Computing and Statistical Data Analysis
Stat 9: Parameter Estimation, Limits

London Postgraduate Lectures on Particle Physics;

University of London MSci course PH4515

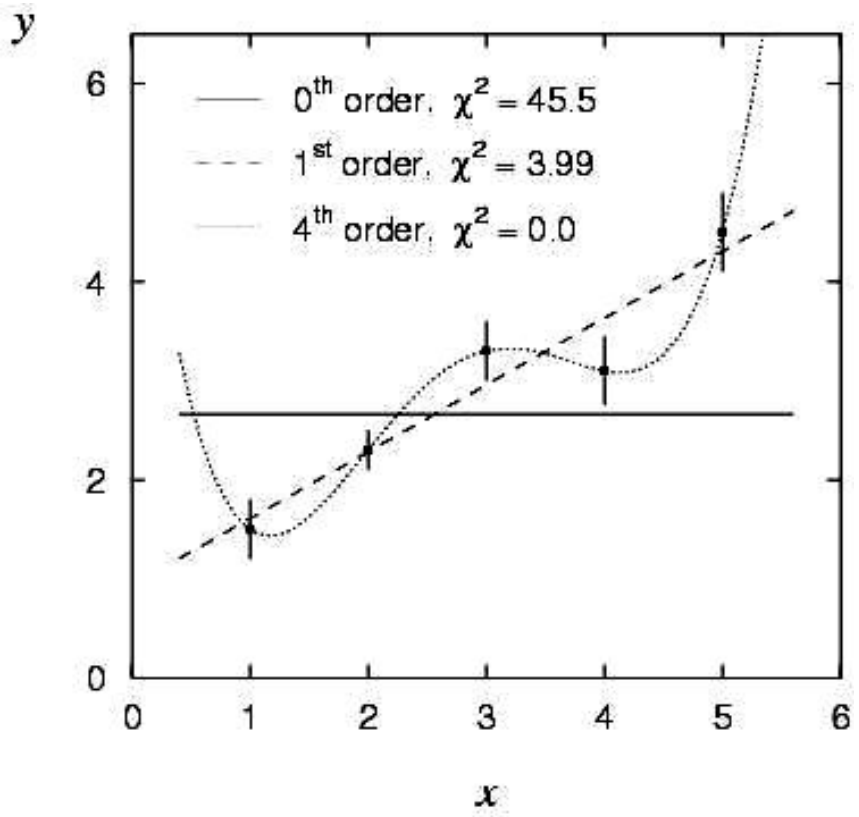
Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Course web page:

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

Example of least squares fit

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



Variance of LS estimators

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

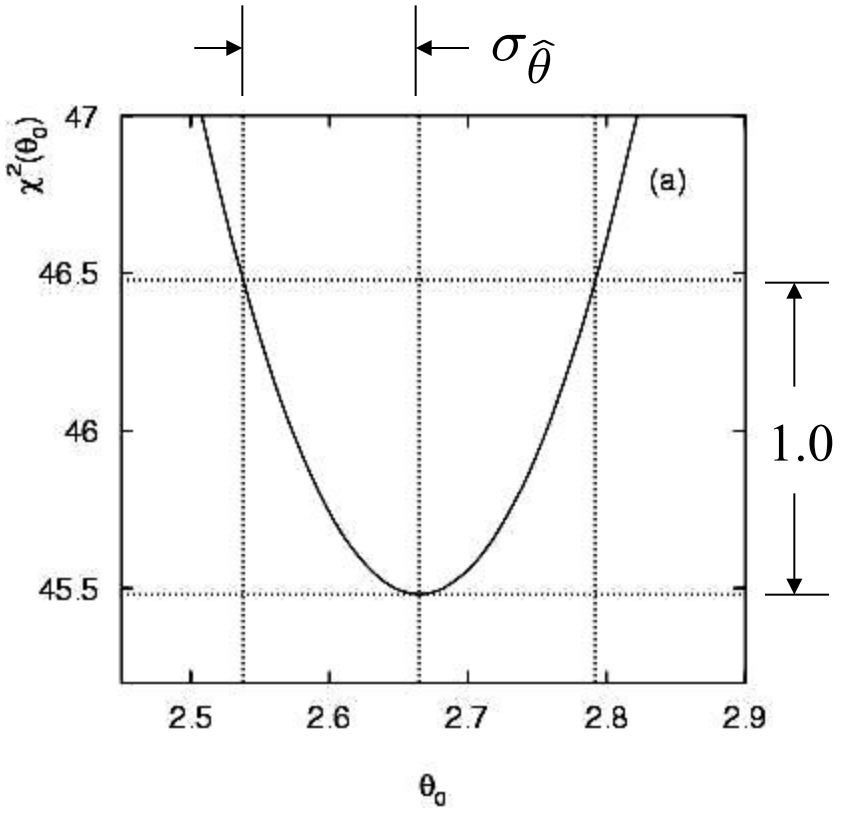
$$\chi^2(\theta) = -2 \ln L(\theta)$$

and so

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

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

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



Two-parameter LS fit

2-parameter case (line with nonzero slope):

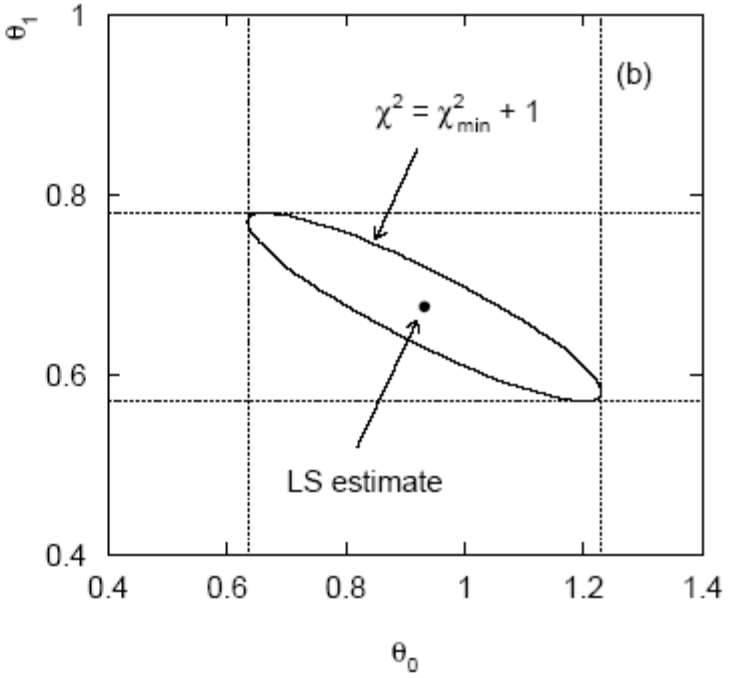
$$\hat{\theta}_0 = 0.93 \pm 0.30,$$

$$\hat{\theta}_1 = 0.68 \pm 0.10$$

$$\widehat{\mathrm{cov}}[\hat{\theta}_0, \hat{\theta}_1] = -0.028$$

$$r = -0.90$$

$$\chi^2 = 3.99$$



Tangent lines $\rightarrow \sigma_{\hat{\theta}_0}$, $\sigma_{\hat{\theta}_1}$.

Angle of ellipse $\rightarrow$ correlation (same as for ML)

Goodness-of-fit with least squares

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

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

It can therefore be employed as a goodness-of-fit statistic to test the hypothesized functional form $\lambda(x; \theta)$.

We can show that if the hypothesis is correct, then the statistic $t = \chi^2_{\text{min}}$ 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

Goodness-of-fit with least squares (2)

The chi-square pdf has an expectation value equal to the number of degrees of freedom, so if $\chi^2_{\text{min}} \approx n_{\text{d}}$ the fit is 'good'.

More generally, find the *p*-value: $p = \int_{\chi^2_{\text{min}}}^{\infty} f(t; n_{\text{d}})\, dt$

This is the probability of obtaining a χ^2_{min} as high as the one we got, or higher, if the hypothesis is correct.

E.g. for the previous example with 1st order polynomial (line),

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

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

$$\chi^2_{\text{min}} = 45.5\,, \qquad n_{\text{d}} = 5-1 = 4\,, \qquad p = 3.1\times10^{-9}$$

Goodness-of-fit vs. statistical errors

Small statistical error does not mean a good fit (nor vice versa).

Curvature of χ^2 near its minimum $\rightarrow$ statistical errors ($\sigma_{\hat{\theta}}$)

Value of χ^2_{min} $\rightarrow$ goodness-of-fit

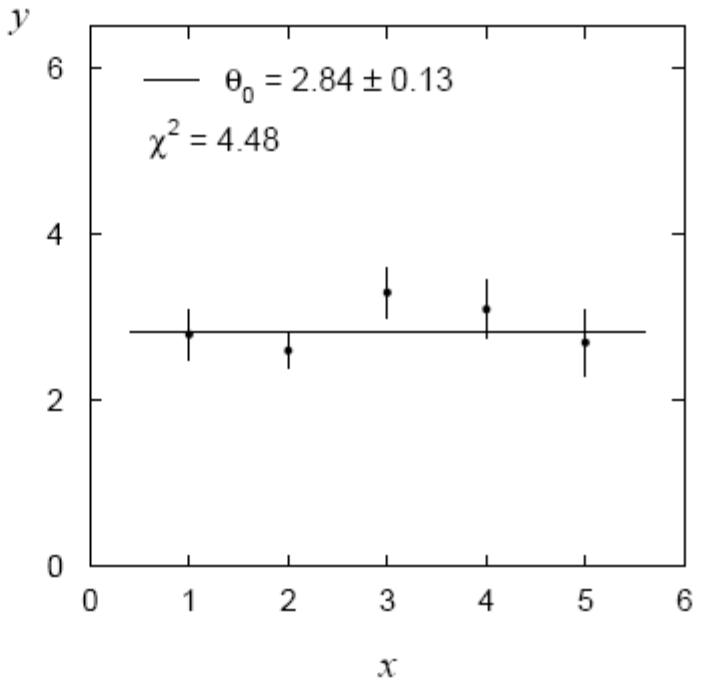
Horizontal line fit, move the data points, keep errors on points same:

$\hat{\theta}_0 = 2.84 \pm 0.13$

$\chi^2_{\text{min}} = 4.48$

Variance same as before,

now χ^2_{min} 'good'.



Goodness-of-fit vs. stat. errors (2)

$\rightarrow$ $\chi^2(\theta_0)$ shifted down, same curvature as before.

Variance of estimator (statistical error) tells us:

if experiment repeated many times, how wide is the distribution of the estimates $\hat{\theta}$. (Doesn't tell us whether hypothesis correct.)

P-value tells us:

if hypothesis is correct and experiment repeated many times, what fraction will give equal or worse agreement between data and hypothesis according to the statistic χ^2_{min}.

Low P-value $\rightarrow$ hypothesis may be wrong $\rightarrow$ systematic error.

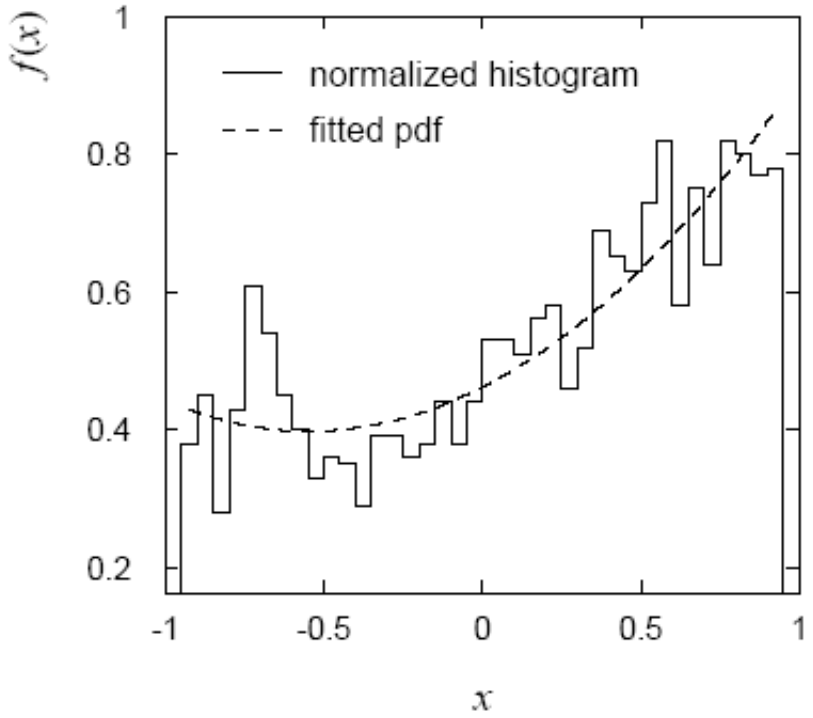
LS with binned data

Histogram:

N bins, n entries.

Hypothesized pdf:

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



We have

$$y_i = \text{number of entries in bin } i,$$

$$\lambda_i(\vec{\theta}) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x;\vec{\theta})dx = np_i(\vec{\theta})$$

LS with binned data (2)

LS fit: minimize

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

where $\sigma_i^2 = V[y_i]$, here not known a priori.

Treat the y_i as Poisson r.v.s, in place of true variance take either

$$\sigma_i^2 = \lambda_i(\vec{\theta}) \qquad \text{(LS method)}$$

$$\sigma_i^2 = y_i \qquad \text{(Modified LS method)}$$

MLS sometimes easier computationally, but χ^2_{min} no longer follows chi-square pdf (or is undefined) if some bins have few (or no) entries.

LS with binned data — normalization

Do not 'fit the normalization':

$$\lambda_i(\vec{\theta}, \nu) = \nu \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = \nu p_i(\vec{\theta})$$

i.e. introduce adjustable ν, fit along with $\vec{\theta}$.

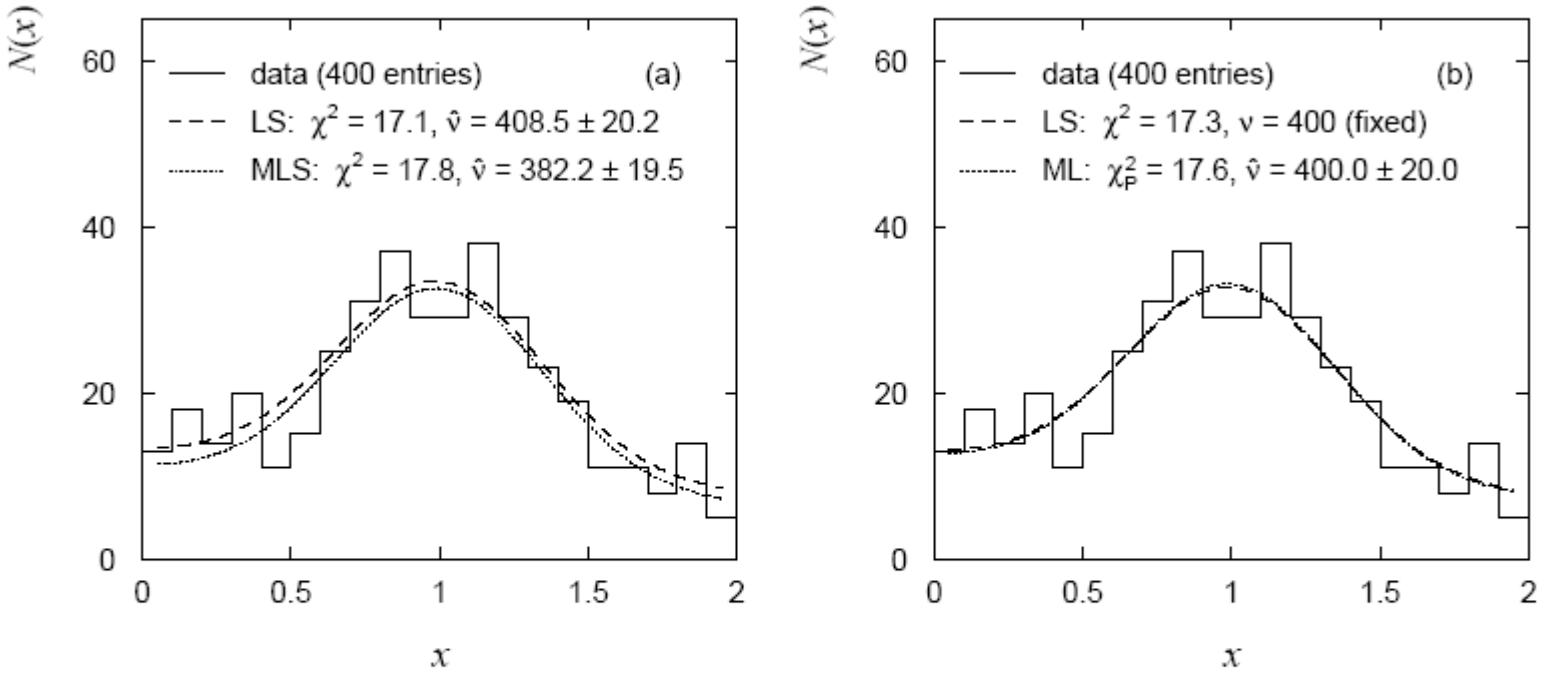
$\hat{\nu}$ is a bad estimator for n (which we know, anyway!)

$$\hat{\nu}_{\rm LS} = n + \frac{\chi^2_{\rm min}}{2}$$

$$\hat{\nu}_{\rm MLS} = n - \chi^2_{\rm min}$$

LS normalization example

Example with $n = 400$ entries, $N = 20$ bins:



Expect $\chi^2_{\rm min}$ around $N - m$,

$\rightarrow$ relative error in $\hat{\nu}$ large when N large, n small

Either get n directly from data for LS (or better, use ML).

Using LS to combine measurements

Use LS to obtain weighted average of N measurements of λ:

y_i = result of measurement i, $i = 1, \ldots, N$;

$\sigma_i^2 = V[y_i]$, assume known;

λ = true value (plays role of θ).

For uncorrelated y_i, minimize

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

Set $\frac{\partial \chi^2}{\partial \lambda} = 0$ and solve,

$$\rightarrow \quad \hat{\lambda} = \frac{\sum_{i=1}^{N} y_i/\sigma_i^2}{\sum_{j=1}^{N} 1/\sigma_j^2} \qquad V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^{N} 1/\sigma_i^2}$$

Combining correlated measurements with LS

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

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

$$\rightarrow \quad \hat{\lambda} = \sum_{i=1}^{N} w_i y_i, \qquad w_i = \frac{\sum_{j=1}^{N}(V^{-1})_{ij}}{\sum_{k,l=1}^{N}(V^{-1})_{kl}}$$

$$V[\hat{\lambda}] = \sum_{i,j=1}^{N} w_i V_{ij} w_j$$

LS $\hat{\lambda}$ has zero bias, minimum variance (Gauss–Markov theorem).

Example: averaging two correlated measurements

Suppose we have y_1, y_2, and $V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

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

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

The increase in inverse variance due to 2nd measurement is

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

$\rightarrow$ 2nd measurement can only help.

Negative weights in LS average

If $\rho > \sigma_1/\sigma_2$, $\rightarrow$ $w < 0$,

$\rightarrow$ weighted average is not between y_1 and y_2 (!?)

Cannot happen if correlation due to common data, but possible for shared random effect; very unreliable if e.g. ρ, σ_1, σ_2 incorrect.

See example in SDA Section 7.6.1 with two measurements at same temperature using two rulers, different thermal expansion coefficients:

average is outside the two measurements; used to improve estimate of temperature.

Interval estimation — introduction

In addition to a 'point estimate' of a parameter we should report an interval reflecting its statistical uncertainty.

Desirable properties of such an interval may include:

communicate objectively the result of the experiment;
have a given probability of containing the true parameter;
provide information needed to draw conclusions about the parameter possibly incorporating stated prior beliefs.

Often use +/− the estimated standard deviation of the estimator.
In some cases, however, this is not adequate:

estimate near a physical boundary,
e.g., an observed event rate consistent with zero.

We will look briefly at Frequentist and Bayesian intervals.

Frequentist confidence intervals

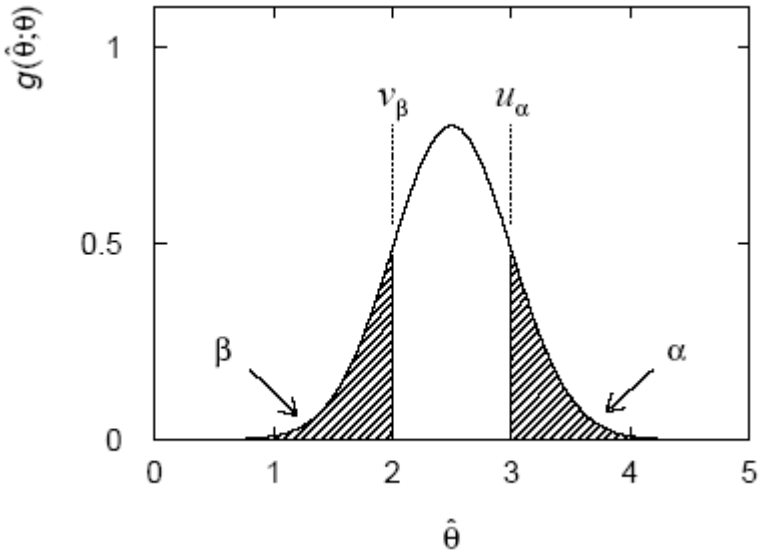
Consider an estimator $\hat{\theta}$ for a parameter θ and an estimate $\hat{\theta}_{\text{obs}}$.

We also need for all possible θ its sampling distribution $g(\hat{\theta}; \theta)$.

Specify upper and lower tail probabilities, e.g., $\alpha = 0.05$, $\beta = 0.05$, then find functions $u_\alpha(\theta)$ and $v_\beta(\theta)$ such that:

$$\begin{aligned} \alpha &= P(\hat{\theta} \geq u_\alpha(\theta)) \\ &= \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta}; \theta)\, d\hat{\theta} \end{aligned}$$

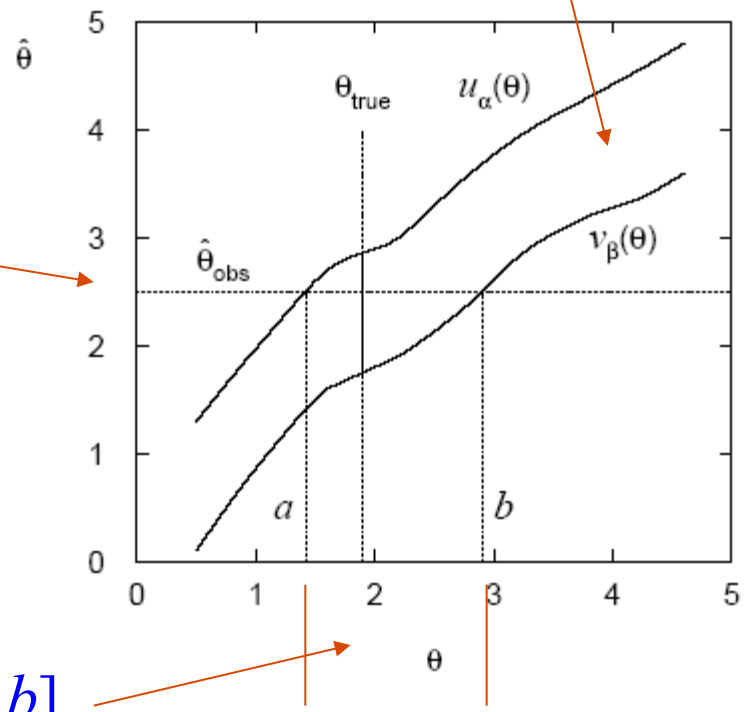
$$\begin{aligned} \beta &= P(\hat{\theta} \leq v_\beta(\theta)) \\ &= \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta}; \theta)\, d\hat{\theta} \end{aligned}$$



Confidence interval from the confidence belt

The region between $u_\alpha(\theta)$ and $v_\beta(\theta)$ is called the confidence belt.

Find points where observed estimate intersects the confidence belt.



This gives the confidence interval $[a, b]$

Confidence level = $1 - \alpha - \beta$ = probability for the interval to cover true value of the parameter (holds for any possible true θ).

Confidence intervals by inverting a test

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

> Specify values of the data that are 'disfavoured' by θ (critical region) such that P(data in critical region) $\leq \gamma$ for a prespecified γ, e.g., 0.05 or 0.1.
>
> If data observed in the critical region, reject the value θ.

Now invert the test to define a confidence interval as:

> set of θ values that would not be rejected in a test of size γ (confidence level is $1 - \gamma$).

The interval will cover the true value of θ with probability $\geq 1 - \gamma$.

Equivalent to confidence belt construction; confidence belt is acceptance region of a test.

Relation between confidence interval and p-value

Equivalently we can consider a significance test for each hypothesized value of θ, resulting in a p-value, p_θ..

If $p_\theta < \gamma$, then we reject θ.

The confidence interval at CL = $1 - \gamma$ consists of those values of θ that are not rejected.

E.g. an upper limit on θ is the greatest value for which $p_\theta \geq \gamma$.

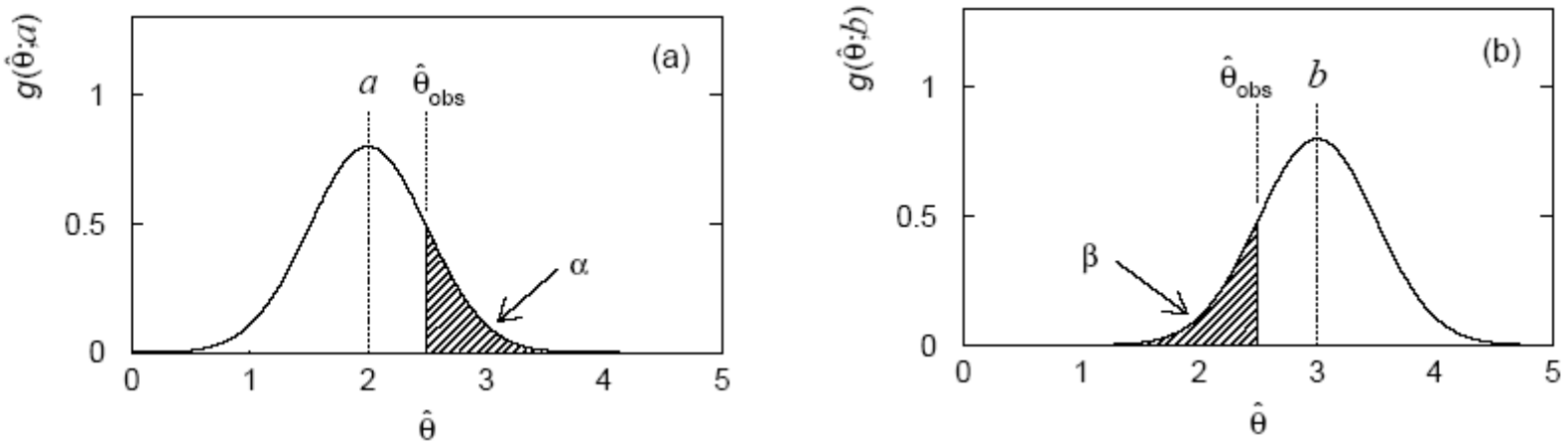
In practice find by setting $p_\theta = \gamma$ and solve for θ.

Confidence intervals in practice

The recipe to find the interval $[a, b]$ boils down to solving

$$\alpha = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\, d\hat{\theta} = \int_{\hat{\theta}_{\rm obs}}^{\infty} g(\hat{\theta};a)\, d\hat{\theta}\,,$$

$$\beta = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\, d\hat{\theta} = \int_{-\infty}^{\hat{\theta}_{\rm obs}} g(\hat{\theta};b)\, d\hat{\theta}\,.$$



→ a is hypothetical value of θ such that $P(\hat{\theta} > \hat{\theta}_{\rm obs}) = \alpha$.

→ b is hypothetical value of θ such that $P(\hat{\theta} < \hat{\theta}_{\rm obs}) = \beta$.

Meaning of a confidence interval

N.B. the interval is random, the true θ is an unknown constant.

Often report interval $[a, b]$ as $\hat{\theta}^{+d}_{-c}$, i.e. $c = \hat{\theta} - a$, $d = b - \hat{\theta}$.

So what does $\hat{\theta} = 80.25^{+0.31}_{-0.25}$ mean? It does not mean:

$P(80.00 < \theta < 80.56) = 1 - \alpha - \beta$, but rather:

repeat the experiment many times with same sample size,

construct interval according to same prescription each time,

in $1 - \alpha - \beta$ of experiments, interval will cover θ.

Central vs. one-sided confidence intervals

Sometimes only specify α or β, $\rightarrow$ one-sided interval (limit)

Often take $\alpha = \beta = \dfrac{\gamma}{2}$ $\rightarrow$ coverage probability $= 1 - \gamma$

$\rightarrow$ central confidence interval

N.B. 'central' confidence interval does not mean the interval is symmetric about $\hat{\theta}$, but only that $\alpha = \beta$.

The HEP error 'convention': 68.3% central confidence interval.

Intervals from the likelihood function

In the large sample limit it can be shown for ML estimators:

$$\hat{\vec{\theta}} \sim N(\vec{\theta}, V) \qquad (n\text{-dimensional Gaussian, covariance } V)$$

$$L(\vec{\theta}) = L_{\max} \exp\left[-\frac{1}{2} Q(\hat{\vec{\theta}}, \vec{\theta})\right], \quad Q(\hat{\vec{\theta}}, \vec{\theta}) = (\hat{\vec{\theta}} - \vec{\theta})^T V^{-1} (\hat{\vec{\theta}} - \vec{\theta})$$

$Q(\hat{\vec{\theta}}, \vec{\theta}) = Q_\gamma$ defines a hyper-ellipsoidal confidence region,

$$P(\text{ellipsoid covers true } \vec{\theta}) = P(Q(\hat{\vec{\theta}}, \vec{\theta}) \leq Q_\gamma)$$

If $\hat{\vec{\theta}} \sim N(\vec{\theta}, V)$ then $Q(\hat{\vec{\theta}}, \vec{\theta}) \sim \text{Chi-square}(n)$

$$\text{coverage probability} \equiv 1 - \gamma = \int_0^{Q_\gamma} f_{\chi^2}(z; n)\, dz = F_{\chi^2}(Q_\gamma; n)$$

Approximate confidence regions from $L(\theta)$

So the recipe to find the confidence region with CL = $1-\gamma$ is:

$$\ln L(\vec{\theta}) = \ln L_{\max} - \frac{Q_\gamma}{2} \quad \text{or} \quad \chi^2(\vec{\theta}) = \chi^2_{\min} + Q_\gamma$$

where $Q_\gamma = F^{-1}_{\chi^2}(1-\gamma; n)$

Q_γ	$1-\gamma$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

For finite samples, these are approximate confidence regions.

Coverage probability not guaranteed to be equal to $1-\gamma$;

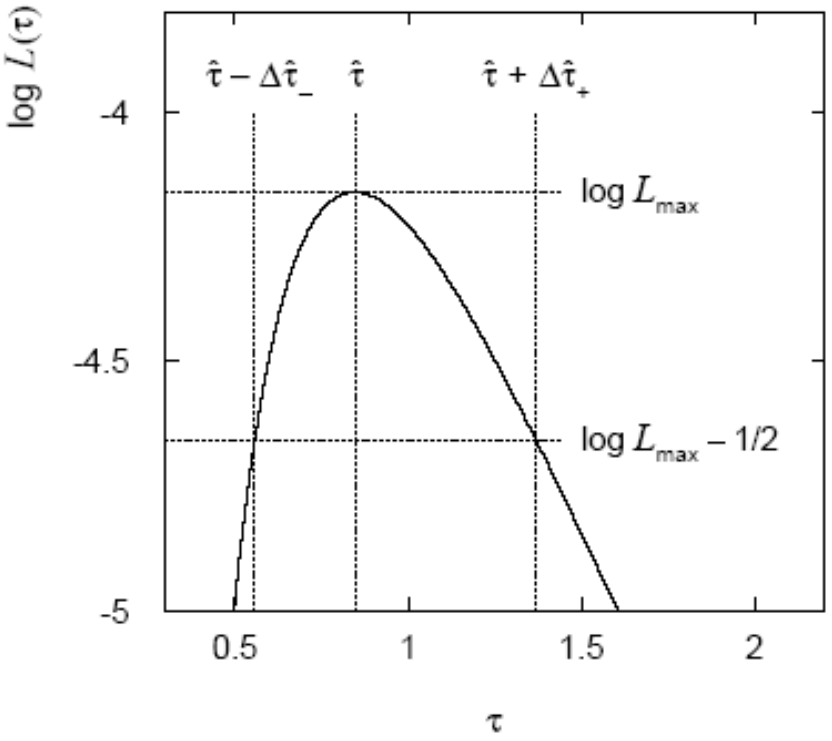
no simple theorem to say by how far off it will be (use MC).

Remember here the interval is random, not the parameter.

Example of interval from ln $L(\theta)$

For n=1 parameter, CL = 0.683, $Q_\gamma = 1$.

Our exponential example, now with $n = 5$ observations:



$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$