Lecture 7:

Confidence vs Credibility

- More Confidence Issues
- Bayesian Credibility Intervals
- CLs Method
- Integration vs Maximisation
- Dealing with Priors
- Display of Frequentist & Bayesian Information

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!

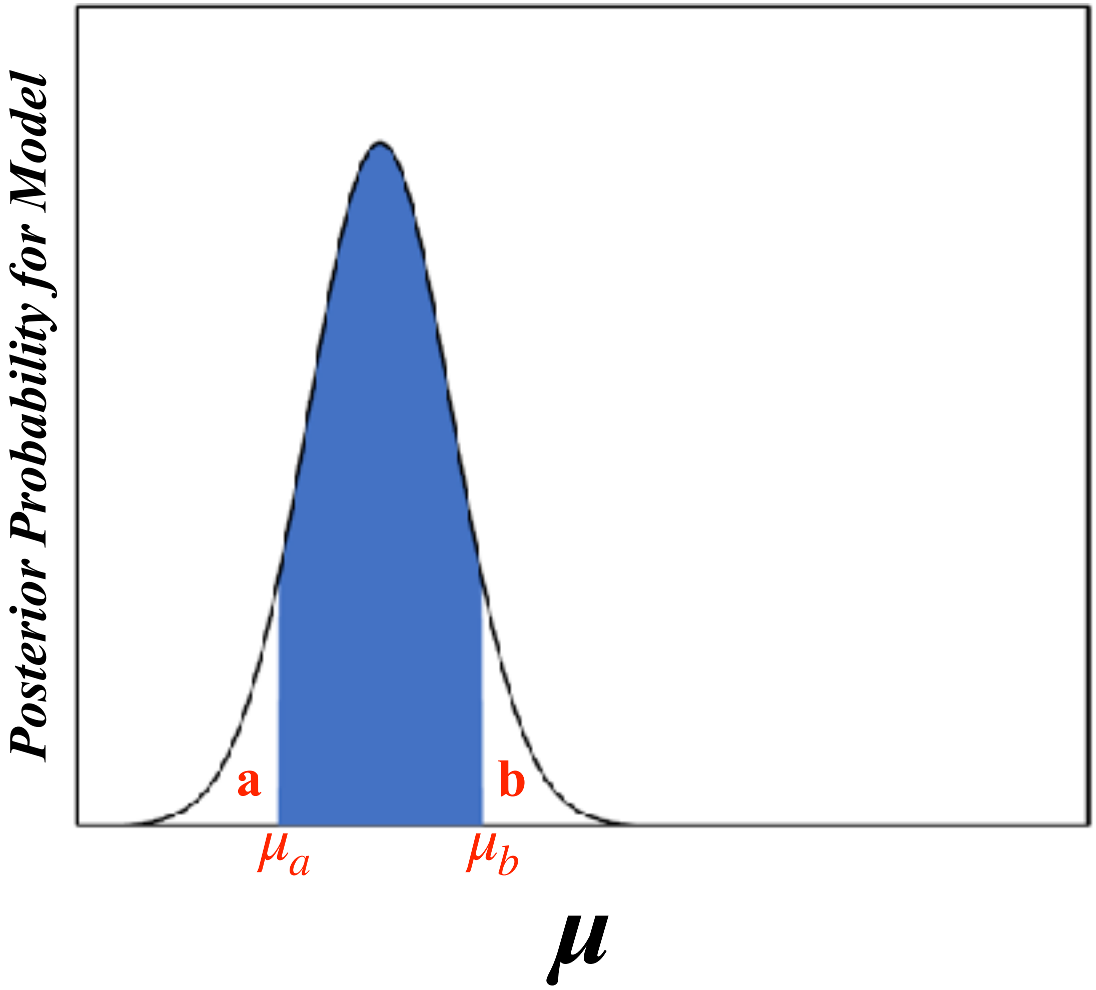
---

* 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:



This is now the model prediction, not an ‘ordering parameter’ based on an observation

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.

Let’s again consider the case of
Poisson statistics as an example…

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.

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

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 in linear space, 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 take the the standard frequentist bounds and renormalise them only up to the range observed in the current data set.

For example, in the case of a Poisson upper limit:

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

“standard” frequentist

normalising only up to observed range

Which is identical to a Bayesian bound with a prior that is constant and non-negative, if used to bound the space of possible models!

But this tends to be interpreted in a **very different** frequentist way: “If I were to repeat this experiment many times, *and only looked at cases where fluctuations of the background are less than the total number seen in this particular observation*, the true model would be bounded in a CL fraction of those cases.”

But this tends to be interpreted in a **very different** frequentist way: "If I were to repeat this experiment many times, *and only looked at cases where fluctuations of the background are less than the total number seen in this particular observation*, the true model would be bounded in a CL fraction of those cases."

What this means is a little unclear: it's not really a "frequentist" paradigm, since the intervals are not based on the frequency of all possible fluctuated measurements. Normalising this way tends to over-cover the range of all allowed occurrences to produce 'more conservative' frequentist bounds… but in a very *ad hoc* way.

It was adopted pragmatically, and is still used by some, because it seems to produce more "sensible results."*

This is probably more of a statement that the question they really want addressed is, in fact, not the frequentist one: the behaviour of standard frequentist bounds is, of course, perfectly "sensible" and self-consistent… if that is the question you're asking! Statements of "sensible behaviour" seem to instead refer to that expected from Bayesian bounds with a constant prior in the physical region.

*Amnon Harel. "Statistical methods in CMS searches" (PDF). indico.cern.ch. Retrieved 2015-04-10.

A useful mantra:

Don’t Kludge!

- Understand the question
- Use a self-consistent framework
- Listen to the math!

(Back to Bayes)

Bayesian Propagation of Systematic Uncertainties

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

Bayesian Integration vs Profile Likelihood Maximisation

The Profile Likelihood method, where dimensionality is reduced by taking the maximum likelihood for marginalised parameters, does not, in fact, yield a true likelihood function in the reduced parameter space*. This basically is because "perfect knowledge" of the marginalised parameter (set to the maximised value) is then assumed in the new distribution without accounting for it's uncertainty, so the effective number of degrees of freedom that may then be used to apply Wilks' Theorem etc. isn't quite right.

Integration over nuisance parameters is the formally correct way to map to a new probability distribution in a lower dimensional space.

But you can only formally do this in a Bayesian paradigm, since frequentism is specifically designed to keep model separated!

* Aitkin, M. (2005). Profile Likelihood. In Encyclopedia of Biostatistics, John Wiley & Sons.

However, in many cases, maximisation can be used to approximate integration, with an argument often attributed to Laplace*:

$$\mathscr{L}(q') = \int \mathscr{L}(q',q)dq = \int \exp\left[\ln\left(\mathscr{L}(q',q)\right)\right] dq$$

fixed parameters

parameters being marginalised over

Taylor expand about the maximum of the likelihood:

$$\mathscr{L}(q') \simeq \int \exp\left[\ln \mathscr{L}(q',q_m) - \frac{1}{2}\left|\frac{d^2 \ln \mathscr{L}(q',q)}{dq^2}\right|_{q=q_m}(q-q_m)^2\right] dq$$

(ignoring higher order terms)

---

* P.S. Laplace, “Memoir on the probability of causes of events,” Memoires de Mathematique et de Physique, Tome Sixieme, (1774). (English translation by S. M. Stigler, Statist. Sci., 1(19):364378, 1986.).

$$\mathscr{L}(q') \simeq \int \exp\left[\ln \mathscr{L}(q', q_m) - \frac{1}{2}\left|\frac{d^2 \ln \mathscr{L}(q', q)}{dq^2}\right|_{q=q_m} (q - q_m)^2\right] dq$$

$$\mathscr{L}(q') \;\propto\; \mathscr{L}(q', q_m) \int \exp\left[-\frac{1}{2}\left|\frac{d^2 \ln \mathscr{L}(q', q)}{dq^2}\right|_{q=q_m} (q - q_m)^2\right] dq$$

$$\propto\; \mathscr{L}(q', q_m) \frac{1}{\sqrt{\left|\frac{d^2 \ln \mathscr{L}(q', q)}{dq^2}\right|_{q=q_m}}}$$

This also works for Poisson distributions, since n! terms cancel in the likelihood ratio

for distributions, such as Gaussians, this is also a constant

$$\ln \mathscr{L}(q') \simeq \ln \mathscr{L}(q', q_m) - \frac{1}{2} \ln \left|\frac{d^2 \ln \mathscr{L}(q', q)}{dq^2}\right|_{q=q_m} + C$$

Don't care about constants for likelihood ratios (or differences in the log)

Differences in the log likelihood marginalised via integrated

$$\Delta \ln \mathscr{L}(q') \simeq \Delta \mathscr{L}(q', q_m)$$

Differences in the log likelihood marginalised via taking maximum

BUT, for example, this doesn't work for Gaussians where σ is nuisance parameter! Marginalising this way gets things wrong! (see Aitkin again)

…RTMENT OF JUSTICE, BUREAU OF INVESTIGATION

IDENTIFICATION DIVISION, WASHINGTON, D. C.

Located at

Received

From ST. LOUIS CITY

Crime Robbery 1st

Sentence 5 yrs. — mos. — days

Date of sentence 12-8-1925

Sentence begins

Sentence expires

Good time sentence expires 11-7-1928

Date of birth 2-3-1904 Occupation Waiter

Birthplace Ga. Nationality Amer.

Age 21 Build M Stout

Height 5' 7 7/16 Hair Black

Eyes Bl Grey Weight 157

Comp. Ruddy

Scars and marks IT 7½ cm vert sc L side top head front

CRIMINAL HISTORY

NAME	NUMBER	CITY	DATE	CHARGE	DISPOSITION

Note: Charles Arthur Floyd is wanted in connection with the murder of Otto Reed, Chief of Police, McAlester, Okla., William J. Grooms, and Frank E. Hermanson, police officers of Kansas City, Mo., Raymond J. Caffrey, Special Agt., U.S. Bu. of Invest., and their prisoner Frank Nash, at Kansas City, Mo., on 6-17-33 (Inf rec I O #1184) wjm 8-10-33.

Priors

What’s the way out??

Pragmatism!

There is no formally “correct” choice of prior! But it is generally possible to define choices that conservatively span the range of reasonable possibilities

Some Guidance on Priors

- Choose simple prior forms that are easy to understand and visualise (*e.g.* constant over the range of interest) 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 an "unconstrained" 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 constant with rate is conservative for setting an upper bound.

- For an unknown phase angle, a prior that is constant with phase angle often makes sense. (Note that if you choose a prior that is instead constant with sin(2θ) in a neutrino oscillation analysis, you are saying that you have an *a priori* preference for smaller angles!)

- If deciding between distinct, discreet hypotheses (*e.g.* normal or inverted neutrino mass orderings), use a prior that gives them equal weight unless there is a very strong argument to prefer one over another.

- 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 constant with either linear or logarithmic scales often provide reasonable bounds.

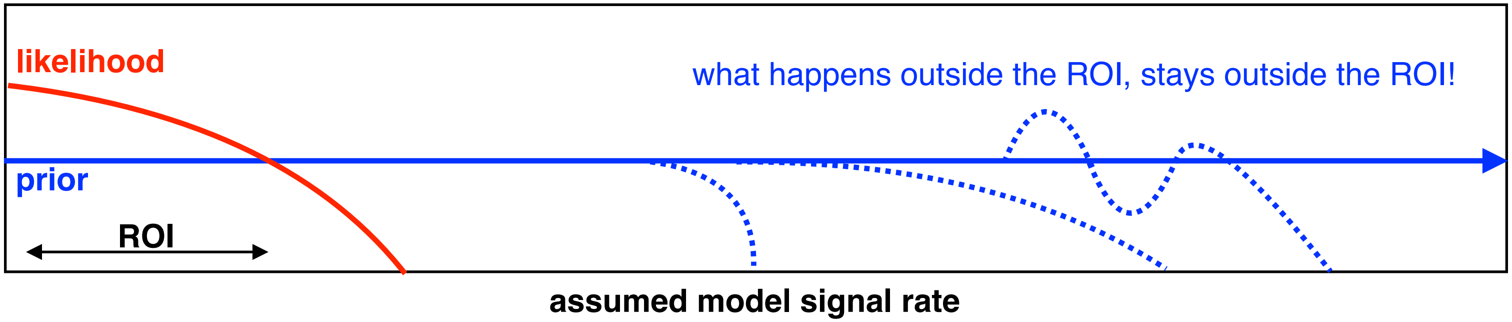
- If there's an ambiguity that leads to a non-conservative bound, show the sensitivity to the choice of prior!

It's really not that hard!

The Use of “Constant” Priors

This refers to using a constant prior for a particular model parameter. For example, priors constant in signal rate mean that you ascribe equal weight to all signal rates.

But is that really realistic? That would allow the possibility of an infinitely large signal and results in a probability distribution that cannot be normalised!!



What we actually mean is that the prior is roughly constant in the vicinity of the region of interest, and then tails off in some way that does not need to be specified because the likelihood crushes its impact as soon as you get much outside the ROI

(more mantras)

The choice of prior only matters if the data itself is not strong enough to unambiguously define the model!

And, if it is ambiguous, you should show the sensitivity of conclusions to <u>reasonable</u> choices of prior!

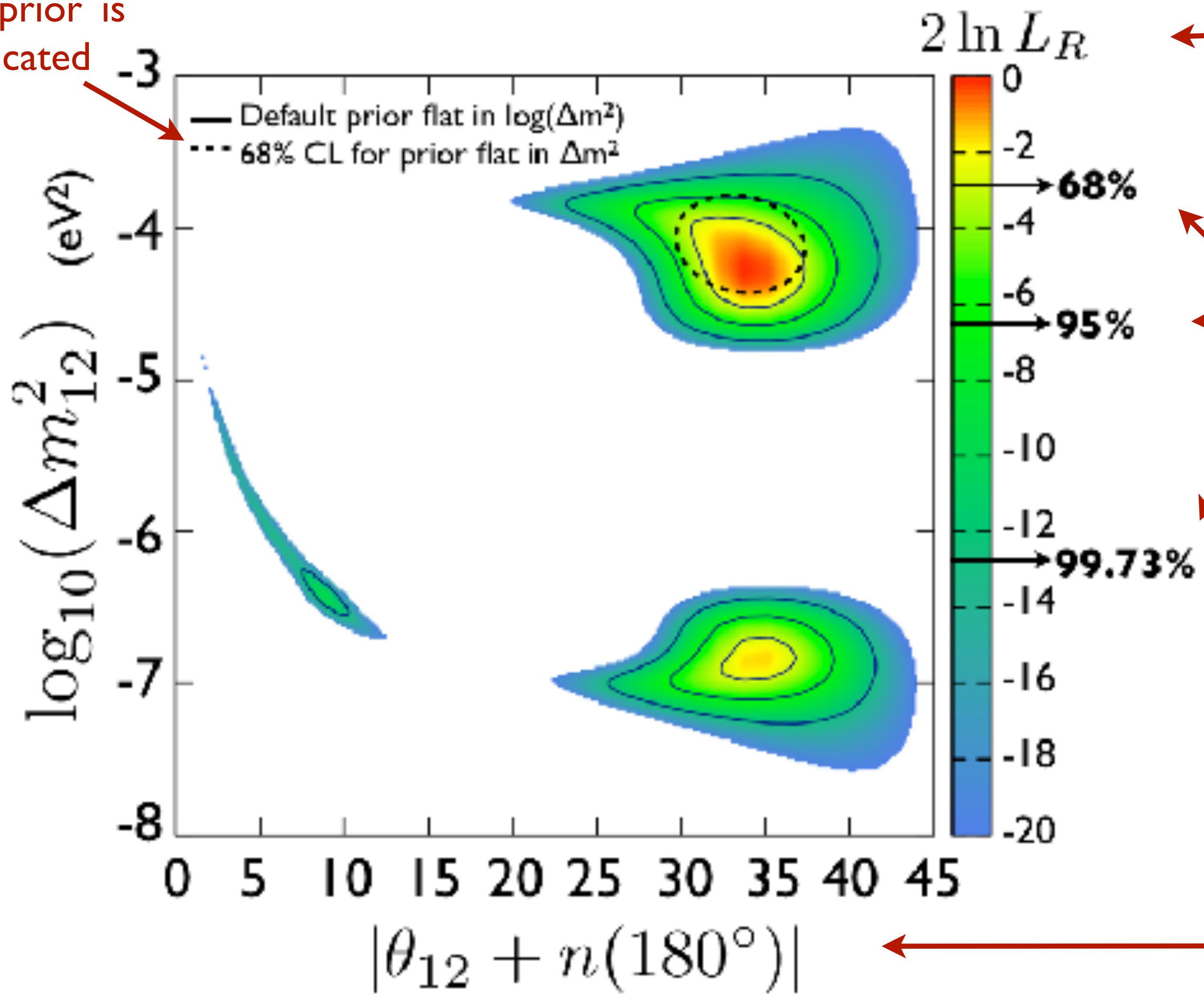
Typically, especially when there’s a clear signal, the differences between bounds derived by frequentist and Bayesian constructions are often minor.

But, in the instances when it does matter, you want to be on the right side of things!

Note: Displaying the likelihood as a function of variables for which the priors are constant, 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)



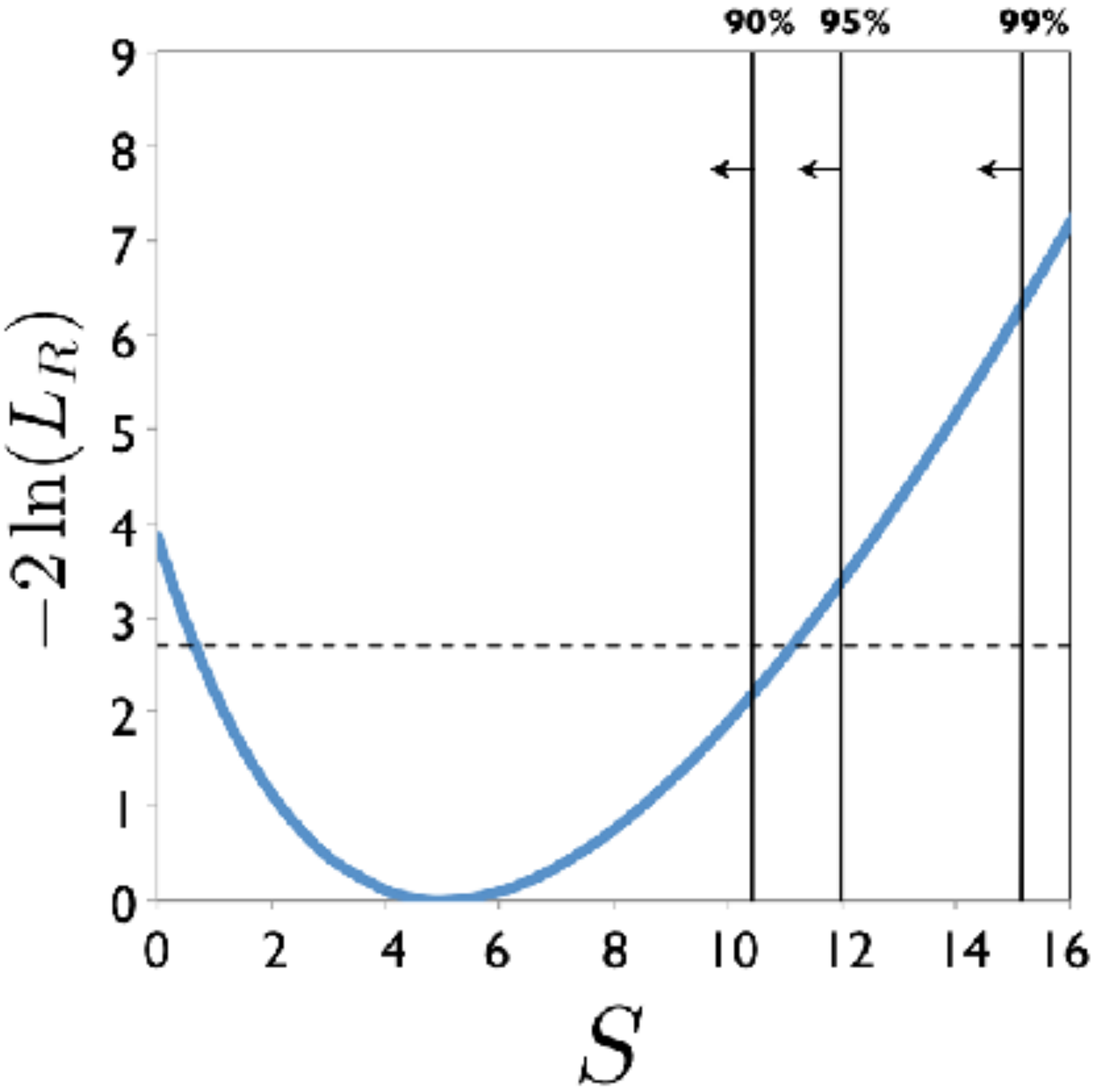
Sensitivity to prior is indicated

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

Bayesian contours from integration of likelihood assuming priors constant in θ and log(Δm^2)

Form for fundamental angle accounts for quadrant ambiguity

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

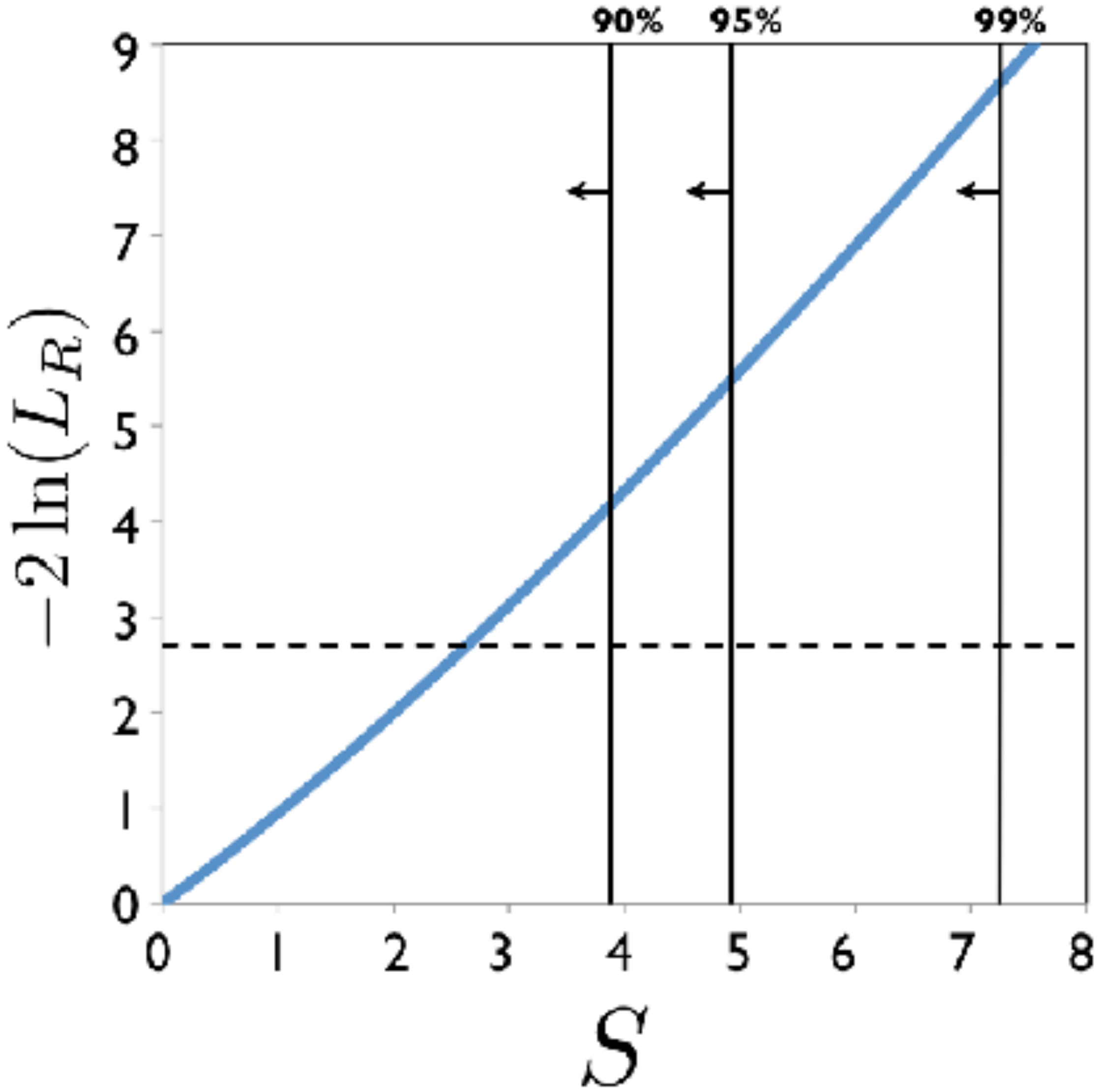


Bayesian upper bounds from integration of likelihood assuming priors constant in S ('conservative')

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)

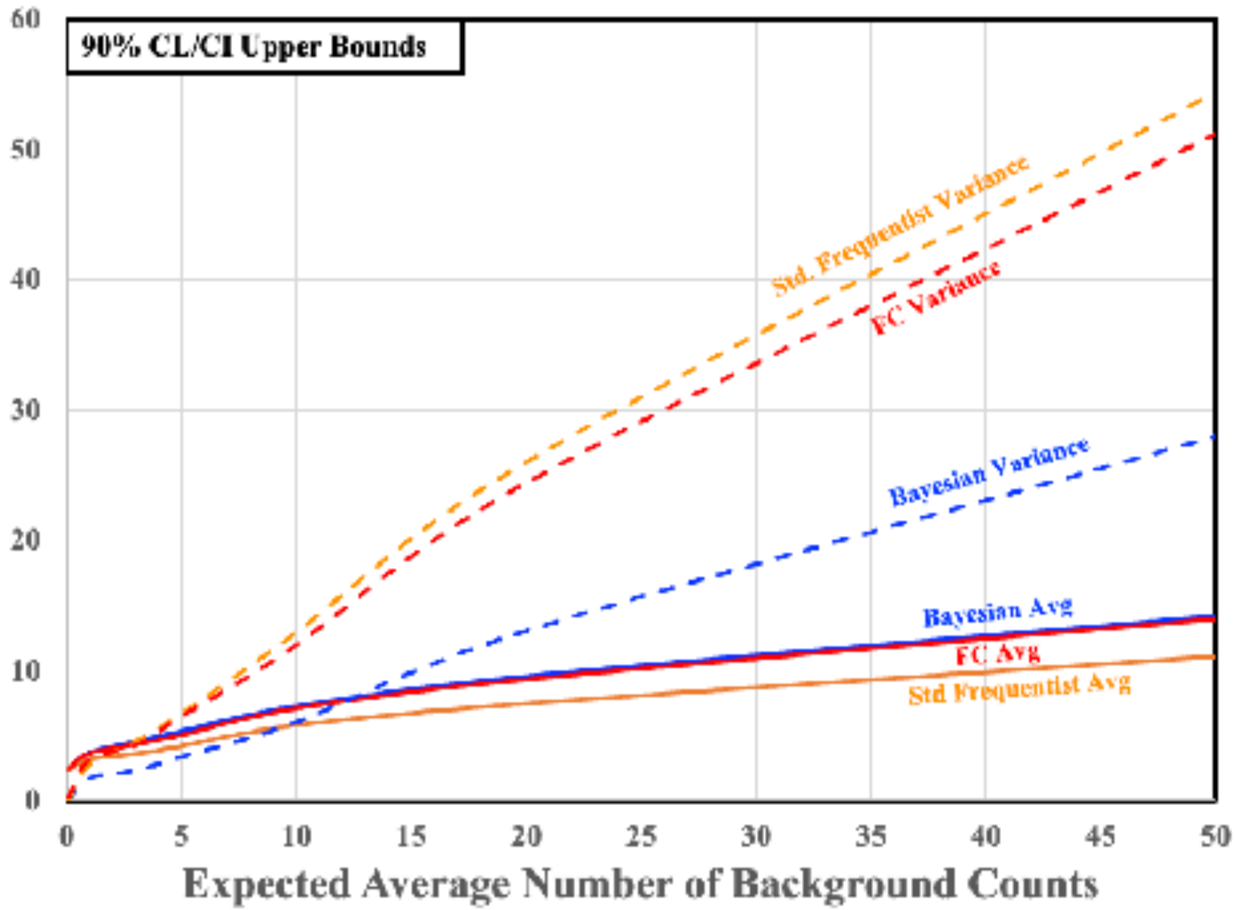


Bayesian upper bounds from integration of likelihood assuming priors constant in S (‘conservative’)

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

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

Robustness of Upper Bounds for Poisson Statistics



The numerical values of Bayesian bounds have notably less variance than frequentist methodologies - more robust for comparison of experimental results!

“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.

As previously stated, frequentist bounds are all about the distribution of the ensemble of hypothetical experiments and not about ascribing meaning to your particular interval.

But what if lots of people do experiments and each defines frequentist bounds, so that you start to have a **real** ensemble. How do you then use this to set bounds on models?

Still Bayesian! Make use of the likelihoods for all these data sets together *(not their frequentist bounds!)* and choose your prior etc. There is no other way! As the ensemble becomes larger and larger, the prior becomes less and less important, and the distinction between frequentist and Bayesian bounds goes away.

Pragmatism: You can use frequentist bounds for models when it gives the same answer as Bayesian bounds.

The Point of Frequentism:

Want to display the results of analyses in a model-independent way that has the most general possible applicability

Absolutely!! Always do this! For example, try to provide sufficient views of the data to allow others to roughly reproduce your results, and show the likelihood distribution, **which gives the full frequentist information content of the data.**

But, if you then want to use this to constrain models, that’s Bayesian!

Both of these are important aspects of data presentation.

Summary

• Bayesian statistics is the <u>**only**</u> 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 from a theoretical ensemble of experiments. *Unfortunately, this is often misinterpreted (or correctly interpreted but then misused).*

• The Feldman-Cousins approach is an **equally** valid reformulation in terms of the frequency of a derived relative quantity, rather than of a direct measurement value, that shares exactly the same caveats (*though may be even more prone to misinterpretation*).

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!