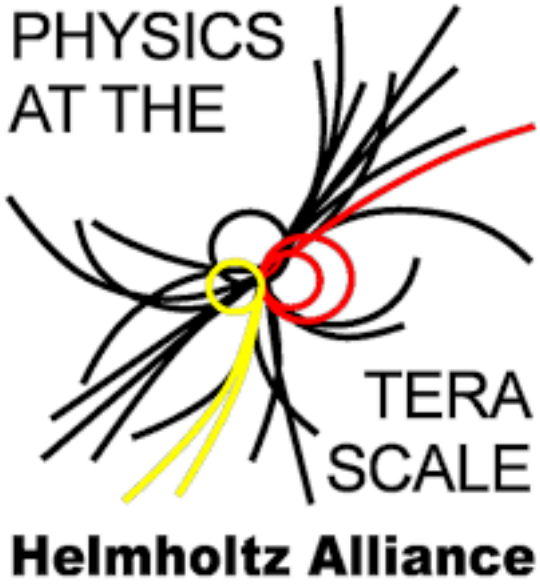
Statistics for Particle Physics

Lecture 2: Frequentist Parameter Estimation



Terascale Statistics School

https://indico.desy.de/event/51468/

DESY, Hamburg

23-27 Feb 2026

Glen Cowan

Physics Department

Royal Holloway, University of London

g.cowan@rhul.ac.uk

www.pp.rhul.ac.uk/~cowan

Outline

Lectures/tutorials from me:

1)	Monday 16:00	Hypothesis testing
→ 2)	Tuesday 9:00	Frequentist parameter estimation
	Tuesday 11:00	
3)	Tuesday 14:00	Confidence limits
	Tuesday 16:00	
4)	Wednesday 9:00	Bayesian parameter estimation
5)	Wednesday 14:00	Errors on errors

Parameter Estimation 2-1

- Introduction to (frequentist) parameter estimation
- The method of Maximum Likelihood
- MLE for exponential distribution

Frequentist parameter estimation

The parameters of a pdf are any constants that characterize it,

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

r.v. parameter

i.e., θ indexes a set of hypotheses.

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

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

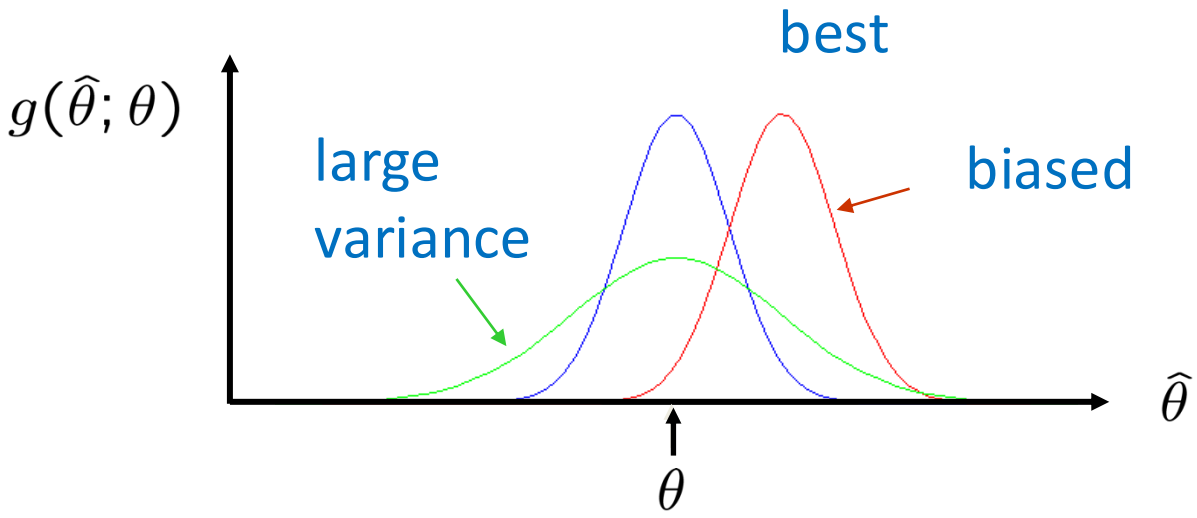
$$\widehat{\theta}(\vec{x})$$

← 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): $b = E[\hat{\theta}] - \theta$

→ 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: $\mu = E[x] = \langle x \rangle = \int_{-\infty}^{\infty} x f(x)\, dx$

Suppose we have a sample of n independent values $x_1,\ldots,x_n$.

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: $\sigma^2 = V[x] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$

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

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

$$P(\mathbf{x}|\boldsymbol{\theta})$$

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

$$L(\boldsymbol{\theta}) = P(\mathbf{x}|\boldsymbol{\theta})$$

($\boldsymbol{x}$ constant)

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

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

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

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

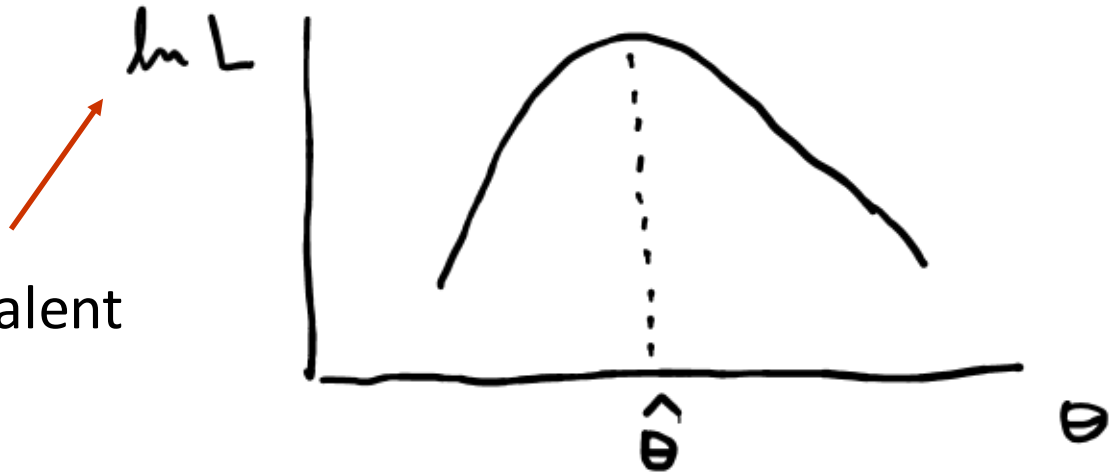
In this case the likelihood function is

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

Maximum Likelihood Estimators (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L equivalent to maximizing $\log L$



Could have multiple maxima (take highest).

MLEs not guaranteed to have any 'optimal' properties, (but in practice they're very good).

MLE 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, $t_1, \ldots, t_n$

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):

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

MLE example: parameter of exponential pdf (2)

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

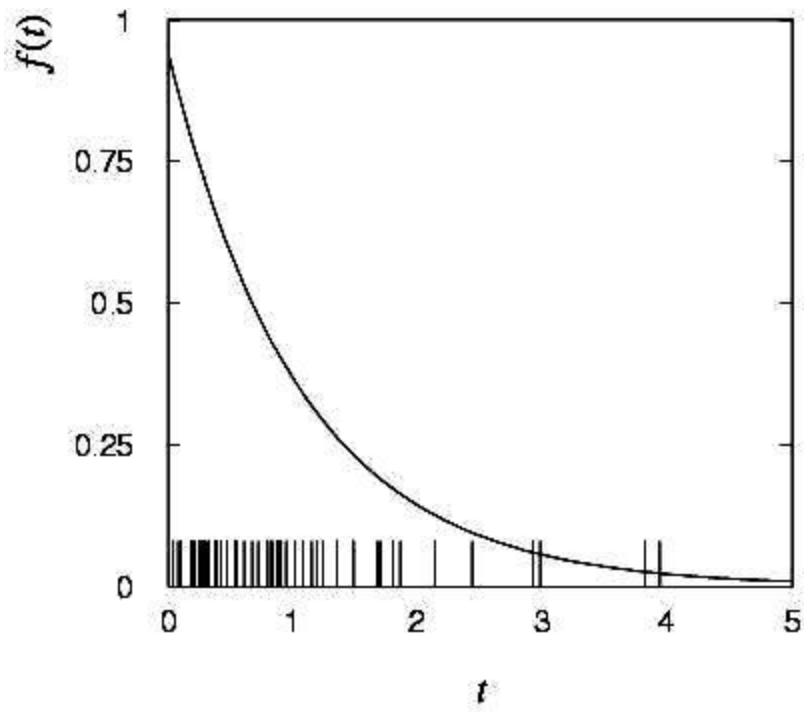
$$\rightarrow \quad \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

Monte Carlo test:

generate 50 values using $\tau = 1$:

We find the ML estimate:

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



MLE example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$$

$$V[t] = \int_0^\infty (t-\tau)^2 \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$$

For the MLE $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

Large-sample (asymptotic) properties of MLEs

Suppose we have an i.i.d. data sample of size n: $x_1,...,x_n$

In the large-sample (or "asymptotic") limit ($n \to \infty$) and assuming regularity conditions one can show that the likelihood and MLE have several important properties.

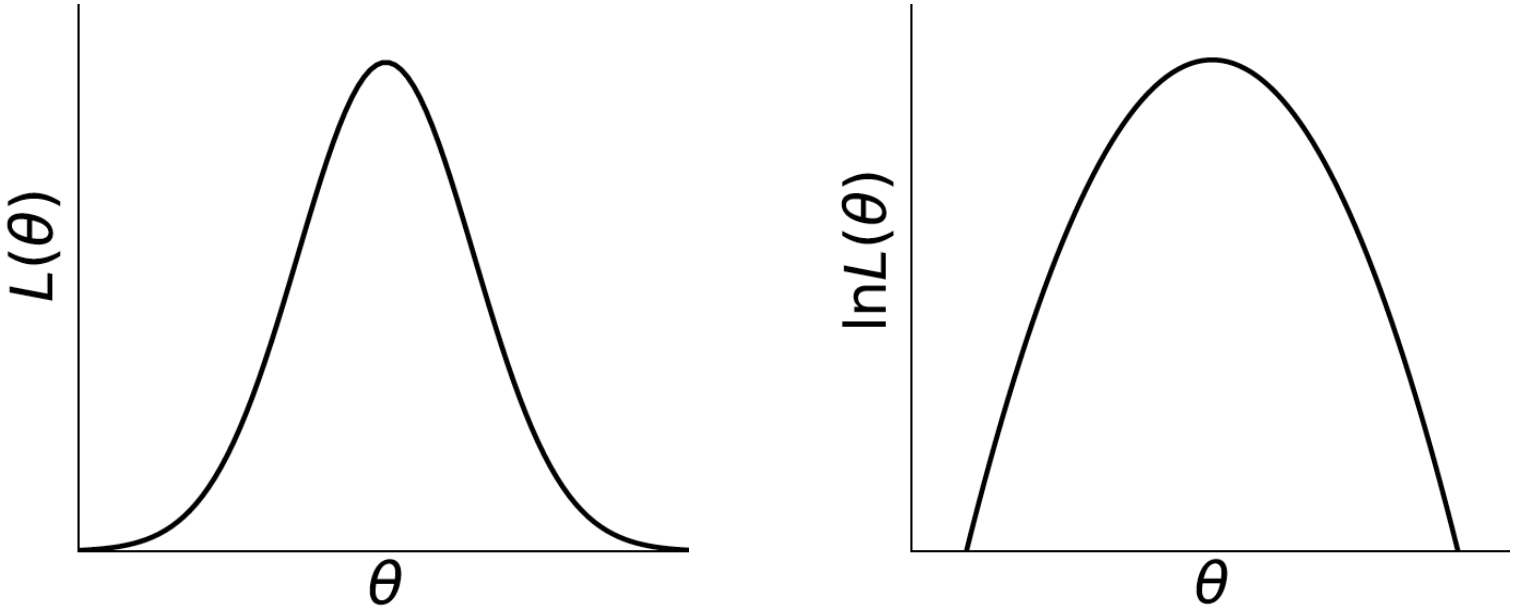
The regularity conditions include:

- the boundaries of the data space cannot depend on the parameter;
- the parameter cannot be on the edge of the parameter space;
- $\ln L(\theta)$ must be differentiable;
- the only solution to $\partial \ln L / \partial \theta = 0$ is $\hat{\theta}$.

In the slides immediately following the properties are shown without proof for a single parameter; the corresponding properties hold also for the multiparameter case, $\boldsymbol{\theta} = (\theta_1,..., \theta_m)$.

log-likelihood becomes quadratic

The likelihood function becomes Gaussian in shape, i.e. the log-likelihood becomes quadratic (parabolic).



The MLE becomes increasingly precise as the (log)-likelihood becomes more tightly concentrated about its peak,
but $L(\theta) = P(\boldsymbol{x}|\theta)$ is the probability for $\boldsymbol{x}$, not a pdf for θ.

The MLE converges to the true parameter value

In the large-sample limit, the MLE converges in probability to the true parameter value.

That is, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

The MLE is said to be *consistent*.

MLE is asymptotically unbiased

In general the MLE can be biased, but in the large-sample limit, this bias goes to zero:

$$\lim_{n\to\infty} E[\hat{\theta}] - \theta = 0$$

(Recall for the exponential parameter we found the bias was identically zero regardless of the sample size n.)

The information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only MLE). For a single parameter:

$$(b = E[\hat{\theta}] - \theta)$$

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

= MVB (Minimum Variance Bound)

where $$E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = \int \frac{\partial^2 \ln P(\mathbf{x}|\theta)}{\partial \theta^2} P(\mathbf{x}|\theta)\, d\mathbf{x}$$

Proof in Exercise 6.6 of SDA, http://www.pp.rhul.ac.uk/~cowan/sda/prob/prob_6.pdf

"Efficiency" of an estimator = MVB / actual variance.

An estimator whose variance equals the MVB is said to be efficient.

The MLE's variance approaches the MVB

In the large-sample limit, the variance of the MLE approaches the minimum-variance bound, i.e., the information inequality becomes an equality (and bias goes to zero):

$$\lim_{n\to\infty} V[\hat{\theta}] = -\frac{1}{E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

The MLE is said to be *asymptotically efficient*.

The MLE's distribution becomes Gaussian

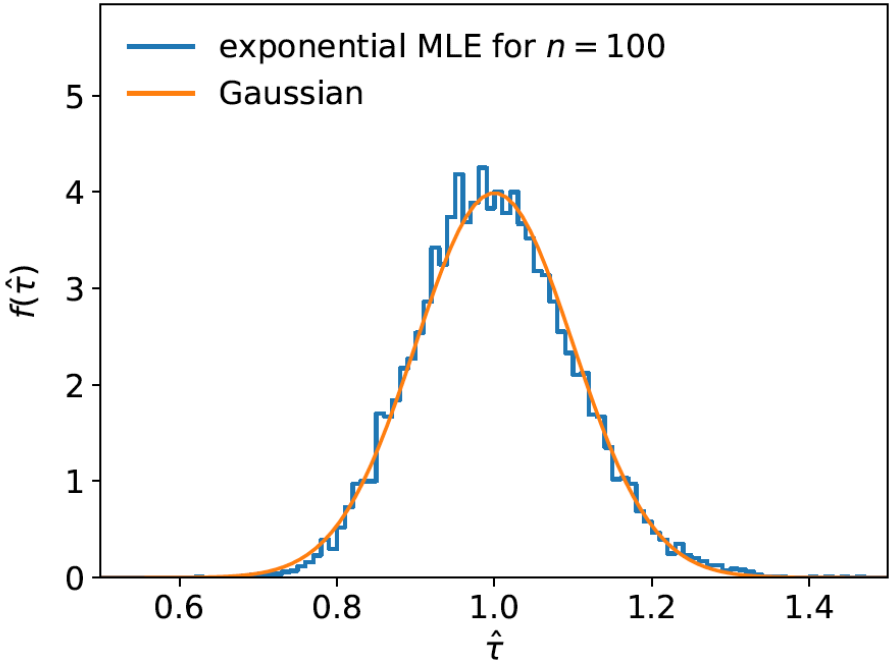
In the large-sample limit, the pdf of the MLE becomes Gaussian,

$$f(\hat{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{\theta}}} e^{-(\hat{\theta}-\theta)^2/2\sigma^2_{\hat{\theta}}}$$

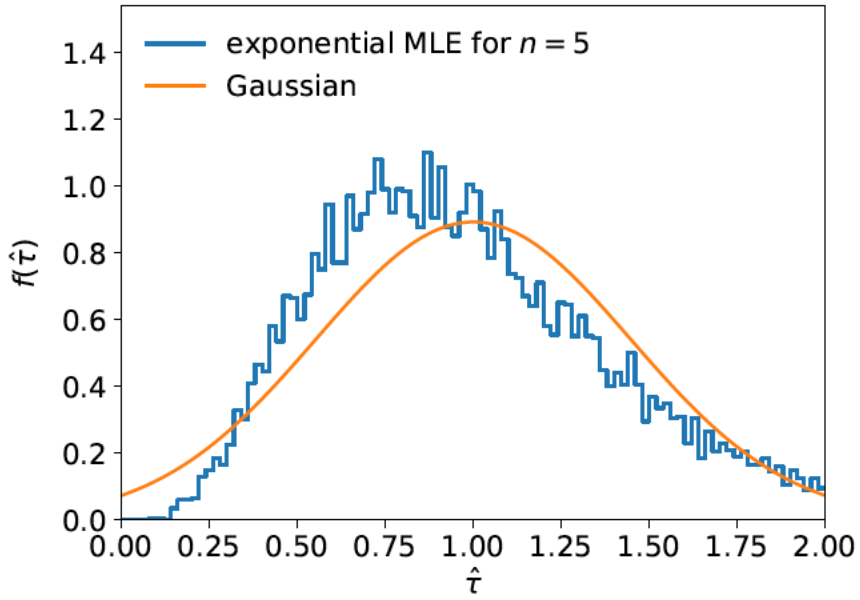
where $\sigma^2_{\hat{\theta}}$ is the minimum variance bound (note bias is zero).

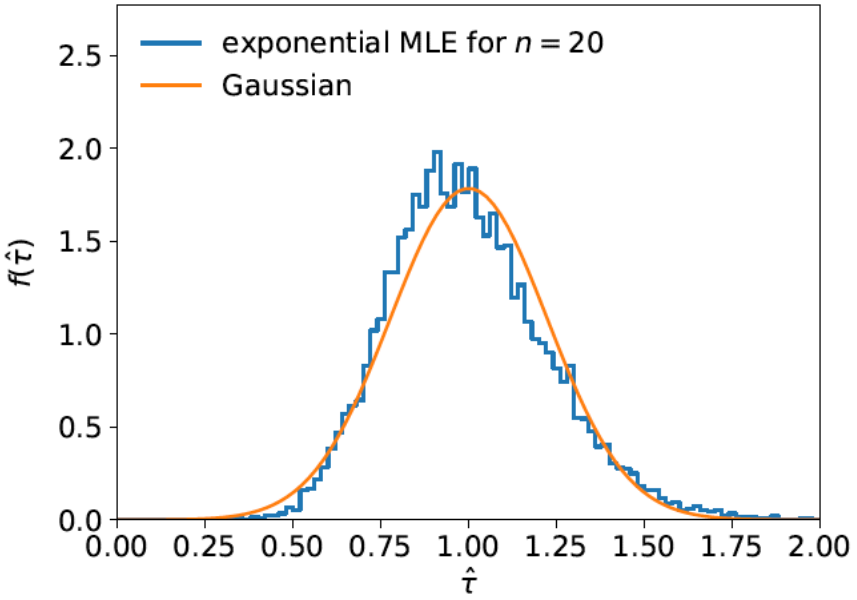
For example, exponential MLE with sample size $n = 100$.

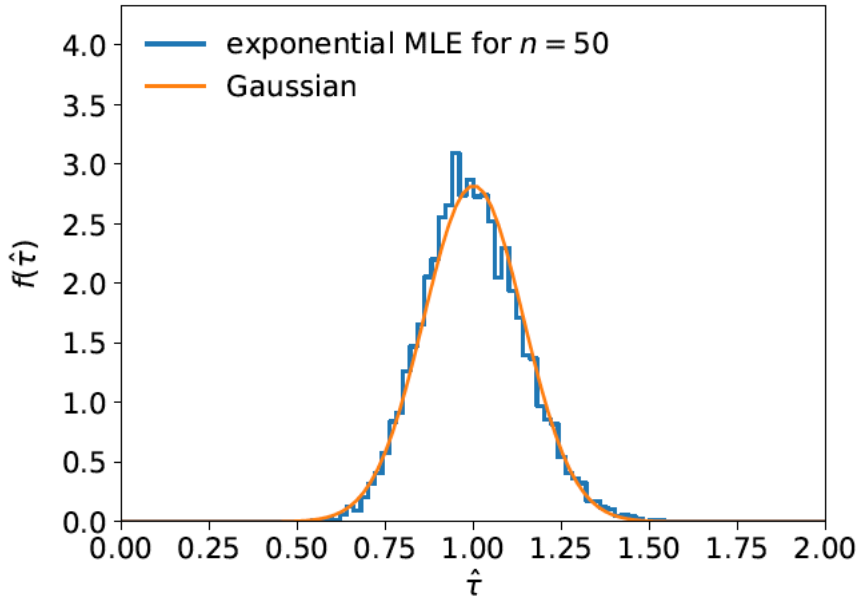
Note that for exponential, MLE is arithmetic average, so Gaussian MLE seen to stem from Central Limit Theorem.

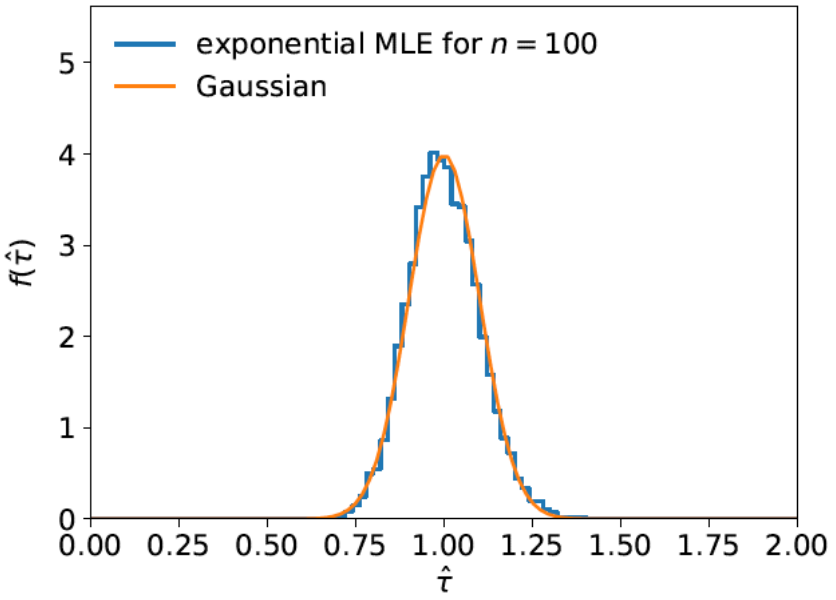


Distribution of MLE of exponential parameter









Parameter Estimation 2-2

- Finding the variance of MLEs
- Information inequality for multiple parameters

Variance of estimators: Monte Carlo method

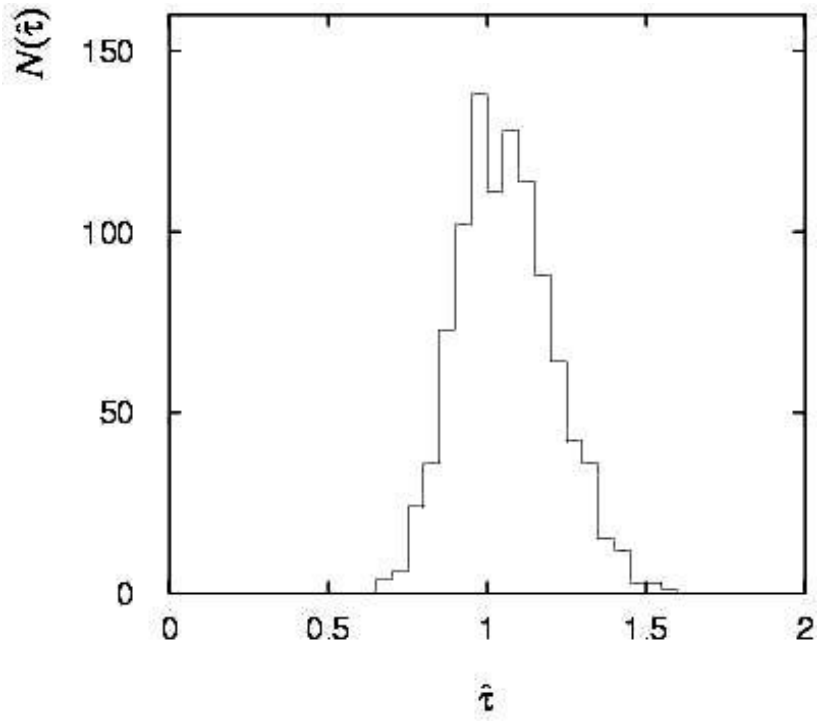
Having estimated our parameter we now need to report its 'statistical error', using e.g. the estimator's standard deviation, or (co)variance.

It is usually not possible to do this with an exact calculation.

Another way is to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example (n=50), from sample variance of estimates we find:

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



Variance of estimators from information inequality

Recall the information inequality (RCF) sets a lower bound on the variance of any estimator (not only MLE):

MVB

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

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:

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

MVB for MLE of exponential parameter

Find $$\mathrm{MVB} = -\left(1 + \frac{\partial b}{\partial \tau}\right)^2 \Bigg/ E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right]$$

We found for the exponential parameter the MLE $$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

and we showed $b = 0$, hence $\partial b/\partial \tau = 0$.

We find $$\frac{\partial^2 \ln L}{\partial \tau^2} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - \frac{2t_i}{\tau^3}\right)$$

and since $E[t_i] = \tau$ for all i, $$E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right] = -\frac{n}{\tau^2}\ ,$$

and therefore $$\mathrm{MVB} = \frac{\tau^2}{n} = V[\hat{\tau}]\ .$$ So here the MLE is efficient.

Variance of estimators: graphical method

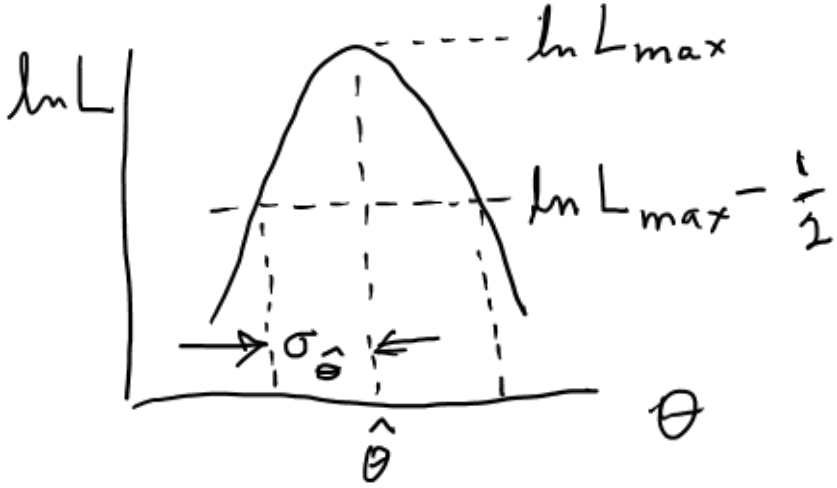
Expand $\ln L(\theta)$ about its maximum:

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

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

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

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



$\rightarrow$ 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

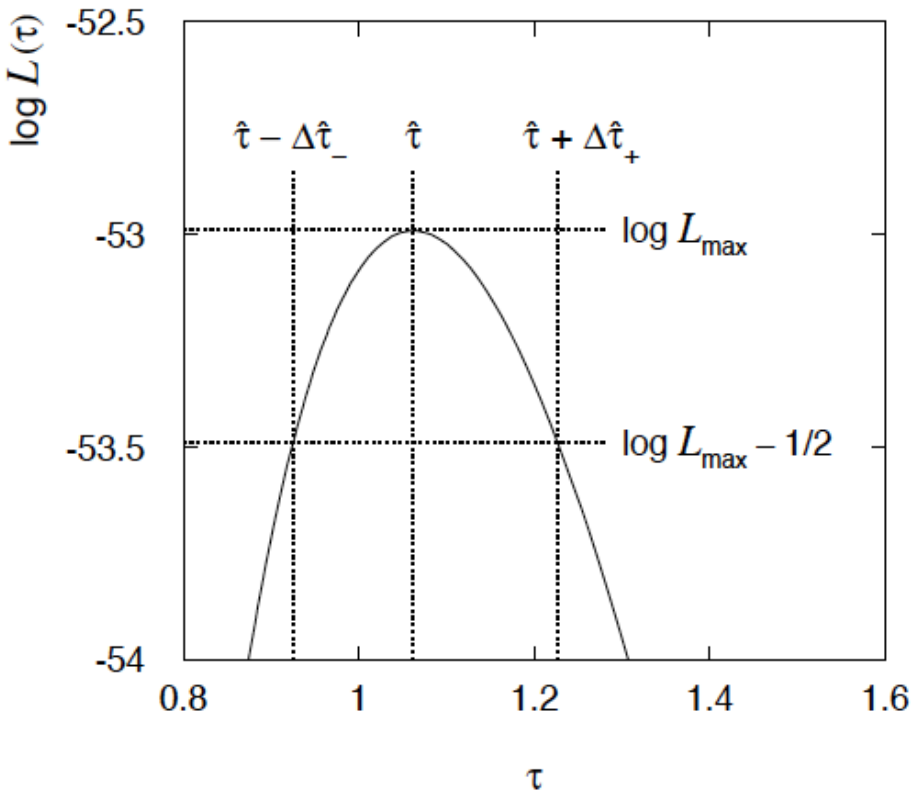
ML example with exponential:

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

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

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

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



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

Parameter Estimation 2-3

- Information inequality for multiple parameters
- Numerical example of 2-D MLE
- The $\ln L = \ln L_{\max} - ½$ contour
- MLE for function of a parameter
- Relation between MLE and Bayesian estimator

Information inequality for N parameters

Suppose we have estimated N parameters $\boldsymbol{\theta} = (\theta_1,...,\theta_N)$

The *Fisher information matrix* is

$$I_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} P(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}$$

and the covariance matrix of estimators $\hat{\boldsymbol{\theta}}$ is $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

The information inequality states that the matrix

$$M_{ij} = V_{ij} - \sum_{k,l} \left(\delta_{ik} + \frac{\partial b_i}{\partial\theta_k}\right) I^{-1}_{kl} \left(\delta_{lj} + \frac{\partial b_l}{\partial\theta_j}\right)$$

is positive semi-definite:

$\boldsymbol{z}^{\mathrm{T}} M \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \neq 0$, diagonal elements ≥ 0

Information inequality for N parameters (2)

In practice the inequality is ~always used in the large-sample limit:

bias $\rightarrow$ 0

inequality $\rightarrow$ equality, i.e, $M = 0$, and therefore $V = I^{-1}$

That is, $$V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

This can be estimated from data using $$\widehat{V}_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\hat{\boldsymbol{\theta}}}$$

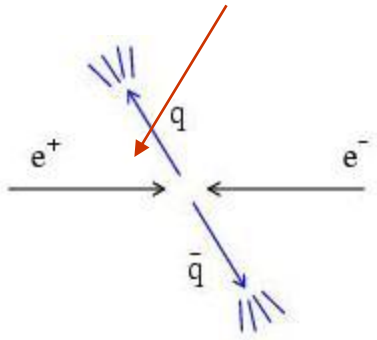
Find the matrix V^{-1} numerically (or with automatic differentiation), then invert to get the covariance matrix of the estimators

$$\widehat{V}_{ij} = \widehat{\mathrm{cov}}[\hat{\theta}_i, \hat{\theta}_j]$$

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 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.

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

need to find maximum numerically

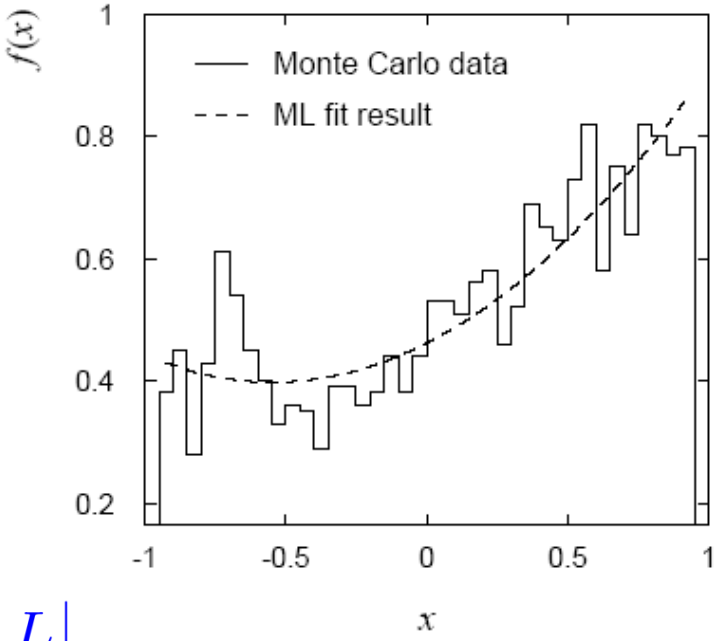
Example of ML with 2 parameters: fit result

Finding maximum of $\ln L(\alpha, \beta)$ numerically 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 χ^2).



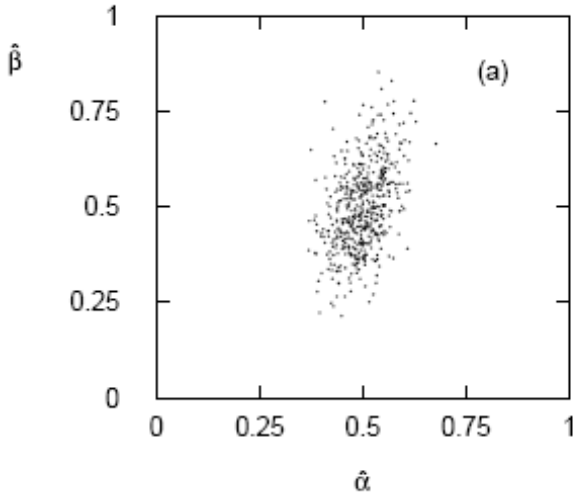
(Co)variances from $(\widehat{V^{-1}})_{ij} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\hat{\vec{\theta}}}$

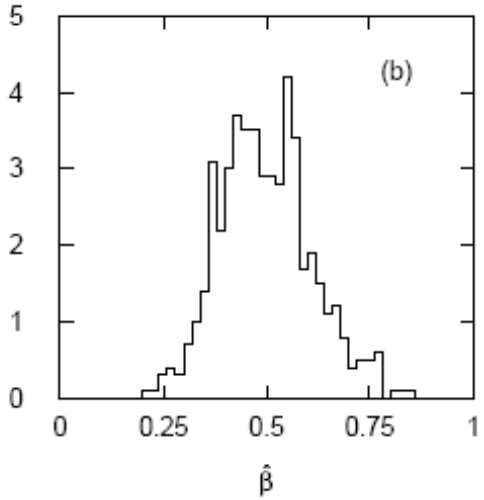
$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \qquad \mathrm{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$

$$\hat{\sigma}_{\hat{\beta}} = 0.11 \qquad r = 0.46 \text{ = correlation coefficient}$$

Two-parameter fit: MC study

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





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

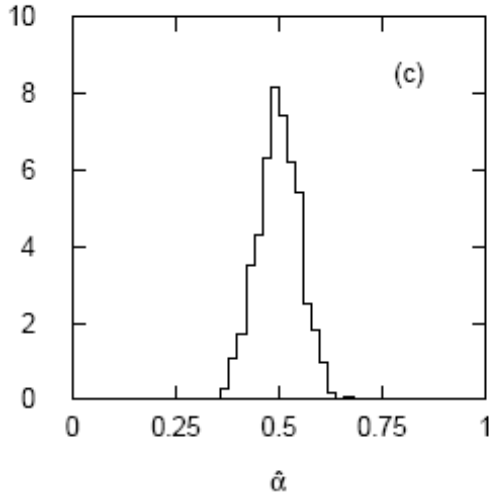
$$s_{\hat{\alpha}} = 0.051$$

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

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

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

$$r = 0.42$$



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

Multiparameter graphical method for variances

Expand $\ln L(\boldsymbol{\theta})$ to 2nd order about MLE:

$$\ln L(\boldsymbol{\theta}) \approx \ln L(\hat{\boldsymbol{\theta}}) + \sum_i \left.\frac{\partial \ln L}{\partial \theta_i}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i) + \frac{1}{2!} \sum_{i,j} \left.\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

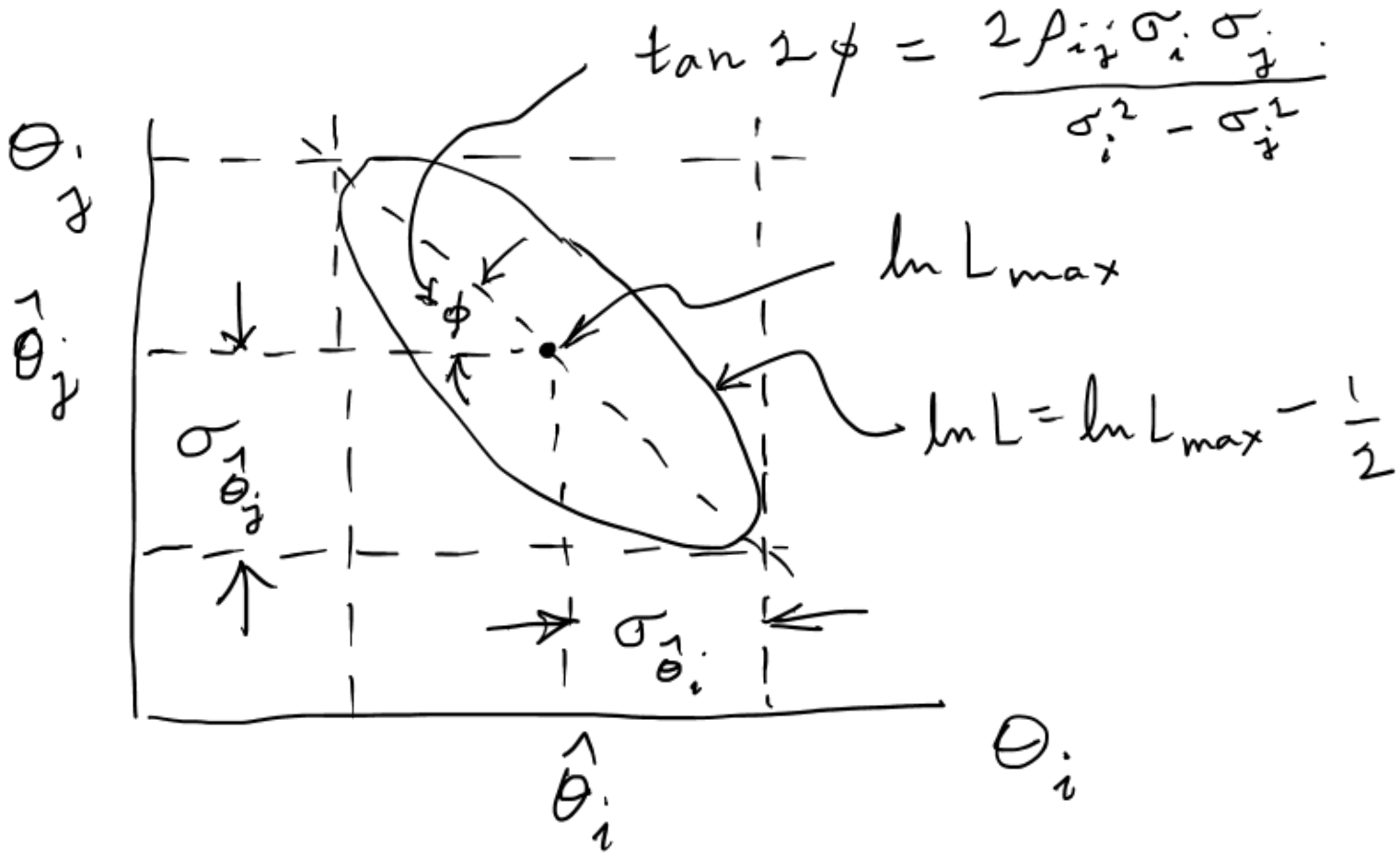
$\ln L_{\max}$ — zero — relate to covariance matrix of MLEs using information (in)equality.

Result: $$\ln L(\boldsymbol{\theta}) = \ln L_{\max} - \frac{1}{2} \sum_{i,j} (\theta_i - \hat{\theta}_i) V_{ij}^{-1} (\theta_j - \hat{\theta}_j)$$

So the surface $\ln L(\boldsymbol{\theta}) = \ln L_{\max} - \frac{1}{2}$ corresponds to

$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T V^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = 1$, which is the equation of a (hyper-) ellipse.

Multiparameter graphical method (2)



Distance from MLE to tangent planes gives standard deviations.

The ln $L_{\max}$ – 1/2 contour for two parameters

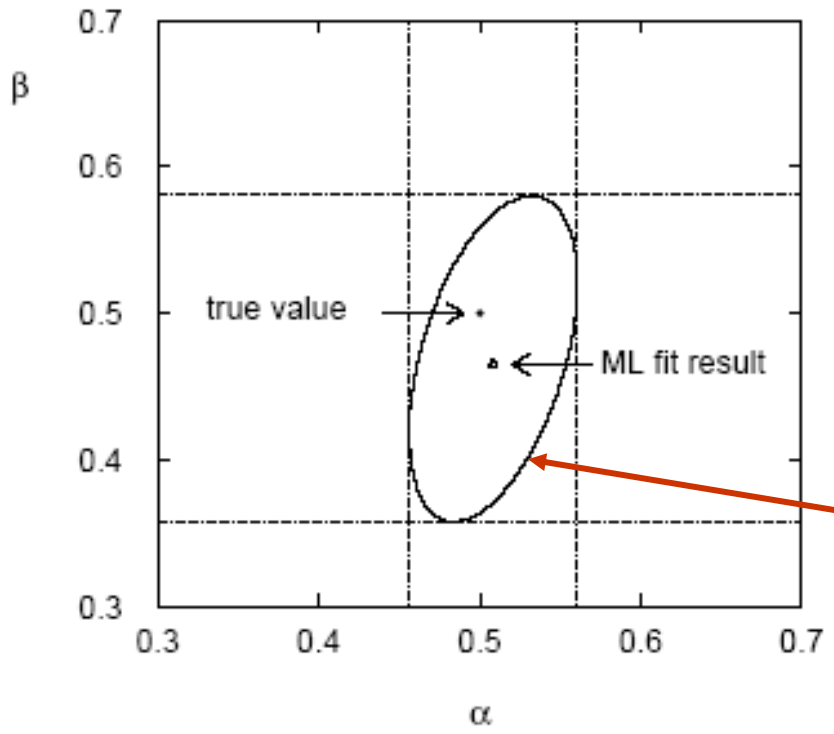
For large n, $\ln L$ takes on quadratic form near maximum:

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

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 α, β 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 φ related to correlation: $\tan 2\phi = \dfrac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma^2_{\hat{\alpha}} - \sigma\hat{\beta}^2}$

Functions of maximum-likelihood estimators

Suppose likelihood has a parameter θ.

Define a new parameter α given by function $\alpha = a(\theta)$.

What is the MLE of α?

Suppose $a(\theta)$ has a unique inverse, so $\theta = a^{-1}(\alpha)$.

The likelihood is $L(\theta) = L(a^{-1}(\alpha))$.

The maximum of the likelihood is $L_{\max} = L(\hat{\theta})$.

So to maximize L, find $\alpha \equiv \hat{\alpha}$ such that

$$a^{-1}(\hat{\alpha}) = \hat{\theta} \quad \longrightarrow \quad \hat{\alpha} = a(\hat{\theta})$$

MLE of a function is the function of the MLE.

Very useful result.

Functions of MLEs: exponential example

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 MLE estimator for λ?

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 $E[\hat{\lambda}] = \lambda \dfrac{n}{n-1}$. (bias $\to 0$ for $n \to \infty$)

In general MLE for a function of an unbiased estimator stays unbiased only for a linear function.

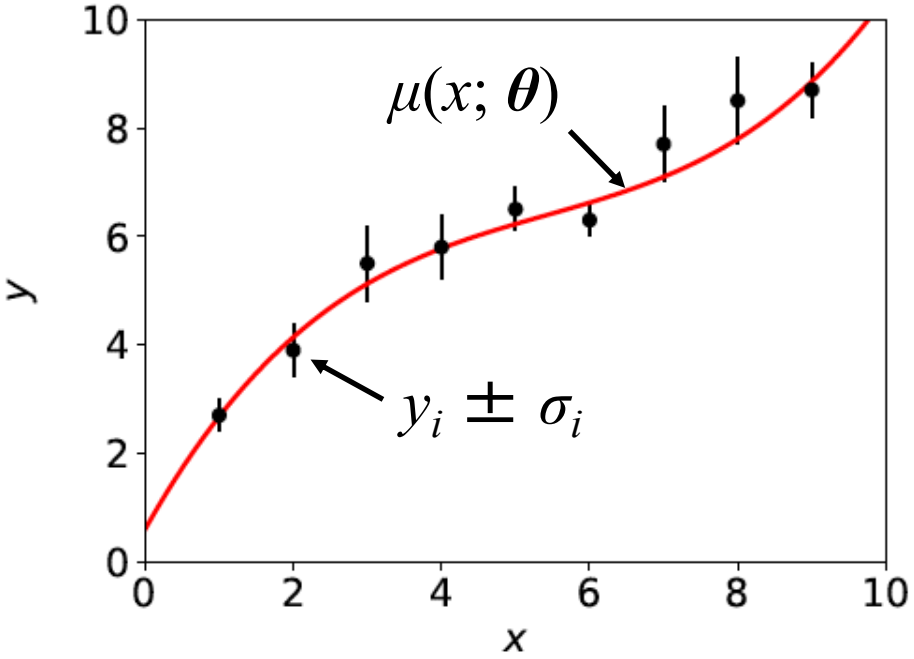
Parameter Estimation 2-4

- The method of Least Squares (LS)
- LS from maximum likelihood
- LS with correlated measurements

Curve fitting: basic idea

Consider N independent measured values y_i, $i = 1,..., N$.

Each y_i has a standard deviation σ_i, and is measured at a value x_i of a control variable x known with negligible uncertainty:



Suppose the functional form of $\mu(x; \theta)$ is given; goal is to estimate its parameters θ .

Gaussian likelihood function → LS estimators

Suppose the measurements $y_1, \ldots, y_N$, are independent Gaussian r.v.s with means $E[y_i] = \mu(x_i; \boldsymbol{\theta})$ and variances $V[y_i] = \sigma_i^2$, so the the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i;\boldsymbol{\theta}))^2/2\sigma_i^2}$$

The log-likelihood function is therefore

$$\ln L(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} + \text{const.}$$

So maximizing the likelihood is equivalent to minimizing

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} = -2\ln L(\boldsymbol{\theta}) + \text{const.}$$

The minimum of $\chi^2(\boldsymbol{\theta})$ defines the least squares (LS) estimators $\hat{\boldsymbol{\theta}}$.

ML ↔ LS

So least-squares (LS) estimators same as maximum likelihood (ML) when the measurements are $y_i \sim \text{Gauss}(\mu(x_i; \boldsymbol{\theta}), \sigma_i)$.

Note that the y_i here are independent but not identically distributed. Do not confuse this case with our previous example of an i.i.d. sample with $x_i \sim \text{Gauss}(\mu, \sigma)$.

If the y_i are not Gaussian distributed the minimum of $\chi^2(\boldsymbol{\theta})$ still defines the LS estimators. But for most applications in practice the y_i are at least approximately Gaussian (a consequence of the Central Limit Theorem).

Often minimize $\chi^2(\boldsymbol{\theta})$, numerically (e.g. programs like curve_fit or MINUIT).

LS with correlated measurements

If $\boldsymbol{y} \sim$ multivariate Gaussian with covariance matrix $V_{ij} = \mathrm{cov}[y_i, y_j]$

$$f(\mathbf{y};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))\right]$$

where $\boldsymbol{\mu}^{\mathrm{T}} = (\mu(x_1),\ldots,\mu(x_N))$, then maximizing the likelihood is equivalent to minimizing

$$\chi^2(\boldsymbol{\theta}) = (\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1}(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\theta}))$$

$$= \sum_{i,j=1}^{N} (y_i - \mu(x_i;\boldsymbol{\theta}))V_{ij}^{-1}(y_j - \mu(x_j;\boldsymbol{\theta}))$$

LS with correlated measurements (2)

For the special case of a diagonal covariance matrix, i.e., uncorrelated measurements. Then

$$V = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \quad \rightarrow \quad V^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & 0 & \cdots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n^2 \end{pmatrix}$$

$V^{-1}{}_{ij} = \delta_{ij}/\sigma_i^2$, carry out one of the sums, $\chi^2(\boldsymbol{\theta})$ same as before:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i,j=1}^{N} (y_i - \mu(x_i;\boldsymbol{\theta}))\frac{\delta_{ij}}{\sigma_i^2}(y_j - \mu(x_j;\boldsymbol{\theta})) = \sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2}$$

Variance of LS estimators for Gaussian data

If $y_i \sim$ Gauss, then we found $\ln L(\boldsymbol{\theta}) = -\frac{1}{2}\chi^2(\boldsymbol{\theta}) + \text{const.}$

To the extent this (approximately) holds, we can use

$$U_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

and so we estimate the inverse covariance matrix with

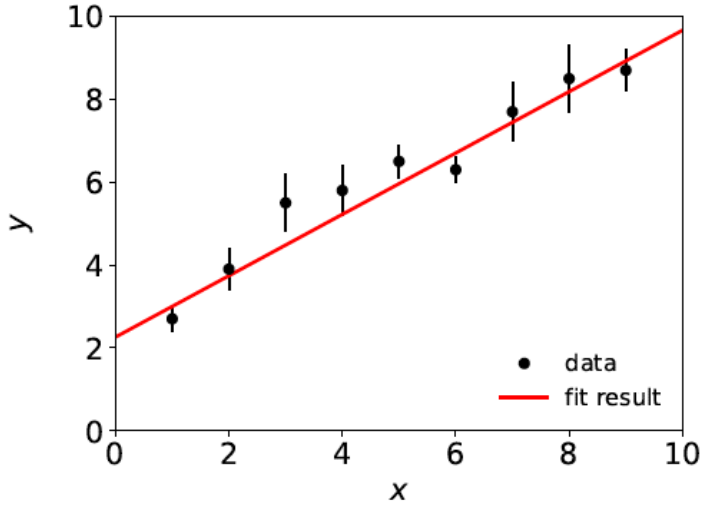
$$\widehat{U}_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \left.\frac{1}{2}\,\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

and invert to estimate the covariance matrix U.

For Gaussian data with the linear LS problem, U is the minimum variance bound (the LS estimators are unbiased and efficient).

Covariance from derivatives of $\chi^2(\boldsymbol{\theta})$

This is what programs like **curve_fit** and **MINUIT** do (derivatives computed numerically). Example with straight-line fit gives:



$$\hat{\theta}_0 = 2.258$$

$$\hat{\theta}_1 = 0.741$$

$$U = \begin{pmatrix} 0.08537 & -0.01438 \\ -0.01438 & 0.003275 \end{pmatrix}$$

$$\sigma_{\hat{\theta}_0} = 0.29\,,$$

$$\sigma_{\hat{\theta}_1} = 0.057\,,$$

$$\mathrm{cov}[\hat{\theta}_0, \hat{\theta}_1] = -0.014\,,$$

$$\rho = -0.86\,.$$

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

If $\mu(x;\, \boldsymbol{\theta})$ is linear in the parameters, then $\chi^2(\boldsymbol{\theta})$ is quadratic:

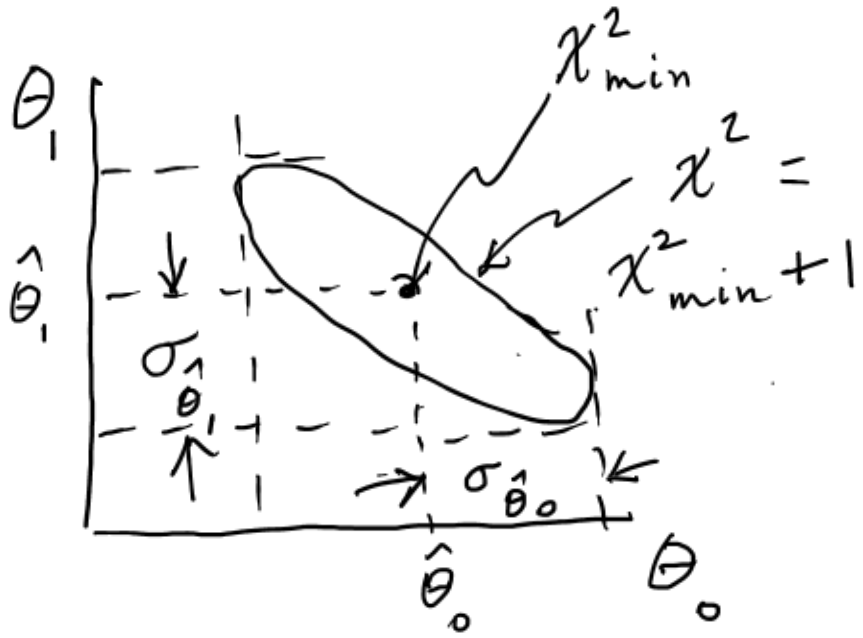
$$\chi^2(\boldsymbol{\theta}) = \chi^2(\hat{\boldsymbol{\theta}}) + \frac{1}{2}\sum_{i,j=1}^{M} \left.\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

$$= \chi^2_{\min} + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T U^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

Standard deviations from tangents to (hyper-) planes of

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

(corresponds to $\ln L(\boldsymbol{\theta}) = \ln L_{\max} - ½$)

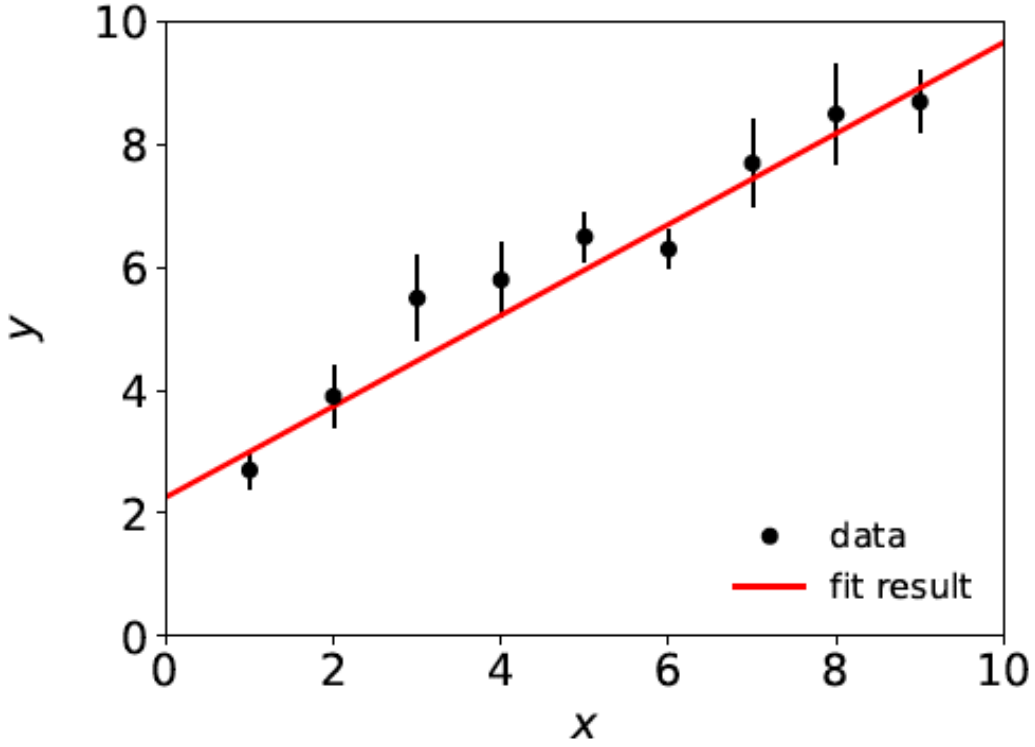


Statistical Data Analysis
Lecture 2-5

- Goodness of fit from χ^2_{min}
- Example of least-squares fit

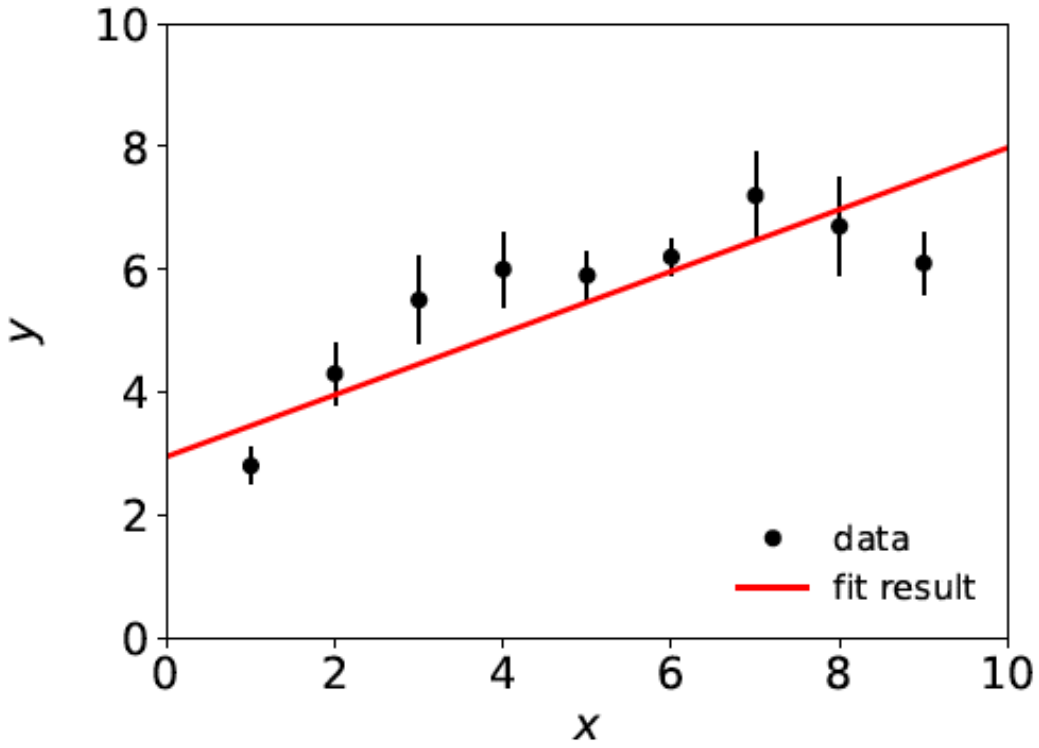
A “good” fit

In an earlier example we fitted data that were reasonably well described by a straight line:



A “bad” fit

But what if a straight-line fit looks like this:



Test hypothesized form of fit function with p-value, if this is below some (user-defined) threshold, reject the hypothesis and try some other function, e.g. a polynomial of higher order.

Goodness-of-fit from χ^2_{min}

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

$$\chi^2_{\text{min}} = \sum_{i=1}^{N} \frac{(y_i - \mu(x_i; \hat{\boldsymbol{\theta}}))^2}{\sigma_i^2} \equiv t(\mathbf{y})$$

It can therefore be used as a goodness-of-fit statistic $t(\boldsymbol{y})$ to test the hypothesized functional form $\mu(x; \boldsymbol{\theta})$.

The p-value of the hypothesized functional form is

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

= the probability, under assumption of $\mu(x; \boldsymbol{\theta})$, to get a χ^2_{min} as high as the one we got or higher.

Distribution of χ^2_{min}

One can show that if the data follow $y \sim \text{Gauss}(\mu(x;\theta),\sigma)$, i.e., if the fit function is correct for some θ, then the statistic $t = \chi^2_{\text{min}}$ follows the chi-square pdf,

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

where the number of degrees of freedom is

n_{d} = number of data points - number of fitted parameters

Note that the composite hypothesis with $E[y] = \mu(x;\theta)$ is only fully specified when we fix θ.

So the p-value is in principle a function of θ, and we should only reject $\mu(x;\theta)$ if $p \leq \alpha$ for all θ.

But here the pdf of our statistic χ^2_{min} is independent of θ, so whatever we get for p holds for any θ.

The "chi-square per degree of freedom"

Recall also the chi-square pdf has an expectation value equal to the number of degrees of freedom, so

$\chi^2_{\text{min}} \sim n_{\text{d}}$ → fit is "good"

$\chi^2_{\text{min}} \gg n_{\text{d}}$ → fit is "bad"

$\chi^2_{\text{min}} \ll n_{\text{d}}$ → fit is better than what one would expect given fluctuations that should be present in the data.

Often this is done using the ratio $\chi^2_{\text{min}}/n_{\text{d}}$, i.e. fit is good if the "chi-square per degree of freedom" comes out not much greater than 1.

But, best to quote both χ^2_{min} and n_{d}, not just their ratio, since e.g.

$\chi^2_{\text{min}} = 15$, $n_{\text{d}} = 10 \rightarrow p$-value = 0.13

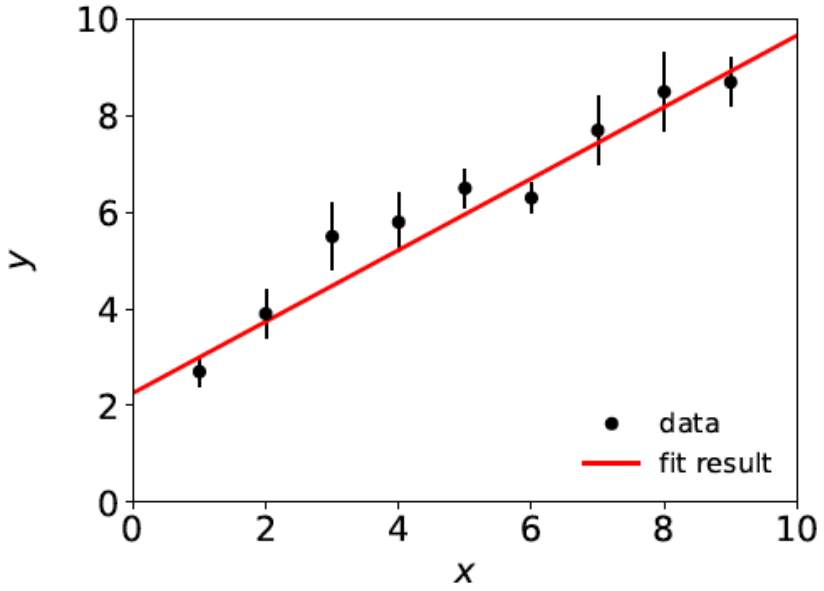
$\chi^2_{\text{min}} = 150$ $n_{\text{d}} = 100 \rightarrow p$-value = 0.00090

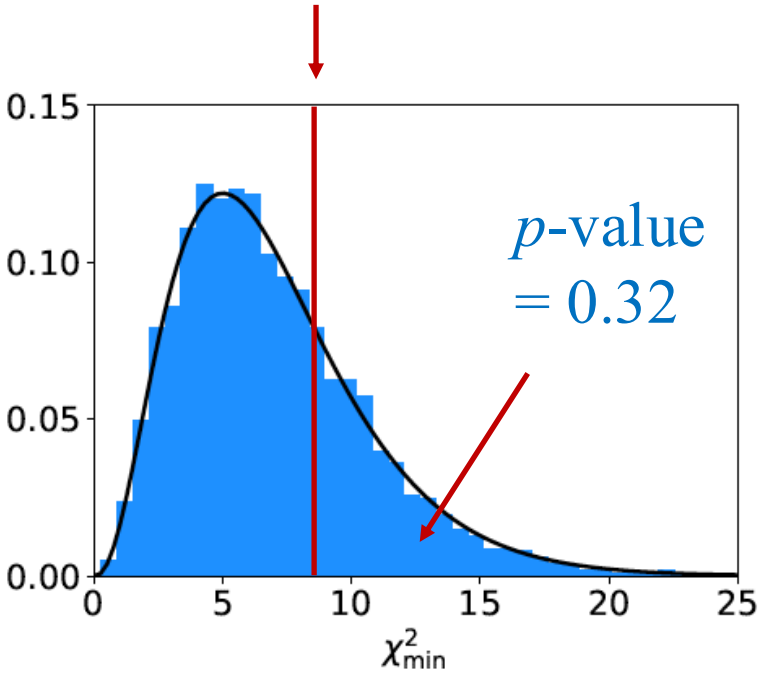
p-value for the "good" fit

$N = 9$ data points, $m = 2$ fitted parameters,

$\chi^2_{\text{min}}/n_{\text{dof}} = 8.2/7 = 1.2$

$\chi^2_{\text{min}} = 8.2$





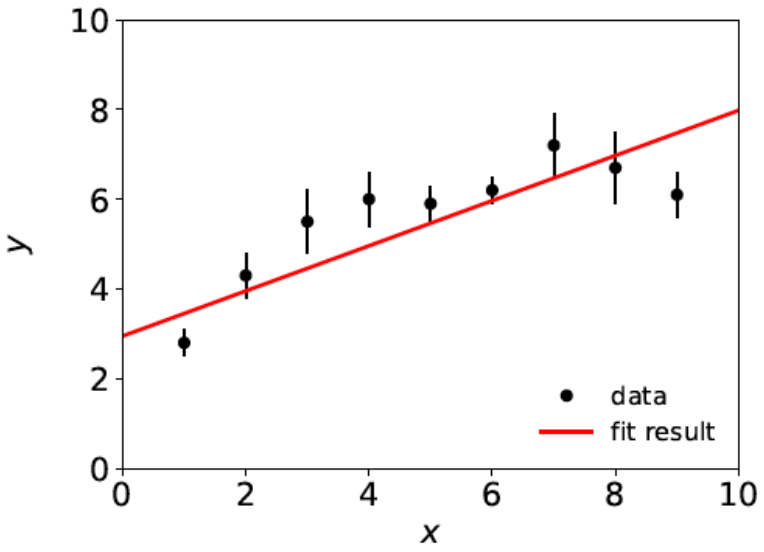
If the straight-line hypothesis is true, expect equal or worse agreement almost 1/3 of the time (i.e. our result is not unusual).

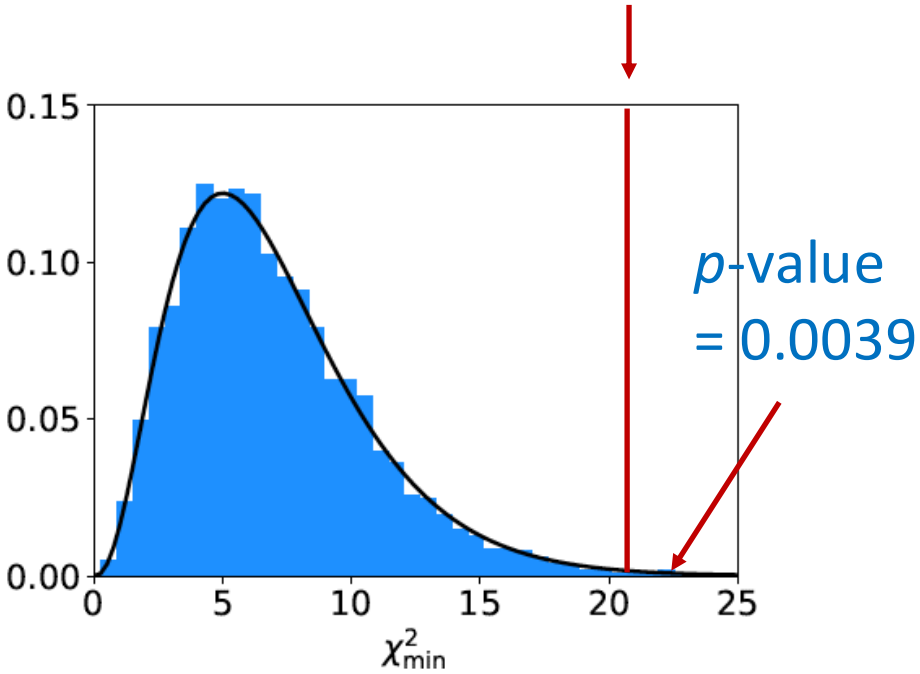
p-value for the "bad" fit

$N = 9$ data points, $m = 2$ fitted parameters,

$\chi^2_{\text{min}}/n_{\text{dof}} = 20.9/7 = 3.0$

$\chi^2_{\text{min}} = 20.9$



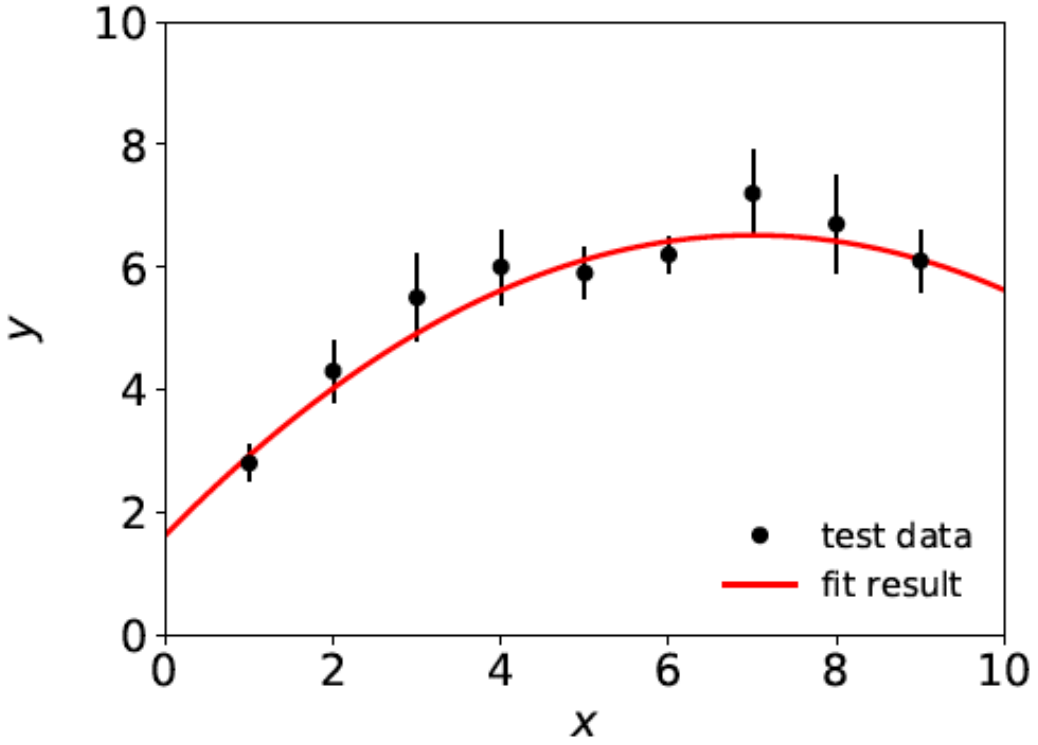


So is the straight-line hypothesis correct? It could be, but if so we would expect a χ^2_{min} as high as observed or higher only 4 times out of a thousand.

A better fit

If we decide the agreement between data and hypothesis is not good enough (exact threshold is a subjective choice), we can try a different model, e.g., a 2nd order polynomial:

$$f(x;\boldsymbol{\theta}) = \theta_0 + \theta_1 x + \theta_2 x^2$$



$\chi^2_{\text{min}} = 3.5$ for $n_{\text{dof}} = 6$

$\chi^2_{\text{min}} / n_{\text{dof}} = 0.58$

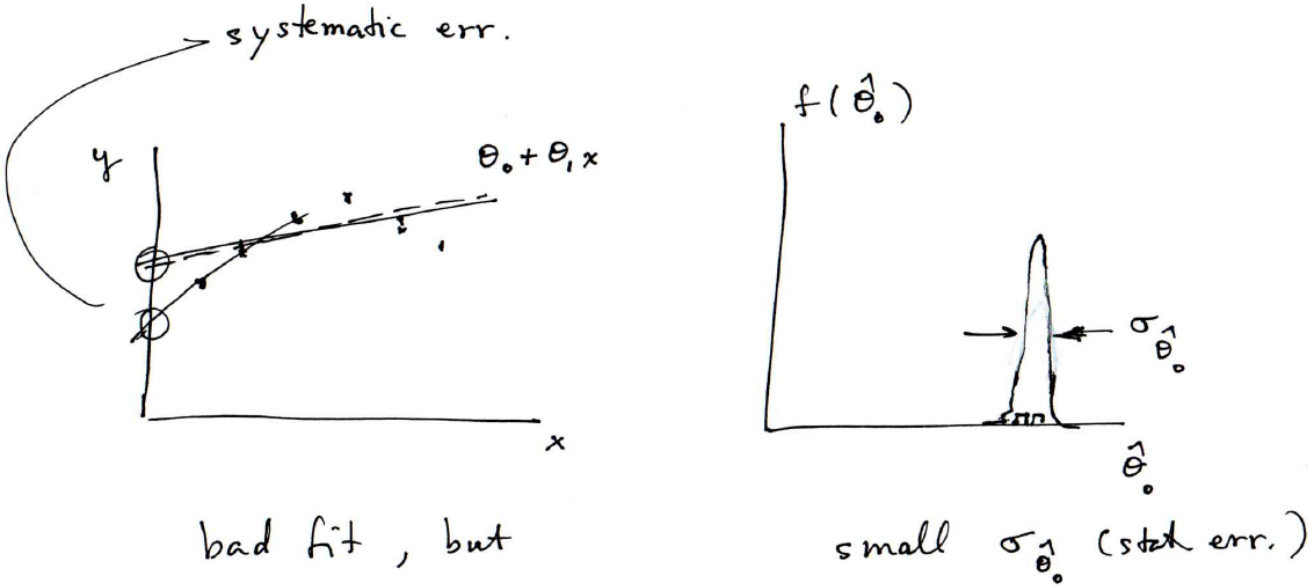
p-value $= 0.75$

Goodness-of-fit vs. statistical errors

If the fit is "bad", something is "wrong" and you may expect large statistical errors for the fitted parameters (std. devs. of estimators). This is not the case.

The statistical errors say how much the parameter estimates should fluctuate when repeating the experiment. This is not the same as the degree to which the fit function can describe the data.

If the hypothesized $\mu(x; \theta)$ is not correct, e.g., the true hypothesis below is curved, not a straight line, then the fitted parameters will have some systematic error – a more complex question that we will take up later.



Extra Slides

LS with histogram data

The fit function in an LS fit is not a pdf, but it could be proportional to one, e.g., when we fit the "envelope" of a histogram.

Suppose for example, we have an i.i.d. data sample of n values $x_1,\ldots, x_n$ sampled from a pdf $f(x;\boldsymbol{\theta})$. Goal is to estimate $\boldsymbol{\theta}$.

Instead of using all n values, put them in a histogram with N bins, i.e., y_i = number of entries in bin i: $\boldsymbol{y} = (y_1,\ldots, y_N)$.

The model predicts mean values:

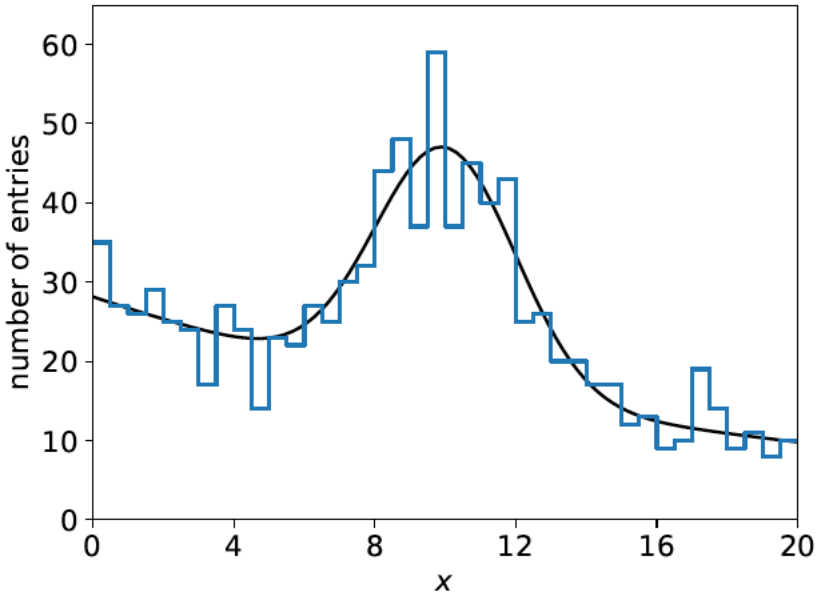
$$E[y_i] = \mu_i(\boldsymbol{\theta})$$

$$= n \int_{\text{bin } i} f(x;\boldsymbol{\theta})\, dx$$

$$\approx n f(x_i;\boldsymbol{\theta})\, \Delta x$$

bin centre

bin width



LS with histogram data (2)

The usual models:

for fixed sample size n, take $\boldsymbol{y} \sim$ multinomial,
if n not fixed, $y_i \sim$ Poisson(μ_i)

Suppose that the expected number of entries in each μ_i are all $\gg 1$ and probability to be in any individual bin $p_i \ll 1$, one can show

$\rightarrow y_i$ indep. and $\sim$ Gauss with $\sigma_i \approx \sqrt{\mu_i}$. ($\rightarrow \sigma_i$ depends on $\boldsymbol{\theta}$).

The (log-) likelihood functions are then

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

$$\ln L(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - \mu_i(\boldsymbol{\theta}))^2}{\sigma_i(\boldsymbol{\theta})^2} - \sum_{i=1}^{N} \ln \sigma_i(\boldsymbol{\theta}) + C$$

LS with histogram data (3)

Still define the least-squares estimators to minimize

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

No longer equivalent to maximum likelihood (equal for $\mu_i \gg 1$).

Two possibilities for σ_i:

$\sigma_i = \sqrt{\mu_i(\boldsymbol{\theta})}$ (LS method)

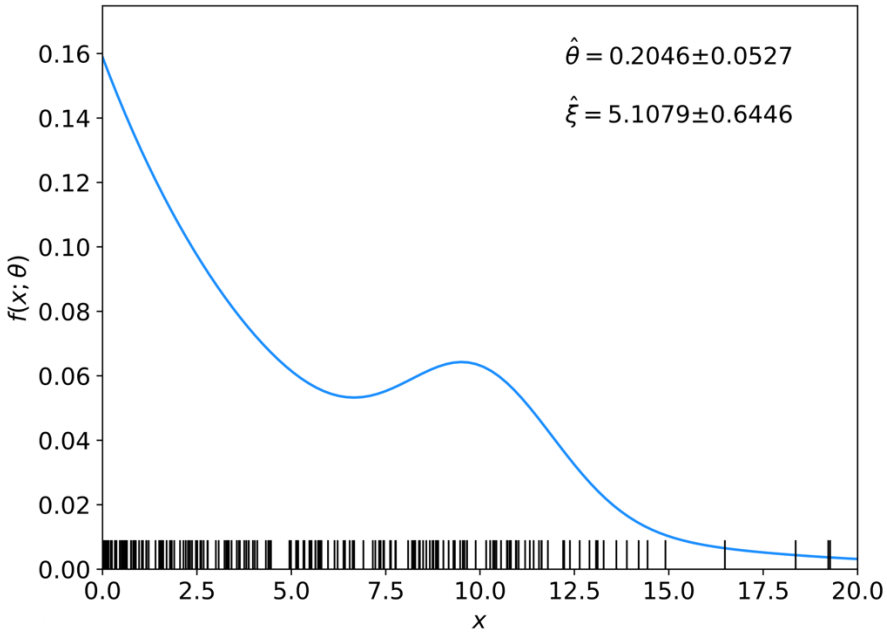
$\sigma_i = \sqrt{y_i}$ (Modified LS method)

Modified LS can be easier computationally but not defined if any $y_i = 0$.

For either method, $\chi^2_{\min} \sim$ chi-square pdf for $\mu_i \gg 1$, but this breaks down for when the μ_i are not large.

Comments on using iminuit

In our earlier iminuit example mlFit.py, the only argument of the log-likelihood function was the parameter array, and the data array xData entered as global (usually not a good idea):



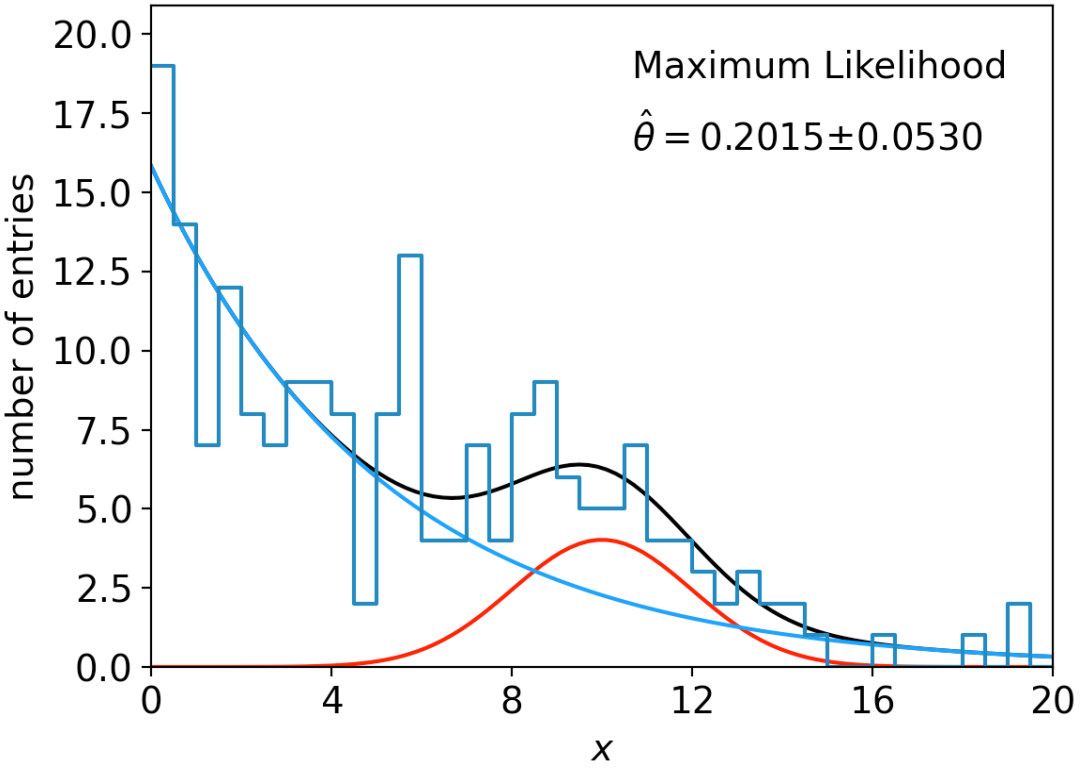

```
def negLogL(par):
    pdf = f(xData, par)
    return -np.sum(np.log(pdf))

    ⋮

m = Minuit(negLogL, par, name=parname)
```

ln$L$ in a class, binned data,...

Sometimes it is convenient to have the function being minimized as a method of a class. An example of this is shown in the program histFit.py, which does the same fit as in mlFit.py but with a histogram of the data:



Commentary on histFit.py

The global data can be avoided if we make the objective function a method of a class:

```
class ChiSquared:                    # function to be minimized

    def __init__(self, xHist, bin_edges, fitType):
        self.setData(xHist, bin_edges)
        self.fitType = fitType

    def setData(self, xHist, bin_edges):
        numVal = np.sum(xHist)
        numBins = len(xHist)
        binSize = bin_edges[1] - bin_edges[0]
        self.data = xHist, bin_edges, numVal, numBins, binSize

    def chi2LS(self, par):            # least squares
        xHist, bin_edges, numVal, numBins, binSize = self.data
        xMid = bin_edges[:numBins] + 0.5*binSize
        binProb = f(xMid, par)*binSize
        nu = numVal*binProb
        sigma = np.sqrt(nu)
        z = (xHist - nu)/sigma
        return np.sum(z**2)
```

class ChiSquared (continued)

```
def chi2M(self, par):                # multinomial maximum likelihood
    xHist, bin_edges, numVal, numBins, binSize = self.data
    xMid = bin_edges[:numBins] + 0.5*binSize
    binProb = f(xMid, par)*binSize
    nu = numVal*binProb
    lnL = 0.
    for i in range(len(xHist)):
        if xHist[i] > 0.:
            lnL += xHist[i]*np.log(nu[i]/xHist[i])
    return -2.*lnL

def __call__(self, par):
    if self.fitType == 'LS':
        return self.chi2LS(par)
    elif self.fitType == 'M':
        return self.chi2M(par)
    else:
        print("fitType not defined")
        return -1
```

Using the ChiSquared class

```
# Put data values into a histogram
numBins=40
xHist, bin_edges = np.histogram(xData, bins=numBins, range=(xMin, xMax))
binSize = bin_edges[1] - bin_edges[0]

# Initialize Minuit and set up fit:
parin   = np.array([theta, mu, sigma, xi])      # initial values (here = true)
parname = ['theta', 'mu', 'sigma', 'xi']
parstep = np.array([0.1, 1., 1., 1.])           # initial setp sizes
parfix  = [False, True, True, False]            # change to fix/free param.
parlim  = [(0.,1), (None, None), (0., None), (0., None)]
chisq = ChiSquared(xHist, bin_edges, fitType)
m = Minuit(chisq, parin, name=parname)
m.errors = parstep
m.fixed = parfix
m.limits = parlim
m.errordef = 1.0                        # errors from chi2 = chi2min + 1
```

For full program see
https://www.pp.rhul.ac.uk/~cowan/stat/exercises/fitting/python/

History

Least Squares fitting also called “regression”

F. Galton, *Regression towards mediocrity in hereditary stature*, The Journal of the Anthropological Institute of Great Britain and Ireland. 15: 246–263 (1886).

Developed earlier by Laplace and Gauss:

C.F. Gauss, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambentium*, Hamburgi Sumtibus Frid. Perthes et H. Besser Liber II, Sectio II (1809);

C.F. Gauss, *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*, pars prior (15.2.1821) et pars posterior (2.2.1823), Commentationes Societatis Regiae Scientiarium Gottingensis Recectiores Vol. V (MDCCCXXIII).

Linear LS Problem

Suppose the fit function is linear in the parameters $\boldsymbol{\theta}^{\mathrm{T}} = (\theta_1, \ldots, \theta_M)$,

$$\mu(x; \boldsymbol{\theta}) = \sum_{i=1}^{M} \theta_i a_i(x)$$

where the $a_i(x)$ are a set of linearly independent basis functions, and write $\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta}) = (\mu(x_1; \boldsymbol{\theta}), \ldots, \mu(x_N; \boldsymbol{\theta}))$.

Define $N \times M$ matrix $A_{ij} = a_j(x_i)$, so $\boldsymbol{\mu}(\boldsymbol{\theta}) = A\boldsymbol{\theta}$.

To find the LS estimators minimize:

$$\chi^2(\boldsymbol{\theta}) = (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

$$= (\mathbf{y} - A\boldsymbol{\theta})^T V^{-1} (\mathbf{y} - A\boldsymbol{\theta})$$

Linear LS Problem (2)

Set derivatives with respect to θ_i to zero,

$$\nabla\chi^2(\boldsymbol{\theta}) = -2(A^T V^{-1}\mathbf{y} - A^T V^{-1} A\boldsymbol{\theta}) = 0$$

$$\nabla = \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_M}\right)$$

Solve system of *M* linear equations to find the LS estimators,

$$\hat{\boldsymbol{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1}\mathbf{y} \equiv B\mathbf{y}$$

Note that the estimators are linear functions of the measured y_i.

Bias of LS estimators

By hypothesis $E[\boldsymbol{y}] = \boldsymbol{\mu} = A\boldsymbol{\theta}$ so for the linear problem, the LS estimators are unbiased:

$$E[\hat{\boldsymbol{\theta}}] = (A^T V^{-1} A)^{-1} A^T V^{-1} E[\mathbf{y}]$$

$$= (A^T V^{-1} A)^{-1} A^T V^{-1} \boldsymbol{\mu}$$

$$= (A^T V^{-1} A)^{-1} A^T V^{-1} A\boldsymbol{\theta} = \boldsymbol{\theta}$$

For the general nonlinear problem the LS estimators can have a bias.

Variance of LS estimators for linear problem

For the linear LS problem, the variance can be found using error propagation. Using

$$V_{ij} = \mathrm{cov}[y_i, y_j]$$

$$\hat{\boldsymbol{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \mathbf{y} \equiv B\mathbf{y}$$

$$U_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$$

We find

$$U = BVB^T = (A^T V^{-1} A)^{-1}$$

Since the estimators are linear in the y_i, error propagation gives an exact result.

Cheap estimator for mass of W boson

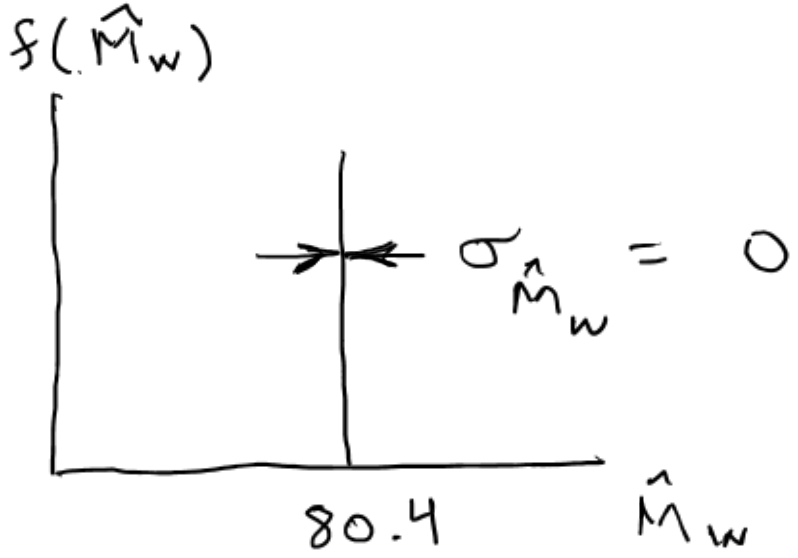
The Particle Physics community has spent huge sums trying to estimate the mass of the W boson with the smallest possible statistical and systematic uncertainty.

Here is an estimator with zero statistical uncertainty. And it's free!

$$\widehat{M_W} = 80.4\,\mathrm{GeV}$$

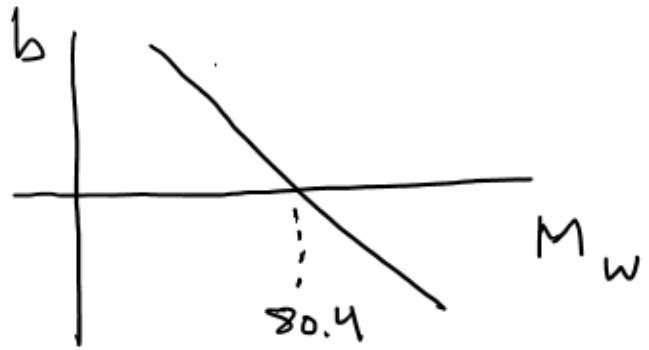
Here is its sampling distribution:

Does this violate the information inequality?



Cheap estimator for mass of W boson (2)

This estimator's bias is $b = E[\widehat{M_W}] - M_W = 80.4\,\text{GeV} - M_W$



Note current best estimate of M_W is 80.379±0.012 GeV, so the numerical value of the bias may be fairly small.

But we have $\dfrac{\partial b}{\partial M_W} = -1$ and so

$$\text{MVB} = -\left(1 + \frac{\partial b}{\partial M_W}\right)^2 \Bigg/ E\left[\frac{\partial^2 \ln L}{\partial M_W^2}\right] = 0$$

So the information inequality is still satisfied.

Example of MLE: parameters of Gaussian pdf

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

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

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 μ, σ^2 to zero and solve,

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

We already know that the estimator for μ is unbiased.

But we find, however, $E[\widehat{\sigma^2}] = \frac{n-1}{n}\sigma^2$, so the MLE for σ^2 has a bias, but $b \to 0$ for $n \to \infty$. Recall, however, that

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

is an unbiased estimator for σ^2. Usually not important whether one uses s^2 or the MLE to estimate σ^2.

Example of ML: parameters of Gaussian pdf (3)

Use 2nd derivatives of $\ln L$ to find covariance.

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2} \quad \longrightarrow \quad E\left[\frac{\partial^2 \ln L}{\partial \mu^2}\right] = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\longrightarrow \quad E\left[\frac{\partial^2 \ln L}{\partial (\sigma^2)^2}\right] = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n} E[(x_i - \mu)^2] = -\frac{n}{2\sigma^4}$$

$$\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) \longrightarrow E\left[\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu}\right] = -\frac{1}{\sigma^4}\sum_{i=1}^{n} E[x_i - \mu] = 0$$