

# Lecture 3

## 1 Probability (90 min.)

Definition, Bayes' theorem, probability densities and their properties, catalogue of pdfs, Monte Carlo

## 2 Statistical tests (90 min.)

general concepts, test statistics, multivariate methods, goodness-of-fit tests

## → 3 Parameter estimation (90 min.)

general concepts, maximum likelihood, variance of estimators, least squares

## 4 Interval estimation (60 min.)

setting limits

## 5 Further topics (60 min.)

systematic errors, MCMC

# Parameter estimation

The parameters of a pdf are constants that characterize its shape, e.g.

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$$

r.v.                      parameter

Suppose we have a **sample** of observed values:  $\vec{x} = (x_1, \dots, x_n)$

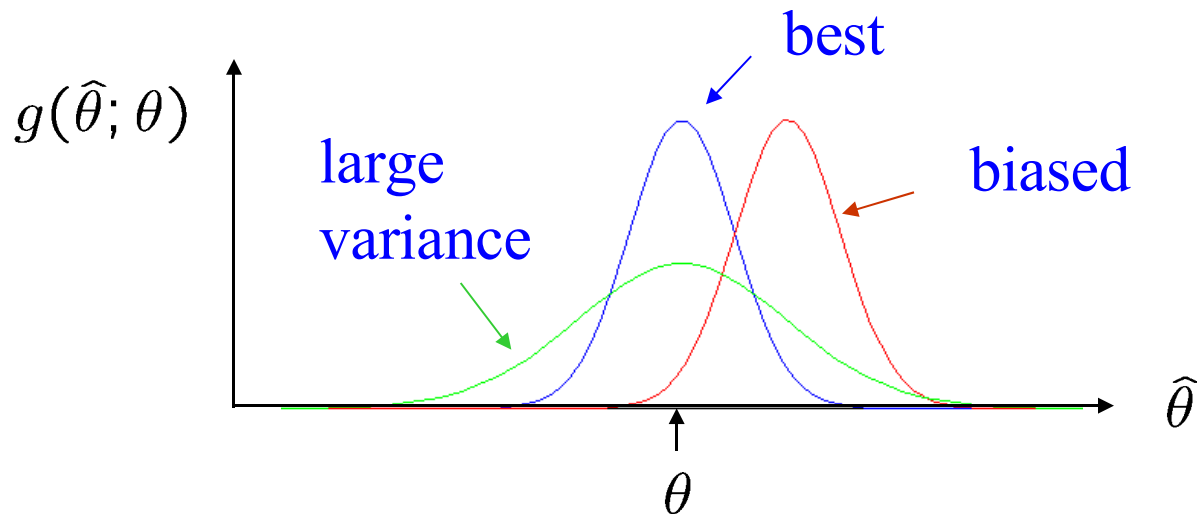
We want to find some function of the data to **estimate** the parameter(s):

$$\hat{\theta}(\vec{x}) \leftarrow \text{estimator written with a hat}$$

Sometimes we say ‘estimator’ for the function of  $x_1, \dots, x_n$ ;  
‘estimate’ for the value of the estimator with a particular data set.

# Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



We want small (or zero) bias (systematic error):  $b = E[\hat{\theta}] - \theta$

→ average of repeated estimates should tend to true value.

And we want a small variance (statistical error):  $V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

# An estimator for the mean (expectation value)

Parameter:  $\mu = E[x]$

Estimator:  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \equiv \bar{x}$  ('sample mean')

We find:  $b = E[\hat{\mu}] - \mu = 0$

$$V[\hat{\mu}] = \frac{\sigma^2}{n} \quad \left( \sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} \right)$$

# An estimator for the variance

Parameter:  $\sigma^2 = V[x]$

Estimator:  $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \equiv s^2$  ('sample variance')

We find:

$$b = E[\hat{\sigma}^2] - \sigma^2 = 0 \quad (\text{factor of } n-1 \text{ makes this so})$$

$$V[\hat{\sigma}^2] = \frac{1}{n} \left( \mu_4 - \frac{n-3}{n-1} \mu_2 \right), \quad \text{where}$$

$$\mu_k = \int (x - \mu)^k f(x) dx$$

# The likelihood function

Consider  $n$  independent observations of  $x$ :  $x_1, \dots, x_n$ , where  $x$  follows  $f(x; \theta)$ . The joint pdf for the whole data sample is:

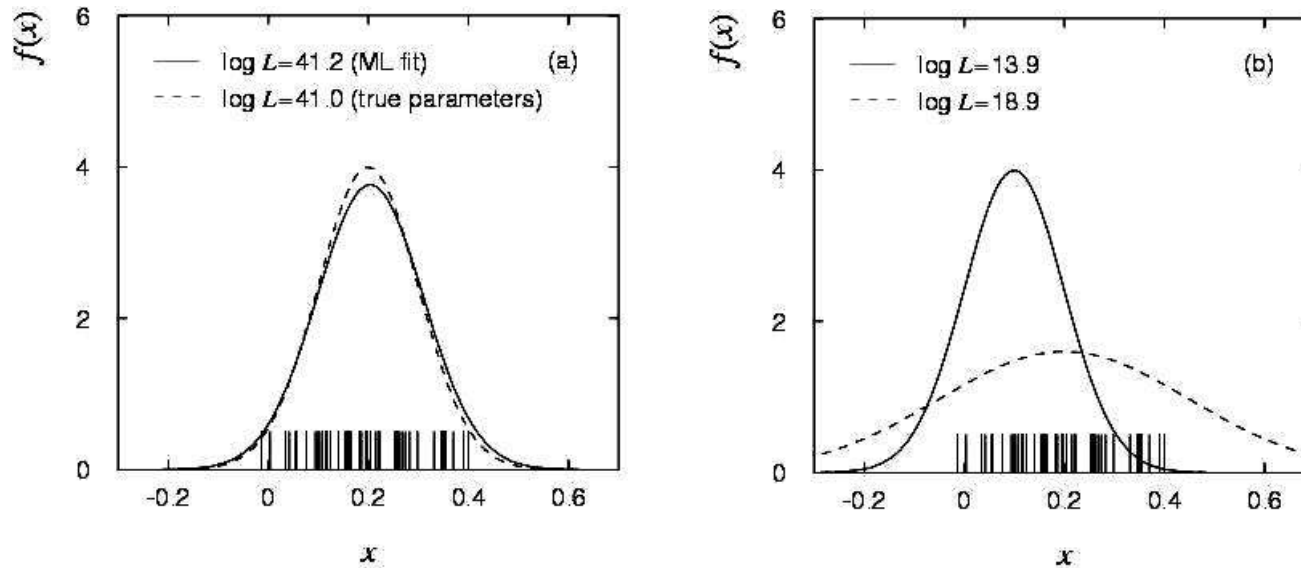
$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

Now evaluate this function with the data sample obtained and regard it as a function of the parameter(s). This is the **likelihood function**:

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta) \quad (x_i \text{ constant})$$

# Maximum likelihood estimators

If the hypothesized  $\theta$  is close to the true value, then we expect a high probability to get data like that which we actually found.



So we define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

ML estimators not guaranteed to have any ‘optimal’ properties, (but in practice they’re very good).

# ML example: parameter of exponential pdf

Consider exponential pdf,  $f(t; \tau) = \frac{1}{\tau} e^{-t/\tau}$

and suppose we have data,  $t_1, \dots, t_n$

The likelihood function is  $L(\tau) = \prod_{i=1}^n \frac{1}{\tau} e^{-t_i/\tau}$

The value of  $\tau$  for which  $L(\tau)$  is maximum also gives the maximum value of its logarithm (the log-likelihood function):

$$\ln L(\tau) = \sum_{i=1}^n \ln f(t_i; \tau) = \sum_{i=1}^n \left( \ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$



# ML example: parameter of exponential pdf (2)

Find its maximum from  $\frac{\partial \ln L(\tau)}{\partial \tau} = 0$ ,

$$\rightarrow \hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$$

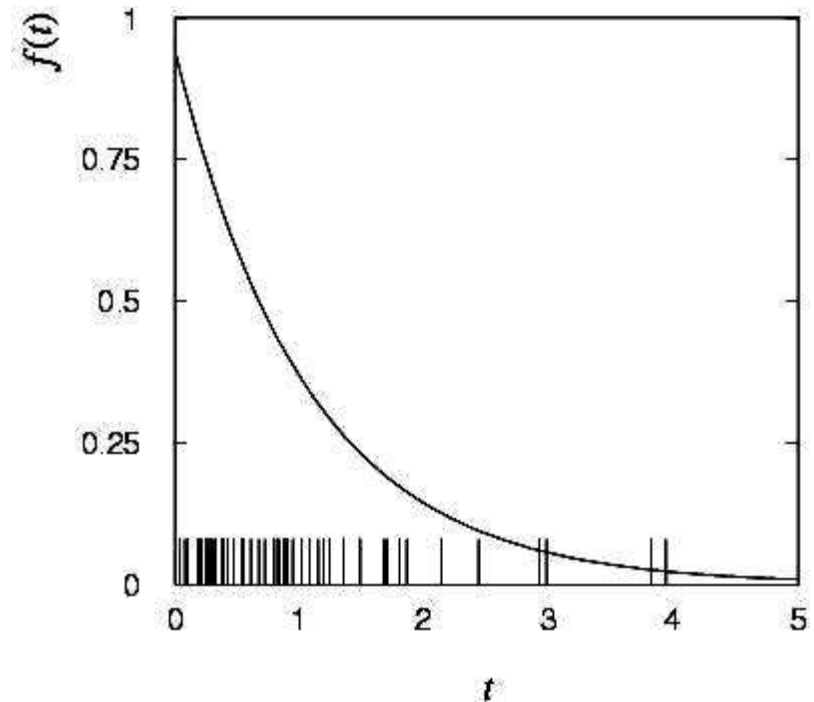
Monte Carlo test:

generate 50 values

using  $\tau = 1$ :

We find the ML estimate:

$$\hat{\tau} = 1.062$$



(Exercise: show this estimator is unbiased.)

# Functions of ML estimators

Suppose we had written the exponential pdf as  $f(t; \lambda) = \lambda e^{-\lambda t}$ ,  
i.e., we use  $\lambda = 1/\tau$ . What is the ML estimator for  $\lambda$ ?

For a function  $\alpha(\theta)$  of a parameter  $\theta$ , it doesn't matter  
whether we express  $L$  as a function of  $\alpha$  or  $\theta$ .

The ML estimator of a function  $\alpha(\theta)$  is simply  $\hat{\alpha} = \alpha(\hat{\theta})$ .

So for the decay constant we have  $\hat{\lambda} = \frac{1}{\hat{\tau}} = \left( \frac{1}{n} \sum_{i=1}^n t_i \right)^{-1}$ .

Caveat:  $\hat{\lambda}$  is biased, even though  $\hat{\tau}$  is unbiased.

Can show  $E[\hat{\lambda}] = \lambda \frac{n}{n-1}$ . (bias  $\rightarrow 0$  for  $n \rightarrow \infty$ )

# Example of ML: parameters of Gaussian pdf

Consider independent  $x_1, \dots, x_n$ , with  $x_i \sim \text{Gaussian}(\mu, \sigma^2)$

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

The log-likelihood function is

$$\begin{aligned} \ln L(\mu, \sigma^2) &= \sum_{i=1}^n \ln f(x_i; \mu, \sigma^2) \\ &= \sum_{i=1}^n \left( \ln \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \ln \frac{1}{\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right). \end{aligned}$$

## Example of ML: parameters of Gaussian pdf (2)

Set derivatives with respect to  $\mu$ ,  $\sigma^2$  to zero and solve,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2.$$

We already know that the estimator for  $\mu$  is unbiased.

But we find, however,  $E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$ , so ML estimator for  $\sigma^2$  has a bias, but  $b \rightarrow 0$  for  $n \rightarrow \infty$ . Recall, however, that

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

is an unbiased estimator for  $\sigma^2$ .

# Variance of estimators: Monte Carlo method

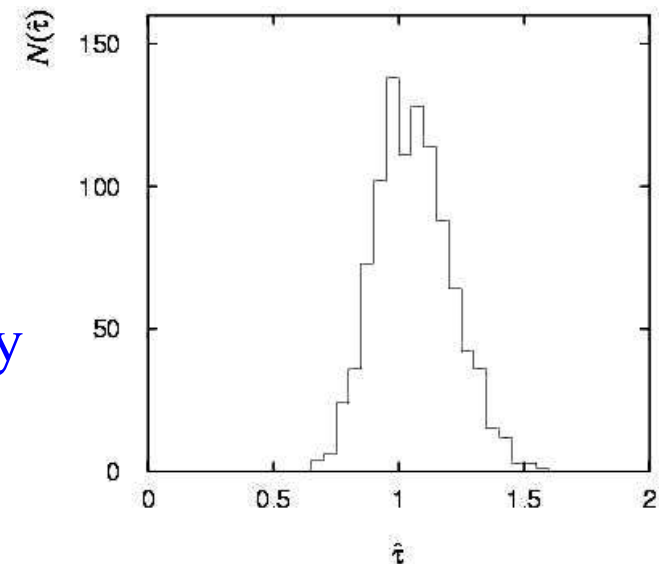
Having estimated our parameter we now need to report its ‘statistical error’, i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\hat{\sigma}_{\hat{\tau}} = 0.151$$

Note distribution of estimates is roughly Gaussian – (almost) always true for ML in large sample limit.



# Variance of estimators from information inequality

The **information inequality** (RCF) sets a minimum variance bound (MVB) for any estimator (not only ML):

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \bigg/ E \left[ -\frac{\partial^2 \ln L}{\partial \theta^2} \right] \equiv \text{MVB} \quad (b = E[\hat{\theta}] - \theta)$$

Often the bias  $b$  is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\hat{\theta}] \approx -1 \bigg/ E \left[ \frac{\partial^2 \ln L}{\partial \theta^2} \right]$$

Estimate this using the 2nd derivative of  $\ln L$  at its maximum:

$$\hat{V}[\hat{\theta}] = - \left( \frac{\partial^2 \ln L}{\partial \theta^2} \right)^{-1} \bigg|_{\theta=\hat{\theta}}$$

# Variance of estimators: graphical method

Expand  $\ln L(\theta)$  about its maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[ \frac{\partial \ln L}{\partial \theta} \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!} \left[ \frac{\partial^2 \ln L}{\partial \theta^2} \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \dots$$

First term is  $\ln L_{\max}$ , second term is zero, for third term use information inequality (assume equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma}_{\hat{\theta}}^2}$$

$$\text{i.e.,} \quad \ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

→ to get  $\hat{\sigma}_{\hat{\theta}}$ , change  $\theta$  away from  $\hat{\theta}$  until  $\ln L$  decreases by 1/2.

# Example of variance by graphical method

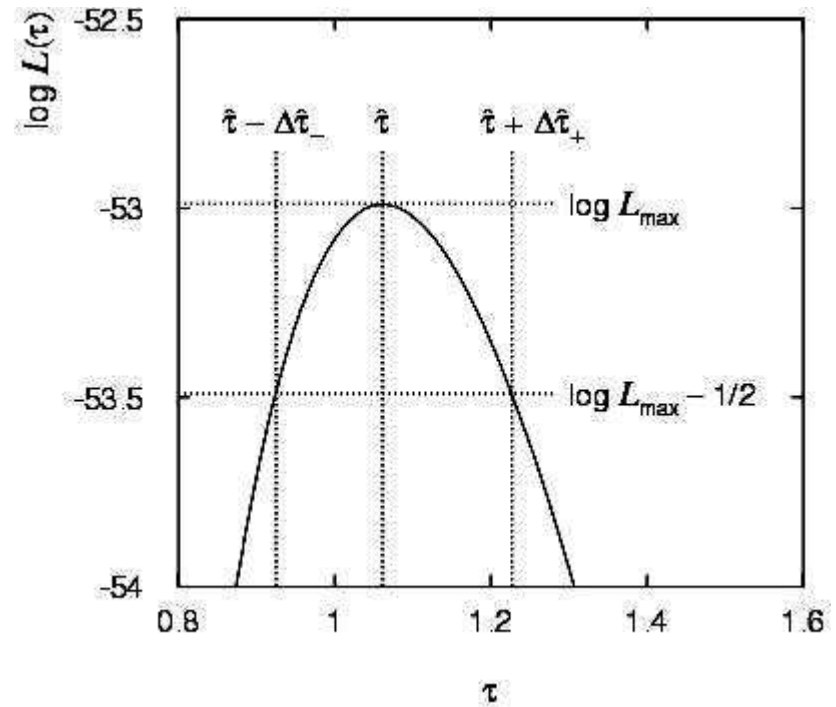
ML example with exponential:

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$



Not quite parabolic  $\ln L$  since finite sample size ( $n = 50$ ).



# Information inequality for $n$ parameters

Suppose we have estimated  $n$  parameters  $\vec{\theta} = (\theta_1, \dots, \theta_n)$ .

The (inverse) minimum variance bound is given by the Fisher information matrix:

$$I_{ij} = E \left[ -\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right] = -n \int f(x; \vec{\theta}) \frac{\partial^2 \ln f(x; \vec{\theta})}{\partial \theta_i \partial \theta_j} dx$$

The information inequality then states that  $V - I^{-1}$  is a positive semi-definite matrix; therefore for the diagonal elements,

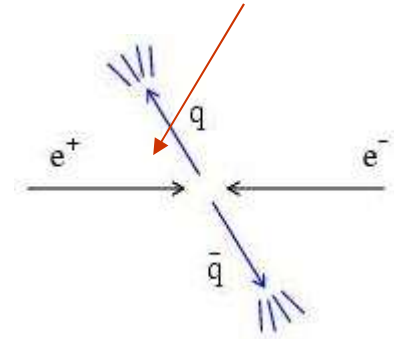
$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use  $I^{-1}$  as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of  $L$ .

# Example of ML with 2 parameters

Consider a scattering angle distribution with  $x = \cos \theta$ ,

$$f(x; \alpha, \beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta/3}$$



or if  $x_{\min} < x < x_{\max}$ , need always to normalize so that

$$\int_{x_{\min}}^{x_{\max}} f(x; \alpha, \beta) dx = 1 .$$

Example:  $\alpha = 0.5$ ,  $\beta = 0.5$ ,  $x_{\min} = -0.95$ ,  $x_{\max} = 0.95$ ,  
generate  $n = 2000$  events with Monte Carlo.

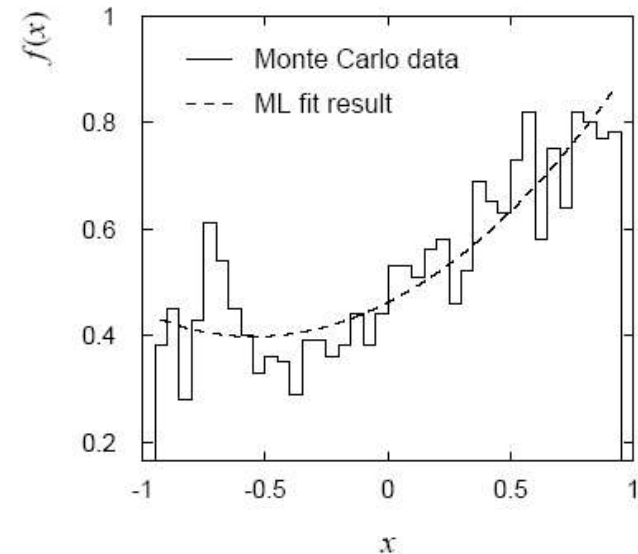
# Example of ML with 2 parameters: fit result

Finding maximum of  $\ln L(\alpha, \beta)$  numerically (MINUIT) gives

$$\hat{\alpha} = 0.508$$

$$\hat{\beta} = 0.47$$

**N.B.** No binning of data for fit, but can compare to histogram for goodness-of-fit (e.g. ‘visual’ or  $\chi^2$ ).



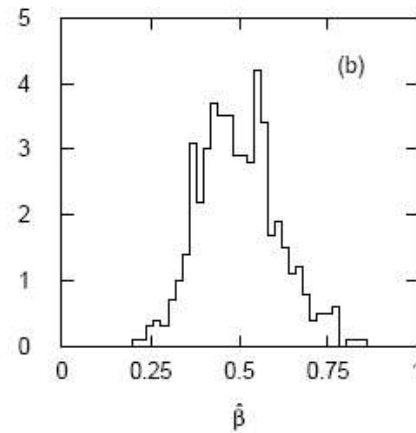
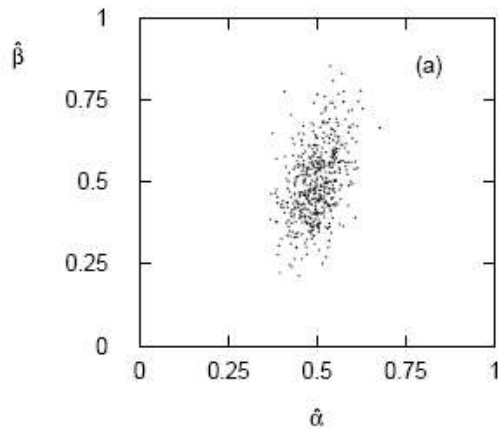
(Co)variances from  $(\widehat{V}^{-1})_{ij} = -\left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\vec{\theta}=\vec{\hat{\theta}}}$  (MINUIT routine HESSE)

$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \quad \text{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$

$$\hat{\sigma}_{\hat{\beta}} = 0.11 \quad r = 0.46$$

# Two-parameter fit: MC study

Repeat ML fit with 500 experiments, all with  $n = 2000$  events:



$$\overline{\hat{\alpha}} = 0.499$$

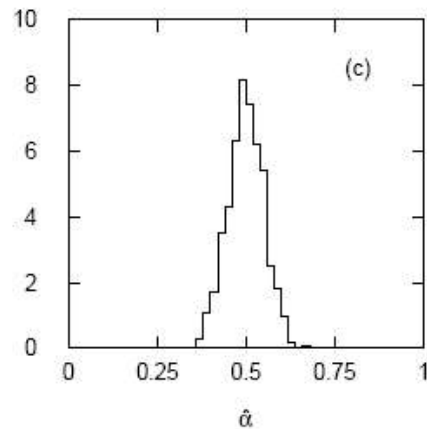
$$s_{\hat{\alpha}} = 0.051$$

$$\overline{\hat{\beta}} = 0.498$$

$$s_{\hat{\beta}} = 0.111$$

$$\text{cov}[\hat{\alpha}, \hat{\beta}] = 0.0024$$

$$r = 0.42$$



Estimates average to  $\sim$  true values;  
(Co)variances close to previous estimates;  
marginal pdfs approximately Gaussian.

# The $\ln L_{\max} - 1/2$ contour

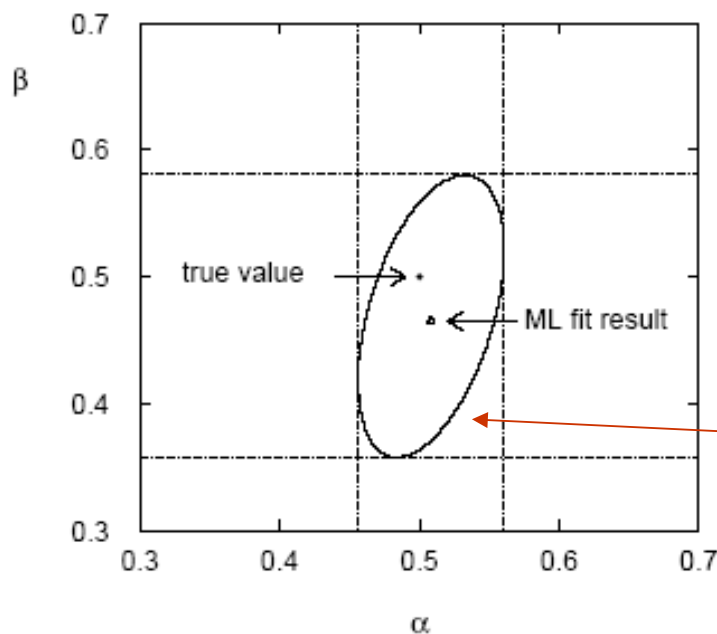
For large  $n$ ,  $\ln L$  takes on quadratic form near maximum:

$$\ln L(\alpha, \beta) \approx \ln L_{\max} - \frac{1}{2(1 - \rho^2)} \left[ \left( \frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right)^2 + \left( \frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2 - 2\rho \left( \frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right) \left( \frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right) \right]$$

The contour  $\ln L(\alpha, \beta) = \ln L_{\max} - 1/2$  is an ellipse:

$$\frac{1}{(1 - \rho^2)} \left[ \left( \frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right)^2 + \left( \frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2 - 2\rho \left( \frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right) \left( \frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right) \right] = 1$$

## (Co)variances from $\ln L$ contour



The  $\alpha, \beta$  plane for the first MC data set

$$\ln L(\alpha, \beta) = \ln L_{\max} - 1/2$$

→ Tangent lines to contours give standard deviations.

→ Angle of ellipse  $\phi$  related to correlation: 
$$\tan 2\phi = \frac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma_{\hat{\beta}}^2}$$

Correlations between estimators result in an increase in their standard deviations (statistical errors).

# Extended ML

Sometimes regard  $n$  not as fixed, but as a Poisson r.v., mean  $\nu$ .

Result of experiment defined as:  $n, x_1, \dots, x_n$ .

The (extended) likelihood function is:

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^n f(x_i; \vec{\theta})$$

Suppose theory gives  $\nu = \nu(\boldsymbol{\theta})$ , then the log-likelihood is

$$\ln L(\vec{\theta}) = -\nu(\vec{\theta}) + \sum_{i=1}^n \ln(\nu(\vec{\theta}) f(x_i; \vec{\theta})) + C$$

where  $C$  represents terms not depending on  $\boldsymbol{\theta}$ .

## Extended ML (2)

Example: expected number of events  $\nu(\vec{\theta}) = \sigma(\vec{\theta}) \int L dt$   
where the total cross section  $\sigma(\boldsymbol{\theta})$  is predicted as a function of the parameters of a theory, as is the distribution of a variable  $x$ .

Extended ML uses more info  $\rightarrow$  smaller errors for  $\hat{\vec{\theta}}$

Important e.g. for anomalous couplings in  $e^+e^- \rightarrow W^+W^-$

If  $\nu$  does not depend on  $\boldsymbol{\theta}$  but remains a free parameter, extended ML gives:

$$\hat{\nu} = n$$

$$\hat{\boldsymbol{\theta}} = \text{same as ML}$$



# Extended ML example

Consider two types of events (e.g., signal and background) each of which predict a given pdf for the variable  $x$ :  $f_s(x)$  and  $f_b(x)$ .

We observe a mixture of the two event types, signal fraction =  $\theta$ , expected total number =  $\nu$ , observed total number =  $n$ .

Let  $\mu_s = \theta\nu$ ,  $\mu_b = (1 - \theta)\nu$ , goal is to estimate  $\mu_s, \mu_b$ .

$$f(x; \mu_s, \mu_b) = \frac{\mu_s}{\mu_s + \mu_b} f_s(x) + \frac{\mu_b}{\mu_s + \mu_b} f_b(x)$$

$$P(n; \mu_s, \mu_b) = \frac{(\mu_s + \mu_b)^n}{n!} e^{-(\mu_s + \mu_b)}$$

$$\rightarrow \ln L(\mu_s, \mu_b) = -(\mu_s + \mu_b) + \sum_{i=1}^n \ln [(\mu_s + \mu_b) f(x_i; \mu_s, \mu_b)]$$

# Extended ML example (2)

Monte Carlo example  
with combination of  
exponential and Gaussian:

$$\mu_s = 6$$

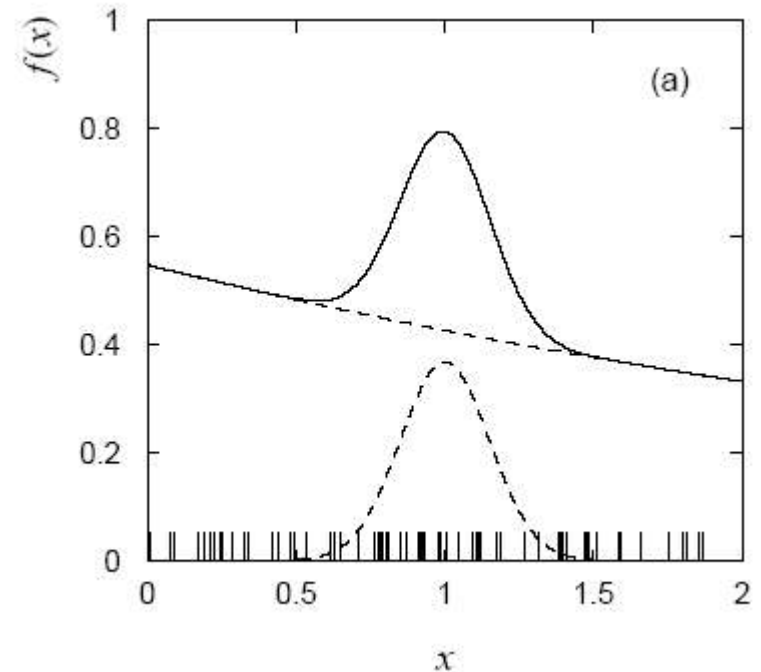
$$\mu_b = 60$$

Maximize log-likelihood in  
terms of  $\mu_s$  and  $\mu_b$ :

$$\hat{\mu}_s = 8.7 \pm 5.5$$

$$\hat{\mu}_b = 54.3 \pm 8.8$$

Here errors reflect total Poisson  
fluctuation as well as that in  
proportion of signal/background.

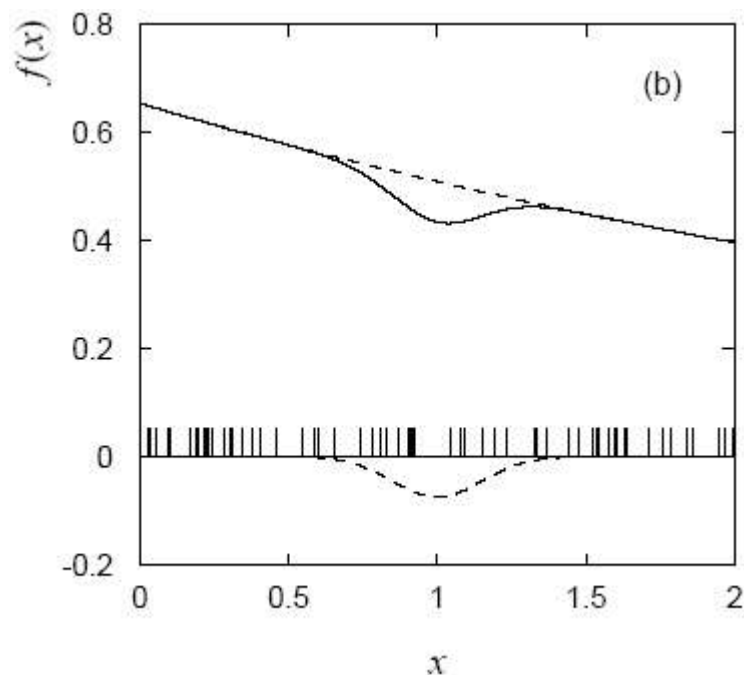


# Extended ML example: an unphysical estimate

A downwards fluctuation of data in the peak region can lead to even fewer events than what would be obtained from background alone.

Estimate for  $\mu_s$  here pushed negative (unphysical).

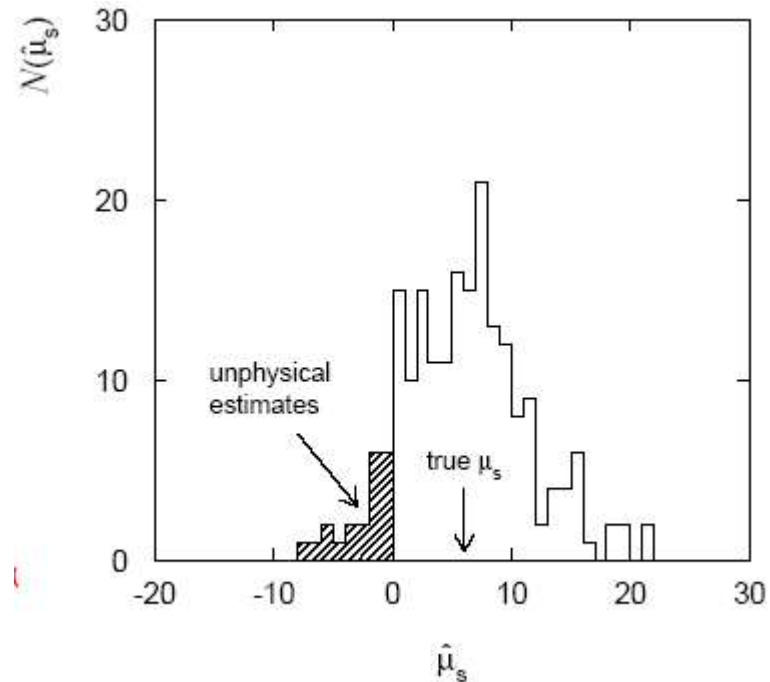
We can let this happen as long as the (total) pdf stays positive everywhere.



## Unphysical estimators (2)

Here the unphysical estimator is unbiased and should nevertheless be reported, since average of a large number of unbiased estimates converges to the true value (cf. PDG).

Repeat entire MC experiment many times, allow unphysical estimates:



# ML with binned data

Often put data into a histogram:  $\vec{n} = (n_1, \dots, n_N)$ ,  $n_{\text{tot}} = \sum_{i=1}^N n_i$

Hypothesis is  $\vec{\nu} = (\nu_1, \dots, \nu_N)$ ,  $\nu_{\text{tot}} = \sum_{i=1}^N \nu_i$  where

$$\nu(\vec{\theta}) = \nu_{\text{tot}} \int_{\text{bin } i} f(x; \vec{\theta}) dx$$

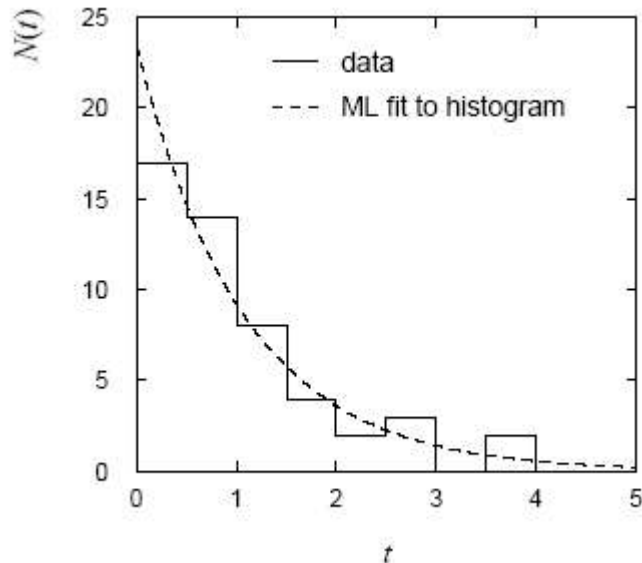
If we model the data as multinomial ( $n_{\text{tot}}$  constant),

$$f(\vec{n}; \vec{\nu}) = \frac{n_{\text{tot}}!}{n_1! \dots n_N!} \left( \frac{\nu_1}{n_{\text{tot}}} \right)^{n_1} \dots \left( \frac{\nu_N}{n_{\text{tot}}} \right)^{n_N}$$

then the log-likelihood function is:  $\ln L(\vec{\theta}) = \sum_{i=1}^N n_i \ln \nu_i(\vec{\theta}) + C$

# ML example with binned data

Previous example with exponential, now put data into histogram:



$$\hat{\tau} = 1.07 \pm 0.17$$

( $1.06 \pm 0.15$  for unbinned

ML with same sample)

Limit of zero bin width  $\rightarrow$  usual unbinned ML.

If  $n_i$  treated as Poisson, we get extended log-likelihood:

$$\ln L(\nu_{\text{tot}}, \vec{\theta}) = -\nu_{\text{tot}} + \sum_{i=1}^N n_i \ln \nu_i(\nu_{\text{tot}}, \vec{\theta}) + C$$

# Relationship between ML and Bayesian estimators

In Bayesian statistics, both  $\theta$  and  $\mathbf{x}$  are random variables:


$$L(\theta) = L(\vec{x}|\theta) = f_{\text{joint}}(\vec{x}|\theta)$$

Recall the Bayesian method:

Use subjective probability for hypotheses ( $\theta$ );

before experiment, knowledge summarized by prior pdf  $\pi(\theta)$ ;

use Bayes' theorem to update prior in light of data:

$$p(\theta|\vec{x}) = \frac{L(\vec{x}|\theta)\pi(\theta)}{\int L(\vec{x}|\theta')\pi(\theta') d\theta'}$$


Posterior pdf (conditional pdf for  $\theta$  given  $\mathbf{x}$ )

## ML and Bayesian estimators (2)

Purist Bayesian:  $p(\theta | x)$  contains all knowledge about  $\theta$ .

Pragmatist Bayesian:  $p(\theta | x)$  could be a complicated function,

→ summarize using an estimator  $\hat{\theta}_{\text{Bayes}}$

Take mode of  $p(\theta | x)$ , (could also use e.g. expectation value)

What do we use for  $\pi(\theta)$ ? No golden rule (subjective!), often represent ‘prior ignorance’ by  $\pi(\theta) = \text{constant}$ , in which case

$$\hat{\theta}_{\text{Bayes}} = \hat{\theta}_{\text{ML}}$$

But... we could have used a different parameter, e.g.,  $\lambda = 1/\theta$ , and if prior  $\pi_{\theta}(\theta)$  is constant, then  $\pi_{\lambda}(\lambda)$  is not!

‘Complete prior ignorance’ is not well defined.



# The method of least squares

Suppose we measure  $N$  values,  $y_1, \dots, y_N$ , assumed to be independent Gaussian r.v.s with

$$E[y_i] = \lambda(x_i; \theta) .$$

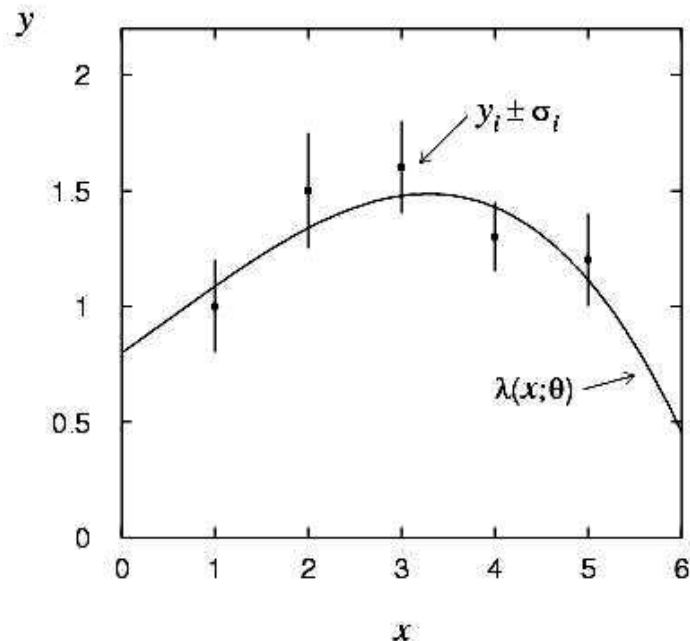
Assume known values of the control variable  $x_1, \dots, x_N$  and known variances

$$V[y_i] = \sigma_i^2 .$$

We want to estimate  $\theta$ , i.e., fit the curve to the data points.

The likelihood function is

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



# The method of least squares (2)

The log-likelihood function is therefore

$$\ln L(\theta) = -\frac{1}{2} \sum_{i=1}^N \frac{(y_i - \lambda(x_i; \theta))^2}{\sigma_i^2} + \text{terms not depending on } \theta$$

So maximizing the likelihood is equivalent to minimizing

$$\chi^2(\theta) = \sum_{i=1}^N \frac{(y_i - \lambda(x_i; \theta))^2}{\sigma_i^2}$$

Minimum defines the least squares (LS) estimator  $\hat{\theta}$ .

Very often measurement errors are  $\sim$ Gaussian and so ML and LS are essentially the same.

Often minimize  $\chi^2$  numerically (e.g. program MINUIT).

# LS with correlated measurements

If the  $y_i$  follow a multivariate Gaussian, covariance matrix  $V$ ,

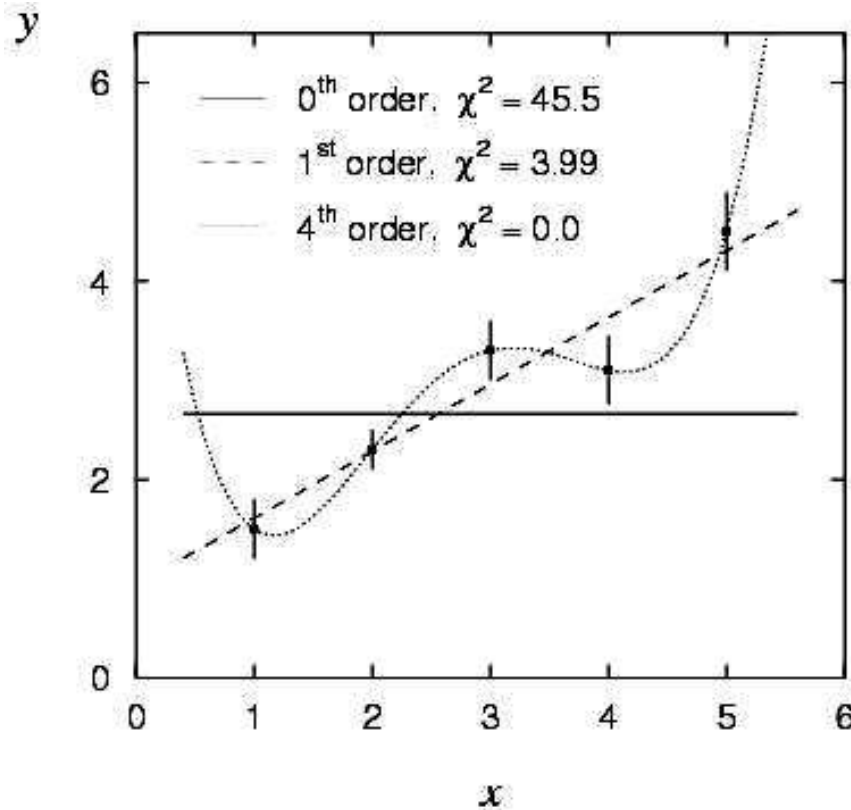
$$g(\vec{y}, \vec{\lambda}, V) = \frac{1}{(2\pi)^{N/2} |V|^{1/2}} \exp \left[ -\frac{1}{2} (\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda}) \right]$$

Then maximizing the likelihood is equivalent to minimizing

$$\chi^2(\vec{\theta}) = \sum_{i,j=1}^N (y_i - \lambda(x_i; \vec{\theta})) (V^{-1})_{ij} (y_j - \lambda(x_j; \vec{\theta}))$$

# Example of least squares fit

Fit a polynomial of order  $p$ :  $\lambda(x; \theta_0, \dots, \theta_p) = \sum_{n=0}^p \theta_n x^n$



# Variance of LS estimators

In most cases of interest we obtain the variance in a manner similar to ML. E.g. for data  $\sim$  Gaussian we have

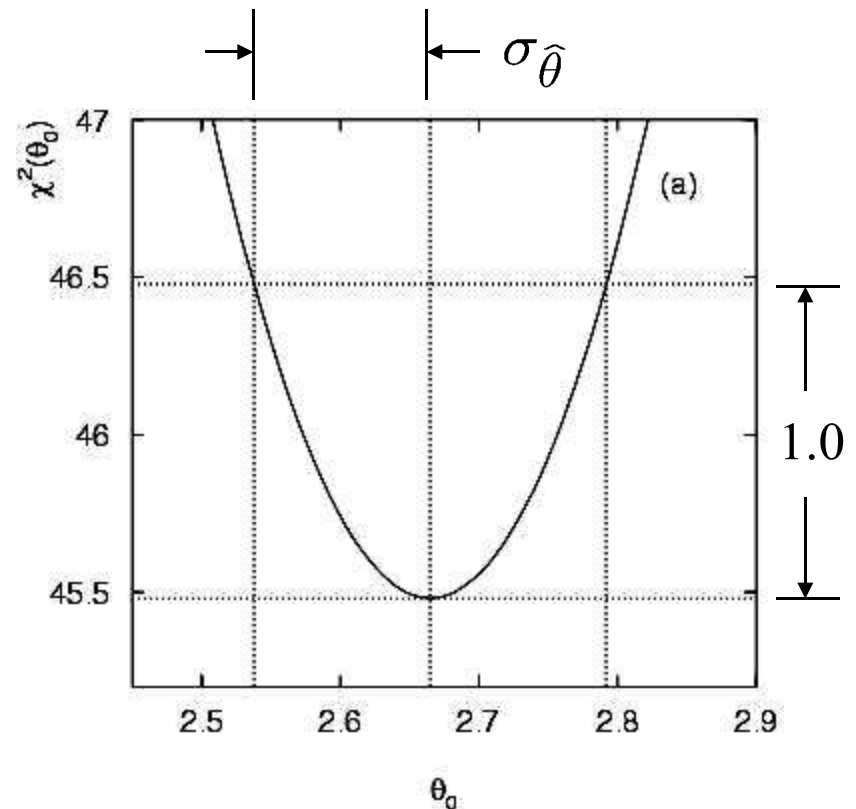
$$\chi^2(\theta) = -2 \ln L(\theta)$$

and so

$$\hat{\sigma}_{\hat{\theta}}^2 \approx 2 \left[ \frac{\partial^2 \chi^2}{\partial \theta^2} \right]_{\theta=\hat{\theta}}^{-1}$$

or for the graphical method we take the values of  $\theta$  where

$$\chi^2(\theta) = \chi_{\min}^2 + 1$$



# Two-parameter LS fit

2-parameter case (line with nonzero slope):

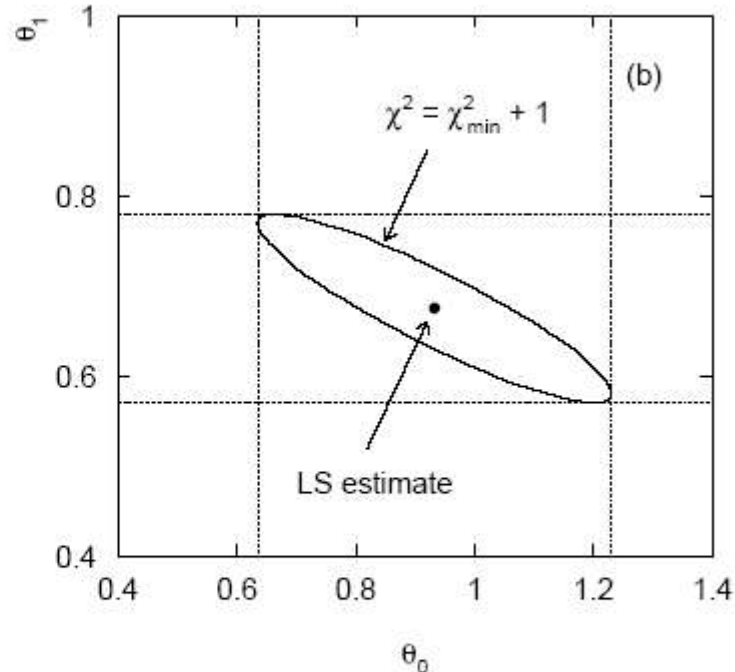
$$\hat{\theta}_0 = 0.93 \pm 0.30,$$

$$\hat{\theta}_1 = 0.68 \pm 0.10$$

$$\widehat{\text{cov}}[\hat{\theta}_0, \hat{\theta}_1] = -0.028$$

$$r = -0.90$$

$$\chi^2 = 3.99$$



Tangent lines  $\rightarrow \sigma_{\hat{\theta}_0}, \sigma_{\hat{\theta}_1}$ .

Angle of ellipse  $\rightarrow$  correlation (same as for ML)

# Goodness-of-fit with least squares

The value of the  $\chi^2$  at its minimum is a measure of the level of agreement between the data and fitted curve:

$$\chi_{\min}^2 = \sum_{i=1}^N \frac{(y_i - \lambda(x_i; \hat{\theta}))^2}{\sigma_i^2}$$

It can therefore be employed as a goodness-of-fit statistic to test the hypothesized functional form  $\lambda(x; \theta)$ .

We can show that if the hypothesis is correct, then the statistic  $t = \chi_{\min}^2$  follows the chi-square pdf,

$$f(t; n_d) = \frac{1}{2^{n_d/2} \Gamma(n_d/2)} t^{n_d/2-1} e^{-t/2}$$

where the number of degrees of freedom is

$$n_d = \text{number of data points} - \text{number of fitted parameters}$$

## Goodness-of-fit with least squares (2)

The chi-square pdf has an expectation value equal to the number of degrees of freedom, so if  $\chi^2_{\min} \approx n_d$  the fit is ‘good’.

More generally, find the  $p$ -value: 
$$p = \int_{\chi^2_{\min}}^{\infty} f(t; n_d) dt$$

This is the probability of obtaining a  $\chi^2_{\min}$  as high as the one we got, or higher, if the hypothesis is correct.

E.g. for the previous example with 1st order polynomial (line),

$$\chi^2_{\min} = 3.99, \quad n_d = 5 - 2 = 3, \quad p = 0.263$$

whereas for the 0th order polynomial (horizontal line),

$$\chi^2_{\min} = 45.5, \quad n_d = 5 - 1 = 4, \quad p = 3.1 \times 10^{-9}$$



# Wrapping up lecture 3

No golden rule for parameter estimation, construct so as to have desirable properties (small variance, small or zero bias, ...)

Most important methods:

- Maximum Likelihood,
- Least Squares

Several methods to obtain variances (stat. errors) from a fit

- Analytically

- Monte Carlo

- From information equality / graphical method

Finding estimator often involves numerical minimization

# Extra slides for lecture 3

Goodness-of-fit vs. statistical errors

Fitting histograms with LS

Combining measurements with LS

# Goodness-of-fit vs. statistical errors

Small statistical error does not mean a good fit (nor vice versa).

Curvature of  $\chi^2$  near its minimum  $\rightarrow$  statistical errors ( $\sigma_{\hat{\theta}}$ )

Value of  $\chi^2_{\min}$   $\rightarrow$  goodness-of-fit

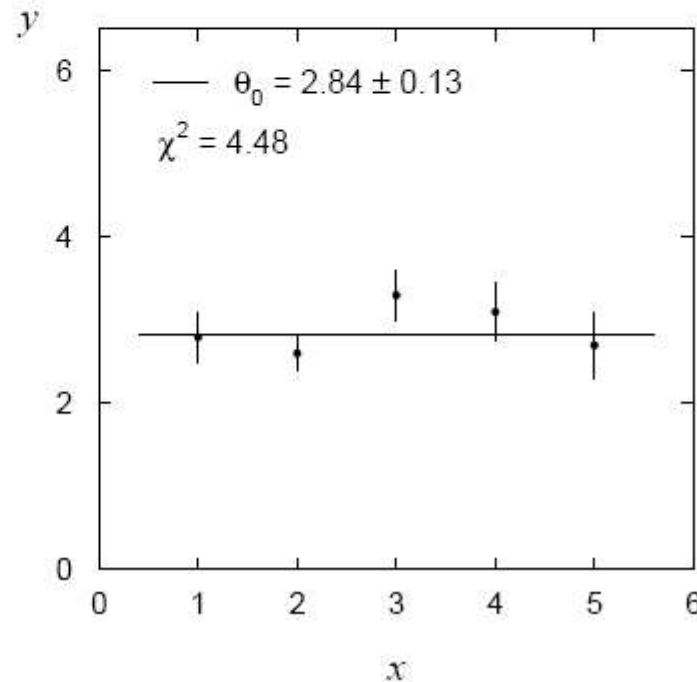
Horizontal line fit, move the data points, keep errors on points same:

$$\hat{\theta}_0 = 2.84 \pm 0.13$$

$$\chi^2_{\min} = 4.48$$

Variance same as before,

now  $\chi^2_{\min}$  'good'.



## Goodness-of-fit vs. stat. errors (2)

→  $\chi^2(\theta_0)$  shifted down, same curvature as before.

Variance of estimator (statistical error) tells us:

if experiment repeated many times, how wide is the distribution of the estimates  $\hat{\theta}$ . (Doesn't tell us whether hypothesis correct.)

$P$ -value tells us:

if hypothesis is correct and experiment repeated many times, what fraction will give equal or worse agreement between data and hypothesis according to the statistic  $\chi_{\min}^2$ .

Low  $P$ -value → hypothesis may be wrong → **systematic error**.

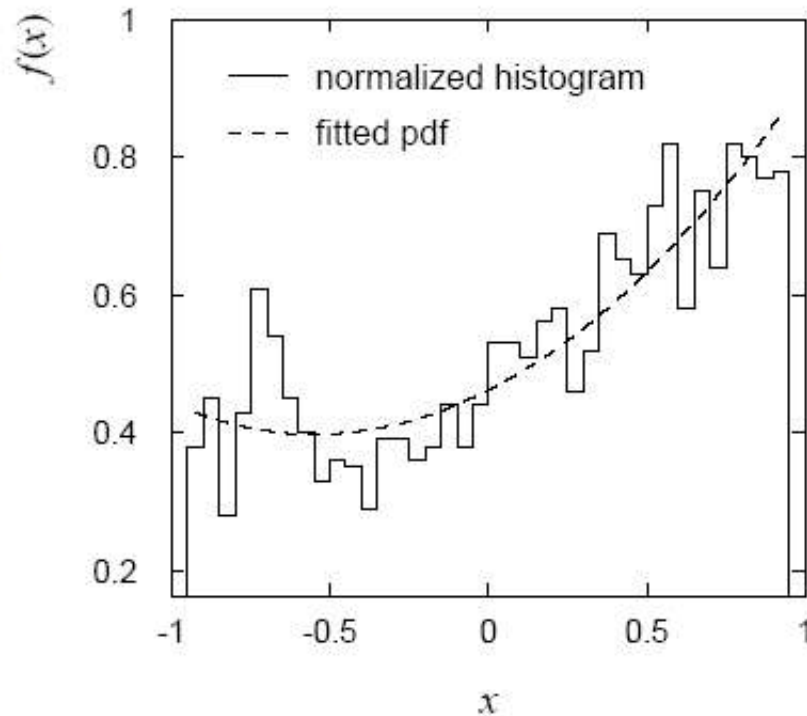
# LS with binned data

Histogram:

$N$  bins,  $n$  entries.

Hypothesized pdf:

$$f(x; \vec{\theta})$$



We have

$y_i$  = number of entries in bin  $i$ ,

$$\lambda_i(\vec{\theta}) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = np_i(\vec{\theta})$$

## LS with binned data (2)

LS fit: minimize

$$\chi^2(\vec{\theta}) = \sum_{i=1}^N \frac{(y_i - \lambda_i(\vec{\theta}))^2}{\sigma_i^2}$$

where  $\sigma_i^2 = V[y_i]$ , here not known a priori.

Treat the  $y_i$  as Poisson r.v.s, in place of true variance take either

$$\sigma_i^2 = \lambda_i(\vec{\theta}) \quad (\text{LS method})$$

$$\sigma_i^2 = y_i \quad (\text{Modified LS method})$$

MLS sometimes easier computationally, but  $\chi_{\min}^2$  no longer follows chi-square pdf (or is undefined) if some bins have few (or no) entries.

# LS with binned data — normalization

Do **not** ‘fit the normalization’:

$$\lambda_i(\vec{\theta}, \nu) = \nu \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = \nu p_i(\vec{\theta})$$

i.e. introduce adjustable  $\nu$ , fit along with  $\vec{\theta}$ .

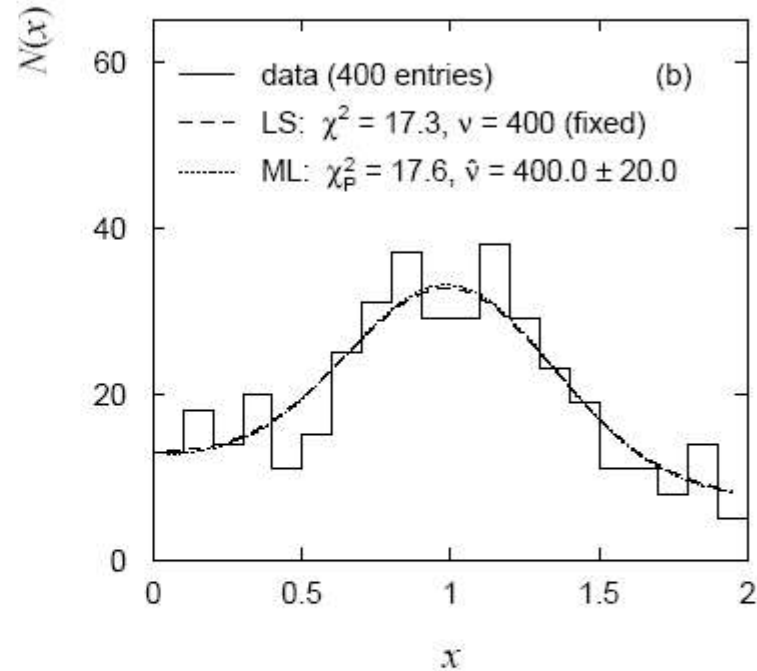
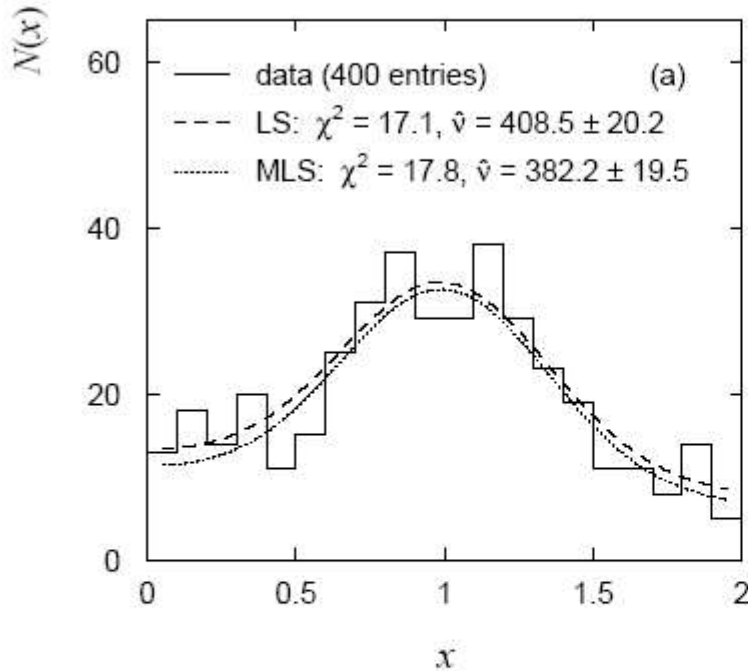
$\hat{\nu}$  is a bad estimator for  $n$  (which we know, anyway!)

$$\hat{\nu}_{\text{LS}} = n + \frac{\chi_{\min}^2}{2}$$

$$\hat{\nu}_{\text{MLS}} = n - \chi_{\min}^2$$

# LS normalization example

Example with  $n = 400$  entries,  $N = 20$  bins:



Expect  $\chi_{\min}^2$  around  $N - m$ ,

→ relative error in  $\hat{v}$  large when  $N$  large,  $n$  small

Either get  $n$  directly from data for LS (or better, use ML).



# Using LS to combine measurements

Use LS to obtain weighted average of  $N$  measurements of  $\lambda$ :

$y_i$  = result of measurement  $i$ ,  $i = 1, \dots, N$ ;

$\sigma_i^2 = V[y_i]$ , assume known;

$\lambda$  = true value (plays role of  $\theta$ ).

For uncorrelated  $y_i$ , minimize

$$\chi^2(\lambda) = \sum_{i=1}^N \frac{(y_i - \lambda)^2}{\sigma_i^2},$$

Set  $\frac{\partial \chi^2}{\partial \lambda} = 0$  and solve,

$$\rightarrow \hat{\lambda} = \frac{\sum_{i=1}^N y_i / \sigma_i^2}{\sum_{j=1}^N 1 / \sigma_j^2} \quad V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^N 1 / \sigma_i^2}$$

# Combining correlated measurements with LS

If  $\text{cov}[y_i, y_j] = V_{ij}$ , minimize

$$\chi^2(\lambda) = \sum_{i,j=1}^N (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda),$$

$$\rightarrow \hat{\lambda} = \sum_{i=1}^N w_i y_i, \quad w_i = \frac{\sum_{j=1}^N (V^{-1})_{ij}}{\sum_{k,l=1}^N (V^{-1})_{kl}}$$

$$V[\hat{\lambda}] = \sum_{i,j=1}^N w_i V_{ij} w_j$$

LS  $\hat{\lambda}$  has zero bias, minimum variance (Gauss–Markov theorem).

# Example: averaging two correlated measurements

Suppose we have  $y_1, y_2$ , and  $V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

$$\rightarrow \hat{\lambda} = wy_1 + (1-w)y_2, \quad w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$V[\hat{\lambda}] = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \sigma^2$$

The increase in inverse variance due to 2nd measurement is

$$\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} = \frac{1}{1-\rho^2} \left( \frac{\rho}{\sigma_1} - \frac{1}{\sigma_2} \right)^2 > 0$$

$\rightarrow$  2nd measurement can only help.

# Negative weights in LS average

If  $\rho > \sigma_1/\sigma_2$ ,  $\rightarrow w < 0$ ,

$\rightarrow$  weighted average is not between  $y_1$  and  $y_2$  (!?)

Cannot happen if correlation due to common data, but possible for shared random effect; very unreliable if e.g.

$\rho$ ,  $\sigma_1$ ,  $\sigma_2$  incorrect.

See example in SDA Section 7.6.1 with two measurements at same temperature using two rulers, different thermal expansion coefficients:

average is outside the two measurements; used to improve estimate of temperature.