Lecture 13: Maximum Likelihood estimation (MLE)

GENOME 560, Spring 2017

Su-In Lee, CSE & GS (suinlee@uw.edu)

Review of Last Lecture

- What did we learn in the last lecture?

Review of Last Lecture

- What did we learn in the last lecture?
  - Bayesian network representation
  - Joint distribution of Bayesian networks
  - Data likelihood

Outline

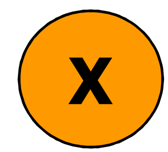
- Basic concepts of parameter estimation
  - Maximum likelihood estimation (MLE)
- MLE for Bayesian networks
- R exercise
- Maximum a posteriori (MAP) estimation

LET’S CONSIDER THE SIMPLEST EXAMPLE.

The *Thumbtack* example

- Parameter estimation for a single variable

- Variable
  - X - an outcome of a thumbtack toss
  - Val(X) = {head, tail}



- Data
  - A set of thumbtack tosses: x[1] … x[M]

Maximum likelihood estimation

- Say that $P(x=head) = \Theta$, $P(x=tail) = 1-\Theta$
  - $P(HHTTHHH\ldots<M_h$ heads, M_t tails$>; \Theta) = \Theta^{Mh} (1-\Theta)^{Mt}$

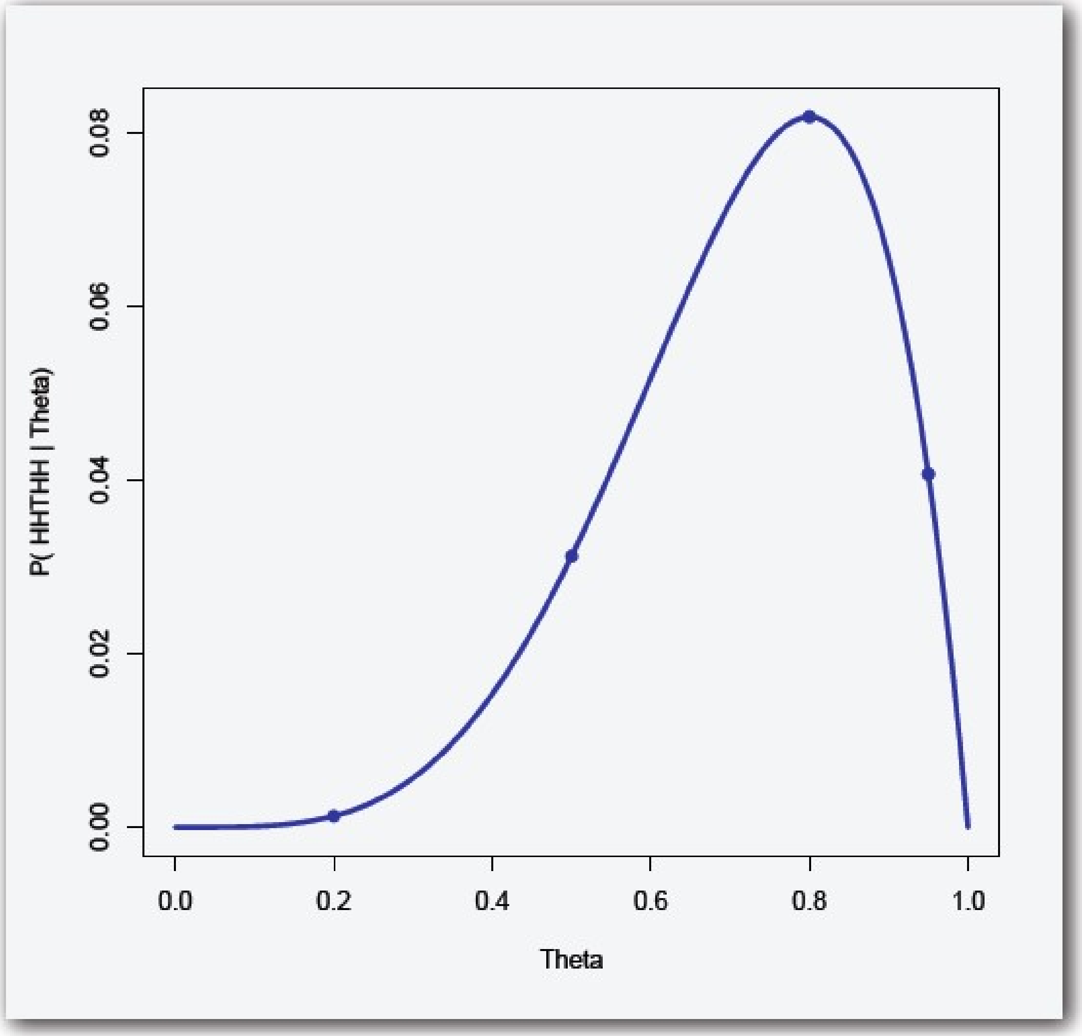
- **Definition:** The likelihood function
  - $L(\Theta : D) = P(D; \Theta)$

- Maximum likelihood estimation (MLE)
  - Given data $D=HHTTHHH\ldots<M_h$ heads, M_t tails$>$, find Θ that maximizes the likelihood function $L(\Theta : D)$.
  - Say that $M_h = 4$ and $M_h = 1$. Write down the likelihood function. $\Theta^4 (1-\Theta)$

Likelihood function

Probability of HHTHH,
given $P(H) = \theta$:

θ	$\theta^4(1-\theta)$
0.2	0.0013
0.5	0.0313
0.8	0.0819
0.95	0.0407



MLE for the *Thumbtack* problem

- Given data D=HHTTHHH…<M_h heads, M_t tails>
  - MLE solution $\theta^* = M_h / (M_h+M_t)$.

- Proof:

MLE for general problems

- Learning problem setting
  - A set of random variables X from unknown distribution P^*
  - Training data D = M instances of X: { d[1] … d[M] }

- A *parametric model* $P(X \mid \theta)$ (a ‘legal’ distribution)

- Define the likelihood function:
  - $L(\theta : D) = P(X \mid \theta)$

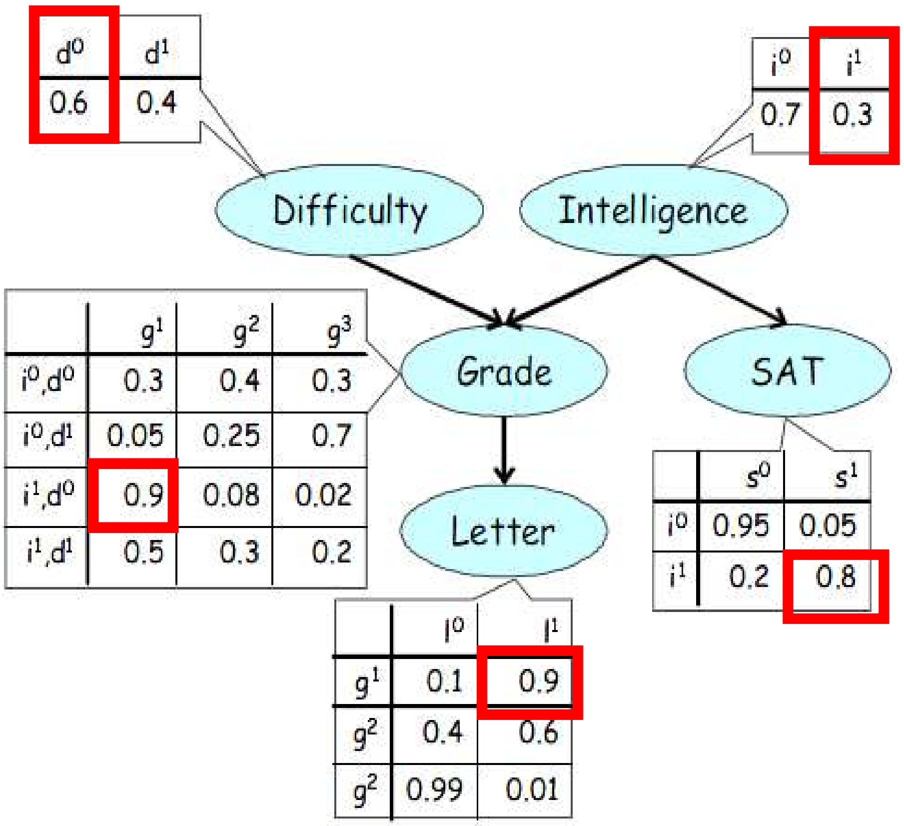
- Maximum likelihood estimation
  - Choose parameters θ^* that satisfy: $\arg\max_{\Theta} L(\Theta : D)$

Outline

- Basic concepts of parameter estimation
  - Maximum likelihood estimation (MLE)
- MLE for Bayesian networks
- R exercise
- Maximum a posteriori (MAP) estimation

Likelihood function for 1 sample

- P(D,I,G,L,S) = P(D) P(I) P(G|D,I) P(S|I) P(L|G)



What is the probability of observing {easy, intelligent, good, strong, high} ?

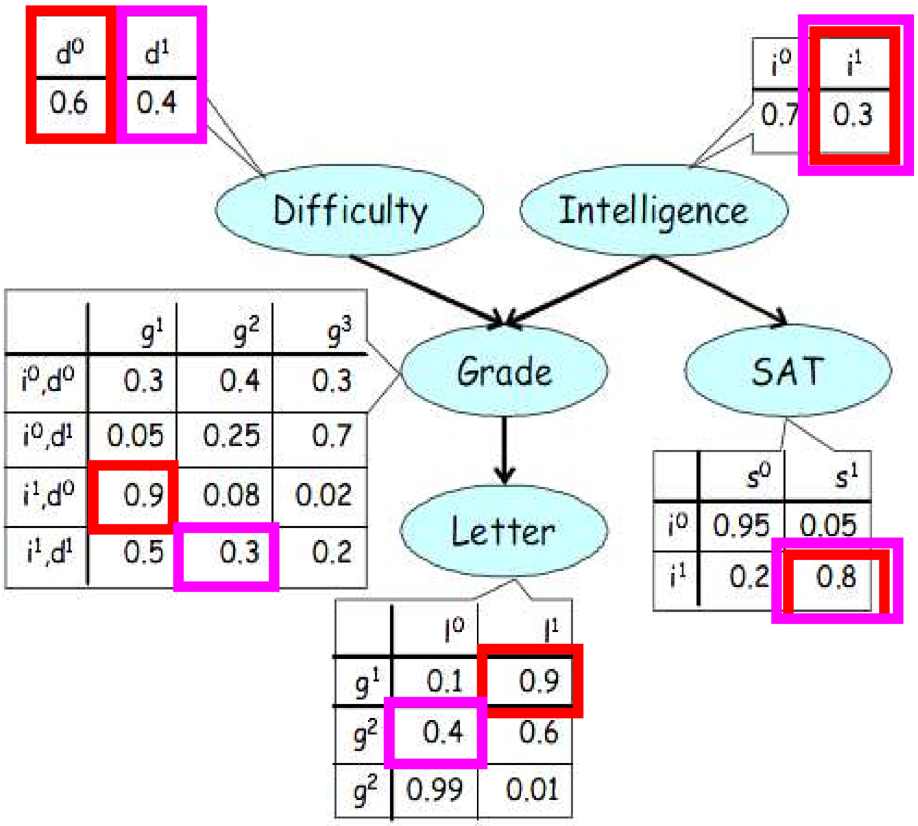
P(D=easy) P(I=intelligent)
P(G=good | D=easy, I=intelligent)
P(S=strong | I=intelligent)
P(L=strong | G=good)

= 0.6 x 0.3 x 0.9 x 0.9 x 0.8

= 0.1166

Likelihood function for 2 samples

- P(D,I,G,L,S) = P(D) P(I) P(G|D,I) P(S|I) P(L|G)



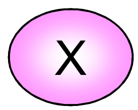
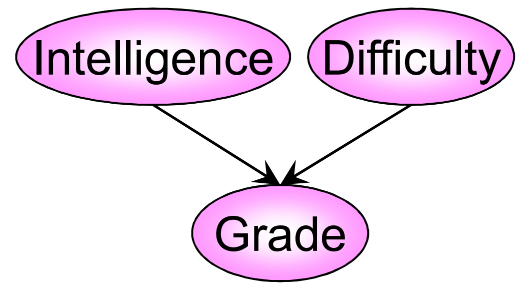
What is the probability of observing {easy, intelligent, good, strong, high} <u>and</u> <u>{difficult, intelligent, medium, bad, high}</u>?

P(D=easy) P(I=intelligent) P(G=good | D=easy, I=intelligent) P(S=strong | I=intelligent) P(L=strong | G=good)

<u>P(D=difficult) P(I=intelligent) P(G=medium | D=difficult, I=intelligent) P(S=strong | I=intelligent) P(L=bad | G=medium)</u>

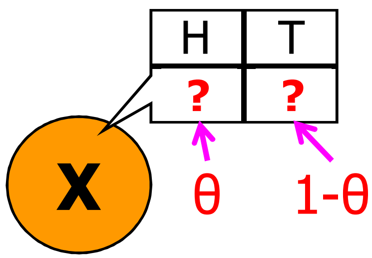
$$= 0.6 \times 0.3 \times 0.9 \times 0.9 \times 0.8$$
$$\underline{\times 0.4 \times 0.3 \times 0.3 \times 0.8 \times 0.4}$$
$$= 0.1166 \ \underline{\times 0.01152}$$
$$= 0.00134$$

Bayesian Network with Table CPDs

	The *Thumbtack* example	vs	**The *Student* example**
			
Joint distribution	P(X)		P(I,D,G) = P(I)P(D)P(G|I,D)
Parameters	θ		θ_I, θ_D, $\theta_{G\|I,D}$
Data	D: {H…x[m]…T}		D: $\{(i^1,d^0,g^1)\ldots(i[m],d[m],g[m])\ldots\}$
Likelihood function **L(θ:D) = P(D;θ)**	$\theta^{M_h}(1-\theta)^{M_t}$		$\theta_{I=i^1}^{M_{I=i^1}} \theta_{I=i^0}^{M_{I=i^0}} \theta_{D=d^1}^{M_{D=d^1}} \theta_{D=d^0}^{M_{D=d^0}} \theta_{G=g^1 \mid I=i^1,D=d^1}^{M_{G=g^1 \mid I=i^1,D=d^1}} \cdots$
MLE solution	$\hat{\theta} = \frac{M_h}{M_h + M_t}$		$\theta_{G=g^1 \mid I=i^1,D=d^0} = \frac{M_{G=g^1,I=i^1,D=d^0}}{M_{I=i^1,D=d^0}}$

MLE in Bayesian networks – easy case

- Let's consider the Bayesian network with 1 variable.



M instances

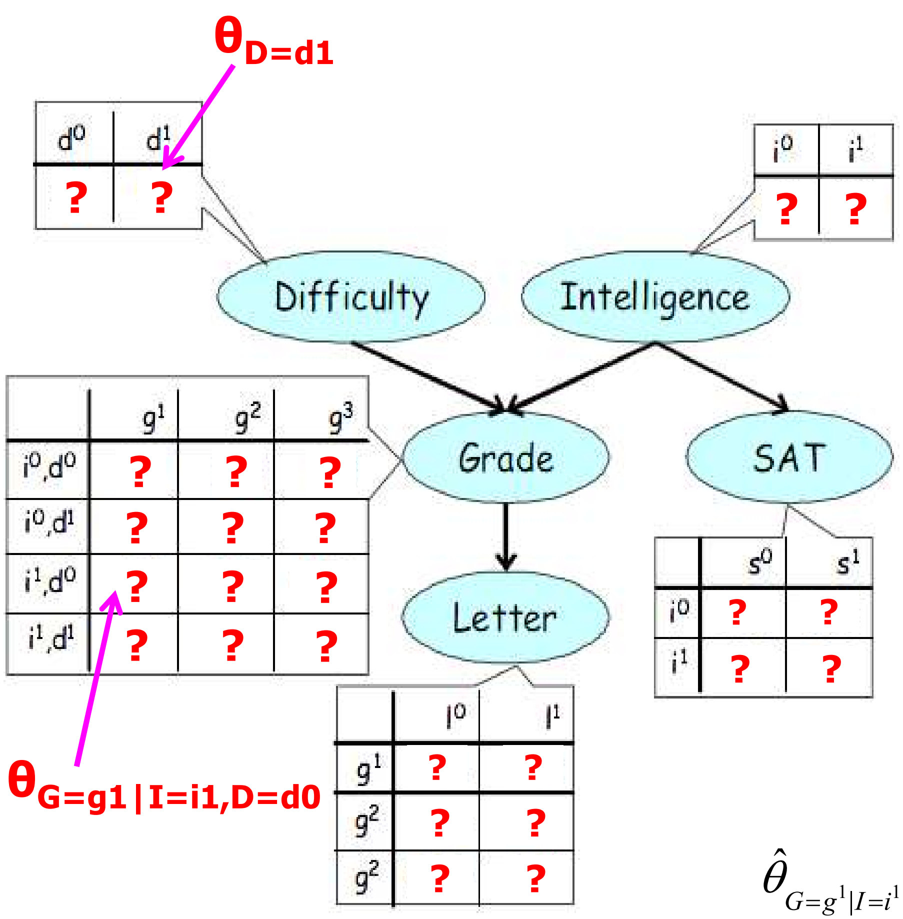
D = H T H T T H H T …

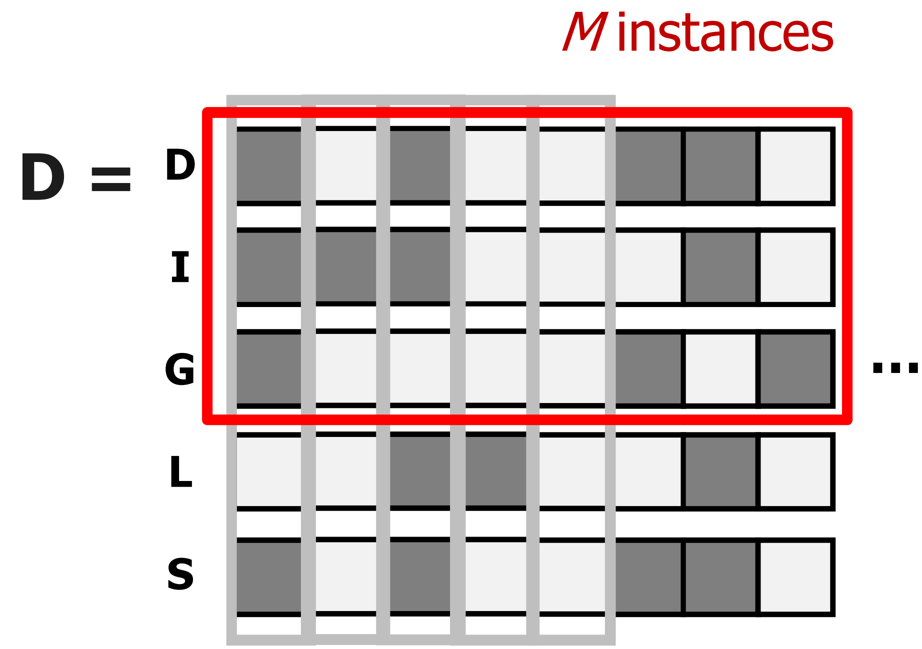
$$\hat{\theta} = \frac{M_h}{M_h + M_t}$$

Number of heads

Total number of tosses

MLE in Bayesian networks – harder case





$$\hat{\theta}_{D=d^1} = \frac{M_{D=d^1}}{M}$$

Number of instances with $D = d^1$

Total number of instances

$$\hat{\theta}_{G=g^1|I=i^1,D=d^0} = \frac{M_{G=g^1,I=i^1,D=d^0}}{M_{I=i^1,D=d^0}}$$

Number of instances with {G=g^1, I=i^1, D=d^0}

Number of instances with {I=i^1, D=d^0}

MLE review

- Find parameter estimates which make observed data most likely – maximize P(**D** | **θ**)
- General approach, as long as tractable likelihood function exists
- Can use all available information
  - Network structure constructed based on prior knowledge
  - Parameterization
  - Training data **D**

Continuous Space

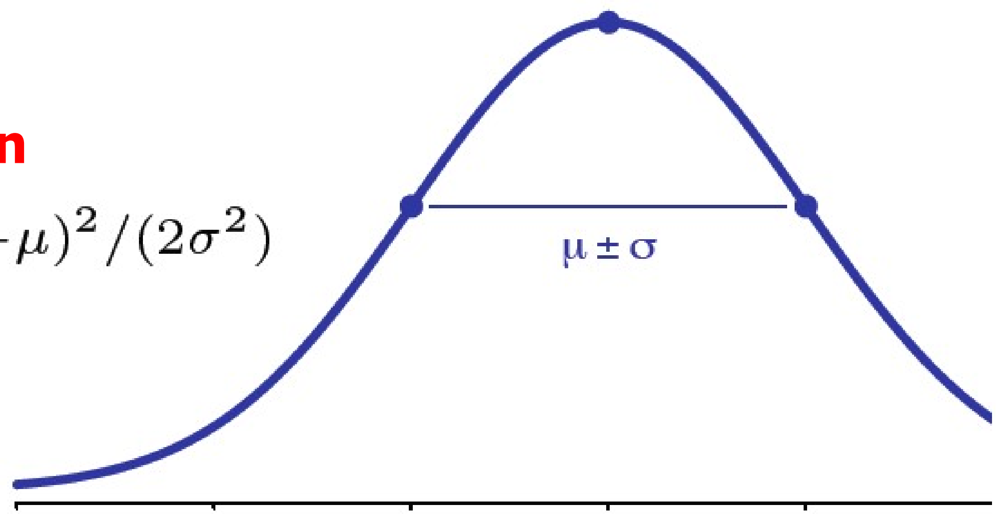
- Assuming sample $x_1, x_2, \ldots, x_n$ is from a parametric probabilistic density function $f(x|\theta)$, estimate θ.

- Say that the n samples are from a normal distribution with mean μ and variance σ^2. (μ, σ^2) are parameters.

Probability density function

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

$$\theta = (\mu, \sigma^2)$$



Continuous Space (cont.)

- Let $\theta_1 = \mu$, $\theta_2 = \sigma^2$

$$L(\theta_1, \theta_2 : x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \exp\left[-\sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}\right]$$

$$\log L(\theta_1, \theta_2 : x_1, x_2, \ldots, x_n) = -n \log\left(\sqrt{2\pi\theta_2}\right) - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial\theta_1} \log L(\theta_1, \theta_2 : x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \frac{(x_i - \theta_1)}{\theta_2} \mathbf{= 0}$$

$$\boxed{\theta_1^* = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

$$\frac{\partial}{\partial\theta_2} \log L(\theta_1, \theta_2 : x_1, x_2, \ldots, x_n) = -\frac{n}{\sqrt{\theta_2}} + \frac{1}{\theta_2\sqrt{\theta_2}} \sum_{i=1}^{n} (x_i - \theta_1)^2 \mathbf{= 0}$$

$$\boxed{\theta_2^* = \frac{1}{n} \sum_{i=1}^{n} (x_i - \theta_1^*)^2}$$

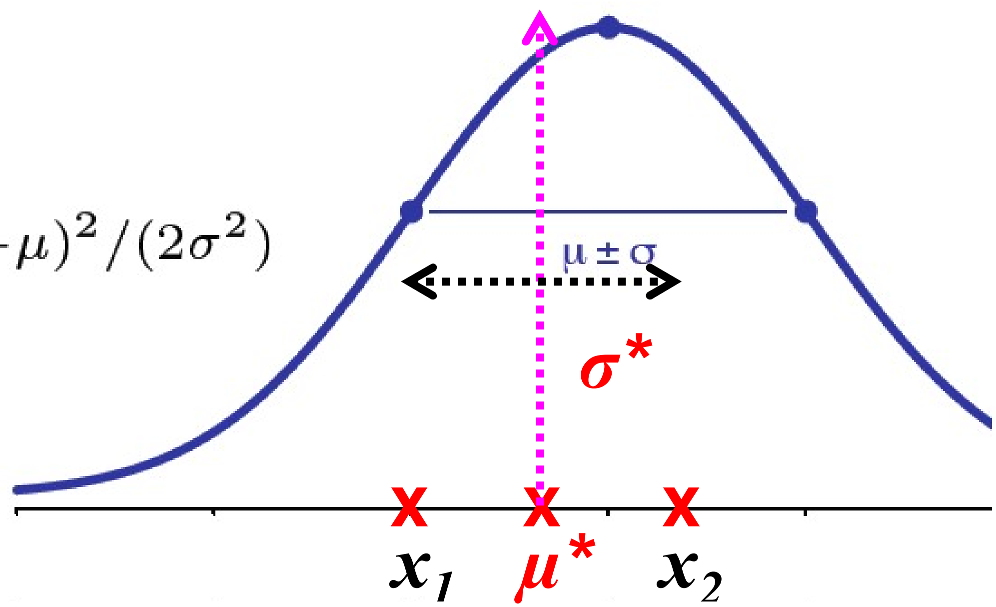
Any Drawback?

- Is it biased?
  - Yes, as an extreme case when n = 1, σ^2* = 0.

- The MLE solution systematically underestimates σ^2*.
  - Let's say n = 2.

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

$$\theta = (\mu, \sigma^2)$$



Outline

- Basic concepts of parameter estimation
  - Maximum likelihood estimation (MLE)
- MLE for Bayesian networks
- R exercise
- Maximum a posteriori (MAP) estimation

The *Halitosis* Example

- **Halitosis**, colloquially called **bad breath**, is a symptom in which a noticeably unpleasant odor is present on the exhaled breath.

- Halitosis is partly genetically determined. The genotype aa has a 40% chance of getting the disease, and the other two possible genotypes, AA and Aa, each has a 10% chance of getting the disease. We want to estimate the frequency of the A allele.

The *Halitosis* Example

- P(getting the disease | AA) = 0.1
  P(getting the disease | Aa) = 0.1
  P(getting the disease | aa) = 0.4

- Now suppose we observe 1000 individuals and find that the 182 of them have the disease.

- What is the allele frequency?

The *Halitosis* Example

- Let's use R to solve this problem.

- The frequency of the disease is expected to be:

$$F(p) = 0.1 \cdot p^2 + 0.1 \cdot 2p(1-p) + 0.4 \cdot (1-p)^2$$

- Define a function:

```
freq.halitosis <- function(p){
return( 0.1*p^2+0.1*2*p*(1-p)+0.4*(1-p)^2 )
}
```

- Define another function:

```
ll.halitosis <- function(f){
return( 182 * log(f) + 818 * log(1-f) )
}
```

The *Halitosis* Example

- What is the value of p that maximizes the likelihood function?

- Find the MLE:
```
p <- seq(0, 1, 0.001)
ll <- ll.halitosis( freq.halitosis( p ) )
```

- Plot the log-likelihood function
```
plot (p, ll, xlim=range(0:1), xlab = "allele frequence p", ylab="log-likelihood")
grid(10,10)
```

- Find the MLE
```
which.max (ll)
p[which.max (ll)]
```

- Add a straight line
```
abline(v= p[which.max (ll)])
abline(v= p[which.max (ll)], col ="red")
```

Outline

- Basic concepts of parameter estimation
  - Maximum likelihood estimation (MLE)
- MLE for Bayesian networks
- R exercise
- Maximum a posteriori (MAP) estimation

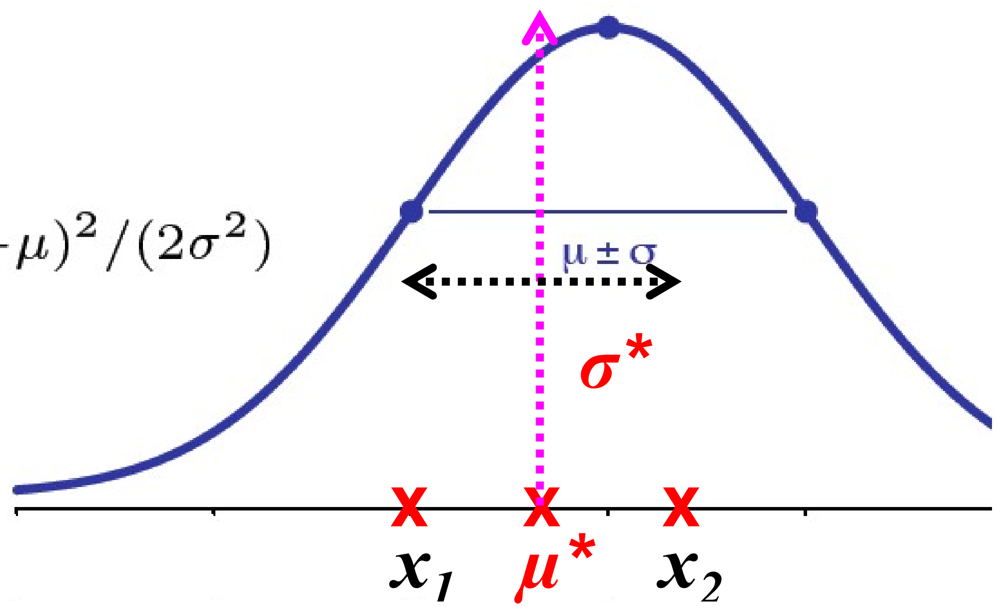
Any Drawback?

- Is it biased?
  - Yes, as an extreme case when $n = 1$, $\sigma^{2*} = 0$.

- The MLE solution systematically underestimates σ^{2*}.
  - Let's say n = 2.

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

$$\theta = (\mu, \sigma^2)$$



Maximum a posteriori (MAP)

- Incorporating "priors"
  - E.g., The chance of "head" is close to 0.5
  - The mean of the normal distribution is close to 0

- MLE vs. MAP estimation
  - **MLE:** maximize P(D | θ)
  - **MAP:** maximize P(θ | D)

$$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)}$$