Lecture 6 outline

The method of maximum likelihood

1. **The likelihood function, ML estimators**
2. **Examples: parameters of exponential and Gaussian pdfs**
3. **Variance of ML estimators:**
   (a) Analytic method
   (b) Monte Carlo method
   (c) The RCF bound
   (d) Graphical method

The likelihood function

Consider data sample $\vec{x} = (x_1, \ldots, x_n)$ where x follows $f(x;\theta)$.

Goal: estimate θ (or in general $\vec{\theta} = (\theta_1, \ldots, \theta_m)$).

If $f(x;\theta)$ is true, then

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

If hypothesis (including value of θ) is true,

$\rightarrow$ expect high probability for the data we actually got.

If hypothesized θ far away from true value,

$\rightarrow$ low probability to have observed what we did.

$\Rightarrow$ true θ should give high value for

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta) \qquad \text{(the likelihood function)}$$

N.B. $L(\theta) = f_{\text{sample}}(\vec{x};\theta)$, but $L(\theta)$ regarded as function of θ, $\vec{x}$ treated as constant (experiment is over).

N.B. In classical statistics, $L(\theta)$ is not a 'pdf' for θ.

$\rightarrow$ θ is not a random variable ($\hat{\theta}$ is).

In Bayesian statistics, treat $L(\theta) = L(\vec{x}|\theta)$ as pdf for $\vec{x}$ given θ, then use Bayes' theorem to get posterior pdf $p(\theta|\vec{x})$.

Maximum likelihood estimators

Define ML estimator $\hat{\theta}$ as the value of θ that maximizes $L(\theta)$.

Write estimators with hat ($\hat{\theta}$) to distinguish from true value θ, which may forever remain unknown.

For m parameters, usually find solution $\hat{\theta}_1, \ldots, \hat{\theta}_m$ by solving

$$\frac{\partial L}{\partial \theta_i} = 0 \quad i = 1, \ldots, m.$$

Sometimes $L(\theta)$ has more than one local maximum,

$\rightarrow$ take highest one.

N.B. no binning of data ('all information used').

N.B. the definition of ML estimators does not guarantee that they are in any way 'optimal'.

$\rightarrow$ investigate properties such as bias, variance.

For many cases of interest and for sufficiently large sample, ML turns out to be about as good as we can do.

Not always optimal for small n, but still usually best practical solution.

Example of ML estimator: parameter of exponential pdf

Consider the exponential pdf,

$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

and suppose we have a data sample $t_1, \ldots, t_n$.

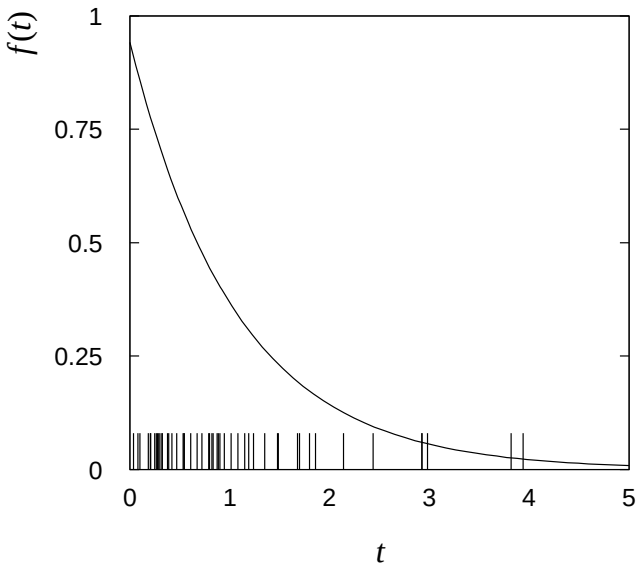
Usually use log-likelihood (maximum at same value of parameter),

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

Set $\dfrac{\partial \log L}{\partial \tau} = 0$ and solve for τ,

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

Example: generate 50 values of t with MC using $\tau = 1$,



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

Bias of ML estimator $\hat{\tau}$

Is $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$ an unbiased estimator for τ?

The hard way to check:

find pdf $g(\hat{\tau};\tau)$, compute $b = E[\hat{\tau}] - \tau$

Or use an easier way to compute $E[\hat{\tau}]$,

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

$$= \int\ldots\int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right) \frac{1}{\tau}e^{-t_1/\tau}\ldots\frac{1}{\tau}e^{-t_n/\tau} dt_1\ldots dt_n$$

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

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

$\rightarrow$ $\hat{\tau}$ is an unbiased estimator for τ.

The really easy way:

We already showed that the sample mean $\overline{t}$ is an unbiased estimator for $E[t]$, and for the exponential pdf, $E[t] = \tau$.

Functions of ML estimators

Suppose we had written the exponential pdf as

$$f(t;\lambda) = \lambda e^{-\lambda t}$$

where $\lambda = 1/\tau$ is the decay constant (inverse lifetime).

What is the ML estimator for λ?

For a function $a(\theta)$ of a parameter θ, it doesn't matter whether we express L as a function of a or θ.

The a that maximizes $L_a(a)$ is $a(\hat{\theta})$, where $\hat{\theta}$ maximizes $L_\theta(\theta)$.

$\rightarrow$ ML estimator of function $a(\theta)$ is $\hat{a} = a(\hat{\theta})$.

So for the decay constant, we have

$$\hat{\lambda} = \frac{1}{\hat{\tau}} = \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)^{-1} \quad .$$

Is $\hat{\lambda}$ an unbiased estimator for λ?

Generally a nonlinear function of an unbiased estimator is a biased estimator for the function of the parameter.

For $\hat{\lambda}$ we can show (cf. SDA Chapter 10)

$$E[\hat{\lambda}] = \lambda \frac{n}{n-1}$$

$\rightarrow$ $\hat{\lambda}$ has bias which goes to zero for $n \rightarrow \infty$.

Example of ML estimators: parameters of Gaussian pdf

Consider a sample from a Gaussian pdf of unknown μ, σ^2.

The log-likelihood function is

$$\log L(\mu, \sigma^2) = \sum_{i=1}^{n} \log f(x_i; \mu, \sigma^2)$$

$$= \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \log \frac{1}{\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right).$$

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

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \,,$$

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

We already know that $\hat{\mu}$ is an unbiased estimator for μ.

For $\widehat{\sigma^2}$ we find

$$E[\widehat{\sigma^2}] = \frac{n-1}{n} \sigma^2 \,,$$

so the ML estimator for σ^2 has a bias, but $b \to 0$ for $n \to \infty$.

Recall, however, that

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

is an unbiased estimator for the variance of any pdf.

Variance of estimator: analyitic method

Recall estimator for mean of exponential: $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$.

How wide is the pdf $g(\hat{\tau}; \tau, n)$?

$$V[\hat{\tau}] = E[\hat{\tau}^2] - (E[\hat{\tau}])^2$$

$$= \int \ldots \int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)^2 \frac{1}{\tau}e^{-t_1/\tau} \ldots \frac{1}{\tau}e^{-t_n/\tau} dt_1 \ldots dt_n$$

$$- \left(\int \ldots \int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right) \frac{1}{\tau}e^{-t_1/\tau} \ldots \frac{1}{\tau}e^{-t_n/\tau} dt_1 \ldots dt_n\right)^2$$

$$= \frac{\tau^2}{n}.$$

$\rightarrow$ variance of $\hat{\tau}$ is n times smaller than variance of t.

(In fact we knew this already, since here $\hat{\tau} = \overline{t}$.)

N.B. $V[\hat{\tau}]$, $\sigma_{\hat{\tau}}$ functions of true (unknown!) τ. Estimate using

$$\hat{\sigma}_{\hat{\tau}} = \frac{\hat{\tau}}{\sqrt{n}}$$

Often given as 'statistical error', e.g. $\hat{\tau} \pm \hat{\sigma}_{\hat{\tau}} = 1.062 \pm 0.150$

This means: ML estimate for τ is 1.062.

ML estimate for σ of $g(\hat{\tau}; \tau, n)$ is 0.150.

If $g(\hat{\tau}; \tau, n)$ is Gaussian, $[\hat{\tau} - \hat{\sigma}_{\hat{\tau}}, \hat{\tau} + \hat{\sigma}_{\hat{\tau}}]$ same as '68% confindence interval' (more on this later).

Variance of estimators: Monte Carlo method

Often form of $\hat{\theta}$, $g(\hat{\theta}; \theta, n)$ not known explicitly,

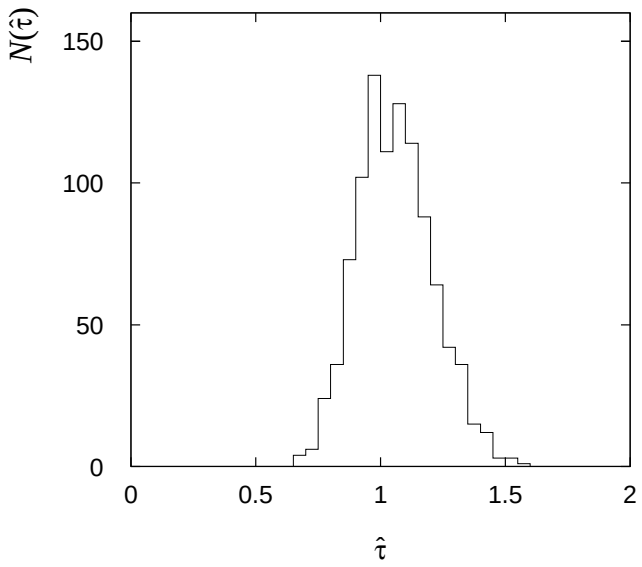
→ get $g(\hat{\theta}; \theta, n)$ from Monte Carlo.

For example with exponential pdf we had $\hat{\tau} = 1.062$.

Use this as 'true' τ in MC,

generate samples of $n = 50$ values (1000 experiments),

compute $\hat{\tau}$ for each experiment and histogram:



Sample standard deviation from MC experiments gives

$$\hat{\sigma}_{\hat{\tau}} = \left[\frac{1}{N_{\text{exp}} - 1} \sum_{i=1}^{N_{\text{exp}}} (\hat{\tau}_i - \bar{\hat{\tau}})^2 \right]^{1/2} = 0.151$$

Similar to previous estimate $\dfrac{\hat{\tau}}{\sqrt{n}} = 0.150$.

N.B. $g(\hat{\tau}; \tau, n)$ approximately Gaussian (cf. central limit theorem)

→ true in general for ML estimators in large sample limit.

The RCF bound (information inequality)

A lower bound on the variance of any estimator (not just ML) is

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

This is the Rao–Cramér–Frechet inequality (information inequality).

If equality is met, $\hat{\theta}$ is said to be efficient.

→ ML estimators are (almost always) efficient for large n,

often assume this to be true and use RCF bound to estimate $V[\hat{\theta}]$.

For the example with the exponential pdf, we obtain

$$\frac{\partial^2 \log L}{\partial \tau^2} = \frac{n}{\tau^2}\left(1 - \frac{2}{\tau}\frac{1}{n}\sum_{i=1}^{n} t_i\right) = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)$$

and we know that $b = 0$, so

$$V[\hat{\tau}] \geq \frac{1}{E\left[-\frac{n}{\tau^2}(1 - \frac{2\hat{\tau}}{\tau})\right]} = \frac{1}{-\frac{n}{\tau^2}\left(1 - \frac{2E[\hat{\tau}]}{\tau}\right)} = \frac{\tau^2}{n}$$

This is equal to the true variance → ML $\hat{\tau}$ is efficient for any n.

For $\vec{\theta} = (\theta_1, \ldots, \theta_m)$ with efficient estimator and zero bias,

$$(V^{-1})_{ij} = E\left[-\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}\right] = -n \int f(x;\vec{\theta})\, \frac{\partial^2 \log f(x;\vec{\theta})}{\partial \theta_i \partial \theta_j} dx$$

→ variance of efficient estimators $\propto \dfrac{1}{n}$.

The RCF bound (continued)

The expectation value of $\dfrac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}$ in the RCF bound is a function of the true parameters.

→ estimate by evaluating with the (single) ML estimate:

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

For a single parameter one has

$$\widehat{\sigma^2}_{\hat{\theta}} = \left(-1 \Big/ \left.\frac{\partial^2 \log L}{\partial\theta^2}\right)\right|_{\theta=\hat{\theta}} .$$

Often maximize $\log L$ numerically, estimate matrix of 2nd derivatives (Hessian matrix) using finite differences.

→ `MINUIT` routine `HESSE`.

Variance of ML estimators: graphical method

Consider single parameter θ, expand $\log L(\theta)$ about $\hat{\theta}$,

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

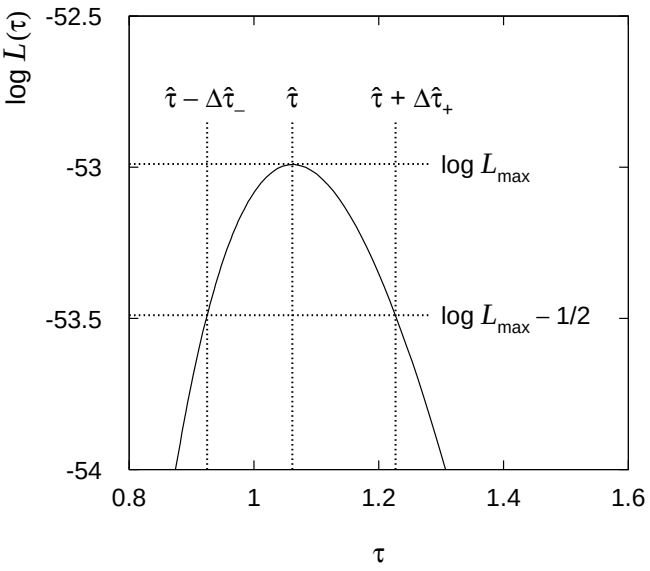
$\log L(\hat{\theta}) = \log L_{\max}$ and the second term is zero, therefore

$$\log L(\theta) = \log L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}},$$

that is,

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

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



Example of exponential:

$\hat{\tau} = 1.062$

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

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

Lecture 6 summary

The method of maximum likelihood

1. **The likelihood function, ML estimators:** $L(\theta)$ is the joint pdf for the data we got. ML estimator $\hat{\theta}$ at maximum of L.
2. **Examples: parameters of exponential and Gaussian pdfs:** for these, everything can be done analytically. $\widehat{\sigma^2}$ has bias ($b \to 0$ for $n \to \infty$).
3. **Variance of ML estimators:** estimate using different techniques
   (a) Analytic method: best when possible
   (b) Monte Carlo method: useful but can be time consuming
   (c) The RCF bound: only an inequality, but equality holds (approximately) for ML with sufficiently large sample.
   (d) Graphical method: change θ away from $\hat{\theta}$ until $\log L$ decreases by $1/2$. (We will generalize this to multiparameter case later.)