CSE 312

Foundations of Computing II

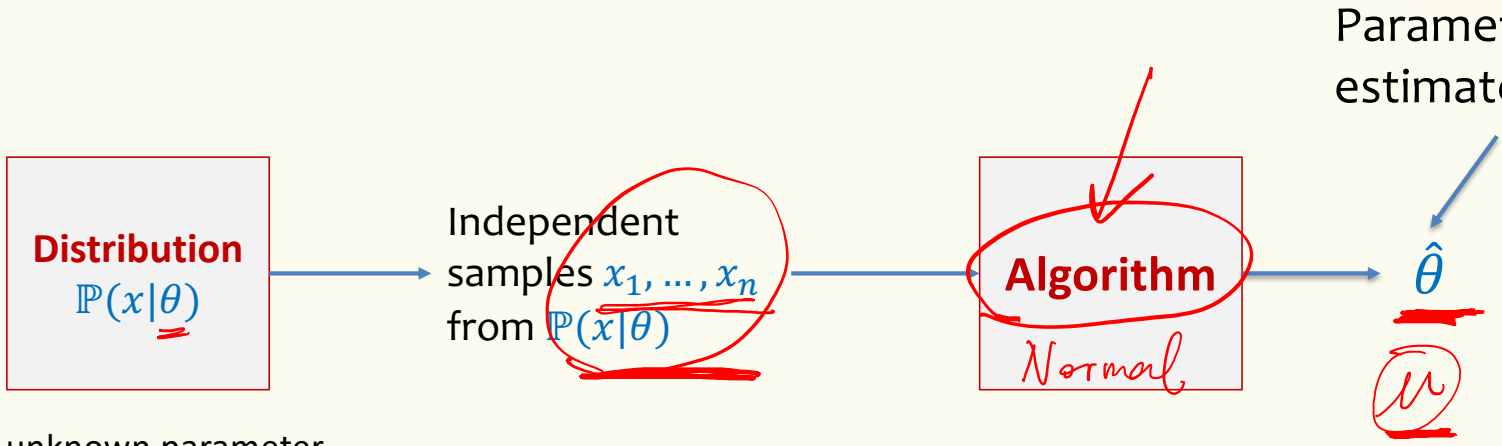
Lecture 25: Maximum Likelihood Estimation (MLE) Cont'd

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Rachel Lin, Hunter Schafer

Slide Credit: Based on Stefano Tessaro's slides for 312 19au
incorporating ideas from Alex Tsun's and Anna Karlin's slides for 312 20su and 20au

Parameter Estimation – Workflow



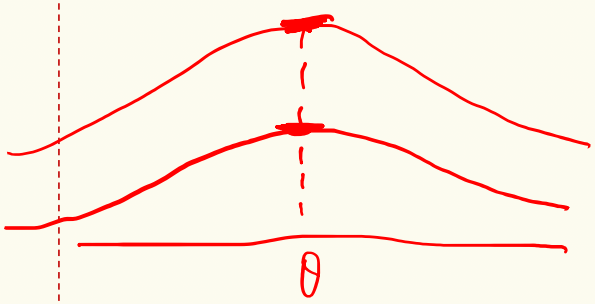
θ = unknown parameter

Maximum Likelihood Estimation (MLE). Given data $x_1, \ldots, x_n$, find $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ ("the MLE") such that $L(x_1, \ldots, x_n|\hat{\theta})$ is maximized!

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(x_1, \ldots, x_n|\theta)$$

Likelihood of Different Observations

Definition. The **likelihood** of independent observations $x_1, \ldots, x_n$ is

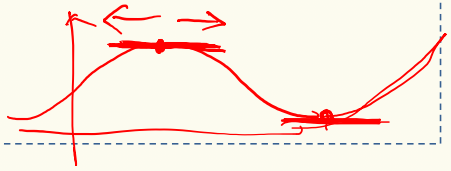


$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \mathbb{P}(x_i; \theta) \quad \text{(Discrete case)}$$

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i; \theta) \quad \text{(continuous case)}$$

Maximum Likelihood Estimation (MLE). Given data $x_1, \ldots, x_n$, find $\hat{\theta}$ ("the MLE") of model such that $L(x_1, \ldots, x_n|\hat{\theta})$ is maximized!

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(x_1, \ldots, x_n|\theta)$$

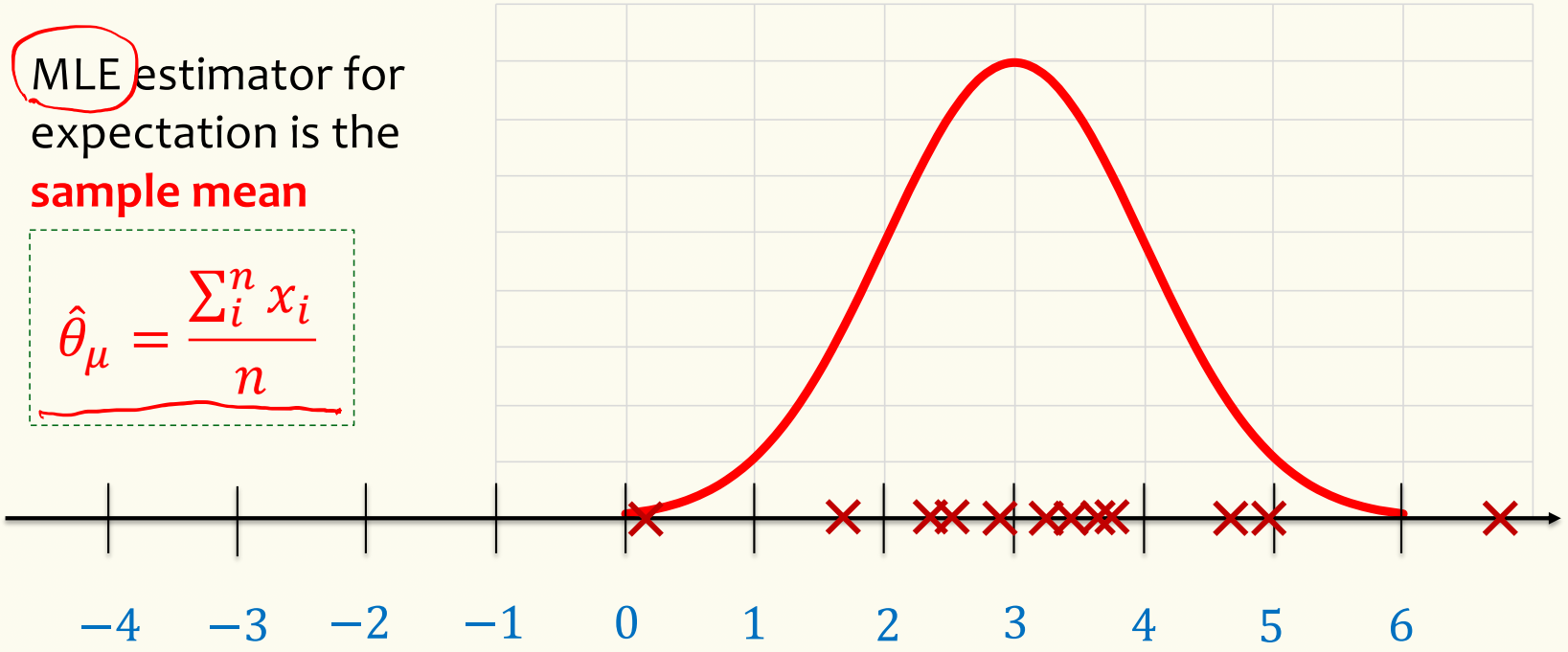


Usually: Solve $\frac{\partial L(x_1, \ldots, x_n|\theta)}{\partial \theta} = 0$ or $\frac{\partial \ln L(x_1, \ldots, x_n|\theta)}{\partial \theta} = 0$ [+check it's a max!]

n samples $x_1, \dots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely μ?

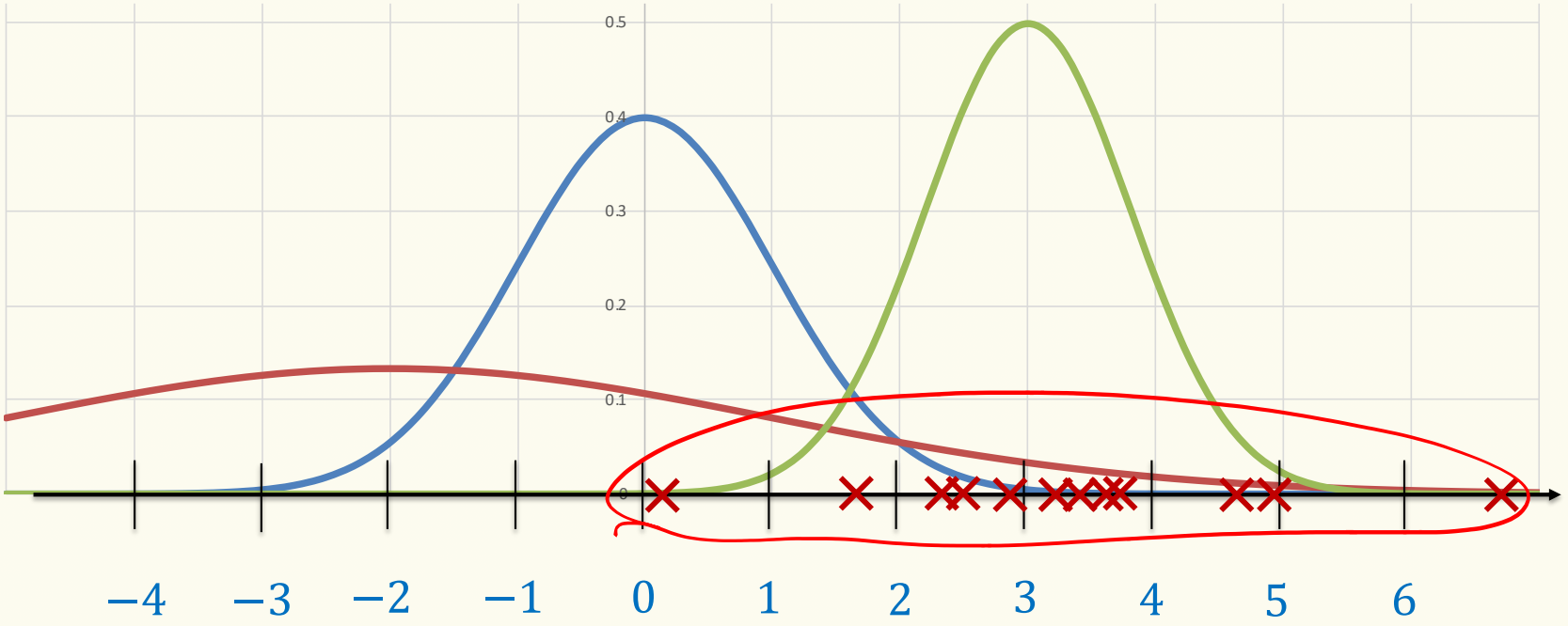
MLE estimator for expectation is the **sample mean**

$$\hat{\theta}_\mu = \frac{\sum_i^n x_i}{n}$$



Next: n samples $x_1, \ldots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, \sigma^2)$.

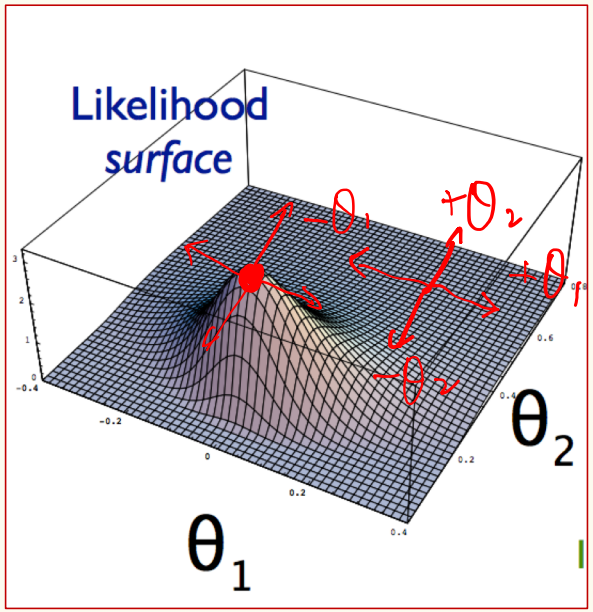
Most likely μ and σ^2?



Two-parameter optimization

Normal outcomes $x_1, \ldots, x_n$

Goal: estimate $\theta_1 = \mu$ = expectation and $\theta_2 = \sigma^2$ = variance



$$L(x_1, \ldots, x_n | \theta_1, \theta_2) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i - \theta_1)^2}{2\theta_2}}$$

$$\ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = -n\frac{\ln(2\pi\,\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

Two-parameter estimation

$$\ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = -\frac{\ln(2\pi\,\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

We need to find a solution $\hat{\theta}_1, \hat{\theta}_2$ to

- $$\frac{\partial}{\partial \theta_1} \ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = 0$$
- $$\frac{\partial}{\partial \theta_2} \ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = 0$$

MLE for Expectation

$$\ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = -n \frac{\ln(2\pi\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_1} \ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = \frac{1}{\theta_2} \sum_{i}^{n} (x_i - \theta_1) = 0$$

$$\hat{\theta}_1 = \frac{\sum_i^n x_i}{n}$$

In other words, MLE of expectation is (again) the *sample mean* of the data, regardless of θ_2

What about the variance?

MLE for Variance

$$\ln L(x_1, \ldots, x_n | \hat{\theta}_1, \theta_2) = -n \frac{\ln(2\pi\,\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \hat{\theta}_1)^2}{2\theta_2}$$

$$= -n \frac{\ln 2\pi}{2} - n \frac{\ln \theta_2}{2} - \frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2$$

$$\frac{\partial}{\partial \theta_2} \ln L(x_1, \ldots, x_n | \hat{\theta}_1, \theta_2) = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2 = 0$$

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2$$

MLE of variance is the *population variance* of the data.

(Note that this is not called sample variance!)

Summary: MLE of Normal Distribution

Definition. The **likelihood** of independent observations $x_1, \dots, x_n$ is

$$L(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \quad \text{(continuous case)}$$

Normal outcomes $x_1, \dots, x_n$

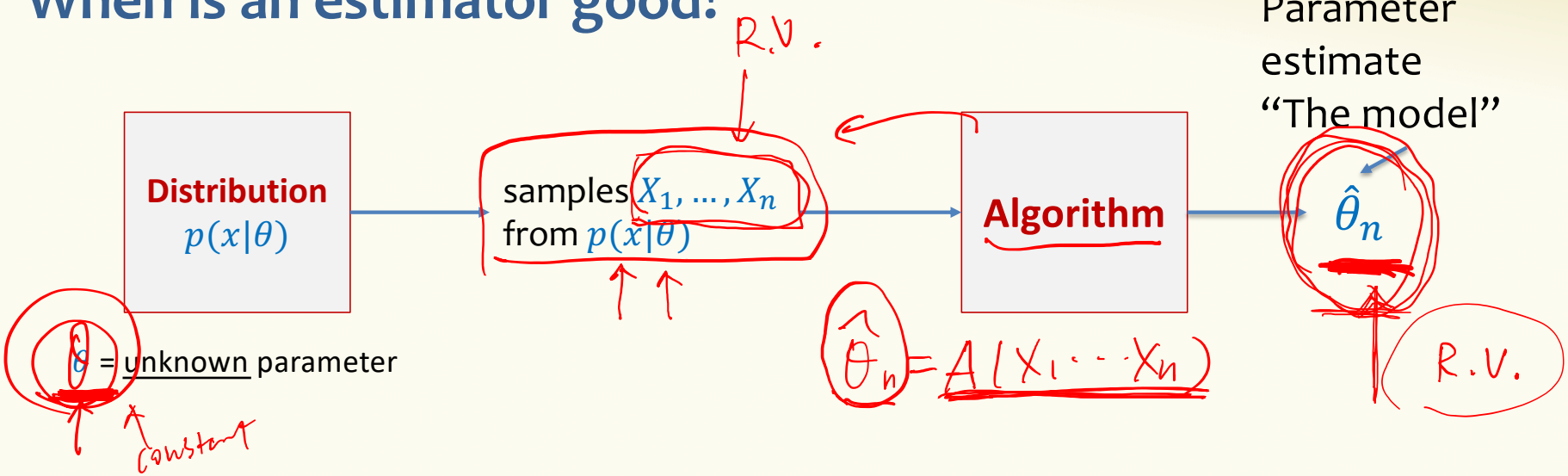
$$\hat{\theta}_\mu = \frac{\sum_i^n x_i}{n}$$

MLE estimator for **expectation**

$$\hat{\theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\theta}_\mu\right)^2$$

MLE estimator for **variance**

When is an estimator good?



Definition. An estimator of parameter θ is an **unbiased estimator**

$$\mathbb{E}(\hat{\theta}_n) = \theta.$$

MLE $\frac{\partial}{\partial \theta} L(x_1 \cdots x_n | \theta) = 0 \rightarrow \hat{\theta}$

constant variable

Example – Coin Flips

Recall: $\hat{\theta}_{\mu} = \frac{n_H}{n}$

Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

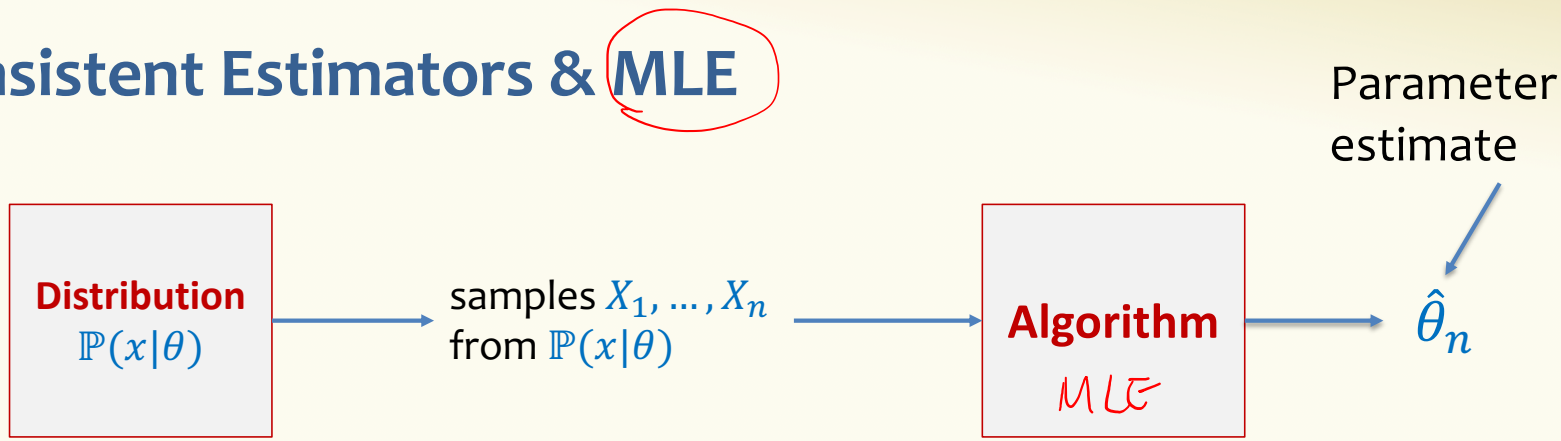
Fact. $\hat{\theta}_{\mu}$ is unbiased

i.e., $\mathbb{E}(\hat{\theta}_{\mu}) = p$, where p is the probability that the coin turns out head.

$$E\left(\frac{n_H}{n}\right) = E\left(\frac{\Sigma x_i}{n}\right) = \frac{E(\Sigma X_i)}{n} = \frac{n \cdot p}{n} = p.$$

Is MLE always an unbiased estimator? NOT!

Consistent Estimators & MLE



θ = unknown parameter

Definition. An estimator is **unbiased** if $\mathbb{E}(\hat{\theta}_n) = \theta$ for all $n \geq 1$.

Definition. An estimator is **consistent** if $\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta$.

Theorem. MLE estimators are consistent.

(But not necessarily unbiased)

Example – Consistency

Normal outcomes $X_1, \ldots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

natural

$$\widehat{\Theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

Population variance – Biased!

strange

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

Sample variance – Unbiased!

$$S_n^2 = \left(\frac{n}{n-1}\right)\widehat{\Theta}_{\sigma^2}$$

$\widehat{\Theta}_{\sigma^2}$ converges to same value as S_n^2, i.e., σ^2, as $n \to \infty$.

$\widehat{\Theta}_{\sigma^2}$ and S_n^2 are "consistent"

Only S_n^2 unbiased

Why is the estimator consistent, but biased?

linearity

$$\mathbb{E}(\hat{\Theta}_{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \hat{\Theta}_{\mu}\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2 - \frac{2}{n}X_i\sum_{j=1}^{n}X_j + \frac{1}{n^2}\sum_{j=1}^{n}X_j\sum_{k=1}^{n}X_k\right]$$

…

Why is the estimator consistent, but biased?

linearity

$$\mathbb{E}(\widehat{\Theta}_{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \widehat{\Theta}_{\mu}\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right]$$

…

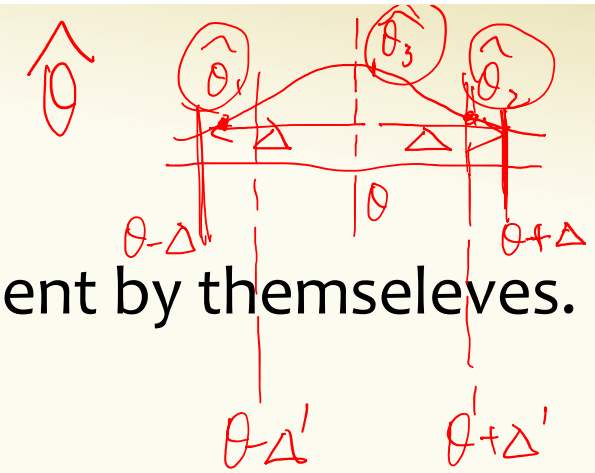
$$= \left(1 - \frac{1}{n}\right)\sigma^2 \qquad \to \sigma^2 \text{ for } n \to \infty$$

$\times \left(\frac{n}{n-1}\right)$

Therefore: $\mathbb{E}(S_n^2) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \widehat{\Theta}_{\mu}\right)^2\right] = \frac{n}{n-1}\mathbb{E}\left(\widehat{\Theta}_{\sigma^2}\right) = \sigma^2$

Bessel's correction

Estimation – Confidence Intervals



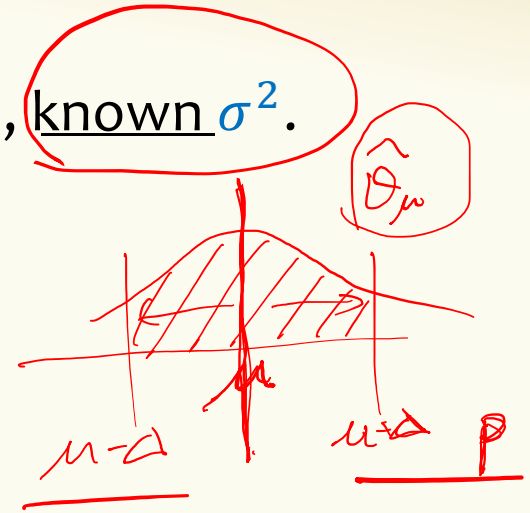
Unbiasedness/consistency are not sufficient by themseleves.

- We want $\mathbb{P}(\widehat{\Theta}_n = \theta) = 1$
  - At least as $n \to \infty$
  - Note that $\widehat{\Theta}_n$ is continuous for Gaussian, so $\mathbb{P}(\widehat{\Theta}_n = \theta) = 0$
- Relaxation: Find smallest Δ such that $\mathbb{P}\left(\left|\widehat{\Theta}_n - \theta\right| \leq \Delta\right) \geq p$ for a given p = e.g. 95% = 0.95
  - We say that Δ gives us the **p-confidence interval**
  - e.g., 95%-confidence interval means $p = 0.95$

Mean Estimator for Normal – Known Variance

Normal outcomes: $X_1, \dots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$, known σ^2.

$$\widehat{\Theta}_\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$



Q: which distribution?

A: Normal!

- Expectation $\frac{1}{n}(n \cdot \mu) = \mu$
- variance $\frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$

Therefore: $\frac{\widehat{\Theta}_\mu - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$

Mean Estimator for Normal – Known Variance

Normal outcomes: $X_1, \dots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$, known σ^2.

$$\widehat{\Theta}_\mu = \frac{\sum_{i=1}^n X_i}{n} \qquad \frac{\widehat{\Theta}_\mu - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$\mathbb{P}\left(-z < \frac{\widehat{\Theta}_\mu - \mu}{\sigma/\sqrt{n}} < z\right) = \Phi(z) - \Phi(-z) = 1 - 2\Phi(-z)$$

Equivalently: $\mathbb{P}\left(\left|\widehat{\Theta}_\mu - \mu\right| < z\sigma/\sqrt{n}\right) = 1 - 2\Phi(-z) = 0.95$

E.g., $\Phi(-1.96) \approx 5\%$ → Estimate is within $\Delta = 1.96\sigma/\sqrt{n}$ of μ with probability $\approx 95\%$ (i.e., "Δ is the 95%-confidence interval")

$[\mu - \Delta, \mu + \Delta]$

Mean Estimator for Normal – Unknown Variance

Normal outcomes: $X_1, \ldots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$, unknown σ^2.

$$\mathbb{P}\left(\left|\widehat{\Theta}_\mu - \mu\right| < \frac{z\sigma}{\sqrt{n}}\right) = 1 - 2\Phi(-z)$$

- Still true, but not that useful, as we cannot evaluate σ
- What about using $S_n = \sqrt{S_n^2}$ instead?

$$\mathbb{P}\left(\left|\widehat{\Theta}_\mu - \mu\right| < \frac{zS_n}{\sqrt{n}}\right) = 1 - 2\Phi(-z) \text{ ?}$$

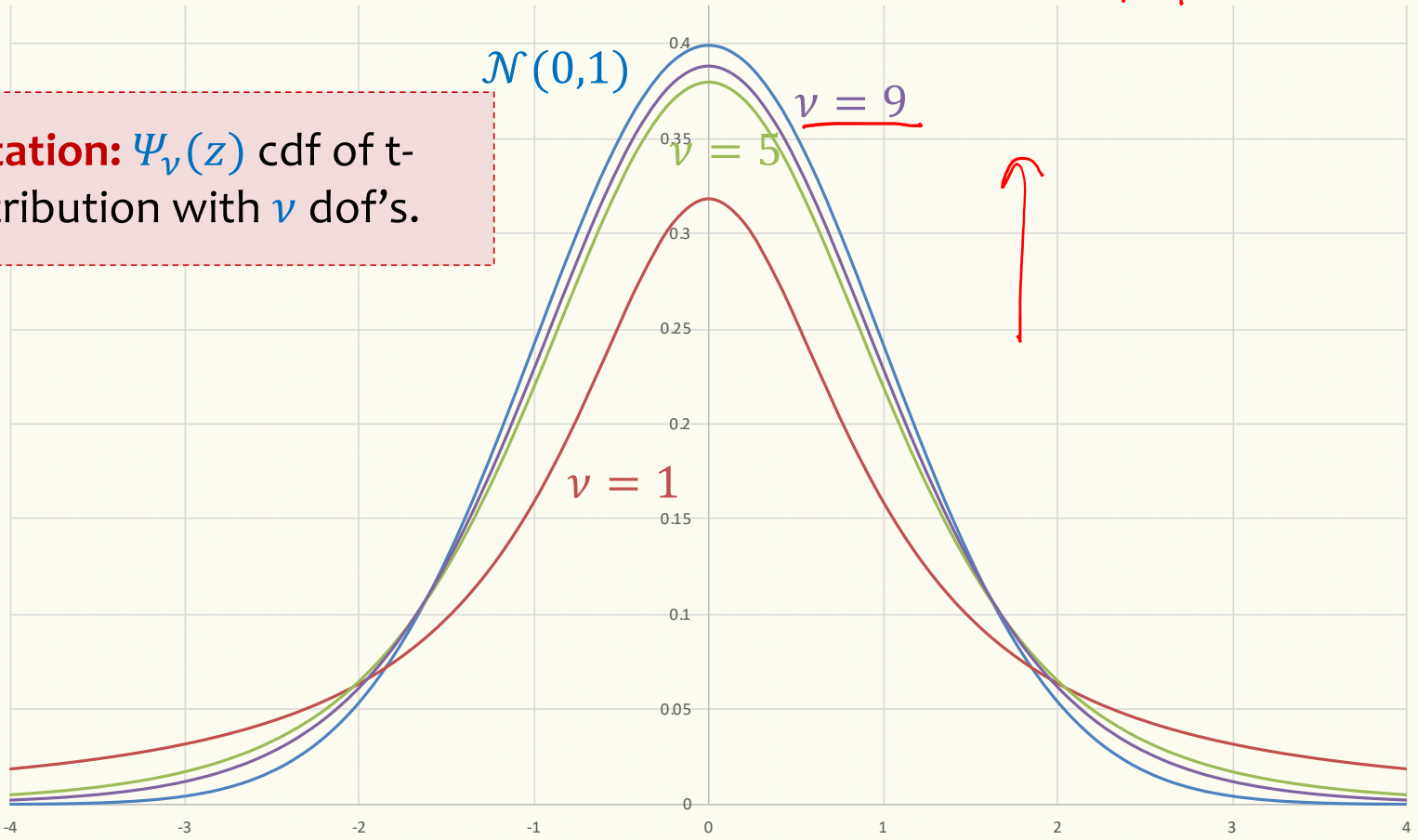
- Not true!

$\frac{\widehat{\Theta}_\mu - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ $\quad$ $\frac{\widehat{\Theta}_\mu - \mu}{S_n/\sqrt{n}} \sim$ t-distribution with $n-1$ degrees of freedom

Student's t-Distribution

Parametrized by ν = degrees of freedom

Notation: $\Psi_\nu(z)$ cdf of t-distribution with ν dof's.



Student,"The probable error of a mean". Biometrika 1908.

Student?

”Student” was a pseudonym for **William Gosset**

- Worked for A. Guinness & Son
- Investigated e.g. brewing and barley yields
- Wasn’t allowed to publish with real name

Source: Wikipedia

Mean Estimator for Normal – Unknown Variance

Therefore: $\mathbb{P}\left(\left|\widehat{\Theta}_\mu - \mu\right| < zS_n/\sqrt{n}\right) = 1 - 2\Psi_{n-1}(-z)$

E.g., $\Psi_9^{-1}(0.05) \approx 2.26$ → Estimate is within $2.26S_n/\sqrt{n}$ of μ with probability $\approx 95\%$