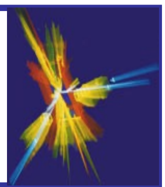


Parameter estimation: general concepts

Hypothesis testing:

Theory $\xrightarrow[\text{Calculus}]{\text{Probability}}$ 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 $\xleftarrow[\text{Inference}]{\text{Statistical}}$ 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 $\hat{\theta}(\vec{x})$?

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

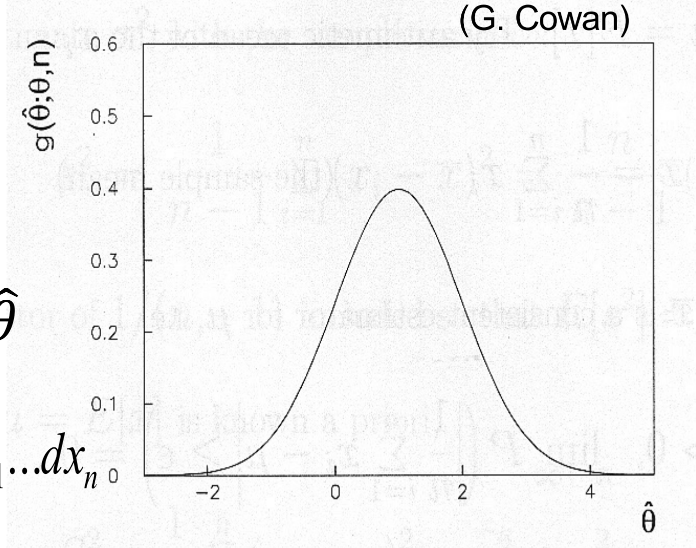
Properties of estimators

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. Cowan)



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

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 $V[\hat{\theta}] = \sigma_{\hat{\theta}}^2$

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

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

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

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, $x_1,\dots, x_n$.
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.

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

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

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

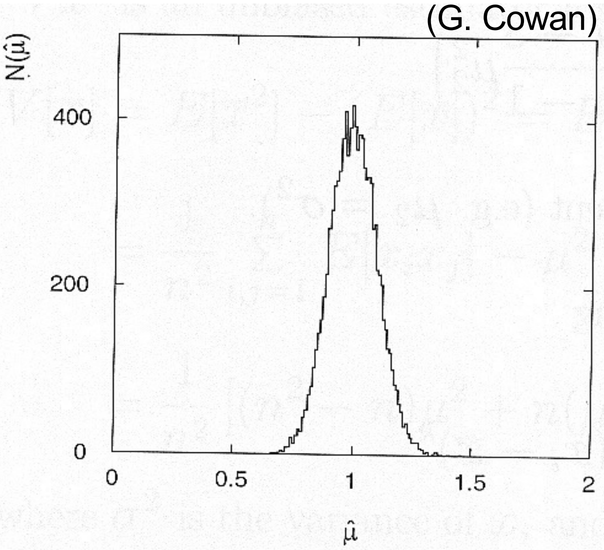
$\rightarrow$ $\bar{x}$ is an unbiased estimator for μ. The variance of $\bar{x}$ is

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

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.

(G. Cowan)



Enter values into a histogram.

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

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

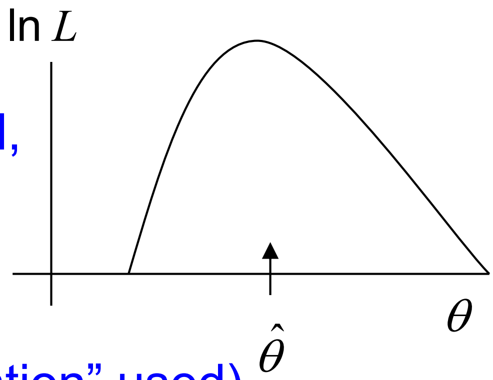
If hypothesis (including value of θ) correct (= true)
$\rightarrow$ expect higher probability for the data
If hypothesized functional form wrong or θ value far away
$\rightarrow$ expect lower probability for the data
$\Rightarrow$ higher value of **likelihood function** close to true θ

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

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.

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

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

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

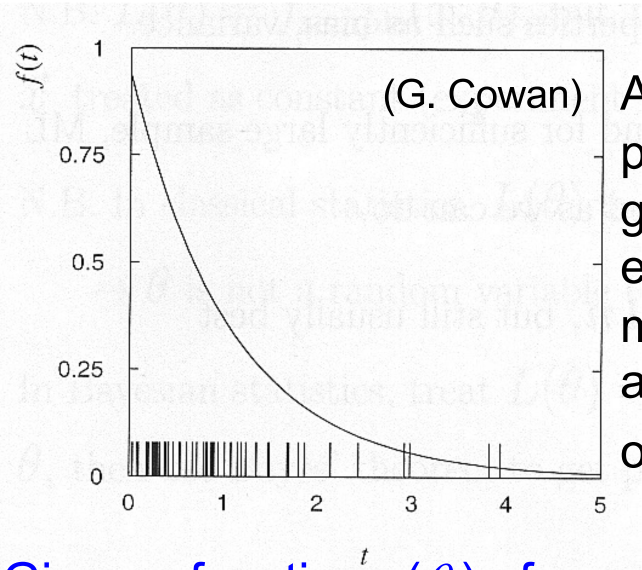
$$\int\ldots\int\left(\sum_{i=1}^{n} t_i\right)\frac{e^{-t_1/\tau}}{\tau}\cdots\frac{e^{-t_n/\tau}}{\tau}\frac{dt_1\ldots dt_n}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(\int\frac{t_i}{\tau}e^{-t_i/\tau}dt_i\prod_{j\neq i}\frac{e^{-t_j/\tau}dt_j}{\tau}\right)$$

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

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 $E[t]$ = τ.

Suppose that one is interested in decay constant λ = ln2 $/\tau$ 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 $\widehat{N}_{\text{taxi}}$ undefined.

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

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 $\rightarrow$

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

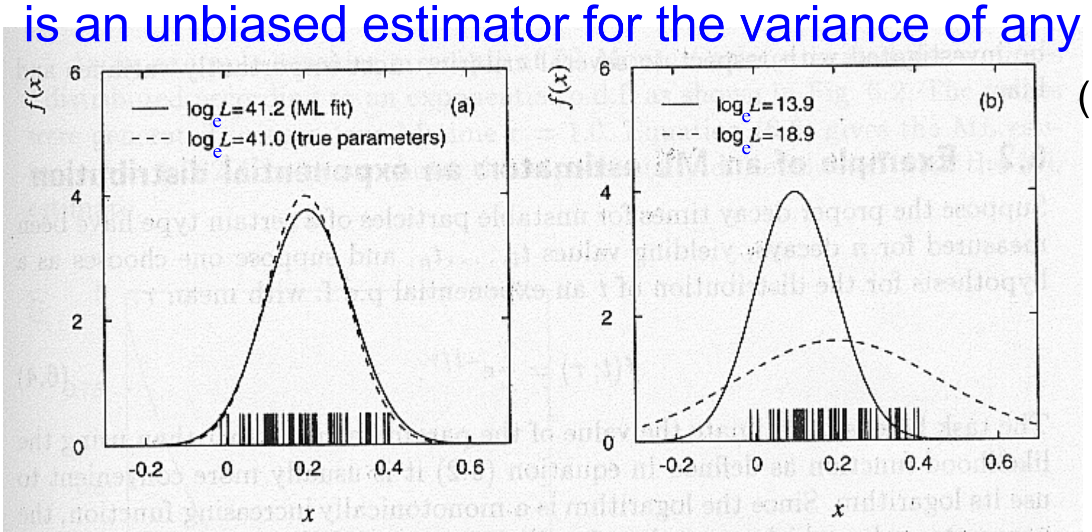
Already known that $\hat{\mu}$ is an unbiased estimator for μ.

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

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.

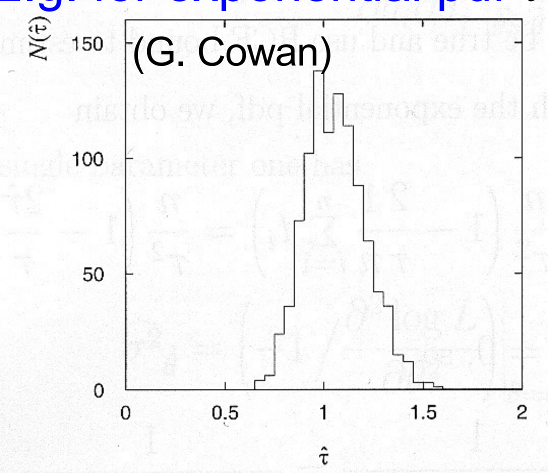
If pdf $g(\hat{\tau};\, \tau, n)$ Gaussian, then $[\hat{\tau} - \widehat{\sigma_{\hat{\tau}}}, \hat{\tau} + \widehat{\sigma_{\hat{\tau}}}]$ equivalent to ”68.3 % confidence interval” for $\hat{\tau}$, generally accepted way to quote uncertainty even when errors non-Gaussian.

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 $\tau = 1.062$ 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)

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

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,

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

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:

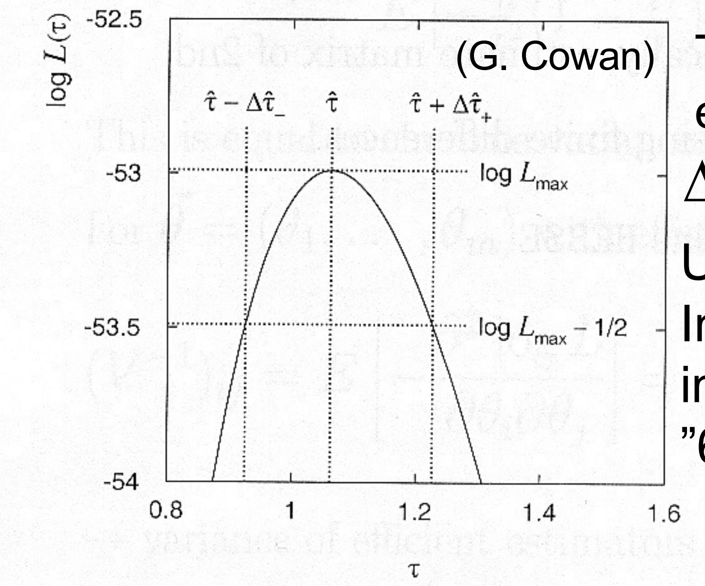
$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta = \hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta = \hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

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

Variance of ML estimators

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

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 $\hat{\tau} = 1.062$

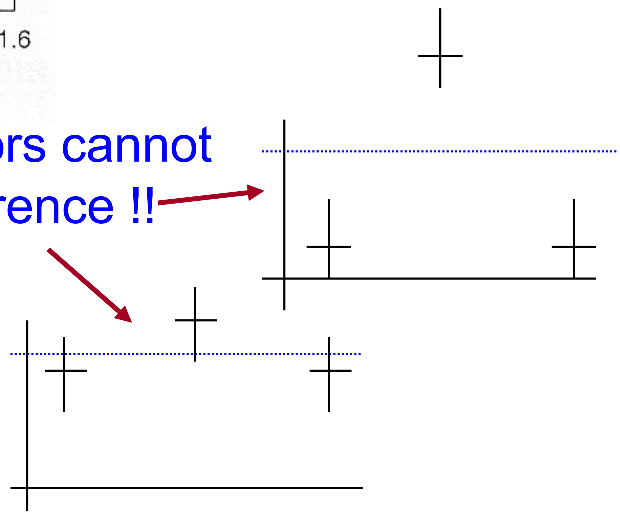
$\Delta\hat{\tau}_- = 0.137,\ \Delta\hat{\tau}_+ = 0.165$

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.