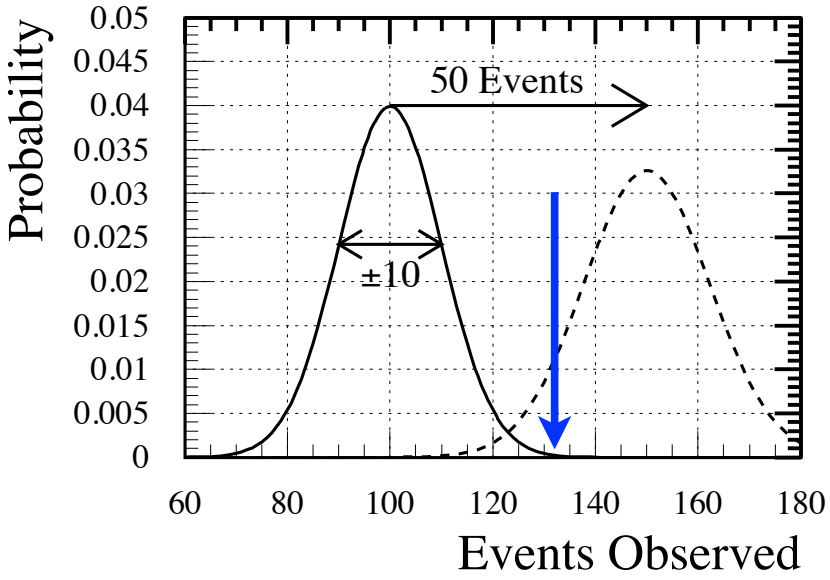
Hypothesis Testing

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

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

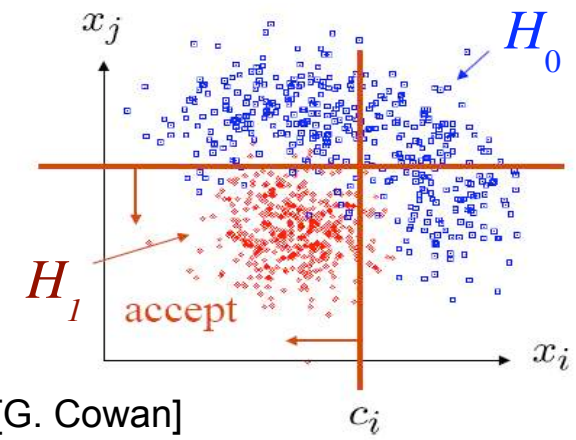
Hypothesis testing

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

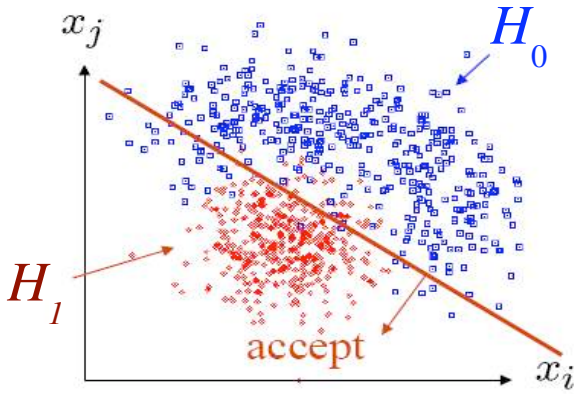
- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

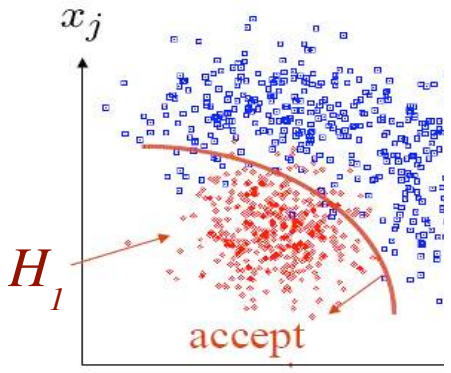
In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- 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

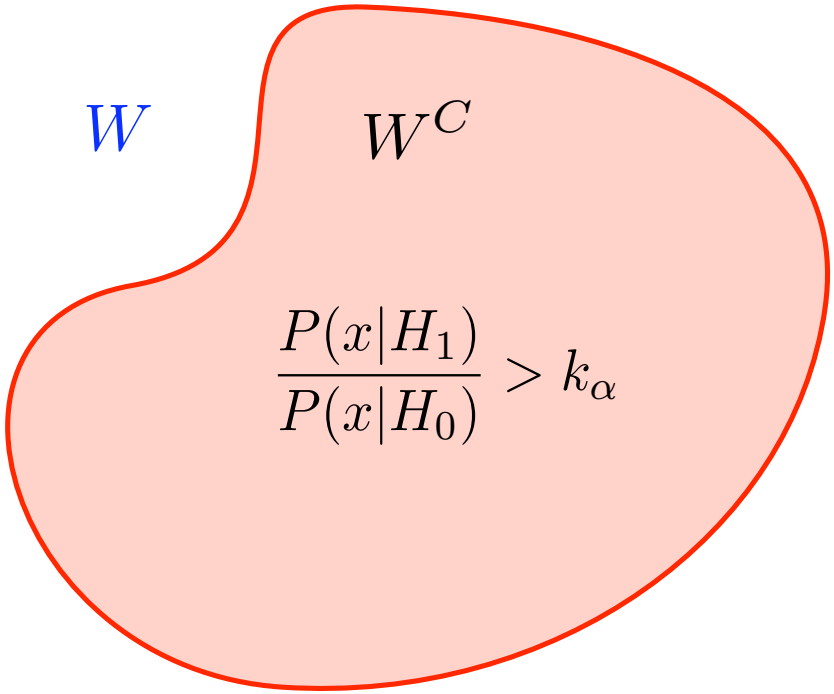
The region W that minimizes the probability of wrongly accepting H_0 is just a contour of the Likelihood Ratio

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

Any other region of the same size will have less power

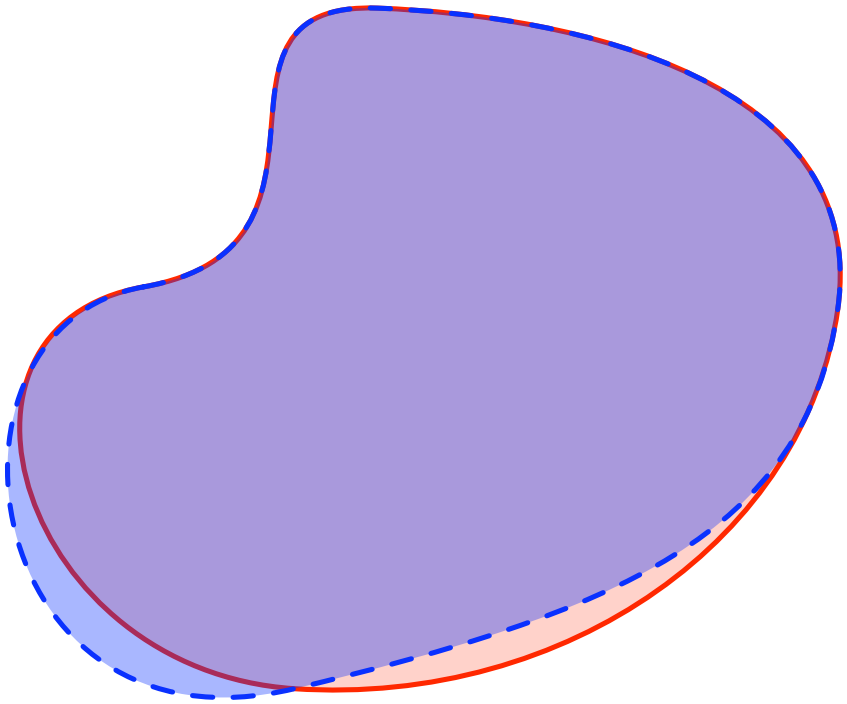
The likelihood ratio is an example of a **Test Statistic**, eg. a real-valued function that summarizes the data in a way relevant to the hypotheses that are being tested

A short proof of Neyman-Pearson



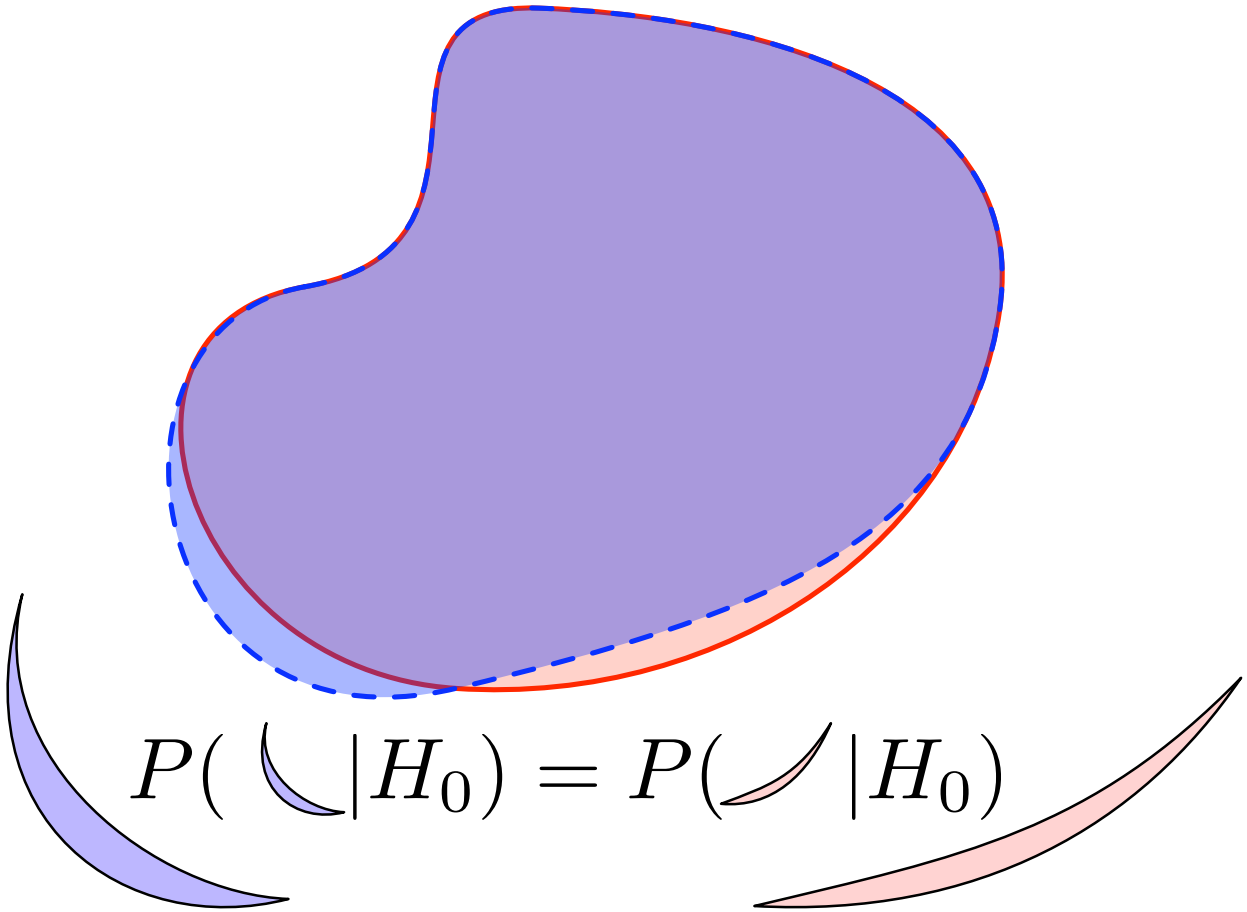
Consider the contour of the likelihood ratio that has size a given size (eg. probability under H_0 is 1-α)

A short proof of Neyman-Pearson



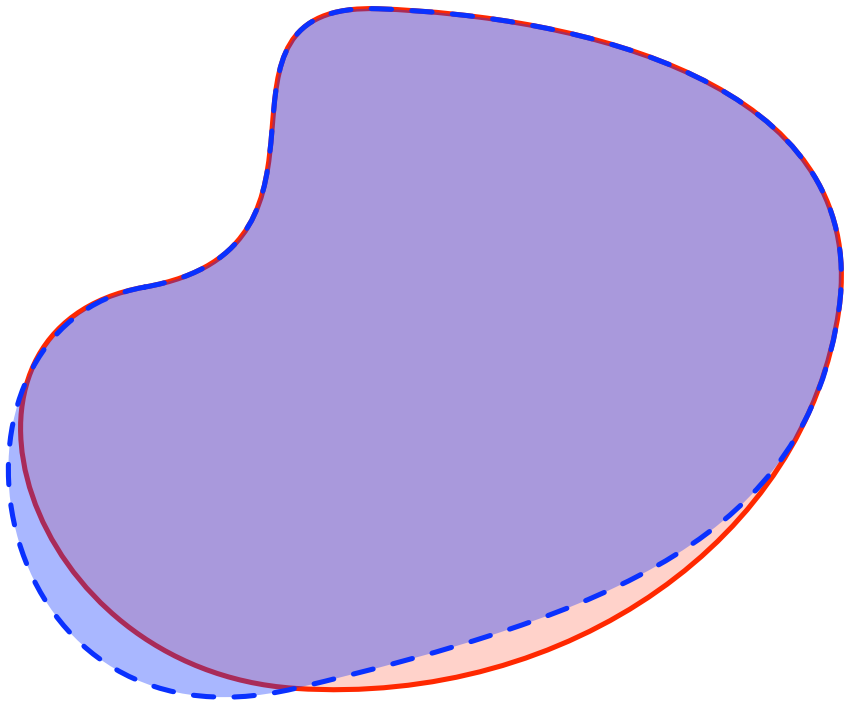
Now consider a variation on the contour that has the same size

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



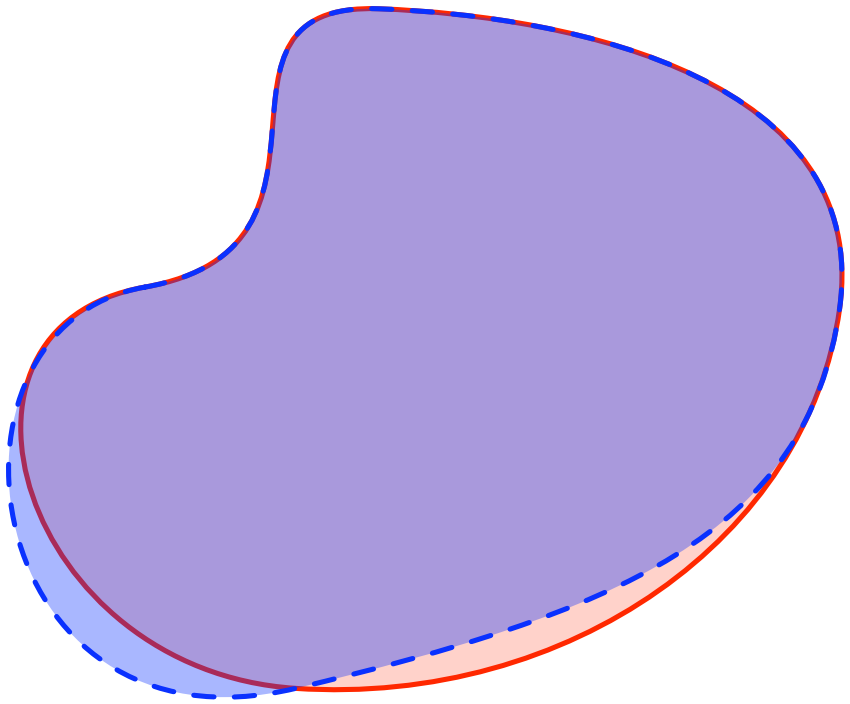
$$P(\ |H_0) = P(\ |H_0)$$

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

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

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

A short proof of Neyman-Pearson



$$P(\;|H_0) = P(\;|H_0)$$

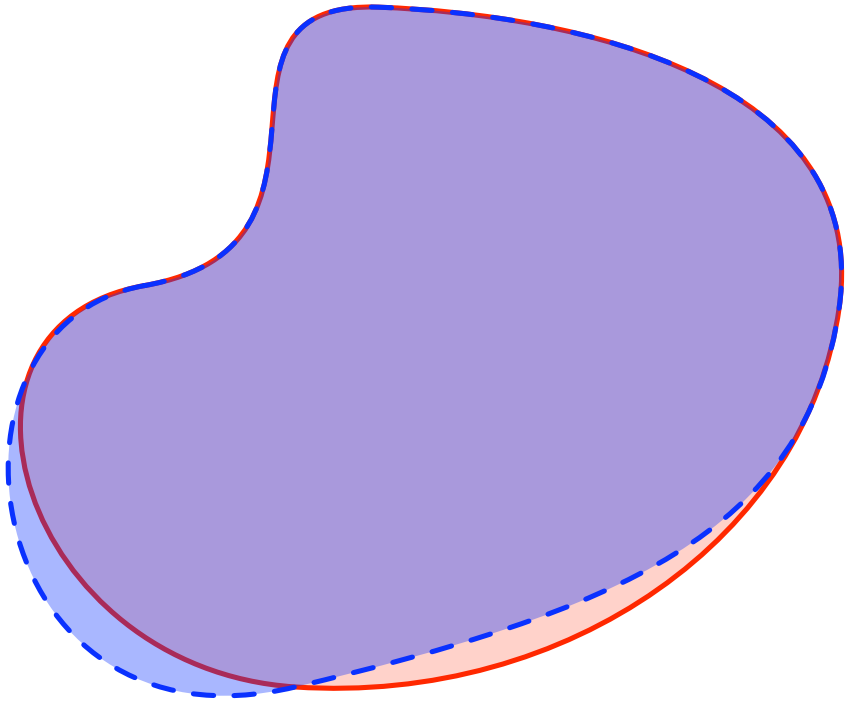
$$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha \qquad \frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

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

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

Together they give...

A short proof of Neyman-Pearson



$$P(\text{ }|H_0) = P(\text{ }|H_0)$$

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

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

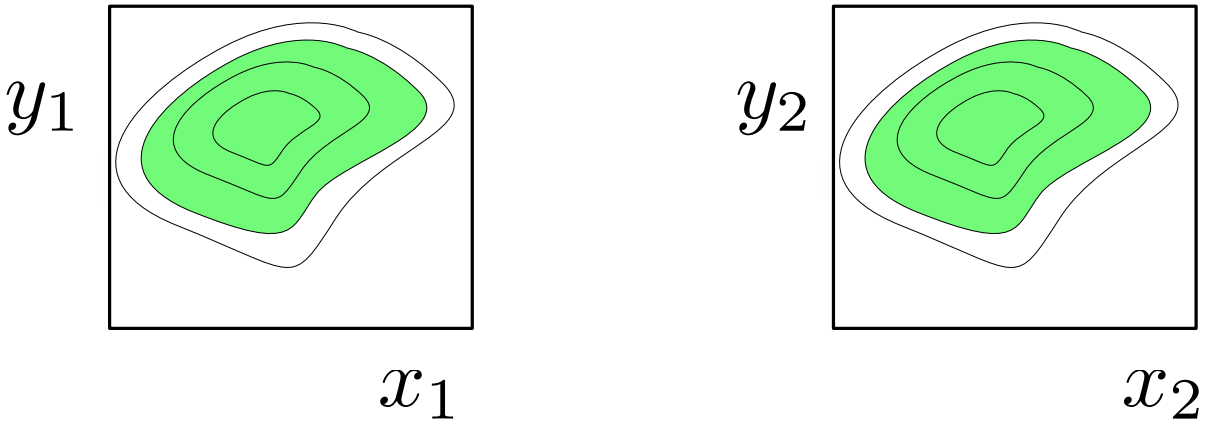
$$P(\text{ }|H_1) < P(\text{ }|H_1)$$

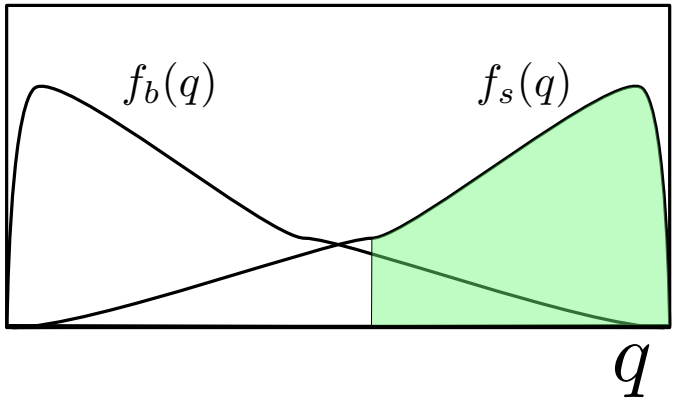
The new region region has less power.

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:





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

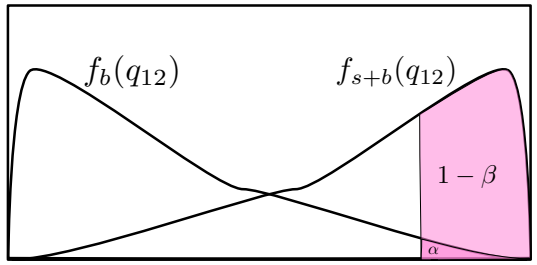
$$q_{12} = \ln L_{12} = \ln L_1 + \ln L_2 = q_1 + q_2$$

Experiments vs. Events

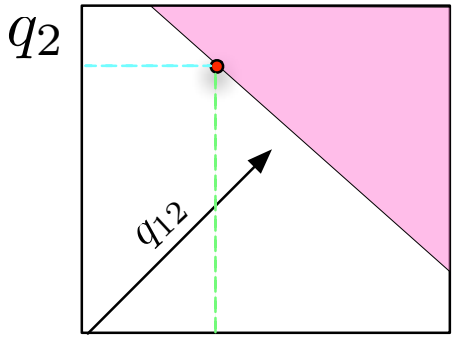
Ideally, you want to cut on the likelihood ratio for your **experiment**

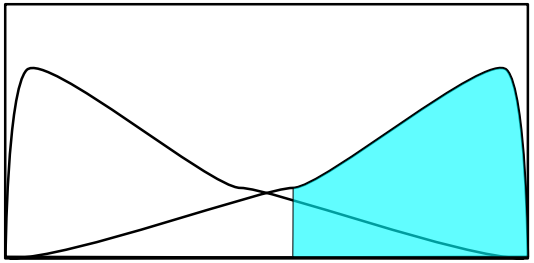
- equivalent to a sum of log likelihood ratios

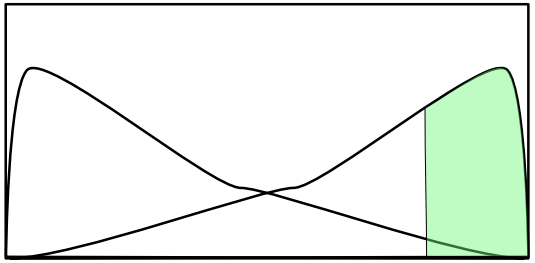
Easy to see that includes experiments where one event had a high likelihood and the other one was relatively small

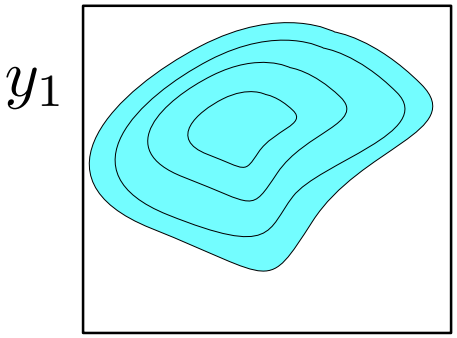


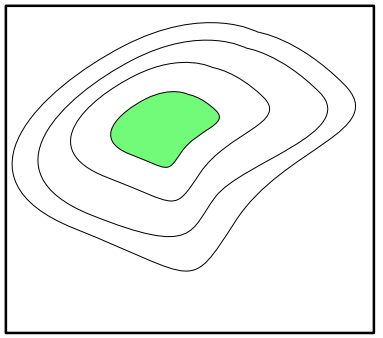
$$q_{12} = q_1 + q_2$$











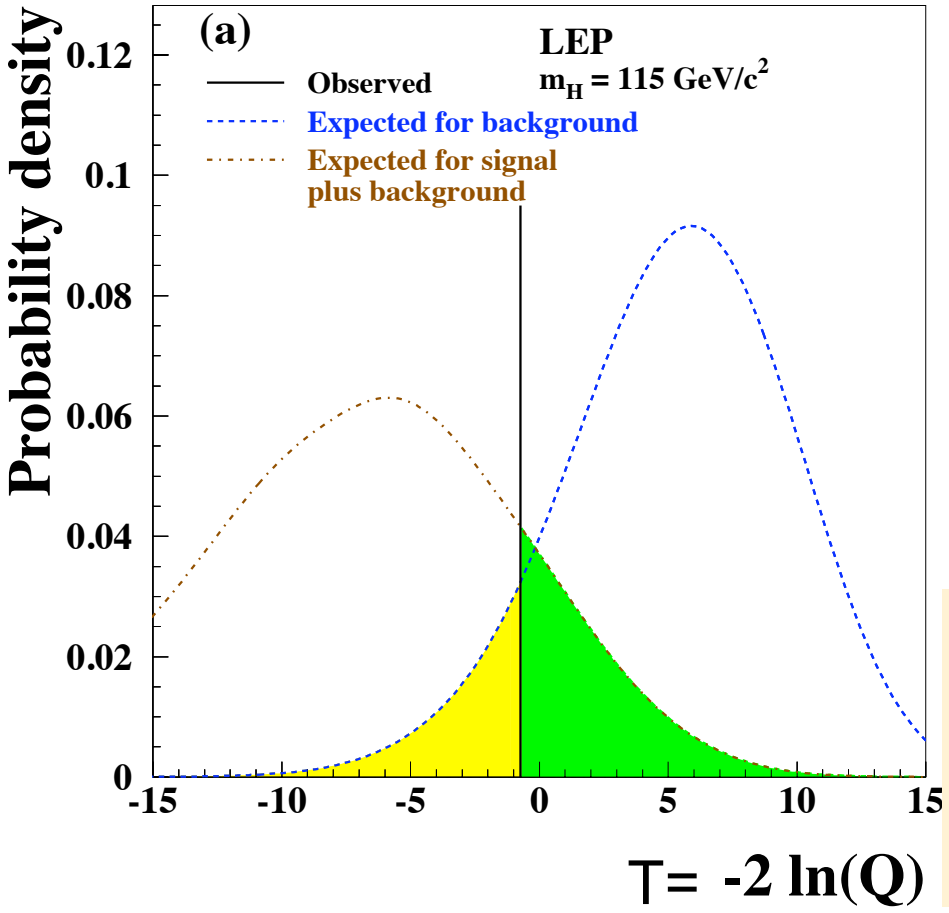
An optimal way to combine

Special case of our general probability model from yesterday

(no nuisance parameters)

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

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



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

ln(1+signal/background ratio)

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

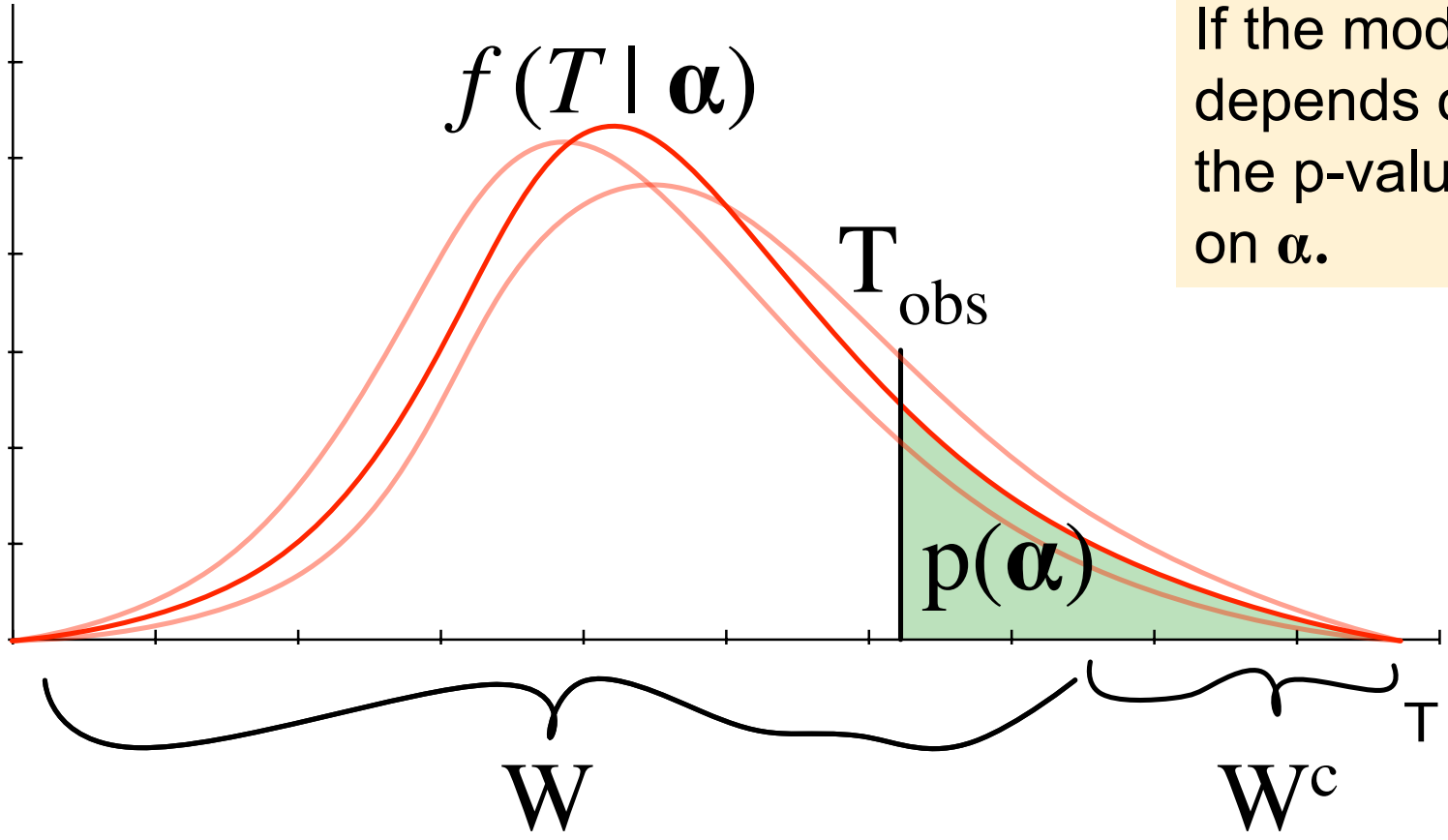
By repeating the experiment many times, you obtain a distribution for T

p-values

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

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

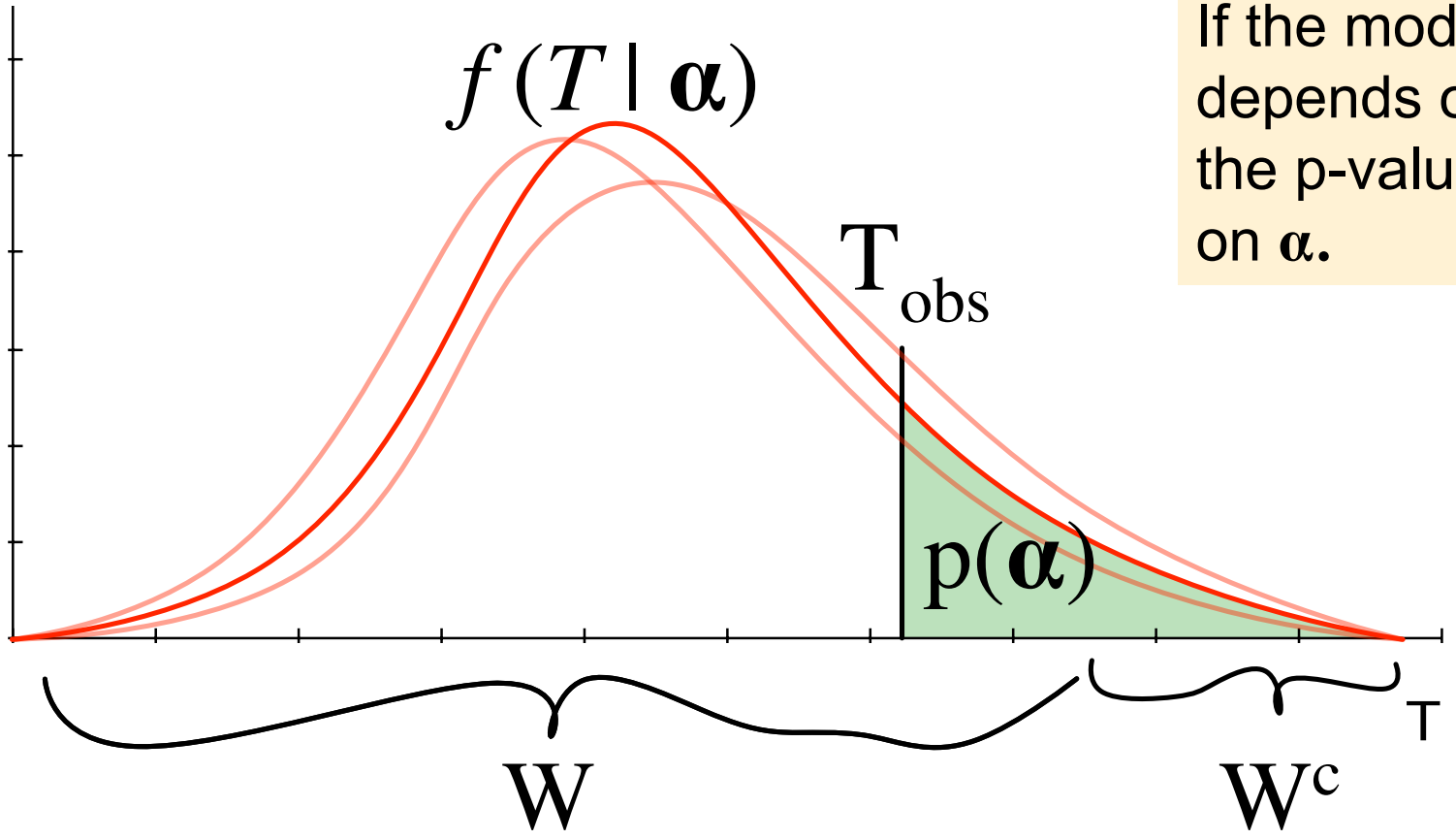
If the model for the data depends on parameters $\boldsymbol{\alpha}$ the p-value also depends on $\boldsymbol{\alpha}$.



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

p-values

When the model has nuisance parameters, only reject the null if p(α) 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.

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

The Profile Likelihood Ratio

Consider our general model with a single parameter of interest μ

- let μ=0 be no signal, μ=1 nominal signal

In the LEP approach the likelihood ratio is equivalent to:

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

- but this variable is sensitive to uncertainty on ν 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 $Q_{\mathrm{Tev}} = \lambda(\mu{=}1)/\lambda(\mu{=}0)$ as generalization of Q_{LEP}

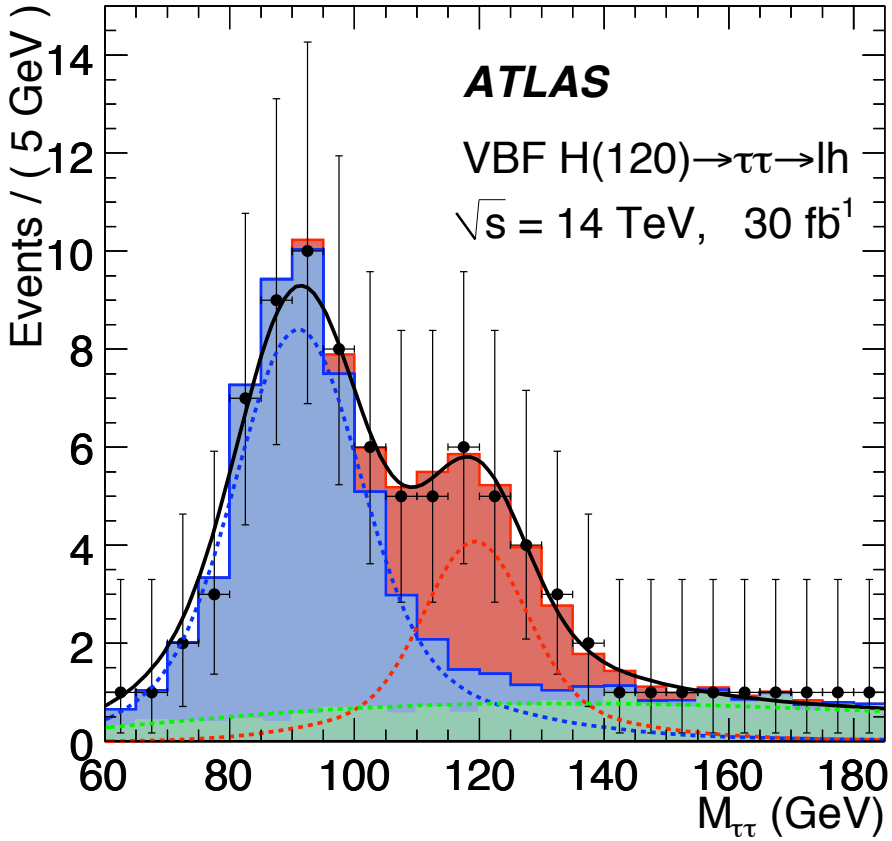
An example

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

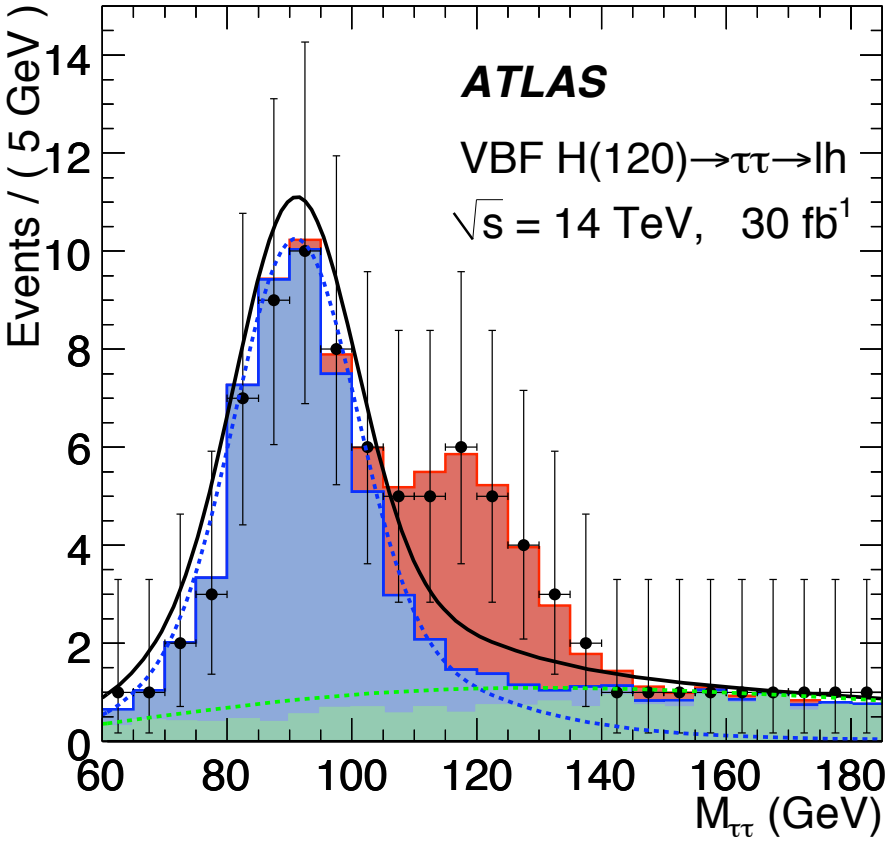
once with everything floating, and once with signal fixed to 0

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

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



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



Properties of the Profile Likelihood Ratio

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!

Toy Monte Carlo

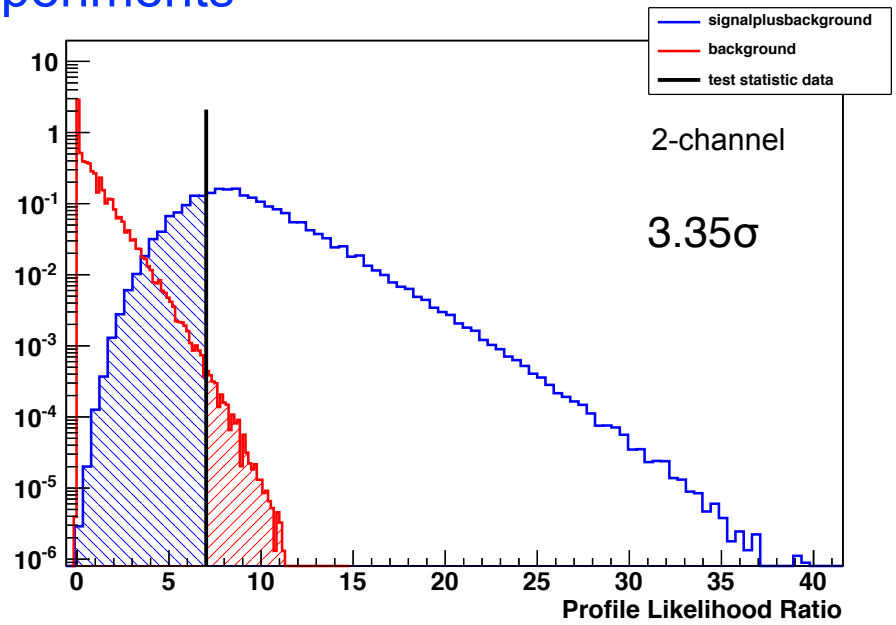
Explicitly build distribution by generating "toys" / pseudo experiments assuming a specific value of μ and ν.

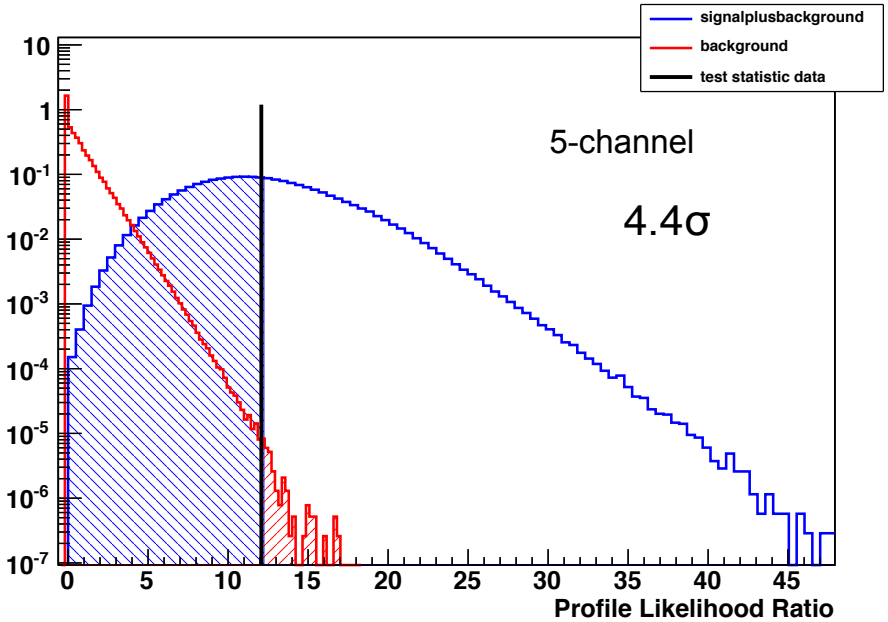
- randomize both main measurements $\mathcal{D}=\{x\}$ and auxiliary measurements $\mathcal{G}=\{\mathbf{a}\}$
- fit the model twice for the numerator and denominator of profile likelihood ratio
- evaluate $-2\ln\lambda(\mu)$ and add to histogram

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



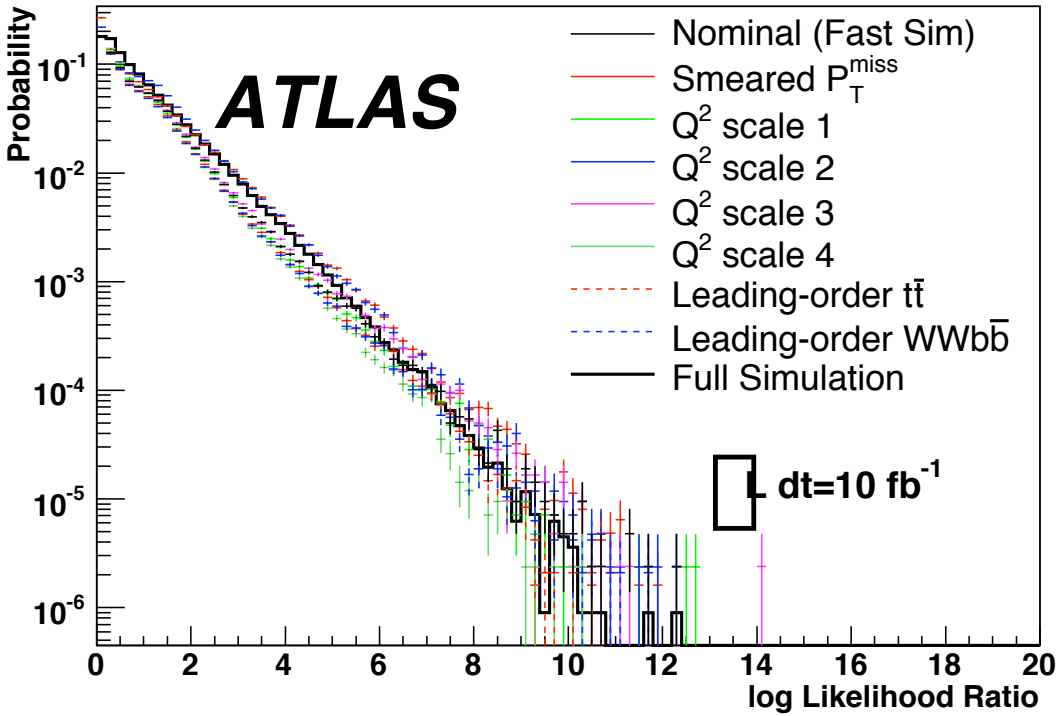


Experimentalist Justification

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.

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 systematics uncertainties (not statistical ones). Statistics can only help us so much after this point. Now we must be physicists!