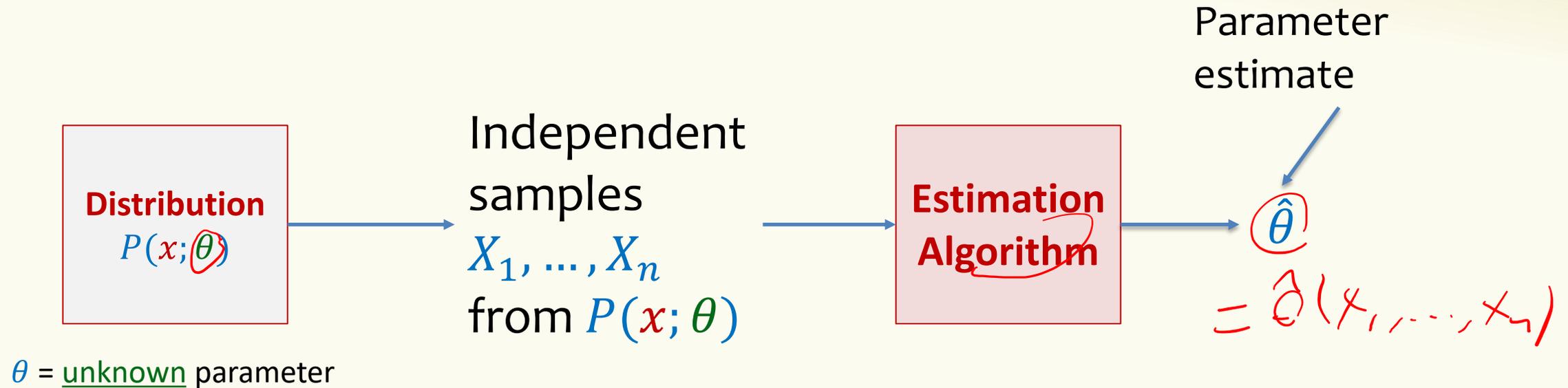


CSE 312

Foundations of Computing II

**Lecture 23: Maximum Likelihood Estimation
Continued**

Review Parameter Estimation – Workflow



Example: coin flip distribution with unknown $\theta =$ probability of heads

Observation: *HTTHHHTHTTTTHTTTTTHT*

Goal: Estimate θ

Review Likelihood of Different Observations

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

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

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

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

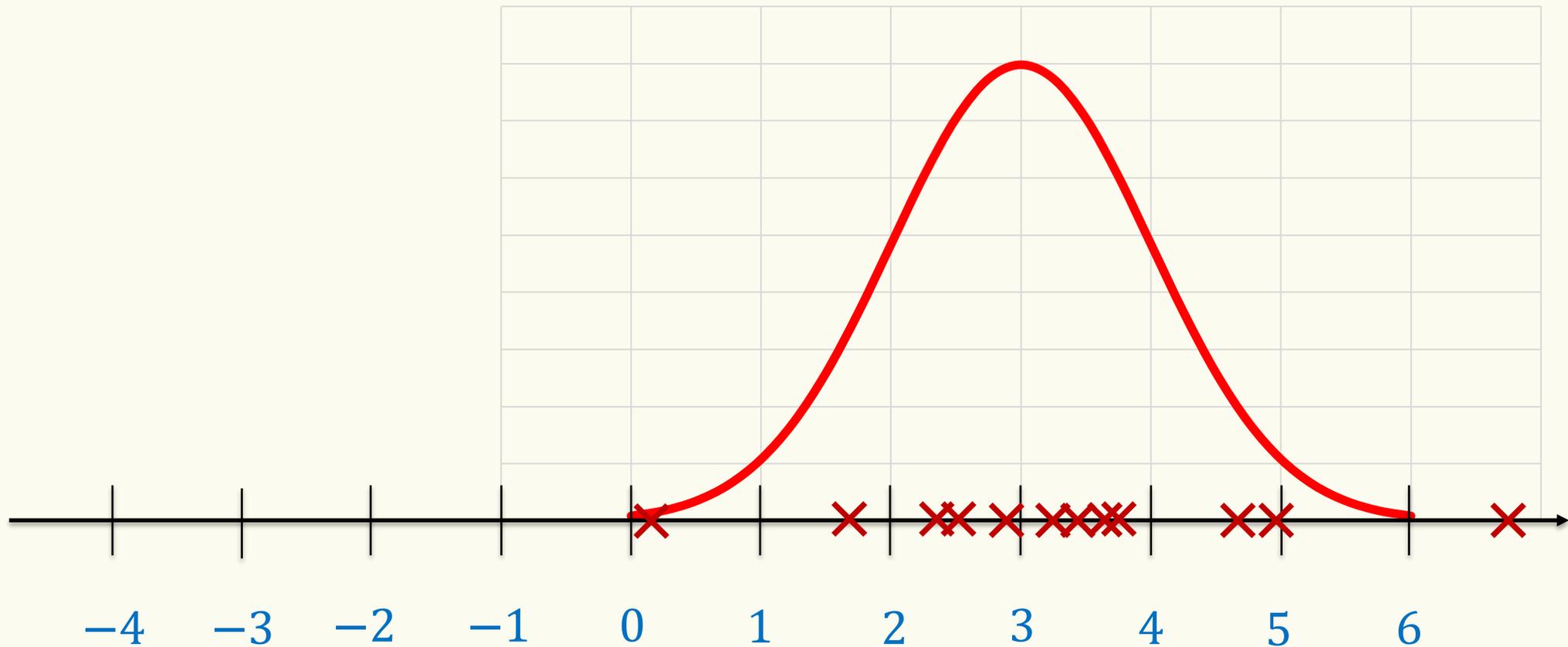
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(x_1, \dots, x_n | \theta)$$

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

Agenda

- MLE for Normal Distribution ◀
- Unbiased and Consistent Estimators
- Odds and ends

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



Example – Gaussian Parameters

Normal outcomes x_1, \dots, x_n , known variance $\sigma^2 = 1$

Goal: estimate θ = expectation

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

Handwritten notes: $(2\pi)^{-1/2}$, $\ln(a^b) = b \ln(a)$, $\ln(ab) = \ln(a) + \ln(b)$

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

Goal: estimate $\theta =$ expectation

Example – Gaussian Parameters

Normal outcomes x_1, \dots, x_n , known variance $\sigma^2 = 1$

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

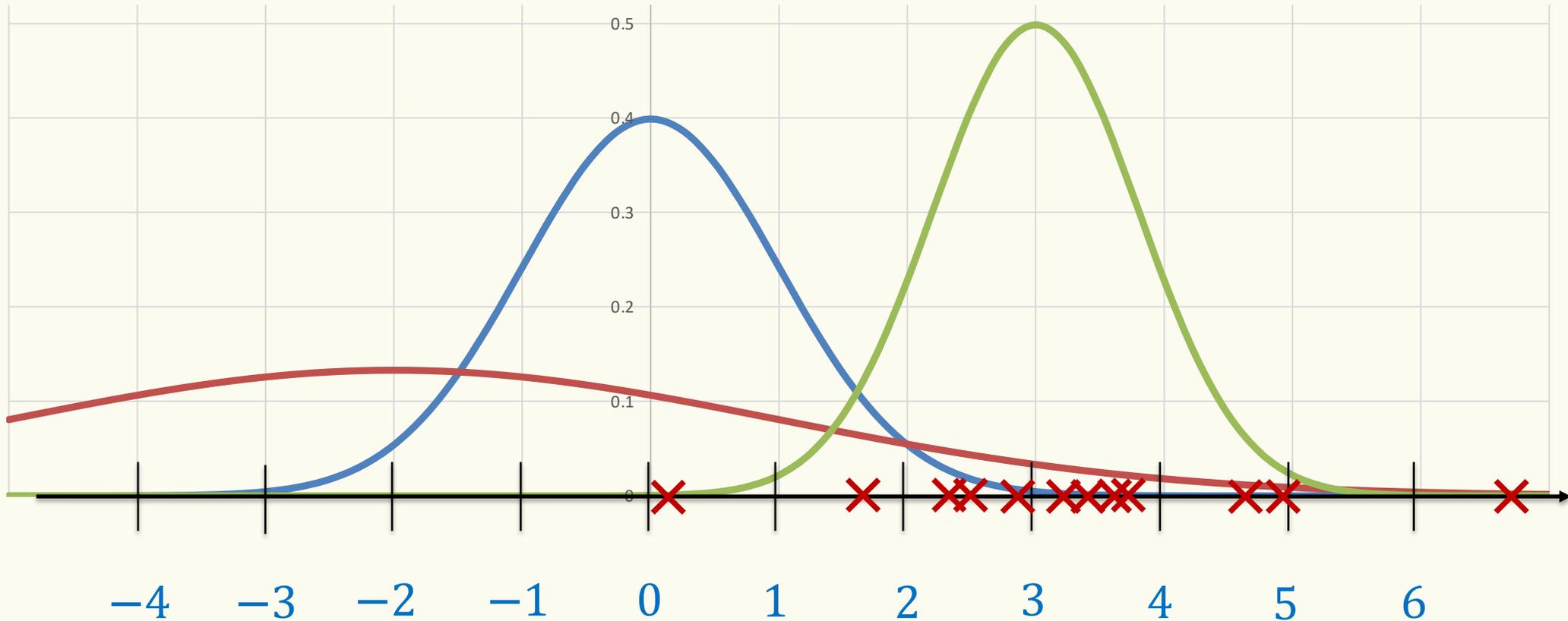
Note: $\frac{\partial}{\partial \theta} \frac{(x_i - \theta)^2}{2} = \frac{1}{2} \cdot 2 \cdot (x_i - \theta) \cdot (-1) = \theta - x_i$

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \dots, x_n | \theta) = \sum_{i=1}^n (x_i - \theta) = \sum_{i=1}^n x_i - n\theta = 0$$

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n}$$

In other words, MLE is the *sample mean* of the data.

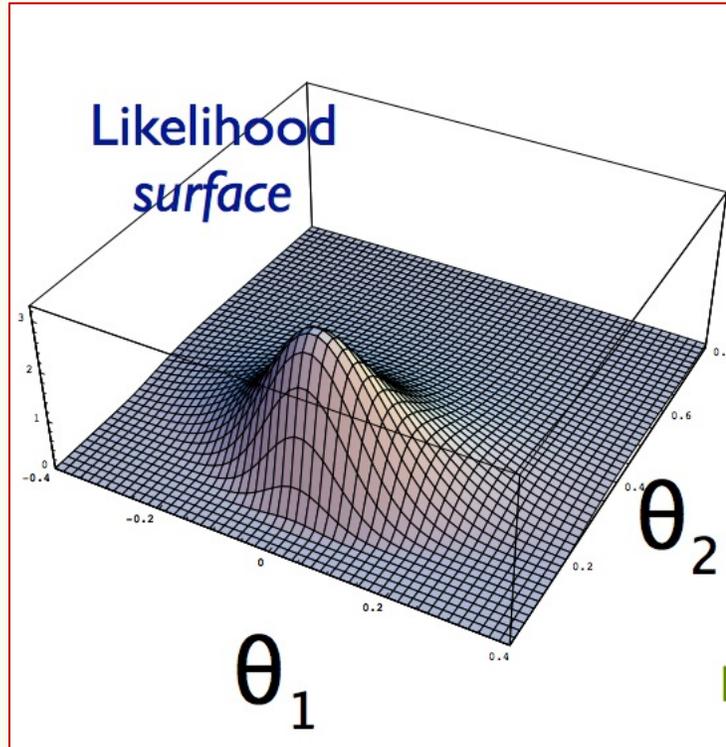
Next: n samples $x_1, \dots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, \sigma^2)$.
Most likely μ and σ^2 ?



Two-parameter optimization

Normal outcomes x_1, \dots, x_n

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



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

Handwritten red annotations: μ and σ^2 above θ_1 and θ_2 respectively. Red circles around θ_2 in the denominator of the exponent and the square root term.

$$\ln \mathcal{L}(x_1, \dots, x_n | \theta_1, \theta_2) =$$

$$= -n \frac{\ln(2\pi\theta_2)}{2} - \sum_{i=1}^n \frac{(x_i - \theta_1)^2}{2\theta_2}$$

Two-parameter estimation

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

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

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

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

MLE for Expectation

$$\ln \mathcal{L}(x_1, \dots, x_n | \theta_1, \theta_2) = \underbrace{-n \frac{\ln(2\pi \theta_2)}{2}}_{\text{?}} - \underbrace{\sum_{i=1}^n \frac{(x_i - \theta_1)^2}{2\theta_2}}_{\text{?}}$$

$$\frac{\partial}{\partial \theta_1} \ln \mathcal{L}(x_1, \dots, x_n | \theta_1, \theta_2) = \frac{1}{\theta_2} \sum_i (x_i - \theta_1) = 0$$

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

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

$$l_n(\mu, \sigma^2) = l_n(\mu) + l_n(\sigma^2)$$

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

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

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

$$\frac{\partial}{\partial \theta_2} \left(\frac{\ln \theta_2}{2} \right)$$

$$= \frac{1}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_2} \frac{1}{\theta_2}$$

$$\frac{\partial}{\partial x} x^{-u} = -u x^{-(u+1)}$$

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

In other words, MLE of variance is the *population variance* of the data.

(Note that this is not called sample variance!)

$$-u + \frac{1}{\theta_2} \sum_{i=1}^n (x_i - \hat{\theta}_1)^2 = 0$$

Likelihood – Continuous Case

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

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

Normal outcomes x_1, \dots, x_n

$$\hat{\theta}_\mu = \frac{\sum_{i=1}^n x_i}{n}$$

MLE estimator for
expectation

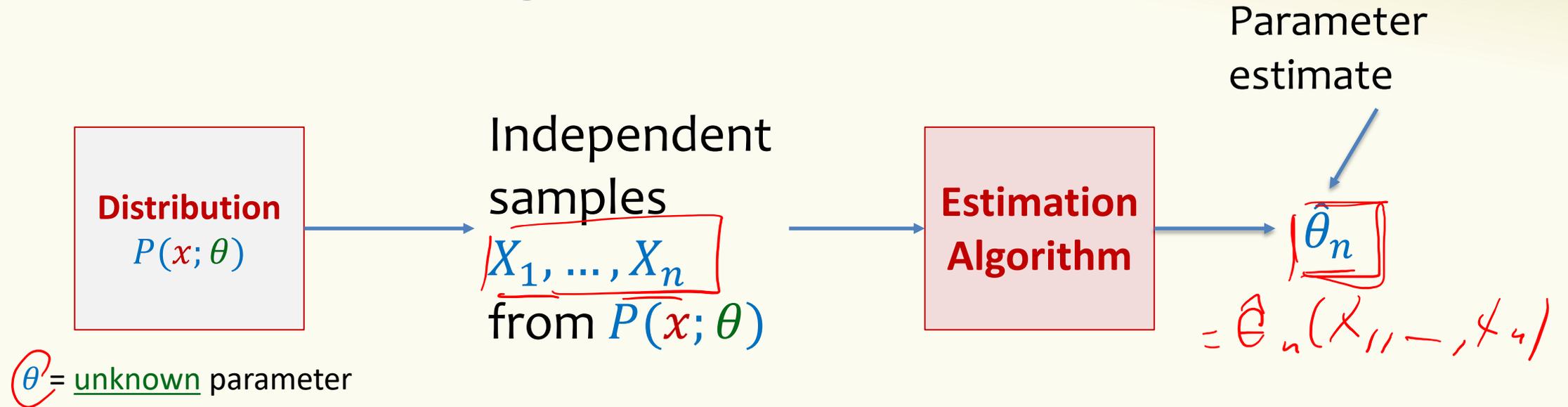
$$\hat{\theta}_{\sigma^2} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\theta}_\mu)^2$$

MLE estimator for
variance

Agenda

- MLE for Normal Distribution
- Unbiased and Consistent Estimators ◀
- Intuition and Bigger Picture

When is an estimator good?



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

$$\mathbb{E}[\hat{\theta}_n] = \theta.$$

Example – Coin Flips

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

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

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

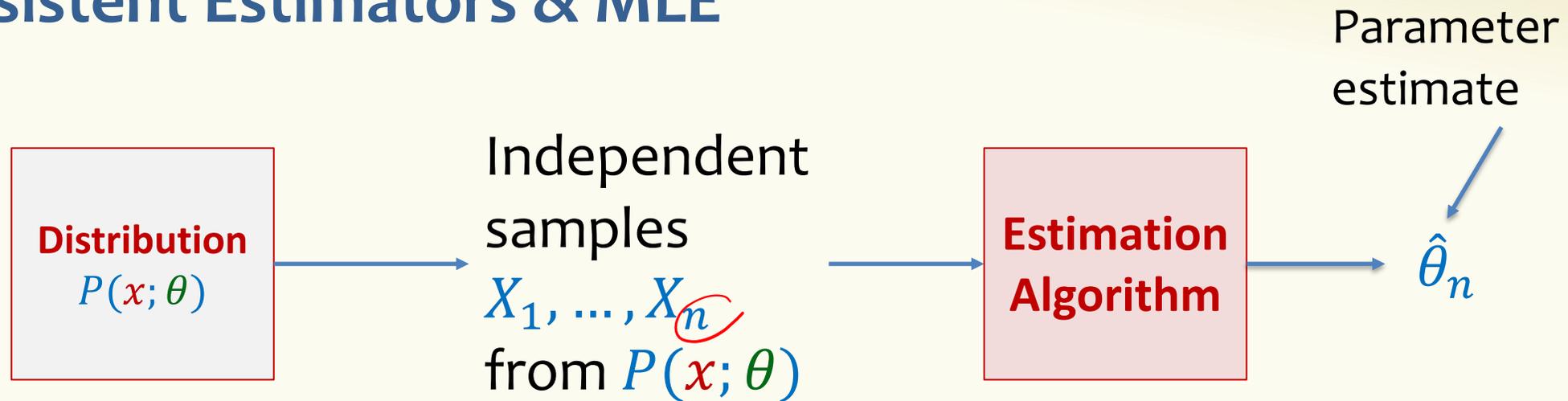
i.e., $\mathbb{E}[\hat{\theta}_\mu]$ = p , where p is the probability that the coin turns out head.

$$\mathbb{E}[\hat{\theta}_\mu] = \mathbb{E}\left[\frac{n_H}{n}\right] = \frac{1}{n} \mathbb{E}[n_H] = \frac{1}{n} (np) = \underline{\underline{p}}$$

Why?

Because $\mathbb{E}[n_H] = np$ when p is the true probability of heads.

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 $\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\theta}_n] = \theta$.

Theorem. MLE estimators are consistent.

(But not necessarily unbiased)

Example – Consistency

Normal outcomes X_1, \dots, X_n i.i.d. according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

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

Population variance – Biased!

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

Sample variance – Unbiased!

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

$\hat{\Theta}_{\sigma^2}$ is “consistent”



Why is the estimator consistent, but biased?

linearity

$X_i \sim \mathcal{N}(\mu, \sigma^2)$

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

... then an algebraic miracle occurs ...

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

Consistent, but biased!

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

Consistent and unbiased!

Bessel's correction

Agenda

- MLE for Normal Distribution
- Unbiased and Consistent Estimators
- Intuition and Bigger Picture ◀

What's with the $n - 1$?

Sooooooooooooo... why is the MLE for variance off?

- **Intuition 1:**

We really want $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i - \mu)^2]$

What we have is $\mathbb{E}[\hat{\Theta}_{\sigma^2}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i - \hat{\Theta}_{\mu})^2]$ for $\hat{\Theta}_{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$
 $\hat{\Theta}_{\mu}$ is not μ !

Each X_i is already included as part of $\hat{\Theta}_{\mu}$ and so is a bit correlated with it
... so X_i is a bit closer to $\hat{\Theta}_{\mu}$ than it would be to the mean μ .

What's with the $n - 1$?


$$\frac{1}{4} \rightarrow x_i$$


$$\underline{x_1, \dots, x_{n-1}}$$

Sooooooooooooo... why is the MLE for variance off?

- **Intuition 2:**

We only have $n - 1$ “degrees of freedom”

– With the sample mean and $n - 1$ of the data points, you know the final data point.

- Only $n - 1$ of the data points have new “information”; the last is fixed by the sample mean.

Why does it matter?

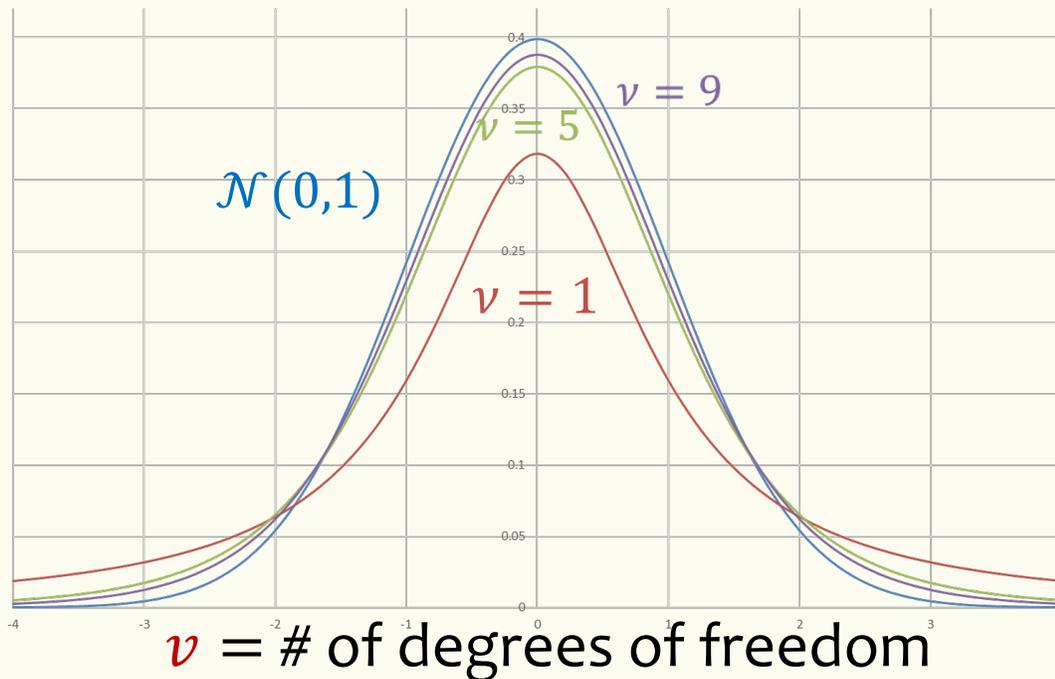
- When statisticians are estimating a variance from a sample, they usually divide by $n-1$ instead of n .
- They and we not only want good estimators (unbiased, consistent)
 - They/we also want **confidence bounds**
 - Upper bounds on the probability that these estimators are far the truth about the underlying distributions
 - Confidence bounds are just like what we wanted for our polling problems, but it turns out that the CLT is not the best thing to use to get them (unless the variance is known)

$$P(|\mu - \hat{\mu}| > \dots) < \dots$$

Why does it matter?

Statisticians do not approximate via the normal distribution, but via Student's (*) t -distribution with $n - 1$ degrees of freedom

– Use t -tables instead of z -tables ...



* "Student" was pseudonym for William Gosset, a statistician who worked for A. Guinness & Son investigating brewing and barley yields



Are there other estimators?

Assume we have prior distribution over what values of θ are likely.

In other words...

assume that we know $P(\theta)$ = probability θ is used, for every θ .

Maximum a-posteriori probability estimation (MAP)

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \operatorname{argmax}_{\theta} \frac{\mathcal{L}(x_1, \dots, x_n | \theta) \cdot P(\theta)}{\sum_{\theta} \mathcal{L}(x_1, \dots, x_n | \theta) \cdot P(\theta)} \\ &= \operatorname{argmax}_{\theta} \mathcal{L}(x_1, \dots, x_n | \theta) \cdot P(\theta)\end{aligned}$$

Note when prior is constant, you get MLE!

MLE and MAP in AI and Machine Learning

- MLE and MAP can be defined over distributions that are not nice well-defined families as we have been considering here
 - e.g. $\vec{\theta}$ might be the vector of parameters in some Neural Net or unknown entries in some Bayes Net.
 - A variety of optimization methods and heuristic methods are used to compute/approximate them.