

# Mathematics for Artificial Intelligence

7강: 통계학 맛보기

---



임성빈

**UNIST**

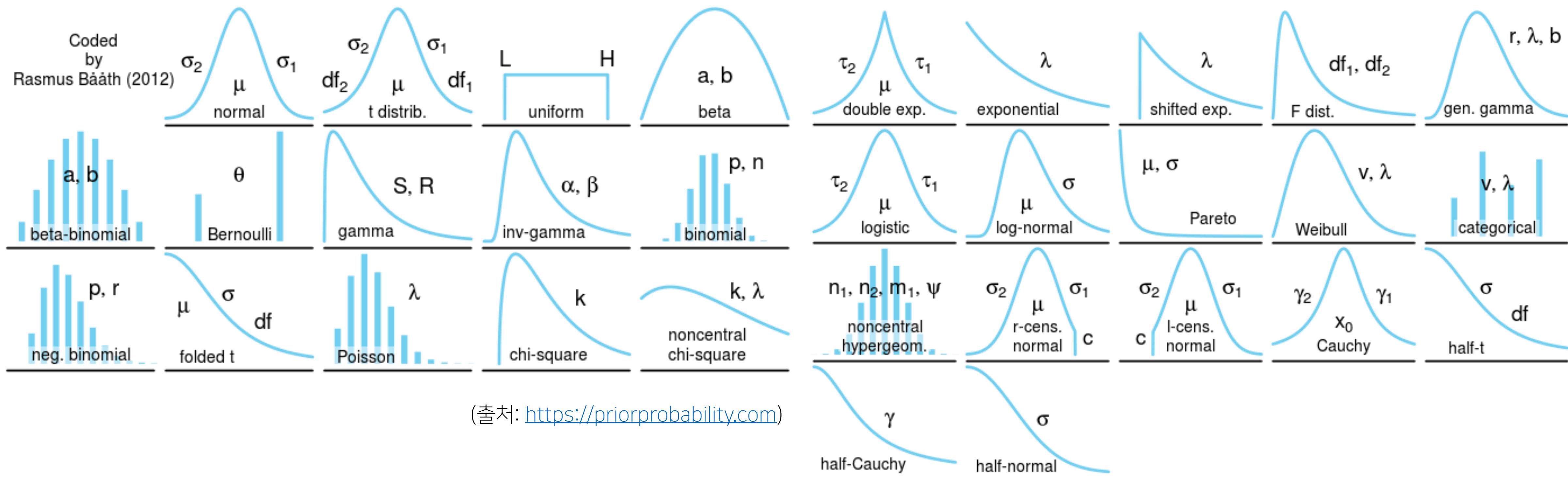
인공지능대학원 & 산업공학과  
Learning Intelligent Machine Lab

X  
**boostcamp** AI Tech

© NAVER Connect Foundation

# 모수가 뭐에요?

- 통계적 모델링은 적절한 가정 위에서 확률분포를 추정(inference)하는 것이 목표이며, 기계학습과 통계학이 공통적으로 추구하는 목표입니다

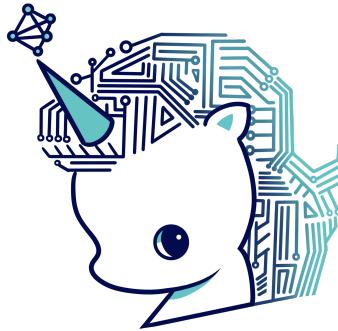


(출처: <https://priorprobability.com>)

# 모수가 뭐에요?

---

- 통계적 모델링은 적절한 가정 위에서 확률분포를 추정(inference)하는 것이 목표이며, 기계학습과 통계학이 공통적으로 추구하는 목표입니다
- 그러나 유한한 개수의 데이터만 관찰해서 모집단의 분포를 정확하게 알아낸다는 것은 불가능하므로, 근사적으로 확률분포를 추정할 수 밖에 없습니다

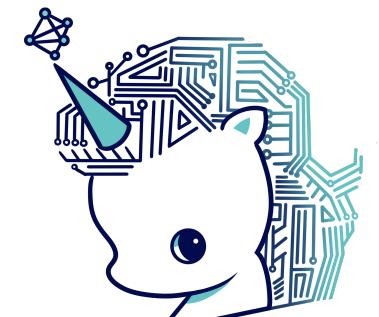


예측모형의 목적은 분포를 정확하게 맞추는 것보다는 데이터와 추정 방법의 불확실성을 고려해서 위험을 최소화하는 것이다

# 모수가 뭐에요?

---

- 통계적 모델링은 적절한 가정 위에서 확률분포를 추정(inference)하는 것이 목표이며, 기계학습과 통계학이 공통적으로 추구하는 목표입니다
- 그러나 유한한 개수의 데이터만 관찰해서 모집단의 분포를 정확하게 알아낸다는 것은 불가능하므로, 근사적으로 확률분포를 추정할 수 밖에 없습니다
- 데이터가 특정 확률분포를 따른다고 선형적으로(a priori) 가정한 후 그 분포를 결정하는 모수(parameter)를 추정하는 방법을 모수적(parametric) 방법론이라 합니다
- 특정 확률분포를 가정하지 않고 데이터에 따라 모델의 구조 및 모수의 개수가 유연하게 바뀌면 비모수(nonparametric) 방법론이라 부릅니다



기계학습의 많은 방법론은 비모수 방법론에 속합니다

X

# 확률분포 가정하기: 예제

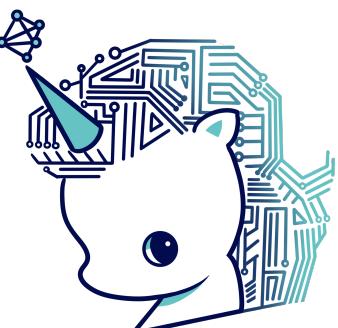
---

- 확률분포를 가정하는 방법: 우선 히스토그램을 통해 모양을 관찰합니다
  - 데이터가 2개의 값(0 또는 1)만 가지는 경우 → 베르누이분포
  - 데이터가  $n$ 개의 이산적인 값을 가지는 경우 → 카테고리분포
  - 데이터가  $[0,1]$  사이에서 값을 가지는 경우 → 베타분포
  - 데이터가 0 이상의 값을 가지는 경우 → 감마분포, 로그정규분포 등
  - 데이터가  $\mathbb{R}$  전체에서 값을 가지는 경우 → 정규분포, 라플라스분포 등

# 확률분포 가정하기: 예제

---

- 확률분포를 가정하는 방법: 우선 히스토그램을 통해 모양을 관찰합니다
  - 데이터가 2개의 값(0 또는 1)만 가지는 경우 → 베르누이분포
  - 데이터가  $n$ 개의 이산적인 값을 가지는 경우 → 카테고리분포
  - 데이터가  $[0,1]$  사이에서 값을 가지는 경우 → 베타분포
  - 데이터가 0 이상의 값을 가지는 경우 → 감마분포, 로그정규분포 등
  - 데이터가  $\mathbb{R}$  전체에서 값을 가지는 경우 → 정규분포, 라플라스분포 등
- 기계적으로 확률분포를 가정해서는 안 되며, **데이터를 생성하는 원리를 먼저 고려하는 것이 원칙입니다**



각 분포마다 검정하는 방법들이 있으므로  
모수를 추정한 후에는 반드시 검정을 해야 한다

X

# 데이터로 모수를 추정해보자!

---

- 데이터의 확률분포를 가정했다면 모수를 추정해볼 수 있습니다
- 정규분포의 모수는 평균  $\mu$  과 분산  $\sigma^2$  으로 이를 추정하는 통계량(statistic)은 다음과 같다:

표본평균

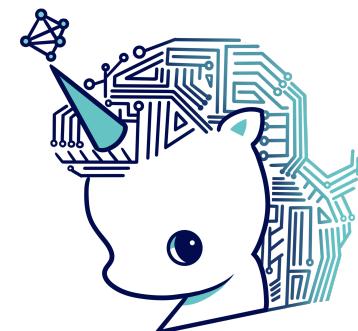
$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

$$\mathbb{E}[\bar{X}] = \mu$$

표본분산

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

$$\mathbb{E}[S^2] = \sigma^2$$



표본분산을 구할 때  $N$  이 아니라  $N - 1$  로 나누는 이유는 불편(unbiased) 추정량을 구하기 위해서다

# 데이터로 모수를 추정해보자!

---

- 데이터의 확률분포를 가정했다면 모수를 추정해볼 수 있습니다
- 정규분포의 모수는 평균  $\mu$  과 분산  $\sigma^2$  으로 이를 추정하는 통계량(statistic)은 다음과 같다:

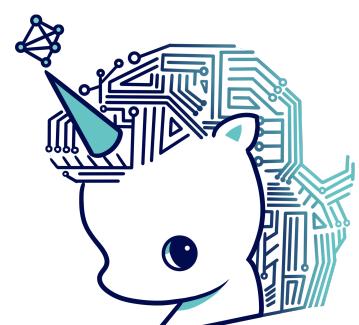
표본평균

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

표본분산

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

- 통계량의 확률분포를 표집분포(sampling distribution)라 부르며, 특히 표본평균의 표집분포는  $N$ 이 커질수록 정규분포  $\mathcal{N}(\mu, \sigma^2/N)$  를 따릅니다



이를 중심극한정리(Central Limit Theorem)이라 부르며,  
모집단의 분포가 정규분포를 따르지 않아도 성립합니다

X

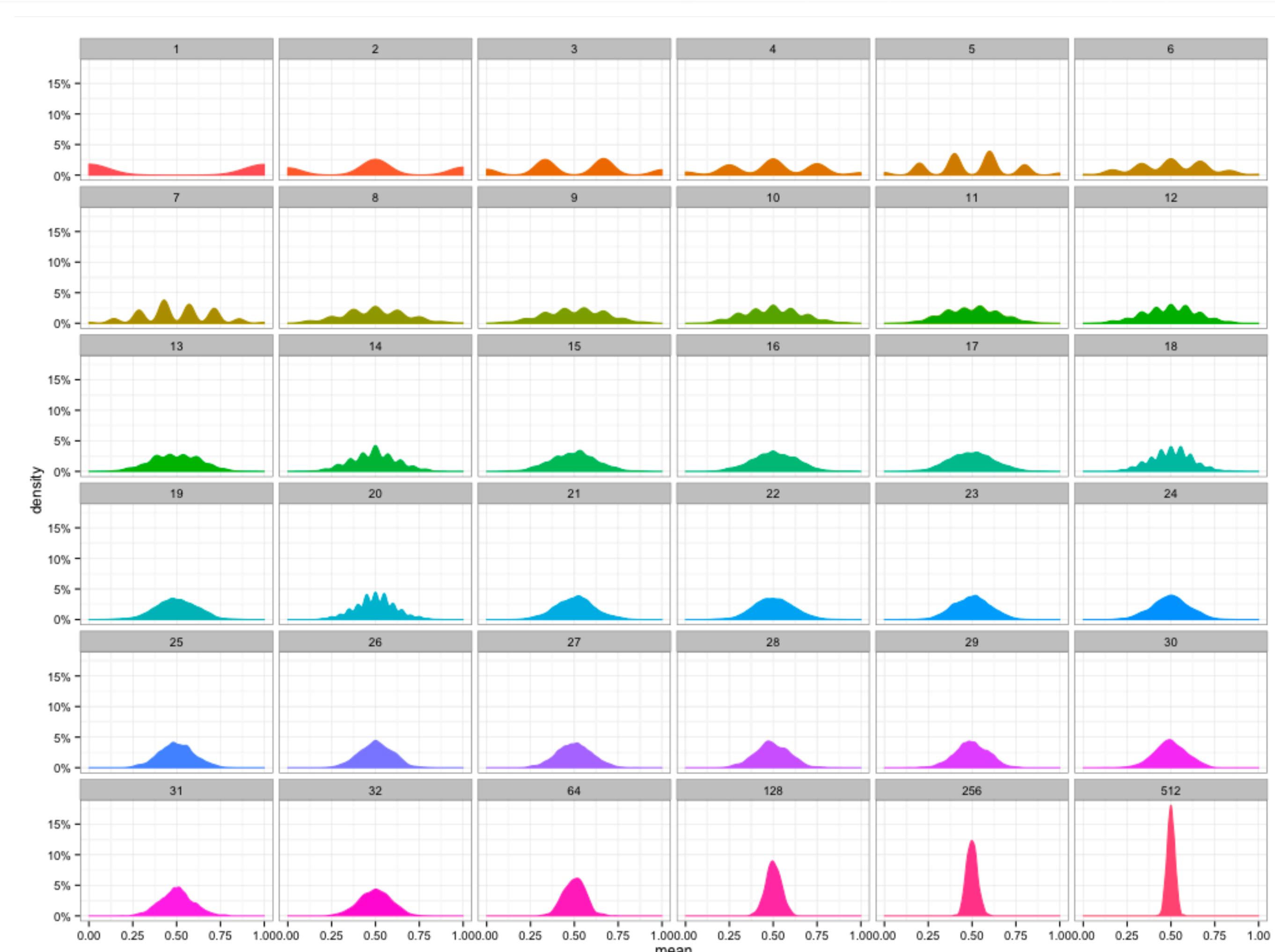
# 데이터로 모수를 추정해보자!

- 데이터의 확률
- 정규분포의 모  
은 다음과 같다

$$\bar{X}$$

- 통계량의 확률  
본평균의 표집

X



다  
계량(statistic)

$$\bar{X})^2$$

부르며, 특히 표  
를 따릅니다

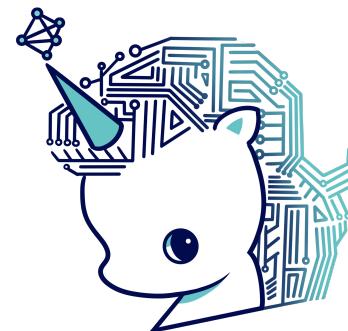
Theorem)이라 부르며,  
| 않아도 성립합니다

# 최대가능도 추정법

---

- 표본평균이나 표본분산은 중요한 통계량이지만 확률분포마다 사용하는 모수가 다르므로 적절한 통계량이 달라지게 됩니다
- 이론적으로 가장 가능성이 높은 모수를 추정하는 방법 중 하나는 **최대가능도 추정법**(maximum likelihood estimation, MLE)입니다

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} L(\theta; \mathbf{x}) = \underset{\theta}{\operatorname{argmax}} P(\mathbf{x}|\theta)$$



가능도(likelihood) 함수는 모수  $\theta$ 를 따르는 분포가  $\mathbf{x}$ 를 관찰할 가능성을 뜻하지만 확률로 해석하면 안됩니다

# 최대가능도 추정법

---

- 표본평균이나 표본분산은 중요한 통계량이지만 확률분포마다 사용하는 모수가 다르므로 적절한 통계량이 달라지게 됩니다
- 이론적으로 가장 가능성이 높은 모수를 추정하는 방법 중 하나는 **최대가능도 추정법**(maximum likelihood estimation, MLE)입니다

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} L(\theta; \mathbf{x}) = \underset{\theta}{\operatorname{argmax}} P(\mathbf{x}|\theta)$$

- 데이터 집합  $\mathbf{X}$  가 **독립적으로** 추출되었을 경우 로그가능도를 **최적화**합니다

$$L(\theta; \mathbf{X}) = \prod_{i=1}^n P(\mathbf{x}_i | \theta) \quad \Rightarrow \quad \log L(\theta; \mathbf{X}) = \sum_{i=1}^n \log P(\mathbf{x}_i | \theta)$$

×

# 왜 로그가능도를 사용하나요?

---

- 로그가능도를 최적화하는 모수  $\theta$ 는 가능도를 최적화하는 MLE 가 됩니다
- 데이터의 숫자가 적으면 상관없지만 만일 데이터의 숫자가 수억 단위가 된  
다면 컴퓨터의 정확도로는 가능도를 계산하는 것은 불가능합니다
- 데이터가 독립일 경우, 로그를 사용하면 가능도의 곱셈을 로그가능도의 덧  
셈으로 바꿀 수 있기 때문에 컴퓨터로 연산이 가능해집니다
- 경사하강법으로 가능도를 최적화할 때 미분 연산을 사용하게 되는데, 로그  
가능도를 사용하면 연산량을  $O(n^2)$ 에서  $O(n)$  으로 줄여줍니다
- 대개의 손실함수의 경우 경사하강법을 사용하므로 음의 로그가능도  
(negative log-likelihood)를 최적화하게 됩니다

×

# 최대가능도 추정법 예제: 정규분포

---

- 정규분포를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{x_1, \dots, x_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} L(\theta; \mathbf{x}) = \underset{\theta}{\operatorname{argmax}} P(\mathbf{x}|\theta)$$

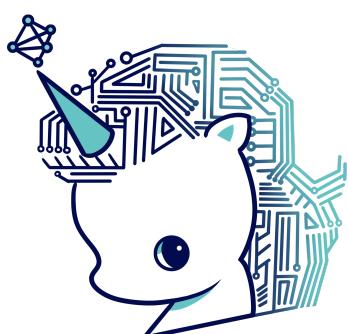
# 최대가능도 추정법 예제: 정규분포

---

- 정규분포를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{x_1, \dots, x_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} L(\theta; \mathbf{x}) = \underset{\mu, \sigma^2}{\operatorname{argmax}} P(\mathbf{X}|\mu, \sigma^2)$$

$$\begin{aligned}\log L(\theta; \mathbf{X}) &= \sum_{i=1}^n \log P(x_i | \theta) = \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|x_i - \mu|^2}{2\sigma^2}} \\ &= -\frac{n}{2} \log 2\pi\sigma^2 - \sum_{i=1}^n \frac{|x_i - \mu|^2}{2\sigma^2}\end{aligned}$$



$\theta = (\mu, \sigma)$ 에 대해 오른쪽 수식을 미분해서 최적화를 할 수 있습니다

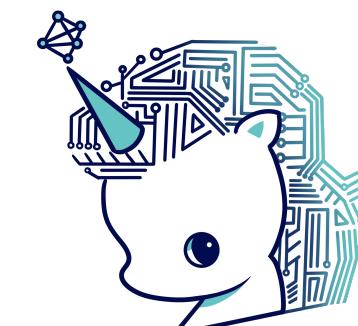
# 최대가능도 추정법 예제: 정규분포

---

- 정규분포를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{x_1, \dots, x_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$0 = \frac{\partial \log L}{\partial \mu} = - \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2}$$

$$0 = \frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n |x_i - \mu|^2$$



두 미분이 모두 0 이 되는  $\mu, \sigma$  를 찾으면 가능도를 최대화하게 된다

# 최대가능도 추정법 예제: 정규분포

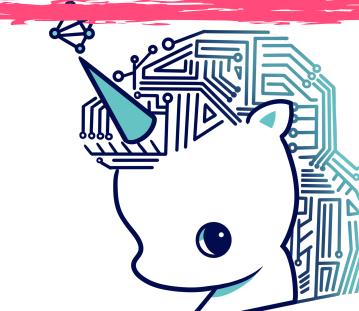
- 정규분포를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{x_1, \dots, x_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$0 = \frac{\partial \log L}{\partial \mu} = - \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2}$$

$$\Rightarrow \hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$0 = \frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n |x_i - \mu|^2$$

$$\Rightarrow \hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$



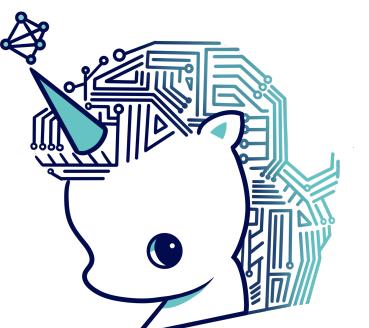
MLE 는 불편추정량을  
보장하진 않습니다

# 최대가능도 추정법 예제: 카테고리 분포

---

- 카테고리 분포  $\text{Multinoulli}(\mathbf{x}; p_1, \dots, p_d)$  를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$\hat{\theta}_{\text{MLE}} = \underset{p_1, \dots, p_d}{\operatorname{argmax}} \log P(\mathbf{x}_i | \theta) = \underset{p_1, \dots, p_d}{\operatorname{argmax}} \log \left( \prod_{i=1}^n \prod_{k=1}^d p_k^{x_{i,k}} \right)$$



카테고리 분포의 모수는 오른쪽  
제약식을 만족해야 합니다

$$\sum_{k=1}^d p_k = 1$$

# 최대가능도 추정법 예제: 카테고리 분포

---

- 카테고리 분포  $\text{Multinoulli}(\mathbf{x}; p_1, \dots, p_d)$  를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$\hat{\theta}_{\text{MLE}} = \underset{p_1, \dots, p_d}{\operatorname{argmax}} \log P(\mathbf{x}_i | \theta) = \underset{p_1, \dots, p_d}{\operatorname{argmax}} \log \left( \prod_{i=1}^n \prod_{k=1}^d p_k^{x_{i,k}} \right)$$

$$\log \left( \prod_{i=1}^n \prod_{k=1}^d p_k^{x_{i,k}} \right) = \sum_{k=1}^d \left( \sum_{i=1}^n x_{i,k} \right) \log p_k$$

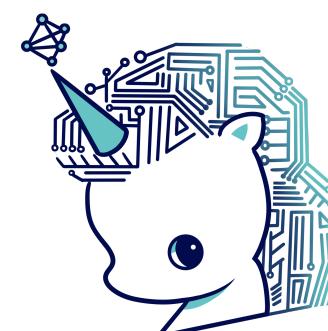
$$n_k = \sum_{i=1}^n x_{i,k}$$

# 최대가능도 추정법 예제: 카테고리 분포

---

- 카테고리 분포  $\text{Multinoulli}(\mathbf{x}; p_1, \dots, p_d)$  를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$\log \left( \prod_{i=1}^n \prod_{k=1}^d p_k^{x_{i,k}} \right) = \sum_{k=1}^d n_k \log p_k \quad \text{with} \quad \sum_{k=1}^d p_k = 1$$



오른쪽 제약식을 만족하면서 왼쪽 목적식을 최대화하는 것이 우리가 구하는 MLE 입니다

# 최대가능도 추정법 예제: 카테고리 분포

---

- 카테고리 분포  $\text{Multinoulli}(\mathbf{x}; p_1, \dots, p_d)$  를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$\log \left( \prod_{i=1}^n \prod_{k=1}^d p_k^{x_{i,k}} \right) = \sum_{k=1}^d n_k \log p_k \quad \text{with} \quad \sum_{k=1}^d p_k = 1$$
$$\Rightarrow \mathcal{L}(p_1, \dots, p_k, \lambda) = \sum_{k=1}^d n_k \log p_k + \lambda \left( 1 - \sum_k p_k \right)$$



라그랑주 승수법을 통해  
최적화 문제를 풀 수 있습니다

# 최대가능도 추정법 예제: 카테고리 분포

---

- 카테고리 분포  $\text{Multinoulli}(\mathbf{x}; p_1, \dots, p_d)$  를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$\Rightarrow \mathcal{L}(p_1, \dots, p_k, \lambda) = \sum_{k=1}^d n_k \log p_k + \lambda(1 - \sum_k p_k)$$

$$0 = \frac{\partial \mathcal{L}}{\partial p_k} = \frac{n_k}{p_k} - \lambda \quad 0 = \frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \sum_{k=1}^d p_k$$

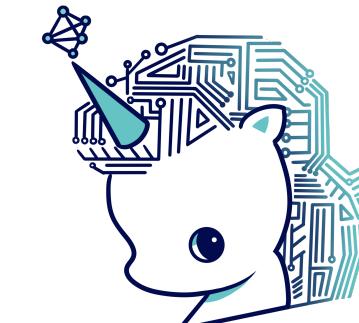
# 최대가능도 추정법 예제: 카테고리 분포

- 카테고리 분포  $\text{Multinoulli}(\mathbf{x}; p_1, \dots, p_d)$  를 따르는 확률변수  $X$ 로부터 독립적인 표본  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$\Rightarrow \mathcal{L}(p_1, \dots, p_k, \lambda) = \sum_{k=1}^d n_k \log p_k + \lambda(1 - \sum_k p_k)$$

$$0 = \frac{\partial \mathcal{L}}{\partial p_k} = \frac{n_k}{p_k} - \lambda \quad 0 = \frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \sum_{k=1}^d p_k$$

$p_k = \frac{n_k}{\sum_{k=1}^d n_k}$



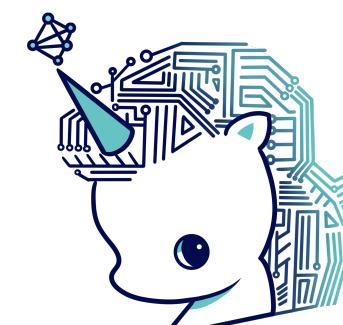
카테고리 분포의 MLE 는 경우의 수  
를 세어서 비율을 구하는 것입니다 22

# 딥러닝에서 최대가능도 추정법

---

- 최대가능도 추정법을 이용해서 기계학습 모델을 학습할 수 있습니다
- 딥러닝 모델의 가중치를  $\theta = (\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L)})$  라 표기했을 때 분류 문제에서 소프트맥스 벡터는 카테고리분포의 모수  $(p_1, \dots, p_K)$ 를 모델링합니다
- 원핫벡터로 표현한 정답레이블  $\mathbf{y} = (y_1, \dots, y_K)$  을 관찰데이터로 이용해 확률분포인 소프트맥스 벡터의 로그가능도를 최적화할 수 있습니다

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K y_{i,k} \log(\text{MLP}_{\theta}(\mathbf{x}_i)_k)$$



위 수식을 잘 기억해두세요

# 확률분포의 거리를 구해보자

---

- 기계학습에서 사용되는 손실함수들은 모델이 학습하는 확률분포와 데이터에서 관찰되는 확률분포의 거리를 통해 유도합니다
- 데이터공간에 두 개의 확률분포  $P(\mathbf{x}), Q(\mathbf{x})$  가 있을 경우 두 확률분포 사이의 거리(distance)를 계산할 때 다음과 같은 함수들을 이용합니다
  - 총변동 거리 (Total Variation Distance, TV)
  - 쿨백-라이블러 발산 (Kullback-Leibler Divergence, KL)
  - 바슈타인 거리 (Wasserstein Distance)

# 쿨백-라이블러 발산

- 쿨백-라이블러 발산(KL Divergence)은 다음과 같이 정의합니다

$$\text{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right)$$

$$\text{KL}(P||Q) = \int_X P(\mathbf{x}) \log \left( \frac{P(\mathbf{x})}{Q(\mathbf{x})} \right) d\mathbf{x}$$

- 쿨백 라이블러는 다음과 같이 분해할 수 있습니다

- 분류 문제에서 정답레이블을  $P$ , 모델 예측을  $Q$  라 두면 최대가능도 추정법은 쿨백-라이블러 발산을 최소화하는 것과 같습니다



# THE END

---

다음 시간에 보아요!