20: Maximum Likelihood Estimation

Lisa Yan
May 20, 2020
Quick slide reference

3  Intro to parameter estimation  20a_intro
14 Maximum Likelihood Estimator  20b_mle
21 argmax and log-likelihood  20c_argmax
30 MLE: Bernoulli  20d_mle_bernoulli
42 MLE exercises: Poisson, Uniform, Gaussian  LIVE
Intro to parameter estimation
Story so far

At this point:

If you are given a **model** with all the necessary probabilities, you can make predictions.

\[ Y \sim \text{Poi}(5) \]
\[ X_1, \ldots, X_n \text{ i.i.d.} \]
\[ X_i \sim \text{Ber}(0.2), \]
\[ X = \sum_{i=1}^{n} X_i \]

But what if you want to **learn** the probabilities in the model?

What if you want to learn the **structure** of the model, too?

(I wish... another day)

**Machine Learning**
ML: Rooted in probability theory
Alright, so Deep Learning now?

Not so fast...
Once upon a time...

...there was parameter estimation.
Recall some estimators

$X_1, X_2, \ldots, X_n$ are $n$ i.i.d. random variables, where $X_i$ drawn from distribution $F$ with $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$.

Sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

unbiased estimate of $\mu$

Sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

unbiased estimate of $\sigma^2$
What are parameters?

def Many random variables we have learned so far are parametric models:

\[
\text{Distribution} = \text{model} + \text{parameter } \theta
\]

ex The distribution Ber(0.2) = Bernoulli model, parameter \( \theta = 0.2 \).

For each of the distributions below, what is the parameter \( \theta \)?

1. Ber\((p)\) \( \theta = p \)
2. Poi\((\lambda)\)
3. Uni\((\alpha, \beta)\)
4. \(N\)(\(\mu, \sigma^2\))
5. \(Y = mX + b\)
What are parameters?

def Many random variables we have learned so far are **parametric models**:  

\[
\text{Distribution} = \text{model} + \text{parameter } \theta
\]

ex The distribution \( \text{Ber}(0.2) \) = Bernoulli model, parameter \( \theta = 0.2 \).

For each of the distributions below, what is the parameter \( \theta \)?

1. \( \text{Ber}(p) \) \( \theta = p \)
2. \( \text{Poi}(\lambda) \) \( \theta = \lambda \)
3. \( \text{Uni}(\alpha, \beta) \) \( \theta = (\alpha, \beta) \)
4. \( \mathcal{N}(\mu, \sigma^2) \) \( \theta = (\mu, \sigma^2) \)
5. \( Y = mX + b \) \( \theta = (m, b) \)

\( \theta \) is the parameter of a distribution. \( \theta \) can be a vector of parameters!
Why do we care?

In the real world, we don’t know the “true” parameters.
• But we do get to observe data: (# times coin comes up heads, lifetimes of disk drives produced, # visitors to website per day, etc.)

def estimator $\hat{\theta}$: random variable estimating parameter $\theta$ from data.

In parameter estimation,

We use the point estimate of parameter estimate (best single value):
• Better understanding of the process producing data
• Future predictions based on model
• Simulation of future processes
Maximum Likelihood Estimator
Defining the likelihood of data: Bernoulli

Consider a sample of $n$ i.i.d. random variables $X_1, X_2, ..., X_n$.

- $X_i$ was drawn from distribution $F = \text{Ber}(\theta)$ with unknown parameter $\theta$.
- Observed data:
  
  $$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \quad (n = 10)$$

How likely was the observed data if $\theta = 0.4$?

$$P(\text{sample}|\theta = 0.4) = (0.4)^8 (0.6)^2 = 0.000236$$

Is there a better parameter $\theta$?
Defining the likelihood of data

Consider a sample of $n$ i.i.d. random variables $X_1, X_2, ..., X_n$.

- $X_i$ was drawn from a distribution with density function $f(X_i|\theta)$.
- Observed data: $(X_1, X_2, ..., X_n)$

Likelihood question:

How likely is the observed data $(X_1, X_2, ..., X_n)$ given parameter $\theta$?

 Likelihood function, $L(\theta)$:

$$L(\theta) = f(X_1, X_2, ..., X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

This is just a product, since $X_i$ are i.i.d.
Defining the likelihood of data

\[ L(\theta) = \prod_{i=1}^{n} f(X_i | \theta) \]
Maximum Likelihood Estimator

Consider a sample of $n$ i.i.d. random variables $X_1, X_2, \ldots, X_n$, drawn from a distribution $f(X_i | \theta)$.

def The Maximum Likelihood Estimator (MLE) of $\theta$ is the value of $\theta$ that maximizes $L(\theta)$.

\[
\theta_{MLE} = \arg \max_{\theta} L(\theta)
\]
Maximum Likelihood Estimator

Consider a sample of \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \), drawn from a distribution \( f(X_i|\theta) \).

The **Maximum Likelihood Estimator (MLE)** of \( \theta \) is the value of \( \theta \) that maximizes \( L(\theta) \).

\[
\theta_{MLE} = \arg \max_{\theta} L(\theta)
\]

Likelihood of your sample

\[
L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)
\]

For continuous \( X_i \), \( f(X_i|\theta) \) is PDF; for discrete \( X_i \), \( f(X_i|\theta) \) is PMF.
Maximum Likelihood Estimator

Consider a sample of \(n\) i.i.d. random variables \(X_1, X_2, ..., X_n\), drawn from a distribution \(f(X_i|\theta)\).

**def** The Maximum Likelihood Estimator (MLE) of \(\theta\) is the value of \(\theta\) that maximizes \(L(\theta)\).

\[
\theta_{MLE} = \arg \max_{\theta} L(\theta)
\]

Stay tuned!
argmax
New function: arg max

arg max \( f(x) \)

The argument \( x \) that maximizes the function \( f(x) \).

Let \( f(x) = -x^2 + 4 \), where \(-2 < x < 2\).

1. \( \max_x f(x) \)?

2. \( \arg \max_x f(x) \)?
New function: arg max

$\arg \max_x f(x)$ The argument $x$ that maximizes the function $f(x)$.

Let $f(x) = -x^2 + 4$, where $-2 < x < 2$.

1. $\max_x f(x) = 4$

2. $\arg \max_x f(x) = 0$

Stanford University
Argmax and log

\[ \text{arg max}_x f(x) \quad \text{The argument } x \text{ that maximizes the function } f(x). \]

\[ = \text{arg max}_x \log f(x) \]

Let \( f(x) = -x^2 + 4 \), where \(-2 < x < 2\).

\[ \text{arg max}_x f(x) = 0 \]
Logs all around

- **Log is increasing**: 
  \[ x < y \iff \log x < \log y \]

- Log of product = sum of logs: 
  \[ \log(ab) = \log a + \log b \]

- Natural logs 
  \[ \log_e x = \ln x \]
Argmax properties

\[
\arg\max_x f(x)
\]

The argument \( x \) that maximizes the function \( f(x) \).

\[
= \arg\max_x \log f(x)
\]

(log is an increasing function: \( x < y \iff \log x < \log y \))

\[
= \arg\max_x (c \log f(x))
\]

\( x < y \iff c \log x < c \log y \)

for any positive constant \( c \)
Argmax properties

\[ \arg \max_x \%
\]

\[ f(x) = \arg \max_x \%
\]

\[ \log f(x) \leq \log y \implies \log f(x) \leq \log y \]

\[ \arg \max_c \%
\]

\[ c \log f(x) \leq c \log y \]

The argument \( x \) that maximizes the function \( f(x) \).

How do we compute argmax?
Finding the argmax with calculus

Let \( f(x) = -x^2 + 4 \), where \(-2 < x < 2\).

\[
\hat{x} = \arg \max_x f(x)
\]

Differentiate w.r.t. argmax’s argument

\[
\frac{d}{dx}f(x) = \frac{d}{dx}(x^2 + 4) = 2x
\]

Set to 0 and solve

\[
2x = 0 \quad \Rightarrow \quad \hat{x} = 0
\]

Make sure \( \hat{x} \) is a maximum

- Check \( f(\hat{x} \pm \epsilon) < f(\hat{x}) \)
- Often ignored in expository derivations
- We’ll ignore it here too
  (and won’t require it in class)
Maximum Likelihood Estimator

Consider a sample of \( n \) i.i.d. random variables \( X_1, X_2, ..., X_n \), drawn from a distribution \( f(X_i | \theta) \).

\( \theta_{MLE} \) maximizes the likelihood of our sample, \( L(\theta) \):

\[
\theta_{MLE} = \arg \max_{\theta} L(\theta)
\]

\( \theta_{MLE} \) also maximizes the log-likelihood function, \( LL(\theta) \):

\[
LL(\theta) = \log L(\theta) = \log \left( \prod_{i=1}^{n} f(X_i | \theta) \right) = \sum_{i=1}^{n} \log f(X_i | \theta)
\]

\( LL(\theta) \) is often easier to differentiate than \( L(\theta) \).
MLE: Bernoulli
Computing the MLE

General approach for finding $\theta_{MLE}$, the MLE of $\theta$:

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i | \theta)$$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$

$$\frac{\partial LL(\theta)}{\partial \theta}$$

3. Solve resulting (simultaneous) equations

To maximize:

$$\frac{\partial LL(\theta)}{\partial \theta} = 0$$

(algebra or computer)

4. Make sure derived $\hat{\theta}_{MLE}$ is a maximum
   - Check $LL(\theta_{MLE} \pm \epsilon) < LL(\theta_{MLE})$
   - Often ignored in expository derivations
   - We’ll ignore it here too (and won’t require it in class)

$LL(\theta)$ is often easier to differentiate than $L(\theta)$. 

$\theta_{MLE} = \arg \max_{\theta} LL(\theta)$
Maximum Likelihood with Bernoulli

Consider a sample of $n$ i.i.d. RVs $X_1, X_2, ..., X_n$. What is $\theta_{MLE} = p_{MLE}$?

1. Determine formula for $LL(\theta)$

   $LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p)$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

3. Solve resulting equations

   • Let $X_i \sim \text{Ber}(p)$.  

   $f(X_i|p) = \begin{cases} p & \text{if } X_i = 1 \\ 1 - p & \text{if } X_i = 0 \end{cases}$
Maximum Likelihood with Bernoulli

Consider a sample of $n$ i.i.d. RVs $X_1, X_2, ..., X_n$.

What is $\theta_{MLE} = p_{MLE}$?

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p)$$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

$$f(X_i|p) = \begin{cases} p & \text{if } X_i = 1 \\ 1 - p & \text{if } X_i = 0 \end{cases}$$

$$f(X_i|p) = p^{X_i}(1 - p)^{1-X_i} \text{ where } X_i \in \{0,1\}$$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

$$\text{If } X_i = 1 \quad f(X_i=1|p) = p^1 (1-p)^{1-1} = p$$

$$\text{If } X_i = 0 \quad f(X_i=0|p) = p^0 (1-p)^{1-0} = 1-p$$

3. Solve resulting equations

- Is differentiable with respect to $p$
- Valid PMF over discrete domain

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1 - p)^{1-X_i}$
Maximum Likelihood with Bernoulli

Consider a sample of $n$ i.i.d. RVs $X_1, X_2, \ldots, X_n$. What is $\theta_{MLE} = p_{MLE}$?

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p) = \sum_{i=1}^{n} \log(p^{X_i}(1-p)^{1-X_i})$$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

$$= \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1 - p)]$$

$$= Y \log p + (n - Y) \log(1 - p), \text{ where } Y = \sum_{i=1}^{n} X_i$$

3. Solve resulting equations

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$
Maximum Likelihood with Bernoulli

Consider a sample of \( n \) i.i.d. RVs \( X_1, X_2, \ldots, X_n \).

What is \( \theta_{MLE} = p_{MLE} \)?

1. Determine formula for \( LL(\theta) \)
   \[
   LL(\theta) = \sum_{i=1}^{n} \left[ X_i \log p + (1 - X_i) \log(1 - p) \right] = Y \log p + (n - Y) \log(1 - p), \text{ where } Y = \sum_{i=1}^{n} X_i
   \]

2. Differentiate \( LL(\theta) \) w.r.t. (each) \( \theta \), set to 0
   \[
   \frac{\partial LL(\theta)}{\partial p} = Y \frac{1}{p} + (n - Y) \frac{-1}{1 - p} = 0
   \]

3. Solve resulting equations

- Let \( X_i \sim \text{Ber}(p) \).
- \( f(X_i|p) = p^{X_i}(1 - p)^{1-X_i} \)
Maximum Likelihood with Bernoulli

Consider a sample of \( n \) i.i.d. RVs \( X_1, X_2, \ldots, X_n \).

What is \( \theta_{MLE} = p_{MLE} \)?

1. Determine formula for \( LL(\theta) \)

\[
LL(\theta) = \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1 - p)]
\]

\[
= Y \log p + (n - Y) \log(1 - p), \text{ where } Y = \sum_{i=1}^{n} X_i
\]

2. Differentiate \( LL(\theta) \) w.r.t. (each) \( \theta \), set to 0

\[
\frac{\partial LL(\theta)}{\partial p} = Y \frac{1}{p} + (n - Y) \frac{-1}{1 - p} = 0
\]

\[
Y = \frac{n - Y}{1 - p} \Rightarrow Y (1 - p) = p (n - Y) \quad p = \frac{Y}{n}
\]

3. Solve resulting equations

\[
np - X_p = n p - X_p
\]

- Let \( X_i \sim \text{Ber}(p) \).
- \( f(X_i | p) = p^{X_i} (1 - p)^{1 - X_i} \)
Maximum Likelihood with Bernoulli

Consider a sample of $n$ i.i.d. RVs $X_1, X_2, ..., X_n$. What is $\theta_{MLE} = p_{MLE}$?

1. Determine formula for $LL(\theta)$
   \[
   LL(\theta) = \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1 - p)]
   \]
   \[
   = Y \log p + (n - Y) \log(1 - p), \text{ where } Y = \sum_{i=1}^{n} X_i
   \]

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0
   \[
   \frac{\partial LL(\theta)}{\partial p} = Y \frac{1}{p} + (n - Y) \frac{-1}{1 - p} = 0
   \]

3. Solve resulting equations
   \[
   p_{MLE} = \frac{1}{n} \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
   \]

   MLE of the Bernoulli parameter, $p_{MLE}$, is the unbiased estimate of the mean, $\bar{X}$ (sample mean)

• Let $X_i \sim \text{Ber}(p)$.
• $f(X_i | p) = p^{X_i} (1 - p)^{1-X_i}$
MLE of Bernoulli is the sample mean

\[ LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p) \]

\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \]

Bernoulli

\[ f(X_i|p) = p^{X_i}(1 - p)^{1-X_i}, \]

where \( X_i \in \{0,1\} \)
Quick check

• You draw \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \) from the distribution \( F \), yielding the following sample:

\[
[0, 0, 1, 1, 1, 1, 1, 1, 1, 1]
\]

\( (n = 10) \)

• Suppose distribution \( F = \text{Ber}(p) \) with unknown parameter \( p \).

1. What is \( p_{MLE} \), the MLE of the parameter \( p \)?

A. 1.0
B. 0.5
C. 0.8
D. 0.2
E. None/other

\[
p_{MLE} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]
Quick check

• You draw \( n \) i.i.d. random variables \( X_1, X_2, ..., X_n \) from the distribution \( F \), yielding the following sample:

\[
[0, 0, 1, 1, 1, 1, 1, 1, 1, 1]
\]

\( (n = 10) \)

• Suppose distribution \( F = \text{Ber}(p) \) with unknown parameter \( p \).

1. What is \( p_{MLE} \), the MLE of the parameter \( p \)?

A. 1.0
B. 0.5
C. 0.8
D. 0.2
E. None/other

\[
p_{MLE} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]
Quick check

- You draw \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \) from the distribution \( F \), yielding the following sample:

\[
[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \quad (n = 10)
\]

- Suppose distribution \( F = \text{Ber}(p) \) with unknown parameter \( p \).

1. What is \( p_{\text{MLE}} \), the MLE of the parameter \( p \)?

   C. 0.8

2. What is the likelihood \( L(\theta) \) of this particular sample?

\[
f(X_i|p) = p^{X_i}(1 - p)^{1-X_i} \quad \text{where } X_i \in \{0,1\}
\]

\[
L(\theta) = \prod_{i=1}^{n} f(X_i|p) \quad \text{where } \theta = p
\]

\[
= p^8(1 - p)^2
\]
20: Maximum Likelihood Estimation

Lisa Yan
May 20, 2020
Computing the MLE

General approach for finding $\theta_{MLE}$, the MLE of $\theta$:

1. Determine formula for $LL(\theta)$

   $$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$

   $$\frac{\partial LL(\theta)}{\partial \theta}$$

3. Solve resulting (simultaneous) equations

   To maximize:
   $$\frac{\partial LL(\theta)}{\partial \theta} = 0$$

4. Make sure derived $\hat{\theta}_{MLE}$ is a maximum
   - Check $LL(\theta_{MLE} \pm \epsilon) < LL(\theta_{MLE})$
   - Often ignored in expository derivations
   - We’ll ignore it here too (and won’t require it in class)

$L L(\theta)$ is often easier to differentiate than $L(\theta)$. 
\[ LL(\theta) = \sum_{i=1}^{n} \log f(X_i | p) \]

\[ \frac{\partial LL(\theta)}{\partial \theta} = 0 \]
Maximum Likelihood with Poisson

Consider a sample of \( n \) i.i.d. RVs \( X_1, X_2, ..., X_n \).

What is \( \theta_{MLE} = \lambda_{MLE} \)?

1. Determine formula for \( LL(\theta) \)

\[
LL(\theta) = \sum_{i=1}^{n} \log \left( \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!)
\]

\[
= -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)
\]

(\text{using natural log, } \ln e = 1)

- Let \( X_i \sim \text{Poi}(\lambda) \).
- PMF: \( f(X_i|\lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \)
Maximum Likelihood with Poisson

Consider a sample of \( n \) i.i.d. RVs \( X_1, X_2, ..., X_n \).

What is \( \theta_{MLE} = \lambda_{MLE} \)?

1. Determine formula for \( LL(\theta) \)

\[
LL(\theta) = \sum_{i=1}^{n} \log \left( \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^{n} \left( -\lambda \log e + X_i \log \lambda - \log X_i! \right)
\]

\[
= -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)
\]

(Using natural log, \( \ln e = 1 \))

2. Differentiate \( LL(\theta) \) w.r.t. (each) \( \theta \), set to 0

\[
\frac{\partial LL(\theta)}{\partial \lambda} = ?
\]

A. \(-n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i + n \log \lambda - \sum_{i=1}^{n} \frac{1}{X_i!} \cdot \frac{\partial X_i!}{\partial X_i} \)

B. \(-n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i \)

C. None/other/don’t know

• Let \( X_i \sim \text{Poi}(\lambda) \).
• PMF: \( f(X_i|\lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \)
Maximum Likelihood with Poisson

Consider a sample of $n$ i.i.d. RVs $X_1, X_2, \ldots, X_n$. What is $\theta_{MLE} = \lambda_{MLE}$?

1. Determine formula for $LL(\theta)$

   \[
   LL(\theta) = \sum_{i=1}^{n} \log \left( \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!)
   \]

   \[
   = -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)
   \]

   (using natural log, $\ln e = 1$)

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

   \[
   \frac{\partial LL(\theta)}{\partial \lambda} = ?
   \]

   \[
   \frac{d}{d\lambda} \left[ -n\lambda \right] + \frac{d}{d\lambda} \log(\lambda) \left( \sum_{i=1}^{n} X_i \right) + \frac{d}{d\lambda} \left( -\sum_{i=1}^{n} \log(X_i!) \right)
   \]

   A. $-n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i + n \log \lambda - \sum_{i=1}^{n} \frac{1}{X_i!} \cdot \frac{\partial X_i!}{\partial X_i}$

   B. $-n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i$

   C. None/other/don’t know
Maximum Likelihood with Poisson

Consider a sample of $n$ i.i.d. RVs $X_1, X_2, ..., X_n$.

What is $\theta_{MLE} = \lambda_{MLE}$?

1. Determine formula for $LL(\theta)$

   
   
   $$
   LL(\theta) = \sum_{i=1}^{n} \log \left( \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!)
   $$

   
   
   
   \begin{align*}
   &= -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)
   
   \text{(using natural log, \(\ln e = 1\))}
   \end{align*}

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

   
   
   $$
   \frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0
   $$

   
   
   $\theta_{MLE} = \lambda_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i$

3. Solve resulting equations

   
   
   Let $X_i \sim \text{Poi}(\lambda)$.

   
   
   PMF: $f(X_i|\lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$
Maximum Likelihood with Poisson

Consider a sample of \( n \) i.i.d. RVs \( X_1, X_2, ..., X_n \).

What is \( \theta_{MLE} = \lambda_{MLE} \)?

1. Determine formula for \( LL(\theta) \)

\[
LL(\theta) = \sum_{i=1}^{n} \log \left( \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!)
\]

\[
= -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)
\]

(using natural log, \( \ln e = 1 \))

2. Differentiate \( LL(\theta) \) w.r.t. (each) \( \theta \), set to 0

\[
\frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0
\]

3. Solve resulting equations

\[
\lambda_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

Let \( X_i \sim \text{Poi}(\lambda) \).

PMF: 
\[
f(X_i|\lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}
\]

MLE of the Poisson parameter, \( \lambda_{MLE} \), is the unbiased estimate of the mean, \( \bar{X} \) (sample mean)
Quick check

1. A particular experiment can be modeled as a Poisson RV with parameter $\lambda$, in terms of events/minute. Collect data: observe 53 events over the next 10 minutes. What is $\lambda_{MLE}$?

2. Is the Bernoulli MLE an unbiased estimator of the Bernoulli parameter $p$?

3. Is the Poisson MLE an unbiased estimator of the Poisson variance?

4. What does unbiased mean?

\[ \lambda_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i \]
Quick check

1. A particular experiment can be modeled as a Poisson RV with parameter $\lambda$, in terms of events/minute. Collect data: observe 53 events over the next 10 minutes. What is $\lambda_{MLE}$?

2. Is the Bernoulli MLE an unbiased estimator of the Bernoulli parameter $p$?  

3. Is the Poisson MLE an unbiased estimator of the Poisson variance?

4. What does unbiased mean?

$E[\text{estimator}] = \text{true thing}$

Unbiased: If you could repeat your experiment, on average you would get what you are looking for.
Interlude for jokes/announcements
Announcements

Quiz #2
Time frame: Thursday 5/21 12:00am-11:59pm PT
Covers: Up to and including Lecture 17

Note: If you have an emergency situation during the quiz, please contact Lisa and Cooper. We will try our best to accommodate.

Problem Set 6: No late days or on-time bonus

Grading clarification
Two examples
https://us.edstem.org/courses/109/discussion/67686
Interesting probability news

Bernoulli’s trials can tell you how many job applications to send

Are these trials independent?

Now let’s say the probabilities of a Yes in each of those steps go something like this:

1. Initial contact: 10%
2. Soft phone interview: 80% (recruiters are optimistic, that’s what pays them)
3. Phone interview: 50% (you’re a good engineer, but companies like to think they’re tough)
4. On-site interviews: 60% (you’re already here, that’s good)
5. Chat with offer giver: 80% (only red flags will mess it up)
6. Job offer

That gives you an overall conversion rate of

\[0.1 \cdot 0.8 \cdot 0.5 \cdot 0.6 \cdot 0.8 = 0.02 \approx 2\%\]

which sounds really tough. I see now what Anastacia meant.

Now if we plug that into the formula for Bernoulli trials on WolframAlpha we see that if you apply for 30 jobs, that gives you a

https://swizec.com/blog/bernoullis-trials-can-tell-many-job-applications-send/swizec/7677  CS109 Current Events Spreadsheet
Maximum Likelihood with Uniform

Consider a sample of $n$ i.i.d. random variables $X_1, X_2, \ldots, X_n$.

Let $X_i \sim \text{Uni}(\alpha, \beta)$.

$$f(X_i | \alpha, \beta) = \begin{cases} 
\frac{1}{\beta - \alpha} & \text{if } \alpha \leq x_i \leq \beta \\
0 & \text{otherwise}
\end{cases}$$

1. Determine the formula for $L(\theta)$

$$L(\theta) = \begin{cases} 
\left(\frac{1}{\beta - \alpha}\right)^n & \text{if } \alpha \leq x_1, x_2, \ldots, x_n \leq \beta \\
0 & \text{otherwise}
\end{cases}$$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

A. Great, let’s do it
B. Differentiation is hard
C. Constraint $\alpha \leq x_1, x_2, \ldots, x_n \leq \beta$ makes differentiation hard
Consider a sample of \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \).

Let \( X_i \sim \text{Uni}(\alpha, \beta) \).

\[
L(\theta) = \begin{cases} 
\left(\frac{1}{\beta - \alpha}\right)^n & \text{if } \alpha \leq x_1, x_2, \ldots, x_n \leq \beta \\
0 & \text{otherwise}
\end{cases}
\]

Suppose \( X_i \sim \text{Uni}(0,1) \). \([0.15, 0.20, 0.30, 0.40, 0.65, 0.70, 0.75]\)

You observe data:

A. \( \text{Uni}(\alpha = 0, \beta = 1) \)

Which parameters would give you maximum \( L(\theta) \)?

B. \( \text{Uni}(\alpha = 0.15, \beta = 0.75) \)

C. \( \text{Uni}(\alpha = 0.15, \beta = 0.70) \)
Example sample from a Uniform

Consider a sample of $n$ i.i.d. random variables $X_1, X_2, ..., X_n$.

Let $X_i \sim \text{Uni}(\alpha, \beta)$.

$$L(\theta) = \begin{cases} 
\left( \frac{1}{\beta - \alpha} \right)^n & \text{if } \alpha \leq x_1, x_2, ..., x_n \leq \beta \\
0 & \text{otherwise}
\end{cases}$$

Suppose $X_i \sim \text{Uni}(0,1)$. [0.15, 0.20, 0.30, 0.40, 0.65, 0.70, 0.75]

You observe data:

A. Uni($\alpha = 0$, $\beta = 1$) \quad (1)^7 = 1

B. Uni($\alpha = 0.15$, $\beta = 0.75$) \quad \left( \frac{1}{0.6} \right)^7 = 59.5

C. Uni($\alpha = 0.15$, $\beta = 0.70$) \quad \left( \frac{1}{0.55} \right)^6 \cdot 0 = 0

Original parameters may not yield maximum likelihood.
Maximum Likelihood with Uniform

Consider a sample of \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \).

Let \( X_i \sim \text{Uni}(\alpha, \beta) \).

\[
L(\theta) = \begin{cases} 
\left(\frac{1}{\beta - \alpha}\right)^n & \text{if } \alpha \leq x_1, x_2, \ldots, x_n \leq \beta \\
0 & \text{otherwise}
\end{cases}
\]

\( \theta_{\text{MLE}} \):

\( \alpha_{\text{MLE}} = \min(x_1, x_2, \ldots, x_n) \quad \beta_{\text{MLE}} = \max(x_1, x_2, \ldots, x_n) \)

Intuition:

- Want interval size \((\beta - \alpha)\) to be as small as possible to maximize likelihood function per datapoint
- Need to make sure all observed data is in interval (if not, then \( L(\theta) = 0 \)) (demo)
Small samples = problems with MLE

Maximum Likelihood Estimator $\theta_{MLE}$:
- Best explains data we have seen
- Does not attempt to generalize to unseen data.

In many cases, $\mu_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i$ (Sample mean)
- Unbiased ($E[\mu_{MLE}] = \mu$ regardless of size of sample, $n$)

For some cases, like Uniform: $\alpha_{MLE} \geq \alpha$, $\beta_{MLE} \leq \beta$
- Biased. Problematic for small sample size
- Example: If $n = 1$ then $\alpha = \beta$, yielding an invalid distribution

$\theta_{MLE} = \arg \max_{\theta} L(\theta)$
Properties of MLE

Maximum Likelihood Estimator: $\theta_{MLE} = \arg \max_{\theta} L(\theta)$

• Best explains data we have seen
• Does not attempt to generalize to unseen data.

• Often used when sample size $n$ is large relative to parameter space
• Potentially biased (though asymptotically less so, as $n \to \infty$)
• **Consistent:** $\lim_{n \to \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$ where $\varepsilon > 0$
  As $n \to \infty$ (i.e., more data), probability that $\hat{\theta}$ significantly differs from $\theta$ is zero
Maximum Likelihood with Normal

Consider a sample of $n$ i.i.d. random variables $X_1, X_2, \ldots, X_n$.

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

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log \left( \frac{1}{\sqrt{2\pi\sigma}} e^{-(X_i-\mu)^2/(2\sigma^2)} \right) = \sum_{i=1}^{n} \left[ -\log(\sqrt{2\pi\sigma}) - (X_i - \mu)^2 / (2\sigma^2) \right]$$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

3. Solve resulting equations

(using natural log)
Maximum Likelihood with Normal

Consider a sample of $n$ i.i.d. random variables $X_1, X_2, \ldots, X_n$.

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

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

3. Solve resulting equations

\[
LL(\theta) = -\sum_{i=1}^{n} \log(\sqrt{2\pi\sigma}) - \sum_{i=1}^{n} \left[ \frac{(X_i - \mu)^2}{2\sigma^2} \right]
\]

\[
\frac{\partial LL(\theta)}{\partial \mu} = \sum_{i=1}^{n} \left[ \frac{2(X_i - \mu)}{2\sigma^2} \right]
\]

\[
= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0
\]
Maximum Likelihood with Normal

Consider a sample of $n$ i.i.d. random variables $X_1, X_2, ..., X_n$.

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

$$f(X_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$ with respect to $\mu$

$$LL(\theta) = -\sum_{i=1}^{n} \log(\sqrt{2\pi\sigma}) - \sum_{i=1}^{n} [(X_i - \mu)^2/(2\sigma^2)]$$

$$\frac{\partial LL(\theta)}{\partial \mu} = \sum_{i=1}^{n} [2(X_i - \mu)/(2\sigma^2)] = \frac{1}{\sigma^2}\sum_{i=1}^{n} (X_i - \mu) = 0$$

2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0

$$\frac{\partial LL(\theta)}{\partial \sigma} = -\sum_{i=1}^{n} \frac{1}{\sigma} + \sum_{i=1}^{n} 2(X_i - \mu)^2/(2\sigma^3)$$

$$\frac{\partial LL(\theta)}{\partial \sigma} = -\sum_{i=1}^{n} \frac{1}{\sigma} + \sum_{i=1}^{n} 2(X_i - \mu)^2/(2\sigma^3) = 0$$

3. Solve resulting equations

$$\sum_{i=1}^{n} (X_i - \mu) = \frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n} (X_i - \mu)^2 = 0$$
Maximum Likelihood with Normal

Consider a sample of \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \).

- Let \( X_i \sim \mathcal{N}(\mu, \sigma^2) \).

What is \( \theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE}) \)?

3. Solve resulting equations

Two equations, two unknowns:

\[
\begin{align*}
\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) &= 0 \\
-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2 &= 0
\end{align*}
\]

First, solve for \( \mu_{MLE} \):

\[
\frac{1}{\sigma^2} \sum_{i=1}^{n} X_i - \frac{1}{\sigma^2} \sum_{i=1}^{n} \mu = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i = n\mu
\]

unbiased

\[
\mu_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]
Maximum Likelihood with Normal

Consider a sample of $n$ i.i.d. random variables $X_1, X_2, \ldots, X_n$.

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

\[
f(X_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i - \mu)^2/(2\sigma^2)}
\]

What is $\theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2)$?

3. Solve resulting equations

Two equations, two unknowns:

\[
\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0
\]
\[
-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2 = 0
\]

First, solve for $\mu_{MLE}$:

\[
\frac{1}{\sigma^2} \sum_{i=1}^{n} X_i - \frac{1}{\sigma^2} \sum_{i=1}^{n} \mu = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i = n\mu 
\]

\[
\Rightarrow \quad \mu_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i 
\]

unbiased

Next, solve for $\sigma_{MLE}$:

\[
\frac{1}{\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2 = \frac{n}{\sigma} \quad \Rightarrow \quad \sum_{i=1}^{n} (X_i - \mu)^2 = \sigma^2 n 
\]

\[
\Rightarrow \quad \sigma_{MLE}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_{MLE})^2 
\]

biased