Statistics

W. Verkerke

Schedule

Time	Monday, 5 December 2022
09:30	**Lecture 1** - Wouter Verkerke (Nikhef)
10:15	Coffee
10:30	**Lecture 2** - Wouter Verkerke (Nikhef)
11:15	Coffee
11:30	**Lecture 3** - Wouter Verkerke (Nikhef)
12:30	Lunch
14:00	**Introduction to hands-on session 1**
14:10	**Hands-on 1**
16:30	**Close-out / Discussion of exercises**

Time	Tuesday, 6 December 2022
09:30	**Lecture 4** - Wouter Verkerke (Nikhef)
10:15	Coffee
10:30	**Lecture 5** - Wouter Verkerke (Nikhef)
11:15	Coffee
11:30	**Lecture 6** - Wouter Verkerke (Nikhef)
12:30	Lunch
14:00	**Introduction to Hands-on 2** - Wouter
14:10	**Hands-on 2**
16:30	**Closeout / Discussion of exercises**

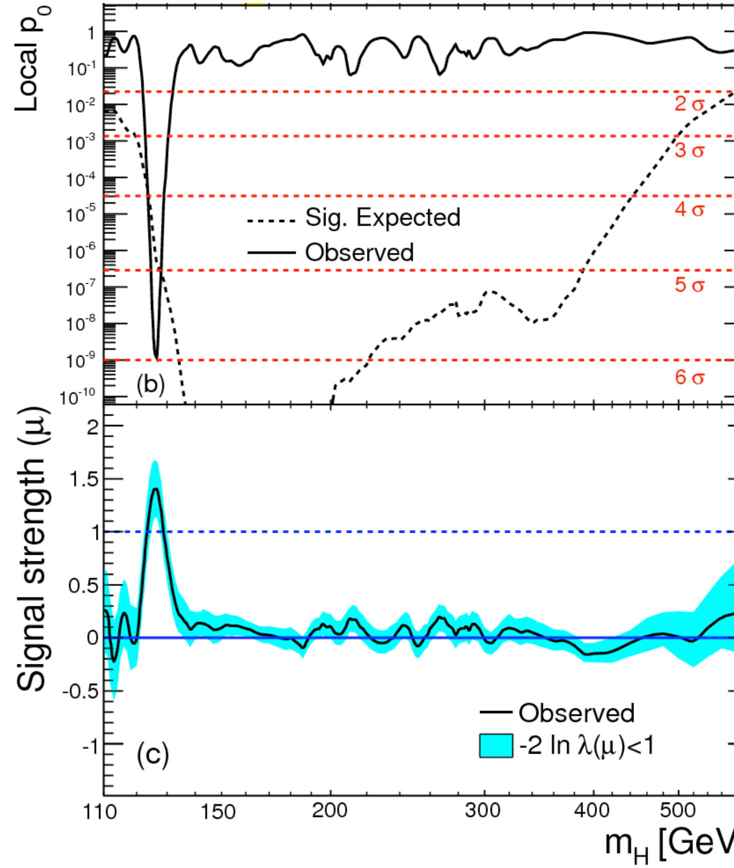
Time	Wednesday, 7 December 2022
09:30	**Lecture 7** - Wouter Verkerke (Nikhef)
10:15	Coffee
10:30	**Lecture 8** - Wouter Verkerke (Nikhef)
11:15	Coffee
11:30	**Guest Lecture** - Max Baak (ING)
12:30	Lunch
14:00	**Hands-on 3**
15:30	**Closeout** - Wouter Verkerke (Nikhef)

Statistics & Modeling

- Statistics → formalism to quantify what you learn about a theory from your data
  - Largely abstract and mathematical in nature
- Modeling → how to write a theory that predicts your specific observed distribution in your experiment
  - I.e how does the SM translate to your 3-jet invariant mass distribution observed in your detector, including all known (systematic) uncertainties
  - Very practical in nature, often not an 'exact science', based on judgements calls. Little text book knowledge on it – but often the central part of your statistical analysis
- Both equally important – both are vast topics
  - Given time available will focus mostly in issues that arise in 'event-based' particle physics (which includes anything ranging from LHC to Neutrino Physics, Dark Matter searches)
  - Physics examples are largely based on LHC physics – but no specific assumption on LHC physics (knowledge) are made

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

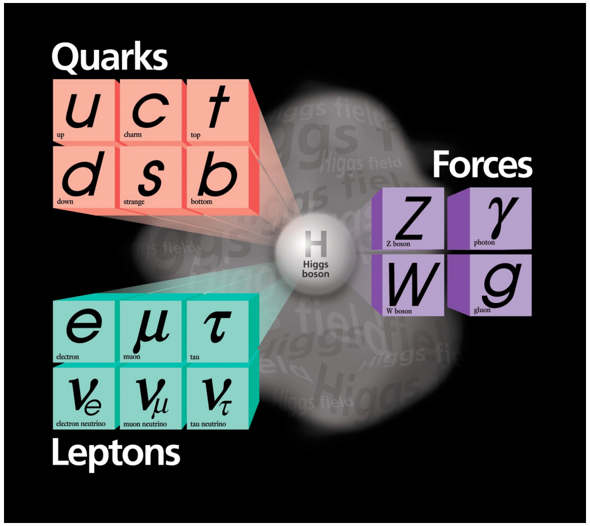
- **Result: *Decision* based on tests**

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

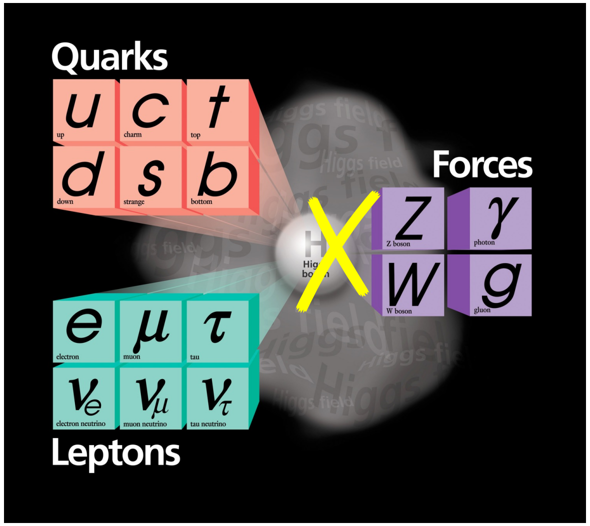
How do we do this?

- **All experimental results start with formulation of a (physics) theory**
- Examples of HEP **physics** models being tested

The Standard Model

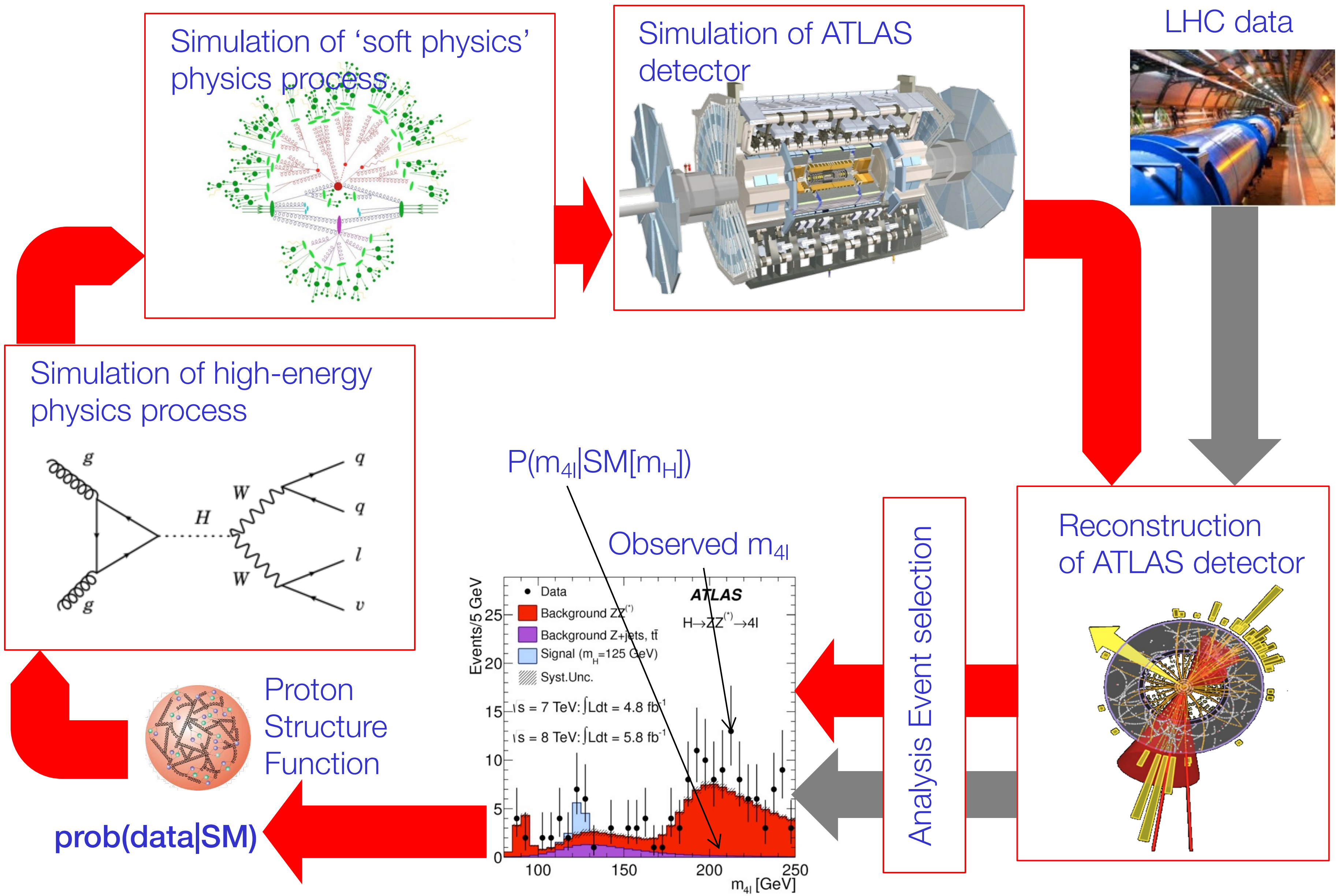


The SM without a Higgs boson



- **Next, you design a measurement to be able to *test* model**
  – Via chain of physics simulation, showering MC, detector simulation and analysis software, a physics model is reduced to a **statistical** model

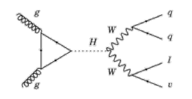
An overview of HEP data analysis procedures



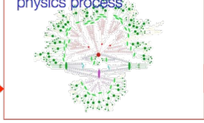
HEP workflow: data analysis in practice

An overview of HEP data analysis procedures

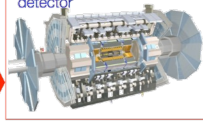
Simulation of high-energy physics process



Simulation of 'soft physics' process



Simulation of ATLAS detector



LHC data

MC Simulated Events (sig,bkg)

All available "real data"

N-tuples
Cut-flows,
Multi-variate analysis (NN,BDT)
ROOT, TMVA, NeuroBayes

Helps to define selection

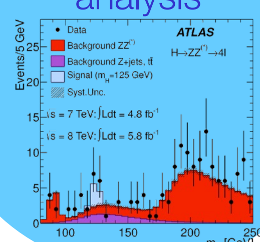
Event selection (cuts, NN, BDT)

Final Event Selection (MC)

Final Event Selection (data)

Signal, background models
Likelihood models,
MINUIT, RooFit
RooStats, MCLimit

Statistical analysis

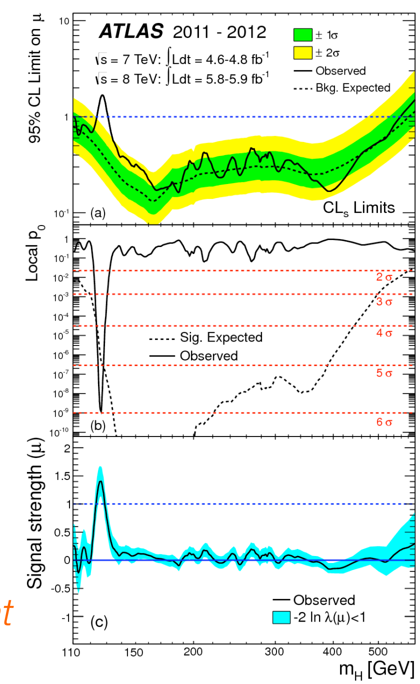


Final Result

Limit

Discovery

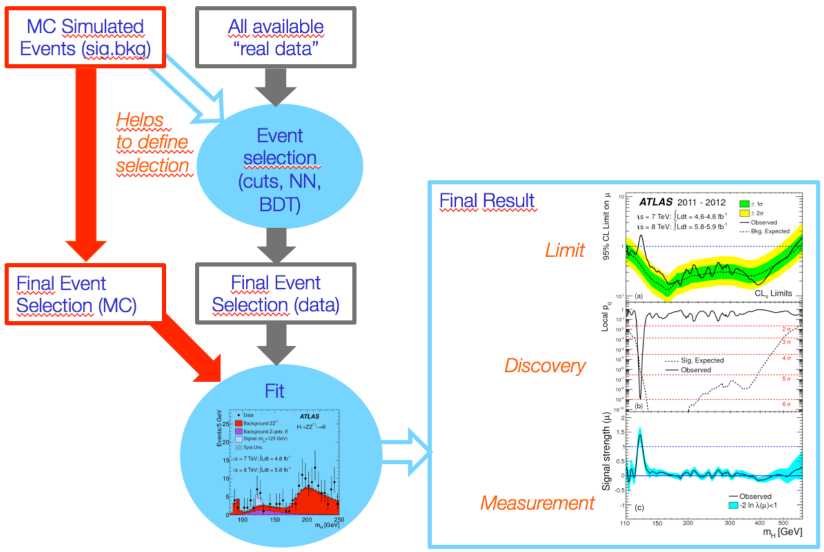
Measurement



From physics theory to statistical model

- HEP “Data Analysis” is for large part
the reduction of a physics theory to a statistical model

Physics Theory: Standard Model with 125 GeV Higgs boson



Statistical Model: *Given a measurement x (e.g. an event count)*
what is the probability to observe each possible value of x,
under the hypothesis that the physics theory is true.

Once you have a statistical model, all physics knowledge has been abstracted into the model, and further steps in statistical inference are ‘procedural’ (no physics knowledge is required in principle)

From statistical model to a result

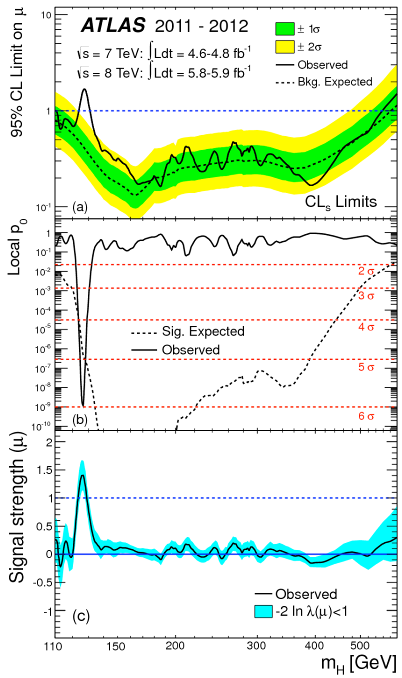
- The next step of the analysis is to confront your model with the data, and summarize the result in a probabilistic statement of some form

Final Result

Limit

Discovery

Measurement



'Confidence/Credible Interval'

$\sigma/\sigma_{SM}\,(H \rightarrow ZZ)\,|_{mH=150} < 0.3$ **@ 95% C.L.**

'p-value'

"Probability to observed this signal or more extreme, under the hypothesis of background-only is 1x10^{9}"

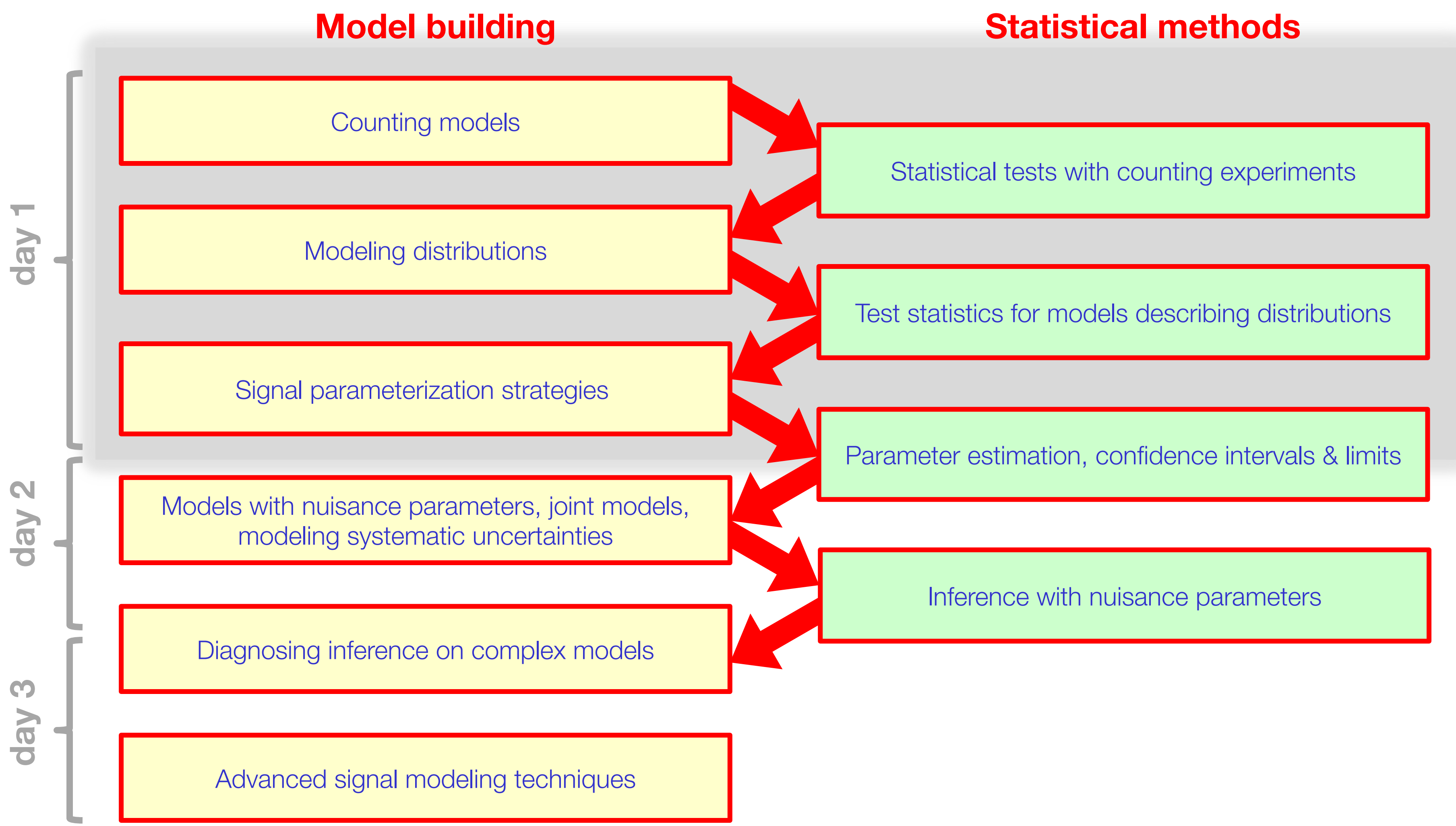
'Measurement with variance estimate'

$\sigma/\sigma_{SM}\,(H \rightarrow ZZ)\,|_{mH=126} = 1.4 \pm 0.3$

- The last step, usually not in a (first) paper, that you, or your collaboration, *decides* if your theory is valid

Roadmap of this course

- Start with basics, gradually build up to complexity



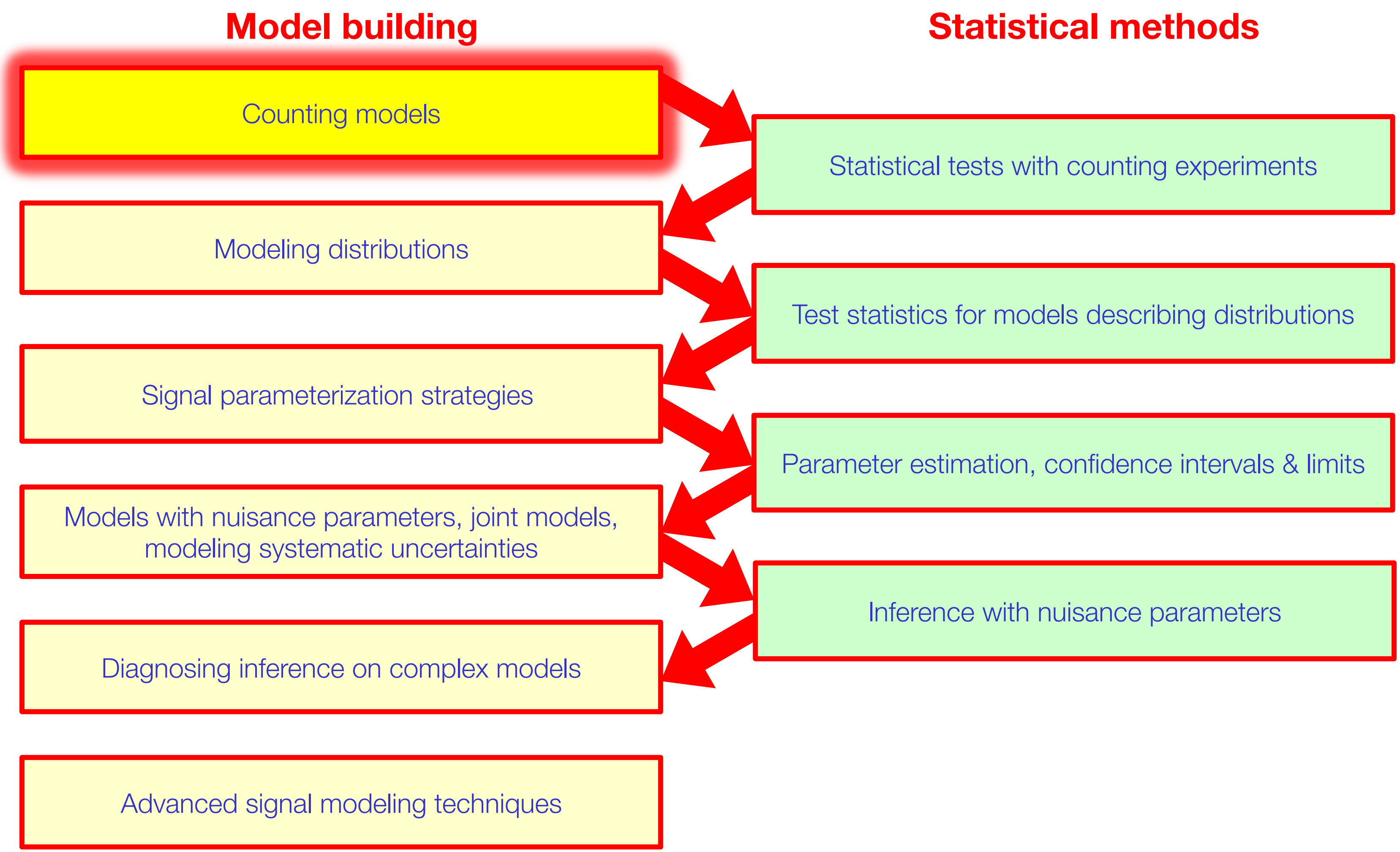
Model building 1

Basic distributions: Binomial, Poisson, Gaussian

Wouter Verkerke, NIKHEF

Roadmap of this course

- Start with basics, gradually build up to complexity



The statistical world

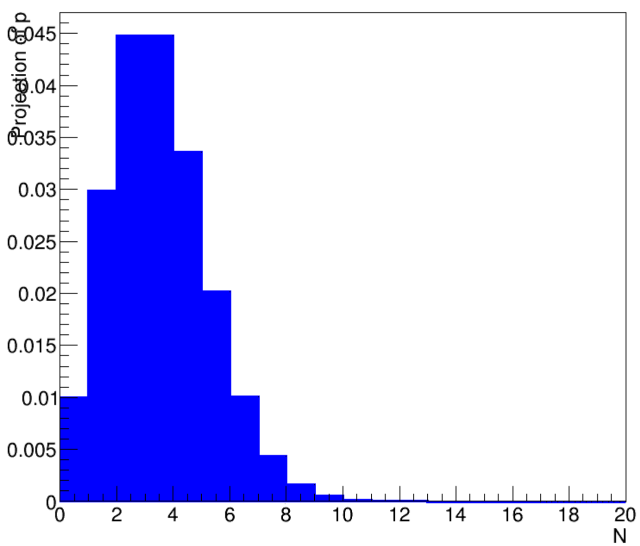
- Central concept in statistics is the '**probability model**'
- *A probability model assigns a probability to each possible experimental outcome*.
- Example: a HEP counting experiment

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

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

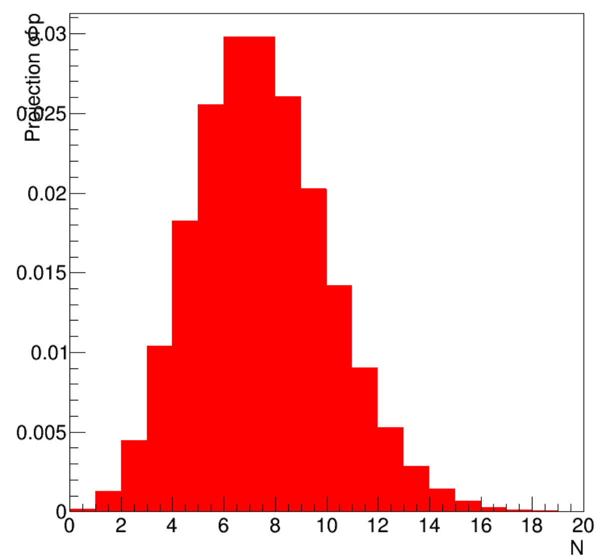
μ=3 ("bkg only")

↑ Probability of outcome



→ Experimental outcome

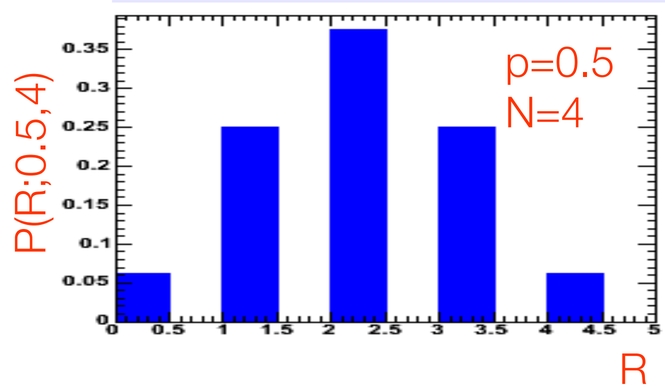
μ=7 ("bkg+signal")



Intermezzo on distributions – The binomial distribution

- Simple **counting** experiment – Drawing marbles from a bowl
  - Bowl with marbles, fraction p are black, others are white
  - Draw N marbles from bowl, *put marble back after each drawing*
  - Distribution of **R** black marbles in drawn sample:

$$P(R;p,N) = \overbrace{p^R(1-p)^{N-R}}^{\text{Probability of a specific outcome e.g. 'BBBWBWW'}} \overbrace{\frac{N!}{R!(N-R)!}}^{\text{Number of equivalent permutations for that outcome}}$$

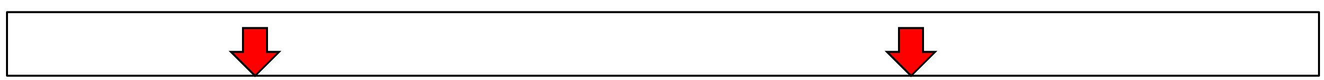


Binomial distribution

Basic Distributions – the Poisson distribution

- Sometimes we don't know the equivalent of the number of drawings
  - Example: Geiger counter
  - Sharp events occurring in a (time) continuum

Begin measurement



End measurement

- What distribution to we expect in measurement over a fixed amount of time?
  - Can be related to Binomial distribution by dividing time interval in fixed number of small intervals, counting #intervals with a collision

Begin measurement

End measurement

A probability model for LHC collisions

- For k expected collisions in measurement, probability of collision in one of N intervals is k/N → Now back to binomial distribution

Begin measurement

End measurement

$$p(r \mid \tfrac{k}{N}, N) = \frac{k^r}{N^r}\left(1 - \frac{k}{N}\right)^{N-r} \frac{N!}{r!(N-r)!}$$

- Now take limit N→∞
  (to avoid possibility of >1 collision per interval)

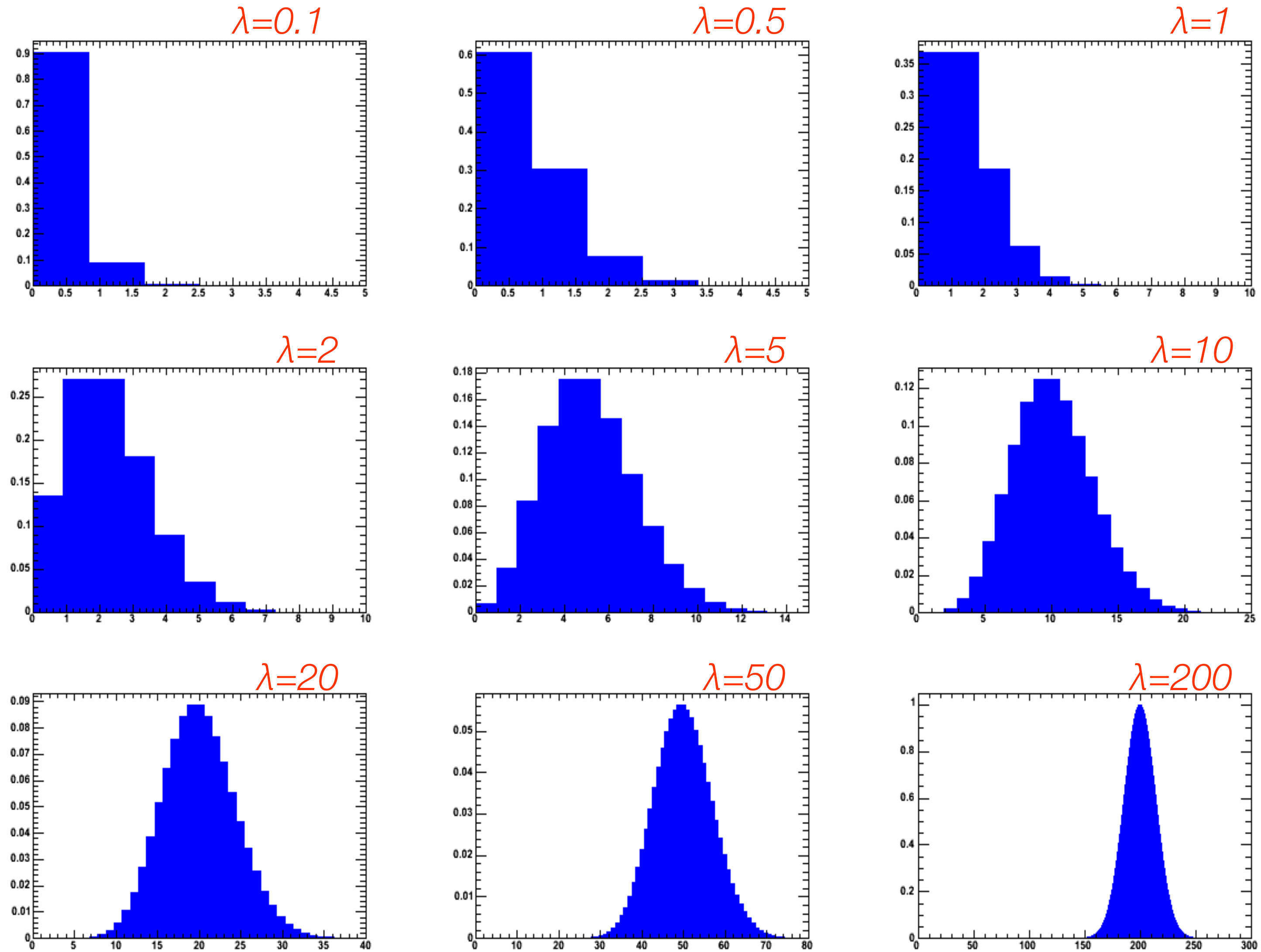
$$\begin{aligned} \lim_{n\to\infty} \frac{n!}{(n-r)!} &= n^r \\ \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n-r} &= e^{-\lambda} \end{aligned}$$

$$p(r \mid k) = \frac{e^{-k} k^r}{r!}$$

Wouter Verkerke, NIKHEF

The Poisson distribution for values value of λ

$$p(r \mid k) = \frac{e^{-k} k^{r}}{r!}$$



Named after Simeon de Poisson – who was investigating the occurence of judgement errors in the French judicial system

More properties of the Poisson distribution

$$P(r;\lambda) = \frac{e^{-\lambda}\lambda^r}{r!}$$

- Mean, variance:

$$\langle r \rangle = \lambda$$

$$V(r) = \lambda \quad \Rightarrow \quad \sigma = \sqrt{\lambda}$$

- Convolution of 2 Poisson distributions is also a Poisson distribution with $\lambda_{ab}=\lambda_a+\lambda_b$

$$P(r) = \sum_{r_A=0}^{r} P(r_A;\lambda_A)P(r-r_A;\lambda_B)$$

$$= e^{-\lambda_A}e^{-\lambda_B}\sum \frac{\lambda_A^{r_A}\lambda_B^{r-r_A}}{r_A!(r-r_A)!}$$

$$= e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}\sum_{r_A=0}^{r}\frac{r!}{(r-r_A)!}\left(\frac{\lambda_A}{\lambda_A+\lambda_B}\right)^{r_A}\left(\frac{\lambda_B}{\lambda_A+\lambda_B}\right)^{r-r_A}$$

$$= e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}\left(\frac{\lambda_A}{\lambda_A+\lambda_B}+\frac{\lambda_B}{\lambda_A+\lambda_B}\right)^r$$

$$= e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}$$

Basic Distributions – The Gaussian distribution

- Look at Poisson distribution in limit of large N

$$P(r;\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}$$

Take log, substitute, r = λ + x,
and use $\ln(r!) \approx r\ln r - r + \ln\sqrt{2\pi r}$

$$\begin{aligned}
\ln(P(r;\lambda)) &= -\lambda + r\ln\lambda - (r\ln r - r) - \ln\sqrt{2\pi r} \\
&= -\lambda + r\left[\ln\lambda - \ln(\lambda(1+\frac{x}{\lambda}))\right] + (\lambda + x) - \ln\sqrt{2\pi\lambda} \\
&\approx x - (\lambda - x)\left(\frac{x}{\lambda} + \frac{x^2}{2\lambda^2}\right) - \ln(2\pi\lambda) \\
&\approx \frac{-x^2}{2\lambda} - \ln(2\pi\lambda)
\end{aligned}$$

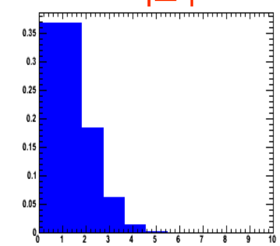
$\ln(1+z) \approx z - z^2/2$

Take exp

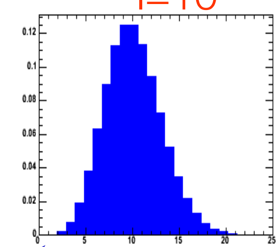
$$P(x) = \frac{e^{-x^2/2\lambda}}{\sqrt{2\pi\lambda}}$$

Familiar Gaussian distribution,
(approximation reasonable for N>10)

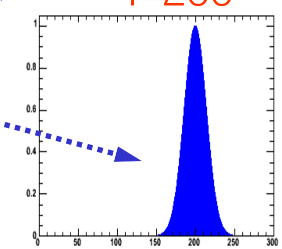
l=1



l=10



l=200



Properties of the Gaussian distribution

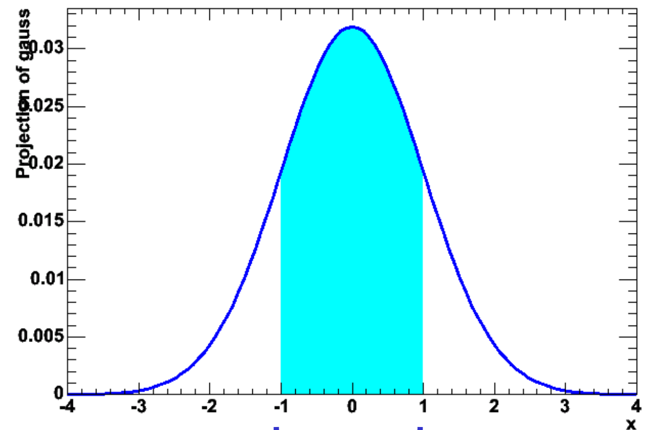
$$P(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}e^{-(x-\mu)^2/2\sigma^2}$$

- *Mean* and *Variance*

$$\langle x\rangle=\int_{-\infty}^{+\infty}xP(x;\mu,\sigma)dx=\mu$$

$$V(x)=\int_{-\infty}^{+\infty}(x-\mu)^2P(x;\mu,\sigma)dx=\sigma^2$$

$$\sigma=\sigma$$



- Integrals of Gaussian

68.27% within 1σ	90% → 1.645σ
95.43% within 2σ	95% → 1.96σ
99.73% within 3σ	99% → 2.58σ
	99.9% → 3.29σ

The Gaussian as 'Normal distribution'

- Why are distributions often Gaussian?

- The ***Central Limit Theorem*** says

- If you take the sum X of N independent measurements x_i, each taken from a distribution of mean m_i, a variance $V_i=\sigma_i^2$, the distribution for x

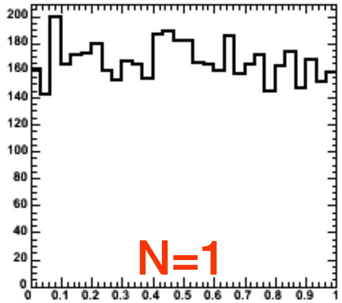
(a) has expectation value $$\langle X \rangle = \sum_i \mu_i$$

(b) has variance $$V(X) = \sum_i V_i = \sum_i \sigma_i^2$$

(c) becomes Gaussian as N → ∞

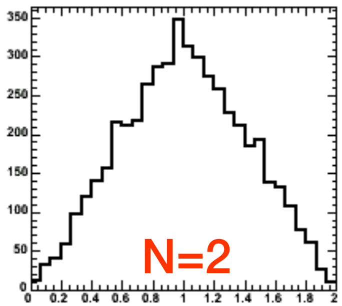
'ex01.C'

Demonstration of Central Limit Theorem



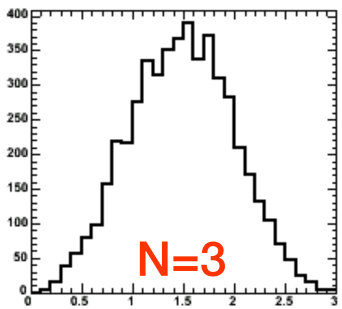
← 5000 numbers taken at random from a uniform distribution between [0,1].

- Mean = $^1/_2$, Variance = $^1/_{12}$

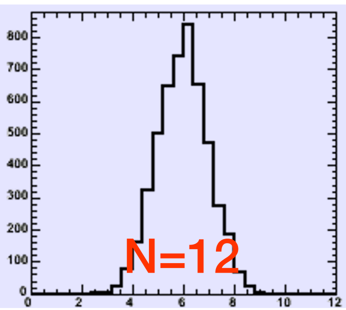


← 5000 numbers, each the sum of 2 random numbers, i.e. $X = x_1+x_2$.

- Triangular shape



← Same for 3 numbers, $X = x_1 + x_2 + x_3$



← Same for 12 numbers, overlaid curve is exact Gaussian distribution

Important: tails of distribution converge very slowly CLT often *not* applicable for '5 sigma' discoveries

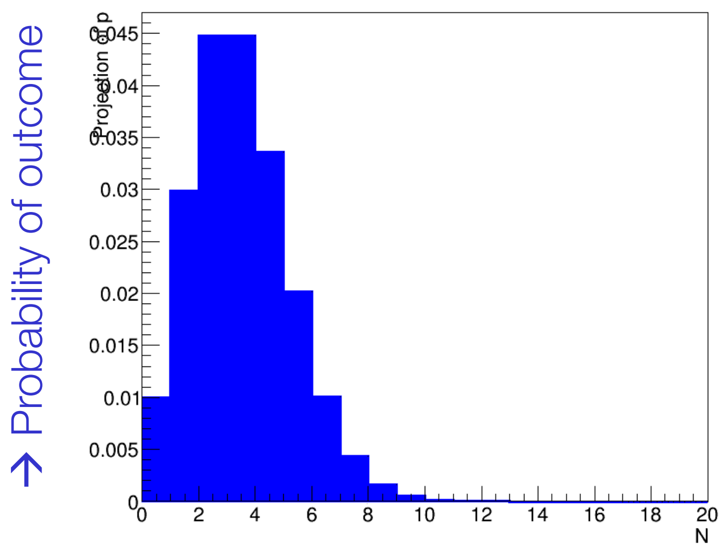
The statistical world

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

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

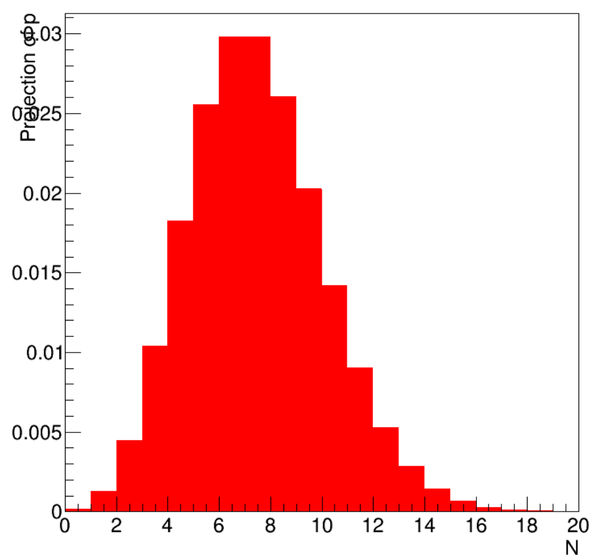
  - Count number of 'events' in a fixed time interval → Poisson distribution
  - Given the *expected event count*, the probability model is fully specified

μ=3 ("bkg only")



→ Experimental outcome

μ=7 ("bkg+signal")



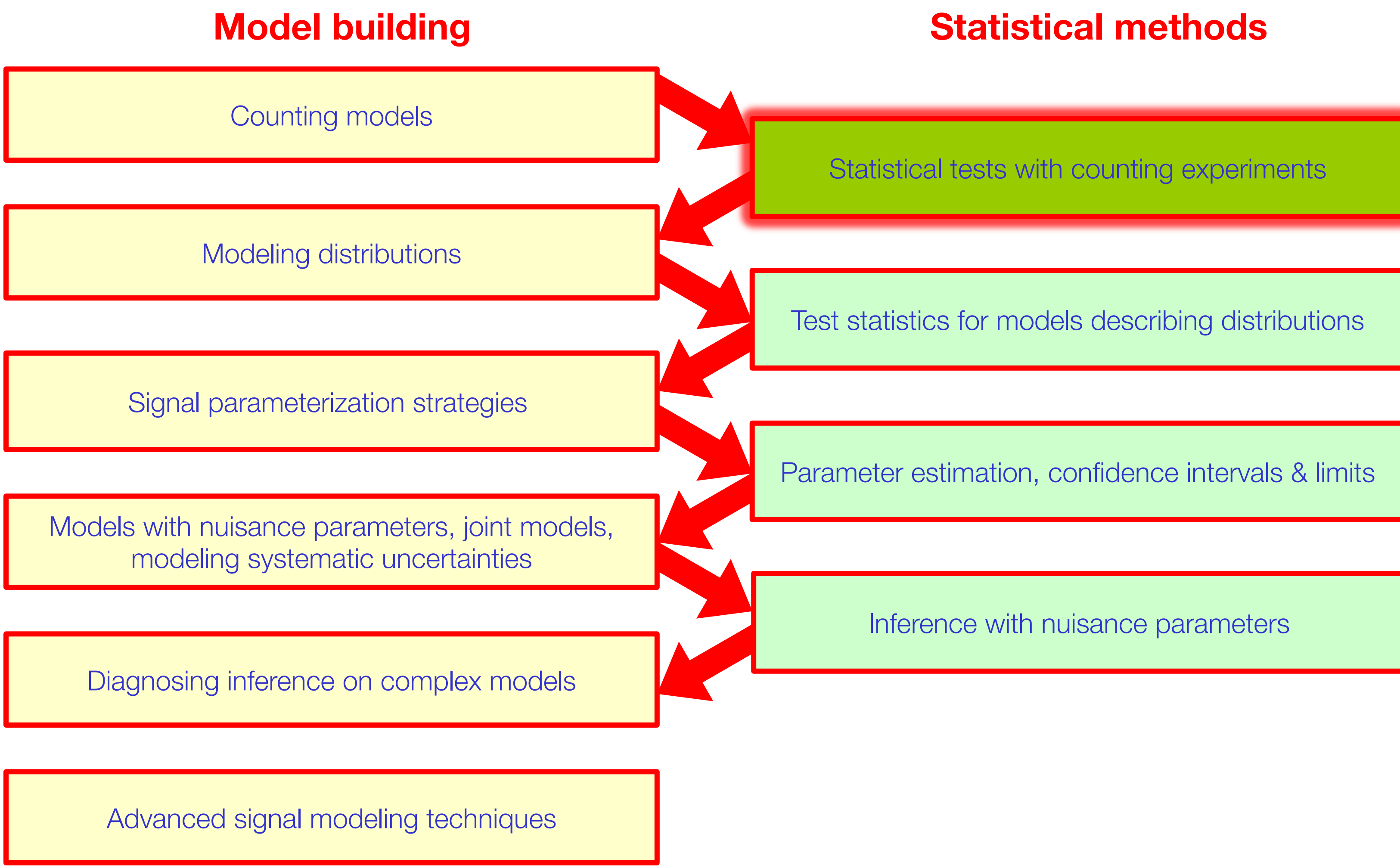
Statistical methods 1

Hypothesis testing, p-values, odds ratios (demonstrated on simple Poisson counting experiments)

Wouter Verkerke, NIKHEF

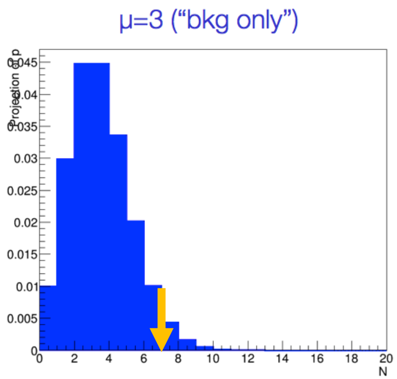
Roadmap of this course

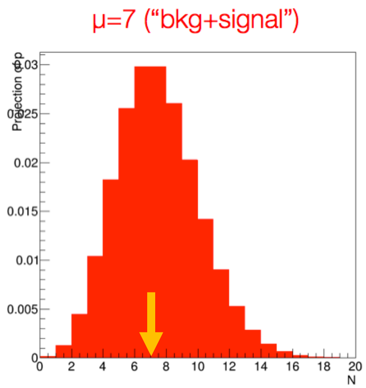
- Start with basics, gradually build up to complexity



Probabilities vs conditional probabilities

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





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

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

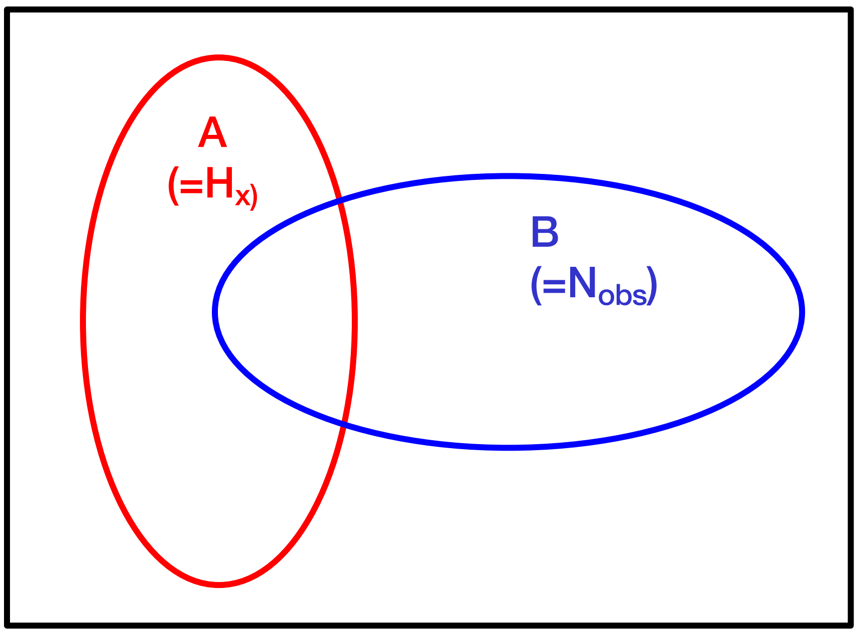
- Suppose we measure N=7 then can calculate

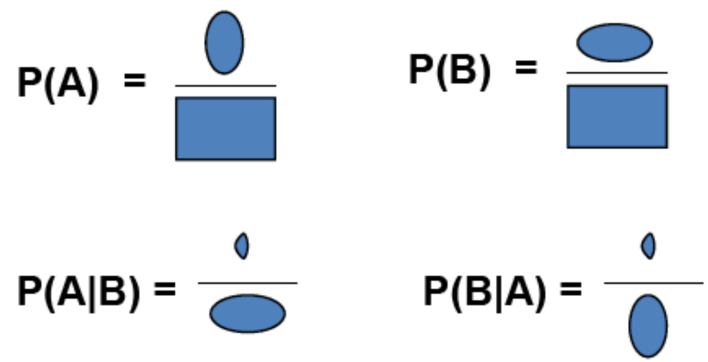
$L(N=7|H_{bkg})=2.2\%$ $\qquad$ $L(N=7|H_{sig+bkg})=14.9\%$

- *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

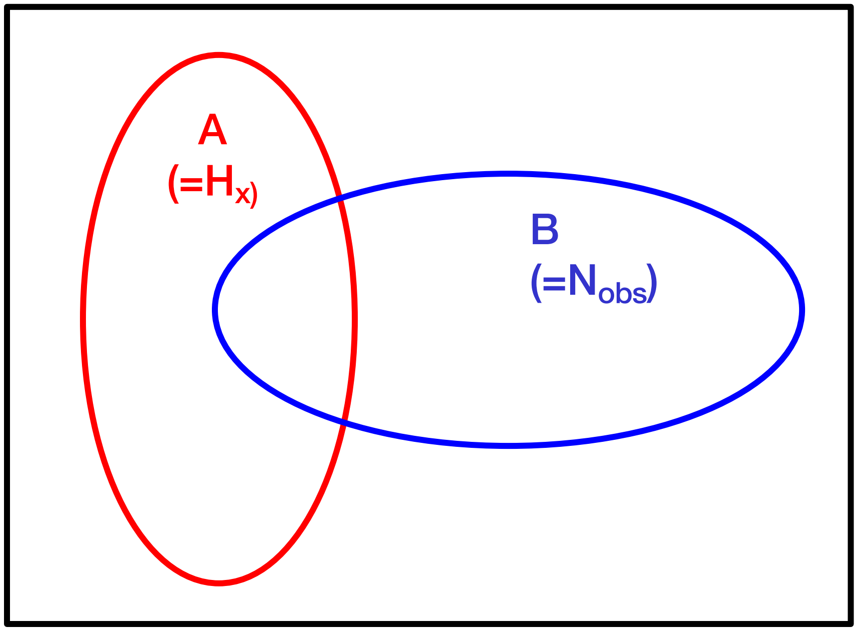


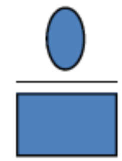


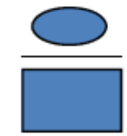
$P(A|B) \neq P(B|A)$

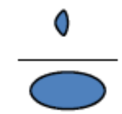
$P(7|H_b) \neq P(H_b|7)$

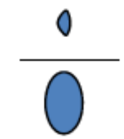
Inverting the conditionality on probabilities



P(A) = 

P(B) = 

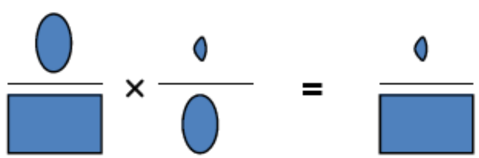
P(A|B) = 

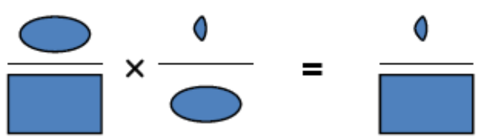
P(B|A) = 

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)?**

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

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

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

and generally for N hypotheses

$$\mathbf{P(N) = \Sigma_i\, P(N|H_i)P(H_i)}$$

- Bayes theorem reformulated using law of total probability

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

- *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 $P(H_b)=50\%$

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

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

- 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

'Bayes Factor' K multiplies prior odds

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

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

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

“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

How a theory becomes text-book physics

Frequentist

Information from experiment

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

P-value threshold from "prior"
(judgment call – no formal theory!)

A: declare discovery at 3σ
B: declare discovery at 5σ

Recent judgements on of 5σ effects:
Higgs – text book
ν(β>1) – rejected

Press release, accept as new 'text book physics'
OR
Wait for more data

Potentially fuzzy information

Prior belief in theory
(can be hard to quantify)

A: $P(H_{sb})=50\%$

B: $P(H_{sb})=0.000001\%$

Cost of wrong decision
(can be hard to quantify)

Cost(FalseDiscovery)
= EternalRidicule/Fired

Cost(UnclaimedDiscovery)
= MissedNobelPrize

Bayesian

Information from experiment

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

Posterior from expt and prior following Bayesian paradigm

A: $P(H_{sb}|data)=0.9999998$
B: $P(H_{sb}|data) = 83\%$

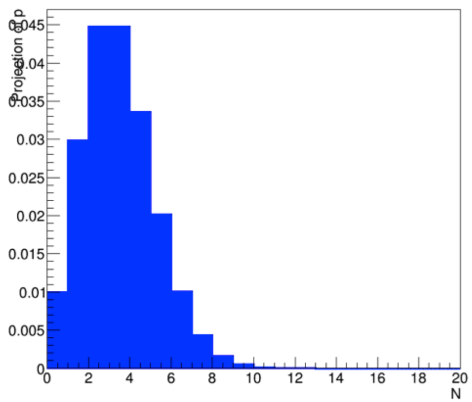
Press release, accept as new 'text book physics'
or
Wait for more data

[Recap]

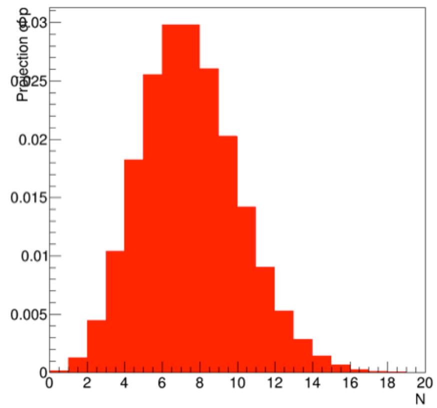
Summary on statistical test with simple hypotheses

- 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

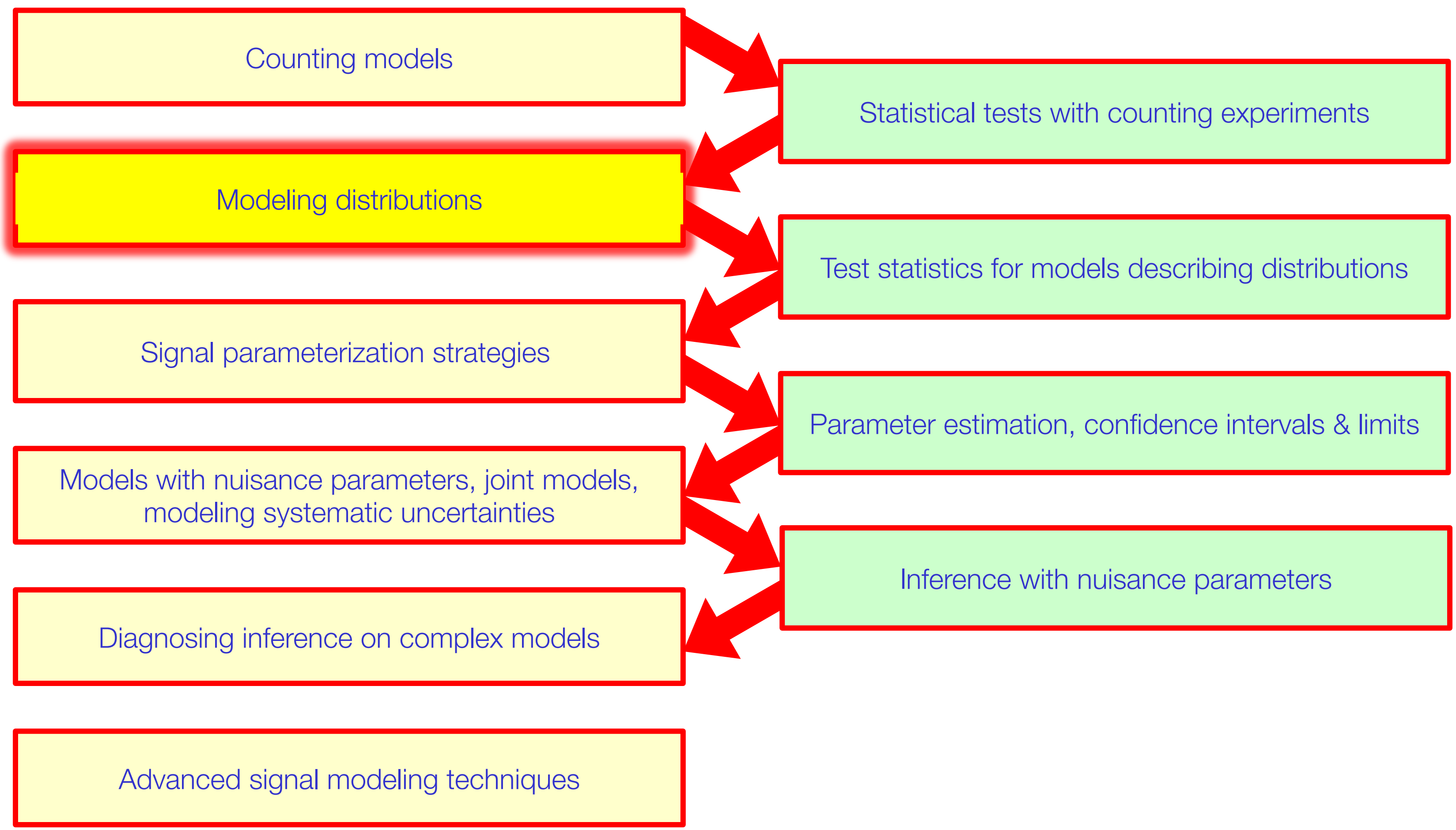
Modelling distributions – template based models or analytical models

Wouter Verkerke, NIKHEF

Roadmap of this course

- Start with basics, gradually build up to complexity

Model building **Statistical methods**



Discriminating observables & counting experiments

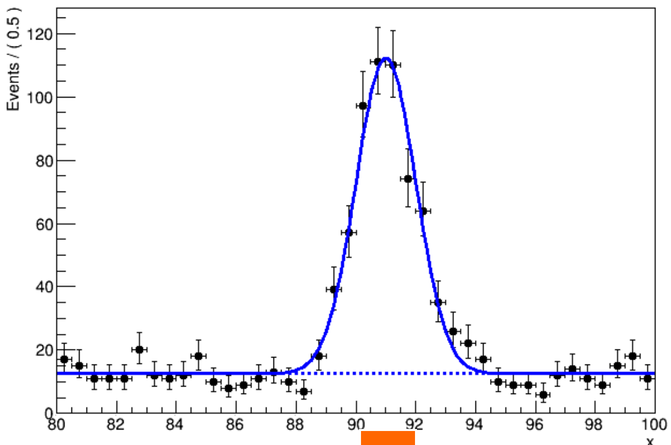
- HEP experimental data usually has many discriminating observables that carry information that can distinguish signal from background hypothesis
- In principle can use them all directly in an elaborate hypothesis test.
  - But would need to formulate a model that describe the expected distribution of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

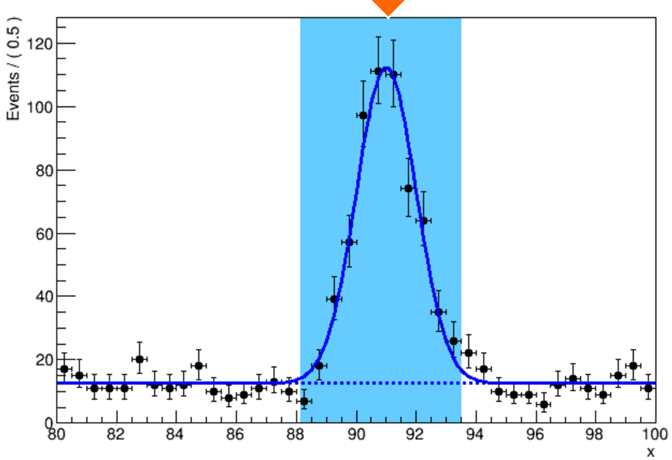
- Example 1 – **Discrimination in selection stage only**

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

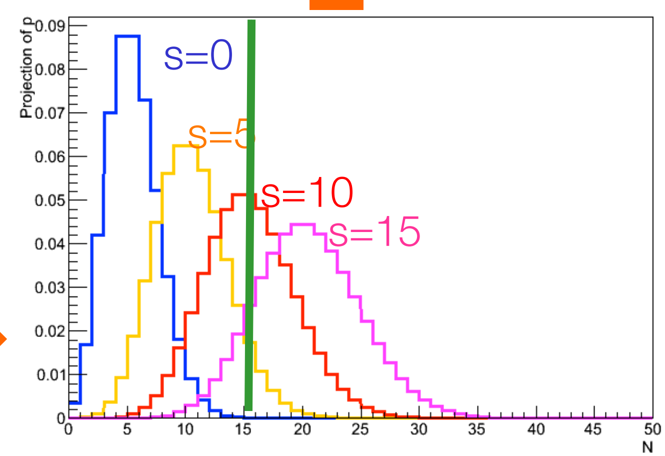
NB2: Selection must be tuned on a 'figure of merit' usually a simplified statistical inference test



Event selection: reduce sample size and dimensionality



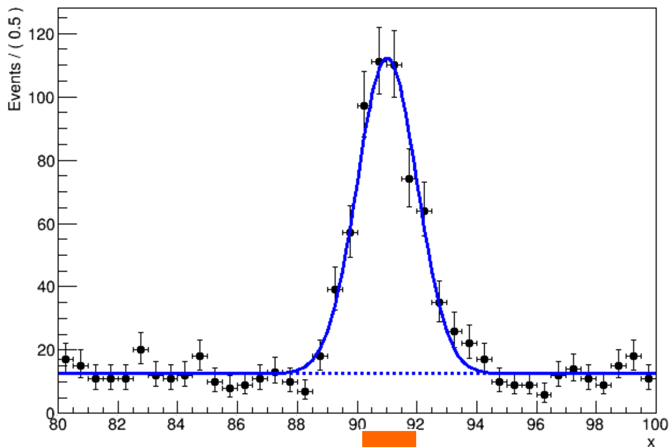
Statistical inference:
$L(15|5) = 1.5\ 10^{-4}$



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

Modeling discriminating observables

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



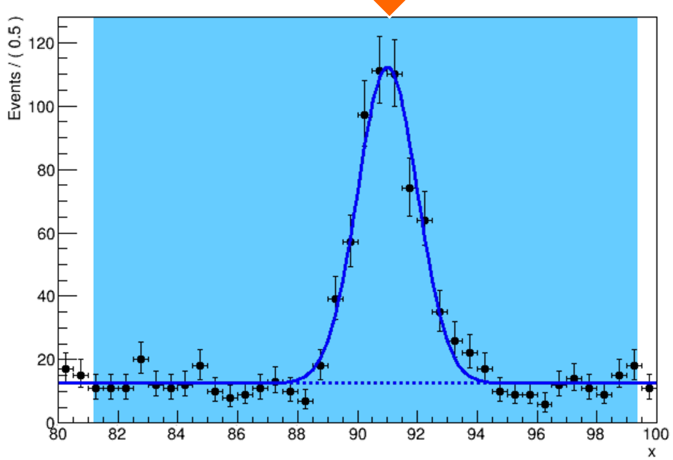
NB1: Most discrimination power in inference step.
This is again design choice!

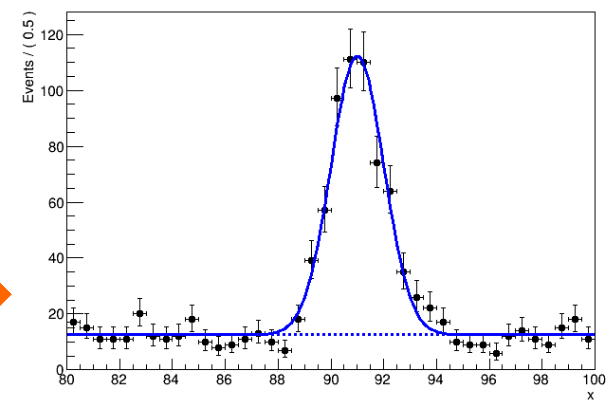
NB2: Optimal selection less critical

NB3: Correct description of selected sample more complex

Event selection: reduce sample size and dimensionality

Statistical inference:
L(data|hypo)=something

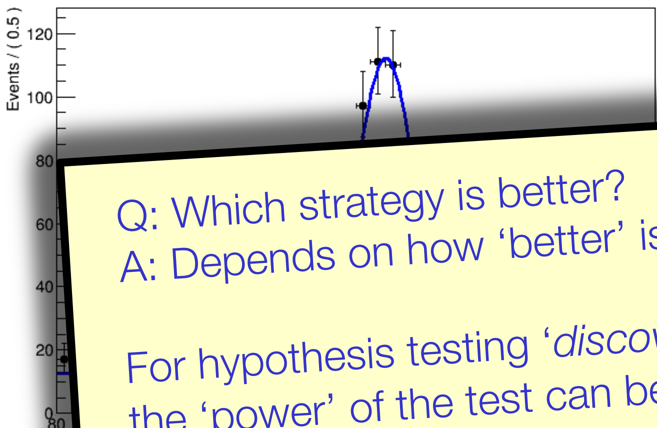




Formulation of probability model of reduced sample:
*Nbkg*Uniform(x) +Nsig*Gaussian(x)*

Modeling discriminating observables

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



NB1: Most discrimination power in inference step.

This is again design choice!

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

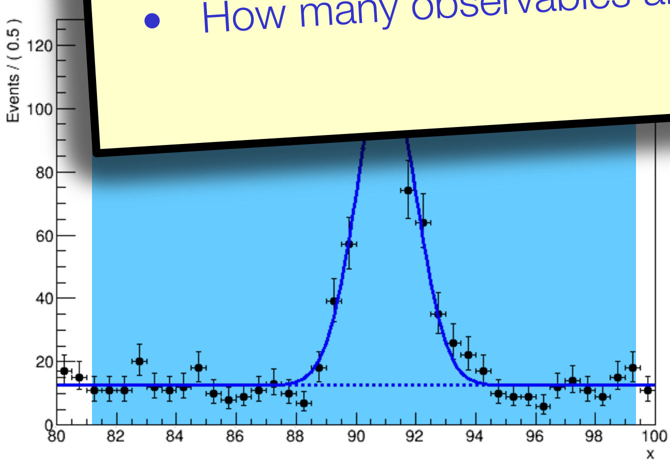
For hypothesis testing '*discovery of a new particle*' the 'power' of the test can be the same, but doesn't need to be

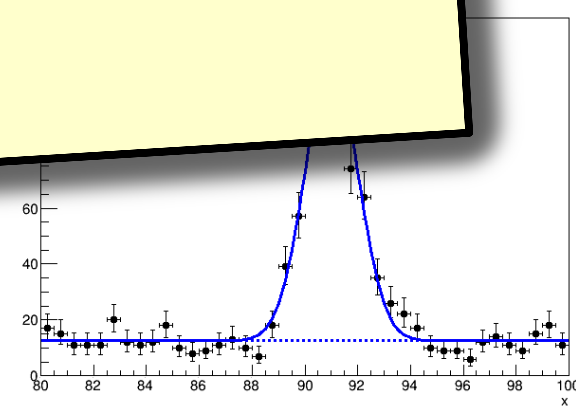
Choice is real life largely dictated by practicalities
- How easy is it to formulate a description of the observables?
- How many observables are important?

mple

:

thing





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

Formulating probability models for discriminating observables

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

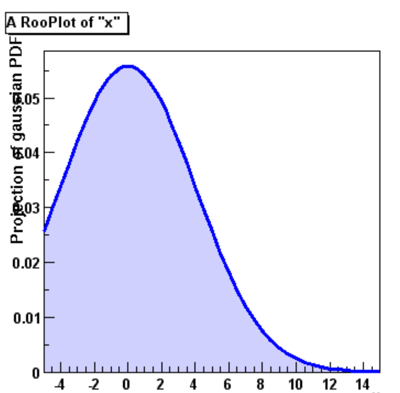
Mathematical formulation of models for observables

- Mathematical description for counting expt is probability model

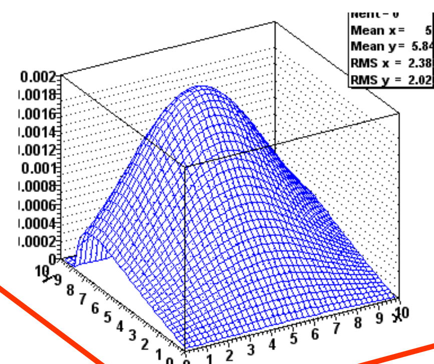
$$P(N) \geq 0 \quad \forall N \qquad\qquad \sum_{N=0}^{\infty} P(N) \equiv 1$$

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

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



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



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

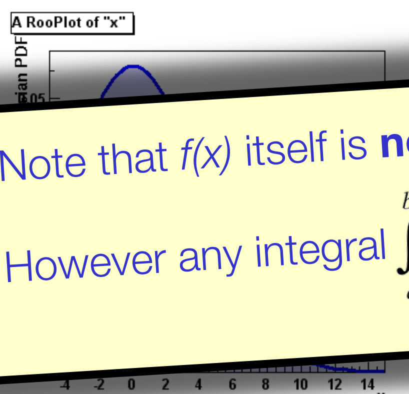
Mathematical formulation of models for observables

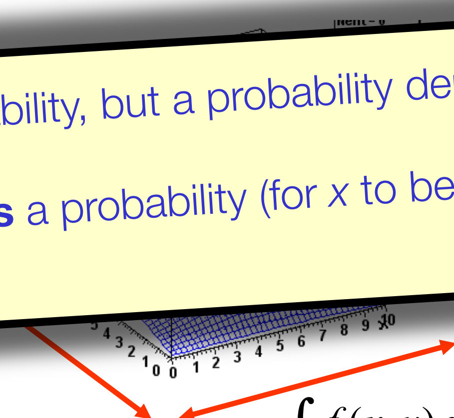
- Mathematical description for counting expt is probability model

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

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

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





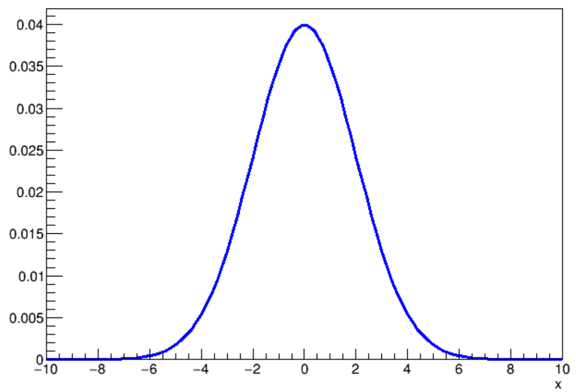
Note that $f(x)$ itself is **not** a probability, but a probability density.

However any integral $\int_a^b f(x)\,dx$ **is** a probability (for x to be in $[a,b]$)

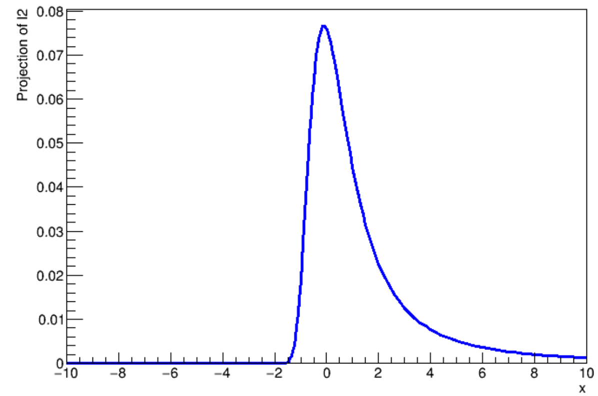
$$\int f(x)\,dx \equiv 1 \qquad\qquad \int f(x,y)\,dx\,dy \equiv 1$$

Some examples of physics-inspired probability density models

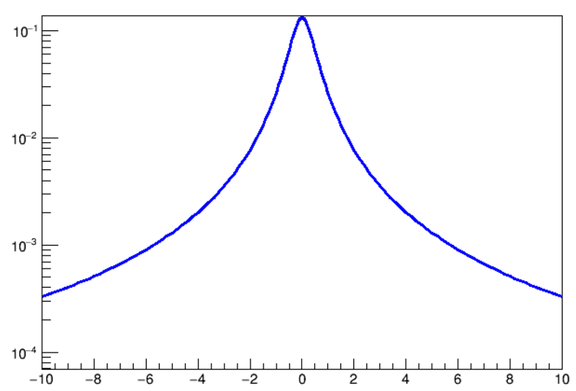
Gaussian
(anything in CLT regime)



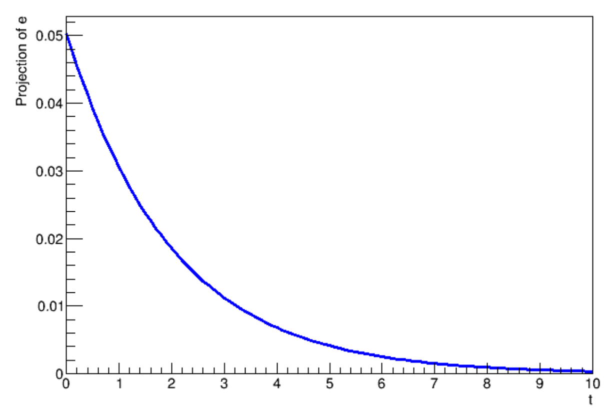
Landau
(energy loss in matter)



Breit-Wigner
(resonant mass)

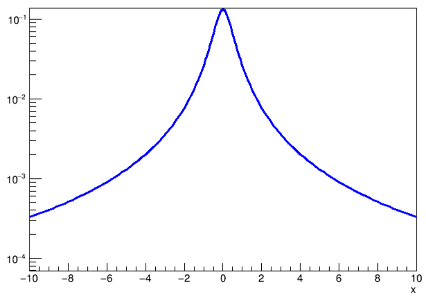


Exponential
(decay time)

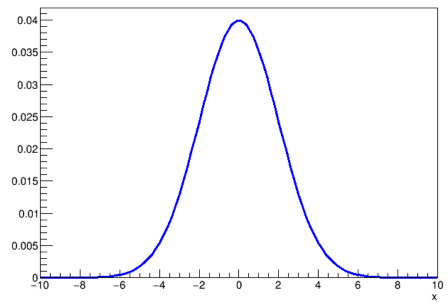


Signal models are often convolutions!

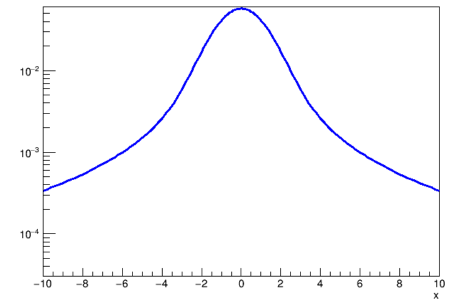
- Observable distributions are often well described by convolutions of physics distributions with (experimental) resolution functions.
  - Often can be calculated analytically, otherwise numerically use FFT
- Example 1: Resonance mass (x) detector resolution



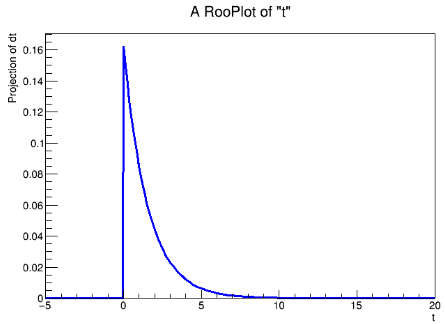
(x)



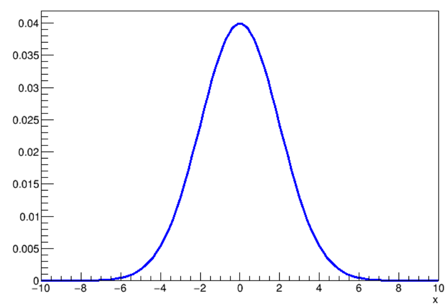
=



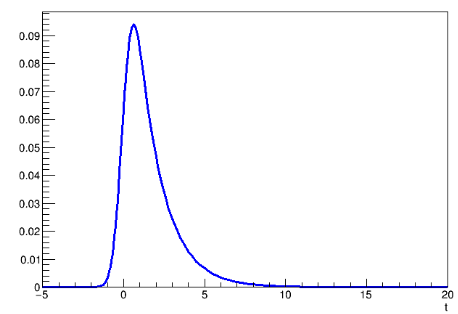
- Example 2: Decay life time (x) detector resolution



(x)



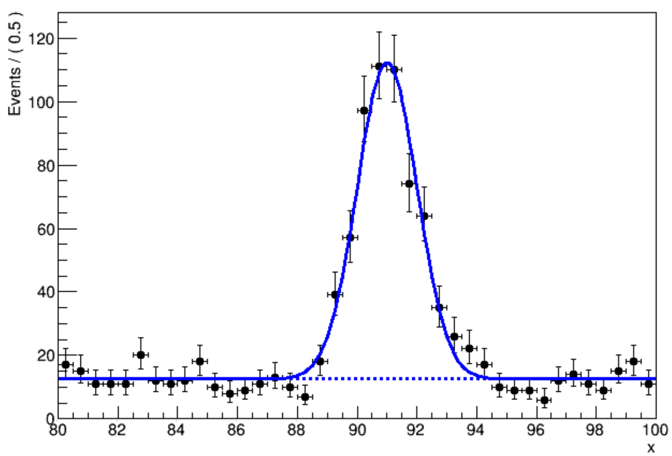
=



PDFs with multiple process contributions

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

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



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

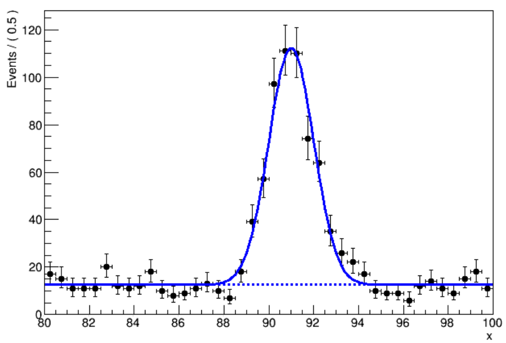
- Given a data sample D(x) of N *independent identically distributed* observations of x, the Likelihood is

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

PDFs with multiple process contributions

- Note that the Likelihood L(x) of a probability density function f(x) for a data sample D(x) with N entries only exploits the differential distribution in x, *but not the event count N of the data*

- In many cases the event count can also distinguish the S/B hypothesis (more events expected if signal is present). If so, the probability model for the event count can be explicitly included in the Likelihood (often called 'extended likelihood')



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

$P(N)$ = Poisson($N \mid N_{exp}$)

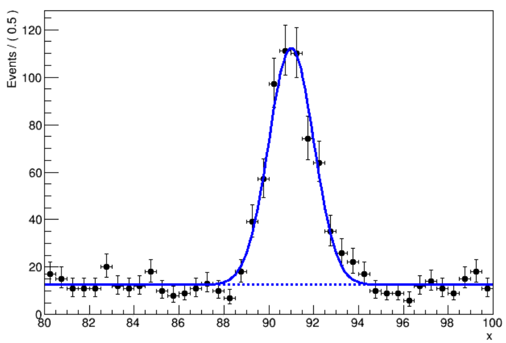
$$L(\vec{x},N) = \prod_{i=0\ldots N} f(x_i \mid f_{sig}) \cdot Poisson(N \mid N_{\exp})$$

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

PDFs with multiple process contributions

- Note that the Likelihood L(x) of a probability density function f(x) for a data sample D(x) with N entries only exploits the differential distribution in x, *but not the event count N of the data*

- In many cases the event count can also distinguish the S/B hypothesis (more events expected if signal is present). If so, the probability model for the event count can be explicitly included in the Likelihood (often called 'extended likelihood')



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

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

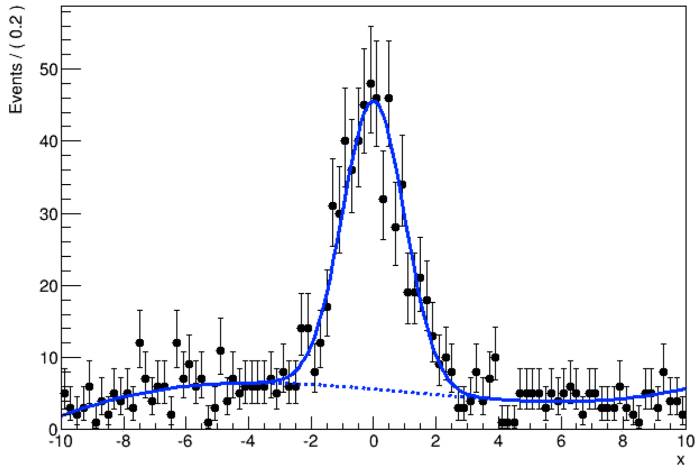
$$L(\vec{x},N) = \prod_{i=0\ldots N} f(x_i \mid S,B)\cdot Poisson(N \mid S+B)$$

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

Empirical probability models

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

Empirical models

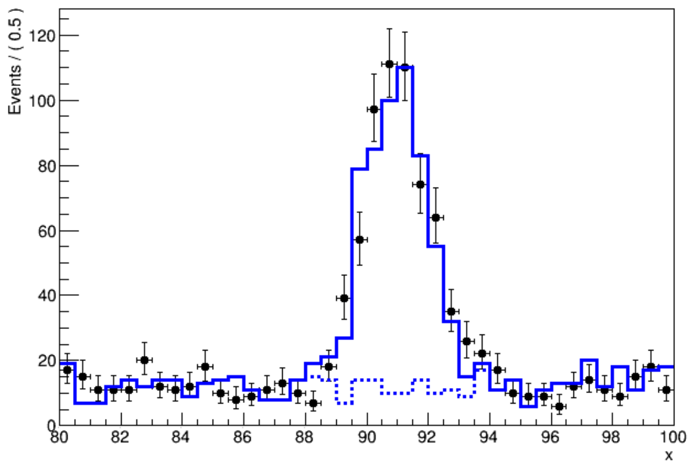


$$B(x) = a_0+a_1x+a_2x^2+a_3x^3\ldots$$

Drawbacks:

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

Simulation-based models



$$B(x) = \text{histogram}$$

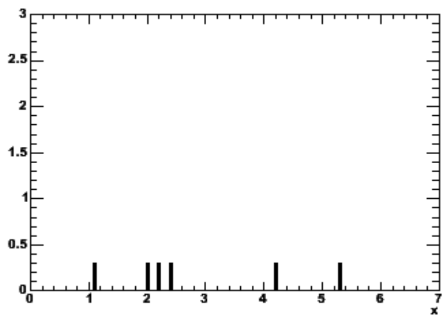
Drawbacks:

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

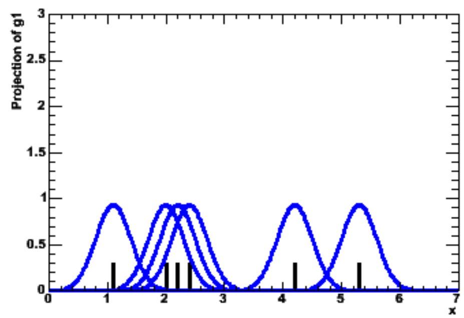
Modeling low-statistics simulation predictions

- 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

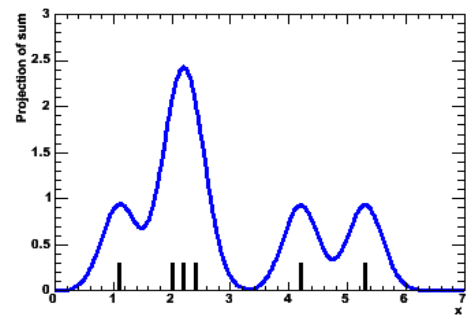
Sample of events



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)