Introduction to Statistics − Day 3

Lecture 1

- Probability
- Random variables, probability densities, etc.
- Brief catalogue of probability densities

Lecture 2

- The Monte Carlo method
- Statistical tests
- Fisher discriminants, neural networks, etc.

→ Lecture 3

- Goodness-of-fit tests
- Parameter estimation
- Maximum likelihood and least squares
- Interval estimation (setting limits)

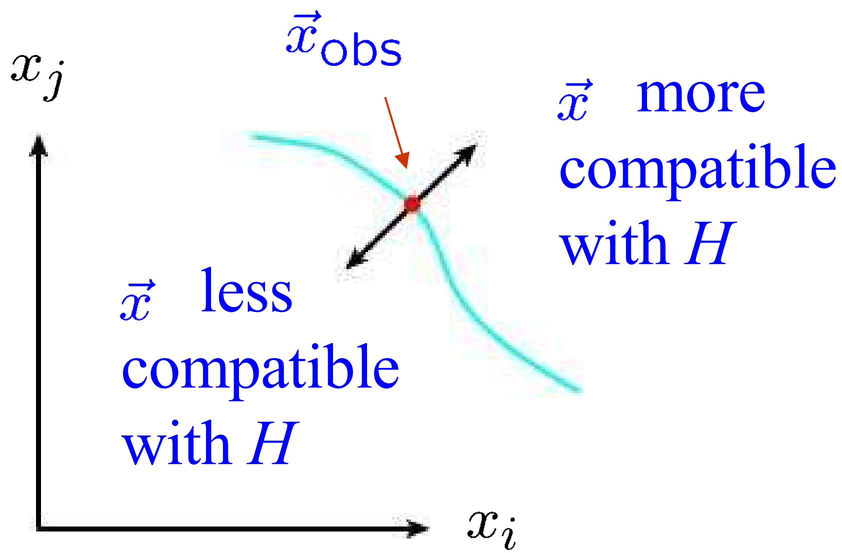
Testing goodness-of-fit

Suppose hypothesis H predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$.

We observe a single point in this space: $\vec{x}_{\mathrm{obs}}$

What can we say about the validity of H in light of the data?

Decide what part of the data space represents less compatibility with H than does the point $\vec{x}_{\mathrm{obs}}$. (Not unique!)



p-values

Express 'goodness-of-fit' by giving the p-value for H:

p = probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data we got.

This is not the probability that H is true!

In frequentist statistics we don't talk about $P(H)$ (unless H represents a repeatable observation). In Bayesian statistics we do; use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for H.

For now stick with the frequentist approach;
result is p-value, regrettably easy to misinterpret as $P(H)$.

p-value example: testing whether a coin is 'fair'

Probability to observe n heads in N coin tosses is binomial:

$$P(n; p, N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Hypothesis H: the coin is fair ($p = 0.5$).

Suppose we toss the coin $N = 20$ times and get $n = 17$ heads.

Region of data space with equal or lesser compatibility with H relative to $n = 17$ is: $n = 17, 18, 19, 20, 0, 1, 2, 3$. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 \ .$$

i.e. $p = 0.0026$ is the probability of obtaining such a bizarre result (or more so) 'by chance', under the assumption of H.

The significance of an observed signal

Suppose we observe n events; these can consist of:

n_b events from known processes (background)

n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means s, b, then $n = n_s + n_b$ is also Poisson, mean $= s + b$:

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

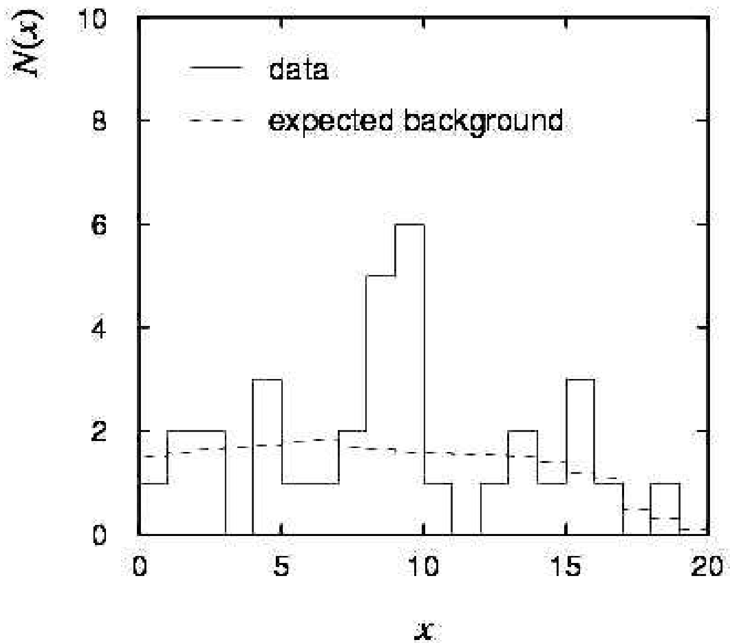
Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$:

$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$

The significance of a peak

Suppose we measure a value x for each event and find:



Each bin (observed) is a Poisson r.v., means are given by dashed lines.

In the two bins with the peak, 11 entries found with $b = 3.2$. The p-value for the $s = 0$ hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

The significance of a peak (2)

But... did we know where to look for the peak?

→ give $P(n \geq 11)$ in any 2 adjacent bins

Is the observed width consistent with the expected x resolution?

→ take x window several times the expected resolution

How many bins × distributions have we looked at?

→ look at a thousand of them, you'll find a 10^{-3} effect

Did we adjust the cuts to 'enhance' the peak?

→ freeze cuts, repeat analysis with new data

How about the bins to the sides of the peak... (too low!)

Should we publish????

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

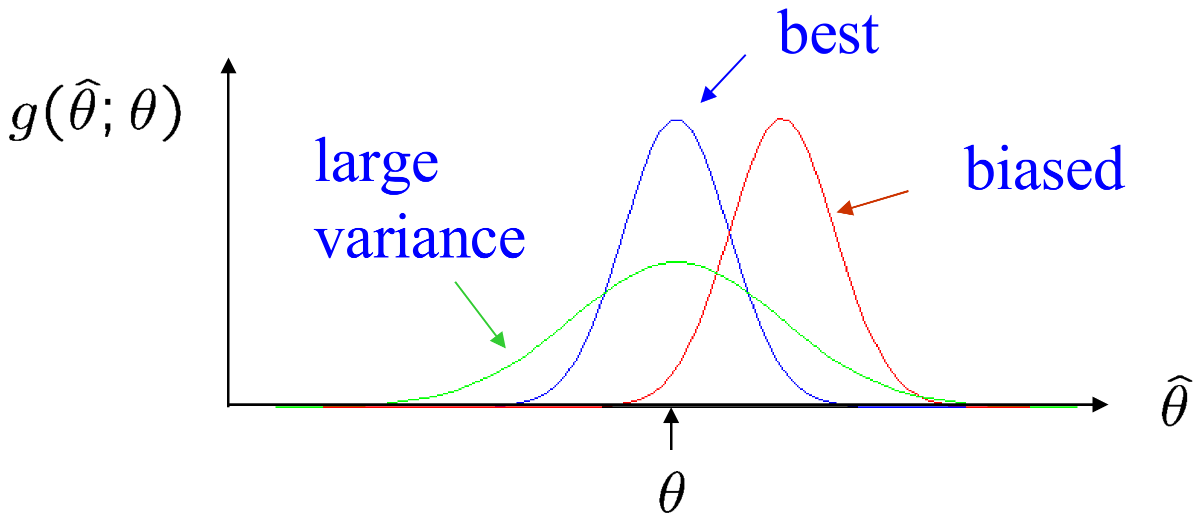
We want to find some function of the data to estimate the parameter(s):

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

Sometimes we say 'estimator' for the function of $x_1, \ldots, x_n$;
'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): $b = E[\widehat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

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

→ small bias & variance are in general conflicting criteria

An estimator for the mean (expectation value)

Parameter: $\mu = E[x]$

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

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

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

An estimator for the variance

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

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

We find:

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

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

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

The likelihood function

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

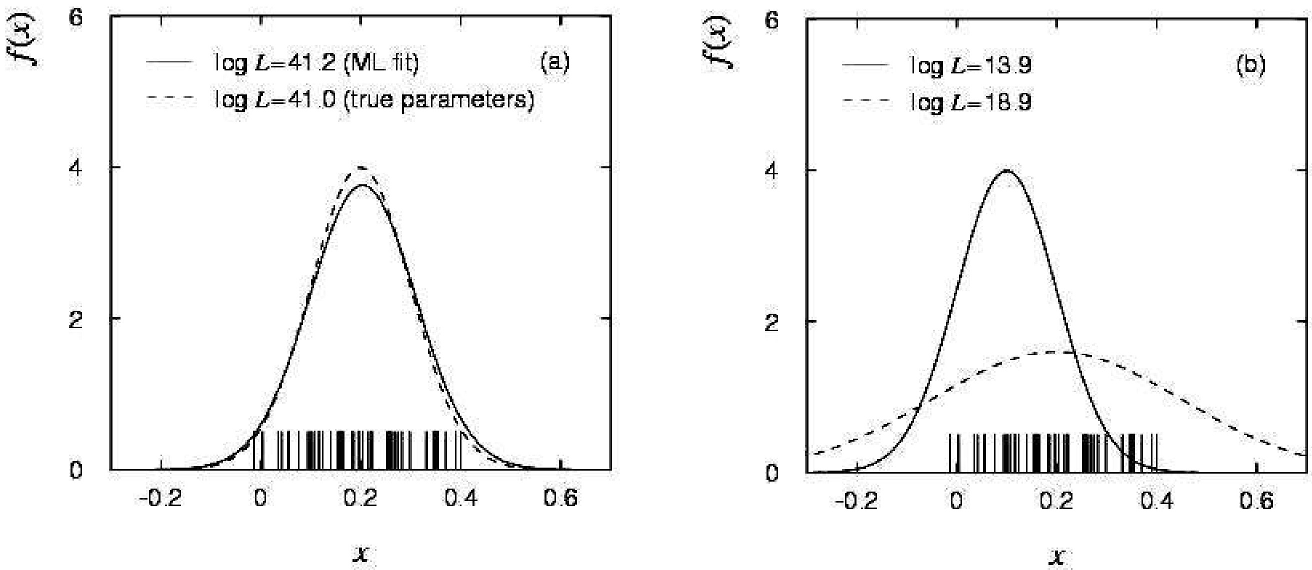
$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

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

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

Maximum likelihood estimators

If the hypothesized θ 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, $$t_1, \ldots, t_n$$

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

The value of τ 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)

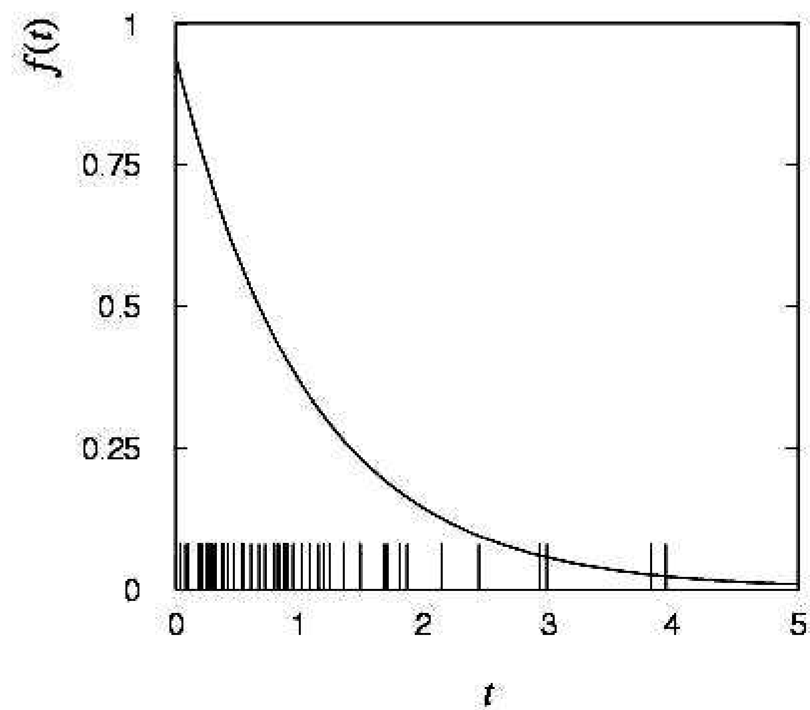
Find its maximum by setting $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0$,

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

Monte Carlo test:

generate 50 values

using $\tau = 1$:



We find the ML estimate:

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

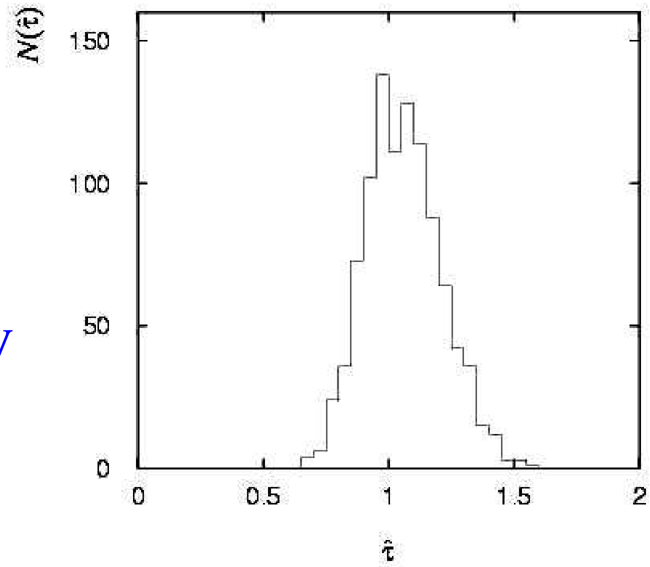
Variance of estimators: Monte Carlo method

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):

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

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

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

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

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

Variance of estimators: graphical method

Expand ln L (θ) about its maximum:

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

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

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

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 θ away from $\hat{\theta}$ until ln L decreases by 1/2.

Example of variance by graphical method

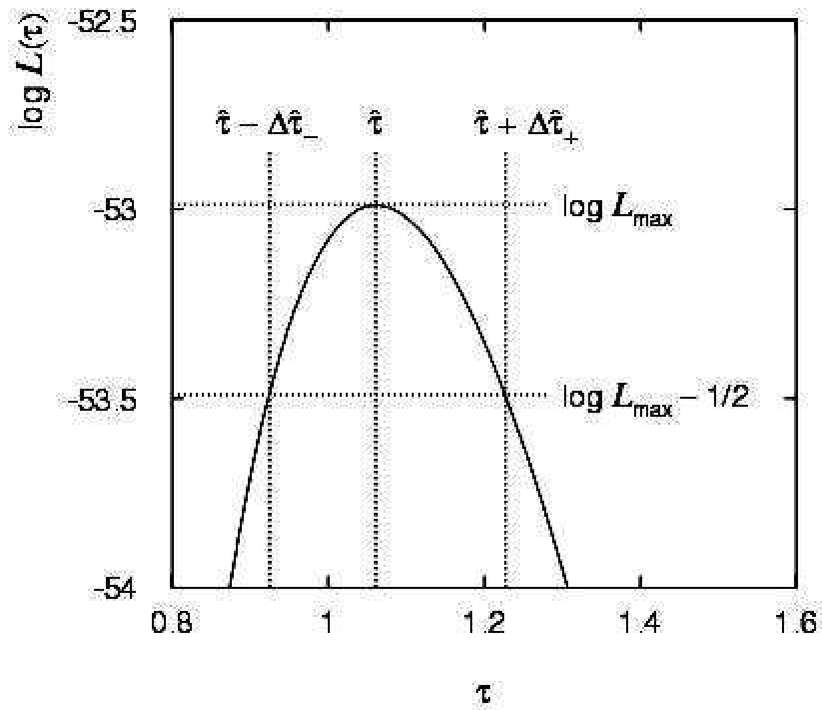
ML example with exponential:

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

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

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

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

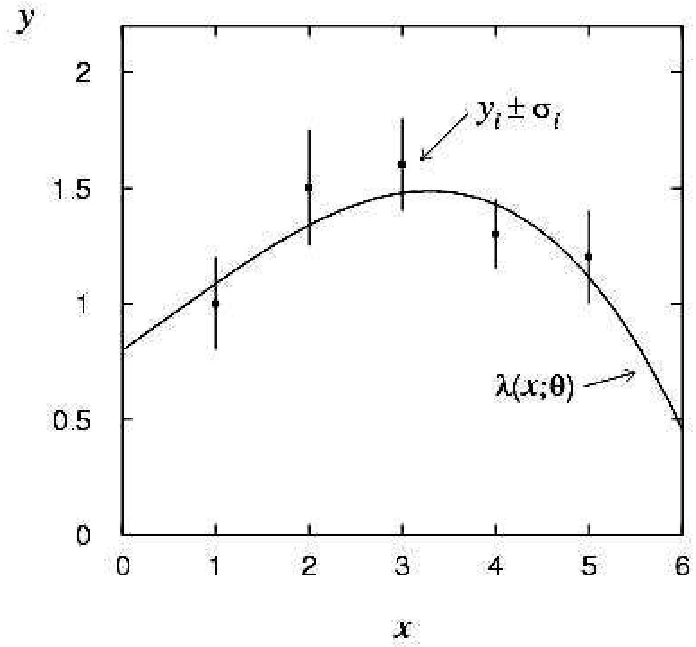


Not quite parabolic ln L since finite sample size (n = 50).

The method of least squares

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

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



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

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

We want to estimate θ, 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

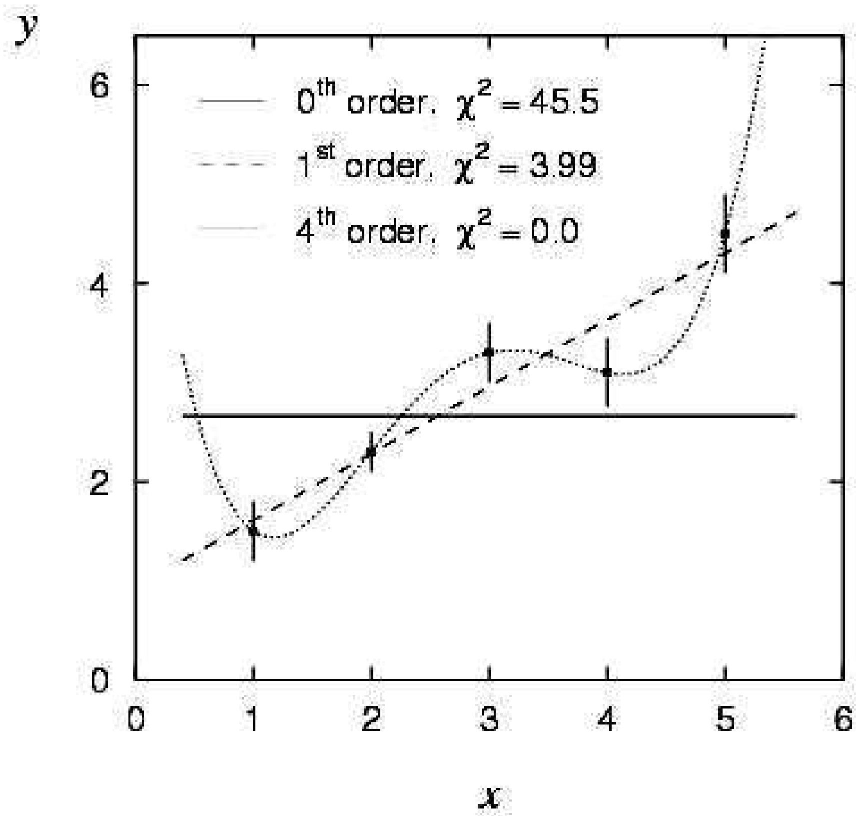
$$\chi^2(\theta) = \sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2}$$

Minimum of this quantity defines the least squares estimator $\hat{\theta}$.

Often minimize χ^2 numerically (e.g. program `MINUIT`).

Example of least squares fit

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



Variance of LS estimators

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

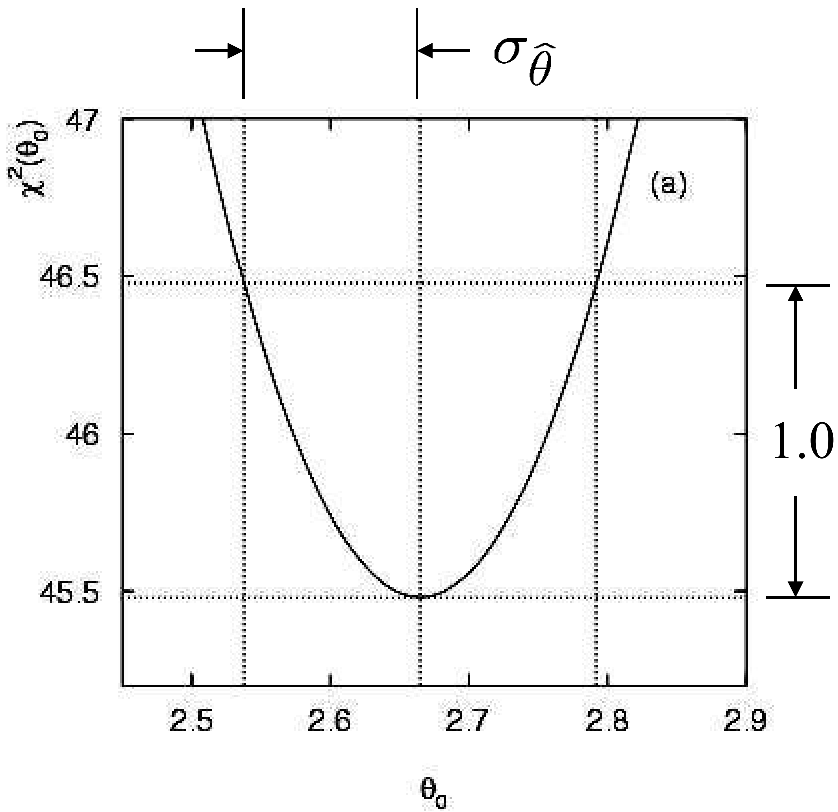
$$\chi^2(\theta) = -2 \ln L(\theta)$$

and so

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

or for the graphical method we take the values of θ where

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



Goodness-of-fit with least squares

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

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

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 $t = \chi^2_{\mathrm{min}}$ follows the chi-square pdf,

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

where the number of degrees of freedom is

$$n_{\mathrm{d}} = \text{number of data points} - \text{number of fitted parameters}$$

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:

$$p = \int_{\chi^2_{\min}}^{\infty} f(t; n_{\rm d})\, dt$$

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),

$$\chi^2_{\min} = 3.99\,, \qquad n_{\rm d} = 5-2 = 3\,, \qquad p = 0.263$$

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

$$\chi^2_{\min} = 45.5\,, \qquad n_{\rm d} = 5-1 = 4\,, \qquad p = 3.1\times 10^{-9}$$

Setting limits

Consider again the case of finding $n = n_s + n_b$ events where

n_b events from known processes (background)

n_s events from a new process (signal)

are Poisson r.v.s with means s, b, and thus $n = n_s + n_b$ is also Poisson with mean $= s + b$. Assume b is known.

Suppose we are searching for evidence of the signal process, but the number of events found is roughly equal to the expected number of background events, e.g., $b = 4.6$ and we observe $n_{obs} = 5$ events.

The evidence for the presence of signal events is not statistically significant,

$\rightarrow$ set upper limit on the parameter s.

Example of an upper limit

Find the hypothetical value of s such that there is a given small probability, say, $\gamma = 0.05$, to find as few events as we did or less:

$$\gamma = P(n \leq n_{\rm obs}; s, b) = \sum_{n=0}^{n_{\rm obs}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Solve numerically for $s = s_{\rm up}$, this gives an upper limit on s at a confidence level of $1-\gamma$.

Example: suppose $b = 0$ and we find $n_{\rm obs} = 0$. For $1-\gamma = 0.95$,

$$\gamma = P(n = 0; s, b = 0) = e^{-s} \quad \rightarrow \quad s_{\rm up} = -\ln \gamma \approx 3.00$$

Many subtle issues here – see e.g. CERN (2000) and Fermilab (2001) workshops on confidence limits.

Wrapping up lecture 3

We've seen how to quantify goodness-of-fit with

p-values,

and we've seen some main ideas about parameter estimation,

ML and LS,
how to obtain/interpret stat. errors from a fit,

and what to do if you don't find the effect you're looking for,

setting limits.

In three days we've only looked at some basic ideas and tools, skipping entirely many important topics. Keep an eye out for new methods, especially multivariate, machine learning, etc.