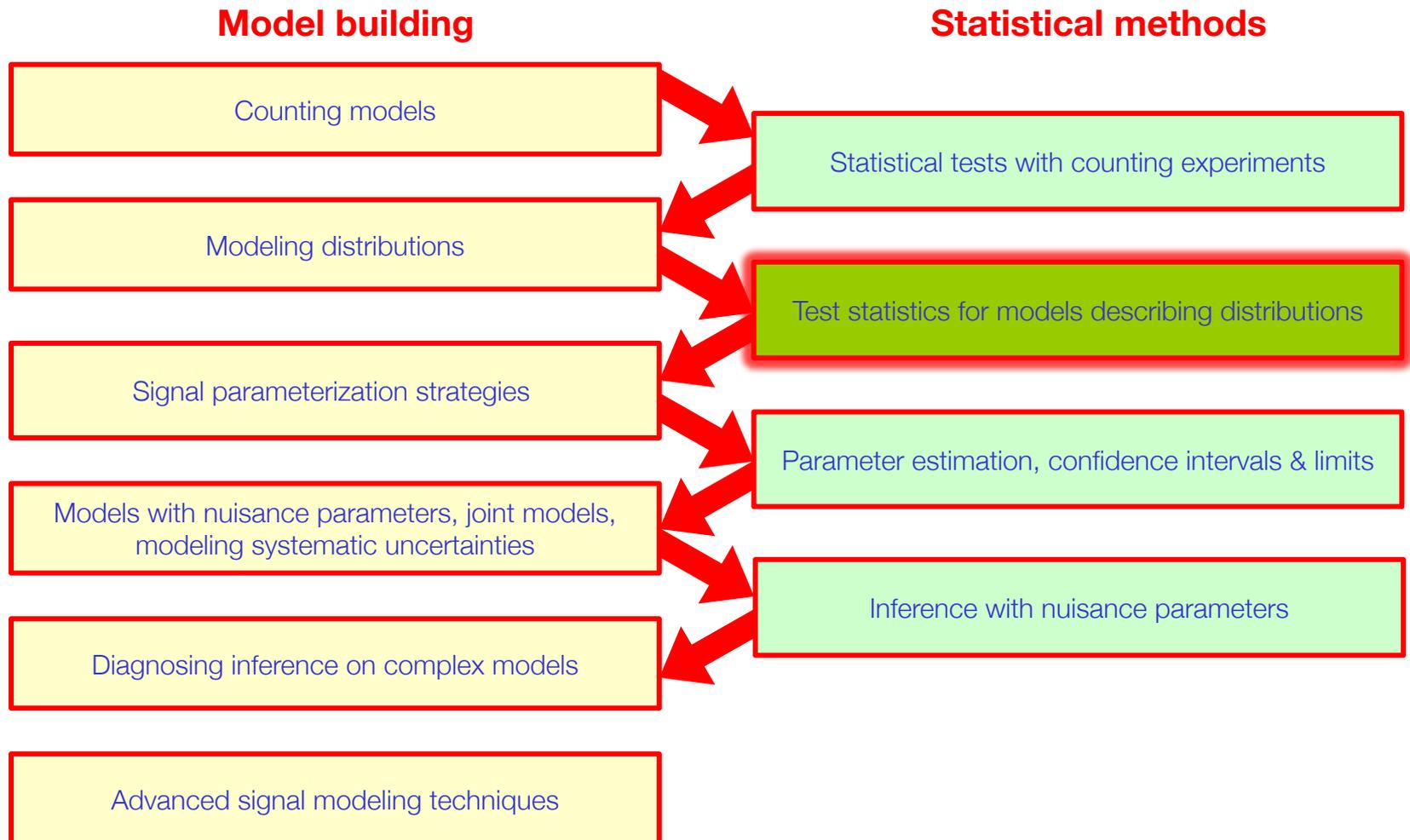


# Statistical methods 2

Adapting statistical methods to use with distributions:  
test statistics as ordering principle, likelihood ratios, contrast with Bayesian methods, the likelihood principle. Practical aspects of toy MC sampling

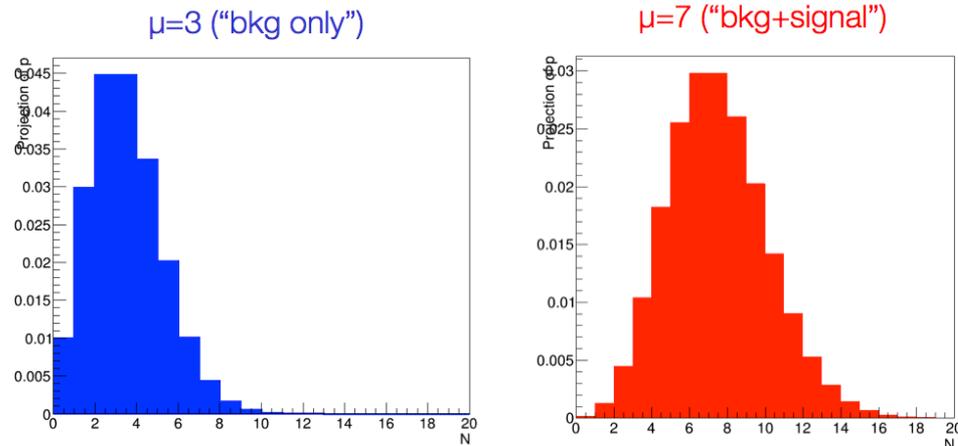
# Roadmap of this course

- Start with basics, gradually build up to complexity



## Summary on statistical test with simple hypotheses

- So far we considered simplest possible experiment we can do: counting experiment
- For a set of 2 or more completely specified (i.e. simple) hypotheses



→ Given probability models  $P(N|\text{bkg})$ , and  $P(N|\text{sig})$   
we can calculate  $P(N_{\text{obs}}|H_x)$  under either hypothesis

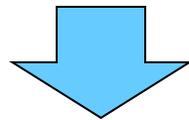
→ With additional information on  $P(H_i)$  we can also calculate  $P(H_x|N_{\text{obs}})$

- In principle, *any potentially complex measurement (for Higgs, SUSY, top quarks) can ultimately take this a simple form.*  
But there is some ‘pre-work’ to get here – examining (multivariate) discriminating distributions → Now try to incorporate that

## Back to $H_b/H_{sb}$ - Formulating evidence for discovery of $H_{sb}$

- Given a scenario with exactly two competing hypotheses
- In the Bayesian school you can cast evidence as an odd-ratio

$$O_{\text{prior}} \equiv \frac{P(H_{sb})}{P(H_b)} = \frac{P(H_{sb})}{1 - P(H_{sb})} \quad \text{If } p(H_{sb})=p(H_b) \rightarrow \text{Odds are 1:1}$$



'Bayes Factor'  $K$  multiplies prior odds

$$O_{\text{posterior}} \equiv \frac{L(x | H_{sb})P(H_{sb})}{L(x | H_b)P(H_b)} = \frac{L(x | H_{sb})}{L(x | H_b)} O_{\text{prior}}$$

If  $P(\text{data}|H_b)=10^{-7}$   
 $P(\text{data}|H_{sb})=0.5$   $K=2.000.000 \rightarrow$  Posterior odds are 2.000.000 : 1

## Formulating evidence for discovery

- In the frequentist school you restrict yourself to  $P(\text{data}|\text{theory})$  and there is no concept of ‘priors’
  - But given that you consider (exactly) 2 competing hypothesis, very low probability for data under  $H_b$  lends credence to ‘discovery’ of  $H_{sb}$  (since  $H_b$  is ‘ruled out’). Example

$$\begin{array}{l} P(\text{data}|H_b)=10^{-7} \\ P(\text{data}|H_{sb})=0.5 \end{array} \quad \rightarrow \quad \text{“}H_b \text{ ruled out”} \rightarrow \text{“Discovery of } H_{sb}\text{”}$$

- Given importance to interpretation of the lower probability, it is customary to quote it in “physics intuitive” form: Gaussian  $\sigma$ .
  - E.g. ‘5 sigma’  $\rightarrow$  probability of 5 sigma Gaussian fluctuation  $=2.87 \times 10^{-7}$
- No formal rules for ‘discovery threshold’
  - Discovery also assumes data is not too unlikely under  $H_{sb}$ . If not, no discovery, but again no formal rules (“your good physics judgment”)
  - NB: In Bayesian case, both likelihoods low reduces Bayes factor  $K$  to  $O(1)$

## Working with Likelihood functions for distributions

- **How do the statistical inference procedures change** for Likelihoods describing distributions?
- Bayesian calculation of  $P(\text{theo}|\text{data})$  they are *exactly the same*.
  - Simply substitute counting model with binned distribution model

$$P(H_{s+b} | \vec{N}) = \frac{L(\vec{N} | H_{s+b})P(H_{s+b})}{L(\vec{N} | H_{s+b})P(H_{s+b}) + L(\vec{N} | H_b)P(H_b)}$$

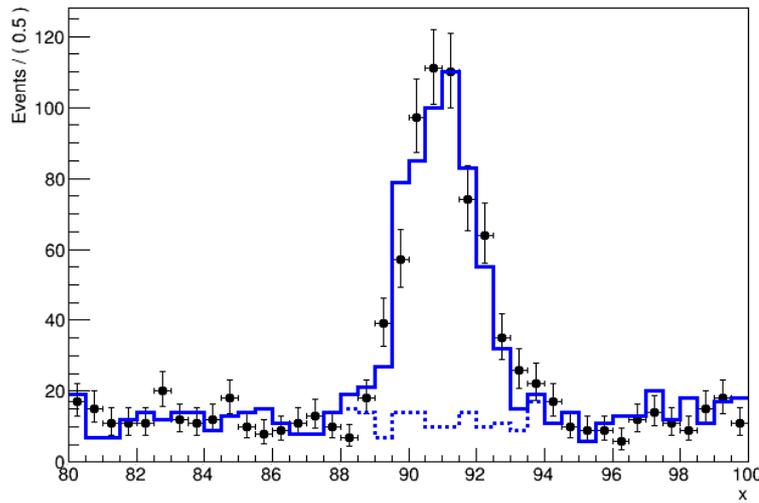


Simply fill in new Likelihood function  
Calculation otherwise unchanged

$$P(H_{s+b} | \vec{N}) = \frac{\prod_i \text{Poisson}(N_i | \tilde{s}_i + \tilde{b}_i)P(H_{s+b})}{\prod_i \text{Poisson}(N_i | \tilde{s}_i + \tilde{b}_i)P(H_{s+b}) + \prod_i \text{Poisson}(N_i | \tilde{b}_i)P(H_b)}$$

## Working with Likelihood functions for distributions

- Frequentist calculation of  $P(\text{data}|\text{hypo})$  also unchanged, but **question arises if  $P(\text{data}|\text{hypo})$  is still relevant?**



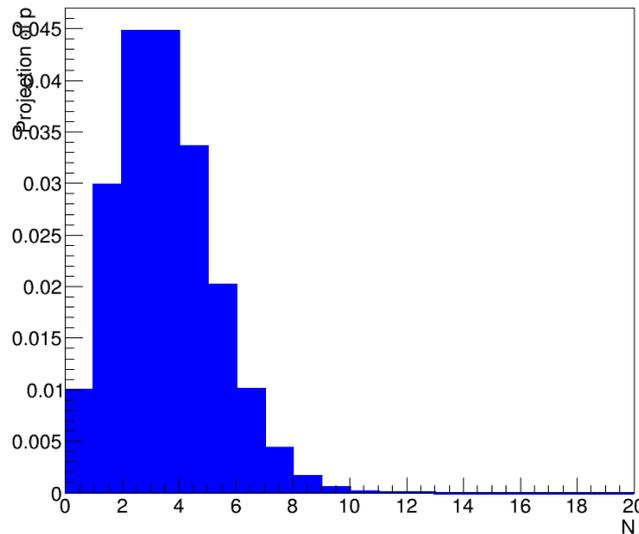
$$L(\vec{N} | H_b) = \prod_i \text{Poisson}(N_i | \tilde{b}_i)$$

$$L(\vec{N} | H_{s+b}) = \prod_i \text{Poisson}(N_i | \tilde{s}_i + \tilde{b}_i)$$

- **$L(N|H)$  is probability to obtain *exactly* the histogram observed.**
- *Is that what we want to know?* Not really.. We are interested in probability to observe any ‘similar’ dataset to given dataset, or in practice dataset ‘similar or more extreme’ than observed data
- **Need a way to quantify ‘similarity’ or ‘extremity’ of observed data**

# Working with Likelihood functions for distributions

- *Definition*: a test statistic  $T(x)$  is *any* function of the data  $x$
- We need a test statistic that will **classify ('order') all possible observations** in terms of 'extremity' (definition to be chosen by physicist)
- NB: For a counting measurement the count itself is already a useful test statistic for such an ordering (i.e.  $T(x) = x$ )



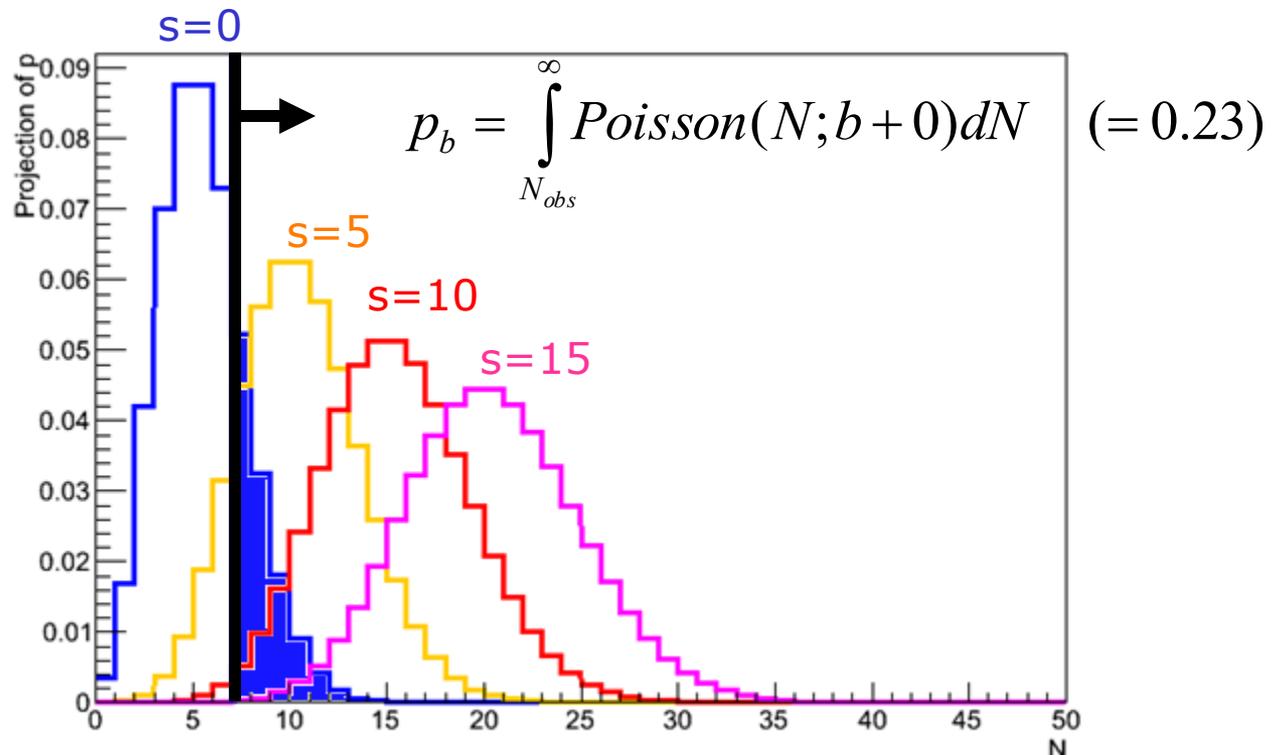
Test statistic  $T(N) = N_{\text{obs}}$  orders observed events count by estimated signal yield

Low  $N \rightarrow$  low estimated signal

High  $N \rightarrow$  large estimated signal

## P-values for counting experiments

- Now make a measurement  $N=N_{\text{obs}}$  (example  $N_{\text{obs}}=7$ )
- **Definition: p-value:**  
probability to obtain the observed data, or more extreme in future repeated identical experiments
  - Example: p-value for background-only hypothesis



# Ordering distributions by 'signal-likeness' aka 'extremity'

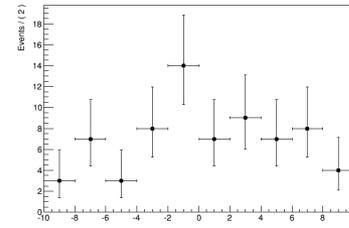
- How to define 'extremity' if observed data is a distribution

Observation

Counting

$$N_{\text{obs}}=7$$

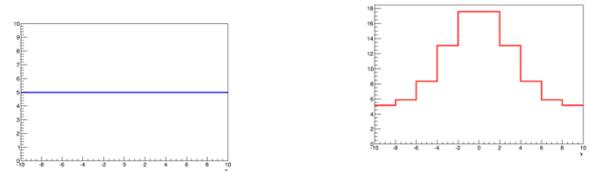
Histogram



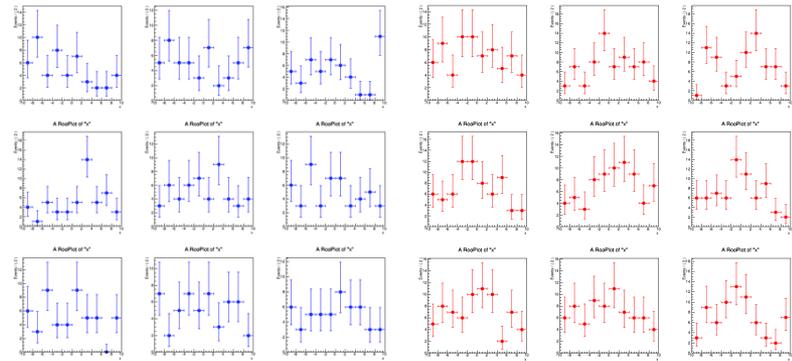
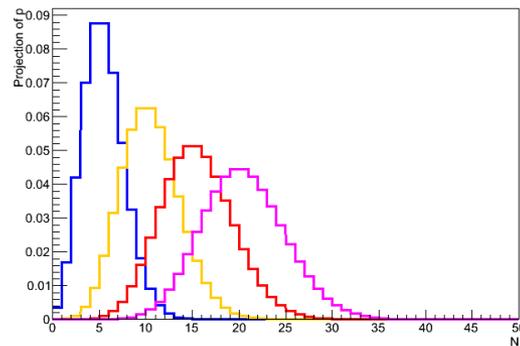
Median expected by hypothesis

$$N_{\text{exp}}(s=0) = 5$$

$$N_{\text{exp}}(s=5) = 10$$



Predicted distribution of observables



Which histogram is more 'extreme'?

## The Likelihood Ratio as a test statistic

- Given two hypothesis  $H_b$  and  $H_{s+b}$  the ratio of likelihoods is a useful test statistic

$$\lambda(\vec{N}) = \frac{L(\vec{N} | H_{s+b})}{L(\vec{N} | H_b)}$$

- Intuitive picture:

→ If data is likely under  $H_b$ ,  
 $L(N|H_b)$  is **large**,  
 $L(N|H_{s+b})$  is smaller

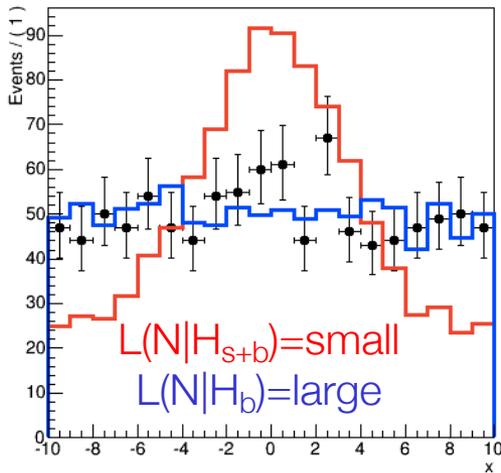
$$\lambda(\vec{N}) = \frac{\text{small}}{\text{large}} = \text{small}$$

→ If data is likely under  $H_{s+b}$ ,  
 $L(N|H_{s+b})$  is **large**,  
 $L(N|H_b)$  is smaller

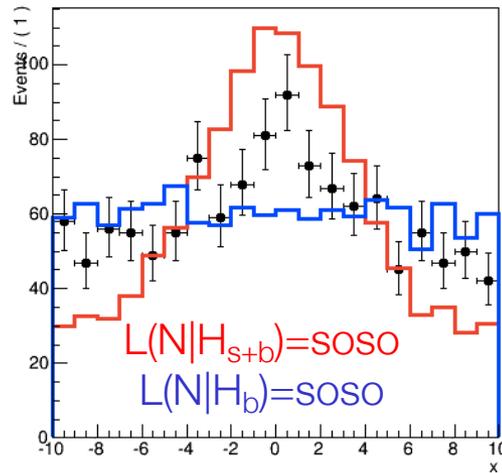
$$\lambda(\vec{N}) = \frac{\text{large}}{\text{small}} = \text{large}$$

# Visualizing the Likelihood Ratio as ordering principle

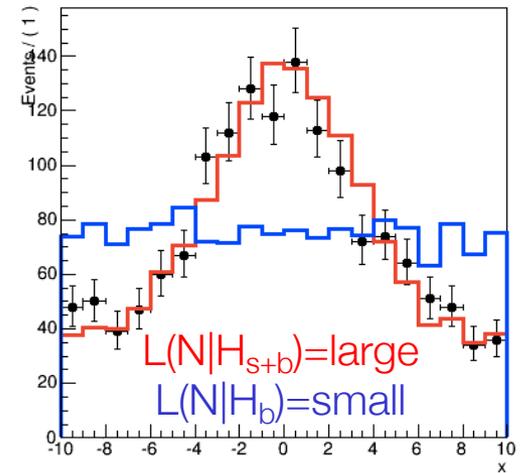
- The Likelihood ratio as ordering principle



$$\lambda(N)=0.0005$$



$$\lambda(N)=0.47$$

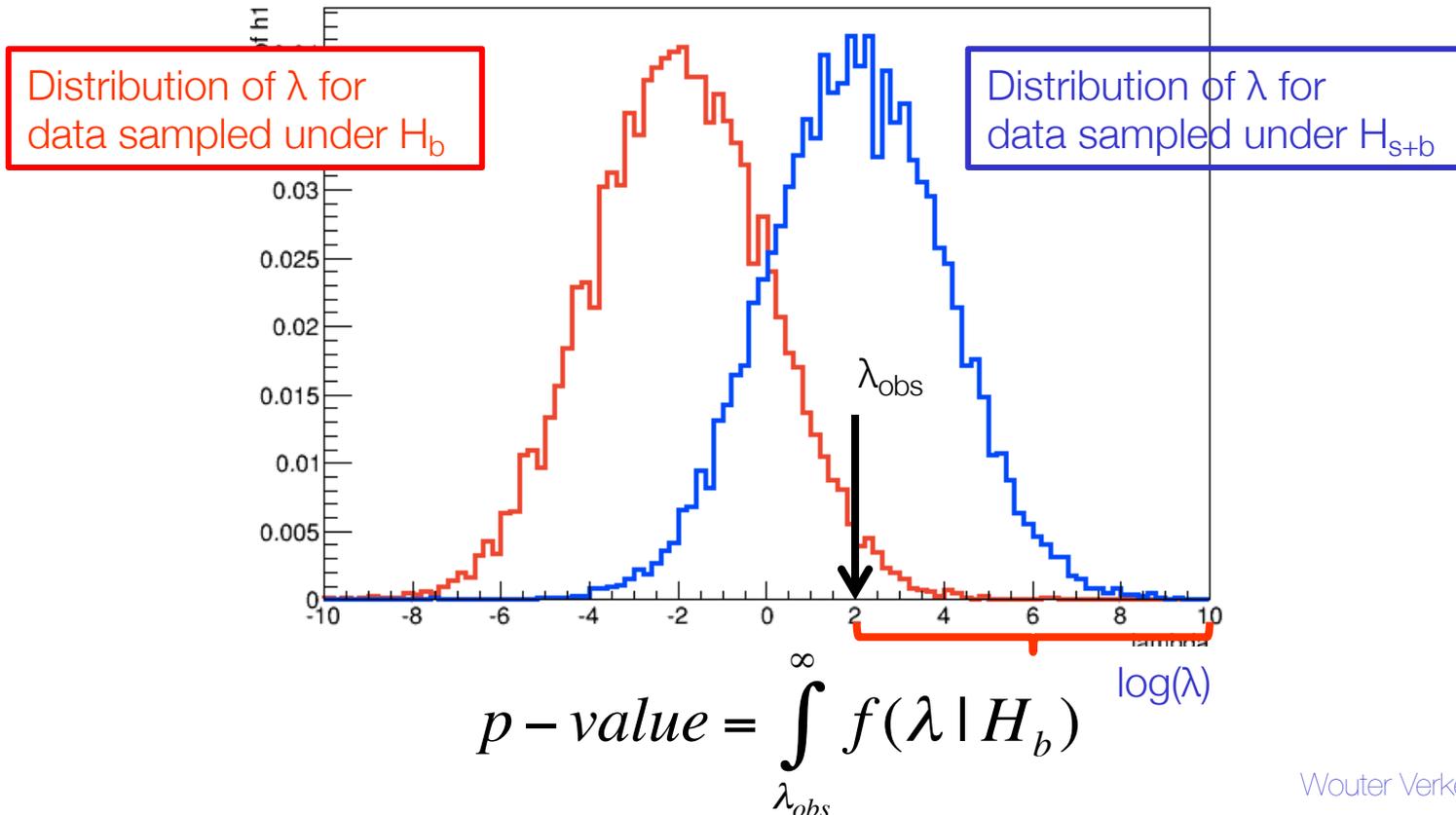


$$\lambda(N)=5000$$

- Frequentist solution to ‘relevance of  $P(\text{data}|\text{theory})$ ’ is to order all observed data samples using a (Likelihood Ratio) test statistic
  - Probability to observe ‘similar data or more extreme’ then amounts to calculating ‘probability to observe test statistic  $\lambda(N)$  as large or larger than the observed test statistic  $\lambda(N_{\text{obs}})$ ’

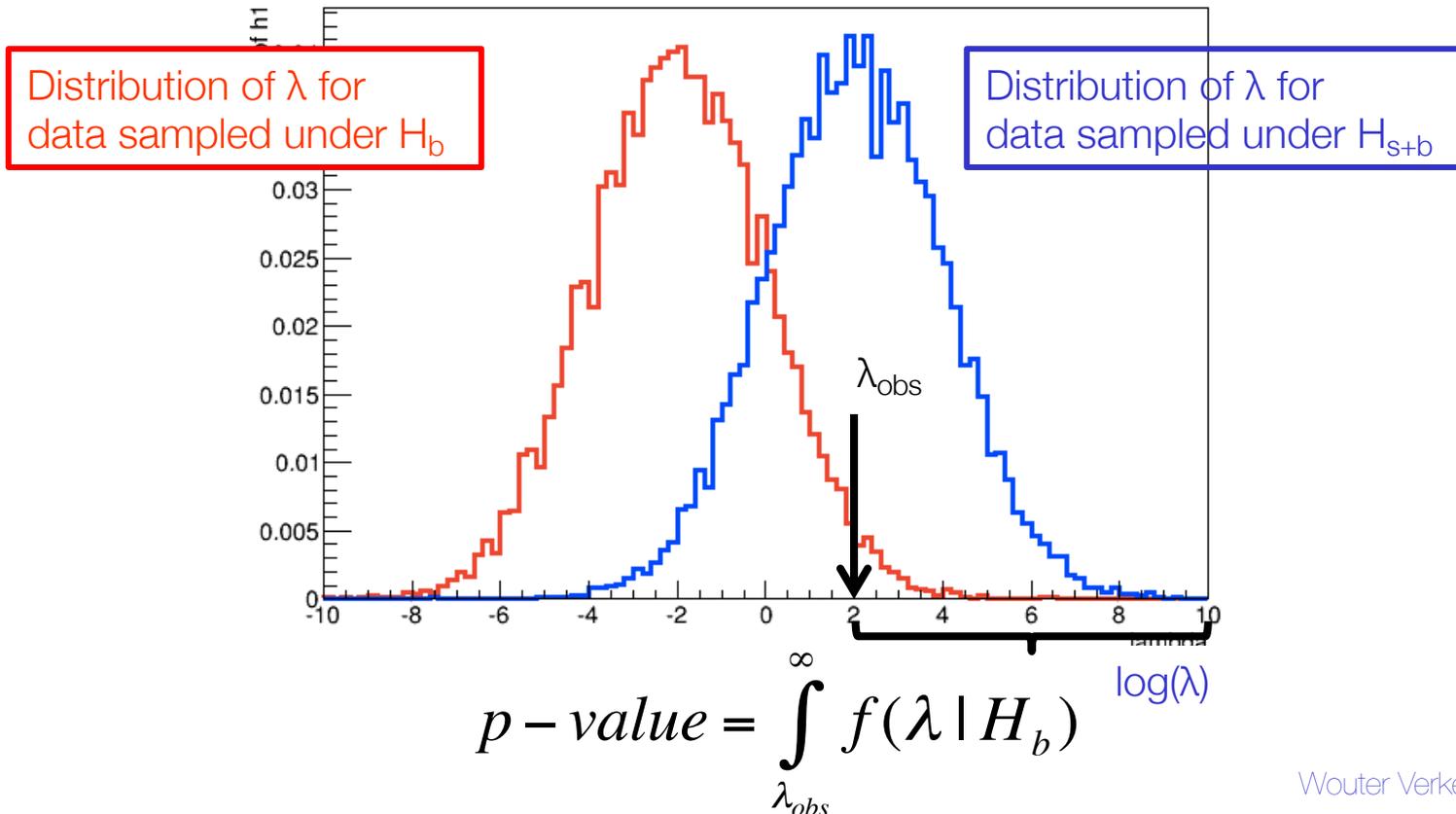
# The distribution of the test statistic

- Distribution of a test statistic is *generally not known*
- Use toy MC approach to approximate distribution
  - Generate many toy datasets  $N$  under  $H_b$  and  $H_{s+b}$  and evaluate  $\lambda(N)$  for each dataset



# The distribution of the test statistic

- **Definition: p-value:**  
probability to obtain the observed data, or more extreme  
in future repeated identical experiments  
(extremity define in the precise sense of the (LR) ordering rule)



# Likelihoods for distributions - summary

- **Bayesian inference unchanged**

→ simply insert L of distribution to calculate  $P(H|\text{data})$

$$P(H_{s+b} | \vec{N}) = \frac{L(\vec{N} | H_{s+b})P(H_{s+b})}{L(\vec{N} | H_{s+b})P(H_{s+b}) + L(\vec{N} | H_b)P(H_b)}$$

- **Frequentist inference procedure *modified***

→ Pure  $P(\text{data}|\text{hypo})$  not useful for non-counting data

→ Order all possible data with a (LR) test statistic in ‘extremity’

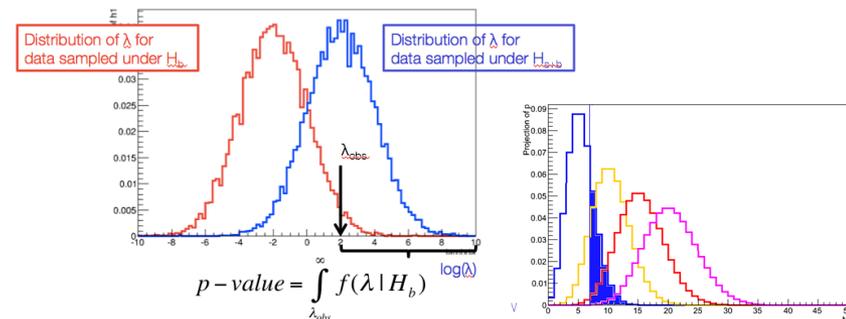
→ Quote  $p(\text{data}|\text{hypo})$  as ‘p-value’ for hypothesis

Probability to obtain observed data, *or more extreme*, is X%

‘Probability to obtain 13 or more 4-lepton events under the no-Higgs hypothesis is  $10^{-7}$ ’

‘Probability to obtain 13 or more 4-lepton events under the SM Higgs hypothesis is 50%’

- **Definition: p-value**



$$p\text{-value} = \int_{\lambda_{obs}}^{\infty} f(\lambda | H_b) \log(\lambda)$$

# The likelihood principle

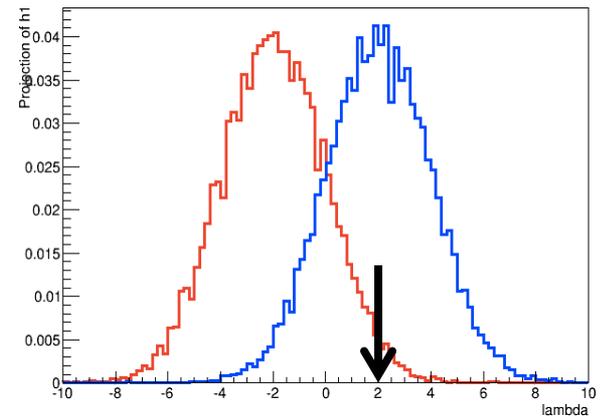
- Note that ‘ordering procedure’ introduced by test statistic also has a profound implication on interpretation
- Bayesian inference only uses the Likelihood of the observed data

$$P(H_{s+b} | \vec{N}) = \frac{L(\vec{N} | H_{s+b})P(H_{s+b})}{L(\vec{N} | H_{s+b})P(H_{s+b}) + L(\vec{N} | H_b)P(H_b)}$$

- While the observed Likelihood Ratio also only uses likelihood of observed data.

$$\lambda(\vec{N}) = \frac{L(\vec{N} | H_{s+b})}{L(\vec{N} | H_b)}$$

- **Distribution  $f(\lambda|N)$ , and thus p-value, also uses likelihood of non-observed outcomes** (in fact Likelihood of every possible outcome is used)



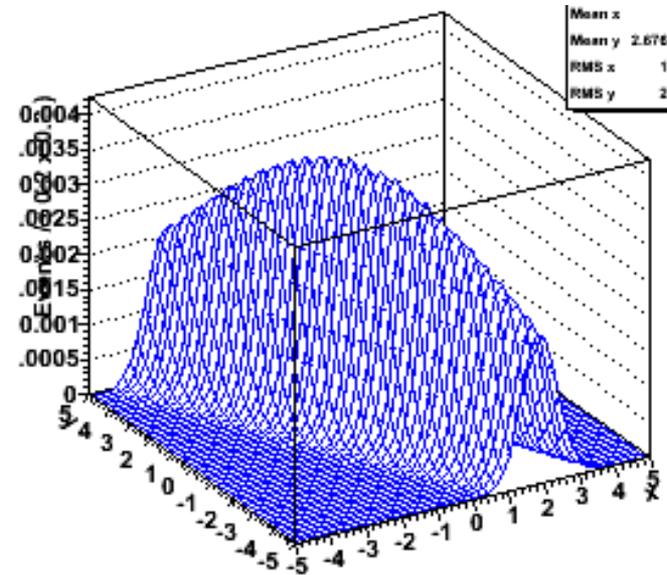
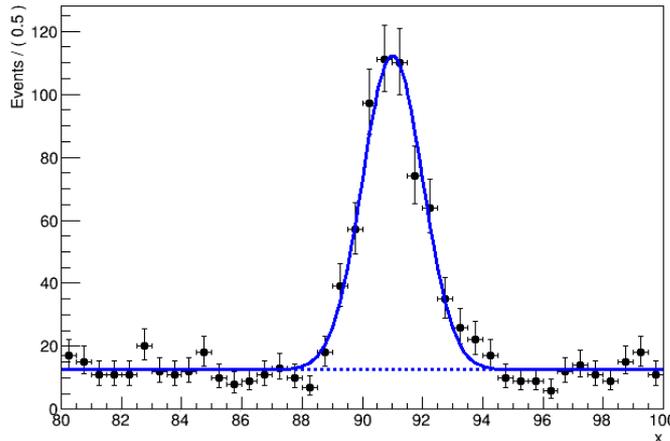
# Likelihood Principle

- In **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.

# Generalizing to multiple dimensions

- Can also generalize likelihood models to distributions in *multiple* observables



$$L(\vec{x}) = \prod_i f(x_i)$$

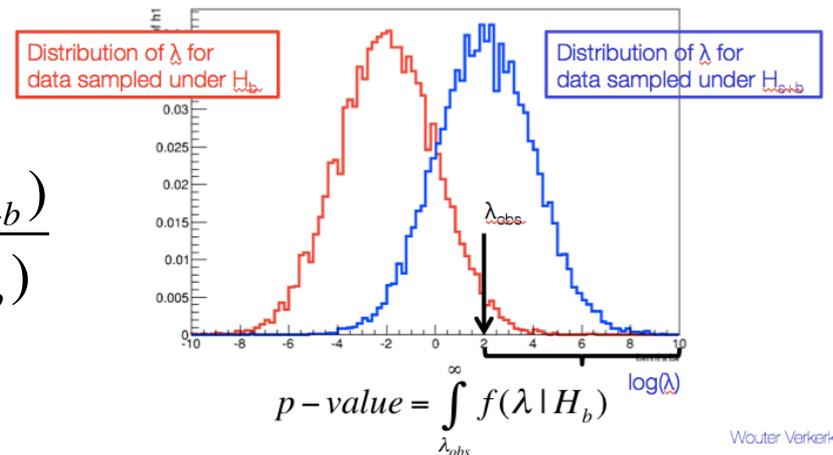
$$L(\vec{x}, \vec{y}) = \prod_i f(x_i, y_i)$$

- Neither generalization (binned  $\rightarrow$  continuous, one  $\rightarrow$  multiple observables) has any further consequences for Bayesian or Frequentist inference procedures

# The Likelihood Ratio test statistic as tool for event selection

- Note that hypothesis testing with two simple hypotheses for observable distributions, exactly describes ‘event selection’ problem
- In fact we have already ‘solved’ the optimal event selection problem! Given two hypothesis  $H_{s+b}$  and  $H_b$  that predict an complex multivariate distribution of observables, **you can always classify all events in terms of ‘signal-likeness’ (a.k.a ‘extremity’) with a likelihood ratio**

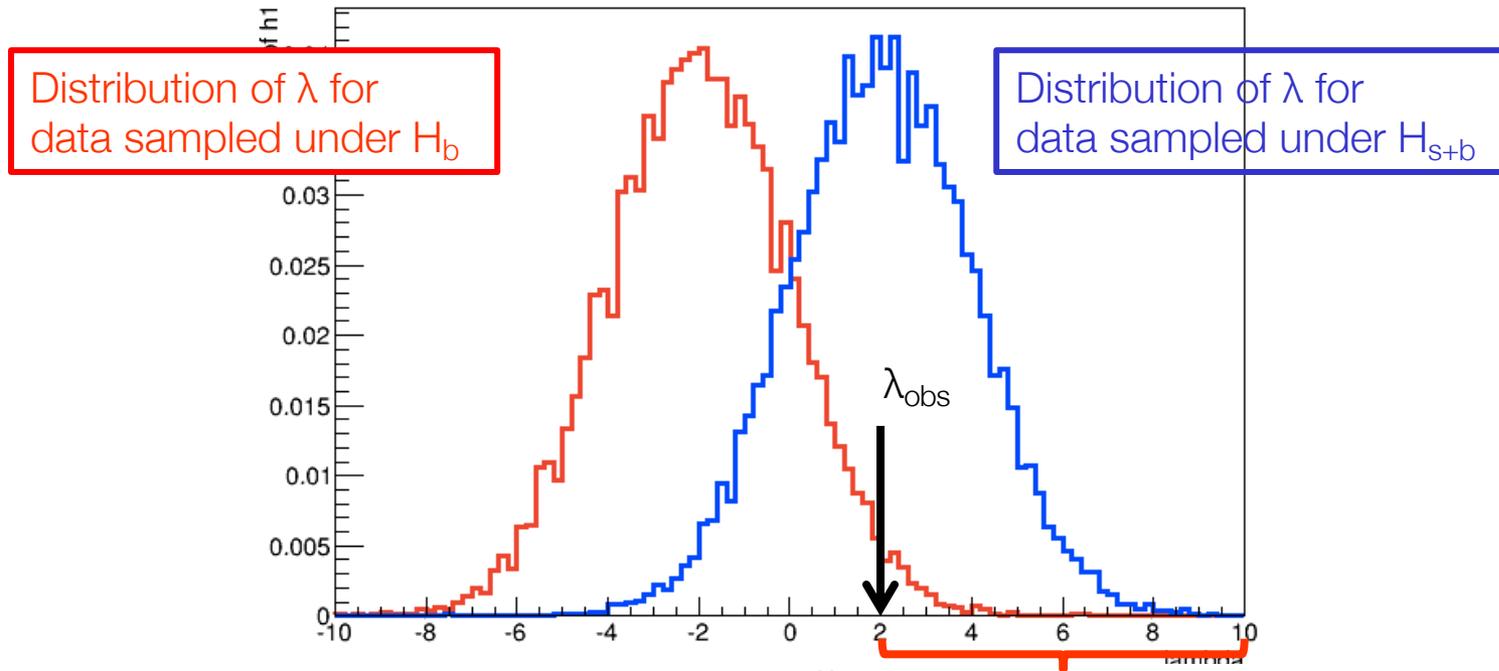
$$\lambda(\vec{x}, \vec{y}, \vec{z}, \dots) = \frac{L(\vec{x}, \vec{y}, \vec{z}, \dots | H_{s+b})}{L(\vec{x}, \vec{y}, \vec{z}, \dots | H_b)}$$



- So far we have exploited  $\lambda$  to calculate a frequentist p-value **now explore properties ‘cut on  $\lambda$ ’ as basis of (optimal) event selection**

# The distribution of the test statistic

- Distribution of a test statistic is *generally not known*
- Use toy MC approach to approximate distribution
  - Generate many toy datasets  $N$  under  $H_b$  and  $H_{s+b}$  and evaluate  $\lambda(N)$  for each dataset



$$p - value = \int_{\lambda_{obs}}^{\infty} f(\lambda | H_b)$$

# Intermezzo – Generating toy data

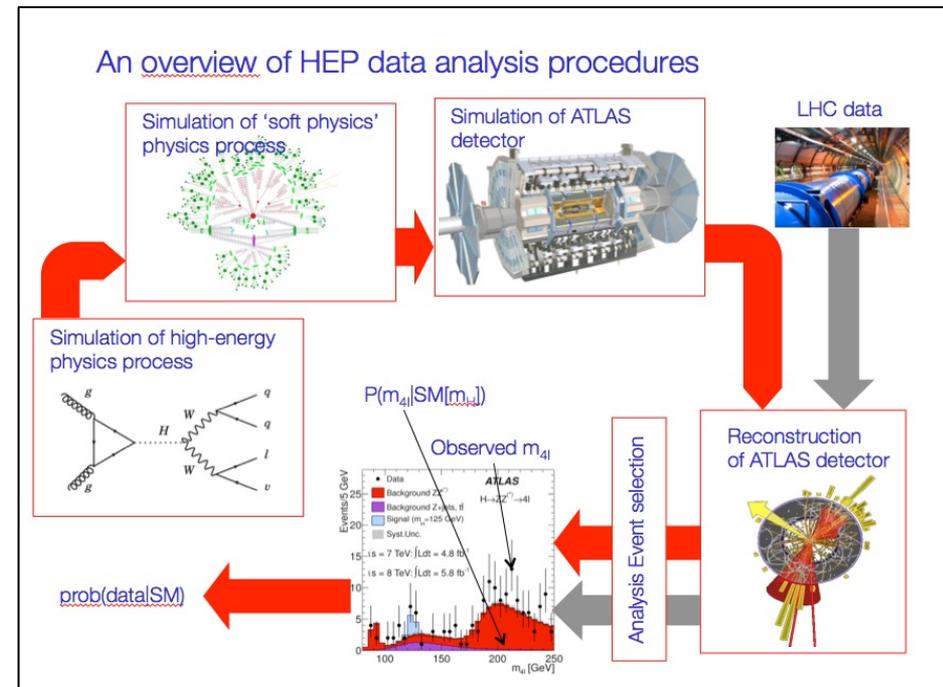
- Two approaches to obtaining simulated data

- First approach is ‘Physics Monte Carlo Chain’, described earlier

- Time consuming, but injects detailed knowledge about physics, detector, output is full collision information, and relation to underlying theory details

- Alternative approach is sample sampling the probability model ‘toy MC’

- Fast (generally), only requires access to probability model
- Can only produce datasets with observables that are described by the probability model → Sufficient to study distribution of test statistics

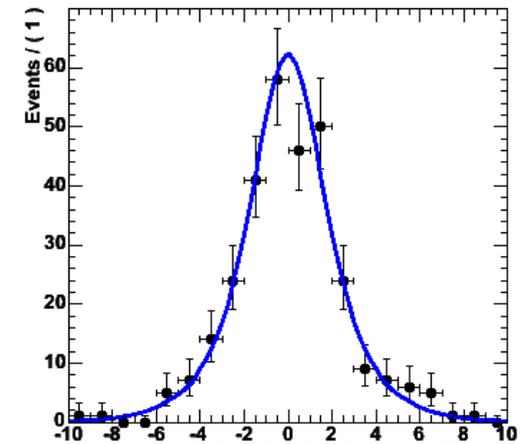
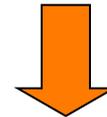
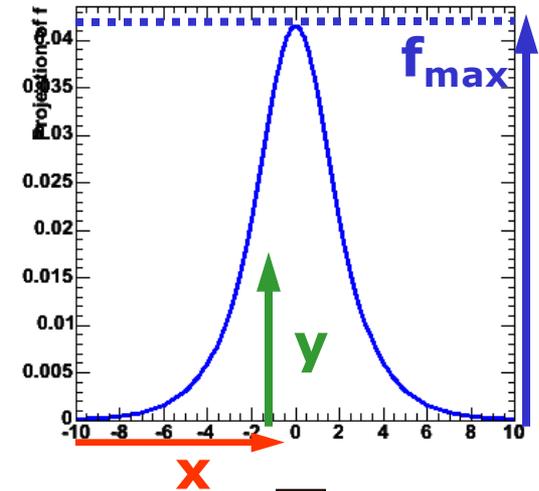


# How do you efficiently generate a toy dataset from a probability model?

- Simplest method is accept/reject sampling

- 1) Determine maximum of function  $f_{\max}$
- 2) Throw random number  $x$
- 3) Throw another random number  $y$
- 4) If  $y < f(x)/f_{\max}$  keep  $x$ ,  
otherwise return to step 2)

- PRO: Easy, always works
- CON: It can be inefficient if function is strongly peaked.  
Finding maximum empirically through random sampling can be lengthy in  $>2$  dimensions

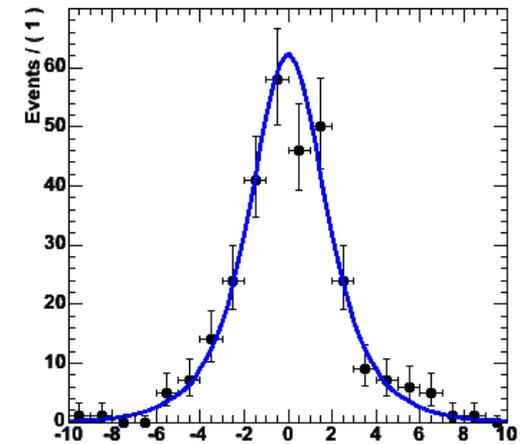
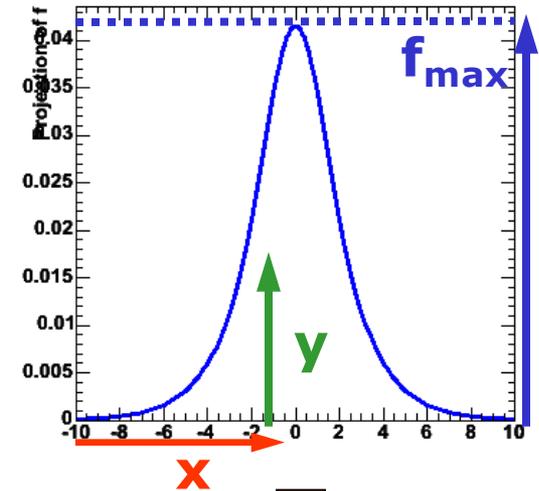


# How do you efficiently generate a toy dataset from a probability model?

- Simplest method is accept/reject sampling

- 1) Determine maximum of function  $f_{\max}$
- 2) Throw random number  $x$
- 3) Throw another random number  $y$
- 4) If  $y < f(x)/f_{\max}$  keep  $x$ , otherwise return to step 2)

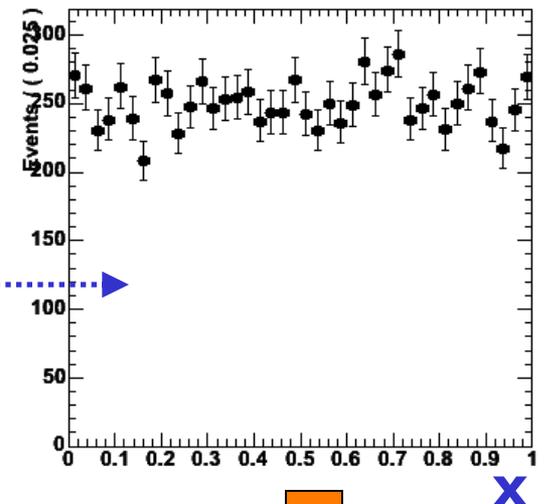
- PRO: Easy, always works
- CON: It can be inefficient if function is strongly peaked.  
Finding maximum empirically through random sampling can be lengthy in  $>2$  dimensions



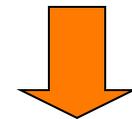
# Toy MC generation – Inversion method

- Fastest: function inversion

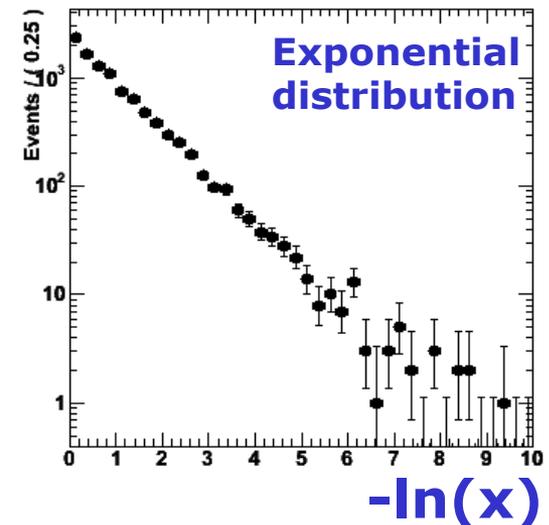
- 1) Given  $f(x)$  find inverted function  $F(x)$  so that  $f(F(x)) = x$
- 2) Throw uniform random number  $x$
- 3) Return  $F(x)$



Take  $-\log(x)$



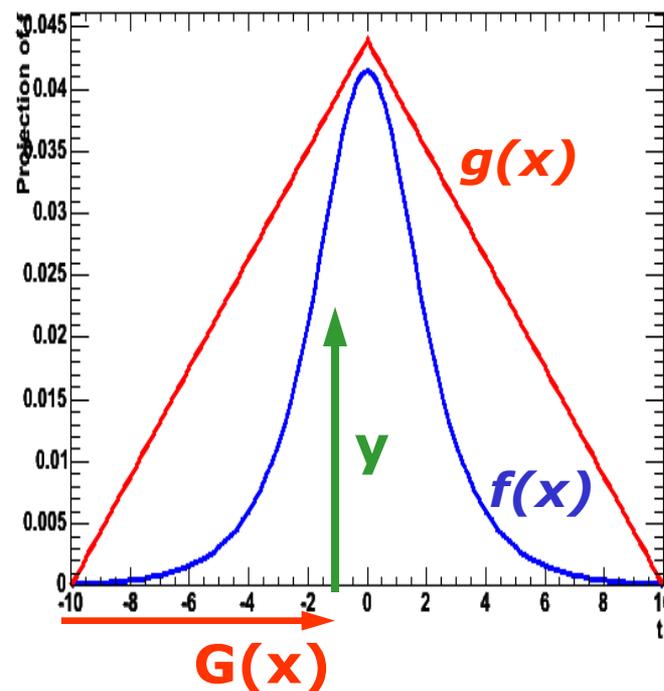
- PRO: Maximally efficient
- CON: Only works for invertible functions



# Toy MC Generation – importance sampling

- Hybrid: Importance sampling

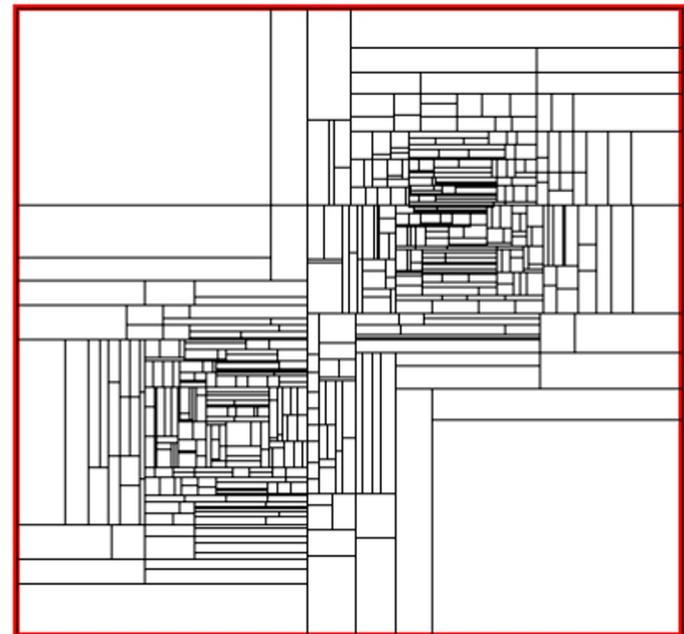
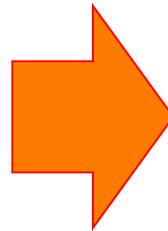
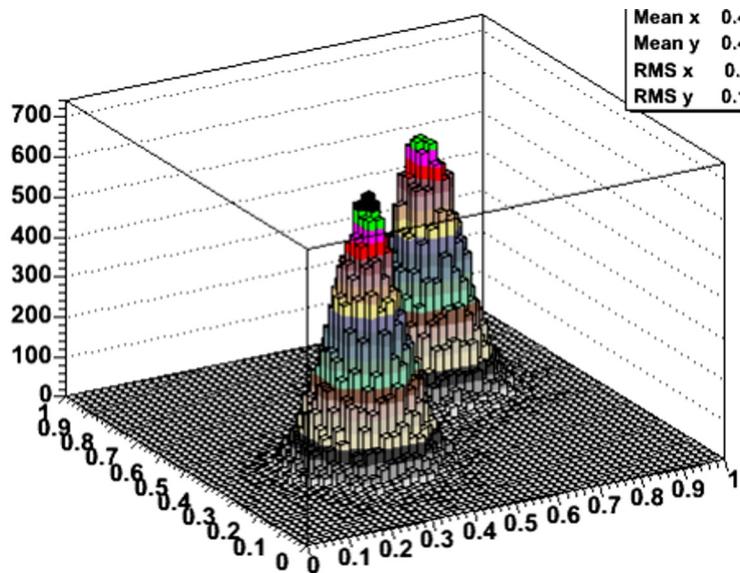
- 1) Find ‘envelope function’  $g(x)$  that is invertible into  $G(x)$  and that fulfills  $g(x) \geq f(x)$  for all  $x$
- 2) Generate random number  $x$  from  $G$  using inversion method
- 3) Throw random number ‘ $y$ ’
- 4) If  $y < f(x)/g(x)$  keep  $x$ , otherwise return to step 2



- PRO: Faster than plain accept/reject sampling  
Function does not need to be invertible
- CON: Must be able to find invertible envelope function

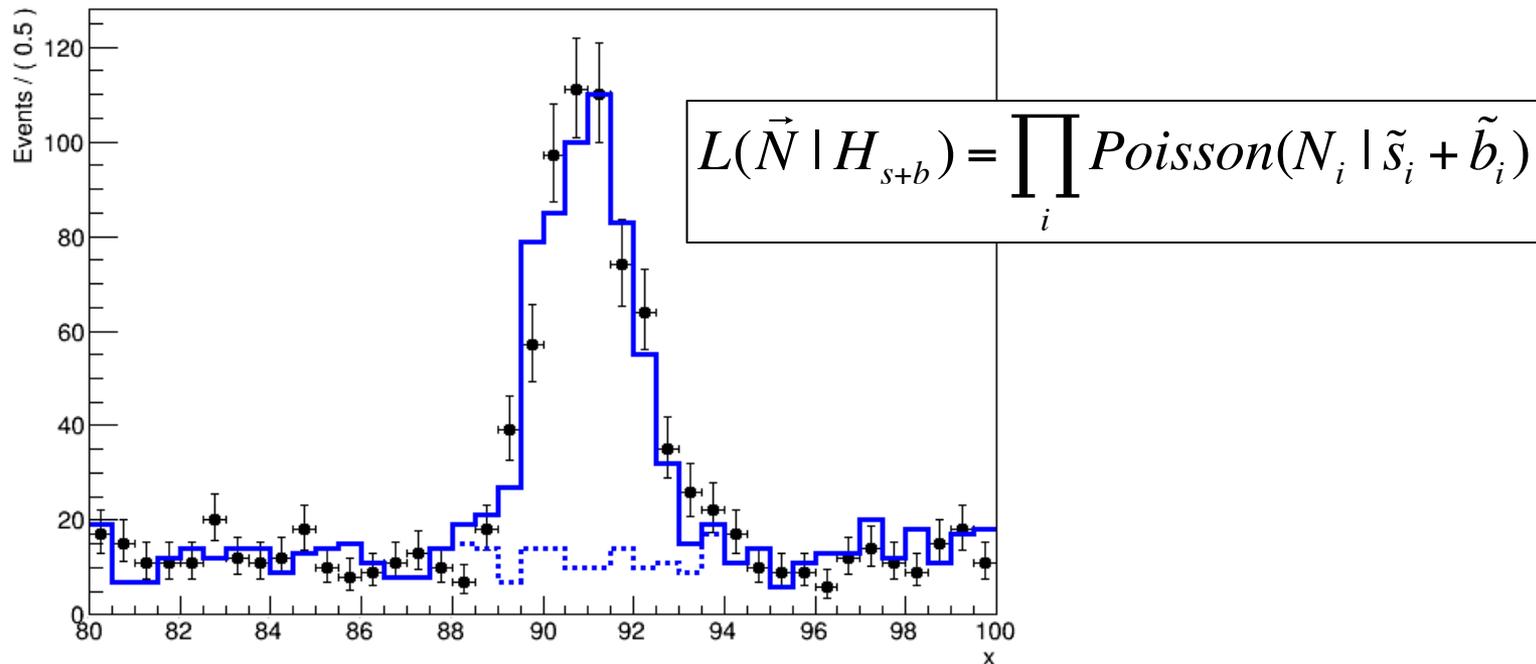
# Toy MC Generation – importance sampling in >1D

- General algorithms exists that can construct empirical envelope function
  - Divide observable space recursively into smaller boxes and take uniform distribution in each box
  - Example shown below from FOAM algorithm



# Toy MC Generation – importance sampling in >1D

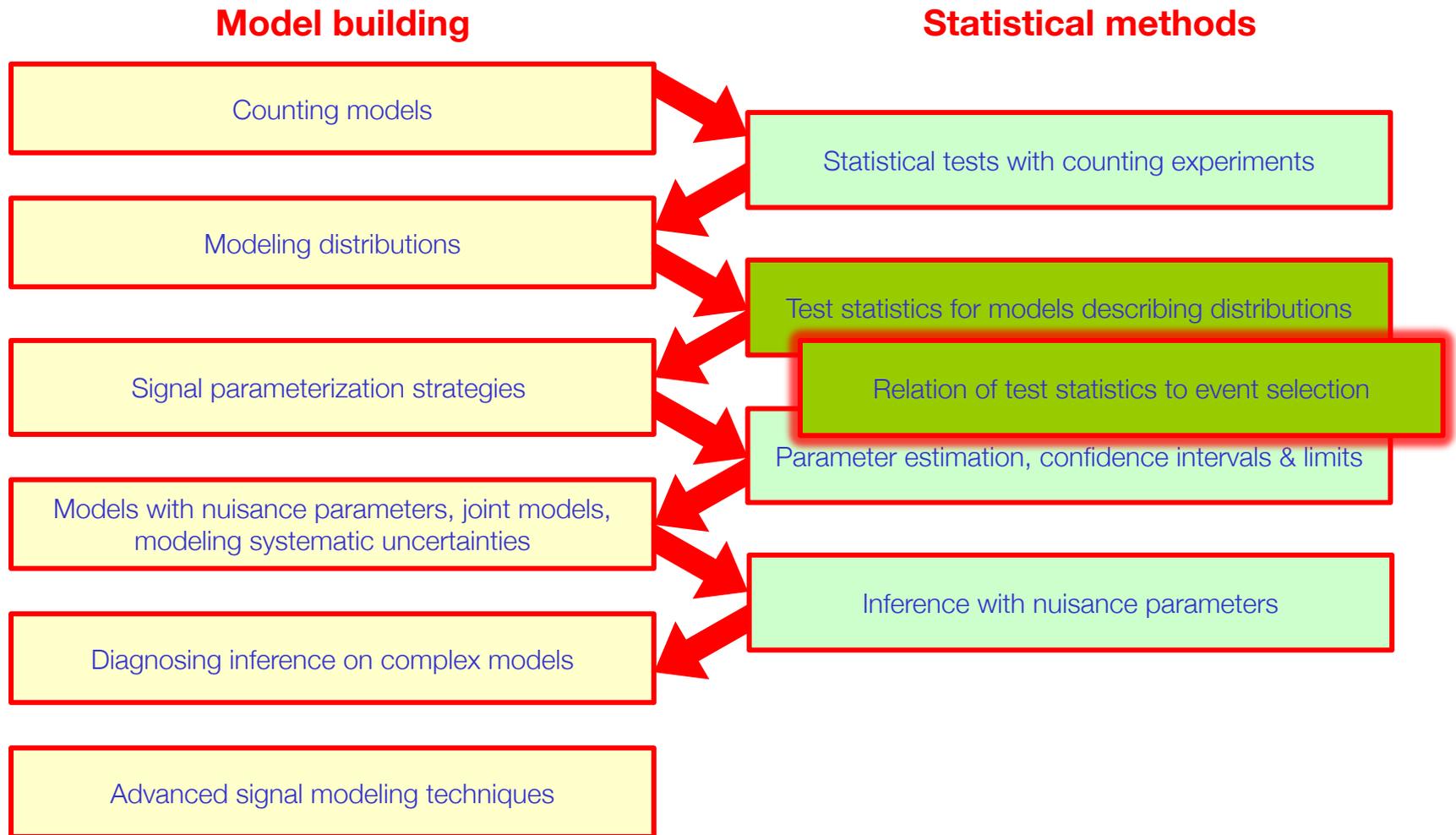
- For *binned distributions*, can generate content of each bin on toy dataset independently, using a Poisson process



- Note that efficient generation of Poisson random number relies on combination of importance sampling (for small  $\mu$ , using exponential envelope, for large  $\mu$  using Cauchy distribution)

# Roadmap of this course

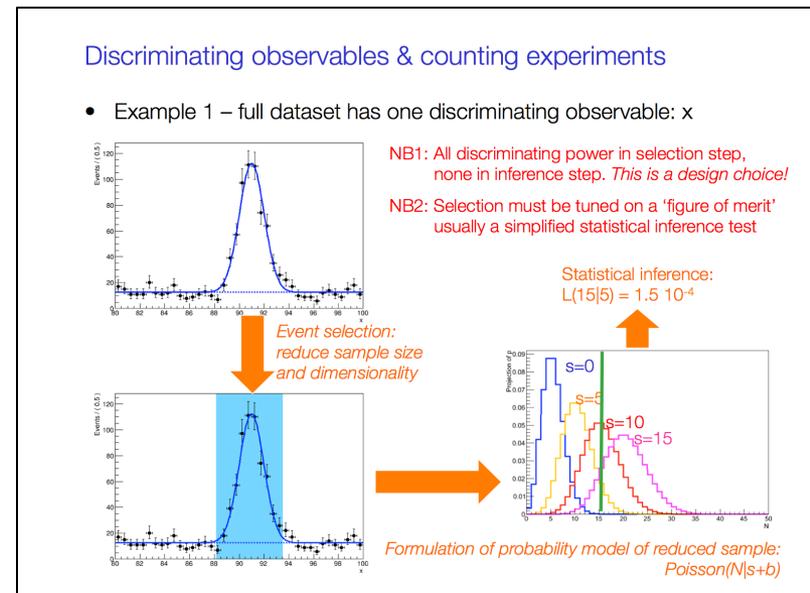
- Start with basics, gradually build up to complexity



# Deciding on a split

- HEP data analysis often a 2-step process:

first selection,  
then inference



- Focus in this course on inference, but Likelihood Ratio as test statistics shows that there is a **general optimal solution for any event selection problem**: the ratio will order all event by signal-likeness

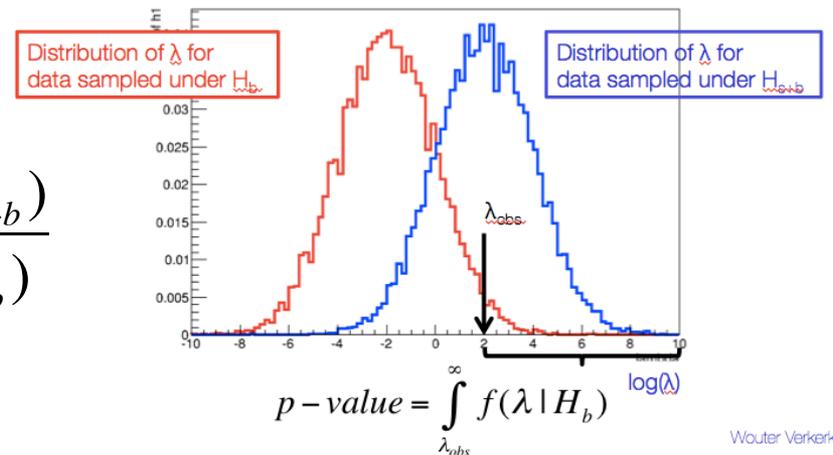
$$\lambda(\vec{x}, \vec{y}, \vec{z}, \dots) = \frac{L(\vec{x}, \vec{y}, \vec{z}, \dots | H_{s+b})}{L(\vec{x}, \vec{y}, \vec{z}, \dots | H_b)}$$

- Hence if we can construct  $\lambda$ , a selection defined by  $\lambda > \lambda_c$  will always be optimal for some stated level of desired purity

# The Likelihood Ratio test statistic as tool for event selection

- Note that hypothesis testing with two simple hypotheses for observable distributions, exactly describes ‘event selection’ problem
- In fact we have already ‘solved’ the optimal event selection problem! Given two hypothesis  $H_{s+b}$  and  $H_b$  that predict an complex multivariate distribution of observables, **you can always classify all events in terms of ‘signal-likeness’ (a.k.a ‘extremity’) with a likelihood ratio**

$$\lambda(\vec{x}, \vec{y}, \vec{z}, \dots) = \frac{L(\vec{x}, \vec{y}, \vec{z}, \dots | H_{s+b})}{L(\vec{x}, \vec{y}, \vec{z}, \dots | H_b)}$$



- So far we have exploited  $\lambda$  to calculate a frequentist p-value **now explore properties ‘cut on  $\lambda$ ’ as basis of (optimal) event selection**

# Event selection

- The event selection problem:
  - Input: Two classes of events “signal” and “background”
  - Output: Two categories of events “selected” and “rejected”
- Goal: select as many signal events as possible, reject as many background events as possible
- Note that optimization goal as stated is ambiguous.
  - But can choose a well-defined by optimization goal by e.g. fixing desired background acceptance rate, and then choose procedure that has highest signal acceptance.
- Relates to “classical hypothesis testing”
  - Two competing hypothesis (traditionally named ‘null’ and ‘alternate’)
  - Here null = background, alternate = signal

# Terminology of classical hypothesis testing

- Definition of terms
  - Rate of type-I error =  $\alpha$
  - Rate of type-II error =  $\beta$
  - Power of test is  $1-\beta$

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

- Treat hypotheses asymmetrically
  - Null hypo is usually special → Fix rate of type-I error
  - Criminal convictions: Fix rate of unjust convictions
  - Higgs discovery: Fix rate of false discovery
  - Event selection: Fix rate of background that is accepted
- Now can define a well stated goal for optimal testing
  - Maximize the power of test (minimized rate of type-II error) for given  $\alpha$
  - Event selection: Maximize fraction of signal accepted

# The Neyman-Pearson lemma

- In 1932-1938 Neyman and Pearson developed a theory in which one must consider competing hypotheses
  - Null hypothesis ( $H_0$ ) = Background only
  - Alternate hypotheses ( $H_1$ ) = e.g. Signal + Background

and proved that

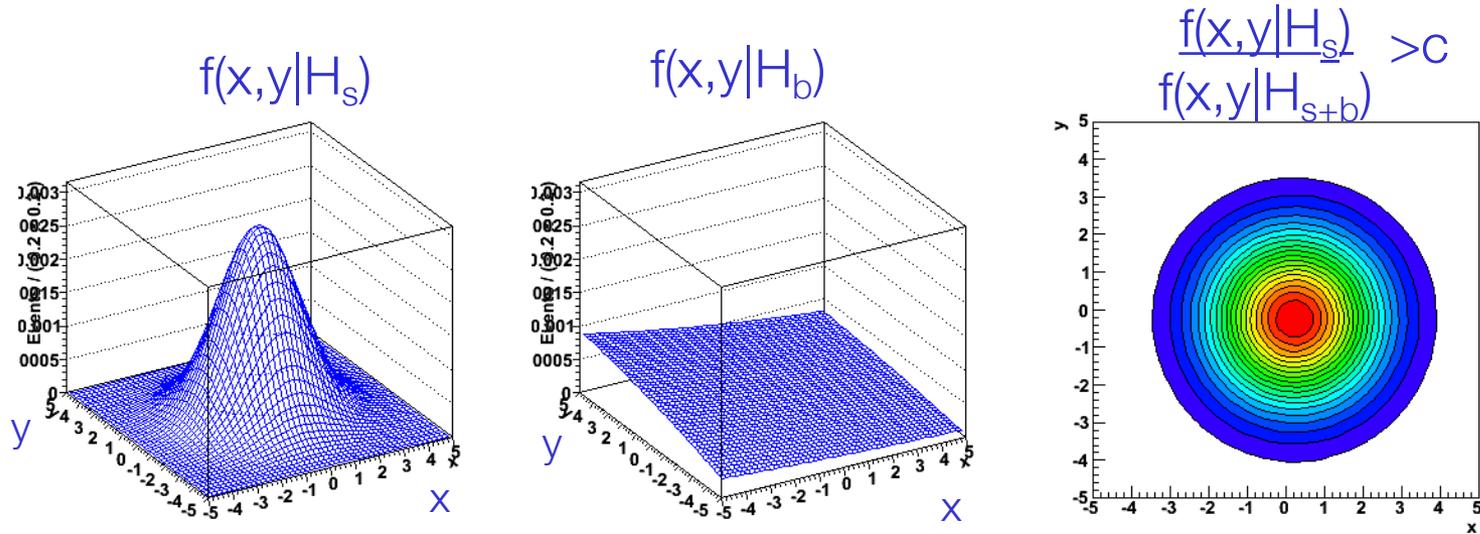
- The region  $W$  that minimizes the rate of the type-II error (not reporting true discovery) is a contour of the Likelihood Ratio

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

- Any other region of the same size will have less power

# The Neyman-Pearson lemma

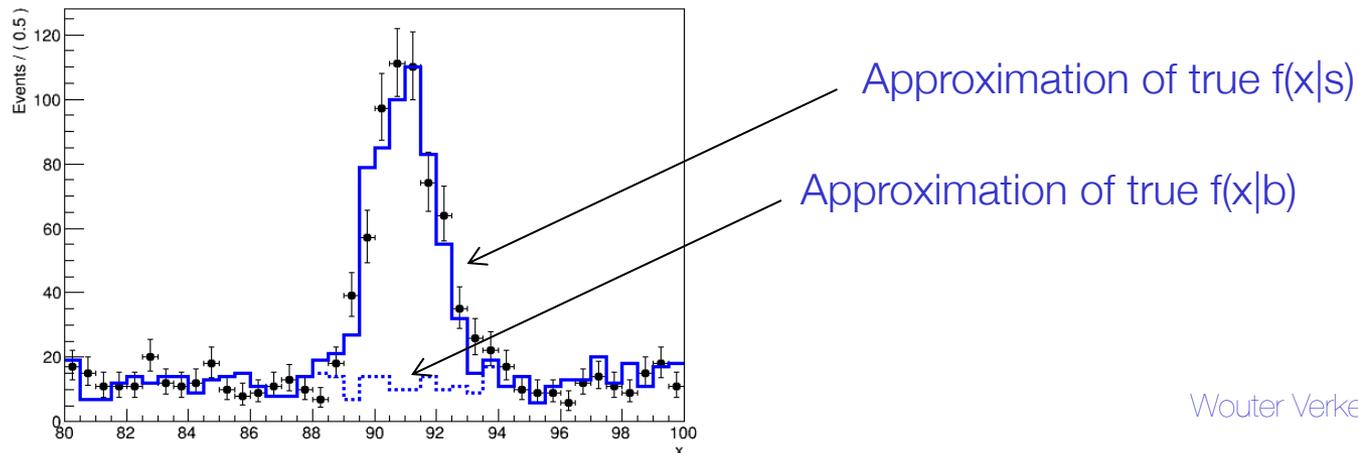
- Example of application of NP-lemma with two observables



- Cut-off value  $c$  controls type-I error rate ('size' = bkg rate)  
Neyman-Pearson: LR cut gives best possible 'power' = signal eff.
- *So why don't we always do this?* (instead of training neural networks, boosted decision trees etc)

# Why Neyman-Pearson doesn't always help

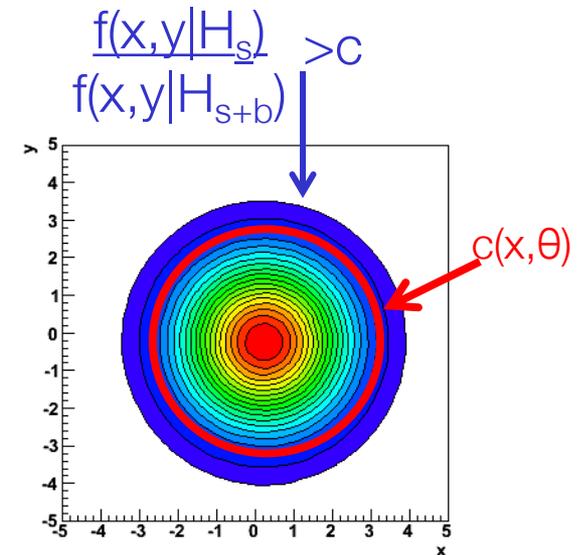
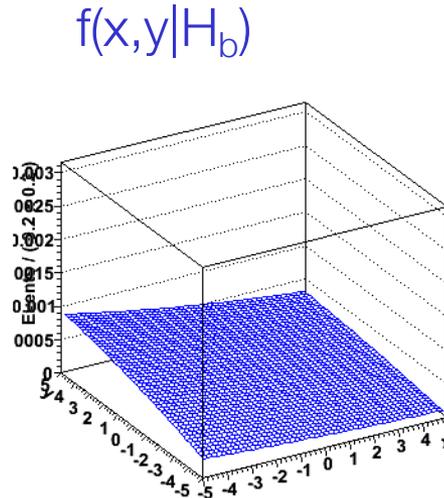
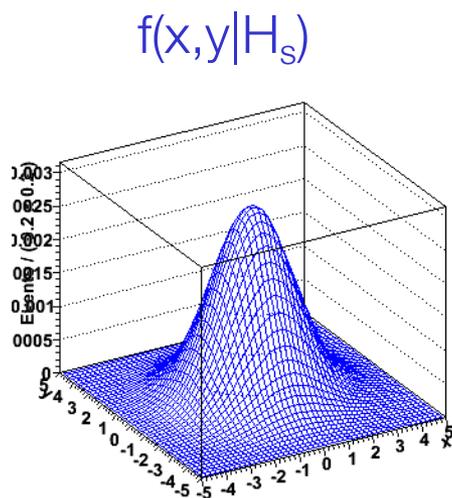
- The problem is that we usually don't have explicit formulae for the pdfs  $f(\vec{x}|\mathbf{s})$ ,  $f(\vec{x}|\mathbf{b})$ .
- Instead we may have Monte Carlo samples for signal and background processes
  - Difficult to reconstruct analytical distributions of pdfs from MC samples, especially if number of dimensions is large
- If physics problem has only few observables can still estimate estimate pdfs with histograms or kernel estimation,
  - But in such cases one can also forego event selection and go straight to hypothesis testing / parameter estimation with all events



# Hypothesis testing with a large number of observables

- When number of observables is large follow different strategy
- Instead of aiming at approximating p.d.f.s  $f(x|s)$  and  $f(x|b)$  aim to approximate decision boundary with an empirical parametric form

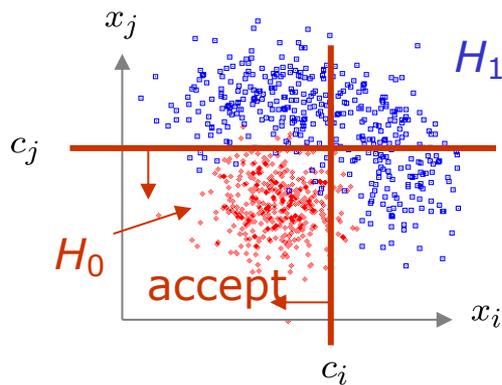
$$A_\alpha(\vec{x}) = \left[ \frac{f(\vec{x}|s)}{f(\vec{x}|s+b)} > \alpha \right] \Rightarrow A_\alpha(\vec{x}) = c(\vec{x}, \vec{\theta})$$



# Empirical parametric forms of decision boundaries

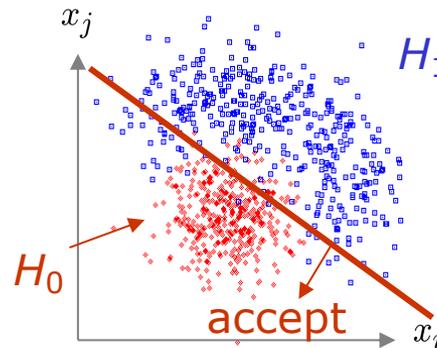
- Can in principle choose any type of Ansatz parametric shape

*Rectangular cut*



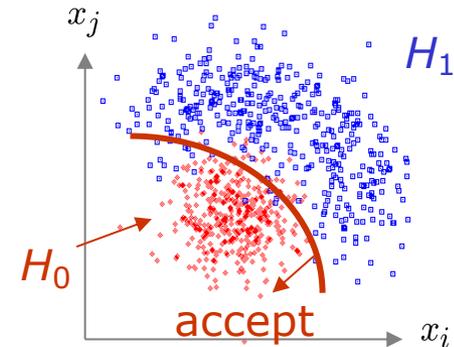
$$t(x) = \theta(x_j - c_j)\theta(x_i - c_i)$$

*Linear cut*



$$t(x) = a_j \cdot x_j + a_i \cdot x_i$$

*Non-linear cut*

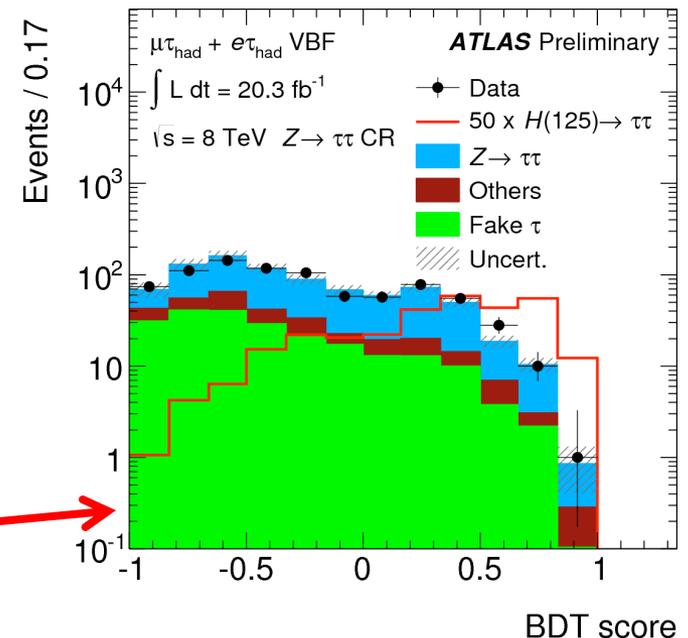


$$t(x) = \vec{a} \cdot \vec{x} + \vec{x}A\vec{x} + \dots$$

- Goal of Ansatz form is estimate of a 'signal probability' for every event in the observable space  $x$  (just like the LR)
- Choice of desired type-I error rate (selected background rate), can be set later by choosing appropriate cut on Ansatz test statistic.

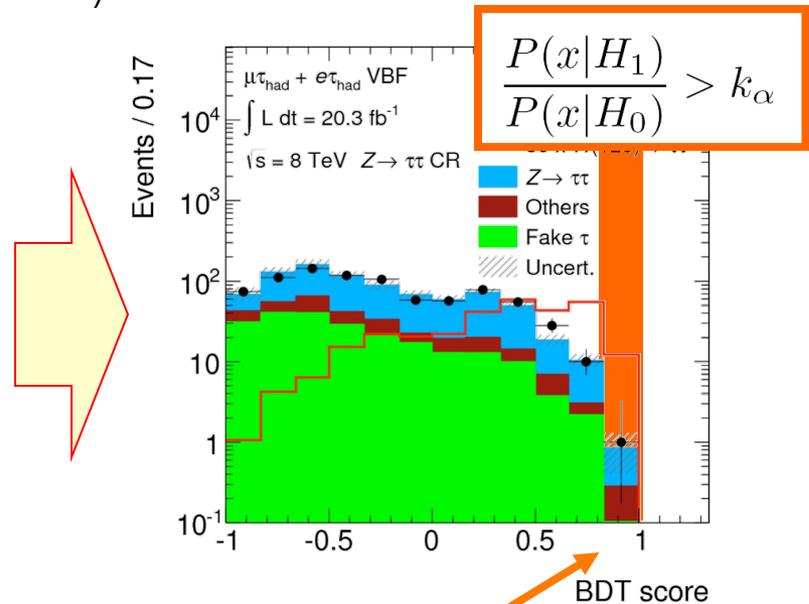
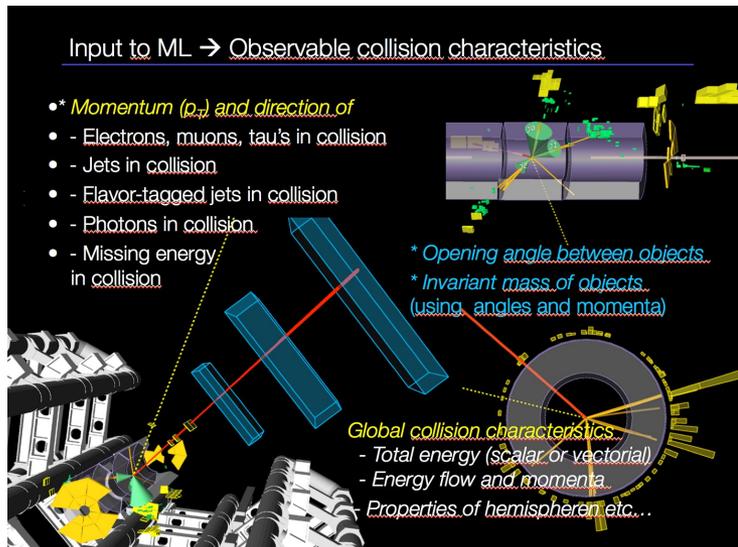
# Machine learning and all that

- A wide range of modern tools exist to perform supervised learning of a multivariate discriminant with the aim to approximate the optimal Neyman-Pearson discriminant.
  - Deep Learning, Boosted Decision Trees, GAN's etc etc.
- Variation in
  - Ansatz (empirical parametric form of discriminant)
  - Learning process (error back propagation, Bayesian)
- Commonality in
  - Input (labeled simulation samples)
  - Output (single function that maps signal probability)
- In all cases output functions is functionally comparable to likelihood ratio discriminant (modulo some trivial transformations)



# Event selection as dimensionality reduction

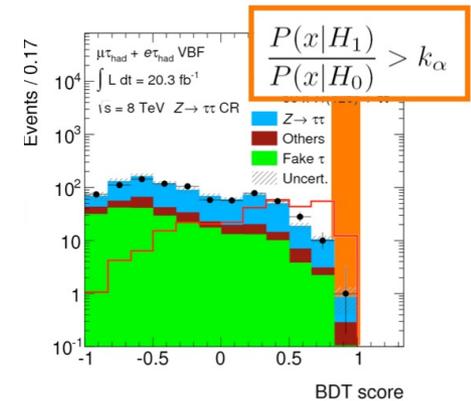
- In the limit of an optimal discriminant – **the event selection step is effectively (and only) a reduction of dimensionality of the data** without loss of information (in the optimal case)



- In case the full discriminant distribution is tested → no loss of information
  - But need for pdf that model distribution
- But can also select high-signal region and perform simplified inference
  - e.g. counting model in that region

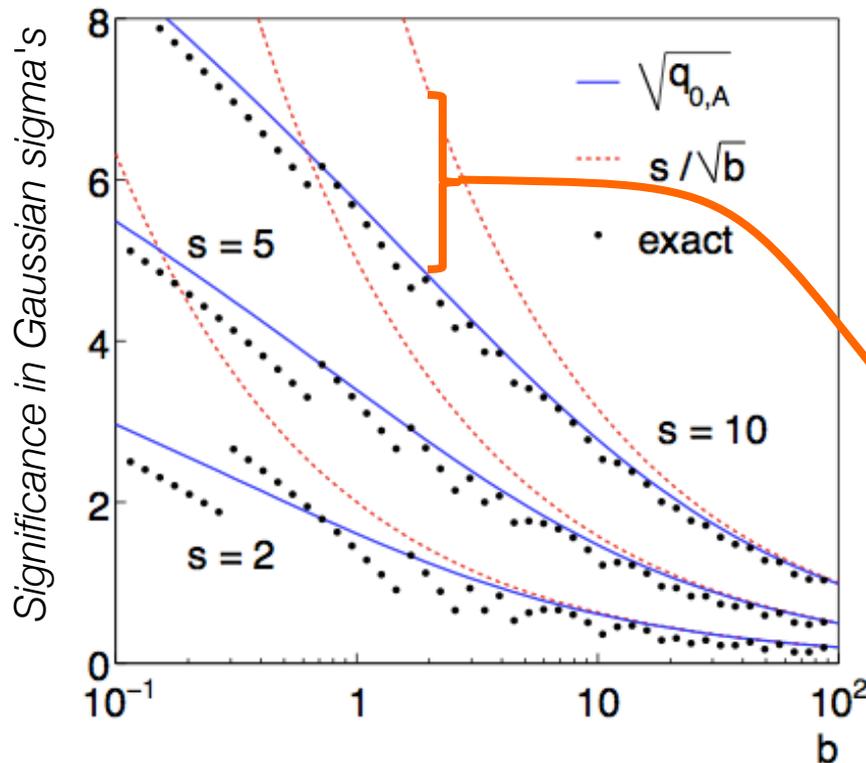
# Choosing the ‘best’ high-signal region

- A common scenario for searches in a low-statistics regime is to perform a simplified analysis
  1. Train MVA to obtain discriminant D
  2. Apply a cut on D
  3. Perform only a counting analysis
- And a common question is then – **what is the ‘optimal cut on D’?**
  - To answer question, a ‘figure of merit’ (FOM) must be chosen that quantifies the optimality of the selection.
  - The FOM for a search is usually the *expected signal significance*.
  - A ‘traditional’ choice is  $FOM = s/\sqrt{b}$ . **For low-statistic searches  $s/\sqrt{b}$  is a bad choice!** It assumes Gaussian distribution, whereas the true distribution is Poisson, which is quite unlike Gaussian especially in the tails at low N
    - A better, and equally easy to use, equation exists based on a Poisson calculation
  - NB: the question arise due to choice for simplified counting in step 3). If a *probability density model* is used for the analysis of the selected data, then the answer is always ‘the full range of the discriminant’



## A better FOM for discovery - the 'Expected Poisson Z'

- The expected counting significance for a Poisson process is analytically calculable:  $\sqrt{2((s+b)\ln(1+s/b) - s)}$ .
- For discovery, the traditional FOM  $s/\sqrt{b}$  *shows significant deviations from the 'exact' expected Poisson significance at low  $b$*



$$\begin{aligned} \sqrt{q_{0,A}} &= \sqrt{2((s+b)\ln(1+s/b) - s)} . \\ &= \frac{s}{\sqrt{b}} (1 + \mathcal{O}(s/b)) . \end{aligned}$$