21: Maximum Likelihood Estimation (MLE)
Agenda

- Wrap up on Law of Total Expectation and Law of Total Probability
- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- Continuous MLE
Conditional Expectation

**Definition.** If $X$ is a discrete random variable then the **conditional expectation** of $X$ given event $A$ is

$$
E[X \mid A] = \sum_{x \in \Omega_X} x \cdot P(X = x \mid A)
$$

**Note:**

- Linearity of expectation still applies here

$$
E[aX + bY + c \mid A] = a \ E[X \mid A] + b \ E[Y \mid A] + c
$$
Law of Total Expectation

Law of Total Expectation (event version). Let $X$ be a random variable and let events $A_1, \ldots, A_n$ partition the sample space. Then,

$$
\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i] \cdot P(A_i)
$$

Law of Total Expectation (random variable version). Let $X$ be a random variable and $Y$ be a discrete random variable. Then,

$$
\mathbb{E}[X] = \sum_{y \in \Omega_Y} \mathbb{E}[X \mid Y = y] \cdot P(Y = y)
$$
Law of total probability

**Definition.** Let $A$ be an event and $Y$ a discrete random variable. Then

$$P[A] = \sum_{y \in \Omega_Y} P(A|Y = y)p_Y(y)$$

**Definition.** Let $A$ be an event and $Y$ a continuous random variable. Then

$$P[A] = \int_{-\infty}^{\infty} P(A|Y = y)f_Y(y)dy$$
Example use of law of total probability

Suppose that the time until server 1 crashes is $X \sim \text{Exp} (\lambda)$ and the time until server 2 crashes is independent, with $Y \sim \text{Exp} (\mu)$.

What is the probability that server 1 crashes before server 2?

$P(X < Y)$
Example use of law of total probability

\( X \sim \text{Exp} (\lambda), Y \sim \text{Exp} (\mu). \)

What is the probability that \( Y > X? \)

\[
P(Y > X) = \int_0^\infty \Pr(Y > X \mid X = x) f_x(x) \, dx
\]

\[
= \int_0^\infty \Pr(Y > x \mid X = x) \lambda e^{-\lambda x} \, dx
\]

\[
= \int_0^\infty \Pr(Y > x) \lambda e^{-\lambda x} \, dx
\]

\[
= \int_0^\infty e^{-\mu x} \lambda e^{-\lambda x} \, dx
\]

\[
= \frac{\lambda}{\lambda + \mu} \int_0^\infty (\lambda + \mu) \cdot e^{-\mu x} e^{-\lambda x} \, dx
\]

\[
= \frac{\lambda}{\lambda + \mu}
\]
Alternative approach

\[ f_{x,y}(x, y) = f_x(x) f_y(y) = \lambda \mu e^{-\lambda x} e^{-\mu y} \]

\[ X \sim \text{Exp} (\lambda), \; Y \sim \text{Exp} (\mu). \]

What is the probability that \( Y > X \)?

\[ P(Y > X) = \int_{x=0}^{\infty} \int_{y=x}^{\infty} f_{x,y}(x, y) \, dy \, dx \]

\[ = \int_{x=0}^{\infty} f_x(x) \cdot f_Y(y) \, dy \, dx \]
### Reference Sheet (with continuous RVs)

<table>
<thead>
<tr>
<th></th>
<th>Discrete</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Joint PMF/PDF</strong></td>
<td>$p_{X,Y}(x, y) = P(X = x, Y = y)$</td>
<td>$f_{X,Y}(x, y) \neq P(X = x, Y = y)$</td>
</tr>
<tr>
<td><strong>Joint CDF</strong></td>
<td>$F_{X,Y}(x, y) = \sum \sum_{t \leq x, s \leq y} p_{X,Y}(t, s)$</td>
<td>$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t, s)dsdt$</td>
</tr>
<tr>
<td><strong>Normalization</strong></td>
<td>$\sum \sum_{x} p_{X,Y}(x, y) = 1$</td>
<td>$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$</td>
</tr>
<tr>
<td><strong>Marginal PMF/PDF</strong></td>
<td>$p_{X}(x) = \sum_{y} p_{X,Y}(x, y)$</td>
<td>$f_{X}(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$</td>
</tr>
<tr>
<td><strong>Expectation</strong></td>
<td>$E[g(X, Y)] = \sum \sum_{x} g(x, y)p_{X,Y}(x, y)$</td>
<td>$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy$</td>
</tr>
<tr>
<td><strong>Conditional PMF/PDF</strong></td>
<td>$p_{X \mid Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_{Y}(y)}$</td>
<td>$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_{Y}(y)}$</td>
</tr>
<tr>
<td><strong>Conditional Expectation</strong></td>
<td>$E[X \mid Y = y] = \sum_{x} x p_{X \mid Y}(x \mid y)$</td>
<td>$E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X \mid Y}(x \mid y)dx$</td>
</tr>
<tr>
<td><strong>Independence</strong></td>
<td>$\forall x, y, p_{X,Y}(x, y) = p_{X}(x)p_{Y}(y)$</td>
<td>$\forall x, y, f_{X,Y}(x, y) = f_{X}(x)f_{Y}(y)$</td>
</tr>
</tbody>
</table>
Agenda

• Idea: Estimation
• Maximum Likelihood Estimation (example: mystery coin)
• Continuous MLE
Probability vs Statistics

Ber(\(p = 0.5\)) ➔ Probability
Given model, predict data ➔ \(P(THHTHH)\)

Ber(\(p = ??\)) ➔ Statistics
Given data, predict model ➔ THHTHH
Recap Formalizing Polls

We assume that poll answers $X_1, \ldots, X_n \sim \text{Ber}(p)$ 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 $\text{Bernoulli}(p)$
    - The arrival rate $\lambda$ for the $\text{Poisson}(\lambda)$ or $\text{Exponential}(\lambda)$
    - The mean $\mu$ and variance $\sigma$ 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 $\theta = \text{probability of heads}$

Observation: $HTTHHHHTHHTTTTTTHHTTTTTTHT$

Goal: Estimate $\theta$
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?

\[
\frac{3}{8}
\]

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 $\text{HHTHH}$. You tell me your best guess about the value of the unknown parameter $\theta$ (a.k.a. $p$) is 4/5. Is there some way that you can argue “objectively” that this is the best estimate?

\[
\frac{d}{d\theta} L(\theta) = 4\theta^3 - 5\theta^4 = 0
\]

\[
4\theta^3 = 5\theta^4
\]

\[
\frac{4}{5} = \theta
\]
Likelihood

Say we see outcome $HHTHH$. 

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

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

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

Max Prob of seeing $HHTHH$
Likelihood of Different Observations (Discrete case)

Definition. The likelihood of independent observations \( x_1, \ldots, x_n \) is

\[
\mathcal{L}(x_1, x_2, \ldots, x_n ; \theta) = \prod_{i=1}^{n} P(x_i; \theta)
\]

Example:
Say we see outcome \( HHTTHH \).

\[
\mathcal{L}(HHTTHH ; \theta) = P(H; \theta) \cdot P(H; \theta) \cdot P(T; \theta) \cdot P(H; \theta) \cdot P(H; \theta) = \theta^4(1 - \theta)
\]
Likelihood vs. Probability

• Fixed $\theta$: probability $\prod_{i=1}^{n} P(x_i; \theta)$ that dataset $x_1, \ldots, x_n$ is sampled by distribution with parameter $\theta$
  – A function of $x_1, \ldots, x_n$

• Fixed $x_1, \ldots, x_n$: likelihood $L(x_1, x_2, \ldots, x_n; \theta)$ that parameter $\theta$ explains dataset $x_1, \ldots, x_n$.
  – A function of $\theta$

These notions are the same number if we fix both $x_1, \ldots, x_n$ and $\theta$, but different role/interpretation
Likelihood of Different Observations

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

$$\mathcal{L}(x_1, x_2, \ldots, x_n ; \theta) = \prod_{i=1}^{n} P(x_i; \theta)$$

**Maximum Likelihood Estimation (MLE).** Given data $x_1, \ldots, x_n$, find $\hat{\theta}$ such that $\mathcal{L}(x_1, x_2, \ldots, x_n ; \hat{\theta})$ is maximized!

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(x_1, x_2, \ldots, x_n ; \theta)$$
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 $\theta = \text{prob. heads.}$

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

**Goal:** find $\theta$ 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 $\theta = \text{prob. heads}$.

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

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

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

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

Useful log properties

\[
\ln(ab) = \ln(a) + \ln(b)
\]
\[
\ln(a/b) = \ln(a) - \ln(b)
\]
\[
\ln(a^b) = b \cdot \ln(a)
\]
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 $\theta = \text{prob. heads}$.

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

$$\ln L = \ln \theta^{n_H} + \ln (1 - \theta)^{n_T}$$

$$= n_H \ln \theta + n_T \ln (1 - \theta)$$
Example – Coin Flips

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

– i.e., $n_H + n_T = n$

**Goal:** estimate $\theta = \text{prob. heads.}$

\[
\mathcal{L}(x_1, ..., x_n ; \theta) = \theta^{n_H} (1 - \theta)^{n_T}
\]

\[
\ln \mathcal{L}(x_1, ..., x_n ; \theta) = n_H \ln \theta + n_T \ln(1 - \theta)
\]

\[
\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, ..., x_n ; \theta) = n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1 - \theta}
\]

Want value $\hat{\theta}$ of $\theta$ s.t.

\[
\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, ..., x_n ; \theta) = 0
\]

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 $x_1, ..., x_n$ from parametric model with parameter $\theta$.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, ..., x_n ; \theta)$.
   - For discrete $\mathcal{L}(x_1, ..., x_n ; \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, ..., x_n ; \theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, ..., x_n ; \theta)$
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

\[ \Pr(x \in X) = \int_X p(x) \, dx \]
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 $\mu$? [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 $\mu$?
Example – Gaussian Parameters

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

**Goal:** estimate $\theta$, the expectation

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

$$
\ln \mathcal{L} = \ln \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^n \right] + \sum_{i=1}^{n} \ln \left[ e^{-\frac{(x_i - \theta)^2}{2}} \right] - n \ln 2\pi\
\ln \mathcal{L}(x_1, \ldots, x_n \mid \theta) = -n \frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2}
$$
Example – Gaussian Parameters

Goal: estimate $\theta = \text{expectation}$

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

Goal: estimate $\theta = \text{expectation}$

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

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

Note:

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

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

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

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

In other words, MLE is the sample mean of the data.
Next: \( n \) samples \( x_1, \ldots, x_n \in \mathbb{R} \) from Gaussian \( \mathcal{N}(\mu, \sigma^2) \).
Most likely \( \mu \) and \( \sigma^2 \)?
Two-parameter optimization

Normal outcomes $x_1, \ldots, x_n$

Goal: estimate $\theta_\mu = \text{expectation}$ and $\theta_{\sigma^2} = \text{variance}$

$$\mathcal{L}(x_1, \ldots, x_n ; \theta_\mu, \theta_{\sigma^2}) = \left(\frac{1}{\sqrt{2\pi \theta_{\sigma^2}}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i - \theta_\mu)^2}{2\theta_{\sigma^2}}}$$

$$\ln \mathcal{L}(x_1, \ldots, x_n ; \theta_\mu, \theta_{\sigma^2}) =$$

$$= -n \ln(2\pi \theta_{\sigma^2}) - \sum_{i=1}^{n} \frac{(x_i - \theta_\mu)^2}{2\theta_{\sigma^2}}$$
Two-parameter estimation

\[
\ln \mathcal{L}(x_1, \ldots, x_n ; \theta_\mu, \theta_{\sigma^2}) = -\frac{\ln(2\pi \theta_{\sigma^2})}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_\mu)^2}{2\theta_{\sigma^2}}
\]

Find pair \( \hat{\theta}_\mu, \hat{\theta}_{\sigma^2} \) that maximizes \( \ln \mathcal{L}(x_1, \ldots, x_n ; \theta_\mu, \theta_{\sigma^2}) \)
Two-parameter estimation

\[ \ln \mathcal{L}(x_1, \ldots, x_n ; \theta_\mu, \theta_\sigma^2) = -\frac{\ln(2\pi \theta_\sigma^2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_\mu)^2}{2\theta_\sigma^2} \]

We need to find a solution \( \hat{\theta}_\mu, \hat{\theta}_\sigma^2 \) to

\[ \frac{\partial}{\partial \theta_\mu} \ln \mathcal{L}(x_1, \ldots, x_n ; \theta_\mu, \theta_\sigma^2) = 0 \]

\[ \frac{\partial}{\partial \theta_\sigma^2} \ln \mathcal{L}(x_1, \ldots, x_n ; \theta_\mu, \theta_\sigma^2) = 0 \]
MLE for Expectation

\[ \ln \mathcal{L}(x_1, \ldots, x_n ; \theta_\mu, \theta_\sigma^2) = -n \frac{\ln(2\pi \theta_\sigma^2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_\mu)^2}{2\theta_\sigma^2} \]

\[ \frac{\partial}{\partial \theta_\mu} \ln \mathcal{L}(x_1, \ldots, x_n ; \theta_\mu, \theta_\sigma^2) = \frac{1}{\theta_\sigma^2} \sum_{i}^{n} (x_i - \theta_\mu) = 0 \]
MLE for Expectation

\[
\ln \mathcal{L}(x_1, \ldots, x_n ; \theta, \theta_\sigma) = -n \ln(2\pi \theta_\sigma) - \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\theta_\sigma}
\]

\[
\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n ; \theta, \theta_\sigma) = \frac{1}{\theta_\sigma} \sum_{i} (x_i - \theta) = 0
\]

\[\hat{\theta}_\mu = \frac{\sum_{i} x_i}{n}\]

In other words, MLE of expectation is (again) the \textit{sample mean} of the data, regardless of \(\theta_\sigma\)

What about the variance?
MLE for Variance

\[ \ln L(x_1, ..., x_n; \hat{\mu}, \theta_\sigma^2) = -n \frac{\ln(2\pi \theta_\sigma^2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{2 \theta_\sigma^2} \]

\[ = -n \frac{\ln 2\pi}{2} - n \frac{\ln \theta_\sigma^2}{2} - \frac{1}{2 \theta_\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \]

\[ \frac{\partial}{\partial \theta_\sigma^2} \ln L(x_1, ..., x_n; \hat{\mu}, \theta_\sigma^2) = -\frac{n}{2 \theta_\sigma^2} + \frac{1}{2 \theta_\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = 0 \]

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

In other words, MLE of variance is the population variance of the data.
**Likelihood – Continuous Case**

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

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

Normal outcomes \( x_1, \ldots, x_n \)

\[
\hat{\theta}_\mu = \frac{\sum_{i}^{n} x_i}{n}
\]

**MLE estimator for expectation**

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

**MLE estimator for variance**
General Recipe

1. **Input** Given $n$ i.i.d. samples $x_1, \ldots, x_n$ from parametric model with parameter $\theta$.

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
• Intuition and Bigger Picture
When is an estimator good?

**Definition.** An estimator of parameter $\theta$ is an **unbiased estimator** if

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

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

$\theta = \text{unknown}$ parameter.
Three samples from $U(0, \theta)$
Example – Coin Flips

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**

- **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:

- Distribution $P(x; \theta)$
- Independent samples $X_1, \ldots, X_n$ from $P(x; \theta)$
- Estimation Algorithm
- Parameter estimate $\hat{\theta}_n$
Example – Consistency

Normal outcomes \( X_1, \ldots, X_n \) 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!

\( \hat{\theta}_{\sigma^2} \) is “consistent”
Example – Consistency

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

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

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., $\sigma^2$, as $n \to \infty$.

$\hat{\Theta}_{\sigma^2}$ 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)