Confidence intervals
Limits & significance

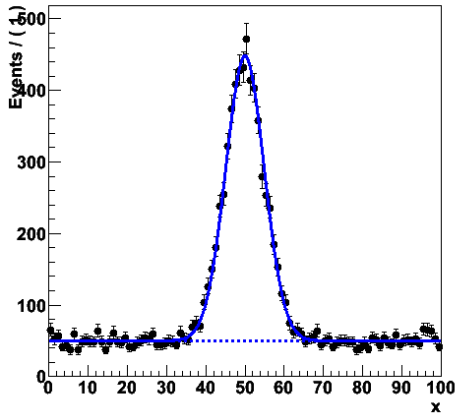
—Null Hypothesis testing – P-values
— Classical or 'frequentist' confidence intervals
— Issues that arise in interpretation of fit result
— Bayesian statistics and intervals

Wouter Verkerke, UCSB

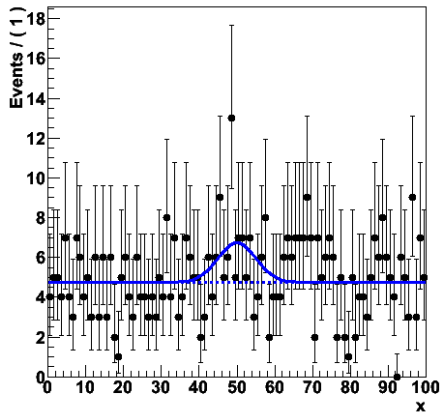
Introduction

- Issues and differences between methods arise when experimental result contains little information

'Easy'



'Difficult'



- Now we focus on the difficult cases
- Most common scenario is establishing the presence of signal in the data (at a certain confidence level), or be able to set limits, in the absence of a convincing signal
  - Connection with hypothesis testing

Hypothesis testing

- Hypothesis testing
  - What to choose as null hypothesis and what as alternate depends on the question you want to ask
  - Example: discovery/exclusion of SuperSymmetry extension of SM
- Version 1
  - H_0 = No supersymmetry
  - H_1 = SuperSymmetry (of some kind)
  - Prediction: Nevt(8 jets>p_T20) = 10 (at *x* fb^{-1})
  - Measurement: N=9
- Version 2
  - H_0 = Standard Model
  - H_1 = Standard model + something else
  - Prediction: Nevt(8 jets>p_T20) = 3 (at *x* fb^{-1})
  - Measurement: N=9

Significance, Probability

- Be sure to formulate the correct question!
  - What we usually want to know for discovery is: what is the probability that the 'background' has an fluctuation that looks like our signal (or better), i.e. Version#2
  - Version #1 quantifies the probability that nature with SUSY would result in an experimental result consistent with no SUSY. You would use this to set a limit
- When making statistical inference on data samples that contain little information, precise formulation of question and assumption made, become very important
- Need to discuss fundamentals of probability and statistics more before proceeding.

Definition of "Probability"

- 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 (see next slides)

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

- Two established* incarnations of P are:

- ***1) Frequentist P**:* limiting frequency in ensemble of imagined repeated samples (as usually taught in Q.M.). P(constant of nature) and P(SUSY is true) do not exist (in a useful way) for this definition of P (at least in one universe).

- ***2) (Subjective) Bayesian P***: subjective degree of belief. (de Finetti, Savage) P(constant of nature) and P(SUSY is true) exist for You. Shown to be basis for coherent personal decision-making.

**It is important to be able to work with either definition of P, and to know which one you are using!*

Frequentist P – the initial example

- Work out initial example (discovery – version 2)
  - 'Signal' = SUSY 'Background' = Standard model

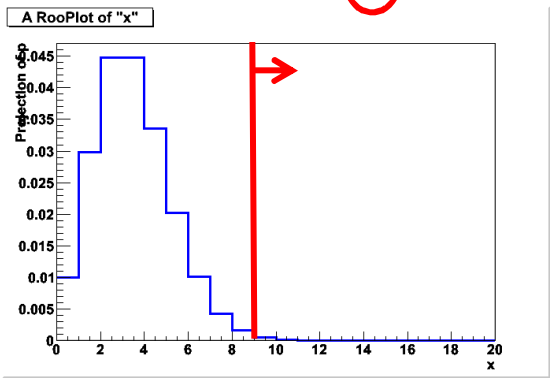
Prediction N=3
Measurement N=9

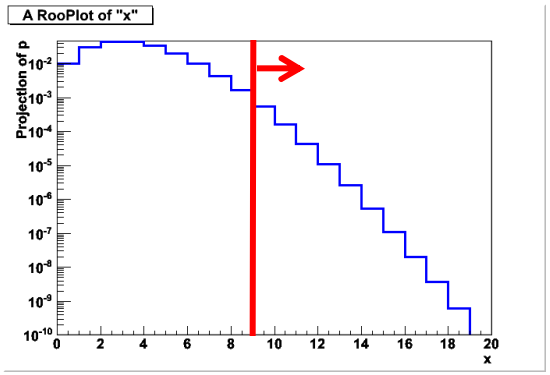
N(bkg) = 3
N(sig+bkg) = 9

- Can we calculate probability that SM mimics SM+SUSY (i.e. result is a 'false positive)?
  - Calculation details depend on how measurement was done (fit, counting etc..)
  - Simplest case: counting experiment, Poisson process

$$p = \int_{9}^{\infty} Poisson(n;\mu = 3)dn = 0.0038$$

='p value'





Frequentist P – working out example #2

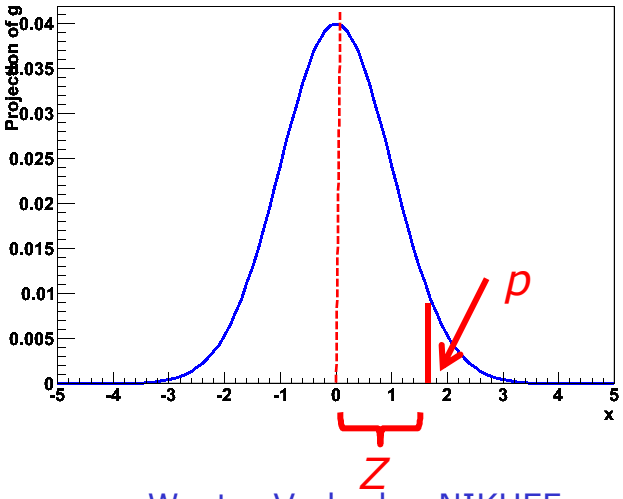
- P-value - If you repeat experiment many times, given fraction of experiments will result in result more extreme that observed value
  - In this example, only 0.38% of experiments will result in an observation of 9 or more events when 3 are expected.
- P-Value vs Z-value (significance)
  - Often defines significance *Z* as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same *p*-value.

`TMath::Erfc`

$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$

$$Z = \Phi^{-1}(1-p)$$

`TMath::NormQuantile`

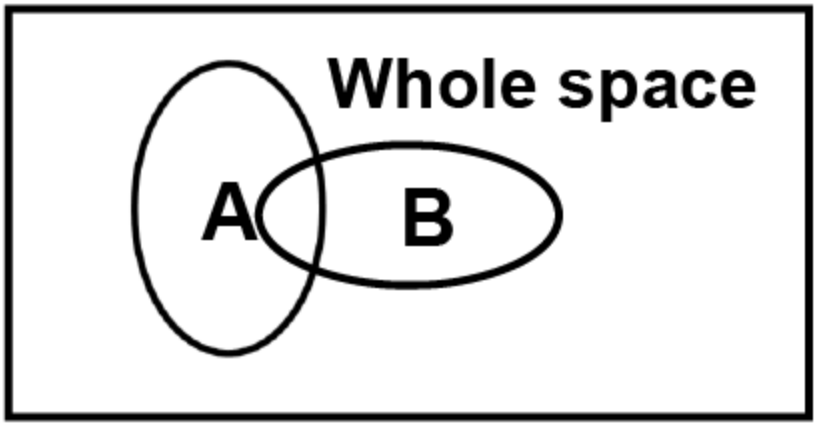


Bayes Theorem in pictures

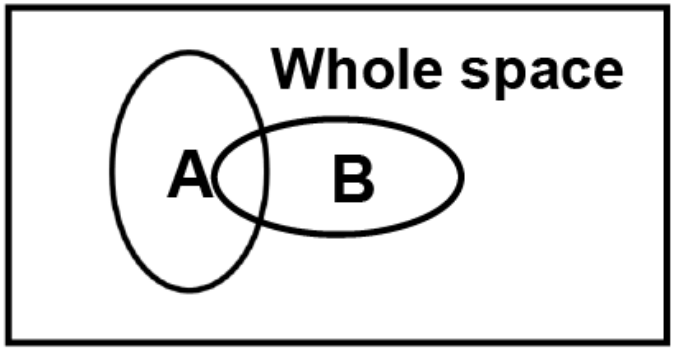
- Rev. Thomas Bayes
- 1702 – 7 April 1761
- Bayes Theorem

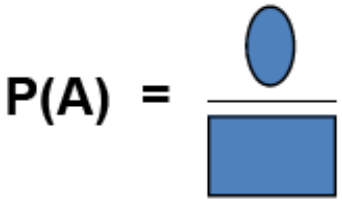
$$P(B|A) = P(A|B)\,P(B) / P(A).$$

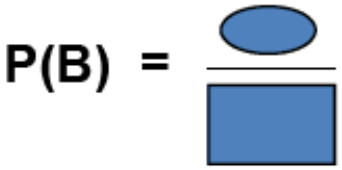
- *Essay "Essay Towards Solving a Problem in the Doctrine of Chances" published in Philosophical Transactions of the Royal Society of London in 1764*

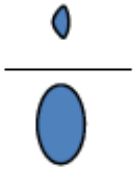


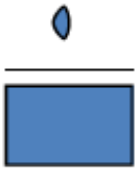
Bayes' Theorem in Pictures

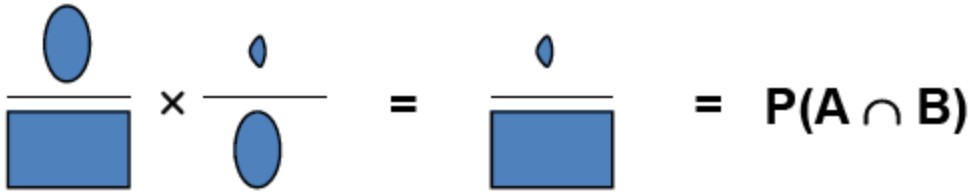


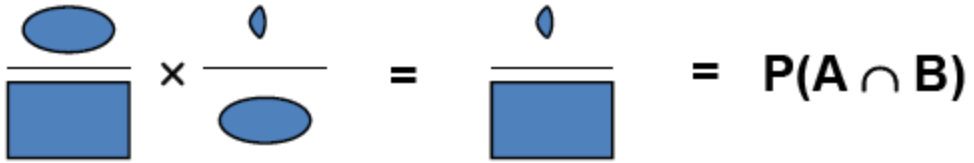












⇒ P(B|A) = P(A|B) × P(B) / P(A)

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 –see the discussion of "Conditioning" in later slides.

[B.Cousins HPCP]

Example of Bayes' Theorem Using Frequentist P

- A b-tagging method is developed and one measures:
  - P(btag| b-jet), i.e., efficiency for tagging b's
  - P(btag| not a b-jet), i.e., efficiency for background
  - P(no btag| b-jet) = 1 -P(btag| b-jet),
  - P(no btag| not a b-jet) = 1 -P(btag| not a b-jet)
- Question: Given a selection of jets tagged as b-jets, what fraction of them is b-jets?
  I.e., what is P(b-jet | btag) ?
- Answer: *Cannot be determined from the given information!*
  - Need also: P(b-jet), the true fraction of *all jets that are b-jets.*
    Then Bayes' Theorem inverts the conditionality:

    $P(\text{b-jet} \mid \text{btag}) \propto P(\text{btag}|\text{b-jet})\, P(\text{b-jet})$

[B.Cousins HPCP]

Example of Bayes' Theorem Using Bayesian P

- In a background-free experiment, a theorist uses a "model" to predict a signal with Poisson mean of 3 events. From Poisson formula we know
  - P(0 events | model true) = $3^0 e^{-3}/0! = 0.05$
  - P(0 events | model false) = 1.0
  - P(>0 events | model true) = 0.95
  - P(>0 events | model false) = 0.0
- The experiment is performed and zero events are observed.
- Question: Given the result of the expt, what is the probability that the model is true?

  I.e., What is P(model true | 0 events) ?

[B.Cousins HPCP]

Example of Bayes' Theorem Using Bayesian P

- Answer: *Cannot be determined from the given information!*
  - *Need in addition: P(model true), the degree of belief in the mode prior to the experiment. Then using Bayes' Thm*
  - P(model true | 0 events) ∝ P(0 events | model true) P(model true)

- If "model" is S.M., then still very high degree of belief after experiment!
- If "model" is large extra dimensions, then low prior belief becomes even lower.
  - N.B. Of course this example is over-simplified

[B.Cousins HPCP]

A Note re *Decisions*

- Suppose that as a result of the previous experiment, your degree of belief in the model is P(model true | 0 events) = 99%, and you need to *decide whether or not to take an action (making a press release, or planning your next experiment), based on the model being true.*
- 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.
- Statisticians often focus on decision-making; in HEP, the tradition thus far is to communicate experimental results (well) short of formal decision calculations. *One thing should become clear: classical "hypothesis testing" is not a complete theory of decision-making!*

[B.Cousins HPCP]

At what p/Z value do we claim 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.,

phenomenon	reasonable *p*-value for discovery
D^0D^0 mixing	~0.05
Higgs	$\sim 10^{-7}$ (?)
Life on Mars	$\sim 10^{-10}$
Astrology	$\sim 10^{-20}$

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

Bayes' Theorem Generalized to Probability Densities

- Original Bayes Thm:

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

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

$$p(\mu|x) \propto p(x|\mu)\, p(\mu).$$

- 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:

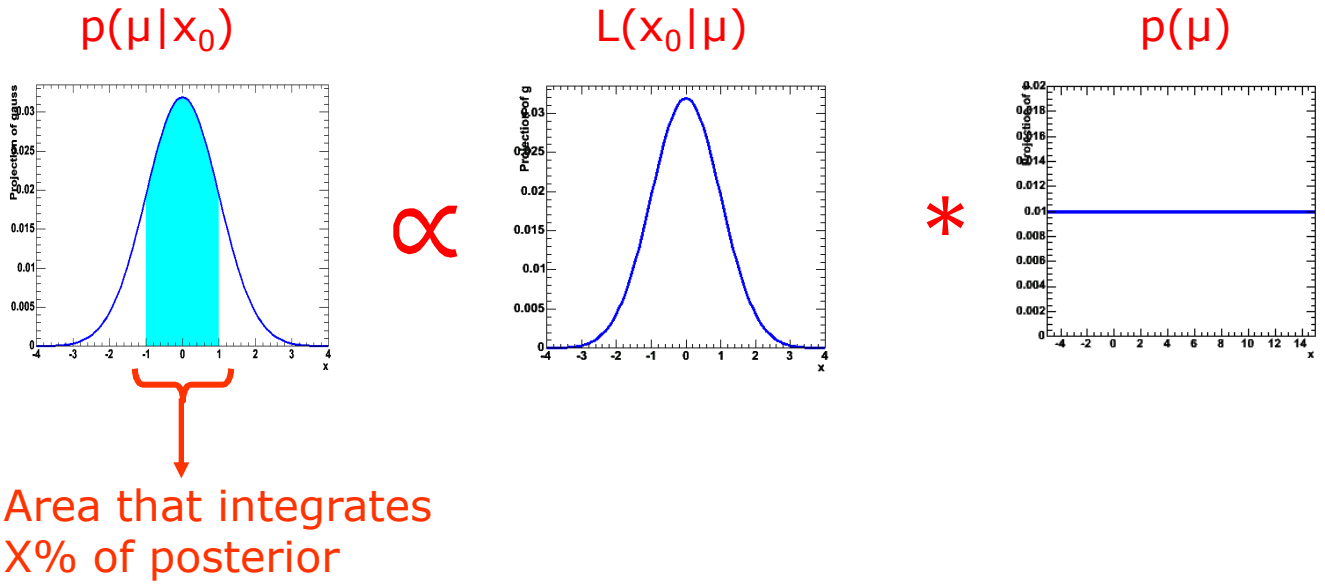
  - $p(\mu|x_0)$ = posterior pdf for μ, given the results of this experiment
  - $L(x_0|\mu)$ = Likelihood function of μ from the experiment
  - $p(\mu)$ = prior pdf for μ, before incorporating the results of this experiment

- 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.

[B.Cousins HPCP]

Bayes' Theorem Generalized to pdfs

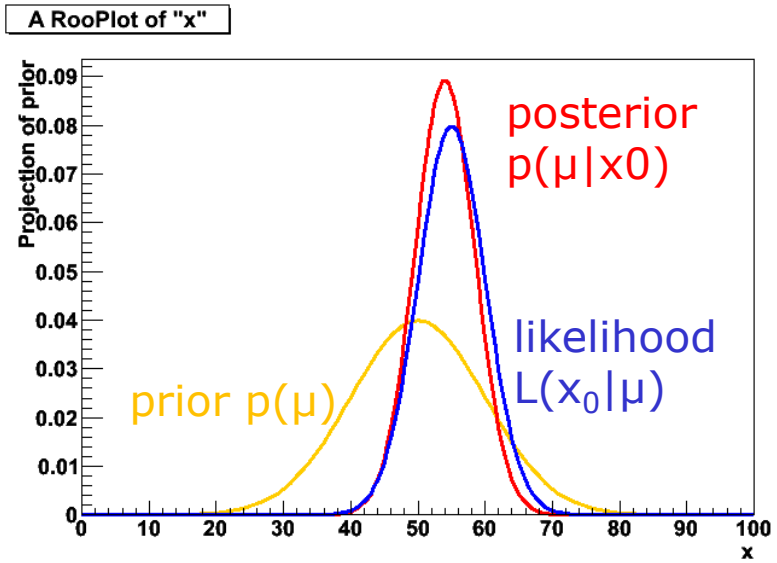
- Graphical illustration of $p(\mu|x_0) \propto L(x_0|\mu)\, p(\mu)$



- Upon obtaining p(μ|x0), the *credibility* of *μ* being in any interval can be calculated by integration.
- To make a *decision* as to whether or not μ is in an interval or not (e.g., whether or not μ>0) , one requires a further subjective input: the cost function (or utility function) for making wrong decisions

Choosing Priors

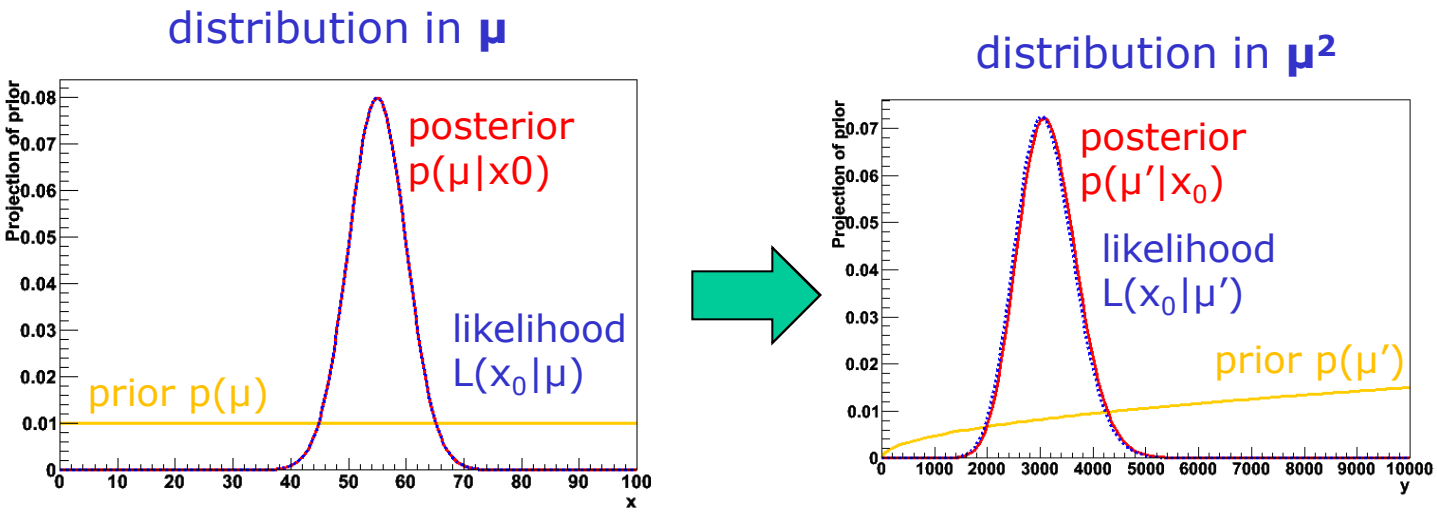
- When using the Bayesian formalism you always have a prior. What should you put in there?
- When there is clear prior knowledge, it is usually straightforward what to choose as prior
  - Example: prior measurement of μ = 50 ± 10



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

Choosing Priors

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



- Flat prior implies choice on given metric
  - Conversely you make any prior flat by a appropriate coordinate transformation (i.e a probability integral transform)
  - 'Preferred metric' has often no clear-cut answer. (E.g. when measuring neutrino-mass-squared, state answer in m or m^2)
  - In multiple dimensions even more issues (flat in x,y or flat in r,φ?)

Probability Integral Transform

- *"…seems likely to be one of the most fruitful conceptions introduced into statistical theory in the last few years" –Egon Pearson (1938)*
- Given continuous x ∈(a,b), and its pdf p(x), let

$$y(x) = \int_a^x p(x')\,dx'.$$

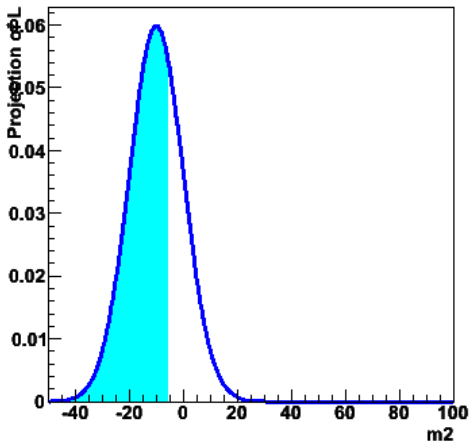
- Then y ∈(0,1) and p(y) = 1 (uniform) for all y. (!)
- So there *always* exists a metric in which the pdf is uniform.
  - The specification of a Bayesian prior pdf p(μ) for parameter μ is equivalent to the choice of the metric f(μ) in which the pdf is uniform.

[B.Cousins HPCP]

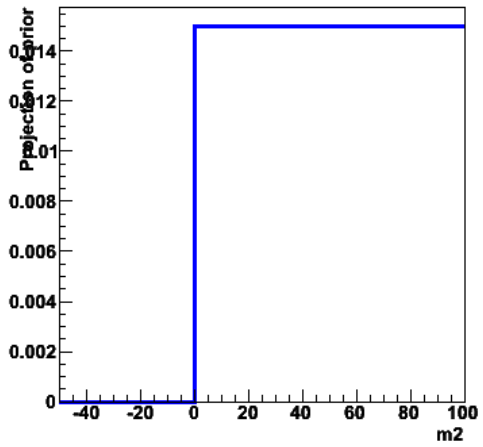
Using priors to exclude unphysical regions

- 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

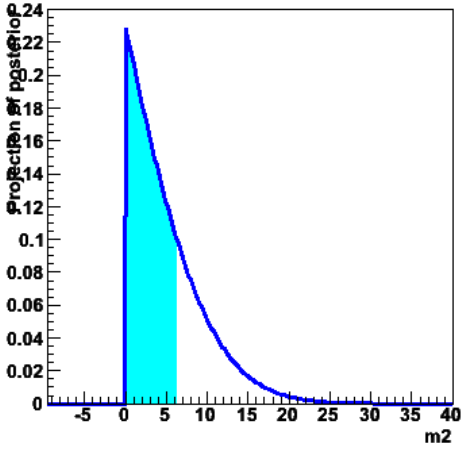
$p(\mu|x_0)$ with flat prior



$p'(\mu)$



$p(\mu|x_0)$ 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

Non-subjective priors?

- The question is: can the Bayesian formalism be used by scientists to report the results of their experiments in an "objective" way (however one defines "objective"), and does any of the coherence remain when subjective P is replaced by something else?

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

- "Jeffreys Prior" answers the question using a prior uniform in a metric related to the Fisher information.
  - Unbounded mean μ of gaussian: p(μ) = 1
  - Poisson signal mean μ, no background: p(μ) = 1/sqrt(μ)

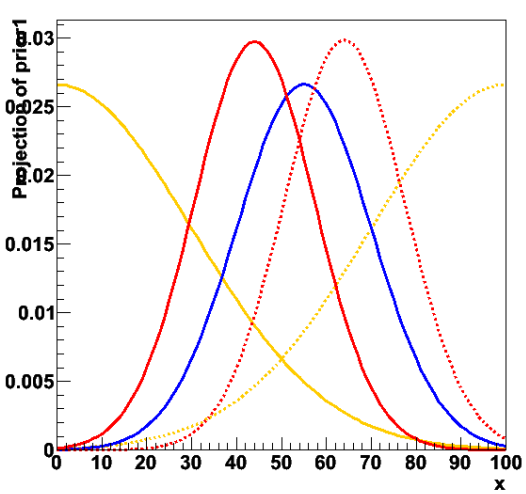
- Many ideas and names around on non-subjective priors
  - Objective priors? *Non-informative* priors? *Un*informative priors?
  - Vague priors? Ignorance priors? Reference priors?

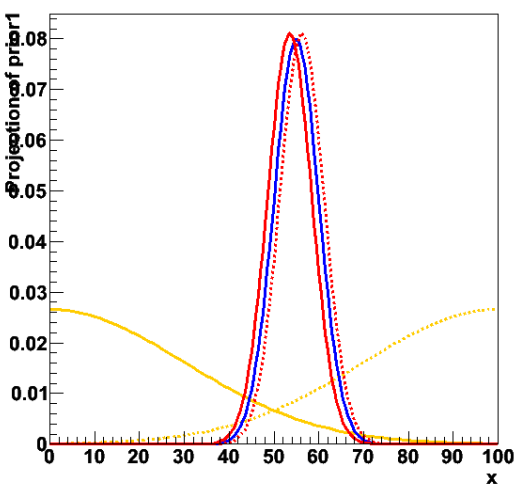
- Kassand & Wasserman who have compiled a list of them, suggest a neutral name : *Priors selected by "formal rules"*.
  - Whatever the name, keep in mind that choice of prior in one metric determines it in all other metrics: be careful in the choice of metric in which it is uniform!
  - N.B. When professional statisticians refer to "flat prior", they usually mean the Jeffreys prior.

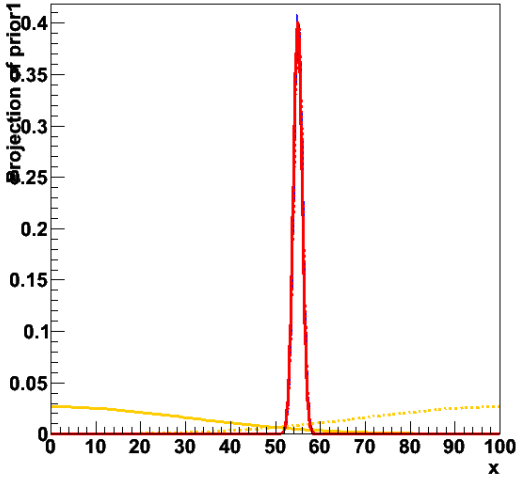
[B.Cousins HPCP]

Sensitivity Analysis

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







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

What Can Be Computed without Using a Prior?

- *Not* P(constant of nature | 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.
- Compare and contrast them with Bayesian methods.

[B.Cousins HPCP]

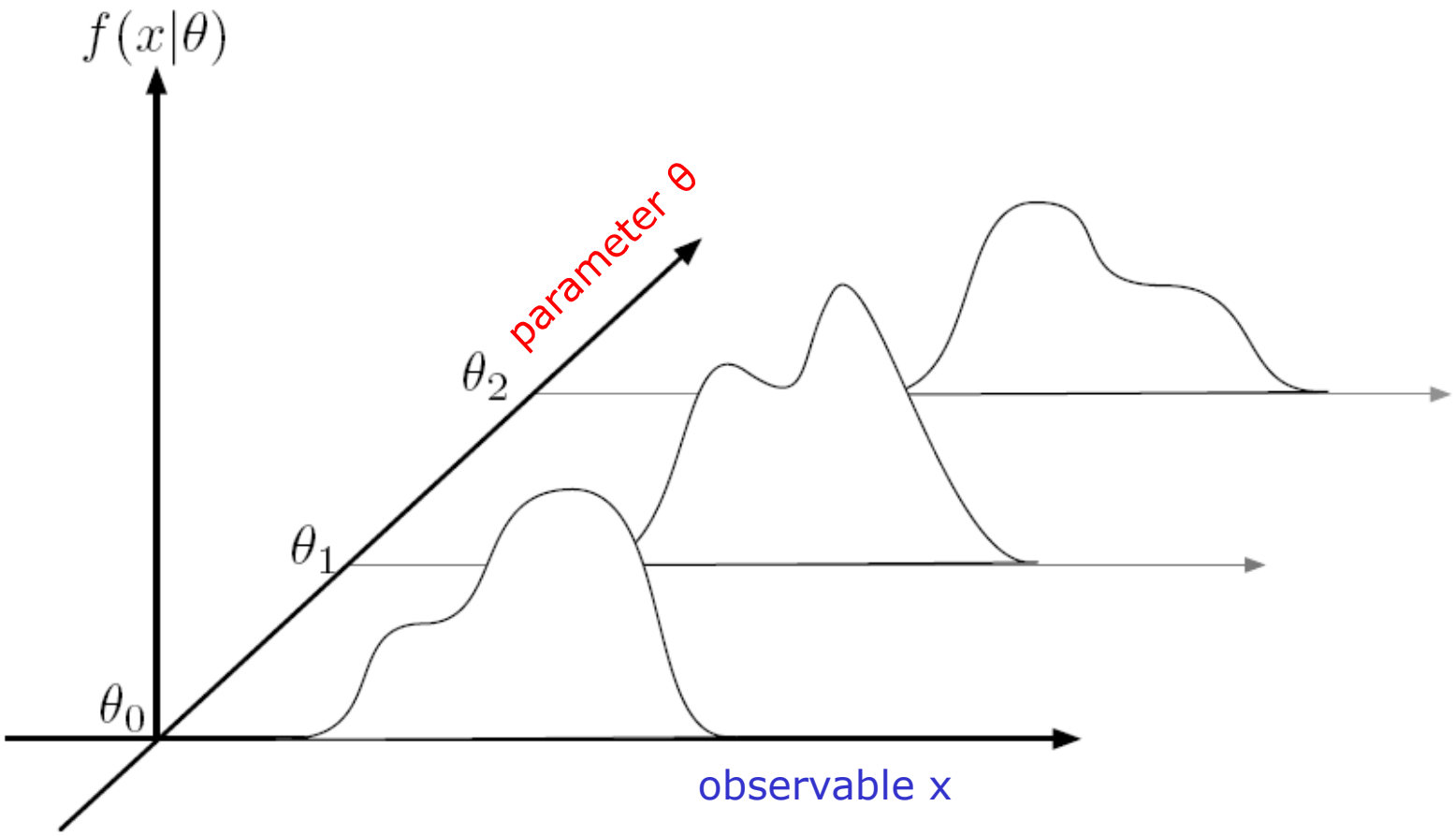
Confidence Intervals

- "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!

[B.Cousins HPCP]

How to construct a Neyman Confidence Interval

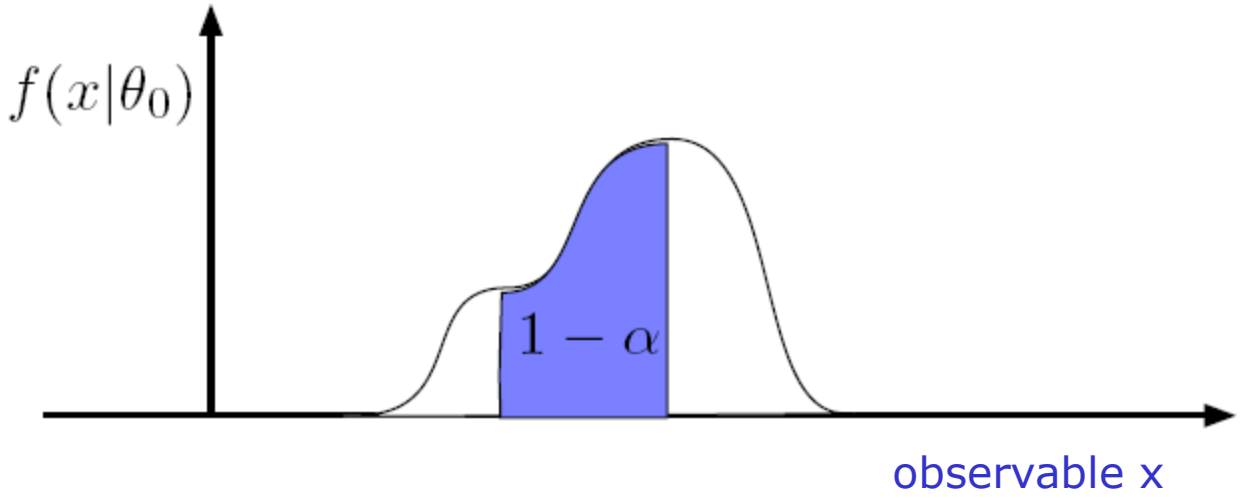
- For each value of parameter θ, determine distribution in in observable x



How to construct a Neyman Confidence Interval

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

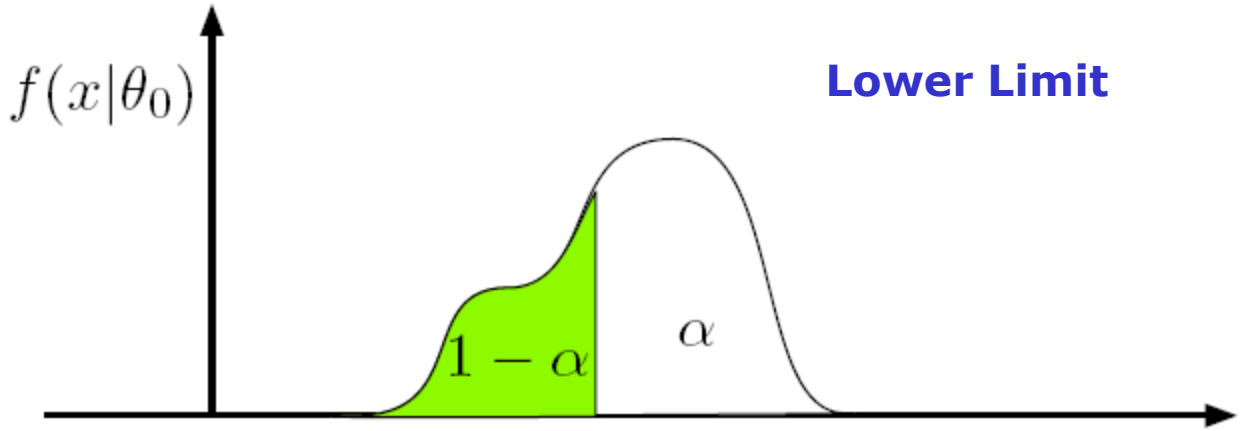
pdf for observable x
given a parameter value θ_0

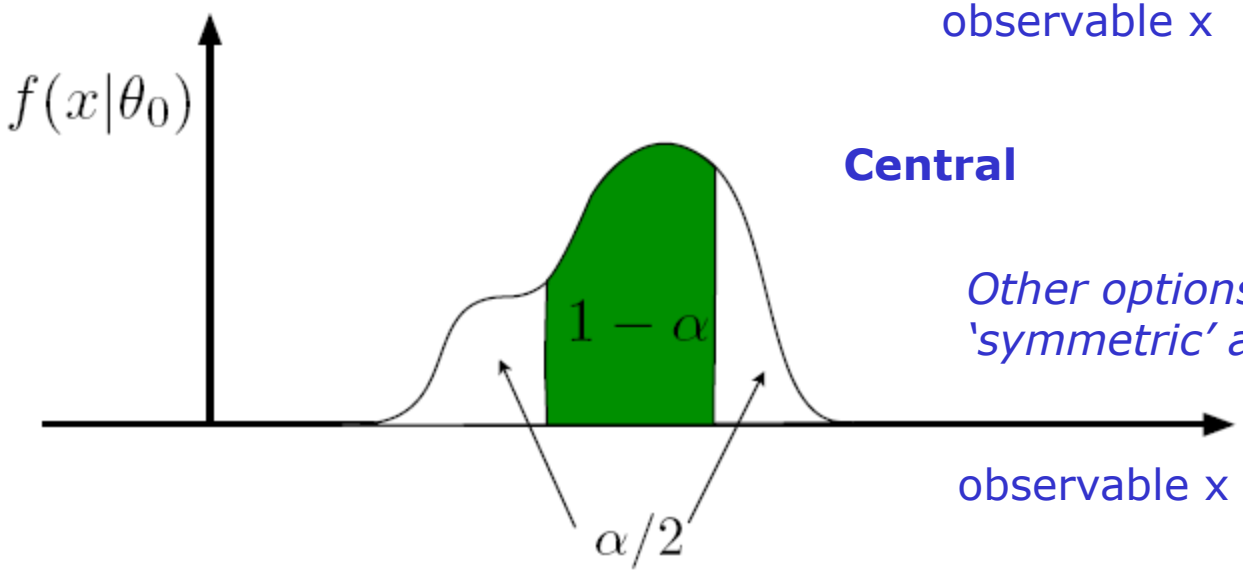


How to construct a Neyman Confidence Interval

- Definition of acceptance interval is not unique

pdf for observable x
given a parameter value θ_0

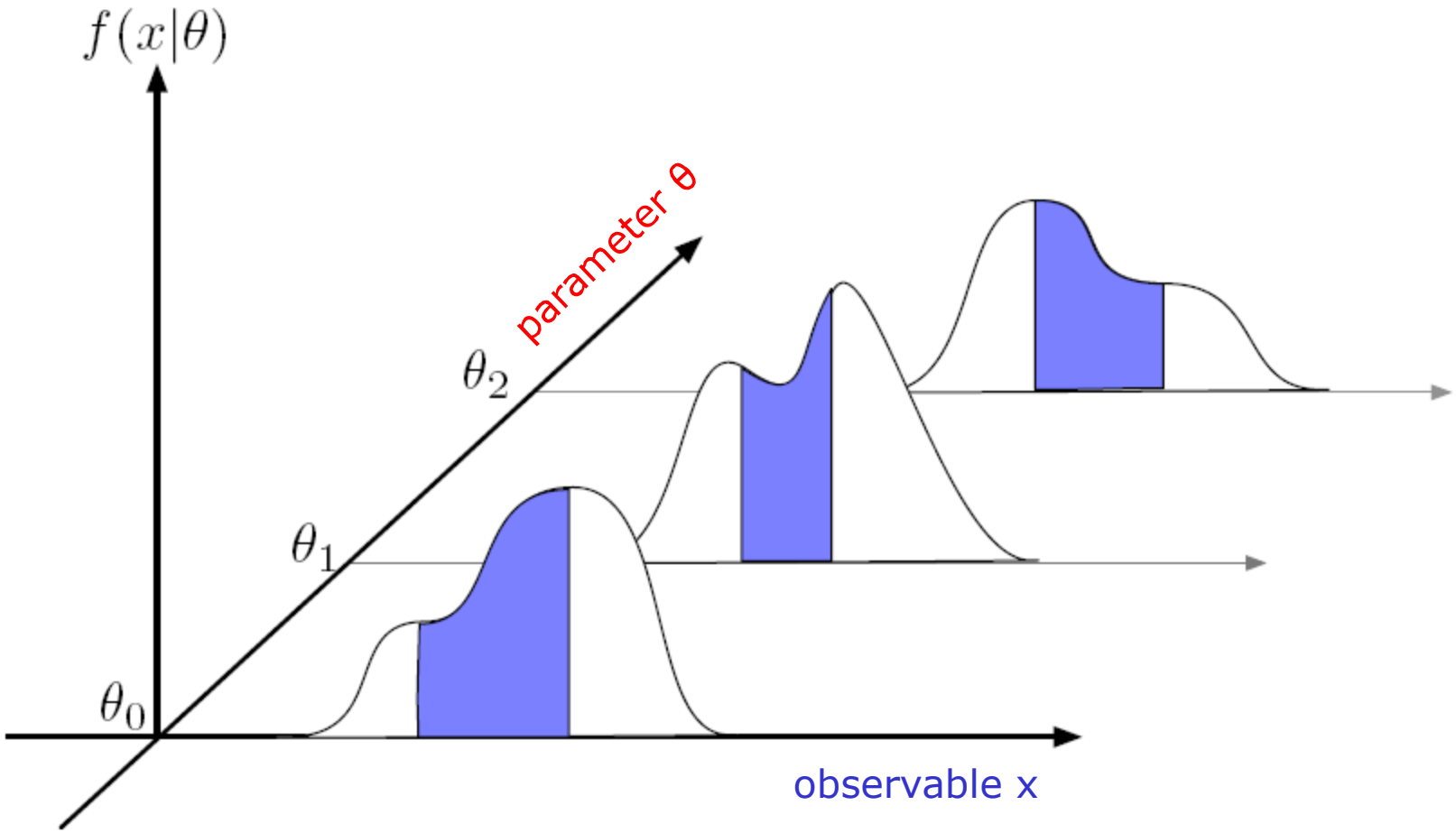




Other options, are e.g. 'symmetric' and 'shortest'

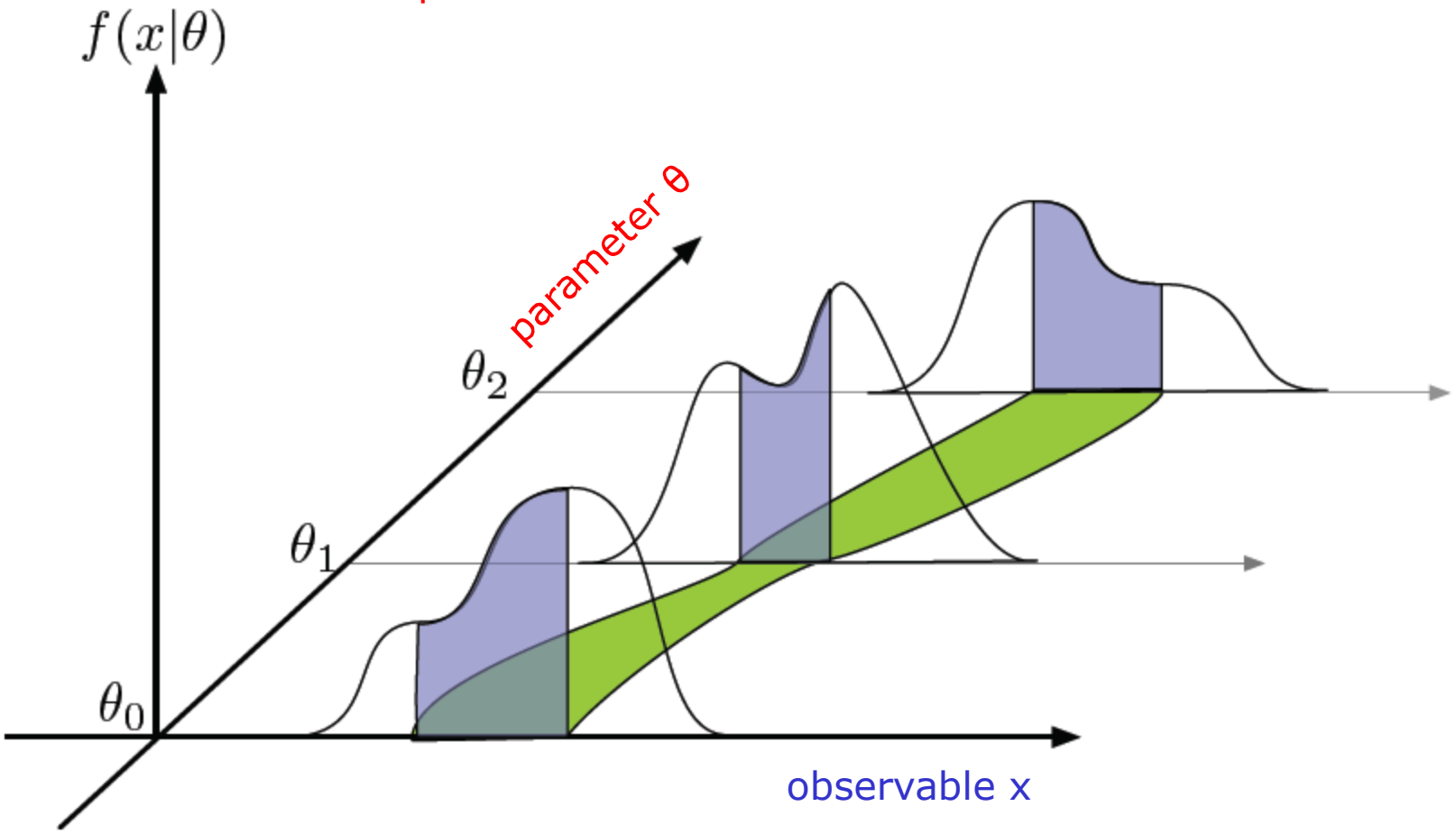
How to construct a Neyman Confidence Interval

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



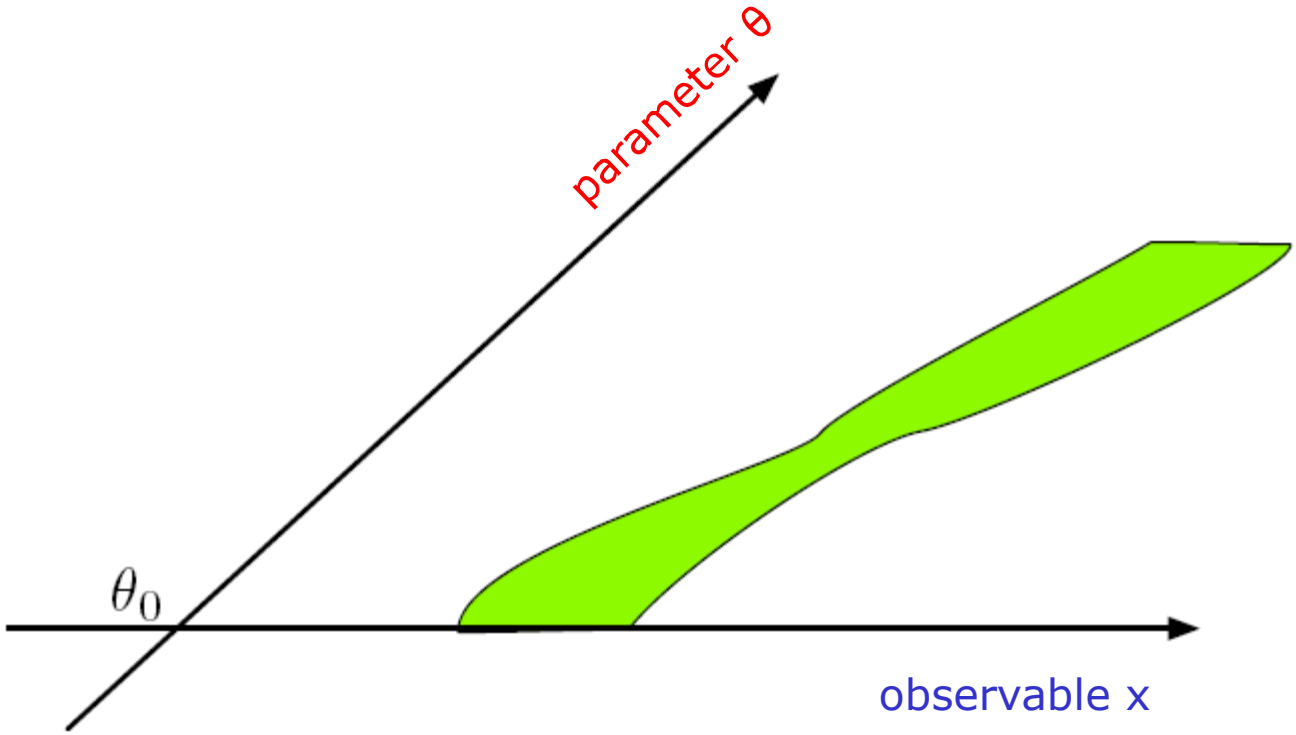
How to construct a Neyman Confidence Interval

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



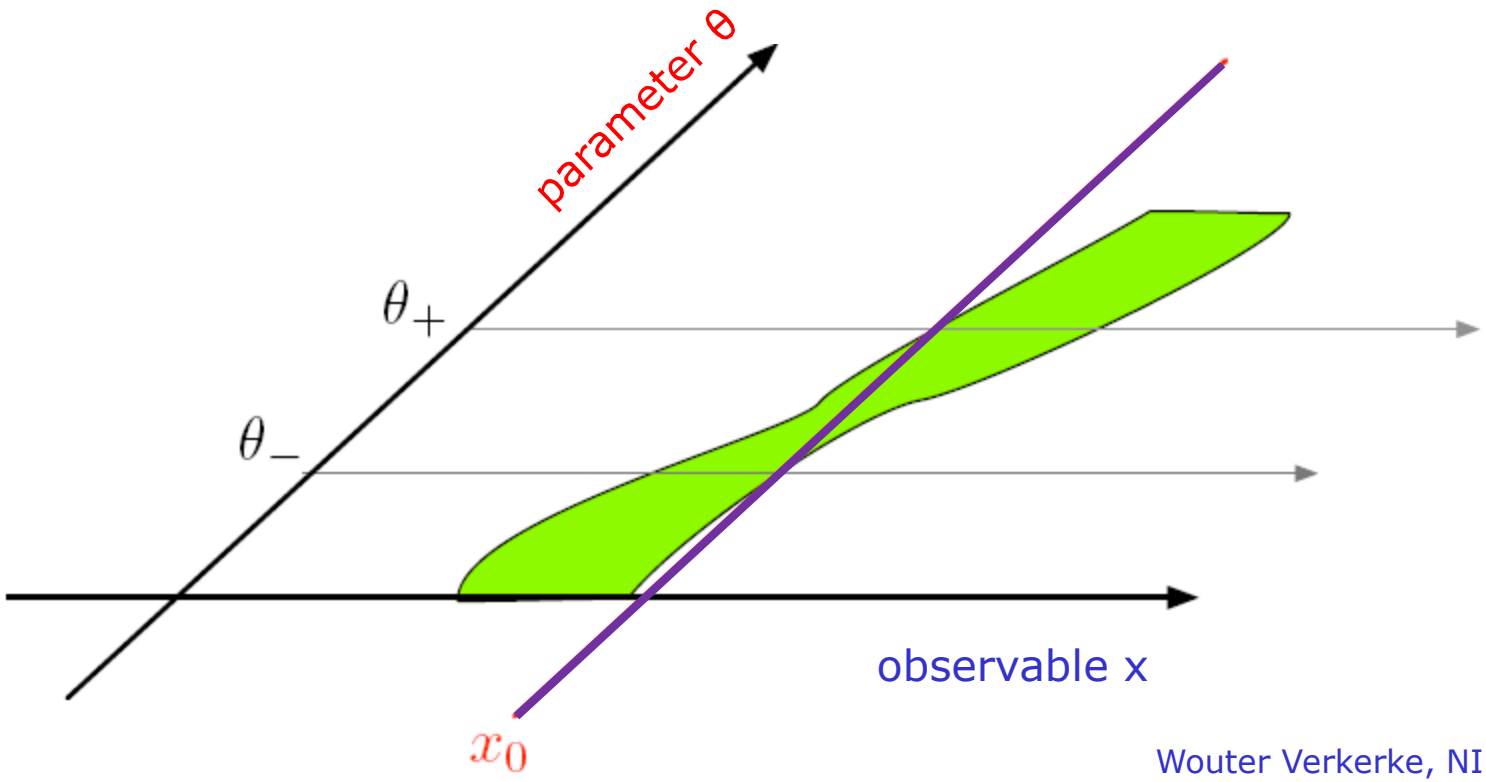
How to construct a Neyman Confidence Interval

- 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 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

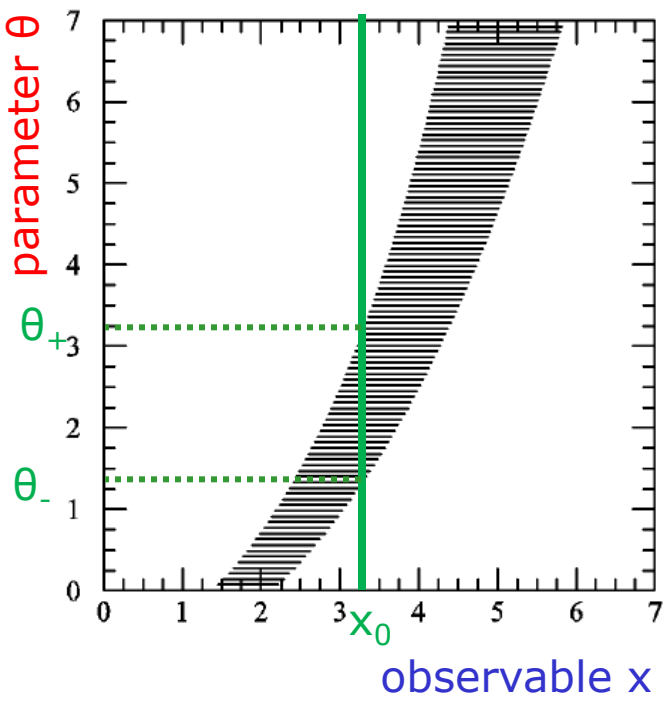
- *What does it mean?*

- Let the unknown true value of μ be μ_t.

  In repeated expt's, the confidence intervals obtained will have different endpoints $[\mu_1, \mu_2]$, since the endpoints are functions of the randomly sampled *x.*

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

  $P(\mu_t \in [\mu_1, \mu_2]) = \text{C.L.} = 1 - \alpha$.

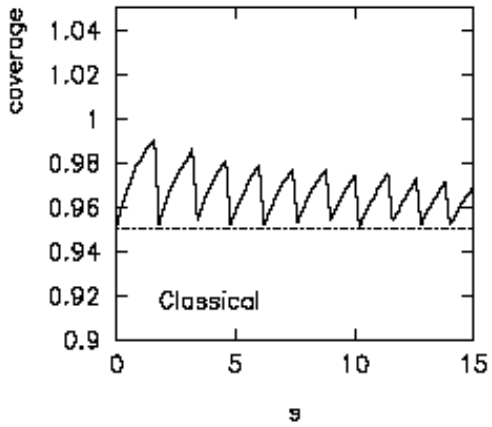


- The random variables in this equation are μ_1 and μ_2, and not μ_t,
- Coverage is a property of the set, not of an individual interval!
- It *is* true that the confidence interval consists of those values of μ for which the observed *x* is among the most probable to be observed.
  - In precisely the sense defined by the ordering principle used in the Neyman construction

[B.Cousins HPCP]

Coverage

- Coverage = Calibration of confidence interval
  - Interval has coverage if probability of true value in interval is α% for all values of mu
  - 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 → 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

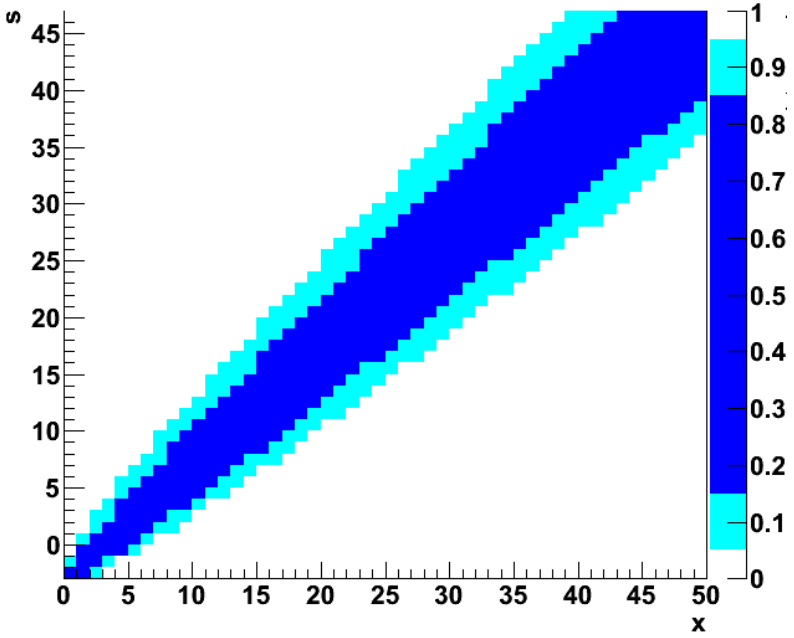


Wo

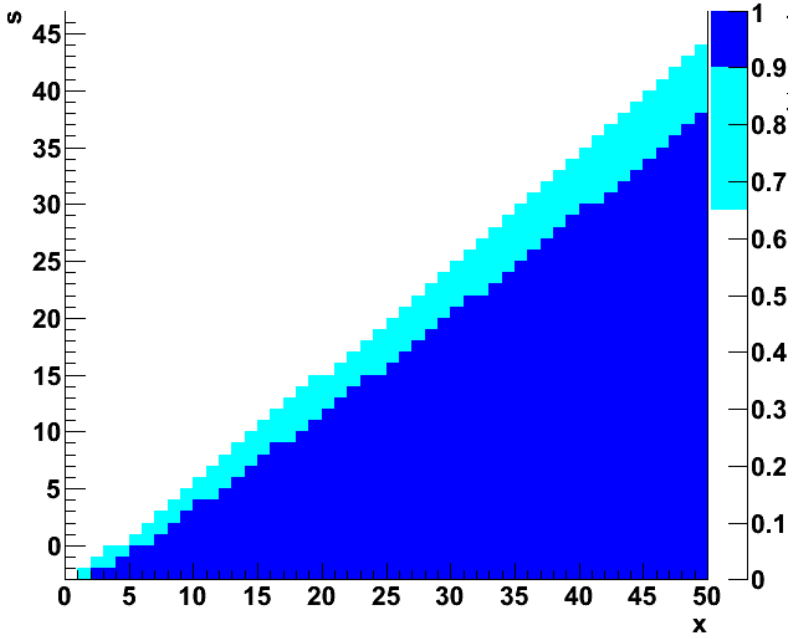
Confidence intervals for Poisson counting processes

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

Confidence belt from
68% and 90% central intervals

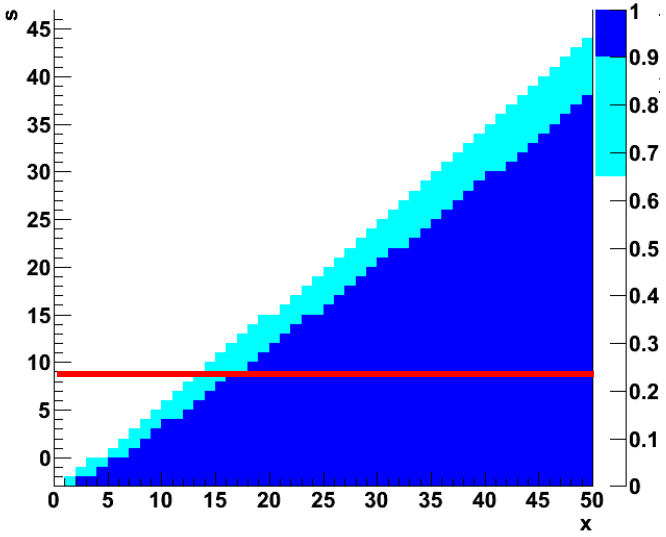


Confidence belt from
68% and 90% upper limit



Connection with hypothesis testing example

- Construction of confidence intervals and hypothesis testing closely connected.
- Going back to opening example:
  worked with P(x|μ) with μ=3 to calculate p-value → Slice at μ=3 of confidence belt



Frequentist P – working out example #2

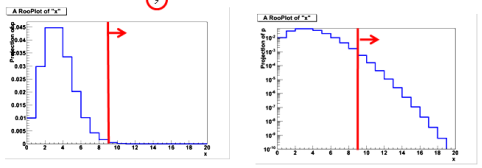
- Work out attempt #2 (aim to disprove Standard Model)
  - 'Signal' = SUSY 'Background' = Standard model

Prediction N=3
Measurement N=9

→ N(bkg) = 3
N(sig+bkg) = 9

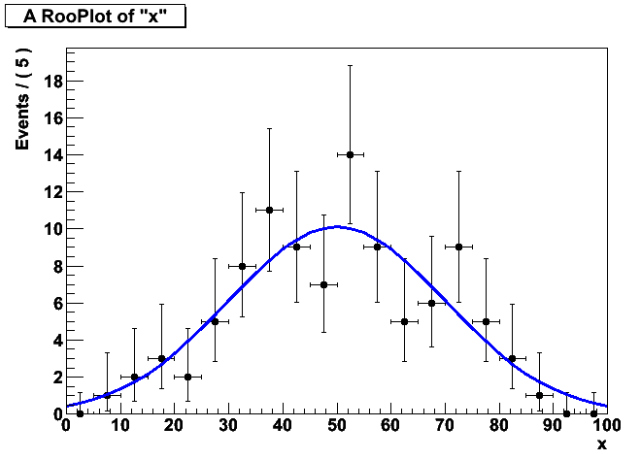
- Can we calculate probability that SM mimics SM+SUSY (i.e. result is a 'false positive)?
  - Calculation details depend on how measurement was done (fit, counting etc..)
  - Simplest case: counting experiment, Poisson process

$$p = \int_{9}^{\infty} Poisson(n, \mu = 3)\,dn = 0.0038$$ ='p value'



Confidence belts for non-trivial data

- How to construct a confidence belt if measurement is not a single observed value 'x'
  - Example: Data is histogram with 20 bins in observable x
    Model is Gaussian in x with fixed width and floating mean



- Compactify N-dimensional data point into 1-D using a 'test statistic T(x,μ)'
  - Common choice Likelihood ratio $LR(\vec{x},\mu) = L(\vec{x},\mu) / L(\vec{x},\hat{\mu})$
    - with L(μ) the likelihood of the data given parameter μ,and
    - μ-hat the value of mu with gives the lowest L(u) (i.e. the 'fitted value' of μ)

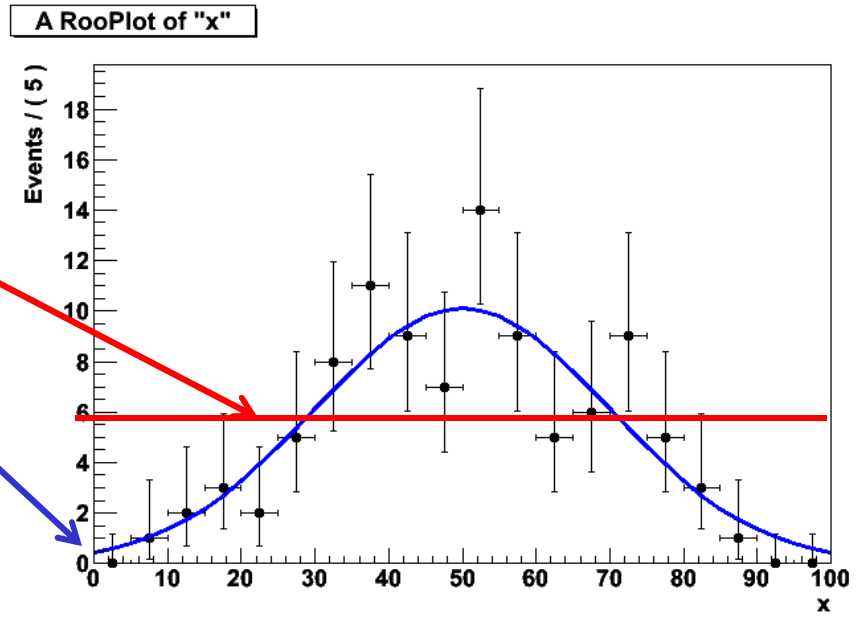
Confidence belts for non-trivial data

- Illustration of meaning of likelihood ratio

Likelihood of data for model for a given value of u

$$LR(\vec{x},\mu) = \frac{L(\vec{x},\mu)}{L(\vec{x},\hat{\mu})}$$

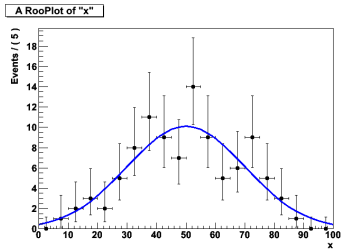
Likelihood of data for model at fitted value of μ



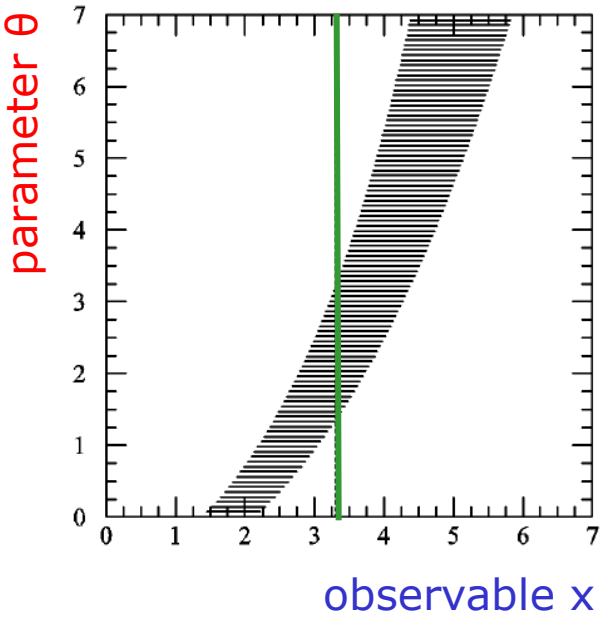
Confidence belts for non-trivial data

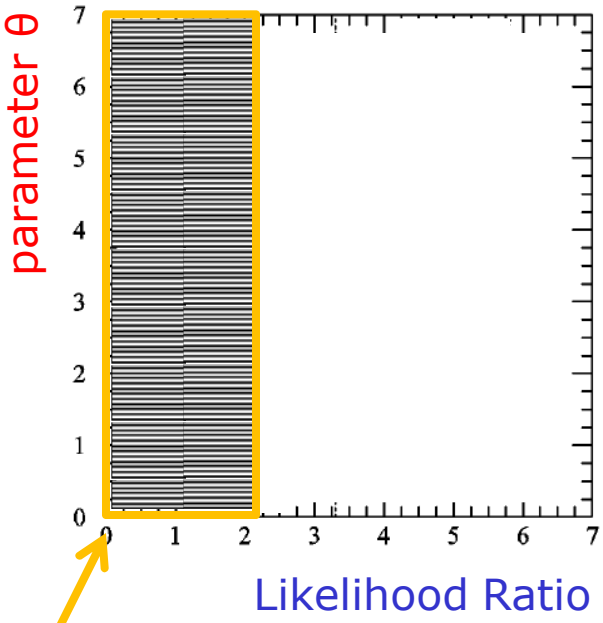
- What will the confidence belt look like when replacing $x \rightarrow LR(\vec{x}, \mu)$

x=3.2



LR(x,μ)



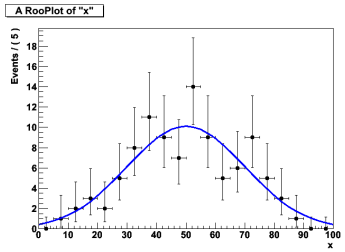


Confidence interval now range in LR

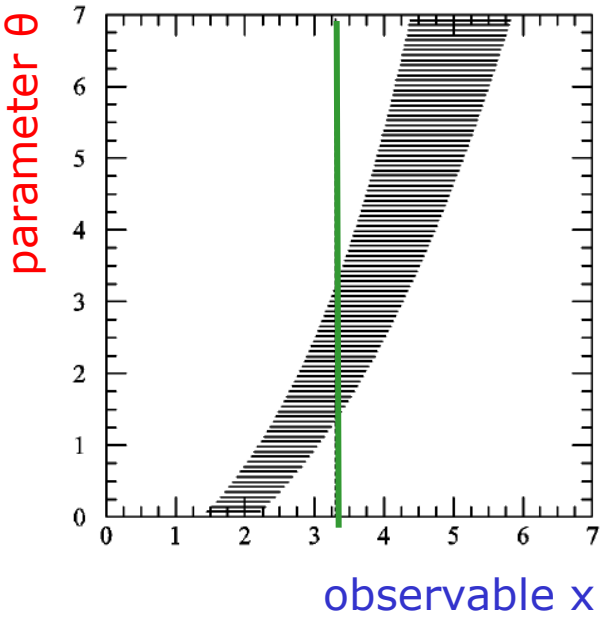
Confidence belts for non-trivial data

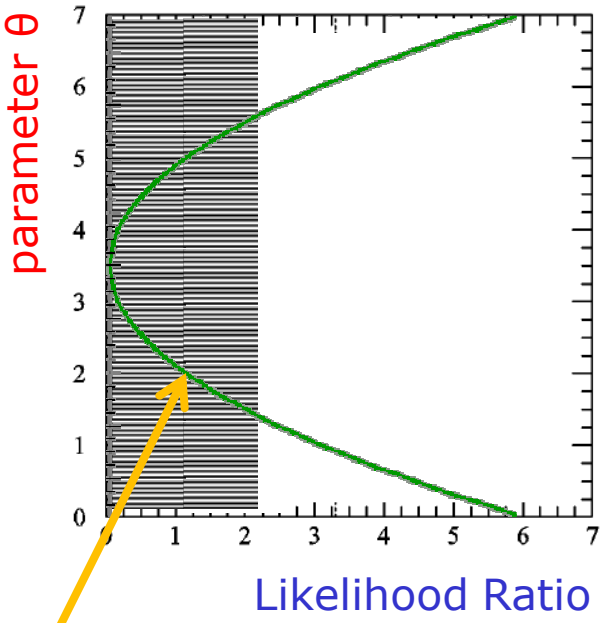
- What will the confidence belt look like when replacing $x \rightarrow LR(\vec{x}, \mu)$

x=3.2



LR(x,μ)





LR(data) is now a function of μ

Confidence belts with Likelihood Ratio ordering rule

- Note that a confidence interval with a Likelihood Ratio ordering rule (i.e. acceptance interval is defined by a range in the LR) is exactly the Feldman-Cousins interval

Unified approach to the classical statistical analysis of small signals

Gary J. Feldman*
Department of Physics, Harvard University, Cambridge, Massachusetts 02138

Robert D. Cousins†
Department of Physics and Astronomy, University of California, Los Angeles, California 90095

- One of the important features of FC that it provides a *unified method* for upper limits and central confidence intervals with good coverage
  - Upper limit at low x, central interval at higher
  - When choosing 'ad hoc' criteria to switch, good chance that your procedure doesn't have good coverage

Flip-flopping

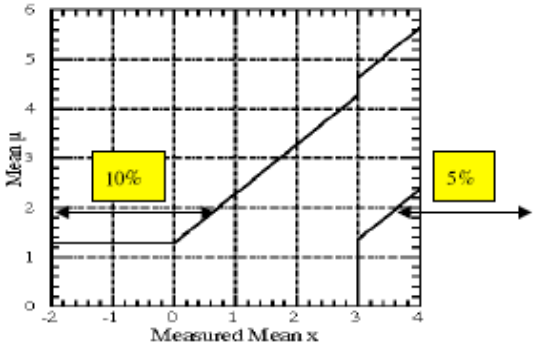
How might a typical physicist use these plots?

(1) "If the result $x < 3\sigma$, I will quote an upper limit."

(2) "If the result $x > 3\sigma$, I will quote central confidence interval."

(3) "If the result $x < 0$, I will pretend I measured zero."

This results in the following:

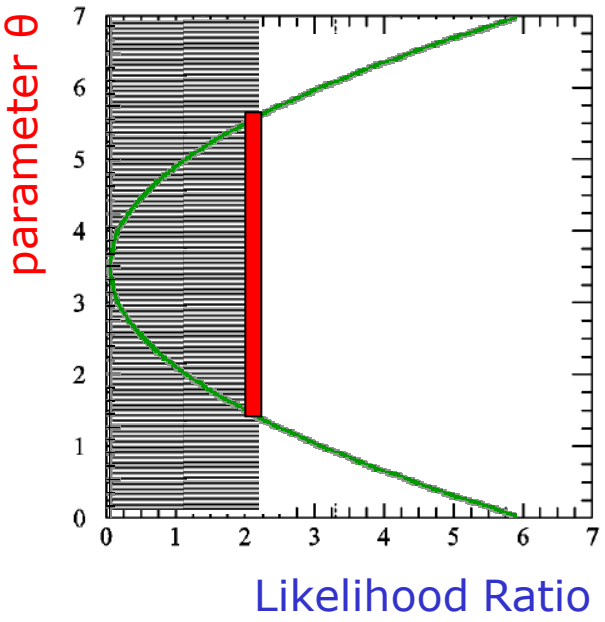


In the range $1.36 \le \mu \le 4.28$, there is only 85% coverage!

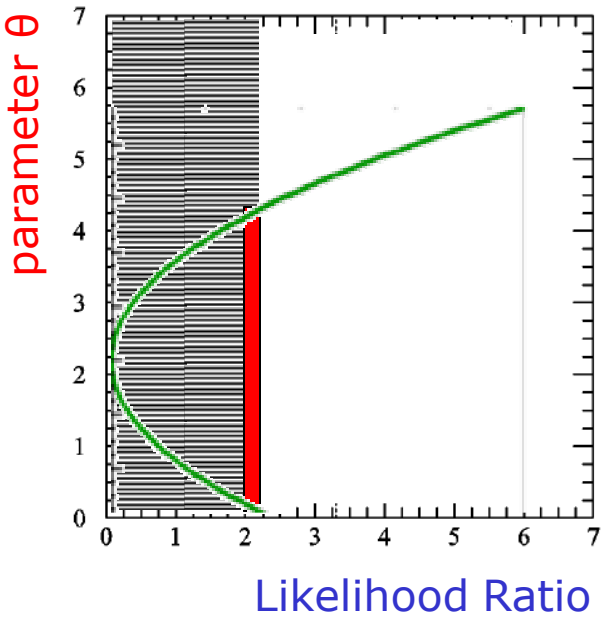
Due to flip-flopping (deciding whether to use an upper limit or a central confidence region based on the data) these are not valid confidence intervals.

Going from central to one-sided interval in FC

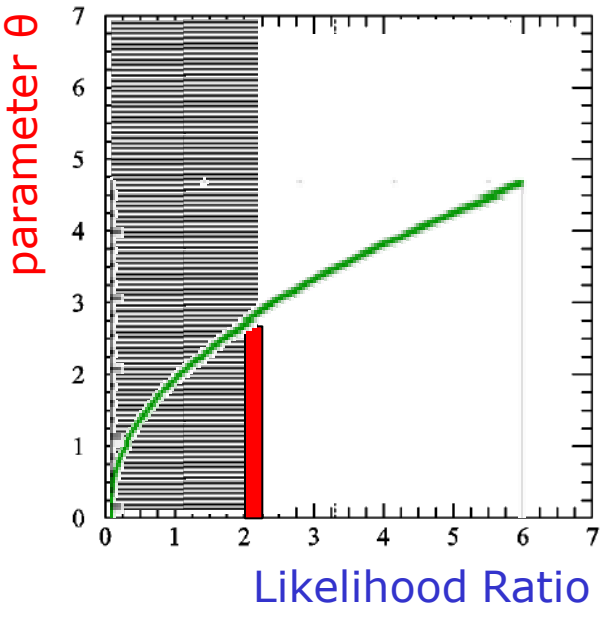
Central Interval



Transition point



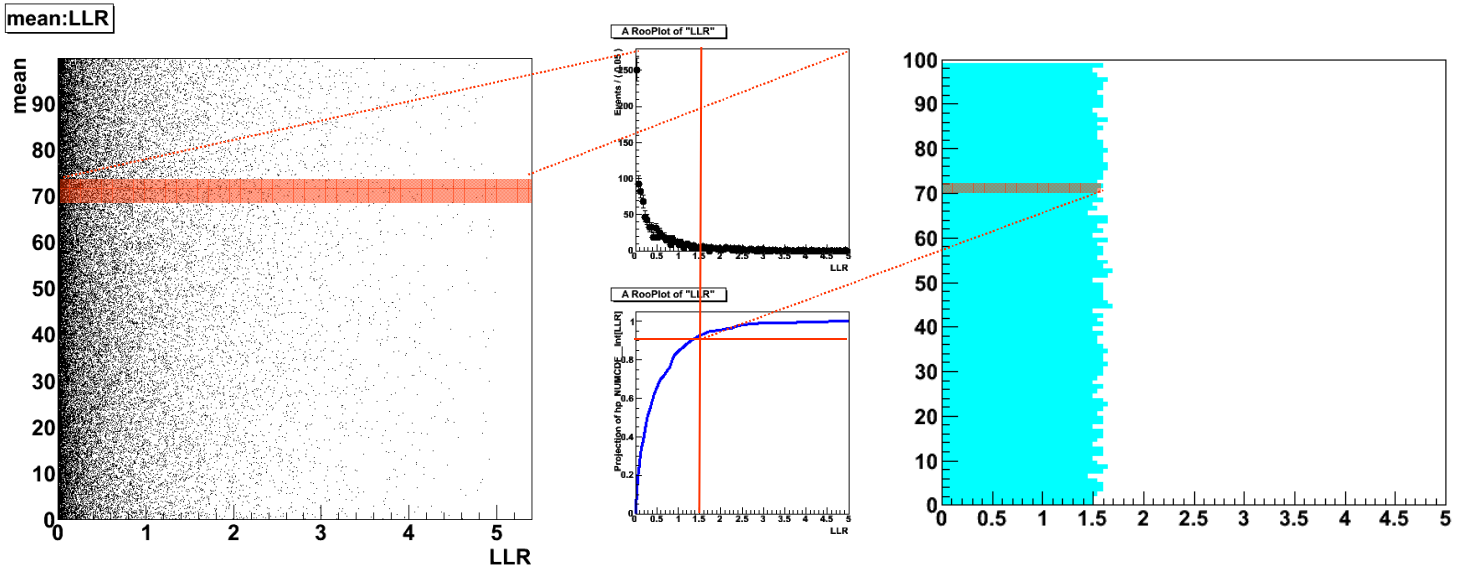
Upper limit point



Confidence belts with Likelihood Ratio ordering rule

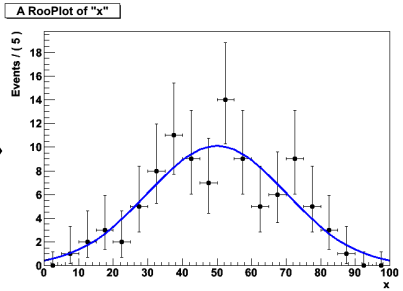
- How can we determine the shape of the confidence belt in (LR,μ) for random problem
  - In the case of the Poisson(x|s+b) confidence belt in (x,s) we could construct the belt directly from the p.d.f.
  - In rare cases you can do the same for a belt in (LR,s)

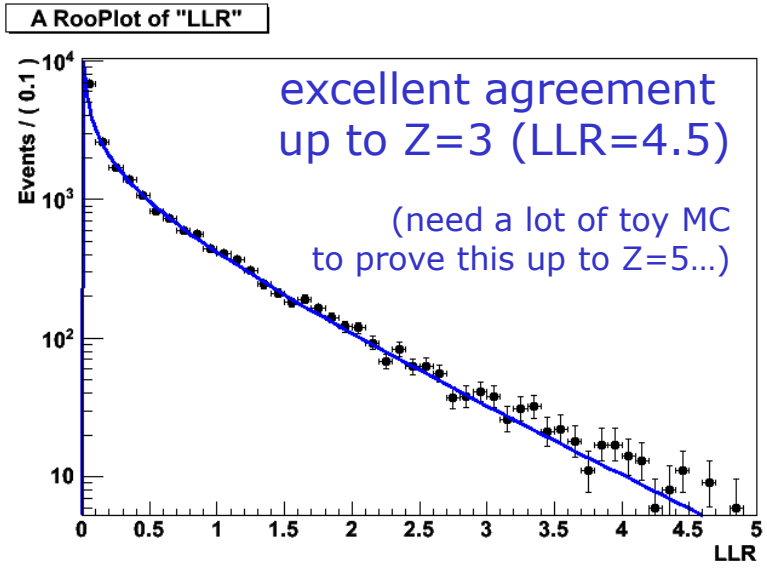
1. Calculation with toy-MC sampling
   - For each μ generate N samples of 'toy' data generated from the model F(x|μ). Calculate LR for each toy and construct distribution

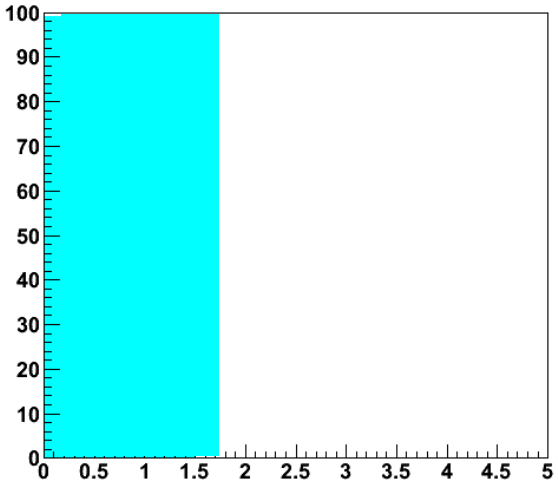


Confidence belts with Likelihood Ratio ordering rule

- Use asymptotic distribution of LR
  - Wilks theorem → Asymptotic distribution of –log(LR) is chi-squared distribution $\chi^2(2 \cdot LLR, n)$, with n the number of parameters of interest (n=1 in example shown)
  - Does ***not*** assume p.d.f.s are Gaussian
  - Example:
    LLR distribution from 100 event,
    20-bin measurement with Gaussian model
    from toy MC (histogram) vs asymptotic p.d.f



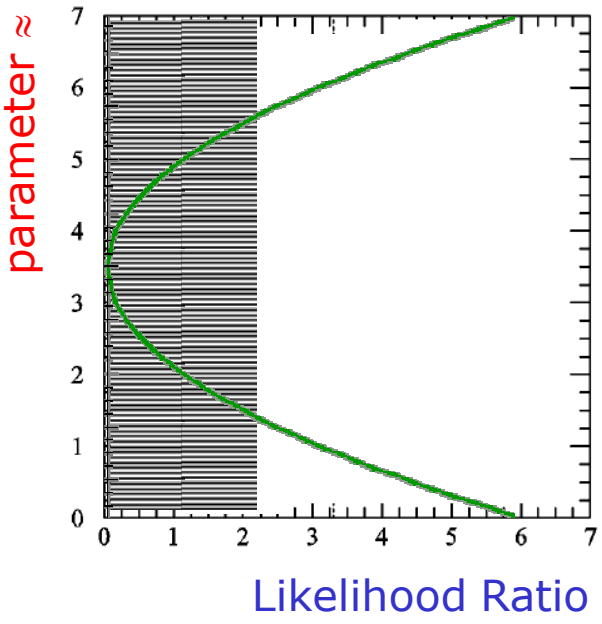




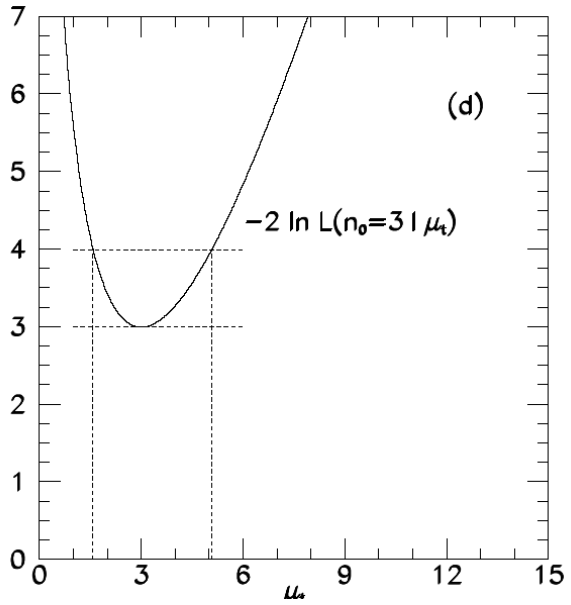
Connection with likelihood ratio intervals

- If you assume the asymptotic distribution for LLR,
  - Then the confidence belt is exactly a box
  - And the constructed confidence interval can be simplified to finding the range in μ where LLR=½·Z^2
    → This is exactly the MINOS error

FC interval with Wilks Theorem



MINOS / Likelihood ratio interval

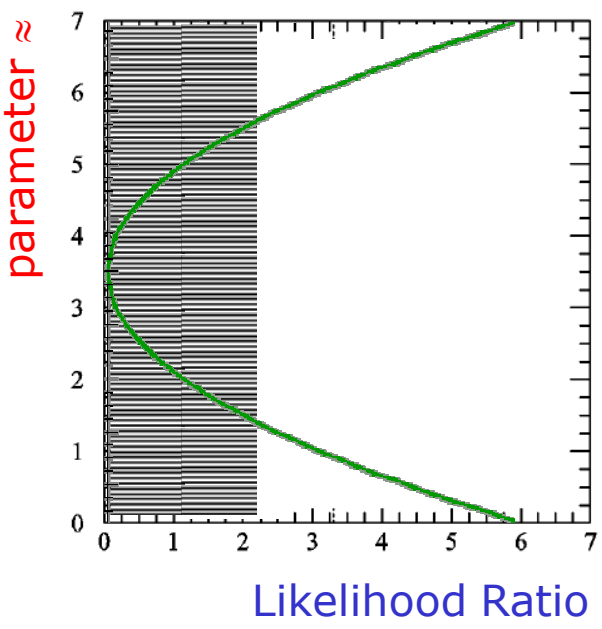


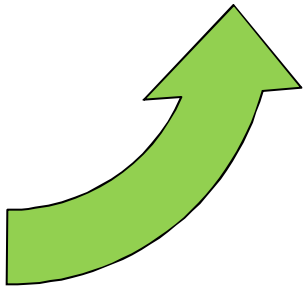
Likelihood (Ratio) Intervals

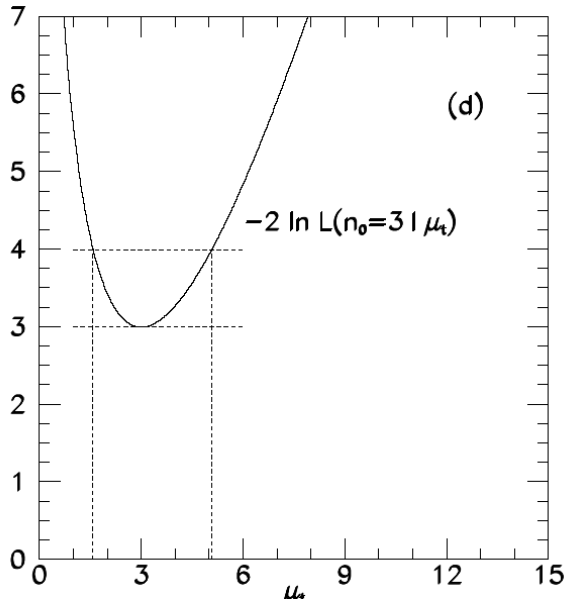
- Thus, after using maximum-likelihood method to obtain estimate û which maximizes L(u), one can obtain a likelihood interval $[u_1,u_2]$ as the union of all u for which

$$LR(u) \leq Z^2, \text{ for Z real.}$$

- But! Regularity conditions, in particular requirement that û not be on the boundary, need to be carefully checked. (E.g., if u≥0 on physical grounds, then û=0 requires care.)







Likelihood-Ratio Interval example

- 68% C.L. likelihood-ratio interval for Poisson process with n=3 observed:
- $L(\mu) = \mu^3 \exp(-\mu)/3!$
- Maximum at $\mu = 3$.
- $\Delta 2\ln L = 1^2$ for approximate ±1 Gaussian standard deviation yields interval [1.58, 5.08]

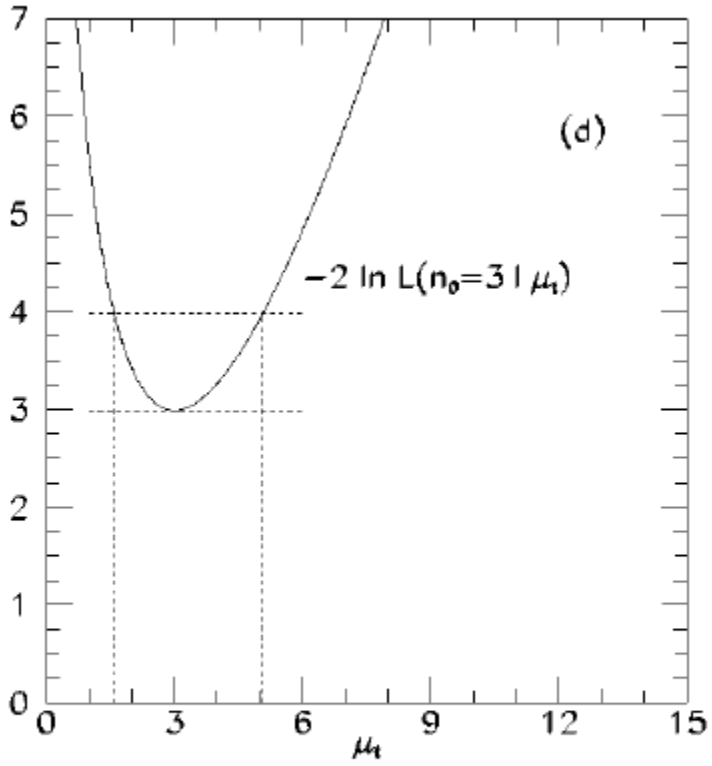
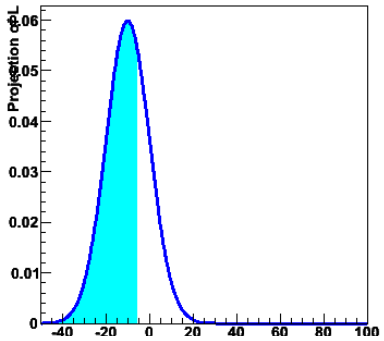


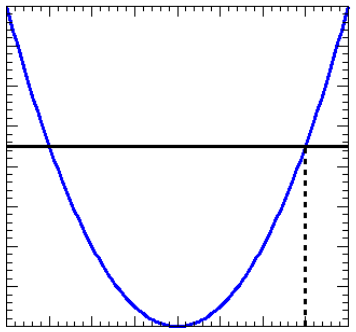
Figure from R. Cousins,
Am. J. Phys. 63 398 (1995)

U.L. in Poisson Process, n=3 observed: 3 ways

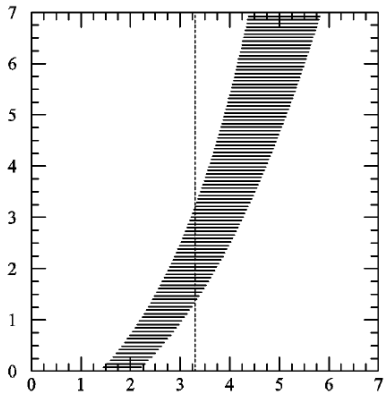
- Bayesian interval
  at 90% credibility:
  find μ_u such that posterior probability $p(\mu>\mu_u) = 0.1$.



- Likelihood ratio method for approximate 90% C.L. U.L.:
  find μ_u such that $L(\mu_u) / L(3)$ has prescribed value.



- Frequentist one-sided 90% C.L. upper limit: find μ_u such that $P(n\leq 3 \mid \mu_u) = 0.1$.



U.L. in Poisson Process, n=3 observed: 3 ways

- Deep foundational issues
  - Only #3 has guaranteed ensemble properties ("coverage") (though issues arise with systematics.) Good ?!?
  - Only #3 uses P(n|μ) for n ≠observed value. Bad?!? (See likelihood principle in next slides)

- These issues will not be resolved: aim to have software for reporting all 3 answers, and sensitivity to prior.

- Note on coverage
  - Bayesian methods do not necessarily cover (it is not their goal), but that also means you shouldn't interpret a 95% Bayesian "Credible Interval" in the same way. Coverage can be thought of as a **calibration of our statistical apparatus.**

[B.Cousins HPCP]

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

- NB: Frequentist intervals over-cover due to discreteness of n in this example
- Note that issues, divergences in outcome are usually more dramatic and important at high Z (e.g. 5σ = 'discovery')

[B.Cousins HPCP]

Expected versus observed limits

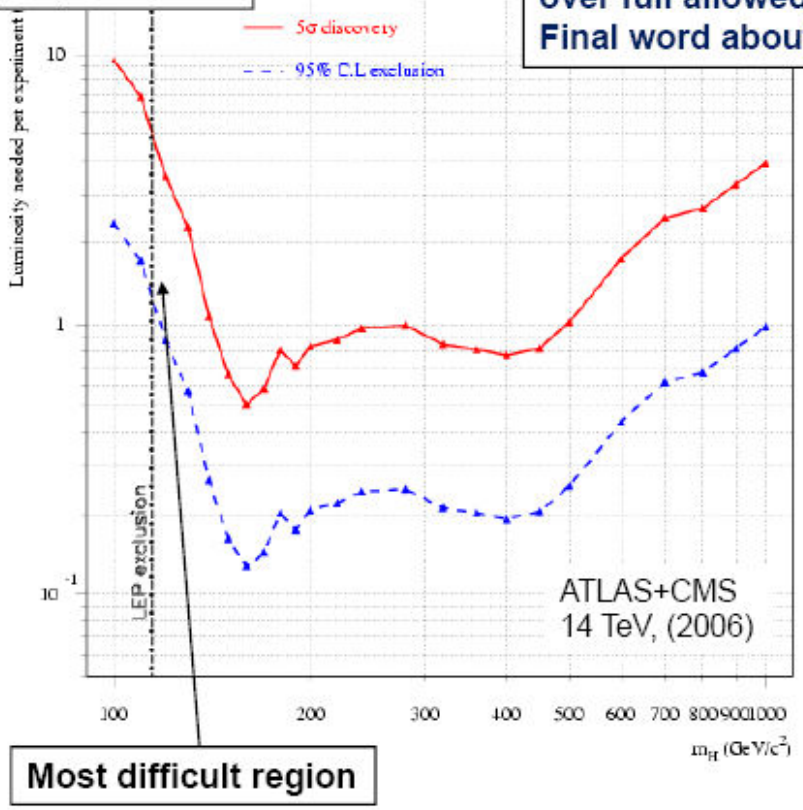
- With knowledge of your detector and the expected background you can calculate the 'expected limit' for any new discovery you'd like to make
- This tells you how sensitive your experiment is to make a discovery.
- Procedure
  - For each discovery type (e.g. Higgs at mass X GeV) run many MC studies, for each construct the limit.
  - Average of limits you get from above procedure = expected limit
  - Works in principle for any type of limit setting procedure (Bayesian, Frequentist of Likelihood)
- Two flavors of output
  - Required amount of data to make N sigma discovery → Customary when you don't have any data yet
  - expected vs observed → Customary when you have data

Example of expected limits – Higgs discovery potential

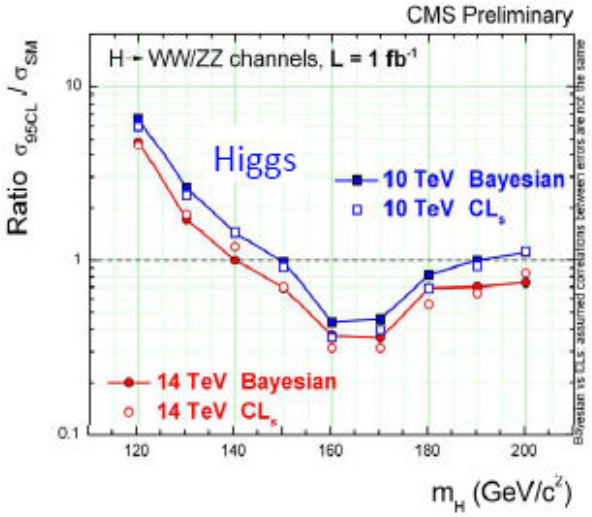
Summary of Higgs discovery potential at the LHC

Needed ∫Ldt per experiment (fb^{-1})

With
1 fb^{-1}: 95% C.L. exclusion
5 fb^{-1}: 5σ discovery
over full allowed mass range
Final word about Higgs mechanism by end 2012 ?

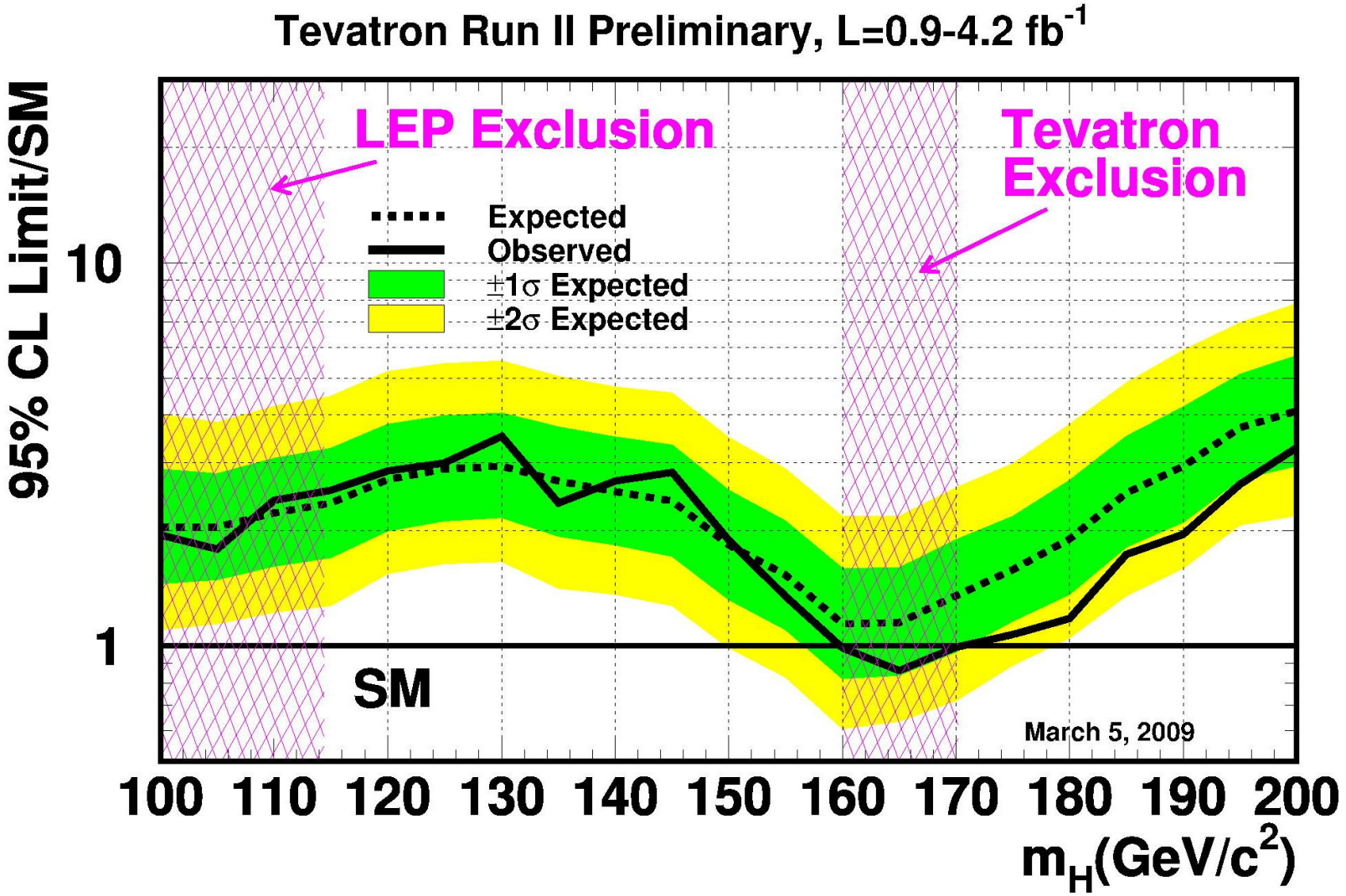


Most difficult region



With 200 pb^{-1}, reach current Tevatron sensitivity for Higgs

Example of expected vs observed



Expected versus observed limit

- If you find less ‘null hypothesis’ events than expected your observed limit will be better then expected
  - You got ‘lucky’ in terms of limit setting
- If you find more ‘null hypothesis’ events than expected your observed limit will be worse than expected
  - You’re unlucky in terms of setting a good limit
  - On the other hand it is also possible that those extra events were actually ‘signal’ → You might get lucky soon with a discovery

Likelihood Principle

- As noted above, in both **Bayesian** methods and **likelihood-ratio** based methods, the probability (density) for obtaining the *data at hand is used (via the likelihood function), but probabilities for obtaining other data are not used!*
- In contrast, in typical **frequentist** calculations (e.g., a p-value which is the probability of obtaining a value as extreme or *more extreme than that observed), one uses probabilities of data not seen.*
- This difference is captured by the *Likelihood Principle**: If two experiments yield likelihood functions which are proportional, then Your inferences from the two experiments should be identical.
- L.P. is built in to Bayesian inference (except e.g., when Jeffreys prior leads to violation).
- L.P. is violated by p-values and confidence intervals.
- Although practical experience indicates that the L.P. may be too restrictive, it is useful to keep in mind. When frequentist results "make no sense" or "are unphysical" the underlying reason might be traced to a bad violation of the L.P.
- *There are various versions of the L.P., strong and weak forms, etc. See Stuart99 and book by Berger and Wolpert.

[B.Cousins HPCP]

Likelihood Principle Example #1

- The "Karmen Problem"
  - You expect background events sampled from a Poisson mean b=2.8, assumed known precisely.
  - For signal mean μ, the total number of events n is then sampled from Poisson mean μ+b.
  - So $P(n) = (\mu+b)^n \exp(-\mu-b) / n!$
  - Then you see no events at all! I.e., n=0.
  - $L(\mu) = (\mu+b)^0 \exp(-\mu-b) / 0! = \exp(-\mu) \exp(-b)$

- Note that changing b from 0 to 2.8 changes L(μ) only by the constant factor exp(-b).
  - This gets renormalized away in any Bayesian calculation, and is irrelevant for likelihood *ratios.*
- So for zero events observed, likelihood-based inference about signal mean μ *is independent of expected b.*

- For essentially all frequentist confidence interval constructions, the fact that n=0 is less likely for b=2.8 than for b=0 results in *narrower* confidence intervals for μ as b increases.
  - Clear violation of the L.P.

[B.Cousins HPCP]

Likelihood Principle Example #2

- Binomial problem famous among statisticians

- Translated to HEP: You want to know the trigger efficiency *e*.
  - You count until reaching n=4000 zero-bias events, and note that of these, m=10 passed trigger.

    Estimate e = 10/4000, compute binomial conf. interval for e.

  - Your colleague (in a different sample!) counts zero-bias events until m=10 have passed the trigger. She notes that this requires n=4000 events.

    Intuitively, e=10/4000 *over-estimates* e because she stopped *just* upon reaching 10 passed events. (The relevant distribution is the negative binomial.)

- Each experiment had a different *stopping rule*. Frequentist confidence intervals depend on the stopping rule.
  - It turns out that the likelihood functions for the binomial problem and the negative binomial problem differ only by a constant! So with same n and m, (the strong version of) the L.P. demands *same* inference about e from the two stopping rules!

[B.Cousins HPCP]

Likelihood Principle Discussion

- We will not resolve this issue, but should be aware of it.
- If you are interested, read the book by Berger & Wolpert, but be prepared for the stopping rule arguments to set your head spinning.
- *Irrelevance* of the Stopping Rule is known as the "Stopping Rule Principle" and has been hotly debated for decades, with some famous statisticians changing their minds, e.g:
  - L.J. "Jimmie" Savage is widely quoted as saying in 1962, "I learned the stopping-rule principle from Professor Barnard in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resent an idea so patently right."

[B.Cousins HPCP]

Conditioning*

- An "ancillary statistic" (see literature for precise math definition) is a function of your data which carries information about the precision of your measurement of the parameter of interest, but no info about parameter's value.
- The classic example is a branching ratio measurement in which the total number of events N can fluctuate if the expt design is to run for a fixed length of time. Then N is an ancillary statistic.
- You perform an experiment and obtain N total events, and then do a toy M.C. of repetitions of the experiment. Do you let N fluctuate, or do you fix it to the value observed?
- It may seem that the toy M.C. should include your *complete* procedure, including fluctuations in N.
- But there are strong arguments, going back to Fisher, that inference should be based on probabilities *conditional* on the value of the ancillary statistic actually obtained!

[B.Cousins HPCP]

Conditioning (cont.)

- The 1958 thought expt of David R. Cox focused the issue:
  - Your procedure for weighing an object consists of flipping a coin to decide whether to use a weighing machine with a 10% error or one with a 1% error; and then measuring the weight. (Coin flip result is ancillary stat.)
  - Then "surely" the error you quote for your measurement should reflect which weighing machine you actually used, and not the average error of the "whole space" of all measurements!
  - But classical most powerful Neyman-Pearson hypothesis test uses the whole space!
- In more complicated situations, ancillary statistics do not exist, and it is not at all clear how to restrict the "whole space" to the relevant part for frequentist coverage.
- In methods obeying the likelihood principle, in effect one conditions on the exact data obtained, giving up the frequentist coverage criterion for the guarantee of relevance

[B.Cousins HPCP]

Summary of Three Ways to Make Intervals

	Bayesian Credible	Frequentist Confidence	Likelihood Ratio
Requires prior pdf?	Yes	No	No
Obeys likelihood principle?	Yes (exception re Jeffreys prior)	No	Yes
Random variable in "$P(\mu_t \in [\mu_1, \mu_2])$":	μ_t	μ_1, μ_2	μ_1, μ_2
Coverage guaranteed?	No	Yes (but over-coverage…)	No
Provides P(parameter|data)?	Yes	No	No