Covariance & Correlation

The covariance between two variables is defined by:

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

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:

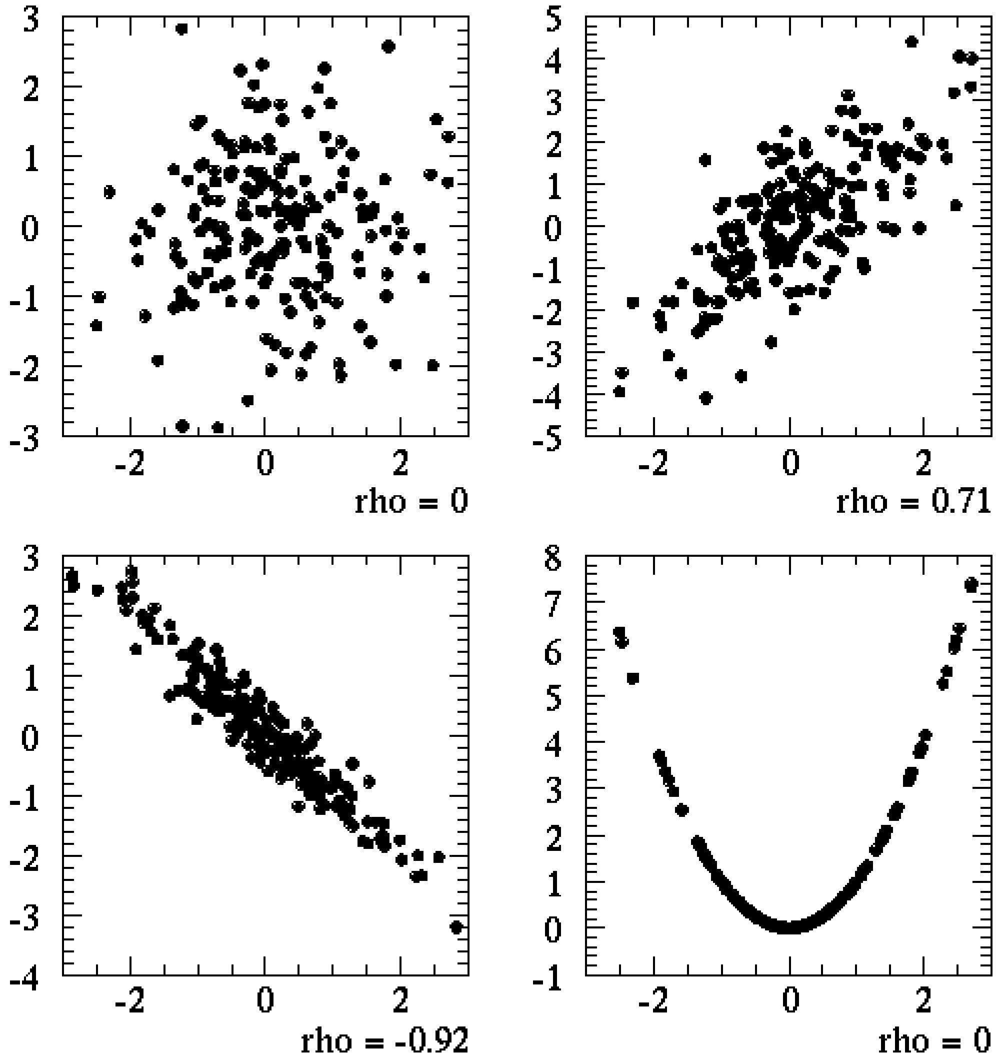
$$\rho=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$$

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

$$\operatorname{cov}(x,y)=\int dx\,dy\,P(x,y)\,xy\;-\;(\int dx\,dy\,P(x,y)\,x)(\int dx\,dy\,P(x,y)\,y)$$

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

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.

Variance and Covariance of Linear Combinations of Variables

Suppose we have two random variable X and Y (not necessarily independent), and that we know cov(X,Y).

Consider the linear combinations W=aX+bY and Z=cX+dY. It can be shown that

$$
\begin{aligned}
\mathrm{cov}(W,Z) &= \mathrm{cov}(aX+bY,cX+dY) \\
&= \mathrm{cov}(aX,cX) + \mathrm{cov}(aX,dY) + \mathrm{cov}(bY,cX) + \mathrm{cov}(bY,dY) \\
&= ac\ \mathrm{cov}(X,X) + (ad + bc)\ \mathrm{cov}(X,Y) + bd\ \mathrm{cov}(Y,Y) \\
&= ac\ V(X) + bd\ V(Y) + (ad+bc)\ \mathrm{cov}(X,Y)
\end{aligned}
$$

Special case is V(X+Y):

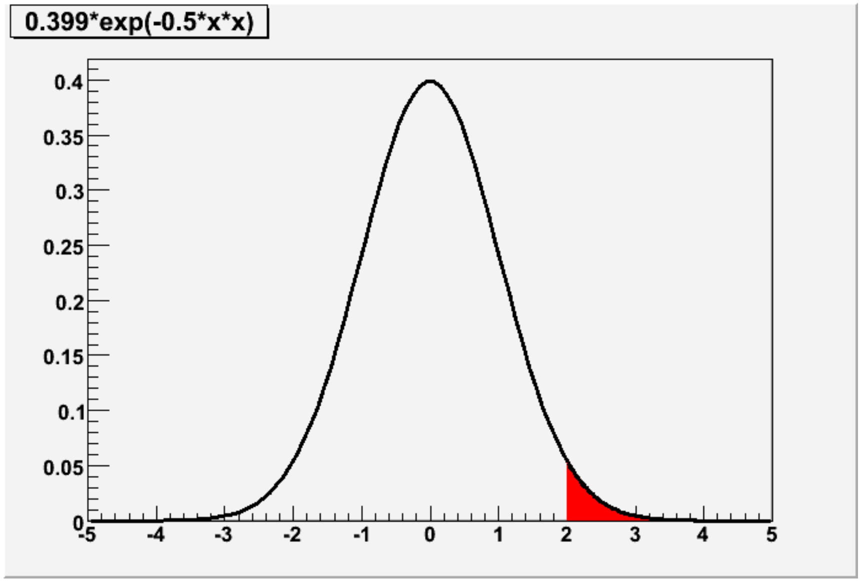
$$V(X+Y) = \mathrm{cov}(X+Y,X+Y) = V(X) + V(Y) + 2\mathrm{cov}(X,Y)$$

Very special case: variance of the sum of independent random variables is the sum of their individual variances!

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 ±1σ
95.45% of area within ±2σ
99.73% of area within ±3σ

90% of area within ±1.645σ
95% of area within ±1.960σ
99% of area within ±2.576σ

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!

Why you should be very careful with Gaussians ..

The major danger of Gaussians is that they are overused. Although many distributions are approximately Gaussian, they often have long non-Gaussian tails.

While 99% of the time a Gaussian distribution will correctly model your data, many foul-ups result from that other 1%.

It's usually good practice to simulate your data to see if the distributions of quantities you think are Gaussian really follow a Gaussian distribution.

Common example: the ratio of two numbers with Gaussian distributions is itself often not very Gaussian (although in certain limits it may be).

Review of covariances of joint PDFs

Consider some multidimensional PDF *p(x₁ ... xₙ)*. We define the covariance between any two variables by:

$$\mathrm{cov}(x_i, x_j) = \int d\vec{x}\, p(\vec{x})\ \ (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle)$$

The set of all possible covariances defines a covariance matrix, often denoted by V_{ij}. The diagonal elements of V_{ij} are the variances of the individual variables, while the off-diagonal elements are related to the correlation coefficients:

$$V_{ij} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \ldots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_1\sigma_n & \sigma_2^2 & \ldots & \rho_{2\mathrm{n}}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_1\sigma_n & \rho_{n2}\sigma_2\sigma_n & \ldots & \sigma_n^2 \end{bmatrix}$$

Properties of covariance matrices

Covariance matrices always:

- are symmetric and square
- are invertible (very important requirement!)

The most common use of a covariance matrix is to invert it then use it to calculate a χ^2:

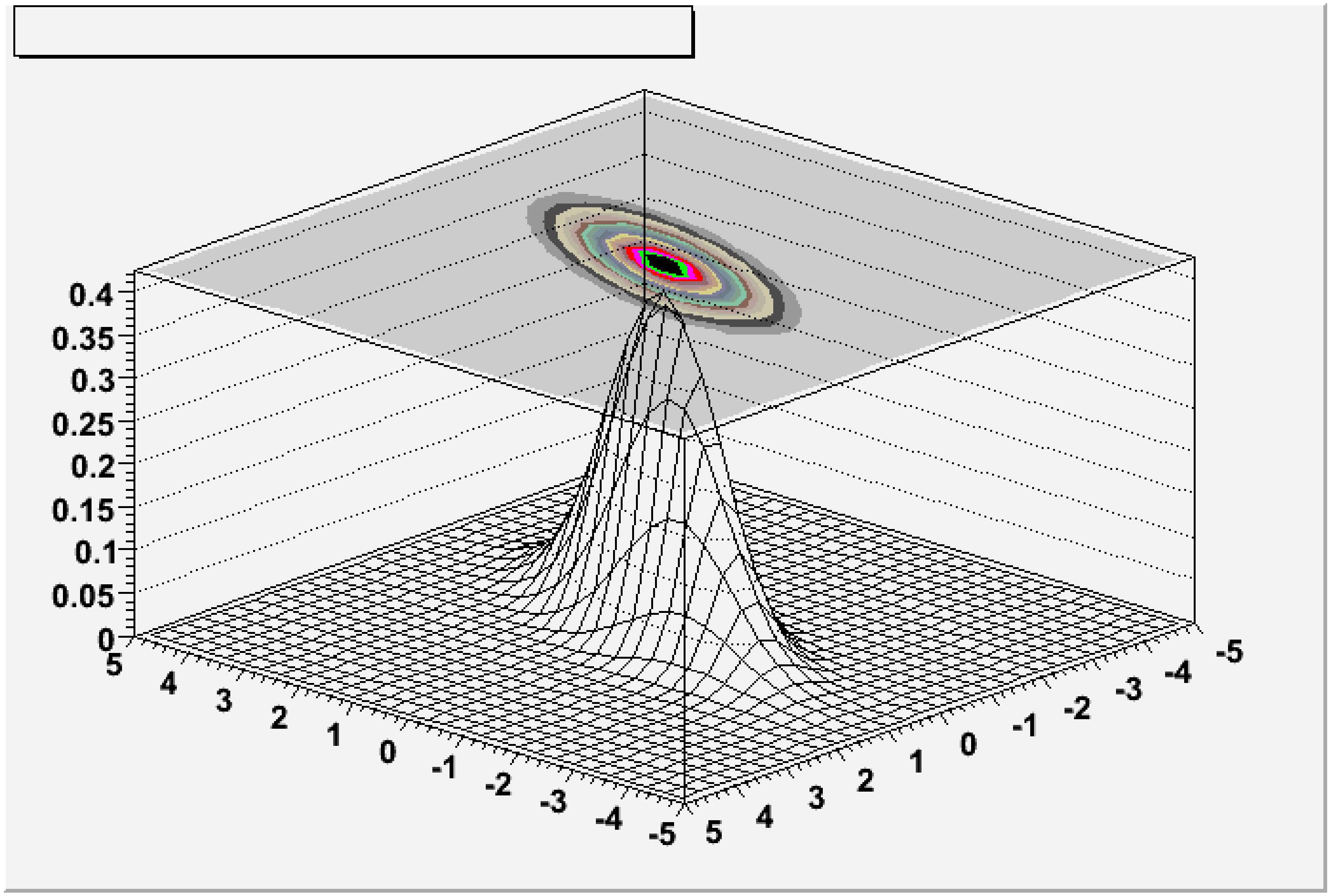
$$\chi^2 = \sum_i \sum_j (y_i - f(x_i)) V_{ij}^{-1} (y_j - f(x_j))$$

If the covariances are zero, then $V_{ij} = \delta_{ij} \sigma_i^2$, and this reduces to:

$$\chi^2 = \sum_i \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

Warning: do NOT use the simplified formula if data points are correlated!

Approximating the peak of a PDF with a multidimensional Gaussian



Suppose we have some complicated-looking PDF in 2D that has a well-defined peak.

How might we approximate the shape of this PDF around its maximum?

Taylor Series expansion

Consider a Taylor series expansion of the logarithm of the PDF around its maximum at (x_0,y_0):

$$\log P(x,y) = P_0 + A(x-x_0) + B(y-y_0) - C(x-x_0)^2 - D(y-y_0)^2 - 2E(x-x_0)(y-y_0)\ldots$$

Since we are expanding around the peak, then the first derivatives must equal zero, so A=B=0. The remaining terms can be written in matrix form:

$$\log P(x,y) \approx P_0 - (\Delta x, \Delta y)\begin{pmatrix} C & E \\ E & D \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

In order for (x_0,y_0) to be a maximum of the PDF (and not a minimum or saddle point), the above matrix must be positive definite, and therefore invertible.

Taylor Series expansion

$$\log P(x,y) \approx P_0 - (\Delta x, \Delta y) \begin{pmatrix} C & E \\ E & D \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

Let me now suggestively denote the inverse of the above matrix by V_{ij}. It's a positive definite matrix with three parameters. In fact, I might as well call these parameters σ_x, σ_y, and ρ.

Exponentiating, we see that around its peak the PDF can be approximated by a multidimensional Gaussian. The full formula, including normalization, is

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-x_0}{\sigma_x}\right)^2 + \left(\frac{y-y_0}{\sigma_y}\right)^2 - 2\rho \left(\frac{x-x_0}{\sigma_x}\right)\left(\frac{y-y_0}{\sigma_y}\right) \right] \right\}$$

This is a good approximation as long as higher order terms in Taylor series are small.

Interpretation of multidimensional Gaussian

$$P(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-x_0}{\sigma_x}\right)^2+\left(\frac{y-y_0}{\sigma_y}\right)^2-2\rho\left(\frac{x-x_0}{\sigma_x}\right)\left(\frac{y-y_0}{\sigma_y}\right)\right]\right\}$$

Can I directly relate the free parameters to the covariance matrix? First calculate *P(x)* by marginalizing over *y:*

$$P(x)\propto\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}\int dy\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-y_0}{\sigma_y}\right)^2-2\rho\left(\frac{x-x_0}{\sigma_x}\right)\left(\frac{y-y_0}{\sigma_y}\right)\right]\right\}$$

$$P(x)\propto\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}\int dy\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-y_0}{\sigma_y}\right)^2-2\rho\left(\frac{x-x_0}{\sigma_x}\right)\left(\frac{y-y_0}{\sigma_y}\right)+\rho^2\left(\frac{x-x_0}{\sigma_x}\right)^2-\rho^2\left(\frac{x-x_0}{\sigma_x}\right)^2\right]\right\}$$

$$P(x)\propto\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}\int dy\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-y_0}{\sigma_y}-\rho\left(\frac{x-x_0}{\sigma_x}\right)\right)^2-\rho^2\left(\frac{x-x_0}{\sigma_x}\right)^2\right]\right\}$$

$$P(x)\propto\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}\exp\left\{+\frac{\rho^2}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}=\exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}$$

So we get a Gaussian with width σ_x. Calculations of σ_y similar, and can also show that ρ is correlation coefficient.

P(x|y)

$$P(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-x_0}{\sigma_x}\right)^2+\left(\frac{y-y_0}{\sigma_y}\right)^2-2\rho\left(\frac{x-x_0}{\sigma_x}\right)\left(\frac{y-y_0}{\sigma_y}\right)\right]\right\}$$

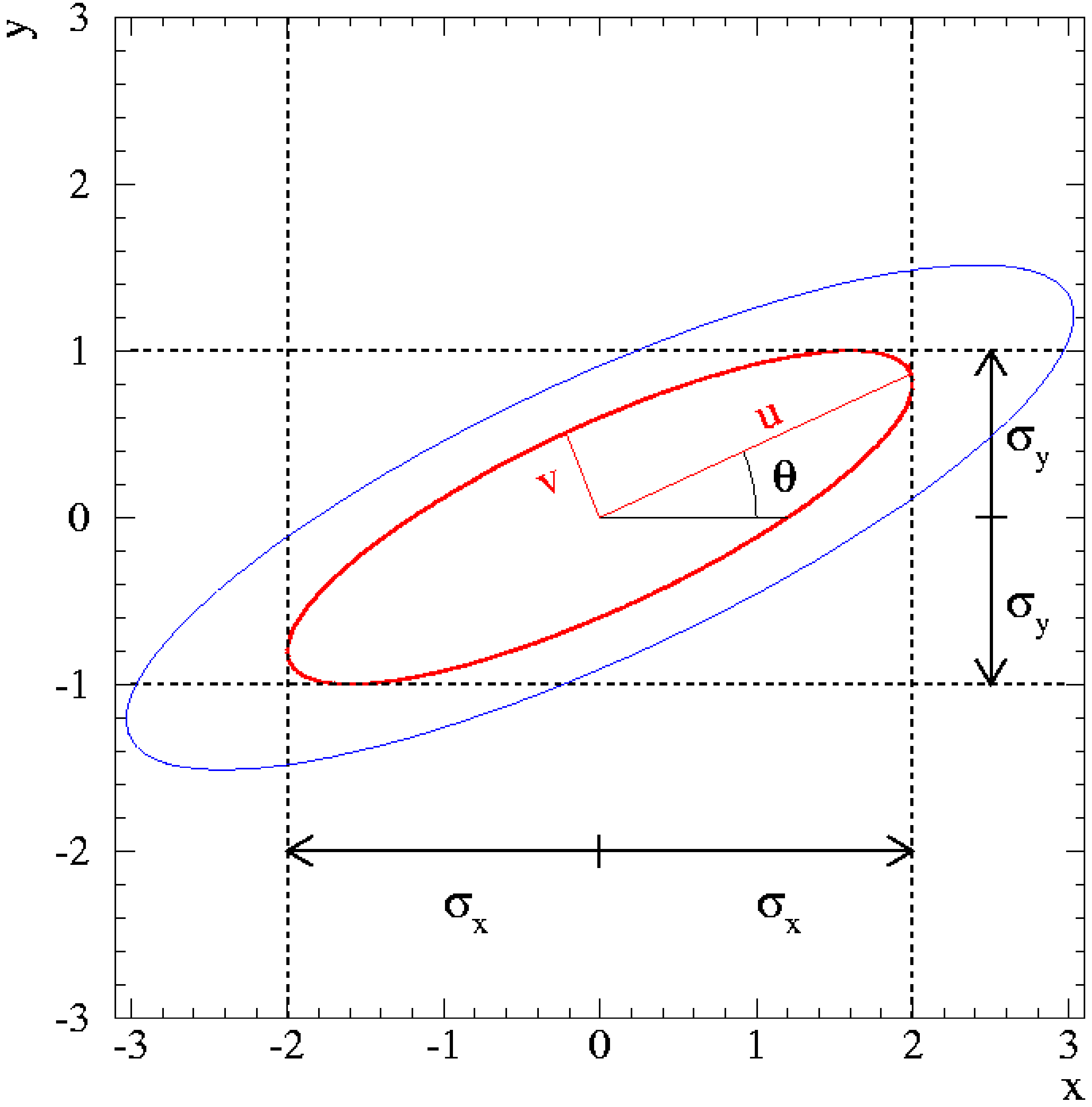
Note: if you view y as a fixed parameter, then the PDF P(x|y) is a Gaussian with width of:

$$\sigma_x\sqrt{1-\rho^2}$$

and a mean value of

$$x_0+\rho\left(\frac{\sigma_x}{\sigma_y}\right)(y-y_0)$$

(It makes sense that the width of P(x|y) is always narrower than the width of the marginalized PDF P(x) (integrated over y). If you know the actual value of y, you have additional information and so a tighter constraint on x.



σ_x=2

σ_y=1

ρ=0.8

Red ellipse: contour with argument of exponential set to equal -1/2

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

Contour ellipses

$$P(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-x_0}{\sigma_x}\right)^2+\left(\frac{y-y_0}{\sigma_y}\right)^2-2\rho\left(\frac{x-x_0}{\sigma_x}\right)\left(\frac{y-y_0}{\sigma_y}\right)\right]\right\}$$

The contour ellipses are defined by setting the argument of the exponent equal to a constant. The exponent equals -1/2 on the red ellipse from the previous graph. Parameters of this ellipse are:

$$\tan 2\theta=\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2}$$

$$\sigma_u=\frac{\cos^2\theta\cdot\sigma_x^2-\sin^2\theta\cdot\sigma_y^2}{\cos^2\theta-\sin^2\theta}$$

$$\sigma_v=\frac{\cos^2\theta\cdot\sigma_y^2-\sin^2\theta\cdot\sigma_x^2}{\cos^2\theta-\sin^2\theta}$$

Probability content inside a contour ellipse

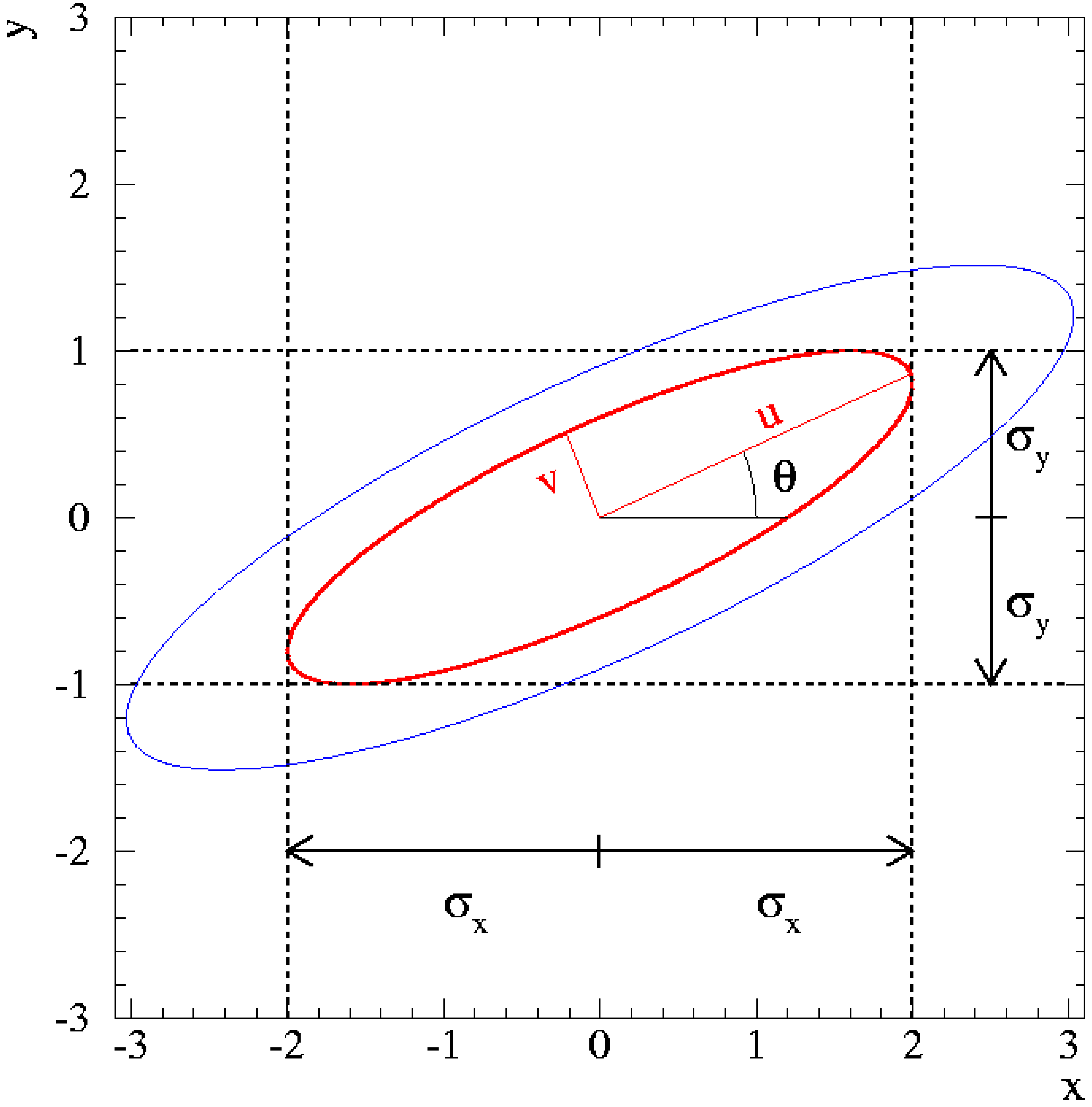
For a 1D Gaussian exp($-x^2/2\sigma^2$), the ±1σ limits occur when the argument of the exponent equals -1/2. For a Gaussian there's a 68% chance of the measurement falling within around the mean.

But for a 2D Gaussian this is not the case. Easiest to see this for the simple case of $\sigma_x=\sigma_y=1$:

$$\frac{1}{2\pi}\int dx\,dy\,\exp\left[-\frac{1}{2}(x^2+y^2)\right]=\int_0^{r_0} dr\,\exp\left[-\frac{1}{2}r^2\right]=0.68$$

Evaluating this integral and solving gives $r_0^2=2.3$. So 68% of probability content is contained within a radius of $\sigma\sqrt{(2.3)}$.

We call this the 2D contour. Note that it's bigger than the 1D version---if you pick points inside the 68% contour and plot their x coordinates, they'll span a wider range than those picked from the 68% contour of the 1D marginalized PDF!



σ_x=2

σ_y=1

ρ =0.8

Red ellipse: contour with argument of exponential set to equal -1/2

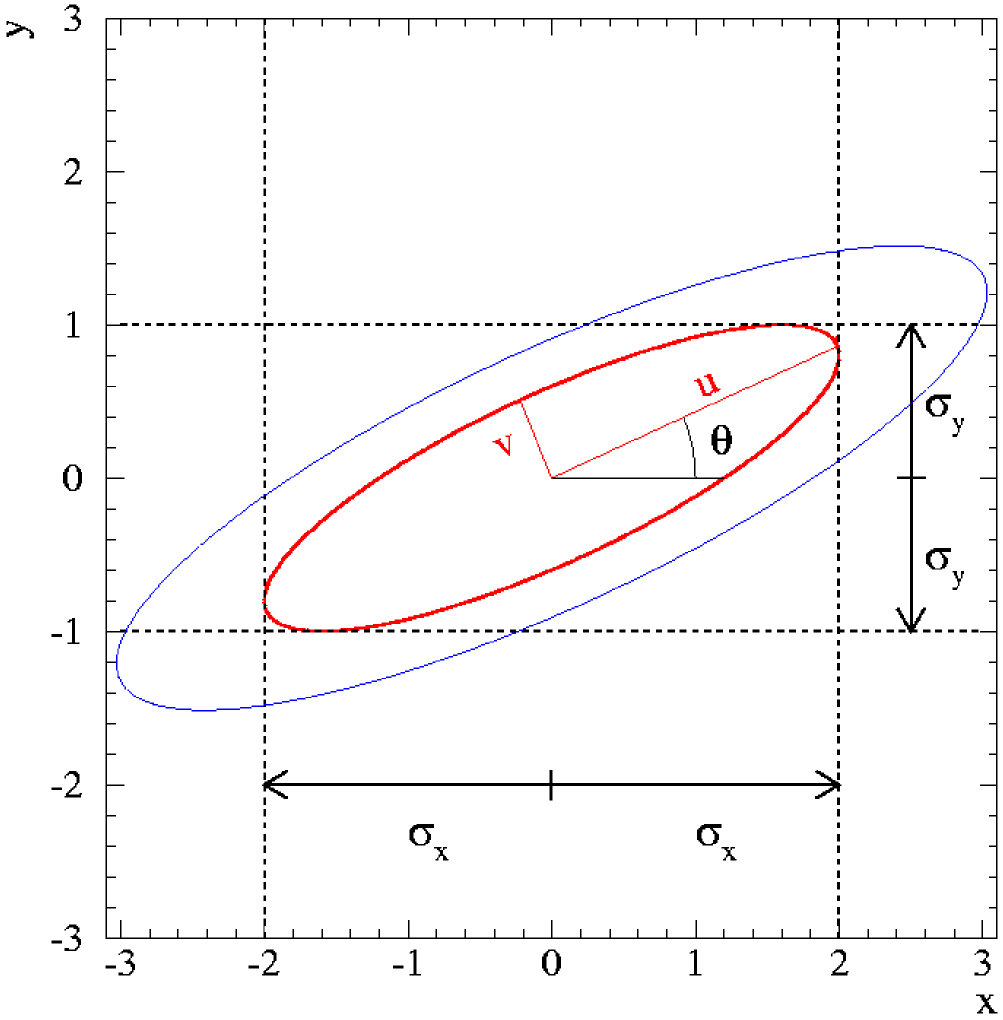
Blue ellipse: contour containing 68% of probability content.

Marginalization by minimization

Normal marginalization procedure: integrate over y.

For a multidimensional Gaussian, this gives the same answer as finding the extrema of the ellipse---for every x, find the the value of y that maximizes the likelihood.

For example, at x=±2 the value of y which maximizes the likelihood is just where the dashed line touches the ellipse. The value of the likelihood at that point then is the value P(x)



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.

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

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

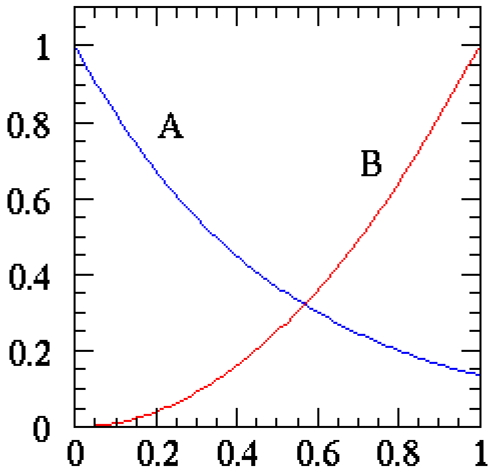
It's more common (and easier) to calculate -ln *L* instead:

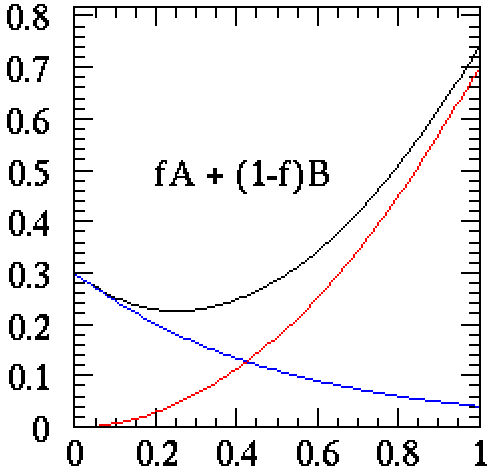
$$-\ln L(x_1 \ldots x_N|\vec{\alpha}) = -\sum_{i=1}^{N} \ln P(x_i|\vec{\alpha})$$

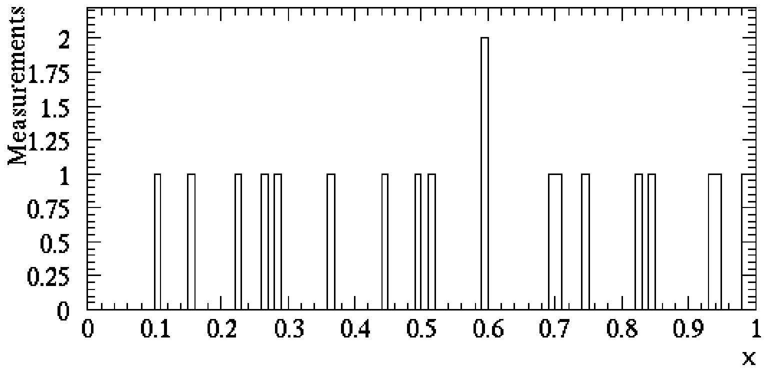
The maximum likelihood estimator is that value of α which maximizes L as a function of α. It can be found by minimizing -ln *L* over the unknown parameters.

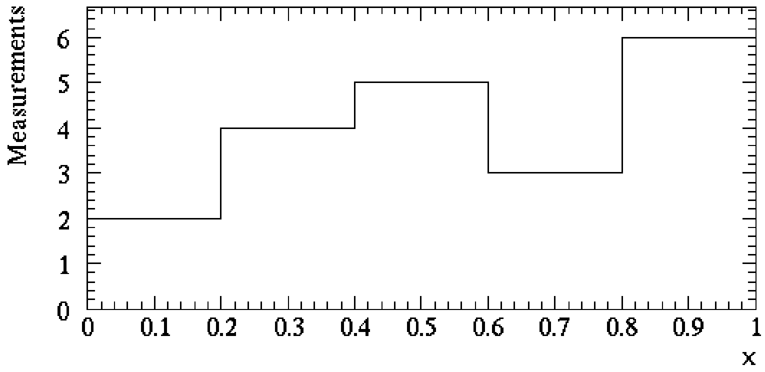
Simple example of an ML estimator

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.









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

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

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.

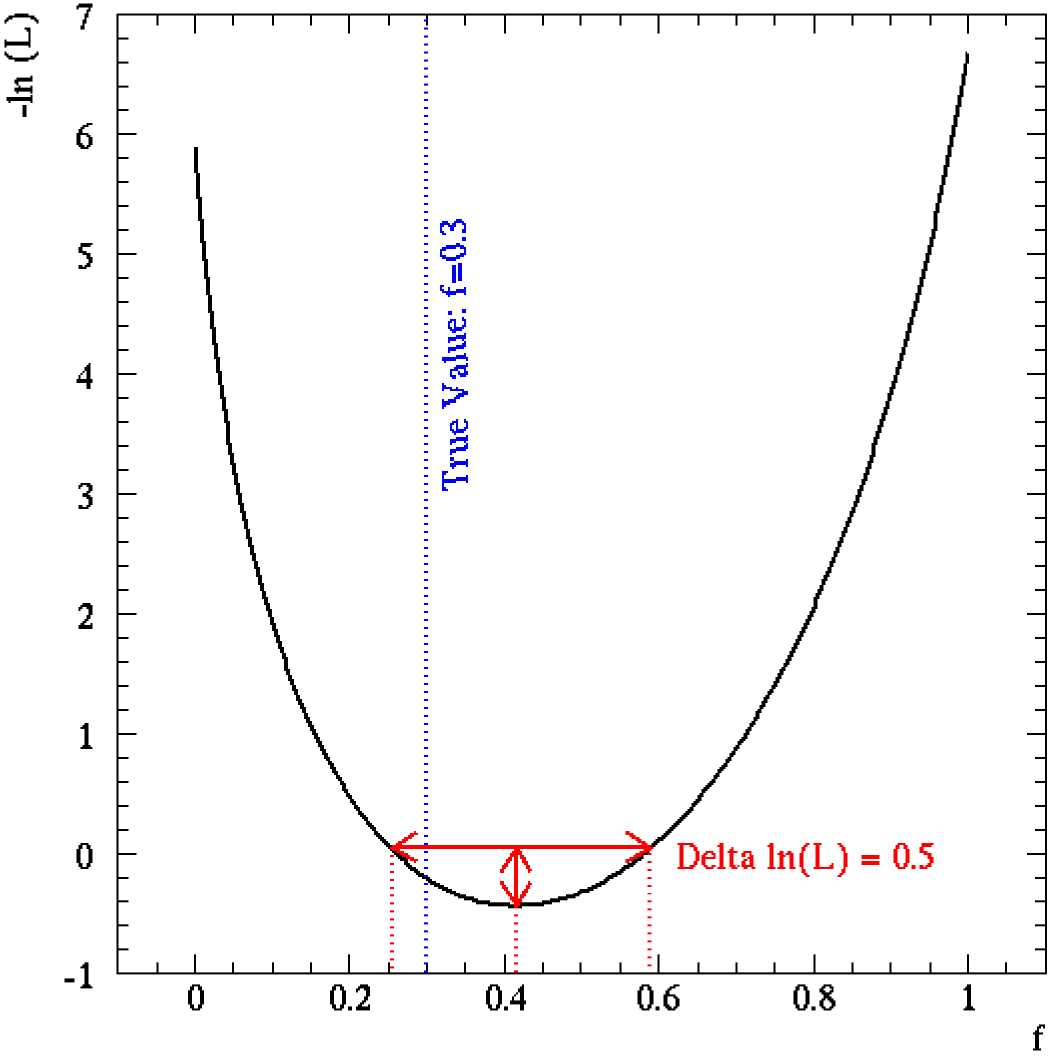
$$P_{tot}(x) = f\,P_A(x) + (1-f)\,P_B(x)$$

Form the negative log likelihood:

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

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

Graph of the log likelihood



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

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

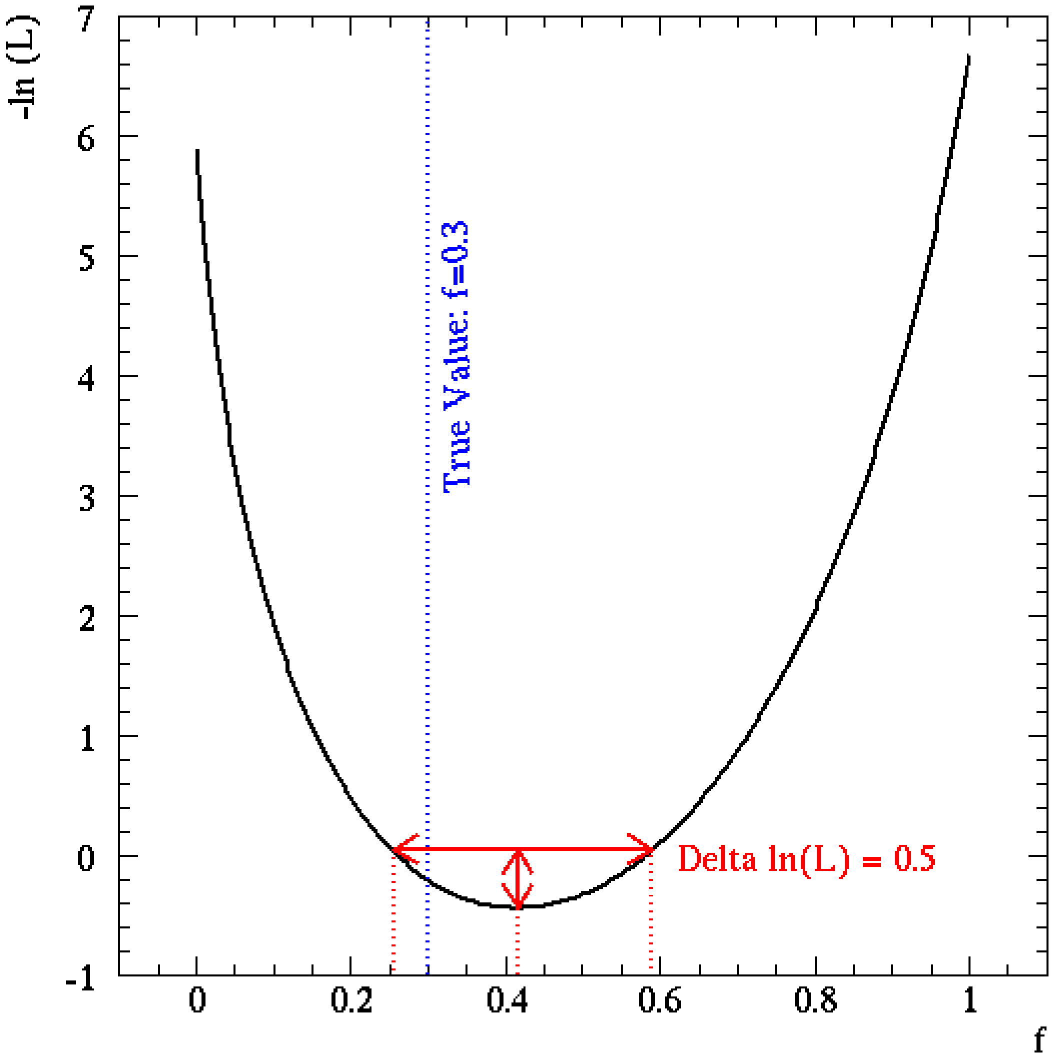
As we'll see, the “1σ” error range is defined by

Δ ln(L)=0.5 above the

minimum.

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

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:

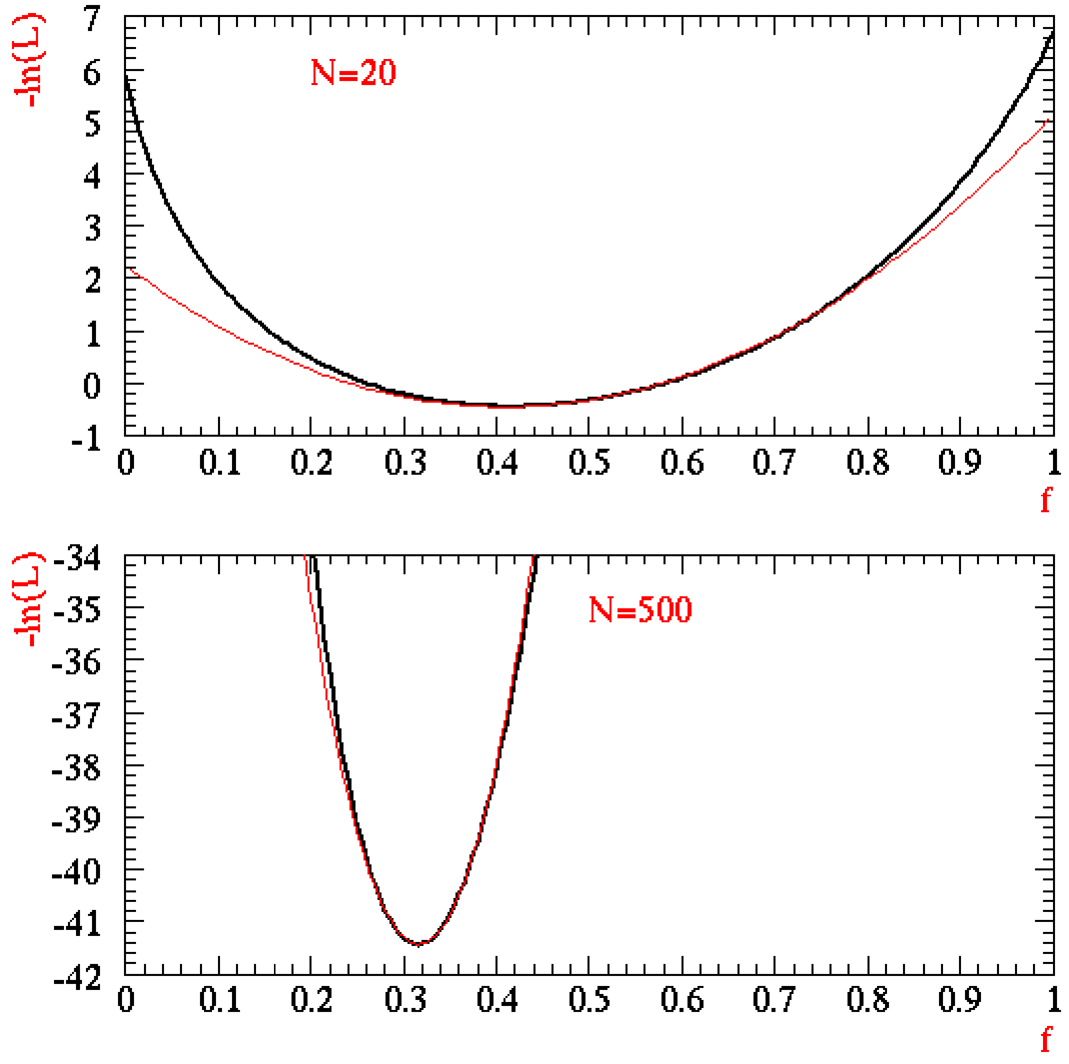
$$-\ln L = L_0 + \frac{1}{2}\left(\frac{f - \langle f \rangle}{\sigma_f}\right)^2$$

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

2σ range: Δ ln(L)=½(2)^2=2
3σ range: Δ ln(L)=½(3)^2=4.5

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

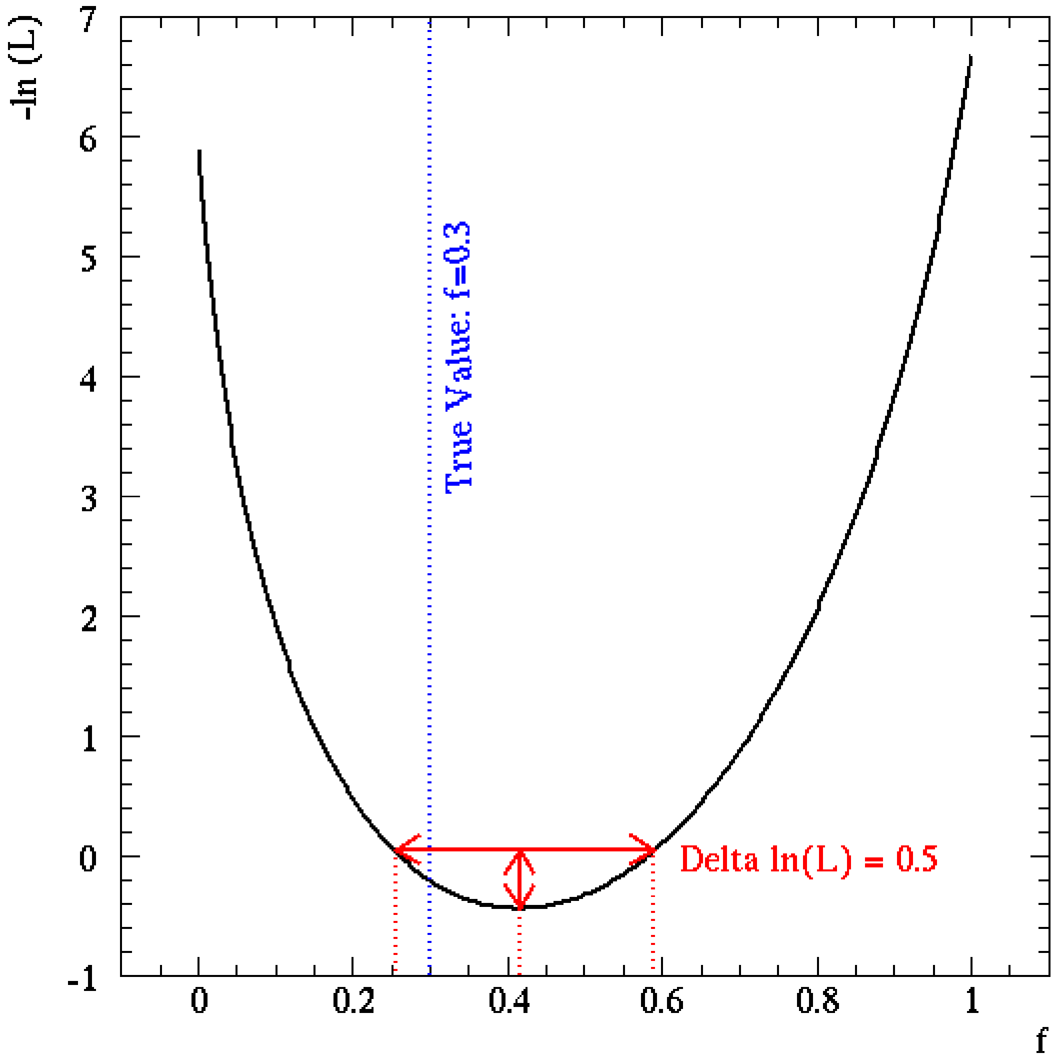
Parabolicity of the log likelihood



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



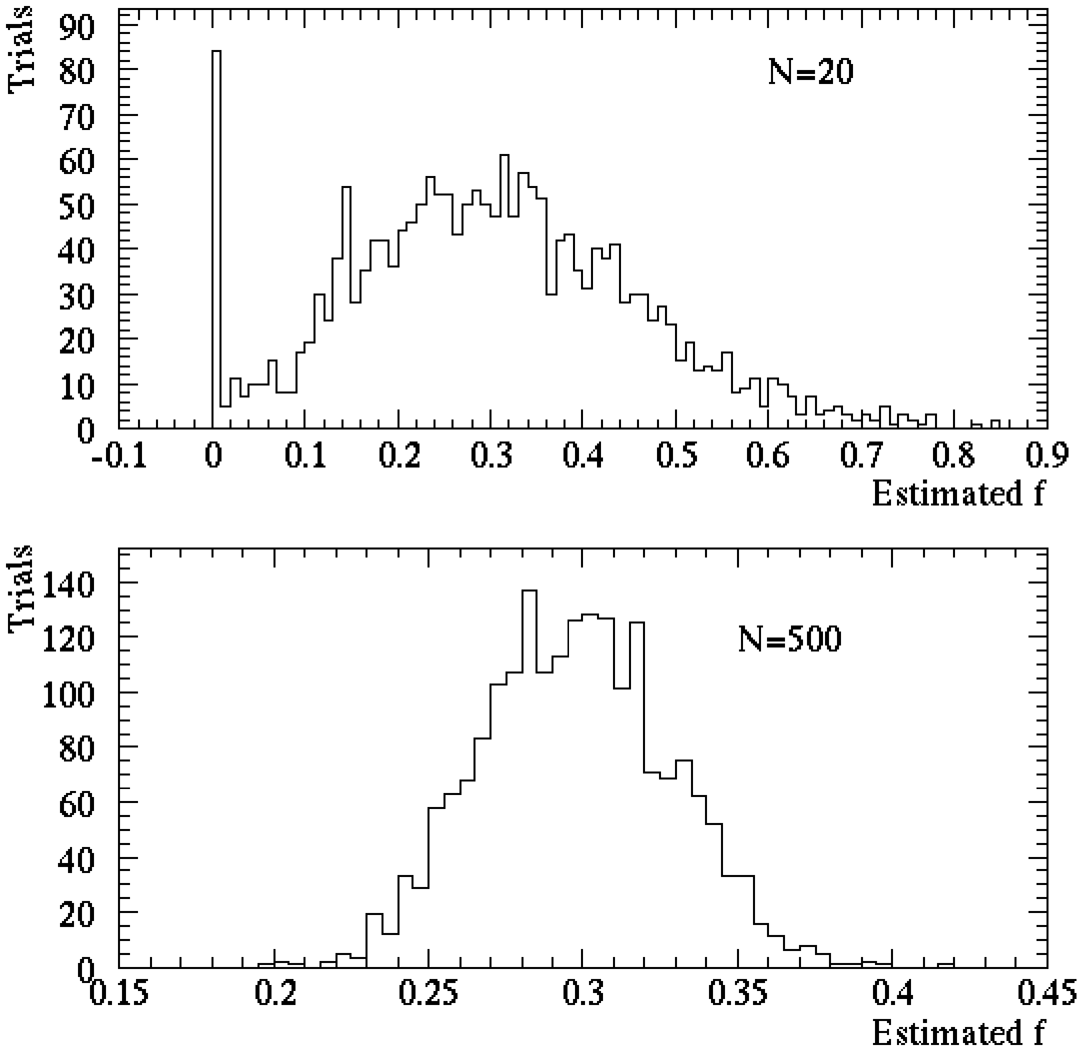
Even when the log likelihood is not Gaussian, it's nearly universal to define the 1σ range by Δ ln(L)=½. This can result in asymmetric error bars, such as:

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

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

Does this procedure actually work?

Coverage of ML estimator errors

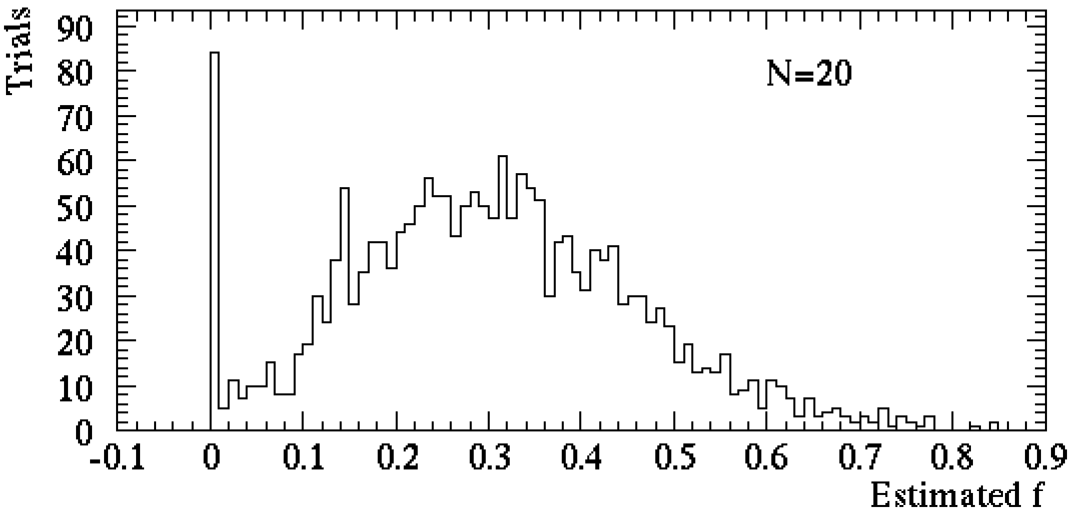


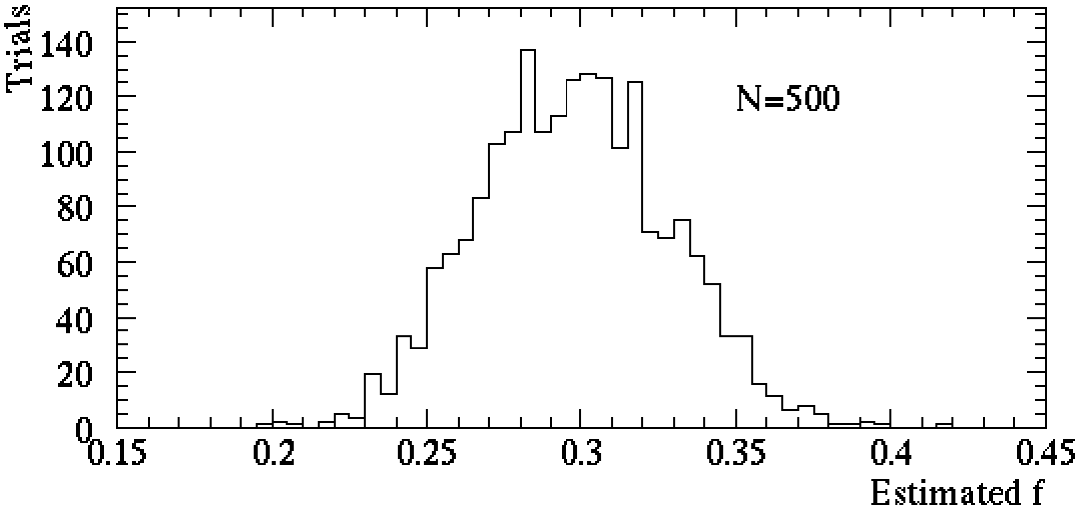
Distribution of ML estimators for two N values

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

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

Errors on ML estimators





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

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

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

Likelihood functions of multiple parameters

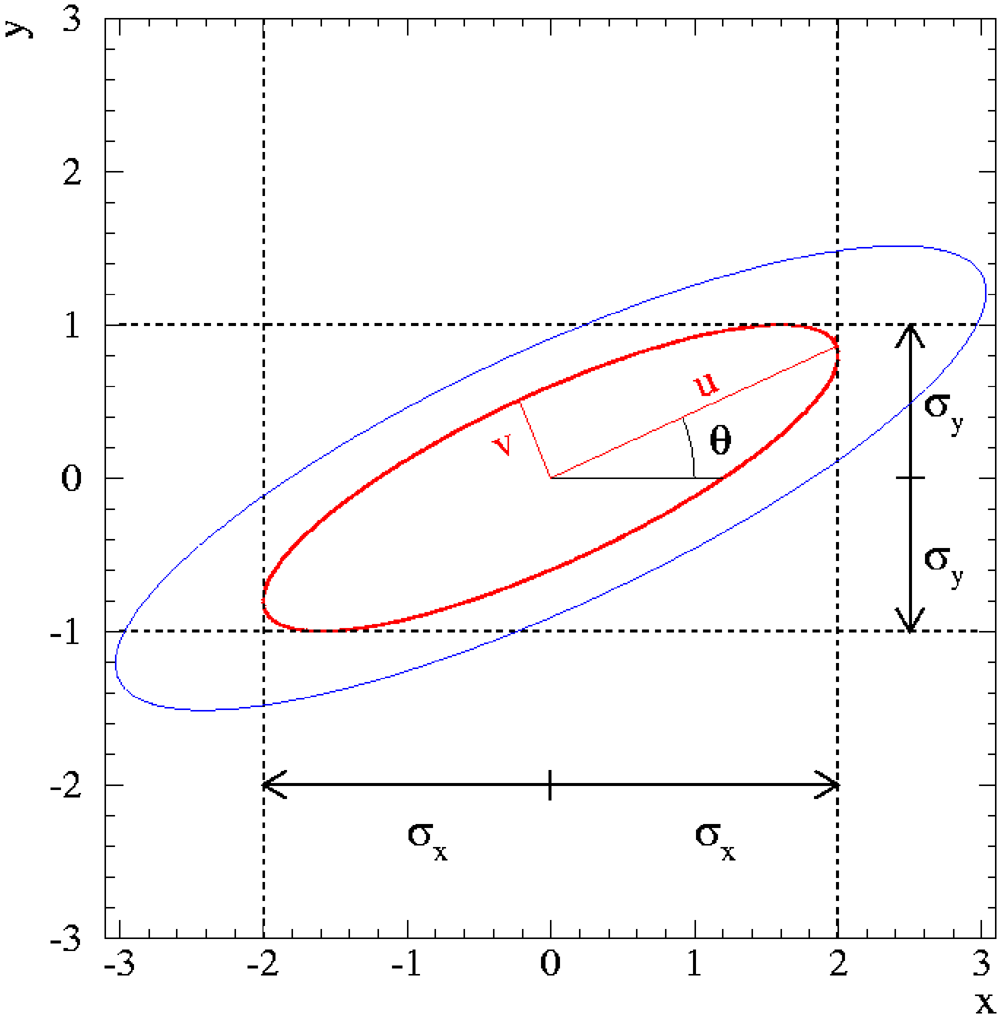
Often there is more than one free parameter. To handle this, we simply minimize the negative log likelihood over all free parameters.

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

Errors determined by (in the Gaussian approximation):

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

Error contours for multiple parameters



We can also find the errors on parameters by drawing contours on
Δ ln L.

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

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

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

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

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

Maximizing this is equivalent to minimizing

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

The Least Squares Method

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

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

Why?

1) Easily implemented.
2) Graphically motivated (see title slide!)
3) Mathematically straightforward---often analytic solution
4) Extension of LS to correlated uncertainties straightforward:

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

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

Δ 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.

Formulas for Errors in the Least Squares Method

We can also derive the errors by relating the χ^2 to the negative log likelihood, and using the error formula:

$$\mathrm{cov}^{-1}(a_i,a_j) = -\left\langle \frac{\partial^2 \ln L}{\partial a_i \partial a_j} \right\rangle = -\frac{\partial^2 \ln L}{\partial a_i \partial a_j}\Big|_{a=\hat{a}} = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial a_i \partial a_j}\Big|_{a=\hat{a}}$$

$$\sigma_{\hat{m}}^2 = \frac{1}{\sum 1/\sigma_i^2}\frac{1}{\langle x^2\rangle - \langle x\rangle^2} = \frac{\sigma^2}{N}\frac{1}{(\langle x^2\rangle - \langle x\rangle^2)}$$

$$\sigma_{\hat{b}}^2 = \frac{1}{\sum 1/\sigma_i^2}\frac{\langle x^2\rangle}{\langle x^2\rangle - \langle x\rangle^2} = \frac{\sigma^2}{N}\frac{\langle x^2\rangle}{(\langle x^2\rangle - \langle x\rangle^2)}$$

(intuitive when <x>=0)

$$\mathrm{cov}(\hat{m},\hat{b}) = -\frac{1}{\sum 1/\sigma_i^2}\frac{\langle x^2\rangle}{\langle x\rangle - \langle x\rangle^2} = -\frac{\sigma^2}{N}\frac{\langle x\rangle}{(\langle x^2\rangle - \langle x\rangle^2)}$$

Nonlinear least squares

The derivation of the least squares method doesn't depend on the assumption that your fitting function is linear in the parameters. Nonlinear fits, such as A + B sin(Ct + D), can be tackled with the least squares technique as well. But things aren't nearly as nice:

- No closed form solution---have to minimize the χ^2 numerically.
- Estimators are no longer guaranteed to have zero bias and minimum variance.
- Contours generated by $\Delta\chi^2$=+1 no longer are ellipses, and the tangents to these contours no longer give the standard deviations. (However, we can still interpret them as giving "1σ" errors---although since the distribution is non-Gaussian, this error range isn't the same thing as a standard deviation
- Be very careful with minimization routines---depending on how badly non-linear your problem is, there may be multiple solutions, local minima, etc.

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

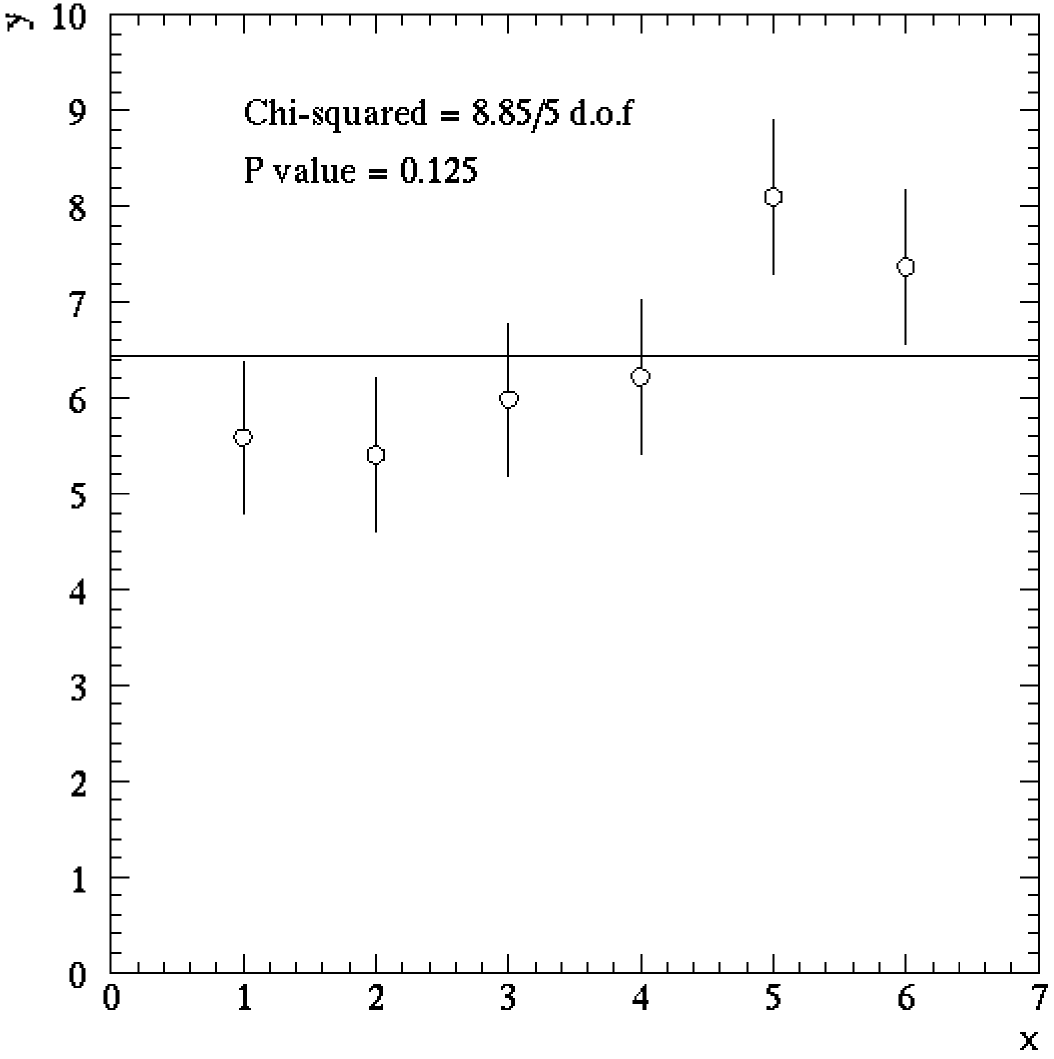
Remember that the sum of N Gaussian variables with zero mean and unit RMS, when squared and added, follows a χ^2 distribution with N degrees of freedom. Compare to the least squares formula:

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

If each y_i is distributed around the function according to a Gaussian, **and** f(x|α) 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.

Goodness of fit: an example



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

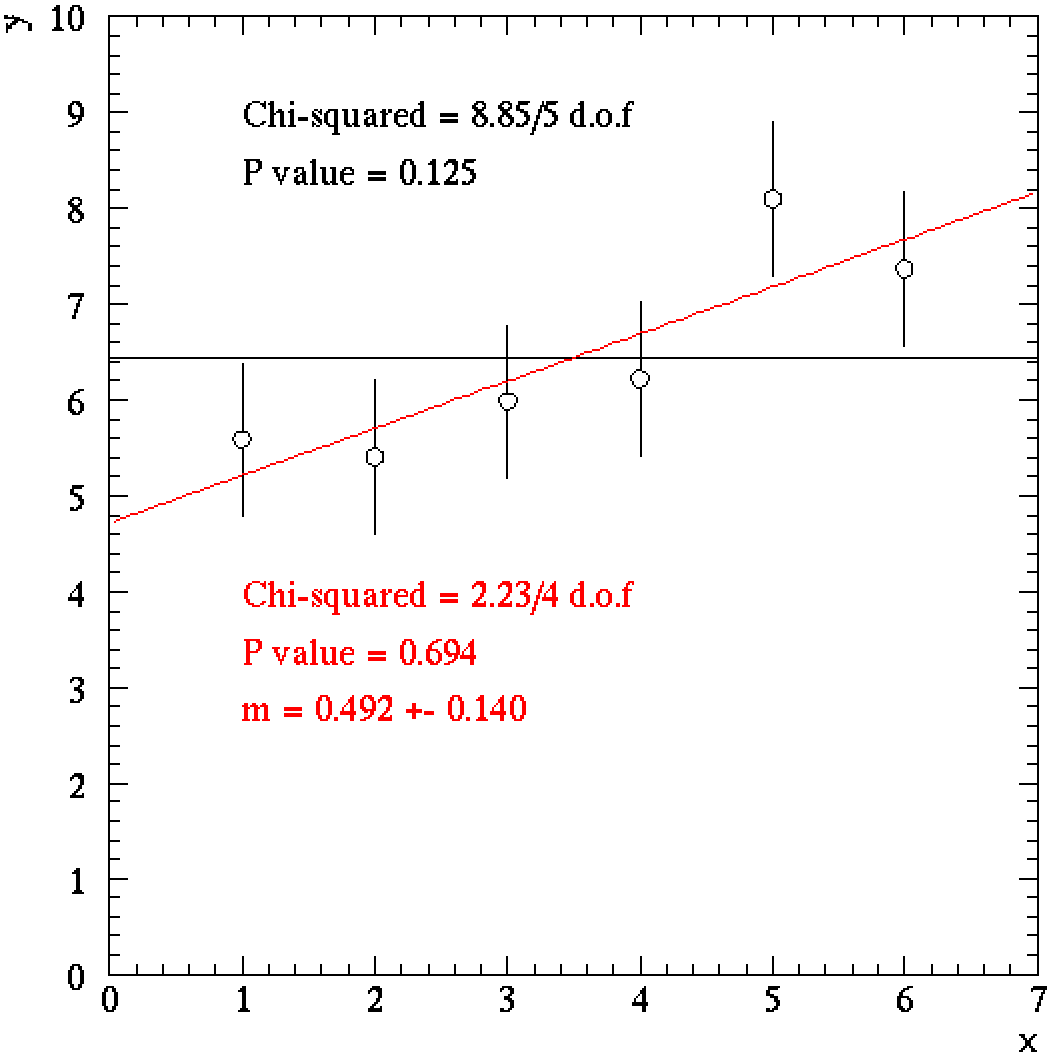
6 data points – 1 free parameter = 5 d.o.f.

χ^2 = 8.85/5 d.o.f.

Chance of getting a larger χ^2 is 12.5%---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?

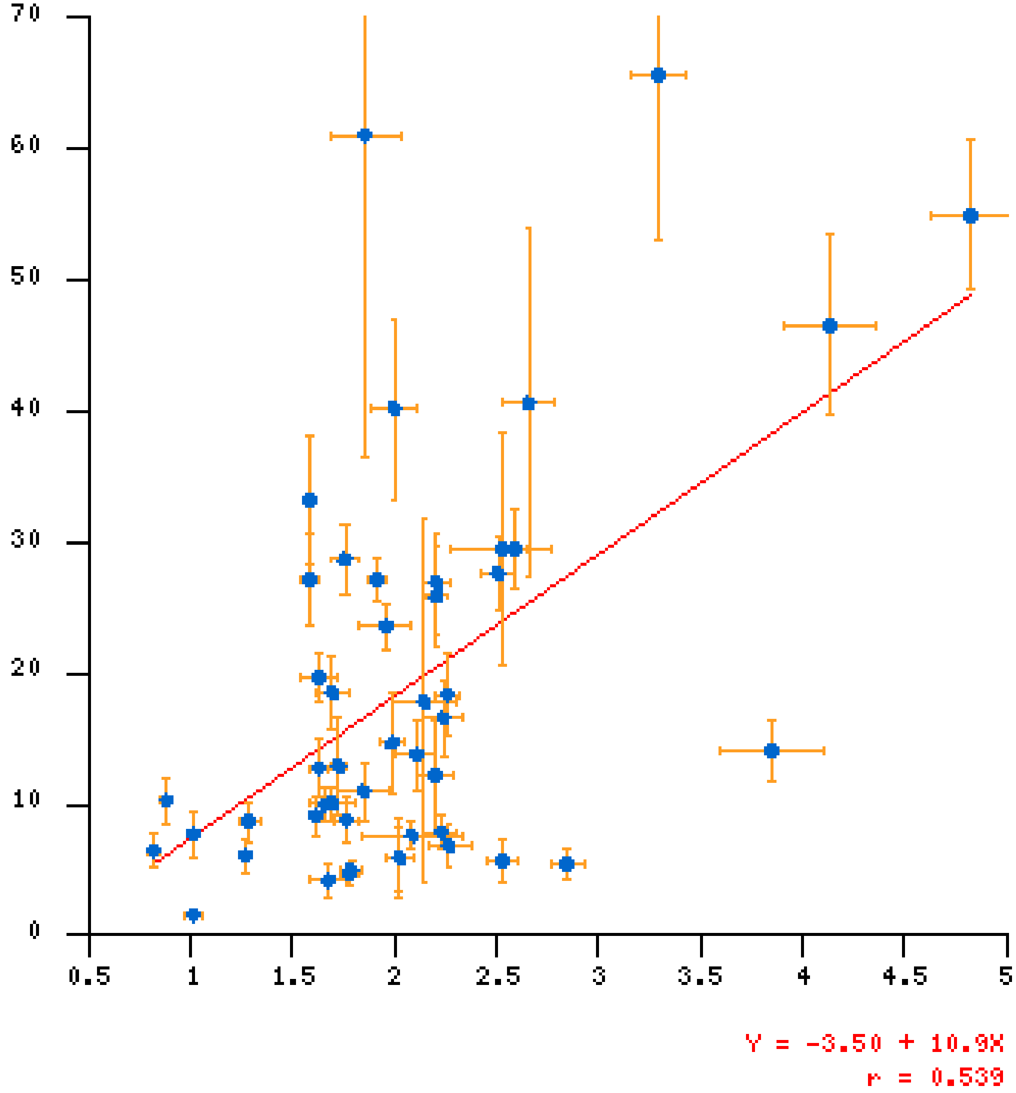
2) Parameter estimation: which of the following limited set of hypotheses is most consistent with the data?

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?”

What is an error bar?



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

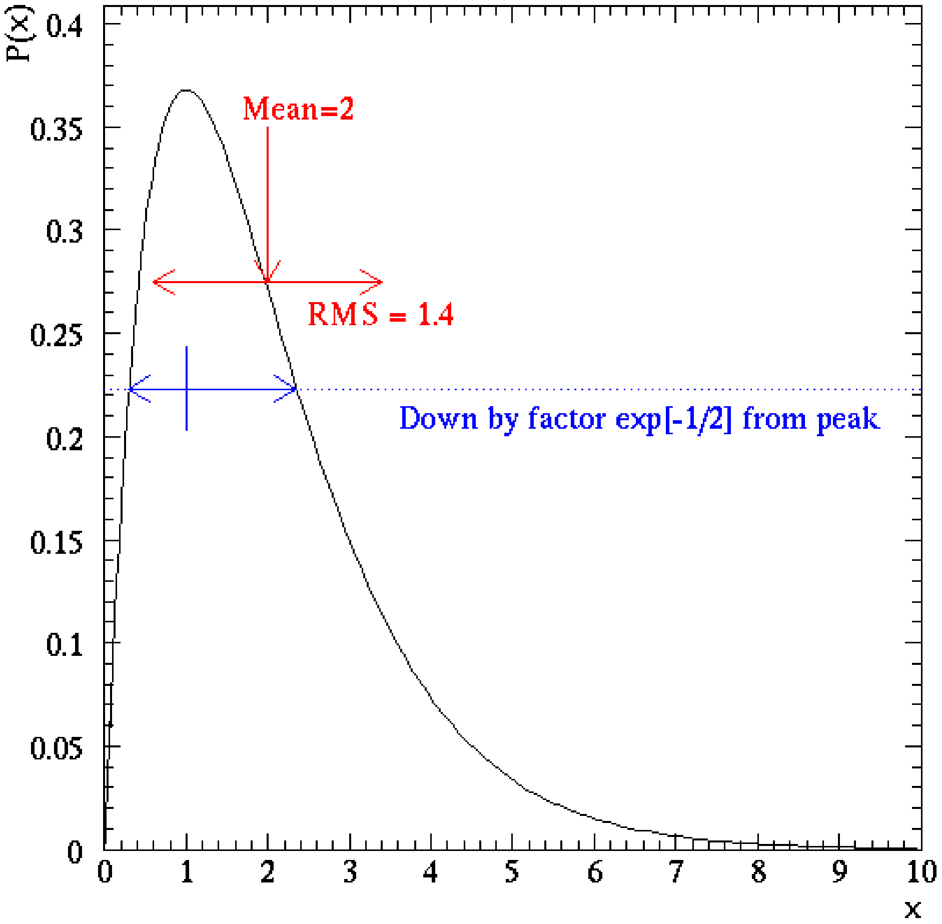
Answer: you can never be sure, unless it's specified!

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

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

Correlations unknown!

Relation of an error bar to PDF shape



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

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

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

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

An error bar is a shorthand approximation to a PDF!

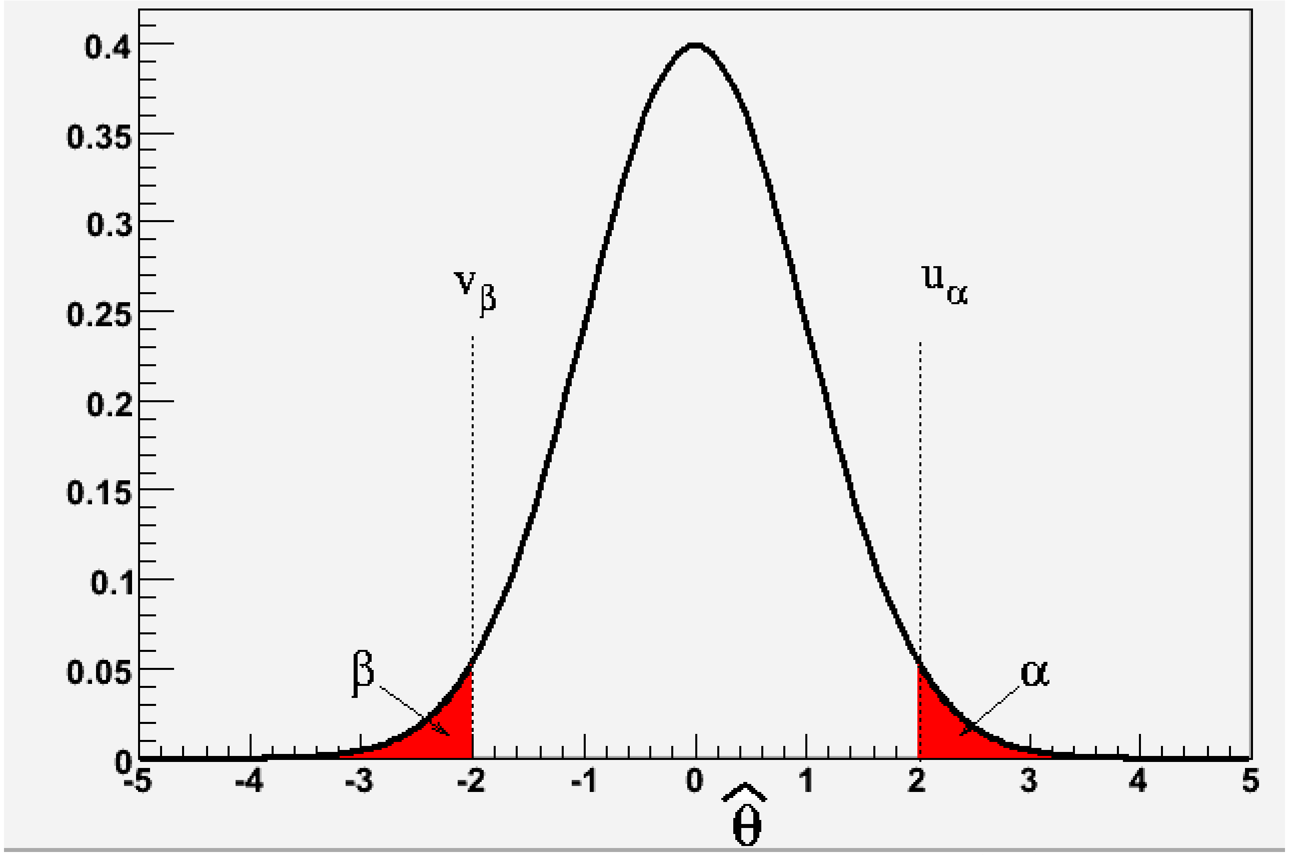
In an ideal Bayesian universe, error bars don't exist. Instead, everyone will use the full prior PDF and the data to calculate the posterior PDF, and then report the shape of that PDF (preferably as a graph or table).

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

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

An error bar as a confidence interval

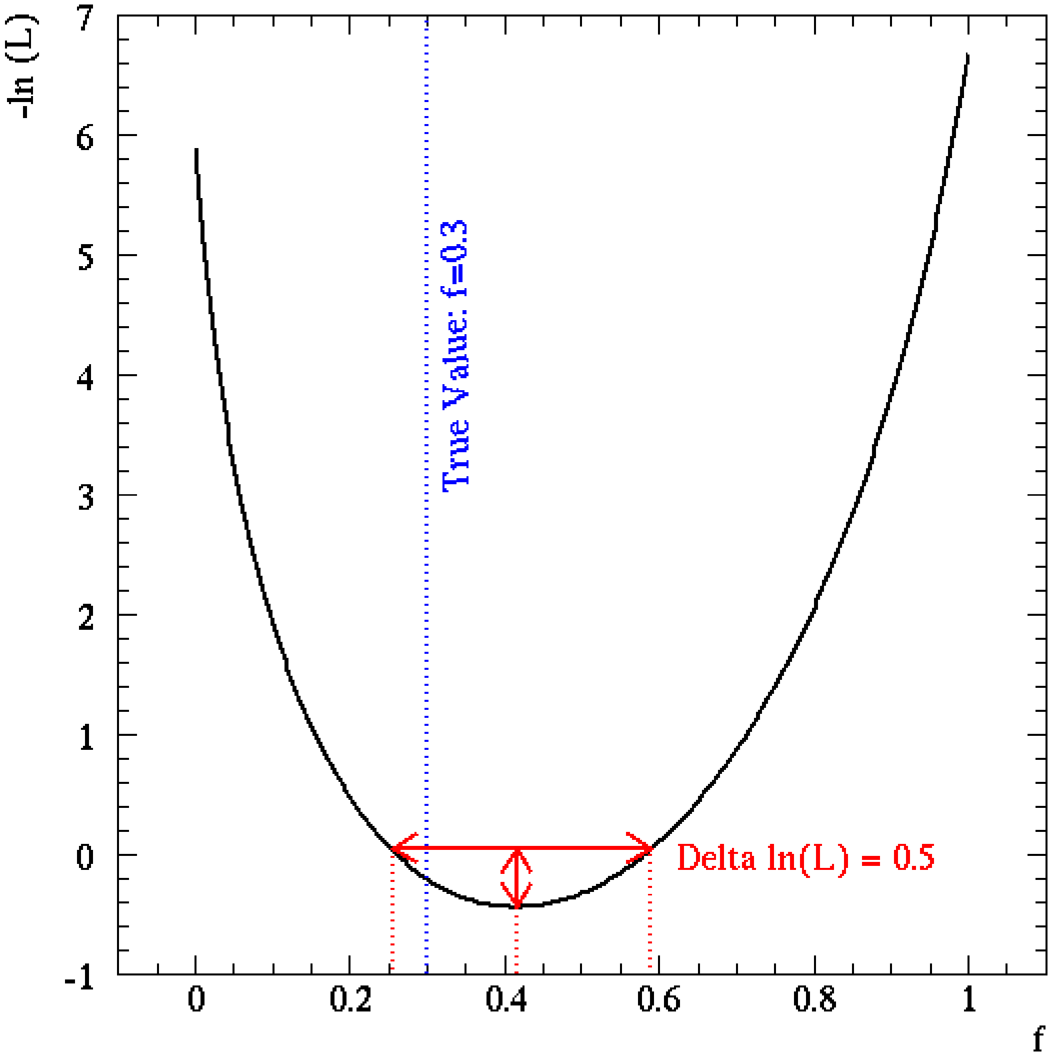
Frequentist techniques don't directly answer the question of what the probability is for a parameter to have a particular value. All you can calculate is the probability of observing your data given a value of the parameter.The confidence interval construction is a dodge to get around this.



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

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

The Δ ln(L) rule



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

This is only an approximation, and does not give exactly the right coverage when N is small.

More generally, if you have d free parameters, then the quantity $\omega = $ “$\Delta\chi^2$” $= 2[\ln(L_{max})-\ln(L)]$ approximates a χ^2 with d degrees of freedom.

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

Error-weighted averages

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

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

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

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

Averaging correlated measurements II

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

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

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:

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

Averaging correlated measurements III

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

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

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

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

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

The error propagation equation

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

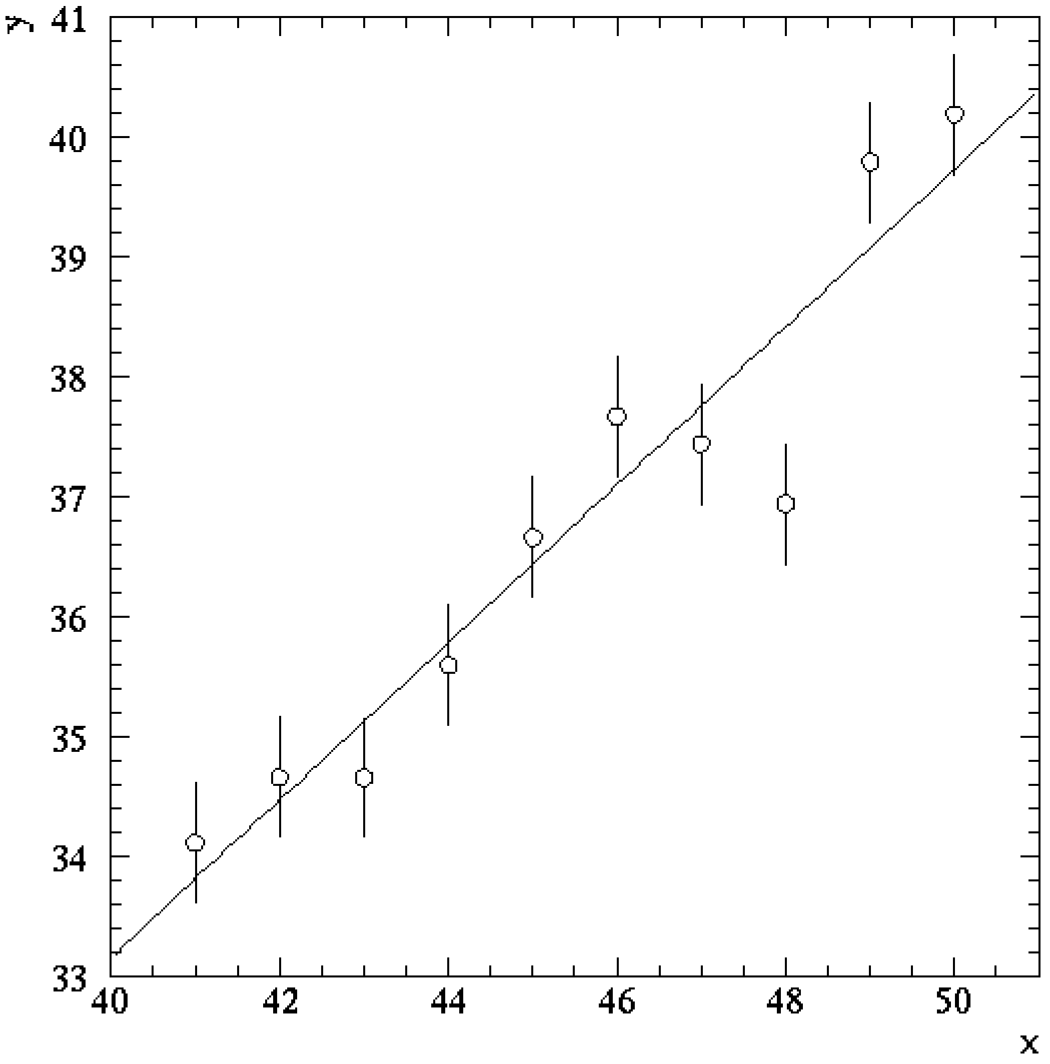
$$\sigma_f^2 = \left(\frac{df}{dx}\right)^2 \sigma_x^2 + \left(\frac{df}{dy}\right)^2 \sigma_y^2 + 2\left(\frac{df}{dx}\right)\left(\frac{df}{dy}\right)\rho\,\sigma_x\sigma_y$$

The assumptions underlying the error propagation equation are:

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

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

Example: interpolating a straight line fit



Straight line fit y=mx+b

Reported values from a standard fitting package:

m = 0.658 ± 0.056
b = 6.81 ± 2.57

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

y=0.658*45.5+6.81=36.75

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

UGH! NONSENSE!

Example: straight line fit, done correctly

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

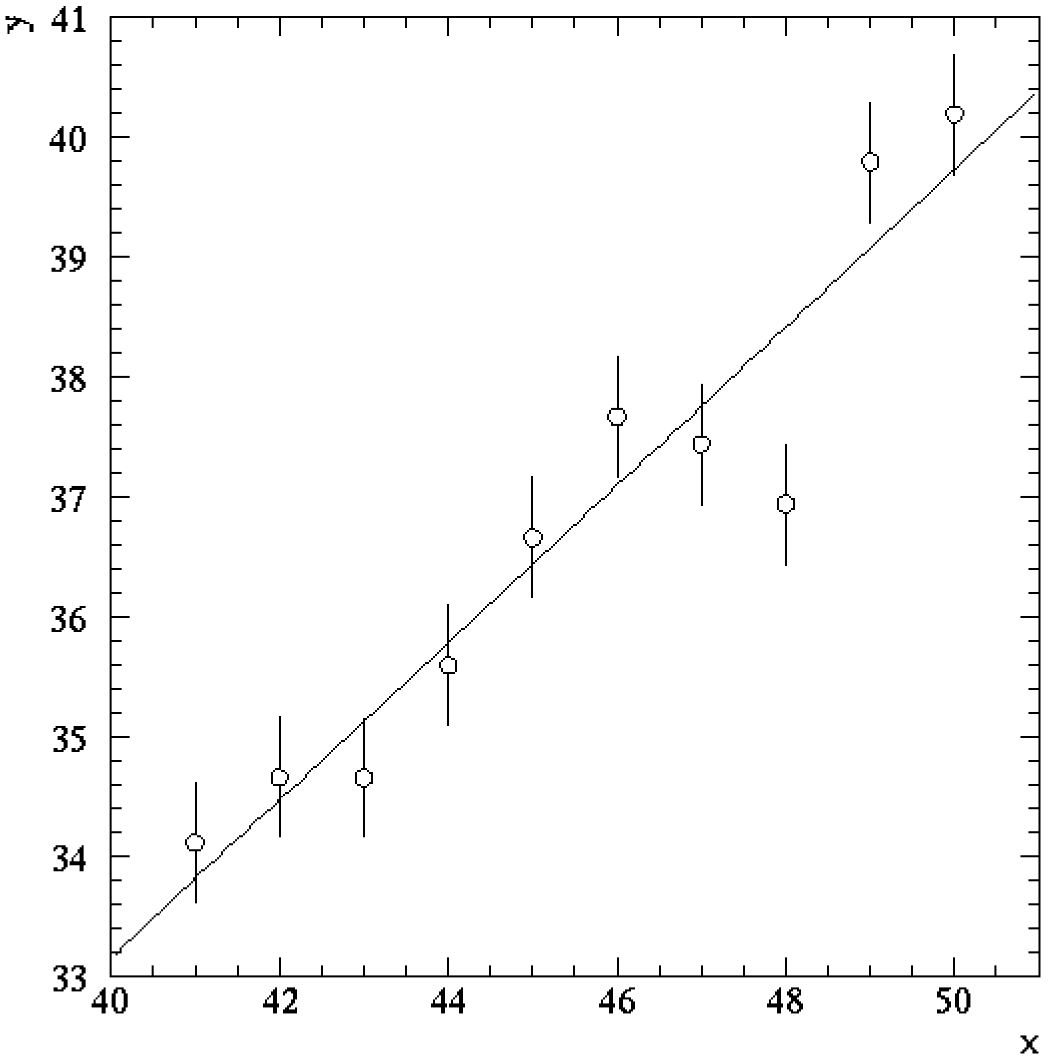
m = 0.0658 + .056
b = 6.81 + 2.57
ρ = -0.9981

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

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

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

Reducing correlations in the straight line fit



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

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

$b' = 36.77 \pm 0.16$
$m = 0.658 \pm .085$
$\rho = 0.43$

dy at $x=45.5 = 0.16$