CSE 312

Foundations of Computing II

22: Maximum Likelihood Estimation (MLE)

www.slido.com/1692973

Agenda

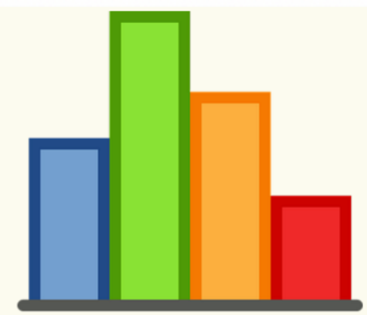
- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- Continuous MLE

Probability vs Statistics

$\mathrm{Ber}(p = 0.5)$

Probability
Given model, predict data

$P(THHTHH)$



$\mathrm{Ber}(p = ??)$

Statistics
Given data, predict model

$THHTHH$

Recap Formalizing Polls

We assume that poll answers $X_1, \dots, X_n \sim \mathrm{Ber}(p)$ i.i.d. for unknown p

Goal: Estimate p

We did this by computing $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$

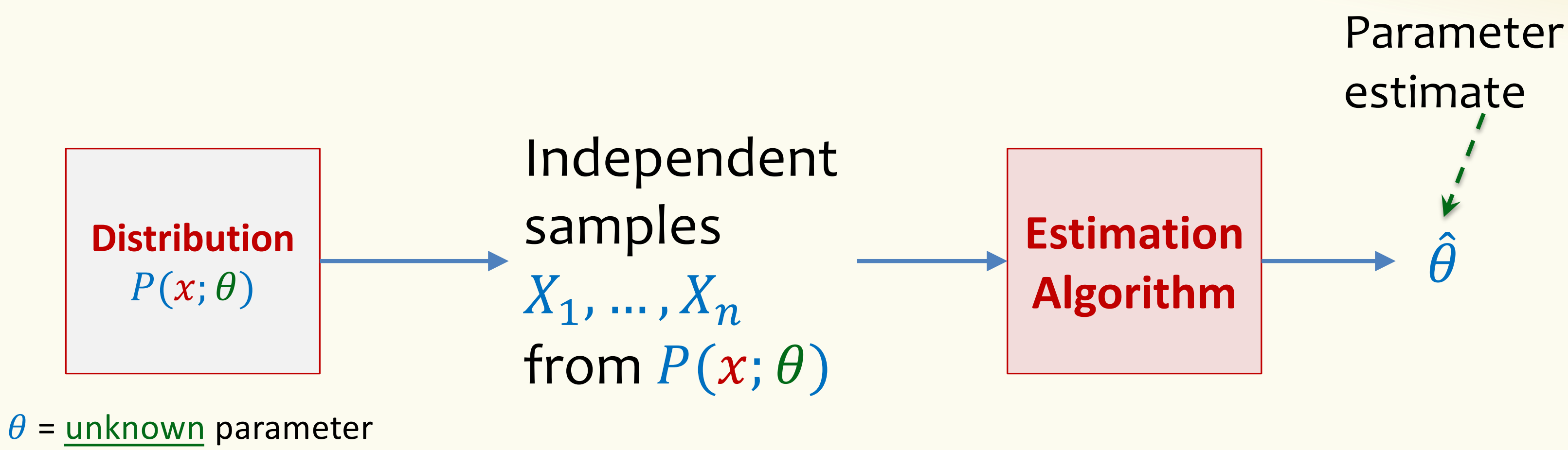
Recap More generally …

In estimation we often … .

- Assume: we know the type of the random variable that we are observing independent samples from
  - We just don't know the parameters, e.g.
    - the bias p of a random coin $\mathrm{Bernoulli}(p)$
    - The arrival rate λ for the $\mathrm{Poisson}(\lambda)$ or $\mathrm{Exponential}(\lambda)$
    - The mean μ and variance σ of a normal $\mathcal{N}(\mu, \sigma)$

- Goal: find the "best" parameters to fit the data

Statistics: Parameter Estimation – Workflow



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

Observation: $HTTHHHTHTHTTTTHTHTTTTTHT$

Goal: Estimate θ

Example

Suppose we have a mystery coin with some probability p of coming up heads. We flip the coin 8 times, independent of other flips, and see the following sequence of flips

$$TTHTHTTH$$

Given this data, what would you estimate p is?

Poll: **www.slido.com/1692973**

a. $1/2$
b. $5/8$
c. $3/8$
d. $1/4$

How can you argue "objectively" that this your estimate is the best estimate?

Agenda

- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- Continuous MLE

Likelihood

Say we see outcome $HHTHH$.

You tell me your best guess about the value of the unknown parameter θ (a.k.a. p) is 4/5. Is there some way that you can argue "objectively" that this is the best estimate?

Likelihood

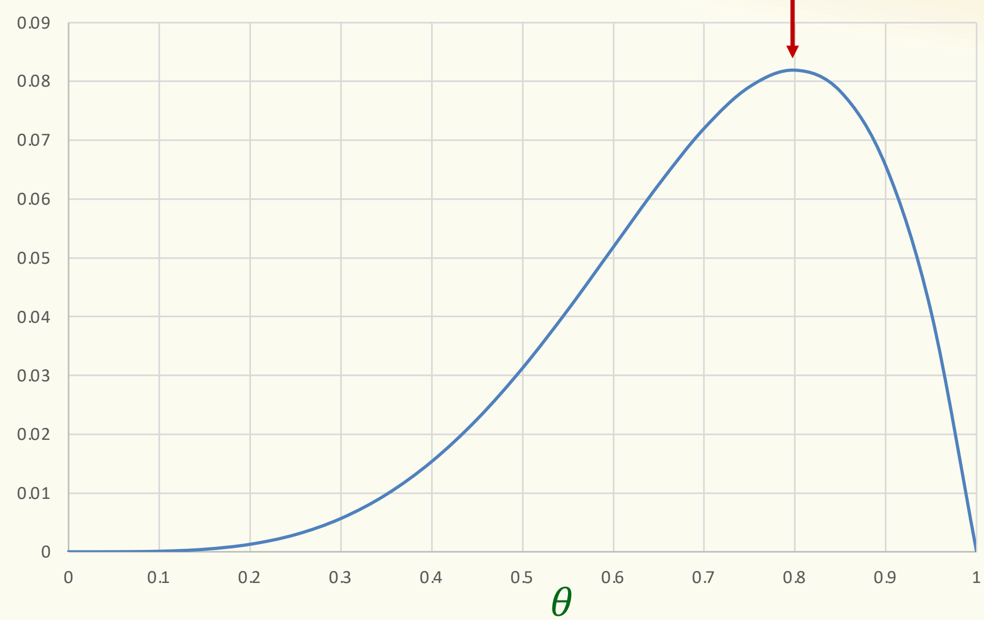
Say we see outcome $HHTHH$.

$$\mathcal{L}(HHTHH \mid \theta) = \theta^4(1-\theta)$$

Probability of observing the outcome $HHTHH$ if $\theta =$ prob. of heads.

For a fixed outcome $HHTHH$, this is a function of θ.

Max Prob of seeing HHTHH



Likelihood of Different Observations

(Discrete case)

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

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

Example:
Say we see outcome $HHTHH$.

$$\mathcal{L}(HHTHH \mid \theta) = P(H;\theta) \cdot P(H;\theta) \cdot P(T;\theta) \cdot P(H;\theta) \cdot P(H;\theta) = \theta^4(1-\theta)$$

Likelihood vs. Probability

- Fixed θ: **probability** $\prod_{i=1}^{n} P(x_i;\theta)$ that dataset $x_1,\dots,x_n$ is sampled by distribution with parameter θ
  - A function of $x_1,\dots,x_n$
- Fixed $x_1,\dots,x_n$: **likelihood** $\mathcal{L}(x_1,\dots,x_n \mid \theta)$ that parameter θ explains dataset $x_1,\dots,x_n$.
  - A function of θ

These notions are the same number if we fix *both* $x_1,\dots,x_n$ and θ, but different role/interpretation

Likelihood of Different Observations

(Discrete case)

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

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

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!

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

Example – Coin Flips

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

– i.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots., x_n | \theta) = \theta^{n_H} (1 - \theta)^{n_T}$$

Goal: find θ that maximizes $\mathcal{L}(x_1, \ldots., x_n | \theta)$

Example – Coin Flips

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

– i.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots., x_n | \theta) = \theta^{n_H} (1-\theta)^{n_T}$$

$$\frac{\partial}{\partial \theta} \mathcal{L}(x_1, \ldots., x_n | \theta) = ???$$

While it is possible to compute this derivative, it's not always nice since we are working with products.

Log-Likelihood

We can save some work if we use the **log-likelihood** instead of the likelihood directly.

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

$$\ln \mathcal{L}(x_1, \ldots, x_n \mid \theta) = \ln \prod_{i=1}^{n} P(x_i; \theta) = \sum_{i=1}^{n} \ln P(x_i; \theta)$$

Useful log properties

$$\ln(ab) = \ln(a) + \ln(b)$$
$$\ln(a/b) = \ln(a) - \ln(b)$$
$$\ln\left(a^b\right) = b \cdot \ln(a)$$

Example – Coin Flips

$\ln(ab) = \ln(a) + \ln(b)$
$\ln(a/b) = \ln(a) - \ln(b)$
$\ln(a^b) = b \cdot \ln(a)$

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

– i.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \dots, x_n | \theta) = \theta^{n_H} (1-\theta)^{n_T}$$

Example – Coin Flips

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

– i.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \theta^{n_H} (1-\theta)^{n_T}$$

$$\ln \mathcal{L}(x_1, \ldots, x_n | \theta) = n_H \ln \theta + n_T \ln(1-\theta)$$

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n | \theta) = n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1-\theta}$$

Want value $\hat{\theta}$ of θ s.t. $\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n | \theta) = 0$

So we need $n_H \cdot \frac{1}{\hat{\theta}} - n_T \cdot \frac{1}{1-\hat{\theta}} = 0$

Solving gives

$$\hat{\theta} = \frac{n_H}{n}$$

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

Brain Break

Agenda

- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- Continuous MLE

The Continuous Case

Given n (independent) samples $x_1, \ldots, x_n$ from (continuous) parametric model $f(x_i; \theta)$ which is now a family of densities

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

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

Replace pmf with pdf!

Why density?

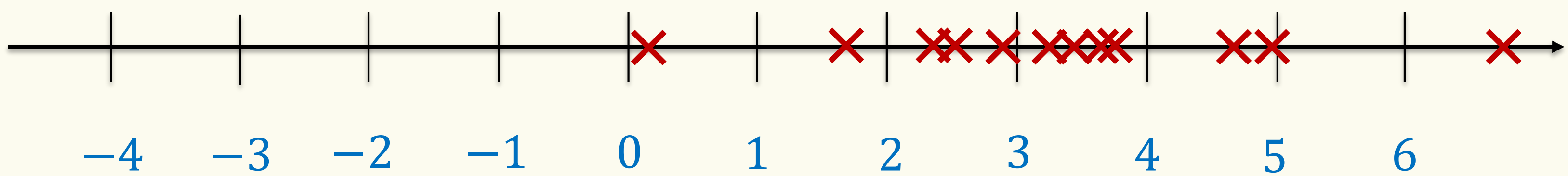
- Density $\neq$ probability, but:
  - For maximizing likelihood, we really only care about relative likelihoods, and density captures that
  - has desired property that likelihood increases with better fit to the model

Agenda

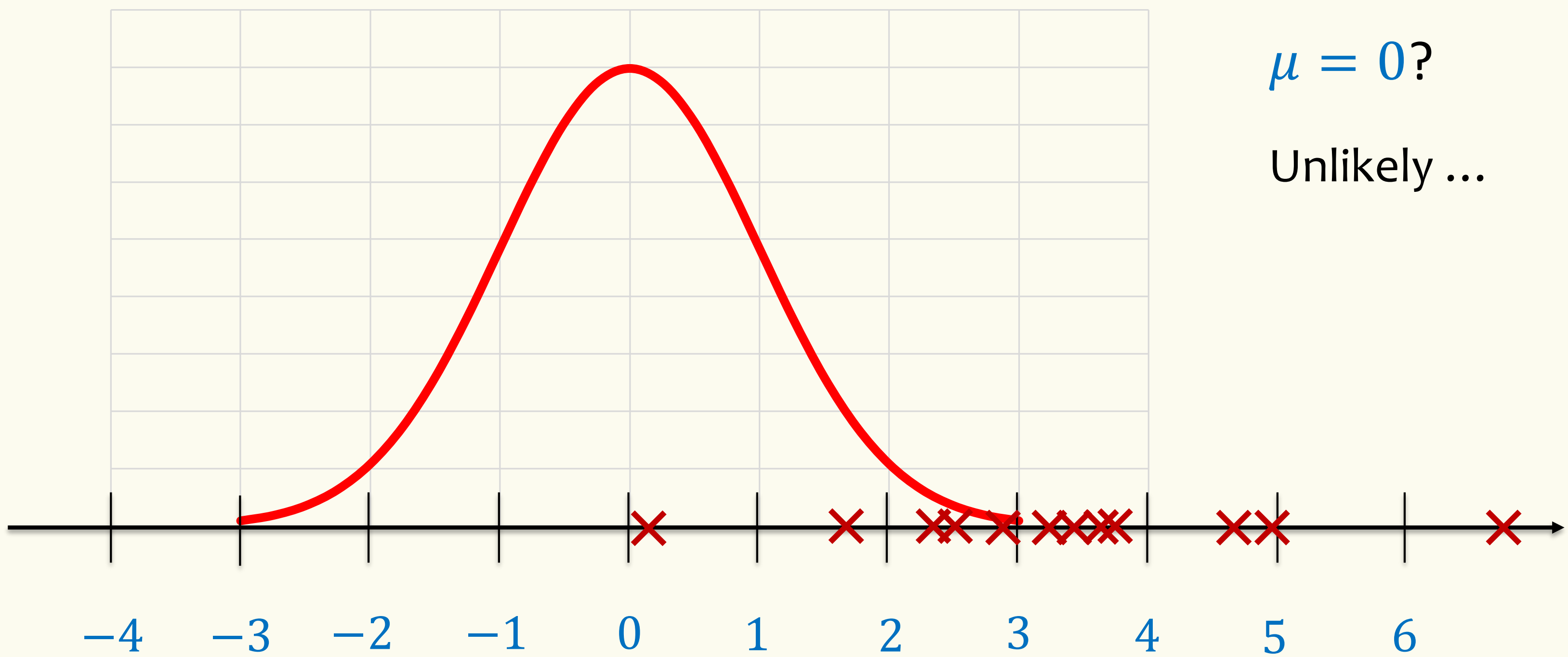
- MLE for Normal Distribution
- Unbiased and Consistent Estimators
- Odds and ends

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

[i.e., we are given the promise that the variance is 1]



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



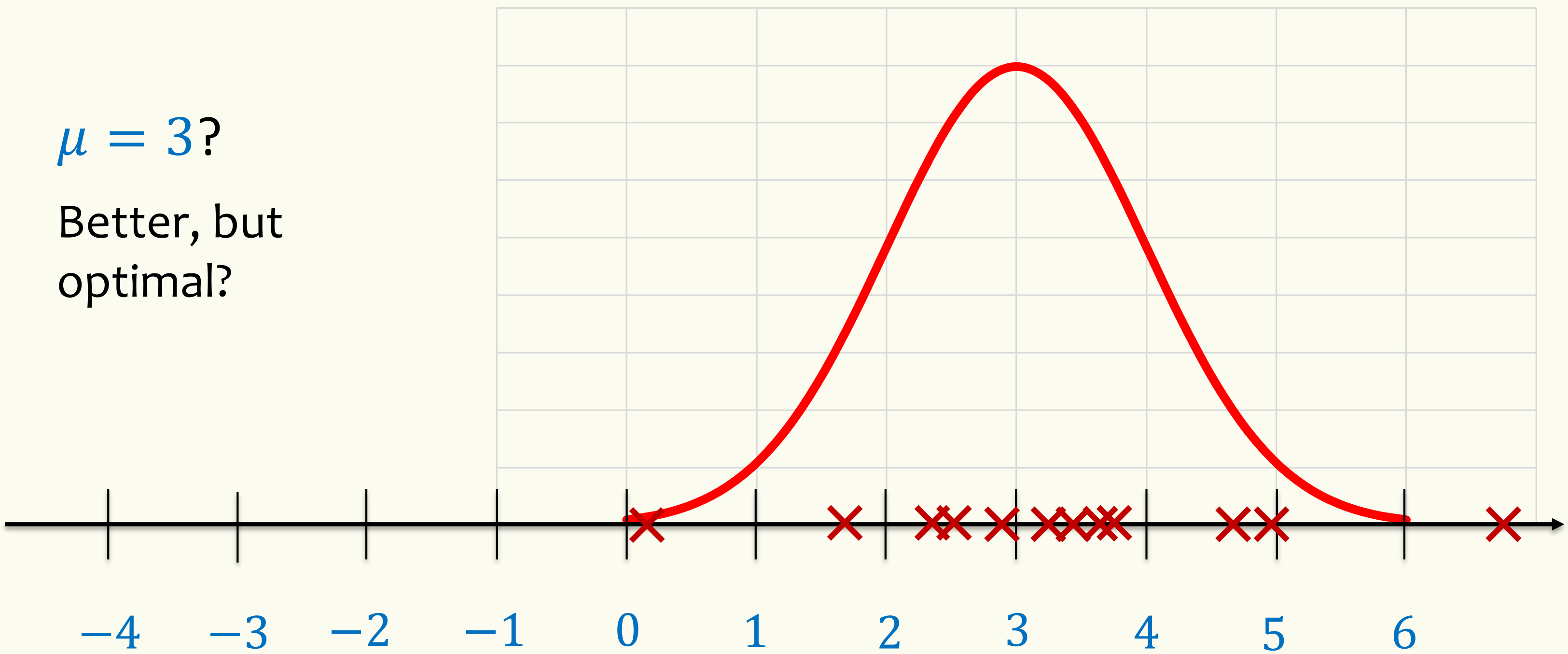
$\mu = 0$?

Unlikely …

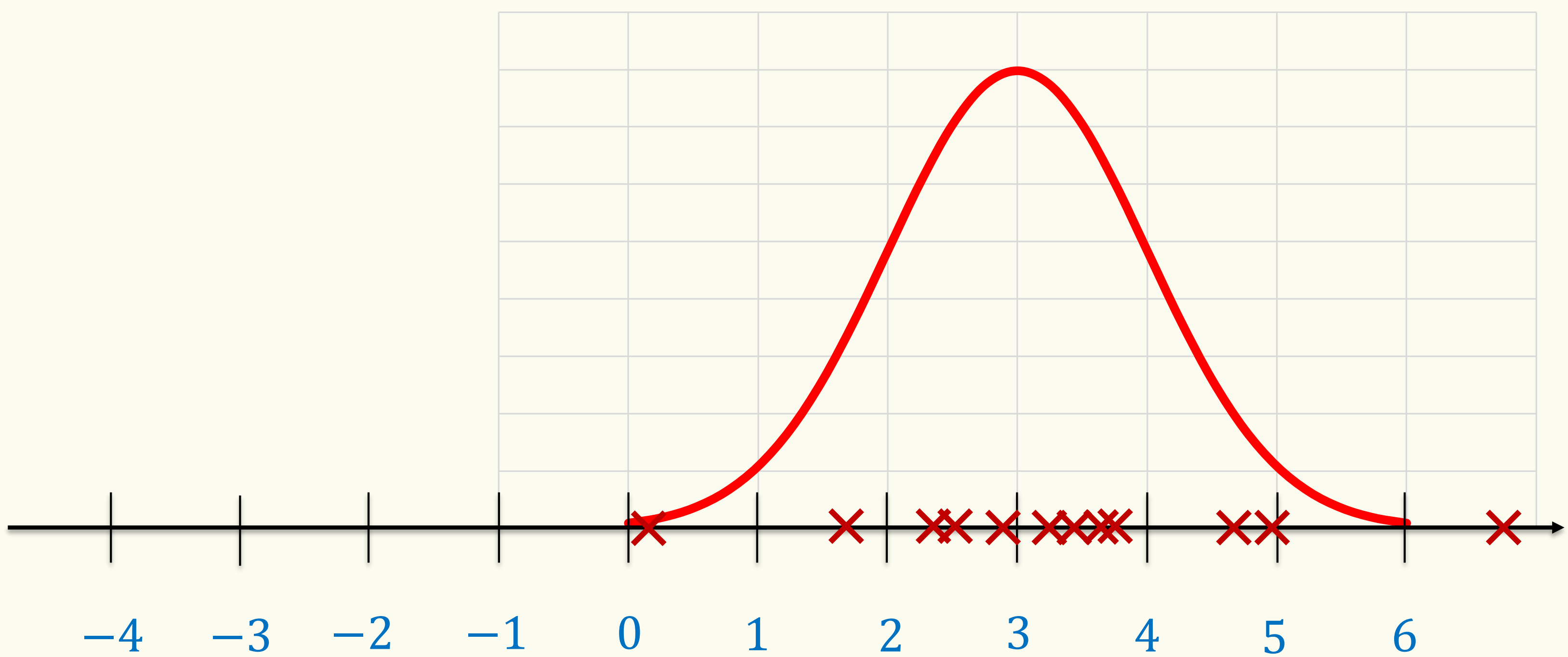
n samples $x_1, \ldots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely μ?

$\mu = 3$?

Better, but optimal?



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



$$\ln(ab) = \ln(a) + \ln(b)$$
$$\ln(a/b) = \ln(a) - \ln(b)$$
$$\ln\left(a^b\right) = b \cdot \ln(a)$$

Example – Gaussian Parameters

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

Goal: estimate θ, the expectation

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

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

Example – Gaussian Parameters

Goal: estimate θ= expectation

Normal outcomes $x_1, \dots, x_n$, 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: $\frac{\partial}{\partial\theta}\frac{(x_i-\theta)^2}{2} = \frac{1}{2} \cdot 2 \cdot (x_i - \theta) \cdot (-1) = \theta - x_i$

Example – Gaussian Parameters

Goal: estimate θ= expectation

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

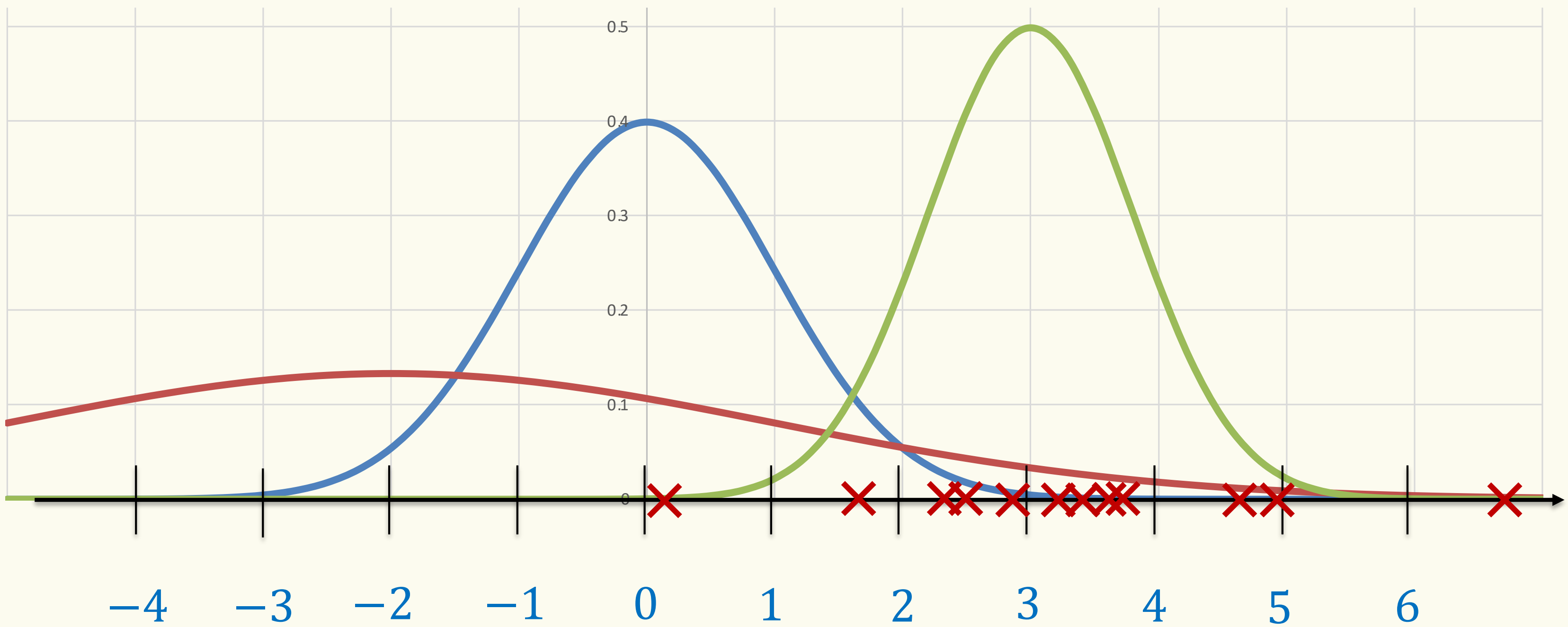
$$\frac{\partial}{\partial\theta} \ln \mathcal{L}(x_1, \dots., x_n | \theta) = \sum_{i=1}^{n} (x_i - \theta) = \sum_{i=1}^{n} x_i - n\theta$$

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

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

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?

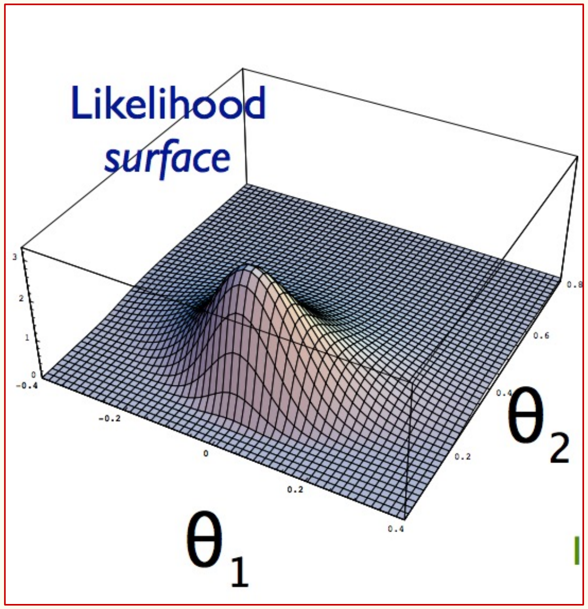


$$\ln(ab) = \ln(a) + \ln(b)$$
$$\ln(a/b) = \ln(a) - \ln(b)$$
$$\ln\left(a^b\right) = b \cdot \ln(a)$$

Two-parameter optimization

Normal outcomes $x_1, \dots, x_n$

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



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

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

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

Two-parameter estimation

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

Find pair $\hat{\theta}_1, \hat{\theta}_2$ that maximizes $\ln \mathcal{L}(x_1, \dots, x_n \mid \theta_1, \theta_2)$

Two-parameter estimation

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

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

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

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

MLE for Expectation

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

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

MLE for Expectation

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

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

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

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

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

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

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

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

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**

General Recipe

1. **Input** Given n i.i.d. samples $x_1, \dots, x_n$ from parametric model with parameter θ.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, \dots., x_n | \theta)$.
   - For discrete $\mathcal{L}(x_1, \dots., x_n | \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
   - For continuous $\mathcal{L}(x_1, \dots., x_n | \theta) = \prod_{i=1}^{n} f(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \dots., x_n | \theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial\theta} \ln \mathcal{L}(x_1, \dots., x_n | \theta)$
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
- Intuition and Bigger Picture

When is an estimator good?

Parameter estimate

Distribution $P(x;\theta)$ → Independent samples $X_1, \dots, X_n$ from $P(x;\theta)$ → **Estimation Algorithm** → $\hat{\theta}_n$

θ = unknown parameter

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

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

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

Three samples from $U(0, \theta)$

Example – Coin Flips

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., $\mathbb{E}\left[\hat{\theta}_\mu\right] = p$, 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

Parameter estimate

Distribution $P(x;\theta)$ → Independent samples $X_1, \dots, X_n$ from $P(x;\theta)$ → **Estimation Algorithm** → $\hat{\theta}_n$

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 $X_1, \dots, X_n$ 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 – <u>Biased!</u>

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

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 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 CLT is usually not the best thing to use to get them (unless the variance is known)

Agenda

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

Another approach to parameter estimation

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)

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

$$= \text{argmax}_{\theta} \; \mathcal{L}(x_1, \dots, x_n \mid \theta) \cdot P(\theta)$$

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 the 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.

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.