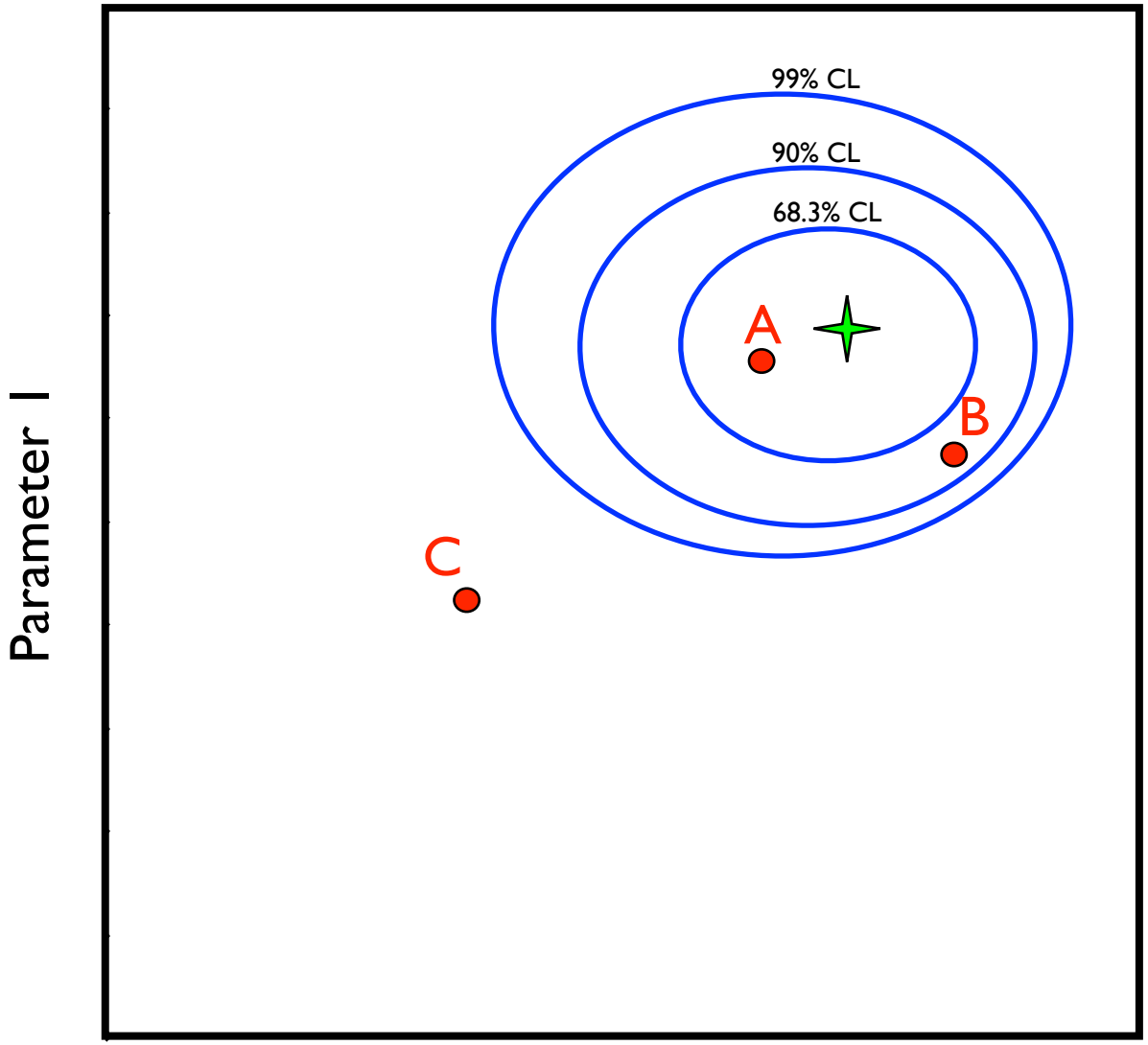
Lecture 7:

- Confidence and Credibility




Consider a single experiment in which 2 parameters are measured (✦) and compared with predictions from 3 different theoretical models (A, B, C)



"Another Look at Confidence Intervals: Proposal for a More Relevant and Transparent Approach"
Biller and Oser,
NIM A 774 (2015) 103-119
arXiv:1405.5010

Frequentist Confidence Intervals

Construction of Frequentist Confidence Intervals via Wilks' Theorem

We've been here before…

$$-2[\ln L(\mathbf{q_o}) - \ln L(\mathbf{q})] = -2\ln\left(\frac{L(\mathbf{q_o})}{L(\mathbf{q})}\right) \equiv -2\ln L_{\mathbf{R}} \sim \chi^2_d$$

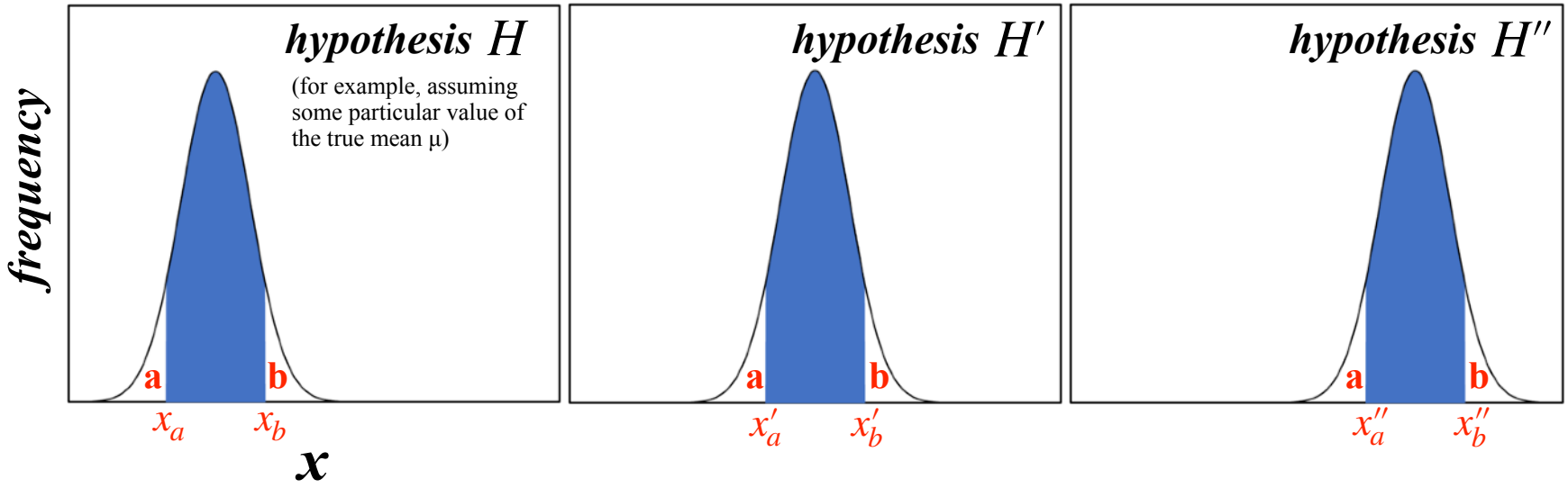
where $\mathbf{q_o}$ are the set of model parameters that define the default (null) hypothesis, and the d = DoF = the difference in the number of model parameters constrained (*i.e.* how many extra degrees of freedom one model has compared to the other)

Legal Statement:

- *For nested hypotheses (i.e. a continuous transition from one hypothesis to the next)*
- *Away from boundaries in likelihood space*
- *In the limit of large amounts of data*

Because this is an approximation, perfect statistical coverage is not guaranteed… but is is usually pretty close for most cases you will encounter, and actually works pretty well for counting statistics even for small numbers. For more unusual cases, the validity can often be "spot-checked" with Monte Carlo calculations.

Neyman Construction of Frequentist Confidence Intervals



$$CL = 1 - a - b$$

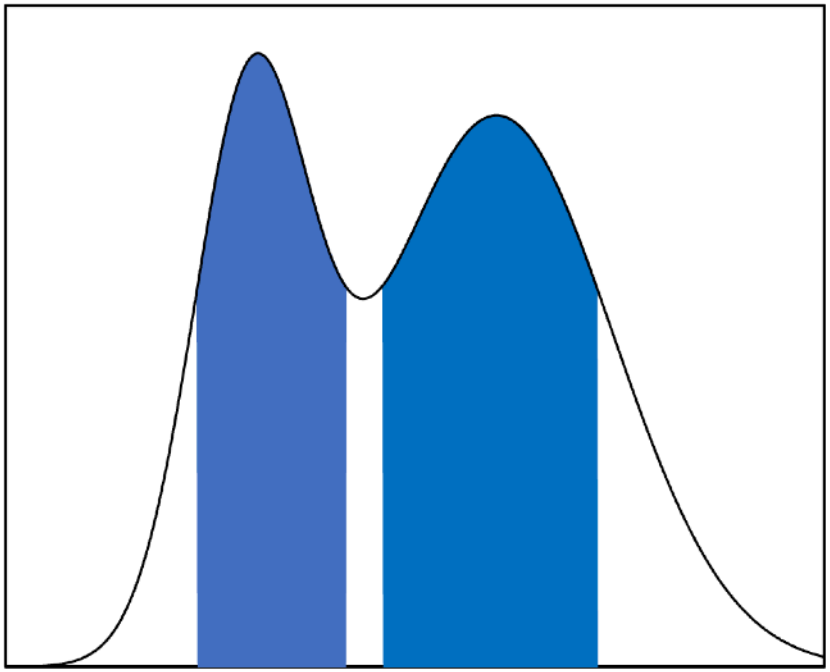
(where "Confidence Level" refers to the frequency of hypothetical measurements landing in the defined region for a given model)

... etc.

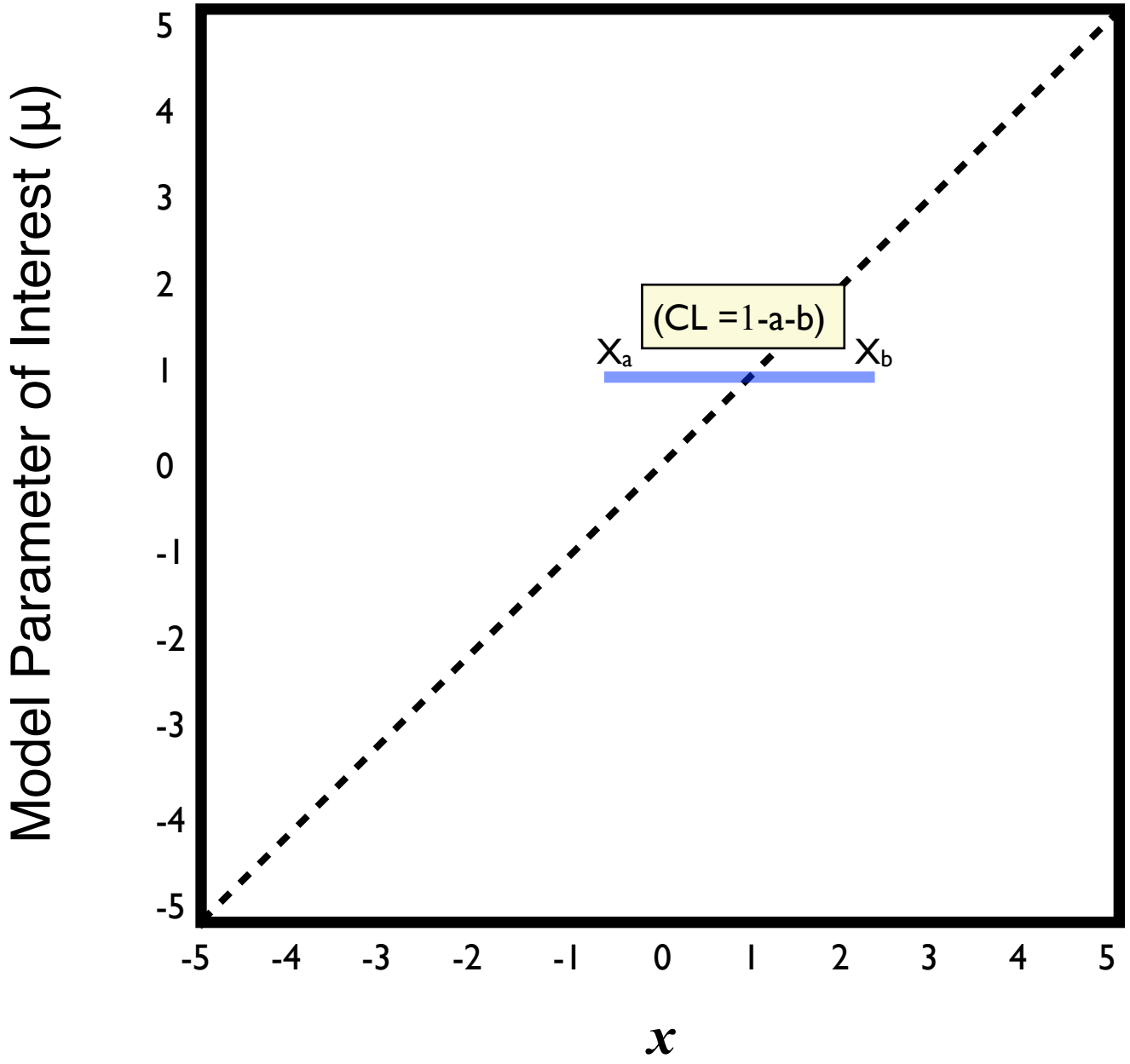
x is an "ordering parameter," which can be a direct measurable (such as the number of counts) or can be a derived quantity (such as a likelihood ratio)

Note that the fraction of models to be included in a particular CL interval can be chosen in a number of different of ways to yield, for example: upper bounds, lower bounds, central intervals, most compact interval, or intervals containing the highest probability densities

useful for more complicated cases, such as multi-modal distributions

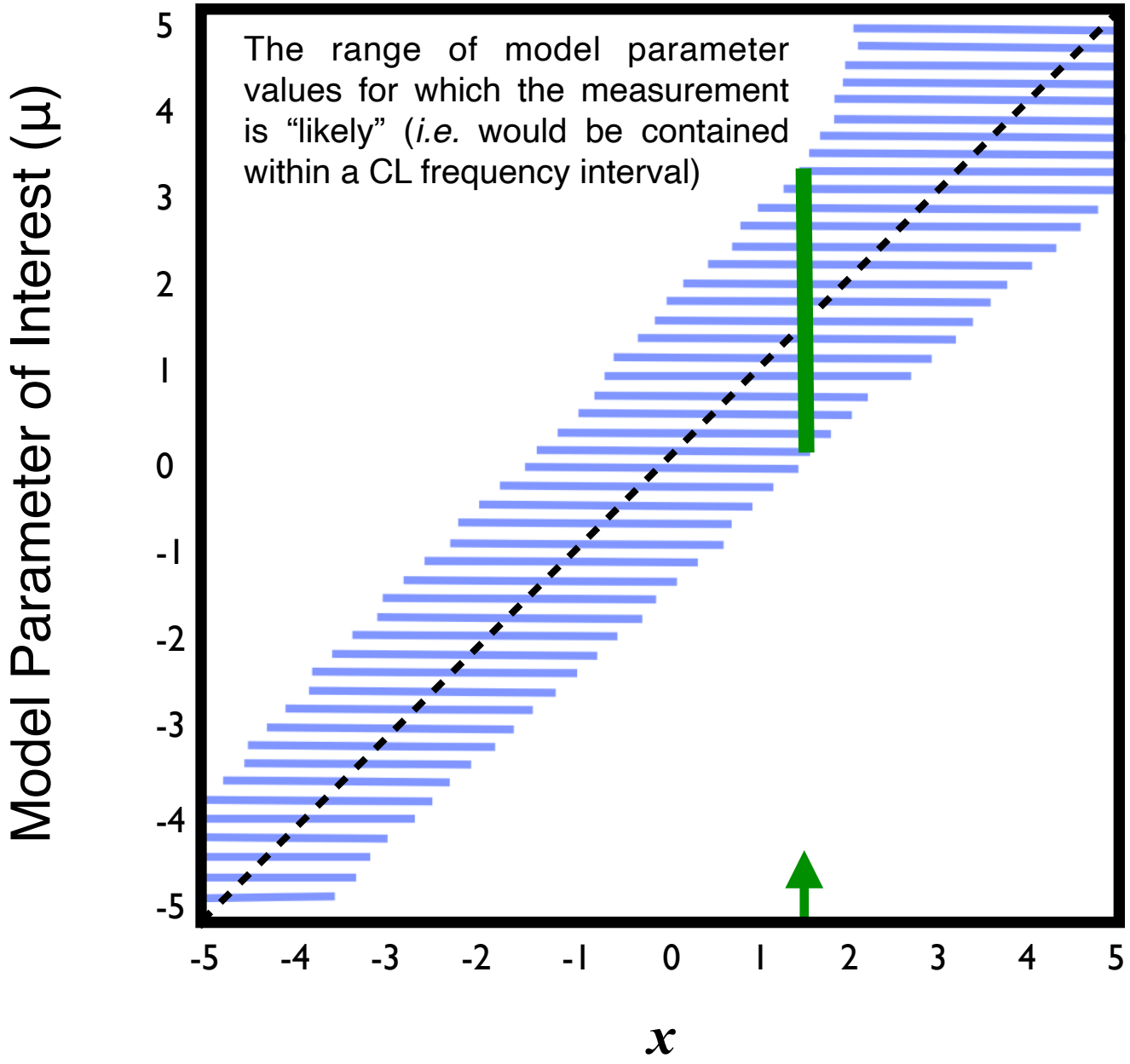


Neyman Construction of Frequentist Confidence Intervals



In the example here, let's assume that the measurement x is an unbiased estimator for the model parameter μ

Neyman Construction of Frequentist Confidence Intervals



In the example here, let's assume that the measurement x is an unbiased estimator for the model parameter μ

Example: Find the standard frequentist CL upper bound on the mean signal strength, **S**, for a counting experiment where the expected background level is **B** and a total of **n** events are observed.

For a given model of signal strength, S, the observable number of counts would follow a Poisson distribution. Given a fixed observed value of **n**, we then want to find the range of models, from S=0 to S_{max}, that would be contained in a CL fraction of repeated experiments:

$$\int_0^{S_{max}} \frac{(S+B)^n e^{-(S+B)}}{n!} = CL$$

It can be shown, from repeated integration by parts, that this is equivalent to:

$$\sum_{m=0}^{n} \frac{(S_{max}+B)^m e^{-(S_{max}+B)}}{m!} = 1 - CL$$

Then solve numerically for S_{max}

Note that there is no constraint to restrict the background from being greater than the observed number of counts!! This is because we are interested in the ***average*** background over an ensemble of experiments, not the particular background for this measurement. **Frequentists only care about the ensemble, not about you!**

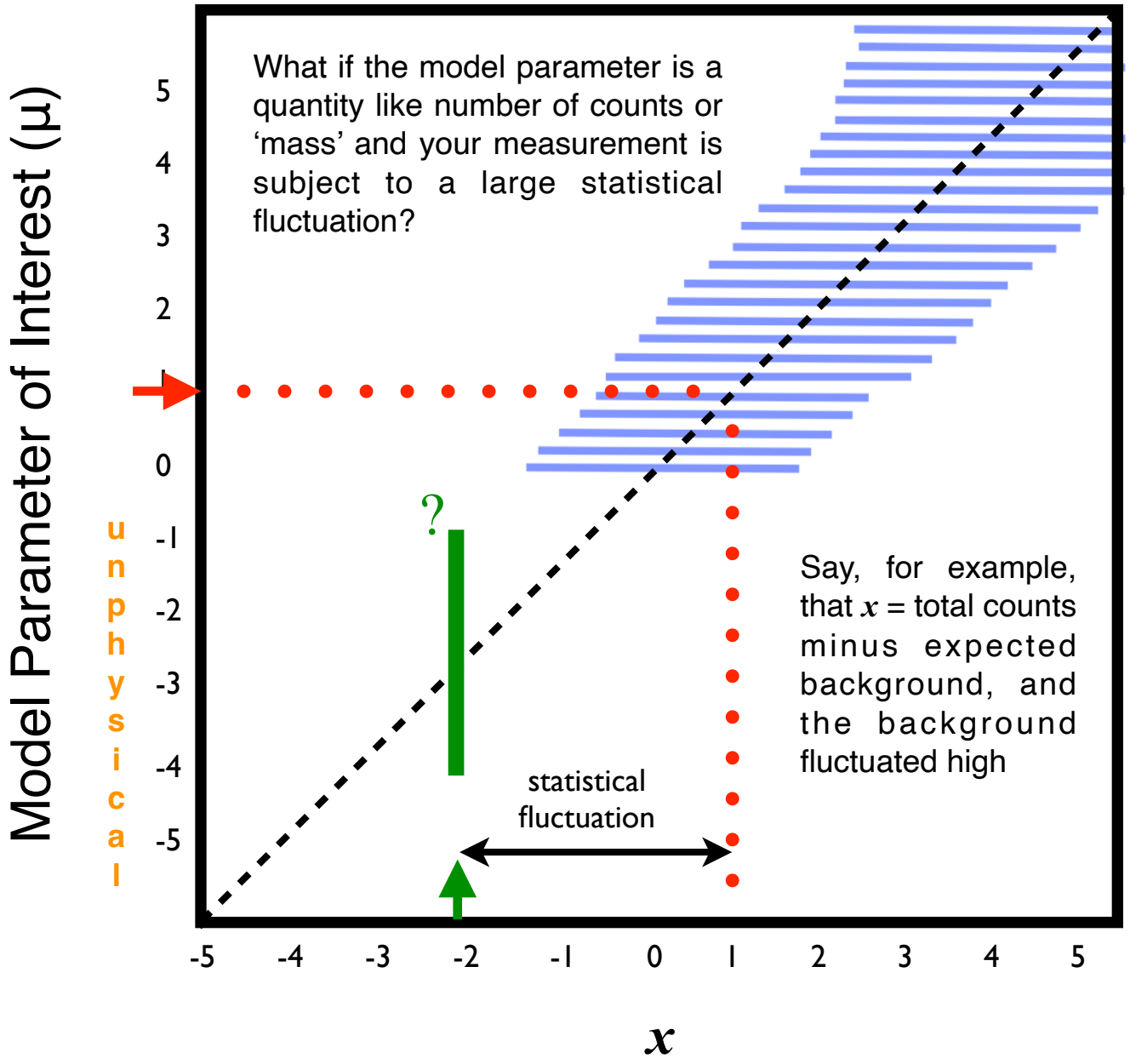
When using likelihoods for CL intervals, you can often appeal to Wilks' Theorem: for each true value of μ, the quantity x = -2log of the likelihood ratio between observed and expected quantities will be asymptotically distributed as a χ^2 distribution for nested hypotheses. Then, for a given observed measure of x, the integral χ^2 distribution for μ can be use to define the CL intervals.

Where this approximation breaks down, you can always resort to Monte Carlo methods to verify/derive the correct interval coverage.

Always a good thing to check: Do my derived contours seem to behave in the correct manner if I repeat the measurement with multiple MC data sets?

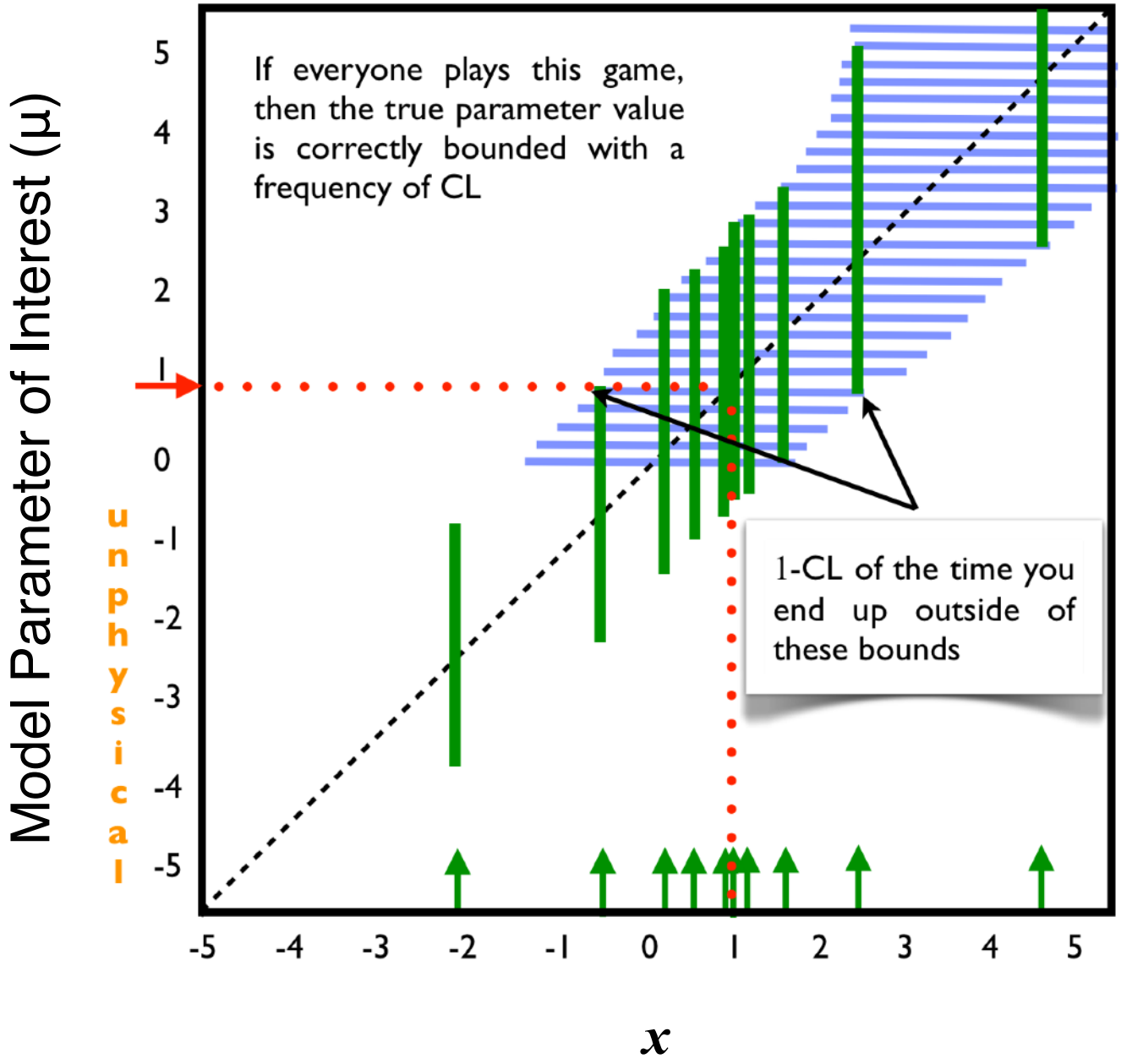
Note: It's a little weird that coverage here is no longer concerned with the frequency of physically observed quantities, but rather with the frequency of arbitrarily constructed mathematical quantities… but the construction is perfectly valid.

Neyman Construction of Frequentist Confidence Intervals



What's gone wrong?

Neyman Construction of Frequentist Confidence Intervals



Nothing!
Frequentists don't care about you, only about the ensemble of many experiments

Frequent Statement About Frequentist Intervals

"There is a 68% chance (for a ±1σ CL interval) that the model parameter lies in this range."

No! There is not a probability distribution associated with the model parameter, that's a Bayesian concept. Either it lies in your interval or not, but your one measurement does not constrain it.

"There is a 68% chance that my interval happens to bound the one, true value of the model parameter."

No! This is just an attempt to say the same thing with a wording that sounds more frequentist. Either it lies in your interval or it doesn't. However, there is a 68% probability that you would have been dealt a set of data that would have lead to an interval (not necessarily this particular one) containing the true parameter.

"If someone else were to repeat the experiment, there is a 68% chance that they would land in this range."

No! Your particular data set could have been a 3σ fluctuation, in which case there is very little chance that the next measurement would land in your interval.

Fre•quent•ist [free-kwuh nt-ist] *noun*

One who espouses the principles of the frequency definition of probability, and then misapplies them to answer the Bayesian question that they actually have in mind.

Qualifier:

This is a generalisation and just a personal opinion.

But check it out - it's really true!

Many physicists don't like the fact that statistical fluctuations can result in a bound extending into an "unphysical" region, or can result in a "null" interval if the unphysical region is rejected.

(but frequentist intervals do not bound physical models, so there really is nothing at all wrong with this!! The concern suggests that you might want to ask a different question form the one you are answering)

This is generally dealt with by either:

1) Truncating the allowed parameter space and renormalising the distributions to the "physical region." *(which corrupts the stated coverage)*
2) Defining the ordering parameter in a way that cannot wander into the "non-physical" region in the first place *(which distorts the interval definitions often in a non-intuitive way)*

Both are effectively trying to introduce a prior for the model parameter, which is not very frequentist!

In addition, Feldman and Cousins* were concerned about "flip-flopping: If experimenters choose for themselves when to quote a given type of interval based on the result, this can lead to a small statistical bias in frequentist coverage.

Worst case (at borderline of CL):
a 90% CL might only have 85% coverage;
a 99% CL might only have 98.5% coverage

A concern over tiny biases in unfiltered surveys of borderline results (!!)

So, F-C intervals use an ordering parameter of the likelihood ratio wrt to the maximum likelihood for parameters in the "physical" regime, and use a highest probability density ordering for this ratio to specify either a one or two-sided interval, based on the CL value. Monte Carlo methods are used to determine intervals with the correct coverage.

In contrast, "Standard Frequentist Intervals" will be defined as those using physical observables as the ordering parameter, without parameter space truncation, with distinct 1-sided and 2-side bounds.

Unified Approach to the Classical Statistical Analysis of Small Signals (Phys.Rev.D 57:3873-3889,1998)

Issues with F-C In Particular

• Conflicts with scientifically well-motivated convention to quote 90% or 95% CL upper/lower bounds for results consistent with the null hypothesis, but only claim a 2-sided discovery interval when the null hypothesis is rejected at a considerably higher confidence level;

• Can't easily cope with look-elsewhere effects: Search for gamma-ray emission from 1000 different astrophysical sources results in no event excess above 3σ, consistent with statistical fluctuations. Most appropriate to quote upper bounds on the possible emission from each source, but unified approach forces 3σ detection interval;

• Even for a clear detection, it may still be relevant to also quote upper and lower bounds in the context of different models. **Different interval constructions can be simultaneously valid and relevant for the same results, they simply address different questions!**

• Intervals do not represent the frequency of physical observables, are asymmetric and can be non-intuitive: observations of physical observables that occur with the same frequency can be included or excluded from the intervals differently;

• Because the construction is designed to always return a value "in the physical region," it fools people into thinking they are setting bounds on model parameters, which they are not! This has not dealt with the underlying issue and frequently leads to interpretation problems;

• Can be incredibly computationally expensive!

• All F-C concerns and methodologies are only relevant for borderline signals, otherwise you are just deriving "standard" parameter contours using likelihood… and **it's worth checking whether Wilks' Theorem is good enough** here (if you are dominated by Poisson statistics and Gaussian constraints, it probably is!).

Propagation of Systematic Uncertainties

There is no mathematically self-consistent way to propagate systematics in a frequentist paradigm!

Systematic uncertainties are exactly like model parameters: they have true fixed but unknown values. So, for a given assumed value of the model parameter and assumed values for the systematic uncertainties, you can define a frequentist confidence interval. That's it!

There are a number of suggested propagation approaches (such as Highland-Cousins) that involve Bayesian integrations over systematic uncertainties, but the interpretation of the resulting bounds are unclear…

The Problem With Zero:

Consider the case where zero events are observed in an experiment and we then wish to set a 90% CL/CI upper bound on the average signal strength.

Bayesian: We know that the number of background events here is exactly zero. The 90% CI upper bound on the average number of signal events is 2.3 (i.e. there is a 10% Poisson probability to fluctuate from this to 0)

Frequentist: It depends on the expected number of background events… even though the known number is zero! That's because frequentists don't care about you, it's all about the ensemble.

If you don't have a model for the background, you can't set a bound… even when you know the background.

The Problem With Zero:

Consider the case where zero events are observed in an experiment and we then wish to set a 90% CL/CI upper bound on the average signal strength.

If you try to ‘propagate’ uncertainties in the background estimate for frequentist bounds using a hybrid approach such as H-C, the derived constraints get better as the uncertainty grows!

This is because there is now some chance that the expected background could be higher*, which makes your observation in the ensemble less likely.

*could also be lower, but the impact on the probability is asymmetric and the higher fluctuations have greater effect

90% CL upper bounds on a possible average signal level from a simple counting experiment

	Initial Test: B=5, n=2	Improved Cuts: B=0.5, n=0
Standard Frequentist	0.32	1.8 **(worse)**
Feldman-Cousins	1.73	1.94 **(worse)**
Bayesian (prior uniform in rate)	3.13	2.3 **(better)**

Can appear to be overly strict bounds on the average signal strength

New analysis technique: suppresses backgrounds by a factor of 10 with no loss in signal efficiency!

F-C: “Should always also quote expected sensitivity”

Not appropriate

Consider the case where you look for a signal from 1000 different astronomical objects and see one with an excess of 3σ. This is not significant given the context of the search, so you just want to set an upper bound on the possible flux from this object.

Those constraints will be worse than the nominal expected sensitivity for this object because of the large excess, which is nonetheless still consistent with the null hypothesis because of the context

F-C automatically transitions from 1-sided to 2-sided bounds based on the p-value to avoid biases* due to “flip-flopping”

Not appropriate

Consider the case where you look for a signal from 1000 different astronomical objects and see one with an excess of 3σ. This is not significant given the context of the search, so you just want to set an upper bound on the possible flux from this object.

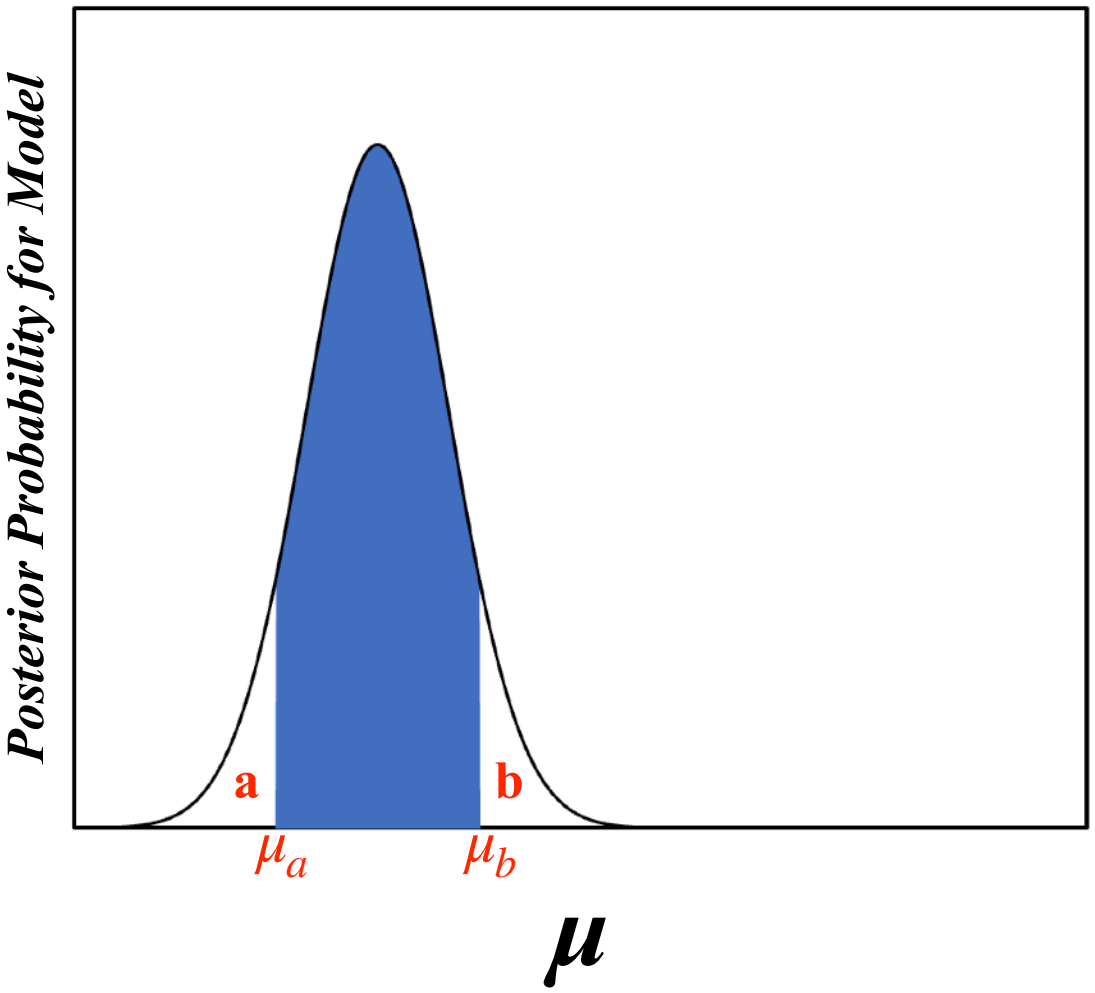
2-sided bounds **ONLY** have meaning once you have rejected the null hypothesis!

* A purely frequentist issue, with the biases being very minor and only relevant for potential signals at the border of being significant.

Bayesian Credibility Intervals

Bayesian Credibility Intervals

Given the measured data set:



Credibility Interval:

CI = 1- a - b

Note: This is **not** trying to represent the 'actual distribution' of the true model parameter (which wouldn't make much sense). This shows how much you'd bet that a given value is true based on the data you have.

Example: Find the Bayesian CI upper bound on the mean signal strength, **S**, for a counting experiment where the expected background level is **B** and a total of **n** events are observed.

$$\int_{-\infty}^{S_{up}} \frac{\overbrace{\frac{(S+B)^n e^{-(S+B)}}{n!}}^{\text{Likelihood}} \overbrace{H(S)}^{\text{Prior}}}{\underbrace{\int_{-\infty}^{+\infty} \frac{(S'+B)^n e^{-(S'+B)}}{n!} H(S')dS'}_{\text{Normalisation}}} dS = CI$$

Posterior probability from signal from Bayes' Theorem

We'll assume there is no *a priori* reason why all values of S shouldn't be considered equally likely, aside from the fact that it must be non-negative. So, take the prior to be zero for S<0 and constant otherwise.

Then just solve for S_{up}

Conveniently, this turns out to be mathematically identical to:

$$\frac{\sum_{m=0}^{n} \frac{(S_{up}+B)^m e^{-(S_{up}+B)}}{m!}}{\sum_{m=0}^{n} \frac{B^m e^{-B}}{m!}} = 1 - CI$$

renormalises allowed range of background counts (which must be less than or equal to n)

Otherwise, same expression as for the "Standard" frequentist approach!

The CLs Method

Introduced by physicists at LEP to get around some of the apparent problems that arise when mis-interpreting frequentist upper bounds in the presence of background fluctuations. The idea is to renormalise the standard frequentist bounds to the range of background values that are consistent with the current data set.

For example:

$$\sum_{m=0}^{n} \frac{(S_{up}+B)^m e^{-(S_{up}+B)}}{m!} \geq 1 - \mathrm{CL}$$

Which is identical to a Bayesian bound with a constant non-zero prior!

But this is interpreted in a frequentist way… though it now does not guarantee frequentist coverage, so it isn't really frequentist. It is defined without a prior, so it isn't formally Bayesian. However, if used to bound the space of models, it is equivalent to a Bayesian bound with a constant and non-zero prior (without admitting to it!).

Bayesian Propagation of Systematic Uncertainties

Just integrate over the posterior probability distribution for the systematic in question.

What's the way out??

There is no "correct" choice of prior!

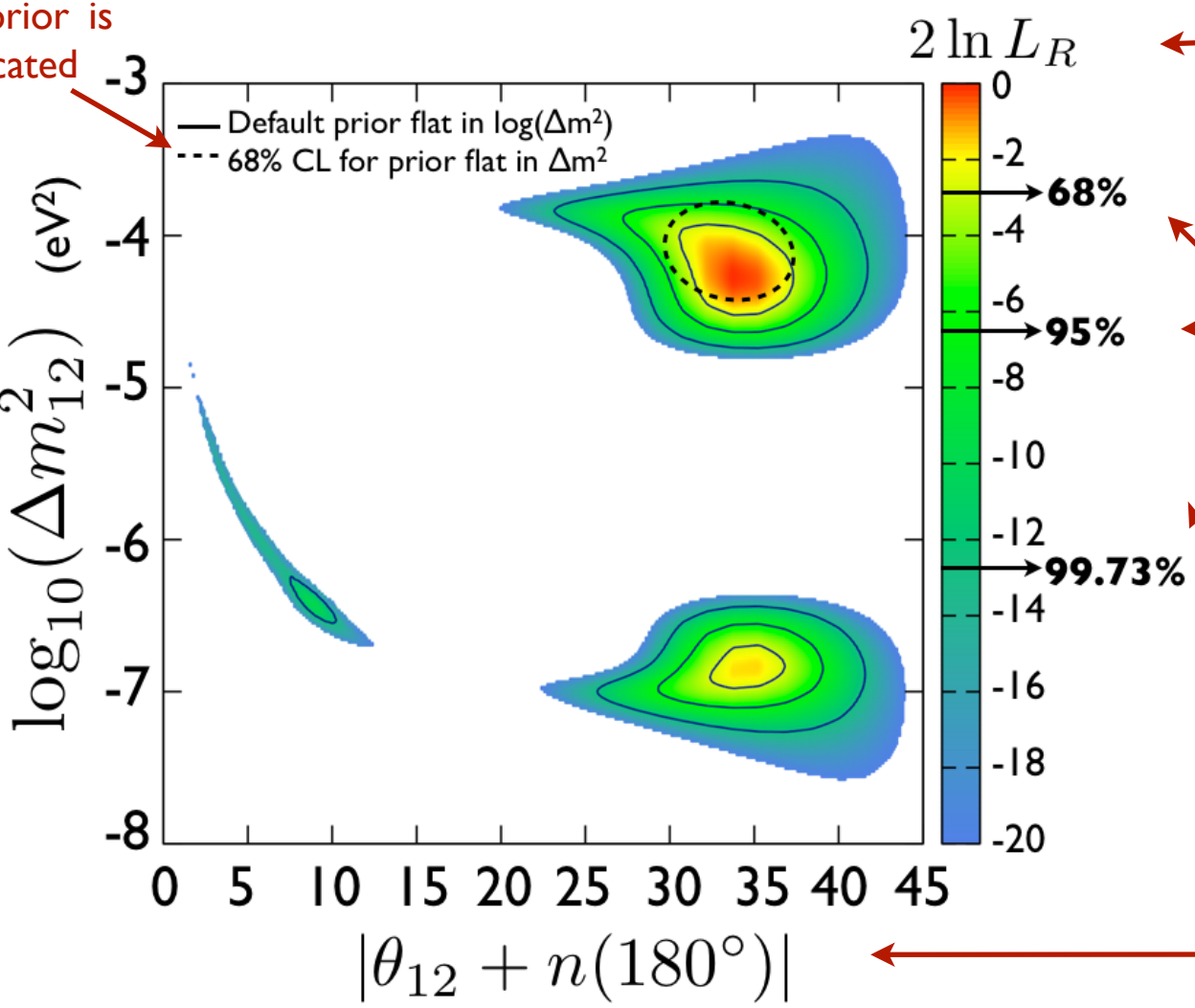
- Choose simple prior forms that are easy to understand and visualise (*e.g.* uniform) and try to use common parameter choices that will "make sense" for these priors.
- If using a more sensitive instrument to look for evidence of a signal that **has not been seen before**, this rules out priors with a probability that rises with the signal rate (because the higher the rate, the more likely it would have been seen before). So using a prior that is uniform in rate is conservative for setting an upper bound.
- Model parameter uncertainties generally tend to be either be about precision (i.e. *I know the parameter is roughly in this range*) or scale (i.e. *I don't really know what order of magnitude this is*). So forms of priors that are uniform on either linear or logarithmic scales often provide reasonable bounds.
- If there's an ambiguity that leads to a non-conservative bound, show the sensitivity to the choice of prior!

Note: Displaying the likelihood as a function of variables for which the priors are uniform, automatically also then plots the Bayesian posterior probability.

Example of "Unified" Likelihood Map: SNO salt phase solar ν data

(using publicly available data associated with Phys. Rev. Lett. **101**, 111301, 2008)

Sensitivity to prior is indicated

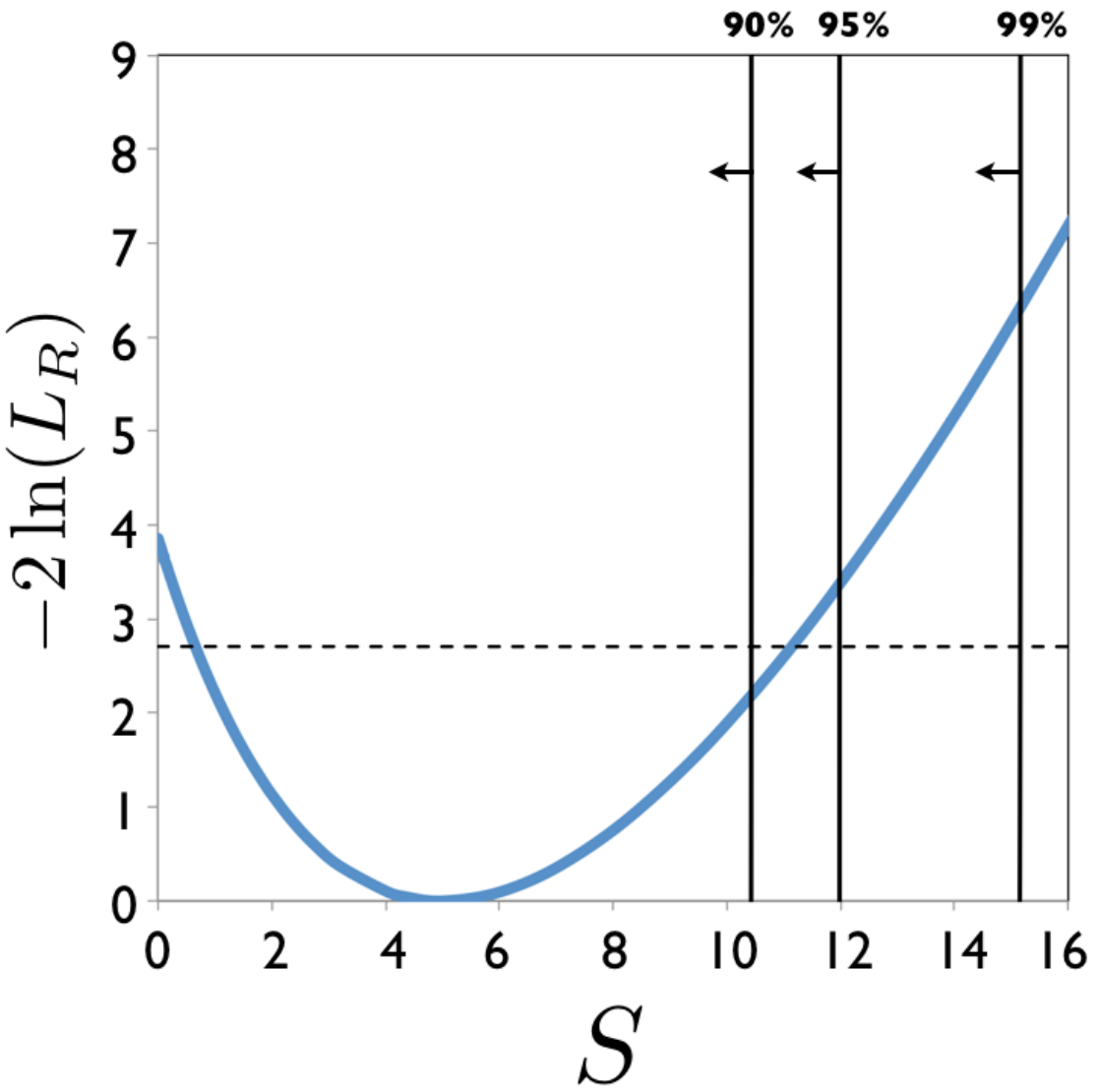


Approximate $\Delta\chi^2$ value from Wilks

Bayesian contours from integration of likelihood assuming priors uniform in θ and log(Δm²)

Form for fundamental angle accounts for quadrant ambiguity

Example 2: Rare Event Search Counting Experiment (B=5, n=10)

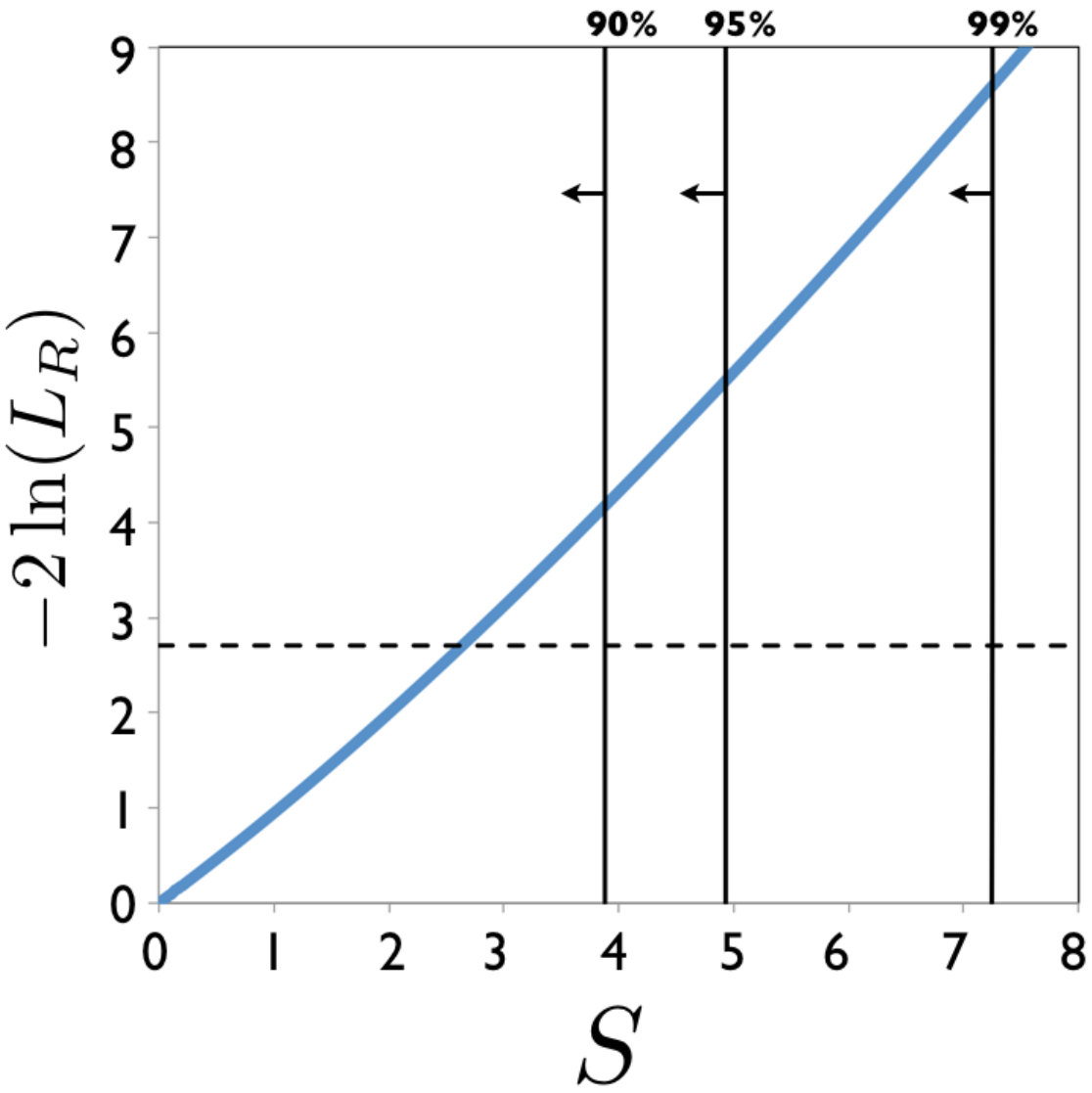


Bayesian upper bounds from integration of likelihood assuming priors uniform in S ('conservative')

Approximate 2-sided 90% CL frequentist bound derived from Wilks' theorem (comparable to FC)

S < 10.4(15.2) at 90%(99%) CI

Example 3: Rare Event Search Counting Experiment (B=9, n=5)



Bayesian upper bounds from integration of likelihood assuming priors uniform in S ('conservative')

Approximate 2-sided 90% CL frequentist bound derived from Wilks' theorem (comparable to FC)

S < 3.88(7.25)
at 90%(99%) CI

"Should I then use the outcome of previous experiments as part of the prior?"

Careful!!

Yes for other experiments that you have performed (e.g. calibrations) to assess certain aspects of detector performance, or related data that can be regarded as **unimpeachable**. Otherwise, generally not because the ability to properly assess systematic uncertainties associated with individual experiments is not generally under your control and can be difficult. This is why each experiment should stand on its own and be independently cross-checked by other experiments.

Summary

- Bayesian statistics is the **only** correct formalism that can address the question, "Given my measurement, what models do I constrain?" *My experience is that this form of the question has been implicit in all discussions of the physical interpretation of experimental data I've seen.*

- The standard frequentist approach is a perfectly valid and self-consistent formalism. However, it answers a different question, where the identification of a model only emerges for a "sufficiently large" ensemble of experiments. *Unfortunately, this is often misinterpreted (or correctly interpreted but then misused).*

- The Feldman-Cousins approach is a mathematically valid, if somewhat arbitrary, formulation of the frequentist method (*though it may be even more prone to misinterpretation by not making the nature of these intervals appear quite so obvious*).

Fortunately, for many cases (especially in the large n limit), these different approaches all give very similar results. However, this is not always the case, so be clear about exactly what your question you are asking!