March 25

Two cities

	infection rate
City I -	5 %
City II -	10 %

A Q. A 100 people arrive from "one" of the cities. After screening it was found that 9 out of 100 people were infected. then the question is

"From which city did they come"?

9% is closer to 10% as compared to the 5%.

So, an "educated guess" is they came from "city 1".

Parameter space

$\theta \in \{5\%, 10\%\}$

$p \in (\quad , \quad)$

Let $X_1, \ldots, X_n$ are IID observations from Bernoulli(p), $p \in (0,1)$ distribution.

The likelihood function

$$L(\hat{p} \mid X_1, \ldots, X_n)$$

$$= P(X_1 = x_1, \ldots, X_n = x_n)$$

$$= \prod_{i=1}^{n} P(X_i = x_i)$$

$$= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

we know

$$P(X_i = x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$$

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

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

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

Denote $T_n = \sum_{i=1}^{n} X_i$.

Then $L(p)$.

$$= p^{T_n}(1-p)^{n-T_n}$$

We want to maximize $L(p)$ for p.

log-likelihood function

$$\ell(p) = \log L(p)$$

$$= T_n \log p + (n - T_n)\log(1-p)$$

The MLE

$$\hat{p} = \arg\max_{p \in (0,1)} \underline{L(p)}$$

$$= \arg\max \underline{\ell(p)}$$

$p \in (0,1)$

Using Calculus, if $\hat{p}$ is a maxima then

$$\frac{d}{dp} l(p)\Big|_{p=\hat{p}} = 0$$

That is

$$\frac{d}{dp}\left[T_n \log p + (n-T_n)\log(1-p)\right]$$

$$= \frac{d}{dp}\left[T_n \log p + n\log(1-p) - T_n \log(1-p)\right]$$

$$= T_n \frac{1}{p} + n \cdot \frac{1}{1-p}(-1)$$

$$- T_n \frac{1}{1-p} (-1)$$

Equating $\frac{d}{dp} \ell(p) = 0$

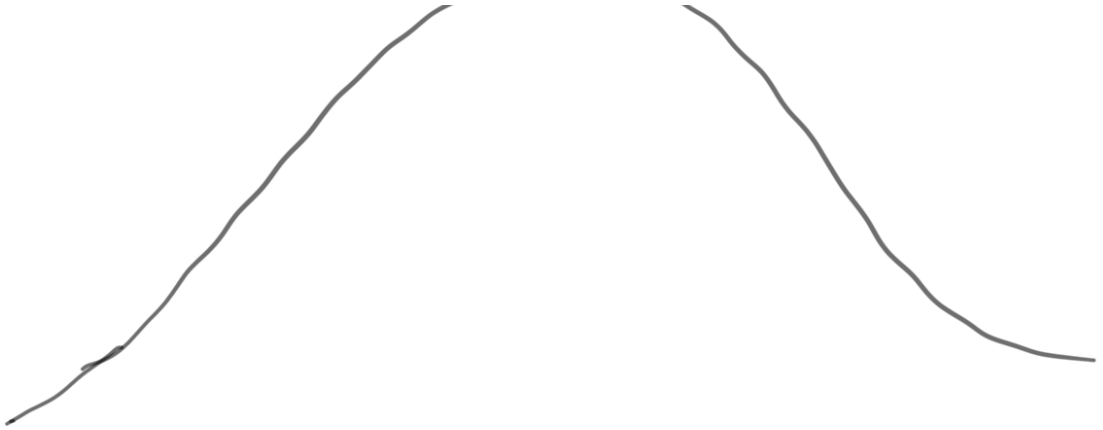
we get

$$\hat{p} = \frac{T_n}{n}$$

we can verify

$$\frac{d^2}{dp^2} \ell(p) \Big|_{p=\hat{p}} < 0$$

Concave function



Next Let
$X_1, \ldots X_n$ are IID
observations from
Binomial (m, p), $p \in (0,1)$.
p is unknown.
we want to find MLE
of p.
The likelihood function is
$L(p)$

$$= \prod_{i=1}^{n} P(X_i = x_i)$$

$$= \prod_{i=1}^{n} \binom{m}{x_i} p^{x_i} (1-p)^{n-x_i}$$

$$= \left[\prod_{i=1}^{n} \binom{m}{x_i} \right] p^{\sum_{i=1}^{n} x_i} (1-p)^{\sum_{i=1}^{n} (m-x_i)}$$

$\downarrow$ independent of p

$$= C \; p^{\sum x_i} (1-p)^{\sum (m-x_i)}$$

ω

MTL108
Point Estimation

Rahul Singh

Maximum Likelihood Estimation (MLE)

Motivation

Before we dive into the calculus, we must understand the core philosophy behind Maximum Likelihood Estimation (MLE). The principle asks a very intuitive question: *"Given the data we actually observed, what are the most plausible parameters of the population that produced it?"*

Definition 1 (The Likelihood Function, $L(\theta)$)**.** Once the data is observed, we treat the data as fixed constants and view the joint probability as a function of the unknown parameter θ. Because the samples are independent, their joint probability is the product of their individual marginal probabili-

ties,

$$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Example 1 (Quality Control in Manufacturing)**.** Suppose you are inspecting a batch of new smartphone batteries. You test 5 batteries at random, and exactly 1 is defective. What is the true defect rate (p) of the entire factory? Could the defect rate be 99%? It's physically possible, but if it were, pulling 4 working batteries out of 5 would be nearly impossible. Could it be 0.0001%? Also unlikely, because you found a defect so quickly. Common sense dictates that the most likely defect rate is $\frac{1}{5} = 0.20$. MLE is the mathematical engine that proves our common-sense intuition is formally optimal.

A coin: the probability of getting head

Consider there is a coin, but you do not know if it is perfectly balanced or weighted. You decide to run an experiment: you flip the coin $n = 10$ times, and

you observe exactly $x = 8$ Heads and 2 Tails.

We know that the number of heads follows a Binomial distribution. The probability of observing exactly 8 heads out of 10 flips is given by the Binomial Probability Mass Function (PMF),

$$P(X = 8 \mid p) = \binom{10}{8} p^8 (1-p)^{10-8} = 45\, p^8 (1-p)^2$$

Here, the data (8 Heads) is fixed and already observed. What is unknown is p, the true probability of flipping a Head. The core idea of Maximum Likelihood Estimation (MLE) is to treat this probability equation as a function of the unknown parameter p, which we call the **Likelihood Function**,

$$L(p) = 45p^8(1-p)^2$$

To find the most plausible value for p, we can plug in different hypothetical values for p and see which one yields the highest likelihood of producing the data we actually saw.

Analyzing the Results:

- If someone claims the coin is heavily biased toward tails ($p = 0.1$), the probability of getting

Table 1: Likelihood of observing exactly 8 Heads in 10 flips for various p.

Hypothetical Probability (p)	**Likelihood:** $45p^8(1-p)^2$	**Likelihood $L(p)$ value**
0.1	$45(0.1)^8(0.9)^2$	0.00000036
0.2	$45(0.2)^8(0.8)^2$	0.000074
0.3	$45(0.3)^8(0.7)^2$	0.00145
0.4	$45(0.4)^8(0.6)^2$	0.01062
0.5 *(Fair Coin)*	$45(0.5)^8(0.5)^2$	0.04395
0.6	$45(0.6)^8(0.4)^2$	0.12093
0.7	$45(0.7)^8(0.3)^2$	0.23347
0.8	$\mathbf{45(0.8)^8(0.2)^2}$	**0.30199**
0.9	$45(0.9)^8(0.1)^2$	0.19371

8 heads is astronomically low (less than 1 in a million). We can safely reject this hypothesis.

- If someone claims the coin is perfectly fair ($p = 0.5$), the probability of getting 8 heads is still quite low, at only about 4.4%.
- However, if we assume the coin has a $p = 0.8$ chance of landing on heads, the probability of observing exactly 8 heads peaks at about 30.2%.

MLE simply formalizes this common-sense logic using calculus. Instead of guessing values in a table, we take the derivative of the likelihood function to mathematically prove that $\hat{p} = 0.8$ (which is exactly $\frac{8}{10}$) is the parameter that strictly *maximizes the likelihood* of seeing the data we just collected.

The Parameter Space (Θ)

The **parameter space**, typically denoted by the capital Greek letter Theta (Θ), is the mathematical set of all possible, valid values that the unknown parameter (or vector of parameters) θ can assume.

When we define a probability distribution, the parameters are inherently subject to logical and mathematical constraints. For example:

- **Bernoulli and Binomial (p):** The parameter p represents a probability, which must be between 0 and 1. Therefore, the parameter space is $\Theta = [0, 1]$.
- **Poisson and Exponential (λ):** A rate of occurrence or an average count must be strictly greater than zero. Therefore,

$$\Theta = (0, \infty)$$

- **Normal (μ, σ^2):** The mean can be any real number, but the variance must be strictly positive. Therefore, the joint two-dimensional parameter space is defined as

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$$

Remark 1. (Connection to MLE) Defining the parameter space is critical because Maximum Likelihood Estimation is a *constrained* optimization problem. When we attempt to find our estimator $\hat{\theta}$, we are strictly searching for the maximum likelihood *within* the boundaries of Θ.

Remark 2. The standard approaches yield a critical point that falls outside of Θ (for instance, if setting the derivative to zero mathematically suggests a negative variance or a probability greater than 1), we must reject that point. In such cases, the true MLE will lie strictly on the boundary of the parameter space.

Mathematical Approach

Let $X_1, X_2, \ldots, X_n$ be a random sample of independent and identically distributed (IID) observations drawn from a population with a probability density function (PDF) or probability mass function (PMF) denoted by $f(x;\theta)$, where θ is an unknown parameter (or vector of parameters).

Definition 2 (The Log-Likelihood Function, $l(\theta)$)**.** Multiplying many small probabilities leads to arithmetic underflow and difficult calculus. Because the natural logarithm is a strictly monotonically increasing function, the θ that maximizes $L(\theta)$ will also perfectly maximize $\ln L(\theta)$. We define the log-likelihood as

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i; \theta).$$

Definition 3 (MLE)**.** Let $x = (x_1, x_2, \ldots, x_n)$ represents our observed data (IID sample) vector, the MLE $\hat{\theta}$ is formally defined as

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta \mid x),$$

where the arg max operator returns the *argument* (the input θ) that produces the *maximum* value of the function, rather than returning the maximum value itself.

Because the natural logarithm is a strictly monotonically increasing function, the parameter that

maximizes the likelihood function $L(\theta \mid x)$ is mathematically guaranteed to be the exact same parameter that maximizes the log-likelihood function $l(\theta \mid x)$. Therefore, the most common working definition in modern statistical computing and algorithmic implementation is

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} l(\theta \mid x) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln f(x_i \mid \theta).$$

Mathematically, Maximum Likelihood Estimation is fundamentally an optimization problem. The Maximum Likelihood Estimator (MLE) is defined as the specific value of the parameter θ that maximizes the likelihood function over the entire allowable parameter space Θ.

Finding MLE, when L is differentiable function of θ

To find the maximum, we must perform two steps from calculus.

1. **First Derivative (The Score Function):** We take the first derivative with respect to θ and

set it to zero to find the critical points, that is,

$$\frac{\partial}{\partial \theta} l(\theta) = 0$$

2. **Second Derivative (Concavity):** Setting the first derivative to zero only guarantees a flat slope (which could be a minimum, a maximum, or an inflection point). To mathematically prove that our estimator $\hat{\theta}$ is a **global maximum**, we must show that the log-likelihood function is strictly concave. We do this by proving the second derivative is strictly less than zero for all valid parameters, that is,

$$\frac{\partial^2}{\partial \theta^2} l(\theta) < 0.$$

MLE for Discrete Distributions

A. Bernoulli Distribution

Let $X_1, \ldots, X_n \sim$ Bernoulli(p). PMF: $P(X = x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$.

$$l(p) = \left(\sum x_i\right) \ln(p) + \left(n - \sum x_i\right) \ln(1-p)$$

$$\frac{\partial l}{\partial p} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0 \implies \hat{p} = \frac{\sum x_i}{n} = \bar{X}$$

Second Derivative Check:

$$\frac{\partial^2 l}{\partial p^2} = -\frac{\sum x_i}{p^2} - \frac{n - \sum x_i}{(1-p)^2}$$

Since $x_i \geq 0 \implies \sum x_i \geq 0$, and the sample size $n \geq \sum x_i$, both terms are strictly negative. Thus, $\frac{\partial^2 l}{\partial p^2} < 0$, confirming $\hat{p} = \bar{X}$ is the global maximum.

B. Binomial Distribution

Let $X_1, \ldots, X_m \sim$ Binomial(k, p), where k is the known number of trials per observation.

$$l(p) = C + \left(\sum x_i\right) \ln(p) + \left(mk - \sum x_i\right) \ln(1-p)$$

$$\frac{\partial l}{\partial p} = \frac{\sum x_i}{p} - \frac{mk - \sum x_i}{1-p} = 0 \implies \hat{p} = \frac{\sum x_i}{mk} = \frac{\bar{X}}{k}$$

Second Derivative Check:

$$\frac{\partial^2 l}{\partial p^2} = -\frac{\sum x_i}{p^2} - \frac{mk - \sum x_i}{(1-p)^2} < 0$$

Since the maximum possible value for $\sum x_i$ is mk, the numerators are non-negative, making the entire expression strictly negative.