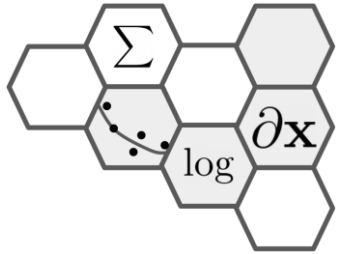
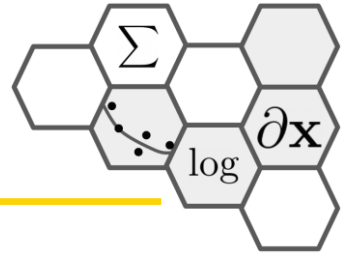


인공지능 이산수학

데이터 정리와 확률

조준우
metamath@gmail.com

1차원 데이터



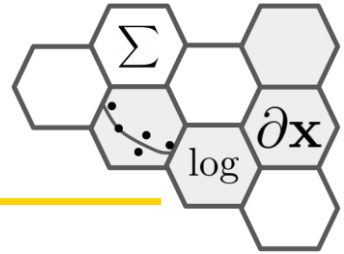
- 시험 점수

[45, 26, 57, 67, 40, 30, 55, 60, 95]

→ 획득된 샘플(표본)

샘플은 숫자 하나로 구성

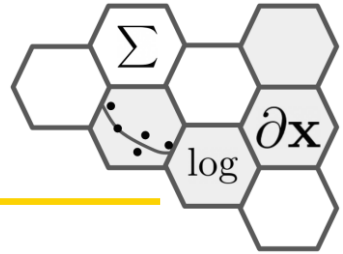
중심의 지표: 평균



- 평균mean
  - 데이터를 모두 더한 후 개수로 나눈 대푯값
  - 가장 많이 사용
  - 이상치에 민감

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_1 + x_2 + \cdots + x_n$$

중심의 지표: 중앙값



- 중앙값median
  - 데이터를 크기 순서로 나열한 후 가장 가운데 있는 값

45, 26, 57, 67, 40, 30, 55, 60, 95 → 정렬

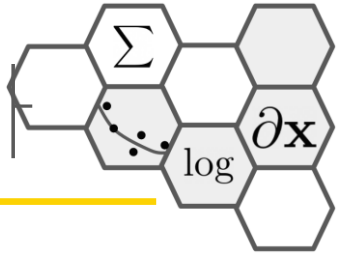
26, 30, 40, 45, 55, 57, 60, 67, 95

45, 26, 57, 67, 40, 30, 55, 60, 95, 500 → 정렬

26, 30, 40, 45, 55, 57, 60, 67, 95, 500

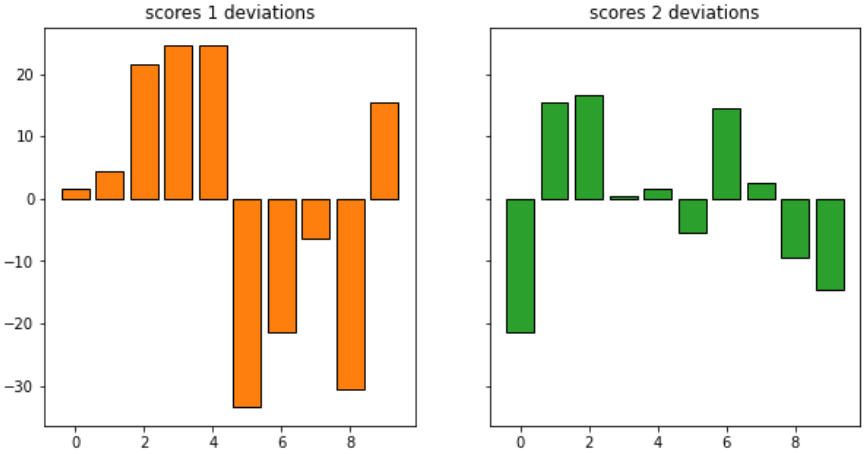
평균

퍼짐의 지표: 분산과 표준편차

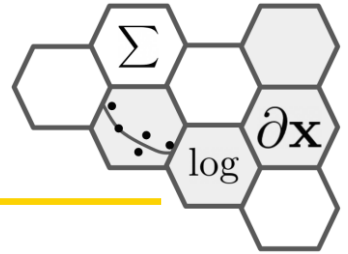


- 분산variance
  - 평균으로 부터 퍼짐의 정도를 숫자로 표현

scores1 = [74 77 94 97 97 39 51 66 42 88] mean=72.5
scores2 = [51 88 89 73 74 67 87 75 63 58] mean=72.5



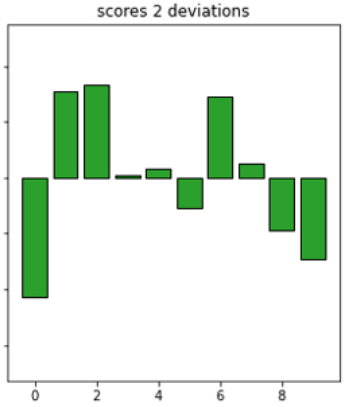
분산의 그림 표현

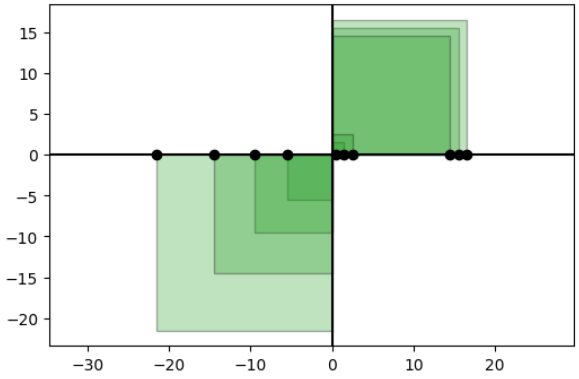


- 분산variance
  - 평균으로 부터 퍼짐의 정도를 한변으로 하는 사각형의 평균 넓이

분산 $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

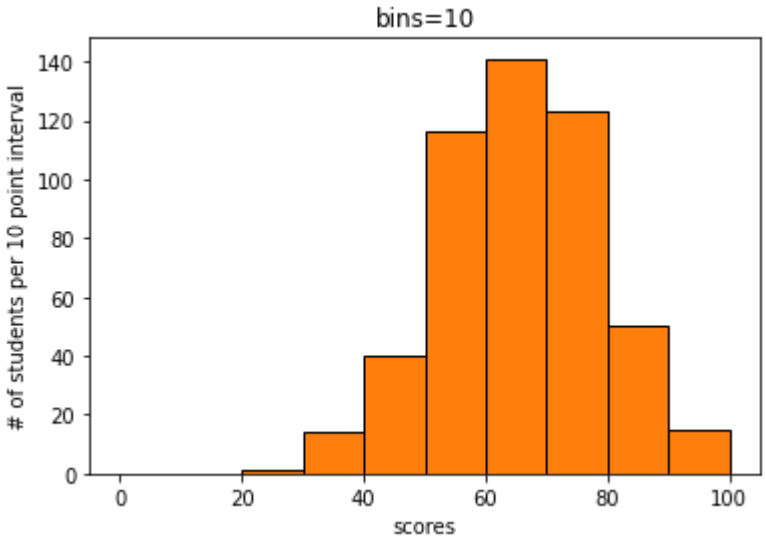
표준편차standard deviation $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

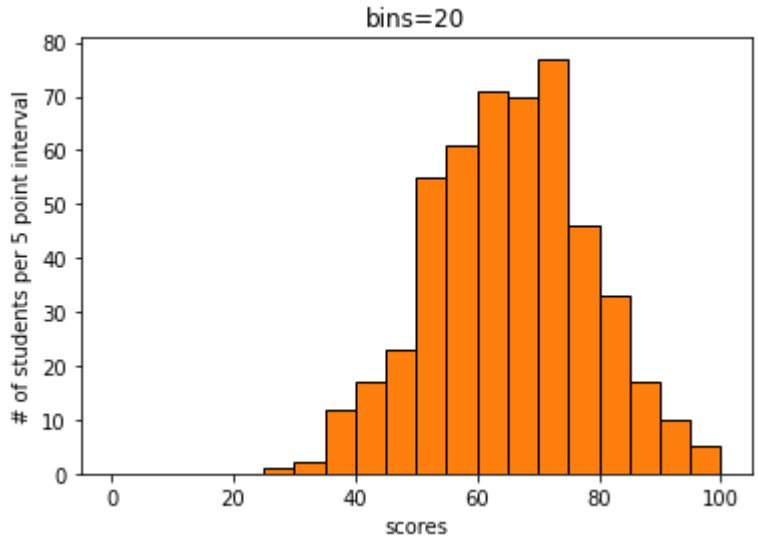




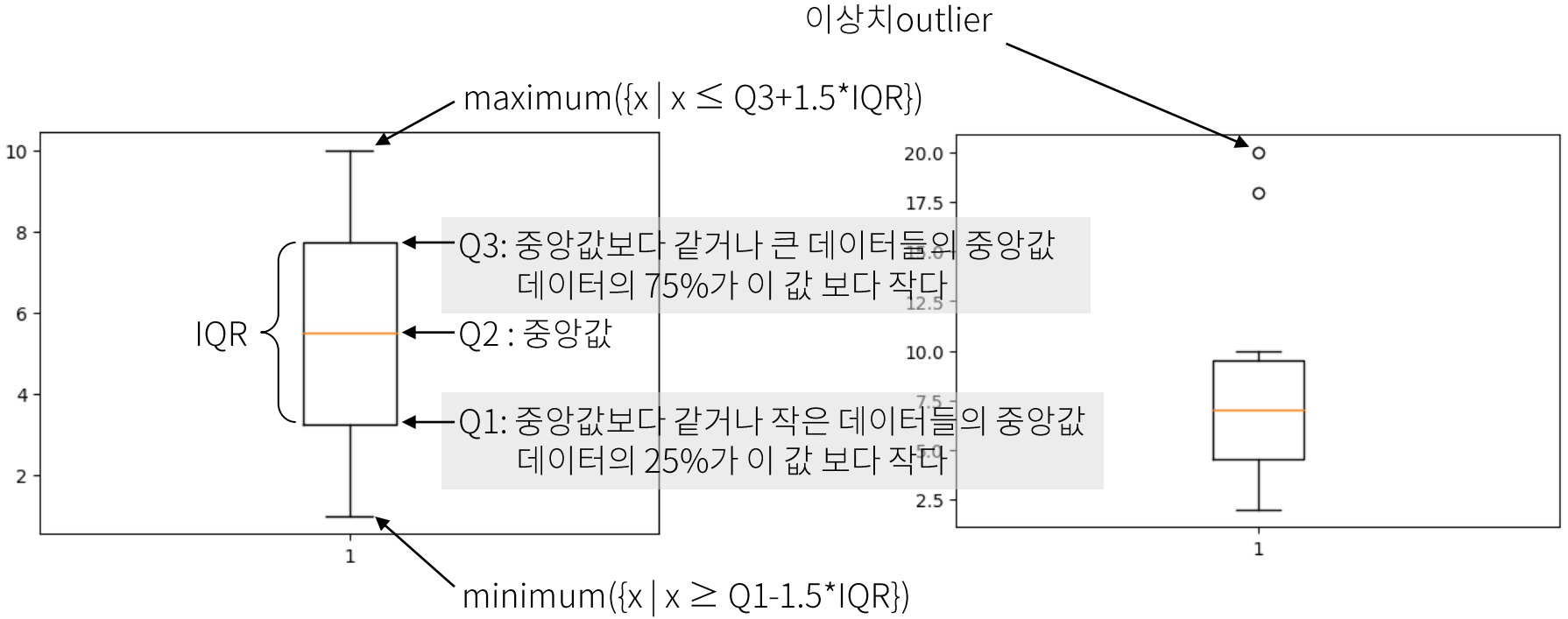
히스토그램

- 데이터를 계급으로 나눠 계급에 해당하는 빈도수(도수)를 막대그래프로 그린 그래프

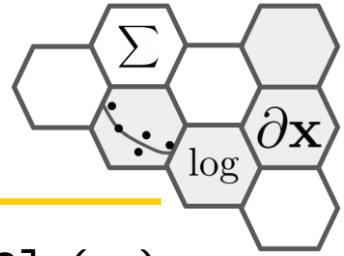




상자그림Box Plot



2차원 데이터



- 키 데이터: [170, 155, 175, 182, 171, 188, 165, 167, 175, 183] (cm)
- 몸무게: [65, 59, 68, 78, 62, 85, 63, 58, 70, 98] (kg)
- 신체 지수

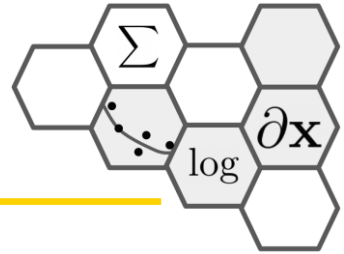
```
X=[[170, 155, 175, 182, 171, 188, 165, 167, 175, 183]
   [ 65,  59,  68,  78,  62,  85,  63,  58,  70,  98]]
```

샘플이 숫자 두 개로 구성

키 데이터, 몸무게 데이터를 열 방향으로

```
X=[[170, 65
    155, 59
    175, 68
    182, 78
    171, 62
    188, 85
    165, 63
    167, 58
    175, 70
    183, 98]]
```

평균과 분산

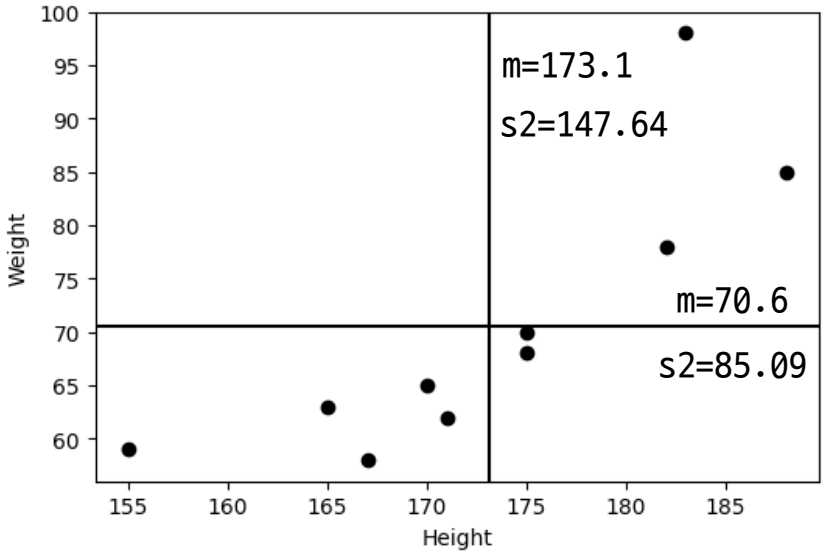



```
X=[[170, 65
   155, 59
   175, 68
   182, 78
   171, 62
   188, 85
   165, 63
   167, 58
   175, 70
   183, 98]]
```

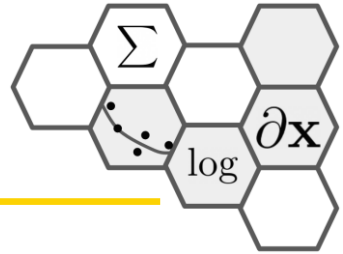
키와 몸무게의 평균?

키 평균 몸무게 평균

키 분산 몸무게 분산



공분산

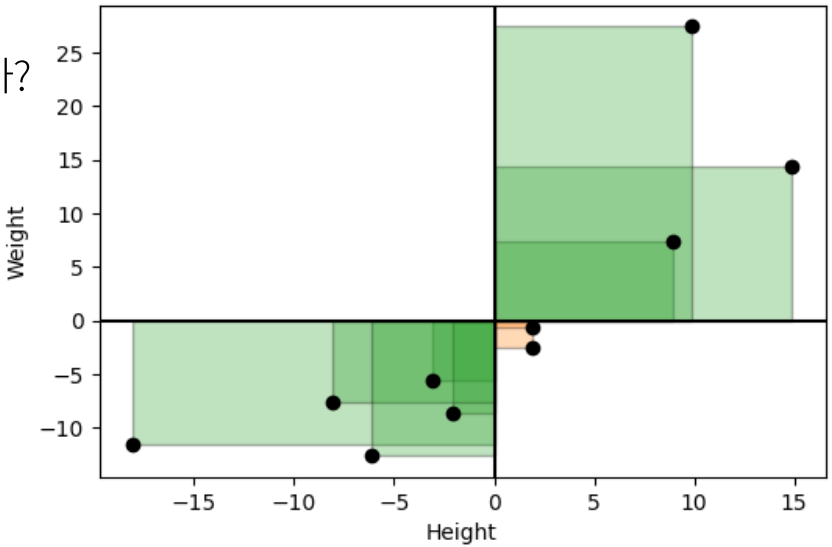


- 공분산covariance
  - 두 평균으로 부터 퍼짐의 정도를 한변으로 하는 사각형의 평균 넓이
  - 양수: 두 데이터는 양의 직선관계, X 증가 → Y 증가, X 감소 → Y 감소
  - 음수: 두 데이터는 음의 직선관계, X 증가 → Y 감소, X 감소 → Y 증가
  - 거의 0: 직선의 관계가 없음
  - 공분산이 크면 직선의 관계가 더 강한가?

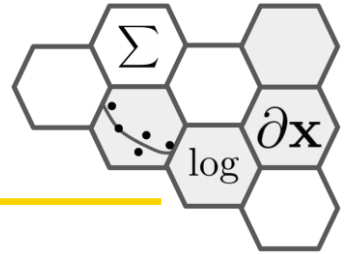
cov_corr.gsheet

공분산 $s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

분산 $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$



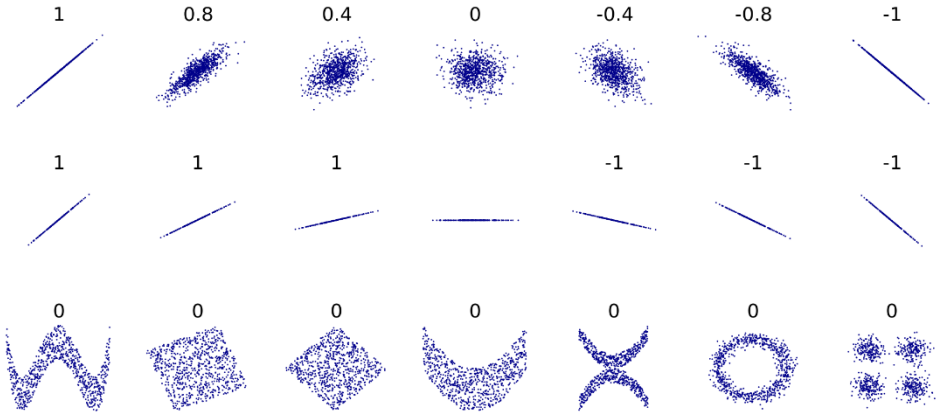
상관계수



- 상관계수correlation coefficient
  - 두 변수의 단위에 관계없이 상관성을 나타내는 지표
  - 1에 가까우면: 두 데이터는 양의 상관관계, X 증가 → Y 증가, X 감소 → Y 감소
  - -1에 가까우면: 두 데이터는 음의 상관관계, X 증가 → Y 감소, X 감소 → Y 증가
  - 거의 0: 직선의 관계가 없음

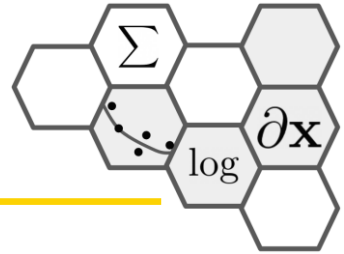
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

$$-1 \le r_{XY} \le 1$$



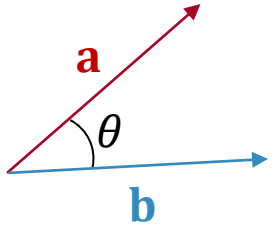
https://en.wikipedia.org/wiki/Correlation, CC0, DenisBoigelot

상관계수



- 두 샘플 편차 벡터 사이각의 코사인 값

```
X=[[170, 65          Xd=[[-3.1,  -5.6
    155, 59               -18.1, -11.6
    175, 68                 1.9,  -2.6
    182, 78                 8.9,   7.4
    171, 62   −X̄ =         -2.1,  -8.6
    188, 85                14.9,  14.4
    165, 63                -8.1,  -7.6
    167, 58                -6.1, -12.6
    175, 70                 1.9,  -0.6
    183, 98]]               9.9,  27.4]]
```



$$-1 \le r_{XY} \le 1$$

$$\| $$

$$-1 \le \cos\theta \le 1$$

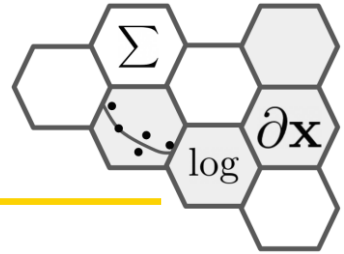
$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta$$

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

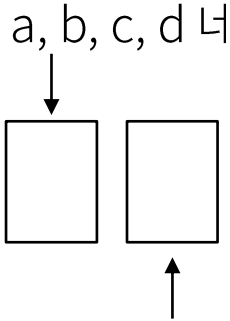
$$\cos\theta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

cov_corr.gsheet

순열



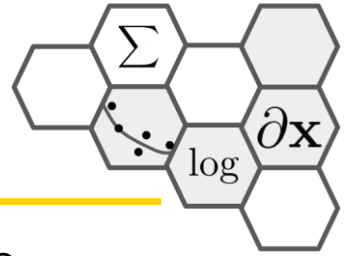
- 줄 세우기
  - a, b, c, d에서 두 개를 골라 줄 세우기
  - (a,b),(a,c),(a,d)
  - (b,a),(b,c),(b,d)
  - (c,a),(c,b),(c,d)
  - (d,a),(d,b),(d,c)



$$4 \times 3 = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} = \frac{4!}{2!} = \frac{4!}{(4-2)!}$$

$$P(n,r) = \frac{n!}{(n-r)!} \qquad n\text{: 음이 아닌 정수}, r\text{: } 0 \le r \le n$$

순열

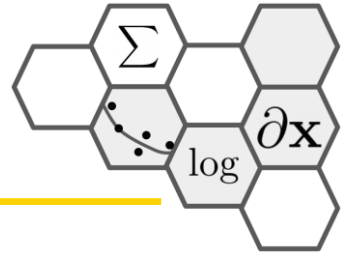


- ABCDEFG의 순열 중 AB 문자열이 포함된 것은 모두 몇 개?

  ABCDEFG

  문자 하나로 보면 여섯 개 문자를 배열하는 것으로 6!

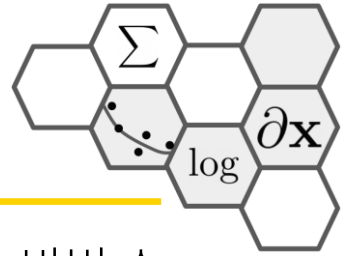
조합



- 선택 하기
  - a, b, c, d에서 두 개를 선택하기
  - (a,b),(a,c),(a,d)
  - (b,a),(b,c),(b,d)
  - (c,a),(c,b),(c,d)
  - (d,a),(d,b),(d,c)
  - 순열에서 같은 요소가 있는 선택들이 하나로 줄어 버림
  - 2!개가 한 개로!

$$\binom{n}{r} = \frac{n!}{(n-r)!} \div r! = \frac{n!}{r!\,(n-r)!} \quad n\text{: 음이 아닌 정수, } r\text{: } 0 \le r \le n$$

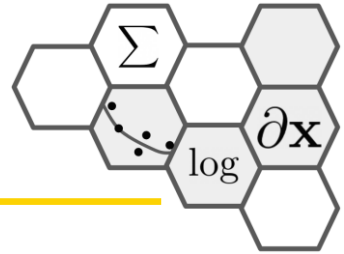
조합



- 남학생 5명, 여학생 7명이 지원한 선발에서 남학생 3명, 여학생 3명 뽑는 방법 수

$$\binom{5}{3} \times \binom{7}{3} = \frac{5!}{(5-3)!\,3!} \times \frac{7!}{(7-3)!\,3!} = 350$$

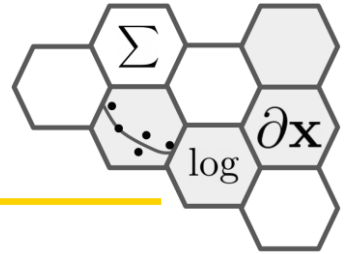
확률의 정의



- 수학적 확률
  - 어떤 시행에서 사건 A 가 일어날 가능성을 수로 나타낸 것: $P(A)$
  - 표본공간 Ω에서 사건 A 가 일어날 수학적 확률
    - 표본공간 Ω인 어떤 시행에서 각 결과가 일어날 가능성이 모두 같은 정도로 기대될 때
    - $P(A) = \frac{n(A)}{n(\Omega)}$

- 통계적 확률
  - 일어날 가능성이 같은 정도로 기대될 수 없을 때
  - 같은 시행을 n 번 반복할 때 사건 A 가 일어난 횟수를 r_n
  - 시행 횟수 n 이 한없이 커짐에 따라 $\frac{r_n}{n}$이 일정한 값 P 에 가까워 지면
  - P 는 사건 A 의 통계적 확률

확률은 면적



- 동전을 던지는 표본공간에서

Ω

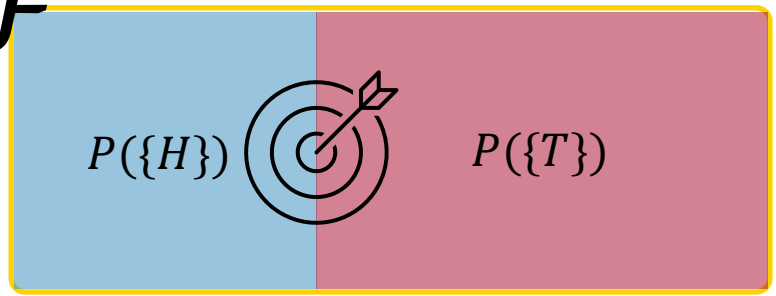
H T

$\mathcal{F}$

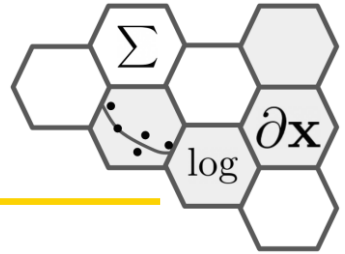
$P(\{\}) = 0$ $P(\Omega) = 1$

$P(\{H\}) = 0.3$ $P(\{T\}) = 0.7$

$\mathcal{F}$



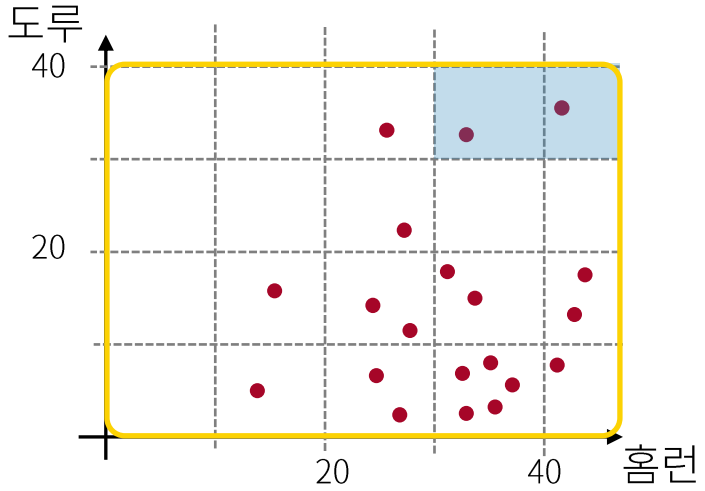
조건부 확률



- 두 사건 A, B에 대해
- 결합확률Joint Probability: 두 사건이 동시에 일어날 확률 $P(A, B)$
- 조건부확률Conditional Probability: 사건 A가 일어났을 때 사건B가 일어날 확률 $P(B|A)$

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

A: 홈런 30개 이상 12번 $P(A) = 12/20$
B: 도루 30개 이상 3번
$A \cap B$: 30,30 이상 2번 $P(A,B) = 2/20$



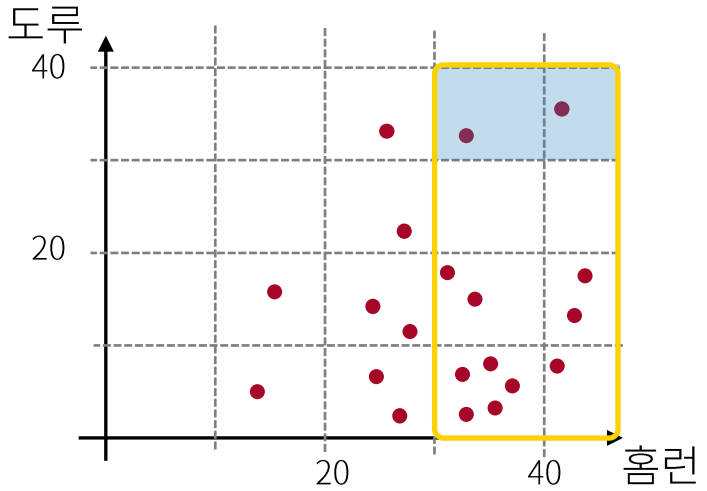
조건부 확률

- 두 사건 A, B에 대해
- 결합확률Joint Probability: 두 사건이 동시에 일어날 확률 $P(A, B)$
- 조건부확률Conditional Probability: 사건 A가 일어났을 때 사건B가 일어날 확률 $P(B|A)$

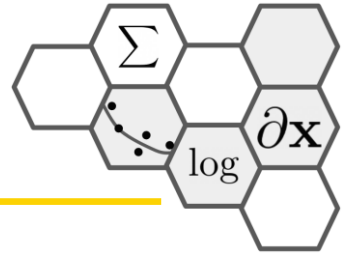
$$P(B|A) = \frac{P(A,B)}{P(A)}$$

A: 홈런 30개 이상 12번 $P(A) = 12/20$
B: 도루 30개 이상 3번
$A \cap B$: 30,30 이상 2번 $P(A,B) = 2/20$

홈런 30개 이상 쳤을 때 도루 30개 이상할 확률
$P(B|A) = 2/12$, 분자 분모를 20으로 나누면



조건부 확률

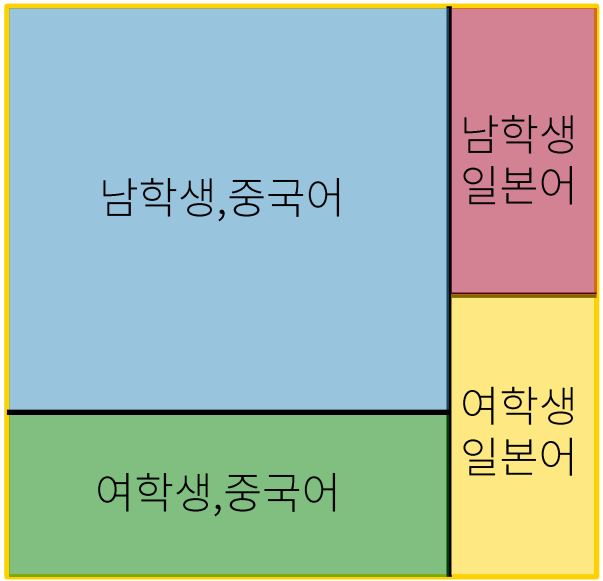


- P(남학생,중국어)=45/100
- P(남학생|중국어)=45/70

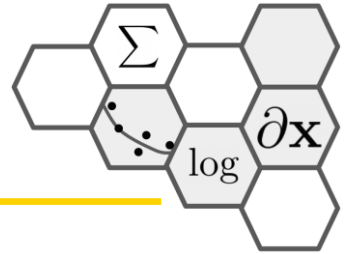
	중국어	일본어
남학생 수	45	15
여학생 수	25	15

조건부 확률

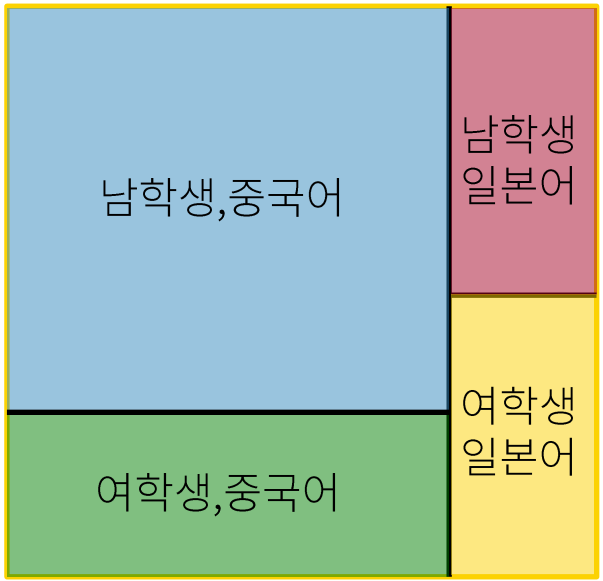
- P(남학생, 중국어)=45/100

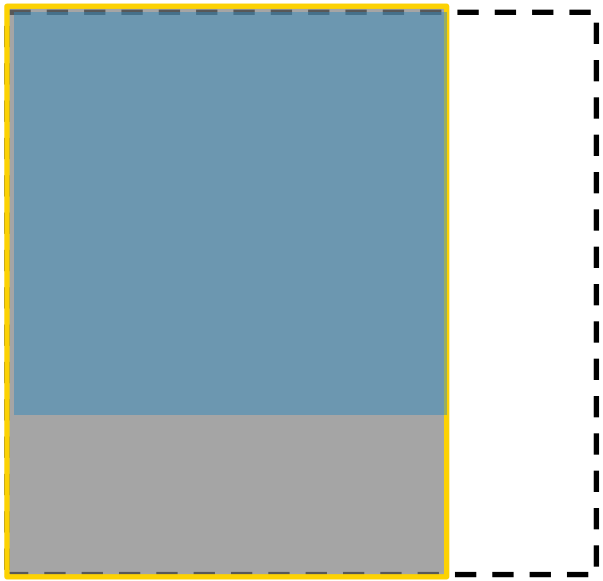


조건부 확률

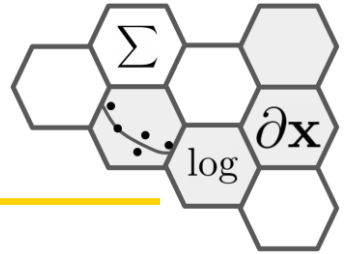


- P(남학생|중국어)=45/70





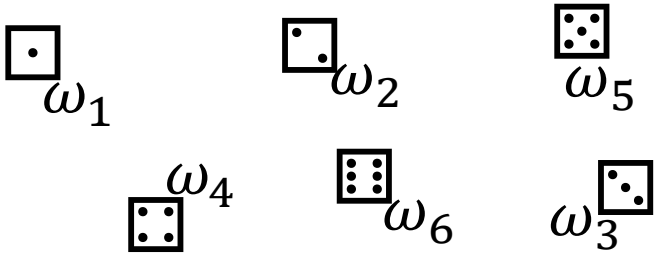
확률변수: 이산 확률변수



- 표본공간의 샘플에 숫자를 할당하는 함수
- 이산 확률변수 $X: \Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\} \rightarrow \mathbb{R}$

Ω

우리가 관심있는 표본sample들이 모여 있는 공간



$X(\boxdot) = 1$ $\quad$ $Y(\boxdot) = 1$ $\quad$ $Z(\boxdot) = 0$

$X(\boxdot) = 2$ $\quad$ $Y(\boxdot) = 0$ $\quad$ $Z(\boxdot) = 0$

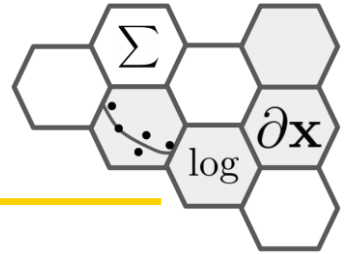
$X(\boxdot) = 3$ $\quad$ $Y(\boxdot) = 1$ $\quad$ $Z(\boxdot) = 1$

$X(\boxdot) = 4$ $\quad$ $Y(\boxdot) = 0$ $\quad$ $Z(\boxdot) = 0$

$X(\boxdot) = 5$ $\quad$ $Y(\boxdot) = 1$ $\quad$ $Z(\boxdot) = 0$

$X(\boxdot) = 6$ $\quad$ $Y(\boxdot) = 0$ $\quad$ $Z(\boxdot) = 1$

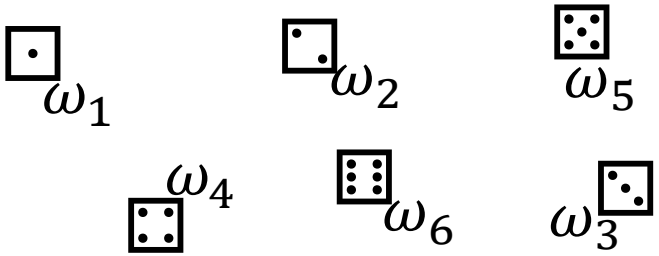
확률분포



- 확률변수 X가 가질 수 있는 값과 확률의 대응 관계
  - 다시 말해 X가 가질 수 있는 값에 확률이 얼마나 할당되었는가를 나타낸 확률의 펼쳐짐 정도
  - $P(X = x)$: 확률변수 X가 x값을 가질 확률
- 이산 확률변수

Ω

우리가 관심있는 표본sample들이 모여 있는 공간



$X(\boxdot) = 1$

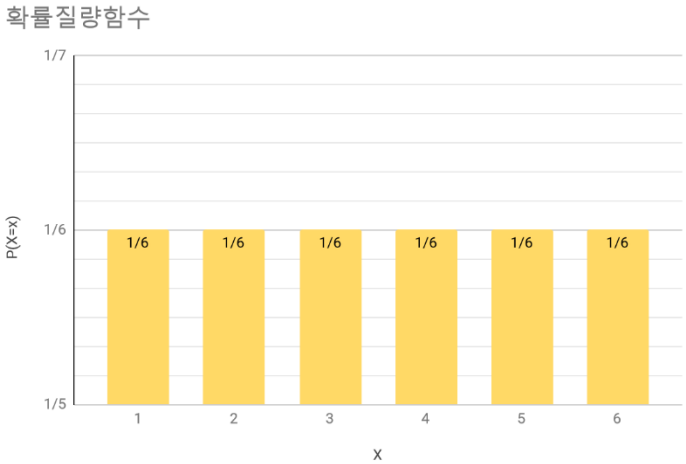
$X(\boxdot) = 2$

$X(\boxdot) = 3$

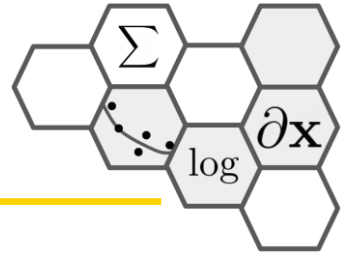
$X(\boxdot) = 4$

$X(\boxdot) = 5$

$X(\boxdot) = 6$



확률 질량함수



- 이산 확률변수의 다른 예

Ω

우리가 관심있는 표본sample들이 모여 있는 공간

ω_1 ω_2 ω_5 ω_4 ω_6 ω_3

$Z(\boxdot) = 0$

$Z(\boxdot) = 0$

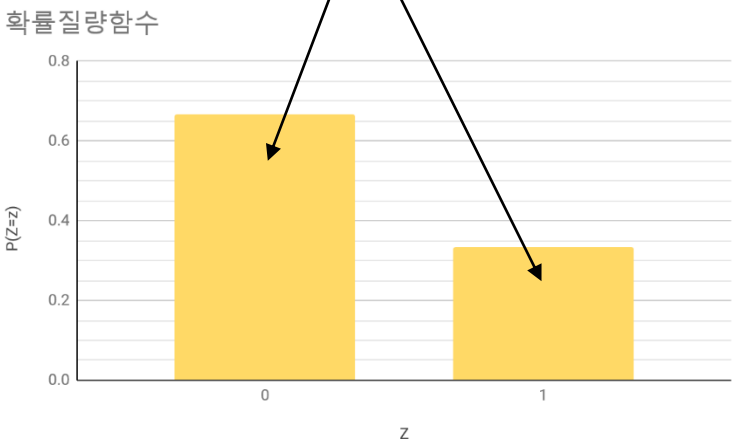
$Z(\boxdot) = 1$

$Z(\boxdot) = 0$

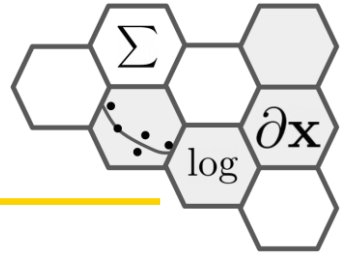
$Z(\boxdot) = 0$

$Z(\boxdot) = 1$

확률이 특정값에 덩어리처럼 몰려 있다. (질량)



기댓값Expected Value



- 평균
  - N개를 다 더해서 N으로 나눔

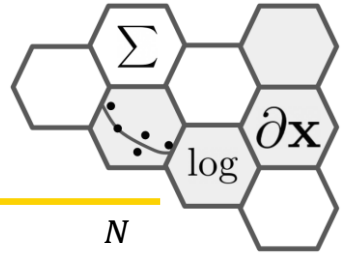
상금(원)	0	10,000	30,000	50,000	합계
행운권 수	872	100	20	8	1,000

  - 행운권 한 장 당 평균 상금: 총 상금/행운권 수

$$\frac{0 \times 872 + 10000 \times 100 + 30000 \times 20 + 50000 \times 8}{1000} = 2000$$

$$0 \times \frac{872}{1000} + 10000 \times \frac{100}{1000} + 30000 \times \frac{20}{1000} + 50000 \times \frac{8}{1000} = 2000$$

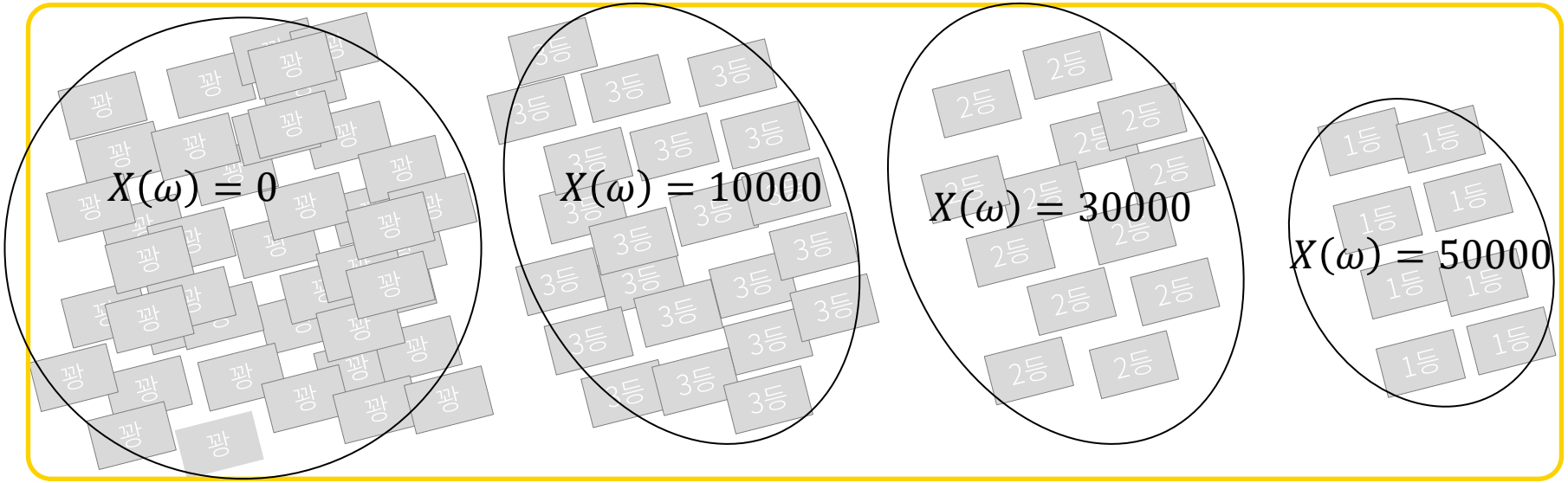
확률변수의 기댓값



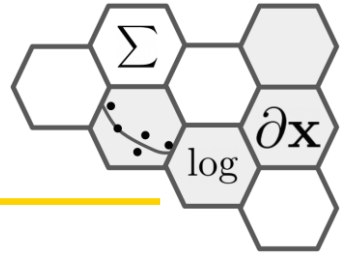
- 평균
  - 행운권 한 장 당 평균 상금: 총 상금/행운권 수

$$E[X] = \sum_{i=1}^{N} x_i p_i$$

$$0 \times \frac{872}{1000} + 10000 \times \frac{100}{1000} + 30000 \times \frac{20}{1000} + 50000 \times \frac{8}{1000} = 2000$$



확률변수의 기댓값, 분산, 표준편차



- 기댓값(평균)

$$E[X] = \sum_{i=1}^{N} x_i p_i$$

- 분산variance: 편차 제곱의 평균

$$Var[X] = E[(X - E[X])^2] = \sum_{i=1}^{N} (x_i - E[X])^2 p_i$$

- 표준편차standard deviation

$$\sigma(X) = \sqrt{Var[X]}$$

분산: 제곱의 평균 – 평균의 제곱

$$\begin{aligned} Var[X] &= \sum_{i=1}^{N} (x_i - E[X])^2 p_i \\ &= \sum_{i=1}^{N} (x_i^2 - 2E[X]x_i + E[X]^2) p_i \\ &= \sum_{i=1}^{N} x_i^2 p_i - 2E[X] \sum_{i=1}^{N} x_i p_i + \mathbb{E}[X]^2 \sum_{i=1}^{N} p_i \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

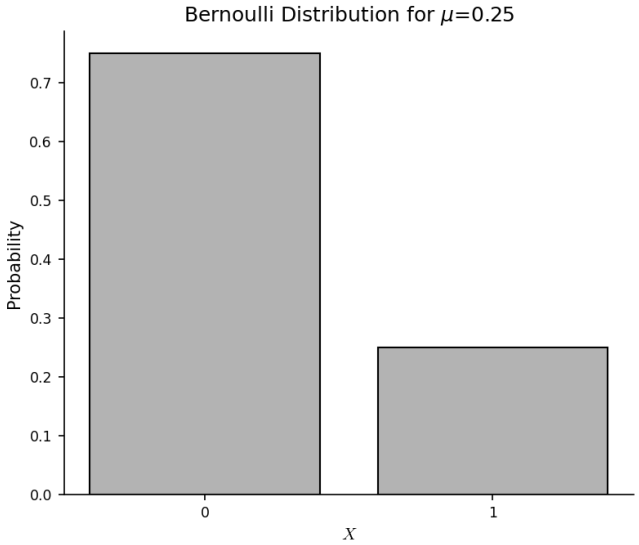
이산확률분포: 베르누이 분포

- 베르누이 분포Bernoulli distribution
  - 0 또는 1을 값으로 가지는 바이너리binary 확률변수 X에 대한 분포. 즉 $x \in \{0, 1\}$
  - 예: 동전 던지기, 동전의 앞면(Head)이 나오면 $x = 1$, 뒷면(Tail)이 나오면 $x = 0$

- 확률질량함수
  - $x \in \{0, 1\}$이므로 $x = 1$일 확률은
    $p(x = 1|\mu) = \mu$, $x = 0$일 확률은 $p(x = 0|\mu) = 1 - \mu$

$$f_X(x|\mu) = \mu^x(1 - \mu)^{1-x}$$

바이너리 확률변수에 값을 할당하게 하는 행위
- 베르누이 시행bernoulli trial,bernoulli experiment



이산확률분포: 멀티누이 분포

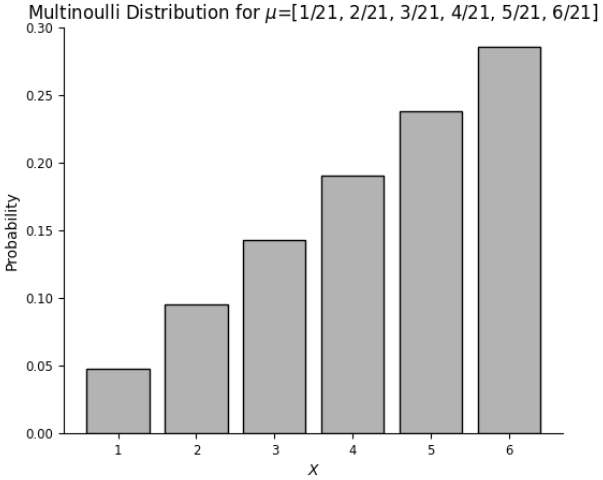
- 멀티누이Multinoulli distribution, Categorical distribution
  - 다항변수multinomial variables:이진 변수의 일반화로 이진 변수가 여러 개가 모인 벡터 변수 X에 대한 확률분포
  - 예: 주사위 던지기, ⚀: $\mathbf{x} = (1,0,0,0,0,0)^{\mathrm{T}}$, ⚁: $\mathbf{x} = (0,1,0,0,0,0)^{\mathrm{T}}$, ⋯
  - 제약조건: $\sum_{k=1}^{K} x_k = 1$
- 확률질량함수

$$f_X(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

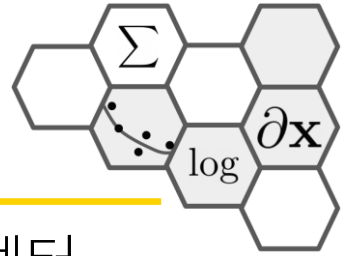
제약조건

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^T \quad \mu_k \geq 0 \quad \sum_k \mu_k = 1$$

  - 파라미터도 벡터
  - k번째 자리가 1인 벡터변수에 할당된 확률은 k번째 자리의 μ_k



확률벡터



- 다변량 확률변수: 확률변수 여러 개 모임, 숫자의 모임 → 벡터

확률변수 X 온도 — 추출된 샘플 벡터 $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

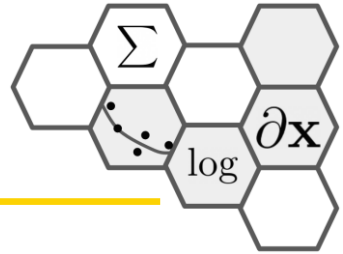
확률변수 Y 습도 — 추출된 샘플벡터 $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

확률변수 $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$ 날씨

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$$

추출된 샘플→행렬

결합확률분포



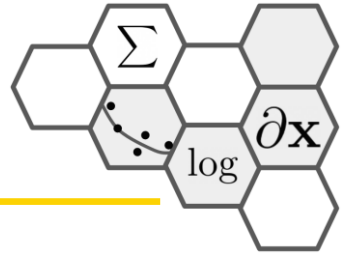
- 결합확률분포joint probability distribution
  - 확률변수 X, Y를 동시에 고려한 확률분포
  - 확률 질량함수는 이변수 스칼라 함수이며 0 이상의 값을 함숫값으로 가지며 모두 더해서 1이 됨

확률변수 $XY = \begin{bmatrix} X \\ Y \end{bmatrix}$

샘플 $xy = \{(x_i, y_j) | i = 1,2, \dots ; j = 1,2, \dots\}$

결합확률질량함수 $f_{XY}(XY) = P(X = x_i, Y = y_j)$

결합확률분포 예



- 결합확률분포joint probability distribution

$f_{XY}(x,y)$		X			$f_Y(y)$
		5	6	7	
Y	8	0	0.4	0.1	0.5
	9	0.3	0	0.2	0.5
$f_X(x)$		0.3	0.4	0.3	1

https://en.wikipedia.org/wiki/Covariance

$(5, 8)$ $f_{XY}(5,8) = 0$

$(5, 9)$ $f_{XY}(5,9) = 0.3$

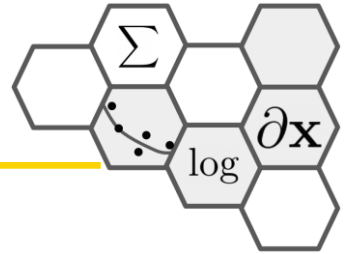
$(6, 8)$ $f_{XY}(6,8) = 0.4$

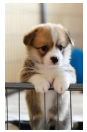
$(6, 9)$ $f_{XY}(6,9) = 0$

$(7, 8)$ $f_{XY}(7,8) = 0.1$

$(7, 9)$ $f_{XY}(7,9) = 0.2$

👀: 베르누이 분포추정

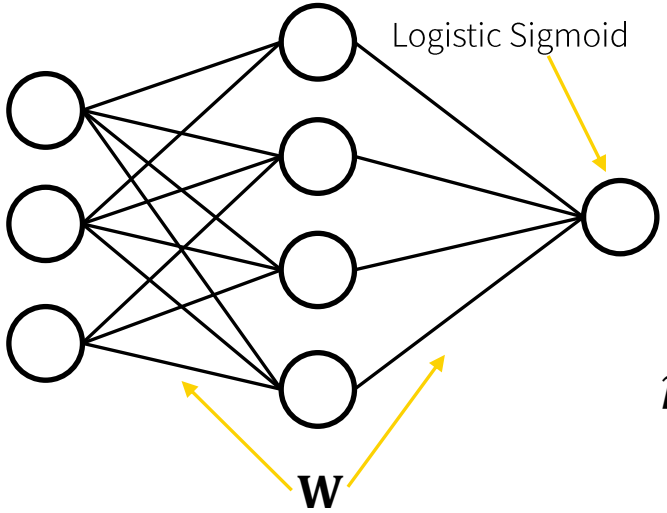


$P(x = $  $, y = 1) = ?$

$P(x = $  $, y = 0) = ?$

$y = 1$

$\mathbf{x}$



Bernoulli distribution

0.3

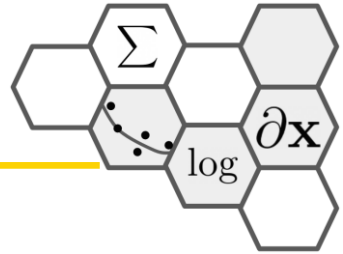
$p(y|\mu, \mathbf{x}; \mathbf{W})$

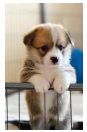
강아지일 확률

1

$p(\mathbf{y}|\mathbf{x} =$  $)$

👀: 멀티누이 분포추정

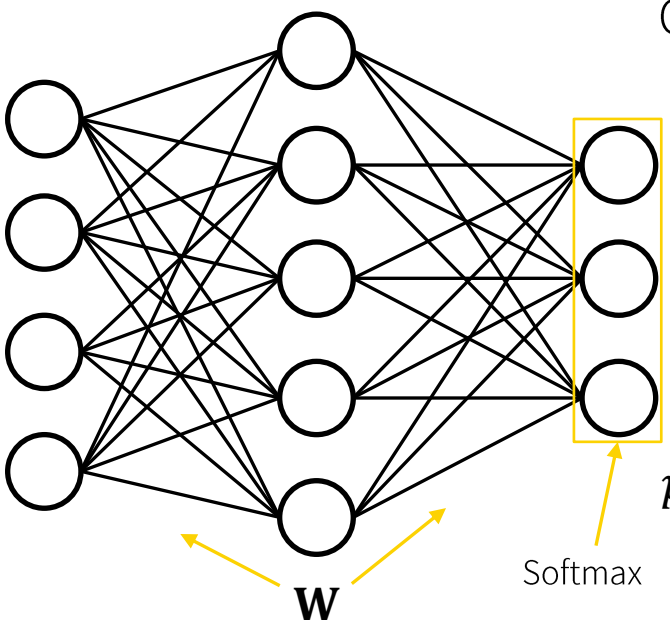


$P(x = $  $, \mathbf{y} = (0,1,0)) = ?$

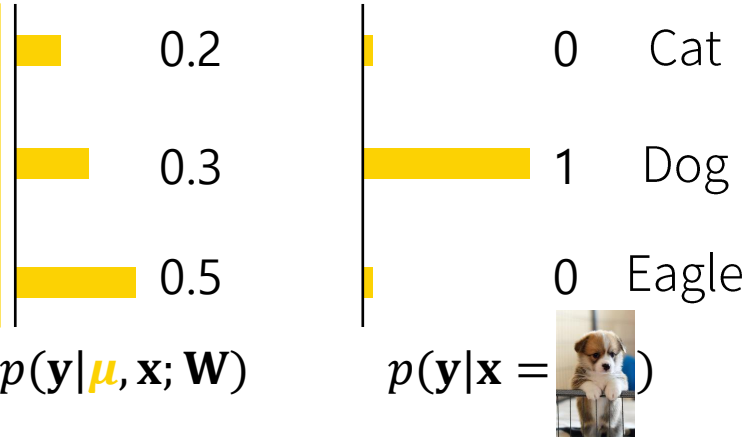
$P(x = $  $, \mathbf{y} = (0,1,0)) = ?$

$\mathbf{y} = (0,1,0)^T$

$\mathbf{x}$



Categorical distribution



$p(\mathbf{y}|\boldsymbol{\mu}, \mathbf{x}; \mathbf{W})$

$p(\mathbf{y}|\mathbf{x} = $ $)$