New slides went up moments ago
(significant notation changes, you'll want the
new version).

Maximum Likelihood Estimation

CSE 312 Spring 21
Lecture 25

Announcements

You might get some $\overbrace{\text{trivial bounds}}^{\mathbb{P}(\ ) \le 1.2,\ \mathbb{P}(\ ) \ge 0}$ on HW8. Don't replace the numbers with 0 or 1. Just continue with the calculations through the end of the problem.

Robbie just made arithmetic mistakes when designing problems this week ☹
But the pedagogical value of the problems is still there.

Also an (accidental) lesson that concentration inequalities don't always work.

Howard's office hours moved to Wed (still others today)

Asking The Opposite Question

So far:

Give you rules for an experiment.

Give you the event/outcome we're interested in.

You calculate/estimate/bound what the probability is.

Today:

Give you some of the rules of the experiment.

Tell you what happened.

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

Example

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

HTTTHHTHHH

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

A. $1/2$

[✓ B. $3/5$]

C. $2/5$

D. $\frac{55}{100}$

~~Fill out the poll everywhere so Robbie knows how long to explain~~
~~Go to pollev.com/cse312~~

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 most likely.

Maximum Likelihood Estimation

$\mathbb{P}(A|B)$

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 $\theta$where the semicolon means "extra rules" rather than conditioning

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

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

Notation comparison

$\mathbb{P}(X|Y)$ 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:*

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

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

We'll also use the notation $\hat{\theta}_{\mathrm{MLE}}$ 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?

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

How do we usually find a maximum?

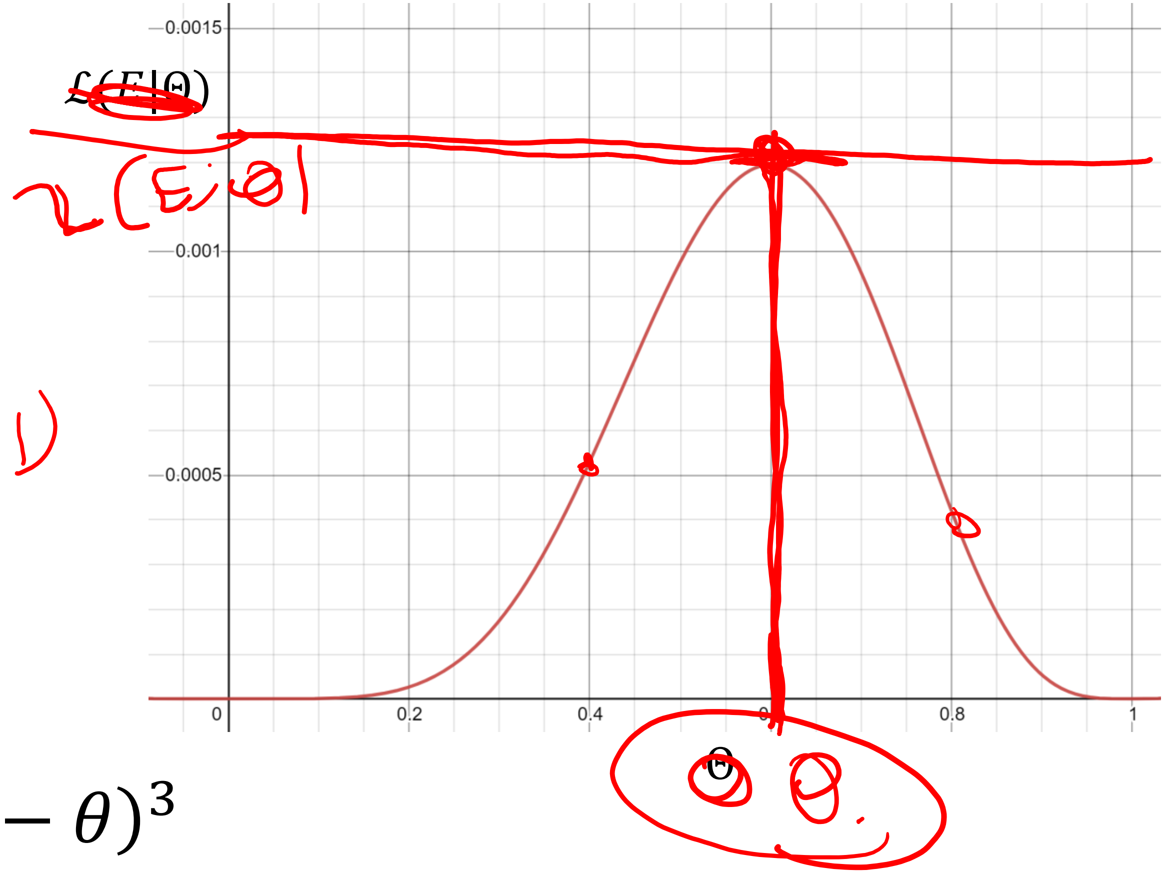
Calculus!! $\frac{\partial}{\partial\theta} fg = f'g + fg'$

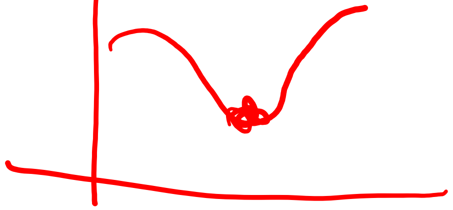
$$\frac{d}{d\theta}\theta^6(1-\theta)^4 = 6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3$$

Set equal to 0 and solve

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

$\theta = \frac{3}{5}$





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.

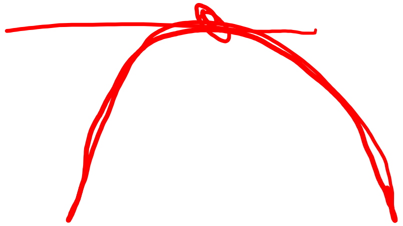
Evaluate $\mathcal{L}(\cdot; 0) = 0, \mathcal{L}(\cdot; 1) = 0$

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

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

$\hat{\theta} = 0.6$

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

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

Coin flips is easier

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

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

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

Set to 0 and solve:

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

$\frac{d^2}{d\theta^2} = \frac{-6}{\theta^2} - \frac{4}{(1-\theta)^2} < 0$ 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 $\mathbf{0}$.

Can use the density!

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

$$\mathcal{L}(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n; \boldsymbol{\theta}) = \prod f_X(\boldsymbol{x}_i; \boldsymbol{\theta})$$

Continuous Example

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

We'll also call these <u>"realizations"</u> of the random variable.

$\ln(e^{?}) = ?$

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

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

"log-likelihood"

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

Finding $\hat{\mu}$

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

$$+ \cancel{\tfrac{1}{2}} \cancel{2}(x_i - \mu)$$

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

Setting ~~μ~~ $\ln(\mathcal{L})$ $= 0$ 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

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.

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

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

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)

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.