

Advanced Statistics Course – Part I

W. Verkerke (NIKHEF)

Outline of this course

- Advances statistical methods – Theory and practice
- Focus on limit setting and discovery for the LHC
- Part I – Fundamentals : Tue morning
 - Interpreting probabilities
 - Bayes theorem : $P(\text{data}|\text{theo})$ vs $P(\text{theo}|\text{data})$
 - Counting experiments in detail: p-values, limits and CLs
 - The Neyman construction
 - Test statistics from likelihood ratios
 - One-side 'LHC' test statistics
- Part II – Software : Tue afternoon
 - Introduction to RooFit and RooStats
 - Model building, the workspace and the factory

Outline of this course

- Part III – Wed morning
 - Introducing nuisance parameters
 - Formulating models with nuisance parameters
 - Template morphing techniques
 - Dealing with nuisance parameters in statistical techniques
 - Frequentist vs Bayesian treatment of nuisance parameters
- Part IV – Wed afternoon
 - Expected limits
 - Asymptotic formulas for test statistics
 - Understanding and evaluating the look-elsewhere-effect
 - Constructing combinations with the Higgs as example

Probabilities and their interpretation

Introduction

- Statistics in particle physics – Quantifying our results in terms of probabilities for theoretical models and their parameters, e.g.
 - Hypothesis testing (“SM is excluded at 95% C.L.”)
 - Interval estimation (“ $170 < m(t) < 175$ at 68% C.L.”)
- Goal of this course: *how to make such statements for practical problems in particle physics*
- Precise interpretation of formulated results often surprisingly subtle
 - What is our interpretation of a probability?
 - Is our result $P(\text{data}|\text{theory})$, or $P(\text{theory}|\text{data})$?
 - Does our statement depend on $P(\text{theory})$ before the measurement?
 - Also relates to question what you want to publish? An updated ‘world view’ on the Higgs boson, or purely the result obtained from a particular experiment, to serve as (independent) ingredient for consideration on the existence of the Higgs boson?

Introduction – what we mean with probabilities

- **Two physicists meet at a bus stop in Brussels.**
After observing busses come and go for about half an hour, each is asked to make a statement on the arrival of the next bus
 - Physicist A says “I believe the next bus will come 10 +/- 5 minutes”
 - Physicist B says “The probability that a buss will pass between 10 and 20 minutes from now is 68%”
- Both physicists have a valid definition of probability (i.e. obeying Kolmogorov axioms), but have an important difference in interpretation
 - Physicist A defines probability as a (personal) degree of belief. His answer is a probability density function in his belief of the true arrival time of the bus (**Bayesian interpretation**)
 - Physicist B defines probability as a frequency in an ensemble of repeated experiments. **His answer makes no statement on the true arrival time**, which is taken as fixed but unknown. His statement is constructed such that in 68% of future observations the bus will arrive in the stated interval (**Frequentist interpretation**)
- When formulating results in terms of probabilities one should always decide which interpretation is used

Probabilities and Conditional Probabilities

- Abstract mathematical probability P can be defined in terms of sets and axioms that P obeys.
- If the axioms are true for P , then P obeys Bayes' Theorem

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

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



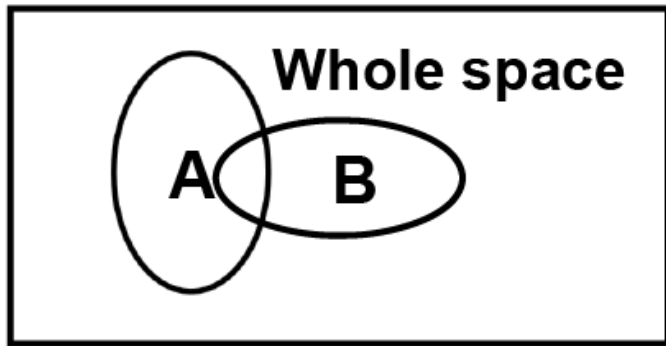
$$P(T|D) = P(D|T) P(T) / P(D) \leftarrow \text{(normalization term)}$$

What is $P(\text{(no)Higgs})$?

This is result of our experiment:
 $P(\text{LHCdata}|\text{(no)Higgs})$

This is what we usually *want* to know:
 $P(\text{(no)Higgs}|\text{LHCdata})$

Bayes' Theorem in Pictures



$$P(A) = \frac{\text{Area of A}}{\text{Area of Whole space}}$$

$$P(B) = \frac{\text{Area of B}}{\text{Area of Whole space}}$$

$$P(A|B) = \frac{\text{Area of } A \cap B}{\text{Area of B}}$$

$$P(B|A) = \frac{\text{Area of } A \cap B}{\text{Area of A}}$$

$$P(A \cap B) = \frac{\text{Area of } A \cap B}{\text{Area of Whole space}}$$

$$P(A) \times P(B|A) = \frac{\text{Area of A}}{\text{Area of Whole space}} \times \frac{\text{Area of } A \cap B}{\text{Area of B}} = \frac{\text{Area of } A \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

$$P(B) \times P(A|B) = \frac{\text{Area of B}}{\text{Area of Whole space}} \times \frac{\text{Area of } A \cap B}{\text{Area of A}} = \frac{\text{Area of } A \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

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

What is the “Whole Space”?

- Note that for probabilities to be well-defined, the “whole space” needs to be defined, which in practice introduces assumptions and restrictions.
- Thus the “whole space” itself is more properly thought of as a conditional space, conditional on the assumptions going into the model (Poisson process, whether or not total number of events was fixed, etc.).
- Furthermore, it is widely accepted that restricting the “whole space” to a relevant subspace can sometimes improve the quality of statistical inference.

Intuitive examples of $P(A|B) \neq P(B|A)$

- Intuitive

$$P(\text{pregnant}|\text{woman}) \neq P(\text{woman}|\text{pregnant})$$

$$P(\text{sunny}|\text{taking photos}) \neq P(\text{taking photos}|\text{sunny})$$

- Less intuitive...

What we say

What ends up in the news paper

$$P(\text{data}|\text{no-higgs}) \neq 1 - P(\text{higgs}|\text{data})$$

Using Bayes theorem

- **$P(T|D) = P(D|T) P(T) / P(D)$**
- Simplest possible experiment: a flue test.
- **Two completing hypotheses: have flue, don't not flue**
- **Suppose we know $P(D|T)$:**
 - $P(D=+|T=+) = 0.98$ (98% of flue cases correctly detected)
 - $P(D=-|T=-) = 0.99$ (99% of healthy people diagnosed healthy)
 - $P(D=-|T=+) = 0.02$, $P(D=+|T=-) = 0.01$
- **Observation: $D=+$ (I test positive)**
- **Question: What are odds I have flue, i.e. $P(T=+|D=+)$**
- **Answer: Can't be answered without $p(T=+)$!**

Using Bayes theorem

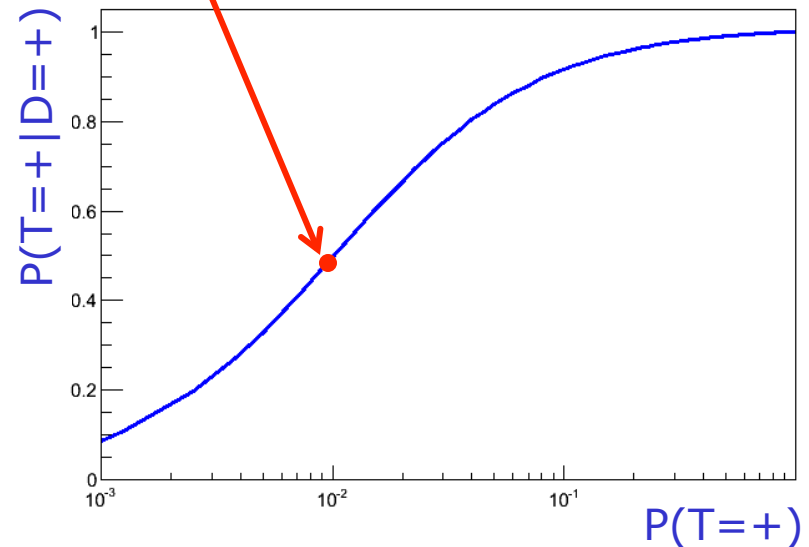
- Question: **What are odds I have flu**, i.e. $P(T=+|D=+)$
- **Suppose $P(T=+)=0.01$** (1% of population has flu), then

$$\begin{aligned} - P(D) &= P(D=+|T=-) * P(T=-) + P(D=+|T=+) * P(T=+) \\ &= 0.01 * 0.99 + 0.98 * 0.01 \\ &= 0.0197 \end{aligned}$$

$$\begin{aligned} - P(T=+|D=+) &= P(D=+|T=+) P(T=+) / P(D=+) \\ &= 0.98 * 0.01 / 0.0197 = 0.497 \end{aligned}$$

- **You can get any answer between 0 and 1 depending on choice of $P(T)$...**

- Makes it difficult to see $P(T|D)$ as an objective summary of your measurement
- But it does provide a coherent framework to update your from before the test $P(T)$ to after the test $P(T|D)$



A Note re *Decisions*

- Suppose that as a result of the previous experiment, your degree of belief in the model is $P(\text{have flu}|\text{positive test}) = 99\%$, and you need to *decide whether or not to take an action*
 - *Visit doctor, cancel vacation etc...*
- **Question: What should you decide?**
- Answer: *Cannot be determined from the given information!*
 - Need in addition: the utility function (or cost function), which gives 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).
- 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.

What do we want to report?

- You cannot state $P(\text{theory}|\text{data})$ without $P(\text{theory})$
 $P(\text{theory}|\text{data})$ may answer the question: “does the Higgs exist?”, but it cannot summarize your experiment result without prior assumptions on $P(\text{theory})$
- Also statements on $P(\text{theory}|\text{data})$ often (but not always) restricted to Bayesian interpretation of probability, as frequentist formulation of $P(\text{theory})$ often impossible
 - Flu test OK $\rightarrow p(\text{theo})$ is fraction of population with flue
 - B-tagging: $p(\text{b-jet}|\text{b-tag}) = p(\text{b-tag}|\text{b-jet})p(\text{b-jet})$
OK $\rightarrow p(\text{b-jet})$ is fraction of all jets that are b-jets
 - Higgs discovery: $p(\text{Higgs}|\text{LHCdata}) = p(\text{LHCdata}|\text{Higgs})p(\text{Higgs})$
Not OK $\rightarrow p(\text{Higgs})$ not defineable as a frequency (there is only one universe)
- In HEP, results are often stated as $P(\text{data}|\text{theo})$ with a Frequentist interpretation of probabilities

What Can Be Computed without Using a Prior?



- *Not* $P(\text{constant of nature} \mid \text{data})$.
 1. *Confidence Intervals* for parameter values, as defined in the 1930's by Jerzy Neyman.
 2. *Likelihood ratios*, the basis for a large set of techniques for point estimation, interval estimation, and hypothesis testing.
- These can both be constructed using frequentist definition of P .

Simple versus composite hypothesis

- Flu example involved a simple hypothesis
- In particle physics composite hypothesis are most common
 - Simply hypothesis 'Standard Model with Higgs boson'
 - Composite hypothesis 'Standard Model with Higgs boson with production cross-section of X pb' where X is a parameter
- Introduces one (or more) model parameters in the hypothesis we're testing and in our statement of the final result
- Can test hypothesis for any value of X , e.g.

- $P(\text{data}|\text{theory}(x=35)) = 3\%$
- $P(\text{data}|\text{theory}(x=20)) = 5\%$
- $P(\text{data}|\text{theory}(x=10)) = 20\%$
- $P(\text{data}|\text{theory}(x=5)) = 70\%$


Hypothesis testing

 $X < 20$ at 95% C.L.

Confidence intervals
(can be constructed without $p(\text{theory})!$)

Understanding the Poisson counting experiment

P-values and limits for counting experiments

- This morning we will focus on the simplest of measurements: counting experiments
 - Data: Event count collected (N) in a fixed time frame:
 - Theory: the expected distribution for N for repeated measurements
 - We assume (for know) an exact prediction for the number of background events: $b=5$

- The observed number n will follow a Poisson distribution:

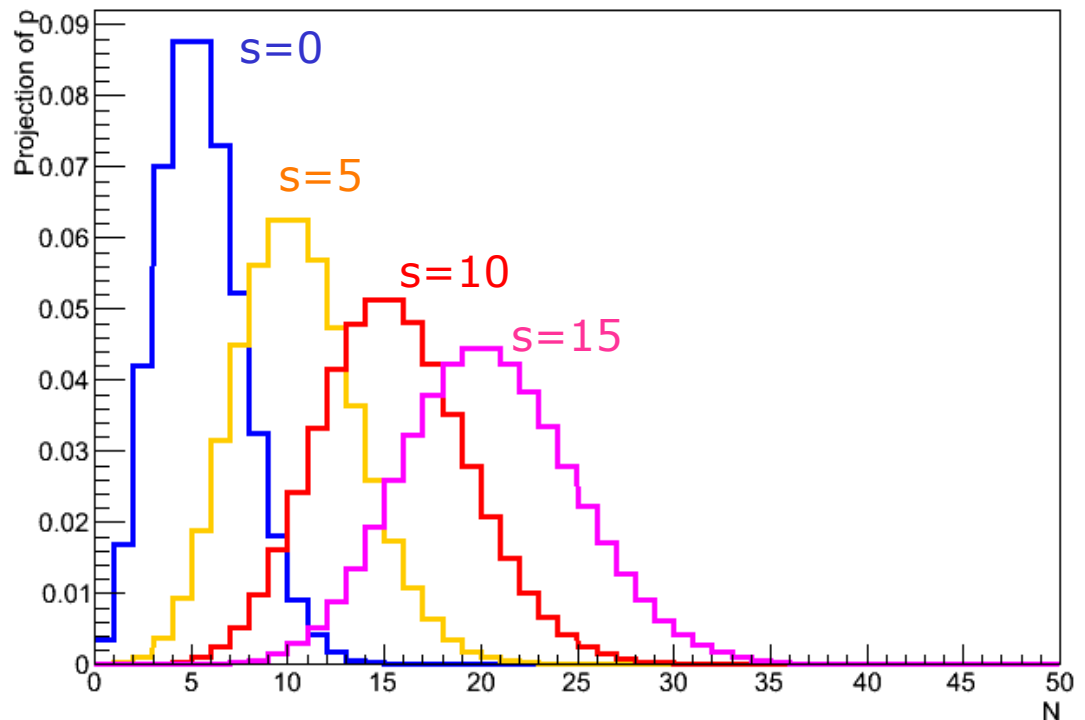
$$P(n|b) = \frac{b^n}{n!} e^{-b} \qquad P(n|s+b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

- The relevant hypotheses are
 - H_0 : all events are of the background time
 - H_1 : the events are a mixture of signal and background
- Rejecting H_0 with $Z > 5$ constitutes discovery of signal

P-values and limits for counting experiments

- Suppose we measure $N=15$
 - Did we discover signal (at $Z=5$)?
- Suppose we measure $N=7$
 - What signal strengths can we exclude
- Suppose we measure $N=2$
 - What do we learn from that?

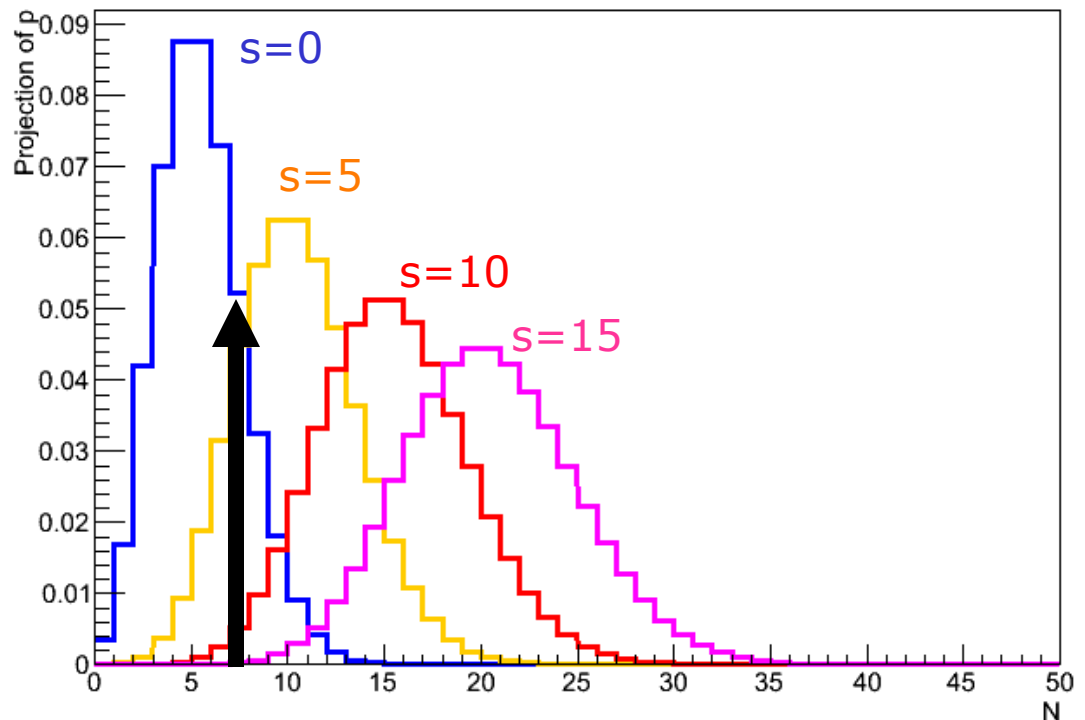
- Given a Poisson distribution with $\mu = s + (b=5)$,
 - we expect the following distributions for N_{obs} for $S=0, S=5, S=10,$



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

- Now make a measurement $N=N_{\text{obs}}$ (example $N_{\text{obs}}=7$)
 - $P(N_{\text{obs}}=7|T(s=0)) = \text{Poisson}(7;5) = 0.104$
 - $P(N_{\text{obs}}=7|T(s=5)) = \text{Poisson}(7;10) = 0.090$
 - $P(N_{\text{obs}}=7|T(s=10)) = \text{Poisson}(7;15) = 0.010$
 - $P(N_{\text{obs}}=7|T(s=15)) = \text{Poisson}(7;20) = 0.001$

- This is great feature of simple counting experiments:
for each observation $P(D|T)$ can be trivially calculated

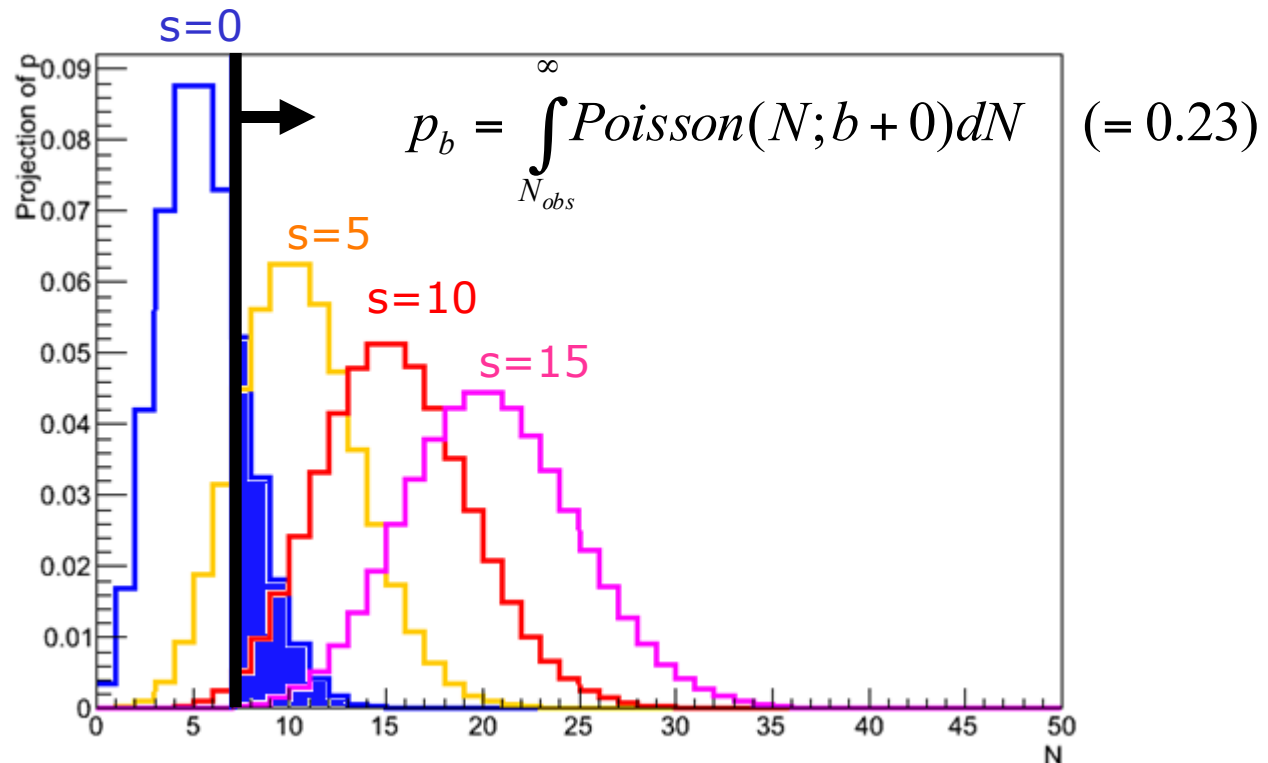


Interpreting $N_{\text{obs}}=7$

- Formulating discovery more precisely:

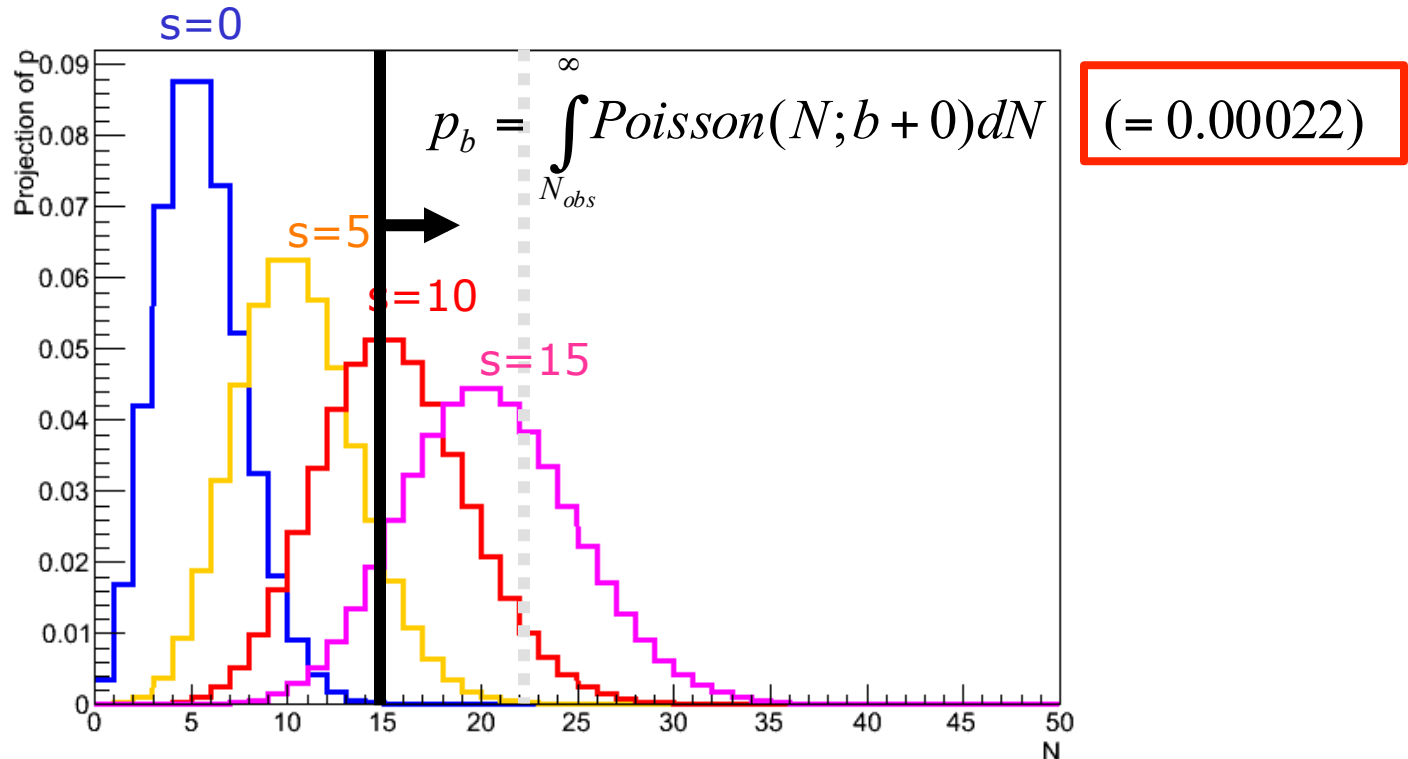
“What fraction of future measurements would result in 7 *or more* events, if the bkg-only hypothesis is true”

= ‘p-value of background hypothesis’



Interpreting N=15

- Another example: $N_{obs}=15$ for same model, what is the p-value for the background?



- Result customarily re-expressed as odds of a *Gaussian fluctuation with equal p-value* (3.5 sigma for above case)

$N_{obs}=22$ gives $p_b < 2.8 \cdot 10^{-7}$ ('5 sigma')

At what p-value does one declare discovery?

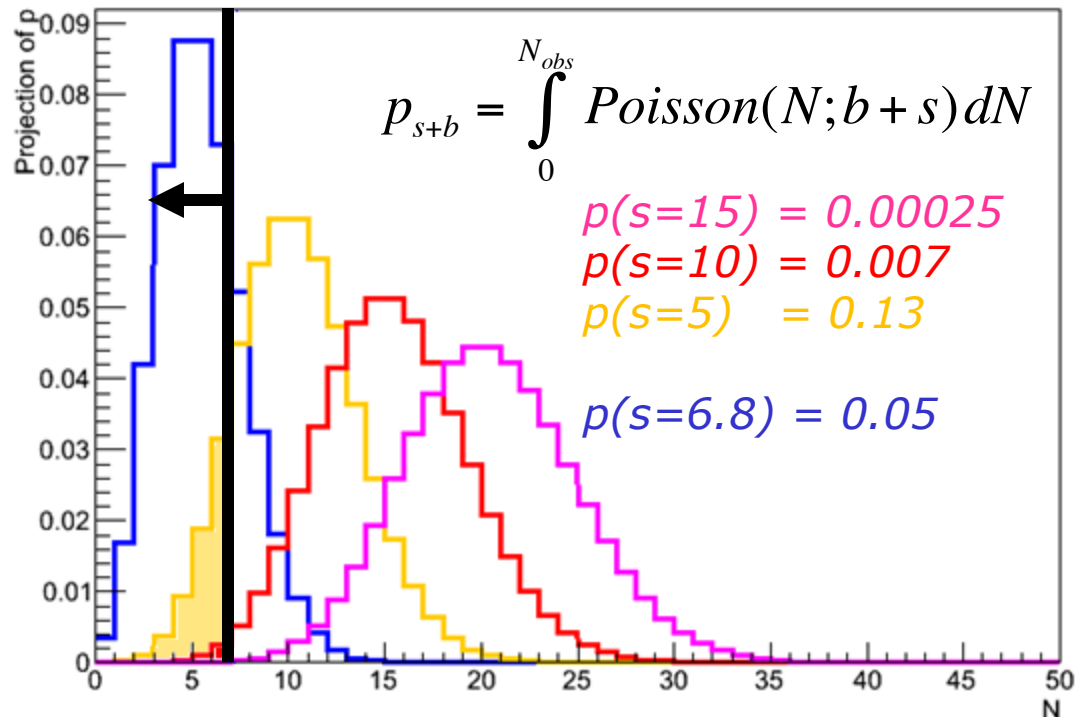
- HEP folklore: claim discovery when p -value of background only hypothesis is 2.87×10^{-7} , corresponding to significance $Z = 5$.
- This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

<u>phenomenon</u>	<u>reasonable p-value for discovery</u>
D ⁰ D ⁰ mixing	~0.05
Higgs	~10 ⁻⁷ (?)
Life on Mars	~10 ⁻¹⁰
Astrology	~10 ⁻²⁰

- Cost of type-I error (false claim of discovery) can be high
 - Remember cold nuclear fusion 'discovery'

Upper limits (one-sided confidence intervals)

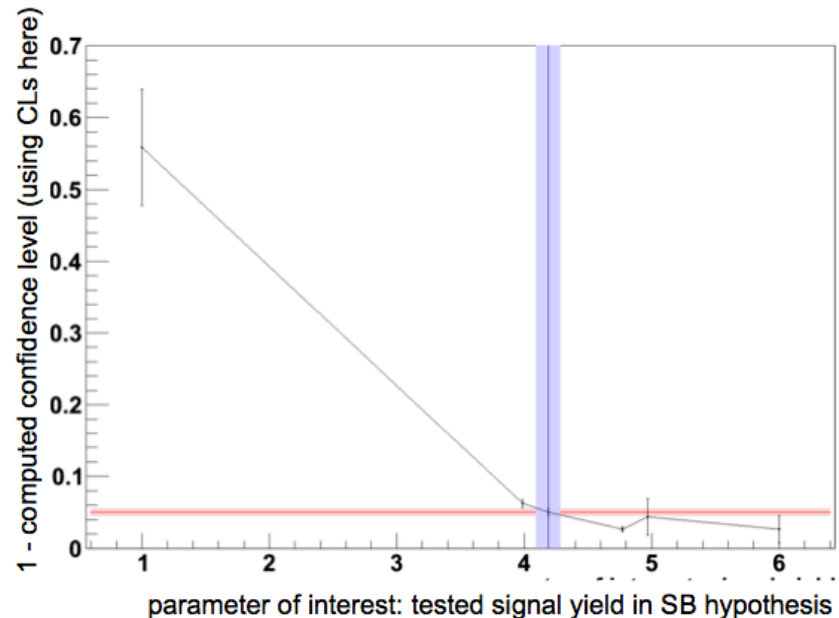
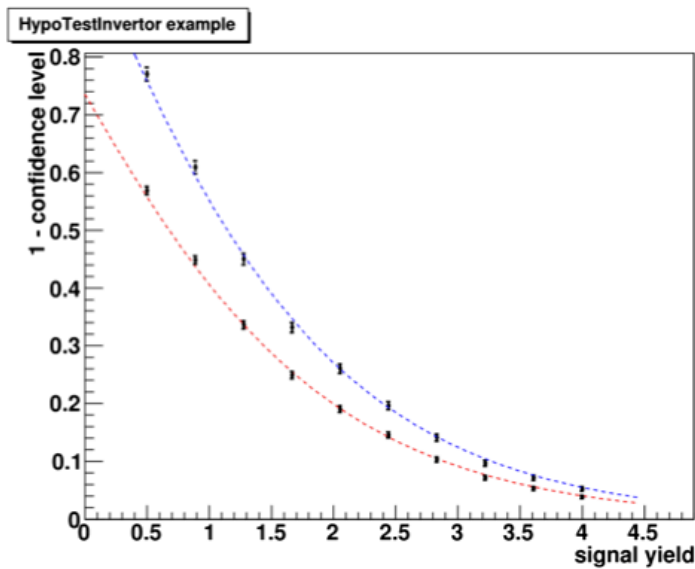
- Can also define p-values for **hypothesis with signal**: p_{s+b}
 - Note convention: integration range in p_{s+b} is flipped



- Convention: express result as value of **s** for which $p(s+b)=5\% \rightarrow$ **"s>6.8 at 95% C.L."**

Upper limits (one-sided confidence intervals)

- Procedure of scanning for the value of s so that $p(s+b)=5\%$ is called "Hypothesis Test Inversion" and invariably involves some numerical method to find that point.
- Can scan 'by hand' in fixed steps [left] in signal yield, or develop smart iterative algorithm [right]



Interpreting $N=1$

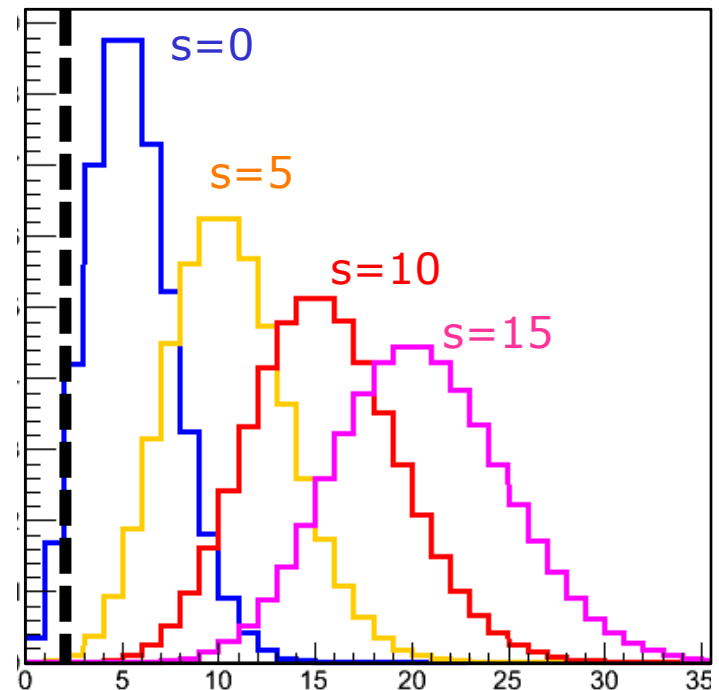
- Need to be careful about interpretation $p(s+b)$ in terms of inference on signal only
 - Since $p(s+b)$ quantifies consistency of signal *plus* background
 - Problem most apparent when observed data has **downward stat. fluctuations w.r.t background** expectation

- Example: $N_{\text{obs}} = 1$

$$\rightarrow p_{s+b}(s=0) = 0.04$$

$s \geq 0$ excluded at >95% C.L. ?!

- 'Spurious exclusion' due to weak sensitivity
 - For low s , distributions for s and $s+b$ are very similar

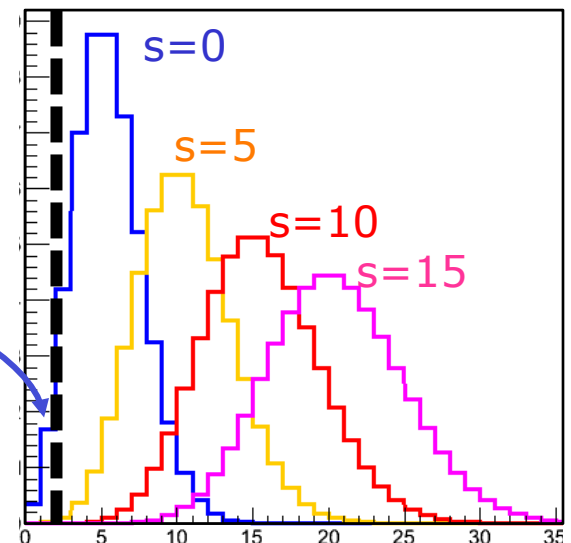


Interpreting N=1 (continued)

- Not problematic in strict frequentist interpretation – we expect this result in 5% of the experiments, but complicates interpretation of result in terms of signal
- **Problem is that we know that s must be ≥ 0**
 - In a Bayesian approach we construct $P(t|d)=P(d|t)P(t)$ and we can include this prior knowledge on s in $p(t)$, e.g. $p(\text{theory})=0$ for $s<0$
 - In Frequentist approach we don't want to use (or formulate) $p(\text{theory})$, so how incorporate this in our result?
- Current LHC solution is called 'CLs'
 - Instead of $p(s+b)$ base test on

$$CL_S \equiv \frac{p_{s+b}}{1 - p_b} = 5\% \text{ (e.g.)}$$

- If observation is also unlikely under bkg-only hypothesis, net effect is increased limit on s (in areas of low sensitivity)
- For $N_{\text{obs}}=1$ exclude $s>3.4$ at 95% (instead of $s>0$)



Modified frequentist approach

- The CL_S method is sometimes also referred to as the 'modified' frequentist approach
- Note that CL_S is a HEP invention (at the time of the LEP experiments), it is not used in the professional statistics literature
- Within HEP it is generally accepted, and is the current recommendation for ATLAS/CMS (Higgs) results, but alternative prescriptions exist with similar effect, e.g.
 - Power constrained limits
 - Feldman-Cousins

The Neyman construction

Introduction

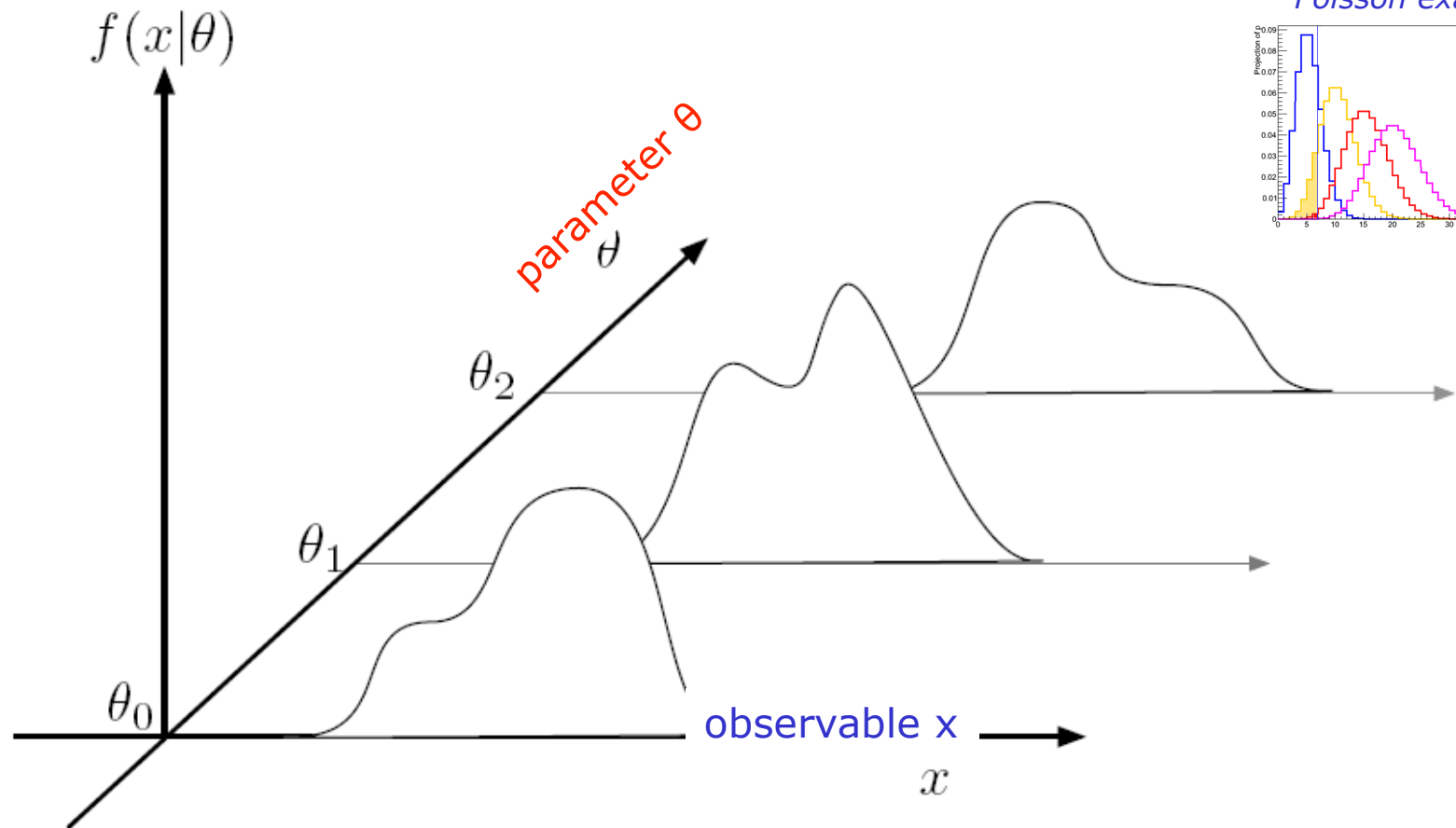
- Poisson counting experiment illustrates procedures to compute frequentist statements
 - P-value: probability that background hypothesis produces observed result, or or more extreme ('discovery')
 - Upper limit: signal strength that has that has a preset probability (usually 5%) produce observed result, or less extreme
→ $s < XX$ at 95% C.L.
- Upper limit is a special case of a frequentist confidence interval
 - Interval is $[0,XX]$ for the above case
- Next: the 'Neyman construction' – a prescription to construct confidence intervals for any measurement

Confidence Intervals

- “**Confidence intervals**”, and this phrase to describe them, were invented by Jerzy Neyman in 1934-37.
 - While statisticians mean Neyman’s intervals (or an approximation) when they say “confidence interval”, in HEP the language tends to be a little loose.
 - Recommend using “confidence interval” only to describe intervals corresponding to Neyman’s construction (or good approximations thereof), described below.
- The slides contain the crucial information, but you will want to cycle through them a few times to “take home” how the construction works, since it is really ingenious – perhaps a bit *too* ingenious given how often confidence intervals are misinterpreted.
- In particular, you will understand that the confidence level does *not* tell you “how confident you are that the unknown true value is in the interval” –only a *subjective* Bayesian credible interval has that property!

How to construct a Neyman Confidence Interval

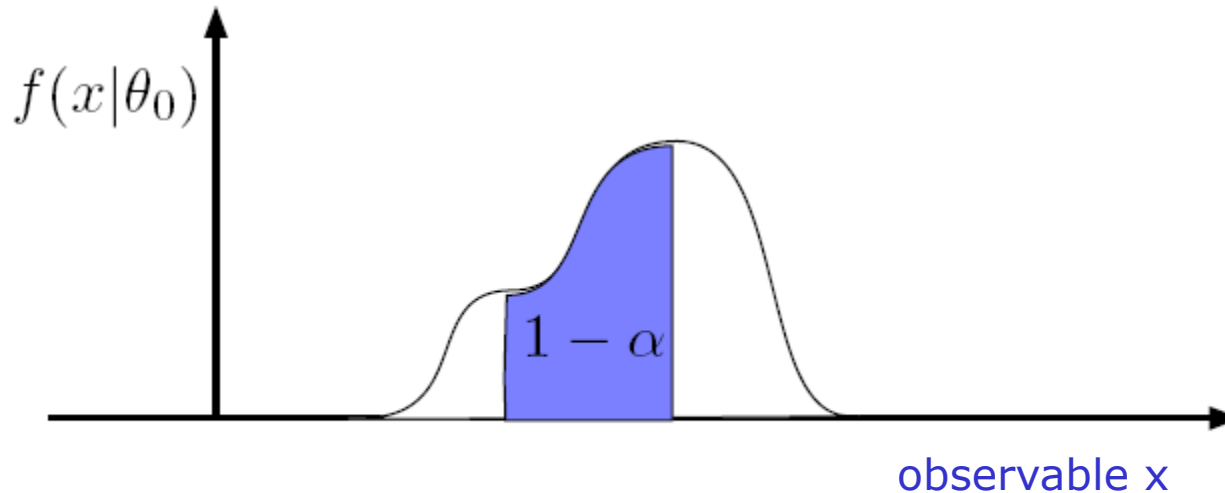
- Simplest experiment: one measurement (x), one theory parameter (θ)
- For each value of **parameter θ** , determine distribution in in **observable x**



How to construct a Neyman Confidence Interval

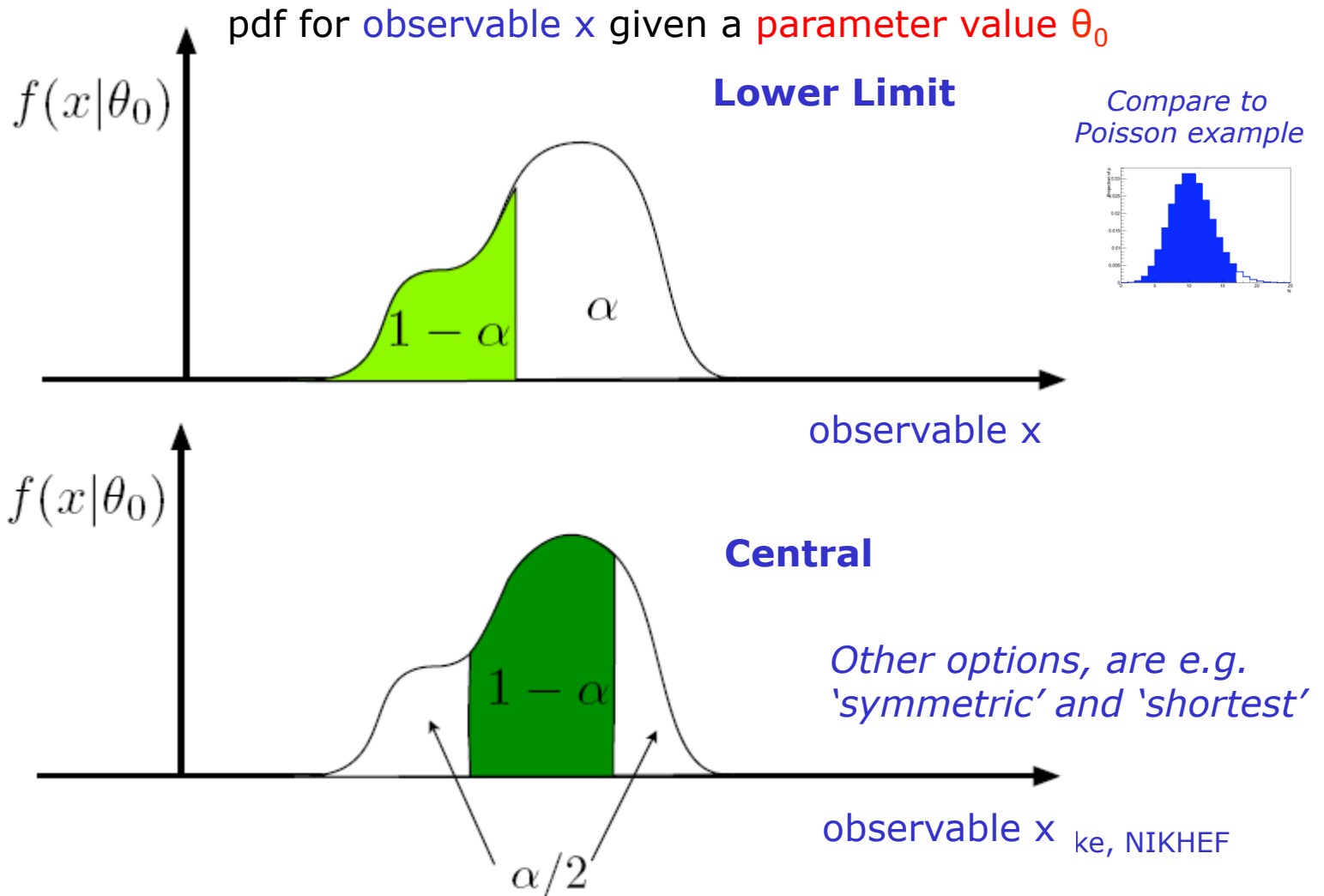
- Focus on a slice in θ
 - For a $1-\alpha\%$ confidence Interval, define **acceptance interval** that contains $100\%-\alpha\%$ of the probability

pdf for **observable** x
given a **parameter value** θ_0



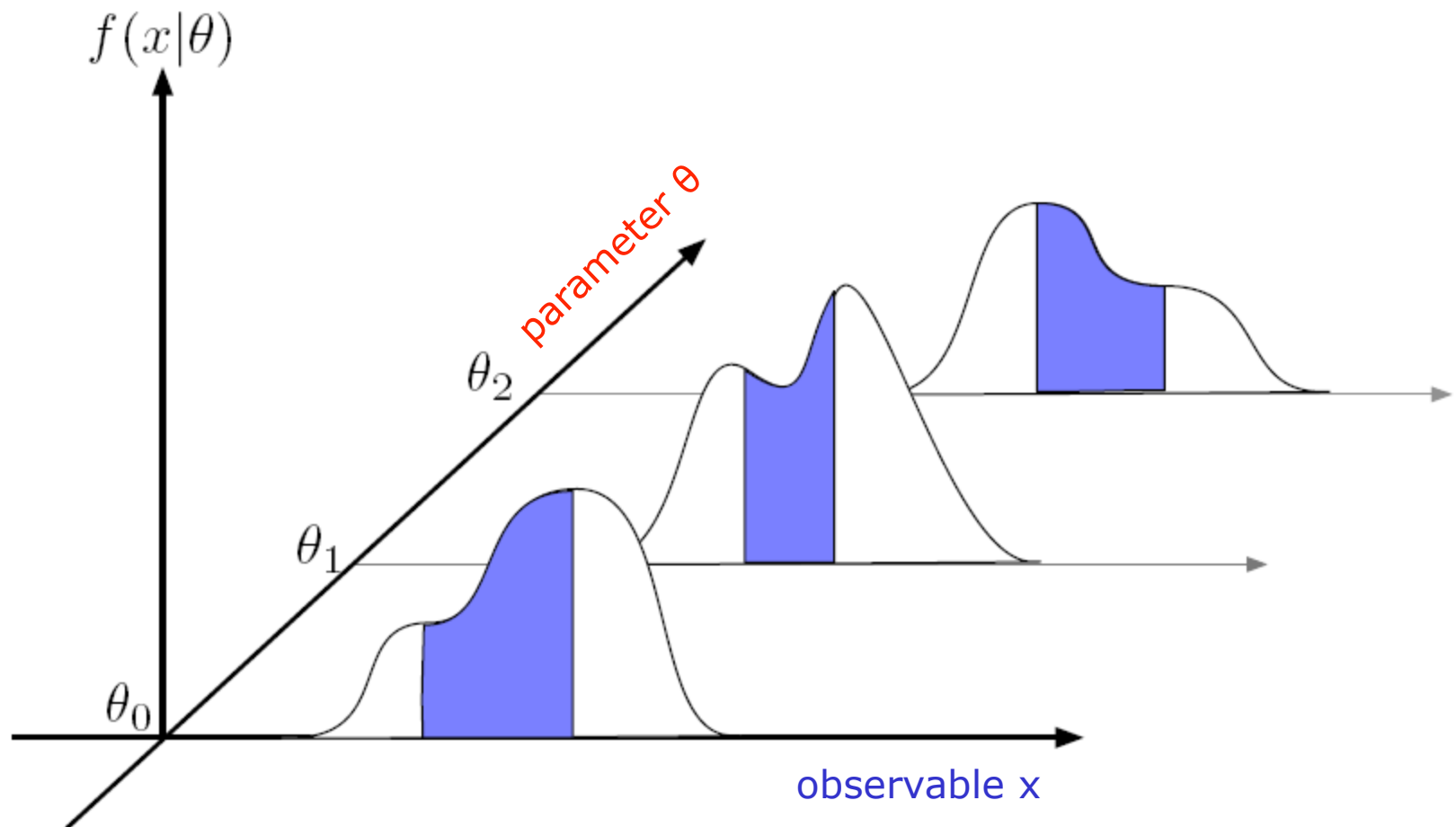
How to construct a Neyman Confidence Interval

- Definition of acceptance interval is not unique
 - Algorithm to define acceptance interval is called '**ordering rule**'



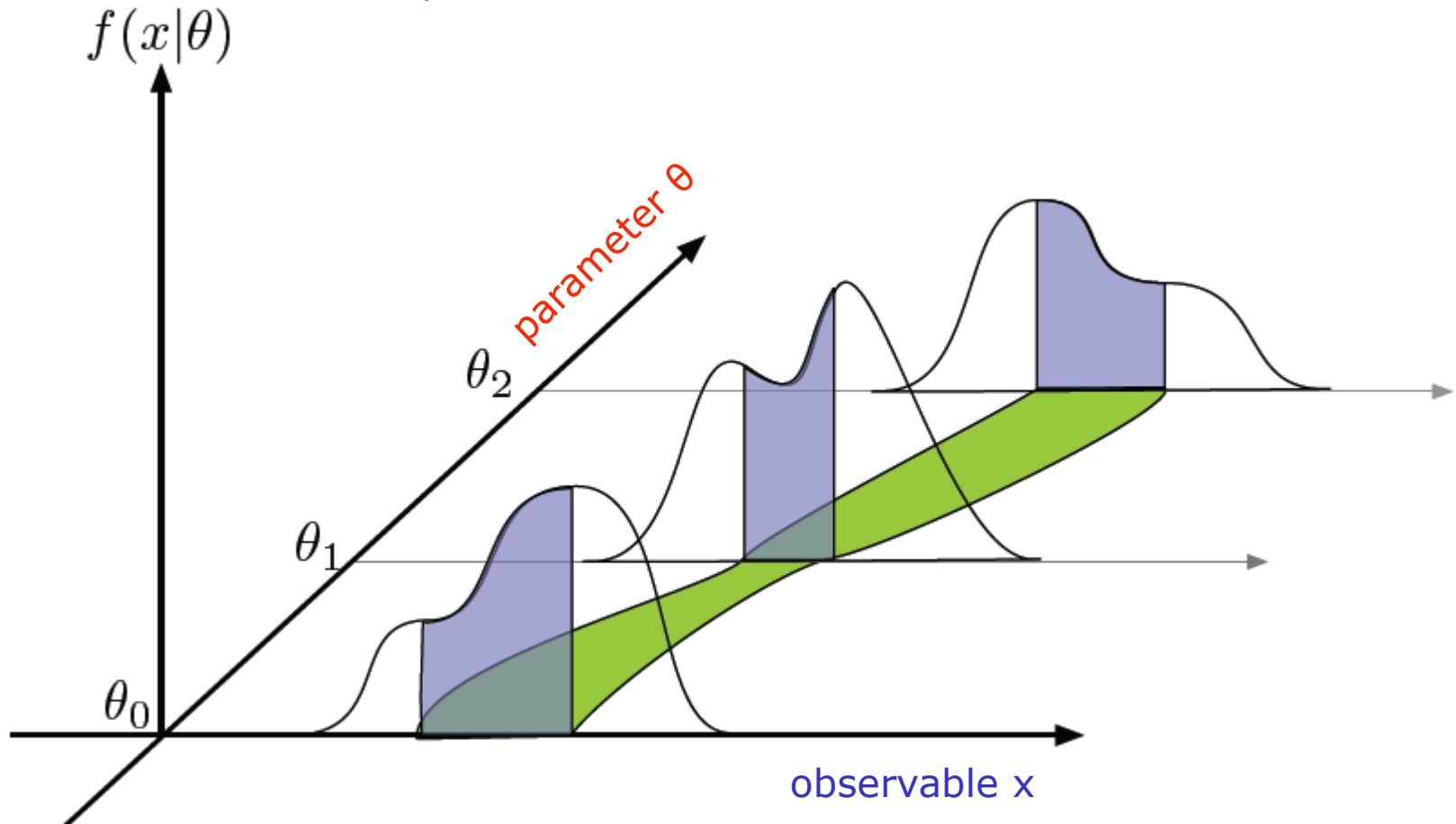
How to construct a Neyman Confidence Interval

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



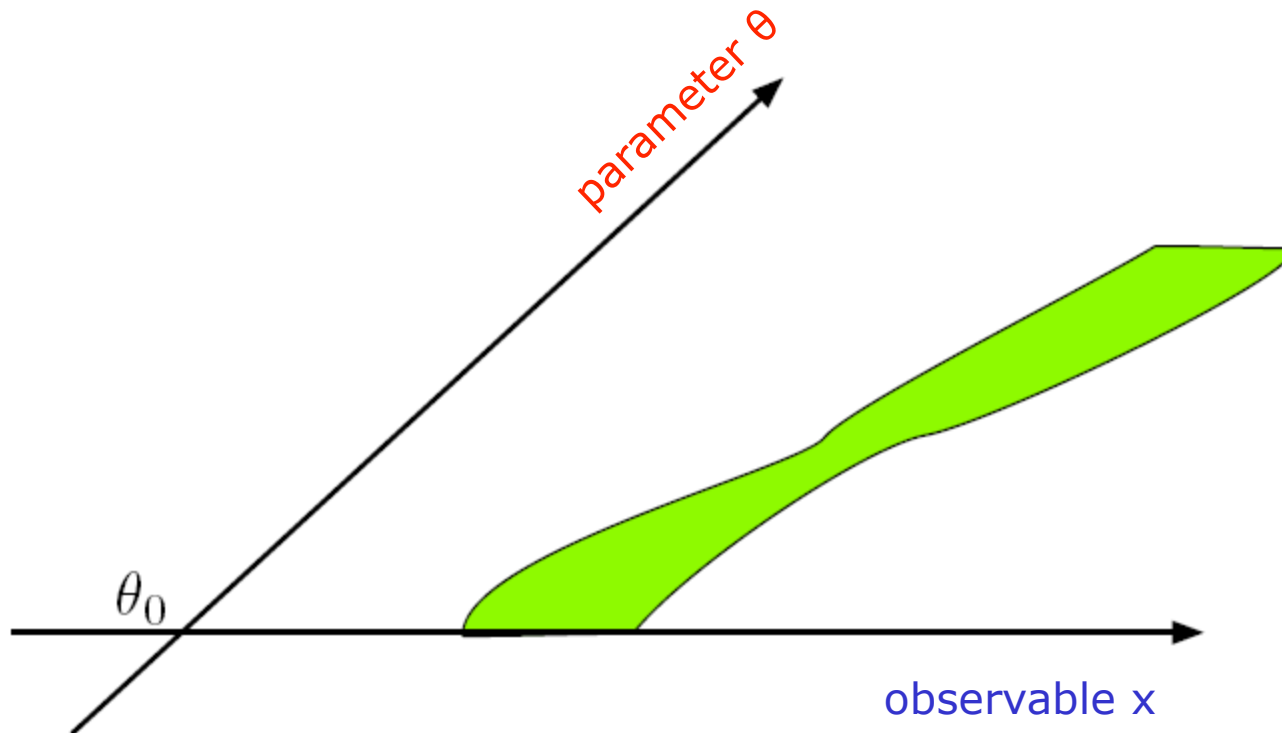
How to construct a Neyman Confidence Interval

- This makes the confidence belt
 - The region of data in the confidence belt can be considered as consistent with **parameter θ**



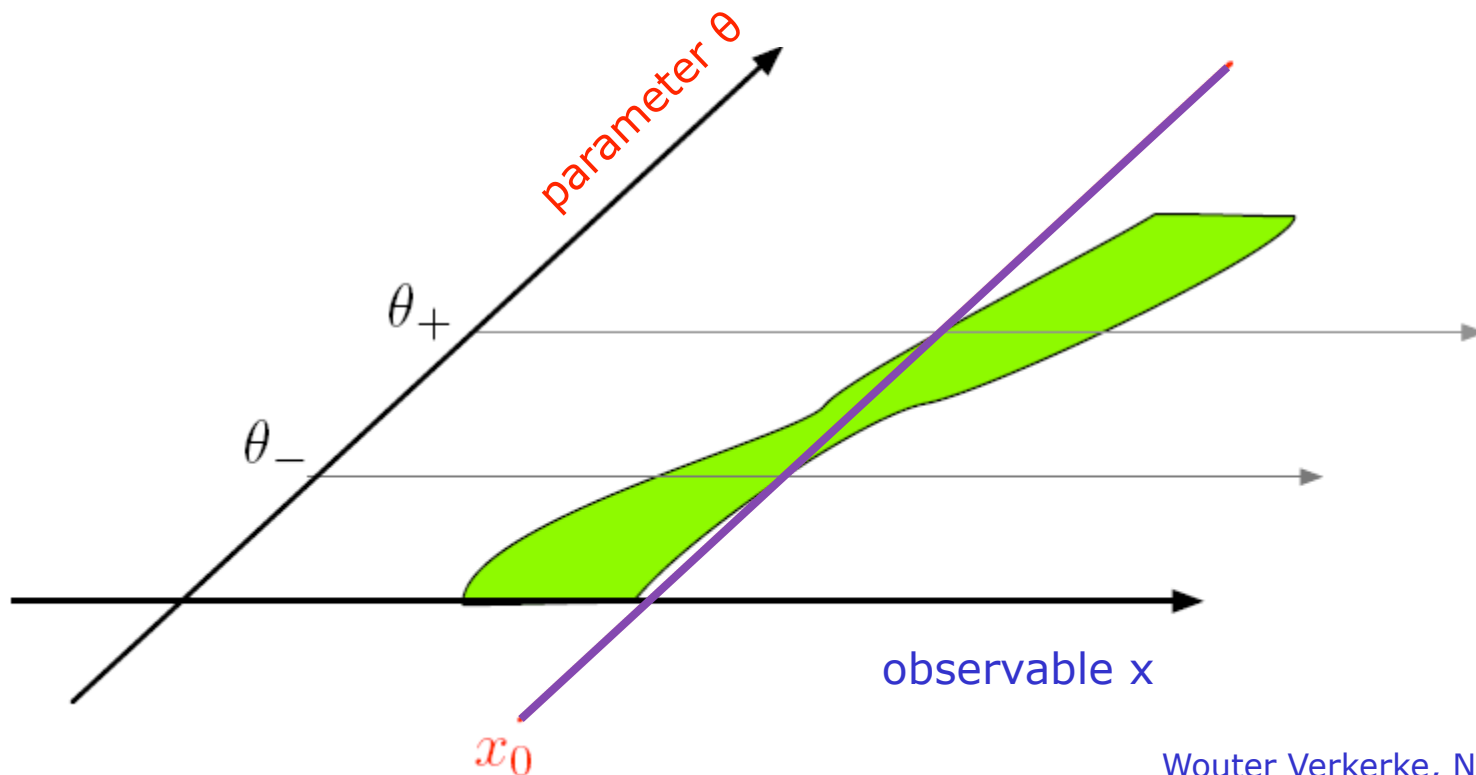
How to construct a Neyman Confidence Interval

- This makes the confidence belt
 - The region of data in the confidence belt can be considered as consistent with **parameter θ**



How to construct a Neyman Confidence Interval

- The confidence belt can be 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



Confidence interval – summary

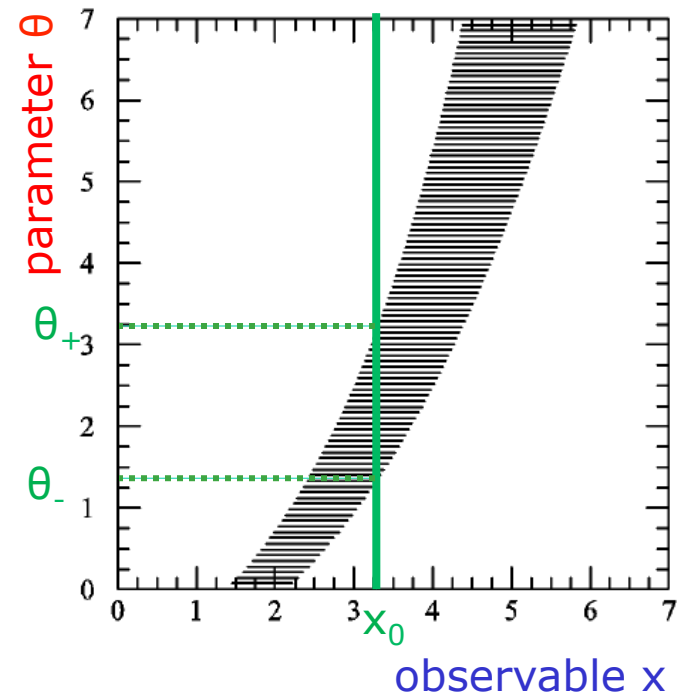
- *Note that this result does NOT amount to a probability density distribution in the true value of θ*
- Let the unknown true value of θ be θ_t .

In repeated expt's, the confidence intervals obtained will have different endpoints $[\theta_1, \theta_2]$, since the endpoints are functions of the randomly sampled x .

A little thought will convince you that a fraction C.L. = $1 - \alpha$ of intervals obtained by Neyman's construction will contain ("cover") the fixed but unknown μ_t . i.e.,

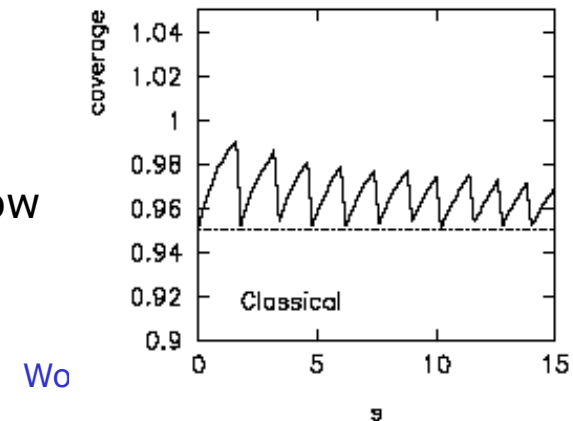
$$P(\theta_t \in [\theta_1, \theta_2]) = \text{C.L.} = 1 - \alpha.$$

- The random variables in this equation are θ_1 and θ_2 , and not θ_{tr}
- Coverage is a property of the set, not of an individual interval!
- It *is* true that the confidence interval consists of those values of θ for which the observed x is among the most probable to be observed.
 - In precisely the sense defined by the ordering principle used in the Neyman construction



Coverage

- Coverage = Calibration of confidence interval
 - Interval has coverage if probability of true value in interval is $\alpha\%$ for all values of μ
 - It is a property of the procedure, not an individual interval
- **Over-coverage** : probability to be in interval $>$ C.L.
 - Resulting confidence interval is conservative
- **Under-coverage** : probability to be in interval $<$ C.L.
 - Resulting confidence interval is optimistic
 - **Under-coverage is undesirable \rightarrow You may claim discovery too early**
- Exact coverage is difficult to achieve
 - For Poisson process impossible due to discrete nature of event count
 - “Calibration graph” for preceding example below



Exact Coverage = Fixing the 'type I error rate'

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

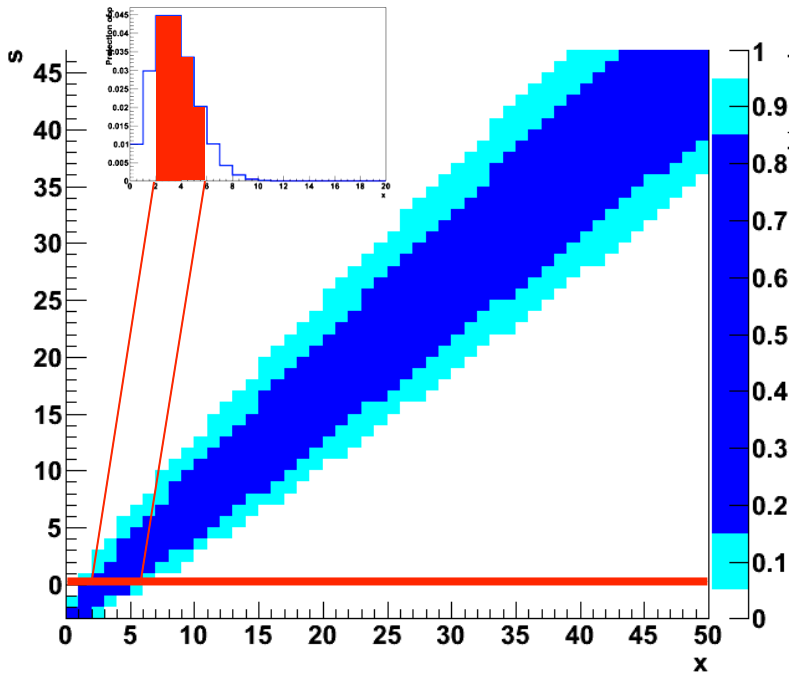
		Actual condition	
		Guilty	Not guilty
Decision	Verdict of 'guilty'	True Positive	False Positive (i.e. guilt reported unfairly) Type I error
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) Type II error	True Negative

- Treat hypotheses asymmetrically
 - Fix rate of type-I error to preset goal

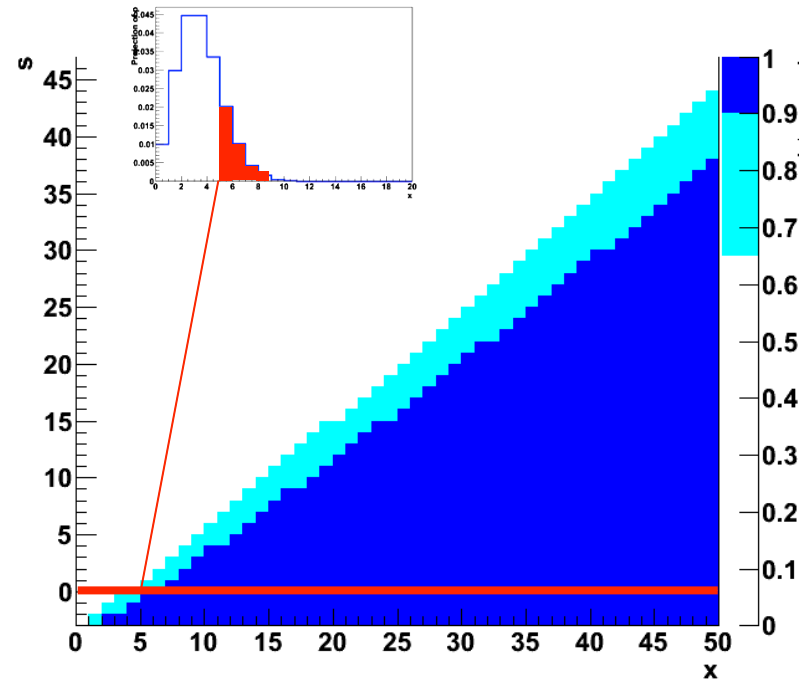
Confidence intervals for Poisson counting processes

- For simple cases, $P(x|\mu)$ is known analytically and the confidence belt can be constructed analytically
 - Poisson counting process with a fixed background estimate,
 - Example: for $P(x|s+b)$ with $b=3.0$ known exactly

Confidence belt from 68% and 90% central intervals



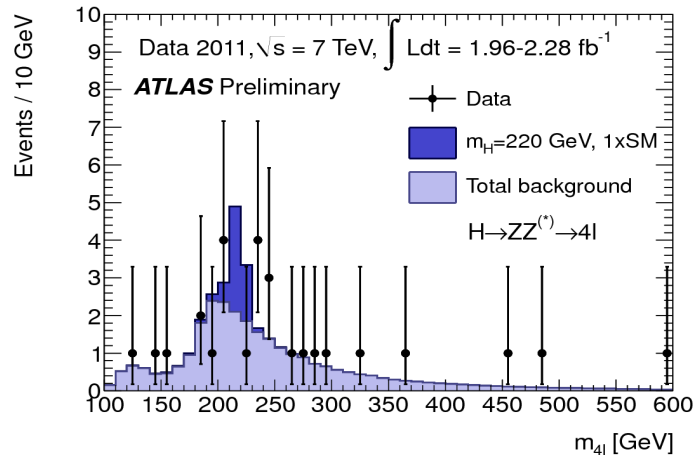
Confidence belt from 68% and 90% upper limit



Likelihood (ratios) and test statistics

Confidence intervals for non-counting experiments

- Typical LHC result is not a simple number counting experiment, but looks like this:



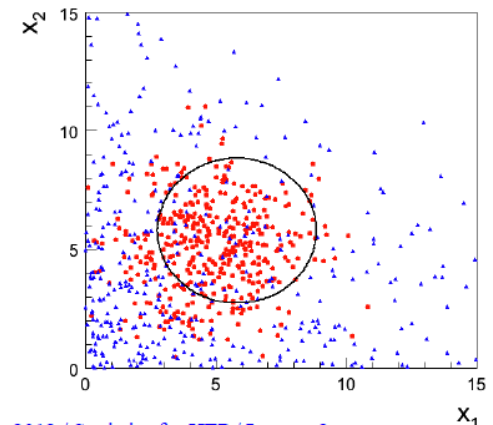
- Result is a distribution, not a single number
- (Models for signal and background have intrinsic uncertainties → for tomorrow)

- Any type of result can be converted into a single number by constructing a 'test statistic'
 - **A test statistic compresses all signal-to-background discrimination power in a single number**

The Neyman-Pearson lemma

- In 1932-1938 Neyman and Pearson developed in which one must consider competing hypotheses
 - Null hypothesis (H_0) = Background only
 - Alternate hypotheses (H_1) = e.g. Signal + Background
- 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
- → Use likelihood ratio as test statistic

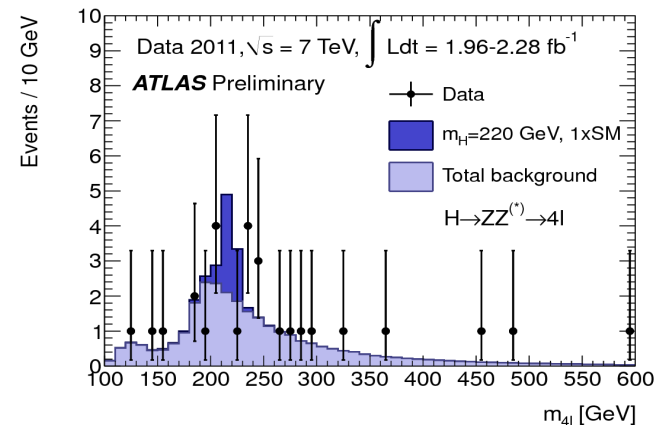
Formulating the likelihood for a distribution

- We observe n instances of x ($x_1 \dots x_n$)
- The likelihood for the entire experiment assuming background hypothesis (H_0) is

$$L_b = \frac{b^n}{n!} e^{-b} \prod_{i=1}^n f(x_i | b)$$

and for the signal-plus-background hypothesis (H_1) it is

$$L_{s+b} = \frac{(s+b)^n}{n!} e^{-(s+b)} \prod_{i=1}^n \left(\frac{s}{s+b} f(x_i | s) + \frac{b}{s+b} f(x_i | b) \right)$$



Formulating the likelihood ratio

- With the likelihood L_{s+b} and L_b the ratio becomes

$$Q = -2 \log \frac{L_{s+b}}{L_b} = -s + \sum_{i=1}^n \log \left(1 + \frac{s}{b} \frac{f(x_i | s)}{f(x_i | b)} \right)$$

- To compute the p-values for the s and $s+b$ hypotheses given an observed value of Q we need the distributions $f(Q|b)$ and $f(Q|s+b)$
 - Note that the $-s$ term is a constant and can be dropped
 - The rest is a sum of contributions for each event, and each term in the sum has the same distribution
 - Can exploit this to relate the distribution of Q to that of a single event using Fourier Transforms (this was done e.g. at LEP)

Using a likelihood ratio as test statistic

- For discovery and exclusion it is common to reformulate hypothesis in terms of signal strength

$\mu = \text{signal strength} / \text{nominal signal strength (e.g. SM)}$

so that $\mu=0$ represents the background hypothesis and $\mu=1$ represent the nominal signal hypothesis (e.g. the SM cross-section)

- In this formulation, likelihood ratio of previous page becomes

$$q = -2 \ln \frac{L(\text{data} | \mu = 1)}{L(\text{data} | \mu = 0)}$$

- This is the 'Tevatron test statistic' (without nuisance parameters)

Using a likelihood ratio as test statistic

- At the LHC experiments a different test statistic is commonly used:

$$t_{\mu} = -2 \ln \lambda(\mu), \quad \lambda(\mu) = \frac{L(\text{data} \mid \mu = \mu)}{L(\text{data} \mid \hat{\mu})}$$

$\hat{\mu}$ is best fit value of μ

- Where μ can be chosen, e.g.

$$t_1 = -2 \ln \lambda(1), \quad \lambda(1) = \frac{L(\text{data} \mid \mu = 1)}{L(\text{data} \mid \hat{\mu})}$$

$$t_0 = -2 \ln \lambda(0), \quad \lambda(0) = \frac{L(\text{data} \mid \mu = 0)}{L(\text{data} \mid \hat{\mu})}$$

Use a likelihood ratio as test statistic

- Illustration of s vs s+b discrimination power: t_1 :

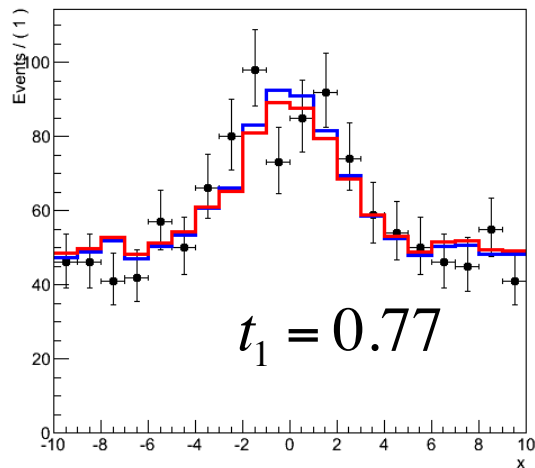
'likelihood assuming nominal signal strength'

$$t_1 = -2 \ln \frac{L(\text{data} \mid \mu = 1)}{L(\text{data} \mid \hat{\mu})}$$

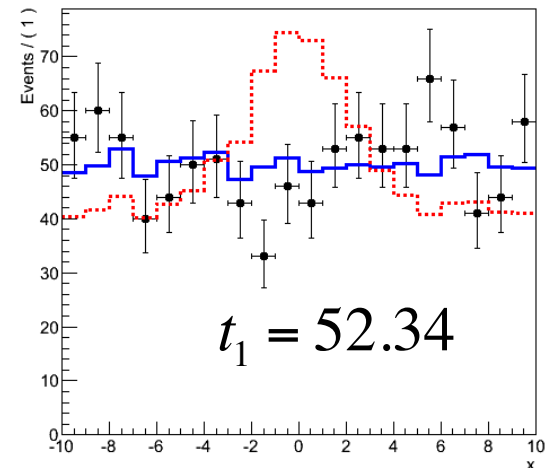
$\hat{\mu}$ is best fit value of μ

'likelihood of best fit'

On signal-like data q_1 is small



On background-like data q_1 is large



Use a likelihood ratio as test statistic

- Illustration of s vs s+b discrimination power: t_0 :

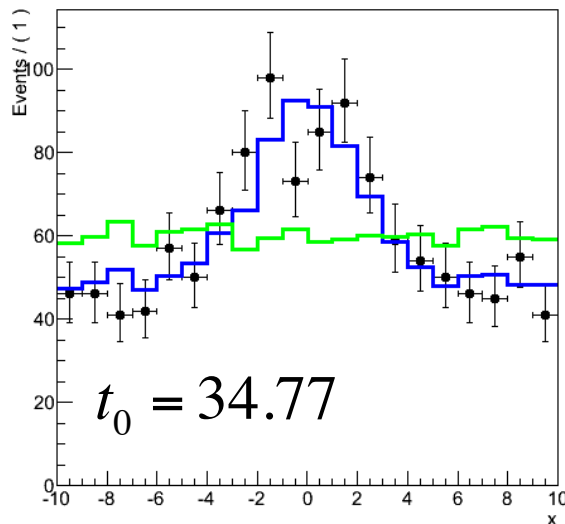
'likelihood assuming zero signal strength'

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

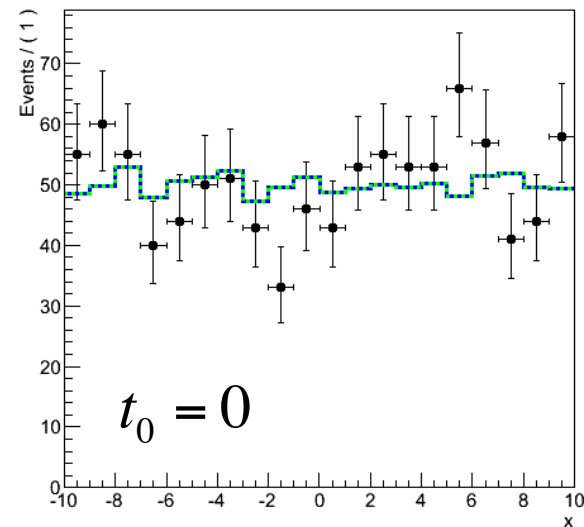
$\hat{\mu}$ is best fit value of μ

'likelihood of best fit'

On signal-like data t_0 is large

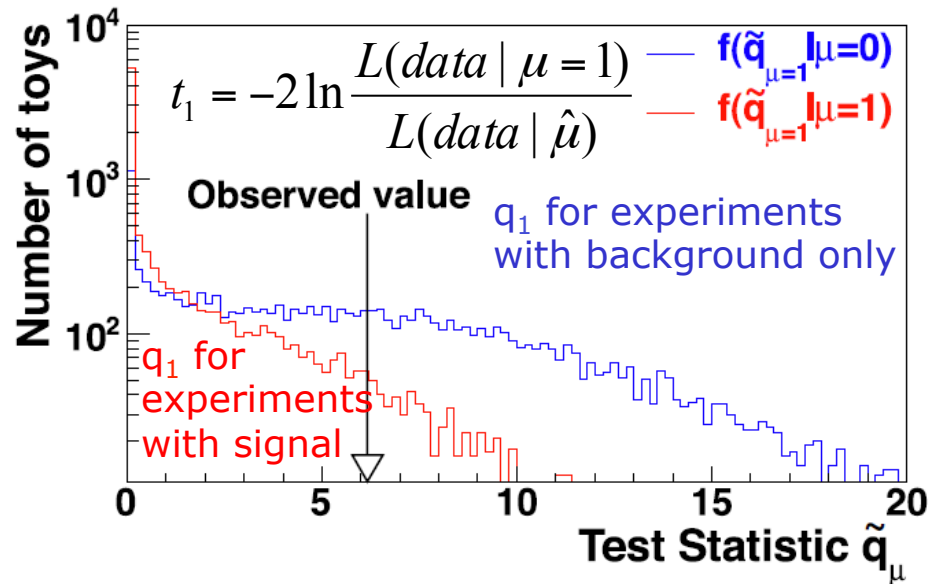
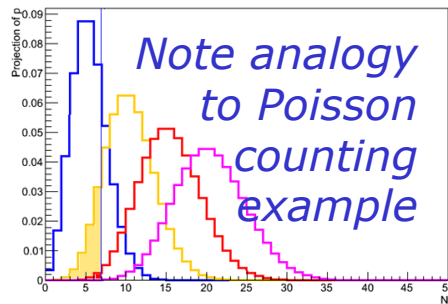


On background-like data t_0 is small



Setting limits with t_1

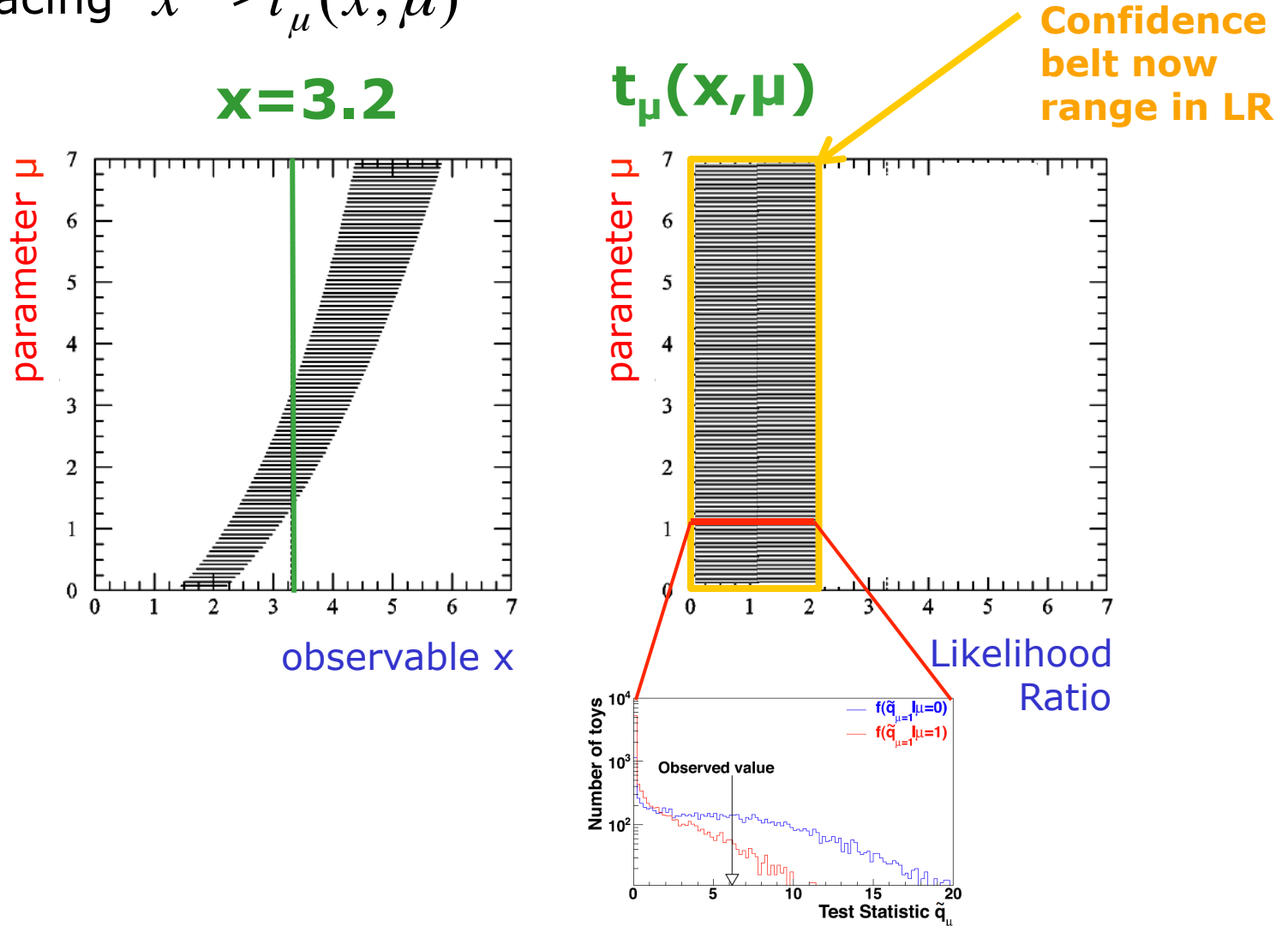
- Observed value of t_1 is now the 'measurement'
- **Distribution of $t_1 = f(t_1 | \mu=1)$ not calculable**
 - But can obtain distribution from toy MC approach
 - Asymptotic form exists for $N \rightarrow \infty$



- Limit on μ (w/o CL_S) : Find t_μ for which $\int_0^\infty f(t_\mu | \mu) = \alpha\%$
- P-value of bkg similarly obtained from t_0 $\int_{t_0}^{t_0^{obs}} f(t_0 | \mu = 0) = \alpha\%$

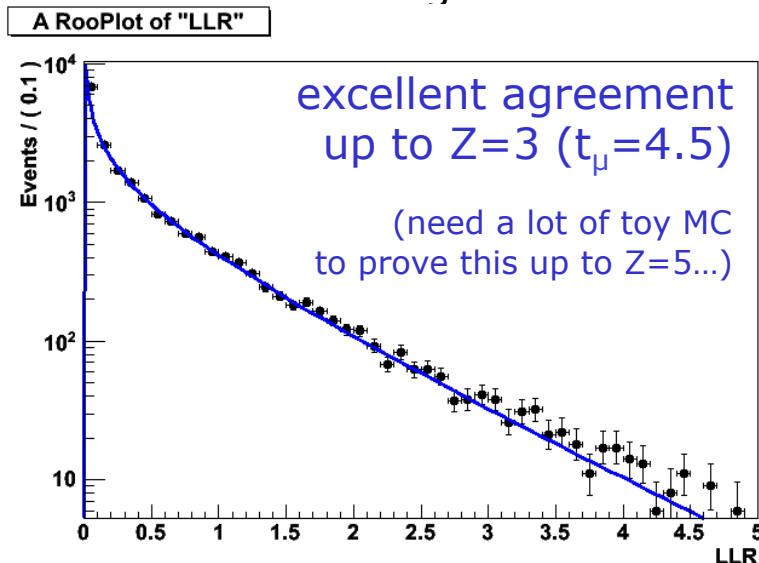
Confidence belts for non-trivial data

- What will the confidence belt look like when replacing $x \rightarrow t_\mu(\vec{x}, \mu)$

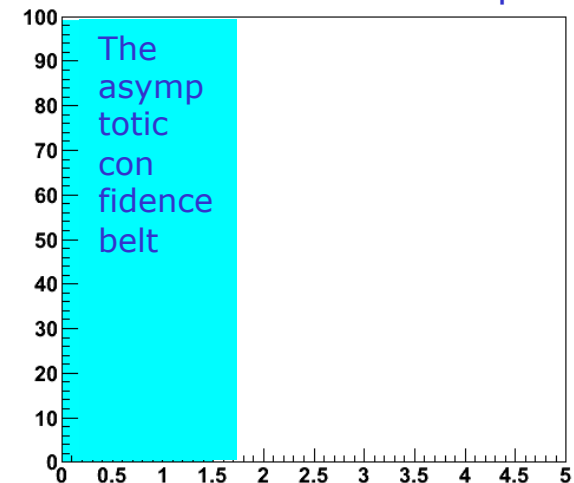


Confidence belts with t_μ as test statistic

- Use asymptotic distribution of t_μ
 - Wilks theorem \rightarrow
Asymptotic form of $f(t_\mu|\mu)$ is chi-squared distribution $f(t_\mu|\mu)=\chi^2(2\cdot t_\mu, n)$, with n the number of parameters of interest ($n=1$ in example shown)
 - Note that $f(t_\mu|\mu)$ is independent of μ ! \rightarrow
For data generated under the hypothesis μ the the distribution of t_μ is asymptotically always $\chi^2(2\cdot t_\mu, n)$
 - Example of convergence to asymptotic behavior
 $f(t_\mu|\mu)$ distribution for measurement consisting of 100 event with Gaussian



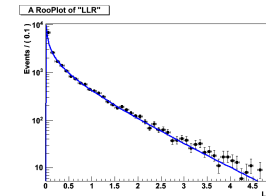
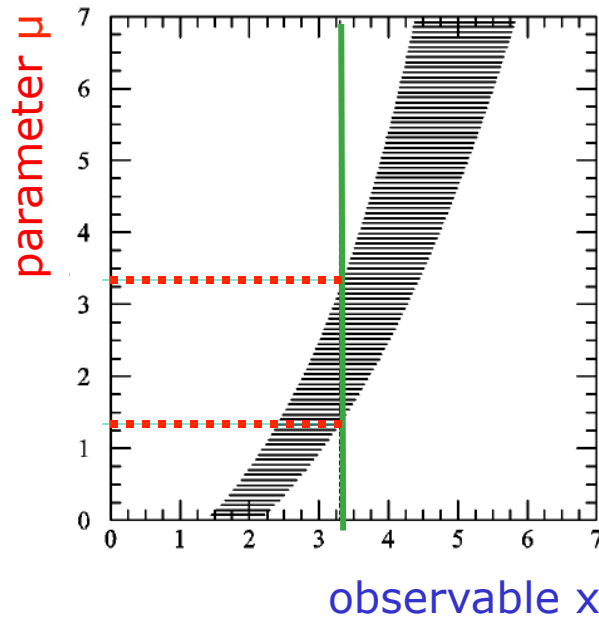
Value of t_μ representing fixed quantile of $f(t_\mu|\mu)$ is the same for all μ



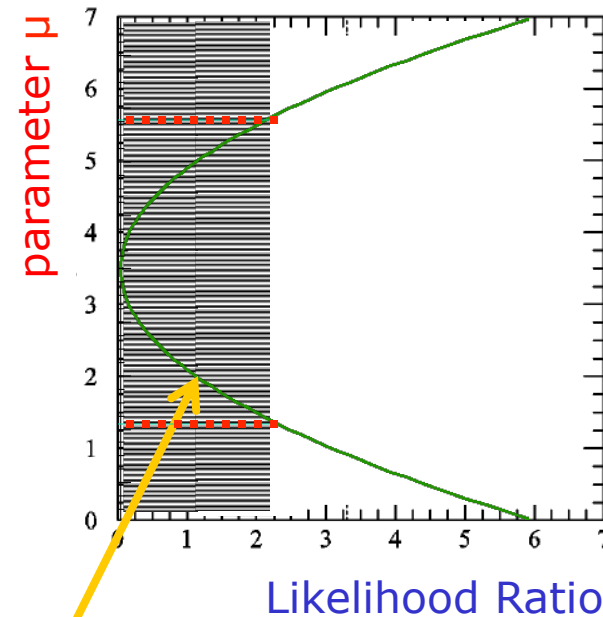
Confidence belts with t_μ as test statistic

- What will the confidence belt look like when replacing $x \rightarrow t_\mu(\vec{x}, \mu)$

$x=3.2$



$t_\mu(x, \mu)$

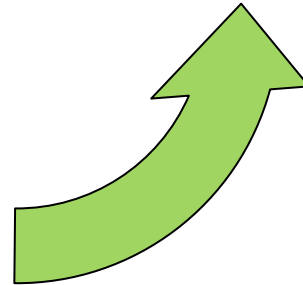
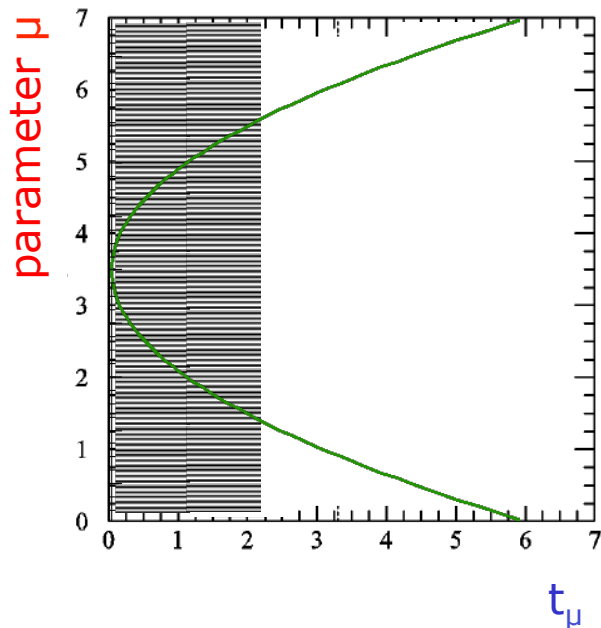


Measurement = $t_\mu(x_{\text{obs}}, \mu)$
is now a function of μ

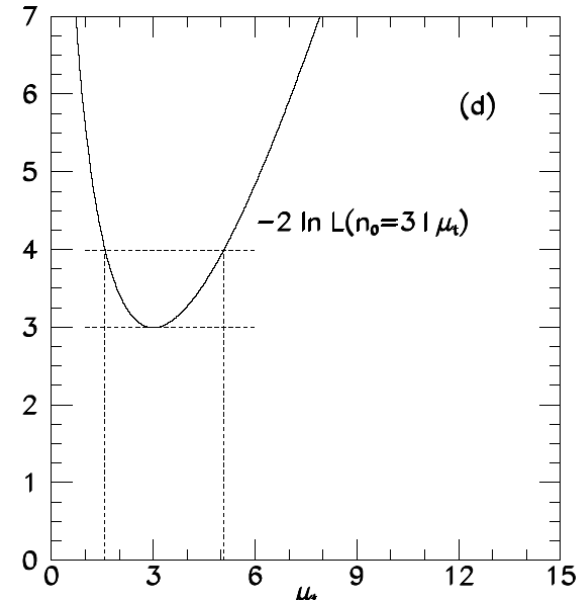
Connection with likelihood ratio intervals

- 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 $t_\mu = 1/2 \cdot Z^2$
 - This is exactly the MINOS error

FC interval with Wilks Theorem



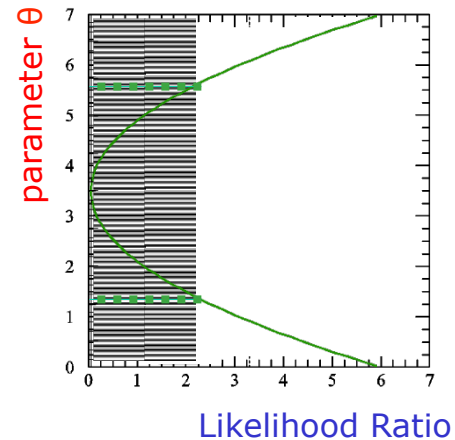
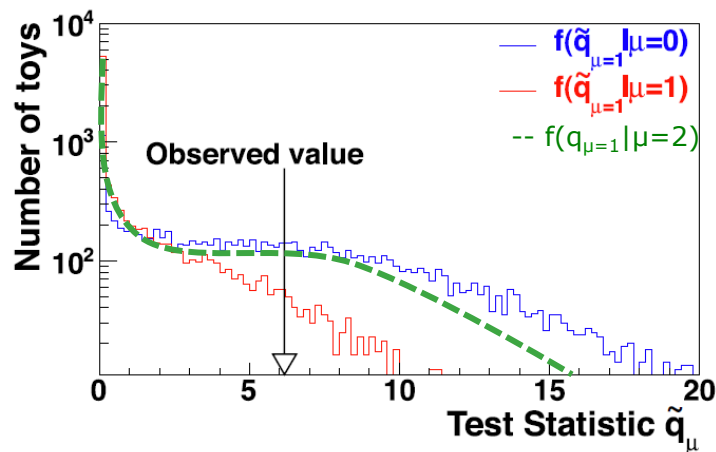
MINOS / Likelihood ratio interval



LHC test statistics for discovery and limit setting

Note on test statistic t_μ

- Note that high values of t_μ (i.e. strong incompatibility of data with hypothesized signal strength μ) can arise in two ways
 - An estimated signal strength $\hat{\mu}$ **greater** than the assumed signal strength μ
 - An estimated signal strength $\hat{\mu}$ **smaller** than the assumed signal strength μ



- May result in a two-sided interval (i.e. rejected values of μ are both below and above those accepted)

Test statistics q_0 for discovery of a positive signal

- Important special case:
test $\mu=0$ for a class of models where we assume $\mu \geq 0$
 - Rejecting $\mu=0$ effectively leads to discovery of new signal
- Define a new test statistic q_0 :

$$q_0 = \begin{cases} -2 \ln \lambda(0) \hat{\mu} \geq 0 \\ 0 \quad \hat{\mu} < 0 \end{cases}$$

$$p_0 = \int_{q_0^{obs}}^{\infty} f(q_0 | 0) dq_0$$

- Here only regard *upward* fluctuation of data as evidence against the background-only hypothesis
- Note that even though here physically $\mu \geq 0$, we allow $\hat{\mu}$ to be negative.
 - In the large sample limit its distributions becomes Gaussian and allows to write a simple expression for the distribution of this test statistic (will cover this tomorrow)

Test statistic q_μ for upper limits

- For the purpose of establishing an upper limit we introduce a new test statistic

$$q_\mu = \begin{cases} -2 \ln \lambda(\mu) \hat{\mu} \leq \mu \\ 0 \hat{\mu} > \mu \end{cases}$$

- With this test statistic one does **not** regard data with $\hat{\mu} > \mu$ as representing less compatibility with μ than the data obtained
- Note that $q_0 \neq q_\mu(\mu=0)$: q_0 is zero if data fluctuate downward ($\hat{\mu} < 0$), q_μ is zero if data fluctuate upward ($\hat{\mu} > \mu$)
- Calculate p-value as usual

$$p_\mu = \int_{q_\mu^{obs}}^{\infty} f(q_\mu | \mu) dq_\mu$$

Asymptotic formulae for p_0 from q_0 for Poisson model

- Well-known Gaussian approximation of significance for counting experiments

For large $s + b$, $n \rightarrow x \sim \text{Gaussian}(\mu, \sigma)$, $\mu = s + b$, $\sigma = \sqrt{s + b}$.

$$\text{median}[Z_0 | s + b] = \frac{s}{\sqrt{b}}$$

- Better approximation using Poisson likelihood in q_0

$$L(s) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

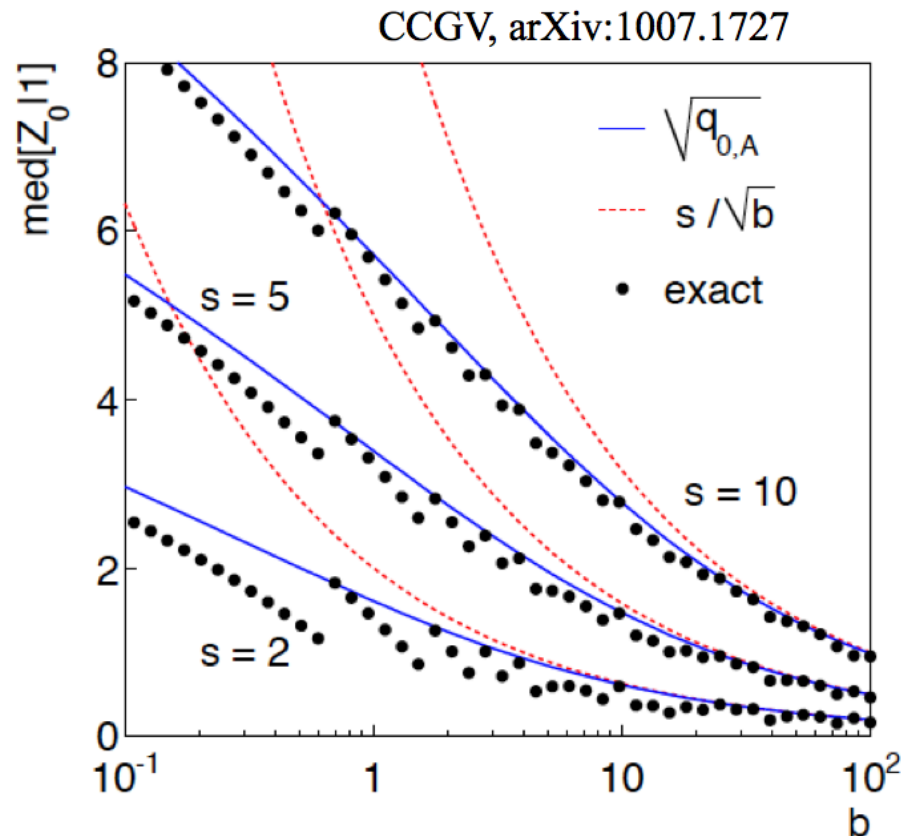
$$q_0 = -2 \ln \frac{L(0)}{L(\hat{s})} = 2 \left(n \ln \frac{n}{b} + b - n \right) \quad \text{for } n > b, 0 \text{ otherwise}$$

$$Z \approx \sqrt{q_0} \quad \text{with } n = s + b$$

$$\text{median}[Z_0 | s + b] \approx \sqrt{2 \left((s + b) \ln(1 + s/b) - s \right)}$$

Testing the approximate Poisson significance

- Model: Poisson($n|\mu s+b$).
- Test hypothesis $\mu=0$ for data generated with $\mu=1$ for $s=2,5,10$ and $b=0.01 \dots 100$
 - Exact solutions 'jumps' due to discreteness of Poisson



Summary on LHC test statistics

- Test statistic t_μ
$$t_\mu = -2 \ln \lambda(\mu), \lambda(\mu) = \frac{L(\text{data} | \mu = \mu)}{L(\text{data} | \hat{\mu})}$$
 - Can result in both 1-sided and 2-sided intervals as high values of t_μ (data incompatible with hypothesis μ) can arise if $\mu > \hat{\mu}$ or if $\mu < \hat{\mu}$
 - Asymptotically relates to MINOS intervals
- Test statistic q_0
$$q_0 = \begin{cases} -2 \ln \lambda(0) \hat{\mu} \geq 0 \\ 0 \hat{\mu} < 0 \end{cases}$$
 - Formulated for discovery – does not count low statistical fluctuations against the background hypothesis
- Test statistic q_μ
$$q_\mu = \begin{cases} -2 \ln \lambda(\mu) \hat{\mu} \leq \mu \\ 0 \hat{\mu} > \mu \end{cases}$$
 - Formulated for limit setting – does not count high statistical fluctuation against the signal hypothesis

Bayesian intervals in 3 slides

Bayes' Theorem Generalized to Probability Densities

- Original Bayes Thm:

$$P(B|A) \propto P(A|B) P(B).$$

- Let probability density function $p(\mathbf{x}|\mu)$ be the conditional pdf for data \mathbf{x} , given parameter μ . Then Bayes' Thm becomes

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu) p(\mu).$$

- Substituting in a set of **observed data**, x_0 , and recognizing the likelihood, written as $L(x_0|\mu)$, $L(\mu)$, then

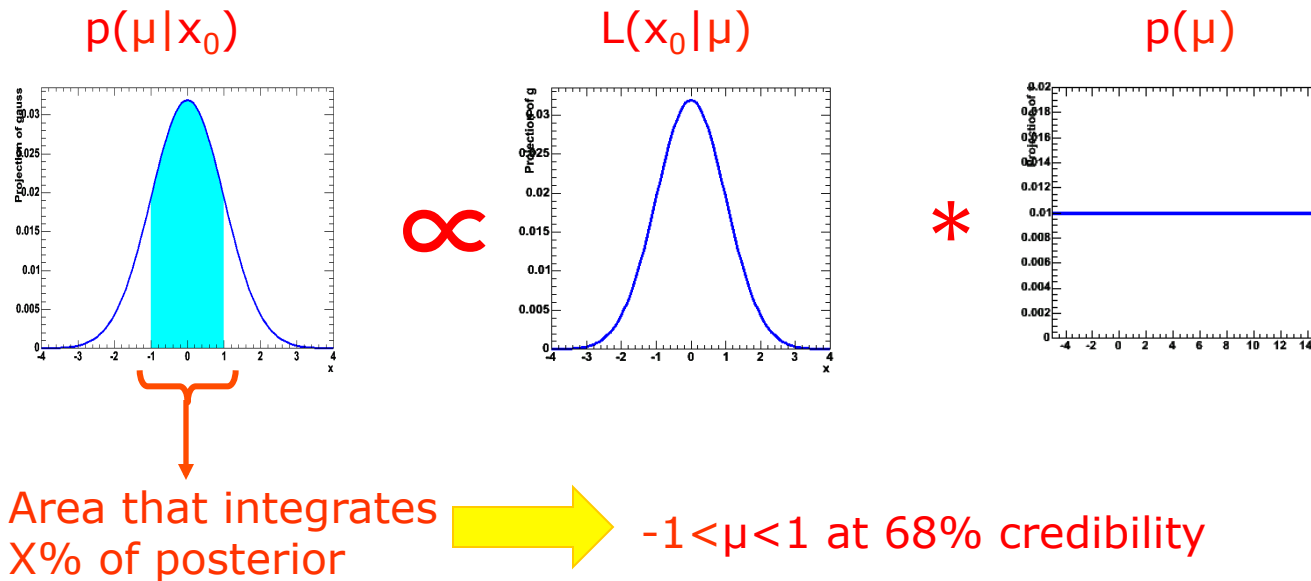
$$p(\mu|x_0) \propto L(x_0|\mu) p(\mu),$$

where:

- $p(\mu|x_0)$ = posterior pdf for μ , given the results of this experiment
 - $L(x_0|\mu)$ = Likelihood function of μ from the experiment
 - $p(\mu)$ = prior pdf for μ , before incorporating the results of this experiment
- Note that there is one (and only one) probability density in μ on each side of the equation, again consistent with the likelihood *not* being a density.

Bayes' Theorem Generalized to pdfs

- Graphical illustration of $p(\mu|x_0) \propto L(x_0|\mu) p(\mu)$

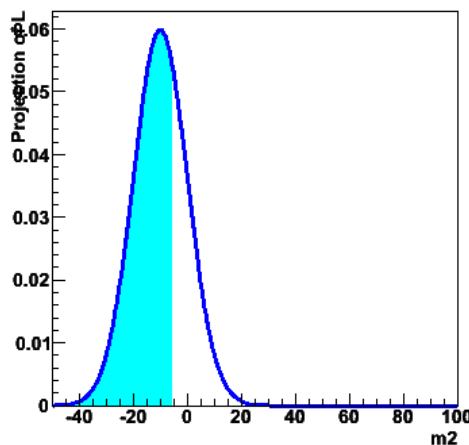


- Upon obtaining $p(\mu|x_0)$, the *credibility* of μ being in any interval can be calculated by integration.

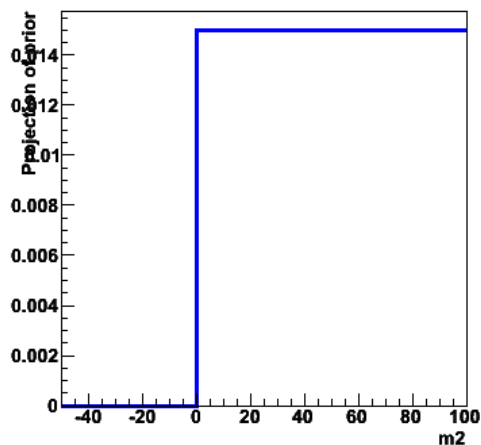
Using priors to exclude unphysical regions

- Priors provide a simple way to exclude unphysical regions from consideration
- Simplified example situations for a measurement of m_ν^2
 1. Central value comes out negative (= unphysical).
 2. Upper limit (68%) may come out negative, e.g. $m^2 < -5.3$, not so clear what to make of that

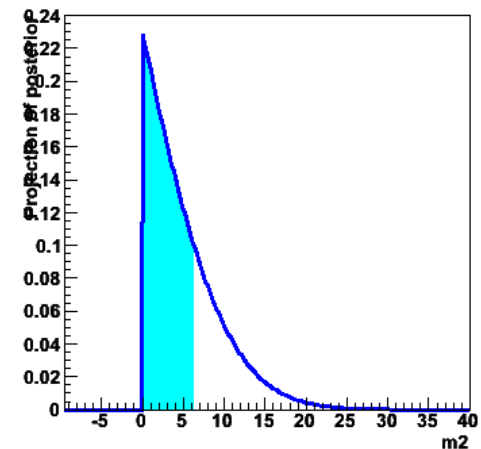
$p(\mu|x_0)$ with flat prior



$p'(\mu)$



$p(\mu|x_0)$ with $p'(\mu)$



- Introducing prior that excludes unphysical region ensure limit in physical range of observable ($m^2 < 6.4$)
- NB: Previous considerations on appropriateness of flat prior for domain $m^2 > 0$ still apply

Comparing Bayesian and Frequentist results

- Frequentist statements use only $P(\text{data}|\text{theory})$ (i.e. the likelihood)
 - Formulate output as p-value for an hypothesis (the probability to measure observed result or more extreme is $\alpha\%$ under the background hypothesis)
 - Formulate output as confidence interval: range of signal values for which p-value is below the stated threshold
- Bayesian statements calculate $P(\text{theory}|\text{data})$
 - Choice of prior will matter
 - Formulate output as Bayesian credible interval integrate $\alpha\%$ of posterior
 - No equivalent of p-values
- Numeric results will usually differ a bit since statistical question posed is different
 - Agreement generally worse at high confidence levels / low p-values
 - Agreement generally better with increasing statistics

68% intervals by various methods for Poisson process with $n=3$ observed

Method	Prior	Interval	Length	Coverage?
rms deviation $n \pm \sqrt{n}$	–	(1.27, 4.73)	3.46	no
Bayesian central	1	(2.09, 5.92)	3.83	no
Bayesian shortest	1	(1.55, 5.15)	3.60	no
Bayesian central	$1/\mu$	(1.37 , 4.64)	3.27	no
Bayesian shortest	$1/\mu$	(0.86, 3.85)	2.99	no
Likelihood ratio	–	(1.58, 5.08)	3.50	no
Frequentist central	–	(1.37 , 5.92)	4.55	yes
Frequentist shortest	–	(1.29, 5.25)	3.96	yes
Frequentist LR ordering	–	(1.10, 5.30)	4.20	yes

Hands-on exercises – Part 1

- Any input files in <http://www.nikhef.nl/~verkerke/brussel>
- The goals of this set of hands-on exercises is to work with a very simple number-counting experiment and and calculate p-values and limits 'by hand' to appreciate the concepts
- The model we will be working with is

$$\text{Poisson}(N; \mu = s + b)$$

with $b=3$ precisely (no uncertainty on the prediction) and s as a free parameter

- Given the very simple nature of these exercises you can do these in plain ROOT.
 - Some visualization of models provided for you in RooFit
 - In plain root you can use function `TMath::Poisson()` which is always normalized on the range $[0, \text{inf}]$

Hands-on exercises - Part 1

- **A)** Visualizing the model (done for you in RooFit → [mod1/ex1A.C](#))
 - The provided macro plots
 - $P(n; \mu=s+b)$ versus N for $s=0$ (background only) $s=5$, $s=10$
 - $P(N=7; \mu)$ versus s (this is the likelihood $L(s)$ for $N=7$).
 - $\log(L(s))$
 - Plot Poisson distributions for other values of s
 - Plot the likelihood for other values of N
- **B)** The p-value of the background hypothesis
 - Given an observation of $N=7$, calculate the probability to see 7 or more events for the background-only hypothesis (using `Tmath::Poisson`)
- **C)** An 95% C.L. upper limit on the signal
 - Write a C++ function `calc_prob(int N, double s)` that returns the probability to observe N or less events for a signal count s .
[[Here you are writing a function that performs a hypothesis test](#)]
 - Calculate the probability to see 7 or less events for the $s=5$ and $s=10$ models.
 - Write a C++ function `find_limit(int N, double CL)` that returns the value of s for which `calc_prob(N,s)==CL`.
For simplicity I suggest you simply scan s in steps of 0.01 to find the right value.
[[Here you are performing an 'hypothesis test inversion' to construct a confidence interval \[0,x\]](#)]
 - Calculate the 95% upper limits for s for $N=0,1,2,3,4,5,6,7,8,9,10$
 - Note that for a statement " $s > X$ excluded at 95% C.L." you construct an interval $[0, X]$ at 5% C.L.

Hands-on exercises – Part 1

- **D)** A 95% C.L. CL_S limit on signal
 - Calculate CL_S for $N=7$. Remember that $CL_S = p_{s+b} / (1-p_b)$.
 - For a discover observable N p_{s+b} integrates $[0,N]$ and $1-p_b$ integrates $[0,N]$ too (so you can calculate the latter with `calc_prob(N,0)`)
 - Write a C++ function `calc_cls(int N, double s)` that calculates CL_S for N observed events and a signal hypothesis of s event. You can use `calc_prob()` of the previous exercise as a starting point
 - Write a C++ function `calc_cls_limit(int N)` that calculates a upper limit using the CL_S procedure for N observed events. You can use `calc_limit()` of the previous exercise as a starting point and simply substitute the call to `calc_prob()` to `calc_cls()`.
- **E)** Comparing CL_S limits and plain frequentist limits
 - Make a plot of the CLS limits on s and plain frequent limit on s for observations of $N=0,1,2,3,\dots,20$ to visualize the impact of the CLS procedure on the limits at low N

Hands-on exercises – Part 1

- **F) Bayesian intervals**

(done for you in RooFit → mod1/ex1F.C)

- A Bayesian interval on the same model is calculated by finding an interval that contains 95% of the posterior. The posterior is calculated as $P(s) = L(s) \cdot \pi(s)$, where $\pi(s)$ is the prior
- A common choice of prior for upper limits where $s \geq 0$, is a flat prior for $s \geq 0$ and $\pi(s) = 0$ for $s < 0$. For this particular case, the Bayesian upper limit can simply be calculated by find the value s_{UL} for which

$$\int_{-\infty}^{s_{UL}} P(s) ds = \frac{\int_0^{s_{UL}} L(s) ds}{\int_0^{\infty} L(s) ds} = 0.95$$

- The provided macro plots the cumulative distribution of the posterior function, normalized over the range $[0, \infty]$ so you can trivially find the value of s_{UL} for which the above equation holds
- Calculate the Bayesian 95% upper limit for $N=2, N=7$ and compare this to the classic and CL_S limit for $N=2, 7$