Module 3 — Point Estimation

Statistics

hbar.university

Estimators, Risk, and Mean Squared Error

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.1

Formal Definition

Let $X_1, \ldots, X_n$ be data generated under a model $\{\mathbb{P}_\theta : \theta \in \Theta\}$.

An **estimator** of a parameter θ (or a function $g(\theta)$) is a statistic:

$$\hat{\theta} = T(X_1, \ldots, X_n).$$

A **loss function** quantifies error of an estimate a when the true parameter is θ:

$$L(\theta, a) \geq 0.$$

A **decision rule** (estimator is a special case) is a map $\delta(X_1, \ldots, X_n) \in \mathcal{A}$.

The **risk function** is the expected loss:

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X_1, \ldots, X_n))]\,.$$

Intuition

An estimator is a rule that turns data into a guess.

Risk is "how bad this rule is on average" if the world is θ.

Statistics is not just computing $\hat{\theta}$; it is designing rules with controlled risk.

Structural Role

Risk is the evaluation metric that makes estimation a well-posed optimization problem.

Everything that follows—bias/variance, optimal estimators, Cramér–Rao bounds—talks about controlling or lower-bounding risk.

This node installs the objective function for inference.

Mathematical Development

Squared Error Loss and MSE

For estimating a scalar $g(\theta)$, a canonical loss is squared error:

$$L(\theta, a) = (a - g(\theta))^2.$$

Then the risk is the **mean squared error (MSE)**:

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta\left[(\hat{\theta} - g(\theta))^2\right].$$

Bias–Variance Decomposition

Let $b_\theta = \mathbb{E}_\theta[\hat{\theta}] - g(\theta)$ be the **bias**. Then:

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) + b_\theta^2.$$

This decomposition is the first fundamental tradeoff in estimation.

Worked Example

Estimating a Bernoulli Mean

Let $X_i \sim \mathrm{Bernoulli}(p)$, $g(\theta) = p$, and estimator $\hat{p} = \bar{X}_n$.

Then:

$$\mathbb{E}[\hat{p}] = p \quad \text{(unbiased)},$$

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n},$$

so:

$$\mathrm{MSE}(\hat{p}) = \frac{p(1-p)}{n}.$$

Cross-Links

- **AI:** Expected loss over a data distribution is risk; ERM approximates it with empirical risk.
- **Economics:** Decision theory and expected utility are risk frameworks with different loss functions.
- **Linear Algebra:** Variance/covariance are quadratic forms; risk often becomes a quadratic objective.
- **Quantum:** Estimation risk bounds connect to (quantum) Fisher information.

Structural Compression

Invariant:
Performance of an estimator is defined by expected loss under $\mathbb{P}_\theta$.

Mechanism:
Choose a loss → take expectation → obtain risk as the optimization target.

Cross-System Echo:
In ML, "training" is minimizing an empirical proxy for risk.

Exercises

1. Prove the bias–variance decomposition for squared error loss.
2. For $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2, compute the MSE of $\bar{X}_n$ as an estimator of μ.
3. Give an example of a loss function other than squared error (e.g., absolute loss) and define the corresponding risk.

Method of Moments and Maximum Likelihood Estimation (MLE)

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.2

Formal Definition

Let $X_1, \ldots, X_n$ be data generated under a model $\{f_\theta : \theta \in \Theta\}$.

Method of Moments (MoM)

Choose k population moments $m_j(\theta) = \mathbb{E}_\theta[X^j]$ (or other moment-type quantities), and set them equal to the corresponding sample moments:

$$\hat{m}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

Solve the system:

$$m_j(\theta) = \hat{m}_j, \quad j = 1, \ldots, k$$

for θ. The solution (if it exists) is the MoM estimator $\hat{\theta}_{\text{MoM}}$.

Maximum Likelihood Estimation (MLE)

Define the likelihood:

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

and log-likelihood:

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

An MLE is any maximizer:

$$\hat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Theta} \ell(\theta).$$

Equivalently (when differentiable), solve the score equation:

$$\nabla_\theta \ell(\theta) = 0$$

and select maxima.

Intuition

MoM: "Match what the model predicts on average to what the data shows on average."
Fast, direct, often algebraic.

MLE: "Choose the parameter under which the observed data would be most plausible."
Optimization-based, often statistically efficient, and the central object of likelihood theory.

Structural Role

These are the two canonical construction principles for estimators:

- MoM provides quick, often closed-form estimators and a gateway to estimating equations.
- MLE is the backbone of classical inference (Wald/Score/LR tests, asymptotics, efficiency, Fisher information).

Understanding their contrast sets up: - consistency and asymptotic normality (Domain 7), - Fisher-information bounds (Node 3.4/3.5), - computational optimization (Domain 9).

Mathematical Development

MoM Example Pattern

If the model implies $\mathbb{E}_\theta[X] = g(\theta)$, then MoM sets:

$$g(\hat{\theta}_{\text{MoM}}) = \bar{X}_n.$$

If g is invertible:

$$\hat{\theta}_{\text{MoM}} = g^{-1}(\bar{X}_n).$$

MLE Example Pattern

MLE uses curvature of $\ell(\theta)$: - gradient (score) gives stationary points, - Hessian indicates maxima/minima, - Fisher information predicts asymptotic variance.

Practical Comparison (Canonical)

- **MoM**: simple, may be less efficient, depends on which moments you pick.
- **MLE**: often efficient under regularity, invariant under reparameterization, but may require numerical optimization and can be sensitive to model misspecification.

Worked Examples

Example 1: Bernoulli(p)

MoM using first moment:

$$\mathbb{E}[X] = p, \quad \hat{m}_1 = \bar{X}_n \Rightarrow \hat{p}_{\text{MoM}} = \bar{X}_n.$$

MLE:

$$\ell(p) = \sum_{i=1}^{n} \left(x_i \log p + (1 - x_i) \log(1 - p) \right).$$

Setting derivative to zero yields:

$$\hat{p}_{\text{MLE}} = \bar{X}_n.$$

Here MoM = MLE.

Example 2: Exponential(λ)

Density: $f_\lambda(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

MoM:

$$\mathbb{E}[X] = \frac{1}{\lambda} \Rightarrow \hat{\lambda}_{\text{MoM}} = \frac{1}{\bar{X}_n}.$$

MLE:

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i \Rightarrow \hat{\lambda}_{\text{MLE}} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}_n}.$$

Again MoM = MLE.

(They diverge in many multi-parameter or constrained models.)

Cross-Links

- **AI:** MLE corresponds to minimizing negative log-likelihood; MoM resembles matching feature statistics (e.g., method-of-moments in latent variable models).
- **Economics:** Structural estimation often uses MLE; MoM generalizes to GMM (bridge to econometrics).
- **Linear Algebra:** MoM often reduces to solving linear systems; MLE uses gradients/Hessians (quadratic forms).
- **Quantum:** Tomography can be posed as MLE; moment-matching appears via expectation-value constraints.

Structural Compression

Invariant:
Both methods define estimators by enforcing model–data agreement.

Mechanism:
MoM matches expectations; MLE maximizes plausibility via likelihood curvature.

Cross-System Echo:
In control systems: match moments (constraints) vs optimize a global objective.

Exercises

1. Compute MoM and MLE for Poisson(λ) and compare.
2. Give an example where MoM and MLE differ (hint: use a multi-parameter model or a constrained parameter space).
3. Explain why maximizing $\ell(\theta)$ is equivalent to maximizing $L(\theta)$.
4. For Normal(μ, σ^2), derive the MLEs for μ and σ^2.

Properties of Estimators: Bias, Consistency, and Efficiency

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.3

Formal Definitions

Let $\hat{\theta}_n$ be an estimator of θ.

1. Bias

$$\text{Bias}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta[\hat{\theta}_n] - \theta.$$

The estimator is **unbiased** if this equals 0 for all θ.

---

2. Consistency

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{as } n \to \infty.$$

That is, the estimator converges in probability to the true parameter.

(Strong consistency replaces convergence in probability with almost sure convergence.)

---

3. Efficiency (Introductory Notion)

Among unbiased estimators, an estimator is **more efficient** if it has smaller variance.

In asymptotic settings:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta)).$$

An estimator is asymptotically efficient if it achieves the minimal possible variance (to be formalized via Cramér–Rao bound).

Intuition

Bias measures systematic error.
Consistency measures long-run correctness.
Efficiency measures precision.

These are orthogonal dimensions:

- An estimator may be biased but consistent.
- It may be unbiased but inefficient.

- It may be consistent but converge slowly.

Inference requires balancing all three.

Structural Role

This node defines evaluation criteria for estimators.

It prepares:

- Cramér–Rao lower bound (finite-sample efficiency)
- Asymptotic normality
- Optimality theory
- Decision-theoretic comparisons

Without these properties, estimators are just formulas.

Mathematical Development

Bias–Variance Perspective

Recall:

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

Thus:

- Unbiasedness controls the second term.
- Efficiency controls the first term.

Consistency via Law of Large Numbers

If:

$$\hat{\theta}_n = g(\bar{X}_n),$$

and:

$$\bar{X}_n \xrightarrow{P} \mu(\theta),$$

then by continuous mapping theorem:

$$\hat{\theta}_n \xrightarrow{P} g(\mu(\theta)).$$

Thus, many estimators are consistent via LLN.

Worked Examples

Example 1: Sample Mean

If $X_i \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{E}[\bar{X}_n] = \mu,$$

$$\mathrm{Var}_\theta(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Thus:

- Unbiased
- Consistent (variance $\to 0$)
- Variance decreases at rate $1/n$

Example 2: Biased but Consistent Estimator

Let:

$$\tilde{\theta}_n = \frac{n-1}{n}\bar{X}_n.$$

Bias:

$$\mathbb{E}[\tilde{\theta}_n] - \mu = -\frac{1}{n}\mu.$$

Bias $\to 0$ as $n \to \infty$.

Thus biased but consistent.

Cross-Links

- **AI:** Bias–variance tradeoff mirrors underfitting vs overfitting.
- **Economics:** Consistency crucial for structural parameter estimation.
- **Linear Algebra:** Variance often expressible as quadratic forms.
- **Quantum:** Efficiency bounds tied to (quantum) Fisher information.

Structural Compression

Invariant:
Estimators are evaluated by expectation and variance under the model.

Mechanism:
Long-run convergence (LLN/CLT) governs consistency and asymptotics.

Cross-System Echo:
In optimization, convergence guarantees mirror statistical consistency.

Exercises

1. Prove that if $\mathrm{Var}(\hat{\theta}_n) \to 0$ and estimator is unbiased, then it is consistent.
2. Give an example of a consistent but inefficient estimator.
3. Show that MLE for Bernoulli is unbiased and consistent.
4. Explain difference between finite-sample efficiency and asymptotic efficiency.

The Cramér–Rao Lower Bound

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.4

Regularity Conditions (Important)

Assume:

1. The model $f_\theta(x)$ is differentiable in θ.
2. Differentiation under the integral sign is valid.
3. The support of f_θ does not depend on θ.

These are required for the identities below.

Theorem (Cramér–Rao Inequality)

Let $\hat{\theta}$ be an unbiased estimator of θ.

Then:

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{\mathcal{I}_n(\theta)},$$

where:

$$\mathcal{I}_n(\theta) = n\mathcal{I}(\theta) = n\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right].$$

Interpretation

No unbiased estimator can have variance smaller than the reciprocal of Fisher information.

Fisher Information sets a fundamental precision limit.

More curvature → smaller lower bound → more precise estimation possible.

Proof Sketch (Core Geometry)

Let:

$$S(\theta) = \frac{\partial}{\partial\theta}\log L(\theta).$$

Then:

$$\mathbb{E}_\theta[S(\theta)] = 0.$$

Differentiate unbiasedness:

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

Differentiating both sides w.r.t. θ:

$$\frac{d}{d\theta}\mathbb{E}_\theta[\hat{\theta}] = 1.$$

Using interchange of derivative and expectation:

$$\mathbb{E}_\theta\left[\hat{\theta}S(\theta)\right] = 1.$$

Now apply Cauchy–Schwarz:

$$1^2 \leq \mathrm{Var}_\theta(\hat{\theta}) \cdot \mathrm{Var}_\theta(S(\theta)).$$

But:

$$\mathrm{Var}_\theta(S(\theta)) = \mathcal{I}_n(\theta).$$

Thus:

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{\mathcal{I}_n(\theta)}.$$

Equality Condition

Equality holds iff:

$$\hat{\theta} = \theta + c(\theta)S(\theta).$$

In exponential families, the MLE often achieves the bound.

Example: Normal Mean (Known Variance)

If $X_i \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathcal{I}_n(\mu) = \frac{n}{\sigma^2}.$$

Thus:

$$\mathrm{Var}(\hat{\mu}) \geq \frac{\sigma^2}{n}.$$

The sample mean satisfies:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

It achieves the bound → efficient.

Cross-Links

- **Linear Algebra:** Bound derived from Cauchy–Schwarz in L^2 space.
- **AI:** Fisher Information underlies natural gradient methods.
- **Economics:** Efficiency bounds in structural estimation.
- **Quantum:** Quantum Cramér–Rao bound generalizes this inequality.

Structural Compression

Invariant:
Variance $\geq$ inverse information.

Mechanism:
Geometry of score function via Cauchy–Schwarz.

Cross-System Echo:
In physics, uncertainty principles bound precision.

Exercises

1. Prove the unbiasedness differentiation step carefully.
2. Compute Fisher information for Poisson and derive CRLB.
3. Show sample mean achieves CRLB for Normal mean.
4. Explain why biased estimators can beat the bound.

Sufficiency, Rao–Blackwell, and Lehmann–Scheffé

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.5

Core Definitions

Statistic

A **statistic** is any measurable function of the data:

$$T = T(X_1, \ldots, X_n).$$

Sufficiency (Factorization Theorem)

A statistic T is **sufficient** for θ if the likelihood can be factored as:

$$f_\theta(x_1, \ldots, x_n) = g_\theta(T(x_1, \ldots, x_n))\, h(x_1, \ldots, x_n),$$

where h does not depend on θ.

Equivalently: conditional distribution of the data given T does not depend on θ.

Intuition

A sufficient statistic is a *lossless compression of the data* for the purpose of estimating θ.

Once you know T, the remaining details of the sample contain no additional information about θ.

This is why exponential families matter: they often admit low-dimensional sufficient statistics.

Structural Role

This node gives you the canonical pipeline:

1. Find a sufficient statistic T.
2. Improve any unbiased estimator by conditioning on T (Rao–Blackwell).
3. If T is complete, the improved estimator is uniquely optimal (Lehmann–Scheffé).

This is "optimality by structure," not by brute-force search.

Theorem 1: Rao–Blackwell

Let $\hat{\theta}$ be an estimator of $g(\theta)$ with finite variance, and let T be sufficient for θ.

Define the Rao–Blackwellized estimator:

$$\hat{\theta}^\star = \mathbb{E}_\theta[\hat{\theta} \mid T].$$

Then:

1. If $\hat{\theta}$ is unbiased for $g(\theta)$, so is $\hat{\theta}^\star$.
2. Variance improves (or stays equal):

$$\mathrm{Var}_\theta(\hat{\theta}^\star) \leq \mathrm{Var}_\theta(\hat{\theta}).$$

Theorem 2: Completeness and Lehmann–Scheffé

A statistic T is **complete** if:

$$\mathbb{E}_\theta[a(T)] = 0 \text{ for all } \theta \quad \Rightarrow \quad \mathbb{P}(a(T) = 0) = 1.$$

Lehmann–Scheffé Theorem

If:

- T is sufficient and complete for θ,
- $\hat{\theta}$ is any unbiased estimator of $g(\theta)$,

then:

$$\hat{\theta}^\star = \mathbb{E}[\hat{\theta} \mid T]$$

is the **unique** uniformly minimum-variance unbiased estimator (UMVU) of $g(\theta)$.

Worked Example: Bernoulli(p)

If $X_i \sim \text{Bernoulli}(p)$,

the likelihood is:

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

Thus:

$$T = \sum_{i=1}^{n} X_i$$

is sufficient for p.

An unbiased estimator of p is $\bar{X}$, which is already a function of T, so Rao–Blackwell does not change it.

This illustrates a key point:

Once you are already using a sufficient statistic, there is no "information left" to gain.

Cross-Links

- **AI:** Sufficient representations = no loss of task-relevant information (representation learning analogue).
- **Linear Algebra:** Conditioning is projection in L^2 (variance reduction via orthogonal projection).
- **Economics:** Sufficient statistics appear in exponential family and discrete choice models.
- **Quantum:** Informationally complete measurements vs sufficient summaries for parameter estimation.

Structural Compression

Invariant:
Conditioning on sufficient information cannot increase variance.

Mechanism:
Rao–Blackwell is variance reduction via conditional expectation (projection).

Cross-System Echo:
In ML, better representations reduce sample complexity and variance.

Exercises

1. Prove the variance reduction statement using the law of total variance.
2. Show that if $\hat{\theta}$ is a function of T, then Rao–Blackwell leaves it unchanged.
3. For $X_i \sim \text{Poisson}(\lambda)$, show that $T = \sum X_i$ is sufficient for λ.
4. Explain (conceptually) why completeness yields uniqueness in Lehmann–Scheffé.

Invariance and Equivariance in Estimation

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.6

Core Principle

Estimation procedures should behave naturally under transformations of the parameter.

If we reparameterize the model, the estimator should transform accordingly.

This is called **invariance** (or more precisely, equivariance).

Definition: Invariance of the MLE

Let $\hat{\theta}_{\text{MLE}}$ be the MLE of θ.

If $\phi = g(\theta)$ is a one-to-one transformation, then:

$$\hat{\phi}_{\text{MLE}} = g(\hat{\theta}_{\text{MLE}}).$$

That is:

> The MLE of a transformed parameter equals the transformation of the MLE.

This is called the **invariance property of the MLE**.

Why It Holds

The likelihood for ϕ is:

$$L_\phi(\phi) = L_\theta(g^{-1}(\phi)).$$

Maximizing over ϕ is equivalent to maximizing over θ.

Thus the argmax transforms accordingly.

No recalculation needed.

Intuition

If you estimate:

- variance σ^2,
- then want standard deviation σ,

you do not re-estimate. You transform the estimate.

MLE respects the structure of the parameter space.

Other estimators (e.g., unbiased estimators) generally do **not** satisfy this property.

Structural Role

This explains why MLE is "natural":

- It respects reparameterization.
- It behaves coherently under model transformations.
- It is defined through the likelihood geometry rather than moment constraints.

This prepares for:

- Asymptotic normality
- Information geometry
- Delta method
- Likelihood-based testing

Worked Example

Example: Normal Variance

Suppose $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with known μ.

MLE for variance:

$$\hat{\sigma}^2 = \frac{1}{n}\sum (X_i - \mu)^2.$$

Now consider $\sigma = \sqrt{\sigma^2}$.

By invariance:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}.$$

No separate optimization needed.

Contrast: Unbiased Estimators

Unbiased estimators do **not** generally preserve invariance.

Example:

If $\hat{\theta}$ is unbiased for θ,

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

it does **not** imply:

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

Thus unbiasedness is not invariant under nonlinear transformations.

This is one reason unbiasedness is not always the dominant principle in modern inference.

Equivariance Under Group Transformations

More generally, if a model is invariant under a transformation group G, estimators can be required to satisfy:

$$\delta(g \cdot x) = g \cdot \delta(x).$$

Such estimators are called **equivariant**.

This becomes important in:

- Location-scale families
- Decision theory
- Invariant testing

Cross-Links

- **Linear Algebra:** Coordinate changes preserve geometric objects.
- **AI:** Parameterization choice affects optimization but not the optimum distribution.
- **Economics:** Reparameterizations in structural models.
- **Quantum:** Different parameterizations of states must yield consistent estimators.

Structural Compression

Invariant:
MLE commutes with smooth one-to-one transformations.

Mechanism:
Argmax of likelihood preserved under reparameterization.

Cross-System Echo:
In geometry, intrinsic properties do not depend on coordinates.

Exercises

1. Prove the invariance property formally.
2. Give an example where unbiasedness fails under transformation.
3. Show that MLE for $\log \theta$ equals $\log(\hat{\theta}_{\text{MLE}})$.
4. Consider a location family $f(x - \theta)$. What form must an equivariant estimator take?