2 Maximum likelihood estimation

When it is not in our power to follow what is true, we ought to follow what is most probable. (Descartes)

2.1 Maximum likelihood estimation

Suppose that the random variable X has probability density function $f(x \mid \theta)$. Given the observed value x of X, the **likelihood** of θ is defined by

$$\mathrm{lik}(\theta) = f(x \mid \theta)\,.$$

Thus we are considering the density as a function of θ, for a fixed x. In the case of multiple observations, i.e., when $x = (x_1, \dots, x_n)$ is a vector of observed values of $X_1, \dots, X_n$, we assume, unless otherwise stated, that $X_1, \dots, X_n$ are IID; in this case $f(x_1, \dots, x_n \mid \theta)$ is the product of the marginals,

$$\mathrm{lik}(\theta) = f(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)\,.$$

It makes intuitive sense to estimate θ by whatever value gives greatest likelihood to the observed data. Thus the **maximum likelihood estimate** $\hat{\theta}(x)$ of θ is defined as the value of θ that maximizes the likelihood. Then $\hat{\theta}(X)$ is called the **maximum likelihood estimator** (**MLE**) of θ.

Of course, the maximum likelihood estimator need not exist, but in many examples it does. In practice, we usually find the MLE by maximizing $\log f(x \mid \theta)$, which is known as the **log-likelihood**.

Examples 2.1

(a) Smarties are sweets which come in k equally frequent colours. Suppose we do not know k. We sequentially examine 3 Smarties and they are red, green, red. The likelihood of this data, $x =$ *the second Smartie differs in colour from the first but the third Smartie matches the colour of the first*, is

$$\begin{aligned}\mathrm{lik}(k) = p(x \mid k) &= \mathbb{P}(\text{2nd differs from 1st})\mathbb{P}(\text{3rd matches 1st}) = \left(\frac{k-1}{k}\right)\frac{1}{k} \\ &= (k-1)/k^2\,,\end{aligned}$$

which equals 1/4, 2/9, 3/16 for $k = 2, 3, 4$, and continues to decrease for greater k. Hence the maximum likelihood estimate is $\hat{k} = 2$.

Suppose a fourth Smartie is drawn and it is orange. Now

$$\mathrm{lik}(k) = (k-1)(k-2)/k^3\,,$$

which equals 2/27, 3/32, 12/125, 5/54 for $k = 3, 4, 5, 6$, and decreases thereafter. Hence the likelihood estimate is $\hat{k} = 5$. Note that although we have seen only 3 colours the maximum likelihood estimate is that there are 2 colours we have not yet seen.

(b) $X \sim B(n, p)$, *n known, p to be estimated.*

Here

$$\log p(x \mid n, p) = \log \binom{n}{x} p^x (1-p)^{n-x} = \cdots + x \log p + (n - x)\log(1-p)\,.$$

This is maximized where

$$\frac{x}{\hat{p}} - \frac{n-x}{1-\hat{p}} = 0\,,$$

so the MLE of p is $\hat{p} = X/n$. Since $\mathbb{E}[X/n] = p$ the MLE is unbiased.

(c) $X \sim B(n, p)$, *p known, n to be estimated.*

Now we want to maximize

$$p(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

with respect to n, $n \in \{x, x+1, \dots\}$. To do this we look at the ratio

$$\frac{p(x \mid n+1, p)}{p(x \mid n, p)} = \frac{\binom{n+1}{x} p^x (1-p)^{n+1-x}}{\binom{n}{x} p^x (1-p)^{n-x}} = \frac{(1-p)(n+1)}{n+1-x}\,.$$

This is monotone decreasing in n. Thus $p(x \mid n, p)$ is maximized by the least n for which the above expression is ≤ 1, i.e., the least n such that

$$(1-p)(n+1) \leq n + 1 - x \iff n + 1 \geq x/p\,,$$

giving a MLE of $\hat{n} = [X/p]$. Note that if x/p happens to be an integer then both $n = x/p - 1$ and $n = x/p$ maximize $p(x \mid n, p)$. Thus the MLE need not be unique.

(d) $X_1, \dots, X_n \sim$ geometric(p), *p to be estimated.*

Because the X_i are IID their joint density is the product of the marginals, so

$$\log f(x_1, \dots, x_n \mid p) = \log \prod_{i=1}^{n} (1-p)^{x_i - 1} p = \left(\sum_{i=1}^{n} x_i - n\right)\log(1-p) + n \log p\,.$$

with a maximum where

$$-\frac{\sum_i x_i - n}{1-\hat{p}} + \frac{n}{\hat{p}} = 0\,.$$

So the MLE is $\hat{p} = \bar{X}^{-1}$. This MLE is **biased**. For example, in the case $n = 1$,

$$\mathbb{E}[1/X_1] = \sum_{x=1}^{\infty} \frac{1}{x}(1-p)^{x-1} p = -\frac{p}{1-p}\log p > p\,.$$

Note that $\mathbb{E}[1/X_1]$ does not equal $1/\mathbb{E}X_1$.

2.2 Sufficient statistics

The MLE, if it exists, is always a function of a **sufficient statistic**. The informal notion of a sufficient statistic $T = T(X_1, \dots, X_n)$ is that it summarises all information in $\{X_1, \dots, X_n\}$ which is relevant to inference about θ.

Formally, the statistic $T = T(X)$ is said to be **sufficient** for $\theta \in \Theta$ if, for each t, $P_\theta\big(X \in \cdot \mid T(X) = t\big)$ does not depend on θ. I.e., the conditional distribution of $X_1, \dots, X_n$ given $T(X) = t$ does not involve θ. Thus to know more about x than that $T(x) = t$ is of no additional help in making any inference about θ.

Theorem 2.2 *The statistic T is sufficient for θ if and only if $f(x \mid \theta)$ can be expressed as*

$$f(x \mid \theta) = g\big(T(x), \theta\big)h(x).$$

This is called the **factorization criterion**.

Proof. We give a proof for the case that the sample space is discrete. A continuous sample space needs measure theory. Suppose $f(x \mid \theta) = \mathbb{P}_\theta(X = x)$ has the factorization above and $T(x) = t$. Then

$$\mathbb{P}_\theta\big(X = x \mid T(X) = t\big) = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta\big(T(X) = t\big)} = \frac{g\big(T(x), \theta\big)h(x)}{\sum_{x:T(x)=t} g\big(T(x), \theta\big)h(x)}$$
$$= \frac{g(t, \theta)h(x)}{\sum_{x:T(x)=t} g(t, \theta)h(x)} = \frac{h(x)}{\sum_{x:T(x)=t} h(x)}$$

which does not depend on θ. Conversely, if T is sufficient and $T(x) = t$,

$$\mathbb{P}_\theta(X = x) = \mathbb{P}_\theta\big(T(X) = t\big)\mathbb{P}_\theta\big(X = x \mid T(X) = t\big)$$

where by sufficiency the second factor does not depend on θ. So we identify the first and second terms on the r.h.s. as $g(t, \theta)$ and $h(x)$ respectively. ■

Examples 2.3

(a) $X_1, \dots, X_n \sim P(\lambda)$, *$\lambda$ to be estimated.*

$$f(x \mid \lambda) = \prod_{i=1}^n \{\lambda^{x_i} e^{-\lambda}/x_i!\} = \lambda^{\sum_i x_i} e^{-n\lambda} \Big/ \prod_{i=1}^n x_i!\,.$$

So $g\big(T(x), \lambda\big) = \lambda^{\sum_i x_i} e^{-n\lambda}$ and $h(x) = 1/\prod_i x_i!$. A sufficient statistic is $t = \sum_i x_i$.

Note that the sufficient statistic is not unique. If $T(X)$ is a sufficient statistic, then so are statistics like $T(X)/n$ and $\log T(X)$.

The MLE is found by maximizing $f(x \mid \lambda)$, and so

$$\frac{d}{d\lambda} \log f(x \mid \lambda)\bigg|_{\lambda=\hat\lambda} = \frac{\sum_i x_i}{\hat\lambda} - n = 0\,.$$

Hence $\hat\lambda = \bar X$. It is easy to check that $\hat\lambda$ is unbiased.

Note that the MLE is always a function of the sufficient statistic. This is because the MLE is the value of θ which maximizes $f(x \mid \theta)$, and $f(x \mid \theta) = g\big(T(x), \theta\big)h(x)$. Thus the MLE is the θ which maximizes $g\big(T(x), \theta\big)$, and hence a function of $T(x)$.

(b) $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, *$\theta = (\mu, \sigma^2)$ to be estimated.*

$$f(x \mid \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_i (x_i-\mu)^2/2\sigma^2}$$
$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\left[\sum_i (x_i-\bar x)^2 + n(\bar x-\mu)^2\right]/2\sigma^2}$$

Thus, with $g\big(T(x), \theta\big)$ taken as the whole r.h.s. and $h(x) = 1$, the sufficient statistic for $\theta = (\mu, \sigma^2)$ is $T(x) = \big(\bar x, \sum_i (x_i - \bar x)^2\big)$.

Note that sometimes the sufficient statistic is not just a single number, but as here, a vector $T(X) = \big(T_1(X), \dots, T_r(X)\big)$. This usually occurs when the parameter is a vector, as $\theta = (\mu, \sigma^2)$.

In this example, if σ^2 had been known, then $\bar x$ would have been sufficient for μ. If μ had been known, then $\sum_i (x_i - \mu)^2$ would have been sufficient for σ^2.

(c) $X_1, \dots, X_k \sim U[0, \theta]$, *$\theta$ to be estimated.*

$$f(x \mid \theta) = \prod_{i=1}^n 1\{0 \le x_i \le \theta\}\frac{1}{\theta} = 1\{0 \le \max_i x_i \le \theta\}\frac{1}{\theta^n}\,,$$

where $1\{\text{condition}\} = 1$ or 0 as 'condition' is true or false. Thus $g\big(T(x), \theta\big) = 1\{0 \le \max_i x_i \le \theta\}/\theta^n$, $h(x) = 1$ and $T(x) = \max_i x_i$ is sufficient for θ. The MLE is $\hat\theta = \max_i X_i$.

To find $\mathbb{E}\hat\theta$ we must find the distribution function of $\max_i x_i$. This is

$$F(t) = \mathbb{P}(\max_i x_i \le t) = \mathbb{P}(x_1 \le t) \cdots \mathbb{P}(x_n \le t) = (t/\theta)^n\,.$$

By differentiation, $f(t) = nt^{n-1}/\theta^n$, and hence

$$\mathbb{E} \max_i x_i = \int_0^\theta t\, \frac{nt^{n-1}}{\theta^n}\, dt = \frac{n}{n+1}\theta\,.$$

So $\hat\theta$ is biased.

However, $\mathbb{E}\hat\theta \to \theta$ as $n \to \infty$. We say that $\hat\theta$ is **asymptotically unbiased**. Under some mild assumptions, MLEs are always asymptotically unbiased. This is one reason why we like the maximum likelihood estimation procedure.