Lecture 3: Properties of Estimators

Bias · Variance · MSE · Consistency · Sufficiency · Cramér–Rao

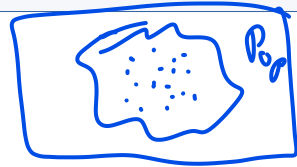
We use estimators every day. Are they any good?

We already use estimators (Lecture 1, plug-in principle):

$\bar{X}$ for μ, $\quad S^2$ for σ^2, $\quad \hat{p} = \frac{\text{count}}{n}$ for p

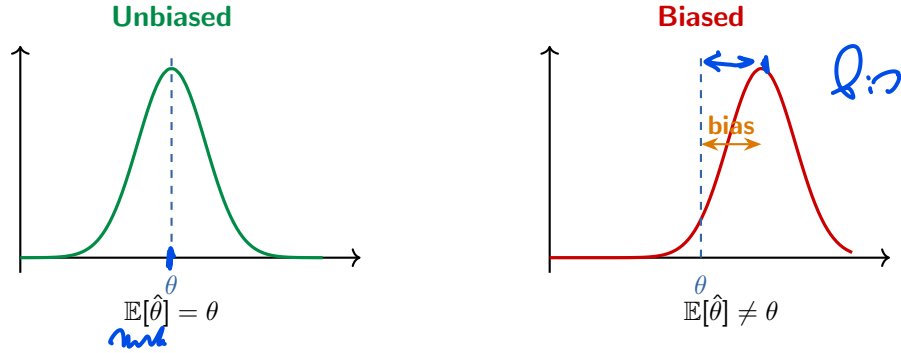
But how do we **judge** an estimator?

Is it close to the truth? How much does it jump around? Can we do better?



Bias: Is the Estimator Centered on the Truth?

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



If $\text{Bias}(\hat{\theta}) = 0$ for all θ, the estimator is **unbiased**.

Worked Example: Is $\bar{X}$ Unbiased for μ?

Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$. Is $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ unbiased?

Step 1: Compute $\mathbb{E}[\hat{\mu}]$:

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

Worked Example: Is $\bar{X}$ Unbiased for μ?

Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$. Is $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ unbiased?

Step 1: Compute $\mathbb{E}[\hat{\mu}]$:

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

Step 2: Check bias:

$$\text{Bias}(\bar{X}) = \mathbb{E}[\bar{X}] - \mu = \mu - \mu = 0$$ ✓ Unbiased!

Recipe for any estimator:
(1) Compute $\mathbb{E}[\hat{\theta}]$ → (2) Subtract the true θ → (3) If the result is 0, it's unbiased.

Worked Example: Why Dividing by n Is Biased

We want to estimate $\sigma^2 = \mathrm{Var}(X_i)$. Natural guess: $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Trick: rewrite each $(X_i - \bar{X})$ by adding and subtracting the true mean μ:

$$X_i - \bar{X} = \underbrace{(X_i - \mu)}_{\text{deviation from truth}} - \underbrace{(\bar{X} - \mu)}_{\text{estimation error}}$$

Squaring and summing gives the **key identity**:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Worked Example: Why Dividing by n Is Biased

We want to estimate $\sigma^2 = \mathrm{Var}(X_i)$. Natural guess: $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Trick: rewrite each $(X_i - \bar{X})$ by adding and subtracting the true mean μ:

$$X_i - \bar{X} = \underbrace{(X_i - \mu)}_{\text{deviation from truth}} - \underbrace{(\bar{X} - \mu)}_{\text{estimation error}}$$

Squaring and summing gives the **key identity**:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Take expectations (using $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$ and $\mathrm{Var}(\bar{X}) = \sigma^2/n$):

$$\mathbb{E}\Big[\sum(X_i - \mu)^2\Big] = n\sigma^2 \qquad (n \text{ terms, each } \sigma^2)$$

$$\mathbb{E}\big[n(\bar{X} - \mu)^2\big] = n \cdot \mathrm{Var}(\bar{X}) = n \cdot \frac{\sigma^2}{n} = \sigma^2 \qquad \text{(one "lost" degree of freedom)}$$

$$\Rightarrow \quad \mathbb{E}\Big[\sum(X_i - \bar{X})^2\Big] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Bessel's Correction: The Fix

From the previous slide: $\mathbb{E}\left[\sum(X_i - \bar{X})^2\right] = (n-1)\sigma^2$, so:

$$\mathbb{E}[\hat{\sigma}_n^2] = \mathbb{E}\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right] = \frac{(n-1)\sigma^2}{n} \neq \sigma^2$$ **Biased!**

It **underestimates** by σ^2/n. Why? We used $\bar{X}$ instead of μ, "using up" one degree of freedom.

Bessel's correction: Divide by $n-1$ instead of n:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \mathbb{E}[S^2] = \sigma^2 \quad \checkmark \text{ Unbiased!}$$

Intuition: We estimated μ from the same data, so the residuals $(X_i - \bar{X})$ are "too small" on average. Dividing by $n-1$ corrects for this.

Bias: Summary

Estimator	Bias	Unbiased?
$\bar{X} = \frac{1}{n}\sum X_i$ for μ	0	Yes
$\hat{\sigma}_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ for σ^2	$-\frac{\sigma^2}{n}$	No
$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ for σ^2	0	Yes
$\hat{p} = \frac{\sum X_i}{n}$ for p (Bernoulli)	0	Yes

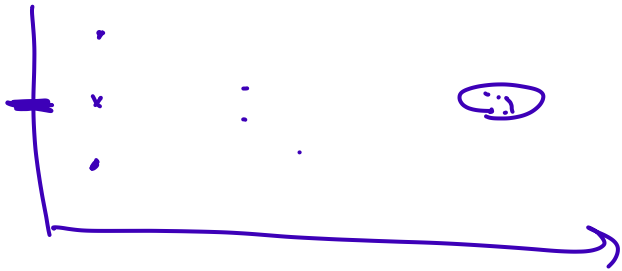
Dividing by n instead of $n-1$ **underestimates** the true variance. Bessel's correction $(n-1)$ fixes this. Recall Lecture 2!

Unbiasedness Alone Isn't Enough

Consider estimating $\mu = \mathbb{E}[X_i]$ from $X_1, \ldots, X_n$.

Surprising fact: $\tilde{\mu} = X_1$ is also **unbiased**!

$$\mathbb{E}[X_1] = \mu \quad \Rightarrow \quad \text{Bias}(X_1) = 0 \quad \checkmark$$



Unbiasedness Alone Isn't Enough

Consider estimating $\mu = \mathbb{E}[X_i]$ from $X_1, \ldots, X_n$.

Surprising fact: $\tilde{\mu} = X_1$ is also **unbiased**!

$$\mathbb{E}[X_1] = \mu \quad \Rightarrow \quad \text{Bias}(X_1) = 0 \quad \checkmark$$

But it's a terrible estimator — it ignores $X_2, \ldots, X_n$ entirely!

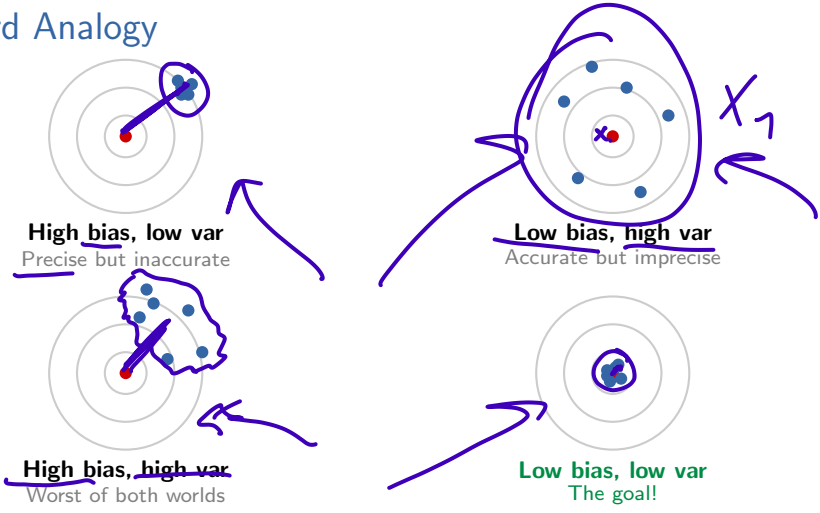
Three statisticians go deer hunting.

The first one shoots **1 meter to the left** of the deer.
The second one shoots **1 meter to the right**.
The third one shouts: *"We got it!"*

On average they hit the target — **unbiased!** But not very useful...

Lesson: Unbiasedness only says $\mathbb{E}[\hat{\theta}] = \theta$. It says nothing about how much $\hat{\theta}$ **varies**. We need more: **variance** and **MSE**.

The Dartboard Analogy



Bullseye = true θ. Blue dots = estimates from repeated samples.

Variance of an Estimator

The **variance** measures how much $\hat{\theta}$ wobbles across samples: $\text{Var}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right]$

Why is $\text{Var}(\bar{X}) = \sigma^2/n$ and not σ^2/n^2?

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) \qquad (\tfrac{1}{n} \text{ comes out as } \tfrac{1}{n^2})$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) \qquad (\text{independent} \Rightarrow \text{variances } \textbf{add})$$

$$= \frac{1}{n^2}\cdot n\sigma^2 = \boxed{\frac{\sigma^2}{n}} \qquad (n \text{ terms cancel one } n)$$

$$\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}} \quad (\textbf{standard error} = \sqrt{\text{Var}})$$

Mean Squared Error: The Total Error

Bias tells us about the **aim**, variance about the **spread**. Can we combine them?

Mean Squared Error: $\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$

The average squared distance from the estimate to the truth.

The trick: add and subtract $\mathbb{E}[\hat{\theta}]$ to decompose the error:

$$\hat{\theta} - \theta = \underbrace{(\hat{\theta} - \mathbb{E}[\hat{\theta}])}_{\text{random fluctuation}} + \underbrace{(\mathbb{E}[\hat{\theta}] - \theta)}_{\text{bias (a constant!)}}$$

This splits the total error into two pieces: the **random part** (how much $\hat{\theta}$ moves around its own mean) and the **systematic part** (how far that mean is from the truth).

MSE = Bias^2 + Variance: The Proof

Now square $\hat{\theta} - \theta = (\hat{\theta} - \mathbb{E}[\hat{\theta}]) + (\mathbb{E}[\hat{\theta}] - \theta)$ and take expectations:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\Big] + 2\underbrace{(\mathbb{E}[\hat{\theta}] - \theta)}_{\text{constant}}\underbrace{\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]]}_{=0 \text{ (always!)}} + (\mathbb{E}[\hat{\theta}] - \theta)^2$$

The cross term vanishes because $\hat{\theta} - \mathbb{E}[\hat{\theta}]$ has mean zero **by definition**.

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

Unbiased means MSE = Var, but a biased estimator can still win if its variance is low enough.

When Biased Beats Unbiased

Example: Estimating σ^2 from $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.

Estimator	Bias	Variance	MSE
$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$	0	$\frac{2\sigma^4}{n-1}$	$\frac{2\sigma^4}{n-1}$
$\hat{\sigma}_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$	$-\frac{\sigma^2}{n}$	$\frac{2(n-1)\sigma^4}{n^2}$	$\frac{(2n-1)\sigma^4}{n^2}$

Compare: $\frac{2n-1}{n^2}$ vs $\frac{2}{n-1}$ $\Rightarrow$ $\hat{\sigma}_n^2$ has **lower MSE** for all $n \geq 2$!

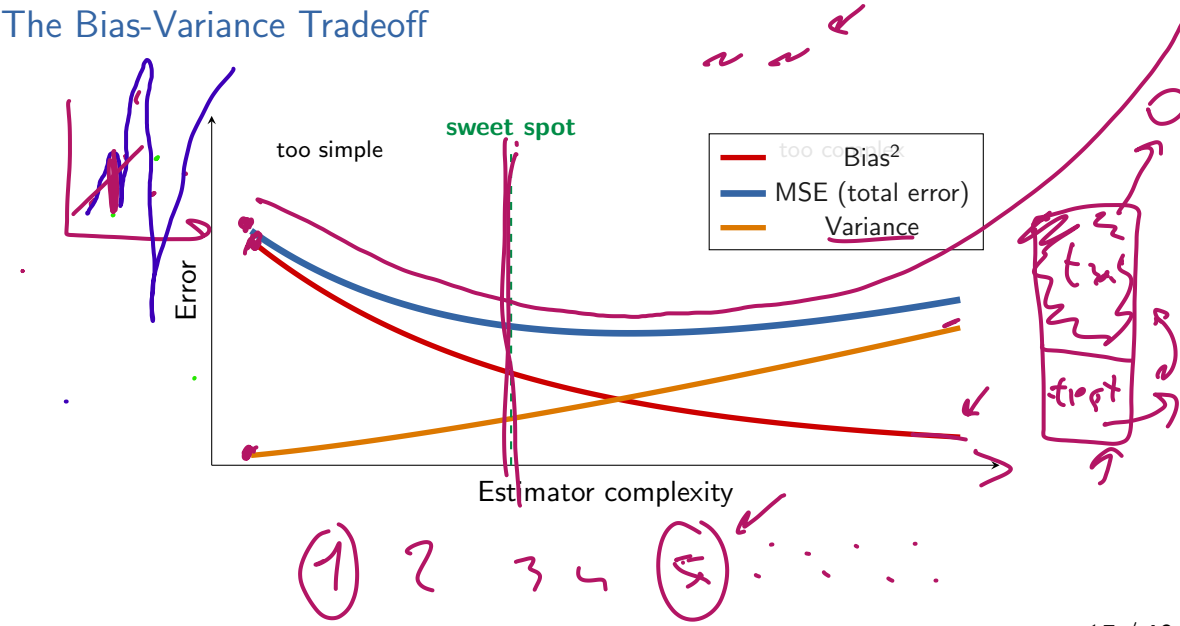
The biased estimator beats the unbiased S^2 because its variance reduction outweighs the small bias.

The Bias-Variance Tradeoff

You can't minimize bias and variance at the same time.

How do we find the **sweet spot**?

The Bias-Variance Tradeoff



Bias-Variance in Machine Learning

This tradeoff is **everywhere** in ML — it's the same principle in different disguises:

Setting	Too simple (high bias)	Too complex (high var)
Polynomial regression	Degree 1 (line)	Degree 20 (wiggly)
KNN	Large k (oversmoothed)	$k = 1$ (memorizes noise)
Decision tree	Shallow tree (underfits)	Deep tree (overfits)
Neural network	Too few neurons	Too many neurons
Regularization	Strong penalty (λ large)	No penalty ($\lambda = 0$)

Key insight: In all these cases, the total error (MSE, test loss) is minimized at an intermediate complexity. This is why we need **cross-validation**, **regularization**, and **held-out test sets** — to find the sweet spot empirically.

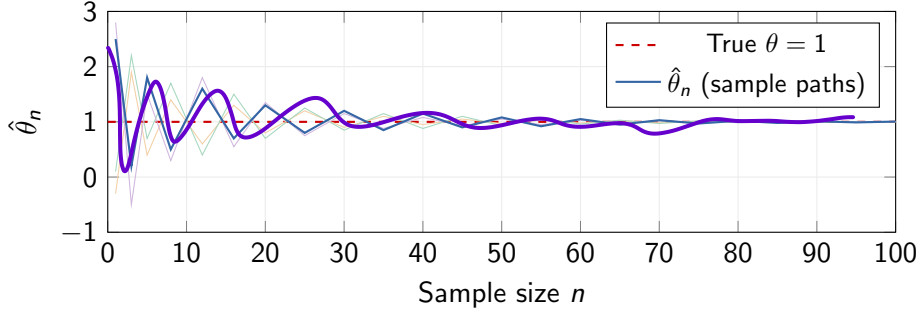
Consistency

Does our estimator converge to the truth as we collect more and more data?

Consistency: Getting It Right Eventually

An estimator $\hat{\theta}_n$ is **consistent** if it converges to the truth as $n \to \infty$:

$$\hat{\theta}_n \xrightarrow{P} \theta \qquad \text{i.e.,} \quad \Pr\Big(|\hat{\theta}_n - \theta| > \varepsilon\Big) \to 0 \ \text{ for all } \varepsilon > 0$$



Consistent vs Inconsistent: A Contrast

Consistent: $\hat{\mu} = \bar{X}_n$

- $\mathbb{E}[\bar{X}_n] = \mu$ (unbiased)
- $\mathrm{Var}(\bar{X}_n) = \sigma^2/n \to 0$
- Uses **all** n observations
- More data $\Rightarrow$ more precise

Not consistent: $\tilde{\mu} = X_1$

- $\mathbb{E}[X_1] = \mu$ (also unbiased!)
- $\mathrm{Var}(X_1) = \sigma^2$ (constant!)
- Uses **only** the first observation
- Ignores all other data forever

Unbiased $\neq$ consistent. X_1 is unbiased but NOT consistent.

Consistent $\neq$ unbiased. $\hat{\sigma}^2_n = \frac{1}{n}\sum(X_i - \bar{X})^2$ is biased but IS consistent (because its bias $\to$ 0 and its variance $\to$ 0).

Sufficient Conditions for Consistency

Chebyshev's inequality gives us a concrete tool:

$$\Pr\Big(|\hat{\theta}_n - \theta| \geq \varepsilon\Big) \leq \frac{\mathbb{E}[(\hat{\theta}_n - \theta)^2]}{\varepsilon^2} = \frac{\mathrm{MSE}(\hat{\theta}_n)}{\varepsilon^2} = \frac{\mathrm{Bias}^2 + \mathrm{Var}}{\varepsilon^2}$$

$\mathrm{Bias}(\hat{\theta}_n) \to 0$ as $n \to \infty$

$\mathrm{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$

$\Downarrow$

$\mathrm{MSE} \to 0 \Rightarrow \Pr(|\hat{\theta}_n - \theta| \geq \varepsilon) \to 0 \Rightarrow$ **consistent!**

Example: $\bar{X}_n$ is consistent for μ: Bias $= 0$, Var $= \sigma^2/n \to 0$, so $\Pr(|\bar{X}_n - \mu| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2) \to 0$.

This is precisely the **(Weak) Law of Large Numbers**: $\bar{X}_n \xrightarrow{P} \mu$

Sufficiency

We have n data points. Do we really need **all** of them?

Can we **compress** without losing information?

$T(X_1, X_2, \cdots, X_n) = \#\{X_i = T\}$ $P(T) = p$

Sufficiency: Can We Compress the Data?

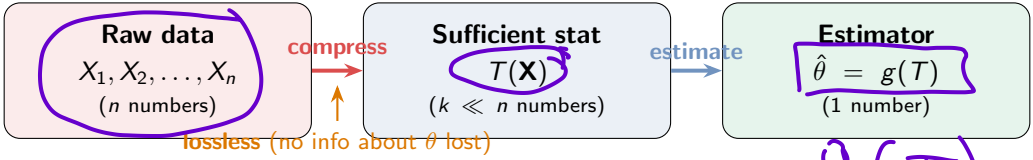
Example: $X_1, \ldots, X_n \sim \text{Bern}(p)$. To estimate p:

- We only need $T = \sum X_i$ (total number of successes)
- The specific order (HHTHT vs THHTH) tells us nothing more about p

Definition: A statistic $T(\mathbf{X})$ is **sufficient** for θ if the conditional distribution of $\mathbf{X} \mid T(\mathbf{X})$ does not depend on θ.

Intuition: Once you know T, the remaining randomness in the data is just noise — it carries **no information** about θ. T is a "lossless summary."

Sufficiency as Data Compression



Example **Bernoulli**

$0, 1, 1, 0, 1, 1, 1, 0, 1, 0 \longrightarrow T = \sum X_i = 6 \longrightarrow \hat{p} = 6/10 = 0.6$

The order $(0, 1, 1, 0, 1, \ldots)$ doesn't matter for estimating p — only the **total count** matters.

How to Check: Fisher–Neyman Factorization

Theorem: $T(\mathbf{X})$ is sufficient for θ if and only if:

$$f(\mathbf{x} \mid \theta) = g(T(\mathbf{x}),\ \theta) \cdot h(\mathbf{x})$$

where g depends on the data **only through** T and h does not depend on θ.

Bernoulli worked example: $X_1, \ldots, X_n \sim \text{Bern}(p)$, let $T = \sum X_i$.

$$f(\mathbf{x} \mid p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = \underbrace{p^{\sum x_i}(1-p)^{n-\sum x_i}}_{g(T,\,p)} \cdot \underbrace{1}_{h(\mathbf{x})}$$

Model	Sufficient statistic	Intuition
Bern(p)	$T = \sum X_i$	1 number for 1 parameter
$N(\mu, \sigma_0^2)$ (σ_0^2 known)	$T = \bar{X}$	1 number for 1 parameter
$N(\mu, \sigma^2)$ (both unknown)	$T = (\bar{X},\ S^2)$	2 numbers for 2 parameters

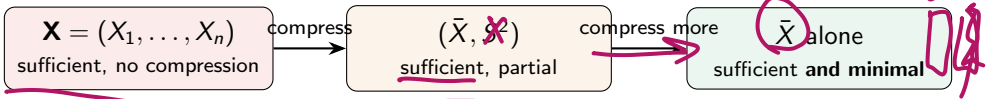
Minimal Sufficiency

The full data $\mathbf{X}$ is always trivially sufficient. But can we compress **further**?

A sufficient statistic is **minimal** if it is a function of every other sufficient statistic.

It achieves the **maximum compression** without losing information about θ.

Example: For $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ with σ_0^2 known:



Since only μ is unknown, S^2 carries no extra information — $\bar{X}$ alone is enough.

The Rao–Blackwell Theorem

Why does sufficiency matter for estimation? Because it lets us **improve** any estimator:

Rao–Blackwell Theorem: Given *any* unbiased estimator $\tilde{\theta}$ and a sufficient statistic T, define $\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T]$. Then:

(1) $\hat{\theta}$ is still **unbiased**: $\mathbb{E}[\hat{\theta}] = \mathbb{E}[\tilde{\theta}] = \theta$

(2) $\hat{\theta}$ has **lower or equal variance**: $\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})$

Conditioning on a sufficient statistic **never hurts, often helps**.

$\tilde{\theta}$ + T

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

6-20
77.352+0.0

The Rao–Blackwell Theorem

Why does sufficiency matter for estimation? Because it lets us **improve** any estimator:

Rao–Blackwell Theorem: Given *any* unbiased estimator $\tilde{\theta}$ and a sufficient statistic T, define $\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T]$. Then:

(1) $\hat{\theta}$ is still **unbiased**: $\mathbb{E}[\hat{\theta}] = \mathbb{E}[\tilde{\theta}] = \theta$

(2) $\hat{\theta}$ has **lower or equal variance**: $\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})$

Conditioning on a sufficient statistic **never hurts, often helps**.

Worked example: $X_1, \ldots, X_n \sim \mathrm{Bern}(p)$, sufficient stat $T = \sum X_i$.

$$\underbrace{\tilde{p} = X_1}_{\text{naive: unbiased, Var}=p(1-p)} \xrightarrow{\mathbb{E}[\,\cdot \mid T\,]} \underbrace{\hat{p} = \mathbb{E}[X_1 \mid T] = T/n = \bar{X}}_{\text{improved: unbiased, Var}=p(1-p)/n} \quad \times \mathbf{n\ better!}$$

What Does $\mathbb{E}[\tilde{\theta} \mid T]$ Actually Mean?

Concrete example: $X_1, X_2, X_3 \sim \text{Bern}(p)$, $T = X_1 + X_2 + X_3$, $\tilde{p} = X_1$.

Suppose someone tells you $T = 2$ (two successes). Which data vectors give $T = 2$?

(X_1, X_2, X_3)	T	$\tilde{p} = X_1$	Equally likely?
$(1, 1, 0)$	2	1	Yes
$(1, 0, 1)$	2	1	Yes
$(0, 1, 1)$	2	0	Yes

$T(x) = 2$

$\tilde{\theta} \mid T$

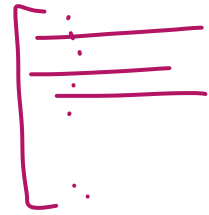
What Does $\mathbb{E}[\tilde{\theta} \mid T]$ Actually Mean?

Concrete example: $X_1, X_2, X_3 \sim \text{Bern}(p)$, $T = X_1 + X_2 + X_3$, $\tilde{p} = X_1$.

Suppose someone tells you $T = 2$ (two successes). Which data vectors give $T = 2$?

(X_1, X_2, X_3)	T	$\tilde{p} = X_1$	Equally likely?
$(1, 1, 0)$	2	1	Yes
$(1, 0, 1)$	2	1	Yes
$(0, 1, 1)$	2	0	Yes

$$\mathbb{E}[X_1 \mid T = 2] = \frac{1+1+0}{3} = \frac{2}{3} = \frac{T}{n} \quad \checkmark$$



"Condition on T" means: average $\tilde{\theta}$ over all data configurations that produce the same value of T. The noise (which specific X_i's are 1 vs 0) gets averaged away. Only the useful part (T) survives.

Why Does Rao–Blackwell Work?

The key is the **law of total variance** (a.k.a. Eve's law):

$$\mathrm{Var}(\tilde{\theta}) = \underbrace{\mathbb{E}\big[\mathrm{Var}(\tilde{\theta} \mid T)\big]}_{\text{“useless” noise} \geq 0} + \mathrm{Var}\Big(\underbrace{\mathbb{E}[\tilde{\theta} \mid T]}_{\hat{\theta}}\Big)$$

$\mathrm{Var}(\hat{\theta} \mid T) + \mathbb{E}(\tilde{\theta} \mid T)$

Why Does Rao–Blackwell Work?

The key is the **law of total variance** (a.k.a. Eve's law):

$$\mathrm{Var}(\tilde{\theta}) = \underbrace{\mathbb{E}\Big[\mathrm{Var}(\tilde{\theta} \mid T)\Big]}_{\text{“useless” noise} \geq 0} + \mathrm{Var}\left(\underbrace{\mathbb{E}[\tilde{\theta} \mid T]}_{\hat{\theta}}\right)$$

Since the first term ≥ 0, we immediately get: $\boxed{\mathrm{Var}(\tilde{\theta}) \;\geq\; \mathrm{Var}(\hat{\theta})}$

$\tilde{\theta}$ **(before)**
Uses data in a noisy way

$\xrightarrow{\mathbb{E}[\,\cdot \mid T\,]}$

$\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T]$ **(after)**
Averages out the noise

Intuition: T captures all the useful information about θ. Conditioning on T removes the “useless” randomness (the part that doesn't tell us about θ). What's left is a cleaner estimator.

Finding Minimal Sufficient Statistics

Theorem (Likelihood Ratio Criterion): $T(\mathbf{X})$ is minimal sufficient iff for all $\mathbf{x}, \mathbf{y}$:

$$T(\mathbf{x}) = T(\mathbf{y}) \iff \frac{f(\mathbf{x} \mid \theta)}{f(\mathbf{y} \mid \theta)} \text{ does not depend on } \theta$$

Bernoulli example: $X_1, \ldots, X_n \sim \text{Bern}(p)$.

$$\frac{f(\mathbf{x} \mid p)}{f(\mathbf{y} \mid p)} = \frac{p^{\sum x_i}(1-p)^{n-\sum x_i}}{p^{\sum y_i}(1-p)^{n-\sum y_i}} = \left(\frac{p}{1-p}\right)^{\sum x_i - \sum y_i}$$

Free of $p \iff \sum x_i = \sum y_i$. So $T = \sum X_i$ is **minimal sufficient** for p. ✓

Recipe: Write the likelihood ratio $f(\mathbf{x} \mid \theta)/f(\mathbf{y} \mid \theta)$.
Find which function of the data must match for the ratio to lose its θ-dependence.
That function is the minimal sufficient statistic.

The Exponential Family: A Unifying Framework

All our examples — Bernoulli, Normal, Poisson, Exponential — share one structure:

$$f(x \mid \theta) = h(x)\ \exp\Big(\eta(\theta)\ T(x) - A(\theta)\Big)$$

Distribution	**Natural param** $\eta(\theta)$	$T(x)$	**Suff. stat** (n obs)
Bern(p)	$\log \frac{p}{1-p}$	x	$\sum X_i$
$N(\mu, \sigma_0^2)$ (σ_0^2 known)	μ/σ_0^2	x	$\sum X_i$
Pois(λ)	$\log \lambda$	x	$\sum X_i$
Exp(λ)	$-\lambda$	x	$\sum X_i$

Pattern: For single-parameter families, $T(x) = x$. The sufficient statistic for n observations is always $\sum T(X_i)$ — straight from the factorization theorem!

Why Exponential Families Are Special

Nearly every nice property we've discussed is **automatic** in exponential families:

Sufficiency: $T(\mathbf{X}) = \sum T(X_i)$ is sufficient **and minimal**

Completeness: the natural sufficient statistic is **complete** (see below)

Regularity: all conditions for the Cramér–Rao bound (coming soon) are satisfied

Optimal estimators exist: we'll see this when we reach the CR bound

Completeness means: if $\mathbb{E}_\theta[g(T)] = 0$ for all θ, then $g(T) = 0$ a.s. $\rightarrow$ **no non-trivial unbiased estimator of zero** based on T.

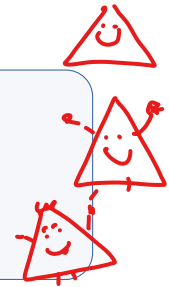
Lehmann–Scheffé: An unbiased estimator based on a **complete** sufficient statistic is the **unique best** unbiased estimator (UMVUE). For exp. families, $\sum T(X_i)$ is always complete $\Rightarrow$ UMVUE exists!

Can We Do Better? The Fundamental Question

We know $\text{Var}(\bar{X}) = \sigma^2/n$ for estimating the mean.

Can **any** unbiased estimator have **lower** variance?

Or is $\bar{X}$ already the best we can do?



To answer this, we need to measure **how much information** one observation carries about θ.

Roadmap:

Why log? $\rightarrow$ **Score function** (sensitivity of the model to θ) $\rightarrow$ **Fisher information** $\rightarrow$ **Cramér–Rao bound** (the variance floor)

Why the Logarithm? From Products to Sums

The likelihood for i.i.d. data is a **product**:

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

Taking the log turns this into a **sum**:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

Products are painful:

- Multiplying tiny numbers $\rightarrow$ underflow
- Product rule for derivatives is messy
- Hard to work with analytically

Sums are friendly:

- Numerically stable
- Derivative of a sum = sum of derivatives
- LLN, CLT apply directly

Key fact: log is monotonically increasing, so $\arg\max_\theta L(\theta) = \arg\max_\theta \ell(\theta)$. Same maximizer!

The Score Function: How Sensitive Is the Model?

Given a model $f(x \mid \theta)$, the **score** measures how the log-probability changes with θ:

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(X \mid \theta)$$

Concrete example: $X \sim \text{Bernoulli}(p)$.
$\log f(x \mid p) = x \log p + (1-x) \log(1-p)$

$$s(p) = \frac{\partial}{\partial p} \left[x \log p + (1-x) \log(1-p) \right] = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-p}{p(1-p)}$$

- If we observe $x = 1$ and p is small, the score is **large positive** $\rightarrow$ "p should be higher"
- If we observe $x = 0$ and p is large, the score is **large negative** $\rightarrow$ "p should be lower"
- On average: $\mathbb{E}[s(p)] = 0$ — the score points in the right direction but **averages out**

Fisher Information: How Informative Is One Observation?

The score averages to zero, but it **varies**. More variation = more information:

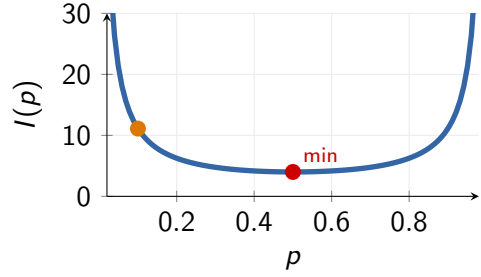
$$I(\theta) = \mathrm{Var}[s(\theta)] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right]$$

Bernoulli derivation: We found $s(p) = \frac{X-p}{p(1-p)}$.

Since $\mathbb{E}[s] = 0$:

$$I(p) = \mathbb{E}[s^2] = \mathbb{E}\left[\frac{(X-p)^2}{p^2(1-p)^2}\right]$$

$$= \frac{\mathrm{Var}(X)}{p^2(1-p)^2} = \frac{p(1-p)}{p^2(1-p)^2} = \boxed{\frac{1}{p(1-p)}}$$



p near 0 or 1: very informative. $p = 0.5$: max noise, min info.

Fisher Information: Two Equivalent Forms

Under regularity conditions, there is an equivalent formula that's often easier to compute:

$$\boxed{I(\theta) = \mathbb{E}\big[s(\theta)^2\big] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right]}$$

Why are these the same? Start from $\mathbb{E}[s(\theta)] = 0$ and differentiate both sides w.r.t. θ:

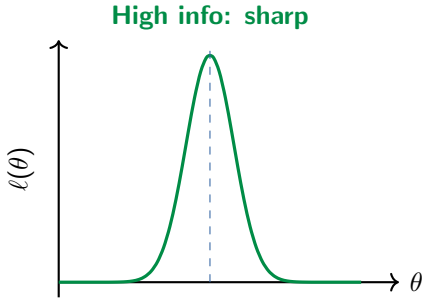
$$0 = \frac{\partial}{\partial\theta}\mathbb{E}[s] = \mathbb{E}\left[\frac{\partial s}{\partial\theta}\right] + \mathbb{E}[s \cdot s] = \mathbb{E}[\ell''] + \mathbb{E}[s^2]$$

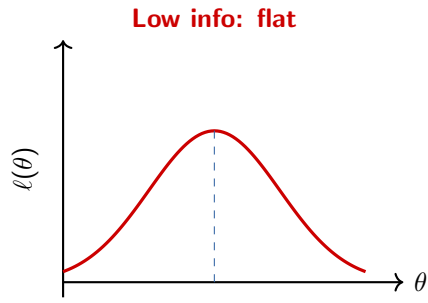
So: $\mathbb{E}[s^2] = -\mathbb{E}[\ell'']$. ✓

Verify for Bernoulli: $\ell(p) = x\log p + (1-x)\log(1-p)$

$$\ell''(p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2} \quad \Rightarrow \quad -\mathbb{E}[\ell''] = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$ ✓

Intuition: Sharp vs Flat Log-Likelihood





$I(\theta)$ measures the **curvature** of the log-likelihood at the true θ.

Sharp curve $\Rightarrow$ high $I(\theta)$ $\Rightarrow$ data is very informative $\Rightarrow$ estimator is precise.

This connects the two forms: $I(\theta) = -\mathbb{E}[\ell'']$ is literally the expected curvature.

Cramér–Rao Lower Bound

Now we can answer the fundamental question. For any **unbiased** estimator $\hat{\theta}$ based on n i.i.d. observations:

$$\boxed{\text{Var}(\hat{\theta}) \geq \frac{1}{n \cdot I(\theta)}}$$

Intuition: Why $\frac{1}{n \cdot I(\theta)}$?

- **More observations (n large)** $\Rightarrow$ bound gets smaller $\Rightarrow$ can estimate more precisely
- **More informative data ($I(\theta)$ large)** $\Rightarrow$ bound gets smaller $\Rightarrow$ each observation tells us more
- The bound is **tight** for many models — it's the actual achievable precision

Verify for Bernoulli:

$$I(p) = \frac{1}{p(1-p)} \quad \Rightarrow \quad \text{CR bound: } \text{Var}(\hat{p}) \geq \frac{1}{n \cdot \frac{1}{p(1-p)}} = \frac{p(1-p)}{n}$$

Actual variance of $\hat{p} = \bar{X}$: $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$ ✓ Hits the bound exactly!

Cramér–Rao: Efficiency and Practical Use

What it says:
There is a **floor** on how precise any unbiased estimator can be

Efficient estimator:
Achieves the bound — the **best possible**

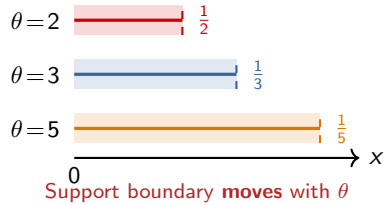
Practical use:
Tells you whether to keep searching for a better estimator

Model	**Estimator**	$\mathrm{Var}(\hat{\theta})$	**CR bound**	**Efficient?**
$\mathrm{Bern}(p)$	$\hat{p} = \bar{X}$	$\frac{p(1-p)}{n}$	$\frac{p(1-p)}{n}$	**Yes**
$N(\mu, \sigma_0^2)$	$\hat{\mu} = \bar{X}$	$\frac{\sigma_0^2}{n}$	$\frac{\sigma_0^2}{n}$	**Yes**
$\mathrm{Exp}(\lambda)$	$\hat{\lambda} = 1/\bar{X}$	$\frac{\lambda^2}{n}$	$\frac{\lambda^2}{n}$	**Yes**

Regularity Conditions: When Does CR Apply?

The Cramér–Rao bound requires **regularity conditions**:

1. **Support** of $f(x \mid \theta)$ doesn't depend on θ
2. θ in the **interior** of the parameter space
3. Can **differentiate under the integral** sign (swap $\frac{\partial}{\partial \theta}$ and $\int$)
4. $0 < I(\theta) < \infty$ (finite, positive info)



Counterexample: Uniform$(0, \theta)$

- Support $[0, \theta]$ depends on θ! (violates #1)
- Suff. stat: $X_{(n)} = \max_i X_i$
- $\mathrm{Var}(X_{(n)}) \sim 1/n^2$ — **faster** than CR!
  (CR would give $1/n$, but $1/n^2$ is possible here)

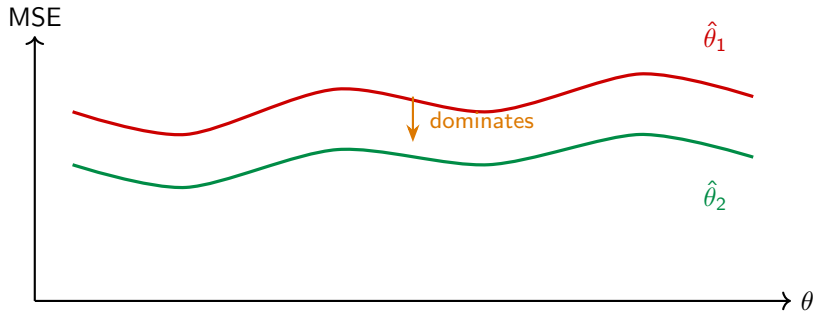
Good news: All exponential family distributions automatically satisfy the regularity conditions. The CR bound always applies to them.

Admissibility

Definition: $\hat{\theta}_1$ is **inadmissible** if $\exists\ \hat{\theta}_2$ that **dominates** it:

$$\text{MSE}(\hat{\theta}_2, \theta) \leq \text{MSE}(\hat{\theta}_1, \theta)\ \ \forall\, \theta, \quad \text{with strict inequality for some } \theta$$

An estimator is **admissible** if no other estimator dominates it.



$\hat{\theta}_1$ is **inadmissible** — $\hat{\theta}_2$ is at least as good everywhere, and strictly better somewhere.

Stein's Paradox (1956)

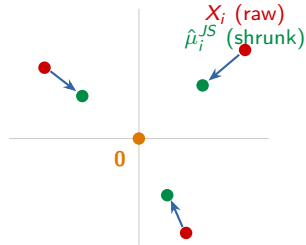
Surprising fact:

When estimating $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ from $X_i \sim N(\mu_i, 1)$,
the sample mean $\hat{\mu}_i = X_i$ is **inadmissible** when $d \geq 3$!

The **James–Stein estimator** dominates it:

$$\hat{\mu}_i^{JS} = \left(1 - \frac{d-2}{\|\mathbf{X}\|^2}\right) X_i$$

- **Shrinks** each X_i toward 0
- Works even if μ_i's are unrelated!
- A little bias buys a lot of variance reduction



Paradox: estimating the average temperature in Yerevan *improves* if you jointly estimate it with the price of tea in China and the height of the Eiffel Tower.

Why Does Stein's Paradox Work?

The MSE comparison tells the whole story:

$\text{MSE}(\mathbf{X}) = d$
(1 per coordinate)

shrinkage helps →

$\text{MSE}(\hat{\boldsymbol{\mu}}^{JS}) < d$
(always, when $d \geq 3$)

Why $d \geq 3$? The shrinkage factor $\frac{d-2}{\|\mathbf{X}\|^2}$ needs to be estimated from data.

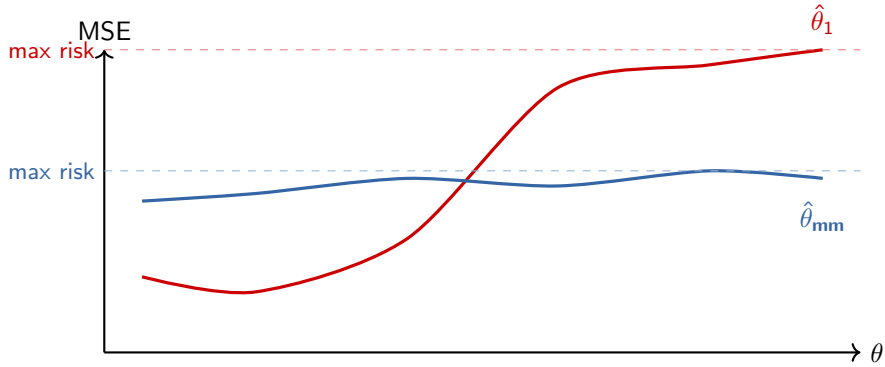
- In $d = 1$ or 2: not enough "room" — estimation error of the shrinkage factor wipes out the gain
- In $d \geq 3$: $\|\mathbf{X}\|^2$ concentrates well enough → shrinkage factor is accurate → net win

Connection to ML: James–Stein shrinkage is an early form of **regularization**.
Ridge regression (L^2 penalty) does the same thing: shrink coefficients toward zero. The bias-variance tradeoff in action: a little bias buys a lot of variance reduction.

Minimax Estimators

A **minimax** estimator minimizes the **worst-case** risk:

$$\hat{\theta}_{\text{minimax}} = \arg\min_{\hat{\theta}} \max_{\theta} \text{MSE}(\hat{\theta}, \theta)$$



Minimax = **conservative**: protects against the worst θ. Minimax hedges.

Three Philosophies of Estimation

Plug-in (unbiased)	**Shrinkage**	**Minimax**
Use sample statistic directly ($\bar{X}$, S^2, $\hat{p}$)	Pull estimates toward a central value (e.g. 0)	Minimize worst-case risk
Admissible in $d = 1$	Biased but lower MSE (James–Stein)	Conservative guarantee
Inadmissible in $d \geq 3$		No single θ can hurt you badly

Takeaway: In high dimensions ($d \geq 3$), shrinkage estimators are provably better than using each sample statistic on its own. We'll see more of this in later lectures.

What We Haven't Covered (Yet)

This lecture focused on **point estimation** — producing a single "best guess" for θ. But there's much more to statistical inference:

Confidence intervals: How uncertain is our estimate? (Lectures 5–6)

Hypothesis testing: Is the effect real or just noise? (Lectures 7–8)

Bayesian estimation: Incorporating prior beliefs (Lecture 5)

Bootstrap: Resampling to estimate uncertainty without formulas (Lecture 6)

Asymptotic theory: What happens as $n \to \infty$ in general? (Lecture 5)

Nonparametric estimation: What if we don't assume a distribution at all?

Today's tools (bias, MSE, CR bound, sufficiency) will be the **foundation** for all of these.

Summary: How to Judge an Estimator

Bias: $\mathbb{E}[\hat{\theta}] - \theta$. Does it aim at the right place?

Variance: $\mathrm{Var}(\hat{\theta})$. How much does it jump around?

MSE $= \mathrm{Bias}^2 + \mathrm{Var}$. Total error. Biased can beat unbiased!

Consistency: $\hat{\theta}_n \xrightarrow{P} \theta$. Converges to truth with enough data.

Sufficiency: $T(\mathbf{X})$ captures everything about θ. Compress without loss.

Cramér–Rao: $\mathrm{Var} \geq 1/(n \cdot I(\theta))$. The efficiency floor.

Admissibility: No other estimator dominates it everywhere.

Minimax: Best worst-case guarantee. Shrinkage often wins.

Homework

1. Show that $\bar{X}$ is unbiased for μ and compute its MSE.
2. Show that $\hat{\sigma}_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ is biased for σ^2. Find the bias.
3. Compute the Fisher information $I(\theta)$ for Poisson(λ).
   Use it to find the Cramér–Rao lower bound for estimating λ.
   Is $\hat{\lambda} = \bar{X}$ efficient?
4. Suppose you shrink $\bar{X}$ toward 0: $\hat{\mu}_c = c\bar{X}$ for $0 < c < 1$.
   Find the bias, variance, and MSE as functions of c.
   For what value of c is MSE minimized? Is the optimal estimator biased?
5. Use the factorization theorem to show that $T = \sum X_i$ is a sufficient statistic for λ when $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$.

Questions?