Probabilities

MLE

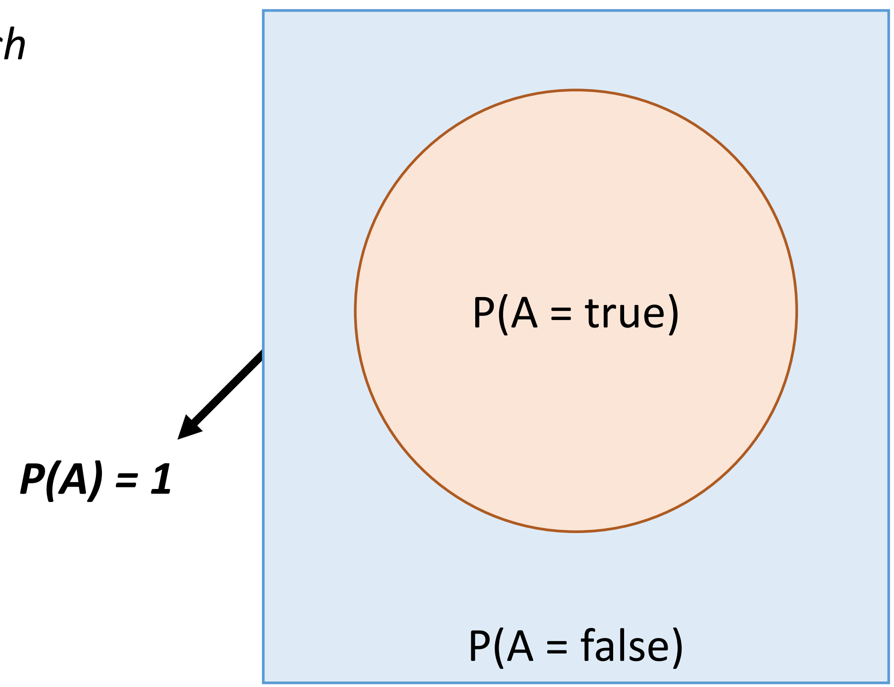
Maximum Likelihood Estimation

확률

P(A = true), 사건 *A* 가 *true* 일 확률

P 는 상대적인 빈도

- *Frequentist Approach*
- *Bayesian Approach*



50%

50%

$\frac{3}{5}$

$\frac{2}{5}$

Binomial Distribution

The binomial distribution with parameters ***n*** and ***p*** is the **discrete probability distribution**

Boolean-valued outcome: success/yes/true/one (with probability ***p***) or

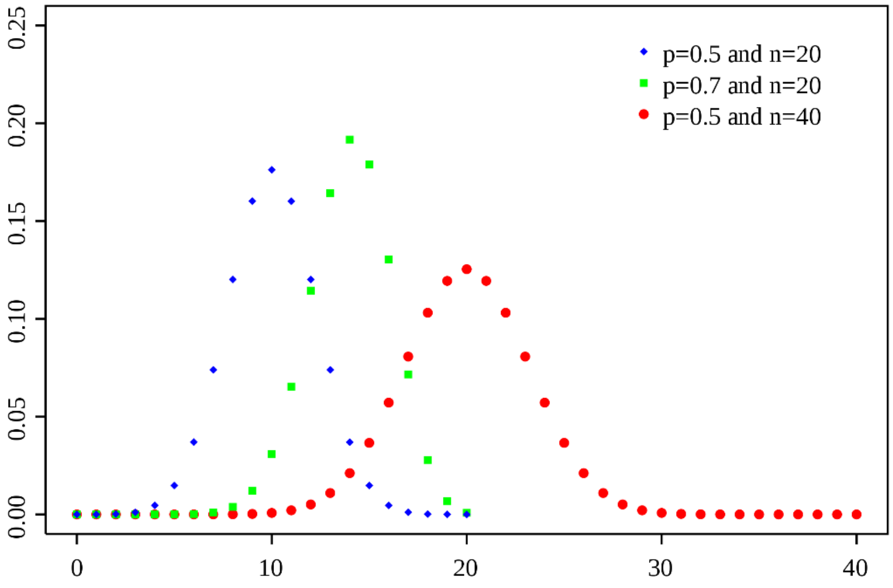
(Bernoulli trial or Bernoulli experiment) failure/no/false/zero (with probability ***q*** = ***1 − p***)

The binomial distribution is a **Bernoulli distribution**

I.I.D condition:

Independent events

Identically **D**istributed according to binomial distribution



$$P(H) = \theta,\ \mathrm{P(T)} = 1 - \theta \quad (P \geq 0)$$

$$P(HHTHT) = \theta \cdot \theta \cdot (1 - \theta) \cdot \theta \cdot (1 - \theta)$$

$$= \theta^3 \cdot (1 - \theta)^2$$

$$D = H, H, T, H, T$$

$$n = 5$$

$$p = \theta$$

$$k = a_H = 3$$

$$P(D|\theta) = \theta^{a_H}(1 - \theta)^{a_T}$$

$$f(k; n, p) = P(K = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!\,(n - k)!}$$

Maximum Likelihood Estimation

$P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$

Hypothesis follows the binomial distribution of θ

How to make our hypothesis strong?

Finding out a better distribution of the observation

Finding out the **best candidate** of θ

Maximum Likelihood Estimation of θ

Choose θ that maximizes the probability of observed data

$$\boldsymbol{\hat{\theta} = argmax_{\theta} P(D|\theta)}$$

$$\hat{\theta} = argmax_{\theta} P(D|\theta) = argmax_{\theta} \theta^{a_H}(1-\theta)^{a_T}$$

$$\hat{\theta} = argmax_{\theta} \ln P(D|\theta) = argmax_{\theta} \ln \theta^{a_H}(1-\theta)^{a_T}$$

$$= argmax_{\theta}\, a_H \ln \theta + a_T \ln(1-\theta)$$

$$\frac{d}{d\theta}(a_H \ln \theta + a_T \ln(1-\theta)) = 0$$

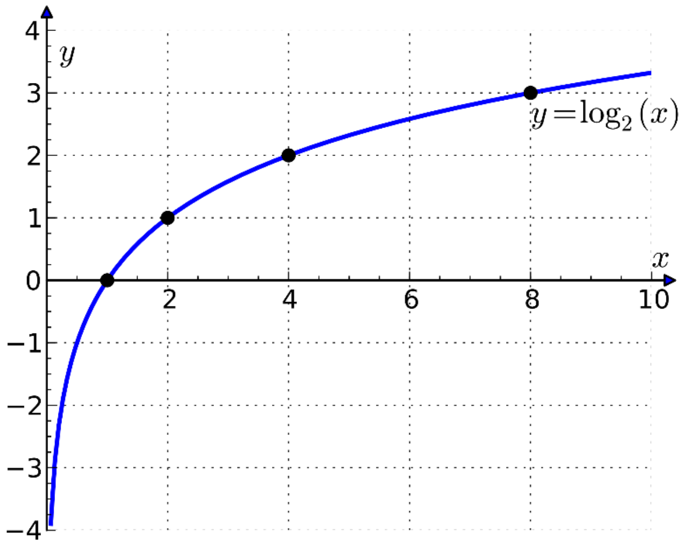
$$\frac{a_H}{\theta} - \frac{a_T}{1-\theta} = 0$$

$$\frac{a_H}{\theta} = \frac{a_T}{1-\theta}$$

$$(1-\theta)a_H = a_T\theta$$

$$a_H - a_H\theta = a_T\theta$$

$$\boldsymbol{\hat{\theta} = \frac{a_H}{a_H + a_T}}$$



PAC Learning

$\widehat{\boldsymbol{\theta}} = \frac{\boldsymbol{a_H}}{\boldsymbol{a_H} + \boldsymbol{a_T}}$ $\boldsymbol{N} = \boldsymbol{a_H} + \boldsymbol{a_T}$

$\boldsymbol{\theta}^*$ is true parameter, $\boldsymbol{\varepsilon} > \mathbf{0}$

Error bound: $P\left(\left|\hat{\theta} - \theta^*\right| \geq \varepsilon\right) \leq 2e^{-2Ne^2}$

Can you calculate the required number of trials, N? $\boldsymbol{\varepsilon} = \mathbf{0.1}$ **with 0.01%**

⟶ **Probably Approximate Correct** learning

MAP

Maximum A Posteriori estimation

50%

50%

Prior Knowledge

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$ ← *Bayes Theorem*

$$Posterior = \frac{Likelihood \times Prior\ Knowledge}{Normalizing\ Constant}$$

MLE: $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$

$P(\theta|D)$ is conclusion influenced by the data and the prior knowledge

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

$$P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$$

$$P(\theta) = \cancel{50\%?} \longrightarrow \textit{Binomial Distribution}$$

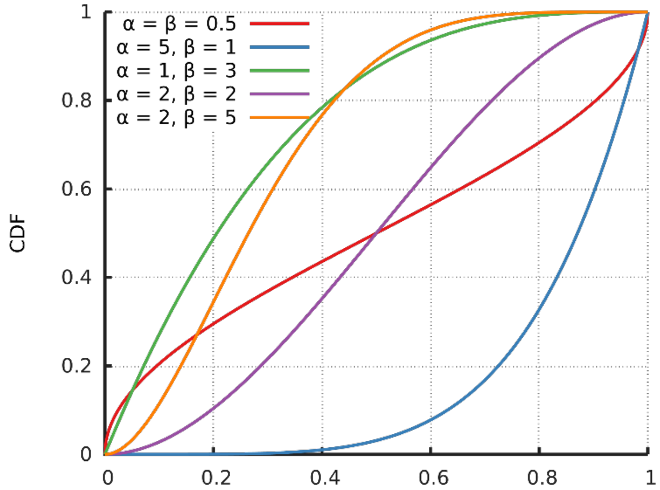
$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \longleftarrow \textit{Beta Distribution} \qquad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\Gamma(\alpha) = (\alpha-1)!$$

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

$$\propto \theta^{a_H}(1-\theta)^{a_T}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$$



Maximum A Posteriori estimation

in MLE, we found θ from $\hat{\theta} = argmax_{\theta} P(D|\theta)$

$P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$

$$\hat{\theta} = \frac{a_H}{a_H + a_T}$$

in MAP, we find θ from $\hat{\theta} = argmax_{\theta} P(\theta|D)$

$P(\theta|D) = \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$

$$\hat{\theta} = \frac{a_H + \alpha - 1}{a_H + \alpha + a_T + \beta - 2}$$

MLE

$$\hat{\theta} = \frac{a_H}{a_H + a_T}$$

MAP

$$\hat{\theta} = \frac{a_H + \alpha - 1}{a_H + \alpha + a_T + \beta - 2}$$

α and β are very important

If a_H and a_T become big, α, β becomes nothing

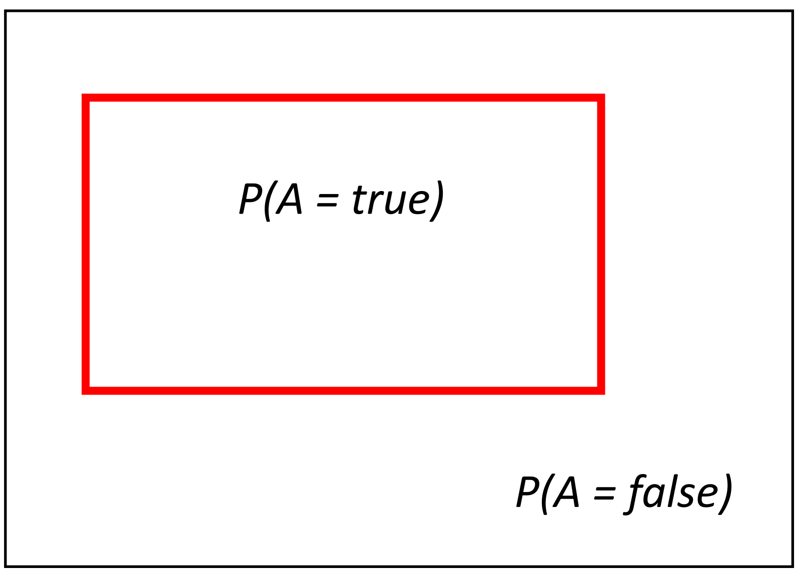
Bayes Theorem

Probability

P(A = true) means the probability that A = true

It is relative frequency with which an outcome would be obtained

If the process were repeated a large number of times under similar conditions



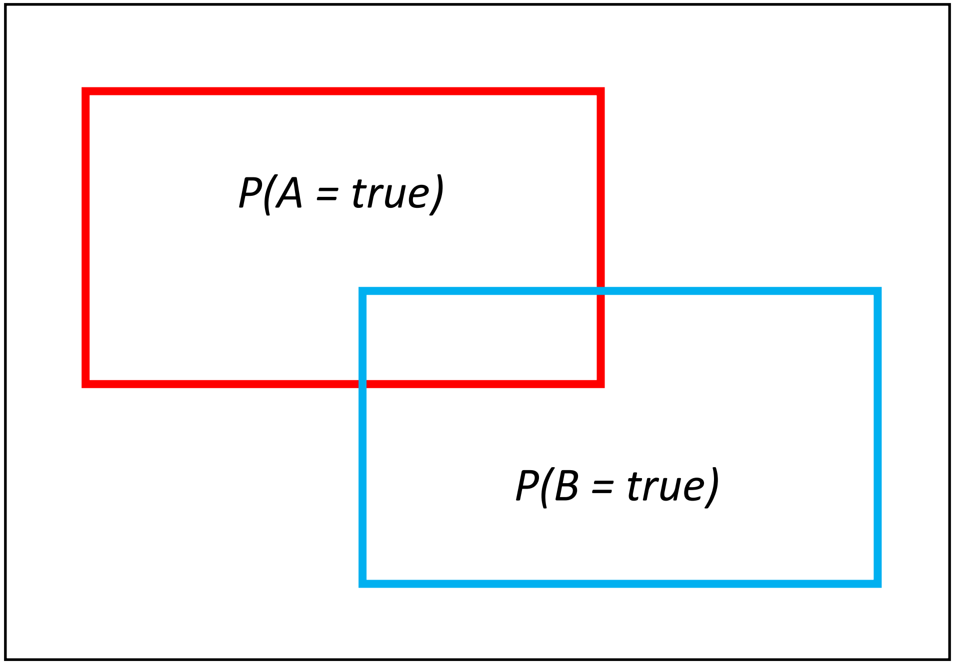
freq(A = true) = 10, so what?

Conditional Probability

확률만으로는 충분한 정보를 획득하기 어려움 (예측이 어려움)

$P(A = true \mid B = true)$, 사건 *B* 가 *true* 일 때(조건), 사건 *A* 가 *true* 일 확률

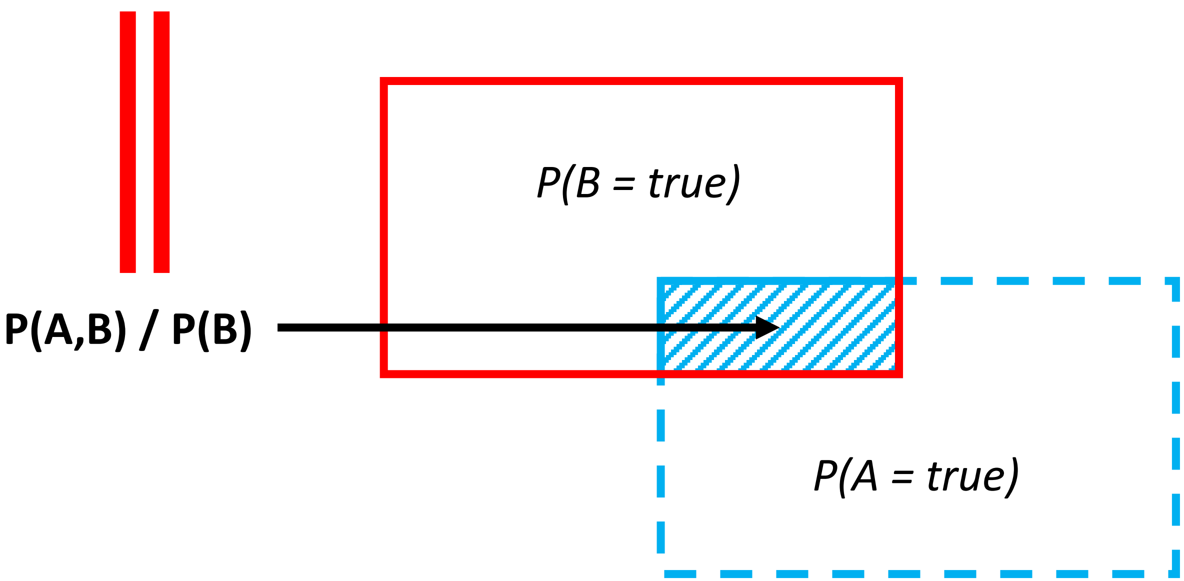
두통이 올 때, 감기가 올 확률?



A = 두통이 있음
B = 감기에 걸림 일 때,

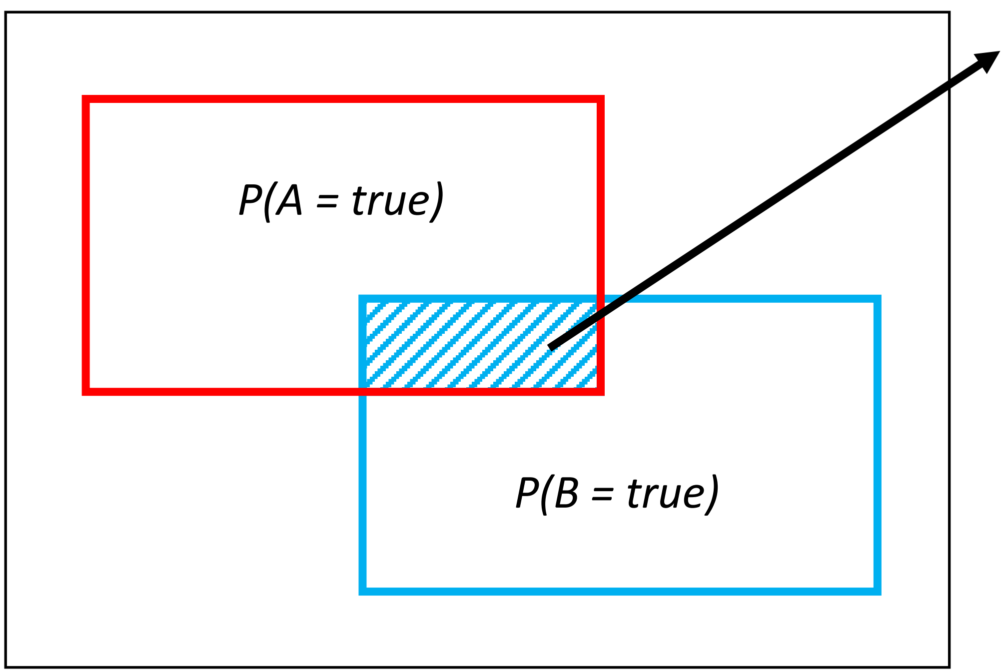
P(A = true) = 1/10
P(B = true) = 1/40

P(A = true | B = true) = 1/2 → *감기에 걸렸을 때, 두통이 있을 확률*



Joint Probability

P(A,B), 사건 *A* 가 *true* 이고, 사건 *B* 도 *true* 일 확률



P(B = true, A = true)

$$= \frac{Area\ of\ A\ and\ B\ region}{Area\ of\ A\ region}$$

$$= \frac{P(A{=}true, B{=}true)}{P(A)}$$

Joint Probability

P(X|Y) = P(X,Y) / P(Y)

Conditional Probability

개별 확률과 **결합 확률**을 알면, **조건부 확률**을 계산할 수 있음!

P(A), P(B)와 **P(A,B)**를 알면, **P(A|B)** 을 계산할 수 있음!

Total Probability *(Summing out, Marginalization)*

$$P(A) = \sum_B P(A,B)$$

(in case of binary) $= P(A, B = true)$

$+P(A, B = false)$

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$= \sum_B P(A|B)/P(B)$$

$$P(A,B) = P(A|B)P(B)$$

결합 확률을 알면, **개별 확률**을 계산할 수 있음!

P(A,B)를 알면, **P(A)**, **P(B)**를 계산할 수 있음!

조건부 확률과 **개별 확률**을 알면, **결합 확률**을 계산할 수 있음!

P(A|B)와, **P(A)**, **P(B)**를 알면, **P(A,B)**를 계산할 수 있음!

결합 확률 P(A,B,C,D)를 알 때,

개별 확률 P(B)는

$$P(B) = \sum_A \sum_C \sum_D P(A,B,C,D)$$

결합 확률 P(A,B,C,D)를 알 때,

조건부 확률 P(C|B)는

$$P(C|B) = \sum_A \sum_D P(A,C,D|B)$$

$$= \frac{\sum_A \sum_D P(A,B,C,D)}{P(B)}$$

$$= \frac{1}{P(B)} \times \sum_A \sum_D P(A,B,C,D)$$

Normalization Constant

결합 확률을 알면, **개별 확률**과 **조건부 확률**을 알 수 있음

그러나, 파라미터 수가 급격하게 증가함

P(A,B,C,D)이고, 각 확률변수에 2가지(true, false) 경우가 있다면?

{true, true, true, true}

•

•

•

{false, false, false, false}

$2^4 = 16$ 가지의 결합 확률이 필요함

확률변수 E, F가 추가 된다면? **불가능에 가까움**

Factorization

$$P(A, B, C, \cdots, Z) = P(A|B, C, \cdots, Z)P(B, C, \cdots, Z)$$

P(X|Y) = P(X,Y) / P(Y)

Conditional Probability

Chain Rule

$$= P(A|B, C, \cdots, Z)P(B|C, \cdots, Z)P(C|\cdots, Z)\cdots P(Z)$$