Laws of MLE - Confidence Intervals

1/ Standard errors

An estimator is calculated from a random simple and is a random variable: it has an expectation and a standard deviation. The standard deviation of an estimator is called its standard error.

Notation: if $\hat{\theta}$ is an estimator of θ, let $s_{\hat{\theta}}$ be its standard error.

Ex: $X_1, \ldots, X_n$ iid with mean μ and variance σ^2.

$\hat{\theta} = \overline{X}$. Then $s_{\hat{\theta}} = \frac{\sigma}{\sqrt{n}}$. When σ is unknown, $s_{\hat{\theta}}$ is estimated by $\hat{s}_{\hat{\theta}} = \frac{\hat{\sigma}}{\sqrt{n}}$.

We have a formula for the standard error of $\overline{X}$. What about standard errors for other estimators?

$\hookrightarrow$ there is a simple method for calculating the standard error of a maximum likelihood estimator (MLE)!

Definition: The Fisher Information of the sample $(X_1, \ldots, X_n)$ is

$$I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log L(\theta)\right]$$

Proposition: The standard error of $\hat{\theta}$ can be approximated (for large n) by

$$s_{\hat{\theta}} = \frac{1}{\sqrt{I(\theta)}}$$

and estimated by

$$\hat{s}_{\hat{\theta}} = \frac{1}{\sqrt{\hat{I}}}, \quad \text{where } \hat{I} = -\hat{H}(\hat{\theta}) = -\frac{\partial^2}{\partial \theta^2} \log L(\hat{\theta})$$

Remarks

- If $\theta \in \mathbb{R}^p$, $I(\theta)$ is a $p \times p$ square matrix and

$$s_{\hat{\theta}_j} = -E\left(\frac{\partial^2}{\partial \theta_j^2} \log L(\theta)\right).$$

- The proposition holds under certain regularity conditions of the model: the range of the X_is (support of the law) does not depend on θ, $\log L$ is 2 times differentiable and $0 < I(\theta) < +\infty$.

- The calculation of standard errors of MLE in learning models are programmed in statistical software.

Example

$X_1, \ldots, X_n$ i.i.d. Pareto(a, c)

CDF of the income distribution: $cdf(x) = 1 - \left(\frac{c}{x}\right)^a$, $x > c$.

c is the minimum income, a is called the tail index.

$$pdf(x) = \frac{a c^a}{x^{a+1}}, \quad x > c.$$

Suppose c is known, we estimate a.

$$\log L(a) = \sum_{i=1}^{n} \{\log(a) + a \log(c) - (a+1)\log(X_i)\}$$

$$\log L'(a) = \frac{n}{a} - \sum_{i=1}^{n} \log\left(\frac{X_i}{c}\right)$$

$$\log L''(a) = -\frac{n}{a^2}$$

Therefore the MLE of a is $\hat{a} = \dfrac{n}{\sum_i \log(X_i/c)}$ and

$s_{\hat{a}} = \dfrac{a}{\sqrt{n}}$, which can be estimated by $\hat{s}_{\hat{a}} = \dfrac{\hat{a}}{\sqrt{n}}$.

2/ Confidence intervals

A confidence interval $[\hat{m}, \hat{M}]$ is defined by

$$\mathbb{P}(\hat{m} \leq \theta \leq \hat{M}) = 1 - \alpha .$$

- $\hat{m}$ and $\hat{M}$ are functions of $X_1, \dots, X_n$ and random variables.
- $1 - \alpha$ = confidence level of the interval

  If a large number of independent 90% intervals are constructed, then 90% of them will contain θ.

  if $\alpha = 0.05$: 95% confidence interval.

a/ Confidence interval for the mean of a normal population.

$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

In the Gaussian model, we have: $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

$$S^2 = \frac{1}{n-1} \Sigma (X_i - \bar{X})^2 \quad , \quad \frac{(n-1) S^2}{\sigma^2} \sim \chi^2(n-1)$$

$\hookrightarrow$ $\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1)$

$\hookrightarrow$ estimate σ by S, unbiased estimator

$\hookrightarrow$ the so-called "t-statistic": $t = \dfrac{\bar{X} - \mu}{S / \sqrt{n}}$

The distribution of t is the Student distribution or t-distribution with $(n-1)$ degrees of freedom.

We denote by $t_{\alpha/2}$ and $t_{1-\alpha/2}$ the quantiles of this distribution.

Because of the symetry of the t_ distribution: $t_{1-\alpha/2} = -t_{\alpha/2}$.

We have $$\mathbb{P}\left(t_{\alpha/2} \leq t \leq t_{1-\alpha/2} \right) = 1-\alpha$$

After a bit of algebra, we obtain that

$$\mathbb{P}\left(\overline{X} - t_{1-\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{1-\alpha/2} \frac{S}{\sqrt{n}} \right) = 1-\alpha.$$

$\Leftrightarrow$ $\boxed{\overline{X} \pm t_{1-\alpha/2} \frac{S}{\sqrt{n}}}$ is a $(1-\alpha)$ confidence interval for μ.

Confidence intervals for the mean $\overline{X}$

b/ What if the sampling law is not normal?

The distribution of t is not the t_ distribution.

But the famous Central Limit Theorem (CLT) states that if the common variance σ^2 of $X_1, \ldots, X_n$ is finite, then the distribution of $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ can be approximated by the $\mathcal{N}(0,1)$ normal law, whatever the common distribution of the X_is, for large n. (say $n \geq 20$ if the population is symmetric, $n \geq 50$ for skewed populations)

The approximation holds true if we replace σ by a consistent estimator $\hat{\sigma}$. Thus,

$$\mathbb{P}\left(-q_{1-\alpha/2} \leq \frac{\overline{X} - \mu}{\hat{\sigma}/\sqrt{n}} \leq q_{1-\alpha/2} \right) \simeq 1-\alpha$$

where q_α is the α-quantile of the $\mathcal{N}(0,1)$.

$$\boxed{\overline{X} \pm q_{1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}}$$ is an approximate $(1-\alpha)$ confidence interval for $E(X_i)$.

c/ Confidence intervals for the variance of a normal population

A $(1-\alpha)$ confidence interval is given by:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \right] = CI(\sigma^2)$$

where $\chi^2_{\alpha/2,(n-1)}$ is the $\frac{\alpha}{2}$ quantile of the chi-square distribution with $(n-1)$ degrees of freedom.

To find a confidence interval for σ, take square roots of both endpoints: $CI(\sigma) = \left[\sqrt{\frac{(n-1)}{\chi^2_{1-\alpha/2,n-1}}}\, S ; \sqrt{\frac{n-1}{\chi^2_{\alpha/2,n-1}}}\, S \right]$.

(because $x \to \sqrt{x}$ is increasing).

Here, the assumption that the sample is normally distributed cannot be dispensed with, even if n is large.
Instead, we can compute the MLE of σ^2 and use the approximate distribution of MLEs (section d/) or use the bootstrap (next course).

d/ Confidence intervals for MLEs

Suppose $X_1, \dots, X_n$ is an iid sample whose distribution depends on a parameter θ.

In "regular" models, it can be shown that for large n, the MLE of θ is approximately normally distributed:

$$\hat{\theta}_{ML} \underset{n \text{ large}}{\sim} \mathcal{N}\left(\theta, I^{-1}(\theta)\right).$$

The approximation is still true if $I^{-1}(\theta)$ is estimated by $\hat{I}$

$\left(\hat{I} = -H(\hat{\theta})\right)$

Thus, if θ is a real parameter,

$\hat{\theta}_{ML} \pm q_{1-\alpha/2}\, \hat{s}_{\hat{\theta}_{ML}}$ is an approximate $(1-\alpha)$ confidence interval for θ, with $\hat{s}_{\hat{\theta}_{ML}} = \dfrac{1}{\sqrt{\hat{I}}}$.

Ex a 90% confidence interval of the Pareto tail index is

$$\hat{a} \pm q_{0.95} \frac{\hat{a}}{\sqrt{n}}.$$