# 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):

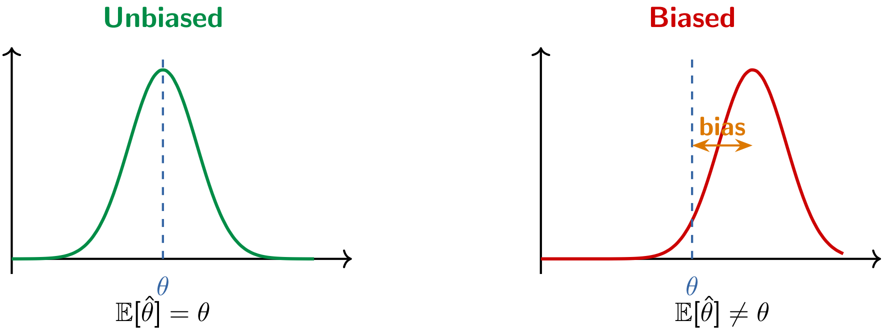
$$\bar{X} \text{ for } \mu, \quad S^2 \text{ for } \sigma^2, \quad \hat{p} = \frac{\text{count}}{n} \text{ for } p$$

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 $\theta$, the estimator is **unbiased**.

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

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 $\mu$?

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:

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

**Recipe for any estimator:**
(1) Compute $\mathbb{E}[\hat{\theta}]$ $\rightarrow$ (2) Subtract the true $\theta$ $\rightarrow$ (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 $\mu$:

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

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 $\mu$:

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

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

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

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

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

# Bessel's Correction: The Fix

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

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

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

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

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

**Intuition:** We estimated $\mu$ 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 $\mu$ | 0 | **Yes** |
| $\hat{\sigma}_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ for $\sigma^2$ | $-\frac{\sigma^2}{n}$ | **No** |
| $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ for $\sigma^2$ | 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!

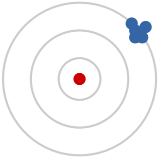
**Three statisticians go deer hunting.**

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!"*

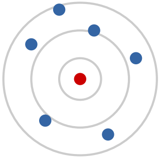
On average they hit the target — **unbiased!** But not very useful...

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

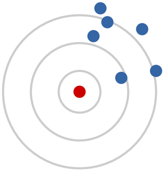
# The Dartboard Analogy



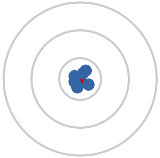
**High bias, low var**
Precise but inaccurate



**Low bias, high var**
Accurate but imprecise



**High bias, high var**
Worst of both worlds



**Low bias, low var**
The goal!

Bullseye $=$ true $\theta$. Blue dots $=$ estimates from repeated samples.

# Variance of an Estimator

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

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

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) && \left(\frac{1}{n} \text{ comes out as } \frac{1}{n^2}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) && (\text{independent} \Rightarrow \text{variances } \textbf{add}) \\
&= \frac{1}{n^2} \cdot n\sigma^2 = \boxed{\frac{\sigma^2}{n}} && (n \text{ terms cancel one } n)
\end{aligned}$$

$\mathrm{SE}(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$ (**standard error** $= \sqrt{\mathrm{Var}}$)

# Mean Squared Error: The Total Error

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

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

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).

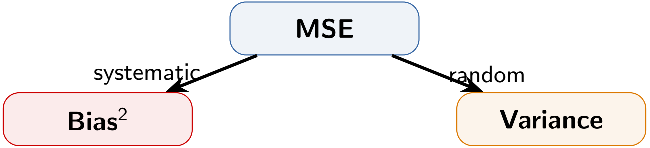
# MSE $=$ Bias$^2$ $+$ Variance: The Proof

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

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

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 $\sigma^2$ from $X_1, \dots, X_n \sim N(\mu, \sigma^2)$.

| 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 $n \geq 2$!

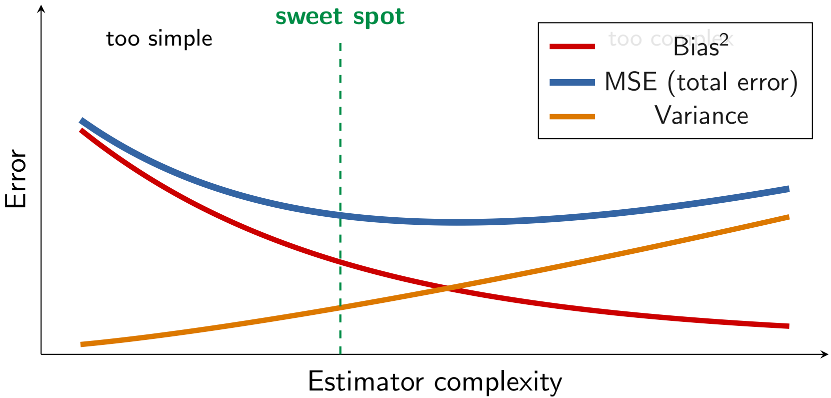
The biased estimator beats the unbiased $S^2$ because its variance reduction outweighs the small bias.

# 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 ($\lambda$ 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.

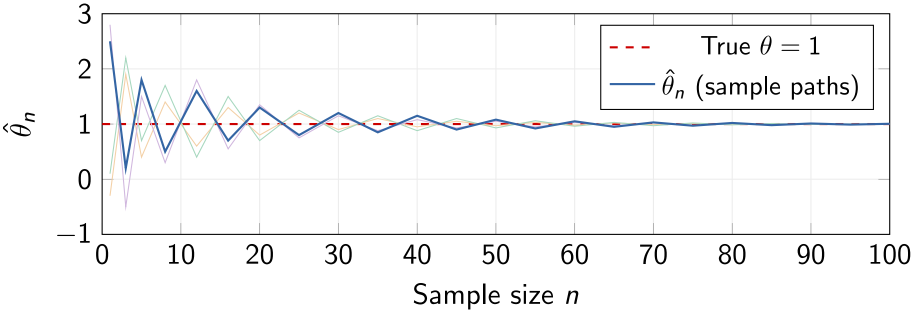
# Consistency

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 $n \to \infty$:

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



# 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:

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

$\mathrm{Bias}(\hat{\theta}_n) \to 0$ 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 $\mu$: $\mathrm{Bias} = 0$, $\mathrm{Var} = \sigma^2/n \to 0$, 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?

# Sufficiency: Can We Compress the Data?

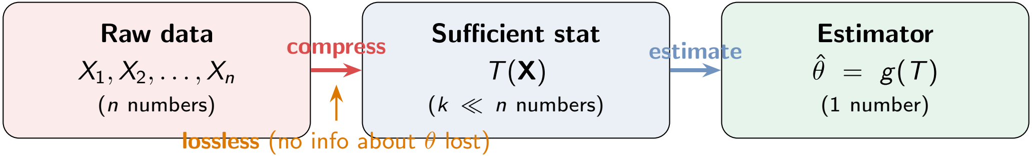
**Example:** $X_1, \ldots, X_n \sim \text{Bern}(p)$. To estimate $p$:

- 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 $\theta$ if
the conditional distribution of $\mathbf{X} \mid T(\mathbf{X})$ does not depend on $\theta$.

**Intuition:** Once you know $T$, the remaining randomness in the data is just noise — it carries **no information** about $\theta$. $T$ 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 $\theta$ 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 $\theta$.

**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 |
|---|---|---|
| $\text{Bern}(p)$ | $T = \sum X_i$ | 1 number for 1 parameter |
| $N(\mu, \sigma_0^2)$ ($\sigma_0^2$ 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 $\theta$.

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

| $\mathbf{X} = (X_1, \ldots, X_n)$ | | $(\bar{X}, S^2)$ | | $\bar{X}$ alone |
|---|---|---|---|---|
| sufficient, no compression | $\xrightarrow{\text{compress}}$ | sufficient, partial | $\xrightarrow{\text{compress more}}$ | sufficient **and minimal** |

Since only $\mu$ 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**.

# 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 $T = \sum X_i$.

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

**×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 |

## 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):

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

# 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 $\geq 0$, we immediately get: $\boxed{\mathrm{Var}(\tilde{\theta}) \ \geq \ \mathrm{Var}(\hat{\theta})}$

$$\text{$\tilde{\theta}$ (before): Uses data in a noisy way} \xrightarrow{\mathbb{E}[\,\cdot\,|\,T\,]} \text{$\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T]$ (after): Averages out the noise}$$

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

# Finding Minimal Sufficient Statistics

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

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

**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 $p$ $\Longleftrightarrow$ $\sum x_i = \sum y_i$. 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 $\theta$-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) |
|---|---|---|---|
| Bern($p$) | $\log \frac{p}{1-p}$ | $x$ | $\sum X_i$ |
| $N(\mu, \sigma_0^2)$ ($\sigma_0^2$ known) | $\mu/\sigma_0^2$ | $x$ | $\sum X_i$ |
| Pois($\lambda$) | $\log \lambda$ | $x$ | $\sum X_i$ |
| 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** (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 $\theta$, then $g(T) = 0$ a.s. $\to$ **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!

# Homework

1. Show that $\bar{X}$ is unbiased for $\mu$ and compute its MSE.
2. Show that $\hat{\sigma}_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ is biased for $\sigma^2$. Find the bias.
3. 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?
4. Use the factorization theorem to show that $T = \sum X_i$ is a sufficient statistic for $\lambda$ when $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$.

# Questions?