Chapter 24

Maximum Likelihood Estimation

Lecture 24 | Devore 9e: 6.2

The method of moments matched summary statistics. Maximum likelihood asks a different and more direct question:

Which parameter value makes the data I actually saw as likely as possible?

24.1 A discrete warm-up

Example 24.1 (Candy jar)**.** A jar holds 10 candies, k of them red, with k unknown. You draw 4 candies without replacement and observe 3 red. Estimate k.

A first instinct is to match proportions: $\frac{k}{10} = \frac{3}{4}$, so $k = 7.5$ — which is not an integer, and not obviously right. Instead, write down the probability of what happened, as a function of k:

$$\mathbb{P}(3 \text{ red out of } 4) = \frac{\binom{k}{3}\binom{10-k}{1}}{\binom{10}{4}}, \qquad 3 \le k \le 9.$$

(The range is forced: we saw 3 reds so $k \ge 3$, and we saw a non-red so $k \le 9$.) The denominator does not involve k, and $\binom{k}{3}\binom{10-k}{1} = \frac{k(k-1)(k-2)}{6}(10-k)$, so we maximize $k(k-1)(k-2)(10-k)$ over the integers:

k	3	4	5	6	7	$\boxed{8}$	9
$k(k-1)(k-2)(10-k)$	42	144	300	480	630	$\boxed{672}$	504

So $\hat{k}_{\text{MLE}} = 8$.

Remark 24.2. The estimate is the value of the parameter under which the observed data is least surprising. Note that this required no derivative — for a discrete parameter you simply tabulate.

24.2 The continuous case, and the likelihood function

Example 24.3 (Light bulbs)**.** Lifetimes follow Exponential(λ). You observe 5 bulbs:

$$x_1 = 2, \quad x_2 = 1.5, \quad x_3 = 3, \quad x_4 = 0.5, \quad x_5 = 4.$$

Estimate λ.

We cannot copy the discrete recipe, because $\mathbb{P}\{X_1 = 2,\ X_2 = 1.5,\ \dots\} = 0$ for a continuous variable. Use the joint *density* instead: it plays exactly the role that probability played above.

Definition 24.4 (Likelihood function)**.** Let $X_1, \dots, X_n$ be a sample from a distribution with parameters $\theta_1, \dots, \theta_m$, and let $x_1, \dots, x_n$ be the observed data. The **likelihood function** is the joint pdf (or pmf) viewed as a function of the parameters:

$$L(\theta_1, \dots, \theta_m) := f(x_1, x_2, \dots, x_n; \theta_1, \dots, \theta_m).$$

The **maximum likelihood estimates** $\hat{\theta}_1, \dots, \hat{\theta}_m$ are the values maximizing L.

Remark 24.5. Read the roles carefully. The joint pdf is normally a function of the data with the parameters fixed. The likelihood is the *same formula* regarded as a function of the parameters with the data fixed. Same expression, opposite point of view.

24.2.1 The log trick

For an i.i.d. sample the joint density is a product, $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$, and products are unpleasant to differentiate. Since log is strictly increasing, maximizing L is the same as maximizing

$$\ell(\theta) := \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta),$$

which turns the product into a sum. Always take logs first.

24.2.2 Exponential, worked

$$L(\lambda) = \prod_{i=1}^{5} \lambda e^{-\lambda x_i}, \qquad \log L(\lambda) = 5 \log \lambda - \lambda \sum_{i=1}^{5} x_i.$$

Differentiate and set to zero:

$$\frac{d}{d\lambda} \log L(\lambda) = \frac{5}{\lambda} - \sum_{i=1}^{5} x_i \overset{\text{set}}{=} 0 \qquad \Longrightarrow \qquad \hat{\lambda}_{\text{MLE}} = \frac{5}{\sum_i x_i}.$$

With $\sum x_i = 11$, $\hat{\lambda} = \frac{5}{11} = \frac{1}{2.2}$. In general

$$\boxed{\hat{\lambda}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}}$$

which is reassuring, since $\mathbb{E}[X] = 1/\lambda$ for the exponential.

24.3 Normal

Example 24.6 (Unknown mean)**.** X_i i.i.d. Normal$(\mu, 1)$ with data $8, 9, 10, 11, 12$.

$$L(\mu) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu)^2/2}, \qquad \log L(\mu) = -\frac{5}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{5} (x_i - \mu)^2.$$

Then

$$\frac{d}{d\mu}\log L(\mu) = \sum_{i=1}^{5}(x_i - \mu) \overset{\text{set}}{=} 0 \qquad \Longrightarrow \qquad \hat{\mu}_{\text{MLE}} = \frac{\sum_i x_i}{5} = \bar{x} = 10.$$

Remark 24.7. Notice that maximizing the likelihood here is the same as *minimizing* $\sum_i (x_i - \mu)^2$. Least squares is maximum likelihood under normal errors — that identity is the foundation of regression.

Example 24.8 (Unknown variance)**.** X_i i.i.d. Normal$(10, \sigma^2)$ with the same data. Writing $v = \sigma^2$,

$$\log L(v) = -\frac{5}{2}\log(2\pi v) - \frac{1}{2v}\sum_{i=1}^{5}(x_i - 10)^2,$$

$$\frac{d}{dv}\log L(v) = -\frac{5}{2}\cdot\frac{1}{v} + \frac{\sum_i (x_i - 10)^2}{2v^2} \overset{\text{set}}{=} 0 \qquad \Longrightarrow \qquad \hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{i=1}^{5}(x_i - 10)^2}{5} = \frac{10}{5} = 2.$$

Remark 24.9 (MLE can be biased)**.** In general, when μ is also estimated from the data, the MLE of the variance is

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2,$$

the *biased* sample variance from Lecture 22. So maximum likelihood does not automatically produce unbiased estimators. It has other virtues — consistency, asymptotic efficiency, and invariance — which is why it is the default method anyway.

24.4 When calculus does not apply

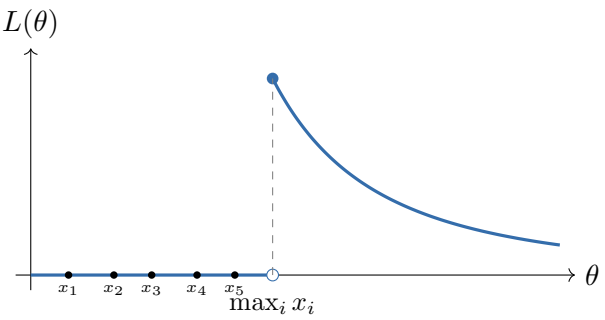
Example 24.10 (Uniform $[0, \theta]$, one more time)**.** $X_1, \ldots, X_n$ i.i.d. Uniform$[0, \theta]$. The likelihood is

$$L(\theta) = f(x_1)\cdots f(x_n) = \begin{cases} \dfrac{1}{\theta^n}, & 0 \le x_i \le \theta \text{ for all } i, \\ 0, & \text{otherwise.} \end{cases}$$

The condition "$x_i \le \theta$ for all i" is just $\theta \ge \max_i x_i$. So

$$L(\theta) = \begin{cases} 0, & \theta < \max_i x_i, \\ \theta^{-n}, & \theta \ge \max_i x_i, \end{cases}$$

which is 0, jumps up at $\theta = \max_i x_i$, and then decreases.



There is no interior critical point — the derivative is never zero. The maximum sits at the left edge of the allowed region:

$$\boxed{\hat{\theta}_{\text{MLE}} = \max(X_1, \ldots, X_n)}$$

Remark 24.11. So maximum likelihood recovers the biased estimator from Lecture 22, not the bias-corrected $\frac{n+1}{n}$ max. Two lessons:

1. Always check whether the support of the distribution depends on the parameter. If it does, do not reach for the derivative — sketch L instead.
2. MLE and unbiasedness are different criteria and can disagree.

24.5 Recipe

1. Write the likelihood $L(\theta) = \prod_i f(x_i; \theta)$, being careful about the range of validity.
2. Take logs: $\ell(\theta) = \sum_i \log f(x_i; \theta)$.
3. Differentiate, set to zero, solve.
4. Check it is a maximum (second derivative, or the shape of ℓ).
5. If the support depends on θ, skip steps 2–4 and argue directly.