7.1, 7.2 Maximum Likelihood Estimation

Anna Karlin
Most slides by Alex Tsun

Agenda

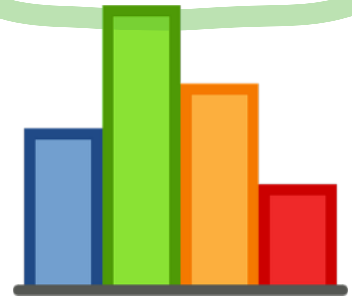
- Probability vs Statistics
- Likelihood
- Maximum Likelihood Estimation (MLE)
- MLE Example (Poisson)
- MLE Example (Normal)

Probability vs Statistics

$Ber(p = 0.5)$

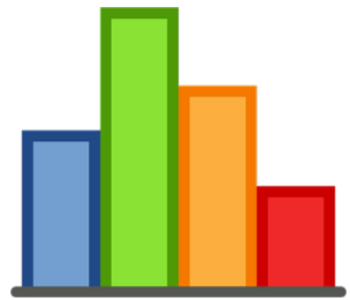
Probability
given model, predict data

$P(THHTHH)$



Probability vs Statistics

$Ber(p = 0.5)$ → **Probability** given model, predict data → $P(THHTHH)$



$Ber(p = ???)$ ← **Statistics** given data, predict model ← $THHTHH$

using parametric model & data.

Bin(n, p), Exp(λ)

θ parameter unknown.

Random Picture

Likelihood (Intuition)

I give you and your classmates each 5 minutes with a coin with unknown probability of heads p. Whoever has the closest estimate will get an A+ in the class. What do you do in your precious 5 minutes, and what do you give as your estimate?

T T H T H T T H

What is your estimate?

a) $\frac{1}{2}$

b) $\frac{5}{8}$

c) $\frac{3}{8}$

d) 1

Likelihood (Intuition)

I give you and your classmates each 5 minutes with a coin with unknown probability of heads p. Whoever has the closest estimate will get an A+ in the class. What do you do in your precious 5 minutes, and what do you give as your estimate?

Flip it as many times as possible, and return # H / (# H + # T)!

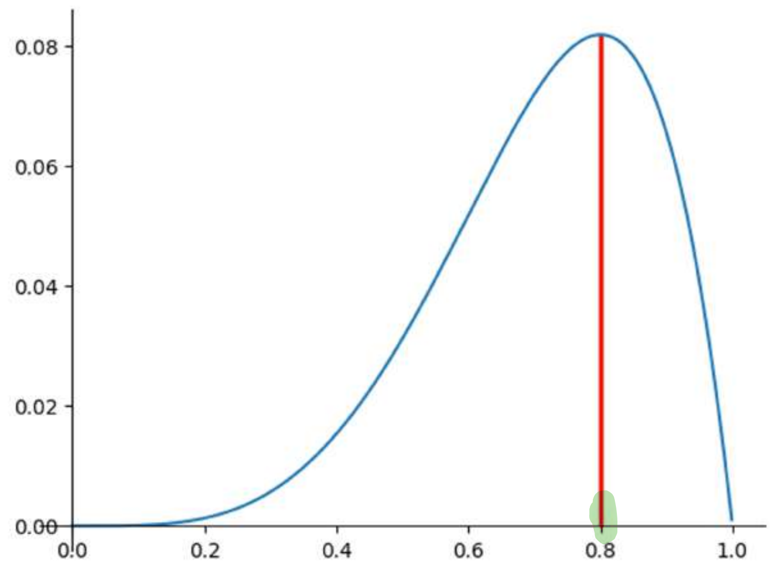
T T H T H T T H

Likelihood (Intuition)

Let's say you saw 4 heads and 1 tail. You tell me $\hat{p} = \frac{4}{5}$. = 0.8

How can you argue, ***objectively***, that this is the "best" estimate?

Is there some objective function it maximizes?



Likelihood (Intuition)

You assume a model (Bernoulli in our case) with unknown parameter θ, and receive iid samples $\boldsymbol{x} = (x_1, \ldots, x_n) \sim Ber(\theta)$. The **likelihood** of the data given a parameter θ is

$$L(\boldsymbol{x}|\theta) = P(\text{seeing data} \mid \theta)$$

$x_1 = H \quad x_2 = H \quad \ldots$

H H H H T

$\theta \cdot \theta \cdot \theta \cdot \theta (1-\theta)$

$= \theta^4 (1-\theta)$

Likelihood (Intuition)

You assume a model (Bernoulli in our case) with unknown parameter θ, and receive iid samples $\boldsymbol{x} = (x_1, \dots, x_n) \sim Ber(\theta)$. The **likelihood** of the data given a parameter θ is

$$\begin{aligned} L(\boldsymbol{x}|\theta) &= P(\text{seeing data} \mid \theta) \\ &= P(x_1, \dots, x_n \mid \theta) \end{aligned}$$

Likelihood (Intuition)

You assume a model (Bernoulli in our case) with unknown parameter θ, and receive iid samples $\boldsymbol{x} = (x_1, \dots, x_n) \sim Ber(\theta)$. The **likelihood** of the data given a parameter θ is

$$L(\boldsymbol{x}|\theta) = P(\text{seeing data} \mid \theta)$$
$$= P(x_1, \dots, x_n \mid \theta)$$
$$= \prod_{i=1}^{n} p_X(x_i; \theta)$$

$p_X(H;\theta) = \theta$ $\qquad$ $p_X(T;\theta) = 1-\theta$

Maximum Likelihood Estimation (Bernoulli)

$Ber(\theta)$

$$L(HHHHT \mid \theta) = \theta^4(1-\theta) = \theta^4 - \theta^5$$

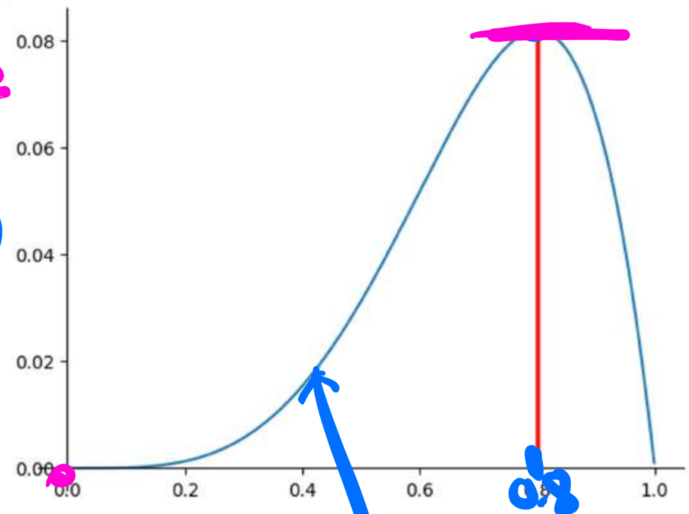
find param θ that maximizes $L(HHHHT \mid \theta)$

$$\frac{d}{d\theta}\left[\theta^4 - \theta^5\right] = 4\theta^3 - 5\theta^4$$

$$\theta^3(4 - 5\theta) = 0$$

$$\Downarrow$$

$$\theta = 0 \quad , \quad \boxed{\theta = \frac{4}{5}}$$

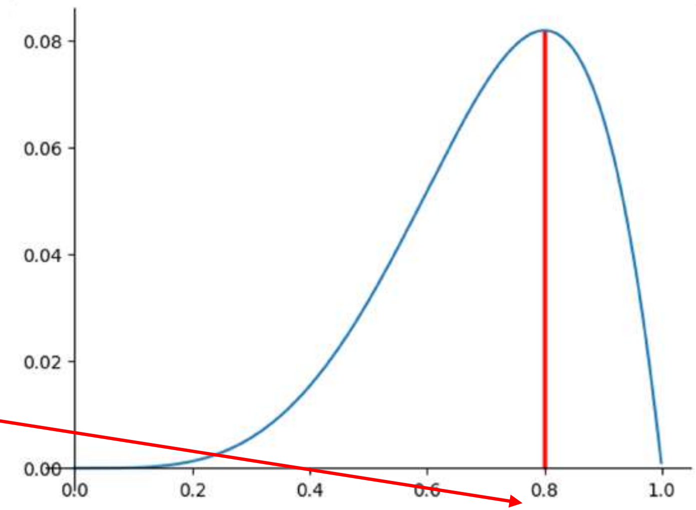


Maximum Likelihood Estimation (Bernoulli)

$$L(HHHHT \mid \theta) = \theta^4(1-\theta) = \theta^4 - \theta^5$$

$$\frac{\partial}{\partial\theta} L(x|\theta) = 4\theta^3 - 5\theta^4 = \theta^3(4-5\theta)$$

$$\hat{\theta}^3(4-5\hat{\theta}) = 0 \rightarrow \hat{\theta} = \frac{4}{5} \text{ or } 0$$



Likelihood vs Probability. Ber(θ)

$P(x; \theta)$ = prob of event x given model Ber(θ)

as fn of x (for fixed θ)

$$\sum_x \Pr(x;\theta) = 1$$

$L(x;\theta)$ as fn of θ (fixed x)

$$\sum_\theta L(x;\theta) \text{ anything.}$$

$$L(HHHT \mid 0.6) > L(HHHT \mid \theta = 0.5)$$

Finds parameters θ that maximize likelihood

⇓

$X_1 X_2 X_3 X_4$ iid. Ber($\frac{1}{3}$)

$L(\underline{HHHT} \mid \theta)$

fixed — variable.

$Ber(\theta)$

Prob(seeing data for param θ)

var.

$P(\underbrace{HHHT}_{x_1 x_2 x_3 x_4} ; p = \frac{1}{3})$

fixed.

$P(\underline{x_1 x_2 x_3 x_4})$

$L(\underline{x_1, \dots, x_n} \mid \theta) \Leftarrow$ function of θ

[assuming the param of distn is θ]

[not cond prob]

Assuming data i.i.d Poisson(θ)

$x_1, \dots, x_n$ samples

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \dots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

unknown.

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

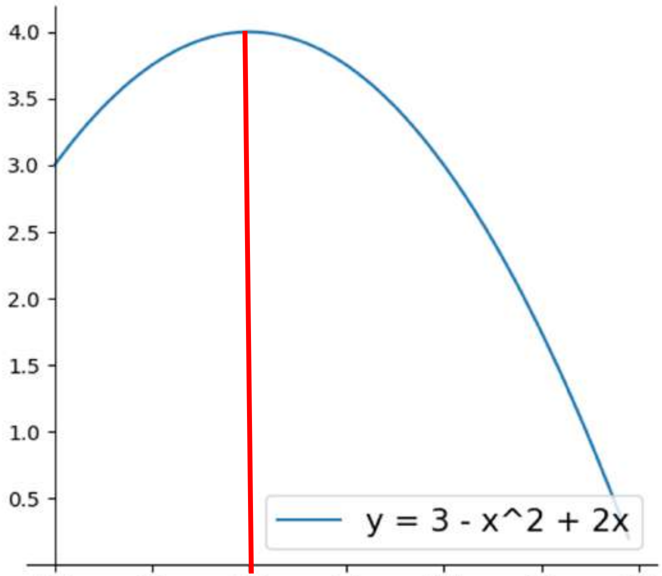
$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

Prob of seeing this data as fn of parameter θ

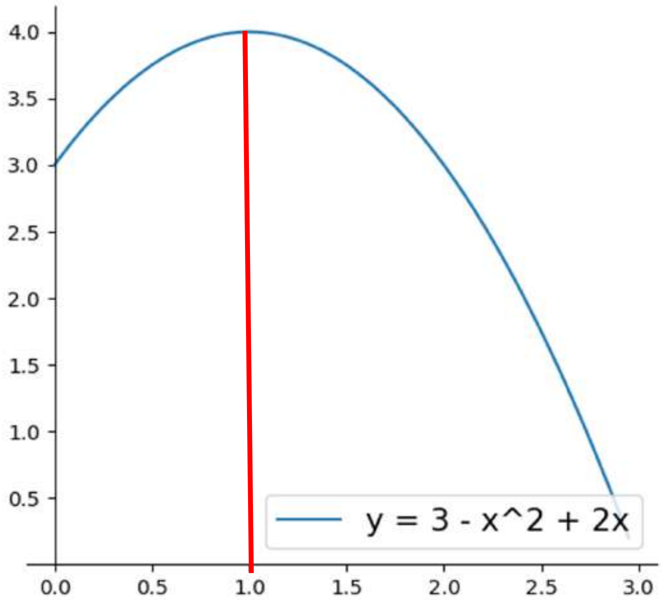
Find θ that maximizes L(x|θ)

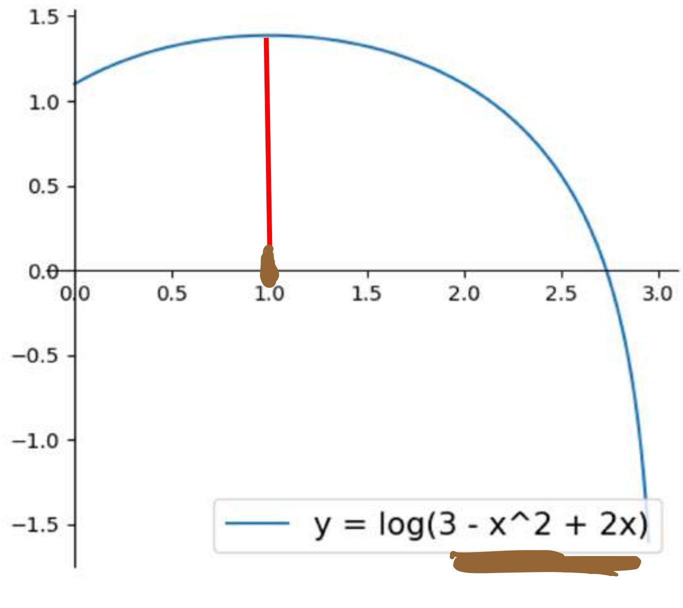
$$e^{-\theta}\frac{\theta^3}{3!}\; e^{-\theta}\frac{\theta^5}{5!}\; e^{-\theta}\frac{\theta^4}{4!}$$

Optimizing Function vs Log(Function)



Optimizing Function vs Log(Function)

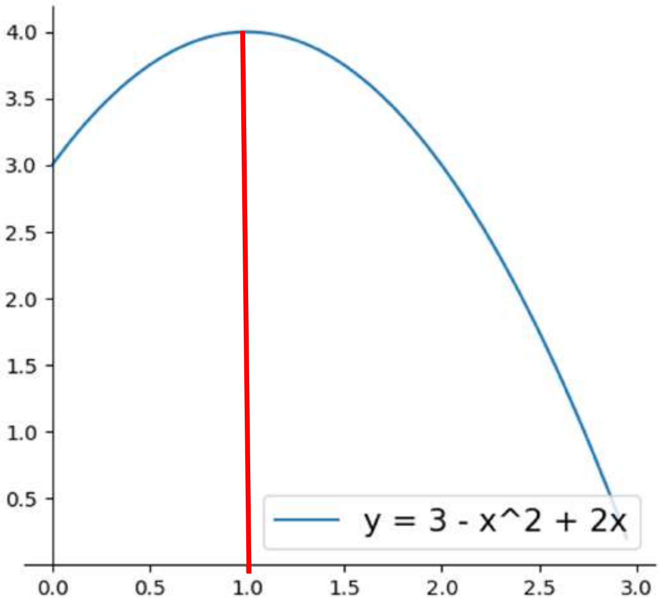


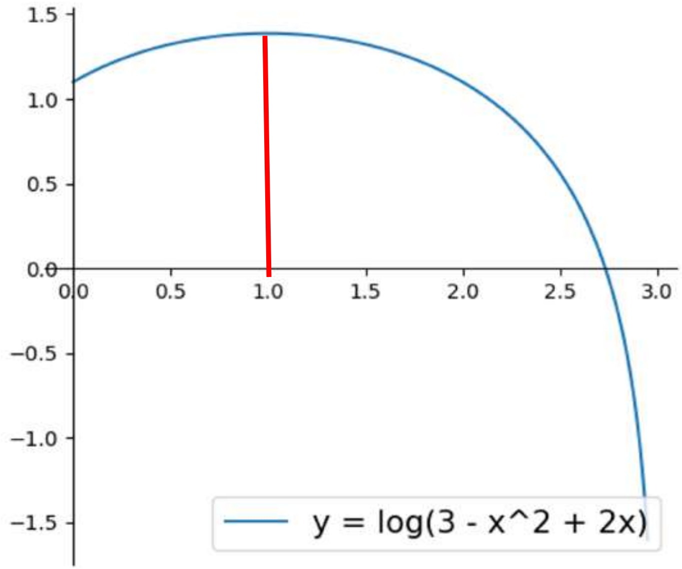


$f(x_1) > f(x_2)$ $\quad$ $\ln f(x_1) > \ln f(x_2)$

x that maximizes $f(x)$ $\equiv$ x that maximizes $\ln f(x)$

Optimizing Function vs Log(Function)





Since $g(x) = \log x$ is **strictly increasing**, it preserves order, and in particular, the argmax.

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

$\log(ab) = \log(a) + \log(b)$

$\log(a/b) = \log(a) - \log(b)$

$\log(a^b) = b \log a$

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

$\log \prod_{i=1}^{n} a_i = \sum_{i=1}^{n} \log a_i$

$LL(x|\theta) = \log L(x|\theta) = \sum_{i=1}^{n} \log\left(e^{-\theta} \frac{\theta^{x_i}}{x_i!}\right)$

log likelihood.

$\ln e^{-\theta} + x_i \ln \theta - \ln(x_i!)$

$-\theta \ln e$, $\ln e = 1$

$$LL(x|\theta) = \sum_{i=1}^{n}\left[-\theta + x_i \ln\theta - \ln(x_i!)\right]$$

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

$\log(ab) = \log(a) + \log(b)$

$\log(a/b) = \log(a) - \log(b)$

$\log(a^b) = b \log a$

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!}$$

$$\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n}[-\theta + x_i \ln\theta - \ln(x_i!)]$$

$$\frac{d}{d\theta} LL(x|\theta) = -n + \sum_{i=1}^{n} x_i \frac{d}{d\theta}\ln\theta$$

$$= -n + \sum_{i=1}^{n}\frac{x_i}{\theta}$$

$$0 = -n + \sum_{i=1}^{n}\frac{x_i}{\theta}$$

$$\frac{\sum x_i}{\theta} = n \equiv \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \dots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

$$\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n} [-\theta + x_i \ln \theta - \ln(x_i!)]$$

$$\frac{\partial}{\partial \theta} \ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n} \left[-1 + \frac{x_i}{\theta}\right]$$

$$\sum_{i=1}^{n} \left[-1 + \frac{x_i}{\hat{\theta}}\right] = 0 \rightarrow -n + \frac{1}{\hat{\theta}} \sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \dots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!}$$

$$\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n} [-\theta + x_i \ln\theta - \ln(x_i!)]$$

$$\frac{\partial}{\partial\theta}\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n}\left[-1 + \frac{x_i}{\theta}\right]$$

$$\sum_{i=1}^{n}\left[-1 + \frac{x_i}{\hat{\theta}}\right] = 0 \rightarrow -n + \frac{1}{\hat{\theta}}\sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial^2}{\partial\theta^2}\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n}\left[-\frac{x_i}{\theta^2}\right] < 0 \rightarrow \text{concave down everywhere}$$

General recipe

Given indep samples $x_1, \ldots, x_n$ from parametric model
→ Ber(θ)
→ Poisson(θ)
Exp(θ)

Find $\hat{\theta}$ that maximizes $L(x_1, \ldots x_n | \theta)$

discrete $= \prod_{i=1}^{n} p_X(x_i; \theta)$ ⟸

cont. $\triangleq \prod_{i=1}^{n} f_X(x_i; \theta)$

find $\hat{\theta}$ that maximizes log likelihood $LL(x_1, \ldots x_n | \theta)$

discrete $= \sum_{i=1}^{n} \ln[p_X(x_i; \theta)]$ ⟸

cont $= \sum_{i=1}^{n} \ln[f_X(x_i; \theta)]$

Find $\hat{\theta}$ to max $LL(\vec{x} \mid \theta)$

$x_1, \ldots, x_n$

distn has 1 parameter

compute $\frac{d\, LL(\theta)}{d\theta}$

set $\frac{d\, LL(\theta)}{d\theta} = 0$

solve for $\hat{\theta}$

[verify soln is max (2^{nd} deriv < 0)] — don't need to do this

multiple params $\theta_1, \ldots, \theta_k$

find $\hat{\theta}_1, \ldots, \hat{\theta}_k$ that are soln to this system

$$\begin{cases} \frac{\partial LL}{\partial \theta_1} = 0 \\ \frac{\partial LL}{\partial \theta_2} = 0 \\ \vdots \\ \frac{\partial LL}{\partial \theta_k} = 0 \end{cases}$$

[Check max
Check Hessian −ve definite]

Likelihood

Realization/Sample: A realization/sample x of a random variable X is the value that is actually observed.

Likelihood

Realization/Sample: A realization/sample x of a random variable X is the value that is actually observed.

Likelihood: Let $\boldsymbol{x} = (x_1, \dots, x_n)$ be iid realizations from probability mass function $p_X(t\,;\theta)$ (if X discrete), or from density $f_X(t\,;\theta)$ (if X continuous), where θ is a parameter (or vector of parameters). We define the likelihood of x as the probability of seeing the data.

Likelihood

Realization/Sample: A realization/sample x of a random variable X is the value that is actually observed.

Likelihood: Let $\boldsymbol{x} = (x_1, \dots, x_n)$ be iid realizations from probability mass function $p_X(t\,;\theta)$ (if X discrete), or from density $f_X(t\,;\theta)$ (if X continuous), where θ is a parameter (or vector of parameters). We define the likelihood of x as the probability of seeing the data.

If X is discrete,

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta)$$

X cont.

$$L(\vec{x} \mid \theta) = \prod_{i=1}^{n} f_X(x_i\,;\,\theta)$$

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE): Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be iid realizations from probability mass function $p_X(t\,;\theta)$ (if X discrete), or from density $f_X(t\,;\theta)$ (if X continuous), where θ is a parameter (or vector of parameters). We define the maximum likelihood estimator $\hat{\theta}_{MLE}$ of θ to be the parameter which maximizes the likelihood (or equivalently, the log-likelihood).

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \ln L(\boldsymbol{x} \mid \theta)$$

Random Picture