Maximum likelihood

Estimate

- Whatever method we use to estimate the parameter of interest θ in a statistical model, the result depends only on our dataset.

- An **estimate** is a value to estimate θ that only on our dataset, i.e.,

$$t = h(x_1, x_2, \ldots, x_n) .$$

Estimator

- Let $t = h(x_1, x_2, \ldots, x_n)$ based on the dataset $x_1, x_2, \ldots, x_n$. Then t is a realization of the RV

$$T = h(X_1, X_2, \ldots, X_n)\,.$$

- The RV T is called an **estimator**. By definition it is also a **statistic**.

- The word **estimator** refers to the method for estimation, while **estimate** refers to the actual value computed from a dataset.

Agenda

- The maximum likelihood estimation (MLE)
- Properties of MLE
- MLE in Python

The maximum likelihood estimation

Likelihood function

Let $X_1, X_2, \ldots, X_n$ be a random sample. The likelihood function is given by

$$L(\boldsymbol{\theta} \,|\, X_1, \ldots, X_n) \triangleq f(X_1, \ldots, X_n; \boldsymbol{\theta})\,. \tag{1}$$

When $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x \,|\, \boldsymbol{\theta})$, equation (1) becomes

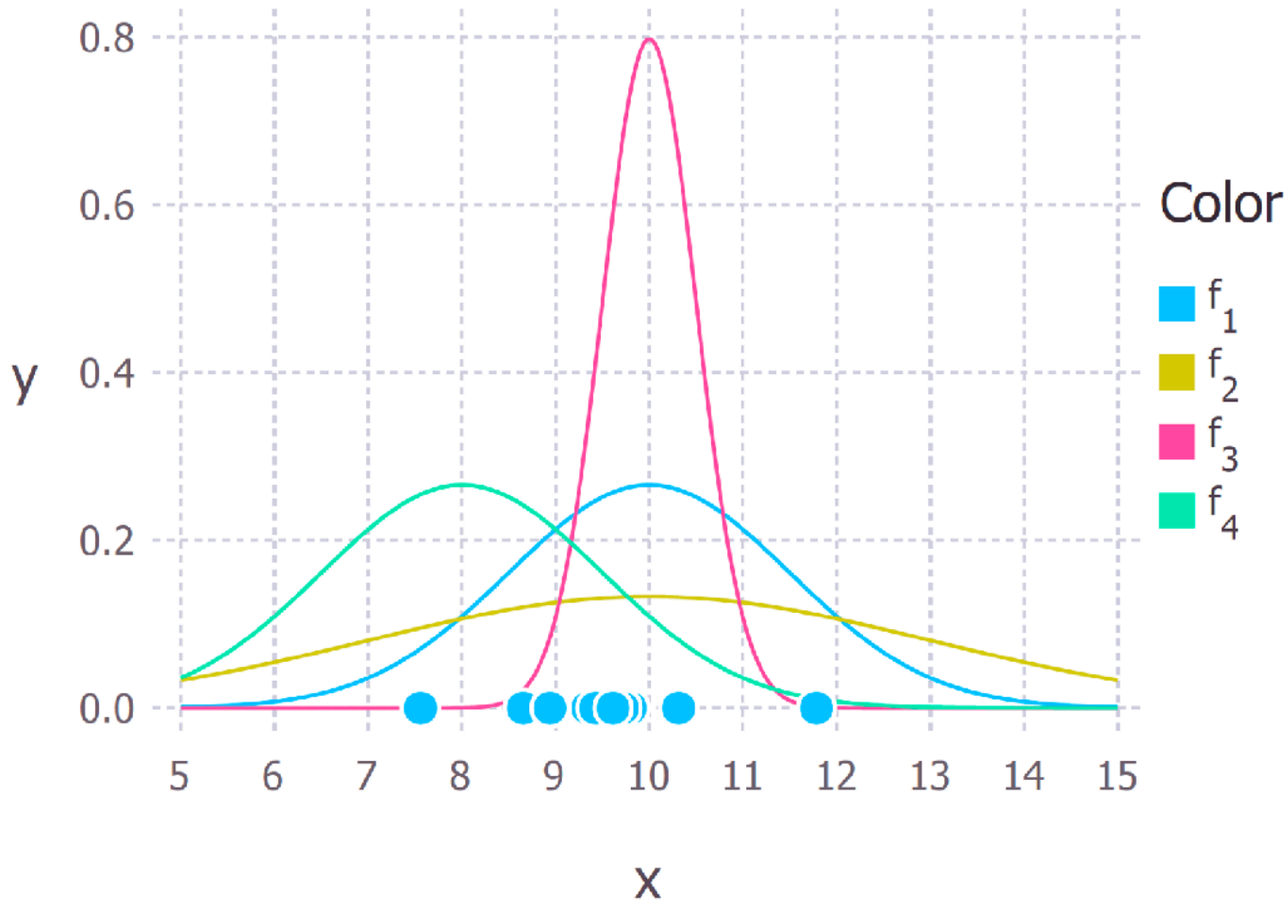
$$\boxed{L(\boldsymbol{\theta} \,|\, X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i; \boldsymbol{\theta})\,.} \tag{2}$$

easier to compute

- The likelihood function is a function of θ
- It is **not** a probability density function
- It measures the "support" (i.e. likelihood) provided by the data for each possible value of the parameter.

MLE

- Find the parameter that is “most likely” to observe your data.
- Maximizing the likelihood function is equivalent to maximizing the log–likelihood function (for computational issues) since logarithm is a monotone function.



https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

Log–likelihood function

Maximizing equation (2) directly is difficult, but maximizing the log–likelihood is much easier:

$$\ell(\boldsymbol{\theta} \mid X_1, \ldots, X_n) \triangleq \log L(\boldsymbol{\theta} \mid X_1, \ldots, X_n)$$

$$= \sum_{i=1}^{n} \log f(X_i; \boldsymbol{\theta}) \,. \qquad (3)$$

How to maximize a function?

Let $\ell(\boldsymbol{\theta})$ be a d–dimensional function on the closed domain Ω. Then, the max (or min) value of $\ell(\boldsymbol{\theta})$ will occur at:

- the boundary of Ω, or
- critical points satisfying $\nabla \ell = 0$ (first order condition)

Second partial derivative test

Let $\boldsymbol{\theta}_0$ be a critical point of $\ell(\boldsymbol{\theta})$ (i.e. $\nabla \ell(\boldsymbol{\theta_0}) = 0$).

- If $\nabla^2 \ell(\boldsymbol{\theta}_0)$ is positive definite, then ℓ attains a local minimum at $\boldsymbol{\theta}_0$.

- If $\nabla^2 \ell(\boldsymbol{\theta}_0)$ is negative definite, then ℓ attains a local maximum at $\boldsymbol{\theta}_0$.

- Otherwise, $\boldsymbol{\theta}_0$ is a saddle point for ℓ.

Example: 罷韓民調

- If $Y_i \overset{iid}{\sim}$ Bernoilli(p), then

$$P\left(Y_i = y_i\right) = \begin{cases} p & \text{if } y_i = 1 \text{ (同意罷韓)} \\ 1-p & \text{if } y_i = 0 \text{ (反對罷韓)} \end{cases}$$

$$= p^{y_i}(1-p)^{1-y_i}$$

- Likelihood function:

$$L(p \mid Y_1, \ldots, Y_n) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

- Log–likelihood:

$$\ell(p) = \sum_{i=1}^{n} \left[Y_i \log p + (1 - Y_i) \log(1 - p) \right]$$

$$= \log p \sum_{i=1}^{n} Y_i + \log(1 - p) \sum_{i=1}^{n} (1 - Y_i)$$

- First order condition:

$$\frac{d\ell}{dp} = \frac{1}{p} \sum_{i=1}^{n} Y_i - \frac{1}{1 - p} \sum_{i=1}^{n} (1 - Y_i) \triangleq 0$$

$$(1-p)\sum_{i=1}^{n} Y_i - p\sum_{i=1}^{n}(1-Y_i) = 0$$

$$\Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$ ← sample mean

- Second derivative test:

$$\frac{d^2\ell}{dp^2}(\hat{p}) = -\frac{1}{\hat{p}^2}\sum_{i=1}^{n} Y_i - \frac{1}{(1-\hat{p})^2}\sum_{i=1}^{n}(1-Y_i) < 0.$$

Example: simple linear regression

- Consider $Y_i | X_i = x_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$ for $i = 1, \ldots, n$ with $Y_i | X_i = x_i$ and $Y_j | X_j = x_j$ being independent for all $i \neq j$.

- Log–likelihood function:

$$\ell\left(\alpha, \beta, \sigma^2\right) = -n \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[y_i - \alpha - \beta x_i\right]^2 \quad (4)$$

- Note that since $\sigma^2 > 0$, maximizing (4) with respect to α and β is equivalent to minimizing

$$\sum_{i=1}^{n} \left[y_i - \alpha - \beta x_i \right]^2 . \tag{5}$$

- Minimizing equation (5) is also referred to least–squares estimation.

Properties of MLE

- Invariance principle: if $\hat{\theta}$ is an MLE of θ, then $g(\hat{\theta})$ is also an MLE of $g(\theta)$.

- Asymptotic unbiasedness:

$$\lim_{n\to\infty} E[\hat{\theta}_n] = \theta\,.$$

- Asymptotic minimum variance (by Cramer–Rao lower bound).

- 詳見數統I

MLE in Python

- scipy.stats
- R by rpy2
- numerical methods

```
from scipy.stats import norm
X = norm.rvs(0.5, 2, 100)
xbar, s = norm.fit(X)
print([xbar,s])
```

```
[0.4672385387662436, 1.9461666200490428]
```

Readings

- Chapters 19.1, 21, and 22 of our textbook.
- Chapter 8 of Rice (our first reference)

Homeworks

21.5, 21.6, 21.7, 21.9, 21.14, 22.3, 22.5, 22.8, 22.9, 22.12.