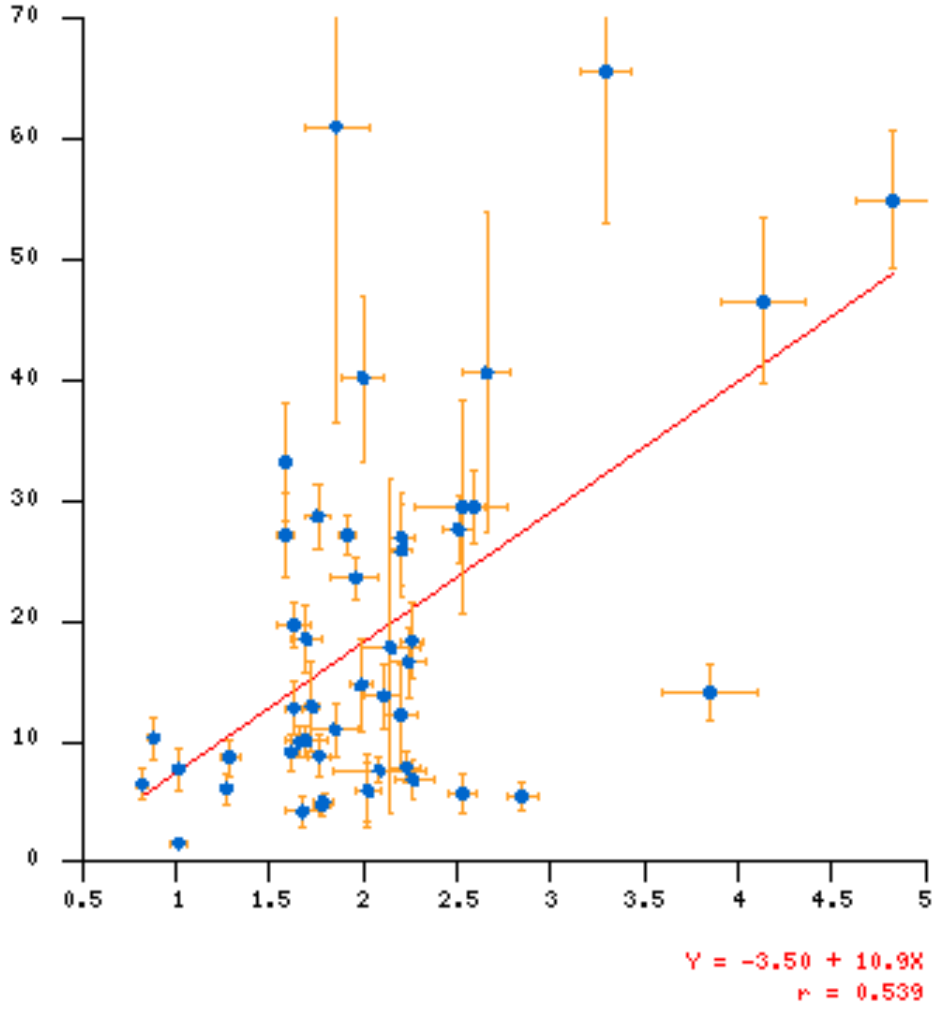
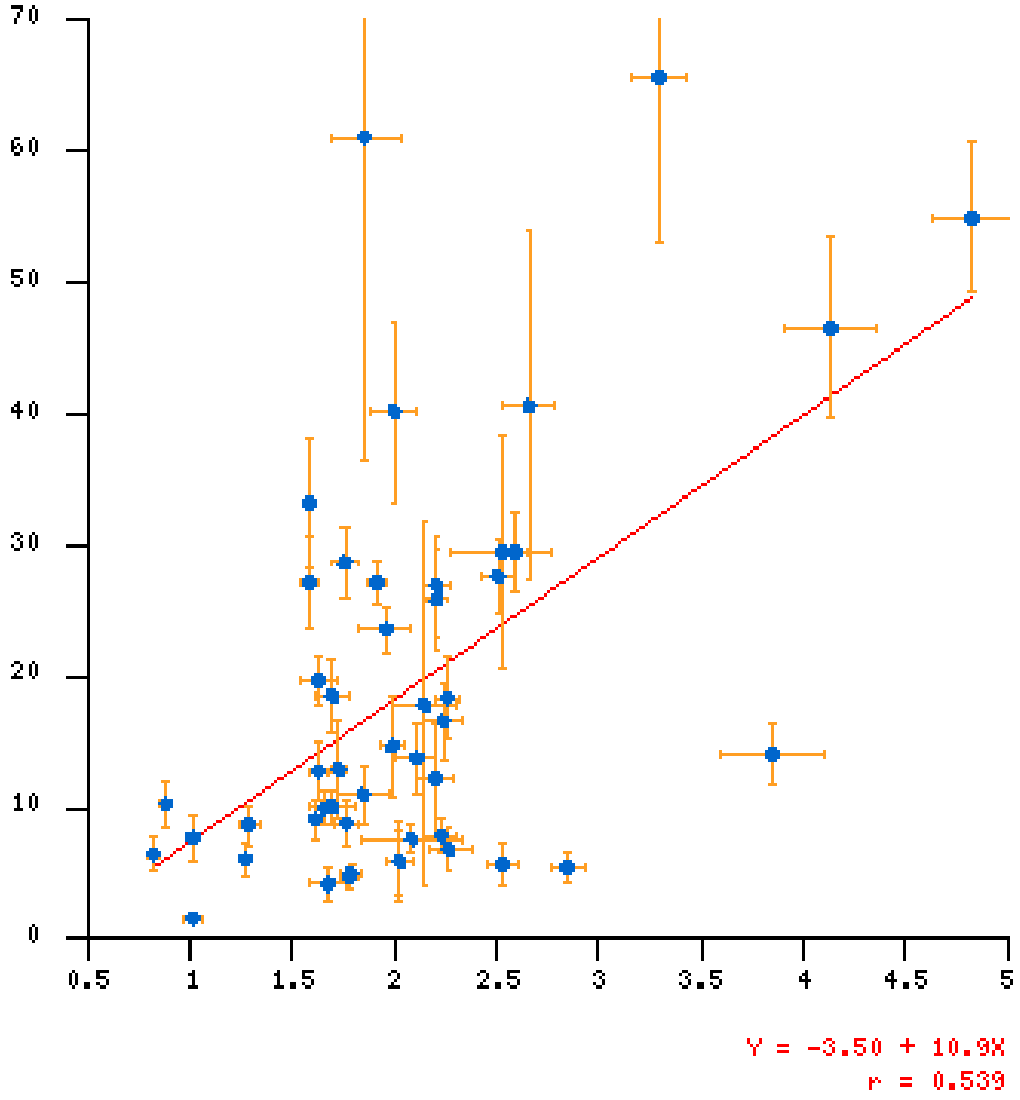
Error bars and confidence intervals

Lecture #3

Scott Oser
TRISEP 2024



What is an error bar?



Someone hands you a plot like this. What do the error bars indicate?

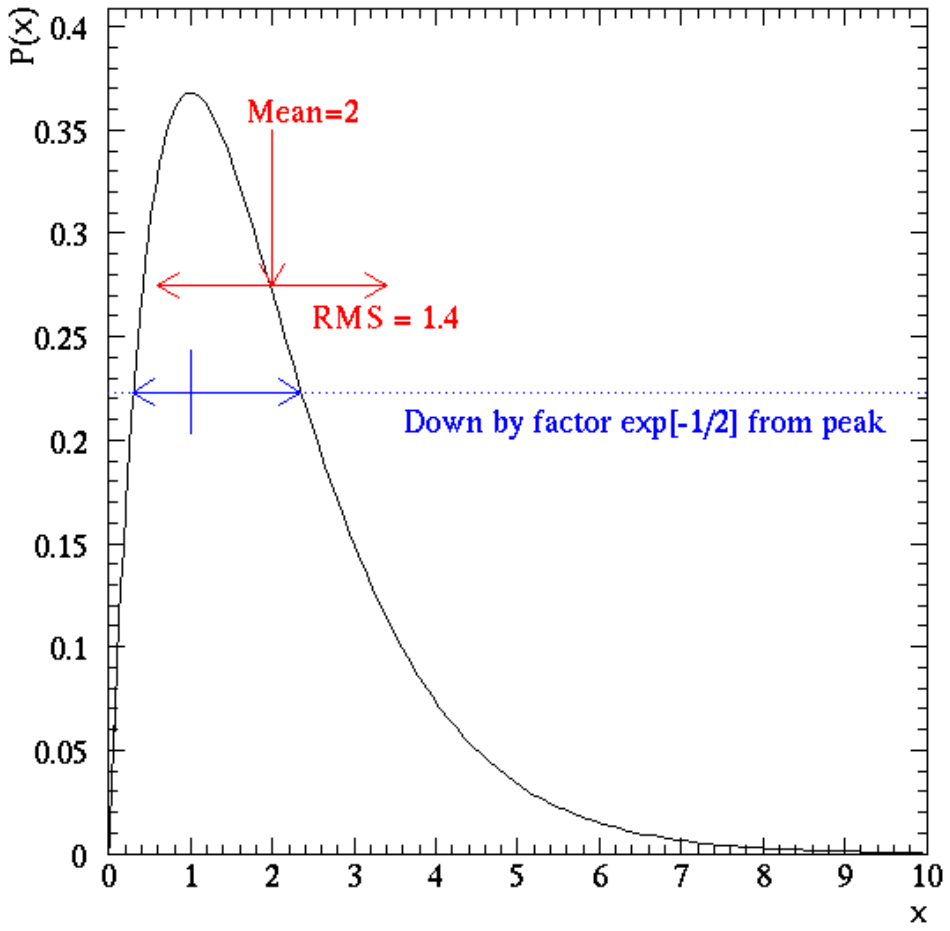
Answer: you can never be sure, unless it's specified!

Most common: vertical error bars indicate "±1σ" uncertainties.

Horizontal error bars can indicate uncertainty on X coordinate, or can indicate binning.

Correlations unknown!

Relation of an error bar to PDF shape



The error bar on a plot is most often meant to represent the $\pm 1\sigma$ uncertainty on a data point. Bayesians and frequentists will disagree on what that means.

If data is distributed normally around "true value", it's clear what is intended:

$$\exp[-(x-\mu)^2/2\sigma^2].$$

But for asymmetric distributions, different things are sometimes meant ...

An error bar is a shorthand approximation to a PDF!

In an ideal Bayesian universe, error bars don't exist. Instead, everyone will use the full prior PDF and the data to calculate the posterior PDF, and then report the shape of that PDF (preferably as a graph or table).

An error bar is really a shorthand way to parametrize a PDF. Most often this means pretending the PDF is Gaussian and reporting its mean and RMS.

Many sins with error bars come from assuming Gaussian distributions when there aren't any.

The error propagation equation

Let *f(x,y)* be a function of two variables, and assume that the uncertainties on *x* and *y* are known and "small". Then:

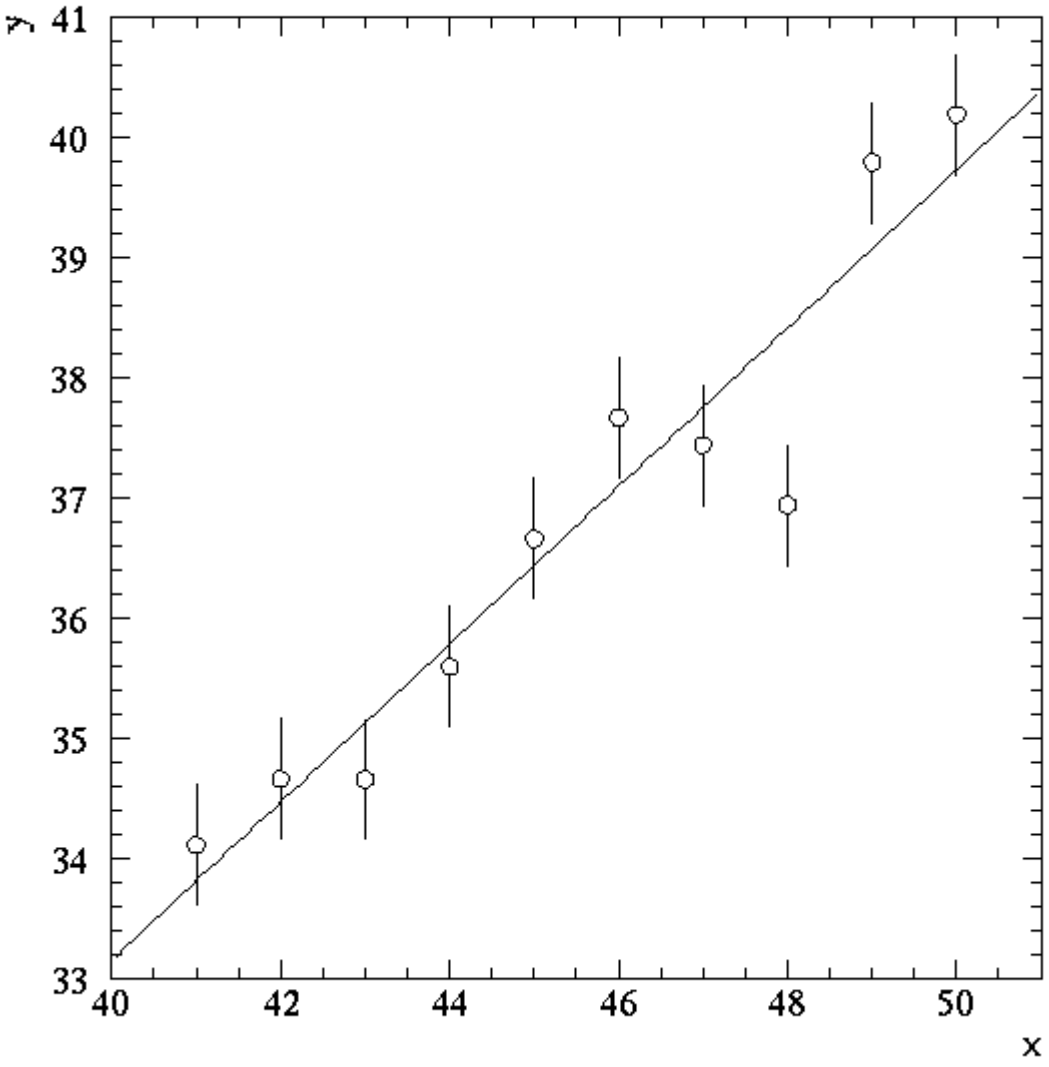
$$\sigma_f^2 = \left(\frac{df}{dx}\right)^2 \sigma_x^2 + \left(\frac{df}{dy}\right)^2 \sigma_y^2 + 2\left(\frac{df}{dx}\right)\left(\frac{df}{dy}\right)\rho\sigma_x\sigma_y$$

The assumptions underlying the error propagation equation are:

- covariances are known
- *f* is an approximately linear function of *x* and *y* over the span of *x*±*dx* or *y*±*dy.*

The most common mistake in the world: ignoring the third term. Intro courses ignore its existence entirely!

Example: interpolating a straight line fit



Straight line fit y=mx+b

Reported values from a standard fitting package:

m = 0.658 ± 0.056
b = 6.81 ± 2.57

Estimate the value and uncertainty of *y* when *x*=45.5:

$y = 0.658 * 45.5 + 6.81 = 36.75$

$dy = \sqrt{2.57^2 + (45.5 \cdot .056)^2} = 3.62$

UGH! NONSENSE!

Example: straight line fit, done correctly

Here's the correct way to estimate y at x=45.5. First, I find a better fitter, which reports the actual covariance matrix of the fit:

m = 0.0658 + .056
b = 6.81 + 2.57
ρ = -0.9981

$$dy = \sqrt{2.57^2 + (0.056 \cdot 45.5)^2 + 2(-0.9981)(0.056 \cdot 45.5)(2.57)} = 0.16$$

(Since the uncertainty on each individual data point was 0.5, and the fitting procedure effectively averages out their fluctuations, then we expect that we could predict the value of y in the meat of the distribution to better than 0.5.)

Food for thought: if the correlations matter so much, why don't most fitting programs report them routinely???

Generalizing the error propagation equation

If we have N functions of M variables, we can calculate their covariance by:

$$\operatorname{cov}\left(f_k, f_l\right) = \sum_i \sum_j \left(\frac{\partial f_k}{\partial x_i}\right)\left(\frac{\partial f_l}{\partial x_j}\right)\operatorname{cov}\left(x_i, x_j\right)$$

We can write this compactly in matrix form as:

$$G = G_{ki} = \left(\frac{\partial f_k}{\partial x_i}\right) \qquad (\text{an N x M matrix})$$

$$V_f = G \cdot V_x \cdot G^T$$

Averaging correlated measurements: example

Consider the following example, adapted from Glen Cowan's book*:

We measure an object's length with two rulers. Both are calibrated to be accurate at $T=T_0$, but otherwise have a temperature dependency: true length y is related to measured length by:

$$y_i = L_i + c_i\left(T - T_0\right)$$

We assume that we know the c_i and the uncertainties, which are Gaussian. We measure L_1, L_2, and T, and so calculate the object's true length y.

$$y_i = L_i + c_i\left(T - T_0\right)$$

We wish to combine the measurements from the two rulers to get our best estimate of the true length of the object.

* “Statistical Data Analysis”, by Glen Cowan (Oxford, 1998)

Averaging correlated measurements: example

We start by forming the covariance matrix of the two measurements:

$$y_i = L_i + c_i\left(T - T_0\right) \qquad \sigma_i^2 = \sigma_L^2 + c_i^2\,\sigma_T^2$$

$$\mathrm{cov}\left(y_1, y_2\right) = c_1\,c_2\,\sigma_T^2$$

We use the method previously described to calculate the weighted average for the following parameters:

c_1 = 0.1	L_1=2.0 ± 0.1	y_1=1.80 ± 0.22	T_0=25
c_2 = 0.2	L_2=2.3 ± 0.1	y_2=1.90 ± 0.41	T = 23 ± 2

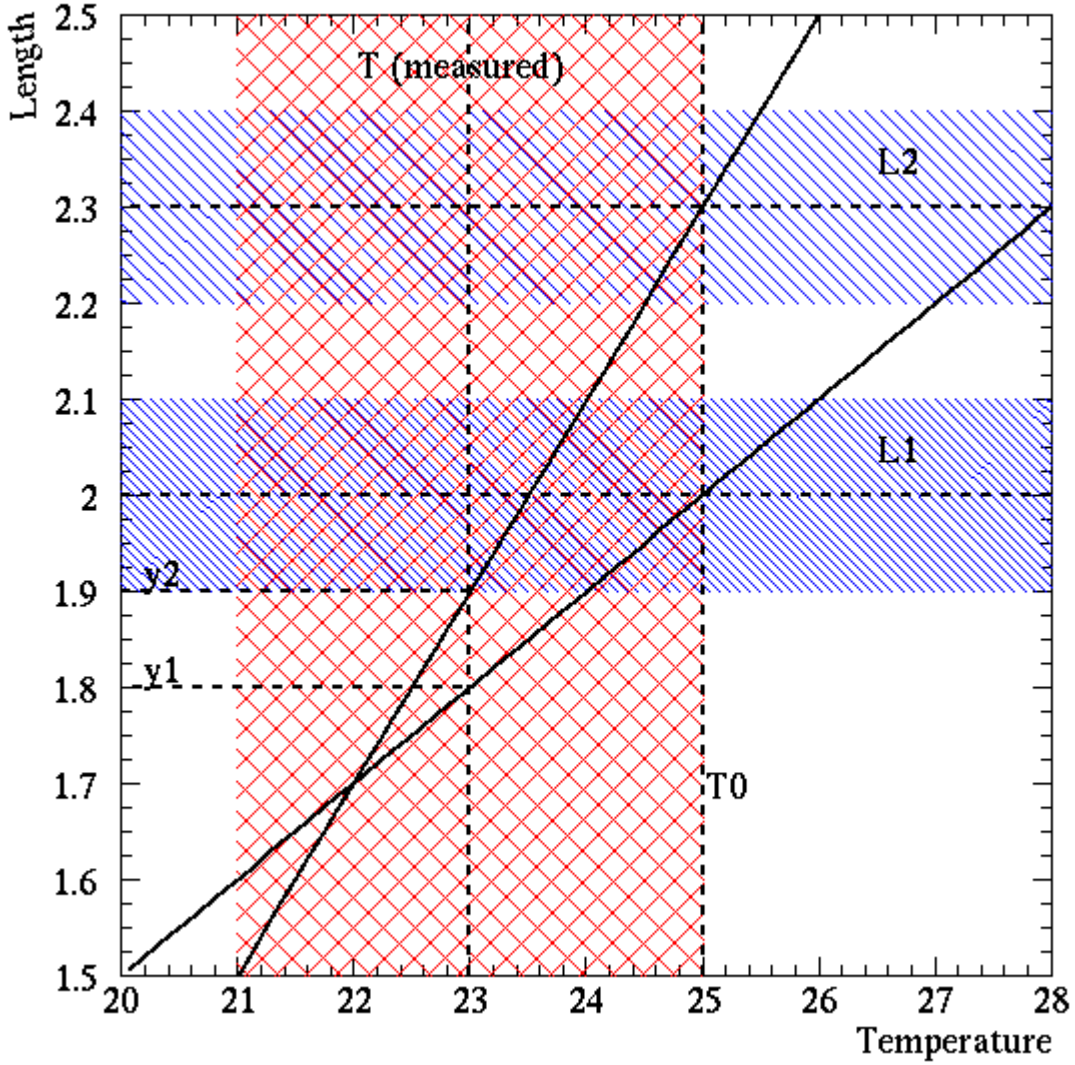
Using the error propagation equations, we get for the weighted average:

$$y_{true} = 1.75 \pm 0.19$$

WEIRD: the weighted average is smaller than either measurement! What's going on??

A case where the data constrained a systematic

$$y_i = L_i + c_i\left(T - T_0\right)$$



c_1 = 0.1
L_1=2.0 ± 0.1
y_1=1.80 ± 0.22
T_0=25

c_2 = 0.2
L_2=2.3 ± 0.1
y_2=1.90±0.41
T = 23 ± 2

The intersection of the two lines in this example from earlier provides a better estimate of the true temperature than that provided from the external calibration of 23 ± 2.

Constraint terms in the likelihood

Working in Bayesian language, the posterior PDF is given by

$$P\left(\theta,\alpha\,|\,D,I\right)\propto P\left(\theta\,|\,I\right)P\left(\alpha\,|\,I\right)P\left(D\,|\,\theta,\alpha,I\right)$$

We saw previously that the ML estimator is same thing as the mode of the Bayesian posterior PDF assuming a flat prior on θ. In that case we maximized ln L(θ)=ln P(D|θ,I), and use the shape of ln L to determine the confidence interval on θ.

This easily generalizes to include systematics by considering the nuisance parameters α to simply be more parameters we're trying to estimate:

$$\ln L\left(\theta,\alpha\right)=\ln L\left(\theta\,|\,D,\alpha\right)+\ln P\left(\alpha\right)$$

The first term is the regular log likelihood---a function of θ, with α considered to be a fixed parameter. The second term is what we call the constraint term---basically it's the prior on α.

Application of constraint terms in likelihood

Remember the problem in which we measured an object using two rulers with different temperature dependencies?

$$y = L_i + c_i\left(T - T_0\right)$$

c_1 = 0.1	L_1=2.0 ± 0.1	y_1=1.80 ± 0.22	T_0=25
c_2 = 0.2	L_2=2.3 ± 0.1	y_2=1.90 ± 0.41	T = 23 ± 2

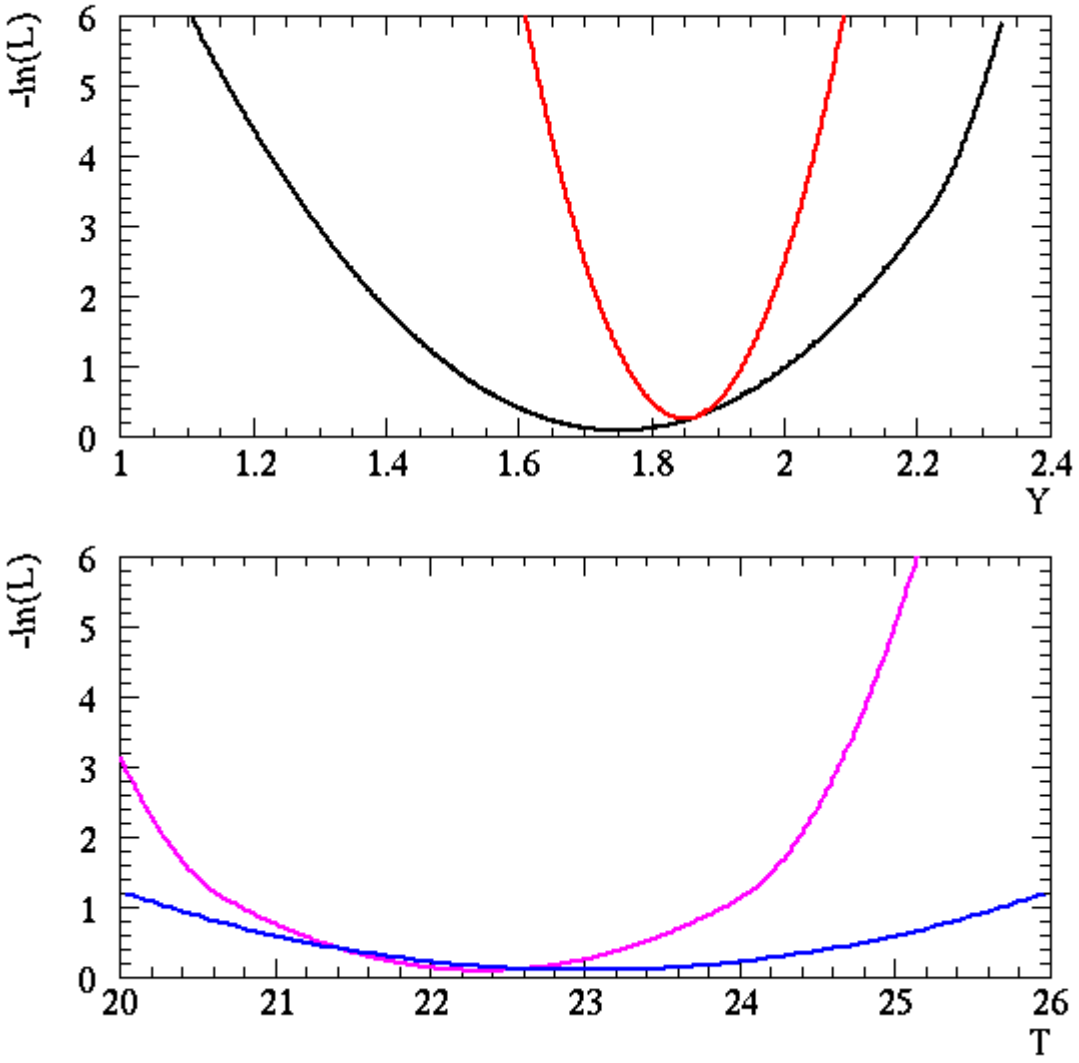
$$\ln L\left(\theta,\alpha\right) = \ln L\left(\theta \mid D,\alpha\right) + \ln P\left(\alpha\right)$$

$$-\ln L\left(y,T\right) = \frac{1}{2}\sum_{i=1}^{2}\left(\frac{y - L_i - c_i\left(T - T_0\right)}{\sigma_L}\right)^2 + \frac{1}{2}\left(\frac{T-23}{2}\right)^2$$

The first term of the likelihood is the usual likelihood containing "statistical errors" on the L_i, with T considered fixed. The second is the constraint term (think: "prior on T"). The joint likelihood is a function of the two unknowns y and T.

Procedure: minimize over T to get shape of likelihood as function of y.

Constraint terms in likelihood: results



Top plot is shape of likelihood as function of y, after marginalizing over T:

Red: T fixed (stat error only)
Black: after minimizing -ln(L) as function of T at each y
1σ range: same as covariance matrix approach

Blue: “a priori” constraint on T (23±2).
Magenta: shape of likelihood as a function of T, after marginalizing over y.

A simple recipe that usually will work

1) Build a quantitative model of how your likelihood function depends on the nuisance parameters.

2) Form a joint negative log likelihood that includes both terms for the data vs. model and for the prior on the nuisance parameter.

3) Treat the joint likelihood as a multidimensional function of both physics parameters and nuisance parameters, treating these equally.

4) Minimize the likelihood with respect to all parameters to get the best-fit.

5) The error matrix for all parameters is given by inverting the matrix of partial derivatives with respect to all parameters:

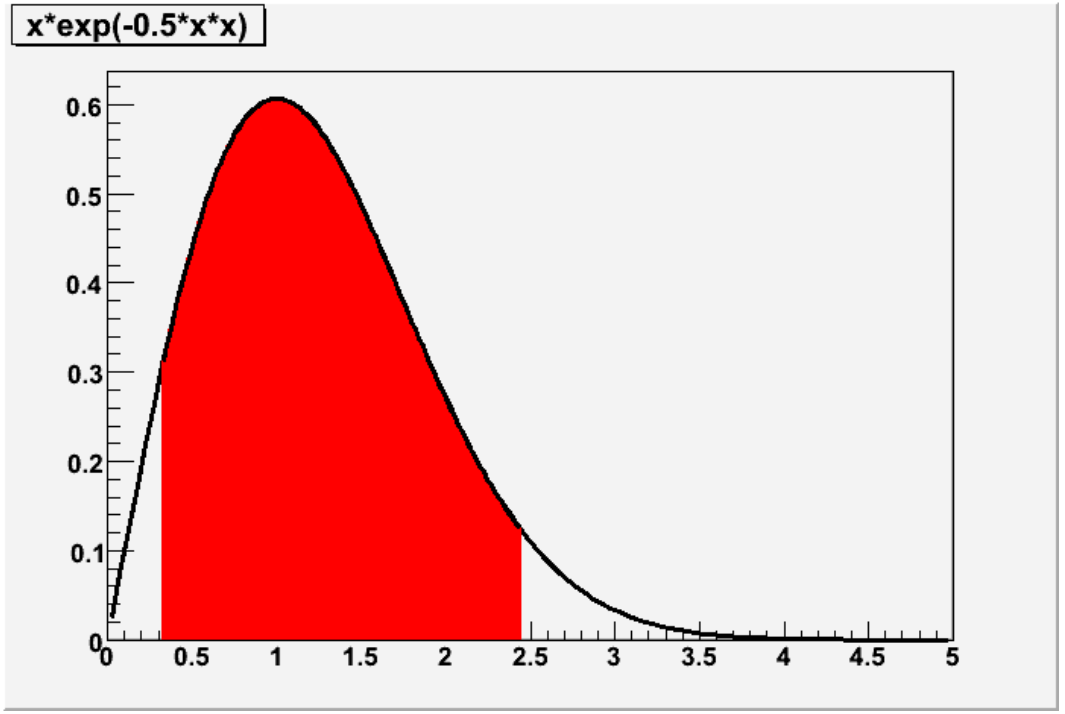
$$\mathbf{V} = -\left(\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right)^{-1}$$

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 .

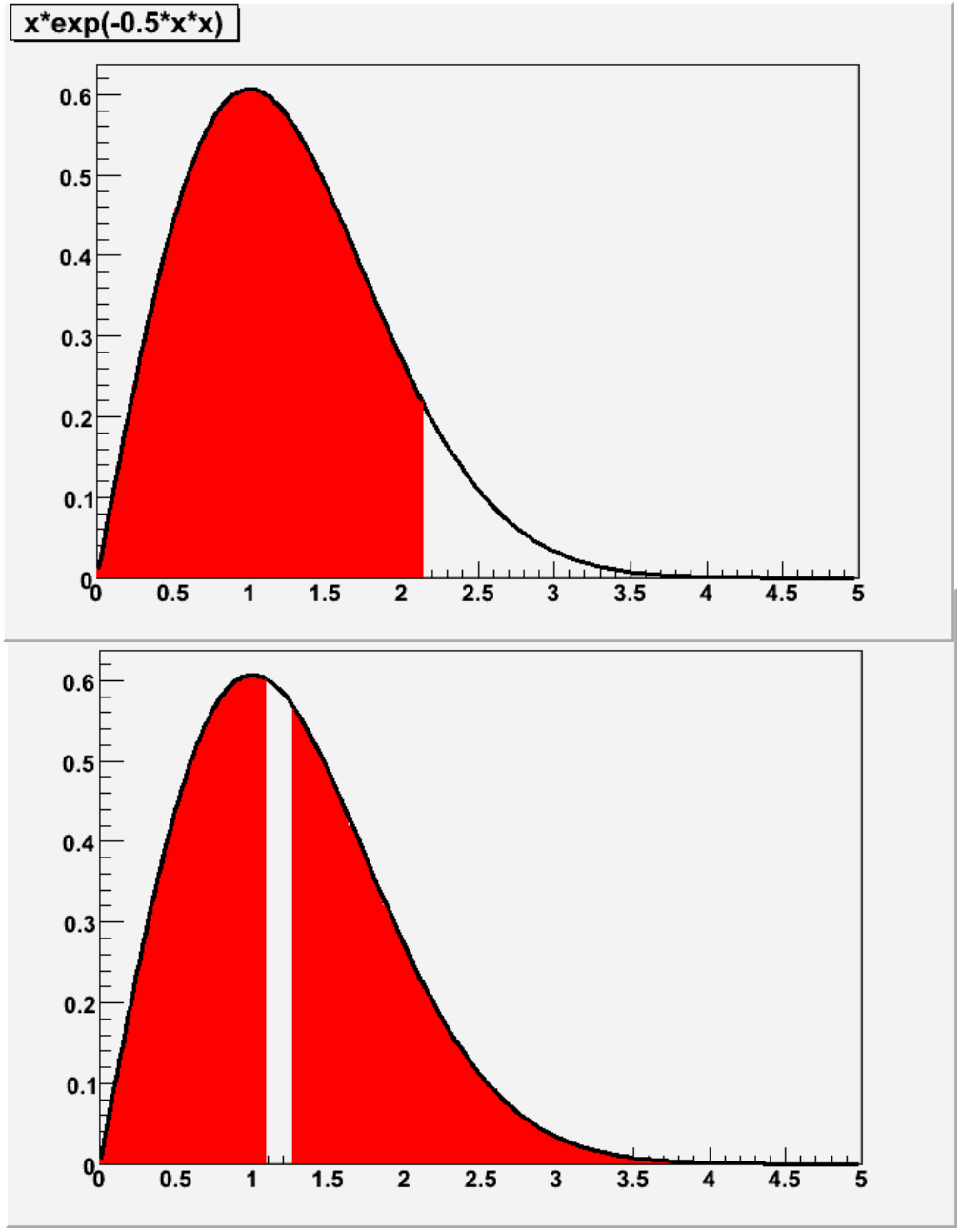
The red area contains 90% of the probability content---the Bayesian credible region is (0.32,2.45)



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

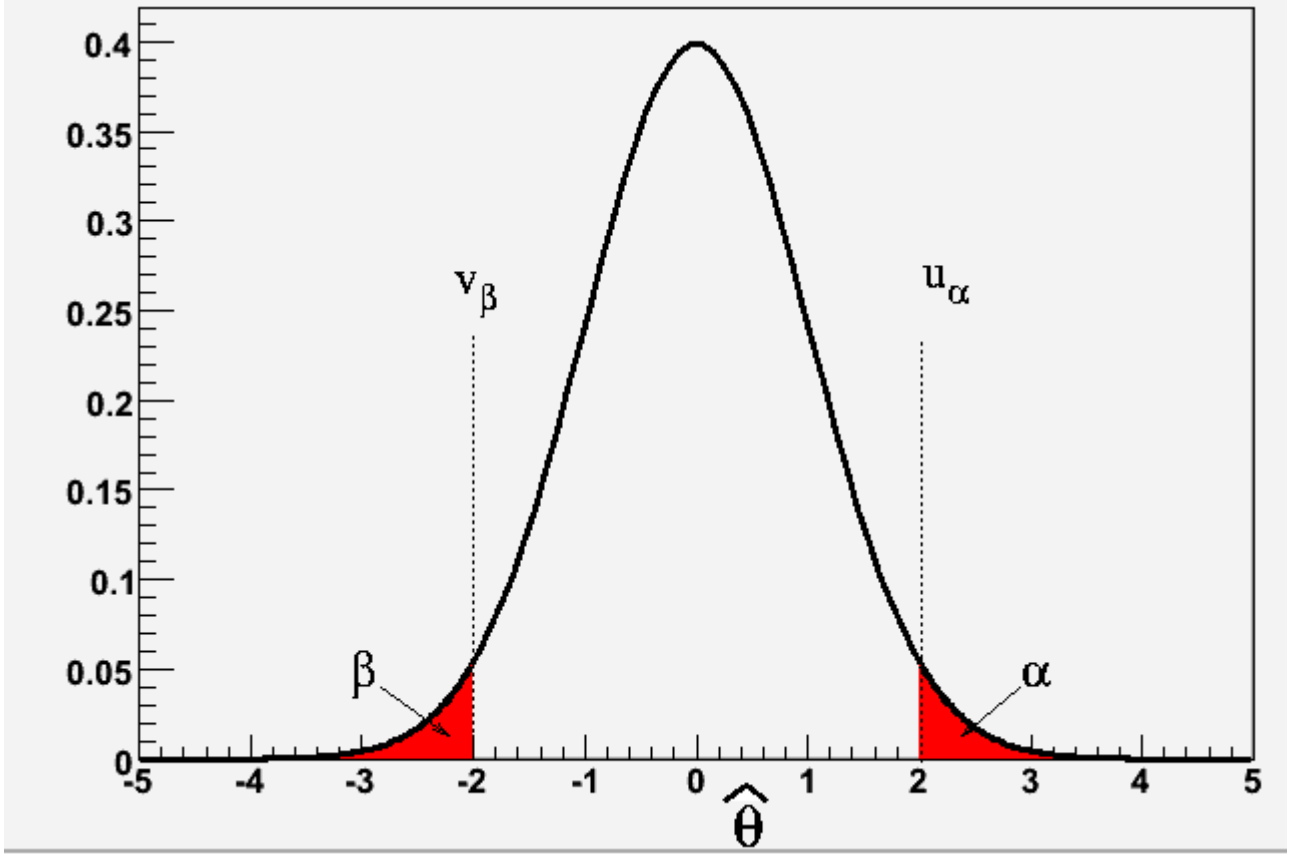
A frequentist confidence interval is a different beast. While a Bayesian credible region is based on the probability that the true parameter lies in the specified, the frequentist interval really refers to the probability of getting the observed data.

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 $\hat{a}$ of that parameter: $P(\hat{a}|a)$.
2) Using some procedure, define the interval in $\hat{a}$ 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.

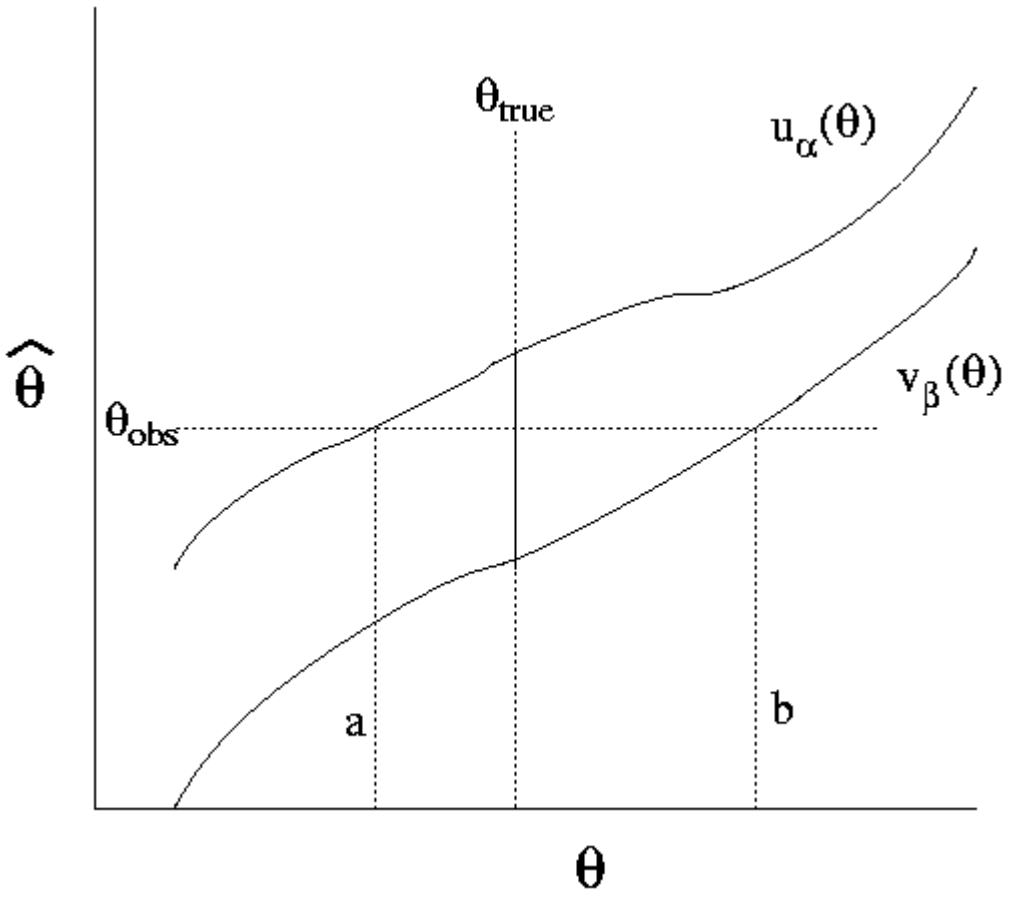


Starting point is the PDF for the estimator, for a fixed value of the parameter.

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

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

Confidence interval construction



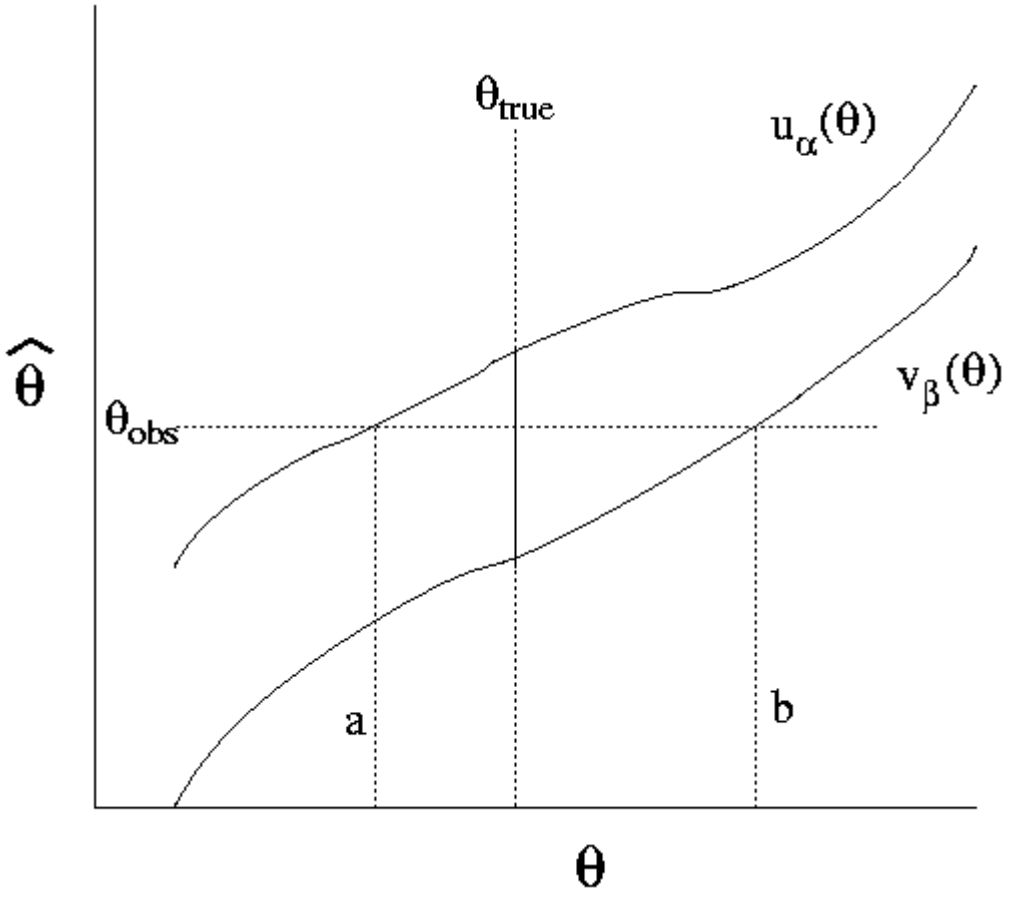
The confidence band is constructed so that the average probability of θ_{true} lying in the confidence interval is $1-\alpha-\beta$.

This should not be interpreted to mean that there is, for example, a 90% chance that the true value of θ is in the interval (a,b) on any given trial.

Rather, it means that if you ran the experiment 1000 times, and produced 1000 confidence intervals, then in 90% of these hypothetical experiments (a,b) would contain the true value.

Obviously a very roundabout way of speaking ...

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

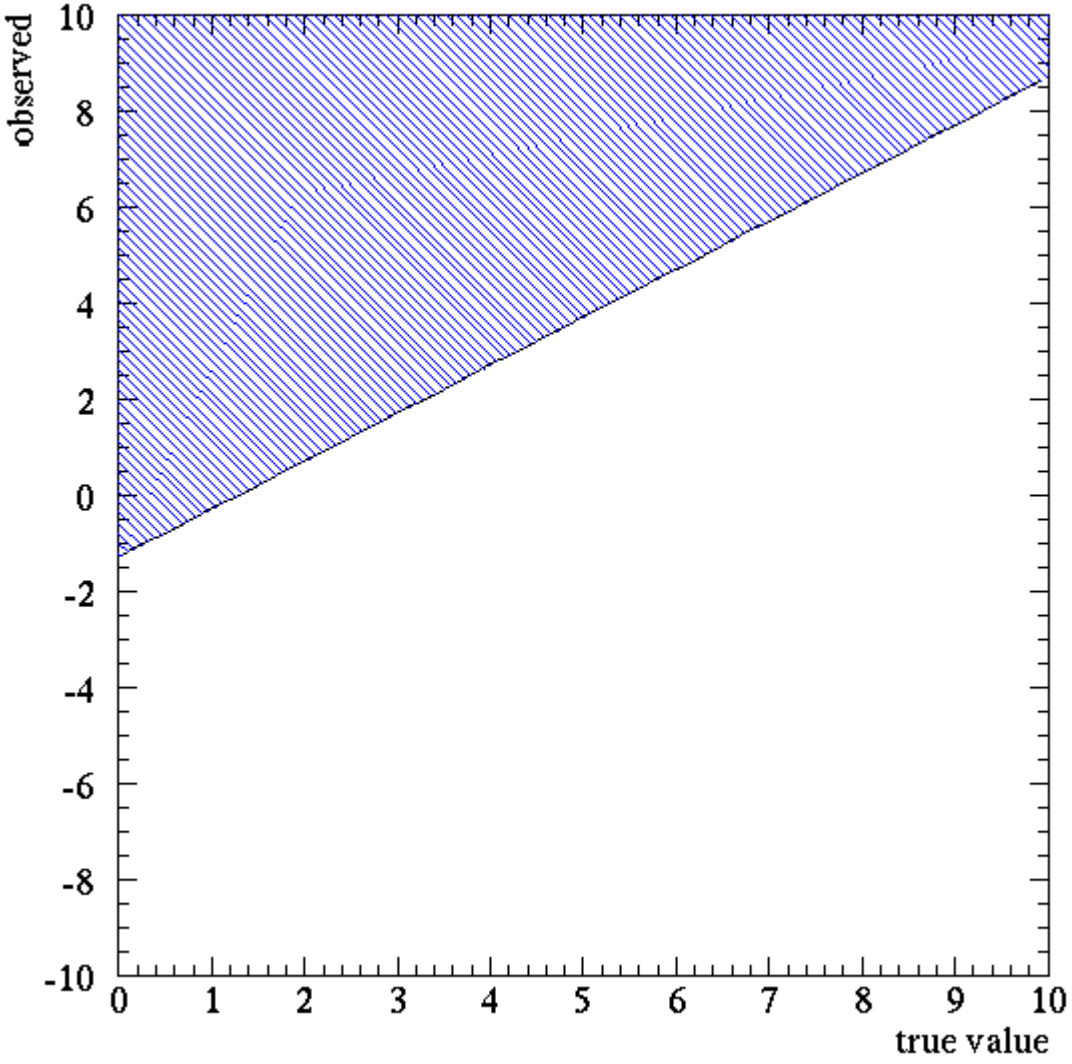
$$\mu-1.28<\hat{\mu}<\infty$$

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.

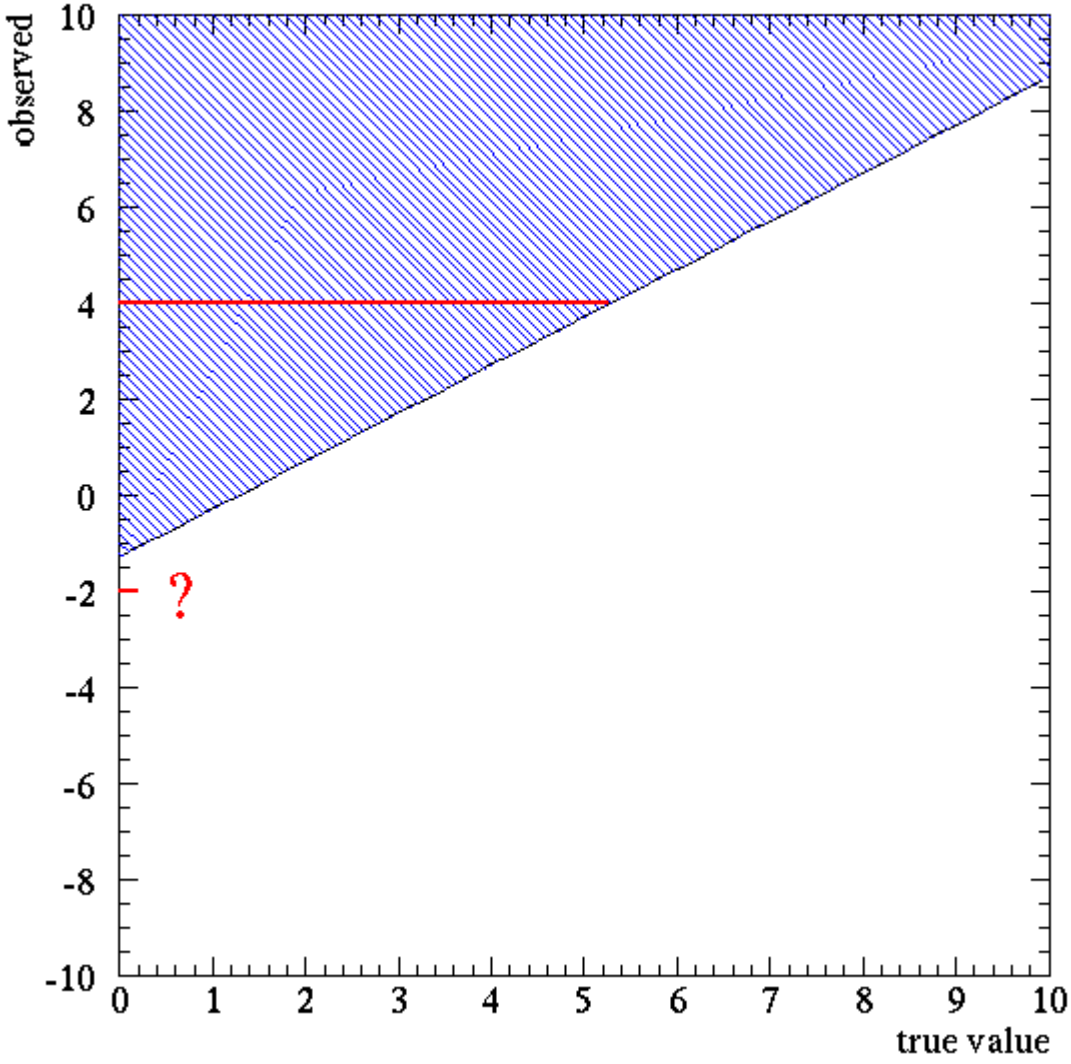


One sided confidence belt

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 $(-\infty, -0.72)$. 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.

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:

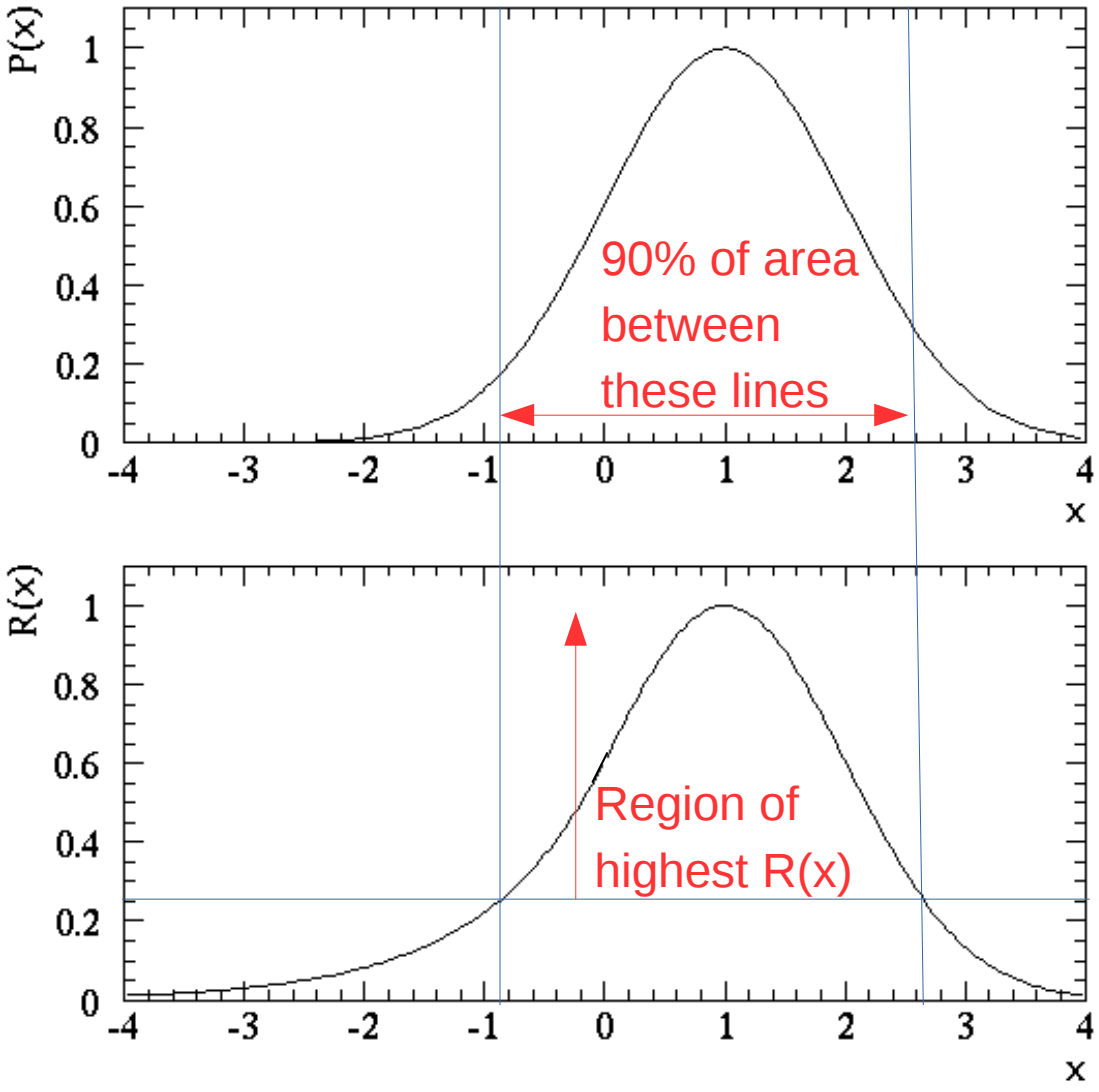
$$R = \frac{P(x|\mu)}{P(x|\mu_{best})}$$

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

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

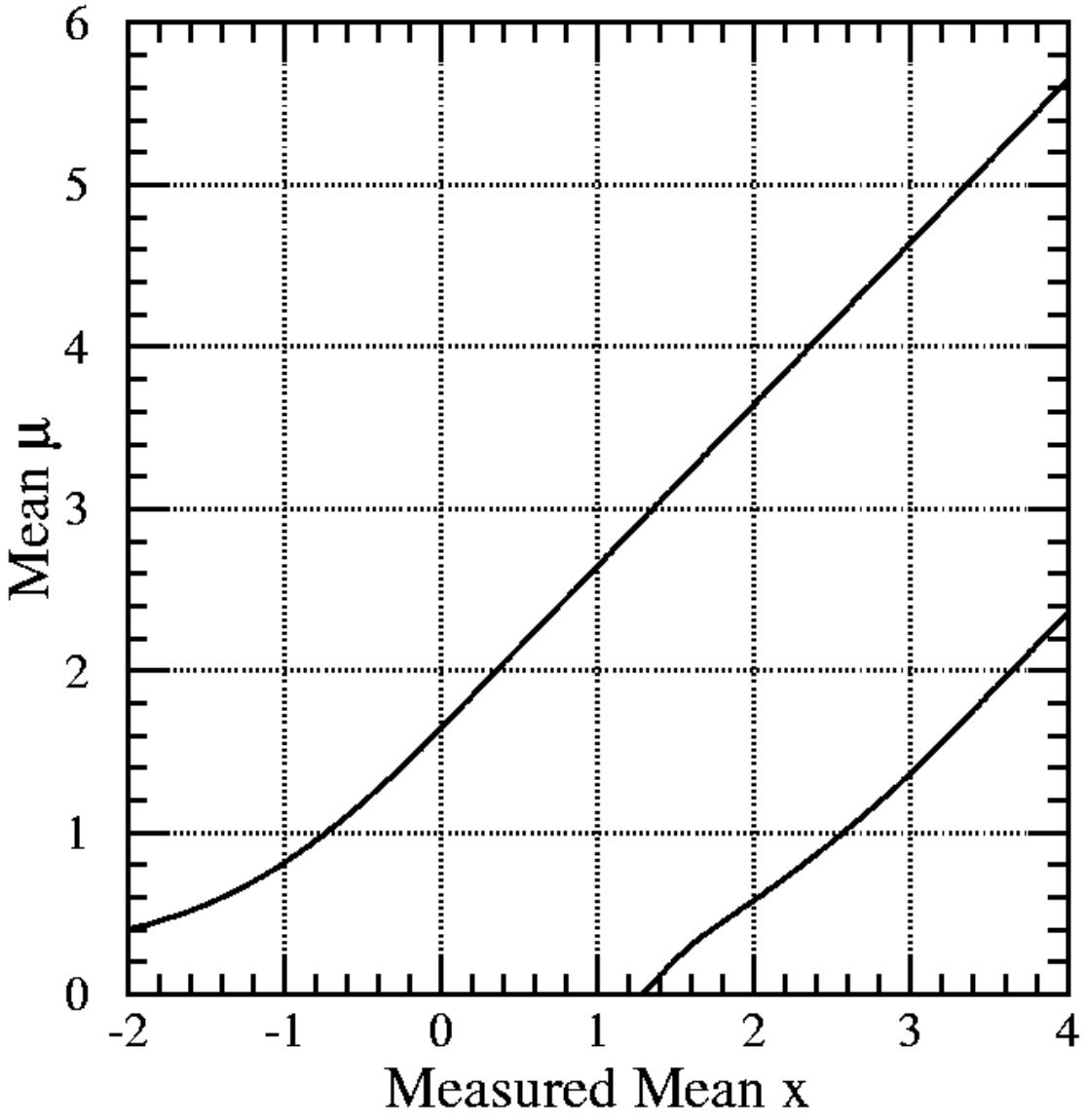
$$R = \frac{P(x|\mu)}{P(x|\mu_{best})} = \frac{\exp\left[-\frac{1}{2}(x-\mu)^2\right]}{\exp\left[-\frac{1}{2}x^2\right]} \quad \text{if } x \leq 0$$

Application of Feldman-Cousins to Gaussian with physical limit



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.

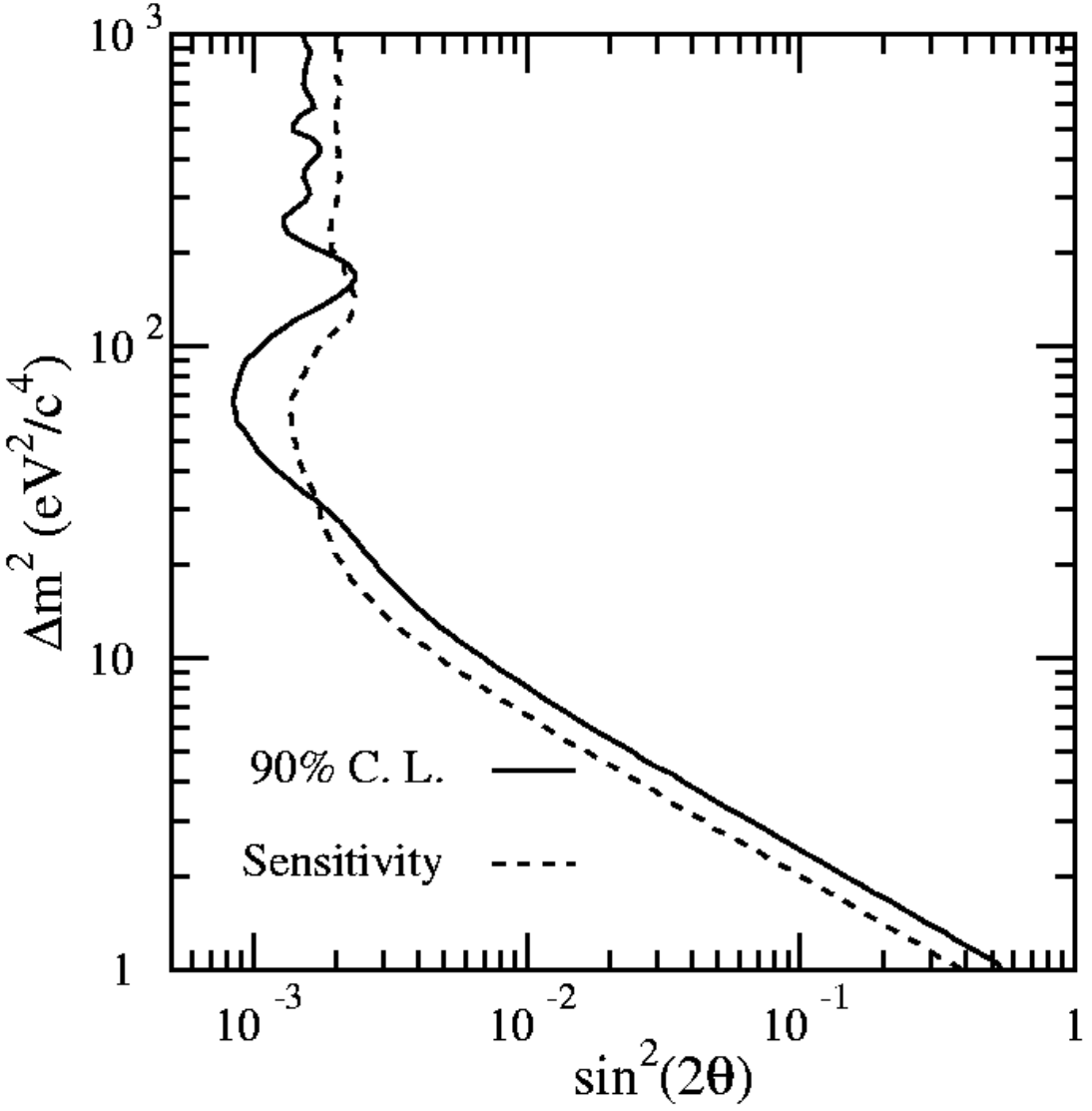
Nonetheless, there are some serious difficulties with Feldman-Cousins:

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) The confidence intervals, while no longer empty, still wind up being misleadingly small in the case of statistical fluctuations. In cases where limit is much better than sensitivity, the limit should not be trusted, by Feldman & Cousins' own admission.

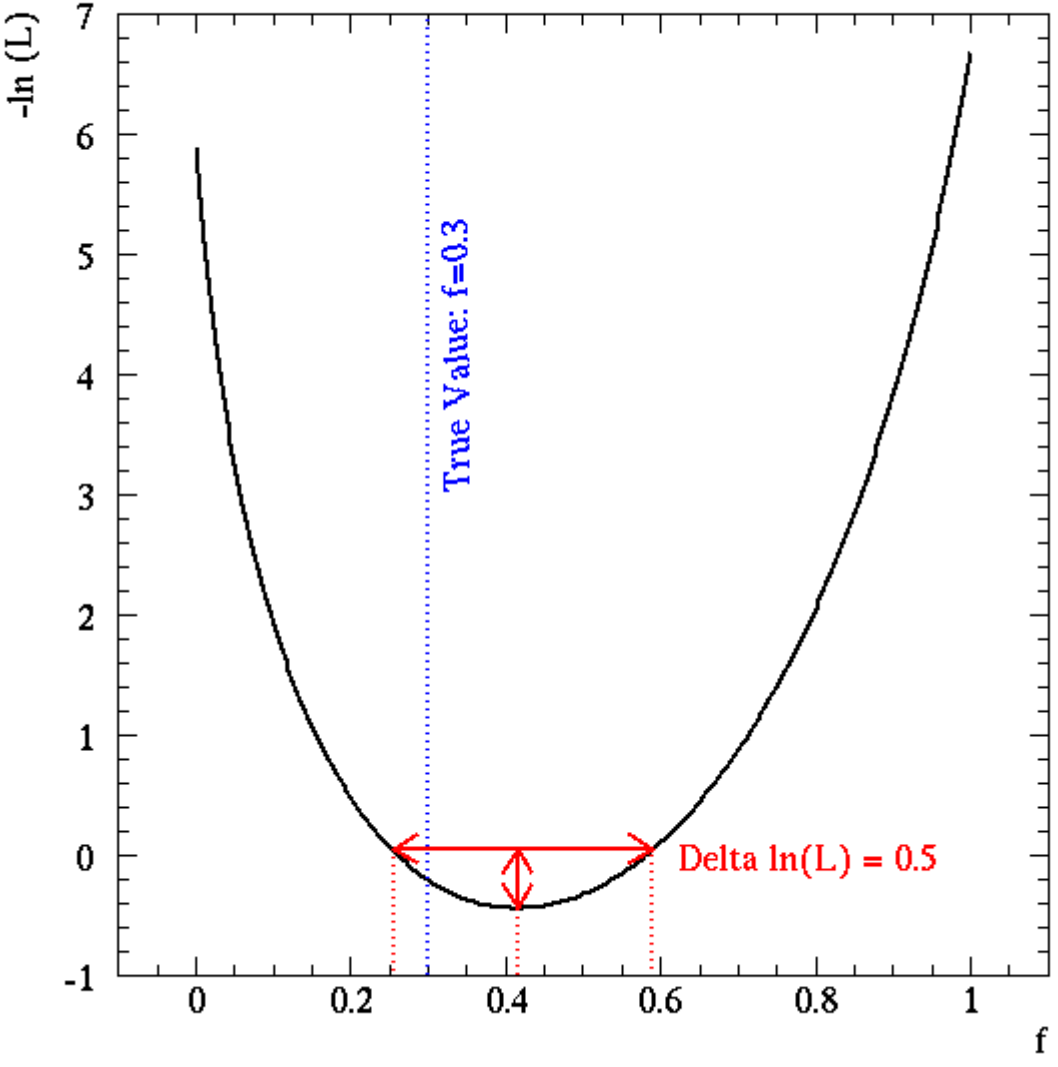
2D Feldman-Cousins contours

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.



The Δ ln(L) rule



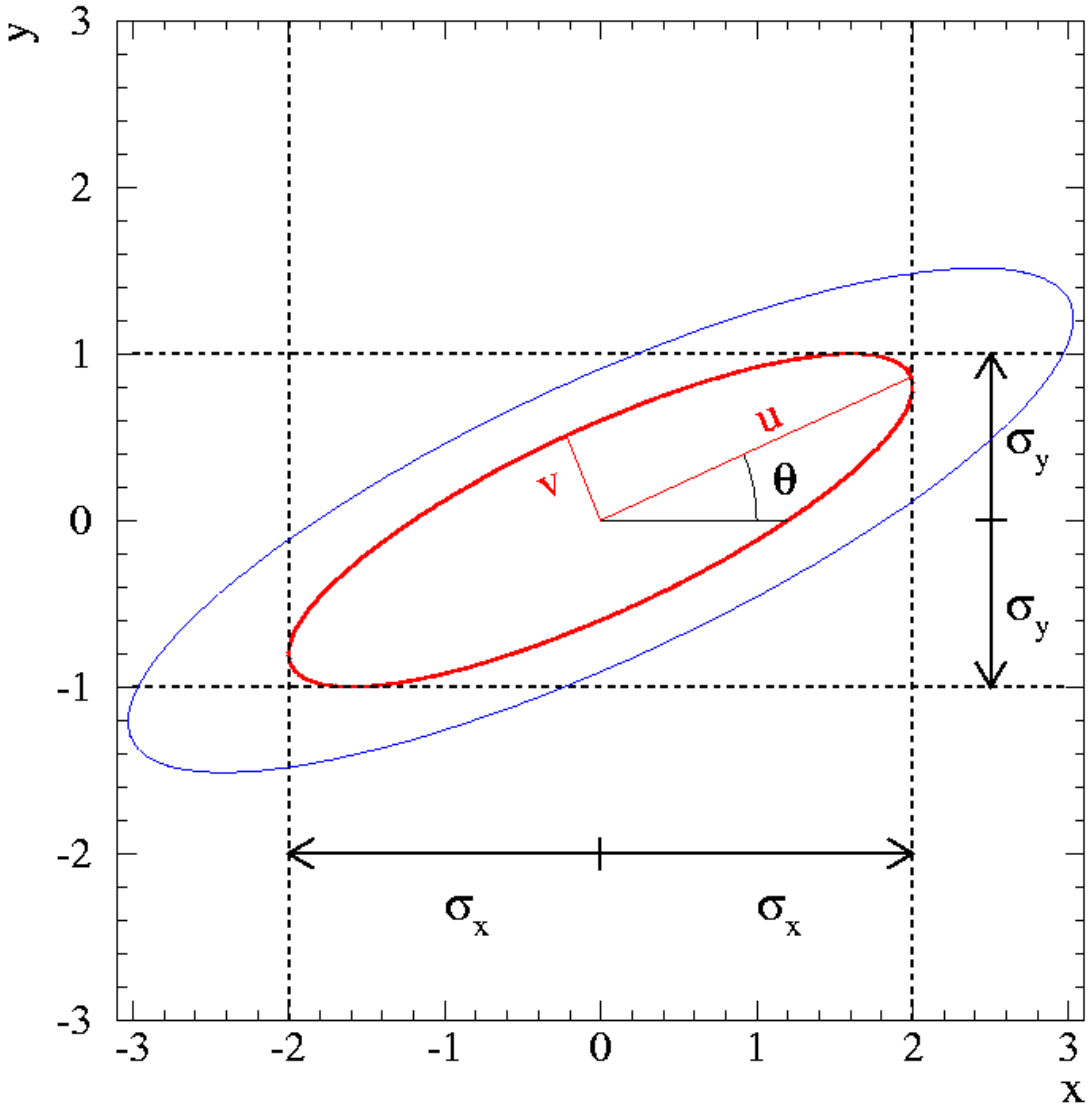
It is not trivial to construct proper frequentist confidence intervals. Most often an approximation is used: the confidence interval for a single parameter is defined as the range in which $\ln(L_{max})-\ln(L)<0.5$

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 ω = "$\Delta\chi^2$" = $2[\ln(L_{max})-\ln(L)]$ approximates a χ^2 with d degrees of freedom.

For experts: there do exist corrections to the Δ ln(L) rule that more accurately approximate coverage---see "Bartlett's correction". Often MC is better way to go.

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 Num Rec Sec 15.6)

Problem to work in groups:

A neutrinoless double-beta decay experiment counts the number of events in a signal region. The expected background is 2 events. For an effective neutrino mass of $m_{\beta\beta}$ = 50 meV the experiment expects 4 signal events. The experiment is done, and no events are seen in the signal region.

- What is the Feldman-Cousins upper limit on $m_{\beta\beta}$?
- Assuming a flat prior, what is the Bayesian upper limit on $m_{\beta\beta}$?
- Suppose a second experiment has the same expected signal rate but an expected background of 5. It also observes zero events. What Feldman-Cousins limit do you get now? Compare to your result from Part A. Do these results make sense?

Backups

How to report systematics

In reality there is no deep fundamental distinction between statistical and systematic errors. (Bayesians will say that both equally reflect our uncertainty about the universe.) Nonetheless, it is traditional, and useful, to separately quote the errors, such as X = 5.2 ± 2.4(stat) ± 1.5(sys).

There is a common tendency to assume that statistical and systematic uncertainties will be uncorrelated. This is often the case, but not always. (For example, if the data itself is providing a meaningful constraint on the nuisance parameter, there will likely be a correlation.) If such a correlation exists, report it explicitly (maybe as contour plots of X vs. the nuisance parameters). Otherwise you can be sure that someone is going to take your data, add the errors in quadrature, and report

$$X = 5.2 \pm \sqrt{2.4^2 + 1.5^2} = 5.2 \pm 2.8$$

Consider making the full form of the joint likelihood (or the priors and posterior PDFs if it's a Bayesian analysis) publicly available---on the web, if it won't fit in the paper itself.

Averaging correlated measurements II

The obvious generalization for correlated uncertainties is to form the χ^2 including the covariance matrix:

$$\chi^2 = \sum_i \sum_j \left(x_i - \mu\right)\left(x_j - \mu\right)\left(V^{-1}\right)_{ij}$$

We find the best value of μ by minimizing this χ^2 and can then find the 1σ uncertainties on μ by finding the values of μ for which $\chi^2 = \chi^2_{\text{min}} + 1$.

This is really parameter estimation with one variable.

The best-fit value is easy enough to find:

$$\mu = \frac{\sum_{i,j} x_j \left(V^{-1}\right)_{ij}}{\sum_{i,j} \left(V^{-1}\right)_{ij}}$$

Error-weighted averages

Suppose you have N independent measurements of a quantity. You average them. The proper error-weighted average is:

$$\langle x \rangle = \frac{\sum x_i / \sigma_i^2}{\sum 1/\sigma_i^2}$$

$$V\left(\langle x \rangle\right) = \frac{1}{\sum 1/\sigma_i^2}$$

If all of the uncertainties are equal, then this reduces to the simple arithmetic mean, with V(<x>) = V(x)/N.

Bayesian derivation of error-weighted averages

Suppose you have N independent measurements of a quantity, distributed around the true value μ with Gaussian distributions. For flat prior on μ we get:

$$P\left(\mu|\vec{x}\right) \propto \prod_i \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma_i}\right)^2\right] = \exp\left[-\frac{1}{2}\sum_i\left(\frac{x_i-\mu}{\sigma_i}\right)^2\right]$$

It's easy to see that this has the form of a Gaussian. To find its peak, set derivative with respect to μ equal to zero.

$$\frac{dP}{d\mu} = \exp\left[-\frac{1}{2}\sum_i\left(\frac{x_i-\mu}{\sigma_i}\right)^2\right]\left[\sum_i\left(\frac{x_i-\mu}{\sigma_i^2}\right)\right] = 0 \quad \rightarrow \quad \mu = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}$$

Calculating the coefficient of μ^2 in the exponent yields:

$$V\left(\langle x\rangle\right) = \frac{1}{\sum 1/\sigma_i^2}$$

Averaging correlated measurements

We already saw how to average N independent measurement. What if there are correlations among measurements?

For the case of uncorrelated Gaussianly distributed measurements, finding the best fit value was equivalent to minimizing the chi-squared:

$$\chi^2 = \sum_i \left(\frac{x_i - \mu}{\sigma_i} \right)^2$$

In Bayesian language, this comes about because the PDF for μ is $\exp(-\chi^2/2)$. Because we know that this PDF must be Gaussian:

$$P(\mu) \propto \exp\left[-\frac{1}{2} \left(\frac{\mu - \mu_0}{\sigma_\mu} \right)^2 \right]$$

then an easy way to find the 1σ uncertainties on μ is to find the values of μ for which $\chi^2 = \chi^2_{\text{min}} + 1$.

Averaging correlated measurements III

Recognizing that the χ^2 really just is the argument of an exponential defining a Gaussian PDF for μ ...

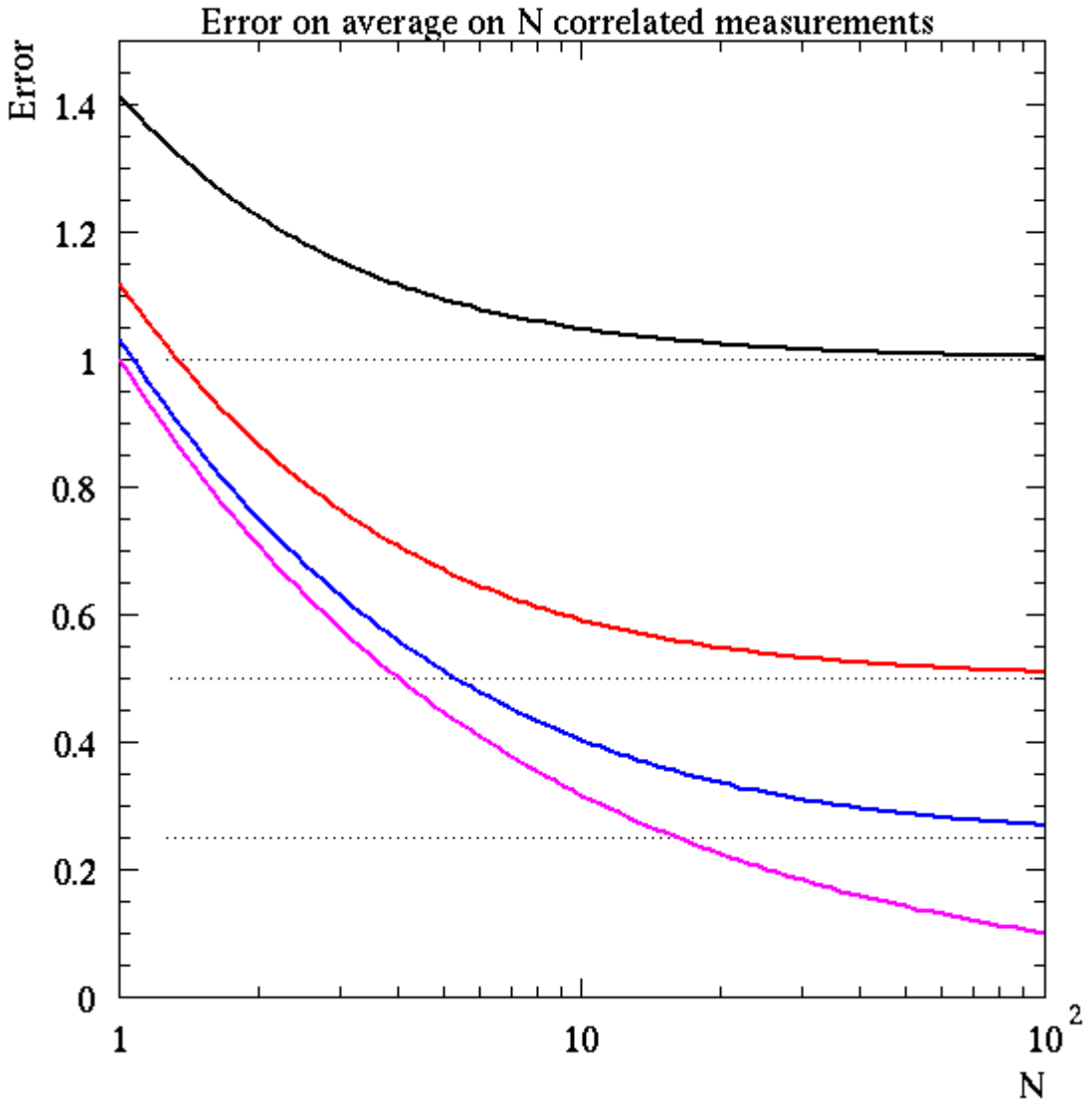
$$\chi^2 = \sum_i \sum_j \left(x_i - \mu\right)\left(x_j - \mu\right)\left(V^{-1}\right)_{ij}$$

we can in fact read off the coefficient of μ^2, which will be $1/V(\mu)$:

$$\sigma_\mu^2 = \frac{1}{\sum_{i,j}\left(V^{-1}\right)_{ij}}$$

In general this can only be computed by inverting the matrix as far as I know.

Averaging correlated measurements IV



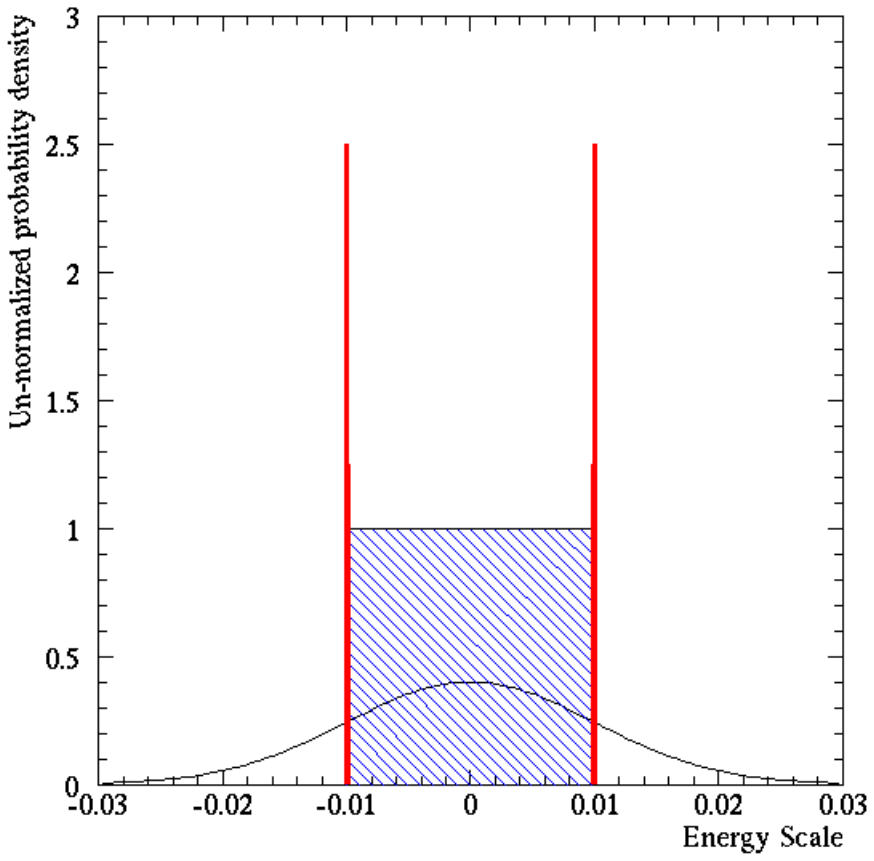
Suppose that we have N correlated measurements. Each has some independent error σ=1 and a common error b that raises or lowers them all together. (You would simulate by first picking a random value for b, then for each measurement picking a new random value c with RMS σ and writing out b+c.

Each curve shows how the error on the average changes with N, for different values of b.

b=1
b=0.5
b=0.25
b=0

Avoid inflating systematics

There is a regrettable tendency to overestimate systematics in the name of CYA or to save effort. For example, perhaps you've concluded that the energy scale of your detector is at worst off by 1%. So you write σ_E=0.01, and proceed to treat this as the RMS of your nuisance parameter.



Black: typical Gaussian PDF implied by σ=0.01. Long tails far beyond "worst case" range.

Red: δ functions---only PDF with σ=0.01 that is fully contained within "worst case" range

Blue: uniform distribution consistent with "worst case range. σ=0.02/$\sqrt{12}$=0.0058

Which should you use?

What is a systematic uncertainty?

There are many meanings of the term “systematic uncertainty”. (I prefer this term to “systematic error”, which means more or less the same thing.)

The most common definition is “any error that's not a statistical error”.

To avoid this definition becoming circular, we'd better be more precise.

Perhaps this works: “A systematic uncertainty is a possible unknown variation in a measurement, or in a quantity derived from a set of measurements, that does not randomly vary from data point to data point.”

Usually you see it listed broken out as: 5.0 ± 1.2 (stat) ± 0.8 (sys)

Examples of systematic uncertainties

“Like sands through an hourglass, so are the systematics of our lives ...”

- You measure the length of an object, but worry that the ruler might have contracted slightly due to it being a cold day.
- You try to infer the brightness of a distant supernova, but worry that intervening dust might make it seem dimmer than you expect.
- Your thermometer is miscalibrated.
- You measure g-2, the anomalous magnetic moment of the muon, and ask whether it agrees with the Standard Model expectation. A theorist tells you that there are higher order corrections to the theory prediction that are too complicated for her to calculate, but she helpfully quotes an uncertainty based on how large she believes these are likely to be.
- You are trying to fit an energy spectrum to an expected shape plus a background component to determine the size of a signal. There are two experimental measurements of the expected shape. They disagree by an amount much larger than their error bars.

Why are systematics problematic for frequentists?

The whole frequentist program is based upon treating the outcomes of experiments as “random variables”, and predicting the probabilities of observing various outcomes. For quantities that fluctuate, this makes sense.

But often we conceive of systematic uncertainties that aren't fluctuations. Maybe your thermometer really IS off by 0.2K, and every time you repeat the measurement you'll have the same systematic bias.

There's both a conceptual problem and a practical problem here. Conceptually, we resort to the dodge of imagining “identical” hypothetical experiments, except that certain features of the setup are allowed to vary. Practically, we usually can't measure the size of a systematic by repeating the measurement 100 times and looking at the distribution. We're almost forced to be pseudo-Bayesian about the whole thing.

Bayesian approach to systematics

Bayesians lose no sleep over systematics. Suppose you want to measure some quantity θ. You have a prior $P(\theta|I)$, you observed some data D, and you need to calculate a likelihood $P(D|\theta,I)$. Let's suppose that the likelihood depends on some systematic parameter α (which could for example be the calibration of the energy scale). We handle the systematic uncertainty by simply treating both θ and α as unknown parameters, assign a prior to each, and write down Bayes theorem:

$$P(\theta,\alpha|D,I)=\frac{P(D|\theta,\alpha,I)P(\theta,\alpha|I)}{\int d\theta\, d\alpha\, P(D|\theta,\alpha,I)P(\theta,\alpha|I)}$$

In the end we get a distribution for θ, whose value we care about, and for α, which may be uninteresting. We marginalize by integrating over α to get $P(\theta|I)$.

The prior $P(\alpha)$ presents our prior knowledge of α and is often the result of a calibration measurement.

Note that since the likelihood $P(D|\theta,\alpha,I)$ depends on as α well, it can provide additional information on α.

Distinction between statistical and systematic uncertainties

A common set of definitions:

A “statistical uncertainty” represents the scatter in a parameter estimation caused by fluctuations in the values of random variables. Typically this decreases in proportion to $1/\sqrt{N}$.

A “systematic uncertainty” represents a constant (not random) but unknown error whose size is independent of N.

DO NOT TAKE THESE DEFINITIONS TOO SERIOUSLY. Not all statistical uncertainties decrease like $1/\sqrt{N}$. And more commonly, taking more data can decrease a systematic uncertainty as well, especially when the systematic affects different parts of the data in different ways, as in the example on the previous page.

Need to have a systematics model

The most important step in dealing with any systematic is to have a quantitative model of how it affects the measurement. This includes:

A. How does the systematic affect the measured data points themselves?
B. How does the systematic appear quantitatively in the calculations applied to the data?

It is essential to have some model, however simplified, in order to quantify the systematic uncertainty.

Systematic error model #1: an offset

Suppose we take N measurements from a distribution, and wish to estimate the true mean of the underlying distribution.

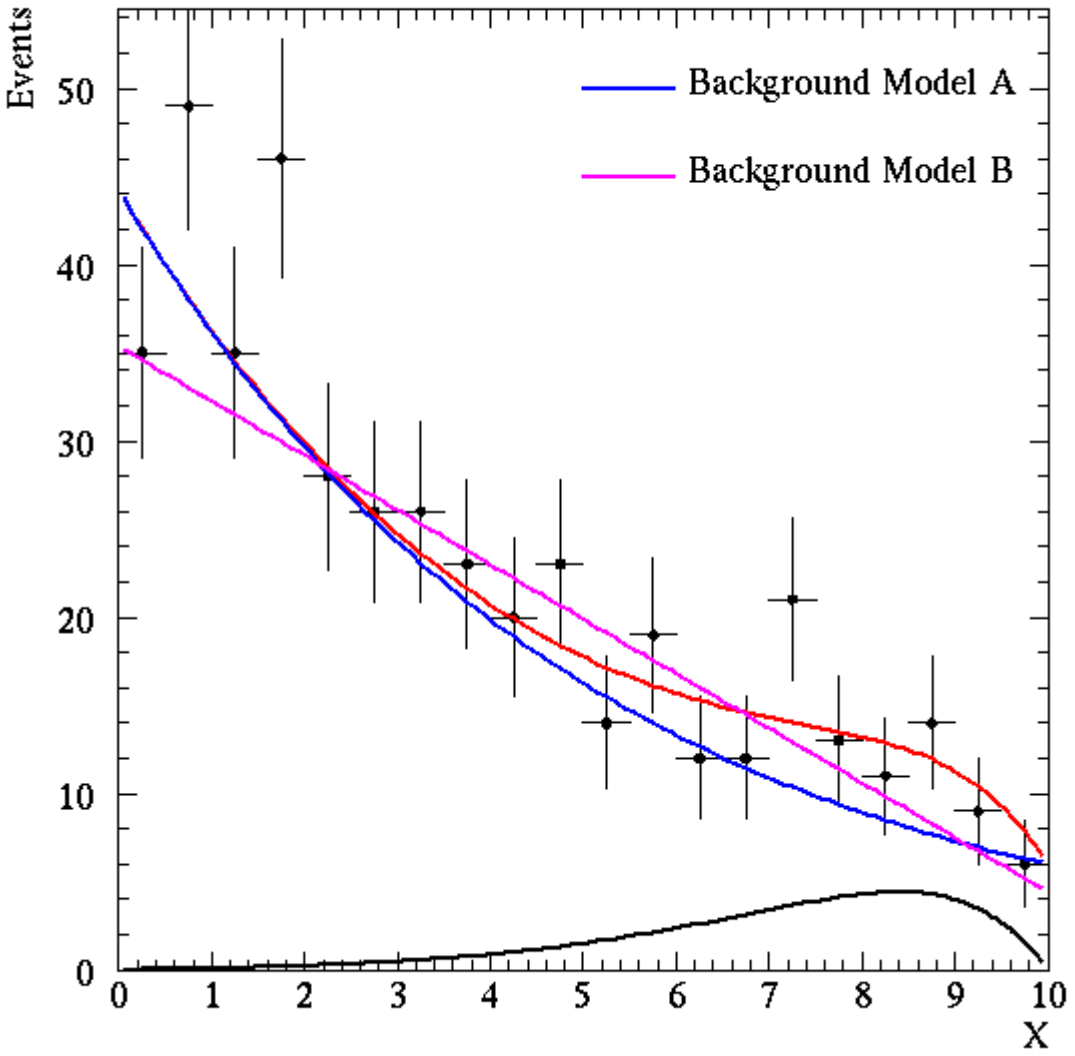
Our measuring apparatus might have an offset *s* from 0. We attempt to calibrate this. Our systematic error model consists of:

1) There is some additive offset *s* whose value is unknown.
2) It affects each measurement identically by $x_i \rightarrow x_i + s$.
3) The true mean is estimated by:

$$\hat{\mu} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i \right) - s$$

4) Our calibration is $s = 2 \pm 0.4$

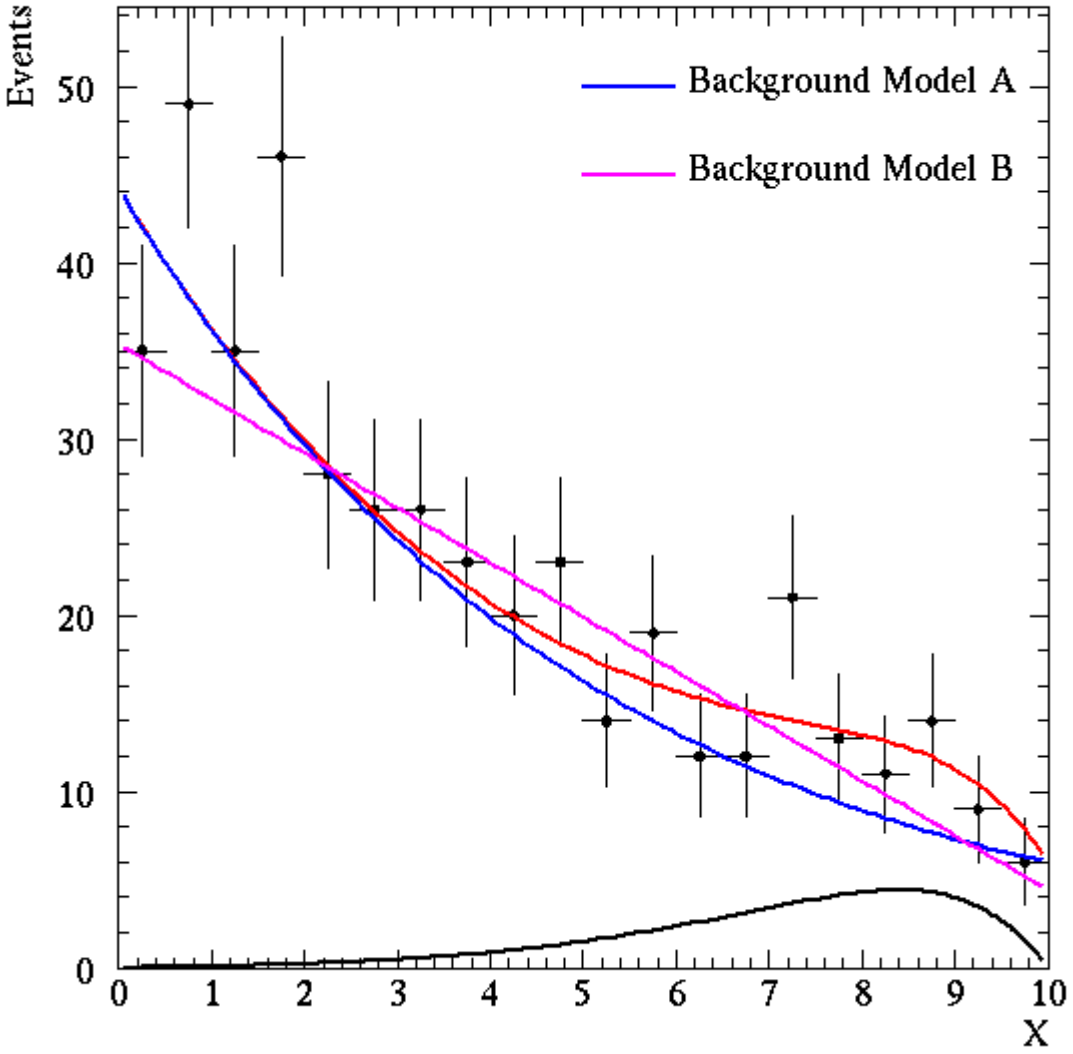
Systematic error model #2: two incompatible models



In order to determine the rate of some process, we fit the data to a two-component model consisting of a signal shape and a background shape.

But there are two different and mutually exclusive background models, which we'll denote as A and B.

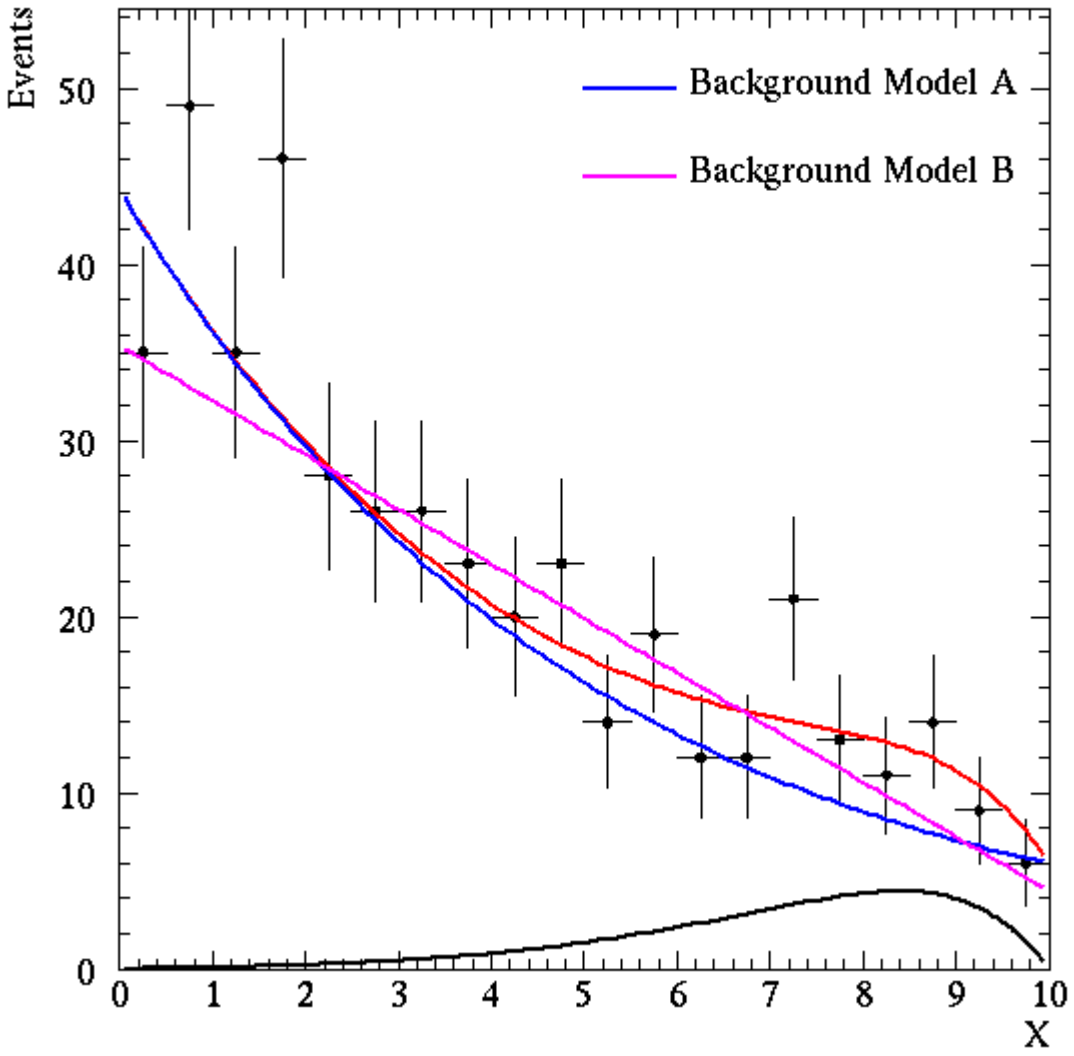
Systematic error model #2: two incompatible models



If the error ranges on the background models are negligible, one possibility is to just do the analysis twice, reporting the result with each model, and hope that future information will determine which model is right.

But in this case the shape of the data actually will tell us something about the two models---data will constrain the systematic (the shape of the background).

Systematic error model #2: two incompatible models



One approach is to make a parameterized background model that interpolates between the two:

$$m(x) = f m_A(x) + (1-f) m_B(x)$$

Here $0 \le f \le 1$. You can define whatever Bayesian prior you like for f (even Dirac delta functions at f=0 and f=1). Your fit to the data will favour some values of f and not others, but the most important thing is you've quantified the problem through nuisance parameter f.

How do you measure a systematic?

So you've quantified the effects of the systematic through some nuisance parameter. How do you determine the value of that nuisance parameter itself? Various approaches:

1) Calibration measurements, taken separately from your main data
2) A priori estimate based on known parameters of the apparatus
3) If data provides useful data about nuisance parameter, fit it from the main data itself.
4) “Theory”: some systematic uncertainties will be what we call “theoretical uncertainties”. There are various causes/interpretations:
   A. Measurement uncertainties in theory parameters
   B. Theorists' estimates of errors due to approximations made
   C. Spread between different theory estimates (careful here!)
5) Data vs. Monte Carlo comparisons: use calibration data to estimate how well Monte Carlo reproduces data, then use spread as an estimate of how well Monte Carlo predicts other quantities

How do you measure a systematic?

This is a black art. I'd argue that 90% of experimental physics is thinking of clever ways to reduce or at least measure systematics.

Severus Snape, dabbler in the black arts

Unfortunately, there is no real magic, merely hard work.

A strong dose of paranoia helps as well.

Why you should avoid inflating systematics...

What's wrong with inflating systematics to cover all bases? Isn't this the “conservative” thing to do?

1) This tends to paper over model inaccuracies, and imply greater support for your model than is warranted. (Think Bayesian-wise: since Bayesian analyses always choose *between* competing hypotheses, being “conservative” with one hypothesis is equivalent to selectively favouring another.

2) Your inflated error might hide a serious problem with your data, or worst of all may miss an important discovery.

3) Tendency to bias: everyone recognizes that it's wrong to “fudge your data” to make your central value agree with expectations. Fewer people recognize that it's equally wrong to inflate your errors to make sure the error bars overlap the expected value!

Propagating systematics with Monte Carlo

So you've listed all of the systematics, mapped them all to nuisance parameters (or decided that they're negligible), and have assigned PDFs to each nuisance parameter. What next?

“Propagating the systematics” means to determine how much uncertainty results in your final value from your systematics model. Toy Monte Carlo is an excellent way to do evaluate this:

1) Randomly choose values for each nuisance parameter according to their respective PDFs.
2) Analyze the data as if those values of the nuisance parameters are the true values for the systematic parameters.
3) Repeat many times.
4) If you're trying to estimate the error on a fit parameter, plot the distribution of the fitted values of that parameter. Take the RMS width as the systematic error.

Propagating systematics with Monte Carlo 2

Advantages of the Monte Carlo method:

- few approximations made---no need to assume Gaussian errors
- considers the effects of all systematics jointly, including nonlinearities
- can easily accommodate correlations between systematics

Disadvantages of the Monte Carlo method:

- method does not allow the data itself to constrain the systematics
- because all systematics are varied at once, the resulting distribution is the convolution of the effects of all nuisance parameters. On the one hand this is a feature---in real life all systematics vary at once, and so Monte Carlo gives an "exact" way of modelling how various systematics interact. On the other hand, if you want to understand the relative importance of each component, you have to either marginalize or project over each parameter, or rerun your Monte Carlos, this time varying just one systematic at a time. (Actually, this is recommended practice in any case.)

Covariance matrix approach

Monte Carlo is not always necessary, and not always the fastest way to propagate systematics. In the "covariance matrix" approach, you treat the nuisance parameter *s* and the data values x_j as a set of correlated random variables. You then calculate their full covariance matrix, and use error propagation to estimate the uncertainties.

Ex. taking the average of a set of measurements with a systematic additive offset:

$$x_j = \mu + X_j + s$$

(Implicitly assuming X_j is independent of s).

$$\mathrm{cov}\left(x_i, x_j\right) = \mathrm{cov}\left(\mu + X_i + s, \mu + X_j + s\right) = \mathrm{cov}\left(X_i, X_j\right) + \mathrm{cov}\left(s, s\right)$$

You can think of this as the sum of two covariance matrices:

$$V_{tot} = V_{stat} + V_{sys}$$

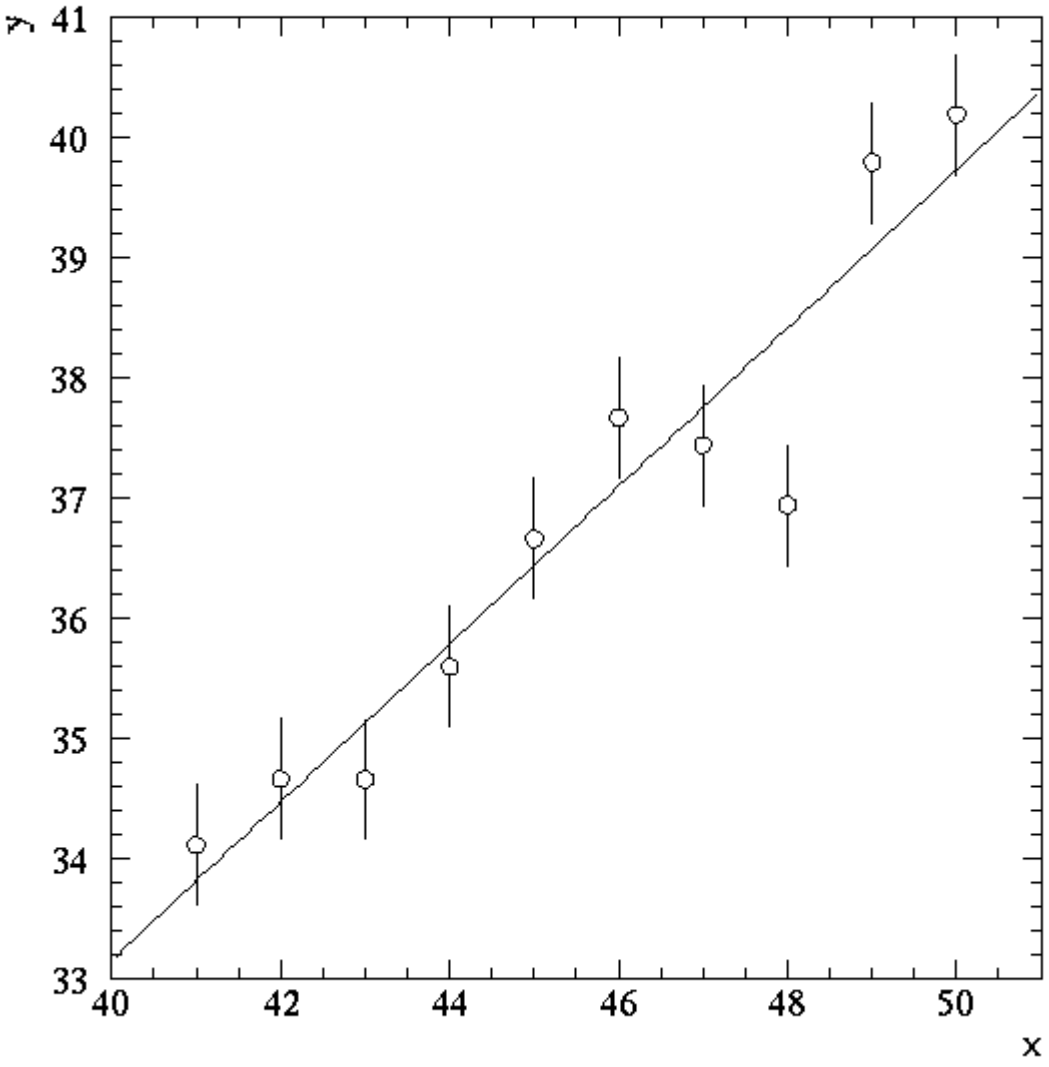
Covariance matrix approach 2

Now just include the new covariance matrix in your analysis wherever you previously had just the statistical error covariances---e.g.

$$\chi^2(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(y_i - f(x_i|\theta)\right) V_{ij}^{-1} \left(y_j - f(x_j|\theta)\right)$$

Note: in this approach you often will consider the value of *s* to be fixed at its central value. In other words, although the covariance matrix V will contain information on how much the uncertainties on the measured values y_i are increased by the systematic, the above formulation doesn't directly yield a refined estimate of *s*. We'll correct this shortly.

Reducing correlations in the straight line fit



The strong correlation between m and b results from the long lever arm--- since you must extrapolate line to x=0 to determine b, a big error on m makes a big error on b.

You can avoid strong correlations by using more sensible parametrizations: for example, fit data to y=b'+m(x-45.5):

b' = 36.77 ± 0.16
m = 0.658 ± 0.085
ρ = 0.43

dy at x=45.5 = 0.16

Non-Gaussian errors

The error propagation can give the false impression that propagating errors is as simple as plugging in variances and covariances into the error propagation equation and then calculating an error on output.

However, a significant issue arises: although the error propagation equation is correct as far as it goes (small errors, linear approximations, etc), it is often not true that the resulting uncertainty has a Gaussian distribution! Reporting the central value and an RMS may be misleading.

Ratio of two Gaussians I

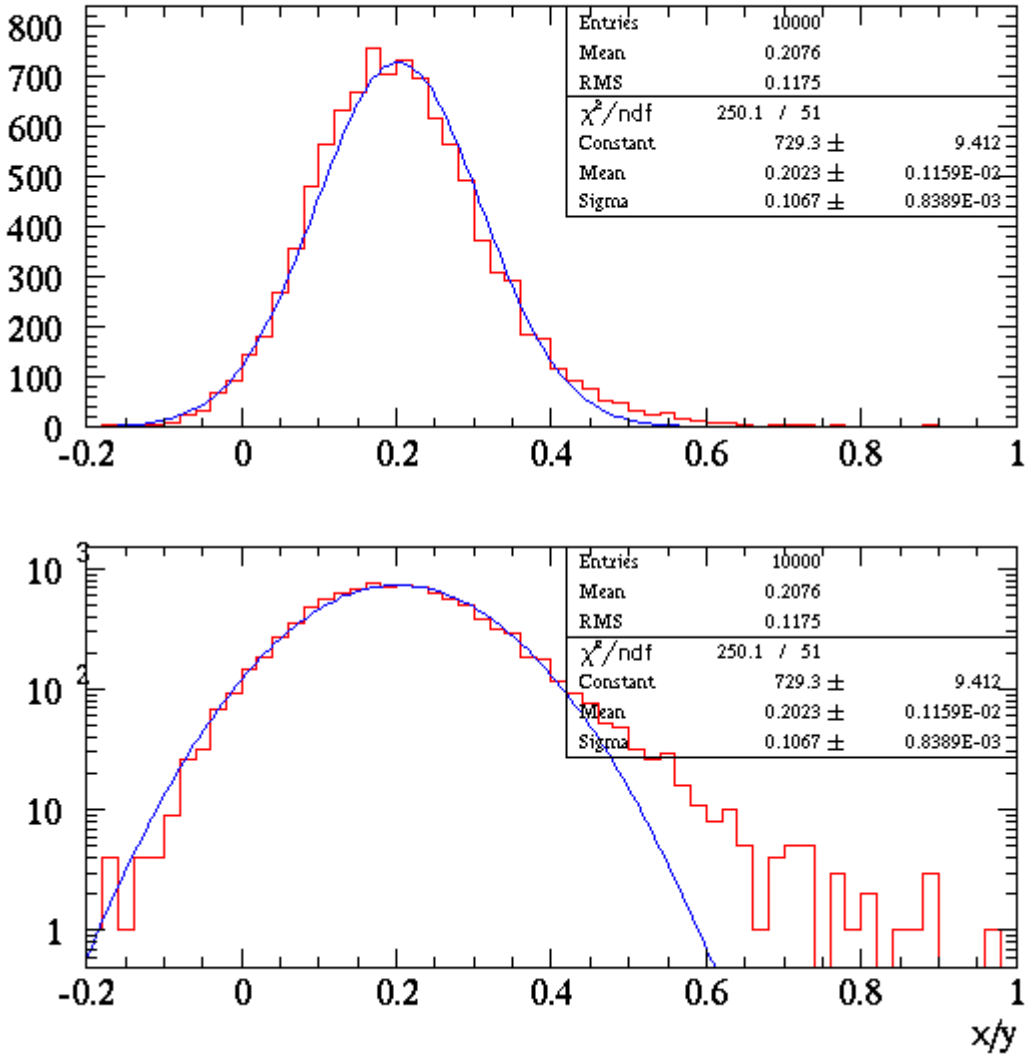
Consider the ratio R=x/y of two independent variables drawn from normal distributions. By the error propagation equation

$$\left(\frac{dR}{R}\right)^2 = \left(\frac{dx}{x}\right)^2 + \left(\frac{dy}{y}\right)^2$$

Let's suppose x = 1 ± 0.5 and y = 5 ± 1. Then the calculated value of R = 0.200 ± .108.

What does the actual distribution for R look like?

Ratio of two Gaussians II



x = 1 ± 0.5 and y = 5 ± 1.

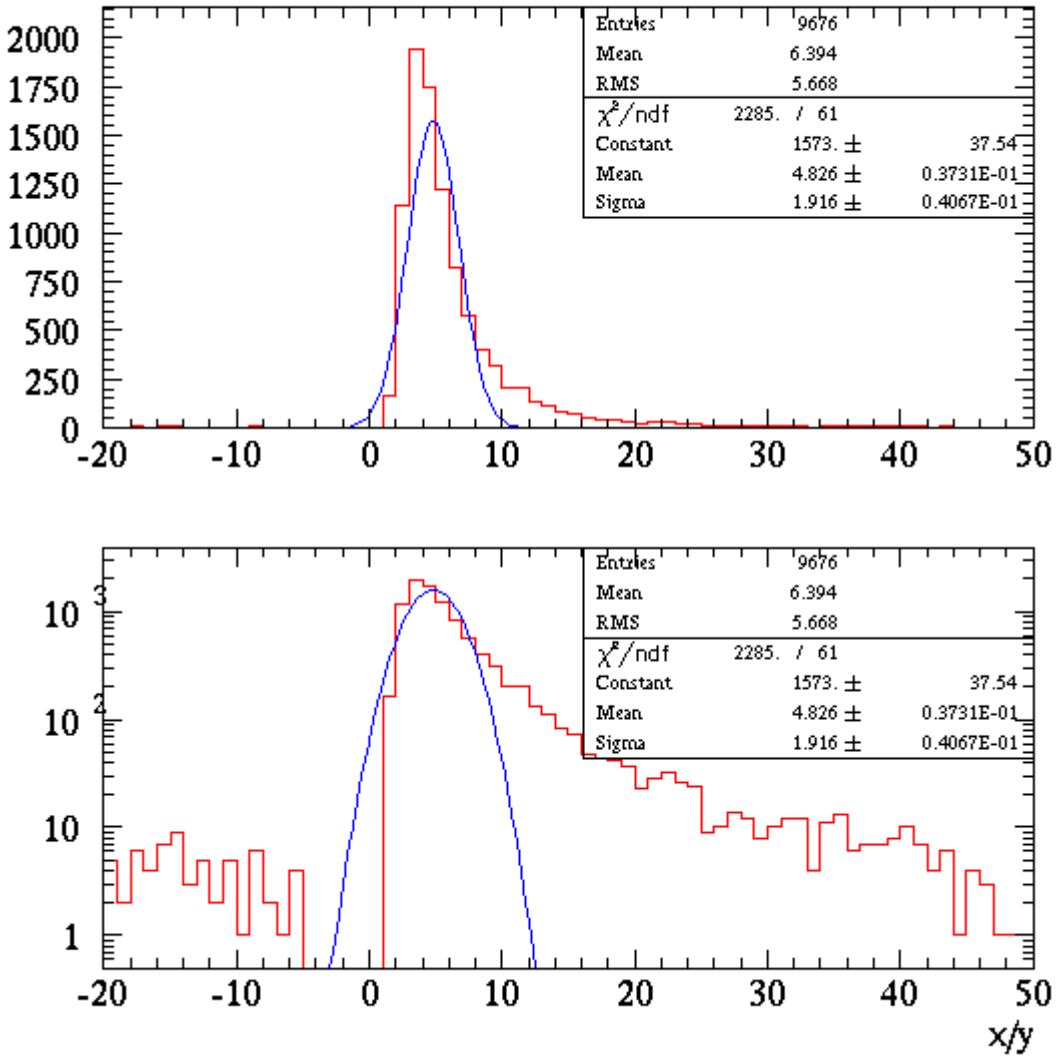
Error propagation prediction:
R = 0.200 ± .108.

Mean and RMS of R:
0.208 ± 0.118

Gaussian fit to peak:
0.202 ± 0.107

Non-Gaussian tails evident, especially towards larger R, but not too terrible.

Ratio of two Gaussians III



x = 5 ± 1 and y = 1 ± 0.5

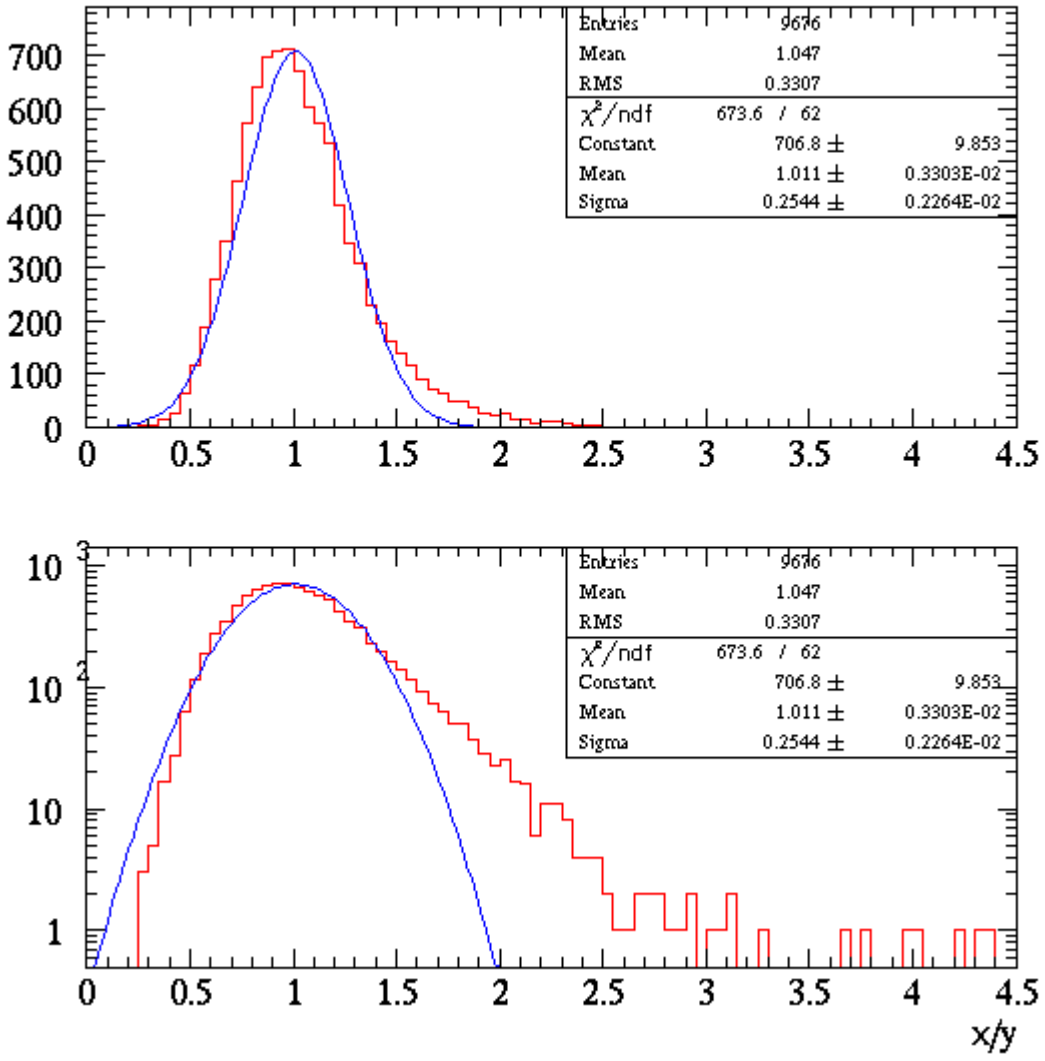
Error propagation:
R = 5.00 ± 2.69.

Mean and RMS of R:
6.39 ± 5.67

Gaussian fit to peak:
4.83 ± 1.92

Completely non-Gaussian in all respects. Occasionally we even divide by zero, or close to it!

Ratio of two Gaussians IV



$x = 5 \pm 1$ and $y = 5 \pm 1$

Error propagation:
$R = 1 \pm 0.28$

Mean and RMS of R:
1.05 ± 0.33

Gaussian fit to peak:
1.01 ± 0.25

More non-Gaussian than first case, much better than second.

Rule of thumb: ratio of two Gaussians will be approximately Gaussian if fractional uncertainty is dominated by numerator, and denominator cannot be small compared to numerator.

Ratio of two Gaussians: testing an asymmetry

A word of caution: often scientists like to form asymmetries---for example, does the measurement from the left half of the apparatus agree with that from the right half? Asymmetry ratios are the usual way to do this, since common errors will cancel in ratio:

$$A = \frac{\phi_L - \phi_R}{\frac{1}{2}\left(\phi_L + \phi_R\right)}$$

Be extremely careful about using the error on the asymmetry as a measure of whether A is consistent with zero. In other words, A=0.5 ± 0.25 is usually not a “2 sigma” result, probability-wise.

Instead, it's better to simply test whether the numerator is consistent with zero or not.

Asymmetric errors

The error propagation equation works with covariances. But the confidence interval interpretation of error bars often reports asymmetric errors. We can't use the error propagation equation to combine asymmetric errors. What do we do?

Quite honestly, the typical physicist doesn't have a clue. The most common procedure is to separately add the negative error bars and the positive error bars in quadrature:

Source	- Error	+ Error
Error A	-0.025	+0.050
Error B	-0.015	+0.010
Error C	-0.040	+0.040
Combined	-0.049	+0.065

Warning: in spite of how common this procedure is and what you may have heard, it has no statistical justification and often gives the wrong answer!

How to handle asymmetric errors

Best case: you know the likelihood function (at least numerically). Report it, and include it in your fits.

More typical: all you know are a central value, with +/- errors. Your only hope is to come up with some parametrization of the likelihood that works well. This is a black art, and there is no generally satisfactory solution.

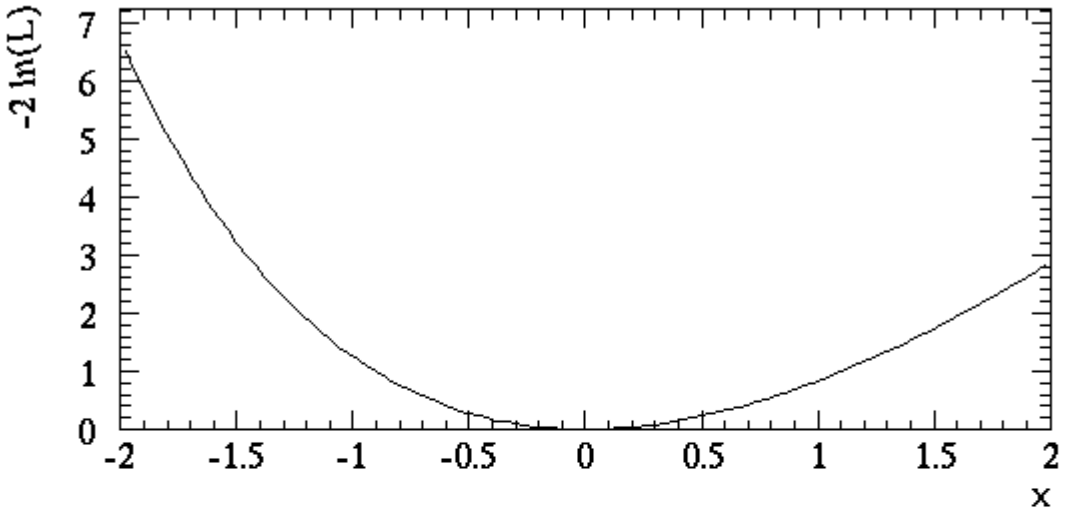
Roger Barlow recommends (in a paper on his web site):

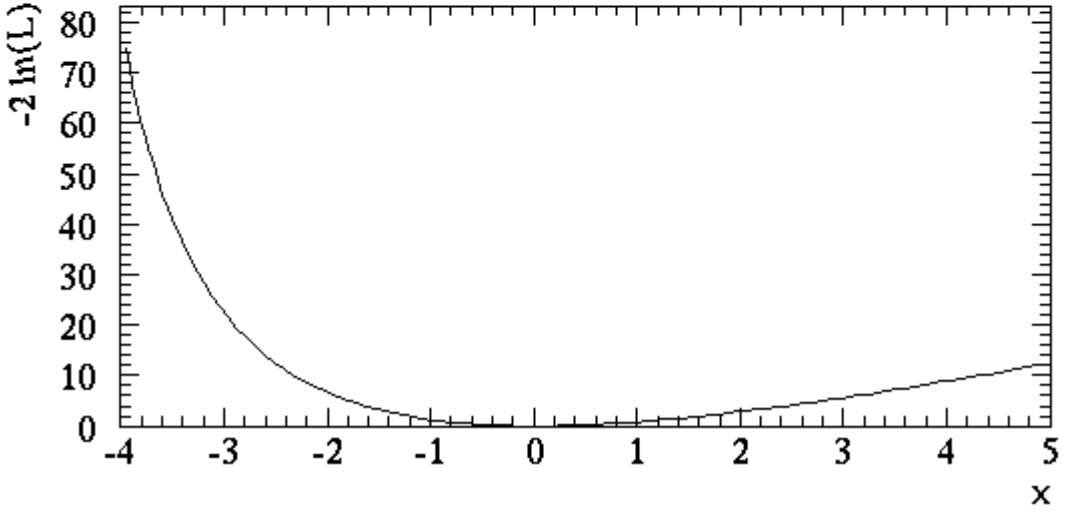
$$-\ln L(\alpha) = \frac{1}{2}\frac{\alpha^2}{\sigma_1\sigma_2 + (\sigma_1 - \sigma_2)\alpha}$$

You can verify that this expression evaluates to ½ when $\alpha = +\sigma_1$ or $\alpha = -\sigma_2$

Parametrizing the asymmetric likelihood function

$$-\ln L(\alpha) = \frac{1}{2}\frac{\alpha^2}{\sigma_1\sigma_2 + (\sigma_1 - \sigma_2)\alpha}$$





Note that parametrization breaks down when denominator becomes negative. It's like saying there is a hard limit on α at some point. Use this only over its appropriate range of validity.

Barlow recommends this form because it worked well for a variety of sample problems he tried---it's completely empirical.

See Barlow, "Asymmetric Statistical Errors", arXiv:physics/0406120

Application: adding two measurements with asymmetric errors.

Suppose $A = 5^{+1}_{-2}$ and $B = 3^{+1}_{-3}$. What is the confidence interval for A+B?

$$-\ln L(A,B) = \frac{1}{2}\frac{(A-5)^2}{(1)(2)+(1-2)(A-5)} + \frac{1}{2}\frac{(B-3)^2}{(1)(3)+(1-3)(B-3)}$$

Let C=A+B. Rewrite likelihood to be a function of C and one of the other variables---eg.

$$-\ln L(C,A) = \frac{1}{2}\frac{(A-5)^2}{(1)(2)+(1-2)(A-5)} + \frac{1}{2}\frac{(C-A-3)^2}{(1)(3)+(1-3)(C-A-3)}$$

To get the likelihood function for C alone, which is what we care about, minimize with respect to nuisance parameter A:

$$\text{-}\ln L(C) = \min_{A} \text{-}\ln(C,A)$$

Asymmetric error results