

# Statistical Methods for Particle Physics

## Lecture 3: parameter estimation

[www.pp.rhul.ac.uk/~cowan/stat\\_aachen.html](http://www.pp.rhul.ac.uk/~cowan/stat_aachen.html)



Graduierten-Kolleg  
RWTH Aachen  
10-14 February 2014



Glen Cowan  
Physics Department  
Royal Holloway, University of London  
[g.cowan@rhul.ac.uk](mailto:g.cowan@rhul.ac.uk)  
[www.pp.rhul.ac.uk/~cowan](http://www.pp.rhul.ac.uk/~cowan)

# Outline

## 1 Probability

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

## 2 Statistical tests

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

## → 3 Parameter estimation

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

## 4 Hypothesis tests for discovery and exclusion

discovery significance, sensitivity, setting limits

## 5 Further topics

systematic errors, Bayesian methods, MCMC

# Frequentist parameter estimation

Suppose we have a pdf characterized by one or more parameters:

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

random variable

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):

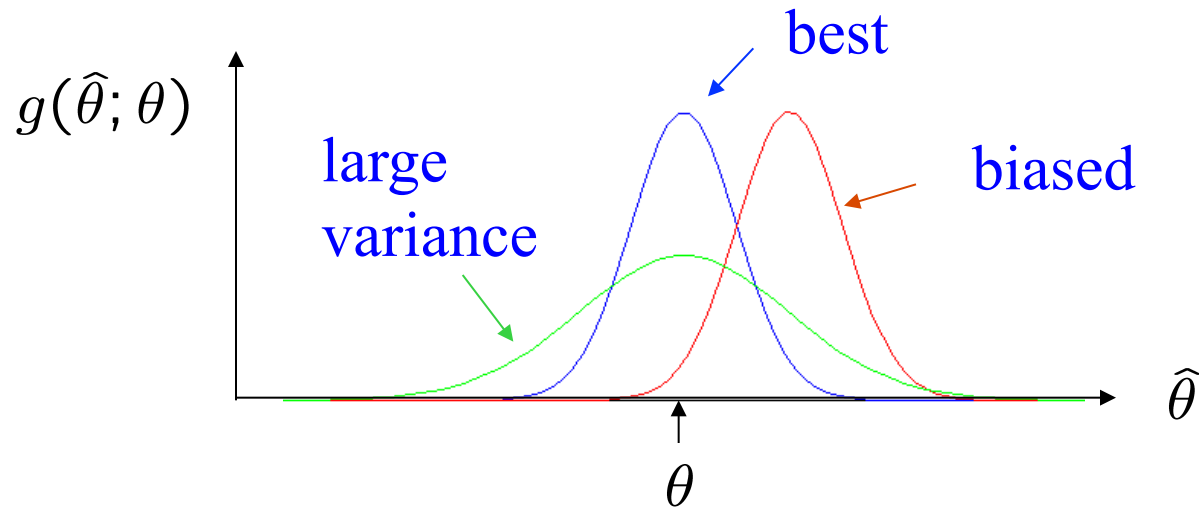
$$\hat{\theta}(\vec{x})$$

← 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 measurements should tend to true value.

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

→ small bias & variance are in general conflicting criteria

# Distribution, likelihood, model

Suppose the outcome of a measurement is  $x$ . (e.g., a number of events, a histogram, or some larger set of numbers).

The probability density (or mass) function or ‘distribution’ of  $x$ , which may depend on parameters  $\theta$ , is:

$$P(x|\theta) \quad (\text{Independent variable is } x; \theta \text{ is a constant.})$$

If we evaluate  $P(x|\theta)$  with the observed data and regard it as a function of the parameter(s), then this is the **likelihood**:

$$L(\theta) = P(x|\theta) \quad (\text{Data } x \text{ fixed; treat } L \text{ as function of } \theta.)$$

We will use the term ‘**model**’ to refer to the full function  $P(x|\theta)$  that contains the dependence both on  $x$  and  $\theta$ .

# Bayesian use of the term ‘likelihood’

We can write Bayes theorem as

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

where  $L(x|\theta)$  is the likelihood. It is the probability for  $x$  given  $\theta$ , evaluated with the observed  $x$ , and viewed as a function of  $\theta$ .

Bayes’ theorem only needs  $L(x|\theta)$  evaluated with a given data set (the ‘likelihood principle’).

For frequentist methods, in general one needs the full model.

For some approximate frequentist methods, the likelihood is enough.

# The likelihood function for i.i.d.\*. data

\* i.i.d. = independent and identically distributed

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:

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

In this case the likelihood function is

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

# Maximum likelihood

The most important frequentist method for constructing estimators is to take the value of the parameter(s) that maximize the likelihood:  $\hat{\theta} = \operatorname{argmax}_{\theta} L(x|\theta)$

The resulting estimators are functions of the data and thus characterized by a sampling distribution with a given (co)variance:  $V_{ij} = \operatorname{cov}[\hat{\theta}_i, \hat{\theta}_j]$

In general they may have a nonzero bias:  $b = E[\hat{\theta}] - \theta$

Under conditions usually satisfied in practice, bias of ML estimators is zero in the large sample limit, and the variance is as small as possible for unbiased estimators.

ML estimator may not in some cases be regarded as the optimal trade-off between these criteria (cf. regularized unfolding).



# ML example: parameter of exponential pdf

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

and suppose we have i.i.d. 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 by setting  $\frac{\partial \ln L(\tau)}{\partial \tau} = 0$ ,

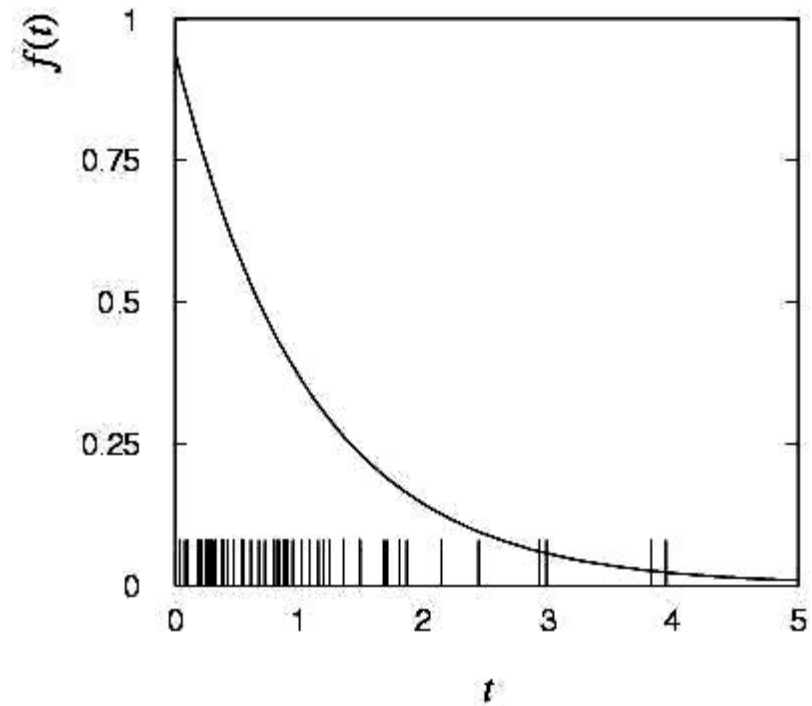
$$\rightarrow \hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$$

Monte Carlo test:

generate 50 values  
using  $\tau = 1$ :

We find the ML estimate:

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



# Variance of estimators: Monte Carlo method

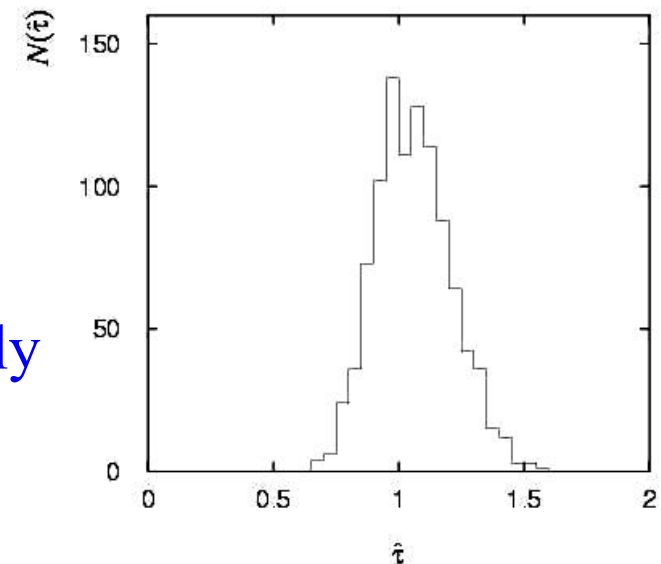
Having estimated our parameter we now need to report its ‘statistical error’, i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

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 lower bound on the variance of any estimator (not only ML):

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

 Minimum Variance Bound (MVB)  
( $b = E[\hat{\theta}] - \theta$ )

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

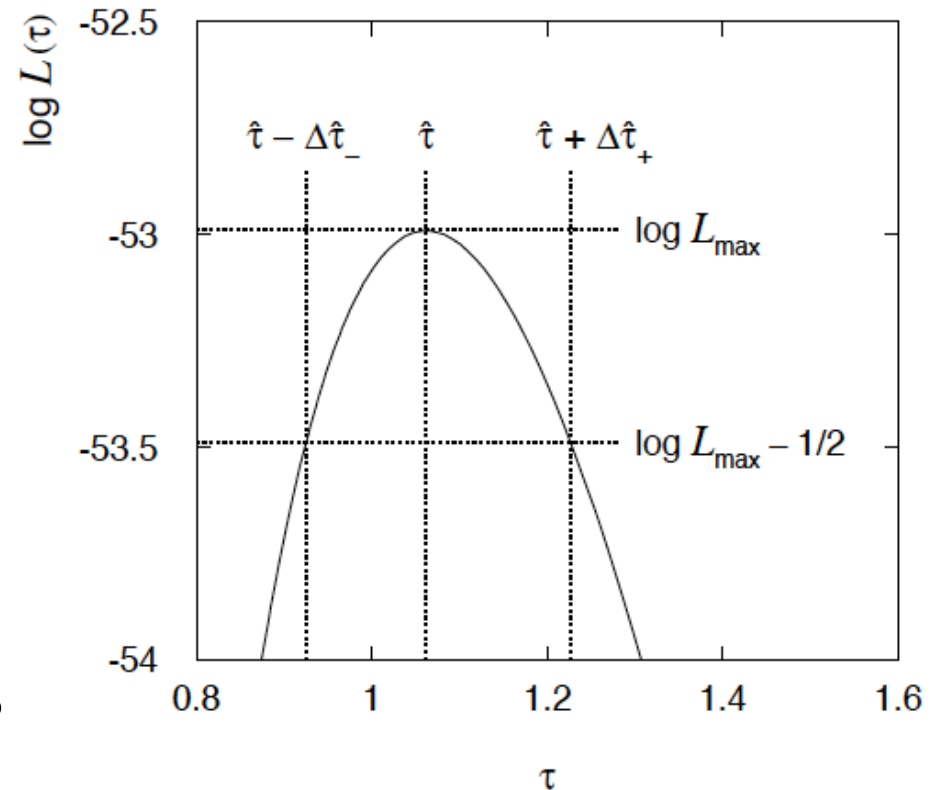
ML example with exponential:

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

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

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

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



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

# 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  $\lambda$ ?

Rewrite the likelihood replacing  $\tau$  by  $1/\lambda$ . The  $\lambda$  that maximizes  $L(\lambda)$  is the  $\lambda$  that corresponds to the  $\tau$  that maximizes  $L(\tau)$ , i.e.,

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

# 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, where  $V_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$ . Therefore

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

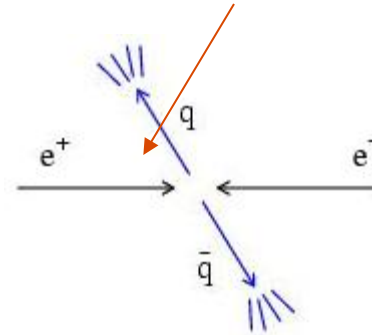
Often use  $I^{-1}$  as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of  $L$ .



# Two-parameter example of ML

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

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



Data:  $x_1, \dots, x_n$ ,  $n = 2000$  events.

As test generate with MC using  $\alpha = 0.5$ ,  $\beta = 0.5$

From data compute log-likelihood:

$$\ln L(\alpha, \beta) = \sum_{i=1}^n \ln f(x_i; \alpha, \beta)$$

Maximize numerically (e.g., program MINUIT)

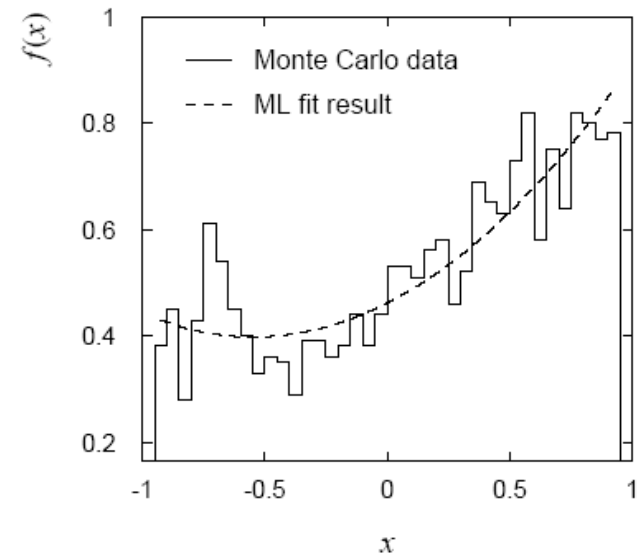
# Example of ML: fit result

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

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

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

**N.B.** Here 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**)

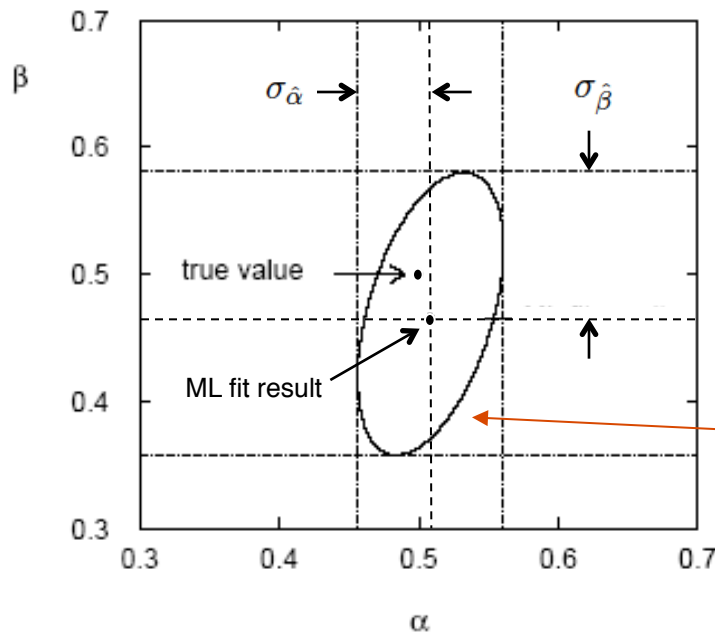
$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \quad \text{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$

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

# Variance of ML estimators: graphical method

Often (e.g., large sample case) one can approximate the covariances using only the likelihood  $L(\theta)$ :

$$\hat{V}_{ij}^{-1} \approx - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\theta=\hat{\theta}}$$



This translates into a simple graphical recipe:

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

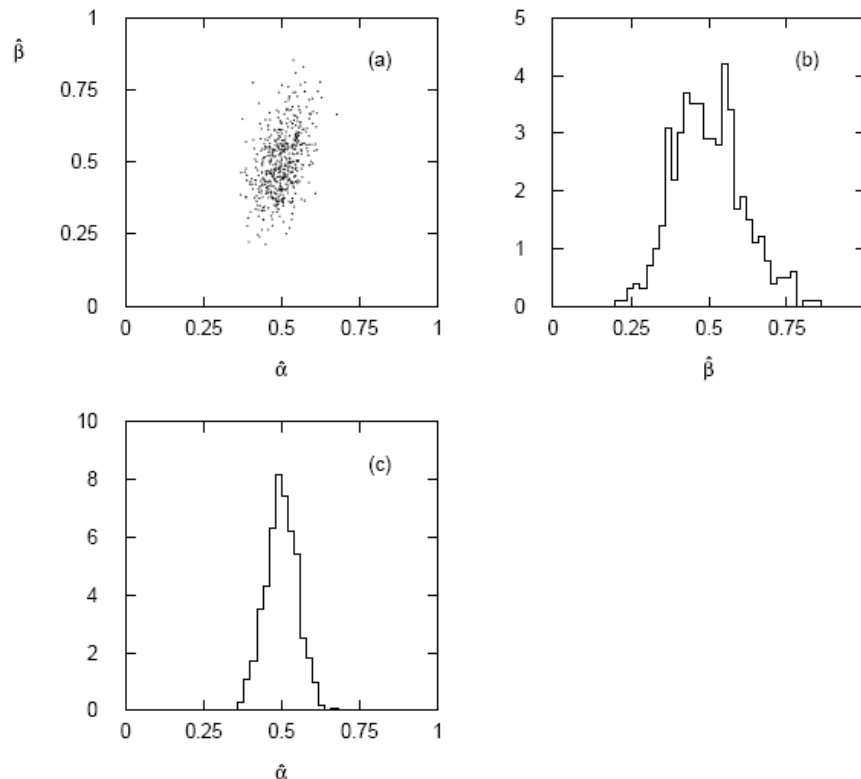
→ Tangent lines to contours give standard deviations.

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

# Variance of ML estimators: MC

To find the ML estimate itself one only needs the likelihood  $L(\theta)$ .

In principle to find the covariance of the estimators, one requires the full model  $L(x|\theta)$ . E.g., simulate many times independent data sets and look at distribution of the resulting estimates:



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

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

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

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

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

$$r = 0.42$$

## 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  $\nu = \nu(\theta)$ , 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  $\theta$ .

## Extended ML (2)

Example: expected number of events  $\nu(\vec{\theta}) = \sigma(\vec{\theta}) \int L dt$   
where the total cross section  $\sigma(\theta)$  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  $\theta$  but remains a free parameter, extended ML gives:

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

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

# 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:

$$\mu_s = 6$$

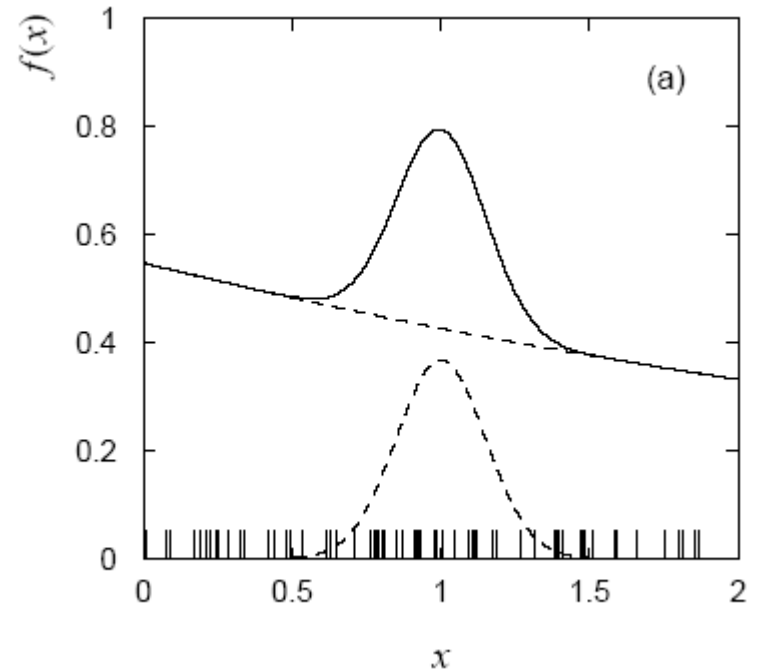
$$\mu_b = 60$$

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.





# 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

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

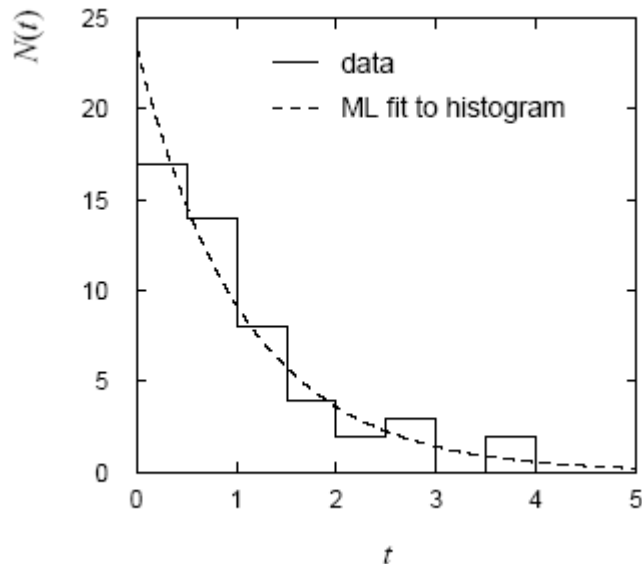
If we model the data as multinomial ( $n_{\text{tot}}$  constant),

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

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

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

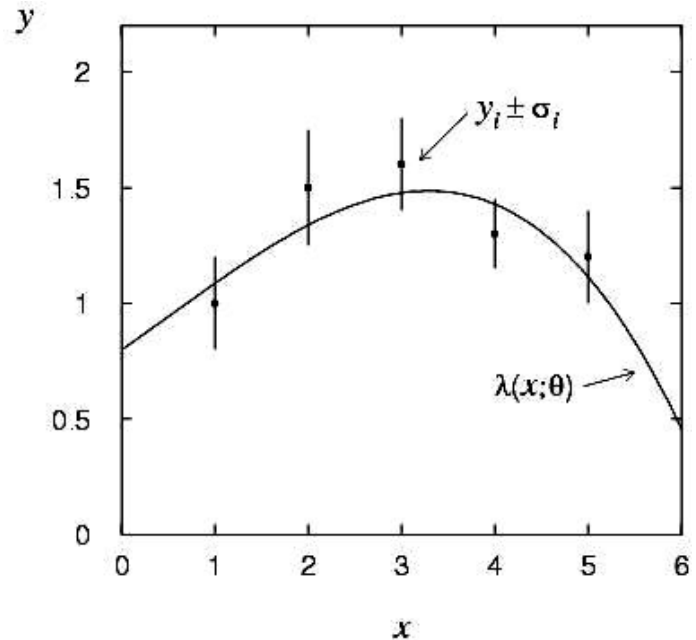
Assume known values of the control variable  $x_1, \dots, x_N$  and known variances

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

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  $V$ ,

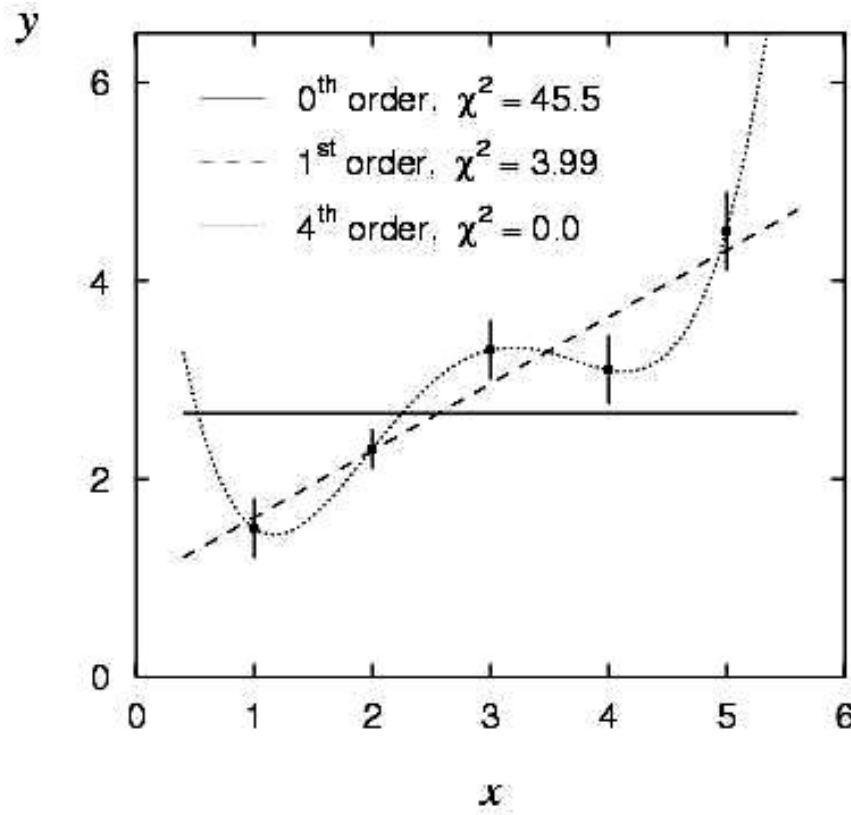
$$g(\vec{y}, \vec{\lambda}, V) = \frac{1}{(2\pi)^{N/2} |V|^{1/2}} \exp \left[ -\frac{1}{2} (\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda}) \right]$$

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

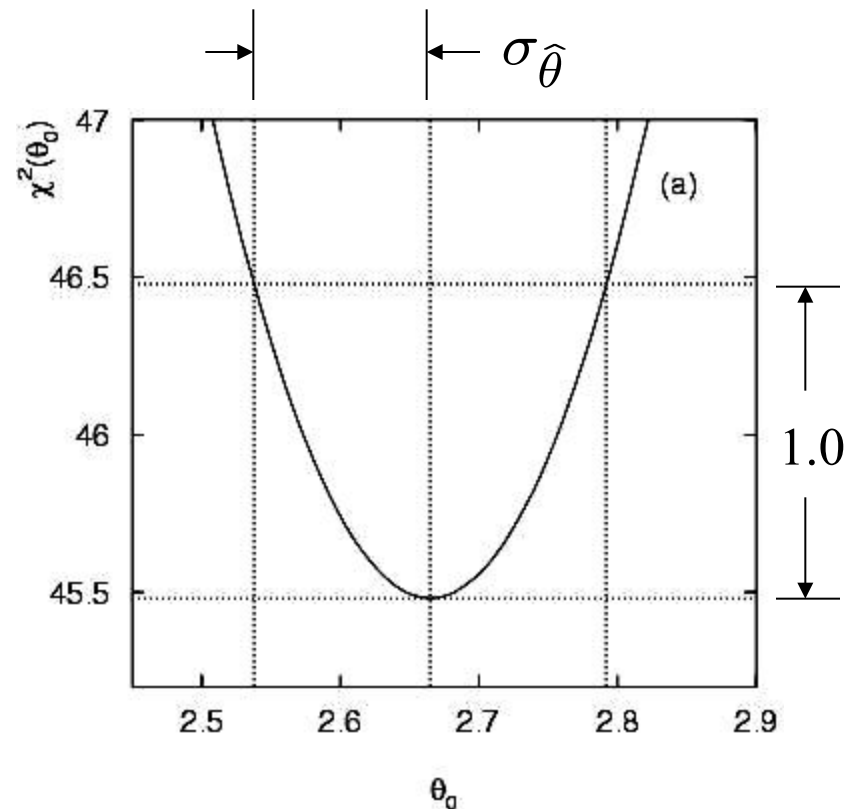
$$\chi^2(\theta) = -2 \ln L(\theta)$$

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):

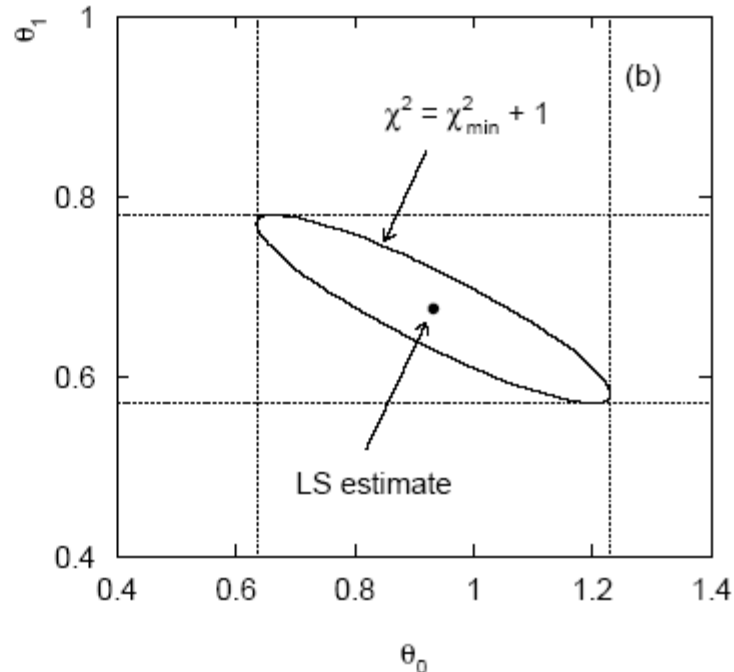
$$\hat{\theta}_0 = 0.93 \pm 0.30,$$

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

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

$$r = -0.90$$

$$\chi^2 = 3.99$$



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

$n_d$  = number of data points – 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),

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

# 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

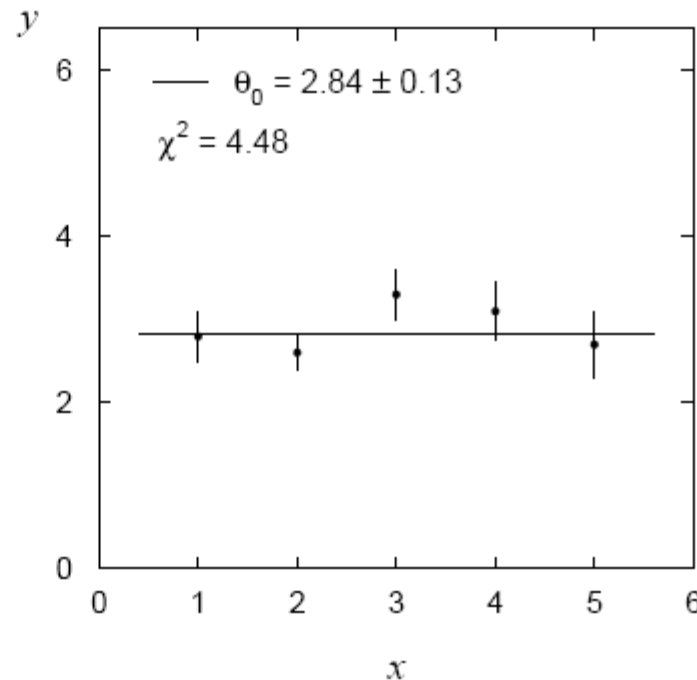
Horizontal line fit, move the data points, keep errors on points same:

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

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

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  $\chi^2_{\min}$ .

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

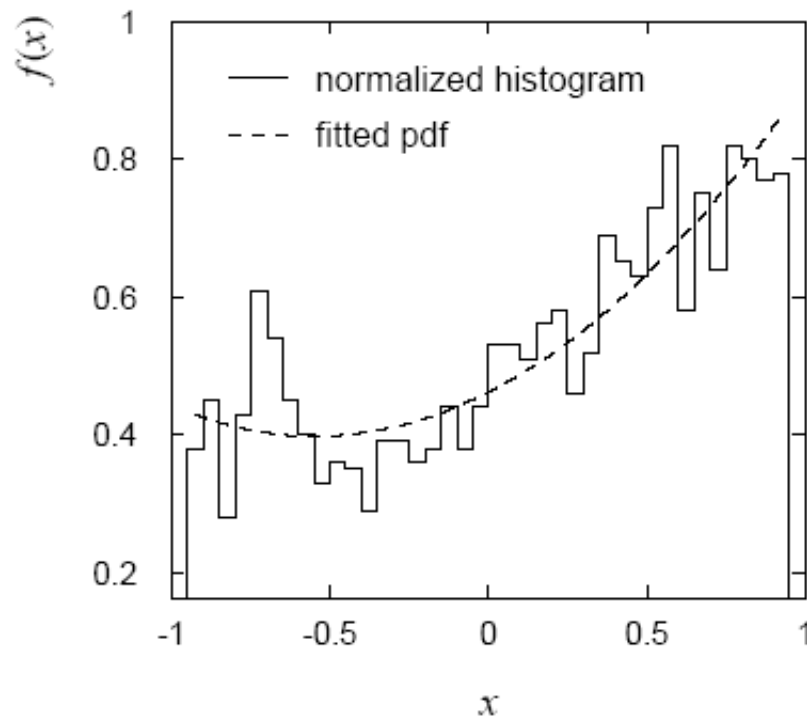
# LS with binned data

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  $\vec{\theta}$ .

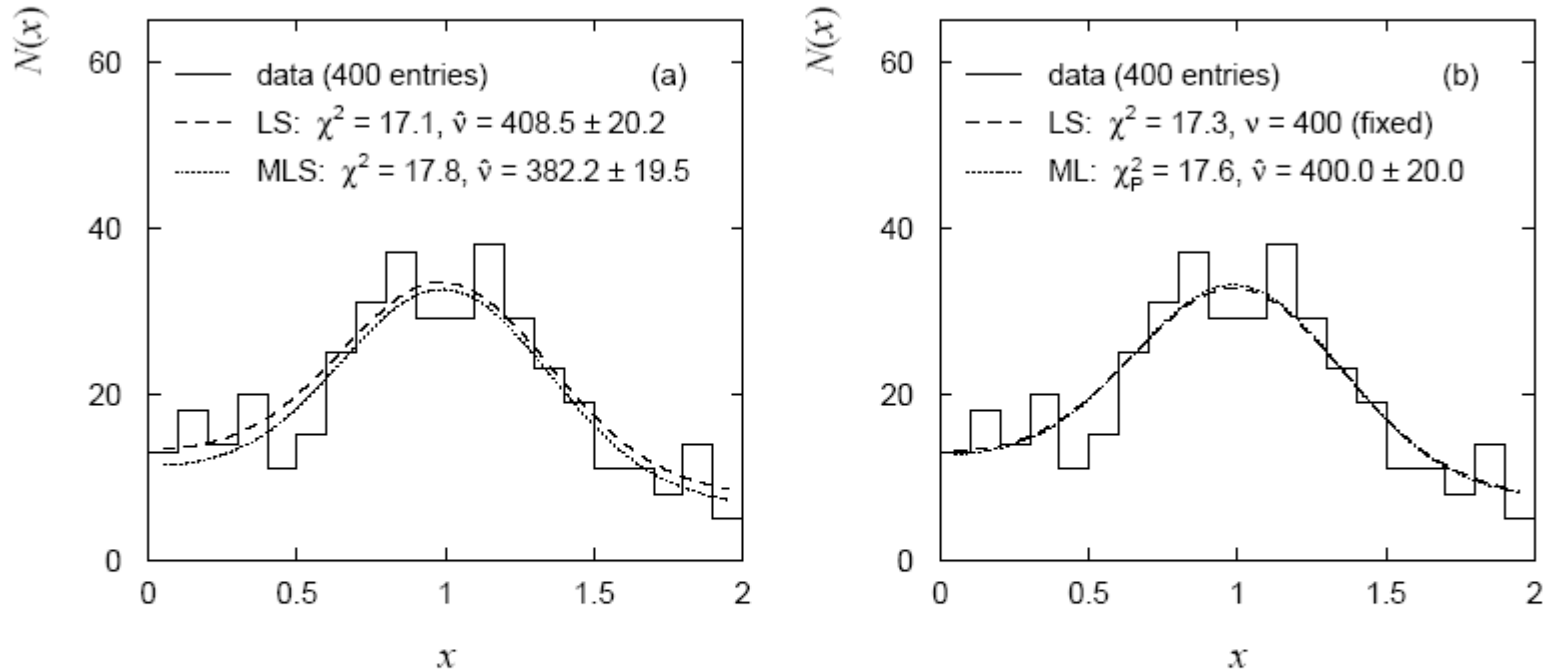
$\hat{\nu}$  is a bad estimator for  $n$  (which we know, anyway!)

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

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

# LS normalization example

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



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

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

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

# Goodness of fit from the likelihood ratio

Suppose we model data using a likelihood  $L(\boldsymbol{\mu})$  that depends on  $N$  parameters  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$ . Define the statistic

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})}$$

Value of  $t_{\boldsymbol{\mu}}$  reflects agreement between hypothesized  $\boldsymbol{\mu}$  and the data.

Good agreement means  $\hat{\boldsymbol{\mu}} \approx \boldsymbol{\mu}$ , so  $t_{\boldsymbol{\mu}}$  is small;

Larger  $t_{\boldsymbol{\mu}}$  means less compatibility between data and  $\boldsymbol{\mu}$ .

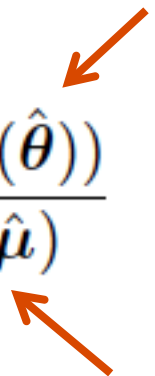
Quantify “goodness of fit” with  $p$ -value:  $p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\text{obs}}}^{\infty} f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}) dt_{\boldsymbol{\mu}}$

## Likelihood ratio (2)

Now suppose the parameters  $\mu = (\mu_1, \dots, \mu_N)$  can be determined by another set of parameters  $\theta = (\theta_1, \dots, \theta_M)$ , with  $M < N$ .

E.g. in LS fit, use  $\mu_i = \mu(x_i; \theta)$  where  $x$  is a control variable.

Define the statistic

$$q_\mu = -2 \ln \frac{L(\mu(\hat{\theta}))}{L(\hat{\mu})}$$


fit  $M$  parameters

fit  $N$  parameters

Use  $q_\mu$  to test hypothesized functional form of  $\mu(x; \theta)$ .

To get  $p$ -value, need pdf  $f(q_\mu|\mu)$ .

# Wilks' Theorem (1938)

Wilks' Theorem: if the hypothesized parameters  $\mu = (\mu_1, \dots, \mu_N)$  are true then in the large sample limit (and provided certain conditions are satisfied)  $t_\mu$  and  $q_\mu$  follow chi-square distributions.

For case with  $\mu = (\mu_1, \dots, \mu_N)$  fixed in numerator:

$$t_\mu = -2 \ln \frac{L(\mu)}{L(\hat{\mu})}$$

$$f(t_\mu | \mu) \sim \chi_N^2$$

Or if  $M$  parameters adjusted in numerator,

$$q_\mu = -2 \ln \frac{L(\mu(\hat{\theta}))}{L(\hat{\mu})}$$

$$f(q_\mu | \mu) \sim \chi_{N-M}^2$$

degrees of  
freedom



S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

# Goodness of fit with Gaussian data

Suppose the data are  $N$  independent Gaussian distributed values:

$$y_i \sim \text{Gauss}(\mu_i, \sigma_i), \quad i = 1, \dots, N$$

want to estimate  known

Likelihood: 
$$L(\mu) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu_i)^2 / 2\sigma_i^2}$$

Log-likelihood: 
$$\ln L(\mu) = -\frac{1}{2} \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\sigma_i^2} + C$$

ML estimators: 
$$\hat{\mu}_i = y_i \quad i = 1, \dots, N$$

# Likelihood ratios for Gaussian data

The goodness-of-fit statistics become

$$t_{\mu} = -2 \ln \frac{L(\mu)}{L(\hat{\mu})} = \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\sigma_i^2} \quad f(t_{\mu} | \mu) \sim \chi_N^2$$

$$q_{\mu} = -2 \ln \frac{L(\mu(\hat{\theta}))}{L(\hat{\mu})} = \sum_{i=1}^N \frac{(y_i - \mu_i(\hat{\theta}))^2}{\sigma_i^2} \quad f(q_{\mu} | \mu) \sim \chi_{N-M}^2$$

So Wilks' theorem formally states the well-known property of the minimized chi-squared from an LS fit.

# Likelihood ratio for Poisson data

Suppose the data are a set of values  $\mathbf{n} = (n_1, \dots, n_N)$ , e.g., the numbers of events in a histogram with  $N$  bins.

Assume  $n_i \sim \text{Poisson}(\nu_i)$ ,  $i = 1, \dots, N$ , all independent.

Goal is to estimate  $\boldsymbol{\nu} = (\nu_1, \dots, \nu_N)$ .

Likelihood: 
$$L(\boldsymbol{\nu}) = \prod_{i=1}^N \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

Log-likelihood: 
$$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^N [n_i \ln \nu_i - \nu_i] + C$$

ML estimators: 
$$\hat{\nu}_i = n_i, \quad i = 1, \dots, N$$



# Goodness of fit with Poisson data

The likelihood ratio statistic (all parameters fixed in numerator):

$$\begin{aligned}t_{\boldsymbol{\nu}} &= -2 \ln \frac{L(\boldsymbol{\nu})}{L(\hat{\boldsymbol{\nu}})} \\&= -2 \sum_{i=1}^N \left[ n_i \ln \frac{\nu_i}{\hat{\nu}_i} - \nu_i + \hat{\nu}_i \right] \\&= -2 \sum_{i=1}^N \left[ n_i \ln \frac{\nu_i}{n_i} - \nu_i + n_i \right]\end{aligned}$$

Wilks' theorem:  $f(t_{\boldsymbol{\nu}}|\boldsymbol{\nu}) \sim \chi_N^2$

## Goodness of fit with Poisson data (2)

Or with  $M$  fitted parameters in numerator:

$$q_{\nu} = -2 \ln \frac{L(\nu(\hat{\theta}))}{L(\hat{\nu})} = -2 \sum_{i=1}^N \left[ n_i \ln \frac{\nu_i(\hat{\theta})}{n_i} - \nu_i(\hat{\theta}) + n_i \right]$$

Wilks' theorem:  $f(q_{\nu}|\nu) \sim \chi_{N-M}^2$

Use  $t_{\mu}$ ,  $q_{\mu}$  to quantify goodness of fit ( $p$ -value).

Sampling distribution from Wilks' theorem (chi-square).

Exact in large sample limit; in practice good approximation for surprisingly small  $n_i$  ( $\sim$ several).

# Goodness of fit with multinomial data

Similar if data  $\mathbf{n} = (n_1, \dots, n_N)$  follow multinomial distribution:

$$P(\mathbf{n}|\mathbf{p}, n_{\text{tot}}) = \frac{n_{\text{tot}}!}{n_1! n_2! \dots n_N!} p_1^{n_1} p_2^{n_2} \dots p_N^{n_N}$$

E.g. histogram with  $N$  bins but fix:  $n_{\text{tot}} = \sum_{i=1}^N n_i$

$$\text{Log-likelihood: } \ln L(\boldsymbol{\nu}) = \sum_{i=1}^N n_i \ln \frac{\nu_i}{n_{\text{tot}}} + C \quad (\nu_i = p_i n_{\text{tot}})$$

ML estimators:  $\hat{\nu}_i = n_i$  (Only  $N-1$  independent; one is  $n_{\text{tot}}$  minus sum of rest.)

# Goodness of fit with multinomial data (2)

The likelihood ratio statistics become:

$$t_{\nu} = -2 \sum_{i=1}^N n_i \ln \frac{\nu_i}{n_i} \qquad f(t_{\nu} | \nu) \sim \chi_{N-1}^2$$
$$q_{\nu} = -2 \sum_{i=1}^N n_i \ln \frac{\nu_i(\hat{\theta})}{n_i} \qquad f(q_{\nu} | \nu) \sim \chi_{N-M-1}^2$$

One less degree of freedom than in Poisson case because effectively only  $N-1$  parameters fitted in denominator.

# Estimators and g.o.f. all at once

Evaluate numerators with  $\theta$  (not its estimator):

$$\chi_P^2(\theta) = -2 \sum_{i=1}^N \left[ n_i \ln \frac{\nu_i(\theta)}{n_i} - \nu_i(\theta) + n_i \right] \quad (\text{Poisson})$$

$$\chi_M^2(\theta) = -2 \sum_{i=1}^N n_i \ln \frac{\nu_i(\theta)}{n_i} \quad (\text{Multinomial})$$

These are equal to the corresponding  $-2 \ln L(\theta)$  plus terms not depending on  $\theta$ , so minimizing them gives the usual ML estimators for  $\theta$ .

The minimized value gives the statistic  $q_\mu$ , so we get goodness-of-fit for free.

Steve Baker and Robert D. Cousins, *Clarification of the use of the chi-square and likelihood functions in fits to histograms*, NIM **221** (1984) 437.

# 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,

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

# 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.

G. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.

# Extra slides

# 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

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

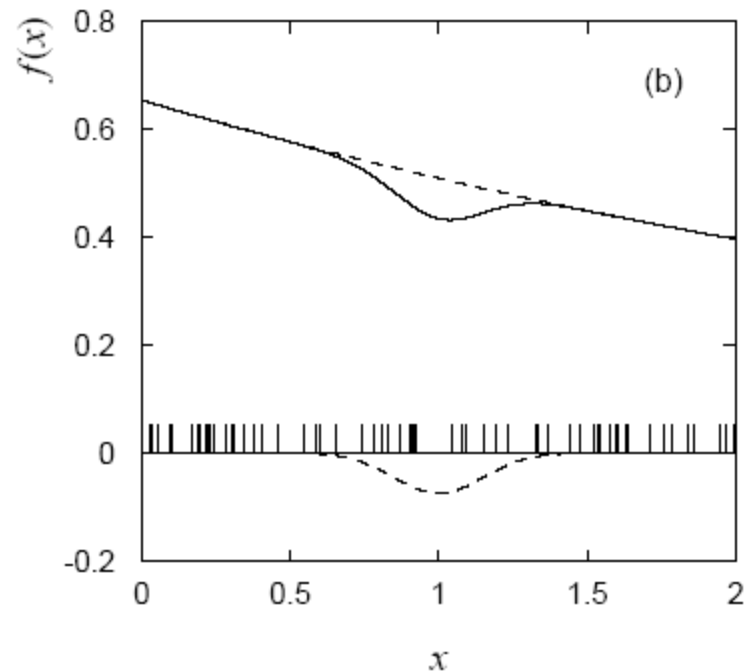
is an unbiased estimator for  $\sigma^2$ .

# 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:

