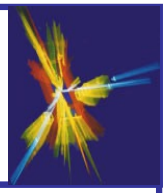


Parameter estimation: general concepts

Hypothesis testing:

Theory —Probability / Calculus→ Data

Given predictions, what can one say about data?

Given data, what can one say about parameters or properties as well as about correctness of predictions?

Parameter estimation:

Theory ←Statistical / Inference— Data

estimator = procedure giving a value for a parameter or a property of distribution (pdf) from actual data values

notation: estimator for θ is $\hat{\theta}$ (a **hat** indicates estimator)

estimate = observed value of an estimator (often $\hat{\theta}_{\text{obs}}$)

How does one construct an estimator $\hat{\theta}(\bar{x})$?

Exists no golden rule how to construct an estimator!

Examples of estimators are arithmetic mean & variance:

$$\hat{\mu}(\{x\}) = \frac{1}{N}\sum_i x_i \qquad \hat{V}(\{x\}) = \frac{1}{N-1}\sum_i (x_i - \hat{\mu})^2$$

N.B. $\hat{\theta}(\bar{x})$ function of random variables & random variable itself, characterized by a pdf $g(\hat{\theta};\theta,n)$, which depends on (true value of) θ & has expectation value, variance, etc...

Properties of estimators

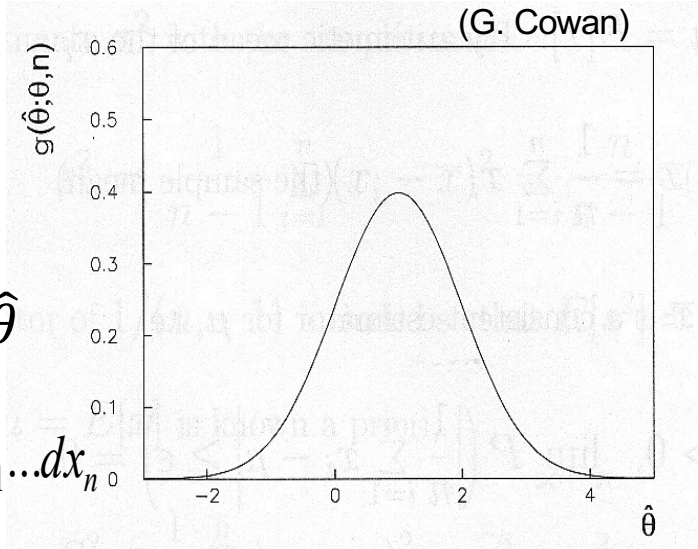
Often start by requiring **consistency**: $\lim_{n\to\infty} \hat{\theta} = \theta$ i.e. as sample size increases, estimate converges to true value:

$$\text{for any } \varepsilon > 0, \quad \lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

NB! convergence in sense of probability, i.e. no guaranty that any particular $\hat{\theta}_{\text{obs}}$ will be within given distance of θ.

$g(\hat{\theta};\, \theta, n)$ is the pdf of $\hat{\theta}$ for a fixed sample size n.

(G. Cowan)



Expectation value of $\hat{\theta}$:

$$E[\hat{\theta}] = \int \hat{\theta}\, g(\hat{\theta};\theta,n)\, d\hat{\theta}$$

$$\int \ldots \int \hat{\theta}(\bar{x})\, f(x_1;\theta) \ldots f(x_n;\theta)\, dx_1 \ldots dx_n$$

variance $V[\hat{\theta}] = \sigma^2_{\hat{\theta}}$

$\sigma_{\hat{\theta}}$ = "statistical" uncertainty
b = "systematic" uncertainty (due to construction of $\hat{\theta}$)

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

For most estimators: $\sigma_{\hat{\theta}} \propto 1/\sqrt{n}$, $b \propto 1/n$

Good estimator: **consistent, unbiased** ($E[\hat{\theta}] = \theta$) and **efficient** (i.e. has minimal possible variance = "RCF bound").

RCF bound: $$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Bigg/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] \quad (b = \text{bias})$$

Also "CR (Cramér-Rao) bound" or "information inequality". L = likelihood function (defined on slide 5), F = Fréchet.

Estimator for the mean

Consider n measurements of random variable x, x_1,..., x_n.
Arithmetic mean a natural choice as estimator of $\mu = E[x]$:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(= the sample mean)}$$

If $V[x]$ finite, then $\bar{x}$ is a consistent estimator for μ, i.e.

$$\text{for any } \varepsilon > 0, \quad \lim_{n\to\infty} P(|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu| > \varepsilon) = 0$$

i.e. the **weak law of large numbers**. Expection value of $\bar{x}$:

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

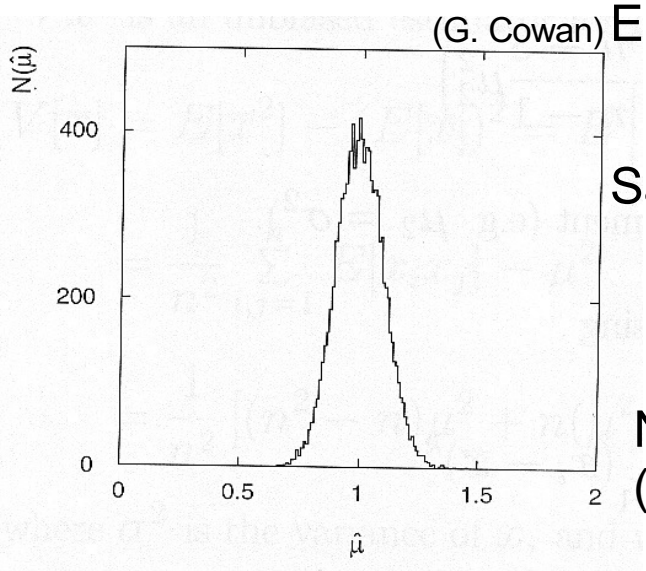
$\rightarrow$ $\bar{x}$ is an unbiased estimator for μ. The variance of $\bar{x}$ is

$$V[\bar{x}] = E[\bar{x}^2] - (E[\bar{x}])^2 = \frac{1}{n^2}\sum_{i,j=1}^{n} E[x_i x_j] - \mu^2 = \frac{\sigma^2}{n},$$

σ^2 = variance of x, $E[x_i x_j] = \mu^2$ for $i \neq j$ and $E[x_i^2] = \mu^2 + \sigma^2$.

Example of estimator for mean: take samples of n = 100 values from a Gaussian MC generator with μ = 1 & σ^2 = 1. Calculate sample mean & repeat procedure many times. Enter values into a histogram.

(G. Cowan)



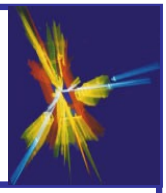
$$\bar{\hat{\mu}} = 0.9981 \quad (\hat{\mu} \text{ unbiased})$$

Sample standard deviation of

$$\hat{\mu} \text{ values} = 0.0995 \approx \frac{\sigma}{\sqrt{n}}$$

NB! pdf of $\hat{\mu}$ $\approx$ Gaussian (result of central limit theorem).

Estimators for the variance



Suppose mean μ and variance $V[x] = \sigma^2$ both unknown.
Then estimate σ^2 using sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}(\overline{x^2} - \bar{x}^2)$$

Factor $1/(n-1)$ introduced to have $E[s^2] = \sigma^2$ (unbiased).
If mean $\mu = E[x]$ known then estimate σ^2 using statistic S^2

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \overline{x^2} - \mu^2$$

also $E[S^2] = \sigma^2$ (i.e. unbiased estimator).

Variance of $s^2(S^2)$ calculated with k^{th} central moments μ_k.
μ_k's estimated from corresponding estimator m_k or M_k.

$$V[s^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right) \qquad V[S^2] = \frac{1}{n}\left(\mu_4 - \mu_2^2\right)$$

$$\hat{\mu}_k = m_k = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^k \qquad \hat{\mu}_k = M_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^k$$

A natural estimator for standard deviation, σ, then

$$\hat{\sigma} = s = \sqrt{s^2} \qquad \left(\text{or in case } \mu \text{ known } \hat{\sigma} = S = \sqrt{S^2}\right)$$

For variance of estimator for standard deviation:

$$V[\hat{s}^2] = \left(d\sigma^2/d\sigma\right)^2 V[\hat{\sigma}] = 4\sigma^2 V[\hat{\sigma}] \Rightarrow V[\hat{\sigma}] = V[\hat{s}^2]/4\sigma^2$$

For a Gaussian pdf $\hat{s}^2(\sigma) = \sigma^2 \Longrightarrow d\hat{s}^2/d\sigma = 2\sigma$

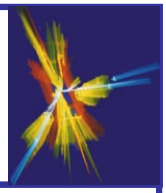
$$E[\hat{\mu}_4] = 3\sigma^4 \Rightarrow V[s^2] = 2\sigma^4/(n-1) \Rightarrow \sigma_s = \sigma/\sqrt{2(n-1)}$$

Another quality measure of estimator: **mean square error**

$$MSE = E[(\hat{\theta} - \theta)^2] = V[\hat{\theta}] + b^2$$

(used as measure in e.g. unfolding methods)

Maximum likelihood method



Random variable x distributed according to pdf $f(x,\theta)$.
Assume functional form of f known but not parameter θ.
maximum likelihood method (suurimman uskottavuuden menetelmä) technique for estimating θ from a data sample.

If $f(x,\theta)$ correct pdf hypothesis, then

$P(x_i \text{ found in } [x_i, x_i+dx_i] \text{ for all } i) = \prod_{i=1}^{n} f(x_i,\theta)\, dx_i$

If hypothesis (including value of θ) correct (= true)
$\rightarrow$ expect higher probability for the data
If hypothesized functional form wrong or θ value far away
$\rightarrow$ expect lower probability for the data
$\Rightarrow$ higher value of **likelihood function** close to true θ

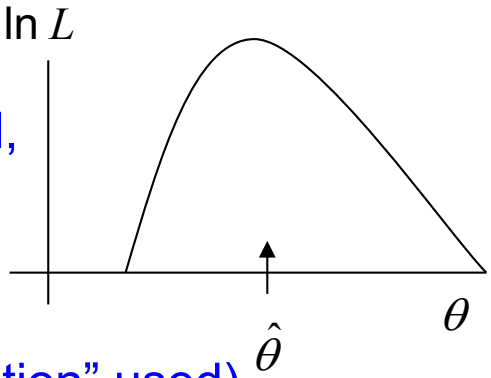
$$L(\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

NB! $L(\theta) = f_{sample}(\bar{x};\theta)$, but $L(\theta)$ regarded only a function of θ, measurements x_i's constants, "experiment" finished.

Define ML estimator $\hat{\theta}$ as value of θ that maximizes $L(\theta)$.
For m parameters, usually find solution $\hat{\theta}_1, \ldots, \hat{\theta}_m$ by solving

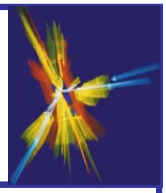
$$\frac{\partial L}{\partial \theta_i} = 0, \quad i = 1,\ldots,m$$

In practice maximize ln $L(\theta)$ instead,
Can then add individual ln $P(x_i)$'s.
ln $L(\theta)$ might have more than one local maximum $\rightarrow$ take highest one.



N.B. No binning of data ("all information" used).
N.B. Definition of ML estimators don't guarantee optimality
$\rightarrow$ investigate properties such as bias, variance ...
In most cases, especially for sufficient large data samples, ML estimators generally most optimal estimator choice.

Example of ML estimator



Suppose proper decay times of a certain type of unstable states measured for n decays, t_1, ..., t_n. Choose as hypothesis for t distribution an exponential pdf with mean τ.

$$f(t;\tau) = e^{-t/\tau}/\tau$$

Task to estimate value of τ. Use **log-likelihood function** instead to find parameter value giving maximum value for function. Equivalent since logarithm a monotonic function ($\rightarrow$ maximum at same value). In addition, products in L becomes sums in ln L and exponentials becomes factors.

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

set $\dfrac{\partial \ln L}{\partial \tau} = 0$ and solve for $\tau \quad \rightarrow \quad \hat{\tau} = \dfrac{1}{n}\sum_{i=1}^{n} t_i$

How find out whether $\hat{\tau}$ is an unbiased estimator for τ?

i) Find pdf $g(\hat{\tau};\tau,n)$ and compute $b = E[\hat{\tau}] - \tau$

ii) Compute $E[\hat{\tau}(t_1,\ldots t_n)] = \int\ldots\int \hat{\tau}(\bar{t})\, f_{\text{joint}}(\bar{t};\tau)\, dt_1 \ldots dt_n =$

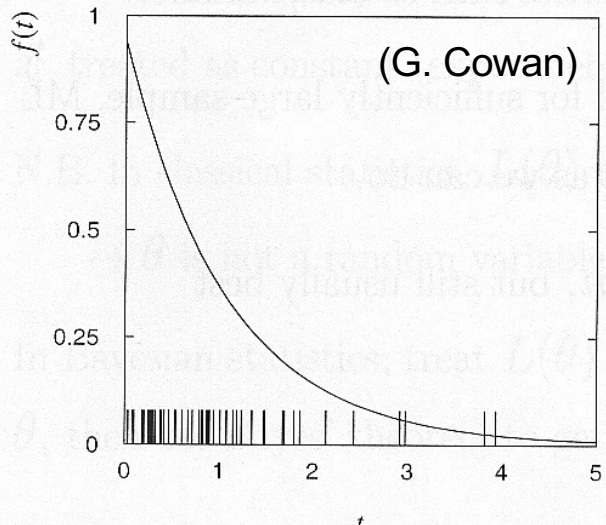
$$\int\ldots\int \left(\sum_{i=1}^{n} t_i\right) \frac{e^{-t_1/\tau}}{\tau} \ldots \frac{e^{-t_n/\tau}}{\tau} \frac{dt_1 \ldots dt_n}{n} = \frac{1}{n}\sum_{i=1}^{n} \left(\int \frac{t_i}{\tau} e^{-t_i/\tau} dt_i \prod_{j\neq i} \frac{e^{-t_j/\tau} dt_j}{\tau} \right)$$

$$= \sum_{i=1}^{n} \tau/n = \tau \quad \rightarrow \quad \hat{\tau} \text{ unbiased estimator for } \tau \text{ !!}$$

iii) Could make same conclusion without any calculation based on the fact that the sample mean an unbiased estimator for $E[t]$ and for exponential pdf $E[t] = \tau$.

Suppose that one is interested in decay constant $\lambda = \ln 2 / \tau$ instead of mean lifetime τ. ML estimator for λ?

Example of ML estimators (cont.)



A sample of 50 observations of proper time, t, ("ticks" on x-axis), generated using MC assuming exponential distribution with mean τ = 1.0. Curve result of a maximum likelihood fit to observations, giving $\hat{\tau} = 1.062$.

Given a function $a(\theta)$ of some parameter θ, one has

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial a}\frac{\partial a}{\partial \theta} = 0 \quad \Rightarrow \quad \left.\frac{\partial L}{\partial a} = 0\right|_{a=a(\theta)} \quad \text{unless} \quad \frac{\partial a}{\partial \theta} = 0$$

So a maximizing $L_a(a)$ is $a(\hat{\theta})$, where $\hat{\theta}$ maximizes $L_\theta(\theta)$.

$\rightarrow$ ML estimator of function $a(\theta)$ is $\hat{a} = a(\hat{\theta})$

This is called invariance. ML estimators are invariant.

So for decay constant, one gets $\hat{\lambda} = \dfrac{\ln 2}{\hat{\tau}} = \ln 2 \cdot n \Big/ \sum_{i=1}^{n} t_i$

Is $\hat{\lambda}$ an unbiased estimator for λ ?

For $\hat{\lambda}$ one can show that $E[\hat{\lambda}] = \dfrac{n\lambda}{n-1} = \dfrac{\ln 2}{\tau}\dfrac{n}{n-1}$

$\rightarrow$ $\hat{\lambda}$ has a bias that goes to zero for $n \rightarrow \infty$.

Above true for ML estimators ($b \rightarrow 0$, when $n \rightarrow \infty$).

Example where ML fails: assume taxis numbered 1 to N_{taxi}; ML estimator for N_{taxi} from m taxi number observations?

$$f(n) = \frac{1}{(N_{\text{taxi}} - 1)} \Rightarrow L = \frac{1}{(N_{\text{taxi}} - 1)^m} \Rightarrow \frac{\partial \ln L}{\partial N_{\text{taxi}}}$$

no local maxima so $\hat{N}_{\text{taxi}}$ undefined.

Ansatz: $\hat{N}_{\text{taxi}} = 2\bar{n} - 1$, where $\bar{n}$ = mean; $E\left[\hat{N}_{\text{taxi}}\right] = N_{\text{taxi}}$.

Example of ML estimators (cont.)

n measurement of a variable x assumed to be Gaussian distributed with unknown μ & σ. Log-likelihood function:

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

set $\dfrac{\partial \ln L}{\partial \mu} = 0$ & $\dfrac{\partial \ln L}{\partial \sigma^2} = 0$ and solve for μ & σ^2 $\rightarrow$

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

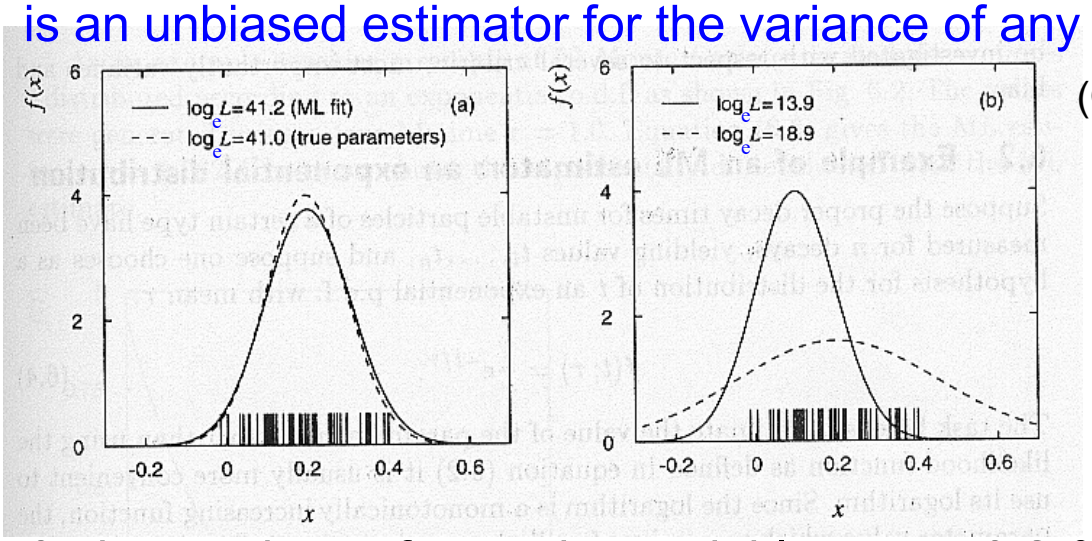
Already known that $\hat{\mu}$ is an unbiased estimator for μ.

What about $\widehat{\sigma^2}$? $\quad E[\hat{\sigma}^2] = \dfrac{n-1}{n}\sigma^2$

So ML estimator for σ^2 has a bias, but $b \rightarrow 0$ for $n \rightarrow \infty$.
Recall, however, that the sample variance

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

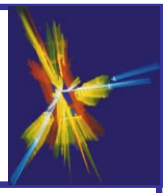
is an unbiased estimator for the variance of any pdf.



(G. Cowan)

50 observations of gaussian variable x; μ_x = 0.2 & σ_x = 0.1.
(a) pdf of parameters maximizing ln L & true parameters.
(b) pdf of parameters far from true ones $\rightarrow$ low ln L values.

What about statistical uncertainty of ML estimates?



Variance of ML estimators: **the analytic method**

A direct way of estimating uncertainty of estimate is to compute variance of estimator, e.g. when $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$ i.e. width of the pdf $g(\hat{\tau};\, \tau, n)$:

$$V[\hat{\tau}] = E[\hat{\tau}^2] - (E[\hat{\tau}])^2 = \int\!\ldots\!\int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)^2 \frac{e^{-t_1/\tau}}{\tau}\ldots\frac{e^{-t_n/\tau}}{\tau}\, dt_1 \ldots dt_n - (E[\hat{\tau}])^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\left(\int (t_i^2 + t_i t_j)\frac{e^{-t_i/\tau}}{\tau}dt_i \frac{e^{-t_j/\tau}}{\tau}dt_j \prod_{k\neq i,j}\int \frac{e^{-t_k/\tau}}{\tau}dt_k\right) - (E[\hat{\tau}])^2$$

$$= \frac{\tau^2(2n + n(n-1))}{n^2} - \tau^2 = \frac{\tau^2}{n} \quad \rightarrow \quad V[\hat{\tau}]$$

n times smaller than $V[t]$

(in fact this result was obvious, since here $\hat{\tau} = \bar{t}$)

N.B. $V[\hat{\tau}]$ & $\sigma_{\hat{\tau}}$ are functions of true (& unknown!) τ.

Estimate standard deviation using $\widehat{\sigma_{\hat{\tau}}} = \hat{\tau}/\sqrt{n}$

Estimated standard deviation often quoted as ”statistical error” of a measurement e.g. $\hat{\tau} \pm \widehat{\sigma_{\hat{\tau}}} = 1.062 \pm 0.150$

should be interpreted as: ML estimate for τ is 1.062.
ML estimate for σ of $g(\hat{\tau};\, \tau, n)$ is 0.150.

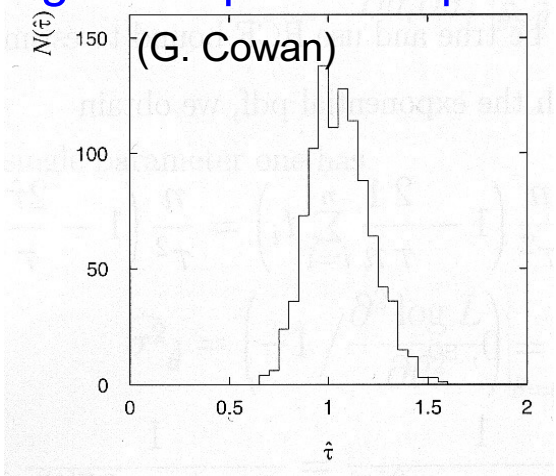
If pdf $g(\hat{\tau};\, \tau, n)$ Gaussian, then $[\hat{\tau} - \widehat{\sigma_{\hat{\tau}}}, \hat{\tau} + \widehat{\sigma_{\hat{\tau}}}]$ equivalent to ”68.3 % confidence interval” for $\hat{\tau}$, generally accepted way to quote uncertainty even when errors non-Gaussian.

NB! very seldom variance explicitly computable as above!!

Variance of ML estimators: **the Monte Carlo method**

Cases too difficult to solve analytically (or $g(\hat{\tau};\ \tau, n)$ not known), ML estimator distribution investigated with MC. Simulate large number of **pseudoexperiments**, compute ML estimates each time, resulting distribution $\approx\ g(\hat{\tau};\ \tau, n)$

Use experimental μ as "true" value & MC to get distribution of sample means, width = unbiased variance estimator. E.g. for exponential pdf $\hat{\tau}$ = 1.062, used as "true" MC value.



A histogram with ML estimates from 1000 MC experiments with 50 observations each time. MC used τ = 1.062 as true lifetime. Calculated standard deviation of histogram entries, s, is 0.151.

Similar to analytical estimate $\hat{\tau}/\sqrt{n} = 1.062/\sqrt{50} = 0.150$

NB! $g(\hat{\tau};\ \tau, n)$ approx. Gaussian ($\Leftrightarrow$ central limit theorem) $\rightarrow$ true in general for ML estimators in large sample limit.

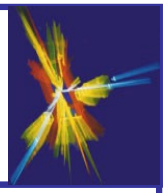
Variance of ML estimators: **the RCF bound**

A lower bound on variance of any estimator (not just ML)

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

Rao-Cramer-Frechet bound (or "information inequality"). If equality true, then corresponding estimator **efficient.** ML estimators efficient in large sample limit. So, assume estimator efficient & use RCF bound to estimate $V[\hat{\theta}]$.

Variance of ML estimators



For the example with exponential pdf, one obtains

$$\frac{\partial^2 \ln L}{\partial \tau^2} = \frac{n}{\tau^2}\left(1 - \frac{2}{n\tau}\sum_{i=1}^{n} t_i\right) = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right) \quad \& \quad b = 0 \quad \text{so}$$

$$V[\hat{\tau}] \geq \left(E\left[-\frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)\right]\right)^{-1} = \left(-\frac{n}{\tau^2}\left(1 - \frac{2E[\hat{\tau}]}{\tau}\right)\right)^{-1} = \frac{\tau^2}{n}$$

same variance as obtained from analytical calculation
$\rightarrow$ ML $\hat{\tau}$ an efficient estimator for parameter τ for any n.

For $\bar{\theta} = (\theta_1, \ldots, \theta_m)$ with efficient estimators and zero bias,

$$(V^{-1})_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] \Rightarrow \boxed{(\hat{V^{-1}})_{ij} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\bar{\theta}=\bar{\hat{\theta}}}}$$

Impractical to compute RCF bound analytically. In case of sufficiently large data sample, estimate V^{-1} by evaluating 2nd derivate at the ML estimates with the measured data.

Procedure: 1st numerically maximize ln L, then determine matrix of 2nd derivates using finite differences evaluated at ML estimates, finally invert result to find covariance matrix.

Variance of ML estimators: **the graphical method**

Extension of RCF bound technique leads to a graphical technique for obtaining the variance of ML estimators.
Expand ln $L(\theta)$ around ML estimate $\hat{\theta}$ of parameter θ:

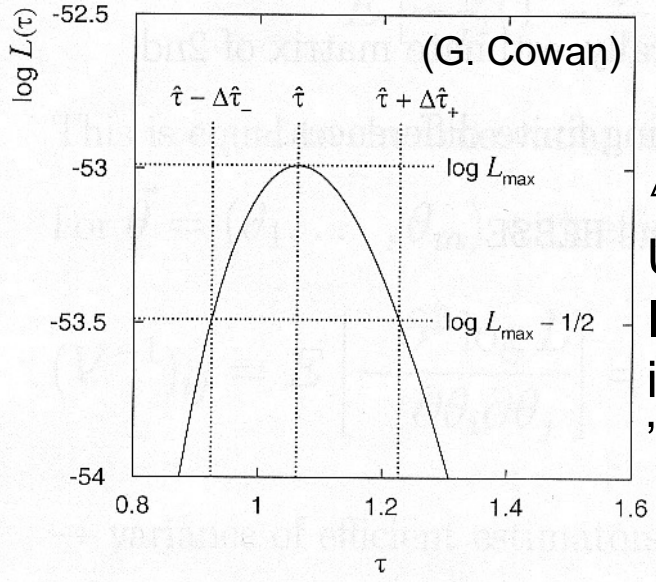
$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[\frac{\partial \ln L}{\partial\theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial\theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

$$\text{now } \ln L(\hat{\theta}) = \ln L_{\max} \quad \& \quad \left[\frac{\partial \ln L}{\partial\theta}\right]_{\theta=\hat{\theta}} = 0 \quad \Rightarrow$$

Variance of ML estimators

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

1 (2) standard deviation change of θ from its ML estimate leads to a ln $L(\theta)$ decrease by 0.5 (2.0) from ln $L_{\max}$.

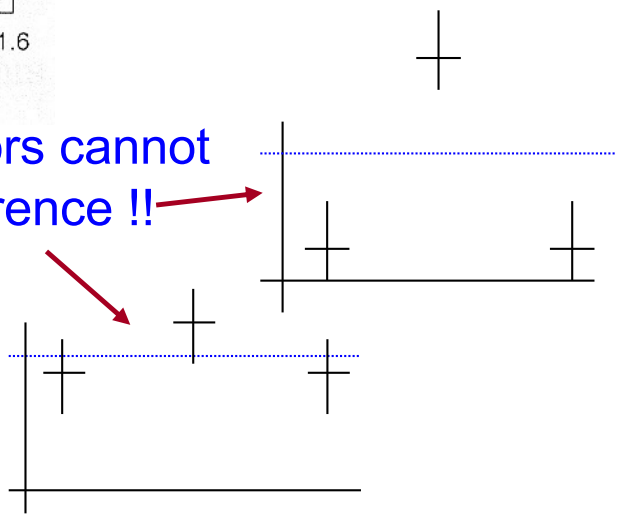


The exponential distribution example $\hat{\tau} = 1.062$

$\Delta\hat{\tau}_- = 0.137, \ \Delta\hat{\tau}_+ = 0.165$

Usually set $\hat{\tau} = 1.062^{-0.137}_{+0.165}$

Interval $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+]$ interpreted as estimate for "68.3 % confidence interval"

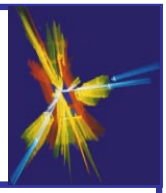


Summary on ML estimators:

ML estimators cannot tell the difference !!

- consistent.
- invariant.
- biased for small n.
- not "right", just sensible.
- don't give "most likely value of θ" but value of θ for which the data is "most likely" (highest likelihood).
- efficient for large n (saturates the RCF bound).
- often imply the use of numerical methods. (analytical ln$L$ maximisation for >1 free variables in practice impractical)
- the ln$L_{\max}$ value in itself contains no valuable information $\Rightarrow$ won't indicate if choosen function for pdf correct or not.

Extended ML



Upto now, normalisation has been fixed; can also leave normalisation free e.g. treat n as Poisson random variable with mean $\nu \rightarrow$ result of experiment: n & n x-values x_1 ... x_n.

Extended likelihood function: $L(\nu,\bar{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i;\bar{\theta})$

2 separate cases: either ν independent or a function of $\bar{\theta}$

theory/model gives $\nu = \nu(\bar{\theta})$, droping constant terms $\rightarrow$

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

Now more information used $\rightarrow$ smaller variances for $\hat{\bar{\theta}}$

Example: particle scattering, expected number of events $\nu = \varepsilon\sigma\int \mathcal{L}\, dt$, where ε detection efficiency, σ scattering cross section ("probability" given by theory) & $\mathcal{L}$ luminosity ("flux"). Use not only event describing variables but also σ.

N.B.before `repetition of experiment´ = same number of events. Here, meaning same ν (e.g. same integrated $\mathcal{L}$).

Suppose ν & $\bar{\theta}$ are (functionally) independent:

$$\frac{\partial \ln L}{\partial \nu} = 0 \rightarrow \hat{\nu} = n; \qquad \frac{\partial \ln L}{\partial \theta_j} = 0 \rightarrow \text{ usual ML } \hat{\bar{\theta}}$$

i.e. an additional fluctuation source. Useful sometimes e.g. when $f(x;\bar{\theta})$ superposition of known components $f(x;\bar{\theta}) = \sum_{j=1}^{m} \theta_j f_j(x)$. Can use usual ML with constrain

$$\theta_m = 1 - \sum_{j=1}^{m-1} \theta_j$$

but then not all θ_j independent & different θ_j treated differently.

With extended ML avoid all of that.

Extended ML (cont.)

Same with extended ML,

$$\ln L(\nu,\bar{\theta}) = \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{m} \nu\,\theta_j f_j(x_i)\right) - \nu$$

define $\mu_j = \nu\theta_j$ as expected number of events of type j,

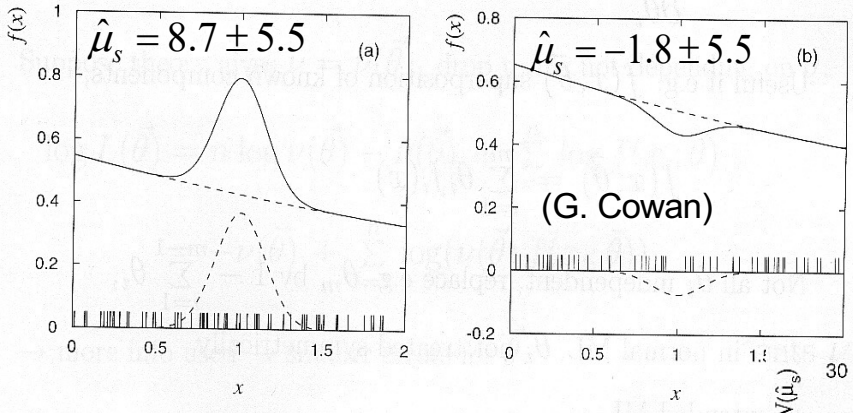
$$\ln L(\bar{\mu}) = \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{m} \mu_j f_j(x_i)\right) - \sum_{j=1}^{m} \mu_j$$

Now all parameters treated symmetrically. Often μ_j's more closely related to wanted final result e.g. production cross section for type j events. NB! ML fitted μ_j's can be < 0 & in case μ_j< 0 unphysical, must decide how to treat that case. Example: 2 types of events, signal(s) & background(b).

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

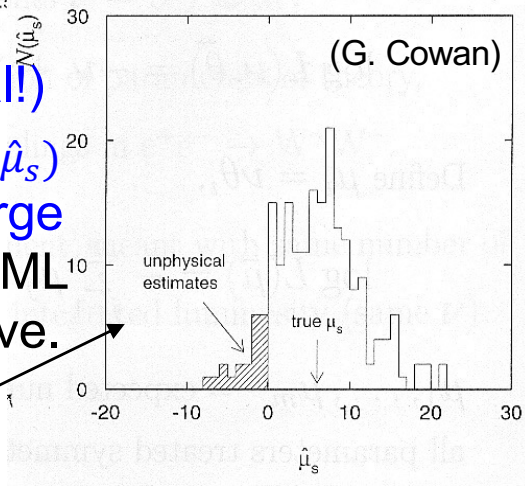
Assume $f_s(x)$ & $f_b(x)$ known,estimate μ_s& μ_b

mostly `works´ well ... but sometimes $\hat{\mu}_s < 0$



extended ML fit to 2 MC samples generated with μ_s = 6 & μ_b = 60. $f_s(x)$ Gaussian & $f_b(x)$ exponential.

Can report negative $\hat{\mu}_s$ (unphysical!) or take as estimator $\hat{\mu}_s^{\text{phys}} = \max(0, \hat{\mu}_s)$ (biased!). Can be a problem for large samples. Example: 200 extended ML fits on similar MC samples as above.

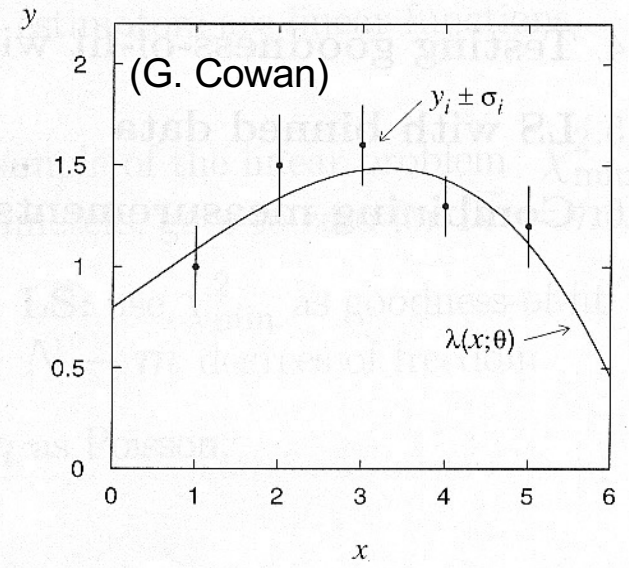


$\hat{\mu}_s = 6.1 \pm 0.4$ (using all $\hat{\mu}_s$)

$\hat{\mu}_s = 6.4 \pm 0.4$ (using only $\hat{\mu}_s^{\text{phys}}$)

The method of least squares (pienimmän neliösumman menetelmä)

N independent Gaussian random variables y_i, i = 1,..., N with expectation values $E[y_i] = \lambda_i = \lambda(x_i; \bar{\theta})$ that depend on unknown parameters $\bar{\theta}$. Each y_i related to a known x_i & known $V[y_i] = \sigma_i^2$ (e.g. estimated measurement uncertainty). E.g. temperature measurement T at positions x_i.



Least squares problem in a nut shell: N values y_1, ..., y_N measured with uncertainties σ_1,..., σ_N at x-values x_1, .., x_N (known without uncertainties). Each value λ_i of y_i given by function $\lambda(x_i, \theta)$. Goal is to minimize χ^2 sum by adjusting parameters θ (i.e. to find most optimal curve through points).

Joint pdf for N independent Gaussians y_i is product of N Gaussians:

$$g(\bar{y}; \bar{\lambda}, \bar{\sigma}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(y_i - \lambda_i)^2}{2\sigma_i^2} \right)$$

Then log-likelihood function (drop terms independent of $\bar{\theta}$):

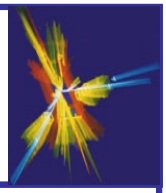
$$\ln L(\bar{\theta}) = -\frac{1}{2} \sum_{i=1}^{N} \left(y_i - \lambda(x_i, \bar{\theta}) \right)^2 / \sigma_i^2$$

However maximizing $\ln L(\bar{\theta})$ equivalent to minimizing

$$\chi^2(\bar{\theta}) = \sum_{i=1}^{N} \left(y_i - \lambda(x_i, \bar{\theta}) \right)^2 / \sigma_i^2$$

Basis of **method of least squares**: minimize quadratic sum of difference between measured & hypothesized values, weighted by inverse of measurement variance.

Linear LS estimators & their variances



So ML justifies somehow method of least-squares (LS). What "proves" validity of ML? Nothing, only an assumption

LS estimators have particularly desirable properties when $\lambda(x;\bar{\theta})$ linear function of θ: $\lambda(x;\bar{\theta}) = \sum_{j=1}^{m} a_j(x)\theta_j$,

where $a_j(x)$ are any linearly independent functions of x (i.e. one term can't be given as linear combination of the others) $\Rightarrow$

• LS estimators have zero bias & are efficient (i.e. have minimal variance) for any N ("Gauss-Markov" theorem).

• LS estimators & their variances can be found analytically although one may still prefer to estimate them numerically.

Variance (also **MC method** valid):

– Calculate **analytically** covariance matrix $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

– Alternatively one can estimate elements **numerically:**

$$(V^{-1})_{ij} = \frac{1}{2}\left[\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right]_{\bar{\theta}=\hat{\bar{\theta}}}$$

$\rightarrow$ coincides with RCF bound if λ linear function of θ & $V[y_i]$ known and Gaussian, then ln L = $-\chi^2/2$.

– Since $\lambda(x;\bar{\theta})$ linear in parameters $\bar{\theta}$, χ^2 quadratic in

$$\chi^2(\bar{\theta}) = \chi^2(\hat{\bar{\theta}}) + \frac{1}{2}\sum_{i,j=1}^{m}\left[\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right]_{\bar{\theta}=\hat{\bar{\theta}}}(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

(an expansion of the χ^2 function around its minima)

$\rightarrow$ 1 standard deviation contour in parameter space given by curve whose tangents are $\hat{\theta}_i \pm \hat{\sigma}_{\hat{\theta}_i}$ that corresponds to

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

graphical method

Even when $\lambda(x;\bar{\theta})$ not linear in $\bar{\theta}$ & formula not really valid, region $\chi^2(\bar{\theta}) \le \chi^2_{\min} + 1$ can still be interpreted as a "confidence region" with a given probability of containing true $\bar{\theta}$.

A usual application of LS method is a polynomial fit. Independent data $y \pm \Delta y$ measured at different x-values:

$$\lambda(x;\theta_0,\ldots,\theta_m) = \sum_{j=0}^{m} \theta_j x^j \quad (m+1 \text{ free parameters})$$

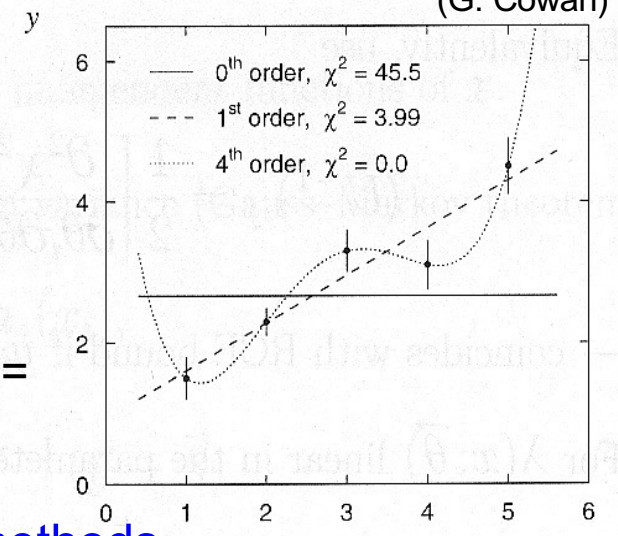
(special case of linear LS fits with $a_j(x) = x^j$)

LS polynomial fit to 5 data points (with χ^2_{min} indicated)

- 0th order (1 parameter)
- 1st order (2 parameters)
- 4th order (5 parameters)

NB! last case meaningless since number of parameters = number of data points

(G. Cowan)

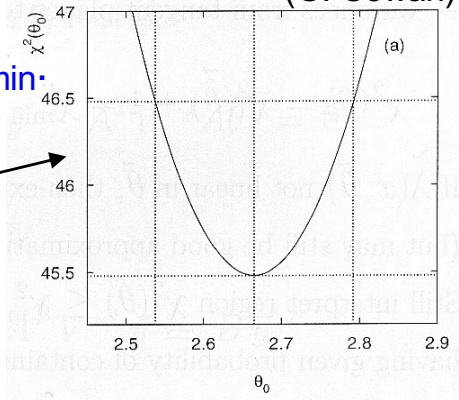


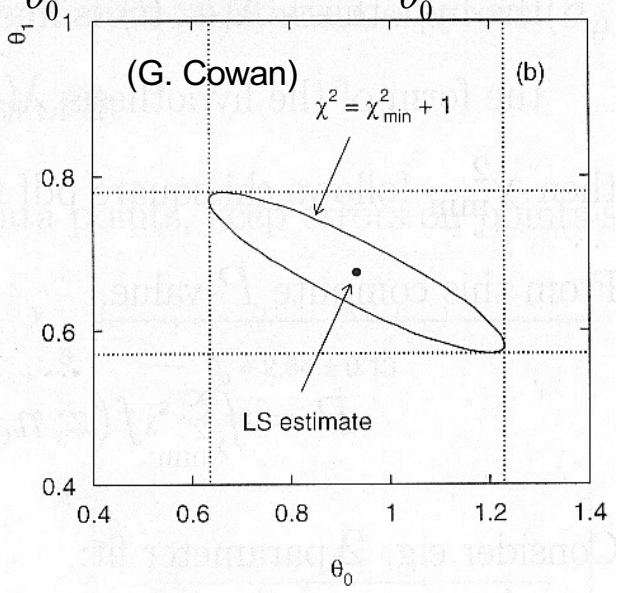
Uncertainties & covariances estimated using any of the 4 methods, all related to χ^2 change when parameters changed from those giving χ^2_{min}.

$$0\text{th order}: \hat{\theta}_0 = 2.66 \pm 0.13, \; \chi^2_{\text{min}} = 45.5$$

$$\sigma_{\hat{\theta}_0} \text{ from } \chi^2(\hat{\theta}_0 \pm \sigma_{\hat{\theta}_0}) \leq \chi^2_{\text{min}} + 1$$

(G. Cowan)

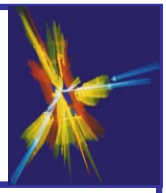




1st order:

$$\hat{\theta}_0 = 0.93 \pm 0.30,$$
$$\hat{\theta}_1 = 0.68 \pm 0.10,$$
$$\chi^2_{\text{min}} = 3.99,$$
$$\text{cov}[\hat{\theta}_0, \hat{\theta}_1] = -0.028, \; \rho = \text{cov}[\hat{\theta}_0, \hat{\theta}_1]/\sigma_{\hat{\theta}_0}\sigma_{\hat{\theta}_1} = -0.93$$

Deriving LS estimators



Let's examine our 1st order polynomial fit in more detail. How would one determine the LS estimators for the 0th & 1st order terms in this case. The expression to minimize is

$$\chi^2(\bar{\theta}) = \sum_{i=1}^{N} \left(y_i - \theta_0 - \theta_1 x_i\right)^2 / \sigma_i^2$$

To get the LS estimators one has to look for a local minima: let's further make the simplification that all σ_i's are equal (= σ).

$$\frac{\partial \chi^2}{\partial \theta_0} = 0 \rightarrow \sum_{i=1}^{N} \left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right) = \bar{y} - \hat{\theta}_0 - \hat{\theta}_1 \bar{x} = 0$$

$$\frac{\partial \chi^2}{\partial \theta_1} = 0 \rightarrow \sum_{i=1}^{N} x_i \left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right) = \overline{xy} - \hat{\theta}_0 \bar{x} - \hat{\theta}_1 \overline{x^2} = 0$$

Combining the two equations, one obtains for the LS estimators:

$$\hat{\theta}_1 = (\overline{xy} - \bar{x}\,\bar{y}) \Big/ (\overline{x^2} - \bar{x}^2) = \mathrm{cov}[x, y]/V[x]$$

$$\hat{\theta}_0 = (\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}) \Big/ (\overline{x^2} - \bar{x}^2) = \bar{y} - \hat{\theta}_1 \bar{x}$$

These are standard equations to get the slope θ_1 & the intercept with the y-axis θ_0 for a straight line fit. With different σ_i's for each point, the formula should only be modified such that each point is given a weight $1/\sigma_i^2$ & normalisation = the total weight $\Sigma 1/\sigma_i^2$.

Obtained $\chi^2_{\min}$ value used to estimate probability of hypothesis, if true, would give observed data. $(y_i - \lambda(x_i;\theta))/\sigma_i$ – a measure of agreement btwn observed data & hypothesis.

$\chi^2_{\min}$ obtained from a LS fit follows the standard chi-square distribution with the degrees-of-freedom = $N-m$ if:

- y_i, i = 1, ... , N Gaussian random variables with known covariance matrix V_{ij} or independent y_i's with known σ_i's.
- hypothesis $\lambda(x;\ \bar{\theta})$ is linear in parameters θ_i, i = 1, ..., m.
- functional form of the hypothesis $\lambda(x;\ \bar{\theta})$ is correct.

If all previous satisfied, one can calculate the P-value:

$$P = \int_{\chi^2_{\min}}^{\infty} f(z; n_d)\, dz$$

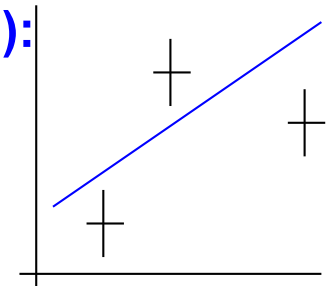
Example: consider our 1st order polynomial fit (m = 2)

$$\chi^2_{\min} = 3.99,\ N - m = 3 \rightarrow P = 0.263$$

i.e. if it is true that $\lambda(x)$ would be a straight line and if the experiment would be repeated many times, then in 26.3 % of the cases, one would obtain a worse (i.e. higher) $\chi^2_{\min}$.

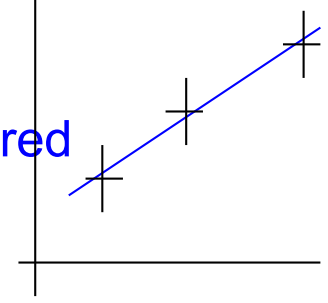
NB! $E[\chi^2_{\min}]$ = n_d (number degrees-of-freedom) → each data point should contribute ≈ 1 to the χ^2

$\chi^2_{\min}/n_d$ »1 (or a very small P–value):
- hypothesis (”= function”) wrong
- or measurement wrong/bad
- or uncertainties underestimated
- or extremely bad luck (unlikely!)



Very common pit fall!

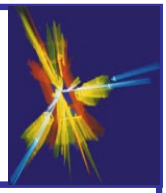
$\chi^2_{\min}/n_d$ «1 (or a large P–value):
- the uncertainties overestimated
- or correlations of uncertainties ignored
- or extremely good luck (unlikely!)



Note distinction btwn small statistical uncertainty on LS estimator & a good LS fit (i.e. small $\chi^2_{\min}$)
Statistical uncertainty estimated from change of χ^2 near its minimum is independent of the absolute value of the $\chi^2_{\min}$.

Variance (statistical uncertainty) of estimator tells us: If experiment repeated many times, how wide is distribution of estimates (doesn’t tell whether hypothesis correct or not).

Combining measurements with LS



Very common use of LS method is to combine a number of measurements of the same quantity. Then one has:

y_i = result of measurement i, i = 1, ..., N;
$\sigma_i^2 = V[y_i]$, assumed to be known;
λ = true value (takes role of θ, no x-dependence).

For independent y_i's, minimize: $\chi^2(\lambda) = \sum_{i=1}^{N} \frac{(y_i - \lambda)^2}{\sigma_i^2}$

$$\frac{\partial \chi^2}{\partial \lambda} = 0 \;\rightarrow\; \hat{\lambda} = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \Big/ \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$$

$$V[\hat{\lambda}] = \left(\frac{1}{2} \left[\frac{\partial \chi^2}{\partial^2 \lambda} \right]_{\lambda=\hat{\lambda}} \right)^{-1} = 1 \Big/ \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$$

Well-known formula for **weighted average**. Variance of average < variances of individual measurements. More precise measurements (i.e. having smaller variances) have larger weight. Generalized to non-independent measurements y_i i.e. cov[y_i,y_j]=V_{ij}. Then have to minimize:

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

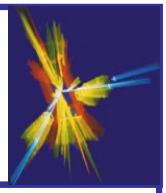
$$\frac{\partial \chi^2}{\partial \lambda} = 0 \;\rightarrow\; \hat{\lambda} = \sum_{i=1}^{N} w_i y_i, \quad w_i = \sum_{j=1}^{N} (V^{-1})_{ij} \Big/ \sum_{k,l=1}^{N} (V^{-1})_{kl}$$

$$V[\hat{\lambda}] = \left(\frac{1}{2} \left[\frac{\partial \chi^2}{\partial^2 \lambda} \right]_{\lambda=\hat{\lambda}} \right)^{-1} = \sum_{i,j=1}^{N} w_i V_{ij} w_j = \mathbf{w}^T \mathbf{V}\, \mathbf{w}$$

$$\sum_{i=1}^{N} w_i = 1 \rightarrow E[\hat{\lambda}] = \sum_{i=1}^{N} w_i E[y_i] = \lambda \sum_{i=1}^{N} w_i = \lambda, \text{ i.e. unbiased}$$

Assumption: individual y_i's unbiased. The weights in LS prescription give RCF bound for variance ("Gauss-Markov").

Combining measurements with LS (cont.)



Averaging 2 correlated measurements: measurements y_1 & y_2 with covariance matrix $\mathbf{V}$:

$$\mathbf{V}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\rightarrow\quad \mathbf{V}^{-1}=\frac{1}{1-\rho^2}\begin{pmatrix}1/\sigma_1^2 & -\rho/\sigma_1\sigma_2\\ -\rho/\sigma_1\sigma_2 & 1/\sigma_2^2\end{pmatrix}$$

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

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

Increase of inverse variance due to the 2nd measurement:

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

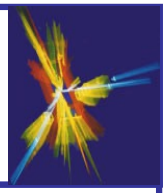
$\rho\leq 1\Longrightarrow$ 2nd measurement only beneficial for average (i.e. new combined variance $\leq$). No variance change when $\rho = \sigma_1/\sigma_2$ (incl. $\rho = 1$ & $\sigma_1 = \sigma_2$, i.e. same measurement twice).

If $\rho > \sigma_2/\sigma_1$ then $w < 0$ & weighted average not btwn y_1 & y_2, due to a large positive correlation btwn y_1 & y_2. Can happen in the case of common normalisation uncertainty. Usable for calibrating common variable see e.g. G.Cowan: *Statistical Data Analysis*, page 109, where temperature (T) estimate improved using measurements at same T with two rulers having different thermal expansion coefficients.

Please use method with care !

Overlaping samples can be used to estimate covariance matrix if not known. Either use MC generated samples or use real data by dividing data sample into a large number of subsamples & determing estimators y_1, ..., y_N for each subsample & from these correlations coefficient matrix.

LS fit with penalty functions (nuisance terms)



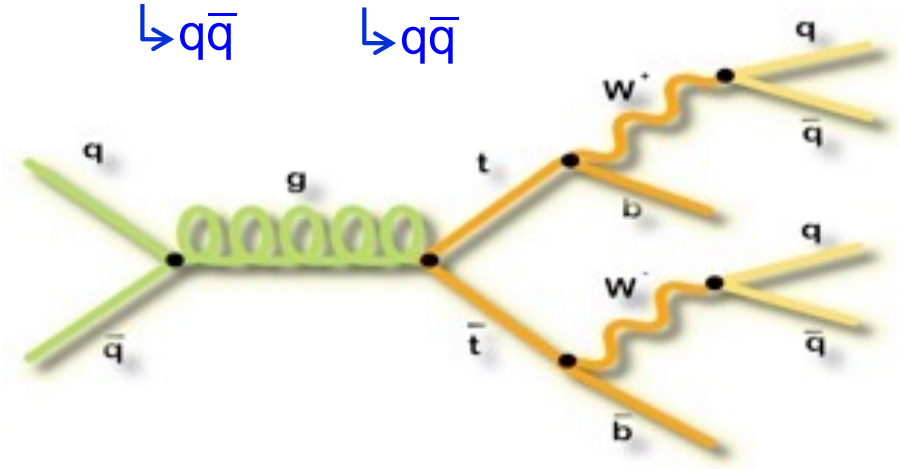
LS fit with constraints (penalty functions/nuisance terms) to improve data quality: Sometimes inputs (x_i's) suffer from significant uncertainties or there are some scalings (nuisance terms) involved that directly effects our result. Then a possible solution can be to include additional terms in the χ^2 sum and look for the global χ^2_{min} that minimazes everything including the variation of inputs (with respect to their uncertainties) or the scalings.

Example: top mass reconstruction at CDF experiment

Proton-antiproton collisions create top-antitop pair:

$p\bar{p} \rightarrow t\,(\rightarrow W^+b)\bar{t}\,(\rightarrow W^-\bar{b})\,X$

$\hookrightarrow q\bar{q}$ $\quad$ $\hookrightarrow q\bar{q}$



experimental challenge: quarks (q, b) hadronize and make jets whose energy (& momentum) measured with poor resolution (10-20 %) $\Rightarrow$ if raw energy (& momentum) measurement used the top mass poorly reconstructed. Make use of additional constraints: two of the jets should make a W resonance ($\mu_W = 80.42 \pm 0.03$ GeV for the W mass) with a width $\Gamma_W = 2.1$ GeV + combined with the third jet should make a top quark with same μ_{top} (= top mass) as the other top (= "triplet" of jets) within a width $\Gamma_{top} = 1.6$ GeV.

LS fit with penalty functions (nuisance terms)

i) The masses are first determined by using the formulae given:

$$m_i = \sqrt{E_i^2 - P_{x,i}^2 - P_{y,i}^2 - P_{z,i}^2}$$

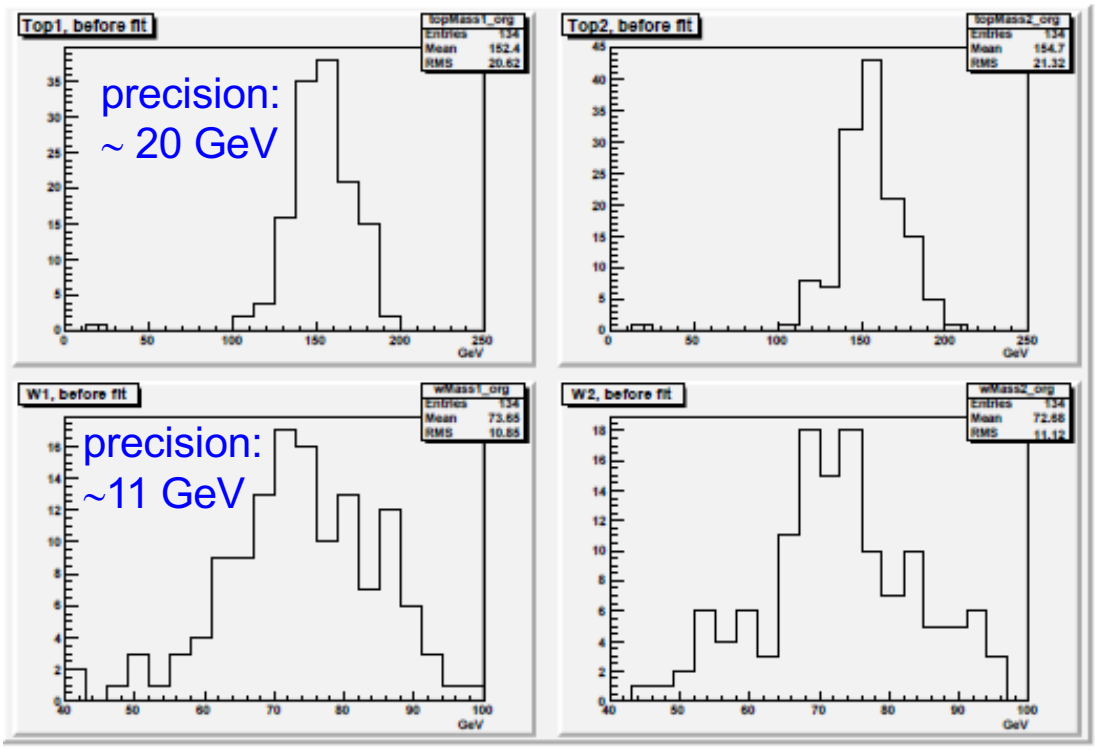
The energies of the j:th ($j = 1, 2$) top quark and the W-boson are (from the channel given):

$E_{bj(qj)} \equiv$ energy of b-jet (q-jet) j
$P_{x(y,z),bj(qj)} \equiv x(y,z)$ momentum component of b-jet (q-jet) j

$$E_{t,j} = E_{bj} + E_{qj1} + E_{qj2}$$

$$E_{w,j} = E_{qj1} + E_{qj2}$$

(same for P_x, P_y and P_z)



ii) Now we test the hypothesis given

$$\chi^2 = \frac{(m_{top1} - m_{top2})^2}{2\Gamma_{top}^2} + \frac{(m_{W1} - m_{Wnom})^2}{\Gamma_W^2} + \frac{(m_{W2} - m_{Wnom})^2}{\Gamma_W^2},$$

Γ_{top} = 1.6 GeV, Γ_W = 2.1 GeV, m_{Wnom} = 80.42 GeV

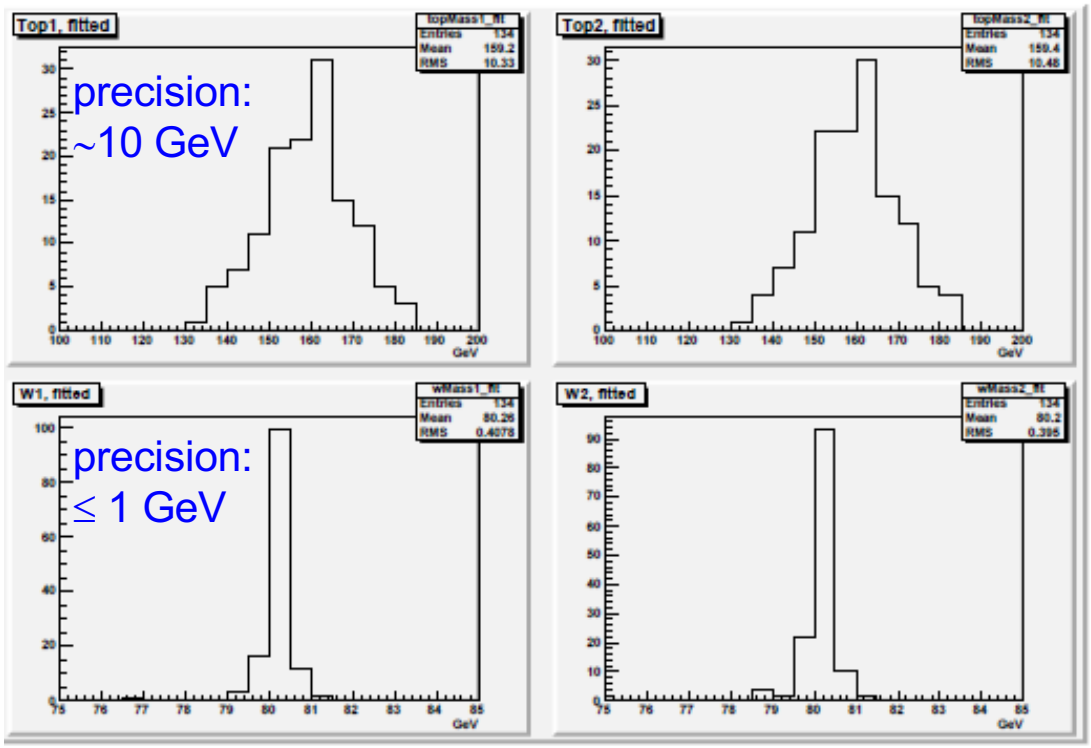
χ^2 large (> 100). Cure: allow energy measurements to vary

$$\chi^2_{new} = \chi^2 + \sum_{i=b1,b2,q11,q12,q21,q22} \frac{(E_i - c_i \cdot E_i)^2}{\sigma_{E_i}^2}$$

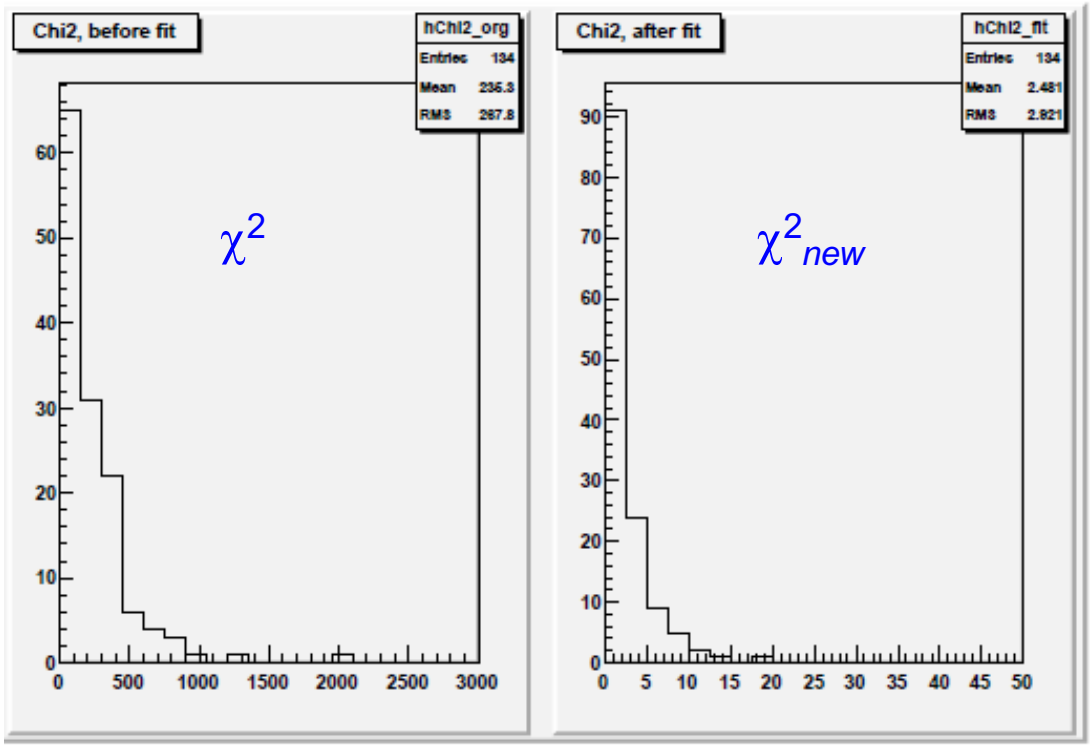
Let each c_i vary in order to minimaze whole χ^2_{new}

LS fit with penalty functions (nuisance terms)

Result:



global χ^2 values significantly improved (by a factor ~100!)



Note: here problem a bit simplied, in reality need also to scale *P*'s