Parameter Estimati

A Complete Visual Guide — From Concepts to Mastery

1 · Parameters & Statistics → 2 · Sampling Distribution →
3 · Estimates & Estimators → 4 · Quality of Estimators →
5 · Estimation Frameworks

01 Parameters & Statistics

The Big Picture

In machine learning and statistics, we almost never have access to the **entire population**. Instead, we collect a **sample** and try to learn about the population from it. This whole topic answers one question: *How do we make good guesses about population characteristics using sample data?*

Population (unknown) → Sample (observed data) →
Statistic (function of sample) → Estimate of Parameter

Parameter (θ)

A **parameter** is a fixed (but unknown) characteristic of the population distribution *F*.

> DEFINITION $\theta = t(F)$ — a function of the distribution

Examples: Population mean μ, population variance σ^2, probability p in a binomial.
Think of parameters as the *truth* we're trying to discover.

Statistic (T)

A **statistic** is any function computed from the sample data — no unknown values allowed.

> DEFINITION $T = s(x)$ — a function of sample data $x = (x_1, \ldots, x_n)$

Examples: Sample mean $\bar{x}$, sample variance s^2, sample proportion $\hat{p}$.
Not all statistics are useful — some are designed to estimate parameters.

> 🧠 MNEMONIC — "P-P-S-S"
> **P**arameters describe **P**opulations · **S**tatistics describe **S**amples

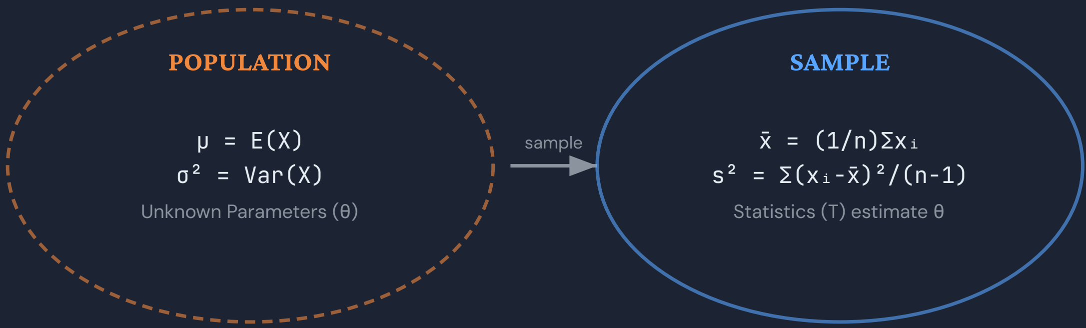
Probability Distribution Reminder

A random variable *X* has a **CDF** $F(x) = P(X \leq x)$ and an associated **PMF** (discrete) or **PDF** (continuous).

> DISCRETE (PMF) $f(x) = P(X = x)$

> CONTINUOUS (PDF) $P(a < X < b) = \int_a^b f(x)\, dx$

These distributions depend on parameters. Our job: figure out those parameters from data!

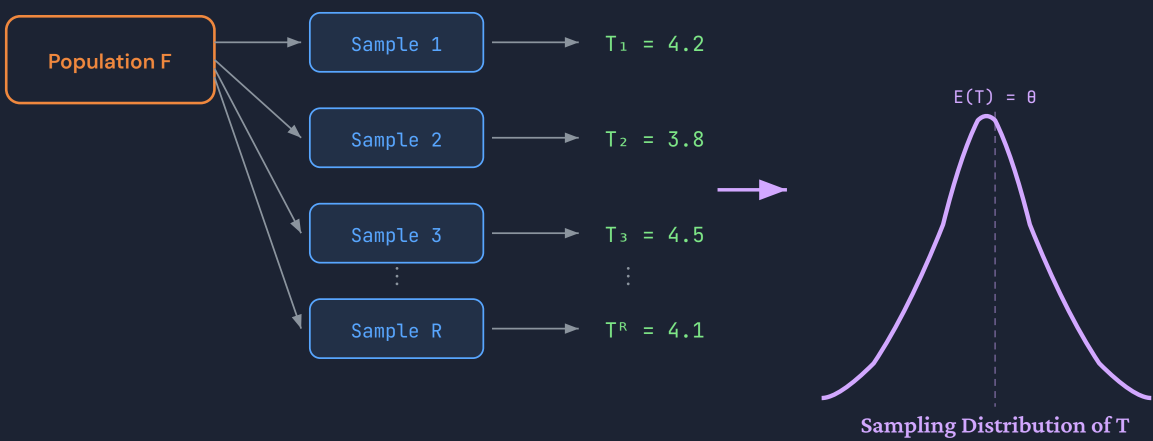


02 Sampling Distribution

Statistics Are Random Variables!

Since our data is random (drawn from *F*), any statistic T = s(**x**) computed from that data is *also* a random variable. If we drew a different sample, we'd get a different value of T.

The sampling distribution of T is what we'd see if we repeated the experiment infinitely many times — collecting R independent samples and computing T each time.



🧠 MNEMONIC — "DARTS"

Data → Apply statistic → Repeat many Times → Sampling distribution emerges

💡 KEY INSIGHT

The sampling distribution depends on *which* population the data comes from. Different F → different sampling distribution for the same statistic. As n grows, many sampling distributions become approximately Normal (thanks to the **Central Limit Theorem**).

03 Estimates & Estimators

Estimator vs Estimate

An **estimator** is the *recipe* (function $g(\cdot)$) you apply to data. An **estimate** is the *number* you get when you plug in actual data.

NOTATION `θ̂ = g(x) — "theta hat" estimates θ`

The "Hat" Convention

In statistics, placing a ˆ **(hat)** on a parameter means "estimate of". So:

$\hat{\theta}$ → estimate of θ

$\hat{\mu} = \bar{x}$ → estimate of μ

$\hat{\sigma}^2 = s^2$ → estimate of σ^2

🧠 MNEMONIC — "RECIPE VS DISH"

Estimator = the recipe (a function waiting for data) · **Estimate** = the dish (the actual number you compute)

Key Examples

EXAMPLE 1 — SAMPLE MEAN

Population parameter: **$\mu = E(X)$**

$$\bar{x} = (1/n) \Sigma_i x_i = \hat{\mu}$$

The sample mean estimates the population mean.

EXAMPLE 2 — SAMPLE VARIANCE

Population parameter: **$\sigma^2 = E[(X-\mu)^2]$**

$$s^2 = \Sigma(x_i - \bar{x})^2 / (n-1) = \hat{\sigma}^2$$

Why n-1? It makes s^2 *unbiased* (see next section!).

⚠ **WATCH OUT — TWO VARIANCE FORMULAS**

$s^2 = \Sigma(x_i - \bar{x})^2 / (n-1)$ → unbiased (divides by n-1)

$\tilde{s}^2 = \Sigma(x_i - \bar{x})^2 / n$ → biased downward (divides by n, the MLE)

Both are consistent (converge to σ^2 as $n\to\infty$), but s^2 is unbiased while $\tilde{s}^2$ underestimates σ^2.

04 Quality of Estimators

How Do We Judge an Estimator?

Not all estimators are created equal. There are four key properties to evaluate:

Bias | Variance | MSE = Bias² + Variance | Consistency

🧠 MNEMONIC — "BVCE" → "BEST VALUES COME EVENTUALLY"

Bias · Variance · Consistency · Efficiency — the four pillars of estimator quality

① Bias

Bias measures the *systematic error* — how far off the estimator is **on average**.

BIAS FORMULA $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$

$\text{Unbiased} \iff E(\hat{\theta}) = \theta \iff \text{Bias} = 0$

✅ UNBIASED: SAMPLE MEAN

$E(\bar{x}) = (1/n)\,\Sigma\, E(x_i) = (1/n)(n\mu) = \boldsymbol{\mu}$

Bias = $\mu - \mu = 0$ ✓

✅ UNBIASED: S² (WITH N-1)

$E(s^2) = \sigma^2$ (proven through algebra)

This is WHY we divide by n-1!

❌ BIASED: S̃² (WITH N)

$E(\tilde{s}^2) = E[(n-1)/n \cdot s^2] = (n-1)/n \cdot \sigma^2$

Since $(n-1)/n < 1$ for any finite n, $\tilde{s}^2$ **underestimates** σ^2 on average.

💡 SURPRISE: BIASED CAN BE GOOD!

Ridge regression, LASSO, and Elastic Net are all *intentionally biased* estimators — they trade a little bias for a big reduction in variance, often producing better predictions overall.

② Variance & Standard Error

Variance measures *how spread out* the estimates would be across repeated samples.

VARIANCE OF ESTIMATOR $\mathrm{Var}(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2]$

Standard Error: $SE(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$

VARIANCE OF SAMPLE MEAN

$\mathrm{Var}(\bar{x}) = \sigma^2 / n$

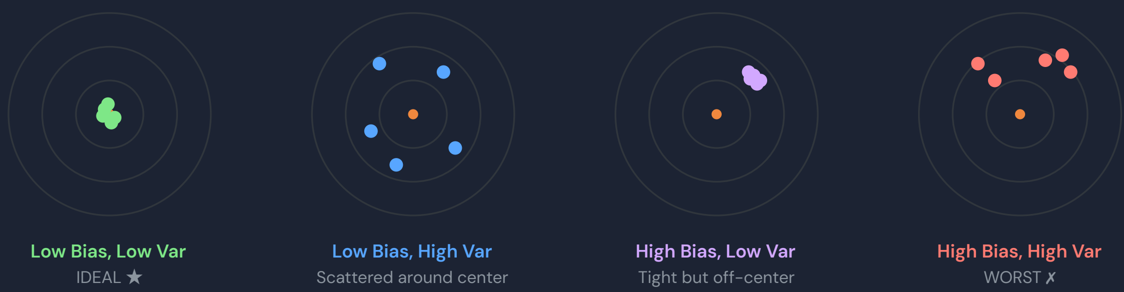
↑ More data (larger n) → smaller variance!

VARIANCE OF SAMPLE VARIANCE

$\mathrm{Var}(s^2) = (1/n)(\mu_4 - (n-3)/(n-1) \cdot \sigma^4)$

where $\mu_4 = E[(X-\mu)^4]$ (4th central moment)

The Bias-Variance Target Analogy



🟠 Bullseye = true parameter θ · Colored dots = estimates from different samples

③ Mean Squared Error (MSE) — The Gold Standard

MSE captures *both* bias and variance in one number. It's the preferred measure of estimator quality.

MSE DECOMPOSITION (THE KEY FORMULA) $\mathrm{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$

$= \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta})$

$= [E(\hat{\theta}) - \theta]^2 + E[(\hat{\theta} - E(\hat{\theta}))^2]$

🧠 **MNEMONIC**

MSE = B²V → "My Squared Error = Bias² + Variance" (like $E = mc^2$ but for estimators!)

1. **Start:** $\mathrm{MSE} = E[(\hat{\theta} - \theta)^2] = E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2$
2. **Note:** $\mathrm{Bias}^2 = [E(\hat{\theta}) - \theta]^2 = E(\hat{\theta})^2 - 2\theta E(\hat{\theta}) + \theta^2$
3. **Note:** $\mathrm{Var}(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2$
4. **Add them:** $\mathrm{Bias}^2 + \mathrm{Var} = E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2 = \mathrm{MSE}$ ✓

④ Consistency

An estimator is **consistent** if it converges to the true parameter as sample size grows.

CONSISTENCY $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$

$P(|\hat{\theta} - \theta| > \varepsilon) \to 0$ for any $\varepsilon > 0$

Any reasonable estimator should be consistent. $\bar{x}$, s^2, and $\tilde{s}^2$ are all consistent.

⑤ Efficiency

An estimator is **efficient** if it has the smallest MSE among all estimators of θ.

COMPARING EFFICIENCY $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)$

If both are unbiased: compare $Var(\hat{\theta}_1)$ vs $Var(\hat{\theta}_2)$

Quality Properties at a Glance

Property	What It Measures	Formula	Want It To Be
Bias	Systematic error (accuracy)	$E(\hat{\theta}) - \theta$	= 0 (ideally)
Variance	Spread/reliability (precision)	$E[(\hat{\theta} - E(\hat{\theta}))^2]$	Small
MSE	Overall quality	$Bias^2$ + Variance	Minimum
Consistency	Improves with more data?	$\hat{\theta} \to^p \theta$ as $n \to \infty$	Yes (always)
Efficiency	Best among competitors?	Lowest MSE	Yes

05 Estimation Frameworks

Three Main Approaches

There are three major frameworks for finding estimators. Each answers: "Given data, how do I compute $\hat{\theta}$?"

Least Squares (LS) | Method of Moments (MoM) | Maximum Likelihood (MLE)

🧠 MNEMONIC — "LMM" → "LEARN MY MODELS"

Least Squares · **M**ethod of Moments · **M**aximum Likelihood — the three estimation pillars

① Least Squares Estimation

Idea: Find the parameter that minimizes the sum of squared differences between data and the parameter.

LEAST SQUARES LOSS $LS(\theta|x) = \Sigma_i (h(x_i) - \theta)^2$

Typically $h(x) = x$, so: $LS(\mu|x) = \Sigma_i (x_i - \mu)^2$

1. **Write loss:** $LS(\mu) = \Sigma x_i^2 - 2\mu\Sigma x_i + n\mu^2$
2. **Differentiate:** $dLS/d\mu = -2\Sigma x_i + 2n\mu$
3. **Set = 0:** $2n\mu = 2\Sigma x_i \rightarrow$ $\hat{\mu} = \bar{x} = (1/n)\Sigma x_i$

📌 WHEN TO USE

Best for **mean parameters** and **regression coefficients**. Not ideal for variance parameters.

② Method of Moments (MoM)

Idea: Set *population moments* equal to *sample moments* and solve for the parameters.

CORE EQUATION Population moment: $\mu_j = E(X^j) = m_j(\theta_1,\ldots,\theta_p)$

Sample moment: $\hat{\mu}_j = (1/n)\,\Sigma_i\, x_i^{\,j}$

Set $\hat{\mu}_j = m_j(\hat{\theta}_1,\ldots,\hat{\theta}_p)$ and solve for $\hat{\theta}$'s

🧠 MNEMONIC

"Match the Moments!" — Population moments = Sample moments → solve for unknowns

EXAMPLE: NORMAL N(M, Σ²)

Population: $\mu_1 = \mu$, $\mu_2 = \mu^2 + \sigma^2$

Sample: $\hat{\mu}_1 = \bar{x}$, $\hat{\mu}_2 = \bar{x}^2 + \tilde{s}^2$

Solution: $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \tilde{s}^2 = (1/n)\Sigma(x_i-\bar{x})^2$

EXAMPLE: UNIFORM U[A,B]

Population: $\mu_1 = (a+b)/2$, $\mu_2 = (a^2+ab+b^2)/3$

Solution (via quadratic formula):

$\hat{a} = \hat{\mu}_1 - \sqrt{3}\cdot\sqrt{(\hat{\mu}_2 - \hat{\mu}_1^{\,2})}$

$\hat{b} = \hat{\mu}_1 + \sqrt{3}\cdot\sqrt{(\hat{\mu}_2 - \hat{\mu}_1^{\,2})}$

③ Maximum Likelihood Estimation (MLE) — The Star Player

Idea: Find the parameter values that make the observed data *most probable*.

LIKELIHOOD & LOG-LIKELIHOOD Likelihood: $L(\theta|x) = \Pi_i f(x_i|\theta)$

Log-Likelihood: $\ell(\theta|x) = \Sigma_i \log f(x_i|\theta)$

MLE: $\hat{\theta}$_MLE = arg max $L(\theta|x)$ = arg max $\ell(\theta|x)$

🧠 MNEMONIC — "MLE = MOST LIKELY EXPLANATION"

Which parameter values would have made this data *most likely* to occur? Those are the MLEs!

⭐ THREE SUPERPOWER PROPERTIES OF MLES

1. Consistent: $\hat{\theta}$_MLE → θ as $n \to \infty$

2. Asymptotically Efficient: Lowest variance as $n \to \infty$

3. Functionally Invariant: If $\hat{\theta}$ is MLE of θ, then $h(\hat{\theta})$ is MLE of $h(\theta)$

MLE Step-by-Step Examples

MLE FOR NORMAL DISTRIBUTION

1. **PDF:** $f(x|\mu,\sigma^2) = (1/\sqrt{(2\pi\sigma^2)}) \cdot \exp[-(x-\mu)^2/(2\sigma^2)]$
2. **Log-likelihood:** $l = -(1/2\sigma^2)\Sigma(x_i-\mu)^2 - (n/2)\log(\sigma^2) - c$
3. **For μ:** Maximize ⟹ minimize $\Sigma(x_i-\mu)^2$ → $\hat{\mu} = \bar{x}$
4. **For σ²:** $dl/d\sigma^2 = 0$ → $\hat{\sigma}^2 = (1/n)\Sigma(x_i-\bar{x})^2 = \hat{s}^2$

Note: MLE of σ² uses n (not n−1), so it's biased but consistent!

MLE FOR BINOMIAL B[N,P]

MLE FOR UNIFORM U[A,B]

Log-likelihood: $l = \log(p)\Sigma x_i + \log(1-p)(nN - \Sigma x_i) + c$

Derivative = 0: $(1-p)n\bar{x} - pn(N-\bar{x}) = 0$

Result: $\hat{p} = \bar{x}/N$

Log-likelihood: $l = -n \cdot \log(b-a)$

Maximize l = minimize $(b-a)$

Subject to $a \le$ all $x_i \le b$

Result: $\hat{a} = \min(x_i)$, $\hat{b} = \max(x_i)$

Framework Comparison

Framework	Core Idea	Strengths	Weaknesses
Least Squares	Minimize squared errors	Simple, intuitive, great for regression	Only good for means/regression; not for variance
Method of Moments	Match moments	Easy to compute, always gives an answer	Can be inefficient; may give impossible values
MLE	Maximize probability of data	Consistent, efficient, invariant; best overall	Can be biased in small samples; may need optimization

📋 Master Cheat Sheet

Sample Mean

```
x̄ = (1/n) Σ xᵢ
```

Unbiased for μ · Var = σ²/n · LS & MLE solution

Sample Variance (unbiased)

```
s² = Σ(xᵢ-x̄)² / (n-1)
```

$E(s^2) = \sigma^2$ · Bessel's correction $(n-1)$

MLE Variance (biased)

$\tilde{s}^2 = \Sigma(x_i - \bar{x})^2 / n$

$E(\tilde{s}^2) = (n-1)/n \cdot \sigma^2$ · Biased but MLE

MSE Decomposition

MSE = Bias² + Variance

THE single most important formula in estimation

Consistency

$\hat{\theta} \to^p \theta$ as $n \to \infty$

More data = better estimate. Non-negotiable property.

MLE Recipe

$\hat{\theta} = \text{argmax}\ \Sigma \log f(x_i|\theta)$

Write log-likelihood → differentiate → set to 0 → solve

🧠 **ALL MNEMONICS RECAP**

P-P-S-S: Parameters → Populations, Statistics → Samples
DARTS: Data → Apply → Repeat → Times → Sampling distribution
Recipe vs Dish: Estimator = recipe, Estimate = the number
BVCE: Best Values Come Eventually (Bias, Variance, Consistency, Efficiency)
MSE = B²V: Mean Squared Error = Bias² + Variance
LMM: Learn My Models (Least Squares, Method of Moments, MLE)
MLE: Most Likely Explanation

🤖 Why This Matters in Machine Learning

Direct Connections

Linear Regression = Least Squares estimation of β coefficients

Logistic Regression = MLE for the

Bias-Variance Tradeoff = MSE decomposition applied to prediction error

Bernoulli/Binomial parameter

Regularization (Ridge, LASSO, Elastic Net) = Intentionally biased estimators that reduce variance → lower MSE

Neural Network Training = Finding MLE (or MAP) via gradient descent on the loss function

Cross-Validation = Estimating the sampling distribution of model performance

Your GeoGebra Visualization

The GeoGebra screenshot you shared shows points scattered around a circle defined by $(x - 12.28)^2 + (y + 0.96)^2 = 30$. This is a perfect visual of **estimation in action**:

The center (12.28, -0.96) and radius $\sqrt{30}$ are *parameters* of the circle. The plotted points (A, B, C, ..., R) are *sample data*. If you were estimating the circle's center from noisy data, you'd be doing parameter estimation — using the sample points to find $\hat{\theta}$ = (center_x, center_y, radius).