

Statistics for Particle Physics

Lecture 3: Parameter Estimation

<https://indico.cern.ch/event/747653/>



INSIGHTS Workshop on
Statistics and Machine Learning
CERN, 17-21 September, 2018



Glen Cowan

Physics Department

Royal Holloway, University of London

g.cowan@rhul.ac.uk

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

Outline

Lecture 1: Fundamentals

Probability

Random variables, pdfs

Lecture 2: Statistical tests

Formalism of frequentist tests

Comments on multivariate methods (brief)

p -values

Discovery and limits

→ Lecture 3: Parameter estimation

Properties of estimators

Maximum likelihood

Parameter estimation

The parameters of a pdf are constants that characterize its shape, e.g.

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

r.v. parameter

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

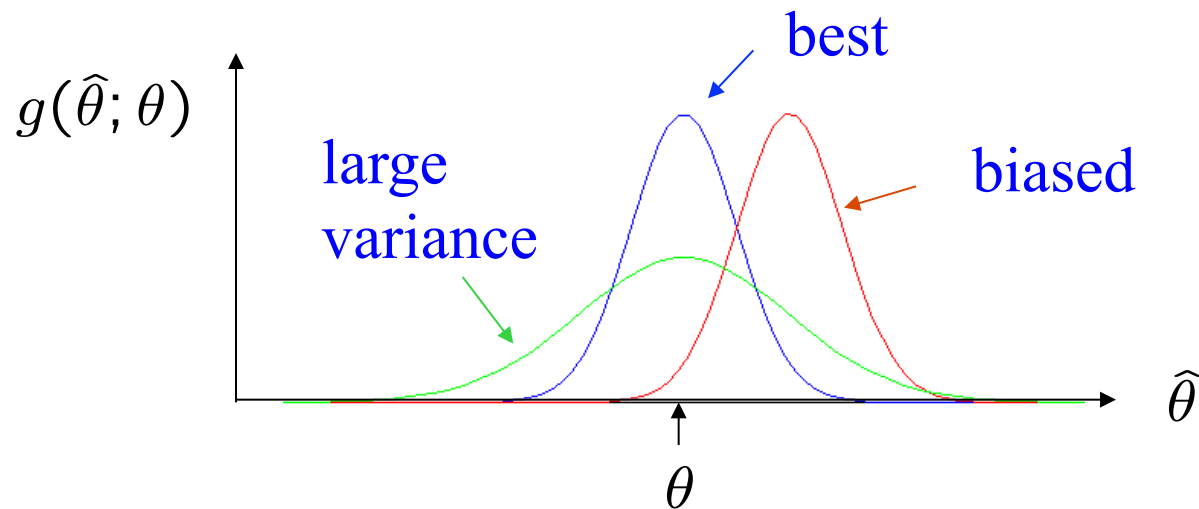
We want to find some function of the data to **estimate** the parameter(s):

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

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

An estimator for the mean (expectation value)

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

Estimator: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \equiv \bar{x}$ ('sample mean')

We find: $b = E[\hat{\mu}] - \mu = 0$

$$V[\hat{\mu}] = \frac{\sigma^2}{n} \quad \left(\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} \right)$$

An estimator for the variance

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

Estimator: $\widehat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \equiv s^2$ ('sample variance')

We find:

$$b = E[\widehat{\sigma}^2] - \sigma^2 = 0 \quad (\text{factor of } n-1 \text{ makes this so})$$

$$V[\widehat{\sigma}^2] = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \mu_2 \right), \quad \text{where}$$

$$\mu_k = \int (x - \mu)^k f(x) dx$$

The likelihood function

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

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

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

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

(\mathbf{x} constant)

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

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

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

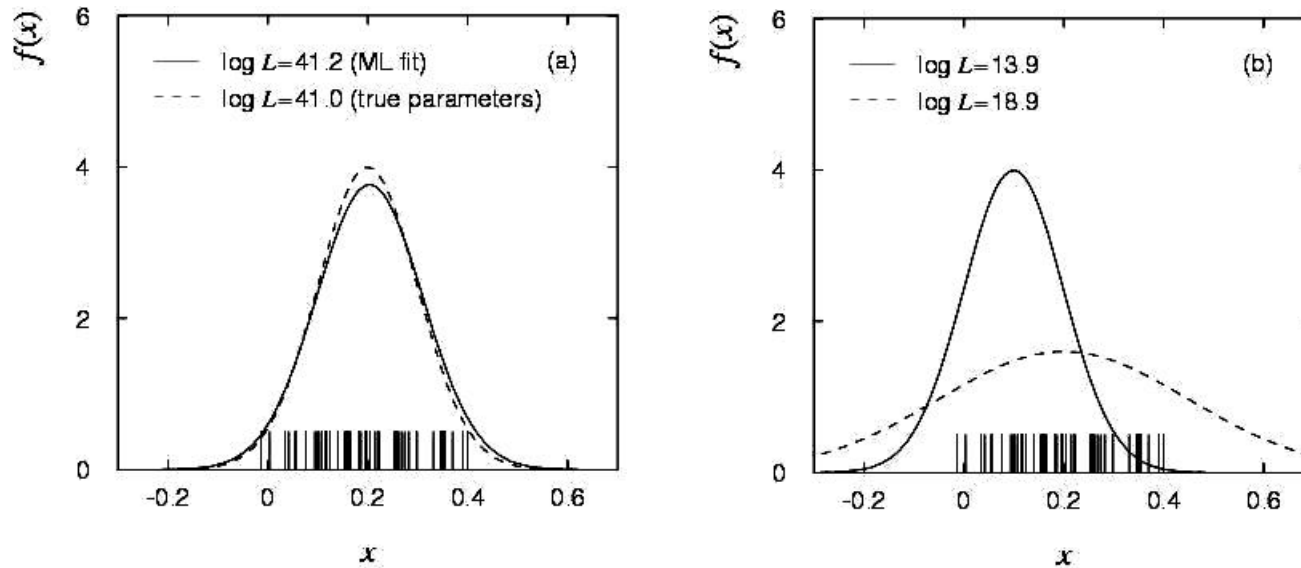
$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

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

If the hypothesized θ is close to the true value, then we expect a high probability to get data like that which we actually found.



So we define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

ML estimators not guaranteed to have any ‘optimal’ properties, (but in practice they’re very good).

ML 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):

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

ML example: parameter of exponential pdf (2)

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

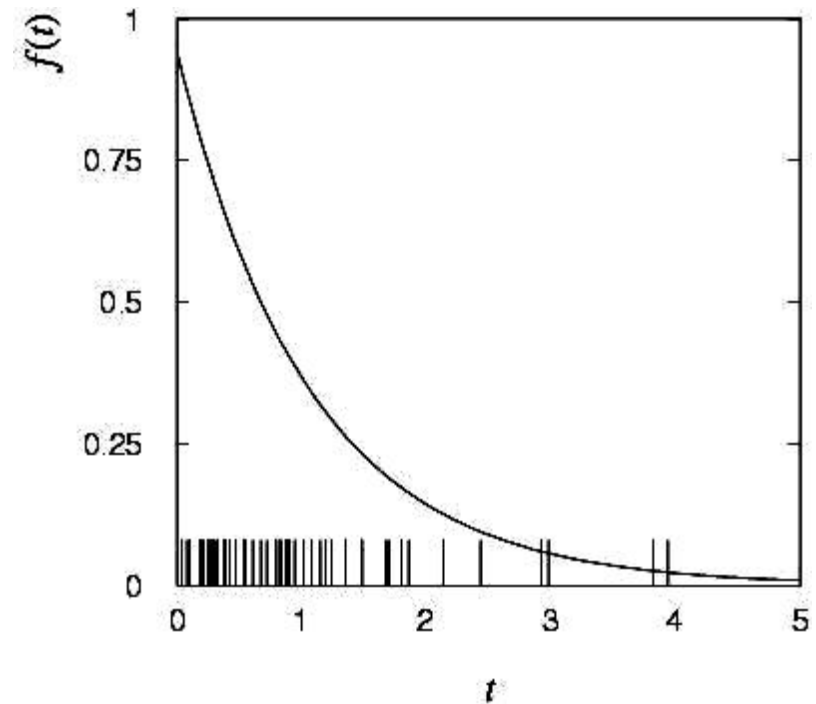
$$\rightarrow \hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$$

Monte Carlo test:

generate 50 values
using $\tau = 1$:

We find the ML estimate:

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



ML 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 ML estimator $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$ we therefore find

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

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

Functions of ML estimators

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 ML estimator for λ ?

For a function (with unique inverse) $\lambda(\tau)$ of a parameter τ , it doesn't matter whether we express L as a function of λ or τ .

The ML estimator of a function $\lambda(\tau)$ is simply $\hat{\lambda} = \lambda(\hat{\tau})$

So 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 $n \rightarrow \infty$)

Example of ML: parameters of Gaussian pdf

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

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

The log-likelihood function is

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

Example of ML: parameters of Gaussian pdf (2)

Set derivatives with respect to μ , σ^2 to zero and solve,

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

We already know that the estimator for μ is unbiased.

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

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

is an unbiased estimator for σ^2 .

Variance of estimators: Monte Carlo method

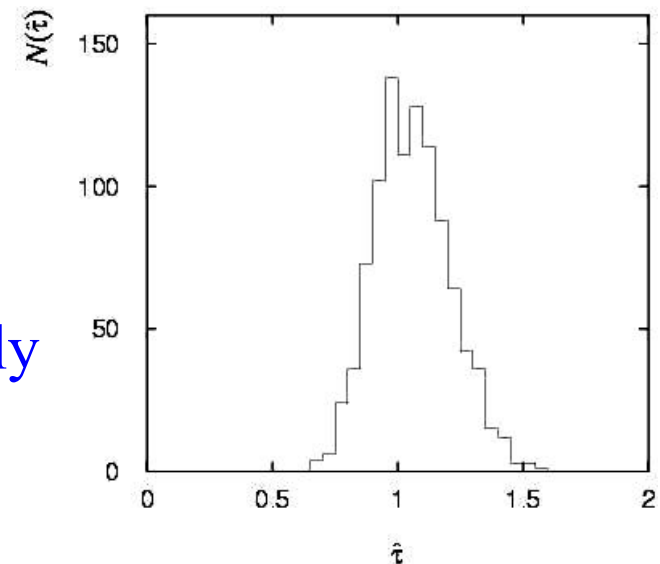
Having estimated our parameter we now need to report its ‘statistical error’, i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

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

Note distribution of estimates is roughly Gaussian – (almost) always true for ML in large sample limit.



Variance of estimators from information inequality

The **information inequality** (RCF) sets a lower bound on the variance of any estimator (not only ML):

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

← Minimum Variance Bound (MVB)
($b = E[\hat{\theta}] - \theta$)

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

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

Estimate this using the 2nd derivative of $\ln L$ at its maximum:

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

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):

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

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

→ 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

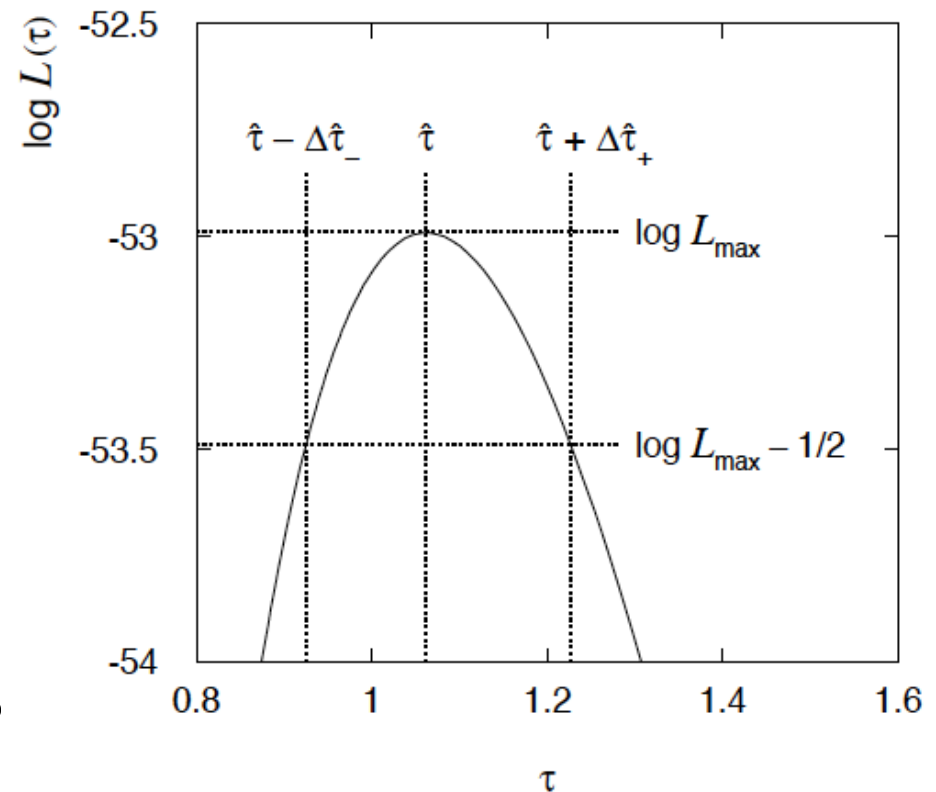
ML example with exponential:

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

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

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

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



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

Information inequality for n parameters

Suppose we have estimated n parameters $\vec{\theta} = (\theta_1, \dots, \theta_n)$.

The (inverse) minimum variance bound is given by the Fisher information matrix:

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

The information inequality then states that $V - I^{-1}$ is a positive semi-definite matrix, where $V_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$. Therefore

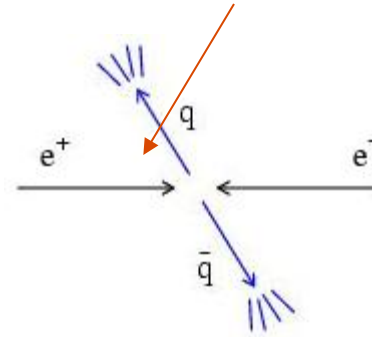
$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use I^{-1} as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of L .

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 always 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.

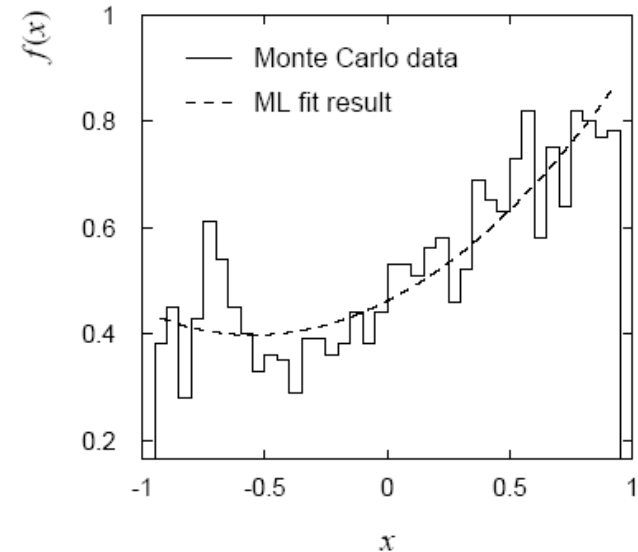
Example of ML with 2 parameters: fit result

Finding maximum of $\ln L(\alpha, \beta)$ numerically (**MINUIT**) 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).



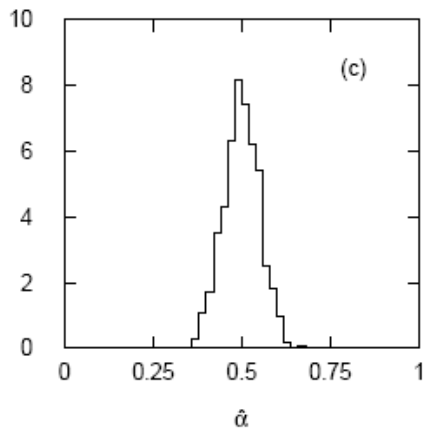
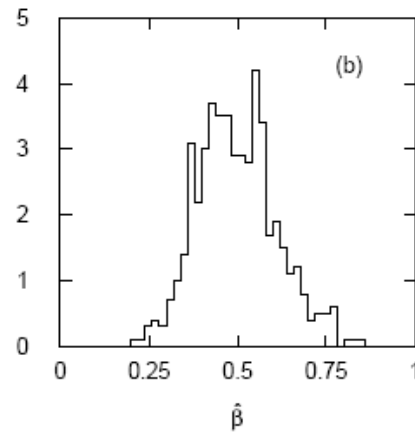
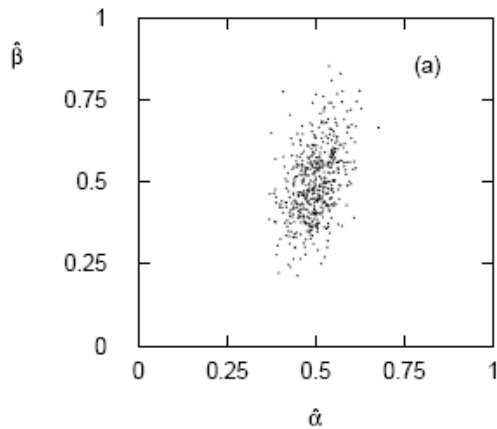
(Co)variances from $(\widehat{V}^{-1})_{ij} = -\left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\vec{\theta}=\vec{\hat{\theta}}}$ (**MINUIT routine HESSE**)

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

$$\hat{\sigma}_{\hat{\beta}} = 0.11 \quad r = 0.46$$

Two-parameter fit: MC study

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



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

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

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

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

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

$$r = 0.42$$

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

The $\ln L_{\max} - 1/2$ contour

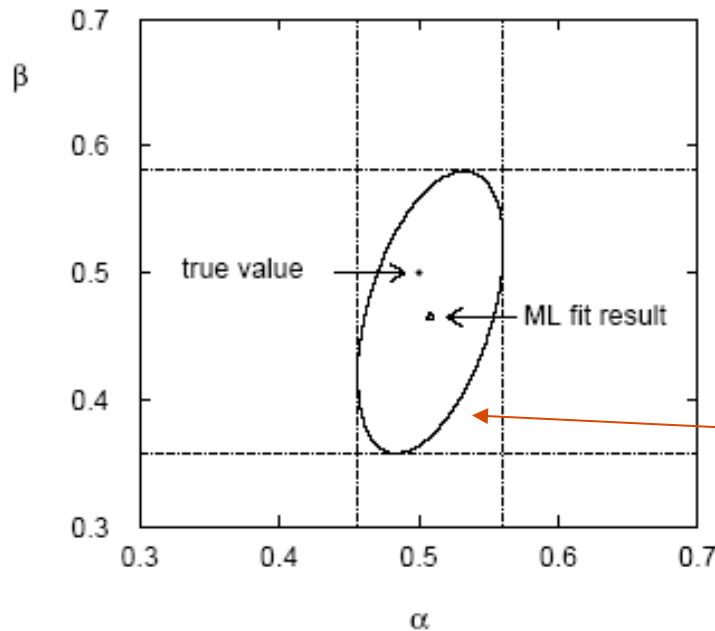
For large n , $\ln L$ takes on quadratic form near maximum:

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

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

Correlations between estimators result in an increase in their standard deviations (statistical errors).

ML with binned data

Often put data into a histogram: $\vec{n} = (n_1, \dots, n_N)$, $n_{\text{tot}} = \sum_{i=1}^N n_i$

Hypothesis is $\vec{\nu} = (\nu_1, \dots, \nu_N)$, $\nu_{\text{tot}} = \sum_{i=1}^N \nu_i$ where

$$\nu_i(\vec{\theta}) = \nu_{\text{tot}} \int_{\text{bin } i} f(x; \vec{\theta}) dx$$

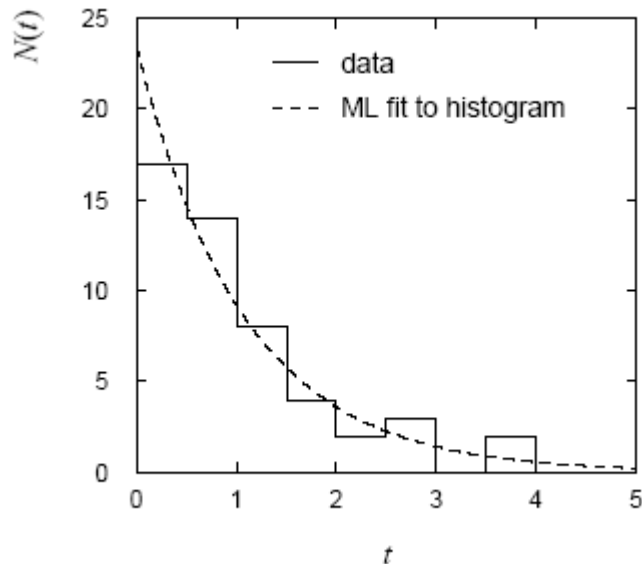
If we model the data as multinomial (n_{tot} constant),

$$f(\vec{n}; \vec{\nu}) = \frac{n_{\text{tot}}!}{n_1! \dots n_N!} \left(\frac{\nu_1}{n_{\text{tot}}} \right)^{n_1} \dots \left(\frac{\nu_N}{n_{\text{tot}}} \right)^{n_N}$$

then the log-likelihood function is: $\ln L(\vec{\theta}) = \sum_{i=1}^N n_i \ln \nu_i(\vec{\theta}) + C$

ML example with binned data

Previous example with exponential, now put data into histogram:



$$\hat{\tau} = 1.07 \pm 0.17$$

(1.06 \pm 0.15 for unbinned

ML with same sample)

Limit of zero bin width \rightarrow usual unbinned ML.

If n_i treated as Poisson, we get extended log-likelihood:

$$\ln L(\nu_{\text{tot}}, \vec{\theta}) = -\nu_{\text{tot}} + \sum_{i=1}^N n_i \ln \nu_i(\nu_{\text{tot}}, \vec{\theta}) + C$$

Relationship between ML and Bayesian estimators

In Bayesian statistics, both θ and \mathbf{x} are random variables:


$$L(\theta) = L(\vec{x}|\theta) = f_{\text{joint}}(\vec{x}|\theta)$$

Recall the Bayesian method:

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

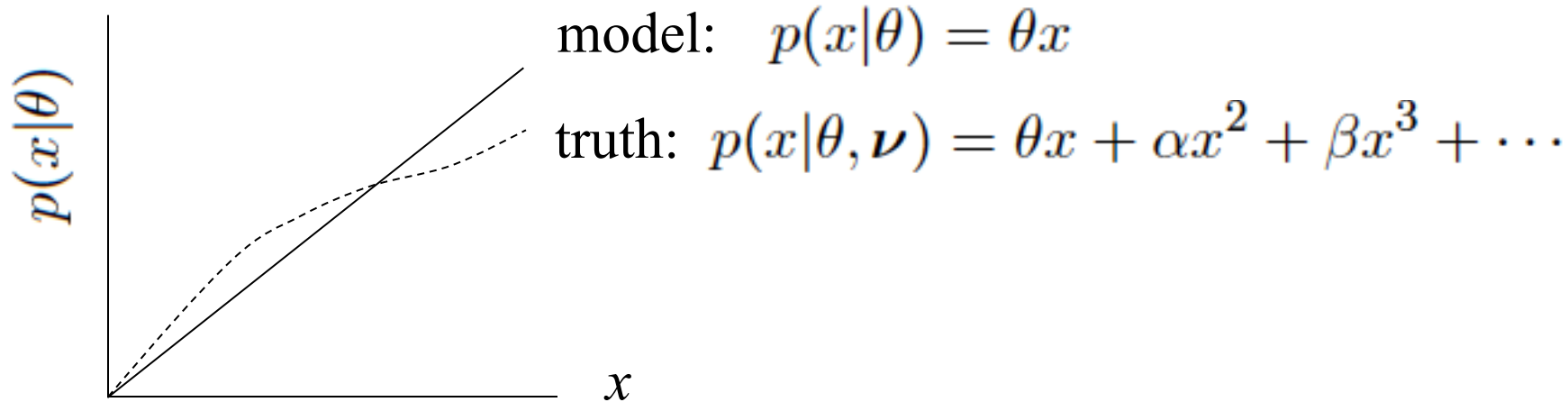
$$\hat{\theta}_{\text{Bayes}} = \hat{\theta}_{\text{ML}}$$

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

‘Complete prior ignorance’ is not well defined.

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).

Example: fitting a straight line

Data: (x_i, y_i, σ_i) , $i = 1, \dots, n$.

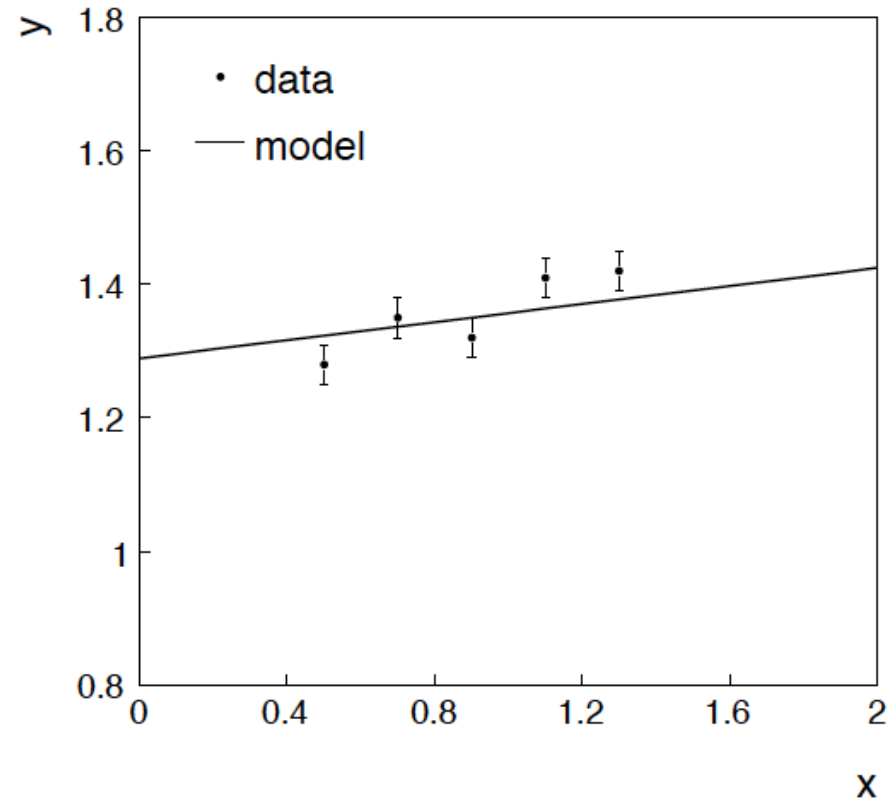
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x,$$

assume x_i and σ_i known.

Goal: estimate θ_0

Here suppose we don't care about θ_1 (example of a “nuisance parameter”)



Maximum likelihood fit with Gaussian data

In this example, the y_i are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right],$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

θ_1 known a priori

$$L(\theta_0) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right].$$

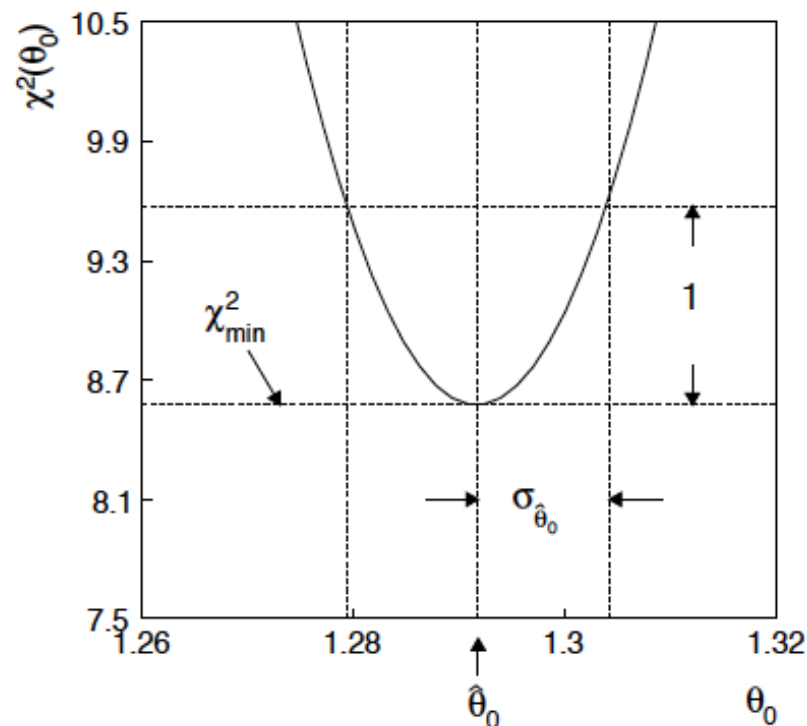
$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

For Gaussian y_i , ML same as LS

Minimize $\chi^2 \rightarrow$ estimator $\hat{\theta}_0$.

Come up one unit from χ_{\min}^2

to find $\sigma_{\hat{\theta}_0}$.



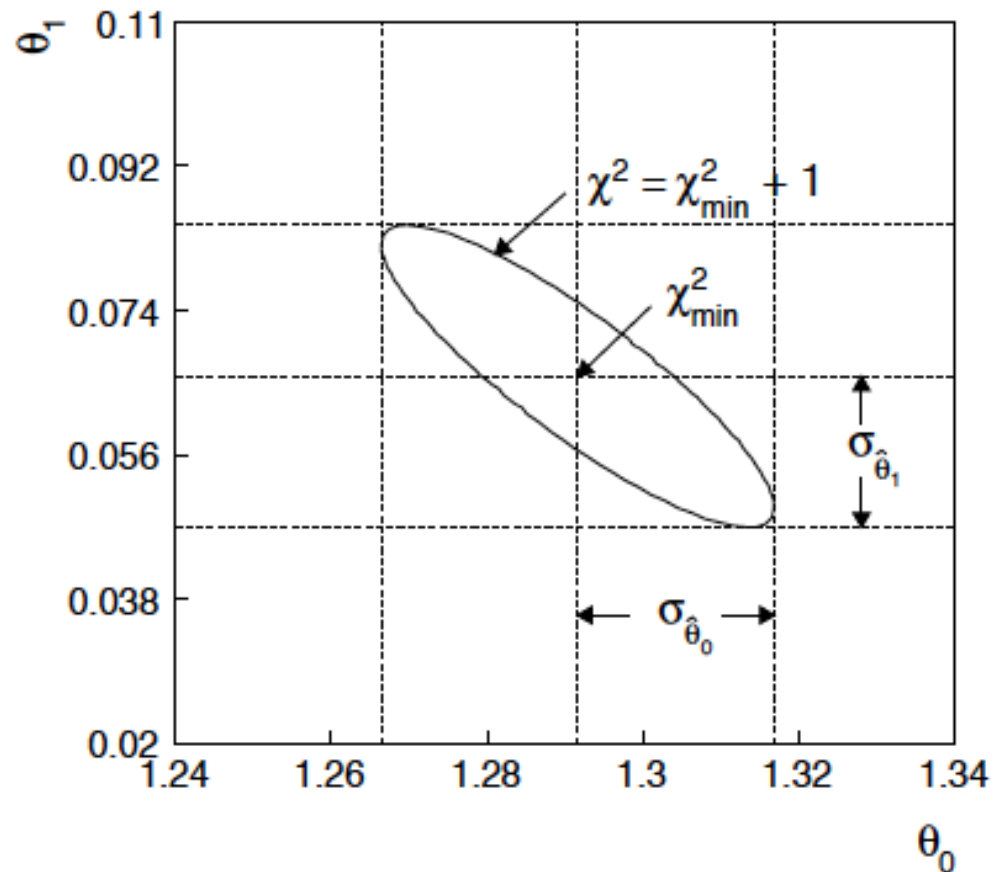
ML (or LS) fit of θ_0 and θ_1

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from
tangent lines to contour

$$\chi^2 = \chi_{\min}^2 + 1 .$$

Correlation between
 $\hat{\theta}_0$, $\hat{\theta}_1$ causes errors
to increase.

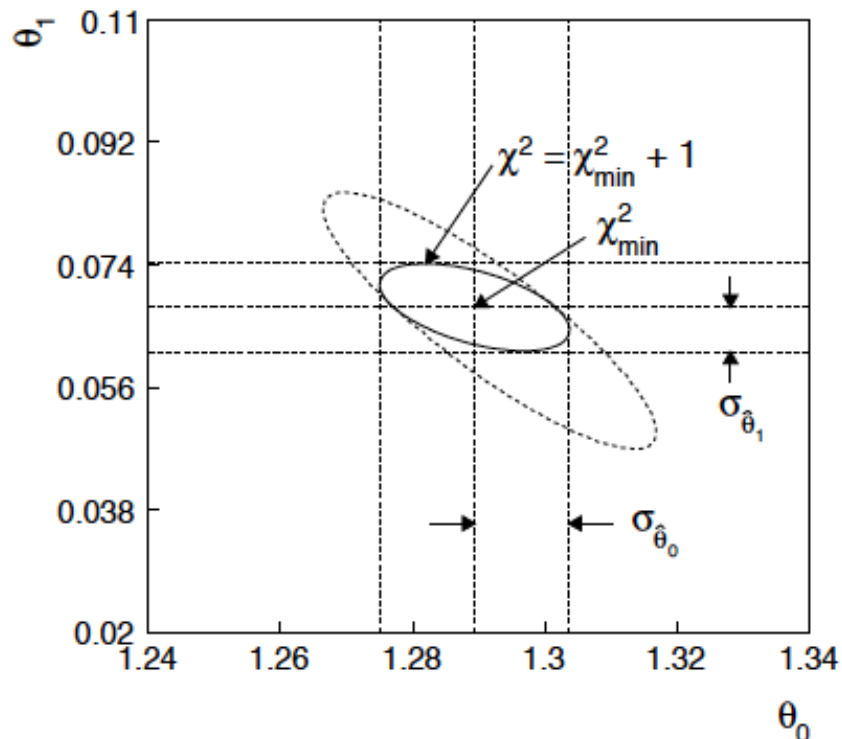


If we have a measurement $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t_1})$

$$L(\theta_0, \theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1 - \theta_1)^2 / 2\sigma_{t_1}^2} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right]$$

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(t_1 - \theta_1)^2}{\sigma_{t_1}^2}$$

The information on θ_1
improves accuracy of $\hat{\theta}_0$.



The Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter value θ .

Interpret probability of θ as ‘degree of belief’ (subjective).

Need to start with ‘**prior pdf**’ $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Our experiment has data x , \rightarrow **likelihood function** $L(x|\theta)$.

Bayes’ theorem tells how our beliefs should be updated in light of the data x :

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

Posterior pdf $p(\theta|x)$ contains all our knowledge about θ .

Bayesian method

We need to associate prior probabilities with θ_0 and θ_1 , e.g.,

$$\begin{aligned}\pi(\theta_0, \theta_1) &= \pi_0(\theta_0) \pi_1(\theta_1) && \text{'non-informative', in any} \\ \pi_0(\theta_0) &= \text{const.} && \text{case much broader than } L(\theta_0) \\ \pi_1(\theta_1) &= \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2} && \leftarrow \text{based on previous} \\ &&& \text{measurement}\end{aligned}$$

Putting this into Bayes' theorem gives:

$$p(\theta_0, \theta_1 | \vec{y}) \propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i; \theta_0, \theta_1))^2 / 2\sigma_i^2} \pi_0 \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2}$$

posterior \propto likelihood \times prior

Bayesian method (continued)

We then integrate (marginalize) $p(\theta_0, \theta_1 | x)$ to find $p(\theta_0 | x)$:

$$p(\theta_0 | x) = \int p(\theta_0, \theta_1 | x) d\theta_1 .$$

In this example we can do the integral (rare). We find

$$p(\theta_0 | x) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2 / 2\sigma_{\theta_0}^2} \quad \text{with}$$

$$\hat{\theta}_0 = \text{same as ML estimator}$$

$$\sigma_{\theta_0} = \sigma_{\hat{\theta}_0} \text{ (same as before)}$$

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x) d\theta_1 .$$

often high dimensionality and impossible in closed form,
also impossible with ‘normal’ acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.




MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than if all values independent .

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n -dimensional pdf $p(\vec{\theta})$,
generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \dots$

- 1) Start at some point $\vec{\theta}_0$
- 2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$  Proposal density $q(\vec{\theta}; \vec{\theta}_0)$
e.g. Gaussian centred
about $\vec{\theta}_0$
- 3) Form Hastings test ratio $\alpha = \min \left[1, \frac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)} \right]$
- 4) Generate $u \sim \text{Uniform}[0, 1]$
- 5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$,  move to proposed point
else $\vec{\theta}_1 = \vec{\theta}_0$  old point repeated
- 6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than if points were independent.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

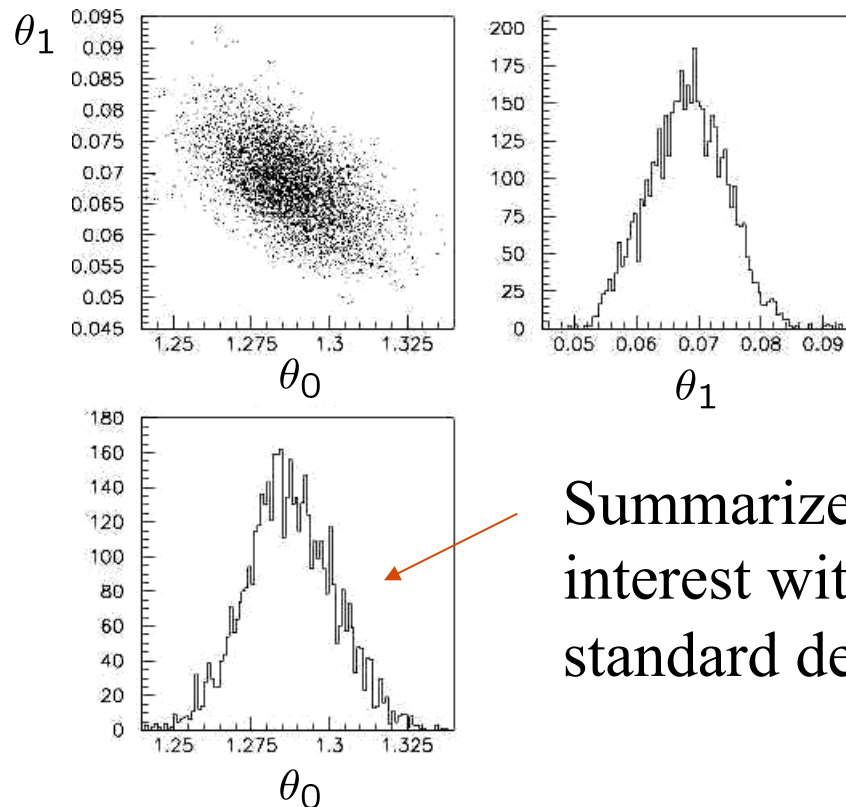
Test ratio is (*Metropolis-Hastings*): $\alpha = \min \left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.

If proposed step rejected, hop in place.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

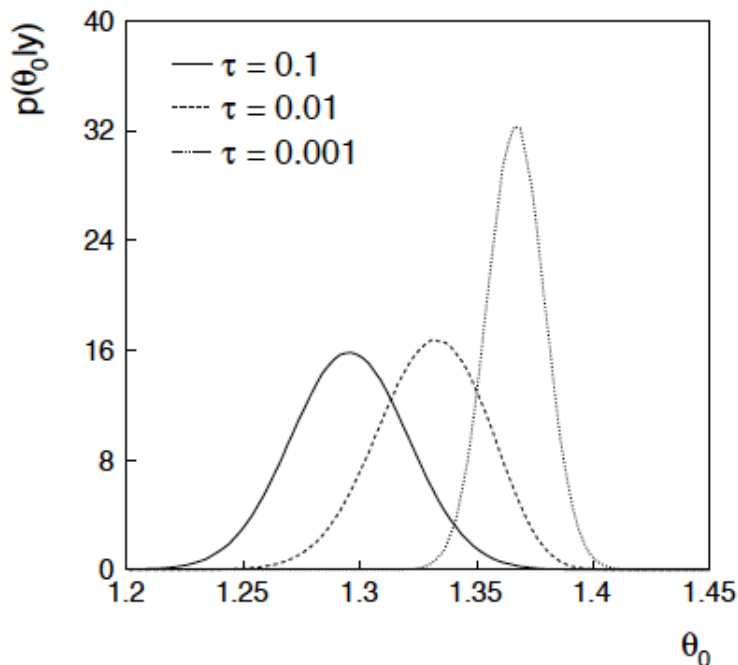
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, e.g., a theorist says it should be positive and not too much greater than 0.1 "or so", i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau}, \quad \theta_1 \geq 0, \quad \tau = 0.1 .$$

From this we obtain (numerically) the posterior pdf for θ_0 :



This summarizes all knowledge about θ_0 .

Look also at result from variety of priors.

Finally

In three introductory lectures we've only had time to touch on the basics of

Probability

Statistical tests

Parameter estimation

Many important topics yet to be covered, e.g.,

asymptotic methods,

experimental sensitivity,

look-elsewhere effect,...

More on this and more in future INSIGHTS training events.

Final thought: once the basic formalism is understood, most of the work focuses on writing down the likelihood, e.g., $P(\mathbf{x}|\boldsymbol{\theta})$, and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches).

Extra slides

Extended ML

Sometimes regard n not as fixed, but as a Poisson r.v., mean ν .

Result of experiment defined as: n, x_1, \dots, x_n .

The (extended) likelihood function is:

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^n f(x_i; \vec{\theta})$$

Suppose theory gives $\nu = \nu(\theta)$, then the log-likelihood is

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

where C represents terms not depending on θ .

Extended ML (2)

Example: expected number of events $\nu(\vec{\theta}) = \sigma(\vec{\theta}) \int L dt$
where the total cross section $\sigma(\theta)$ is predicted as a function of the parameters of a theory, as is the distribution of a variable x .

Extended ML uses more info \rightarrow smaller errors for $\hat{\vec{\theta}}$

Important e.g. for anomalous couplings in $e^+e^- \rightarrow W^+W^-$

If n does not depend on θ but remains a free parameter, extended ML gives:

$$\hat{\nu} = n$$

$$\hat{\theta} = \text{same as ML}$$

Extended ML example

Consider two types of events (e.g., signal and background) each of which predict a given pdf for the variable x : $f_s(x)$ and $f_b(x)$.

We observe a mixture of the two event types, signal fraction = θ , expected total number = ν , observed total number = n .

Let $\mu_s = \theta\nu$, $\mu_b = (1 - \theta)\nu$, goal is to estimate μ_s, μ_b .

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

$$P(n; \mu_s, \mu_b) = \frac{(\mu_s + \mu_b)^n}{n!} e^{-(\mu_s + \mu_b)}$$

$$\rightarrow \ln L(\mu_s, \mu_b) = -(\mu_s + \mu_b) + \sum_{i=1}^n \ln [(\mu_s + \mu_b) f(x_i; \mu_s, \mu_b)]$$

Extended ML example (2)

Monte Carlo example
with combination of
exponential and Gaussian:

$$\mu_s = 6$$

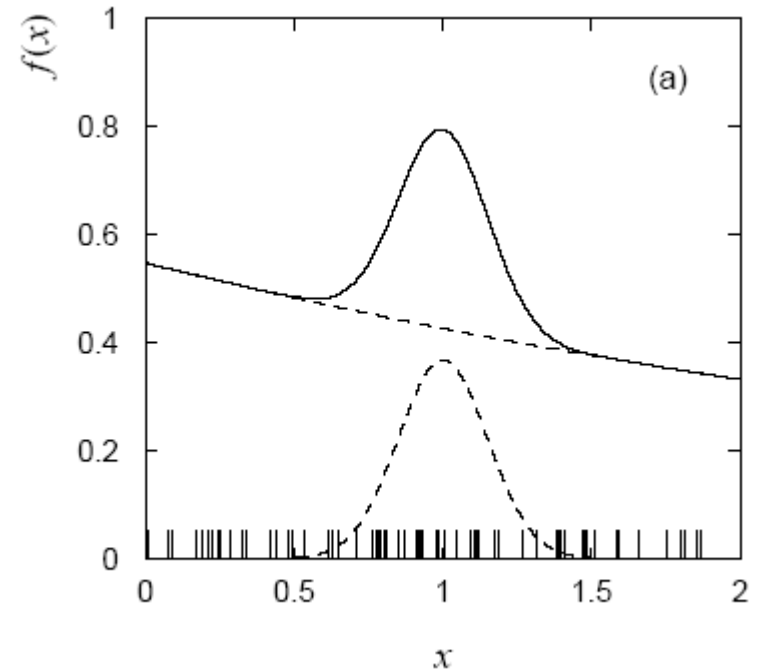
$$\mu_b = 60$$

Maximize log-likelihood in
terms of μ_s and μ_b :

$$\hat{\mu}_s = 8.7 \pm 5.5$$

$$\hat{\mu}_b = 54.3 \pm 8.8$$

Here errors reflect total Poisson
fluctuation as well as that in
proportion of signal/background.



Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called “objective priors”

Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In a Subjective Bayesian analysis, using objective priors can be an important part of the sensitivity analysis.

Priors from formal rules (cont.)

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes' theorem as a recipe to produce an interval with certain coverage properties. For a review see:

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in HEP, but there is recent interest in this direction; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, arxiv:1002.1111 (Feb 2010)

Jeffreys' prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

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

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters.

For a Gaussian mean, the Jeffreys' prior is constant; for a Poisson mean μ it is proportional to $1/\sqrt{\mu}$.

“Invariance of inference” with Jeffreys’ prior

Suppose we have a parameter θ , to which we assign a prior $\pi_\theta(\theta)$.

An experiment gives data x , modeled by $L(\theta) = P(x|\theta)$.

Bayes’ theorem then tells us the posterior for θ :

$$P(\theta|x) \propto P(x|\theta)\pi_\theta(\theta)$$

Now consider a function $\eta(\theta)$, and we want the posterior $P(\eta|x)$.

This must follow from the usual rules of transformation of random variables:

$$P(\eta|x) = P(\theta(\eta)|x) \left| \frac{d\theta}{d\eta} \right|$$

“Invariance of inference” with Jeffreys’ prior (2)

Alternatively, we could have just starting with η as the parameter in our model, and written down a prior pdf $\pi_\eta(\eta)$.

Using it, we express the likelihood as $L(\eta) = P(x|\eta)$ and write Bayes’ theorem as

$$P(\eta|x) \propto P(x|\eta)\pi_\eta(\eta)$$

If the priors really express our degree of belief, then they must be related by the usual laws of probability $\pi_\eta(\eta) = \pi_\theta(\theta(\eta)) |d\theta/d\eta|$, and in this way the two approaches lead to the same result.

But if we choose the priors according to “formal rules”, then this is not guaranteed. For the Jeffrey’s prior, however, it does work!

Using $\pi_\theta(\theta) \propto \sqrt{I(\theta)}$ and transforming to find $P(\eta|x)$ leads to the same as using $\pi_\eta(\eta) \propto \sqrt{I(\eta)}$ directly with Bayes’ theorem.

Jeffreys' prior for Poisson mean

Suppose $n \sim \text{Poisson}(\mu)$. To find the Jeffreys' prior for μ ,

$$L(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \qquad \frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\mu^2}$$

$$I = -E \left[\frac{\partial^2 \ln L}{\partial \mu^2} \right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for $\mu = s + b$, this means the prior $\pi(s) \sim 1/\sqrt{(s + b)}$, which depends on b . But this is not designed as a degree of belief about s .