Parameter Estimation

Complete Visual Guide · Every Graph · Every Proof · Every Formula

10 Interactive Graphs | Full Derivations | Step-by-Step Proofs | 7 Mnemonics | ML Connections

01 Probability Distribution Foundations

CDF, PMF & PDF

CDF — CUMULATIVE DISTRIBUTION FUNCTION

$F(x) = P(X \leq x)$

Properties: $F(-\infty)=0$, $F(+\infty)=1$, non-decreasing, right-continuous

PMF (DISCRETE)

$f(x) = P(X = x)$

$\sum_x f(x) = 1$

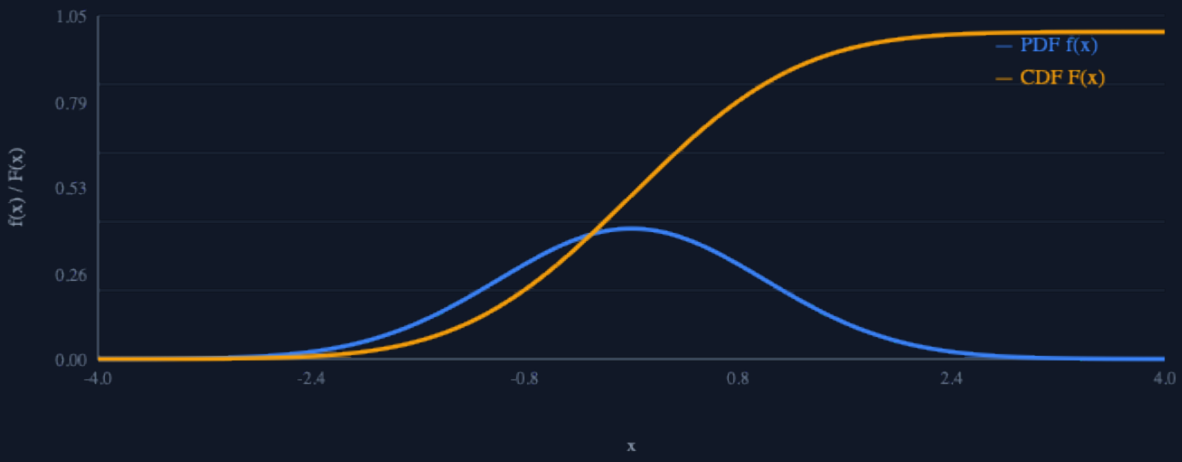
PDF (CONTINUOUS)

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

$\int_{-\infty}^{\infty} f(x)\,dx = 1$

$f(x) \geq 0$ but can exceed 1 (density!)

📊 Graph 1: Normal Distribution — PDF & CDF

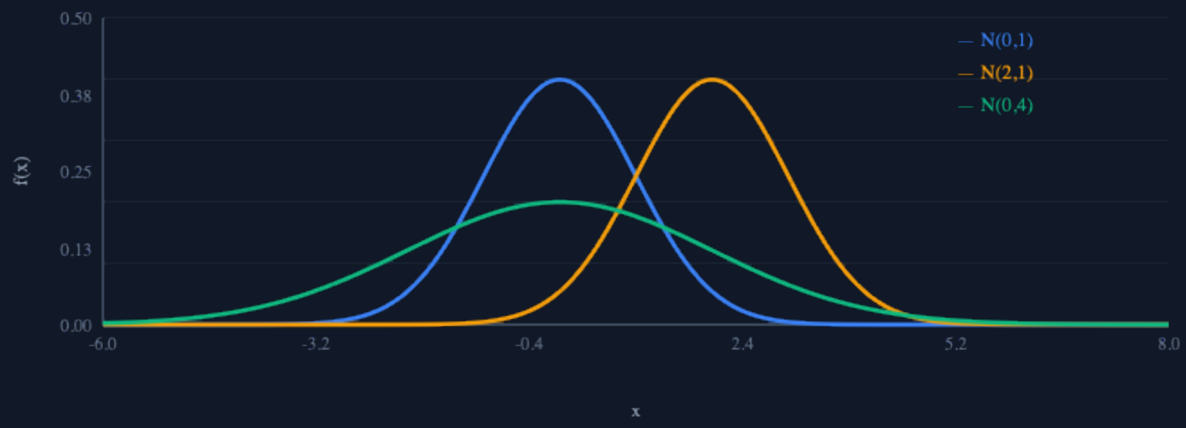


Blue = PDF f(x) bell curve · Orange = CDF F(x) cumulative probability · N(0,1)

Distribution Parameters Table

Distribution	Parameters	PDF/PMF	Mean	Variance
Normal N(μ, σ²)	μ, σ²	(1/√(2πσ²)) exp[−(x−μ)²/(2σ²)]	μ	σ²
Binomial B[N,p]	N, p	C(N,x) pˣ(1−p)^(N−x)	Np	Np(1−p)
Uniform U[a,b]	a, b	1/(b−a)	(a+b)/2	(b−a)²/12
Poisson(λ)	λ	e^(−λ) λˣ/x!	λ	λ
Exponential(λ)	λ	λe^(−λx)	1/λ	1/λ²

📊 Graph 2: How Parameters Change Distribution Shape

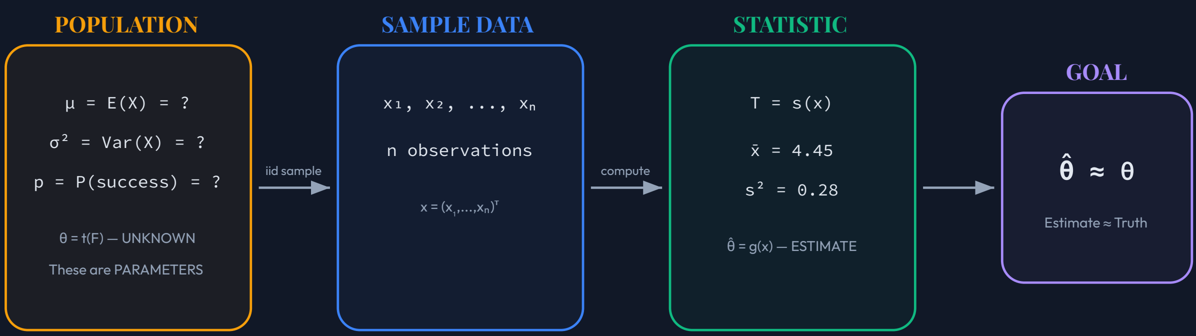


N(0,1) blue · N(2,1) orange — shifted mean · N(0,4) green — wider variance

02

Parameters vs Statistics

🎯 Visual: The Complete Pipeline



🧠 **MNEMONIC — P-P-S-S**

Parameters → Populations (fixed unknown) · Statistics → Samples (computed from data)

IID Assumption

INDEPENDENT & IDENTICALLY DISTRIBUTED

```
xᵢ ~iid~ F   for i = 1, ..., n

Independent:  E(xᵢ · xⱼ) = E(xᵢ) · E(xⱼ)   for i ≠ j
Identically:  E(xᵢ) = μ,  Var(xᵢ) = σ²      for ALL i
```

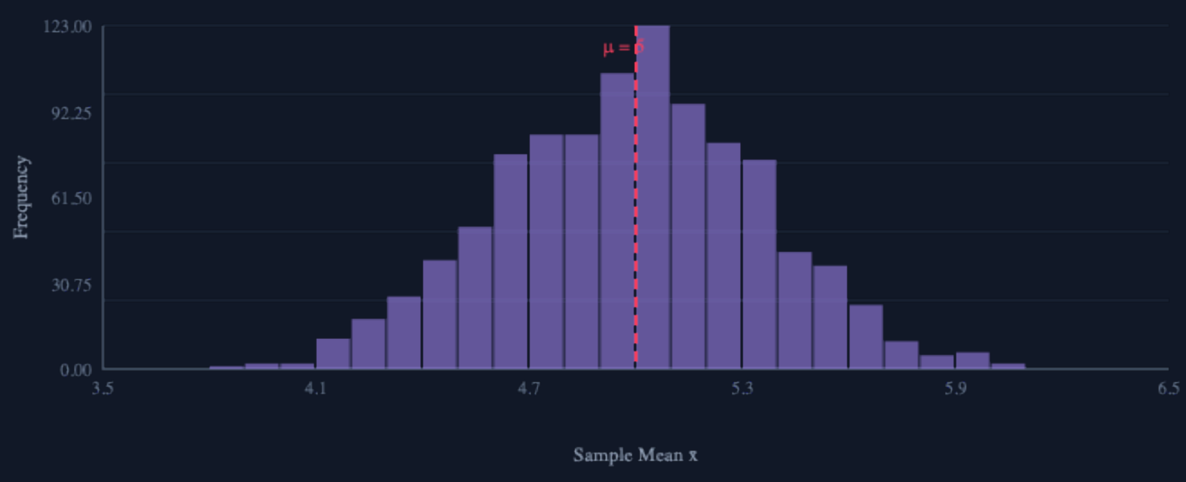
03

Sampling Distribution

Statistics Are Random Variables!

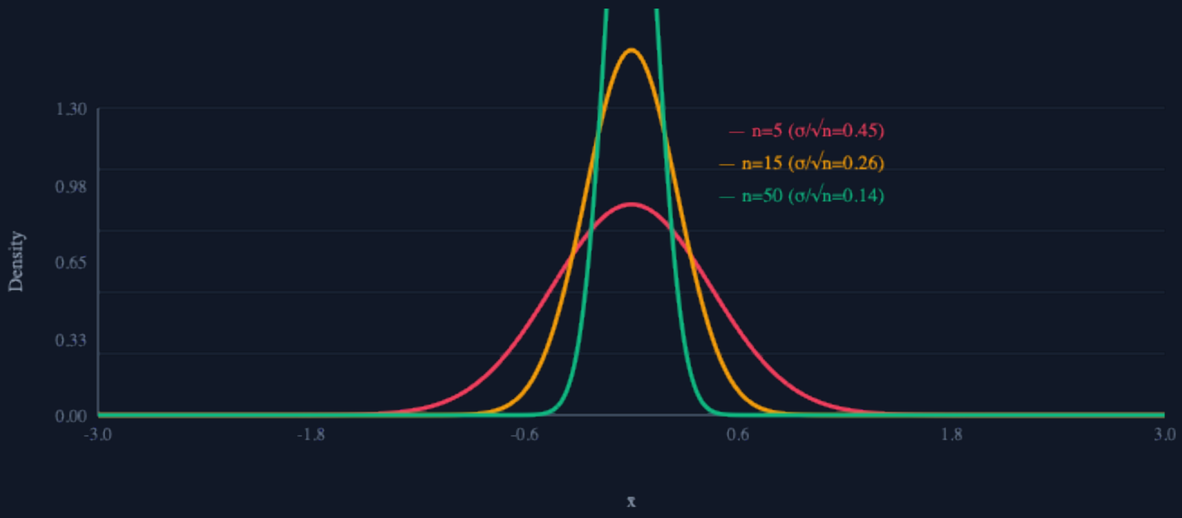
Since x is random, T = s(x) is random too. Draw different sample → get different T. The **sampling distribution** is the distribution of T across all possible samples.

📊 Graph 3: Sampling Distribution of the Mean (1000 Simulations)



1000 samples of n=30 from N(5,4). Means cluster around μ=5 with spread $\sigma/\sqrt{n} \approx 0.365$

📊 Graph 4: CLT — Variance Shrinks with Sample Size



Sampling dist of $\bar{x}$ for n=5 (wide red), n=15 (medium orange), n=50 (tight green). All at μ=0.

🧠 **MNEMONIC — DARTS**

Data → Apply statistic → Repeat many Times → Sampling distribution

04 Estimates & Estimators

Estimator (the recipe)

The **function** g(·) applied to data.

EXAMPLE

```
g(x) = (1/n) Σᵢ₌₁ⁿ xᵢ   ← waiting for data
```

Estimate (the dish)

The **number** you get with real data.

EXAMPLE

```
x = (4.2, 3.8, 5.1, 4.7)
θ̂ = g(x) = 17.8/4 = 4.45
← a number!
```

🧠 MNEMONIC — RECIPE VS DISH

Estimator = recipe (function) · Estimate = the actual dish (number)

⚠️ Two Variance Formulas

✅ UNBIASED S²

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

$E(s^2) = \sigma^2$ ← Bessel's correction

⚠️ MLE $\tilde{S}^2$

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

$E(\tilde{s}^2) = (n-1)/n \cdot \sigma^2$ ← biased down!

💡 WHY N−1?

Using $\bar{x}$ instead of μ "uses up" 1 degree of freedom. We have n data points but only n−1 independent deviations from $\bar{x}$ (since they sum to 0). Dividing by n−1 corrects this.

05 Quality of Estimators — All Proofs

① Bias · ② Variance · ③ MSE · ④ Consistency · ⑤ Efficiency

🧠 MNEMONIC — BVCE → "BEST VALUES COME EVENTUALLY"

Bias · Variance · Consistency · Efficiency

① Bias — Systematic Error

BIAS FORMULA

```
Bias(θ̂) = E(θ̂) − θ
Unbiased ⟺ E(θ̂) = θ ⟺ Bias = 0
```

PROOF: X̄ IS UNBIASED FOR Μ

1. $E(\bar{x}) = E[(1/n)\ \Sigma_i\ x_i]$
2. $= (1/n)\ \Sigma_i\ E(x_i)$ — linearity of expectation
3. $= (1/n)\ \Sigma_i\ \mu = (1/n)(n\mu)$ — identically distributed
4. $= \mu$ ∴ $\text{Bias} = \mu - \mu = 0$ ✓ **UNBIASED**

PROOF: S² IS UNBIASED FOR Σ² (FULL 3-PART DERIVATION)

1. **Key identity:** $\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2$

   Proof: $\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - 2\bar{x}\Sigma x_i + n\bar{x}^2 = \Sigma x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 = \Sigma x_i^2 - n\bar{x}^2$
2. **So:** $E(s^2) = 1/(n-1) \cdot [\Sigma\ E(x_i^2) - n \cdot E(\bar{x}^2)]$
3. **Find $E(x_i^2)$:** $\text{Var}(X) = E(X^2) - [E(X)]^2 \rightarrow E(x_i^2) = \sigma^2 + \mu^2$
4. **Find $E(\bar{x}^2)$:**

   $\bar{x}^2 = (1/n^2)[\Sigma x_i^2 + 2 \cdot \Sigma_{i>j}\ x_i x_j]$

   $E(\bar{x}^2) = (1/n^2)[n(\sigma^2+\mu^2) + 2 \cdot (n(n-1)/2) \cdot \mu^2]$

   $= (1/n^2)[n\sigma^2 + n\mu^2 + n(n-1)\mu^2] = (1/n^2)[n\sigma^2 + n^2\mu^2]$

   $= \sigma^2/n + \mu^2$
5. **Combine:**

   $E(s^2) = 1/(n-1) \cdot [n(\sigma^2+\mu^2) - n(\sigma^2/n + \mu^2)]$

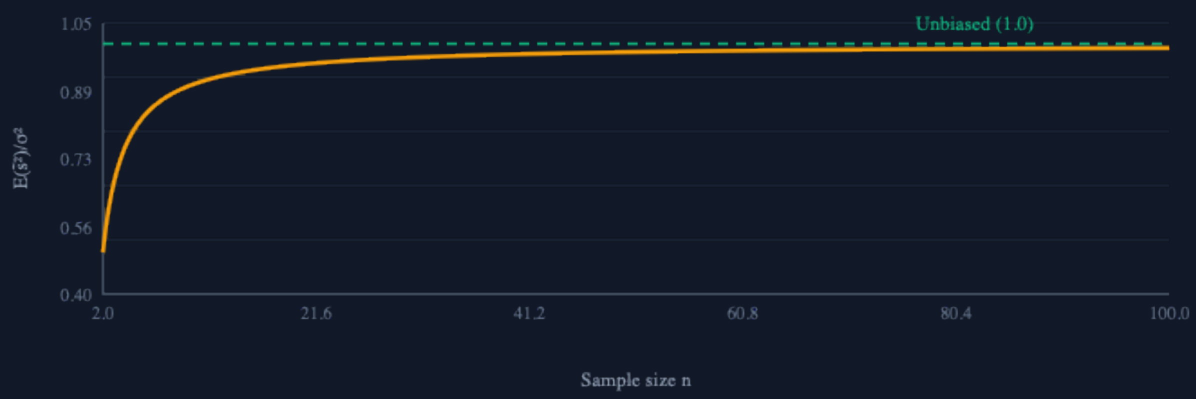
   $= 1/(n-1) \cdot [n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2]$

   $= 1/(n-1) \cdot (n-1)\sigma^2 = \sigma^2$ ✓ **UNBIASED**

◺ PROOF: $\tilde{s}^2$ IS BIASED

1. $\tilde{s}^2 = (n-1)/n \cdot s^2$
2. $E(\tilde{s}^2) = (n-1)/n \cdot E(s^2) = (n-1)/n \cdot \sigma^2$
3. Bias = $(n-1)/n \cdot \sigma^2 - \sigma^2 =$ **$-\sigma^2/n$ (BIASED DOWNWARD)**

📊 Graph 5: Bias of $\tilde{s}^2$ Vanishes as n Grows



$E(\tilde{s}^2)/\sigma^2 = (n-1)/n \to 1$. At n=5 underestimates by 20%, at n=100 only 1%.

---

② Variance & Standard Error

DEFINITIONS

```
Var(θ̂) = E[(θ̂ - E(θ̂))²]
SE(θ̂) = √Var(θ̂)
```

📐 **PROOF: VAR($\bar{X}$) = Σ²/N**

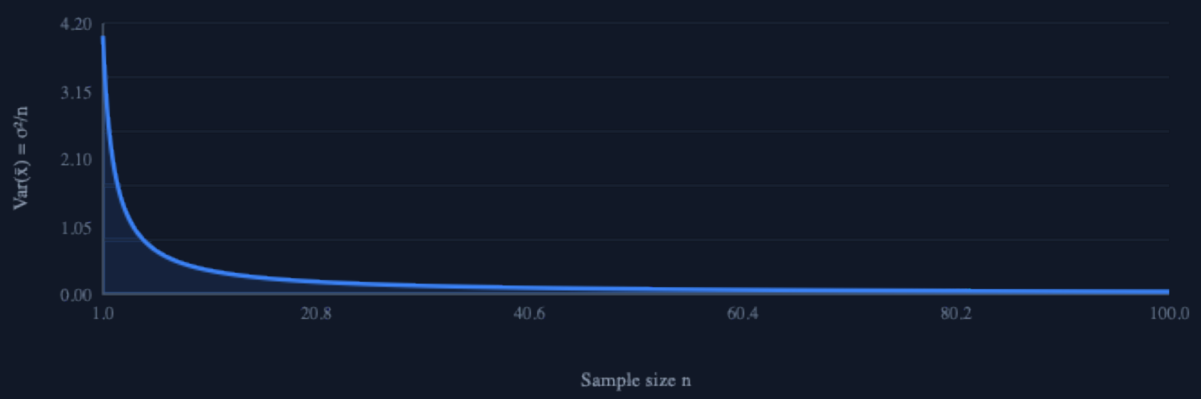
1. $\mathrm{Var}(\bar{x}) = \mathrm{Var}[(1/n)\,\Sigma x_i]$
2. $= (1/n^2)\,\mathrm{Var}(\Sigma x_i)$ — $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$
3. $= (1/n^2)\,\Sigma\,\mathrm{Var}(x_i)$ — independence: no covariance
4. $= (1/n^2) \cdot n\sigma^2 =$ **σ^2/n**

VARIANCE OF SAMPLE VARIANCE (ADVANCED)

```
Var(s²) = (1/n)(μ₄ - (n-3)/(n-1) · σ⁴)
where μ₄ = E[(X-μ)⁴] = 4th central moment

For Normal data: μ₄ = 3σ⁴ → Var(s²) = 2σ⁴/(n-1)
```

📊 Graph 6: Variance of x̄ Shrinks as n Increases



Var(x̄) = σ²/n (σ²=4). More data → more precise estimates!

③ MSE = Bias² + Variance — THE Gold Standard

MSE DECOMPOSITION

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

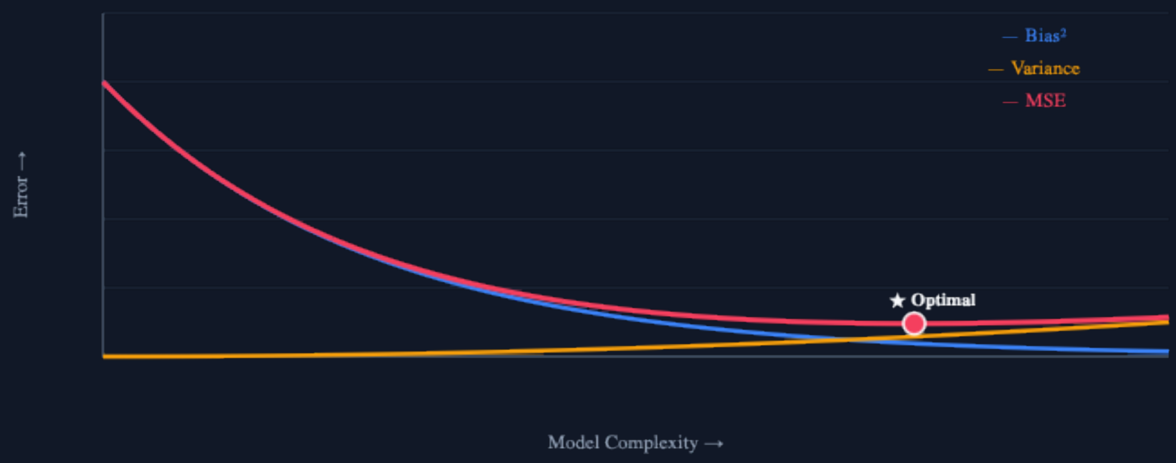
📐 PROOF: MSE = BIAS² + VARIANCE

1. **Expand MSE:** $E[(\hat{\theta}-\theta)^2] = E(\hat{\theta}^2) - 2\theta \cdot E(\hat{\theta}) + \theta^2$
2. **Bias²:** $[E(\hat{\theta})-\theta]^2 = E(\hat{\theta})^2 - 2\theta \cdot E(\hat{\theta}) + \theta^2$
3. **Var:** $E(\hat{\theta}^2) - E(\hat{\theta})^2$
4. **Add:** $Bias^2 + Var = E(\hat{\theta})^2 - 2\theta E(\hat{\theta}) + \theta^2 + E(\hat{\theta}^2) - E(\hat{\theta})^2$
   $= E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2 = MSE$ ✓ **Q.E.D.**

🧠 MNEMONIC — MSE = B²V

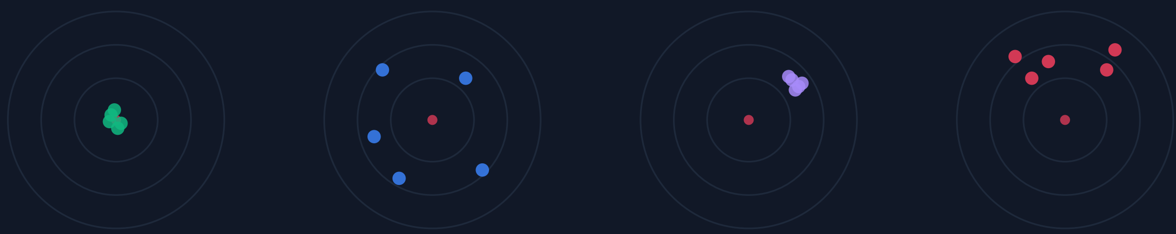
"My Squared Error = Bias² + Variance" — like E=mc² for estimators!

📊 Graph 7: Bias-Variance Tradeoff Curve



Blue = Bias² (decreases) · Orange = Variance (increases) · Red = MSE total. ★ = optimal complexity.

🎯 Bias-Variance Target Analogy



Low Bias + Low Var
⭐ IDEAL

Low Bias + High Var
Scattered around center

High Bias + Low Var
Tight but off-center

High Bias + High Var
✗ WORST

🔴 Center = true θ · Colored dots = estimates from different samples

---

④ Consistency

DEFINITION

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

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

⑤ Efficiency

DEFINITION

$\hat{\theta}_1$ more efficient than

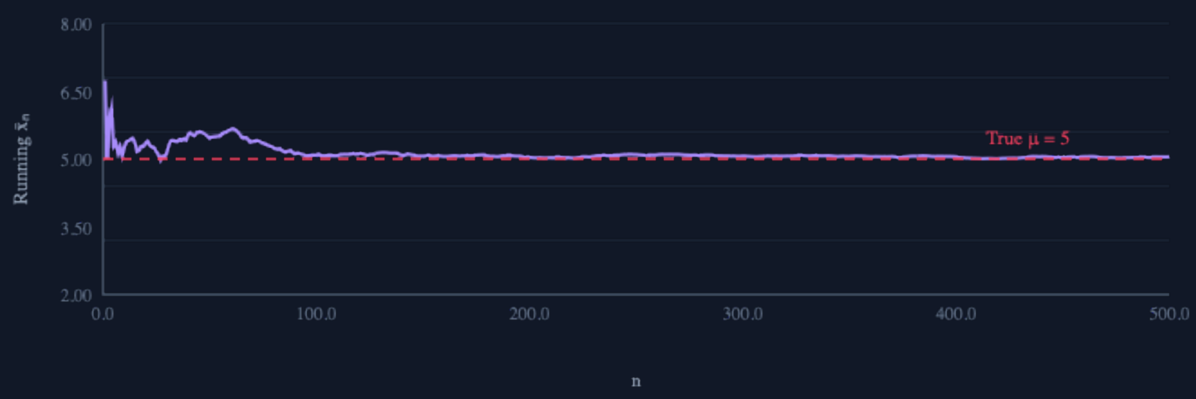
$\hat{\theta}_2$ if $MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)$

```
any ε > 0
"More data = closer to
truth"
```

$\bar{x}$, s^2, $\tilde{s}^2$ are ALL consistent.

```
If both unbiased:
compare variances
Cramér-Rao bound =
minimum possible
variance
```

📊 Graph 8: Consistency — Running Mean Converges



Running $\bar{x}_n$ as n grows from 1→500. True μ=5 (red dashed). Convergence = consistency!

Quality Properties Summary Table

Property	Formula	$\bar{x}$ for μ	s^2 for σ^2	$\tilde{s}^2$ for σ^2
Bias	$E(\hat{\theta}) - \theta$	0 ✓	0 ✓	$-\sigma^2/n$ ✗
Variance	$E[(\hat{\theta}-E(\hat{\theta}))^2]$	σ^2/n	complex	complex
MSE	Bias²+Var	σ^2/n	—	—
Consistent	$\hat{\theta}\to\theta$	Yes ✓	Yes ✓	Yes ✓

06

Estimation Frameworks — Full Derivations

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

Least Squares · Method of Moments · Maximum Likelihood

① Least Squares

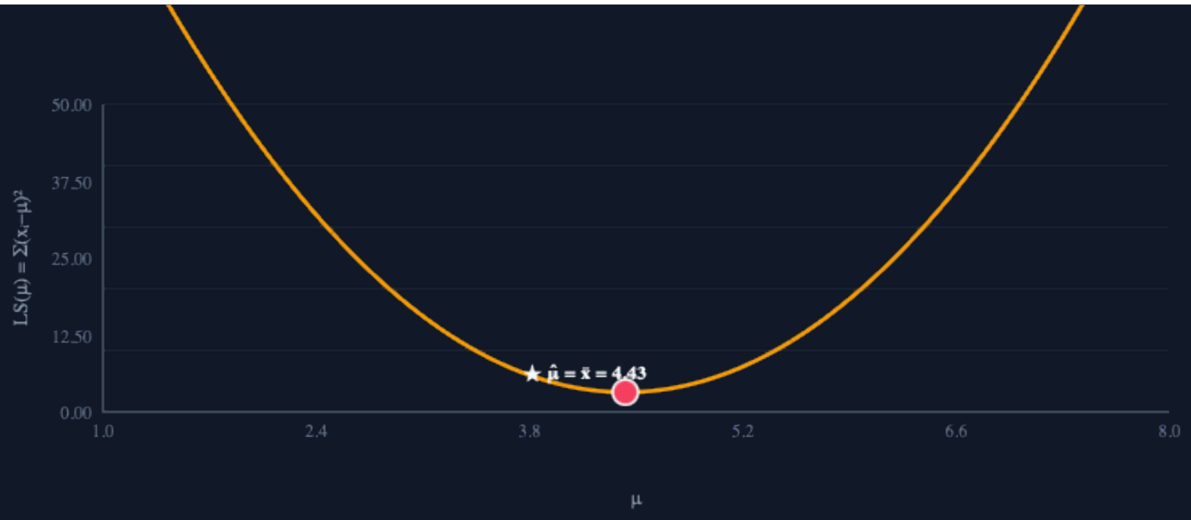
LOSS FUNCTION

$LS(\theta|x) = \Sigma_i (x_i - \theta)^2$

📐 **FULL DERIVATION: LS ESTIMATE OF M**

1. **Expand:** $LS(\mu) = \Sigma x_i^2 - 2\mu\Sigma x_i + n\mu^2$
2. **Differentiate:** $dLS/d\mu = -2\Sigma x_i + 2n\mu$
3. **Set = 0:** $2n\mu = 2\Sigma x_i$
4. **Solve:** $\hat{\mu} = (1/n)\Sigma x_i = \bar{x}$
5. **Verify min:** $d^2LS/d\mu^2 = 2n > 0$ ✓

📊 Graph 9: Least Squares Loss — Parabola



LS(μ) is a parabola. Minimum at ★ = $\bar{x}$ (sample mean).

---

② Method of Moments

RECIPE

1. Write population moments: $\mu_j = E(X^j) = m_j(\theta_1,...,\theta_p)$
2. Compute sample moments: $\hat{\mu}_j = (1/n)\Sigma x_i^j$
3. Set $\hat{\mu}_j = m_j(\hat{\theta}_1,...,\hat{\theta}_p)$
4. Solve for $\hat{\theta}_1,...,\hat{\theta}_p$

📐 MOM: NORMAL $N(M,\Sigma^2)$

1. $\mu_1 = \mu$, $\mu_2 = \mu^2 + \sigma^2$
2. $\hat{\mu}_1 = \bar{x}$, $\hat{\mu}_2 = \bar{x}^2 + \tilde{s}^2$
3. Set equal → $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \tilde{s}^2$

📐 **MOM: UNIFORM U[A,B]**

1. $\mu_1 = (a+b)/2$, $\mu_2 = (a^2+ab+b^2)/3$
2. From eq1: $b = 2\mu_1 - a$ → substitute into eq2
3. Get quadratic: $a^2 - 2\mu_1 a + (4\mu_1^2 - 3\mu_2) = 0$
4. Quadratic formula → $\hat{a} = \hat{\mu}_1 - \sqrt{3}\sqrt{(\hat{\mu}_2 - \hat{\mu}_1^2)}$, $\hat{b} = \hat{\mu}_1 + \sqrt{3}\sqrt{(\hat{\mu}_2 - \hat{\mu}_1^2)}$

---

③ Maximum Likelihood Estimation ⭐

CORE FRAMEWORK

Likelihood: $L(\theta|x) = \prod_i f(x_i|\theta)$

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

MLE: $\hat{\theta} = \text{argmax } \ell(\theta|x)$

WHY log? Products → sums, avoids underflow, same maximum

MLE 4-STEP RECIPE

Step 1: Write PDF/PMF $f(x_i|\theta)$

Step 2: Log-likelihood $\ell = \sum \log f(x_i|\theta)$

Step 3: Differentiate $d\ell/d\theta$

Step 4: Set = 0, solve for $\hat{\theta}$ (check $d^2\ell/d\theta^2 < 0$)

🧠 **MNEMONIC — MLE = MOST LIKELY EXPLANATION**

Which parameter values would have made this data *most likely*?

FULL MLE: NORMAL N(M,Σ²)

1. **PDF:** $f(x|\mu,\sigma^2) = (2\pi\sigma^2)^{(-\frac{1}{2})} \cdot \exp[-(x-\mu)^2/(2\sigma^2)]$
2. **Log-lik:** $\ell = -(1/2\sigma^2)\Sigma(x_i-\mu)^2 - (n/2)\log(\sigma^2) - (n/2)\log(2\pi)$
3. **$\partial\ell/\partial\mu = 0$:** $(1/\sigma^2)[\Sigma x_i - n\mu] = 0 \rightarrow$ **$\hat{\mu} = \bar{x}$**
4. **$\partial\ell/\partial\sigma^2 = 0$:** $(1/2\sigma^4)\Sigma(x_i-\bar{x})^2 - n/(2\sigma^2) = 0$
   $\rightarrow \Sigma(x_i-\bar{x})^2 = n\sigma^2 \rightarrow$ **$\hat{\sigma}^2 = (1/n)\Sigma(x_i-\bar{x})^2 = \tilde{s}^2$**

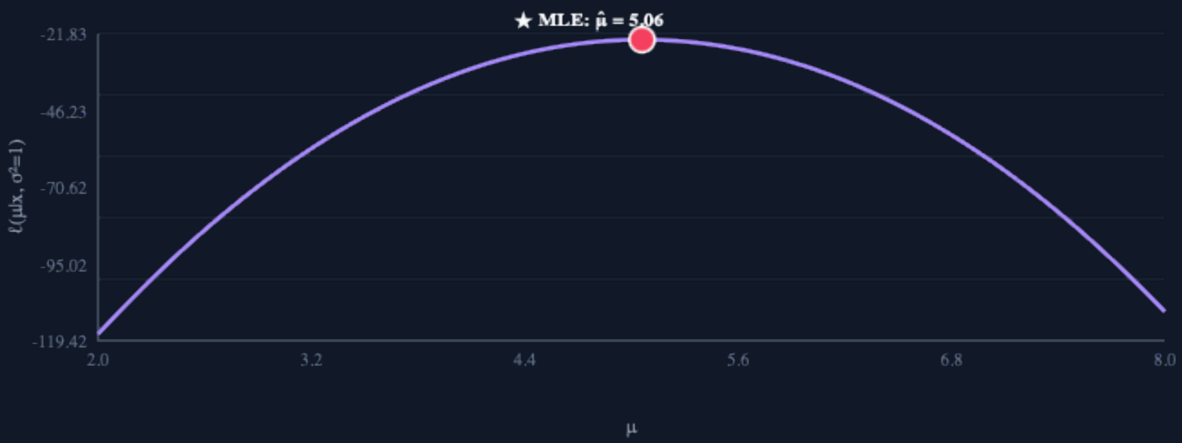
FULL MLE: BINOMIAL B[N,P]

1. PMF: $f(x|N,p) = C(N,x)p^x(1-p)^{(N-x)}$
2. $\ell = \log(p)\Sigma x_i + \log(1-p)(nN-\Sigma x_i) + \text{const}$
3. $d\ell/dp = (\Sigma x_i)/p - (nN-\Sigma x_i)/(1-p) = 0$
4. Multiply by $p(1-p)$: $(1-p)n\bar{x} - pn(N-\bar{x}) = 0 \rightarrow n\bar{x} - pnN = 0$
5. $\hat{p} = \bar{x}/N$

FULL MLE: UNIFORM U[A,B]

1. PDF: $f(x|a,b) = 1/(b-a)$ for $a \le x \le b$
2. $\ell = -n \cdot \log(b-a)$, subject to $a \le$ all $x_i \le b$
3. Maximize ℓ = minimize $(b-a)$ = smallest interval containing all data
4. **$\hat{a} = \min(x_i)$, $\hat{b} = \max(x_i)$**

📊 Graph 10: Log-Likelihood for Normal Distribution



ℓ (μ) for data from N(5,1). Peak ★ at $\hat{\mu} = \bar{x}$ = MLE.

⭐ **MLE SUPERPOWER PROPERTIES**

1. Consistent: $\hat{\theta} \to \theta$ as $n \to \infty$ · **2. Asymptotically Efficient:** lowest variance for large n · **3. Functionally Invariant:** $h(\hat{\theta})$ is MLE of $h(\theta)$

Framework Comparison Table

Aspect	Least Squares	MoM	MLE
Core idea	Min $\Sigma(x_i-\theta)^2$	Match moments	Max $\Pi f(x_i\|\theta)$
Requires	Loss function	Moment equations	Full PDF/PMF
Efficient?	Sometimes	Usually not	Yes (asymptotic)
Best for	Means, regression	Quick estimates	Everything ⭐
Normal $\hat{\mu}$	$\bar{x}$	$\bar{x}$	$\bar{x}$
Normal $\hat{\sigma}^2$	N/A	$\tilde{s}^2$ (÷n)	$\tilde{s}^2$ (÷n)
Uniform	N/A	$\bar{x} \pm \sqrt{3}\cdot\tilde{s}$	min/max(x_i)

All Mathematical Rules Used

EXPECTATION RULES

E(c) = c

E(cX) = c·E(X)

E(X+Y) = E(X) + E(Y) ← ALWAYS

E(XY) = E(X)·E(Y) ← only if independent!

$E(X^2) = Var(X) + [E(X)]^2$

VARIANCE RULES

Var(c) = 0

$Var(cX) = c^2 \cdot Var(X)$ ← note c^2!

Var(X+c) = Var(X)

Var(X+Y) = Var(X)+Var(Y) ← if independent

$Var(X) = E(X^2) - [E(X)]^2$

CALCULUS RULES

$d/dx\ [\Sigma(x_i-\mu)^2] = -2\Sigma(x_i-\mu)$

d/dx [log(x)] = 1/x

$d/dx\ [x^n] = nx^{n-1}$

$d/dx\ [e^x] = e^x$

Optimization: f'(x)=0, check f''(x)

LOGARITHM RULES

log(AB) = log(A) + log(B) ← product→sum

log(A/B) = log(A) − log(B)

$\log(A^n) = n \cdot \log(A)$

$\log(e^x) = x$

In MLE: $\log[\Pi f(x_i|\theta)] = \Sigma \log[f(x_i|\theta)]$

SUMMATION IDENTITIES

$\Sigma_i\ c = nc$ $\Sigma(x_i-\bar{x}) = 0$ (deviations sum to zero!)

$\Sigma(x_i-\bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2$ (computational shortcut)

$\Sigma_{i=2}^{n} \Sigma_{j=1}^{i-1}\ c = n(n-1)/2 \cdot c$ (counting pairs)

🤖 Machine Learning Connections

ML Algorithm	Estimation Method	What's Estimated
Linear Regression	Least Squares	β coefficients
Logistic Regression	MLE	Log-odds β
Naive Bayes	MLE	Class priors, likelihoods
Ridge/LASSO	Penalized LS/MLE	β (intentionally biased → lower MSE)
Gaussian Mixtures	MLE via EM	μ_k, σ_k^2, π_k per cluster
Neural Networks	MLE via SGD	Weights & biases
Cross-Validation	Sampling distribution	Generalization error

🧠 **ALL 7 MNEMONICS RECAP**

P-P-S-S: Parameters→Populations, Statistics→Samples

DARTS: Data→Apply→Repeat→Times→Sampling dist.

Recipe vs Dish: Estimator=recipe, Estimate=number

BVCE: Best Values Come Eventually

MSE=B²V: Bias²+Variance

LMM: Learn My Models (LS, MoM, MLE)

MLE: Most Likely Explanation