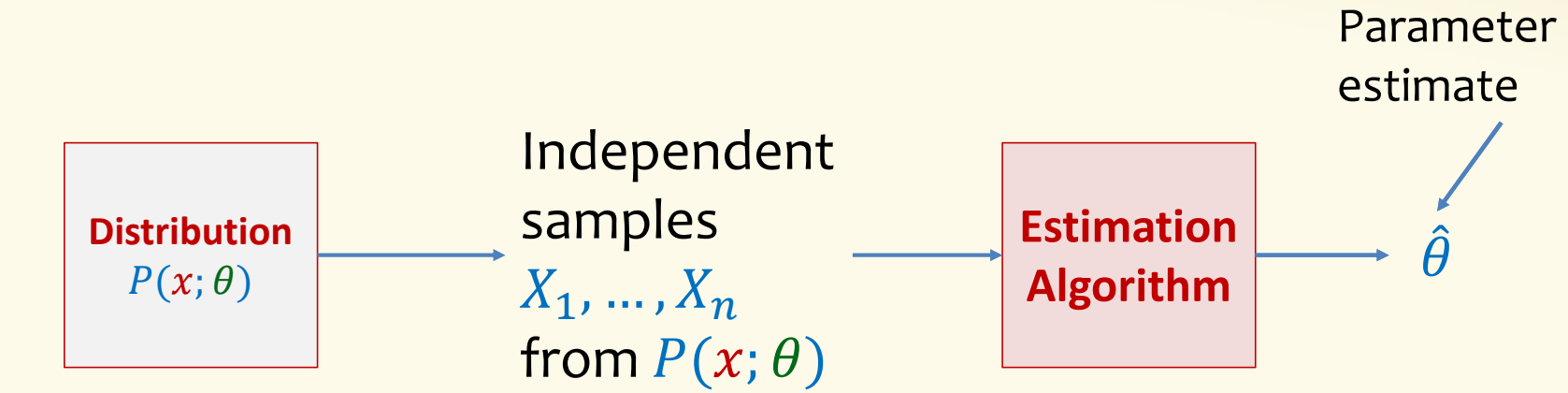
CSE 312

Foundations of Computing II

Lecture 23: Maximum Likelihood Estimation Continued

Review Parameter Estimation – Workflow



θ = unknown parameter

Example: coin flip distribution with unknown θ = probability of heads

Observation: $HTTHHHTHTHTTTTHTHTTTTTHT$

Goal: Estimate θ

Review Likelihood of Different Observations

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

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

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

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

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \mathcal{L}(x_1, \ldots, x_n \mid \theta)$$

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

Review General Recipe

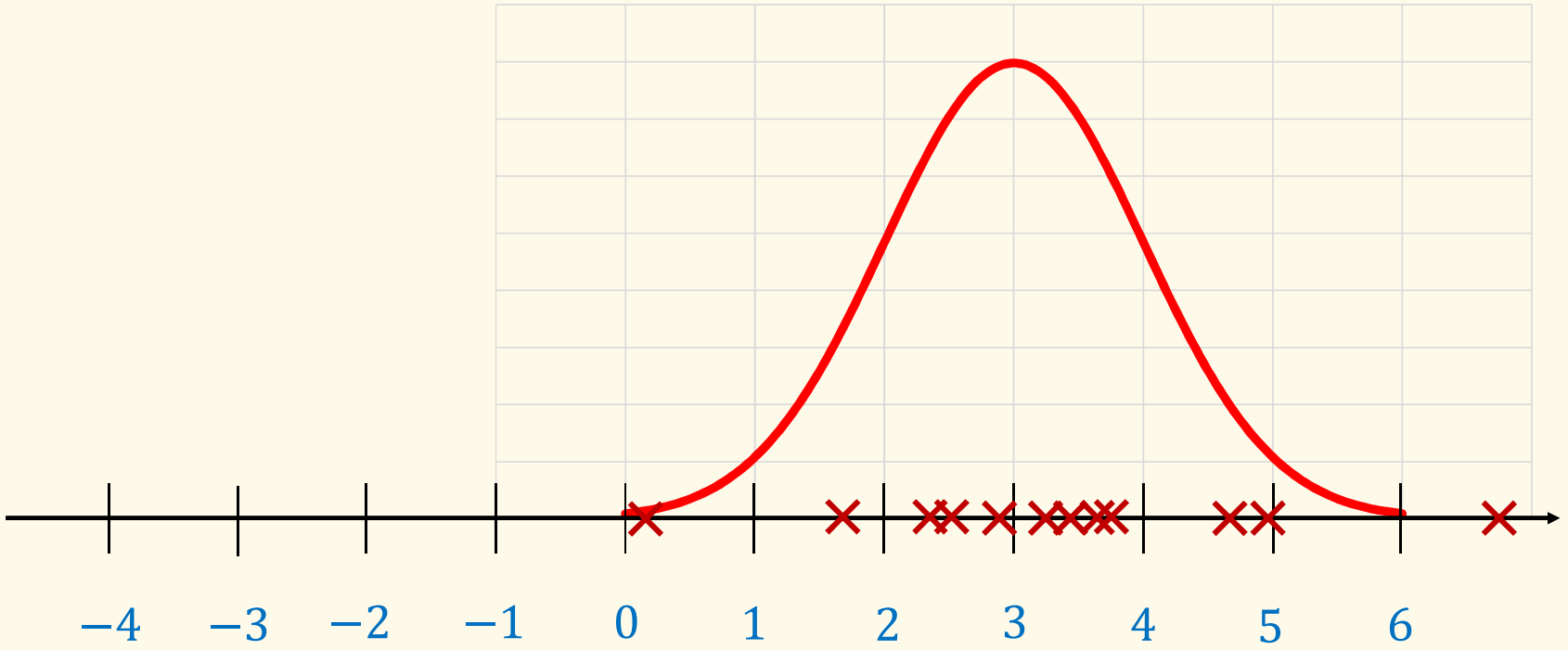
1. **Input** Given n i.i.d. samples $x_1, \ldots, x_n$ from parametric model with parameter θ.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, \ldots., x_n | \theta)$.
   - For discrete $\mathcal{L}(x_1, \ldots., x_n | \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
   - For continuous $\mathcal{L}(x_1, \ldots., x_n | \theta) = \prod_{i=1}^{n} f(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \ldots., x_n | \theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots., x_n | \theta)$
5. **Solve for** $\hat{\theta}$ by setting derivative to 0 and solving for max.

Generally, you need to do a second derivative test to verify it is a maximum, but we won't ask you to do that in CSE 312.

Agenda

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

n samples $x_1, \ldots, x_n \in \mathbb{R}$ 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

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

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

Example – Gaussian Parameters

Goal: estimate θ = expectation

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

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

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

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

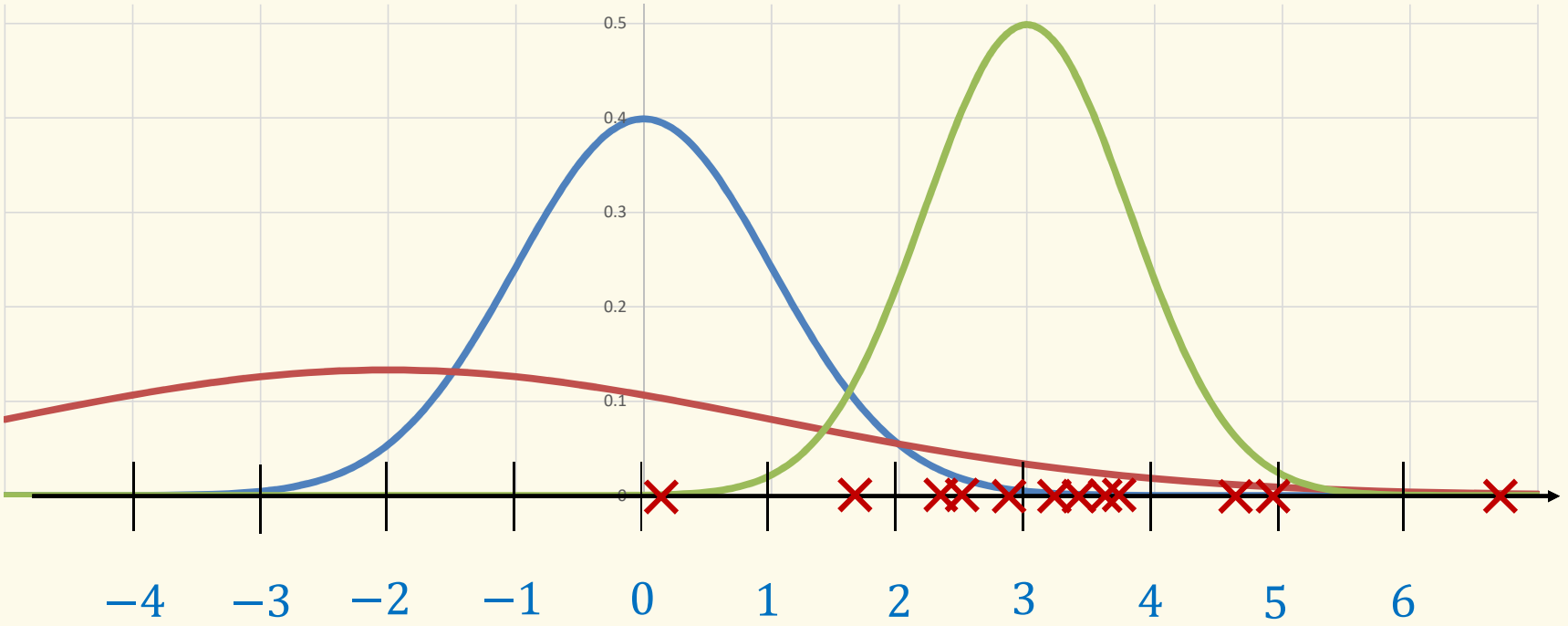
So… solve $\sum_{i=1}^{n} x_i - n\hat{\theta} = 0$ for $\hat{\theta}$

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

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

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

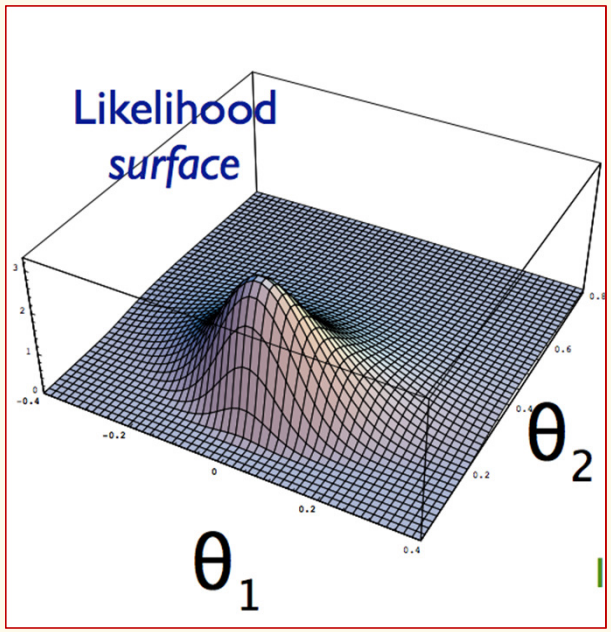
Most likely μ and σ^2?



Two-parameter optimization

Normal outcomes $x_1, \dots, x_n$

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



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

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

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

Two-parameter estimation

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

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

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

MLE for Expectation

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

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

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

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

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

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

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

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

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

(Note that this is not called sample variance!)

Likelihood – Continuous Case

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

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

Normal outcomes $x_1, \ldots, x_n$

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

MLE estimator for **expectation**

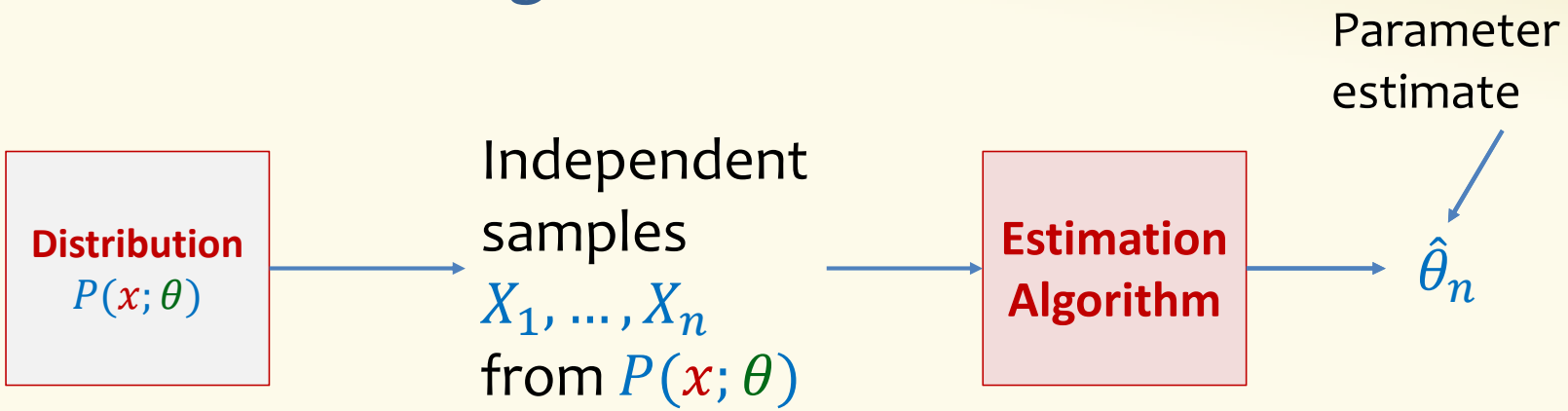
$$\hat{\theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\theta}_\mu\right)^2$$

MLE estimator for **variance**

Agenda

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

When is an estimator good?



θ = unknown parameter

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

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

Note: This expectation is over the samples $X_1, \ldots, X_n$

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

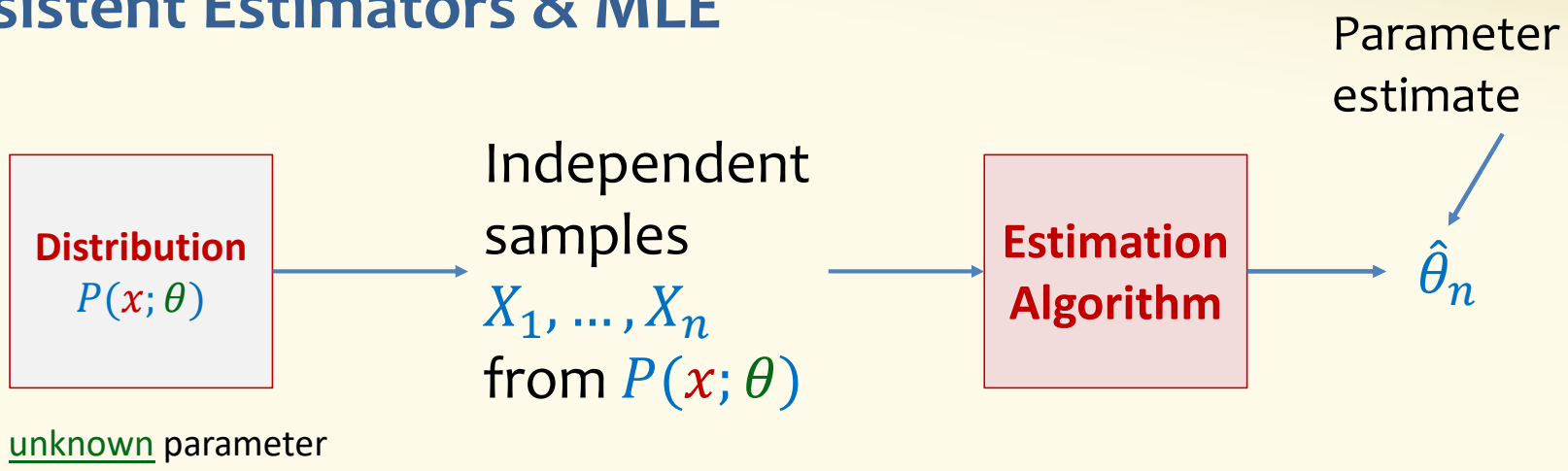
Fact. $\hat{\theta}_\mu$ is unbiased

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

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 $\mathbb{E}\left[\hat{\theta}_n\right] = \theta$ for all $n \geq 1$.

Definition. An estimator is **consistent** if $\lim_{n\to\infty} \mathbb{E}\left[\hat{\theta}_n\right] = \theta$.

Theorem. MLE estimators are consistent.

(But not necessarily unbiased)

Example – Consistency

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

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

Population variance – Biased!

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

Sample variance – Unbiased!

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

$\widehat{\Theta}_{\sigma^2}$ is "consistent"

Why is the estimator consistent, but biased?

linearity

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

… then an algebraic miracle occurs …

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

Consistent, but biased!

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

Consistent and unbiased!

Bessel's correction

Agenda

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

What's with the $n - 1$?

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

  $\widehat{\Theta}_{\mu}$ is not μ !

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

What's with the $n - 1$?

Soooooooooo… 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.

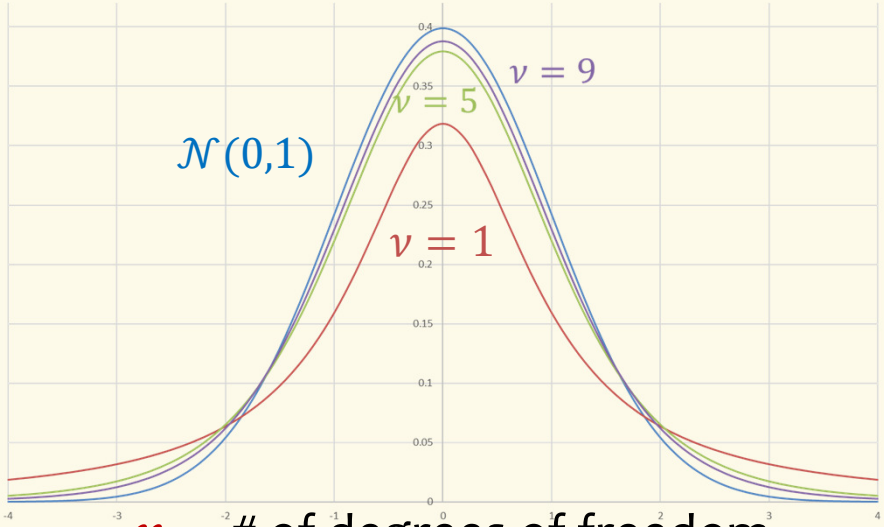
- They and we not only want good estimators (unbiased, consistent)
  - They/we also want **confidence bounds**
    - Upper bounds on the probability that these estimators are far the truth about the underlying distributions
  - Confidence bounds are just like what we wanted for our polling problems, but it turns out that the CLT is not the best thing to use to get them (unless the variance is known)

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 …



ν = # of degrees of freedom

* "Student" was pseudonym for William Gosset, a statistician who worked for A. Guiness & 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)

$$\hat{\theta}_{\text{MAP}} = \text{argmax}_{\theta} \frac{\mathcal{L}(x_1, \ldots, x_n | \theta) \cdot P(\theta)}{\sum_{\theta} \mathcal{L}(x_1, \ldots, x_n | \theta) \cdot P(\theta)}$$

$$= \text{argmax}_{\theta} \ \mathcal{L}(x_1, \ldots, x_n | \theta) \cdot P(\theta)$$

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 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.