Statistical Data Analysis: Lecture 9

1 Probability, Bayes' theorem, random variables, pdfs
2 Functions of r.v.s, expectation values, error propagation
3 Catalogue of pdfs
4 The Monte Carlo method
5 Statistical tests: general concepts
6 Test statistics, multivariate methods
7 Significance tests
8 Parameter estimation, maximum likelihood
→ 9 More maximum likelihood
10 Method of least squares
11 Interval estimation, setting limits
12 Nuisance parameters, systematic uncertainties
13 Examples of Bayesian approach
14 tba

Information inequality for *n* parameters

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

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

$$I_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right] = -n \int f(x;\vec{\theta}) \frac{\partial^2 \ln f(x;\vec{\theta})}{\partial\theta_i \partial\theta_j} \, dx$$

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

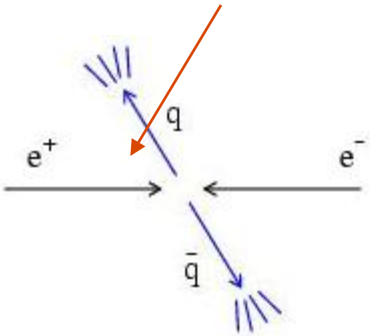
$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

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

$$\int_{x_{\min}}^{x_{\max}} f(x;\alpha,\beta)\,dx = 1 \ .$$

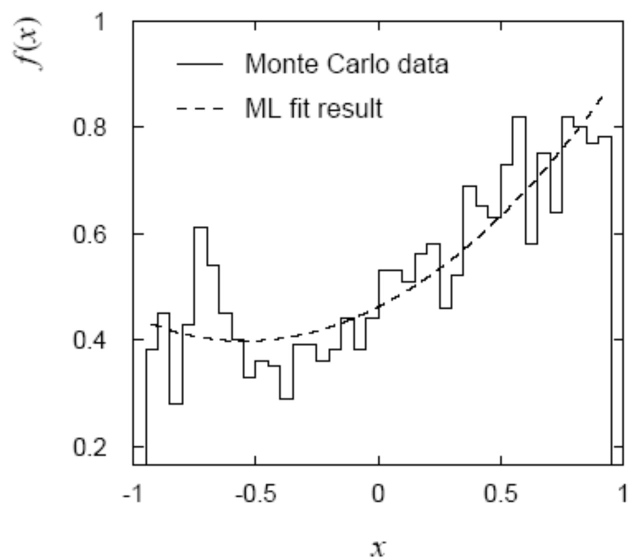
Example: $\alpha = 0.5$, $\beta = 0.5$, $x_{\min} = -0.95$, $x_{\max} = 0.95$, generate $n = 2000$ events with Monte Carlo.

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).

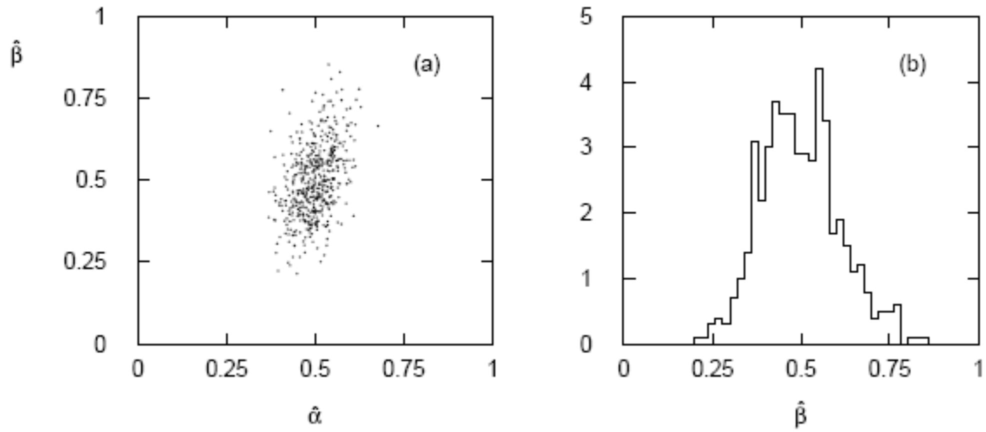


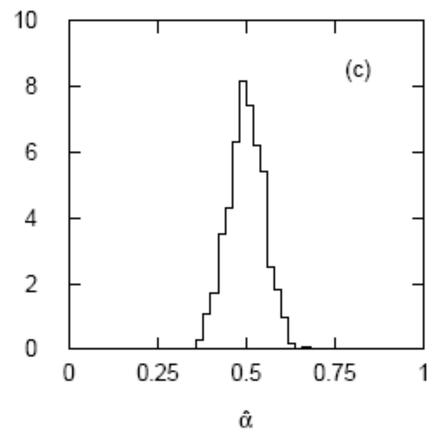
(Co)variances from $\widehat{(V^{-1})}_{ij} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\hat{\vec{\theta}}}$ (**`MINUIT`** routine **`HESSE`**)

$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \qquad \mathrm{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$
$$\hat{\sigma}_{\hat{\beta}} = 0.11 \qquad r = 0.46$$

Two-parameter fit: MC study

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





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

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

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

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

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

$$r = 0.42$$

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

The ln $L_{\max}$ − 1/2 contour

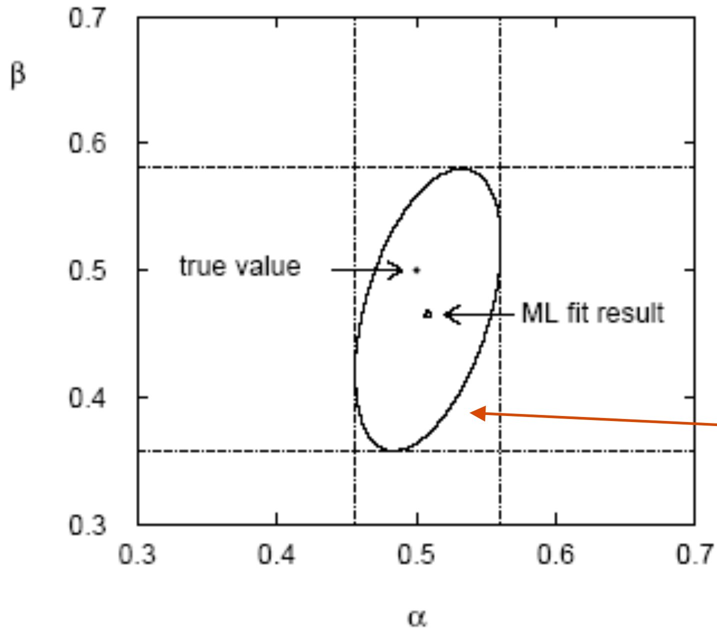
For large n, ln L takes on quadratic form near maximum:

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

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

$$\ln L(\alpha, \beta) = \ln L_{\mathrm{max}} - 1/2$$

→ Tangent lines to contours give standard deviations.

→ Angle of ellipse ϕ related to correlation: $\tan 2\phi = \dfrac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma\hat{\beta}^2}$

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

Extended ML

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

Result of experiment defined as: $n, x_1, ..., x_n$.

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 $\nu = \nu(\boldsymbol{\theta})$, 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 $\boldsymbol{\theta}$.

Extended ML (2)

Example: expected number of events $\nu(\vec{\theta}) = \sigma(\vec{\theta}) \int L \, dt$

where the total cross section $\sigma(\boldsymbol{\theta})$ is predicted as a function of the parameters of a theory, as is the distribution of a variable x.

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

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

If ν does not depend on $\boldsymbol{\theta}$ but remains a free parameter, extended ML gives:

$$\begin{aligned} \hat{\nu} &= n \\ \hat{\theta} &= \text{same as ML} \end{aligned}$$

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 $\mu_\mathrm{s} = \theta\nu$, $\mu_\mathrm{b} = (1-\theta)\nu$, goal is to estimate μ_s, μ_b.

$$f(x;\mu_\mathrm{s},\mu_\mathrm{b}) = \frac{\mu_\mathrm{s}}{\mu_\mathrm{s}+\mu_\mathrm{b}} f_\mathrm{s}(x) + \frac{\mu_\mathrm{b}}{\mu_\mathrm{s}+\mu_\mathrm{b}} f_\mathrm{b}(x)$$

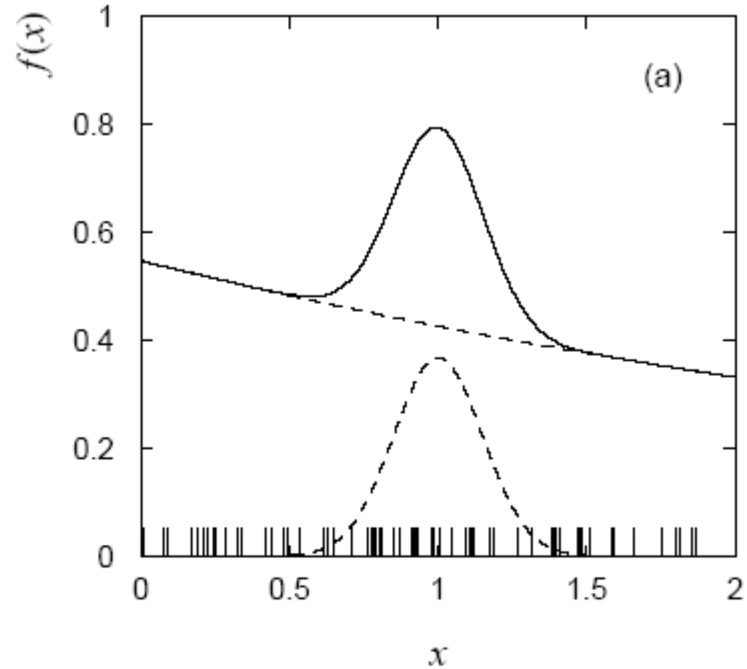
$$P(n;\mu_\mathrm{s},\mu_\mathrm{b}) = \frac{(\mu_\mathrm{s}+\mu_\mathrm{b})^n}{n!} e^{-(\mu_\mathrm{s}+\mu_\mathrm{b})}$$

$$\rightarrow \quad \ln L(\mu_\mathrm{s},\mu_\mathrm{b}) = -(\mu_\mathrm{s}+\mu_\mathrm{b}) + \sum_{i=1}^{n} \ln\left[(\mu_\mathrm{s}+\mu_\mathrm{b}) f(x_i;\mu_\mathrm{s},\mu_\mathrm{b})\right]$$

Extended ML example (2)

Monte Carlo example with combination of exponential and Gaussian:

$$\mu_{\rm s} = 6$$
$$\mu_{\rm b} = 60$$



Maximize log-likelihood in terms of μ_s and μ_b:

$$\hat{\mu}_{\rm s} = 8.7 \pm 5.5$$
$$\hat{\mu}_{\rm b} = 54.3 \pm 8.8$$

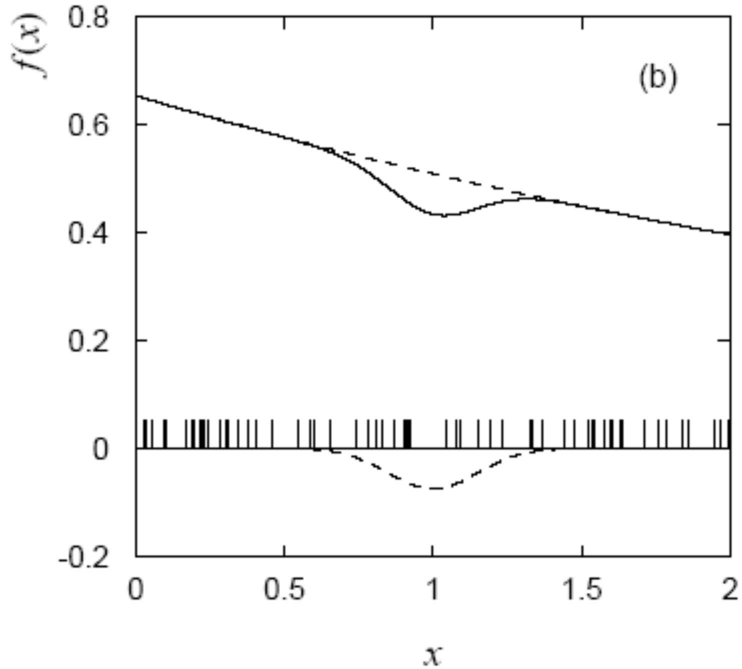
Here errors reflect total Poisson fluctuation as well as that in proportion of signal/background.

Extended ML example: an unphysical estimate

A downwards fluctuation of data in the peak region can lead to even fewer events than what would be obtained from background alone.

Estimate for μ_s here pushed negative (unphysical).

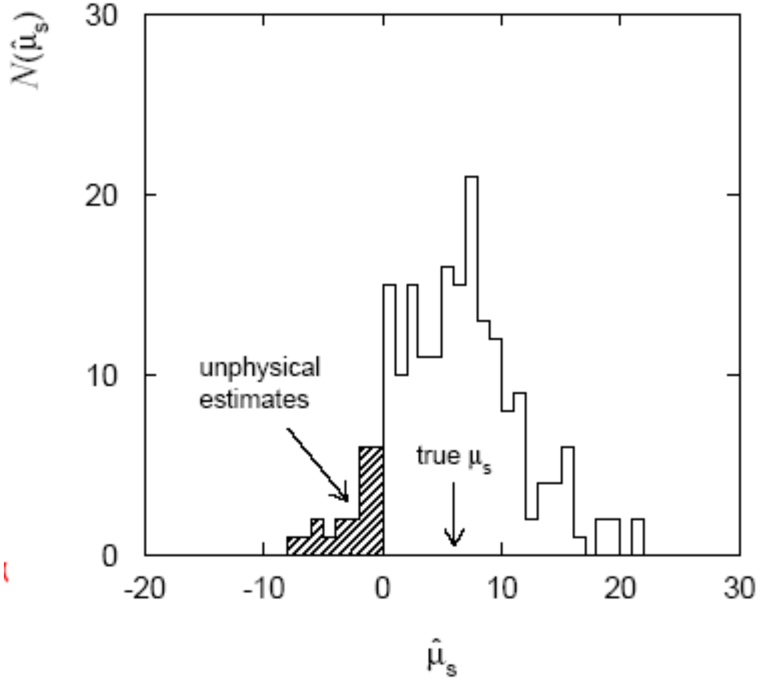
We can let this happen as long as the (total) pdf stays positive everywhere.



Unphysical estimators (2)

Here the unphysical estimator is unbiased and should nevertheless be reported, since average of a large number of unbiased estimates converges to the true value (cf. PDG).

Repeat entire MC experiment many times, allow unphysical estimates:



ML with binned data

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

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

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

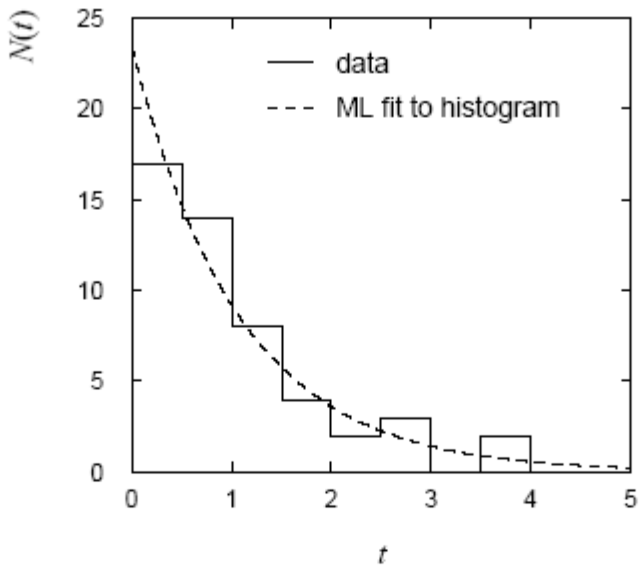
If we model the data as multinomial (n_{tot} constant),

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

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 ± 0.15 for unbinned ML with same sample)

Limit of zero bin width → 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 $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are random variables:

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

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 $\boldsymbol{x}$)

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 $\hat{\theta}_{\mathrm{Bayes}}$

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 $\pi(\theta)$ = constant, in which case

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

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

'Complete prior ignorance' is not well defined.

Wrapping up lecture 9

We've now seen several examples of the method of Maximum Likelihood:

multiparameter case
variable sample size (extended ML)
histogram-based data

and we've seen the connection between ML and Bayesian parameter estimation.

Next we will consider a special case of ML with Gaussian data and show how this leads to the method of Least Squares.

Extra slides

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

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

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 $1/\sqrt{\mu}$.

Jeffreys' prior for Poisson mean

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

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

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

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

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.