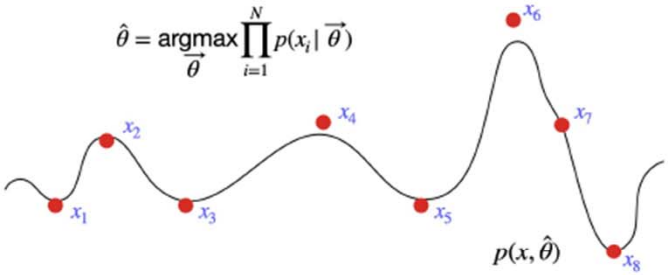
Table of Contents



20: Maximum Likelihood Estimation

Jerry Cain

February 26, 2024

Lecture Discussion on Ed

Parameter Estimation

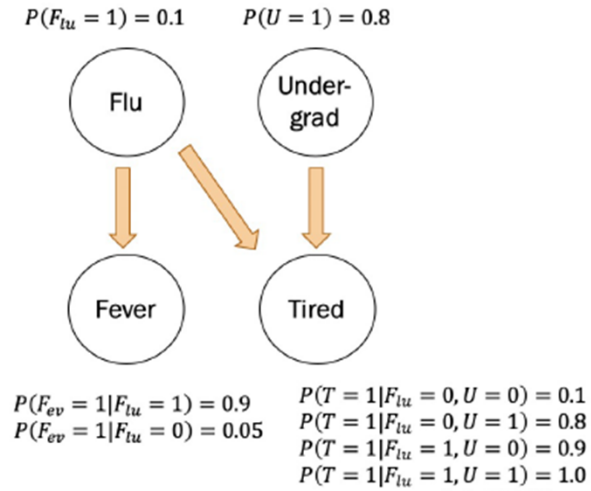
Story so far

At this point:

If you are provided with a **model** and all the necessary probabilities, you can make predictions!

$$Y \sim \text{Poi}(5)$$

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



But how do we infer the probabilities for a given model?

this is today's focus!

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

Machine Learning

you need entire classes to understand machine learning, not just one week!

Some estimators

introduced last Wednesday and Friday

$X_1, X_2, \ldots, X_n$ are n iid random variables, underlying (i.e. unknown)
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 μ

Sample variance: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$ unbiased estimate of σ^2

What are parameters?

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

Distribution = model + parameter θ

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

For each of the distributions below, what is the parameter θ?

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

θ is the parameter of a distribution.
θ 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, offer amount for a used bike

whenever you see ^ over a parameter, it almost always means an estimate!

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

In parameter estimation,

We use the point estimate of parameter estimate (best single value):

- Provides an understanding of the process generating the data
- Can make future predictions based that model
- Can even 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 $F = \text{Ber}(\theta)$ with unknown parameter θ.
- Observed sample:

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

intuitive tell: use $\hat{p} = 0.8$.
but is our intuition correct?

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

conditioned on a belief that $\theta = 0.4$! this is technically an event.

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

Likelihood of data given parameter $\theta = 0.4$

Is there a better choice for θ?

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 $f(X_i|\theta)$. (or mass)
- 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 θ?

Likelihood function, $L(\theta)$: this is the definition of $L(\theta)$ in all scenarios

this follows from generic definition when X_i 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 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 θ is the value of θ that maximizes $L(\theta)$. → i.e. maximizes the likelihood of the observed data.

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

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 θ is the value of θ 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, 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 θ is the value of θ that maximizes $L(\theta)$.

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

The argument θ that maximizes $L(\theta)$

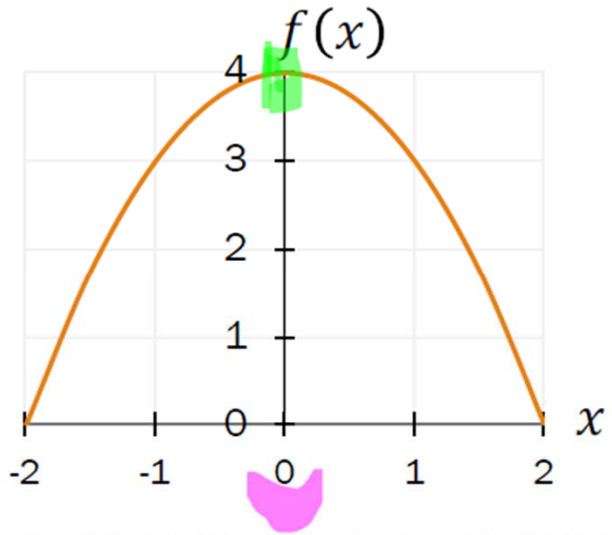
argmax and log likelihood

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

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

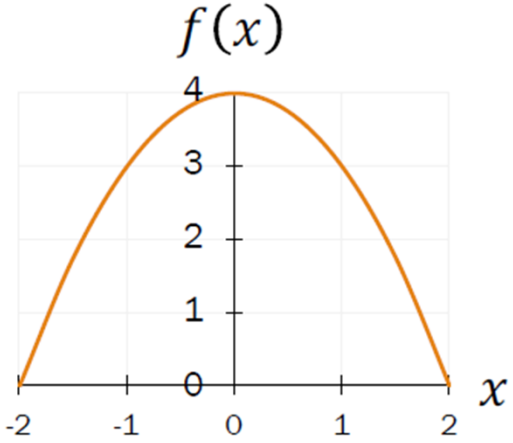
Let $f(x) = -x^2 + 4$, where $-2 < x < 2$.

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 here too (and won't require it in class)

Maximum Likelihood Estimator

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

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

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

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

θ_{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 θ_{MLE} , the MLE of θ:

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ
3. Solve resulting equations

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

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

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 $L(\theta)$.

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) θ, set to 0
3. Solve resulting equations

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

?

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

functi'm as expressed is not differentiable! not what we want! ∴

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 $LL(\theta)$

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

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

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

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

expanded
$X_i = 1?$ $f(X_i=1|p) = p^1(1-p)^{1-1} = p^1(1-p)^0 = p$
$X_i = 0?$ $f(X_i=0|p) = p^0(1-p)^{1-0} = p^0(1-p)^1 = 1-p$

3. Solve resulting equations

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

Maximum Likelihood with Bernoulli

$\log ab = \log a + \log b$, $\log c^d = d \log c$] properties of log

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) θ, 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) θ, 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) θ, set to 0

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

$\frac{Y}{p} = \frac{n-Y}{1-p}$

$Y - Yp = np - Yp \rightarrow p = \frac{Y}{n}$

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 unbiased estimate of the mean, $\bar{X}$ (sample 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?

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

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

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

MLE: Poisson and Uniform

Maximum Likelihood with Poisson

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

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

recall that $\log ab = \log a + \log b$

$\log \frac{a}{b} = \log a - \log b$

- 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, $\ln e = 1$)

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, $\ln e = 1$)

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

$$\frac{\partial LL(\theta)}{\partial \lambda} = ? \quad \frac{d}{d\lambda}(-n\lambda) + \frac{d}{d\lambda}\log\lambda \sum_{i=1}^{n} x_i \; \underbrace{-}_{*} \; \underbrace{\frac{d}{d\lambda}\sum_{i=1}^{n}\log X!}_{0}$$

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 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, $\ln e = 1$)

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, 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, λ_{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 λ, in terms of events/minute.
   Collect data: observe 53 events over the next 10 minutes. What is λ_{MLE}? $\lambda_{MLE} = 5.3$

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

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

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

2. Is the Bernoulli MLE an unbiased estimator of the Bernoulli parameter p? ✅ $X \sim Ber(p)$

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

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 could repeat your experiment, on average you would 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 $L(\theta)$

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

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

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

A. Great, let's do it
B. Differentiation is hard
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 $X_i \sim \text{Uni}(\alpha, \beta)$.

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

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:

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

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

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

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

Suppose 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 would give you maximum $L(\theta)$?

A. $\text{Uni}(\alpha = 0 \quad, \beta = 1 \quad)$ $(1)^7 = 1$ } on behalf of original parameters

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

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

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

θ_{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 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 θ_{MLE} :

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

- Best explains data we have 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 λ, Normal μ)

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

⚠️ For some cases, like Uniform: $\alpha_{MLE} \geq \alpha$, $\beta_{MLE} \leq \beta$ ← α of underlying distribution (presumably unknown) ← β, same story

- Biased. Problematic for small sample size
- Example: If $n = 1$ then $\alpha = \beta$, yielding an invalid distribution

Properties of MLE

Maximum Likelihood Estimator θ_{MLE}:

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

- Best explains data we have seen
- Does not attempt to generalize to data not yet observed.

- Often used when sample size n is large relative to parameter space
- Potentially biased (though asymptotically less so, as $n \to \infty$)
- Consistent: $\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1$ where $\varepsilon > 0$

  As $n \to \infty$ (i.e., more data), probability that $\hat{\theta}$ significantly differs from θ is zero

MLE: Gaussian

Maximum Likelihood with Normal

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

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

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

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

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

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

with respect to μ

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

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

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 $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

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

with respect to μ

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

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

with respect to σ

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

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

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 μ_{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 μ_{MLE}:

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

unbiased

Next, solve for σ_{MLE}:

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

biased!!

MLE: Multinomial

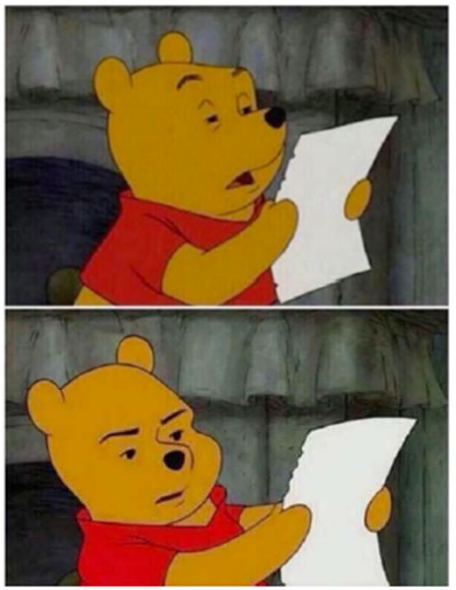
Okay, just one more MLE with the Multinomial

Consider a sample of n iid random variables where:

- Each element is drawn from one of m outcomes. $P(\text{outcome } i) = p_i$, where $\sum_{i=1}^{m} p_i = 1$
- X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

this is the classic description of multinomial

Staring at my math homework like



Let's give an example!

Okay, just one more MLE with the Multinomial

Consider a sample of n iid random variables where:

- Each element is drawn from one of m outcomes. $P(\text{outcome } i) = p_i$, where $\sum_{i=1}^{m} p_i = 1$
- X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

m is the number of possible outcomes

p_3 is probability of seeing third of 6 outcomes. $p_1, p_2, p_4, \ldots$ all similar

Example: Suppose each RV is outcome of 6-sided die. $m = 6, \sum_{i=1}^{6} p_i = 1$

note: We're assuming know nothing about its fairness.

- Roll the dice $n = 12$ times.
- Observe data: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

$$X_1 = 3, X_2 = 2, X_3 = 0,$$
$$X_4 = 3, X_5 = 1, X_6 = 3$$

Check: $X_1 + X_2 + \cdots + X_6 = 12$ (3, 2, 0,1, 3)

Okay, just one more MLE with the Multinomial

Consider a sample of n iid random variables where:

- Each element is drawn from one of m outcomes. $P(\text{outcome } i) = p_i$, where $\sum_{i=1}^{m} p_i = 1$
- X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

1. What is the likelihood of observing the sample$(X_1, X_2, \ldots, X_m)$, given the probabilities $p_1, p_2, \ldots, p_m$?

A. $\dfrac{n!}{X_1! X_2! \cdots X_m!} p_1^{X_1} p_2^{X_2} \cdots p_m^{X_m}$ $\Rightarrow$ $\binom{n}{X_1, X_2, \ldots X_m} p_1^{X_1} p_2^{X_2} \cdots p_m^{X_m}$

B. $p_1^{X_1} p_2^{X_2} \cdots p_m^{X_m}$

C. $\dfrac{n!}{X_1! X_2! \cdots X_m!} X_1^{p_1} X_2^{p_2} \cdots X_m^{p_m}$

if p_1, p_2, p_3, etc. are unknown, then they are the parameters in any MLE problem try to maximize likelihood of seeing data!

Okay, just one more MLE with the Multinomial

Consider a sample of n iid random variables where:

- Each element is drawn from one of m outcomes.
  $P(\text{outcome } i) = p_i$, where $\sum_{i=1}^{m} p_i = 1$
- X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

here, $\theta = (p_1, p_2, \cdots, p_m)$

1. What is the likelihood of observing the sample$(X_1, X_2, \ldots, X_m)$, given the probabilities $p_1, p_2, \ldots, p_m$?

$$L(\theta) = \frac{n!}{X_1!\, X_2! \cdots X_m!} p_1^{X_1} p_2^{X_2} \cdots p_m^{X_m}$$

recall that $\log p_1^{X_1} = X_1 \log p_1$

2. What is θ_{MLE}?

$$LL(\theta) = \log(n!) - \sum_{i=1}^{m} \log(X_i!) + \sum_{i=1}^{m} X_i \log(p_i), \text{ such that } \sum_{i=1}^{m} p_i = 1$$

Optimize with Lagrange multipliers in extra slides → θ_{MLE}: $p_i = \frac{X_i}{n}$

Intuitively, probability p_i = proportion of outcomes

When MLEs attack!

MLE for Multinomial: $p_i = \frac{X_i}{n}$

Consider a 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

What is θ_{MLE}?

When MLEs attack!

MLE for Multinomial: $p_i = \frac{X_i}{n}$

Consider a 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

θ_{MLE}:

$$p_1 = 3/12$$
$$p_2 = 2/12$$
$$p_3 = 0/12$$
$$p_4 = 3/12$$
$$p_5 = 1/12$$
$$p_6 = 3/12$$

- MLE say you just never roll threes.
- Do you really believe that?

was 12 rolls enough to dismiss 3 as a possibility?

Roll more!
prob = frequency in limit

$p_1 = 3/12$
$p_2 = 2/12$
$p_3 = 0/12$
$p_4 = 3/12$
$p_5 = 1/12$
$p_6 = 3/12$

But what if you cannot observe anymore rolls?

Frequentist

Bayesian Statistics

Starting Today!

Today we are going to learn something unintuitive, beautiful, and useful!

We are going to think of probabilities as random variables.

A new definition of probability

Flip a coin $n + m$ times, produce n heads.

We don't know the probability θ that the coin comes up heads.

The world's first coin

Frequentist

θ is a single value.

$$\theta = \lim_{n+m\to\infty} \frac{n}{n+m} \approx \frac{n}{n+m}$$

Bayesian

θ is a random variable.

θ's continuous support: (0, 1)

Let's play a game

Roll 2 dice. If neither roll is a 6,
you win (event W). Else, I win (event W^C).

- Before you play, what's the probability that you win?
- Play once. What's the probability that you win?
- Play three more times. What's the probability that you win?

Frequentist

$$P(W) = \left(\frac{5}{6}\right)^2$$

Bayesian

I am constantly re-evaluating the situation

Bayesian statistics: Constantly update your prior beliefs.

Bayesian probability

Bayesian statistics: Probability represents our ever-evolving understanding of the world.

Mixing discrete and continuous random variables, combined with Bayes' Theorem, allows us to reason about probabilities as random variables.

Mixing discrete and continuous

Let X be a continuous random variable, and N be a discrete random variable.

Bayes' Theorem:

$$f_{X|N}(x|n) = \frac{p_{N|X}(n|x) f_X(x)}{p_N(n)}$$

Intuition: $P(X \cong x | N = n) = \dfrac{P(N = n | X = x) P(X \cong x)}{P(N = n)}$

$$P(X \cong x) = \int_{x-\epsilon/2}^{x+\epsilon/2} f_X(x)\,dx \approx f_X(x)\,\epsilon_X$$

$$f_{X|N}(x|n)\varepsilon_X = \frac{P(N = n | X = x) f_X(x) \varepsilon_X}{P(N = n)} \Rightarrow f_{X|N}(x|n) = \frac{p_{N|X}(n|x) f_X(x)}{p_N(n)}$$

these cancel

Bayes' Theorem: All Flavors

Let X, Y be continuous and M, N be discrete random variables.

dates back to Lecture 4!

Original Bayes: $$p_{M|N}(m|n) = \frac{p_{N|M}(n|m)p_M(m)}{p_N(n)}$$

from previous slide

Mix Bayes #1: $$f_{X|N}(x|n) = \frac{p_{N|X}(n|x)f_X(x)}{p_N(n)}$$

Mix Bayes #2: $$p_{N|X}(n|x) = \frac{f_{X|N}(x|n)p_N(n)}{f_X(x)}$$ ← second mixed form

All continuous: $$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

Will covered this in Lecture 17 (remember satellites!)

Mixing discrete and continuous

Let θ be a random variable for the probability your coin comes up heads, and N be the number of heads you observe in an experiment. assume n+m trials :)

$$\text{posterior} \quad f_{\theta|N}(x|n) = \frac{p_{N|\theta}(n|x) f_\theta(x)}{p_N(n)} \quad \begin{matrix}\text{likelihood} \\ \text{prior}\end{matrix}$$

normalization constant

- Prior belief of parameter θ — $f_\theta(x)$
- Likelihood of $N = n$ heads, given parameter $\theta = x$. — $p_{N|\theta}(n|x)$
- Posterior updated belief of parameter θ. — $f_{\theta|N}(x|n)$