Lecture 17 Maximum Likelihood Estimation (MLE)

BIO210 Biostatistics

Xi Chen

Spring, 2024

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

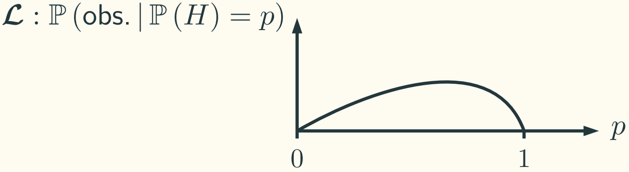
Intuition over MLE

Experiment: A coin, with an unknown $\mathbb{P}(H) = p$, was flipped 10 times. The outcome is $HHHTHHHTHH$.

Question: What is your best guess for p ?

Thinking: Given the data/observation we have, what values should p take such that our data/observation is most likely to occur ?

Aim: find the value that maximise our chance of observing the data, and use that value as our best guess/estimate for p.



Estimators of Parameters

- **Parameter space** Ω: the set of all possible values of a parameter θ or of a vector of parameters $(\theta_1, \theta_2, \theta_3, ..., \theta_k)$ is called the parameter space.

- Bernoulli: $\theta = p$, $\Omega = \{p \mid 0 \leqslant p \leqslant 1\}$
- Binomial: $\theta_1 = n, \theta_2 = p$, $\Omega = \{(n, p) \mid n = 2, 3, ...,$ a finite number$; 0 \leqslant p \leqslant 1\}$
- Poisson: $\theta = \lambda$, $\Omega = \{\lambda \mid \lambda \geqslant 0\}$
- Normal (Gaussian): $\theta_1 = \mu, \theta_2 = \sigma^2$, $\Omega = \{(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 \geqslant 0\}$

- We refer to an estimator of a parameter θ as $\hat{\boldsymbol{\theta}}$. An estimator $\hat{\boldsymbol{\theta}}$ of a parameter θ is unbiased if $\mathbb{E}\left[\hat{\boldsymbol{\theta}}\right] = \theta$. For example, $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ is an unbiased estimator for μ.

Maximum Likelihood Estimation (MLE)

- **Maximum likelihood estimation (MLE)** is a technique used for estimating the parameters of a given distribution, using some observed data.
- Introduced by R.A. Fisher in 1912.
- MLE can be used to estimate parameters using a limited sample of the population, by finding particular values so that the observation is the most likely result to have occurred.

Maximum Likelihood Estimation (MLE)

Formal definition

Let $x_1, x_2, x_3, ..., x_n$ be observations from n **i.i.d** random variables $(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, ..., \boldsymbol{X}_n)$ drawn from a probability distribution f_0, where f_0 is known to be from a family of distributions $\boldsymbol{f}$ that depend on some paramters θ. For example, f_0 could be known to be from the family of normal distributions $\boldsymbol{f}$, which depend on parameters μ and σ^2, and $x_1, x_2, x_3, ..., x_n$ would be observations from f_0. The goal of MLE is to maximise the likelihood function:

$$\begin{aligned}\boldsymbol{\mathcal{L}}(\theta; x_1, x_2, x_3, ..., x_n) &= f(x_1, x_2, x_3, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \\ &= f(x_1; \theta) \cdot f(x_2; \theta) \cdots f(x_n; \theta)\end{aligned}$$

The log-likelihood function:

$$\boldsymbol{\ell} = \ln \boldsymbol{\mathcal{L}} = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

Probability vs. Likelihood

$$\mathcal{L}(\theta; x_1, x_2, x_3, ..., x_n) = f(x_1, x_2, x_3, ..., x_n; \theta)$$

the likelihood of the parameter(s) θ taking certain values given that a bunch of data $x_1, x_2, ..., x_n$ are observed.

the joint probability mass/density of observing the data $x_1, x_2, ..., x_n$ with model parameter(s) θ.

from Wolfram:

Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a **probability** refers to the occurrence of future events, while a **likelihood** refers to past events with known outcomes.

Maximum Likelihood Estimation (MLE): Example 1

- Other notation: $\boldsymbol{\mathcal{L}}(\theta|x_1, x_2, x_3, ..., x_n) = f(x_1, x_2, x_3, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$

- **Example 1**: A coin, with an unknown $\mathbb{P}(H) = p$, was flipped 10 times. The outcome is $HHHTHHHTHH$. What is the MLE for p?
  - 1. Specify the parameter - $\theta : p$
  - 2. Specify the parameter space - $\boldsymbol{\Omega} : \{p \mid 0 \leqslant p \leqslant 1\}$
  - 3. Write out the probability function - $\mathbb{P}_{\boldsymbol{X}}(k) = \begin{cases} p & \text{, when } k = 1 \\ 1 - p & \text{, when } k = 0 \end{cases}$
  - 4. Write out the likelihood function:

$$\boldsymbol{\mathcal{L}}(p; 1110111011) = f(1110111011; p) = \prod_{i=1}^{10} f(x_i; p)$$

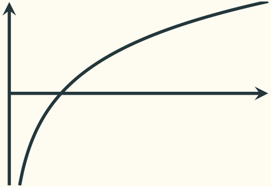
$$= f(1; p) \cdot f(1; p) \cdot f(1; p) \cdot f(0; p) \cdot f(1; p) \cdot f(1; p) \cdot f(1; p) \cdot f(0; p) \cdot f(1; p) \cdot f(1; p)$$

$$= p \cdot p \cdot p \cdot (1 - p) \cdot p \cdot p \cdot p \cdot (1 - p) \cdot p \cdot p = p^8(1 - p)^2$$

Maximum Likelihood Estimation (MLE): Example 2

- **Example 2 A more generalised case of coin flipping**: A (possibly unfair) coin is flipped m times, and k heads are observed. Let $\mathbb{P}(H) = p$. What is the MLE for p?

- 1. $\theta : (n, p)$
- 2. $\boldsymbol{\Omega} : \{(n, p) \mid n = m, 0 < p < 1\}$
- 3. $\mathbb{P}_{\boldsymbol{X}}(k) = \binom{n}{k} p^k (1-p)^{n-k}$
- 4. $\boldsymbol{\mathcal{L}}(n, p; k) = f(k; n, p) = \binom{m}{k} p^k (1-p)^{m-k}$



$$\boldsymbol{\ell} = \ln \boldsymbol{\mathcal{L}} = \ln \binom{m}{k} p^k (1-p)^{m-k}$$

$$= \ln \binom{m}{k} + k \ln p + (m-k) \ln (1-p)$$

What value should p take to maximise $\boldsymbol{\ell}$?

Let $\dfrac{\mathrm{d}\boldsymbol{\ell}}{\mathrm{d}p} = 0 \;\Rightarrow\; \hat{p} = \dfrac{k}{m}$

Maximum Likelihood Estimation (MLE): Example 3

- **Example 3 DNA synthesis errors**: The genetic material is copied and synthesised by DNA polymerase. One high-fidelity DNA polymerase, *Pfu*, originally isolated from the hyperthermophilic archae *Pyrococcus furiosus*, is believed to have very low error rate. Assume the errors generated by *Pfu* follow a Poisson distribution with λ mutations per 10^6 base pairs (Mb). We have examined n newly synthesised DNA fragments and observed that the nubmer of mutations per Mb is $k_1, k_2, k_3, ..., k_n$. What is the MLE for λ?

- 1. $\theta : \lambda$

- 2. $\mathbf{\Omega} : \{\lambda \mid \lambda > 0\}$

- 3. $\mathbb{P}_{\boldsymbol{X}}(k) = \dfrac{\lambda^k}{k!} e^{-\lambda}$

- 4. $\boldsymbol{\mathcal{L}}(\lambda; k_1, k_2, ..., k_n) = f(k_1, k_2, ..., k_n; \lambda) = \prod_{i=1}^{n} \dfrac{\lambda^{k_i}}{k_i!} e^{-\lambda}$

Advantages and Disadvantages of MLE

Advantages:

- Intuitive and straightforward to understand.
- If the model is correctly assumed, the MLE is efficient (meaning small variance or mean squared error).
- Can be extended to do other useful things.

Disadvantages:

- Relies on assumptions of a model (need to know the PMF/PDF).
- Sometimes difficult or impossible to solve the derivate of $\mathcal{L}$ or ℓ.
- Sometimes leads to the wrong or biased conclusions