

Lecture 6:

Priors & Confidence Intervals

- Mandatory Nature of Priors
- Bernstein - von Mises Theorem
- Self-Iteration and “Unfolding”
- Confidence Intervals - Wilks' and Neyman
- Meaning and Misinterpretation
- Issues with Confidence Intervals

Blue
and
Black

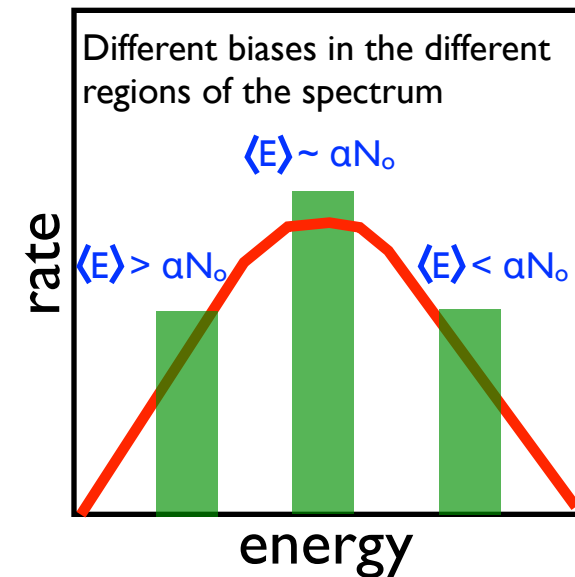
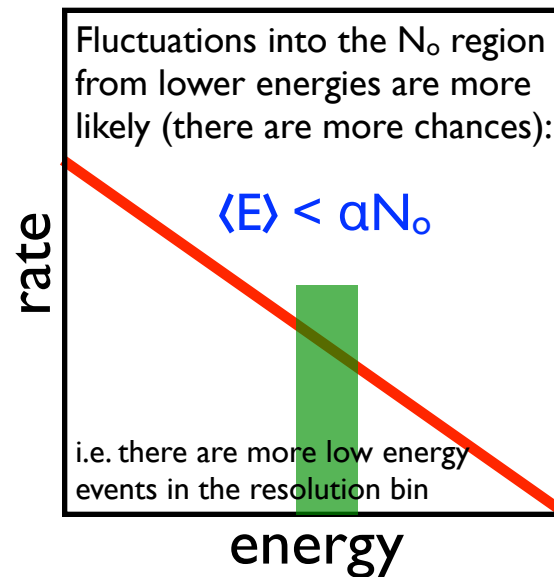
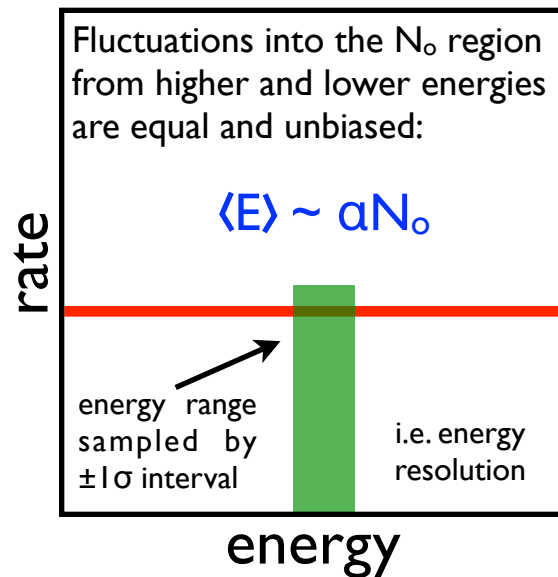


White
and
Gold

Another example:

Charged particles produce light as they pass through plastic scintillators, which can be detected by photomultiplier tubes and used as an estimator for the energy deposition. Say that that you calibrate such an instrument using known gamma line energies from various radioactive sources and determine that the energy can be very well described by taking the mean number (N) of detected photons (drawn from a Gaussian distribution of width σ) and multiplying it a proportionality constant, α .

Now you measure emission from some continuous spectrum and detect N_0 photons from an interaction. What is the best estimate of the gamma ray energy?



Relating data to model parameters **requires** a context (i.e a prior)!

Any inference about models based on an observation is an inherently Bayesian undertaking as it requires an assessment of the posterior probability $P(H_i|D)$ and, thus, **requires** the choice of a prior!

rarely

This is ~~often~~ not appreciated! The assumption that the relative likelihoods for two hypotheses alone is the same as the betting odds for which hypothesis is correct tacitly assumes an odds ratio of 1.

People often view priors as a problematic aspect of Bayesian statistics; a nuisance that they have to find a way around.

But this is **WRONG!** An ambiguity in the form of a prior represents a **REAL** ambiguity in the interpretation of data! *The choice of prior should only matter when the data itself isn't strong enough to provide an unambiguous interpretation.*

If there is an ambiguity in the choice of prior that can lead to notably different conclusions, you should show this!

Science & Environment

Cosmic inflation: 'Spectacular' discovery hailed

By Jonathan Amos
Science correspondent, BBC News

17 March 2014 [Science & Environment](#)

Scientists say they have extraordinary new evidence to support a Big Bang Theory for the origin of the Universe.

Researchers believe they have found the signal left in the sky by the super-rapid expansion of space that must have occurred just fractions of a second after everything came into being.

It takes the form of a distinctive twist in the oldest light detectable with telescopes.

The work will be scrutinised carefully, but already there is talk of a Nobel.

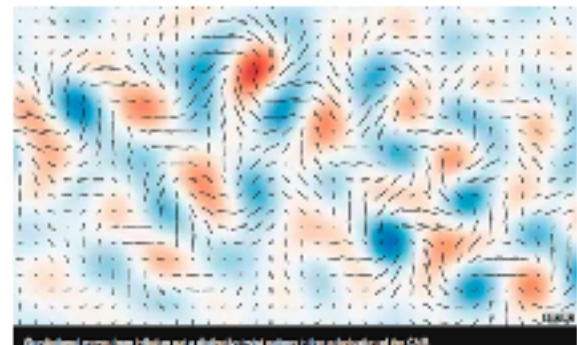
"This is spectacular," commented Prof Marc Kamionkowski, from Johns Hopkins University.

"I've seen the research; the arguments are persuasive, and the scientists involved are among the most careful and conservative people I know," he told BBC News.

The breakthrough was announced by an American team working on a project known as **BICEP2**.

This has been using a telescope at the South Pole to make detailed observations of a small patch of sky.

The aim has been to try to find a residual marker for "inflation" - the idea that the cosmos experienced an exponential growth spurt in its first trillionth, of a trillionth of a trillionth of a second.



Downloaded from the Internet and put a distribution pattern in the distribution of the CMB

Science & Environment

Cosmic inflation: BICEP 'underestimated' dust problem

By Jonathan Amos
Science correspondent, BBC News

22 September 2014 [Science & Environment](#)



BICEP2's South Pole telescope targeted areas the team hoped was a relatively clean part of the sky

One of the biggest scientific claims of the year has received another set-back.

Example:

As the result of a **random** blood test, you are diagnosed with “Saturday Night Fever,” a disease suffered by 0.5% of the population that results in convulsions when exposed to anything associated with John Travolta. The blood test reliably diagnoses the disease in 80% of cases and yields a false positive 5% of the time. Should you avoid listening to BeeGees albums?

$$\begin{aligned} P(SNF | B) &= \frac{P(B | SNF)P(SNF)}{P(B | SNF)P(SNF) + P(B | no\ SNF)P(no\ SNF)} \\ &= \frac{(0.8)(0.005)}{(0.8)(0.005) + (0.05)(0.995)} = 0.074 \end{aligned}$$

What if the reason you went to your GP for a blood test was that you got splitting headaches whenever someone mentioned the word “Grease?”

These are basically the same numbers as for COVID-19 (early Oct 2020).

What if you feel ill and get a positive test?

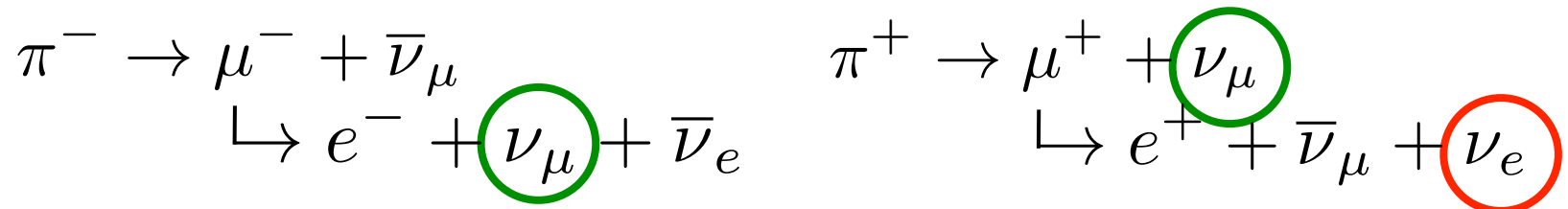
Say the average person is typically ill 10 days per year, so the odds of currently being ill from the common cold is $\sim 10/365 = 0.027$. With social distancing, reduce this by a factor of ~ 10 to 0.0027. So, the fraction of people feeling ill that have COVID-19 is perhaps something like $0.005/(0.005+0.0027) = 0.65$ (this, then, is the prior instead of 0.005).

$$\begin{aligned} P(CV19 | +T) &= \frac{P(+T | CV19)P(CV19)}{P(+T | CV19)P(CV19) + P(+T | no\ CV19)P(no\ CV19)} \\ &= \frac{(0.8)(0.65)}{(0.8)(0.65) + (0.05)(0.35)} = 0.97 \end{aligned}$$

Priors are important!

Example 2:

Atmospheric neutrinos result from the decay of charged pions produced by hadronic interactions in the atmosphere. The characteristic decay sequences are:



You are detecting these neutrinos coming from directly overhead with an underground water Cherenkov detector. From the fuzziness of the ring pattern of observed light from a particular event, simulations tell you that 70% of ν_e 's will produce a ring at least this fuzzy, whereas only 50% of ν_μ 's will do this. What is the probability that this event is a ν_e ?

$$\begin{aligned} P(\nu_e|R) &= \frac{P(R|\nu_e)P(\nu_e)}{P(R|\nu_e)P(\nu_e) + P(R|\nu_\mu)P(\nu_\mu)} \\ &= \frac{(0.7)(1/3)}{(0.7)(1/3) + (0.5)(2/3)} = 0.41 \end{aligned}$$

Bernstein – von Mises Theorem

In the limit of an infinitely large data set, the posterior probability is independent of the exact form of the prior probability.

(the likelihood function that multiplies the prior crushes its impact away from the region of interest)

For example, if you instead asked for the probability for a large number Cherenkov events to be v_e out of a big data set, the information contained in the distribution of ring fuzziness within the data itself carries more weight than the form of any previously assumed prior.

Priors carry greater weight for weaker data sets

“Should I then use the outcome (i.e. posterior probabilities) from previous experiments to form the prior for this one?”



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.

Self-Iteration and “Unfolding”

You might wonder what happens if we iteratively update the priors using the posterior probabilities that emerge from the same data set. Does this converge to something meaningful in a way that doesn't depend so much on the initial choice of priors?

Let's take the simple case of a single bin in a histogram, where a number of counts, n , is observed, a background, b , is expected, and we wish to determine the best estimate for the number of signal counts, s .

Say we want to do this in a Bayesian way, so we'll start with some prior, $P(s_i)$, as a function of signal value, and then iterate...

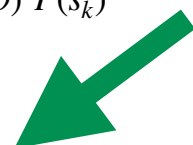
1st estimate:

$$P^{(1)}(s_i | n) = \frac{P(n | s_i + b) P(s_i)}{\sum_j P(n | s_j + b) P(s_j)}$$

2nd estimate:

$$P^{(2)}(s_i | n) = \frac{P(n | s_i + b) \frac{P(n | s_i + b) P(s_i)}{\sum_k P(n | s_k + b) P(s_k)}}{\sum_j P(n | s_j + b) \frac{P(n | s_j + b) P(s_j)}{\sum_k P(n | s_k + b) P(s_k)}} = \frac{P^2(n | s_i + b) P(s_i)}{\sum_j P^2(n | s_j + b) P(s_j)}$$

Nth estimate:

$$P^{(n)}(s_i | n) = \frac{P^N(n | s_i + b) P(s_i)}{\sum_j P^N(n | s_j + b) P(s_j)}$$


We can see what's happening: this process simply accentuates features (including fluctuations) that are already in the likelihood. As $N \rightarrow \infty$, the posterior converges to 1 for the maximum likelihood value and zero elsewhere. In the case of a degeneracy, the convergence value is determined by the original choice of prior. Features outside the maximum likelihood values become **artificially suppressed**, but **no additional information has been gained... because there is none!**

One popular approach* used by some in particle physics to try to deconvolve or “unfold” underlying model distributions with minimum reliance on assumption from priors involves self-iteration of priors such as this... which then suffers from exactly these issues. **It gets you nowhere.**

At it's heart, “unfolding” is, fundamentally, a Bayesian undertaking. A number of approaches have been suggested for different cases but, ultimately, it necessarily comes down to the use of the likelihood function guided, in some way, by prior probabilities to help break degeneracies, insure continuity, and generally constrain the solution to a physically meaningful and realistic form.
So deal with this explicitly!

DID THE SUN JUST EXPLODE? (IT'S NIGHT, SO WE'RE NOT SURE.)

THIS NEUTRINO DETECTOR MEASURES
WHETHER THE SUN HAS GONE NOVA.

THEN, IT ROLLS TWO DICE. IF THEY
BOTH COME UP SIX, IT LIES TO US.
OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY.

DETECTOR! HAS THE
SUN GONE NOVA?



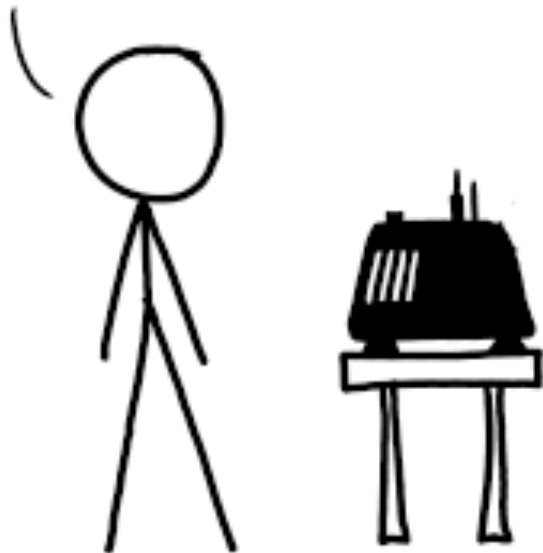
FREQUENTIST STATISTICIAN:

BAYESIAN STATISTICIAN:

FREQUENTIST STATISTICIAN:

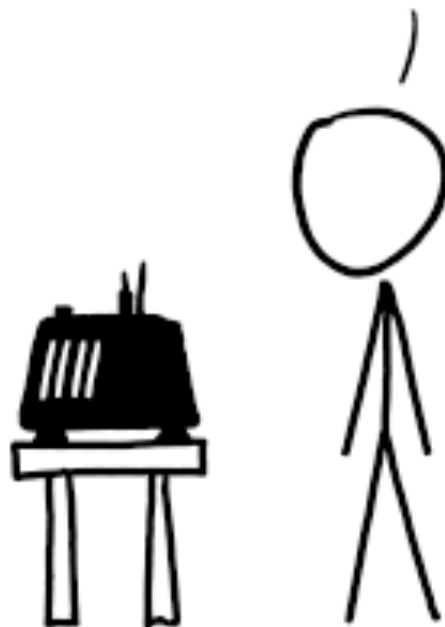
THE PROBABILITY OF THIS RESULT
HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.

SINCE $p < 0.05$, I CONCLUDE
THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU \$50
IT HASN'T.



Confidence and Credibility I: Frequentist Confidence Intervals (an attempt to avoid priors)



Construction of Frequentist Confidence Intervals via Wilks' Theorem

We've been here before...

$$-2[\ln L(\mathbf{q}_0) - \ln L(\mathbf{q})] = -2 \ln \left(\frac{L(\mathbf{q}_0)}{L(\mathbf{q})} \right) \equiv -2 \ln L_R \sim \chi_d^2$$

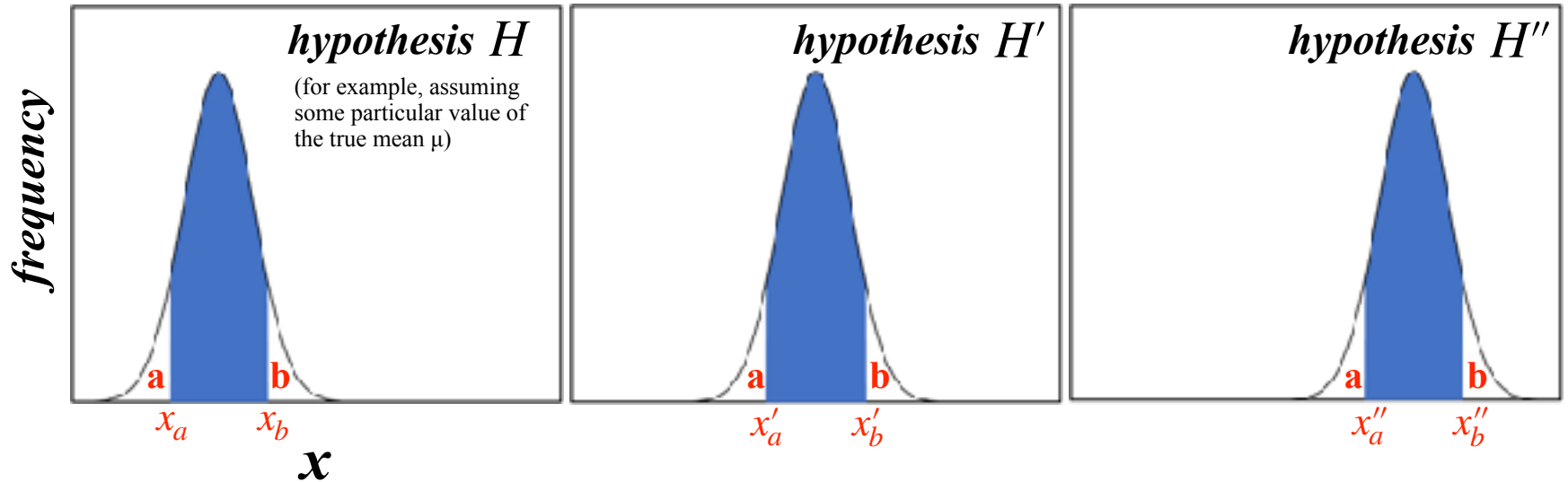
where \mathbf{q}_0 are the set of model parameters that define the default (null) hypothesis, and the $d = \text{DoF}$ = the difference in the number of model parameters constrained (i.e. how many extra degrees of freedom one model has compared to the other)

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 it 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



$$CL = 1 - a - b$$

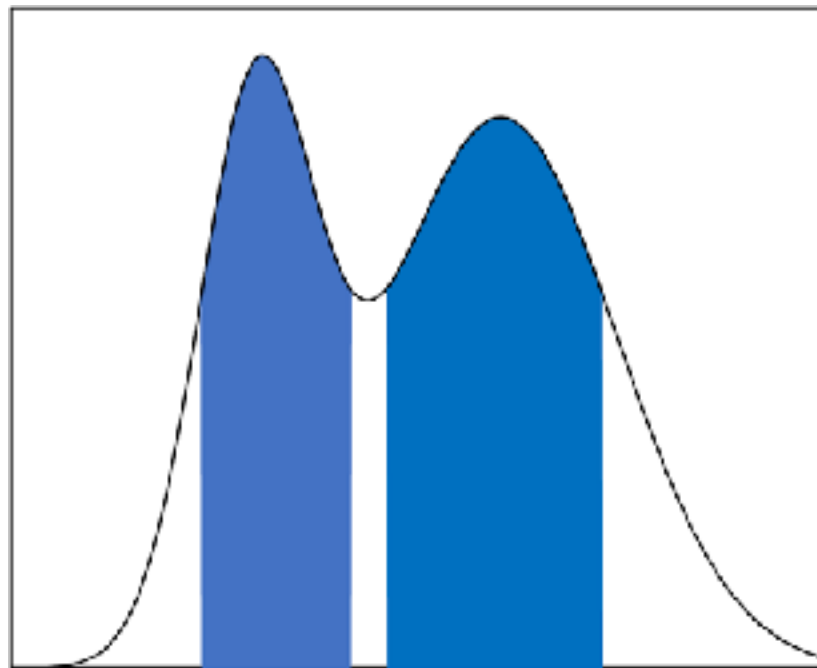
(where “Confidence Level” refers to the frequency of hypothetical measurements landing in the defined region for a given model)

... etc.

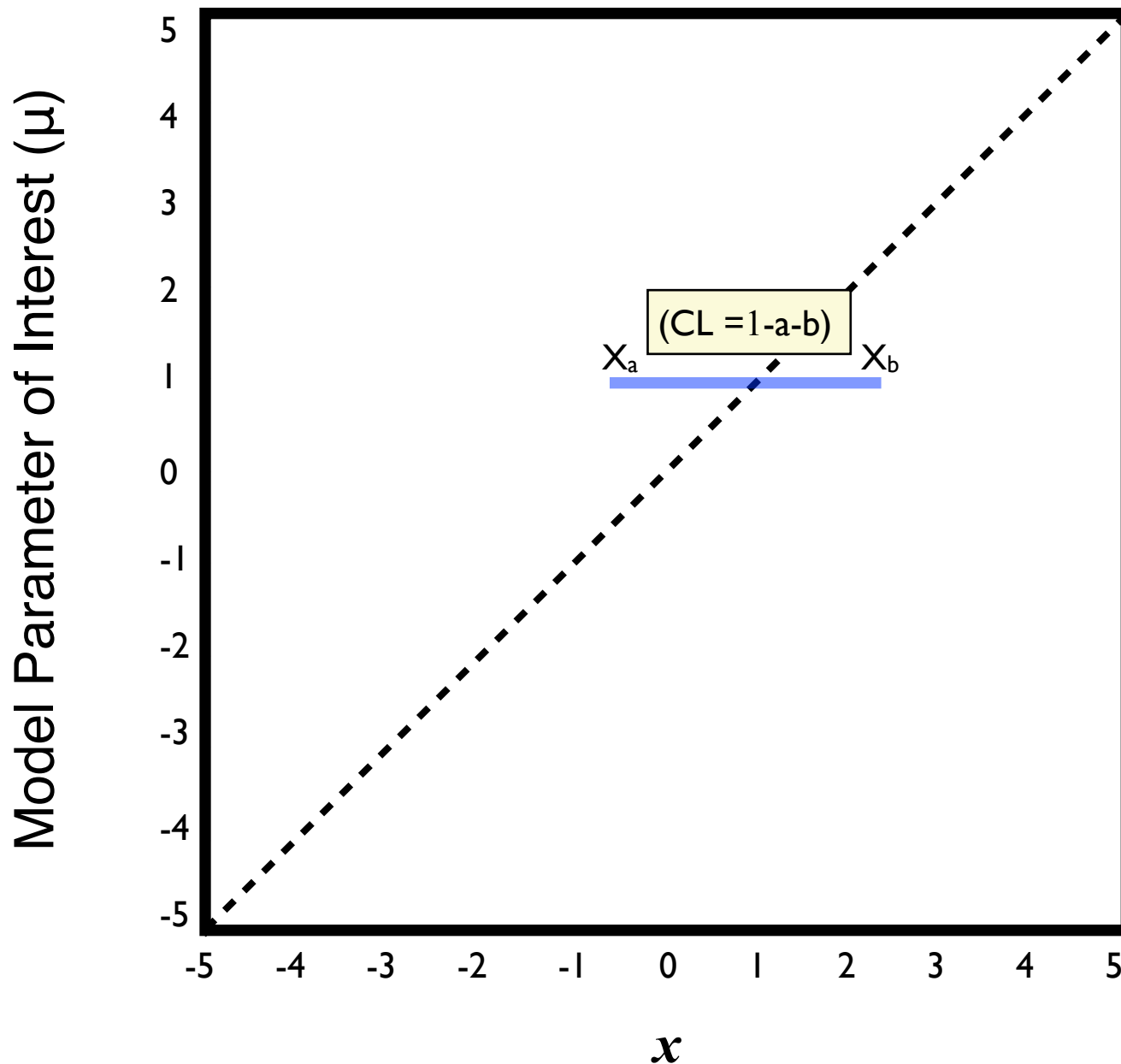
x is an “ordering parameter,” which can be a direct measurable (such as the number of counts) or can be a derived quantity (such as a likelihood ratio)

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

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

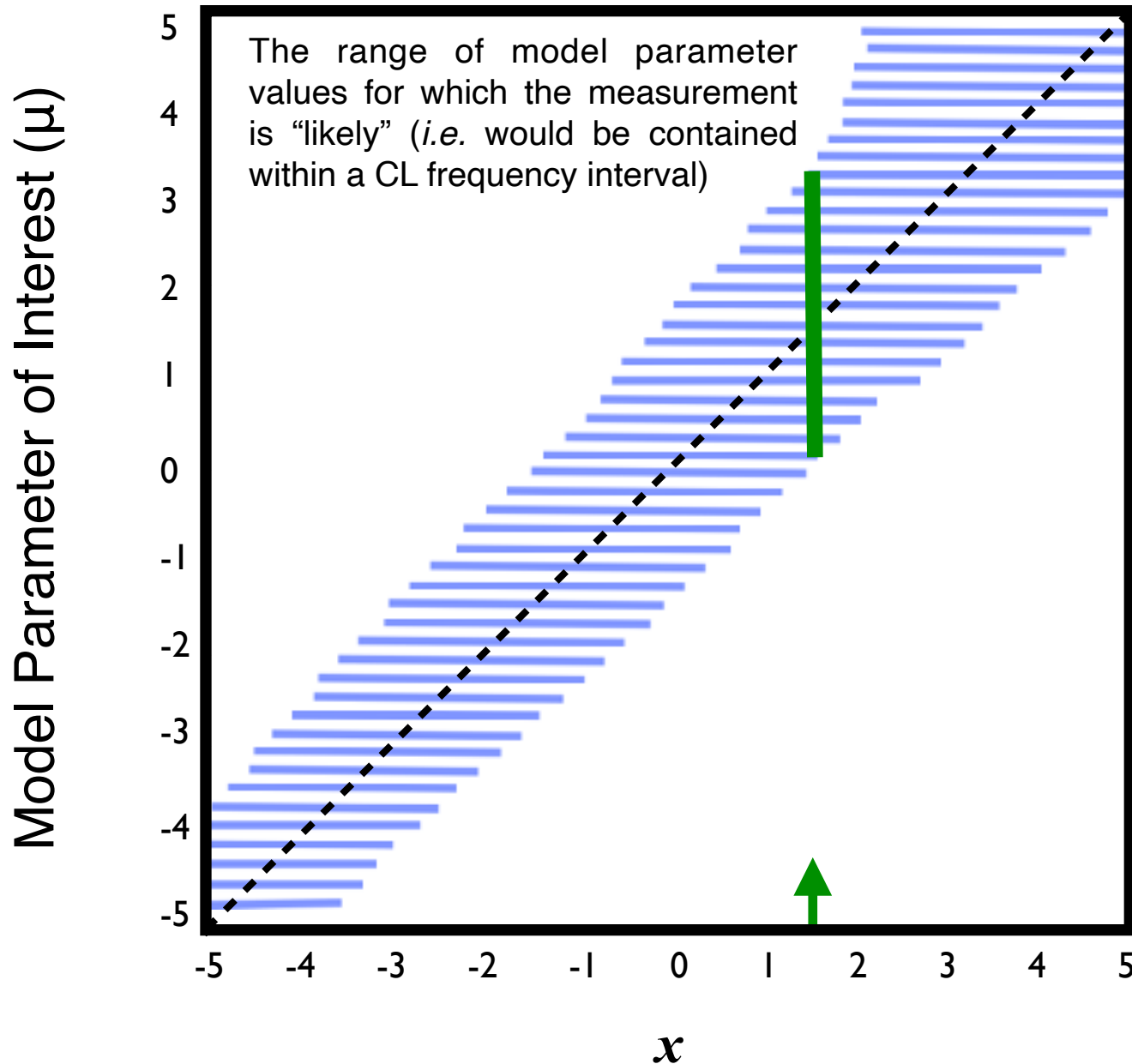


Neyman Construction of Frequentist Confidence Intervals



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

Let's consider the specific case of
Poisson statistics as an example...

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!**

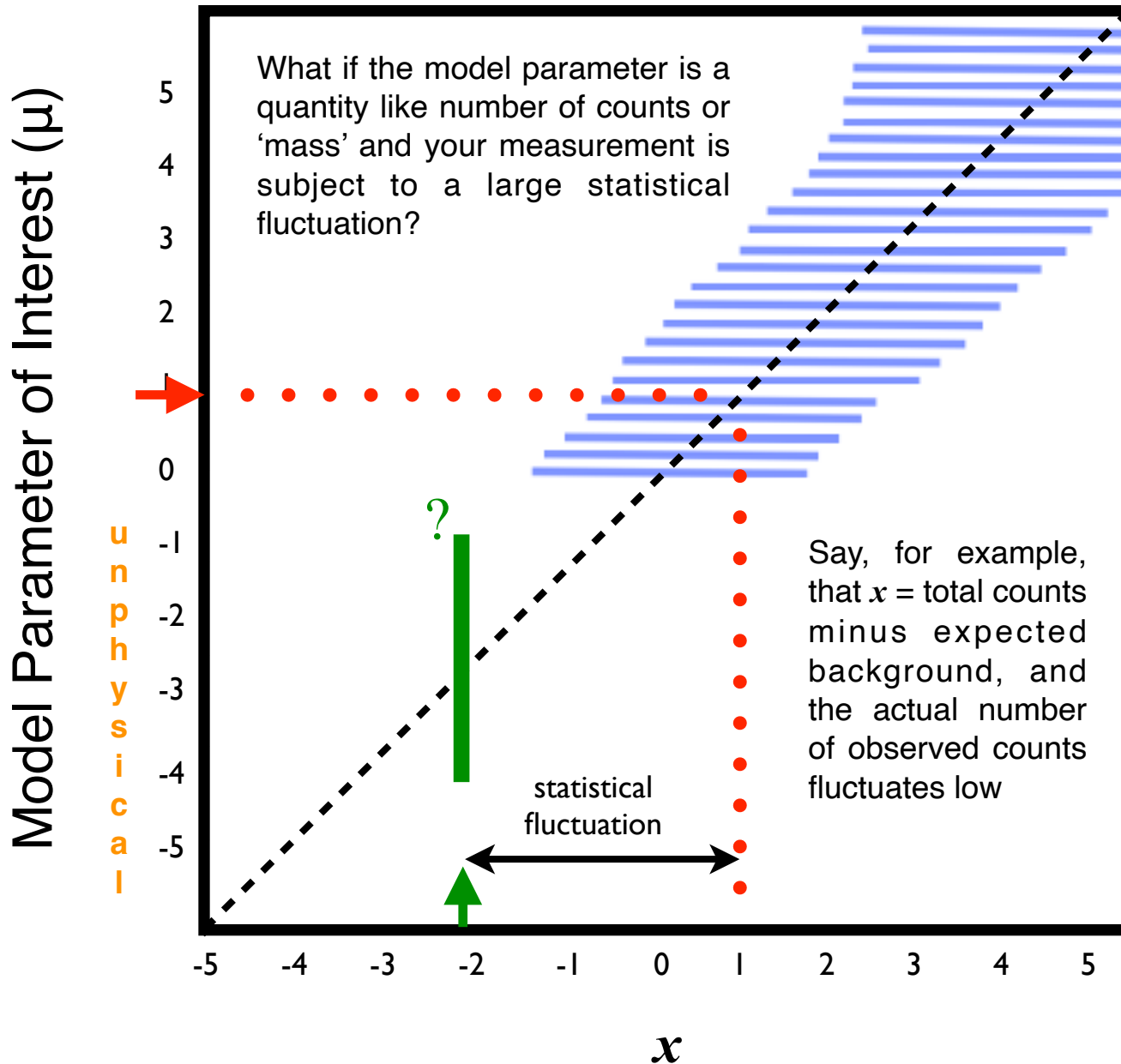
When using likelihoods for CL intervals, you can often appeal to Wilks' Theorem: for each true value of μ , the quantity $x = -2\log$ of the likelihood ratio between observed and expected quantities will be asymptotically distributed as a χ^2 distribution for nested hypotheses. Then, for a given observed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

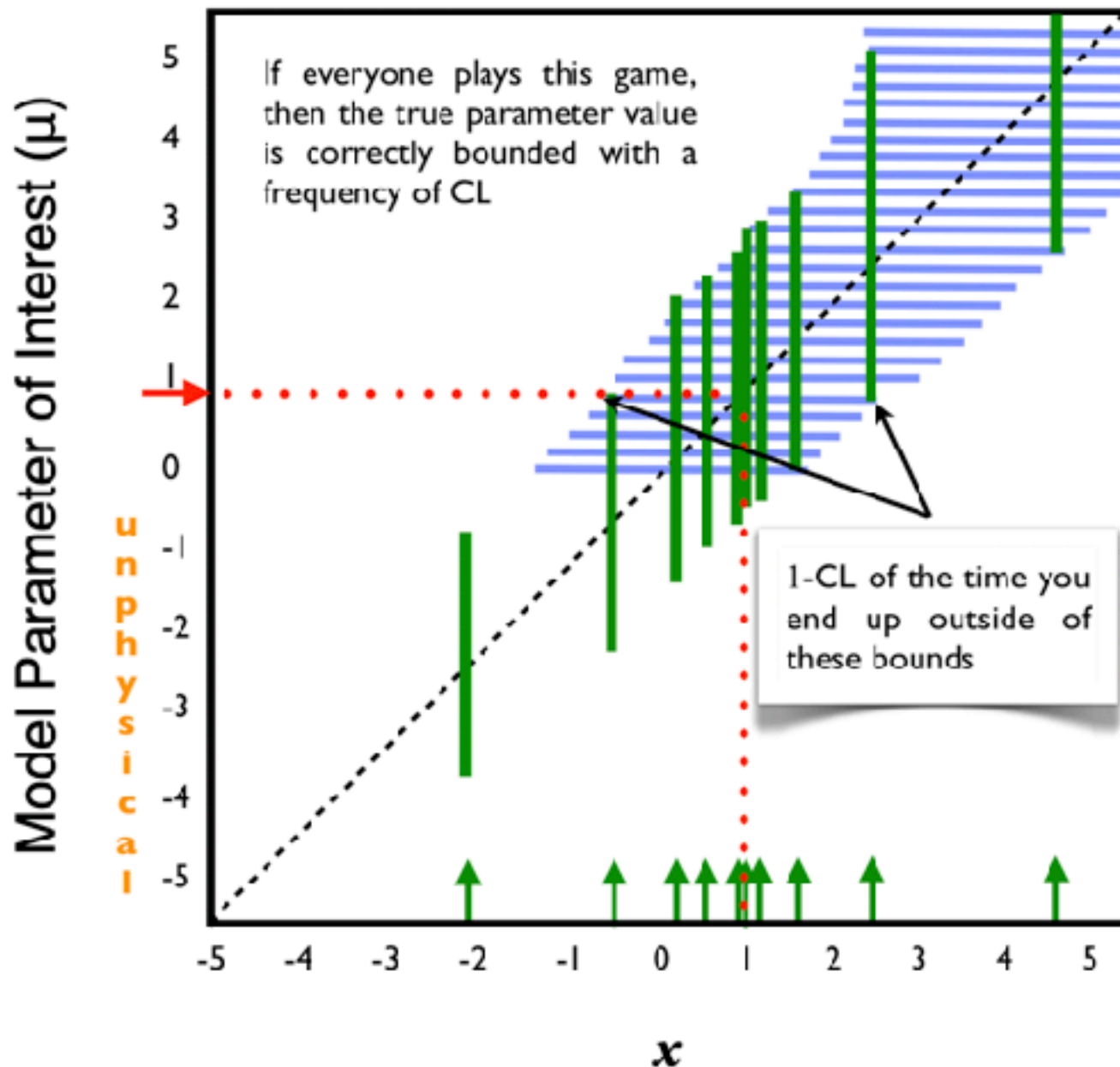
Note: It's a little weird that coverage here is no longer concerned with the frequency of physically observed quantities, but rather with the frequency of arbitrarily constructed mathematical quantities... but the construction is perfectly valid.

Neyman Construction of Frequentist Confidence Intervals



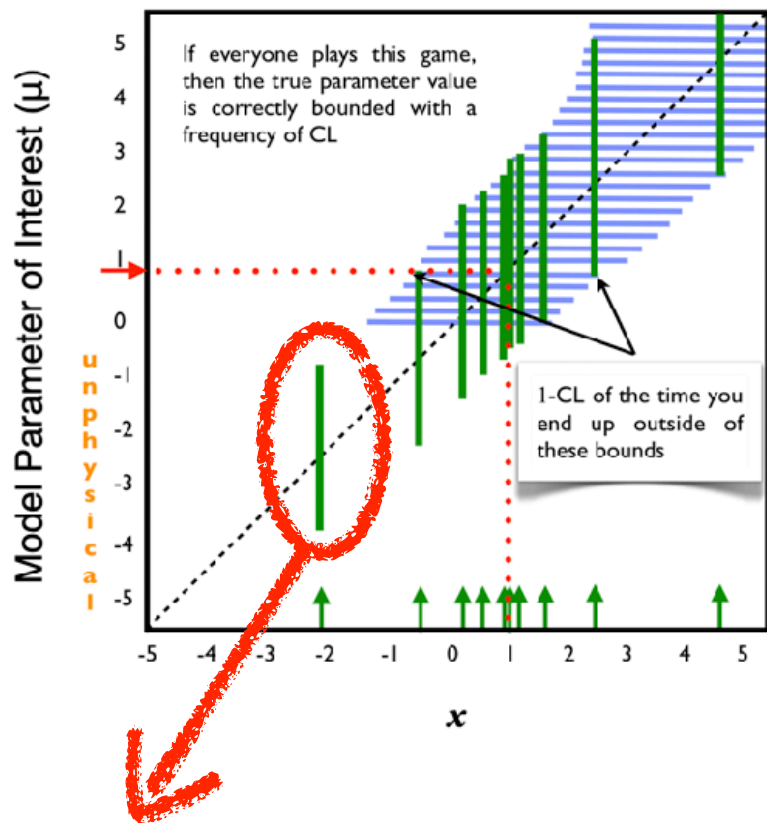
What's
gone
wrong?

Neyman Construction of Frequentist Confidence Intervals



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

Say we are setting a 90% CL:



The correct model is bounded 90% of the time in the ensemble



So, if you were to randomly choose a data set and its associated CL interval, the chance that you would pick one that bounds the correct model is 90%



~~It then sounds like you should be able to say that the probability that my CL interval (taken as a random sample) bounds the correct model is 90%...~~

But what if this is your interval?

What is the probability that your interval bounds the correct model?

0%... and you KNOW it is 0% !!

All you've got is the ticket to say you've played the game

Weather forecasts are accurate in the sense that, when they say there is a 10% chance of rain, it rains 10% of the time.

The weather forecast says there is a 10% chance of rain.
Without looking outside, what is the probability it is raining?
(Btw, you hear thunder and a gentle tapping on the roof)

Do you take your raincoat?

When you use the term ‘probability,’ perhaps you don’t actually care about whether it’s raining somewhere else or how often it rained on similar days. You may instead mean “Given THIS location and THIS day, what is the degree of my belief that it is raining”

Frequent Statement About Frequentist Intervals

“There is a 68% chance (for a $\pm 1\sigma$ 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.

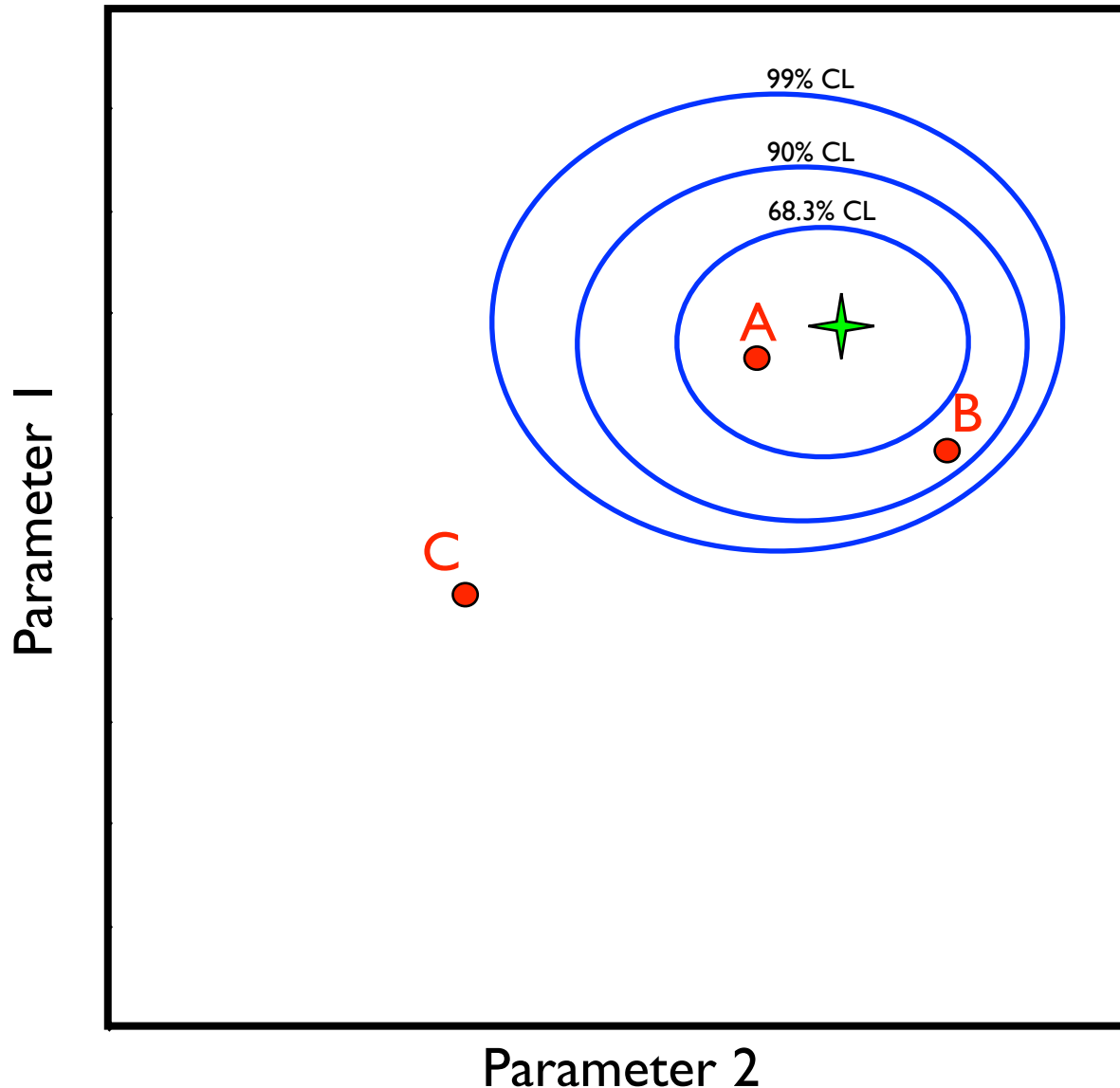
“There is a 68% chance that my interval happens to bound the one, true value of the model parameter.”

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.

Consider a single experiment in which 2 parameters are measured (\star) 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

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 from 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” region, 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 the frequency of physical observables as the ordering parameter, without parameter space truncation and 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 (being neither fully Bayesian nor guaranteeing statistical coverage)

The Problem With Zero:

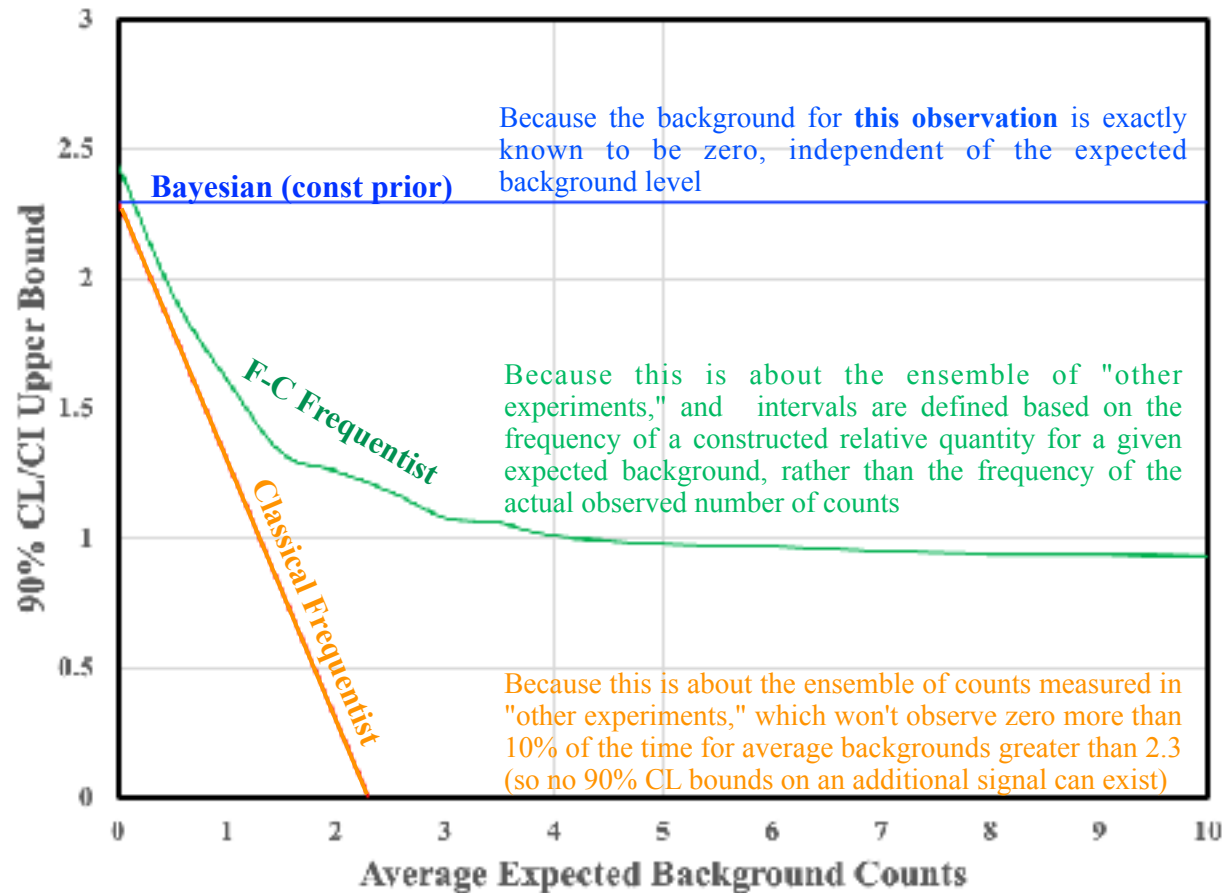
Consider the case where zero events are observed in an experiment and we then wish to set a 90% CL/CI upper bound on the average signal strength.

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.

90% CL/CI Upper Bounds on a Signal when Zero Counts are Observed



The specific frequentist intervals derived in this region are increasingly less likely to bound the true value of average signal flux because they are increasingly less representative of what would be seen by most of the ensemble (which will yield less restrictive bounds that carry much greater weight for determining the correct coverage)

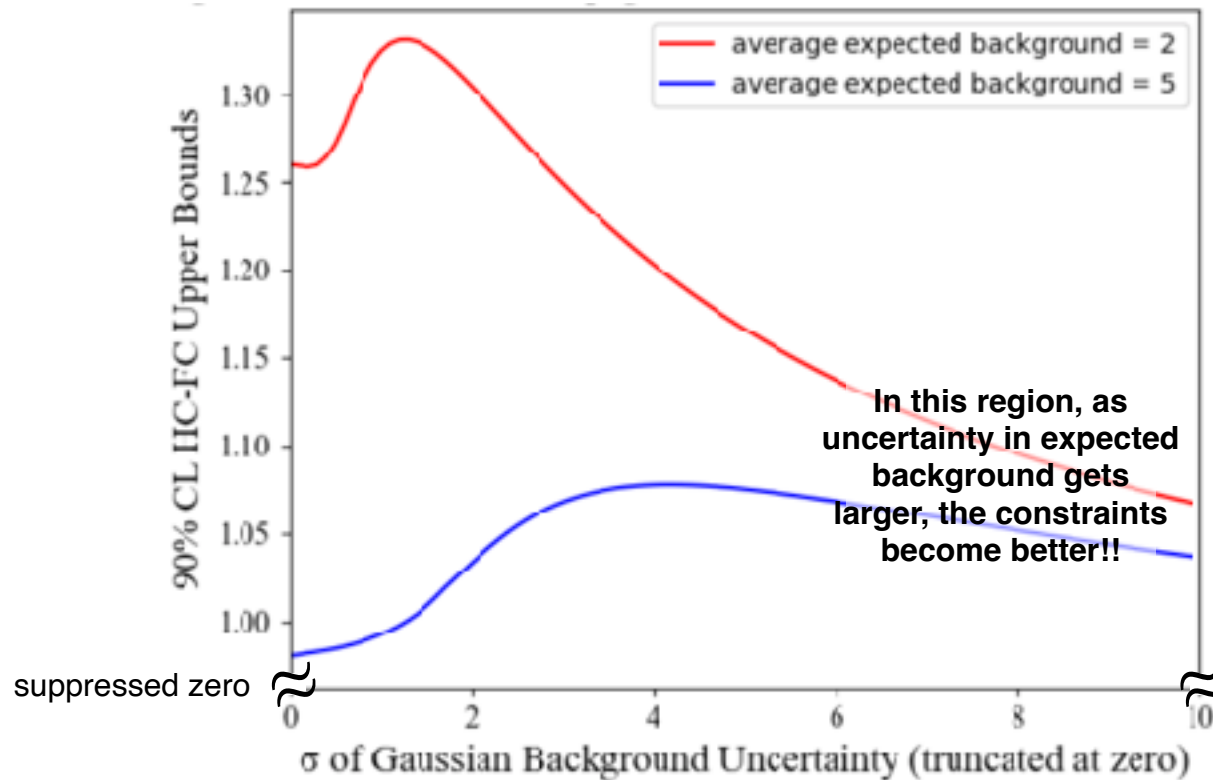
The Problem With Zero:

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 can behave peculiarly, and sometimes even get better as the uncertainty grows!

This is a consequence of integrating over possible expected backgrounds, which have a downward trend for the obtained limits as the expected background increases

Highland-Cousins Error Propagation for FC Intervals when Zero is Observed



(Bayesian bounds don't care about the uncertainty in the expected background... because there is no uncertainty for this measurement: it is exactly zero!)