Likelihood Fits

Craig Blocker
Brandeis
August 23, 2004

Outline

I. What is the question?
II. Likelihood Basics
III. Mathematical Properties
IV. Uncertainties on Parameters
V. Miscellaneous
VI. Goodness of Fit
VII. Comparison With Other Fit Methods

What is the Question?

Often in HEP, we make a series of measurements and wish to deduce the value of a fundamental parameter.

For example, we may measure the mass of many B ® J/y K_S decays and then wish to get the best estimate of the B mass.

Or we might measure the efficiency for detecting such events as a function of momentum and then wish to derive a functional form.

The question is: What is the best way to do this?
(And what does "best" mean?)

Likelihood Method

P(X|a) º Probability of measuring X on a given event.
a is a parameter or set of parameters on which P depends.

Suppose we make a series of measurements, yielding a set of X_i's.
The likelihood function is defined as

$$L = \prod_{i=1}^{N} \mathbf{P}(\mathbf{X}_i \mid a)$$

The value of a that maximizes *L* is known as the Maximum Likelihood Estimator (MLE) of a, which we will denote as a^*.

Note that we often work with ln(*L*).

Example

Suppose we have N measurements of a variable x, which we believe is Gaussian, and wish to get the best estimate of the mean and width. **[Note that this and other examples will be doable analytically - usually we must use numerical methods].**

$$P(x \mid m, s) = \frac{1}{\sqrt{2p}\, s} e^{-\frac{(x-m)^2}{2s^2}}$$

$$\ln L = \sum_{i=1}^{N} \ln P(x_i) = -N \ln\left(\sqrt{2p}\, s\right) - \sum_{i=1}^{N} \frac{(x_i - m)^2}{s^2}$$

Maximizing with respect to m and s gives

$$m^* = \frac{\sum x_i}{N} = \bar{x}$$

$$s^{*2} = \frac{\sum (x_i - m^*)^2}{N} = \overline{x^2} - \bar{x}^2$$

Discussed later.

Warning

Neither $L(a)$ nor $\ln(L)$ is a probability distribution for a.

A frequentist would say that such a statement is just nonsense, since parameters of nature have definite values.

A Bayesian would say that you can convert $L(a)$ to a "probability" distribution in a by applying Bayes thereom, which includes a prior probability distribution for a.

Bayesian versus Frequentist statistics is a can of worms that I won't open further in this talk.

Bias, Consistency, and Efficiency

What does "best" mean?

We want estimator to be close to the true value.

Unbiased Þ $\overline{a^*} = a_0$

Consistent Þ **Unbiased for large N**

Efficient Þ $\overline{(a^* - a_0)^2}$ **is minimal for large N** WOW!

Maximum Likelihood Estimators are NOT necessarily unbiased but are consistent and efficient for large N.
This makes MLE's powerful and popular (although we must be aware that we may not be in the large N limit).

Bias Example

Consider again, the mean and width of a Gaussian.

$$\overline{m^*} = \overline{\frac{1}{N}\sum x_i} = \frac{1}{N}\sum \overline{x_i} = \frac{1}{N}\sum m_0 = m_0$$

$$\overline{s^{*2}} = \overline{\frac{1}{N}\sum \left(x_i - m^*\right)^2} = \frac{N}{N-1} s_0^2$$

Note that the MLE of the mean is unbiased, but for the width is not (although it is consistent). However,

$$\tilde{s}^2 = \frac{\sum \left(x_i - m^*\right)^2}{N-1}$$

is unbiased.

In this case, we could find the bias analytically - in most cases we must look for it numerically.

Bias Example 2

Bias can depend upon choice of parameter. Consider an exponential lifetime distribution. We can use either the averge lifetime τ or the decay width Γ as the parameter.

$$P(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}} = \Gamma e^{-\Gamma t}$$

$$\tau^* = \frac{\sum t_i}{N}$$ **is unbiased.**

$$\Gamma^* = \frac{N}{\sum t_i}$$ **is biased.**

Uncertainty on Parameters

Just as important as getting an estimate of a parameter is knowing the uncertainty of that estimate.

The maximum likelihood method also provides an estimate of the uncertainty.

For one parameter, L becomes Gaussian for large N. Thus,

$$\ln L \approx \ln L^* + \frac{1}{2}\left.\frac{\partial^2 \ln L}{\partial a^2}\right|_{a=a^*}(a - a^*)^2$$

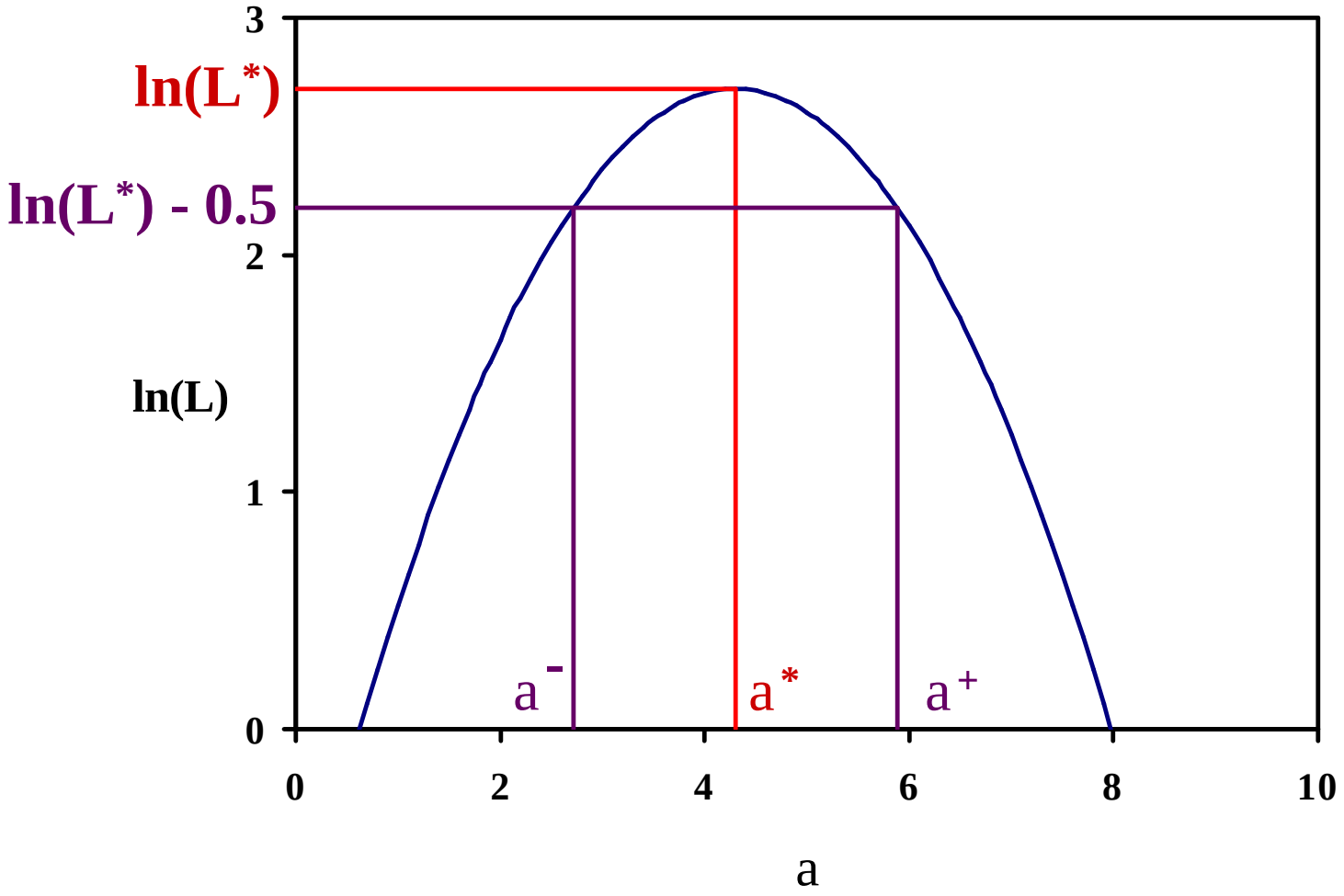
$$\Rightarrow \Delta a^2 \approx \overline{(a^* - a_0)^2} = -\frac{1}{\left.\frac{\partial^2 \ln L}{\partial a^2}\right|_{a=a^*}}$$

We usually write this as $\alpha = \alpha^* \pm \Delta\alpha$

If $a = a^* \pm \Delta a$, $\ln L = \ln L^* - \frac{1}{2}$

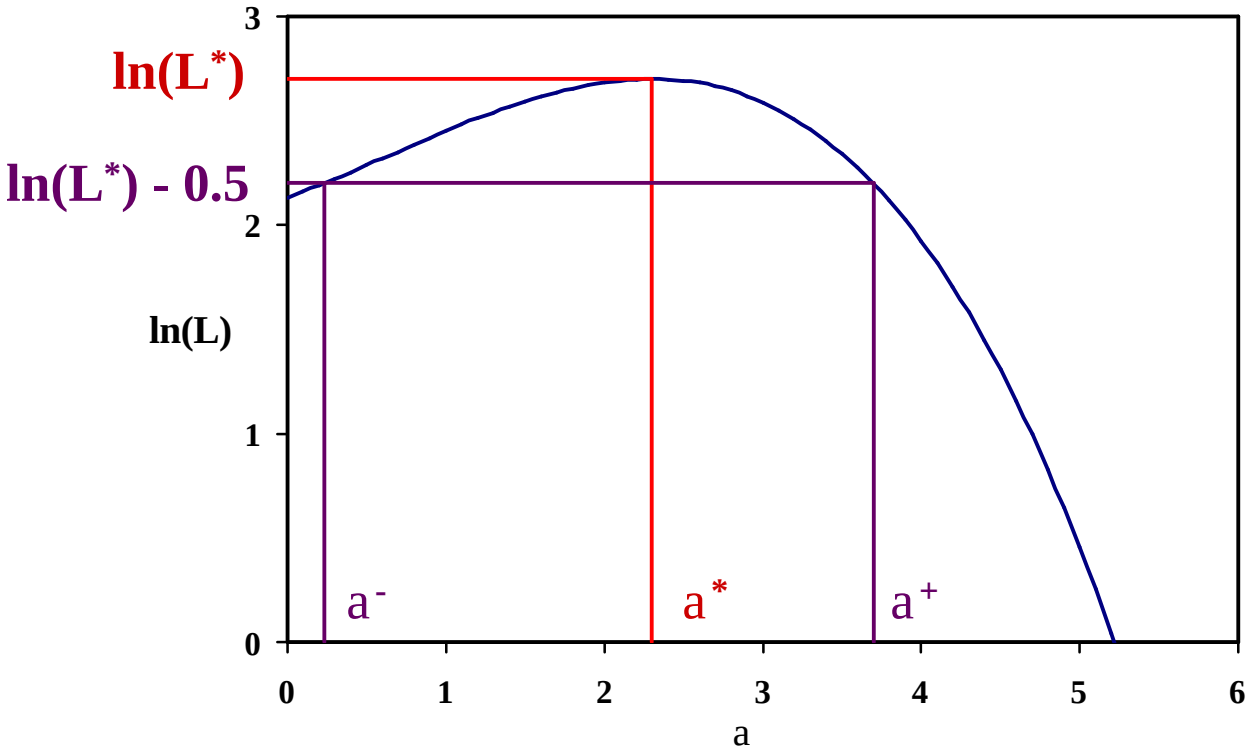
Note that this is a statement about the probability of the measurements, not the probability of the true value.

Uncertainty Example



Asymmetric Uncertainties

Sometimes ln$L$ may not be parabolic and there may be asymmetric uncertainties.



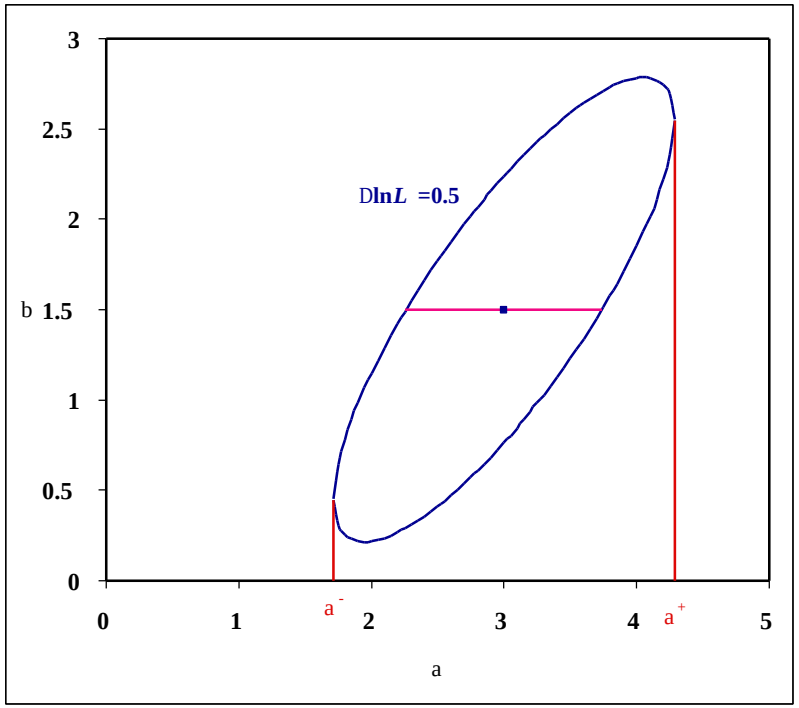
We write

$$a = a^{*}{}^{+\Delta a_+}_{-\Delta a_-}$$

Note: the D ln$L$ = 1/2 interval does NOT always give a 68% confidence interval (see counterexample in handout).

Correlations

If there are multiple parameters, things are more complicated due to possible correlations.



For example, a fit to the linear function y = a x + b will have correlations between parameters a and b

Numerically, $a^{\pm}$ are given by where $\Delta\ln(L) = 1/2$ and ln(L) is maximized wrt other parameters

Covariance matrix V is given by $V_{ij} = \overline{(a_i - \overline{a_i})(a_j - \overline{a_j})}$

V is equal to U^{-1}, where $U_{ij} = -\dfrac{\partial^2 \ln L}{\partial a_i \partial a_j}$

Normalization

Sometimes, people will say they don't need to normalize their probability distributions. This is sometimes true.

For the Gaussian example, if we omitted the normalization factor of $1/\sqrt{2\pi}s$ we get the mean correct but not the width. In general, if the normalization depends on any of the parameters of interest, it must be included.

My advice is always normalize (and always check the normalization).

Extended Likelihood

Suppose we have a Gaussian mass distribution with a flat background and wish to determine the number of events in the Gaussian.

$$P = f_S \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(M - M_0)^2}{2\sigma^2}} + (1 - f_S)\frac{1}{\Delta M}$$

where f_S is the fraction of signal events and ΔM is the mass range of the fit.

We can fit for f_S and get Δf_S. $N f_S$ is a good estimate of the number of events in the Gaussian, but $N \Delta f_S$ is not a good estimate of the variation on the number of signal events.

We can fix this by adding a Poisson term in the total number of events. This is called an Extended Likelihood fit.

We could also use $\Delta N_S^2 = N^2 \Delta f_S^2 + f_S^2 \Delta N^2 = N^2 \Delta f_S^2 + N f_S^2$

Extended Likelihood 2

Instead of f_S, we use m_S and m_{BG}, the expected number of signal and background events. N is the observed total number of events.

$$L = \frac{e^{-(m_S + m_{BG})}(m_S + m_{BG})^N}{N!} \prod_{i=1}^{N} \left[\frac{m_S}{m_S + m_{BG}} G(M_i | M_0, s) + \frac{m_{BG}}{m_S + m_{BG}} \frac{1}{DM} \right]$$

$$= \frac{e^{-(m_S + m_{BG})}}{N!} \prod_{i=1}^{N} \left[m_S G(M_i | M_0, s) + m_{BG} \frac{1}{DM} \right]$$

$$\ln(L) = -m_S - m_{BG} - \ln(N!) + \sum_i \ln\left[m_S G + \frac{m_{BG}}{DM} \right]$$

If you are not interested in the uncertainty on N_S (for example, your are measuring a lifetime and not a cross section), I recommend not doing an extended likelihood fit.

Constrained Fits

Suppose there is a parameter in the likelihood that is somewhat known from elsewhere. This information can be incorporated in the fit.

For example, we are fitting for the mass of a particle decay with resolution σ. Suppose the Particle Data Book lists the mass as $M_0 \pm \sigma_M$. We can incorporate this into the likelihood function as

$$L = \frac{1}{\sqrt{2\pi}\,\sigma_M} e^{-\frac{(M - M_0)^2}{2\sigma_M^2}} \prod_i^N \left[\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(m_i - M)^2}{2\sigma^2}} \right]$$

This is known as a constrained fit.

Constrained Fits 2

Let ΔM be the uncertainty on M that could be determined by the fit alone.

If $\Delta M \gg \sigma_M$, constraint will dominate, and you might as well just fix M to M_0. For example, you never see a constrained fit to $\hbar$ in an HEP experiment.

If $\Delta M \ll \sigma_M$, constraint does very little. You have a better measurement than the PDG. You should do an unconstrained fit and PUBLISH.

Constrained fit is most useful if σ_M and ΔM are comparable.

Simple Monte Carlo Tests

It is possible to write simple, short, fast Monte Carlo programs that generate data for fitting. Can then look at fit values, uncertainties, and pulls. These are often called "toy" Monte Carlos to differentiate them from complicated event and detector simulation programs.

★ Tests likelihood function.
★ Tests for bias.
★ Tests that uncertainty from fit is correct.

This does NOT test the correctness of the model of the data. For example, if you think that some data is Gaussian distributed, but it is really Lorentzian, then the simple Monte Carlo test will not reveal this.

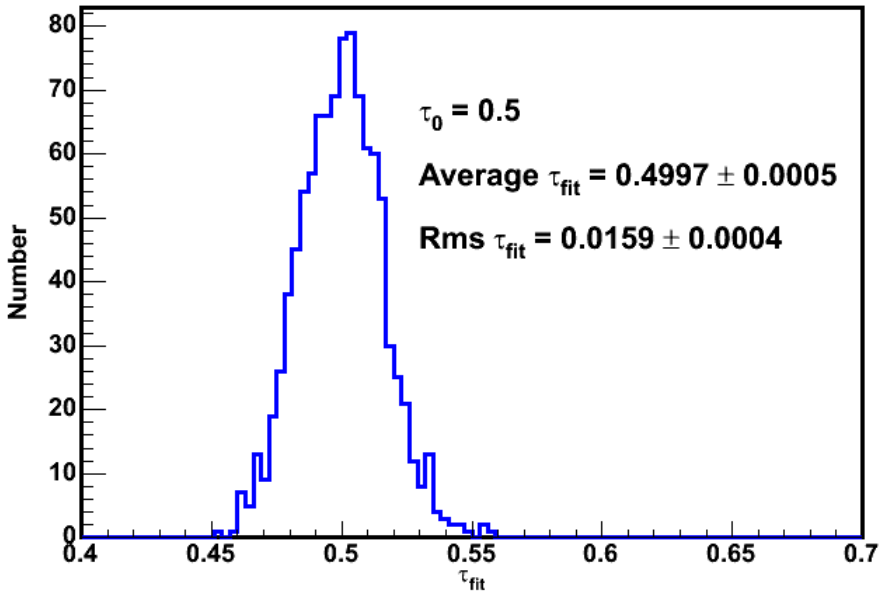
Simple Monte Carlo Tests 2

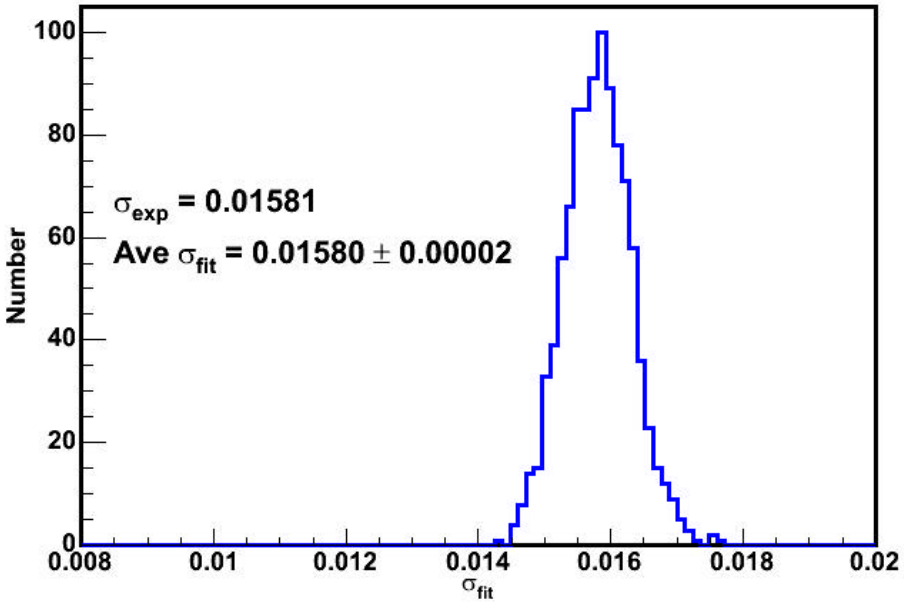
Generate exponential (τ = 0.5 and N = 1000).

Fit.

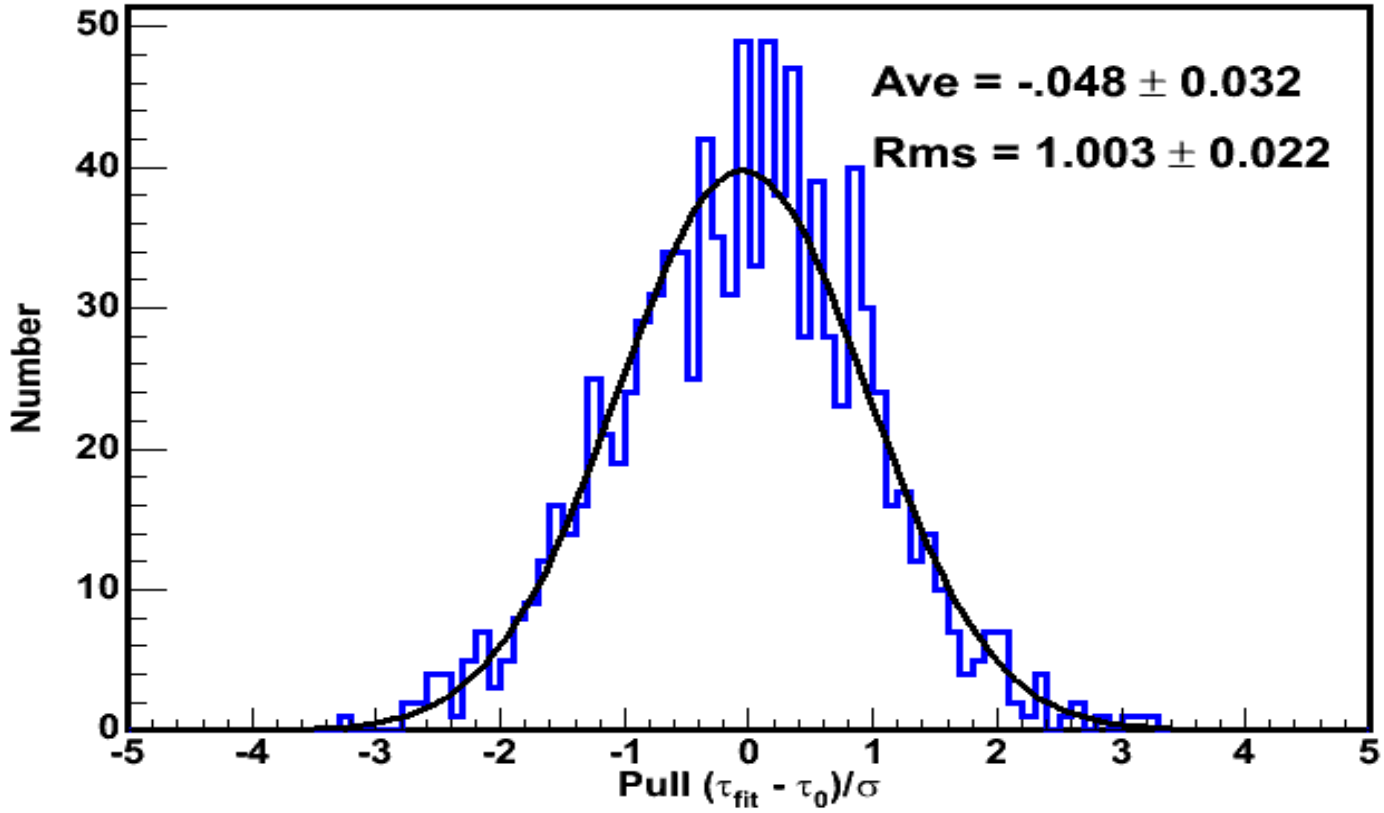
Repeat many times (1000 times here).

Histrogram τ, s_τ, and pulls.





Simple Monte Carlo Tests 3



Goodness of Fit

Unfortunately, the likelihood method does not, in general, provide a measure of the goodness of fit (as a χ^2 fit does).

For example, consider fitting lifetime data to an exponential.

$$L = \prod_i^N \frac{1}{t} e^{-\frac{t_i}{t}}$$

$$t^* = \frac{\sum t_i}{N}$$

$$L(t^*) = -N\left(1 + \ln \frac{\sum t_i}{N}\right)$$

Thus the value of *L* at the maximum depends only on the number of events and average value of the data.

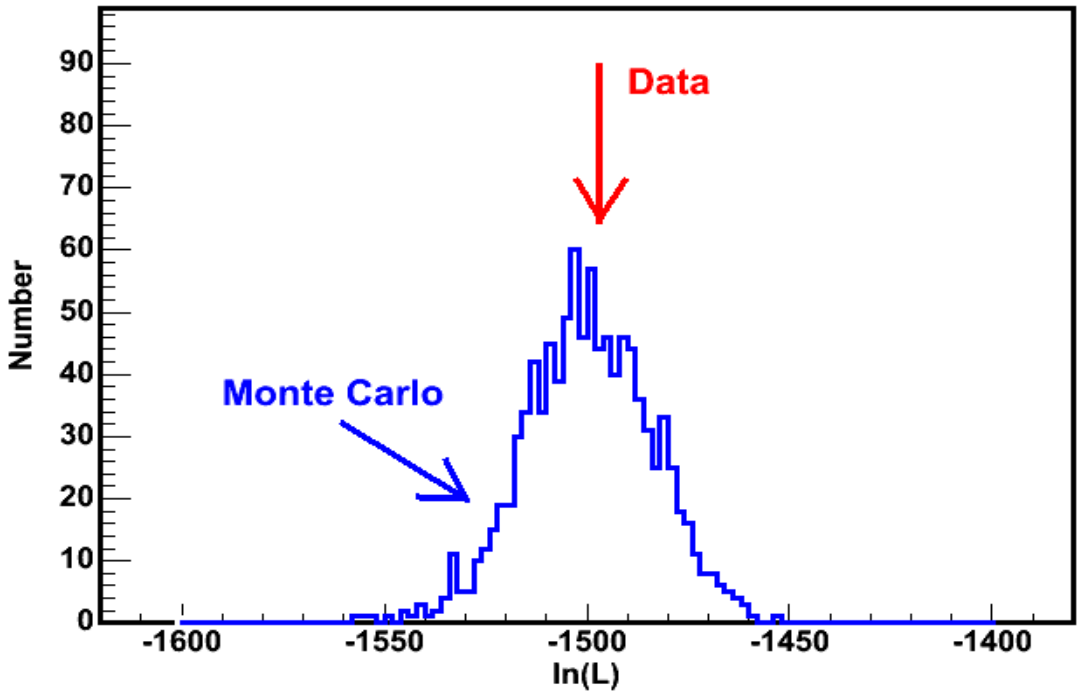
Goodness of Fit 2

Fit to exponential

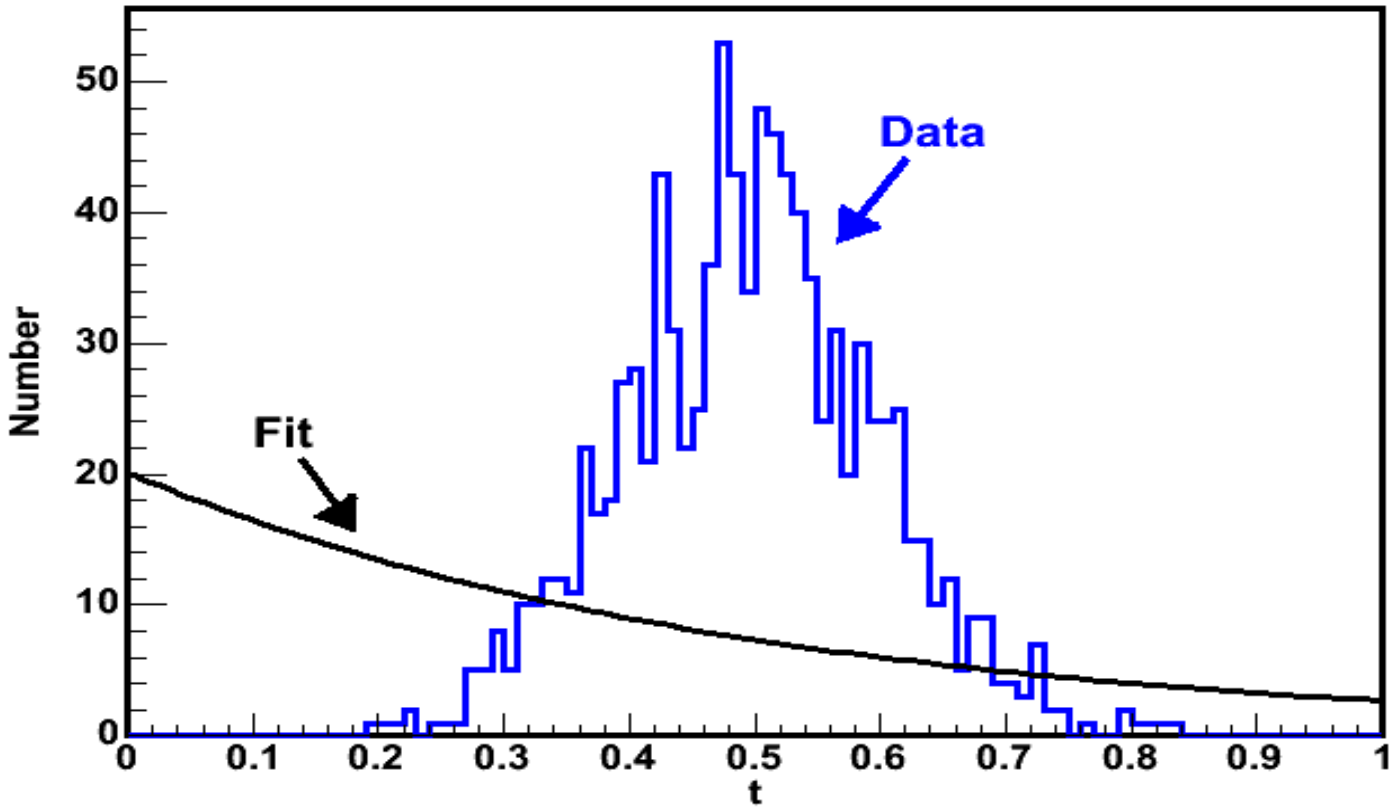
Plot log(L*) for

(1) exponential Monte Carlo and

(2) Gaussian data



Goodness of Fit 3



Other Types of Fits

Chi-square:

If data is binned and uncertainties are Gaussian, then maximum likelihood is equivalent to a c^2 fit.

Binned Likelihood:

If data is binned and not Gaussian, can still do a binned likelihood fit. Common case is when data are Poisson distributed.

$$P_i = \frac{e^{-\bar{m}_i}\left(\bar{m}_i\right)^{n_i}}{n_i!}$$

$$\ln L = \sum_{\text{bins}} \ln P_i$$

Comparison of Fits

Chi-square:

- **Goodness of fit.**
- **Can plot function with binned data.**
- **Data should be Gaussian, in particular, c^2 doesn't work well with bins with a small number of events.**

Binned likelihood:

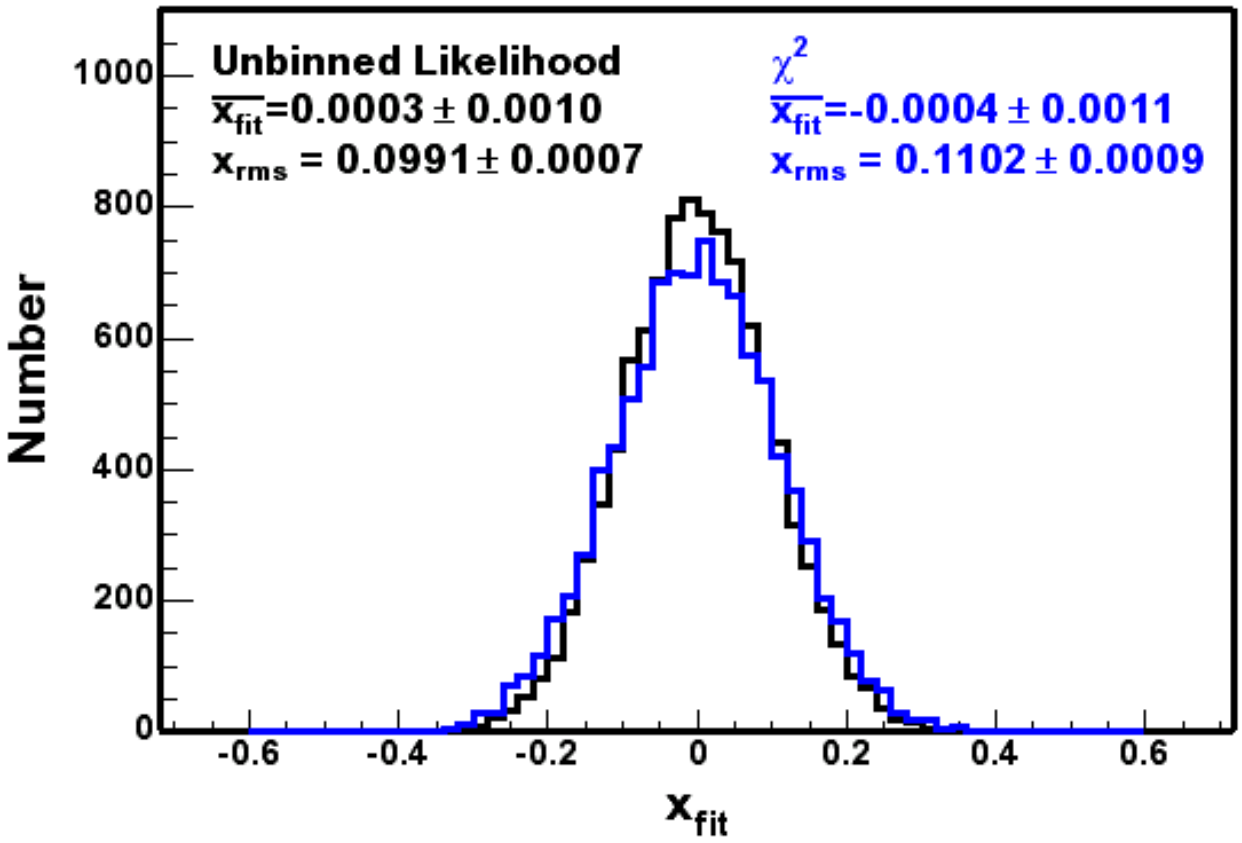
- **Goodness of fit**
- **Can plot function with binned data.**
- **Still need to be careful of bins with small number of events (don't add in too many zero bins).**

Unbinned likelihood:

- **Usually most powerful.**
- **Don't need to bin data.**
- **Works well for multi-dimensional data.**
- **No goodness of fit estimate.**
- **Can't plot fit with data (unless you bin data).**

Comparison of Fits 2

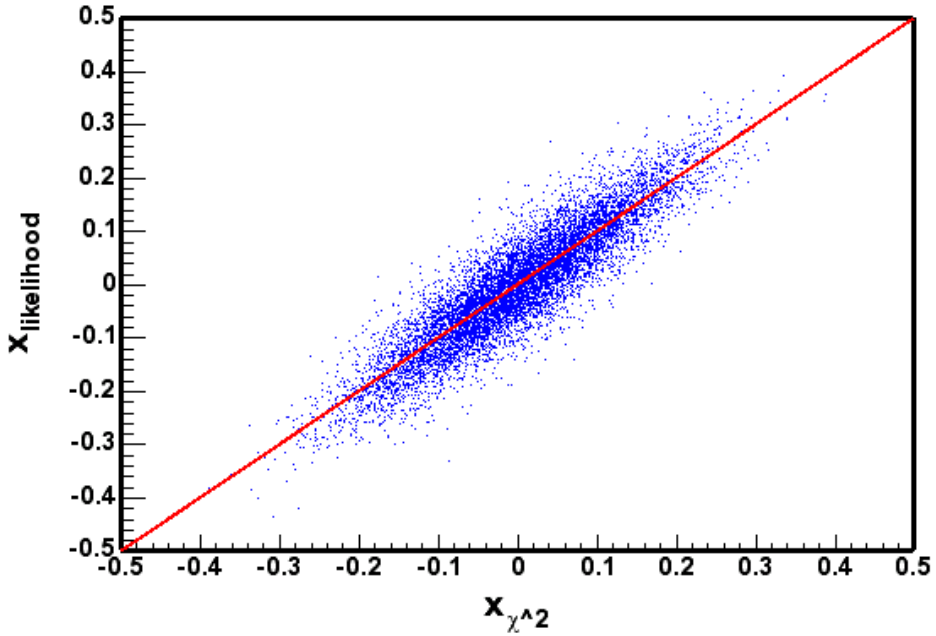
Generate 100 values for Gaussian with $\mu = 0$, $\sigma = 1$.
Fit unbinned likelihood and χ^2 to SAME data.
Repeat 10,000 times.



Both are unbiased.

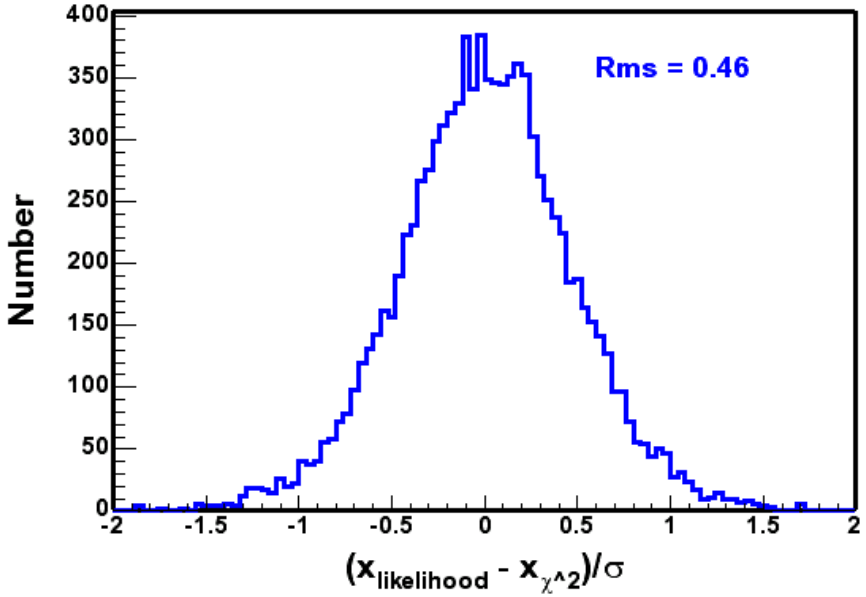
Unbinned likelihood is more efficient.

Comparison of Fits 3

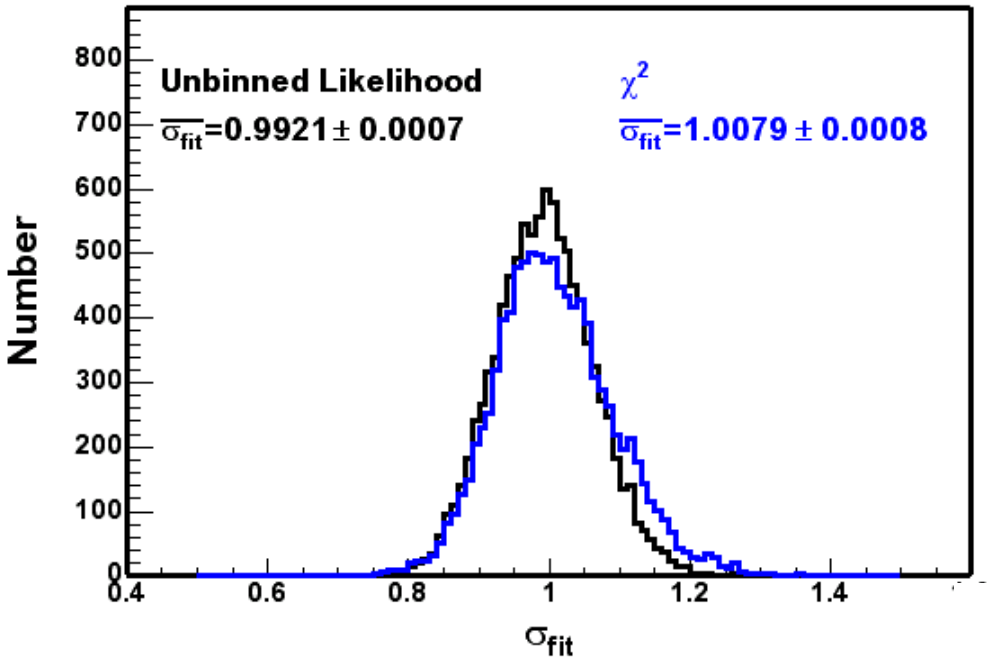


Fit values are correlated, but not completely.

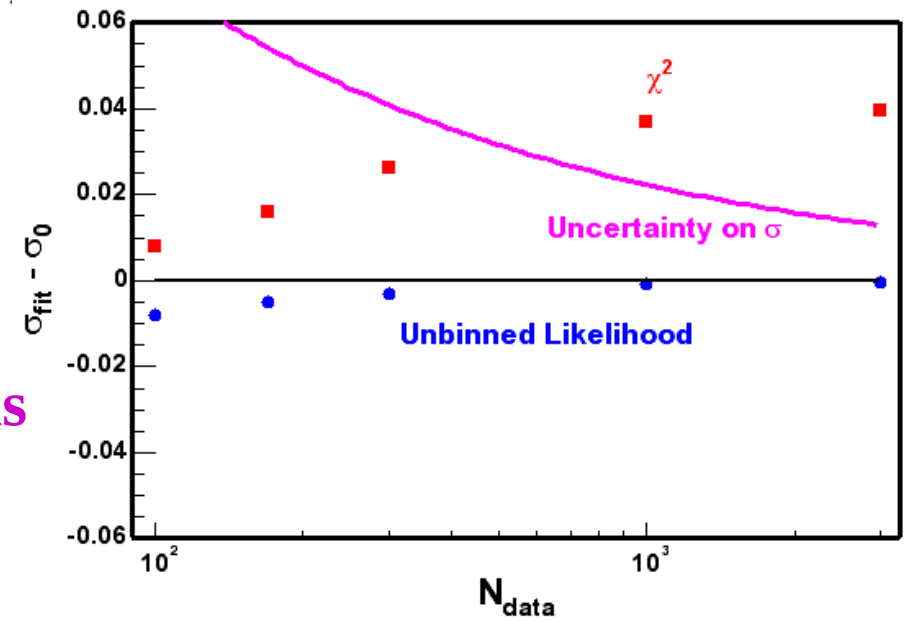
Difference is of the order of half of the uncertainty.



Comparison of Fits 4



Fit to width is biased for both.



But, unbinned likelihood widths tend to true value for large N.

Numerical Methods

Even slight complications to the probability make analytic methods intractable.

Also, likelihood fits often have many parameters (perhaps scores) and can't be done analytically.

However, numerical methods are still very effective.

MINUIT is a powerful program from CERN for doing maximum likelihood fits (see references in handout).

Systematic Uncertainties

When fitting for one parameter, there often are other parameters that are imperfectly known.

It is tempting to estimate the systematic uncertainty due to these parameters by varying them and redoing the fit.

Because of statistical variations, this overestimates the systematic uncertainty (often called double counting).

Best way to estimate such systematics is probably with a high statistics Monte Carlo program.

Potentially Interesting Web Sites

CDF Statistics Committee page:
www-cdf.fnal.gov/physics/statistics/statistics_home.html

Lectures by Louis Lyons:
www-ppd.fnal.gov/EPPOffice-w/Academic_Lectures

Summary

Maximum Likelihood methods are a powerful tool for extracting measured parameters from data.

However, it is important to understand their proper use and avoid potential problems.