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):

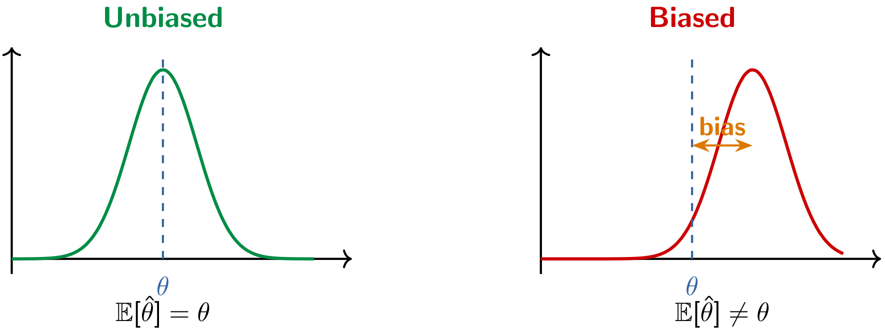
$\bar{X}$ for μ, $\quad S^2$ for σ^2, $\quad \hat{p} = \frac{\text{count}}{n}$ for p

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?

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



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

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

Let $X_1, \dots, X_n$ 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:

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

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

Worked Example: Why Dividing by n Is Biased

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

Key identity (add and subtract μ):

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

Take expectations:

$$\begin{aligned} \mathbb{E}\left[\sum (X_i - \mu)^2\right] &= n\sigma^2 && (n \text{ terms, each } \sigma^2) \\ \mathbb{E}\left[n(\bar{X} - \mu)^2\right] &= \sigma^2 && (\text{since } \mathrm{Var}(\bar{X}) = \sigma^2/n) \end{aligned}$$

So: $\mathbb{E}\left[\sum (X_i - \bar{X})^2\right] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$

$\mathbb{E}[\hat{\sigma}_n^2] = \frac{(n-1)\sigma^2}{n} \neq \sigma^2$ **Biased!** It underestimates by σ^2/n.

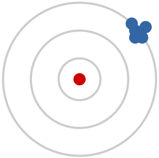
Bessel's correction: Divide by $n-1$: $S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$ ✓
Unbiased

Bias: Summary

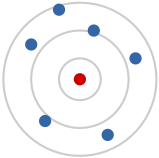
Estimator	Bias	Unbiased?
$\bar{X} = \frac{1}{n}\sum X_i$ for μ	0	**Yes**
$\hat{\sigma}_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ for σ^2	$-\frac{\sigma^2}{n}$	**No**
$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ for σ^2	0	**Yes**
$\hat{p} = \frac{\sum X_i}{n}$ for p (Bernoulli)	0	**Yes**

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

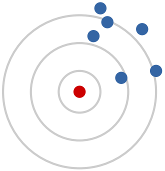
The Dartboard Analogy



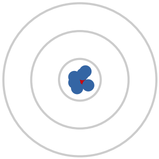
High bias, low var
Precise but inaccurate



Low bias, high var
Accurate but imprecise



High bias, high var
Worst of both worlds



Low bias, low var
The goal!

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

Variance of an Estimator

The **variance** measures how much $\hat{\theta}$ wobbles across samples:

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right]$$

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

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

More data $\Rightarrow$ smaller variance $\Rightarrow$ more precise estimate.
Variance shrinks at rate $1/n$; standard error at rate $1/\sqrt{n}$.

MSE = Bias^2 + Variance: Derivation

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

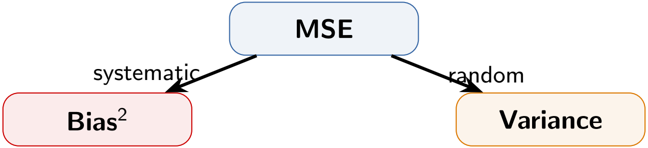
Trick: add and subtract $\mathbb{E}[\hat{\theta}]$:

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

Square and take expectations:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right] + 2\,\underbrace{(\mathbb{E}[\hat{\theta}] - \theta)}_{\text{constant}} \cdot \underbrace{\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]]}_{=0} + (\mathbb{E}[\hat{\theta}] - \theta)^2$$

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



When Biased Beats Unbiased

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

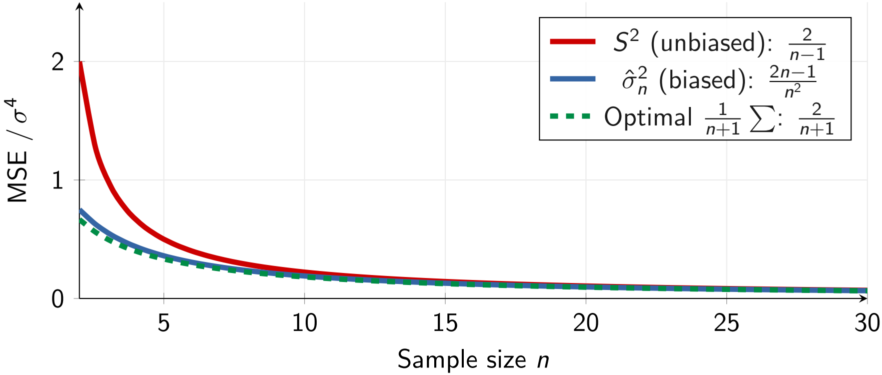
Estimator	**Bias**	**Variance**	**MSE**
$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$	0	$\frac{2\sigma^4}{n-1}$	$\frac{2\sigma^4}{n-1}$
$\hat{\sigma}_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$	$-\frac{\sigma^2}{n}$	$\frac{2(n-1)\sigma^4}{n^2}$	$\frac{(2n-1)\sigma^4}{n^2}$

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.

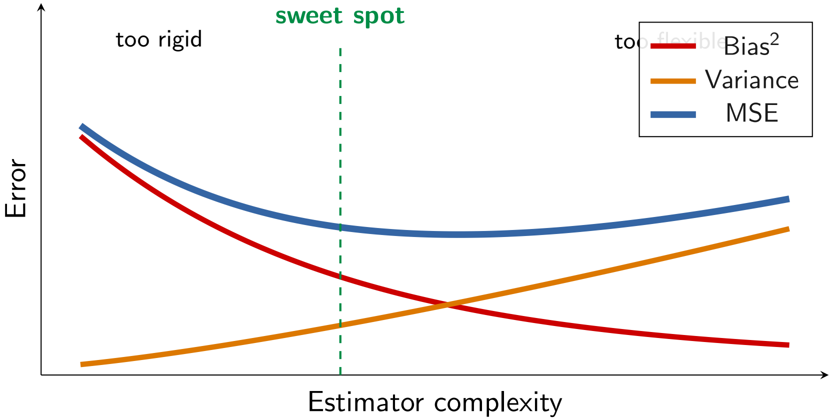
MSE Comparison: Visualized

How do S^2 (unbiased, divides by $n-1$) and $\hat{\sigma}_n^2$ (biased, divides by n) compare as n grows?



The **biased** estimator (blue) always beats the unbiased one (red).
The optimal divisor is actually $n+1$, not n or $n-1$ — even more biased, even lower MSE!

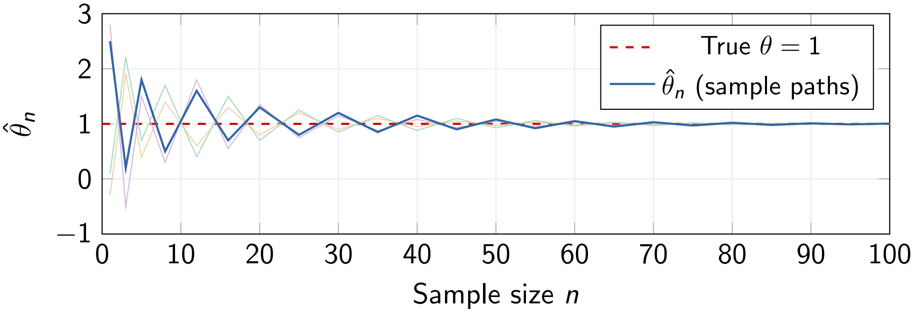
The Bias-Variance Tradeoff



Consistency: Getting It Right Eventually

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

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



Sufficient Conditions for Consistency

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

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

$$\Downarrow$$

$\hat{\theta}_n$ is **consistent** (by Chebyshev's inequality)

- The **sample mean** $\bar{X}_n$ is consistent for μ (by the Law of Large Numbers)
- The **sample variance** S_n^2 is consistent for σ^2
- **Unbiased + vanishing variance** $\Rightarrow$ consistent. But consistency does *not* require unbiasedness!

Sufficiency: Can We Compress the Data?

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

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

Raw data
$X_1, X_2, \dots, X_n$
(n numbers)

compress

Sufficient stat
$T(\mathbf{X})$
($k \ll n$ numbers)

estimate

Estimator
$\hat{\theta} = g(T)$
(1 number)

lossless (no info about θ lost)

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 $(0,1,1,0,1,\dots)$ 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: $X_1, \dots, X_n \sim \text{Bern}(p)$, 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
$\text{Bern}(p)$	$T = \sum X_i$	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 and Why It Matters

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:

- $\mathbf{X} = (X_1, \ldots, X_n)$ — sufficient (trivially), but no compression
- $(\bar{X}, S^2)$ — sufficient, some compression
- $\bar{X}$ alone — sufficient **and minimal**. Maximum compression!

Rao–Blackwell: Any unbiased $\tilde{\theta}$ can be improved: $\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T]$ has $\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})$.

In action: $X_1, \ldots, X_n \sim \mathrm{Bern}(p)$, $T = \sum X_i$: $\underbrace{\tilde{p} = X_1}_{\mathrm{Var}=p(1-p)} \xrightarrow{\mathbb{E}[\,\cdot\,|\,T\,]} \underbrace{\hat{p} = \bar{X}}_{\mathrm{Var}=p(1-p)/n}$ **$\times$n better!**

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:

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

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

Fisher info: $I(\eta) = A''(\eta)$ — just differentiate A twice

Regularity: all conditions for Cramér–Rao are satisfied

Efficiency: the CR bound is achievable — optimal estimators exist

Completeness: T is **complete** if $\mathbb{E}_\theta[g(T)] = 0 \; \forall\, \theta \;\Rightarrow\; g(T) = 0$ a.s.

Lehmann–Scheffé: An unbiased estimator based on a complete sufficient statistic is the **unique best** unbiased estimator (UMVUE).

Can We Do Better? The Fundamental Question

We know $\mathrm{Var}(\bar{X}) = \sigma^2/n$ 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:

Score function (sensitivity of the model to θ) $\rightarrow$
Fisher information (how informative the data is)
$\rightarrow$ **Cramér–Rao bound** (the variance floor)

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:

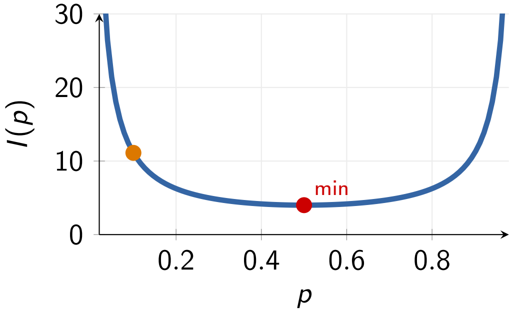
$$\boxed{I(\theta) = \mathrm{Var}[s(\theta)] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right]}$$

Bernoulli example:

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

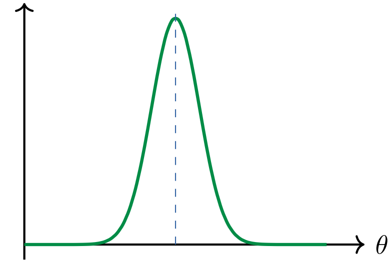
- $p = 0.5$: $I = 4$ (least informative)
- $p = 0.1$: $I = 11.1$ (more informative)
- $p = 0.01$: $I = 101$ (most informative)

Near $p = 0$ or 1: each flip tells you a lot.
At $p = 0.5$: max noise, min information.



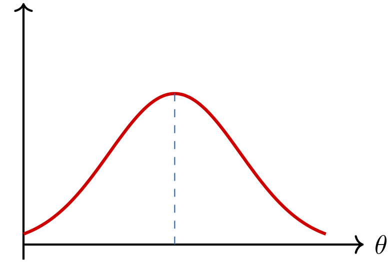
Intuition: Sharp vs Flat Log-Likelihood

High Fisher Information



Sharp peak $\Rightarrow$ precise estimate

Low Fisher Information



Flat peak $\Rightarrow$ uncertain estimate

$I(\theta)$ = **curvature** of the log-likelihood. Sharp curve $\Rightarrow$ high $I(\theta)$ $\Rightarrow$ data is very informative.

Equivalently: $I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right]$ (expected negative curvature).

Cramér–Rao Lower Bound

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

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

Verify for the Bernoulli example:

- We computed $I(p) = \frac{1}{p(1-p)}$
- CR bound: $\mathrm{Var}(\hat{p}) \geq \frac{1}{n \cdot \frac{1}{p(1-p)}} = \frac{p(1-p)}{n}$
- Actual variance of $\hat{p} = \bar{X}$: $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$ ✓ Hits the bound exactly!

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

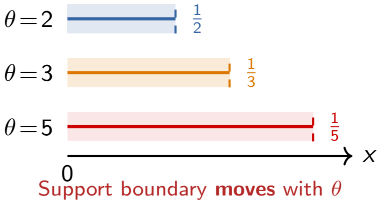
Efficient estimator:
Achieves the bound — the **best possible**

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

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
4. $0 < I(\theta) < \infty$



Counterexample: Uniform$(0, \theta)$

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

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

Cramér–Rao: Checking Efficiency

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

These natural plug-in estimators achieve the bound — they are the **best possible** unbiased estimators.

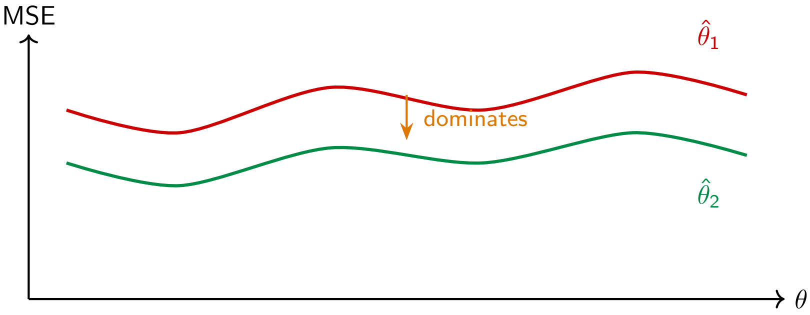
Not every estimator is efficient, but the Cramér–Rao bound tells us how close we can get.

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)

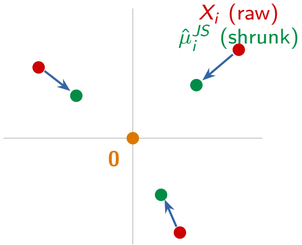
Surprising fact:

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

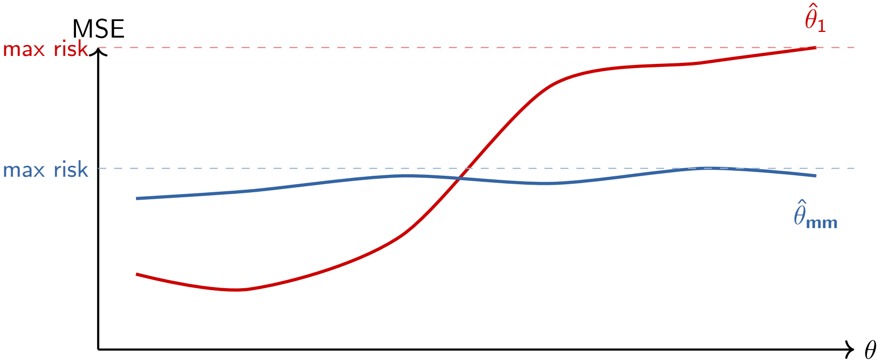


Paradox: estimating the average temperature in Yerevan *improves* if you jointly estimate it with the price of tea in China and the height of the Eiffel Tower.

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, $\hat{p}$) Admissible in $d = 1$ Inadmissible in $d \geq 3$	Pull estimates toward a central value (e.g. 0) Biased but lower MSE (James–Stein)	Minimize worst-case risk Conservative guarantee 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.

Summary: How to Judge an Estimator

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

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

MSE $= \text{Bias}^2 + \text{Var}$. 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: $\text{Var} \geq 1/(n \cdot I(\theta))$. 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 $X_1, \ldots, X_n \sim$ Poisson(λ).

Questions?