Lecture 4. Maximum Likelihood Estimation - confidence intervals.

Igor Rychlik

Chalmers
Department of Mathematical Sciences

Probability, Statistics and Risk, MVE300 • Chalmers • March 2011. 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

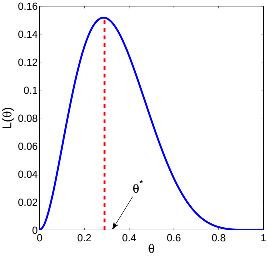
$$q_i^n = \mathrm{P}(B \mid A_i) \cdot q_i^0 = \mathrm{P}(B \mid A_i) \cdot 1 = P(B_1|A_i) \cdot \ldots \cdot P(B_n|A_i).$$

  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, $\mathrm{P}(B_i) \approx f_X(x_i) \cdot 2\epsilon$ thus

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

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

Example 2 Exponential cdf.

[1] Since $\mathrm{P}(X = x_i) = 0$ 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 $x_1, \dots, x_n$ 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:

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

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

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

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

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

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

Example - Earthquake data:

Recall - the ML-estimate is $a^* = 437.2$ days and, with the $\alpha = 0.05$,

$$e_{1-\alpha/2} = -1.96 \cdot \sqrt{3083} = -108.8, \quad e_{\alpha/2} = 1.96 \cdot \sqrt{3083} = 108.8.$$

and hence, with approximate confidence $1 - \alpha$,

$$a \in [437.25 - 108.8,\ 437.2 + 108.8] = [328,\ 546].$$

For exponential distribution with parameter a there is also **exact** interval: with confidence $1 - \alpha$

$$\theta \in \left[\frac{2na^*}{\chi^2_{\alpha/2}(2n)},\ \frac{2na^*}{\chi^2_{1-\alpha/2}(2n)} \right],$$

where $\chi^2_\alpha(f)$ is the α quantile of the $\chi^2(f)$ distribution. For the data $\alpha = 0.05$, $n = 62$, $\chi^2_{1-\alpha/2}(2n) = 95.07$, $\chi^2_{\alpha/2}(2n) = 156.71$ gives

$$a \in [346,\ 570].$$

Example - normal cdf:

Suppose we have independent observations $x_1, \dots, x_n$ from $\mathrm{N}(m, \sigma^2)$, σ *unknown*. Here one can construct an **exact** interval for m, viz. estimate σ^2 by

$$(\sigma^2)^* = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2 = s_{n-1}^2,$$

then the exact confidence interval for m is given by

$$\left[\bar{\mathbf{x}} - t_{\alpha/2}(n-1)\frac{s_{n-1}}{\sqrt{n}},\ \bar{\mathbf{x}} + t_{\alpha/2}(n-1)\frac{s_{n-1}}{\sqrt{n}}\right]$$

where $t_{\alpha/2}(f)$ are quantiles of the so-called **Student's t distribution** with $f = n-1$ degrees of freedom.

The asymptotic interval is

$$\left[\bar{\mathbf{x}} - \lambda_{\alpha/2}\frac{s_n}{\sqrt{n}},\ \bar{\mathbf{x}} + \lambda_{\alpha/2}\frac{s_n}{\sqrt{n}}\right].$$

Consider $\alpha = 0.05$. Then $\lambda_{\alpha/2} = 1.96$ and for $n = 10$, one has $t_{\alpha/2}(9) = 2.26$ while for $n = 25$, $t_{\alpha/2}(24) = 2.06$, which is closer to $\lambda_{\alpha/2} = 1.96$.

Quantiles of Student's t-distribution :

n	α						
	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

"The derivation of the t-distribution was first published in 1908 by William Sealy Gosset, while he worked at a Guinness Brewery in Dublin. He was prohibited from publishing under his own name, so the paper was written under the pseudonym Student."

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

If we have time: 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]

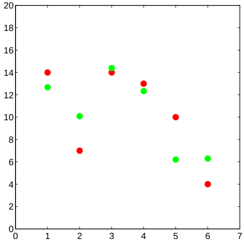
[5] As a rule of thumb one should check that $np_k > 5$ for all k.

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 examples of exact conf. int..
- Student's t distribution.
- χ^2 test for continuous cdf.

Examples in this lecture "click"