Estimators, Risk, and Mean Squared Error

Statistics · Module 3 — Point Estimation

Node 3.1

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.

Statistics · Module 3 — Point Estimation · Node 3.1
Concepts: expectation-value, cost-function, variance-and-covariance
Prerequisites: 2.1, 2.2, 1.5, 1.6
Enables: 3.2, 3.3

hbar.university — own.your.cognition