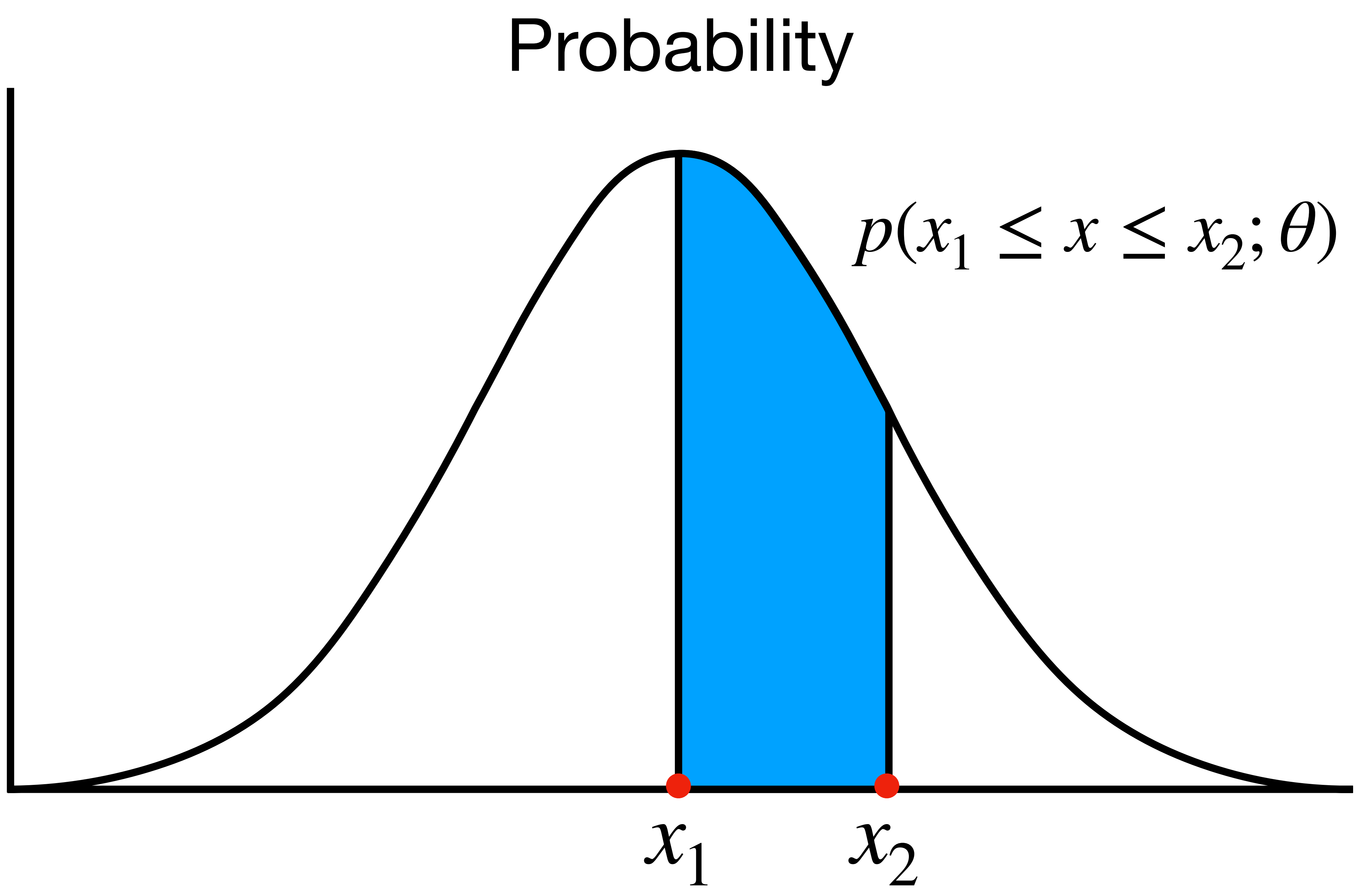
ICT이노베이션스퀘어 AI복합교육 고급 언어과정

자연어처리를 위한 Negative Log Likelihood

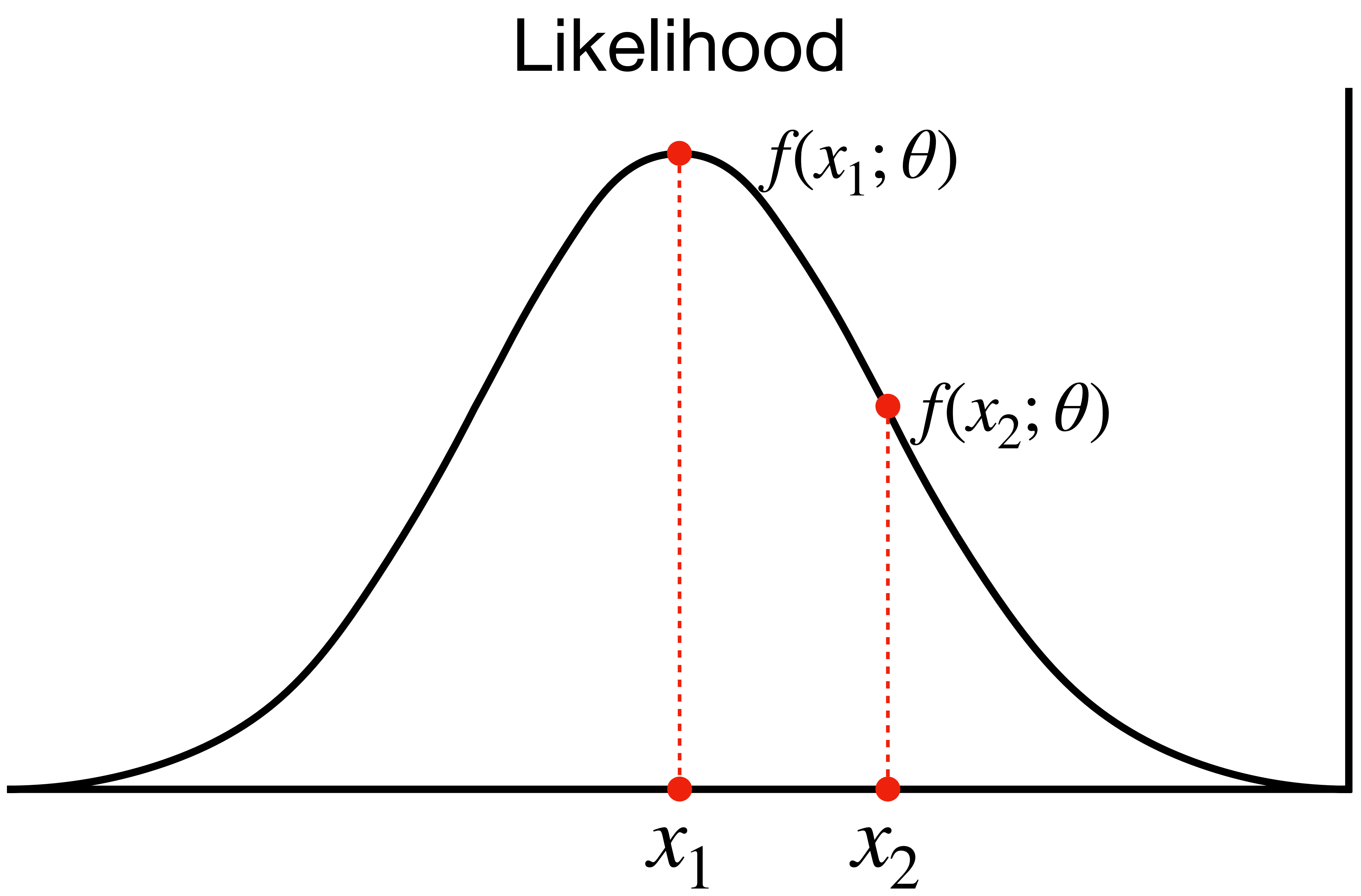
현청천

2021.04.19

Likelihood (연속확률분포)



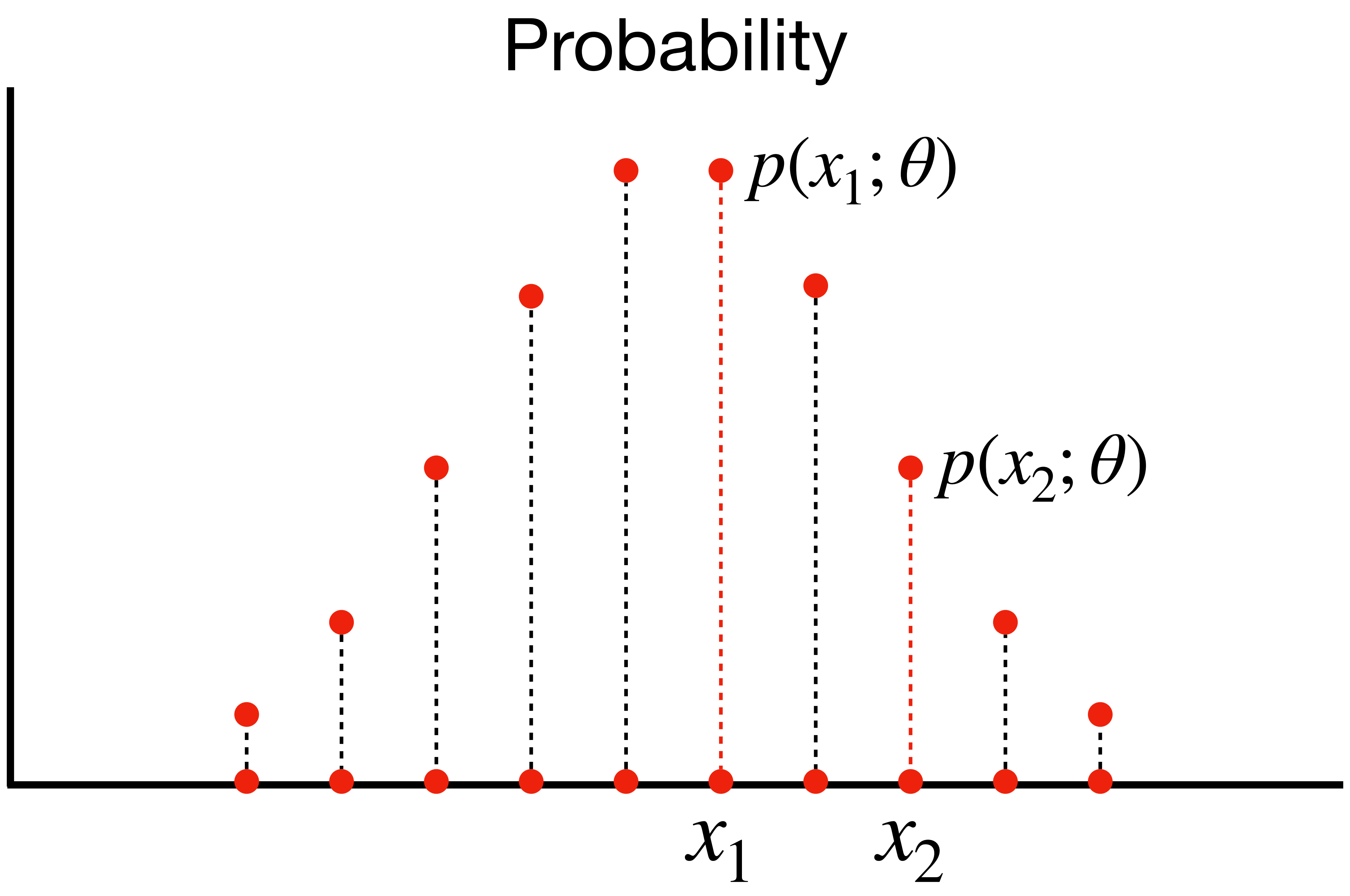
값보단 범위 $x_1 < x < x_2$

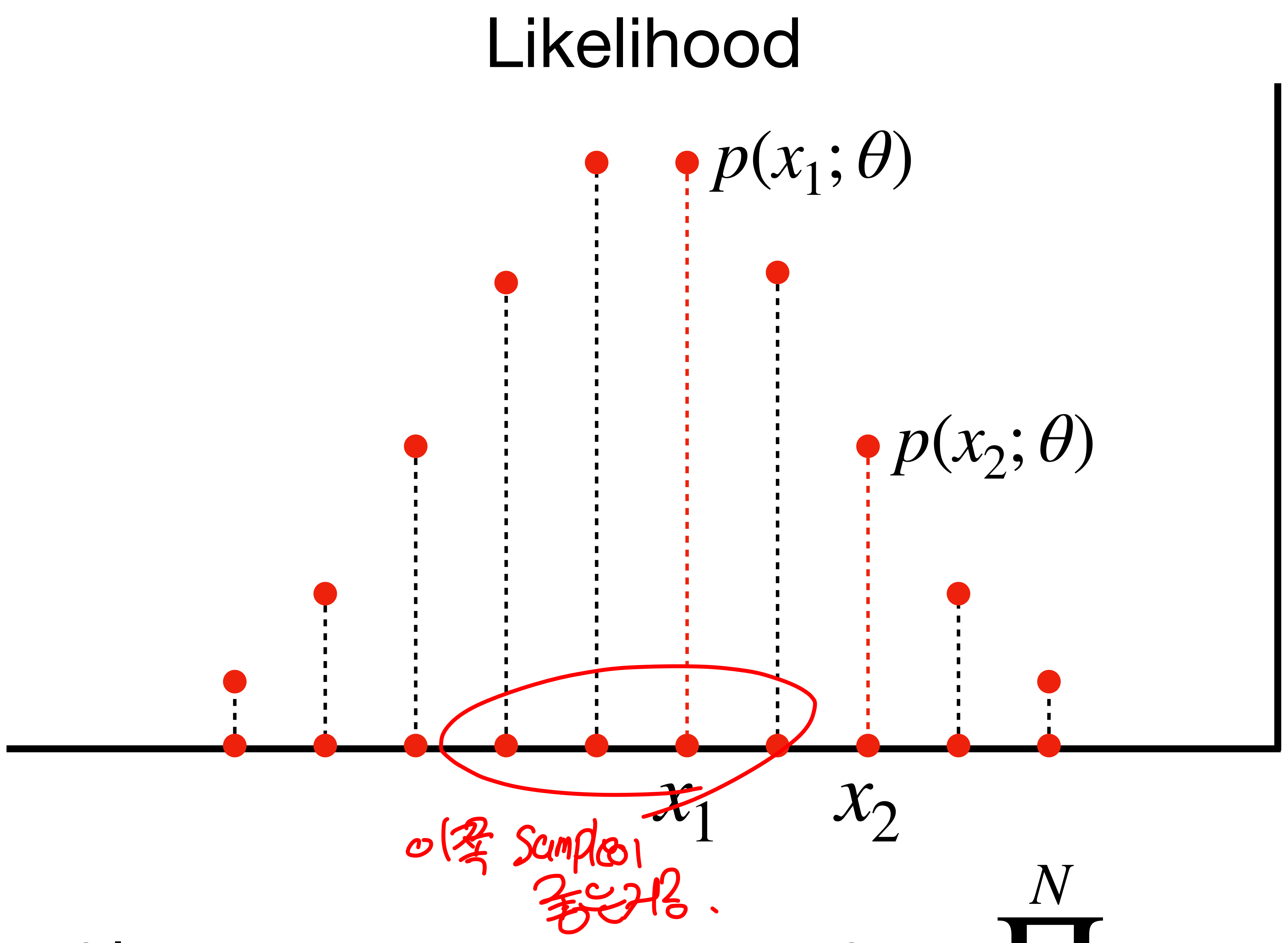


$$\mathscr{L}(\theta \mid x) = f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{N} f(x_i; \theta)$$

가능도 (특정 사건들이 일어날 가능성)

Likelihood (이산확률분포)



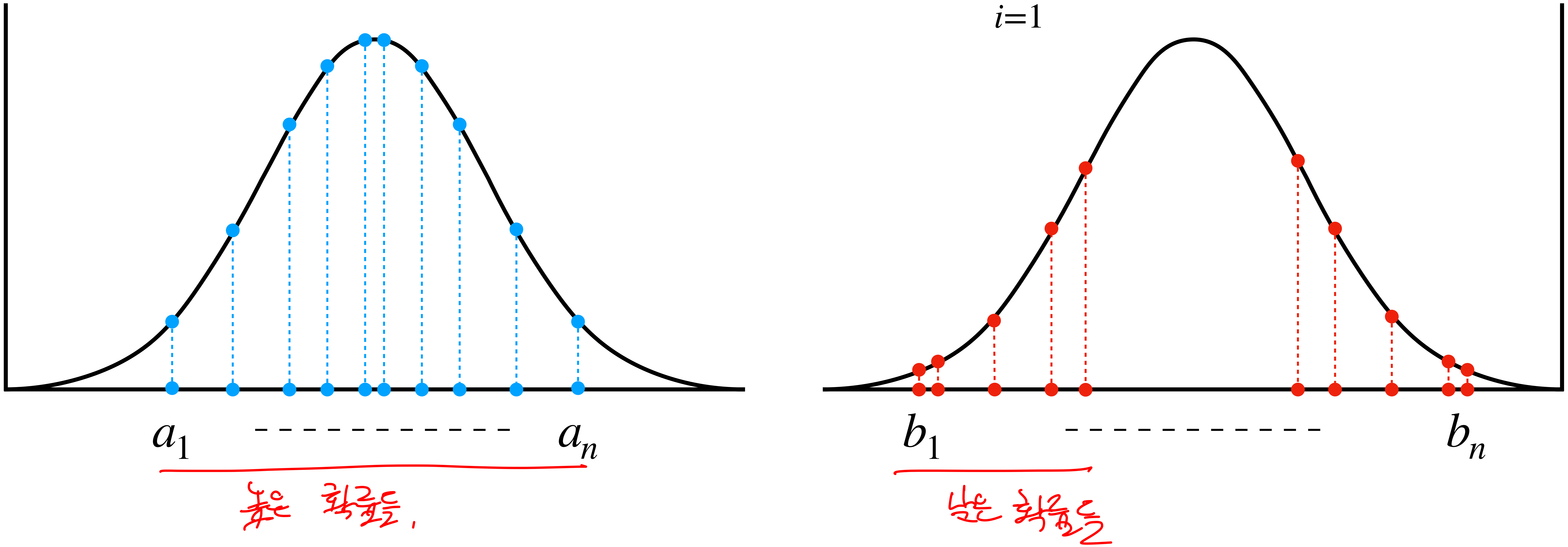


$$\mathcal{L}(\theta \mid x) = p(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

가능도 (특정 사건들이 일어날 가능성)

Likelihood

$$\mathscr{L}(\theta \mid x) = f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{N} f(x_i; \theta)$$



$$\mathscr{L}(\theta \mid a) > \mathscr{L}(\theta \mid b)$$

Maximum Likelihood Estimation

분포를 모르고 사건으로부터

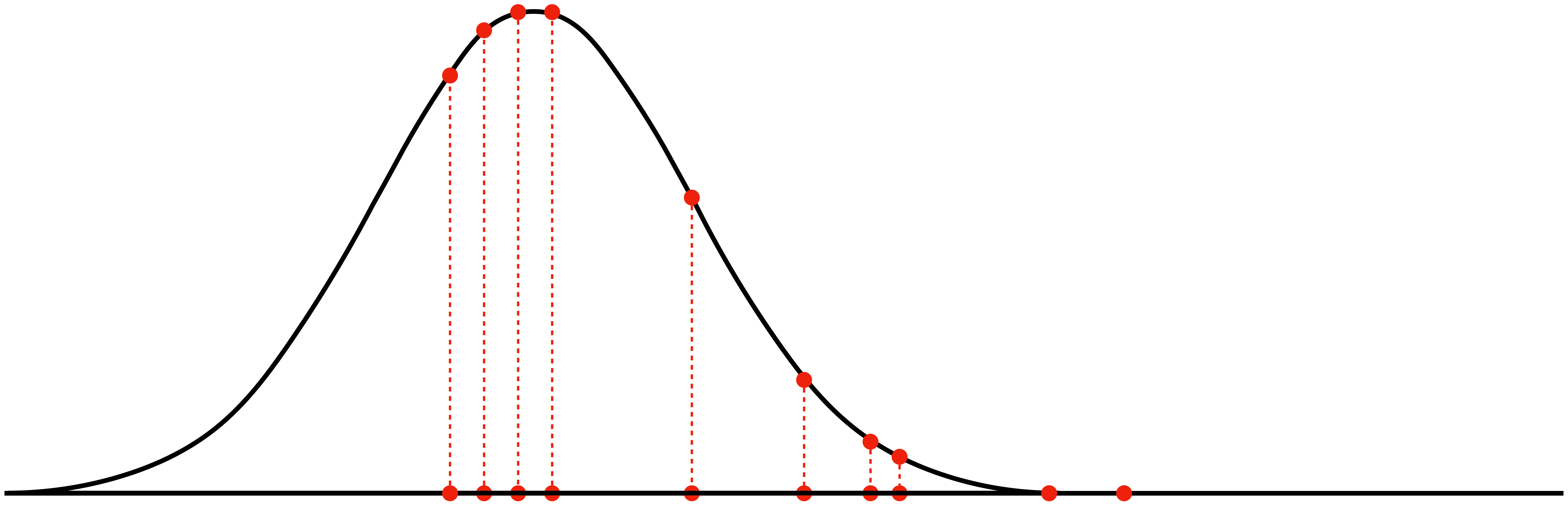
키
164.95
165.35
165.76
166.16
167.78
168.99
169.80
170.20
171.82
172.63

사건으로부터 확률분포를 예측

Maximum Likelihood Estimation

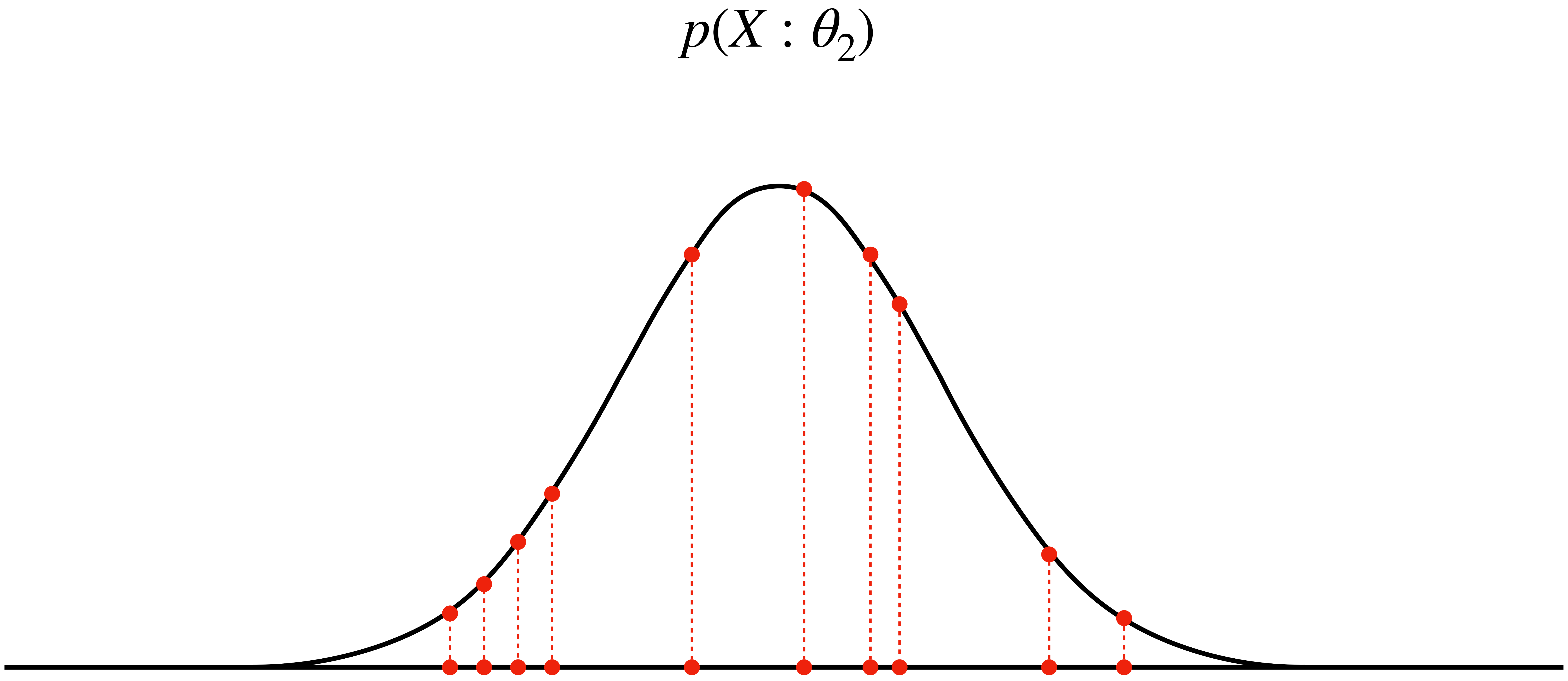
$p(X; \theta_1)$

키
164.95
165.35
165.76
166.16
167.78
168.99
169.80
170.20
171.82
172.63



Maximum Likelihood Estimation

키
164.95
165.35
165.76
166.16
167.78
168.99
169.80
170.20
171.82
172.63

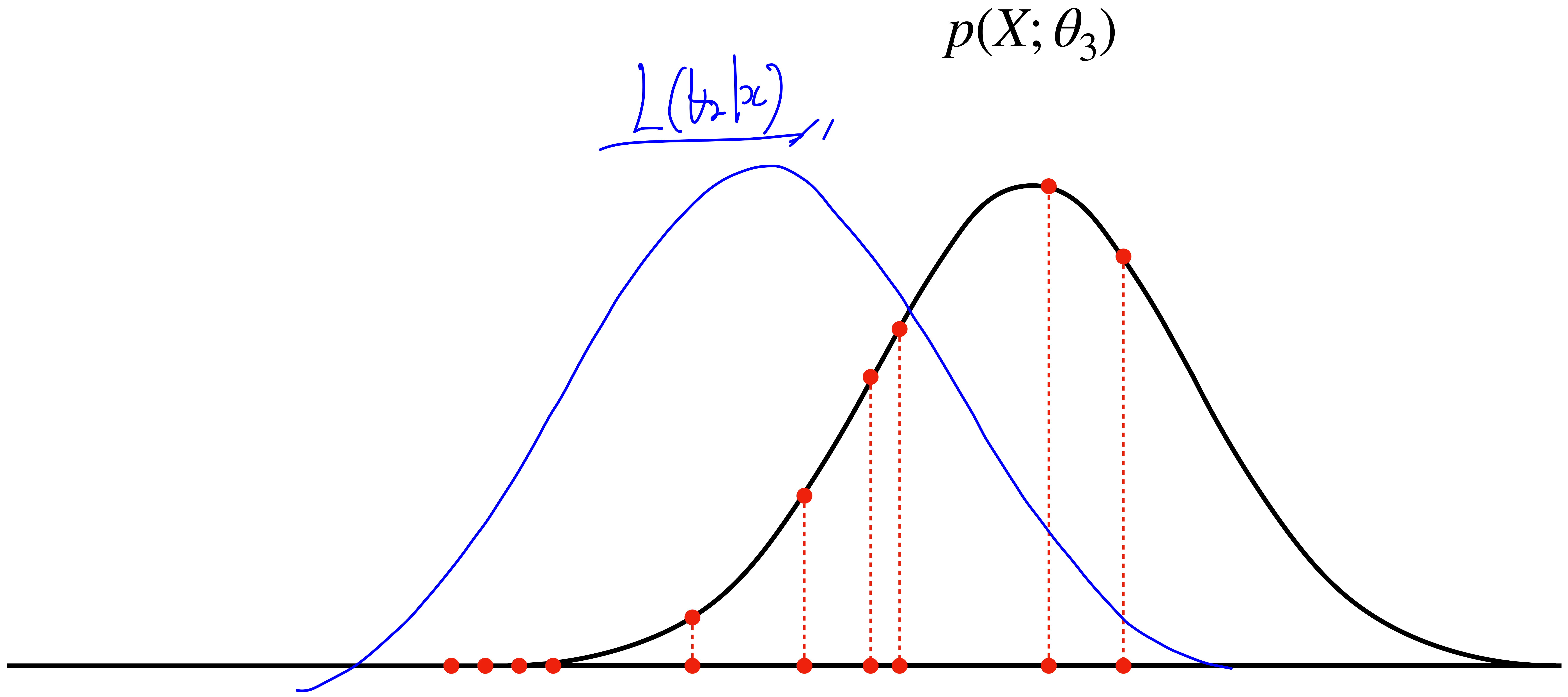


$$\mathscr{L}(\theta_1 | x) < \mathscr{L}(\theta_2 | x)$$

θ_1보다 θ_2가 likelihood가 높다.

Maximum Likelihood Estimation

키
164.95
165.35
165.76
166.16
167.78
168.99
169.80
170.20
171.82
172.63



sample한 게 원래의 확률분포를 잘 반영할거라고 가정!

$$\mathcal{L}(\theta_1 \,|\, x) < \mathcal{L}(\theta_2 \,|\, x) > \mathcal{L}(\theta_3 \,|\, x)$$

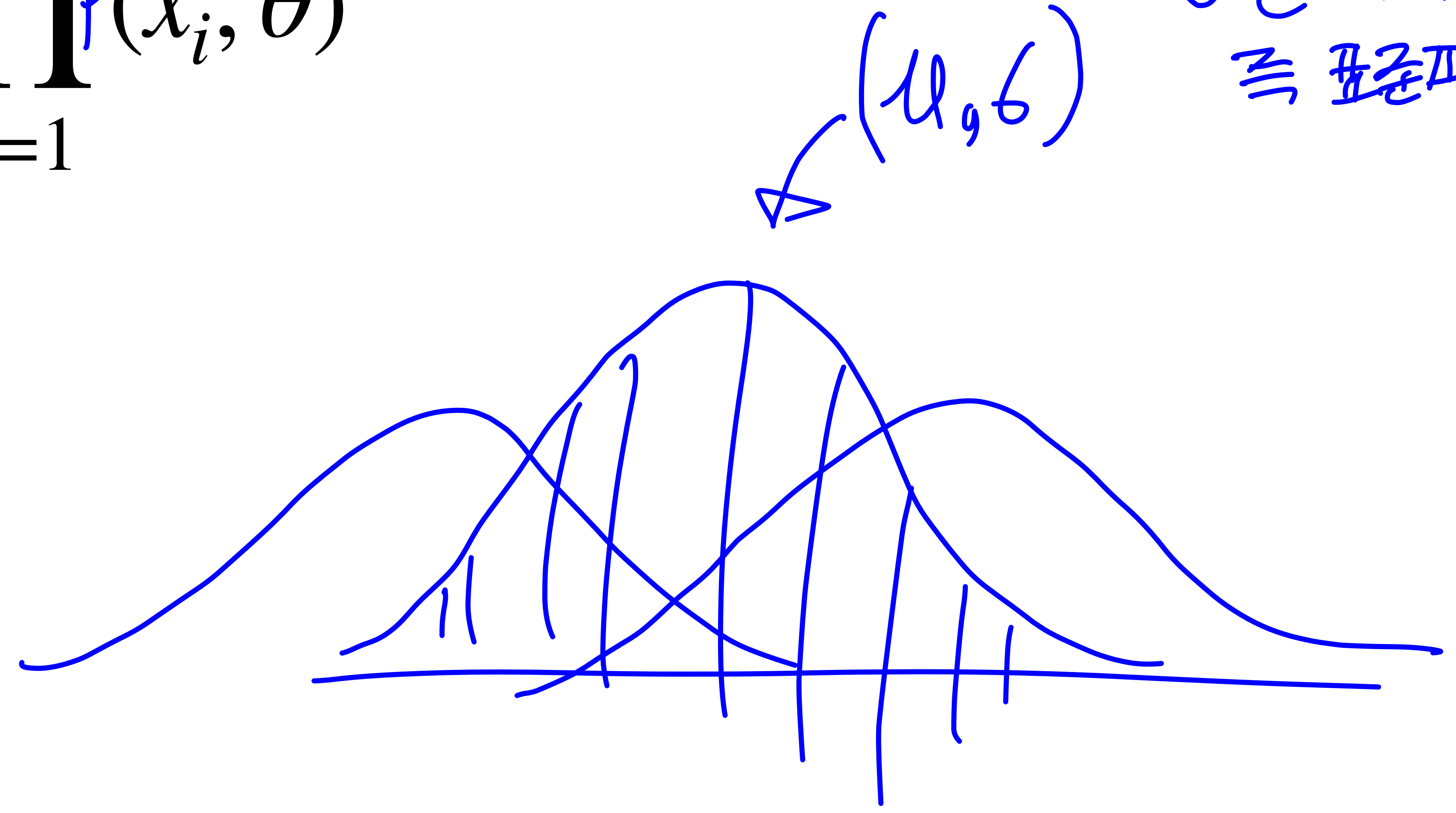
Maximum Likelihood Estimation

Likelihood

$$\mathscr{L}(\theta\,|\,x) = \prod_{i=1}^{N} p(x_i;\theta)$$

이 θ를
최대화 시킬때

θ는 이때 σ.
즉 표준편차를 말함

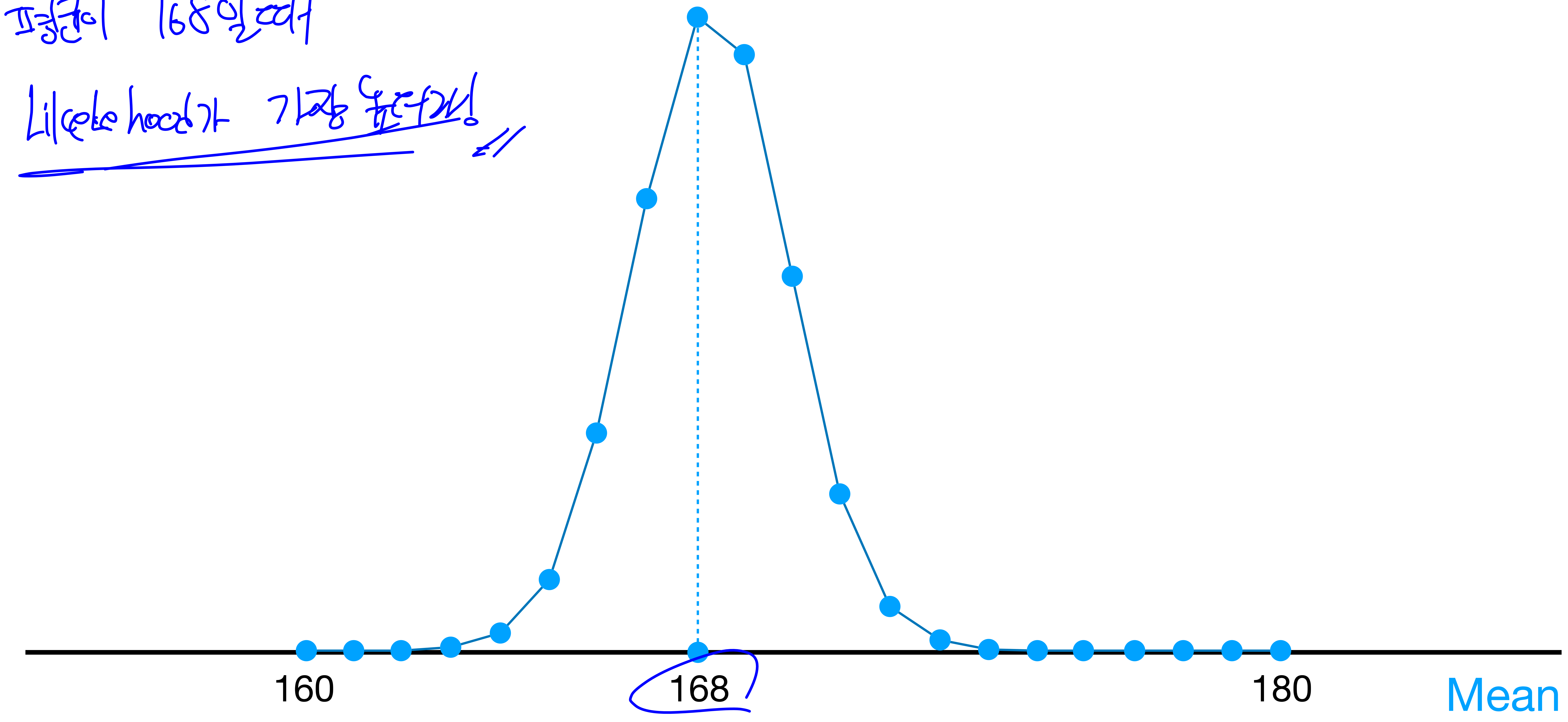


$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\,\mathscr{L}(\theta\,|\,x)$$

Maximum Likelihood Estimation

평균이 168일때
Likelihood가 가장 높더라!

키
164.95
165.35
165.76
166.16
167.78
168.99
169.80
170.20
171.82
172.63



$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \mathscr{L}(\theta \mid x)$$

Log Likelihood

Likelihood

$$\mathscr{L}(\theta|x) = \prod_{i=1}^{N} f(x_i;\theta)$$

f는 기본적으로 0~1 사이므로
곱하게 되면 매우 작아지므로

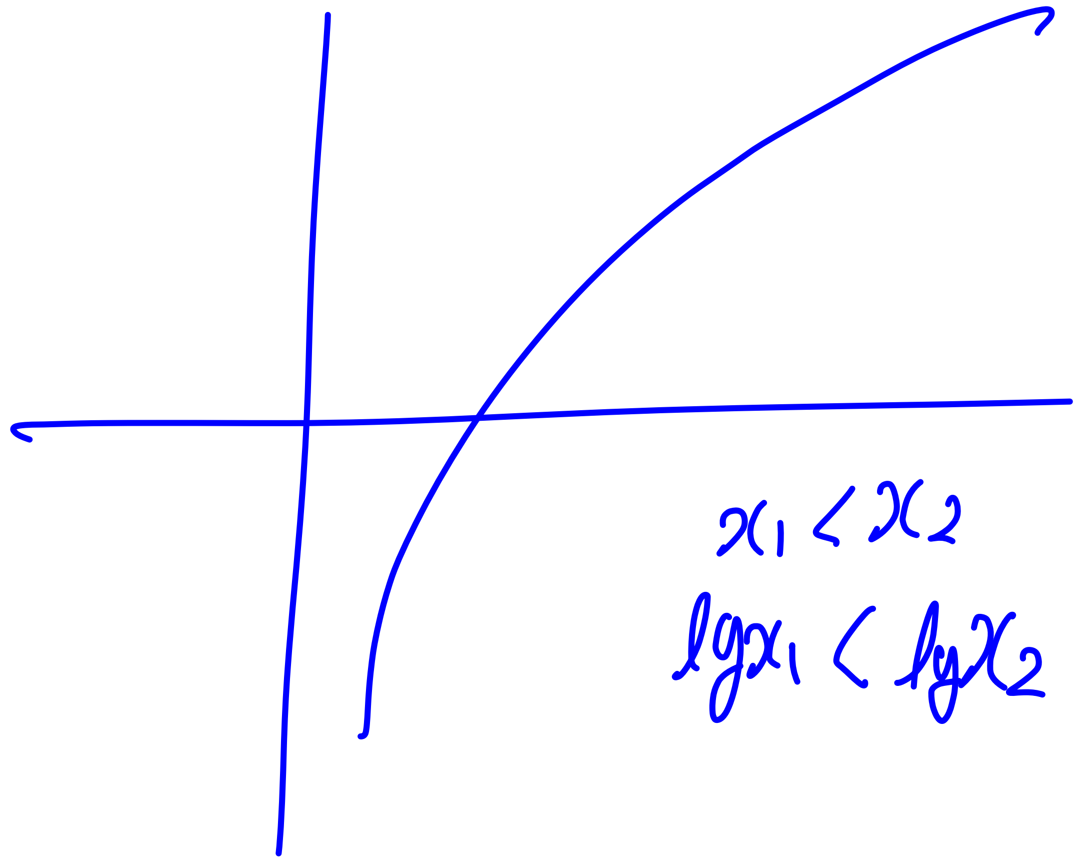
Log Likelihood

$$\log \mathscr{L}(\theta|x) = \sum_{i=1}^{N} \log f(x_i;\theta)$$

함수의(f)의 log를 취해서
더해주게 되면 위랑 같은거나
마찬가지가 되지

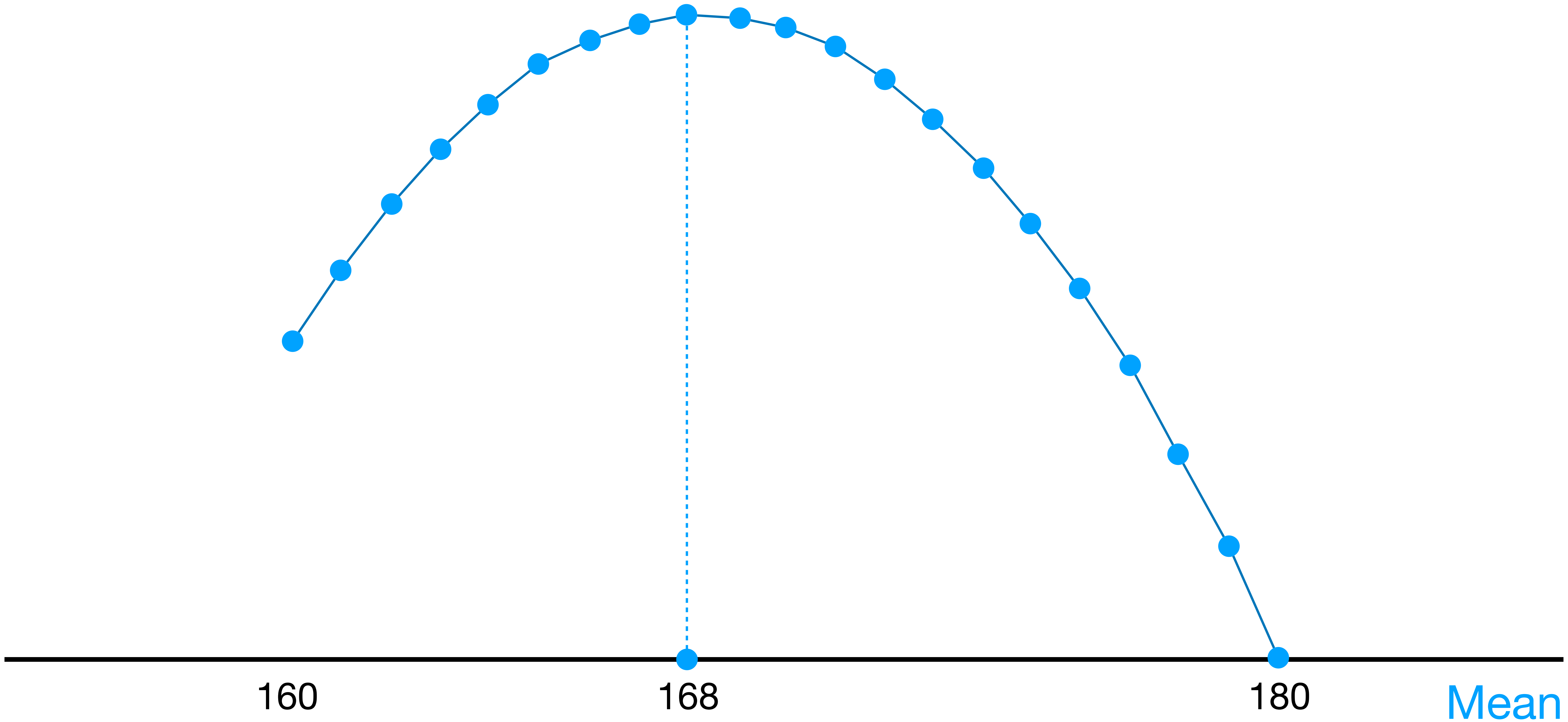
θ가 최대가 되는 값을 찾아야 되는데
다(동일)한 값이지

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log \mathscr{L}(\theta|x)$$



Log Likelihood

키
164.95
165.35
165.76
166.16
167.78
168.99
169.80
170.20
171.82
172.63



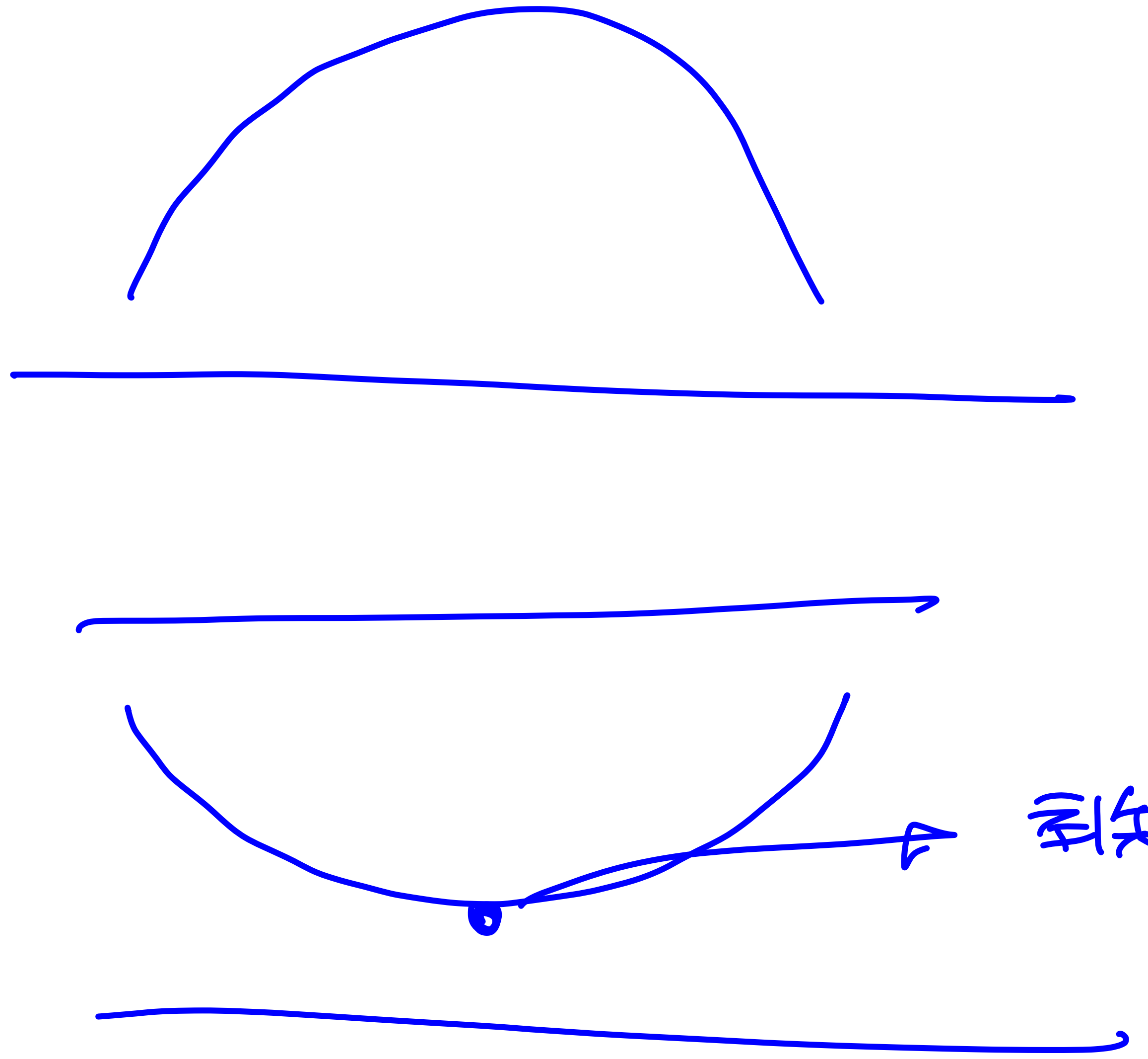
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log \mathcal{L}(\theta \mid x)$$

Negative Log Likelihood

Likelihood $$\mathscr{L}(\theta|x) = \prod_{i=1}^{N} f(x_i;\theta)$$

Log Likelihood $$\log \mathscr{L}(\theta|x) = \sum_{i=1}^{N} \log f(x_i;\theta)$$

Negative Log Likelihood $$-\log \mathscr{L}(\theta|x) = -\sum_{i=1}^{N} \log f(x_i;\theta)$$



최솟값이 됨.

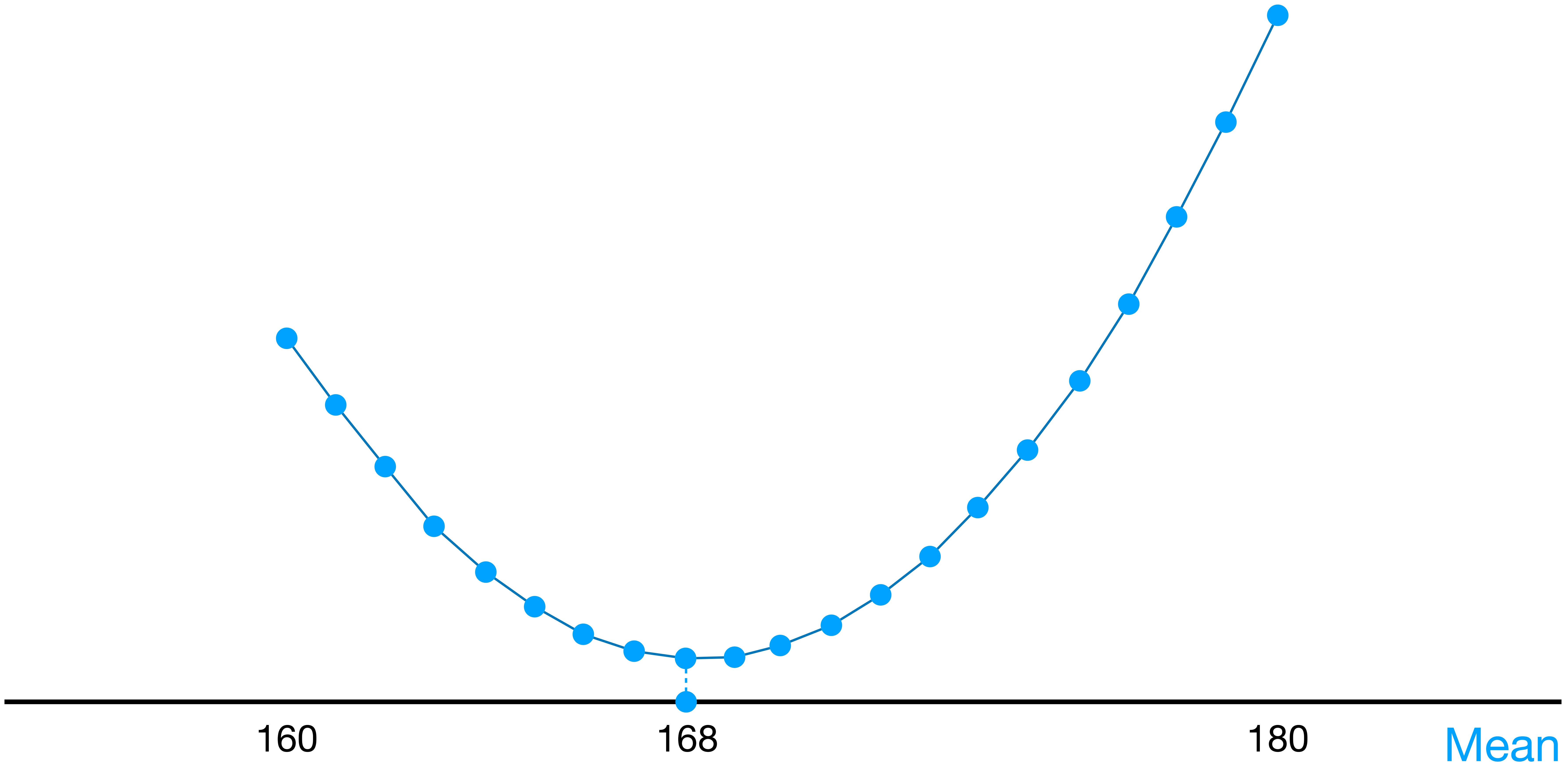
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} -\log \mathscr{L}(\theta|x)$$

Loss를 통해 미니멈을 찾지.

즉, Negative Log Likelihood를 Loss function을 통해 업데이트할 수 있지.

Negative Log Likelihood

키
164.95
165.35
165.76
166.16
167.78
168.99
169.80
170.20
171.82
172.63



$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} - \log \mathscr{L}(\theta \mid x)$$

Negative Log Likelihood (Example)

- Simple Example
  - 길가는 사람 10명의 핸드폰 운영체제를 조사했다.
    - Android 7명
    - iOS 3명

핸드폰 운영체제 점유율 추정

7:3이 지배

vs

Negative Log Likelihood (Example)

이 방식으로 접근해보자.

Android probability p

IOS probability $1 - p$

Sampling probability ${}_{10}C_7 p^7 (1-p)^3$ → 확률분포

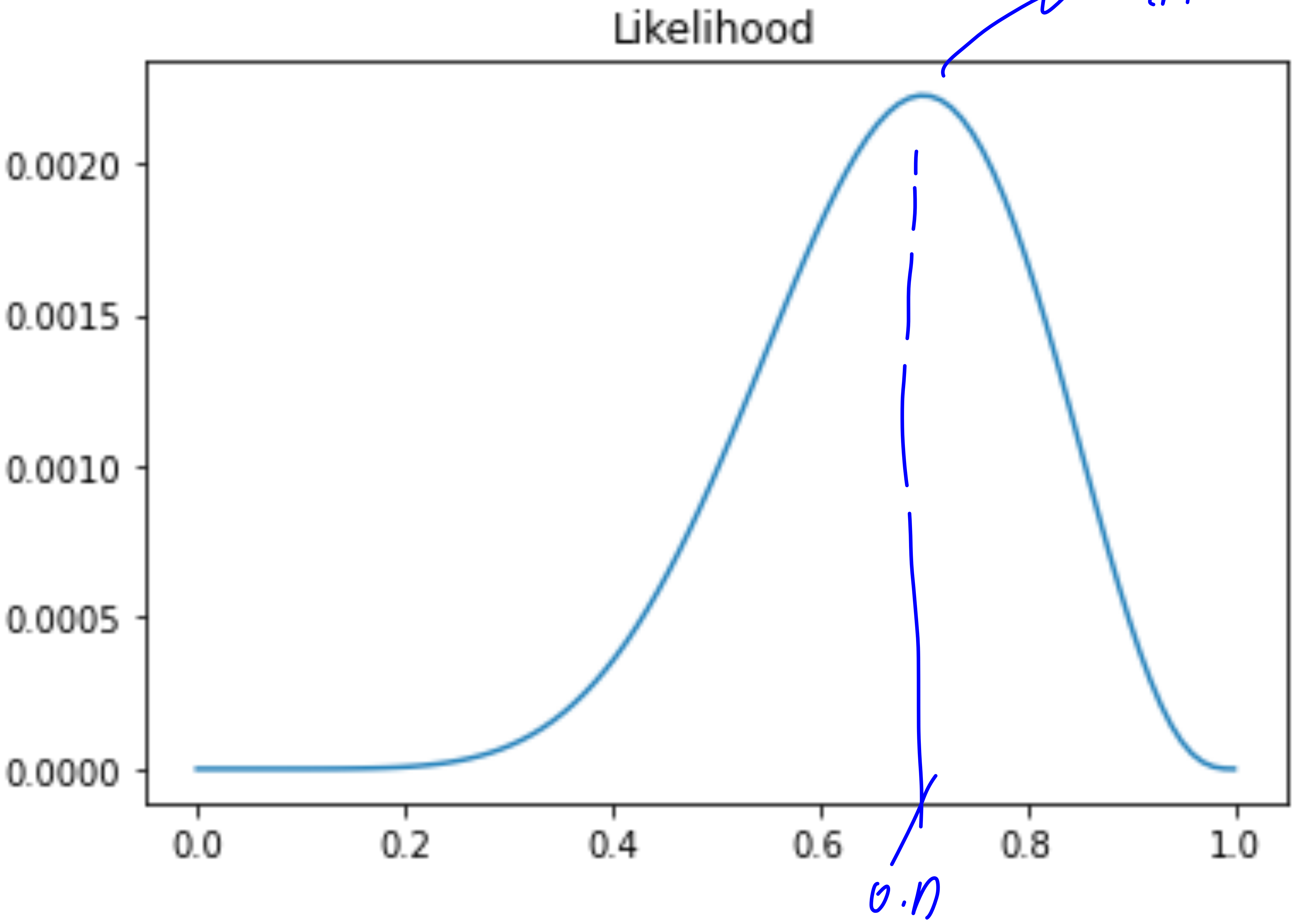
10개중 7개 뽑는 확률

Negative Log Likelihood (Example)

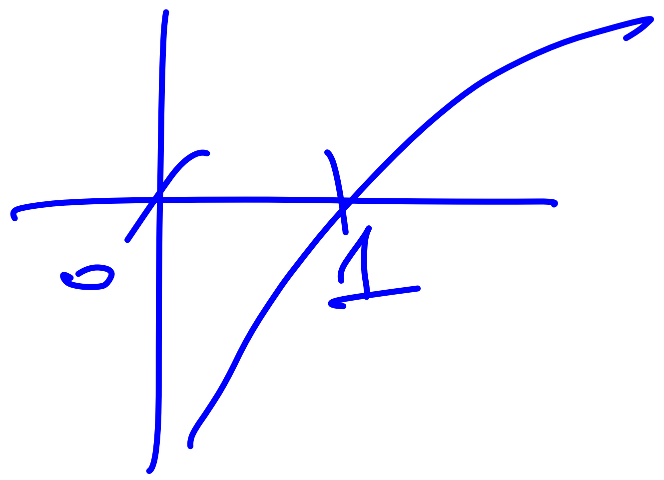
log 취해볼게

Likelihood $\mathscr{L}(p) = p^7(1-p)^3$

0.7에서 최댓값이 되지.



Negative Log Likelihood (Example)



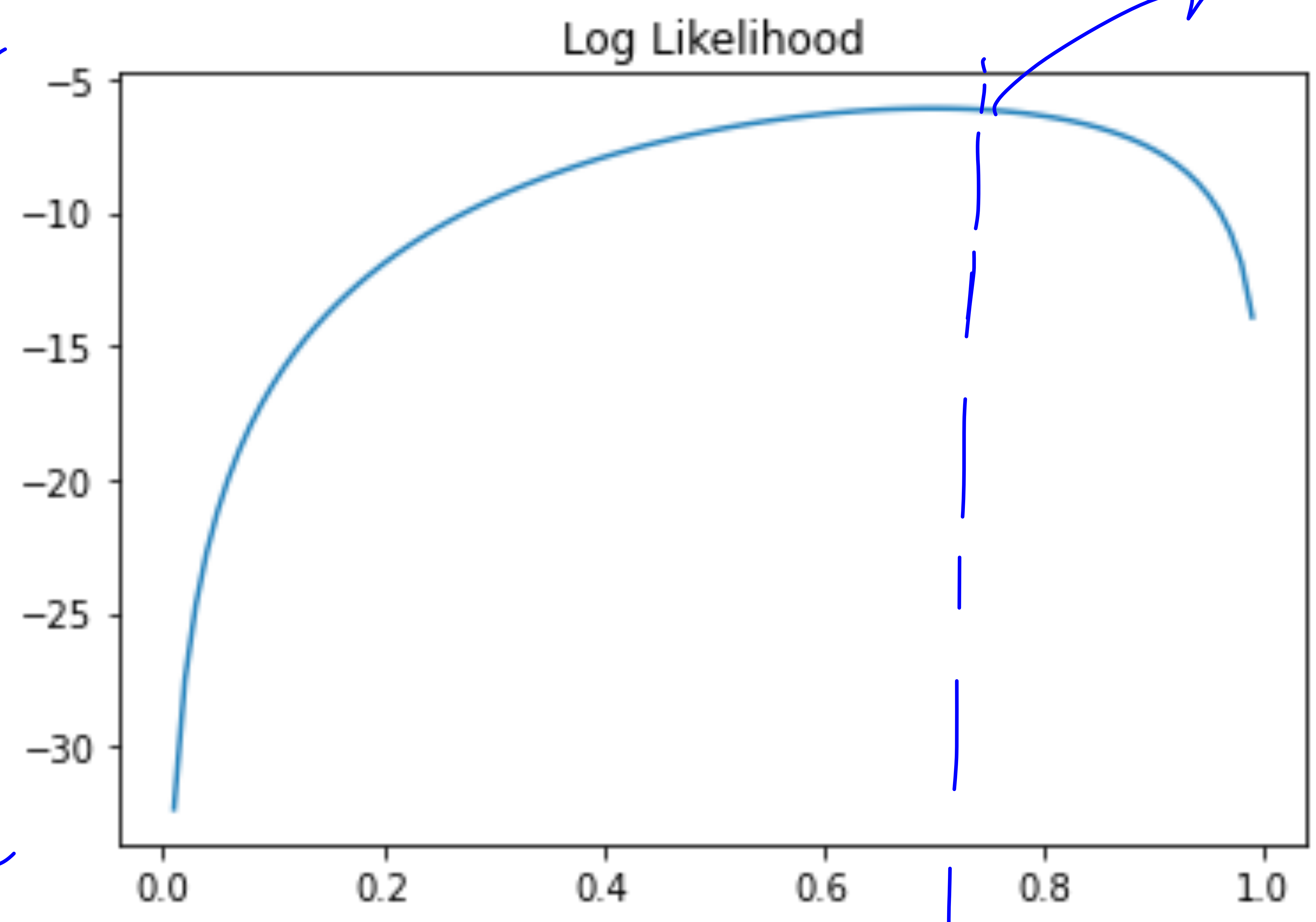
x는 0~1 사이니까
항상 값은 음수지!!

Log Likelihood $\log \mathscr{L}(p) = 7\log p + 3\log(1-p)$

- 취해줄게

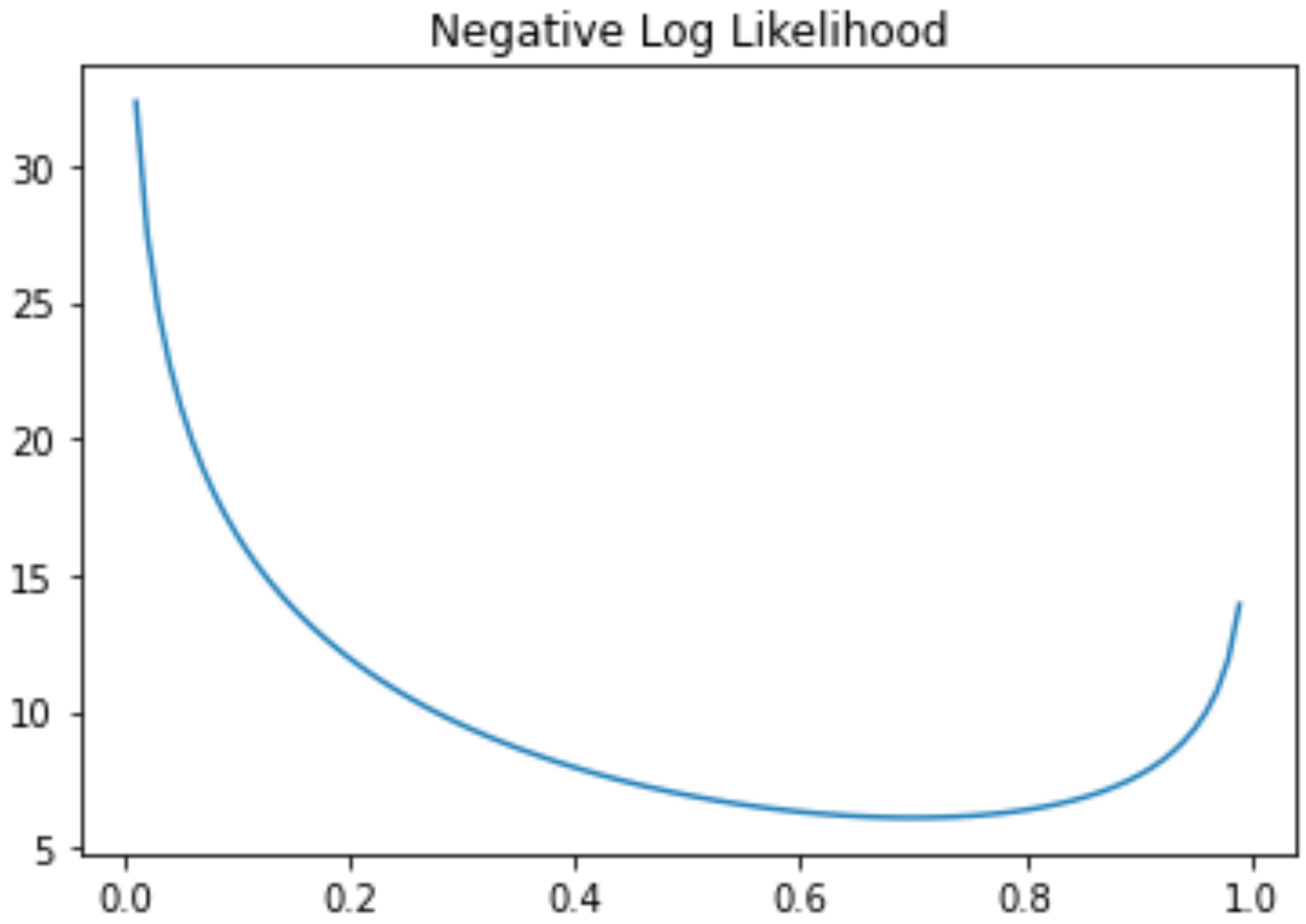
log 최대
0.7에서
최대네

음수



Negative Log Likelihood (Example)

NLL $$-\log \mathscr{L}(p) = -7\log p - 3\log(1-p)$$

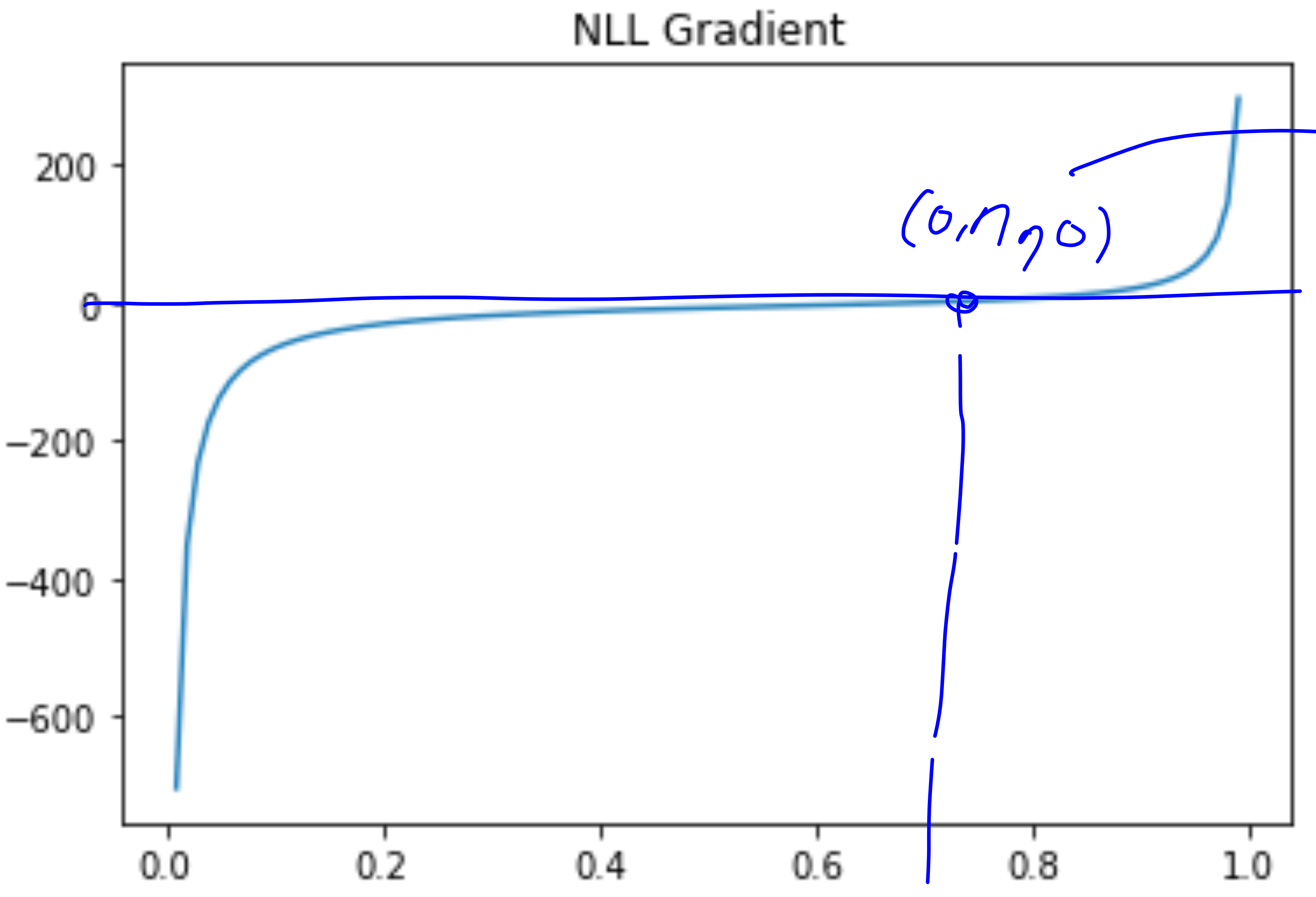


Negative Log Likelihood (Example)

NLL Gradient

$$-\frac{\log \mathscr{L}(p)}{dp} = -7\frac{1}{p} + 3\frac{1}{1-p}$$

p=0.7일 때
gradient가 0이징



기울기가 0인
극소점이 되는거지.

감사합니다.