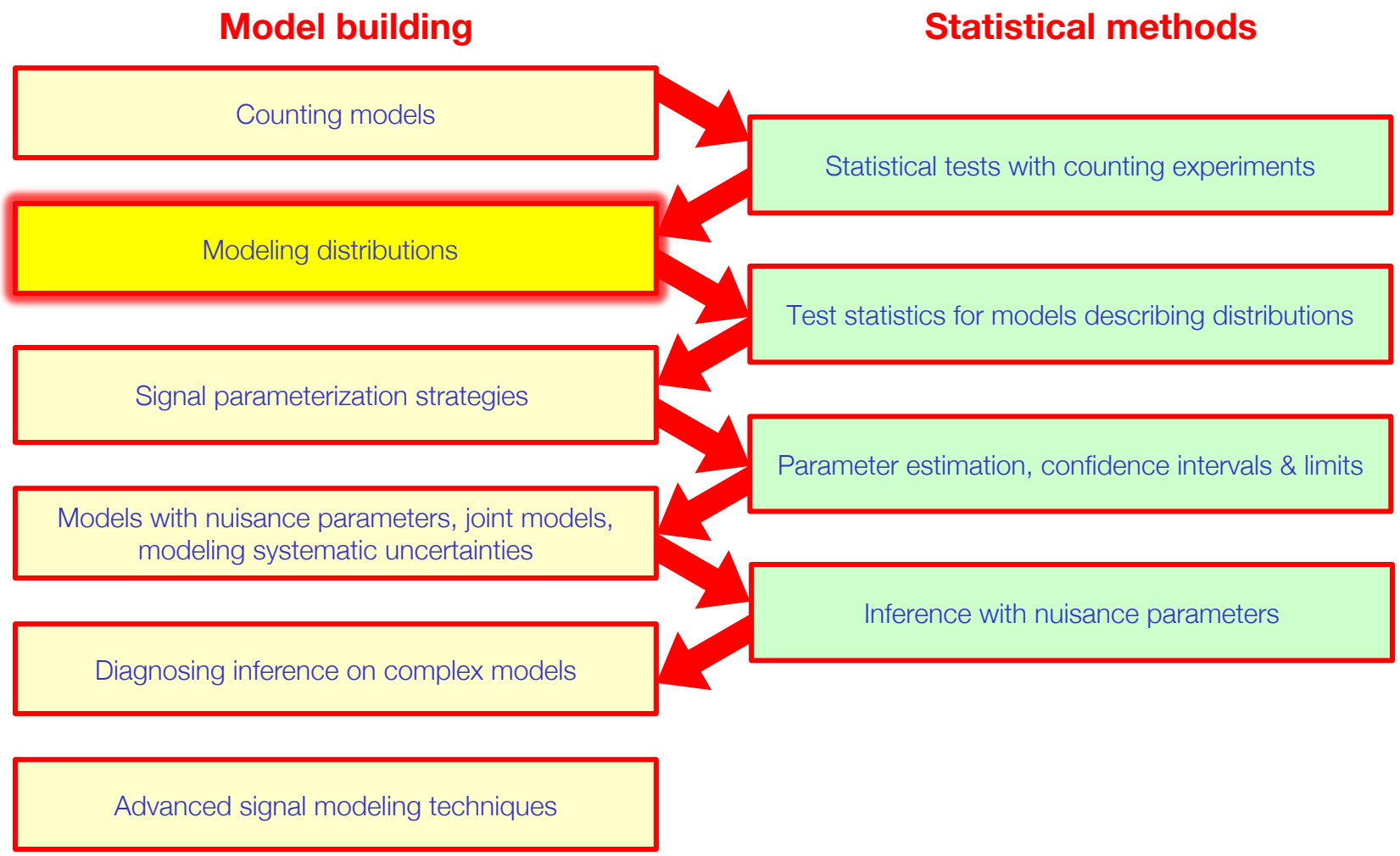
Model building 2

Modelling distributions – template based models or analytical models

Wouter Verkerke, NIKHEF

Roadmap of this course

- Start with basics, gradually build up to complexity

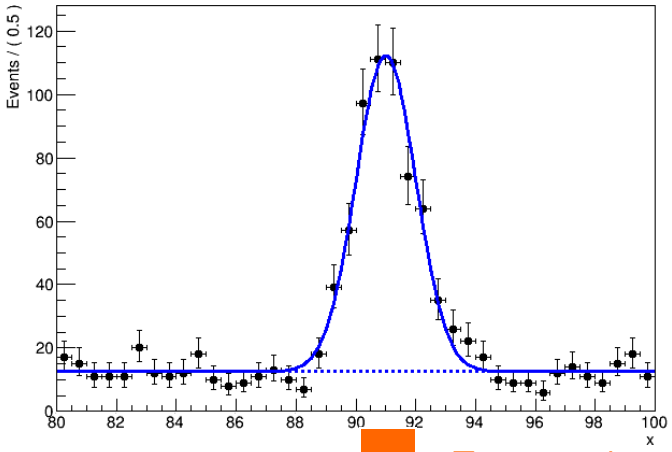


Discriminating observables & counting experiments

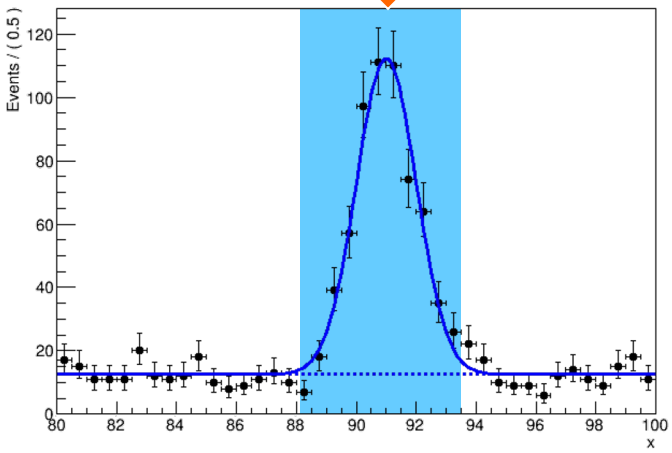
- HEP experimental data usually has many discriminating observables that carry information that can distinguish signal from background hypothesis
- In principle can use them all directly in an elaborate hypothesis test.
  - But would need to formulate a model that describe the expected distribution of all of these → Complicated
  - If expectations are uncertain (from simulation or theory) process of modeling becomes even more complex
- A pragmatic solution to reduce complexity is to split task in two
  - Define empirical selection of events enriched in signal using one or more observable properties of the event (invariant masses, distributions, angles etc)
  - Perform statistical test (hypothesis test, parameter estimation etc) on sample that reduced in size and in dimensionality of discriminating observables that are modeled
  - Most extreme reduction of dimensionality is to zero → counting experiment

Discriminating observables & counting experiments

- Example 1 – **Discrimination in selection stage only**



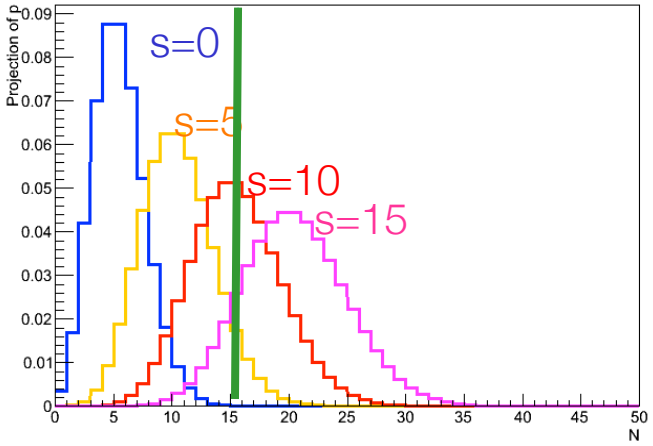
Event selection: reduce sample size and dimensionality



NB1: All discriminating power in selection step, none in inference step. *This is a design choice!*

NB2: Selection must be tuned on a 'figure of merit' usually a simplified statistical inference test

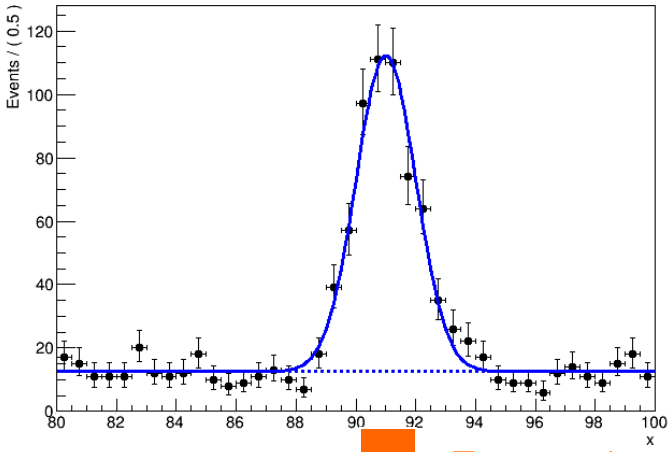
Statistical inference:
$L(15|5) = 1.5\ 10^{-4}$



Formulation of probability model of reduced sample: Poisson(N|s+b)

Modeling discriminating observables

- Example 2 – **Discrimination in inference stage**

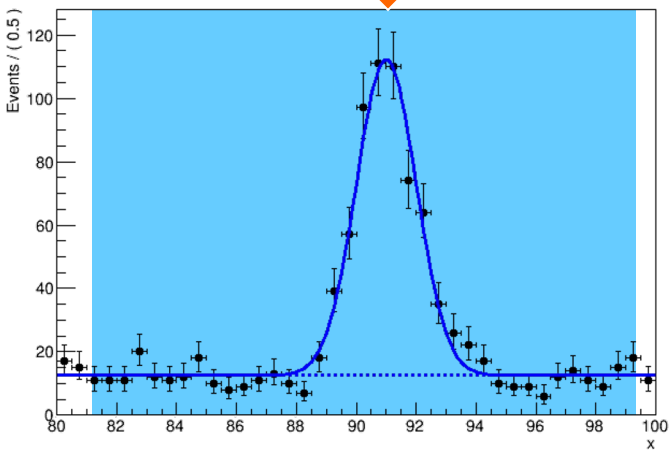


NB1: Most discrimination power in inference step.
This is again design choice!

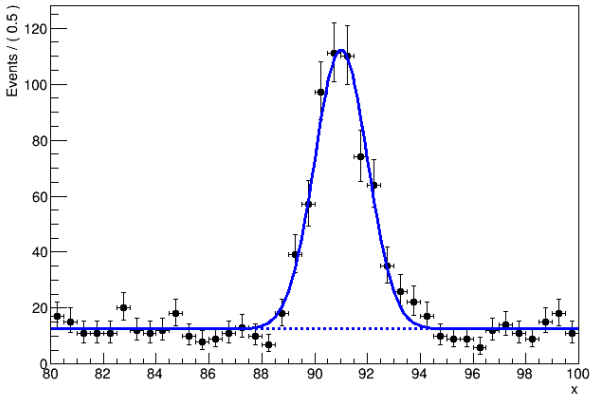
NB2: Optimal selection less critical

NB3: Correct description of selected sample more complex

Event selection: reduce sample size and dimensionality



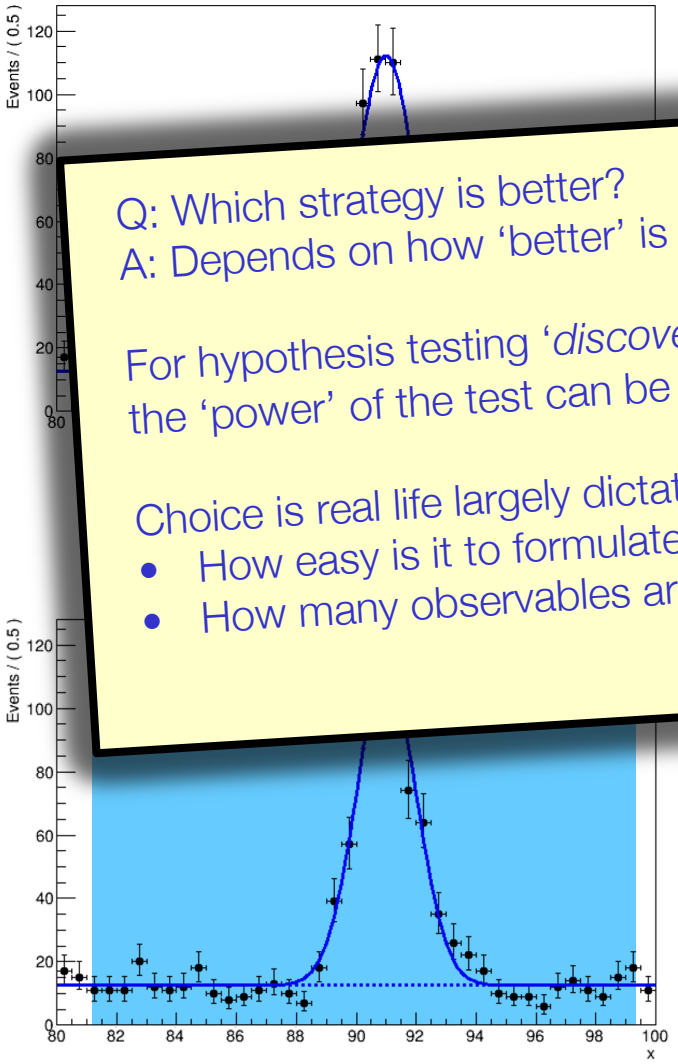
Statistical inference:
L(data|hypo)=something



Formulation of probability model of reduced sample:
*Nbkg*Uniform(x) +Nsig*Gaussian(x)*

Modeling discriminating observables

- Example 2 – full dataset has one discriminating observable: x

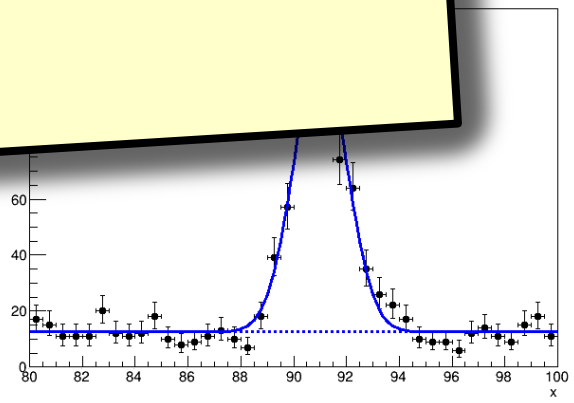


NB1: Most discrimination power in inference step.
This is again design choice!

Q: Which strategy is better?
A: Depends on how 'better' is defined?

For hypothesis testing '*discovery of a new particle*'
the 'power' of the test can be the same, but doesn't need to be

Choice is real life largely dictated by practicalities
- How easy is it to formulate a description of the observables?
- How many observables are important?



Formulation of probability model of reduced sample:
*Nbkg*Uniform(x) +Nsig*Gaussian(x)*

Formulating probability models for discriminating observables

- For counting experiments could derive Poisson(N|μ) from first principles ('random discrete events measured in fixed time interval)
- For experiments with discriminating observables, description should ideally also derive from underlying (physics) hypothesis/theory
  - In many cases this is possible, but not always without assumptions.
  - Assumptions lead to uncertainties in predictions → we'll revisit later how to deal with those.
- Example: common underlying principle in (signal) model is that discriminating observable is sum/average of many components
  - E.g. light collected by photomultiplier has contributions from >>1 photons
  - Tracks reconstructed in detector have contributions >>1 hits
  - Central Limit Theorem: for large N → Can be analytically described by Gaussian
- In case there is no easy analytical solution → empirical models (polynomial) or numerical solution (simulation-based histogram)

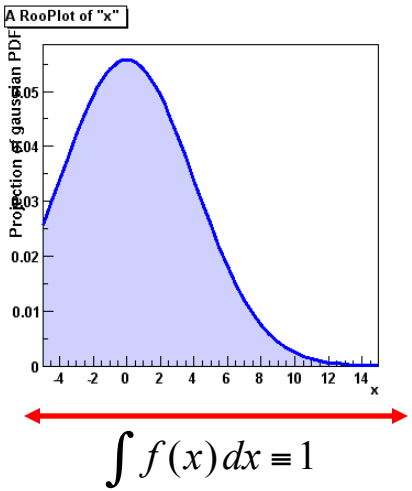
Mathematical formulation of models for observables

- Mathematical description for counting expt is probability model

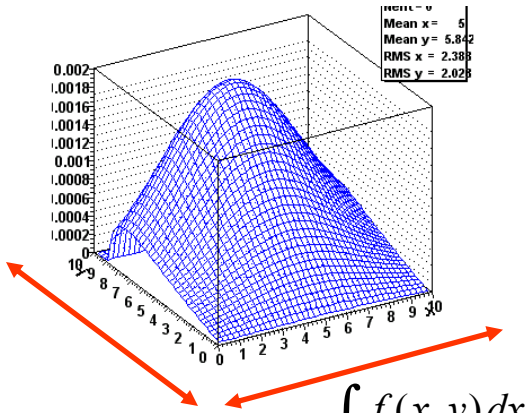
$$P(N) \geq 0 \quad \forall N \qquad\qquad \sum_{N=0}^{\infty} P(N) \equiv 1$$

- Mathematical description for distribution of discriminating observable is a *probability density model*:

$$f(\vec{x}) \geq 0 \quad \forall \vec{x} \qquad\qquad \int f(\vec{x})\, d\vec{x} \equiv 1$$



$$\int f(x)\, dx \equiv 1$$



$$\int f(x,y)\, dx\, dy \equiv 1$$

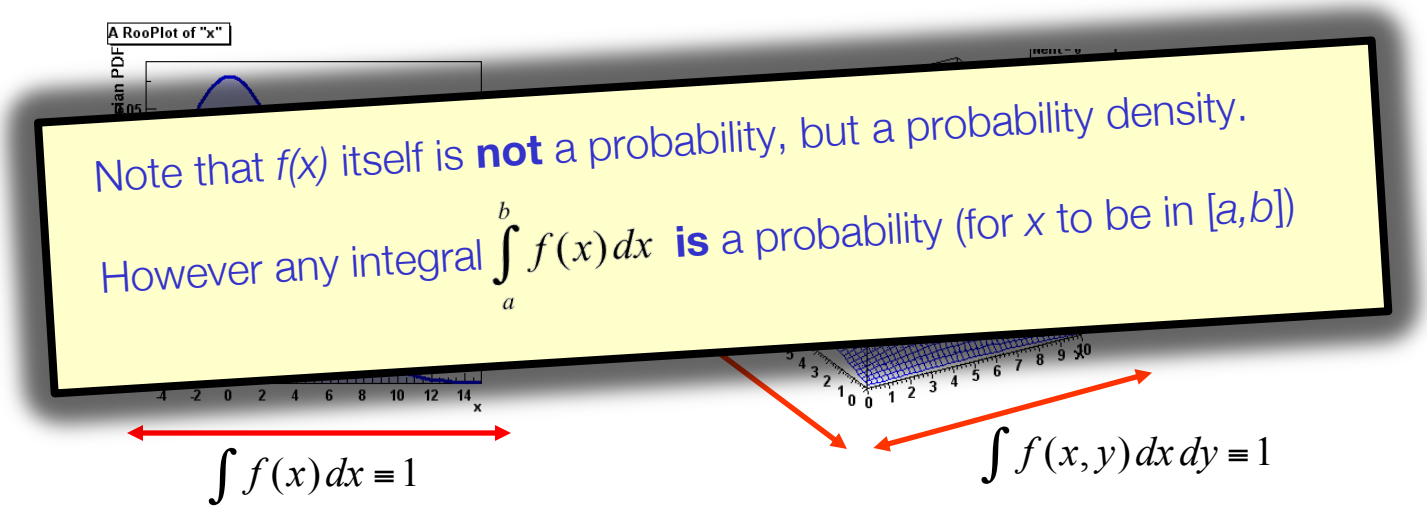
Mathematical formulation of models for observables

- Mathematical description for counting expt is probability model

$$P(N) \geq 0 \quad \forall N \qquad\qquad \sum_{N=0}^{\infty} P(N) \equiv 1$$

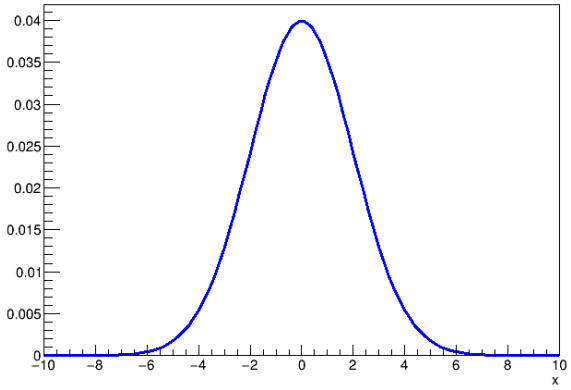
- Mathematical description for distribution of discriminating observable is a *probability density model*:

$$f(\vec{x}) \geq 0 \quad \forall \vec{x} \qquad\qquad \int f(\vec{x})\, d\vec{x} \equiv 1$$

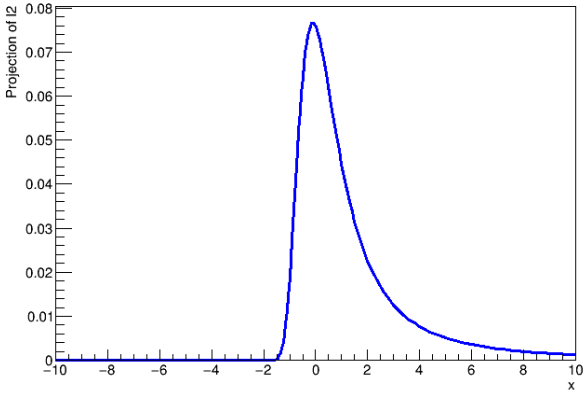


Some examples of physics-inspired probability density models

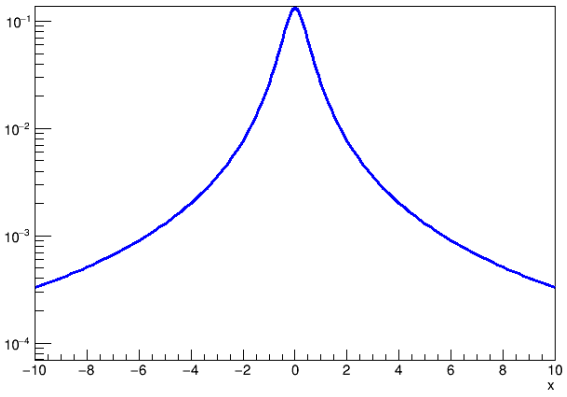
Gaussian
(anything in CLT regime)



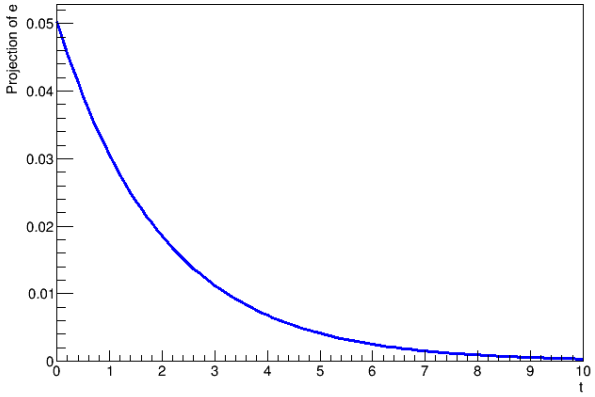
Landau
(energy loss in matter)



Breit-Wigner
(resonant mass)

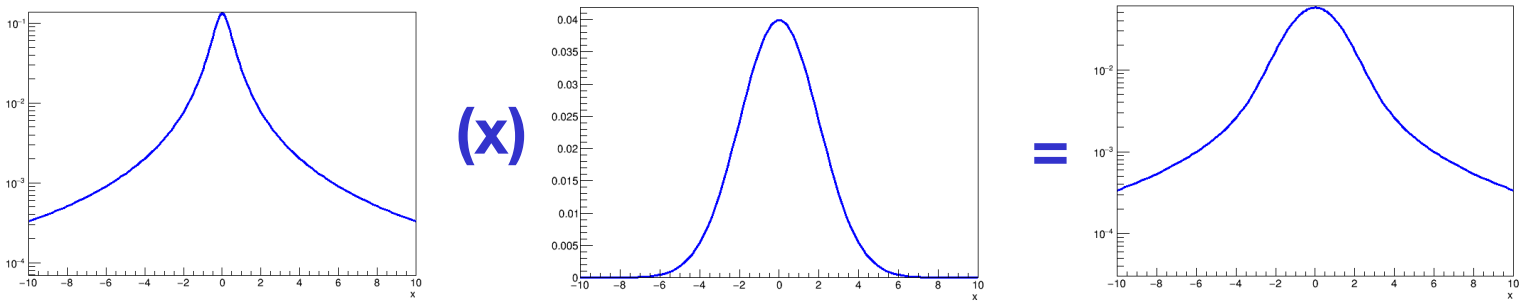


Exponential
(decay time)

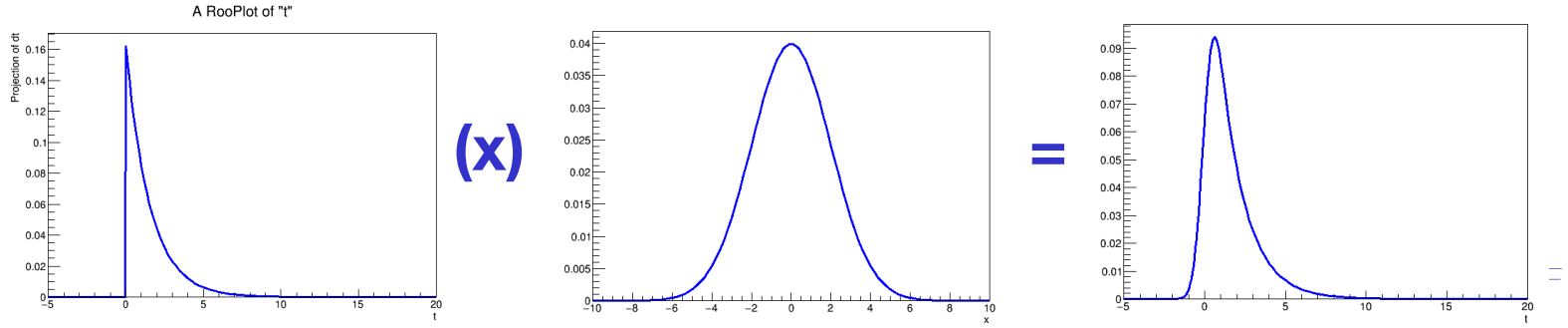


Signal models are often convolutions!

- Observable distributions are often well described by convolutions of physics distributions with (experimental) resolution functions.
  - Often can be calculated analytically, otherwise numerically use FFT
- Example 1: Resonance mass (x) detector resolution



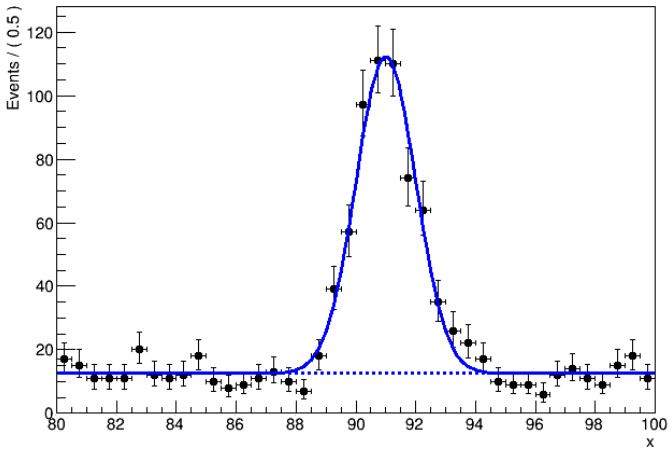
- Example 2: Decay life time (x) detector resolution



PDFs with multiple process contributions – aka mixture models

- Analogous to the counting model Poisson(N|S+B), probability density models can describe the distribution of such hypothesis through simple addition

$$f(x) = f_{sig}\ \text{Gaussian}(x) + (1-f_{sig})\ \text{Uniform}(x)$$



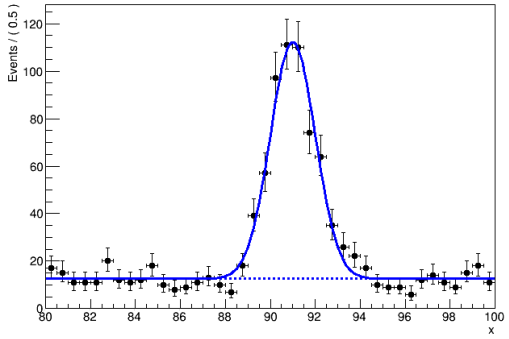
If Gaussian(x) and Uniform(x) are pdfs, then their sum is also a pdf, provided the sum of the coefficients is also 1

- Given a data sample D(x) of N *independent identically distributed* observations of x, the Likelihood is

$$L(\vec{x}) = \prod_{i=0\ldots N} f(x_i)$$

PDFs with multiple process contributions

- Note that the Likelihood L(x) of a probability density function f(x) for a data sample D(x) with N entries only exploits the differential distribution in x, *but not the event count N of the data*
- In many cases the event count can also distinguish the S/B hypothesis (more events expected if signal is present). If so, the probability model for the event count can be explicitly included in the Likelihood (often called 'extended likelihood')



f(x) = f_{sig} Gaussian(x) + (1-f_{sig}) Uniform(x)

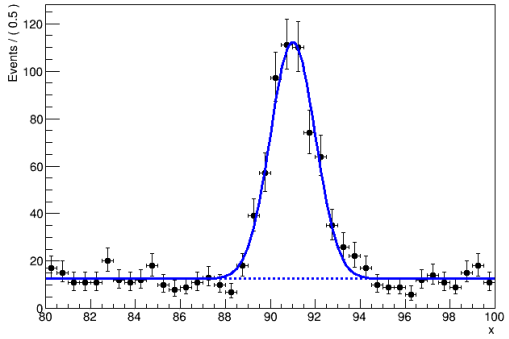
P(N) = Poisson(N | N_{exp})

$$L(\vec{x}, N) = \prod_{i=0\ldots N} f(x_i \mid f_{sig}) \cdot Poisson(N \mid N_{\exp})$$

- In the common case of a signal and background, with a respective expected event S and B, one can reparameterize (f_{sig},N_{exp}) → (S,B)

PDFs with multiple process contributions

- Note that the Likelihood L(x) of a probability density function f(x) for a data sample D(x) with N entries only exploits the differential distribution in x, *but not the event count N of the data*
- In many cases the event count can also distinguish the S/B hypothesis (more events expected if signal is present). If so, the probability model for the event count can be explicitly included in the Likelihood (often called 'extended likelihood')



f(x) = S/(S+B)Gaussian(x) + B/(S+B)Uniform(x)

P(N) = Poisson(N | S+B)

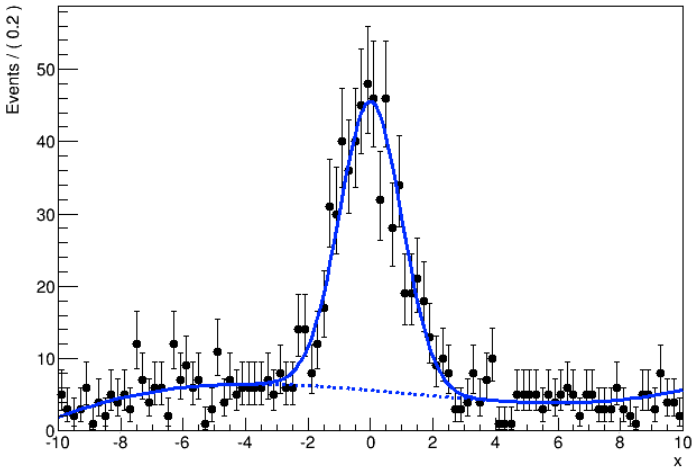
$$L(\vec{x},N) = \prod_{i=0\ldots N} f(x_i \mid S,B) \cdot Poisson(N \mid S+B)$$

- In the common case of a signal and background, with a respective expected event S and B, one can reparameterize (f_{sig},N_{exp}) → (S,B)

Empirical probability models

- In case no description from first principles exists for a differential distribution, empirical or simulation-based models can be deployed

Empirical models

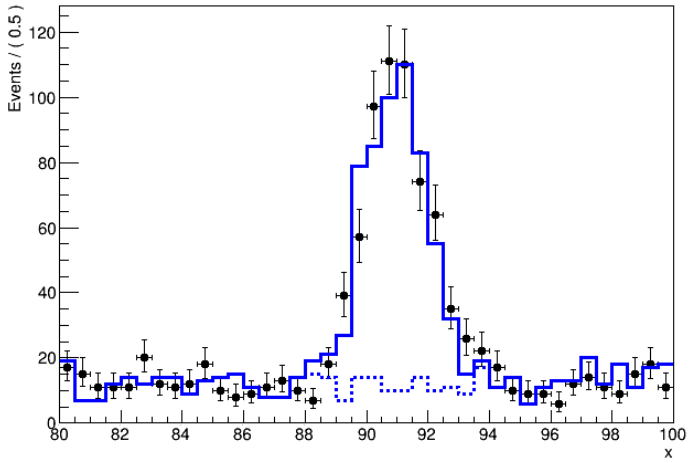


$B(x) = a_0+a_1x+a_2x^2+a_3x^3...$

Drawbacks:

- Arbitrariness in parameterization, e.g. which order to choose for a polynomial

Simulation-based models



B(x) = histogram

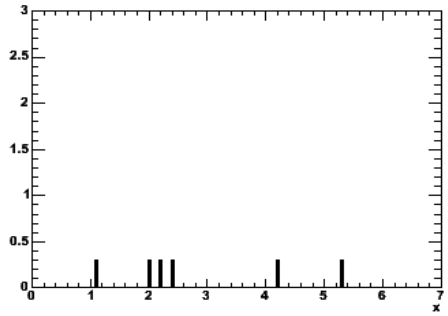
Drawbacks:

- Quantization of model prediction in bins
- Poor modeling in regions with low simulation statistics

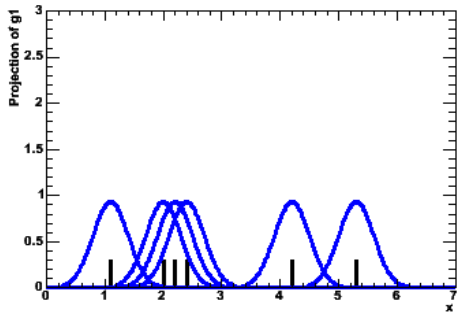
Modeling low-statistics simulation predictions

- For low-statistics simulation predictions,
  kernel estimation techniques can improve modeling substantially
- Procedure:
  - Assign a Gaussian probability density distribution to each simulated event.
  - Sum Gaussian probability densities of all events
  - Started from unbinned data → no binning effects

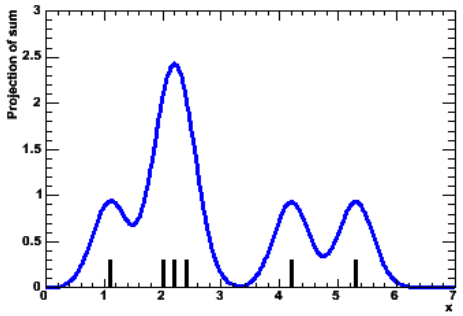
Sample of events



Gaussian probability distributions for each event



Summed probability distribution for all events in sample

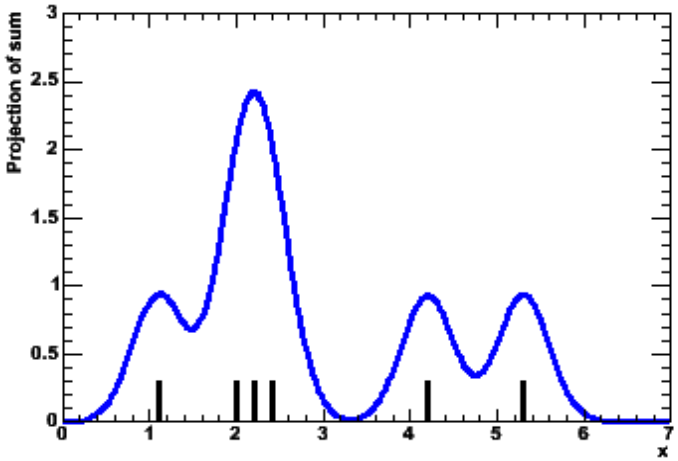


Modeling low-statistics simulation predictions

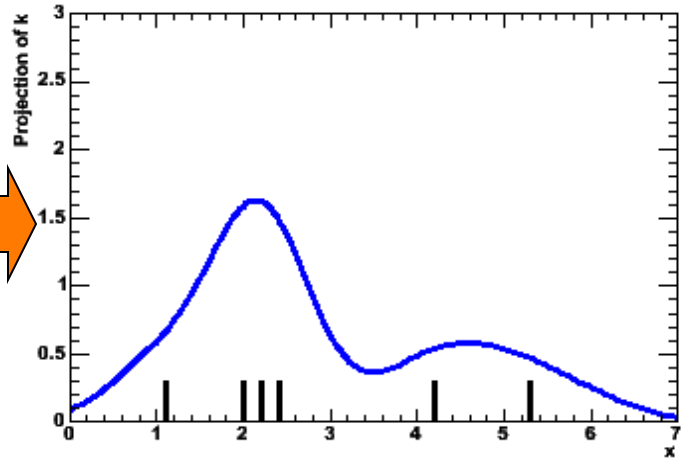
- Technique does *not* require that all Gaussian kernels have same width
- Improved procedure: 'adaptive kernel'
  - Adjust with of Gaussian kernels depending on local event density
  - High density → narrow kernels → preserve more detail
  - Low density → wide kernels → promote smoothness

*Kernel estimation is a simplified form of a '**Gaussian processes**' techniques*

Static Kernel
(with of all Gaussian identical)



Adaptive Kernel
(width of all Gaussian depends on local density of events)



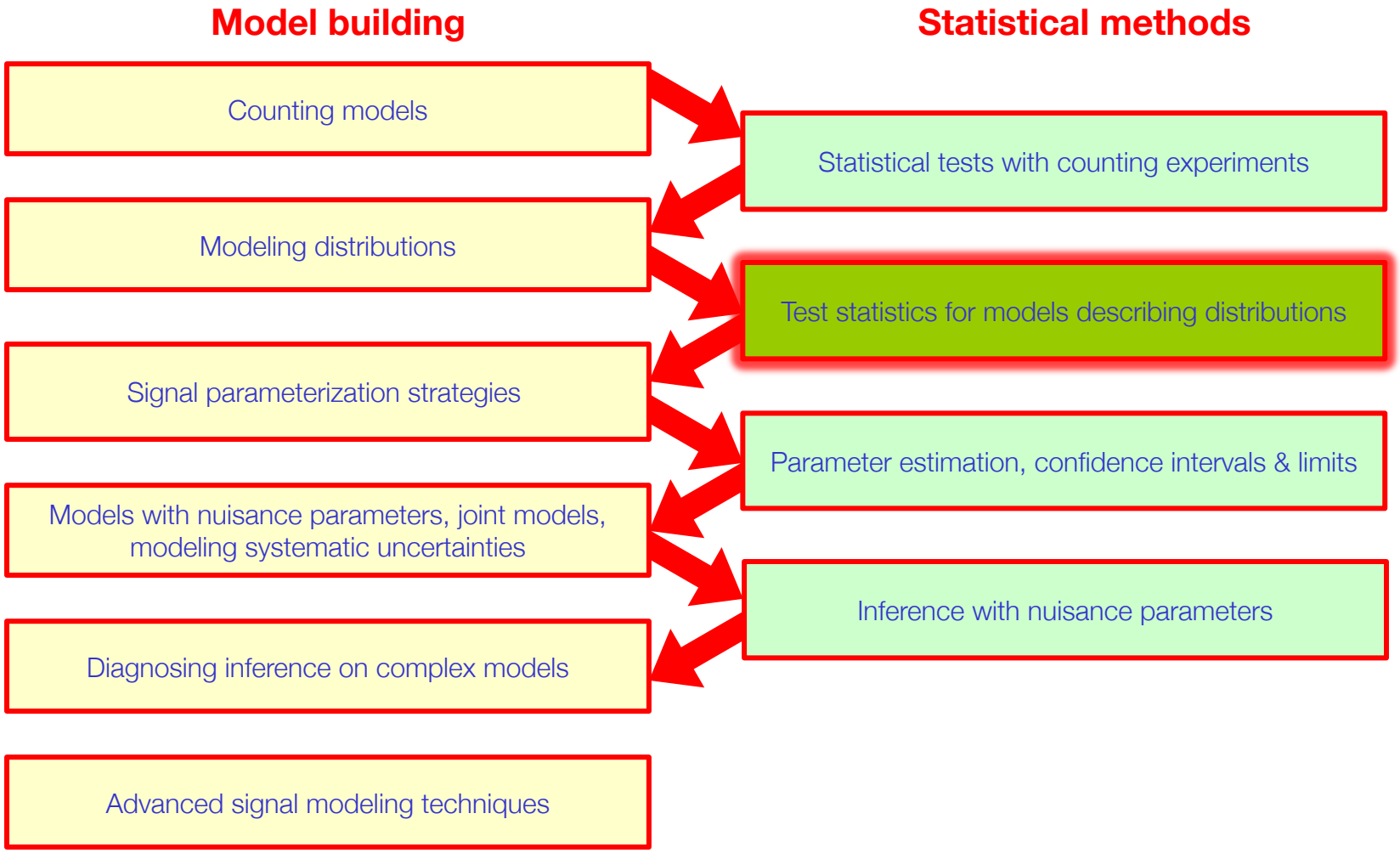
Statistical methods 2

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

Wouter Verkerke, NIKHEF

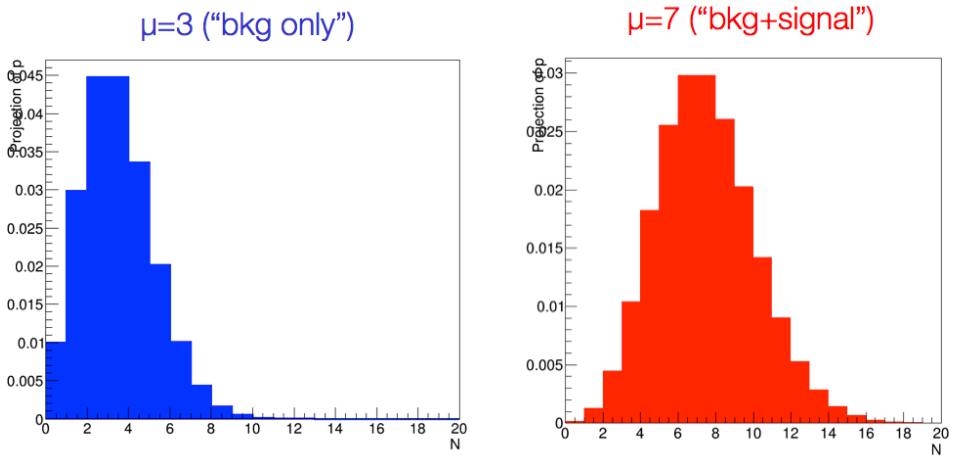
Roadmap of this course

- 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|bkg), and P(N|sig) we can calculate $P(N_{obs}|Hx)$ under either hypothesis

→ With additional information on P(Hi) we can also calculate P(Hx|Nobs)

- 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

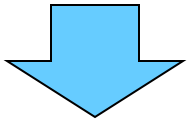
[Recap]

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

$$O_{prior} \equiv \frac{P(H_{sb})}{P(H_{b)}} = \frac{P(H_{sb})}{1 - P(H_{sb})}$$

If $p(H_{sb})=p(H_b)$ → Odds are 1:1



'Bayes Factor' K multiplies prior odds

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

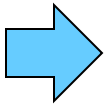
If $P(data|H_b)=10^{-7}$, $P(data|H_{sb})=0.5$ K=2.000.000 → Posterior odds are 2.000.000 : 1

[Recap]

Formulating evidence for discovery

- In the frequentist school you restrict yourself to P(data|theory) and there is no concept of 'priors'
  - But given that you consider (exactly) 2 competing hypothesis, very low probability for data under Hb lends credence to 'discovery' of Hsb (since Hb is 'ruled out'). Example

$P(data|H_b)=10^{-7}$
$P(data|H_{sb})=0.5$



"H_b ruled out" → "Discovery of H_{sb}"

- Given importance to interpretation of the lower probability, it is customary to quote it in "physics intuitive" form: Gaussian σ.
  - E.g. '5 sigma' → probability of 5 sigma Gaussian fluctuation $=2.87\times10^{-7}$
- 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(theo|data) they are *exactly the same.*
  - Simply substitute counting model with binned distribution model

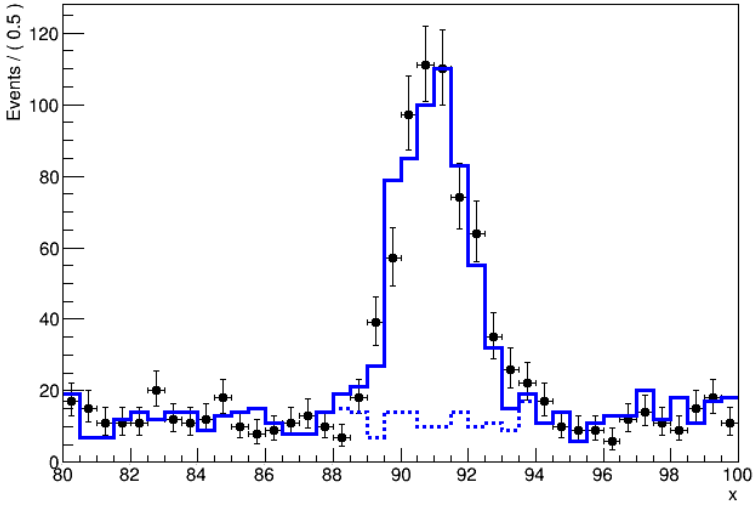
$$P(H_{s+b} \mid \vec{N}) = \frac{L(\vec{N} \mid H_{s+b})P(H_{s+b})}{L(\vec{N} \mid H_{s+b})P(H_{s+b}) + L(\vec{N} \mid H_b)P(H_b)}$$

Simply fill in new Likelihood function
Calculation otherwise unchanged

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

Working with Likelihood functions for distributions

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



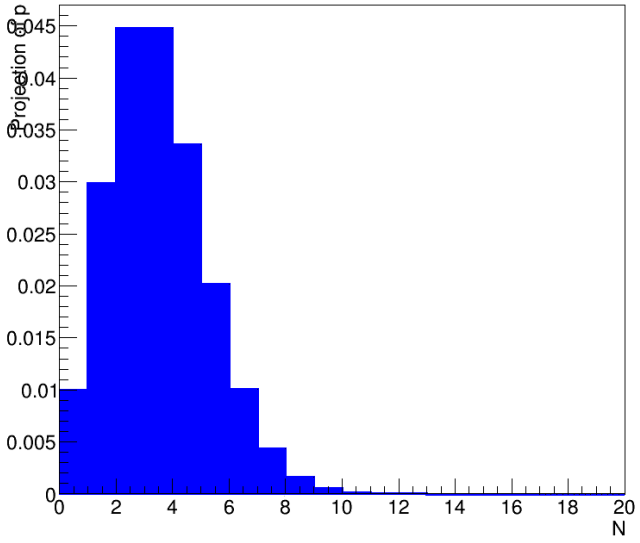
$$L(\vec{N} \mid H_b) = \prod_i Poisson(N_i \mid \tilde{b}_i)$$

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

- **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' that 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)

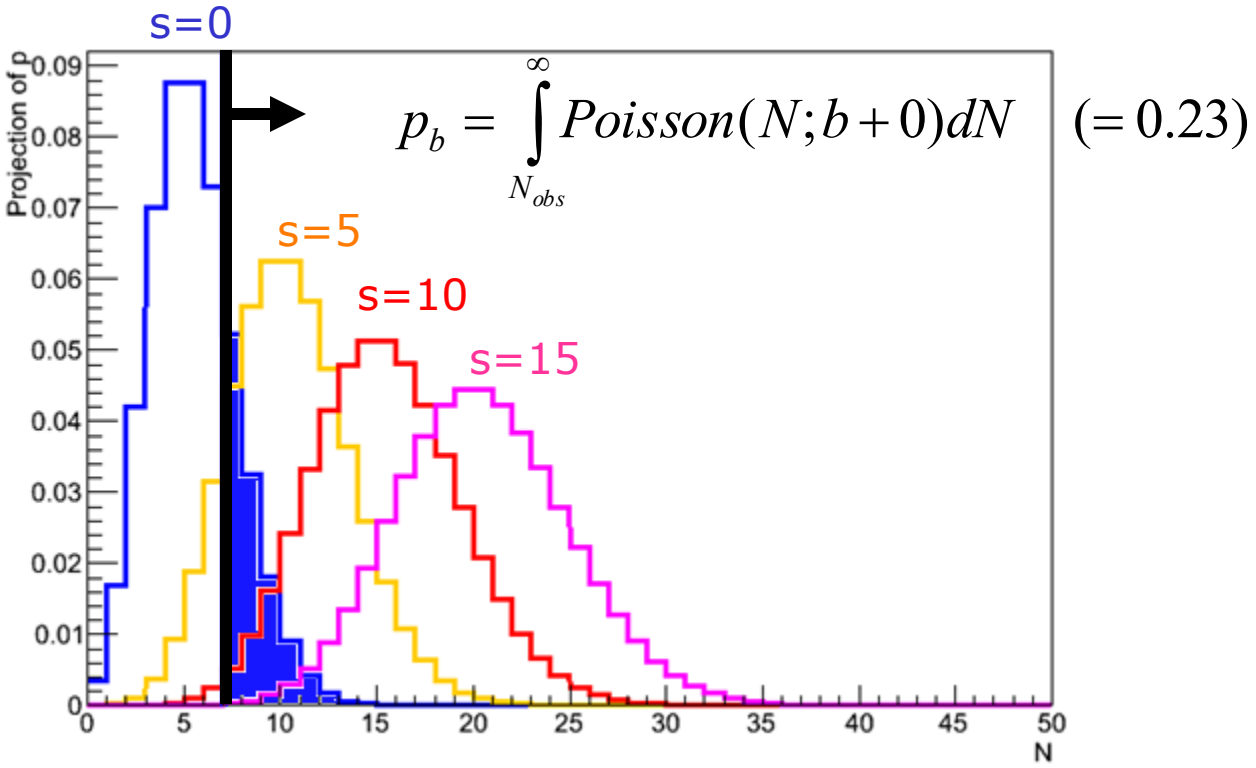


Test statistic T(N)=Nobs orders observed events count by estimated signal yield

Low N → low estimated signal
High N → large estimated signal

P-values for counting experiments

- Now make a measurement $N=N_{obs}$ (example $N_{obs}=7$)
- **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’

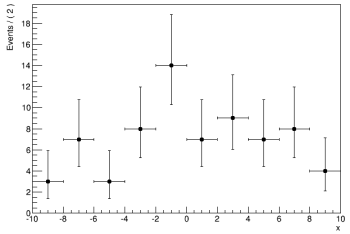
- How to define ‘extremity’ if observed data is a distribution

Counting

Histogram

Observation

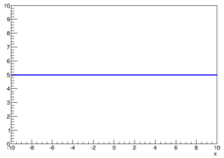
$N_{obs}=7$

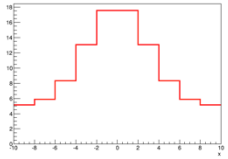


Median expected by hypothesis

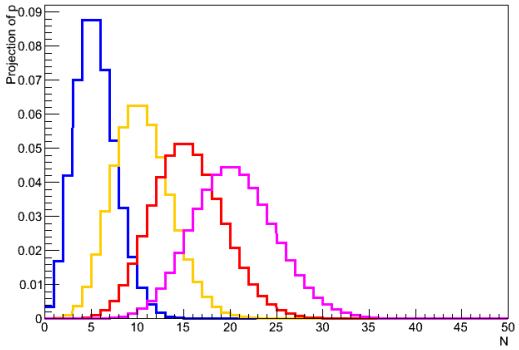
$N_{exp}(s=0) = 5$

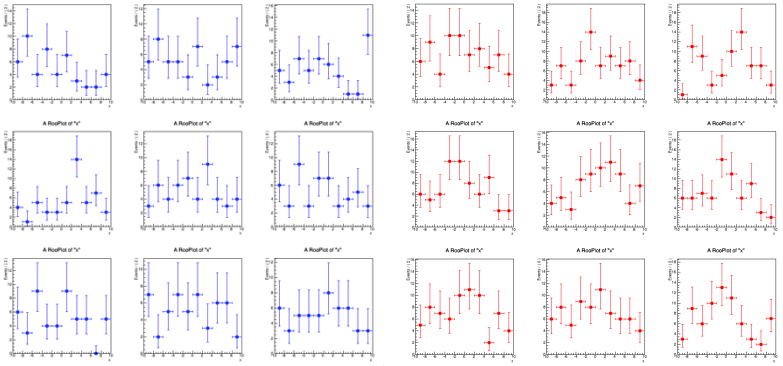
$N_{exp}(s=5) = 10$





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

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

- Intuitive picture:

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

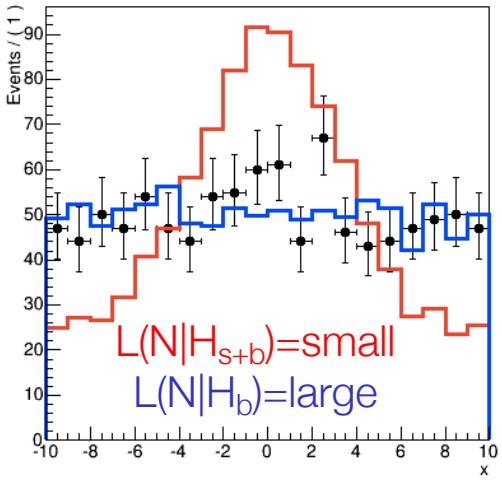
$$\lambda(\vec{N}) = \frac{\text{small}}{\text{large}} = \text{small}$$

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

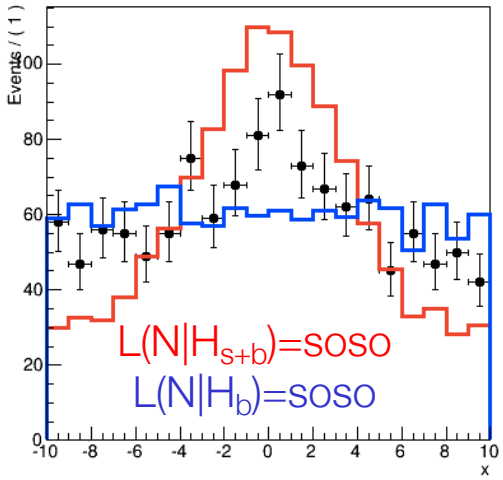
$$\lambda(\vec{N}) = \frac{\text{large}}{\text{small}} = \text{large}$$

Visualizing the Likelihood Ratio as ordering principle

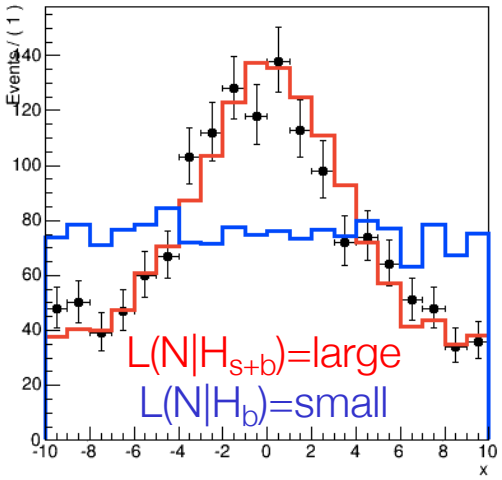
- The Likelihood ratio as ordering principle



λ(N)=0.0005



λ(N)=0.47

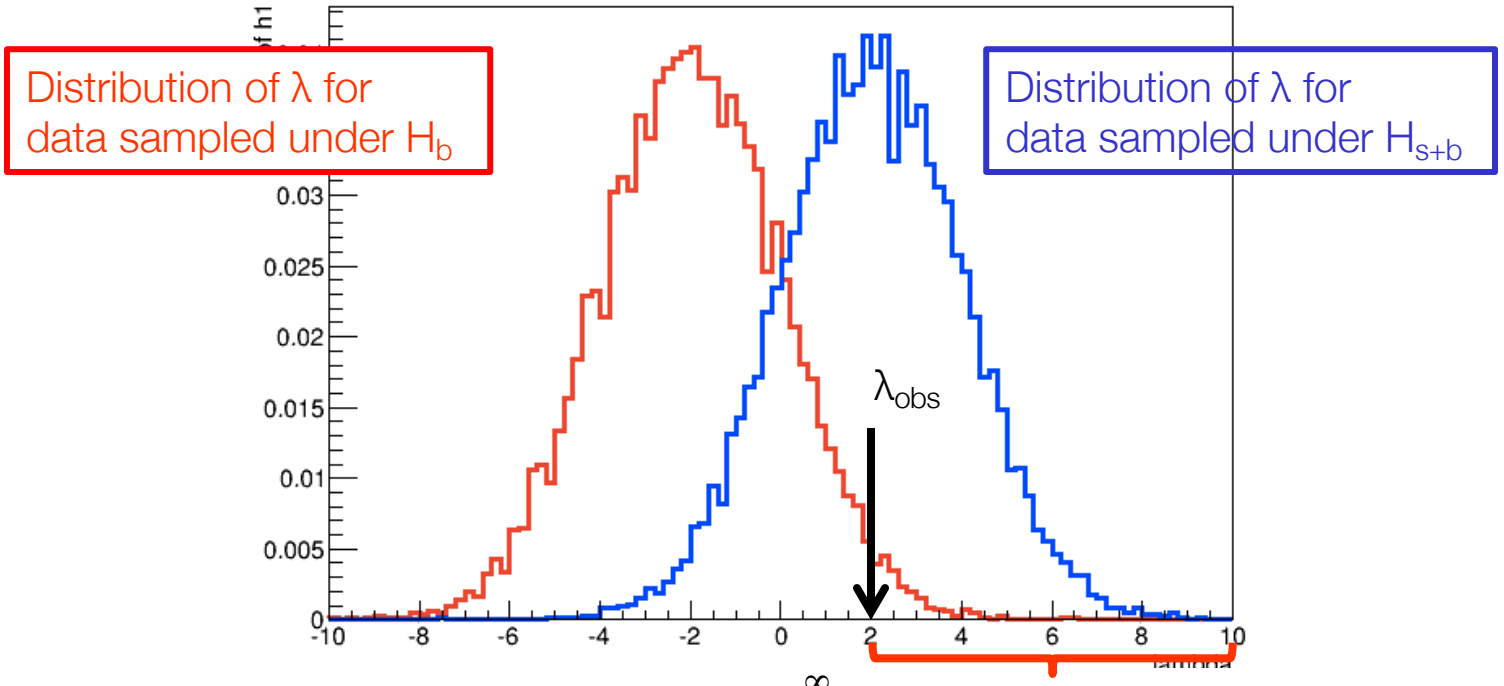


λ(N)=5000

- **Frequentist solution to 'relevance of P(data|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 λ(N) as large or larger than the observed test statistic λ(N$_{obs}$)**

The distribution of the test statistic

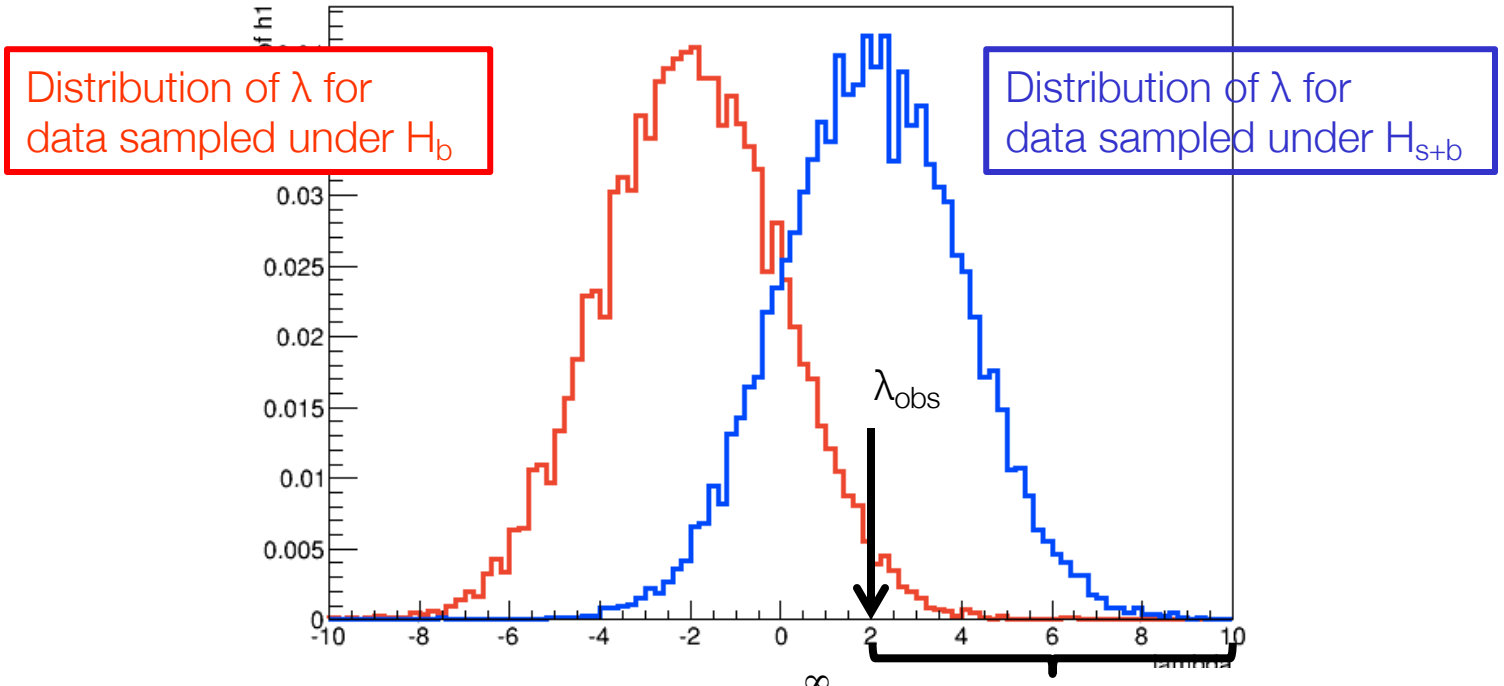
- Distribution of a test statistic is *generally not known*
- Use toy MC approach to approximate distribution
  - Generate many toy datasets N under H_b and H_{s+b} and evaluate λ(N) for each dataset



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

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)



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

Likelihoods for distributions - summary

- **Bayesian inference unchanged**

  → simply insert L of distribution to calculate P(H|data)

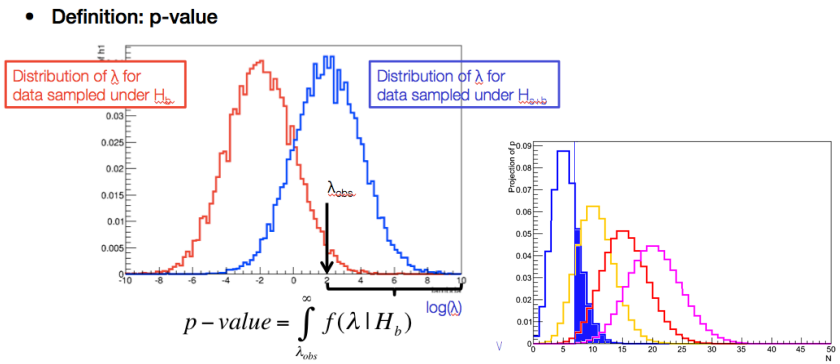
$$P(H_{s+b} \mid \vec{N}) = \frac{L(\vec{N} \mid H_{s+b})P(H_{s+b})}{L(\vec{N} \mid H_{s+b})P(H_{s+b}) + L(\vec{N} \mid H_b)P(H_b)}$$

- **Frequentist inference procedure *modified***

  → Pure P(data|hypo) not useful for non-counting data
  → Order all possible data with a (LR) test statistic in 'extremity'
  → Quote p(data|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%'



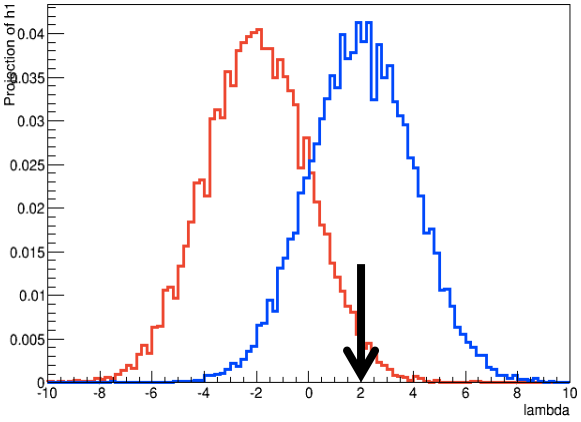
The likelihood principle

- 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

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

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

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



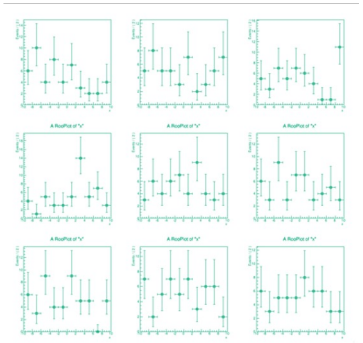
- **Distribution f(λ|N), and thus p-value, also uses likelihood of non-observed outcomes** (in fact Likelihood of every possible outcome is used)

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**:

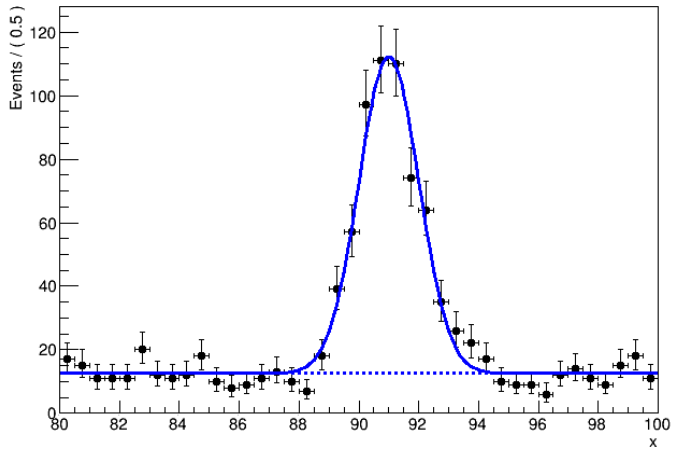
  If two experiments yield likelihood functions which are proportional, then Your inferences from the two experiments should be identical.

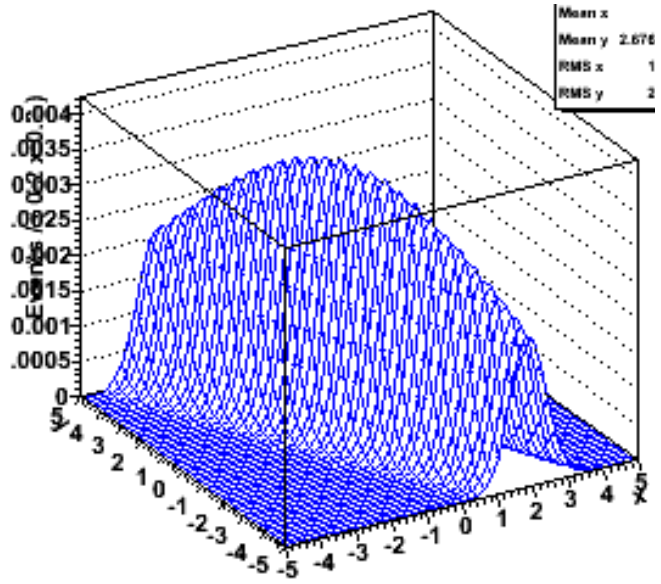
"Does it matter what the observable outcome of other experiments is?"
"Suppose you learn that your observation is an outlier, should that somehow be factored into your conclusions?"



Generalizing to multiple dimensions

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





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

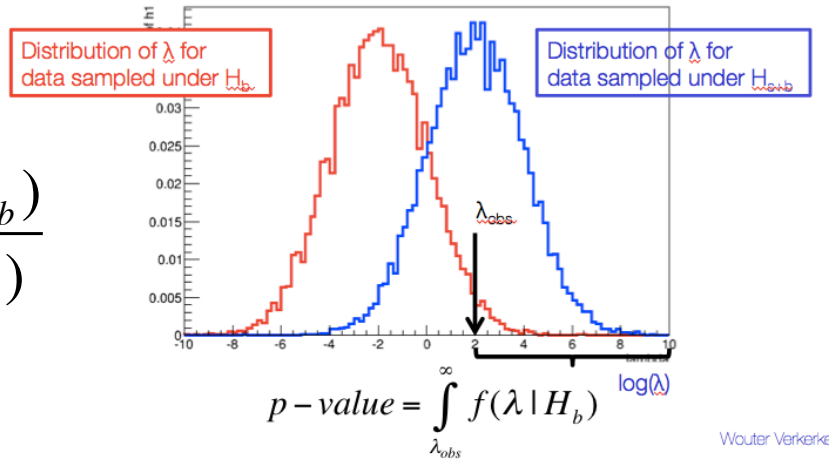
$$L(\vec{x},\vec{y}) = \prod_i f(x_i, y_i)$$

- Neither generalization (binned→continuous, one→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**

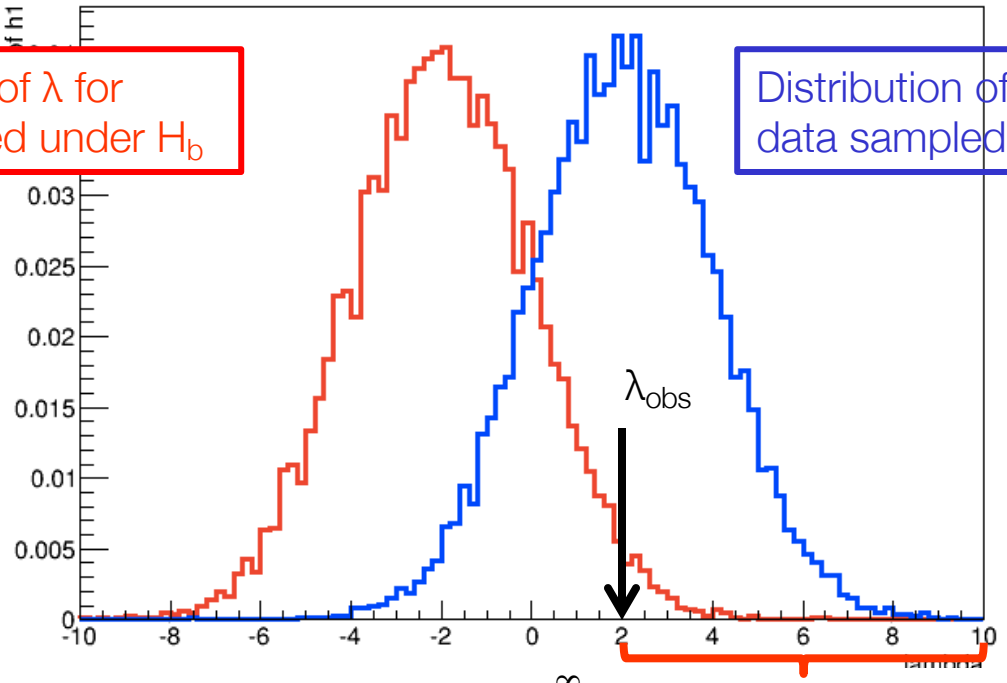
$$\lambda(\vec{x},\vec{y},\vec{z},...) = \frac{L(\vec{x},\vec{y},\vec{z},... \mid H_{s+b})}{L(\vec{x},\vec{y},\vec{z},... \mid H_b)}$$



- So far we have exploited λ to calculate a frequentist p-value **now explore properties 'cut on λ' as basis of (optimal) event selection**

The distribution of the test statistic

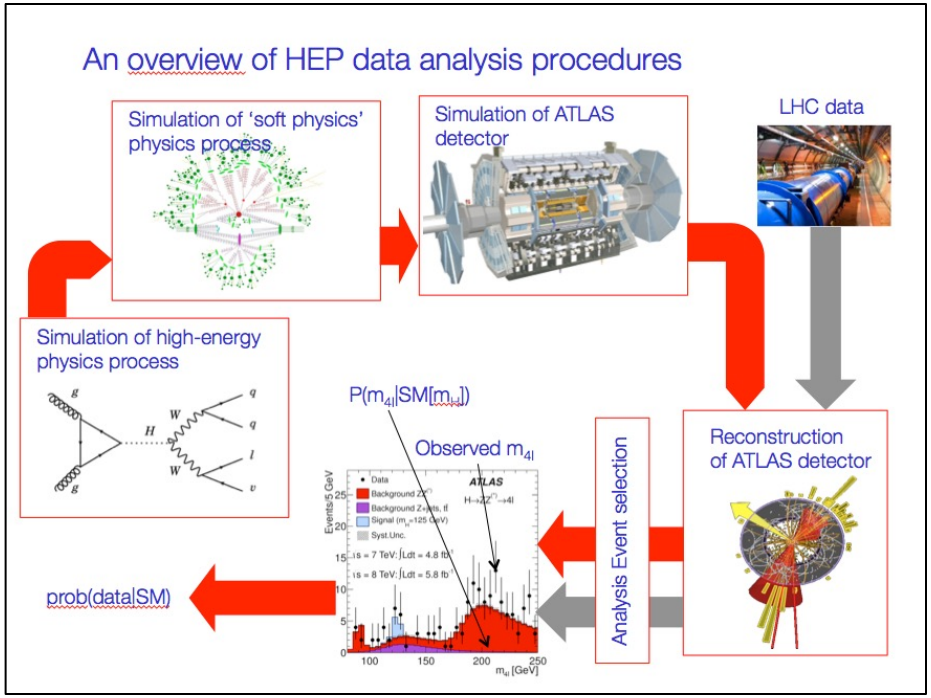
- Distribution of a test statistic is *generally not known*
- Use toy MC approach to approximate distribution
  - Generate many toy datasets N under H_b and H_{s+b} and evaluate λ(N) for each dataset



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

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

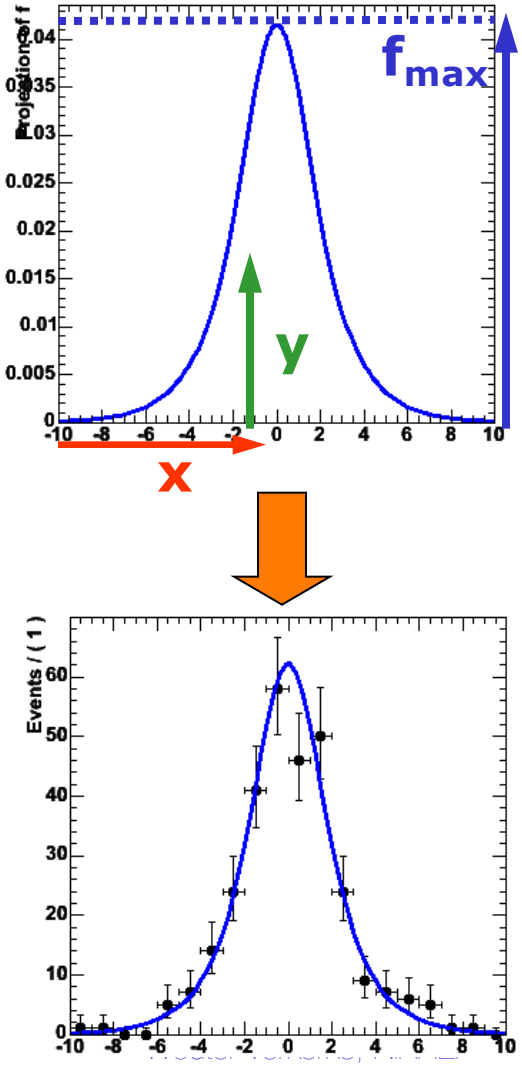


- Alternative approach is sample sampling the probability model 'toy MC'
  - Fast (generally), only requires access to probability model
  - Can only produce datasets with observables that are described by the probability model → Sufficient to study distribution of test statistics

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)



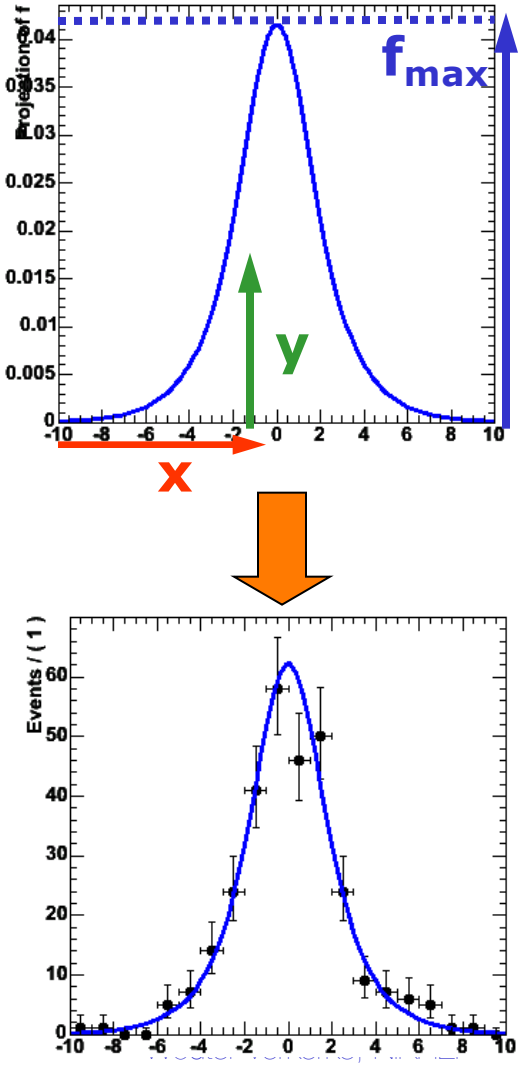
– PRO: Easy, always works

– 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)



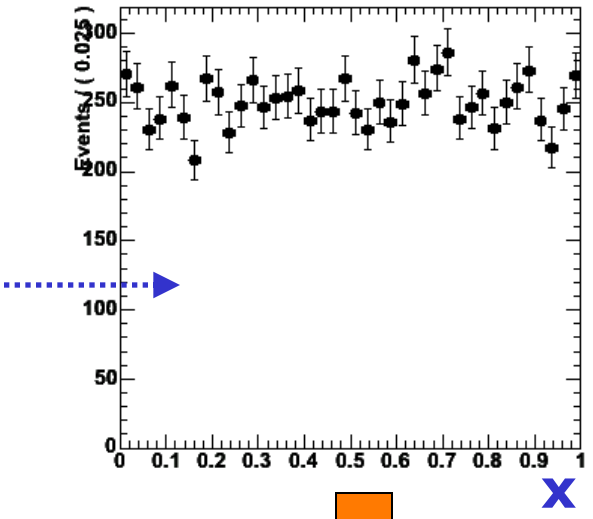
– PRO: Easy, always works

– CON: It can be inefficient if function is strongly peaked.
Finding maximum empirically through random sampling can be lengthy in >2 dimensions

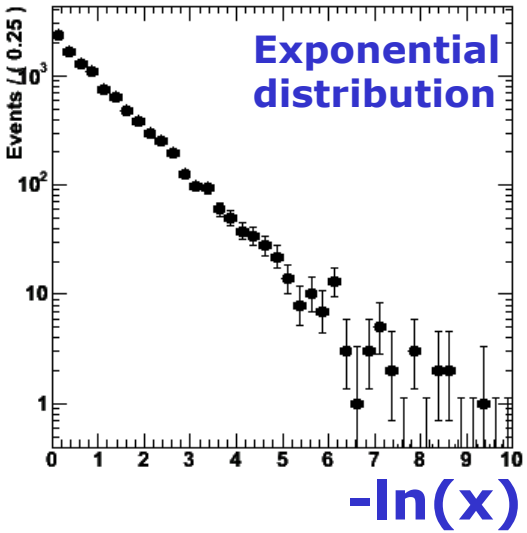
Toy MC generation – Inversion method

- 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)

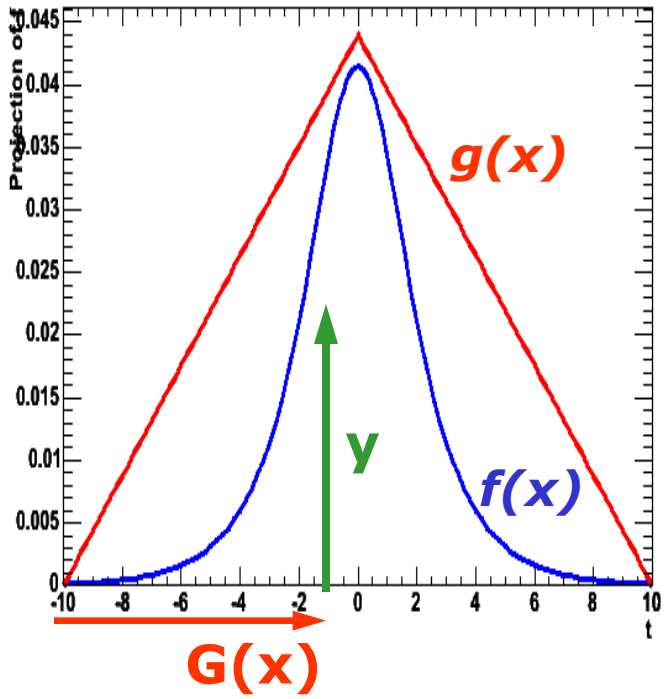


- PRO: Maximally efficient
- CON: Only works for invertible functions

Toy MC Generation – importance sampling

- Hybrid: Importance sampling

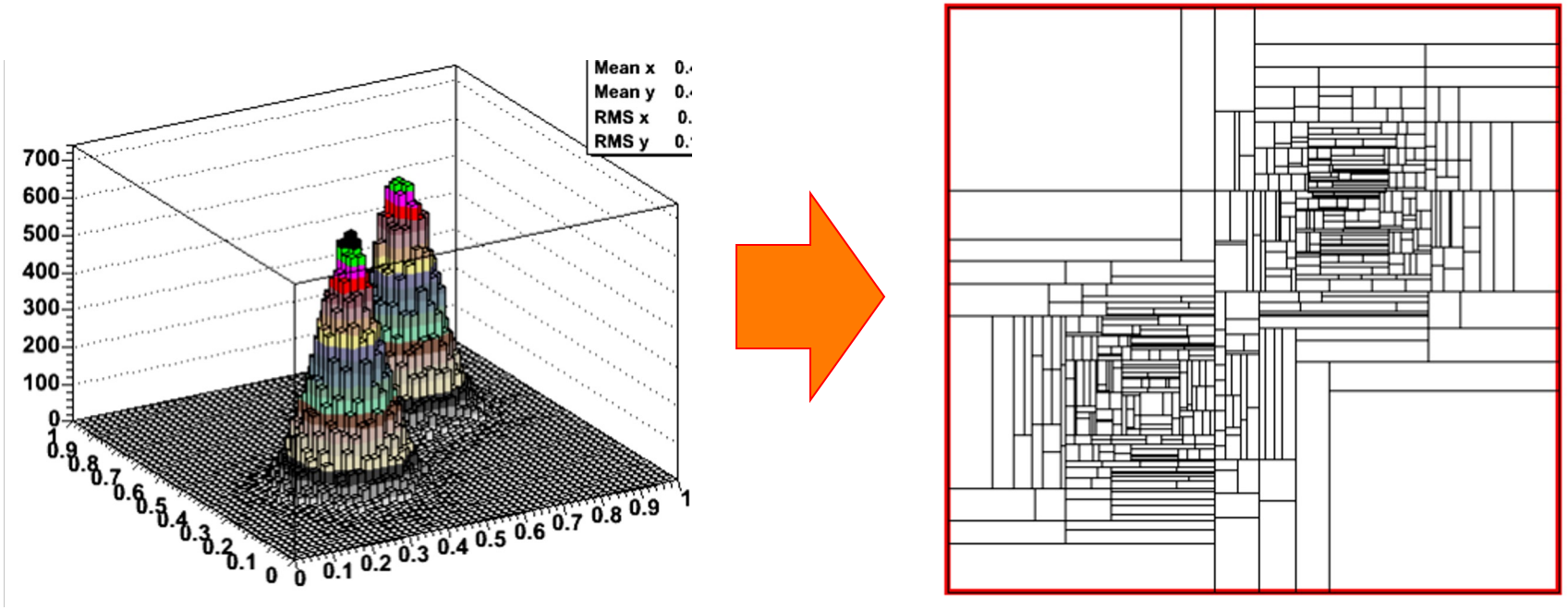
1) Find 'envelope function' g(x) that is invertible into G(x) and that fulfills g(x)>=f(x) for all x
2) Generate random number x from G using inversion method
3) Throw random number 'y'
4) If y<f(x)/g(x) keep x, otherwise return to step 2



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

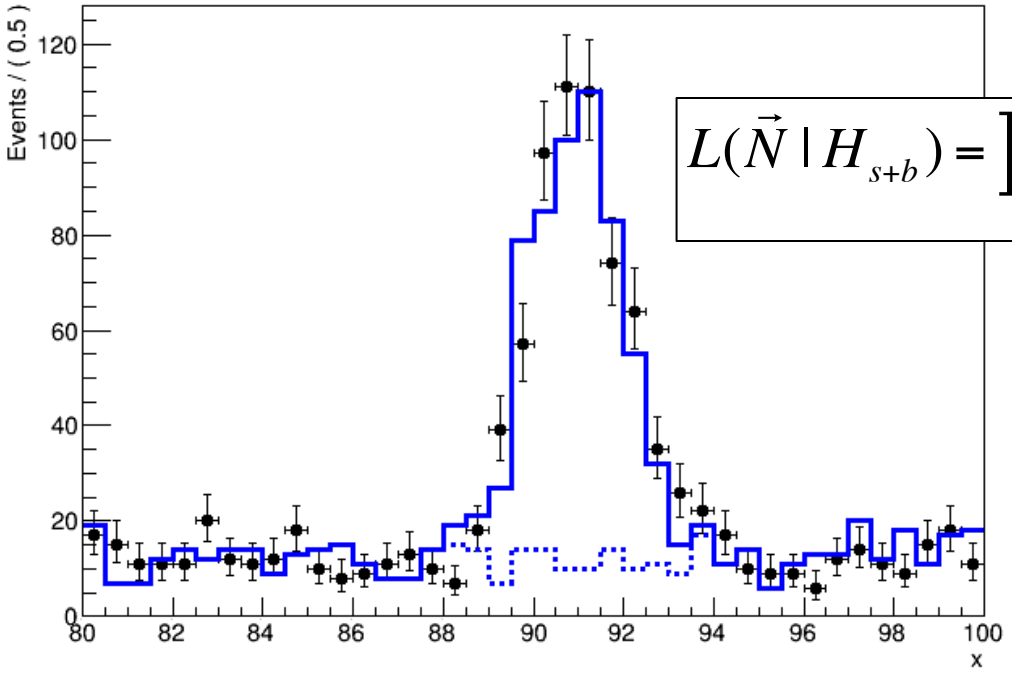
Toy MC Generation – importance sampling in >1D

- 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

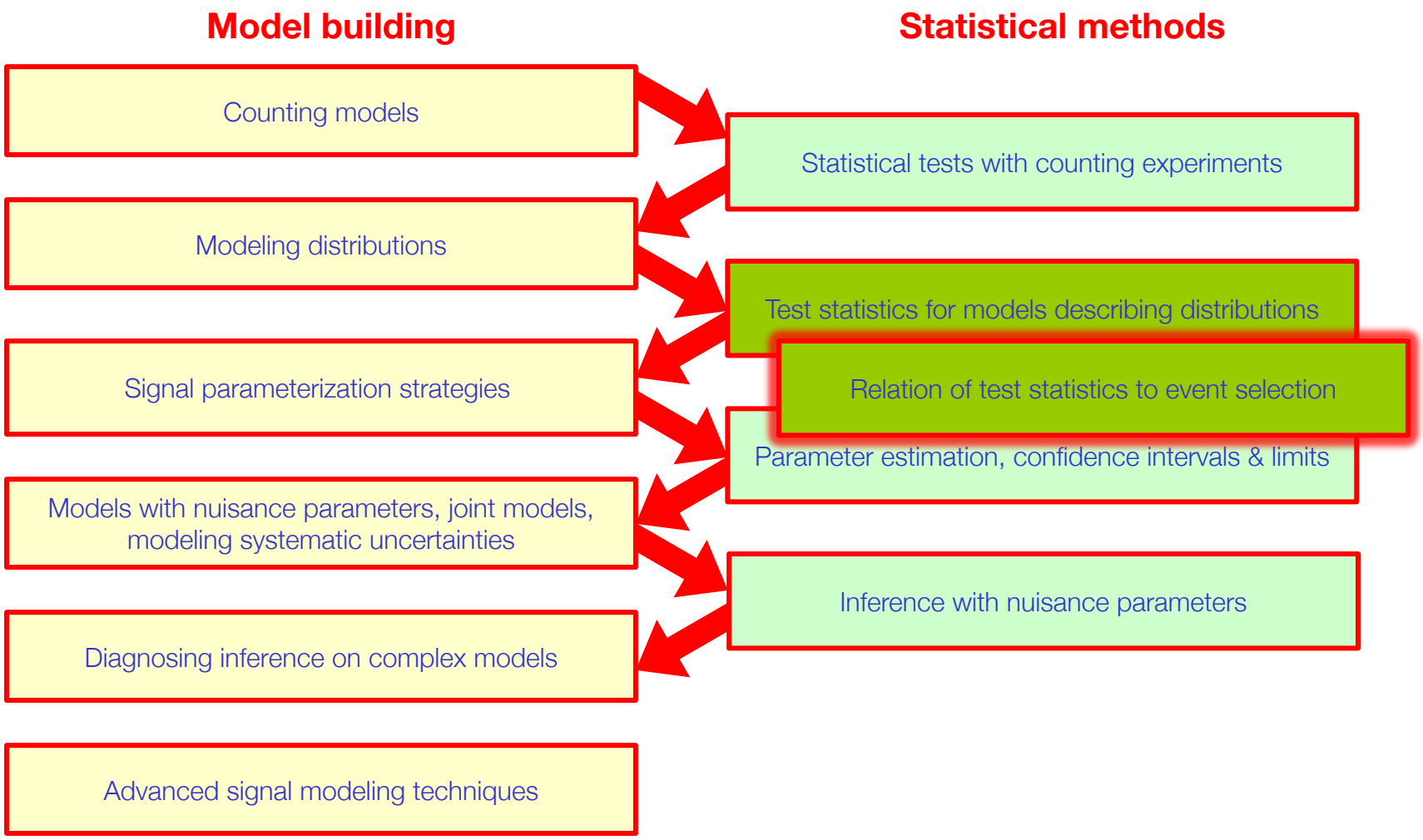


$$L(\vec{N} \mid H_{s+b}) = \prod_i Poisson(N_i \mid \tilde{s}_i + \tilde{b}_i)$$

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

Roadmap of this course

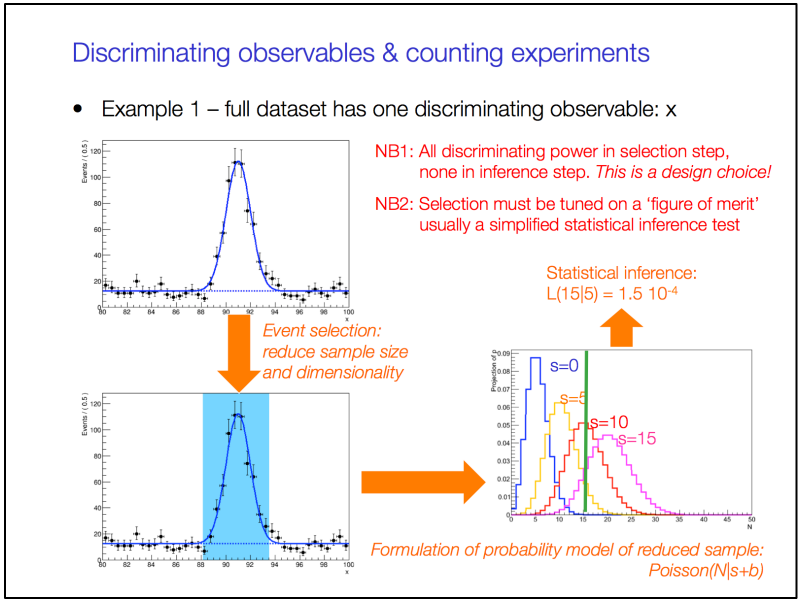
- Start with basics, gradually build up to complexity



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

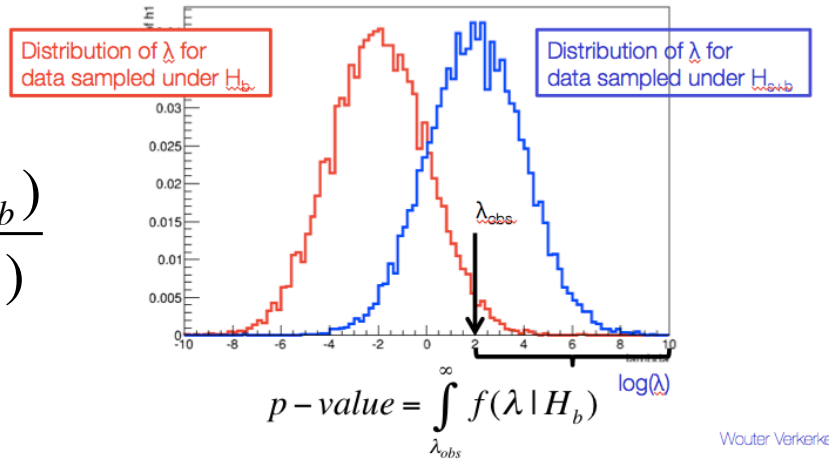
$$\lambda(\vec{x},\vec{y},\vec{z},...) = \frac{L(\vec{x},\vec{y},\vec{z},...\,|\,H_{s+b})}{L(\vec{x},\vec{y},\vec{z},...\,|\,H_{b})}$$

- Hence if we can construct λ, 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**

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



- So far we have exploited λ to calculate a frequentist p-value **now explore properties 'cut on λ' 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 = α
  - Rate of type-II error = β
  - Power of test is 1-β

		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 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 α
  - 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

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

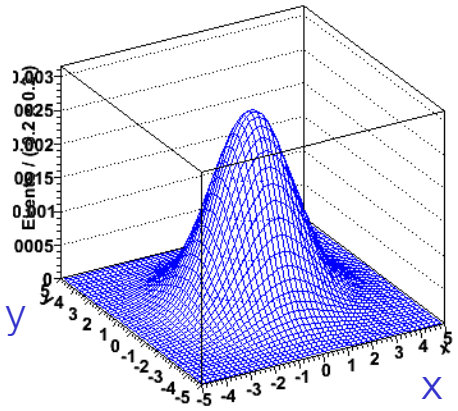
- Any other region of the same size will have less power

'ex03.C'

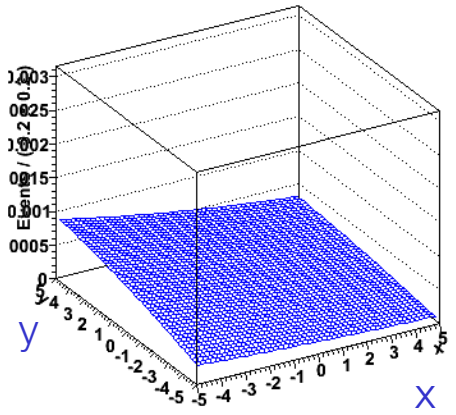
The Neyman-Pearson lemma

- Example of application of NP-lemma with two observables

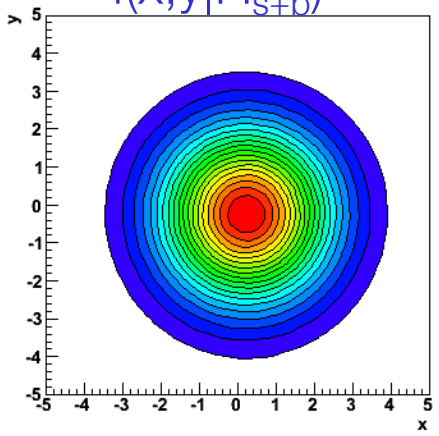
$f(x,y|H_s)$



$f(x,y|H_b)$



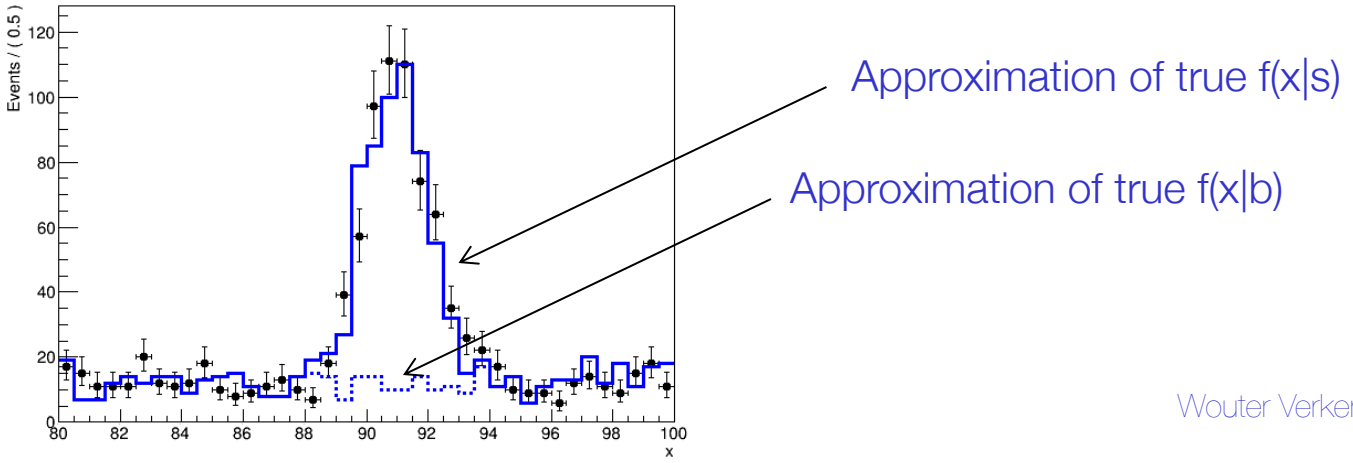
$$\frac{f(x,y|H_s)}{f(x,y|H_{s+b})} > c$$



- 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 $f(\vec{x}|\mathrm{s})$, $f(\vec{x}|\mathrm{b})$.
- 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 / paramater estimation with all events

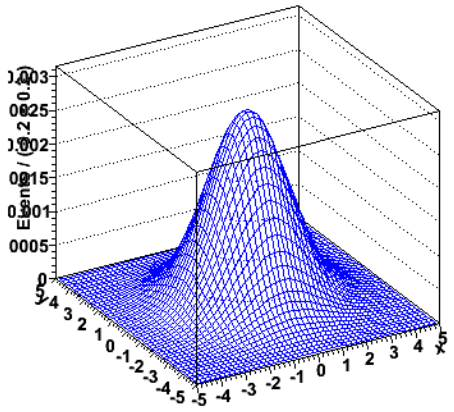


Hypothesis testing with a large number of observables

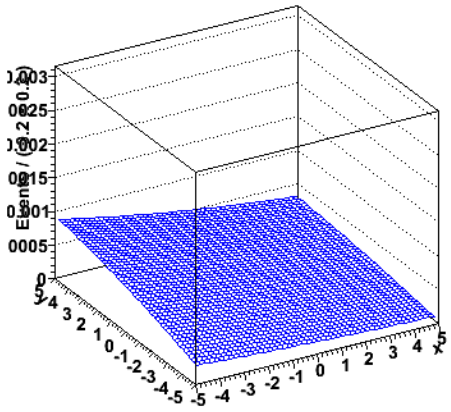
- When number of observables is large follow different strategy
- Instead of aiming at approximating p.d.f.s f(x|s) and f(x|b) aim to approximate decision boundary with an empirical parametric form

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

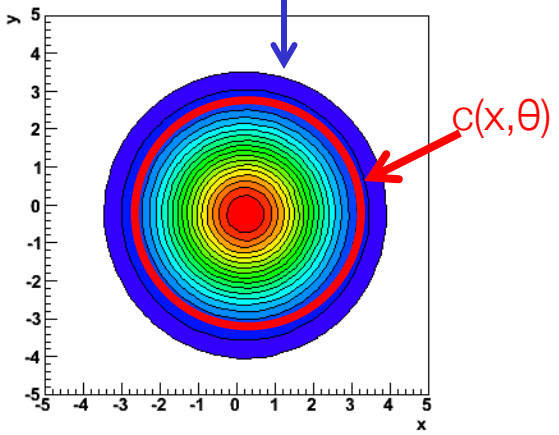
$f(x,y|H_s)$



$f(x,y|H_b)$



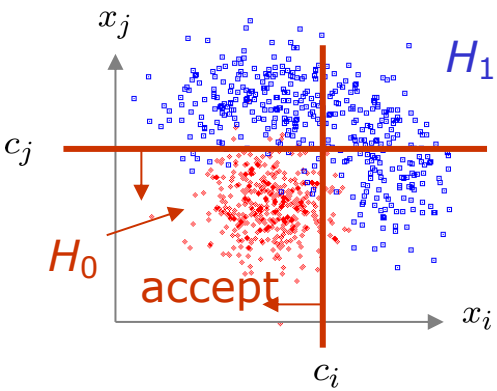
$\frac{f(x,y|H_s)}{f(x,y|H_{s+b})} > c$



Empirical parametric forms of decision boundaries

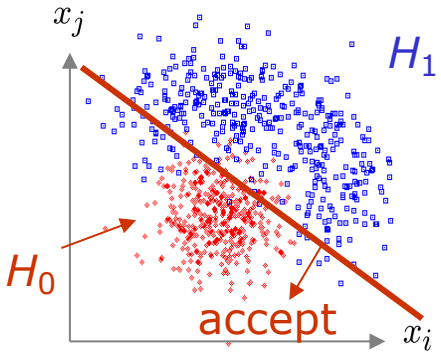
- Can in principle choose any type of Ansatz parametric shape

Rectangular cut



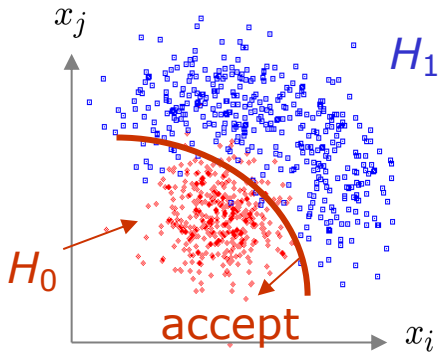
$$t(x) = \theta(x_j - c_j)\theta(x_i - c_i)$$

Linear cut



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

Non-linear cut

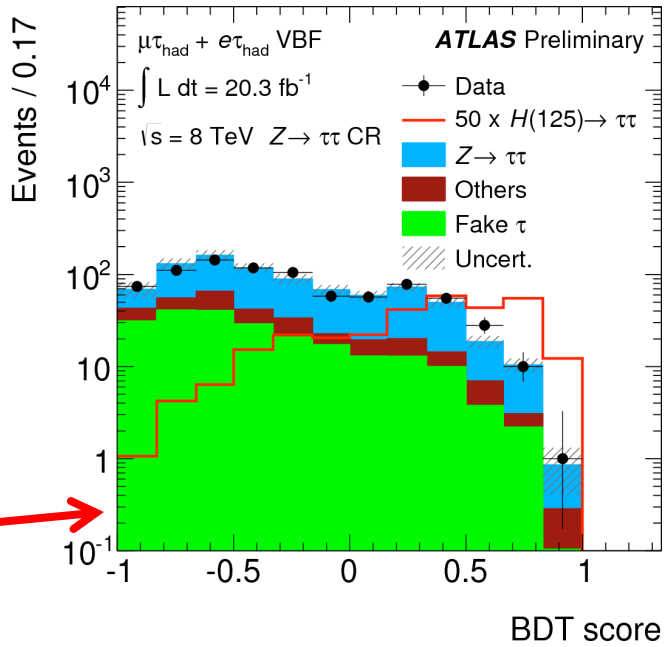


$$t(x) = \vec{a} \cdot \vec{x} + \vec{x} A \vec{x} + \ldots$$

- 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

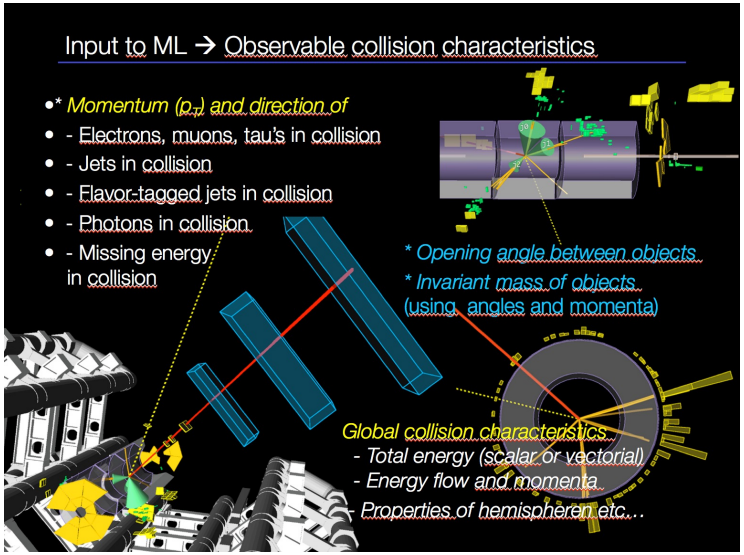
- A wide range of modern tools exist to perform supervised learning of a multivariate discriminant with the aim to approximate the optimal Neyman-Pearson discriminant.
  - Deep Learning, Boosted Decision Trees, GAN's etc etc.
- Variation in
  - Ansatz (empirical parametric form of discriminant)
  - Learning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

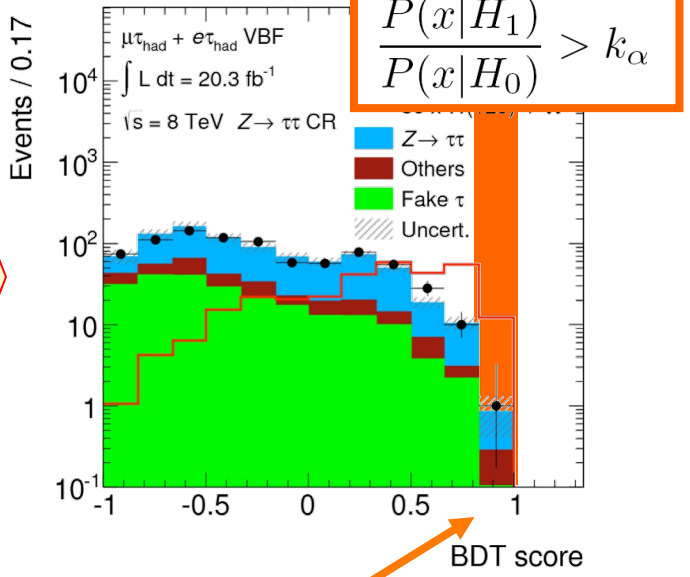


- In all cases output functions is functionally comparable to likelihood ratio discriminant (modulo some trivial transformations)

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 this data compression

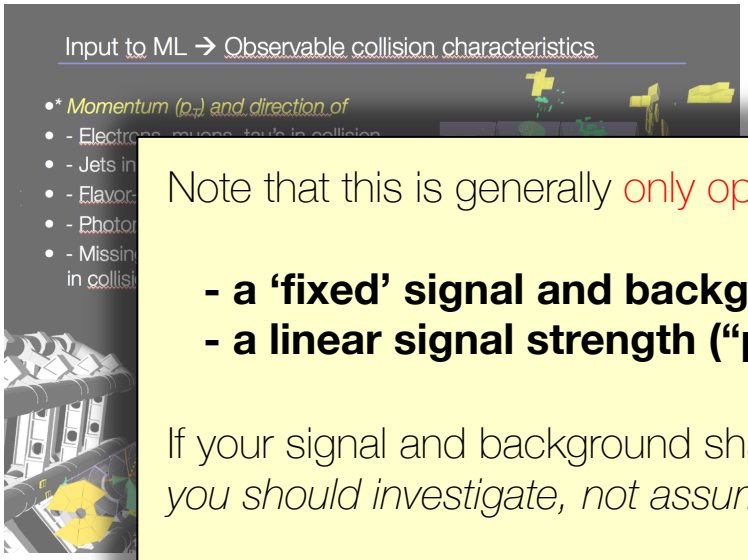


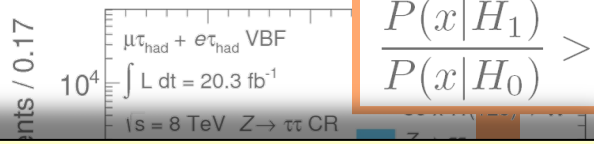


- In case the full discriminant distribution is tested → no loss of information
  - But need for pdf that model distribution
- But can also select high-signal region and perform simplified inference
  - e.g. counting model in that region

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 this data compression





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

Note that this is generally only optimal when you have

- a 'fixed' signal and background prediction
- a linear signal strength ("μ*S+B") in the entire phase space

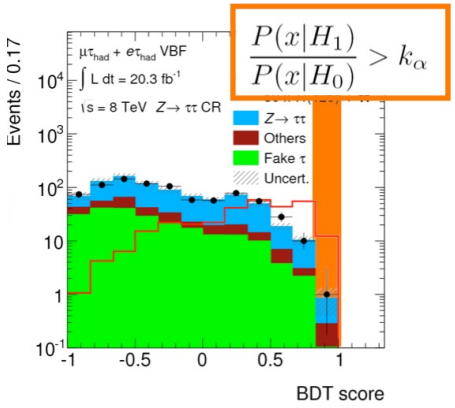
If your signal and background shapes have significant uncertainties → *you should investigate, not assume*

If there are significant interference effects with the signal (Effective Field Theory models, Offshell Higgs measurements) → *you should investigate, you should investigate, not assume*

- In cas
  is tes
  - Bu
- But can also select high-signal region and perform simplified inference
  - e.g. counting model in that region

Choosing the 'best' high-signal region

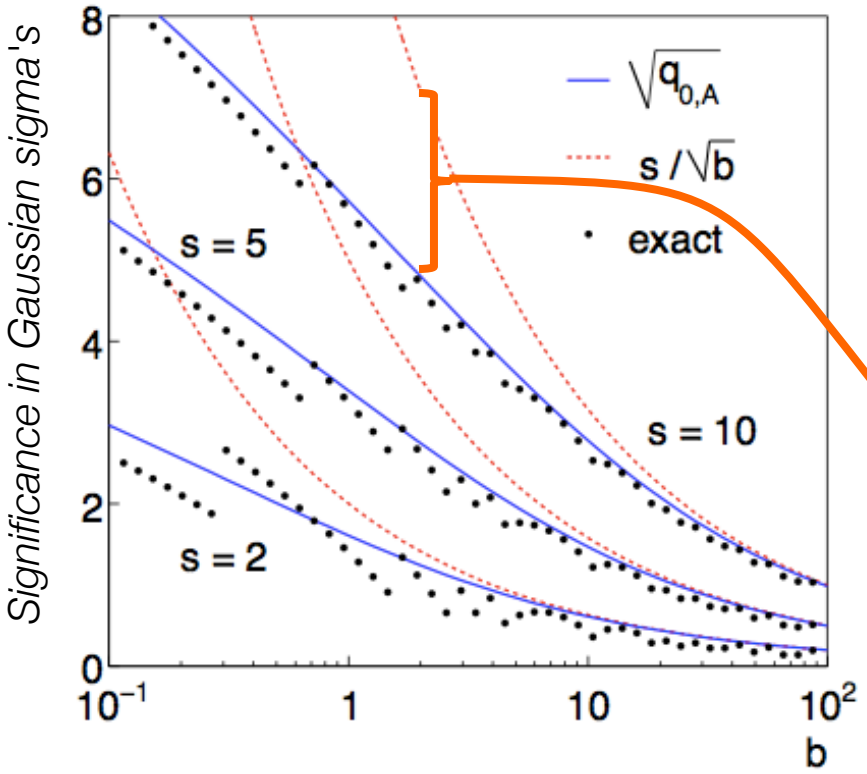
- A common scenario for searches in a low-statistics regime is to perform a simplified analysis
  1. Train MVA to obtain discriminant D
  2. Apply a cut on D
  3. Perform only a counting analysis



- 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/√b. For low-statistic searches s/√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: $\sqrt{2\left((s+b)\ln(1+s/b)-s\right)}$.
- For discovery, the traditional FOM s/√b *shows significant deviations from the 'exact' expected Poisson significance at low b*



$$\sqrt{q_{0,\mathrm{A}}} = \sqrt{2\left((s+b)\ln(1+s/b)-s\right)} .$$

$$= \frac{s}{\sqrt{b}}\left(1+\mathcal{O}(s/b)\right) .$$