Statistics & (a bit of) Machine Learning

W. Verkerke

Statistics & Machine Learning

- A really vast topic on which you could spend an entire week lecturing
- Have only 4 hours - so will make some selection of topics here
- Will mostly focus on statistics methods and model building – with a modest excursion into machine learning
- General idea of the course
  - start with simple models – focus on fundamental concepts for those (p-values, bayes vs frequentist)
  - then gradually make models more complex and look at how statistical procedures deal with these (nuisance parameters, systematic uncertainties), but also look on the practical side for phycisists – do you understand what happens and how can you debug and validate your complex fits
  - Excursion in multivariate methods and machine learning, when discussing multi-dimensional models (nice connection via NP lemma)

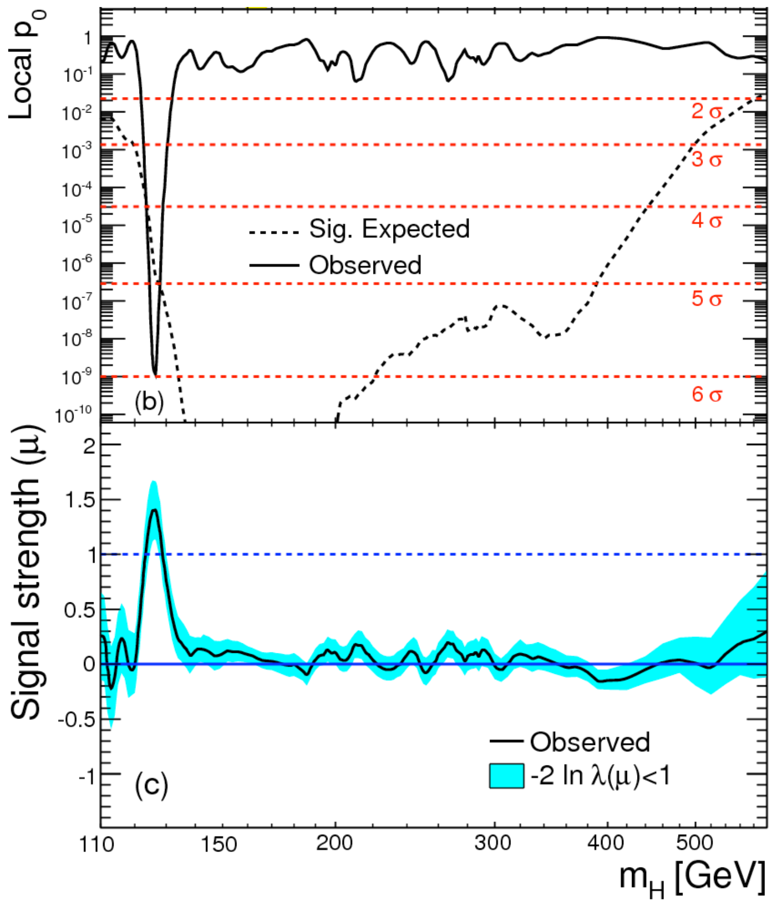
What do we want to know?

- **Physics questions we have…**
  - Does the (SM) Higgs boson exist?
  - What is its production cross-section?
  - What is its boson mass?

- **Statistical tests construct probabilistic statements: p(theo|data), or p(data|theo)**
  - Hypothesis testing (discovery)
  - (Confidence) intervals
    Measurements & uncertainties

- **Result: *Decision* based on tests**

  “As a layman I would now say: I think we have it”



How do we do this?

- Statistics: if you know distribution $f(x|\mu,\theta)$ for your observable(s) x in terms of your parameter of interest μ (and other parameters θ) then in principle solvable problem
  - In other words if $f(x|\mu,\theta)$ is known then problem is 'simple' ('just' follow prescription of statistical procedures)

- Particle physics: connection of theory (SM or its extension) with your parameter μ is highly non-trivially connected to your observables x.

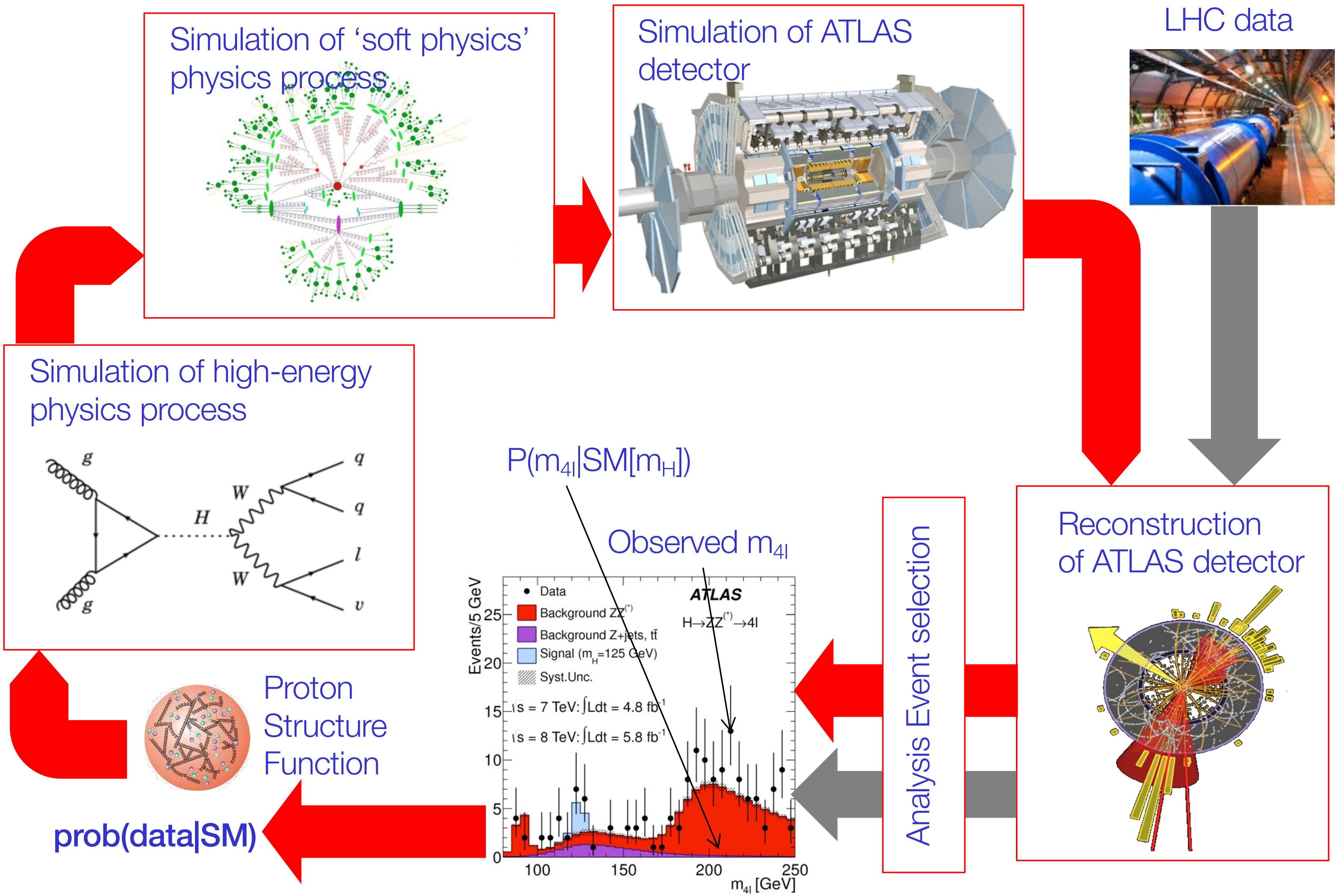
- Relation between x and μ can in almost all cases not be analytically formulated, but distribution $f(x|\mu,\theta)$ can be sampled through (chain of simulation packages)

Root of many of the complexities of HEP data analysis

- Simulation-based knowledge of $f(x|\mu,\theta)$ often approximate, often in ways that cannot be exactly quantified

Often the really thorny problems ('theory systematics') – root of problem is not statistical in nature, but must somehow be accounted for in statistical procedures

Particle physics data analysis – chain of simulation steps



Simulation of ‘soft physics’ physics process

Simulation of ATLAS detector

LHC data

Simulation of high-energy physics process

Reconstruction of ATLAS detector

Analysis Event selection

$P(m_{4l}|SM[m_H])$

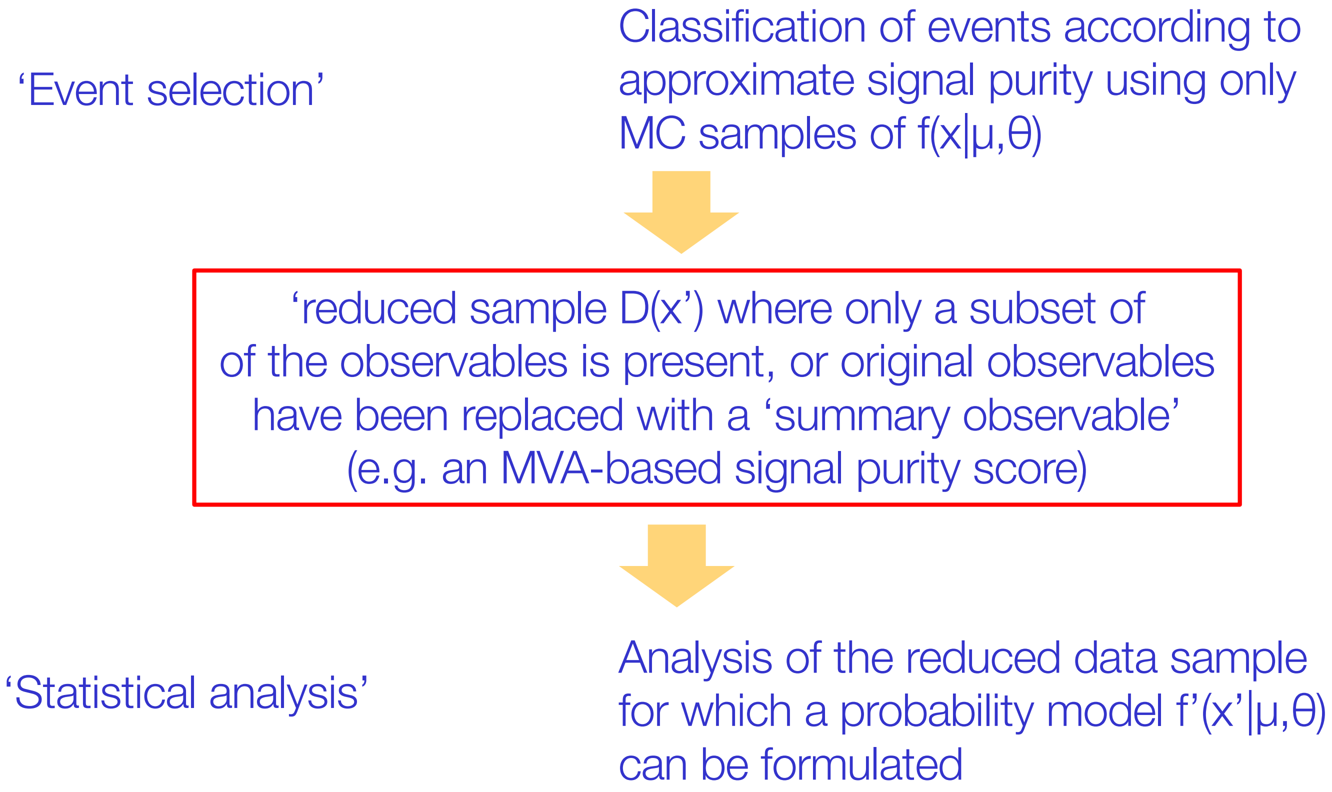
Observed m_{4l}

Proton Structure Function

prob(data|SM)

Statistical analysis with HEP data – a multi-step approach

- To make HEP data analysis tractable with only samples of $f(x|\mu,\theta)$ known (instead of function itself), inference of μ from $f(x|\mu,\theta)$ often performed as a multi-step process



'Event selection'

Classification of events according to approximate signal purity using only MC samples of $f(x|\mu,\theta)$

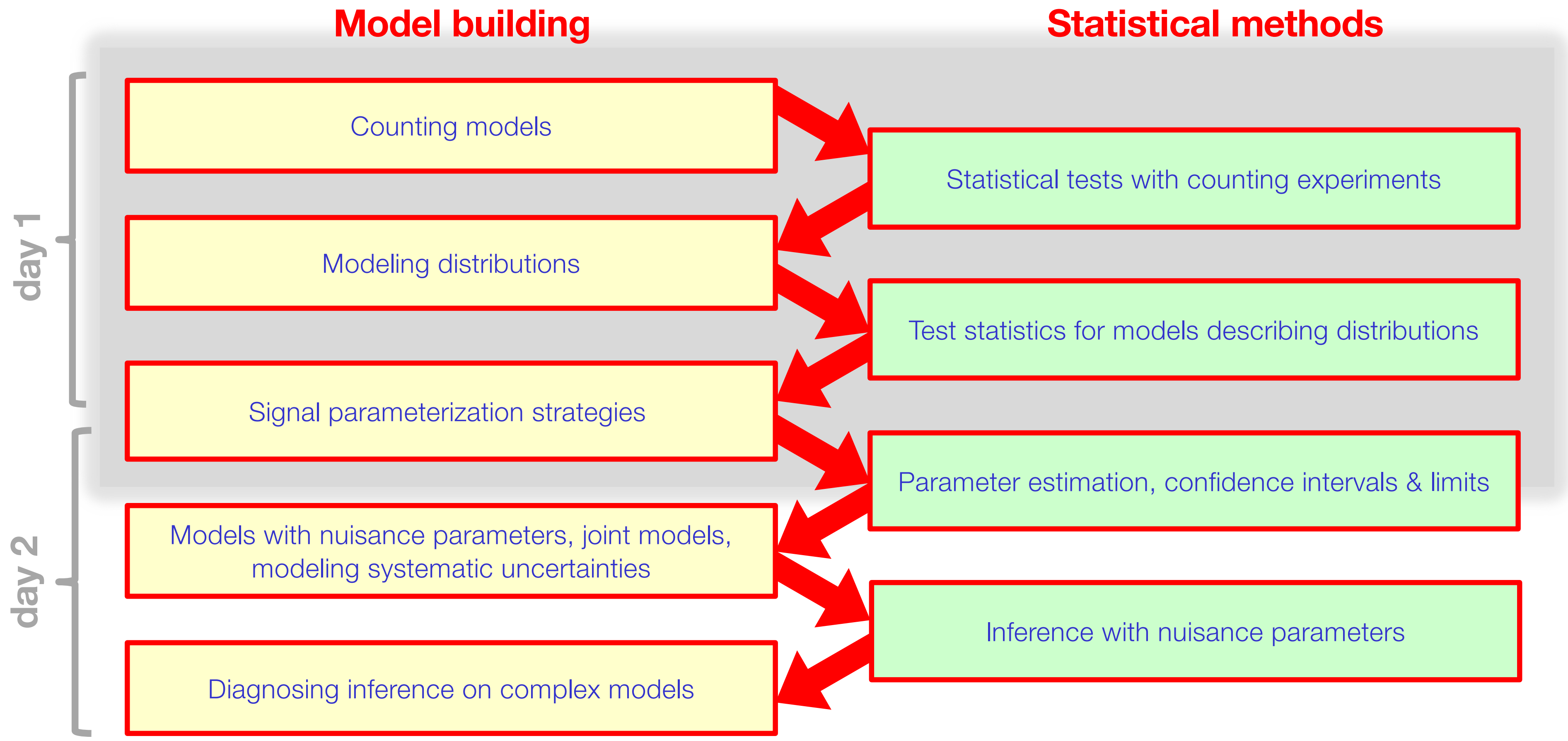
'reduced sample D(x') where only a subset of of the observables is present, or original observables have been replaced with a 'summary observable' (e.g. an MVA-based signal purity score)

'Statistical analysis'

Analysis of the reduced data sample for which a probability model $f'(x'|\mu,\theta)$ can be formulated

Roadmap of this course

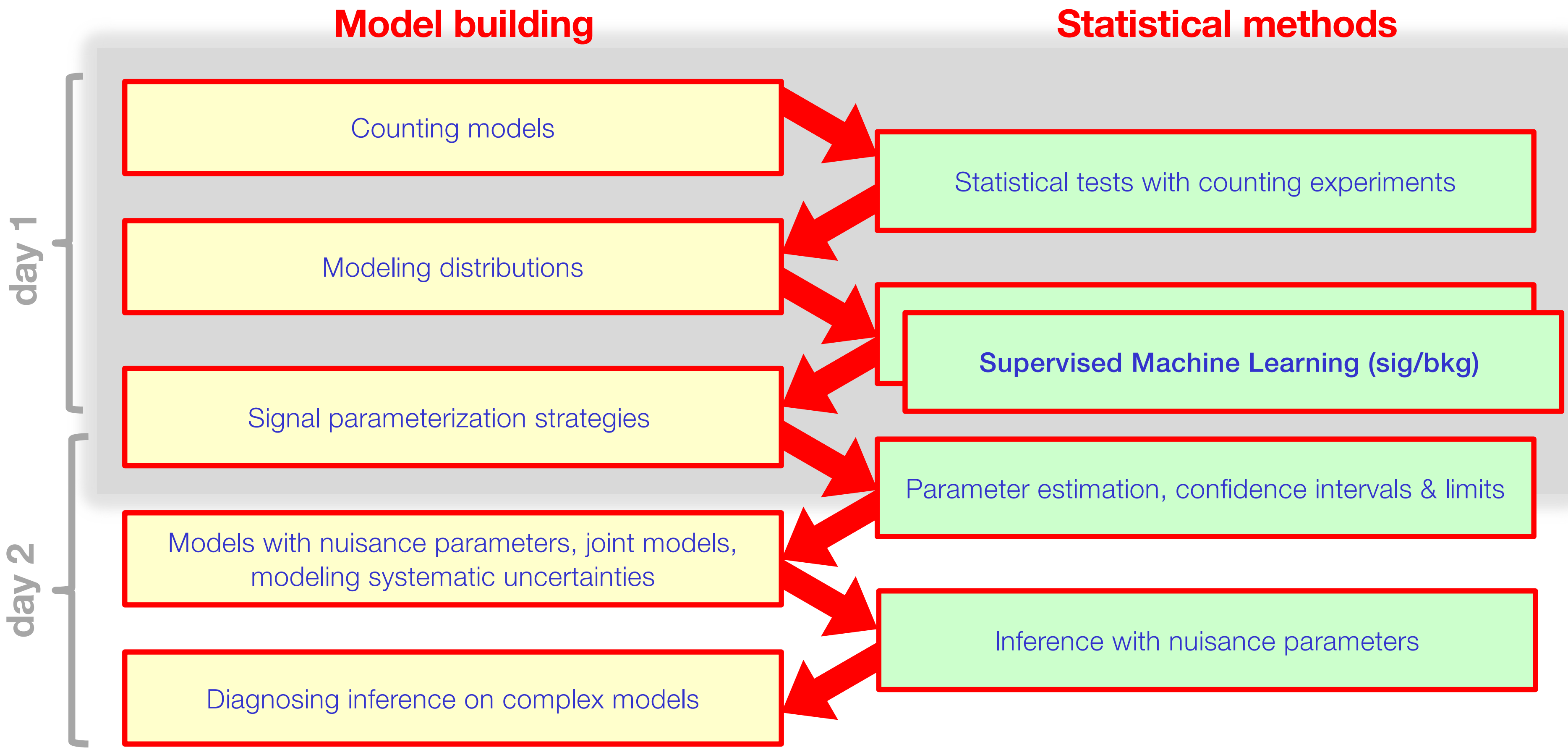
- Start with basics, gradually build up to complexity



Model building
Statistical methods
day 1
day 2
Counting models
Statistical tests with counting experiments
Modeling distributions
Test statistics for models describing distributions
Signal parameterization strategies
Parameter estimation, confidence intervals & limits
Models with nuisance parameters, joint models, modeling systematic uncertainties
Inference with nuisance parameters
Diagnosing inference on complex models

Roadmap of this course

- Start with basics, gradually build up to complexity



Model building

Statistical methods

day 1

Counting models

Statistical tests with counting experiments

Modeling distributions

Supervised Machine Learning (sig/bkg)

Signal parameterization strategies

day 2

Parameter estimation, confidence intervals & limits

Models with nuisance parameters, joint models, modeling systematic uncertainties

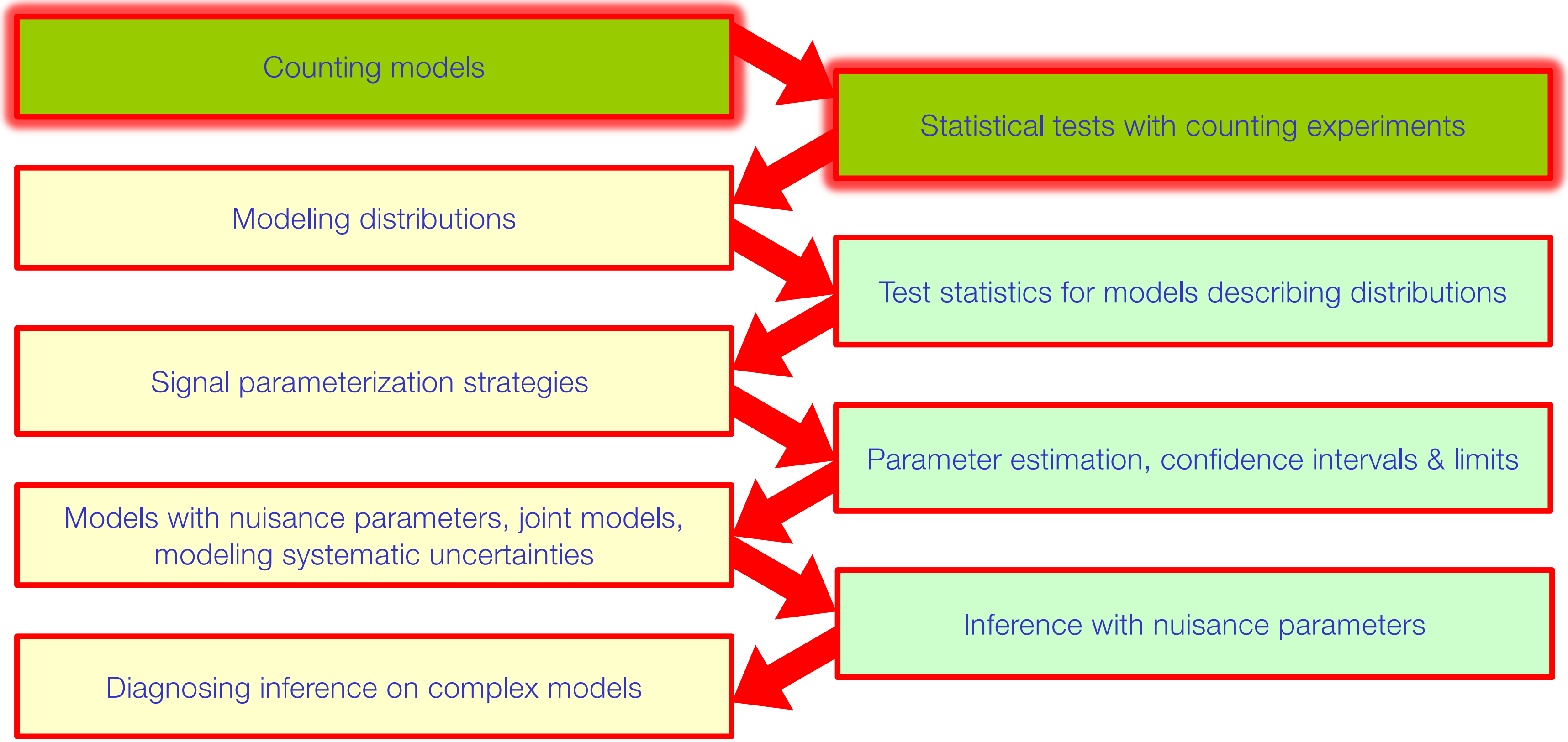
Inference with nuisance parameters

Diagnosing inference on complex models

Roadmap of this course

- Start with basics, gradually build up to complexity

Model building **Statistical methods**



Counting models

Statistical tests with counting experiments

Modeling distributions

Test statistics for models describing distributions

Signal parameterization strategies

Parameter estimation, confidence intervals & limits

Models with nuisance parameters, joint models, modeling systematic uncertainties

Inference with nuisance parameters

Diagnosing inference on complex models

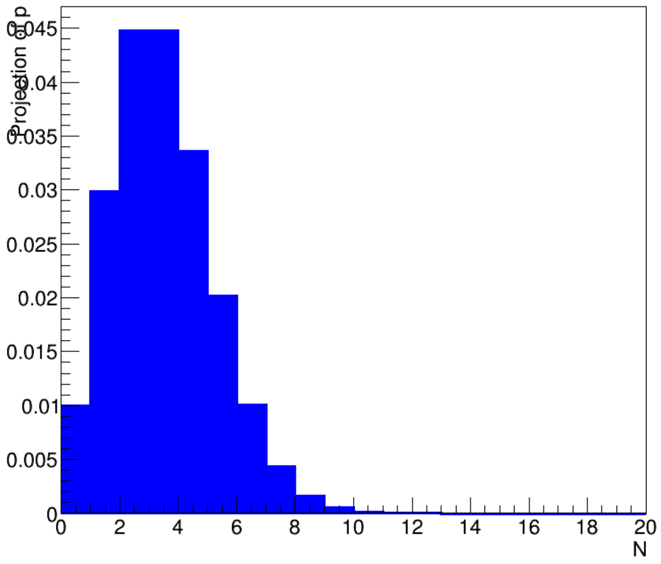
Counting models

- Central concept in statistics is the ‘**probability model**’
- *A probability model assigns a probability to each possible experimental outcome*.
- Example: a HEP counting experiment

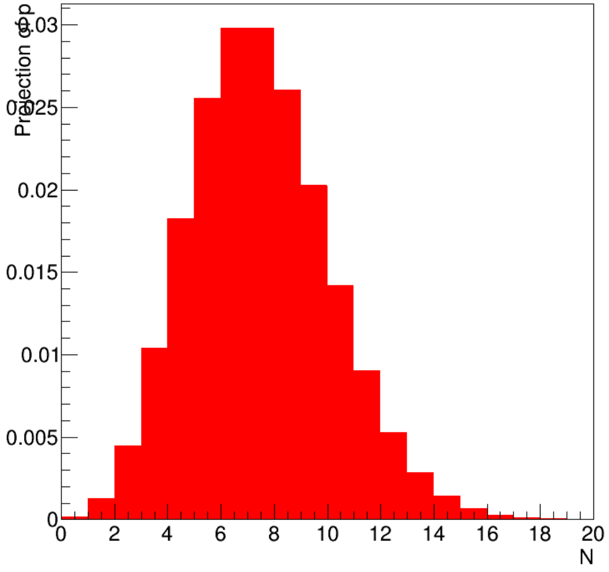
$$P(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$$

  - Count number of ‘events’ in a fixed time interval → Poisson distribution
  - Given the *expected event count*, the probability model is fully specified

μ=3 (“bkg only”)



μ=7 (“bkg+signal”)



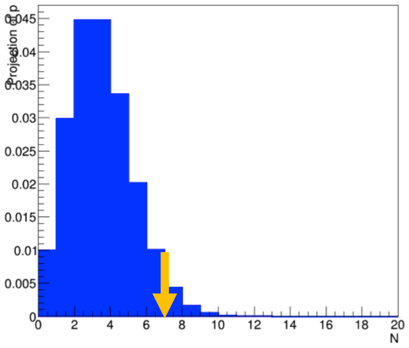
→ Probability of outcome

→ Experimental outcome

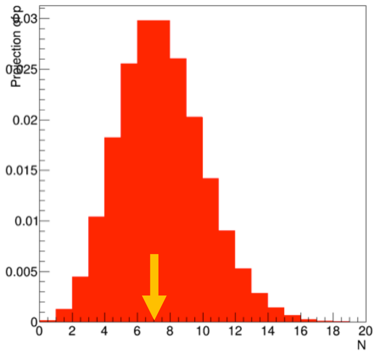
Probabilities vs conditional probabilities

- Note that probability models strictly give *conditional* probabilities (with the condition being that the underlying hypothesis is true)

μ=3 (“bkg only”)



μ=7 (“bkg+signal”)



Definition:
P(data|hypo) is called the ***likelihood***

$$P(N) \rightarrow P(N \mid H_{bkg}) \qquad P(N) \rightarrow P(N \mid H_{sig+bkg})$$

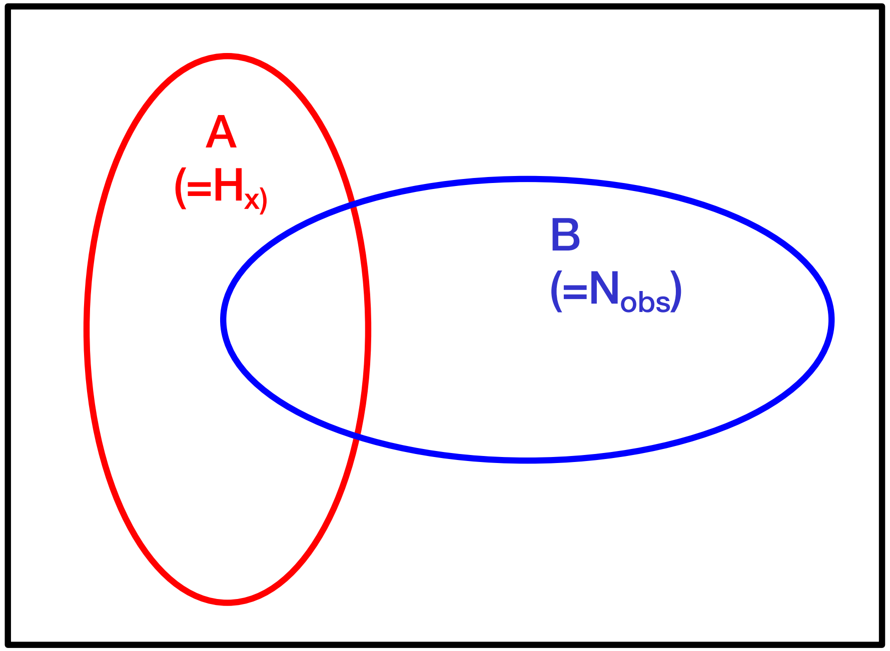
- Suppose we measure N=7 then can calculate

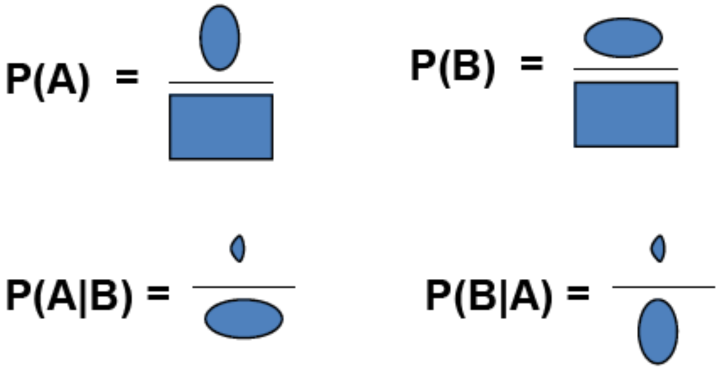
$L(N=7|H_{bkg})=2.2\%$ $L(N=7|H_{sig+bkg})=14.9\%$

- *Data is more likely under sig+bkg hypothesis than bkg-only hypo*
- Is this what we want to know? Or do we want to know $L(H_{s+b}|N=7)$?

Inverting the conditionality on probabilities

- Do $L(7|H_b)$ and $L(7|H_{sb})$ provide you enough information to calculate $P(H_b|7)$ and $P(H_{sb}|7)$
- **No!**
- Image the ‘whole space’ and two subsets A and B

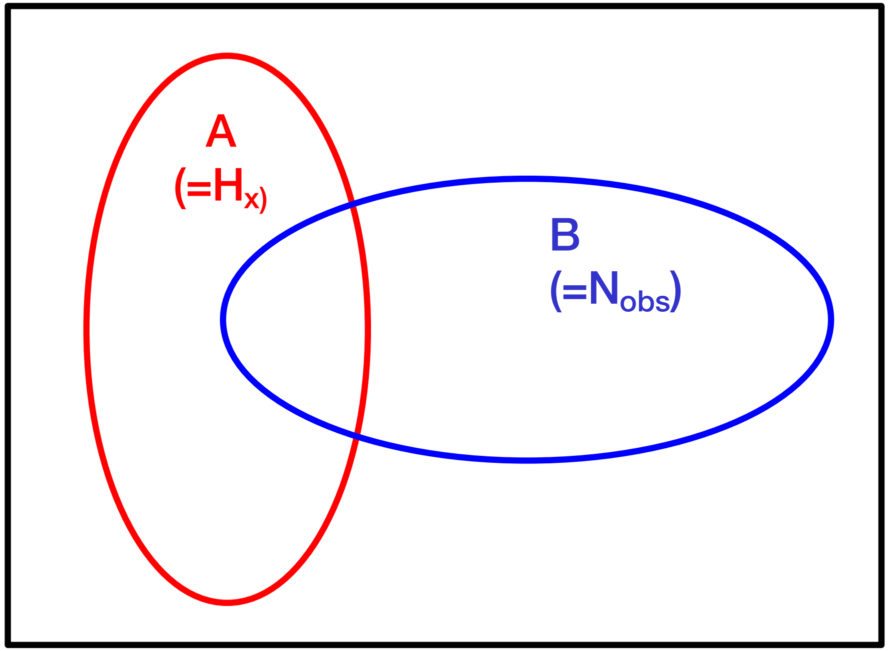


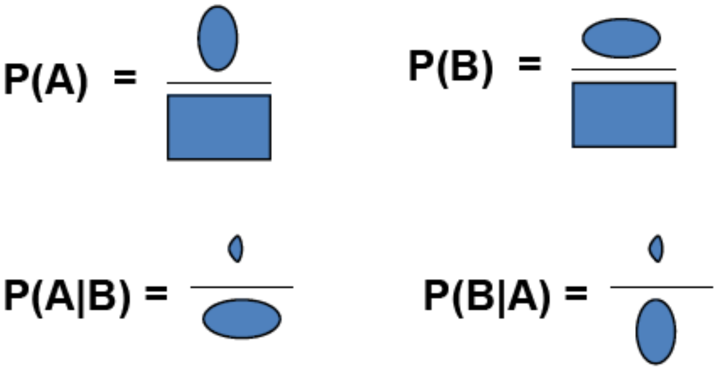


$P(A|B) \neq P(B|A)$

$P(7|H_b) \neq P(H_b|7)$

Inverting the conditionality on probabilities





P(A) =

P(B) =

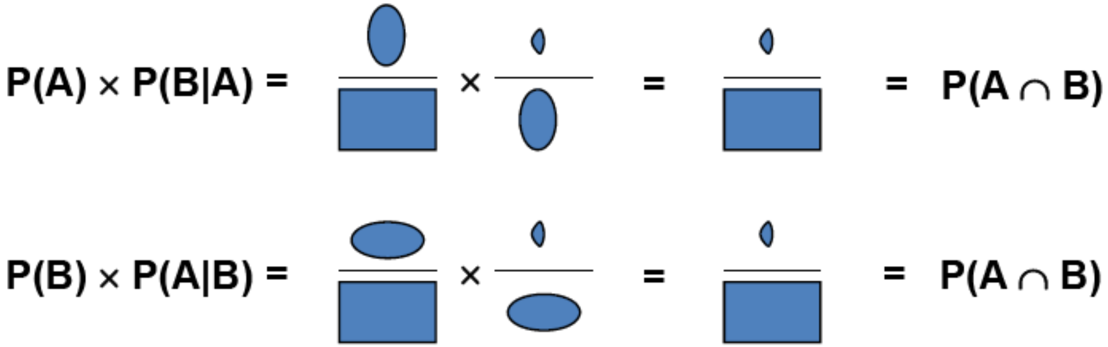
P(A|B) =

P(B|A) =

P(A|B) ≠ P(B|A)

but you can deduce their relation

⇒ P(B|A) = P(A|B) × P(B) / P(A)



P(A) × P(B|A) = × = = P(A ∩ B)

P(B) × P(A|B) = × = = P(A ∩ B)

Inverting the conditionality on probabilities

- This conditionality inversion relation is known as **Bayes Theorem**

$$P(B|A) = P(A|B) \times P(B)/P(A)$$

Essay "Essay Towards Solving a Problem in the Doctrine of Chances" published in Philosophical Transactions of the Royal Society of London in 1764

Thomas Bayes (1702-61)

- And choosing A=data and B=theory

$$P(\text{theo}|\text{data}) = P(\text{data}|\text{theo}) \times P(\text{theo}) / P(\text{data})$$

- *Return to original question:*

  Do you $L(7|H_b)$ and $L(7|H_{sb})$ provide you enough information to calculate $P(H_b|7)$ and $P(H_{sb}|7)$

- **No!** → Need P(A) and P(B) → **Need $P(H_b)$, $P(H_{sb})$ and P(7)**

Inverting the conditionality on probabilities

- **What is P(data)?**

P(theo|data) = P(data|theo) × P(theo) / **P(data)**

- It is the probability of the data under *any* hypothesis
  - For Example for two competing hypothesis H_b and H_{sb}

$$P(N) = L(N|H_b)P(H_b) + L(N|H_{sb})P(H_{sb})$$

and generally for N hypotheses

$$P(N) = \Sigma_i\, P(N|H_i)P(H_i)$$

- Bayes theorem reformulated using law of total probability

$$P(\text{theo}|\text{data}) = \frac{L(\text{data}|\text{theo}) \times P(\text{theo})}{\Sigma_i\, L(\text{data}|\text{theo-i})P(\text{theo-i})}$$

- *Return to original question:* Do you $L(7|H_b)$ and $L(7|H_{sb})$ provide you enough information to calculate $P(H_b|7)$ and $P(H_{sb}|7)$
No! → Still need $P(H_b)$ and $P(H_{sb})$

Prior probabilities

- What is the **meaning** of $P(H_b)$ and $P(H_{sb})$?
  - They are the probability assigned to hypothesis H_b *prior to the experiment*.
- What are the **values** of $P(H_b)$ and $P(H_{sb})$?
  - Can be result of an earlier measurement
  - Or more generally (e.g. when there are no prior measurement) they quantify *a prior degree of belief* in the hypothesis

- Example – suppose prior belief $P(H_{sb})$=50% and $P(H_b)$=50%

$$P(H_{sb}|N=7) = \frac{P(N=7|H_{sb}) \times P(H_{sb})}{[\ P(N=7|H_{sb})P(H_{sb})+P(N=7|H_b)P(H_b)\ ]}$$

$$= \frac{0.149 \times 0.50}{[\ 0.149\times0.5+0.022\text{x}0.5\ ]} = 87\%$$

- Observation N=7 strengthens belief in hypothesis H_{sb} (and weakens belief in H_b → 13%)

Interpreting probabilities

- We have seen

  probabilities assigned observed experimental outcomes
  (probability to observed 7 events under some hypothesis)

  probabilities assigned to hypotheses
  (prior probability for hypothesis H_{sb} is 50%)

  which are conceptually different.

- How to interpret probabilities – two schools

  Bayesian probability = (subjective) degree of belief

  P(theo|data)
  P(data|theo)

  Frequentist probability = fraction of outcomes in future repeated identical experiments

  P(data|theo)

  "If you'd repeat this experiment identically many times, in a fraction P you will observe the same outcome"

Interpreting probabilities

- Frequentist:
  Constants of nature are fixed – you cannot assign a probability to these. Probability are restricted to observable experimental results
  - "The Higgs either exists, or it doesn't" – you can't assign a probability to that
  - Definition of P(data|hypo) is objective (and technical)
- Bayesian:
  Probabilities can be assigned to constants of nature
  - Quantify your *belief* in the existence of the Higgs – can assign a probablity
  - But is can very difficult to assign a meaningful number (e.g. Higgs)
- Example of weather forecast

  Bayesian: "*The probability it will rain tomorrow is 95%*"
  - Assigns probability to constant of nature ("rain tomorrow")
    P(rain-tomorrow|satellite-data) = 95%

  Frequentist: *"If it rains tomorrow,*
  95% of time satellite data looks like what we observe now"
  - Only states P(satellite-data|rain-tomorrow)

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

$$O_{posterior} \equiv \frac{L(x|H_{sb})P(H_{sb})}{L(x|H_b)P(H_b)} = \overbrace{\frac{L(x|H_{sb})}{L(x|H_b)}}^{\text{‘Bayes Factor’ K multiplies prior odds}} O_{prior}$$

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

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

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

Taking decisions based on your result

- What are you going to do with the results of your measurement?
- Usually basis for a *decision*
  - Science: declare discovery of Higgs boson (or not), make press release, write new grant proposal
  - Finance: buy stocks or sell
- Suppose you believe P(Higgs|data)=99%.
- Should declare discovery, make a press release?
  A: *Cannot be determined from the given information!*
- Need in addition: the utility function (or cost function),
  - The cost function specifies the relative costs (to You) of a Type I error (declaring model false when it is true) and a Type II error (not declaring model false when it is false).

Taking decisions based on your result

- Thus, your *decision*, such as where to invest your time or money, requires two subjective inputs:

  Your prior probabilities, and

  the relative costs to You of outcomes.

- Statisticians often focus on decision-making;
  in HEP, the tradition thus far is to communicate experimental results (well) short of formal decision calculations.

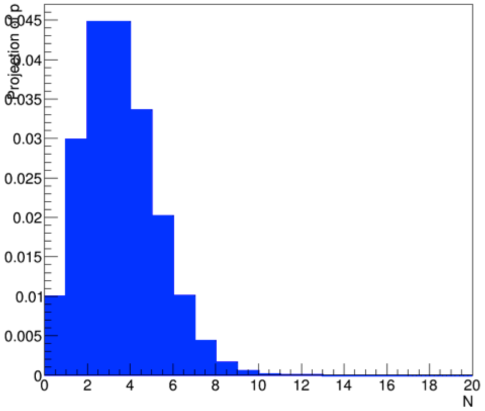
- Costs can be difficult to quantify in science.
  - What is the cost of declaring a false discovery?
  - Can be high (“Fleischman and Pons”), but hard to quantify
  - What is the cost of missing a discovery (“Nobel prize to someone else”), but also hard to quantify

[Recap]

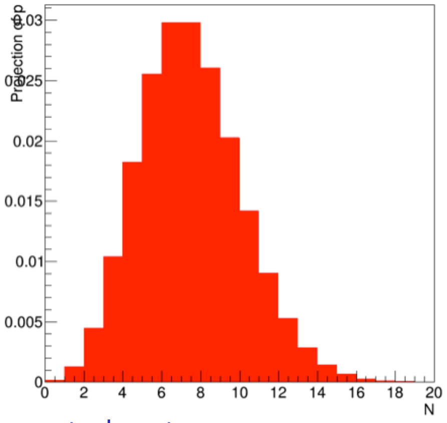
Summary on statistical test with simple hypotheses

- 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

μ=3 ("bkg only")



μ=7 ("bkg+signal")



→ 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)

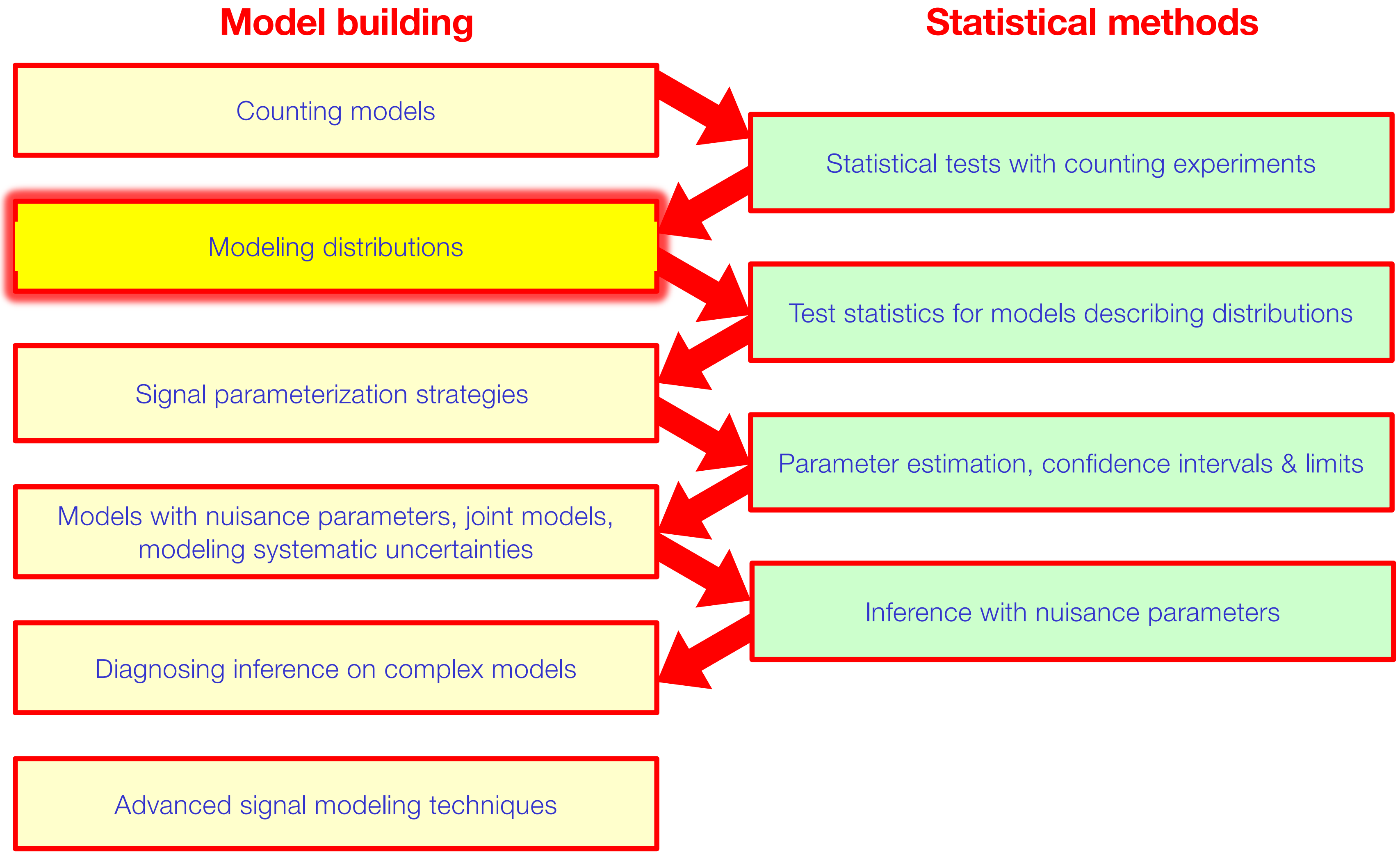
- In principle, *any potentially complex measurement (for Higgs, SUSY, top quarks) can ultimately take this a simple form.* But there is some 'pre-work' to get here – examining (multivariate) discriminating distributions → Now try to incorporate that

Model building 2

Modelling distributions – template based models or analytical models

Roadmap of this course

- Start with basics, gradually build up to complexity



Model building

Statistical methods

Counting models

Statistical tests with counting experiments

Modeling distributions

Test statistics for models describing distributions

Signal parameterization strategies

Parameter estimation, confidence intervals & limits

Models with nuisance parameters, joint models, modeling systematic uncertainties

Inference with nuisance parameters

Diagnosing inference on complex models

Advanced signal modeling techniques

Discriminating observables & counting experiments

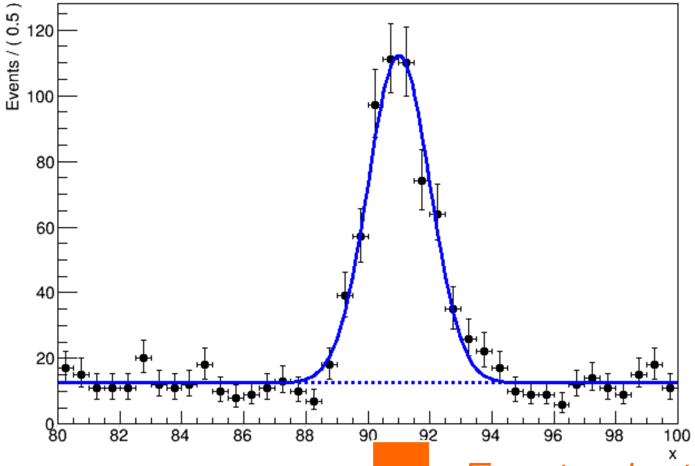
- HEP experimental data usually has many discriminating observables that carry information that can distinguish signal from background hypothesis
- In principle can use them all directly in an elaborate hypothesis test.
  - But would need to formulate a model that describe the expected distribution of 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

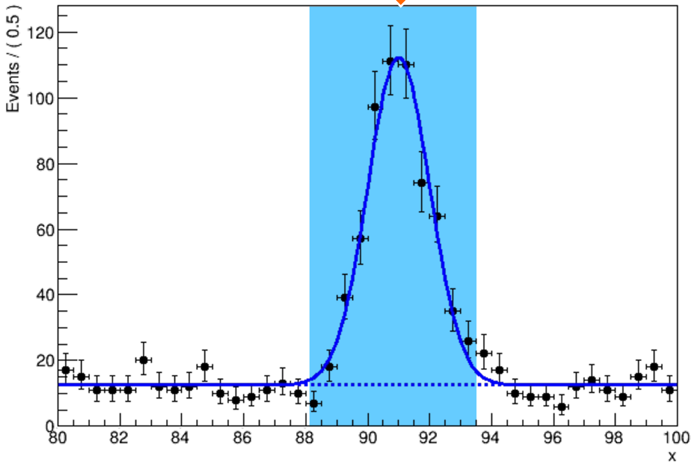
- Example 1 – **Discrimination in selection stage only**

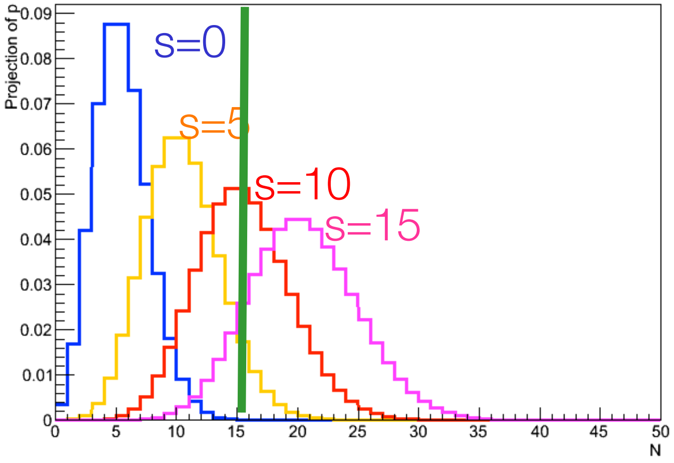
NB1: All discriminating power in selection step, none in inference step. *This is a design choice!*

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



Event selection: reduce sample size and dimensionality



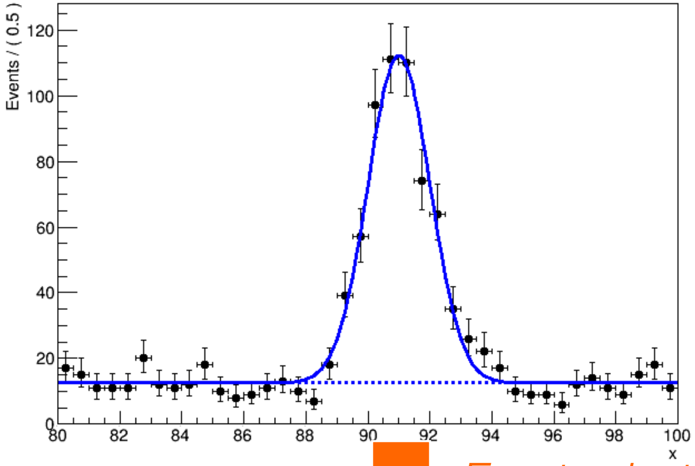


Formulation of probability model of reduced sample: Poisson(N|s+b)

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

Modeling discriminating observables

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



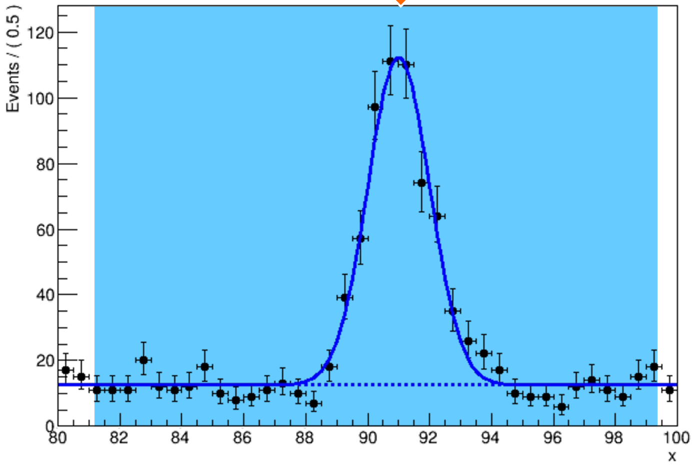
NB1: Most discrimination power in inference step.
This is again design choice!

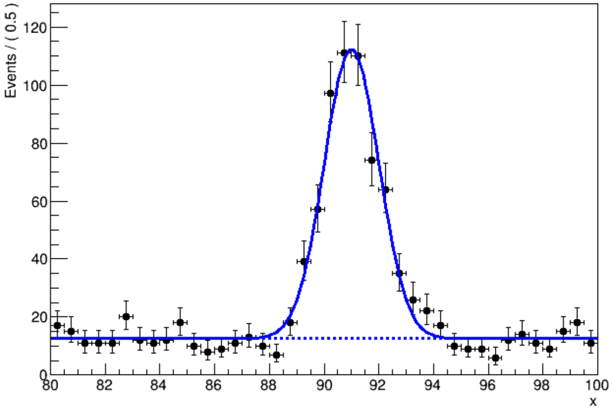
NB2: Optimal selection less critical

NB3: Correct description of selected sample more complex

Event selection: reduce sample size and dimensionality

Statistical inference:
L(data|hypo)=something

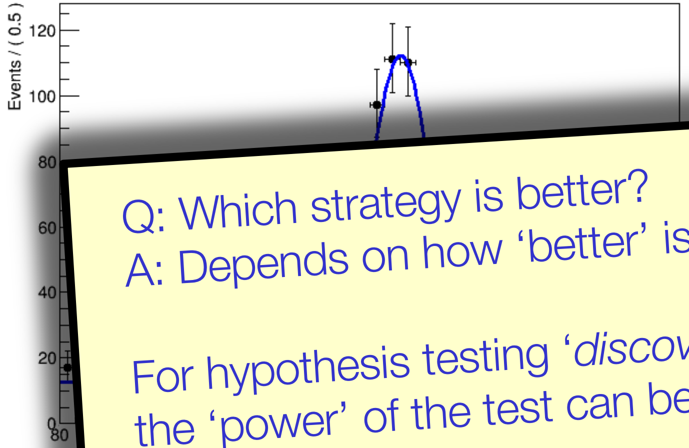




Formulation of probability model of reduced sample:
*Nbkg*Uniform(x) +Nsig*Gaussian(x)*

Modeling discriminating observables

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



NB1: Most discrimination power in inference step.

This is again design choice!

mple

:

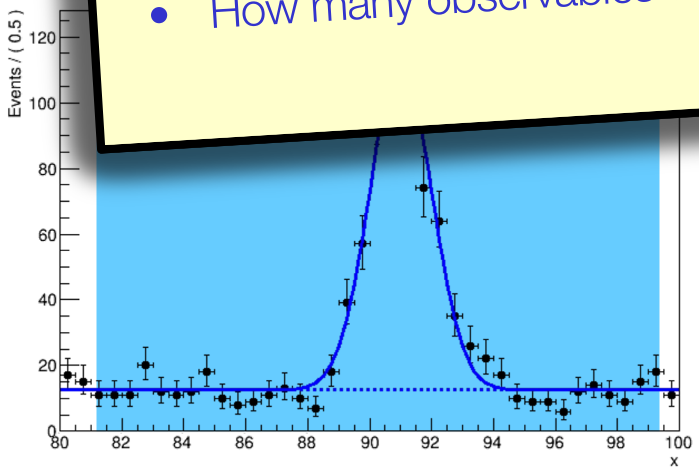
thing

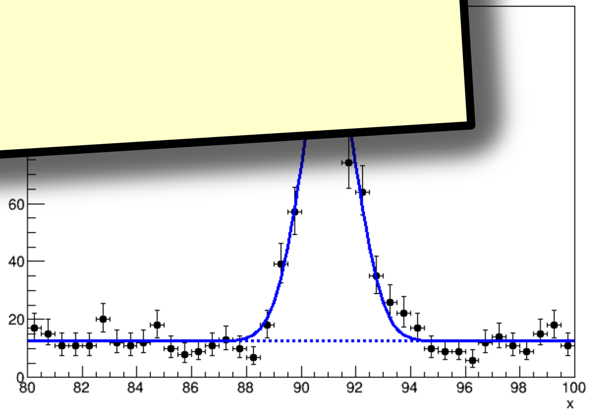
Q: Which strategy is better?
A: Depends on how 'better' is defined?

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)*

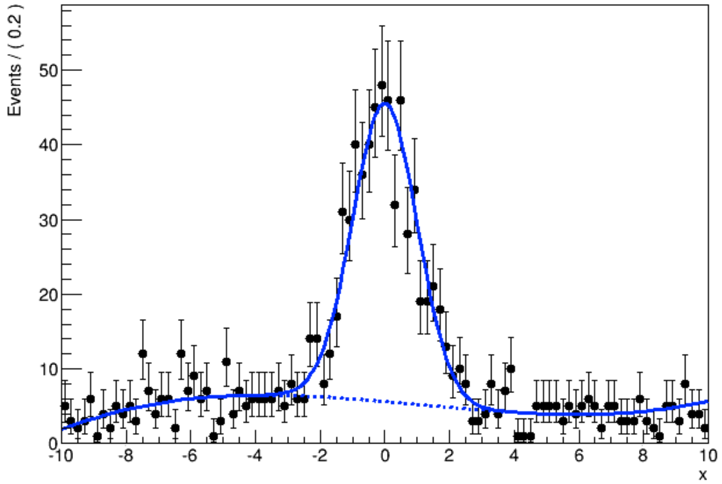
Formulating probability models for discriminating observables

- 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)

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

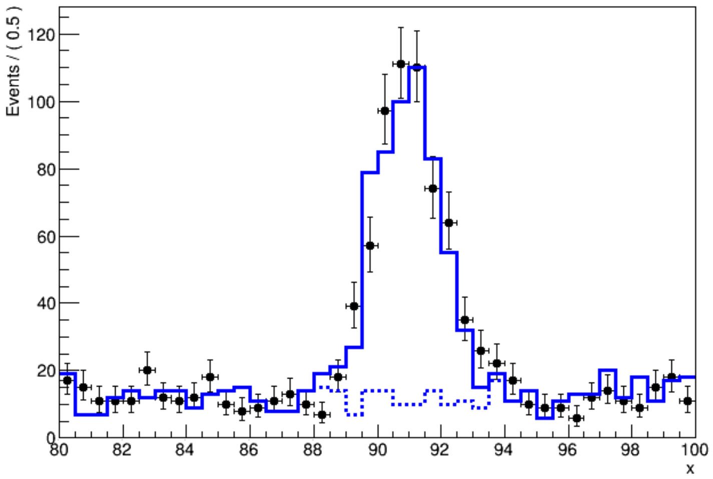


$$B(x) = a_0+a_1x+a_2x^2+a_3x^3\ldots$$

Drawbacks:

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

Simulation-based models



B(x) = histogram

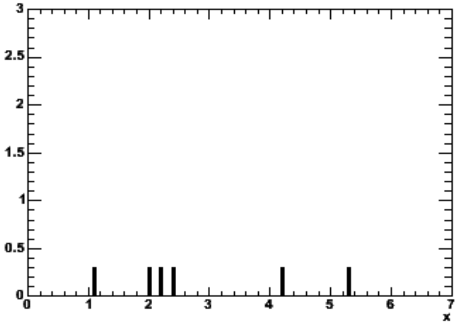
Drawbacks:

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

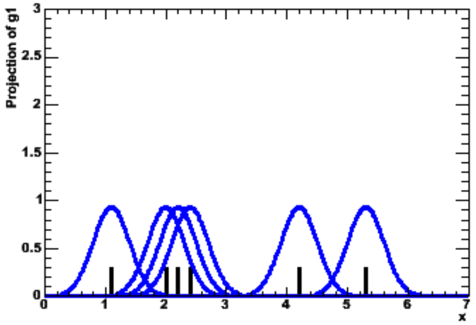
Modeling low-statistics simulation predictions

- 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

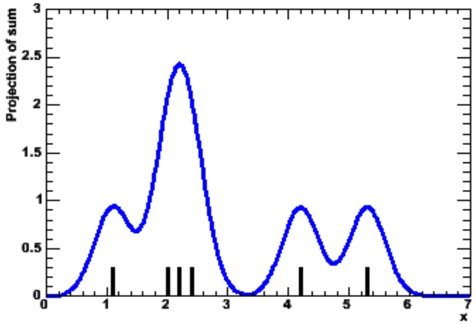
Sample of events



Gaussian probability distributions for each event

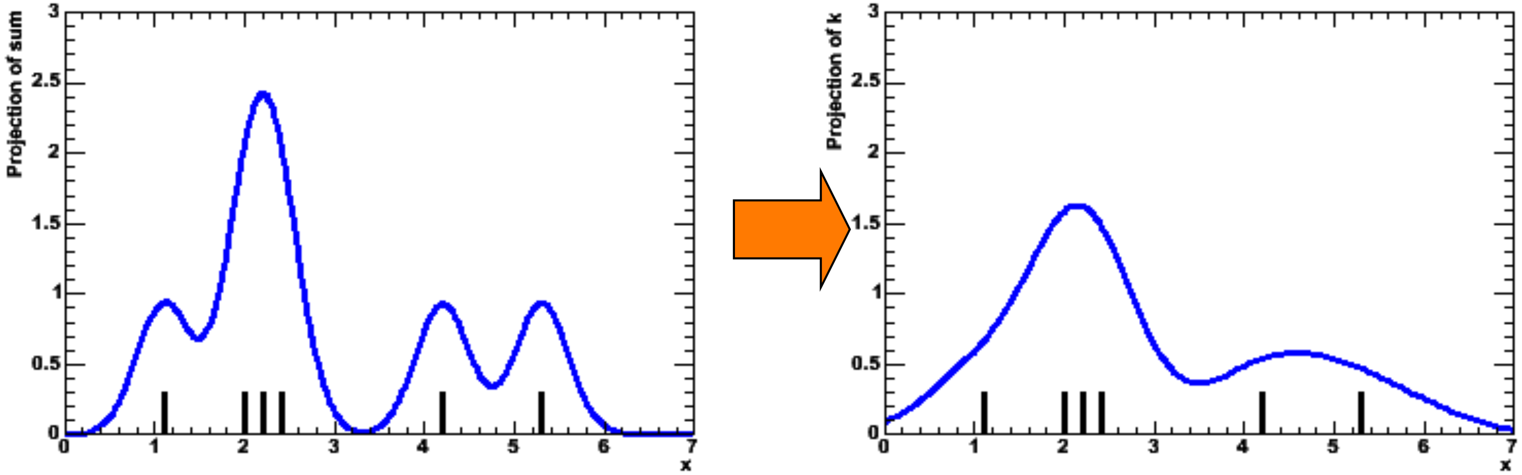


Summed probability distribution for all events in sample



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



Static Kernel (with of all Gaussian identical)

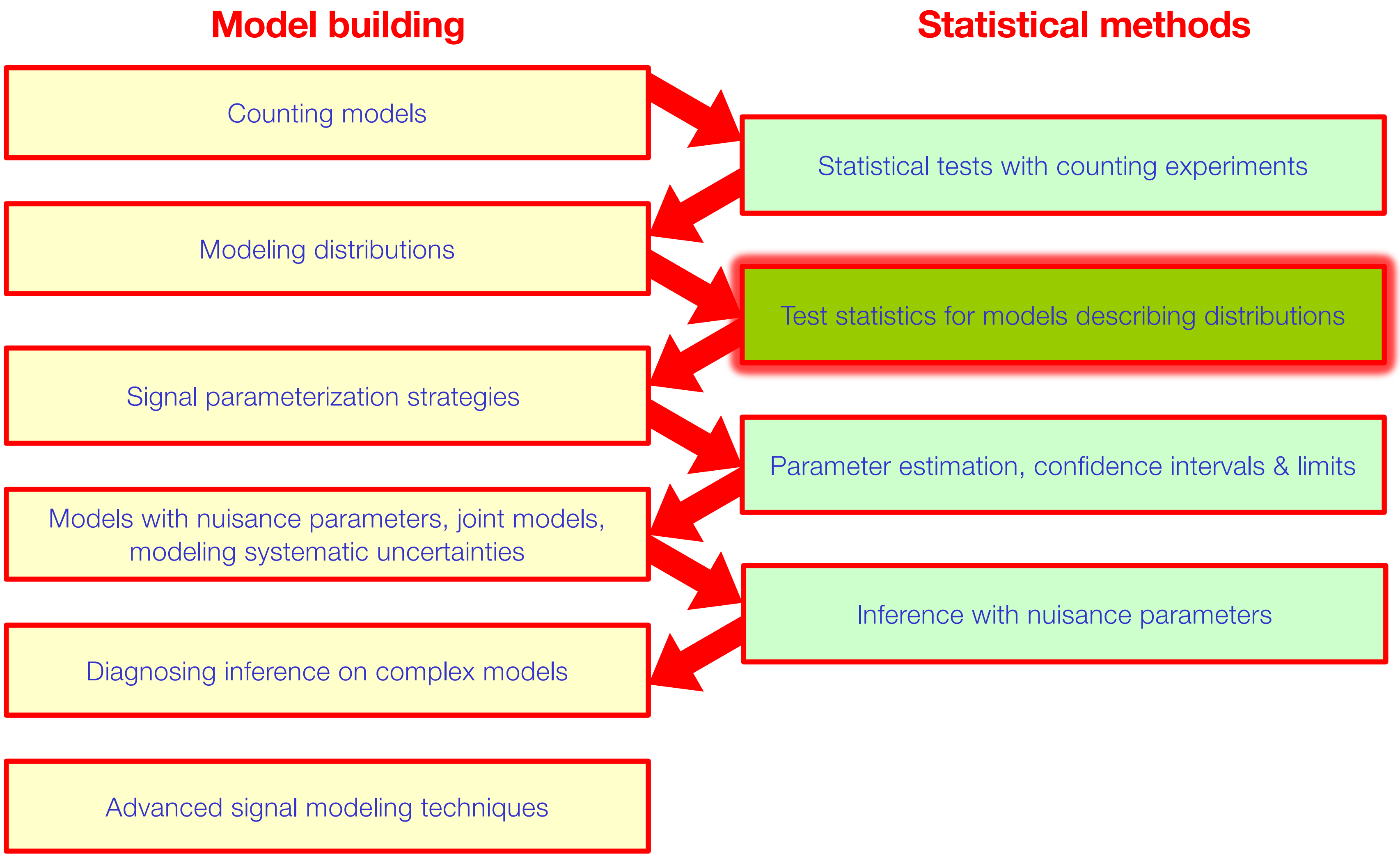
Adaptive Kernel (width of all Gaussian depends on local density of events)

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

Roadmap of this course

- Start with basics, gradually build up to complexity



Model building

Statistical methods

Counting models

Statistical tests with counting experiments

Modeling distributions

Test statistics for models describing distributions

Signal parameterization strategies

Parameter estimation, confidence intervals & limits

Models with nuisance parameters, joint models, modeling systematic uncertainties

Inference with nuisance parameters

Diagnosing inference on complex models

Advanced signal modeling techniques

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

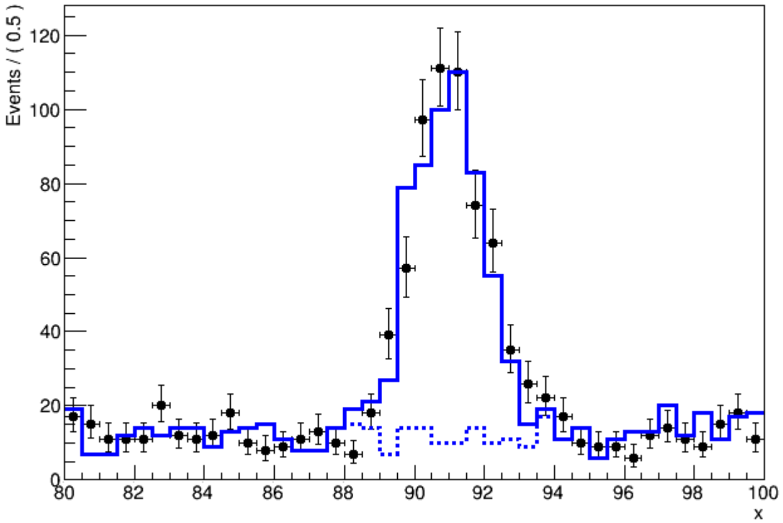
$$P(H_{s+b}\mid\vec{N})=\frac{L(\vec{N}\mid H_{s+b})P(H_{s+b})}{L(\vec{N}\mid H_{s+b})P(H_{s+b})+L(\vec{N}\mid H_{b})P(H_{b})}$$

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?**



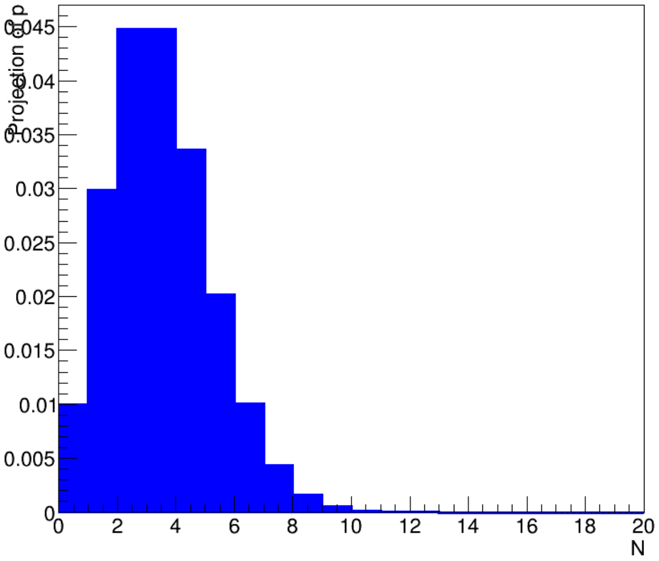
$$L(\vec{N} \mid H_b) = \prod_i Poisson(N_i \mid \tilde{b}_i)$$

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

- **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)

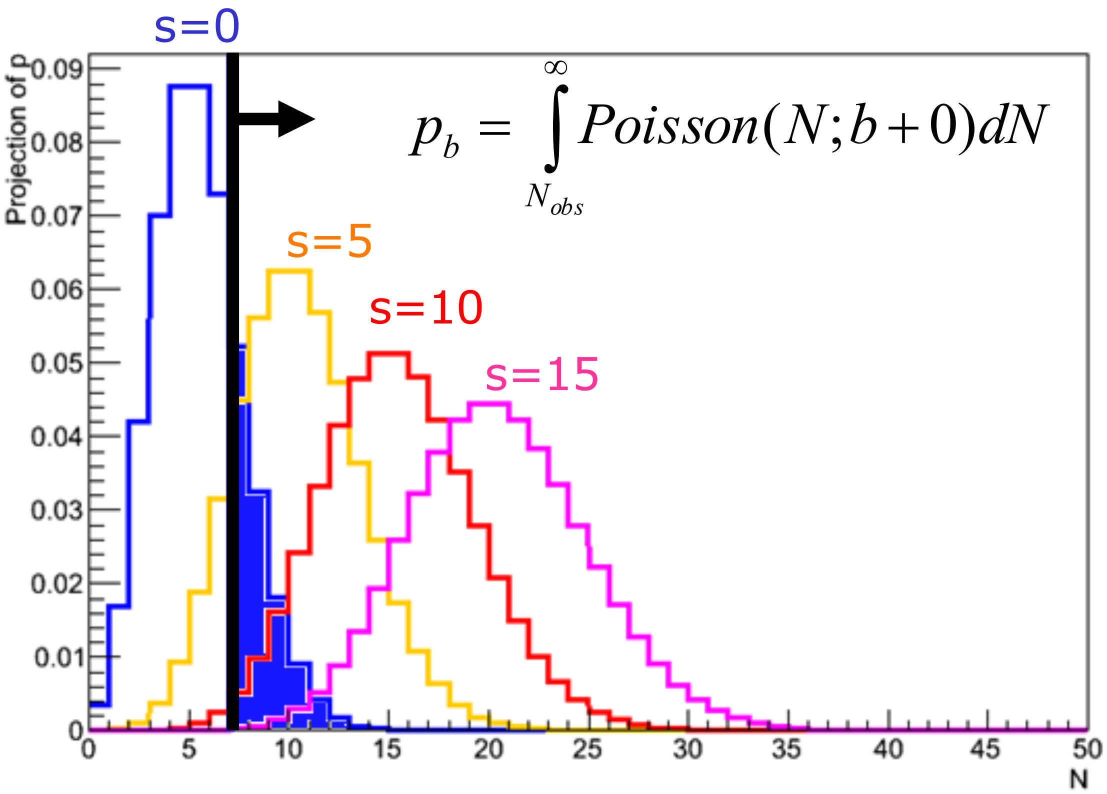


Test statistic T(N)=Nobs orders observed events count by estimated signal yield

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



$$p_b = \int_{N_{obs}}^{\infty} Poisson(N; b+0)dN \quad (=0.23)$$

Ordering distributions by ‘signal-likeness’ aka ‘extremity’

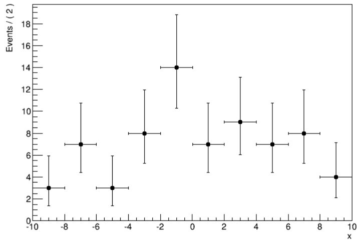
- How to define ‘extremity’ if observed data is a distribution

Counting

Histogram

Observation

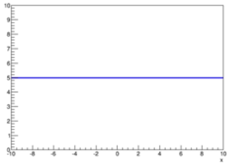
N_{obs}=7

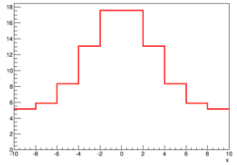


Median expected by hypothesis

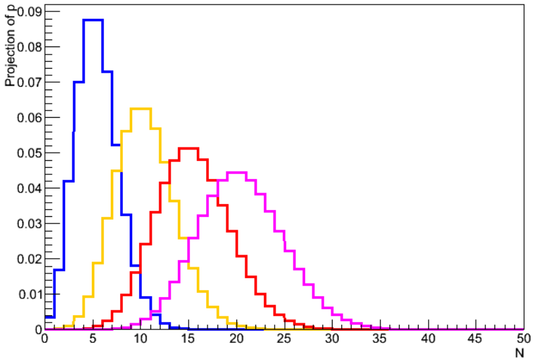
N_{exp}(s=0) = 5

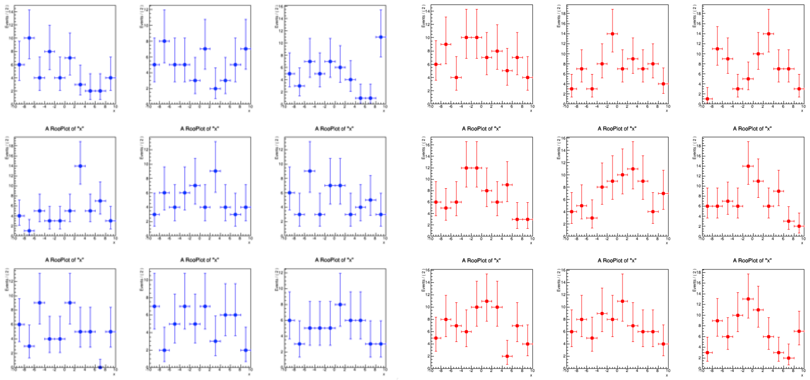
N_{exp}(s=5) = 10





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

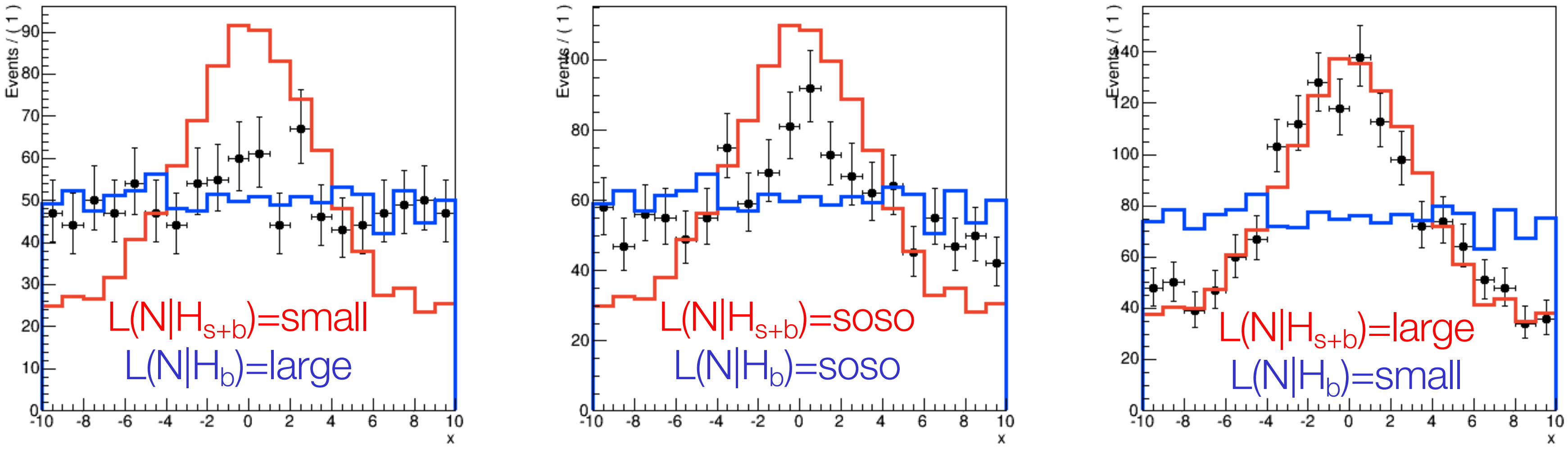
$$\lambda(\vec{N}) = \frac{\text{small}}{\text{large}} = \text{small}$$

→ 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

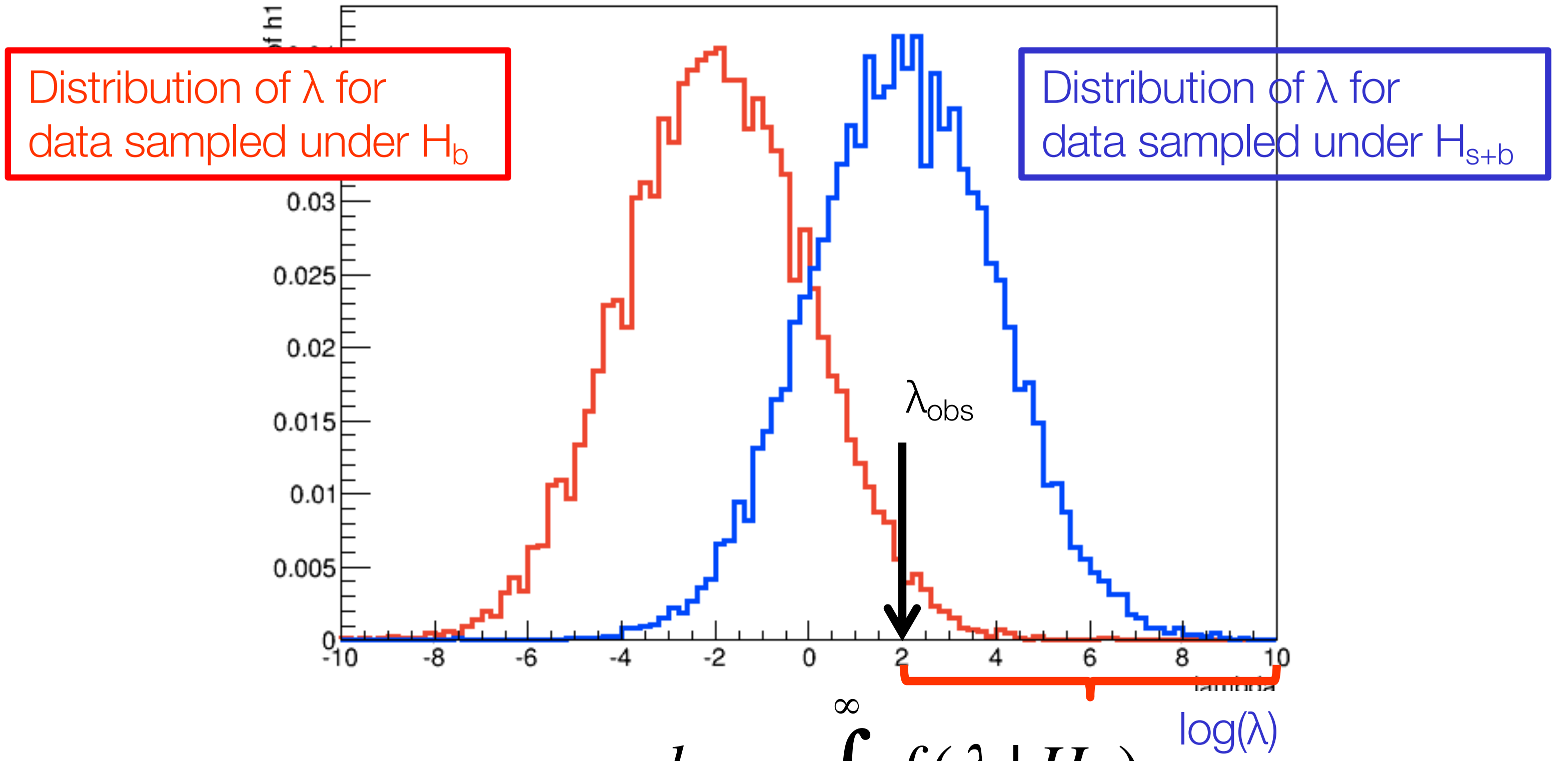


$\lambda(N)$=0.0005 $\lambda(N)$=0.47 $\lambda(N)$=5000

- **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 $\lambda(N)$ as large or larger than the observed test statistic $\lambda(N_{obs})$**

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



Distribution of λ for data sampled under H_b

Distribution of λ for data sampled under H_{s+b}

λ_{obs}

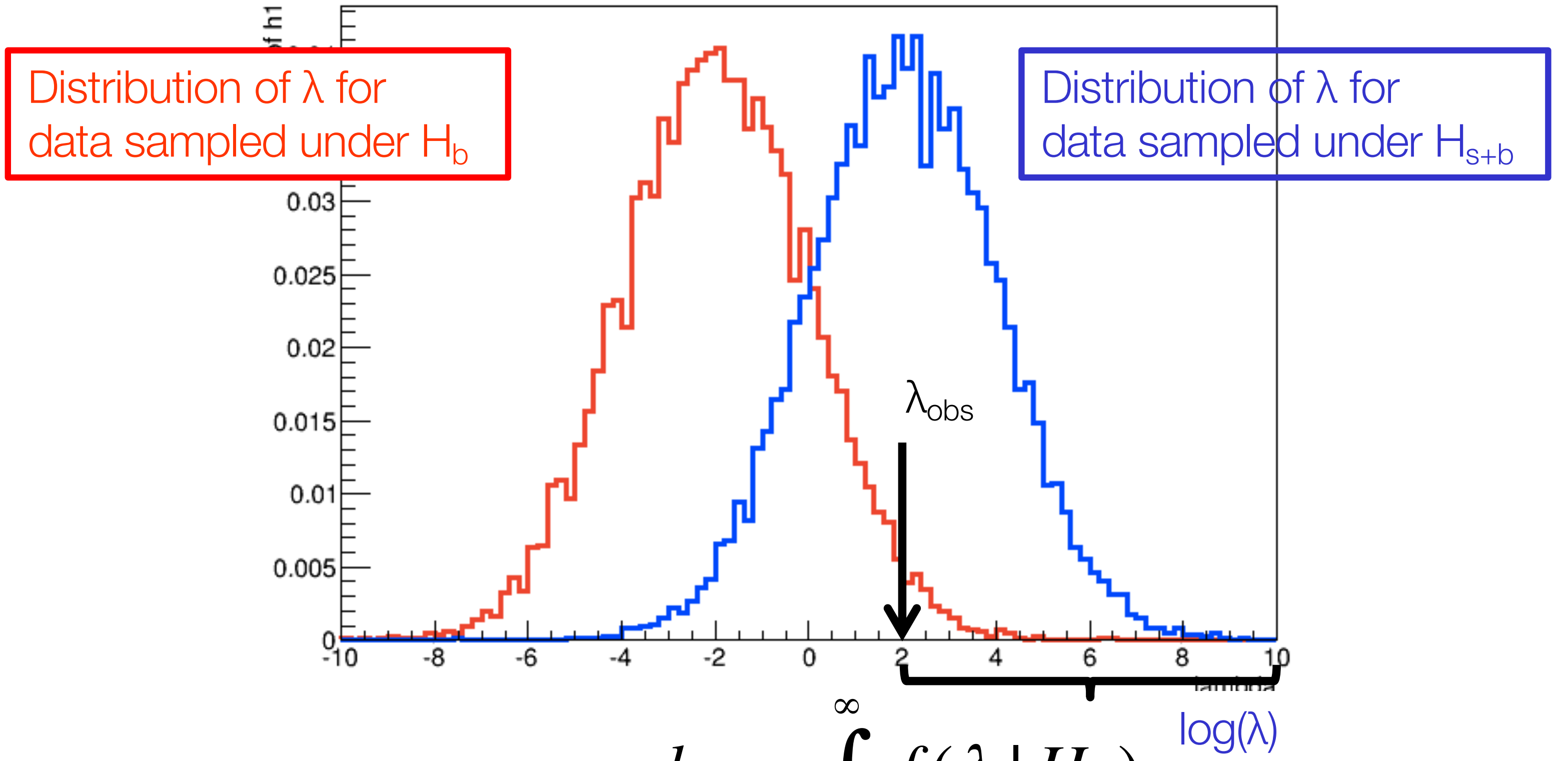
log(λ)

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

Wouter Verkerke, NIKHEF

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)



Distribution of λ for data sampled under H_b

Distribution of λ for data sampled under H_{s+b}

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

Wouter Verkerke, NIKHEF

Likelihoods for distributions - summary

- **Bayesian inference unchanged**

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

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

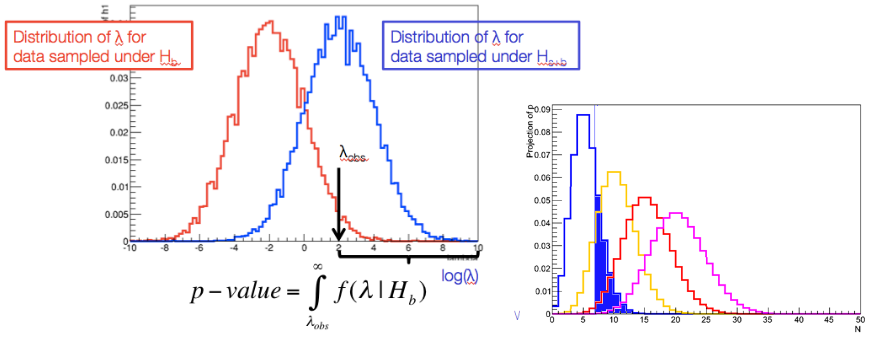
- **Frequentist inference procedure *modified***

  → 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%'

- **Definition: p-value**



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

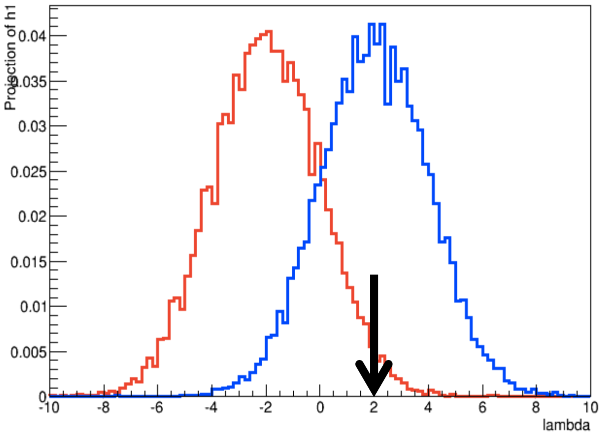
The likelihood principle

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

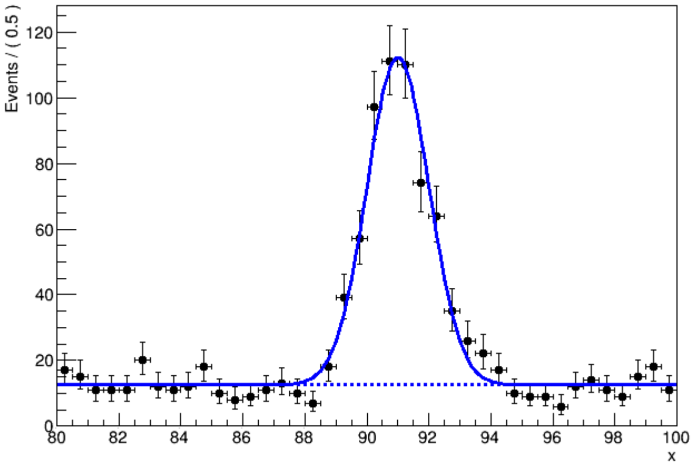
Likelihood Principle

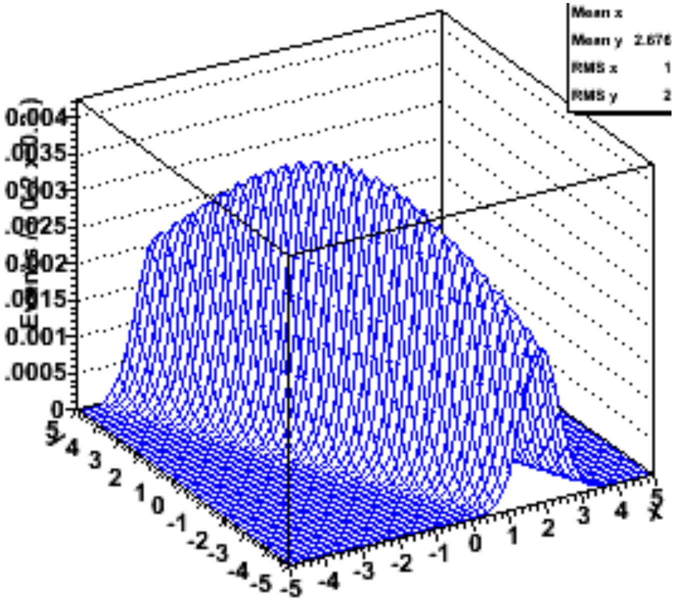
- In **Bayesian** methods and **likelihood-ratio** based methods, the probability (density) for obtaining the *data at hand is used (via the likelihood function), but probabilities for obtaining other data are not used!*
- In contrast, in typical **frequentist** calculations (e.g., a p-value which is 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.

Generalizing to multiple dimensions

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





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

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

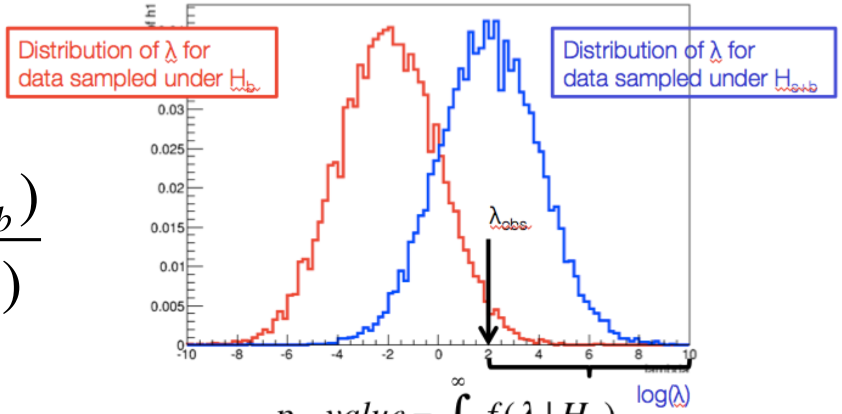
- Neither generalization (binned→continuous, one→multiple observables) has any further consequences for Bayesian or Frequentist inference procedures

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**

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



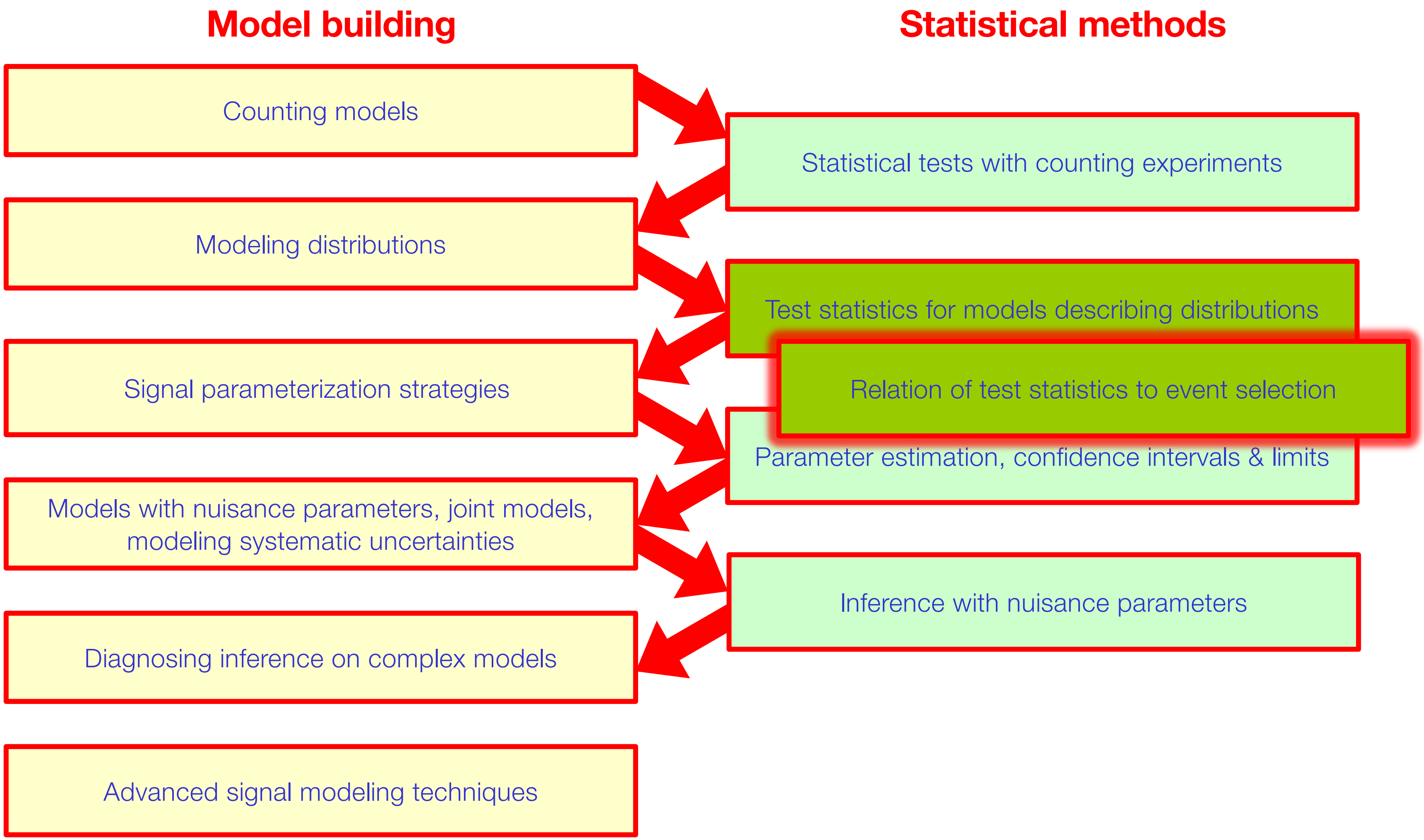
$$p - value = \int_{\lambda_{obs}}^{\infty} f(\lambda \mid H_b)$$

Wouter Verkerke

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

Roadmap of this course

- Start with basics, gradually build up to complexity

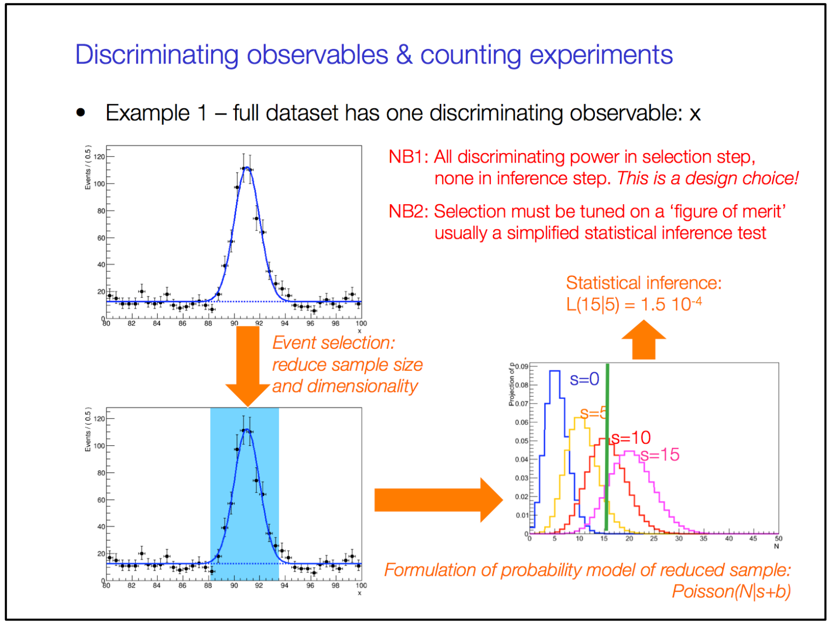


Model building
Statistical methods
Counting models
Statistical tests with counting experiments
Modeling distributions
Test statistics for models describing distributions
Relation of test statistics to event selection
Signal parameterization strategies
Parameter estimation, confidence intervals & limits
Models with nuisance parameters, joint models, modeling systematic uncertainties
Inference with nuisance parameters
Diagnosing inference on complex models
Advanced signal modeling techniques

Deciding on a split

- HEP data analysis often a 2-step process:

  first selection,
  then inference



Discriminating observables & counting experiments

- Example 1 – full dataset has one discriminating observable: x

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}

Event selection: reduce sample size and dimensionality

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

- 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

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

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

Event selection

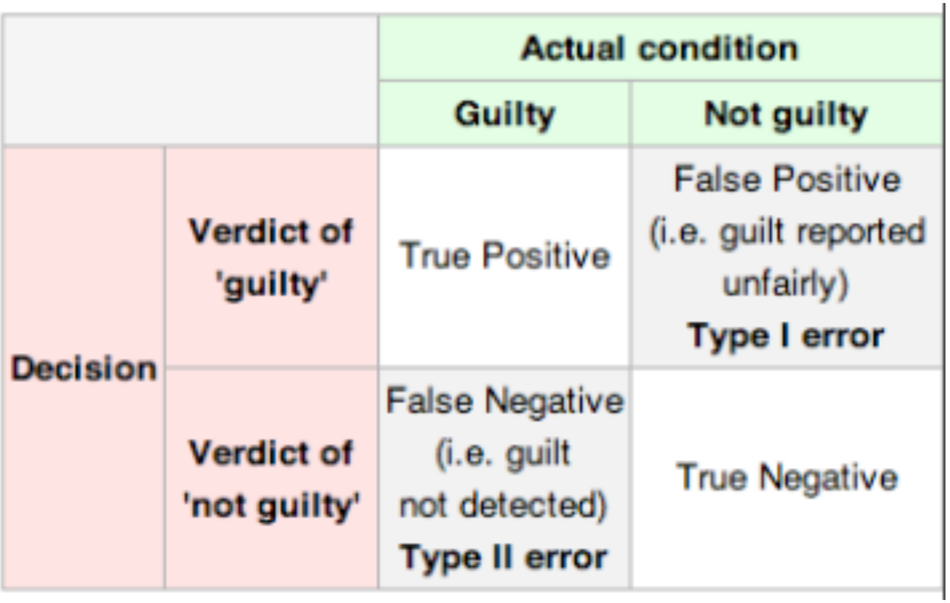
- The event selection problem:
  - Input: Two classes of events “signal” and “background”
  - Output: Two categories of events “selected” and “rejected”
- Goal: select as many signal events as possible,
  reject as many background events as possible

- 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

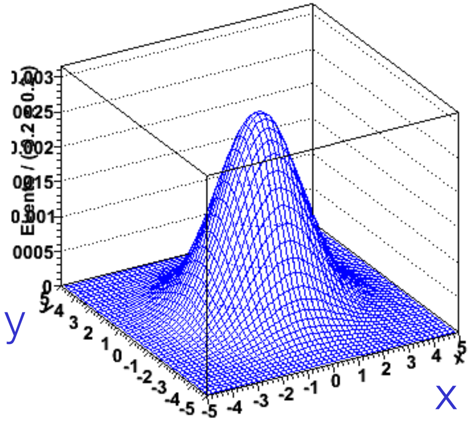
$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

- Any other region of the same size will have less power 'ex03.C'

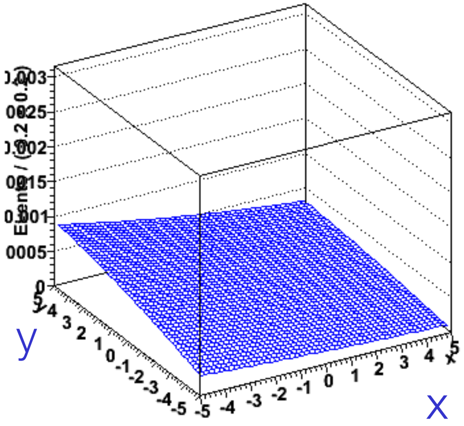
The Neyman-Pearson lemma

- Example of application of NP-lemma with two observables

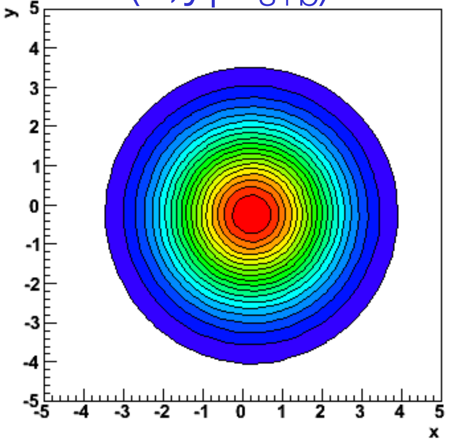
$f(x,y|H_s)$



$f(x,y|H_b)$



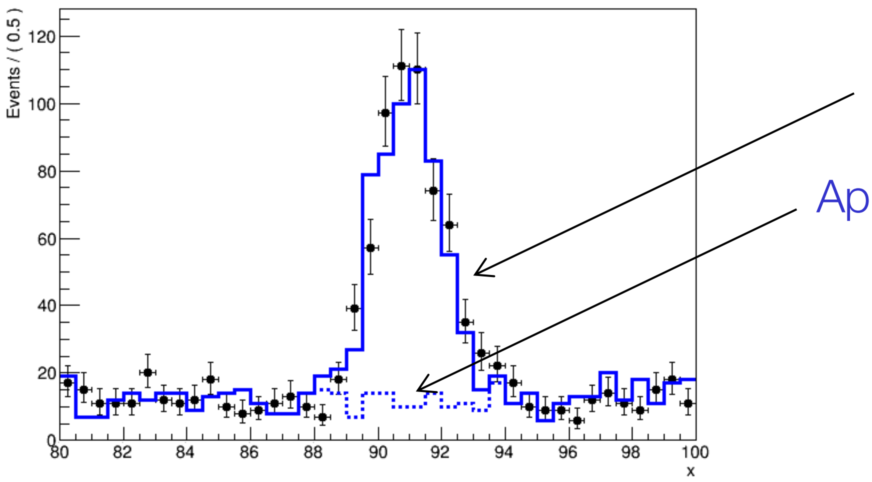
$$\frac{f(x,y|H_s)}{f(x,y|H_{s+b})} > c$$



- 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



Approximation of true f(x|s)

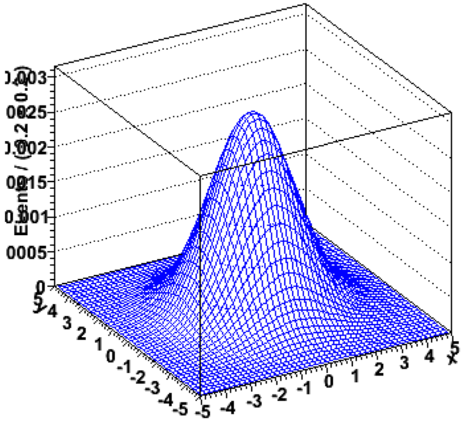
Approximation of true f(x|b)

Hypothesis testing with a large number of observables

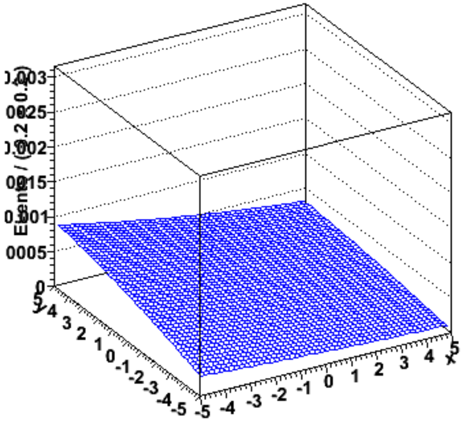
- When number of observables is large follow different strategy
- Instead of aiming at approximating p.d.f.s f(x|s) and f(x|b) aim to approximate decision boundary with an empirical parametric form

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

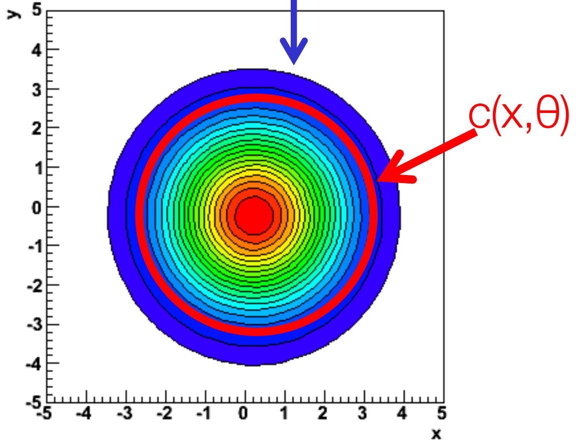
f(x,y|H$_s$)



f(x,y|H$_b$)



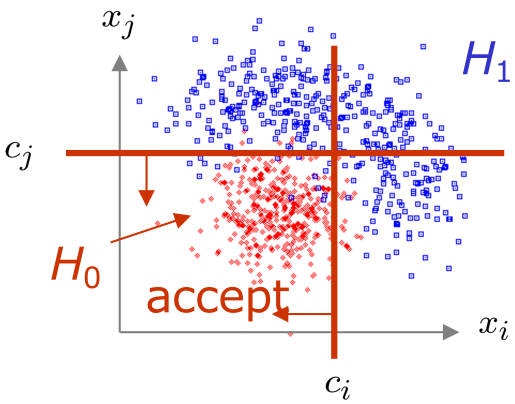
$\frac{f(x,y|H_s)}{f(x,y|H_{s+b})} > c$



Empirical parametric forms of decision boundaries

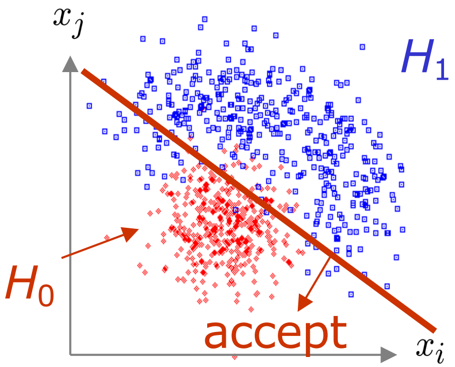
- Can in principle choose any type of Ansatz parametric shape

Rectangular cut



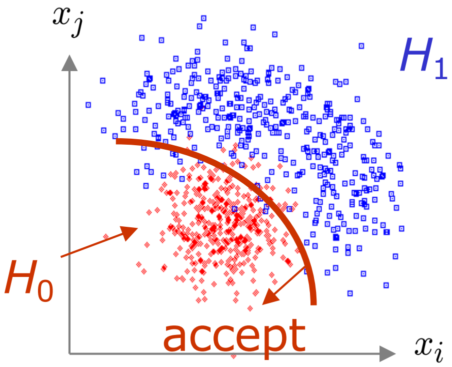
$$t(x) = \theta(x_j - c_j)\theta(x_i - c_i)$$

Linear cut



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

Non-linear cut

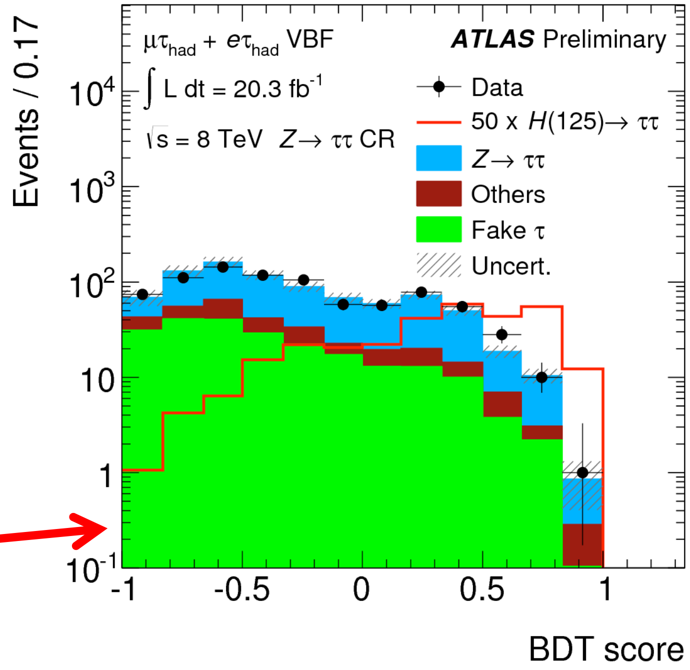


$$t(x) = \vec{a} \cdot \vec{x} + \vec{x} A \vec{x} + \ldots$$

- 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)

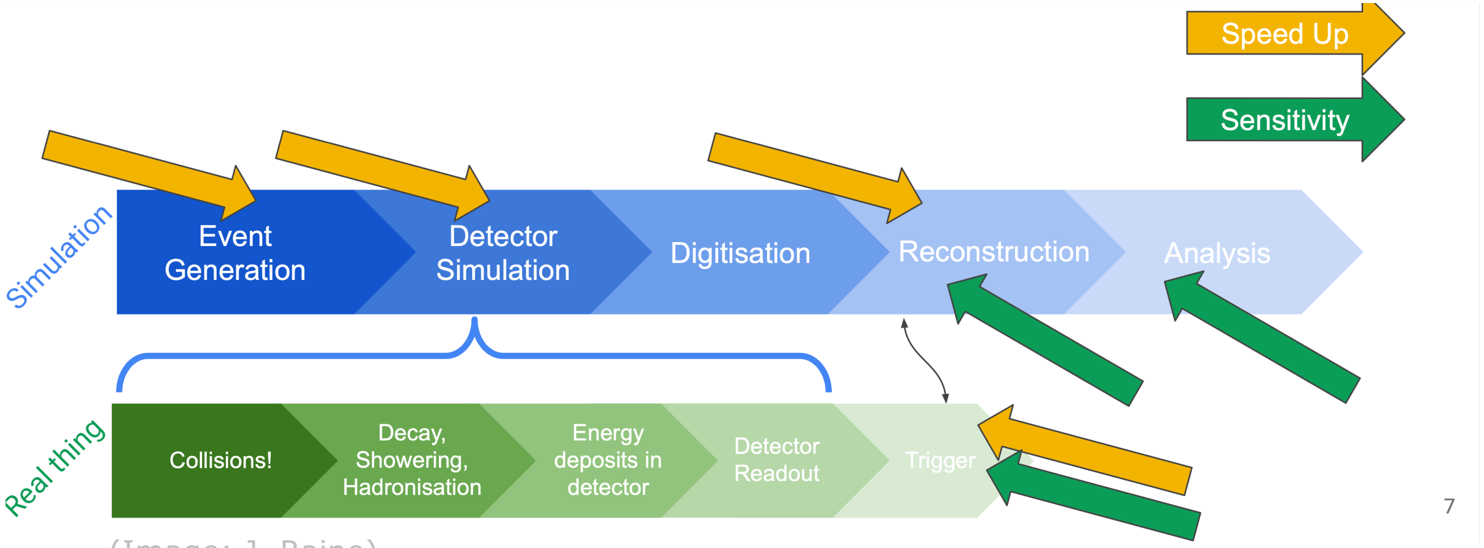
Classification with Machine Learning

Machine Learning

- What is Machine Learning?
  - Giving computers the ability to learn without explicitly programming them (Arthur Samuel, 1959)
  - Mathematical models learnt from data that characterize the patterns, regularities, and relationships amongst variables in the system
- Huge variety of choices in goals, formulations, training procedures
- Mathematical structure of model: (deep) neural networks, convolutional networks, transformer models, (boosted) decision trees, etc etc
- Input data: supervised learning (learn from simulation with truth-labels), unsupervised learning (learn from data without truth labels)
- Learning goal: classification, regression
- Scope of model: discrimination or generative

Machine Learning in HEP

- ML used in HEP in many places any in many ways

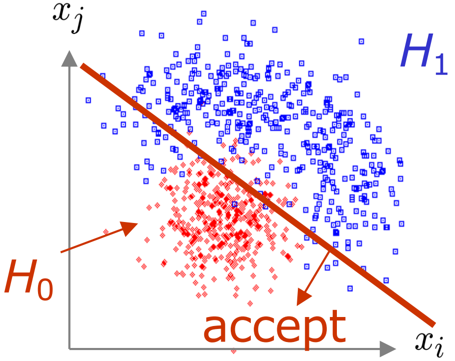


(Image: J. Raine)

- Dominant use case: Supervised learning for classification
  - Signal/background separation, object tagging – trained on simulation samples
  - Will largely focus on *this* use case today
- But also many other uses
  - Unsupervised learning: anomaly detection
  - Regression: improving mass estimate of e.g. jets in events

Multivariate Discriminants – the simplest case: the linear discriminant

- A linear discriminant constructs t(x) from a linear combination of the variables x_i



$$t(\vec{x}) = \sum_{i=1}^{N} a_i x_i = \vec{a} \cdot \vec{x}$$

  - Optimize discriminant by chosing a_i to maximize separation between signal and background

- Most common form of the linear discriminant is the Fisher discriminant

$$F(\vec{x}) = \overbrace{(\vec{\mu}_S - \vec{\mu}_B)^T V^{-1}}^{\vec{a}} \vec{x}$$

Mean values in x_i for sig,bkg

Inverse of variance matrix of signal/background (assumed to be the same)

R.A. Fisher
Ann. Eugen. 7(1936) 179.

Ansatz test statistics – The Fisher discriminant

$$F(\vec{x}) = \overbrace{(\vec{\mu}_S - \vec{\mu}_B)^T V^{-1}}^{\vec{a}}\,\vec{x}$$

R.A. Fisher
Ann. Eugen. 7(1936) 179.

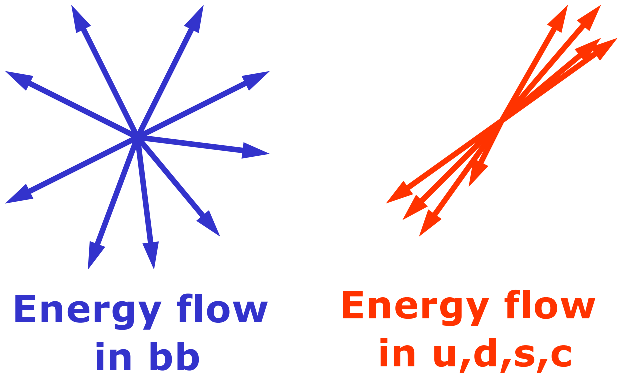
Mean values in x_i for sig,bkg

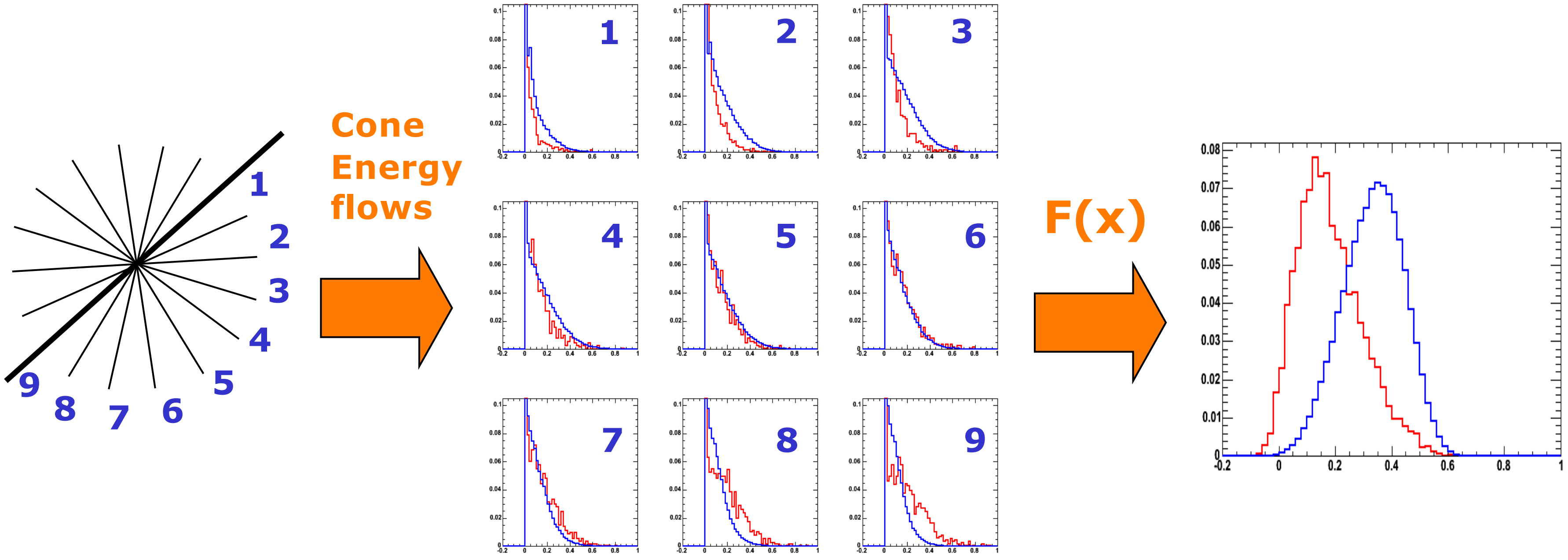
Inverse of variance matrix of signal/background (assumed to be the same)

- Advantage of Fisher Discriminant:
  - Ingredients μ_s,μ_b,V can all be calculated directly from data or simulation samples. No ‘training’ or ‘tuning’
- Disadvantages of Fisher Discriminant
  - Fisher discriminant only exploits difference in means.
  - If signal and background have different variance, this information is not used.

Example of Fisher discriminant

- The “CLEO” Fisher discriminant
  - Goal: distinguish between e+e- → Y4s → $b\bar{b}$ and $u\bar{u}$,$d\bar{d}$,$s\bar{s}$,$c\bar{c}$
  - Method: Measure energy flow in 9 concentric cones around direction of B candidate





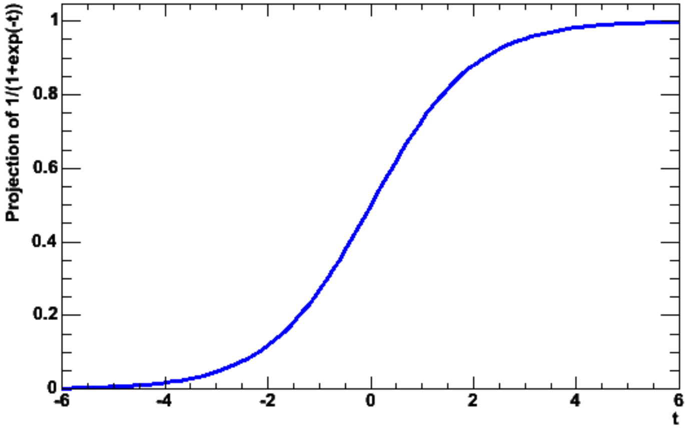
When is Fisher discriminant is the optimal discriminant?

- A very simple dataset

$$\left.\begin{aligned} S &= \prod_i Gauss(x_i;\mu_i^S,\sigma_i) \\ B &= \prod_i Gauss(x_i;\mu_i^B,\sigma_i) \end{aligned}\right\}$$

Multivariate Gaussian distributions with **different means** but **same width** for signal and background

- Fisher is optimal discriminant for this case
  - In this case we can also directly correlate F(x) to **absolute signal probability**



$$P(F) = \frac{1}{1+e^{-F}}$$

'Logistic sigmoid function'

Multivariate data selection – Neural networks

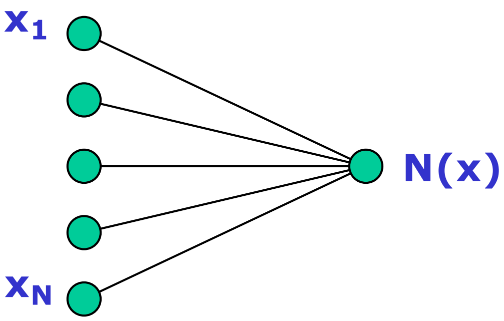
- Neural networks are used in neurobiology, pattern recognition, financial forecasting (and also HEP)

$$N(\vec{x}) = s\left(a_0 + \sum_i a_i x_i \right)$$

s(t) is the *activation function*, usually a logistic sigmoid

$$s(t) = \frac{1}{1+e^{-t}}$$

- This formula corresponds to the 'single layer perceptron'
  - Visualization of single layer network topology

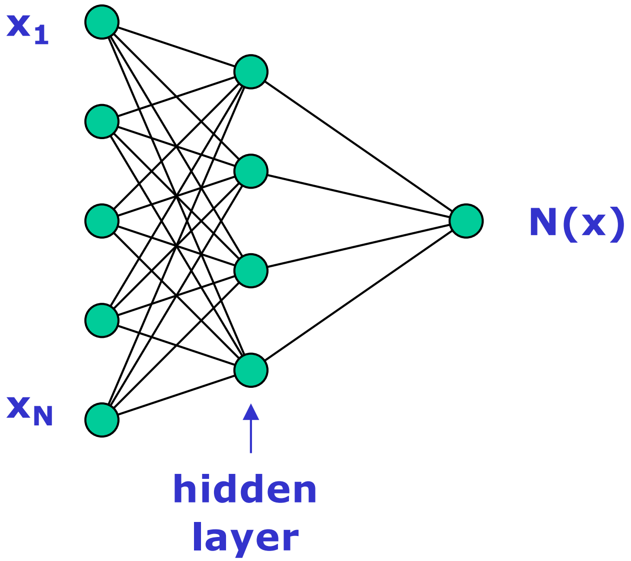


Since activation function s(t) is monotonic,
the single layer N(x) is equivalent to the Fisher discriminant F(x)

Wouter Verkerke, UCSB

Neural networks – general structure

- The single layer model and easily be generalized to a *multilayer* perceptron



$$N(\vec{x}) = s(a_0 + \sum_{i=1,}^{m} a_i h_i(\vec{x}))$$

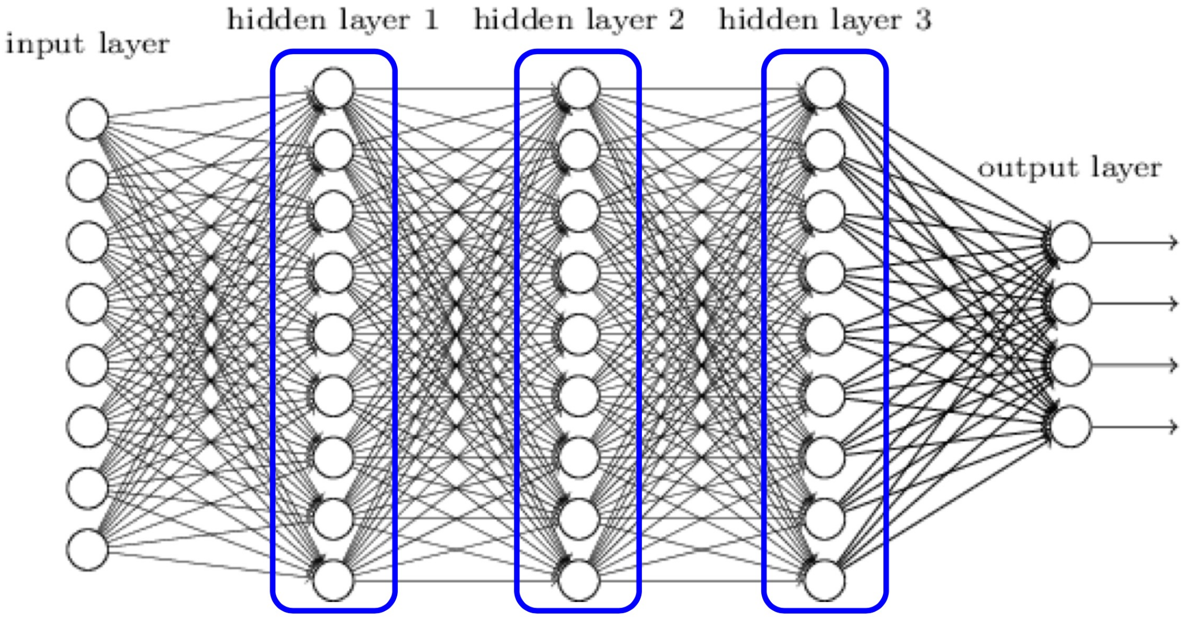
$$\textit{with } h_i(\vec{x}) = s(w_{i0} + \sum_{j=1}^{n} w_{ij} x_j)$$

with a_i and w_{ij} weights (connection strengths)

- Easy to generalize to arbitrary number of layers
- Feed-forward net: values of a node depend only on earlier layers (usually only on preceding layer) ‘the network architecture’
- More nodes bring N(x) close to optimal t(x)=S(x)/B(x) but with much more parameters to be determined

Deep Neural Networks

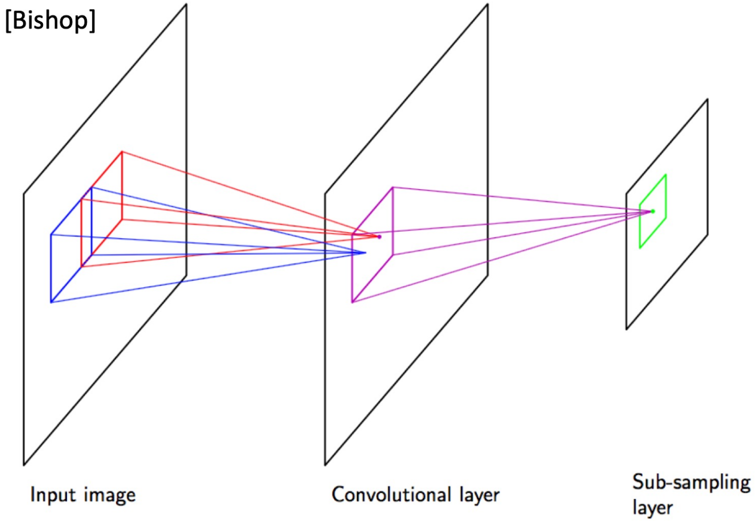
- Availability of much more computing power, and notably GPUs have, and intense efforts worldwide outside HEP on algorithm and archivecture development have increased ambitions (and results) of ML by many orders of magnitude



- Generally networks are labeled ‘deep’ if they reach a level of complexity where specialized nodes inside the network serve to extract specific ’features’ of the data

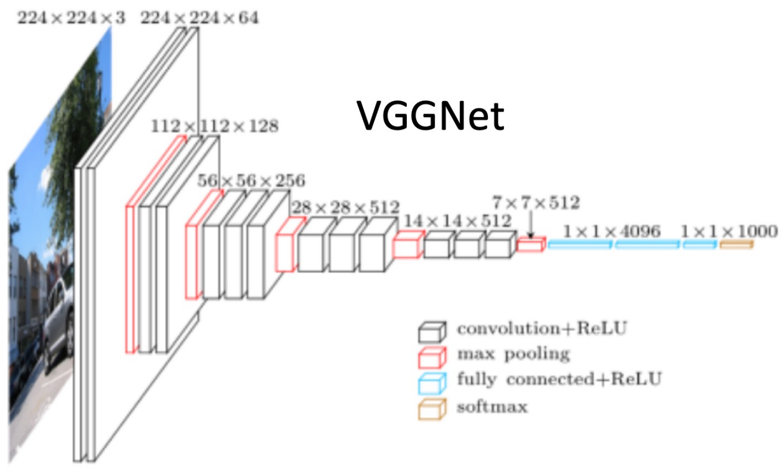
Example deep NN - convolutional NLL

- Convolutions NNs primarily designed to process ‘image data’
  - Scan for features defined by convolutions of local image data with a specific kernel function. Network designed to be insensitive to spatial location of feature
  - Structure allows for capturing local structure in early convolutions, and long range structure in later stage convolutions and in fully connected layers

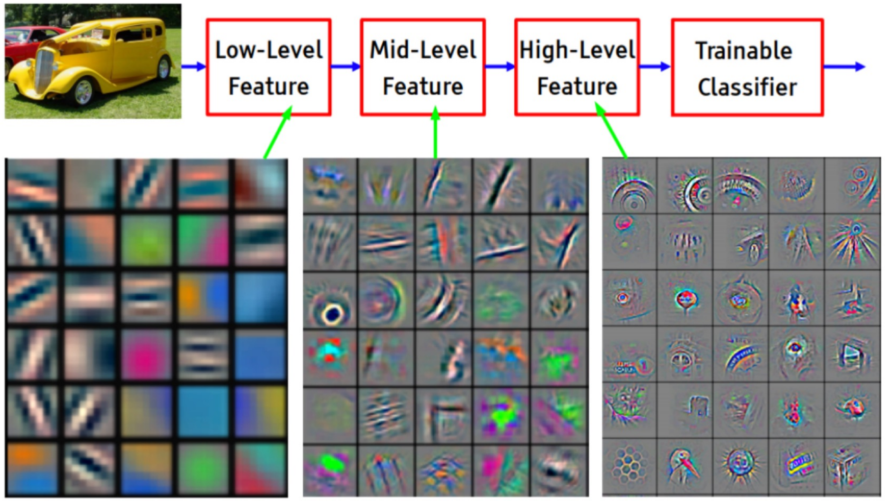


Example deep NN - convolutional NLL

- Example here on generic image data, but many HEP problems are similar to image data (e.g. jet particles projected on a calorimeter surface)



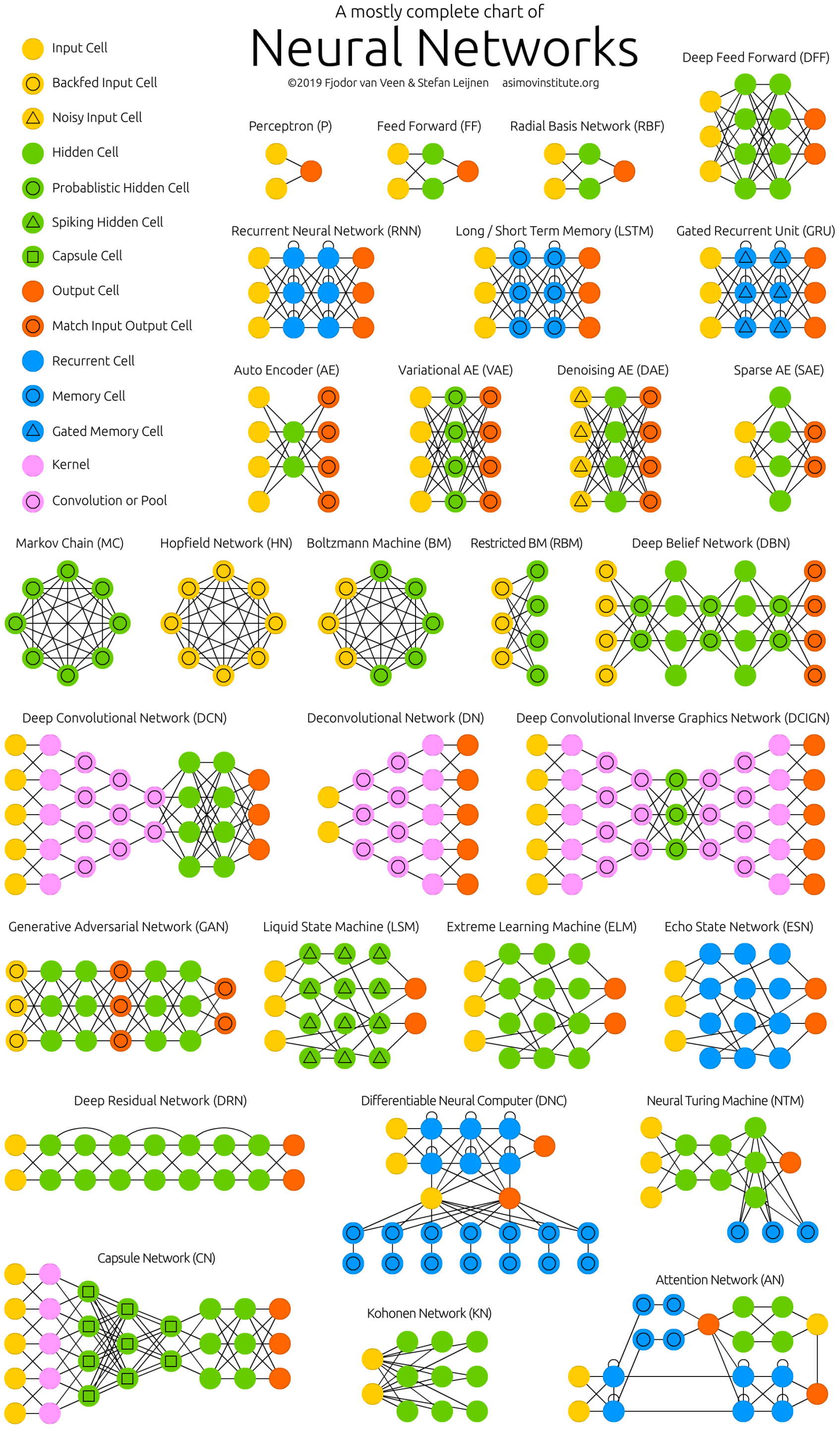
(Simonyan and Zisserman, 2014)



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Many network structures possible for many purposes

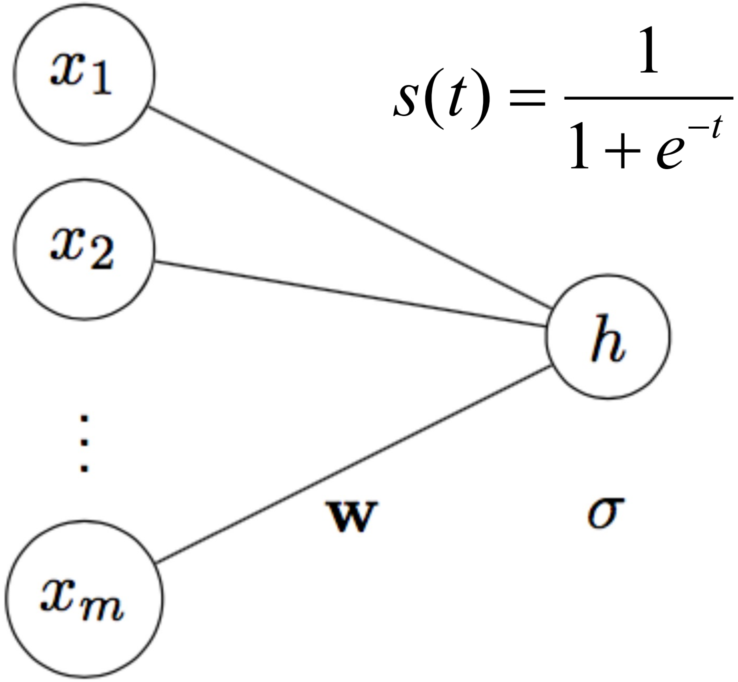
- For more information see specialized lectures in e.g. IN2P3 school of statistics, Terascale school of statistics, CERN academic lectures by M. Kagan…
- Generally, continuous rapid developments in ML/AI community



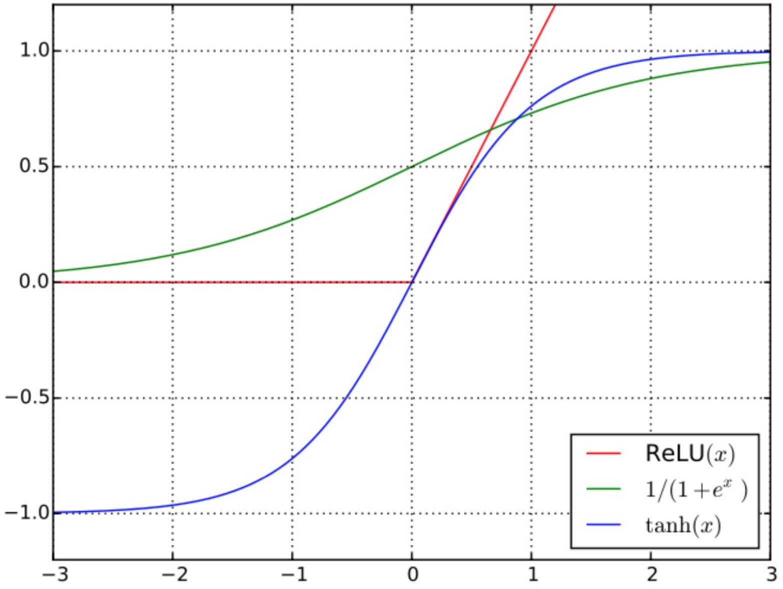
A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

Activation functions

- So far only discussion logistic sigmoid as activation function



$$s(t) = \frac{1}{1+e^{-t}}$$

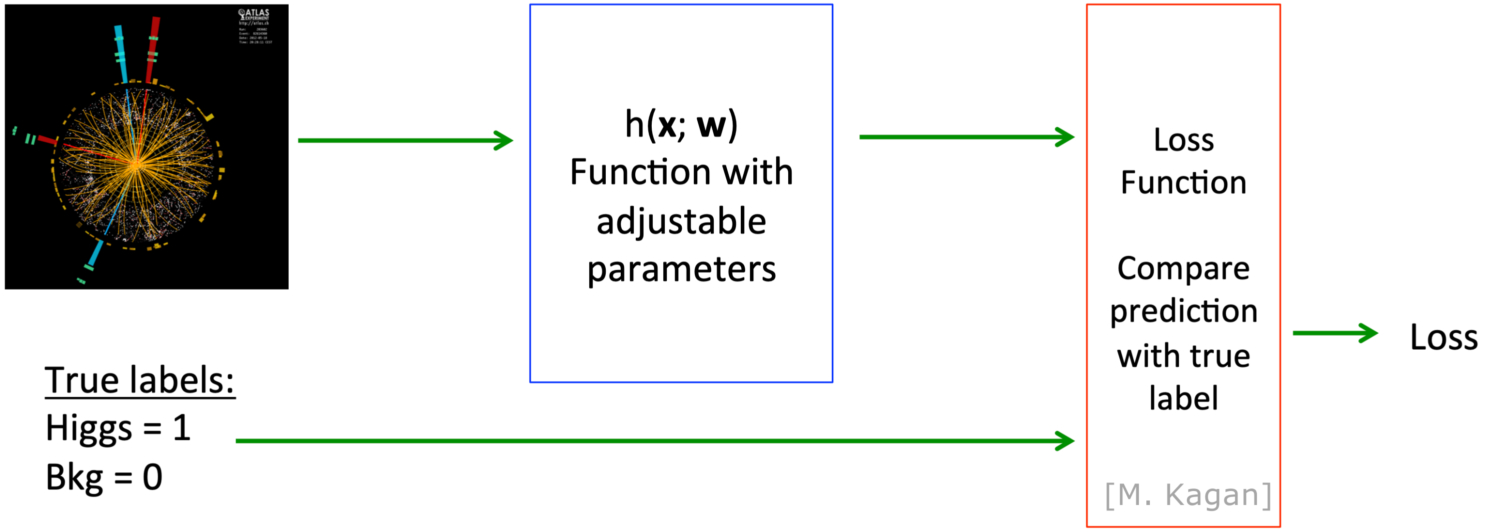


- Other functions are generally possible and useful. In particular the Rectified Linear Unit (ReLU) activiation function improves Deep NN training as its derivative is constant (rather than vanishing) is therefore often used

Training (Deep) Neural Networks

- Training Deep Networks computationally very challenging
- Why did this take off in the last decade (or so)?
- Big data → large training sets
- Wide availability of cheap GPUs has increased computational power by orders of magnitude
- Many assorted improvedment in many areas: Improved optimization algorithms, new regulatization techniques, new activations functions

Supervised learning – general procedure



h(**x**; **w**)
Function with adjustable parameters

Loss Function

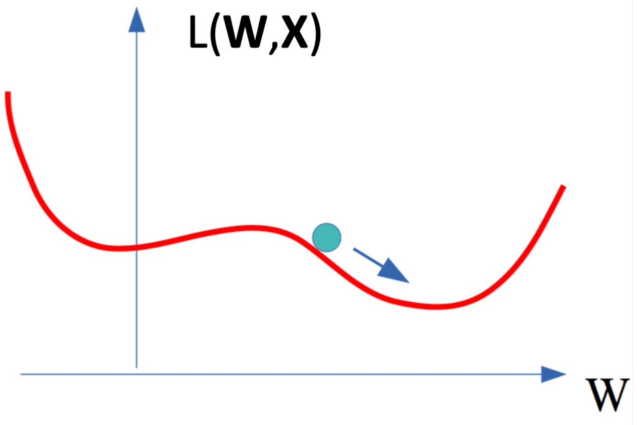
Compare prediction with true label

Loss

True labels:
Higgs = 1
Bkg = 0

[M. Kagan]

1. Design a (D)NN with adjustable parameters
2. Design a Loss function
3. Find best parameters which minimize loss



Supervised learning – Loss function

- General form

$$\arg\min_{\mathbf{w}} \underbrace{\frac{1}{N}\sum_{i=1}^{N} L(\underline{h(\mathbf{x}_i;\mathbf{w}), y_i})}_{\boxed{\text{Average expected loss}}} + \underbrace{\lambda\Omega(\mathbf{w})}_{\boxed{\text{Model regularization}}}$$

compares MVA prediction h() with (training) target data y

Additional term Ω(w) penalizes certain values of parameters w with the aim to regularize the minimization process.

- Specific choice of loss function depends among others on training goal

Square error loss

$$\boxed{L(h(\mathbf{x};\mathbf{w}), y) = \big(h(\mathbf{x};\mathbf{w}) - y\big)^2}$$

(often used in regression)

Cross-entropy loss

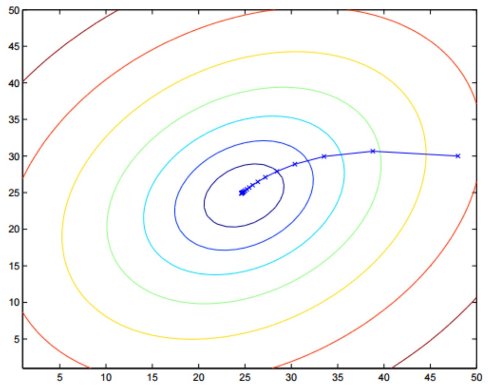
$$\boxed{\begin{aligned} L(h(\mathbf{x};\mathbf{w}), y) = &- y\log h(\mathbf{x};\mathbf{w}) \\ &- (1-y)\log(1-h(\mathbf{x};\mathbf{w})) \end{aligned}}$$

(often used in classification)

Supervised learning – minimization of loss function

- Minimize loss function using back-propagation
  - Compute gradient on each training set or batch

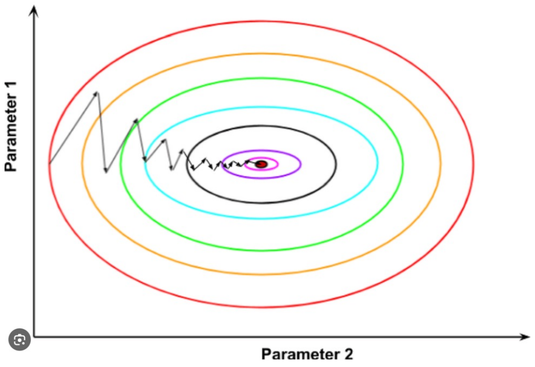
$$\boxed{\nabla_{w_j} L = \frac{\partial L}{\partial f}\frac{\partial f}{\partial g_n}\frac{\partial g_n}{\partial g_{n-1}}\cdots\frac{\partial g_{k+1}}{\partial g_k}\frac{\partial g_k}{\partial w_j}}$$



- Update weights with gradient descent
  - where α is the learning rate

$$\boxed{w_j \leftarrow w_j - \alpha \nabla_{w_j} L}$$

- Often advantageous to compute gradient only on subset of data (mini batch) 'Mini Batch Gradient Descent'
  - Less computation required
  - Noise in gradient descent helps to avoid local minima



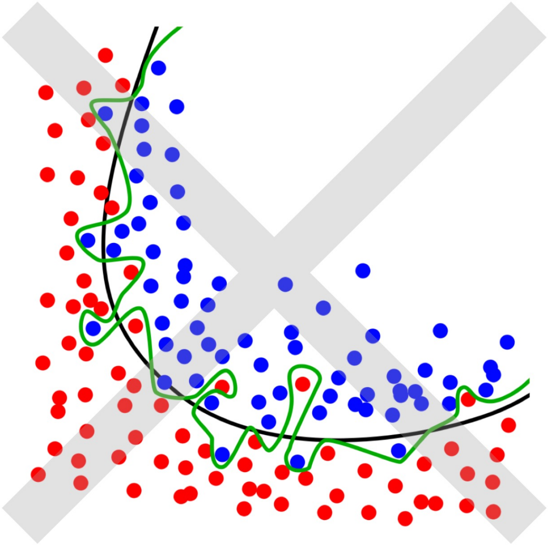
Supervised learning – Loss function

- General form

$$\arg\min_{\mathbf{w}} \underbrace{\frac{1}{N}\sum_{i=1}^{N} L(\underline{h(\mathbf{x}_i;\mathbf{w}), y_i})}_{\text{Average expected loss}} + \underbrace{\lambda\Omega(\mathbf{w})}_{\text{Model regularization}}$$

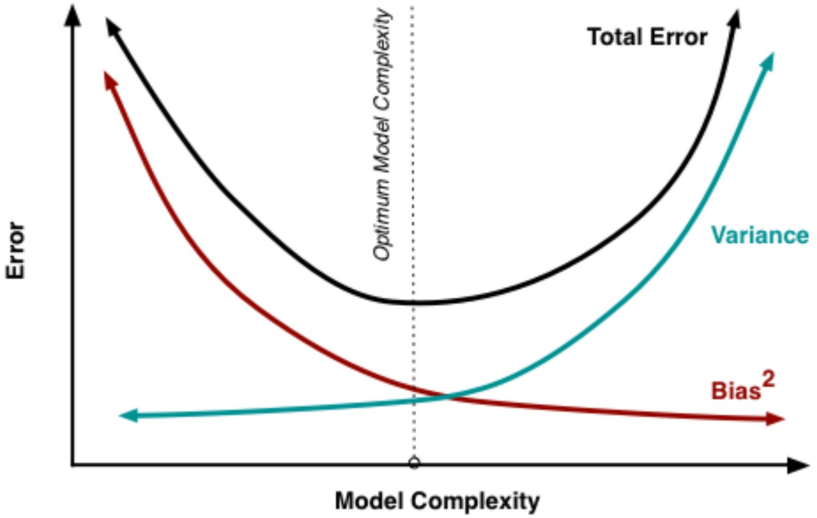
- Add regularization term to penalize overly complex models

$$\boxed{L' = L + \frac{1}{2}\sum_j w_j^2}$$



Some notes on model complexity

- Simple models under-fit: will deviate from data (high bias) but will not be influenced by peculiarities of data (low variance).
- Complex models over-fit: will not deviate systematically from data (low bias) but will be very sensitive to data (high variance).



Regularization term can help to control model complexity

$$\begin{array}{ccccc} \text{generalization error} = & \text{systematic error} & + & \text{sensitivity of prediction} \\ & \text{(bias)} & & \text{(variance)} \end{array}$$

$$\begin{aligned} E[(y-h(x))^2] &= E[(y-\bar{y})^2] \quad + \quad (\bar{y}-\bar{h}(x))^2 \quad + \quad E[(h(x)-\bar{h}(x))^2] \\ &= \text{noise} \quad + \quad (\text{bias})^2 \quad + \quad \text{variance} \end{aligned}$$

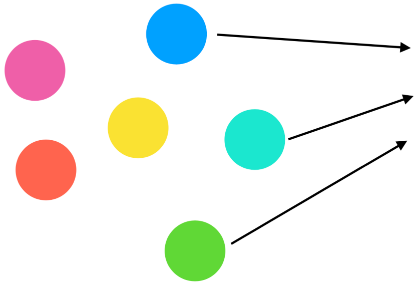
Modeling unordered data

- So far (implicitly) assumed input data is *structured, i.e.* similar to a fixed-length vector where each element has same meaning for every event.
  - *Maps well to classical ML approach in HEP* → inputs are precalculated quantities, often using some physics input (invariant masses, highest-pT of lepton in event etc etc)
- But with ever increasing power and success of automatic feature extraction by deep networks on HEP data, question arises, why not simply give *all* event information to the DNN?
  - Apart from scale of problem, presents a small logistical challenge in the data format: full event reconstructed event record of events is not 'structured' in the sense above: 4-vectors are not ordered in a particular way, nor is the data set fixed size.
  - Other network structures can help here, notable Graph Neural Networks are very suitable for this type of data

Graph Neural Networks

- Modeling of *ordered* data as matrix or vector

set of inputs with N constituents, M features
{..., (pT, η, ϕ, particle ID), ...}



jet constituents

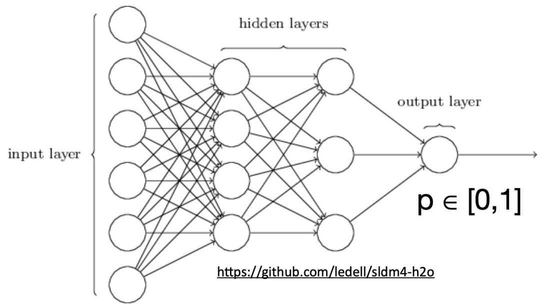
feature matrix (N, M)

pT (GeV)	η	ϕ	particle ID
12.3	*1.2*	*0.5*	*pi+*
11.8	*1.24*	*0.45*	*K0*
10.4	*1.18*	*0.43*	*pi-*
9.8	*1.39*	*...*	*e-*
6.4	*...*	*...*	*...*
5.3	*...*	*...*	

flat NxM feature vector

particle 1, pT
particle 2, pT
particle 3, pT
particle 1, η
particle 2, η
...
...

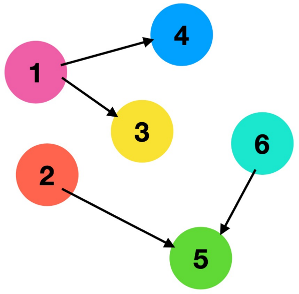
map feature vector to an output



https://github.com/ledell/sldm4-h2o

- Modeling of graph data

graph = set of nodes/vertices/elements + edges between them



Edges represented as a index pairs
edges = [(1,4), (1,3), (2,5), (6,5)]

Or as an NxN adjacency matrix

$$A_{ij} = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

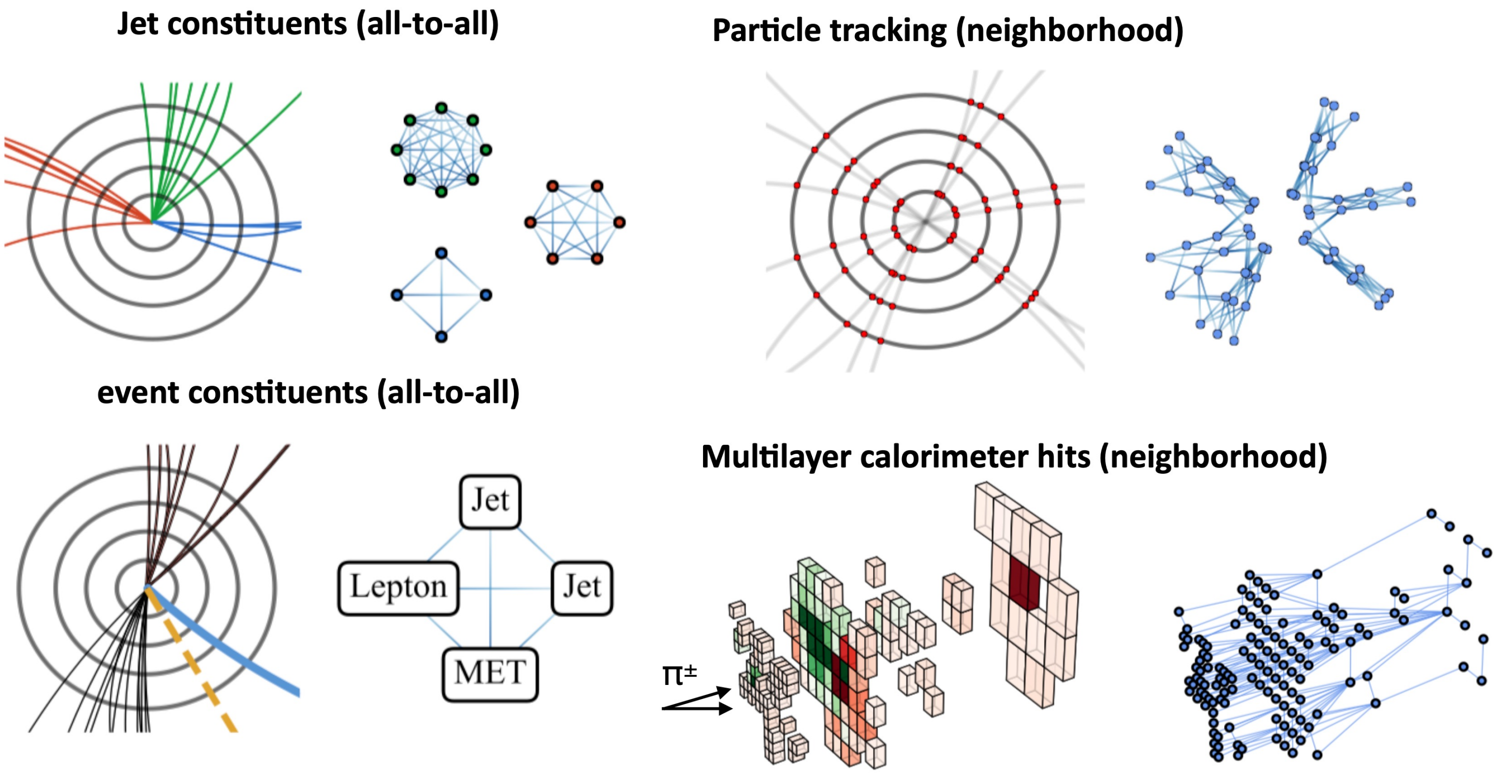
Where do we get this graph structure?
1. All-to-all connections, in case of small input sets.
2. From physics priors: connect "nearby" elements in advance
3. Optimize as a part of the learning process (Graph Structure Learning)

[Images: J. Pata]

Wouter Verkerke, NIKHEF

Graph Neural Networks

- Graph structures common in HEP data



Jet constituents (all-to-all)

Particle tracking (neighborhood)

event constituents (all-to-all)

Multilayer calorimeter hits (neighborhood)

Graph Neural Networks in Particle Physics, Jonathan Shlomi, Peter Battaglia, Jean-Roch Vlimant, 2007.13681, 10.1088/2632-2153/abbf9a

→ GNNs very succesful in flavor tagging for LHC experiments

[Image: J. Pata]

ML Classification in HEP – some summarizing thoughts

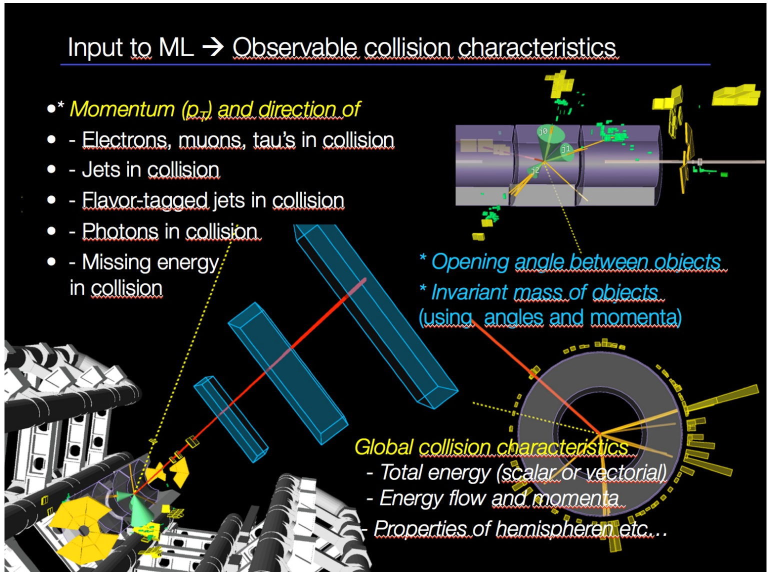
- Tremendous progress over past decade in better, more powerful techniques to use ML for classification, regression.
  - Only discussed in the briefest possible terms at the conceptual level here
  - Many practical online courses available for many of these tools
  - Tendency to move away from letting ML deal with 'pre-cooked' physics observables, to letting ML analyse complete event records
- Many other aspects of ML not discussed at all here (unsupervised learning, generative models (not discussed), decision trees)
  - Also with many use cases other than sig/bkg classification (regression, fast simulation)
- It is quite likely that in a few years even newer ML techniques in will replace the current best performing ones..

ML Classification in HEP – some summarizing thoughts

- But keep in mind that ML/AI techniques are *not* magic
  → their performance is also bound by the Neyman-Pearson limit
  - There is a well-defined upper bound on the reachable performance by any algorithm. This bound is not calculable for many complex models, but it is nevertheless there
- ML/AI techniques take their inputs quite literally → simulation samples used in supervised training are known to be subject to uncertainties.
  - If simulation differs from data, ML-based results may be suboptimal, or even wrong, depending on how ML was used
  - Impact of ML training on imperfect simuation depends on analysis design → more on this later
  - There is also room to take some uncertainty on input samples into account, sometimes this is trivial (e.g. if the total cross-section is uncertain)
  - But in other cases this is exceeding difficult if the specification of what is uncertain is fuzzy or incomplete. (e.g. hadronization uncertainties that are only expressed as different outcomes for two different generators)
- Bottom line – if systematic effects are non-negligible, great care must be taken in the use of ML discriminants in the analysis

Event selection as dimensionality reduction

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



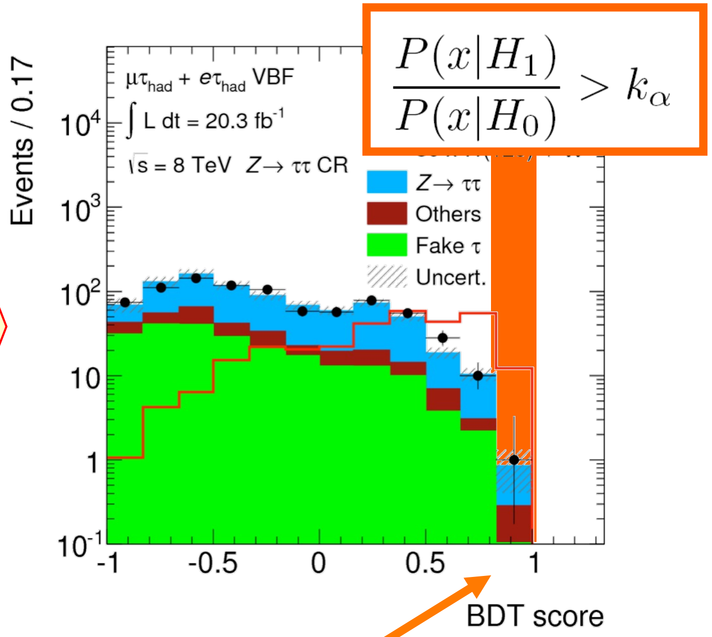
Input to ML → Observable collision characteristics

- * Momentum (p_T) and direction of
- - Electrons, muons, tau's in collision
- - Jets in collision
- - Flavor-tagged jets in collision
- - Photons in collision
- - Missing energy in collision

* Opening angle between objects

* Invariant mass of objects (using angles and momenta)

Global collision characteristics
- Total energy (scalar or vectorial)
- Energy flow and momenta
- Properties of hemispheren etc…

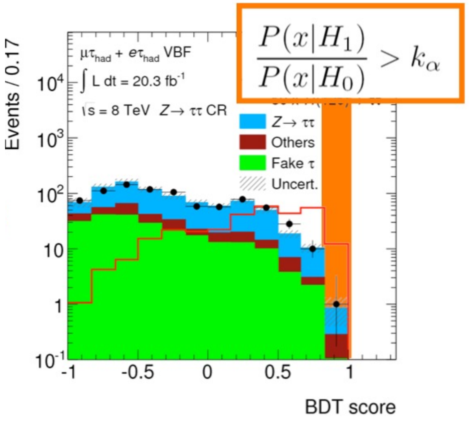


$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

- 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

The simplest analysis design: *cut-and-count*

- 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

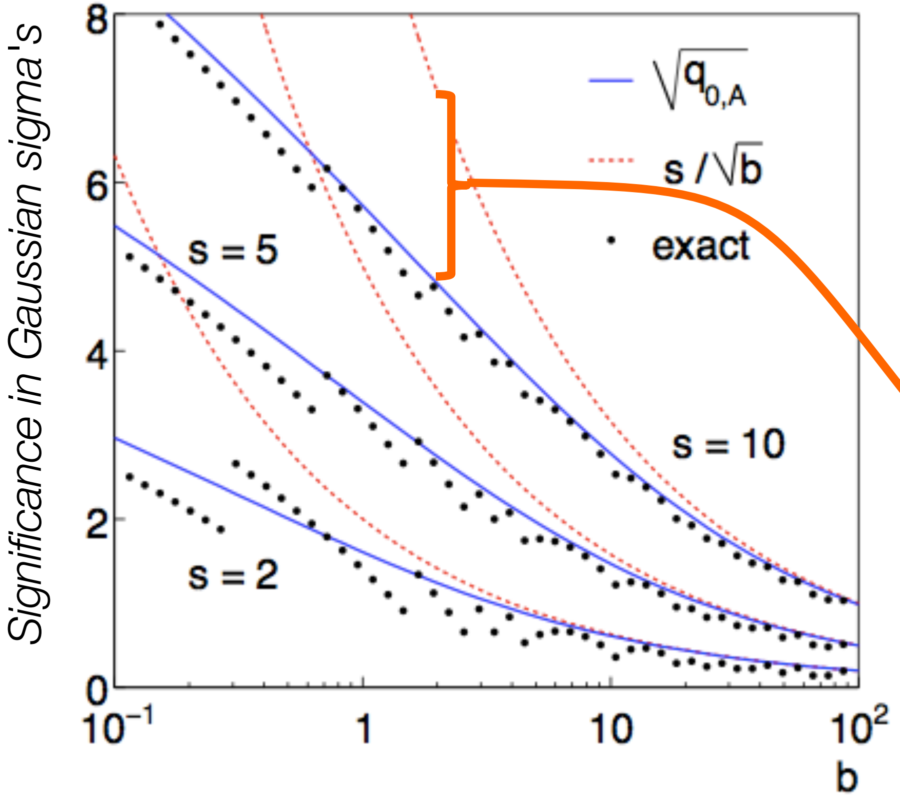


$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

- 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*



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

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

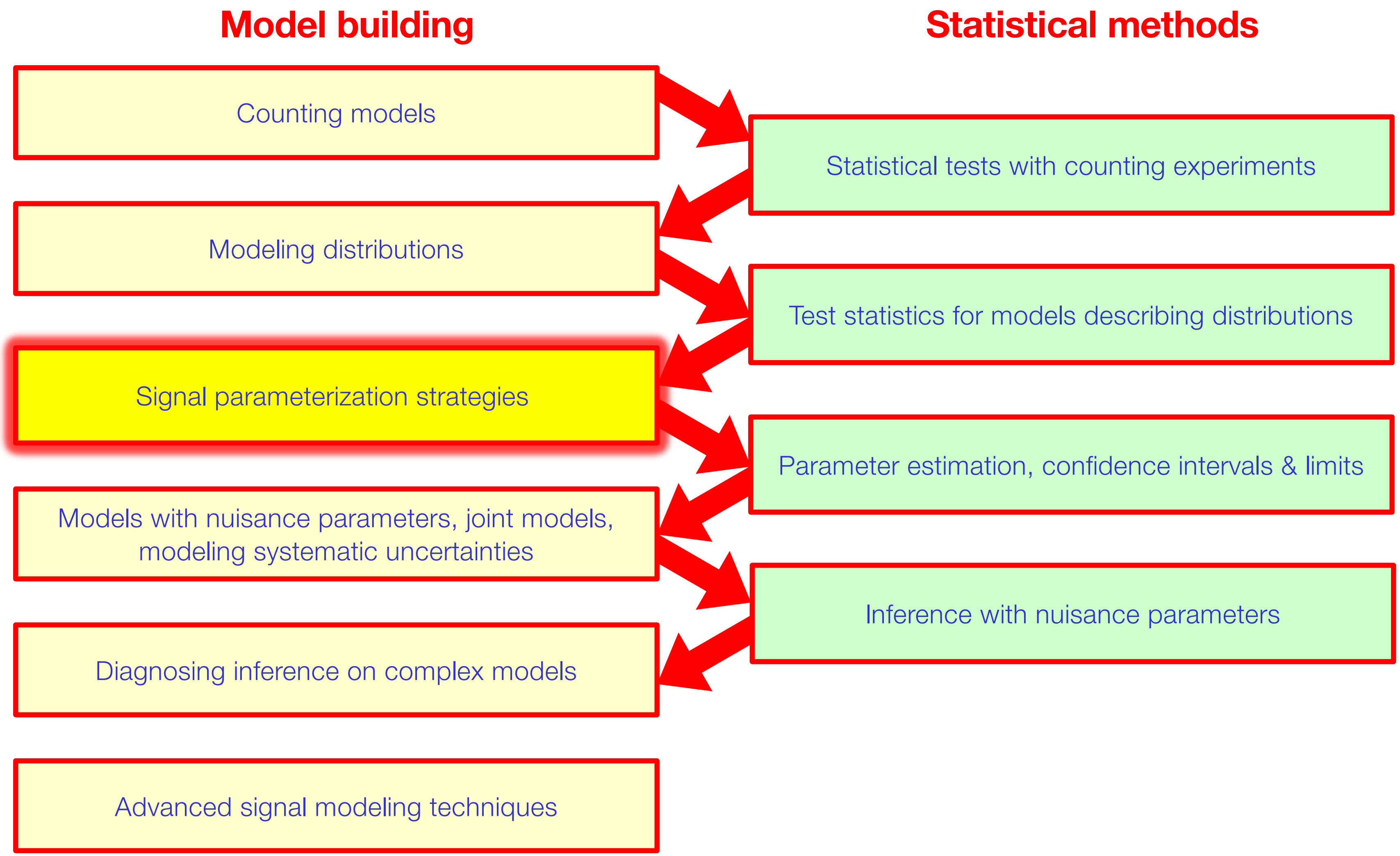
Model building 3

Models with parameters I - analytical parametric models, template morphing approach for histogram-based models

Wouter Verkerke, NIKHEF

Roadmap of this course

- Start with basics, gradually build up to complexity



Model building

Statistical methods

Counting models

Statistical tests with counting experiments

Modeling distributions

Test statistics for models describing distributions

Signal parameterization strategies

Parameter estimation, confidence intervals & limits

Models with nuisance parameters, joint models, modeling systematic uncertainties

Inference with nuisance parameters

Diagnosing inference on complex models

Advanced signal modeling techniques

Introduce concept of composite hypotheses

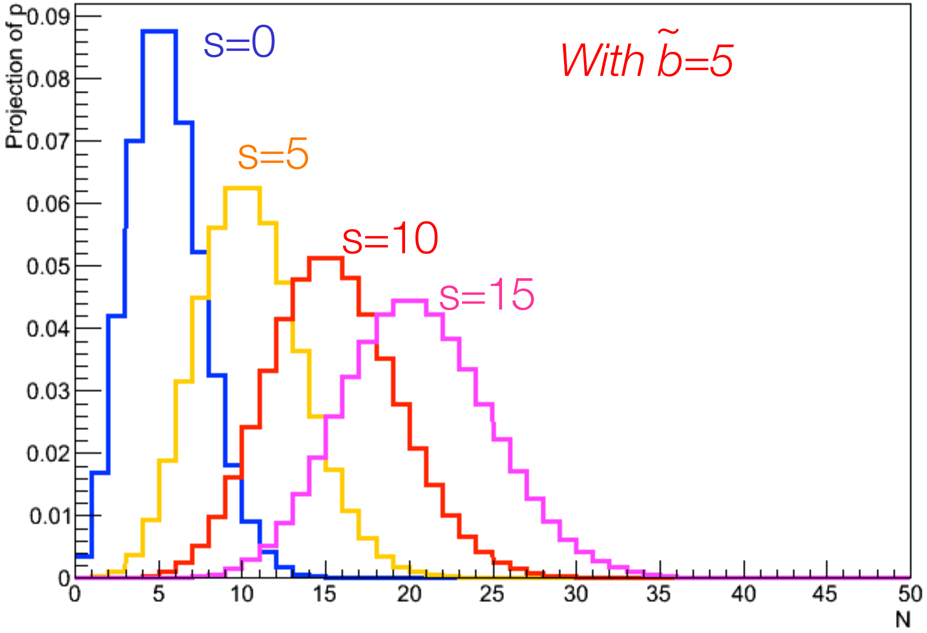
- In most cases in physics, a hypothesis is not “simple”, but “composite”
- **Composite hypothesis** = Any hypothesis which does *not* specify the population distribution completely
- Example: counting experiment with signal and background, that leaves signal expectation unspecified

Simple hypothesis

$$L = Poisson(N \mid \tilde{s} + \tilde{b})$$

$$L(s) = Poisson(N \mid s + \tilde{b})$$

Composite hypothesis



(My) notation convention: all symbols with ~ are constants

A common convention in the meaning of model parameters

- A common convention is to recast signal rate parameters into a normalized form (e.g. w.r.t the Standard Model rate)

Simple hypothesis

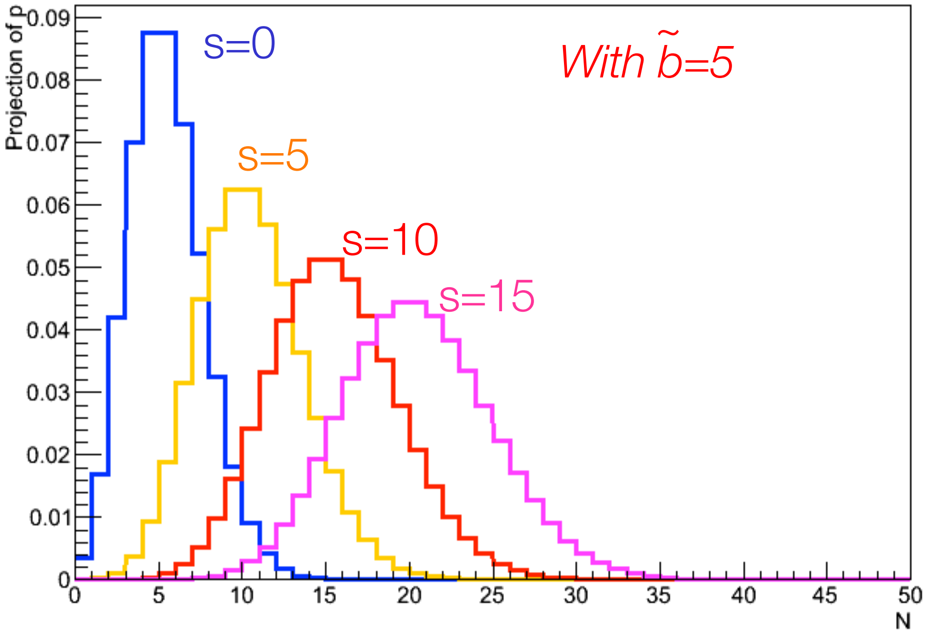
$$L = Poisson(N \mid \tilde{s} + \tilde{b})$$

$$L(s) = Poisson(N \mid s + \tilde{b})$$

Composite hypothesis

$$L(\mu) = Poisson(N \mid \mu \cdot \tilde{s} + \tilde{b})$$

Composite hypothesis
with normalized rate parameter



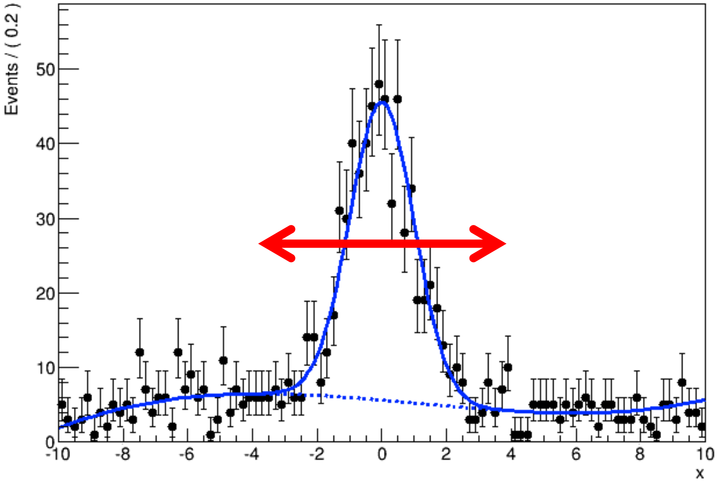
'Universal' parameter interpretation makes it easier to work with your models

μ=0 → no signal
μ=1 → expected signal
μ>1 → more than expected signal

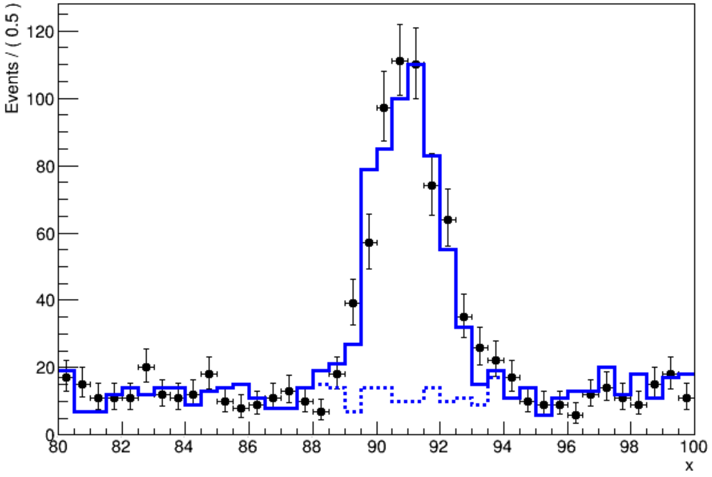
Model building for measurements → shape parameter

- Beyond discovery/rate measurements, can also build models to measure properties of particles (e.g mass)
→ introduce shape parameters
- Often trivial for analytical models,
less so for simulation-based models

F(x|**m**) = Gaussian(x,**m**,σ)+bkg

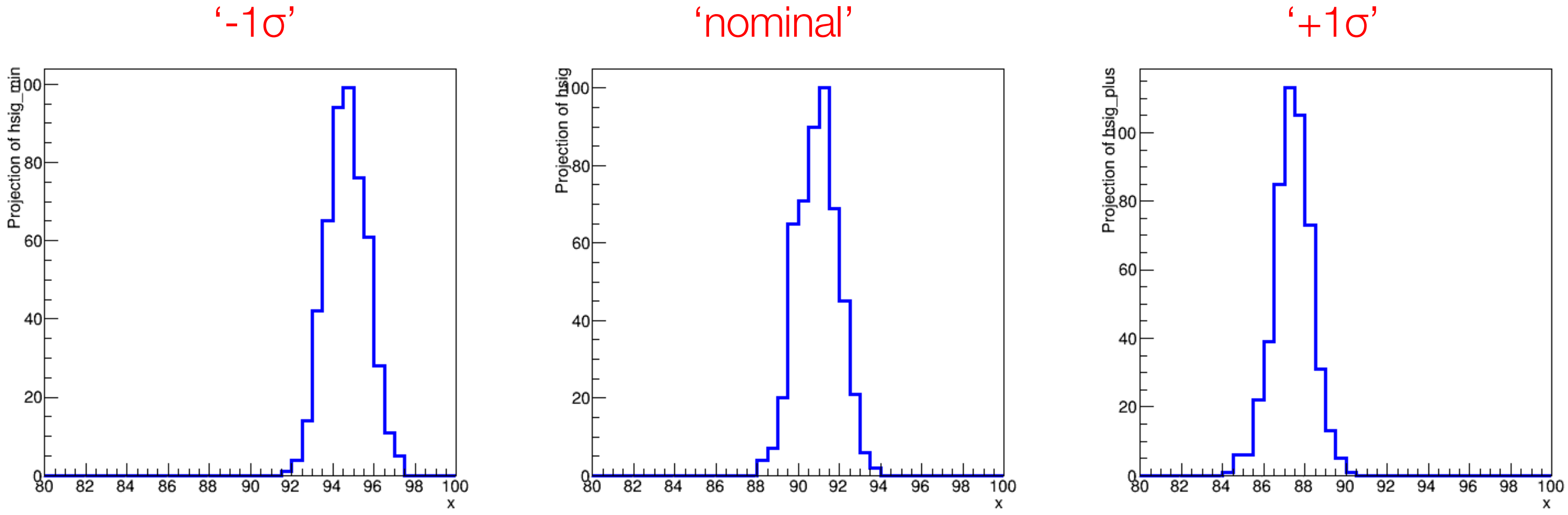


F(x|**m**) = ??



Modeling of shape variations in the likelihood

- If underlying simulation has free parameter θ, can assess impact on reconstructed shapes by rerunning simulation at different values
  - Obtain histogram templates for distributions at '+1σ' and '-1σ' settings of systematic effect



'-1σ'

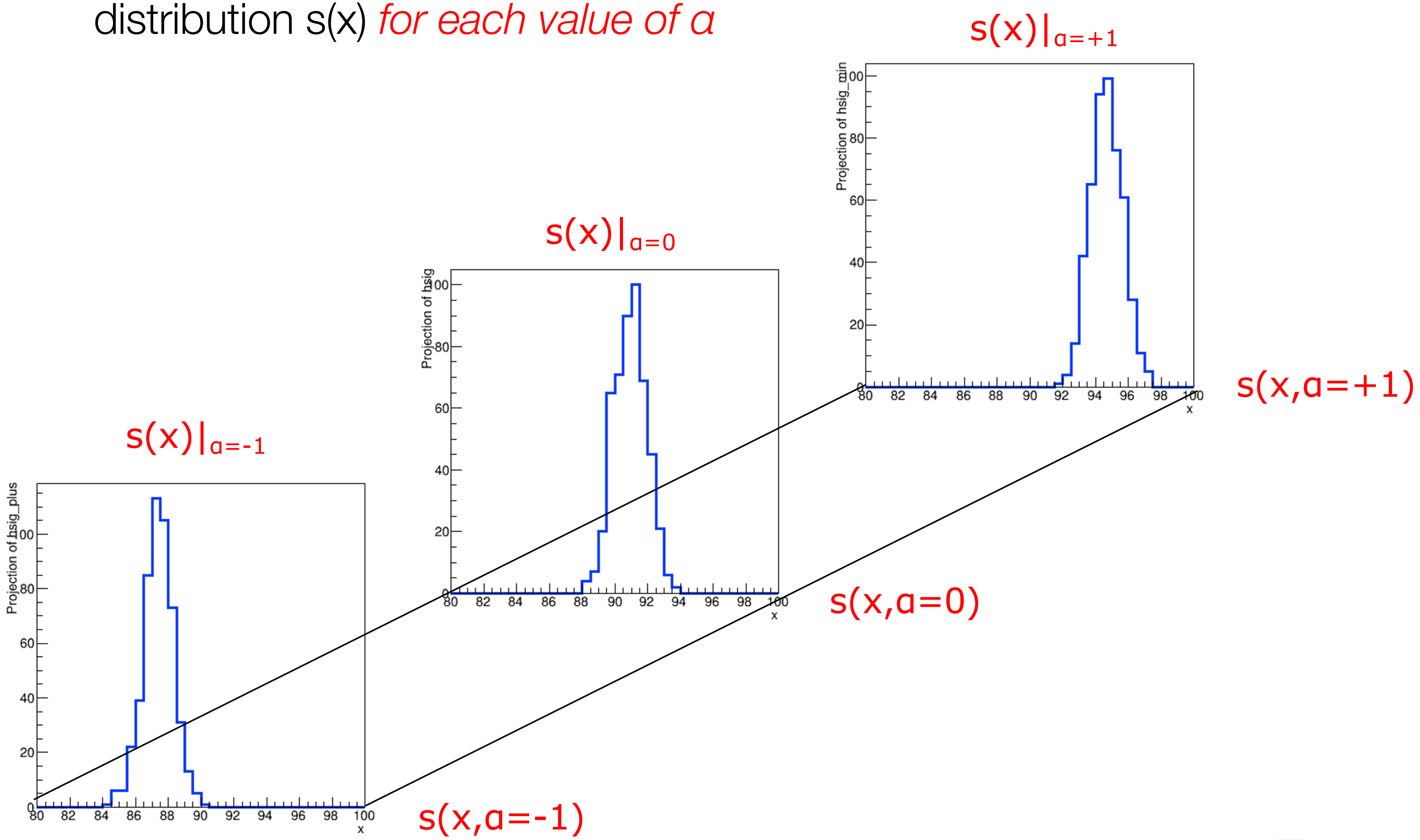
'nominal'

'+1σ'

- Challenge: **construct an empirical response function based on the interpolation of the shapes of these three templates**.

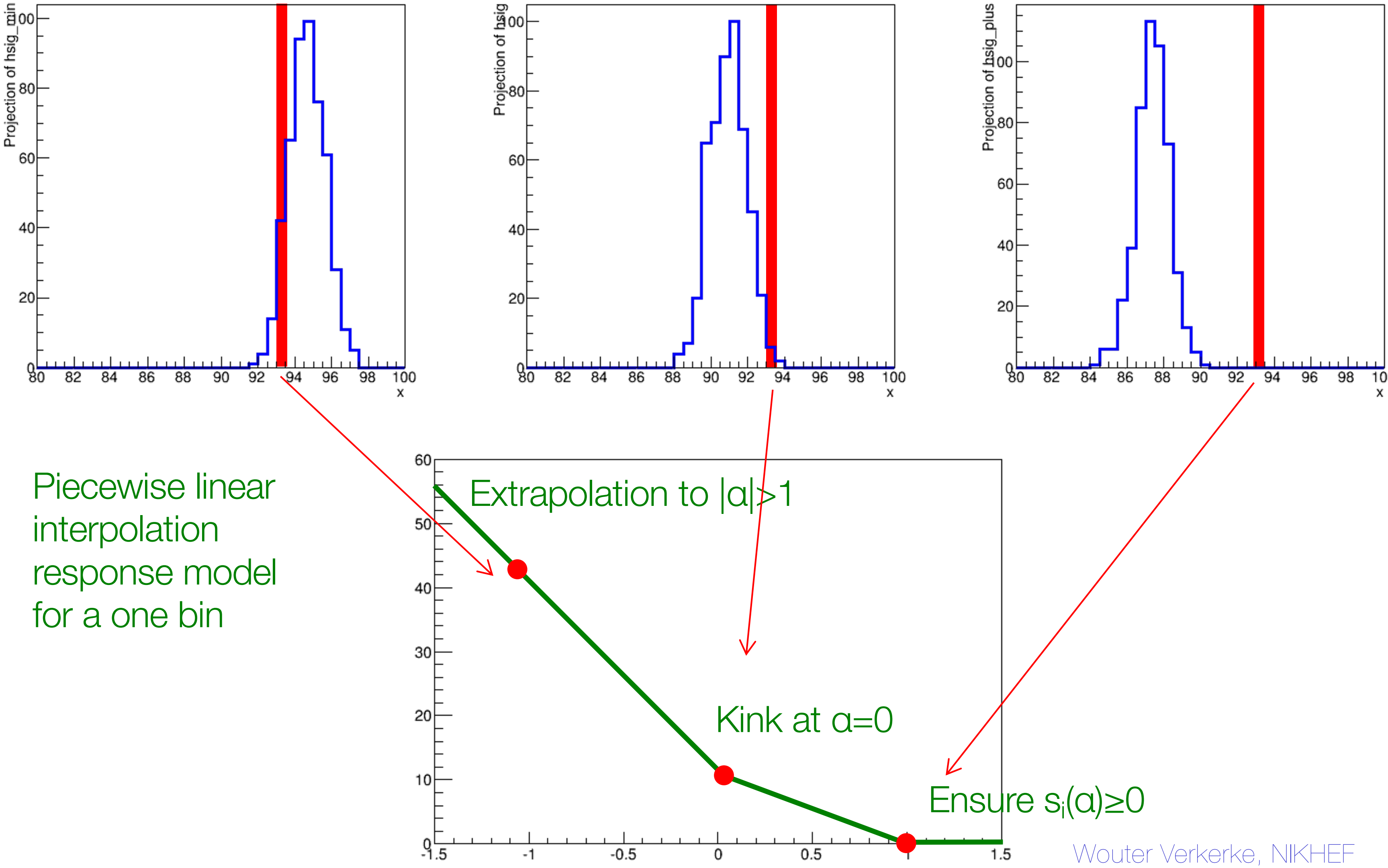
Need to interpolate between template models

- Need to define 'morphing' algorithm to define distribution s(x) *for each value of α*



Piecewise linear interpolation

- Simplest solution is piece-wise linear interpolation for each bin



Piecewise linear interpolation response model for a one bin

Extrapolation to |α|>1

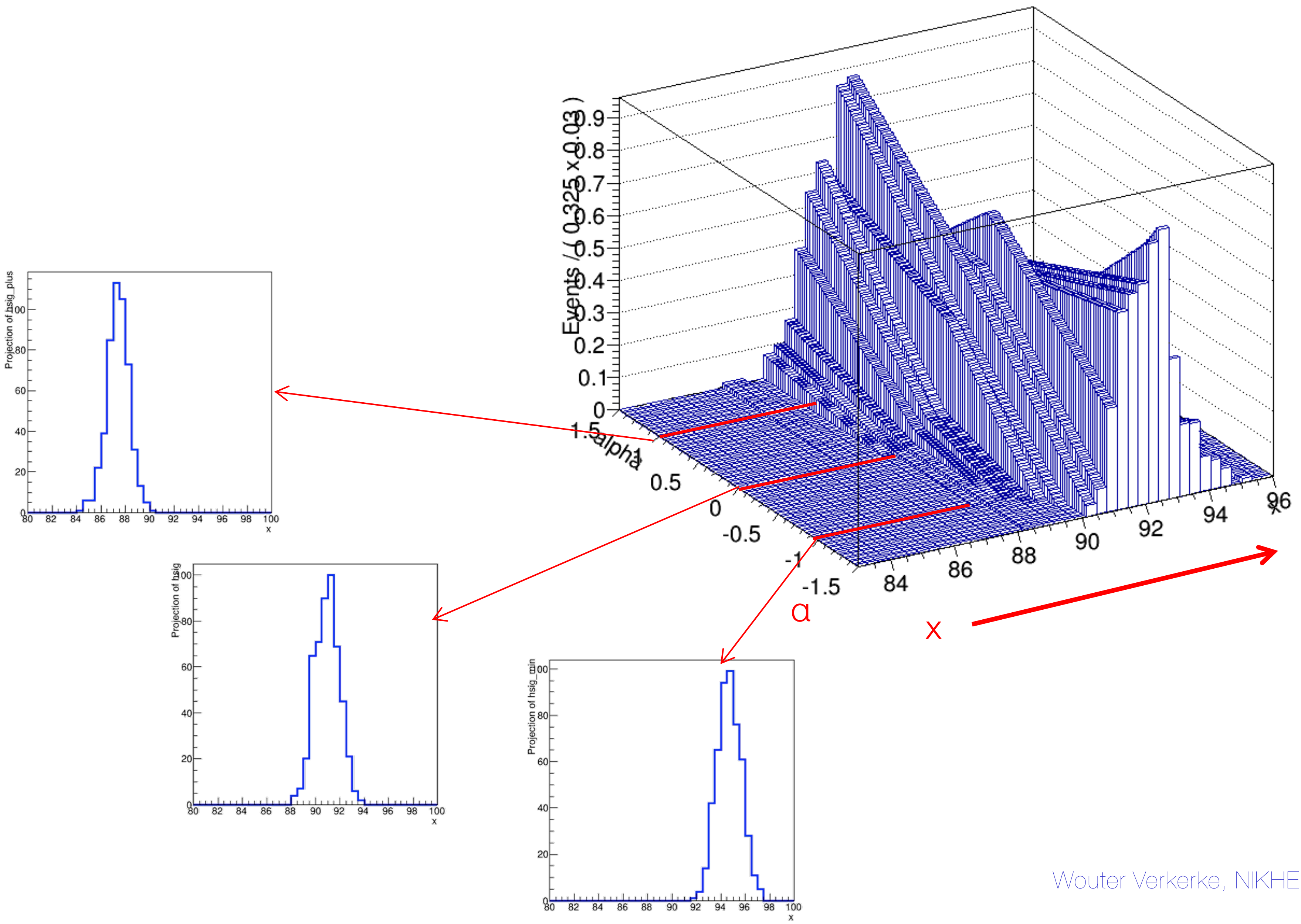
Kink at α=0

Ensure $s_i(\alpha) \geq 0$

Wouter Verkerke, NIKHEF

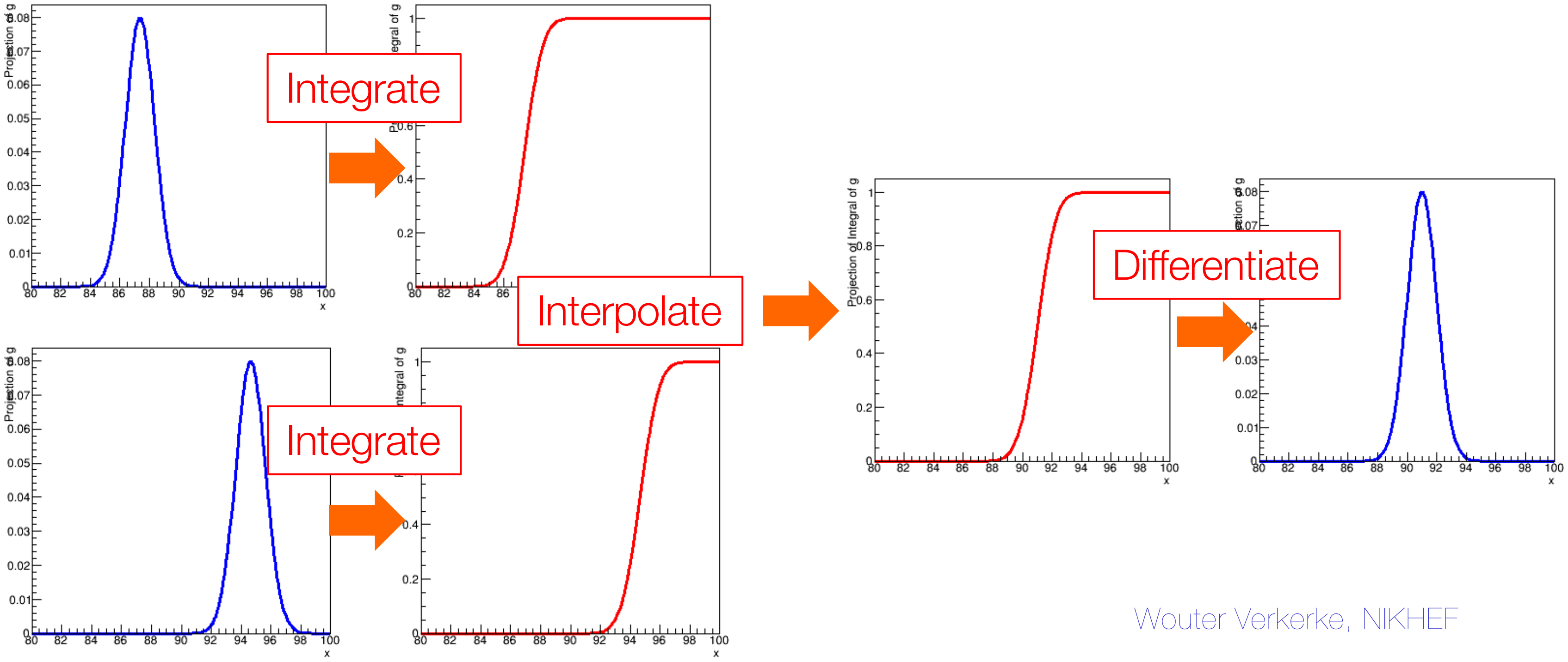
‘ex15_build_binned_morphing.C’

Visualization of bin-by-bin linear interpolation of distribution



Other morphing strategies – ‘horizontal morphing’

- Other template morphing strategies exist that are less prone to unintended side effects
- A ‘horizontal morphing’ strategy was invented by Alex Read.
  - Interpolates the cumulative distribution function instead of the distribution
  - Especially suitable for shifting distributions
  - Here shown on a continuous distribution, but also works on histograms
  - Drawback: computationally expensive, algorithm only worked out for 1 NP



Integrate

Integrate

Interpolate

Differentiate

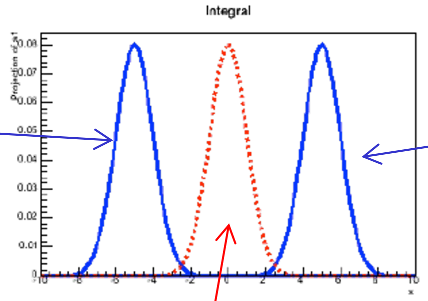
Yet another morphing strategy – ‘Moment morphing’

M. Baak & S. Gadatsch

- Given two template model *f₋(x)* and *f₊(x)* the strategy of moment morphing considers first two moment of template models (mean and variance)

$$\mu_- = \int x \cdot f_-(x)dx$$

$$V_- = \int (x-\mu_-)^2 \cdot f_-(x)dx$$



$$\mu_+ = \int x \cdot f_+(x)dx$$

$$V_+ = \int (x-\mu_+)^2 \cdot f_+(x)dx$$

- The goal of moment morphing is to construct an interpolated function that has linearly interpolated moments

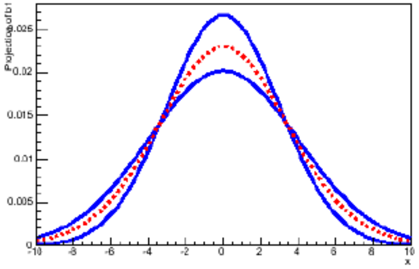
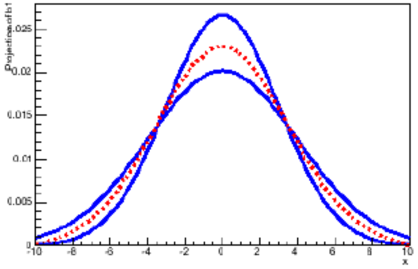
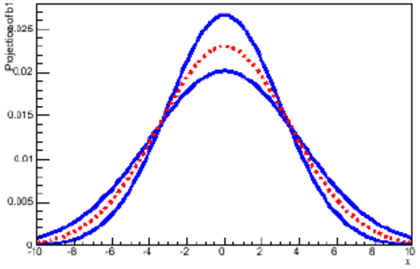
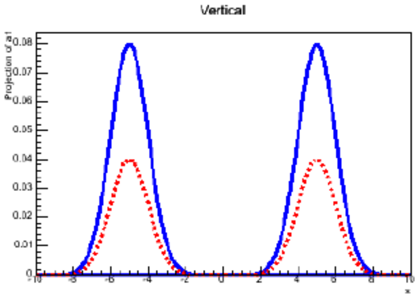
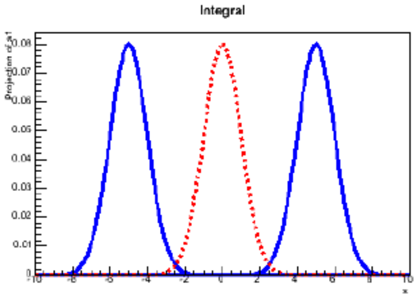
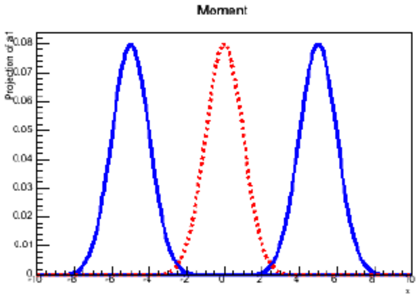
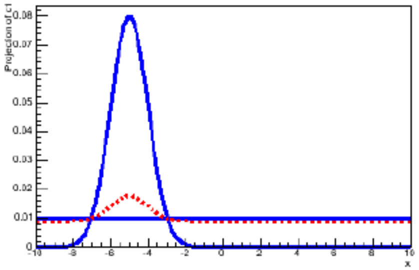
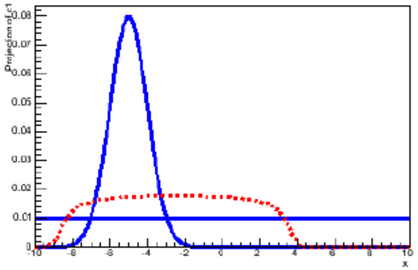
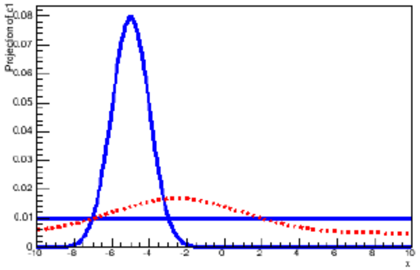
$$\mu(\alpha) = \alpha\mu_- + (1-\alpha)\mu_+$$
$$V(\alpha) = \alpha V_- + (1-\alpha)V_+ \quad [1]$$

- It constructs this morphed function as combination of linearly transformed input models

$$f(x,\alpha) \rightarrow \alpha f_-(ax+b) + (1-\alpha) f_+(cx-d)$$

  - Where constants a,b,c,d are chosen such so that f(x,α) satisfies conditions [1]

There are other morphing algorithms to choose from

	Vertical Morphing	Horizontal Morphing	Moment Morphing
Gaussian varying width			
Gaussian varying mean			
Gaussian to Uniform (this is conceptually ambigous!)			
n-dimensional morphing?	✓	✗	✓

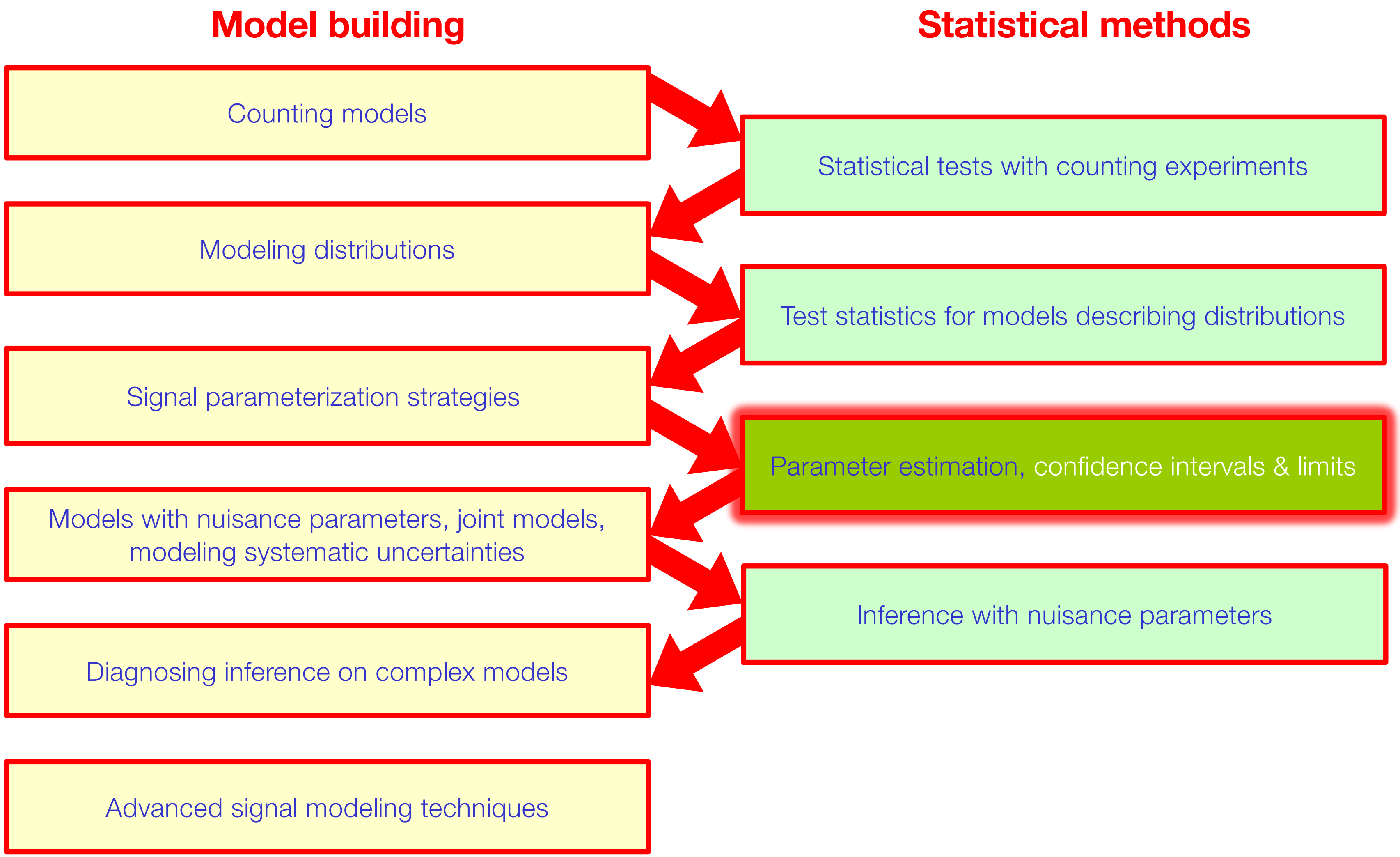
Statistical methods 3

Inference with parameters: maximum likelihood, confidence intervals, upper limits, likelihood ratio and asymptotic formulae

Wouter Verkerke, NIKHEF

Roadmap of this course

- Start with basics, gradually build up to complexity



Model building

Statistical methods

Counting models

Statistical tests with counting experiments

Modeling distributions

Test statistics for models describing distributions

Signal parameterization strategies

Parameter estimation, confidence intervals & limits

Models with nuisance parameters, joint models, modeling systematic uncertainties

Inference with nuisance parameters

Diagnosing inference on complex models

Advanced signal modeling techniques

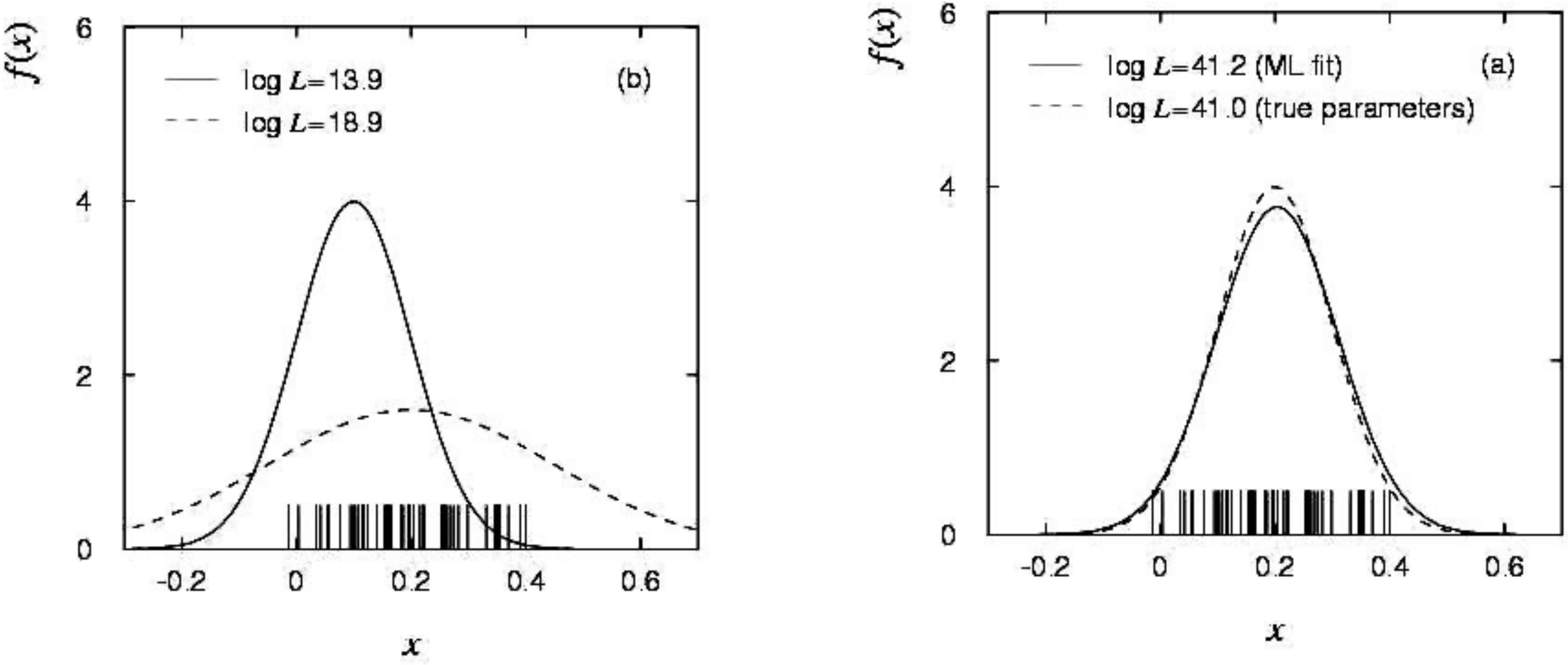
What can we do with composite hypothesis

- With simple hypotheses – inference is restricted to making statements about P(D|hypo) or P(hypo|D)
- With composite hypotheses – many more options
- 1 Parameter estimation and variance estimation
  - What is value of ***s*** for which the observed data is most probable?
  - What is the variance (std deviation squared) in the estimate of ***s?***

  } s=5.5 ± 1.3
- 2 Confidence intervals
  - Statements about model parameters using frequentist concept of probability
  - s<12.7 at 95% confidence level
  - 4.5 < s < 6.8 at 68% confidence level
- 3 Bayesian credible intervals
  - Bayesian statements about model parameters
  - s<12.7 at 95% credibility

Parameter estimation using Maximum Likelihood

- Likelihood is high for values of p that result in distribution similar to data



- Define the maximum likelihood (ML) estimator to be the procedure that finds the parameter value for which the likelihood is maximal.

‘ex06.C’

Parameter estimation – Maximum likelihood

- Practical estimation of maximum likelihood performed by minimizing the negative log-Likelihood

$$L(\vec{p}) = \prod_i f(\vec{x}_i;\vec{p})$$

$$\Downarrow$$

$$-\ln L(\vec{p}) = -\sum_i \ln F(\vec{x}_i;\vec{p})$$

  - Advantage of log-Likelihood is that contributions from events can be summed, rather than multiplied (computationally easier)

- In practice, find point where derivative of –logL is zero

$$\left.\frac{d\ln L(\vec{p})}{d\vec{p}}\right|_{p_i=\hat{p}_i} = 0$$

- Standard notation for ML estimation of p is $\hat{p}$

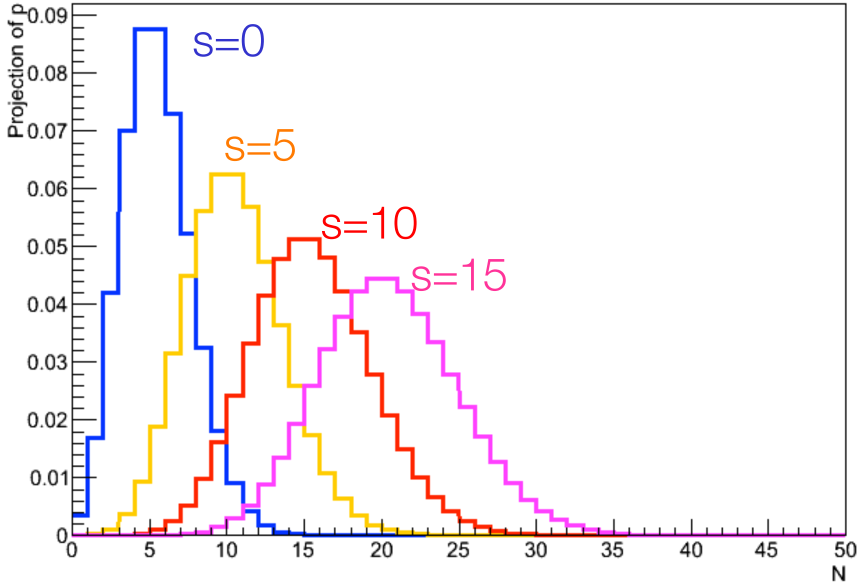
'ex09.C'

Example of Maximum Likelihood estimation

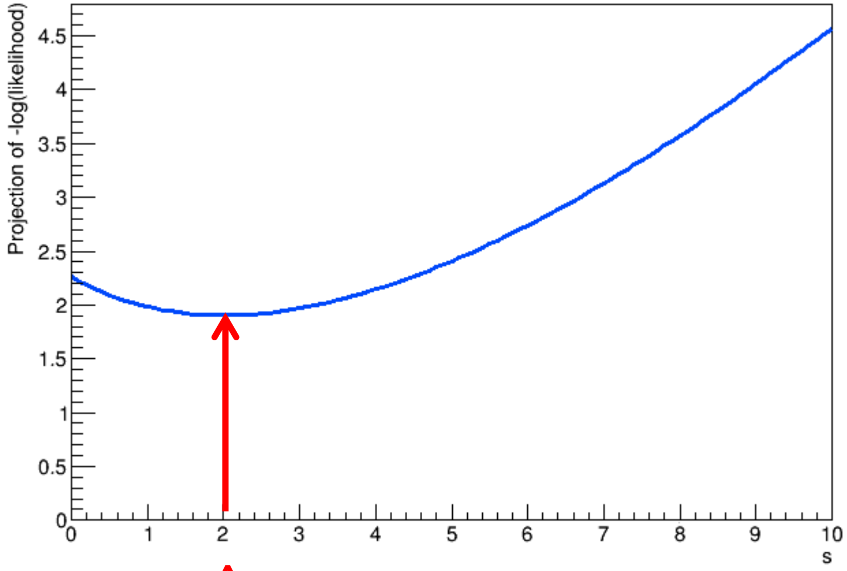
- Illustration of ML estimate on Poisson counting model

$$L(N\,|\,s) = Poisson(N\,|\,s+\tilde{b})$$

-log *L*(*N*|*s*) versus *N* [*s*=0,5,10,15]



-log *L*(*N*|*s*) versus *s* [N=7]



$\hat{s}=2$

- Note that Poisson model is discrete in N, *but continuous in s!*

Properties of Maximum Likelihood estimators

- In general, Maximum Likelihood estimators are
  - Consistent (gives right answer for N→∞)
  - Mostly unbiased (bias $\propto 1/N$, may need to worry at small N) 'ex05.C'
  - Efficient for large N (you get the smallest possible error)
  - Invariant: (a transformation of parameters will Not change your answer, e.g $(\hat{p})^2 = \widehat{\left(p^2\right)}$

- MLE efficiency theorem: the MLE will be *unbiased* and *efficient* if an unbiased efficient estimator exists
  - Proof not discussed here
  - Of course this does not guarantee that any MLE is unbiased and efficient for any given problem

Relation between Likelihood and χ^2 estimators

- Properties of χ^2 estimator follow from properties of ML estimator using *Gaussian probability density functions*

$$F(x_i, y_i, \sigma_i; \vec{p}) = \prod_i \exp\left[-\left(\frac{y_i - f(x_i;\vec{p})}{\sigma_i} \right)^2 \right]$$

Gaussian Probability Density Function in p for single measurement y±σ from a predictive function f(x|p)

Take log,
Sum over all points (***x_i*** **,** ***y_i*** **,** ***σ_i***)

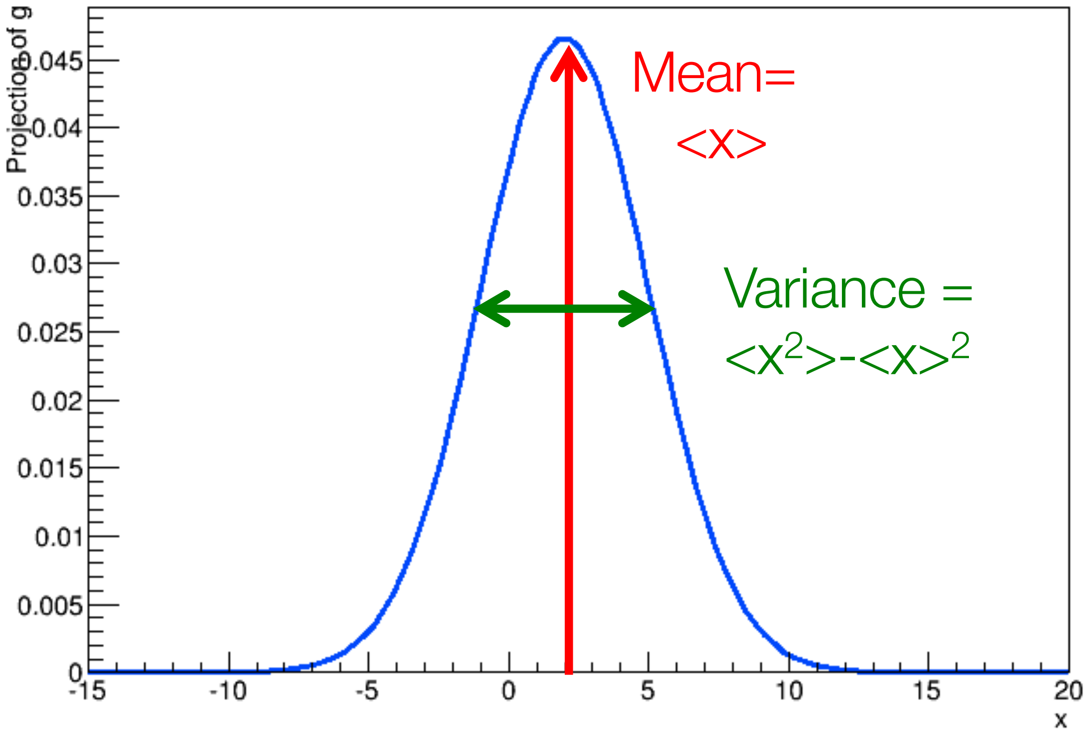
$$-\ln L(\vec{p}) = \tfrac{1}{2} \sum_i \left(\frac{y_i - f(x_i;\vec{p})}{\sigma_i} \right) = \tfrac{1}{2} \chi^2$$

The Likelihood function in p for given points $x_i(s_i)$ and function $f(x_i;p)$

- The χ^2 estimator follows from ML estimator, i.e it is
  - Efficient, consistent, bias 1/N, invariant,
  - But only in the limit that the error **on *x_i*** is truly Gaussian

Estimating parameter variance

- Note that 'uncertainty' on a parameter estimate is an ambiguous statement
- Can either mean an interval with a stated confidence or credible, level (e.g. 68%), or simply assume it is the square-root of the variance of a distribution



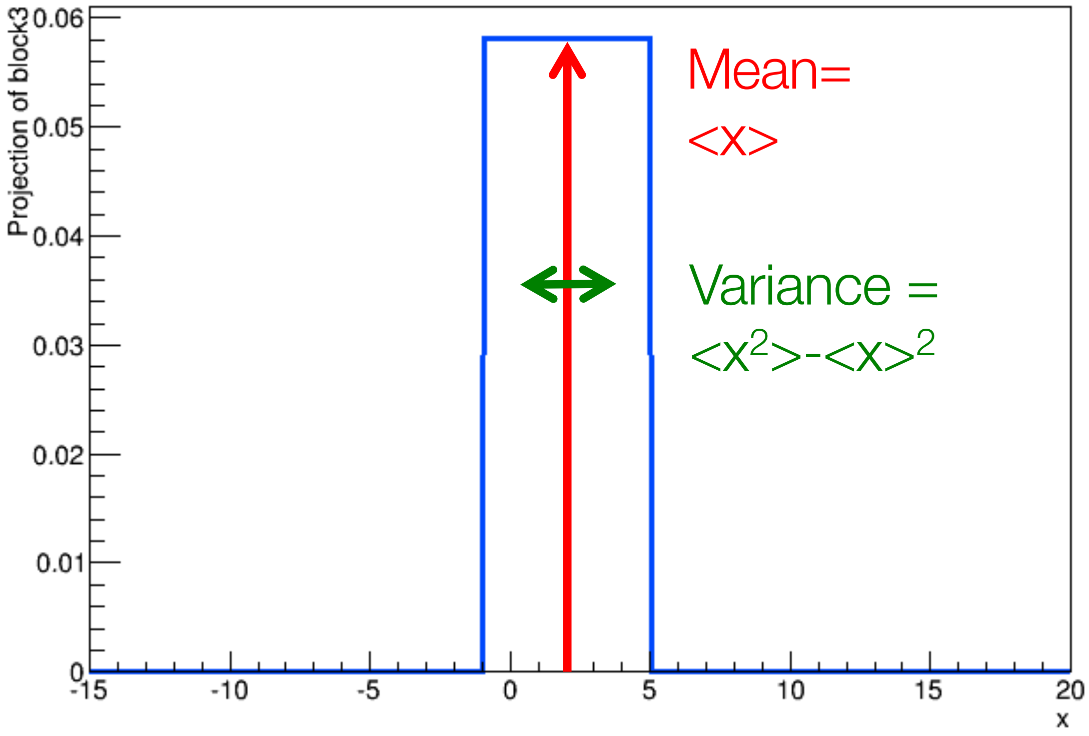
For a Gaussian distribution mean and variance map to parameters for *mean* and *sigma*^{2}

and interval defined by $\sqrt{V}$ contains 68% of the distribution (='1 sigma' by definition)

Thus for Gaussian distributions all common definitions of 'error' work out to the same numeric value

Estimating parameter variance

- Note that ‘uncertainty’ on a parameter estimate is an ambiguous statement
- Can either mean an interval with a stated confidence or credible, level (e.g. 68%), or simply assume it is the square-root of the variance of a distribution



Mean= <x>

Variance = $\langle x^2\rangle - \langle x\rangle^2$

For other distributions intervals by √V do not necessarily contain 68% of the distribution

Estimating variance on parameters

- Variance on of parameter can also be estimated from Likelihood using the variance estimator

$$\hat{\sigma}(p)^2 = \hat{V}(p) = \left(\frac{d^2 \ln L}{d^2 p}\right)^{-1}$$

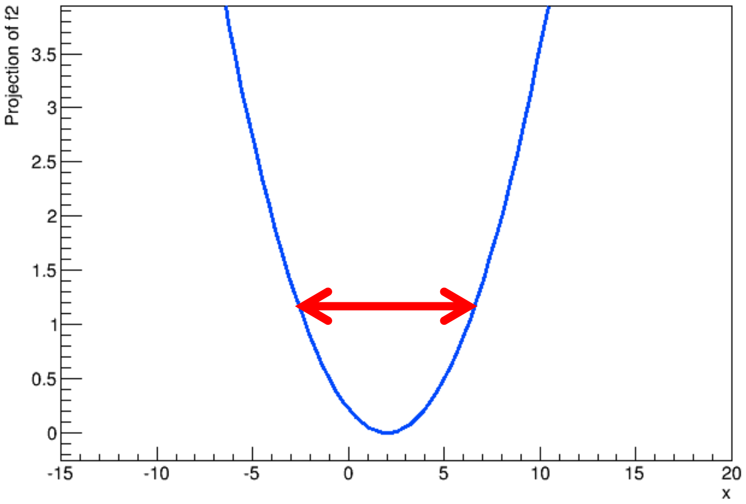
From Rao-Cramer-Frechet inequality

$$V(\hat{p}) \geq \left(1 + \frac{db}{dp}\right) \Big/ \left(\frac{d^2 \ln L}{d^2 p}\right)$$

b = bias as function of p, inequality becomes equality in limit of efficient estimator

- Valid if estimator is efficient and unbiased!

- Illustration of Likelihood Variance estimate on a Gaussian distribution



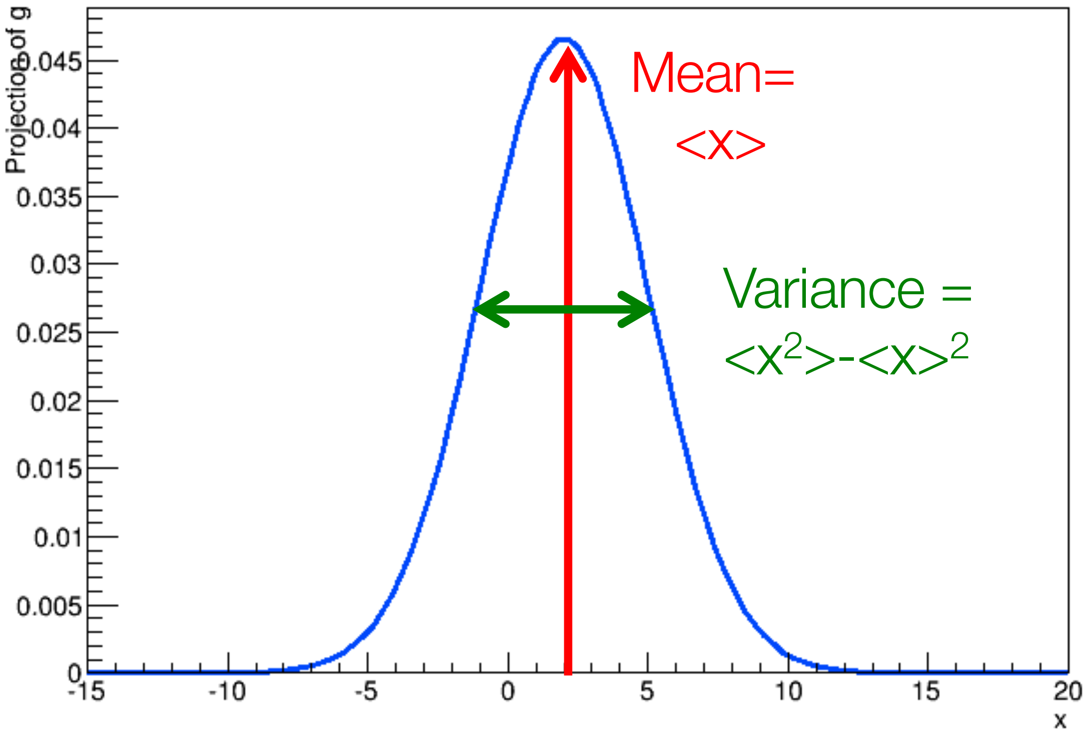
$$f(x|\mu,\sigma) = \frac{1}{\enclose{circle}{\sigma}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\ln f(x|\mu,\sigma) = -\ln \enclose{circle}{\sigma} - \ln\sqrt{2\pi} + \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

$$\left.\frac{d\ln f}{d\sigma}\right|_{x=\mu} = \frac{-1}{\enclose{circle}{\sigma}} \quad \Rightarrow \quad \left.\frac{d^2 \ln f}{d^2\sigma}\right|_{x=\mu} = \frac{1}{\sigma^2}$$

Bayesian parameter estimation

- Bayesian parameter estimate is the posterior mean
- Bayesian variance is the posterior variance



$$\hat{\mu} = \int \mu P(\mu \mid N) d\mu$$

$$\hat{V} = \int (\hat{\mu} - \mu)^2 P(\mu \mid N) d\mu$$

What can we do with composite hypothesis

- With simple hypotheses – inference is restricted to making statements about P(D|hypo) or P(hypo|D)
- With composite hypotheses – many more options
- 1 Parameter estimation and variance estimation
  - What is value of ***s*** for which the observed data is most probable?
  - What is the variance (std deviation squared) in the estimate of ***s?***

  } s=5.5 ± 1.3
- 2 Confidence intervals
  - Statements about model parameters using frequentist concept of probability
  - s<12.7 at 95% confidence level
  - 4.5 < s < 6.8 at 68% confidence level
- 3 Bayesian credible intervals
  - Bayesian statements about model parameters
  - s<12.7 at 95% credibility

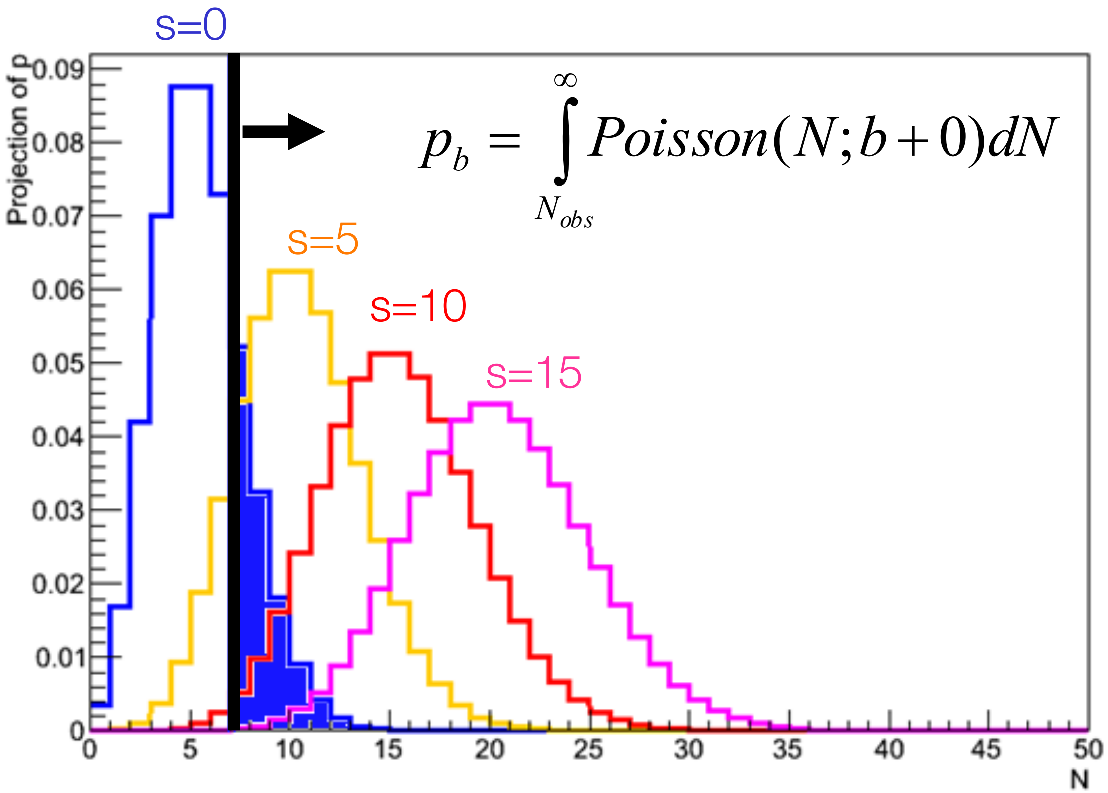
Interval estimation with fundamental methods

- Can also construct parameters intervals using ‘fundamental’ methods explored earlier (Bayesian or Frequentist)
- Construct Confidence Intervals or Credible Intervals with defined probabilistic meaning, independent of assumptions on normality of distribution (Central Limit Theorem) → “95% C.L.”
- With fundamental methods you greater flexibility in types of interval. E.g when no signal observed → usually wish to set an upper limit (construct ‘upper limit interval’)

Reminder - Frequentist test statistics and p-values

- Definition of ‘p-value’: *Probability to observe this outcome or more extreme in future repeated measurements is x%,* if hypothesis is true

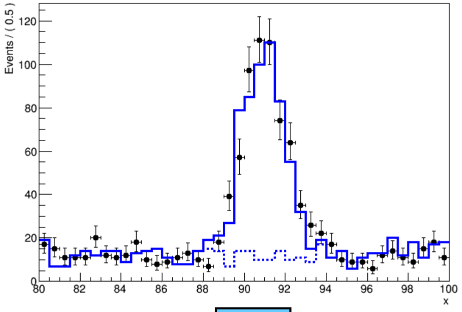
- Note that the definition of p-value assumes an explicit ordering of possible outcomes in the ‘or more extreme’ part



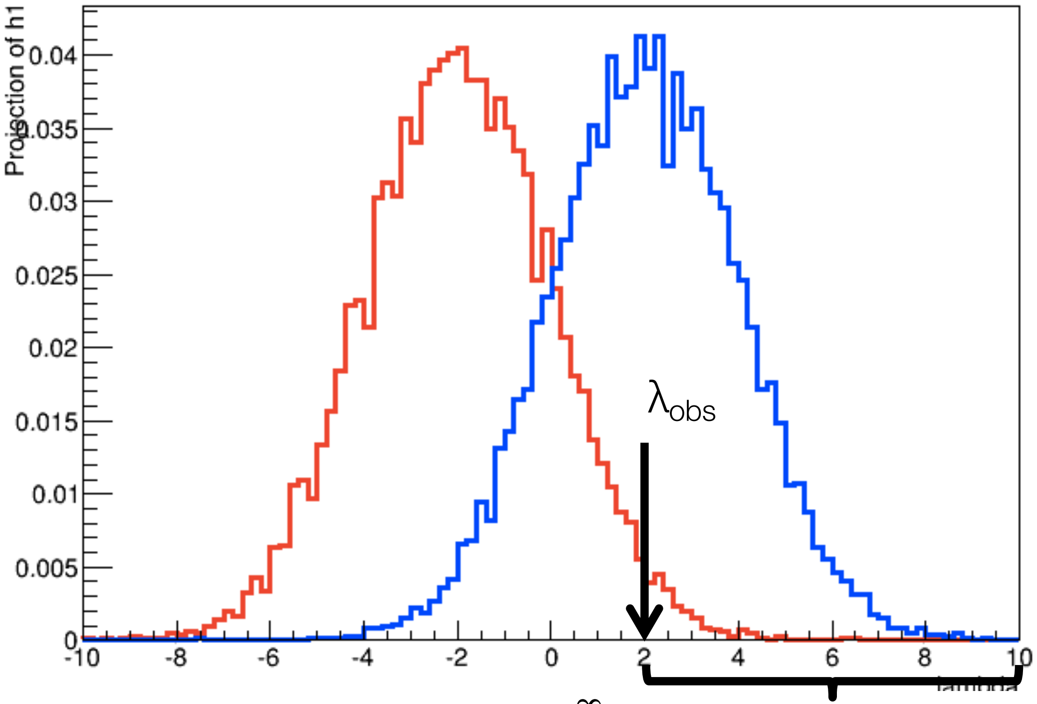
$$p_b = \int_{N_{obs}}^{\infty} Poisson(N;b+0)dN \quad (=0.23)$$

P-values with a likelihood ratio test statistic

- With the introduction of a (likelihood ratio) test statistic, hypothesis testing of models of arbitrary complexity is now reduced to the same procedure as the Poisson example



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



log(λ)

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

- *Except that we generally don't know distribution f(λ)…*

A different Likelihood ratio for composite hypothesis testing

- On *composite hypotheses,* where both null and alternate hypothesis map to values of μ, we can define an alternative likelihood-ratio test statistics that has better properties

‘simple hypothesis’

‘composite hypothesis’

Hypothesis μ that is being tested

$$\lambda(\vec{N}) = \frac{L(\vec{N}\,|\,H_0)}{L(\vec{N}\,|\,H_1)} \quad \Longrightarrow \quad \lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N}\,|\,\mu)}{L(\vec{N}\,|\,\hat{\mu})}$$

‘Best-fit value’

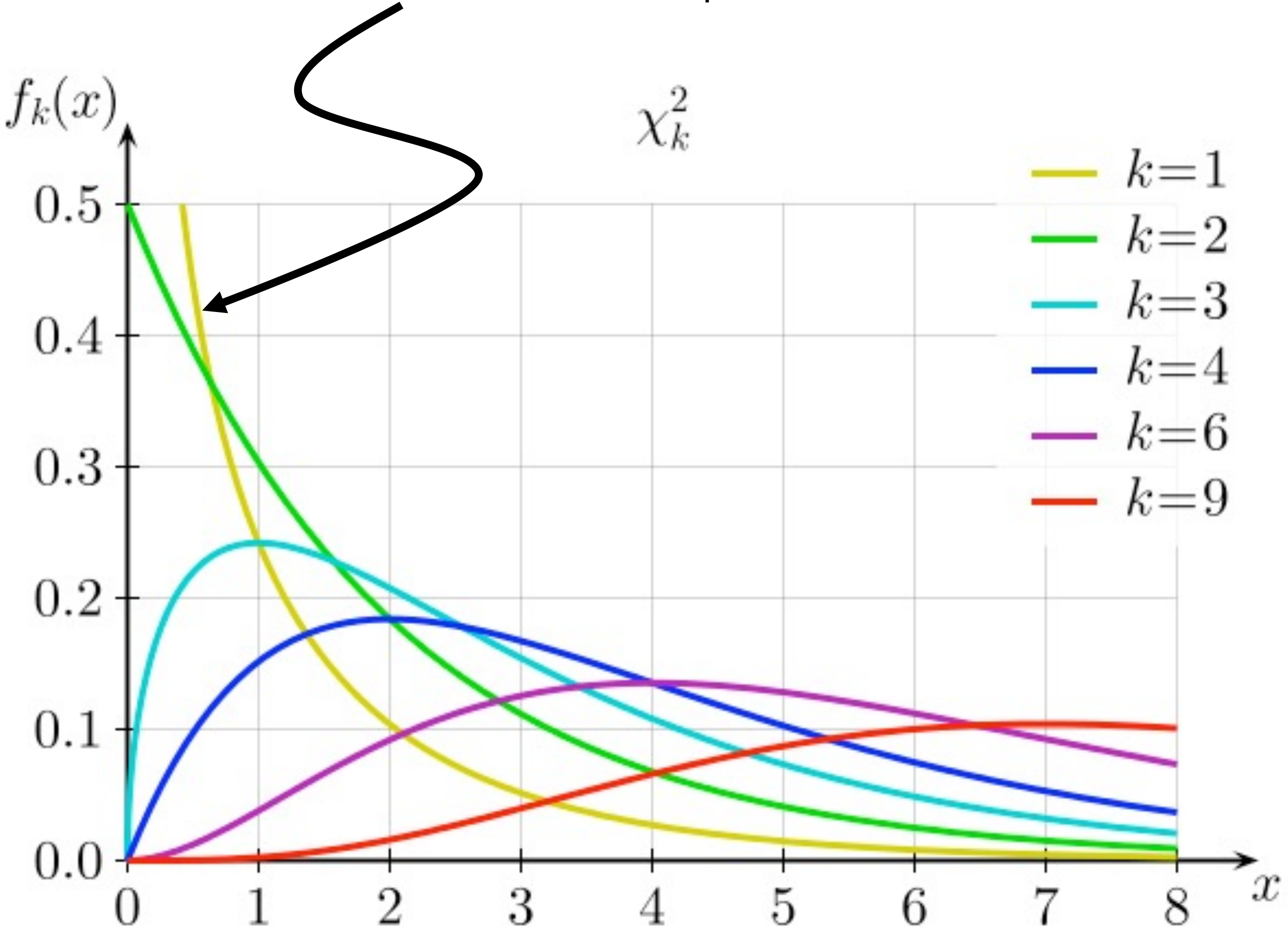
- **Advantage: distribution of new λ_μ has <u>known asymptotic form</u>**
- Wilks theorem: distribution of –log(λ_μ) is asymptotically distribution as a χ^2 with N_{param} degrees of freedom*

*Some regularity conditions apply

- → Asymptotically, we can *directly* calculate p-value from λ_μ^{obs}

What does a χ^2 distribution look like for n=1?

- Note that it for n=1, it does not peak at 1, but rather at 0…



Composite hypothesis testing in the asymptotic regime

- For ‘histogram example’: what is p-value of null-hypothesis

‘likelihood assuming zero signal strength’

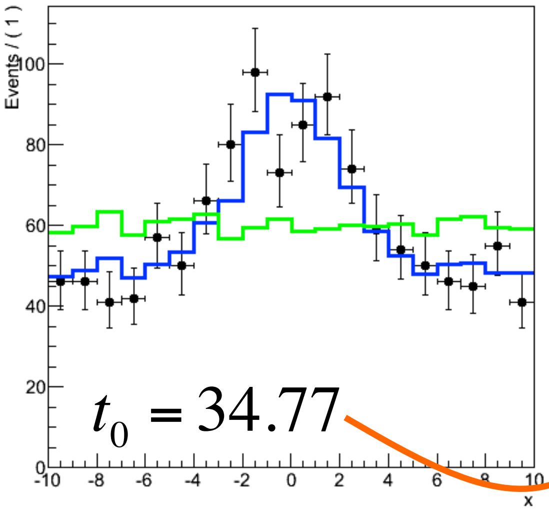
$$t_0 = -2\ln\frac{L(data\,|\,\mu=0)}{L(data\,|\,\hat{\mu})}$$

$\hat{\mu}$ is best fit value of μ

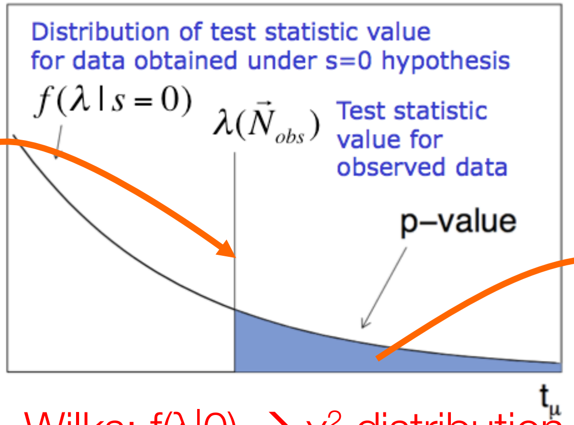
‘likelihood of best fit’

$-\log\mu$

On signal-like data t_0 is large



$t_0 = 34.77$



Distribution of test statistic value for data obtained under s=0 hypothesis

Test statistic value for observed data

Wilks: f(λ|0) → χ^2 distribution

P-value = TMath::Prob(34.77,1)
= $3.7\text{x}10^{-9}$

Composite hypothesis testing in the asymptotic regime

- For 'histogram example': what is p-value of null-hypothesis

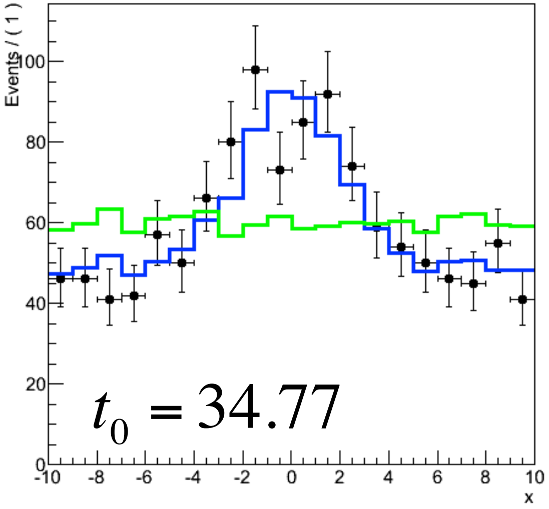
'likelihood assuming zero signal strength'

$$t_0 = -2\ln\frac{L(data\,|\,\mu=0)}{L(data\,|\,\hat{\mu})}$$

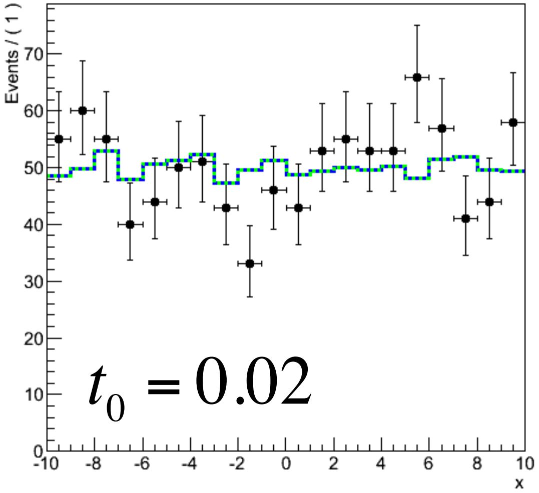
$\hat{\mu}$ is best fit value of μ

'likelihood of best fit'

On signal-like data t_0 is large



On background-like data t_0 is small



Use Wilks Theorem

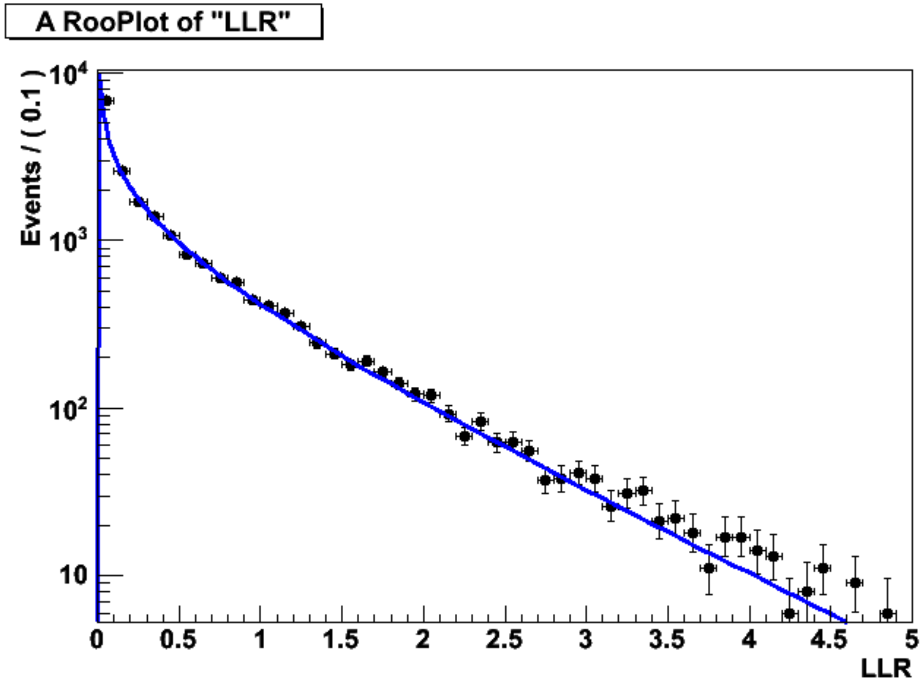
P-value = TMath::Prob(34.77,1)
= 3.7×10^{-9}

P-value = TMath::Prob(0.02,1)
= 0.88

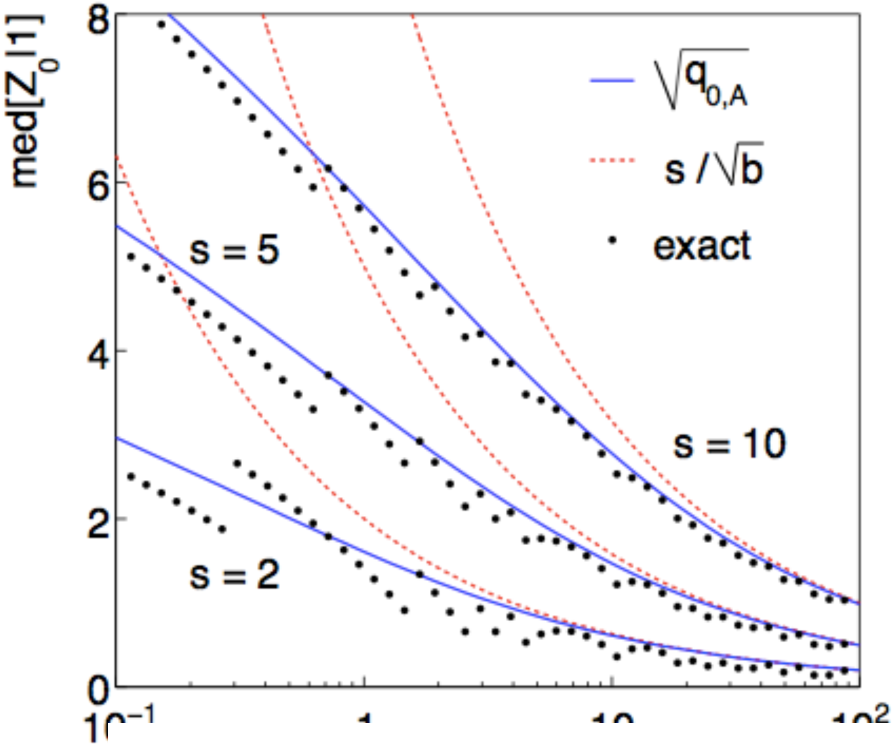
How quickly does $f(\lambda_{\mu}|\mu)$ converge to its asymptotic form

- Pretty quickly –

Here is an example of likelihood function for 10-bin distribution with 200 events



Here is an example for event counting at various s,b



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

From hypothesis testing to confidence intervals

- Next step for composite hypothesis is to go from p-values for a hypothesis defined by fixed value of μ to *an interval statement on μ*

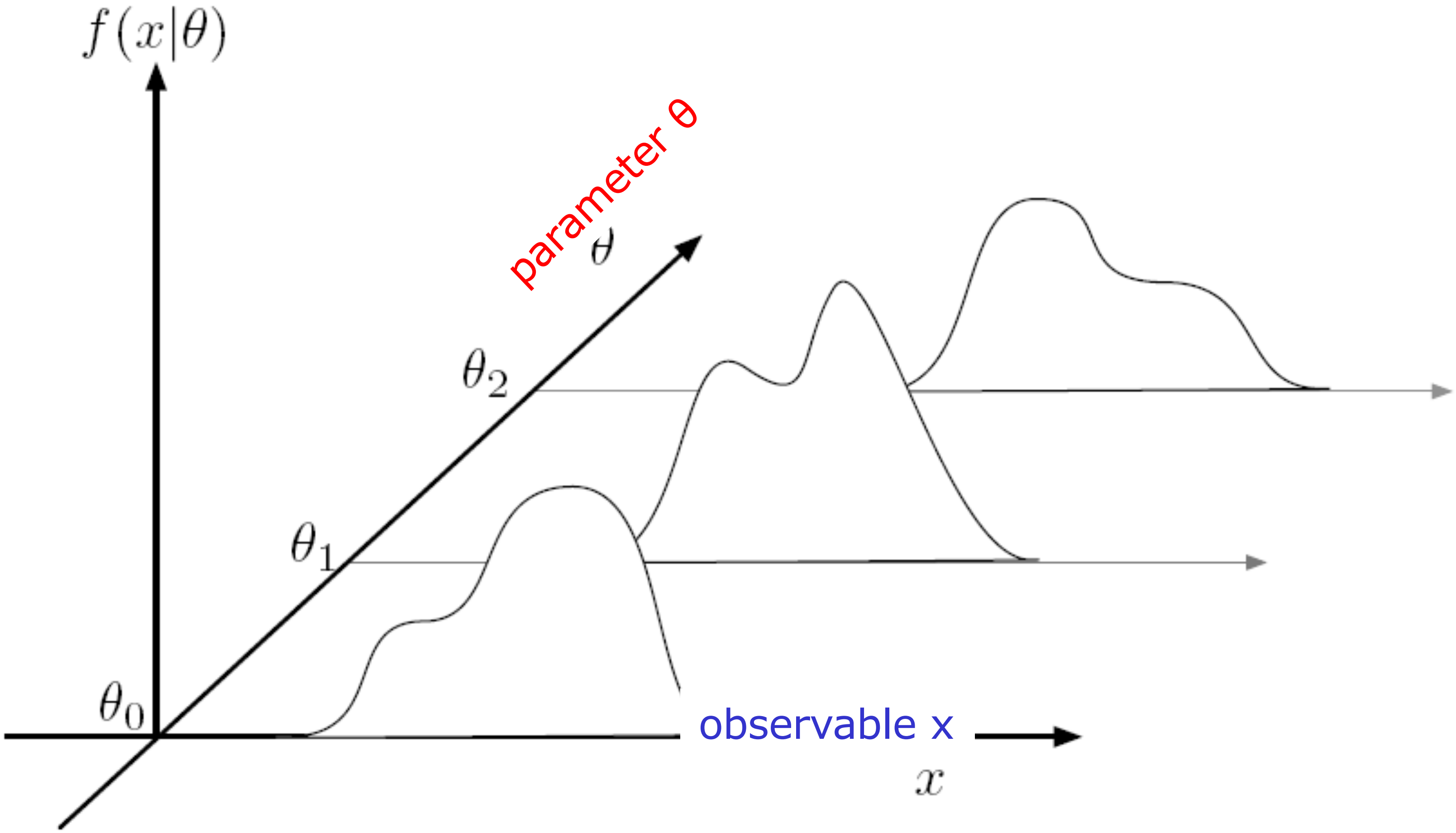
- Definition: A interval on *μ* at X% confidence level is defined such that the true of value of *μ* is contained X% of the time in the interval.
  - Note that the output is *not* a probabilistic statement on the true s value
  - The true μ is fixed but unknown – each observation will result in an estimated interval $[\mu_-,\mu_+]$. X% of those intervals will contain the true value of μ
  - Coverage = guarantee that probabilistic statements is true (i.e. repeated future experiments do reproduce results in X% of cases)

- Definition of confidence intervals does not make any assumption on shape of interval

  → Can choose one-sided intervals ('limits'),
  two-sided intervals ('measurements'),
  or even disjoint intervals ('complicated measurements')

Exact confidence intervals – the Neyman construction

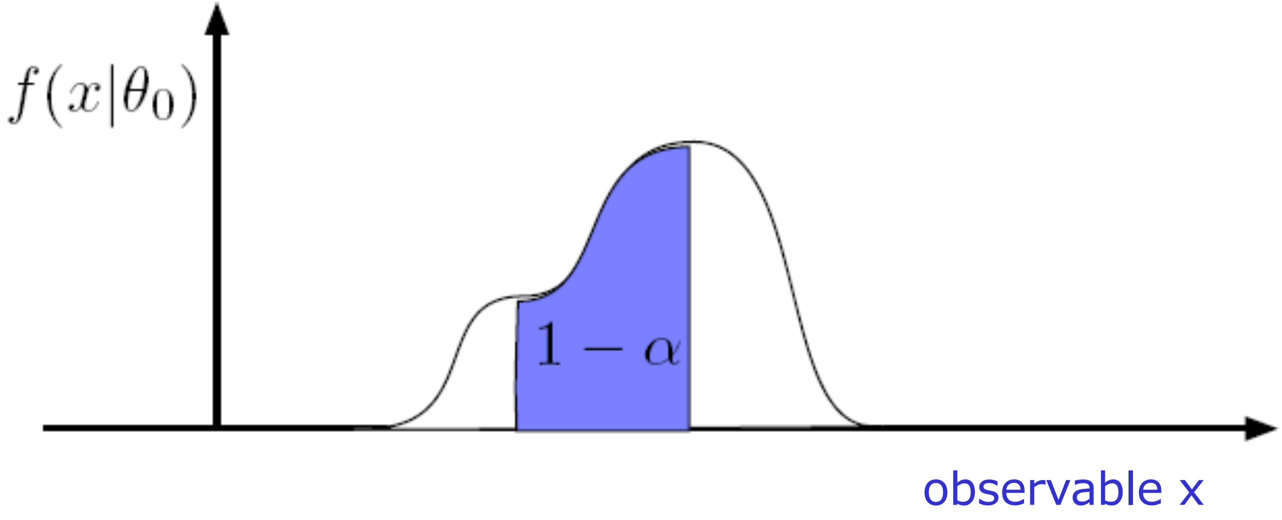
- Simplest experiment: one measurement (x), one theory parameter (θ)
- For each value of parameter θ, determine distribution in in observable x



How to construct a Neyman Confidence Interval

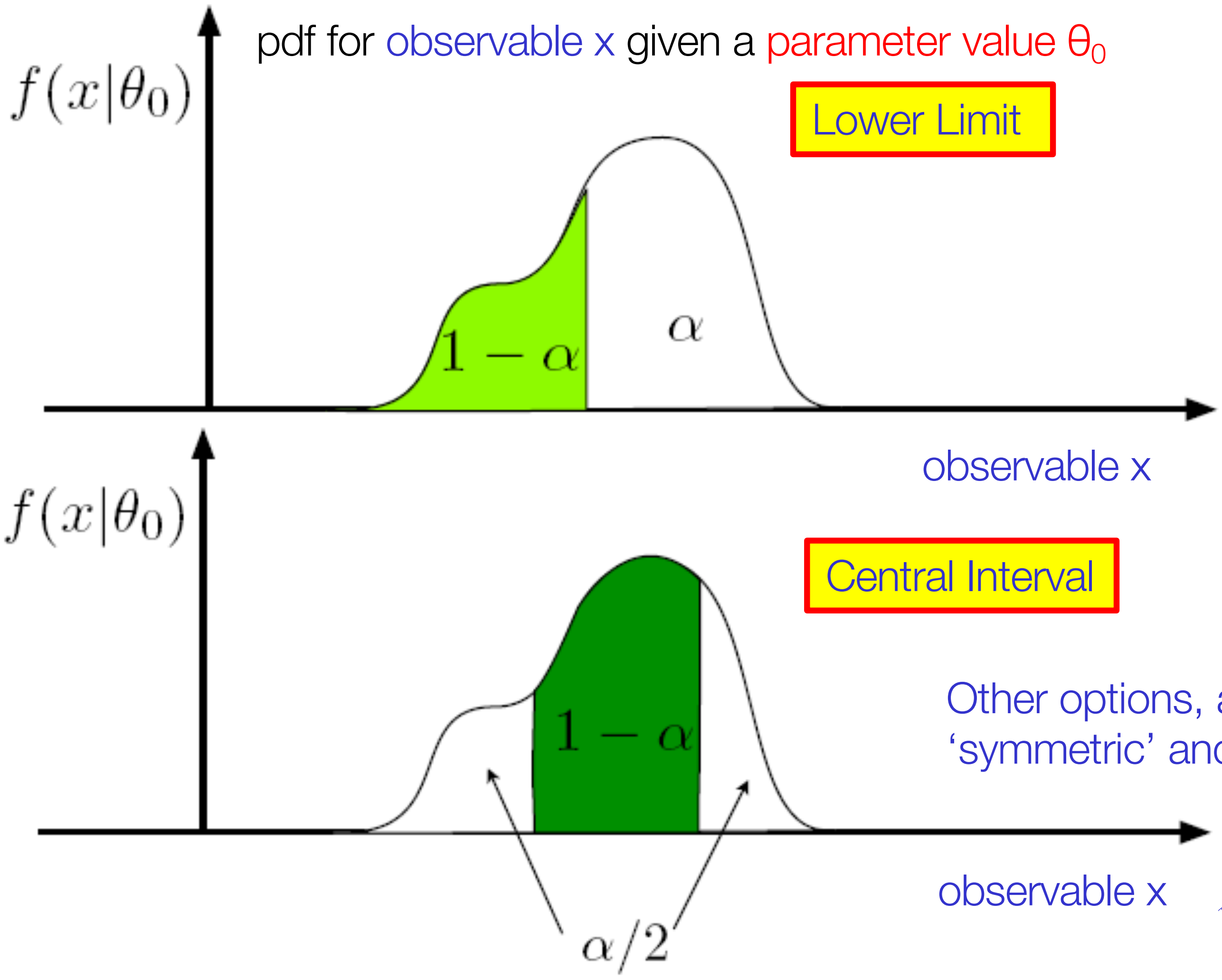
- Focus on a slice in θ
  - For a 1-α% confidence Interval, define ***acceptance interval*** that contains 100%-α% of the distribution

pdf for observable x
given a parameter value θ_0



How to construct a Neyman Confidence Interval

- Definition of acceptance interval is not unique
  → Choose shape of interval you want to set here.
  - Algorithm to define acceptance interval is called 'ordering rule'



pdf for observable x given a parameter value θ_0

Lower Limit

observable x

Central Interval

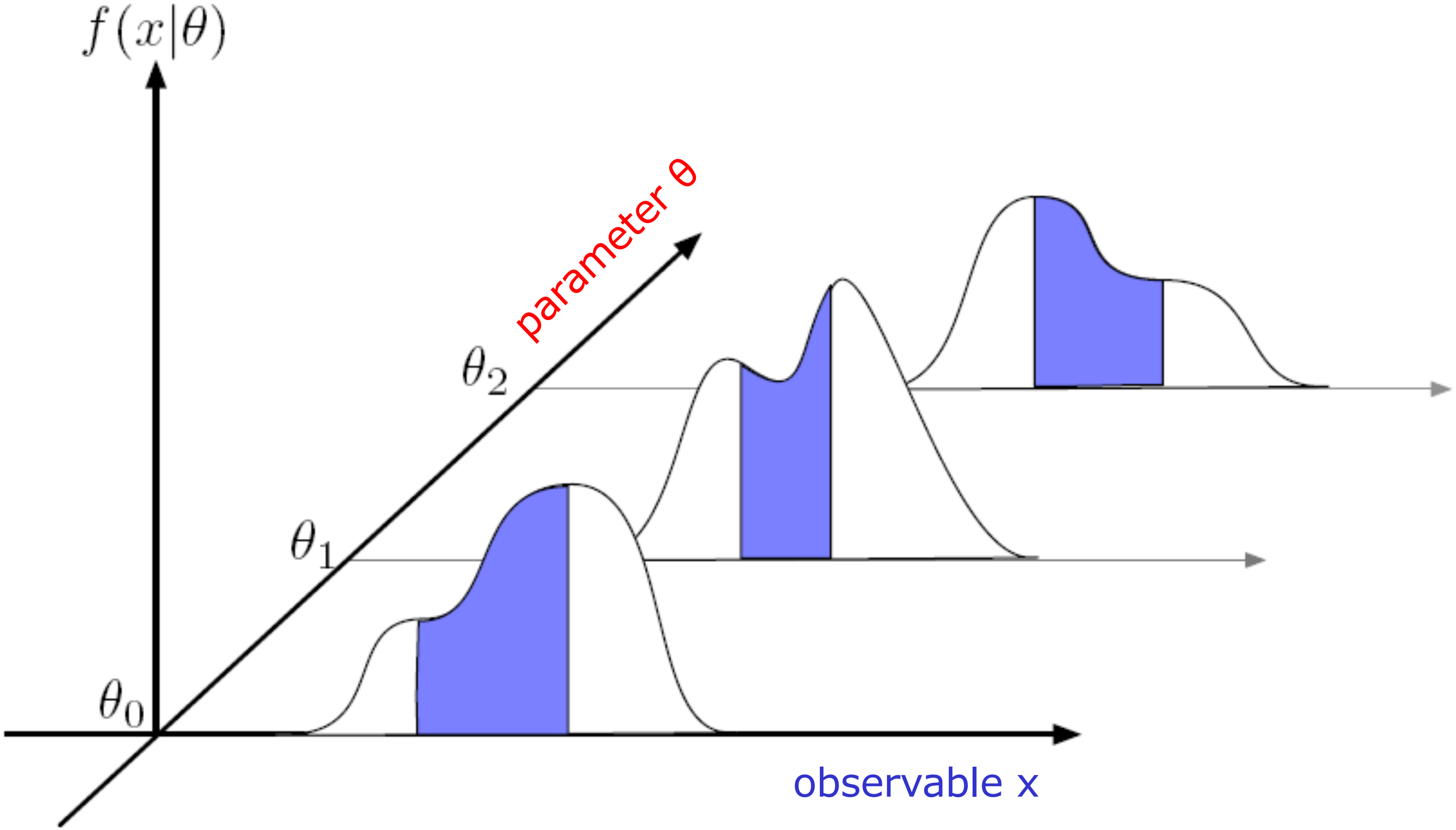
Other options, are e.g. 'symmetric' and 'shortest'

observable x

Kerke, NIKHEF

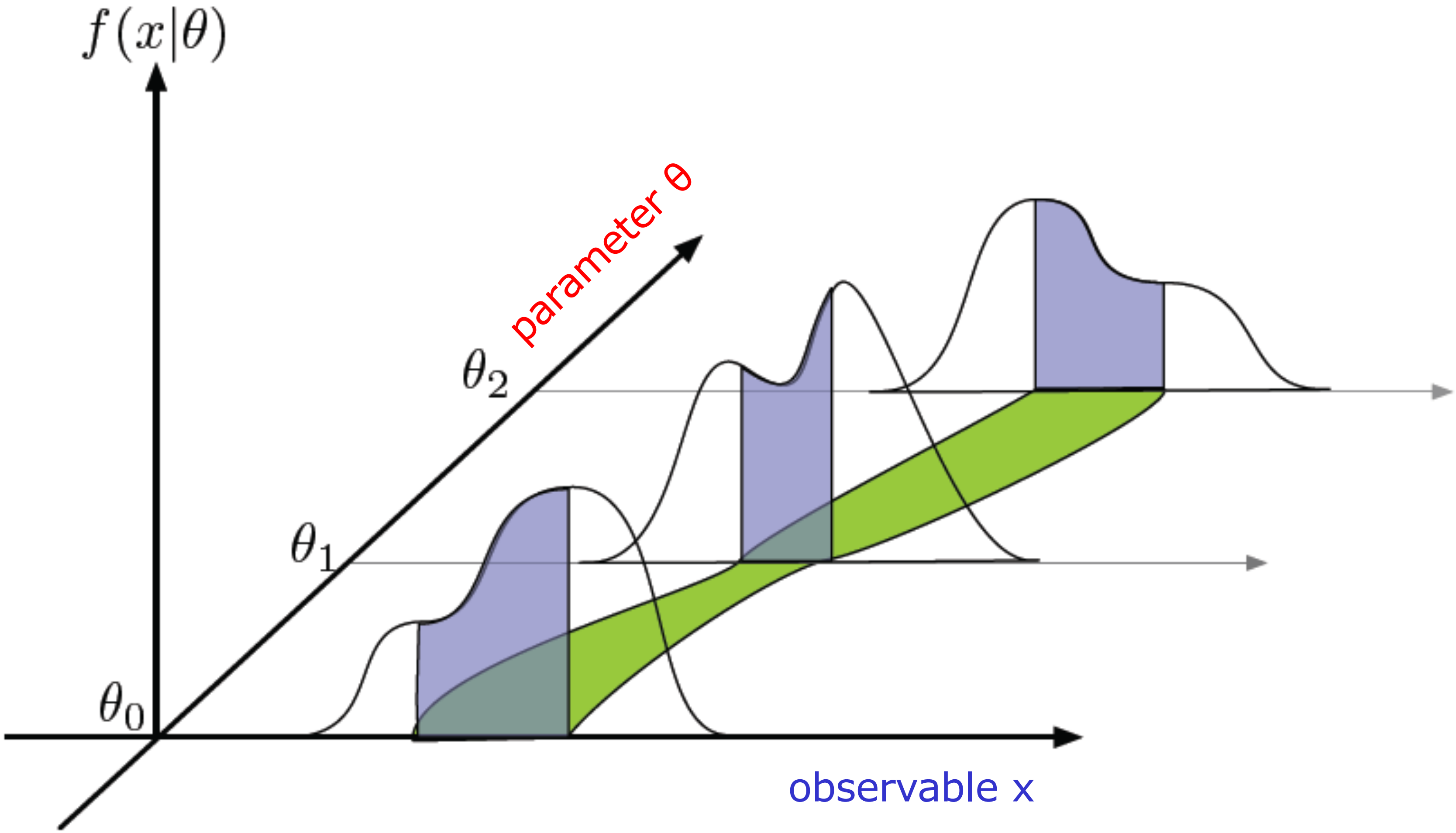
How to construct a Neyman Confidence Interval

- Now make an acceptance interval in observable x for each value of parameter θ



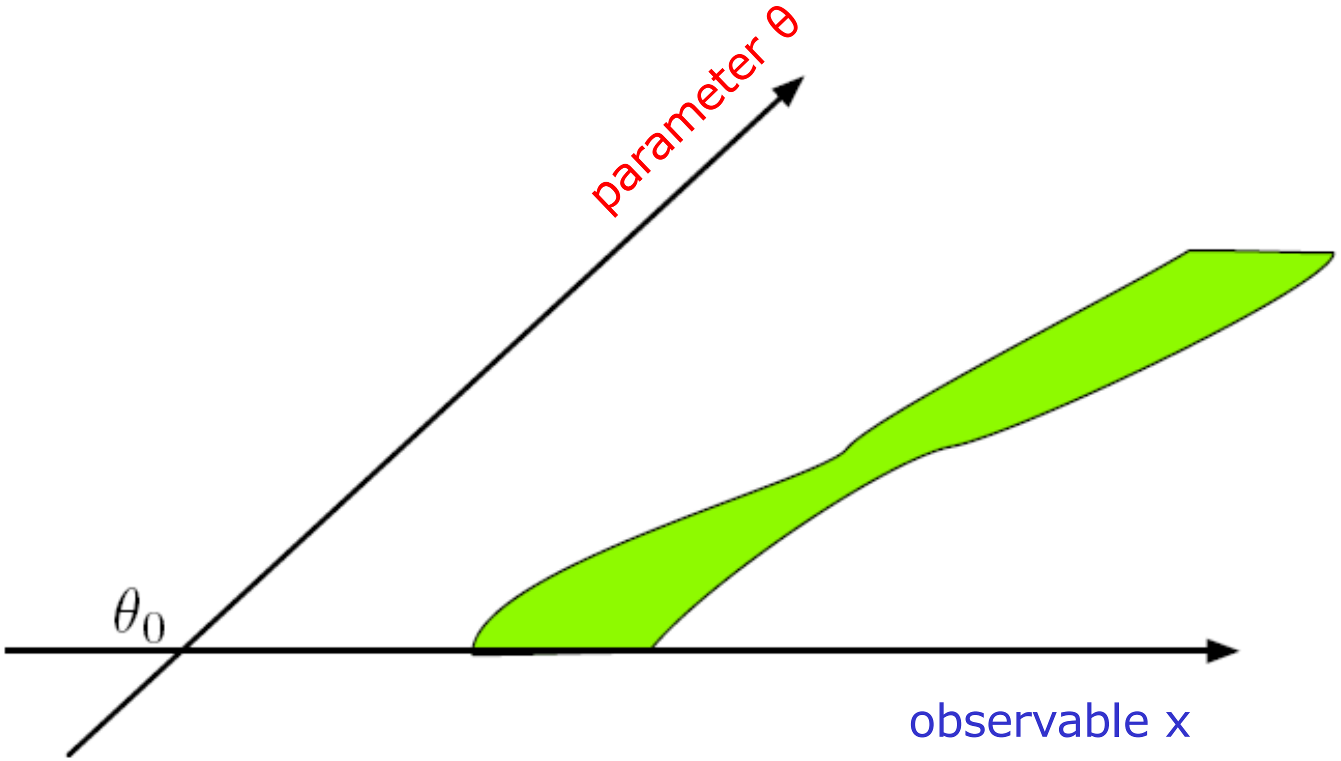
How to construct a Neyman Confidence Interval

- This makes the confidence belt



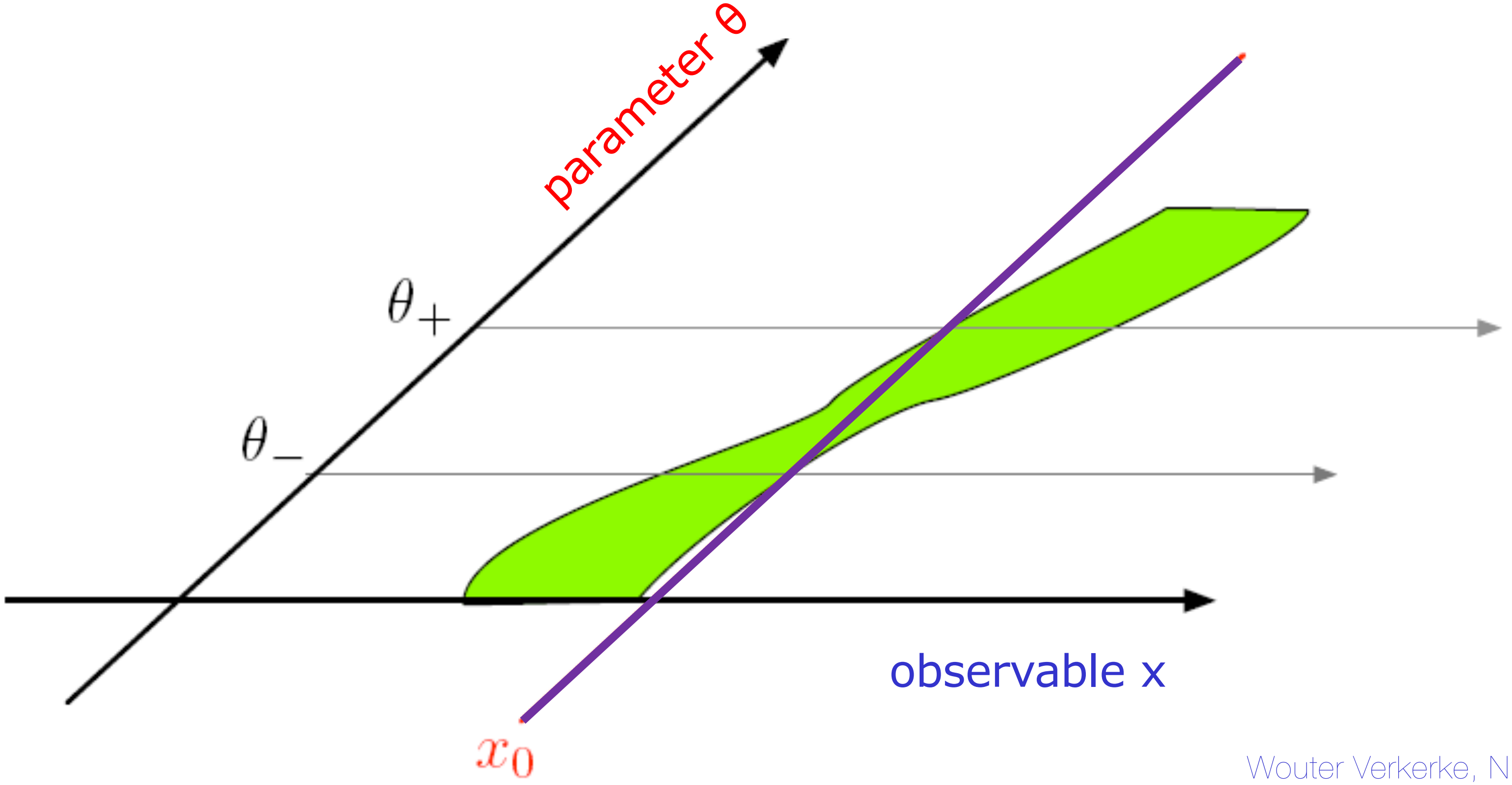
How to construct a Neyman Confidence Interval

- This makes the confidence belt



How to construct a Neyman Confidence Interval

- The confidence belt can constructed *in advance of any measurement*, it is a property of the model, not the data
- Given a measurement x_0, a confidence interval $[\theta_+,\theta_-]$ can be constructed as follows
- The interval $[\theta_-,\theta_+]$ has a 68% probability to cover the true value



Wouter Verkerke, NIKHEF

What confidence interval means & concept of coverage

- A confidence interval is an interval on a parameter that contains the true value X% of the time
- This is a property of the procedure, and should be interpreted in the concept of repeated identical measurements:

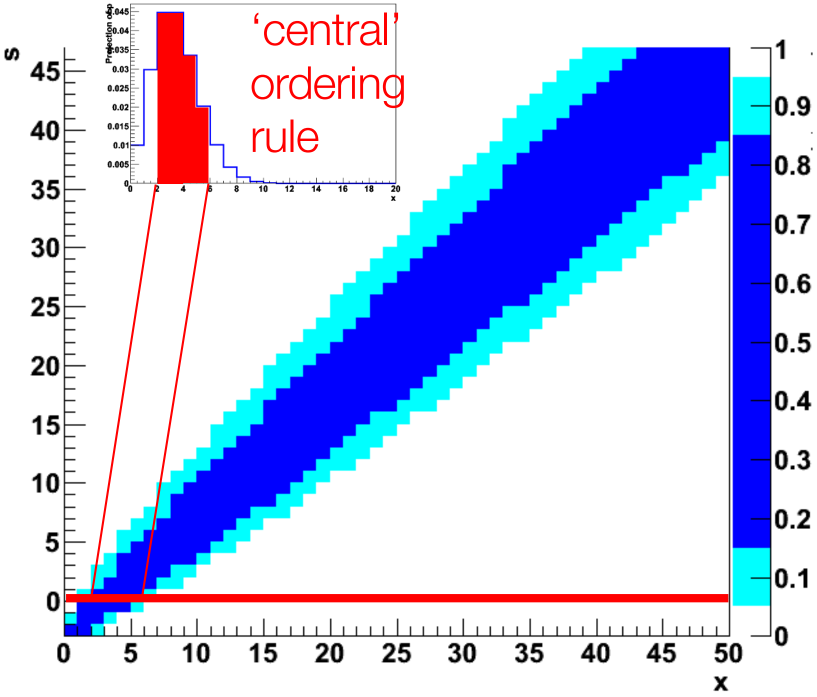
  Each future measurement will result a confidence interval that has somewhat different limits every time
  (*‘confidence interval limits are a random variable’)*

  But procedure is constructed such that true value is in X% of the intervals in a series of repeated measurements
  (*this calibration concept is called ‘coverage*’. The Neyman constructions guarantees coverage)
- **It is explicitly not a probability statement on the true value** *you are trying to measure. In the frequentist the true value is fixed (but unknown)*

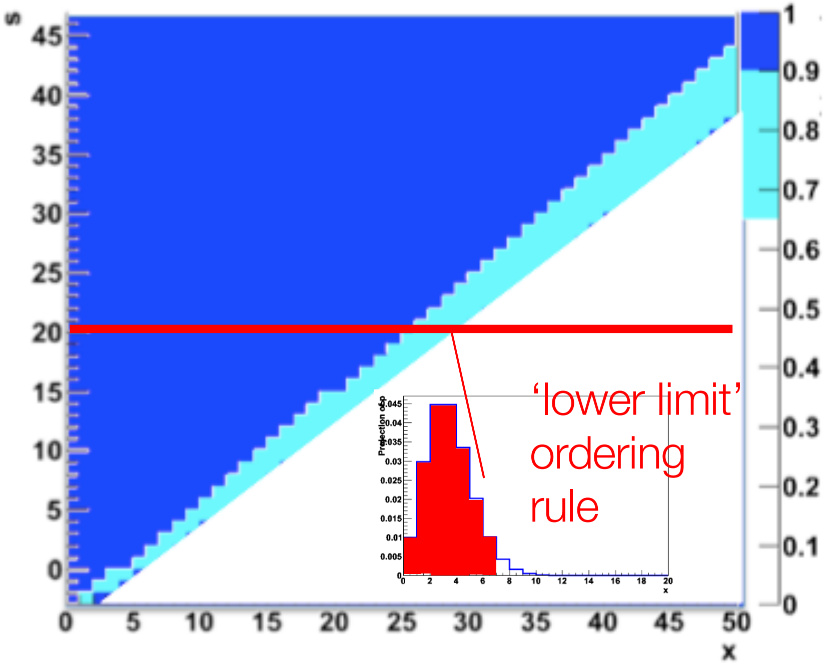
The confidence interval – Poisson counting example

- Given the probability model for Poisson counting example: for every hypothesized value of *s*, plot the expected distribution *N*

Confidence belt for
68% and 90% central intervals



Confidence belt for
68% and 90% lower limit

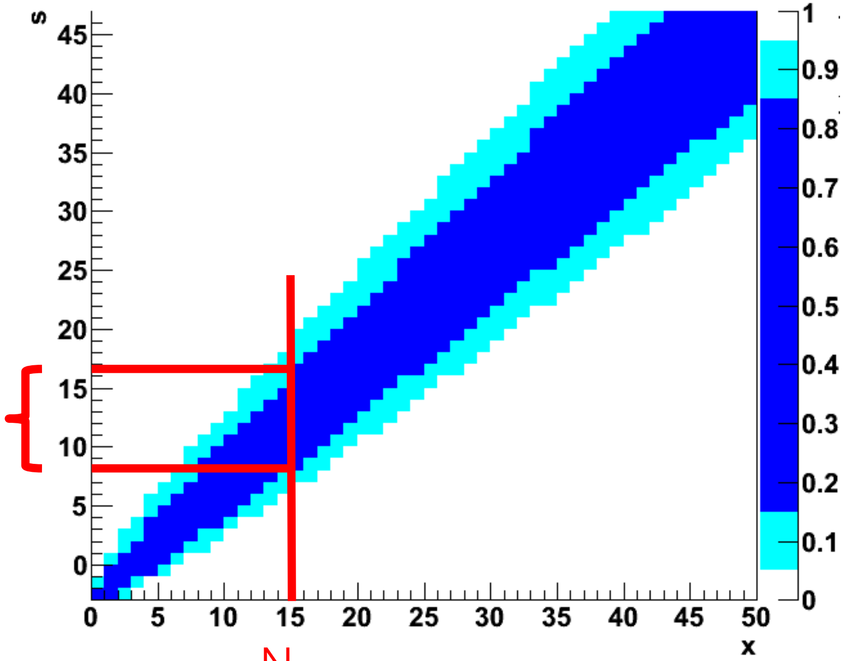


Wouter Verkerke, NIKHEF

The confidence interval – Poisson counting example

- Given confidence belt and observed data, confidence interval on parameter is defined by belt intersection

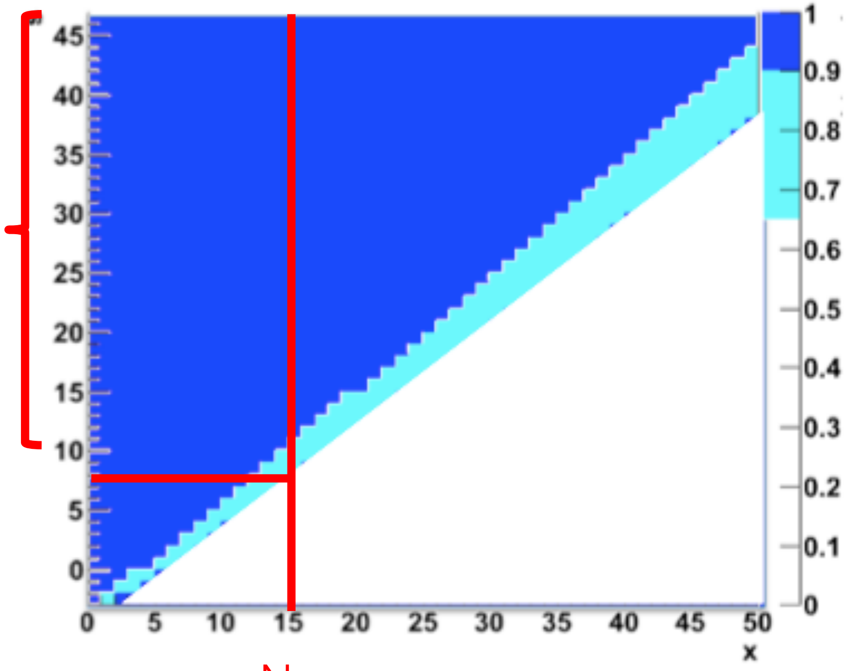
Confidence belt for
68% and 90% central intervals



N_{obs}

Central interval on s at 68% C.L.

Confidence belt for
68% and 90% lower limit



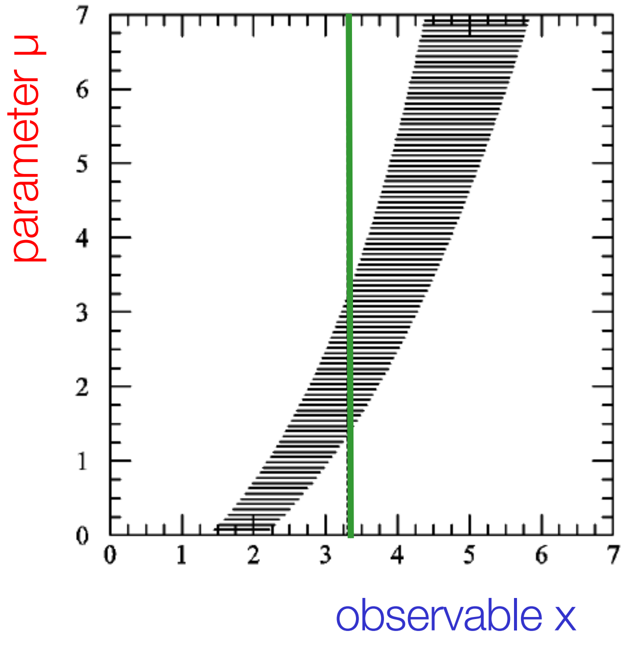
N_{obs}

Lower limit on s at 90% C.L.

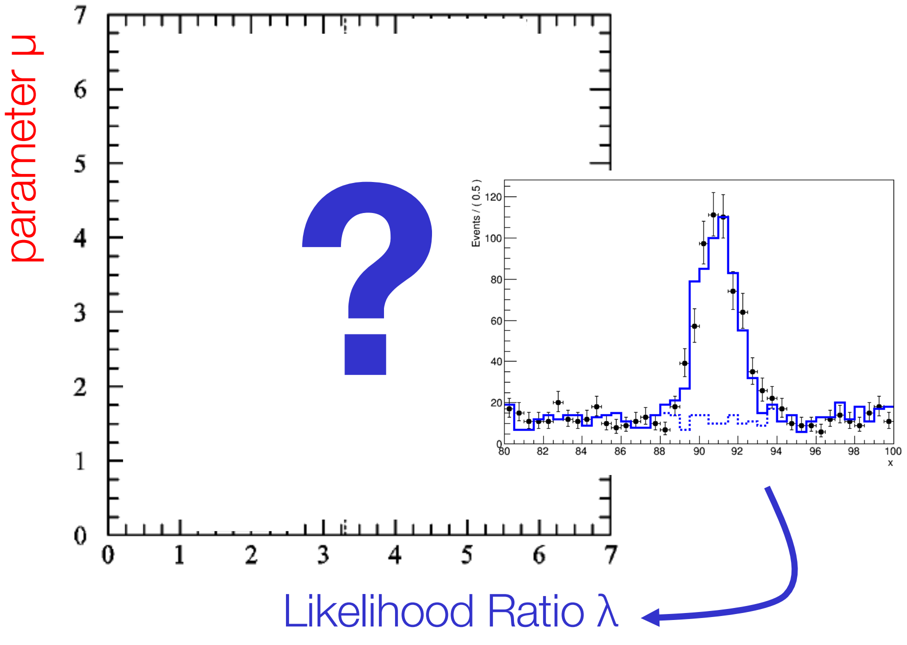
Confidence intervals using the Likelihood Ratio test statistic

- Neyman Construction on Poisson counting looks like ‘textbook’ belt.
- In practice we’ll use the Likelihood Ratio test statistic to summarize the measurement of a (multivariate) distribution for the purpose of hypothesis testing.
- Procedure to construct belt with LR is identical:
  obtain distribution of λ for every value of μ to construct confidence belt

x=3.2



$\lambda_\mu(x,\mu)$



The asymptotic distribution of the likelihood ratio test statistic

- Given the likelihood ratio

$$t_\mu = -2\log\lambda_\mu(x) = -2\log\frac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})}$$

  Q: What do we know about asymptotic distribution of $\lambda(\mu)$?

- A: Wilks theorem → Asymptotic form of $f(t|\mu)$ is a χ^2 distribution

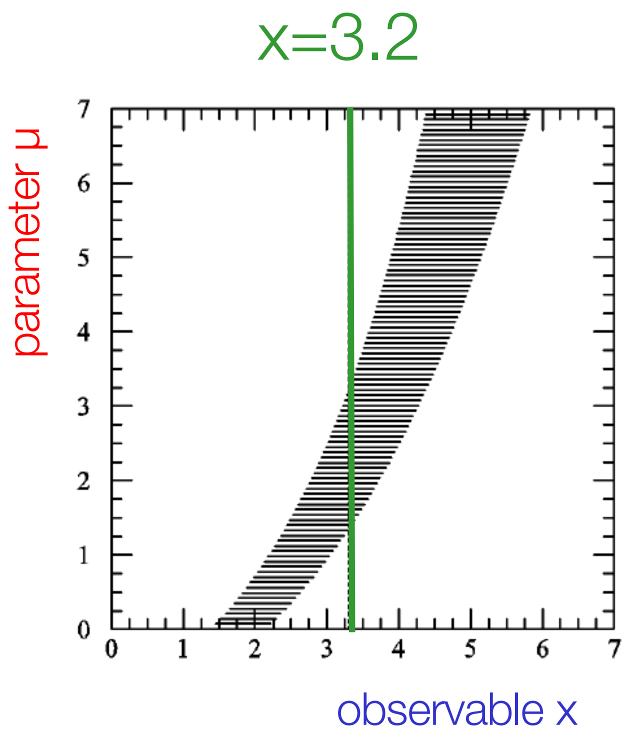
$$f(t_\mu|\mu) = \chi^2(t_\mu,n)$$

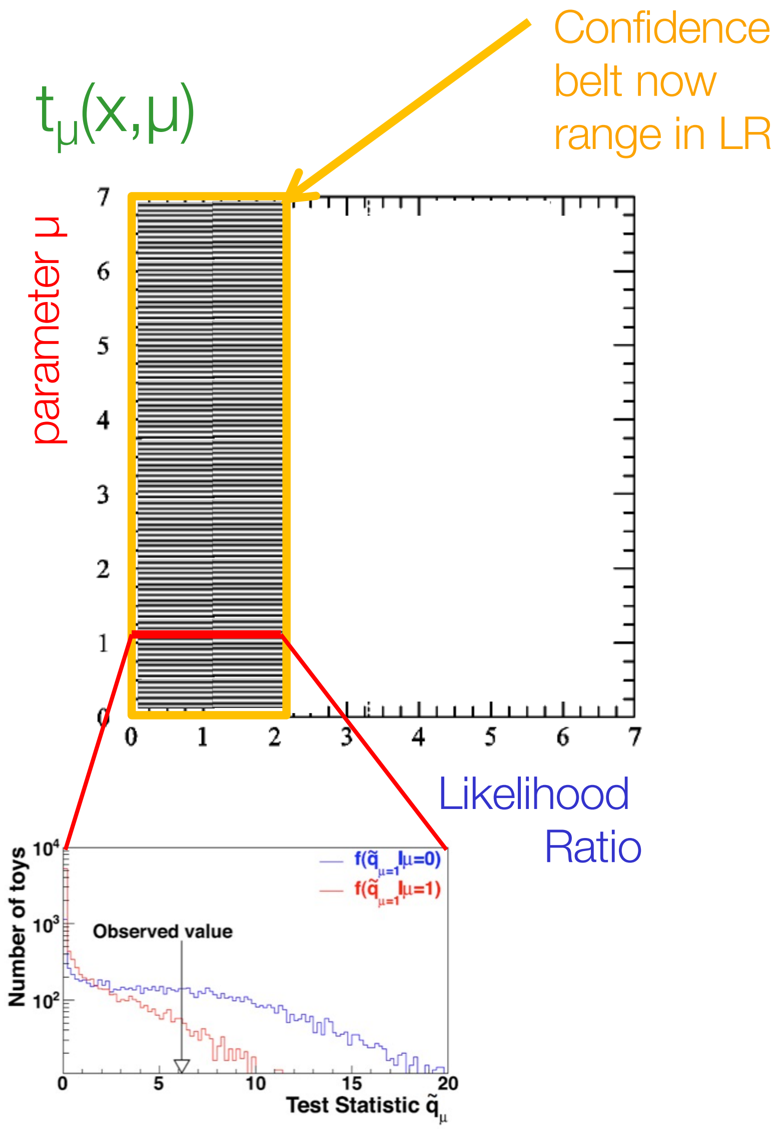
  Where
  μ is the hypothesis being tested and
  n is the number of parameters (here 1: μ)

- **Note that $f(t_\mu|\mu)$ is independent of μ!**
  → Distribution of t_μ is the *same* for every ‘horizontal slice’ of the belt

Confidence intervals using the Likelihood Ratio test statistic

- Procedure to construct belt with LR is identical:
  obtain distribution of λ for every value of μ to construct belt



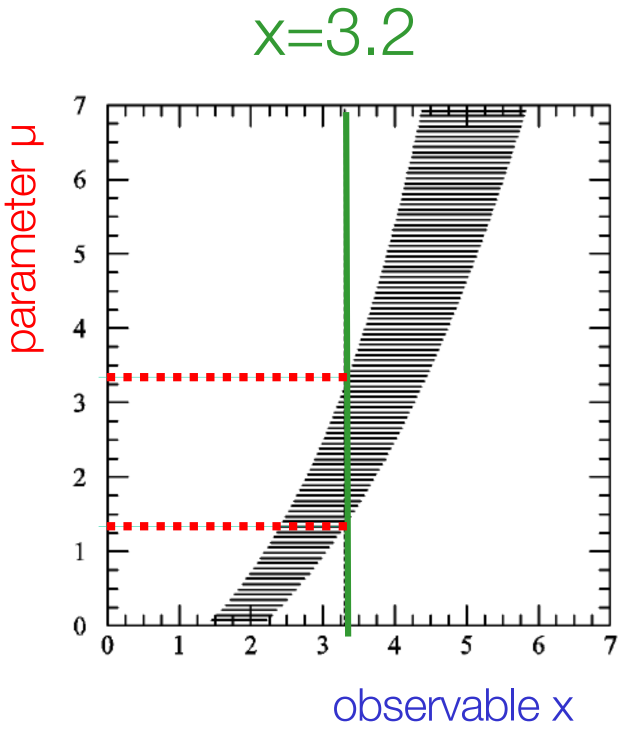


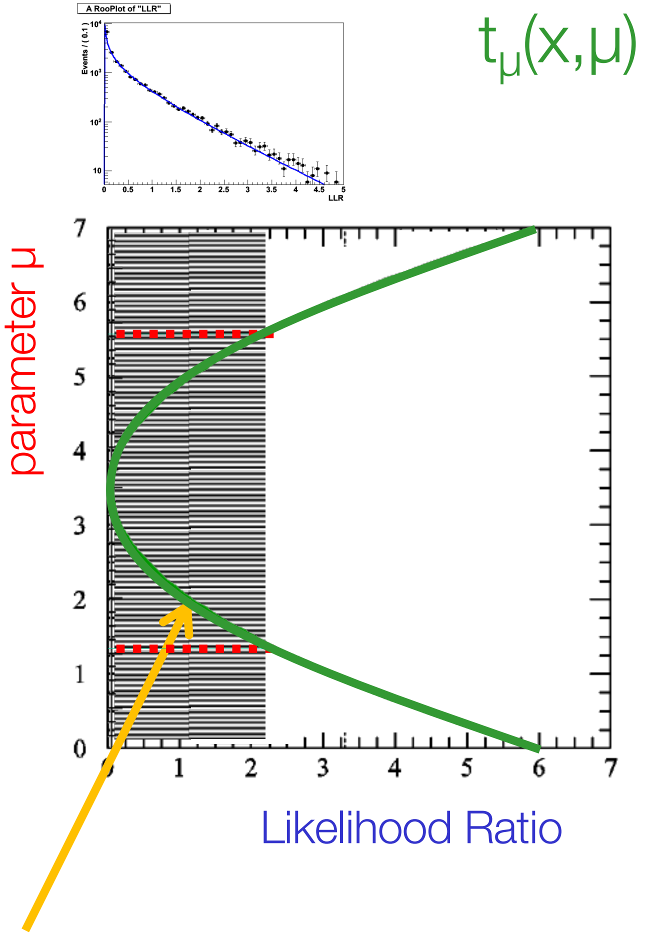
Confidence belt now range in LR

Likelihood Ratio

What does the observed data look like with a LR?

- Note that while belt is (asymptotically) independent of parameter μ, observed quantity now is dependent of the assumed μ





Measurement = $t_\mu(x_{obs},\mu)$
is now a function of μ

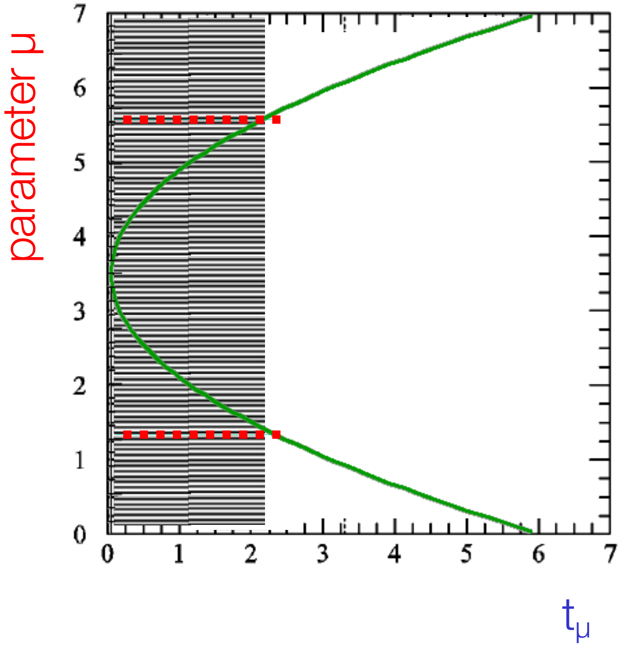
‘ex10_roostats_plr_interval.C’

Connection with likelihood ratio intervals

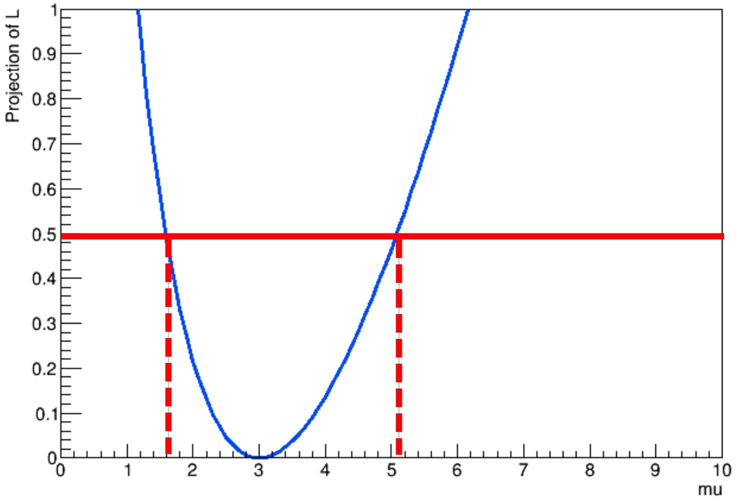
- If you assume the asymptotic distribution for t_μ,
  - Then the confidence belt is exactly a box
  - And the constructed confidence interval can be simplified to finding the range in μ where $t_\mu=½\cdot Z^2$

→ **This is exactly the MINOS error**

FC interval with Wilks Theorem

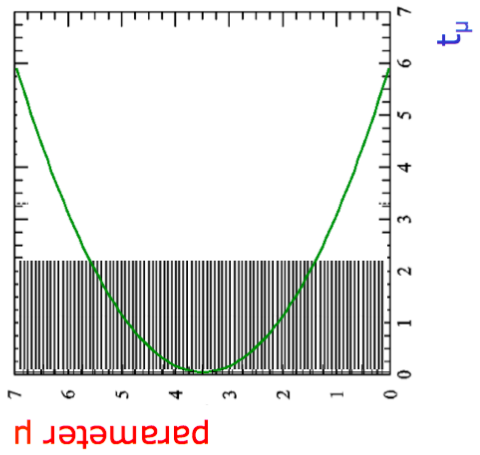


MINOS / Likelihood ratio interval



Recap on confidence intervals

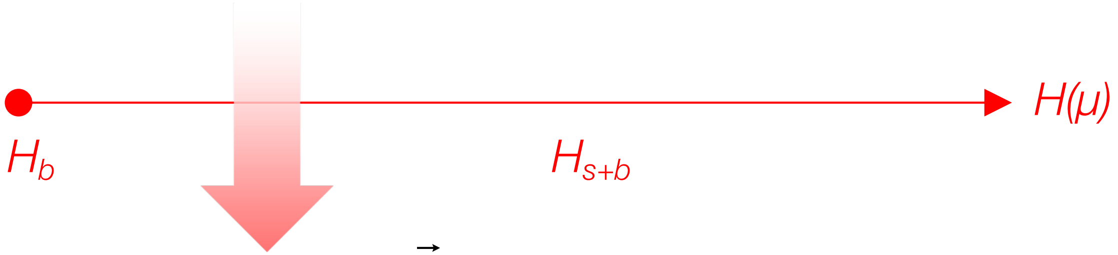
- Confidence intervals on parameters are constructed to have precisely defined probabilistic meaning
  - This calibration is called "coverage"
    The Neyman Construction has coverage by construction
  - This is different from parameter variance estimates (or Bayesian methods) that don't have (a guaranteed) coverage
  - For most realistic models confidence intervals are calculated using (Likelihood Ratio) test statistics to define the confidence belt
- Asymptotic properties
  - In the asymptotic limit (Wilks theorem), Likelihood Ratio interval converges to a Neyman Construction interval (with guaranteed coverage) "Minos Error"
    *NB: the likelihood does **not** need to be parabolic for Wilks theorem to hold*
  - Separately, in the limit of normal distributions the likelihood becomes exactly parabolic and the ML Variance estimate converges to the Likelihood Ratio interval



Bayesian inference with composite hypothesis

- With change L→L(μ) the prior and posterior model probabilities become probability density functions

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



$$P(\mu\,|\,\vec{N}) = \frac{L(\vec{N}\,|\,\mu)P(\mu)}{\int L(\vec{N}\,|\,\mu)P(\mu)\,d\mu}$$

Posterior probability *density*

Prior probability *density*

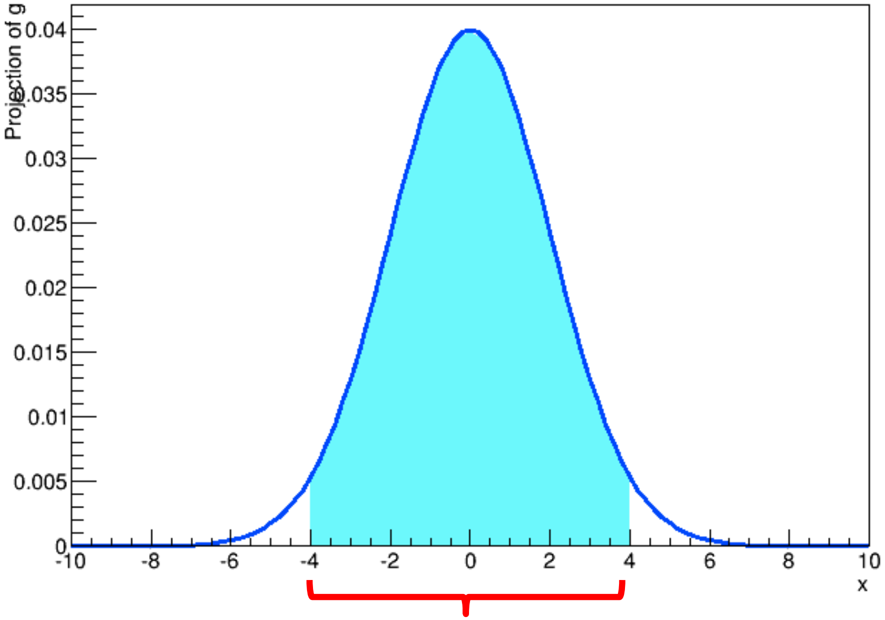
$$P(\mu\,|\,\vec{N}) \propto L(\vec{N}\,|\,\mu)P(\mu)$$

NB: Likelihood is not a probability density ‘ex10_roostats_bayes_interval.C’

Bayesian credible intervals

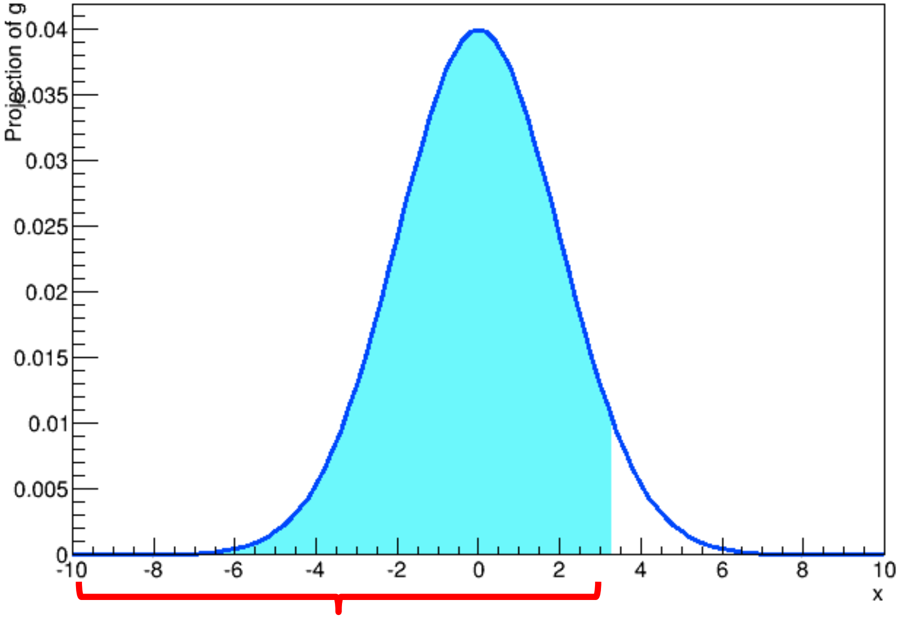
- From the posterior density function, a credible interval can be constructed through integration

Posterior on μ



95% credible central interval

Posterior on μ

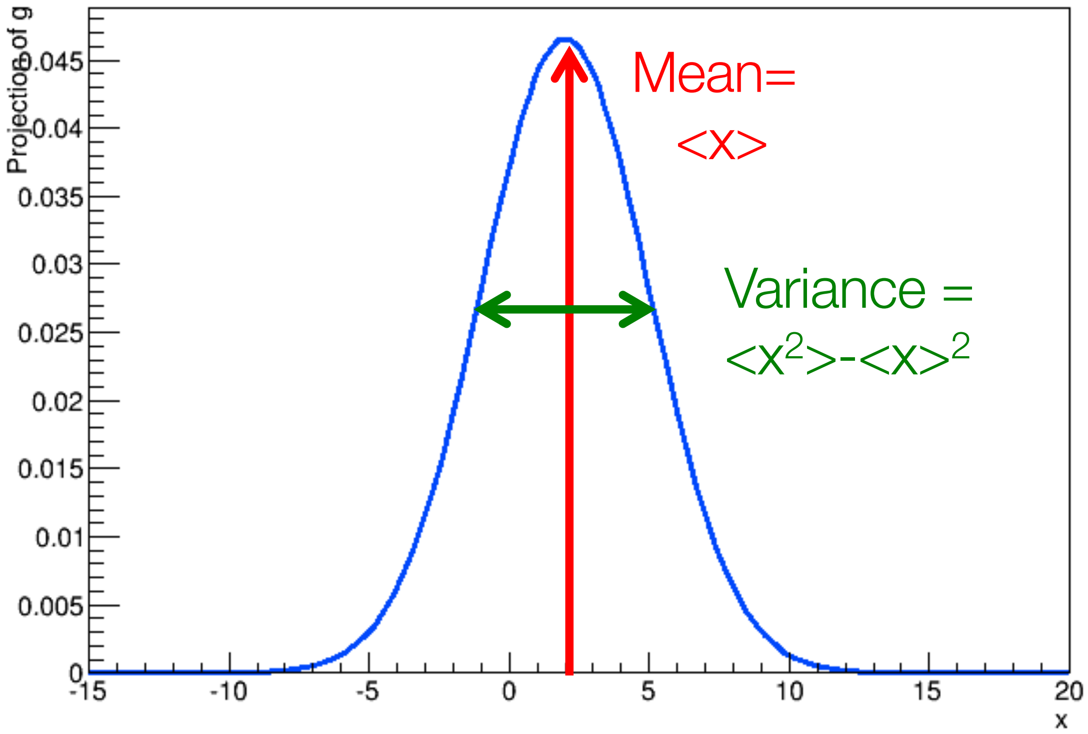


95% credible upper limit

- Note that Bayesian interval estimation require *no minimization* of –logL, just integration

Bayesian parameter estimation

- Bayesian parameter estimate is the posterior mean
- Bayesian variance is the posterior variance



Mean= $\langle x \rangle$

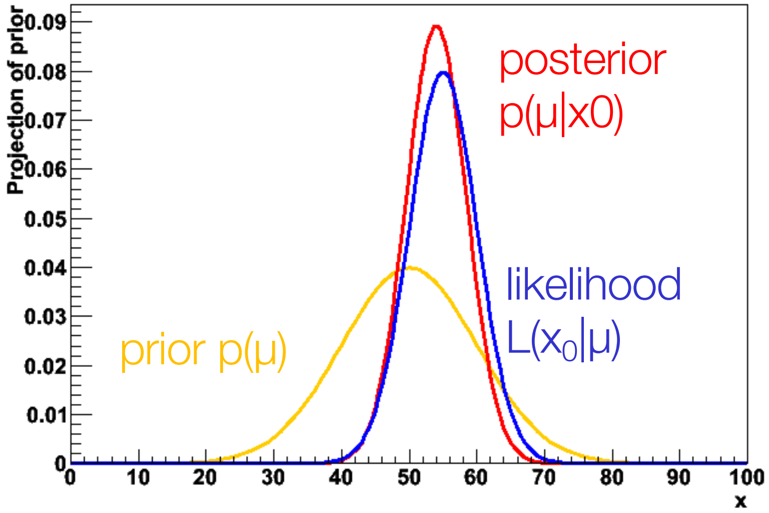
Variance = $\langle x^2 \rangle - \langle x \rangle^2$

$$\hat{\mu} = \int \mu P(\mu \mid N) d\mu$$

$$\hat{V} = \int (\hat{\mu} - \mu)^2 P(\mu \mid N) d\mu$$

Choosing Priors

- As for simple models, Bayesian inference always in involves a prior
  → now a prior probability density on your parameter
- When there *is* clear prior knowledge, it is usually straightforward to express that knowledge as prior density function
  - Example: prior measurement of μ = 50 ± 10



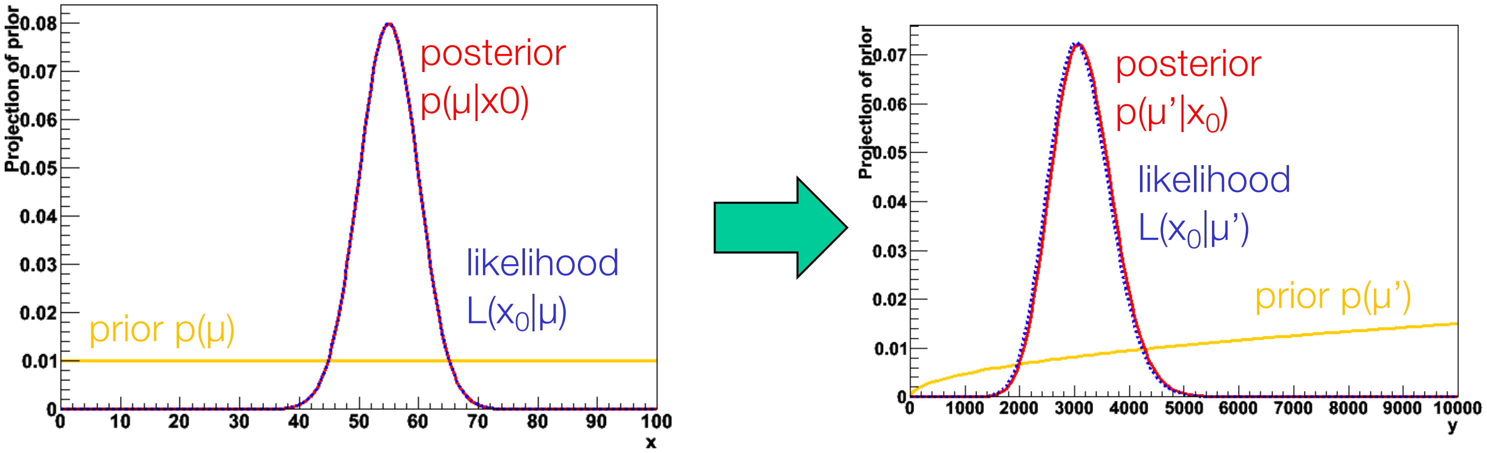
  - **Posterior represents updated belief →** It incorporates information from measurement *and* prior belief
  - But sometimes we only want to publish result of *this* experiment, or there is no prior information. What to do?

Choosing Priors

- Common but thoughtless choice: a flat prior
  - Flat implies choice of metric. Flat in x, is not flat in x^2

distribution in μ

distribution in μ^2



- Flat prior implies choice on of metric
  - A prior that is flat in μ is not flat in μ^2
  - **‘Preferred metric’ has often no clear-cut answer.** (E.g. when measuring neutrino-mass-squared, state answer in m or m^2)
  - **In multiple dimensions even complicated** (prior flat in x,y or is prior flat in r,ϕ?)

Is it possible to formulate an ‘objective’ prior?

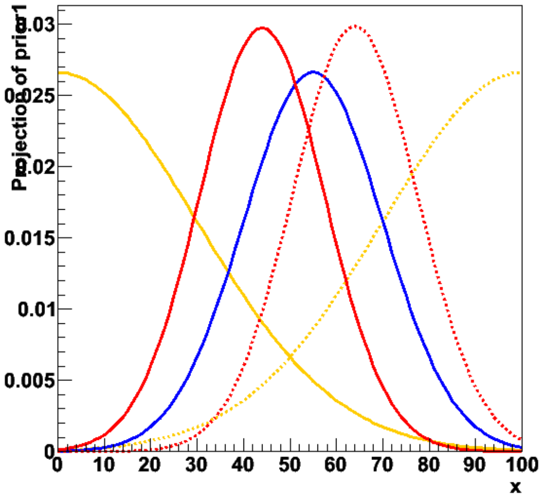
- *Can one define a prior p(μ) which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?*
  - A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20thcentury:
  - This is a really *really* thoughtless idea, recognized by Jeffreys as such, but dismayingly common in HEP: just choose p(μ) uniform in whatever metric you happen to be using!
- “Jeffreys Prior” answers the question using a prior uniform in a metric related to the Fisher information.

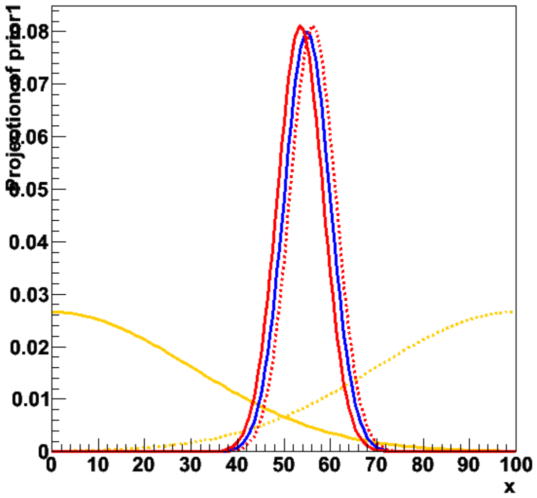
$$I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(x \mid \theta) \middle| \theta \right]$$

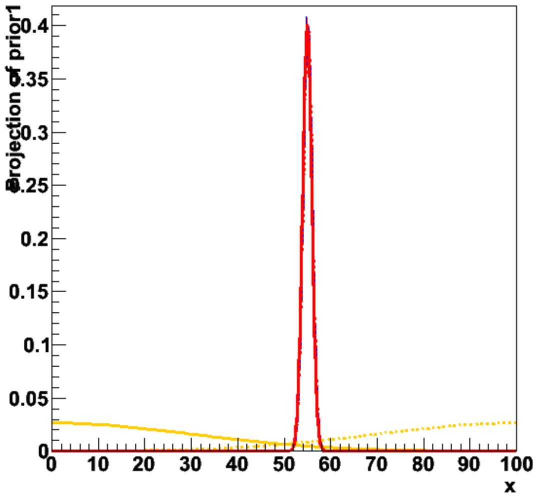
  - Unbounded mean μ of gaussian: p(μ) = 1
  - Poisson signal mean μ, no background: p(μ) = 1/√μ
- Many ideas and names around on non-subjective priors
  - Advanced subject well beyond scope of this course.
  - Many ideas (see e.g. summary by Kass & Wasserman), but very much an open/active in area of research

Sensitivity Analysis

- Since a Bayesian result depends on the prior probabilities, which are either personalistic or with elements of arbitrariness, it is widely recommended by Bayesian statisticians to study the sensitivity of the result to varying the prior.
- Sensitivity generally decreases with precision of experiment



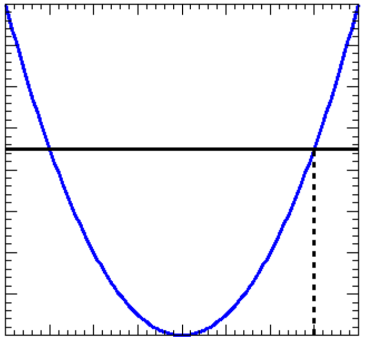




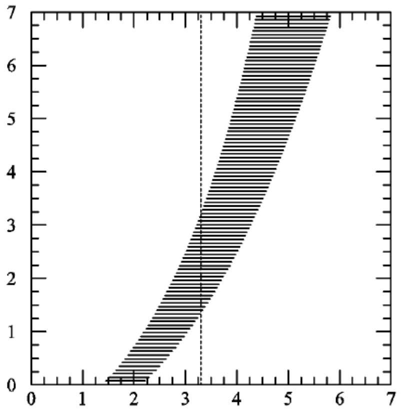
- Some level of arbitrariness – what variations to consider in sensitivity analysis

Summary

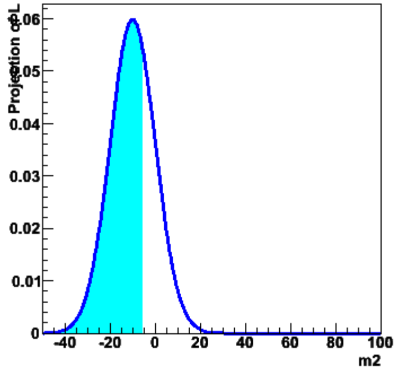
- Maximum Likelihood
  - Point and variance estimation
  - Variance estimate assumes normal distribution. No upper/lower limits



- Frequentist confidence intervals
  - Extend hypothesis testing to composite hypothesis
  - Neyman construction provides exact “coverage” = calibration of quoted probabilities
  - Strictly p(data|theory)
  - Asymptotically identical to likelihood ratio intervals (MINOS errors, *does not assume parabolic L*)



- Bayesian credible intervals
  - Extend P(theo) to p.d.f. in model parameters
  - Integrals over posterior density → credible intervals
  - Always involves prior density function in parameter space


Projection of L
m2