Physics 509: Introduction to Parameter Estimation

Scott Oser
Lecture #8

Outline

Last time: we reviewed multi-dimensional Gaussians, the Central Limit Theorem, and talked in a general way about Gaussian error ellipses.

Today:

1) Introduction to parameter estimation
2) Some useful basic estimators
3) Error estimates on estimators
4) Maximum likelihood estimators
5) Estimating errors in the ML method
6) Extended ML method

What is an estimator?

Quite simple, really ... an estimator is a procedure you apply to a data set to estimate some property of the parent distribution from which the data is drawn.

This could be a recognizable parameter of a distribution (eg. the *p* value of a binomial distribution), or it could be a more general property of the distribution (eg. the mean of the parent distribution).

The procedure can be anything you do with the data to generate a numerical result. Take an average, take the median value, multiply them all and divide by the GDP of Mongolia ... all of these are estimators. You are free to make up any estimator you care to, and aren't restricted to standard choices. (Whether an estimator you make yourself is a useful estimator or not is a completely separate question!)

Bayesian estimators

You're already seen the Bayesian solution to parameter estimation ... if your data is distributed according to a PDF depending on some parameter *a*, then Bayes' theorem gives you a formula for the PDF of *a*:

$$P(a|D,I)=\frac{P(a|I)P(D|a,I)}{\int da\,P(a|I)P(D|a,I)}=\frac{P(a|I)P(D|a,I)}{P(D|I)}$$

The PDF P(*a*|D,I) contains all the information there is to have about the true value of *a*. You can report it any way you like---preferably by publishing the PDF itself, or else if you want to report just a single number you can calculate the most likely value of *a*, or the mean of its distribution, or whatever you want.

There's no special magic: Bayesian analysis directly converts the observed data into a PDF for any free parameters.

Frequentist estimators

Frequentists have a harder time of it ... they say that the parameters of the parent distribution have some fixed albeit unknown values. “It doesn't make sense to talk about the probability of a fixed parameter having some other value---all we can talk about is how likely or unlikely was it that we would observe the data we did given some value of the parameter. Let's try to come up with estimators that are as close as possible to the true value of the parameter.”

Most of what we will be talking about in this lecture is frequentist methodology, although I'll try to relate it to Bayesian language as much as possible.

Desired properties of estimators

What makes a good estimator? Consider some $\hat{a} = \hat{a}(x_1, x_2, \ldots x_n)$

1) Consistent: a consistent estimator will tend to the true value as the amount of data approaches infinity:

$$\lim_{N \to \infty} \hat{a} = a$$

2) Unbiased: the expectation value of the estimator is equal to its true value, so its bias *b* is zero.

$$b = \langle \hat{a} \rangle - a = \int dx_1 \ldots dx_n P(x_1 \ldots x_n | a) \hat{a}(x_1 \ldots x_n) - a = 0$$

3) Efficient: the variance of the estimator is as small as possible (as we'll see, there are limitations on how small it can be)

$$V(\hat{a}) = \int dx_1 \ldots dx_n P(x_1 \ldots x_n | a) \ (\hat{a}(x_1 \ldots x_n) - \langle \hat{a} \rangle)^2$$

$$(\text{Mean square error})^2 = \langle (\hat{a} - a)^2 \rangle = b^2 + V(\hat{a})$$

It's not always possible to satisfy all three of these requirements.

Common estimators

1) Mean of a distribution---obvious choice is to use the average:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Consistent and unbiased if measurements are independent. Not necessarily the most efficient---its variance depends on the distribution under consideration, and is given by

$$V(\hat{\mu}) = \frac{\sigma^2}{N}$$

There may be more efficient estimators, especially if the parent distribution has big tails. But in many circumstances the sample mean is the most efficient.

Estimating the variance

If you know the true mean μ of a distribution, one useful estimator (consistent and unbiased) of the variance is

$$\widehat{V(x)} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

What if μ is also unknown?

A biased estimator:

$$\widehat{V(x)} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$\langle \widehat{V(x)} \rangle = \frac{N-1}{N} V(x)$$

An unbiased estimator:

$$\widehat{V(x)} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

But its square root is a biased estimator of σ!

Estimating the standard deviation

The square root of an estimate of the variance is the obvious thing to use as an estimate of the standard deviation:

$$\widehat{V(x)} = s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

We can use *s* as our estimator for σ. It will generally be biased---we don't worry a lot about this because we're more interested in having an unbiased estimate of s^2.

For samples from a Gaussian distribution, the RMS on our estimate for σ is given by

$$\sigma_s = \frac{\sigma}{\sqrt{2(N-1)}}$$

Think of this as the "error estimate on our error bar".

Likelihood function and the minimum variance bound

Likelihood function: probability of data given the parameters

$$L(x_1 \ldots x_n | a) = \prod P(x_i | a)$$

(The likelihood is actually one of the factors in the numerator of Bayes theorem.)

A remarkable result---for any unbiased estimator for *a*, the variance of the estimator satisfies:

$$V(\hat{a}) \geq \frac{-1}{\left\langle \frac{d^2 \ln L}{da^2} \right\rangle}$$

If estimator is biased with bias *b*, then this becomes

$$V(\hat{a}) \geq \frac{-\left(1 + \frac{db}{da}\right)^2}{\left\langle \frac{d^2 \ln L}{da^2} \right\rangle}$$

The minimum variance bound

The minimum variance bound is a remarkable result---it says that there is some “best case” estimator which, when averaged over thousands of experiments, will give parameter estimates closer to the true value, as measured by the RMS error, than any other.

An estimator that achieves the minimum variance bound is maximally efficient.

Information theory is the science of how much information is encoded in data set. The MVB comes out of this science.

Maximum likelihood estimators

By far the most useful estimator is the maximum likelihood method. Given your data set x_1 ... x_N and a set of unknown parameters α, calculate the likelihood function

$$L(x_1 \ldots x_N|\vec{\alpha}) = \prod_{i=1}^{N} P(x_i|\vec{\alpha})$$

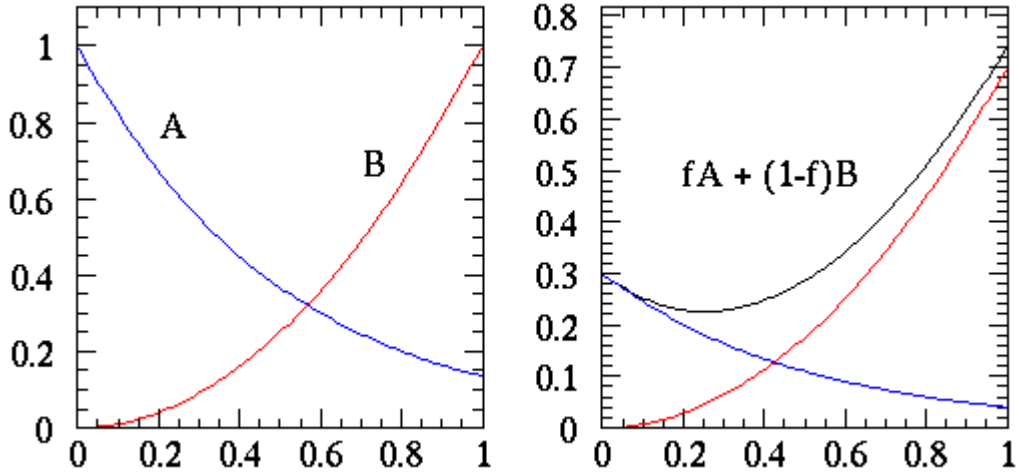
It's more common (and easier) to calculate -ln *L* instead:

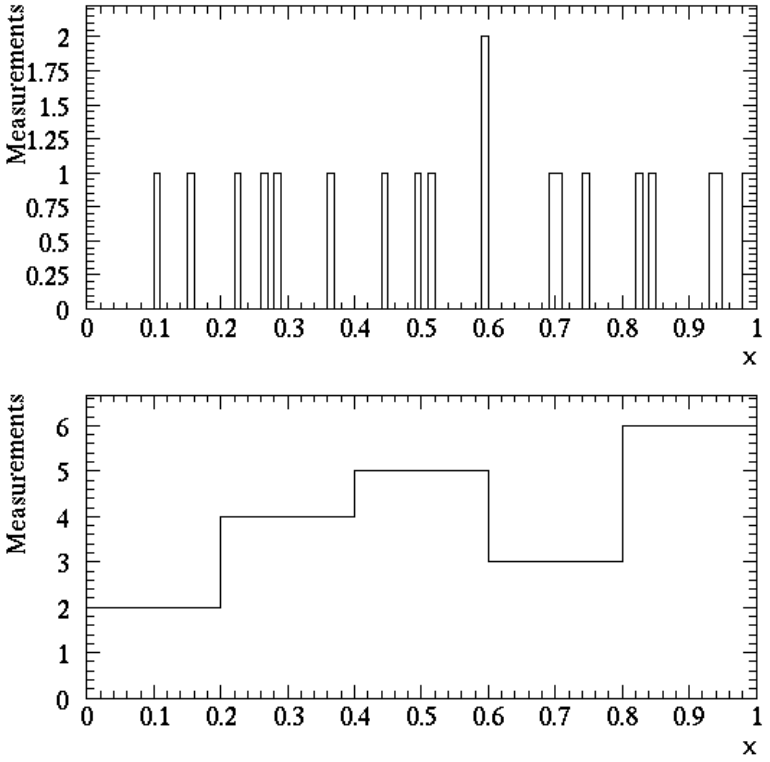
$$-\ln L(x_1 \ldots x_N|\vec{\alpha}) = -\sum_{i=1}^{N} \ln P(x_i|\vec{\alpha})$$

The maximum likelihood estimator is that value of α which maximizes L as a function of α. It can be found by minimizing -ln *L* over the unknown parameters.

Simple example of an ML estimator

Suppose that our data sample is drawn from two different distributions. We know the shapes of the two distributions, but not what fraction of our population comes from distribution A vs. B. We have 20 random measurements of X from the population.





$$P_A(x) = \frac{2}{1-e^{-2}} e^{-2x} \qquad P_B(x) = 3x^2$$

$$P_{tot}(x) = f\,P_A(x) + (1-f)\,P_B(x)$$

Form for the log likelihood and the ML estimator

Suppose that our data sample is drawn from two different distributions. We know the shapes of the two distributions, but not what fraction of our population comes from distribution A vs. B. We have 20 random measurements of X from the population.

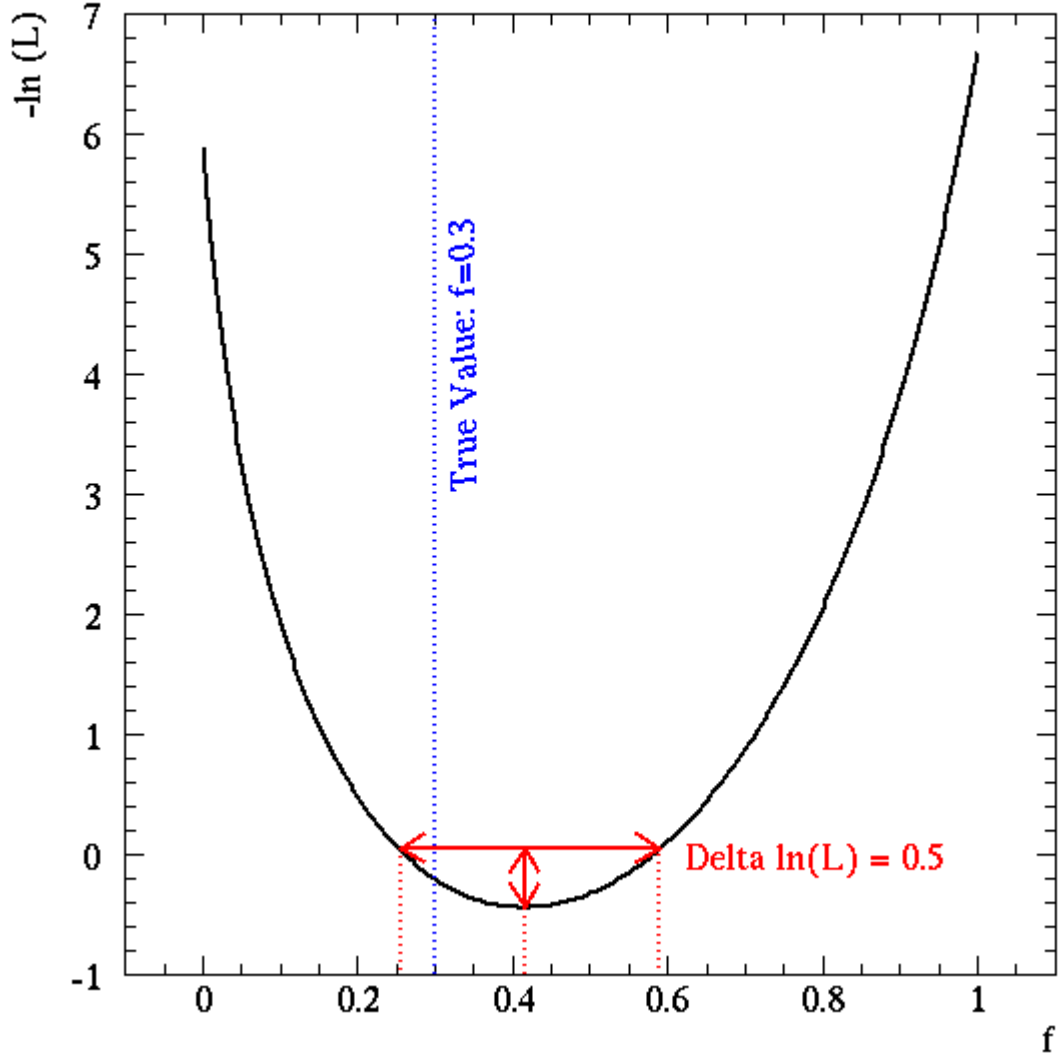
$$P_{tot}(x) = f\,P_A(x) + (1-f)\,P_B(x)$$

Form the negative log likelihood:

$$-\ln L(f) = -\sum_{i=1}^{N} \ln\left(P_{tot}(x_i|f)\right)$$

Minimize -ln(L) with respect to *f*. Sometimes you can solve this analytically by setting the derivative equal to zero. More often you have to do it numerically.

Graph of the log likelihood



The graph to the left shows the shape of the negative log likelihood function vs. the unknown parameter f.

The minimum is f=0.415. This is the ML estimate.

As we'll see, the "1σ" error range is defined by Δ ln(L)=0.5 above the minimum.

The data set was actually drawn from a distribution with a true value of f=0.3

Properties of ML estimators

Besides its intrinsic intuitiveness, the ML method has some nice (and some not-so-nice) properties:

1) ML estimator is usually consistent.

2) ML estimators are usually biased, although if also consistent then the bias approaches zero as N goes to infinity.

3) Estimators are invariant under parameter transformations:

$$\widehat{f(a)} = f(\hat{a})$$

4) In the asymptotic limit, the estimator is efficient. The Central Limit Theorem kicks on in the sum of the terms in the log likelihood, making it Gaussian:

$$\sigma^2_{\hat{a}} = \frac{-1}{\left(\left. \frac{d^2 \ln L}{da^2} \right|_{a_0} \right)}$$

Relation to Bayesian approach

There is a close relation between the ML method and the Bayesian approach.

The Bayesian posterior PDF for the parameter is the product of the likelihood function P(D|a,I) and the prior P(a|I).

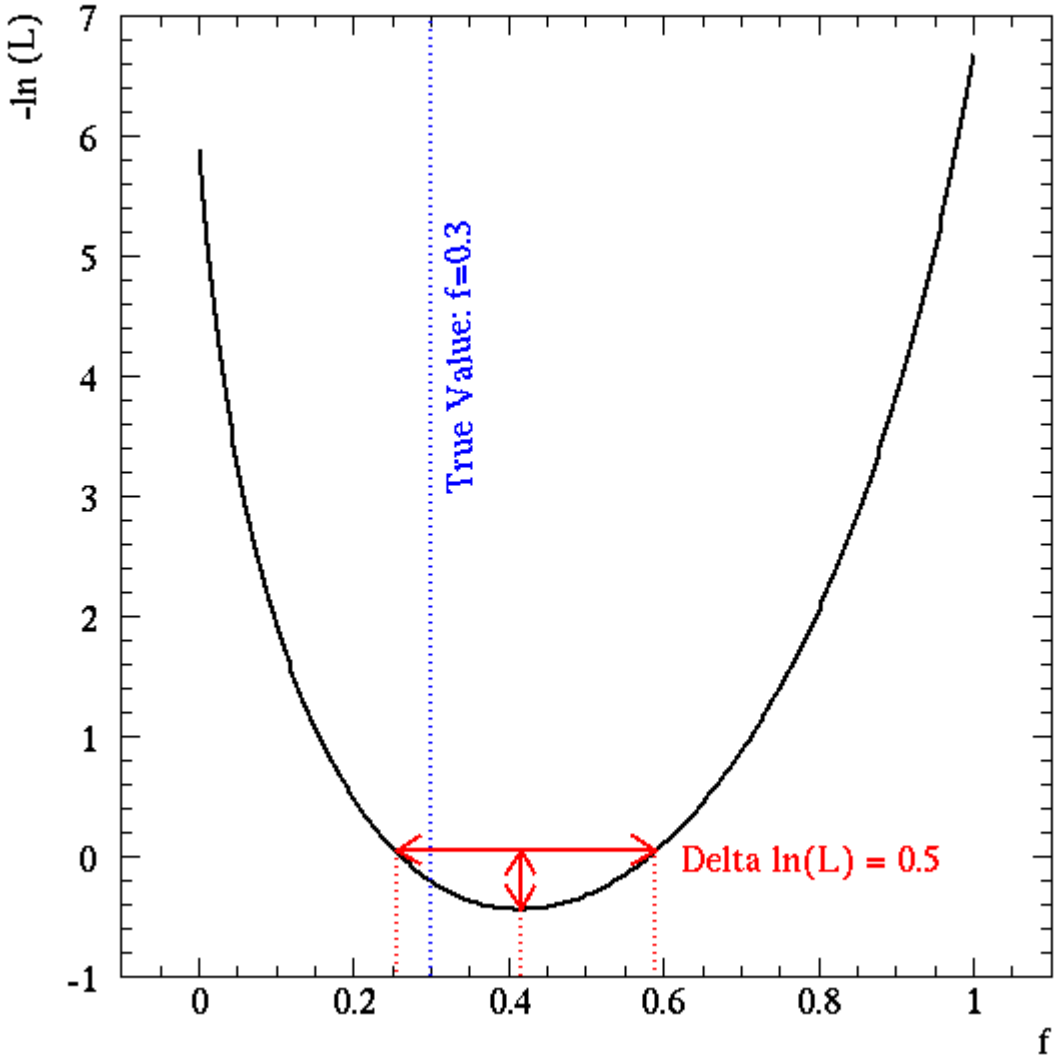
So the ML estimator is actually the peak location for the Bayesian posterior PDF assuming a flat prior P(a|I)=1.

The log likelihood is related to the Bayesian PDF by:

$$P(a|D,I) = \exp[\ \ln(L(a))\ ]$$

This way of viewing the log likelihood as the logarithm of a Bayesian PDF with uniform prior is an excellent way to intuitively understand many features of the ML method.

Errors on ML estimators



In the limit of large N, the log likelihood becomes parabolic (by CLT). Comparing to ln(L) for a simple Gaussian:

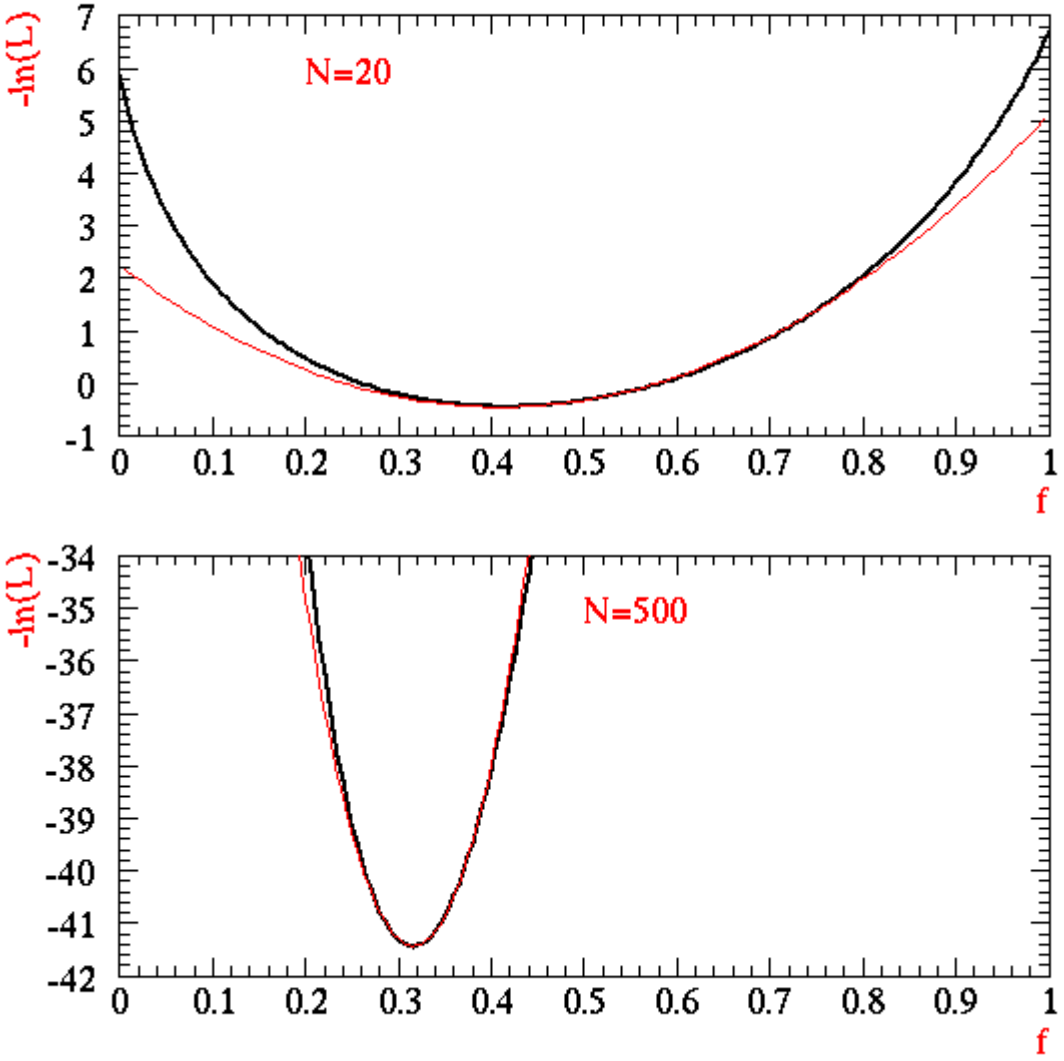
$$-\ln L = L_0 + \frac{1}{2}\left(\frac{f - \langle f \rangle}{\sigma_f}\right)^2$$

it is natural to identify the 1σ range on the parameter by the points as which $\Delta \ln(L)=½$.

2σ range: $\Delta \ln(L)=½(2)^2=2$
3σ range: $\Delta \ln(L)=½(3)^2=4.5$

This is done even when the likelihood isn't parabolic (although at some peril).

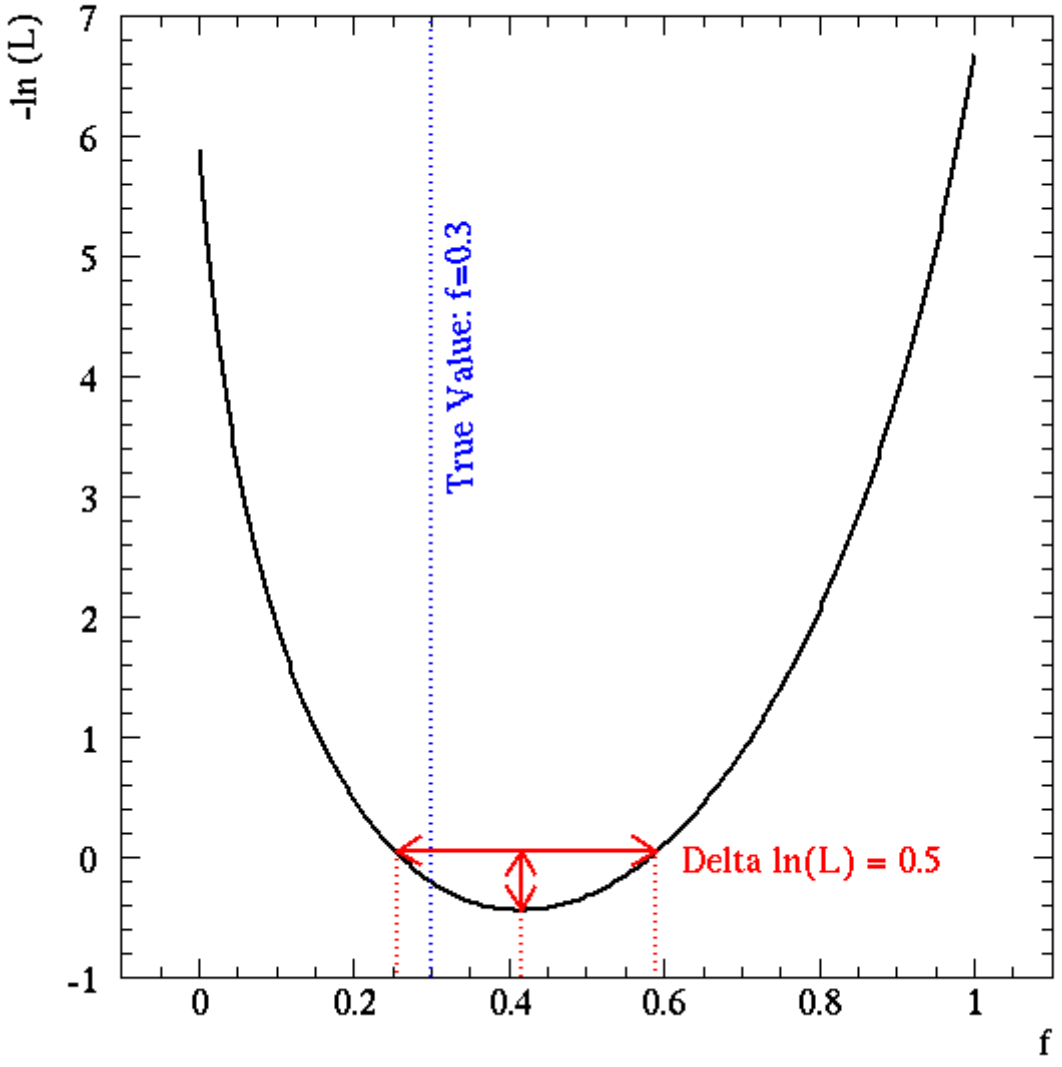
Parabolicity of the log likelihood



In general the log likelihood becomes more parabolic as N gets larger. The graphs at the right show the negative log likelihoods for our example problem for N=20 and N=500. The red curves are parabolic fits around the minimum.

How large does N have to be before the parabolic approximation is good? That depends on the problem---try graphing -ln(L) vs your parameter to see how parabolic it is.

Asymmetric errors from ML estimators



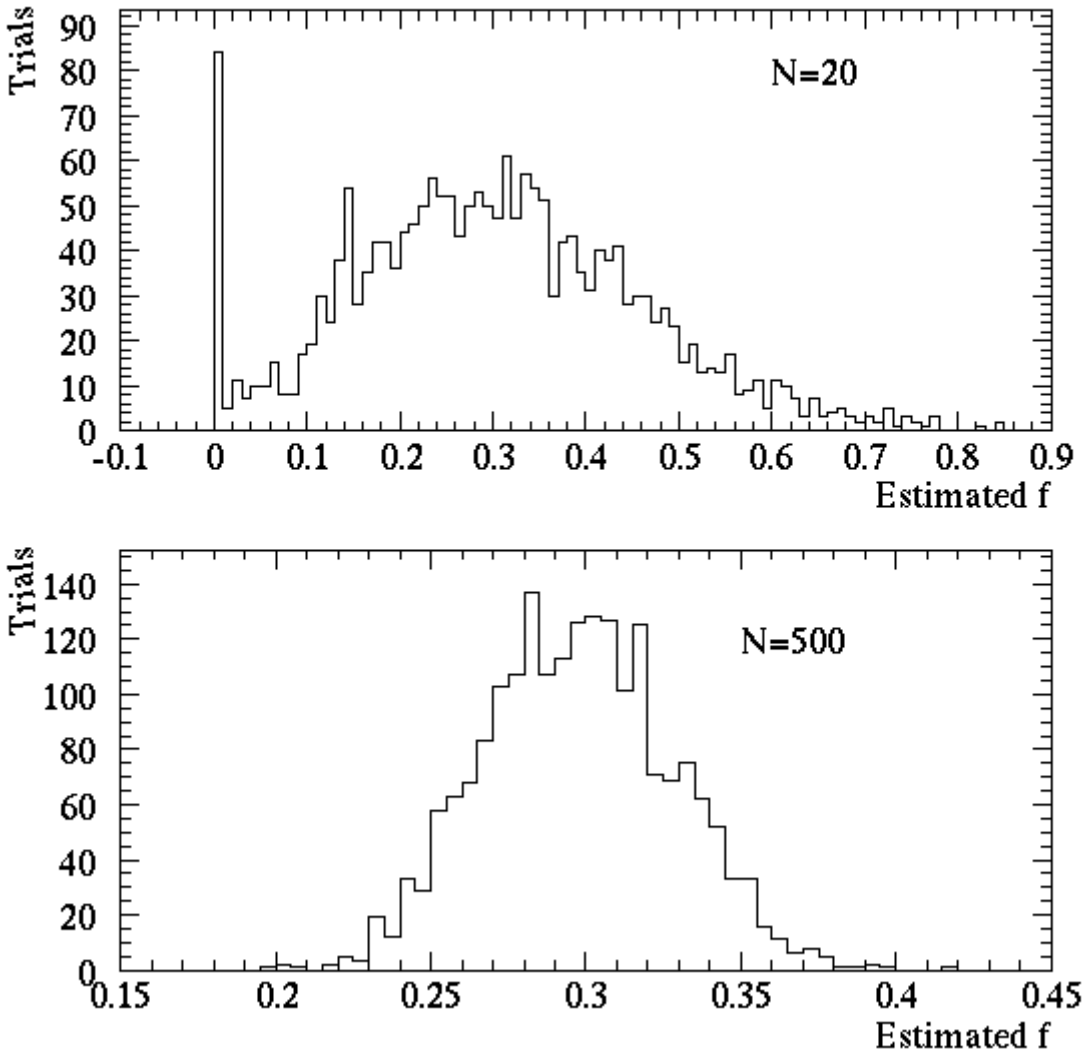
Even when the log likelihood is not Gaussian, it's nearly universal to define the 1σ range by Δ ln(L)=½. This can result in asymmetric error bars, such as:

$$0.41^{+0.17}_{-0.15}$$

The justification often given for this is that one could always reparametrize the estimated quantity into one which does have a parabolic likelihood. Since ML estimators are supposed to be invariant under reparametrizations, you could then transform back to get asymmetric errors.

Does this procedure actually work?

Coverage of ML estimator errors

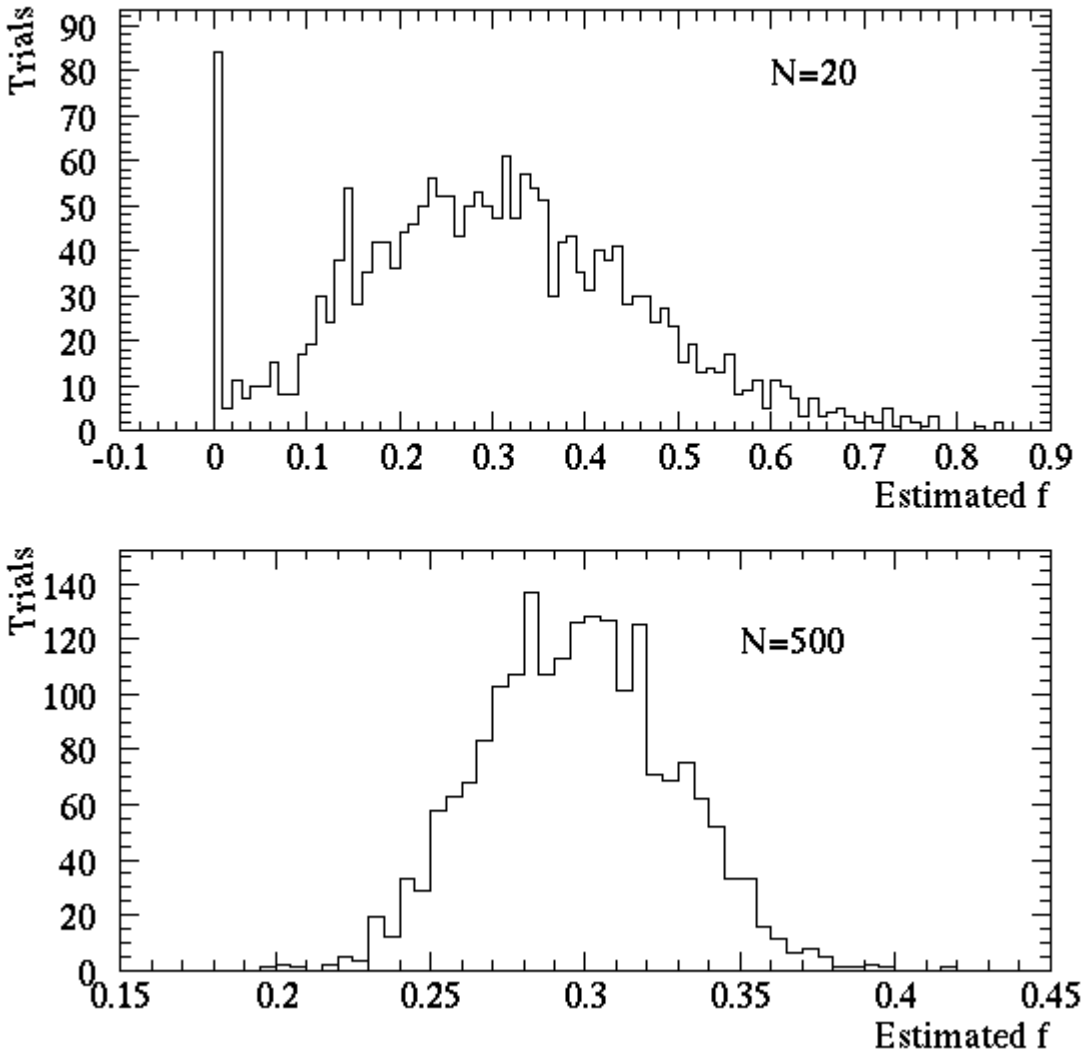


Distribution of ML estimators for two N values

What do we really want the ML error bars to mean? Ideally, the 1σ range would mean that the true value has 68% chance of being within that range.

N	Fraction of time 1σ range includes true value
5	56.7%
10	64.8%
20	68.0%
500	67.0%

Errors on ML estimators



Simulation is the best way to estimate the true error range on an ML estimator: assume a true value for the parameter, and simulate a few hundred experiments, then calculate ML estimates for each.

N=20:
Range from likelihood function: -0.16 / +0.17
RMS of simulation: 0.16

N=500:
Range from likelihood function: -0.030 / +0.035
RMS of simulation: 0.030

Likelihood functions of multiple parameters

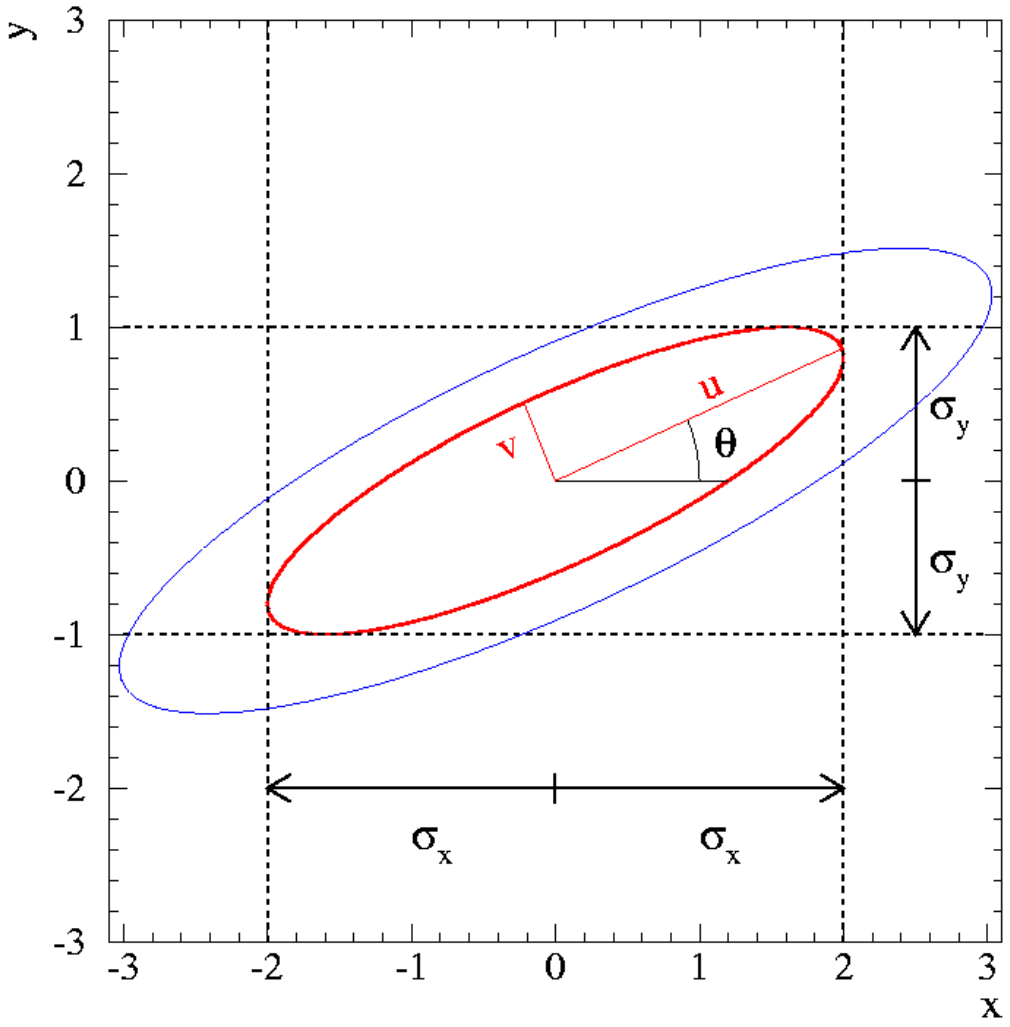
Often there is more than one free parameter. To handle this, we simply minimize the negative log likelihood over all free parameters.

$$\frac{\partial \ln L(x_1 \ldots x_N | a_1 \ldots a_m)}{\partial a_j} = 0$$

Errors determined by (in the Gaussian approximation):

$$cov^{-1}(a_i, a_j) = -\frac{\partial^2 \ln L}{\partial a_i \partial a_j} \quad \text{evaluated at minimum}$$

Error contours for multiple parameters



We can also find the errors on parameters by drawing contours on
$\Delta \ln L$.

1σ range on a single parameter a: the smallest and largest values of a that give $\Delta \ln L = ½$, minimizing ln L over all other parameters.

But to get joint error contours, must use different values of $\Delta \ln L$ (see Num Rec Sec 15.6):

	m=1	m=2	m=3
68.00%	0.5	1.15	1.77
90.00%	1.36	2.31	3.13
95.40%	2	3.09	4.01
99.00%	3.32	4.61	5.65

Extended maximum likelihood estimators

Sometimes the number of observed events is not fixed, but also contains information about the unknown parameters. For example, maybe we want to fit for the rate. For this purpose we can use the extended maximum likelihood method.

Normal ML method:

$$\int P(x|\vec{\alpha})\,dx = 1$$

Extended ML method:

$$\int Q(x|\vec{\alpha})\,dx = \nu = \text{predicted number of events}$$

Extended maximum likelihood estimators

$$\int P(x|\vec{\alpha})\,dx = 1$$

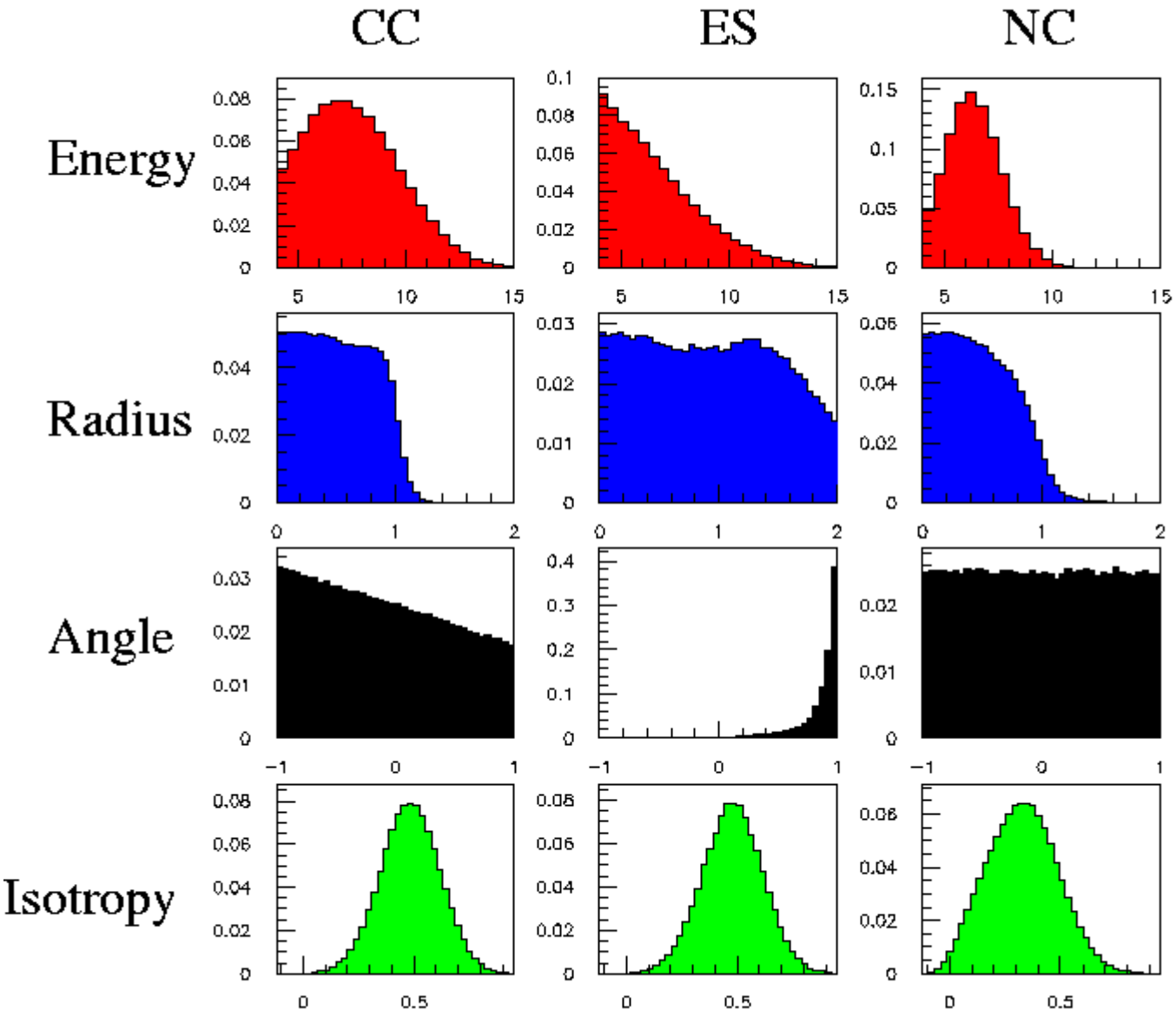
$$\text{Likelihood} = \frac{e^{-\nu}\nu^N}{N!}\cdot\prod_{i=1}^{N} P(x_i|\vec{\alpha}) \quad [\text{note that } \nu = \nu(\vec{\alpha})]$$

$$\ln L(\vec{\alpha}) = \sum \ln P(x_i|\vec{\alpha}) - \nu(\vec{\alpha}) + N\ln\nu(\vec{\alpha})$$

$$= \sum \ln[\nu(\vec{\alpha})\,P(x_i|\vec{\alpha})] - \nu(\vec{\alpha})$$

The argument of the logarithm is the number density of events predicted at x_i. The second term (outside the summation sign) is the total predicted number of events.

Example of the extended maximum likelihood in action: SNO flux fits



$$P(E,R,\Theta,\beta) = CC\ P_{CC}(E,R,\Theta,\beta) + ES\ P_{ES}(E,R,\Theta,\beta) + NC\ P_{NC}(E,R,\Theta,\beta)$$

Fit for the numbers of CC, ES, and NC events.

Careful: because every event must be either CC, ES, or NC, the three event totals are anti-correlated with each other.