Statistics

W. Verkerke

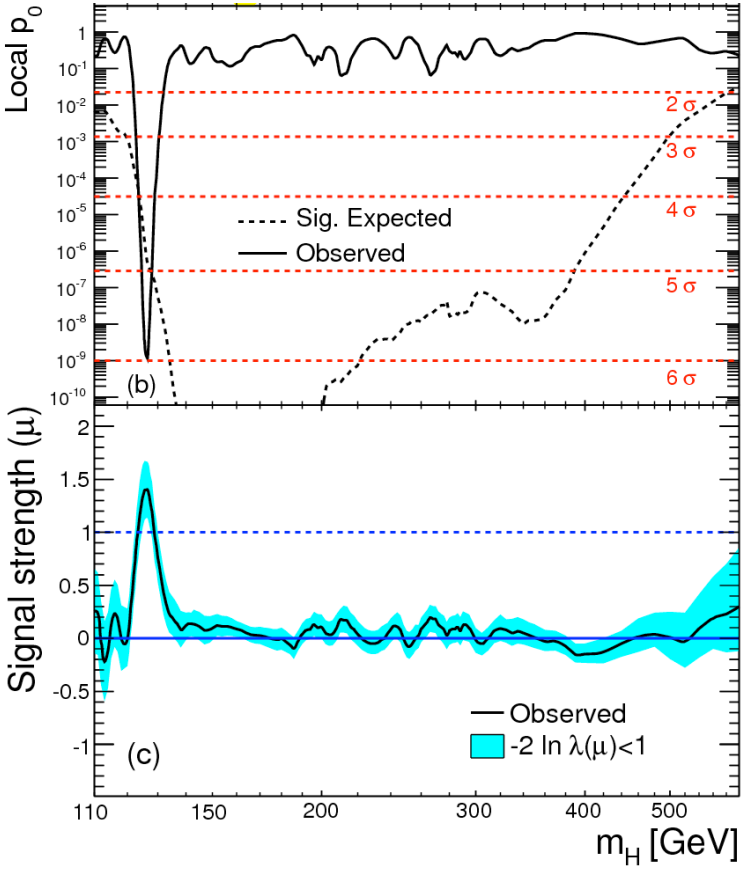
Schedule

- Lectures Monday
  - 10.00-10.45 Basic Concepts
  - 11.00-11.45 Basic Concepts & Event Selection
  - 12.00-12.45 Event Selection
- Exercises Monday 14.00-17.00
- Lectures Tuesday
  - 09.30 - 10.15 Composite Hypotheses
  - 10.30 - 11.15 Composite Hypotheses & Nuisance Parameters
  - *11.30-12.15 (Statistical Issues in point sources – C. Timmermans)*
- Exercises Tuesday 14.00-17.00
- Lectures Wednesday
  - 09.30 - 10.15 Modeling systematics
  - 10.30 - 11.15 Modeling systematics & Fit Diagnostics
  - *11.30-12.15 (Applied statistics outside HEP – M. Baak)*
- Exercises Wednesday 13.30-15.30

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**

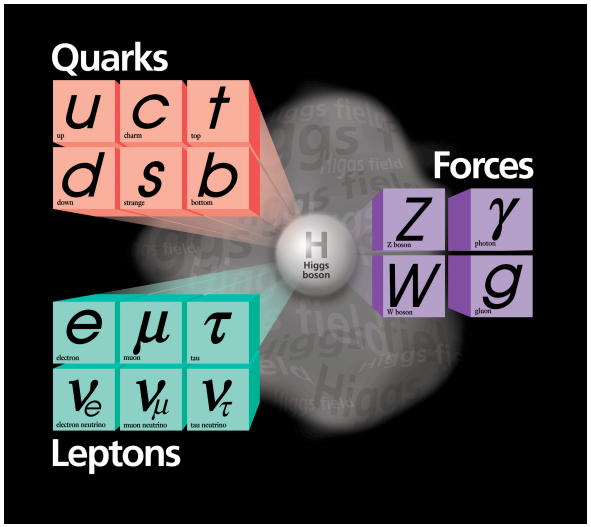
"As a layman I would now say: I think we have it"



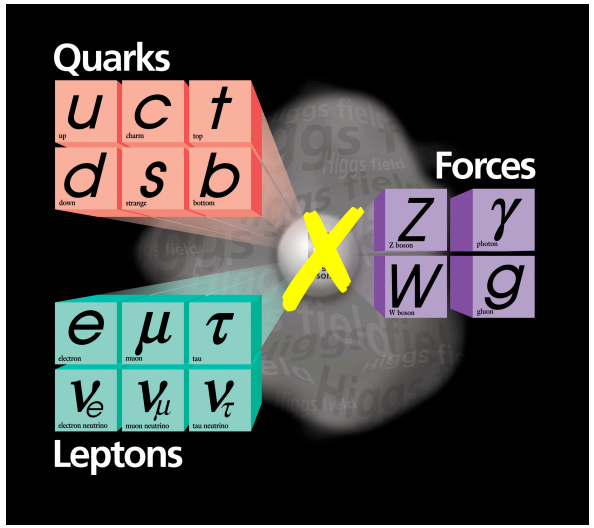
How do we do this?

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

The Standard Model



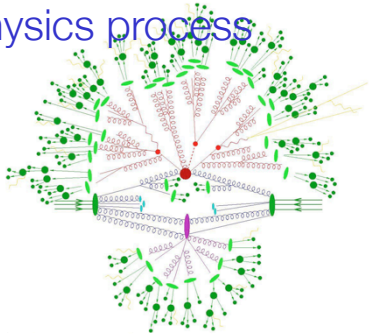
The SM without a Higgs boson



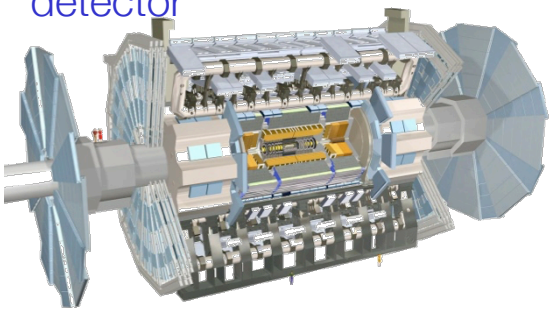
- **Next, you design a measurement to be able to *test* model**
  - Via chain of physics simulation, showering MC, detector simulation and analysis software, a physics model is reduced to a **statistical** model

An overview of HEP data analysis procedures

Simulation of 'soft physics' physics process

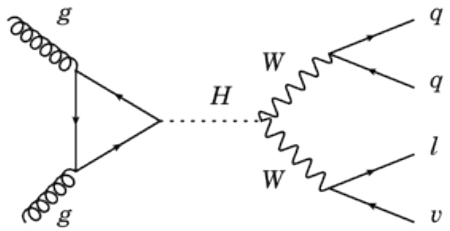


Simulation of ATLAS detector

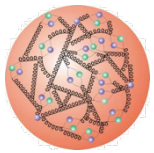


LHC data

Simulation of high-energy physics process



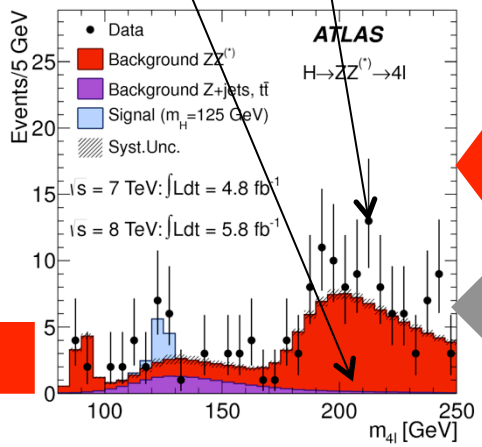
Proton Structure Function



prob(data|SM)

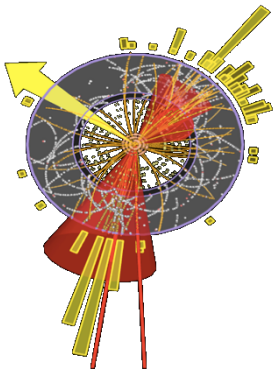
$P(m_{4l}|SM[m_H])$

Observed m_{4l}

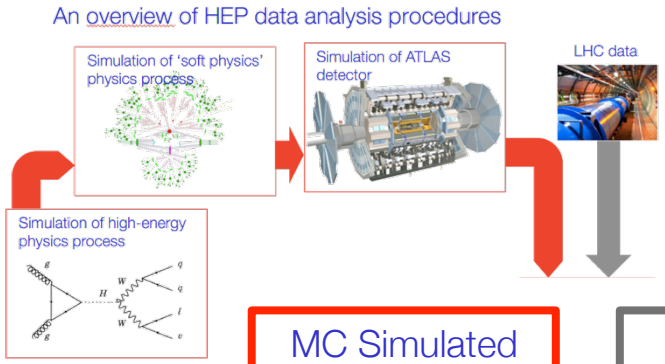


Analysis Event selection

Reconstruction of ATLAS detector



HEP workflow: data analysis in practice



MC Simulated Events (sig,bkg)

All available "real data"

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

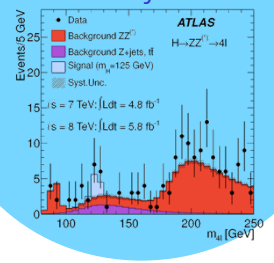
Helps to define selection

Event selection (cuts, NN, BDT)

Final Event Selection (MC)

Final Event Selection (data)

Statistical analysis



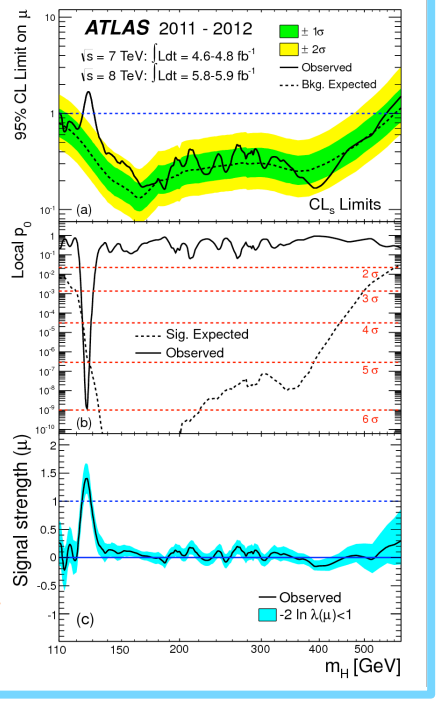
Signal, background models
Likelihood models,
MINUIT, RooFit
RooStats, MCLimit

Final Result

Limit

Discovery

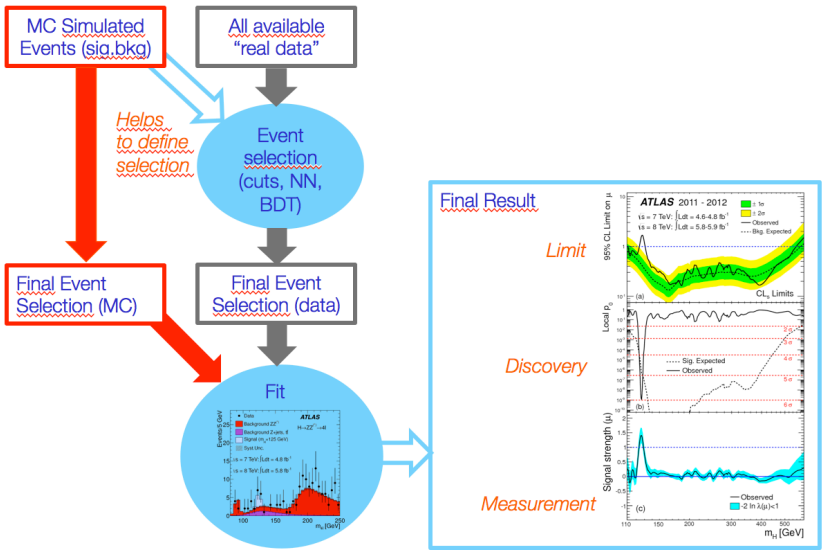
Measurement



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

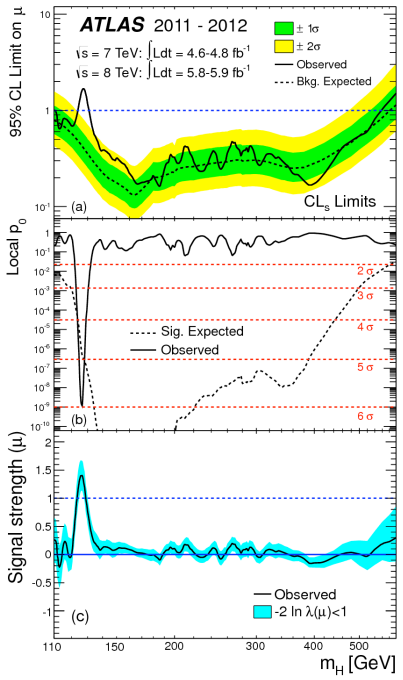
- The next step of the analysis is to confront your model with the data, and summarize the result in a probabilistic statement of some form

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 $1x10^{9}$"

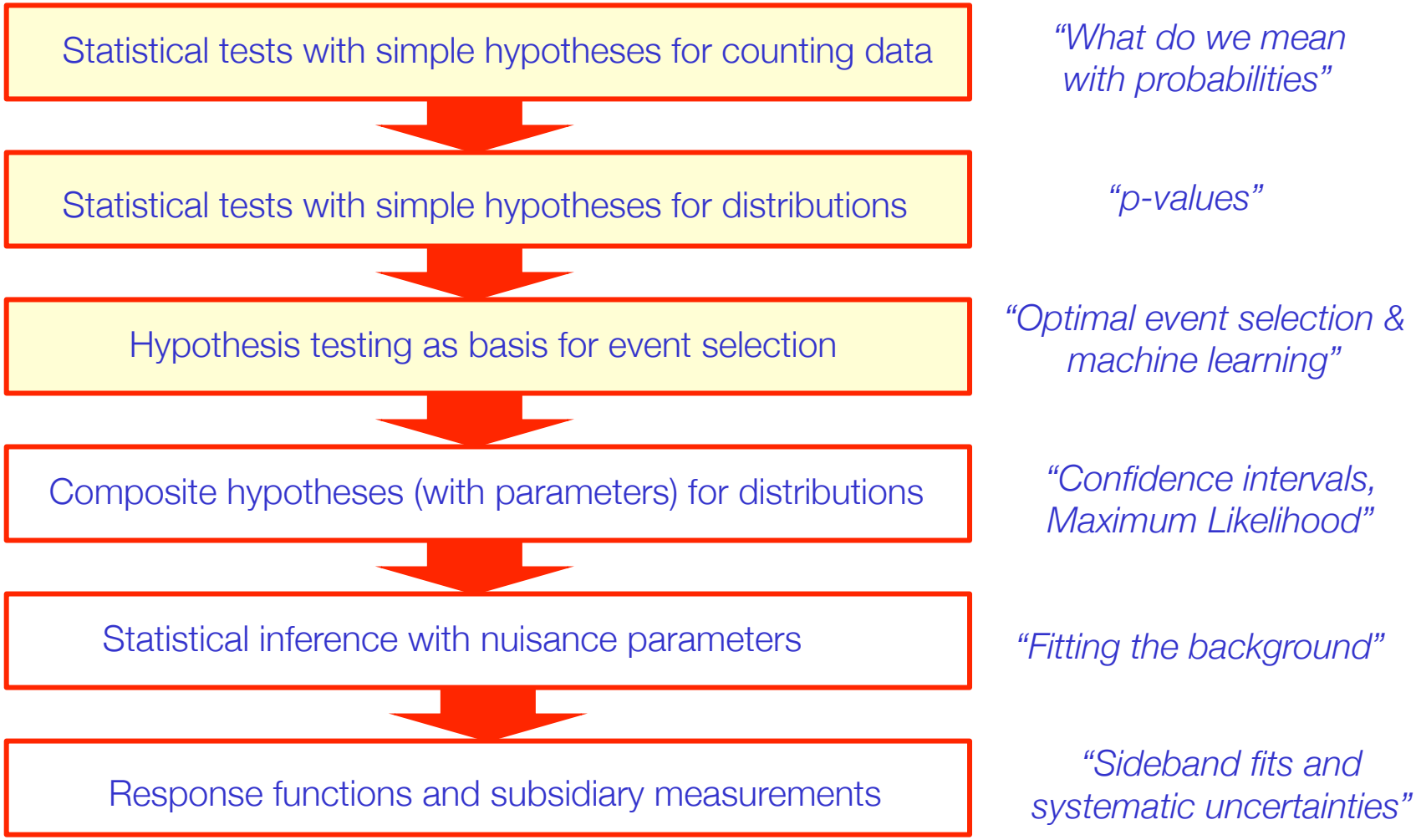
'Measurement with variance estimate'

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

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

Roadmap for this course

- Start with basics, gradually build up to complexity of

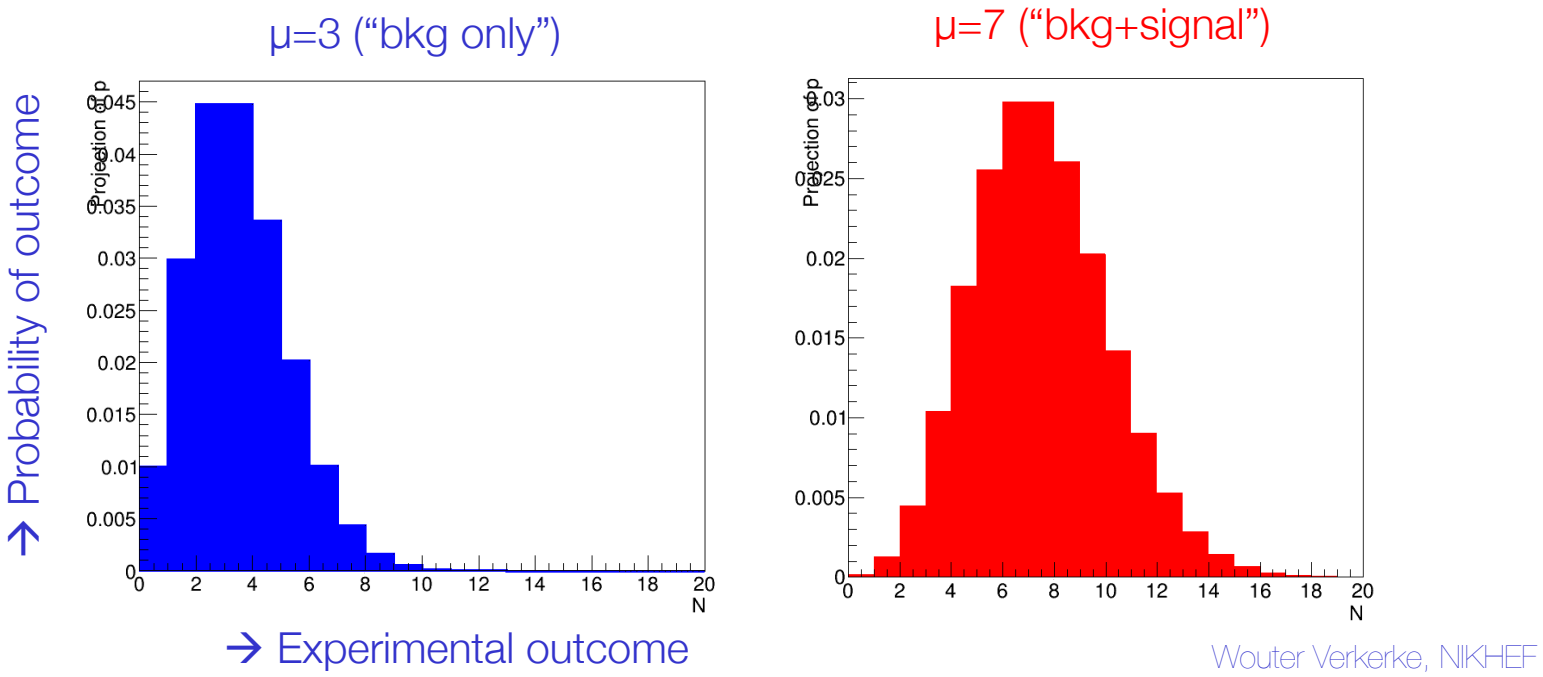


The statistical world

- 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



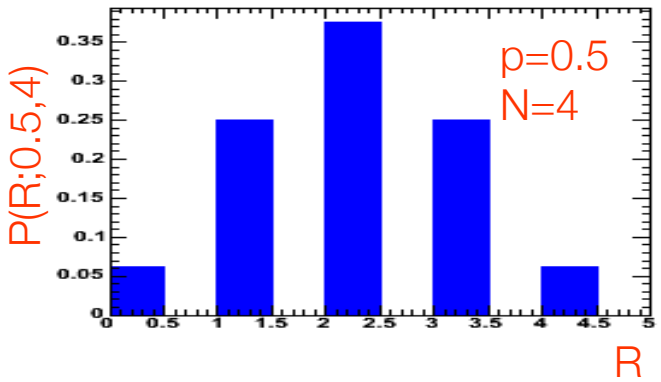
Intermezzo on distributions – The binomial distribution

- Simple 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:

Probability of a specific outcome e.g. 'BBBWBWW'

Number of equivalent permutations for that outcome

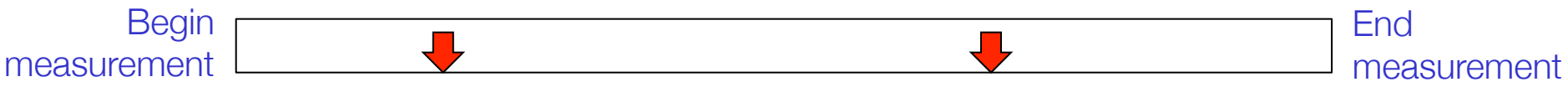
$$P(R;p,N) = p^{R}(1-p)^{N-R}\frac{N!}{R!(N-R)!}$$



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



- 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

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

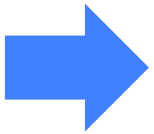
Eind measurement

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

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

$$\lim_{n\to\infty}\frac{n!}{(n-r)!\cdot} = n^r$$

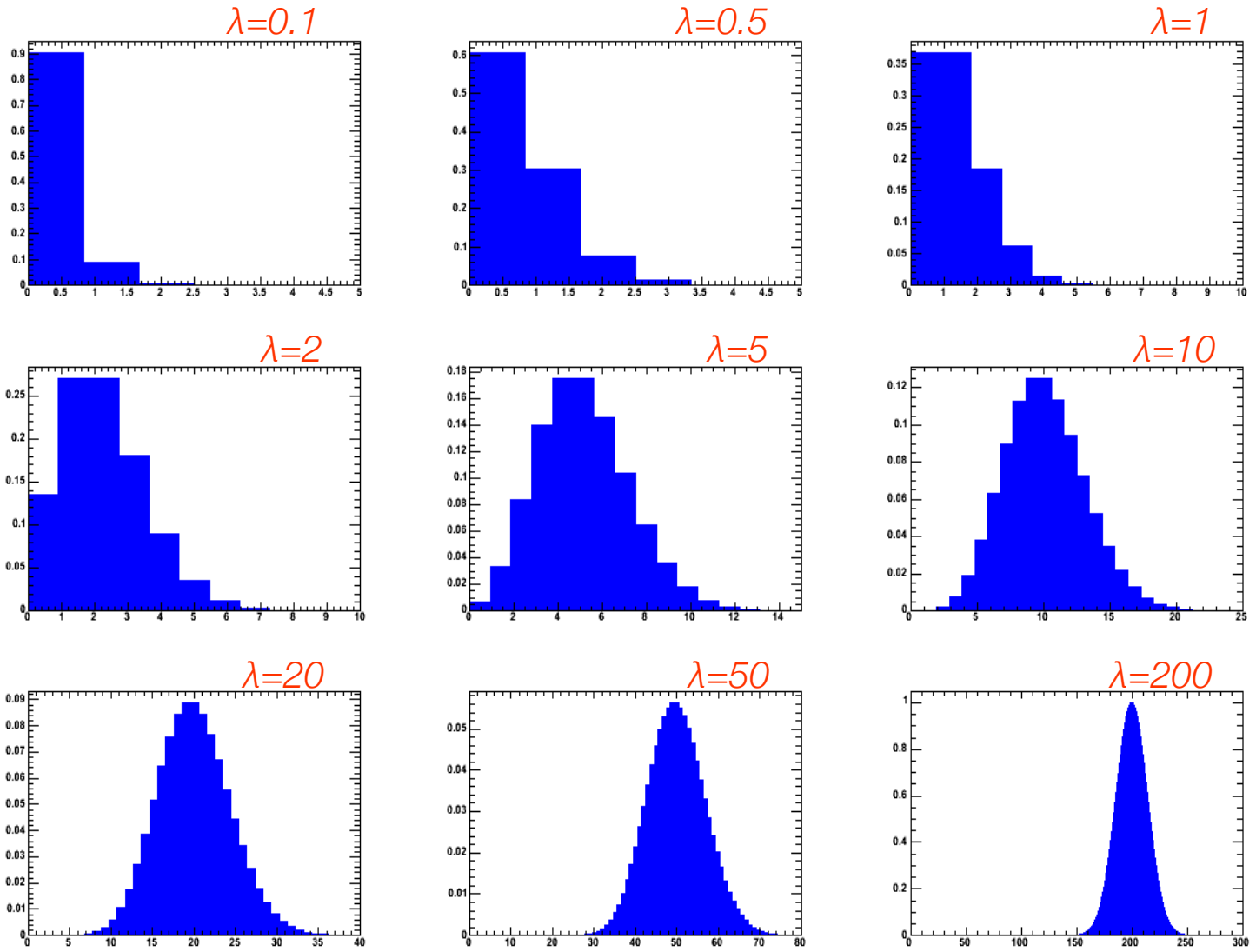
$$\lim_{n\to\infty}(1-\frac{\lambda}{n})^{n-r} = e^{-\lambda}$$



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

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

$$\ln(P(r;\lambda)) = -\lambda + r\ln\lambda - (r\ln r - r) - \ln\sqrt{2\pi r}$$

$$= -\lambda + r\left[\ln\lambda - \ln(\lambda(1+\frac{x}{\lambda}))\right] + (\lambda + x) - \ln\sqrt{2\pi\lambda}$$

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

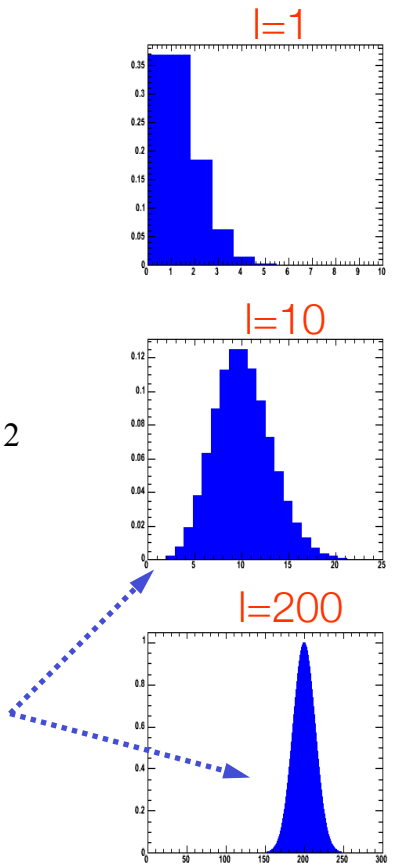
$$\approx x - (\lambda - x)\left(\frac{x}{\lambda} + \frac{x^2}{2\lambda^2}\right) - \ln(2\pi\lambda)$$

$$\approx \frac{-x^2}{2\lambda} - \ln(2\pi\lambda)$$

Take exp

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

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



Properties of the Gaussian distribution

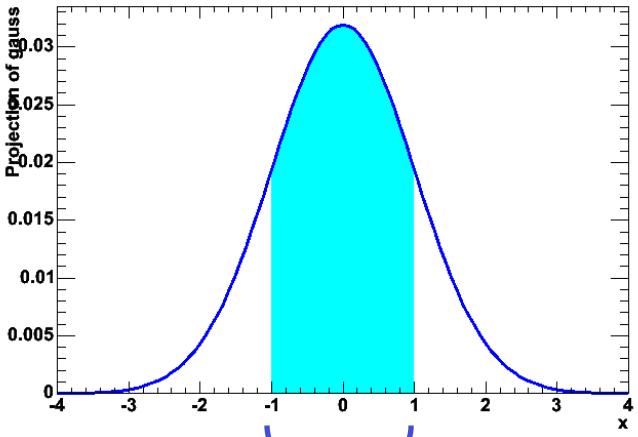
$$P(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

- *Mean* and *Variance*

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

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

$$\sigma = \sigma$$



- 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?

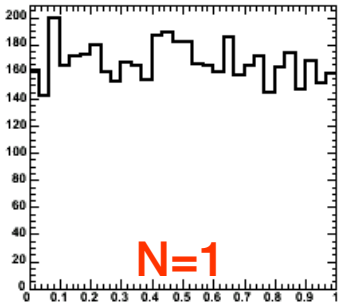
- The ***Central Limit Theorem*** says
- If you take the sum X of N independent measurements x_i, each taken from a distribution of mean m_i, a variance $V_i=\sigma_i^2$, the distribution for x

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

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

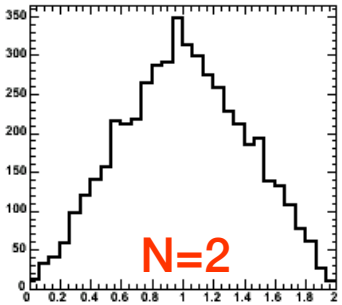
(c) becomes Gaussian as N → ∞

Demonstration of Central Limit Theorem



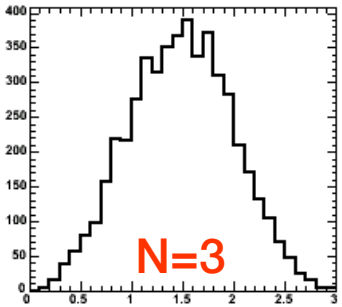
← 5000 numbers taken at random from a uniform distribution between [0,1].

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

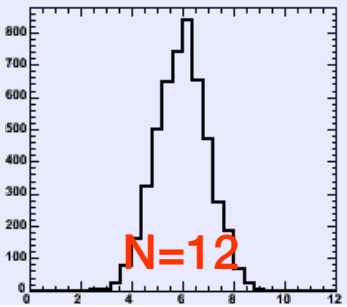


← 5000 numbers, each the sum of 2 random numbers, i.e. $X = x_1 + x_2$.

- Triangular shape



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



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

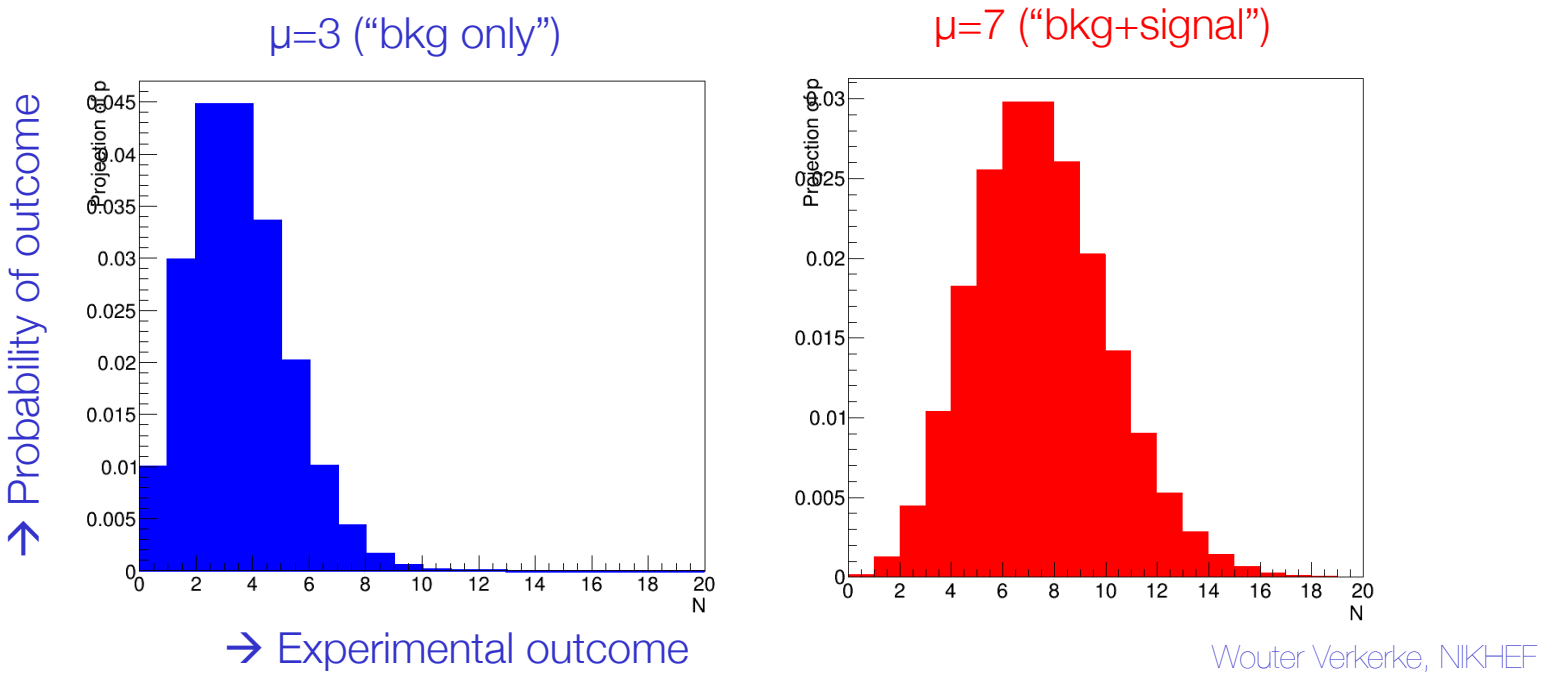
Important: tails of distribution converge very slowly CLT often *not* applicable for '5 sigma' discoveries

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

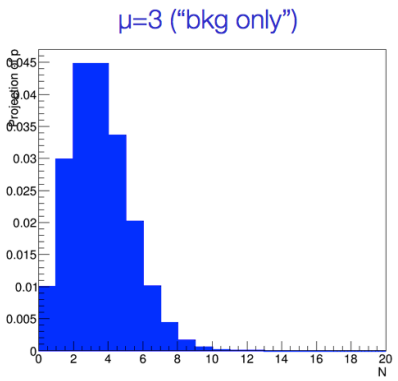
$$P(N \mid \mu) = \frac{\mu^N e^{-\mu}}{N!}$$

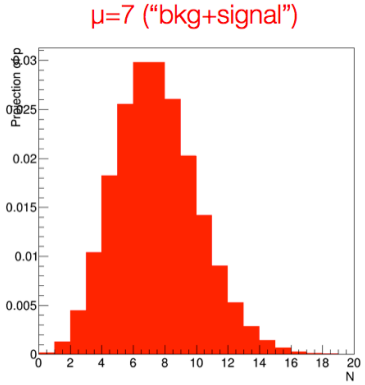
  - Count number of 'events' in a fixed time interval → Poisson distribution
  - Given the *expected event count*, the probability model is fully specified



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***

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

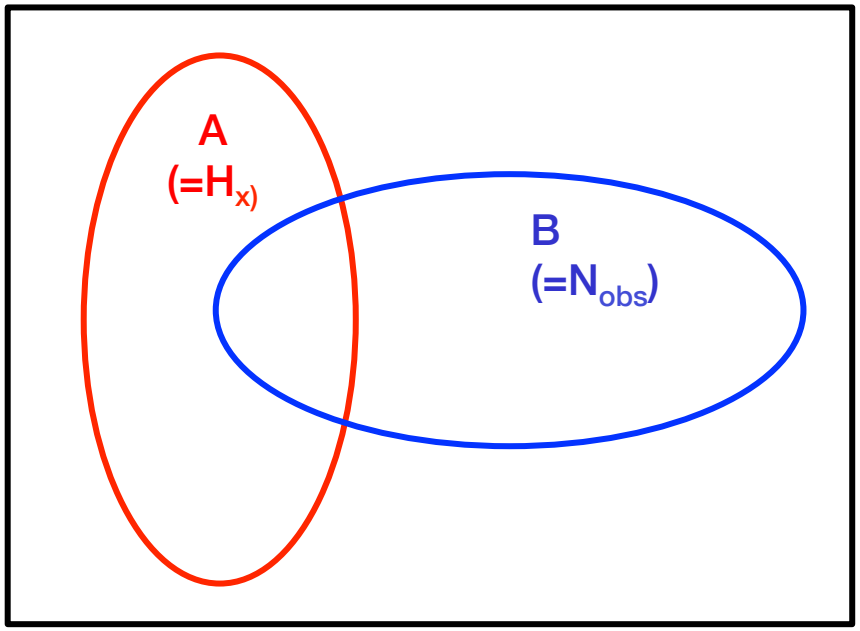
- Suppose we measure N=7 then can calculate

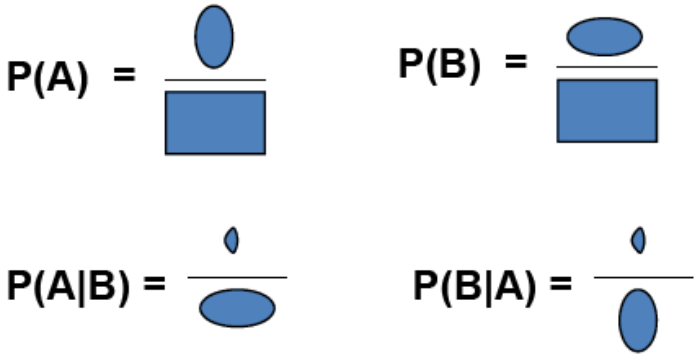
$L(N=7|H_{bkg})=2.2\%$ $L(N=7|H_{sig+bkg})=14.9\%$

- *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

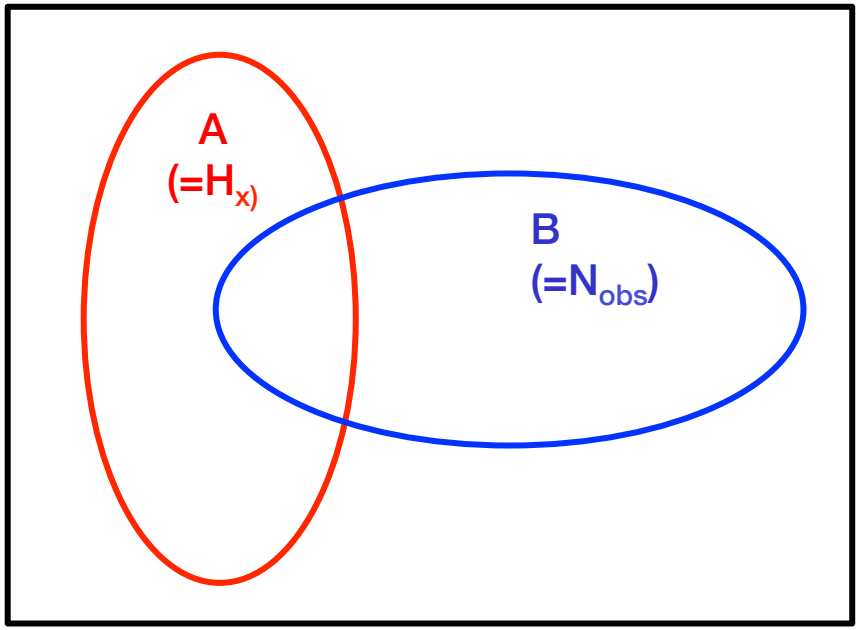


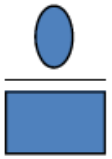
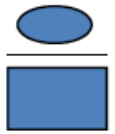


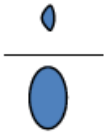
$P(A|B) \neq P(B|A)$

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

Inverting the conditionality on probabilities

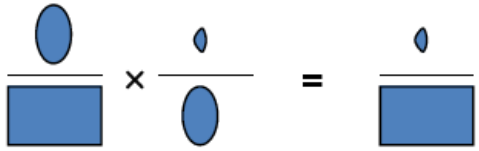


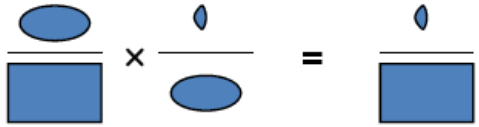
$P(A) =$  $P(B) =$ 

$P(A|B) =$  $P(B|A) =$ 

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

but you can deduce their relation

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

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

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

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

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

- **What is 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)

  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)

Bayesians and Frequentists

- A slide from a professional statistician found when Googling…

ACCP 37th Annual Meeting, Philadelphia, PA [2]

Differences Between Bayesians and Non-Bayesians
According to my friend Jeff Gill

Typical Bayesian

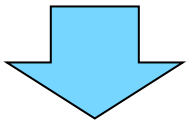
Typical Non-Bayesian

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

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

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

If $p(H_{sb})=p(H_b)$ → Odds are 1:1



'Bayes Factor' K multiplies prior odds

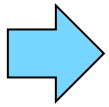
$$O_{posterior} \equiv \frac{L(x \mid H_{sb})P(H_{sb})}{L(x \mid H_{sb})P(H_b)} = \frac{L(x \mid H_{sb})}{L(x \mid H_b)} O_{prior}$$

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

Formulating evidence for discovery

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

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



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

- Given importance to interpretation of the lower probability, it is customary to quote it in "physics intuitive" form: Gaussian σ.
  - E.g. '5 sigma' → probability of 5 sigma Gaussian fluctuation $=2.87\times10^{-7}$
- No formal rules for 'discovery threshold'
  - Discovery also assumes data is not too unlikely under H_{sb}. If not, no discovery, but again no formal rules ("your good physics judgment")
  - NB: In Bayesian case, both likelihoods low reduces Bayes factor K to O(1)

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:

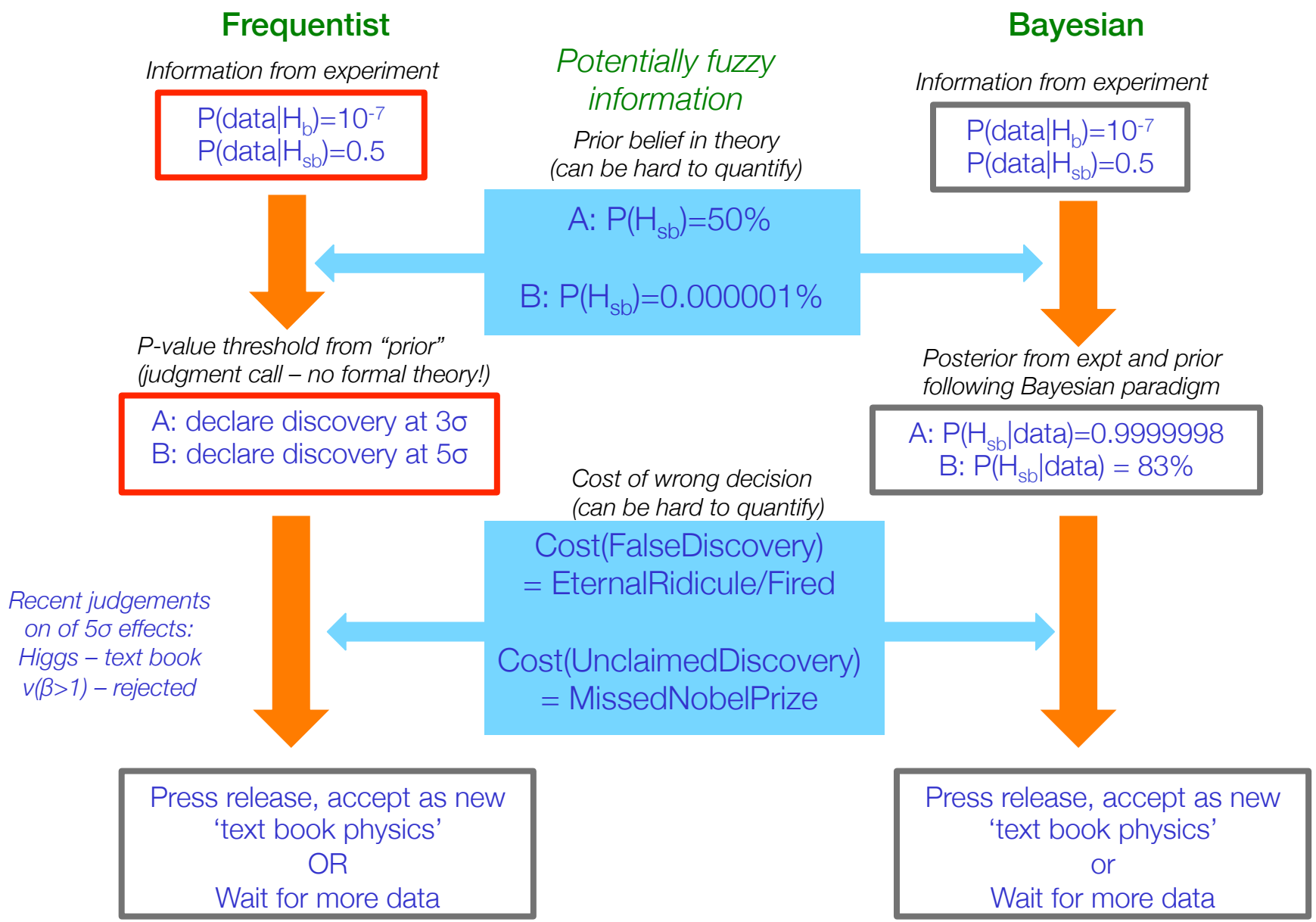
  Your prior probabilities, and

  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.

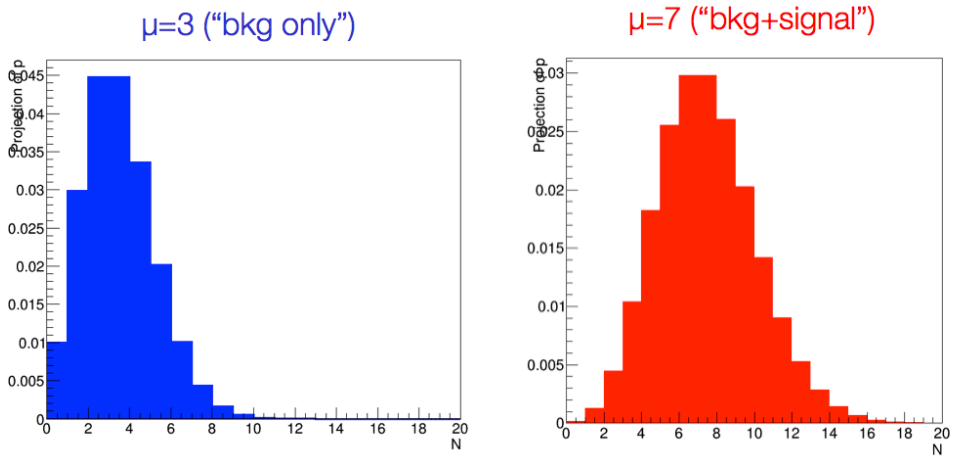
- Costs can be difficult to quantify in science.
  - What is the cost of declaring a false discovery?
  - Can be high ("Fleischman and Pons"), but hard to quantify
  - What is the cost of missing a discovery ("Nobel prize to someone else"), but also hard to quantify

How a theory becomes text-book physics



Summary on statistical test with simple hypotheses

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



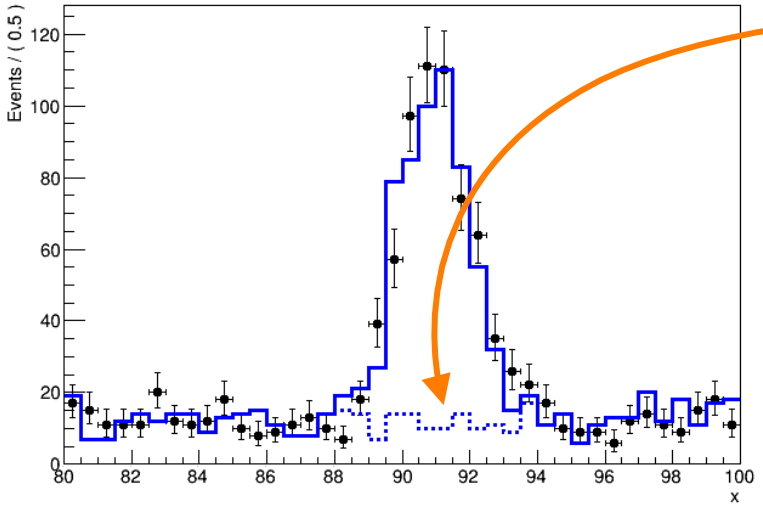
→ Given probability models P(N|bkg), and P(N|sig) we can calculate $P(N_{obs}|Hx)$ under either hypothesis

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

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

Practical statistics – (Multivariate) distributions

- Most realistic HEP analysis are not like simple counting expts at all
  - Separation of signal-like and background-like is a complex task that involves study of many observable distributions
- **How do we deal with distributions in statistical inference?**
  → Construct a probability model for the distribution
- Case 1 – Signal and background distributions from MC simulation
  - Typically have *histograms* for signal and background
  - In effect each histogram is a Poisson counting experiment
    → Likelihood for distribution is product of Likelihoods for each bin



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

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

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

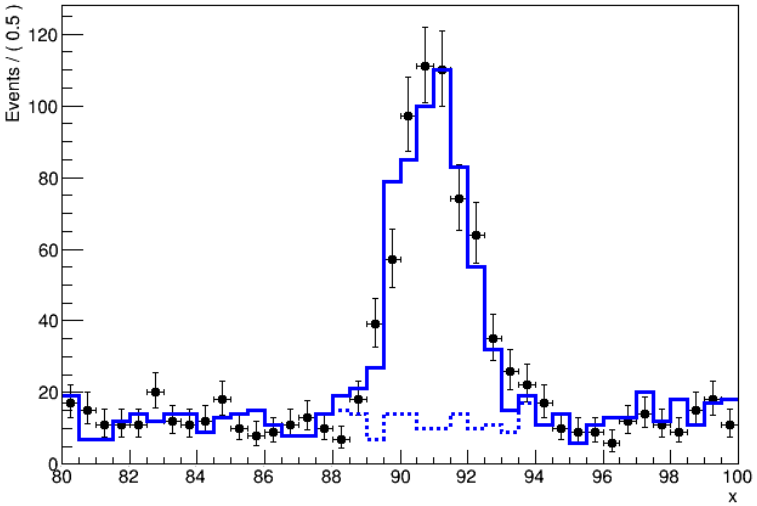
$$P(H_{s+b} \mid \vec{N}) = \frac{L(\vec{N} \mid H_{s+b})P(H_{s+b})}{L(\vec{N} \mid H_{s+b})P(H_{s+b}) + L(\vec{N} \mid H_b)P(H_b)}$$

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?**



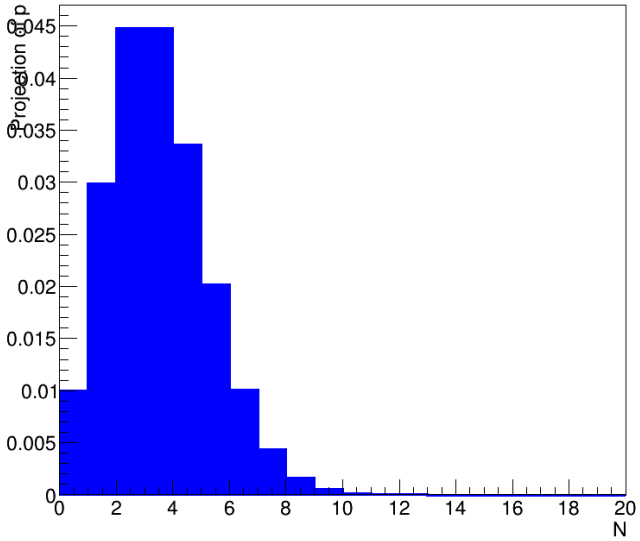
$$L(\vec{N} \mid H_b) = \prod_i Poisson(N_i \mid \tilde{b}_i)$$

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

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

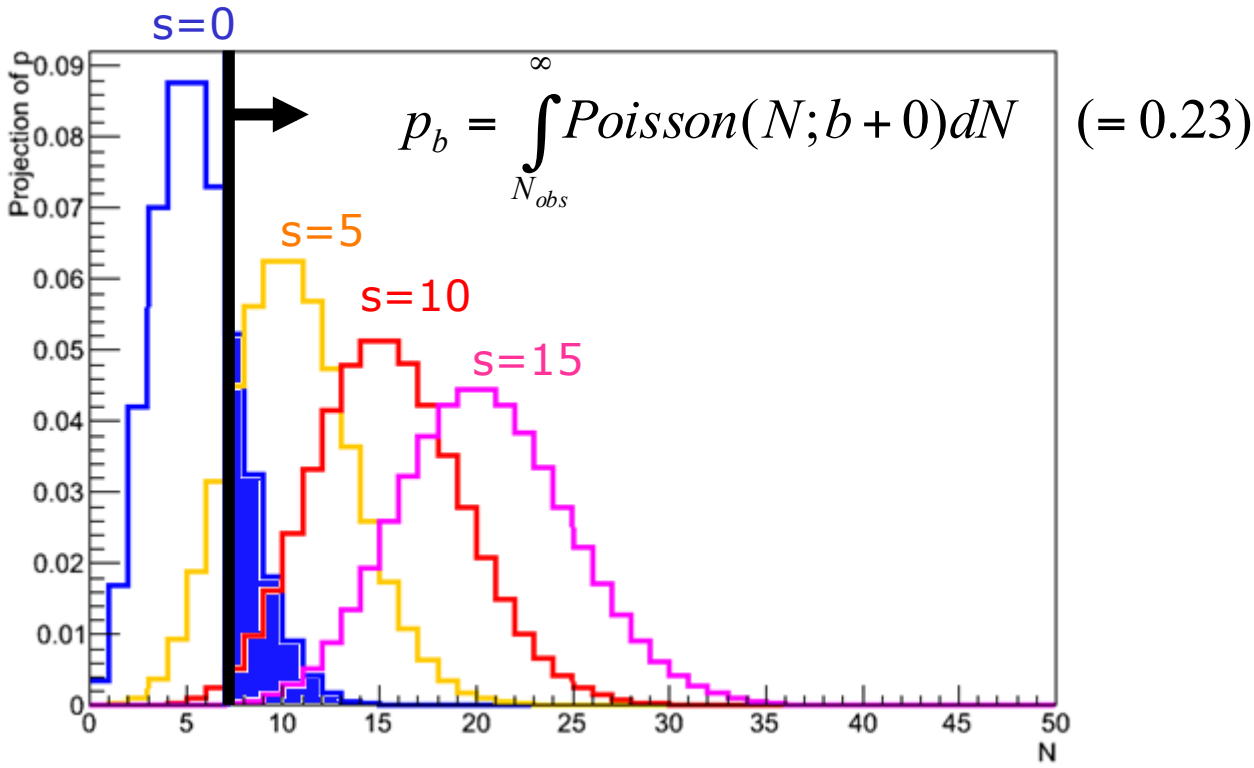


Test statistic T(N)=Nobs orders observed events count by estimated signal yield

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

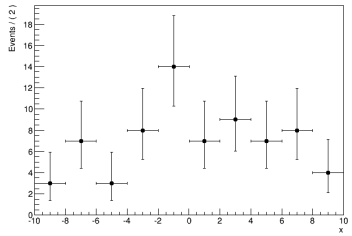
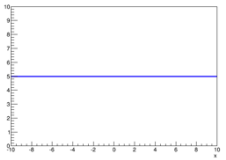
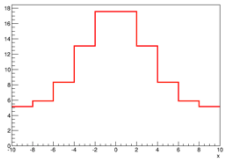
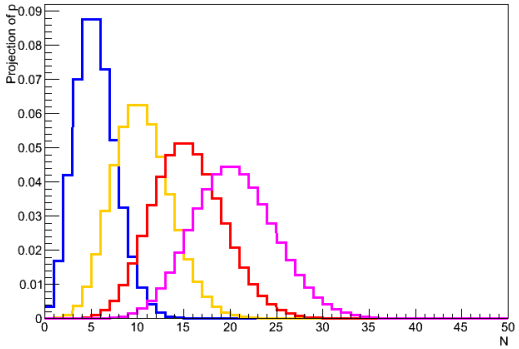
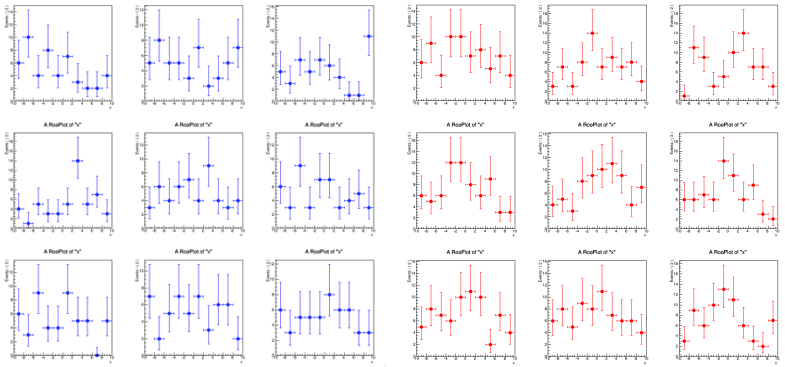
P-values for counting experiments

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



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

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

	Counting	Histogram
Observation	$N_{obs}=7$	
Median expected by hypothesis	$N_{exp}(s=0) = 5$ $N_{exp}(s=5) = 10$	 
Predicted distribution of observables		

Which histogram is more ‘extreme’?

The Likelihood Ratio as a test statistic

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

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

- Intuitive picture:

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

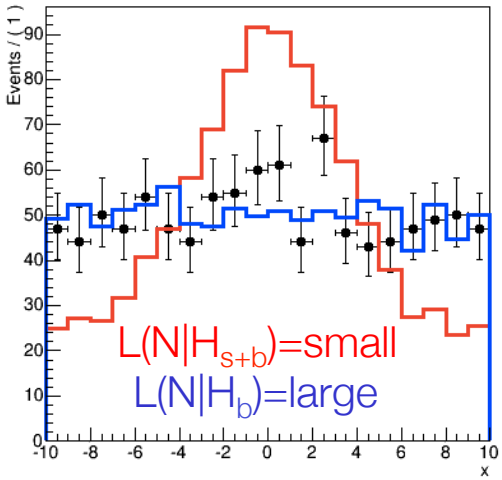
$$\lambda(\vec{N}) = \frac{\text{small}}{\text{large}} = \text{small}$$

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

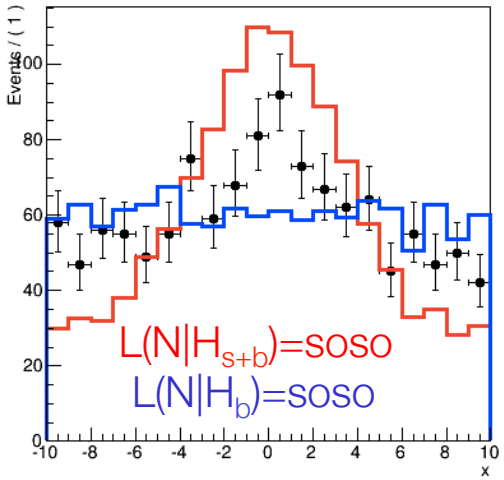
$$\lambda(\vec{N}) = \frac{\text{large}}{\text{small}} = \text{large}$$

Visualizing the Likelihood Ratio as ordering principle

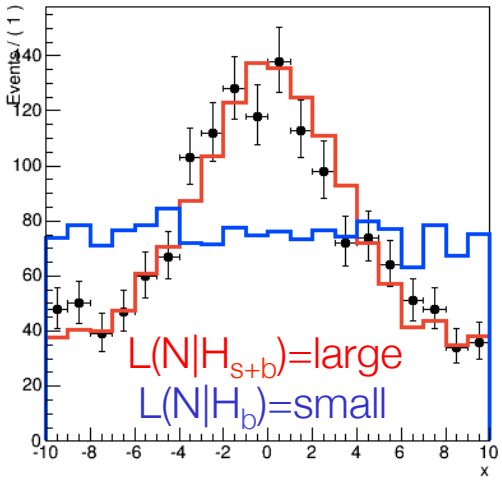
- The Likelihood ratio as ordering principle



λ(N)=0.0005



λ(N)=0.47

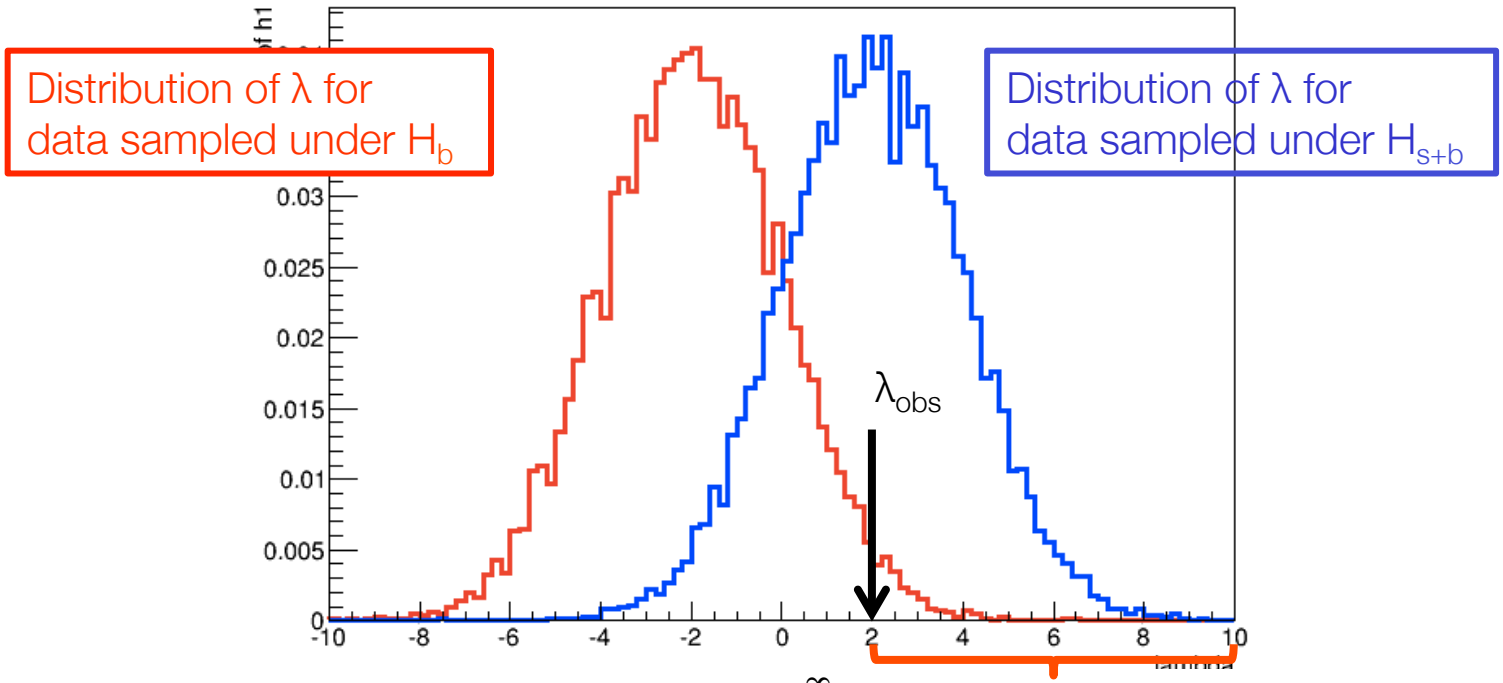


λ(N)=5000

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

The distribution of the test statistic

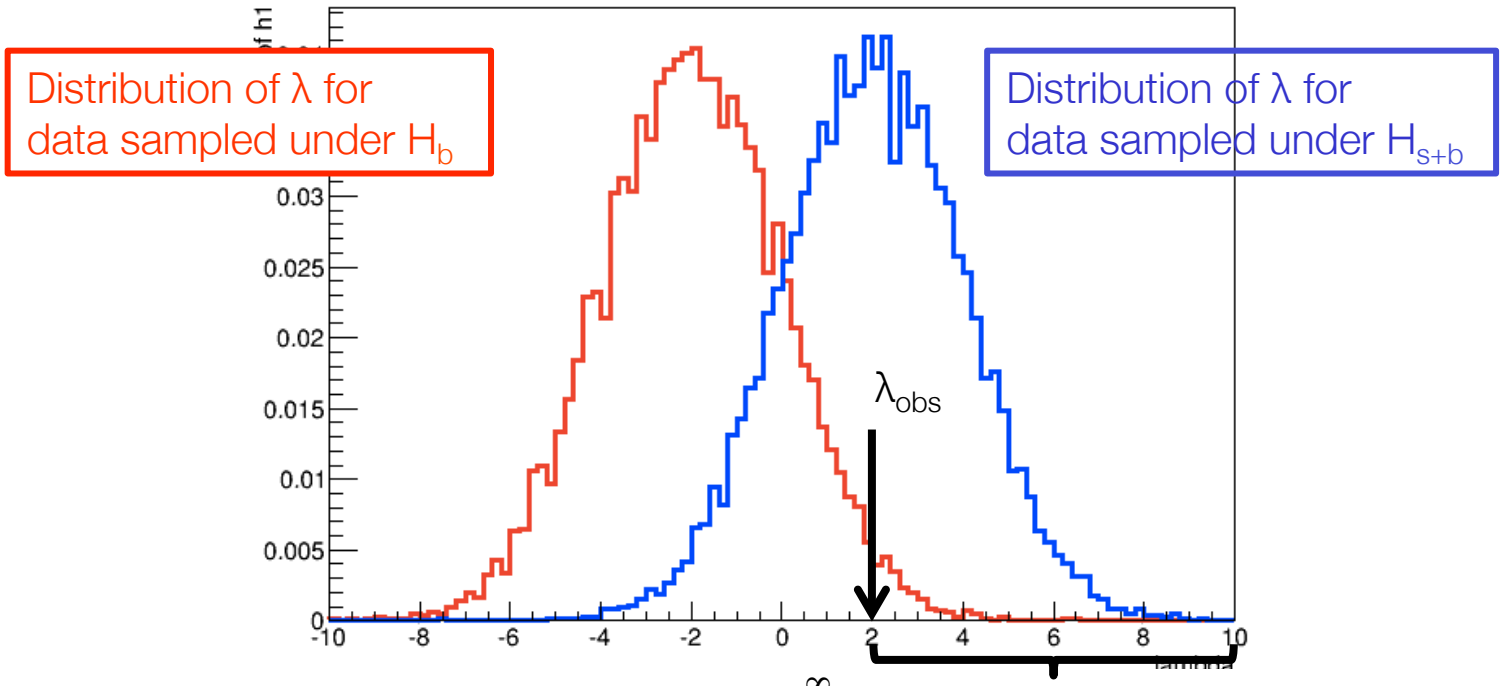
- Distribution of a test statistic is *generally not known*
- Use toy MC approach to approximate distribution
  - Generate many toy datasets N under H_b and H_{s+b} and evaluate λ(N) for each dataset



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

The distribution of the test statistic

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



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

Likelihoods for distributions - summary

- **Bayesian inference unchanged**

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

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

- **Frequentist inference procedure *modified***

  → Pure P(data|hypo) not useful for non-counting data
  → Order all possible data with a (LR) test statistic in 'extremity'
  → Quote p(data|hypo) as 'p-value' for hypothesis
  Probability to obtain observed data, *or more extreme,* is X%

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

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