STA 130A: Mathematical Statistics: Brief Course

Lecture 26: Maximum Likelihood Estimation

Dogyoon Song

Fall 2026, UC Davis

Agenda

Last time: Point estimation

- Method of moments
- Estimator properties: bias, MSE, and consistency

Today: Maximum likelihood estimation

- Likelihood and log likelihood
- MLE calculation: maximization and boundary cases
- Examples: Bernoulli, exponential, and uniform models
- MoM and MLE: comparison and assessment

Motivation: Which defect rate best explains the data?

Example (Inspecting manufactured items)

A manufacturer wants to estimate its defect probability q to forecast defects in future batches. An inspection finds **3 defective items out of 20**.

Under a candidate value of q, the binomial model gives

$$P_q(K = 3) = \binom{20}{3} q^3 (1-q)^{17}, \qquad K = \text{number of defective items.}$$

Candidate defect rate q	0.05	0.15	0.20	0.25	0.30
$P_q(K = 3)$, rounded	0.0596	0.2428	0.2054	0.1339	0.0716

Question: Among all $0 \leq q \leq 1$, which value makes the observed count most probable?

Idea: Keep the data fixed and use the probability model to compare parameter values.

Probability and likelihood

The same PMF can be viewed in two ways:

$$p(k; q) = \binom{20}{k} q^k (1-q)^{20-k}.$$

Viewpoint	What is fixed?	What varies?
Probability	Parameter value $q = 0.15$	Possible outcomes $k = 0, 1, \ldots, 20$
Likelihood	The observed outcome $k = 3$	Candidate parameters $0 \leq q \leq 1$

Two criteria for point estimation

- **MoM:** Match selected model moments to sample moments.
- **MLE:** Maximize $L(q) = p(3; q)$, the likelihood of the observed count.

Why learn these methods? They provide general construction rules that, under suitable conditions, yield **consistent estimators**.

Likelihood for an independent sample

Definition (Likelihood)

Suppose $X_1, \ldots, X_n$ are i.i.d. with PMF or PDF f_θ. For observed values $x_1, \ldots, x_n$, the **likelihood** is

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i), \qquad \theta \in \Theta,$$

where Θ is the set of allowed parameter values.

Keep the support: A parameter value that excludes an observation has likelihood zero.

Example (An unknown upper limit)

For a single observation $X \sim \mathrm{Uniform}[0, \theta]$, suppose $x = 3$. The likelihood is

$$L(\theta) = f_\theta(3) = \begin{cases} 0, & 0 < \theta < 3, \\ \frac{1}{\theta}, & \theta \geq 3. \end{cases}$$

Values below 3 exclude the observation. For $\theta \geq 3$, a larger upper limit spreads the density over a wider interval, giving a smaller likelihood.

Maximum likelihood estimation

Definition (Maximum likelihood estimate)

A **maximum likelihood estimate** (MLE) is a parameter value that maximizes the likelihood of the observed data:

$$\widehat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} L(\theta).$$

Here arg max returns the *maximizing parameter value*.

Before and after observing the sample

- Applied to the random sample $X_1, \ldots, X_n$, the rule gives an **estimator**: a random variable.
- Applied to observed values $x_1, \ldots, x_n$, it gives an **estimate**: a computed parameter value.

Interpretation: Choose the parameter whose model assigns the largest probability or density to the observed sample.

Example: Finding an MLE

Example (Two inspected items)

Let X_1, X_2 be independent $\mathrm{Bernoulli}(q)$ variables, where 1 means defective. We observe

$$x_1 = 1, \qquad x_2 = 0.$$

1. Compute the likelihood. By independence,

$$L(q) = P_q(X_1 = 1)\, P_q(X_2 = 0) = q(1-q), \qquad 0 \le q \le 1.$$

2. Find its maximizer. Rewrite

$$L(q) = \frac{1}{4} - \left(q - \frac{1}{2}\right)^2 \le \frac{1}{4}.$$

The squared term equals zero at $q = 1/2$, so

$$\widehat{q}_{\mathrm{MLE}} = \frac{1}{2}, \qquad L(\widehat{q}_{\mathrm{MLE}}) = \frac{1}{4}.$$

Interpretation: The estimated defect rate is 50%, matching one defective item out of two.

A recipe for finding an MLE

Goal: Find the parameter value that maximizes the likelihood.

1. **Model and constraints:** State the PMF/PDF, parameter range, and support restrictions.
2. **Likelihood:** Multiply the individual PMFs or PDFs.
3. **Maximize:** Find candidates by differentiation or monotonicity, then check boundaries and the maximum.
4. **Interpret:** Give the estimator and the numerical estimate, with units when relevant.

Simplification: Drop only positive factors in L that do not depend on the parameter.

Example: Bernoulli likelihood

Example (Defective items)

Let $X_i = 1$ for a defective item and $X_i = 0$ otherwise, with

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(q), \qquad 0 \le q \le 1.$$

For the observed data, write $k = \sum_{i=1}^{n} x_i$.

$$\begin{aligned} L(q) &= \prod_{i=1}^{n} q^{x_i}(1-q)^{1-x_i} \\ &= q^k(1-q)^{n-k}. \end{aligned}$$

For our inspection data, $n = 20$ and $k = 3$, so

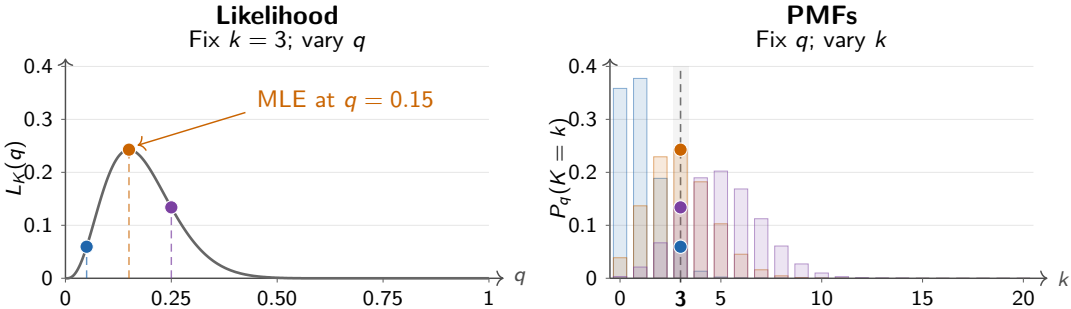
$$L(q) = q^3(1-q)^{17}.$$

Using the count: Its likelihood is $L_K(q) = \binom{n}{k} L(q)$. The factor $\binom{n}{k}$ does not depend on q, so both likelihoods have the same maximizer.

Visualization: Likelihood and binomial PMFs

For $K \sim \mathrm{Binomial}(20, q)$, the observed count is $k = 3$.

$$L_K(q) = P_q(K = 3) = \binom{20}{3} q^3 (1-q)^{17}.$$



Matching heights at $k = 3$: • $q = 0.05$: 0.0596 • $q = 0.15$: 0.2428 • $q = 0.25$: 0.1339

Connection: At $k = 3$, each PMF supplies one point on the likelihood curve.

Caution: The likelihood curve is not a probability density for q.

Likelihood maximization using the log likelihood

Our next task: Locate the peak of $L(q) = q^3(1-q)^{17}$ by calculation.

For parameter values with $L(\theta) > 0$, define the **log likelihood**

$$\ell(\theta) = \log L(\theta) = \log\left(\prod_{i=1}^{n} f_\theta(x_i)\right) = \sum_{i=1}^{n} \log f_\theta(x_i).$$

Why it helps

- The logarithm is strictly increasing, so it preserves the maximizing parameter.
- Products become sums, making differentiation easier.

Example (The same 3 defects out of 20)

For $0 < q < 1$,

$$\ell(q) = \log\left(q^3(1-q)^{17}\right) = 3\log q + 17\log(1-q).$$

Simplification: Additive terms independent of the parameter can be dropped.

Example: Maximizing the Bernoulli likelihood

Example (When $0 < k < n$)

For $0 < q < 1$,

$$\ell(q) = k \log q + (n-k) \log(1-q).$$

1. Find a candidate. Set the derivative to zero:

$$\ell'(q) = \frac{k}{q} - \frac{n-k}{1-q} = 0 \quad \Longrightarrow \quad k(1-q) = (n-k)q$$

$$\Longrightarrow \quad k = nq \quad \Longrightarrow \quad \widehat{q} = \frac{k}{n}.$$

2. Check the maximum. The log likelihood curves downward:

$$\ell''(q) = -\frac{k}{q^2} - \frac{n-k}{(1-q)^2} < 0.$$

The likelihood is zero at both endpoints when $0 < k < n$.

Our data: $\widehat{q} = 3/20 = 0.15$, an estimated defect rate of 15%.

Example: Bernoulli estimates at the boundary

Example (All items have the same outcome)

The parameter range includes both $q = 0$ and $q = 1$.

Observed sample	Likelihood	Maximizer
No defective items: $k = 0$	$(1-q)^n$	$\widehat{q} = 0$
All defective items: $k = n$	q^n	$\widehat{q} = 1$

Thus the same estimator covers every possible sample:

$$\widehat{q}_{\mathrm{MLE}} = \bar{X}_n.$$

Lesson: A maximum can occur at a boundary, with no interior solution to $\ell'(q) = 0$.

Pop-up quiz: Interpreting the MLE

We observe **3 defective items out of 20** and obtain $\widehat{q} = 0.15$.

Question: Which statement correctly interprets this estimate?

A. Among candidate values of q, $q = 0.15$ gives the largest probability of observing 3 defects.

B. There is a 15% probability that the true defect rate equals 0.15.

C. Under the fitted model, the probability of exactly 3 defects is 0.15.

D. A future batch of 20 items will contain exactly 3 defects.

Answer: A. We hold the observed count $K = 3$ fixed and compare candidate values of q.

$$\widehat{q} = 0.15, \qquad L_K(\widehat{q}) = P_{0.15}(K = 3) \approx 0.2428.$$

The estimated defect rate and the probability of this count are different quantities.

Example: Likelihood for exponential waiting times

Example (Estimating an arrival rate)

For independent exponential waiting times with rate $\lambda > 0$,

$$f_\lambda(x) = \lambda e^{-\lambda x}, \qquad x \geq 0.$$

We observe $1, 2, 3, 4$ minutes, with $\sum_i x_i = 10$ and $\bar{x} = 2.5$.

Likelihood: Multiply the density values for the fixed observations:

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i \right). \end{aligned}$$

For these four observations,

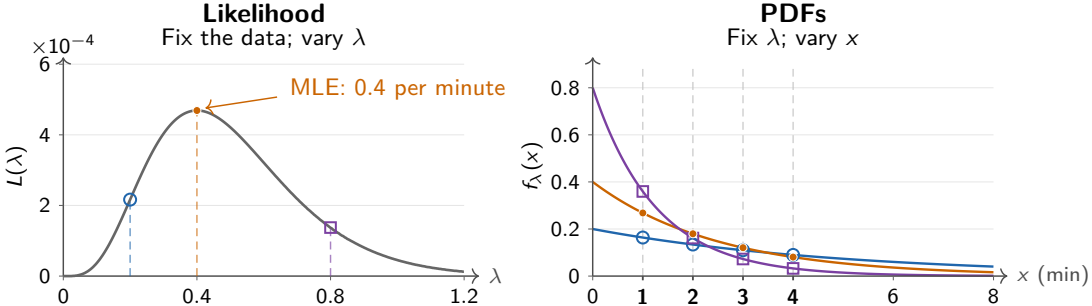
$$L(\lambda) = \lambda^4 e^{-10\lambda}, \qquad \lambda > 0.$$

Continuous data: Exact observations have probability zero; likelihood uses **density values**.

Visualization: Exponential likelihood and PDFs

The observed waiting times are $1, 2, 3, 4$ minutes.

$$L(\lambda) = \prod_{i=1}^{4} f_\lambda(x_i) = \lambda^4 e^{-10\lambda}, \qquad \lambda > 0.$$



$\circ\ \lambda = 0.2:\ L = 0.000217$ $\quad \bullet\ \lambda = 0.4:\ L = 0.000469$ $\quad \square\ \lambda = 0.8:\ L = 0.000137$

Connection: Multiply the four same-colored density heights to obtain that candidate's likelihood.

Caution: Density heights are not probabilities; L is not a PDF for λ.

Example: Maximizing the exponential likelihood

Example (The same four waiting times)

1. Find a candidate. Take logs and set the derivative to zero:

$$\ell(\lambda) = n \log \lambda - \lambda \sum_i x_i,$$

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_i x_i = 0 \quad \Longrightarrow \quad \widehat{\lambda} = \frac{n}{\sum_i x_i} = \frac{1}{\bar{x}}.$$

2. Check the maximum. The log likelihood curves downward:

$$\ell''(\lambda) = -\frac{n}{\lambda^2} < 0, \qquad \lambda > 0.$$

Thus the positive candidate is the unique maximizer.

Our data: For the waiting times $1, 2, 3, 4$ minutes,

$$\widehat{\lambda} = \frac{4}{10} = 0.4 \text{ per minute}, \qquad \frac{1}{\widehat{\lambda}} = 2.5 \text{ minutes}.$$

Estimator: $\widehat{\lambda}_{\mathrm{MLE}} = 1/\bar{X}_n$. A longer average wait gives a smaller estimated arrival rate.

Pop-up quiz: Interpreting the exponential MLE

For the waiting times $1, 2, 3, 4$ minutes,

$$L(\lambda) = \lambda^4 e^{-10\lambda}, \qquad \lambda > 0.$$

Question: Which statement is correct?

A. The factor $e^{-10\lambda}$ can be dropped because the data are fixed.

B. Taking the log likelihood can change the maximizing rate.

C. The estimated rate is 0.4 per minute, giving a fitted mean wait of 2.5 minutes.

D. The likelihood is a probability density for the unknown rate λ.

Answer: C.

$$\widehat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{2.5} = 0.4 \text{ per minute}, \qquad \frac{1}{\widehat{\lambda}} = 2.5 \text{ minutes}.$$

Both factors in $L(\lambda)$ depend on λ. Logs preserve the maximizer, and likelihood compares parameter values for the fixed data.

Method of moments versus maximum likelihood

Questions

- Do the two methods always give the same estimator?
- If they differ, how do their mean squared errors compare?

	Method of moments	**Maximum likelihood**
Principle	Match population and sample moments	Maximize the likelihood of the observed data
Bernoulli parameter q	$\widehat{q} = \bar{X}_n$	$\widehat{q} = \bar{X}_n$
Exponential rate λ	$\widehat{\lambda} = 1/\bar{X}_n$	$\widehat{\lambda} = 1/\bar{X}_n$

In the two examples above, these principles give the same estimator, with the same bias, MSE, and consistency.

Let's see an example where these estimators differ.

Example: Estimating an unknown upper limit

Example (A uniform model)

Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{Uniform}[0, \theta]$, with $\theta > 0$. For nonnegative observations, write $x_{\max} = \max_i x_i > 0$.

$$f_\theta(x) = \frac{1}{\theta} \quad \text{for } 0 \le x \le \theta, \qquad 0 \text{ otherwise.}$$

Method of moments

Match the population and sample means:

$$\mathbb{E}[X] = \frac{\theta}{2}, \qquad \frac{\widehat{\theta}_{\mathrm{MM}}}{2} = \bar{x}.$$

Solve the moment equation for θ.

$$\widehat{\theta}_{\mathrm{MM}} = 2\bar{x}.$$

Maximum likelihood

Keep the support restrictions:

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

On $\theta \ge x_{\max}$, $L(\theta)$ decreases.

$$\widehat{\theta}_{\mathrm{MLE}} = x_{\max}.$$

Message: MoM fits the mean. MLE selects the smallest upper limit covering all observations.

Example: Two methods can give different estimates

Example (The same observations, two fitting criteria)

Under the $\mathrm{Uniform}[0, \theta]$ model, we observe $1, 2, 3, 4$, so

$$\bar{x} = 2.5, \qquad x_{\max} = 4.$$

Method of moments: Match the population and sample means:

$$\frac{\widehat{\theta}_{\mathrm{MM}}}{2} = 2.5 \quad \Longrightarrow \quad \widehat{\theta}_{\mathrm{MM}} = 5.$$

Maximum likelihood: Choose the smallest upper limit covering the data:

$$\widehat{\theta}_{\mathrm{MLE}} = x_{\max} = 4.$$

Message: The two methods need not agree, even with the same model and data.

Next question: Which rule has smaller average squared error across repeated samples?

Example: MLE can have smaller MSE

Example (The uniform model, with sample size $n = 4$)

For $X_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0, \theta]$, compare $2\bar{X}_4$ (MoM) and $M = \max_i X_i$ (MLE).

Recall $\text{MSE}(T) = \text{Var}(T) + \big(\text{Bias}(T)\big)^2$.

For $0 \le t \le \theta$, independence gives $P(M \le t) = P(X_1 \le t, \ldots, X_4 \le t) = (t/\theta)^4$.

Its density gives $\mathbb{E}[M] = 4\theta/5$ and $\mathbb{E}[M^2] = 2\theta^2/3$.

Estimator	**Bias**	**Variance**	**MSE**
MoM: $2\bar{X}_4$	0	$\theta^2/12$	$\theta^2/12$
MLE: M	$-\theta/5$	$2\theta^2/75$	$\theta^2/15$

Result: MLE has **20% smaller MSE than MoM** here. Its smaller variance more than offsets its squared bias.

Example: MoM can have smaller MSE

Example (A different uniform model, still with $n = 4$)

Now let $X_1, \ldots, X_4$ be i.i.d. $\mathrm{Uniform}[\theta, 2\theta]$, with $\theta > 0$.

MoM: Matching $\mathbb{E}[X] = 3\theta/2$ gives $\widehat{\theta}_{\mathrm{MM}} = 2\bar{X}_4/3$.

MLE: $L(\theta) = \theta^{-4}$ for $\max_i x_i/2 \le \theta \le \min_i x_i$, and 0 otherwise.

It decreases on this interval, giving $\widehat{\theta}_{\mathrm{MLE}} = \max_i X_i/2$.

Estimator	**Bias**	**Variance**	**MSE**
MoM: $2\bar{X}_4/3$	0	$\theta^2/108$	$\theta^2/108$
MLE: $\max_i X_i/2$	$-\theta/10$	$\theta^2/150$	$\theta^2/60$

Result: MoM has smaller MSE here. MLE's squared bias outweighs its variance advantage.

Takeaway: Neither method always has smaller finite-sample MSE.

Wrap-up

Maximum likelihood estimation

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i), \qquad \widehat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} L(\theta).$$

- **Construct:** Fix the data and compare parameter values; logs simplify the calculation.
- **Check:** Respect the support and parameter range, and verify the maximum.
- **Assess:** Study bias, MSE, and consistency through the sampling distribution.

Relation to MoM: The methods may give the same estimator.

- When they differ, **either can have smaller finite-sample MSE**.

Suggested reading: [Ric07, Chapter 8].

Next lecture: Final review.

References



John A Rice.
Mathematical statistics and data analysis, volume 9.
Thomson brooks/cole, Belmont, CA, 2007.