

Statistical & Mathematical Foundations

Lecture 2: Parameter and interval estimation



STEAM Academy
CERN
6,7 July 2026

<https://indico.cern.ch/event/1699443/>



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

Outline

Lecture 1:

Introduction

Probability

Hypothesis tests

→ Lecture 2:

Parameter estimation

Confidence limits

Systematic uncertainties

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: $\mathbf{x} = (x_1, \dots, x_n)$

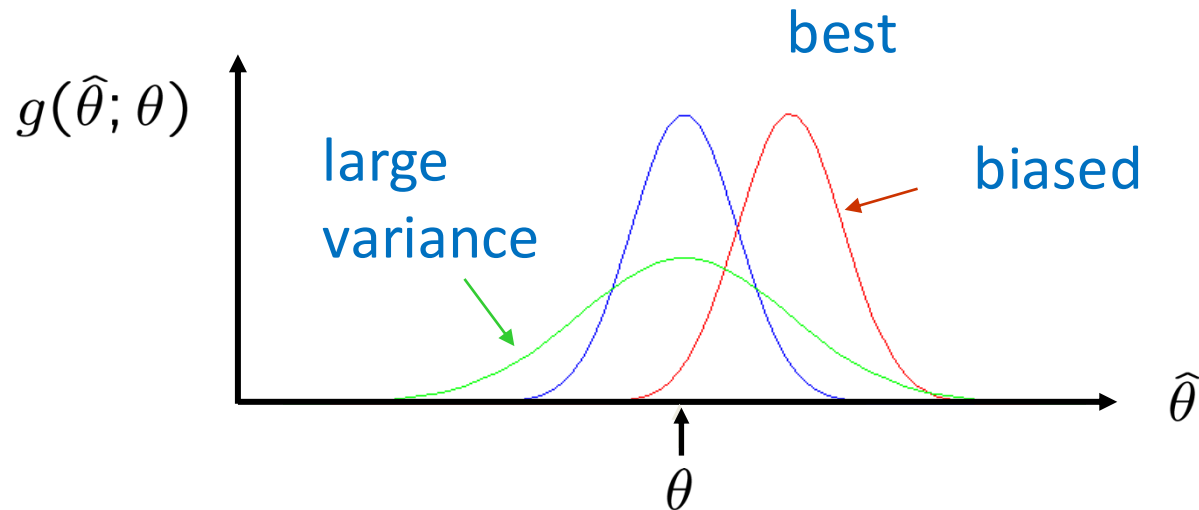
We want to find some function of the data to estimate the parameter(s):

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

Sometimes we say ‘estimator’ for the function of x_1, \dots, 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

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

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

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

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

In this case the likelihood function is

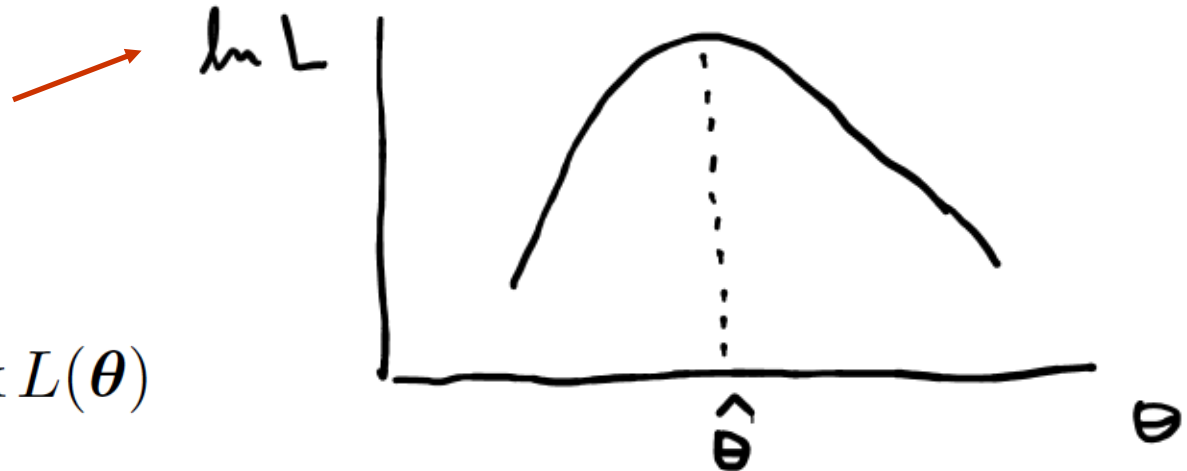
$$L(\vec{\theta}) = \prod_{i=1}^n f(x_i; \vec{\theta}) \quad (x_i \text{ constant})$$

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 $\log L$

$$\hat{\theta} = \operatorname{argmax}_{\theta} L(\theta)$$



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, \dots, 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)

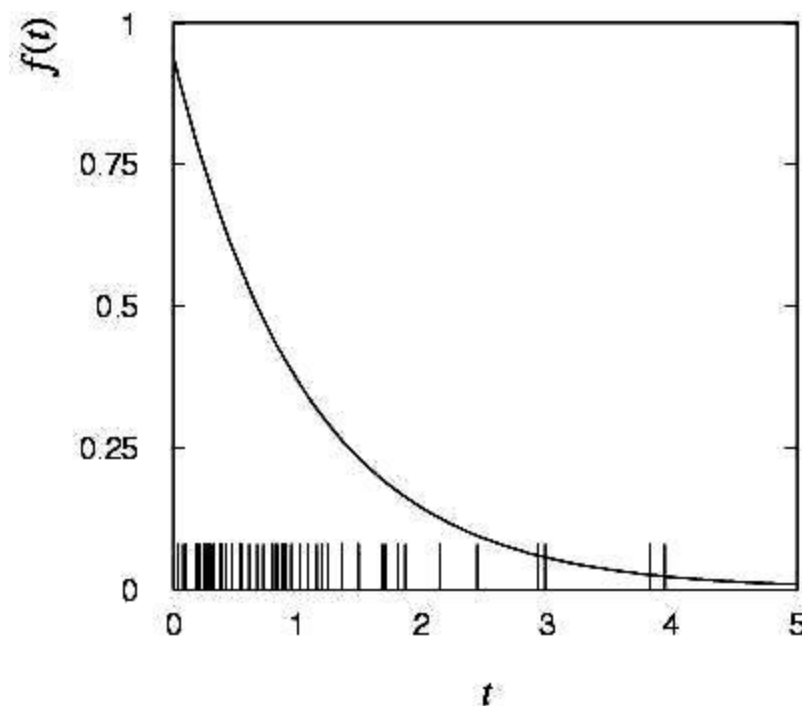
Find its maximum by setting $\frac{\partial \ln L(\tau)}{\partial \tau} = 0$,

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

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, \dots, x_n

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

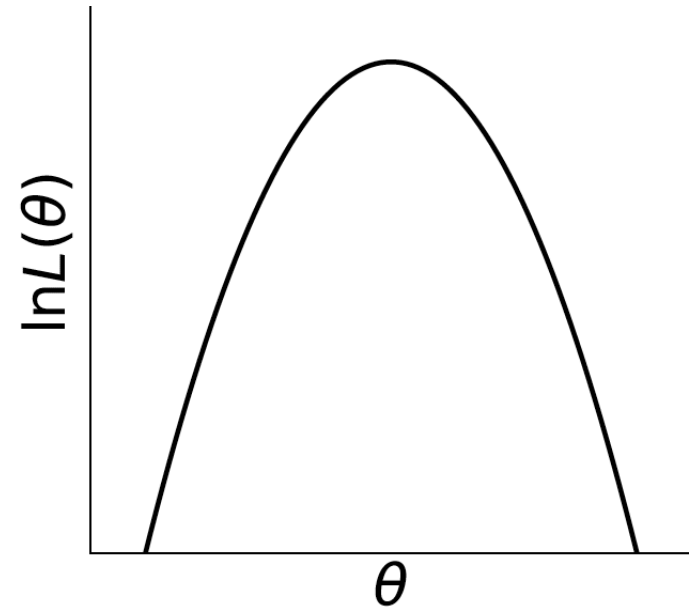
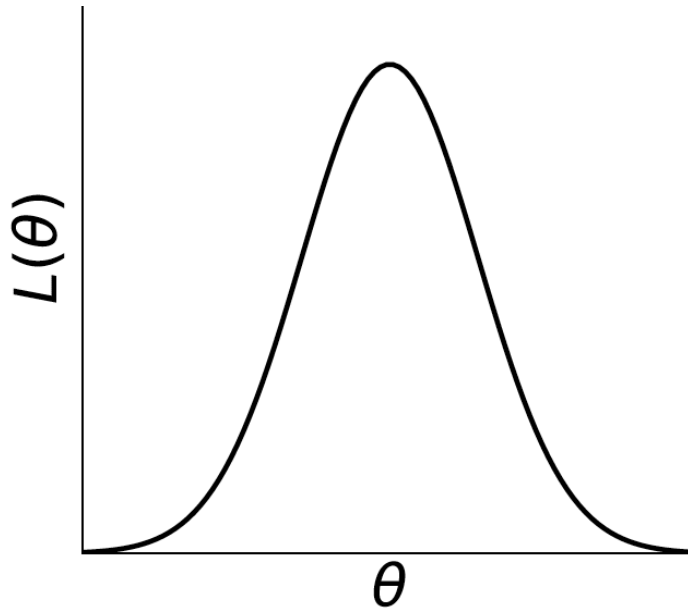
The regularity conditions include:

- 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, $\theta = (\theta_1, \dots, \theta_m)$.

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 $L(\theta) = P(\mathbf{x}|\theta)$ is the probability for \mathbf{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 $\epsilon > 0$,

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

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:

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


(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:

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 / E \left[-\frac{\partial^2 \ln L}{\partial \theta^2} \right] = \text{MVB} \quad (\text{Minimum Variance Bound})$$

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



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):

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

The MLE is said to be *asymptotically efficient*.

The MLE's distribution becomes Gaussian

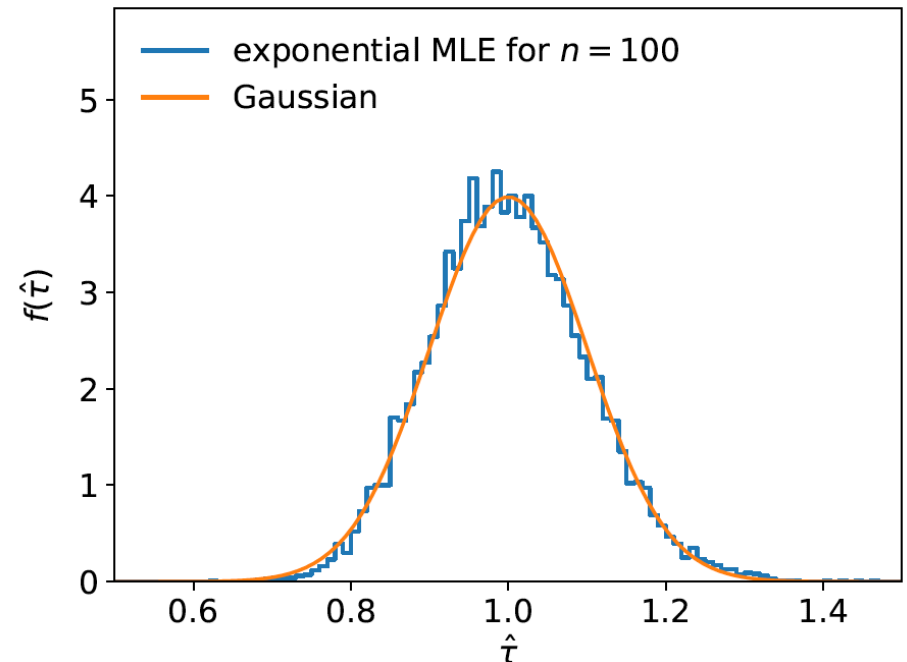
In the large-sample limit, the pdf of the MLE becomes Gaussian,

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

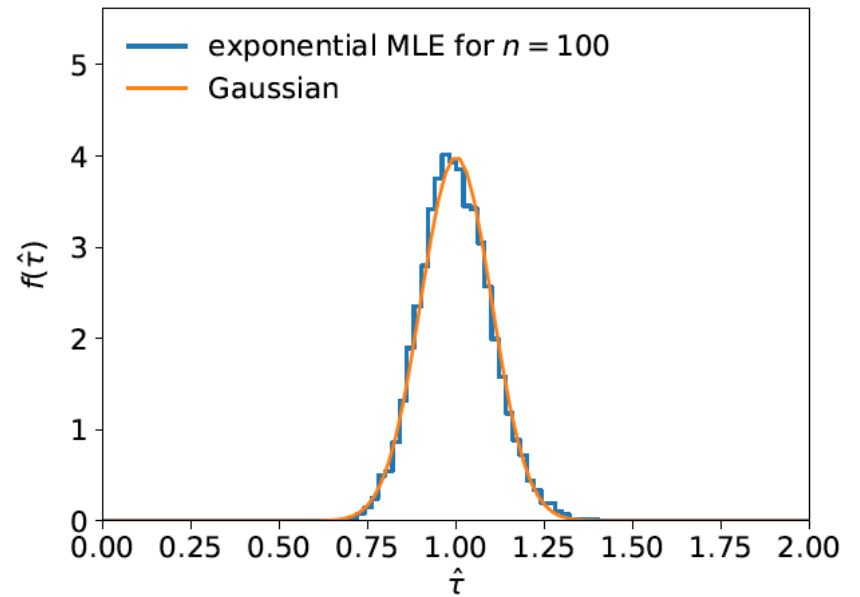
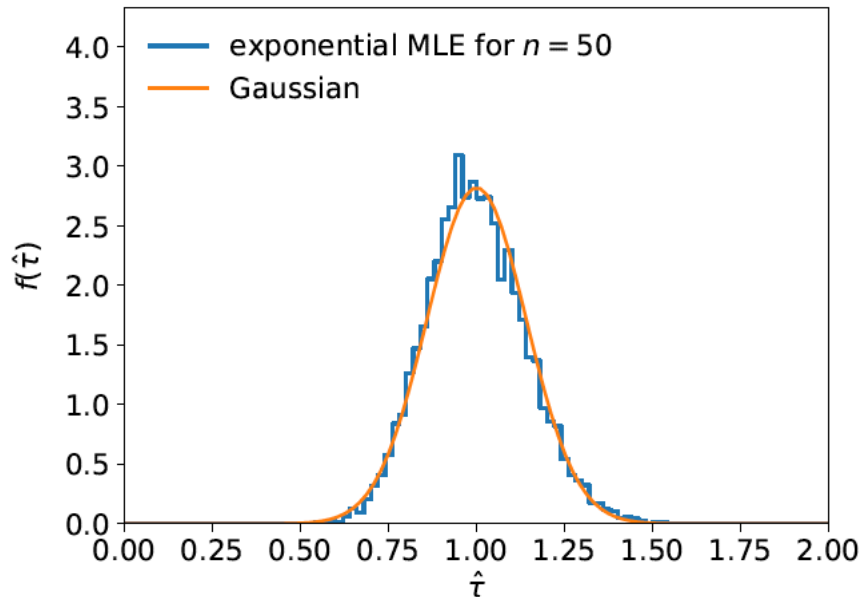
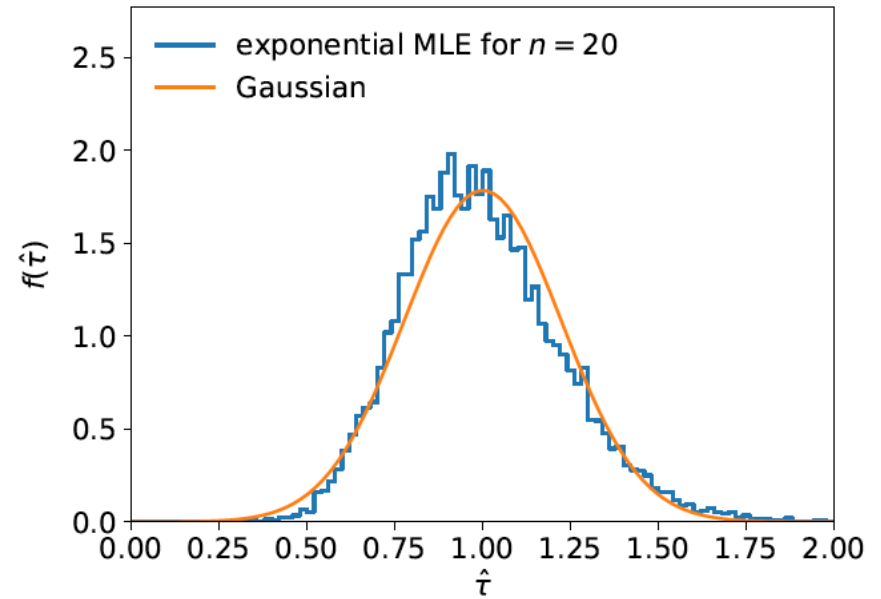
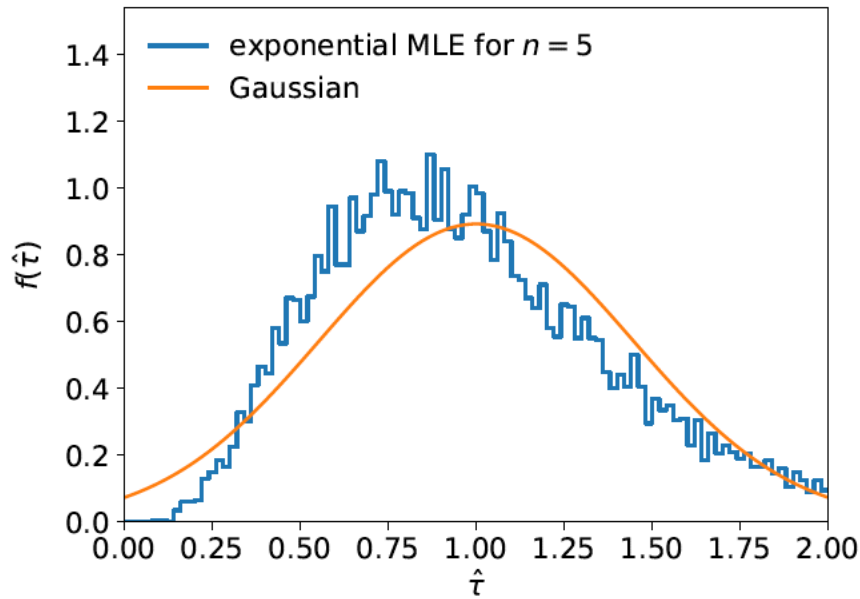
where $\sigma_{\hat{\theta}}^2$ is the minimum variance bound (note bias is zero).

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



Variance of estimators: Monte Carlo method

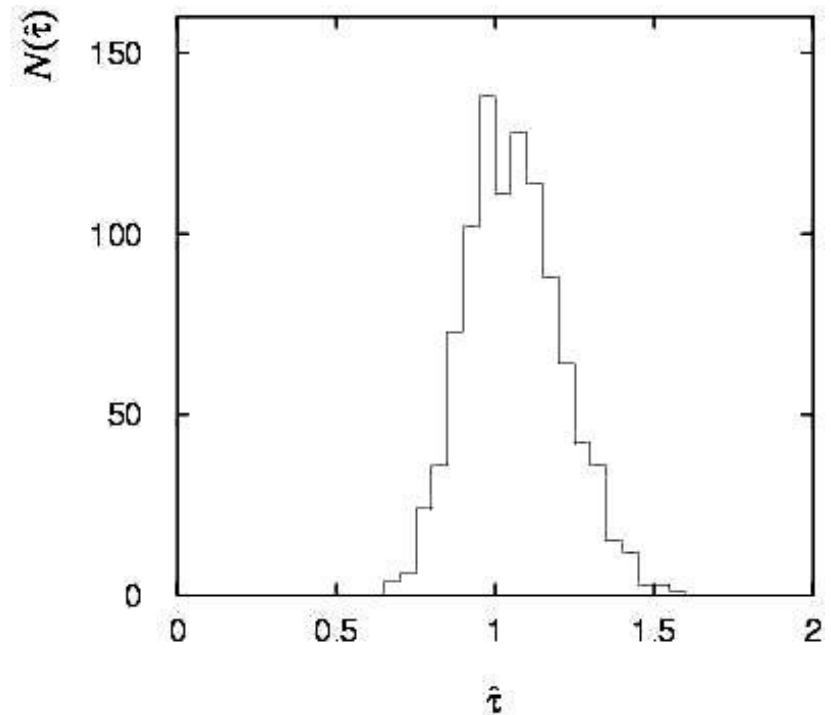
Having estimated our parameter we now need to report its ‘statistical error’, using e.g. the estimator’s standard deviation, or (co)variance.

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):

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

MVB 

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

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

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

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

Variance of estimators: graphical method

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

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

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

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

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



→ to get $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until $\ln L$ decreases by $1/2$.

Example of variance by graphical method

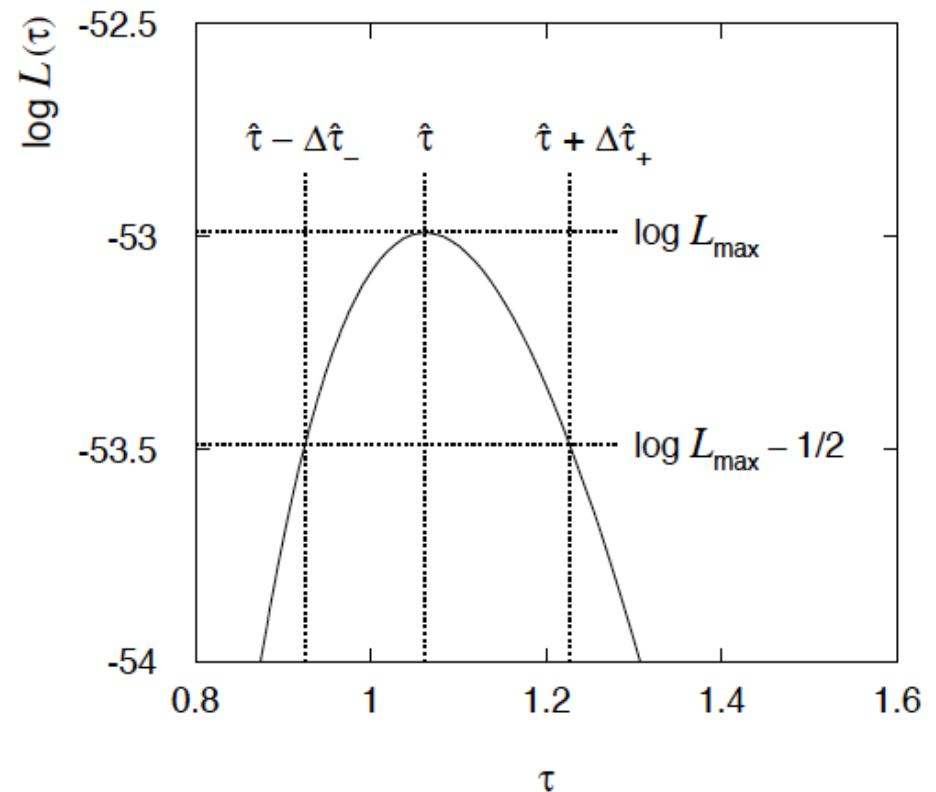
ML example with exponential:

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

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

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

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



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

Information inequality for N parameters

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

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 $V_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$

The information inequality states that the matrix

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

is positive semi-definite:

$$\mathbf{z}^T M \mathbf{z} \geq 0 \text{ for all } \mathbf{z} \neq 0, \text{ diagonal elements } \geq 0$$

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

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

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

Confidence intervals by inverting a test

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

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 $P(\text{data in critical region} | \theta) \leq \alpha$ for a prespecified α , 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 are not rejected in a test of size α (confidence level CL is $1 - \alpha$).

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 $p_\theta \leq \alpha$, then we reject θ .

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

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

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

For a multidimensional parameter space $\theta = (\theta_1, \dots, \theta_M)$ use same idea – result is a confidence “region” with boundary determined by $p_\theta = \alpha$.

Coverage probability of confidence interval

If the true value of θ is rejected, then it's not in the confidence interval. The probability for this is by construction (equality for continuous data):

$$P(\text{reject } \theta | \theta) \leq \alpha = \text{type-I error rate}$$

Therefore, the probability for the interval to contain or “cover” θ is

$$P(\text{conf. interval “covers” } \theta | \theta) \geq 1 - \alpha$$

This assumes that the set of θ values considered includes the true value, i.e., it assumes the composite hypothesis $P(\mathbf{x}|H, \theta)$.

Frequentist upper limit on Poisson parameter

Consider again the case of observing $n \sim \text{Poisson}(s + b)$.

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

Relevant alternative is $s = 0$ (critical region at low n)

p -value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$

Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from

$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}} + b)}$$

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F_{\chi^2}^{-1}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\theta = (\theta_1, \dots, \theta_n)$ using the ratio

$$\lambda(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \quad 0 \leq \lambda(\theta) \leq 1$$

Lower $\lambda(\theta)$ means worse agreement between data and hypothesized θ . Equivalently, usually define

$$t_\theta = -2 \ln \lambda(\theta)$$

so higher t_θ means worse agreement between θ and the data.

p -value of θ therefore

$$p_\theta = \int_{t_{\theta, \text{obs}}}^{\infty} f(t_\theta | \theta) dt_\theta$$

need pdf

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and provided certain conditions hold...)

$$f(t_{\theta}|\theta) \sim \chi_n^2$$

chi-square dist. with # d.o.f. =
of components in $\theta = (\theta_1, \dots, \theta_n)$.

Assuming this holds, the p -value is

$$p_{\theta} = 1 - F_{\chi_n^2}(t_{\theta}) \quad \leftarrow \text{set equal to } \alpha$$

To find boundary of confidence region set $p_{\theta} = \alpha$ and solve for t_{θ} :

$$t_{\theta} = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Recall also

$$t_{\theta} = -2 \ln \frac{L(\theta)}{L(\hat{\theta})}$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in θ space is where

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2} F_{\chi_n^2}^{-1}(1 - \alpha)$$

For example, for $1 - \alpha = 68.3\%$ and $n = 1$ parameter,

$$F_{\chi_1^2}^{-1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

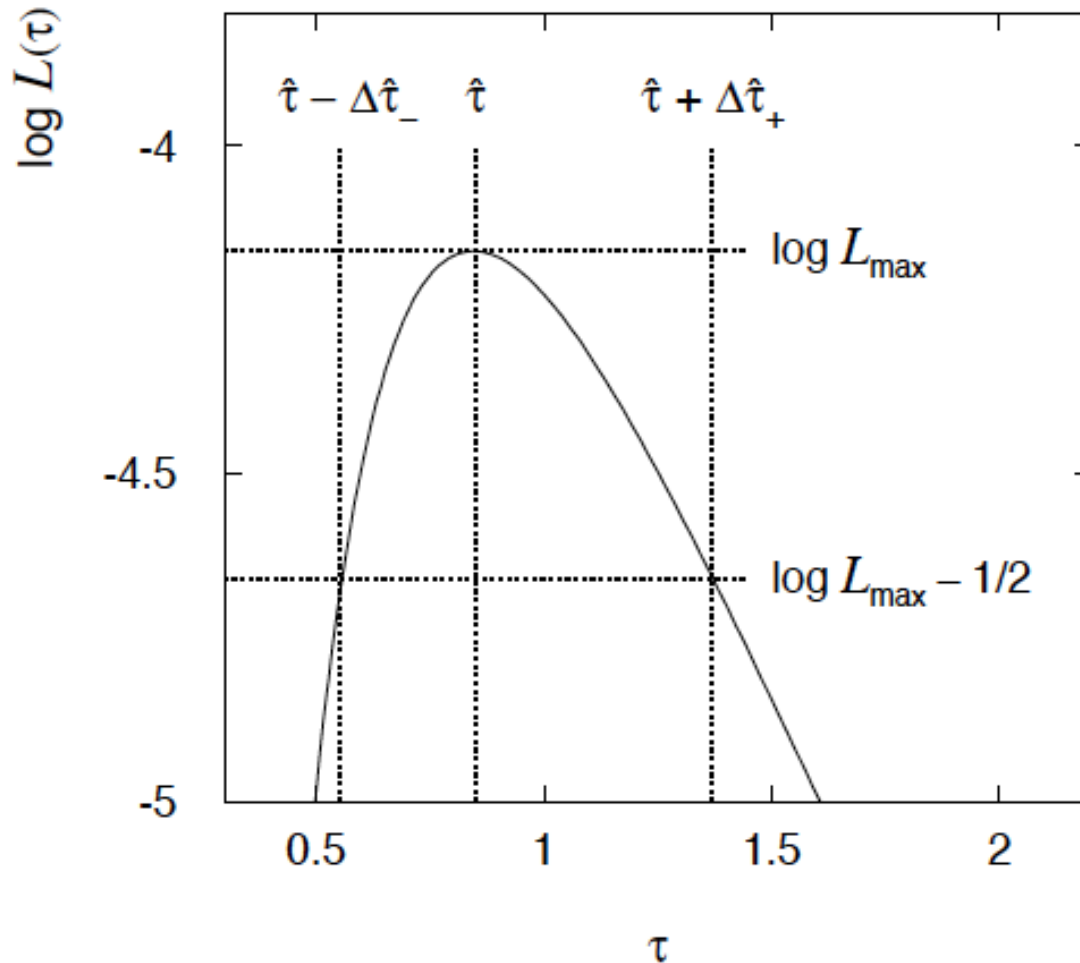
$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

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

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



Our exponential example, now with only $n = 5$ events.

Can report ML estimate with approx. confidence interval from $\ln L_{\max} - 1/2$ as “asymmetric error bar”:

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

Multiparameter case

For increasing number of parameters, $CL = 1 - \alpha$ decreases for confidence region determined by a given

$$Q_\alpha = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Q_α	$1 - \alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
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

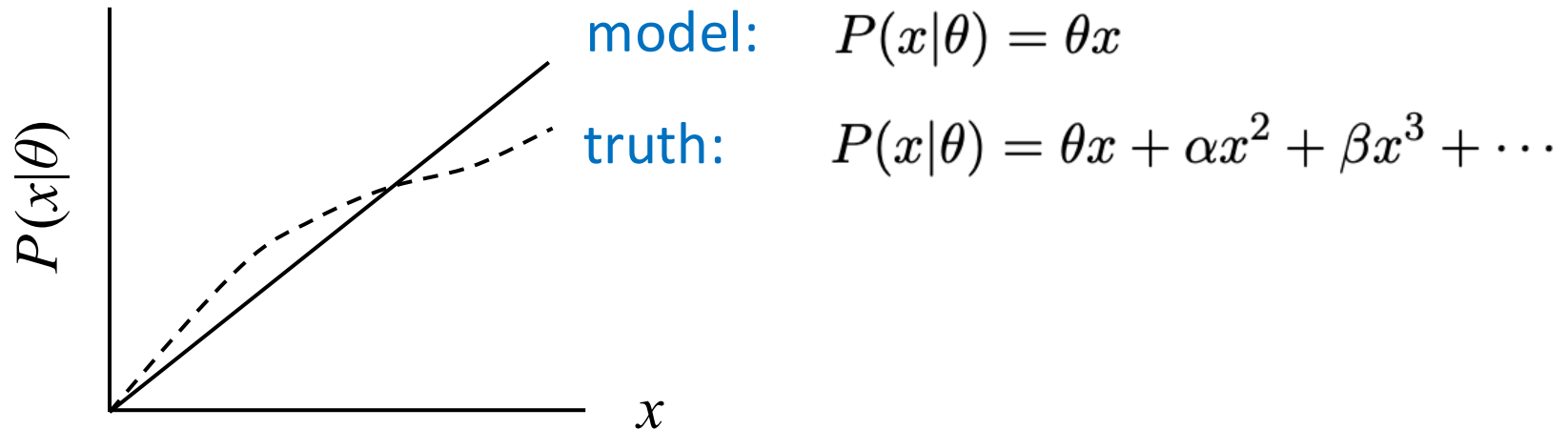
Multiparameter case (cont.)

Equivalently, Q_α increases with n for a given $CL = 1 - \alpha$.

$1 - \alpha$	\tilde{Q}_α				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
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

Systematic uncertainties and nuisance parameters

In general, our model of the data is not perfect:



Can improve model by including additional adjustable parameters.

$$P(x|\theta) \rightarrow P(x|\theta, \nu)$$

Nuisance parameter \leftrightarrow systematic uncertainty. Some point in the parameter space of the enlarged model should be “true”.

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

Profile Likelihood

Suppose we have a likelihood $L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\theta})$ with N parameters of interest $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$ and M nuisance parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$. The “profiled” (or “constrained”) values of $\boldsymbol{\theta}$ are:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\mu}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\mu}, \boldsymbol{\theta})$$

and the profile likelihood is: $L_p(\boldsymbol{\mu}) = L(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}})$

The profile likelihood depends only on the parameters of interest; the nuisance parameters are replaced by their profiled values.

The profile likelihood can be used to obtain confidence intervals/regions for the parameters of interest in the same way as one would for all of the parameters from the full likelihood.

Profile Likelihood Ratio – Wilks theorem

Goal is to test/reject regions of μ space (param. of interest).

Rejecting a point μ should mean $p_\mu \leq \alpha$ for all possible values of the nuisance parameters θ .

Test μ using the “profile likelihood ratio”:
$$\lambda(\mu) = \frac{L(\mu, \hat{\theta})}{L(\hat{\mu}, \hat{\theta})}$$

Let $t_\mu = -2 \ln \lambda(\mu)$. Wilks' theorem says in large-sample limit:

$$t_\mu \sim \text{chi-square}(N)$$

where the number of degrees of freedom is the number of parameters of interest (components of μ). So p -value for μ is

$$p_\mu = \int_{t_{\mu,\text{obs}}}^{\infty} f(t_\mu | \mu, \theta) dt_\mu = 1 - F_{\chi_N^2}(t_{\mu,\text{obs}})$$

Profile Likelihood Ratio – Wilks theorem (2)

If we have a large enough data sample to justify use of the asymptotic chi-square pdf, then if μ is rejected, it is rejected for any values of the nuisance parameters.

The recipe to get confidence regions/intervals for the parameters of interest at $CL = 1 - \alpha$ is thus the same as before, simply use the profile likelihood:

$$\ln L_p(\mu) = \ln L_{\max} - \frac{1}{2} F_{\chi_N^2}^{-1}(1 - \alpha)$$

where the number of degrees of freedom N for the chi-square quantile is equal to the number of parameters of interest.

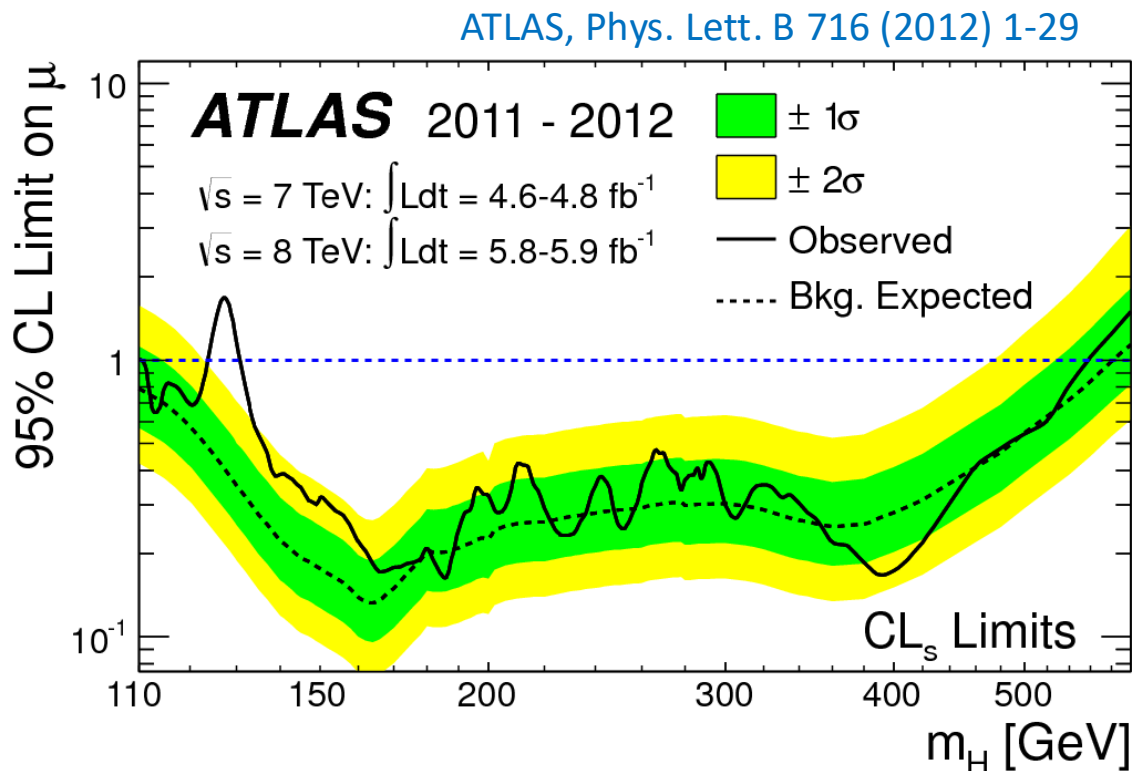
If the large-sample limit is not justified, then use e.g. Monte Carlo to get distribution of t_μ .

How to read the green and yellow limit plots

For every value of m_H , find the upper limit on μ .

Also for each m_H , determine the distribution of upper limits μ_{up} one would obtain under the hypothesis of $\mu = 0$.

The dashed curve is the median μ_{up} , and the green (yellow) bands give the $\pm 1\sigma$ (2σ) regions of this distribution.

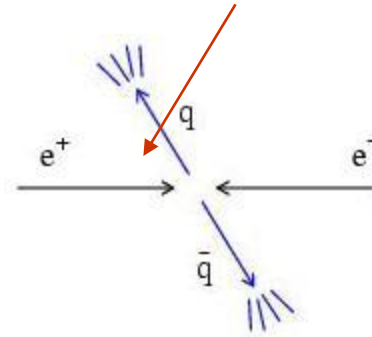


Extra slides

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.

$$\ln L(\alpha, \beta) = \sum_{i=1}^n \ln f(x_i; \alpha, \beta) \quad \leftarrow \text{need to find maximum numerically}$$

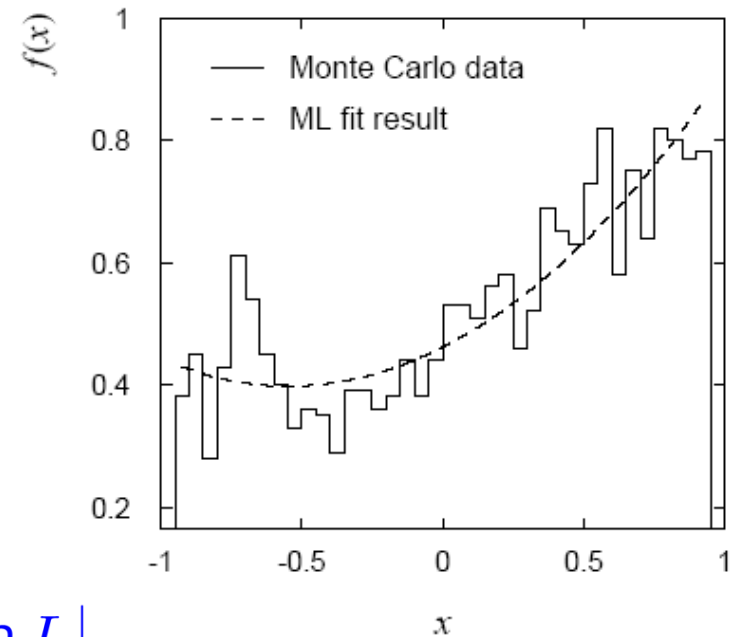
Example of ML with 2 parameters: fit result

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

$$\hat{\sigma}_{\hat{\alpha}} = 0.052$$

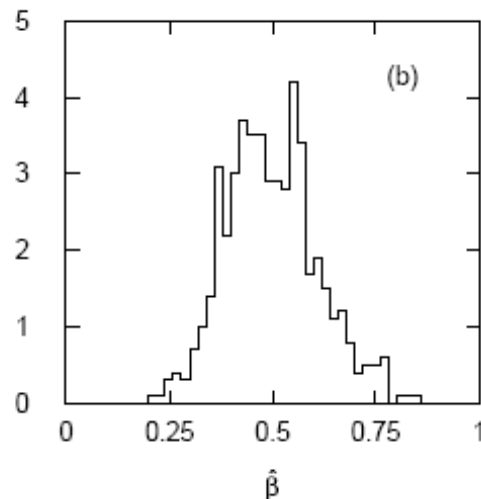
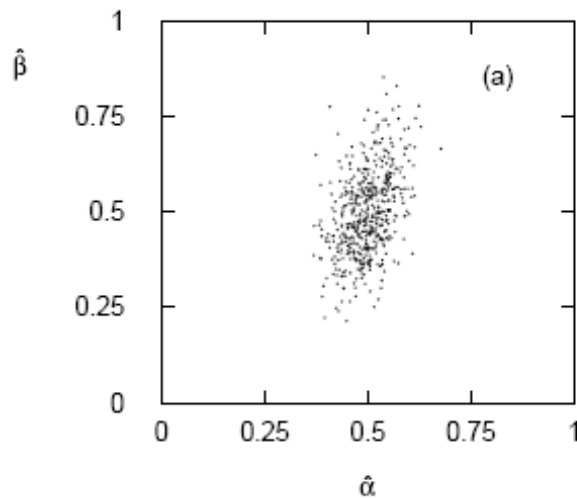
$$\text{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$

$$\hat{\sigma}_{\hat{\beta}} = 0.11$$

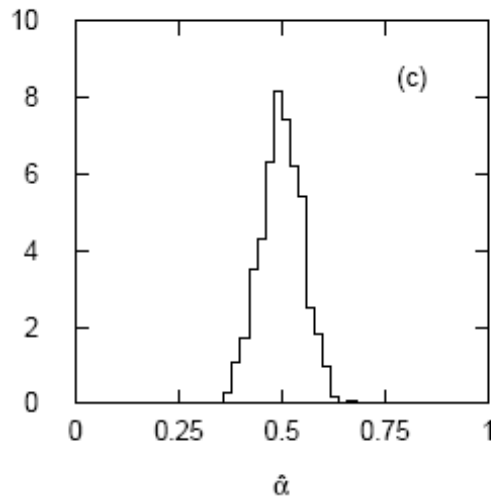
$$r = 0.46 = \text{correlation coefficient}$$

Two-parameter fit: MC study

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



$$\begin{aligned}\overline{\hat{\alpha}} &= 0.499 \\ s_{\hat{\alpha}} &= 0.051 \\ \overline{\hat{\beta}} &= 0.498 \\ s_{\hat{\beta}} &= 0.111 \\ \text{cov}[\hat{\alpha}, \hat{\beta}] &= 0.0024 \\ r &= 0.42\end{aligned}$$



Estimates average to \sim 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:

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

$\ln L_{\max}$ zero

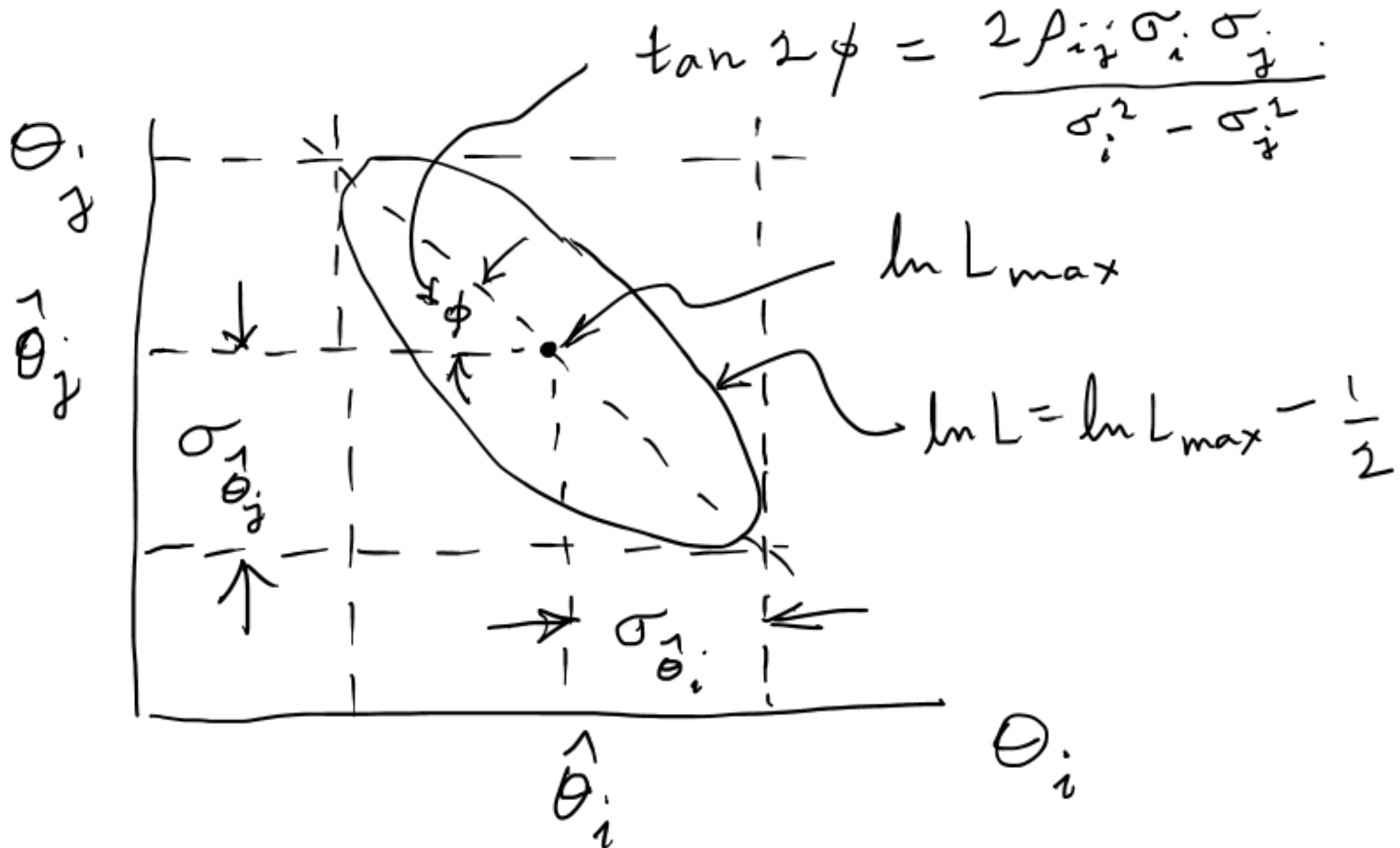
relate to covariance matrix of MLEs using information (in)equality.

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

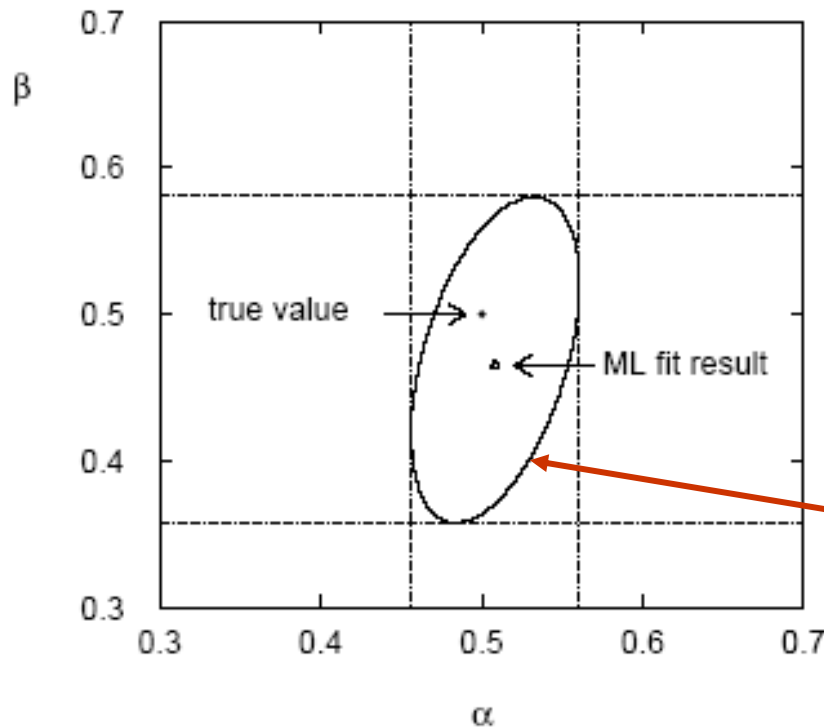
For large n , $\ln L$ takes on quadratic form near maximum:

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

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: $\tan 2\phi = \frac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma_{\hat{\beta}}^2}$

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 ?

For now 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.

If function many-to-one, some complications but related recipe.
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 $E[\hat{\lambda}] = \lambda \frac{n}{n-1}$. (bias $\rightarrow 0$ for $n \rightarrow \infty$)

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

Relationship between ML and Bayesian estimators

Recall the Bayesian approach:

Both θ and \mathbf{x} are random variables.

Use subjective probability for hypotheses (θ);

before experiment, knowledge summarized by prior pdf $\pi(\theta)$;

use Bayes' theorem to update prior in light of data:

$$p(\theta|\vec{x}) = \frac{L(\vec{x}|\theta)\pi(\theta)}{\int L(\vec{x}|\theta')\pi(\theta') d\theta'}$$



Posterior pdf (conditional pdf for θ given \mathbf{x})

ML and Bayesian estimators (2)

Purist Bayesian: $p(\theta|x)$ contains all knowledge about θ .

Pragmatist Bayesian: $p(\theta|x)$ could be a complicated function,

→ summarize using an estimator $\hat{\theta}_{\text{Bayes}}$

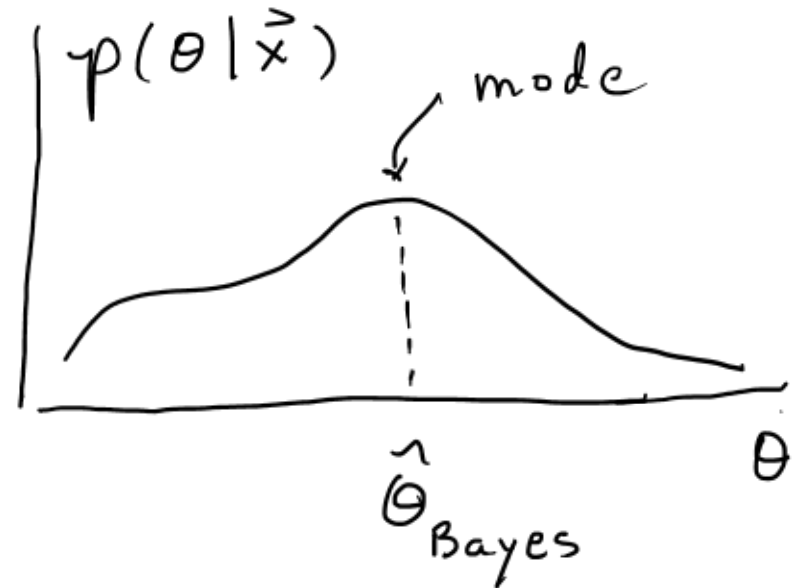
Take mode of $p(\theta|x)$, (could also use e.g. expectation value)

What do we use for $\pi(\theta)$?

No golden rule (subjective!),
often represent 'prior ignorance'
by $\pi(\theta) = \text{constant}$, in which case

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

$$\longrightarrow \hat{\theta}_{\text{Bayes}} = \hat{\theta}_{\text{ML}}$$



ML and Bayesian estimators (3)

Note $\pi_\theta(\theta) = \text{const.}$ cannot be normalized – “improper prior”.

Can be allowed for some problems; prior always appears multiplied by likelihood, so product $L(\theta)\pi_\theta(\theta)$ can result in normalizable posterior probability.

But... we could have used a different parameter, e.g., $\lambda = 1/\theta$, and if prior $\pi_\theta(\theta)$ is constant, then $\pi_\lambda(\lambda)$ is not:

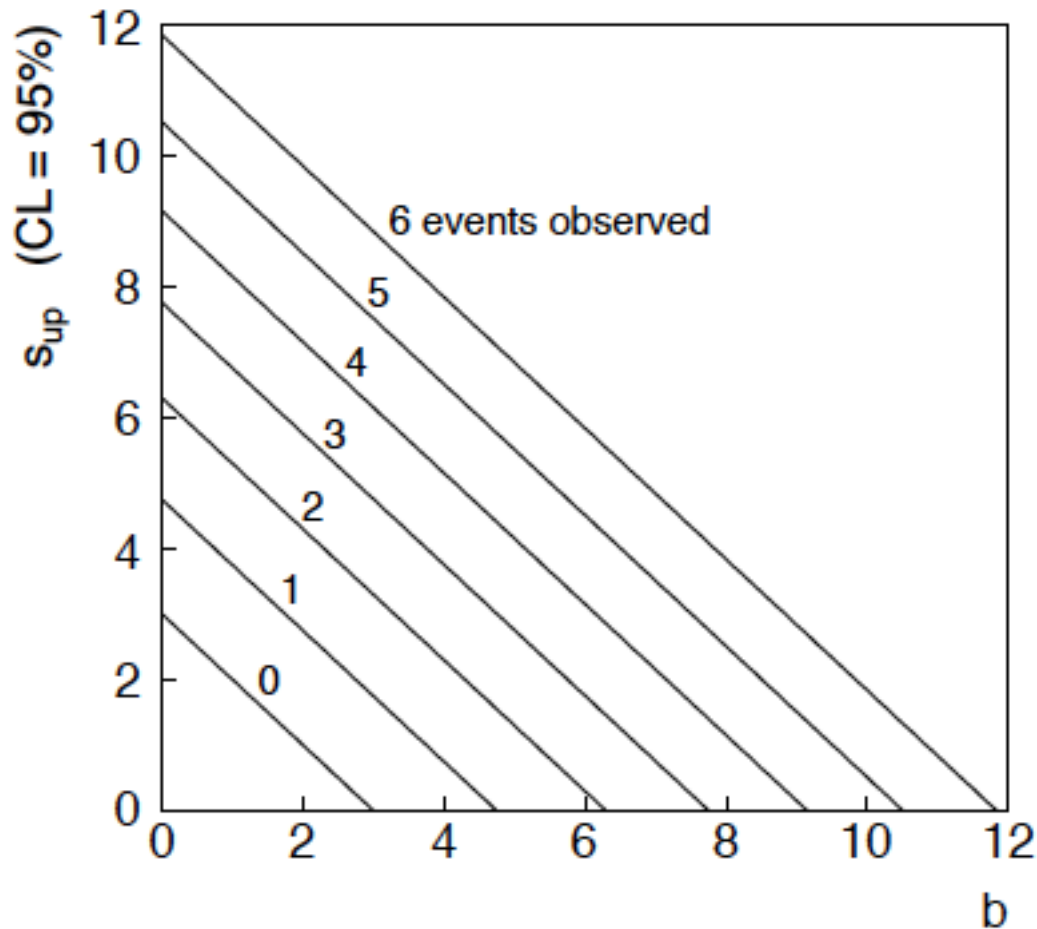
$$\pi_\lambda(\lambda) = \pi_\theta(\theta) \left| \frac{d\theta}{d\lambda} \right| \propto \frac{1}{\lambda^2}$$

Maybe we know say we nothing about λ , so take $\pi_\lambda(\lambda) = \text{const.}$

Then $\hat{\lambda}_{\text{Bayes}} = \hat{\lambda}_{ML} \neq \frac{1}{\hat{\theta}_{\text{Bayes}}}$ ‘Complete prior ignorance’ is not well defined.

$n \sim \text{Poisson}(s+b)$: frequentist upper limit on s

For low fluctuation of n , formula can give negative result for s_{up} ; i.e. confidence interval is empty; all values of $s \geq 0$ have $p_s \leq \alpha$.



Limits near a boundary of the parameter space

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose $CL = 0.9$, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (CL = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when testing parameter values for which one has very little experimental sensitivity, e.g., very small s .

Expected limit for $s = 0$

Physicist: I should have used $CL = 0.95$ — then $s_{up} = 0.496$

Even better: for $CL = 0.917923$ we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44

