Maximum Likelihood Estimation

CSE 312 Spring 24
Lecture 24

Asking The Opposite Question

So far:

Given rules for an experiment.

Given the event/outcome we're interested in.

You calculate/estimate/bound what the probability is.

Today:

Given **some** of the rules of the experiment.

Given what happened.

You estimate what the rest of the rules of the experiment were.

Example

P

$\frac{6}{10}$

Suppose you flip a coin independently 10 times, and you see

HTTTHHTHHH

6H
4T

What is your estimate of the probability the coin comes up heads?

pollev.com/robbie

Maximum Likelihood Estimation

Idea: we got the results we got.

High probability events happen more often than low probability events.

So, guess the rules that maximize the probability of the events we saw (relative to other choices of the rules).

Since that event happened, might as well guess the set of rules for which that event was a ‘high probability event’.

Maximum Likelihood Estimation

Formally, we are trying to estimate a parameter of the experiment (here: the probability of a coin flip being heads).

The likelihood of an event E given a parameter θ is

$\mathcal{L}(E;\theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

We'll use the notation $\mathbb{P}(E;\theta)$ for probability when run with parameter θ where the semicolon means "extra rules" rather than conditioning

We will choose $\hat{\theta} = \text{argmax}_{\theta}\ \mathcal{L}(E;\theta)$

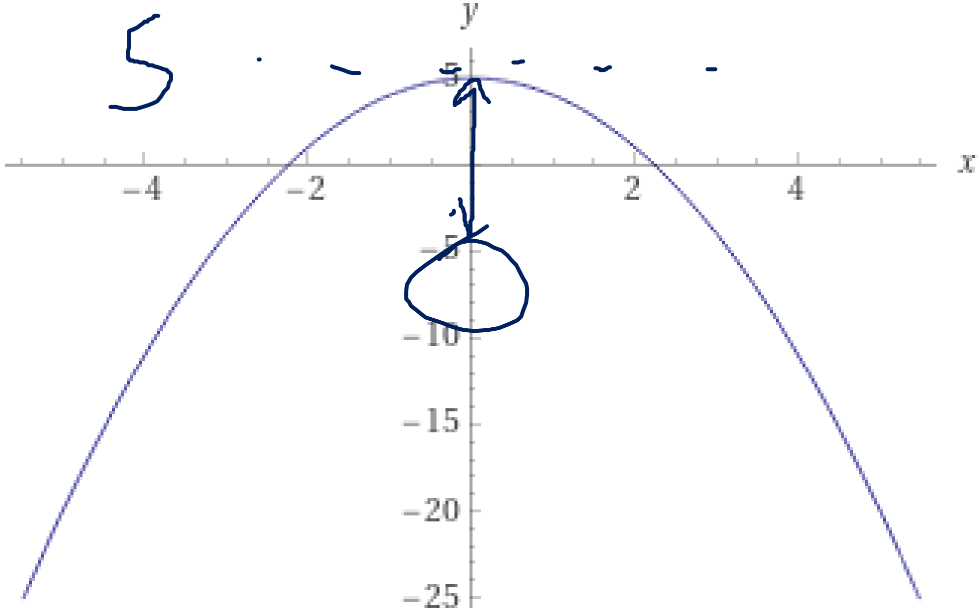
argmax is the argument that produces the maximum so the θ that causes $\mathcal{L}(E;\theta)$ to be maximized.

Argmax?

For example

$\text{argmax}_x\left(5 - \text{x}^2\right) = 0$

$\max_x\left(5 - x^2\right) = 5$, the input (argument) which produces $\mathbf{5}$ is $\mathbf{0}$, so the argmax is 0



Notation comparison $\mathcal{L}(H; 0.5) = \mathbb{P}(H; .5)$

$\mathbb{P}(X|Y)$ probability of X, conditioned on the **event** Y having happened (Y is a subset of the sample space).

$\mathbb{P}(X;\theta)$ probability of X, where to properly define our probability space we need to know the extra piece of information θ. Since θ isn't an event, this is not conditioning.

$\mathcal{L}(X;\theta)$ the likelihood of event X, given that an experiment was run with parameter θ. Likelihoods don't have all the properties we associate with probabilities (e.g. they don't all sum up to 1) and this isn't conditioning on an event (θ is a parameter/rule of how the event could be generated).

MLE

Maximum Likelihood Estimator

The maximum likelihood estimator of the parameter θ is:

$$\hat{\theta} = \text{argmax}_{\theta}\ \mathcal{L}(E;\theta)$$

θ is a variable, $\hat{\theta}$ is a number (or formula given the event).

We'll also use the notation $\hat{\theta}_{\mathbf{MLE}}$ if we want to emphasize how we found this estimator.

The Coin Example

$\mathcal{L}(\text{HTTTHHTHHH}\,;\theta) = \theta^6(1-\theta)^4$

Where is θ maximized?

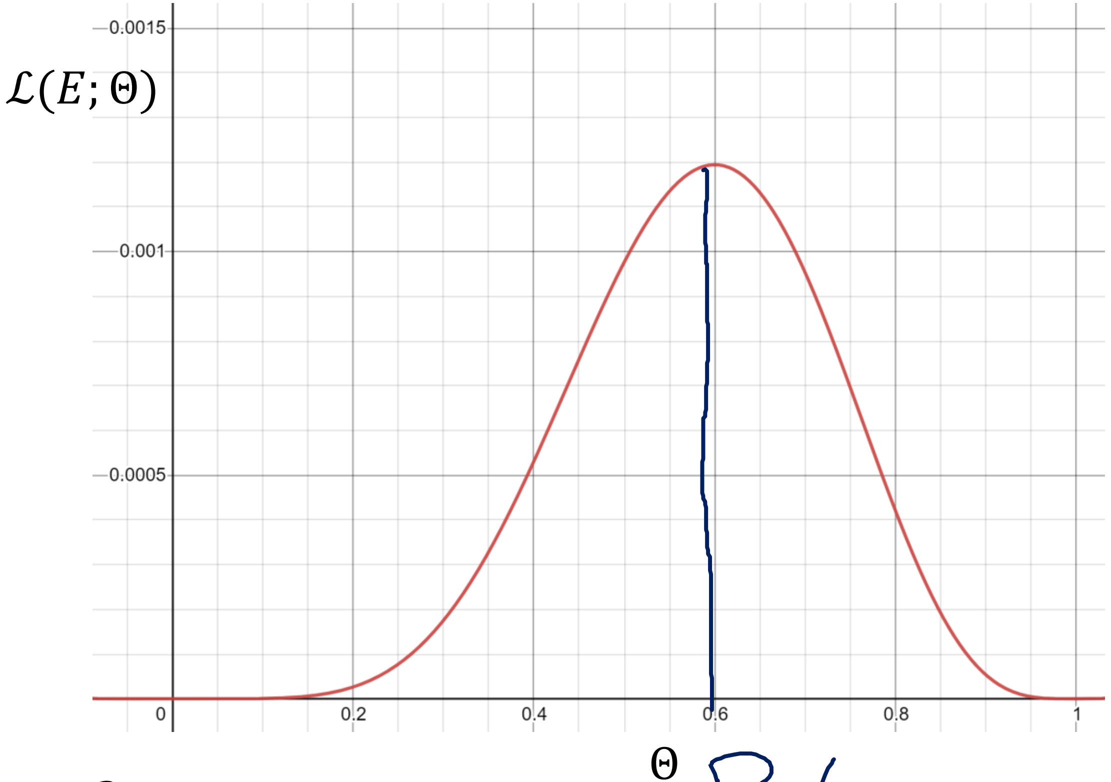
How do we usually find a maximum?

Calculus!!

$$\frac{d}{d\theta}\theta^6(1-\theta)^4 = 6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3$$

Set equal to 0 and solve

$$6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3 = 0 \Rightarrow 6(1-\theta) - 4\theta = 0 \Rightarrow -10\theta = -6 \Rightarrow \theta = \frac{3}{5}$$



3/5

The Coin Example

For this problem, θ must be in the closed interval $[0,1]$. Since $\mathcal{L}()$ is a continuous function, the maximum must occur at and endpoint or where the derivative is 0.

Evaluate $\mathcal{L}(\cdot; 0) = 0, \mathcal{L}(\cdot; 1) = 0$

at $\theta = 0.6$ we get a positive value,

so $\theta = \underline{0.6}$ is the maximizer on the interval $[0,1]$.

Maximizing a Function

CLOSED INTERVALS

Set derivative equal to 0 and solve.

Evaluate likelihood at endpoints and any critical points.

Maximum value must be maximum on that interval.

SECOND DERIVATIVE TEST

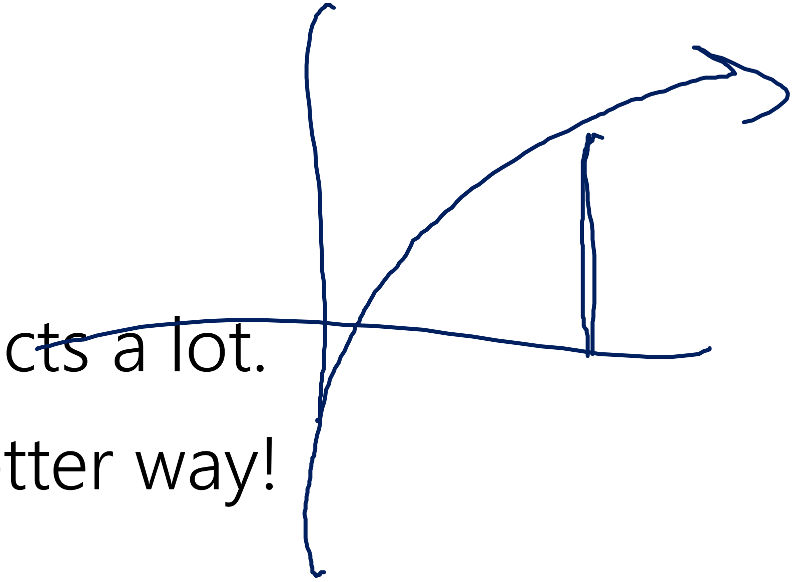
Set derivative equal to 0 and solve.

Take the second derivative. If negative everywhere, then the critical point is the maximizer.

A Math Trick

We're going to be taking the derivative of products a lot.

The product rule is not fun. There has to be a better way!



Take the log!

$$\ln(a \cdot b) = \ln(a) + \ln(b)$$

We don't need the product rule if our expression is a sum!

Can we still take the max? $\ln()$ is an increasing function, so

$$\text{argmax}_\theta \ln(\mathcal{L}(E;\theta)) = \text{argmax}_\theta \mathcal{L}(E;\theta)$$

Coin flips is easier

$\ln(\theta^6) + \ln((1-\theta)^4)$

$$\underline{\mathcal{L}(\text{HTTTHHTHHH};\theta)} = \theta^6(1-\theta)^4$$

$$\underline{\ln(\mathcal{L}(\text{HTTTHHTHHH};\theta)} = \underline{6\ln(\theta)} + \underline{4\ln(1-\theta)}$$

$$\frac{d}{d\theta}\ln\big(\mathcal{L}(\cdot)\big) = \frac{6}{\theta} - \frac{4}{1-\theta}$$

Set to 0 and solve:

$$\frac{6}{\theta} - \frac{4}{1-\theta} = 0 \Rightarrow \frac{6}{\theta} = \frac{4}{1-\theta} \Rightarrow 6 - 6\theta = 4\theta \Rightarrow \theta \boxed{= \frac{3}{5}}$$

$\frac{d^2}{d\theta^2} = \frac{-6}{\theta^2} - \frac{4}{(1-\theta)^2} < 0$ everywhere, so any critical point must be a maximum.

What about continuous random variables?

Can't use probability, since the probability is going to be 0.

Can use the density!

It's supposed to show relative chances, that's all we're trying to find anyway.

$$\underline{\mathcal{L}(\overbrace{x_1, x_2, \dots, x_n}^{E}; \theta)} = \underline{\prod f_X(x_i; \theta)}$$

Continuous Example

Suppose you get values $x_1, x_2, \dots x_n$ from independent draws of a normal random variable $\mathcal{N}(\mu, 1)$ (for μ unkown)

We'll also call these "realizations" of the random variable.

$$\mathcal{L}(x_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right)$$

$$\ln(\mathcal{L}(x_i; \mu)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x_i - \mu)^2$$

$X_1, X_2, \dots X_n$

Finding $\hat{\mu}$

$$\ln(\mathcal{L}) = \sum_{i=1}^{n} \cancel{\ln\left(\frac{1}{\sqrt{2\pi}}\right)} \underset{-}{\cancel{+}} \frac{1}{2}(x_i - \mu)^2$$

$$\frac{d}{d\mu}\ln(\mathcal{L}) = \sum_{i=1}^{n} x_i - \mu$$

Setting $\cancel{\mu} = 0$ and solving:

$$\sum_{i=1}^{n} x_i - \mu = 0 \Rightarrow \sum_{i=1}^{n} x_i = \mu \cdot n \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Check using the second derivative test:

$$\frac{d^2}{d\mu^2}\ln(\mathcal{L}) = -n$$

Second derivative is negative everywhere, so log-likelihood is concave down and average of the x_i is a maximizer.

Summary

Given: an event E (usually n i.i.d. samples from a distribution with unknown parameter θ).

1. Find likelihood $\mathcal{L}(E;\theta)$

Usually $\prod \mathbb{P}\,(x_i;\theta)$ for discrete and $\prod f\,(x_i;\theta)$ for continuous

2. Maximize the likelihood. Usually:

A. Take the log (if it will make the math easier)

B. Set the derivative to 0 and solve

C. Use the second derivative test to confirm you have a maximizer

What about X and Y from last lecture

We polled n people and said, Flip a coin:

If coin is heads OR you have cheated on a partner, tell me “heads”

If coin is tails AND you have never cheated on a partner, tell me “tails”

X was the number of people polled who said “heads”

Y was the number of people polled who cheated on a partner.

We’re trying to find an estimator for Y.

$\mathcal{L}(X = k; Y) =$

What about X and Y from last lecture

X was the number of people polled who said “heads”

Y was the number of people polled who cheated on a spouse.

We’re trying to find an estimator for Y.

The binomial coefficient is maximized when it’s $\binom{m}{m/2}$
Analysis is more complicated because we can’t use calculus (defined only on integers)

$$\mathcal{L}(X = k; Y) = \binom{n-Y}{k-Y}.5^{k-Y}.5^{n-k} = \binom{n-Y}{k-Y}.5^{n-Y}$$

$$k - Y = \frac{n-Y}{2} \Rightarrow k - \frac{n}{2} = \frac{Y}{2} \Rightarrow Y = 2\left(k - \frac{n}{2}\right)$$

So this is also an MLE!
This estimator is only handling the randomness in the coin flips, not the randomness in who was selected. You get the same answer if you back up that far.