Lecture 4. Maximum Likelihood Estimation - confidence intervals.

Igor Rychlik

Chalmers

Department of Mathematical Sciences

Probability, Statistics and Risk, MVE300 • Chalmers • March 2014. Click on red text for extra material.

Maximum Likelihood method

It is *parametric* estimation procedure of F_X consisting of two steps: **choice of a model**; **finding the parameters**:

- Choose a model, i.e. select one of the standard distributions $F(x)$ (normal, exponential, Weibull, Poisson ...). Next postulate that

$$F_X(x) = F\left(\frac{x-b}{a}\right).$$

- Find estimates (a^*, b^*) such that $F_X(x) \approx F\big((x-b^*)/a^*\big)$. The **maximum likelihood** estimates (a^*, b^*) will be presented.

Finding likelihood, review from Lecture 1:

- Let $A_1, A_2, \ldots, A_k$ be a partition of the sample space, i.e. k excluding alternatives such that one of them is true. Suppose that it is equally probable that any of A_i is true, i.e. prior odds $q_i^0 = 1$.
- Let $B_1, \ldots, B_n$ be true statements (evidences) and let B be the event that all B_i are true, i.e. $B = B_1 \cap B_2 \cap \ldots \cap B_n$.
- The new odds q_i^n for A_i after collecting B_i evidences are

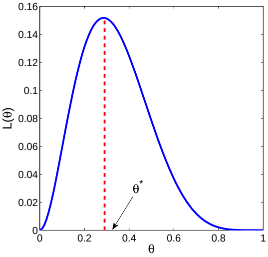
$$q_i^n = \mathrm{P}(B \mid A_i) \cdot q_i^0 = \mathrm{P}(B \mid A_i) \cdot 1 = P(B_1|A_i) \cdot \ldots \cdot P(B_n|A_i).$$

Function $L(A_i) = \mathrm{P}(B \mid A_i)$ is called likelihood that A_i is true.

The ML estimate - discrete case:

The maximum likelihood method recommends to choose the alternative A_i^* having highest likelihood, i.e. find i for which the likelihood $L(A_i)$ is highest.

Example 1 Binomial cdf.



ML estimate - continuous variable:

Model: Let consider a continuous rv. and postulate that $F_X(x)$ is exponential cdf, i.e. $F_X(x) = 1 - \exp(-x/a)$ and pdf

$$f_X(x) = \exp(-x/a)/a = f(x; a).$$

Data: $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are observations of X. (Example: the earthquake data where $n = 62$ obs.)

Likelihood function:[1] In practice data is given with finite number of digits, hence one only knows that events $B_i ='' x_i - \epsilon < X \leq x_i + \epsilon''$ is true. For small ϵ, $\mathsf{P}(B_i) \approx f_X(x_i) \cdot 2\epsilon$ thus

$$L(a) = \mathsf{P}(B_1|a) \cdot \ldots \cdot \mathsf{P}(B_n|a) = (2\epsilon)^n f(x_1; a) \cdot \ldots \cdot f(x_n; a).$$

ML-estimate: a^* maximizes $L(a)$ or **log-likelihood** $l(a) = \ln L(a)$.

Example 2 Exponential cdf.

[1] Since $\mathsf{P}(X = x_i) = 0$ for all values of parameter a it is not obvious how to define the likelihood function $L(a)$.

Sumarizing - Maximum Likelihood Method.

For n independent observations $x_1, \ldots, x_n$ the *likelihood function*

$$L(\theta) = \begin{cases} f(x_1;\theta) \cdot f(x_2;\theta) \cdot \ldots \cdot f(x_n;\theta) & \text{(continuous r.v.)} \\ p(x_1;\theta) \cdot p(x_2;\theta) \cdot \ldots \cdot p(x_n;\theta) & \text{(discrete r.v.)} \end{cases}$$

where $f(x;\theta)$, $p(x;\theta)$ is probability density and probability-mass function, respectively.

The value of θ which maximizes $L(\theta)$ is denoted by θ^* and called the ML estimate of θ.

Example 3 Censored data.

Example: Estimation Error $\mathcal{E}$

Suppose that position of moving equipment is measured periodically using GPS. Example of sequence of positions p^{GPS} is 1.16, 2.42, 3.55, ..., km. Calibration procedure of the GPS states that the **error**

$$\mathcal{E} = p^{true} - p^{GPS}$$

is approximately normal; is in average zero (no bias) and has standard deviation $\sigma = 50$ meters. What does it means in practice?

Quantiles of the standard normal distribution.

α	0.10	0.05	0.025	0.01	0.005	0.001
λ_α	1.28	1.64	1.96	2.33	2.58	3.09

Example 4

$e_\alpha = \sigma\lambda_\alpha$.

Confidence interval:

Clearly error $\mathcal{E} = p^{true} - p^{GPS}$ is with probability $1-\alpha$ in the interval:

$$\mathrm{P}(e_{1-\alpha/2} \leq \mathcal{E} \leq e_{\alpha/2}) = 1-\alpha.$$

For $\alpha = 0.05$, $e_{\alpha/2} \approx 1.96\,\sigma$, $e_{1-\alpha/2} \approx -1.96\,\sigma$, $\sigma = 50$ m, hence

$$\begin{aligned} 1-\alpha &\approx \mathrm{P}\big(p^{GPS} - 1.96\cdot 50 \leq p^{true} \leq p^{GPS} + 1.96\cdot 50\big) \\ &= \mathrm{P}\big(p^{true} \in [p^{GPS} - 1.96\cdot 50,\ \ p^{GPS} + 1.96\cdot 50]\big). \end{aligned}$$

If we measure many times positions using the same GPS and errors are independent then frequency of times statement

$$A = "\,p^{true} \in [p^{GPS} - 1.96\cdot 50,\ \ p^{GPS} + 1.96\cdot 50]"$$

is true will be close to 0.95.[2]

[2]Often, after observing an outcome of an experiment, one can tell whether a statement about outcome is true or not. Observe that this is not possible for A!

Asymptotic normality of error $\mathcal{E}$:

When unknown parameter θ, say, is estimated by mean of observations then by Central Limit Theorem the error $\mathcal{E} = \theta - \theta^*$ has mean zero and is asymptotically (as number of observations n tends to infinity) normally distributed.[3]

Distribution	ML estimates	$(\sigma^2_{\mathcal{E}})^*$
$X \in \mathrm{Po}(\theta)$	$\theta^* = \bar{x}$	$\dfrac{\theta^*}{n}$
$K \in \mathrm{Bin}(n, \theta)$	$\theta^* = \dfrac{k}{n}$	$\dfrac{\theta^*(1-\theta^*)}{n}$
$X \in \mathrm{Exp}(\theta)$	$\theta^* = \bar{x}$	$\dfrac{(\theta^*)^2}{n}$
$X \in \mathrm{N}(\theta, \sigma^2)$	$\theta^* = \bar{x}$	$\dfrac{s_n^2}{n}$

Example 5

[3] Similar result was valid for GPS estimates of positions.

Confidence interval for unknown parameter:

As for GPS measurements, probability that statement

$$A = "\theta \in [\theta^* - \lambda_{\alpha/2}\sigma^*_{\mathcal{E}},\ \theta^* + \lambda_{\alpha/2}\sigma^*_{\mathcal{E}}]",$$

is true is approximately $1 - \alpha$. Since we can not tell whether A is true or not the probability measures **lack of knowledge**. Hence one call the probability **confidence**[4].

Under some assumptions, the ML estimation error $\mathcal{E} = \theta - \theta^*$ is asymptotically normal distributed. With $\sigma^*_{\mathcal{E}} = 1/\sqrt{-\ddot{l}(\theta^*)}$

$$\theta \in [\theta^* - \lambda_{\alpha/2}\sigma^*_{\mathcal{E}},\ \theta^* + \lambda_{\alpha/2}\sigma^*_{\mathcal{E}}],$$

with approximately $1 - \alpha$ confidence.

[4]However if we use confidence intervals to measure uncertainty of estimated parameters values then in long run the statements A will be true with $1 - \alpha$ frequency

Exact confidence interval an example - Horse kicks data:

In 1898, von Bortkiewicz published a dissertation about a law of low numbers where he proposed to use the Poisson probability-mass function in studying accidents.

A part of his famous data is the number of soldiers killed by horse-kicks 1875-1894 in corps of the Prussian army. Here the data from corps II will be used:

0	0	0	2	0	2	0	0	1	1	0	0	2	1	1	0	0	2	0	0

As Bortkiewicz we assumed a Poisson distribution and found the ML estimate $m^* = \bar{\mathbf{x}} = 0.6$. The total number of victims is 12 (in 20 years, $n = 20$) which we consider sufficiently large to apply asymptotic normality.

Confidence interval - Horse kicks data:

For a Poisson variable, $(\sigma^2_{\mathcal{E}})^* = m^*/n$, hence $\sigma^*_{\mathcal{E}} = \sqrt{m^*/20} = 0.173$.
The **asymptotic confidence interval** having approximately confidence 0.95, for the true intensity of killed people due to horse kicks

$$\theta \in \left[\, 0.6 - 1.96 \cdot 0.173,\ 0.6 + 1.96 \cdot 0.173 \,\right] = [0.26,\ 0.94].$$

The **exact confidence interval** having confidence $1 - \alpha$ is

$$m \in \left[\frac{\chi^2_{1-\alpha/2}(2n\, m^*)}{2n}, \quad \frac{\chi^2_{\alpha/2}(2n\, m^* + 2)}{2n} \right].$$

For the Horse kicks data $m^* = 0.6$ and we get

$$\theta \in [0.32,\ 1.05]$$

since $\chi^2_{1-\alpha/2}(2n\theta^*) = \chi^2_{0.975}(24) = 12.40$, $\chi^2_{0.025}(26) = 41.92$.

The χ^2 test for continuous X

- Since the parameter θ is unknown we wish to test hypothesis

$$H_0: \ F_X(x) = F(x, \theta^*).$$

- In order to use χ^2 test the variability of X is described by discrete function $K = f(X)$.
- Definition of K: choose a partition $c_0 < c_1 < \ldots < c_{r-1} < c_r$ and let $K = k$ if $c_{k-1} < X \leq c_k$.
- Observed X, $(x_1, \ldots, x_n)$, are transformed into frequencies n_k, how many times K took value k, and $\mathrm{P}(K = k)$ is estimated by $p_k^* = n_k/n$. Finally p_k^* is compared with

$$p_k = \mathrm{P}(K = k) = \mathrm{P}(c_{k-1} < X \leq c_k) = F(c_k, \theta^*) - F(c_{k-1}, \theta^*).$$

- H_0 is rejected if $Q = \sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k} > \chi^2_\alpha(f)$. Here $f = r - m - 1$, where m is the number of parameters that have been estimated.[5]

---

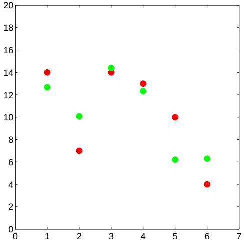
[5]As a rule of thumb one should check that $np_k > 5$ for all k.

Times between serious earthquakes - exponential cdf?

- Hypothesis $H_0 : F(x;\theta) = 1 - \exp(-x/\theta^*)$ with $\theta^* = 437.2$.
- Defining K: $c_0 = 0$, $c_1 = 100$, $c_2 = 200$, $c_3 = 400$, $c_4 = 700$, $c_5 = 1000$, and $c_6 = \infty$ and finding n_k "click".
- Probabilities $p_k = \mathrm{P}(K = k)$;

  $p_1 = 1 - \mathrm{e}^{-100/437.2} = 0.2045, \quad p_2 = \mathrm{e}^{-100/437.2} - \mathrm{e}^{-200/437.2} = 0.1627,$

  and $p_3 = 0.2323$, $p_4 = 0.1989$, $p_5 = 0.1001$ and $p_6 = 0.1015$.
- Computing Q statistics and testing:



Green dots np_i red dots n_i.
$Q = 0.1376 + 0.9449 + 0.0113 + 0.0362 + 2.3191 + 0.8355 = 4.285.$

Testing H_0: Now $f = 6 - 1 - 1$ and with $\alpha = 0.05$, $\chi^2_{0.05}(4) = 9.49$. Hence the exponential model can not be rejected.

In this lecture we met following concepts:

- Maximum Likelihood Method.
- CDF for estimation error.
- Confidence intervals, asymptotic based on ML methodology and example of exact conf. int..
- χ^2 test for continuous cdf.

Examples in this lecture "click"