MTL108
Point Estimation

Rahul Singh

Up to this point, we have calculated parameters (like μ and σ^2) assuming we knew the entire population. In reality, we almost never have access to the full population. Instead, we must use sample data to make **educated guesses** about these unknown parameters.

In many situations, we assume populations whose underlying probability distributions are known in form (e.g., Normal, Poisson) but depend on one or more unknown parameters, denoted by θ. The set of all possible, theoretically valid values for this parameter is called the parameter space, denoted by Θ.

Definition 1 (The Point Estimation Problem)**.** The Point Estimation Problem is the fundamental mathematical task of using the limited information contained within a random sample $X_1, X_2, \ldots, X_n$ to select a single "best" value from the parameter space Θ. This single value serves as our best, informed guess for the unknown parameter θ.

Solving this problem requires defining a rule for how to process the sample data, which leads to the critical distinction between an estimator and an estimate.

Point Estimator vs. Point Estimate

- **Point Estimator ($\hat{\theta}$):** A point estimator is the mathematical formula, rule, or function applied to the random sample to estimate θ. Formally, it is a statistic: $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$. Because it is a function of unobserved random variables, the estimator itself is a *random variable*. It possesses its own probability distribution (the sampling distribution), expected value, and variance.
- **Point Estimate:** A point estimate is the specific, realized numerical value computed by the estimator once the sample data has actually been observed. If our observed data is $x_1, x_2, \ldots, x_n$, the estimate is a fixed, deterministic constant: $\hat{\theta}_{\text{obs}} = t(x_1, x_2, \ldots, x_n)$. It contains no randomness.

Example 1. Suppose we wish to determine the true, unknown average lifespan of a specific mechanical component (μ).

1. **The Estimator:** We decide to use the sample mean as our point estimator: $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Before the test begins, the individual lifespans X_i are unknown, meaning $\bar{X}$ is a random variable subject to sampling error.
2. **The Estimate:** We test $n = 5$ components and observe failure times of $105, 110, 98, 102$, and 115 hours. Plugging these fixed constants into our rule yields 106 hours. The single

number 106 is our point estimate.

How do we know if an estimator is "good"? A good estimator should be accurate (centered on the true value) and precise (having low variance). We formalize these concepts through two primary properties: Unbiasedness and Consistency.

Unbiasedness: "Right on Average"

Definition 2. An estimator $\hat{\theta}$ is an unbiased estimator of a population parameter θ if its expected value is exactly equal to the true parameter, that is,

$$\mathbb{E}[\hat{\theta}] = \theta.$$

If $\mathbb{E}[\hat{\theta}] \neq \theta$, the estimator is biased, and the *bias* is defined as $\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$.

Example 2 (The Sample Mean)**.** We have learned that the sample mean $\bar{X}$ is an unbiased estimator of the population mean μ. Let $X_1, \ldots, X_n$ be a random sample where $\mathbb{E}[X_i] = \mu$. Then,

$$\mathbb{E}[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}(n\mu) = \mu.$$

Why Should We Care About Unbiasedness?

Unbiasedness guarantees that there is no systematic error in your measurement tool or mathematical formula. It ensures your estimator is "aiming" at the correct target.

Example 3 (Miscalibrated Digital Scale in Manufacturing)**.** Suppose you are a quality control engineer estimating the true average weight (θ) of titanium bolts produced by a new milling machine. You use a digital scale to measure a random sample of bolts, calculating the sample mean to serve as your estimator ($\hat{\theta}$).
However, suppose the digital scale was not properly "zeroed" and systematically overestimates every single weight by exactly 0.2 grams. Because of this mechanical bias, the expected value of your estimator is permanently shifted away from the truth,

$$\mathbb{E}[\hat{\theta}] = \theta + 0.2.$$

Here is the critical problem: collecting more data will not save you. Even if you spend an entire week weighing one million bolts, the Law of Large Numbers will simply cause your estimate to converge perfectly to the *wrong* number ($\theta + 0.2$). You will have an incredibly precise, yet fundamentally inaccurate estimate.
Unbiasedness ensures that your instruments and algorithms are not systematically lying to you, guaranteeing that, on average, your mathematical estimates reflect physical reality.

Example 4 (Home Blood Pressure Monitoring)**.** A biomedical engineer designing a digital home blood pressure cuff to estimate a patient's true average systolic blood pressure (θ). The

patient uses the cuff multiple times a week, calculating the sample mean to serve as their health estimator ($\hat{\theta}$).
However, suppose the cuff has a slightly flawed pressure valve that systematically underestimates the blood pressure by exactly 5 mmHg. Because of this mechanical bias, the expected value of the estimator is permanently shifted below the clinical truth, that is,

$$\mathbb{E}[\hat{\theta}] = \theta - 5.$$

The critical problem here is patient safety. Collecting more data will not fix the underlying error. Even if the patient takes their blood pressure $1,000$ times over the course of a year, the Law of Large Numbers will simply cause their average estimate to converge perfectly to the *wrong* number ($\theta - 5$). A patient with dangerous hypertension might be falsely classified as perfectly healthy because they possess an incredibly precise, yet fundamentally inaccurate estimate.
In biomedical engineering and clinical trials, unbiasedness is paramount. It ensures that medical devices and diagnostic algorithms are not systematically masking the truth, guaranteeing that, on average, the mathematical estimates reflect the patient's actual physiological state.

Consistency: The Value of "Big Data"

While unbiasedness describes the behavior of an estimator on average across many samples, it does not guarantee that any *single* estimate is actually close to the true parameter. For that, we need to know what happens as we gather more information.

Definition 3. An estimator $\hat{\theta}_n$ (where n is the sample size) is a consistent estimator of θ if it converges in probability to θ as the sample size approaches infinity. For any arbitrarily small $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

Alternatively, in terms of convergence of probability, this can be written as $\hat{\theta}_n \xrightarrow{p} \theta$.

Why Should We Care About Consistency?

Consistency is the mathematical justification for collecting "Big Data." It guarantees that if you invest the time, money, and computing power to collect a massive sample, your estimate will eventually lock onto the absolute truth.

Example 5 (Optical Tracking in Sports Analytics)**.** Consider an automated optical tracking system (like Hawk-Eye) calculating the exact 3D spatial trajectory of a basketball to evaluate a shooter's mechanics. The system captures the ball's position using sample frames. If the spatial estimator is *consistent*, then increasing the camera's frame rate from 30 frames per second (fps) to 300 fps, and eventually towards infinity, guarantees that the estimated physical trajectory $\hat{\theta}_n$ converges perfectly to the true physical trajectory θ.
If an algorithm is *inconsistent*, it means there is a fundamental flaw in the math. No matter how much money you spend upgrading to 10,000 fps cameras, the calculated trajectory will still retain a margin of error. We should discard inconsistent estimators because they render

additional data collection entirely useless.

Unbiasedness vs. Consistency: The Critical Differences

Students often confuse these two properties. While an estimator is often both unbiased and consistent (like the sample mean $\bar{X}$), they measure completely different phenomena.

- **Unbiased but Inconsistent:** Imagine trying to estimate the average height of a population, but your estimator rule is to simply measure the height of the very first person in your sample and ignore the rest: $\hat{\theta} = X_1$. Because $\mathbb{E}[X_1] = \mu$, this estimator is completely unbiased! However, as $n \to \infty$, your estimate never improves. It does not converge to μ. It is unbiased, but inconsistent.
- **Biased but Consistent:** Recall the unadjusted sample variance: $S_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$. We know its expected value is $\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2$, meaning it systematically underestimates the true variance (it is biased). However, as $n \to \infty$, the fraction $\frac{n-1}{n}$ approaches 1, and the bias disappears. With infinite data, it converges perfectly to σ^2. It is biased, but consistent.

Remark 1. If forced to choose, one should prefer a *consistent* estimator over an *unbiased* one. A consistent estimator can be fixed simply by collecting more data; an inconsistent estimator cannot be saved by data.

When is Unbiasedness Preferred?

While consistency is the gold standard in the era of "Big Data," there are critical real-world scenarios where consistency is mathematically useless, and unbiasedness becomes the absolute priority. As statisticians, we must recognize when to prioritize being "right on average" over asymptotic guarantees.

A. Strict Small-Sample Constraints (The Asymptotic Illusion)

Consistency is a limit property; it only holds true as $n \to \infty$. If you are working in a domain where data is fundamentally limited, expensive, or destructive, asymptotic guarantees mean nothing.

Example 6 (Destructive Testing)**.** Imagine testing the tensile strength of a new, highly expensive material. You only have the budget to destroy $n = 4$ samples. A consistent but biased estimator might only become accurate when $n > 100$. In this scenario, an unbiased estimator is preferred because it guarantees that your estimate is perfectly centered on the true parameter right now, even for $n = 4$.

B. Aggregation of Independent Estimates (The Compounding Error Problem)

Unbiasedness is absolutely critical when many separate, independent estimates are going to be summed together to form a macroscopic total. If you use a biased estimator, the bias will accumulate linearly, leading to a massive systemic error. If you use an unbiased estimator, the random overestimations and underestimations will naturally cancel each other out.

Example 7 (Aggregating Spatial Data in Satellite Imagery)**.** Suppose you are estimating the total vegetation loss across an entire subcontinent. You divide the map into 10,000 small, independent geographic grids. Due to persistent cloud cover, you only have a very small sample of clear pixels ($n \approx 10$) for each grid.
Let $\hat{\theta}_i$ be your estimator for the vegetation loss in grid i, and the true value is θ_i. If you use a biased estimator with a tiny, seemingly negligible systematic bias of just +0.5 hectares per grid,

$$\mathbb{E}[\hat{\theta}_i] = \theta_i + 0.5$$

When you aggregate the total vegetation loss across all 10,000 grids, the total expected value is

$$E\left[\sum_{i=1}^{10000} \hat{\theta}_i\right] = \sum_{i=1}^{10000} \mathbb{E}[\hat{\theta}_i] = \sum_{i=1}^{10000} (\theta_i + 0.5) = \text{True Total} + 5000 \text{ hectares}$$

Because the estimator was biased, the tiny microscopic error compounded into a massive, study-ruining macroscopic error of 5,000 hectares.
However, if you use an *unbiased* estimator, $\mathbb{E}[\hat{\theta}_i] = \theta_i$. By the linearity of expectation,

$$E\left[\sum_{i=1}^{10000} \hat{\theta}_i\right] = \sum_{i=1}^{10000} \mathbb{E}[\hat{\theta}_i] = \sum_{i=1}^{10000} \theta_i = \text{True Total}$$

Even though the estimate for any single grid might have high variance (due to the small $n = 10$ sample size), the positive and negative errors cancel out perfectly when aggregated. The final continental estimate will be highly accurate.

Remark 2. In fields like spatial modeling, financial accounting, or sports analytics—where macroscopic totals are calculated by summing thousands of microscopic estimates, unbiasedness should be preferred to prevent systematic drift.

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 4 (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 probabilities,

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

Example 8 (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 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%.

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

- 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 3. (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 4. 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 5 (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 6 (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 twice 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

To derive the Maximum Likelihood Estimator (MLE) for a given distribution, we generally follow a systematic process: construct the likelihood function from the sample, take its natural logarithm to simplify the math, find the critical point by setting the first derivative to zero, and finally, verify it is a global maximum using the second derivative test.

A. Bernoulli Distribution

Let $X_1, \ldots, X_n$ be a random sample where each $X_i \sim \text{Bernoulli}(p)$. The probability mass function (PMF) for a single observation is $P(X = x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$.

First, we construct the likelihood function, $L(p)$, by taking the product of the individual PMFs:

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

Next, we take the natural logarithm to obtain the log-likelihood function, $l(p)$. This allows us to convert the exponents into multipliers:

$$l(p) = \left(\sum_{i=1}^{n} x_i\right) \ln(p) + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

To find the maximum, we take the first derivative with respect to p and set it equal to zero:

$$\frac{\partial l}{\partial p} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0$$

Now, we use basic algebra to solve for p. We move the negative term to the right side and cross-multiply:

$$\frac{\sum x_i}{p} = \frac{n - \sum x_i}{1-p} \implies (1-p)\sum x_i = p(n - \sum x_i)$$

Expanding both sides:

$$\sum x_i - p\sum x_i = pn - p\sum x_i$$

The $p\sum x_i$ terms cancel out, leaving us with our estimator:

$$pn = \sum x_i \implies \hat{p} = \frac{\sum x_i}{n} = \bar{X}$$

Second Derivative Check: To ensure this critical point is a global maximum, we take the second derivative of the log-likelihood function:

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

Since our data consists of counts, $x_i \geq 0$, which implies $\sum x_i \geq 0$. Furthermore, the total sample size n is always greater than or equal to the sum of successes, meaning $(n - \sum x_i) \geq 0$. Because the numerators are non-negative and the denominators are squared (strictly positive), both terms being subtracted are positive. Thus, $\frac{\partial^2 l}{\partial p^2} < 0$, confirming $\hat{p} = \bar{X}$ is the global maximum.

B. Binomial Distribution

Let $X_1, \ldots, X_m \sim \text{Binomial}(k, p)$, where k is the known number of trials per observation, and m is the number of observations. The PMF is $\binom{k}{x} p^x (1-p)^{k-x}$.

The likelihood function is the product across all m observations:

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

When we take the natural logarithm, we can group the combinatorial constants into a single term, $C = \ln\left(\prod \binom{k}{x_i}\right)$.

$$l(p) = C + \left(\sum_{i=1}^{m} x_i\right) \ln(p) + \left(mk - \sum_{i=1}^{m} x_i\right) \ln(1-p)$$

Taking the derivative with respect to p, the constant C disappears:

$$\frac{\partial l}{\partial p} = \frac{\sum x_i}{p} - \frac{mk - \sum x_i}{1-p} = 0$$

Solving for p follows the exact same cross-multiplication logic as the Bernoulli distribution:

$$(1-p)\sum x_i = p(mk - \sum x_i) \implies \sum x_i - p\sum x_i = pmk - p\sum x_i$$

$$\hat{p} = \frac{\sum x_i}{mk} = \frac{\bar{X}}{k}$$

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 (if every trial is a success), the numerators are guaranteed to be non-negative, making the entire expression strictly negative.

C. Poisson Distribution

Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. The PMF is $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$.

We construct the likelihood function:

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}$$

Taking the log-likelihood allows us to separate the parameter λ from the data's factorial term:

$$l(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right) \ln(\lambda) - \ln\left(\prod_{i=1}^{n} x_i!\right)$$

We take the derivative with respect to λ (the factorial term becomes zero):

$$\frac{\partial l}{\partial \lambda} = -n + \frac{\sum x_i}{\lambda} = 0$$

Rearranging to solve for λ:

$$n = \frac{\sum x_i}{\lambda} \implies \hat{\lambda} = \frac{\sum x_i}{n} = \bar{X}$$

Second Derivative Check:

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{\sum x_i}{\lambda^2}$$

Because Poisson counts are non-negative ($\sum x_i \geq 0$) and $\lambda^2 > 0$, the second derivative is always strictly negative, ensuring a global maximum.

D. Geometric Distribution

Let $X_1, \ldots, X_n \sim \text{Geometric}(p)$, representing the number of trials *up to and including* the first success. The PMF is $P(X = x) = (1-p)^{x-1}p$.

The likelihood function is:

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

Taking the natural logarithm yields:

$$l(p) = n \ln(p) + \left(\sum_{i=1}^{n} x_i - n\right) \ln(1-p)$$

We set the first derivative to zero:

$$\frac{\partial l}{\partial p} = \frac{n}{p} - \frac{\sum x_i - n}{1-p} = 0$$

Cross-multiplying to solve for p:

$$n(1-p) = p\left(\sum x_i - n\right) \implies n - np = p\sum x_i - np$$

$$n = p\sum x_i \implies \hat{p} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}}$$

Second Derivative Check:

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

Since each $x_i \geq 1$ (you must have at least one trial to get a success), the sum $\sum x_i \geq n$. This ensures both subtracted terms are positive, resulting in overall concavity.

E. Negative Binomial Distribution

Let $X_1, \ldots, X_n \sim \text{NB}(r, p)$, representing the total number of trials required to achieve r successes (where r is a known constant).

The likelihood function is proportional to the probability parameters:

$$L(p) \propto \prod_{i=1}^{n} p^r (1-p)^{x_i - r} = p^{nr}(1-p)^{\sum x_i - nr}$$

The log-likelihood is:

$$l(p) = nr\ln(p) + \left(\sum_{i=1}^{n} x_i - nr\right)\ln(1-p)$$

Taking the derivative and setting it to zero:

$$\frac{\partial l}{\partial p} = \frac{nr}{p} - \frac{\sum x_i - nr}{1-p} = 0$$

Solving for p:

$$nr(1-p) = p\left(\sum x_i - nr\right) \implies nr - pnr = p\sum x_i - pnr$$

$$\hat{p} = \frac{nr}{\sum x_i} = \frac{r}{\bar{X}}$$

Second Derivative Check:

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

Since each observation requires at least r trials to get r successes ($x_i \geq r$), the total sum $\sum x_i \geq nr$, ensuring the second derivative is strictly negative.

F. Discrete Uniform Distribution

Let $X_1, \ldots, X_n \sim DisUnif(1, N)$, where N is an unknown integer representing the maximum possible value. The PMF is $P(X = x) = \frac{1}{N}$.

The likelihood function is simply:

$$L(N) = \prod_{i=1}^{n} \frac{1}{N} = \frac{1}{N^n}$$

Crucial Note: We cannot use our standard calculus steps here. The parameter N dictates the upper boundary of the support, which violates the continuity assumptions required for derivatives. Instead, we must use logic. For the likelihood to be valid (non-zero), our parameter N must be at least as large as the largest data point we actually observed.

To maximize the fraction $\frac{1}{N^n}$, we need the denominator to be as small as possible. Therefore, we logically select the smallest valid N, which is the maximum observation in our sample (the n-th order statistic):

$$\hat{N} = \max(X_1, X_2, \ldots, X_n) = X_{(n)}$$

MLE for Continuous Distributions

The process for continuous distributions is nearly identical to discrete distributions, except we build our likelihood function using the Probability Density Function (PDF) instead of the PMF.

G. Continuous Uniform Distribution

Let $X_1, \ldots, X_n \sim U(0, \theta)$. The PDF is $f(x) = \frac{1}{\theta}$ for $0 \leq x \leq \theta$.

Just like the discrete uniform case, the likelihood function is:

$$L(\theta) = \frac{1}{\theta^n} \quad \text{for } \theta \geq \max(x_i)$$

Calculus fails here for the exact same reason: the parameter θ defines the boundary of the probability space. Because $\frac{1}{\theta^n}$ is a strictly decreasing function, its maximum occurs at the smallest possible value θ can take without contradicting the observed data.

$$\hat{\theta} = \max(X_1, X_2, \ldots, X_n) = X_{(n)}$$

H. Exponential Distribution

Let $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$. The PDF is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

We construct the likelihood function:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$

Taking the natural logarithm to bring down the exponents:

$$l(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

We find the critical point by setting the derivative to zero:

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum x_i = 0$$

Rearranging to solve for λ:

$$\frac{n}{\lambda} = \sum x_i \implies \hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}}$$

Second Derivative Check:

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0$$

Because $n > 0$ and $\lambda^2 > 0$, the second derivative is strictly negative, guaranteeing our estimator is a global maximum.

I. Normal Distribution

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. The PDF is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Because the Normal distribution has two parameters, we must establish the log-likelihood function and then take partial derivatives with respect to each parameter.

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$l(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

1. MLE for the Mean (μ): We take the partial derivative with respect to μ, treating σ^2 as a constant:

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)(-1) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

Multiplying by σ^2 removes the denominator:

$$\sum_{i=1}^{n} x_i - n\mu = 0 \implies \hat{\mu} = \frac{\sum x_i}{n} = \bar{X}$$

Second Derivative Check: $\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0$, confirming $\bar{X}$ is a global maximum for μ.

2. MLE for the Variance (σ^2): To make the calculus cleaner, let us substitute $v = \sigma^2$. We take the partial derivative with respect to v:

$$\frac{\partial l}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

To solve for v, we multiply the entire equation by $2v^2$:

$$-nv + \sum_{i=1}^{n}(x_i-\mu)^2 = 0 \implies \hat{v} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

Therefore, the MLE for the variance is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{X})^2$.

Second Derivative Check for Variance: We evaluate the second derivative with respect to v at the critical point $\hat{v}$:

$$\frac{\partial^2 l}{\partial v^2} = \frac{n}{2v^2} - \frac{1}{v^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

We know from our critical point that $\sum(x_i - \mu)^2 = n\hat{v}$. Substituting this back into the second derivative:

$$\left.\frac{\partial^2 l}{\partial v^2}\right|_{v=\hat{v}} = \frac{n}{2\hat{v}^2} - \frac{n\hat{v}}{\hat{v}^3} = \frac{n}{2\hat{v}^2} - \frac{2n}{2\hat{v}^2} = -\frac{n}{2\hat{v}^2} < 0$$

Because the second derivative evaluated at the critical point is negative, $\hat{\sigma}^2$ is confirmed to be the global maximum.

Mean Square Error (MSE)

Definition 7 (MSE)**.** For an estimator $\hat{\theta}$ for a parameter θ, the MSE is defined as

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

Lemma 1. *For an estimator $\hat{\theta}$ for a parameter θ,*

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \Big(Bias(\hat{\theta})\Big)^2,$$

where $\Big(Bias(\hat{\theta})\Big) = \mathbb{E}(\hat{\theta}) - \theta$. Moreover, if an estimator is unbiased, its MSE is simply equal to its variance.

Proof. We have

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}\left[\Big((\hat{\theta} - \mathbb{E}(\hat{\theta})) + (\mathbb{E}(\hat{\theta}) - \theta)\Big)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + (\mathbb{E}(\hat{\theta}) - \theta)^2 + 2(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)\right].
\end{aligned}$$

Using the linearity of expectation,

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right] + \mathbb{E}\left[(\mathbb{E}(\hat{\theta}) - \theta)^2\right] + 2\mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)\right]. \tag{1}$$

Observe that $(\mathbb{E}(\hat{\theta}) - \theta)$ is **not** a random variable, so

$$\mathbb{E}\left[(\mathbb{E}(\hat{\theta}) - \theta)^2\right] = (\mathbb{E}(\hat{\theta}) - \theta)^2 = \Big(\text{Bias}(\hat{\theta})\Big)^2. \tag{2}$$

Next, using the linearity of expectation, we get

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)\right] &= (\mathbb{E}(\hat{\theta}) - \theta)\,\mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))\right] \\
&= (\mathbb{E}(\hat{\theta}) - \theta)\,\left[\mathbb{E}(\hat{\theta}) - \mathbb{E}(\mathbb{E}(\hat{\theta}))\right] \\
&= (\mathbb{E}(\hat{\theta}) - \theta)\,\left[\mathbb{E}(\hat{\theta}) - \mathbb{E}(\hat{\theta})\right] = 0.
\end{aligned} \tag{3}$$

Combining (1), (2) and (3), we obtain

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \Big(\text{Bias}(\hat{\theta})\Big)^2.$$

This proves the first part. Next, if an estimator is unbiased, then $\text{Bias}(\hat{\theta}) = 0$, consequently

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}).$$

□

Remark 5. While unbiasedness is a desirable property, it is not the only criterion for evaluating an estimator. Sometimes, we might prefer an estimator with a tiny amount of bias if it significantly reduces the variance. The MSE is a comprehensive metric that evaluates an estimator by combining both its variance and its bias.

Combining Independent Unbiased Estimators

Let T_1 and T_2 denote independent unbiased estimators of θ, having known variances σ_1^2 and σ_2^2. That is,

$$\mathbb{E}[T_i] = \theta, \quad \text{Var}(T_i) = \sigma_i^2 \quad \text{for } i = 1, 2$$

Any combined estimator of the form

$$T = \lambda T_1 + (1 - \lambda) T_2$$

is also an unbiased estimator for θ. Because, using linearity of expectation

$$\begin{aligned}\mathbb{E}(T) =& \mathbb{E}[\lambda T_1 + (1-\lambda)T_2] \\ =& \lambda \mathbb{E}[T_1] + (1-\lambda)\mathbb{E}[T_2] \\ =& \lambda\theta + (1-\lambda)\theta = \theta.\end{aligned}$$

To determine the value of λ that results in T having the smallest possible MSE (which, since T is unbiased, is just its variance), we compute

$$r(\lambda) = \text{Var}(T) = \lambda^2 \text{Var}(T_1) + (1-\lambda)^2 \text{Var}(T_2) = \lambda^2 \sigma_1^2 + (1-\lambda)^2 \sigma_2^2.$$

To minimize this variance, we differentiate with respect to λ

$$\frac{d}{d\lambda} r(\lambda) = 2\lambda\sigma_1^2 - 2(1-\lambda)\sigma_2^2.$$

Setting this to 0 and solving for λ (let's call it $\hat{\lambda}$), we have

$$2\hat{\lambda}\sigma_1^2 = 2(1-\hat{\lambda})\sigma_2^2 \implies \hat{\lambda}(\sigma_1^2 + \sigma_2^2) = \sigma_2^2.$$

Check for the second derivative and show it's actually the global minimum. Consequently,

$$\hat{\lambda} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}.$$

Thus, the optimal weight given to an estimator is inversely proportional to its variance.

Comparing Estimators for a Uniform Distribution

Let $X_1, \ldots, X_n \sim U(0, \theta)$, where $U(0, \theta)$ is continuous uniform distribution with support $[0, \theta]$. We want to evaluate two estimators for the parameter θ.

Estimator 1: Since $\mathbb{E}[X_i] = \theta/2$, a natural unbiased estimator is based on the sample mean,

$$T_1 = 2\bar{X} = \frac{2}{n}\sum_{i=1}^{n} X_i$$

Since $\mathbb{E}[T_1] = \theta$, its MSE is just its variance. We know that $\text{Var}(X_i) = \theta^2/12$, so

$$\text{MSE}(T_1) = \text{Var}(T_1) = \frac{4}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{4}{n^2}\left(n\frac{\theta^2}{12}\right) = \frac{\theta^2}{3n}$$

Estimator 2: Maximum Likelihood Estimator The MLE for θ is $T_2 = \max(X_i) = X_{(n)}$. The CDF of T_2 is $F_2(x) = (x/\theta)^n$ for $x \leq \theta$, which gives the PDF

$$f_2(x) = \begin{cases} \dfrac{nx^{n-1}}{\theta^n}, & if \quad 0 \leq x \leq \theta \\ 0, & otherwise. \end{cases}$$

Now we find its mean and variance; observe that

$$\mathbb{E}[T_2] = \int_0^\theta x\frac{nx^{n-1}}{\theta^n}\,dx = \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n}{n+1}\theta,$$

$$\mathbb{E}[T_2^2] = \int_0^\theta x^2\frac{nx^{n-1}}{\theta^n}\,dx = \frac{n}{n+2}\theta^2.$$

So,

$$\text{Var}(T_2) = \mathbb{E}[T_2^2] - (\mathbb{E}[T_2])^2 = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \frac{n\theta^2}{(n+2)(n+1)^2}$$

Now, the MSE of T_2 (since T_2 is biased),

$$\begin{aligned}\text{MSE}(T_2) &= (\mathbb{E}[T_2] - \theta)^2 + \text{Var}(T_2) \\ &= \left(\frac{n}{n+1}\theta - \theta\right)^2 + \frac{n\theta^2}{(n+2)(n+1)^2} \\ &= \frac{\theta^2}{(n+1)^2} + \frac{n\theta^2}{(n+2)(n+1)^2} = \frac{(n+2)\theta^2 + n\theta^2}{(n+2)(n+1)^2} = \frac{2\theta^2}{(n+1)(n+2)}\end{aligned}$$

Conclusion: Because $\dfrac{2\theta^2}{(n+1)(n+2)} \leq \dfrac{\theta^2}{3n}$ for all $n \geq 1$, the Maximum Likelihood Estimator, T_2, is fundamentally superior to the estimator T_1, even though T_2 is biased.

Asymptotic properties of MLE

The Rao-Cramer Lower Bound (RCLB)

Theorem 1 (without proof)**.** *For an unbiased estimator $\hat{\theta}$, under standard regularity conditions,*

$$Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

where n is the sample size, and $I(\theta)$ is the Fisher Information of a single observation, defined as

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \ln f(X;\theta)\right]$$

An estimator that achieves this bound is called efficient.

The Rao-Cramer Lower Bound establishes the absolute minimum possible variance for any *unbiased* estimator of a parameter θ. It represents the theoretical limit of estimation precision.

Distribution of MLE under Standard Assumptions

Maximum Likelihood Estimators possess incredibly powerful asymptotic (large-sample) properties.

Theorem 2 (without proof)**.** *Under standard regularity conditions, as $n \to \infty$, the MLE $\hat{\theta}$ is*

1. ***Consistent:*** *$\hat{\theta} \xrightarrow{p} \theta$.*
2. ***Asymptotically Normal:*** *The distribution of the MLE approaches a normal distribution centered on the true parameter.*
3. ***Asymptotically Efficient:*** *The variance of the MLE reaches the Rao-Cramer Lower Bound.*

Formally, we write this asymptotic distribution as:

$$\hat{\theta}_{MLE} \sim N\left(\theta, \frac{1}{nI(\theta)}\right) \quad as\ n \to \infty$$

Healthcare Application

The Problem: An oncologist is leading a clinical trial for a revolutionary new targeted therapy for leukemia. The team tracks a sample of $n = 50$ patients to measure the time until remission ends (the "survival time," T). Biological survival times are strictly positive and typically right-skewed, suppose they follow an Exponential distribution, $T \sim \text{Exp}(\lambda)$, where λ represents the underlying hazard rate (the rate at which patients relapse).

Application: The doctors observe specific relapse times for the patients: $t_1, t_2, \ldots, t_{50}$. To understand the drug's true efficacy, they must estimate the unknown hazard rate, λ. They construct the likelihood function, which represents the joint probability of seeing exactly these 50 survival times:

$$L(\lambda) = \prod_{i=1}^{50} \lambda e^{-\lambda t_i}$$

Using calculus to maximize the log-likelihood, the Maximum Likelihood Estimator is

$$\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\bar{t}}$$

Conclusion:
By mathematically proving that this specific $\hat{\lambda}$ makes the observed clinical data more probable than any other possible parameter, the oncologist can confidently model the entire survival curve. This allows the hospital to give future patients highly accurate prognoses, such as calculating the exact probability that a patient will remain in remission for more than 5 years, $P(T > 5) = e^{-5\hat{\lambda}}$.

References

[1] Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability.* Chapman and Hall/CRC.

[2] Ross, Sheldon M. (2020). *Introduction to probability and statistics for engineers and scientists.* Academic press.

Disclaimer