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

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

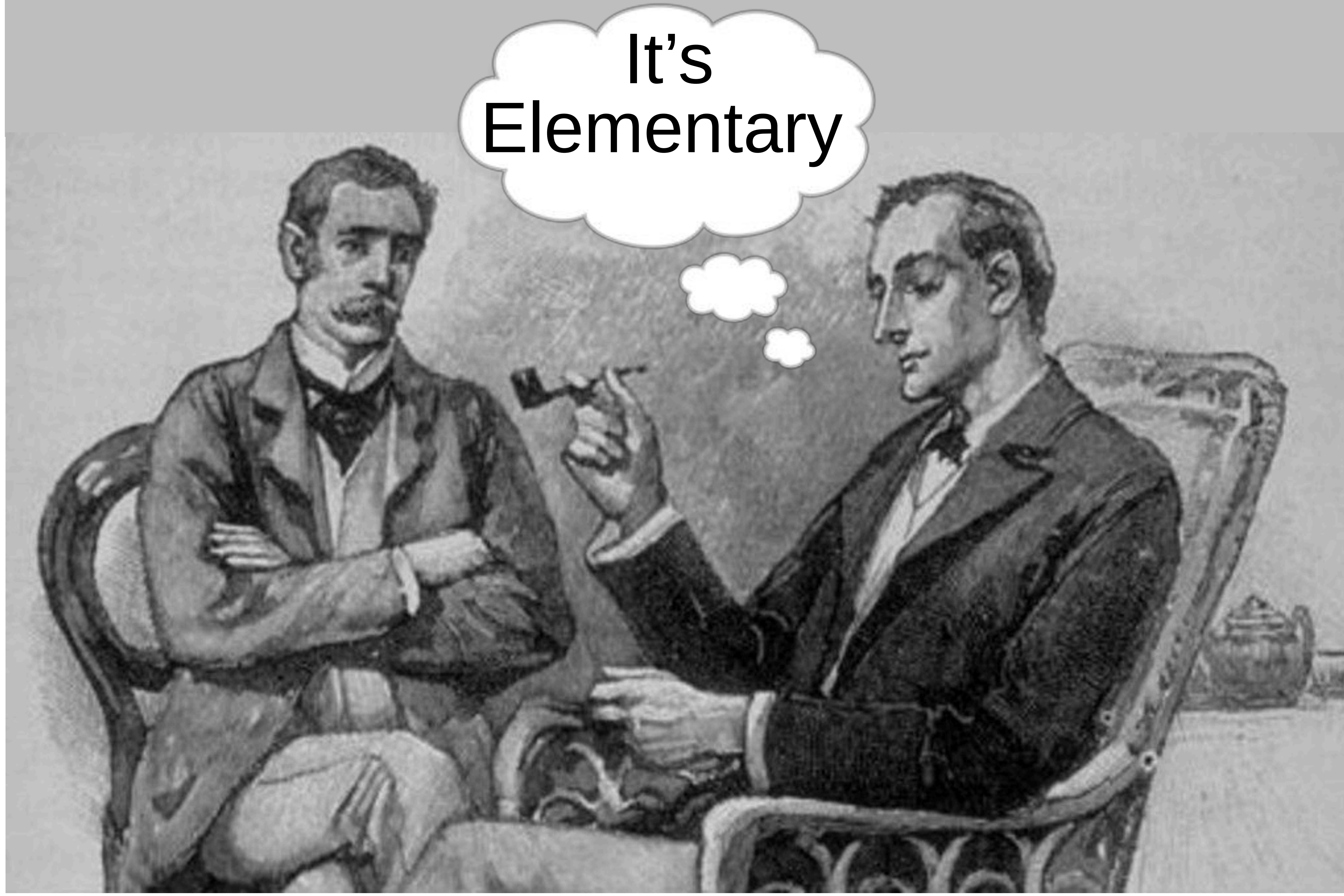
$$\text{“}S^2\text{”} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$ $\overline{X}$ is exact average Lower sum

“S^2” under-estimates σ^2

Explains

Not $\frac{n-1}{n}$ Nor capture whole reason complex

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



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

$$
\begin{aligned}
E(\text{“}S^2\text{”}) &= E(X_1 - \overline{X})^2 \\
&= \sum_{x_1, x_2} p(x_1, x_2) \cdot (x_1 - \overline{x})^2 \\
&= 2 \cdot pq \cdot \frac{1}{4} = \frac{1}{2} pq = \frac{1}{2}\sigma^2
\end{aligned}
$$ ✅

Simpler Easier to analyze

Simplified Formulation

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

Asymmetric, unclear

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

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

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

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

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

$$E(X_1 - \overline{X})^2 = \underbrace{\frac{1}{4}}_{\text{gain ¼ from proximity}} \cdot E(X_1 - X_2)^2 = \frac{1}{4} \cdot \underbrace{2}_{\text{lose 2 for randomness}} \cdot E(X_1 - \mu)^2 = \frac{\sigma^2}{2}$$

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

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

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

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

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

$$= E\left(\frac{X_1-X_2}{2}\right)^2$$

$$\overset{\text{LoE}}{=} \frac{1}{4}\cdot E(X_1-X_2)^2$$

¼ from $\overline{X}$ being closer than μ to X_1

$$\overset{\text{0-mean}}{=} \frac{1}{4}\cdot V(X_1-X_2)$$

$$\overset{\perp\!\!\!\perp}{=} \frac{1}{4}\cdot (V(X_1)+V(X_2))$$

$$\overset{\text{iid}}{=} \frac{1}{4}\cdot 2\cdot V(X_1)$$

2 from $\overline{X}$ being random

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

1/2 together

General n

$$
\begin{aligned}
E(\text{“}S^2\text{”}) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\
&= E(X_1 - \overline{X})^2 \\
&= E\left(\frac{n-1}{n}\left(X_1 - \frac{X_2+\ldots+X_n}{n-1}\right)\right)^2 \\
&\overset{\text{LoE}}{=} \left(\frac{n-1}{n}\right)^2 \cdot E\left(X_1 - \frac{X_2+\ldots+X_n}{n-1}\right)^2 \\
&\overset{\text{0-mean}}{=} \left(\frac{n-1}{n}\right)^2 \cdot V\left(X_1 - \frac{X_2+\ldots+X_n}{n-1}\right) \\
&\overset{\perp\!\!\!\perp}{=} \left(\frac{n-1}{n}\right)^2 \cdot \left[V(X_1) + V\left(\frac{X_2+\ldots+X_n}{n-1}\right)\right] \\
&\overset{\text{iid, var. scaling}}{=} \left(\frac{n-1}{n}\right)^2 \cdot \left[\sigma^2 + \frac{\sigma^2}{n-1}\right] \\
&= \left(\frac{n-1}{n}\right)^2 \cdot \frac{n}{n-1} \cdot \sigma^2 \\
&= \frac{n-1}{n} \cdot \sigma^2
\end{aligned}
$$

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

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

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

$\frac{n-1}{n}$ 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 $\times \frac{5}{4}$ $\boxed{\times \frac{n}{n-1}}$

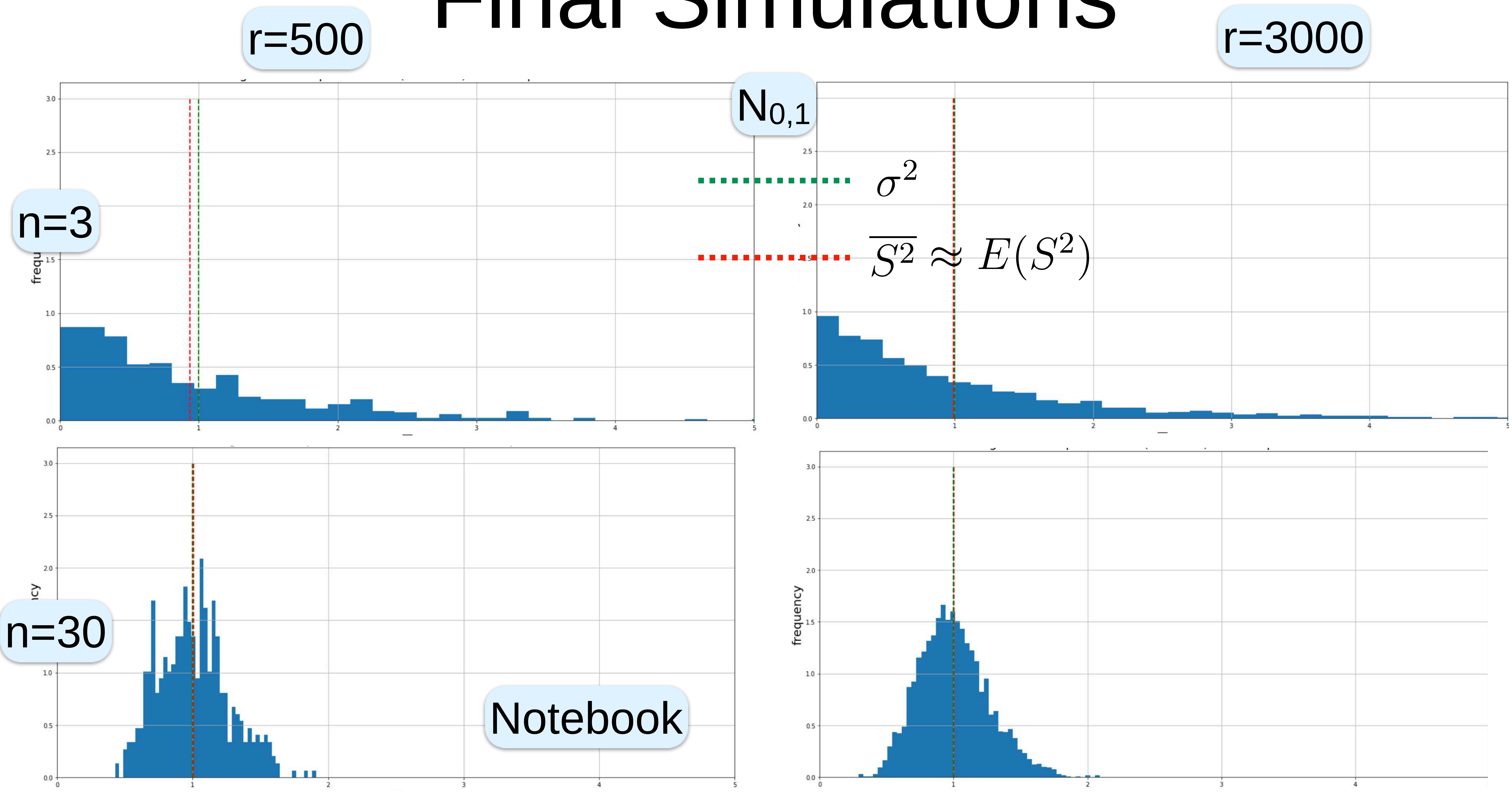
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1+4+1+1+9}{4} = \frac{16}{4} = 4$$

Unbiased estimate of σ²

One-pass calculation

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

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