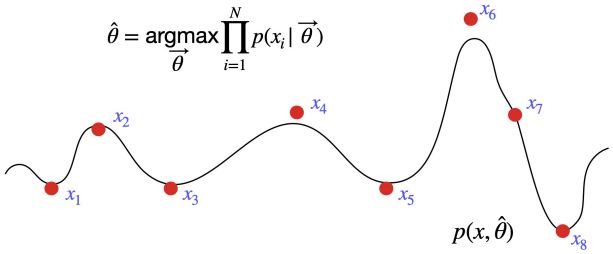
**Table of Contents**






# 20: Maximum Likelihood Estimation


Jerry Cain


February 26, 2024

Lecture Discussion on Ed

# Parameter Estimation

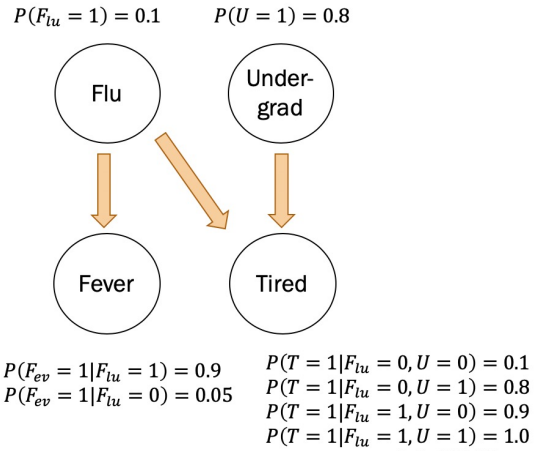
# Story so far

At this point:

If we're provided with a **model** and all the necessary parameters, we can make predictions!

$$Y \sim \text{Exp}\left(\frac{1}{5}\right)$$

$$X_1, \ldots, X_n \text{ iid}$$
$$X_i \sim \text{Ber}(0.2),$$
$$X = \sum_{i=1}^{n} X_i$$




But how do we infer and estimate the parameters for a given model?

this is the focus of today's lecture.

~~What if you want to learn the **structure** of the model, too?~~ Glimpse: Week 10

# Machine Learning

we'll just give you a taste, but there's no way we can do all of machine learning justice in one week. We'd need a whole quarter!

# Some estimators

iid introduced last Wednesday and reviewed some on Friday.

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

in general, we never truly know the values of these, or we make educated guesses.

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 $\sigma^2$

*potentially useful estimates if trying to infer parameters of a Gaussian*

# What are parameters?

def Most random variables we've seen thus far are **parametric models**:

$$\text{Distribution} = \text{model} + \text{parameter } \theta$$

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

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

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

# What are parameters?

def Most random variables we've seen thus far are **parametric models**:

$$\text{Distribution} = \text{model} + \text{parameter } \theta$$

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

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

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

$\theta$ is the parameter of a distribution.
$\theta$ can be a vector.

# Why do we care?

In the real world, we don't know the true parameters.

- But we **observe data**: # times coin comes up heads, # requests for RydeShare per minute, # visitors to website per day, offer amount for that used bike you can't sell

whenever you see a caret over a parameter, it generally means 'it's an estimate.'

def **estimator** $\hat{\theta}$: a **random variable** estimating the true parameter $\theta$.

In parameter estimation,

We'll initially and often rely on **point estimates**—i.e., the best single value

- Provides an understanding of why data looks the way it does
- Can make future **predictions** using that model
- Can run simulations to generate more data

# Maximum Likelihood Estimator

# Defining the likelihood of data: Bernoulli

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

- $X_i$ was drawn from distribution $F \sim \text{Ber}(\theta)$ with unknown parameter $\theta$.
- Observed sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

intuition tells us to assume $\hat{p} = 0.8$, but is my intuition correct?

How likely is this sample if, say, $\theta = 0.4$?

conditioned on the premise / belief that $\theta = 0.4$

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

Likelihood of data given parameter $\theta = 0.4$

```
jerry — Python — 148×16
>>> for step in range(11):
...     theta = 0.1 * step
...     print("P(sample | theta = {:.1f}) = ({:.1f})^8 * ({:.1f})^2 = {:.6f}".format(theta, theta, 1 - theta, (theta ** 8) * ((1 - theta) ** 2)))
...
P(sample | theta = 0.0) = (0.0)^8 * (1.0)^2 = 0.000000
P(sample | theta = 0.1) = (0.1)^8 * (0.9)^2 = 0.000000
P(sample | theta = 0.2) = (0.2)^8 * (0.8)^2 = 0.000002
P(sample | theta = 0.3) = (0.3)^8 * (0.7)^2 = 0.000032
P(sample | theta = 0.4) = (0.4)^8 * (0.6)^2 = 0.000236
P(sample | theta = 0.5) = (0.5)^8 * (0.5)^2 = 0.000977
P(sample | theta = 0.6) = (0.6)^8 * (0.4)^2 = 0.002687
P(sample | theta = 0.7) = (0.7)^8 * (0.3)^2 = 0.005188
P(sample | theta = 0.8) = (0.8)^8 * (0.2)^2 = 0.006711
P(sample | theta = 0.9) = (0.9)^8 * (0.1)^2 = 0.004305
P(sample | theta = 1.0) = (1.0)^8 * (0.0)^2 = 0.000000
>>>
```

Is there a better choice for $\theta$?

# Defining the likelihood of data

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

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

Likelihood question:

How likely is the sample $(X_1, X_2, \ldots, X_n)$ given the parameter $\theta$?

**Likelihood function, $L(\theta)$:** this is the generic definition of likelihood

we can say this when all $X_i$ are independent (which is certainly true if they are iid)

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

This is just a product, since the $X_i$ are iid.

# Maximum Likelihood Estimator

Consider a sample of $n$ iid 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 $L(\theta)$. → restated, the value of $\theta$ that maximizes the probability of observing the data you see.

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

# Maximum Likelihood Estimator

Consider a sample of $n$ iid 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 $L(\theta)$.

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

Likelihood of your sample

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

$\rightarrow P(X_i|\theta)$ is used as well

For continuous $X_i$, $f(X_i|\theta)$ is PDF, and for discrete $X_i$, $f(X_i|\theta)$ is PMF

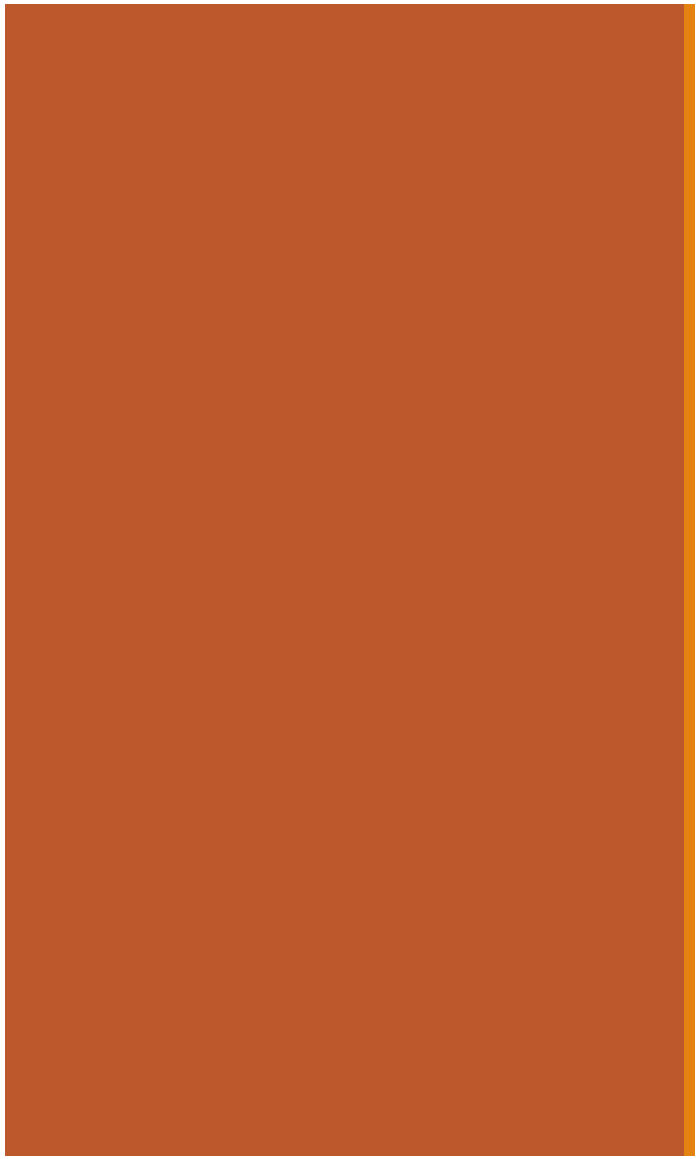
# Maximum Likelihood Estimator

Consider a sample of $n$ iid 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 $L(\theta)$.

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

the argument $\theta$
that maximizes $L(\theta)$



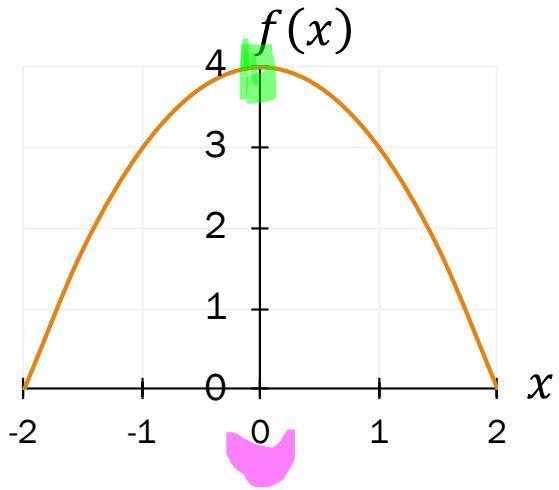
# argmax and log likelihood

# 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. $\arg\max_x f(x)$ ?

   $= 0$

# 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: $x < y \Leftrightarrow \log x < \log y$)

$$= \arg\max_x (c \log f(x))$$

$(x < y \Leftrightarrow c \log x < c \log y)$

for any positive constant $c$

# Finding the argmax with calculus

$$\hat{x} = \arg\max_x f(x)$$

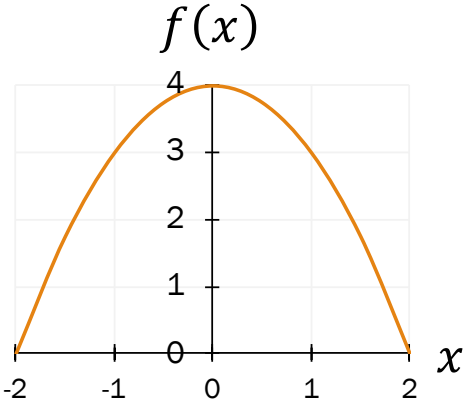
Let $f(x) = -x^2 + 4$, where $-2 < x < 2$.

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 as well (and we won't require it in class or on problem sets and exams)

# Maximum Likelihood Estimator

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

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

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

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

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

$$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$$

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

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

# MLE: Bernoulli

# Computing the MLE

$$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$$

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) $\theta$

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

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

3. Solve resulting equations

*algebra or computer*

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

# Maximum Likelihood with Bernoulli

Consider a sample of $n$ iid RVs $X_1, X_2, \ldots, X_n$.

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

- Let $X_i \sim \text{Ber}(p)$.

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ wrt (each) $\theta$, set to 0
3. Solve resulting equations

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

?

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

$f(X_i|p)$ in its current form is not differentiable, and $\log f(X_i|p)$ isn't either.

sadness

# Maximum Likelihood with Bernoulli

Consider a sample of $n$ iid RVs $X_1, X_2, \ldots, X_n$.

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

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

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

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

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

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

verify this works for $X_i = 0$ and $X_i = 1$

$$\begin{cases} X_i = 1 \Rightarrow f(X_i = 1|p) = p^1(1-p)^0 = p \checkmark \\ X_i = 1 \Rightarrow f(X_i = 0|p) = p^0(1-p)^1 = (1-p) \checkmark \end{cases}$$

2. Differentiate $LL(\theta)$ wrt (each) $\theta$, set to 0

3. Solve resulting equations

✅

- differentiable with respect to $p$
- valid PMF over discrete domain

# Maximum Likelihood with Bernoulli

properties of logarithms: $\log ab = \log a + \log b$, $\log x^y = y \log x$

Consider a sample of $n$ iid RVs $X_1, X_2, \ldots, X_n$.

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

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

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ wrt (each) $\theta$, set to 0
3. Solve resulting equations

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

$$= \log p^{X_i} + \log (1-p)^{1-X_i}$$

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

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

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

# Maximum Likelihood with Bernoulli

Consider a sample of $n$ iid RVs $X_1, X_2, \ldots, X_n$.

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

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

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 $LL(\theta)$ wrt (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

# Maximum Likelihood with Bernoulli

Consider a sample of $n$ iid RVs $X_1, X_2, \ldots, X_n$.

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

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

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 $LL(\theta)$ wrt (each) $\theta$, set to 0

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

$$Y - Yp = np - Yp$$

$$p = Y/n$$

3. Solve resulting equations

# Maximum Likelihood with Bernoulli

Consider a sample of $n$ iid RVs $X_1, X_2, \ldots, X_n$.

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

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

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 $LL(\theta)$ wrt (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

$$p_{MLE} = \frac{1}{n}Y = \frac{1}{n}\sum_{i=1}^{n} X_i$$

MLE of the Bernoulli parameter, $p_{MLE}$, is the sample mean, $\bar{X}$, which is an unbiased estimator of the true mean.

# Quick check

- You draw $n$ iid random variables $X_1, X_2, \ldots, X_n$ from the distribution $F$, yielding the following sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

- 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 $n$ iid random variables $X_1, X_2, \ldots, X_n$ from the distribution $F$, yielding the following sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

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

1. What is $p_{MLE}$, the MLE of the parameter $p$? C. 0.8
2. What is the likelihood $L(\theta)$ of this specific sample?

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

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

$$= p^8(1-p)^2 = 0.8^8 0.2^2 = 0.0067$$

it's very small, but the probability of any specific sequence of 10 Bernoulli's is going to be small.

# MLE: Poisson and Uniform

# Maximum Likelihood with Poisson

Consider a sample of $n$ iid RVs $X_1, X_2, \ldots, X_n$.

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

reminder: $\log ab = \log a + \log b$, $\log \frac{a}{b} = \log a - \log b$

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

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, i.e., $\ln e = 1$

# Maximum Likelihood with Poisson

Consider a sample of $n$ iid RVs $X_1, X_2, \ldots, X_n$.

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

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \dfrac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

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, i.e., $\ln e = 1$

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

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

$$\frac{d}{d\lambda}(-n\lambda) + \frac{d}{d\lambda}\log\lambda \sum_{i=1}^{n} X_i + \underbrace{\frac{d}{d\lambda}\sum_{i=1}^{n}\log X_i!}_{\text{zero}}$$

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. Stop trying

# Maximum Likelihood with Poisson

Consider a sample of $n$ iid RVs $X_1, X_2, \ldots, X_n$.

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

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \dfrac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

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, i.e., $\ln e = 1$

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

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

3. Solve resulting equations

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

MLE of the Poisson parameter, $\lambda_{MLE}$, is the sample mean, $\bar{X}$, which is an unbiased estimator of the true mean.

# Quick Review

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 $\lambda_{MLE}$?

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

sample is $(X_1 = x_1, X_2 = x_2, \ldots, X_{10} = x_{10})$

$\sum_{i=1}^{10} X_i = 53 \Rightarrow \lambda_{MLE} = \frac{1}{10}53 = 5.3$

$\lambda_{MLE} = 5.3$

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

$X \sim Ber(p)$

$E[p_{MLE}] = p$?

$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = E[\bar{X}] = \mu = p$

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

$X \sim Poi(\lambda)$ $E[\lambda_{MLE}] = E[\bar{X}] = \lambda = \sigma^2$

4. What does unbiased mean?

$E[\text{estimator}] = \text{the truth}$

Unbiased: If you repeat your experiment multiple times, on average, you'll get what you are looking for.

# Maximum Likelihood with Uniform

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

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

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

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

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

$LL(\theta) = n \ln \frac{1}{\beta-\alpha}$ provided all $x_i$ are such that $\alpha \le x_i \le \beta$

2. Differentiate $L(\theta)$ wrt each $\theta$, set to 0

A. Great, let's do it
B. Use $LL(\theta)$ instead
C. Constraint $\alpha \le x_1, x_2, \ldots, x_n \le \beta$ makes differentiation hard

# Maximum Likelihood with Uniform: Sample

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

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

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

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

You observe data:

Which parameters maximize $L(\theta)$?

A. $\text{Uni}(\alpha = 0.00, \beta = 1.00)$

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

C. $\text{Uni}(\alpha = 0.15, \beta = 0.70)$

# Maximum Likelihood with Uniform: Sample

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

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

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

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

You observe data:

Which parameters maximize $L(\theta)$?

A. $\text{Uni}(\alpha = 0.00, \beta = 1.00)$ $(1)^7 = 1$ likelihood value using original, underlying parameters.

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

C. $\text{Uni}(\alpha = 0.15, \beta = 0.70)$ $\left(\frac{1}{0.55}\right)^6 \cdot 0 = 0$ included on behalf of the 0.75

⚠ Original parameters may not yield maximum likelihood.

# Maximum Likelihood with Uniform

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

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

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

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

Intuition:

- Want interval size $\beta - \alpha$ to be as narrow as possible to maximize likelihood function.
- Need to ensure all datapoints are included in interval. Otherwise, $L(\theta) = 0$.

(demo)

# Small samples = problems with MLE

Maximum Likelihood Estimator $\theta_{MLE}$ :

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

- Best explains the data we've seen
- Does not attempt to generalize to data not yet observed.

✅ In many cases, $\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$ Sample mean (MLE for Bernoulli $p$, Poisson $\lambda$, Normal $\mu$)

- Unbiased ($E[\mu_{MLE}] = \mu$, regardless of sample size)

⚠️ For some cases, like Uniform: $\alpha_{MLE} \geq \alpha$, $\beta_{MLE} \leq \beta$ ← same exact story.

↙ $\alpha$ of underlying distribution (presumably unknown)

- Ad hoc, biased, and problematic for small sample sizes
- Example: If $n = 1$, then $\alpha = \beta$, and our estimates yield an invalid distribution

# MLE: Gaussian

# Maximum Likelihood with Normal

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

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

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

two parameters

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

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ wrt (each) $\theta$, set to 0
3. Solve resulting equations

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

(using natural log)

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

# Maximum Likelihood with Normal

Consider a sample of $n$ iid 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 $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ wrt (each) $\theta$, set to 0
3. Solve resulting equations

with respect to $\mu$

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

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

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0$$

# Maximum Likelihood with Normal

Consider a sample of $n$ iid 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 $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ wrt (each) $\theta$, set to 0
3. Solve resulting equations

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

with respect to $\mu$

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

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0$$

with respect to $\sigma$

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

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

# Maximum Likelihood with Normal

Consider a sample of $n$ iid 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 $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

3. Solve resulting equations

Two equations, two unknowns:

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

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

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i - \frac{1}{\sigma^2}\sum_{i=1}^{n}\mu = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i = n\mu \quad \Rightarrow \quad \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

unbiased

another sample mean!

# Maximum Likelihood with Normal

Consider a sample of $n$ iid 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 $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

3. Solve resulting equations

Two equations, two unknowns:

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

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

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i - \frac{1}{\sigma^2}\sum_{i=1}^{n}\mu = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i = n\mu \quad \Rightarrow \quad \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

unbiased

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

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{n}{\sigma} \quad \Rightarrow \sum_{i=1}^{n}(X_i-\mu)^2 = \sigma^2 n \quad \Rightarrow \sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{MLE})^2$$

biased !!