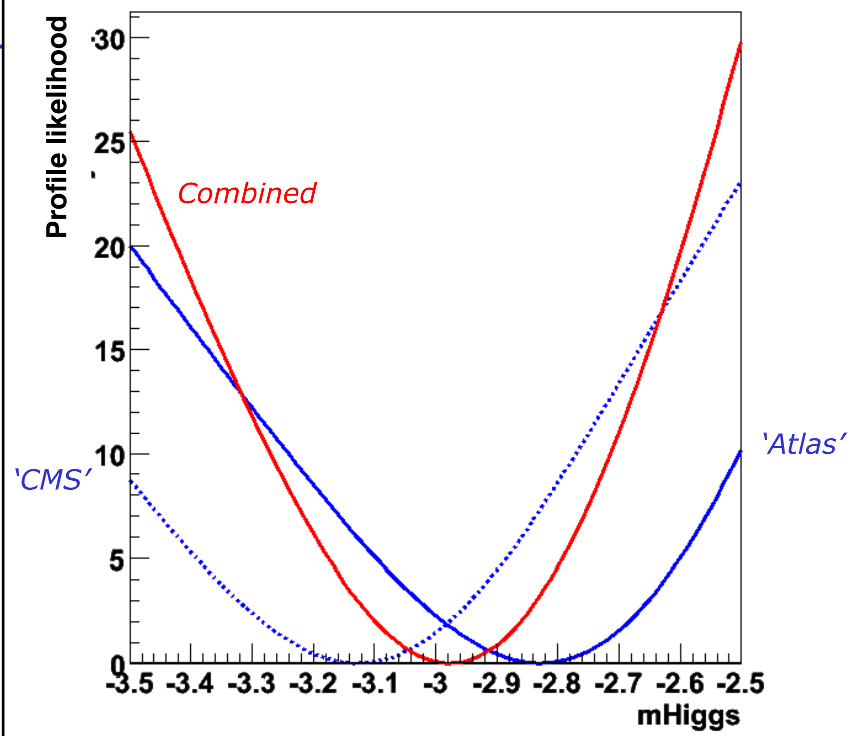
Construct profile LH in mHiggs

```
RooProfileLL pllCombi("pllCombi","pll",nllCombi,*atlas->var("mHiggs")) ;
```

Plot Atlas,CMS, combined profile LH

```
RooPlot* mframe = atlas->var("mHiggs")->frame(-3.5,-2.5) ;
atlas->function("nll")->plotOn(mframe)) ;
cms->function("nll")->plotOn(mframe),LineStyle(kDashed)) ;
pllCombi.plotOn(mframe,LineColor(kRed)) ;
mframe->Draw() ; // result on next slide
```

Wouter Verkerke, NIKHEF



By using the workspace, it is easy to share results, ideal for combinations.

Example above shows opening an 'atlas' and 'cms' workspace, and performing a combined fit to a common parameter with profile likelihood.

Michelangelo's Likelihood Mandate (MLM):

A general assessment of the status and needs of the tools for setting limits on (or fitting) parameters of BSM models, using the multitude of data from searches at the LHC

Two related communities and ongoing discussions

- Characterization & Simplified Models
- Fitting Model Parameters

Introduction

Michelangelo's Likelihood Mandate (MLM):

A general assessment of the status and needs of the tools for setting limits on (or fitting) parameters of BSM models, using the multitude of data from searches at the LHC

Two related communities and ongoing discussions

- Characterization & Simplified Models $\rightarrow$ parametrization
- Fitting Model Parameters $\rightarrow$ interpretation

Michelangelo's Likelihood Mandate (MLM):

A general assessment of the status and needs of the tools for setting limits on (or fitting) parameters of BSM models, using the multitude of data from searches at the LHC

Two related communities and ongoing discussions

- Characterization & Simplified Models → parametrization
- Fitting Model Parameters → interpretation

Potential new tasks

- **Input for the Strategy Group**
  - LPCC and experiments required to produce combined assessment of the 2010-11(-12) findings in Higgs and BSM searches
  - TH community, and other expl communities (e.g. LinCol, SuperB, ...), will use this to assess the implications of LHC data for BSM and future exptl projects

➡ We need to prepare the framework/tools to enable:

  - combination of limits/evidence from ATLAS/CMS(/LHCb)
  - use of the results by the rest of the community (e.g. SUSY-models' fitters)

- This will require coordination with
  - ATLAS-CMS statistics forum
  - Fitters' groups
  - all LHC "search " efforts (Higgs, B decays, exotica of all sorts ....)
  - ...

CENTER FOR COSMOLOGY AND PARTICLE PHYSICS

Michelangelo's Likelihood Mandate (MLM):

A general assessment of the status and needs of the tools for setting limits on (or fitting) parameters of BSM models, using the multitude of data from searches at the LHC

Two related communities and ongoing discussions

- Characterization & Simplified Models → parametrization
- Fitting Model Parameters → interpretation

Potential new tasks

- **Input for the Strategy Group**
  - LPCC and experiments required to produce combined assessment of the 2010-11(-12) findings in Higgs and BSM searches
  - TH community, and other expl communities (e.g. LinCol, SuperB, ...), will use this to assess the implications of LHC data for BSM and future exptl projects

➡ We need to prepare the framework/tools to enable:

  - combination of limits/evidence from ATLAS/CMS(/LHCb)
  - use of the results by the rest of the community (e.g. SUSY-models' fitters)
- This will require coordination with
  - ATLAS-CMS statistics forum
  - Fitters' groups
  - all LHC "search " efforts (Higgs, B decays, exotica of all sorts ....)
  - ...

Goals for this meeting

- Review the progress made by the experiments
- Status report on the SLAC WG
- Collect further input from all fields (TH + exps)
- In the context of simplified models, start outlining the roadmap and the workflow to go from analysis, to publication, to combination of the results of different experiments, to conclude with the exploitation of the published results by a random theorist.

analysis

format of the published result

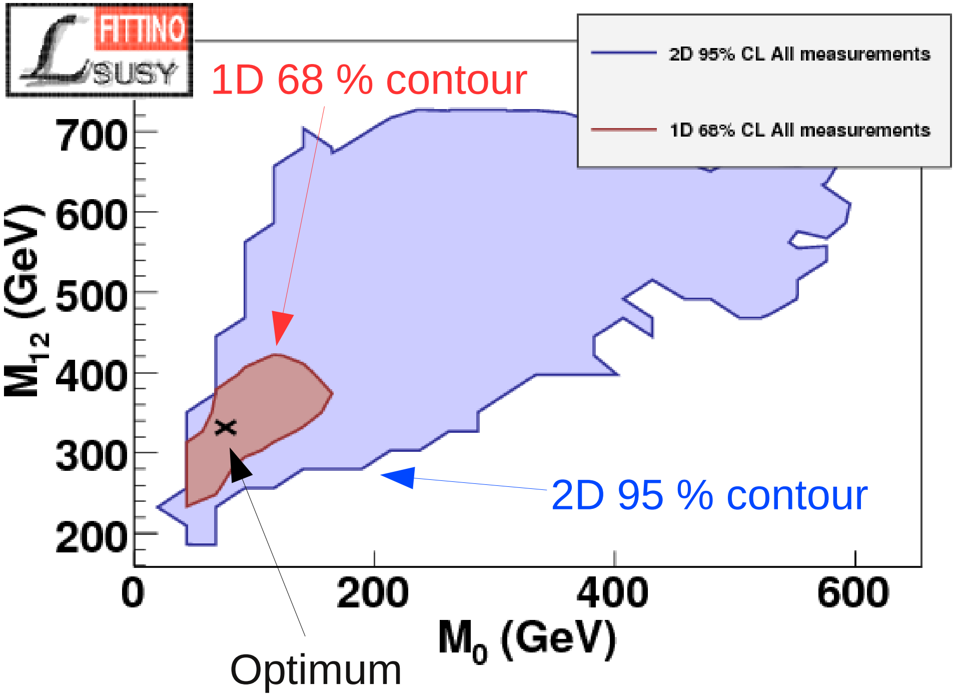
combination among experiments

use of the results by a theorist, in the context of a new model

SUSY Fitting tools

Usually simplify input from experiments to be a single Gaussian

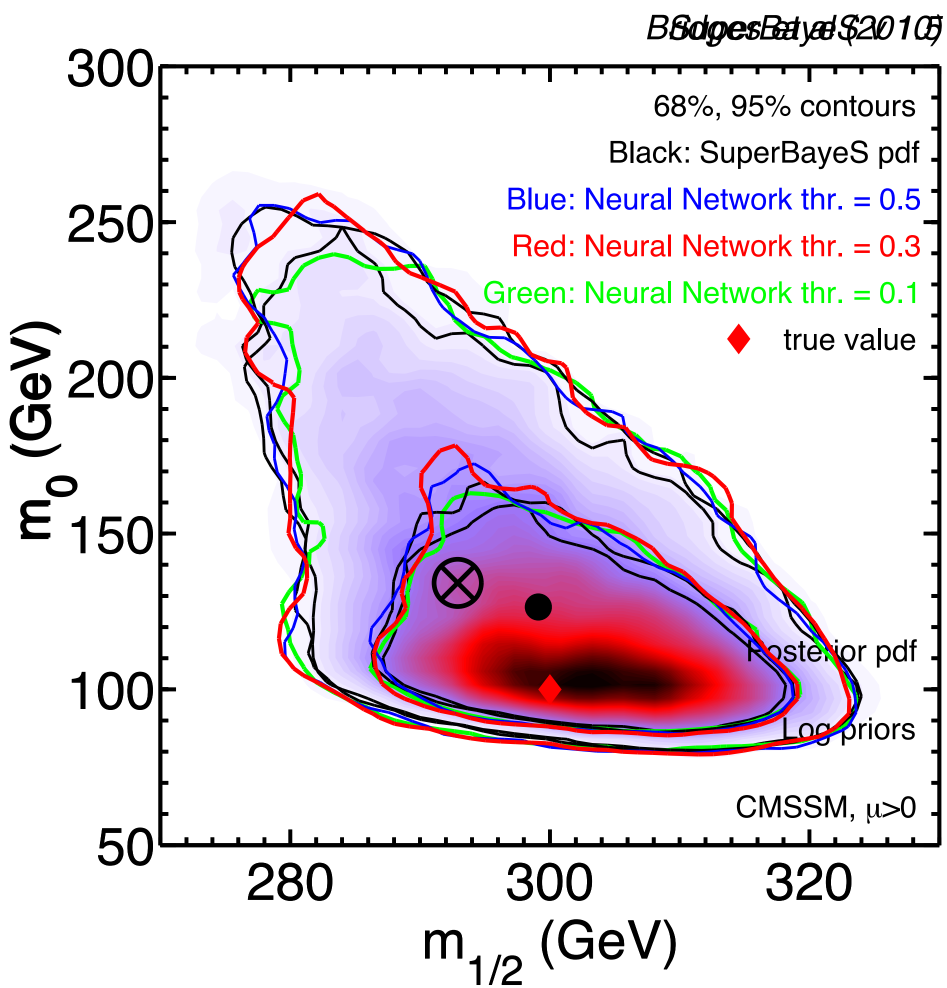
Observable	Experimental Value	Uncertainty stat	Uncertainty syst	Exp. Reference
$\mathcal{B}(B \to s\gamma)/\mathcal{B}(B \to s\gamma)_{SM}$	1.117	0.076	0.096	[47]
$\mathcal{B}(B_s \to \mu\mu)$	$< 4.7\times10^{-8}$			[47]
$\mathcal{B}(B_d \to \ell\ell)$	$< 2.3\times10^{-8}$			[47]
$\mathcal{B}(B \to \tau\nu)/\mathcal{B}(B \to \tau\nu)_{SM}$	1.15	0.40		[48]
$\mathcal{B}(B_s \to X_s\ell\ell)/\mathcal{B}(B_s \to X_s\ell\ell)_{SM}$	0.99	0.32		[47]
$\Delta m_{B_s}/\Delta m_{B_s}^{SM}$	1.11	0.01	0.32	[49]
$\frac{\Delta m_{B_s}/\Delta m_{B_s}^{SM}}{\Delta m_{B_d}/\Delta m_{B_d}^{SM}}$	1.09	0.01	0.16	[47,49]
$\Delta\epsilon_K/\Delta\epsilon_K^{SM}$	0.92	0.14		[49]
$\mathcal{B}(K \to \mu\nu)/\mathcal{B}(K \to \mu\nu)_{SM}$	1.008	0.014		[50]
$\mathcal{B}(K \to \pi\nu\bar{\nu})/\mathcal{B}(K \to \pi\nu\bar{\nu})_{SM}$	< 4.5			[51]
$a_\mu^{exp} - a_\mu^{SM}$	30.2×10^{-10}	8.8×10^{-10}	2.0×10^{-10}	[52,53]
$\sin^2\theta_{eff}$	0.2324	0.0012		[46]
Γ_Z	2.4952 GeV	0.0023 GeV	0.001 GeV	[46]
R_l	20.767	0.025		[46]
R_b	0.21629	0.00066		[46]
R_c	0.1721	0.003		[46]
$A_{fb}(b)$	0.0992	0.0016		[46]
$A_{fb}(c)$	0.0707	0.0035		[46]
A_b	0.923	0.020		[46]
A_c	0.670	0.027		[46]
A_l	0.1513	0.0021		[46]
A_τ	0.1465	0.0032		[46]
$A_{fb}(l)$	0.01714	0.00095		[46]
σ_{had}	41.540 nb	0.037 nb		[46]
m_h	> 114.4 GeV		3.0 GeV	[54,55,56]
$\Omega_{CDM}h^2$	0.1099	0.0062	0.012	[57]
$1/\alpha_{em}$	127.925	0.016		[58]
G_F	$1.16637\times10^{-5}\mathrm{GeV}^{-2}$	$0.00001\times10^{-5}\mathrm{GeV}^{-2}$		[58]
α_s	0.1176	0.0020		[58]
m_Z	91.1875 GeV	0.0021 GeV		[46]
m_W	80.399 GeV	0.025 GeV	0.010 GeV	[58]
m_b	4.20 GeV	0.17 GeV		[58]
m_t	172.4 GeV	1.2 GeV		[59]
m_τ	1.77684 GeV	0.00017 GeV		[58]
m_c	1.27 GeV	0.11 GeV		[46]

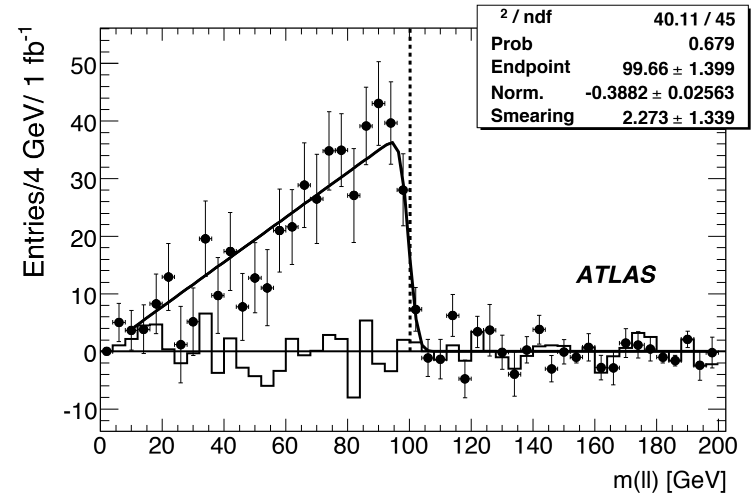


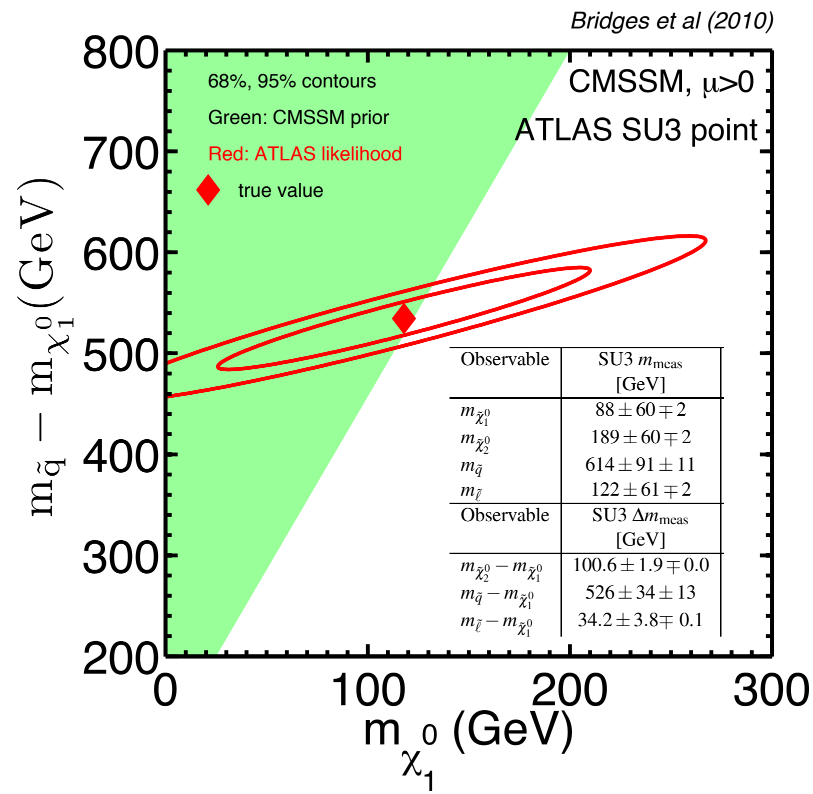
First interface with SuperBayes

Repeated same analysis as Bridges, KC, Trotta et al (1011.4306) with RooStats likelihood

- see consistent results!



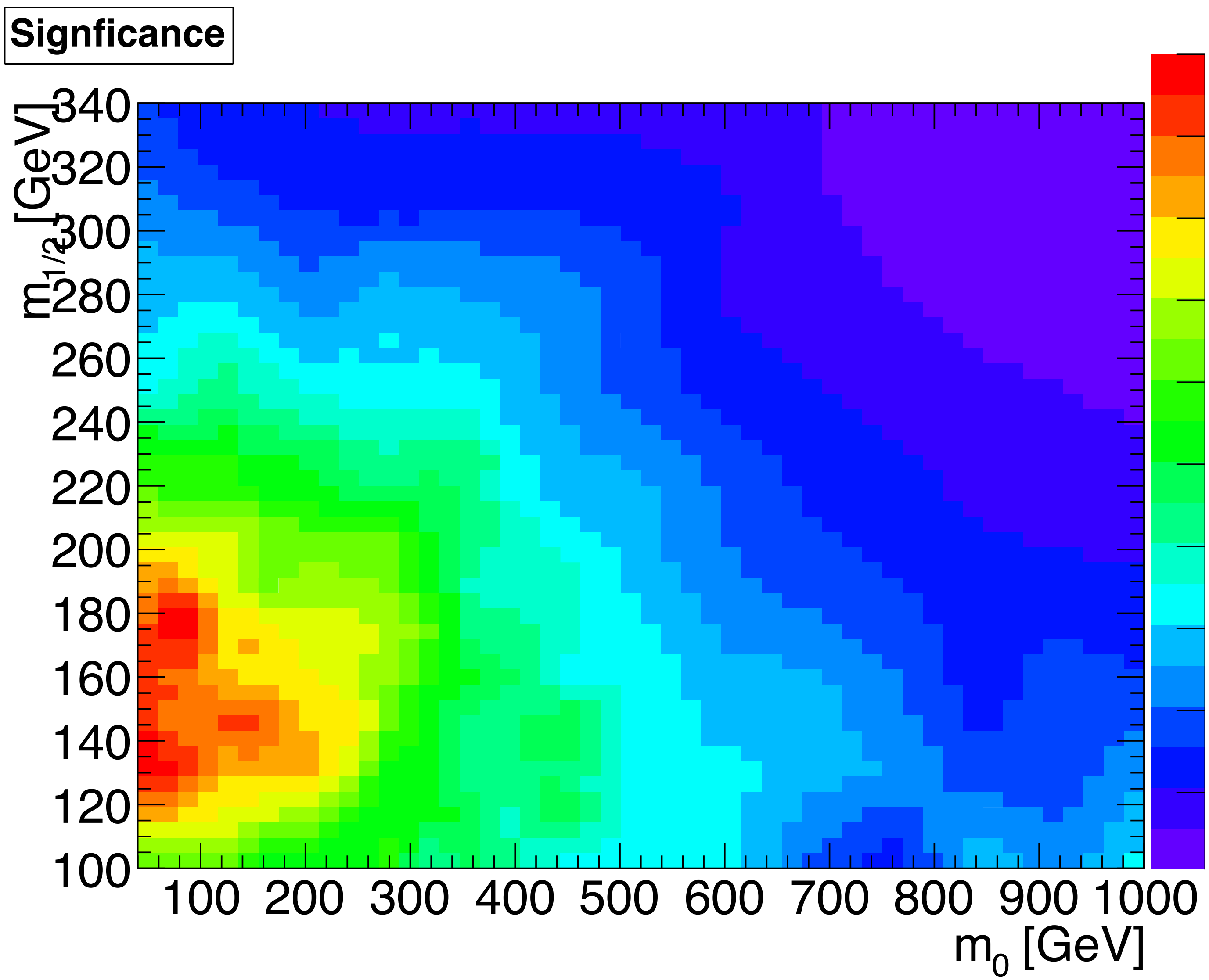




Benchmark based on counting

Center for Cosmology and Particle Physics

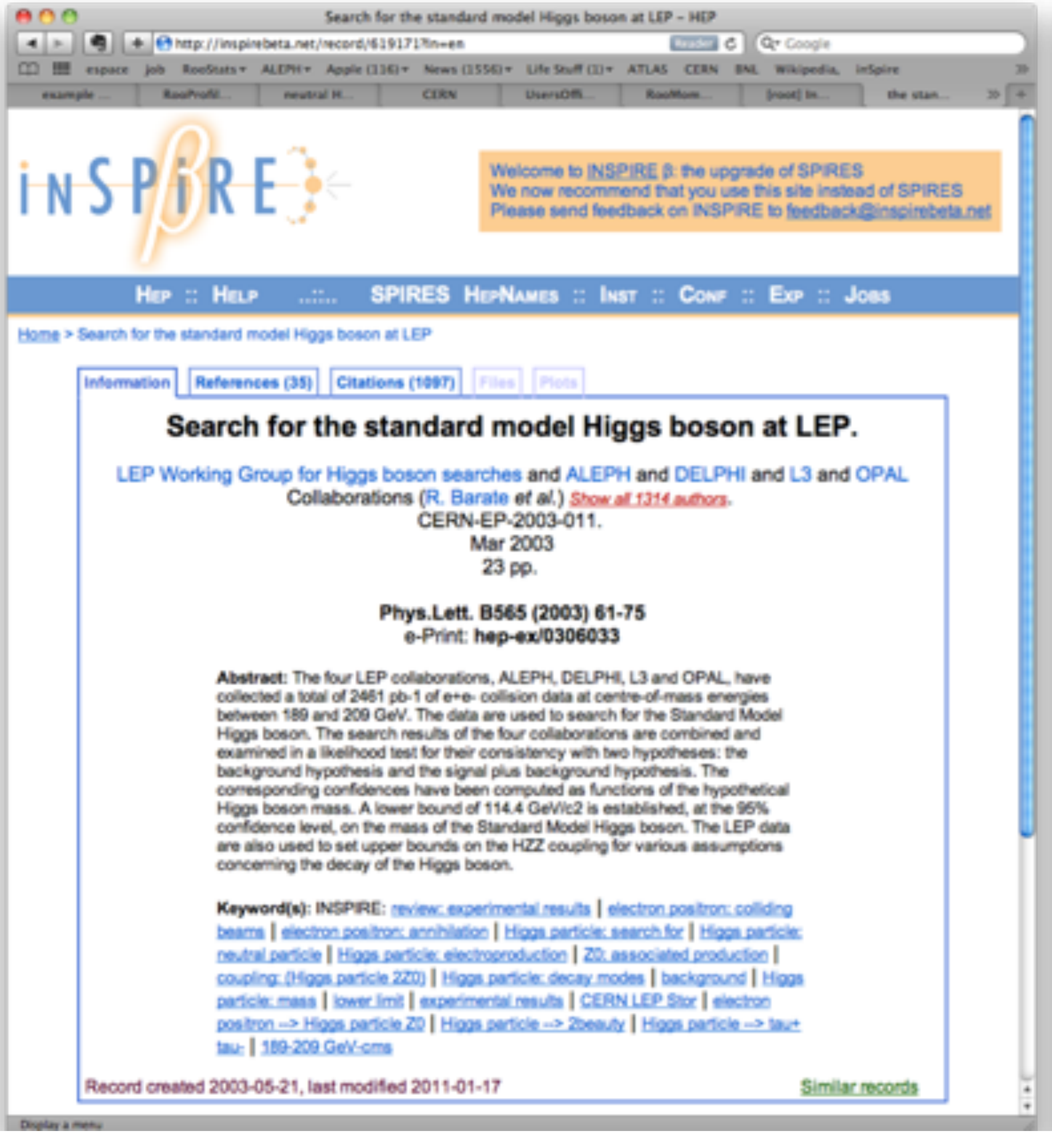
Max Baak's demonstrated interpolation of signal yield and uncertainties in a 3-d mSUGRA scan with a simple number counting analysis



Ultimate Goal

Publish likelihoods along with papers

- first goal, the LEP Higgs


Search for the standard model Higgs boson at LEP - HEP

Welcome to INSPIRE β: the upgrade of SPIRES
We now recommend that you use this site instead of SPIRES
Please send feedback on INSPIRE to feedback@inspirebeta.net

HEP :: HELP .::. SPIRES HEPNAMES :: INST :: CONF :: EXP :: JOBS

Home > Search for the standard model Higgs boson at LEP

Information | References (35) | Citations (1097) | Files | Plots

Search for the standard model Higgs boson at LEP.

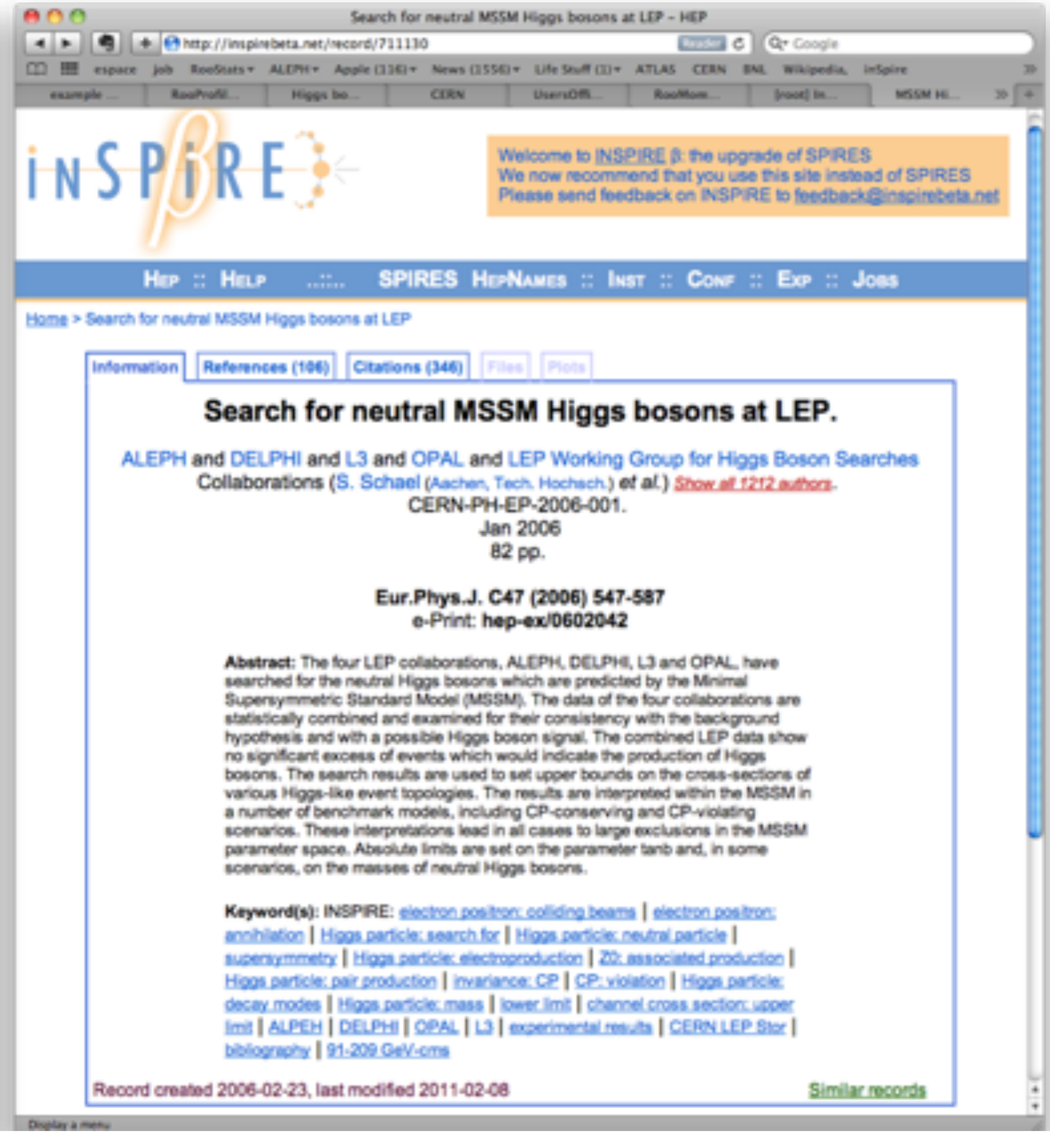
LEP Working Group for Higgs boson searches and ALEPH and DELPHI and L3 and OPAL Collaborations (R. Barate *et al.*) Show all 1314 authors.
CERN-EP-2003-011.
Mar 2003
23 pp.

Phys.Lett. B565 (2003) 61-75
e-Print: **hep-ex/0306033**

Abstract: The four LEP collaborations, ALEPH, DELPHI, L3 and OPAL, have collected a total of 2461 pb-1 of e+e- collision data at centre-of-mass energies between 189 and 209 GeV. The data are used to search for the Standard Model Higgs boson. The search results of the four collaborations are combined and examined in a likelihood test for their consistency with two hypotheses: the background hypothesis and the signal plus background hypothesis. The corresponding confidences have been computed as functions of the hypothetical Higgs boson mass. A lower bound of 114.4 GeV/c2 is established, at the 95% confidence level, on the mass of the Standard Model Higgs boson. The LEP data are also used to set upper bounds on the HZZ coupling for various assumptions concerning the decay of the Higgs boson.

Keyword(s): INSPIRE: review: experimental results | electron positron: colliding beams | electron positron: annihilation | Higgs particle: search for | Higgs particle: neutral particle | Higgs particle: electroproduction | Z0: associated production | coupling: (Higgs particle 2Z0) | Higgs particle: decay modes | background | Higgs particle: mass | lower limit | experimental results | CERN LEP Stor | electron positron --> Higgs particle Z0 | Higgs particle --> 2beauty | Higgs particle --> tau+ tau- | 189-209 GeV-cms

Record created 2003-05-21, last modified 2011-01-17 Similar records


Search for neutral MSSM Higgs bosons at LEP - HEP

Welcome to INSPIRE β: the upgrade of SPIRES
We now recommend that you use this site instead of SPIRES
Please send feedback on INSPIRE to feedback@inspirebeta.net

HEP :: HELP .::. SPIRES HEPNAMES :: INST :: CONF :: EXP :: JOBS

Home > Search for neutral MSSM Higgs bosons at LEP

Information | References (106) | Citations (346) | Files | Plots

Search for neutral MSSM Higgs bosons at LEP.

ALEPH and DELPHI and L3 and OPAL and LEP Working Group for Higgs Boson Searches Collaborations (S. Schael (Aachen, Tech. Hochsch.) *et al.*) Show all 1212 authors.
CERN-PH-EP-2006-001.
Jan 2006
82 pp.

Eur.Phys.J. C47 (2006) 547-587
e-Print: **hep-ex/0602042**

Abstract: The four LEP collaborations, ALEPH, DELPHI, L3 and OPAL, have searched for the neutral Higgs bosons which are predicted by the Minimal Supersymmetric Standard Model (MSSM). The data of the four collaborations are statistically combined and examined for their consistency with the background hypothesis and with a possible Higgs boson signal. The combined LEP data show no significant excess of events which would indicate the production of Higgs bosons. The search results are used to set upper bounds on the cross-sections of various Higgs-like event topologies. The results are interpreted within the MSSM in a number of benchmark models, including CP-conserving and CP-violating scenarios. These interpretations lead in all cases to large exclusions in the MSSM parameter space. Absolute limits are set on the parameter tanb and, in some scenarios, on the masses of neutral Higgs bosons.

Keyword(s): INSPIRE: electron positron: colliding beams | electron positron: annihilation | Higgs particle: search for | Higgs particle: neutral particle | supersymmetry | Higgs particle: electroproduction | Z0: associated production | Higgs particle: pair production | invariance: CP | CP: violation | Higgs particle: decay modes | Higgs particle: mass | lower limit | channel cross section: upper limit | ALPEH | DELPHI | OPAL | L3 | experimental results | CERN LEP Stor | bibliography | 91-209 GeV-cms

Record created 2006-02-23, last modified 2011-02-08 Similar records

CERN Colloquium and Library Science Talk

SPEAKER: Lawrence Lessig (Edmond J. Safra Center for Ethics and Harvard Law School, Cambridge, MA, US)

TITLE: **"The architecture of access to scientific knowledge: just how badly we have messed this up"**

DATE: Mon 18/04/2011 16:30

PLACE: Council Chamber

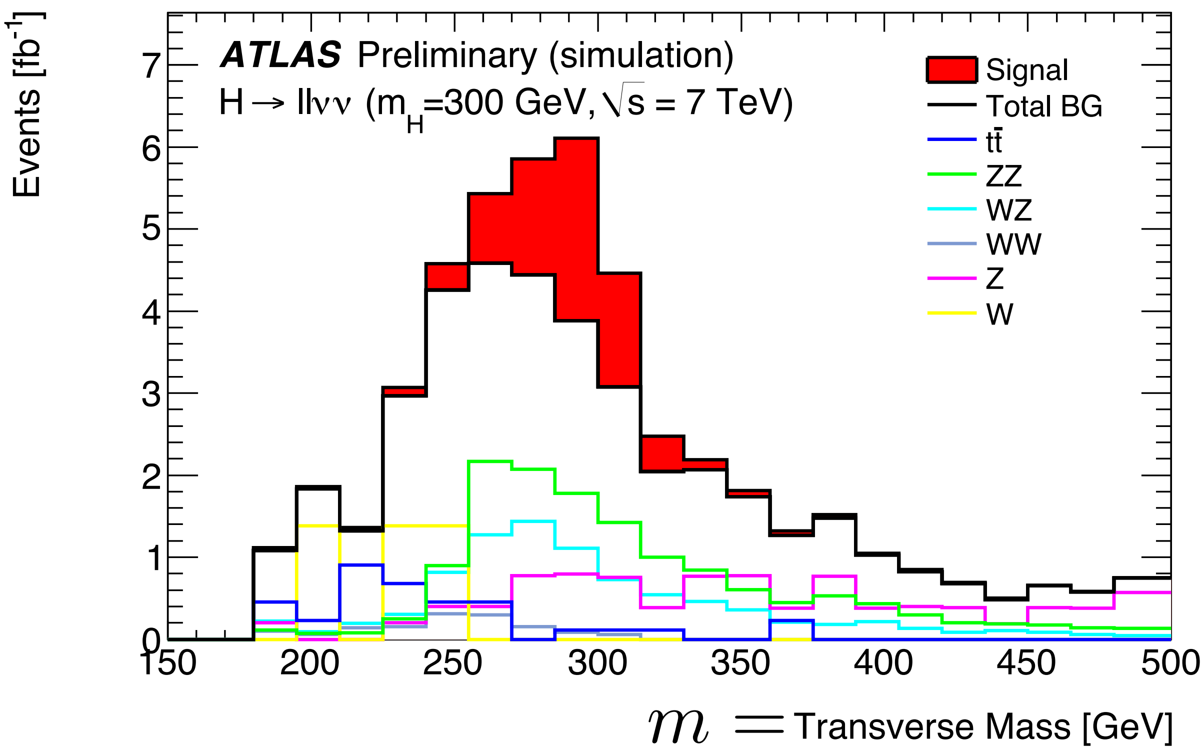
ABSTRACT

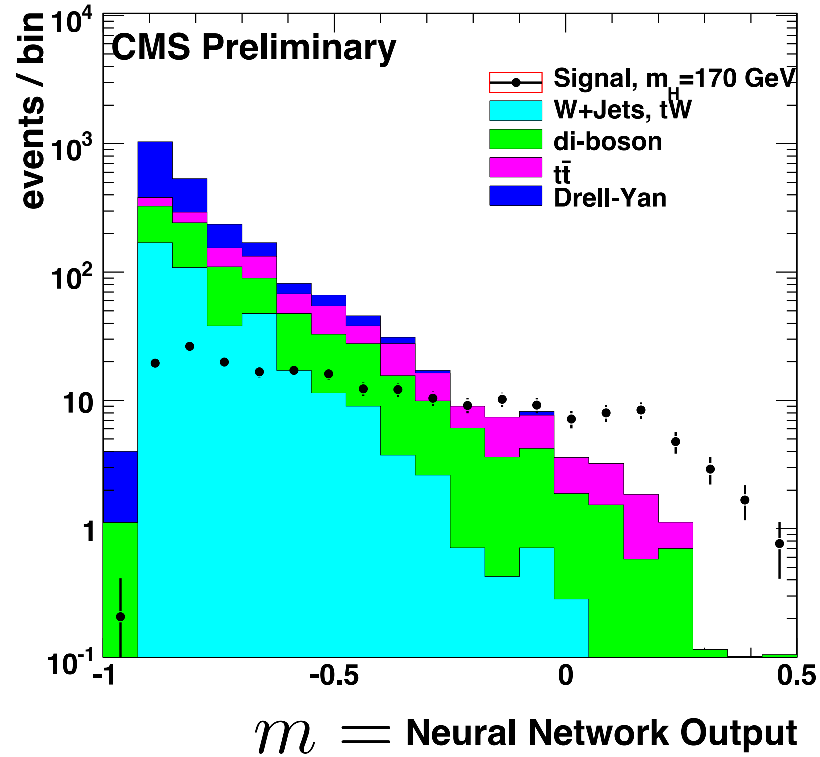
In this talk, Professor Lessig will review the evolution of access to scientific scholarship, and evaluate the success of this system of access against a background norm of universal access. While copyright battles involving artists has gotten most of the public's attention, the real battle should be over access to knowledge, not culture. That battle we are losing.

Lecture 2

Modeling:
The Scientific Narrative
(continued)

In Monte Carlo Simulation approach, use simulated events to build histograms and construct the “Marked Poisson” model below

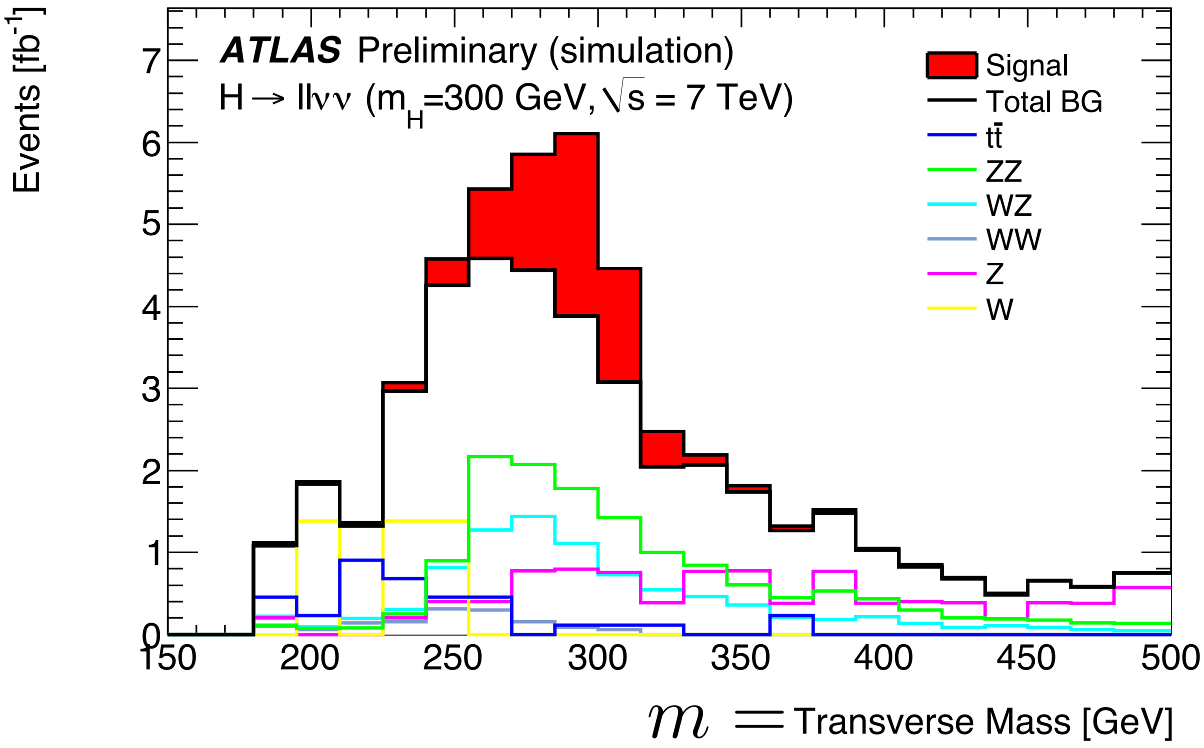




$$P(\mathbf{m}|s) = \text{Pois}(n|s+b) \prod_j^n \frac{s f_s(m_j) + b f_b(m_j)}{s+b}$$

Tabulate effect of individual variations of sources of systematic uncertainty

- use some form of interpolation to parametrize i^{th} variation in terms of nuisance parameter α_i

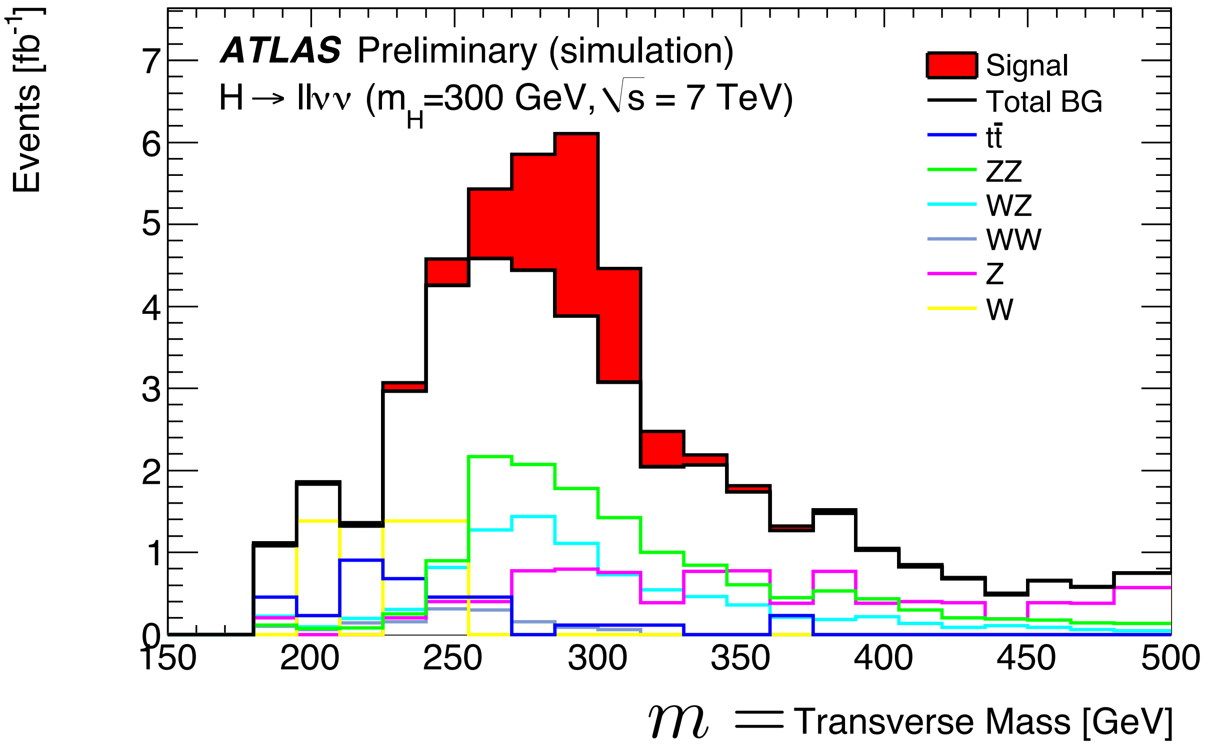


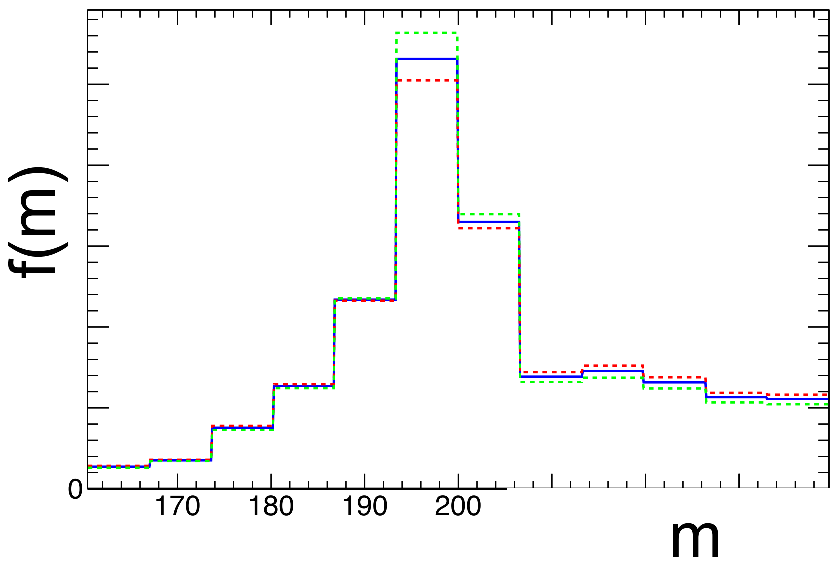
	sig	bkg 1	bkg 2	...
syst 1				
syst 2				
...				

$$P(\mathbf{m}|\boldsymbol{\alpha}) = \text{Pois}(n|s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})) \prod_j^n \frac{s(\boldsymbol{\alpha}) f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha}) f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})}$$

Tabulate effect of individual variations of sources of systematic uncertainty

- use some form of interpolation to parametrize i^{th} variation in terms of nuisance parameter α_i



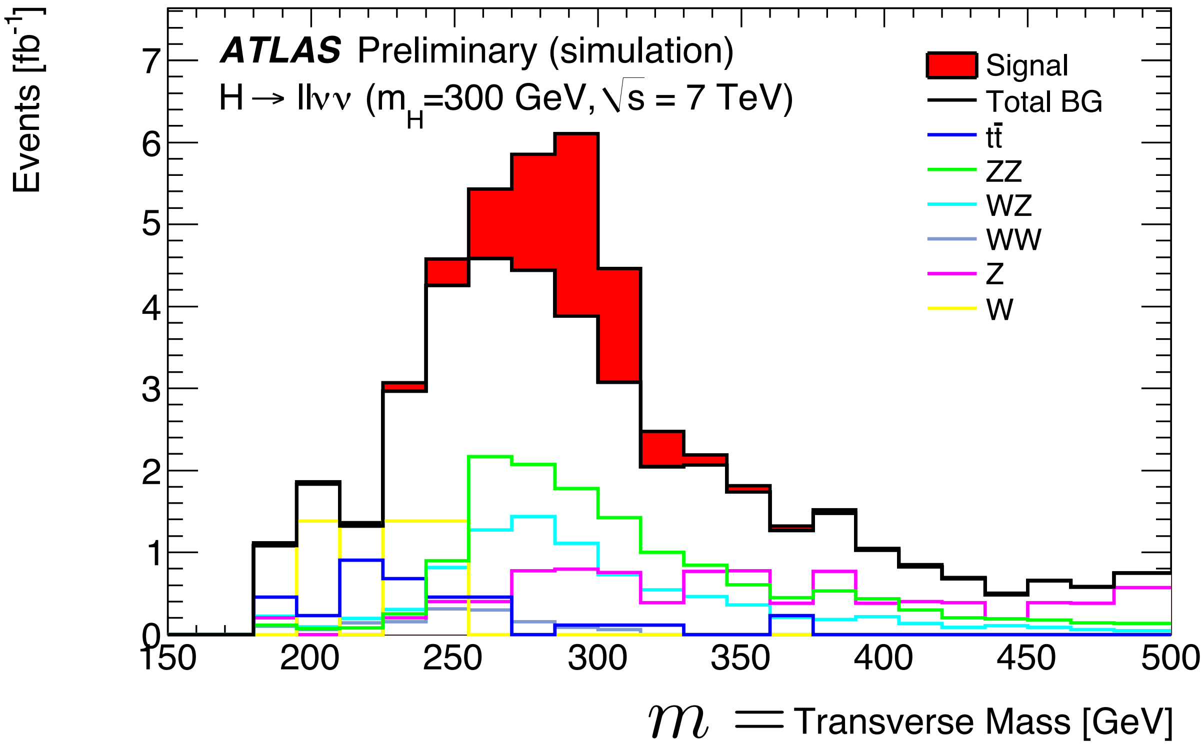


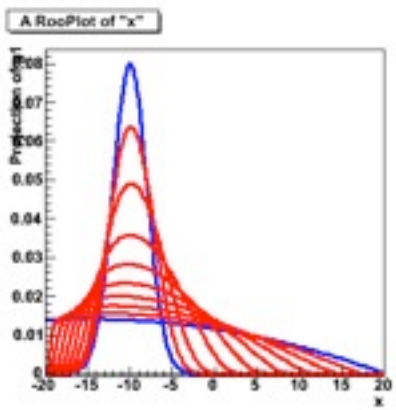
$$P(\mathbf{m}|\boldsymbol{\alpha}) = \text{Pois}(n|s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})) \prod_{j}^{n} \frac{s(\boldsymbol{\alpha}) f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha}) f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})}$$

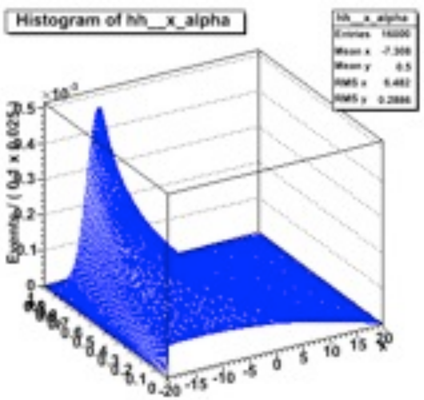
Center for Cosmology and Particle Physics

Tabulate effect of individual variations of sources of systematic uncertainty

- use some form of interpolation to parametrize i^{th} variation in terms of nuisance parameter α_i





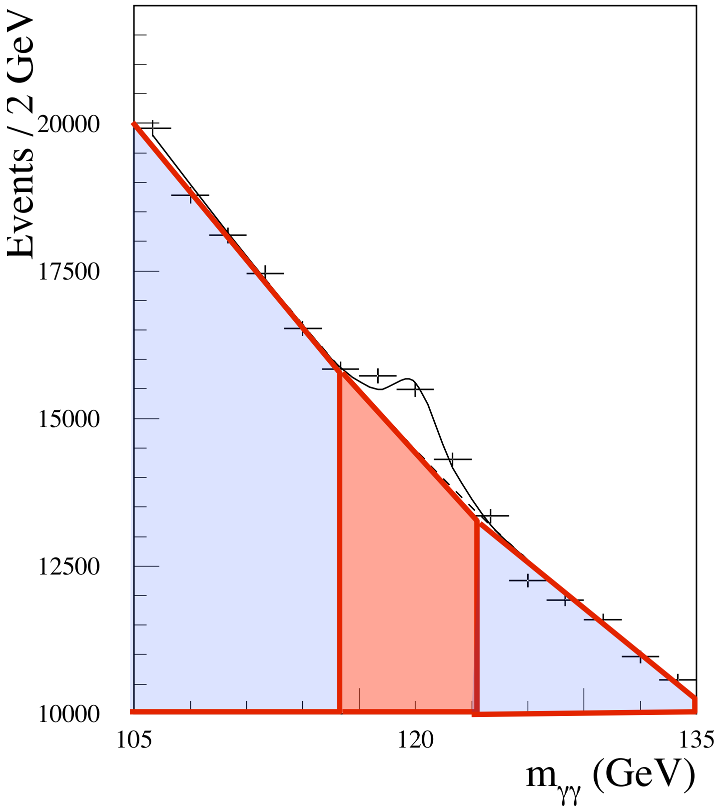


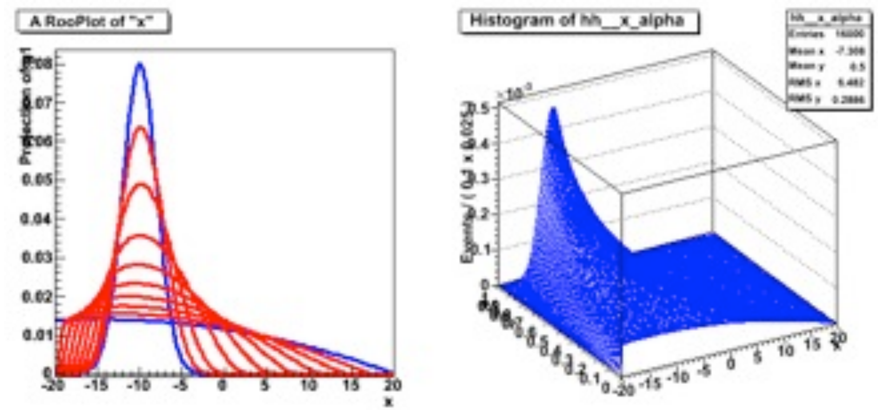
$$P(\mathbf{m}|\boldsymbol{\alpha}) = \text{Pois}(n|s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})) \prod_{j}^{n} \frac{s(\boldsymbol{\alpha}) f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha}) f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})}$$

Review

Something must 'constrain' the α

- the data itself: sidebands; some control region
- constraint term: idealized form of auxiliary measurement or ad hoc 'prior'

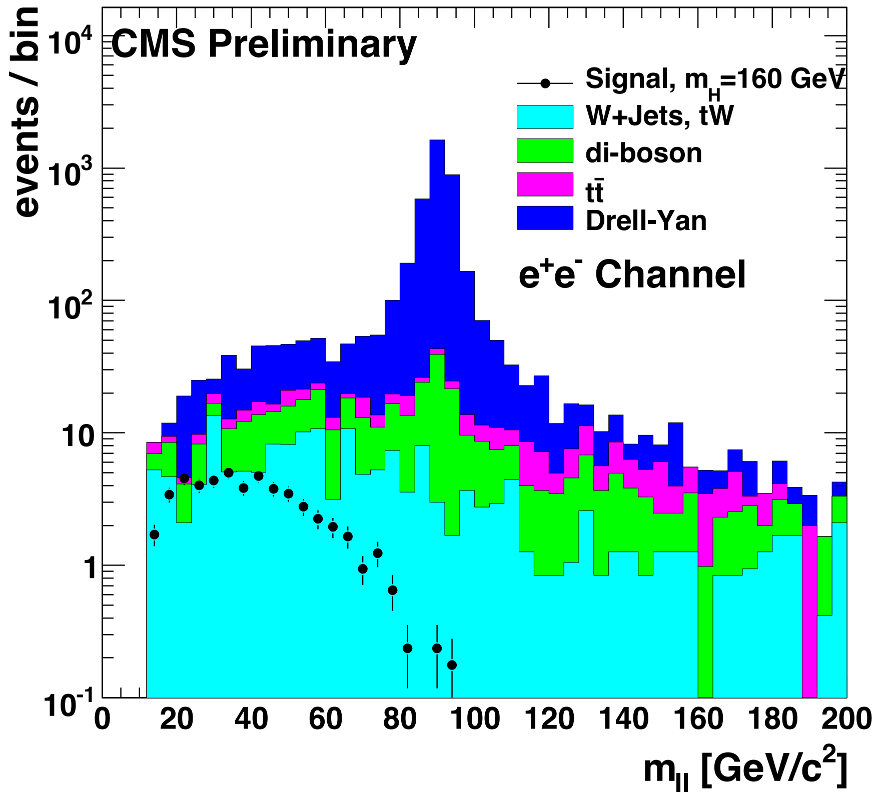


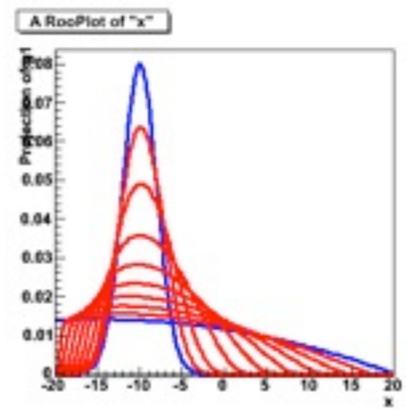


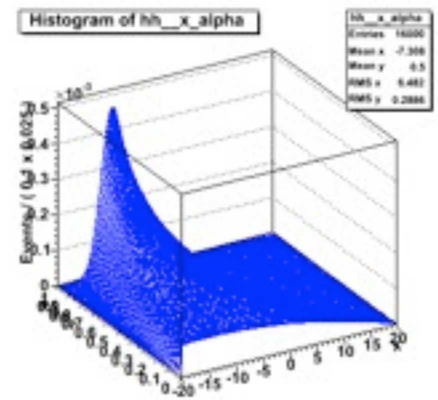
$$P(\mathbf{m}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})) \prod_{j}^{n} \frac{s(\boldsymbol{\alpha}) f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha}) f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})}$$

Something must 'constrain' the α

- the data itself: sidebands; some control region
- constraint term: idealized form of auxiliary measurement or ad hoc 'prior'





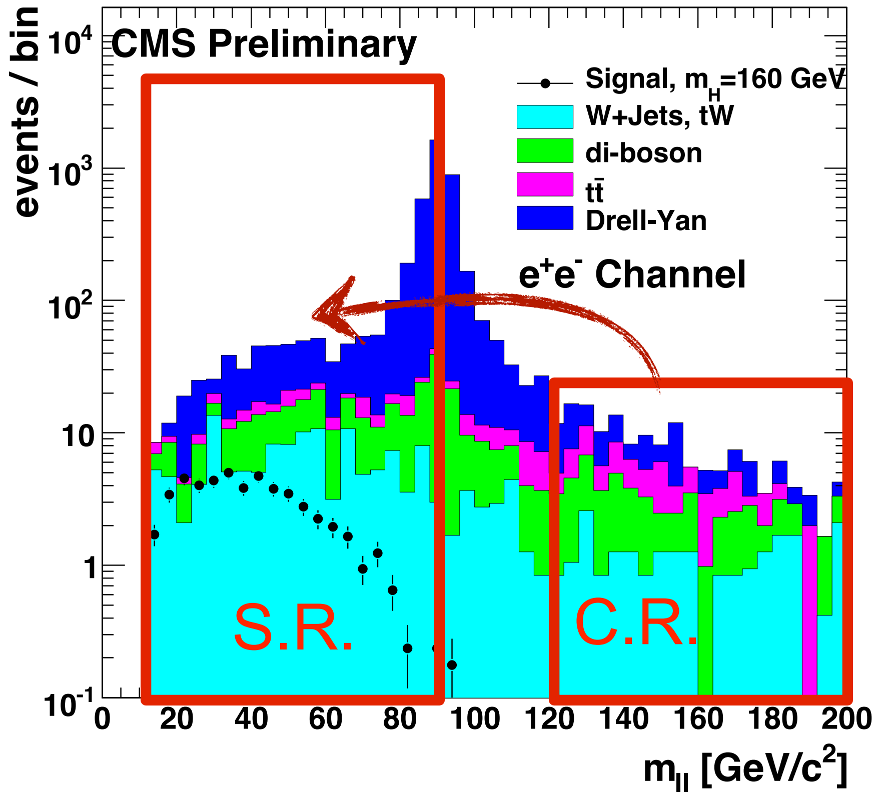


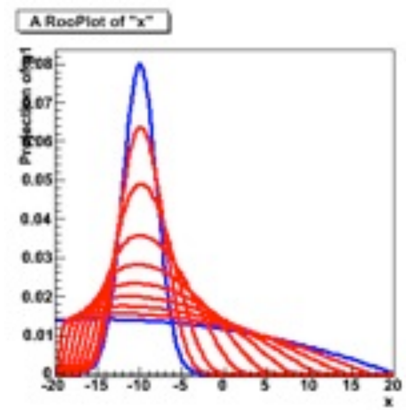
$$P(\mathbf{m}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})) \prod_{j}^{n} \frac{s(\boldsymbol{\alpha}) f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha}) f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})}$$

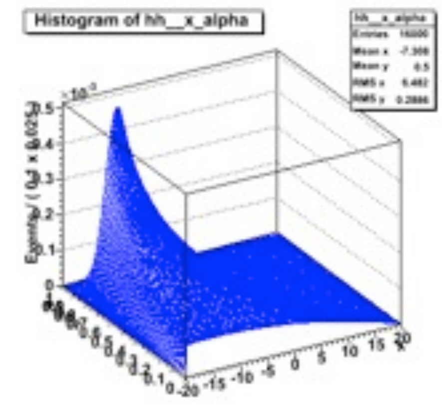
Review

Something must ‘constrain’ the α

- the data itself: sidebands; some control region
- constraint term: idealized form of auxiliary measurement or ad hoc ‘prior’





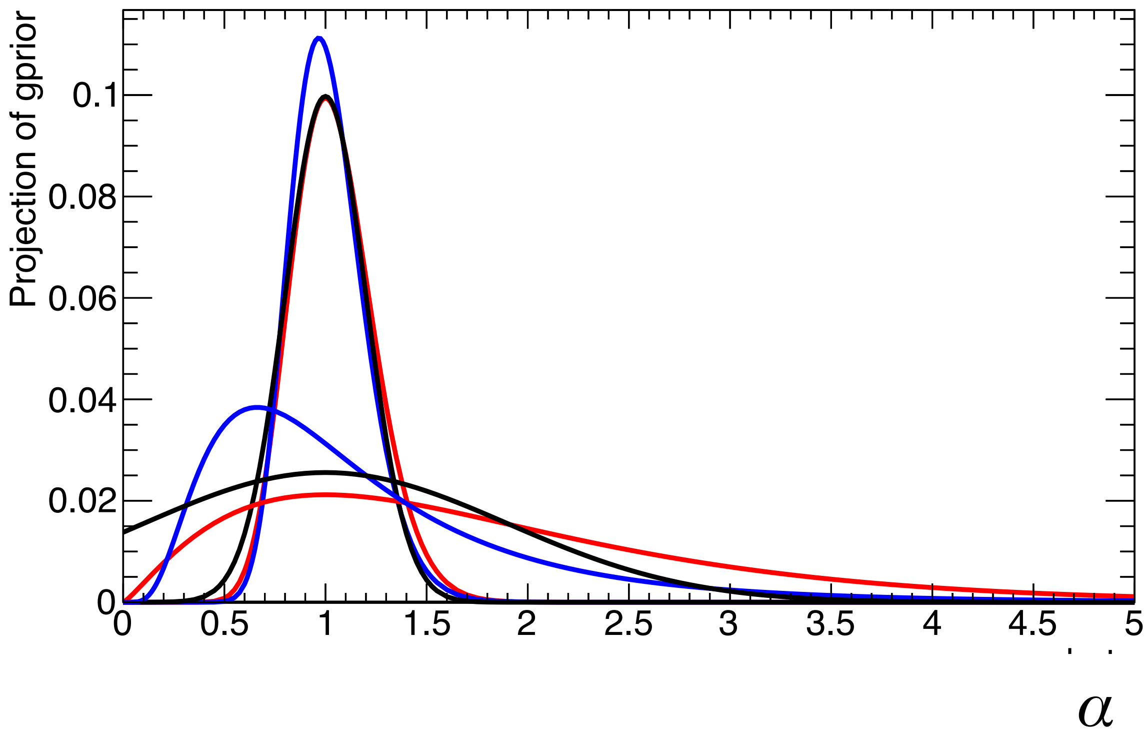


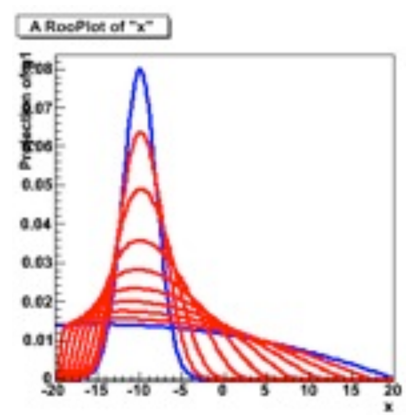
$$P(\mathbf{m}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})) \prod_j^n \frac{s(\boldsymbol{\alpha}) f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha}) f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})}$$

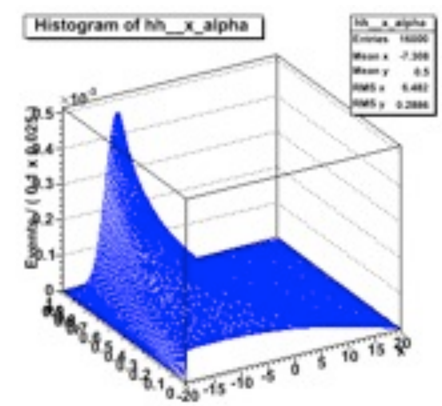
Center for Cosmology and Particle Physics

Something must ‘constrain’ the α

- the data itself: sidebands; some control region
- constraint term: idealized form of auxiliary measurement or ad hoc ‘prior’







$$P(\mathbf{m}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha})) \prod_j^n \frac{s(\boldsymbol{\alpha})f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha})}$$
$$\times G(a|\alpha,\sigma)$$

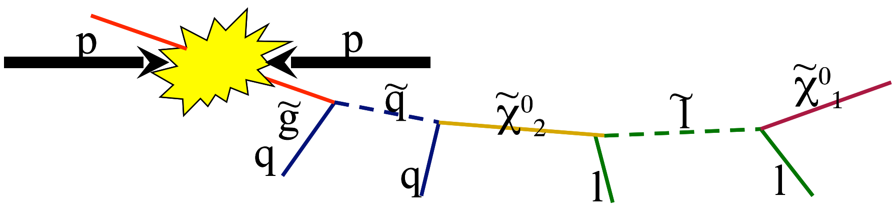
The Data-Driven narrative

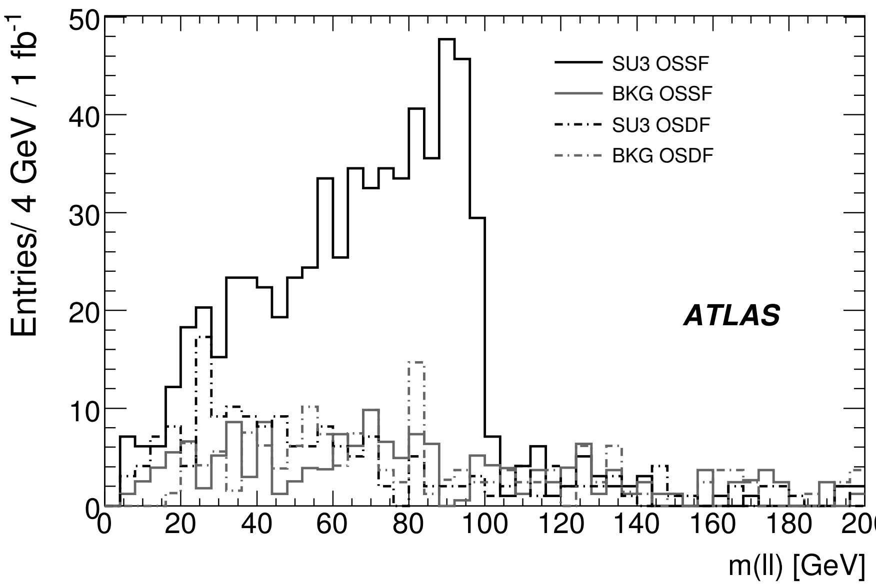
In the data-driven approach, backgrounds are estimated by assuming (and testing) some relationship between a control region and signal region

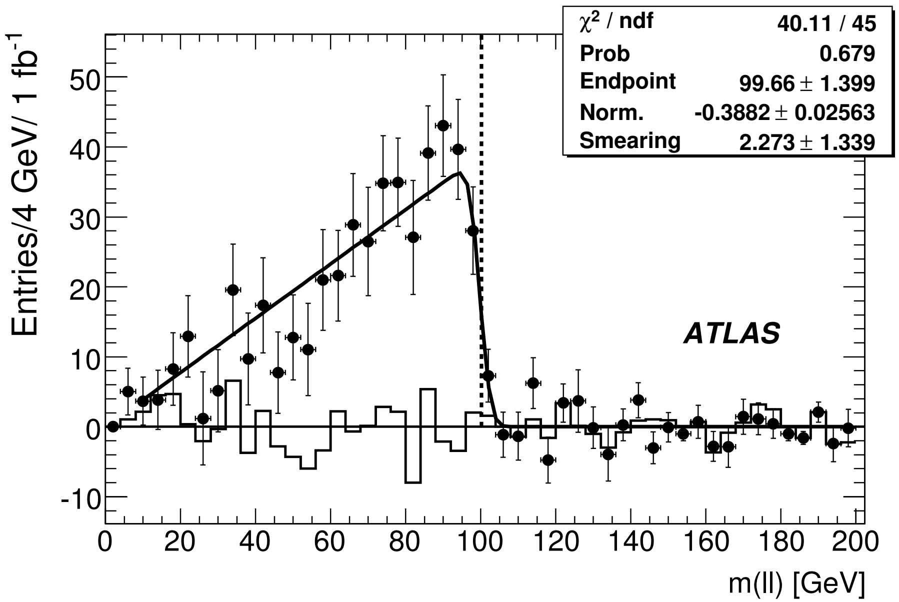
- flavor subtraction, same-sign samples, fake matrix, tag-probe, ...

Pros: Initial sample has "all orders" theory :-) and all the details of the detector

Cons: assumptions made in the transformation to the signal region can be questioned





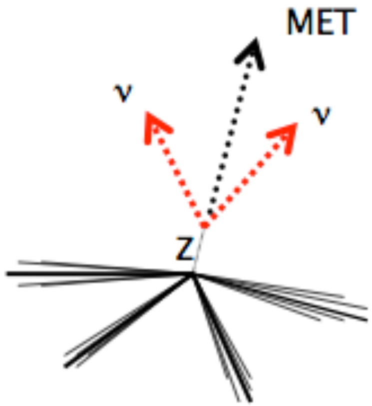


Other Examples of data-driven narrative

All-hadronic searches with MHT

Search for high pT jets, high HT and high MHT (= vector sum of jets)

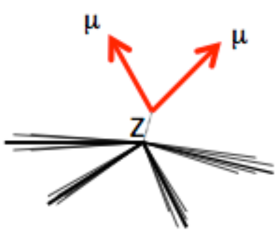
3 jets, E_T>50 $|\eta|$<2.5



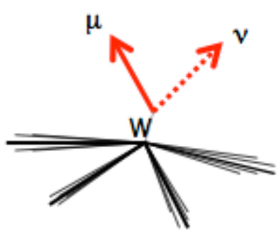
HT > 350 and MHT > 150

Event cleaning cuts.

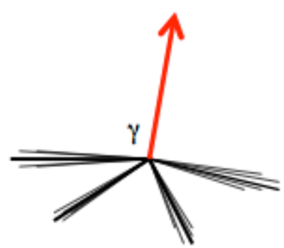
Predict each bkgd separately
- QCD: rebalance & smear
- W & ttbar from μ control
- $Z\rightarrow\nu\nu$ from γ+jets and $Z\rightarrow\mu\mu$



Z → ll + jets
Strength: very clean
Weakness: low statistics



W → lν + jets
Strength: larger statistics
Weakness: background from SM and SUSY



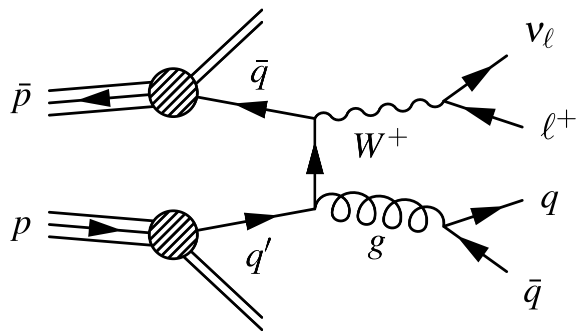
γ + jets
Strength: large statistics and clean at high E_T
Weakness: background at low E_T, theoretical errors

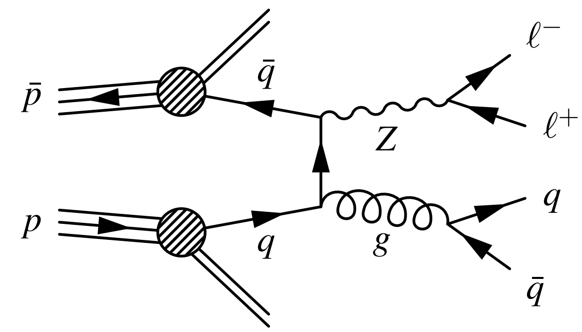
CMS SUSY Results, D. Stuart, April 2011, SUSY Recast, UC Davis

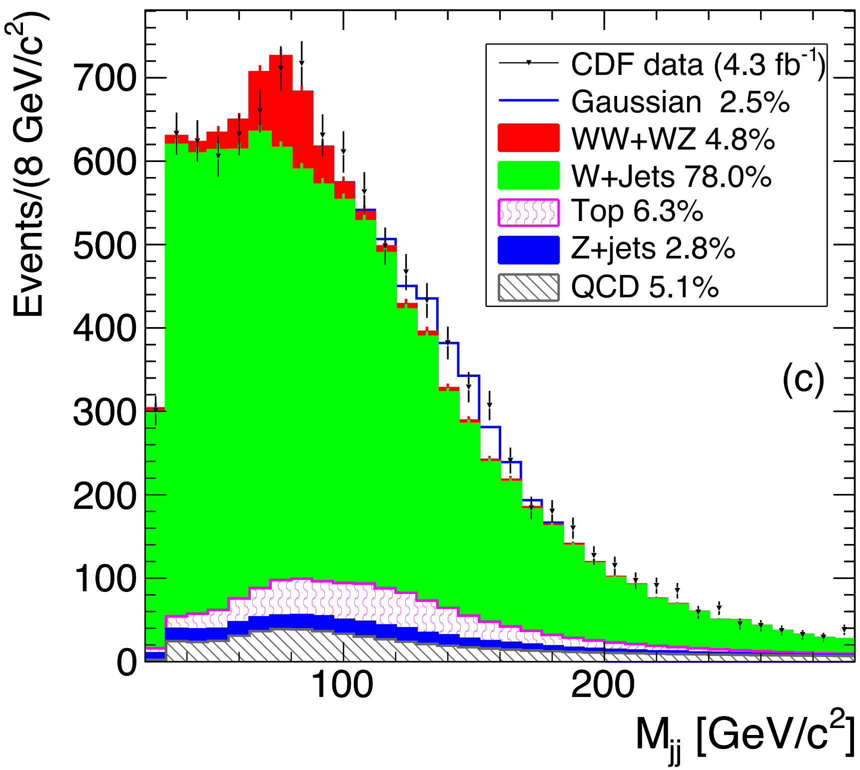
Data driven estimates

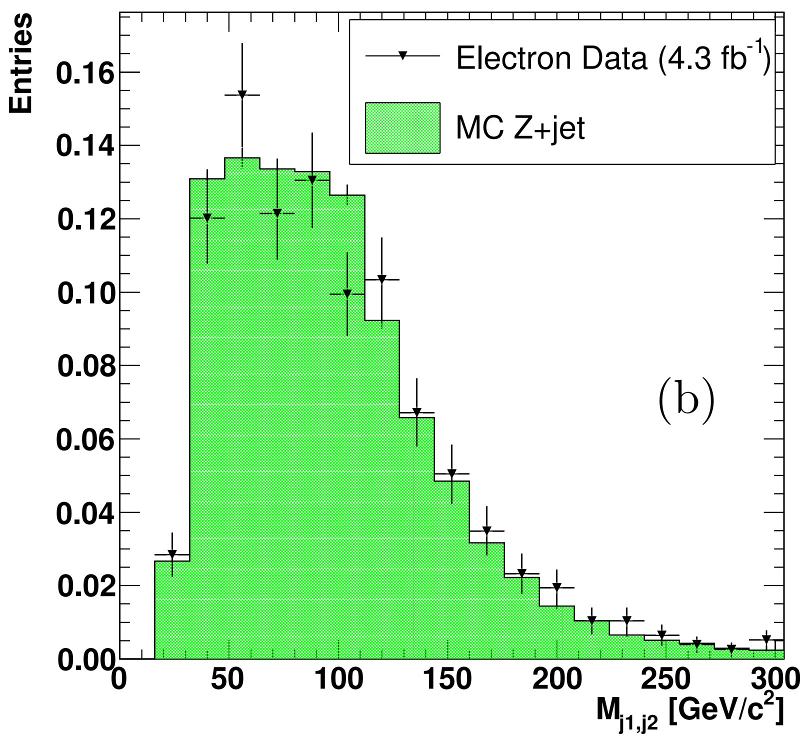
Center for Cosmology and Particle Physics

In the case of the CDF bump, the Z+jets control sample provides a data-driven estimate, but limited statistics. Using the simulation narrative over the data-driven is a **choice**. If you trust that narrative, it's a good choice.







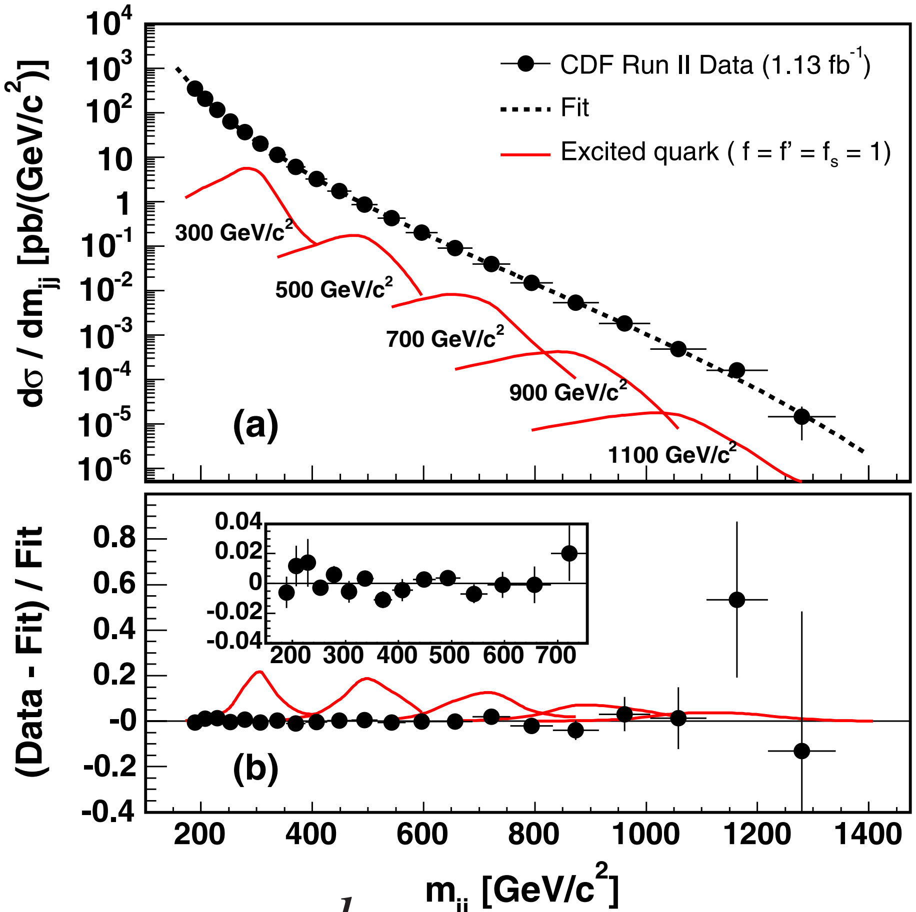


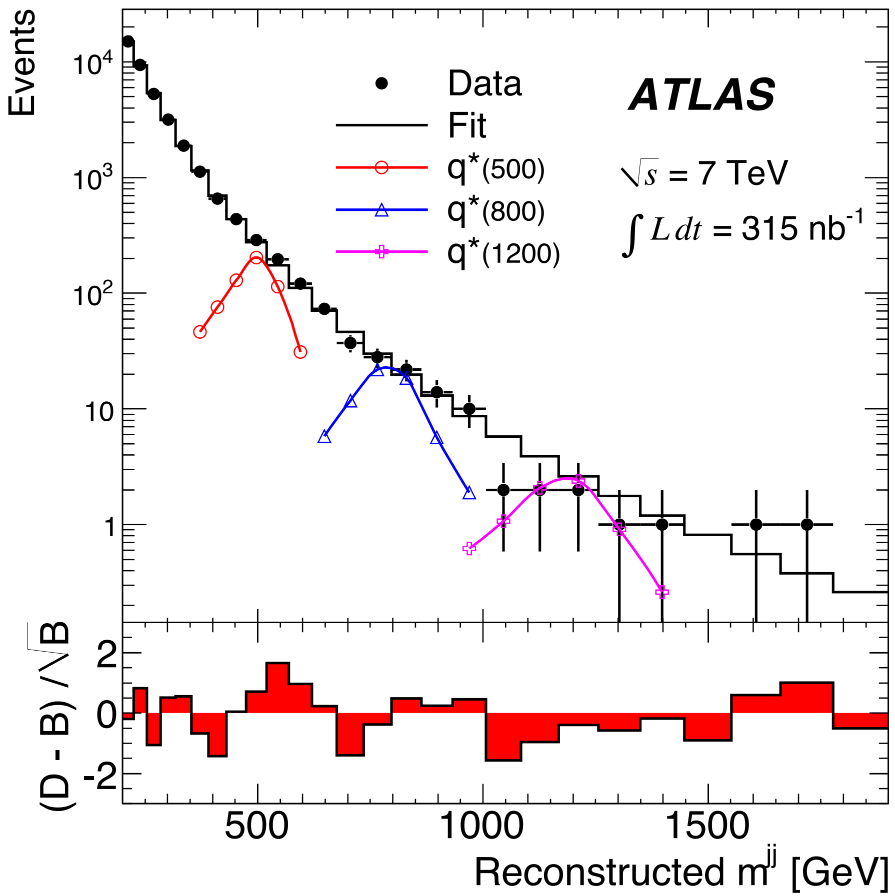
The Effective Model Narrative

It is common to describe a distribution with some parametric function

- "fit background to a polynomial", exponential, ...
- While this is convenient and the fit may be good, the narrative is weak

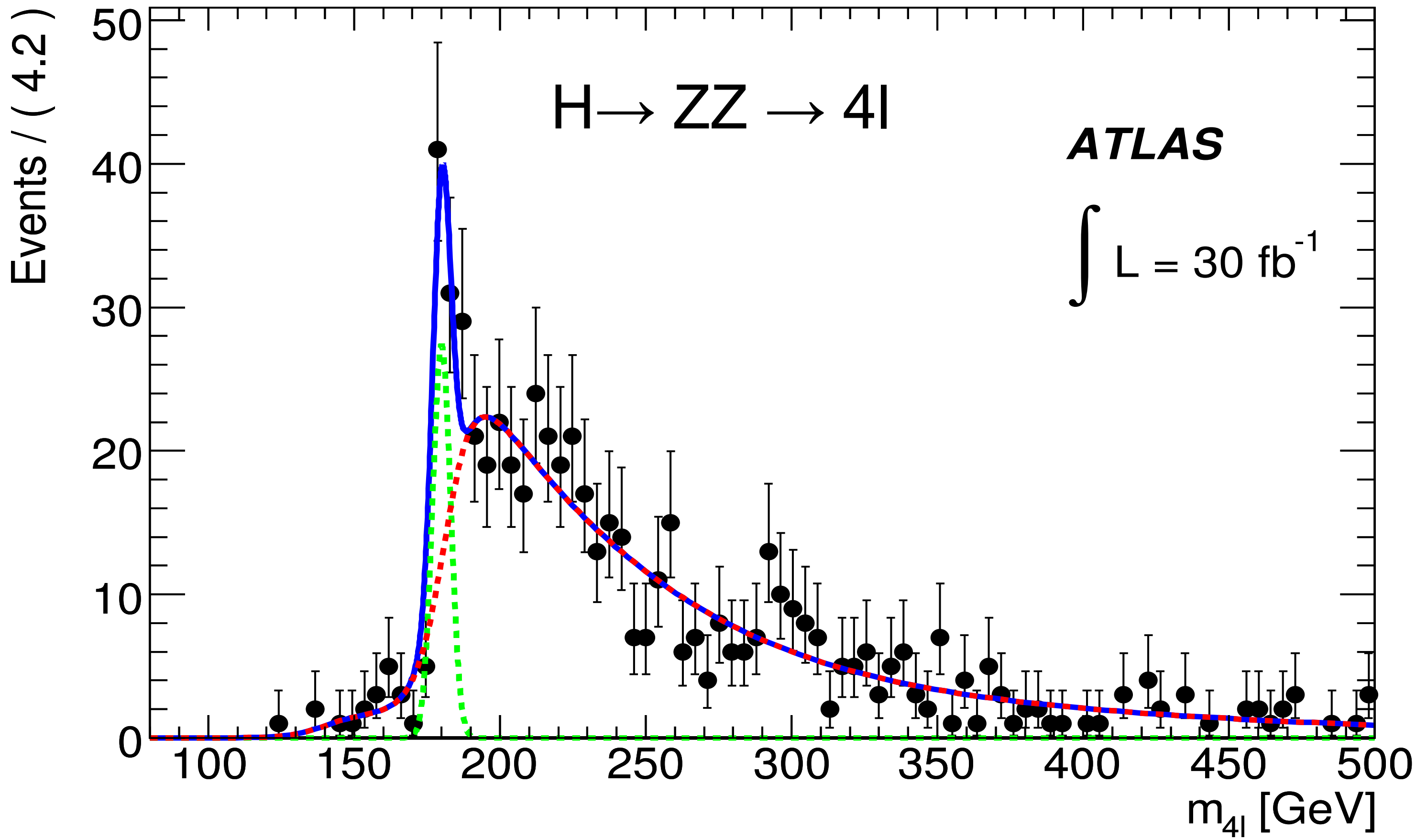
PHYSICAL REVIEW D **79,** 112002 (2009)





$$\frac{d\sigma}{dm_{jj}} = p_0(1-x)^{p_1}/x^{p_2+p_3\cdot\ln(x)}, \qquad x = m_{jj}/\sqrt{s},$$

The Effective Model Narrative

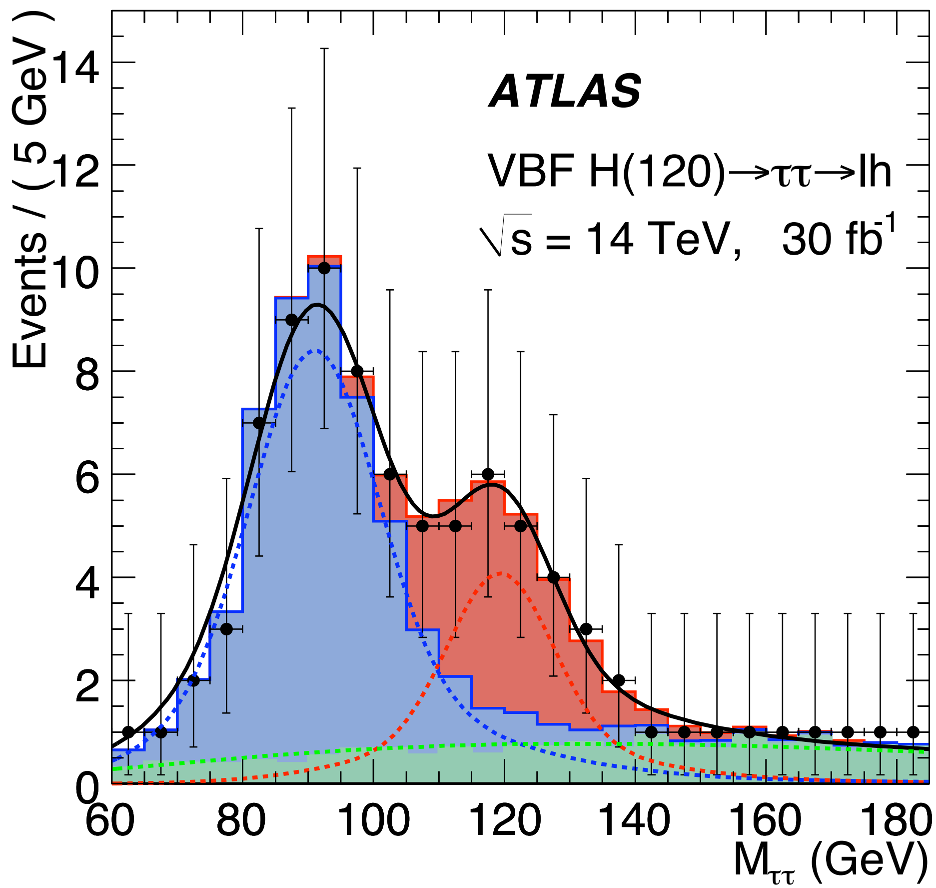


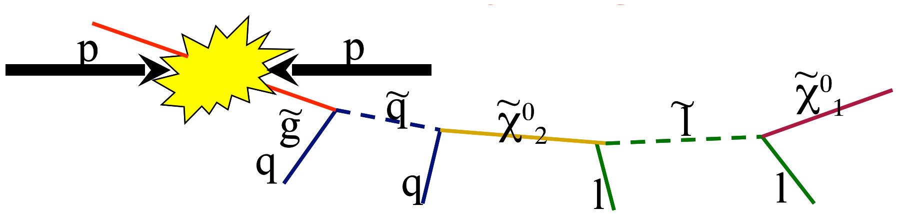
$$f(m_{ZZ}) = \frac{p0}{(1+e^{\frac{p6-m_{ZZ}}{p7}})(1+e^{\frac{m_{ZZ}-p8}{p9}})} + \frac{p1}{(1+e^{\frac{p2-m_{ZZ}}{p3}})(1+e^{\frac{p4-m_{ZZ}}{p5}})}$$

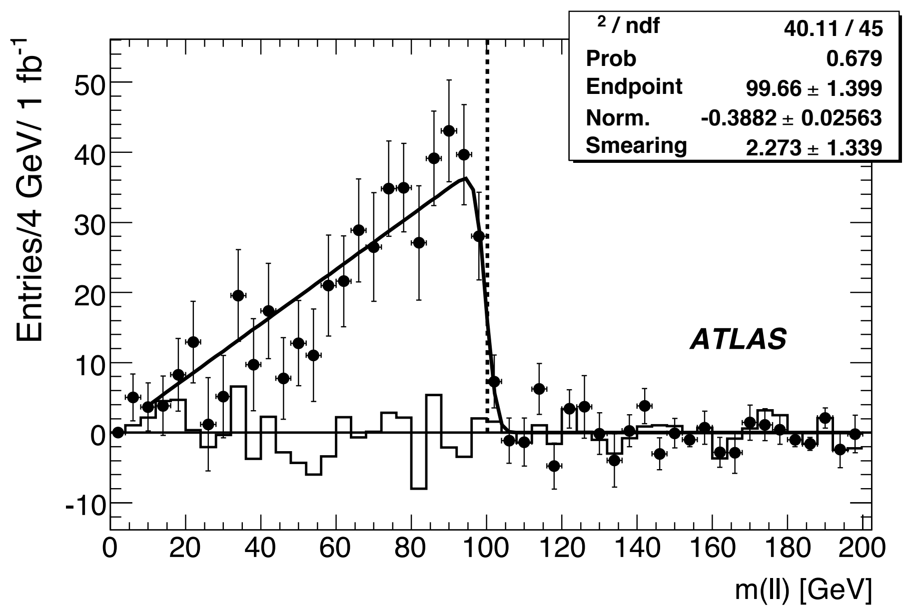
The Effective Model Narrative

Sometimes the effective model comes from a convincing narrative

- convolution of detector resolution with known distribution
  - Ex: MissingET resolution propagated through $M_{\tau\tau}$ in collinear approximation
  - Ex: lepton resolution convoluted with triangular M_{ll} distribution



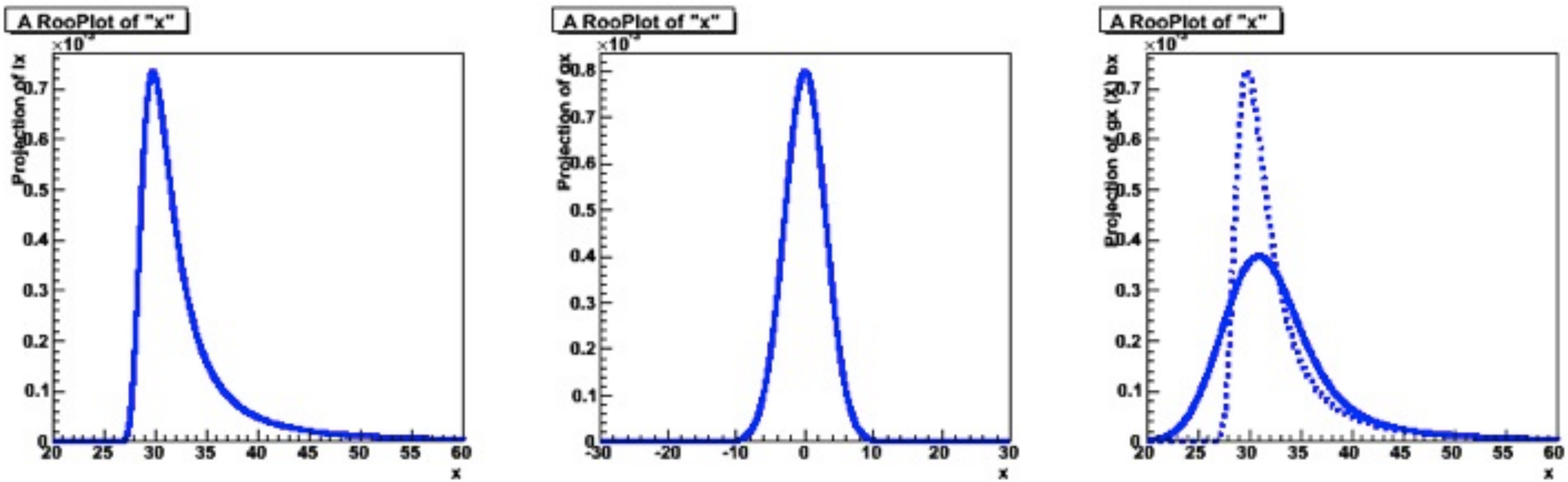




Tools for building effective models

- RooFit's convolution PDFs can aid in building more effective models with a more convincing narrative

```
// Construct landau (x) gauss (10000 samplings 2nd order interpolation)
t.setBins(10000,"cache") ;
RooFFTConvPdf lxg("lxg","landau (X) gauss",t,landau,gauss,2) ;
```



Wouter Verkerke, NIKHEF

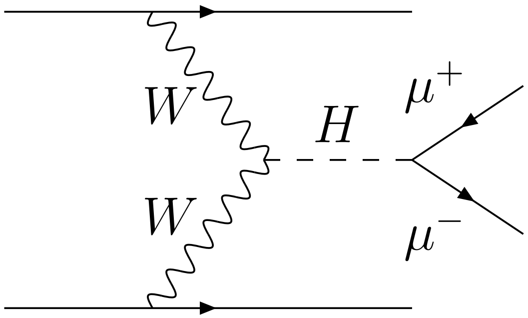
The parametrized response narrative

Center for Cosmology and Particle Physics

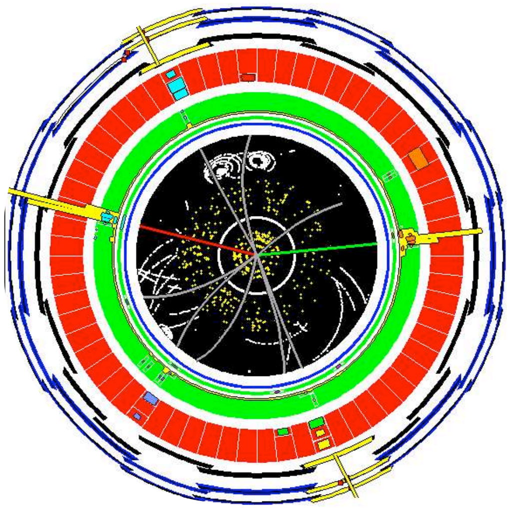
The Matrix-Element technique is conceptually similar to the simulation narrative, but the detector response is parametrized.

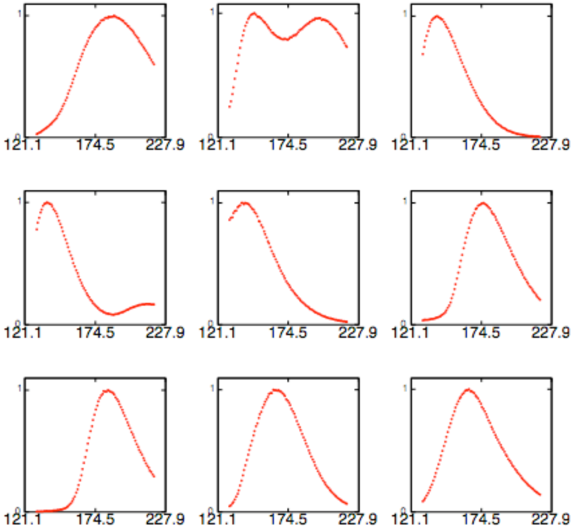
- Doesn't require building parametrized PDF by interpolating between non-parametric templates.

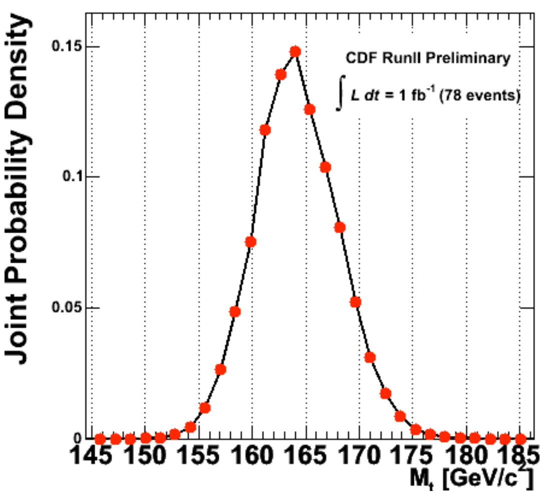
$$L(x|H_0) =$$



$\oplus$







The parametrized response narrative

The Matrix-Element technique is conceptually similar to the simulation narrative, but the detector response is parametrized.

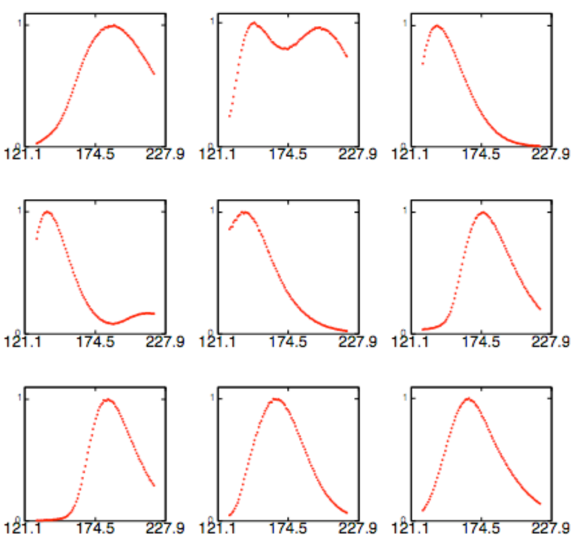
- Doesn’t require building parametrized PDF by interpolating between non-parametric templates.

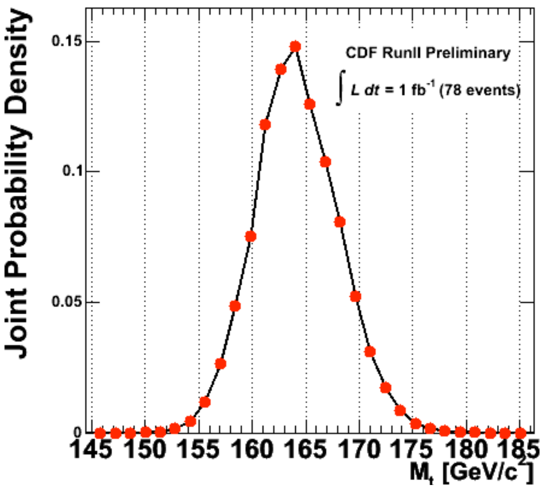
$$P(\mathbf{x}|M_t) = \frac{1}{N}\int d\Phi \; |\mathcal{M}_{t\bar{t}}(p;M_t)|^2 \prod_{jets} f(p_i, j_i) f_{PDF}(q_1) f_{PDF}(q_2)$$

Phase-space Integral

Matrix Element

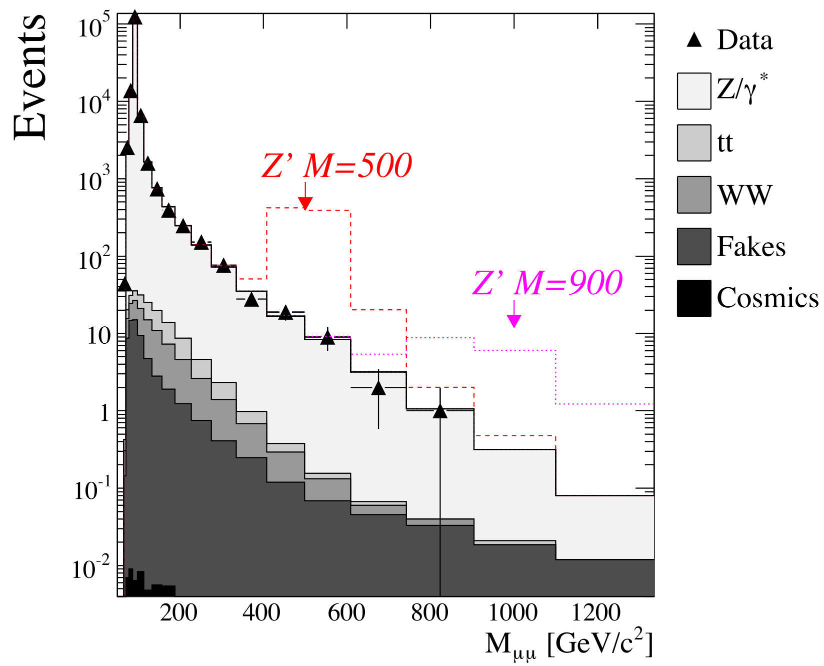
Transfer Functions

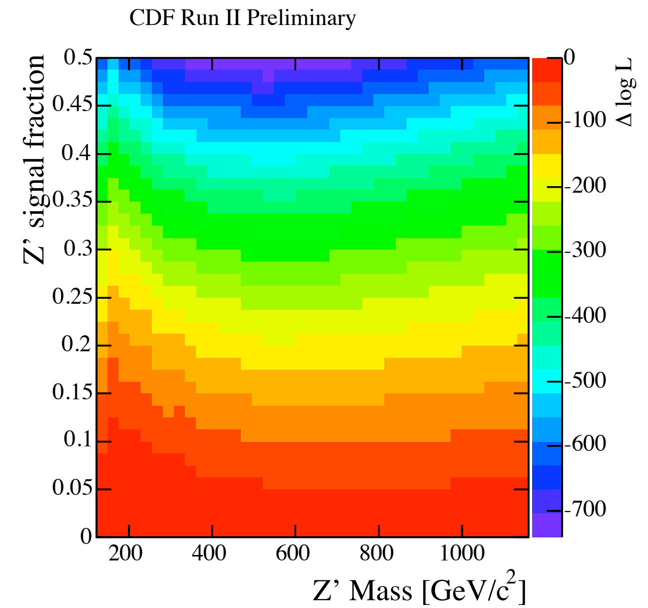


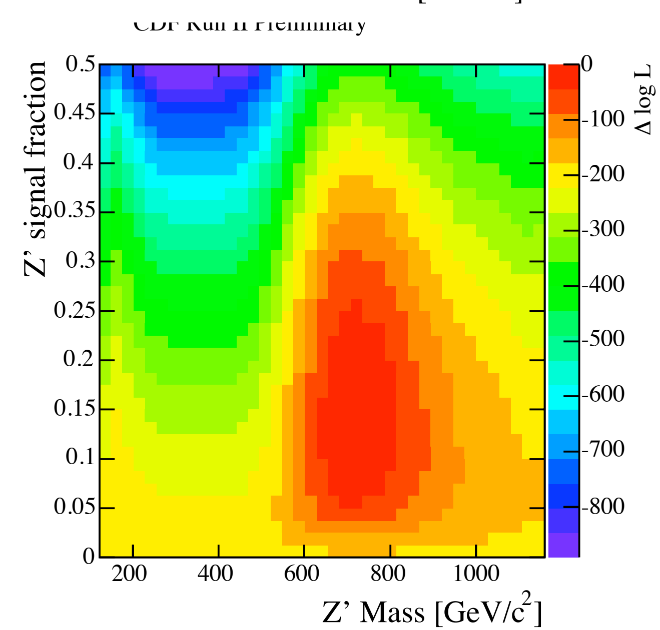


Example: CDF Z' → μμ

"a matrix element based likelihood providing an approximately 20% relative increase in cross section sensitivity at large Z′ mass"







still stronger than ATLAS & CMS

TABLE I: Mass limits on specific spin-1 Z' models [12] in data with 4.6 fb^{-1} of integrated luminosity at 95% confidence level.

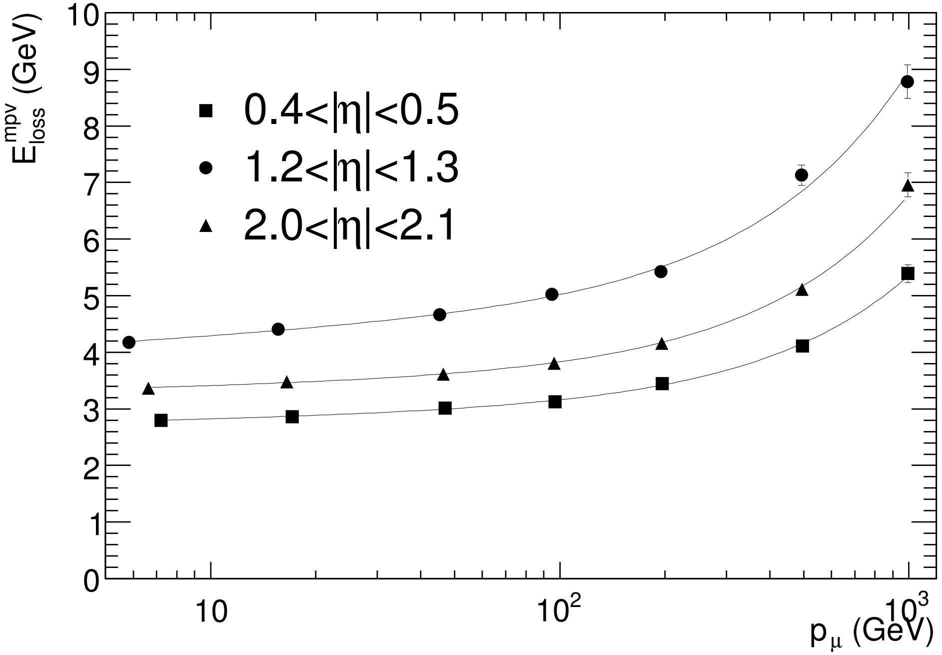
Model	Z'_l	Z'_{sec}	Z'_N	Z'_ψ	Z'_χ	Z'_η	Z'_{SM}
Mass Limit (GeV/c^2)	817	858	900	917	930	938	1071

Examples of parametrized response

While we often see the parametrized response as overly simplistic, the parametrizations are often based on some deeper understanding

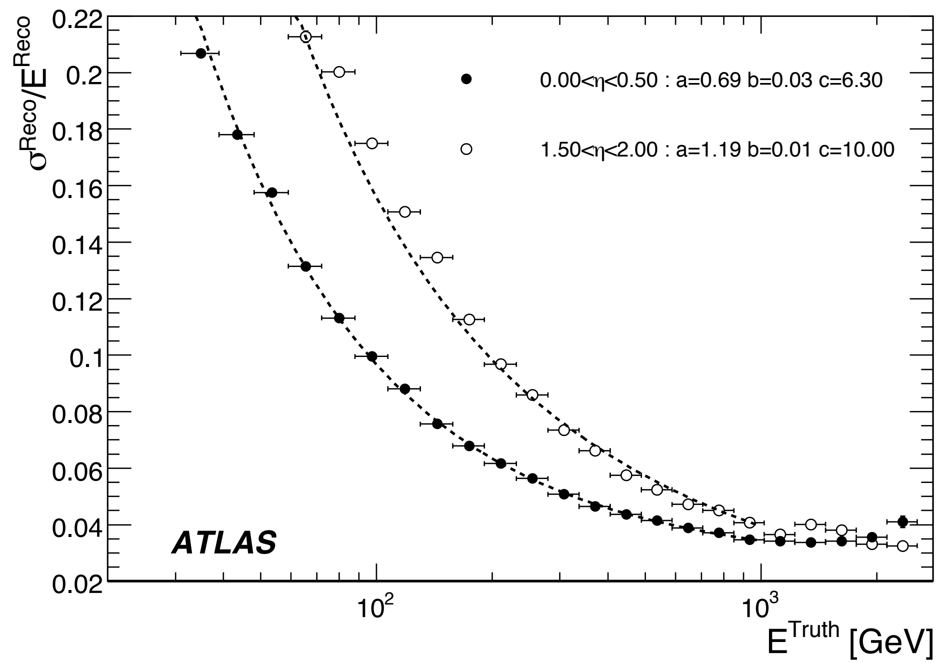
- and parameters can often be measured in data with in situ calibration strategies. No reason we can't propagate uncertainty to next stage.

Muon Energy Loss (Landau)



$$E^{\mathrm{mpv}}_{\mathrm{loss}}(p_\mu) = a^{\mathrm{mpv}}_0 + a^{\mathrm{mpv}}_1 \ln p_\mu + a^{\mathrm{mpv}}_2 p_\mu$$

Jet Resolution

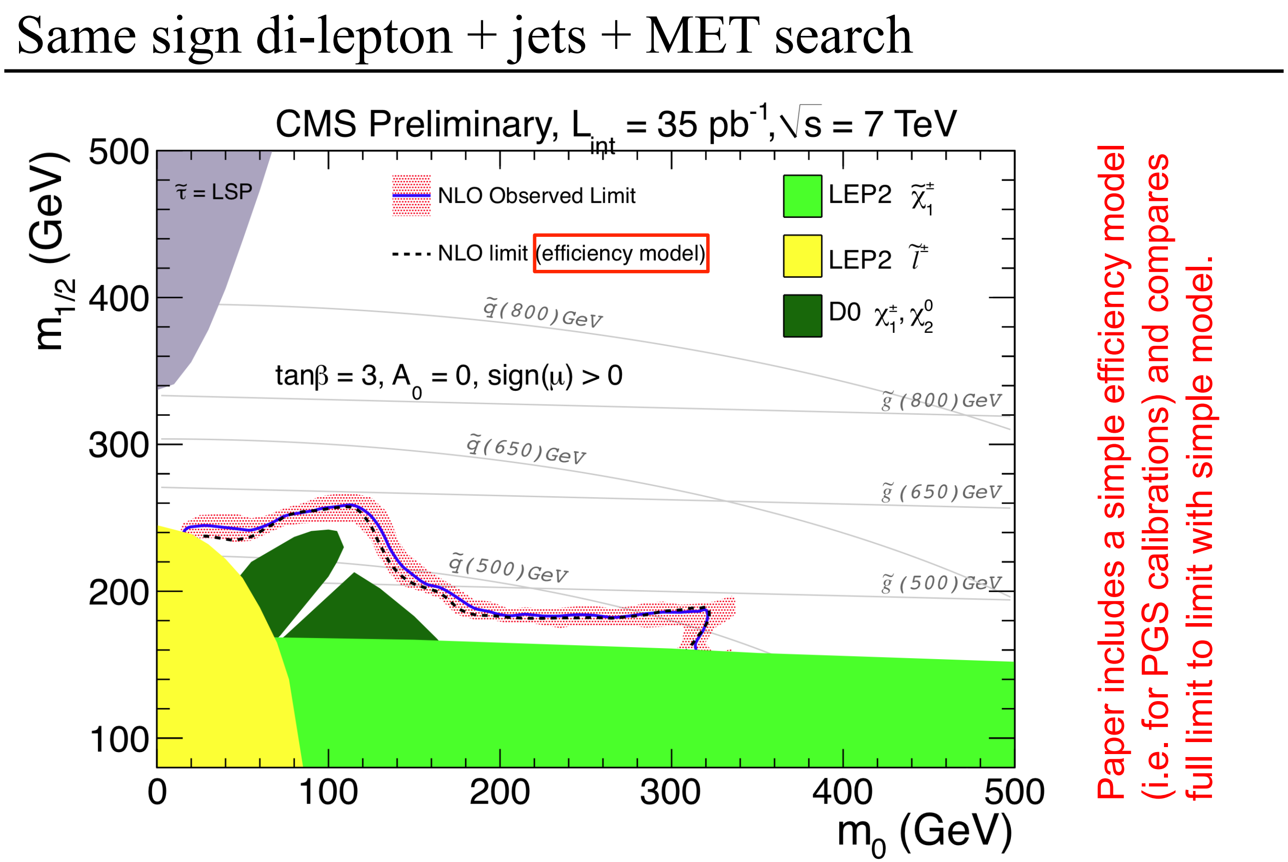


$$\frac{\sigma}{E} = \frac{a}{\sqrt{E\ (\mathrm{GeV})}} \oplus b \oplus \frac{c}{E}.$$

Fast Simulation

Fast simulations based on parametrized detector response are very useful and can often be tuned to perform quite well in a specific analysis context

- For example: tools like PGS, Delphis, ATLFAST, ...



Paper includes a simple efficiency model (i.e. for PGS calibrations) and compares full limit to limit with simple model.

CMS SUSY Results, D. Stuart, April 2011, SUSY Recast, UC Davis 36

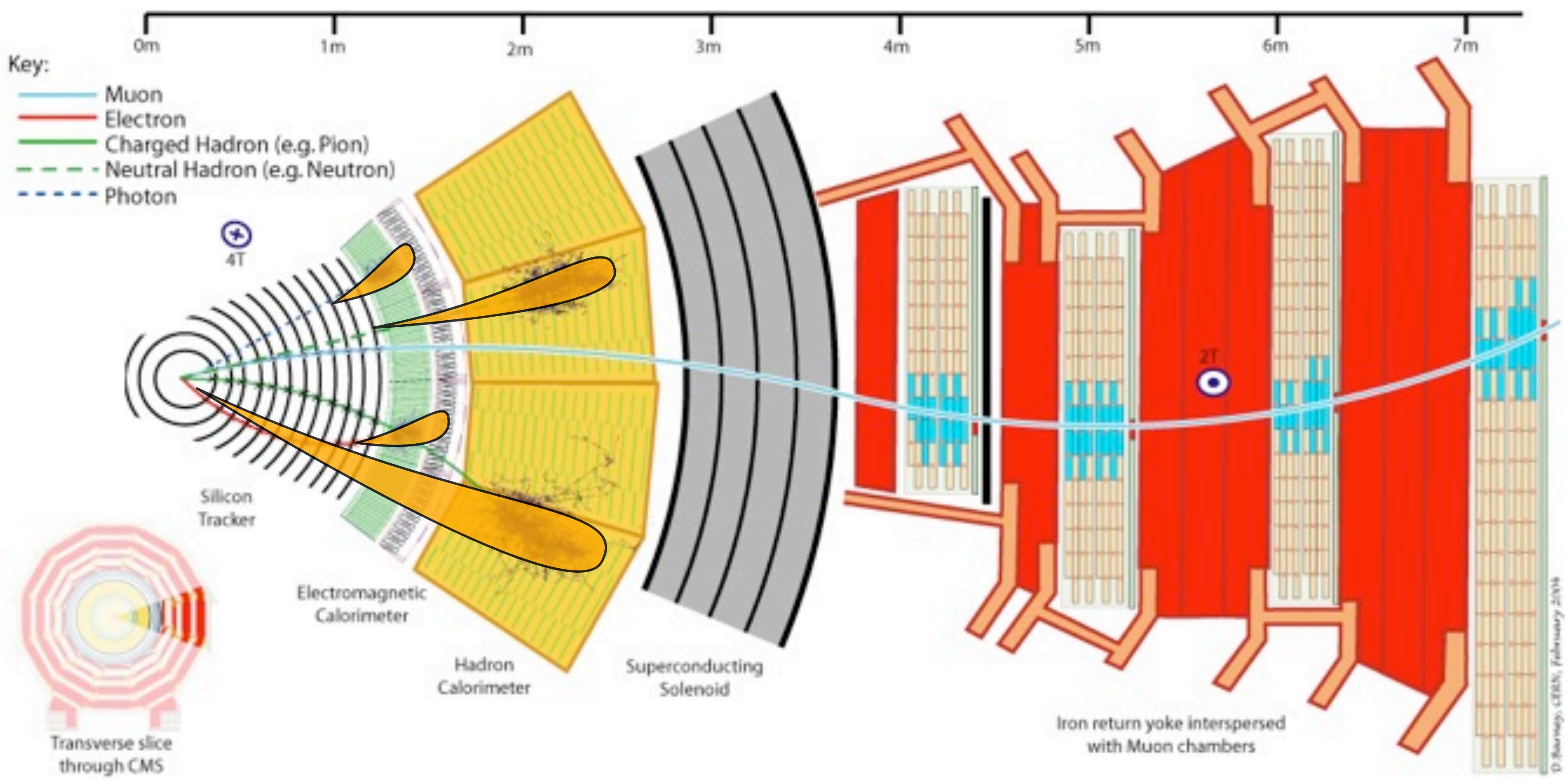
Fast Simulation

Fast simulations based on parametrized detector response are very useful and can often be tuned to perform quite well in a specific analysis context

- For example: tools like PGS, Delphis, ATLFAST, ...

But these tools still use accept/reject Monte Carlo.

- Would be much more useful if the parmaetrized detector response could be used as a transfer function in Matrix-Element approach



Narrative styles

The Monte Carlo Simulation narrative (MC narrative)

- each stage is an accept/reject Monte Carlo based on P(out|in) of some microscopic process like parton shower, decay, scattering
- PDFs built from non-parametric estimator like histograms or kernel estimation
  - need to supplement with interpolation procedures to incorporate systematics
  - smearing approach fundamentally Bayesian
- **pros:** most detailed understanding of micro-physics
- **cons:** computationally demanding, loose analytic scaling properties, relies on accuracy of simulation
- **new ideas:** improved interpolation, Radford Neal's machine learning, "design of experiments"

The Data-driven narrative

- independent data sample that either acts as a proxy for some process or can be transformed to do so
- **pros**: nature includes "all orders", uses real detector
- **cons**: extrapolation from control region to signal region requires assumptions, introduces systematic effects. Appropriate transformation may depend on many variables, which becomes impractical

Narrative styles

Effective modeling narrative

- parametrized functional form: eg. Gaussian, falling exponential para polynomial fit to distribution, etc.
- **pros**: fast, has analytic scaling, parametric form may be well justified (eg. phase space, propagation of errors, convolution)
- **cons**: approximate, parametric form may be ad hoc (eg. polynomial from)
- new ideas: using non-parametric statistical methods

Parametrized detector response narrative (eg. kinematic fitting, Matrix-Element method, ~fast simulation)

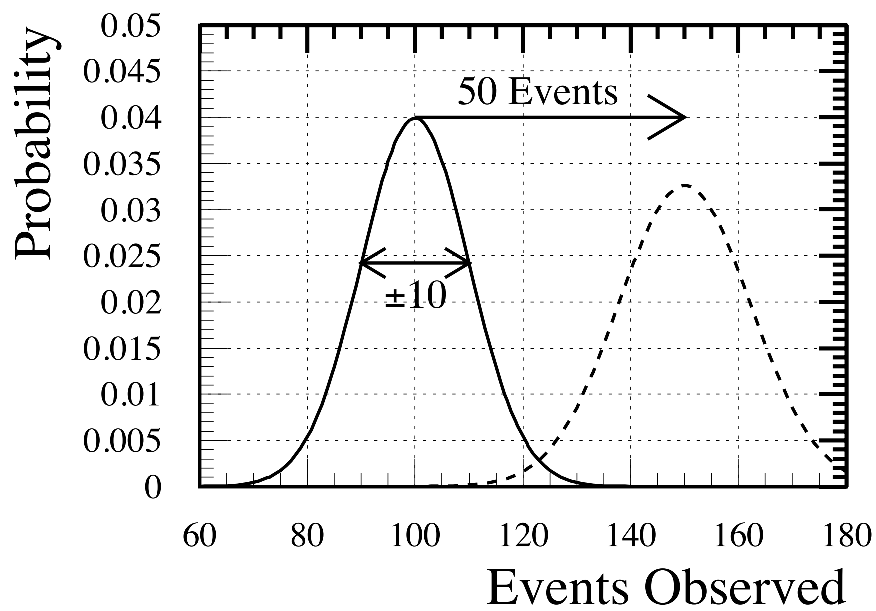
- **pros**: fast, maintains analytic scaling, response usually based on good understanding of the detector, possible to incorporate some types of uncertainty in the response analytically, can evaluate P(out|in) for arbitrary out,in.
- **cons**: approximate, best parametrized detector response is often not available in convenient form
- new ideas: fast simulation is typically parametrized, but we use it in an accept/reject framework (see Geant5)

Hypothesis Testing

Hypothesis testing

One of the most common uses of statistics in particle physics is Hypothesis Testing (e.g. for discovery of a new particle)

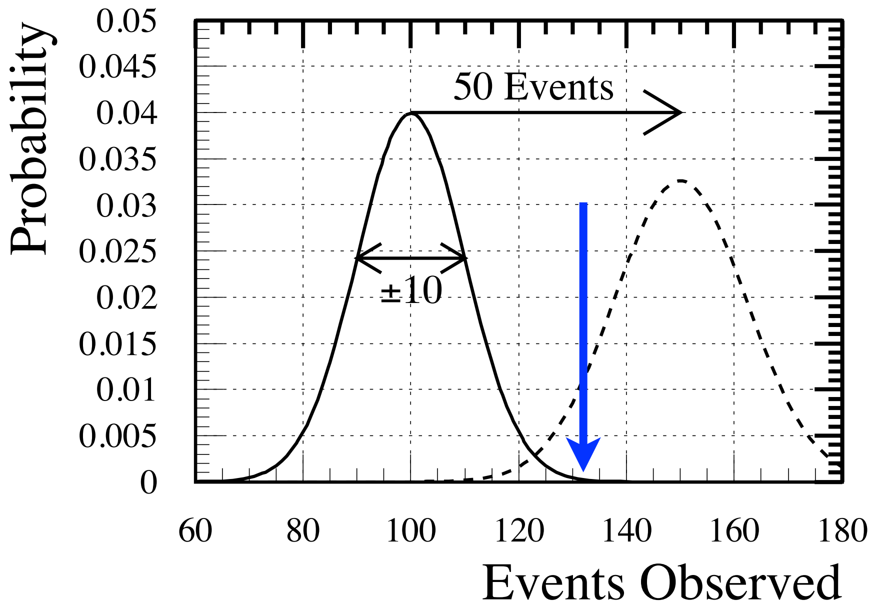
- assume one has pdf for data under two hypotheses:
  - Null-Hypothesis, H_0: eg. background-only
  - Alternate-Hypothesis H_1: eg. signal-plus-background
- one makes a measurement and then needs to decide whether to **reject** or **accept** H_0



Hypothesis testing

One of the most common uses of statistics in particle physics is Hypothesis Testing (e.g. for discovery of a new particle)

- assume one has pdf for data under two hypotheses:
  - Null-Hypothesis, H_0: eg. background-only
  - Alternate-Hypothesis H_1: eg. signal-plus-background
- one makes a measurement and then needs to decide whether to **reject** or **accept** H_0



Hypothesis testing

Before we can make much progress with statistics, we need to decide what it is that we want to do.

- first let us define a few terms:
  - Rate of Type I error α
  - Rate of Type II β
  - Power = $1 - \beta$

		Actual condition	
		Guilty	**Not guilty**
Decision	**Verdict of 'guilty'**	True Positive	False Positive (i.e. guilt reported unfairly) **Type I error**
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) **Type II error**	True Negative

Hypothesis testing

Before we can make much progress with statistics, we need to decide what it is that we want to do.

- first let us define a few terms:
  - Rate of Type I error α
  - Rate of Type II β
  - Power = $1-\beta$

		Actual condition	
		Guilty	Not guilty
Decision	Verdict of 'guilty'	True Positive	False Positive (i.e. guilt reported unfairly) **Type I error**
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) **Type II error**	True Negative

Treat the two hypotheses asymmetrically

- the Null is special.
  - Fix rate of Type I error, call it “the size of the test”

Hypothesis testing

Before we can make much progress with statistics, we need to decide what it is that we want to do.

- first let us define a few terms:
  - Rate of Type I error α
  - Rate of Type II β
  - Power = $1 - \beta$

		Actual condition	
		Guilty	**Not guilty**
Decision	**Verdict of 'guilty'**	True Positive	False Positive (i.e. guilt reported unfairly) **Type I error**
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) **Type II error**	True Negative

Treat the two hypotheses asymmetrically

- the Null is special.
  - Fix rate of Type I error, call it "the size of the test"

Now one can state "a well-defined goal"

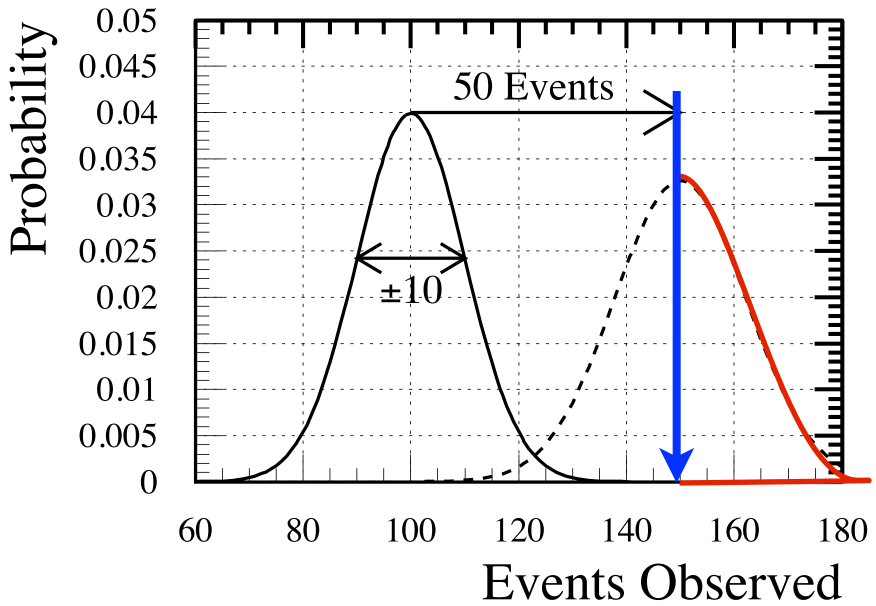
- Maximize power for a fixed rate of Type I error

The idea of a “5σ“ discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy



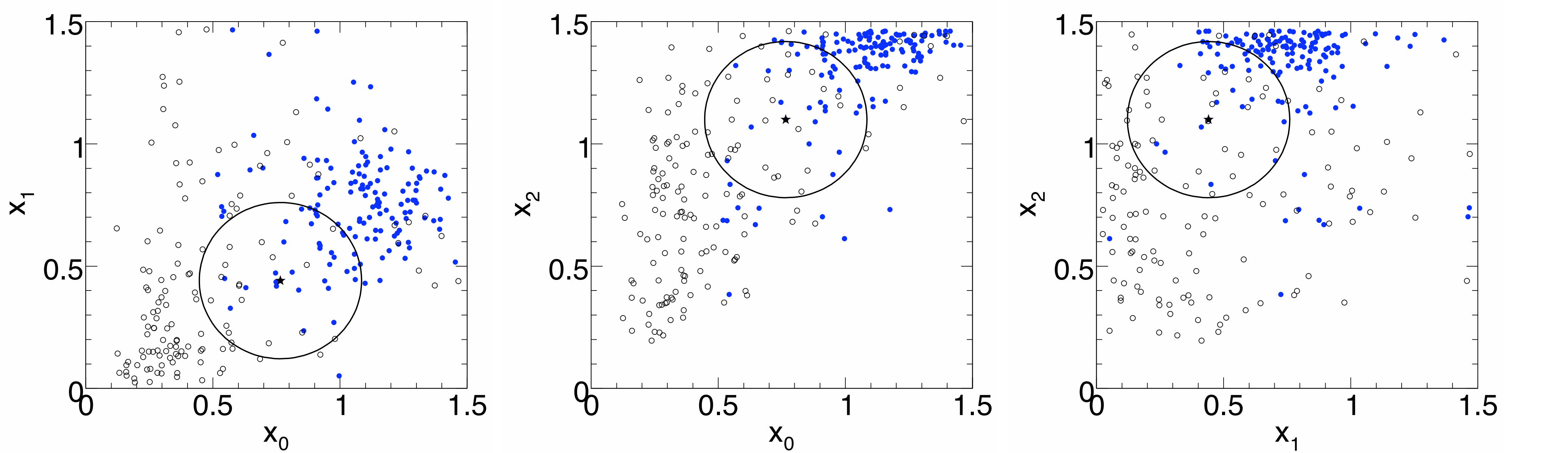
Hypothesis testing

The idea of a "5σ" discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy

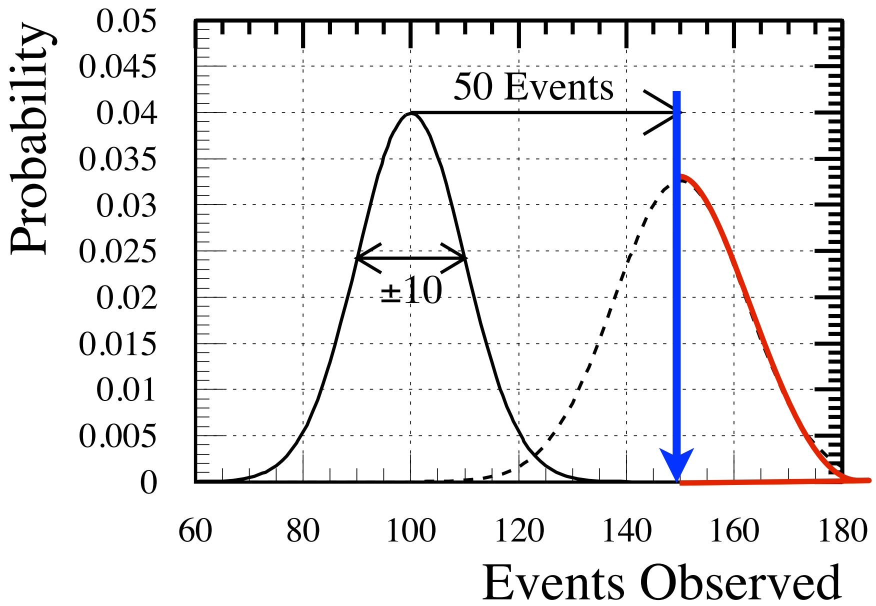


The idea of a "5σ" discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy



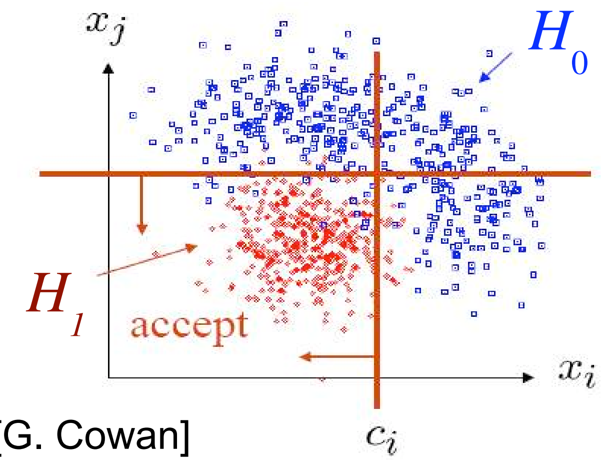
Hypothesis testing

The idea of a "5σ" discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy



[G. Cowan]

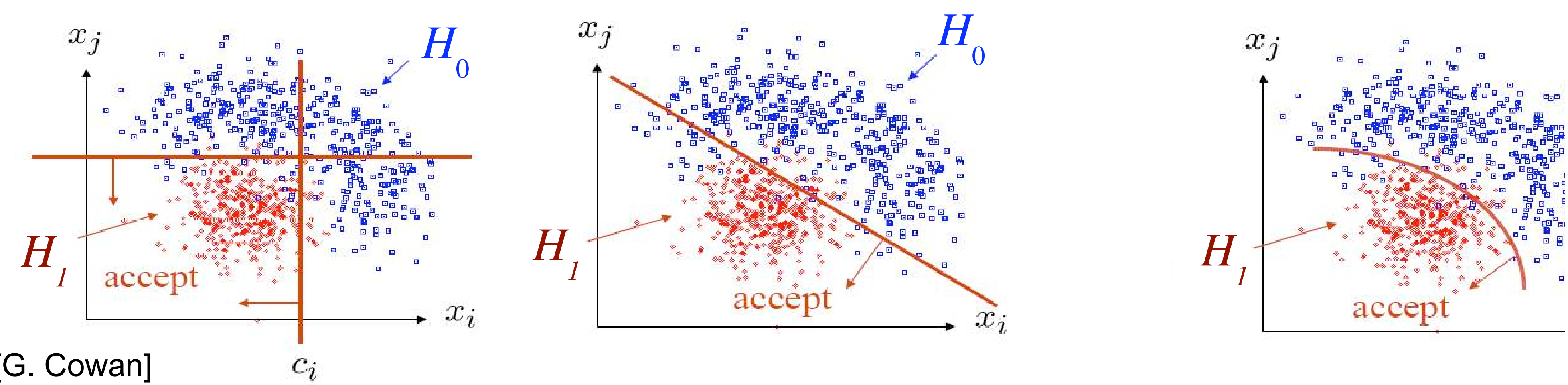
Hypothesis testing

The idea of a “5σ“ discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy



[G. Cowan]

The Neyman-Pearson Lemma

In 1928-1938 Neyman & Pearson developed a theory in which one must consider competing Hypotheses:

- the Null Hypothesis H_0 (background only)
- the Alternate Hypothesis H_1 (signal-plus-background)

Given some probability that we wrongly reject the Null Hypothesis

$$\alpha = P(x \notin W | H_0)$$

(Convention: if data falls in W then we accept H_0)

Find the region W such that we minimize the probability of wrongly accepting the H_0 (when H_1 is true)

$$\beta = P(x \in W | H_1)$$

The Neyman-Pearson Lemma

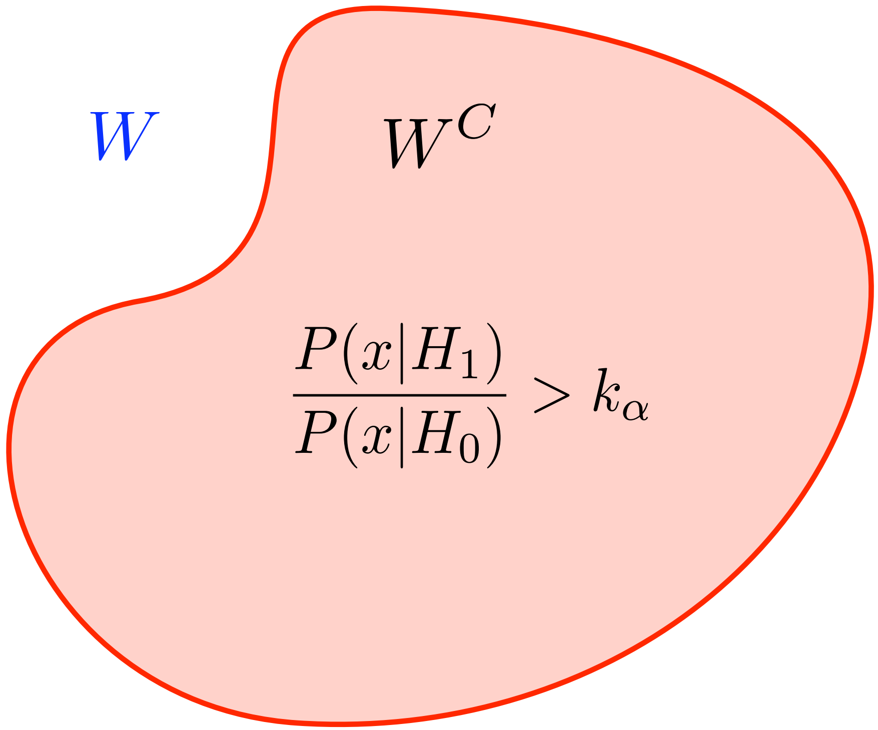
The region W that minimizes the probability of wrongly accepting H_0 is just a contour of the Likelihood Ratio

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

Any other region of the same size will have less power

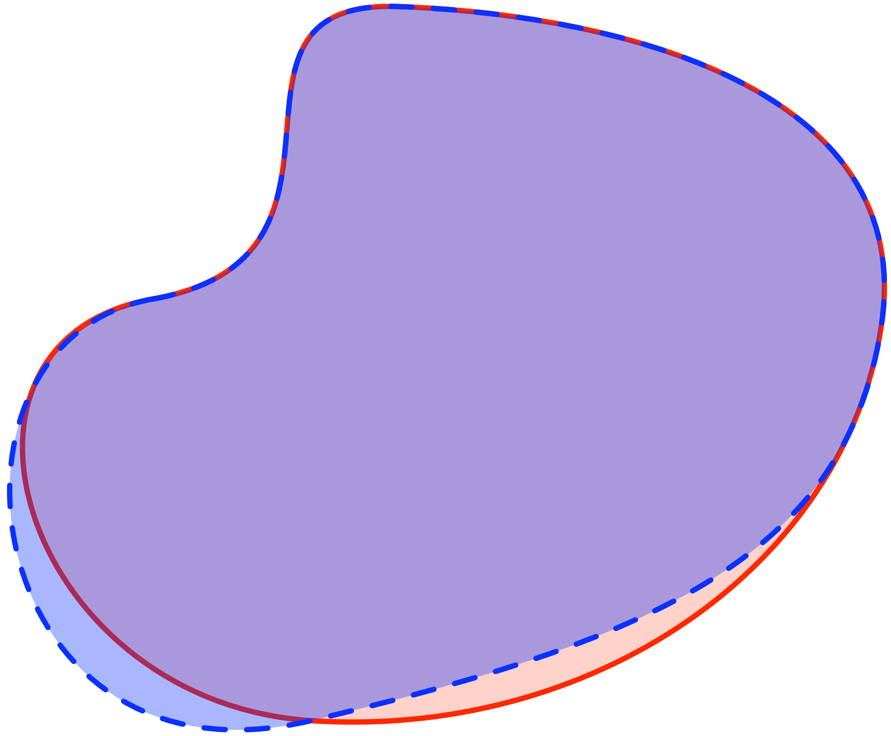
The likelihood ratio is an example of a Test Statistic, eg. a real-valued function that summarizes the data in a way relevant to the hypotheses that are being tested

A short proof of Neyman-Pearson



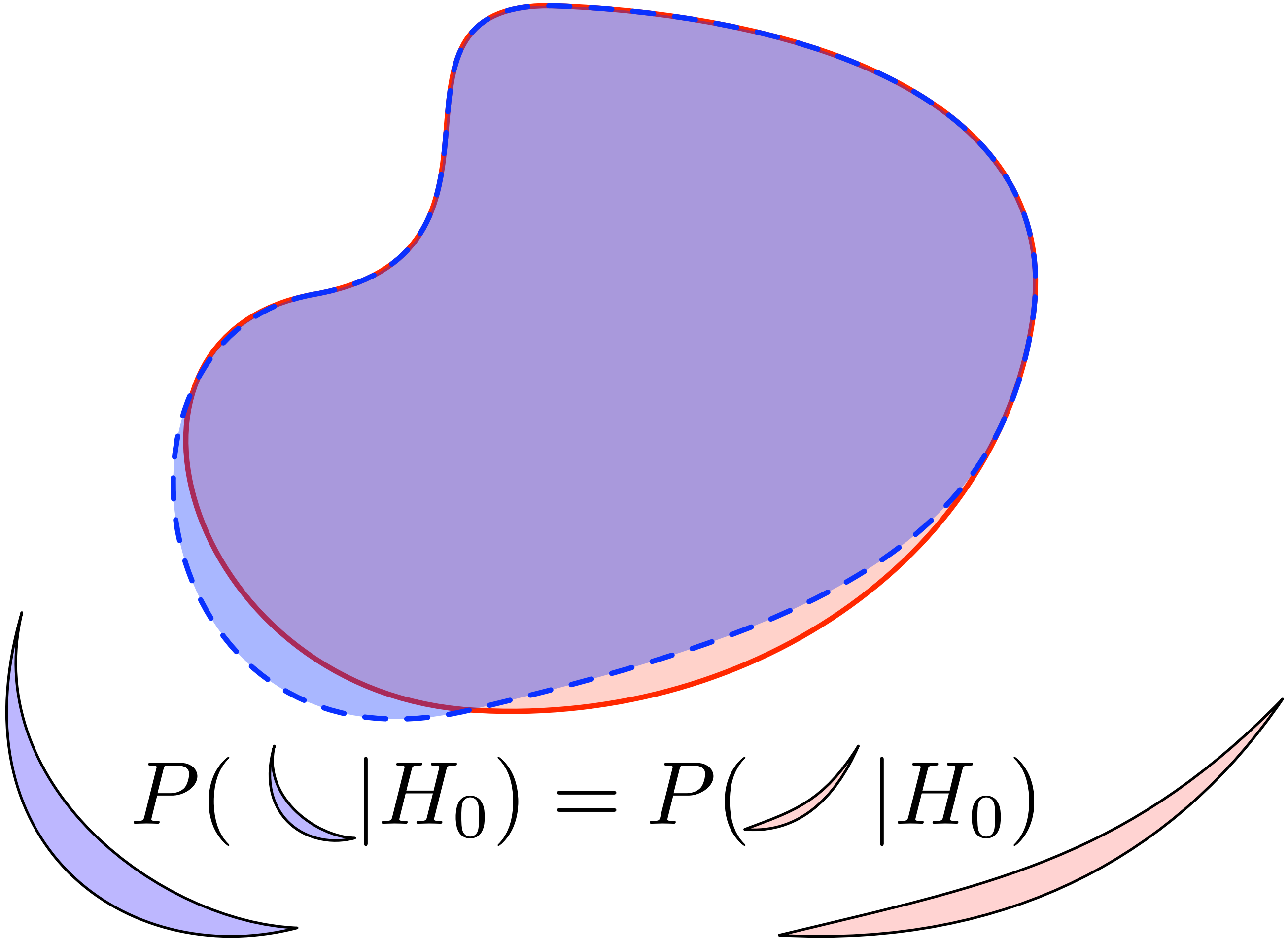
Consider the contour of the likelihood ratio that has size a given size (eg. probability under H_0 is 1-α)

A short proof of Neyman-Pearson



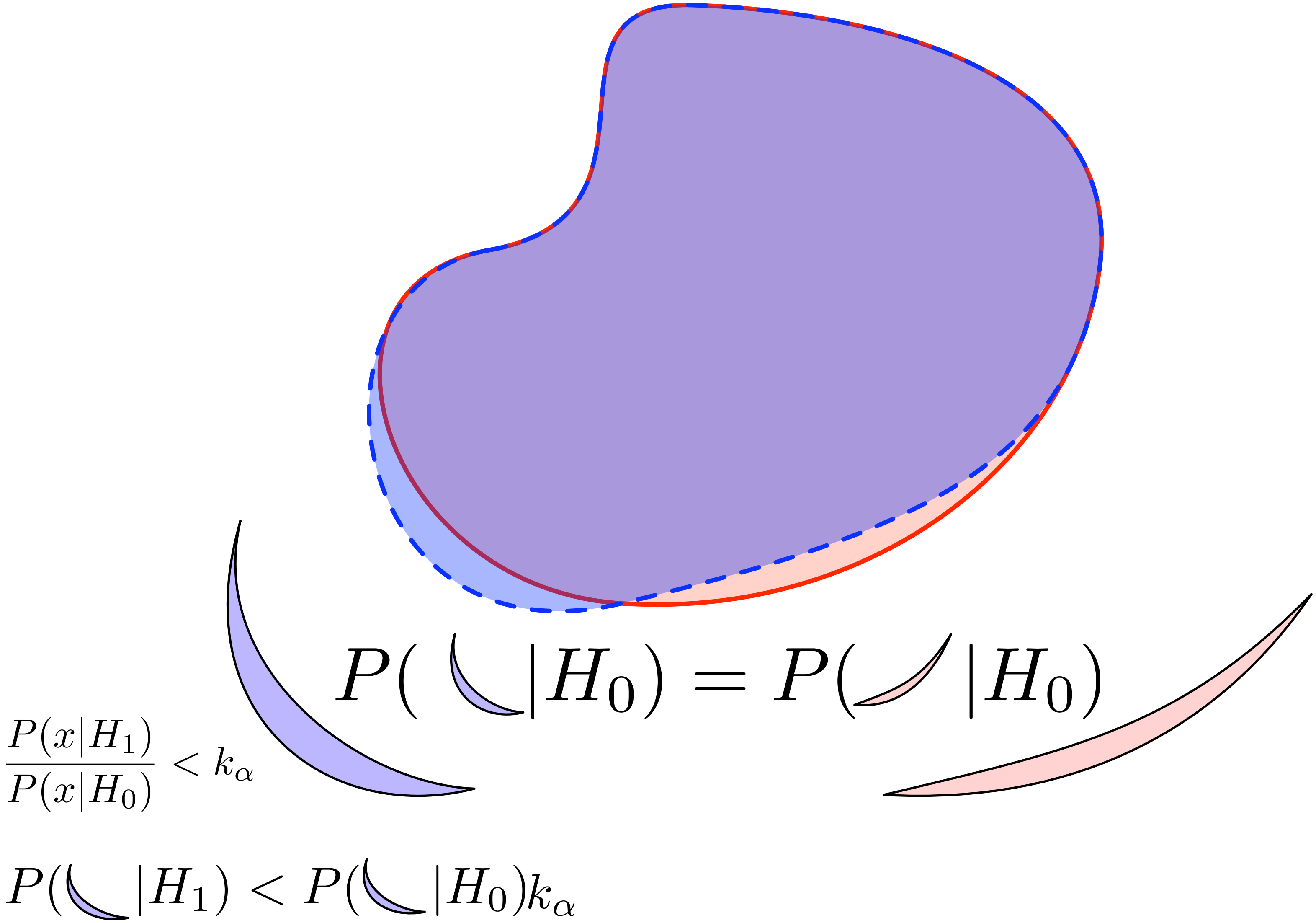
Now consider a variation on the contour that has the same size

A short proof of Neyman-Pearson



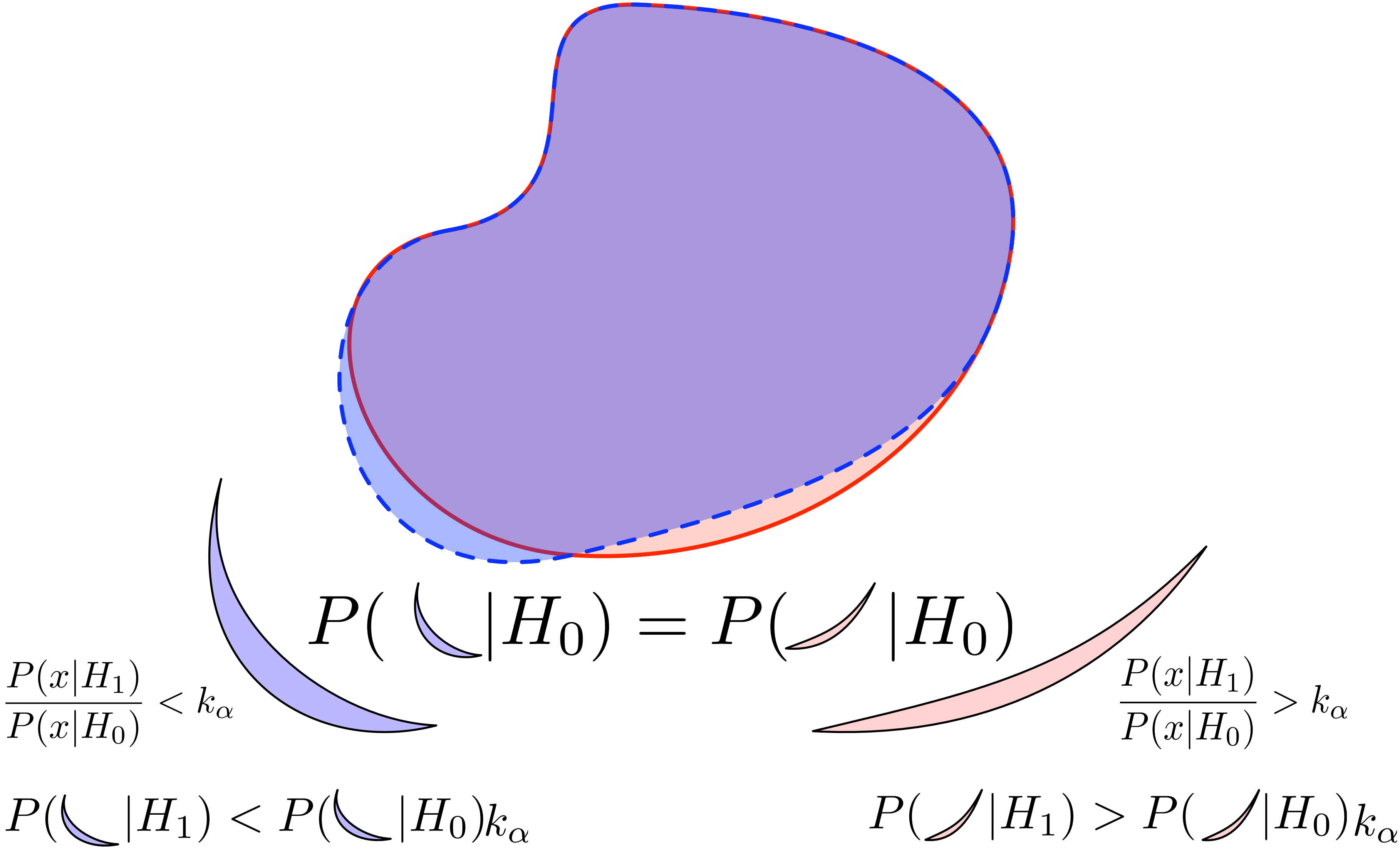
Now consider a variation on the contour that has the same size (eg. same probability under H_0)

A short proof of Neyman-Pearson



Because the new area is outside the contour of the likelihood ratio, we have an inequality

A short proof of Neyman-Pearson

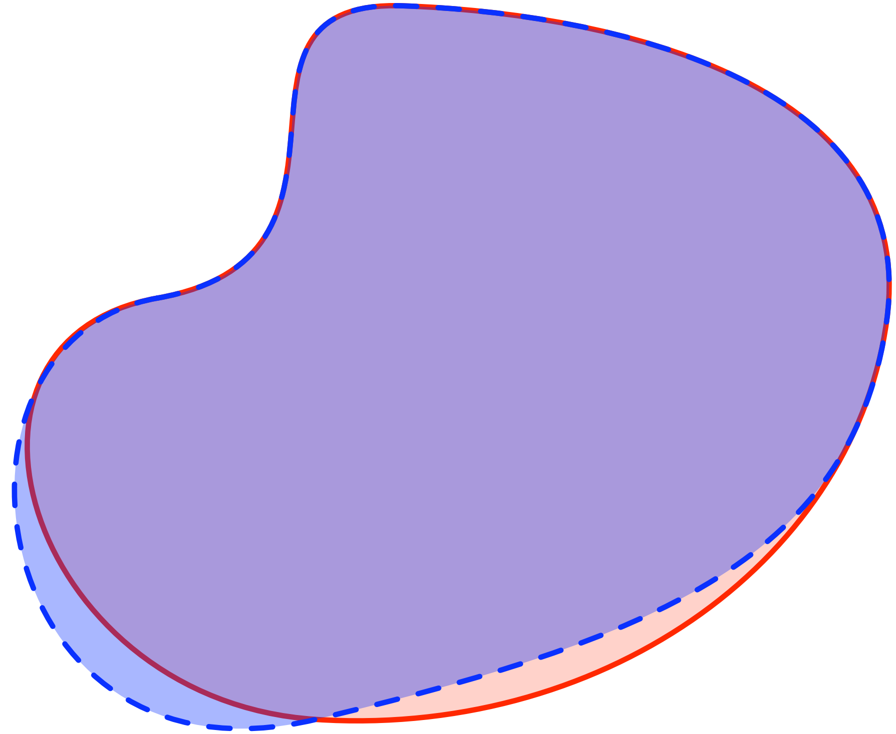


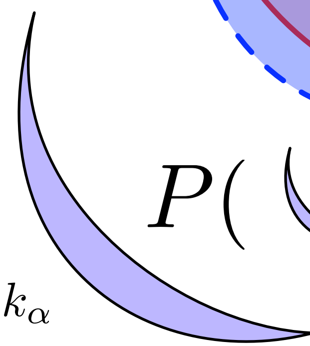
And for the region we lost, we also have an inequality

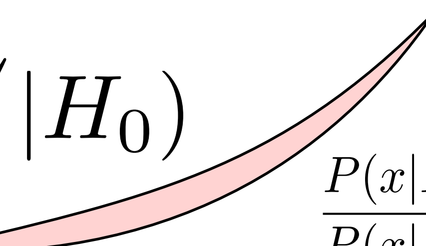
Together they give...

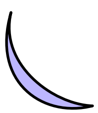
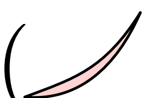
A short proof of Neyman-Pearson

CENTER FOR COSMOLOGY AND PARTICLE PHYSICS

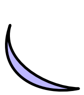
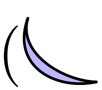
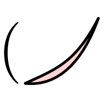
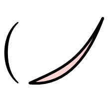


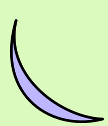
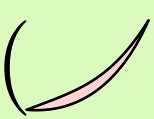




$P($  $|H_0) = P($  $|H_0)$

$$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha \qquad\qquad \frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

$P($  $|H_1) < P($  $|H_0)k_\alpha$ $\qquad$ $P($  $|H_1) > P($  $|H_0)k_\alpha$

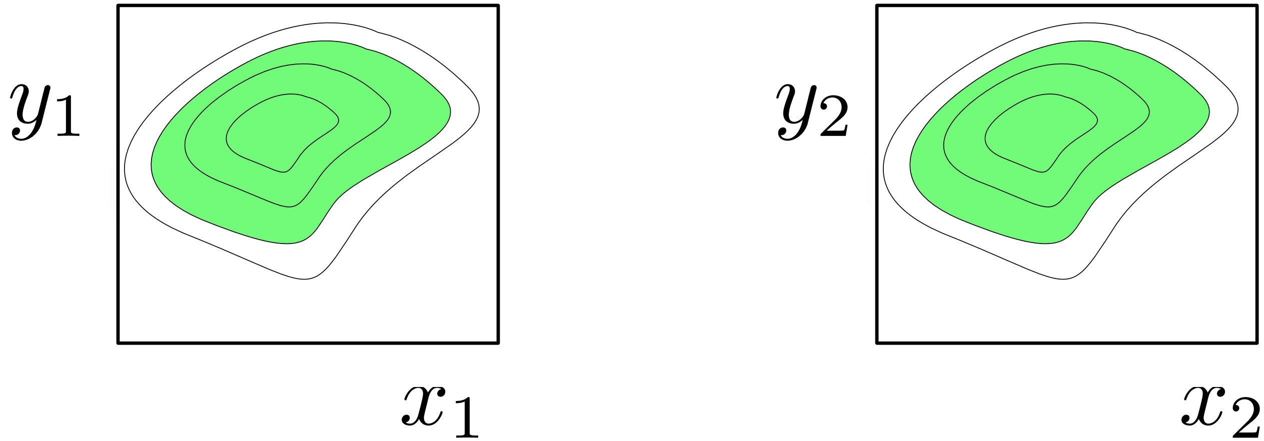
$P($  $|H_1) < P($  $|H_1)$

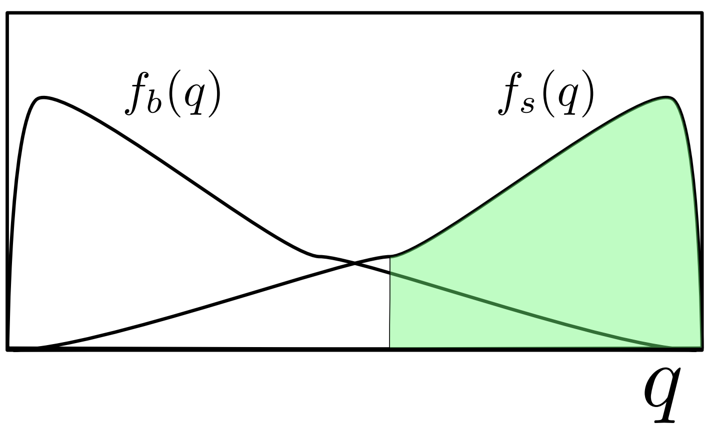
The new region region has less power.

2 discriminating variables

Often one uses the output of a neural network or multivariate algorithm in place of a true likelihood ratio.

- That's fine, but what do you do with it?
- If you have a fixed cut for all events, this is what you are doing:





$$q = \ln Q = -s + \ln\left(1 + \frac{s f_s(x,y)}{b f_b(x,y)}\right)$$

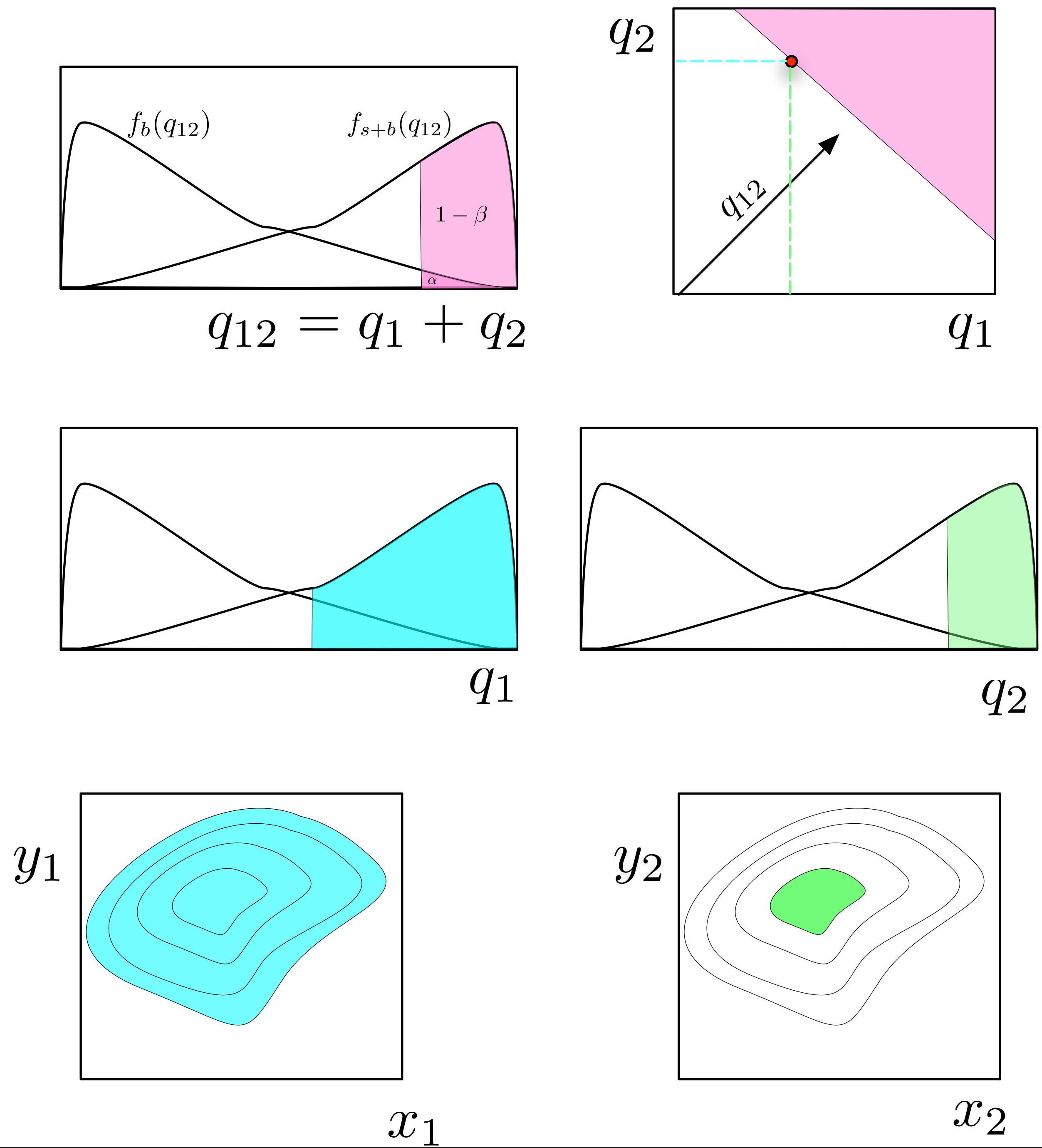
Experiments vs. Events

CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS

Ideally, you want to cut on the likelihood ratio for your experiment

- equivalent to a sum of log likelihood ratios

Easy to see that includes experiments where one event had a high LR and the other one was relatively small

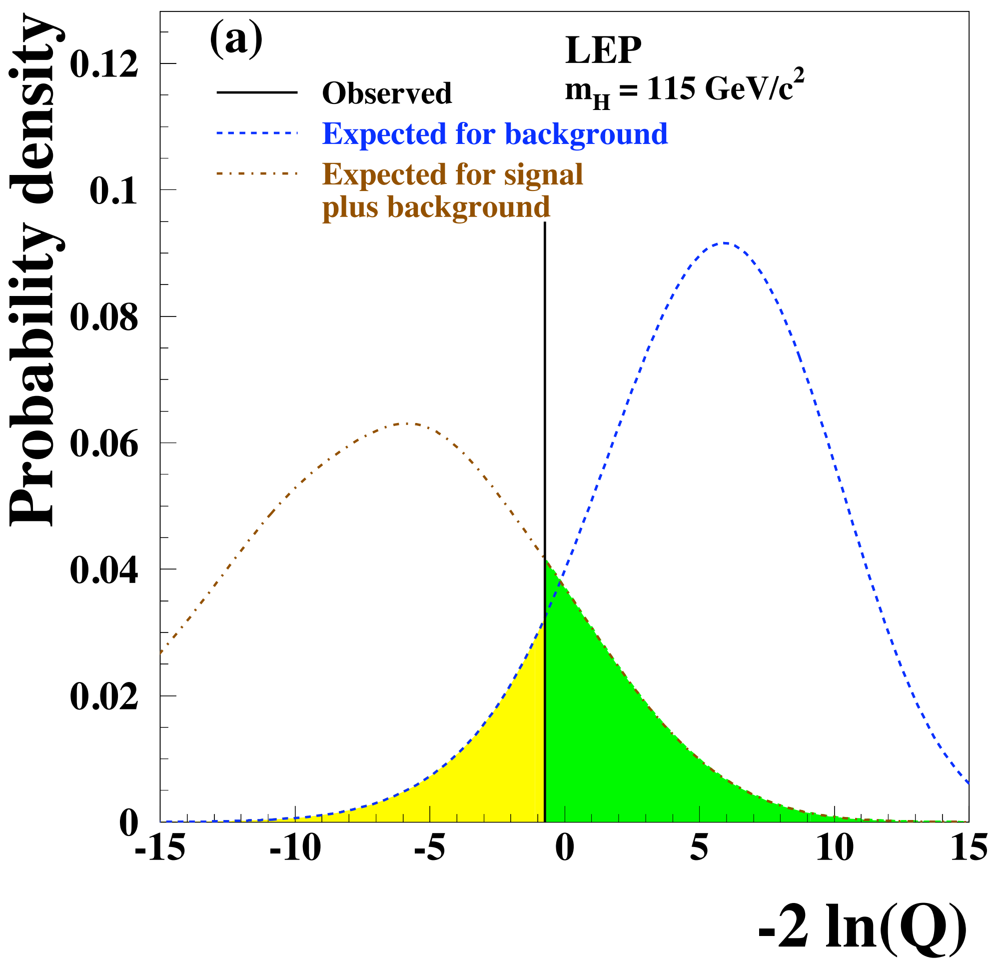


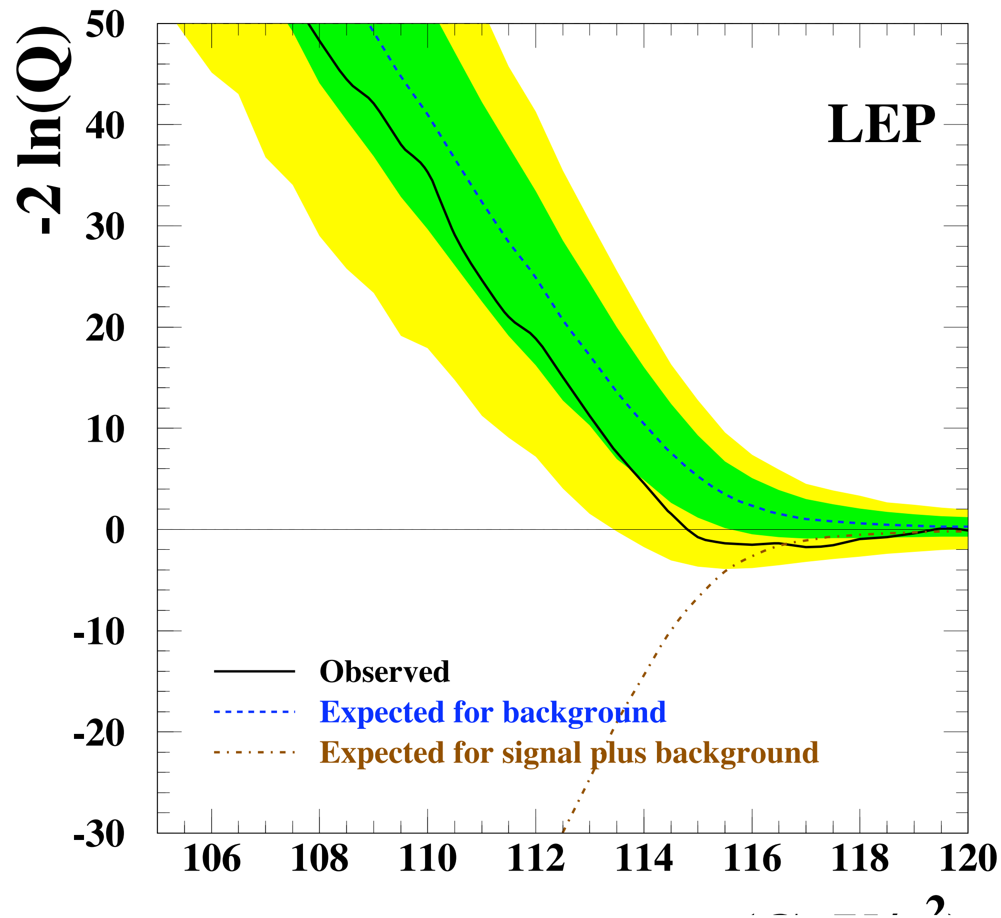
LEP Higgs

In that case:

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

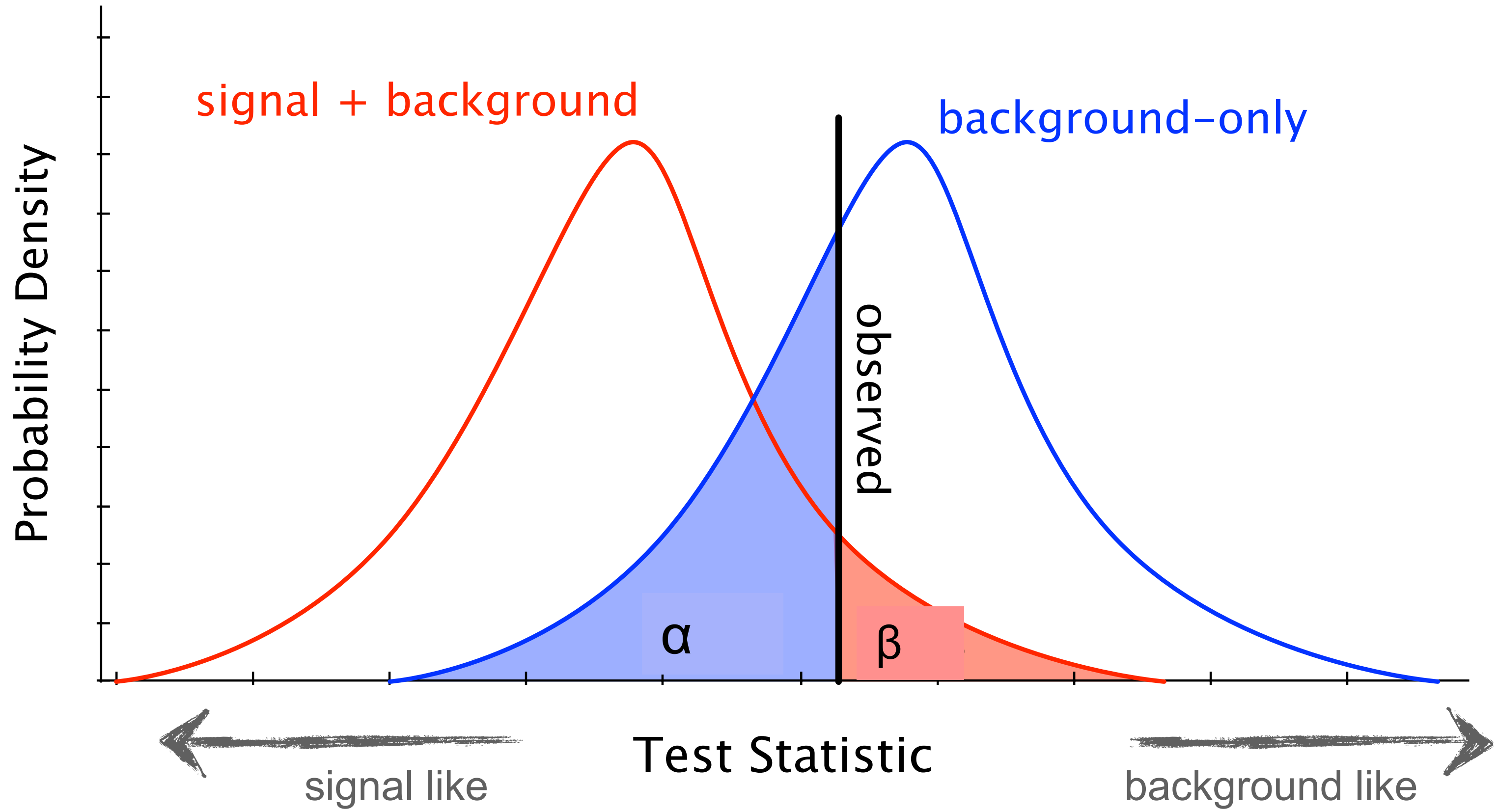
$$q = \ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$





The Test Statistic and its distribution

To get a feel for the different approaches, consider this schematic diagram



The "**test statistic**" is a single number that quantifies the entire experiment, it could just be number of events observed, but often its more sophisticated, like a likelihood ratio. What test statistic do we choose?

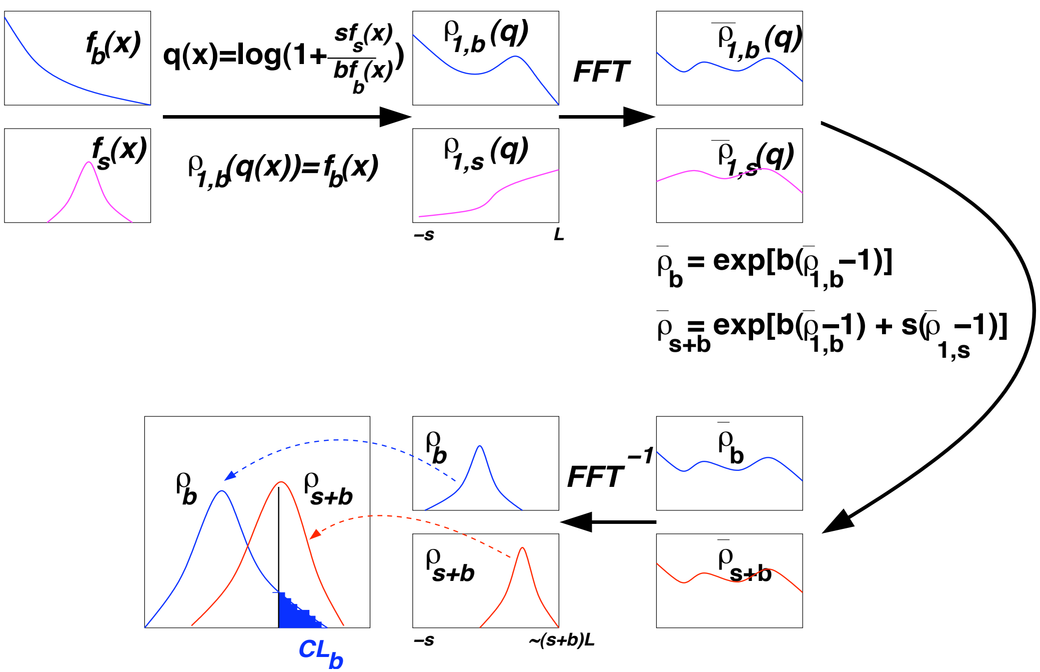
And how do we build the **distribution**? Usually "toy Monte Carlo", but what about the uncertainties... what do we do with the nuisance parameters?

Building the distribution of the test statistic

LEP Higgs Working group developed formalism to combine channels and take advantage of discriminating variables in the likelihood ratio.

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

$$q = \ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$



Hu and Nielsen's CLFFT used Fourier Transform and exponentiation trick to transform the log-likelihood ratio distribution for one event to the distribution for an experiment

Cousins-Highland was used for systematic error on background rate.

Getting this to work at the LHC is tricky numerically because we have channels with n_i from 10-10000 events (physics/0312050)

Building the distribution of the test statistic

LEP Higgs Working group developed formalism to combine channels and take advantage of discriminating variables in the likelihood ratio.

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

$$q = \ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln \left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$

For N events, use Fourier transform to perform N convolutions

$$\rho_{N,i}(q) = \underbrace{\rho_{N,i}(q) \oplus \cdots \oplus \rho_{N,i}(q)}_{N \text{ times}} = \mathcal{F}^{-1}\left\{\left[\mathcal{F}\left(\rho_{1,i}\right)\right]^N\right\}$$

To include Poisson fluctuations on N for a given luminosity, one can exponentiate

$$\rho_i(q) = \sum_{N=0}^{\infty} P(N; L\sigma_i) \cdot \rho_{N,i}(q) = \mathcal{F}^{-1}\left\{e^{L\sigma_i\left[\mathcal{F}(\rho_{1,i}(q))-1\right]}\right\}$$

With nuisance parameters: Hybrid Solutions

Goal of Bayesian-frequentist hybrid solutions is to provide a frequentist treatment of the main measurement, while eliminating nuisance parameters (deal with systematics) with an intuitive Bayesian technique.

$$P(n_{\rm on}|s) = \int db \, {\rm Pois}(n_{\rm on}|s+b)\, \pi(b), \qquad p = \sum_{n=n_{obs}}^{\infty} P(n|s)$$

Tracing back the origin of $\pi(b)$

- clearly state prior $\eta(b)$; identify control samples (sidebands) and use:

$$\pi(b) = P(b|n_{\rm off}) = \frac{P(n_{\rm off}|b)\eta(b)}{\int db P(n_{\rm off}|b)\eta(b)}.$$

Note, if we do not want to use the Hybrid Bayesian-Frequentist approach for the nuisance parameters, then we **must consider both n_{on} and n_{off} when generating our toy Monte Carlo**

$$P(n_{\rm on}, n_{\rm off}|s,b) = {\rm Pois}(n_{\rm on}|s+b)\, {\rm Pois}(n_{\rm off}|\tau b).$$

Coverage as calibration

This prototype problem has been studied extensively.

- instead of arguing about the merits of various methods, just go and check their rate of Type I error (coverage)
- Results indicated large discrepancy in "claimed" coverage and "true" coverage for various methods
- eg. 5σ is really ~4σ for some points

Introduce idea of coverage as a calibration of our statistical apparatus

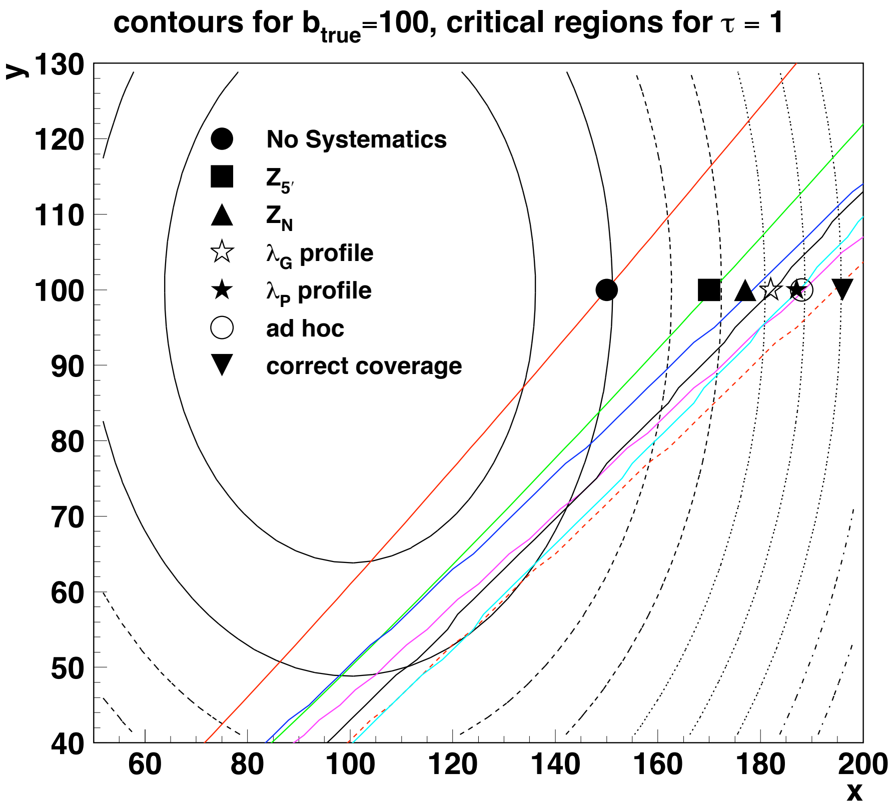


Figure 7. A comparison of the various methods critical boundary $x_{crit}(y)$ (see text). The concentric ovals represent contours of L_G from Eq. 15.

$$L_P(x, y|\mu, b) = Pois(x|\mu + b) \cdot Pois(y|\tau b).$$

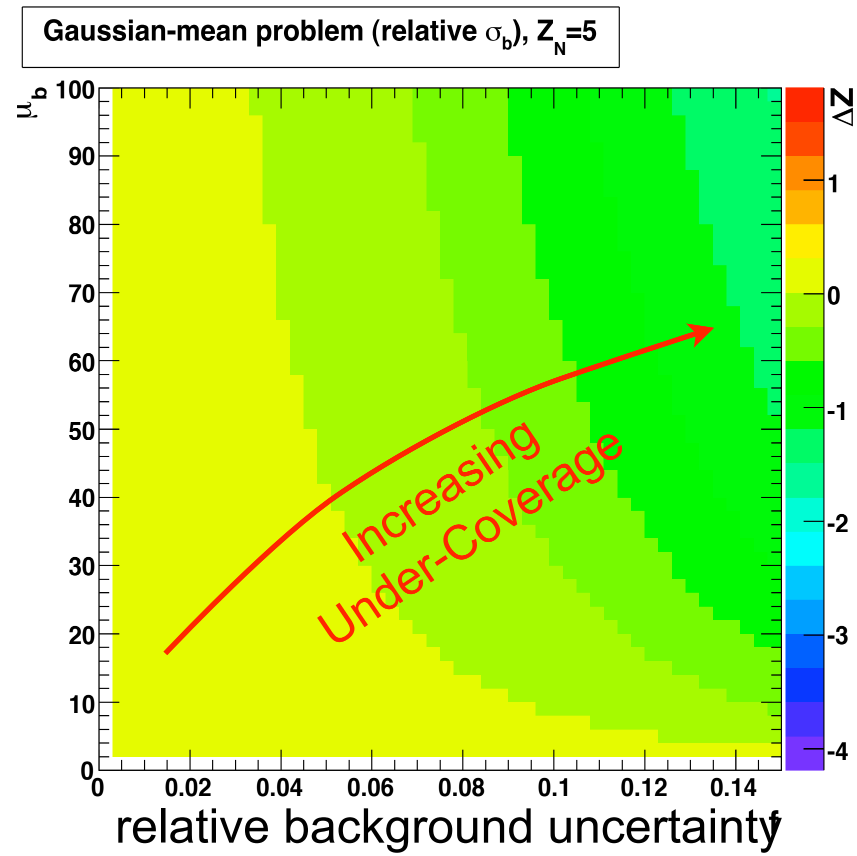
http://www.physics.ox.ac.uk/phystat05/proceedings/files/Cranmer_LHCStatisticalChallenges.ps

Coverage as calibration

This prototype problem has been studied extensively.

- instead of arguing about the merits of various methods, just go and check their rate of Type I error (coverage)
- Results indicated large discrepancy in “claimed” coverage and “true” coverage for various methods
- eg. 5σ is really ~4σ for some points

Introduce idea of coverage as a calibration of our statistical apparatus



Recent work by Bob Cousins & Jordan Tucker, [physics/0702156]

$$L_P(x, y|\mu, b) = Pois(x|\mu + b) \cdot Pois(y|\tau b).$$

http://www.physics.ox.ac.uk/phystat05/proceedings/files/Cranmer_LHCStatisticalChallenges.ps

The Profile Likelihood Ratio

Define μ to be signal rate in units of SM expectation

Define ν to be the shape parameters (nuisance parameters)

In the LEP approach the likelihood ratio is equivalent to:

$$Q_{LEP} = \frac{L(data|\mu = 1, b, \nu)}{L(data|\mu = 0, b, \nu)}$$

- but this variable is sensitive to uncertainty on ν

Alternatively, one can define **profile likelihood ratio**

$$\lambda(\mu = 0) = \frac{L(data|\mu = 0, \hat{\hat{b}}(\mu = 0), \hat{\hat{\nu}}(\mu = 0))}{L(data|\hat{\mu}, \hat{b}, \hat{\nu})},$$

- where $\hat{\hat{\nu}}$ is best fit with μ fixed to 0
- and $\hat{\nu}$ is best fit with μ left floating
- conventional ratio is reciprocal in hypo test <-> limit

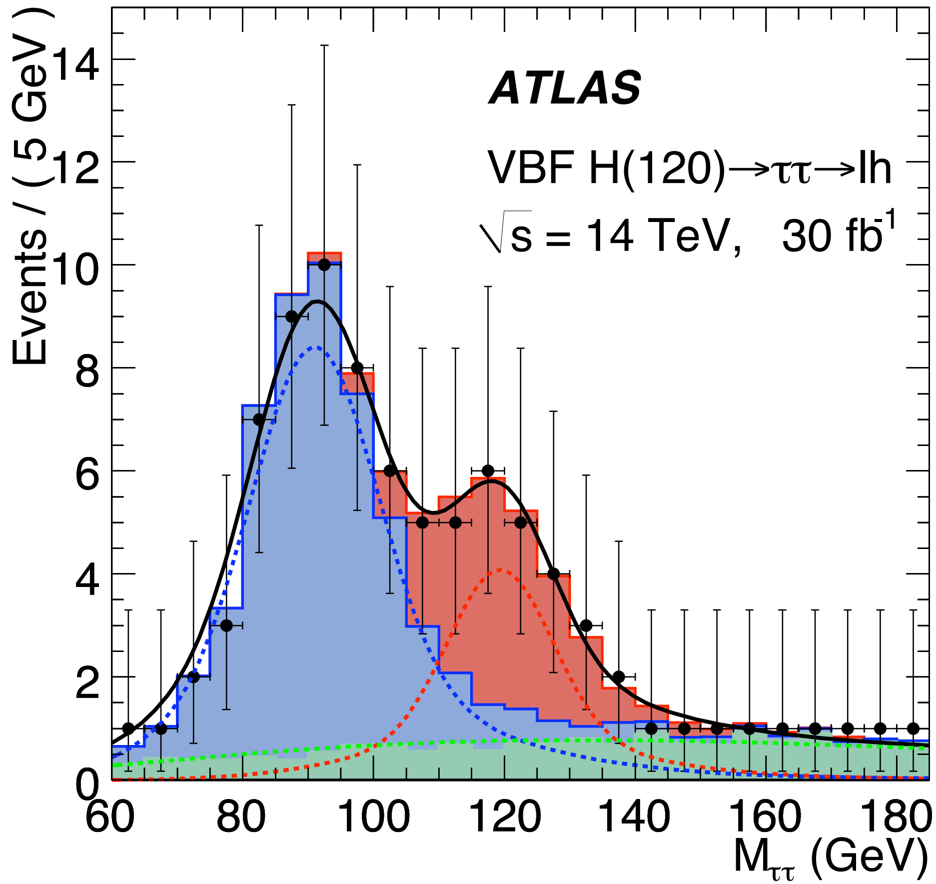
An example

Essentially, you need to fit your model to the data twice:

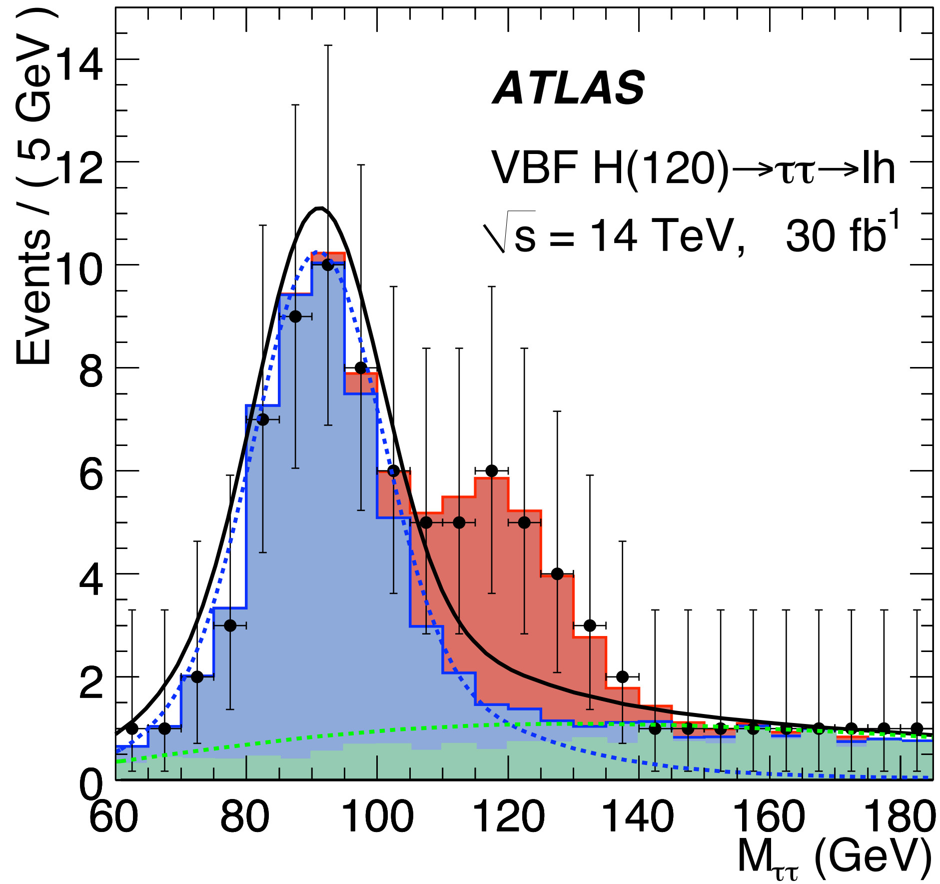
once with everything floating, and once with signal fixed to 0

$$\lambda(\mu = 0) = \frac{L(data|\mu = 0, \hat{\hat{b}}(\mu = 0), \hat{\hat{\nu}}(\mu = 0))}{L(data|\hat{\mu}, \hat{b}, \hat{\nu})},$$

$L(data|\hat{\mu}, \hat{b}, \hat{\nu})$



$L(data|\mu = 0, \hat{\hat{b}}, \hat{\hat{\nu}})$



Properties of the Profile Likelihood Ratio

After a close look at the profile likelihood ratio

$$\lambda(\mu=0) = \frac{L(data|\mu=0, \hat{\hat{b}}(\mu=0), \hat{\hat{\nu}}(\mu=0))}{L(data|\hat{\mu}, \hat{b}, \hat{\nu})},$$

one can see the function is independent of true values of ν

- though its distribution might depend indirectly

Wilks's theorem states that under certain conditions the distribution of the profile likelihood ratio has an asymptotic form

$$-2\log\lambda(\mu=0) \sim \chi^2_1$$

Thus, we can calculate the p-value for the background-only hypothesis by calculating

$$-2\log\lambda(\mu=0)$$

or equivalently:

$$Z = \sqrt{-2\log\lambda(\mu=0)}$$

Hypothesis Testing

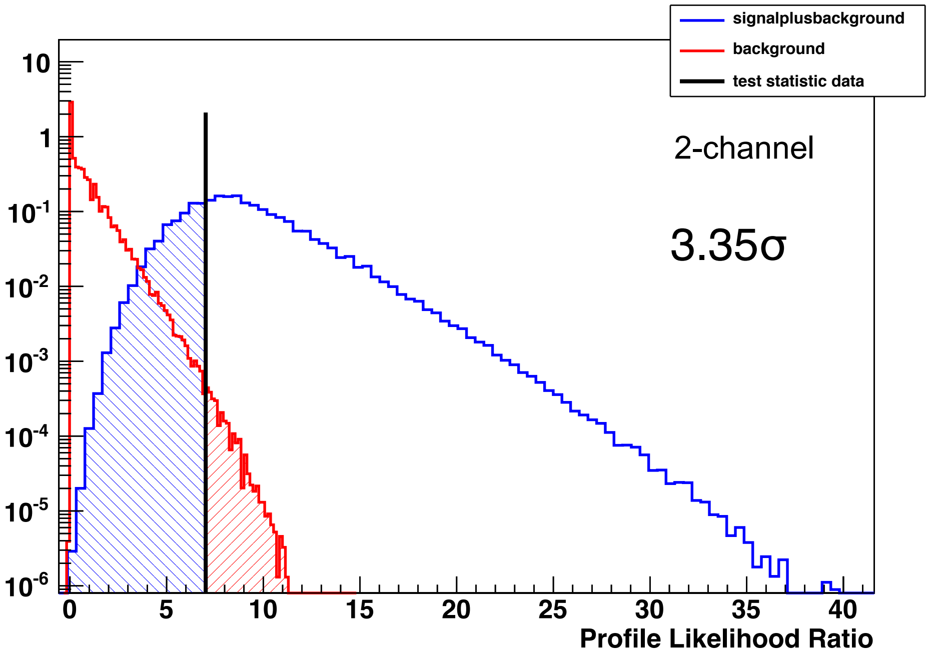
CENTER FOR COSMOLOGY AND PARTICLE PHYSICS

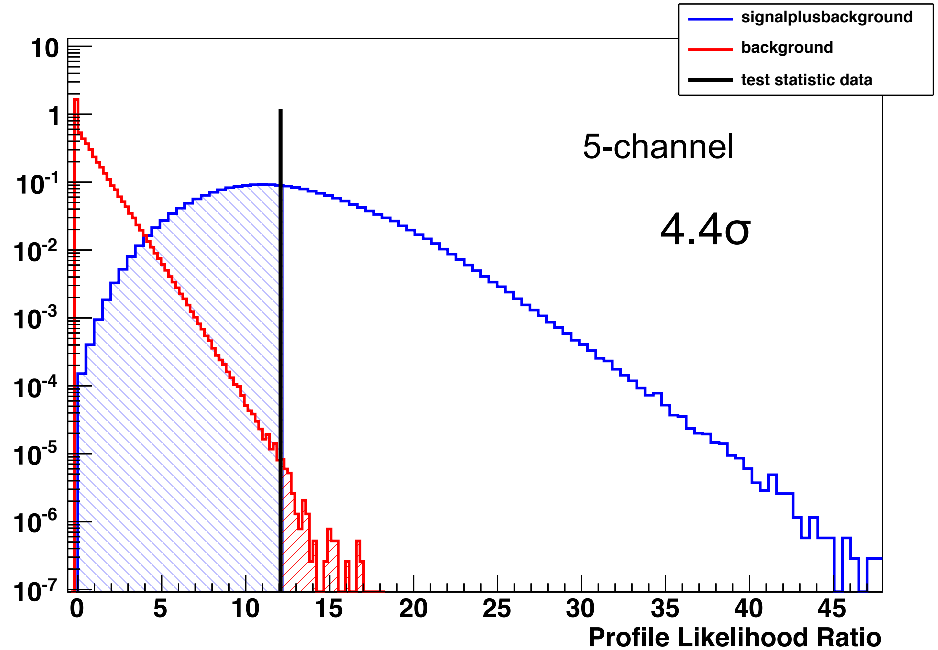
Now on a real PROOF cluster with 30 machines

- real world example throws millions of toys experiments, does full fit on 50 parameters for each toy.
- also supports producing simple shells scripts for use with GRID or batch queues

Now **importance sampling** is also implemented,

- following presentation at Banff with particle physics & statistics experts
- allows for 1000x speed increase!
- Still being tested in detail



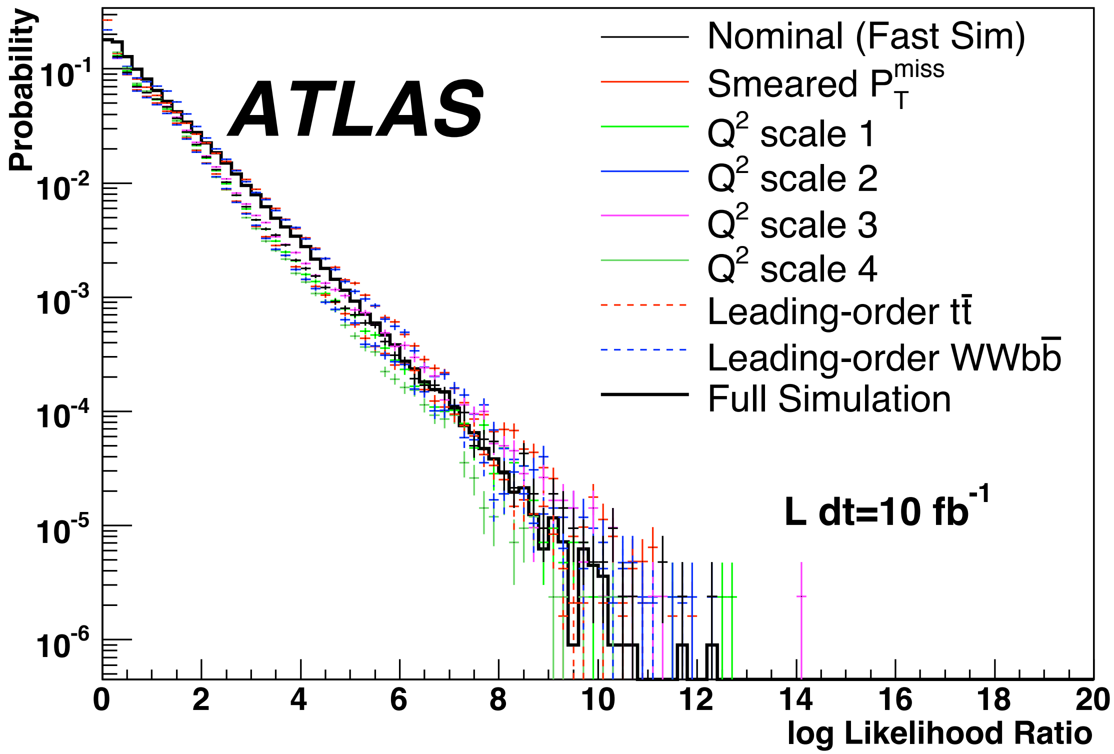


Experimentalist Justification

So far this looks a bit like magic. How can you claim that you incorporated your systematic just by fitting the best value of your uncertain parameters and making a ratio?

It won't unless the the parametrization is sufficiently flexible.

So check by varying the settings of your simulation, and see if the profile likelihood ratio is still distributed as a chi-square



Here it is pretty stable, but it's not perfect (and this is a log plot, so it hides some pretty big discrepancies)

For the distribution to be independent of the nuisance parameters your parametrization must be sufficiently flexible.

Ingredients to Frequentist methods

RooStats supports several statistical methods used in high energy physics

- Choose a test statistic
  - simple likelihood ratio (LEP) $Q_{LEP} = L_{s+b}(\mu = 1)/L_b(\mu = 0)$
  - ratio of profiled likelihoods (Tevatron) $Q_{TEV} = L_{s+b}(\mu = 1, \hat{\hat{\nu}})/L_b(\mu = 0, \hat{\hat{\nu}}')$
  - profile likelihood ratio (LHC) $\lambda(\mu) = L_{s+b}(\mu, \hat{\hat{\nu}})/L_{s+b}(\hat{\mu}, \hat{\nu})$
- Define your ensemble (sampling strategy)
  - toy MC randomizing nuisance parameters according to $\pi(\nu)$
    - aka Bayes-frequentist hybrid, prior-predictive, Cousins-Highland
  - toy MC with nuisance parameters fixed (Neyman Construction)
  - assuming asymptotic distribution (Wilks and Wald)

Lecture 3

Confidence Intervals (Limits)

Simple vs. Compound Hypotheses

The Neyman-Pearson lemma is **the answer** for simple hypothesis testing

- a hypothesis is **simple** if it has no free parameters and is totally fixed $f(x|H_0)$ vs. $f(x|H_1)$

What about cases when there are free parameters?

- eg. the mass of the Higgs boson $f(x|H_0)$ vs. $f(x|H_1, m_H)$

A test is called **similar** if it has size α for all values of the parameters

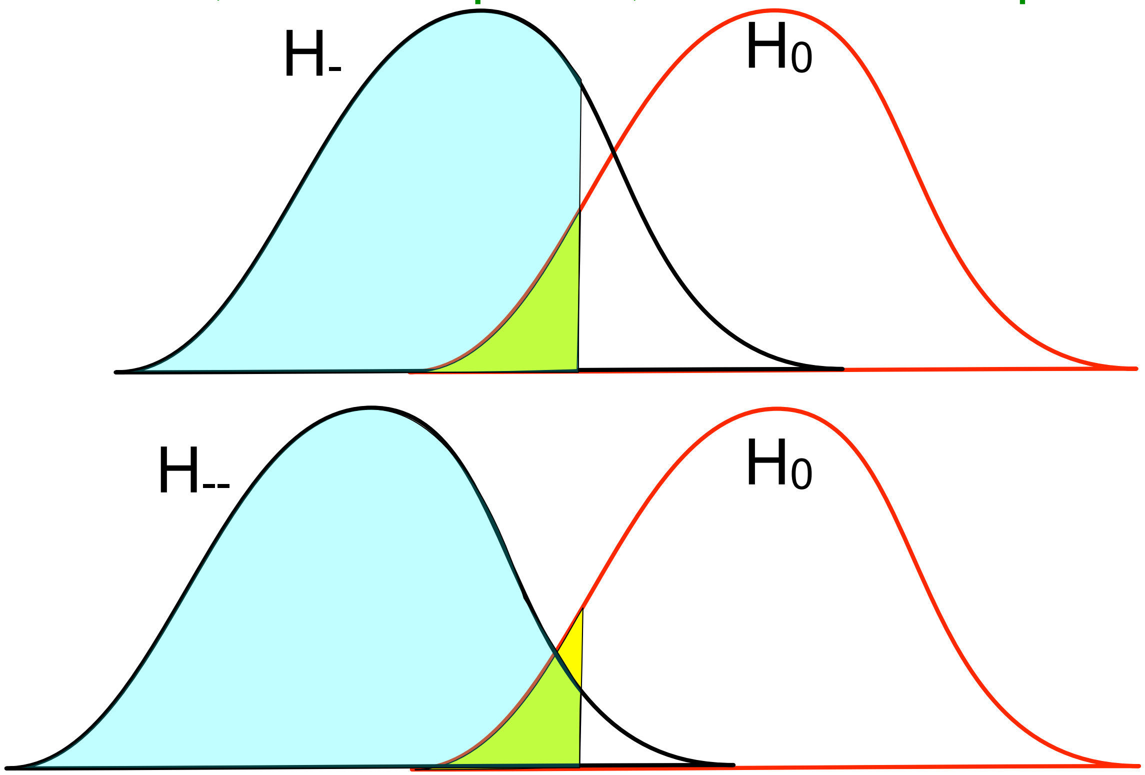
A test is called **Uniformly Most Powerful** if it maximizes the power for all values of the parameter

Uniformly Most Powerful tests don't exist in general

Similar Test Examples

In some cases Uniformly Most Powerful tests do exist:

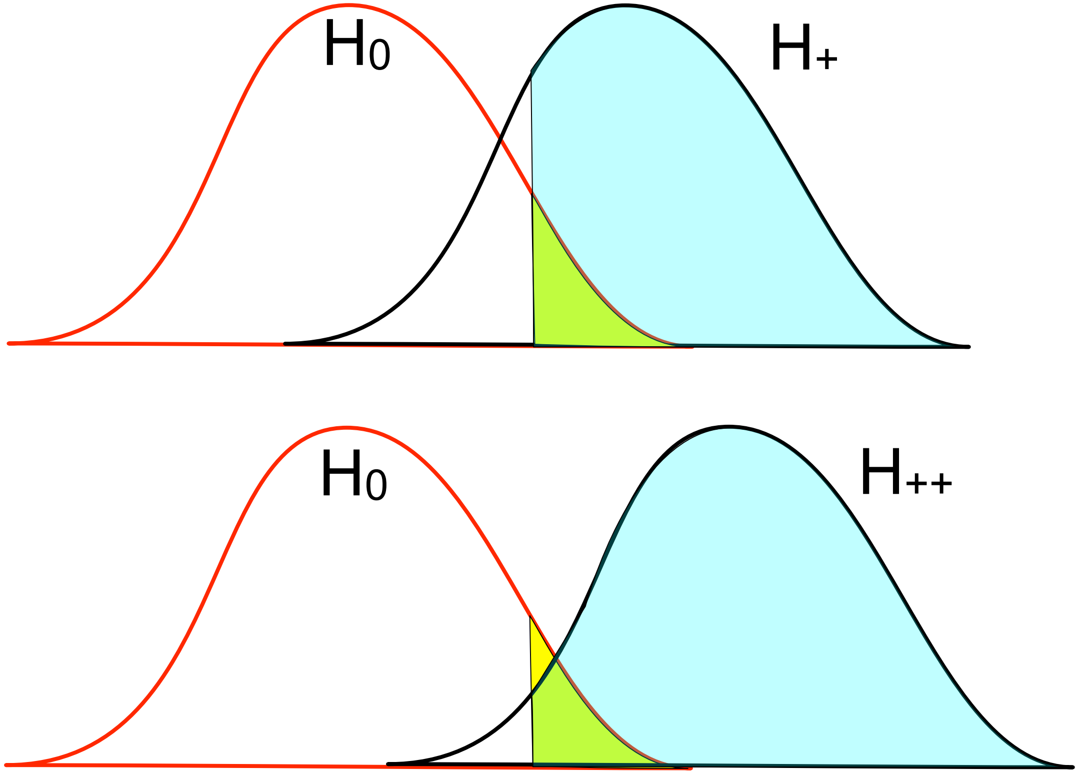
- some examples just to clarify the concept:
- H_0 is simple: a Gaussian with a fixed $\mu = \mu_0, \sigma = \sigma_0$
- H_1 is composite: a Gaussian with $\mu < \mu_0, \sigma = \sigma_0$
  - consider H_- and H_{--}
  - same size, different power, but both max power



Similar Test Examples

In some cases Uniformly Most Powerful tests exists:

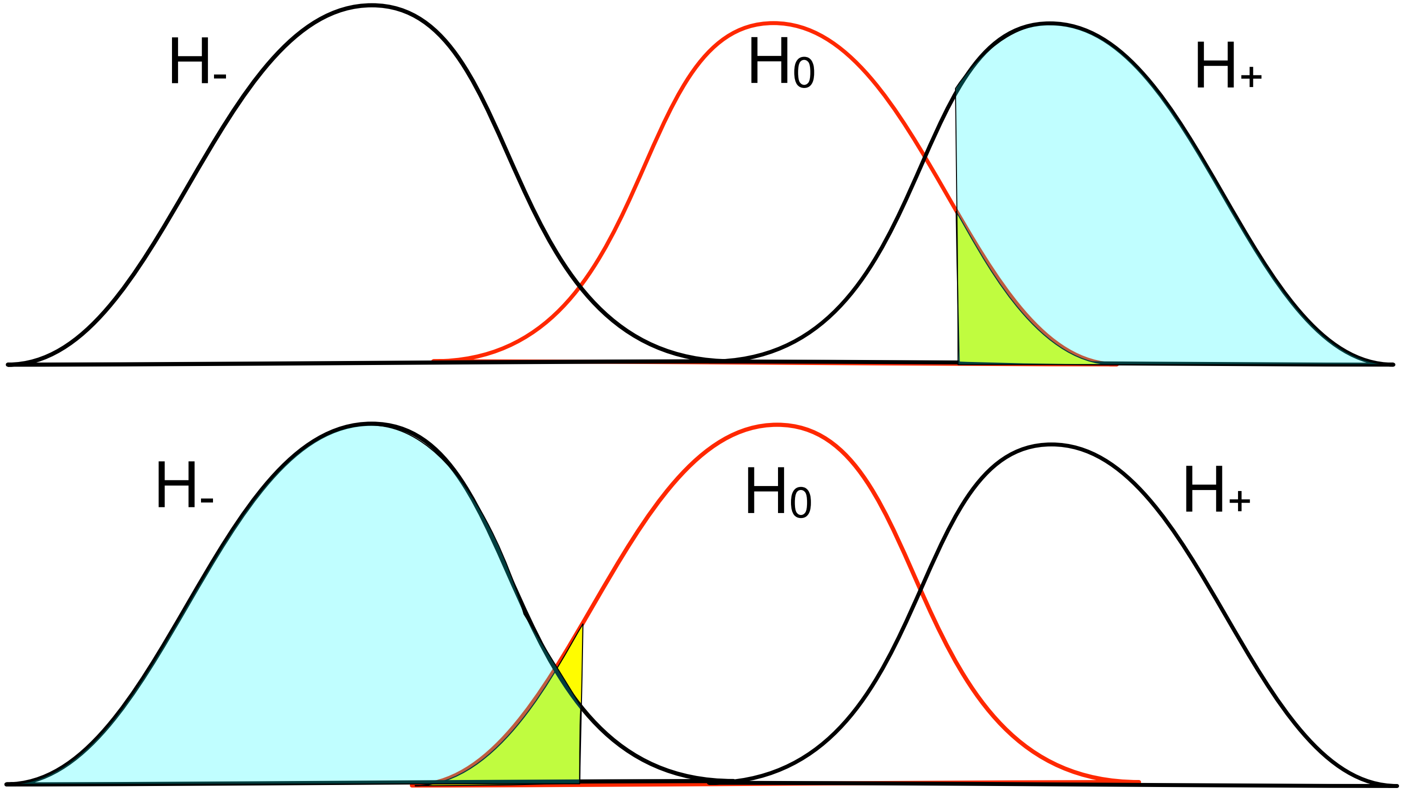
- some examples just to clarify the concept:
- H_0 is simple: a Gaussian with a fixed $\mu = \mu_0, \sigma = \sigma_0$
- H_1 is composite: a Gaussian with $\mu > \mu_0, \sigma = \sigma_0$
  - consider H_+ and H_{++}
  - same size, different power, but both max power



Similar Test Examples

Slight variation, a Uniformly Most Powerful test doesn't exit:

- some examples just to clarify the concept:
- H_0 is simple: a Gaussian with a fixed $\mu = \mu_0, \sigma = \sigma_0$
- H_1 is composite: a Gaussian with $\mu = \mu_0, \sigma \neq \sigma_0$
  - Either H_+ has good power and H_- has bad power
  - or vice versa



Composite Hypothesis & the Likelihood Function

When a hypothesis is composite typically there is a pdf that can be parametrized $f(\vec{x}|\theta)$

- for a fixed θ it defines a pdf for the random variable x
- for a given measurement of x one can consider $f(\vec{x}|\theta)$ as a function of θ called the **Likelihood function**
- Note, this is not Bayesian, because it still only uses P(data | theory) and
  - **the Likelihood function is not a pdf!**

Sometimes θ has many components, generally divided into:

- **parameters of interest**: eg. masses, cross-sections, etc.
- **nuisance parameters**: eg. parameters that affect the shape but are not of direct interest (eg. energy scale)

A simple example:

A Poisson distribution describes a discrete event count n for a real-valued mean μ.

$$Pois(n|\mu) = \mu^n \frac{e^{-\mu}}{n!}$$

The likelihood of μ given n is the same equation evaluated as a function of μ

- Now it's a continuous function
- But it is not a pdf!

$$L(\mu) = Pois(n|\mu)$$

Common to plot the -2 ln *L*

- helps avoid thinking of it as a PDF
- connection to χ^2 distribution

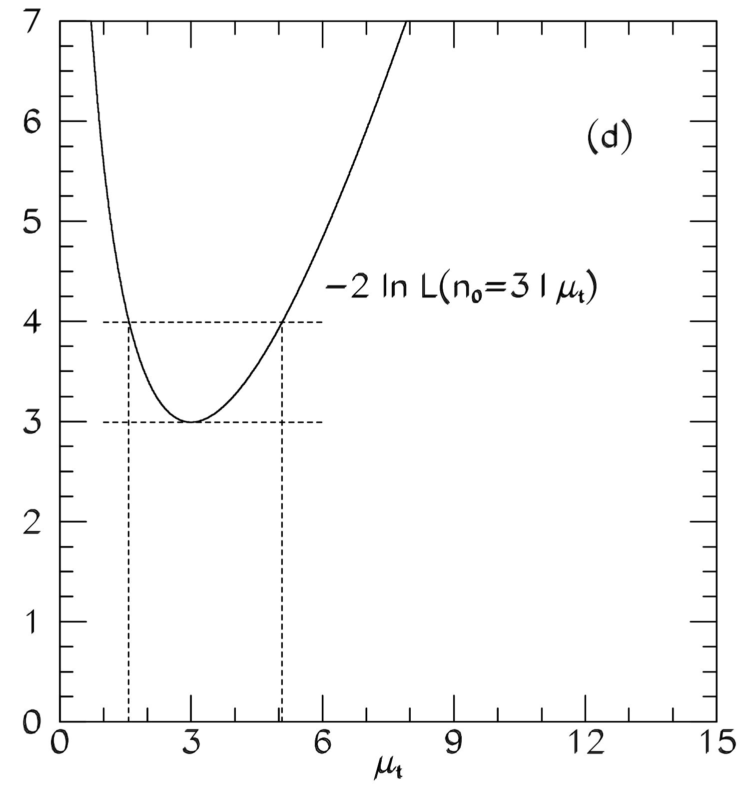
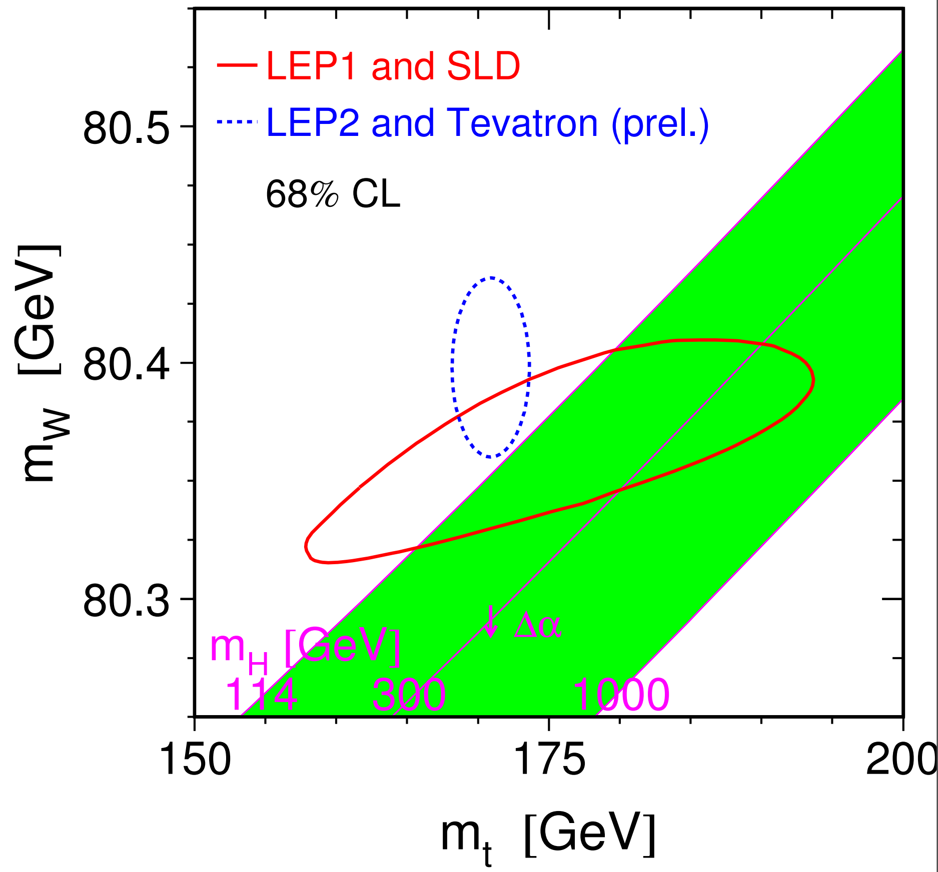


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

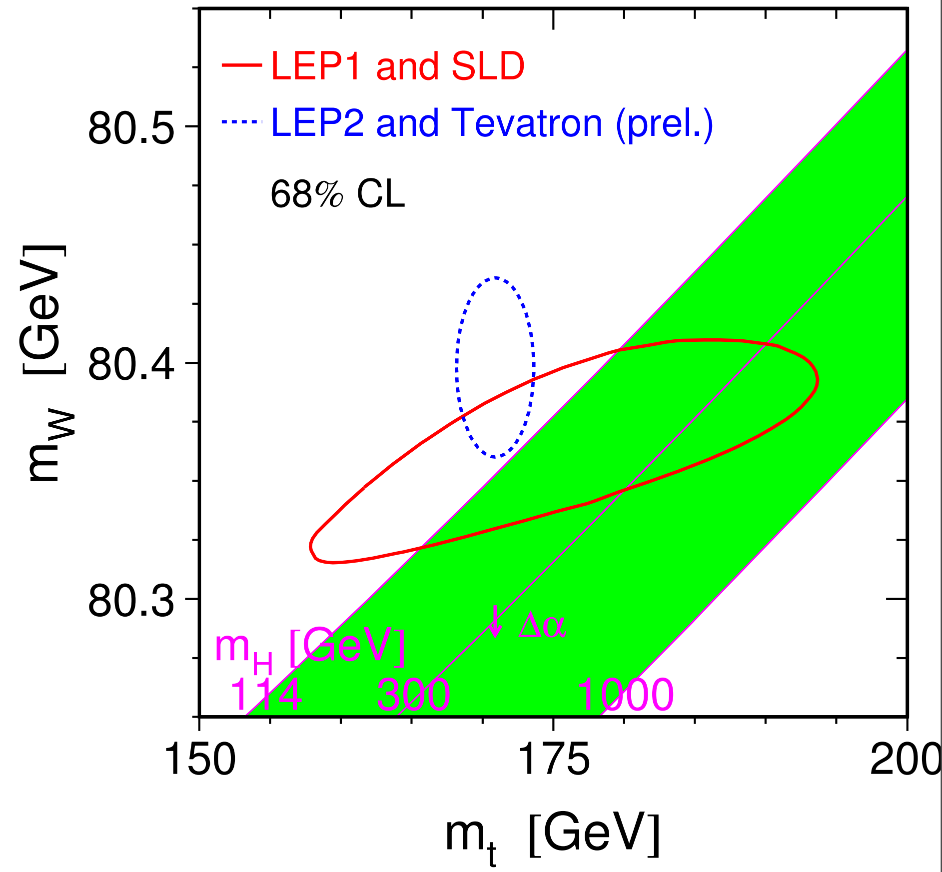
Confidence Interval

Center for
Cosmology and
Particle Physics



Confidence Interval

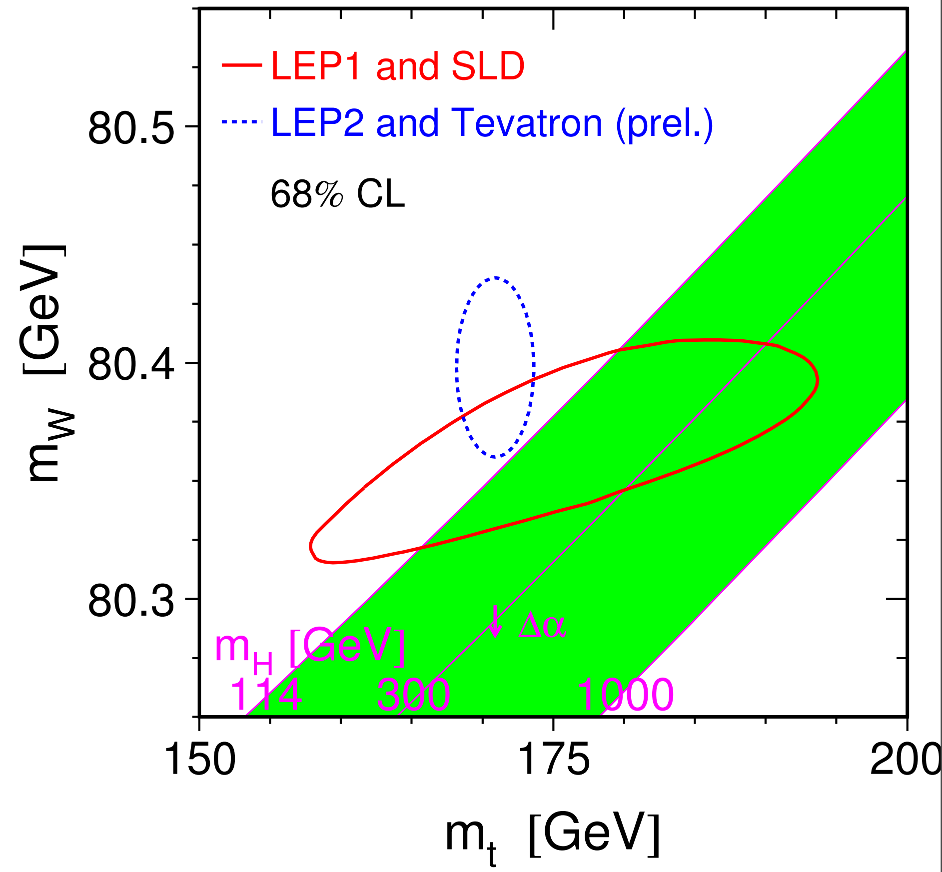
What is a "Confidence Interval?



Confidence Interval

What is a “Confidence Interval?

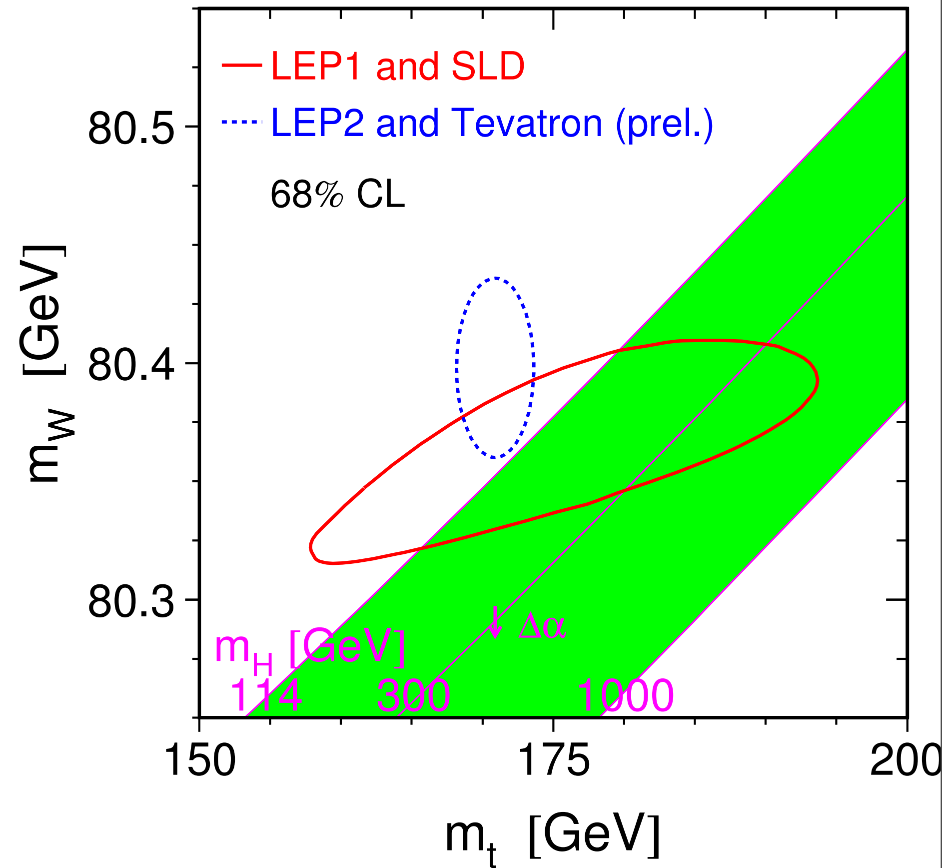
- you see them all the time:



What is a "Confidence Interval?

- you see them all the time:

Want to say there is a 68% chance that the true value of (m_W, m_t) is in this interval



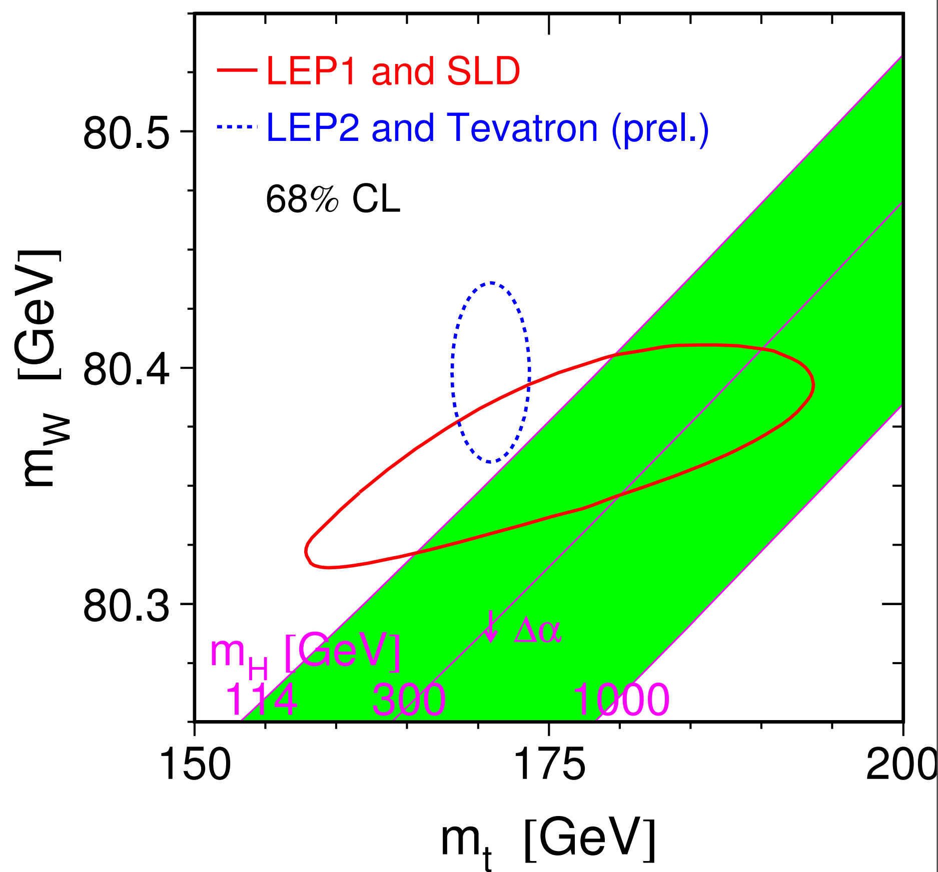
Confidence Interval

What is a "Confidence Interval?

- you see them all the time:

Want to say there is a 68% chance that the true value of (m_W, m_t) is in this interval

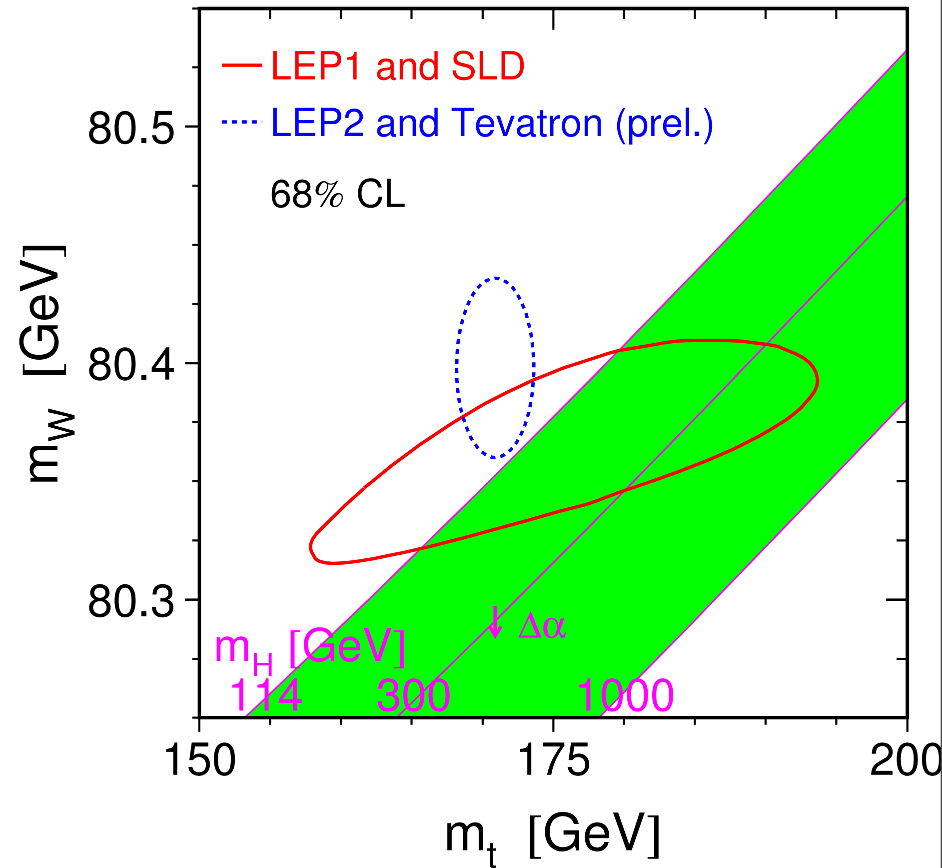
- but that's P(theory|data)!



Confidence Interval

What is a "Confidence Interval?

- you see them all the time:



Want to say there is a 68% chance that the true value of (m_W, m_t) is in this interval

- but that's P(theory|data)!

Correct frequentist statement is that the interval **covers** the true value 68% of the time

Confidence Interval

What is a "Confidence Interval?

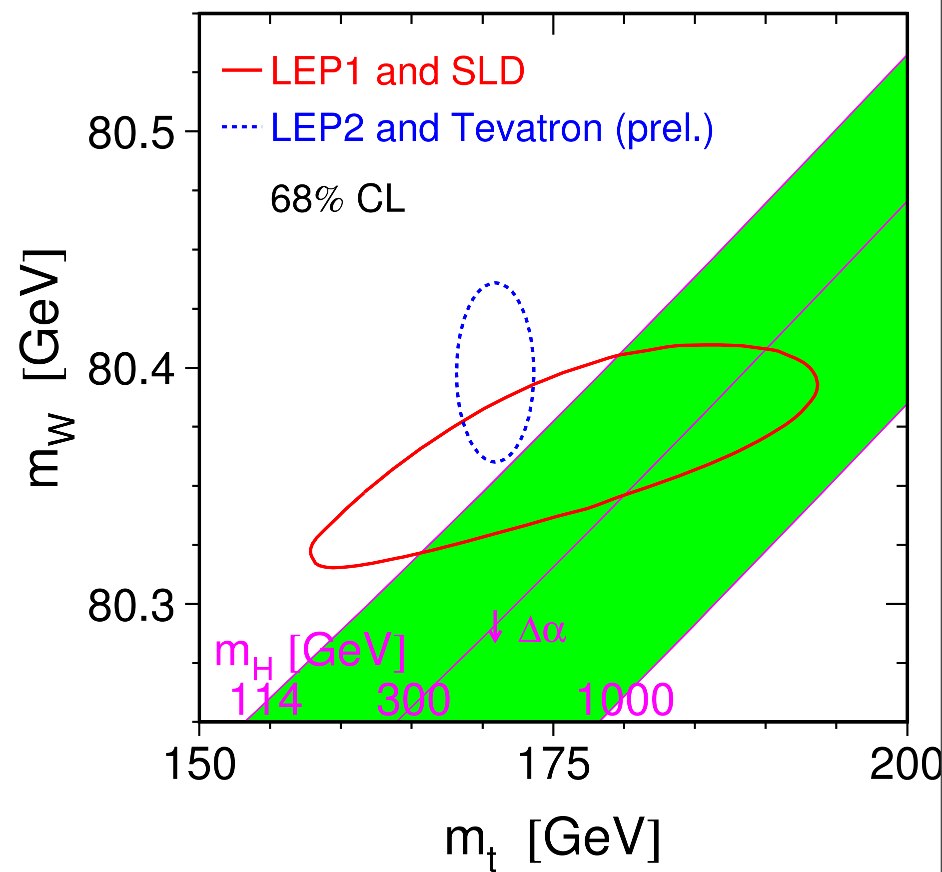
- you see them all the time:

Want to say there is a 68% chance that the true value of (m_W, m_t) is in this interval

- but that's P(theory|data)!

Correct frequentist statement is that the interval **covers** the true value 68% of the time

- remember, the contour is a function of the data, which is random. So it moves around from experiment to experiment



Confidence Interval

Center for
Cosmology and
Particle Physics

What is a "Confidence Interval?

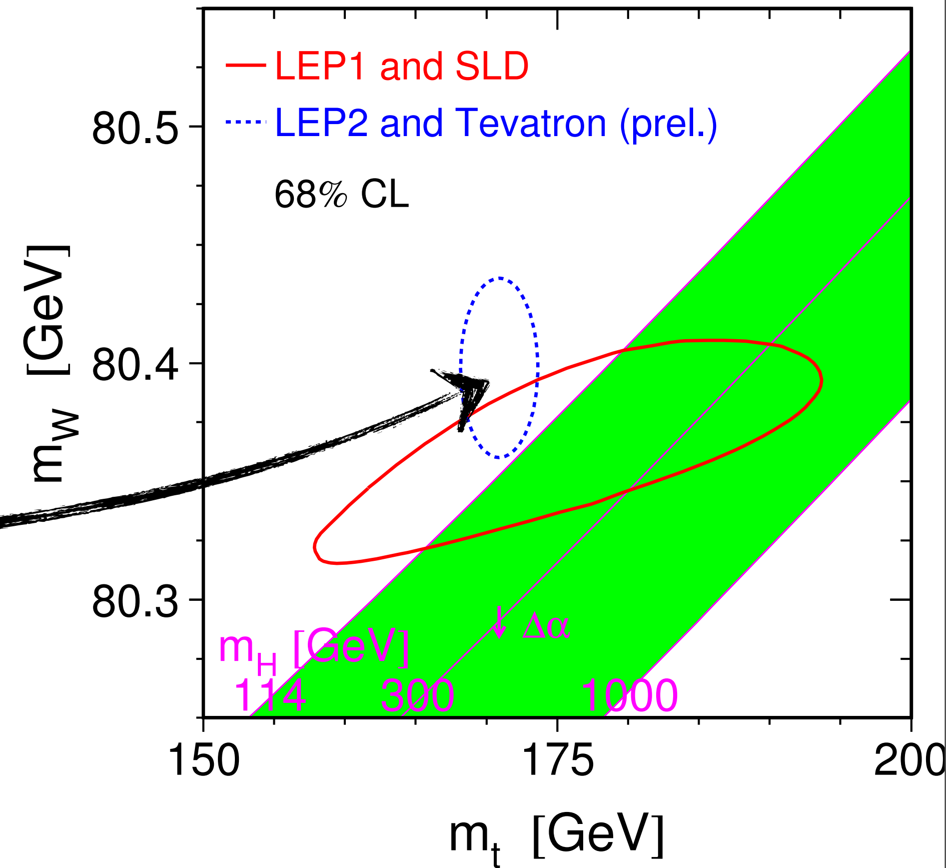
- you see them all the time:

Want to say there is a 68% chance that the true value of (m_W, m_t) is in this interval

- but that's P(theory|data)!

Correct frequentist statement is that the interval **covers** the true value 68% of the time

- remember, the contour is a function of the data, which is random. So it moves around from experiment to experiment



Confidence Interval

What is a "Confidence Interval?

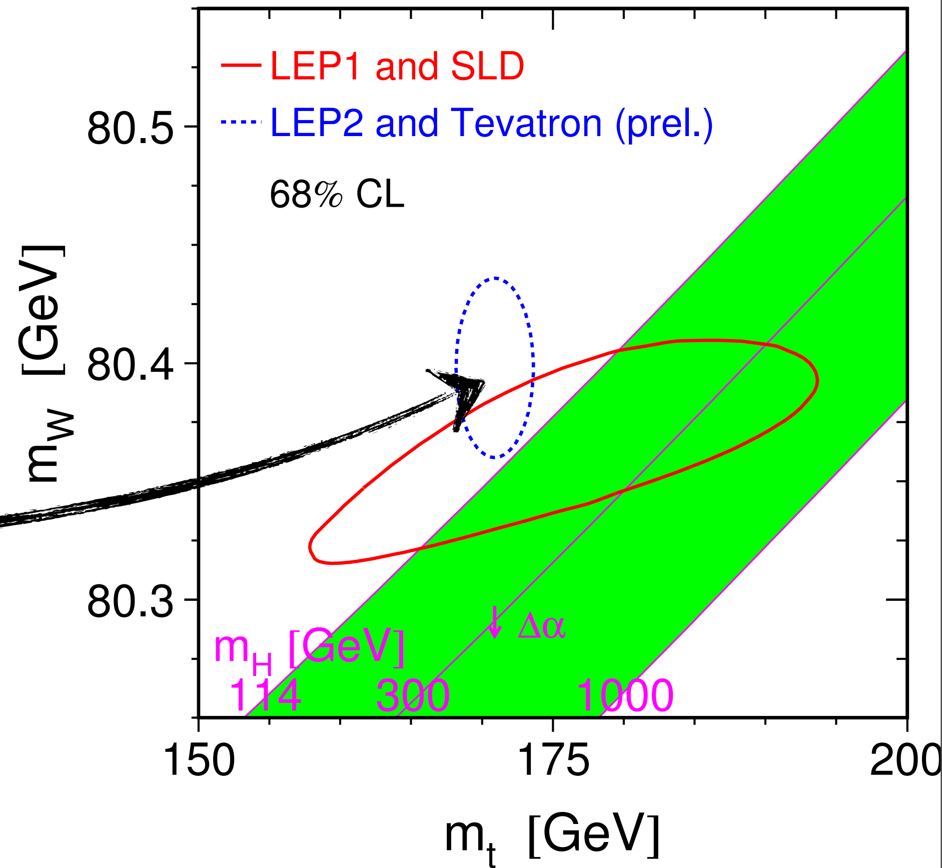
- you see them all the time:

Want to say there is a 68% chance that the true value of (m_W, m_t) is in this interval

- but that's P(theory|data)!

Correct frequentist statement is that the interval **covers** the true value 68% of the time

- remember, the contour is a function of the data, which is random. So it moves around from experiment to experiment

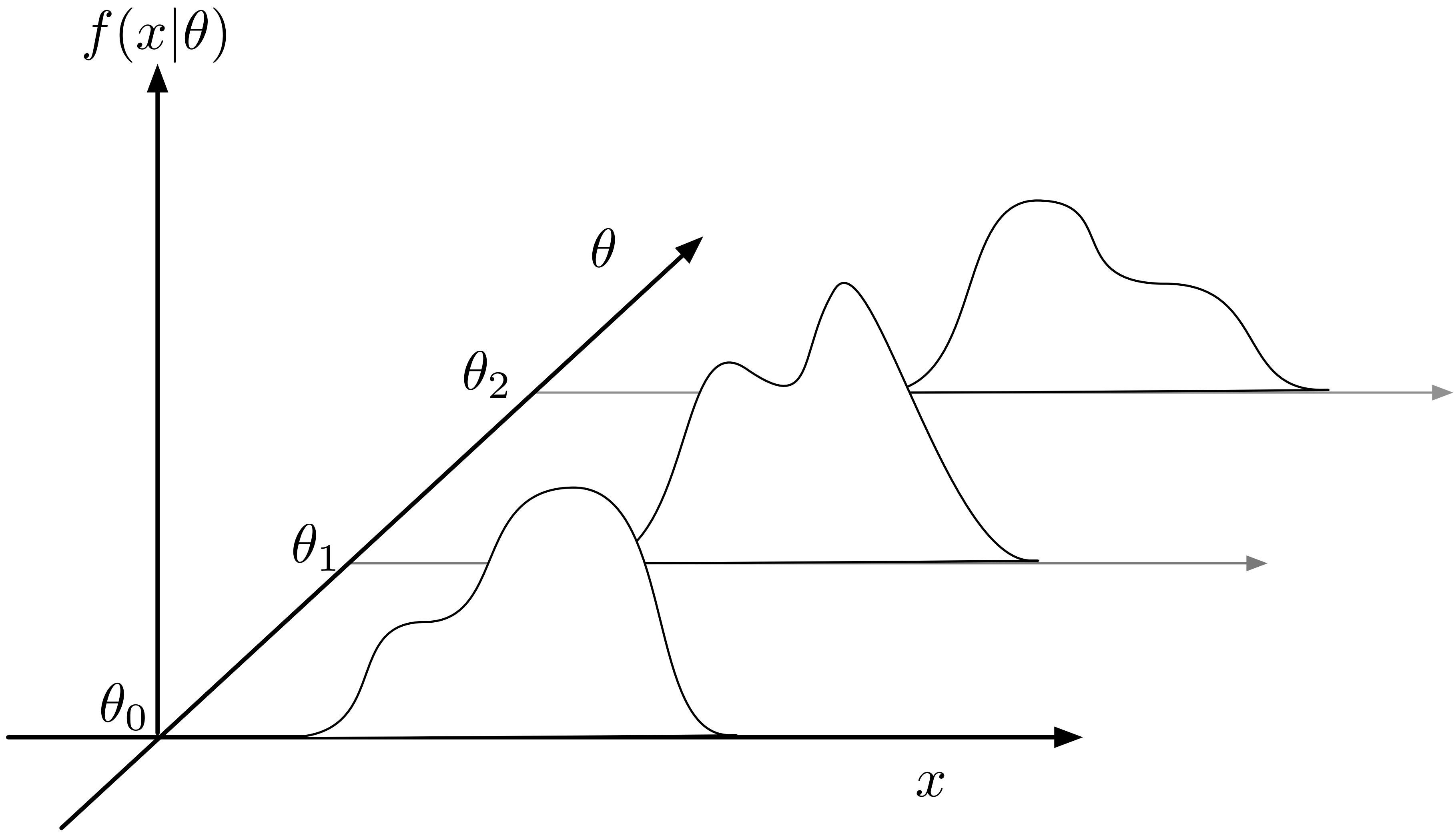


- Bayesian "credible interval" does mean probability parameter is in interval. The procedure is very intuitive:

$$P(\theta \in V) = \int_V \pi(\theta|x) = \int_V d\theta \frac{f(x|\theta)\pi(\theta)}{\int d\theta f(x|\theta)\pi(\theta)}$$

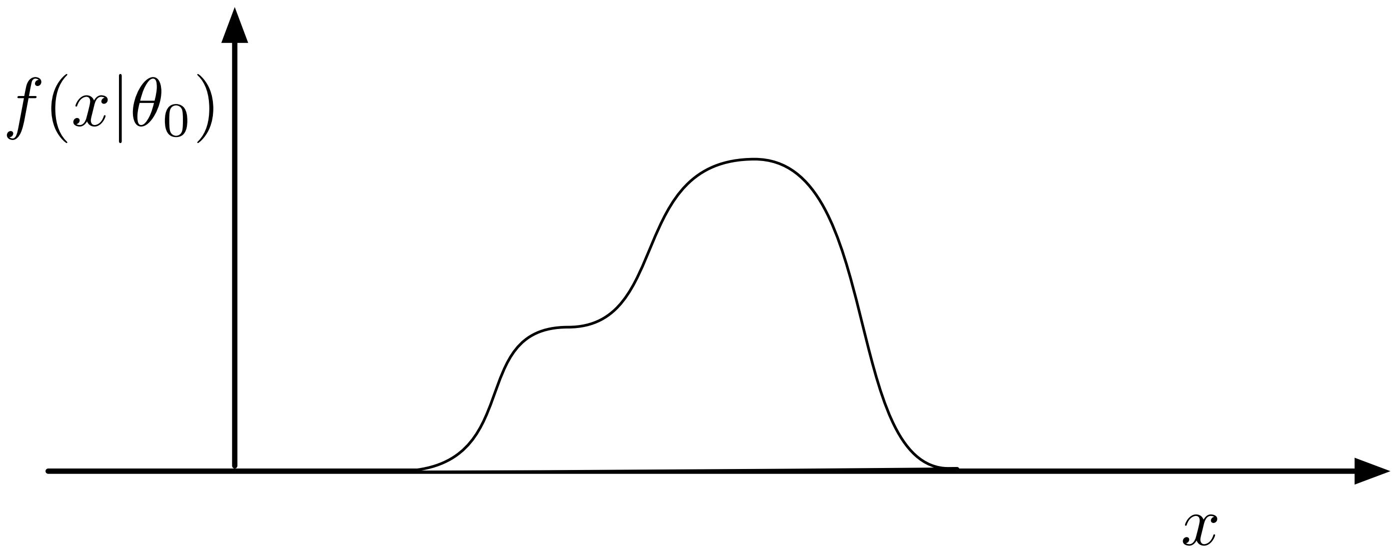
Neyman Construction example

For each value of θ consider $f(x|\theta)$



Neyman Construction example

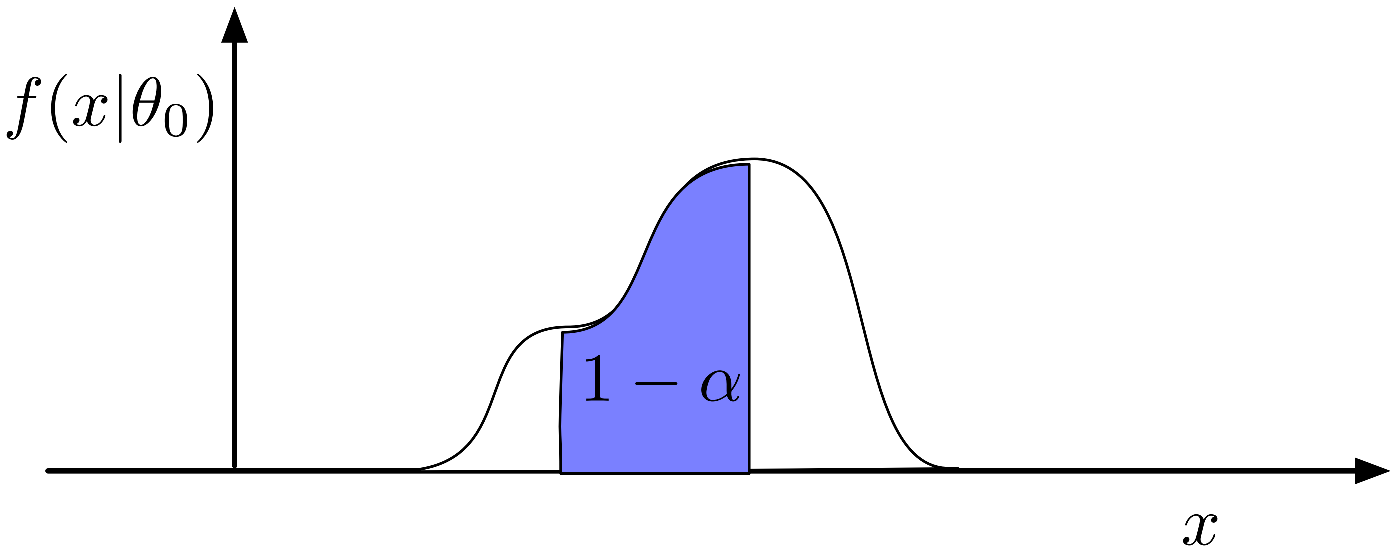
Let's focus on a particular point $f(x|\theta_o)$



Neyman Construction example

Let's focus on a particular point $f(x|\theta_o)$

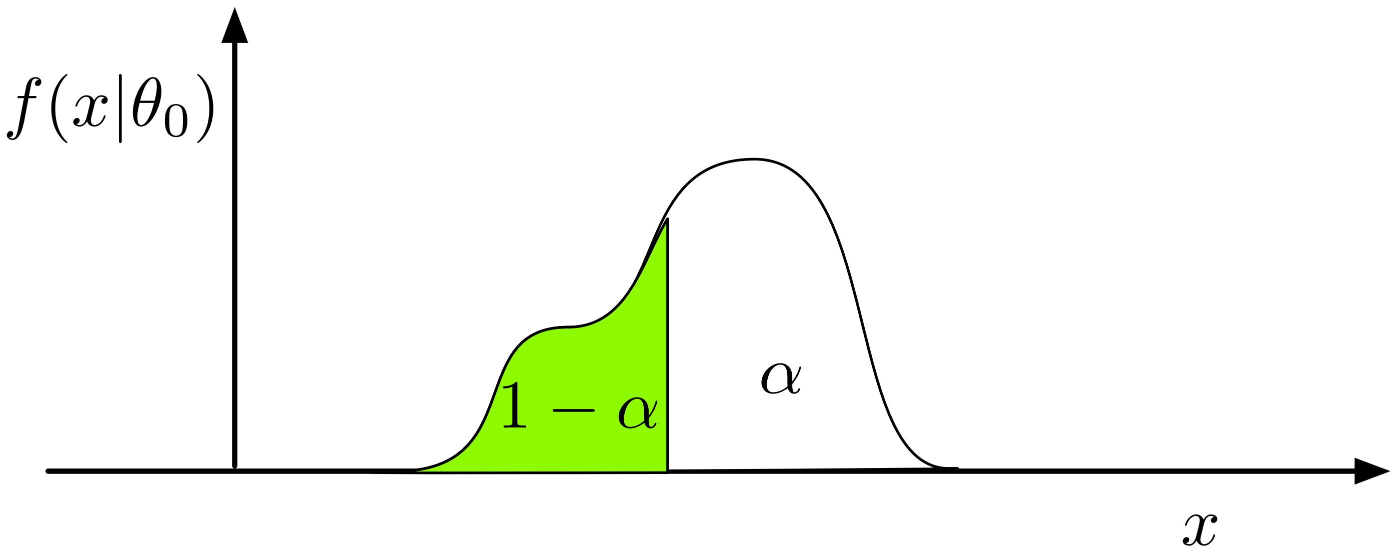
- we want a test of size α
- equivalent to a $100(1-\alpha)\%$ confidence interval on θ
- so we find an **acceptance region** with $1-\alpha$ probability



Neyman Construction example

Let's focus on a particular point $f(x|\theta_o)$

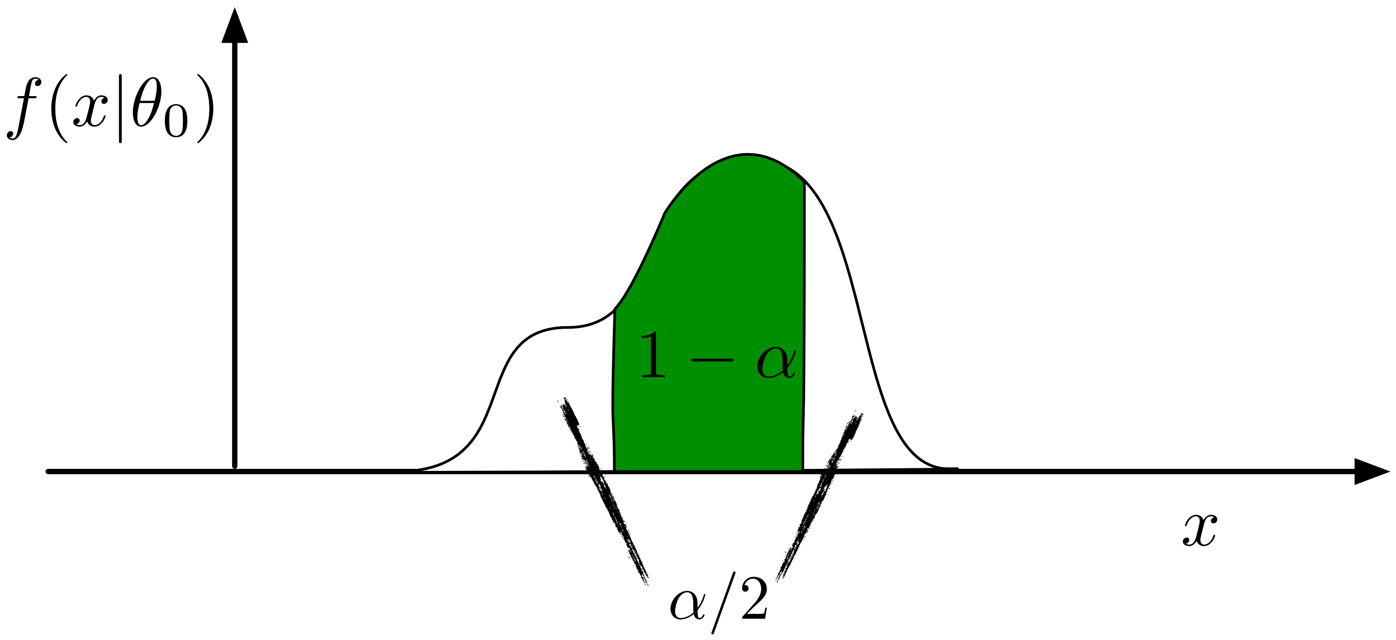
- No unique choice of an acceptance region
- here's an example of a lower limit



Center for Cosmology and Particle Physics

Let's focus on a particular point $f(x|\theta_o)$

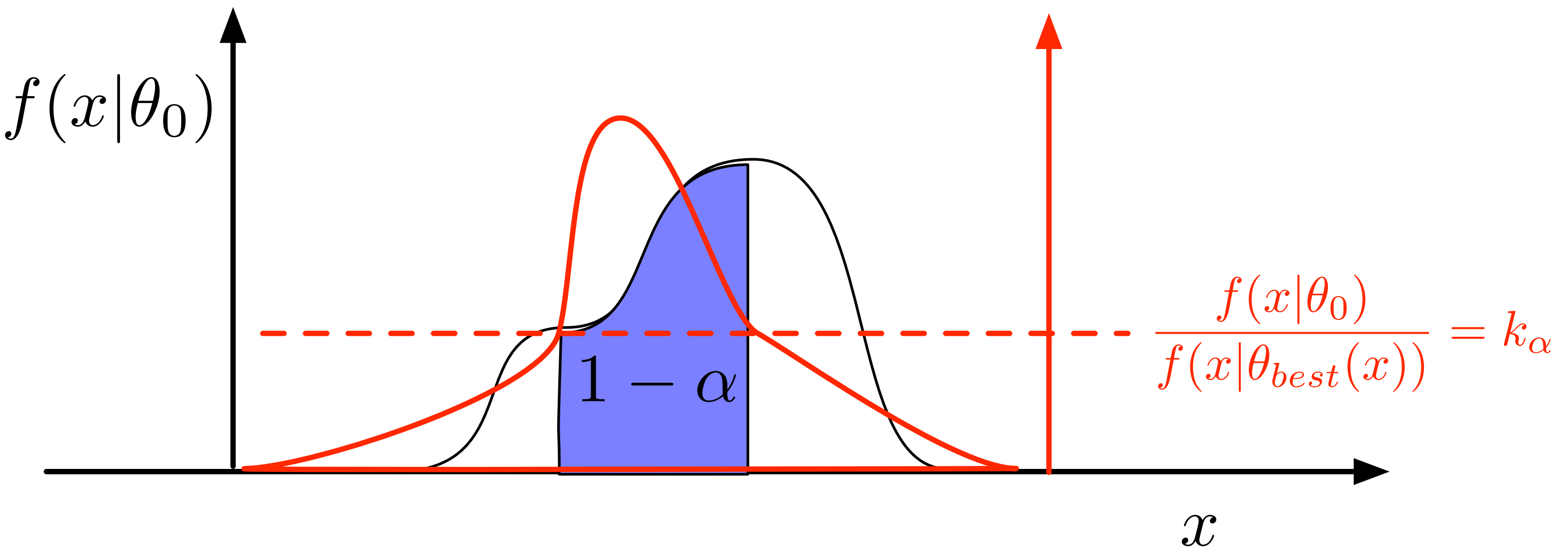
- No unique choice of an acceptance region
- and an example of a central limit



Neyman Construction example

Let's focus on a particular point $f(x|\theta_o)$

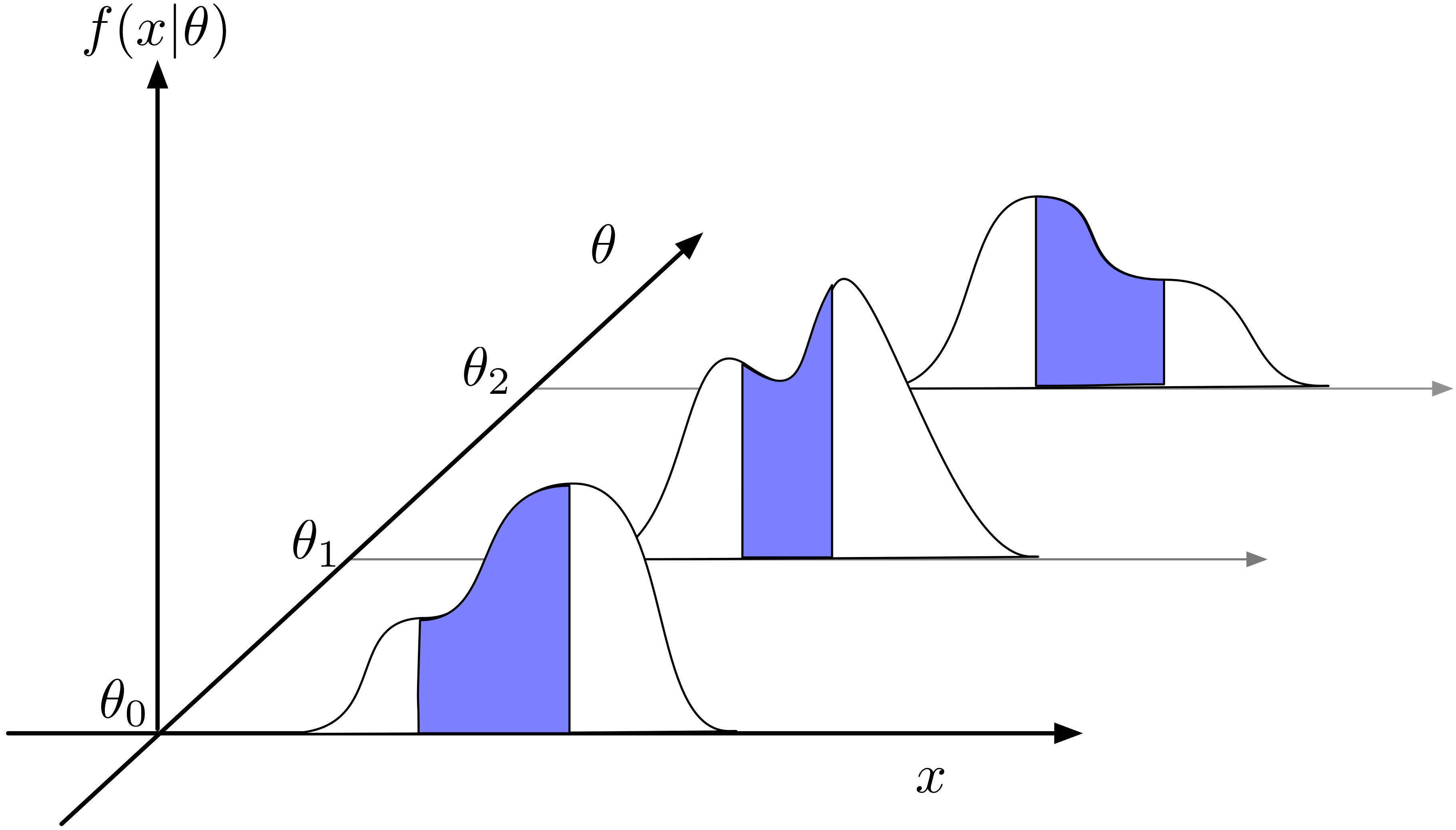
- choice of this region is called an **ordering rule**
- In Feldman-Cousins approach, ordering rule is the likelihood ratio. Find contour of L.R. that gives size α



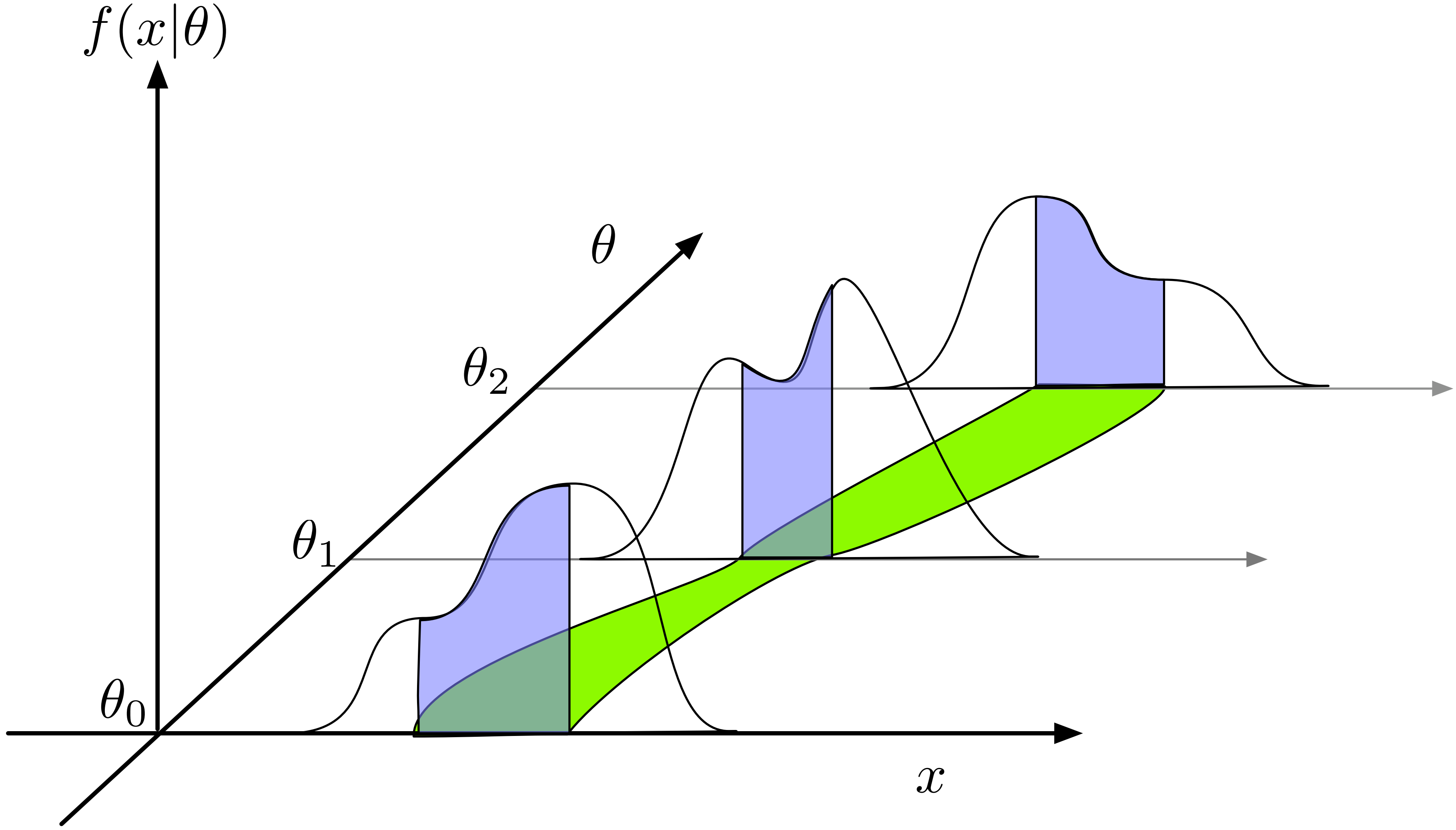
Neyman Construction example

Center for Cosmology and Particle Physics

Now make acceptance region for every value of θ



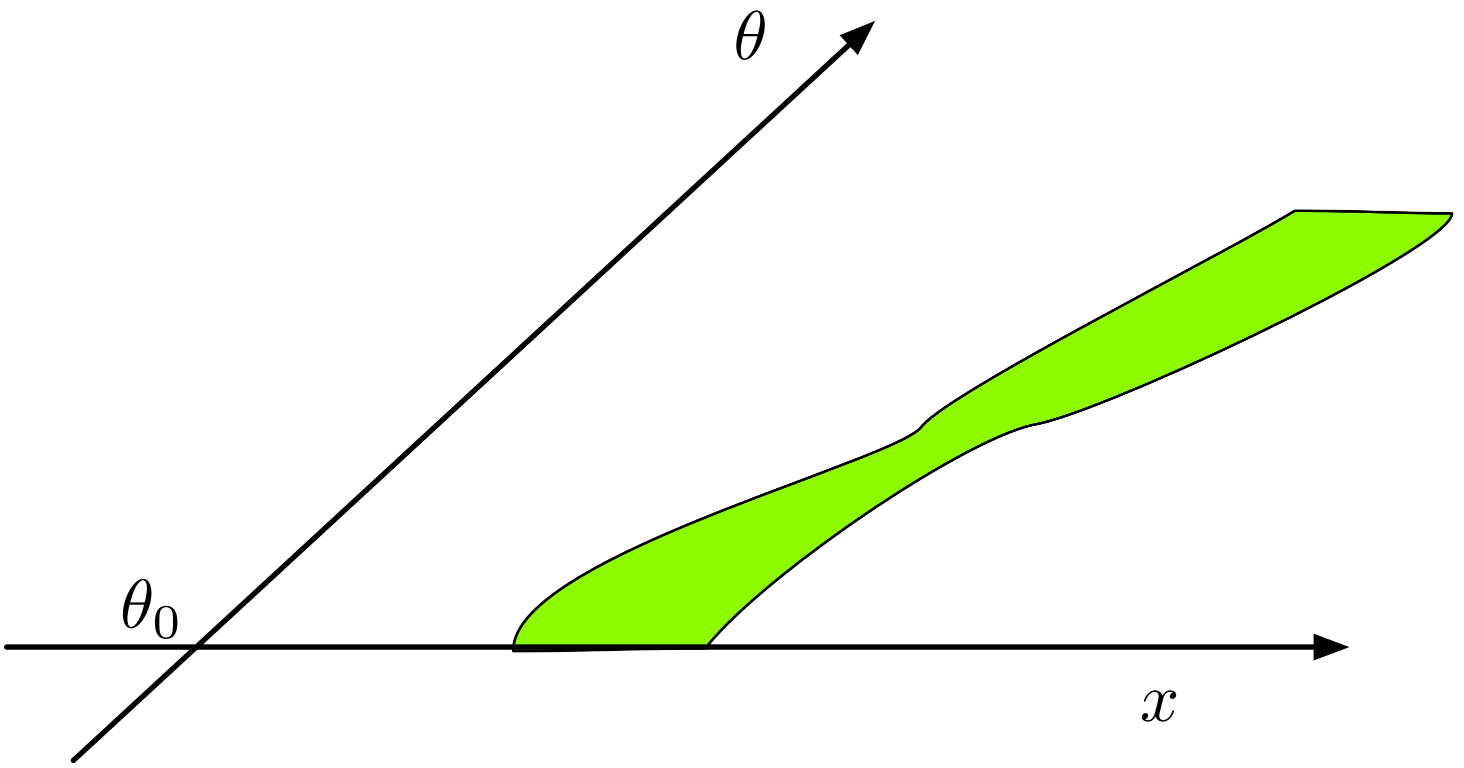
This makes a **confidence belt** for θ



Neyman Construction example

This makes a **confidence belt** for θ

the regions of **data** in the confidence belt can be considered as **consistent** with that value of θ



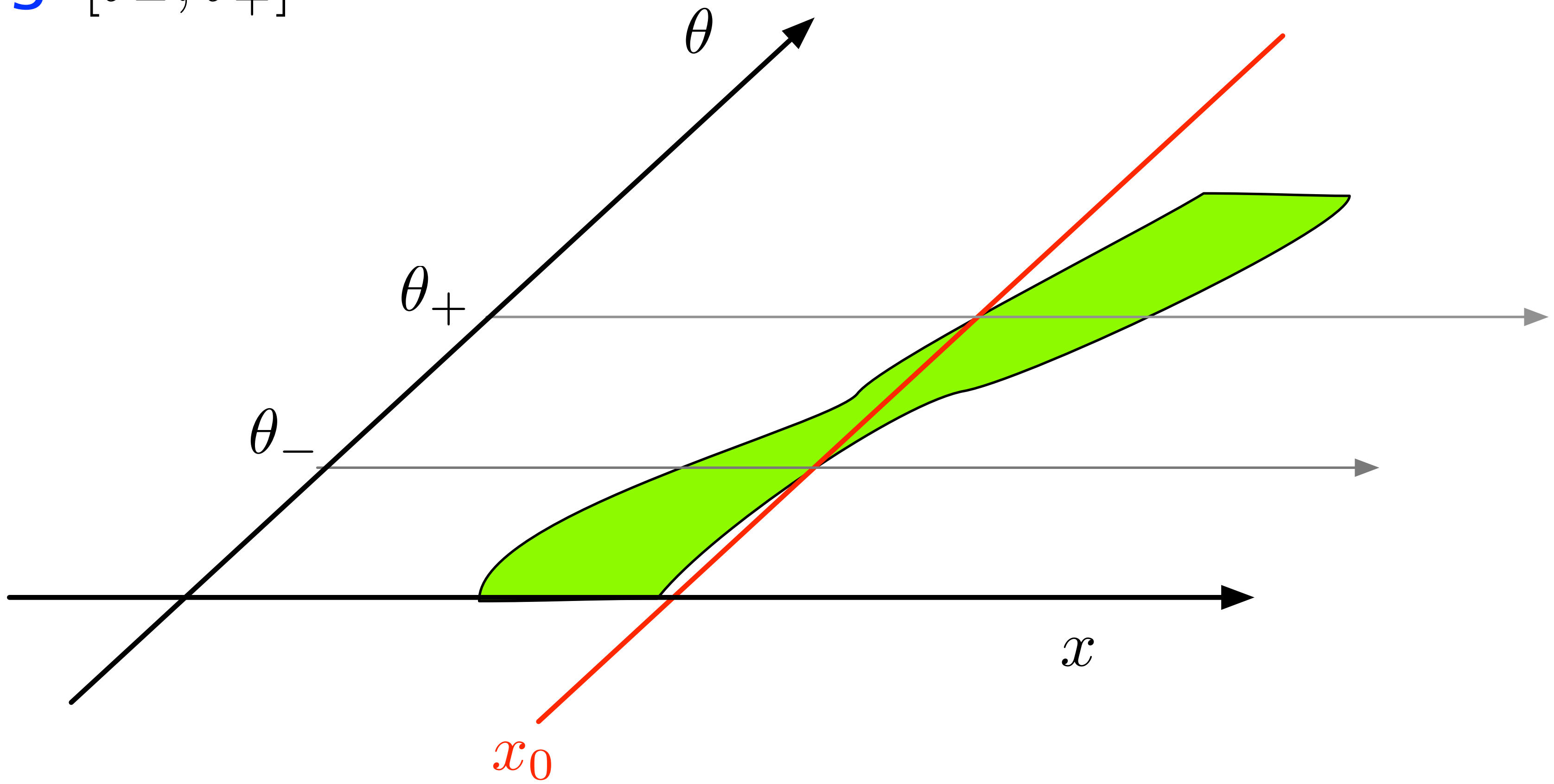
Neyman Construction example

Center for Cosmology and Particle Physics

Now we make a measurement x_0

the points θ where the belt intersects x_0 a part of the **confidence interval** in θ for this measurement

eg. $[\theta_-, \theta_+]$

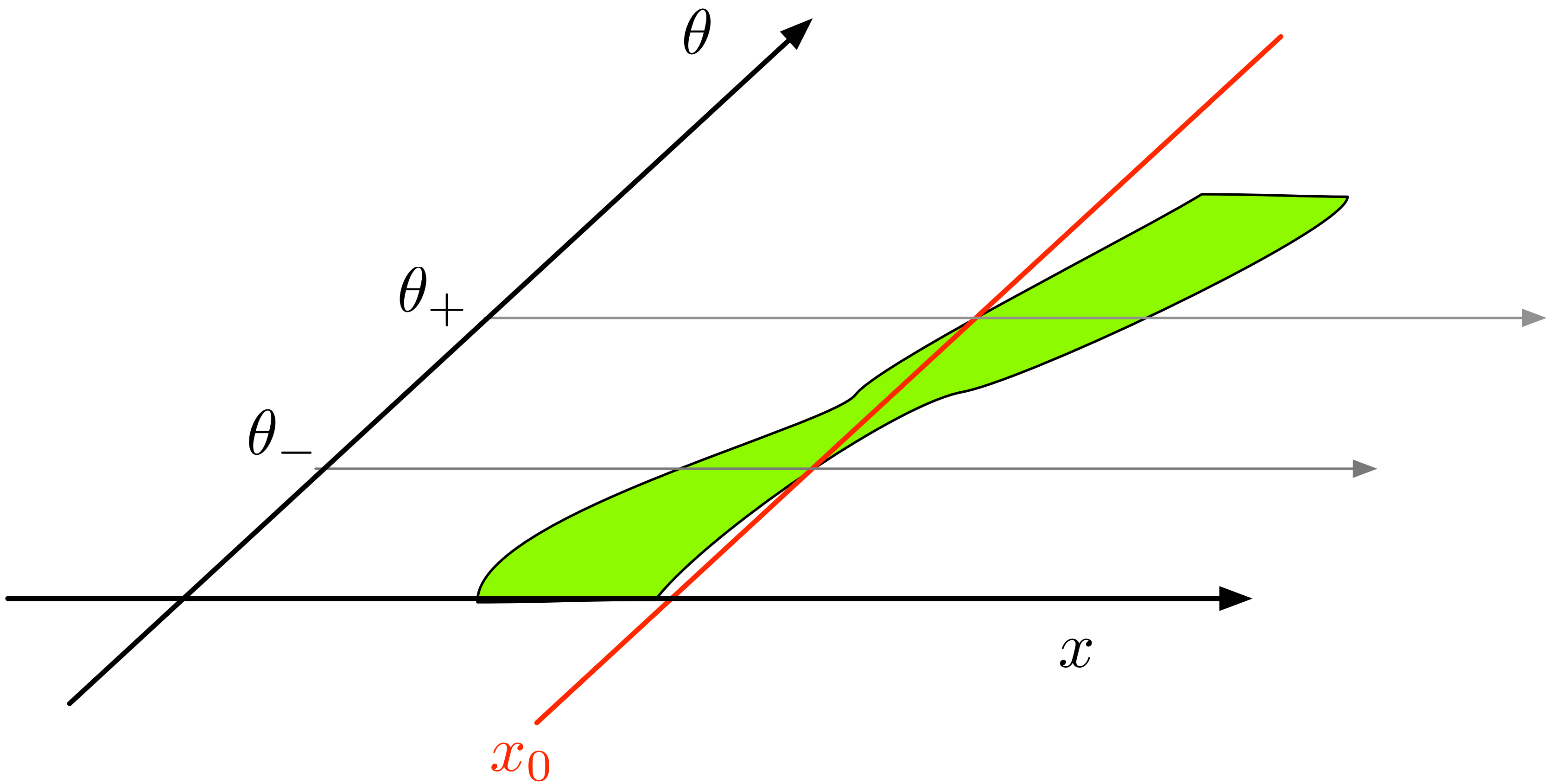


Neyman Construction example

For every point θ, if it were true, the data would fall in its acceptance region with probability $1 - \alpha$

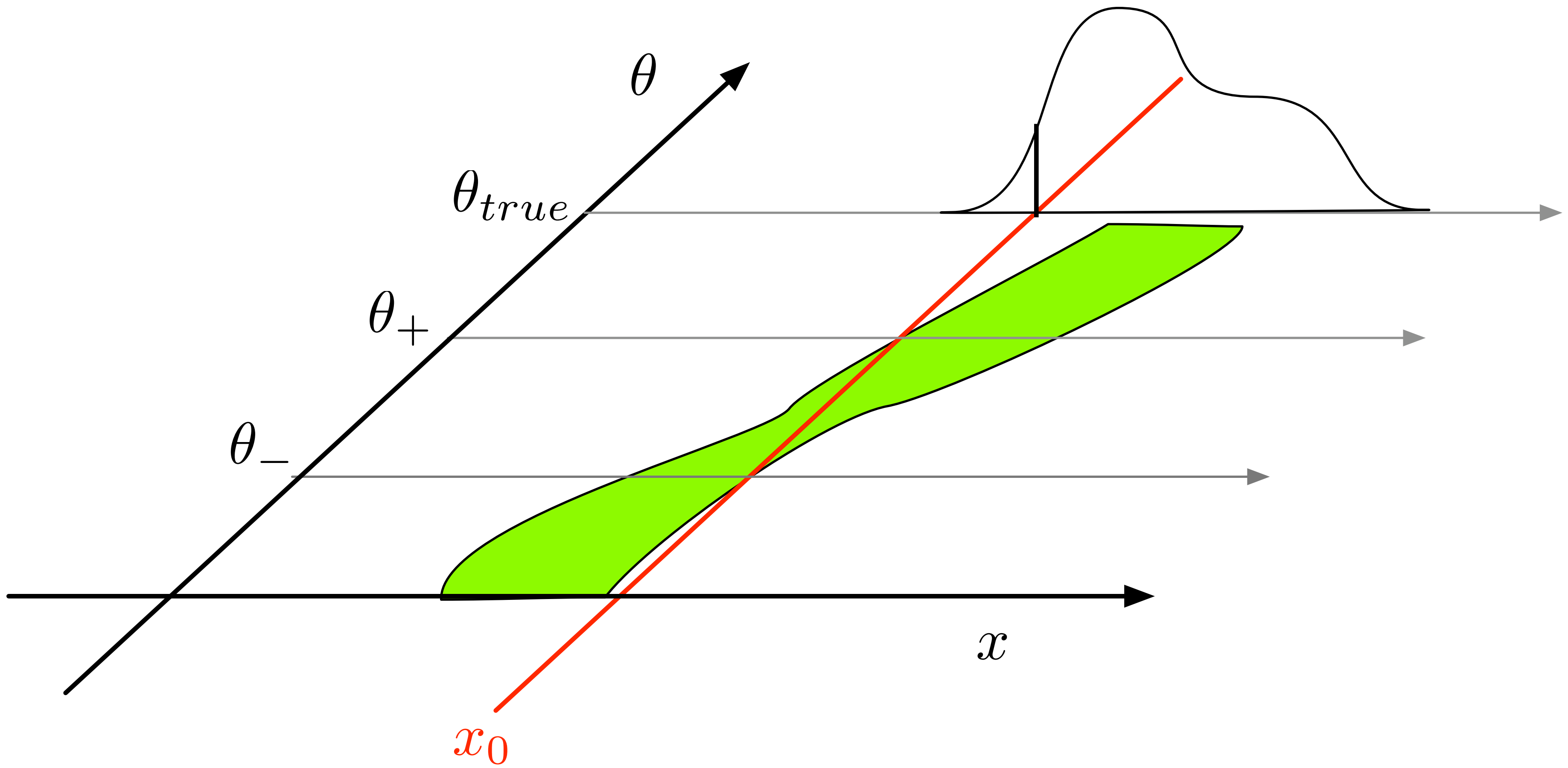
If the data fell in that region, the point θ would be in the interval $[\theta_-, \theta_+]$

So the interval $[\theta_-, \theta_+]$ covers the true value with probability $1 - \alpha$



A Point about the Neyman Construction

This is not Bayesian... it doesn't mean the probability that the true value of θ is in the interval is $1-\alpha$!



Inverting Hypothesis Tests

There is a precise dictionary that explains how to move from from hypothesis testing to parameter estimation.

- **Type I error:** probability interval does not cover true value of the parameters (eg. it is now a function of the parameters)
- **Power** is probability interval does not cover a false value of the parameters (eg. it is now a function of the parameters)
  - We don't know the true value, consider each point θ_0 as if it were true

What about null and alternate hypotheses?

- when testing a point θ_0 it is considered the null
- all other points considered "alternate"

So what about the Neyman-Pearson lemma & Likelihood ratio?

- as mentioned earlier, there are no guarantees like before
- a common generalization that has good power is:

$$\frac{f(x|H_0)}{f(x|H_1)} \Rightarrow \frac{f(x|\theta_0)}{f(x|\theta_{best}(x))}$$

The Dictionary

There is a formal 1-to-1 mapping between hypothesis tests and confidence intervals:

- some refer to the Neyman Construction as an "inverted hypothesis test"

Table 20.1 Relationships between hypothesis testing and interval estimation

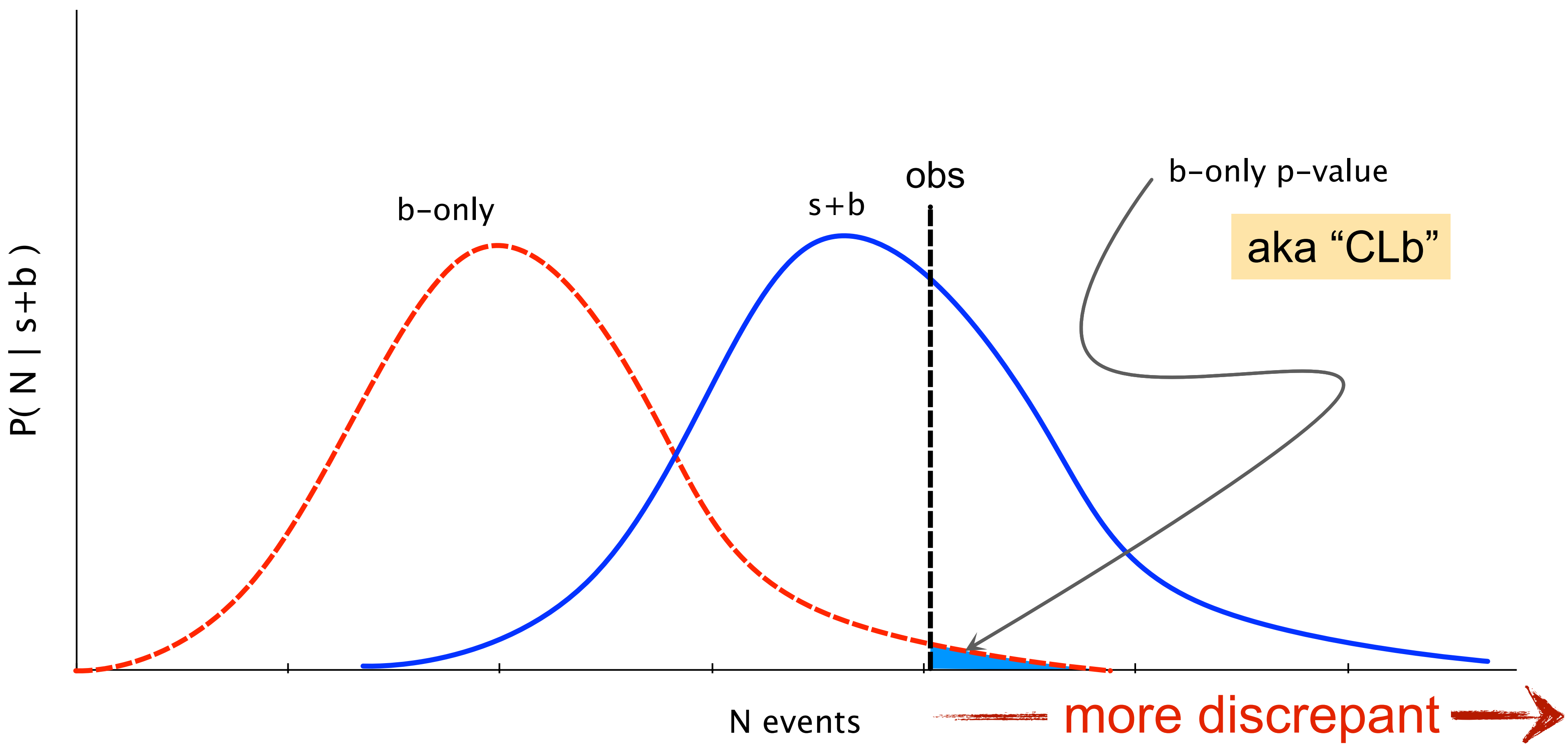
Property of test		Property of corresponding confidence interval
Size $= \alpha$		Confidence coefficient $= 1 - \alpha$
Power $=$ probability of rejecting a false value of $\theta = 1 - \beta$		Probability of not covering a false value of $\theta = 1 - \beta$
Most powerful		Uniformly most accurate
$\longleftarrow$	$\left\{ \begin{array}{c} \textit{Unbiased} \\ 1 - \beta \geq \alpha \end{array} \right\}$	$\longrightarrow$
Equal-tails test $\alpha_1 = \alpha_2 = \frac{1}{2}\alpha$		Central interval

Discovery in pictures

Center for Cosmology and Particle Physics

Discovery: test b-only (null: s=0 vs. alt: s>0)

- note, **one-sided** alternative. larger N is “more discrepant”

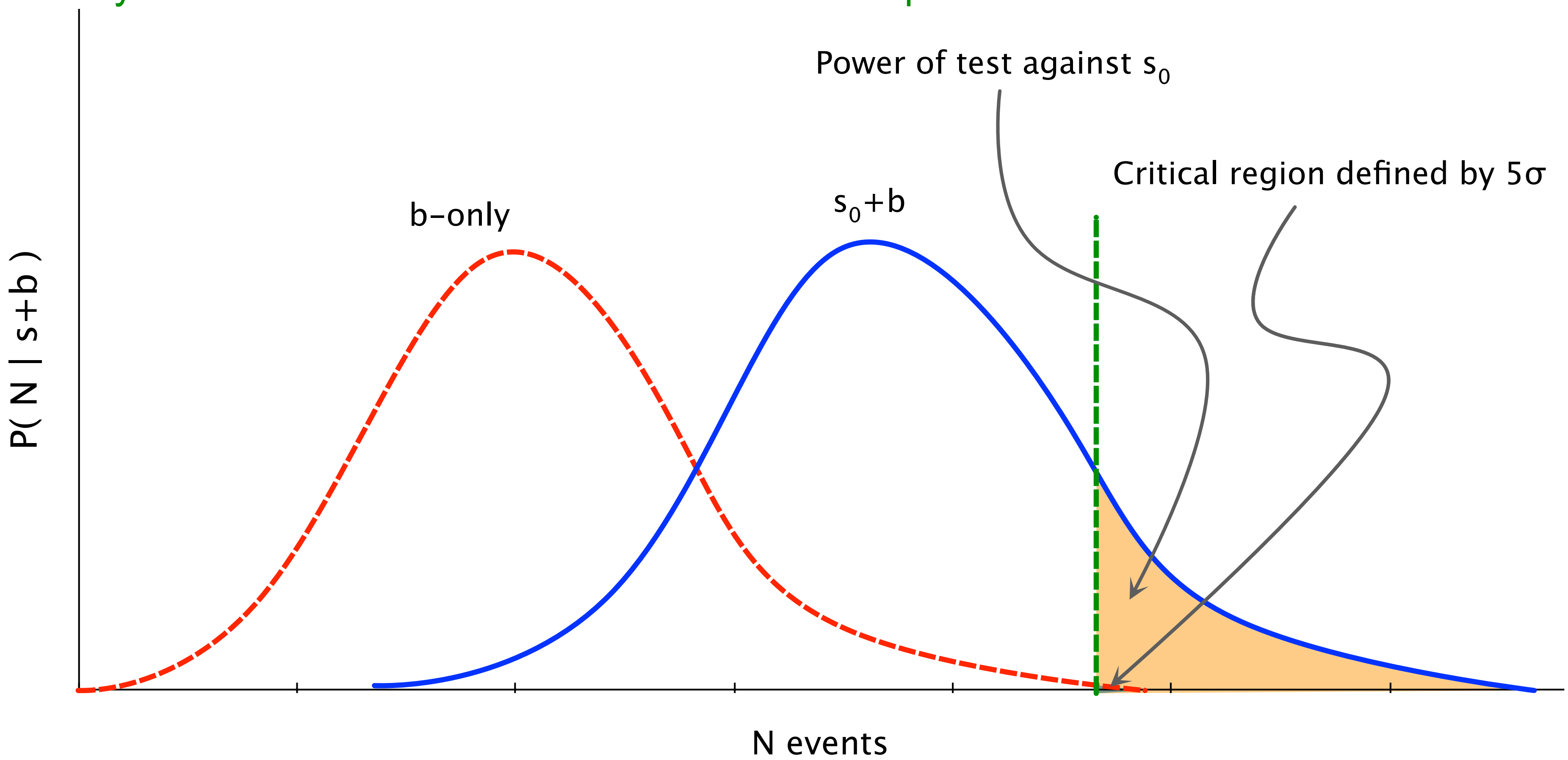


Sensitivity for discovery in pictures

When one specifies 5σ one specifies a critical value for the data before "rejecting the null".

Leaves open a question of sensitivity, which is quantified as "power" of the test against a specific alternative

- In Frequentist setup, one chooses a "test statistic" to maximize power
  - Neyman-Pearson lemma: likelihood ratio most powerful test for one-sided alternative



Measurements in pictures

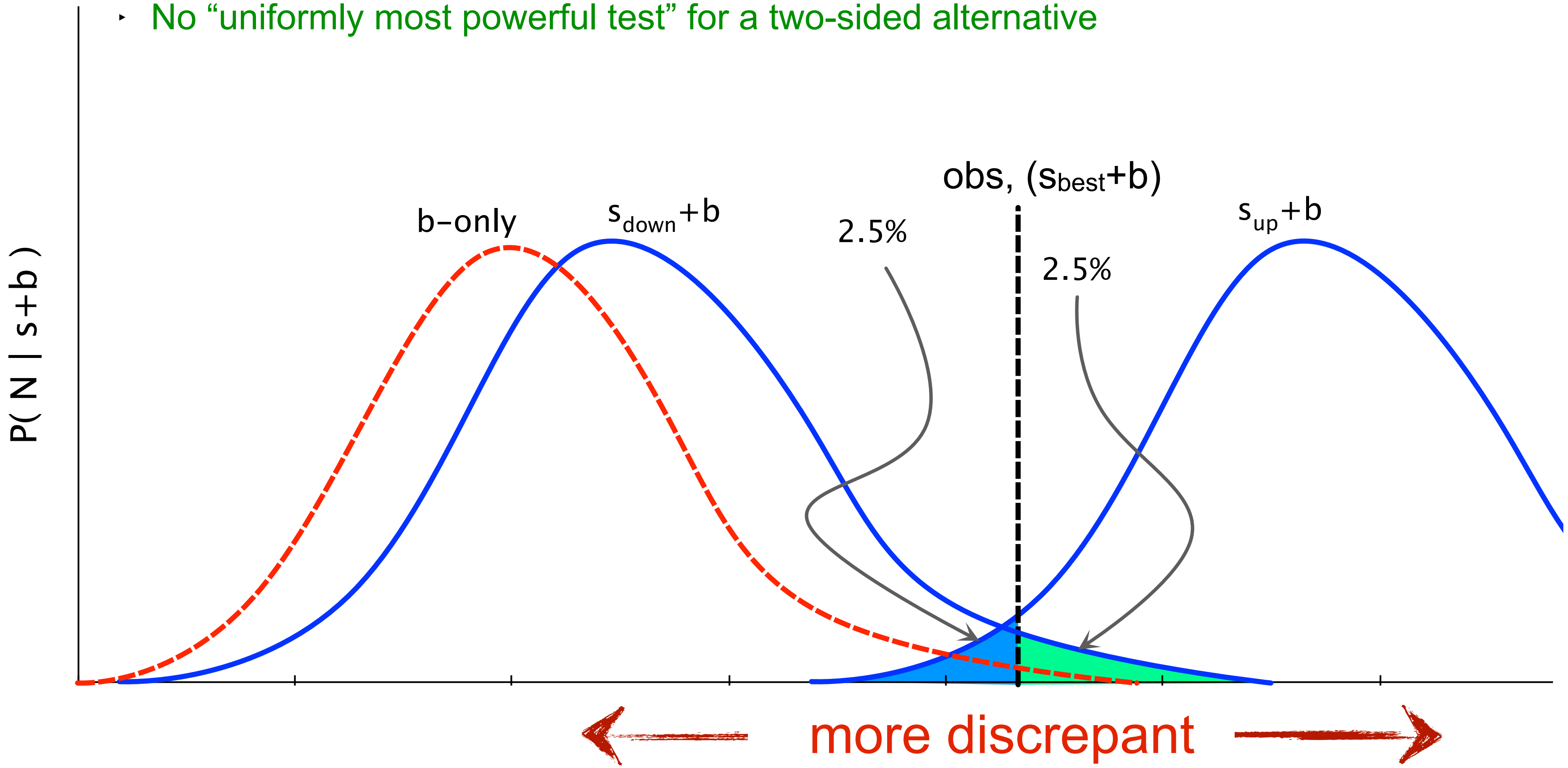
Center for
Cosmology and
Particle Physics

Measurement typically denoted σ = X± Y.

- X is usually the “best fit” or maximum likelihood estimate
- ±Y usually means [X-Y, X+Y] is a 68% confidence interval

Intervals are formally “inverted hypothesis tests”: (null: $s=s_0$ vs. alt: $s \neq s_0$)

- One hypothesis test for each value of s_0 against a **two-sided** alternative
  - No “uniformly most powerful test” for a two-sided alternative

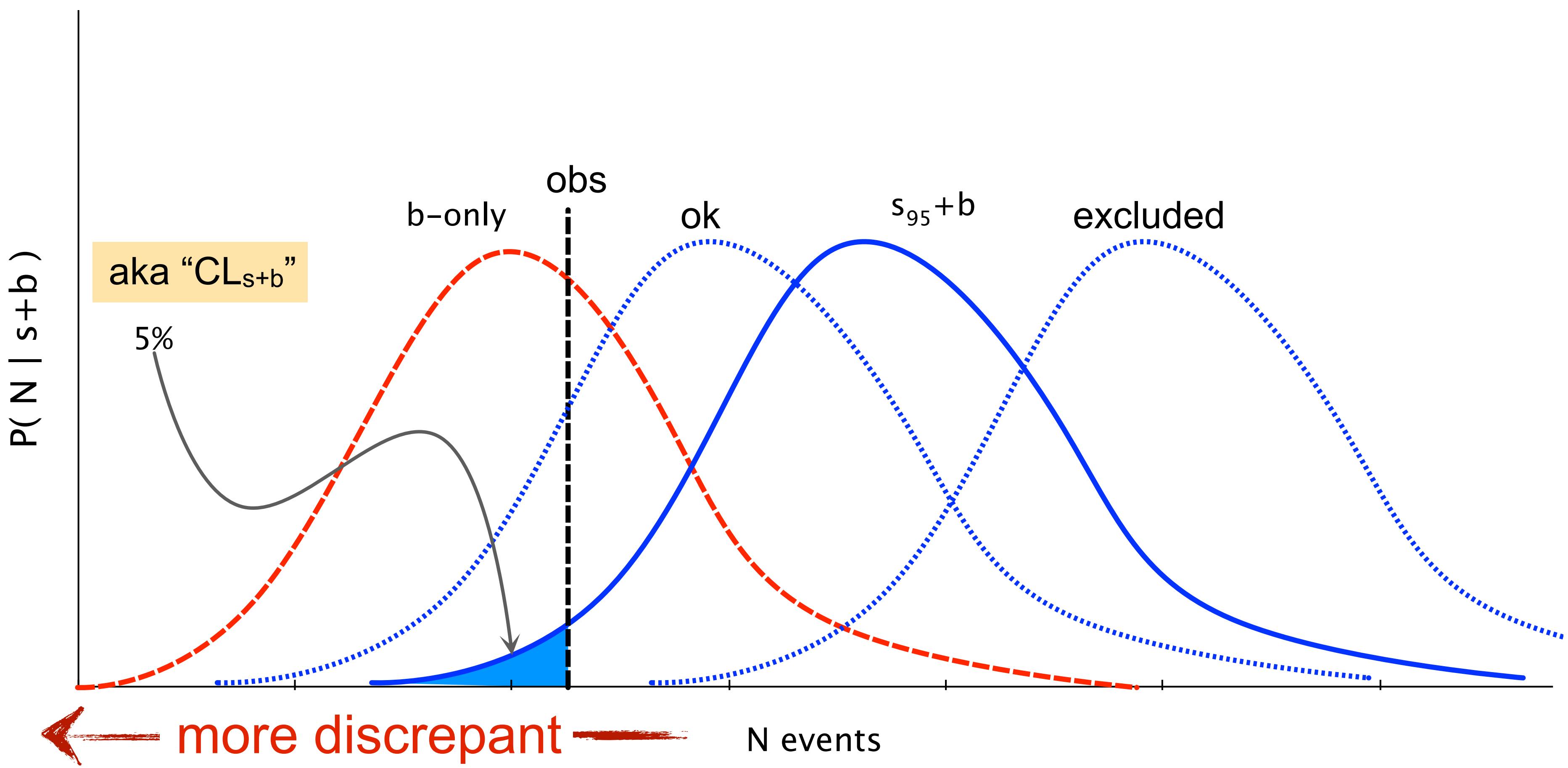


Center for Cosmology and Particle Physics

What do you think is meant by "95% upper limit" ?

Is it like the picture below?

- ie. increase s, until the probability to have data "more discrepant" is < 5%



Upper limits in pictures

Upper-limits are trying to exclude large signal rates.

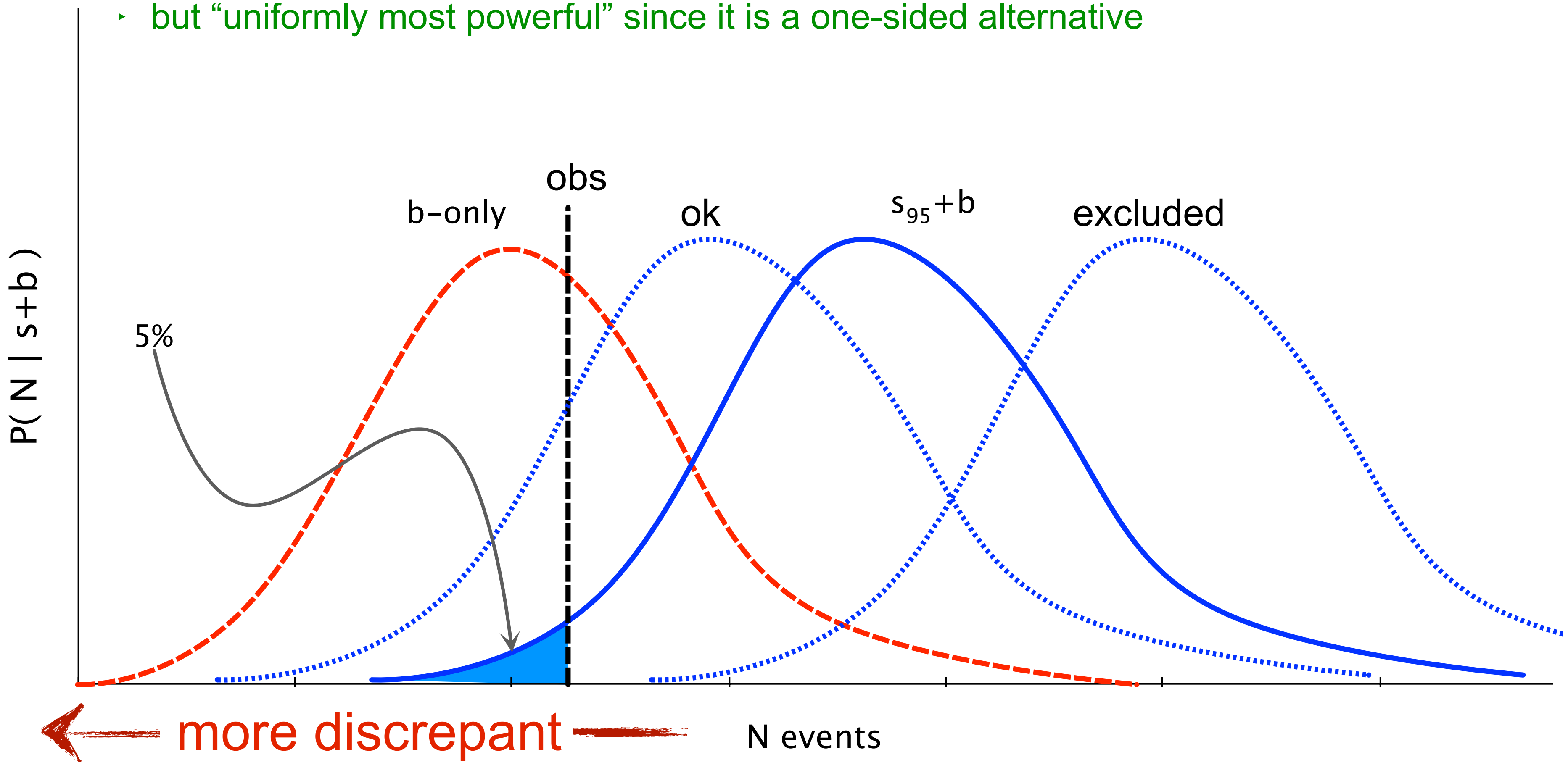
- form a 95% "confidence interval" on s of form $[0, s_{95}]$

Intervals are formally "inverted hypothesis tests": (null: $s=s_0$ vs. alt: $s<s_0$)

- One hypothesis test for each value of s_0 against a **one-sided** alternative

Power of test depends on specific values of null s_0 and alternate s'

- but "uniformly most powerful" since it is a one-sided alternative

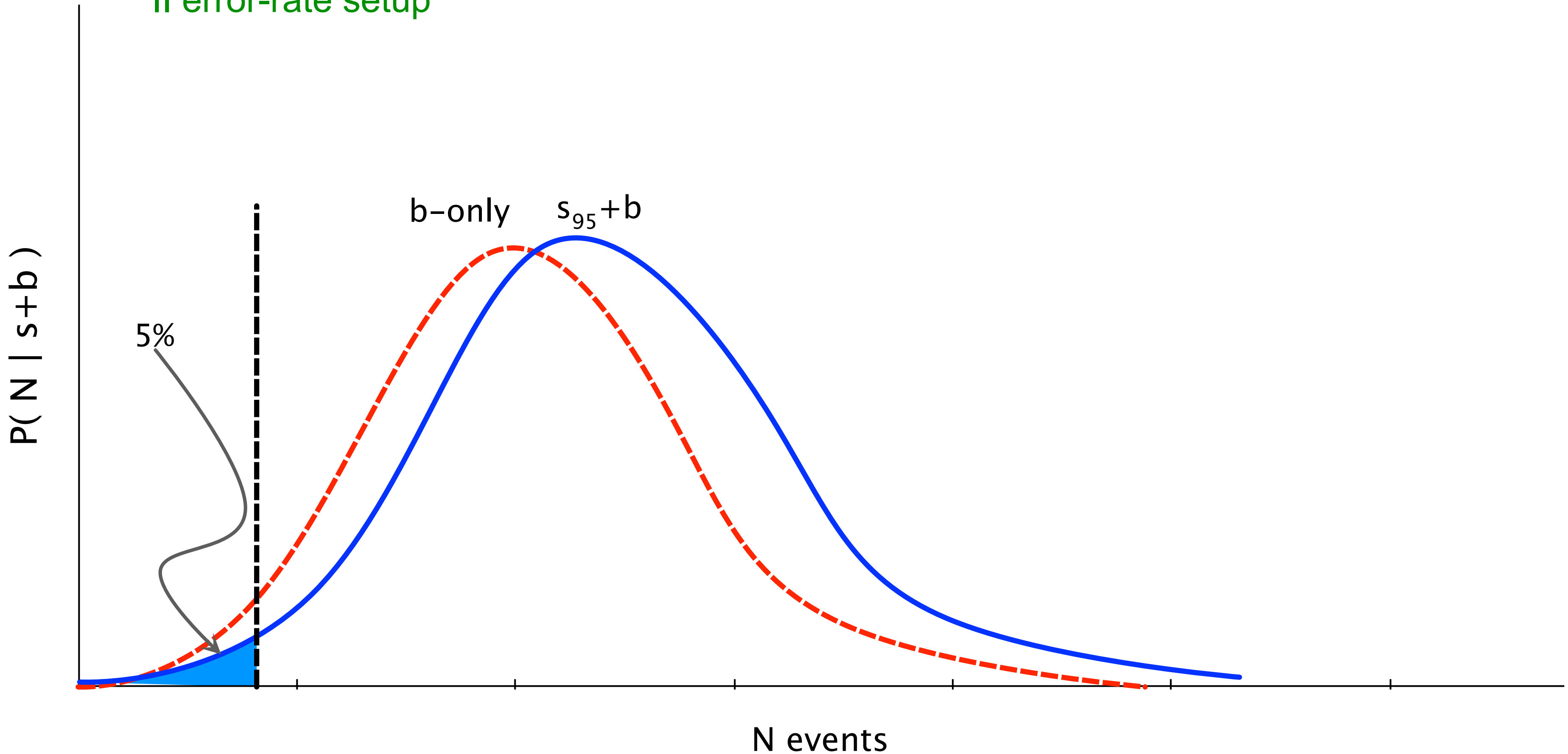


The sensitivity problem

Center for Cosmology and Particle Physics

The physicist's worry about limits in general is that if there is a strong downward fluctuation, one might exclude arbitrarily small values of s

- with a procedure that produces proper frequentist 95% confidence intervals, one should expect to exclude the true value of s 5% of the time, no matter how small s is!
  - This is not a problem with the procedure, but an undesirable consequence of the Type I / Type II error-rate setup

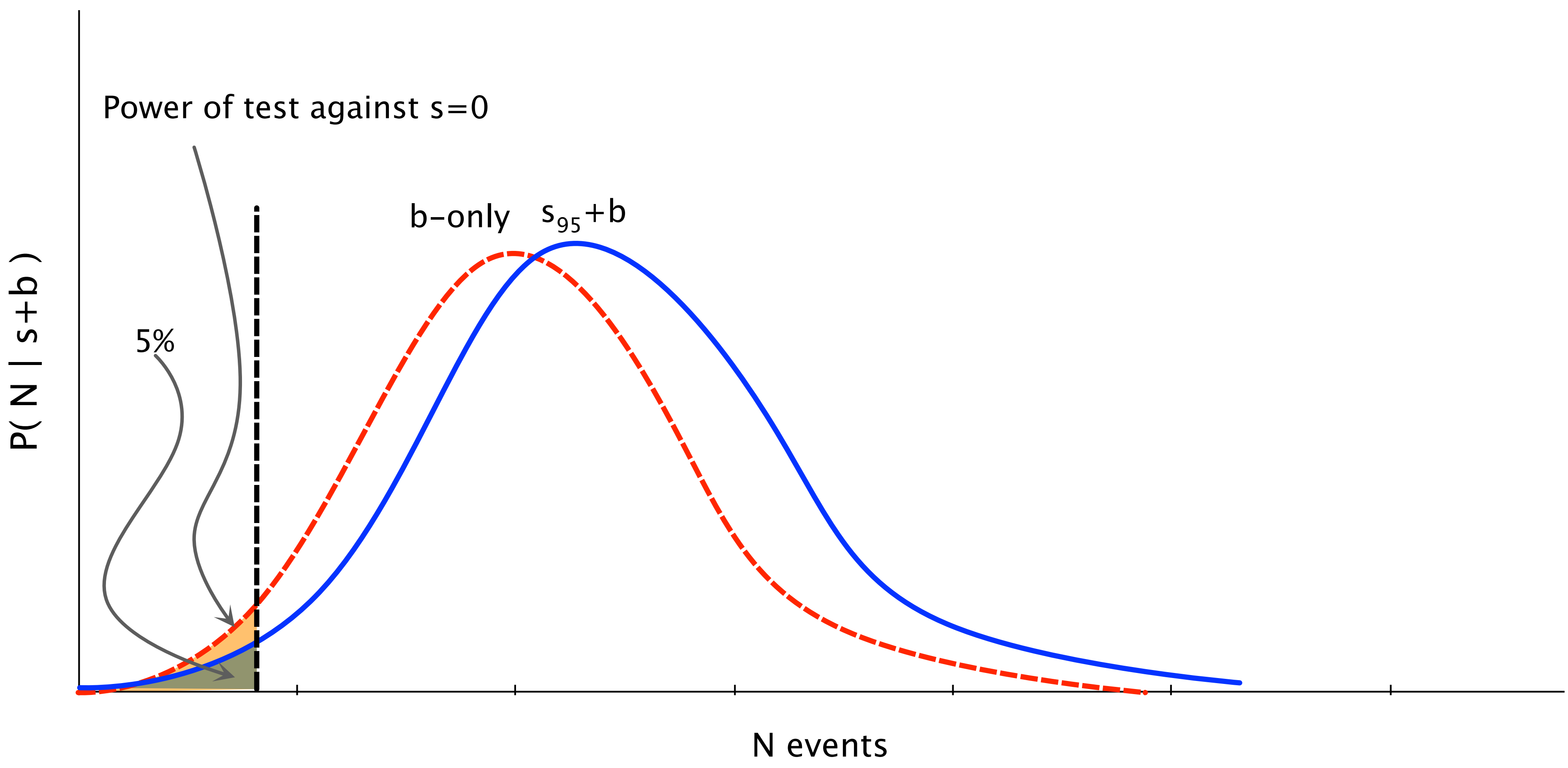


Power in the context of limits

Center for Cosmology and Particle Physics

Remember, when creating confidence intervals the null is $s=s_0$

- and power is defined under a specific alternative (eg. s=0)



CL_s

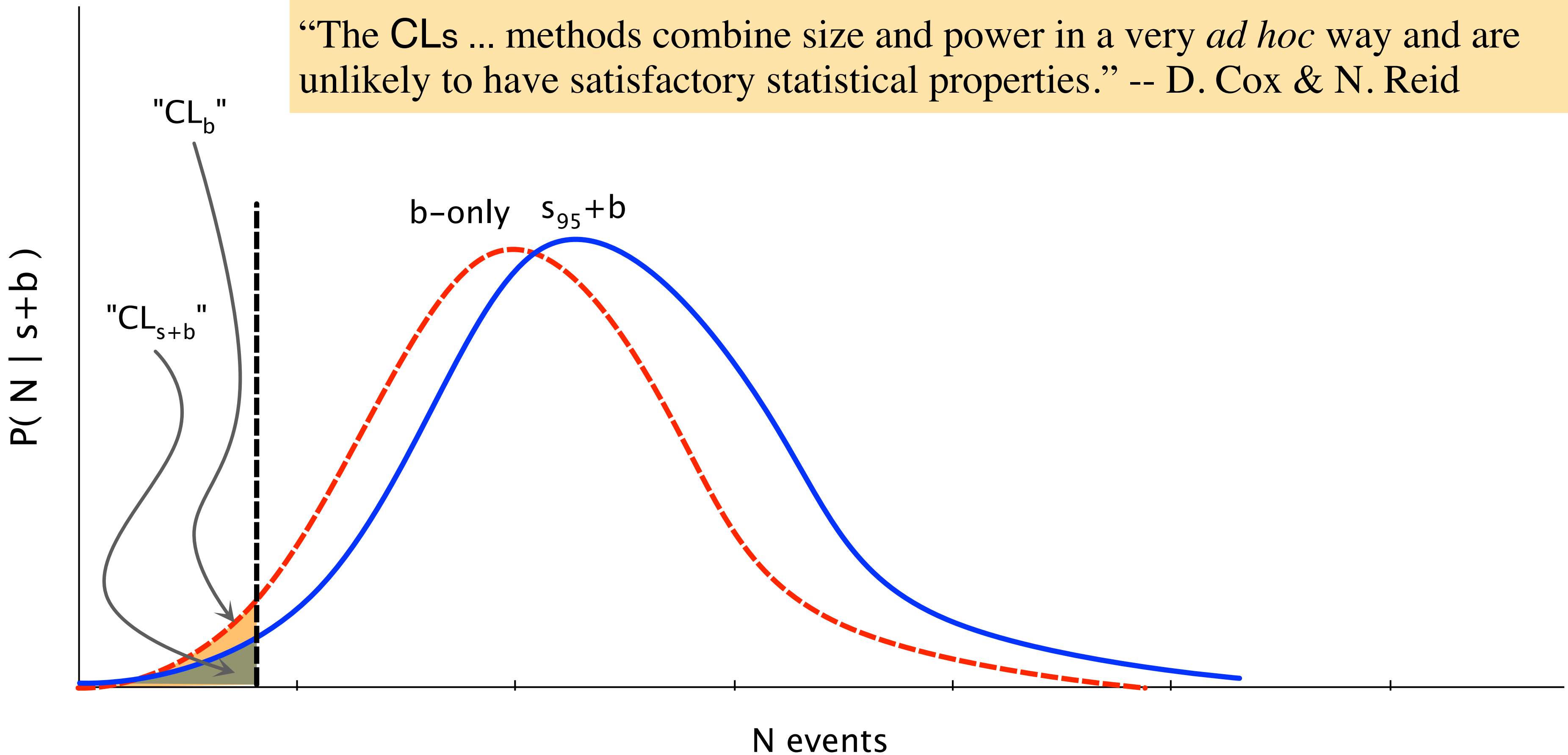
To address the sensitivity problem, CLs was introduced

- common (misused) nomenclature: $CL_s = CL_{s+b}/CL_b$
- idea: only exclude if CL_s<5% (if CL_b is small, CL_s gets bigger)

CL_s is known to be "conservative" (over-cover): expected limit covers with 97.5%

- Note: CL_s is NOT a probability

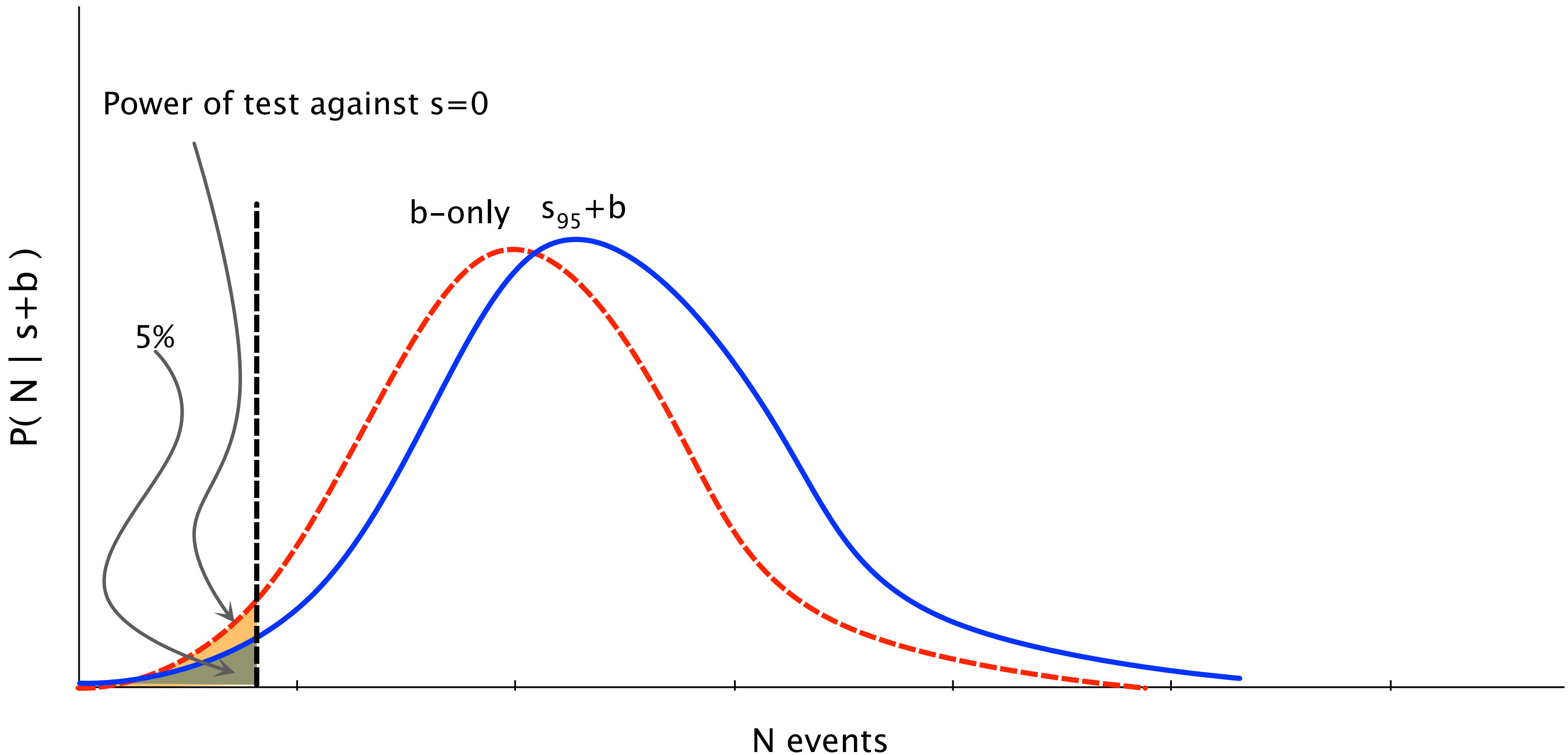
"The CLs ... methods combine size and power in a very *ad hoc* way and are unlikely to have satisfactory statistical properties." -- D. Cox & N. Reid



The Power Constraint

An alternative to CLs that protects against setting limits when one has no sensitivity is to explicitly define the sensitivity of the experiment in terms of power.

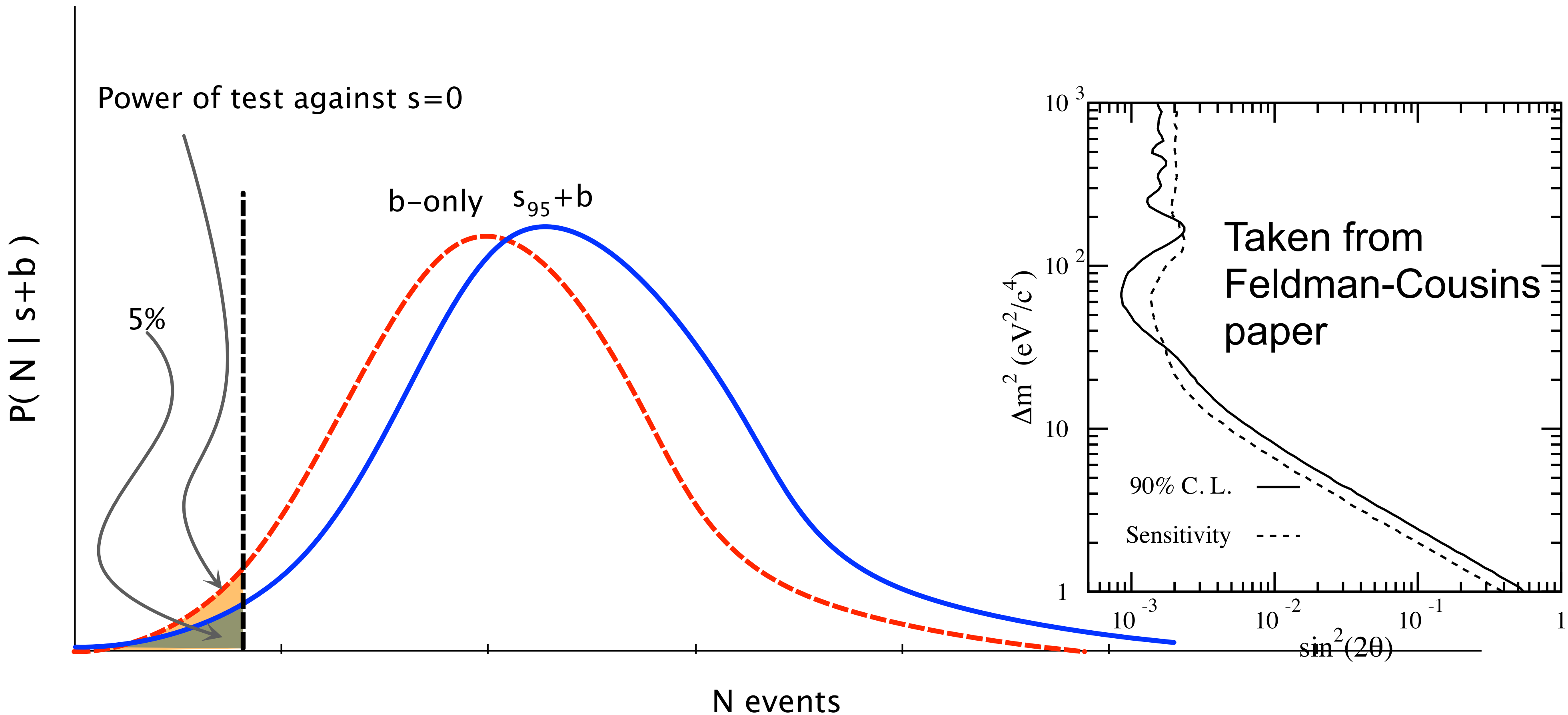
- A clean separation of size and power. (a new, arbitrary threshold for sensitivity)
- Feldman-Cousins foreshadowed the recommendation sensitivity defined as 50% power against b-only
- David van Dyk presented similar idea at PhyStat2011 [arxiv.org:1006.4334]



The Power Constraint

An alternative to CLs that protects against setting limits when one has no sensitivity is to explicitly define the sensitivity of the experiment in terms of power.

- A clean separation of size and power. (a new, arbitrary threshold for sensitivity)
- Feldman-Cousins foreshadowed the recommendation sensitivity defined as 50% power against b-only
- David van Dyk presented similar idea at PhyStat2011 [arxiv.org:1006.4334]

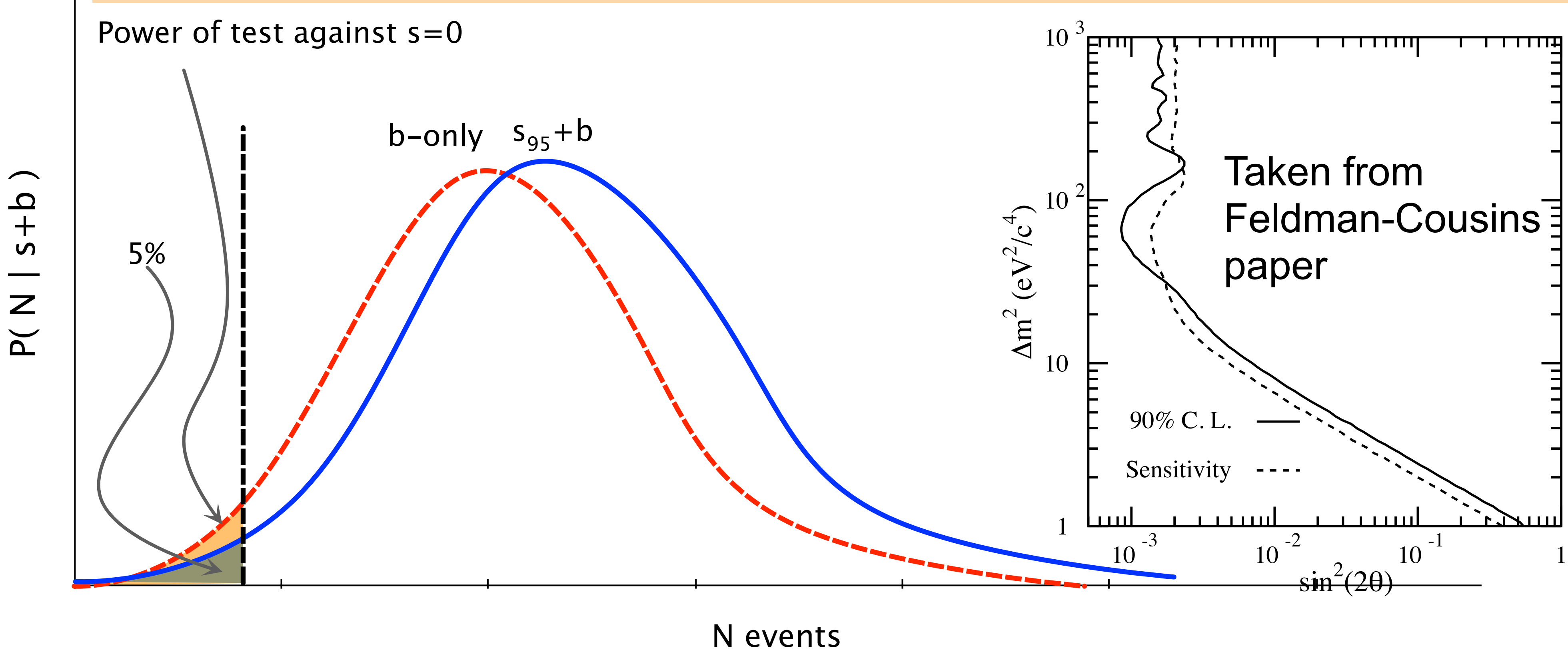


The Power Constraint

An alternative to CLs that protects against setting limits when one has no sensitivity is to explicitly define the sensitivity of the experiment in terms of power.

- A clean separation of size and power. (a new, arbitrary threshold for sensitivity)
- Feldman-Cousins foreshadowed the recommendation sensitivity defined as 50% power against b-only
- David van Dyk presented similar idea at PhyStat2011 [arxiv.org:1006.4334]

“ Both measures are useful quantities that should be reported in order to extract the most science from catalogs”

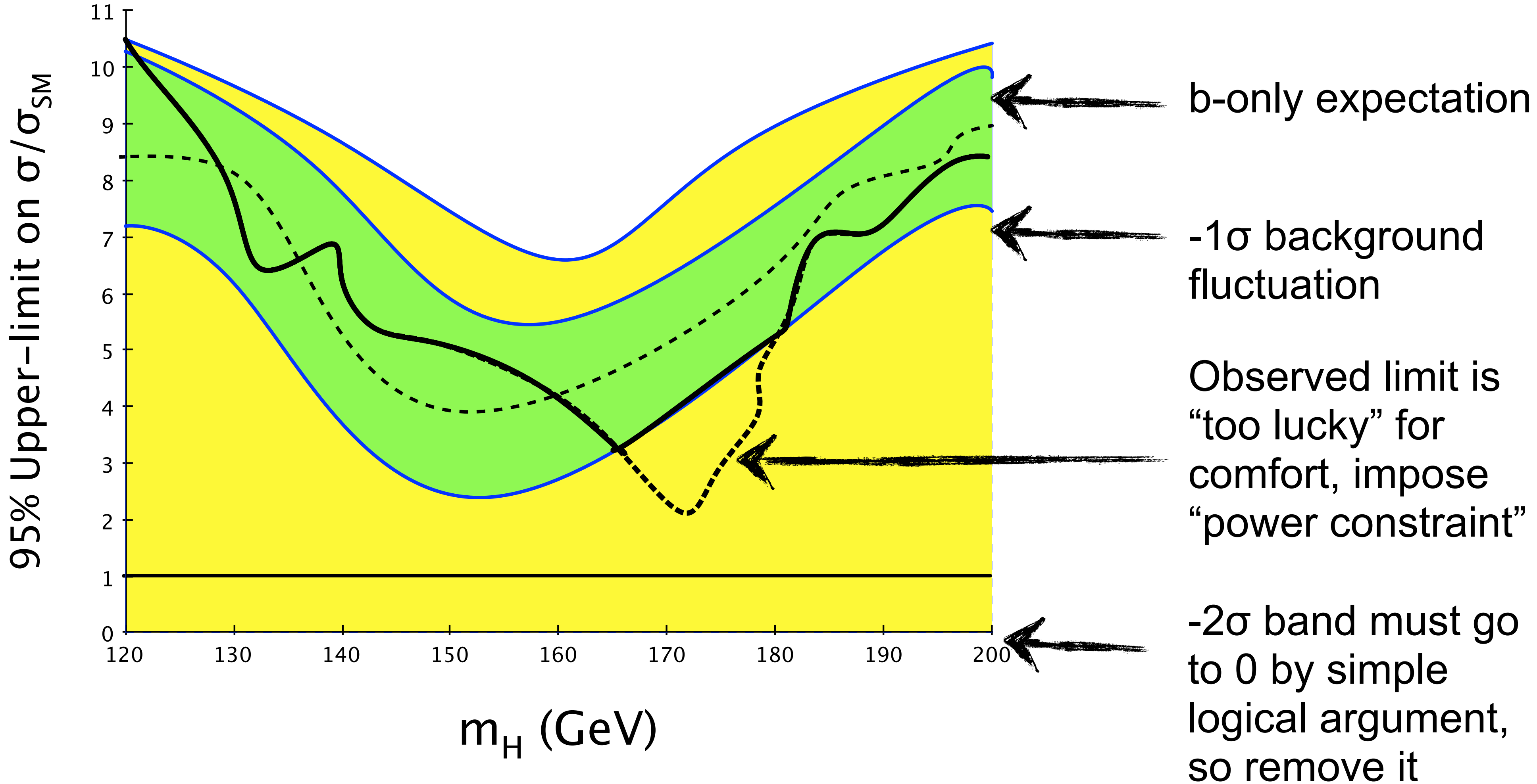


“Power-Constrained” CL_{s+b} limits

CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS

Even for s=0, there is a 5% chance of a strong downward fluctuation that would exclude the background-only hypothesis

- we don’t want to exclude signals for which we have no sensitivity
- idea: don’t quote limit below some threshold defined by an N-σ downward fluctuation of b-only pseudo-experiments (Choose -1σ by convention)

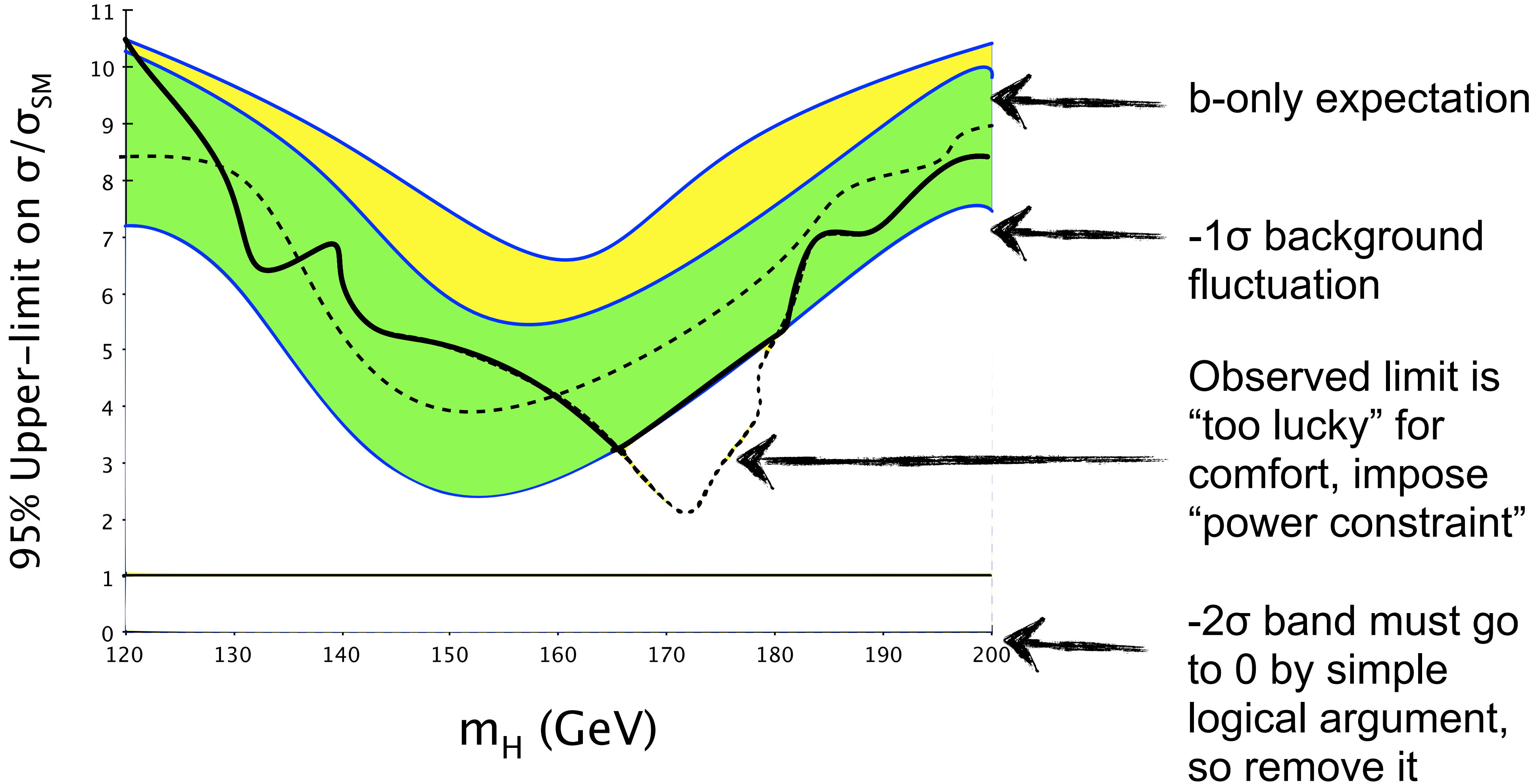


“Power-Constrained” CL_{s+b} limits

Center for
Cosmology and
Particle Physics

Even for s=0, there is a 5% chance of a strong downward fluctuation that would exclude the background-only hypothesis

- we don’t want to exclude signals for which we have no sensitivity
- idea: don’t quote limit below some threshold defined by an N-σ downward fluctuation of b-only pseudo-experiments (Choose -1σ by convention)



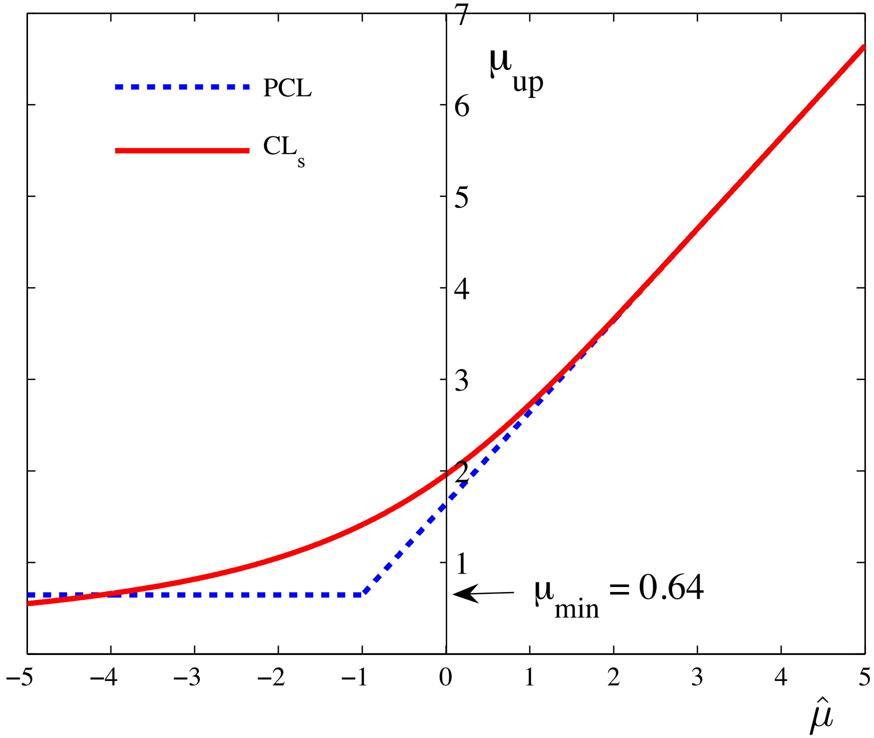
Coverage Comparison with CLs

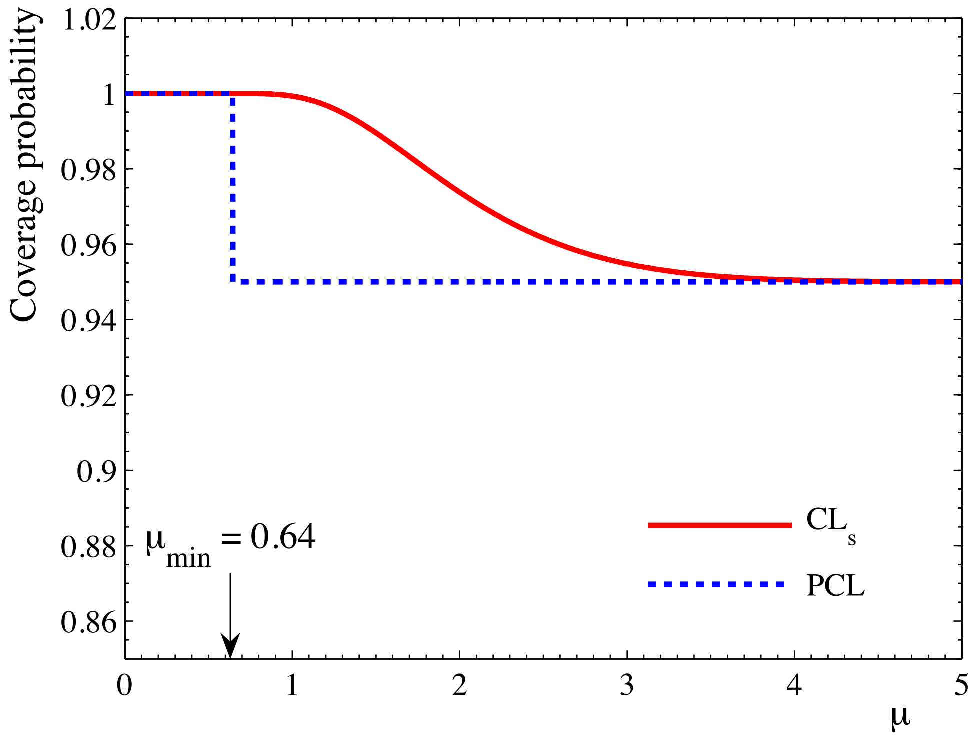
The CLs procedure purposefully over-covers ("conservative")

- and it is not possible for the reader to determine by how much

The power-constrained approach has the specified coverage until the constraint is applied, at which point the coverage is 100%

- limits are not 'aggressive' in the sense that they under-cover





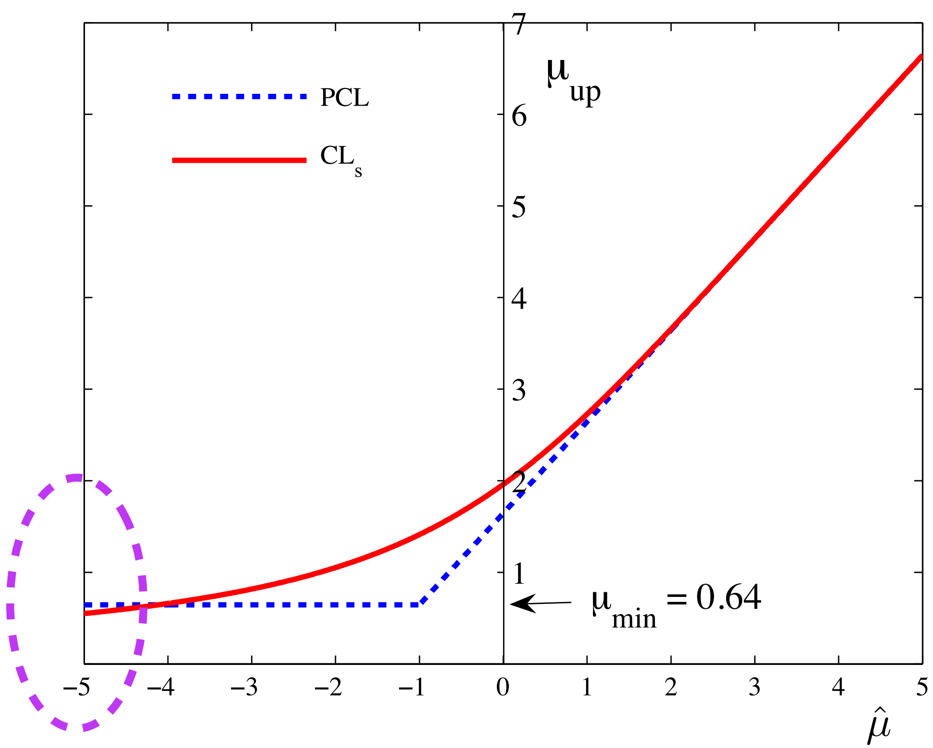
Coverage Comparison with CLs

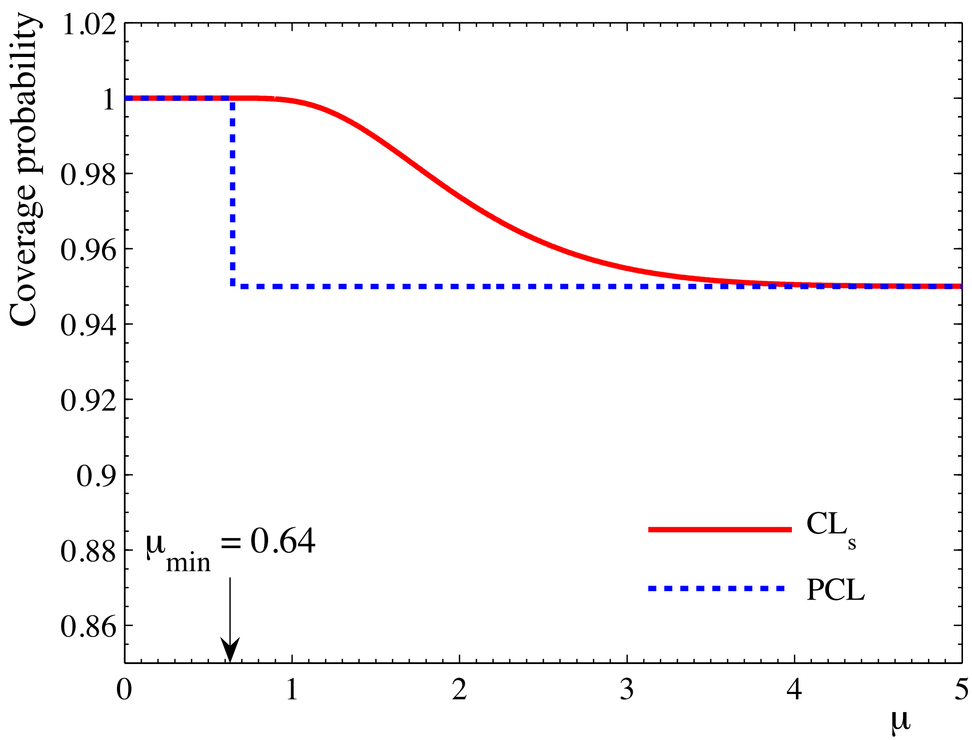
The CLs procedure purposefully over-covers ("conservative")

- and it is not possible for the reader to determine by how much

The power-constrained approach has the specified coverage until the constraint is applied, at which point the coverage is 100%

- limits are not 'aggressive' in the sense that they under-cover

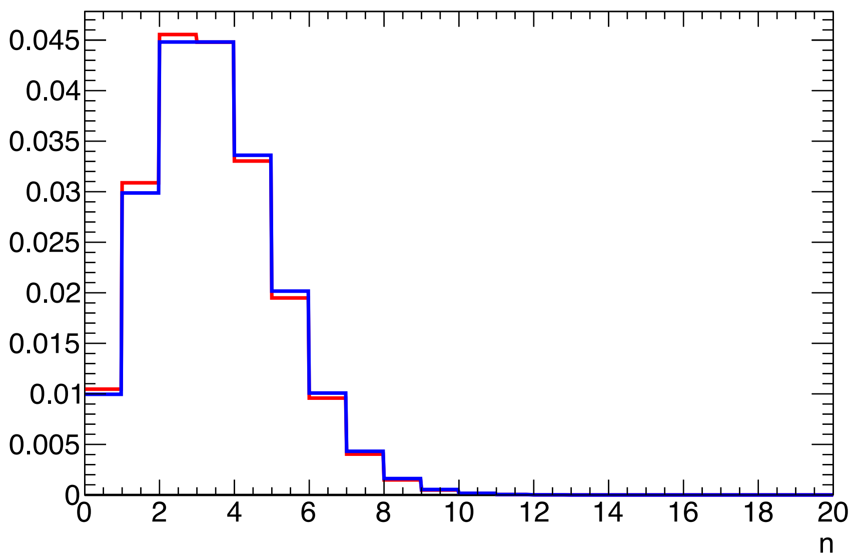


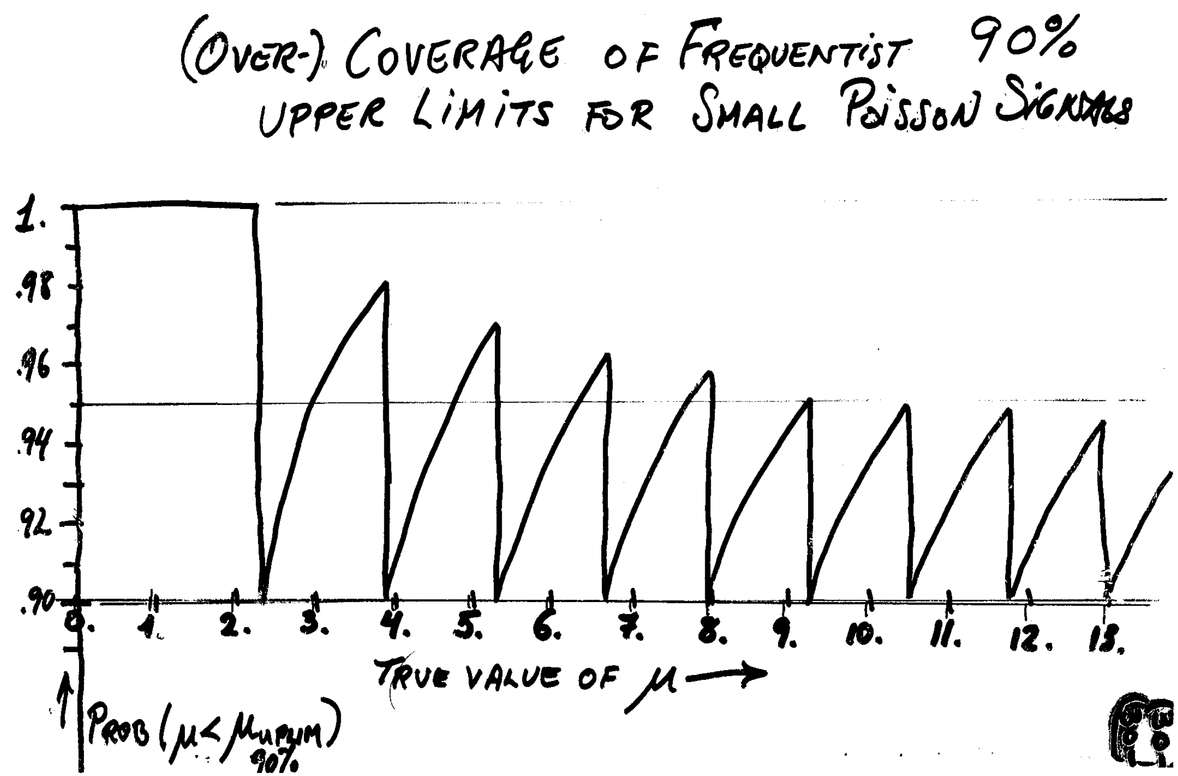


Discrete Problems

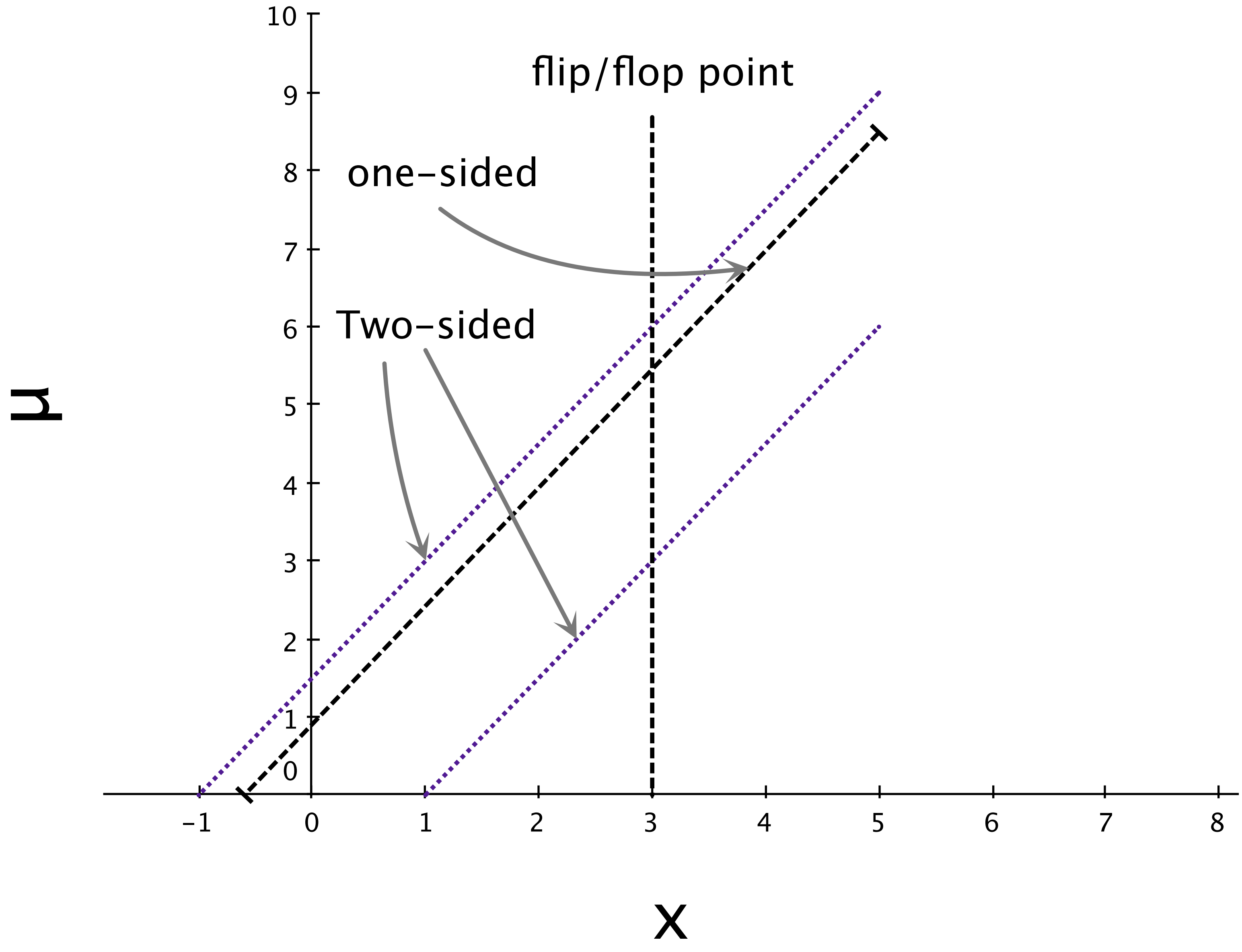
In discrete problems (eg. number counting analysis with counts described by a Poisson) one sees:

- discontinuities in the coverage (as a function of parameter)
- over-coverage (in some regions)
- Important for experiments with few events. There is a lot of discussion about this, not focusing on it here

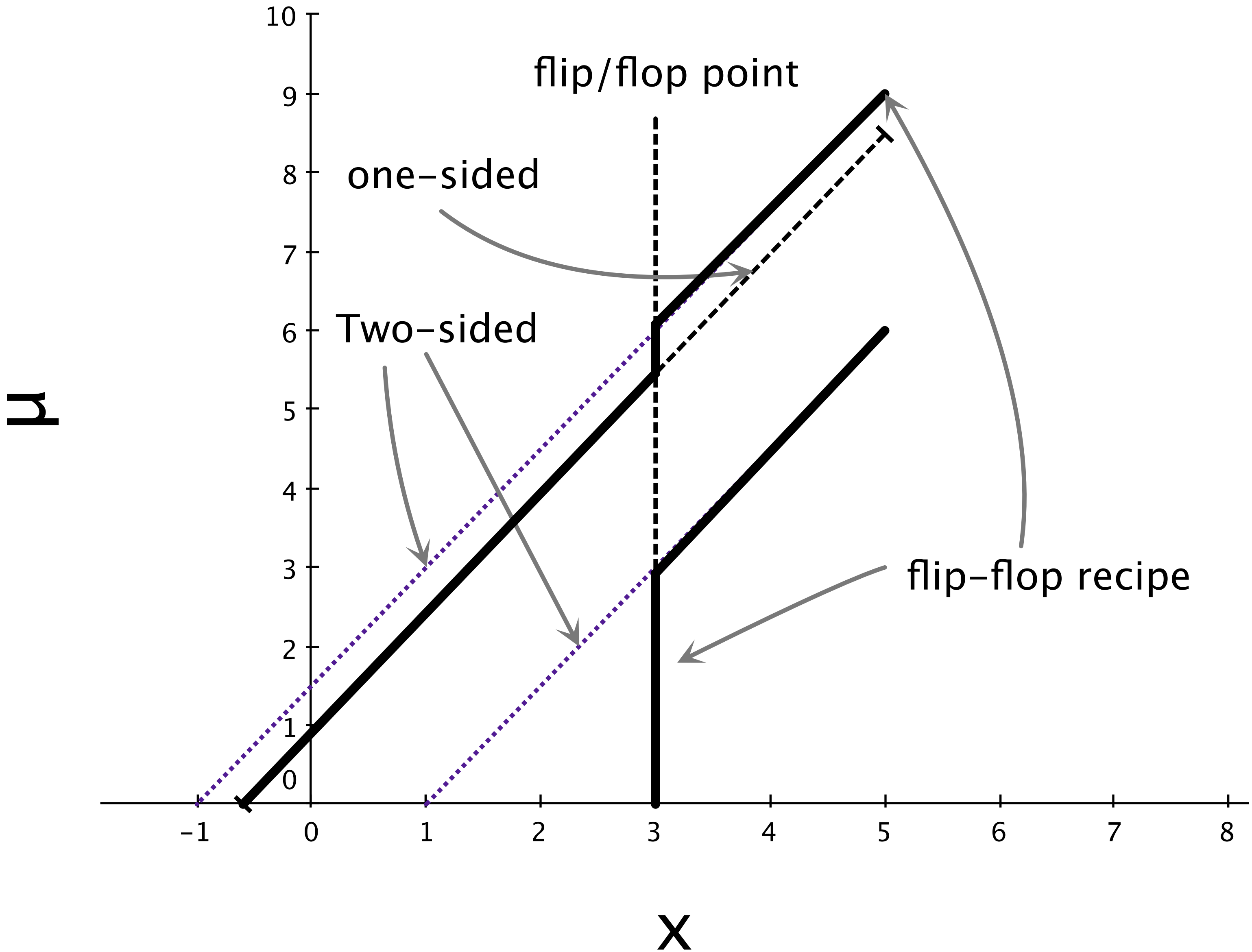


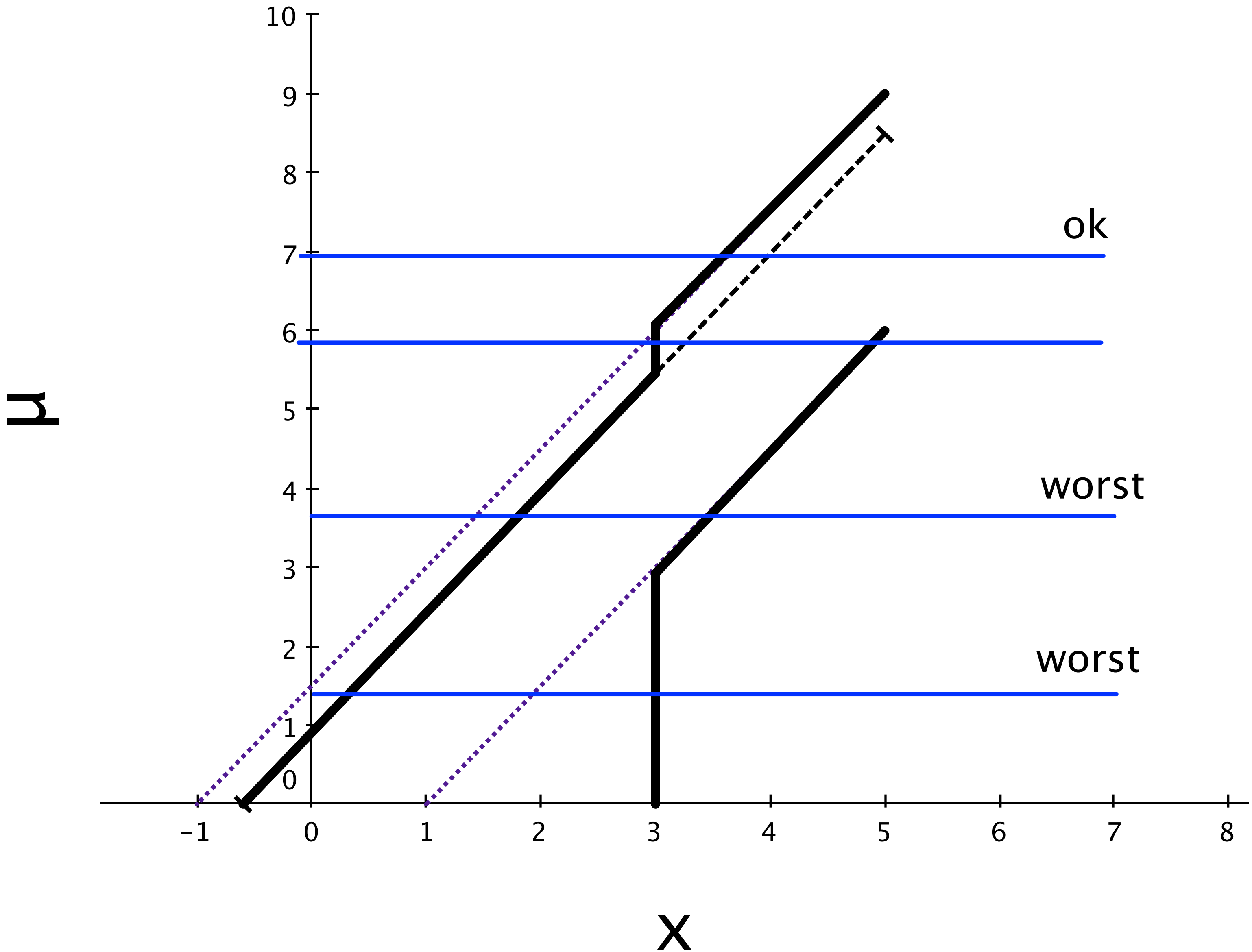


Flip-Flopping



flip/flop point
one-sided
Two-sided
μ
x
10
9
8
7
6
5
4
3
2
1
0
−1
0
1
2
3
4
5
6
7
8


flip/flop point
one-sided
Two-sided
flip-flop recipe
μ
x
10
9
8
7
6
5
4
3
2
1
0
-1
0
1
2
3
4
5
6
7
8


10
9
8
7
6
5
4
3
2
1
0
−1
0
1
2
3
4
5
6
7
8
ok
worst
worst
μ
x

Flip-flopping coverage

Center for Cosmology and Particle Physics

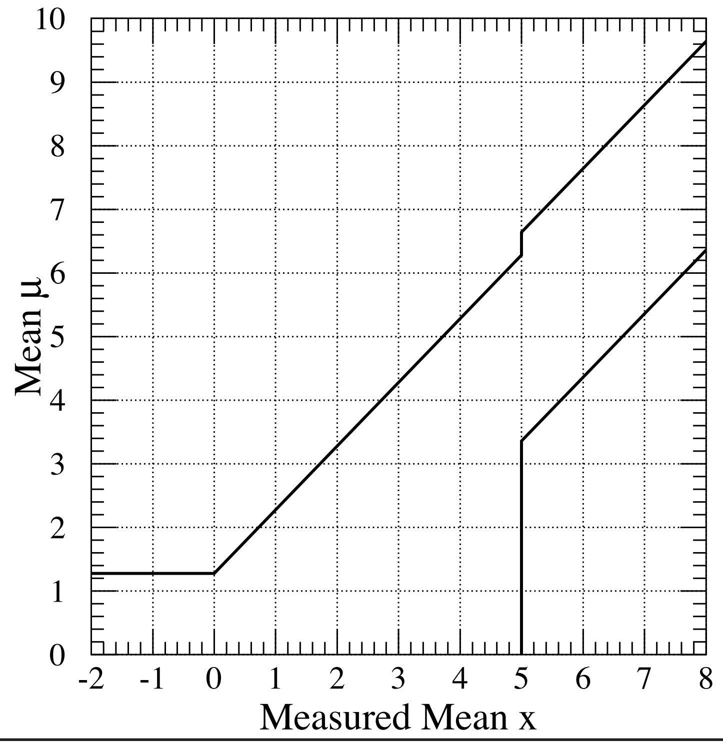
The flip-flopping procedure will under-cover

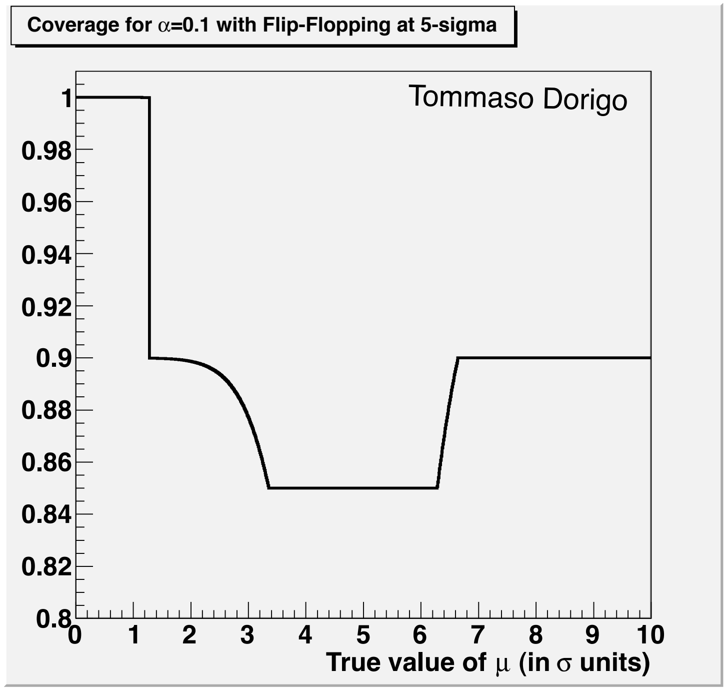
- can be avoided with a 'unified method' or if we always provide both p-value for b-only and 1-sided upper-limit

"As is emphasized in Neal [4], upper and lower one-sided confidence limits should replace confidence intervals, and a full plot of the log-likelihood function is better still." - D. Cox, N. Reid

In practice, we care about coverage on physical parameters (eg. a cross-section, not the number of events). This leads to a subtle semi-philosophical point

- So the relevant 'ensemble' of experiments may be different. With 100x more data one might quickly leave the regions effected by flip-flopping





Feldman & Cousins "Unified Approach" looks like this:

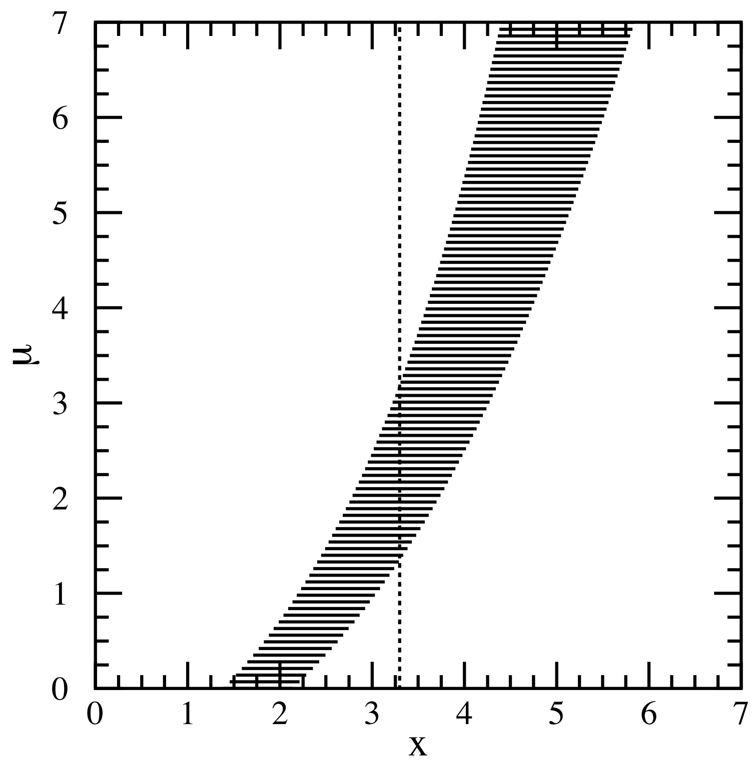
Neyman Construction

- For each μ: find region R_μ with probability $1-\alpha$
- Confidence Interval includes all μ consistent with observation at x_0

Ordering Rule specifies what region

F-C ordering rule is the Likelihood Ratio

$$R_\mu = \left\{ x \;\middle|\; \frac{L(x|\mu)}{L(x|\mu_{\rm best})} > k_\alpha \right\}$$



The F-C ordering rule follows naturally from Neyman-Pearson Lemma

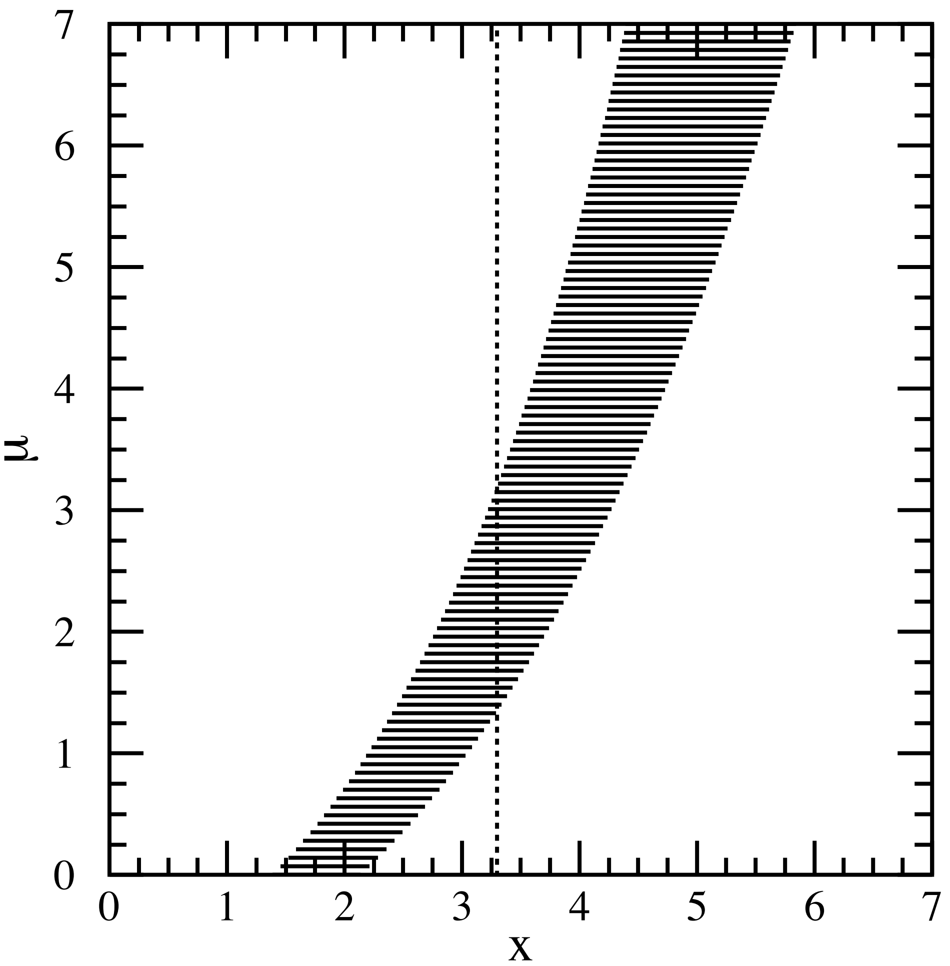
A different way to picture Feldman-Cousins

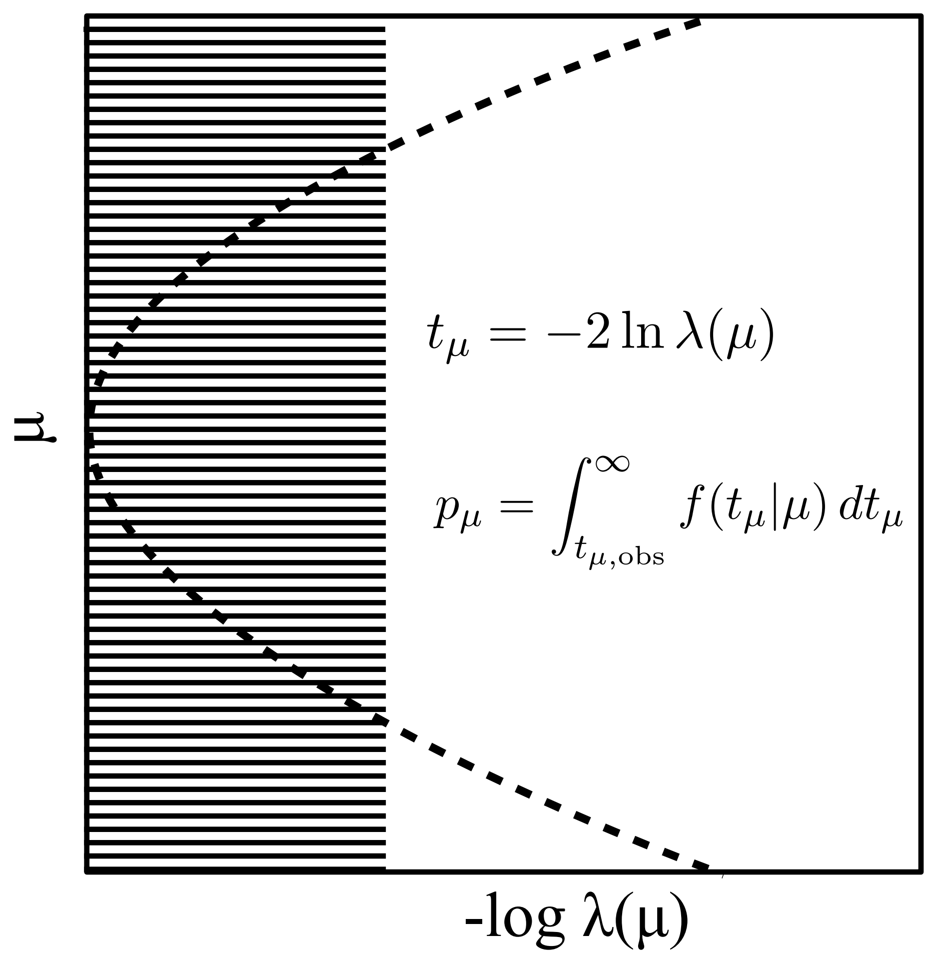
Most people think of plot on left when thinking of Feldman-Cousins

- bars are regions "ordered by" $R = P(n|\mu)/P(n|\mu_{\text{best}})$, with $\int_{x_1}^{x_2} P(x|\mu)dx = \alpha.$

But this picture doesn't generalize well to many measured quantities.

- Instead, just use R as the test statistic... and R is λ(μ)





Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman-Pearson)

Then we generalized it to composite hypotheses.

$$\frac{f(x|H_0)}{f(x|H_1)} \Rightarrow \frac{f(x|\theta_0)}{f(x|\theta_{best}(x))}$$

Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman-Pearson)

Then we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman-Pearson)

Then we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x\|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x\|\theta_{r0}, \hat{\hat{\theta}}_s)$

From Kendall

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman-Pearson)

Then we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x\|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x\|\theta_{r0}, \hat{\hat{\theta}}_s)$

$$(H_0 : \theta_r = \theta_{r0})$$
$$(H_1 : \theta_r \neq \theta_{r0})$$

From Kendall

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

Then we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x\|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x\|\theta_{r0}, \hat{\hat{\theta}}_s)$

$(H_0 : \theta_r = \theta_{r0})$
$(H_1 : \theta_r \neq \theta_{r0})$

Now consider the Likelihood Ratio

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

Intuitively l is a reasonable test statistic for H_0: it is the maximum likelihood under H_0 as a fraction of its largest possible value, and large values of l signify that H_0 is reasonably acceptable.

From Kendall

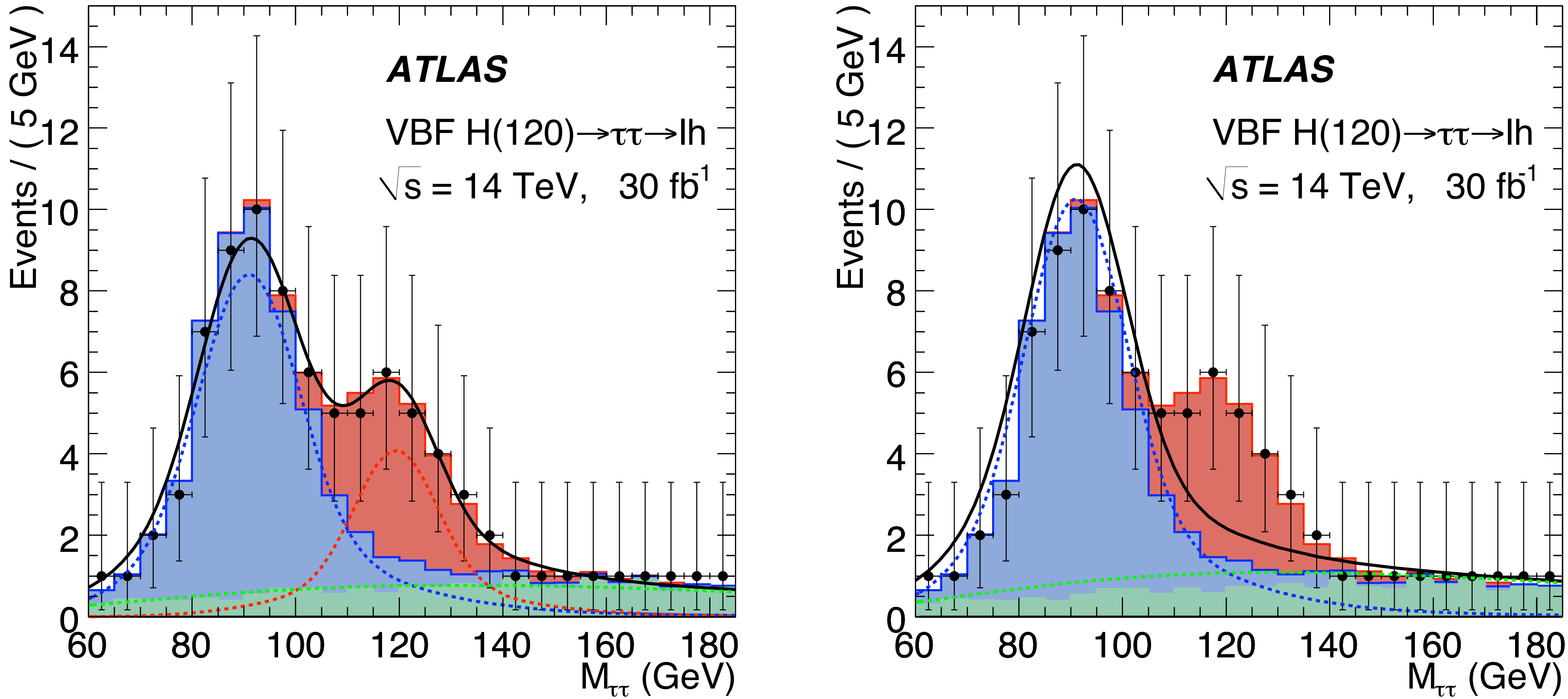
An example

Essentially, you need to fit your model to the data twice:

once with everything floating, and once with signal fixed to 0

$$\lambda(\mu=0) = \frac{L(data|\mu=0,\hat{\hat{b}}(\mu=0),\hat{\hat{\nu}}(\mu=0))}{L(data|\hat{\mu},\hat{b},\hat{\nu})},$$

$L(data|\hat{\mu},\hat{b},\hat{\nu})$ $\qquad$ $L(data|\mu=0,\hat{\hat{b}},\hat{\hat{\nu}})$

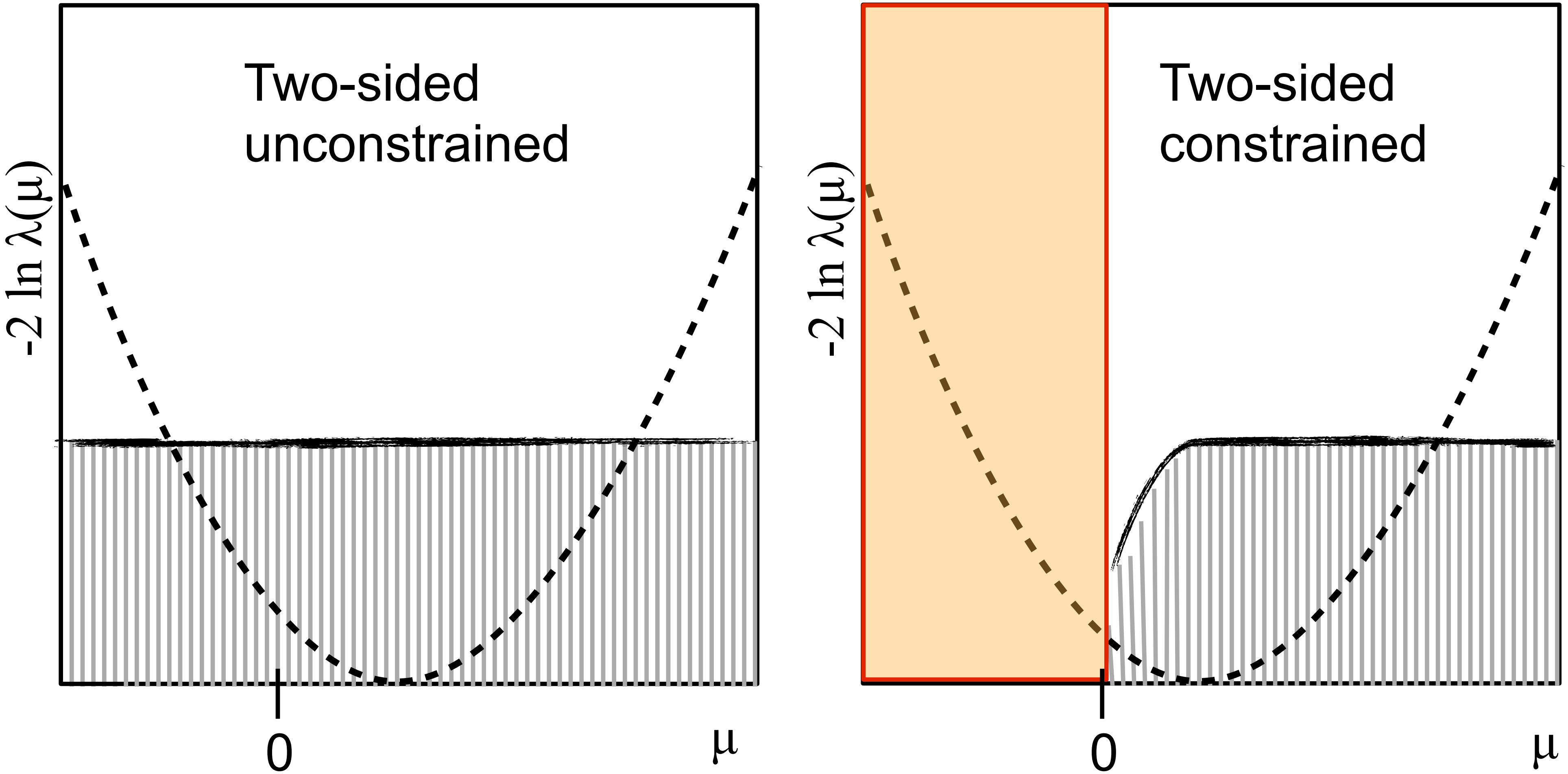


Feldman-Cousins with and without constraint

With a physical constraint (μ>0) the confidence band changes, but conceptually the same. Do not get empty intervals.

$$t_\mu = -2\ln\lambda(\mu)$$

$$\tilde{t}_\mu = -2\ln\tilde{\lambda}(\mu) = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\theta}}(0))} & \hat{\mu} < 0\,, \\ -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0\,. \end{cases}$$

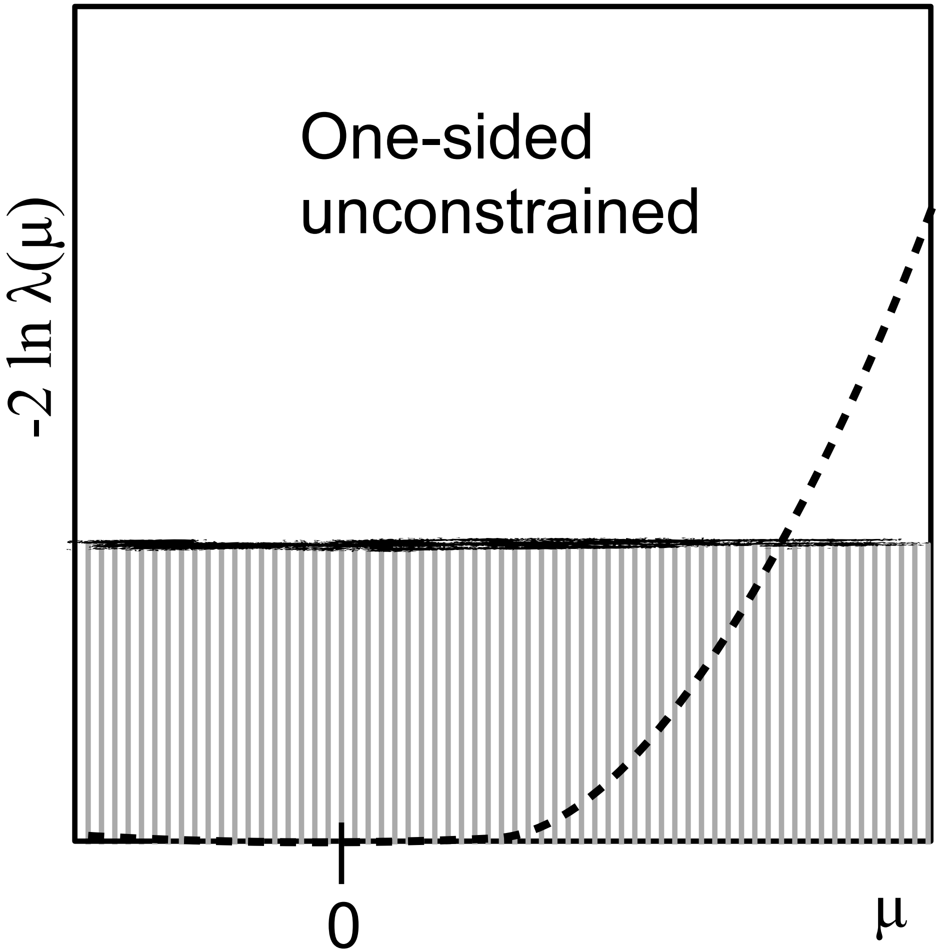


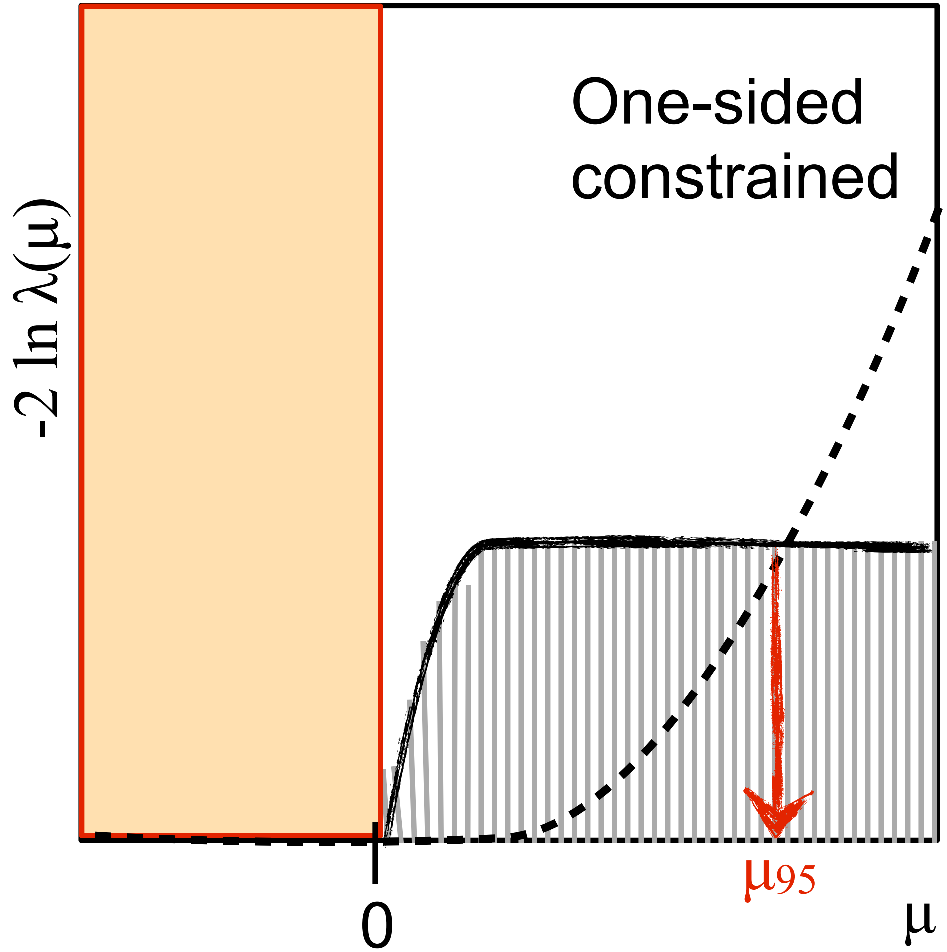
Modified test statistic for 1-sided upper limits

For 1-sided upper-limit one construct a test that is more powerful for all μ>0 (but has no power for μ=0) simply by discarding "upward fluctuations"

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu\ , \\ 0 & \hat{\mu} > \mu\ , \end{cases}$$

$$\tilde{q}_\mu = \begin{cases} -2\ln \frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0 \\ -2\ln \frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})} & 0 \le \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu\ . \end{cases}$$



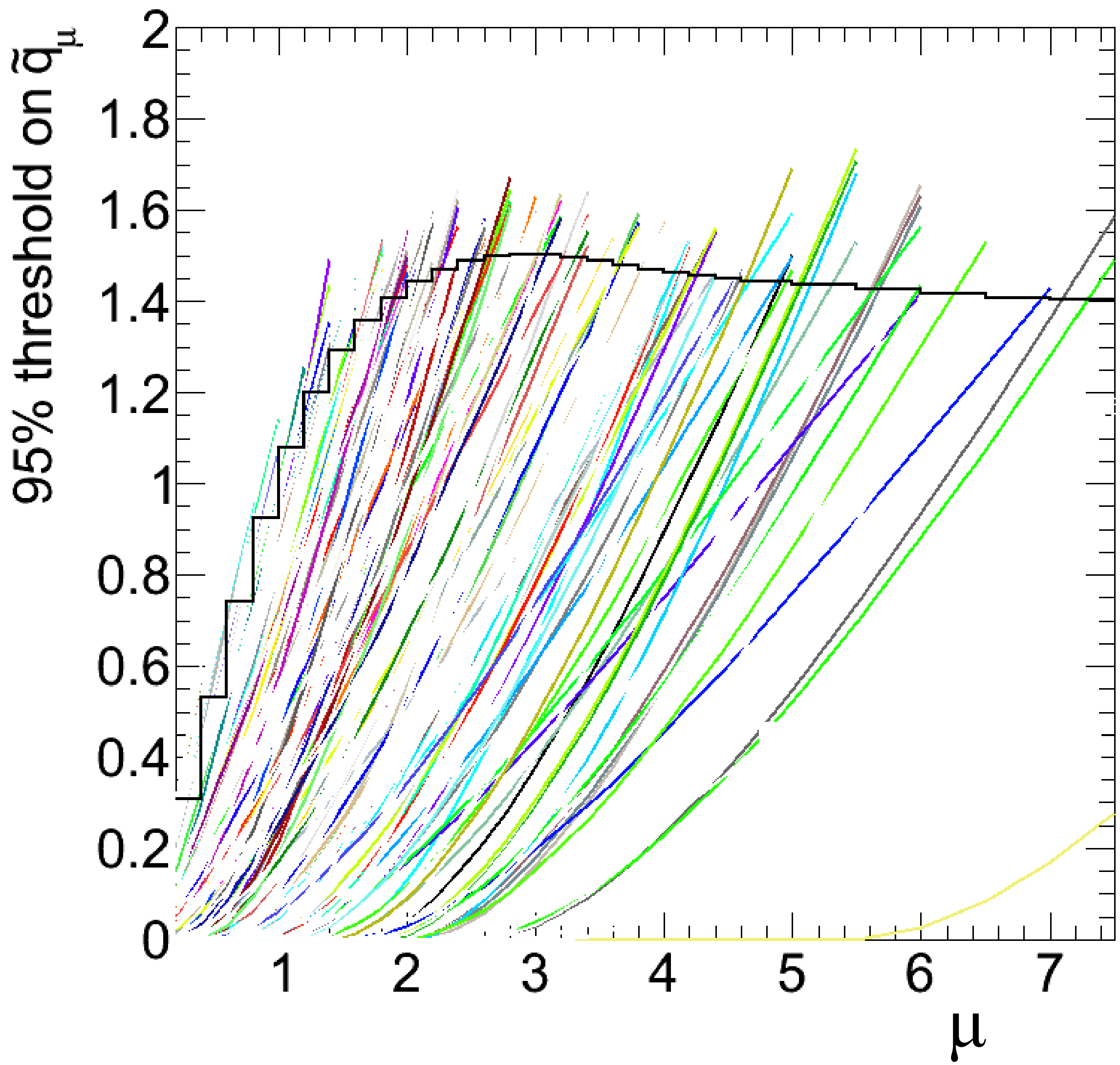


A real life example

Center for Cosmology and Particle Physics

Each colored curve is represents a single pseudo-experiment

- the test statistic is changing as μ, the parameter of interest, changes



Recall: Hybrid Solutions

Goal of Bayesian-frequentist hybrid solutions is to provide a frequentist treatment of the main measurement, while eliminating nuisance parameters (deal with systematics) with an intuitive Bayesian technique.

$$P(n_{\rm on}|s) = \int db\, {\rm Pois}(n_{\rm on}|s+b)\, \pi(b), \qquad p = \sum_{n=n_{obs}}^{\infty} P(n|s)$$

Tracing back the origin of $\pi(b)$

- clearly state prior $\eta(b)$; identify control samples (sidebands) and use:

$$\pi(b) = P(b|n_{\rm off}) = \frac{P(n_{\rm off}|b)\eta(b)}{\int db P(n_{\rm off}|b)\eta(b)}.$$

Note, if we do not want to use the Hybrid Bayesian-Frequentist approach for the nuisance parameters, then we **must consider both *n_{on}* and *n_{off}* when generating our toy Monte Carlo**

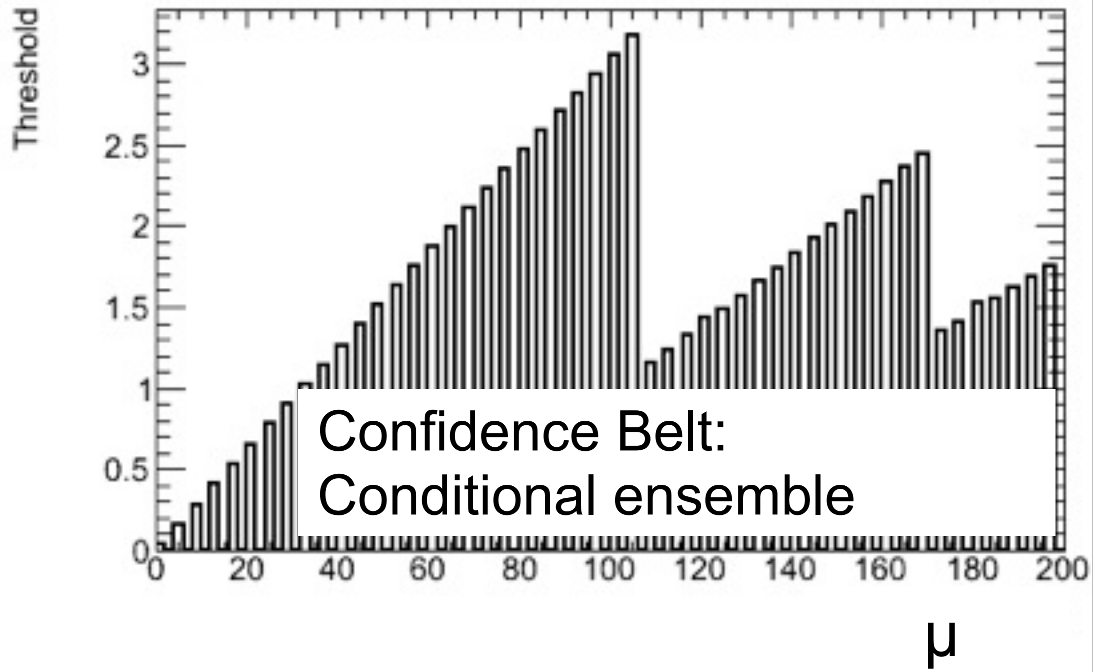
$$P(n_{\rm on}, n_{\rm off}|s,b) = {\rm Pois}(n_{\rm on}|s+b)\, {\rm Pois}(n_{\rm off}|\tau b).$$

Conditional vs. Unconditional Ensemble

Center for Cosmology and Particle Physics

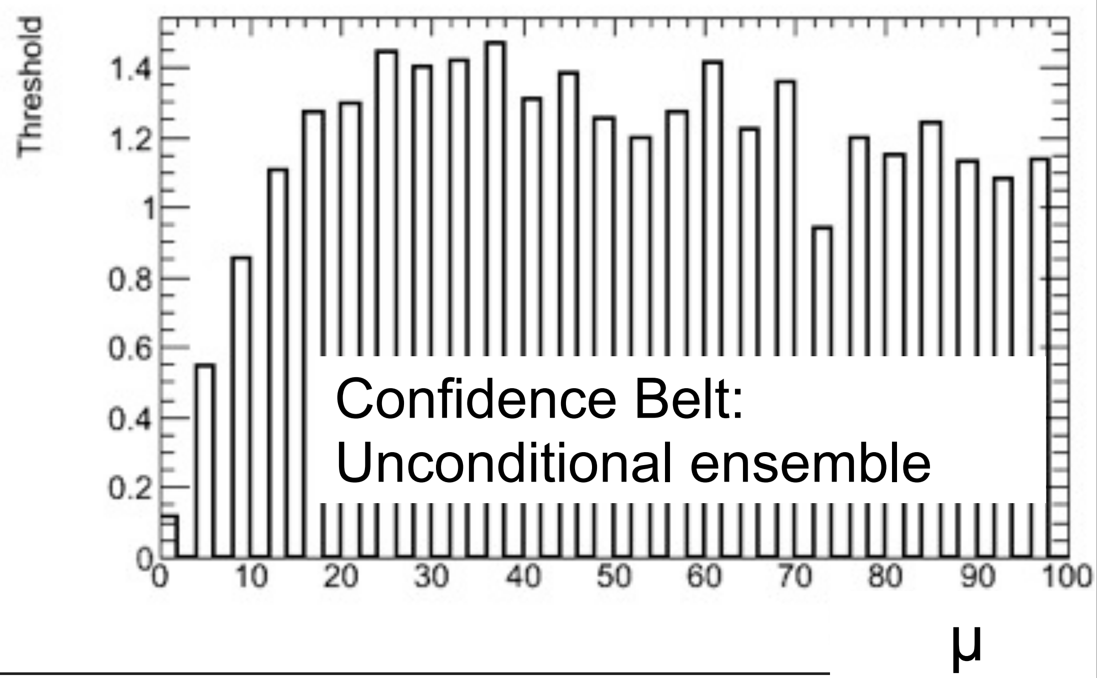
In the Conditional ensemble the global observables / auxiliary measurements are always the same

- if there are very few events expected, the test statistic takes on discrete values
- discreteness leads to over-coverage in some areas



In the Unconditional ensemble the global observables / auxiliary measurements fluctuate "smearing out" the value of the test statistic.

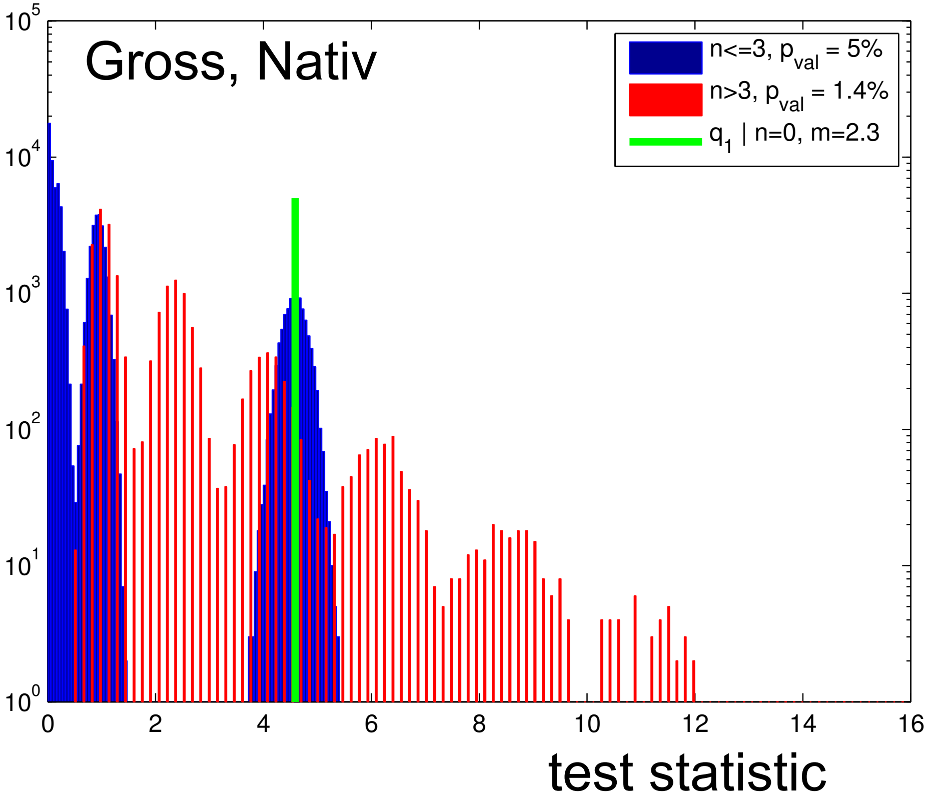
- also more fluctuations in results

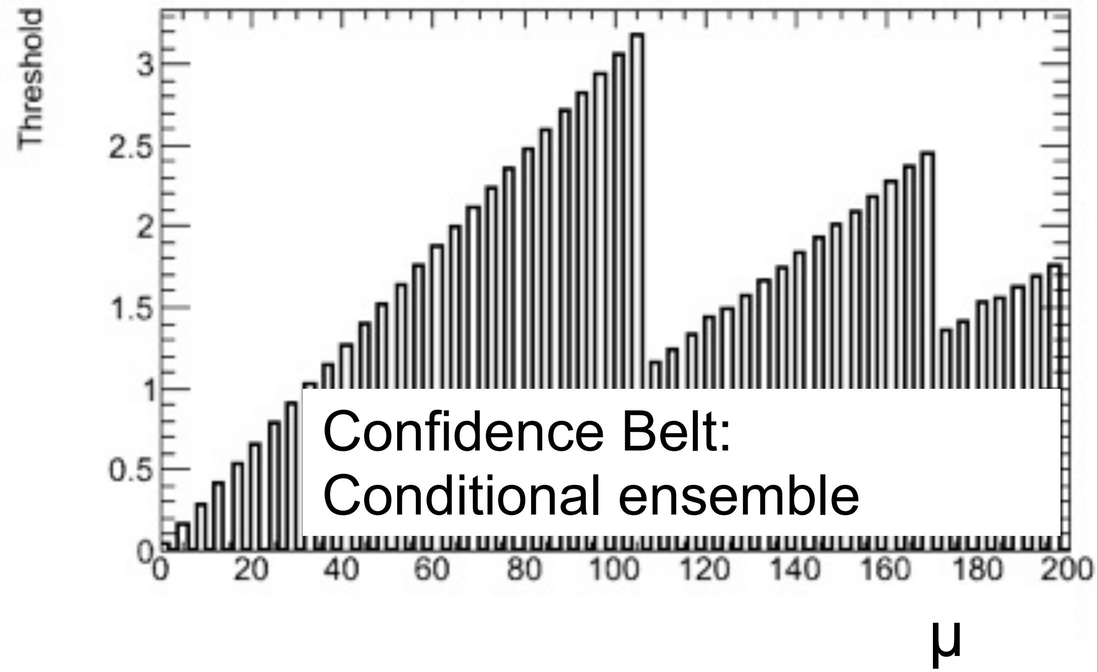


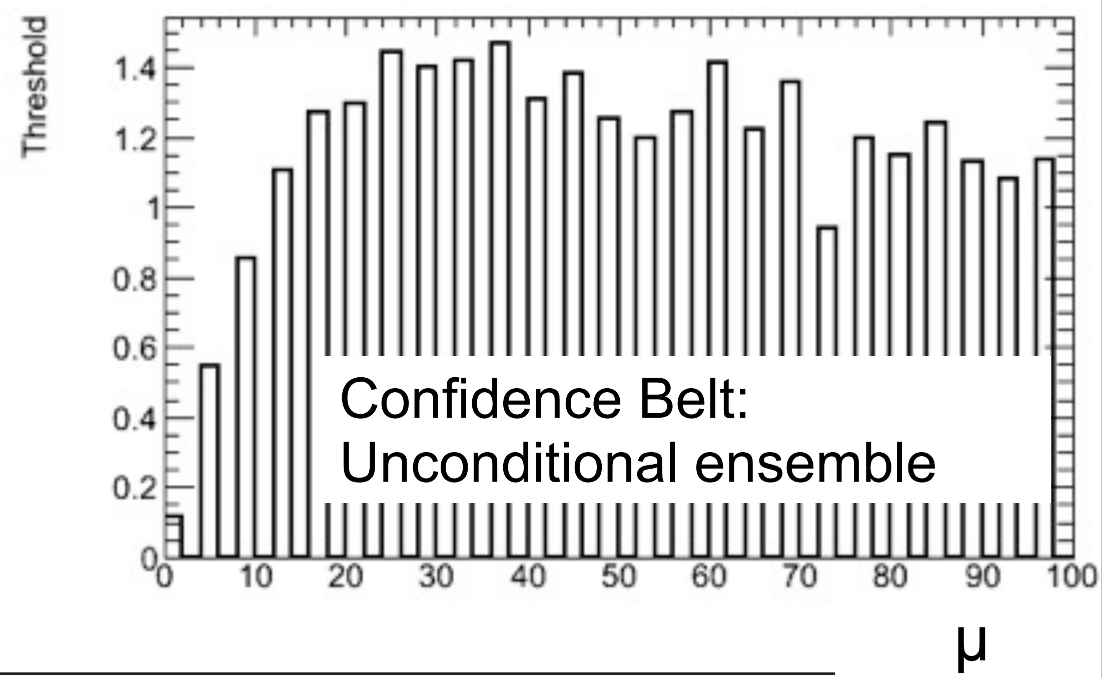
More on conditioning tomorrow!

Conditional vs. Unconditional Ensemble

Center for
Cosmology and
Particle Physics







In the Unconditional ensemble the global observables / auxiliary measurements fluctuate “smearing out” the value of the test statistic.

- also more fluctuations in results

More on conditioning tomorrow!

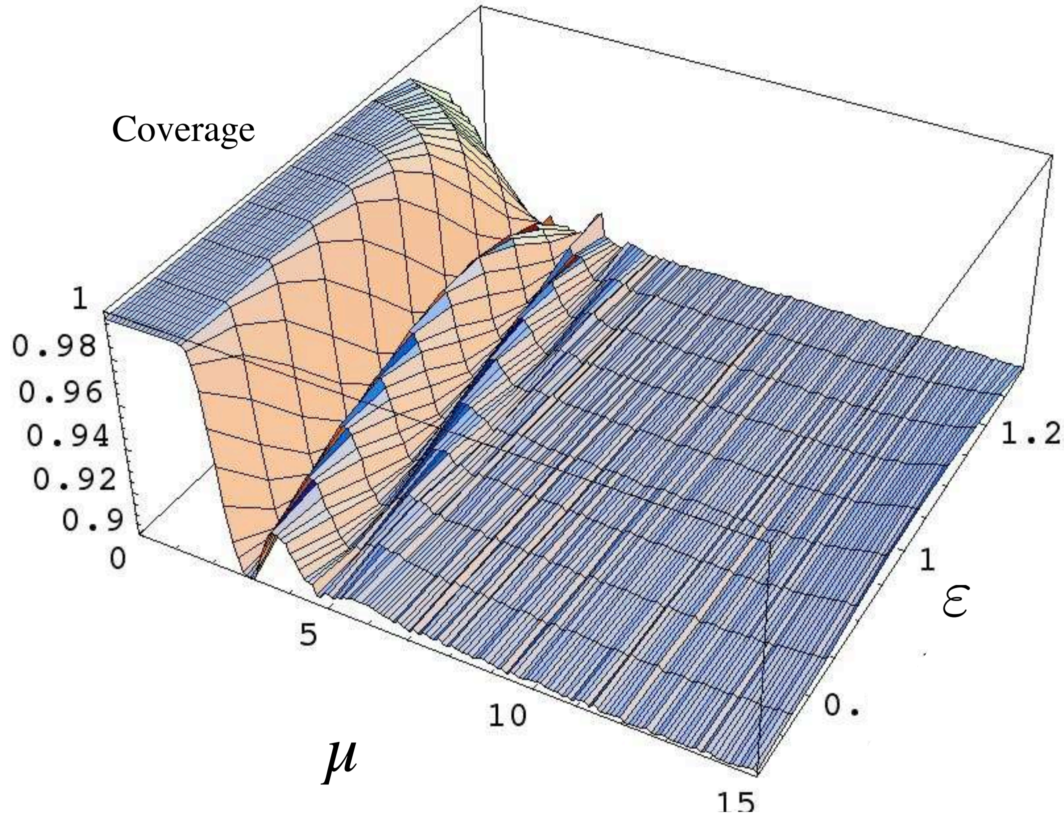
Coverage

Center for Cosmology and Particle Physics

Coverage can be different at each point in the parameter space

Example:

G. Punzi - PHYSTAT 05 - Oxford, UK



Poisson(+background), with a systematic uncertainty on efficiency:

$$x \sim Pois(\varepsilon\mu + b) \quad e \sim G(\varepsilon,\sigma)$$

e is a measurement of the unknown efficiency ε, with resolution σ

ε is the efficiency (a "normalization factor", can be larger than 1).

Neyman Construction with Nuisance parameters

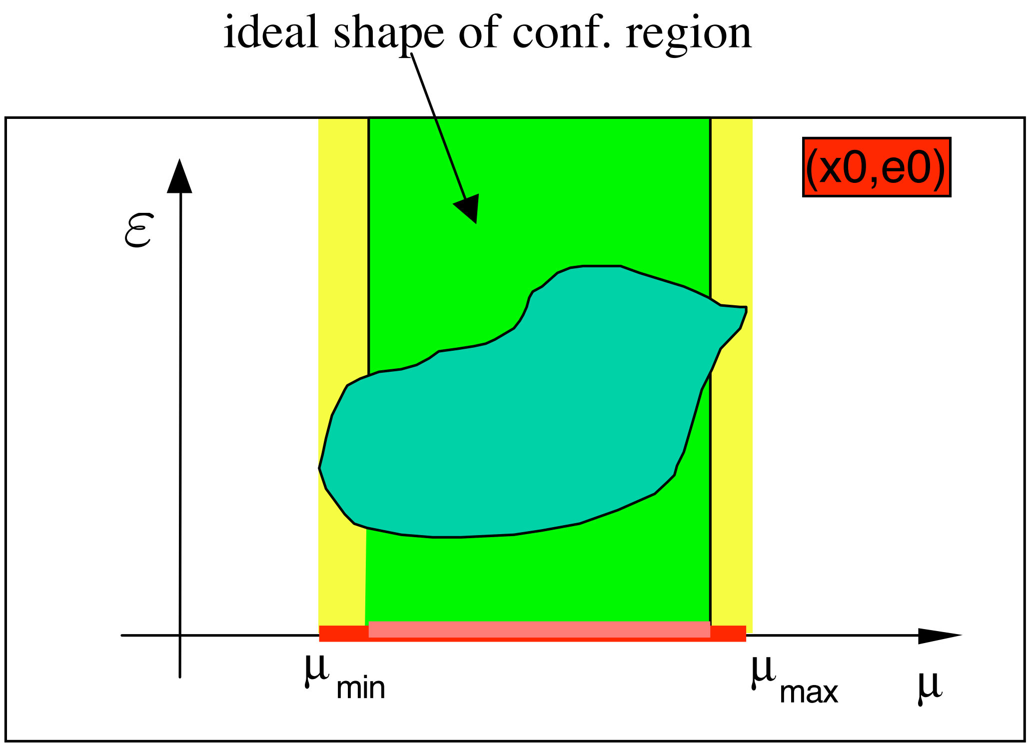
CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS

In the strict sense, one wants coverage for μ **for all** values of the nuisance parameters (here ε)

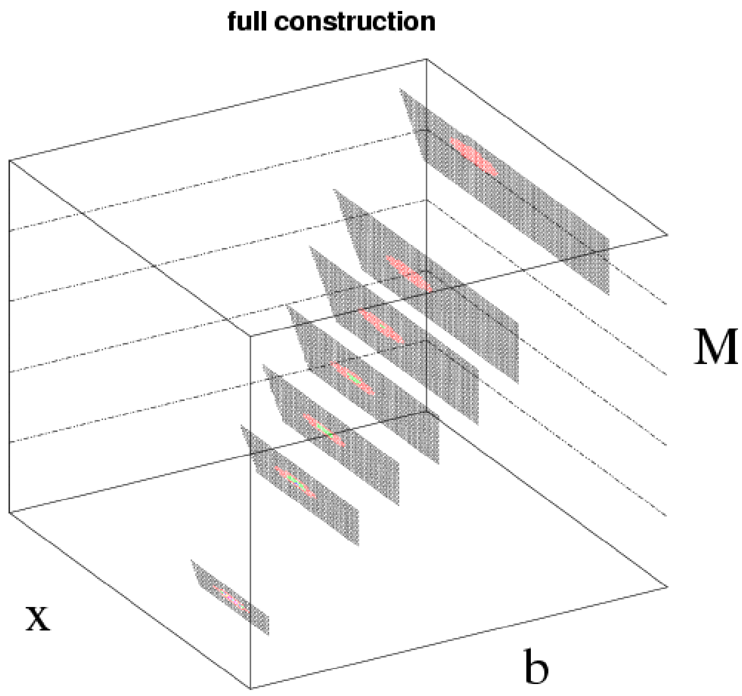
- The "full construction" one n

Challenge for full Neyman Construction is computational time (scan in 50-D isn't practical) and to avoid significant over-coverage

- note: projection of nuisance parameters is a union (eg. set theory) not an integration (Bayesian)



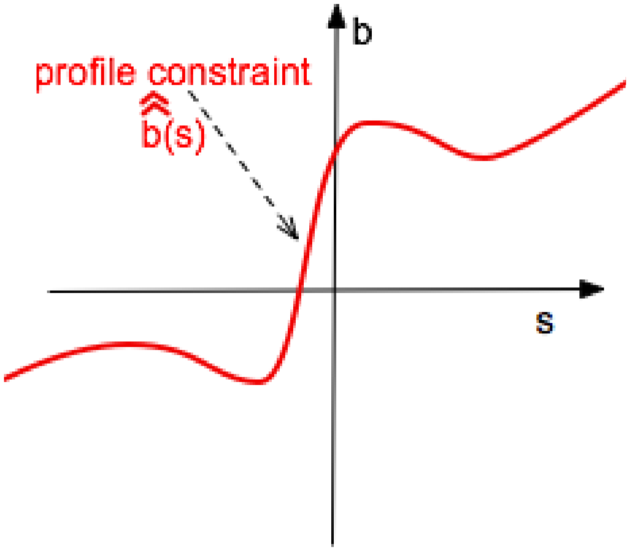
G. Punzi - PHYSTAT 05 - Oxford, UK

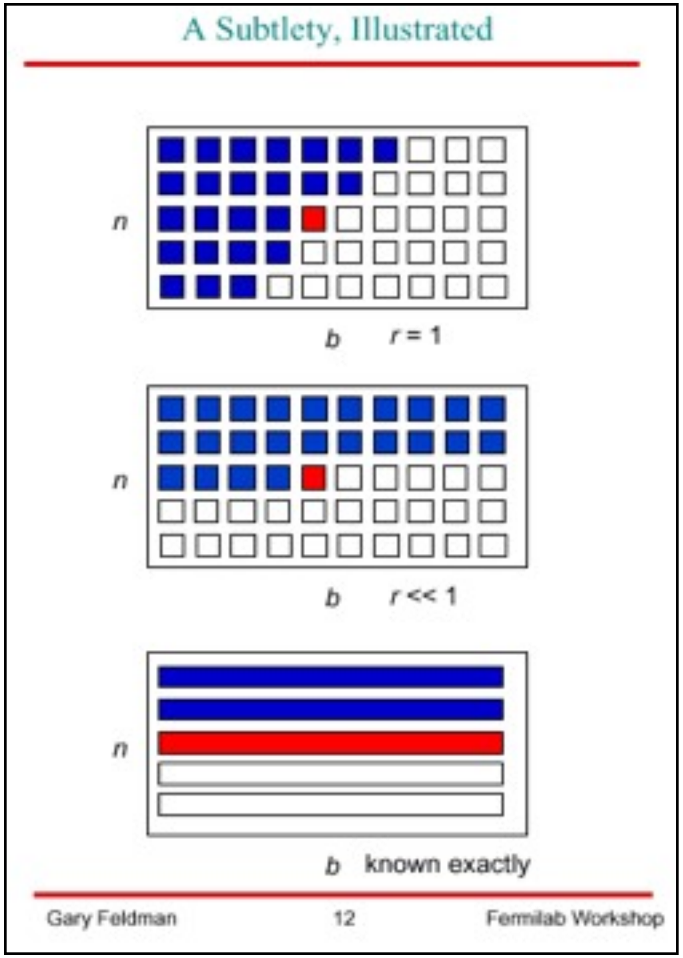


K. Cranmer - PHYSTAT 03 - SLAC

Profile Construction

Gary Feldman presented an approximate Neyman Construction, based on the profile likelihood ratio as an ordering rule, but only performing the construction on a subspace (eg. their conditional maximum likelihood estimate)





The **profile construction** means that one does not need to scan each nuisance parameter (keeps dimensionality constant)

- easier computationally

This approximation does not guarantee exact coverage, but

- tests indicate impressive performance
- one can expand about the profile construction to improve coverage, with the limiting case being the full construction

Profile Construction: professional literature

While I have been calling it the “profile construction”, it has been called a “hybrid resampling” technique by professional statisticians

- Note: ‘hybrid’ here has nothing to do with Bayesian-Frequentist Hybrid, but a connection to “boot-strapping”

Resampling methods for confidence intervals in group sequential trials

By CHIN-SHAN CHUANG

Department of Statistics, University of Wisconsin at Madison, Madison, Wisconsin 53706, U.S.A.

cchuang@stat.wisc.edu

AND TZE LEUNG LAI

Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.

lait@leland.stanford.edu

Chuang, C. and Lai, T. L. (1998). Resampling methods for confidence intervals in group sequential trials. *Biometrika* **85**, 317-332.

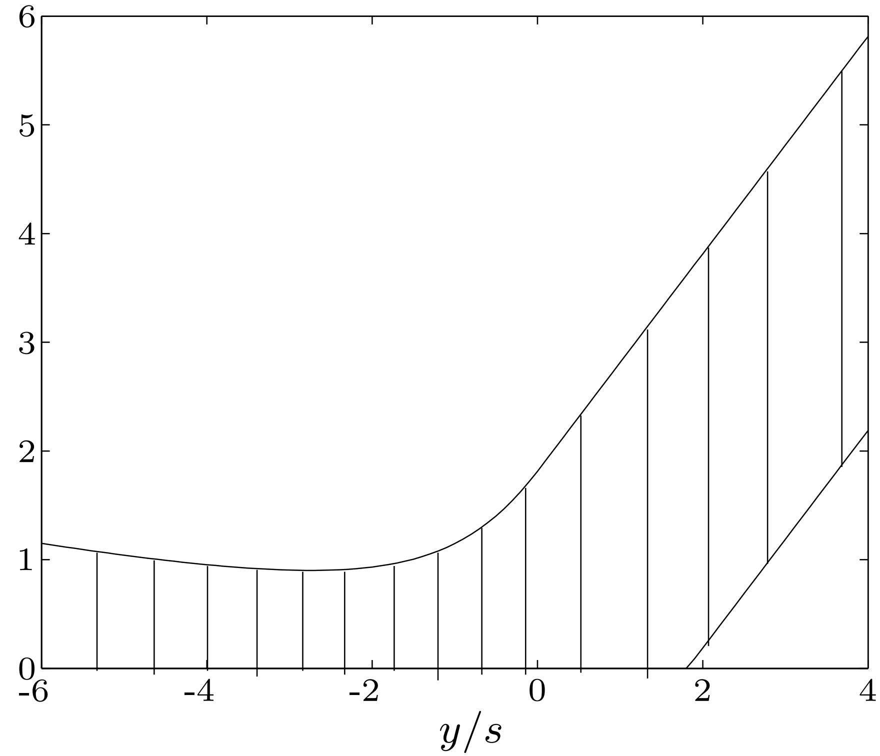
Chuang, C. and Lai, T. L. (2000). Hybrid resampling methods for confidence intervals. *Statist. Sinica* **10**, 1-50.

Statistica Sinica **19** (2009), 301-314

ON THE UNIFIED METHOD WITH NUISANCE PARAMETERS

Bodhisattva Sen, Matthew Walker and Michael Woodroofe

The University of Michigan



Previous ways of addressing spurious exclusion

The problem of excluding parameter values to which one has no sensitivity known for a long time; see e.g.,

Virgil L. Highland, *Estimation of Upper Limits from Experimental Data*, July 1986, Revised February 1987, Temple University Report C00-3539-38.

In the 1990s this was re-examined for the LEP Higgs search by Alex Read and others

T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

and led to the "CL_s" procedure.

Lecture 4

What drives the choice of statistical method?

Do we insist on addressing Prob(theory | data)?

- if yes, then some form of Bayesian (requires priors)

Do we want to be able to incorporate subjective information in our inference?

- if yes, then subjective Bayesian

Do we insist on Coverage OR the Likelihood Principle? (Can't have both)

- If we insist on Coverage, then must use Frequentist
- If we insist on Likelihood Principle, two options:
  - Likelihood-based inference (no prior, approximate coverage, MINOS)
  - Bayesian (need prior, can be objective, can try for approximate coverage)

Do we want to provide the most information or go straight to inference?

- If we do, then we should publish probability model / likelihood function
  - Allows for all types of statistical analysis. Avoids the comparison problem.

What do we want to conclude?

- is a signal present?
- what production rate and model parameters of the new signal are still allowed?
- what is the best estimate and allowed range of rate and model parameters?

Asymptotic Properties of likelihood based tests

&

Likelihood-based methods

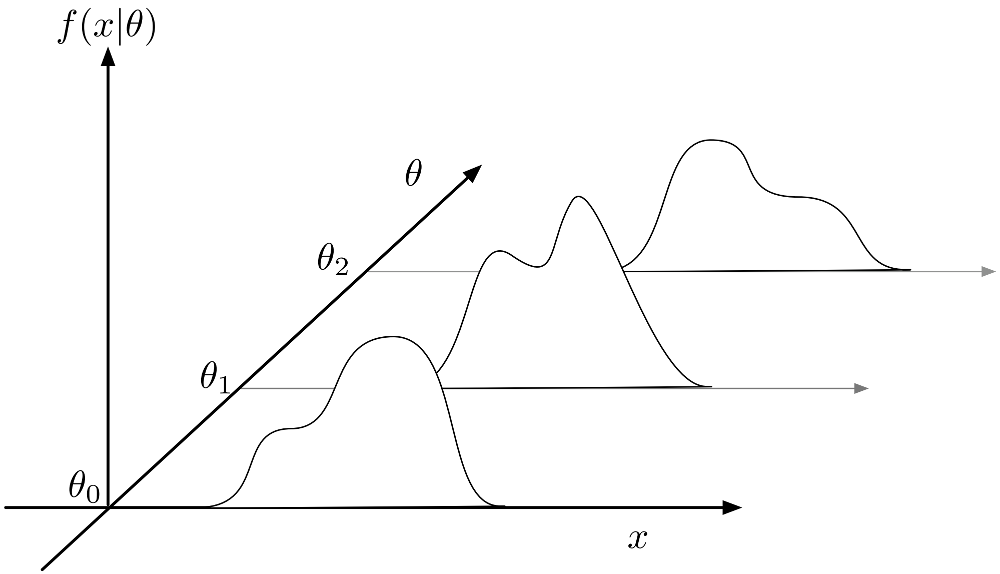
Wilks's theorem

Wilks's theorem says that asymptotically the distribution of

$$-2\log\lambda(\theta_0) = -2\log\frac{f(x|\theta_0)}{f(x|\hat{\theta}(x))}$$

when θ_0 is true approaches a chi-square distribution, with the number of degrees of freedom equal to the number of parameters of interest

$$-2\log\lambda(\theta) \sim \chi^2_n$$



It does not assume that the pdf is Gaussian!

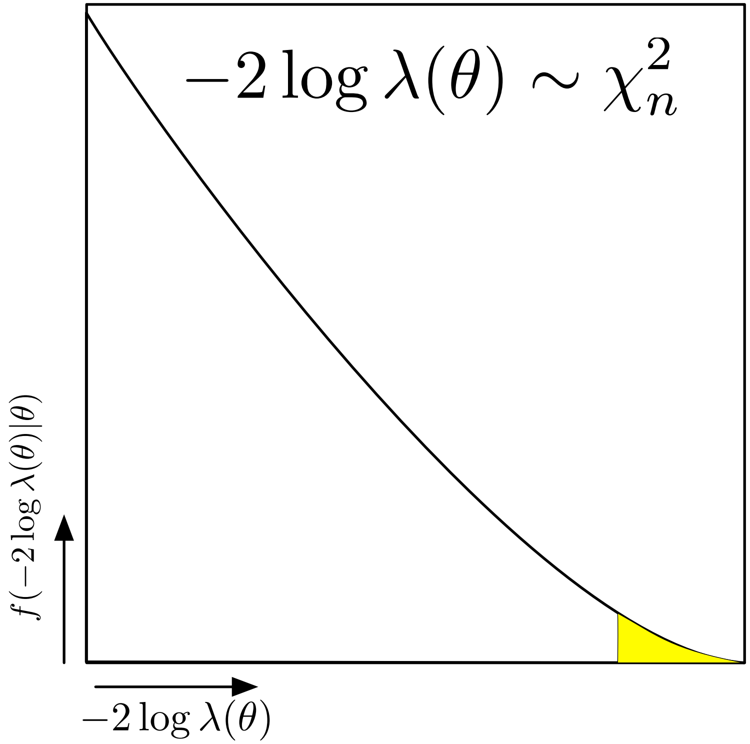
It is true for every value of θ

eg. "distribution free"

Likelihood-based Intervals

Wilks's theorem tells us how the profile likelihood ratio evaluated at θ is "asymptotically" distributed **when θ is true**

- asymptotically means there is sufficient data that the log-likelihood function is parabolic
- does NOT require the model **f(x|θ)** to be Gaussian
- there are some conditions that must be met for this to true



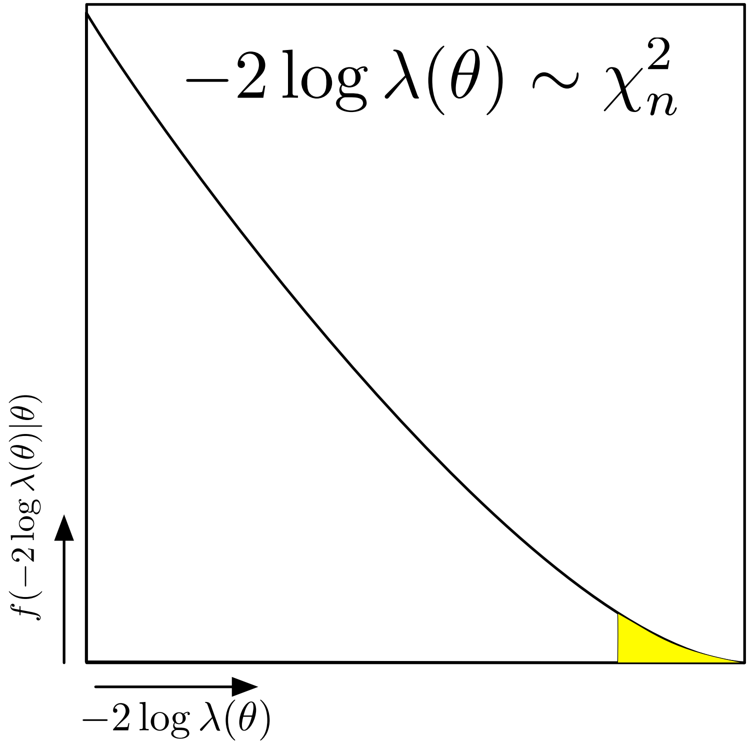
Note common exceptions:

- a parameter has no effect on the likelihood (eg. m_H when testing s=0) related to look-elsewhere effect
- require s≥0, but this just leads to a δ-function at 0 + ½χ^2

Likelihood-based Intervals

Wilks's theorem tells us how the profile likelihood ratio evaluated at θ is "asymptotically" distributed **when θ is true**

- asymptotically means there is sufficient data that the log-likelihood function is parabolic
- does NOT require the model **f(x|θ)** to be Gaussian
- there are some conditions that must be met for this to true



Note common exceptions:

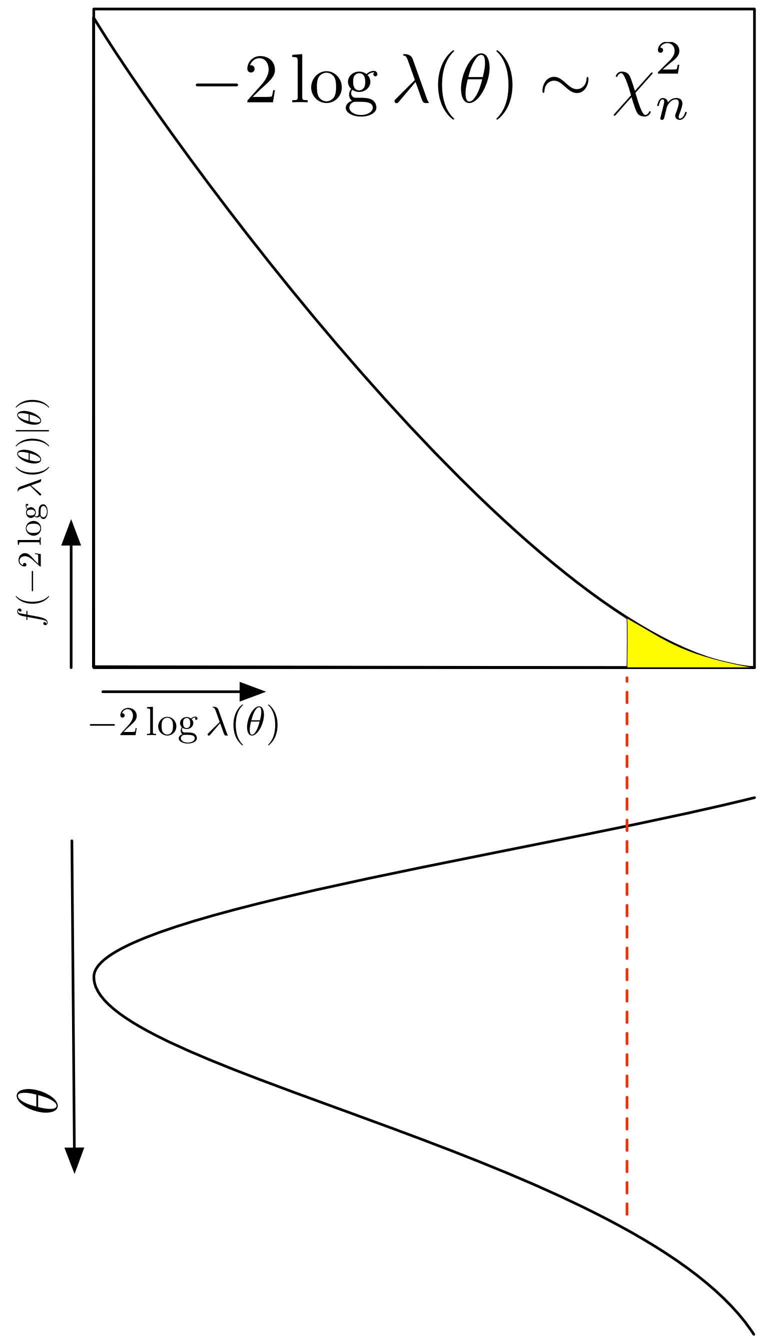
- a parameter has no effect on the likelihood (eg. m_H when testing s=0) related to look-elsewhere effect
- require s≥0, but this just leads to a δ-function at 0 + ½χ^2

Trial factors or the look elsewhere effect in high energy physics.
Eilam Gross, Ofer Vitells

Eur.Phys.J. C70 (2010) 525-530
e-Print: **arXiv:1005.1891 [physics.data-an]**

Wilks's theorem tells us how the profile likelihood ratio evaluated at θ is "asymptotically" distributed **when θ is true**

- asymptotically means there is sufficient data that the log-likelihood function is parabolic
- does NOT require the model **f(x|θ)** to be Gaussian



So we don't really need to go to the trouble to build its distribution by using Toy Monte Carlo or fancy tricks with Fourier Transforms

We can go immediately to the threhsold value of the profile likelihood ratio

Demonstration of Central Limit Theorem

The basic reason it works is due to an **asymptotic** limit

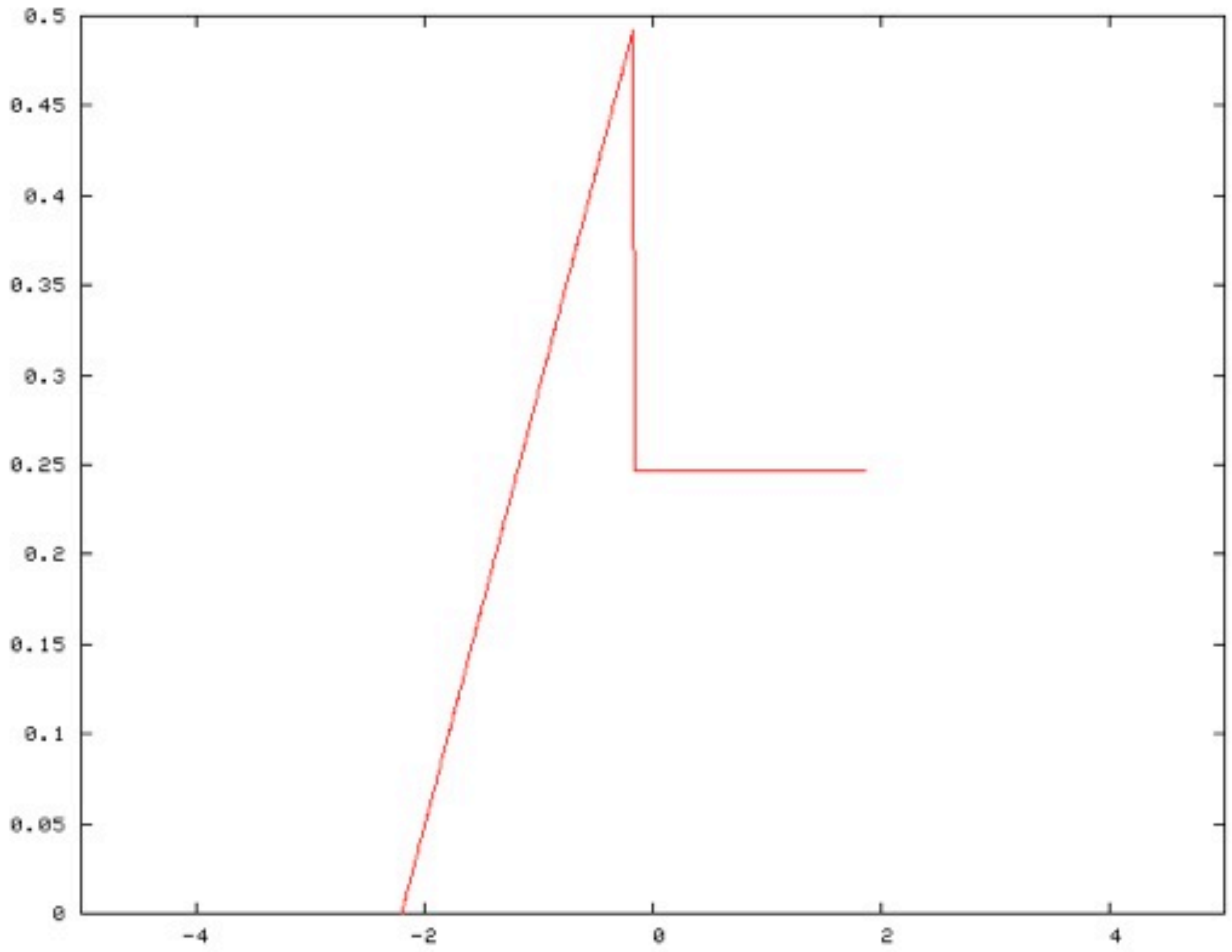
- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood

Demonstration of Central Limit Theorem

Center for Cosmology and Particle Physics

The basic reason it works is due to an **asymptotic** limit

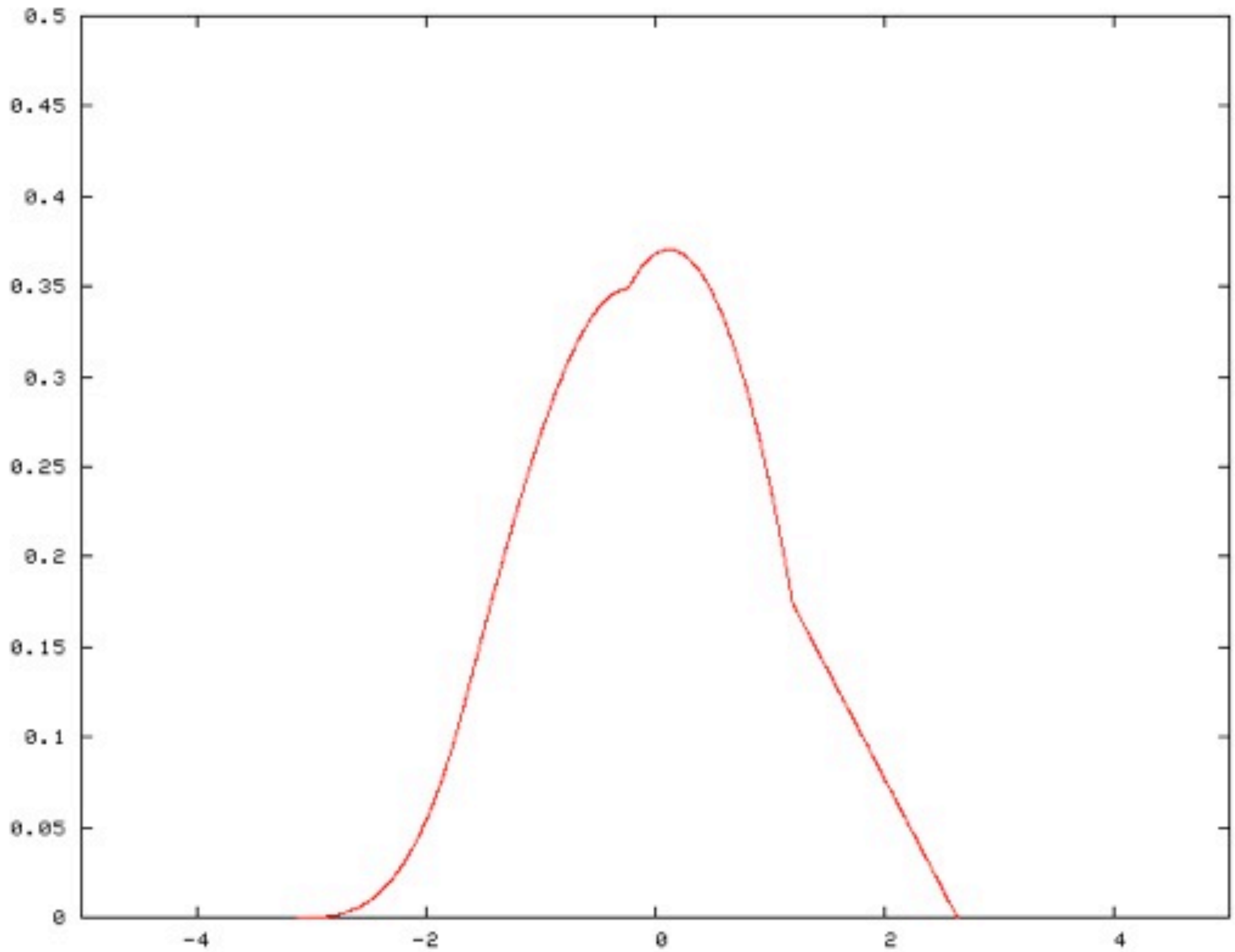
- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood



Demonstration of Central Limit Theorem

The basic reason it works is due to an **asymptotic** limit

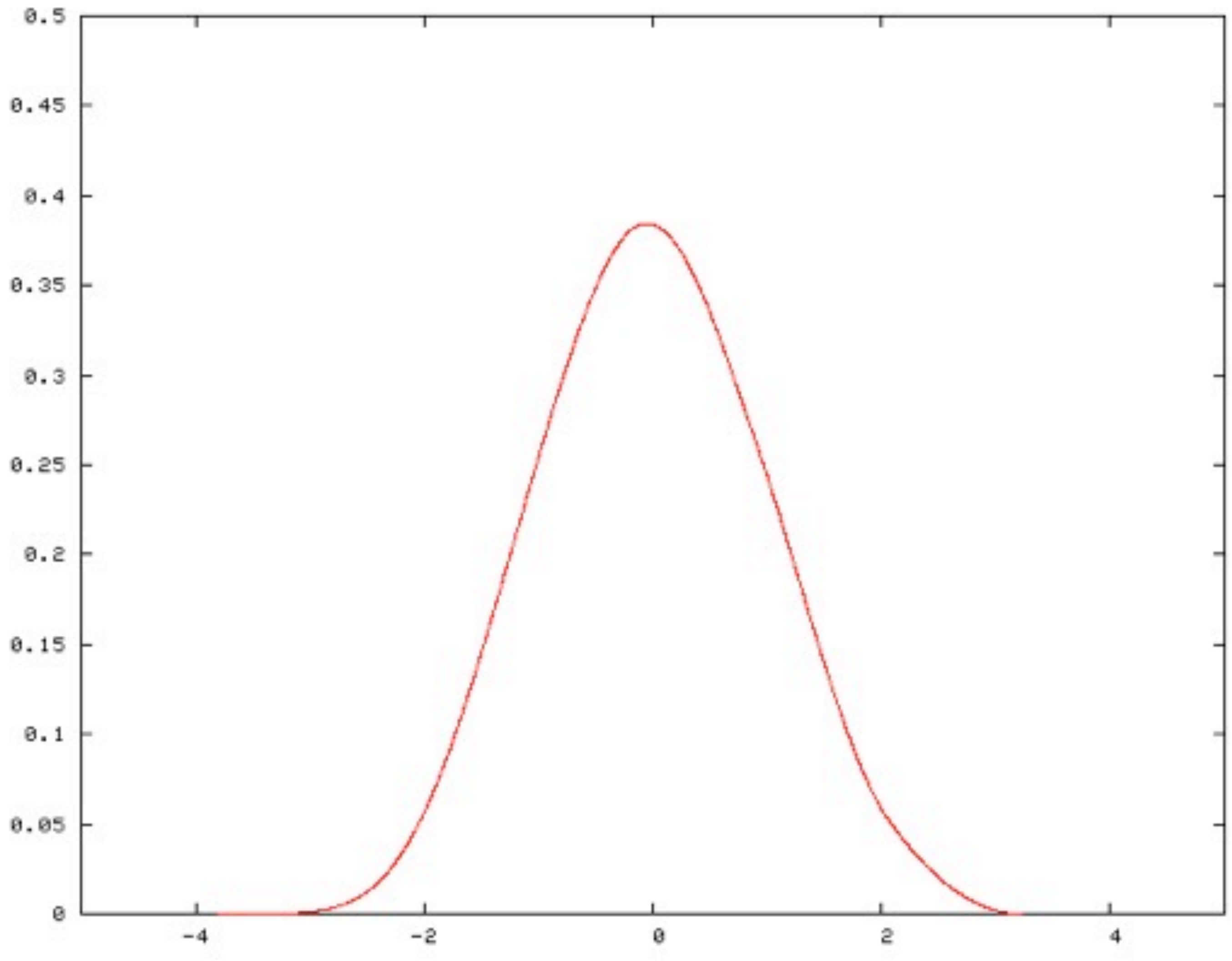
- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood



Demonstration of Central Limit Theorem

The basic reason it works is due to an **asymptotic** limit

- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood

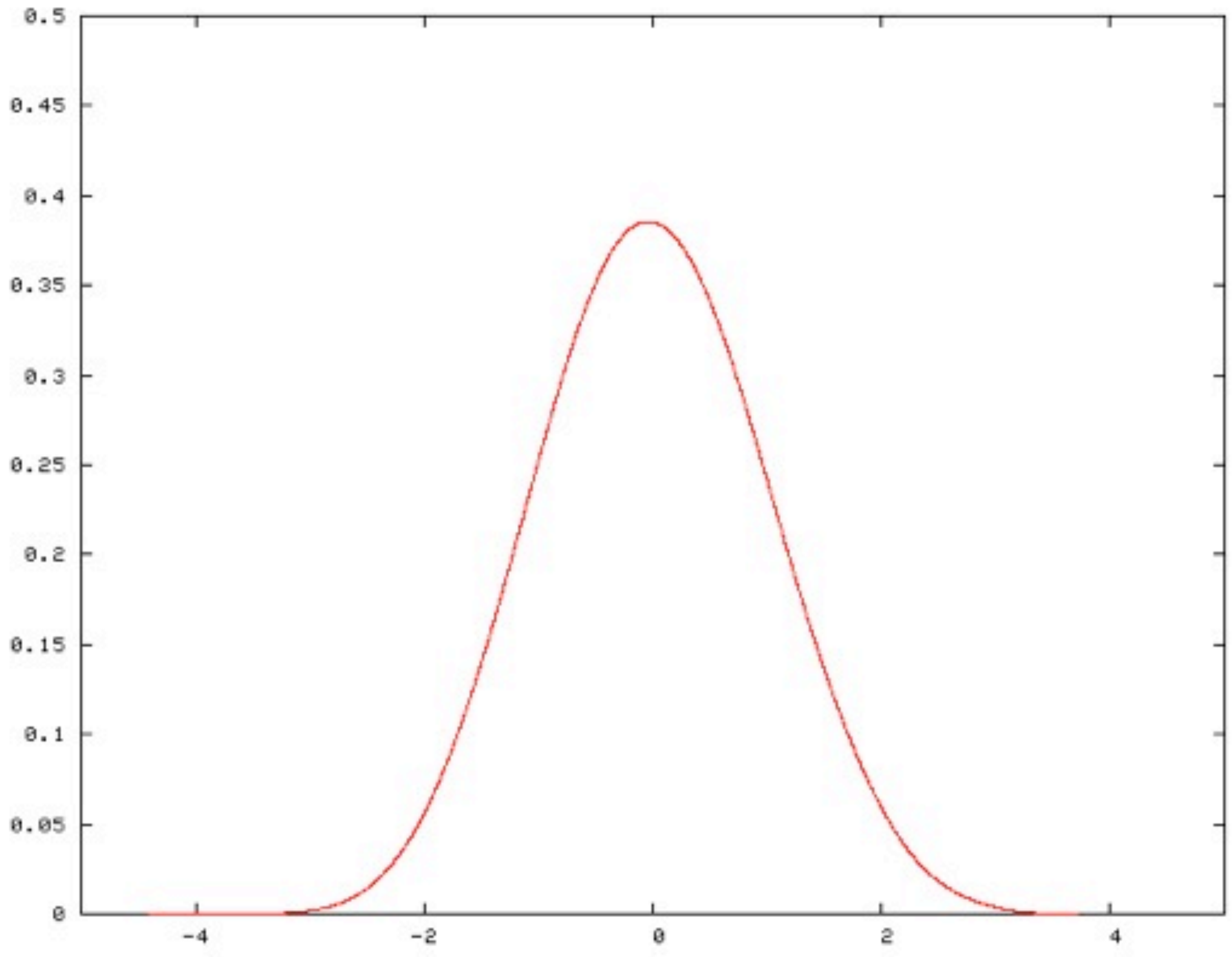


Demonstration of Central Limit Theorem

Center for Cosmology and Particle Physics

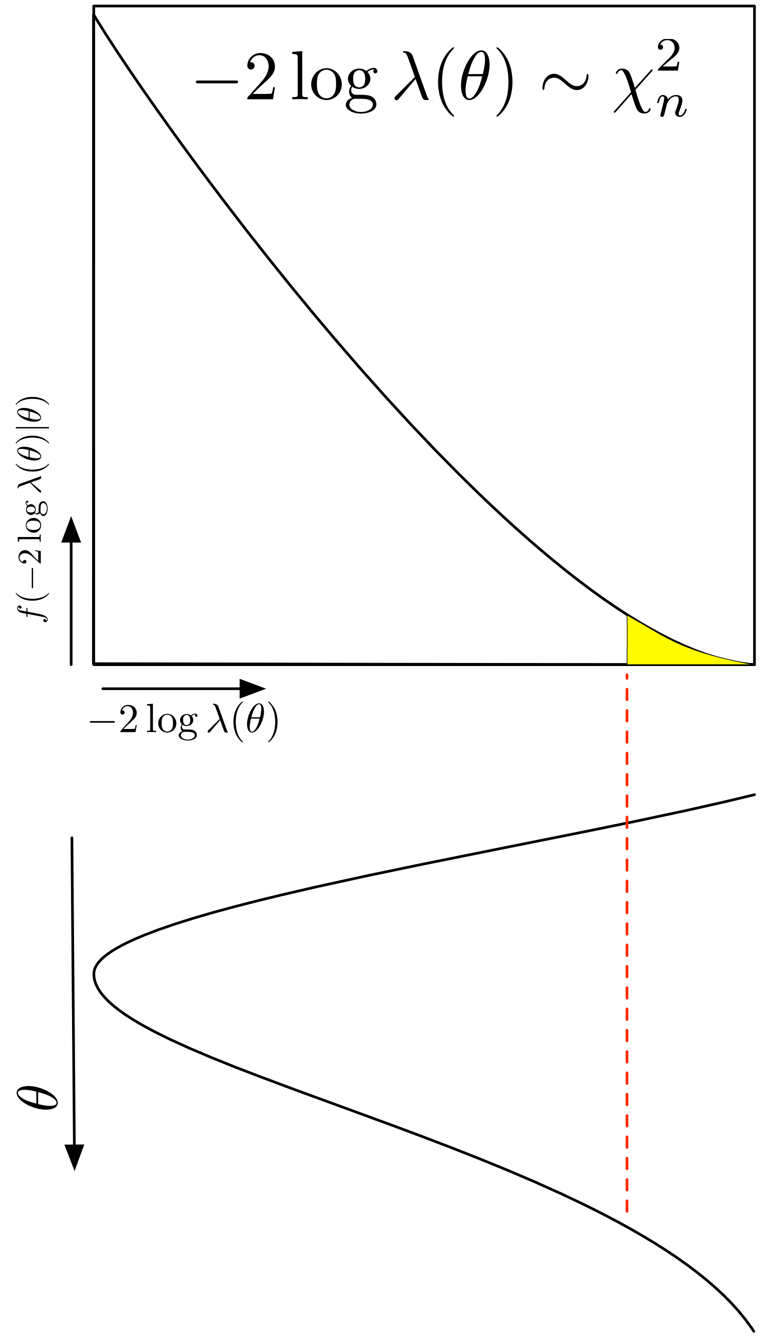
The basic reason it works is due to an **asymptotic** limit

- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood



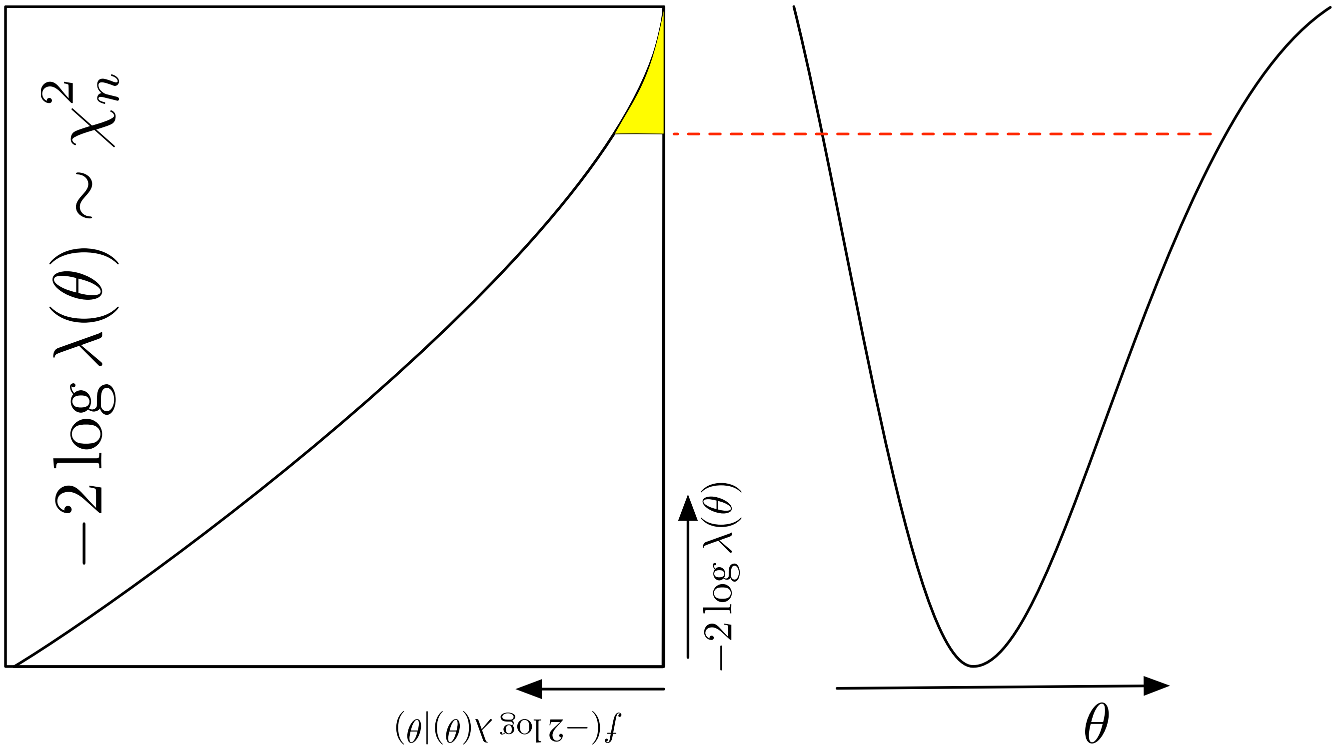
Likelihood-based Intervals

Center for Cosmology and Particle Physics



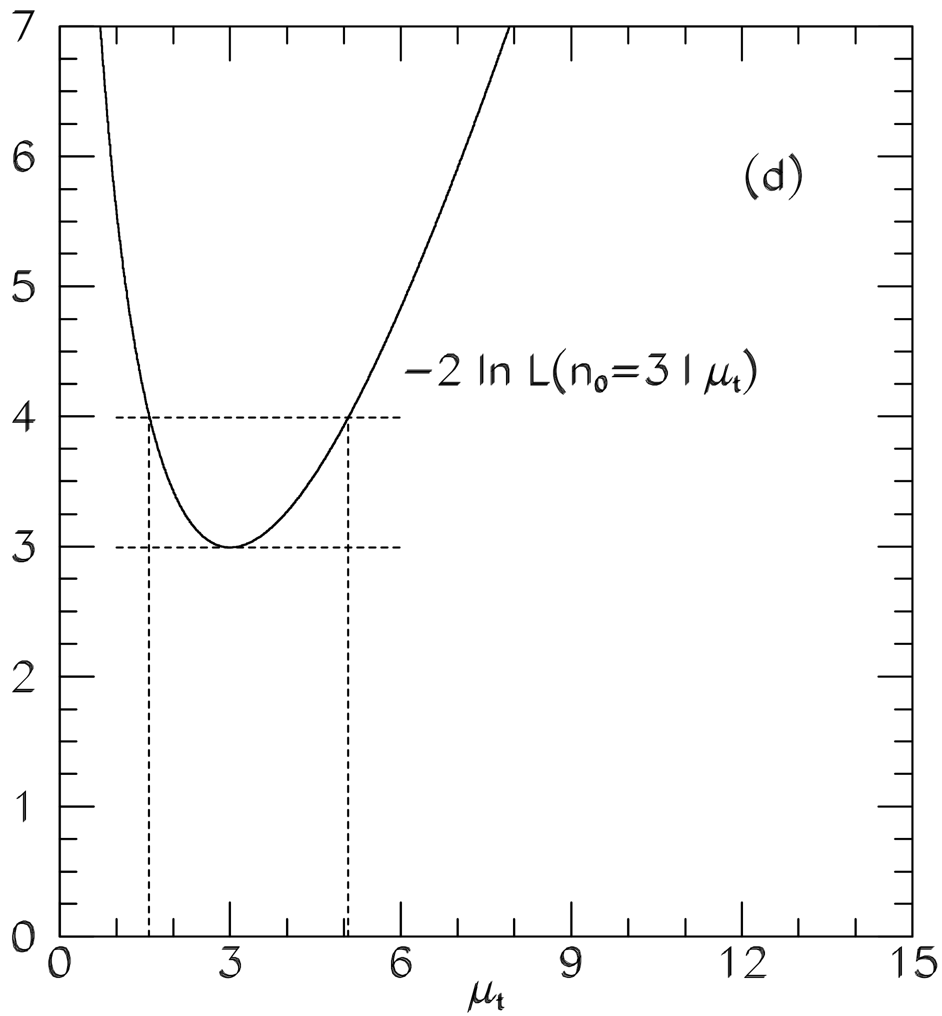
Likelihood-based Intervals

Center for Cosmology and Particle Physics



And typically we only show the likelihood curve and don't even bother with the implicit (asymptotic) distribution

Likelihood-based Intervals



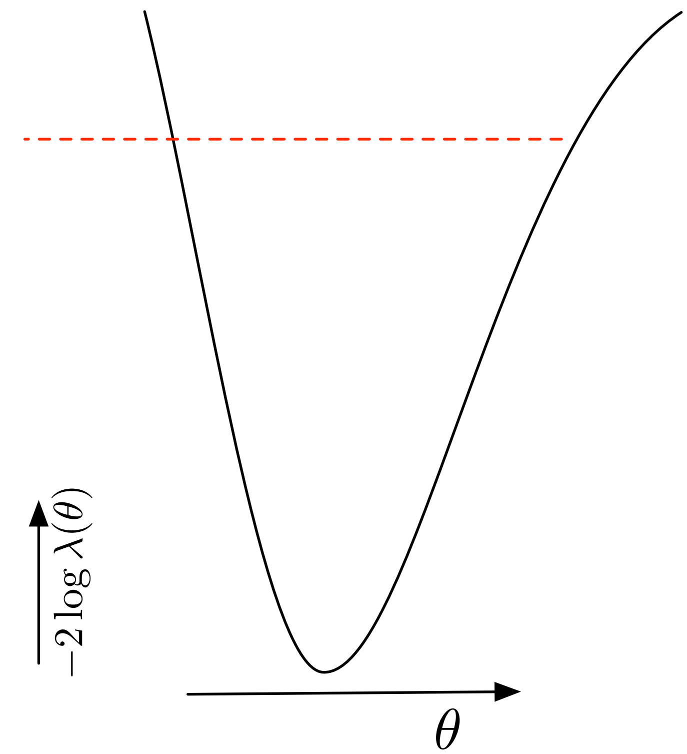


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

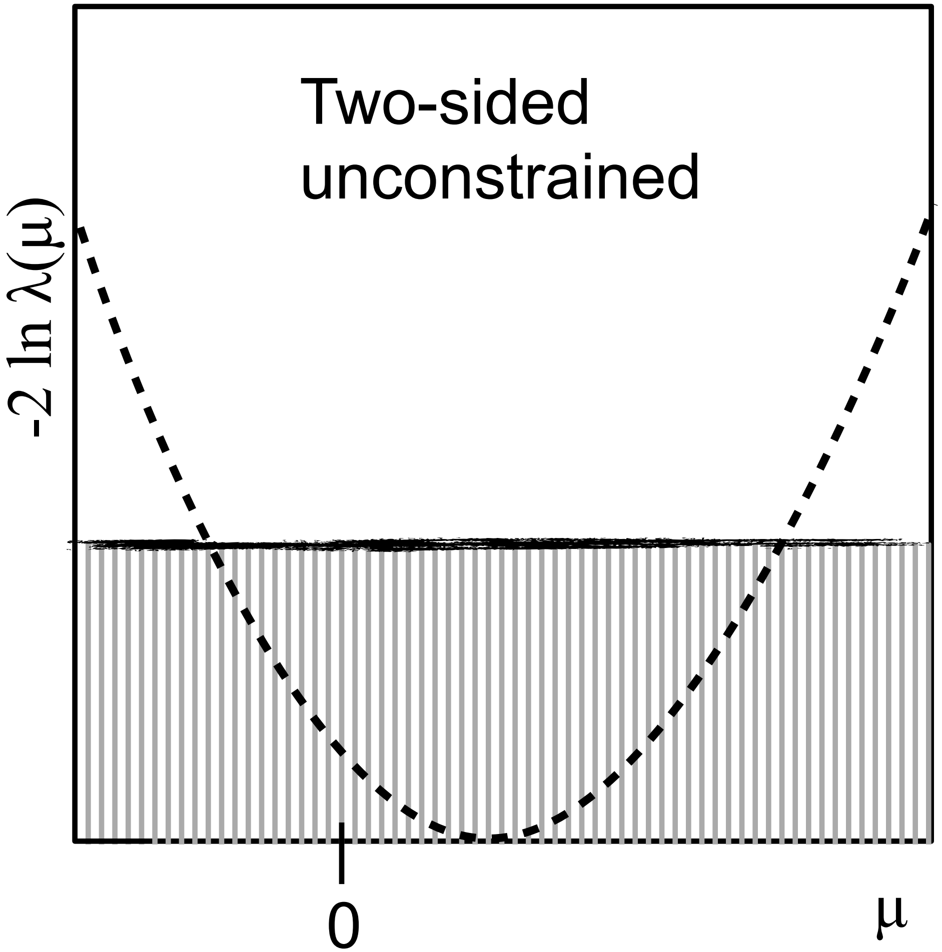
And typically we only show the likelihood curve and don't even bother with the implicit (asymptotic) distribution

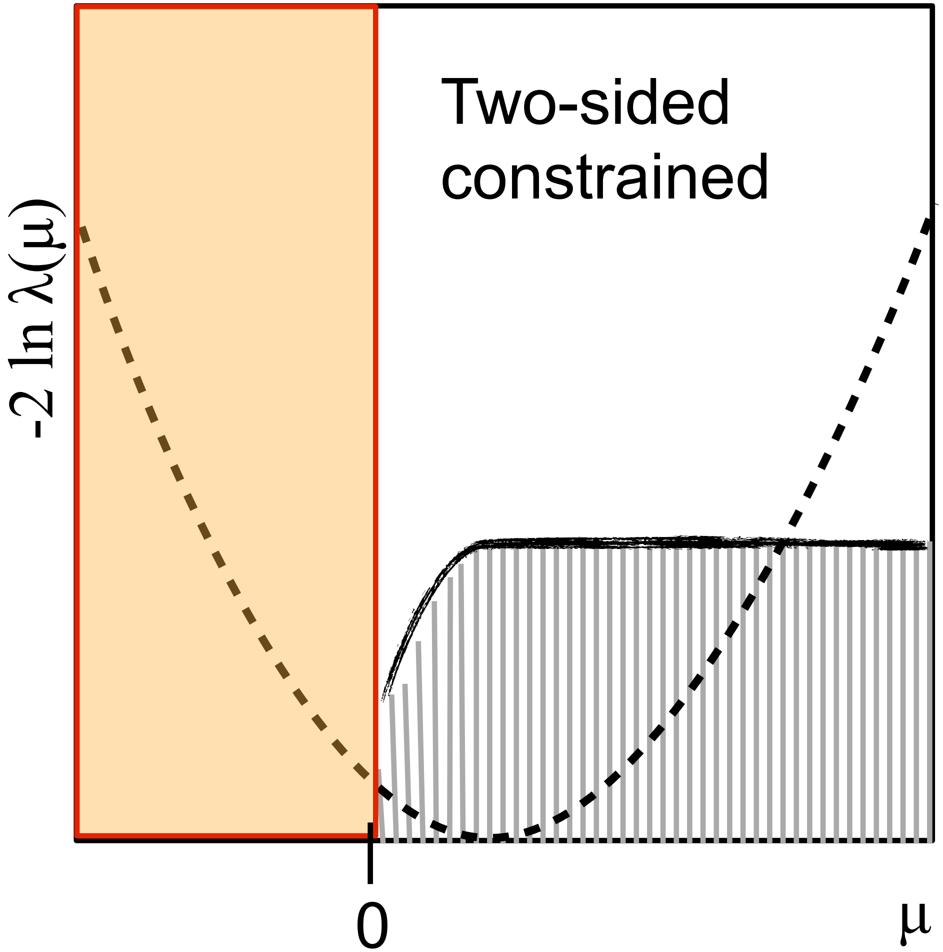
Wilks's theorem gives a short-cut for the Monte Carlo procedure used to find threshold on test statistic ⇒ MINOS is asymptotic approximation of Feldman-Cousins

- With a physical constraint (μ>0) the confidence band changes

$$t_\mu = -2\ln\lambda(\mu)$$

$$\tilde{t}_\mu = -2\ln\tilde{\lambda}(\mu) = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0\,, \\ -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0\,. \end{cases}$$





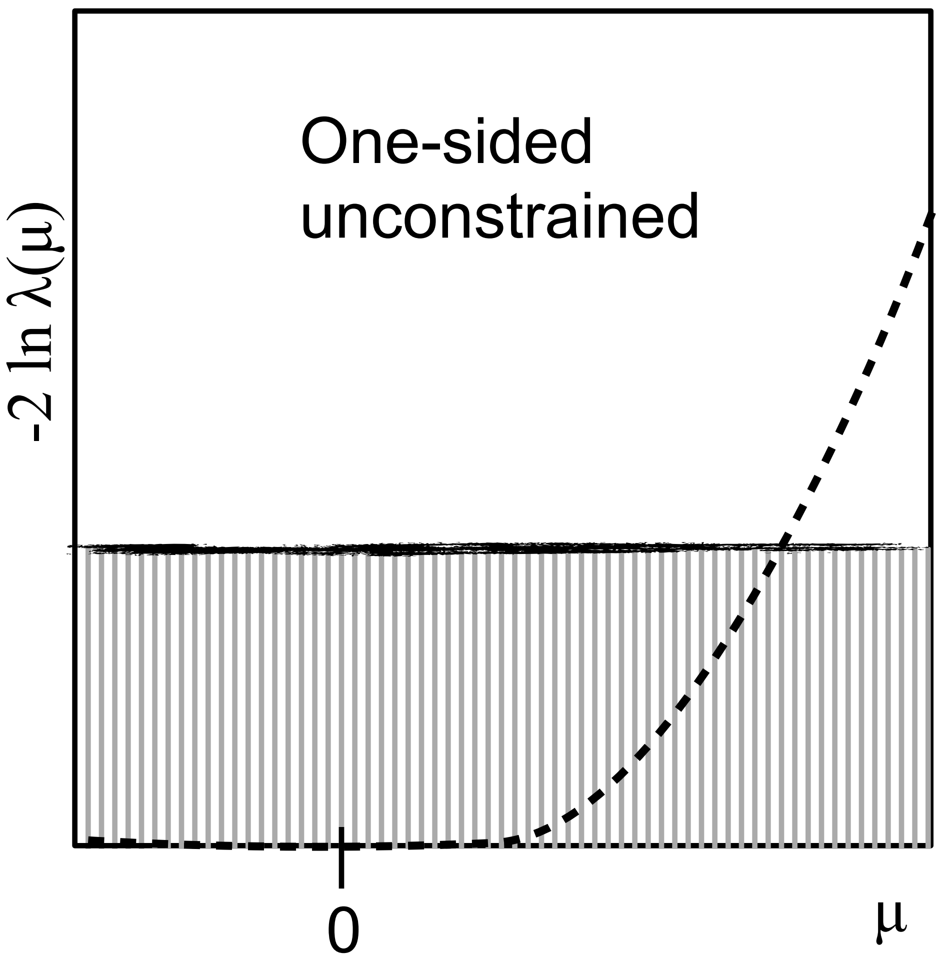
Modified test statistic for 1-sided upper limits

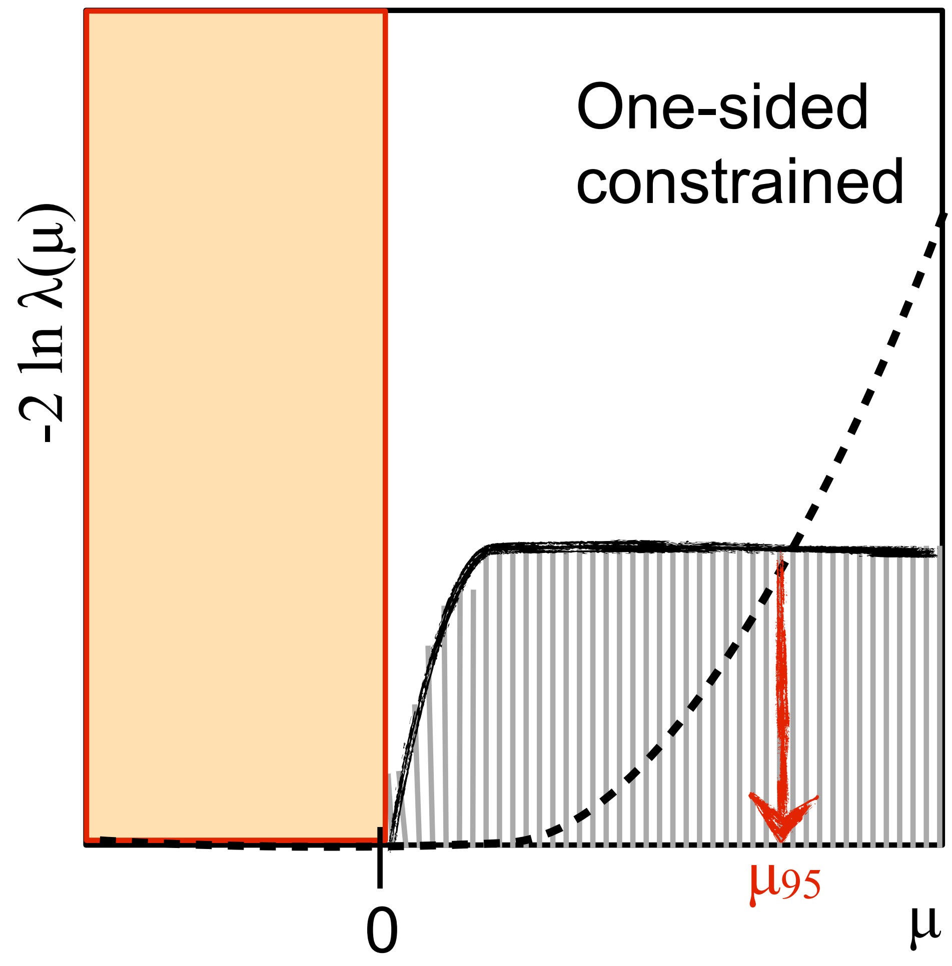
For 1-sided upper-limit the threshold on the test statistic is different

- and with physical boundaries, it is again more complicated

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu\ , \\ 0 & \hat{\mu} > \mu\ , \end{cases}$$

$$\tilde{q}_\mu = \begin{cases} -2\ln \frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0 \\ -2\ln \frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})} & 0 \le \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu\ . \end{cases}$$





"The Asimov paper"

Center for Cosmology and Particle Physics

Recently we showed how to generalize this asymptotic approach

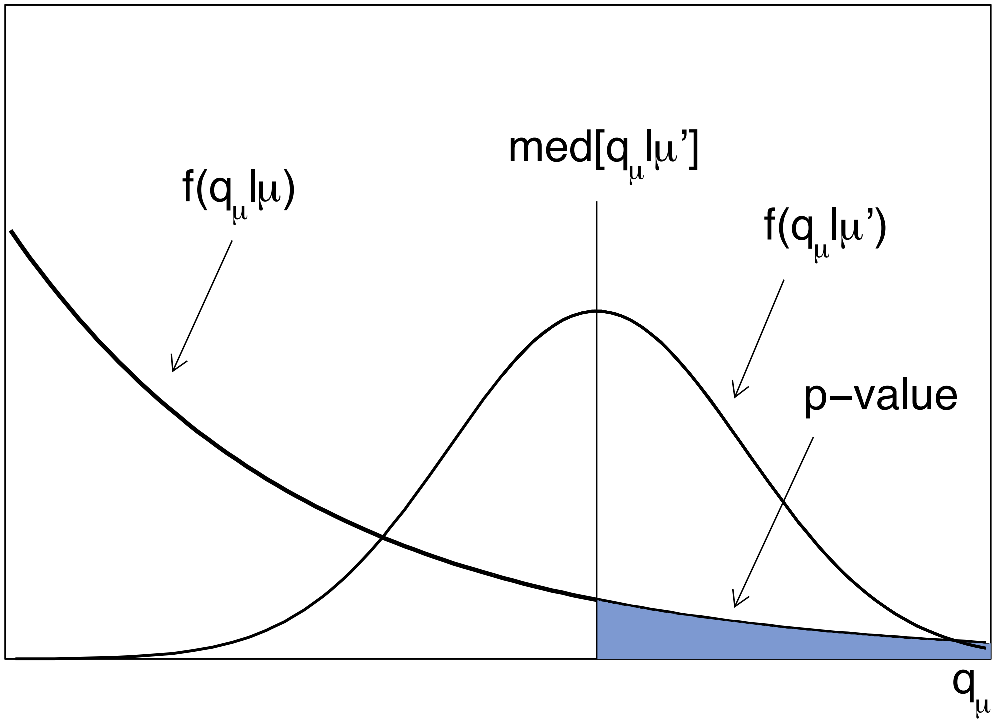
- generalize Wilks's theorem when boundaries are present
- use result of Wald to get $f(-2\log\lambda(\mu) \mid \mu')$

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells

Eur.Phys.J.C71:1554,2011

http://arxiv.org/abs/1007.1727v2



Wald's theorem allows one to find the distribution of -2logλ(μ) when μ is not true -- the result is a non-central chi-square distribution

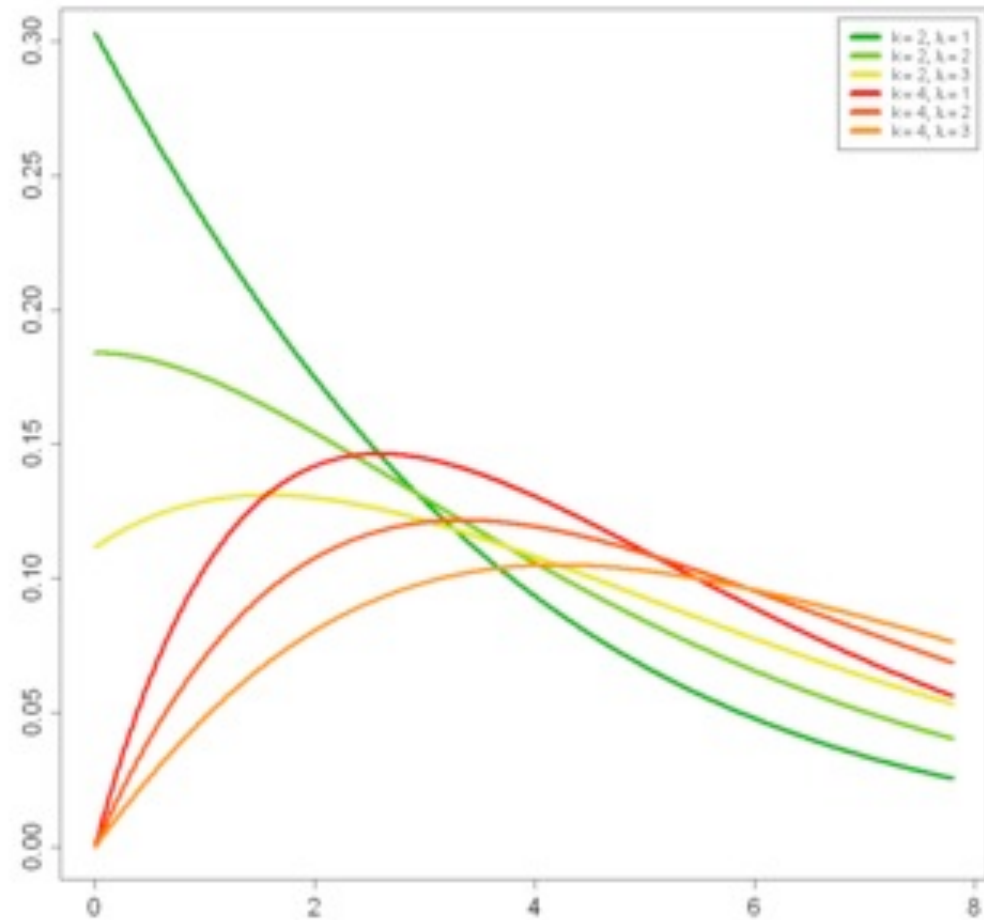
Let X_i be *k* independent, normally distributed random variables with means μ_i and variances . Then the random variable

$$\sum_{i=1}^{k} \left(\frac{X_i}{\sigma_i}\right)^2$$

is distributed according to the noncentral chi-square distribution. It has two parameters: k which specifies the number of degrees of freedom (i.e. the number of X_i), and λ which is related to the mean of the random variables X_i by:

$$\lambda = \sum_{i=1}^{k} \left(\frac{\mu_i}{\sigma_i}\right)^2 .$$

λ is sometime called the noncentrality parameter. Note that some references define λ in other ways, such as half of the above sum, or its square root.



The main results

The Model is just a binned version of the marked Poisson we have considered

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta}_s)\, dx\,,$$

$$b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta}_b)\, dx\,.$$

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

$$E[m_i] = u_i(\boldsymbol{\theta})$$

The "Asimov Data" is an artificial dataset where the "observations" are set equal to the expected values given the parameters of the model

$$n_{i,\text{A}} = E[n_i] = \nu_i = \mu' s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta})\,,$$

$$m_{i,\text{A}} = E[m_i] = u_i(\boldsymbol{\theta})\,.$$

We proved that fits to the Asimov data can be used to get the non-centrality parameter needed for Wald's theorem

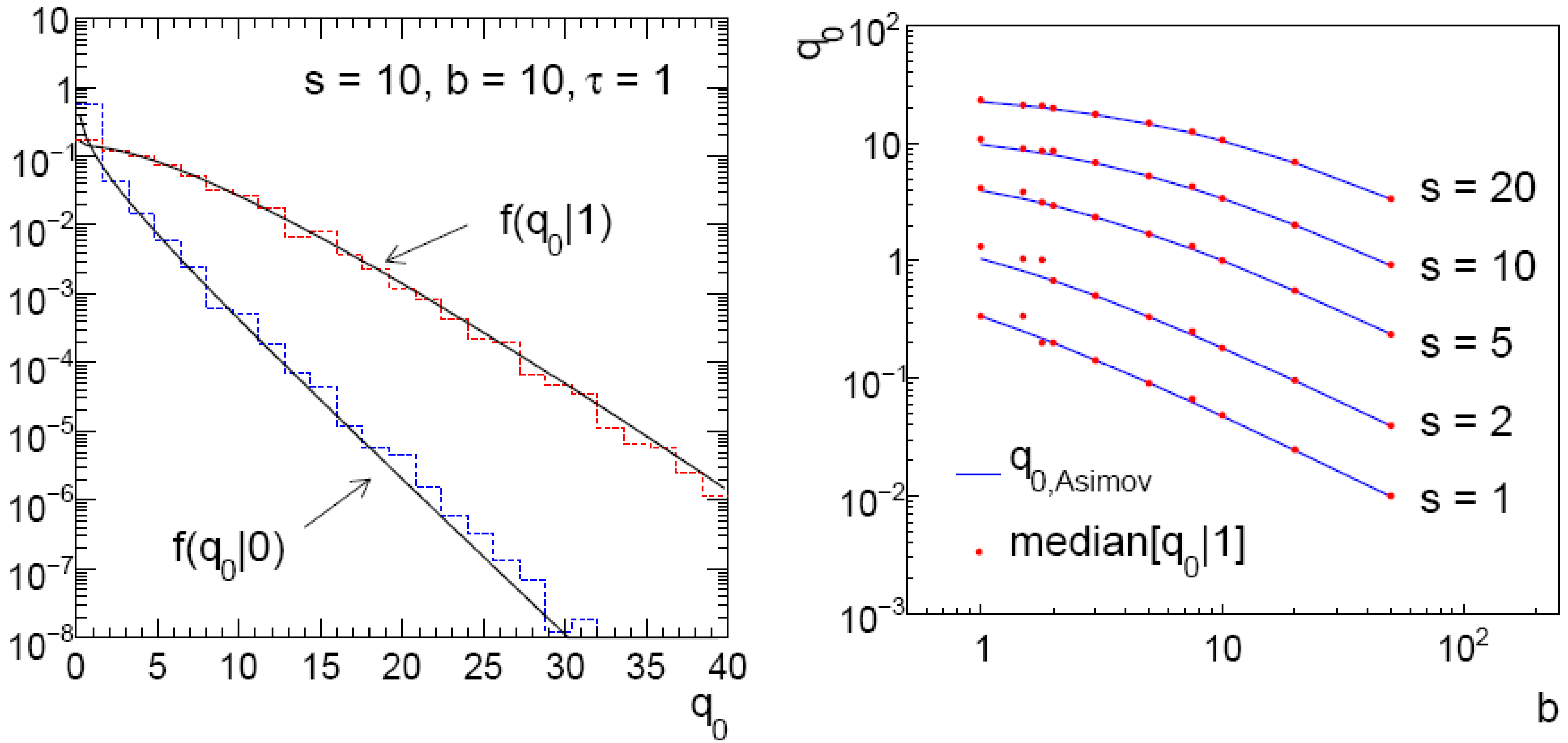
$$\frac{\partial^2 \ln L}{\partial\theta_j \partial\theta_k} = \sum_{i=1}^{N} \left[\left(\frac{n_i}{\nu_i} - 1 \right) \frac{\partial^2 \nu_i}{\partial\theta_j \partial\theta_k} - \frac{\partial\nu_i}{\partial\theta_j} \frac{\partial\nu_i}{\partial\theta_k} \frac{n_i}{\nu_i^2} \right] + \sum_{i=1}^{M} \left[\left(\frac{m_i}{u_i} - 1 \right) \frac{\partial^2 u_i}{\partial\theta_j \partial\theta_k} - \frac{\partial u_i}{\partial\theta_j} \frac{\partial u_i}{\partial\theta_k} \frac{m_i}{u_i^2} \right]$$

$$-2 \ln \lambda_{\text{A}}(\mu) \approx \frac{(\mu - \mu')^2}{\sigma^2} = \Lambda$$

Monte Carlo test of asymptotic formulae

Asymptotic $f(q_0|1)$ good already for fairly small samples.

Median$[q_0|1]$ from Asimov data set; good agreement with MC.



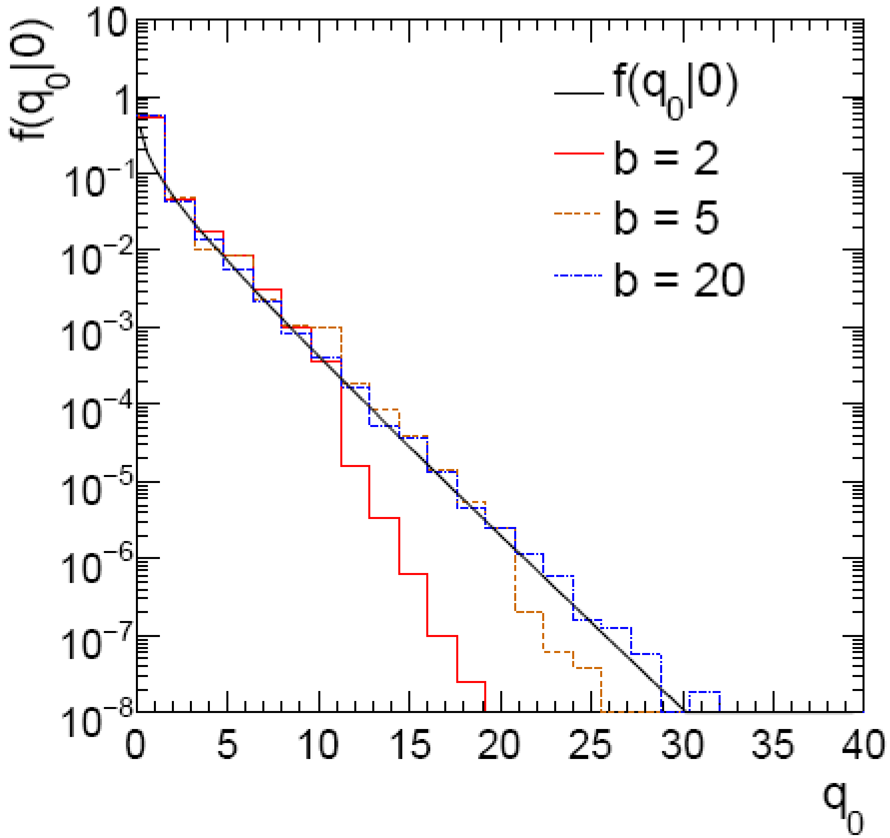
Monte Carlo test of asymptotic formula

$$n \sim \text{Poisson}(\mu s + b)$$

$$m \sim \text{Poisson}(\tau b)$$

Here take $\tau = 1$.

Asymptotic formula is good approximation to 5σ level ($q_0 = 25$) already for $b \sim 20$.

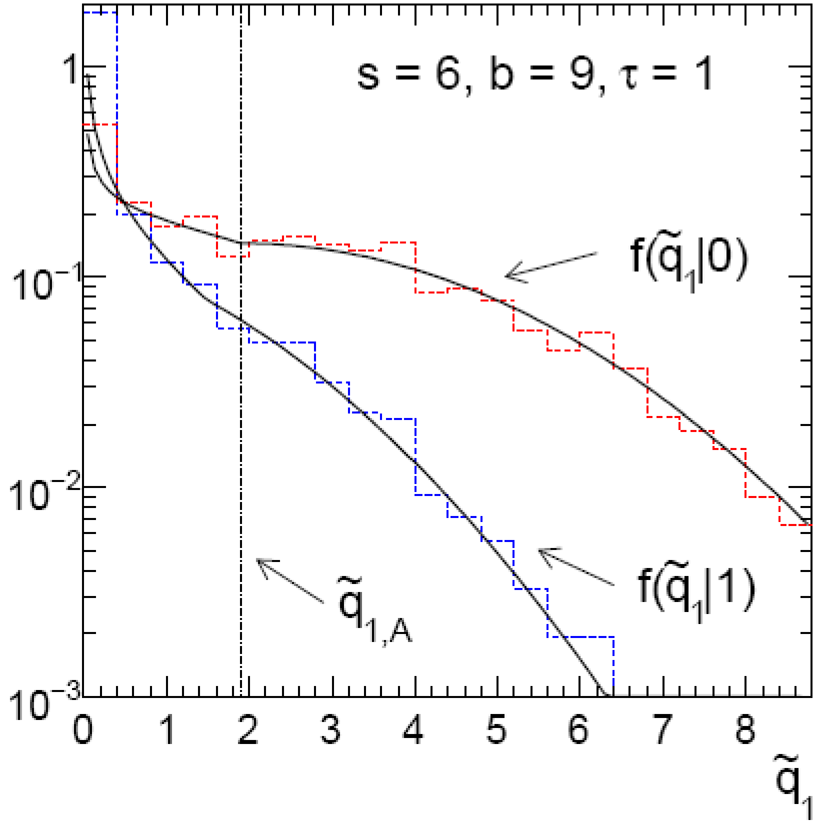


Center for Cosmology and Particle Physics

Monte Carlo test of asymptotic formulae

Same message for test based on $\tilde{q}_\mu$.

q_μ and $\tilde{q}_\mu$ give similar tests to the extent that asymptotic formulae are valid.



G. Cowan Using the Profile Likelihood in Searches for New Physics / Banff 2010 24

Some non-trivial tests: boundaries

Center for
Cosmology and
Particle Physics

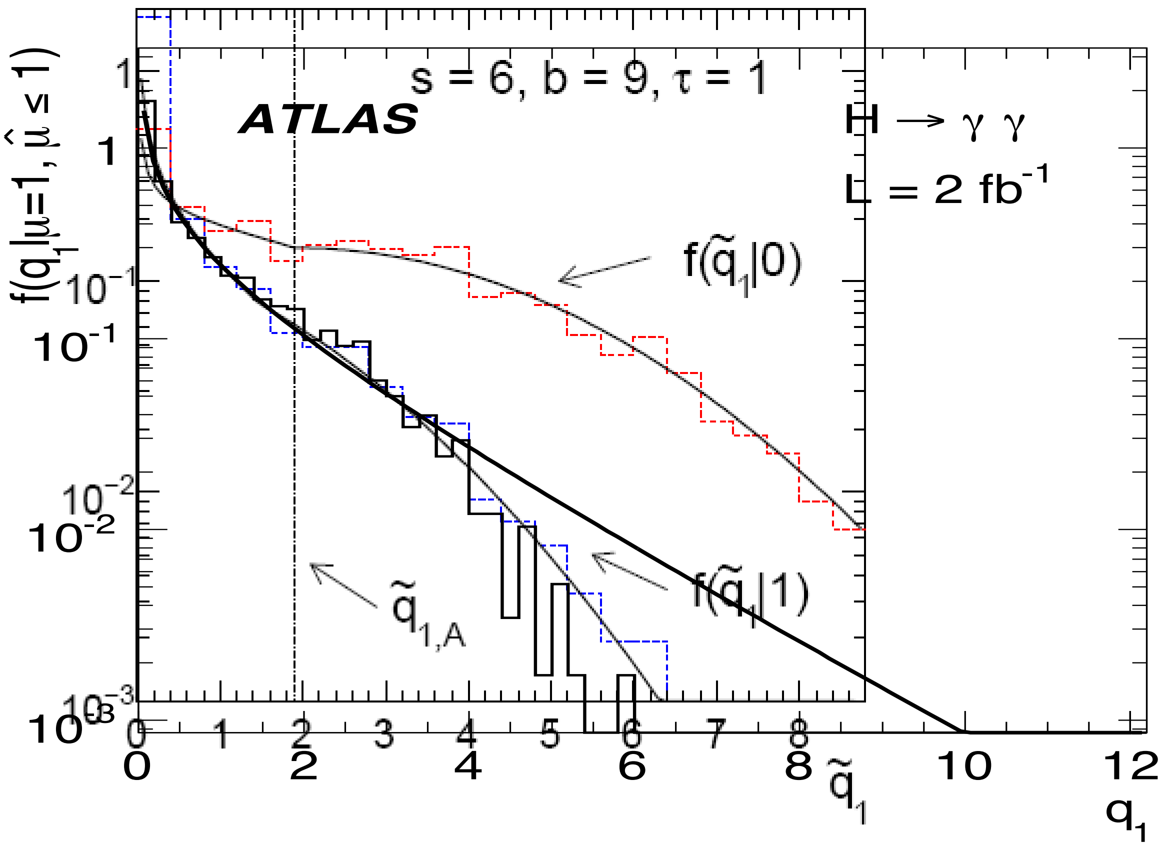
Monte Carlo test of asymptotic formulae

$$
f(\tilde{q}_\mu|\mu') = \Phi\left(\frac{\mu'-\mu}{\sigma}\right)\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{\tilde{q}_\mu}-\frac{\mu-\mu'}{\sigma}\right)^2\right] & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2\,, \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2\,. \end{cases}
$$

Same message for test based on $\tilde{q}_\mu$.

q_μ and $\tilde{q}_\mu$ give similar tests to the extent that asymptotic formulae are valid.

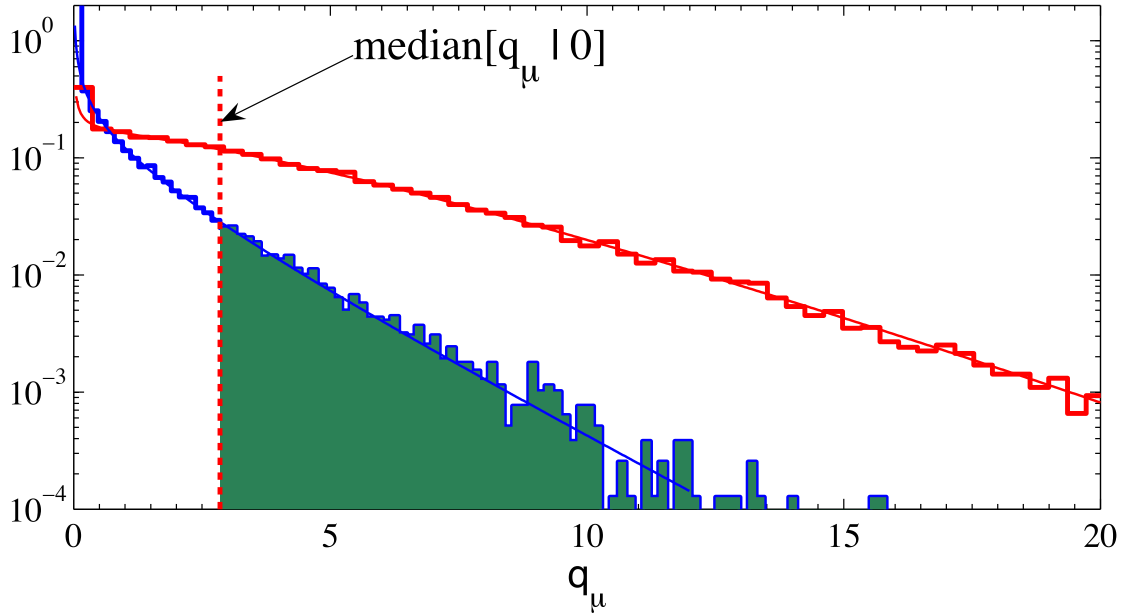
We now can describe effect of the boundary on the distribution of the test statistic.

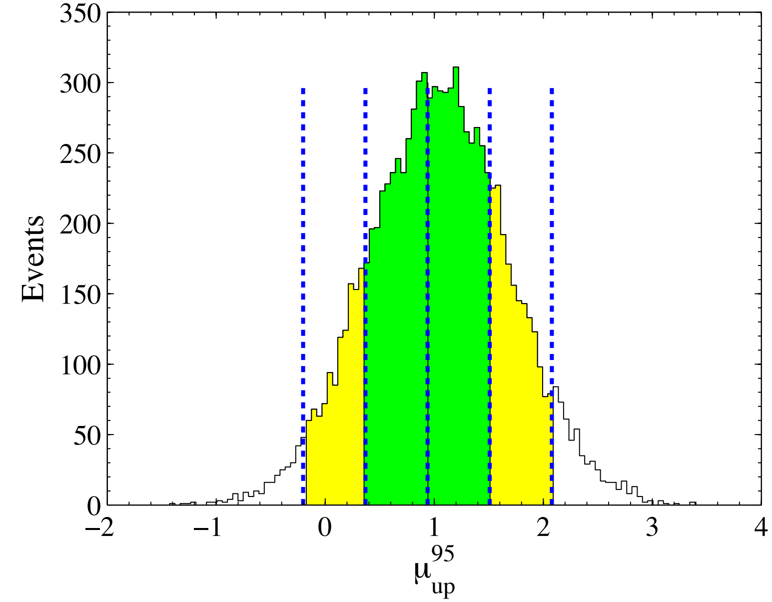


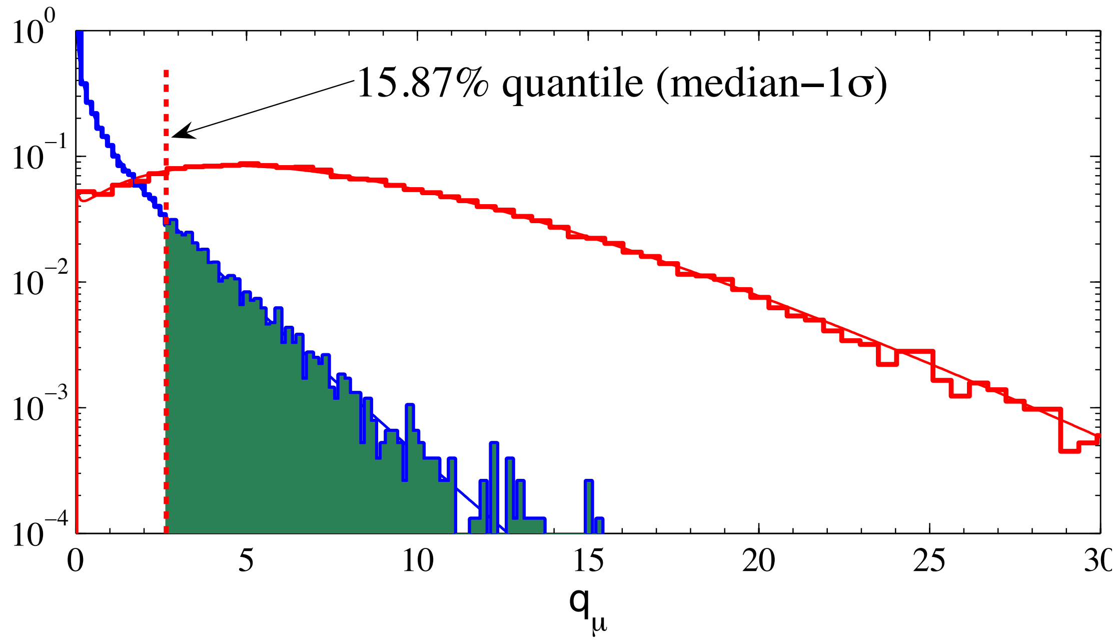
Median & bands from asymptotics

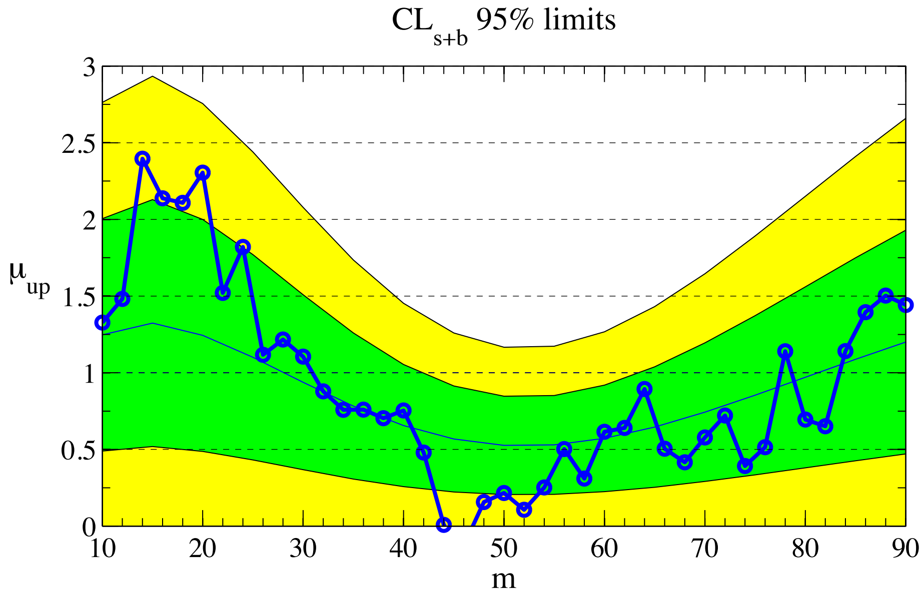
Center for Cosmology and Particle Physics

Get Median and bands in seconds, not days!





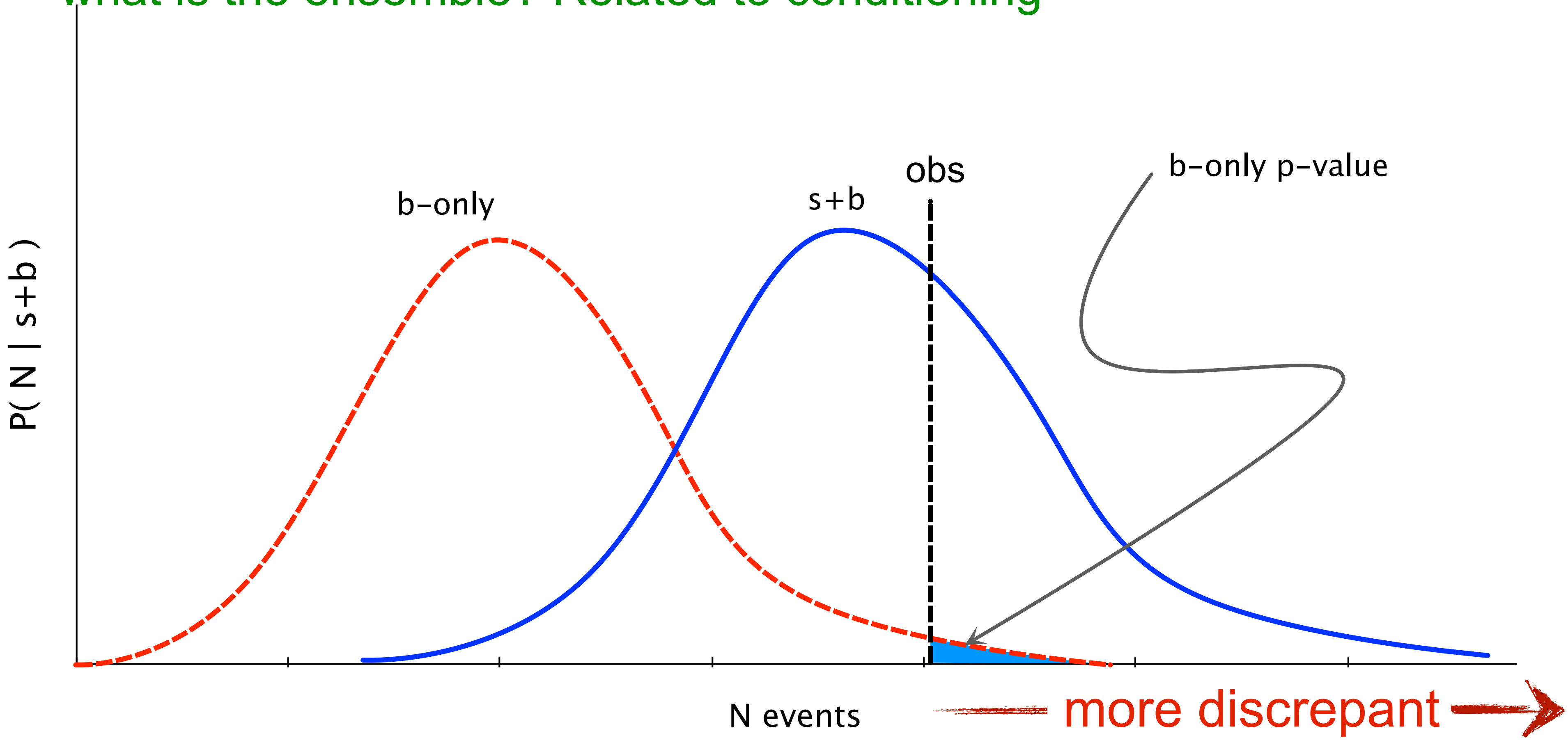




The problem with p-values

The decision to reject the null hypothesis is based on the probability for data you didn't get to agree less well with the hypothesis...

- doesn't sound very convincing when you put it that way. Other criticisms:
  - test statistic is "arbitrary" (not really, it is designed to be powerful against an alternative)
  - what is the ensemble? Related to conditioning



Conditioning (cont.)

- The 1956 thought expt of David R. Cox focused the issue:
  - Your procedure for weighing an object consists of flipping a coin to decide whether to use a weighing machine with a 10% error or one with a 1% error; and then measuring the weight.
  - Then "surely" the error you quote for your measurement should reflect which weighing machine you actually used, and not the average error of the "whole space" of all measurements!
  - But classic most powerful N-P hypothesis test uses the whole space!
- In more complicated situations, ancillary statistics do not exist, and it is not at all clear how to restrict the "whole space" to the relevant part for frequentist coverage.
- …in methods obeying the likelihood principle, in effect one conditions on the exact data obtained, giving up the frequentist coverage criterion for the guarantee of relevance.

Bob Cousins, CMS, 2008 51

Cox's Conditioning argument

[6], Sir David Cox gave a simple convincing example in 1958 that the most powerful test is not always the most relevant test. A version of the argument adapted to HEP is as follows. (Cox's arguments is often applied to "weighing machines", although that phrase is not actually in Ref. [6].)

Suppose that one is "weighing" an elementary particle, i.e., measuring the mass m of a particle that happens to have two decay modes i, each with 50% branching fraction. Suppose that the mass measurement for decay mode $i = 1$ has mass resolution with rms $\sigma_1 = 10$ GeV, and for decay mode $i = 2$, it is $\sigma_2 = 1$ GeV. (The modes are distinguishable; one could be decay to neutrals and the other to charged tracks.) One is testing the null hypothesis that predicts the mass to be 100 GeV in a one-sided test against larger alternative masses. We set the significance level to be 0.05. I.e., from the data, we use a recipe to calculate a 95% C.L. lower limit on m and compare to 100 GeV.

We do the experiment and get one decay sampled randomly from the two modes, with measured mass x sampled randomly from a Gaussian with the resolution for that mode. Using the fact that the one-tailed probability for $x > 1.64\sigma$ is 0.05 for a Gaussian, the "obvious" recipe for testing the null hypothesis ($m = 100$ GeV) is:

- If decay mode 1 is observed, reject the null if $x > 100 \text{ GeV} + 1.64\,\sigma_1 = 116.4$ GeV;
- If decay mode 2 is observed, reject the null if $x > 100 \text{ GeV} + 1.64\,\sigma_2 = 101.64$ GeV.

I.e., we use the σ which is *relevant* for the mode actually observed in performing the one-sided test. One says that the tail probabilities that are calculated are *conditional probabilities*, calculated *conditionally on the mode that was actually observed.*

except from a note by Bob Cousins

Cox's Conditioning argument

Center for Cosmology and Particle Physics

It is easy to see that this is *not* the same result that one obtains by using the *unconditional* probability for obtaining x, which is the sum of two Gaussians (one with σ_1 and one with σ_2), each weighted by 0.5.

Now let us consider a specific alternative hypothesis $m_A = 110$ GeV, and ask what is the power $(1-\beta)$ of the above conditional test. The probability β of accepting the null (100 GeV) when the alternative (110 GeV) is true is $p(x < 116.4 \,|\, m = 110) \approx 0.75$ for mode 1 and $p(x < 101.64 \,|\, m = 110) \approx 0$ for mode 2. Recalling that the probability of each mode is 50%, $\beta \approx 0.38$ and the power is $1 - \beta = 0.62$.

Remarkably, it is easy to show that, among tests with significance level 0.05, this is not the most powerful test for the whole sample space, i.e., for the unconditional ensemble which includes both decay modes. A test which is more powerful against the alternative 110 GeV is:

- If decay mode 1 is observed, reject the null if $x > 100$ GeV $+ 1.28\,\sigma_1 = 112.8$ GeV;
- If decay mode 2 is observed, reject the null if $x > 100$ GeV $+ 5\,\sigma_2 = 105$ GeV.

The significance level is again 0.05. The Type 1 errors are not divided equally between the two modes, but rather occur $\sim 10\%$ of the time in decay mode 1, and by comparison negligibly in decay mode 2.

The probability β of accepting the null (100 GeV) when the alternative m_A (110 GeV) is true is $p(x < 112.8 \,|\, m = 110) \approx 0.4$ for mode 1 and $p(x < 105 \,|\, m = 110) \approx 0$ for mode 2. Recalling that the probability of each mode is 50%, $\beta \approx 0.2$ and the power is 0.8.

except from a note by Bob Cousins

Likelihood Principle

- **As noted above, in both Bayesian methods and likelihood-ratio based methods, the probability (density) for obtaining the *data at hand* is used (via the likelihood function), *but probabilities for obtaining other data are not used!***
- **In contrast, in typical frequentist calculations (e.g., a p-value which is the probability of obtaining a value as extreme or *more extreme* than that observed), one uses probabilities of data *not seen*.**
- **This difference is captured by the *Likelihood Principle**: If two experiments yield likelihood functions which are proportional, then Your inferences from the two experiments should be identical.**
- **L.P. is built in to Bayesian inference (except e.g., when Jeffreys prior leads to violation).**
- **L.P. is violated by p-values and confidence intervals.**
- **Although practical experience indicates that the L.P. may be too restrictive, it is useful to keep in mind. When frequentist results "make no sense" or "are unphysical", in my experience the underlying reason can be traced to a bad violation of the L.P.**

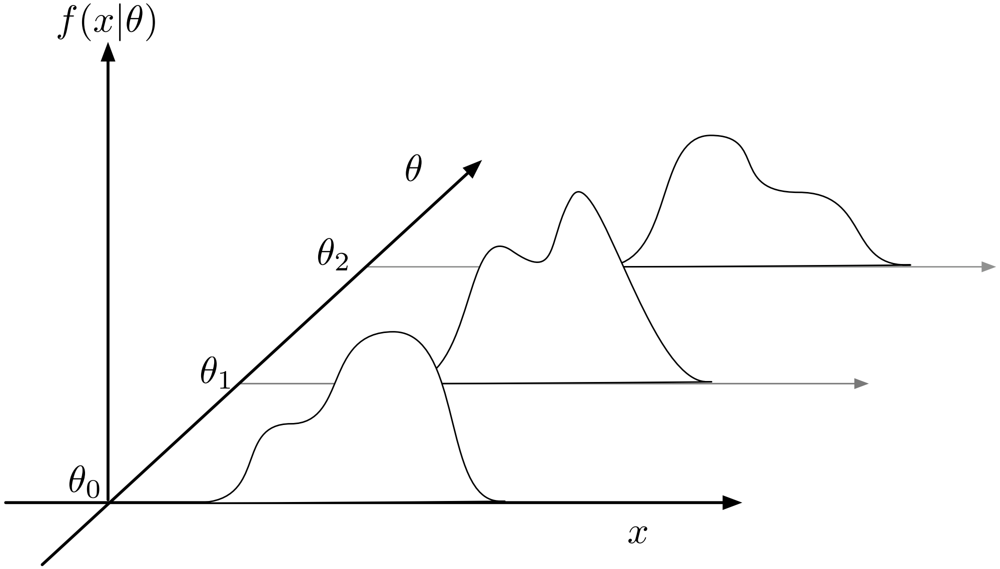
***There are various versions of the L.P., strong and weak forms, etc.**

Bob Cousins, CMS, 2008 46

Goal of Likelihood-based Methods

Likelihood-based methods settle between two conflicting desires:

- We want to obey the likelihood principle because it implies a lot of nice things and sounds pretty attractive
- We want nice frequentist properties (and the only way we know to incorporate those properties "by construction" will violate the likelihood principle)



The asymptotic results give us a a way to approximately cover while simultaneously obeying the likelihood principle and NOT using a prior

Bayesian methods

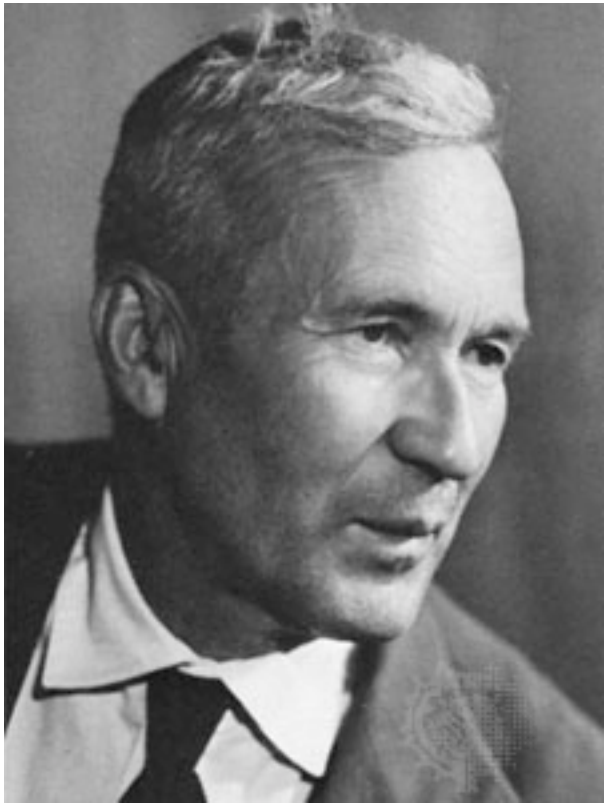
Some personal history

Center for Cosmology and Particle Physics

Archbishop of Canterbury Thomas Cranmer (born: 1489, executed: 1556) author of the "Book of Common Prayer"

Two centuries later (when this Book had become an official prayer book of the Church of England) Thomas Bayes was a non-conformist minister (Presbyterian) who refused to use Cranmer's book

Axioms of Probability



Kolmogorov axioms (1933)

These Axioms are a mathematical starting point for probability and statistics

1. probability for every element, E, is non-negative $P(E) \geq 0 \quad \forall E \subseteq \mathcal{F} = 2^{\Omega}$
2. probability for the entire space of possibilities is 1 $P(\Omega) = 1.$
3. if elements E_i are disjoint, probability is additive $P(E_1 \cup E_2 \cup \cdots) = \sum_i P(E_i).$

Consequences:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(\Omega \setminus E) = 1 - P(E)$$

Different definitions of Probability

Center for Cosmology and Particle Physics

Frequentist

- defined as limit of long term frequency
- probability of rolling a 3 := limit of (# rolls with 3 / # trials)
  - you don’t need an infinite sample for definition to be useful
  - sometimes ensemble doesn’t exist
    - eg. P(Higgs mass = 120 GeV), P(it will snow tomorrow)
- Intuitive if you are familiar with Monte Carlo methods
- compatible with orthodox interpretation of probability in Quantum Mechanics. Probability to measure spin projected on x-axis if spin of beam is polarized along +z

$$|\langle \rightarrow | \uparrow \rangle|^2 = \frac{1}{2}$$

Subjective Bayesian

- Probability is a degree of belief (personal, subjective)
  - can be made quantitative based on betting odds
  - most people’s subjective probabilities are not ***coherent*** and do not obey laws of probability

http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/#3.1

Bayes' Theorem

Bayes' theorem relates the conditional and marginal probabilities of events A & B

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

- P(*A*) is the prior probability or marginal probability of *A*. It is "prior" in the sense that it does not take into account any information about *B*.
- P(*A*|*B*) is the conditional probability of *A*, given *B*. It is also called the posterior probability because it is derived from or depends upon the specified value of *B*.
- P(*B*|*A*) is the conditional probability of *B* given *A*.
- P(*B*) is the prior or marginal probability of *B*, and acts as a normalizing constant

Derivation from conditional probabilities

To derive the theorem, we start from the definition of conditional probability. The probability of event *A* given event *B* is

$$P(A|B) = \frac{P(A\cap B)}{P(B)}.$$

Equivalently, the probability of event *B* given event *A* is

$$P(B|A) = \frac{P(A\cap B)}{P(A)}.$$

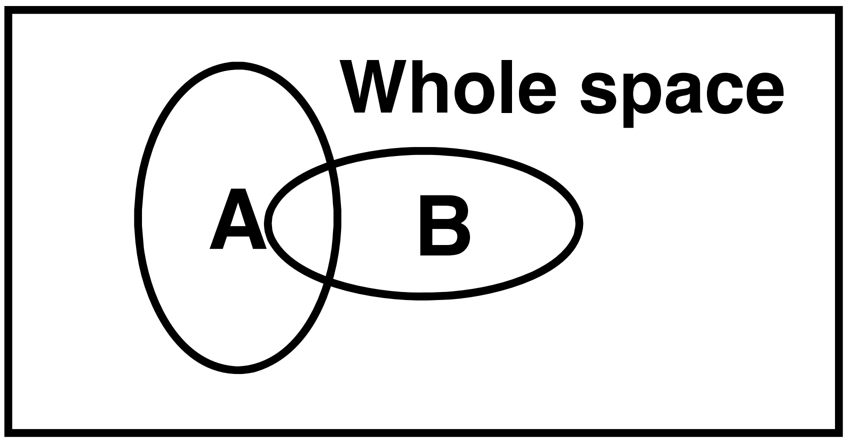
Rearranging and combining these two equations, we find

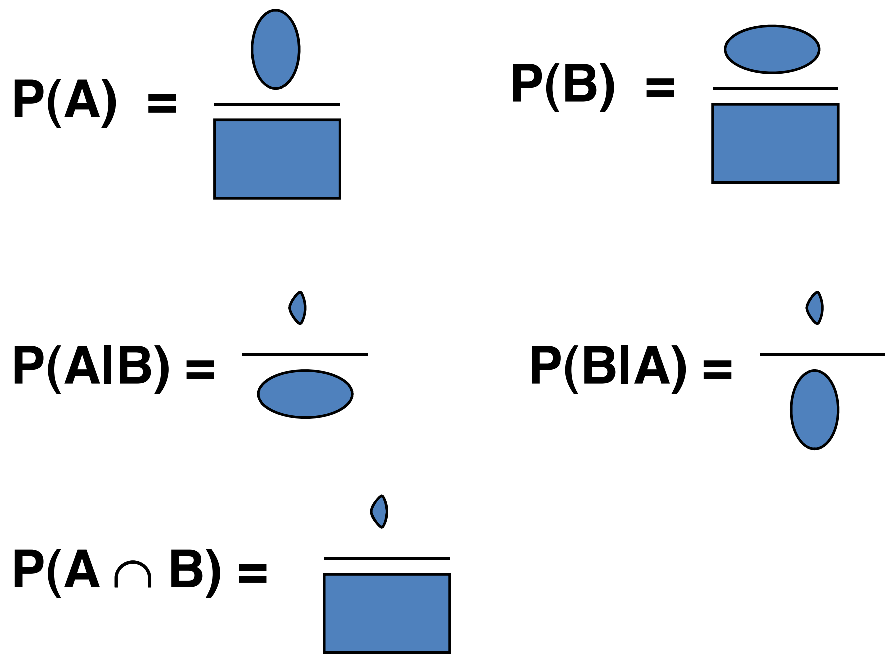
$$P(A|B)\,P(B) = P(A\cap B) = P(B|A)\,P(A).$$

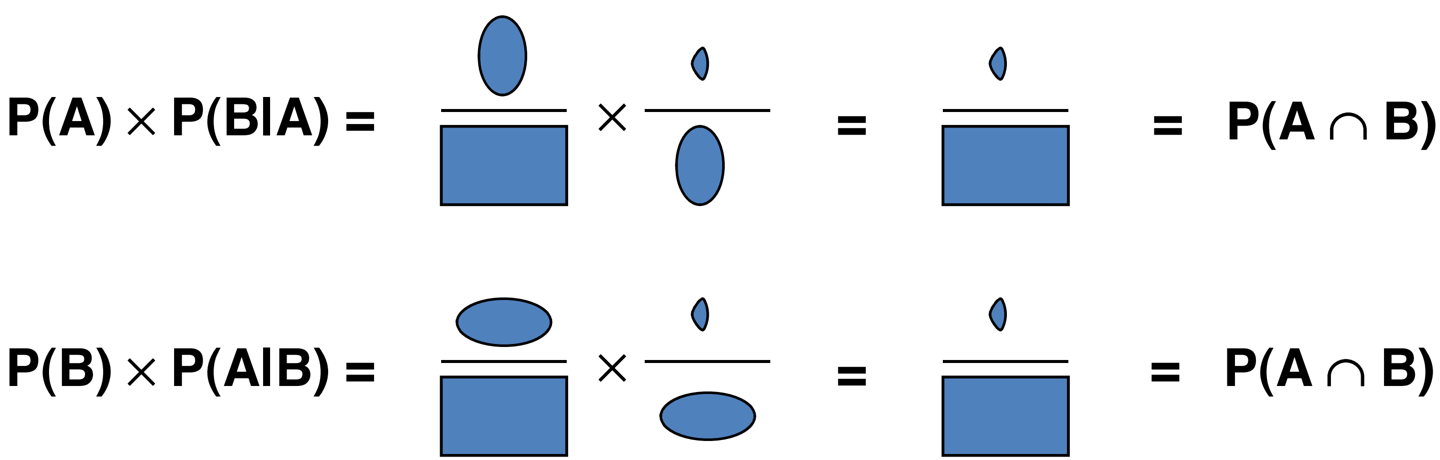
This lemma is sometimes called the product rule for probabilities. Dividing both sides by P(*B*), providing that it is non-zero, we obtain Bayes' theorem:

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}.$$

P, Conditional P, and Derivation of Bayes' Theorem in Pictures



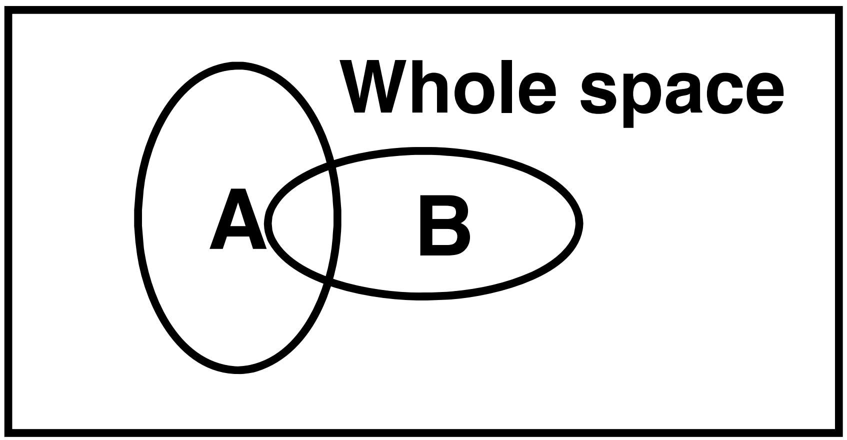


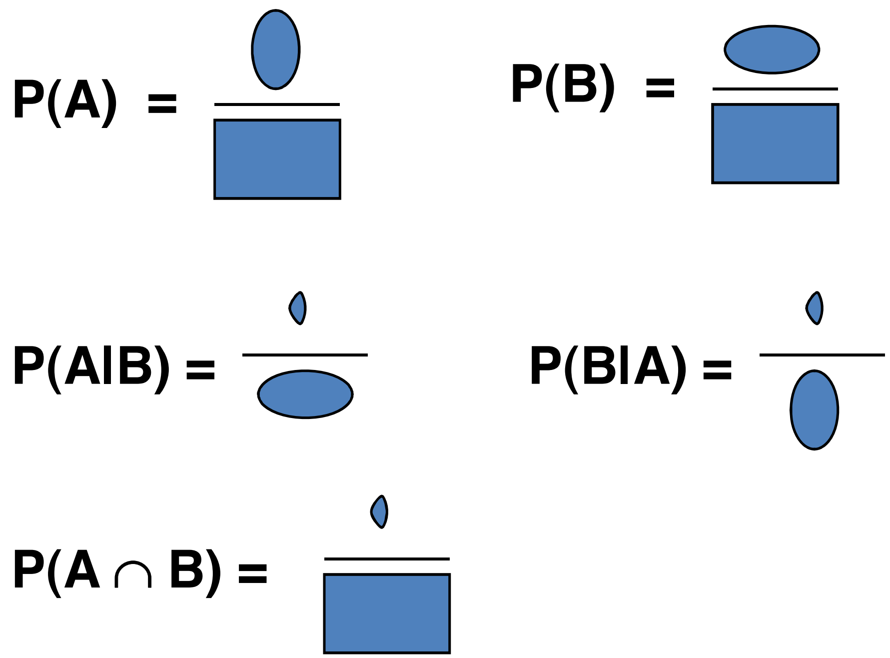


$\Rightarrow$ **P(B|A) = P(A|B) × P(B) / P(A)**

Bob Cousins, CMS, 2008

P, Conditional P, and Derivation of Bayes' Theorem in Pictures





Don't forget about "Whole space" Ω. I will drop it from the notation typically, but occasionally it is important.

Bob Cousins, CMS, 2008

$\Rightarrow$ **P(B|A) = P(A|B) × P(B) / P(A)**

Louis's Example

$$P\,(\text{Data;Theory}) \neq P\,(\text{Theory;Data})$$

Theory = male or female

Data = pregnant or not pregnant

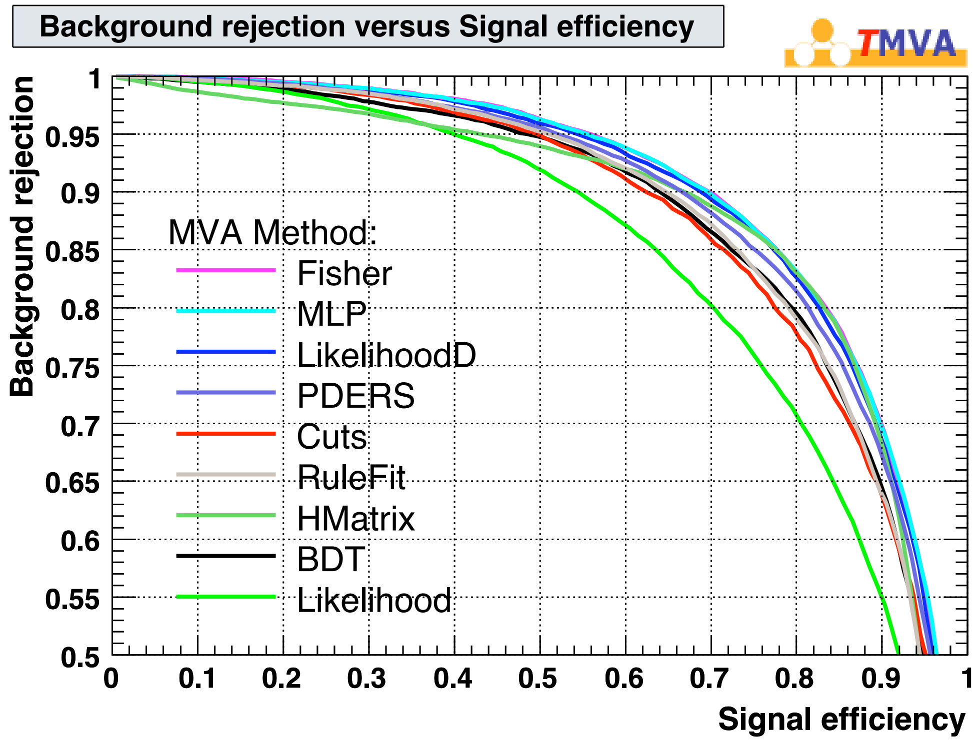
P (pregnant ; female) ~ 3%

but

P (female ; pregnant) >>>3%

Bob's Example

A b-tagging algorithm gives a curve like this



One wants to decide where to cut and to optimize analysis

- For some point on the curve you have:
  - P(btag| b-jet), i.e., efficiency for tagging b's
  - P(btag| not a b-jet), i.e., efficiency for background

Bob's example of Bayes' theorem

Now that you know:

- P(btag| b-jet), i.e., efficiency for tagging b's
- P(btag| not a b-jet), i.e., efficiency for background

Question: Given a selection of jets with btags, what fraction of them are b-jets?

- I.e., **what is P(b-jet | btag) ?**

Answer: Cannot be determined from the given information!

- Need to know **P(b-jet)**: fraction of all jets that are b-jets.
- Then Bayes' Theorem inverts the conditionality:
  - P(b-jet | btag) $\propto$ P(btag|b-jet) P(b-jet)

Note, this use of Bayes' theorem is fine for Frequentist

An different example of Bayes' theorem

Center for
Cosmology and
Particle Physics

An different example of Bayes' theorem

An analysis is developed to search for the Higgs boson

- background expectation is 0.1 events
  - you know P(N | no Higgs)
- signal expectation is 10 events
  - you know P(N | Higgs)

An different example of Bayes' theorem

An analysis is developed to search for the Higgs boson

- background expectation is 0.1 events
  - you know P(N | no Higgs)
- signal expectation is 10 events
  - you know P(N | Higgs)

An different example of Bayes' theorem

An analysis is developed to search for the Higgs boson

- background expectation is 0.1 events
  - you know P(N | no Higgs)
- signal expectation is 10 events
  - you know P(N | Higgs)

Question: one observes 8 events, **what is P(Higgs | N=8) ?**

An different example of Bayes' theorem

An analysis is developed to search for the Higgs boson

- background expectation is 0.1 events
  - you know P(N | no Higgs)
- signal expectation is 10 events
  - you know P(N | Higgs)

Question: one observes 8 events, **what is P(Higgs | N=8) ?**

An different example of Bayes' theorem

An analysis is developed to search for the Higgs boson

- background expectation is 0.1 events
  - you know P(N | no Higgs)
- signal expectation is 10 events
  - you know P(N | Higgs)

Question: one observes 8 events, **what is P(Higgs | N=8) ?**

Answer: Cannot be determined from the given information!

- Need in addition: P(Higgs)
  - no ensemble! no frequentist notion of P(Higgs)
  - prior based on degree-of-belief would work, but it is subjective. This is why some people object to Bayesian statistics for particle physics

Change of variable x, change of parameter θ

- For pdf $p(x|\theta)$ and change of variable from x to y(x):

  $$p(y(x)|\theta) = p(x|\theta) / |dy/dx|.$$

  Jacobian modifies probability *density*, guaranties that

  $$P(\, y(x_1) < y < y(x_2)\,) \;=\; P(x_1 < x < x_2\,),$$ i.e., that

  Probabilities are invariant under change of variable x.

  – Mode of probability *density* is *not* invariant (so, e.g., criterion of maximum probability density is ill-defined).

  – Likelihood *ratio* is invariant under change of variable x. (Jacobian in denominator cancels that in numerator).

- For likelihood $\mathcal{L}(\theta)$ and reparametrization from θ to $u(\theta)$:

  $$\mathcal{L}(\theta) \;=\; \mathcal{L}(u(\theta)) \quad (!).$$

  – Likelihood $\mathcal{L}(\theta)$ is invariant under reparametrization of parameter θ (reinforcing fact that $\mathcal{L}$ is *not* a pdf in θ).

Bob Cousins, CMS, 2008

Probability Integral Transform

"…seems likely to be one of the most fruitful conceptions introduced into statistical theory in the last few years" – Egon Pearson (1938)

Given continuous $x \in (a,b)$, and its pdf $p(x)$, let

$$y(x) = \int_a^x p(x')\,dx' .$$

Then $y \in (0,1)$ and $p(y) = 1$ (uniform) for all y. (!)

So there always exists a metric in which the pdf is uniform.

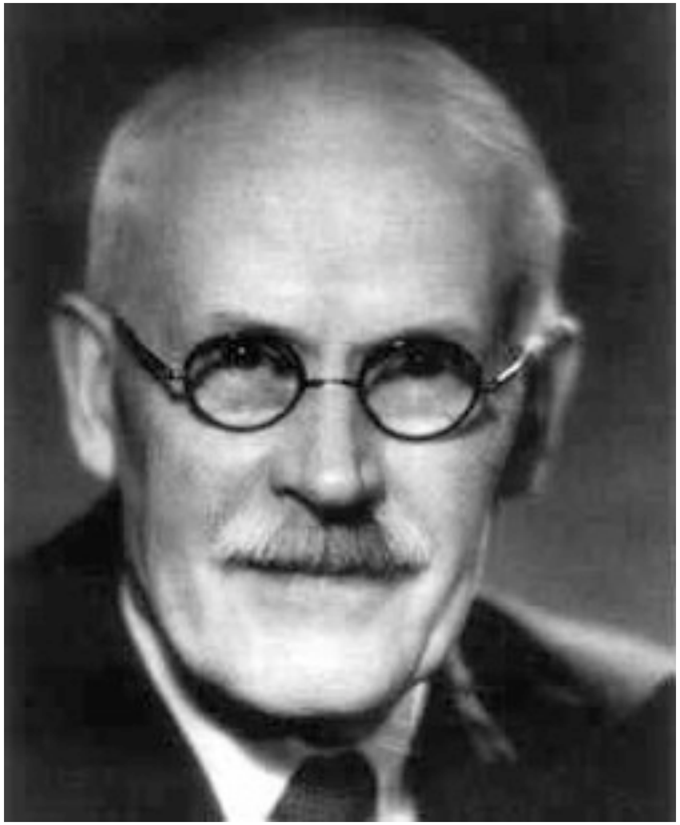
Many issues become more clear (or trivial) after this transformation*. (If x is discrete, some complications.)

The specification of a Bayesian prior pdf $p(\mu)$ for parameter μ is equivalent to the choice of the metric $f(\mu)$ in which the pdf is uniform. This is a *deep* issue, not always recognized as such by users of flat prior pdf's in HEP!

*And the inverse transformation provides for efficient M.C. generation of p(x) starting from RAN().

Bob Cousins, CMS, 2008

The Jeffreys Prior



Physicist Sir Harold Jeffreys had the clever idea that we can "**objectively**" create a flat prior uniform in a metric determined by $I(\theta)$

Adds "minimal information" in a precise sense, and results in: $p(\vec{\theta}) \propto \sqrt{I(\vec{\theta})}$.

It has the key feature that it is invariant under reparameterization of the parameter vector $\vec{\varphi}$ In particular, for an alternate parameterization $\vec{\theta}$ we can derive

$$
\begin{aligned}
p(\vec{\varphi}) &= p(\vec{\theta}) \left| \det\left(\frac{\partial\theta_i}{\partial\varphi_j}\right) \right| \\
&\propto \sqrt{I(\vec{\theta}) \det{}^2\left(\frac{\partial\theta_i}{\partial\varphi_j}\right)} \\
&= \sqrt{\det\left(\frac{\partial\theta_k}{\partial\varphi_i}\right) \det\left(E\left[\frac{\partial \ln L}{\partial\theta_k}\frac{\partial \ln L}{\partial\theta_l}\right]\right) \det\left(\frac{\partial\theta_l}{\partial\varphi_j}\right)} \\
&= \sqrt{\det\left(E\left[\sum_{k,l}\frac{\partial\theta_k}{\partial\varphi_i}\frac{\partial \ln L}{\partial\theta_k}\frac{\partial \ln L}{\partial\theta_l}\frac{\partial\theta_l}{\partial\varphi_j}\right]\right)} \\
&= \sqrt{\det\left(E\left[\frac{\partial \ln L}{\partial\varphi_i}\frac{\partial \ln L}{\partial\varphi_j}\right]\right)} = \sqrt{I(\vec{\varphi})}.
\end{aligned}
$$

Unfortunately, the Jeffreys prior in multiple dimensions causes some problems, and in certain circumstances gives undesirable answers.

Jeffreys's Prior

Center for Cosmology and Particle Physics

Jeffreys's Prior is an "objective" prior based on formal rules (it is related to the Fisher Information and the Cramér-Rao bound]

$$\pi(\vec{\theta}) \propto \sqrt{\det \mathcal{I}\left(\vec{\theta}\right)}.$$

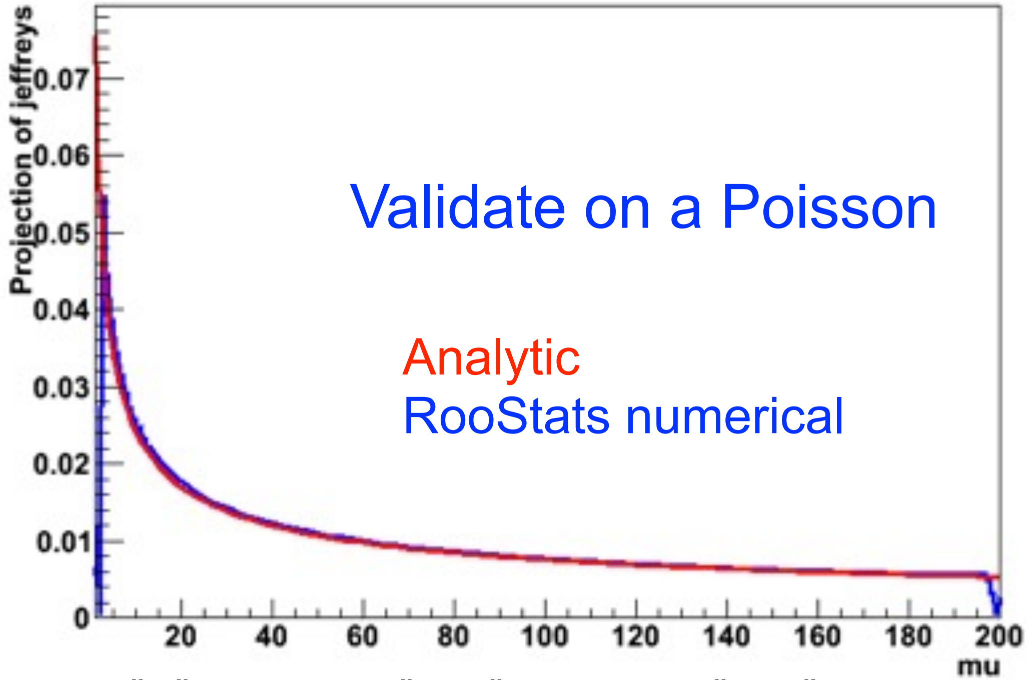
$$\left(\mathcal{I}\left(\theta\right)\right)_{i,j} = -\operatorname{E}\left[\left.\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ln f(X;\theta)\right|\theta\right].$$

Eilam, Glen, Ofer, and I showed in arXiv:1007.1727 that the Asimov data provides a fast, convenient way to calculate the Fisher Information

$$V_{jk}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_j\partial\theta_k}\right] = -\frac{\partial^2 \ln L_{\rm A}}{\partial\theta_j\partial\theta_k} = \sum_{i=1}^{N}\frac{\partial\nu_i}{\partial\theta_j}\frac{\partial\nu_i}{\partial\theta_k}\frac{1}{\nu_i} + \sum_{i=1}^{M}\frac{\partial u_i}{\partial\theta_j}\frac{\partial u_i}{\partial\theta_k}\frac{1}{u_i}$$

Use this as basis to calculate Jeffreys's prior for an arbitrary PDF!

```
RooWorkspace w("w");
w.factory("Uniform::u(x[0,1])");
w.factory("mu[100,1,200]");
w.factory("ExtendPdf::p(u,mu)");

w.defineSet("poi","mu");
w.defineSet("obs","x");
//  w.defineSet("obs2","n");
```

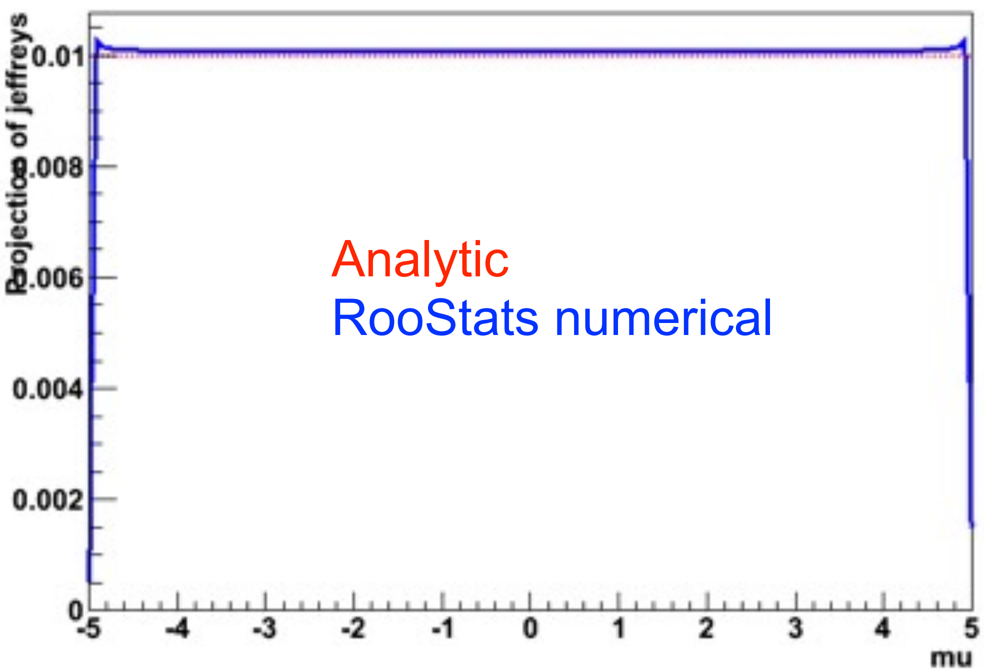


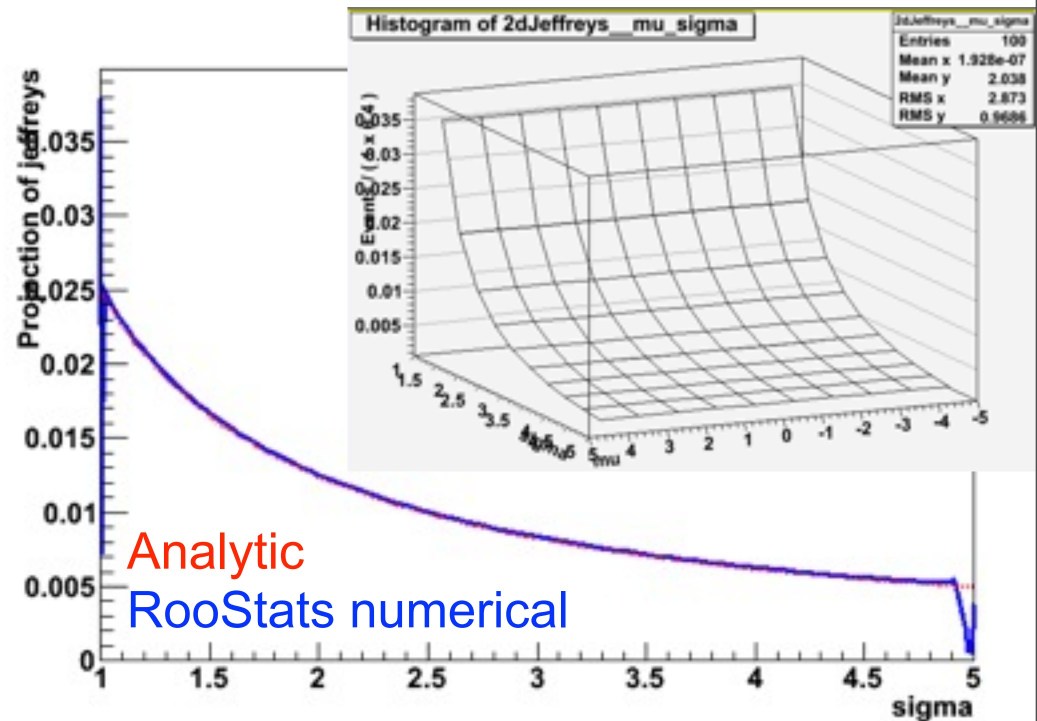

```
RooJeffreysPrior pi("jeffreys","jeffreys",*w.pdf("p"),*w.set("poi"),*w.set("obs"));
```

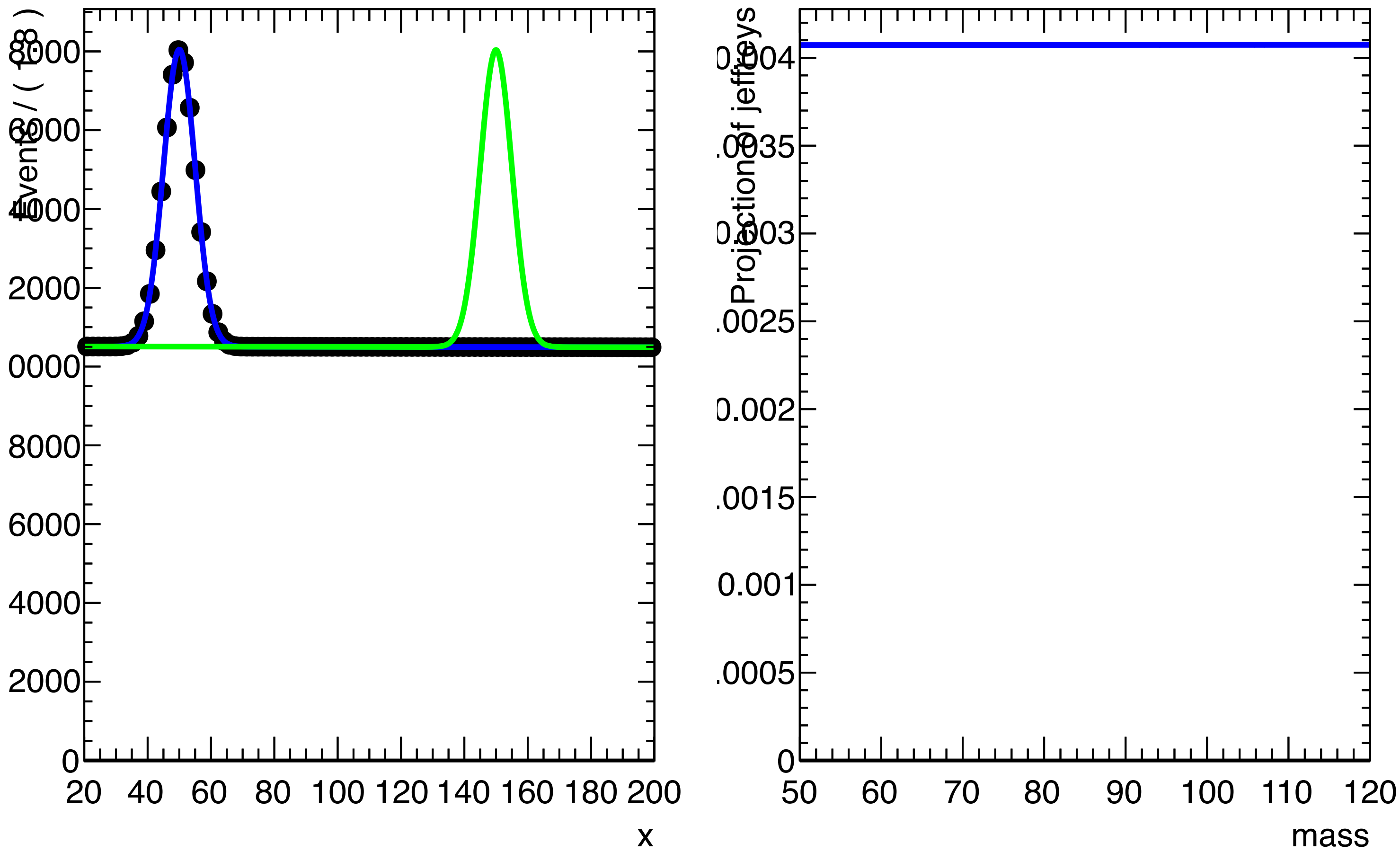
Jeffreys's Prior

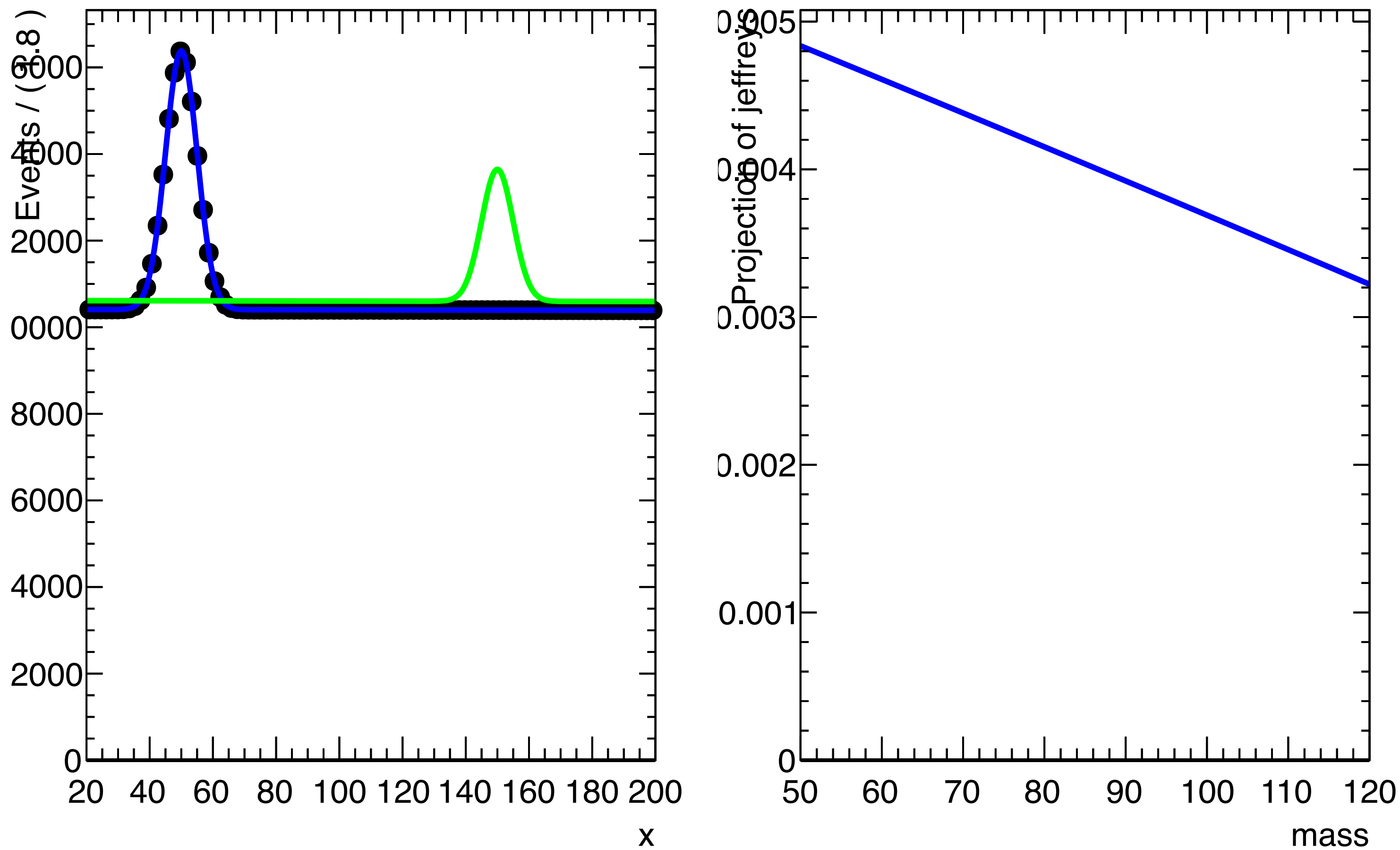
Validate Jeffreys's Prior on a Gaussian μ, σ, and (μ,σ)

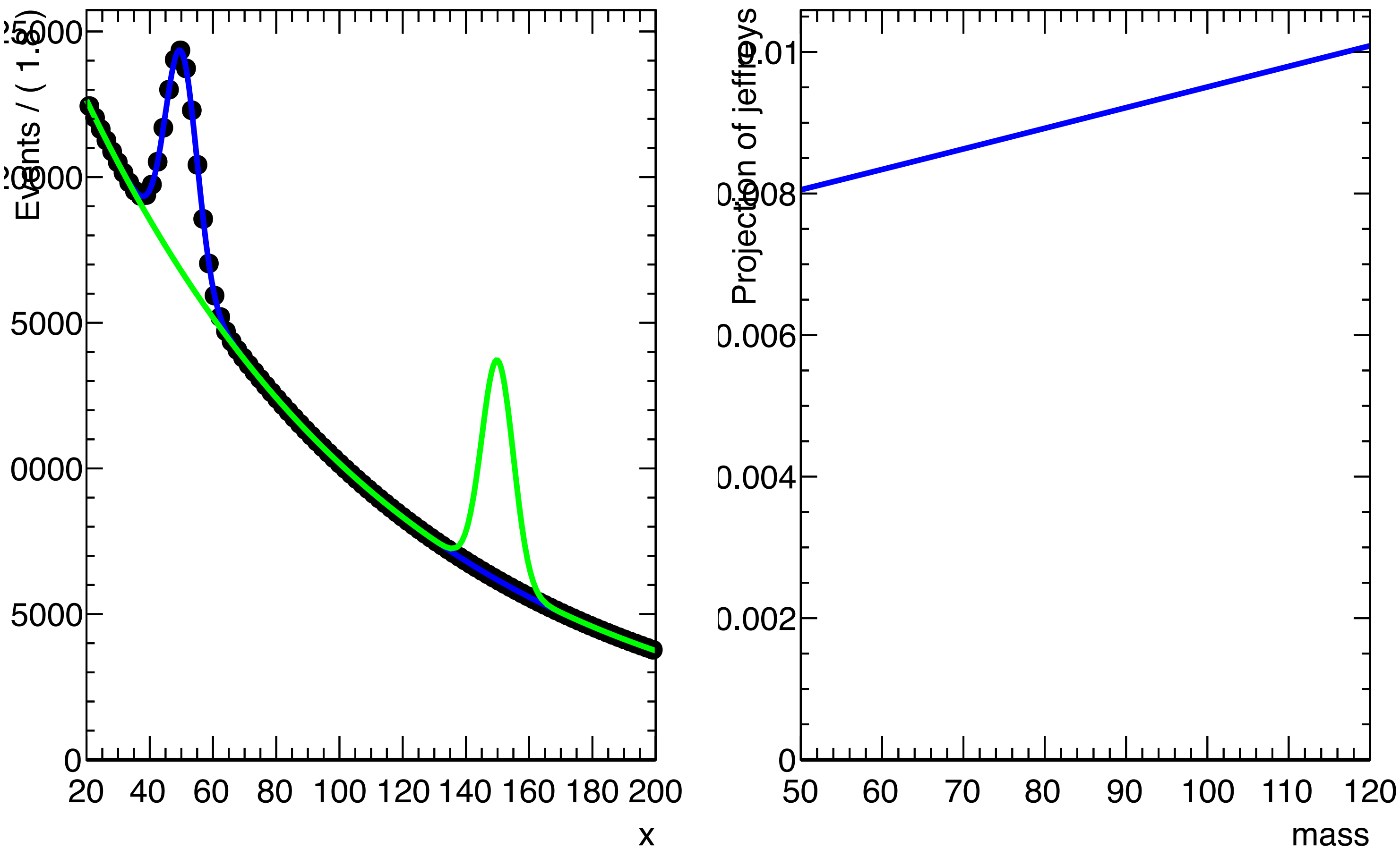
```
RooWorkspace w("w");
 w.factory("Gaussian::g(x[0,-20,20],mu[0,-5,5],sigma[1,0,10])");
 w.factory("n[10,.1,200]");
 w.factory("ExtendPdf::p(g,n)");
 w.var("n")->setConstant();

 w.var("sigma")->setConstant();
 w.defineSet("poi","mu");
 w.defineSet("obs","x");
 RooJeffreysPrior pi("jeffreys","jeffreys",*w.pdf("p"),*w.set("poi"),*w.set("obs"));
```

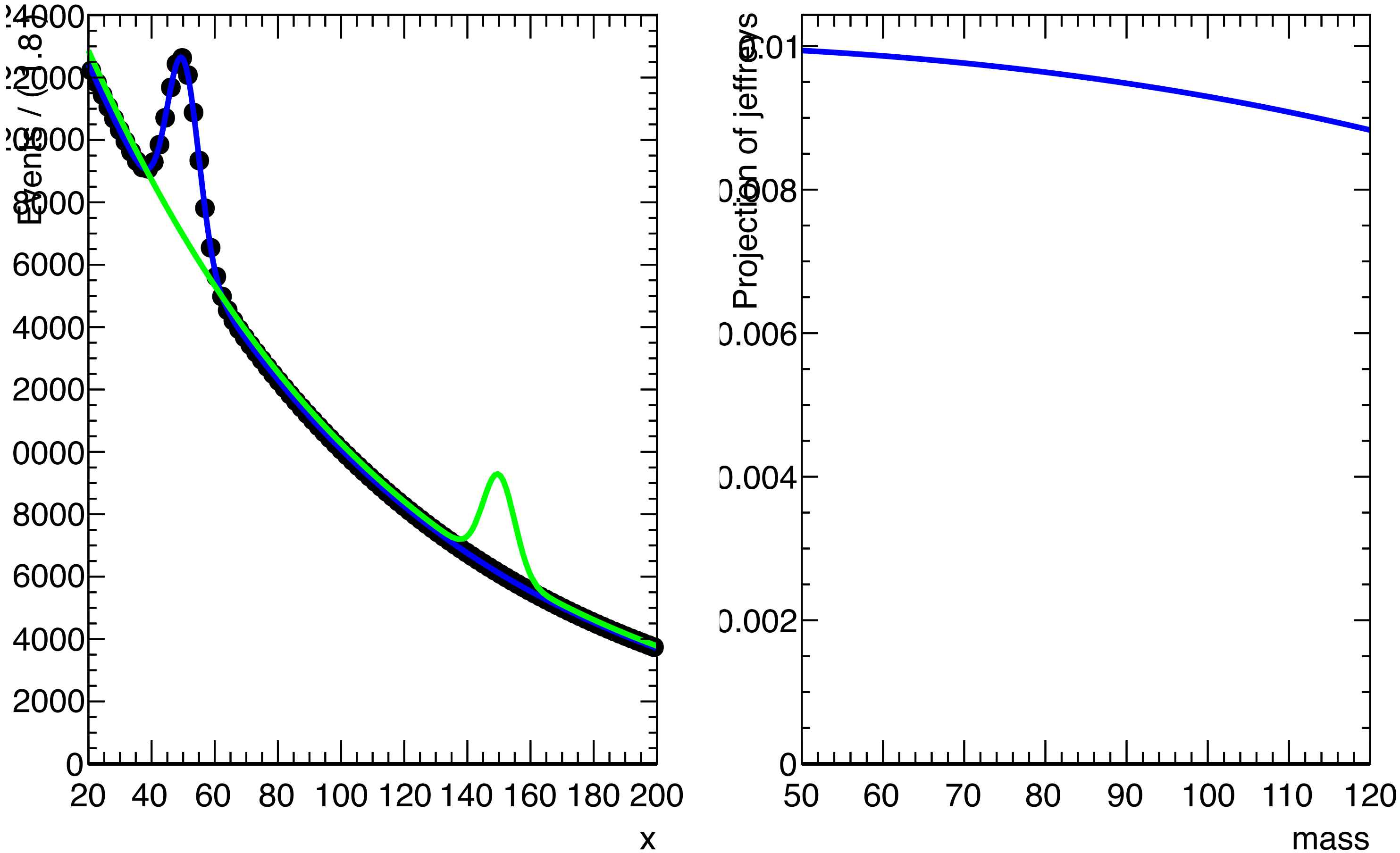





Events / (18)
8000
6000
4000
2000
0000
8000
6000
4000
2000
0
20 40 60 80 100 120 140 160 180 200
x
Projection of jeffreys
0.004
.0035
0.003
.0025
0.002
.0015
0.001
.0005
0
50 60 70 80 90 100 110 120
mass


20 40 60 80 100 120 140 160 180 200
x
Projection of jeffreys
0.001
0
50 60 70 80 90 100 110 120
mass


Events / (1.8)
5000
0000
5000
0
20
40
60
80
100
120
140
160
180
200
x
Projection of jeffreys
0.01
0.008
0.006
0.004
0.002
0
50
60
70
80
90
100
110
120
mass


0
2000
4000
6000
8000
0000
2000
4000
6000
8000
0000
2000
4000
20
40
60
80
100
120
140
160
180
200
x
Projection of jeffreys
0
0.002
0.004
0.006
0.008
0.01
50
60
70
80
90
100
110
120
mass

Reference Priors

Refrerence priors are another type of “objective” priors, that try to save Jeffreys’ basic idea.

Noninformative priors have been studied for a long time and most of them have been found defective in more than one way. Reference analysis arose from this study as the only *general* method that produces priors that have the required *invariance* properties, deal successfully with the *marginalization* paradoxes, and have consistent *sampling* properties.

Ideally, such a method should be very general, applicable to all kinds of measurements regardless of the number and type of parameters and data involved. It should make use of *all* available information, and coherently so, in the sense that if there is more than one way to extract all relevant information from data, the final result will not depend on the chosen way. The desiderata of generality, exhaustiveness and coherence are satisfied by Bayesian procedures, but that of objectivity is more problematic due to the Bayesian requirement of specifying prior probabilities in terms of degrees of belief. Reference analysis[2], an objective Bayesian method developed over the past twenty-five years, solves this problem by replacing the question “what is our prior degree of belief?” by “what would our posterior degree of belief be, if our prior knowledge had a minimal effect, relative to the data, on the final inference?”

See Luc Demortier’s PhyStat 2005 proceedings

http://physics.rockefeller.edu/luc/proceedings/phystat2005_refana.ps

The Bayesian Solution

Bayesian solution generically have a prior for the parameters of interest as well as nuisance parameters

- 2010 recommendations largely echoes the PDG's stance.

Recommendation: When performing a Bayesian analysis one should separate the objective likelihood function from the prior distributions to the extent possible.

Recommendation: When performing a Bayesian analysis one should investigate the sensitivity of the result to the choice of priors.

Warning: Flat priors in high dimensions can lead to unexpected and/or misleading results.

Recommendation: When performing a Bayesian analysis for a single parameter of interest, one should attempt to include Jeffreys's prior in the sensitivity analysis.

Words of wisdom on Bayesian methods

To support the points raised above, here are some quotes from professional statisticians (taken from selected PhyStat talks and selections from Bob Cousins lectures):

- "Perhaps the most important general lesson is that the facile use of what appear to be uninformative priors is a dangerous practice in high dimensions." – Brad Effron
- "meaningful prior specification of beliefs in probabilistic form over very large possibility spaces is very difficult and may lead to a lot of arbitrariness in the specification." – Michael Goldstein
- "Sensitivity Analysis is at the heart of scientific Bayesianism." – Michael Goldstein
- "Non-subjective Bayesian analysis is just a part – an important part, I believe of a healthy sensitivity analysis to the prior choice..." J.M. Bernardo
- "Objective Bayesian analysis is the best frequentist tool around" – Jim Berger

Coverage & Likelihood principle

Methods based on the Neyman-Construction always cover.... by construction.

- this approach violates the likelihood principle

Bayesian methods obey likelihood principle, but do not necessarily cover

- that doesn't mean Bayesians shouldn't care about coverage

Coverage can be thought of as a **calibration of our statistical apparatus**. [explain under-/over-coverage]

What should be the view today; Objective Bayesian analysis is the best frequentist tool around. -Jim Berger

Bayesian and Frequentist results answer different questions

- major differences between them may indicate severe coverage problems and/or violations of the likelihood principle

Joke

"Bayesians address the question everyone is interested in, by using assumptions no-one believes"

"Frequentists use impeccable logic to deal with an issue of no interest to anyone"

-L. Lyons

The End

Thank You!

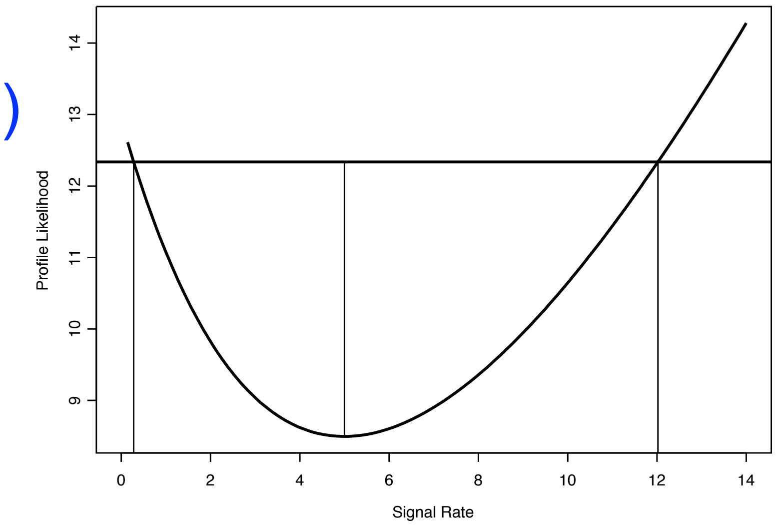
Supplemental Slides

Profile Likelihood Ratio & MINUIT

CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS

Rolke, Lopez, Conrad published a method based on the profile likelihood ratio (NIM A551) before the term was used much in HEP

- noticed identical results with MINOS limits, extensive numerical tests



MINUIT long writeup explains algorithm

- limits based on extreme values of the contour
- algorithm does not sound much like the profile likelihood ratio,

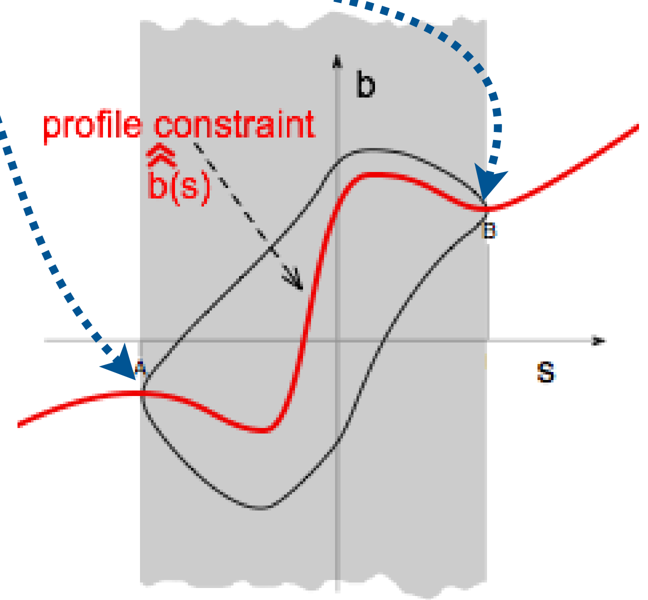


Figure 7.2: MINOS error confidence region for parameter 1

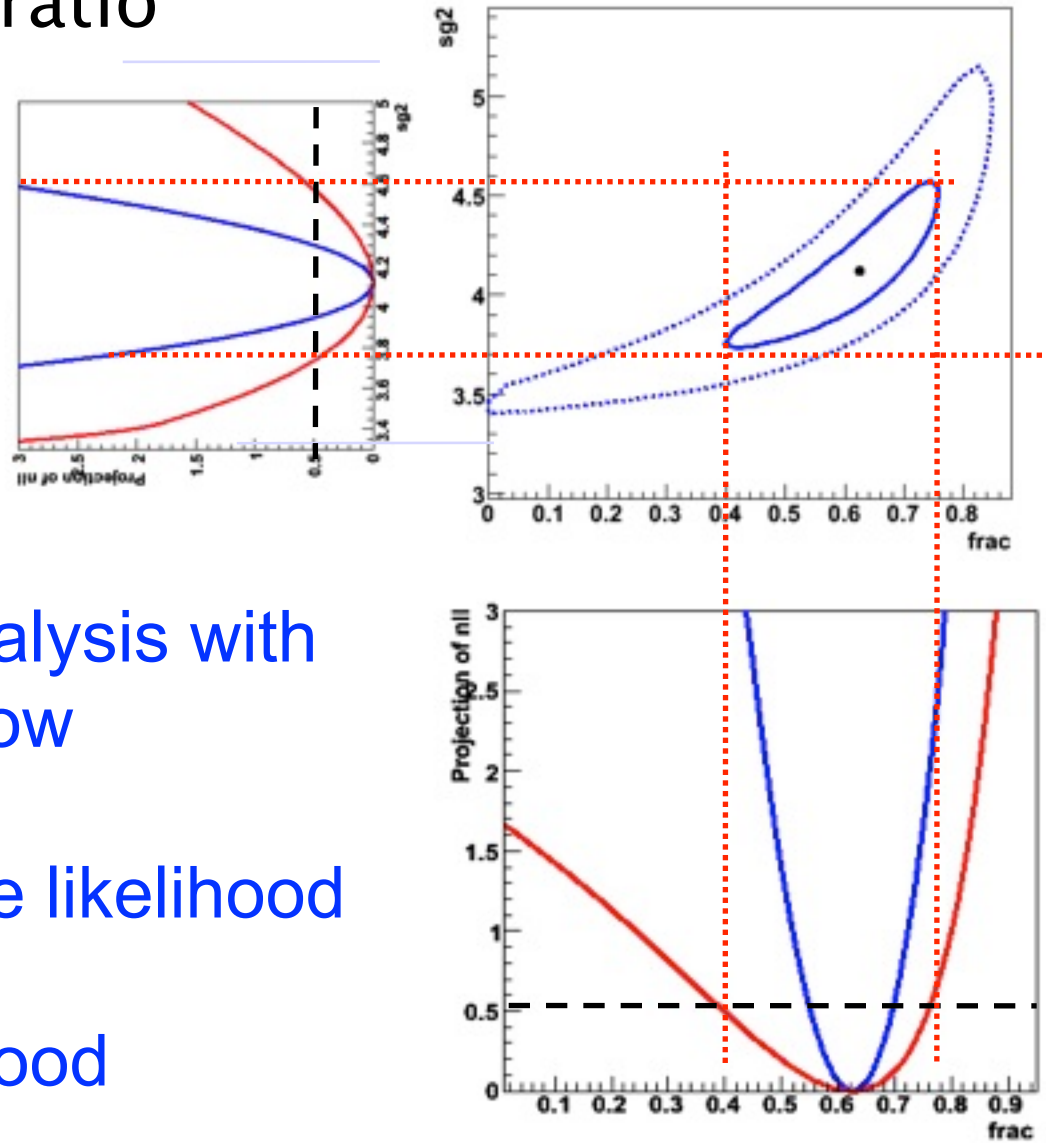
But it's not hard to show extreme points must lie on profile constraint and lie on same likelihood contour

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

The Profile Likelihood Ratio in RooFit/RooStats

An early request from RooStats to RooFit was to provide a profile likelihood ratio

```
root [0] RooAbsPdf* pdf = ...;
root [1] RooRealVar* parameter = ...;
root [2] RooAbsData* data = ...;

root [3] nll = pdf->createNLL(*data)
root [4] profile = nll->createProfile(*parameter)
root [5] frame = parameter->frame()
root [6] profile->plotOn(frame)
root [7] frame->Draw()
```



- Very easy to perform an analysis with the profile likelihood ratio now

- MINOS error box and profile likelihood give same error
for multi-dimensional likelihood

Taken from Wouter Verkerke, NIKHEF

Decision Theory

One of the deficiencies of the Neyman-Pearson approach is that one must specify the size of the test α

- But where does α come from?
  - is it purely conventional or is there a reason?

A great deal of literature related to statistics (and economics, etc.) is devoted to making **decisions**.

- need to consider **Utility** or **Risk** of different outcomes

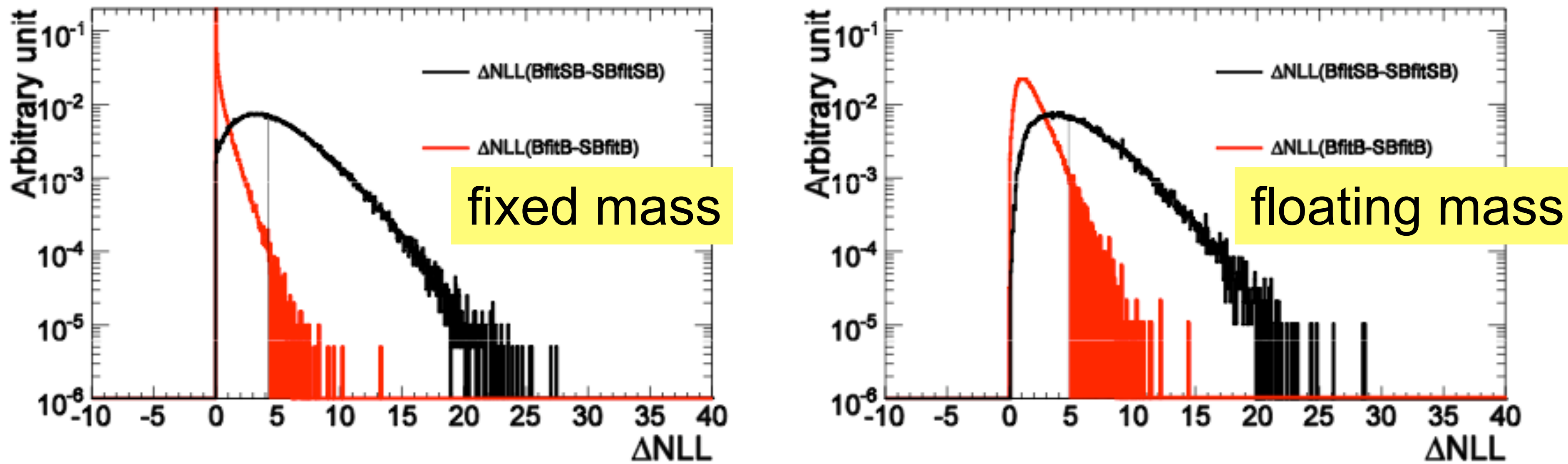
In the context of decision and utility theory there can be a justification, but this is rarely done in particle physics

Floating mass & look-elsewhere effect

In the floating mass case, it is clear that there is a degradation in significance due to the look-elsewhere effect (aka “trials factor”)

- naive estimate of factor is Range/(mass resolution)

Formally, the conditions required for Wilks’s theorem do not hold because floating mass parameter makes no sense in a background-only model. See a Higgs example below.



The effect depends on range that the fit considers (non-local): eg. a 120 GeV Higgs pays price for considering 1TeV

For another example, see L. Demortier, p-vaues: http://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf

Decision Theory

From Fred James lectures

DECISION THEORY [GREATLY SIMPLIFIED] (6)

EXAMPLE DECISION: WHETHER OR NOT TO TAKE AN UMBRELLA TO WORK TOMORROW

OBSERVABLE SPACE Θ : $\begin{cases} R = \text{it rains tomorrow} \\ \bar{R} = \text{no rain} \end{cases}$

DECISION SPACE $\mathcal{D}$: $\begin{cases} u = \text{take umbrella} \\ \bar{u} = \text{do not take it} \end{cases}$

LOSS FUNCTION $\mathcal{L}(\mathcal{D}, \Theta)$:

	$\bar{R}$	R
u	1	1
$\bar{u}$	0	3

DECISION RULE:

1. BAYESIAN DECISION RULE: MINIMIZE EXPECTED LOSS

$E(\mathcal{L})_u = 1 \cdot P(\bar{R}) + 1 \cdot P(R) = 1$

$E(\mathcal{L})_{\bar{u}} = 0 \cdot P(\bar{R}) + 3 \cdot P(R) = 3 \cdot P(R)$

$\Rightarrow$ TAKE UMBRELLA IF $P(R) > 1/3$

2. MINIMAX DECISION RULE: MINIMIZE MAXIMUM LOSS

$\Rightarrow$ ALWAYS TAKE UMBRELLA

Decisions: Bayesian & Frequentist

Center for Cosmology and Particle Physics

F. James, Ch. 6

Structure of $P(x|H_0)$ & $P(x|H_1)$ puts limits on allowable ranges of alpha, beta

- Bayesians want to minimize expected risk based on priors and risk/utility of outcomes

Frequentists don't have priors to work with, so they only have risk/utility in two situations

- "minimax" approach aims to minimize maximum risk
  - most conservative
  - paranoid for games against nature

Frequentist choice of α interpreted in Bayesian framework implies this ratio:

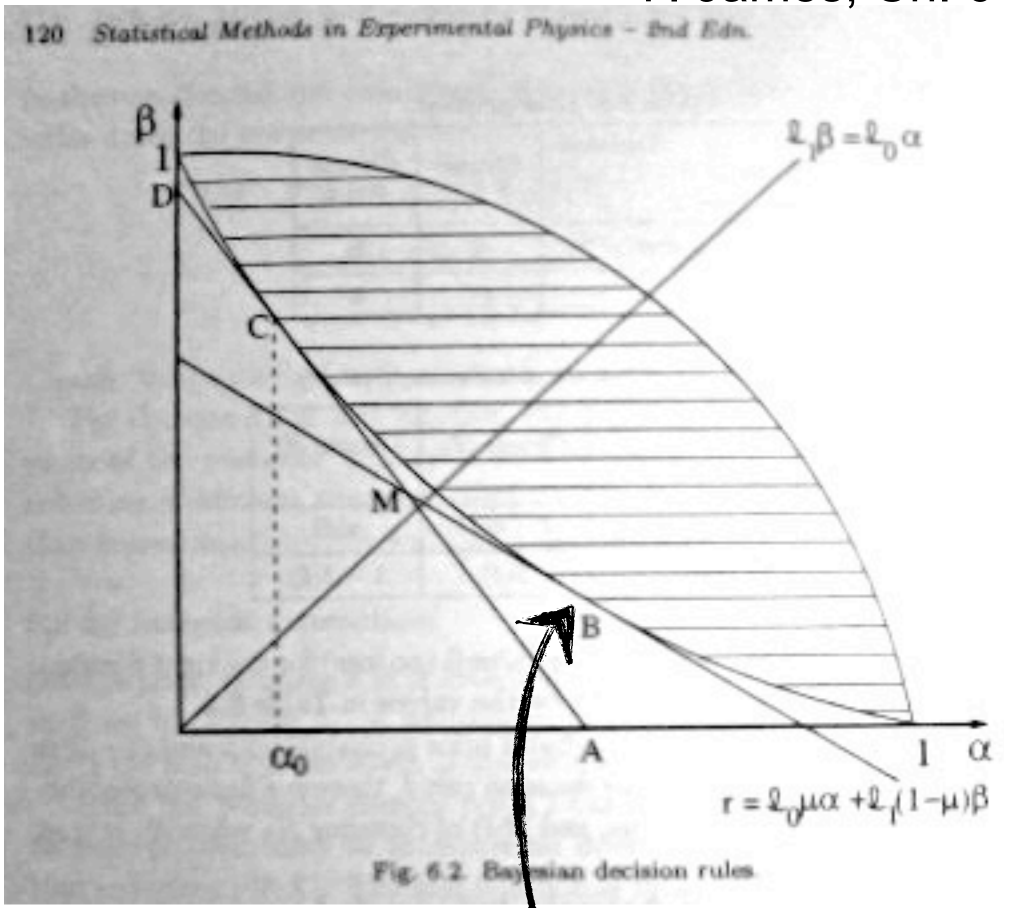


Fig. 6.2 Bayesian decision rules

$$l_1(1-\mu)P(X|H_1) < l_0\mu P(X|H_0)$$

$$\frac{P(X|H_1)}{P(X|H_0)} < \frac{l_0\mu}{l_1(1-\mu)}$$

$$\frac{OD}{OA} = \frac{l_0\mu}{l_1(1-\mu)}$$

Type III Systematics

Type III Systematics are related to variations in inference from uncertainty in the overall theoretical framework

- Bayesian approach: assign priors over the "framework space"
- Sinervo suggests Frequentist can't incorporate them because one cannot find an ensemble associated to the theories
  - but theoretical framework can be thought of as an additional nuisance parameter (possibly discrete) - can be incorporated!
  - only need an ensemble of some observable if one wants to ***constrain*** the space of the theories, not to incorporate them
  - if theoretical framework influences our experimental result, then we don't really know what we are doing!

Taken from Cousins' Phystat05 talk:

- A.W.F. Edwards (in Kalbfleisch 1970): "Let me say at once that I can see no reason why it should always be possible to eliminate nuisance parameters. Indeed, one of the many objections to Bayesian inference is that is always permits this elimination."