Bayesian and Frequentist Parameter Estimation Techniques

Lecture #2

Scott Oser
INSS 2023

Thomas Bayes

Bayes' Theorem adapted for parameter estimation

$$P(H|D,I) = \frac{P(H|I)\,P(D|H,I)}{P(D|I)}$$

H = a hypothesis (e.g. “the Higgs mass is 127 GeV”)
I = prior knowledge or data about H
D = the data

P(H|I) = the “prior probability” for H

P(D|H,I) = the probability of measuring D, given H and I. Also called the “likelihood”

P(D|I) = a normalizing constant: the probability that we would have measured D anyway, averaged over values of H.

End result: a posterior probability distribution for the parameter(s).

An example with parameter estimation: coin flip

Someone hands you a coin and asks you to estimate the *p* value for the coin (probability of getting heads on any given flip).

You flip the coin 20 times and get 15 heads.

What do you conclude?

Bayesian coin flipping

Someone hands you a coin and asks you to estimate the *p* value for the coin (probability of getting heads on any given flip).

You flip the coin 20 times and get 15 heads.

What do you conclude?

$$P(H \mid D,I) = \frac{P(H \mid I)\,P(D \mid H,I)}{P(D \mid I)}$$

Here H is the hypothesis that p has some particular value. To proceed we must evaluate each term.

Evaluating the terms in the Bayesian coin flip

$$P(H|D,I) = \frac{P(H|I)\,P(D|H,I)}{P(D|I)}$$

First, some notation. Let me use *p* in place of H.

Prior: let's assume a uniform prior for *p*. So P(H|I) = P(*p*) = 1.

Likelihood factor: P(D|*p*). This is the probability of observing our data, given p. We model this as a binomial distribution:

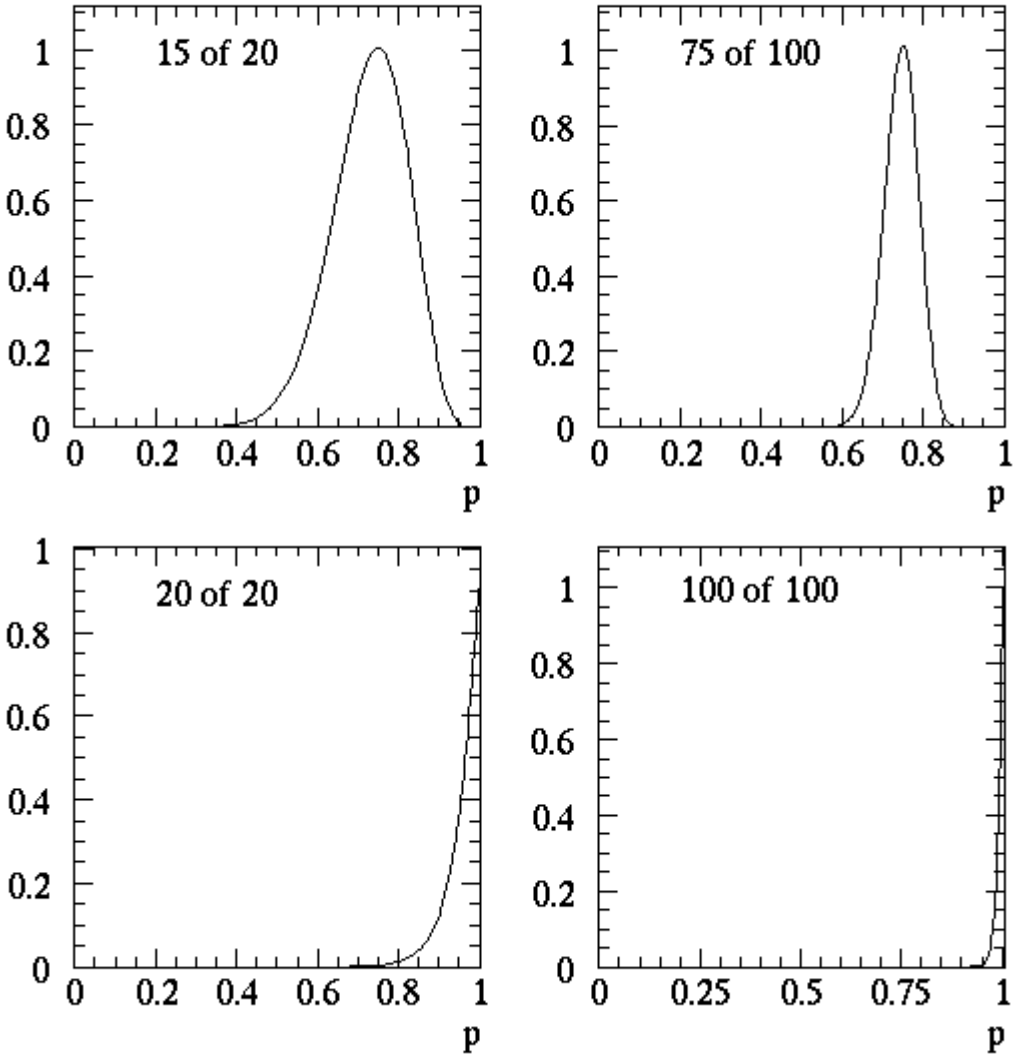
$$P(D|p) = \frac{N!}{m!(N-m)!}\,p^m(1-p)^{N-m}$$

Finally P(D|I). This is the probability of observing the data, summed over all hypotheses (here, all possible values of *p*).

$$P(D|I) = \int_0^1 dp\,P(p)\,\frac{N!}{m!(N-m)!}\,p^m(1-p)^{N-m}$$

Solution for P(*p*|D,I): uniform prior

$$P(p\,|\,D) \propto P(p)\,P(D\,|\,p) = \frac{N!}{m!\,(N-m)!}\,p^{m}(1-p)^{N-m}$$



Bayesian coin flip: alternate prior

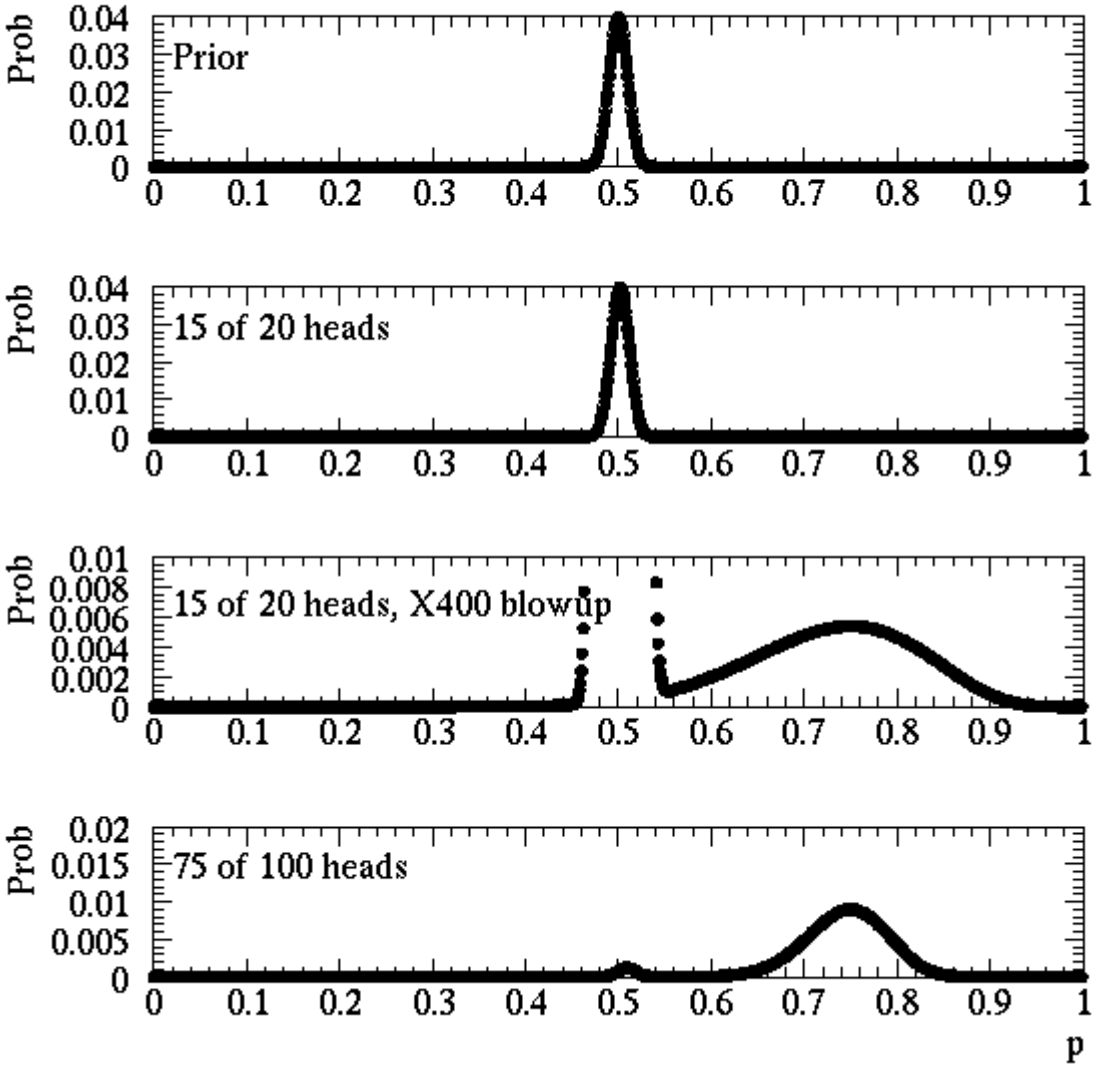
If a friend hands you a coin in the lunchroom, is it really reasonable to assume a uniform prior for *p*? Unbalanced coins must be really rare!

Consider a more plausible prior:

1) You're 99.9% sure this is a normal coin. A normal coin has *p=0.5*. But even normal coins might be a little off-kilter, so model its distribution as a Gaussian with mean 0.5 and width σ=0.01.
2) There's a 0.1% chance this is a trick coin. If so, you have no idea what its true *p* value would be, so use a uniform distribution.

$$P(p) = 0.999 \times \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(p-0.5)^2}{2\sigma^2}} + 0.001 \times 1$$

Solution for P(*p*|D,I): more realistic prior



Prob in peak at 0.5 = 0.999

Prob in peak at 0.5 = 0.997

Prob in peak at 0.5 = 0.030

Dependence on choice of prior

Clearly you get a different answer depending on which prior you choose! This is a big point of controversy for critics.

A Bayesian's reply: “Tough.”

In Bayesian analysis, dependence on choice of priors is a feature, not a bug. The prior is a quantitative means of incorporating external information about the quantities being measured. If the answer depends strongly on the choice of prior, this just means that the data is not very constraining.

In contrast, classical frequentist analysis doesn't require you to spell out assumptions so clearly---what are you implicitly assuming or ignoring?

Good habits for Bayesian analysis:

- be explicit about your choice of prior, and justify it
- try out different priors, and show how result changes

Contrast with frequentist approach

A frequentist would use the data to directly estimate *p* from the data, without invoking prior. Best estimate is *p*=15/20=0.75.

Frequentist would probably try to assign an "error bar" to this value. Perhaps noting that variance of binomial is *Np(1-p)*, we could calculate Var=20(0.75)(0.25)=3.75, or σ=sqrt(Var)=1.94. So the error on *p* might be 1.94/20 = 0.097, so *p*=0.75 ± 0.10. (What would a frequentist do if she observed 20/20 heads?)

But interpretation is very different. Frequentist would not speak of the probability of various *p* values being true. Instead we talk about whether the data is more likely or less likely given any specific *p* value. Very roundabout way of speaking!

Note that the *p* value estimation did not:

- yield a probability distribution for *p*
- did not incorporate any background information (eg. the fact that almost any coin you regularly encounter will be a fair coin)

Practical advantages of a Bayesian approach

Using Bayes theorem has a number of practical advantages:

1) It's conceptually simple. Every problem amounts to:
   A. list all of the possible hypotheses
   B. assign a prior to each hypothesis based upon what you already know
   C. calculate the likelihood of observing the data for each hypothesis, and then use Bayes' theorem
2) It gives an actual probability estimate for each hypothesis
3) It makes it easy to combine different measurements and to include background information
4) It's guaranteed to be self-consistent and in accord with "common sense"
5) It makes handling systematic errors very easy

But the whole thing fails if you don't know how to do A or B. In that case, you probably fall back on frequentist alternatives. These use only C, but at a cost: they cannot directly tell you the relative probabilities of different hypotheses.

Nuisance parameters

A “nuisance parameter” is a parameter model that affects the probability distributions but which we don't care about for its own sake. An example would be a calibration constant of an apparatus---not the sort of thing you report in the abstract, but important nonetheless.

Bayesian analysis gives a simple procedure for handling these: assign priors to all parameters, calculate the joint posterior PDF for all parameters, then marginalize over the unwanted parameters.

If θ is an interesting parameter, while α is a calibration constant, we write:

$$P(\theta \mid D,I) = \int d\alpha\, P(\theta,\alpha \mid D,I) = \int d\alpha \left[\frac{P(\alpha \mid I)\, P(\theta \mid I)\, P(D \mid \theta,\alpha,I)}{P(D \mid I)} \right]$$

(I've assumed independent priors on α and θ, but this is not necessary.)

Systematic uncertainties

$$P(\theta \mid D,I) = \int d\alpha\, P(\theta,\alpha \mid D,I) = \int d\alpha \left[\frac{P(\alpha \mid I)\, P(\theta \mid I)\, P(D \mid \theta,\alpha,I)}{P(D \mid I)} \right]$$

Nuisance parameters provide an obvious way to include systematic uncertainties. Introduce a parameter characterizing the systematic, specify a prior for the true values of that systematic, then integrate over the nuisance parameter to get the PDF for the quantity you do care about.

The frequentist version is much nastier---without the language of a “prior”, the marginalization procedure, and the philosophy of treating the data as generating a PDF for the parameters, it's much harder to handle systematics.

Justifying priors: the principle of Ignorance

In the absence of any reason to distinguish one outcome from another, assign them equal probabilities.

Example: you roll a 6-sided die. You have no reason to believe that the die is loaded. It's intuitive that you should assume that all 6 outcomes are equally likely (p=1/6) until you discover a reason to think otherwise.

Example: a primordial black hole passing through our galaxy hits Earth. We have no reason to believe it's more likely to come from one direction than any other. So we assume that the impact point is uniformly distributed over the Earth's surface.

Parametrization note: this is not the same as assuming that all latitudes are equally likely!

Uniform Prior

Suppose an unknown parameter refers to the location of something (e.g. a peak in a histogram). All positions seem equally likely.

Imagine shifting everything by x'=x+c. We demand that $p(X|I)\,dX = P(X'|I)\,dX' = P(X'|I)\,dX$. This is only true for all c if P(X) is a constant.

Really obvious, perhaps ... if you are completely ignorant about the location of something, use a uniform prior for your initial guess of that location.

Note: although a properly normalized uniform prior has a finite range, you can often get away with using a uniform prior from $-\infty$ to $+\infty$ as long as the product of the prior and the likelihood is finite.

Uniform Prior on Log

Suppose an unknown parameter measures the size of something, and that we have no good idea how big the thing will be (1mm? 1m? 1km?). We are ignorant about the *scale*. Put another way, our prior should have the same form no matter what units we use to measure the parameter with. If T'=βT, then

$$P(T\,|\,I)\,dT = P(T'\,|\,I)\,dT' = p(T'\,|\,I)\,\beta dT$$

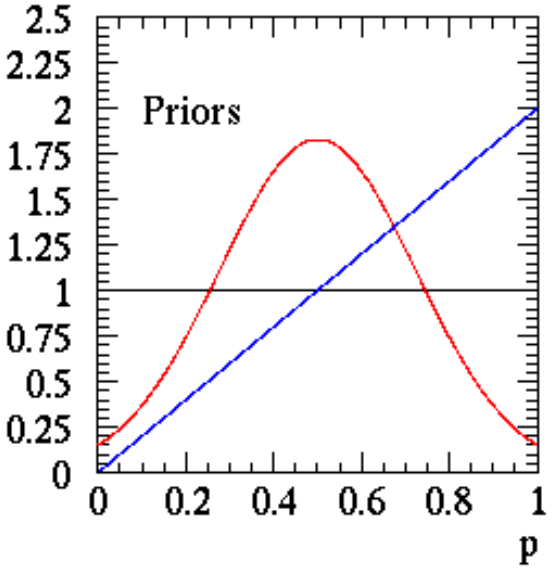
$P(T\,|\,I) = \beta P(\beta T\,|\,I)$, which is only true for all β if

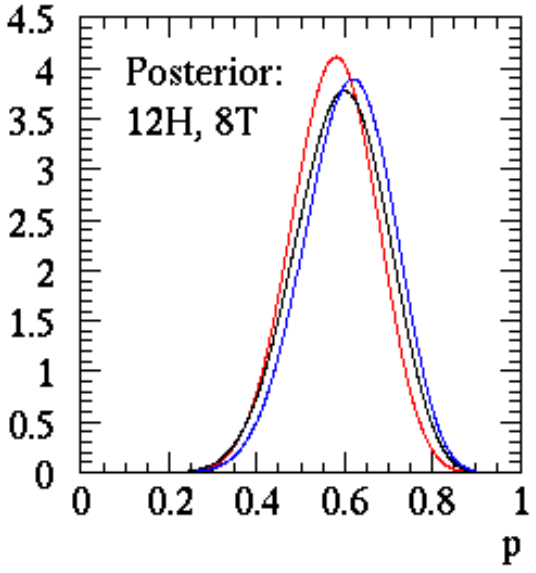
$$P(T\,|\,I) = \frac{\text{constant}}{T}$$

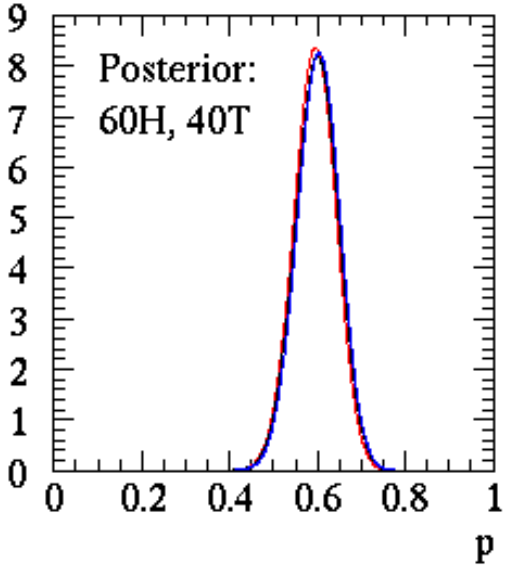
Properly normalized from T_{min} to T_{max} this is:

$$P(T\,|\,I) = \frac{1}{T\,\ln\left(T_{max}/T_{min}\right)}$$

Given enough data, priors don't matter







The more constraining your data becomes, the less the prior matters.

When posterior distribution is your much narrower than prior, the prior won't vary much over the region of interest. Most priors approximate to flat in this case.

Consider the case of estimating *p* for a binomial distribution after observing 20 or 100 coin flips.

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 (functions of the data) that are as close as possible to the true value of the parameter.”

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)

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

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

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

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

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

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

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

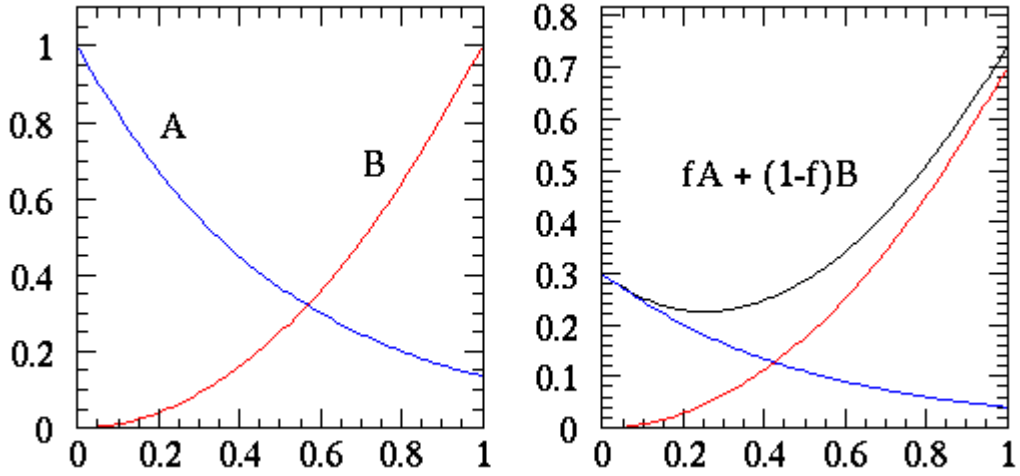
It's more common (and easier) to calculate -ln *L* instead:

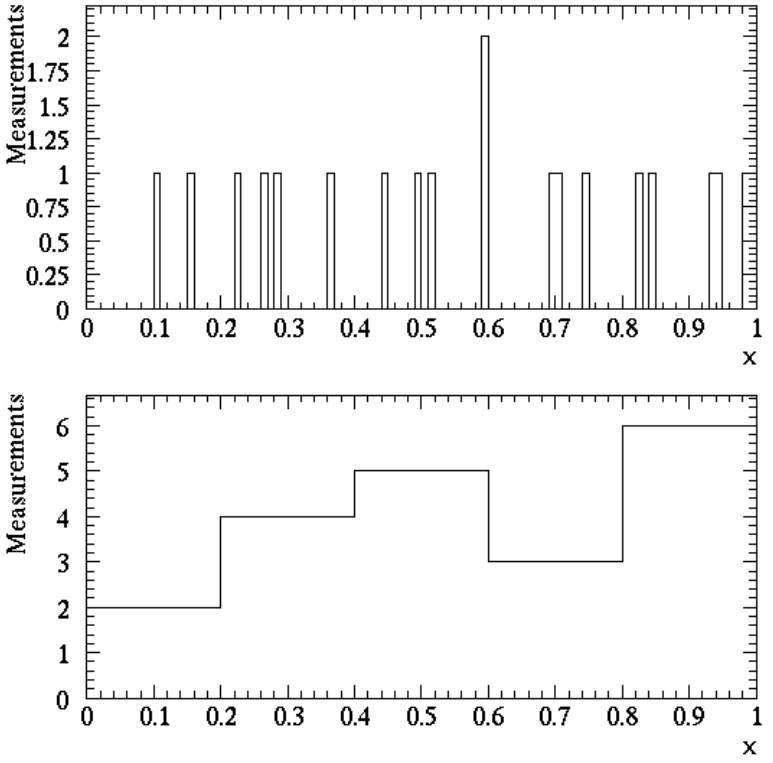
$$-\ln L(x_1 \ldots x_N|\vec{\alpha}) = -\sum_{i=1}^{N} \ln P(x_i|\vec{\alpha})$$

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.

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

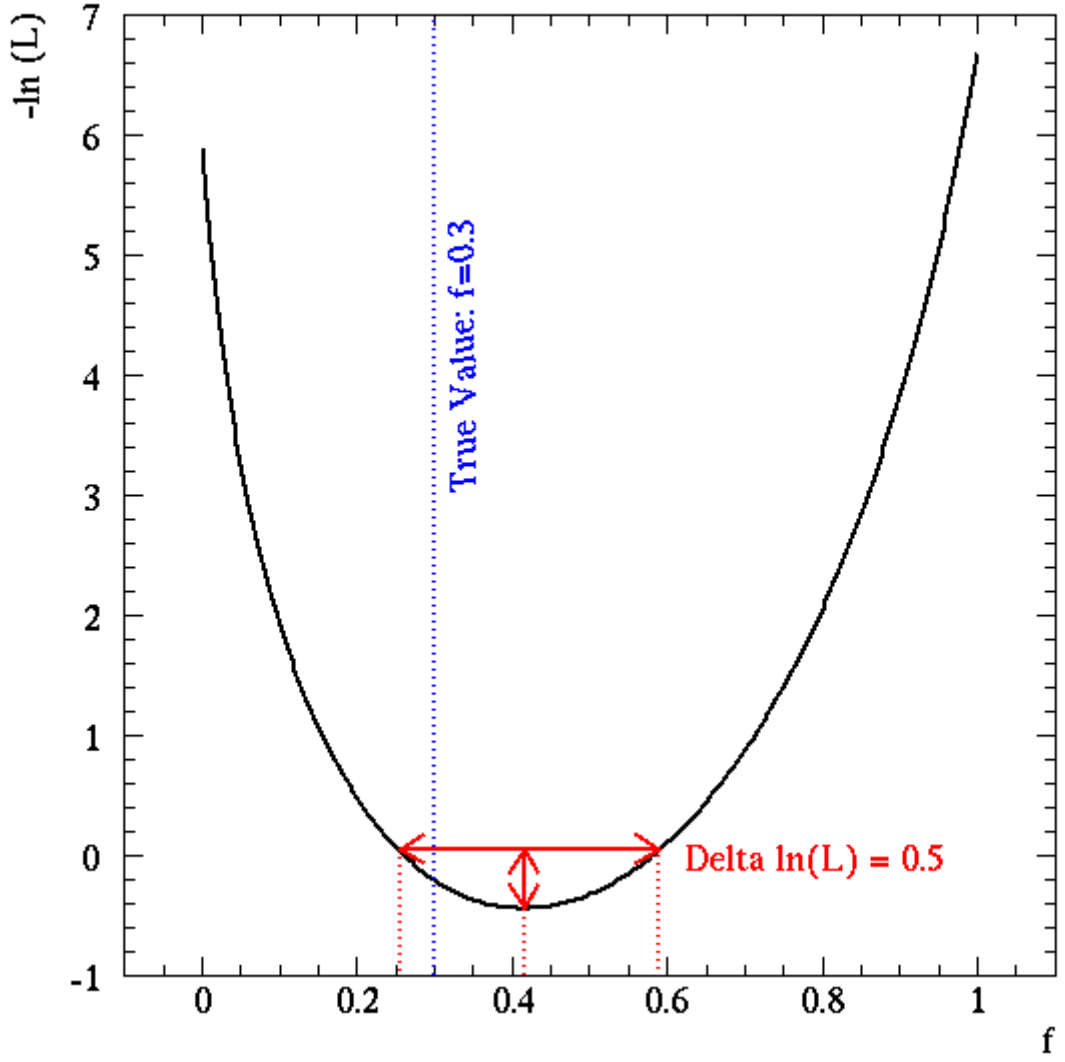
Form the negative log likelihood:

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

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.

Notice: binning is not necessary!

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.

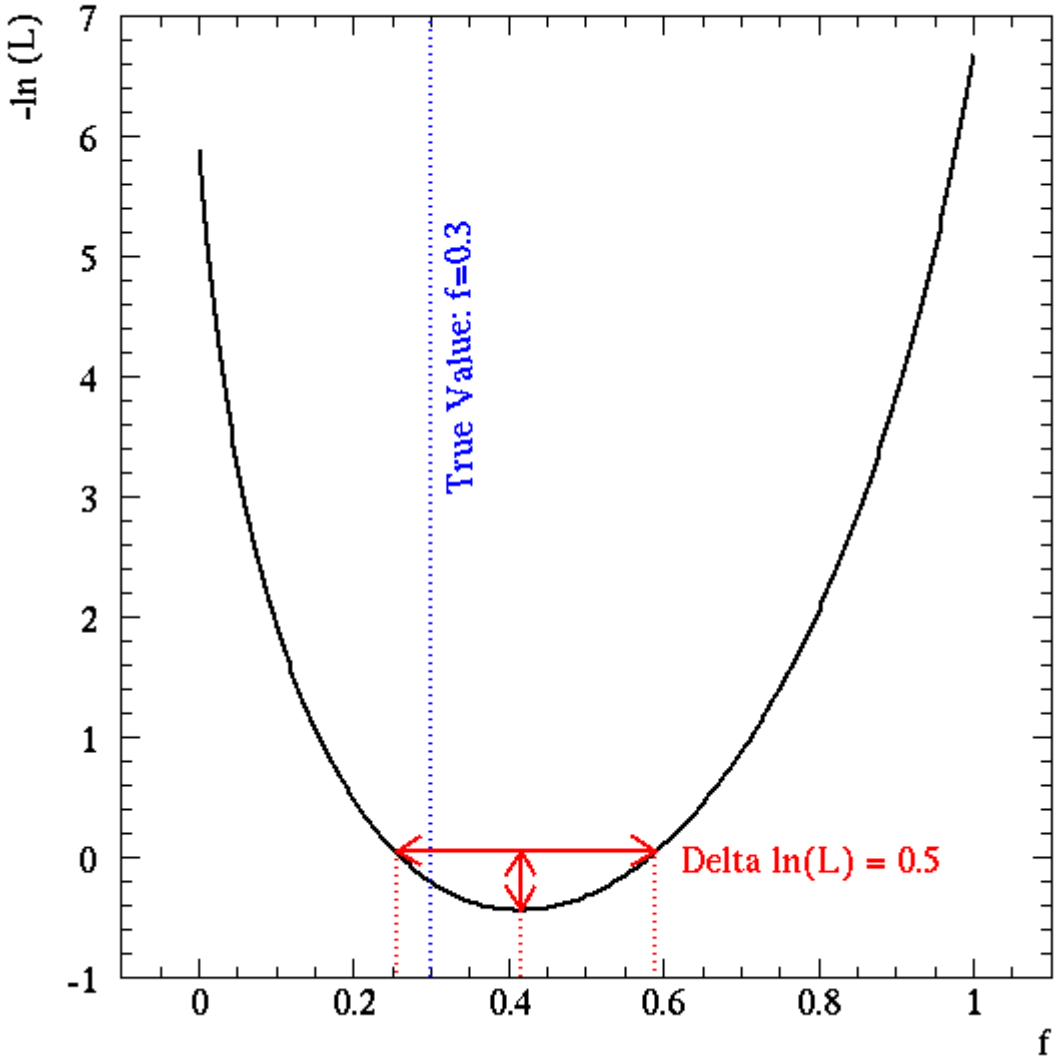
3) Estimators are invariant under parameter transformations:

$$f\left(a\right)=f\left(\hat{a}\right)$$

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:

$$\sigma_{\hat{a}}^{2}=\frac{-1}{\left\langle \left.\frac{d^{2}\ln L}{da^{2}}\right|_{a_0}\right\rangle}$$

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:

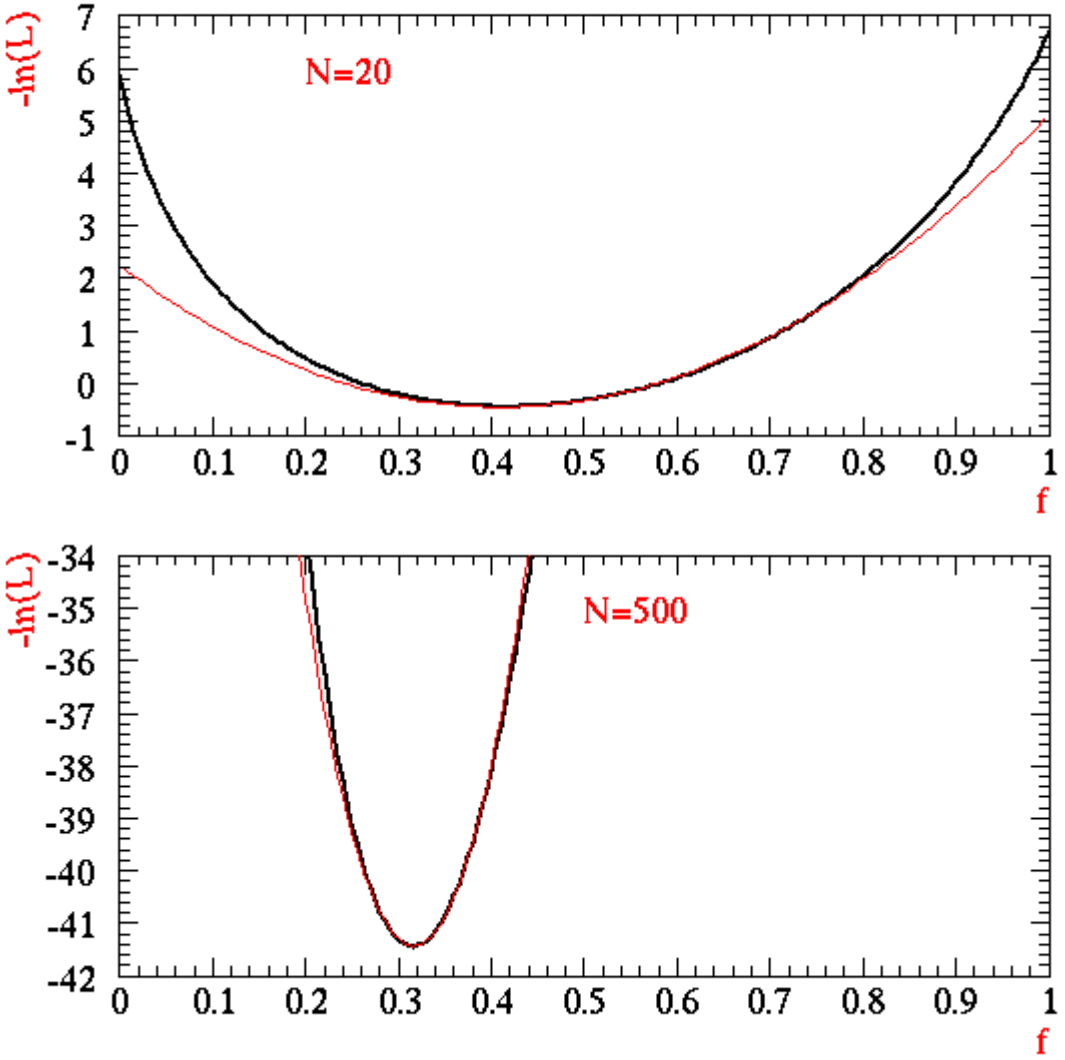
$$-\ln L = L_0 + \frac{1}{2}\left(\frac{f - \langle f \rangle}{\sigma_f}\right)^2$$

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).

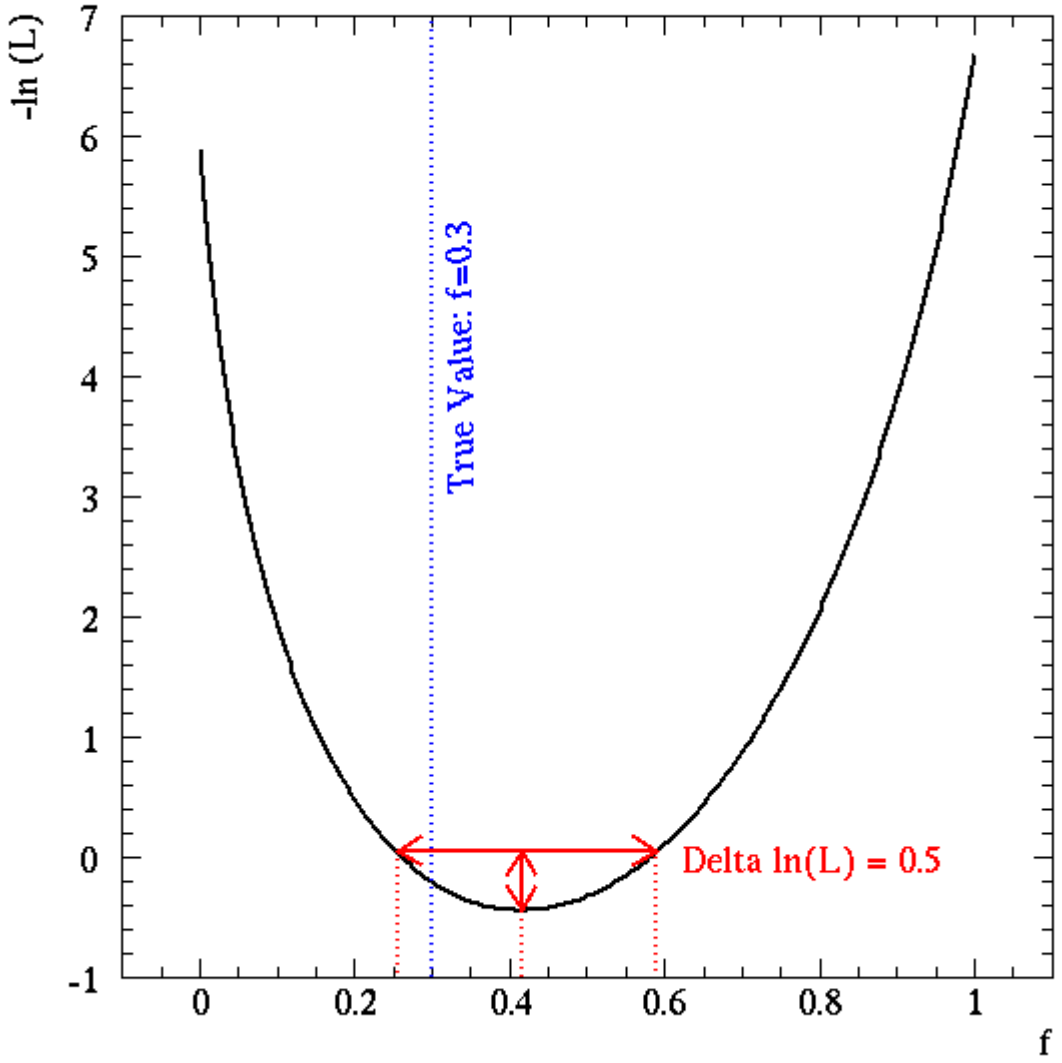
Parabolicity of the log likelihood



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?

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).

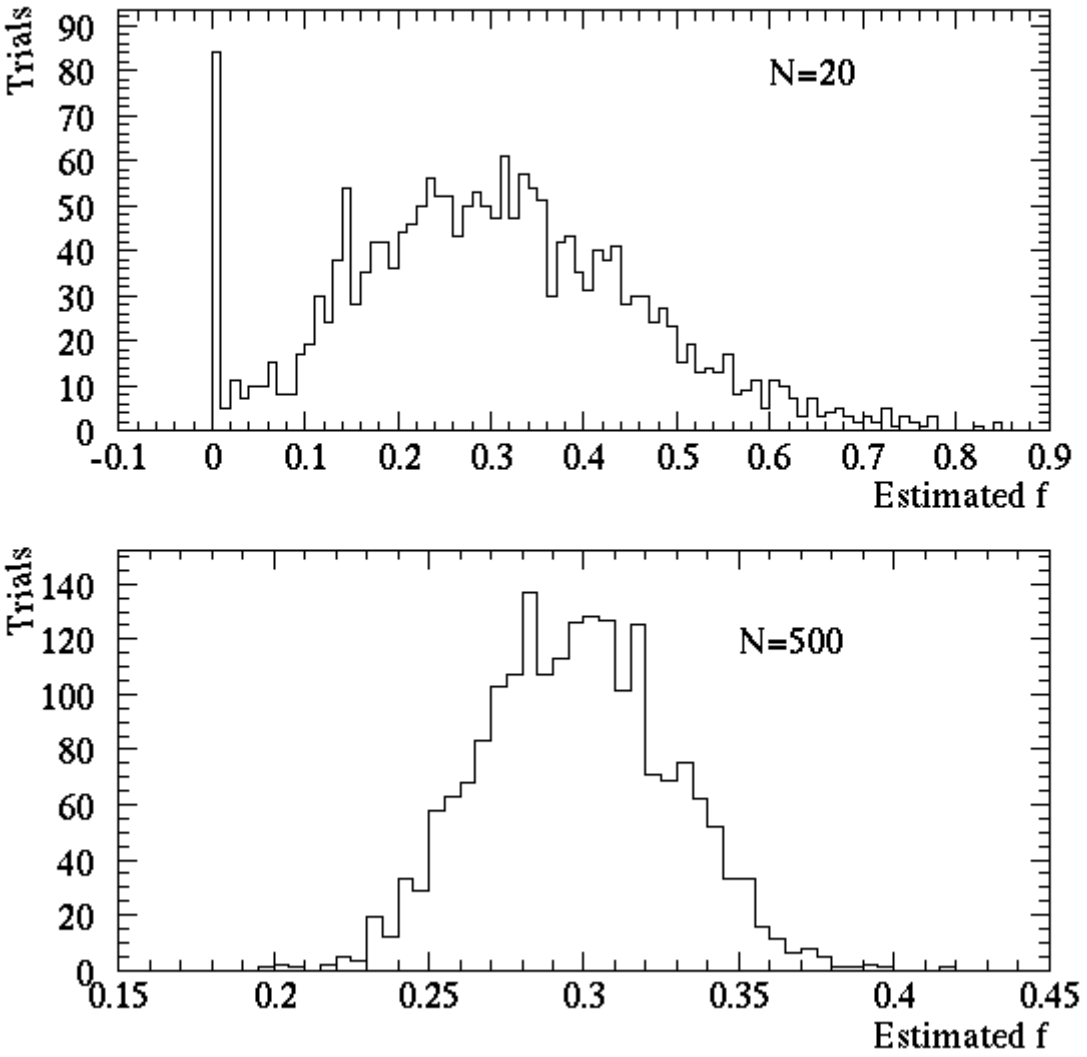
So the ML estimator is actually the peak location for the Bayesian posterior PDF assuming a flat prior P(a|I)=1.

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.

Coverage of ML estimator errors

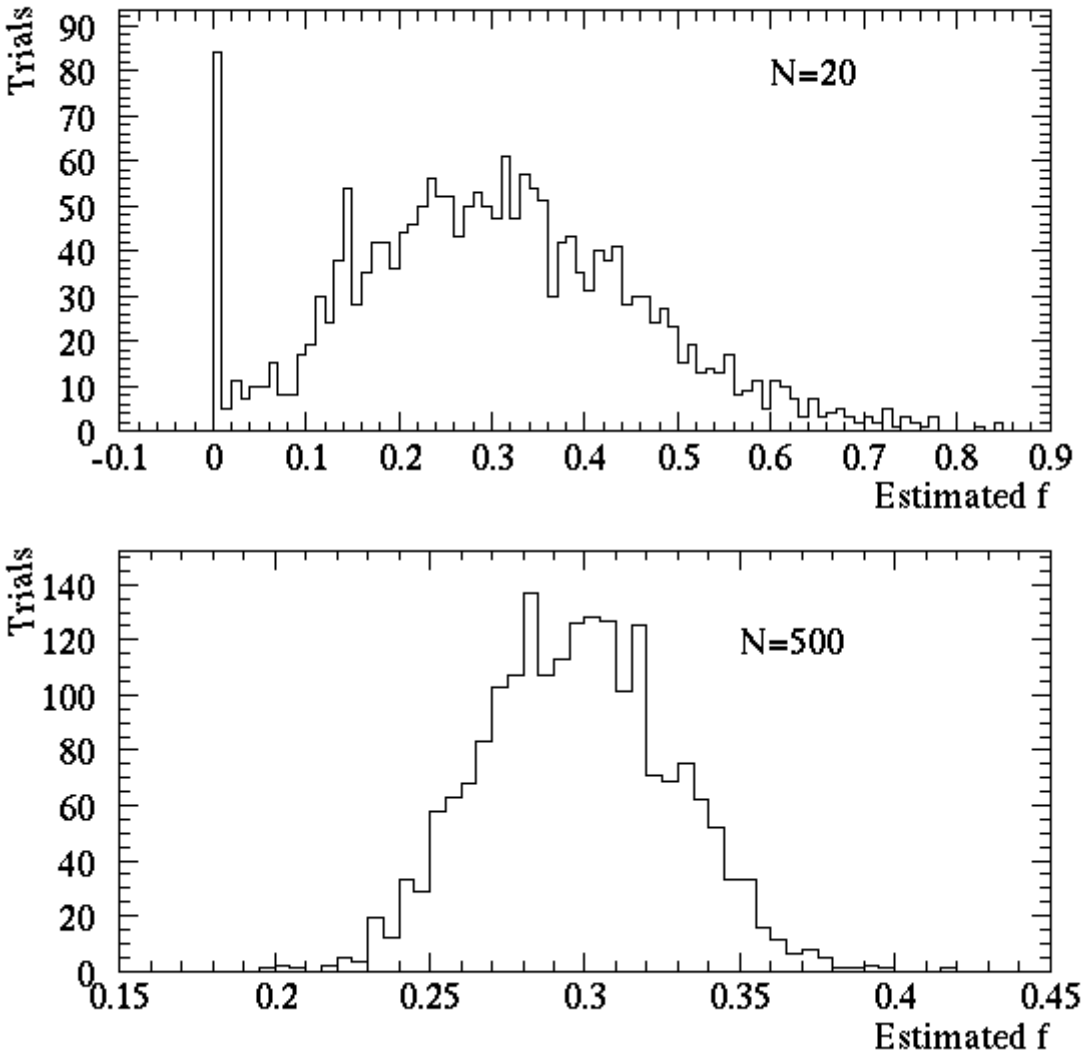


Distribution of ML estimators for two N values

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	How often 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.

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

Errors determined by (in the Gaussian approximation):

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

Maximum Likelihood with Gaussian Errors

Suppose we want to fit a set of points (x_i,y_i) to some model $y=f(x|\alpha)$, in order to determine the parameter(s) α. Often the measurements will be scattered around the model with some Gaussian error. Let's derive the ML estimator for α.

$$L=\prod_{i=1}^{N}\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y_i-f(x_i|\alpha)}{\sigma_i}\right)^2\right]$$

The log likelihood is then

$$\ln L=-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i-f(x_i|\alpha)}{\sigma_i}\right)^2-\sum_{i=1}^{N}\ln\left(\sigma_i\sqrt{2\pi}\right)$$

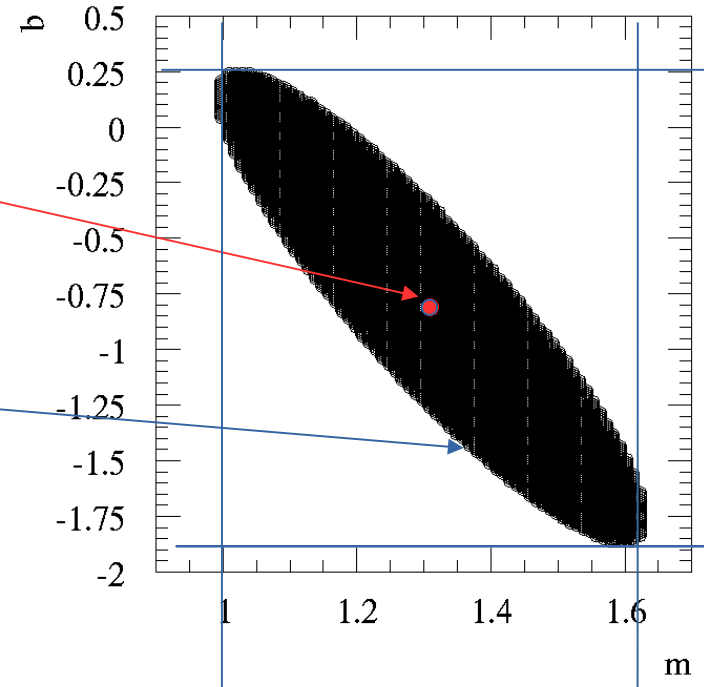
Maximizing this is equivalent to minimizing

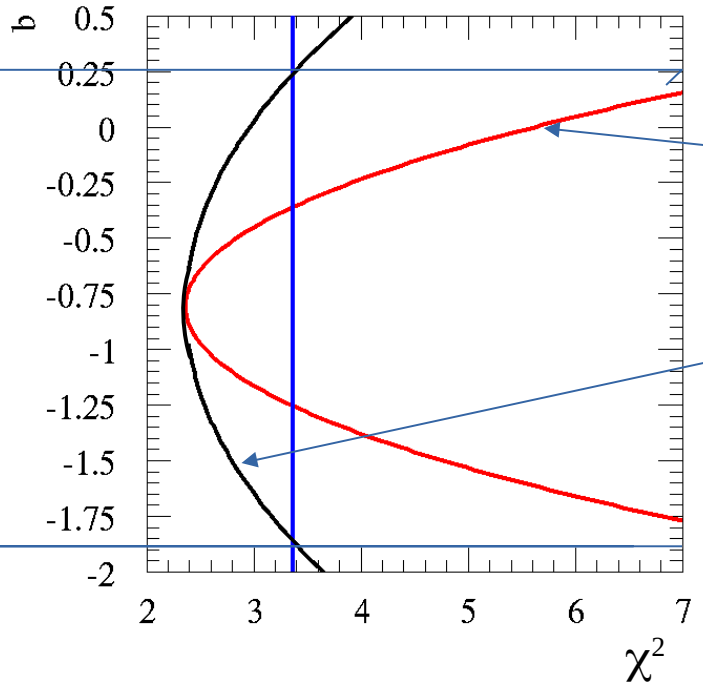
$$\chi^2=\sum_{i=1}^{N}\left(\frac{y_i-f(x_i|\alpha)}{\sigma_i}\right)^2$$

Contours and marginalization

Best fit point

Black ellipse: $\Delta\chi^2<+1$ from best fit



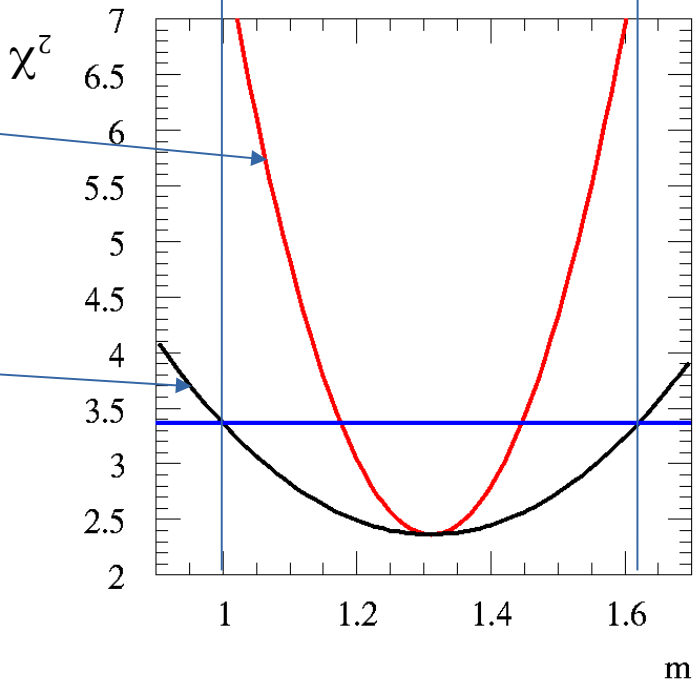


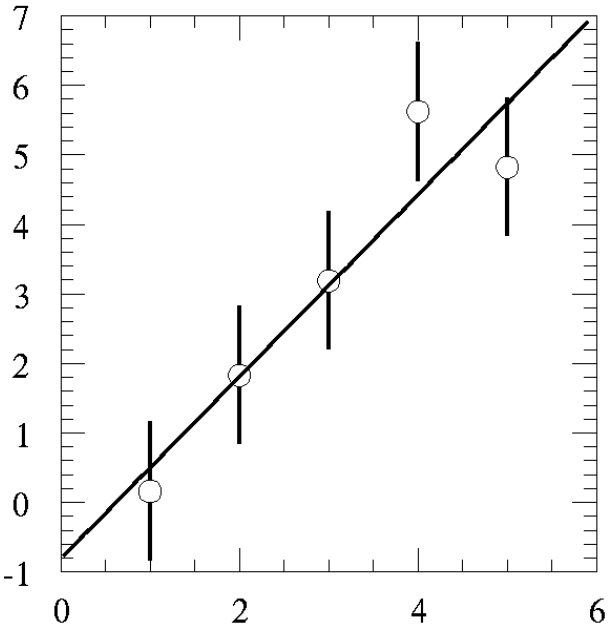
χ^2 vs b for fixed value of m

χ^2 vs b, minimizing χ^2 w.r.t. m at each value of b

χ^2 vs m for fixed value of b

χ^2 vs m, minimizing χ^2 w.r.t. b at each value of m





Best linear fit to the data

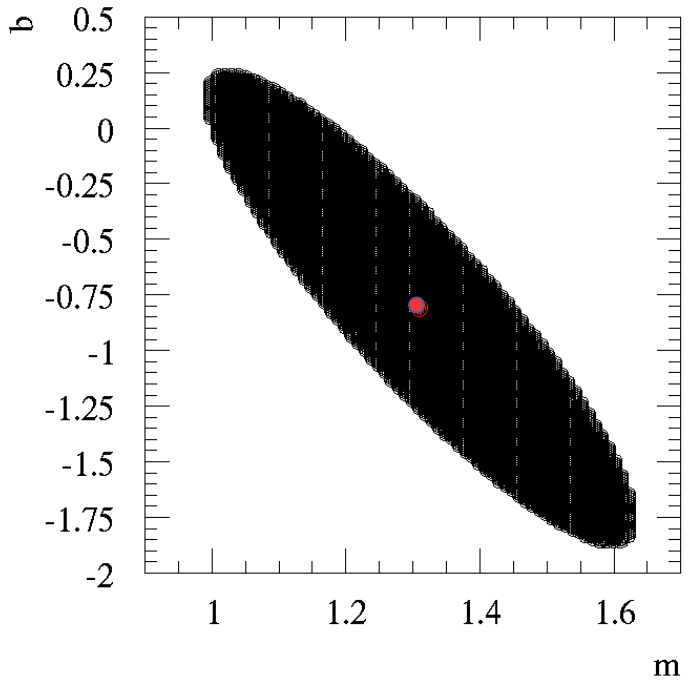
Errors on each individual parameter

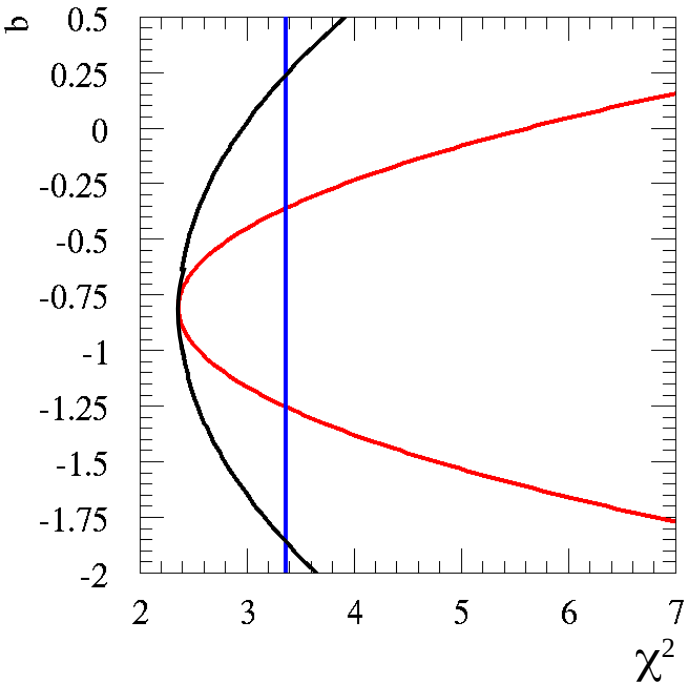
To find 1σ error on any parameter, scan over that parameter while minimizing the χ^2 as a function of all other free parameters.

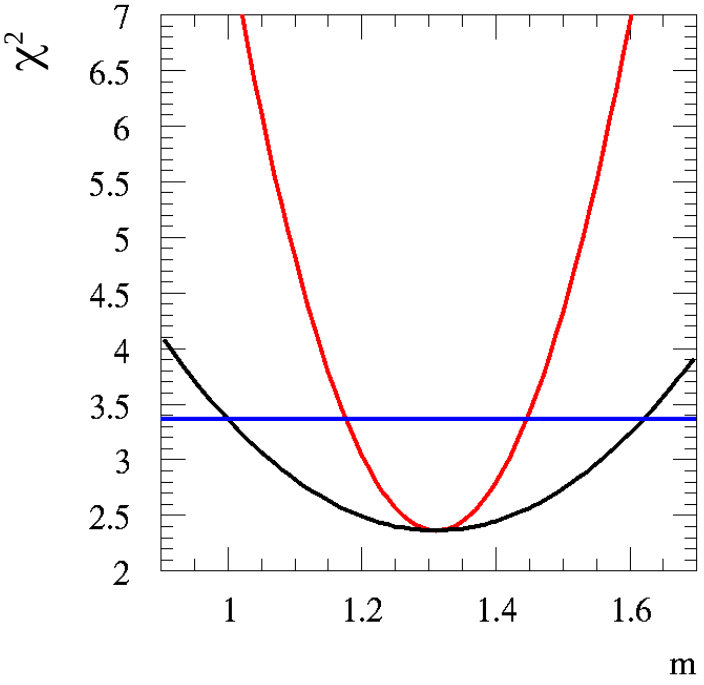
The points at which the χ^2 (minimized with respect to all other free parameters) has increased by +1 from its global minimum give the 1σ errors on the parameter.

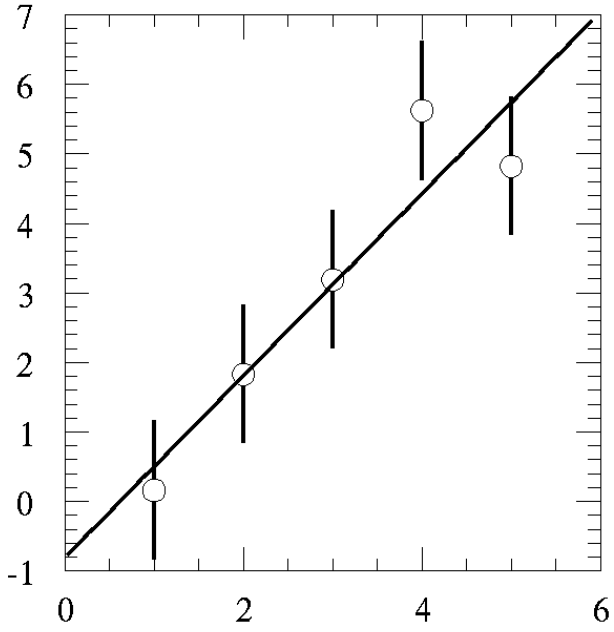
Do NOT leave the other parameters fixed at their best-fit values while scanning!

If minimizing -ln L instead of χ^2, increase by +1/2 instead of +1.

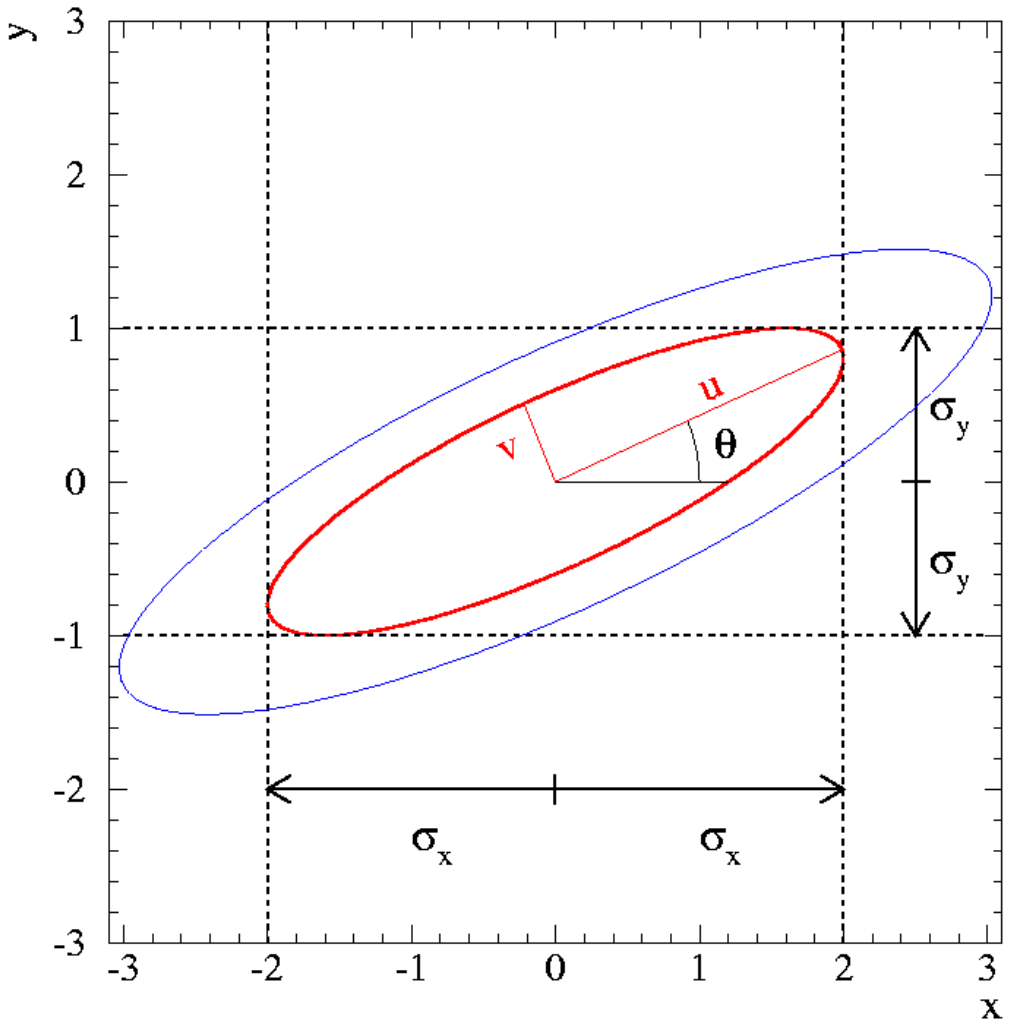








Error contours for multiple parameters



We can also find the errors on parameters by drawing contours on $\Delta \ln L$ or χ^2.

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). Multiply by 2 if using χ^2.

	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%	3.3236	4.61	5.65

Two marginalization procedures

Normal marginalization procedure: integrate over nuisance variables:

$$P(x) = \int dy\, P(x,y)$$

Alternate marginalization procedure: maximize the likelihood as a function of the nuisance variables, and return the result:

$$P(x) = \max_{y}\ P(x,y)$$

(It is not necessarily the case that the resulting PDF is normalized.)

I can prove for Gaussian distributions that these two marginalization procedures are equivalent, but cannot prove it for the general case (In fact they give different results).

Bayesians always follow the first prescription. Frequentists most often use the second.

Sometimes it will be computationally easier to apply one, sometimes the other, even for PDFs that are approximately Gaussian.

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:

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

Extended ML method:

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

Extended maximum likelihood estimators

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

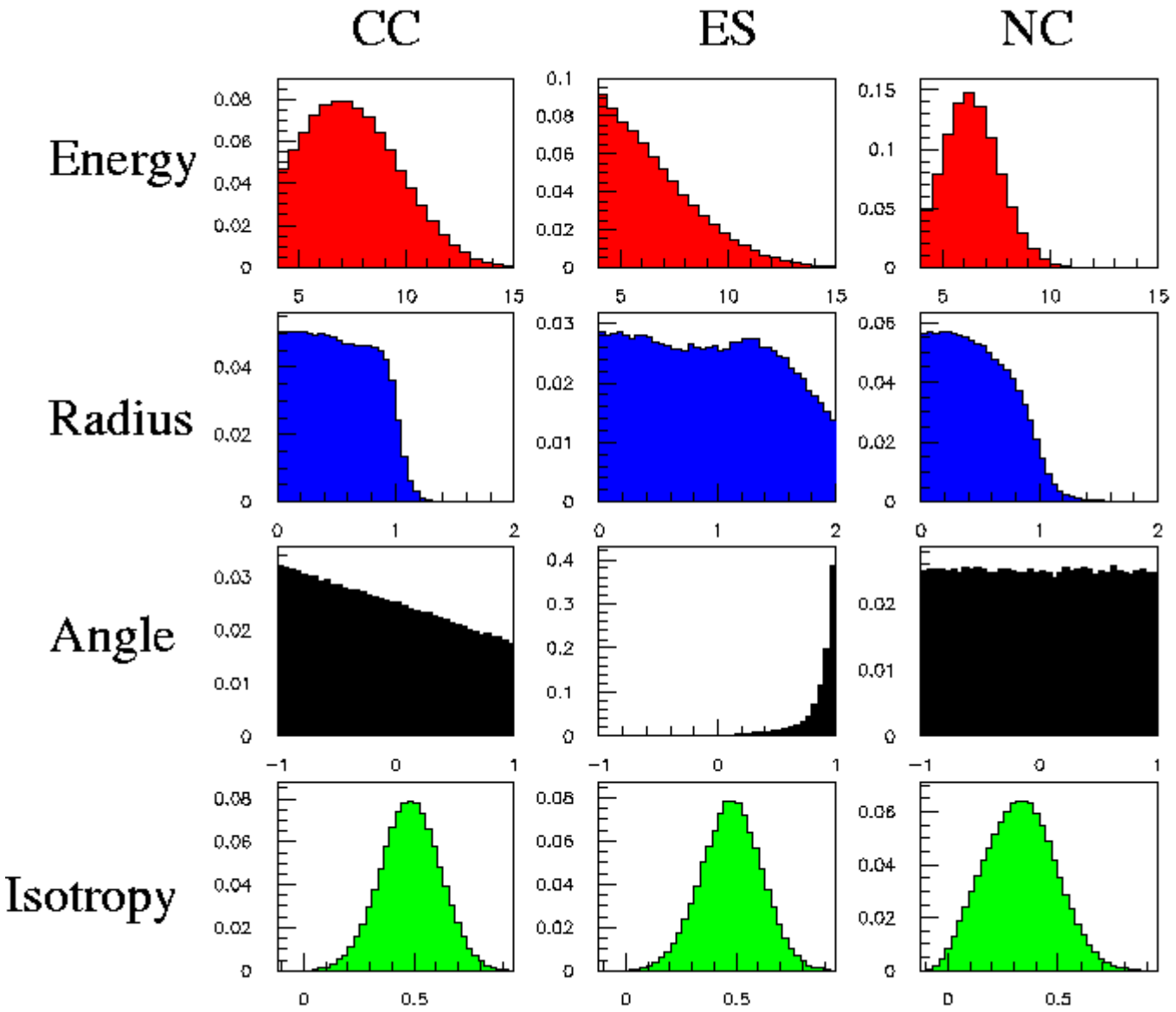
$$\text{Likelihood} = \frac{e^{-\nu}\nu^{N}}{N!}\cdot\prod_{i=1}^{N} P(x_i|\vec{\alpha}) \quad [\text{note that } \nu = \nu(\vec{\alpha})]$$

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

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

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



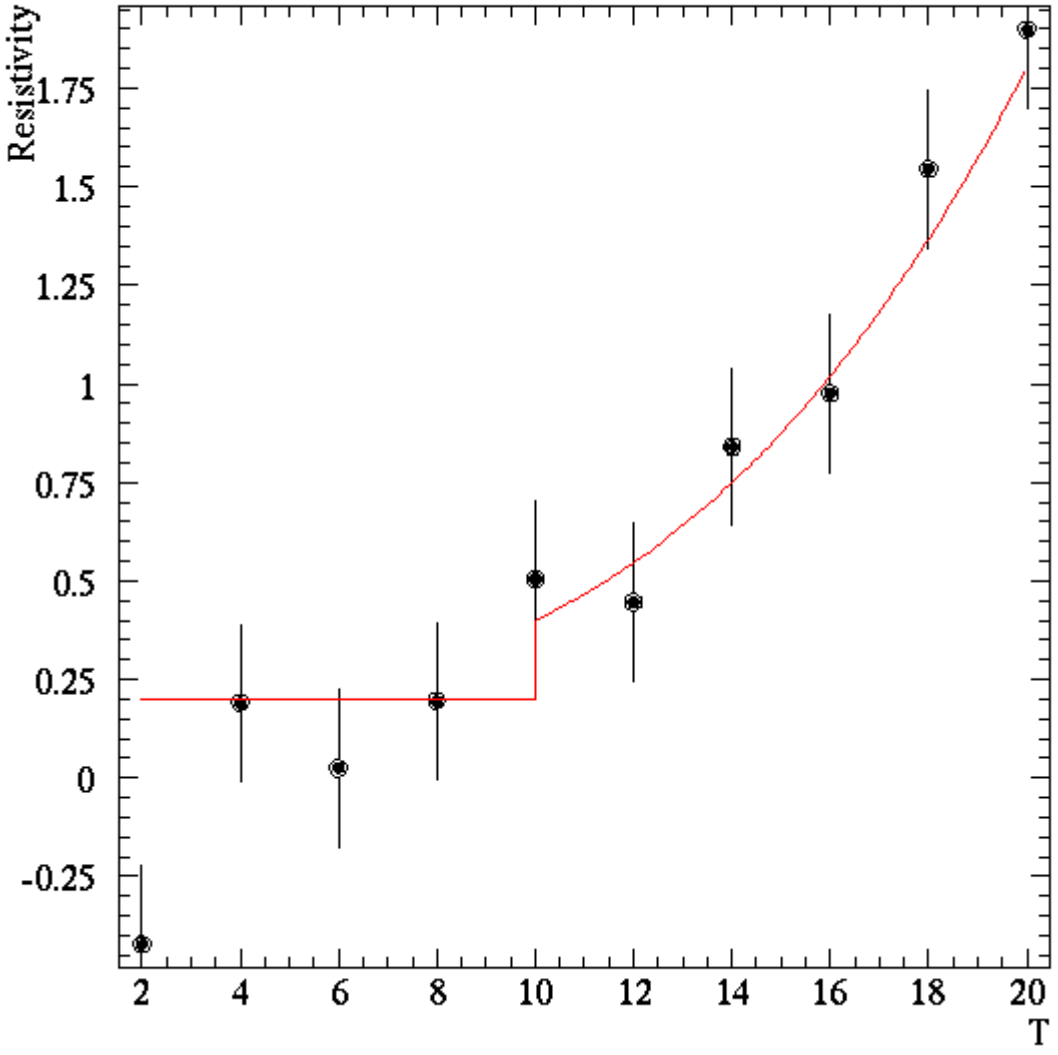
$P(E,R,\Theta,\beta) =$
$\quad CC\ P_{CC}(E,R,\Theta,\beta)$
$+ ES\ P_{ES}(E,R,\Theta,\beta)$
$+ NC\ P_{NC}(E,R,\Theta,\beta)$

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.

Extra material

An involved example: estimating a superconductor's critical temperature



Superconductor has sudden drop in resistivity below its critical temperature. Model it as:

$R = B$ (if $T<T_c$)

$R = B+A(T/T_c)^3$ (if $T>T_c$)

Here B is a calibration offset, T_c is the critical temperature, and A is an uninteresting material parameter.

Data at right drawn from true distribution shown in red.

Superconductor: define the model

There are three parameters, only one of which we really care about. Let's assume uniform priors for each:

$P(B) = 1 \quad (0<B<1)$
$P(A) = 1 \quad (0<A<1)$
$P(T_c)=1/20 \quad (0<T_c<20)$

And now we define the model. The model will be that the data are scattered around the theoretical curve

$R = B \quad (\text{if } T<T_c)$
$R = B+A(T/T_c)^3 \quad (\text{if } T>T_c)$

with Gaussian errors having σ=0.2 (we assume this is known from characterization of the apparatus).

Superconductor: the form of the likelihood

Need to write down a form for P(D|A,B,T_c,I)

$$P\left(D \mid A,B,T_c,I\right)=\prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}\left(D_i-R\left(T_i\right)\right)^2\right]$$

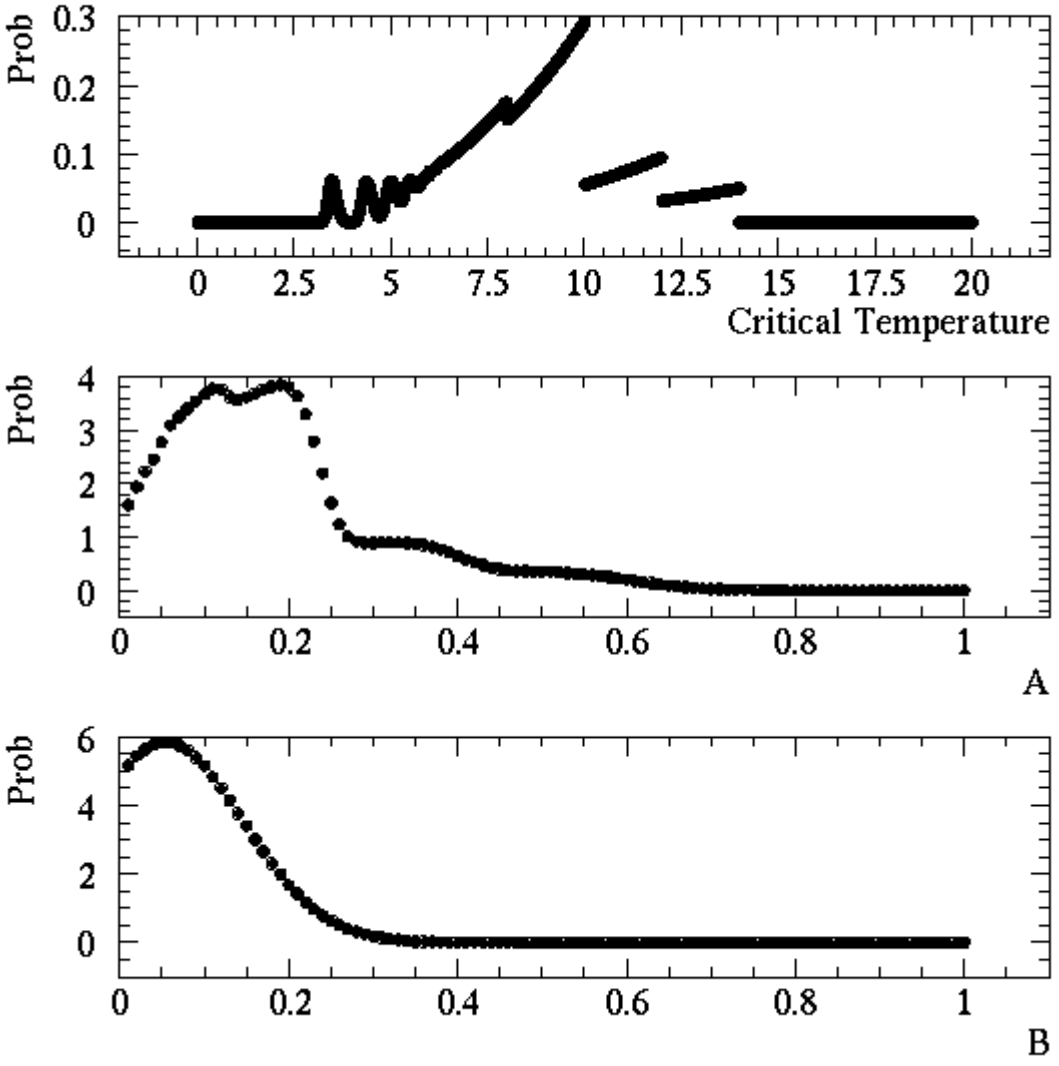
where R(T_i) is the piecewise-defined function given previously. All the dependence on model parameters is contained in R(T).

Bayes theorem now immediately defines a joint PDF for the parameters by

$$P\left(A,B,T_c \mid D,I\right) \propto P\left(A,B,T_c \mid I\right) P\left(D \mid A,B,T_c,I\right)$$

All there is left to do is to normalize the PDF, and marginalize over the unwanted variables to get the PDFs on any parameter you care about.

Superconductor: marginalized PDFs



Here I show the marginalized PDFs for T_c, A, and B. A is perhaps like you would have expected. B is OK---low, but data was quite a bit low as well.

(True values: A=0.2, B=0.2, T_c=10)

PDF for T_c puzzled me at first. It spikes near true value, but is not very smooth. The reason is that the model being fitted is discontinuous, so you get discontinuities at the data points.

Advantages of a Bayesian approach

If you start with some probability distribution for the value of a parameter, or an estimate of the likelihood of a hypothesis, and then you learn some new piece of information (“the data”), Bayes' theorem immediately tells you how to update your distribution.

The strongest benefit of Bayesian statistics is that it directly answers the question you're really asking: how likely is your hypothesis? For example, you can calculate probabilities for things like: what is the probability that there's a new particle with a mass between 200-205 GeV?

You can ONLY directly calculate the odds of a hypothesis being true if you assume some prior, and if your interpretation of probability allows you to think of probability as a measure of credibility (rather than just frequency).

The Normalization Term (aka the denominator)

$$P(H|D,I) = \frac{P(H|I)\,P(D|H,I)}{P(D|I)}$$

The term P(D|I) is first of all a normalization term. It's the probability of the data summed over all considered hypotheses. (Really it's the integral of the numerator over all values of H).

It's also a check on the validity of your assumptions. If P(D|I) is very, very small, then either you got unlucky, or your prior was far off, or your hypothesis set (denoted by I) doesn't include the true hypothesis.

Sherlock Holmes on hypotheses

"How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?"

Bayesian analysis enforces this, since the renormalization procedure demands that one of the hypothesis explicitly under consideration must be the correct one.

If your set of hypotheses is incorrect, your analysis is too.

Writing P(H|D,I) instead of just P(H|D) is one reminder that background assumptions are *always* being made.

Odds Ratio

$$O_{12} = \frac{P(M_1|D,I)}{P(M_2|D,I)} = \frac{P(M_1|I)P(D|M_1,I)}{P(M_2|I)P(D|M_2,I)}$$

$$O_{12} = \frac{P(M_1|I)}{P(M_2|I)} \frac{P(D|M_1,I)}{P(D|M_2,I)} \equiv \frac{P(M_1|I)}{P(M_2|I)} B_{12}$$

The odds ratio is useful because the normalization factors cancel. It's the ratio of the prior probability estimates times the Bayes factor (ratio of the global likelihoods given the data D).

Odds ratios can be easily converted back into probabilities by restoring the normalization factors:

$$P(M_i|D,I) = \frac{O_{i1}}{\sum_{i=1}^{N} O_{i1}}$$

Bayesian justification of Occam's Razor

“Plurality ought never be imposed without necessity.”---William of Ockham

“Of two equivalent theories or explanations, all other things being equal, the simpler one is to be preferred.”

“We are to admit no more causes of natural things than such are both true and sufficient to explain their appearances.”---Isaac Newton

Consider these two hypotheses:

1) Model M_0 is true. It has no free parameters, but there is one parameter θ whose value is fixed by theory to θ_0.

2) Model M_1 is true. It has a single free parameter θ.

Which of these models should we favour given the data?

At first glance, M_1 is more powerful in a sense. After all, it includes $\theta=\theta_0$ as one special case, so shouldn't it always be more likely than the more restricted hypothesis?

Intuitively this can't be right, because this would say we should always favour the more complicated hypothesis, even when the data are perfectly consistent with both.

Odds ratio and the Occam factor

Let's calculate the Bayes factor for the two hypotheses.

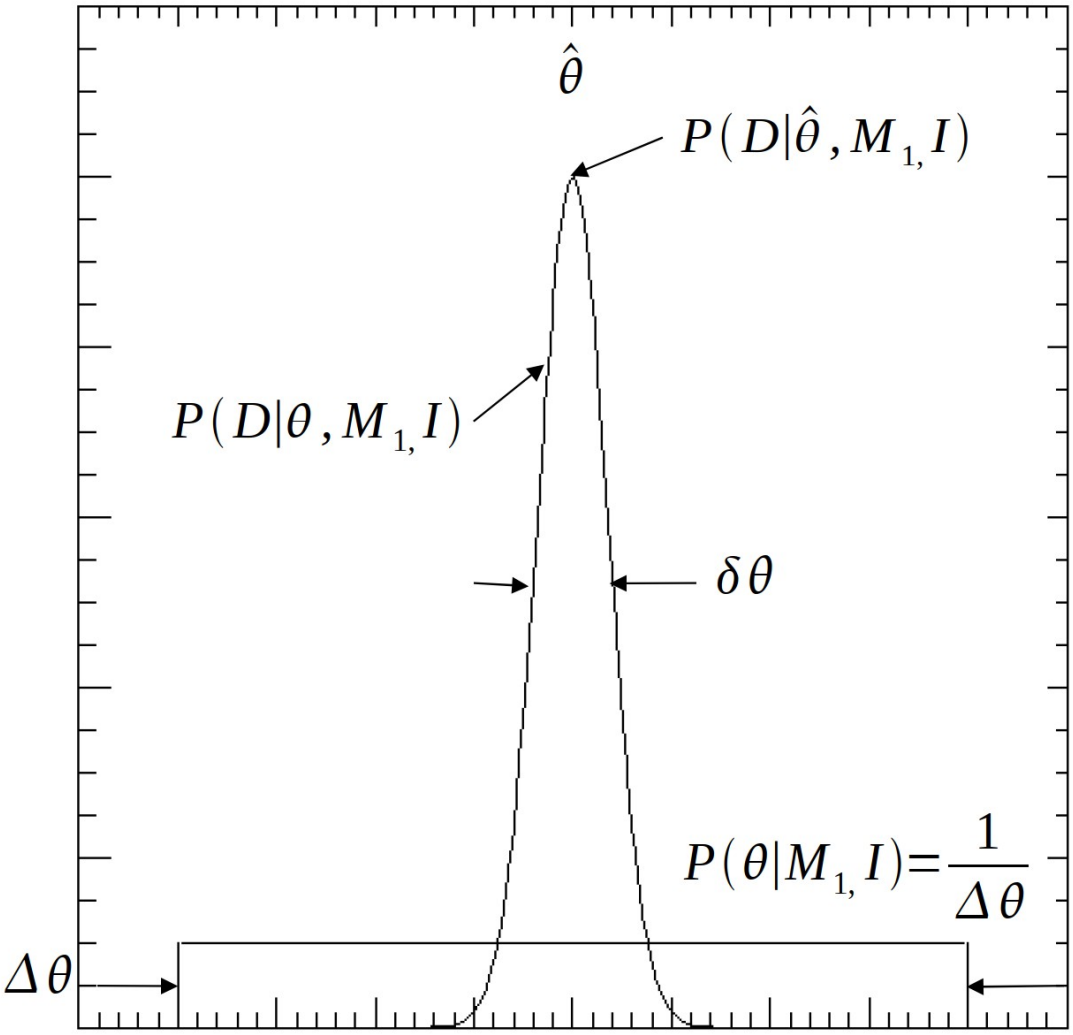
1) For M_0, $P(D|M_0,I)=P(D|\theta_0,M_1,I)=\mathscr{L}(\theta_0)$. Simple to evaluate.

2) For M_1 we need to marginalize over all possible values of θ, including the prior for θ.

$$P(D|M_1,I)=\int d\theta\, P(\theta|M_1,I)\, P(D|\theta,M_1,I)$$

We can approximate this integral. First, let's assume the data is actually pretty constraining compared to the prior, so that $P(\theta|M_1,I)$ is approximately flat over the range for which $P(D|\theta,M_1,I)$ is non-zero.

Odds ratio and the Occam factor



$$\int_{\Delta\theta} d\theta\, P(\theta|M_1,I) = P(\theta|M_1,I)\,\Delta\theta = 1$$

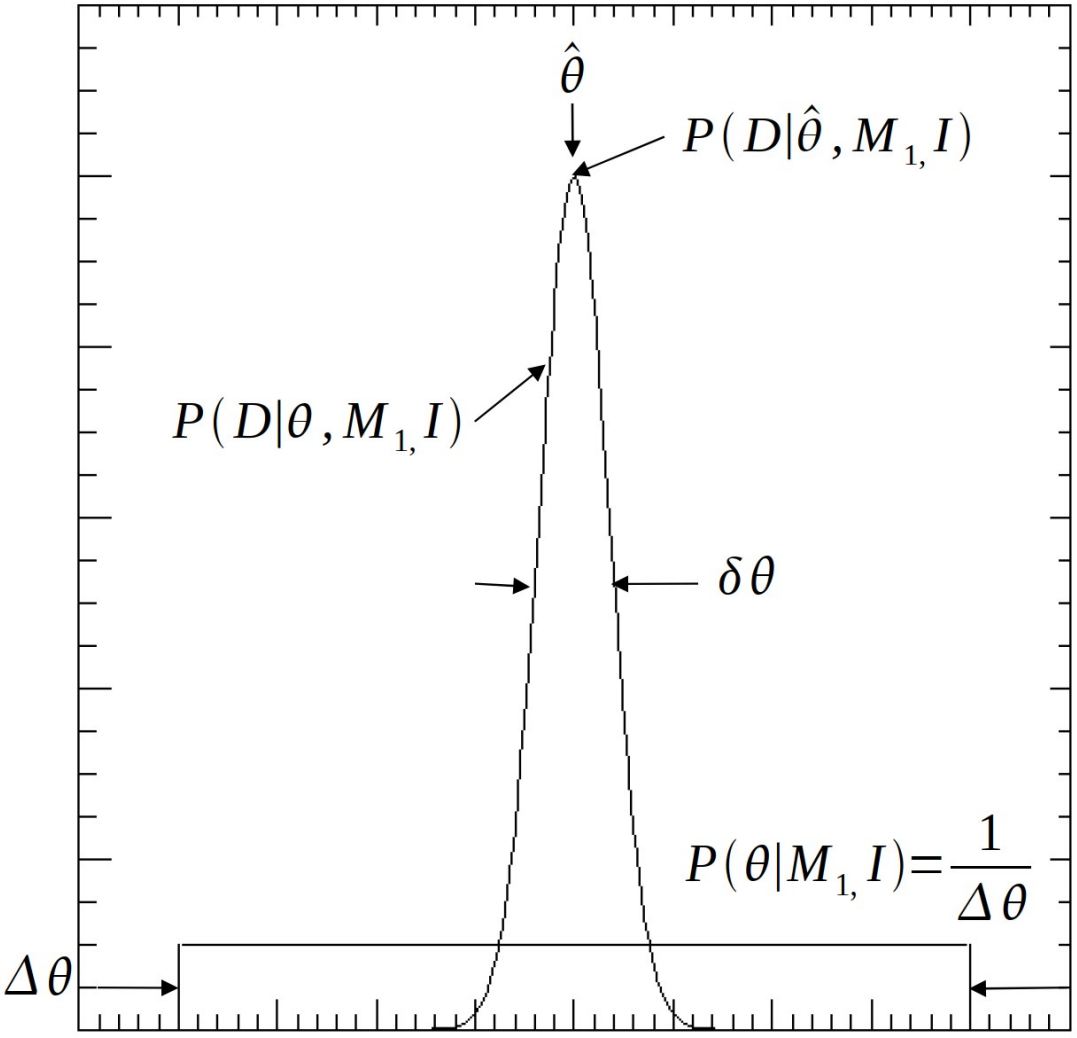
$$\int_{\Delta\theta} d\theta\, P(D|\theta,M_1,I) \equiv p(D|\hat{\theta},M_1,I)\,\delta\theta$$

$$P(D|M_1,I) = \int d\theta\, P(\theta|M_1,I)\,P(D|\theta,M_1,I)$$

$$= \frac{1}{\Delta\theta}\int d\theta\, P(D|\theta,M_1,I)$$

$$\approx \frac{\delta\theta}{\Delta\theta} P(D|\hat{\theta},M_1,I)$$

Odds ratio and the Occam factor



Bayes factor:

$$B \approx \frac{P(D|\hat{\theta}, M_1, I)}{P(D|\theta_0, M_1, I)} \frac{\delta\theta}{\Delta\theta}$$

The first part is always bigger than or equal to 1.

But the second factor is smaller than 1. Since the posterior width $\delta\theta$ is narrower than the prior width $\Delta\theta$, the second parameter is penalized for the "wasted" parameter space.

The Bayes factor will only favour the more complicated model if the likelihood factor (blue) is much better, overwhelming the Occam factor (red).

Summary of the Occam factor

Bayesian analysis automatically penalizes more complicated models in a quantitative way compared to simpler models.

This happens in the process of marginalizing over free parameters in the model. The more free parameters you have to marginalize over, the larger the penalty.

It is still of course possible that a more complicated model fits the data better. If the probability of the data under the simpler model is very small, but much larger under the more complicated model, then the complicated model will still be favoured in spite of penalty factor.

The penalty factor is perhaps intuitively obvious. The more free parameters you have, the more likely it is that your model matched the data just by blind luck. The model that makes more specific predictions (has *fewer* free parameters) will tend to be favoured, so long as it is consistent with data.

Prior from a prior analysis

The best solution to any problem is to let someone else solve it for you.

If there exist prior measurements of the quantities you need to estimate, why not use them as *your* prior? (Duh!)

Be careful, of course---if you have reason to believe that the previous measurement is actually a mistake (not just a statistical fluctuation) you wouldn't want to include it.

Even the most complicated statistical analysis does not eliminate the need to apply good scientific judgement and common sense.

Dependence on parametrization

Two theorists set out to predict the mass of a new particle

Carla (writes down theory):

"There should be a new particle whose mass is greater than 0 but less than 1, in appropriate units. I have absolutely no other knowledge about the mass, so I'll assume it has equal chances of having any value between zero and 1---i.e. P(m) = 1."

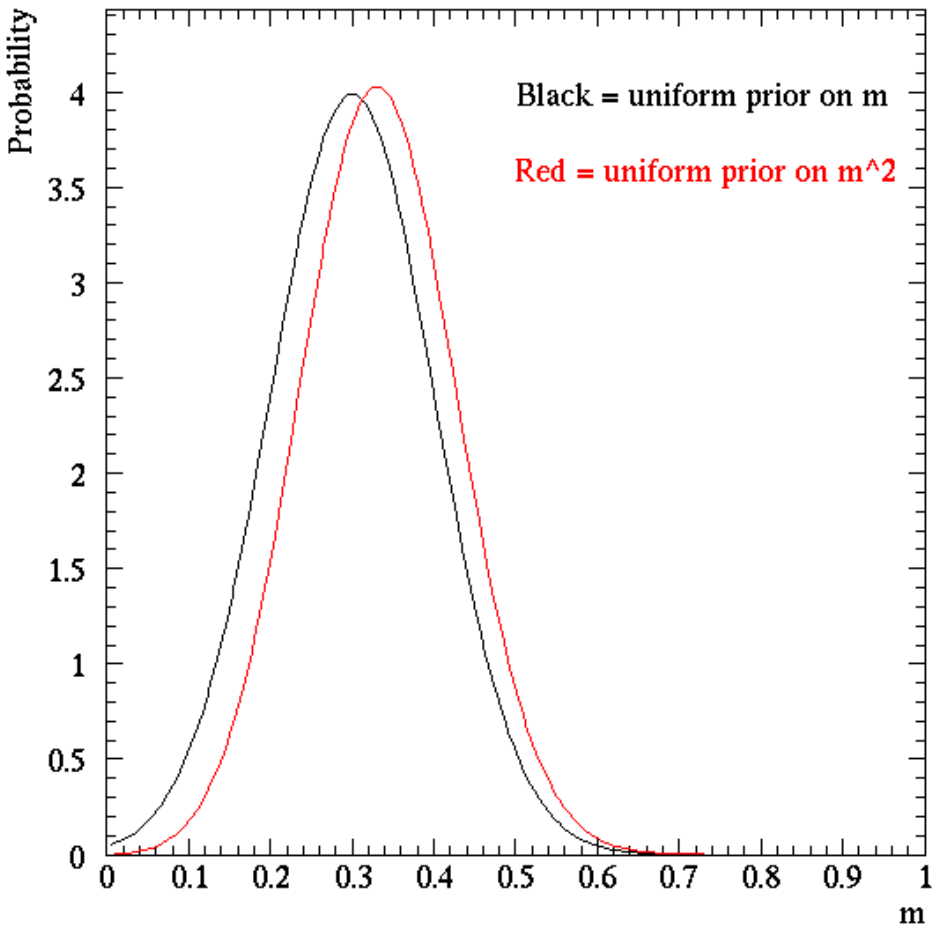
Heidi (writes down the exact same theory):

"There is a new particle described by a single free parameter $y=m^2$ in the Klein-Gordon equation. I'm sure that the true value of y must lie between 0 and 1. Since y is the quantity that appears in my theory, and I know nothing else about it, I'll assume a uniform prior on y---i.e. P(y) = 1."

These are two valid statements of ignorance about the same theory, but with different parametrizations.

An experiment reports: m=0.3±0.1

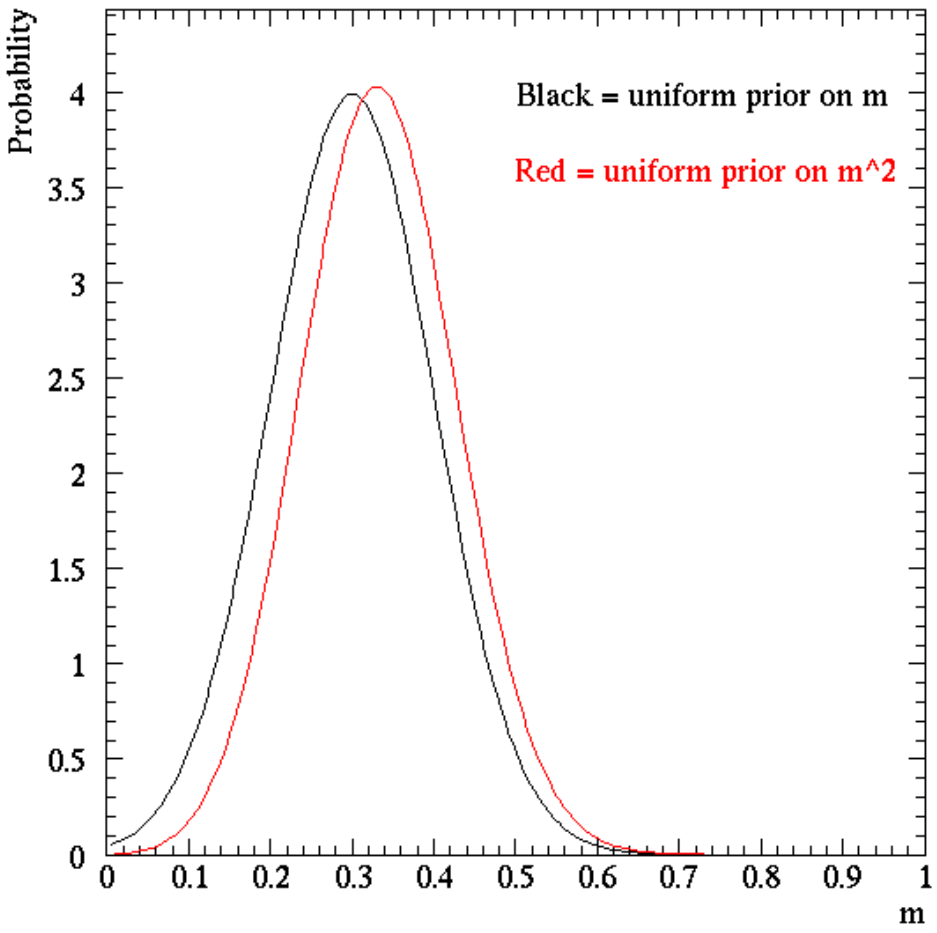
The experimental apparatus naturally measures m, so the experiment reports that (rather than y). Our two theorists incorporate this new knowledge into their theory. Carla calculates a new probability distribution P(m|D,I) for m. Heidi converts the measurement into a statement about the quantity $y=m^2$, and calculates P(y|D,I). They then get together to compare results. Heidi does a change of variables on her PDF so she can directly compare to Carla's result.



The sad truth: choice of parametrization matters

It's quantitatively different to say that all values of m are equally likely versus all values of m^2 are equally likely. The latter will favour larger values of m (if it's 50/50 that m^2 is larger than 0.5, then it's 50/50 than m is larger than 0.707).

Which is right? Statistics alone cannot decide. Only you can, based on physical insight, theoretical biases, etc.



If in doubt, try it both ways.

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:

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

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.

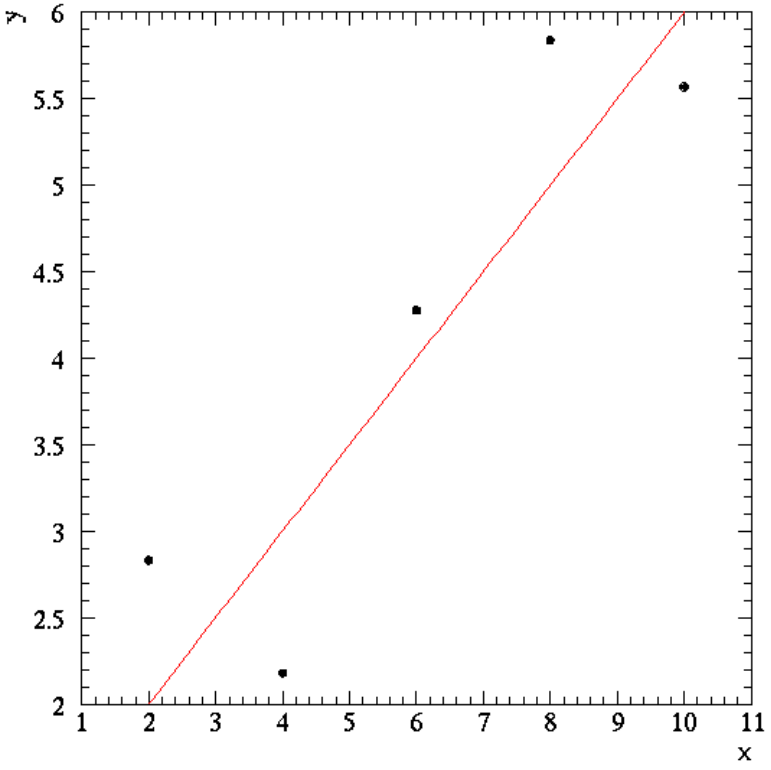
For samples from a Gaussian distribution, the RMS on our estimate for σ is given by

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

Think of this as the “error estimate on our error bar”.

A Gaussian is the least constraining assumption for the error distribution

A very useful and surprising result follows from this maximum entropy argument. Suppose your data is scattered around your model with an unknown error distribution:



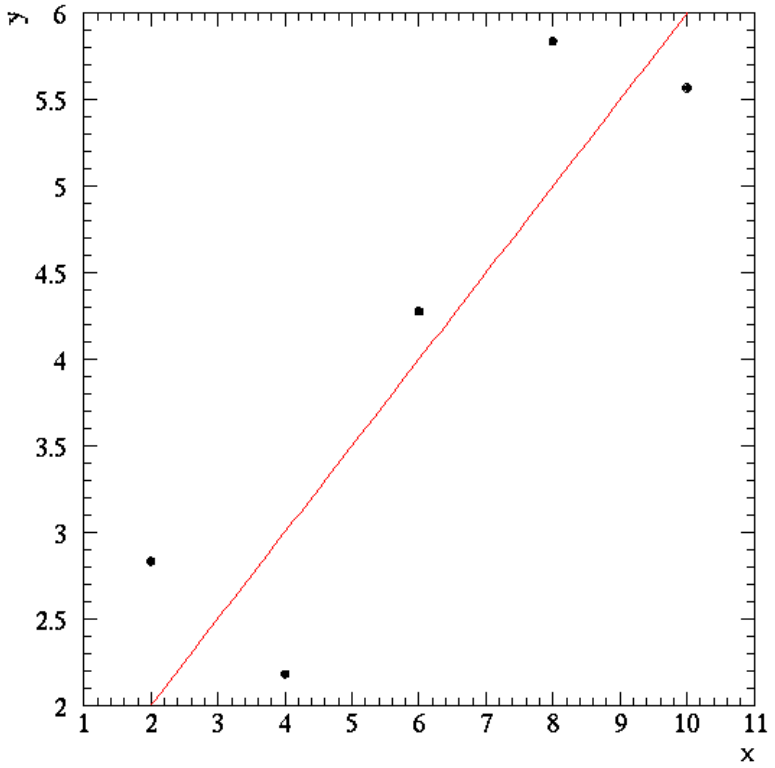
In this example each point is scattered around the model by an error uniformly distributed between -1 and +1.

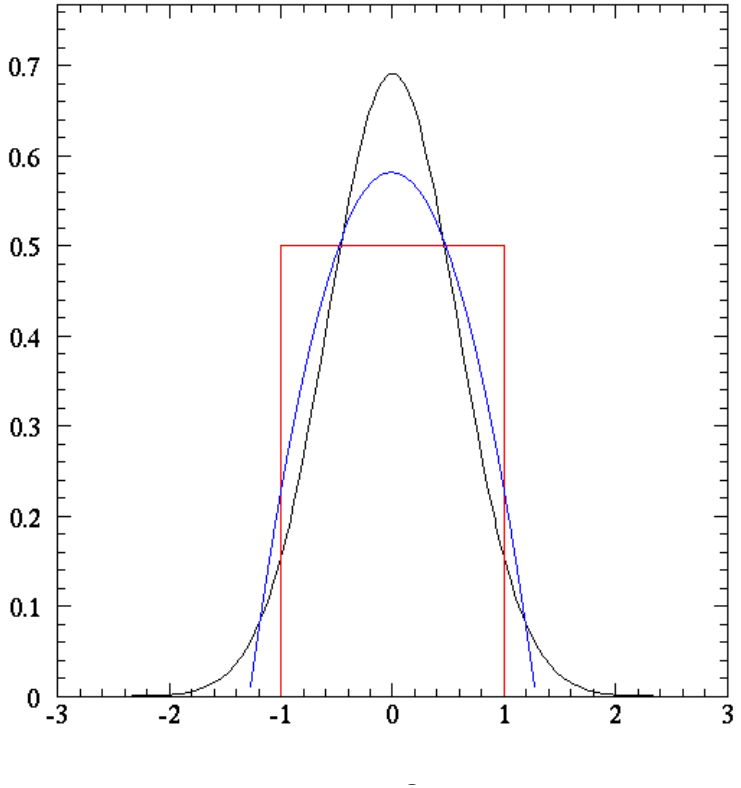
But suppose I don't know how the errors are distributed. What's the most conservative thing I can assume?

A Gaussian error distrib.

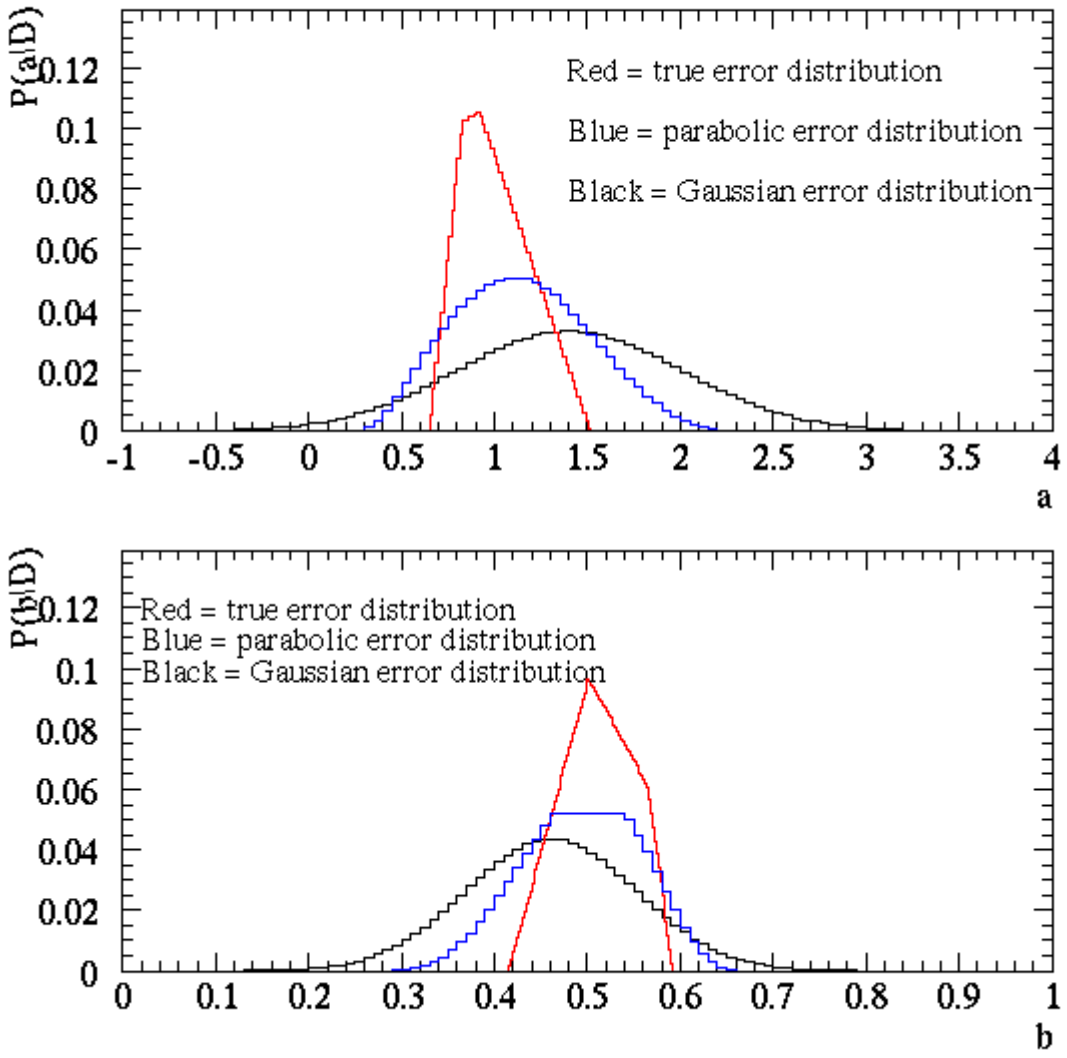
Consider three possible error models

I don't know how the errors are distributed, but I happen to know the RMS of the data around the model by some means. (Maybe Zeus told me.) I consider three possible models for the error: uniform, Gaussian, and parabolic.





Posterior probability distributions for the three error models



These are marginalized PDFs.

Caveat: although in this case the true error distribution gave the tightest parameter constraints, it's perfectly possible for an incorrect assumption about the error distribution to give inappropriately tight constraints!

What if you don't know the RMS?

Imagine that the data is so sparse that you don't already know the scatter of the data around the model.

One possibility is to assume a Gaussian distribution for the errors a la the maximum entropy principle, but to leave σ^2 as a free parameter. Assign it a physically plausible prior (possibly a Jeffreys prior over physically plausible range) and just treat it as a nuisance parameter.

This is more or less like “fitting” for the size of the error.

A prior gotcha

Maybe an obvious point ... if your prior ever equals zero at some value, then your posterior distribution must equal zero at that value as well, no matter what your data says.

Be cautious about choosing priors that are identically zero over any range of interest.

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), or it could be a theoretical parameter like a mass.

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.

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

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

What if μ is also unknown?

A biased estimator:

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

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

An unbiased estimator:

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

But its square root is a biased estimator of σ!