Unbiased Variance Estimation

(The mystery of the missing man)

- Evaluate bias
- Understand behavior
- Unbiased estimator
- Resolve mystery
- Dispel half-truth

Mystery of the Missing Man

Sample mean $\overline{X} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$ $E(\overline{X}) = \mu$ Unbiased

"Raw" sample variance $\text{“}S^2\text{”} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Experiments $E(\text{“}S^2\text{”}) \approx \frac{n-1}{n} \cdot \sigma^2$ Biased

Mystery

Mean

Height of 10 people

Add

Normalize by

10

9

Variance

Show $E(\text{“}S^2\text{”}) = \frac{n-1}{n} \cdot \sigma^2$

Why

How to fix

Partial Explanation

$\text{“}S^2\text{”} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ Show $E(\text{“}S^2\text{”}) = \frac{n-1}{n} \cdot \sigma^2$ “S^2” under-estimates σ^2

Given n points $x_1,\ldots,x_n$ $\sum_{i=1}^{n}(x_i - a)^2$ minimized for $a = \frac{x_1+\ldots+x_n}{n}$

1, -1 $(1-a)^2 + (-1-a)^2 = 2 + 2a^2$ minimized for a=0 average

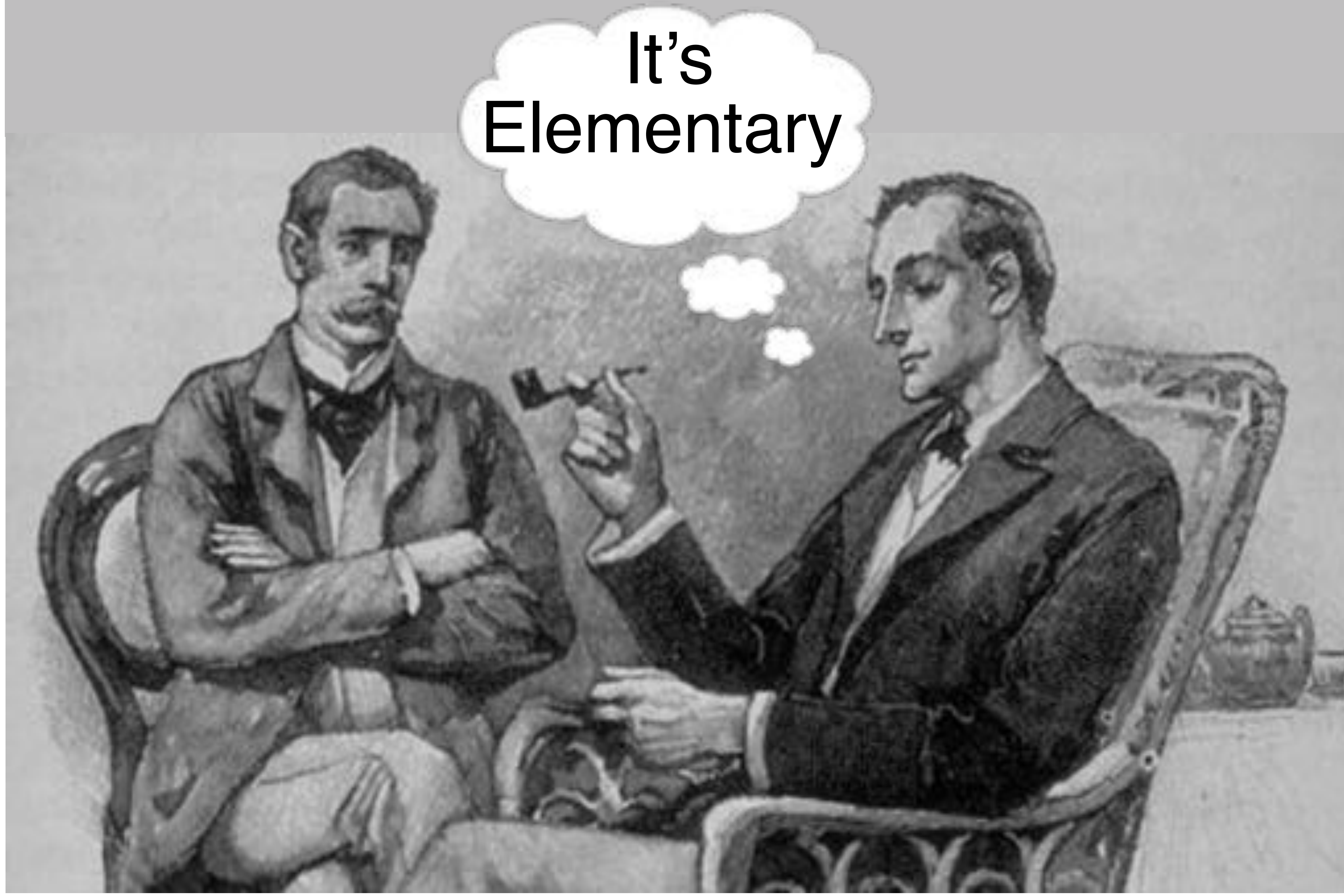
$\sigma^2 \stackrel{\text{def}}{=} E(X-\mu)^2$ $\mu \approx$ average of observations, not exactly

$\text{“}S^2\text{”} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ $\overline{X}$ is exact average Lower sum

Explains “S^2” under-estimates σ^2

Does not Explain $\frac{n-1}{n}$ Nor capture whole reason

$$
\begin{aligned}
E(\text{“}S^{2}\text{”}) &= E\left(\frac{1}{n}\overbrace{\sum_{i=1}^{n}}^{\text{complex}}(X_i-\overline{X})^2\right) \\
&\overset{\text{LoE}}{=} \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&\overset{\text{LoE}}{=} \frac{1}{n}\sum_{i=1}^{n}E(X_i-\overline{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}E(X_1-\overline{X})^2 \\
&= E(X_1-\overline{X})^2
\end{aligned}
$$



Intuitive Simple **Elementary**

Easier understand explain

Recall: Bernoulli

B_p $P(1) = p$ $P(0) = 1\text{-}p = q$ $\sigma^2 = pq$

n=2 x_1, x_2 $\overline{x} = \frac{x_1+x_2}{2}$ $\text{``}S^2\text{''}(x_1, x_2) = \frac{1}{2}((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2)$

x_1,x_2	$P(x_1,x_2)$	$\overline{x}$	"s^2"
0,0	q^2	0	$\frac{1}{2}((0-0)^2 + (0-0)^2) = 0$
0,1	qp	½	$\frac{1}{2}((0-\frac{1}{2})^2 + (1-\frac{1}{2})^2) = \frac{1}{2} \cdot (\frac{1}{4} + \frac{1}{4}) = \frac{1}{4}$
1,0	pq	½	¼
1,1	p^2	1	0

Could get unwieldy!

$$E(\text{``}S^2\text{''}) = \sum_{x_1,x_2} p(x_1, x_2) \cdot \text{``}S^2\text{''}(x_1, x_2)$$

$$= q^2 \cdot 0 + qp \cdot \frac{1}{4} + pq \cdot \frac{1}{4} + p^2 \cdot 0 = \frac{pq}{2} = \frac{\sigma^2}{2}$$

!

Bernoulli

B_P | $P(1) = p$ | $P(0) = 1-p = q$ | $\sigma^2 = E(X-\mu)^2 = p(1-p) = pq$

Simplified calculation

$n=2$ | X_1, X_2

x_1, x_2	$p(x_1, x_2)$	$\overline{x}$	$(x_1 - \overline{x})^2$
0,0	q^2	0	0
0,1	qp	½	¼
1,0	pq	½	¼
1,1	p^2	1	0

$$E(\text{“}S^2\text{”}) = E(X_1 - \overline{X})^2$$

$$= \sum_{x_1, x_2} p(x_1, x_2) \cdot (x_1 - \overline{x})^2$$

$$= 2 \cdot pq \cdot \frac{1}{4} = \frac{1}{2} pq = \frac{1}{2}\sigma^2$$ ✅

Simpler | Easier to analyze

Simplified Formulation

Want to show $E(\text{“}S^2\text{”}) = \frac{n-1}{n} \cdot \sigma^2$

Asymmetric, unclear

$$\overset{\text{def}}{=} E(X_1 - \mu)^2 \qquad X_1 \sim p$$

$$\overset{\text{def}}{=} E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = E(X_1 - \overline{X})^2$$

Show $E(X_1 - \overline{X})^2 = \frac{n-1}{n} \cdot E(X_1 - \mu)^2$

Symmetric, shows difference

Simplistic Argument

$\overline{X}$ includes X_1, hence closer than μ

Doesn't explain $\frac{n-1}{n}$

Not whole story

First n=2

General n

n=2 $E(X_1 - \overline{X})^2 = \frac{1}{2} \cdot E(X_1 - \mu)^2 = \frac{\sigma^2}{2}$

De-couple X_1 from $\overline{X}$ $\quad X_1 - \overline{X} = X_1 - \frac{X_1+X_2}{2} = \frac{X_1-X_2}{2}$

$$E(X_1 - \overline{X})^2 = E(\tfrac{X_1-X_2}{2})^2 = \tfrac{1}{4} \cdot E(X_1 - X_2)^2$$

$X_2 \perp\!\!\!\perp X_1$ If difference was just from correlation between X_1 and $\overline{X}$ we would get $\frac{1}{4} \cdot E(X_1 - \mu)^2 = \frac{\sigma^2}{4}$. Even smaller!

Not whole story. Randomness of X_2 reverses half of decrease. Show $E(X_1 - X_2)^2 = 2 \cdot E(X_1 - \mu)^2$

$$E(X_1 - \overline{X})^2 = \underbrace{\tfrac{1}{4}}_{\text{gain 1/4 from proximity}} \cdot E(X_1 - X_2)^2 = \tfrac{1}{4} \cdot \underbrace{2}_{\text{lose 2 for randomness}} \cdot E(X_1 - \mu)^2 = \frac{\sigma^2}{2}$$

$$E(X_1 - X_2)^2 = 2 \cdot E(X_1 - \mu)^2$$

$$E(X_1 - X_2) = \mu - \mu = 0 \qquad E(X_1 - \mu) = \mu - \mu = 0$$

Both 0-mean

For 0-mean random variable Z

$$E(Z^2) = V(Z)$$

$$E(X_1 - X_2)^2 = 2 \cdot E(X_1 - \mu)^2 \longleftrightarrow V(X_1 - X_2) = 2 \cdot V(X_1)$$

$$V(X_1 - X_2) \overset{\perp\!\!\!\perp}{=} V(X_1) + V(X_2) = 2 \cdot V(X_1)$$

Summary for n=2

$$
\left.
\begin{aligned}
E(``S^2") &\overset{\text{def}}{=} E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&= E(X_1-\overline{X})^2
\end{aligned}
\right\}\text{any n}
$$

$$X_1-\overline{X} = \frac{X_1-X_2}{2}$$

$$
\begin{aligned}
&= E\left(\frac{X_1-X_2}{2}\right)^2 \\
&\overset{\text{LoE}}{=} \frac{1}{4}\cdot E(X_1-X_2)^2 \qquad \boxed{\text{¼ from } \overline{X} \text{ being closer than μ to } X_1} \\
&\overset{\text{0-mean}}{=} \frac{1}{4}\cdot V(X_1-X_2) \\
&\overset{\perp\!\!\!\perp}{=} \frac{1}{4}\cdot (V(X_1)+V(X_2)) \\
&\overset{\text{iid}}{=} \frac{1}{4}\cdot 2\cdot V(X_1) \qquad \boxed{\text{2 from } \overline{X} \text{ being random}} \\
&= \frac{1}{4}\cdot 2\cdot \sigma^2 \;=\; \frac{\sigma^2}{2} \qquad \boxed{\text{1/2 together}}
\end{aligned}
$$

General n

$$E(\text{“}S^2\text{”}) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)$$

$$= E(X_1 - \overline{X})^2$$

$$= E\left(\frac{n-1}{n}\left(X_1 - \frac{X_2+\ldots+X_n}{n-1}\right)\right)^2$$

$$X_1 - \overline{X} = X_1 - \frac{X_1+\ldots+X_n}{n}$$
$$= \frac{(n-1)X_1 - X_2 - \ldots - X_n}{n}$$
$$= \frac{n-1}{n}\left(X_1 - \frac{X_2+\ldots+X_n}{n-1}\right)$$

LoE

$$= \left(\frac{n-1}{n}\right)^2 \cdot E\left(X_1 - \frac{X_2+\ldots+X_n}{n-1}\right)^2$$

$\left(\frac{n-1}{n}\right)^2$ as $\overline{X}$ closer than μ to X_1

0-mean

$$= \left(\frac{n-1}{n}\right)^2 \cdot V\left(X_1 - \frac{X_2+\ldots+X_n}{n-1}\right)$$

$\perp\!\!\!\perp$

$$= \left(\frac{n-1}{n}\right)^2 \cdot \left[V(X_1) + V\left(\frac{X_2+\ldots+X_n}{n-1}\right)\right]$$

iid, var. scaling

$$= \left(\frac{n-1}{n}\right)^2 \cdot \left[\sigma^2 + \frac{\sigma^2}{n-1}\right]$$

$$= \left(\frac{n-1}{n}\right)^2 \cdot \frac{n}{n-1} \cdot \sigma^2$$

$\frac{n}{n-1}$ from $\overline{X}$ being random

$$= \frac{n-1}{n} \cdot \sigma^2$$

$\frac{n-1}{n}$ together

Unbiased Variance Estimate

"Raw" sample variance

$$\text{“}S^2\text{”} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$E(\text{“}S^2\text{”}) = \frac{n-1}{n} \cdot \sigma^2$$

Bessel's Correction

$$S^2 = \frac{n}{n-1} \cdot \text{“}S^2\text{”} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$E(S^2) = \sigma^2$ Unbiased estimator of variance

S^2 typically called sample variance

theoretically interesting

Large sample

Small difference

ExSample

n = 5   2, 1, 4, 2, 6

Saw $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{2+1+4+2+6}{5} = 3$

"S²" = 3.2 $\boxed{\times \frac{5}{4}}$ $\boxed{\times \frac{n}{n-1}}$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1+4+1+1+9}{4} = \frac{16}{4} = 4$$

Unbiased estimate of σ^2

One-pass calculation

$$``S^2" = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \longrightarrow S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$$

Final Simulations

Unbiased Variance Estimation

(The mystery of the missing man)

Evaluate bias

Understand behavior

Unbiased estimator

Bessel Correction

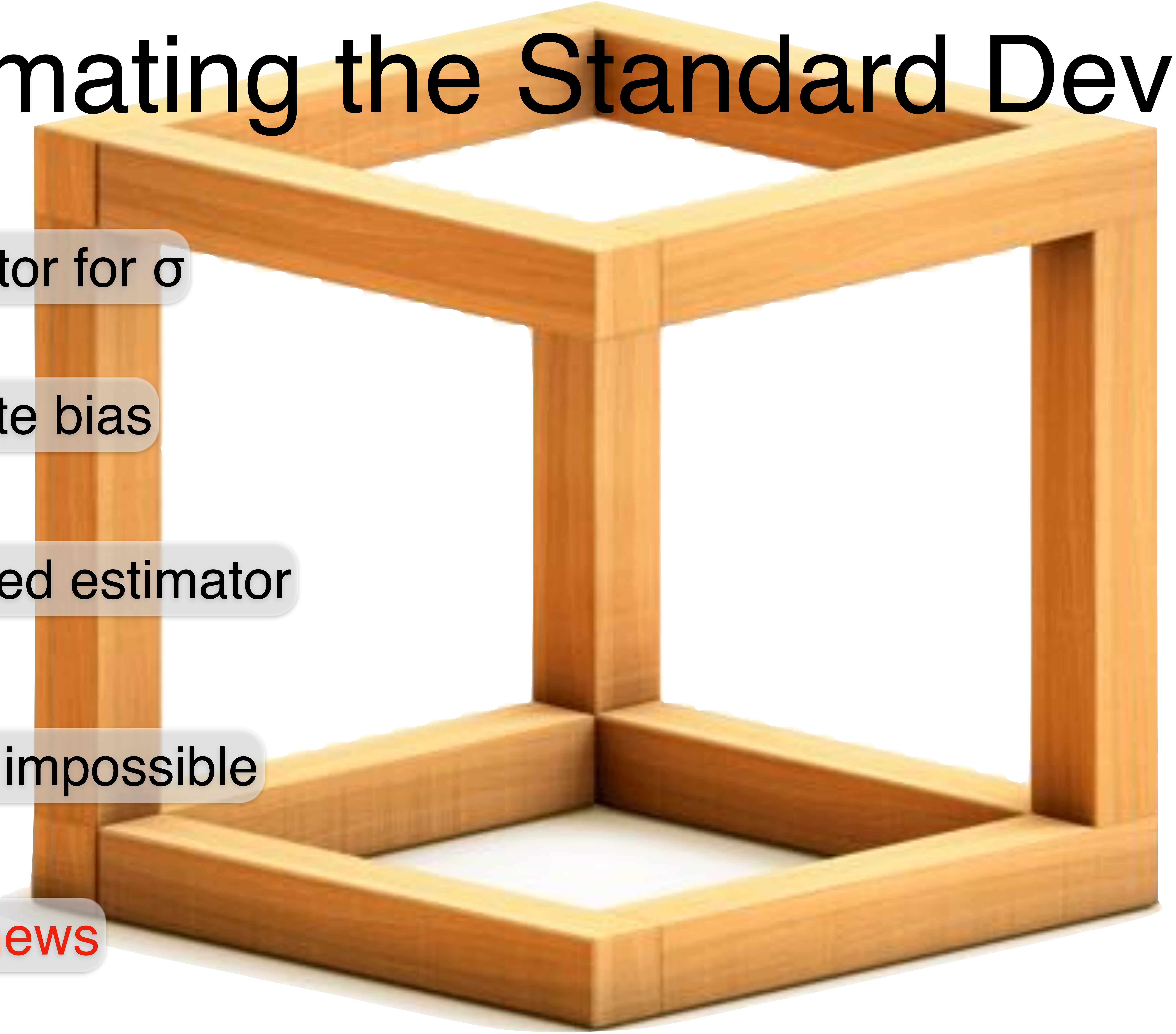
$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Resolve mystery

Dispel half-truth

Estimating σ

Estimating the Standard Deviation



Estimator for σ

Evaluate bias

Unbiased estimator

Easy x impossible

Good news

$\sigma^2 \rightarrow \sigma$

Variance estimator $$S^2 \stackrel{\text{def}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Showed $E(S^2) = \sigma^2$ S^2 is an unbiased estimator for σ^2

Estimating σ $\sigma = \sqrt{\sigma^2}$

$$S \stackrel{\text{def}}{=} +\sqrt{S^2} = +\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Standard Standard-Deviation estimator

Example Evaluation Possible alternatives

ExSample

n = 5   2, 1, 4, 2, 6

Saw
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{2+1+4+2+6}{5} = 3$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1+4+1+1+9}{4} = \frac{16}{4} = 4$$

Estimate for σ²

Estimate for σ

$$S = \sqrt{S^2} = \sqrt{4} = 2$$

Unbiased?

Is S an unbiased estimator for σ?

S² is an unbiased variance estimator

$$E(S^2) = (ES)^2 + V(S) \geq (ES)^2$$

= iff V(S)=0 iff S is a constant

$$(ES)^2 \leq E(S^2) = \sigma^2$$

$$ES \leq \sigma$$

< whenever X is not a constant

On average S underestimates σ

Concrete example

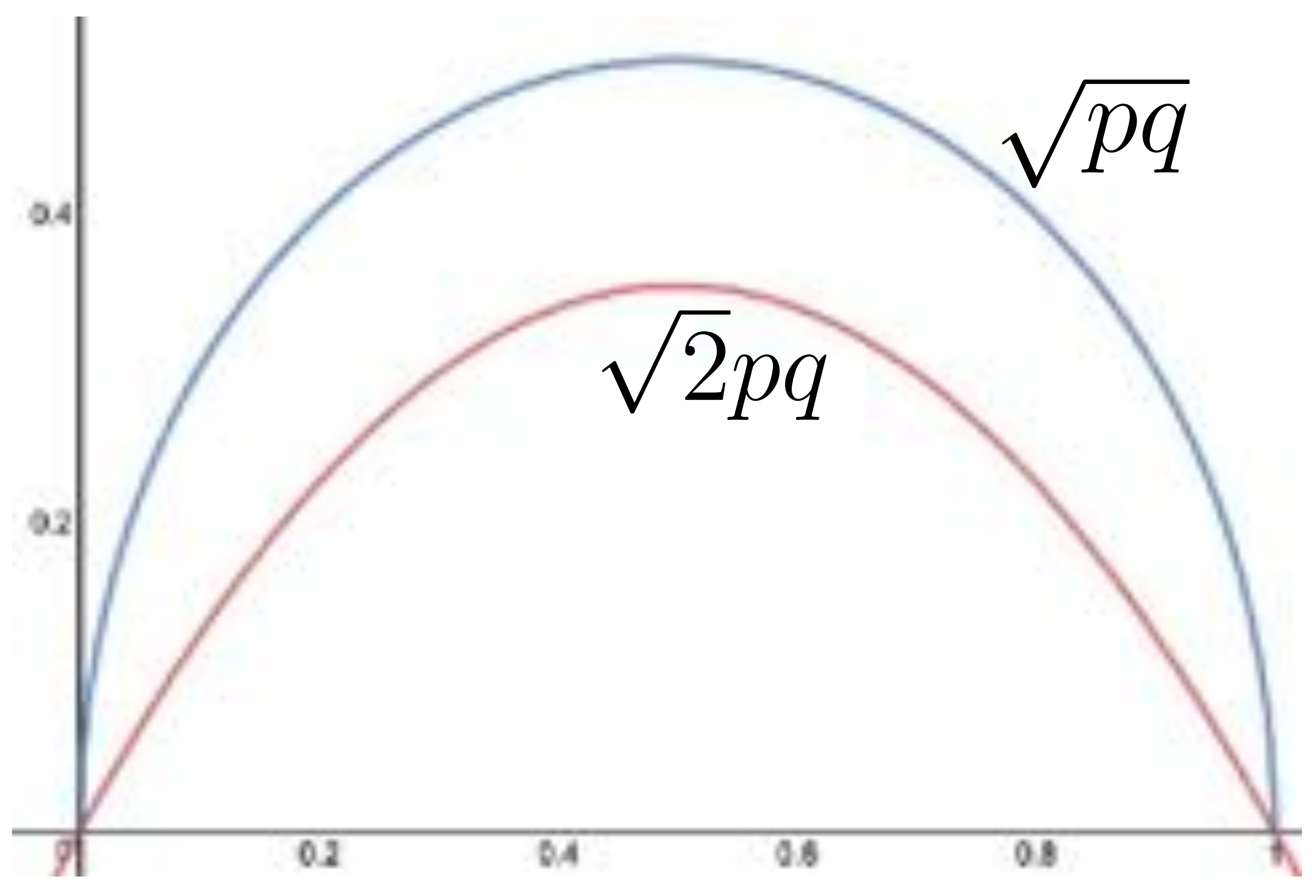
S Strictly Underestimates σ

B_p $\qquad \sigma = \sqrt{p(1-p)} = \sqrt{pq}$

n=2 $\qquad$ Show $\quad E(S) < \sqrt{pq}$

$$S^2 = \frac{1}{1}\left((0-\tfrac{1}{2})^2 + (1-\tfrac{1}{2})^2\right) = \frac{1}{2}$$

x_1,x_2	$P(x_1,x_2)$	$\overline{x}$	s^2	s
0,0	q^2	0	0	0
0,1	qp	½	½	$\frac{1}{\sqrt{2}}$
1,0	pq	½	½	$\frac{1}{\sqrt{2}}$
1,1	p^2	1	0	0



$$E(S) = q^2 \cdot 0 + qp \cdot \frac{1}{\sqrt{2}} + pq \cdot \frac{1}{\sqrt{2}} + p^2 \cdot 0 = \sqrt{2} \cdot pq < \sqrt{pq}$$

Unbiased Estimator for σ?

Is there an unbiased estimator for σ?

If p is known, so is σ, so nothing to estimate

Estimator must work for all distributions

For all p $\quad E(\overline{X}) = \mu \qquad E(S^2) = \sigma^2$

Is there estimator $\hat{\sigma}$ s.t. for all distributions $\quad E(\hat{\sigma}(X^n)) = \sigma$

NO $\quad$ There is no general unbiased estimator for σ !

How do you prove the impossible?

Proof Techniques

- **Proof by obviousness:** "The proof is so clear that it need not be mentioned."
- **Proof by general agreement:** "All in favor?..."
- **Proof by imagination:** "Well, we'll pretend it's true..."
- **Proof by convenience:** "It would be very nice if it were true, so..."
- **Proof by necessity:** "It had better be true, or the entire structure of mathematics would crumble to the ground."
- **Proof by plausibility:** "It sounds good, so it must be true."
- **Proof by intimidation:** "Don't be stupid; of course it's true!"
- **Proof by lack of sufficient time:** "Because of the time constrait, I'll leave the proof to you."
- **Proof by postponement:** "The proof for this is long and arduous, so it is given to you in the appendix."
- **Proof by accident:** "Hey, what have we here?!"
- **Proof by insignificance:** "Who really cares anyway?"
- **Proof by mumbo-jumbo:** $\forall \alpha \in \Phi, \exists \beta \ni \alpha \bullet \beta = \varepsilon, \therefore$ ...
- **Proof by profanity:** (example omitted)
- **Proof by definition:** "We define it to be true."
- **Proof by tautology:** "It's true because it's true."
- **Proof by plagarism:** "As we see on page 289,..."
- **Proof by lost reference:** "I know I saw it somewhere...."
- **Proof by calculus:** "This proof requires calculus, so we'll skip it."
- **Proof by terror:** When intimidation fails...
- **Proof by lack of interest:** "Does anyone really want to see this?"
- **Proof by illegibility:** QED
- **Proof by logic:** "If it is on the problem sheet, it must be true!"
- **Proof by majority rule:** Only to be used if general agreement is impossible.
- **Proof by clever variable choice:** "Let A be the number such that this proof works..."
- **Proof by tessellation:** "This proof is the same as the last."
- **Proof by divine word:** "...And the Lord said, 'Let it be true,' and it was true."
- **Proof by stubbornness:** "I don't care what you say- it is true."
- **Proof by simplification:** "This proof reduced to the statement 1 + 1 = 2."
- **Proof by hasty generalization:** "Well, it works for 17, so it works for all reals."
- **Proof by deception:** "Now everyone turn their backs..."
- **Proof by supplication:** "Oh please, let it be true."
- **Proof by poor analogy:** "Well, it's just like..."
- **Proof by avoidance:** Limit of proof by postponement as it approaches infinity
- **Proof by design:** If it's not true in today's math, invent a new system in which it is.
- **Proof by authority:** "Well, Don Knuth says it's true, so it must be!"
- **Proof by intuition:** "I have this gut feeling."

Handwaving | As you can see…

Induction | True for 1, 2, 3, so must be true

Example | True for this trivial example so must be true

No Unbiased σ Estimator

Even for B_p

p unknown

No unbiased estimator

No unbiased estimators for general distributions

Show for n=2 samples

Similar for any n

How do you prove the impossible?

$\hat{\sigma}$ Any estimator fo σ for B_p distributions

$\hat{\sigma}(x_1, x_2)$ Estimate of σ when observing x_1,x_2 Predetermined constants

$$\begin{aligned} E\left(\hat{\sigma}(X_1, X_2)\right) &= \sum_{x_1,x_2} p(x_1,x_2)\hat{\sigma}(x_1,x_2) \\ &= P(0,0)\hat{\sigma}(0,0) + P(0,1)\hat{\sigma}(0,1) + P(1,0)\hat{\sigma}(1,0) + P(1,1)\hat{\sigma}(1,1) \\ &= (1-p)^2\hat{\sigma}(0,0) + (1-p)p\hat{\sigma}(0,1) + p(1-p)\hat{\sigma}(1,0) + p^2\hat{\sigma}(1,1) \end{aligned}$$

Polynomial in p degree-2 polynomial

$\sigma = \sqrt{p(1-p)}$ Not a polynomial in p

The two functions differ For some p $E(\hat{\sigma}(X_1, X_2)) \neq \sigma$

Impossibility

How did we prove the impossible? Easily!

Estimators for B_p

Showed that for any estimator $\hat{\sigma}$ $E(\hat{\sigma}(X_1, X_2))$ polynomial in p

$\sigma = \sqrt{p(1-p)}$ not polynomial in p

Except: How do you prove?

For some p $E(\hat{\sigma}(X_1, X_2)) \neq \sigma$ "Don't be stupid; of course it's true!"

Therefore $\hat{\sigma}$ not unbiased

Despite joke Complete proof Give up?

Good News

Bias not so bad

Provides more freedom

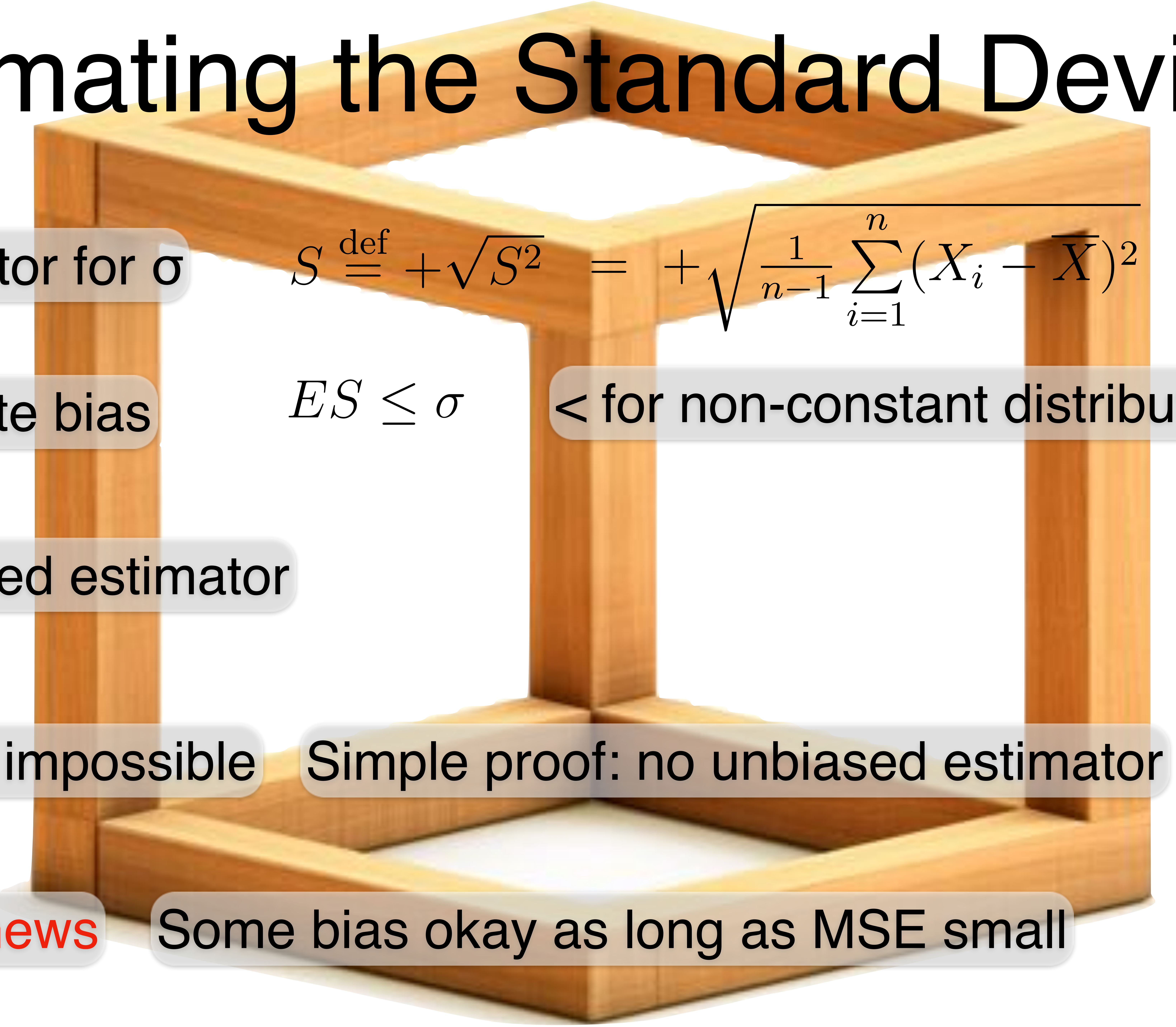
Best estimator (MSE) often biased

As the number of samples n increases

$S \to \sigma$

Consistent

Estimating the Standard Deviation



Estimator for σ $\quad S \overset{\text{def}}{=} +\sqrt{S^2} = +\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

Evaluate bias $\quad ES \le \sigma \quad$ < for non-constant distributions

Unbiased estimator

Easy x impossible $\quad$ Simple proof: no unbiased estimator

Good news $\quad$ Some bias okay as long as MSE small