CSE 312

Foundations of Computing II

Lecture 22: Maximum Likelihood Estimation (MLE)

Announcement

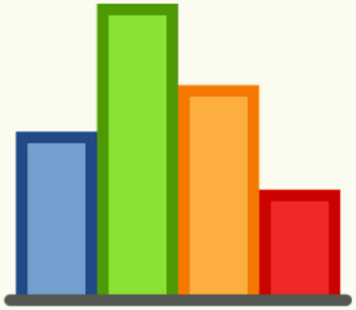
- Lecture on Wed is cancel.
- No lecture on Friday. Happy thanks giving!
- Pset 7 is due on Wed
- Pset 8 is out on Wed, due on next Friday

Agenda

- Chernoff Bound
  - Example: Server Load
  - The Union Bound
- Probability vs statistics
  - Estimation

Probability vs Statistics

$\text{Ber}(p = 0.5)$ ⇨ **Probability** Given model, predict data ⇨ $P(THHTHH)$



$\text{Ber}(p = ??)$ ⇦ **Statistics** Given data, predict model ⇦ $THHTHH$

Recall Formalizing Polls

What type of r.v. is X_i?

		$\mathbb{E}[X_i]$	$\text{Var}(X_i)$
a.	Bernoulli	p	$p(1-p)$

Population size N, true fraction of voting in favor p, sample size n.

Problem: We don't know p

Polling Procedure

for $i = 1, \dots, n$:

1. Pick uniformly random person to call (prob: $1/N$)
2. Ask them how they will vote

$$X_i = \begin{cases} 1, & \text{voting in favor} \\ 0, & \text{otherwise} \end{cases}$$

Report our estimate of p: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Recall 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 $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Why is that a good estimate for p?

More generally …

In estimation we….

- 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 $\text{Bernoulli}(p)$
    - The arrival rate λ for the $\text{Poisson}(\lambda)$ or $\text{Exponential}(\lambda)$
    - The mean μ and variance σ of a normal $\mathcal{N}(\mu, \sigma)$

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

Notation – Parametric Model (discrete case)

Definition. A **(parametric) model** is a family of distributions indexed by a parameter θ, described by a two-argument function

$$P(x;\theta) = \text{prob. of outcome } x \text{ when distribution has parameter } \theta$$

[i.e., every θ defines a different distribution $\sum_x P(x;\theta) = 1$]

Examples

- "Bernoullis": $P(x;\theta = p) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$
- "Geometrics": $P(i;\theta = p) = (1-p)^{i-1}p \quad$ for $i \in \mathbb{N}$

Statistics: Parameter Estimation – Workflow

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

Parameter estimate

θ = unknown parameter

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 flips

$$TTHTHTTH$$

Given this data, what would you estimate p is?

Poll: pollev.com/rachel312

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

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?

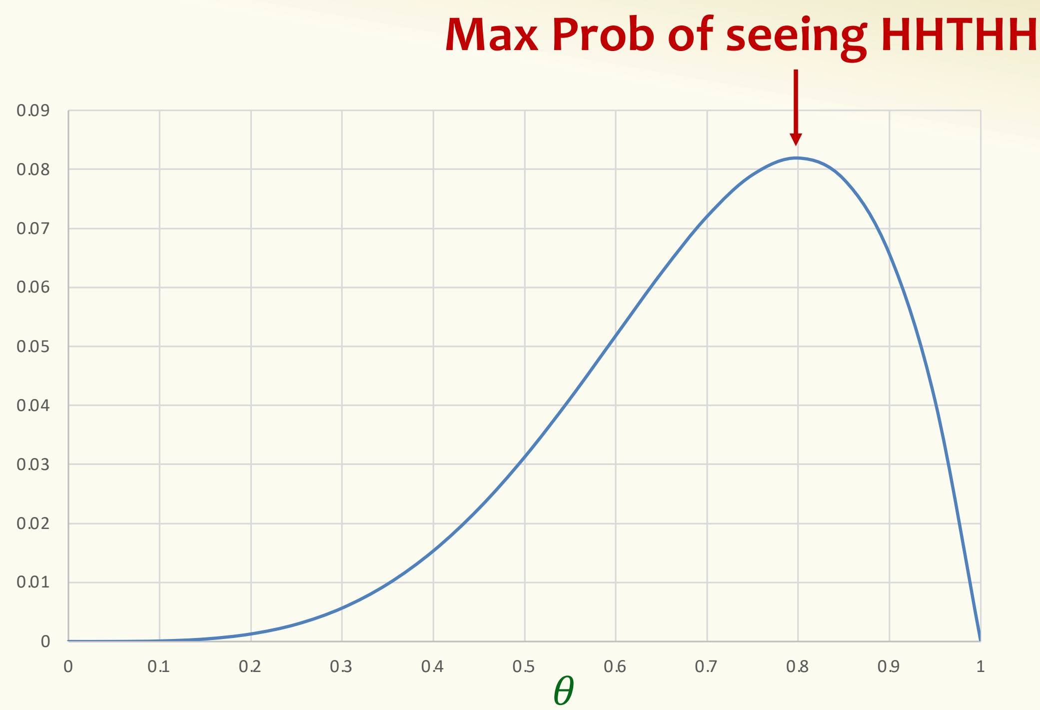
Likelihood

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.



Likelihood of Different Observations

(Discrete case)

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

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

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

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

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

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** $\mathcal{L}(x_1, \dots, x_n | \, \theta)$ 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

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.

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

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

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

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.

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

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

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

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

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 $x_1, \dots, x_n$ from parametric model with parameter θ.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, \dots., x_n | \theta)$.
   - For discrete $\mathcal{L}(x_1, \dots., x_n | \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
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.

Brain Break

Agenda

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

The Continuous Case

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

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

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

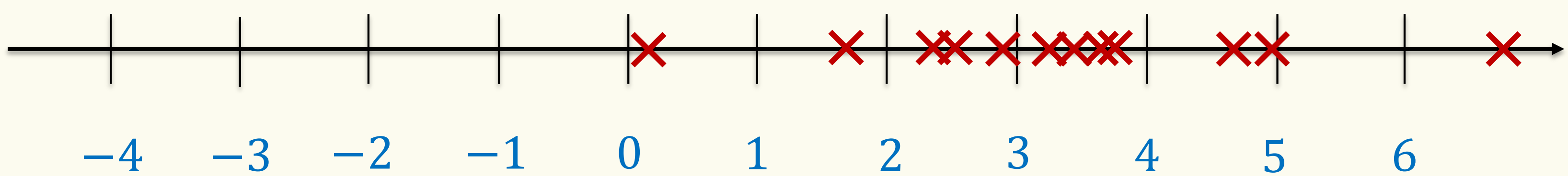
Density function! (Why?)

Why density?

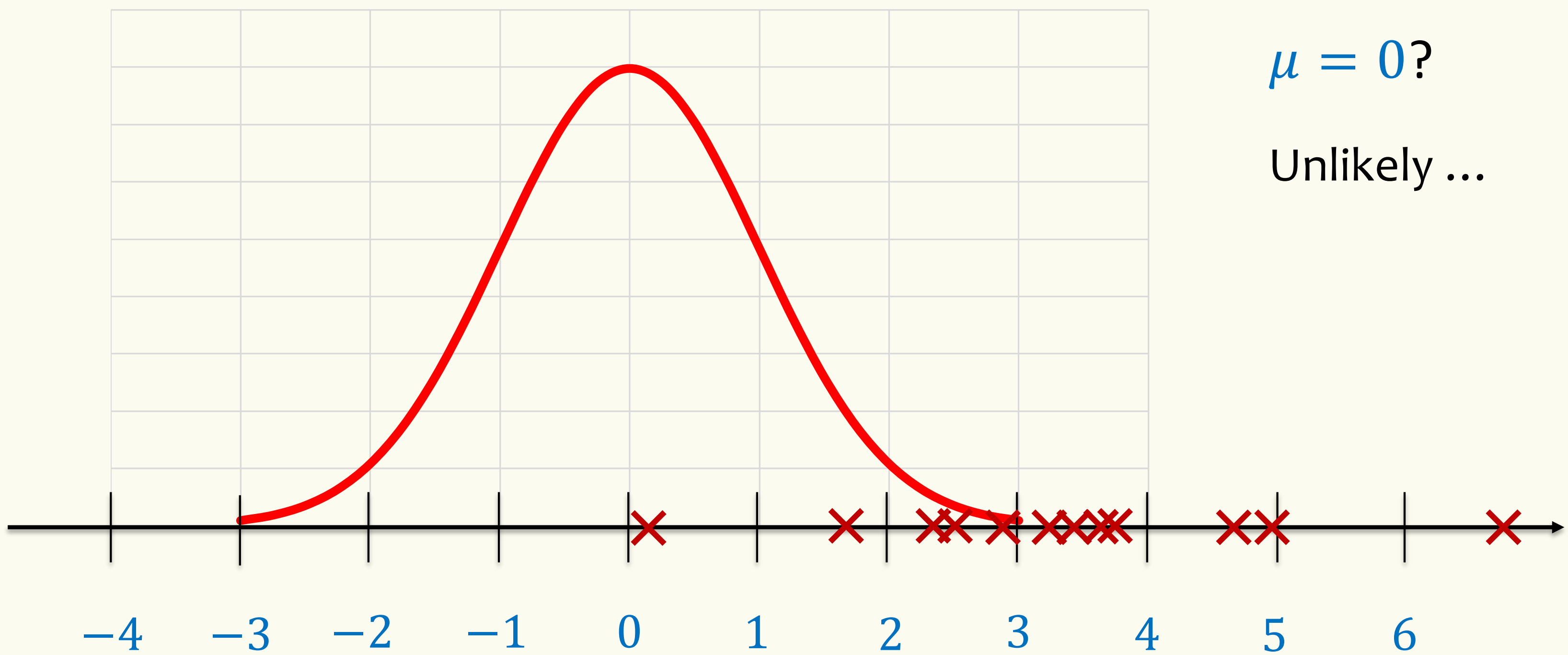
- Density $\neq$ probability, but:
  - For maximizing likelihood, we really only care about relative likelihoods, and density captures that
  - has desired property that likelihood increases with better fit to the model

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

[i.e., we are given the <u>promise</u> that the variance is 1]



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



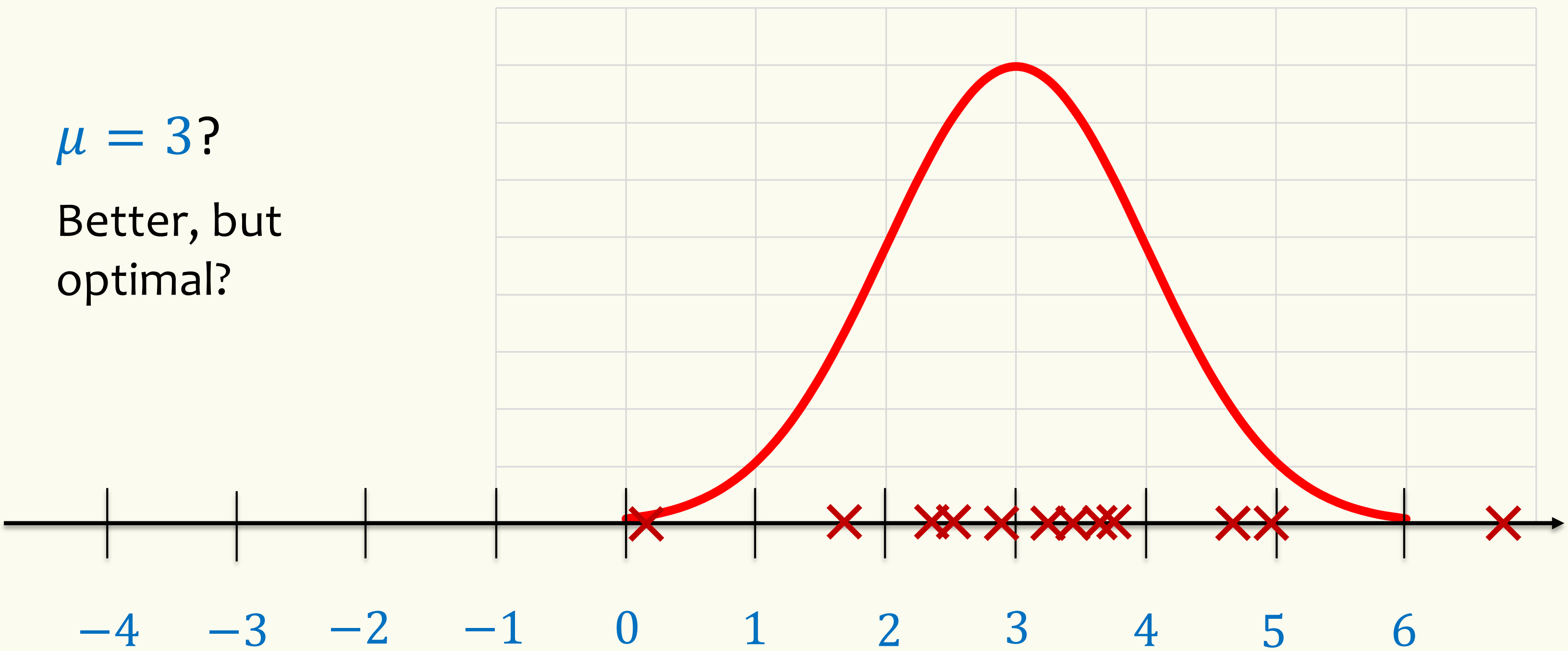
$\mu = 0$?

Unlikely …

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

$\mu = 3$?

Better, but optimal?



Example – Gaussian Parameters

Normal outcomes $x_1, \ldots, x_n$, known variance $\sigma^2 = 1$ but *unknown* mean μ

Goal: estimate θ = mean

Next time:

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

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

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 $\mathcal{L}(x_1, \dots, x_n \mid \theta)$.
   - For discrete $\mathcal{L}(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
   - For continuous $\mathcal{L}(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \dots, x_n \mid \theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \dots, x_n \mid \theta)$
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.