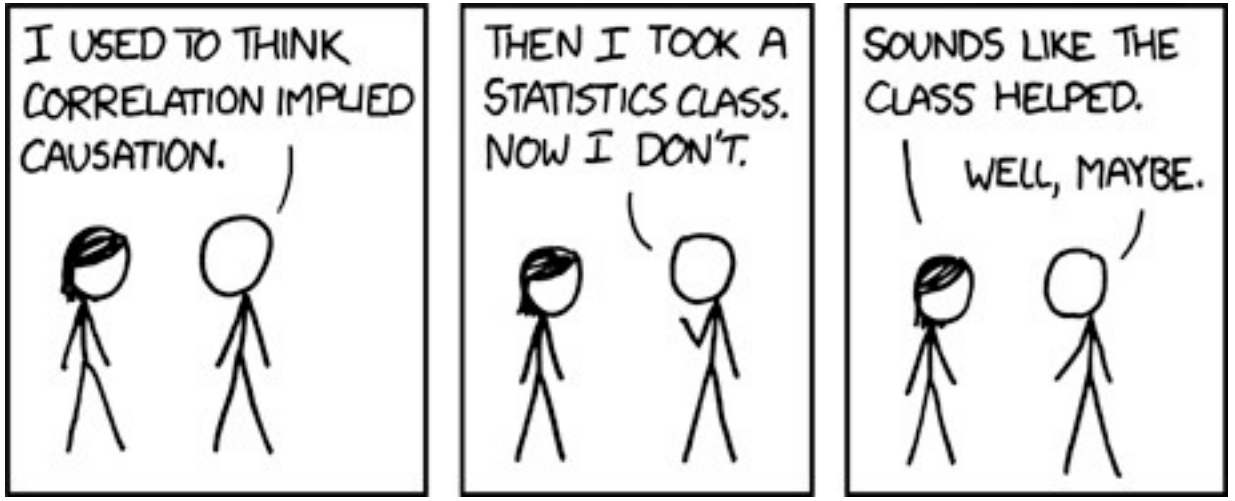
A Crash Course In Statistics

Summer Student Lecture

Scott Oser
UBC/TRIUMF
May 15, 2020



Basic mathematics of probability

1) Probabilities are numbers between 0 and 1.

2) P(A or B) = P(A) + P(B) – P(A & B)

3) Conditional probability: P(A & B) = P(B) P(A|B). Read as “the probability of B times the probability of A given B”.

4) A special case of conditional probability: if A and B are *independent* of each other (nothing connects them), then

P(A & B) = P(A) P(B)

Probability Distribution Functions

Discrete distribution (cleanly separated outcomes):

P(H) = probability of H being true

Ex. H = “rolling two dice gives a total of 7”

Continuous distribution:

P(x) dx = probability that x lies in the range (x, x+dx)

Ex. probability of mean of N measurements being between 5.00 and 5.01

NORMALIZATION CONDITION:

$$\sum P\left(H_i\right)=1 \qquad \text{or} \qquad \int dx\, P\left(x\right)=1$$

Joint PDFs

We often have multi-dimensional probability distributions: *P(x,y)*, where X and Y are two random variables.

These have the obvious interpretation that
P(x,y) dx dy = probability that X is the range x to x+dx while simultaneously Y is in the range y to y+dy. This can trivially be extended to multiple variables, or to the case where one or more variables are discrete and not continuous.

Normalization condition still applies:

$$\int d\,\vec{x}_i\, P\left(\vec{x}_i\right) = 1$$

To generate a probability distribution for one variable only, we *marginalize* by integrating over the unwanted variable(s):

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

The Centre of the Data: Mean, Median, & Mode

Mean of a data set:

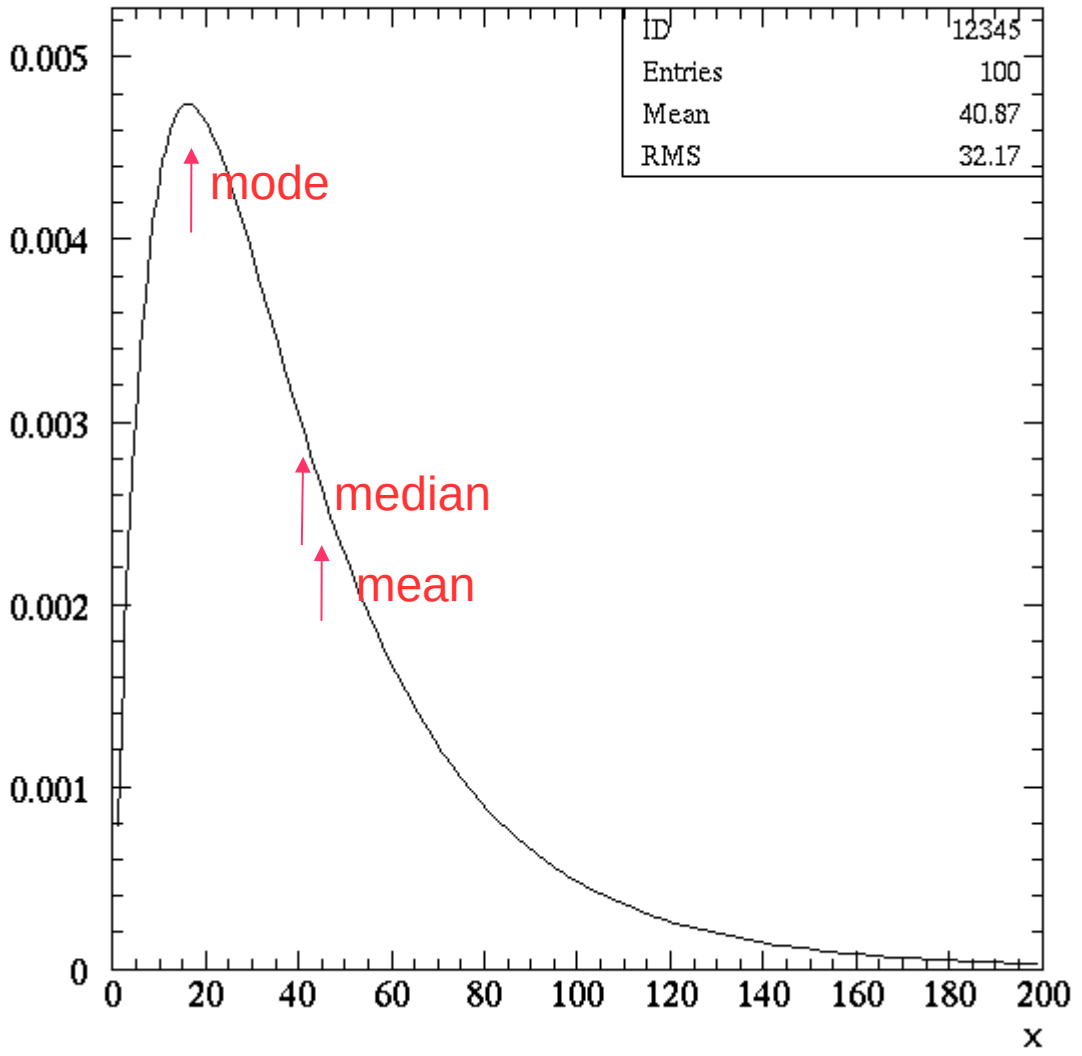
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Mean of a PDF = expectation value of x

$$\mu \equiv \langle x \rangle \equiv \int dx\, P(x)\, x$$

Median: the point with 50% probability above & 50% below. Less sensitive to tails!

Mode: the most likely value



The Width: Variance *V* / standard Deviation σ

Variance of a distribution: $V(x) \equiv \sigma^2 = \int dx\, P(x)(x-\mu)^2$

$$V(x) = \int dx\, P(x) x^2 - 2\mu \int dx\, P(x) x + \mu^2 \int dx\, P(x) = \langle x^2 \rangle - \mu^2 = \langle x^2 \rangle - \langle x \rangle^2$$

Variance of a data sample (regrettably has same notation as variance of a distribution---be careful!):

$$V(x) = \sigma^2 = \frac{1}{N} \sum_i (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

An important point: the above formula underestimates the variance of the underlying distribution, since it uses the mean calculated from the data instead of the true mean μ of the true distribution.

$$\hat{V}(x) = \sigma^2 = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

This is unbiased if you must estimate the mean from the data.

$$V(x) = \sigma^2 = \frac{1}{N} \sum_i (x_i - \mu)^2$$

Use this if you know the true mean of the underlying distribution.

Covariance & Correlation

The covariance between two variables is defined by:

$$\mathrm{cov}(x,y) = \langle (x-\mu_x)(y-\mu_y) \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

This is the most useful thing they never tell you in most lab courses! Note that cov(x,x)=V(x).

The correlation coefficient is a unitless version of the same thing:

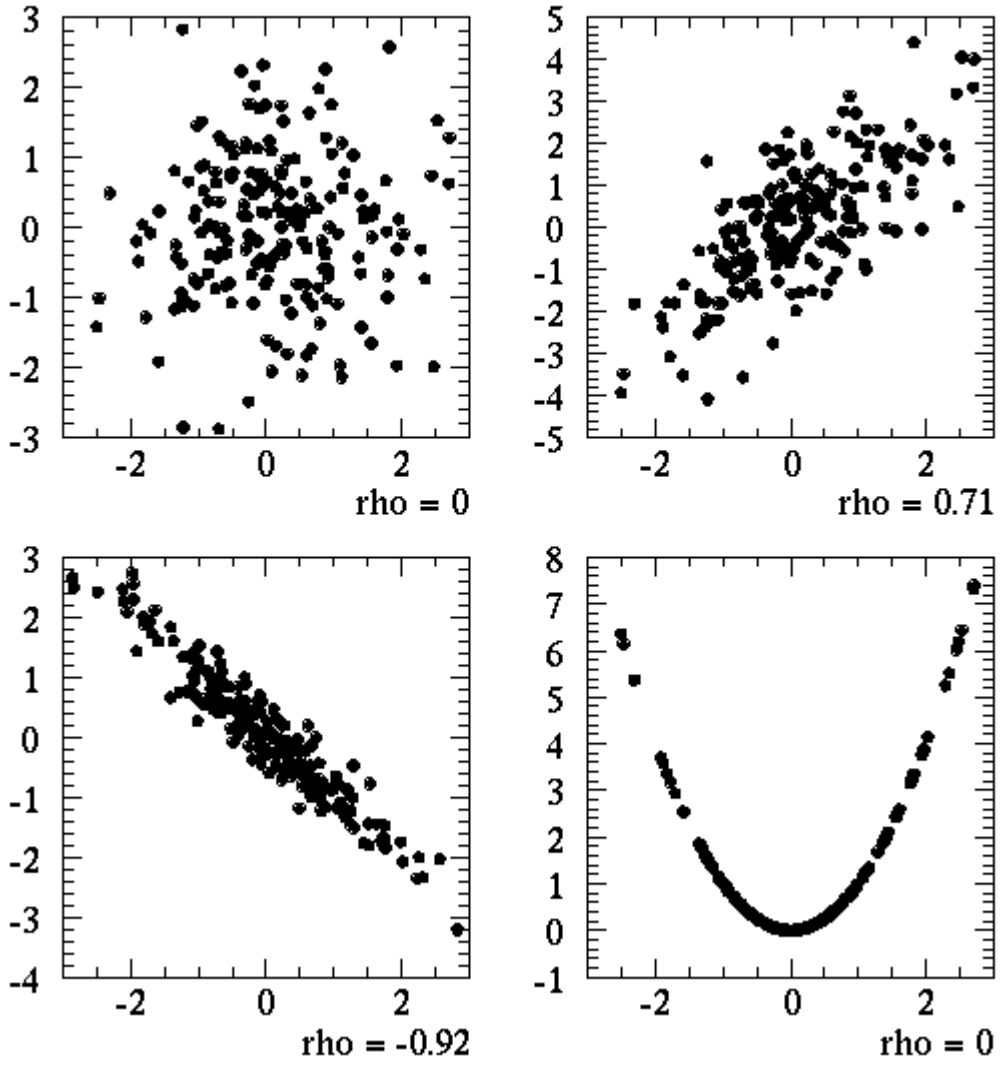
$$\rho = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$$

If x and y are independent variables *(P(x,y) = P(x)P(y))*, then

$$\mathrm{cov}(x,y) = \int dx\,dy\, P(x,y)\, xy - \left(\int dx\,dy\, P(x,y)\, x \right) \left(\int dx\,dy\, P(x,y)\, y \right)$$

$$\int dx\, P(x)\, x \int dy\, P(y)\, y - \left(\int dx\, P(x)\, x \right) \left(\int dy\, P(y)\, y \right) = 0$$

More on Covariance



Correlation coefficients for some simulated data sets.

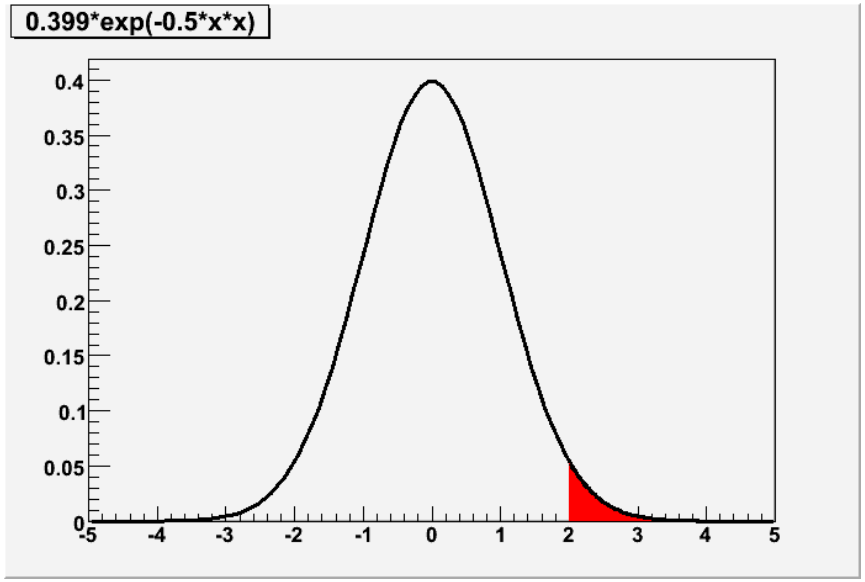
Note the bottom right---while independent variables must have zero correlation, the reverse is not true!

Correlation is important because it is part of the error propagation equation, as we'll see.

Gaussian Distributions

By far the most useful distribution is the Gaussian (normal) distribution:

$$P(x\,|\,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Mean = μ, Variance=σ^2

Note that width scales with σ.

Area out on tails is important---use lookup tables or cumulative distribution function.

In plot to left, red area (>2σ) is 2.3%.

68.27% of area within $\pm 1\sigma$
95.45% of area within $\pm 2\sigma$
99.73% of area within $\pm 3\sigma$

90% of area within $\pm 1.645\sigma$
95% of area within $\pm 1.960\sigma$
99% of area within $\pm 2.576\sigma$

Why are Gaussian distributions so critical?

They occur very commonly---the reason is that the average of several independent random variables often approaches a Gaussian distribution in the limit of large N.

Nice mathematical properties---infinitely differentiable, symmetric. Sum or difference of two Gaussian variables is always itself Gaussian in its distribution.

Many complicated formulas simplify to linear algebra, or even simpler, if all variables have Gaussian distributions.

Gaussian distribution is often used as a shorthand for discussing probabilities. A "5 sigma result" means a result with a chance probability that is the same as the tail area of a unit Gaussian:

$$2\int_{5}^{\infty} dt\, P\left(t \,|\, \mu = 0, \sigma = 1\right)$$

This way of speaking is used even for non-Gaussian distributions!

The Central Limit Theorem

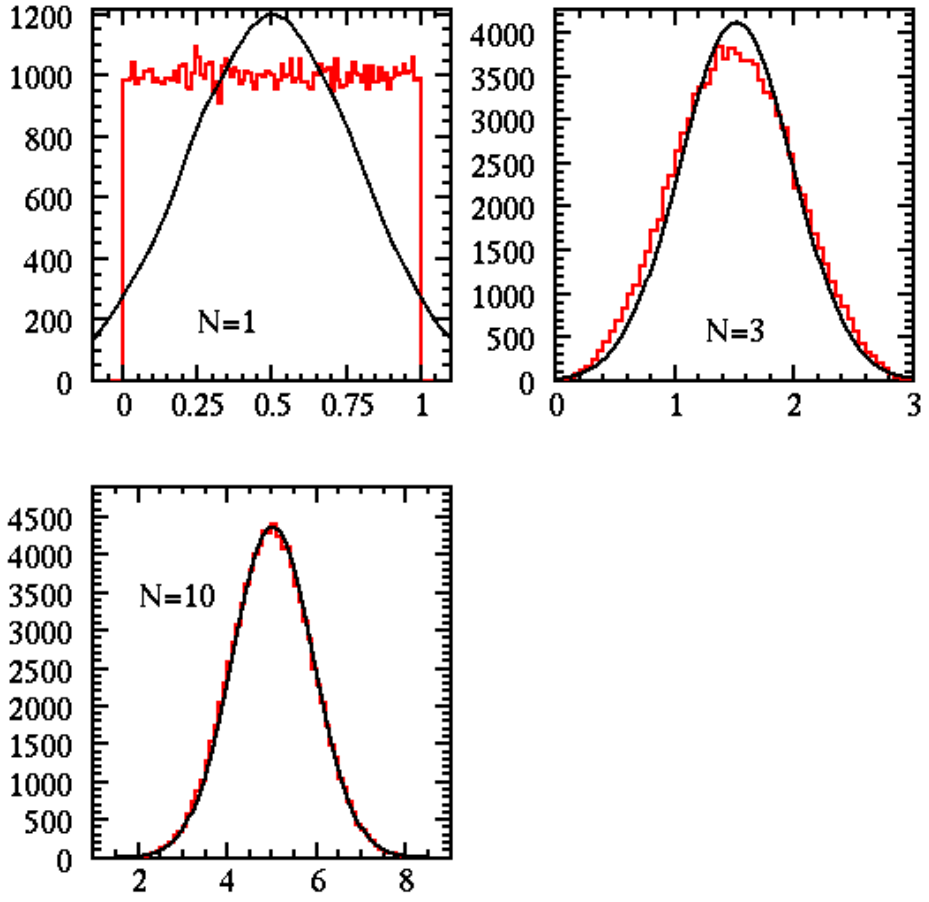
If X is the sum of N independent random variables x_i, each taken from a distribution with mean μ_i and variance σ_i^2, then the distribution for X approaches a Gaussian distribution in the limit of large N. The mean and variance of this Gaussian are given by:

$$\langle X \rangle = \sum \mu_i$$

$$V(X) = \sum V_i = \sum \sigma_i^2$$

The Central Limit Theorem: the caveats

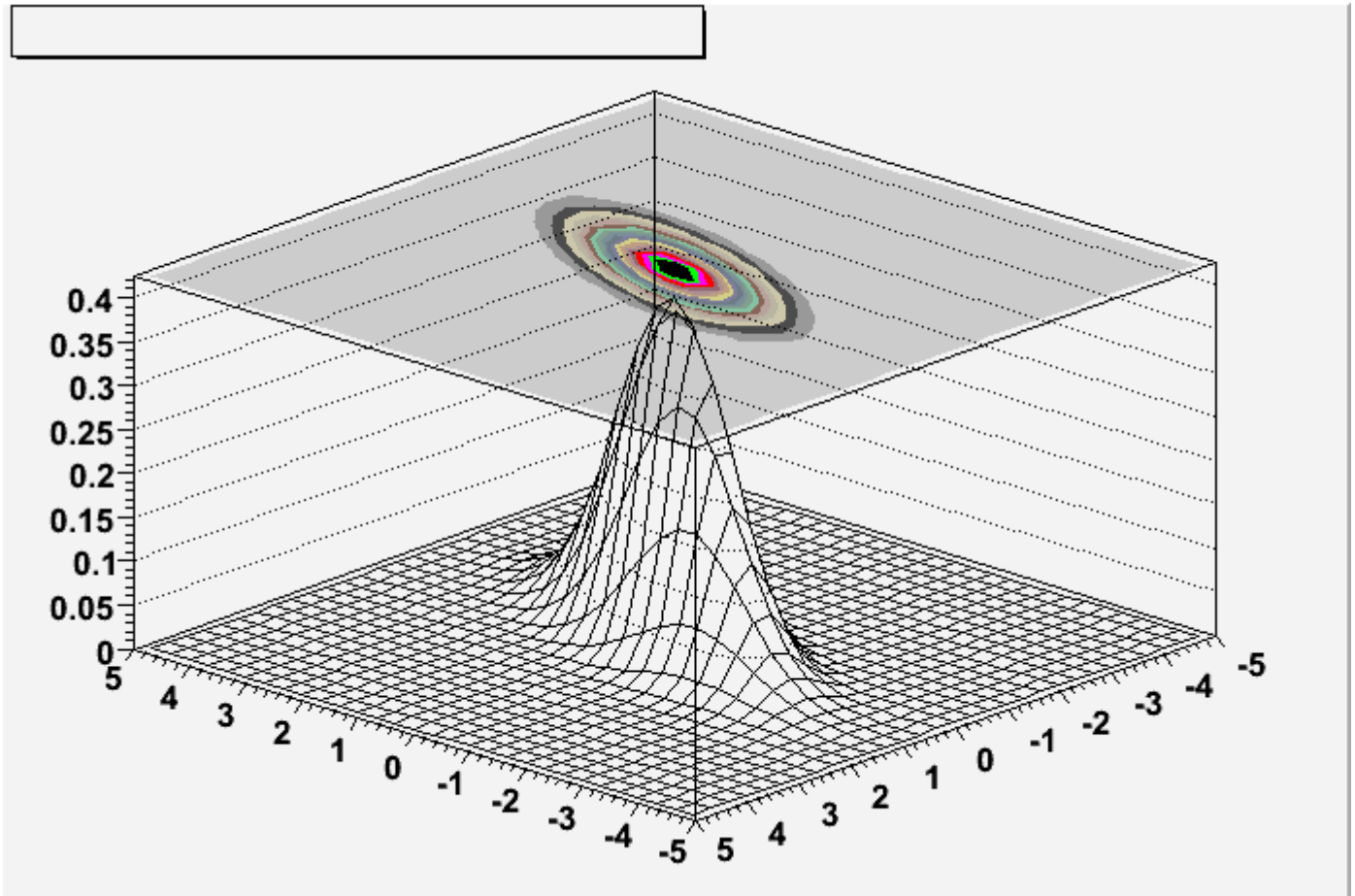
- I said N *independent* variables!
- Obviously the variables must individually have finite variances.
- I've said nothing about *how fast* the distribution approaches a Gaussian as N goes to infinity. But it can be *fast*! Consider below the sum of N uniformly distributed variables



The General Multidimensional Gaussian ...

$$P(\vec{x}) \propto \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T \cdot V^{-1} \cdot (\vec{x}-\vec{\mu})\right]$$

Parametrized by vector of means, μ, and covariance matrix V.



Interpretations of probability

Two interpretations of what we mean by probability:

1) The frequentist school: Probability is a statement about frequency.
If you repeat a measurement 1000 times and get the same outcome
200 times, the probability of that outcome is 0.2.

2) The Bayesian school: Probability quantifies our certainty about
a statement, and hence is a statement about our knowledge.
People with different knowledge may assign different probabilities ...
while I say the probability of rain tomorrow is 1/3, you may have reason
to believe otherwise and may rightfully assign a different probability.
In this sense probability estimates depend on the information we possess.

Problems with the frequentist interpretation

1) We naturally want to talk about the probability of events that are not repeatable even in principle. Tomorrow only happens once--- can we meaningfully talk about frequencies of its weather? Maybe we want to talk about the probability of some cosmological parameter, but we only have one universe! A strict interpretation of probability as frequency says that we cannot use the concept of probability in this way.

2) Probability depends on the choice of ensemble you compare to, which may be non-obvious. The probability of someone in a crowd of people being a physicist depends on whether you are talking about a crowd at a hockey game, a crowd at a university club, or a crowd at in this audience.

In spite of these conceptual problems, the “frequentist interpretation” is the most usual interpretation used in science.

The Bayesian interpretation

This goes most commonly by the name “Bayesian statistics”. In this view probability is a way of quantifying our knowledge of a situation. P(E)=1 means that it is 100% certain that E is the case. Our estimation of P depends on how much information we have available, and is subject to revision.

The Bayesian interpretation is the cleanest conceptually, and actually is the oldest interpretation. Although it is gaining in popularity in recent years, it's still a minority view. The main objections are:

1) As a statement about our knowledge of a situation, Bayesian probabilities seem “subjective”. Science is supposed to be an objective subject.
2) It is not always obvious how to quantify the prior state of our knowledge upon which we base our probability estimate.

Increasingly many people see these objections as not serious.

Frequentist vs. Bayesian Comparison

Bayesian Approach

"The probability of the particle's mass being between 1020 and 1040 MeV is 98%."

Considers the data to be known and fixed, and calculates probabilities of hypotheses or parameters.

Requires *a priori* estimation of the model's likelihood, naturally incorporating prior knowledge.

Well-defined, automated "recipe" for handling almost all problems.

Frequentist Approach

- "If the true value of the particle's mass is 1030 MeV, then if we repeated the experiment 100 times only twice would we get a measurement smaller than 1020 or bigger than 1040."
- Considers the model parameters to be fixed (but unknown), and calculates the probability of the data given those parameters.
- Many "ad hoc" approaches required depending on question being asked. Not all consistent!

Bayes' Theorem

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

H = a hypothesis (e.g. “the electron's mass is in the range 9.10 – 9.11 x10^{-31} kg”)

D = the data

P(H) = the “prior probability” for H

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

P(D) = 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?

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

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

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

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

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

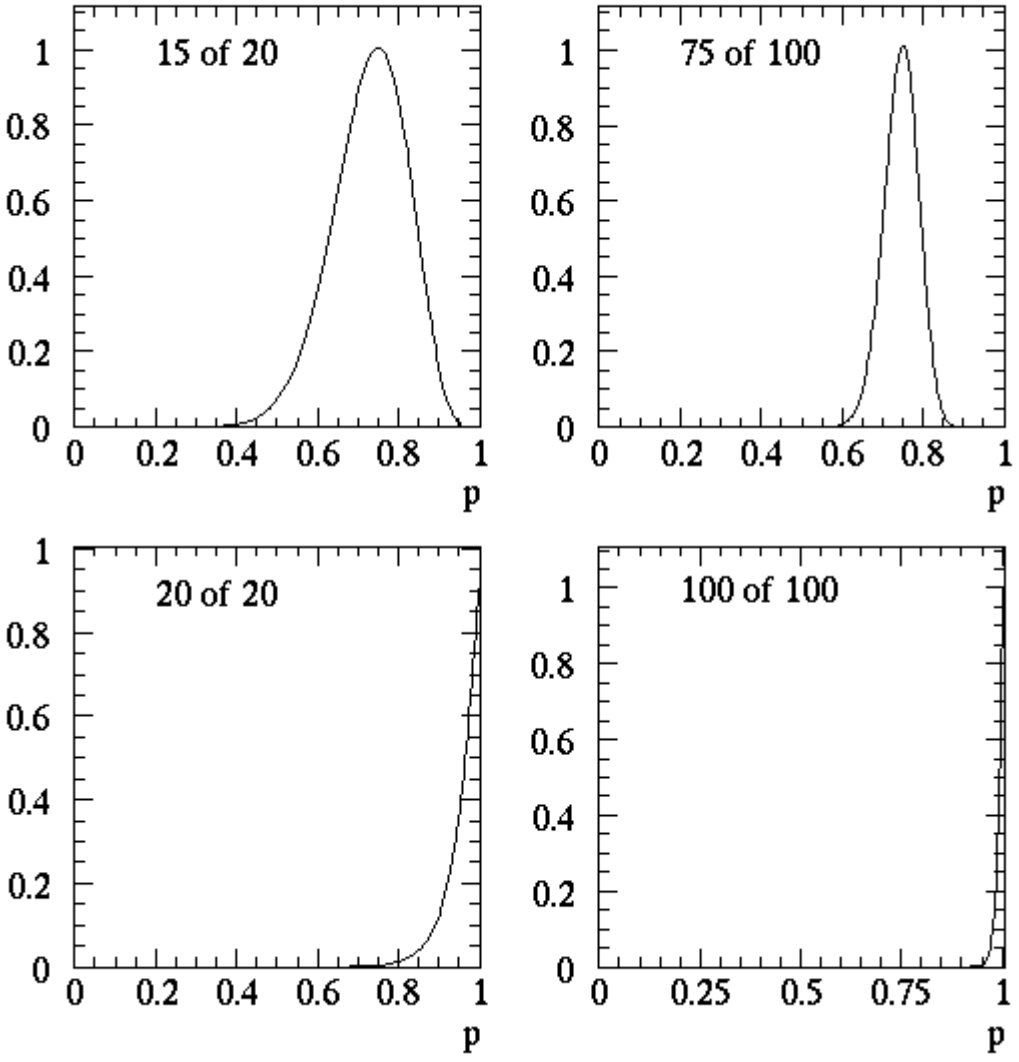
$$P(D|p) = \frac{N!}{m!(N-m)!} p^m (1-p)^{N-m}$$

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

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

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

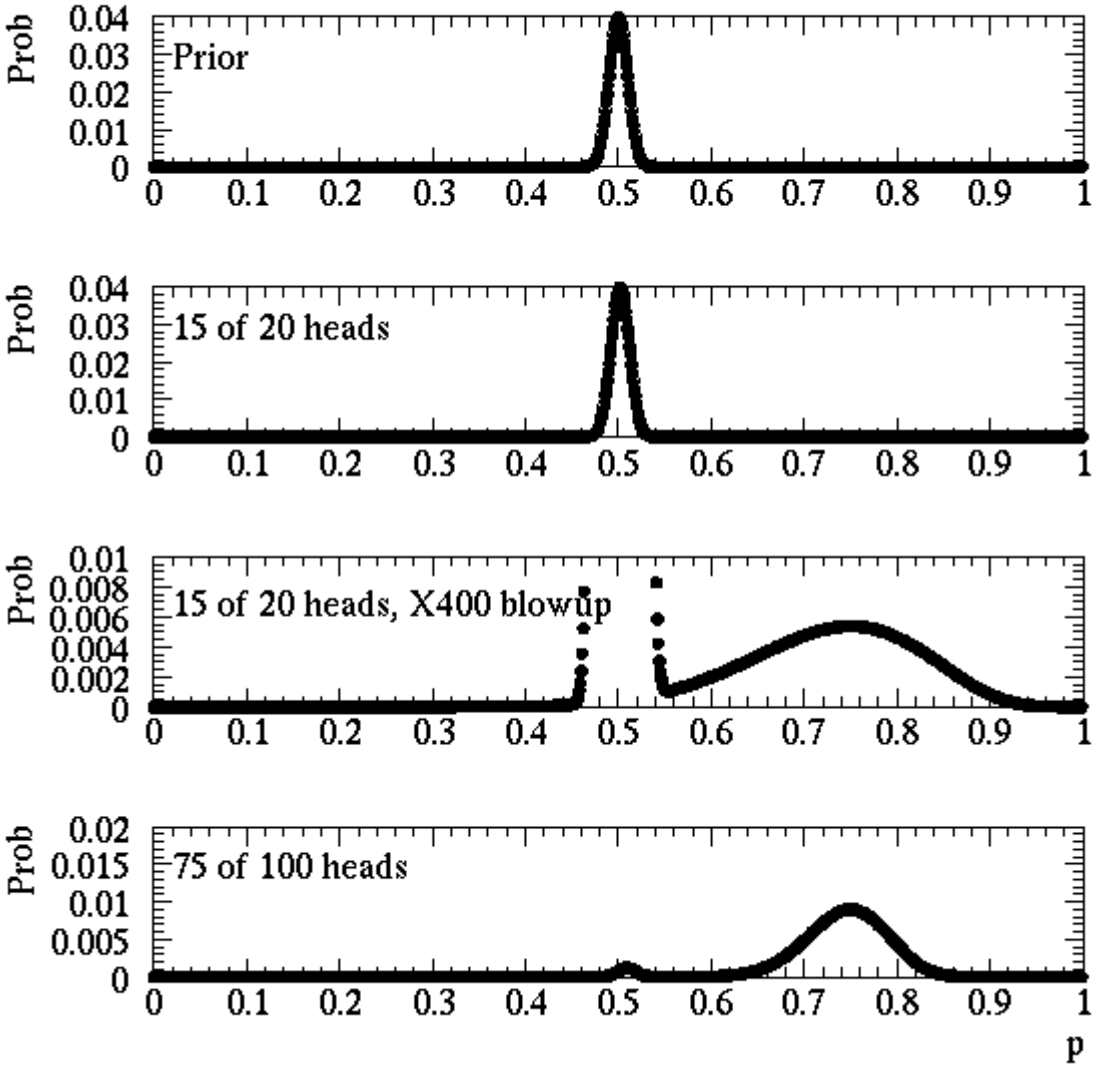
If a cashier hands you a coin as change, is it really reasonable to assume a uniform prior for *p*? Unbalanced coins must be really rare!

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): 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)

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 $p(X)\,dX = P(X')\,dX' = P(X')\,dX$. 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.

Jeffreys Prior

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

$$P(T)\,dT = P(T')\,dT' = P(T')\,\beta\,dT$$

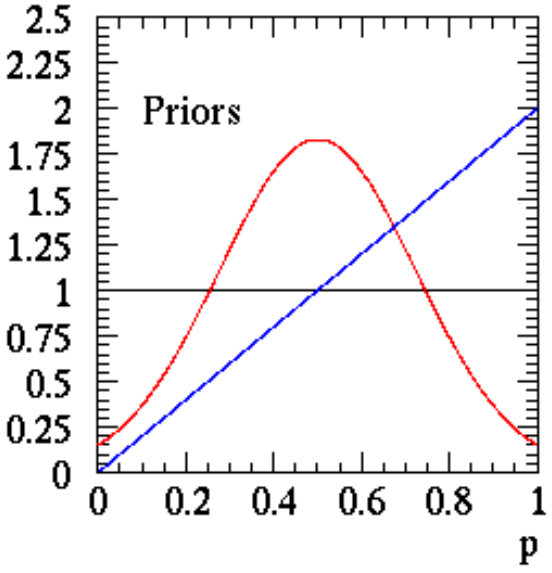
$$\therefore P(T) = \beta P(\beta T), \text{ which is only true for all } \beta \text{ if}$$

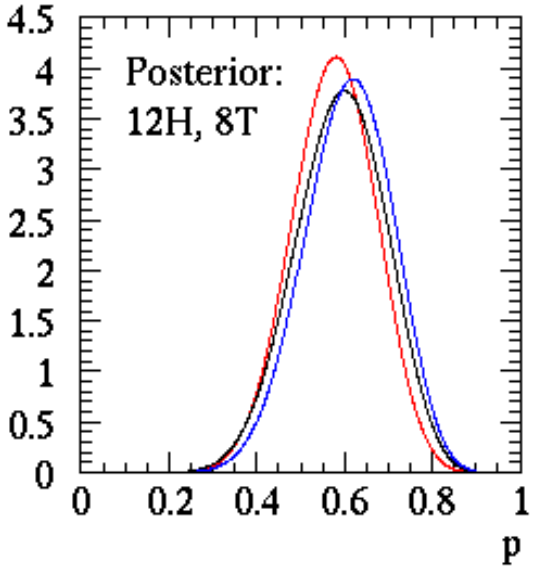
$$P(T) = \frac{\text{constant}}{T}$$

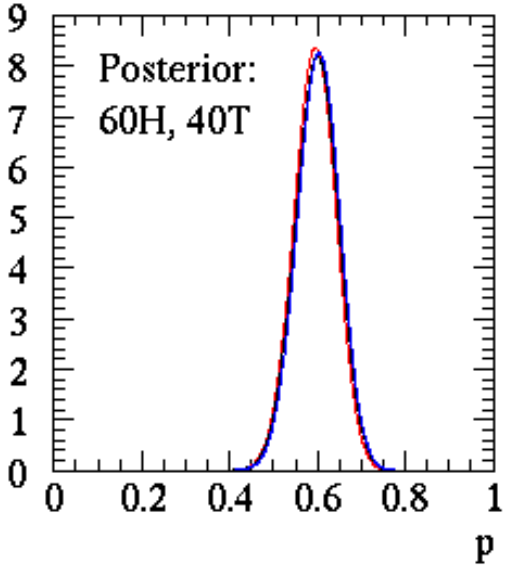
Properly normalized from T_{min} to T_{max} this is:

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

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.

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*:

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

The PDF P(*a*|D) 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.”

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

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

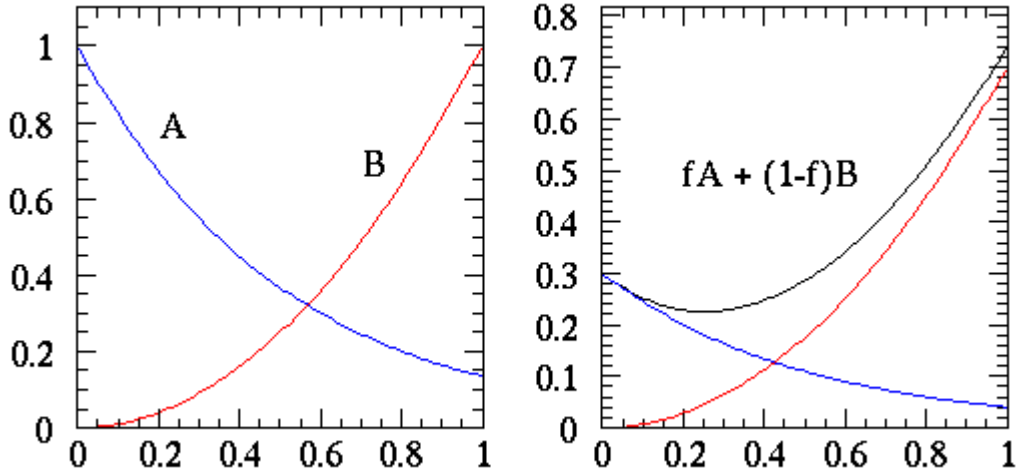
It's more common (and easier) to calculate -ln *L* instead:

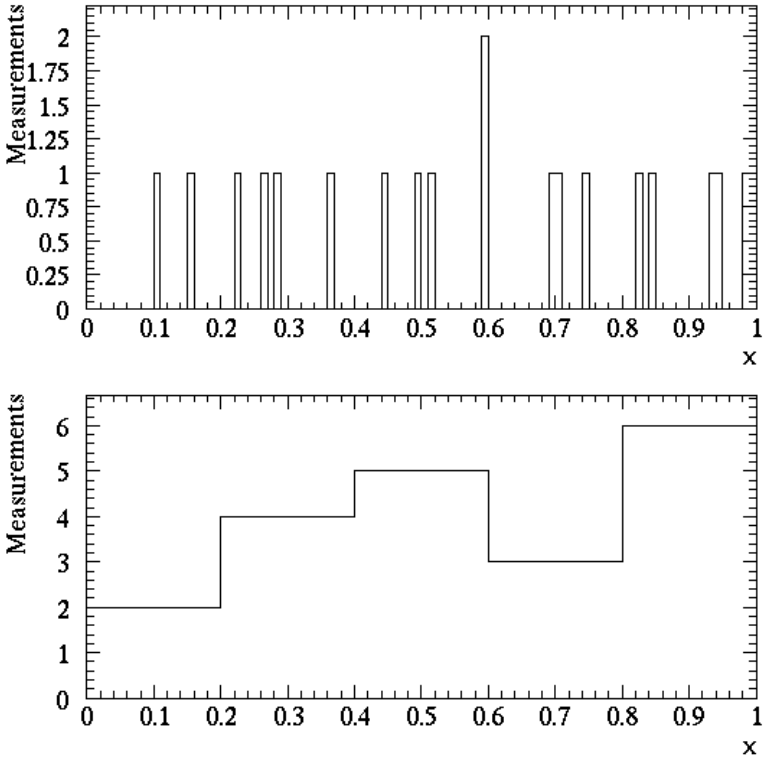
$$-\ln L\left(x_1 \ldots x_N | \alpha\right) = -\sum_{i=1}^{N} \ln P\left(x_i | \alpha\right)$$

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.

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

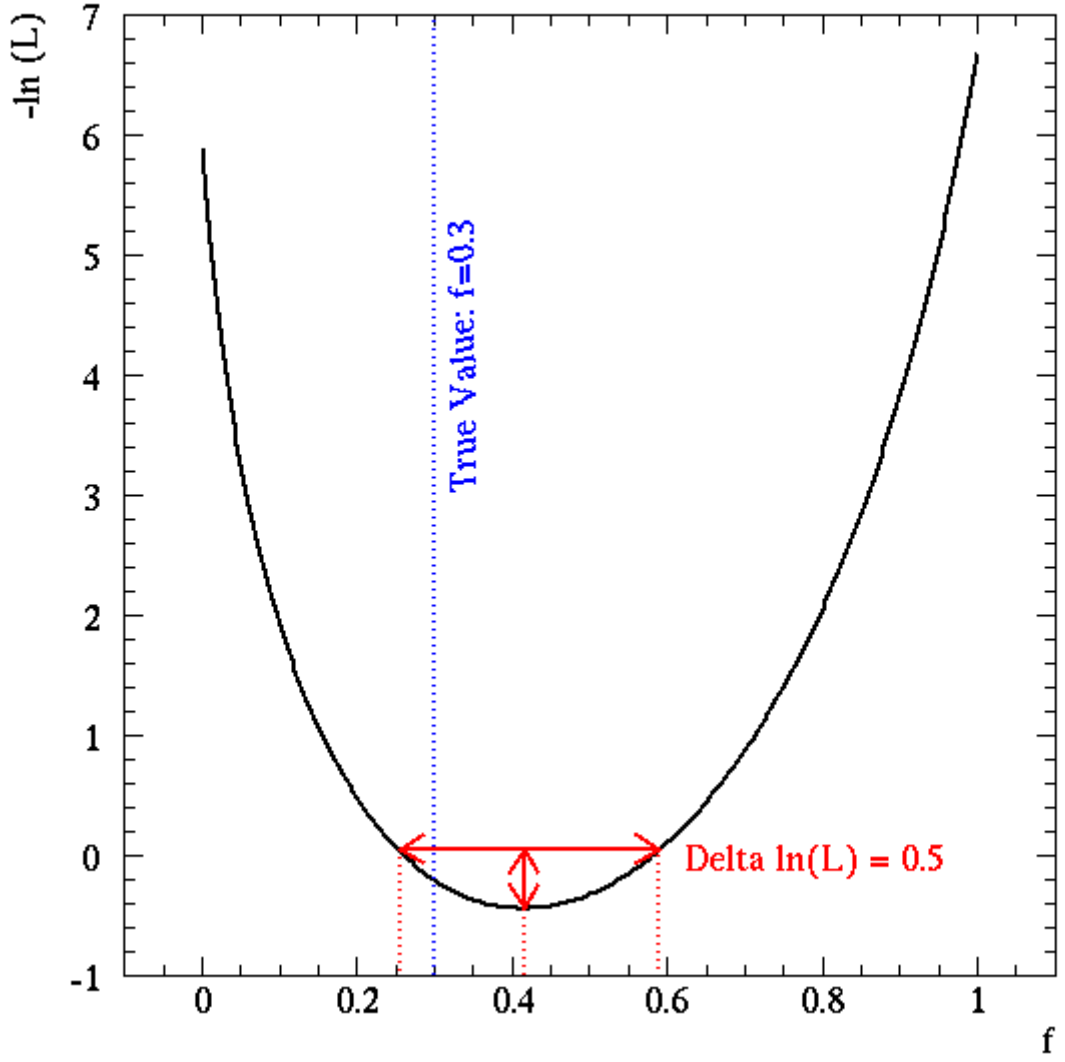
Form the negative log likelihood:

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

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

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.

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

The log likelihood is then

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

Maximizing this is equivalent to minimizing

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

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) and the prior P(a).

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

The log likelihood is related to the Bayesian PDF by:

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

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.

The Least Squares Method

Taken outside the context of the ML method, the least squares method is the most commonly known estimator.

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

Why?

1) Easily implemented.
2) Mathematically straightforward---often analytic solution
3) Extension of LS to correlated uncertainties
   straightforward:

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

Least Squares Straight Line Fit

The most straightforward example is a linear fit: y=mx+b.

$$\chi^2 = \sum \left(\frac{y_i - mx_i - b}{\sigma_i} \right)^2$$

Least squares estimators for m and b are found by differentiating χ^2 with respect to m & b.

$$\frac{d\chi^2}{dm} = -2 \sum \left(\frac{y_i - mx_i - b}{\sigma_i^2} \right) \cdot x_i = 0$$

$$\frac{d\chi^2}{db} = -2 \sum \left(\frac{y_i - mx_i - b}{\sigma_i^2} \right) = 0$$

This is a linear system of simultaneous equations with two unknowns.

Solving for m and b

The most straightforward example is a linear fit: y=mx+b.

$$\frac{d\chi^2}{dm} = -2\sum\left(\frac{y_i - mx_i - b}{\sigma_i^2}\right)\cdot x_i = 0 \qquad \frac{d\chi^2}{db} = -2\sum\left(\frac{y_i - mx_i - b}{\sigma_i^2}\right) = 0$$

$$\sum\left(\frac{x_i y_i}{\sigma_i^2}\right) = m\sum\left(\frac{x_i^2}{\sigma_i^2}\right) + b\sum\left(\frac{x_i}{\sigma_i^2}\right) \qquad \sum\left(\frac{y_i}{\sigma_i^2}\right) = m\sum\left(\frac{x_i}{\sigma_i^2}\right) + b\sum\left(\frac{1}{\sigma_i^2}\right)$$

$$\hat{m} = \frac{\left(\sum\frac{y_i}{\sigma_i^2}\right)\left(\sum\frac{x_i}{\sigma_i^2}\right) - \left(\sum\frac{1}{\sigma_i^2}\right)\left(\sum\frac{x_i y_i}{\sigma_i^2}\right)}{\left(\sum\frac{x_i}{\sigma_i^2}\right)^2 - \left(\sum\frac{x_i^2}{\sigma_i^2}\right)\left(\sum\frac{1}{\sigma_i^2}\right)}$$

$$\hat{b} = \frac{\left(\sum\frac{y_i}{\sigma_i^2}\right) - \hat{m}\left(\sum\frac{x_i}{\sigma_i^2}\right)}{\left(\sum\frac{1}{\sigma_i^2}\right)}$$

(Special case of equal σ's.)

$$\left(\hat{m} = \frac{\langle y\rangle\langle x\rangle - \langle xy\rangle}{\langle x\rangle^2 - \langle x^2\rangle}\right)$$

$$\left(\hat{b} = \langle y\rangle - \hat{m}\langle x\rangle\right)$$

Solution for least squares m and b

There's a nice analytic solution---rather than trying to numerically minimize a χ^2, we can just plug in values into the formulas! This worked out nicely because of the very simple form of the likelihood, due to the linearity of the problem and the assumption of Gaussian errors.

$$\hat{m} = \frac{\left(\sum \frac{y_i}{\sigma_i^2}\right)\left(\sum \frac{x_i}{\sigma_i^2}\right) - \left(\sum \frac{1}{\sigma_i^2}\right)\left(\sum \frac{x_i y_i}{\sigma_i^2}\right)}{\left(\sum \frac{x_i}{\sigma_i^2}\right)^2 - \left(\sum \frac{x_i^2}{\sigma_i^2}\right)\left(\sum \frac{1}{\sigma_i^2}\right)}$$

$$\hat{b} = \frac{\left(\sum \frac{y_i}{\sigma_i^2}\right) - \hat{m}\left(\sum \frac{x_i}{\sigma_i^2}\right)}{\left(\sum \frac{1}{\sigma_i^2}\right)}$$

(Special case of equal errors)

$$\left(\hat{m} = \frac{\langle y \rangle \langle x \rangle - \langle xy \rangle}{\langle x \rangle^2 - \langle x^2 \rangle}\right)$$

$$\left(\hat{b} = \langle y \rangle - \hat{m} \langle x \rangle\right)$$

Errors in the Least Squares Method

What about the errors and correlations between m and b? Simplest way to derive this is to look at the chi-squared, and remember that this is a special case of the ML method:

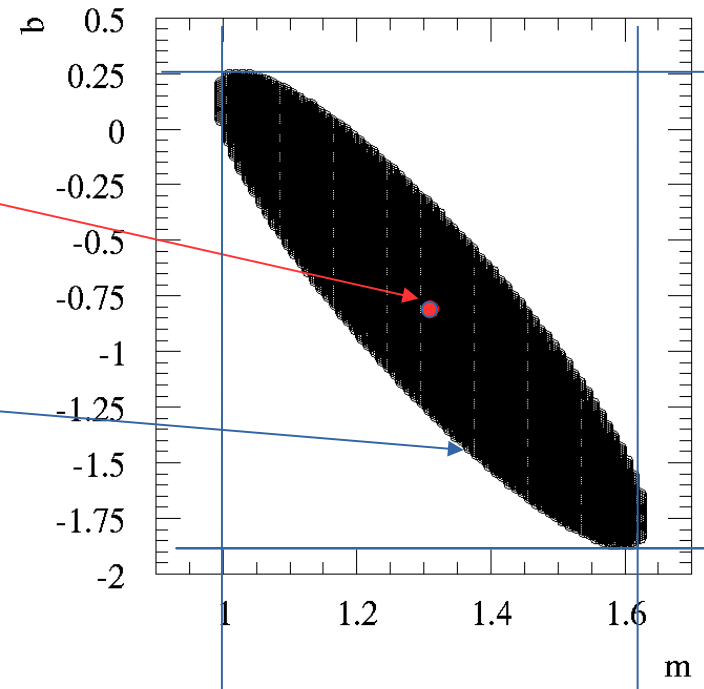
$$-\ln L = \frac{1}{2}\chi^2 = \frac{1}{2}\sum\left(\frac{y_i - mx_i - b}{\sigma_i}\right)^2$$

In the ML method, we define the 1σ error on a parameter by the minimum and maximum value of that parameter satisfying $\Delta \ln L = ½$.

In LS method, this corresponds to $\Delta\chi^2 = +1$ above the best-fit point. Two sigma error range corresponds to $\Delta\chi^2 = +4$, 3σ is $\Delta\chi^2 = +9$, etc.

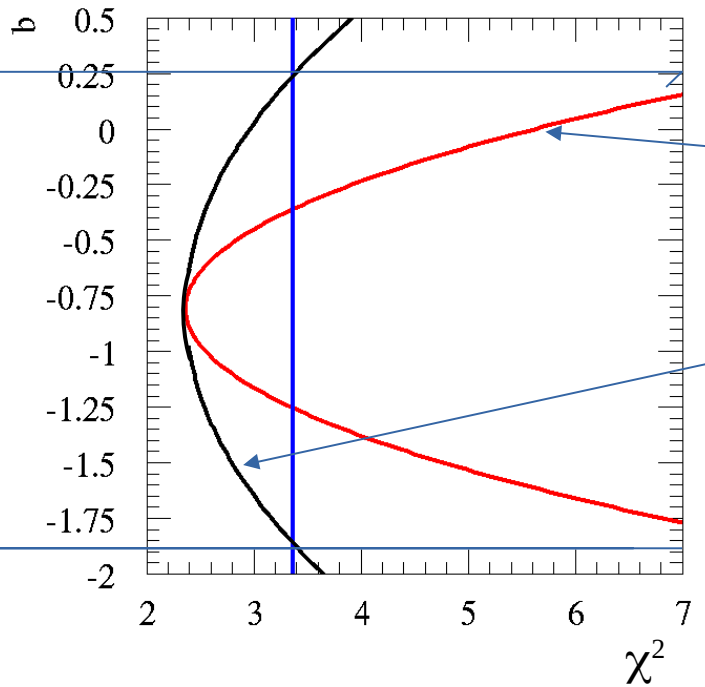
But notice one thing about the dependence of the χ^2---it is quadratic in both m and b, and generally includes a cross-term proportional to mb. Conclusion: Gaussian uncertainties on m and b, with a covariance between them.

Contours and marginalization



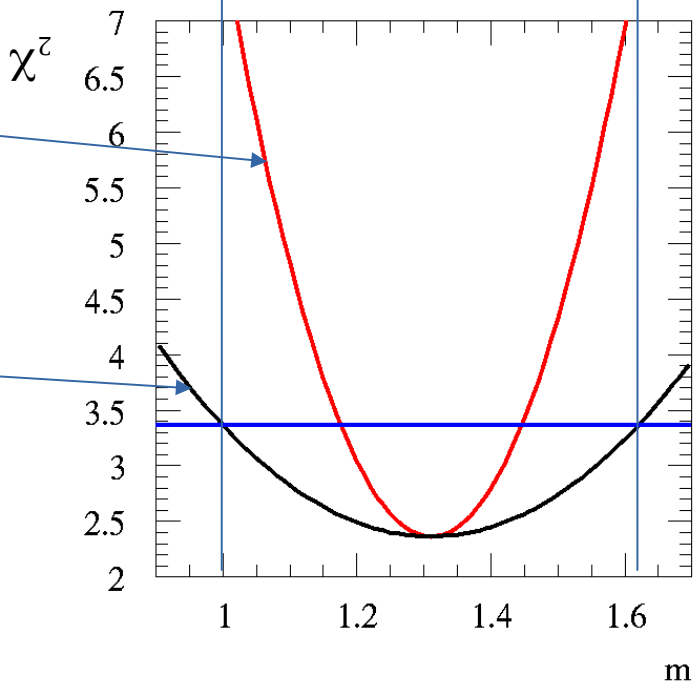
Best fit point

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



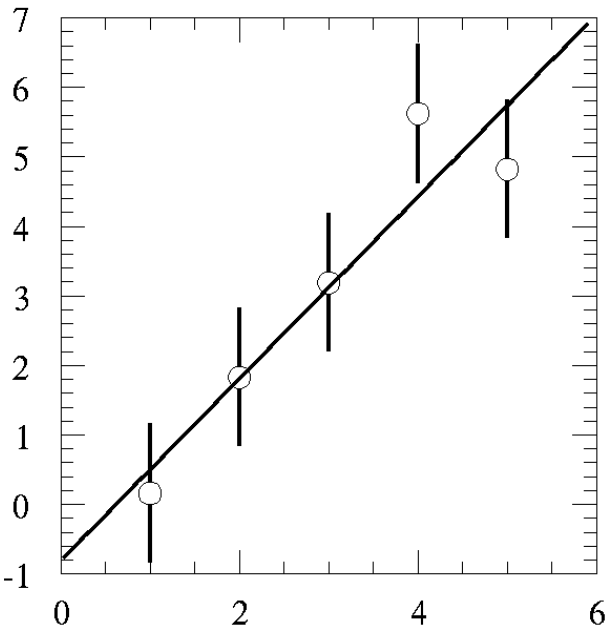
χ^2 vs b for fixed value of m

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

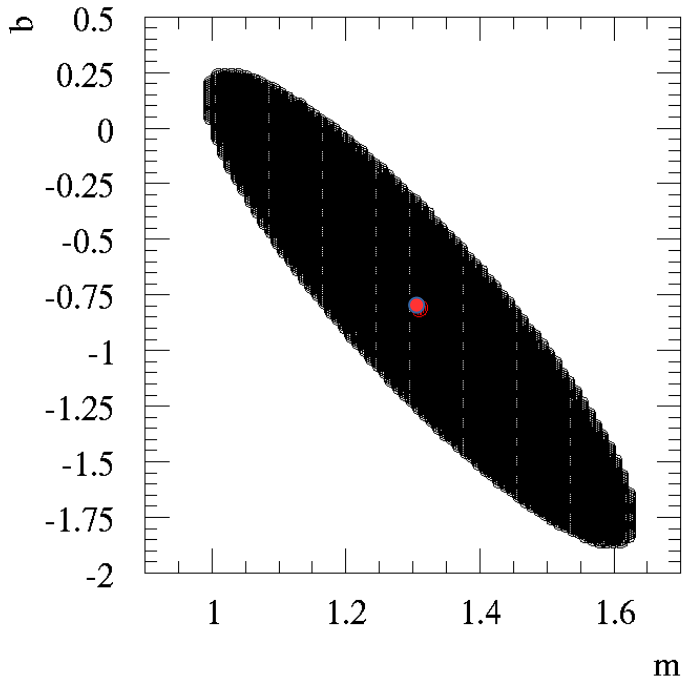
Errors on each individual parameter

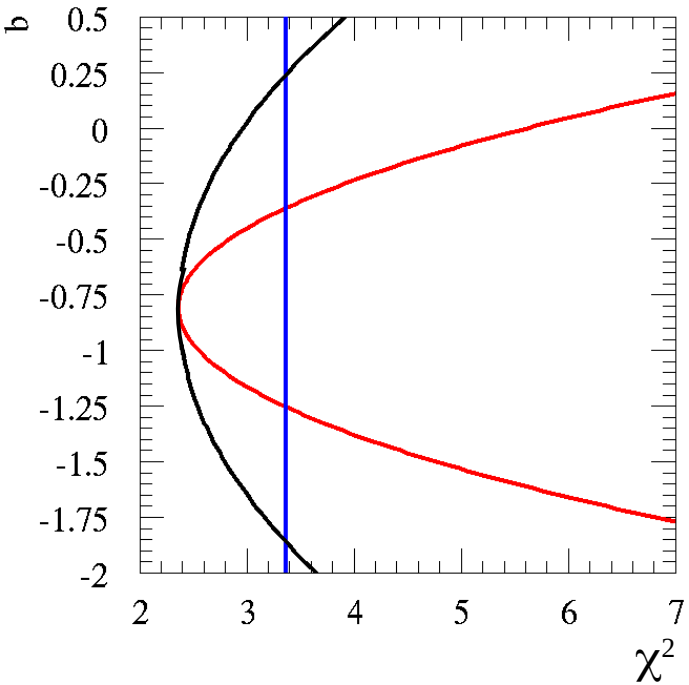
To find 1σ error on any parameter, scan over that parameter while minimizing the χ^2 as a function of all other free parameters.

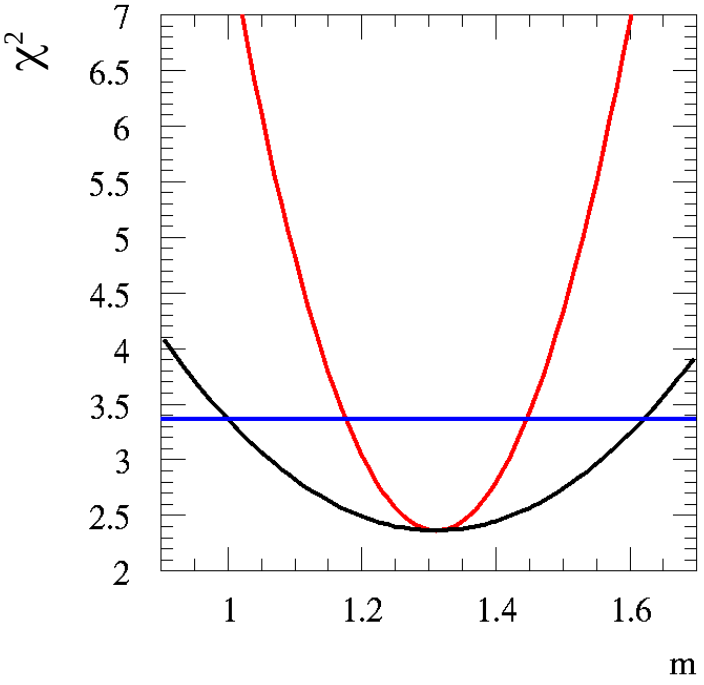
The points at which the χ^2 (minimized with respect to all other free parameters) has increased by +1 from its global minimum give the 1σ errors on the parameter.

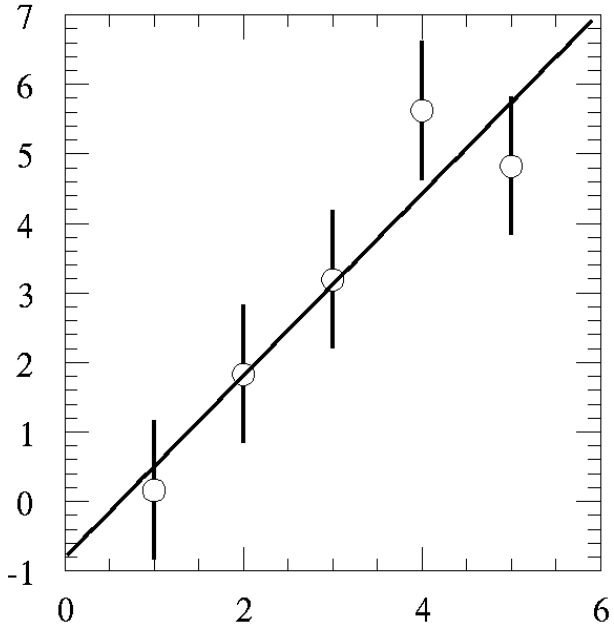
Do NOT leave the other parameters fixed at their best-fit values while scanning!

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

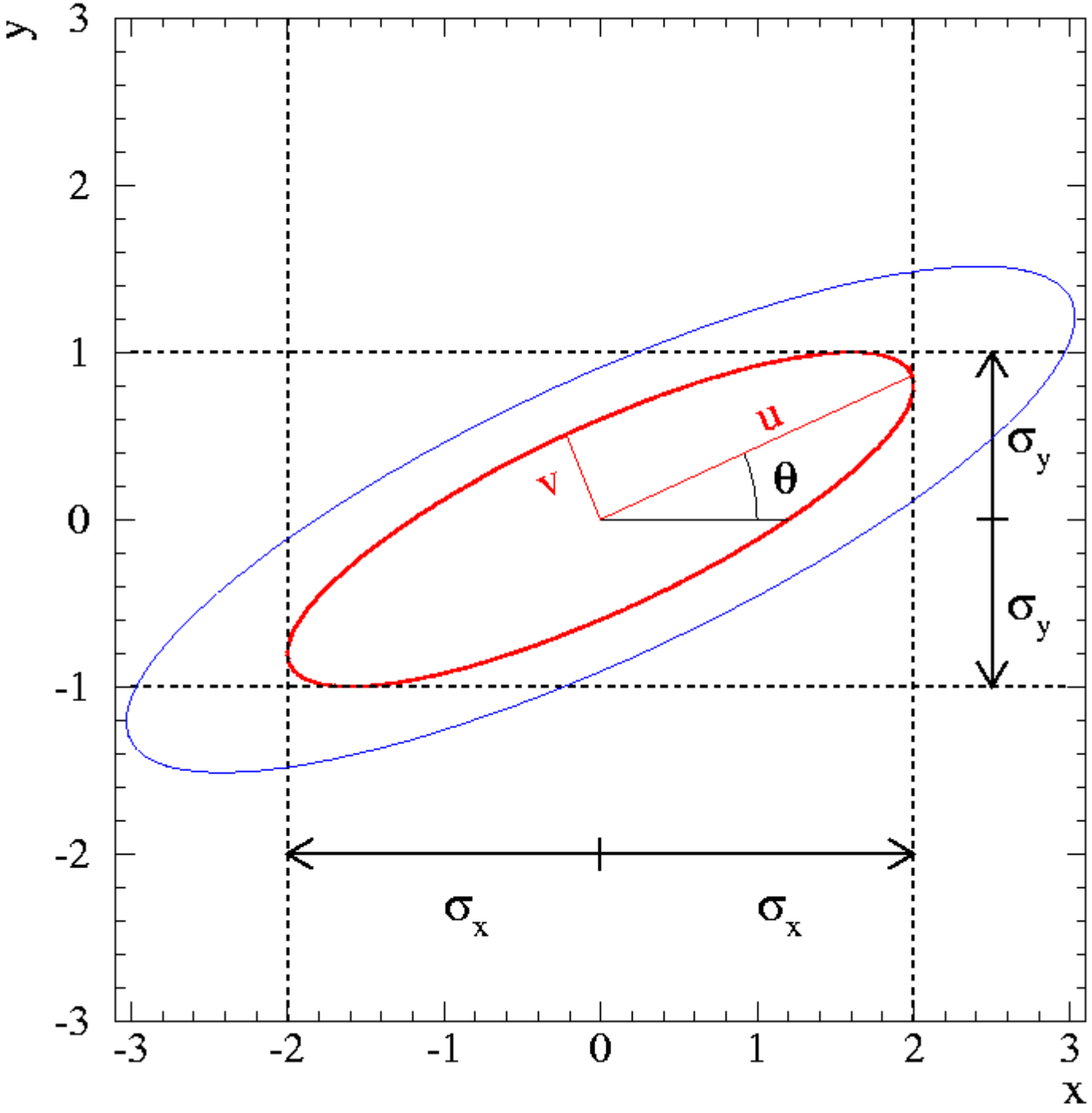








1D vs. 2D confidence regions



σ_x=2

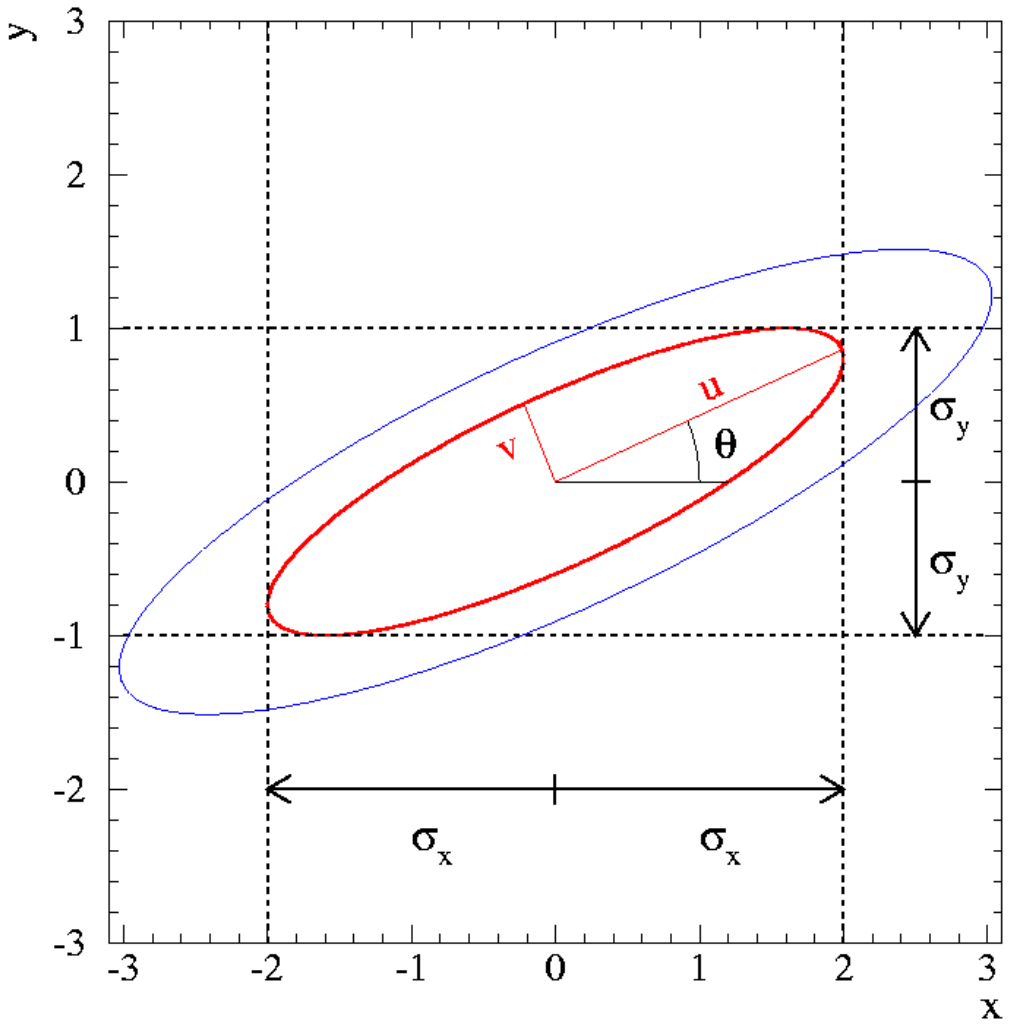
σ_y=1

ρ=0.8

Red ellipse: contour with $\Delta \ln L = +0.5$

Blue ellipse: contour containing 68% of 2D probability content.

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%	3.32	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:

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

(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.

Linear least squares and matrix algebra

Least squares fitting really shines in one area: linear parameter dependence in your fit function:

$$y\left(x\,|\,\vec{\alpha}\right)=\sum_{j=1}^{m}\alpha_j\cdot f_j\left(x\right)$$

In this special case, LS estimators for the α are unbiased, have the minimum possible variance of any linear estimators, and can be solved analytically, even when N is small, and independent of the individual measurement PDFs.†

$$A_{ij}=\begin{bmatrix} f_1\left(x_1\right) & f_2\left(x_1\right) & \cdots \\ f_1\left(x_2\right) & f_2\left(x_2\right) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

$$y_{\text{pred}}=A\cdot\vec{\alpha}$$

$$\chi^2=\left(\vec{y}_{meas}-\vec{y}_{pred}\right)^T\cdot V^{-1}\cdot\left(\vec{y}_{meas}-\vec{y}_{pred}\right)$$

$$\chi^2=\left(\vec{y}_{meas}-A\cdot\vec{\alpha}\right)^T\cdot V^{-1}\cdot\left(\vec{y}_{meas}-A\cdot\vec{\alpha}\right)$$

†Some conditions apply---see Gauss-Markov theorem for exact statement.

Linear least squares: exact matrix solution

$$y(x|\vec{\alpha}) = \sum_{j=1}^{m} \alpha_j \cdot f_j(x)$$

$$y_{\text{pred}} = A \cdot \vec{\alpha}$$

$$A_{ij} = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots \\ f_1(x_2) & f_2(x_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

$$\chi^2 = \left(\vec{y}_{meas} - A \cdot \vec{\alpha}\right)^T \cdot V^{-1} \cdot \left(\vec{y}_{meas} - A \cdot \vec{\alpha}\right)$$

Best fit estimator:

$$\vec{\alpha} = \left(A^T V^{-1} A\right)^{-1} A^T V^{-1} \cdot \vec{y}$$

Covariance matrix of estimators:

$$U_{ij} = \text{cov}\left(\alpha_i, \alpha_j\right) = \left(A^T V^{-1} A\right)^{-1}$$

Nice in principle, but requires lots of matrix inversions---rather nasty numerically. Might be simpler to just minimize χ^2!

χ^2 values used as a goodness of fit

The dominant use of the χ^2 statistics is for least squares fitting.

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

The "best fit" values of the parameters α are those that minimize the χ^2.

If there are *m* free parameters, and the deviation of the measured points from the model follows Gaussian distributions, then this statistic often follows a particular functional form called a χ^2 distribution with N-*m* degrees of freedom.

χ^2 is thus used to test the goodness of the fit. A good fit will have χ^2 about equal to N-m.

The χ^2 distribution

Suppose that you generate N random numbers from a Gaussian (normal) distribution with μ=0, σ=1: Z_1 ... Z_N.

Let *X* be the sum of the squared variables:

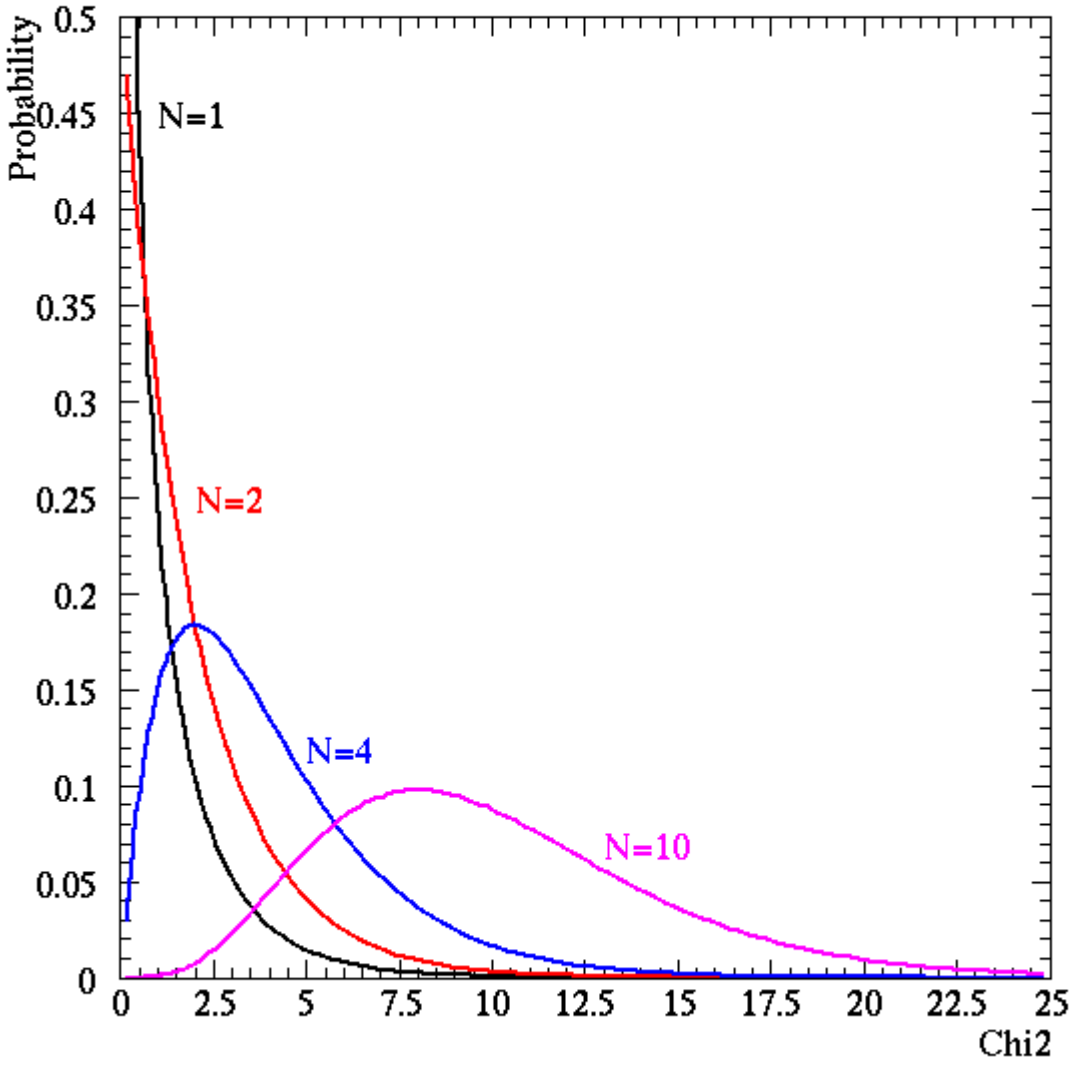
$$X = \sum_{i=1}^{N} Z_i^2$$

The variable *X* follows a χ^2 distribution with N degrees of freedom:

$$P\langle \chi^2 | N \rangle = \frac{2^{-N/2}}{\Gamma(N/2)} \left(\chi^2\right)^{(N-2)/2} e^{-\chi^2/2}$$

Recall that $\Gamma(N) = (N-1)!$ if *N* is an integer.

Properties of the χ^2 distribution

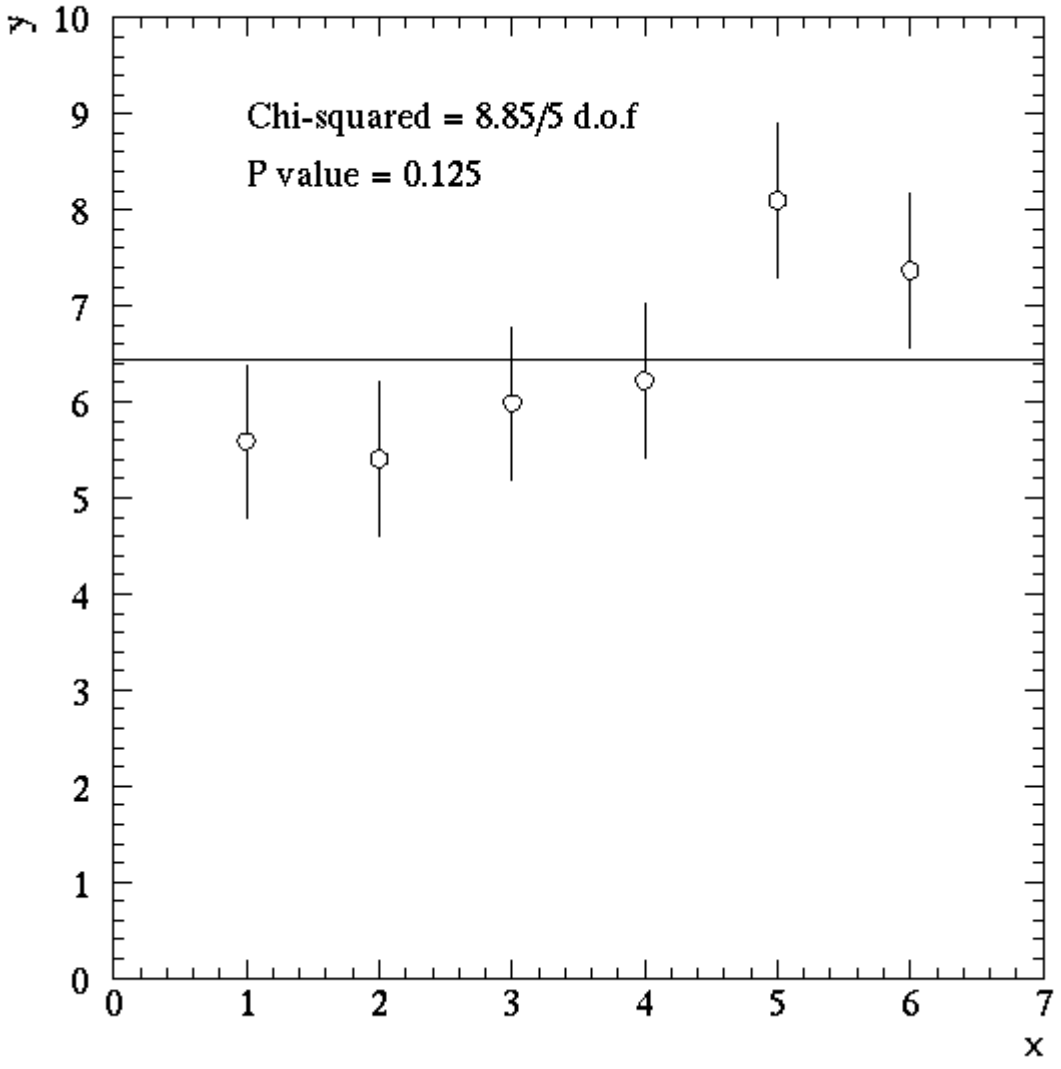


A χ^2 distribution has mean=N, but variance=2N.

In the context of χ^2 fits, this means you expect to get a χ^2 value about equal to the number of degrees of freedom in the problem.

For example, for 10 degrees of freedom, very small chance of getting χ^2 as large as 20 (bad fit) or as low as 3 (fit is too good!)

Goodness of fit: an example



Does the data sample, known to have Gaussian errors, fit acceptably to a constant (flat line)?

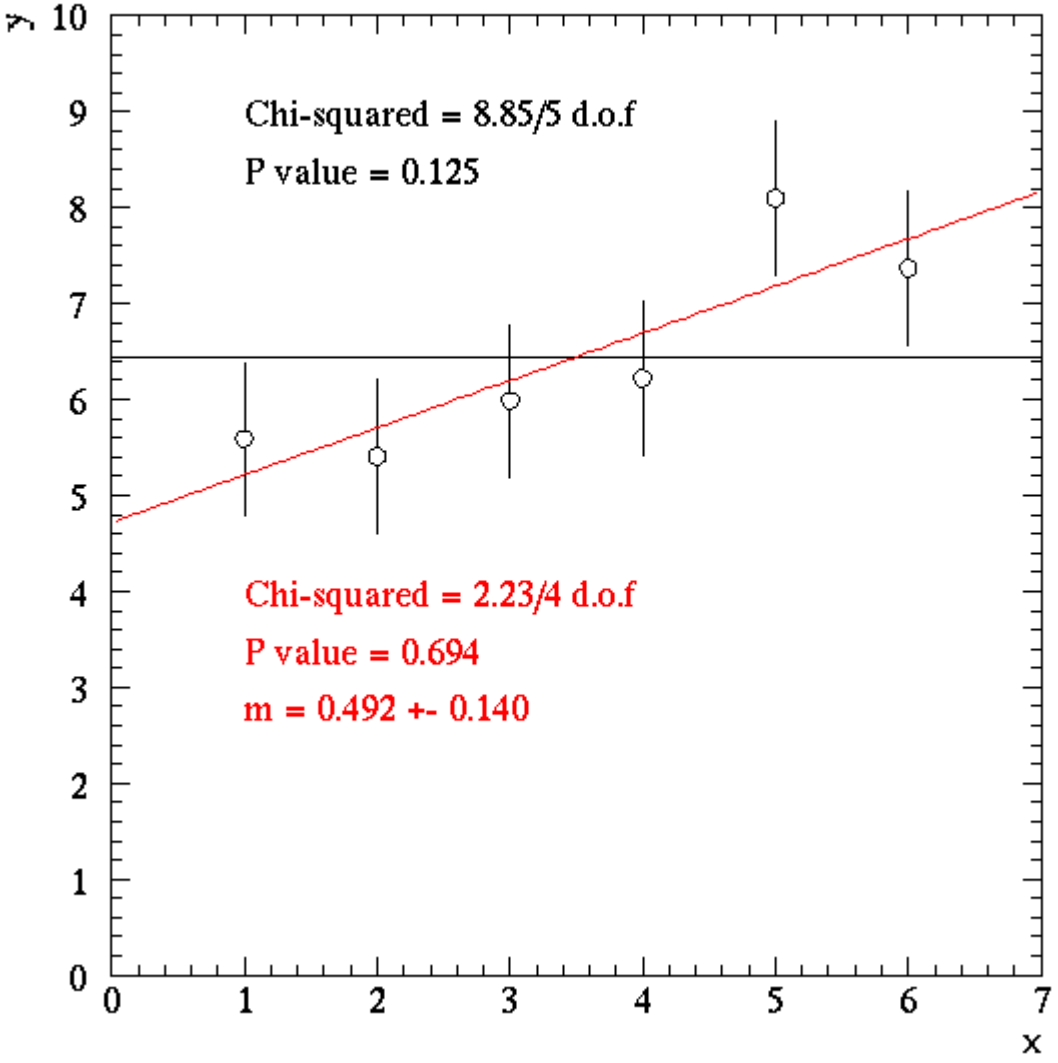
6 data points – 1 free parameter = 5 d.o.f.

χ^2 = 8.85/5 d.o.f.

Chance of getting a larger χ^2 is 12.5% for 5 degrees of freedom: an acceptable fit by almost anyone's standard.

Flat line is a good fit.

Distinction between goodness of fit and parameter estimation



Now if we fit a sloped line to the same data, is the slope consistent with flat?

χ^2 is obviously going to be somewhat better.

But slope is 3.5σ different from zero! Chance probability of this is 0.0002.

How can we simultaneously say that the same data set is “acceptably fit by a flat line” and “has a slope that is significantly larger than zero”???

Distinction between goodness of fit and parameter estimation

Goodness of fit and parameter estimation are answering two different questions.

1) Goodness of fit: is the data consistent with having been drawn from a specified distribution?

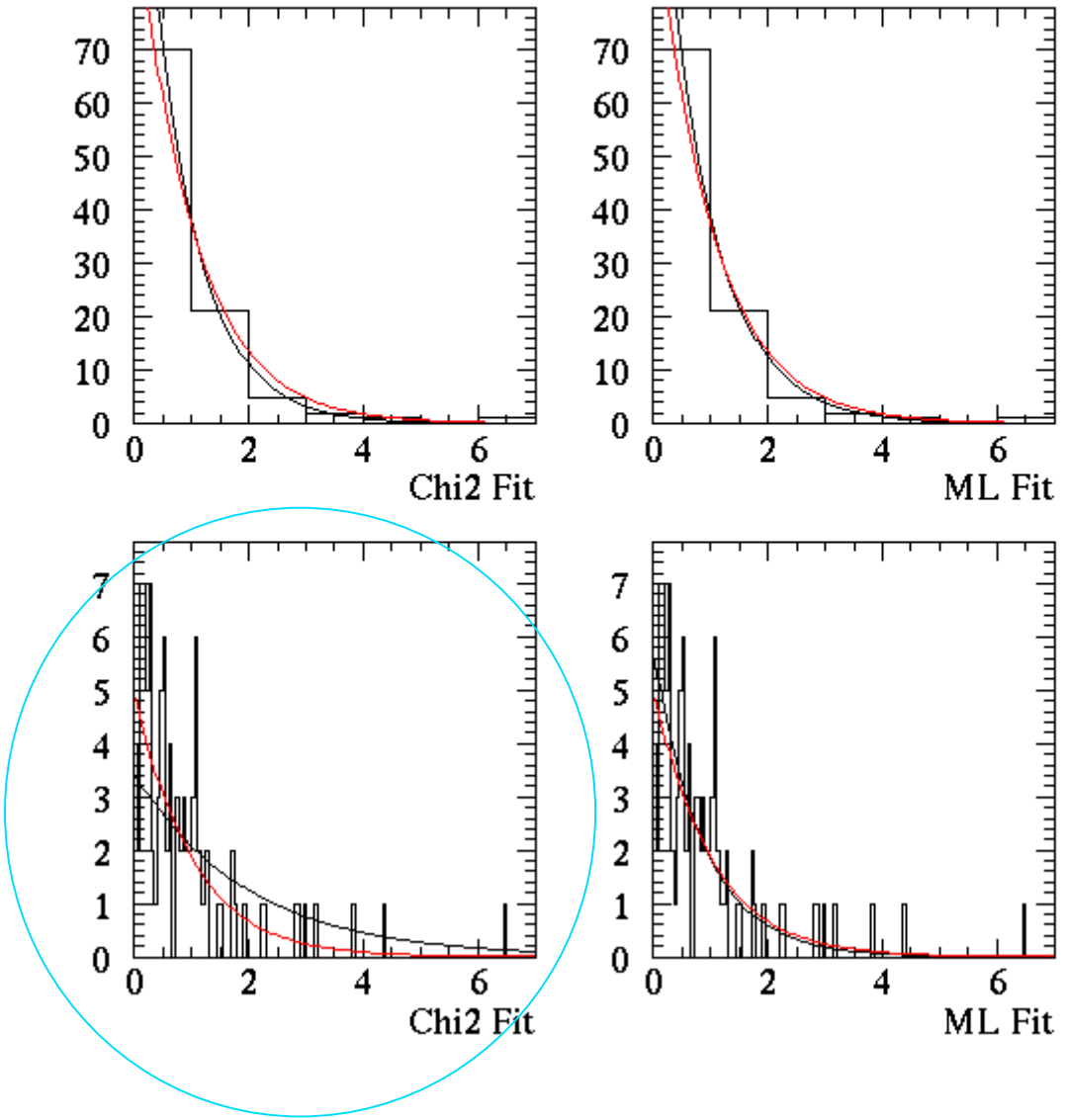
2) Parameter estimation: which of the following limited set of hypotheses is most consistent with the data?

One way to think of this is that a χ^2 goodness of fit compares the data set to all the possible ways that random Gaussian data might fluctuate. Parameter estimation chooses the best of a more limited set of hypotheses.

Parameter estimation is generally more powerful, at the expense of being more model-dependent.

Complaint of the statistically illiterate: “Although you say your data strongly favours solution A, doesn't solution B also have an acceptable χ^2/dof close to 1?”

Limitations of χ^2 fits



χ^2 /least squared fits are based on the assumption of Gaussian errors.

Beware of using these in cases where this doesn't apply.

To the left, the black line is the fit while the red is the true parent distribution.

Don’t use least squares fits with binned data, when bins have few events.

Goodness of fit: ML method

Sadly, the ML method does not yield a useful goodness of fit parameter. This is perhaps surprising, and is not commonly appreciated.

First of all, the quantity that plays the role of the χ^2 in the minimization, -ln(L), doesn't follow a standard distribution.

One sometimes recommended approach is to generate many simulated data sets with the same number of data points as your real data, and to fit them all with ML. Then make a histogram of the resulting minimum values of -ln(L) from all of the fits. Interpret this as a PDF for -ln(L) and see where the -ln(L) value for your data lies. If it lies in the meat of the distribution, it's a good fit. If it's way out on the tail, you can report that the fit is poor, and that the probability of getting a larger value of -ln(L) than that seen from your data is tiny.

This is a necessary condition to conclude that your model is a good fit to your data, but it is not sufficient ...

When to use Least Squares vs. Maximum Likelihood

My general advice: use maximum likelihood whenever you can. To use it, you must know how to calculate the PDFs of the measurements. But always remember that the ML estimators are often biased (although bias is usually negligible if N is large).

Consider using least squares if:

- your problem is linear in all parameters, or
- the errors are known to be Gaussian, or else you don't know the form of the measurement PDFs but only know the covariances, or
- for computational reasons, you need to use a simplified likelihood that may have a closed form solution

In general, the ML method has more general applicability, and makes use of more of the available information.

And avoid fitting histograms with LS whenever possible.

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 uncertainty that's not a statistical uncertainty”.

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)

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 $P(\theta)$, you observed some data D, and you need to calculate a likelihood $P(D|\theta)$. Let's suppose that the likelihood depends on some systematic parameter α (which could for example be the calibration of your thermometer). 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 joint distribution for θ, whose value we care about, and for α, which may be uninteresting. We marginalize by integrating over α to get $P(\theta)$.

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

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

Error-weighted averages

Suppose you have N independent measurements of a quantity. You average them. The proper error-weighted average, and its variance, are:

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

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

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:

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

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:

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

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 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:

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

Averaging correlated measurements III

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

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

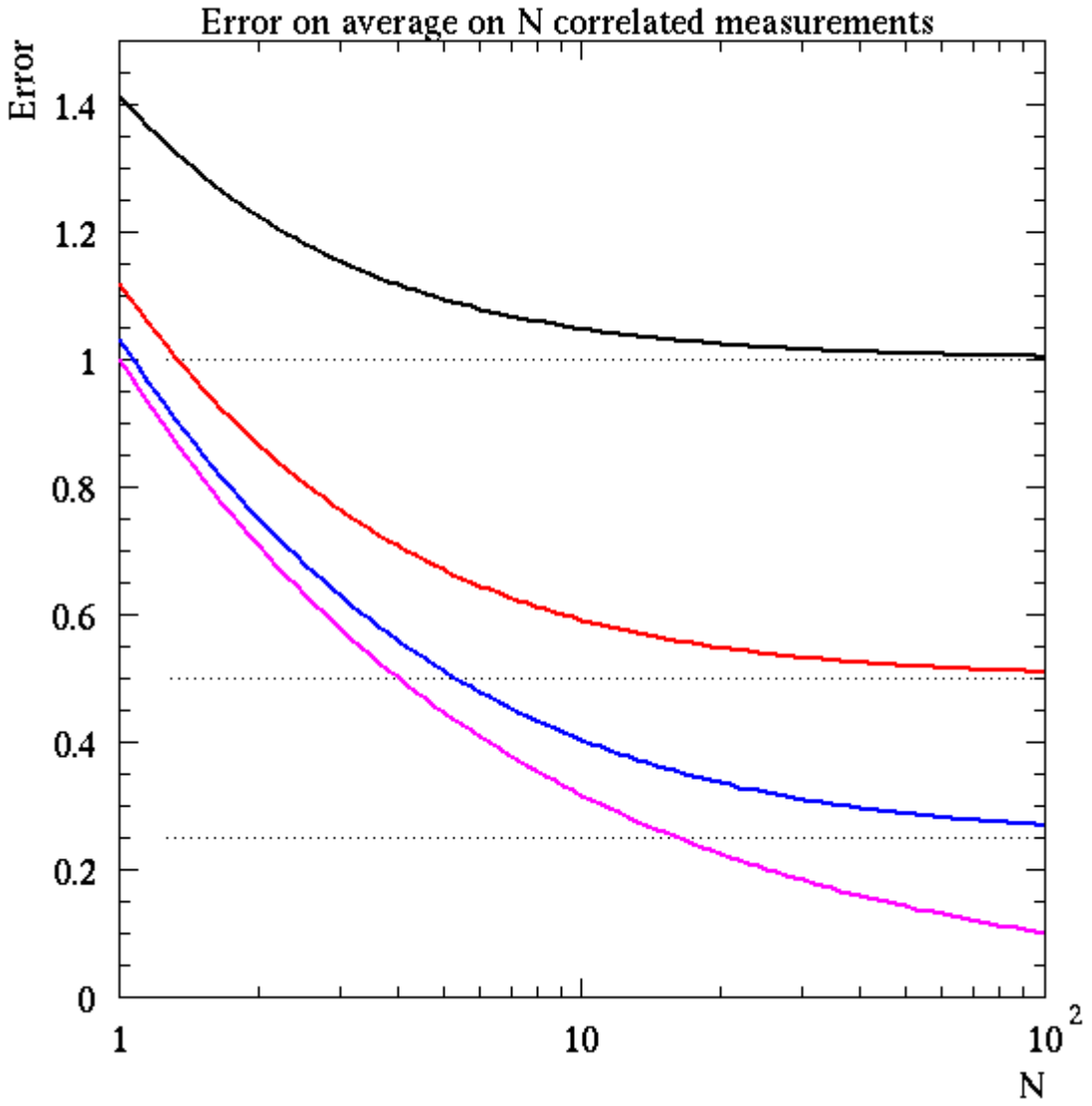
we can in fact read off the coefficient of μ^2, which will be $1/V(\mu)$:

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

For a systematic that affects all data points equally, V is:

$$V = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} + \begin{bmatrix} \sigma_{sys}^2 & \sigma_{sys}^2 & \cdots & \sigma_{sys}^2 \\ \sigma_{sys}^2 & \sigma_{sys}^2 & & \sigma_{sys}^2 \\ \vdots & & & \vdots \\ \sigma_{sys}^2 & \sigma_{sys}^2 & \cdots & \sigma_{sys}^2 \end{bmatrix}$$

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 σ_b.

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

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!

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:

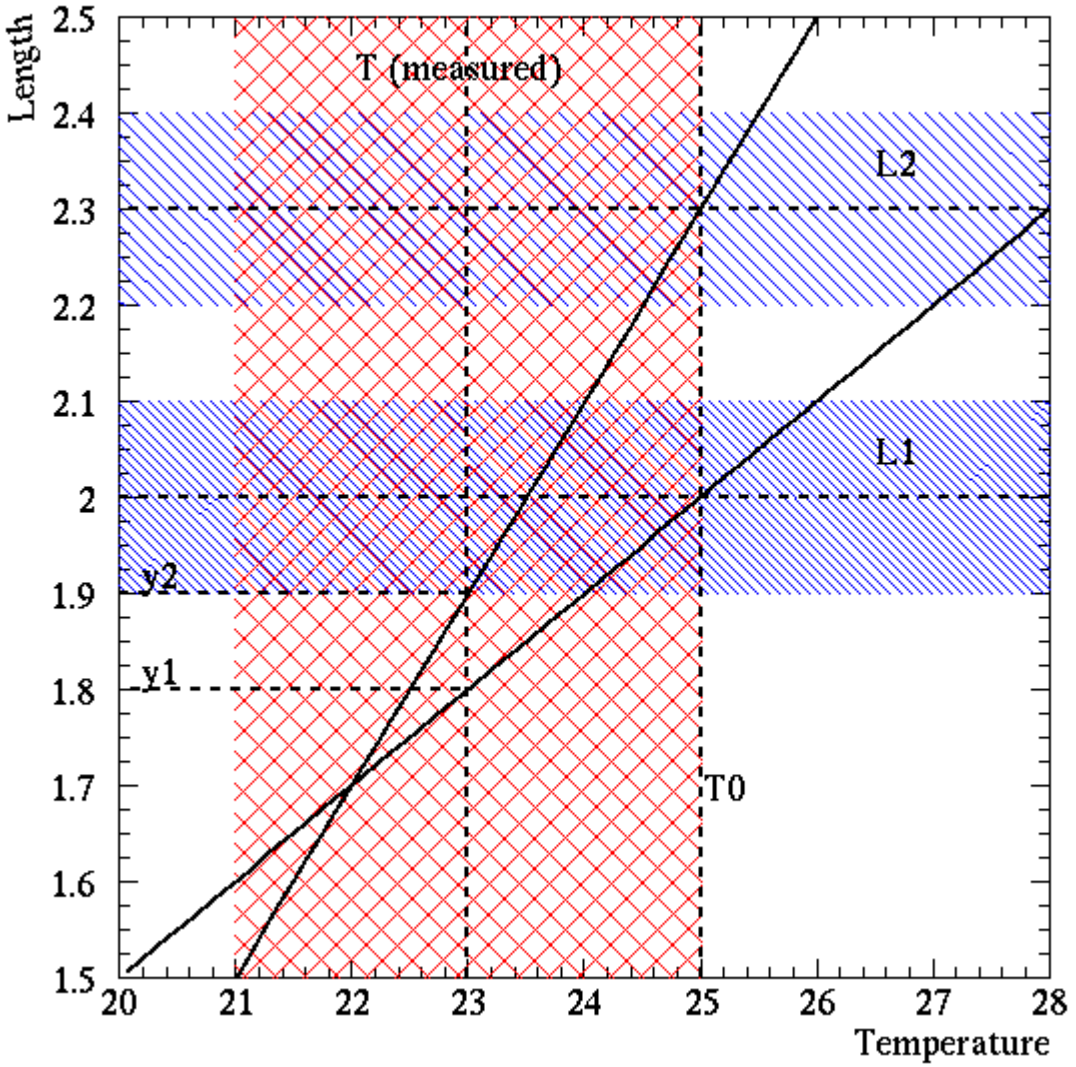
c_1 = 0.1	L_1=2.0 ± 0.1	y_1=1.80 ± 0.22	T_0=25
c_2 = 0.2	L_2=2.3 ± 0.1	y_2=1.90 ± 0.41	T = 23 ± 2

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??

Averaging correlated measurements: example



Because y_1 and y_2 disagree, fit attempts to adjust temperature to make them agree. This pushes fitted length lower.

This is one of many cases in which the data itself gives you additional constraints on the value of a systematic (here, what is the true T).

Constraint terms in the likelihood

Working in Bayesian language, the posterior PDF is given by

$$P(\theta,\alpha|D) \propto P(\theta)P(\alpha)P(D|\theta,\alpha)$$

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:

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

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

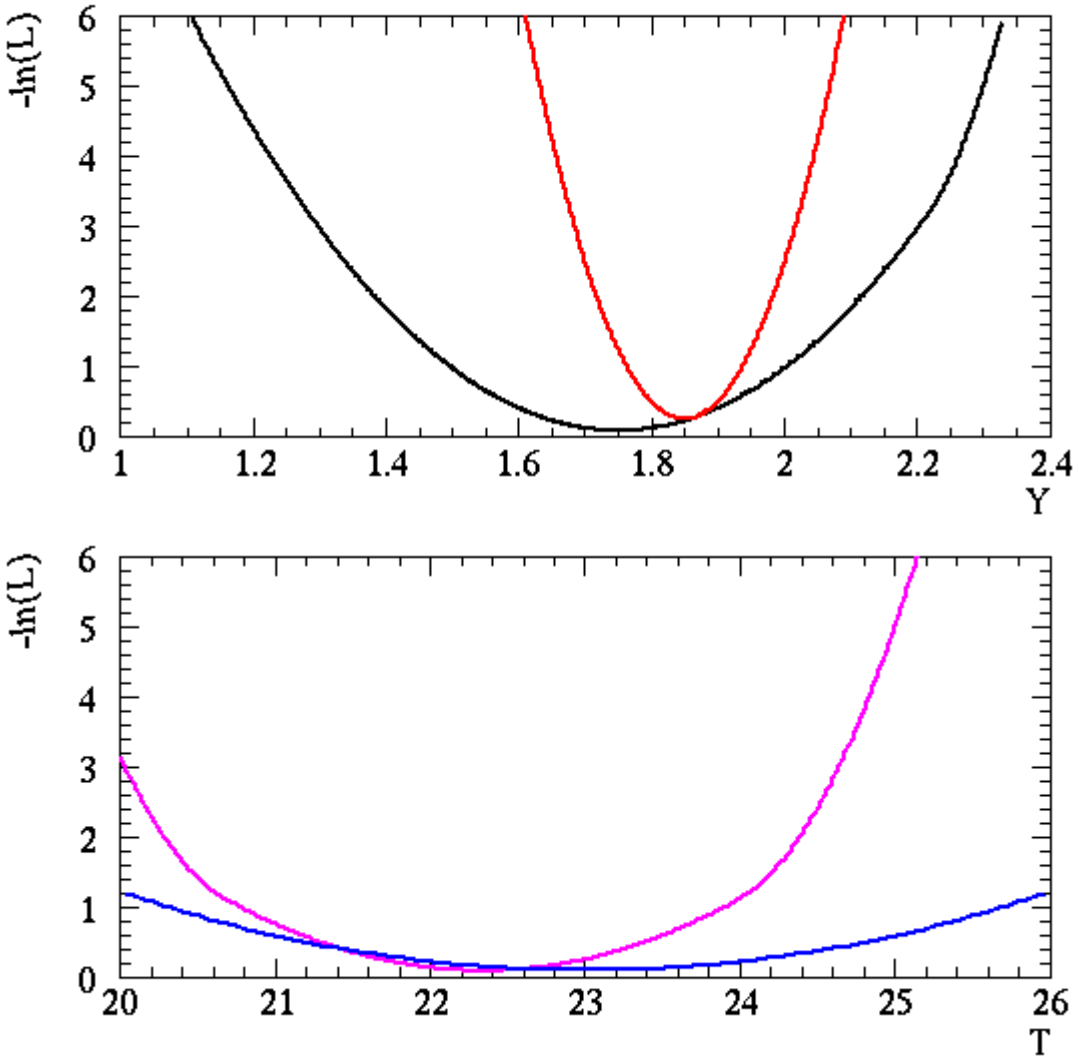
$$\ln L\left(\theta,\alpha\right) = \ln L\left(\theta \mid D,\alpha\right) + \ln P\left(\alpha\right)$$

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

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.

Marginalization 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.

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.

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 (good fitting software will do this for you):

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

For More Information

Notes from my graduate level data analysis course:

http://www.phas.ubc.ca/~oser/p509/

Bayes' Theorem

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

This just follows from laws of conditional probability---even frequentists agree, but they give it a different interpretation.

H = a hypothesis (e.g. “SUSY exists at the TeV scale”)
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 D would have happened anyway, whether or not H is true.

Note: you can only calculate P(D|I) if you have a “hypothesis space” you're comparing to. A hypothesis is only “true” relative to some set of alternatives.

Example: Triple Screen Test

The incidence of Down's syndrome is 1 in 1000 births. A triple screen test is a test performed on the mother's blood during pregnancy to diagnose Down's. The manufacturer of the test claims an 85% detection rate and a 1% false positive rate.

You (or your partner) test positive. What are the chances that your child actually has Down's?

Discussion: Triple Screen Test

The incidence of Down's syndrome is 1 in 1000 births. A triple screen test is a test performed on the mother's blood during pregnancy to diagnose Down's. The manufacturer of the test claims an 85% detection rate and a 1% false positive rate.

You (or your partner) test positive. What are the chances that your child actually has Down's?

Consider 100,000 mothers being tested. Of these, 100,000/1000=100 actually carry a Down's child, while 99,900 don't. For these groups:

85 are correctly diagnosed with Down's.
15 are missed by the test
999 are incorrectly diagnosed with Down's
98901 are correctly declared to be free of Down's

Fraction of fetuses testing positive who really have the disorder:
85/(85+999) = 7.8%

Bayes' Theorem applied to Down's syndrome screening

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

$$P(H|D,I) = \frac{0.001 \times 0.85}{0.001 \times 0.85 + 0.999 \times 0.01}$$

$$P(H|D,I) = 0.078$$

Hypothesis H: fetus has Down's syndrome

Data D = a positive test result

P(H|I) = the "prior probability" for H = 0.001 (rate in general population)

P(D|H,I) = the probability of measuring D, given H and I. Also called the "likelihood". P(D|H,I)= 0.85 in this case

P(D|I) = a normalizing constant: the probability that D would have happened anyway, whether or not H is true.

= 0.001 x 0.85 + 0.999 x 0.01

= P(H) P(D|H) + P(~H) P(D|~H)

Goodness of fit for least squares

By now you're probably wondering why I haven't discussed the use of χ^2 as a goodness of fit parameter. Partly this is because parameter estimation and goodness of fit are logically separate things---if you're CERTAIN that you've got the correct model and error estimates, then a poor χ^2 can only be bad luck, and tells you nothing about how accurate your parameter estimates are.

Carefully distinguish between:

1) Value of χ^2 at minimum: a measure of goodness of fit

2) How quickly χ^2 changes as a function of the parameter: a measure of the uncertainty on the parameter.

Nonetheless, a major advantage of the χ^2 approach is that it does automatically generate a goodness of fit parameter as a byproduct of the fit. As we'll see, the maximum likelihood method doesn't.

How does this work?

χ^2 as a goodness of fit parameter

Remember that the sum of N Gaussian variables with zero mean and unit RMS, when squared and added, follows a χ^2 distribution with N degrees of freedom. Compare to the least squares formula:

$$\chi^2 = \sum_i \sum_j \left(y_i - f\left(x_i|\alpha\right)\right)\left(y_j - f\left(x_j|\alpha\right)\right)\left(V^{-1}\right)_{ij}$$

If each y_i is distributed around the function according to a Gaussian, **and** $f(x|\alpha)$ is a linear function of the m free parameters α, **and** the error estimates don't depend on the free parameters, then the best-fit least squares quantity we call χ^2 actually follows a χ^2 distribution with N-m degrees of freedom.

People usually ignore these various caveats and assume this works even when the parameter dependence is non-linear and the errors aren't Gaussian. Be very careful with this, and check with simulation if you're not sure.

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:

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

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

Systematic uncertainties

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

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.

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

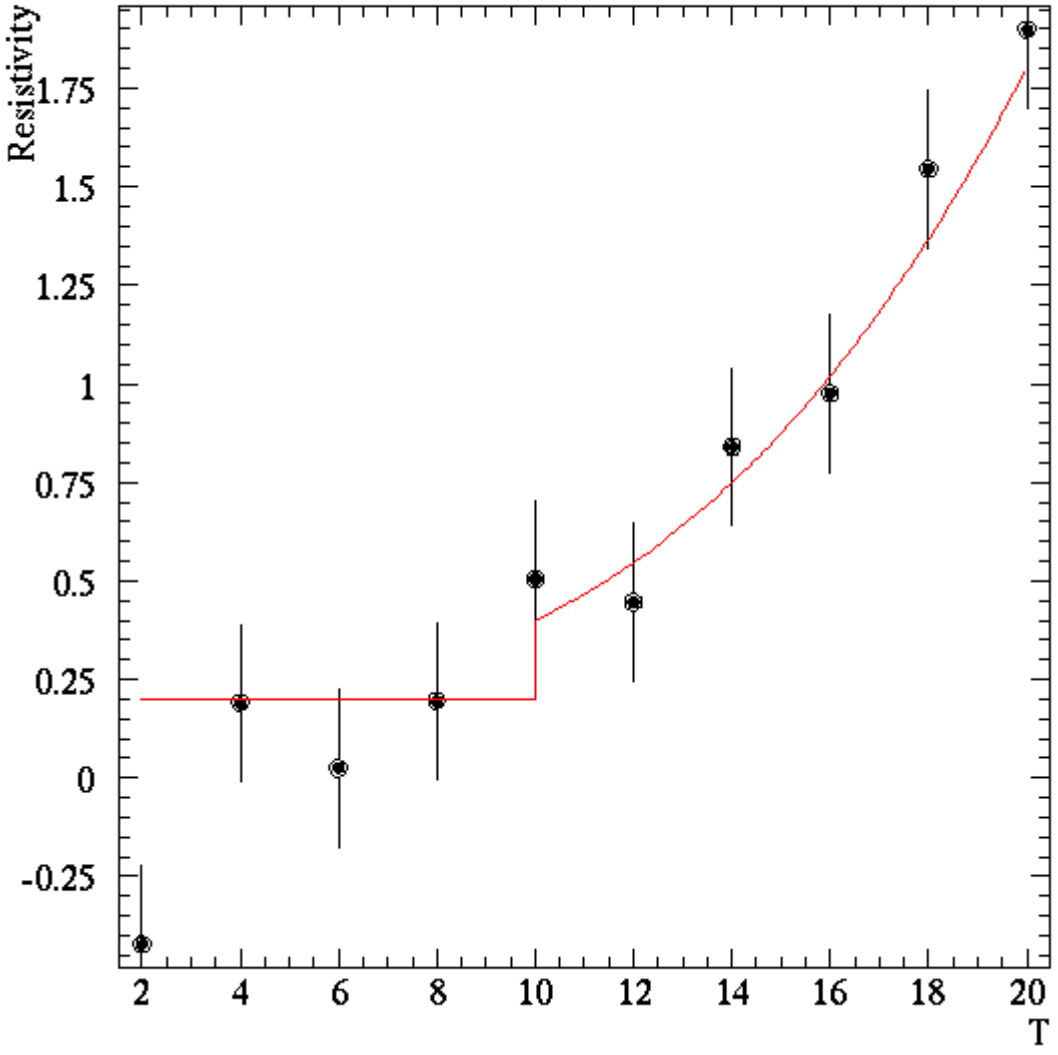
Advantages of the Monte Carlo method:

- 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.)

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)

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

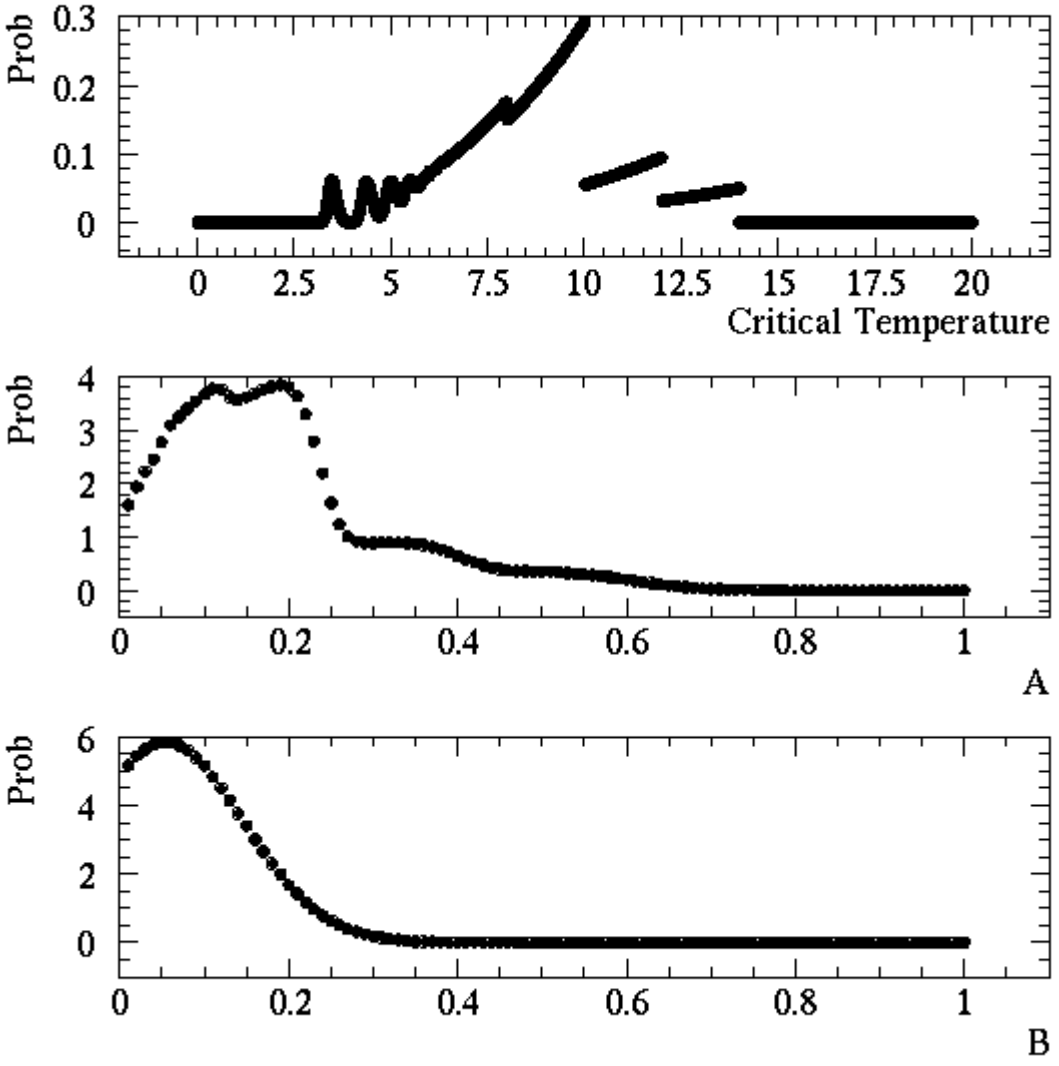
where R(T_i) is the piecewise-defined function given previously. All the dependence on model parameters is contained in R(T).

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

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

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.

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:

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

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

Covariance matrix approach

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:

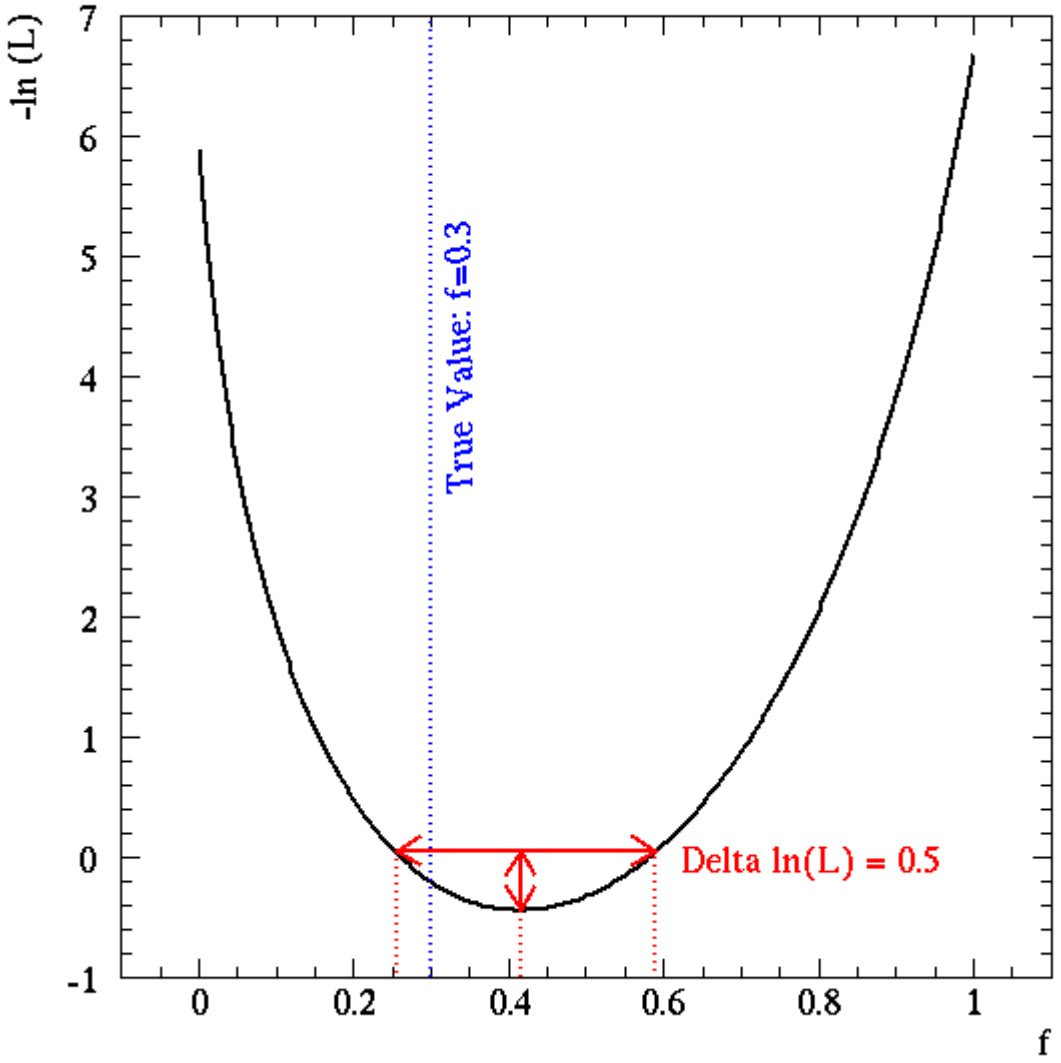
$$V_{tot} = V_{stat} + V_{sys}$$

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.

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

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).

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.

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).

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.