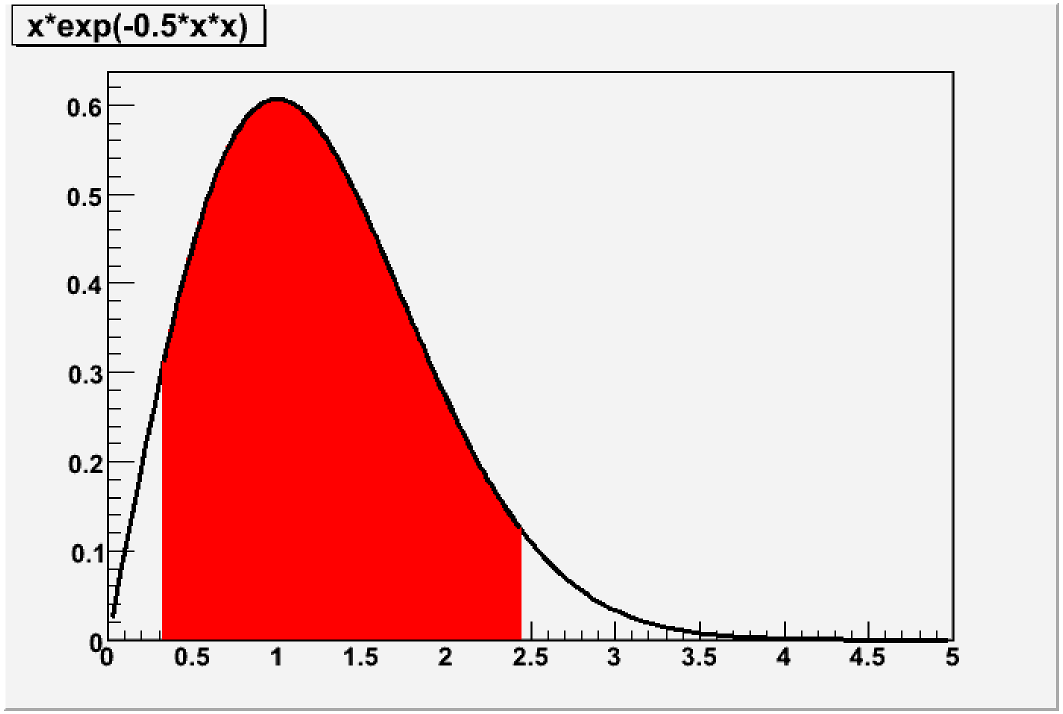
Physics 509: Fun With Confidence Intervals

Scott Oser
Lecture #16

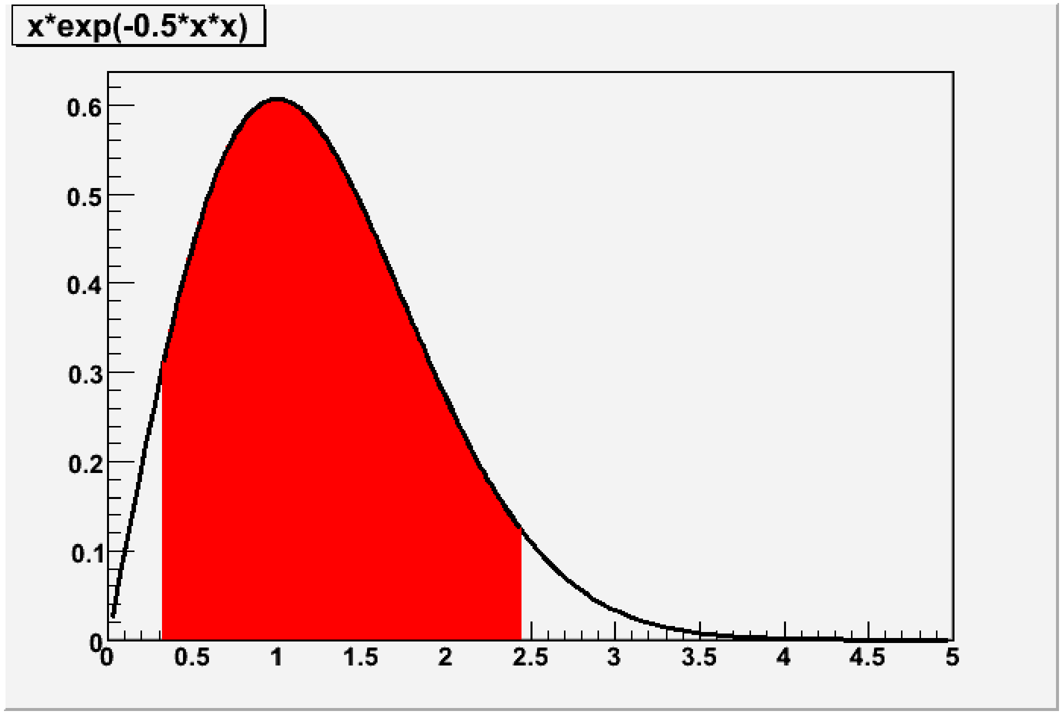


Bayesian credible region

Bayesians generally prefer to report the full PDF for the posterior distribution of a quantity.

If desired to report a range for the parameter, an obvious solution is to integrate the PDF .

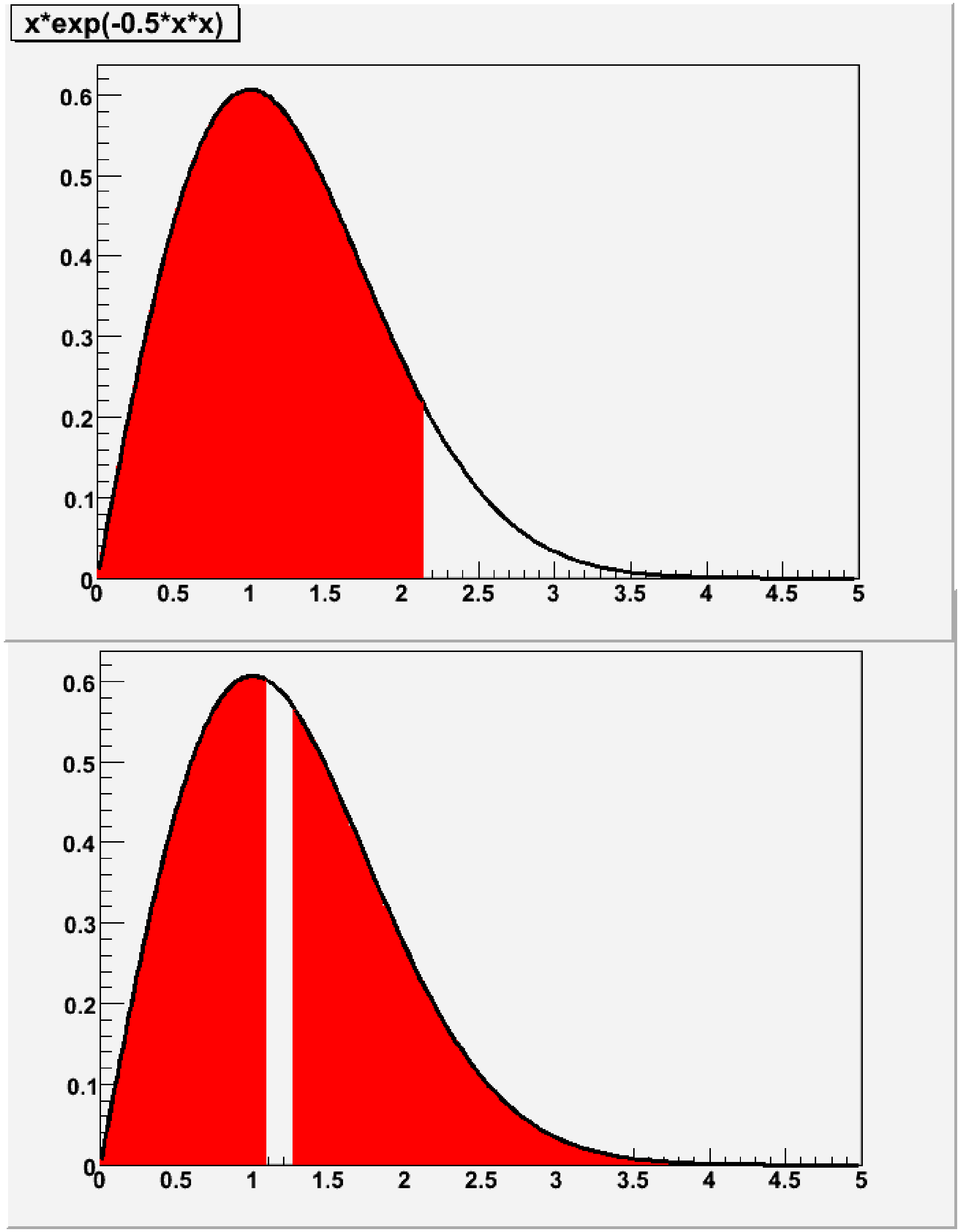
The red area contains 90% of the probability content---the Bayesian credible region is (0.32,2.45)



These are also Bayesian credible regions

The top plot might be appropriate if you were asked to quote an upper limit on the parameter: (0,2.15)

The red region on the bottom also contains 90% of the probability content. You might quote the disconnected credible region (0,1.09) & (1.26,∞) if you were on crack.



Exact Neyman confidence intervals

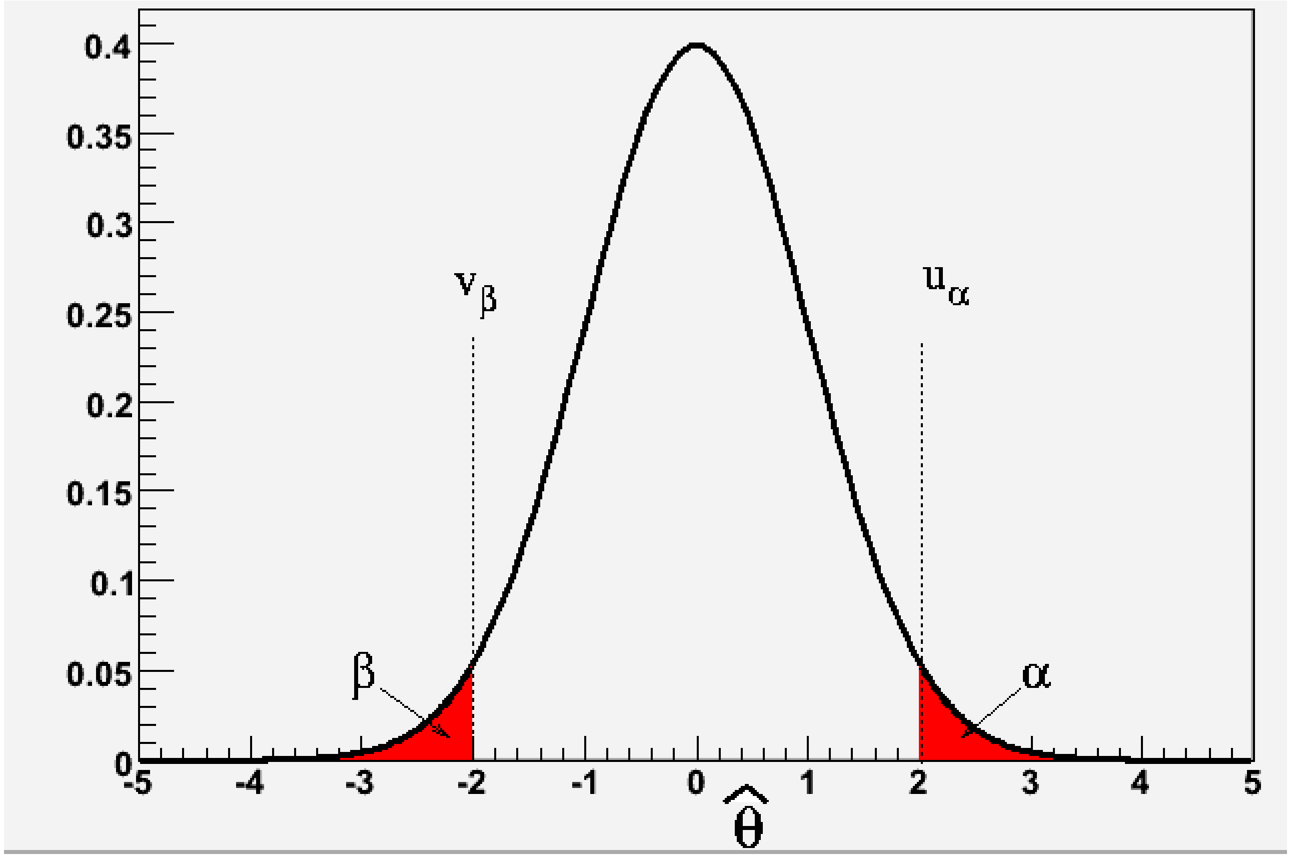
A frequentist confidence interval is a different beast. While a Bayesian credible region is based on the probability that the true parameter lies in the specified, the frequentist interval really refers to the probability of getting the observed data.

The Neyman construction is a procedure for building classical frequentist confidence intervals:

1) Given a true value *a* for the parameter, calculate the PDF for your estimator *â* of that parameter: P(*â*|*a*).
2) Using some procedure, define the interval in *â* that has a specified probability (say, 90%) of occurring.
3) Do this for all possible true values of *a*, and build a confidence belt of these intervals.

A two-sided confidence interval

Frequentist techniques don't directly answer the question of what the probability is for a parameter to have a particular value. All you can calculate is the probability of observing your data given a value of the parameter. The confidence interval construction is a dodge to get around this.

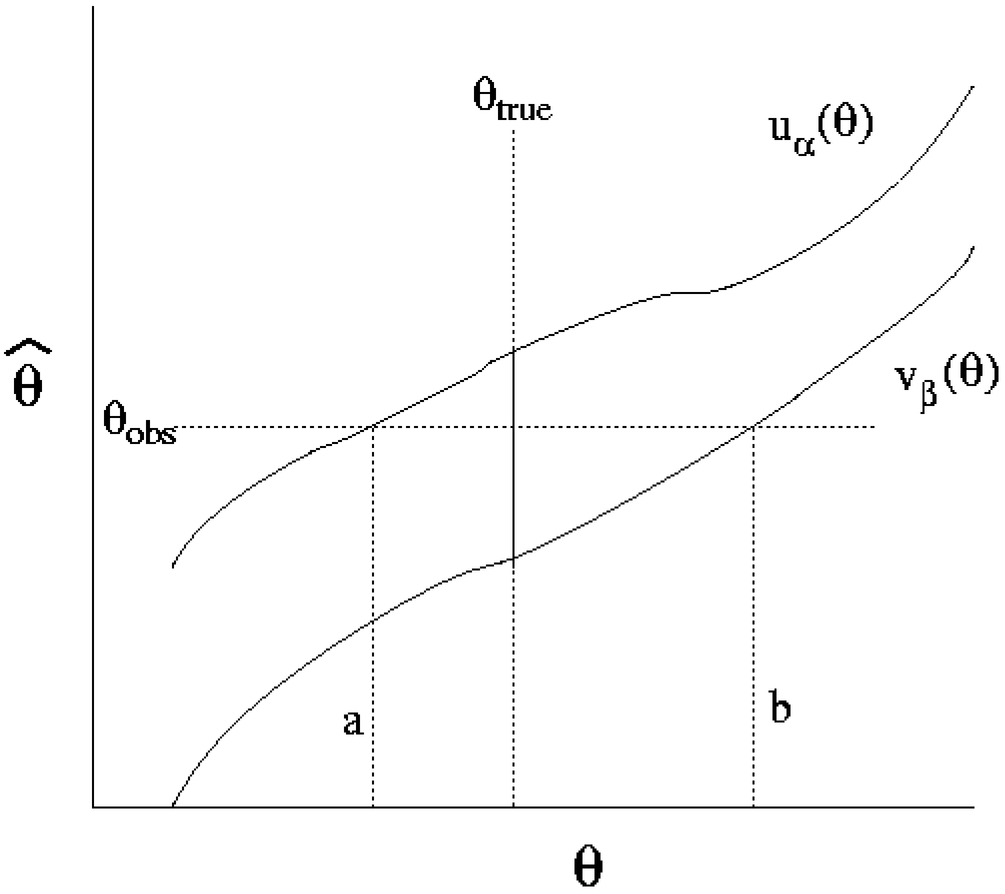


Starting point is the PDF for the estimator, for a fixed value of the parameter.

The estimator has probability $1-\alpha-\beta$ to fall in the white region.

For the obvious choice $\alpha=\beta$ we call this region a central confidence interval.

Confidence interval construction



The confidence band is constructed so that the average probability of θ_{true} lying in the confidence interval is $1-\alpha-\beta$.

Consider any true value of the parameter, such as θ_{true}. The probability that the measured value of the estimator lies on the vertical segment is $1-\alpha-\beta$.

The interval (a,b) will cover θ_{true} if θ_{obs} intersects this vertical line segment, and not otherwise.

By construction, the probability of the confidence interval from this method containing the true value of the parameter is $1-\alpha-\beta$. This sounds like a statement about the true value of θ, but it's really a statement about how (a,b) is generated.

Arbitrariness of confidence interval construction: one-sided vs. two-sided

There is no single way to build the confidence interval. You can make one-sided, two-sided, or even more complicated confidence belts depending on what parts of the PDF you include inside the belt.

As an example, let's build a one-sided confidence belt for a parameter $\mu>0$ whose estimator has a Gaussian distribution. Suppose that:

$$P(\hat{\mu}|\mu)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\mu-\hat{\mu})^2\right]$$

For any fixed μ, 90% of the probability is contained within

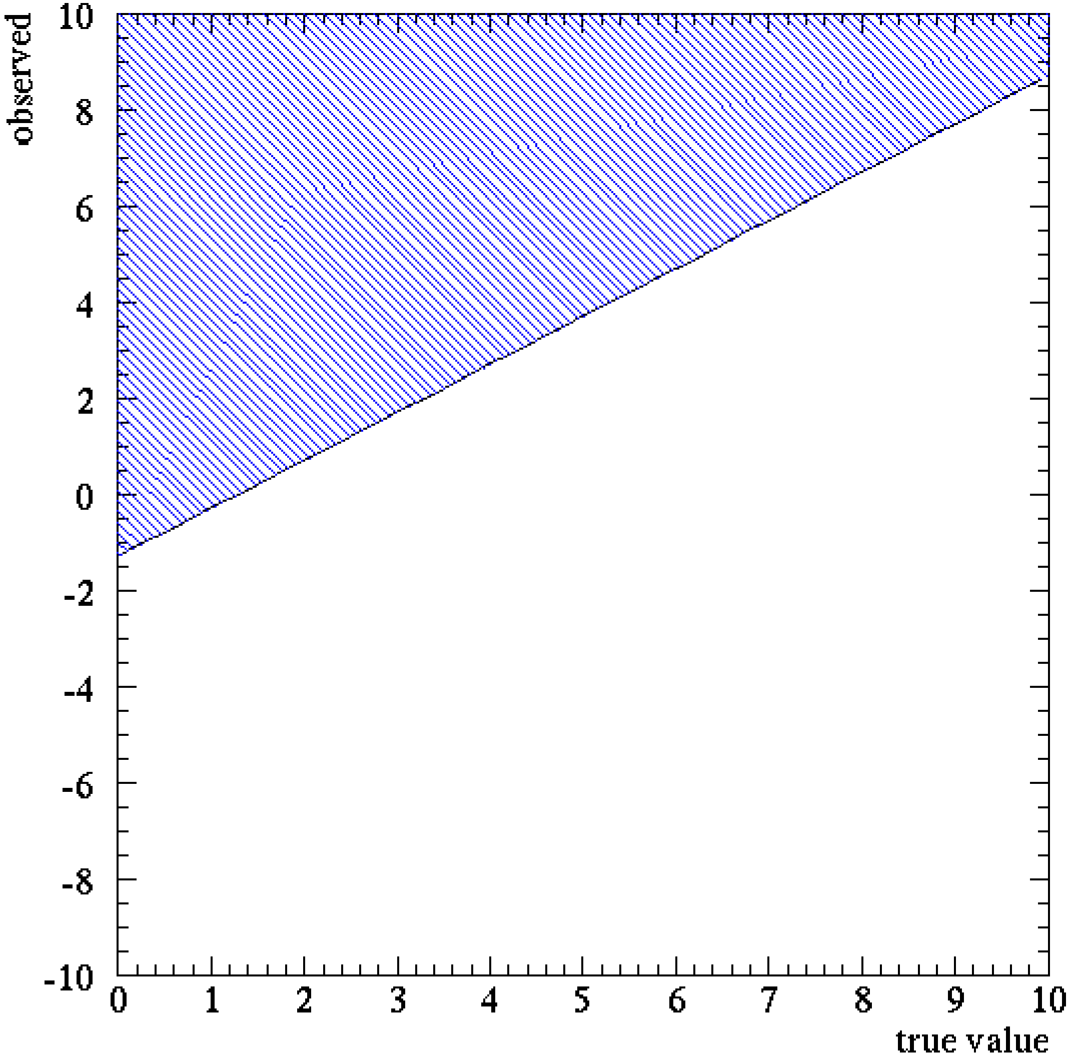
$$\mu-1.28<\hat{\mu}<+\infty$$

Again, there is a 90% probability of the *measured* value falling in this range.

One sided confidence belt

The shaded region is the confidence belt. Read this as saying that for any given true value of m, there's a 90% that the measured value will lie above the line in shaded region.

This in turn generates a confidence region for any observed value.

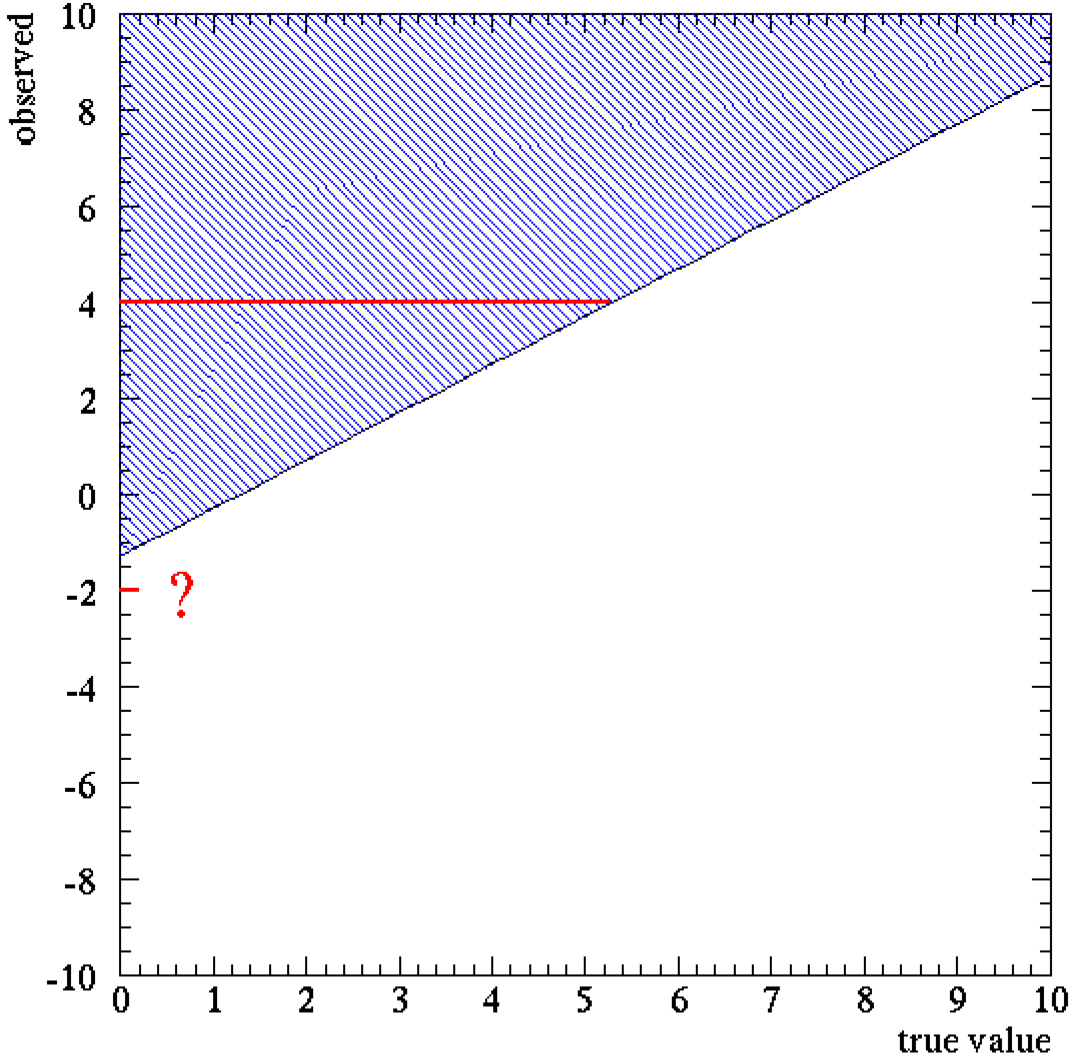


One sided confidence belt

For example, if we measure μ=4, the confidence belt says that the true value of μ lies between 0 and 5.28. We'd say our 90% C.L. upper limit on μ was 5.28.

If we had measured μ=-1.27, then our region would be (0,0.01)---pretty small.

But what if we had measured μ=-2? You might expect the region (-∞, -0.72). But remember: we stipulated μ had to be positive. (Maybe μ is a mass.) ***The confidence interval is an empty set.***



Interpretation of the one-sided confidence belt

Suppose we measured μ=-1.27, and generated the confidence interval (0,0.01). This sounds really strange---we measured a negative, non-physical value for μ, and as a result we get an extremely tight confidence interval for the true value.

Does this really mean that if we measure μ=-1.27 then there is a 90% chance that the true value of μ is between 0 and 0.01?

Interpretation of the one-sided confidence belt

Suppose we measured μ=-1.27, and generated the confidence interval (0,0.01). This sounds really strange---we measured a negative, non-physical value for μ, and as a result we get an extremely tight confidence interval for the true value.

Does this really mean that if we measure μ=-1.27 then there is a 90% chance that the true value of μ is between 0 and 0.01?

No. The confidence belt is constructed so that in 90% of experiments it will contain the true value of the parameter. In this case, getting a value so close to the physical limit can only mean that this particular experimental outcome is likely to be one of the 10% which doesn't contain the true value.

If we had measured μ to be even smaller, the confidence region would be the empty set. This doesn't mean that all values of μ are ruled out---it means that we're definitely in the 10% of experiments which fail to contain the true value.

Be very careful with the interpretation of frequentist confidence intervals

Most of us are psychologically inclined to think of confidence intervals as Bayesian creatures. That is, if someone says their 90% C.L. for μ is (2.5,2.9), then we tend to think that means there's a 90% chance that the true value of the parameter lies between 2.5 and 2.9.

But that's not right---frequentist confidence intervals are designed to give proper coverage only for a hypothetical ensemble of many experiments. It means that if you did the experiment 100 times, then on average 90 of the generated confidence intervals would contain the true value.

It is not necessarily the case that *for your particular data set*, the probability that *your* confidence interval will contain the true value is 90%. Depending on your data, the probability could be less. In fact, you might even KNOW that the confidence interval doesn't contain the true value---for example, if the confidence interval is in the unphysical region.

The “flip-flop” problem

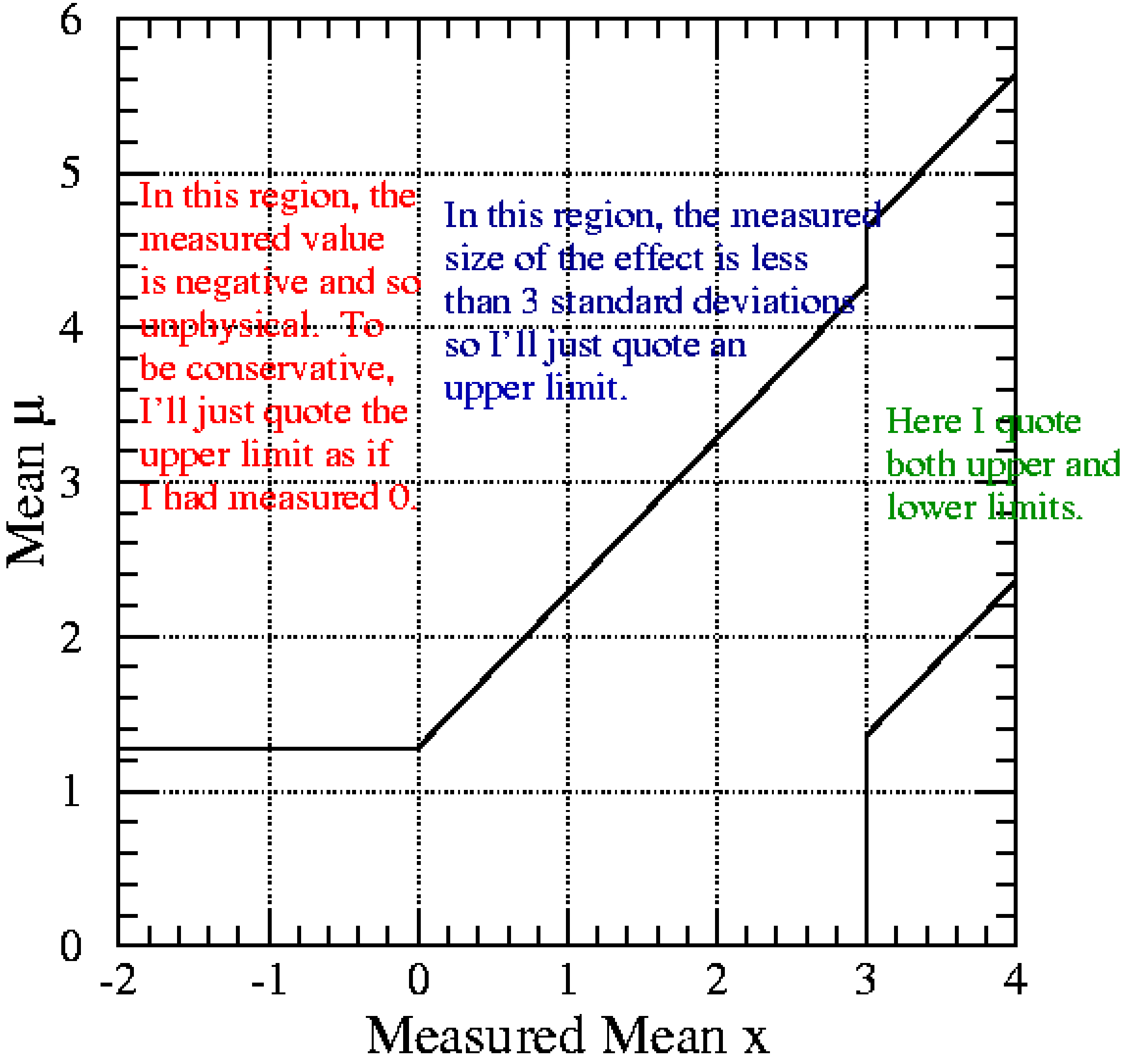
Imagine an experimenter who thinks the following:

“I'm looking for some new effect. I don't know if it exists or not. If it doesn't exist, I should probably report a 90% upper limit on the size of the effect. If it does exist, I'll instead want to report the measured value for the effect, by which I mean the central confidence interval with both upper and lower limits on the effect (not just an upper limit).

I don't know which is the case, though. So what I'll do is take my data and see. If I detect a non-zero value at greater than 3σ, I'll quote both an upper and lower limit, while if my measured value is less than 3σ from zero, I'll just report an upper limit.”

Let's see what the confidence bands look like.

The “flip-flop” confidence belt

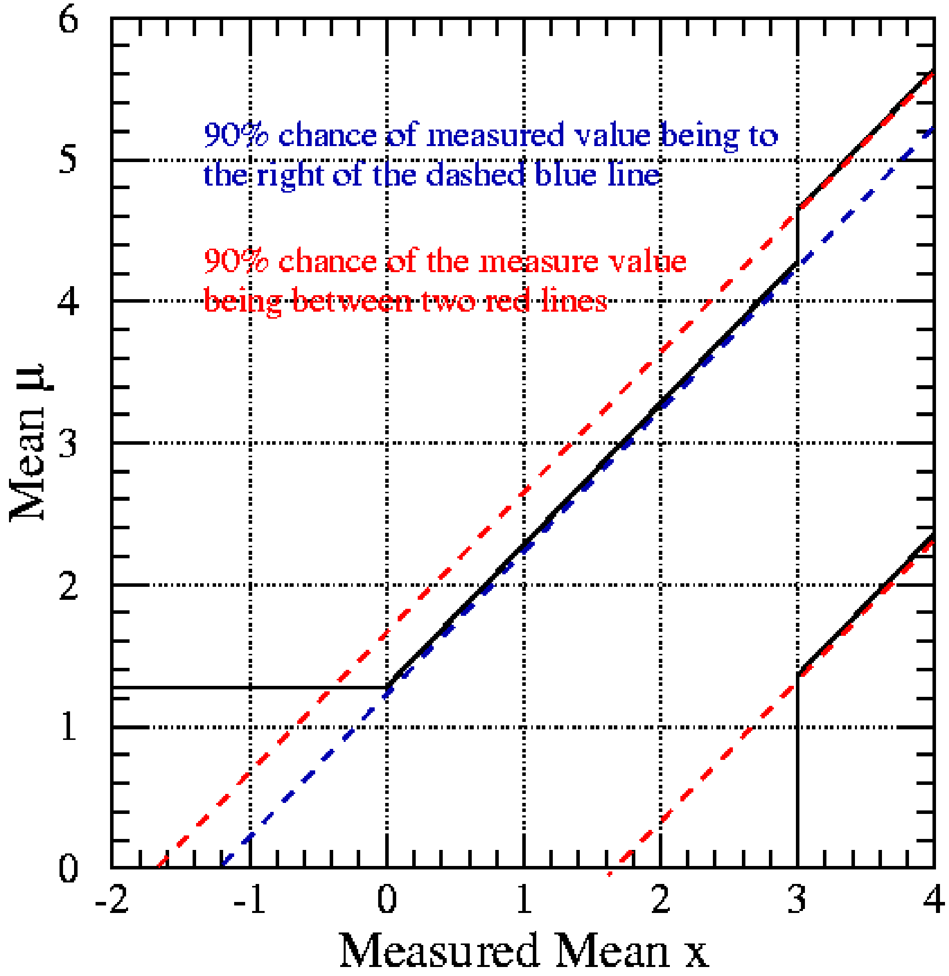


In this region, the measured value is negative and so unphysical. To be conservative, I’ll just quote the upper limit as if I had measured 0.

In this region, the measured size of the effect is less than 3 standard deviations so I’ll just quote an upper limit.

Here I quote both upper and lower limits.

Coverage of the “flip-flop” confidence belt



The coverage of the intervals is wrong. For example, for 1.36<μ<4.28, the measured value has only an 85% chance of being in the region, not the 90% chance we designed for.

For small μ the confidence interval overcovers---we deliberately tried to be conservative when we measure a negative number to avoid absurdly small regions.

For coverage to be meaningful, must decide ahead of time what kind of limit to quote.

Is there an alternative?

The classical confidence intervals shown previously have some regrettable properties:

- at least some fraction of the time the confidence interval can be an empty set
- they do not elegantly handle unphysical regions
- they do not continuously vary between giving upper limits vs. giving upper and lower limits, but instead change discontinuously depending on which you choose.

In a paper by Feldman & Cousins (arXiV:physics/9711021 v2) these issues are explored in some detail, and a solution is proposed. The result is what is known as a Feldman-Cousins confidence interval, which we'll now examine.

Ordering principle

The Neyman confidence interval construction does not specify how you should draw, at fixed μ, the interval over the measured value that contains 90% of the probability content.

There are various different prescriptions:

- 1) add all parameter values greater than or less than a given value (upper limit or lower limit)
- 2) draw a central region with equal probability of the measurement falling above the region as below
- 3) starting with the parameter value which has maximum probability, keep adding points from more probable to less probable until the region contains 90% of the probability
- 4) The Feldman-Cousins prescription (next slide!)

Feldman-Cousins confidence intervals

Feldman-Cousins introduces a new ordering principle based on the likelihood ratio:

$$R = \frac{P(x|\mu)}{P(x|\mu_{best})}$$

Here x is the measured value, μ is the true value, and μ_{best} is the best-fit (maximum likelihood) value of the parameter given the data and the physically allowed region for μ.

The order procedure for fixed μ is to add values of x to the interval from highest R to lower R until you reach the total probability content you desire.

Taking a ratio “renormalizes” the probability when the measured value is unlikely for any value of μ. The Feldman-Cousins confidence interval is therefore never empty.

Application of Feldman-Cousins to Gaussian with physical limit

Feldman-Cousins introduces a new ordering principle based on the likelihood ratio:

$$R = \frac{P(x|\mu)}{P(x|\mu_{best})}$$

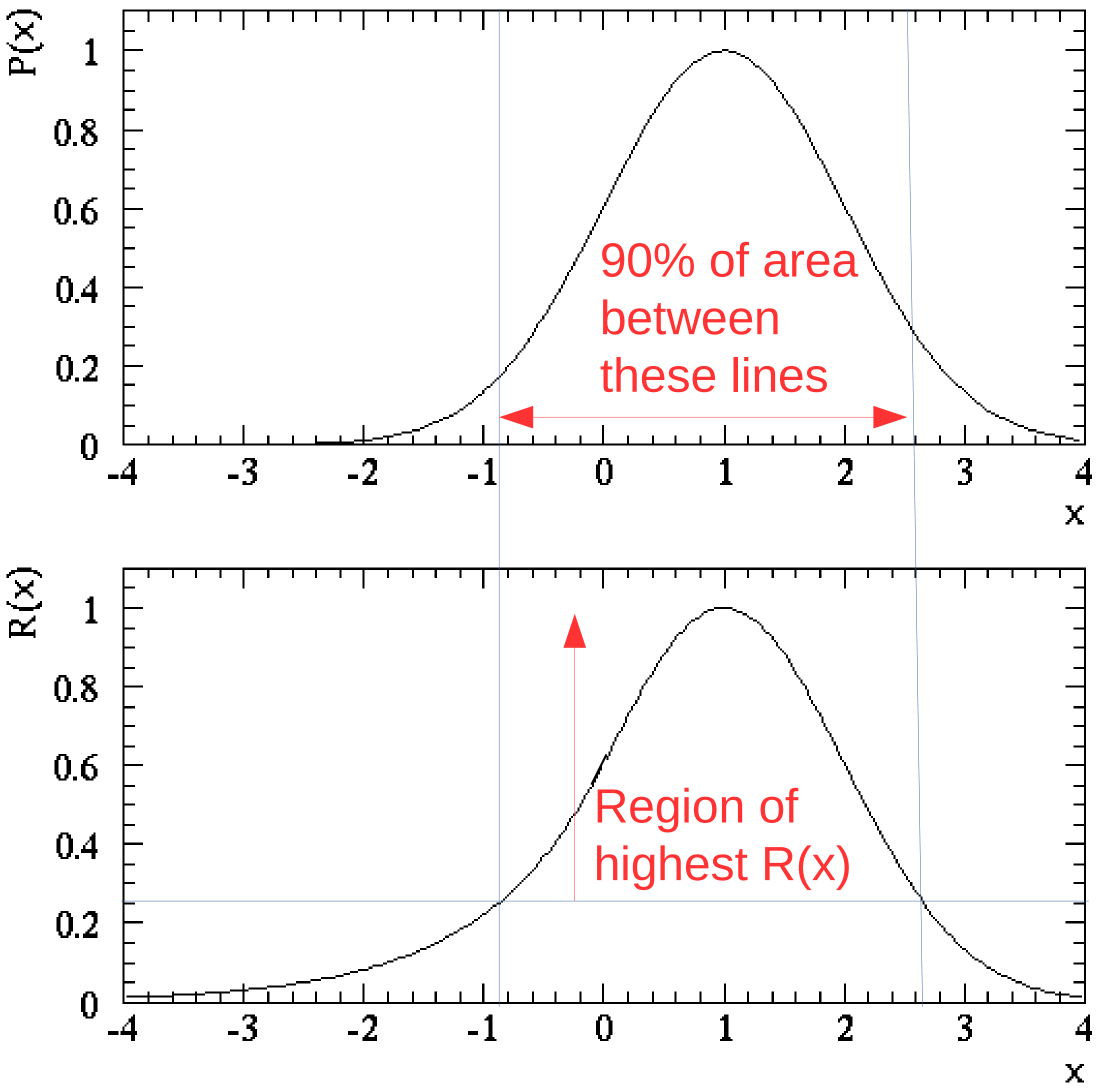
For our example with a Gaussian measurement with unit RMS, we have

μ_{best}=x if x>0 or μ_{best}=0 if x≤0. So the ratio R is given by

$$R = \frac{P(x|\mu)}{P(x|\mu_{best})} = \frac{\exp\left[-\frac{1}{2}(x-\mu)^2\right]}{1} \quad \text{if } \ x>0$$

$$R = \frac{P(x|\mu)}{P(x|\mu_{best})} = \frac{\exp\left[-\frac{1}{2}(x-\mu)^2\right]}{\exp\left[-\frac{1}{2}x^2\right]} \quad \text{if } \ x\leq 0$$

Application of Feldman-Cousins to Gaussian with physical limit

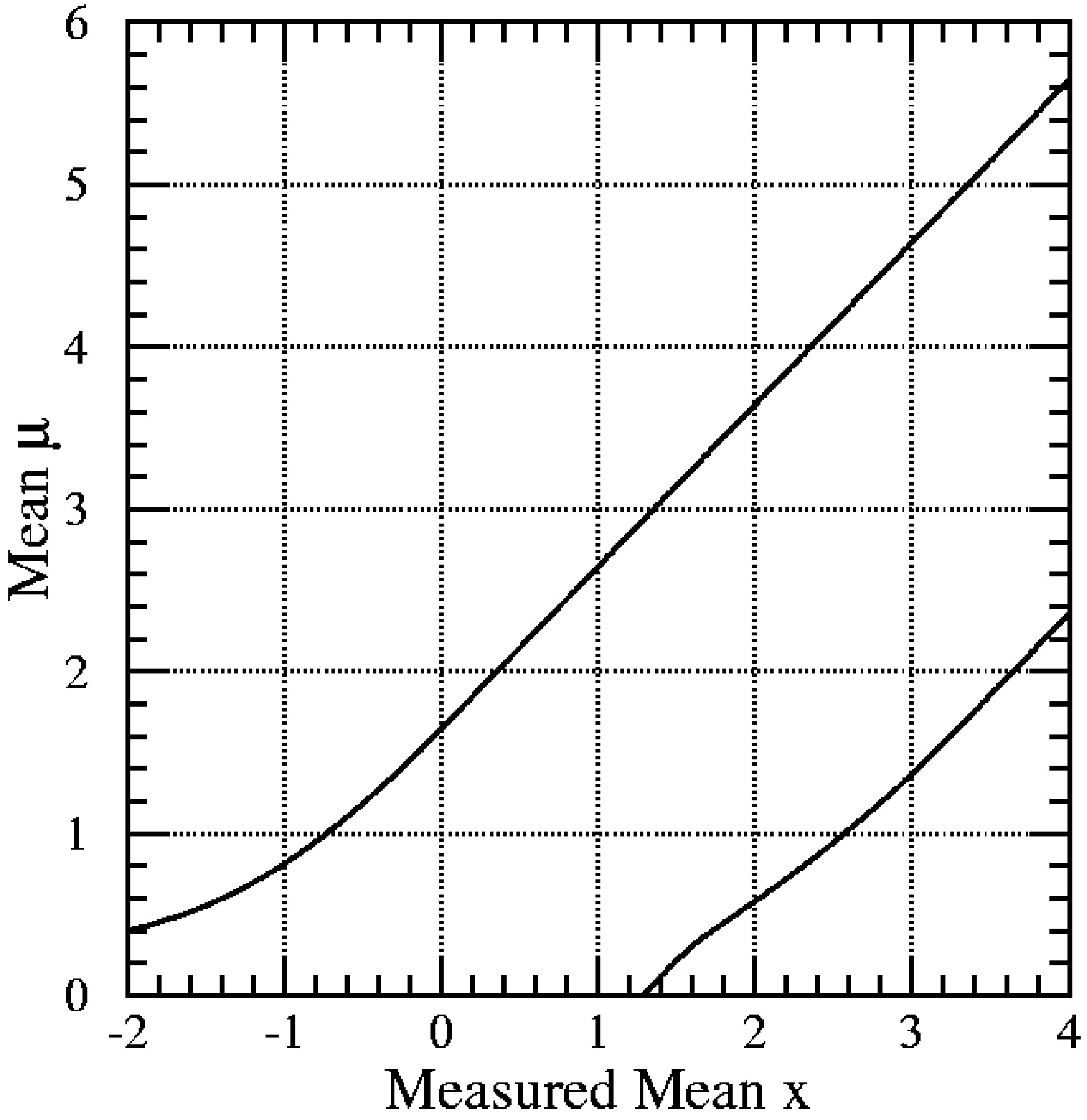


90% of area between these lines

Region of highest R(x)

Example for μ=1

Application of Feldman-Cousins to Gaussian with physical limit



To the left is the Feldman-Cousins confidence belt. Some nice features:

1) Confidence region is never empty, no matter what you measure.
2) It smoothly transitions between an upper limit (lower limit=0 at physical boundary), and a two-sided limit! It gives correct coverage and decides for you when to quote one-sided vs. two-sided limit!

Application of Feldman-Cousins to Poisson signals

The most common use of Feldman-Cousins is for quoting limits on the size of a signal given a known background. For example, you are looking for dark matter particles. The expected background is b=4 events, and you observe N=6 events. What is the confidence interval on the signal rate s?

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

Traditional methods can sometimes give negative values for s when N<b, which is silly. Feldman-Cousins addresses this.

Feldman & Cousins' paper contains lookup tables to help you with this.

Feldman-Cousins lookup table for Poisson signals and backgrounds

To the right is a Feldman-Cousins lookup table at the 90% C.L. for a Poisson signal and background when the expected number of background events is 4.

We have to observe at least 8 events before the lower limit is non-zero. We'd then say that we exclude s=0 at the 90% C.L.

The 99% C.L. table shows that we get a non-zero lower limit when N is 10 or more.

N	Limit (b=4)
0	0.00,1.01
1	0.00,1.39
2	0.00,2.33
3	0.00,3.53
4	0.00,4.60
5	0.00,5.99
6	0.00,7.47
7	0.00,8.53
8	0.66,9.99
9	1.33,11.30
10	1.94,12.50

Limitations of Feldman-Cousins

Feldman-Cousins is probably the best recipe for producing Neyman confidence intervals. It deals with physical boundaries on parameters, never gives an empty confidence interval, and avoids the flip-flop problem.

Nonetheless, there are some obvious difficulties with Feldman-Cousins:

1) Constructing the confidence intervals is complicated, and usually has to be done numerically, or even with Monte Carlo.
2) Systematics are not easily incorporated into the procedure---you basically have to marginalize by Monte Carlo. A literature exists on how to handle this.
3) There's something peculiar I call the “small numbers paradox”

The small numbers paradox

Consider two hypothetical experiments to look for dark matter. One has an expected background of zero. The other expects 15 background events. Which is better?

Experiment #1: b=0
Observes 0 events

Experiment #2: b=15
Observes 0 events

The small numbers paradox

Consider two hypothetical experiments to look for dark matter. One has an expected background of zero. The other expects 15 background events. Which is better?

Experiment #1: $b=0$
Observes 0 events
F-C limit: $s<2.44$ at 90% C.L.

Experiment #2: $b=15$
Observes 0 events
F-C limit: $s<0.92$ at 90% C.L.

The Feldman-Cousins prescription says that the second experiment, which anticipated a large background that was not seen, gave a better limit.

Yet this makes no sense! The only way to get zero events when you expect 15 is a big fluctuation in the background. But background and signal are independent. How can a fluctuation in background result in a better limit on the signal?!?

Feldman & Cousins reply:

"The origin of these concerns lies in the natural tendency to want to interpret these results as the probability $P(\mu_t|x_0)$ of a hypothesis given data, rather than what they are really related to, namely the probability $P(x_0|\mu_t)$ of obtaining data given a hypothesis. It is the former that a scientist may want to know in order to make a decision, but the latter which classical confidence intervals relate to. As we discussed in Sec IIA, scientists may make Bayesian inferences of $P(\mu_t|x_0)$ based on experimental results combined with their personal, subjective prior probability distribution functions. It is thus incumbent on the experimenter to provide information that will assist in this assessment."

In spite of this, Feldman & Cousins maintain that you should be using frequentist confidence intervals, not Bayesian analyses.

A personal remark

My personal conclusion: the continuing difficulties with confidence intervals demonstrates that frequentist statistics is bullshit.

Be a Bayesian.

A helpful aid: stating the sensitivity

Feldman and Cousins recommend that in addition to the limit, you should quote the sensitivity (the average upper limit expected from your experiment). For example:

Experiment #1: b=0
Observes 0 events
F-C limit: s<2.44 at 90% C.L.
Sensitivity: 2.44

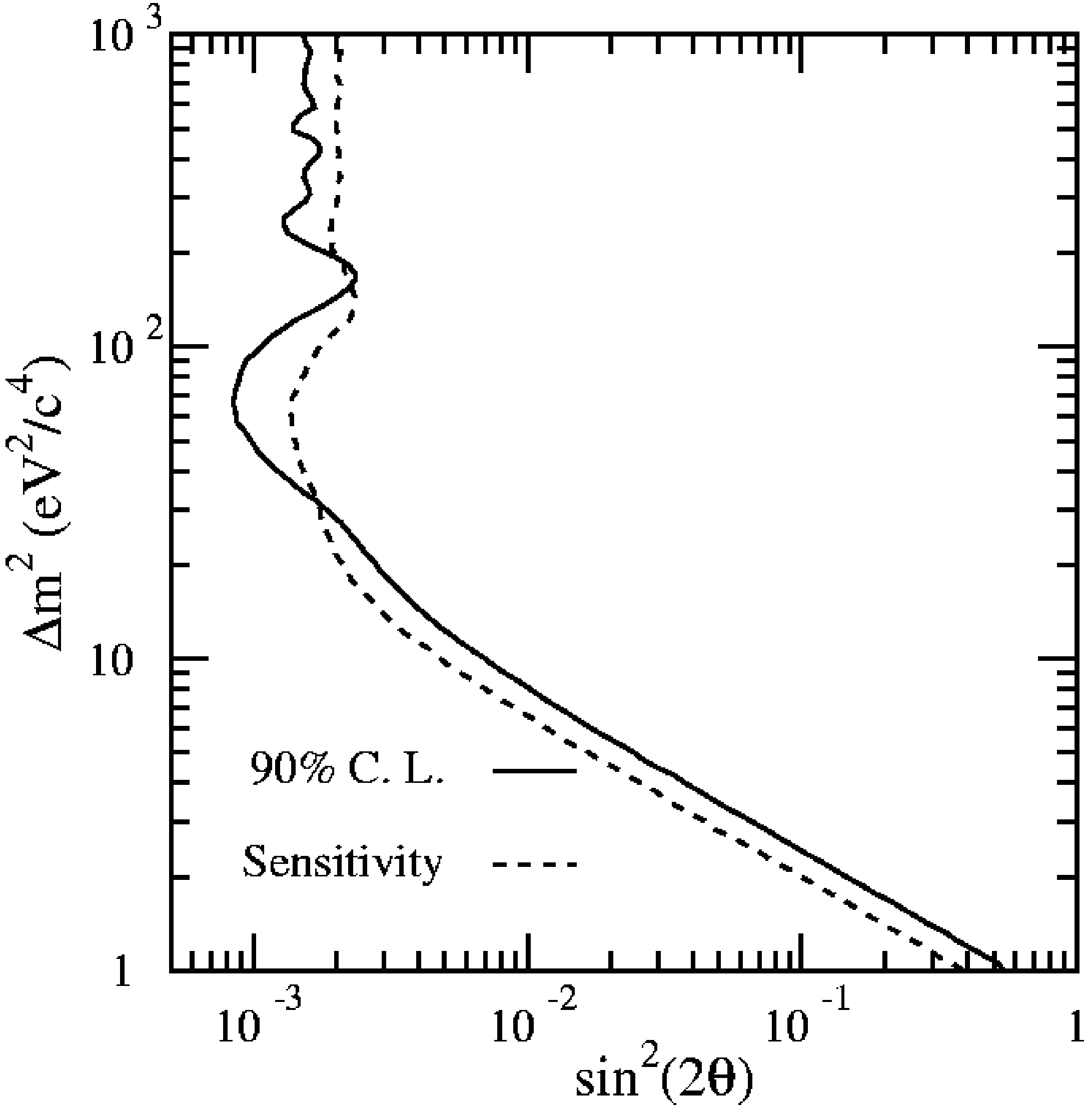
Experiment #2: b=15
Observes 0 events
F-C limit: s<0.92 at 90% C.L.
Sensitivity: 4.83

The fact that the limit for experiment #2 is a lot better than the sensitivity means that this is a “lucky” result, and the limit is much better than it has any right to be. It's a warning sign that the Bayesian limit would likely be very different (much worse).

2D Feldman-Cousins contours

This all works in multiple dimensions as well. For example, here's a 2D confidence region from Feldman and Cousins for a neutrino oscillation experiment.

Notice how they include the sensitivity curve to demonstrate that their limit is in accord with what they expected to get.



Confidence intervals near physical boundaries

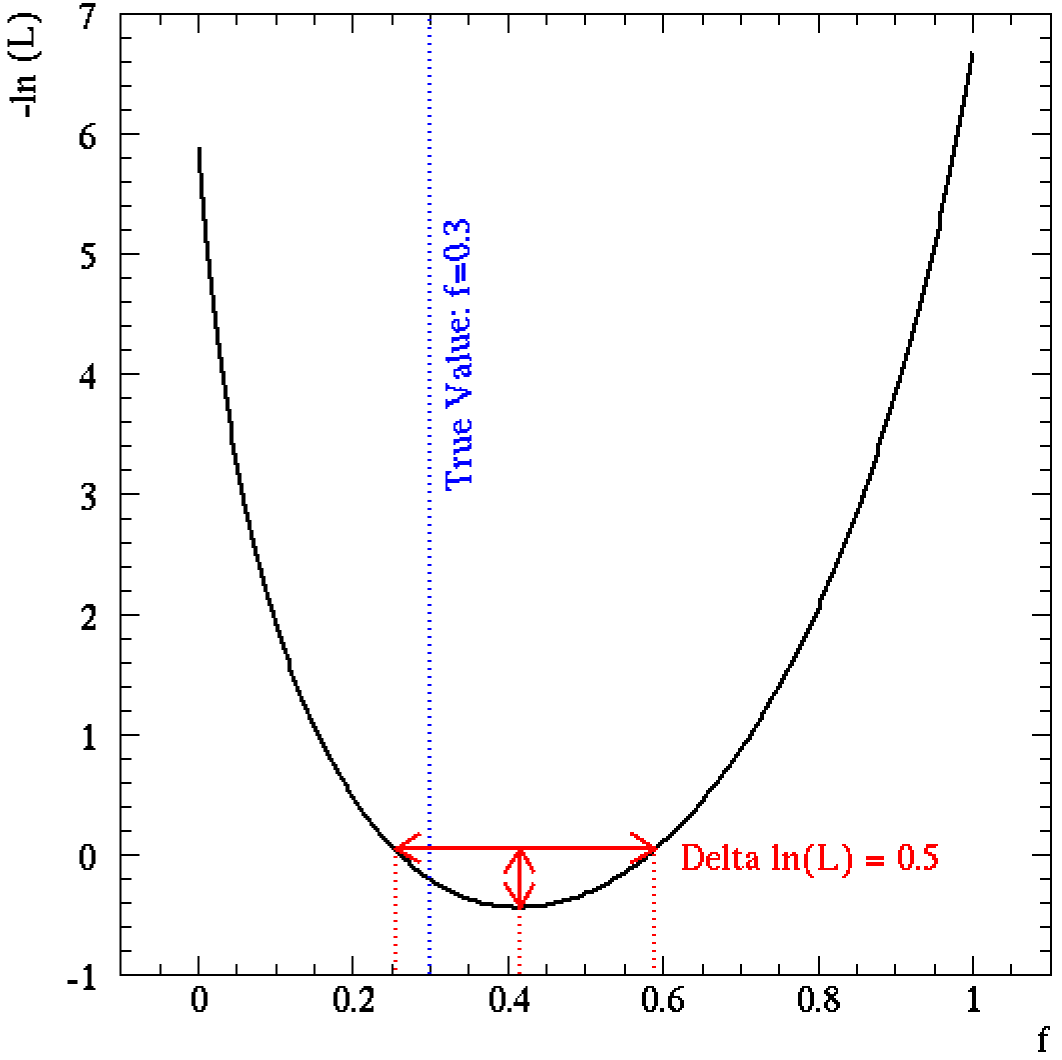
Confidence intervals can get you in trouble near physical boundaries. For example, what do you do if you try to measure the mass of an object and get a negative value (e.g. m=-0.5±0.6 g)?

A quandry: you feel silly reporting a nonsensical value. If you truncate the interval from (-1.1,0.1) to (0,0.1), you get a misleading error bar. It's common to “shift” the whole result up until the central value is zero, and to report something like “m<0.6g”. But this will bias the result if your measurement is averaged with other measurements.

Feldman-Cousins and Bayesian methods both provide one way out---the FC confidence belt is constructed in such a way as to account for physical boundaries, while the Bayesian prior will easily exclude non-physical outcomes.

If your analysis DOES return a confidence interval overlapping the non-physical region, it's important to result the naked, unadulterated result so that it won't bias future meta-analyses.

The Δ ln(L) rule



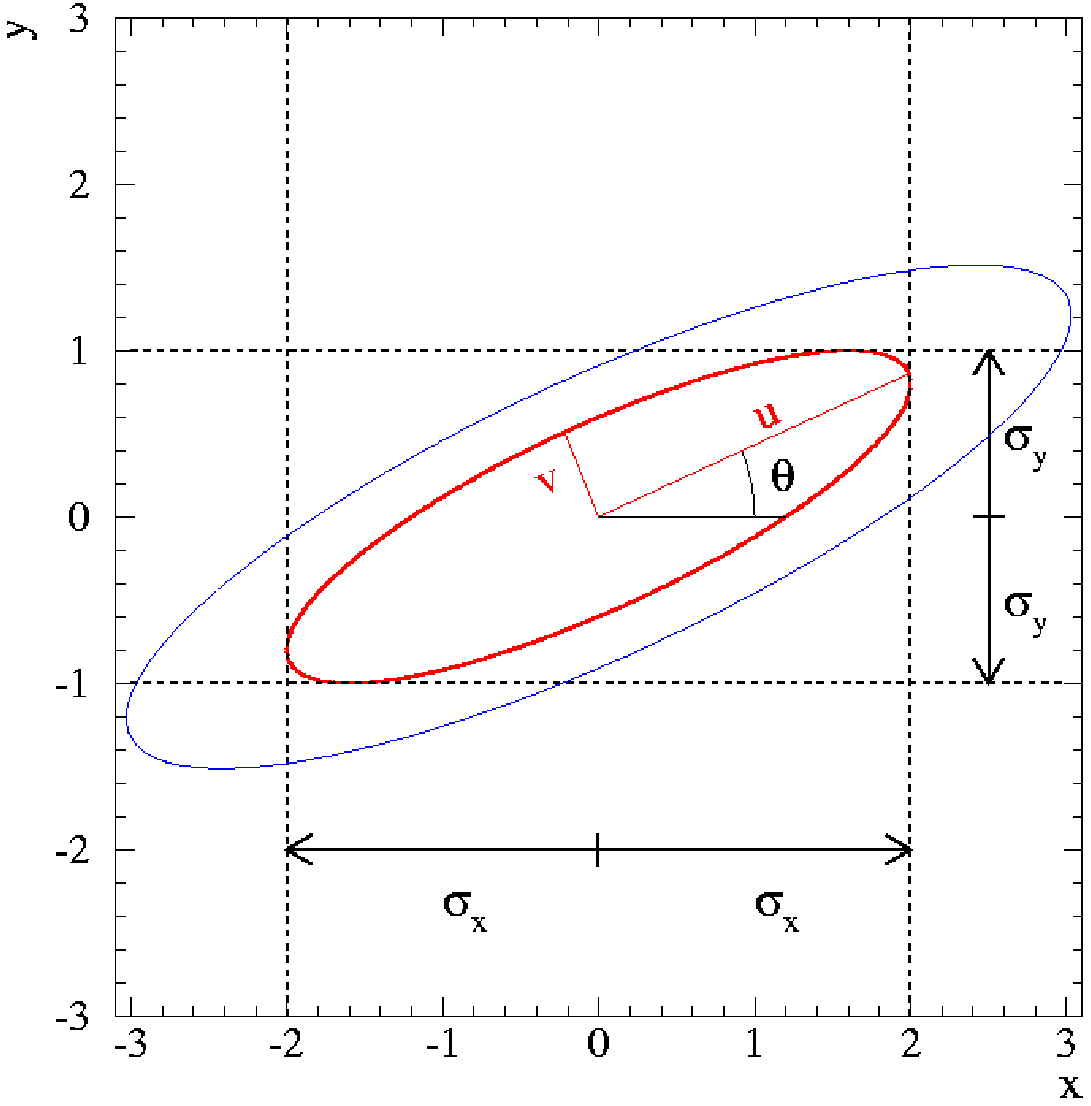
It is not trivial to construct proper frequentist confidence intervals. Most often an approximation is used: the confidence interval for a single parameter is defined as the range in which $\ln(L_{max})$-$\ln(L)<0.5$

This is only an approximation, and does not give exactly the right coverage when N is small.

More generally, if you have d free parameters, then the quantity $\omega = $ "$\Delta\chi^2$" $= 2[\ln(L_{max})-\ln(L)]$ approximates a χ^2 with d degrees of freedom.

Coverage and limits will resemble Feldman-Cousins except when N is small, or near a physical boundary.

Multi-dimensional confidence intervals



Don't forget that the value of Δ ln L you use to draw the contour depends on the dimension of the plot.

Red ellipse: contour with Δ ln L= -1/2 ($\Delta\chi^2=1$). Gives correct 1D limits on a single parameter.

Blue ellipse: contour contains 68% of probability content in 2D. Δ ln L= -1.15 ($\Delta\chi^2=2.30$).

The contour value is based on the probability content of a χ^2 with d degrees of freedom (see NumRec Sec 15.6)