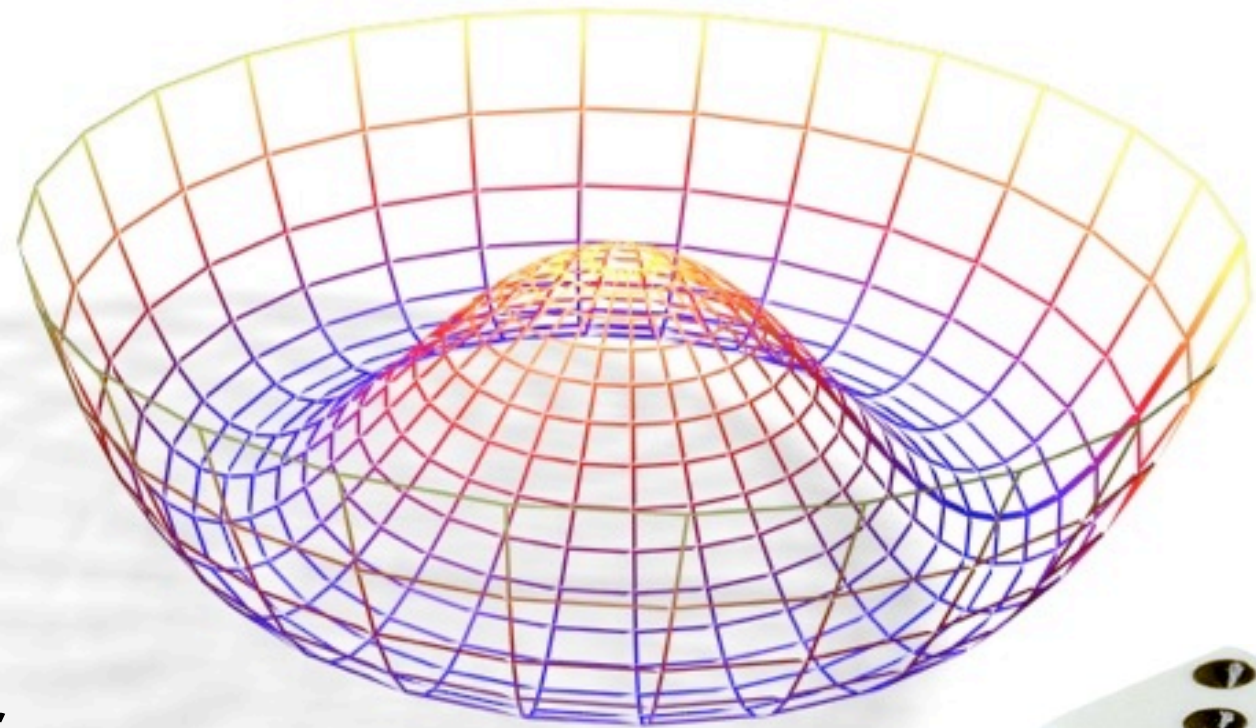




Practical Statistics for Particle Physics

Kyle Cranmer,
New York University



Lecture 1: Preliminaries

- ▶ Probability Density Function vs. Likelihood
- ▶ Monte Carlo
- ▶ Point estimates and maximum likelihood estimators

Lecture 2: Building a probability model

- ▶ A generic template for high energy physics
- ▶ Examples of different “narratives”

Lecture 3: Hypothesis testing

- ▶ The Neyman-Pearson lemma and the likelihood ratio
- ▶ Composite models and the profile likelihood ratio
- ▶ Review of ingredients for a hypothesis test

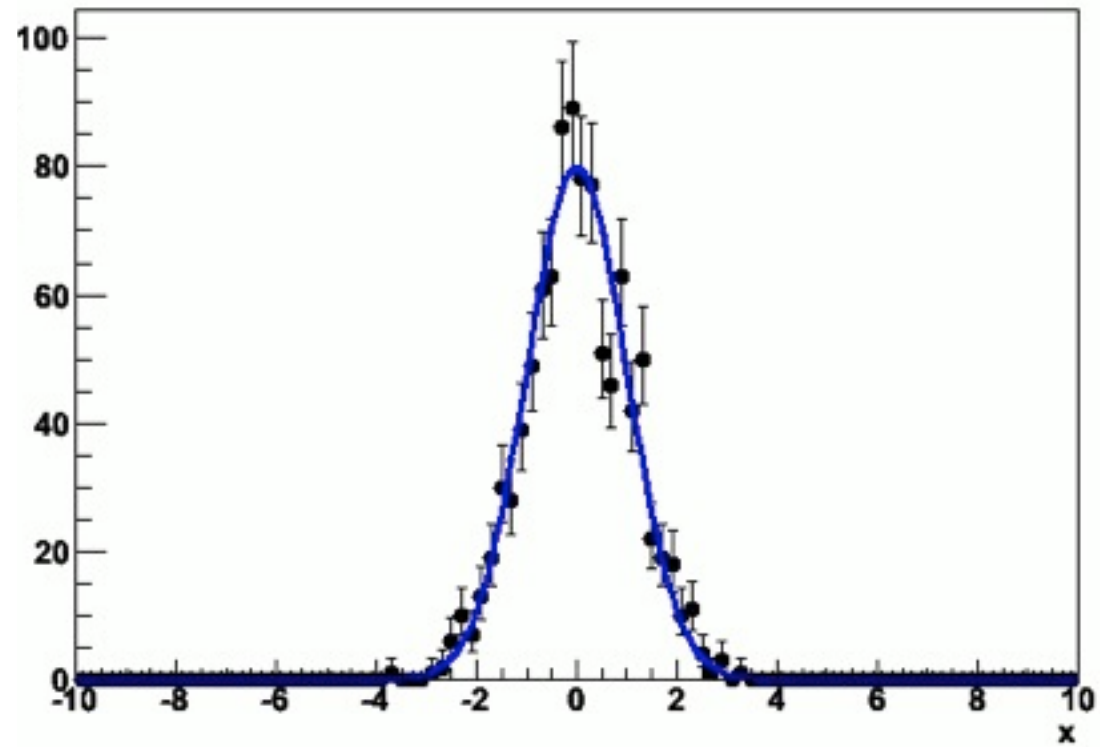
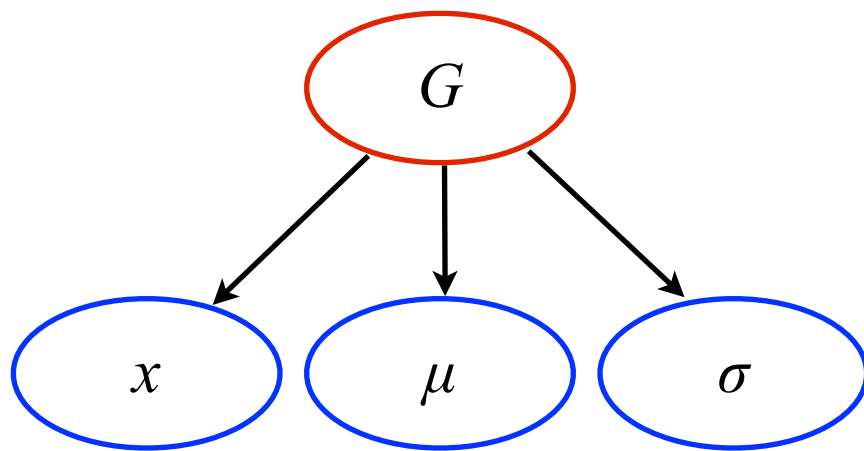
Lecture 4: Limits & Confidence Intervals

- ▶ The meaning of confidence intervals as inverted hypothesis tests
- ▶ Asymptotic properties of likelihood ratios
- ▶ Bayesian approach

A General Purpose Statistical Model (beginning of lecture 3)

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

- ▶ eg. a Gaussian $G(x|\mu, \sigma)$ is parametrized by (μ, σ)
- ▶ every node is a real-valued function of the nodes below



Channel: a subset of the data defined by some selection requirements.

- ▶ eg. all events with 4 electrons with energy > 10 GeV
- ▶ n : number of events observed in the channel
- ▶ ν : number of events expected in the channel

Discriminating variable: a property of those events that can be measured and which helps discriminate the signal from background

- ▶ eg. the invariant mass of two particles
- ▶ $f(x)$: the p.d.f. of the discriminating variable x

$$\mathcal{D} = \{x_1, \dots, x_n\}$$

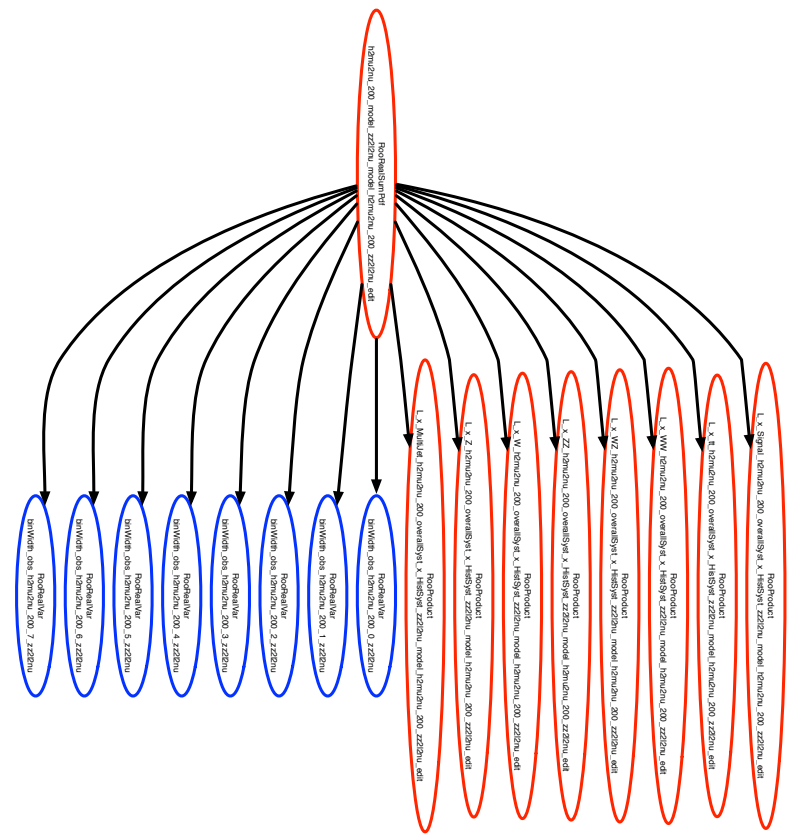
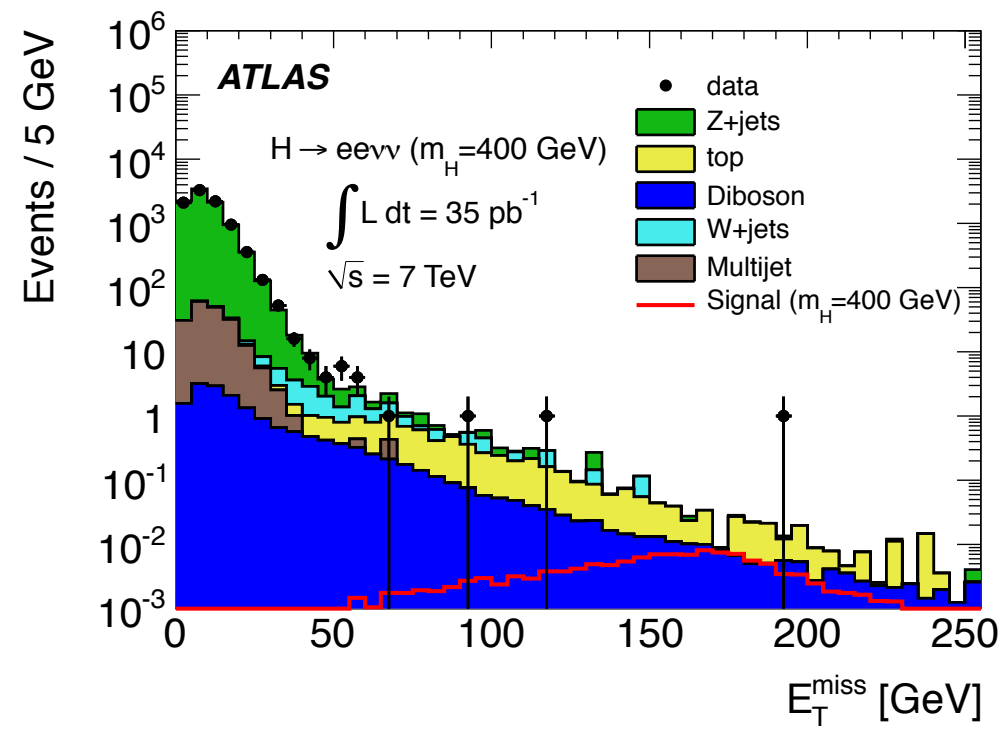
Marked Poisson Process:

$$\mathbf{f}(\mathcal{D}|\nu) = \text{Pois}(n|\nu) \prod_{e=1}^n f(x_e)$$

Sample: a sample of simulated events corresponding to particular type interaction that populates the channel.

▶ statisticians call this a mixture model

$$f(x) = \frac{1}{\nu_{\text{tot}}} \sum_{s \in \text{samples}} \nu_s f_s(x), \quad \nu_{\text{tot}} = \sum_{s \in \text{samples}} \nu_s$$



Parameters of interest (μ): parameters of the theory that modify the rates and shapes of the distributions, eg.

- ▶ the mass of a hypothesized particle
- ▶ the “signal strength” $\mu=0$ no signal, $\mu=1$ predicted signal rate

Nuisance parameters (θ or α_p): associated to uncertainty in:

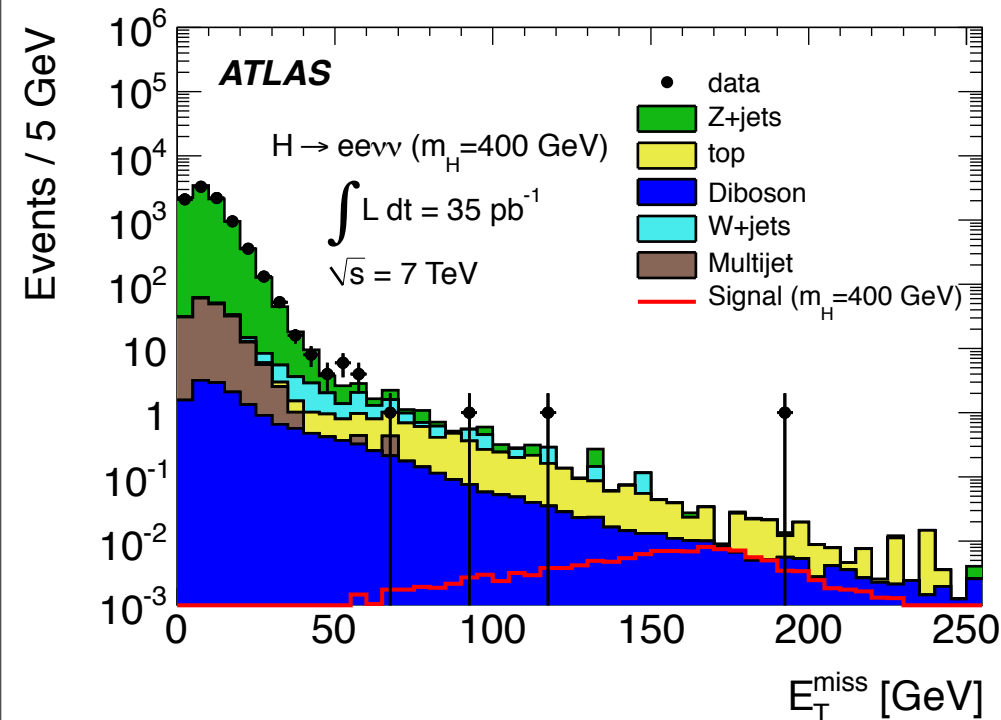
- ▶ response of the detector (calibration)
- ▶ phenomenological model of interaction in non-perturbative regime

Lead to a parametrized model: $\nu \rightarrow \nu(\alpha), f(x) \rightarrow f(x|\alpha)$

$$\mathbf{f}(\mathcal{D}|\alpha) = \text{Pois}(n|\nu(\alpha)) \prod_{e=1}^n f(x_e|\alpha)$$

Tabulate effect of individual variations of sources of systematic uncertainty

- typically one at a time evaluated at nominal and “ $\pm 1 \sigma$ ”
- use some form of interpolation to parametrize p^{th} variation in terms of **nuisance parameter** α_p

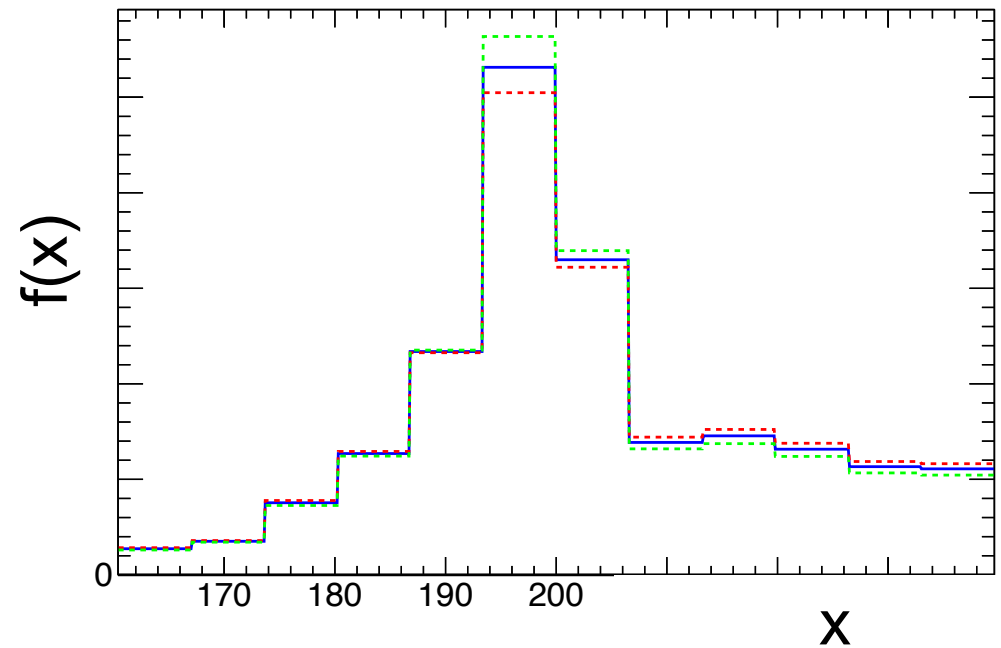
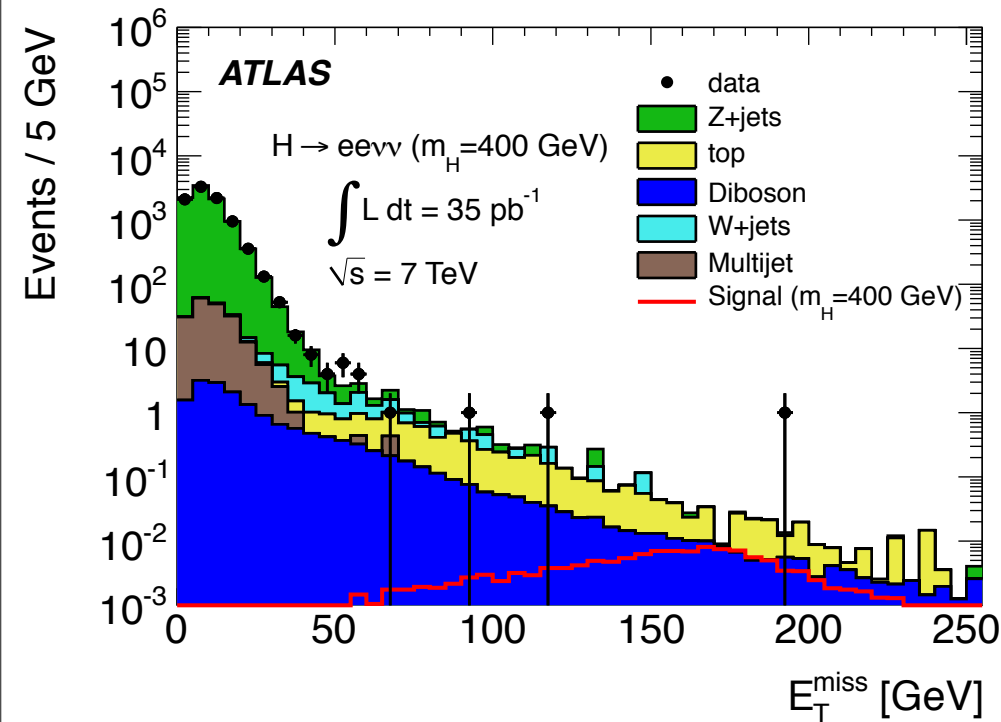


	Z+jets	top	Diboson	...
syst 1				
syst 2				
...				

$$f(\mathcal{D}|\alpha) = \text{Pois}(n|\nu(\alpha)) \prod_{e=1}^n f(x_e|\alpha)$$

Tabulate effect of individual variations of sources of systematic uncertainty

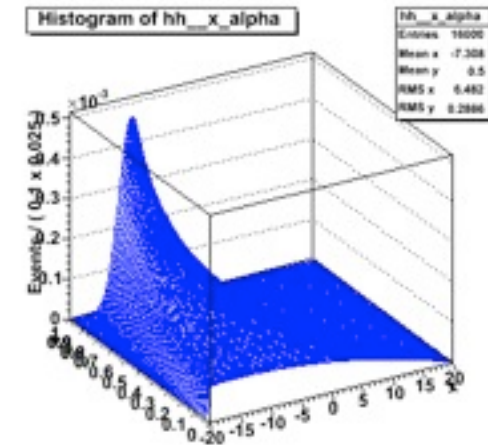
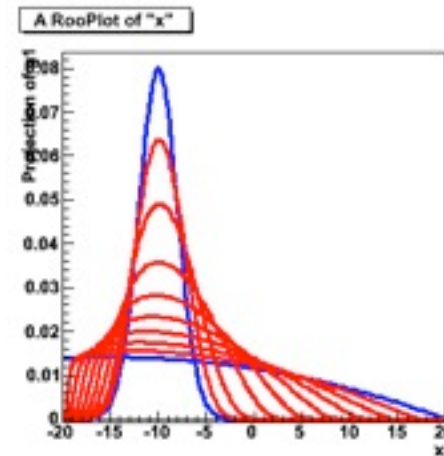
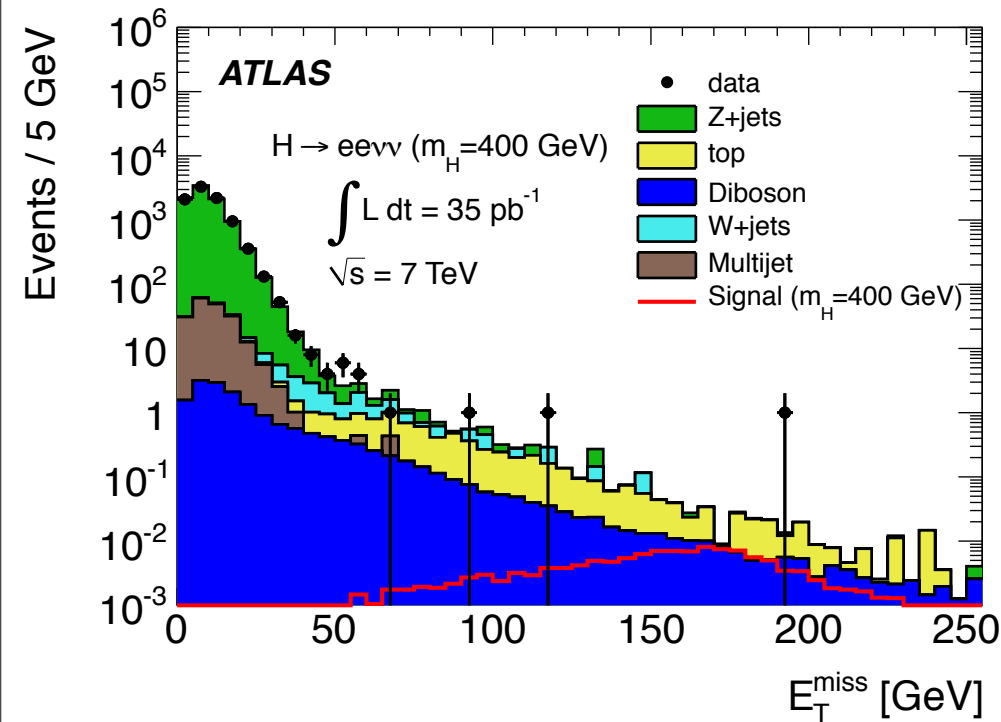
- typically one at a time evaluated at nominal and “ $\pm 1 \sigma$ ”
- use some form of interpolation to parametrize p^{th} variation in terms of **nuisance parameter** α_p



$$f(\mathcal{D}|\alpha) = \text{Pois}(n|\nu(\alpha)) \prod_{e=1}^n f(x_e|\alpha)$$

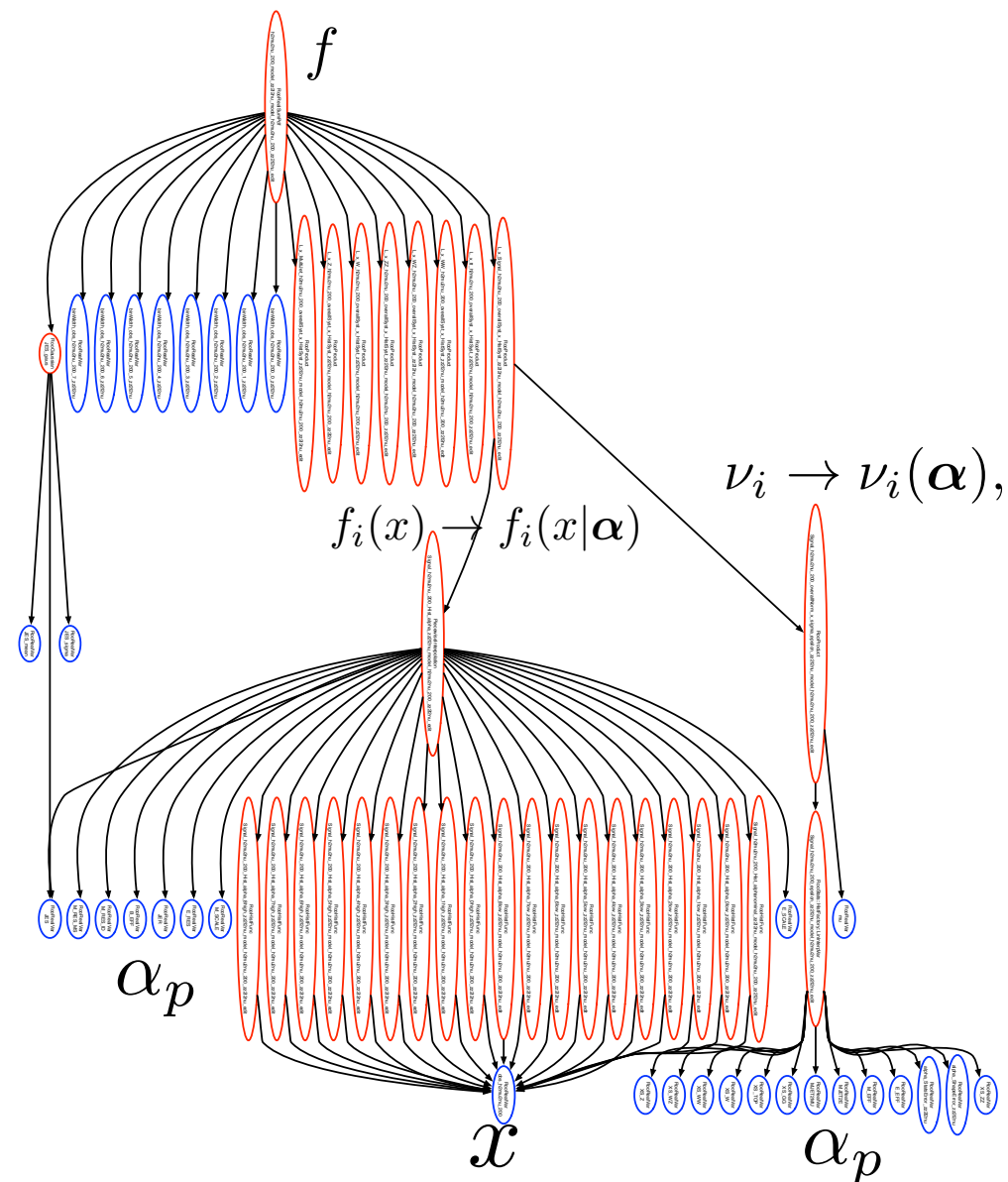
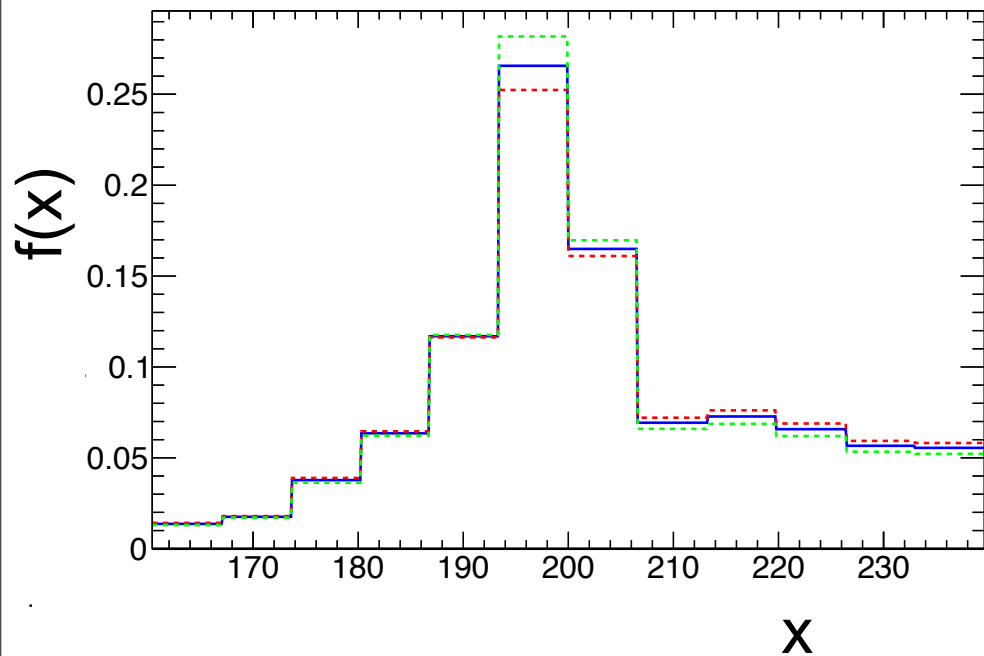
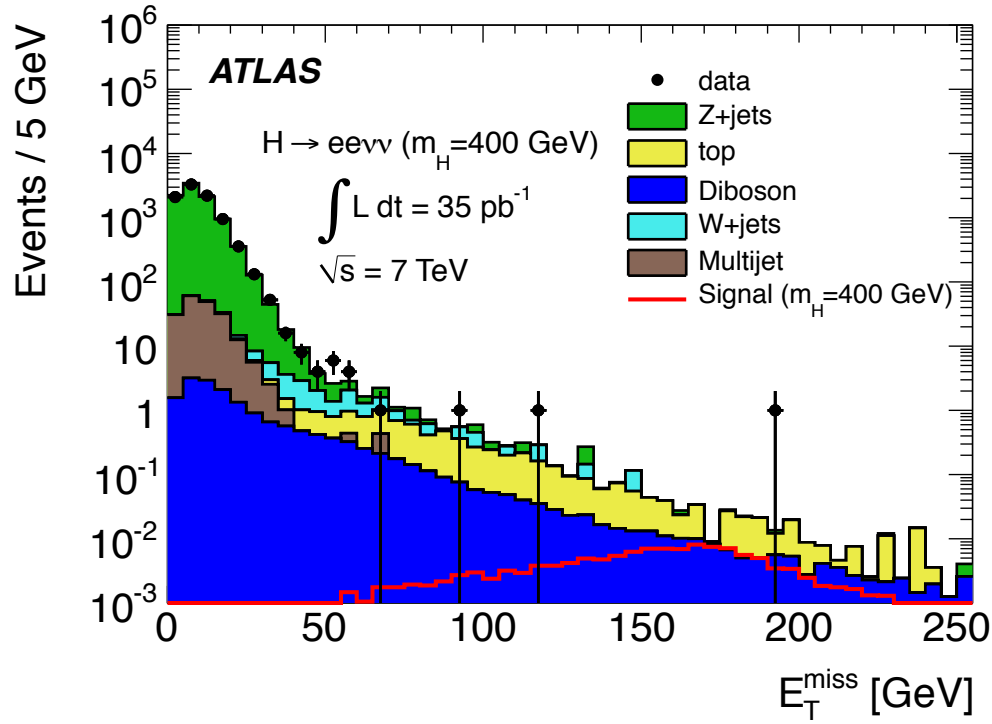
Tabulate effect of individual variations of sources of systematic uncertainty

- typically one at a time evaluated at nominal and “ $\pm 1 \sigma$ ”
- use some form of interpolation to parametrize p^{th} variation in terms of **nuisance parameter** α_p

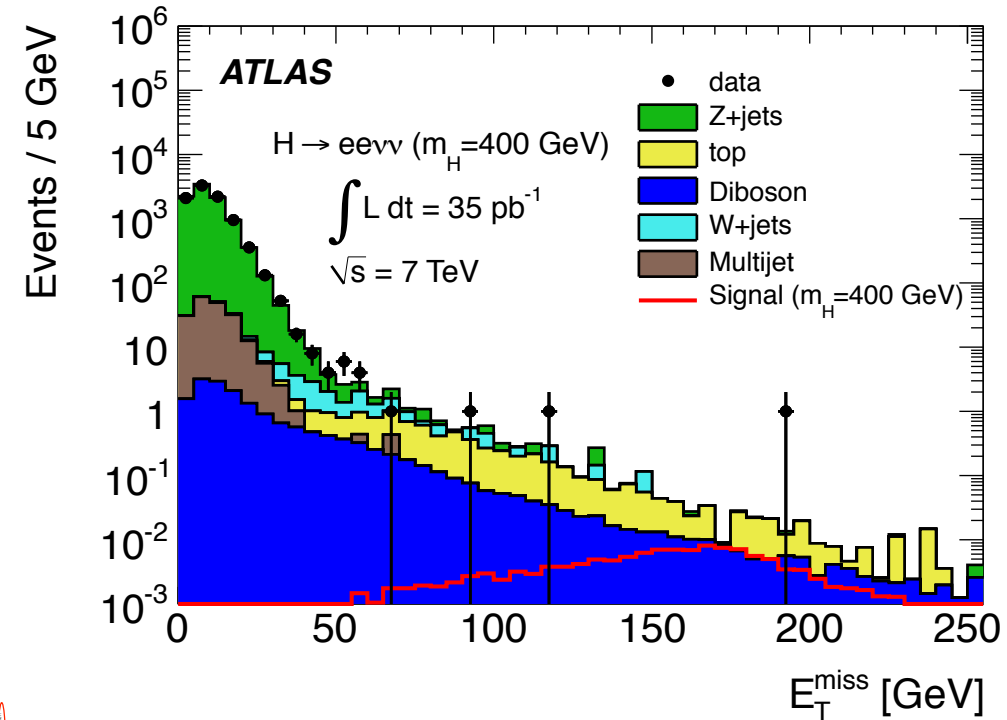
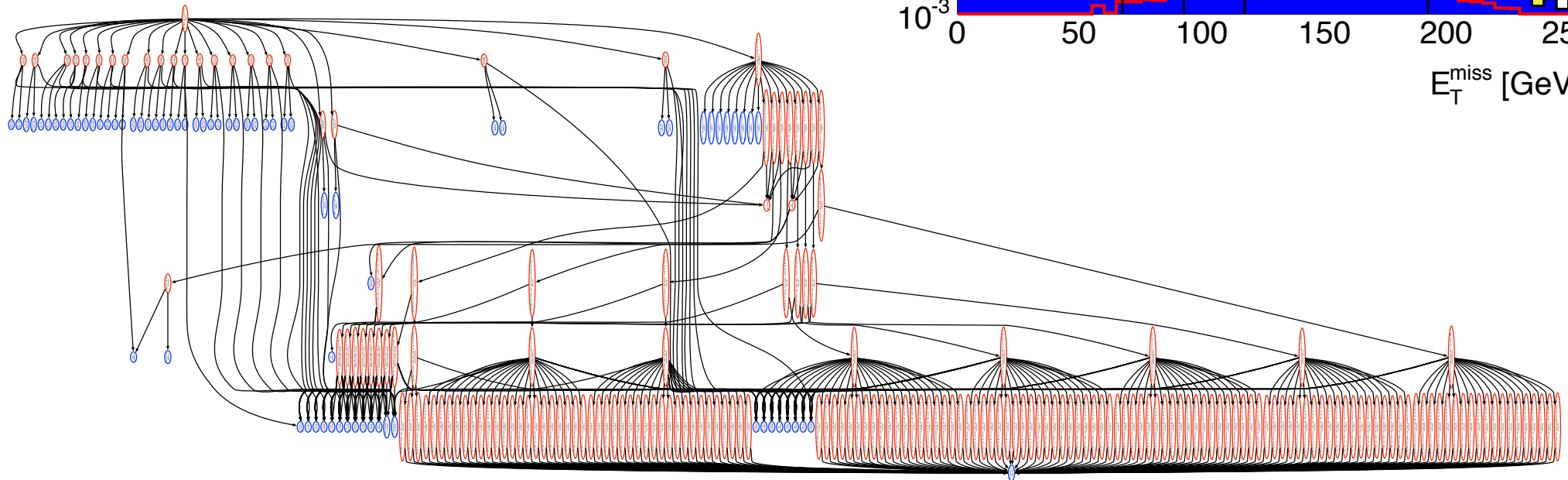


$$f(\mathcal{D}|\alpha) = \text{Pois}(n|\nu(\alpha)) \prod_{e=1}^n f(x_e|\alpha)$$

Visualizing the model for one channel



After parametrizing each component of the mixture model, the pdf for a single channel might look like this



Simultaneous Multi-Channel Model: Several disjoint regions of the data are modeled simultaneously. Identification of common parameters across many channels requires coordination between groups such that meaning of the parameters are really the same.

$$\mathbf{f}_{\text{sim}}(\mathcal{D}_{\text{sim}}|\boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c | \nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce} | \boldsymbol{\alpha}) \right]$$

where $\mathcal{D}_{\text{sim}} = \{\mathcal{D}_1, \dots, \mathcal{D}_{c_{\text{max}}}\}$

Control Regions: Some channels are not populated by signal processes, but are used to constrain the nuisance parameters

- ▶ attempt to describe systematics in a statistical language
- ▶ Prototypical Example: “on/off” problem with unknown ν_b

$$\mathbf{f}(n, m | \mu, \nu_b) = \underbrace{\text{Pois}(n | \mu + \nu_b)}_{\text{signal region}} \cdot \underbrace{\text{Pois}(m | \tau \nu_b)}_{\text{control region}}$$

Often detailed statistical model for auxiliary measurements that measure certain nuisance parameters are not available.

- ▶ one typically has MLE for α_p , denoted a_p and standard error

Constraint Terms: are idealized pdfs for the MLE.

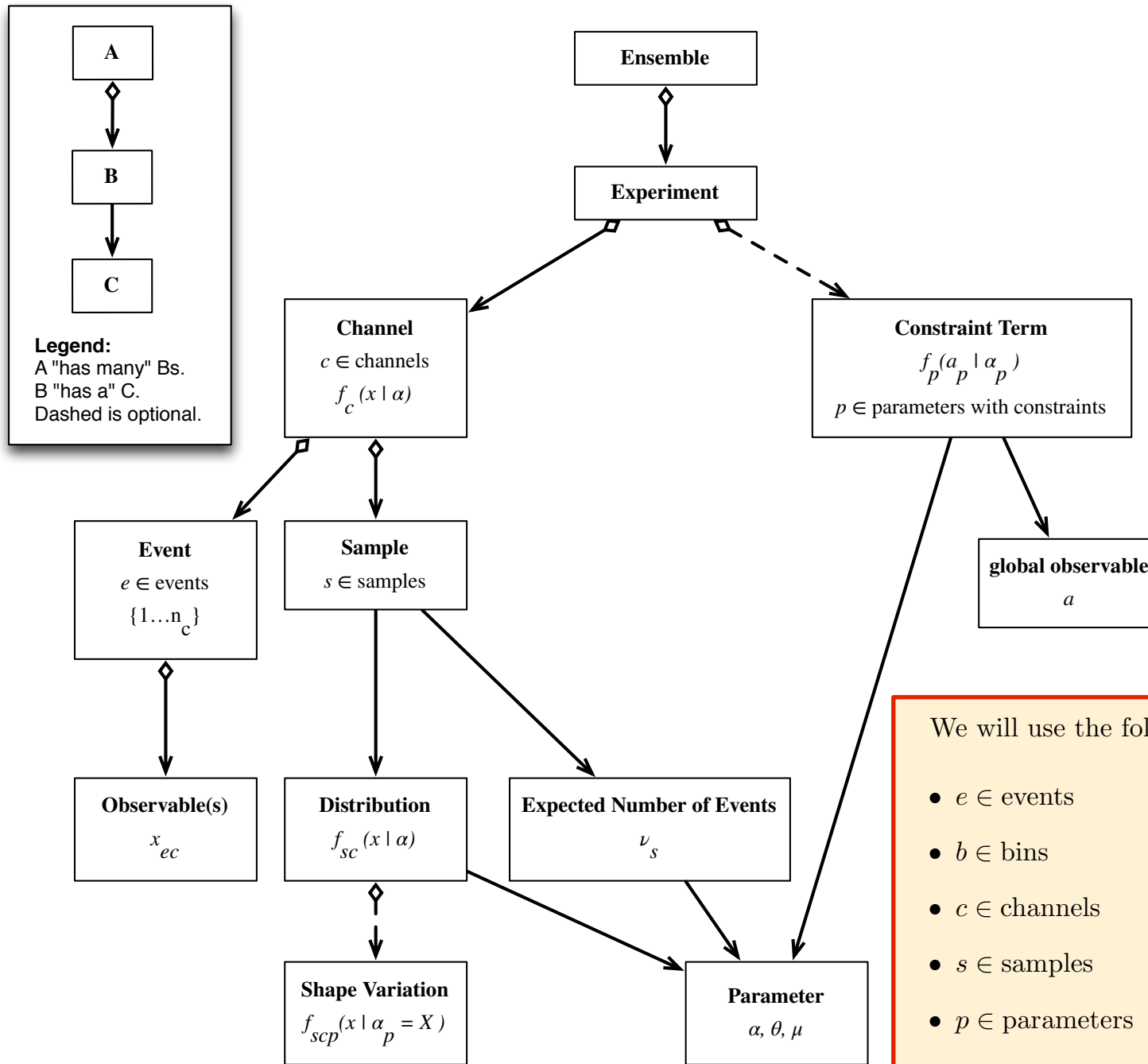
$$f_p(a_p|\alpha_p) \text{ for } p \in \mathbb{S}$$

- ▶ common choices are Gaussian, Poisson, and log-normal
- ▶ New: careful to write constraint term a frequentist way
- ▶ Previously: $\pi(\alpha_p|a_p) = f_p(a_p|\alpha_p)\eta(\alpha_p)$ with uniform η

Simultaneous Multi-Channel Model with constraints:

$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p|\alpha_p)$$

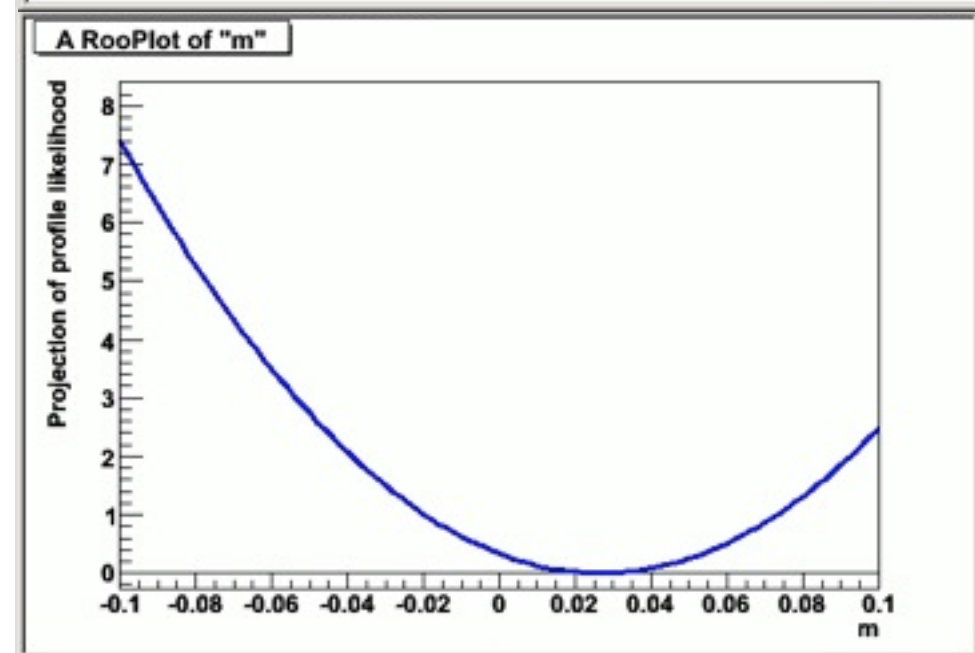
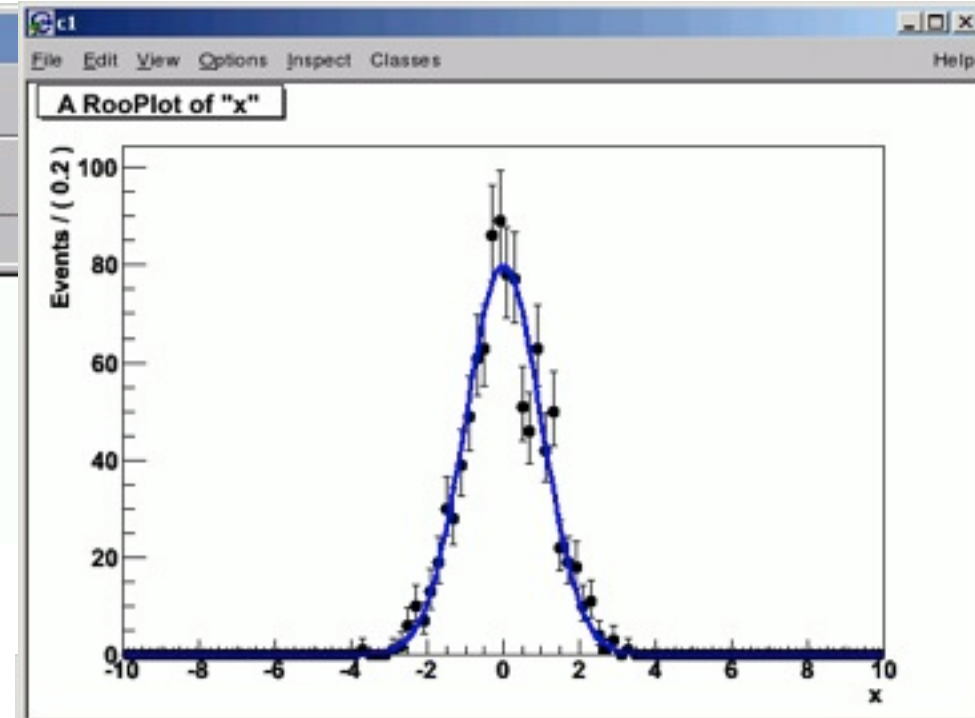
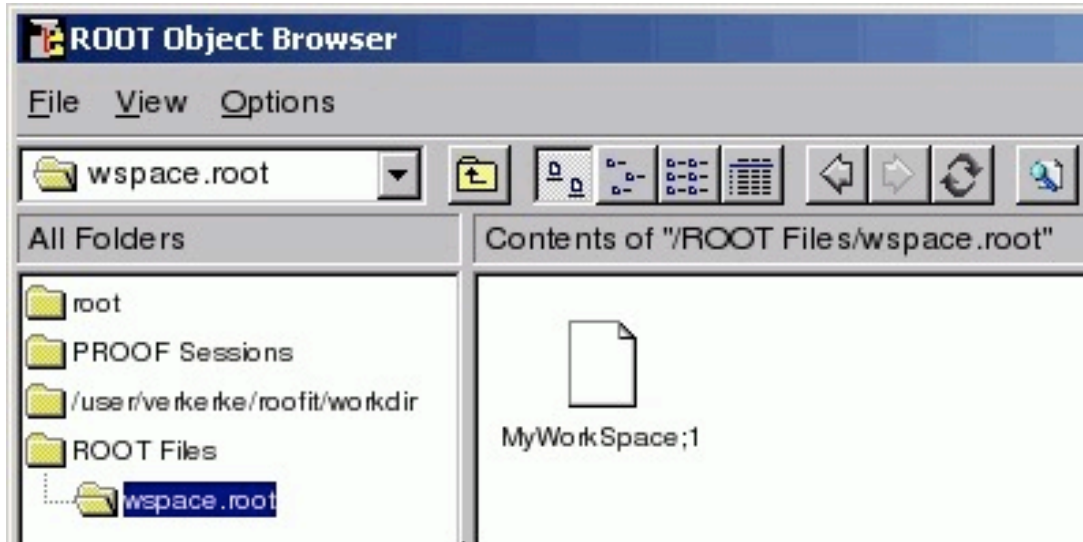
where $\mathcal{D}_{\text{sim}} = \{\mathcal{D}_1, \dots, \mathcal{D}_{c_{\text{max}}}\}$, $\mathcal{G} = \{a_p\}$ for $p \in \mathbb{S}$



We will use the following mnemonic index conventions:

- $e \in \text{events}$
- $b \in \text{bins}$
- $c \in \text{channels}$
- $s \in \text{samples}$
- $p \in \text{parameters}$


Example of Digital Publishing



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.

<http://cds.cern.ch/record/1456844>

Information	Discussion (0)	Files	Linkbacks
 Preprint			
Report number	CERN-OPEN-2012-016		
Title	HistFactory: A tool for creating statistical models for use with RooFit and RooStats		
Author(s)	Cranmer, Kyle (New York U.) ; Lewis, George (New York U.) ; Moneta, Lorenzo (CERN) ; Shibata, Akira (New York U.) ; Verkerke, Wouter (NIKHEF, Amsterdam)		
Collaboration	ROOT Collaboration		
Abstract	<p>The HistFactory is a tool to build parametrized probability density functions (pdfs) in the RooFit/RooStats framework based based on simple ROOT histograms organized in an XML file. The pdf has a restricted form, but it is sufficiently flexible to describe many analyses based on template histograms. The tool takes a modular approach to build complex pdfs from more primitive conceptual building blocks. The resulting PDF is stored in a RooWorkspace which can be saved to and read from a ROOT file. This document describes the defaults and interface in HistFactory 5.32.</p>		

32 page documentation of HistFactory tool + manual

▶ currently a “living document”

State of the art: At the time of the discovery, the combined Higgs search included 100 disjoint channels and >500 nuisance parameters

- ▶ Models for individual channels come from about 11 sub-groups performing dedicated searches for specific Higgs decay modes
- ▶ In addition low-level performance groups provide tools for evaluating systematic effects and corresponding constraint terms

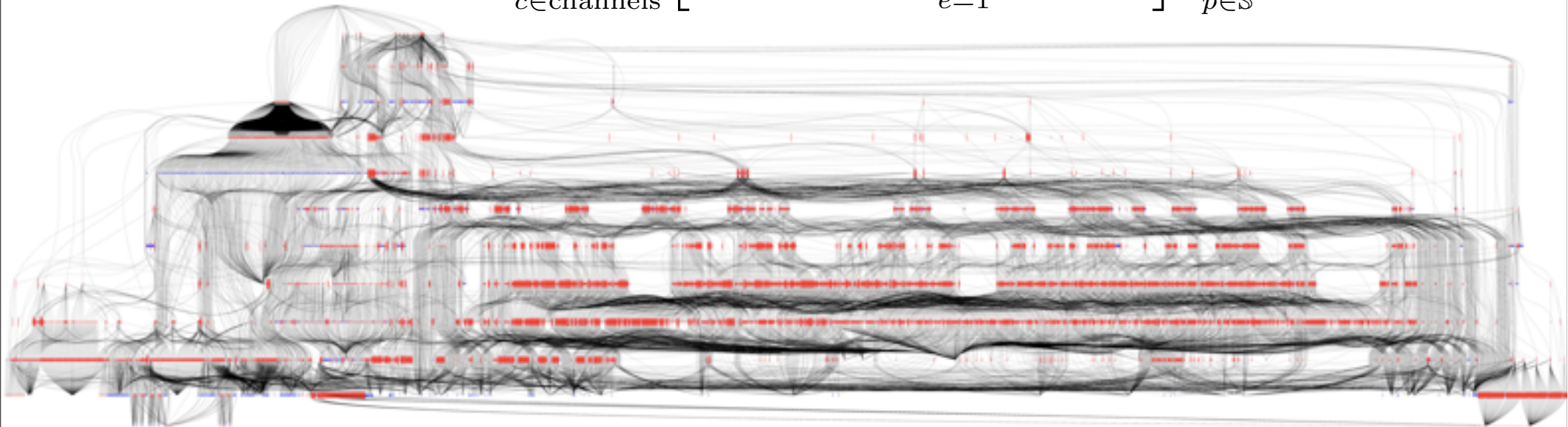
Higgs Decay	Subsequent Decay	Additional Sub-Channels	m_H Range	L [fb ⁻¹]
$H \rightarrow \gamma\gamma$	–	9 sub-channels ($p_{T_i} \otimes \eta_\gamma \otimes$ conversion)	110-150	4.9
$H \rightarrow ZZ$	$lll'l'$	$\{4e, 2e2\mu, 2\mu2e, 4\mu\}$	110-600	4.8
	$ll\nu\nu$	$\{ee, \mu\mu\} \otimes \{\text{low pile-up, high pile-up}\}$	200-280-600	4.7
	$llqq$	$\{b\text{-tagged, untagged}\}$	200-300-600	4.7
$H \rightarrow WW$	$lvlv$	$\{ee, e\mu, \mu\mu\} \otimes \{0\text{-jet, 1-jet, VBF}\}$	110-300-600	4.7
	$lvqq'$	$\{e, \mu\} \otimes \{0\text{-jet, 1-jet}\}$	300-600	4.7
$H \rightarrow \tau^+\tau^-$	$ll4\nu$	$\{e\mu\} \otimes \{0\text{-jet}\} \oplus \{1\text{-jet, VBF, VH}\}$	110-150	4.7
	$l\tau_{\text{had}}3\nu$	$\{e, \mu\} \otimes \{0\text{-jet}\} \otimes \{E_T^{\text{miss}} \geq 20 \text{ GeV}\} \oplus \{e, \mu\} \otimes \{1\text{-jet, VBF}\}$	110-150	4.7
	$\tau_{\text{had}}\tau_{\text{had}}2\nu$	$\{1\text{-jet}\}$	110-150	4.7
$VH \rightarrow b\bar{b}$	$Z \rightarrow \nu\bar{\nu}$	$E_T^{\text{miss}} \in \{120 - 160, 160 - 200, \geq 200 \text{ GeV}\}$	110-130	4.6
	$W \rightarrow l\nu$	$p_T^W \in \{< 50, 50 - 100, 100 - 200, \geq 200 \text{ GeV}\}$	110-130	4.7
	$Z \rightarrow ll$	$p_T^Z \in \{< 50, 50 - 100, 100 - 200, \geq 200 \text{ GeV}\}$	110-130	4.7

State of the art: At the time of the discovery, the combined Higgs search included 100 disjoint channels and >500 nuisance parameters

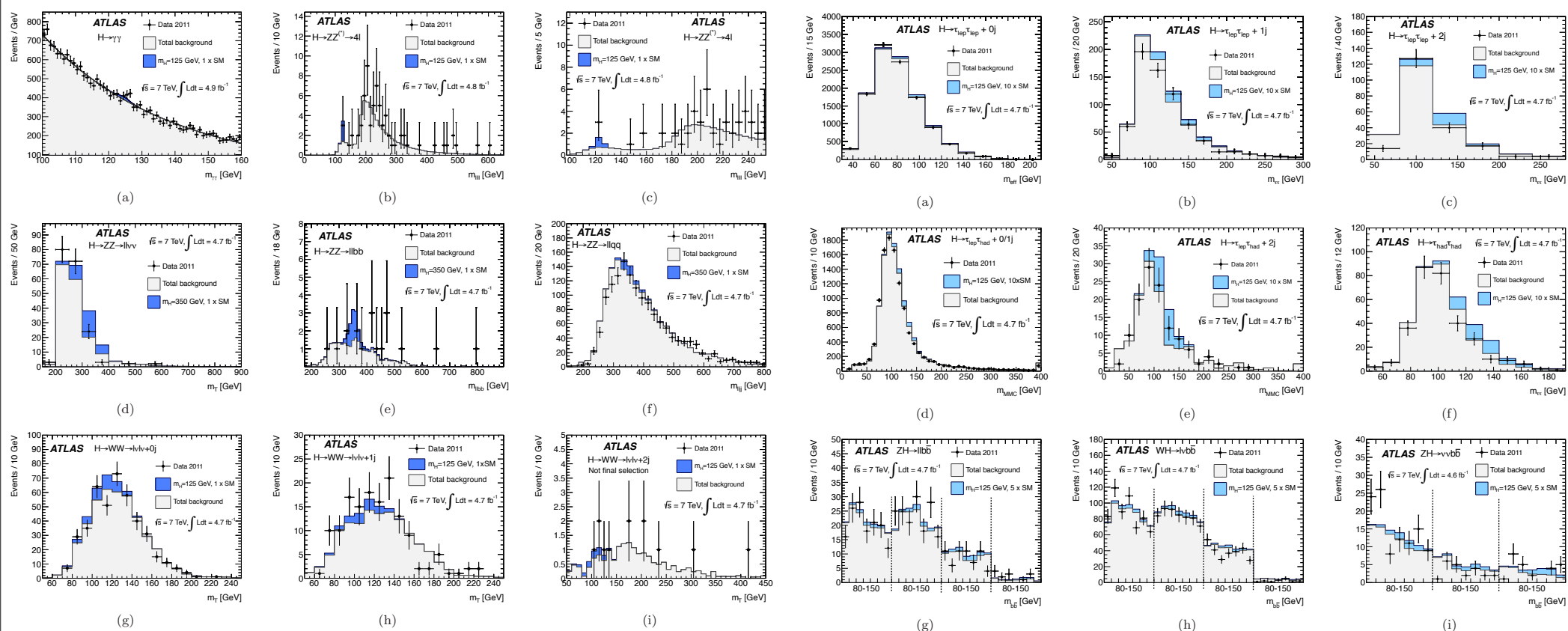
Roofit / RooStats: is the modeling language (C++) which provides technologies for collaborative modeling

- ▶ provides technology to publish likelihood functions digitally
- ▶ and more, it's the full model so we can also generate pseudo-data

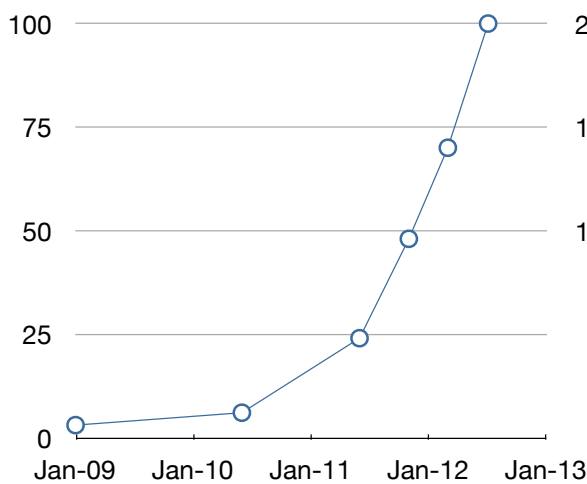
$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c | \nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce} | \boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p | \alpha_p)$$



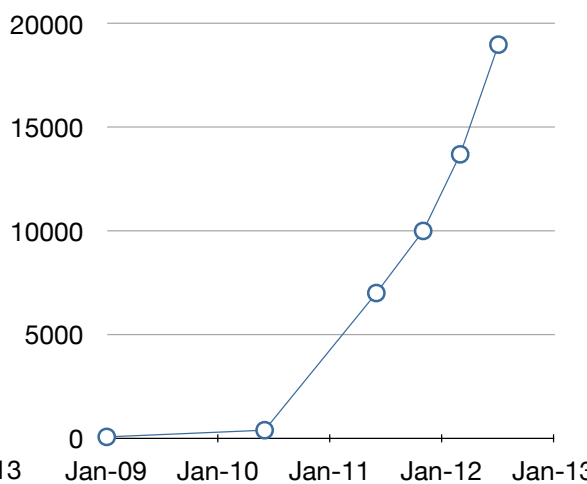
Evolution of Model Complexity



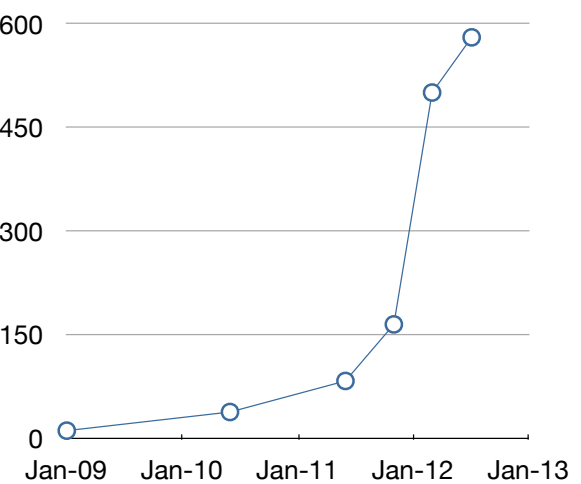
Number of Datasets Combined



Number of Model Components



Number of Parameters in Likelihood





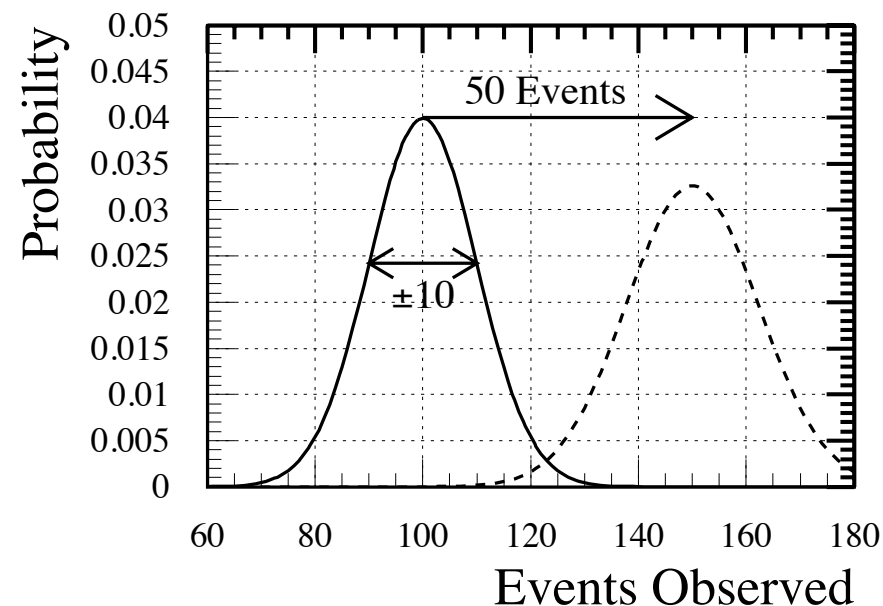
Lecture 3.5



Hypothesis Testing

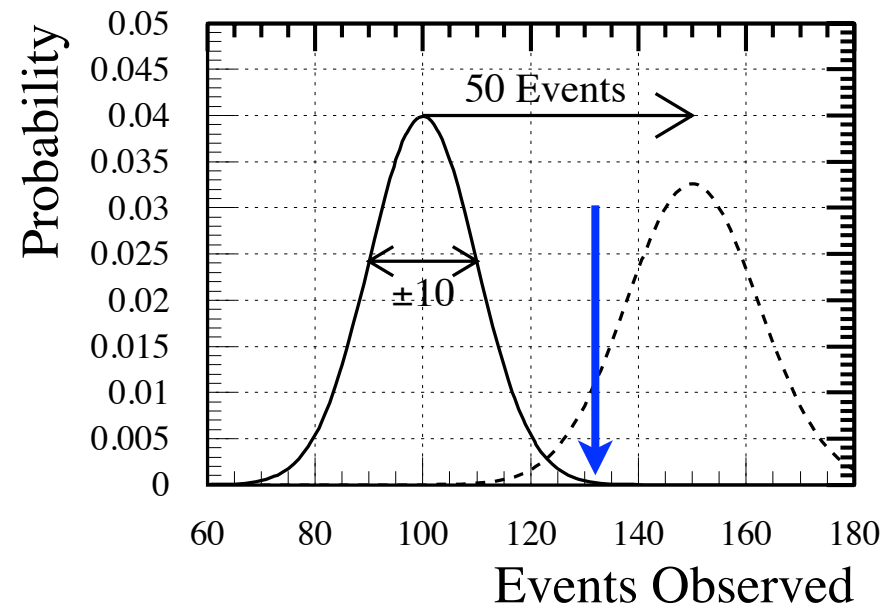
One of the most common uses of statistics in particle physics is Hypothesis Testing (e.g. for discovery of a new particle)

- ▶ 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



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



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

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”

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”

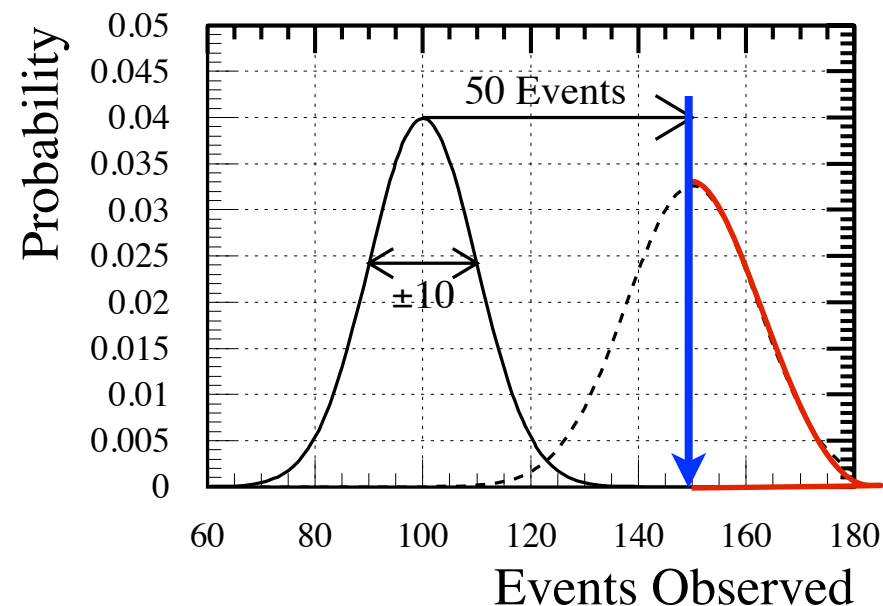
▶ Maximize power for a fixed rate of Type I error

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

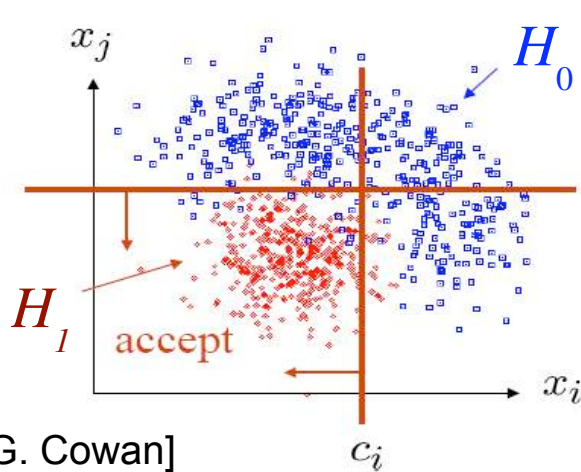


The idea of a “ 5σ ” discovery criteria for particle physics is really a conventional way to specify the size of the test

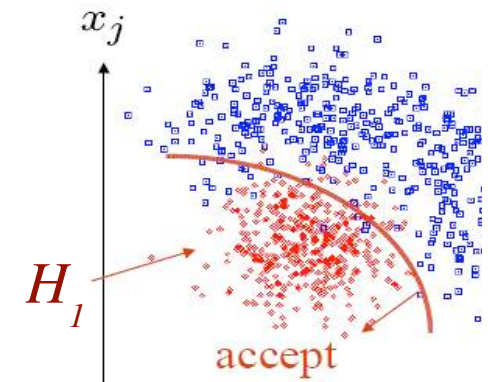
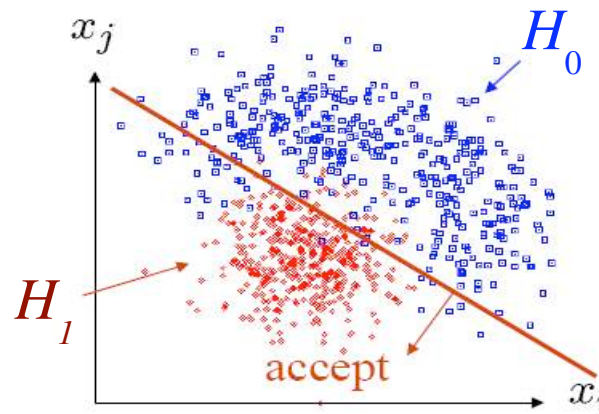
- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
 - eg. a very small chance we reject the standard model

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]



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)

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

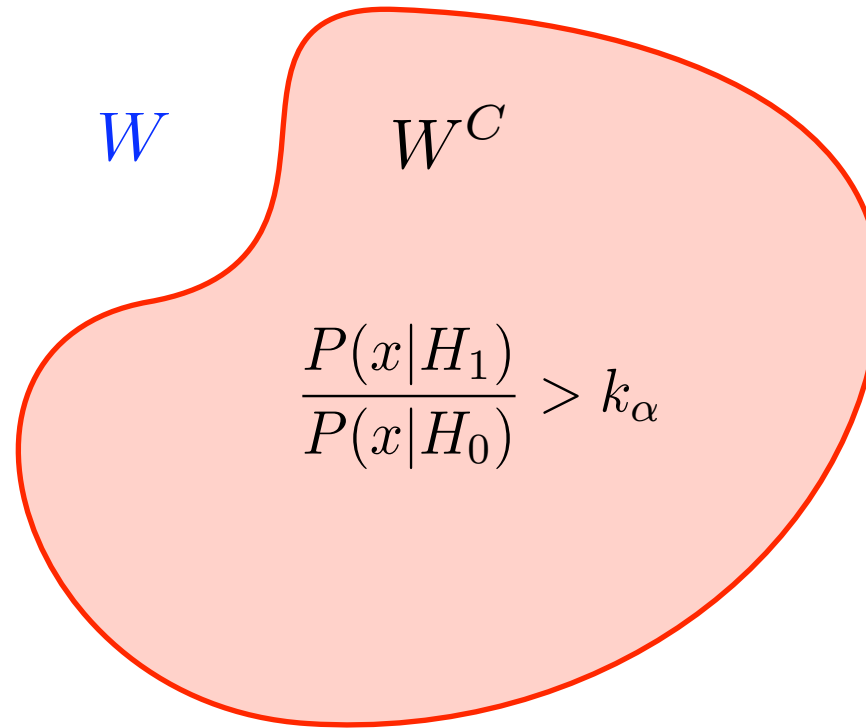
The region W that minimizes the probability of wrongly accepting H_0 is just a contour of the Likelihood Ratio

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

Any other region of the same size will have less power

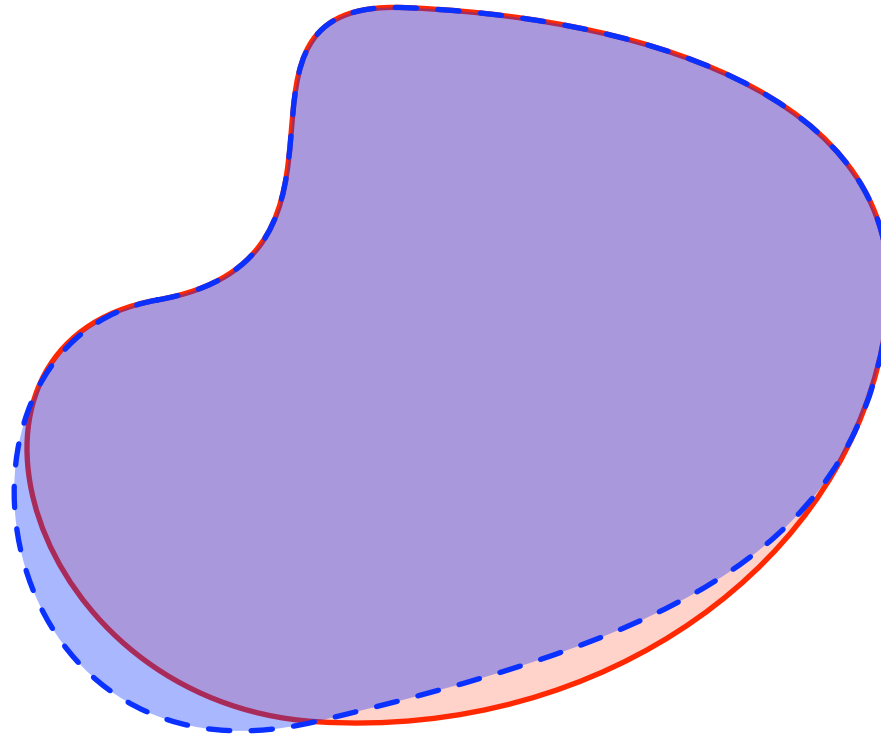
The likelihood ratio is an example of a **Test Statistic**, eg. a real-valued function that summarizes the data in a way relevant to the hypotheses that are being tested

A short proof of Neyman-Pearson

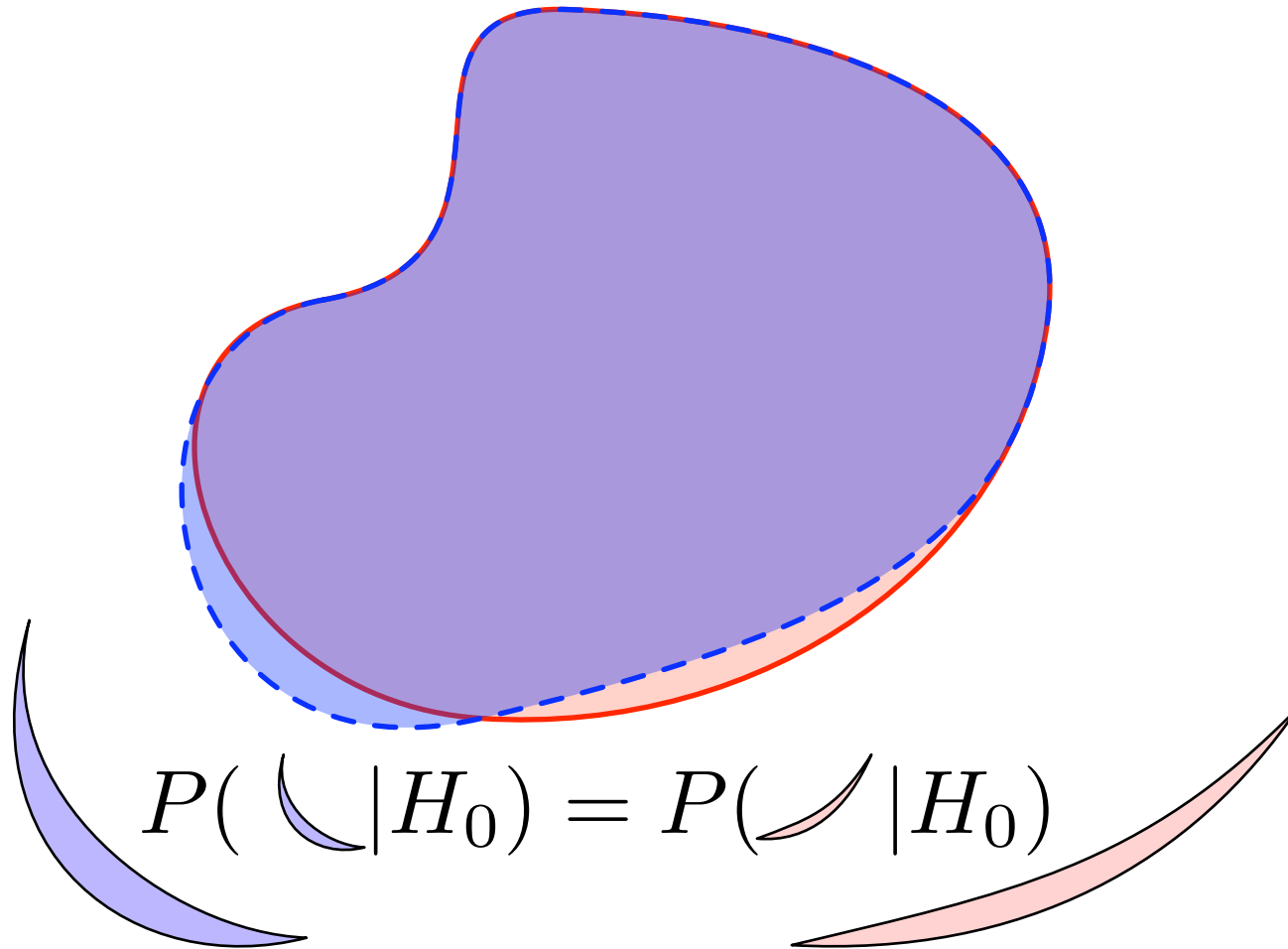


Consider the contour of the likelihood ratio that has size a given size (eg. probability under H_0 is $1-\alpha$)

A short proof of Neyman-Pearson

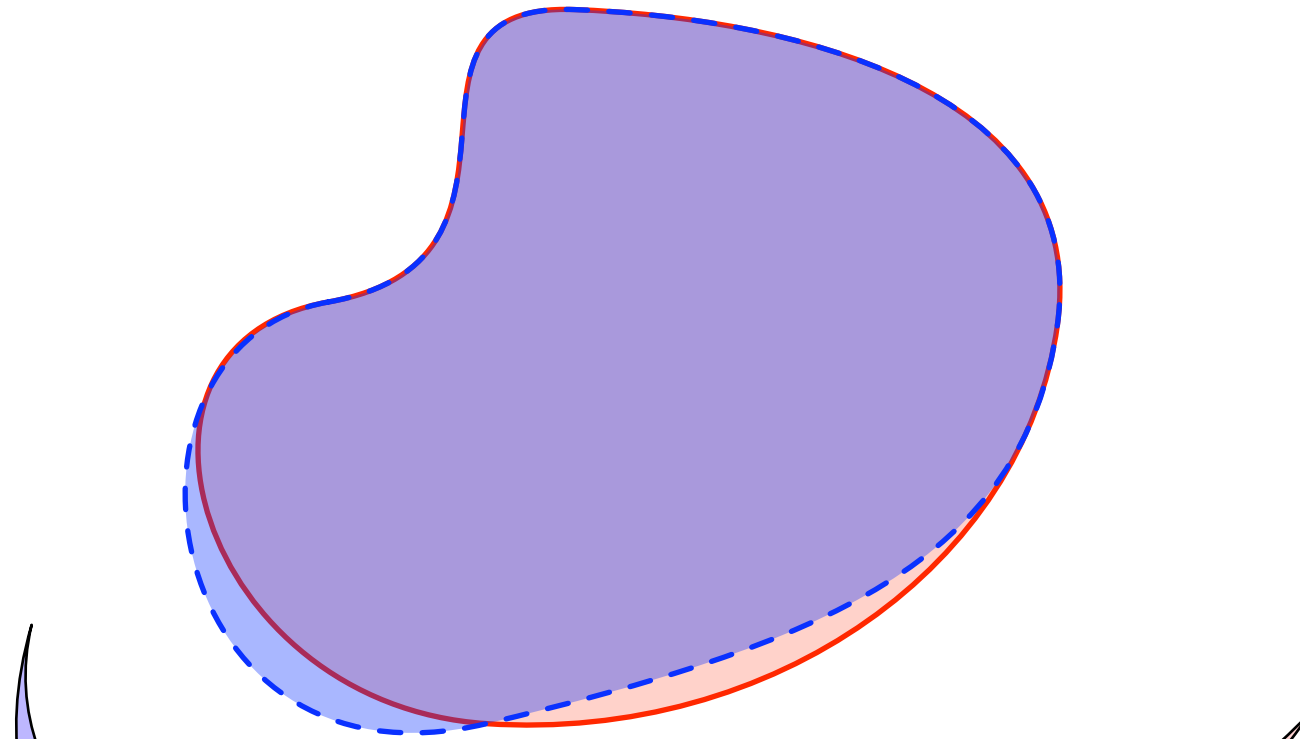


Now consider a variation on the contour that has the same size



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

A short proof of Neyman-Pearson



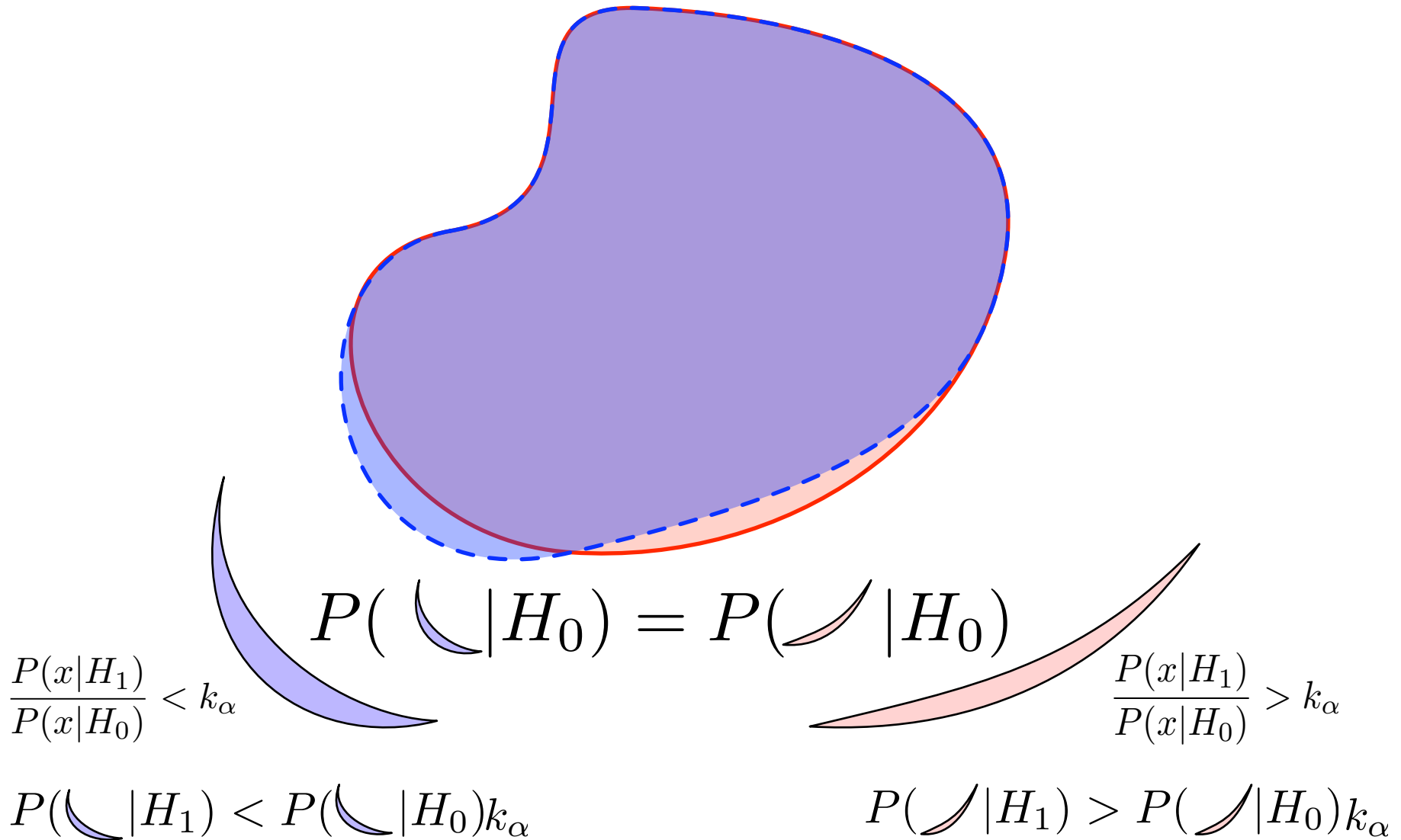
$$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha$$

$$P(\text{purple crescent} | H_0) = P(\text{red crescent} | H_0)$$

$$P(\text{purple crescent} | H_1) < P(\text{purple crescent} | H_0) k_\alpha$$

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

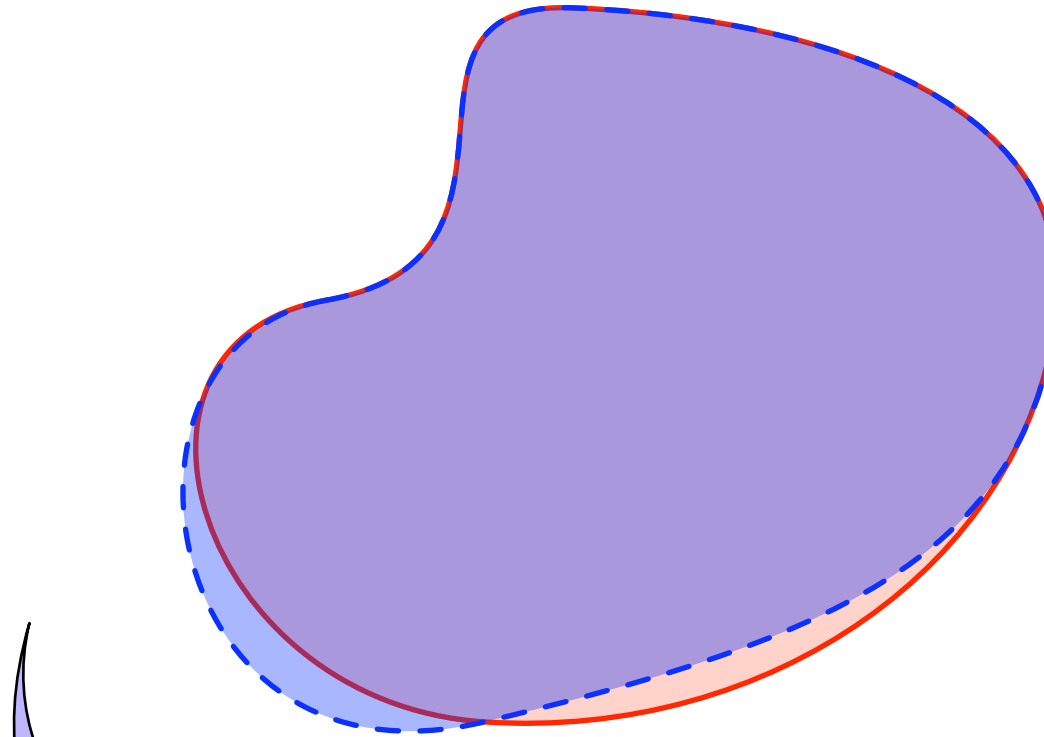
A short proof of Neyman-Pearson



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

Together they give...

A short proof of Neyman-Pearson



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

$$P(\text{blue crescent} | H_0) = P(\text{red crescent} | H_0)$$

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

$$P(\text{blue crescent} | H_1) < P(\text{blue crescent} | H_0)k_\alpha$$

$$P(\text{red crescent} | H_1) > P(\text{red crescent} | H_0)k_\alpha$$

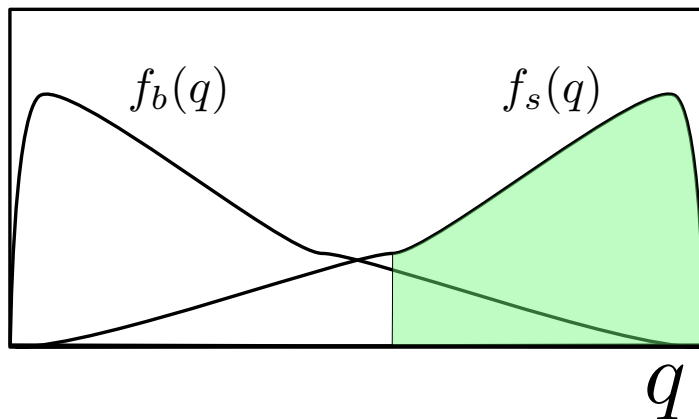
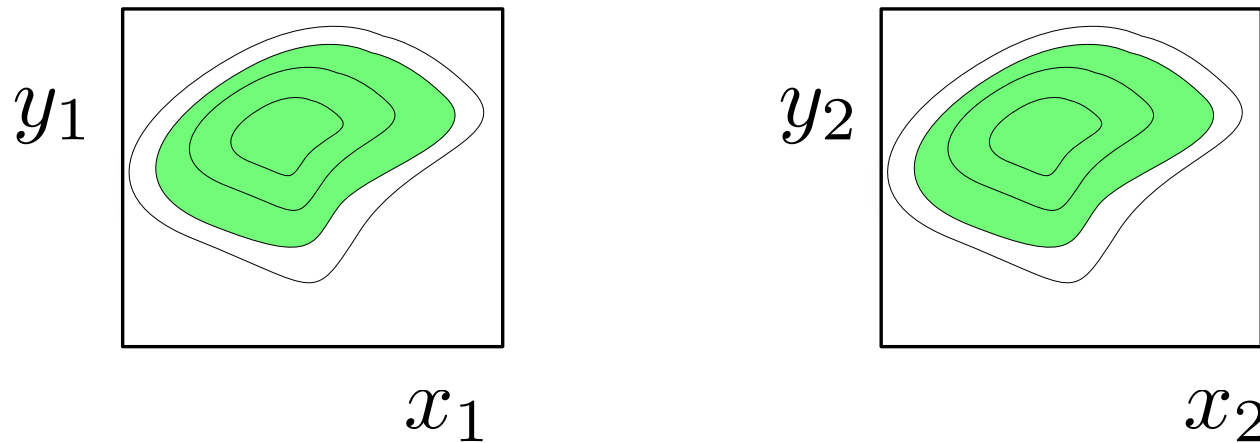
$$P(\text{blue crescent} | H_1) < P(\text{red crescent} | H_1)$$

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:



$$L_{tot} = L_1 \cdot L_2$$

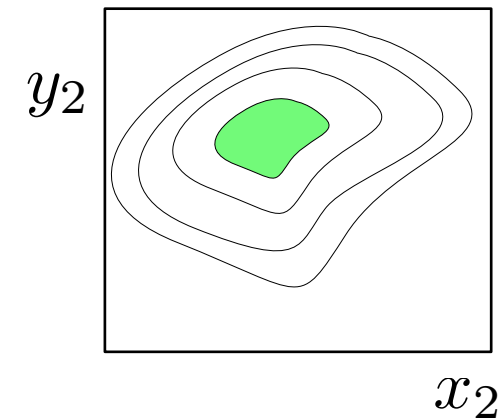
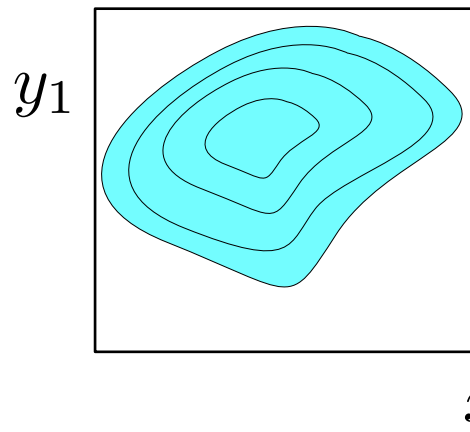
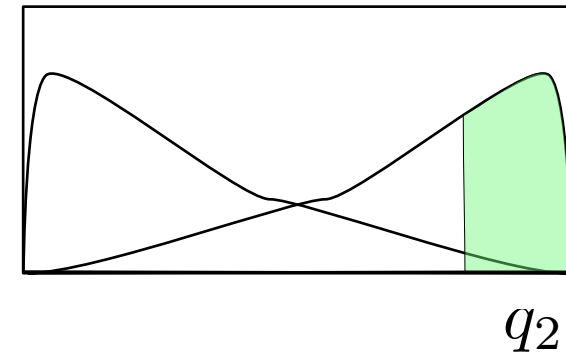
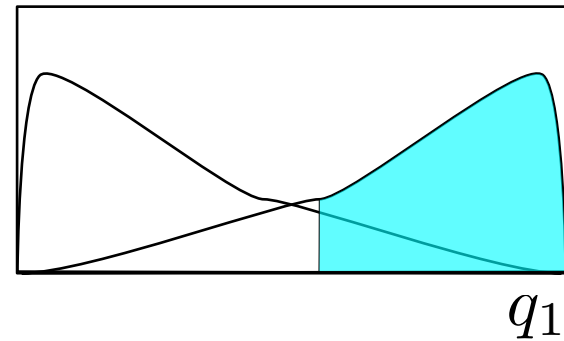
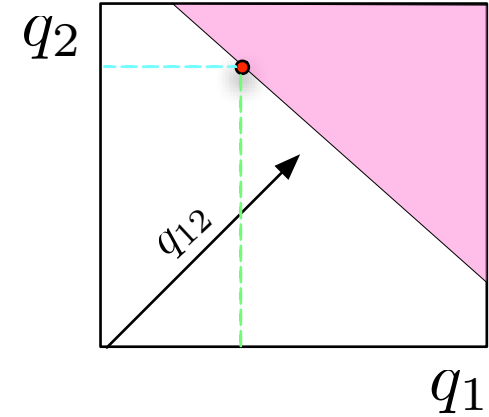
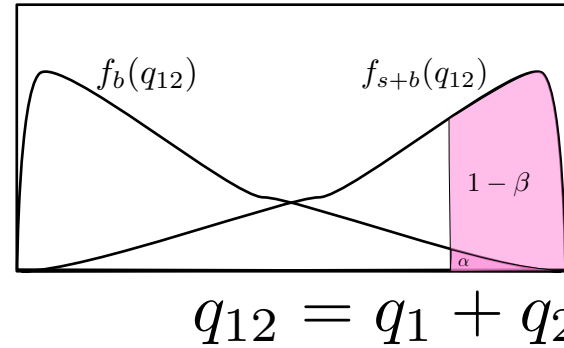
$$q_{12} = \ln L_{12} = \ln L_1 + \ln L_2 = q_1 + q_2$$

Experiments vs. Events

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 likelihood and the other one was relatively small



An optimal way to combine

Special case of our general probability model from yesterday

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

(no nuisance parameters)

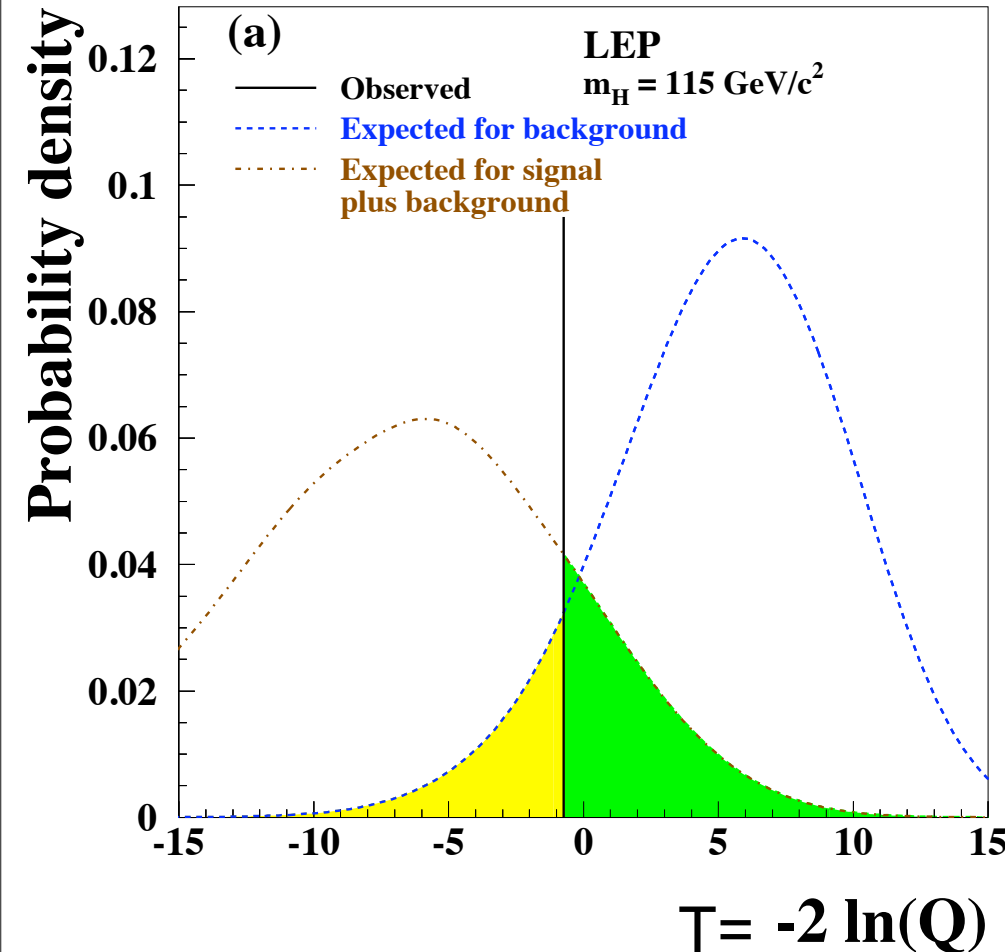
$$\ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln \left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})} \right)$$

Instead of simply counting events, the optimal test statistic is equivalent to adding events weighted by

$\ln(1 + \text{signal/background ratio})$

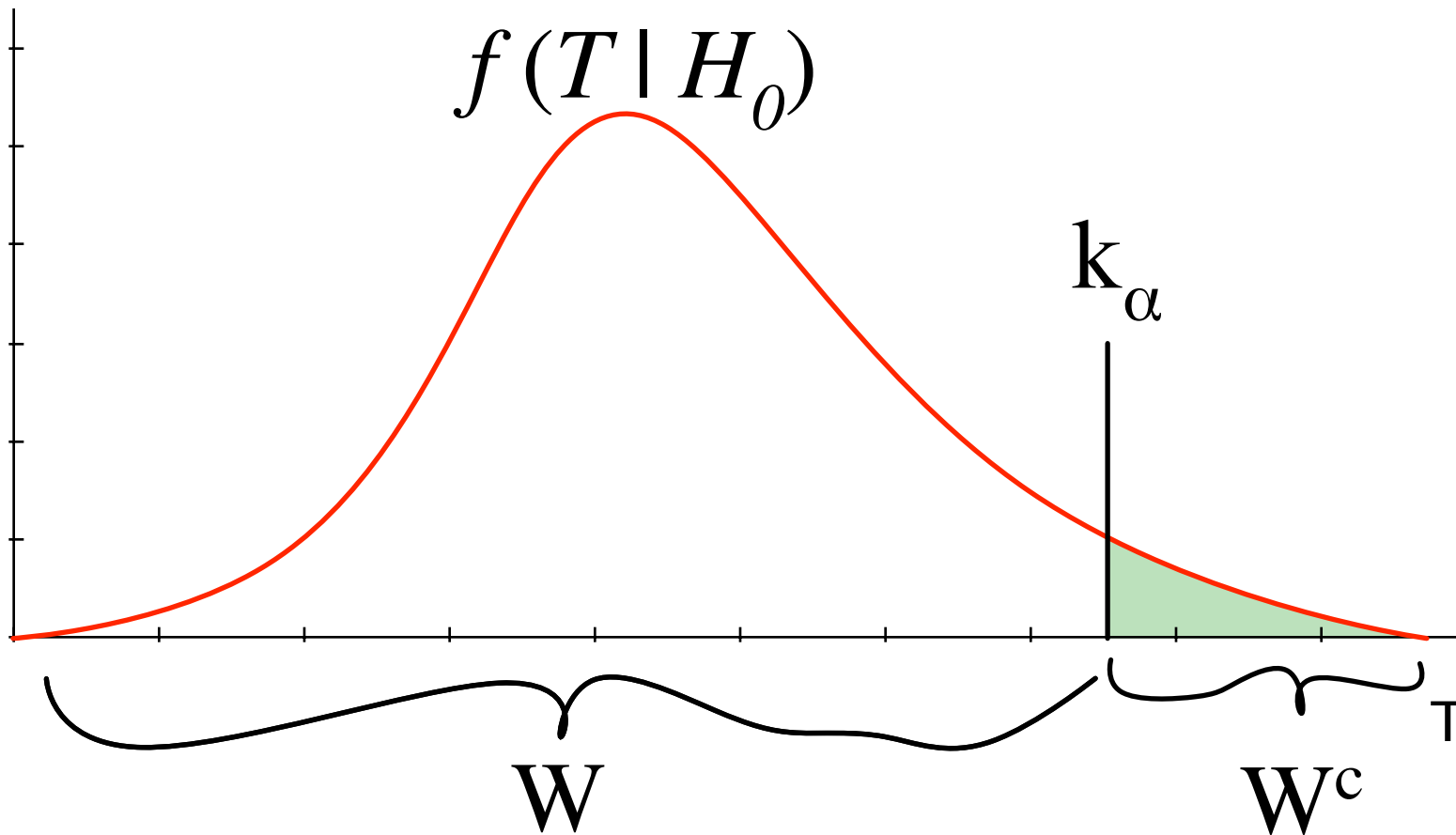
The test statistic is a map $T: \text{data} \rightarrow \mathbb{R}$

By repeating the experiment many times, you obtain a distribution for T



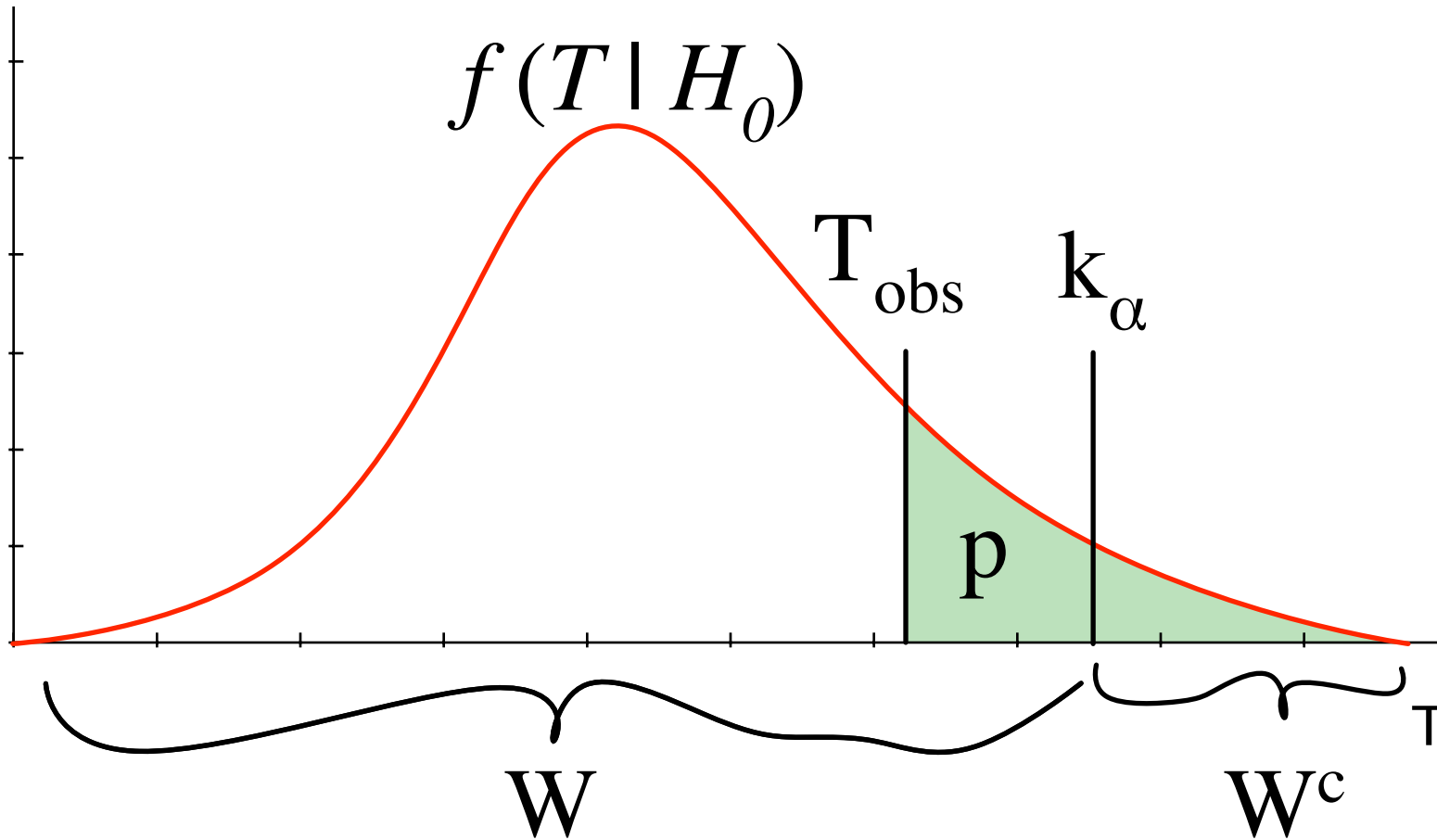
Instead of choosing to accept/reject H_0
one can compute the p-value

$$p = \int_{T_o}^{\infty} f(T|H_0)$$



Instead of choosing to accept/reject H_0
one can compute the p-value

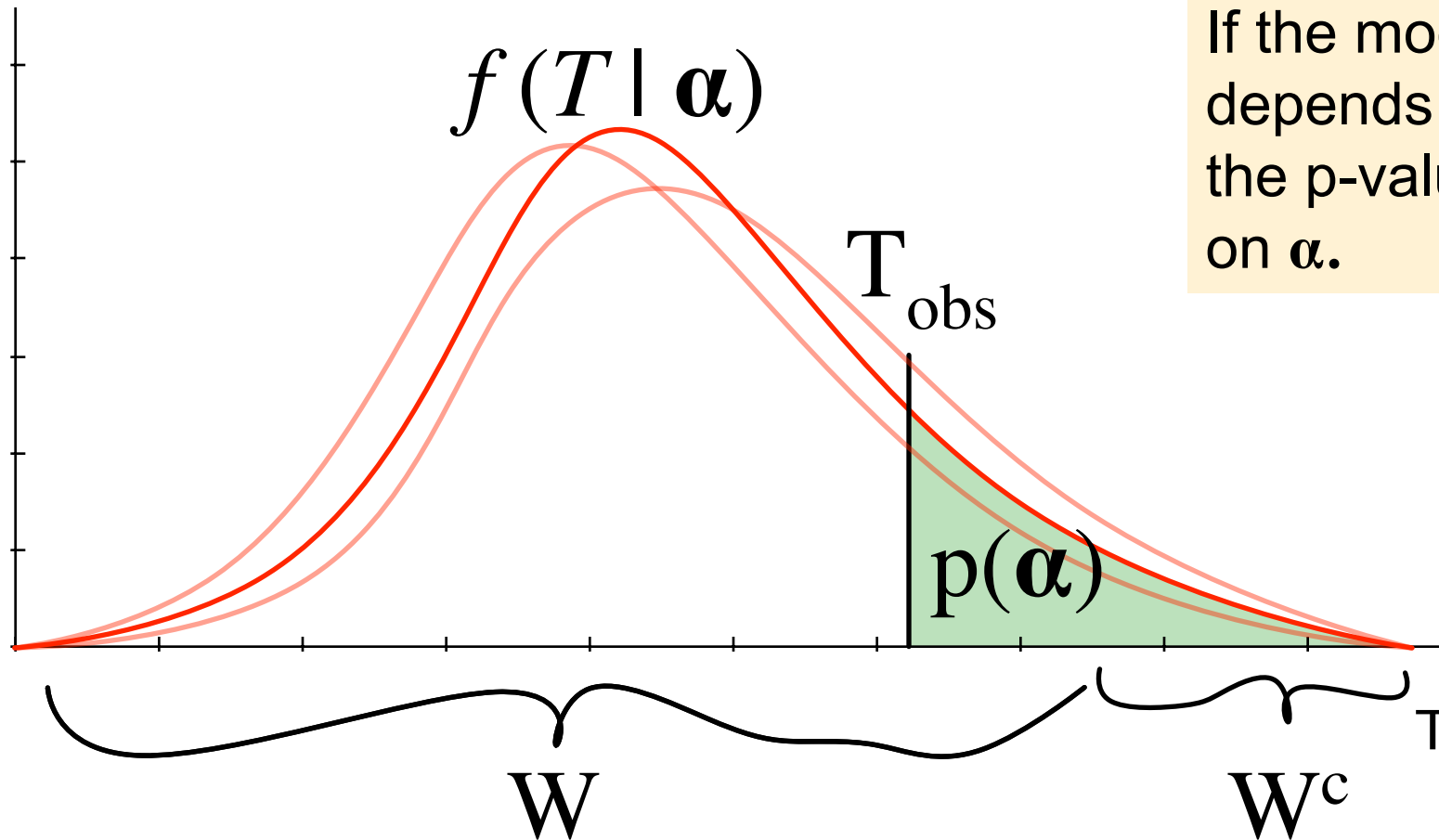
$$p = \int_{T_o}^{\infty} f(T|H_0)$$



Instead of choosing to accept/reject H_0
one can compute the p-value

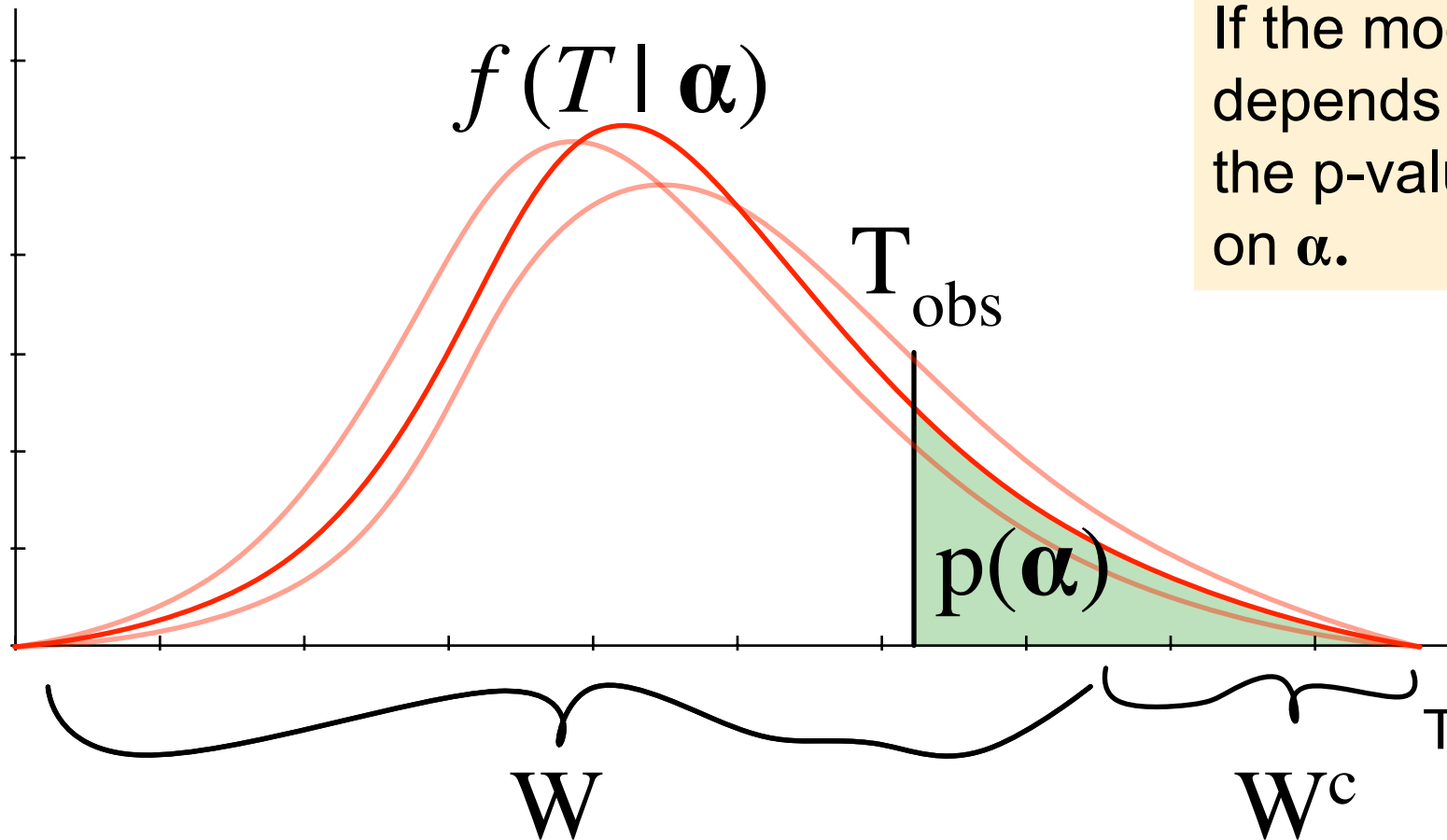
$$p = \int_{T_0}^{\infty} f(T|H_0)$$

If the model for the data depends on parameters α the p-value also depends on α .



$$p(\alpha) = \int_{T_0}^{\infty} f(T|\alpha) dT = \int \mathbf{f}(\mathcal{D}|\alpha) \theta(T(\mathcal{D}) - T_0) d\mathcal{D} = P(T \geq T_0|\alpha)$$

When the model has nuisance parameters, only reject the null if $p(\alpha)$ sufficiently small **for all values** of the nuisance parameters.



If the model for the data depends on parameters α the p-value also depends on α .

$$p(\alpha) = \int_{T_0}^{\infty} f(T|\alpha) dT = \int \mathbf{f}(\mathcal{D}|\alpha) \theta(T(\mathcal{D}) - T_0) d\mathcal{D} = P(T \geq T_0 | \alpha)$$

Consider our general model with a single parameter of interest μ

- ▶ let $\mu=0$ be no signal, $\mu=1$ nominal signal

In the LEP approach the likelihood ratio is equivalent to:

$$Q_{\text{LEP}} = \frac{L(\mu = 1, \theta)}{L(\mu = 0, \theta)} = \frac{f(\mathcal{D}|\mu = 1, \theta)}{f(\mathcal{D}|\mu = 0, \theta)}$$

- ▶ but this variable is sensitive to uncertainty on θ and makes no use of auxiliary measurements **a**

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

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(\mathcal{D}, \mathcal{G}|\mu, \hat{\hat{\theta}}(\mu; \mathcal{D}, \mathcal{G}))}{f(\mathcal{D}, \mathcal{G}|\hat{\mu}, \hat{\theta})}$$

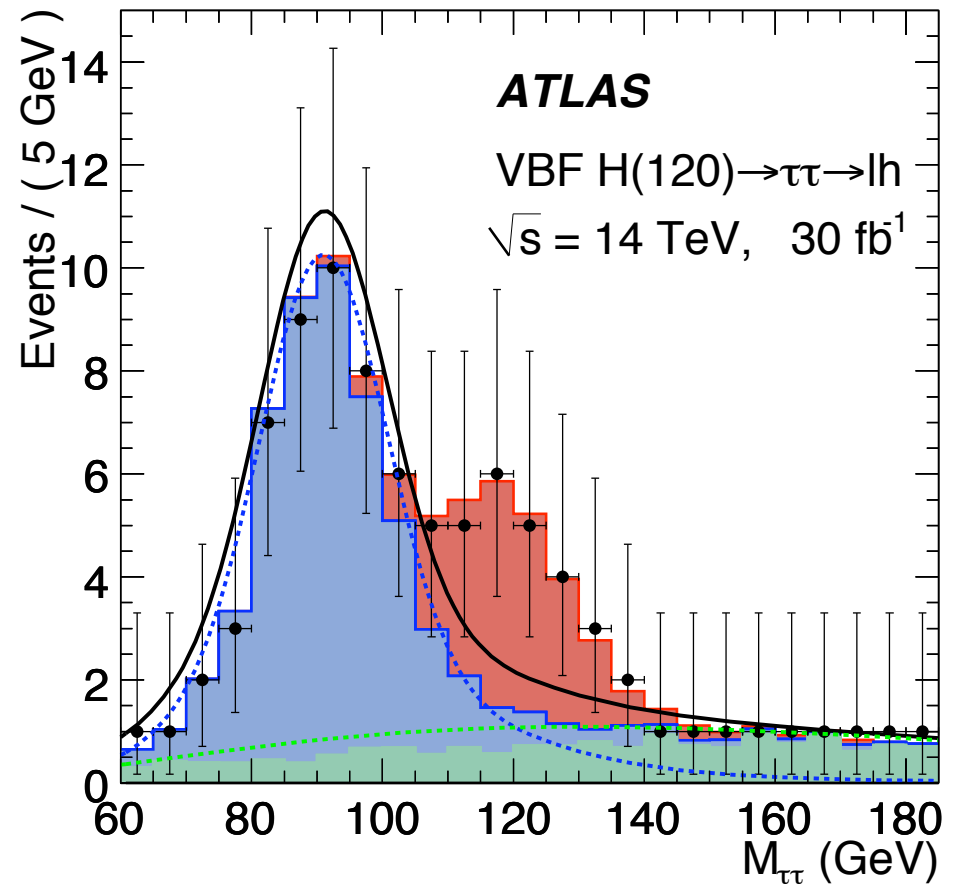
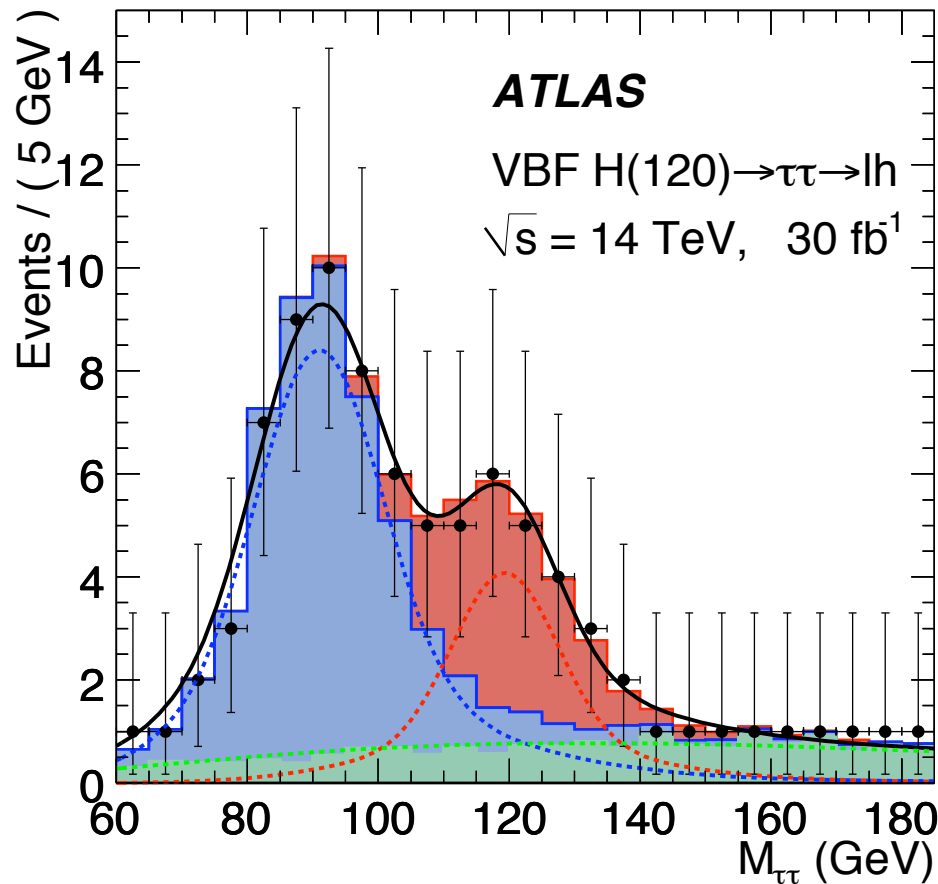
- ▶ where $\hat{\hat{\theta}}(\mu; \mathcal{D}, \mathcal{G})$ is best fit with μ fixed (the constrained maximum likelihood estimator, depends on data)
- ▶ and $\hat{\theta}$ and $\hat{\mu}$ are best fit with both left floating (unconstrained)
- ▶ Tevatron used $Q_{\text{Tev}} = \lambda(\mu=1)/\lambda(\mu=0)$ as generalization of Q_{LEP}

Essentially, you need to fit your model to the data twice:
once with everything floating, and once with signal fixed to 0

$$\lambda(\mu = 0) = \frac{L(\mu = 0, \hat{\hat{\theta}}(\mu = 0))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(\mathcal{D}, \mathcal{G} | \mu = 0, \hat{\hat{\theta}}(\mu = 0; \mathcal{D}, \mathcal{G}))}{f(\mathcal{D}, \mathcal{G} | \hat{\mu}, \hat{\theta})}$$

$$f(\mathcal{D}, \mathcal{G} | \hat{\mu}, \hat{\theta})$$

$$f(\mathcal{D}, \mathcal{G} | \mu = 0, \hat{\hat{\theta}}(\mu = 0; \mathcal{D}, \mathcal{G}))$$



After a close look at the profile likelihood ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(\mathcal{D}, \mathcal{G} | \mu, \hat{\hat{\theta}}(\mu; \mathcal{D}, \mathcal{G}))}{f(\mathcal{D}, \mathcal{G} | \hat{\mu}, \hat{\theta})}$$

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

- ▶ though its distribution might depend indirectly

Wilks's theorem states that under certain conditions the distribution of $-2 \ln \lambda (\mu=\mu_0)$ given that the true value of μ is μ_0 converges to a chi-square distribution

- ▶ more on this tomorrow, but the important points are:
- ▶ “asymptotic distribution” is known and it is independent of θ !
 - more complicated if parameters have boundaries (eg. $\mu \geq 0$)

Thus, we can calculate the p-value for the background-only hypothesis without having to generate Toy Monte Carlo!

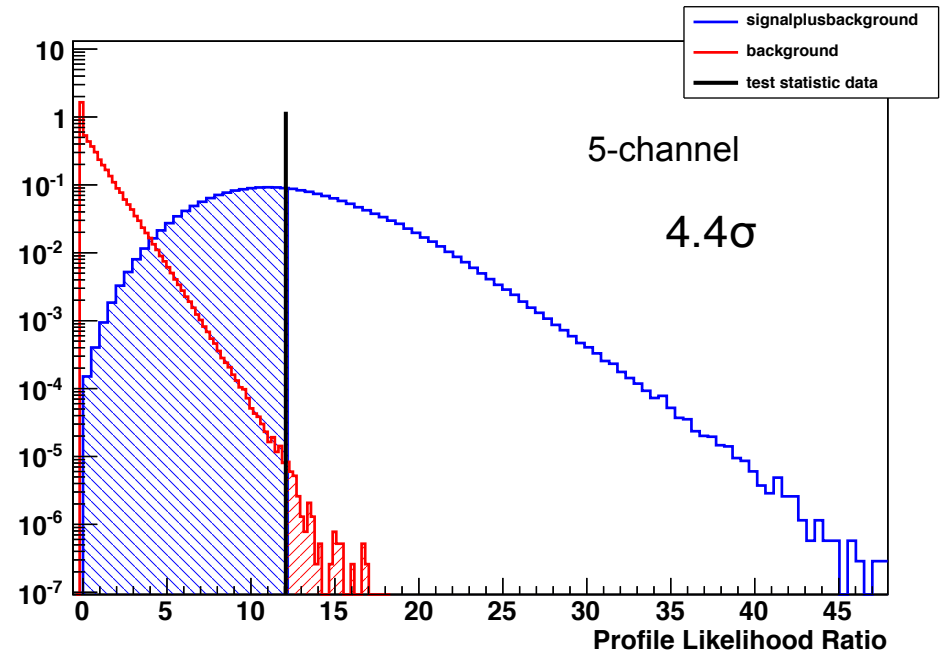
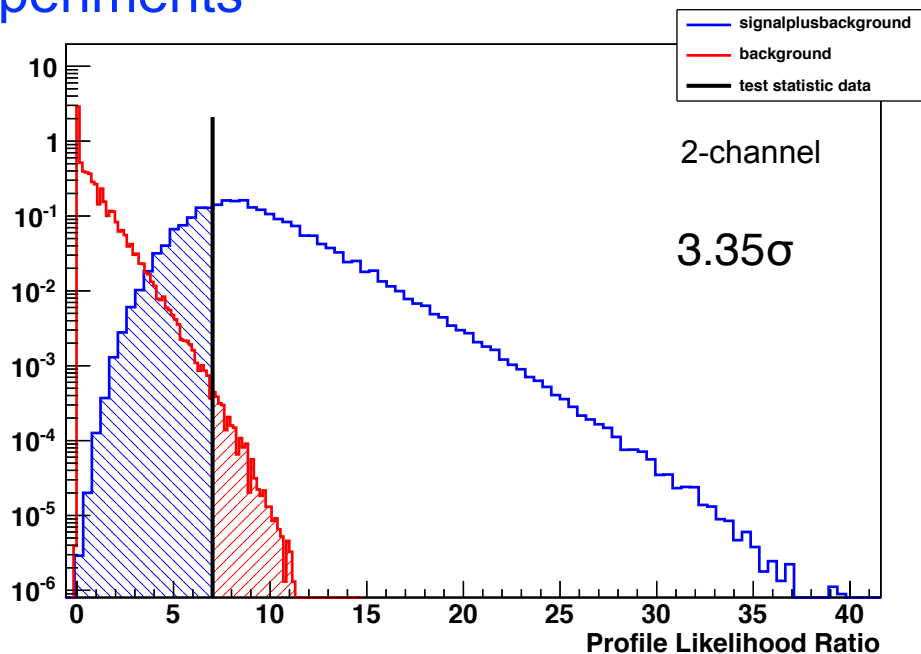
Explicitly build distribution by generating “toys” / pseudo experiments assuming a specific value of μ and ν .

- ▶ randomize both main measurements $\mathcal{D}=\{x\}$ and auxiliary measurements $\mathcal{C}=\{a\}$
- ▶ fit the model twice for the numerator and denominator of profile likelihood ratio
- ▶ evaluate $-2\ln \lambda(\mu)$ and add to histogram

Choice of μ is straight forward: typically $\mu=0$ and $\mu=1$, but choice of θ is less clear

- ▶ more on this tomorrow

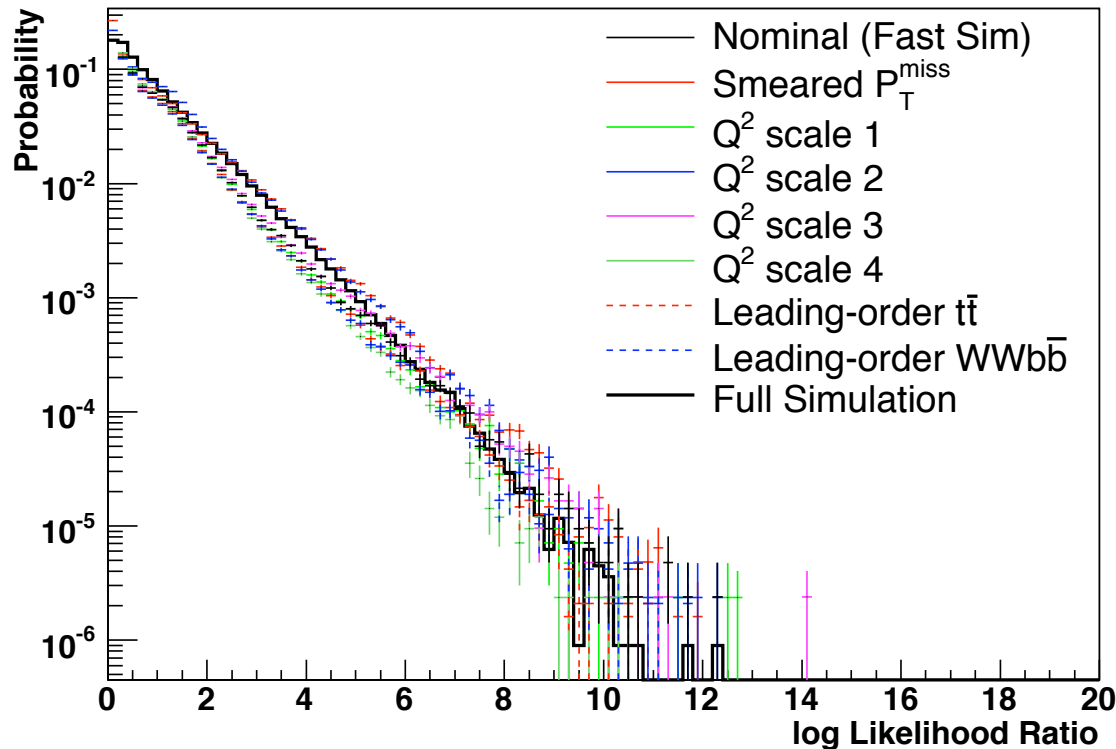
This can be very time consuming. Plots below use millions of “toy” pseudo-experiments



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.

A very important point

If we keep pushing this point to the extreme, the physics problem goes beyond what we can handle practically

The p-values are usually predicated on the assumption that the **true distribution** is in the family of functions being considered

- ▶ eg. we have sufficiently flexible models of signal & background to incorporate all systematic effects
- ▶ but we don't believe we simulate everything perfectly
- ▶ ..and when we parametrize our models usually we have further approximated our simulation.
 - nature -> simulation -> parametrization

At some point these approaches are limited by honest systematic uncertainties (not statistical ones). Statistics can only help us so much after this point. Now we must be physicists!