Ricard Garcia Isern
2023
Statistical Methods & Stochastic Processes

Maximum Likelihood Estimation

1. INTRODUCTION

- Scientists use **models** to understand the phenomena they study.
- Any number computed with sample data → **statistics.**

- We can distinguish between **DETERMINISTIC MODELS & STATISTICAL MODELS.**

- **Probability:**
    - Known Population → **DEDUCTION** → Sample.

- **Statistics:**
    - Unknown Population ←**INDUCTION** ←Sample.

- **Parameters** → are fixed, unknown quantities that specify the population.
- **Estimators** → statistics that are used to estimate the unknown parameters.

2. MODEL ESTIMATION

- Statistical models have **unknown parameters** that need to be specified:
    - **Point Estimation** → estimating a population parameter with a **single value.**
        - MAXIMUM LIKELIHOOD ESTIMATOR.
        - METHOD OF MOMENTS.
    - **Interval Estimation** →estimating a population parameter with a **range of plausible values.**

3. MAXIMUM LIKELIHOOD ESTIMATOR

Example

- Let $X_1, \ldots, X_n$ be a random sample from a distribution $f(x|\theta_1, \ldots, \theta_k)$. k= number of parameters
- The likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$ is defined as

$$L(\boldsymbol{\theta}|\mathbf{x}) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \overbrace{\prod_{i=1}^{n}}^{\text{product}} f(x_i|\theta_1, \ldots, \theta_k)$$

*must maximize
↳ $\frac{df}{d\theta}$ find value
↳ $\frac{d^2l}{d\theta^2} < 0$ max

- This is in fact, $\boxed{\text{the joint density function}}$ considering the data as given.
- We will often work with the $\boxed{\text{log-likelihood function } \ell(\boldsymbol{\theta}|\mathbf{x}),}$ defined correspondingly as

↳ change $\Pi \to \Sigma$

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \ln(L(\boldsymbol{\theta}|\mathbf{x})) = \ln(L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n)) = \overbrace{\sum_{i=1}^{n}}^{\text{sum}} \ln(f(x_i|\theta_1, \ldots, \theta_k))$$

- EXAMPLE **BERNOULLI DISTRIBUTION**:

Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim Bern(p)$

↳ 1 parameter

$$P(X_1 = x_1 \mid p) = p^{x_1}(1-p)^{1-x_1}$$

$$P(X_1 = x_1, \ldots, X_n = x_n \mid p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

$$L(p \mid x_1, \ldots, x_n) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

tita

- The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta}|x)$ as a function of $\boldsymbol{\theta}$.
- The method selects a value for $\boldsymbol{\theta}$ such that the sample is most likely.
- Obtaining a maximum likelihood estimator is an optimization problem.
- In practice, it is often easier (and equivalent) to maximize the natural logarithm of the likelihood function, thus maximize $\ell(\boldsymbol{\theta}|x)$.

- To find candidates for MLE:
    1. Partial derivative of the parameter and equal to 0.
    2. Do the second derivative and check if it's negative →**maximum.**

find point

check if it's maximum

$$\frac{\partial}{\partial\theta}L(\theta|\mathbf{x}) = 0, \qquad i = 1, \ldots, k. \text{ and } \qquad \frac{\partial^2}{\partial\theta^2}L(\theta|\mathbf{x})|_{\theta=\hat{\theta}} < 0$$

- A point estimate obtained by ML is, by itself, not very informative.
- We need to specify its precision. with indicators
- The precision depends on the variance or the Fisher information of the ML estimator.

- **FISCHER INFORMATION:**

Let $X_1, \ldots, X_n$ be a random sample with

theoretic formula

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

The Fisher information about θ contained in $\mathbf{x}$ is defined by

derivative of log-likelihood depending on θ

$$\boxed{I_{\mathbf{x}}(\theta)} = E_\theta\left[\left(\frac{\partial}{\partial\theta}\ln\left(f(\mathbf{x} \mid \theta)\right)\right)^2\right]$$

expected value

FISHER INFORMATION

- **CRAMÉR-RAO LOWER BOUND:**

tells which is the lowest variance

expected value = real value

- For any unbiased estimator $\left(E\left(\hat{\theta}\right) = \theta\right)$, there exists a lower bound on its variance.
- This bound equals the reciprocal (inverse) of the Fisher information.

$$\boxed{V\left(\hat{\theta}\right) \geq \frac{1}{I_{\mathbf{x}}(\theta)}}$$

* chose the estimator with lowest variance → best estimator

Definition
- An unbiased estimator that attains the Cramér-Rao lower bound is called efficient.

- **CONFIDENCE-INTERVALS:**

- Having the variance and the distribution of the ML estimator, we can now say something about uncertainty.
- A confidence interval is an expression of the uncertainty (precisió) of the estimate.
- A classical result, with $X_i \sim N(\mu, \sigma^2)$, is

$$CI(\mu)_{\underbrace{1-\alpha}_{\text{confidence}}} = \overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{1}$$

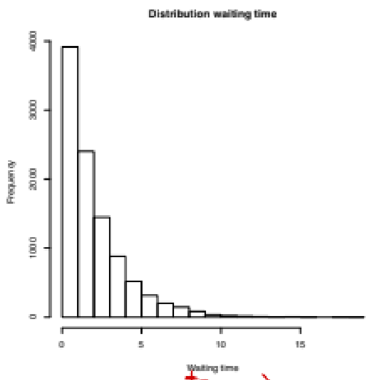
sample mean

where $\overline{X} = \hat{\mu}_{ML}$, and $\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{V\left(\hat{\mu}\right)}$.

- Term $\boxed{\frac{\sigma}{\sqrt{n}}}$ (σ estimated by s) is called the standard error of the mean.
- Term $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \approx 2\frac{\sigma}{\sqrt{n}}$ when $\alpha = 0.05$ is the error margin. $\begin{cases} \alpha = \text{error} \\ 1-\alpha = \text{confidence} \end{cases}$
- Equation (1) holds in general for ML estimators:

$$\boxed{CI(\theta)_{1-\alpha} = \hat{\theta} \pm z_{\alpha/2}\sqrt{V\left(\hat{\theta}\right)}} \rightarrow \frac{1}{I_{\mathbf{x}}(\theta)} \tag{2}$$

- R EXAMPLE OF MLE:



taxa = λ

- What is the rate of decay?
- What is the precision of a rate estimate?

CI, V, SD

```
> fitdistr(x,"exponential")
      rate
  0.498116487
 (0.004981165)
```

WHAT We must answer ??

Density and likelihood:

$$f(x|\lambda) = \lambda e^{-\lambda x} \quad L(\lambda|\mathbf{x}) = \lambda^n e^{-\lambda \sum x_i}$$

With some algebra, it follows that

$$\hat{\lambda} = 1/\bar{x}, \qquad I_n(\lambda) = n/\lambda^2 \qquad V(\hat{\lambda}) = \lambda^2/n$$

rate of decay

MLE

fisher info

variance

$$CI_{1-\alpha}(\lambda) = \hat{\lambda} \pm z_{\alpha/2}\sqrt{V\left(\hat{\lambda}\right)} = \hat{\lambda} \pm z_{\alpha/2}\frac{\hat{\lambda}}{\sqrt{n}}$$

Descriptive statistics of a sample of $n = 10.000$ waiting times

```
      N  N*  Mean   Stdev   Med    Q1     Q3     Min    Max
X 10000   0  2.0075     2  1.397  0.579  2.768  0.001  18.163
```

$$\hat{\lambda} = 1/2.0075 = 0.49812$$

$$CI_{0.95}(\lambda) = 0.49812 \pm 1.96\frac{0.49812}{\sqrt{10000}} = (0.4884; 0.5079)$$