
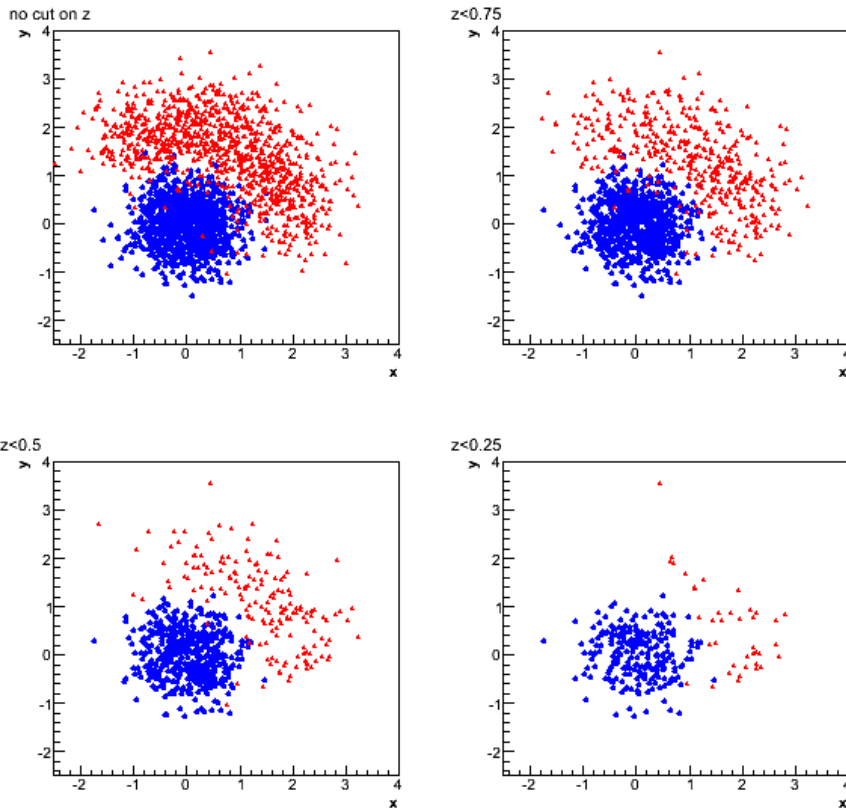


# Statistical Data Analysis: Lecture 8

- 1 Probability, Bayes' theorem, random variables, pdfs
- 2 Functions of r.v.s, expectation values, error propagation
- 3 Catalogue of pdfs
- 4 The Monte Carlo method
- 5 Statistical tests: general concepts
- 6 Test statistics, multivariate methods
- 7 Significance tests
-  8 **Parameter estimation, maximum likelihood**
- 9 More maximum likelihood
- 10 Method of least squares
- 11 Interval estimation, setting limits
- 12 Nuisance parameters, systematic uncertainties
- 13 Examples of Bayesian approach
- 14 tba

# Pre-lecture comments on problem sheet 7

Problem sheet 7 involves modifying some C++ programs to create a Fisher discriminant and neural network to separate two types of events (signal and background):



Each event is characterized by 3 numbers:  $x$ ,  $y$  and  $z$ .

Each "event" (instance of  $x, y, z$ ) corresponds to a "row" in an  $n$ -tuple. (here, a 3-tuple).

In ROOT,  $n$ -tuples are stored in objects of the `TTree` class.

# 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:  $\vec{x} = (x_1, \dots, x_n)$

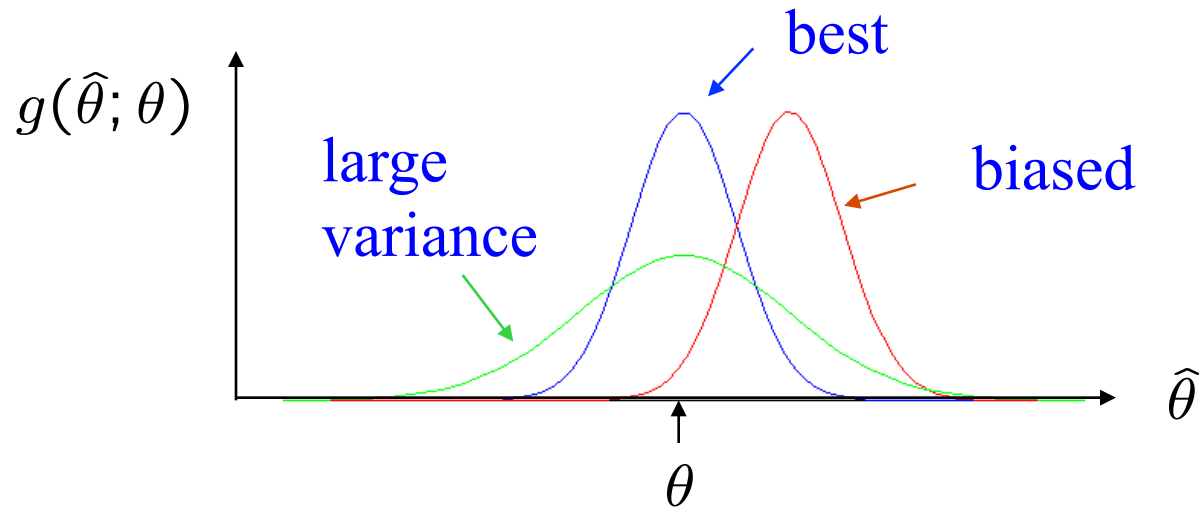
We want to find some function of the data to **estimate** the parameter(s):

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

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 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

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

We find:

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

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

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

# The likelihood function

Suppose the entire result of an experiment (set of measurements) is a collection of numbers  $\mathbf{x}$ , and suppose the joint pdf for the data  $\mathbf{x}$  is a function that depends on a set of parameters  $\theta$ :

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

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

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

( $\mathbf{x}$  constant)

# 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:

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

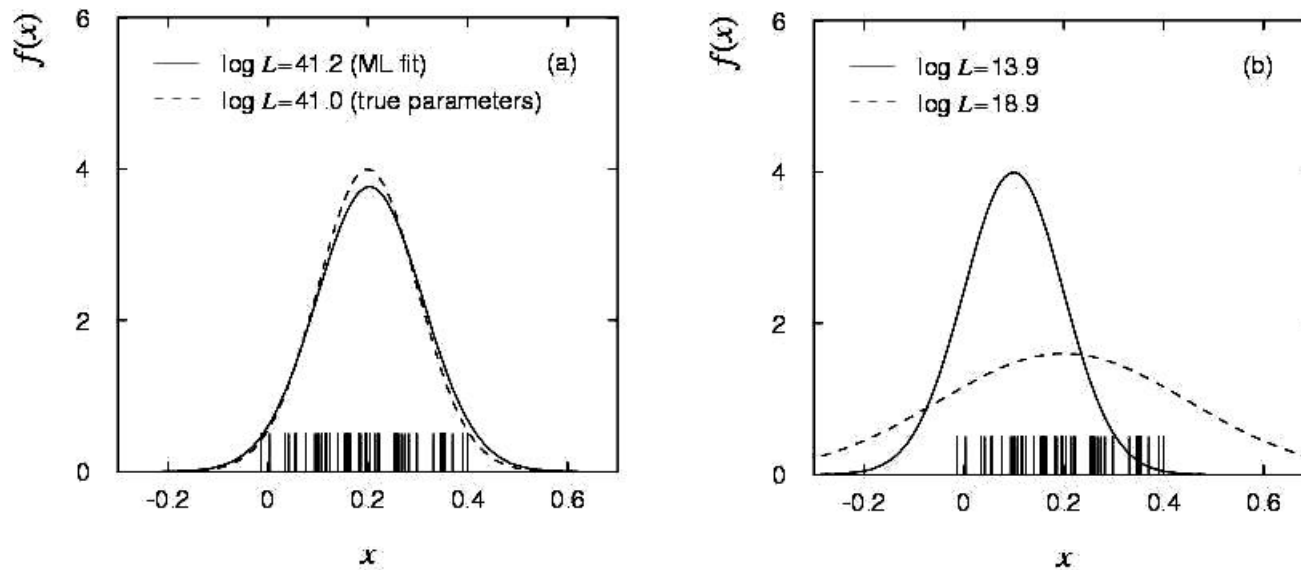
In this case the likelihood function is

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



# Maximum likelihood 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 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  $\frac{\partial \ln L(\tau)}{\partial \tau} = 0$ ,

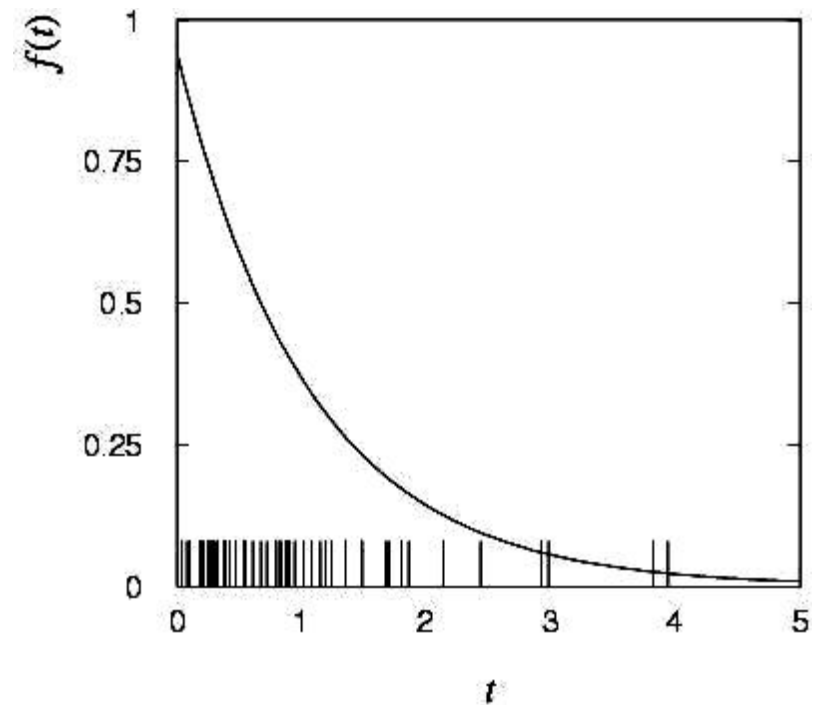
$$\rightarrow \hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$$

Monte Carlo test:

generate 50 values  
using  $\tau = 1$ :

We find the ML estimate:

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



## 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

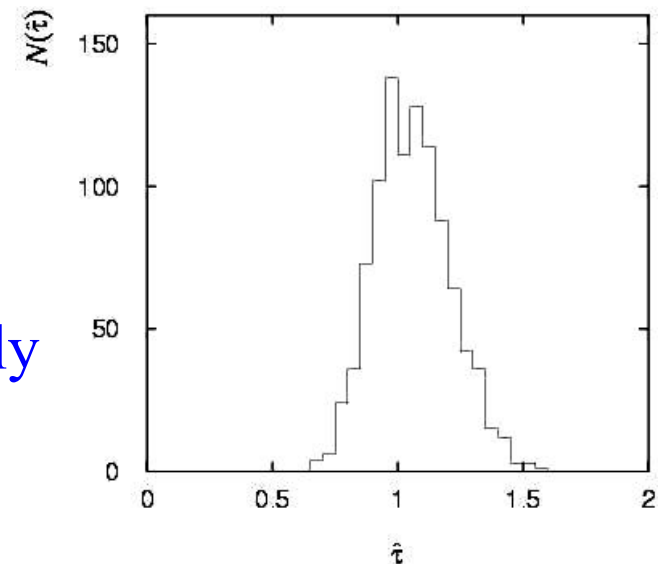
Having estimated our parameter we now need to report its ‘statistical error’, i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

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 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):

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

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

→ 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

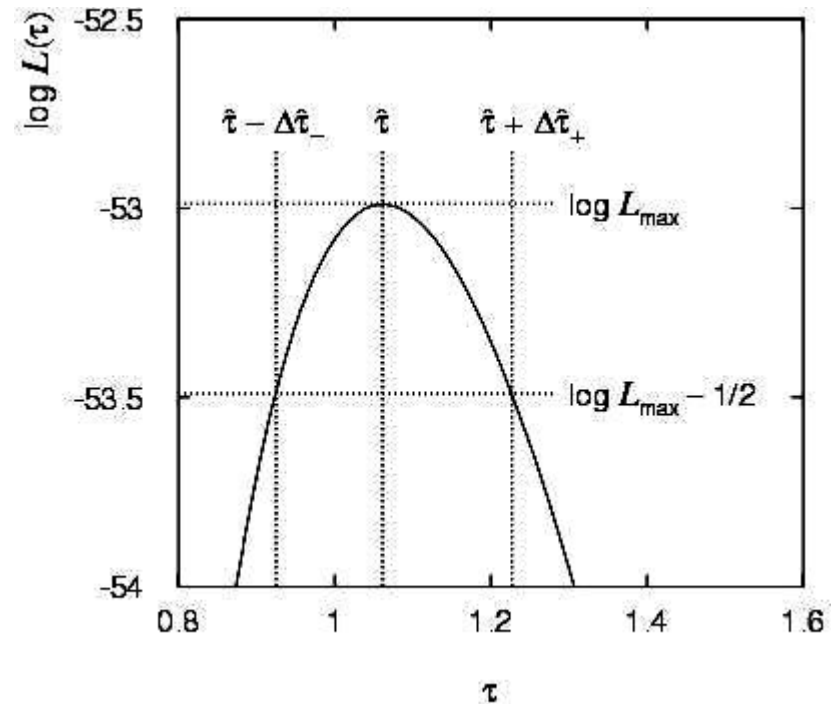
ML example with exponential:

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

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

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

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



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

# Wrapping up lecture 8

We've seen some main ideas about parameter estimation:

estimators, bias, variance,

and introduced the likelihood function and ML estimators.

Also we've seen some ways to determine the variance (statistical error) of estimators:

Monte Carlo method

Using the information inequality

Graphical Method

Next we will extend this to cover multiparameter problems, variable sample size, histogram-based data, ...