확률 통계 기초

고려대학교
전기전자공학부
석준희 교수

목차

빅데이터, 인공지능 그리고 확률통계

4차 산업 혁명

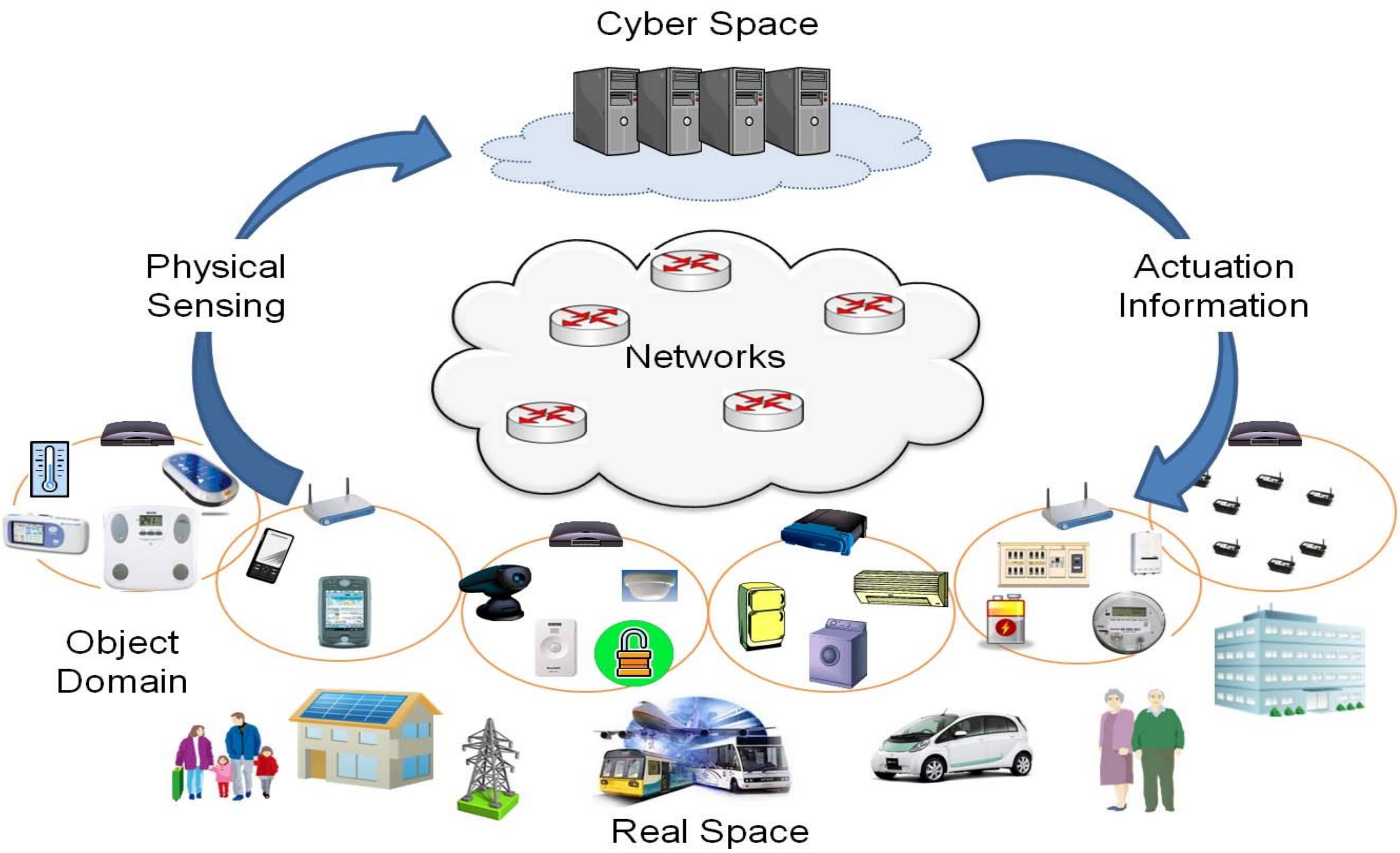
- 새로운 과학기술의 도입을 통한 생산성(=부)의 급속한 증가

4차산업혁명의 과학/기술

- **사이버 물리 시스템 (CPS: Cyber Physical Systems)**
  - 가상세계(컴퓨터)가 현실세계와 융합하여 다양한 서비스를 제공
- **인공지능 (AI: Artificial Intelligence)**
  - 기계가 스스로 판단하고 인식할 수 있는 능력
- **사물인터넷 (IoT: Internet of Things)**
  - 개별 사물이 인터넷에 연결되어 필요한 정보를 주고 받는 환경
- **빅데이터 (Big Data)**
  - **이전에 없었던 대규모로 축적된 데이터의 관리와 활용**
- **클라우드 컴퓨팅 (Cloud Computing)**
  - 데이터와 데이터 처리 능력이 중앙에 집중되어 있는 형태
- **가상현실/증강현실 (VR/AR: Virtual Reality/Augmented Reality)**
  - 가상세계에서의 경험을 현실세계에서 제공하거나, 현실세계의 정보에 가상세계의 정보를 덧붙여 제공
- **로봇/자율주행(Robot/Autonomous Car)**
  - 인간이 수행하던 작업을 대신 수행하는 기계

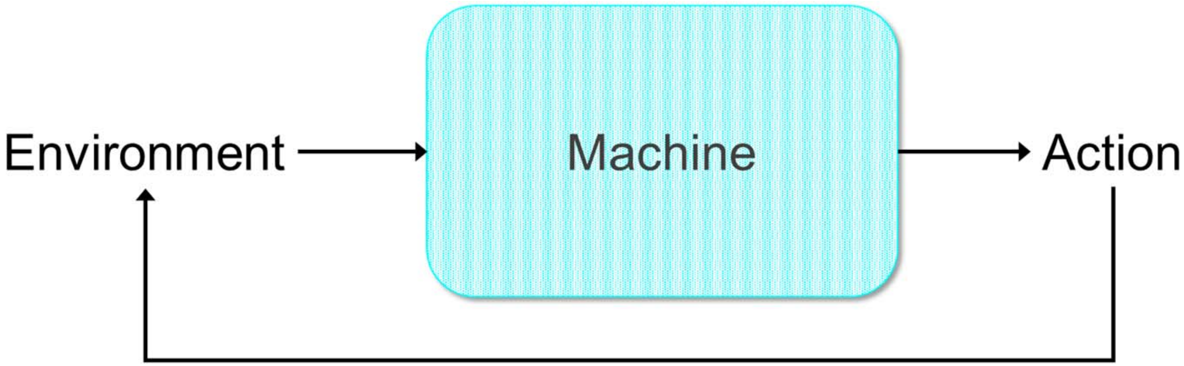
4차산업혁명의 과학/기술

- **사이버 물리 시스템 (CPS: Cyber Physical Systems)**



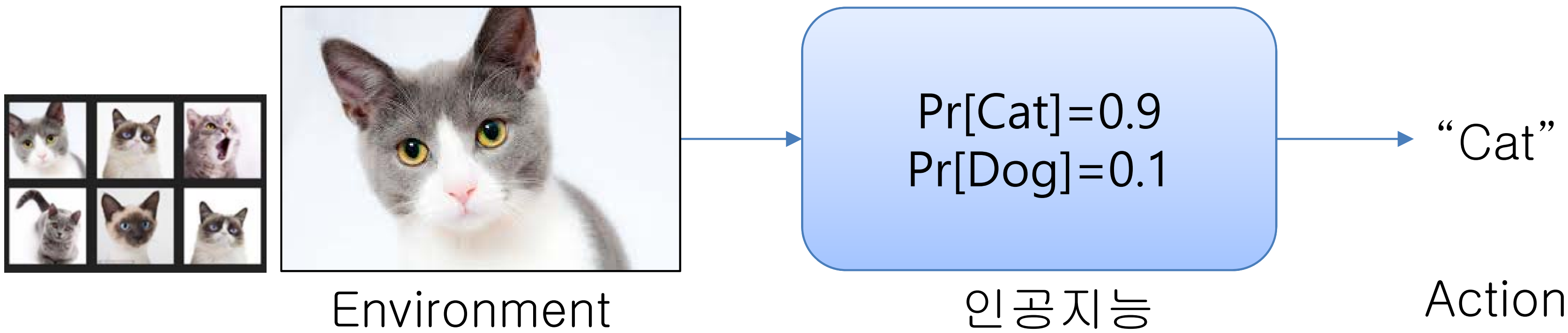
인공지능에서의 확률

- **인공지능에서의 확률론적 모델링**
  - 인공지능은 “관측된 데이터”를 통해 “인지하고 결정하여 “행동”을 함



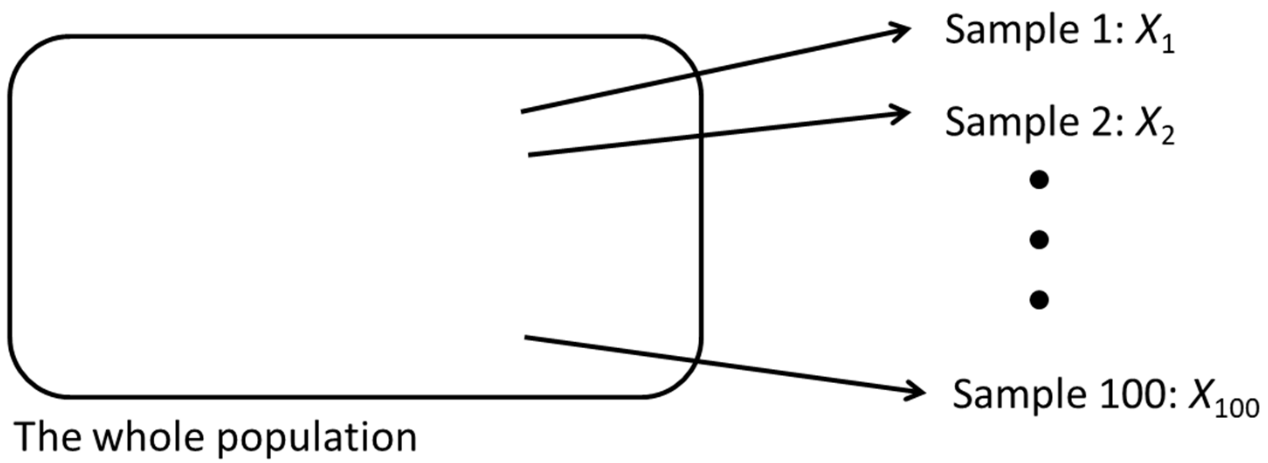
$$\Pr[Y|X]$$

  - Environment: 특정 분포로부터 추출된 데이터(예: 날씨, 음성, …)
  - Action: 주로 확률론적인 결정



임의의 관측된 데이터

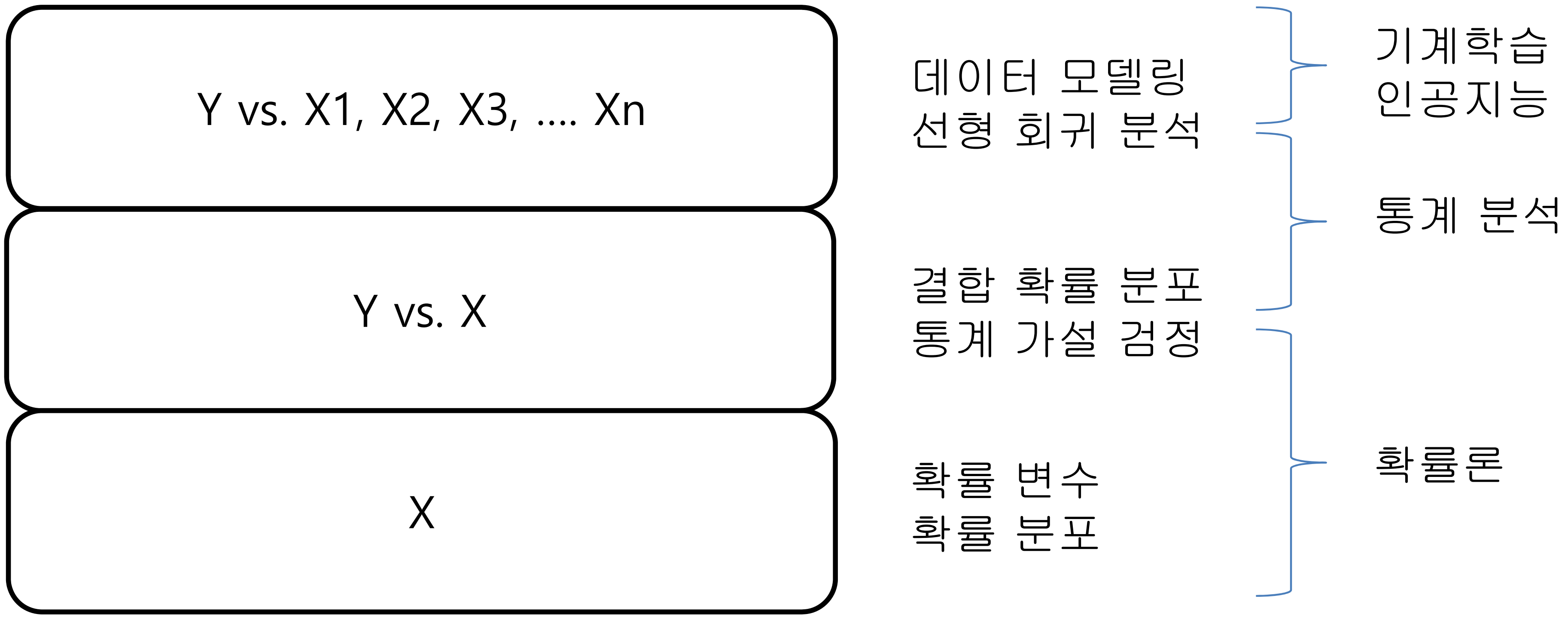
- **데이터란?**



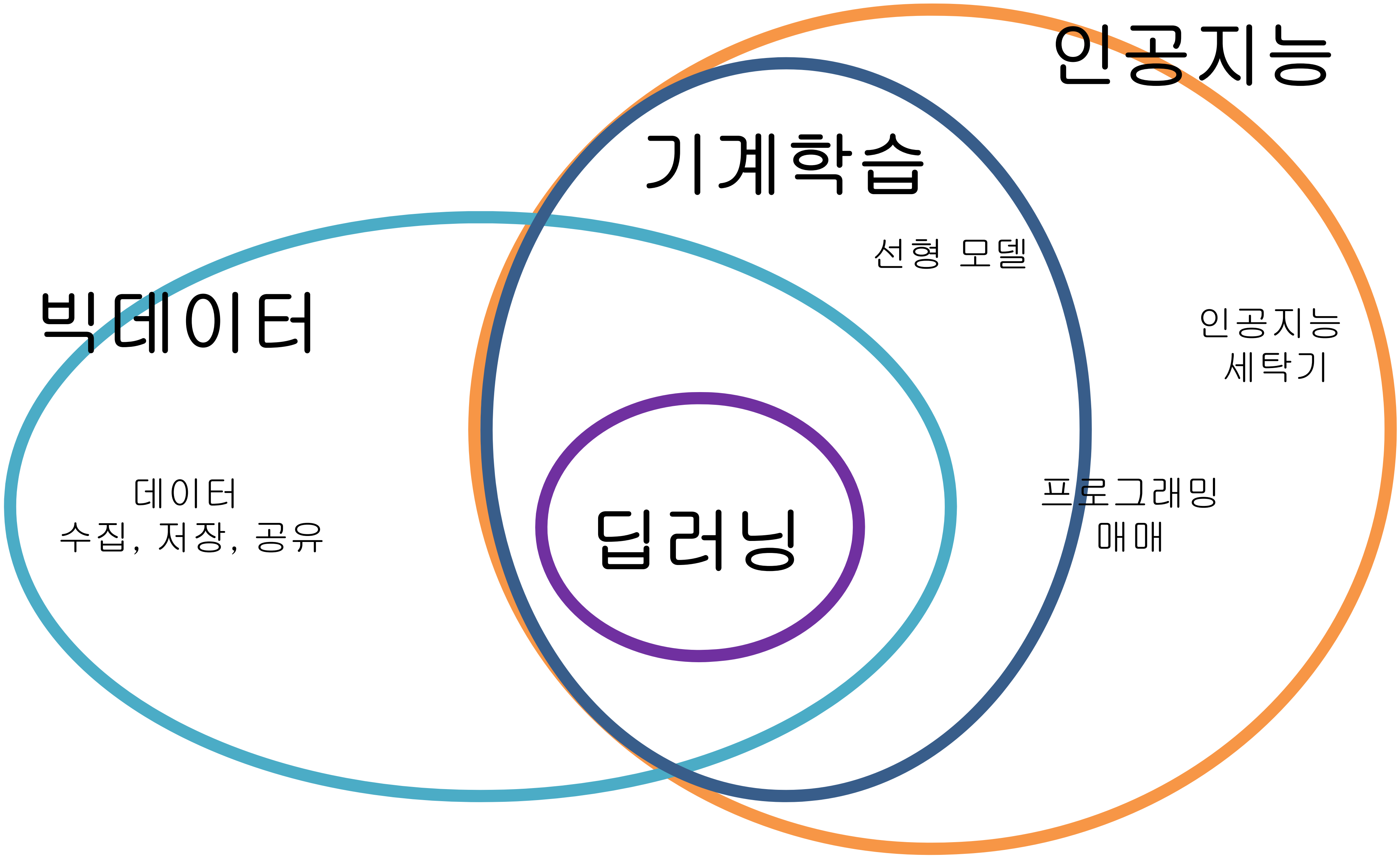
- 모든 관측가능한 데이터는 모집단으로부터 임의로 추출된 랜덤한 데이터
- 데이터를 다시 모은다면 성격은 유사하지만 전혀 다른 데이터가 수집됨
- 하나의 변수(e.g. 키, 판매량, 수율 등)는 하나의 확률 변수로 표현됨
- 인공지능, 빅데이터 분석, 데이터 과학 등은 기본적으로 확률 변수로 표현되는 데이터 사이의 관계를 찾는 것을 목적으로 함

데이터 분석

- **Y: 인공지능의 Action을 결정하는 데이터 (e.g. 내일의 주가, 번역의 결과, 이미지 인식 결과)**
- **X: Y를 예측하는데 도움이 될 것으로 생각되는 데이터 (e.g. 오늘의 주가, 원문, 이미지 자체)**
- **데이터 분석? X와 Y 사이의 관계를 찾는 것**



빅데이터, 인공지능, 기계학습??



확률

확률이란?

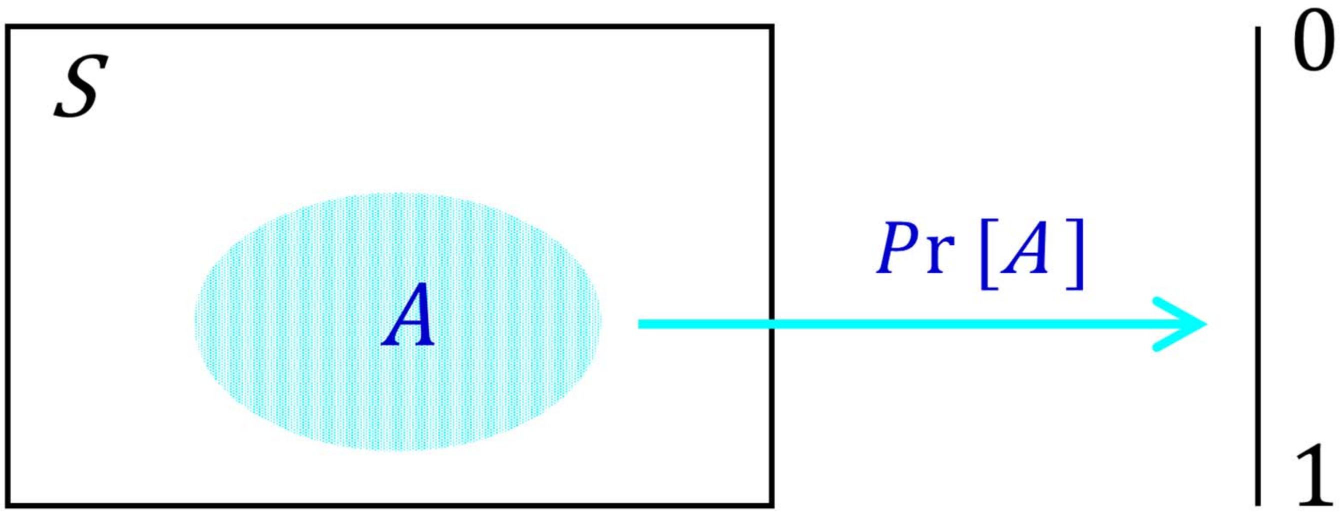
- **Q: 동전 던지기를 할 때, 앞면 혹은 뒷면 중 어떤 결과가 나올까?**
  - 방법 1 (결정론적 방법): 동전 던지기에 대한 모든 정보를 모은다. (동전 던지기 역학, 날씨, 장소 등). 그리고, 복잡한 역학 문제를 푼다.
    - 장점: 결정되어 있는 답을 얻을 수 있음.
    - 단점: 너무 복잡함. 또한 문제를 풀 수 없을 수도 있음.
  - 방법 2 (확률론적 방법): 동전을 모델링 한 후, 각 결과에 확률을 할당한다.
    - 장점: 매우 간단함.
    - 단점: 확정된 답을 얻을 수 없음.
- **공학의 많은 문제들이 확률을 통해 단순화 될 수 있기 때문에 확률을 배움.**
  - 인터넷으로 영화를 다운로드 할 때의 오차율은? 오차율은 다운로드 속도를 제한함.
  - NUGU는 얼마나 정확히 사람의 목소리를 인식 할 수 있을까?

확률론

- **“확률”이 무엇인지는 모두가 알고 있음.**
  - Q: 동전을 던졌을 때 앞면이 나올 확률은?
  - Q: 주사위를 던졌을 때 짝수가 나올 확률은?
  - Q: 소녀시대 멤버와 소개팅을 할 때 태티서의 멤버가 나올 확률은?
  - Q: 같은 소개팅에서, 제시카가 나올 확률은?
  - Q: 리그오브레전드에서 게임을 이길 확률은? 확률을 어떻게 추정할 수 있을까?
  - Q: 리그오브레전드에서 탑 챔피언이 티모일 때 게임을 이길 확률은? 티모가 아닐 때의 확률과 비교하여 높을까 낮을까 혹은 같을까?

확률론

- **확률론은 확률을 수학적으로 정의하기 위한 방법.**
  - 실험: 무작위한 결과를 만드는 시행
  - 표본: 시행에 따른 결과
  - 표본 공간: 가능한 모든 표본의 집합
  - 사건: 표본 공간의 부분집합
  - 확률: 어떠한 사건을 0과 1사이의 숫자에 대응시키는 함수. 확률의 공리에 해당함.



확률론

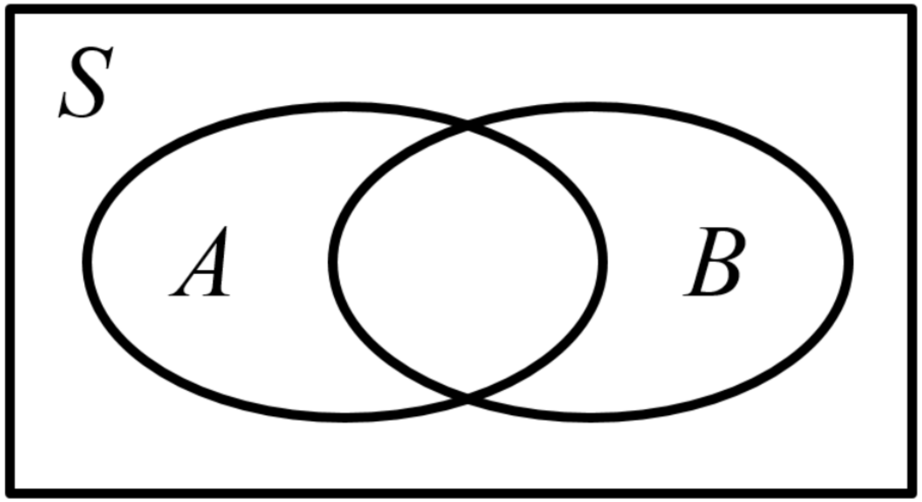
- **확률 Pr[]은 다음의 조건을 만족시키는 함수로써, 어떠한 사건을 특정 실수에 대응하도록 함**
  - 1. 모든 사건에 대하여 $\Pr[A] \geq 0$.
  - 2. $\Pr[S] = 1$.
  - 3. 구별되는 사건들 $A_1, A_2, \ldots, A_n$ 에 대해 $\Pr[A_1 \cup A_2 \ldots \cup A_n] = \Pr[A_1] + \Pr[A_2] + \ldots + \Pr[A_n]$.
  - 위의 조건들은 확률의 공리로 정의됨.
- **Examples**

실험	동전던지기	주사위던지기
표본	앞면, 앞면, 뒷면	2, 6, 3
표본공간	{앞면, 뒷면}	{1, 2, 3, 4, 5, 6}
사건	{앞면}, {뒷면}, { }	$A = \{2, 4, 6\}$ ...
확률	Pr [{앞면}] = ?	$Pr\,[A] = ?$ $Pr\,[\{1,2\}] = ?$

- **세개의 동전을 던질 때, 표본공간은?**

확률론

- **공리에 따른 몇가지 명백한 결과**
  - 1. $\Pr[\phi] = 0$.
  - 2. $\Pr[A^c] = 1 - \Pr[A]$.
  - 3. $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B] = \Pr[A] + \Pr[B] - \Pr[AB]$.
  - 4. If $A \subset B$, $\Pr[\mathrm{A}] \leq \Pr[B]$.
- **집합론과의 관계**



조건부 확률

- **사건 B가 발생했을 때, 사건 A가 발생할 확률 (혹은 줄여서 $A|B$) 은 다음과 같이 주어진다**
  - Pr[B] > 0이 성립할 때에만 정의됨.
  - B가 이미 발생했을 때의 A에 대한 상대적인 확률.
  - A에 대해서는 확률이지만, B는 이미 일어난 사건이기 때문에 B에 대한 확률이 아님.
- ***$A|B$는 $B|B$를 표본공간으로 가정했을 때의 사건으로 해석가능함*. 이 때, 확률의 공리는 다음과 같음**
  - 1. 모든 사건에 대하여 $\Pr[A|B] \geq 0$.
  - 2. $\Pr[B|B] = 1$.
  - 3. 구별되는 사건들 $A_1, A_2, \ldots, A_n$에 대해 $\Pr[A_1 \cup A_2 \ldots \cup A_n |B] = \Pr[A_1|B] + \Pr[A_2|B] + \ldots + \Pr[A_n|B]$.
- **Pr[A]는 조건부 확률과 대비하여 주변확률로 불림.**
- **Examples**
  - Q: 리그오브레전드 게임을 이길 확률은? (주변 확률)
  - Q: 탑 챔피언이 티모일 때, 리그오브레전드 게임을 이길 확률은? (조건부 확률)

조건부 확률

- **베이즈 정리**
  - Pr[*B*|*A*]를 Pr[*A*|*B*]로 나타내는 방법.

$$\mathbf{Pr}[\boldsymbol{B}|\boldsymbol{A}] = \frac{\mathbf{Pr}[\boldsymbol{A}|\boldsymbol{B}]\,\mathbf{Pr}[\boldsymbol{B}]}{\mathbf{Pr}[\boldsymbol{A}]}$$

  - 모든 *i*에 대해 $\Pr[B_i]>0$ 와 $S = \cup_i B_i$을 만족하는 구별되는 사건들 $B_1, B_2, \ldots, B_m$에 대해

$$\Pr[B_i|A] = \frac{\Pr[A|B_i]\Pr[B_i]}{\sum_{i=1}^{m}\Pr[A|B_i]\Pr[B_i]}$$

  - 얻기 쉬운 정보로부터 얻기 어려운 정보를 추정하기 위해 베이즈 정리를 사용함.

- **Examples**
  - 20%의 한국인, 10%의 중국인, 5%의 일본인이 잘 생겼다고 가정하자. 또한, 한국, 중국, 일본의 인구가 각각 1억, 13억, 2억이라고 할 때, 잘 생긴 사람이 한국인일 확률은 얼마인가?

독립

- **Pr[AB] = Pr[A]Pr[B]는 사건 *A* 와 *B* 는 독립 ($A \perp B$) 과 동치.**
- **따라서, $A \perp B$이면, Pr[$A|B$] = Pr[A] 그리고 Pr[$B|A$] = Pr[B].**

- **독립의 의미**
  - 두 사건이 관계가 없음
  - 한 사건의 결과를 아는 것이 다른 사건의 결과를 예측하는 데에 도움이 되지 않음

- **Examples**
  - 20%의 한국인, 10%의 중국인, 5%의 일본인이 잘 생겼다고 가정할 때, 한국인인 것을 아는 것은 어떤 사람이 잘 생겼을 확률을 추정할 수 있게 함. 국적과 잘 생김은 독립이 아님.
  - 동전던지기의 결과를 아는 것은 주사위던지기의 결과예측에 영향을 주지 않음. 두 사건은 독립이기 때문.

요약

- **실험, 사건, 표본 공간.**
- **확률은 사건을 어떠한 수로 대응시키는 함수.**
- **조건부 확률:** $\Pr[A|B] = \frac{\Pr[AB]}{\Pr[B]}$
- **베이즈 정리:** $\Pr[B|A] = \frac{\Pr[A|B]\Pr[B]}{\Pr[A]}$
- **독립:** $\boldsymbol{A \perp B}$ **은 Pr[AB] = Pr[A]Pr[B]와 동치.**

랜덤 변수

랜덤 변수

- **상수: 고정된 값 (예: 1, 3.5, π)**
- **보통의 변수: 고정된 값에 대한 변수**
- **랜덤 변수: 무작위 값에 대한 변수 혹은 개념**
  - 일반적으로, 랜덤변수는 숫자

- **(예) X: 동전던지기의 결과**
  - H → 1, T → 0
  - X = 0 혹은 1; Pr[X =0] = ?, Pr[뒷면] = ?
- **(예) X: 수지의 몸무게**
  - X의 표본공간?
  - Pr[X=-1] = ?
  - Pr[X = 47kg] = ? (네이버 프로필에 의하면 수지의 몸무게는 47kg).
  - Pr[46 < X < 48] = ?

랜덤 변수

- **랜덤 변수는 표본 공간(sample space)에 대한 확률 분포에 의해 정의된다.**

- **예시: 이상적인 동전 던지기 (H(앞면) -> 1, T(뒷면) -> 0)**
  - Pr[X=0] = 0.5, Pr[X=1] = 0.5
  - Pr[X=2] = 0,   Pr[X=-1] = 0
  - Pr[X= x] 는 표본 공간 S 내의 x에 대한 랜덤 변수 X의 확률분포이다.

- **이산 (discrete) 랜덤 변수: 이산 표본 공간**
  - 유한 이산 랜덤 변수 : S가 유한한 집합일 때, 예시: 동전던지기, 주사위 굴리기
  - 무한 이산 랜덤 변수 : S가 무한한 집합일 때, 예시: 카카오톡 메시지 수
- **연속 (continuous) 랜덤 변수 : 연속 표본 공간**

확률 질량 함수

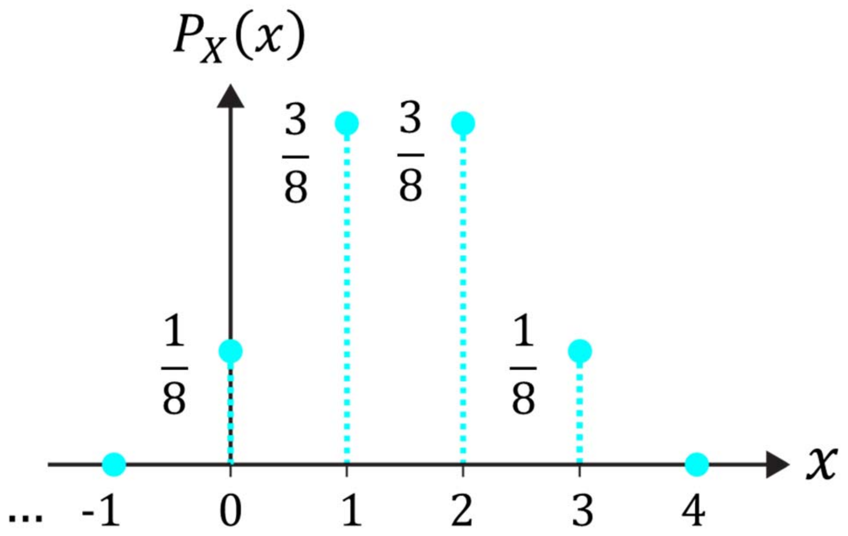
- **이산 랜덤 변수 X는 이산값인 x의 확률 분포에 의해 정의된다.**

- **PMF (확률 질량 함수; probability mass function)**
  - 이산 랜덤 변수에 대한 분포 함수
  - $\boldsymbol{P_X(x)} = \mathbf{Pr}[\boldsymbol{X} = \boldsymbol{x}]$
  - S에 x 가 없으면, $P_X(x) = 0$

- **PMFs의 특징**
  - (1) $P_X(x) \geq 0$
  - (2) $\sum_{x \in S_X} P_X(x) = 1$
  - (3) for $B \subset S_X$, $\Pr[B] = \sum_{x \in B} P_X(x)$

확률 질량 함수

- X: 동전 세 개를 던졌을 때 나온 앞면의 개수
- $S_X = \{ 0, 1, 2, 3 \}$
- $P_X(x) = \begin{cases} 1/8 & \text{for } x = 0 \text{ or } 3 \\ 3/8 & \text{for } x = 1 \text{ or } 2 \\ 0 & \text{otherwise} \end{cases}$



- **Quiz**
  - N 이 무한 이산 랜덤 변수이고 PMF는 $P_N(n) = \begin{cases} c/n^2, & n = 1, 2, 3 \cdots \\ 0, & \text{ow} \end{cases}$ 이다.
    - (1) c = ?
    - (2) Pr[N≥3] = ?

누적 분포 함수

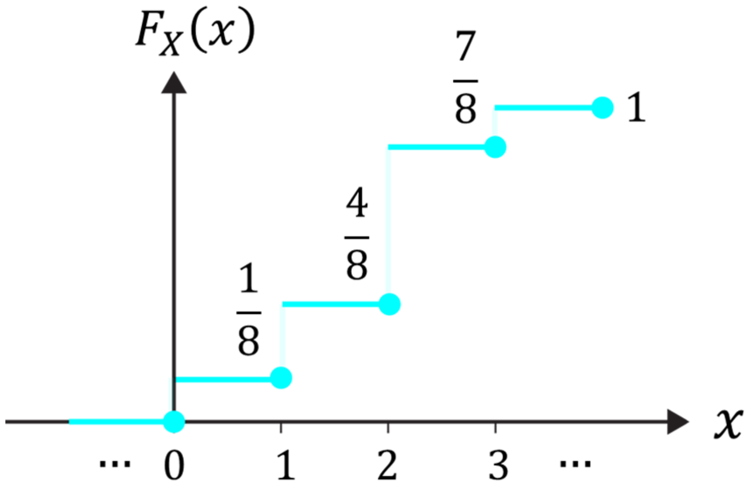
- **PMF 는 연속 랜덤 변수에 적합하지않다**
  - X: 버스를 기다리는 시간
  - $Pr\,[X = 1] = 0$ 이고, $Pr\,[X < 1]$ 는 0이 아니다

- **CDF (누적 분포 함수; cumulative distribution function)**
  - 이산 랜덤 변수와 연속 랜덤 변수 모두에 대한 확률 분포
  - $\boldsymbol{F_X(x) = \mathrm{Pr}[X \leq x]}$

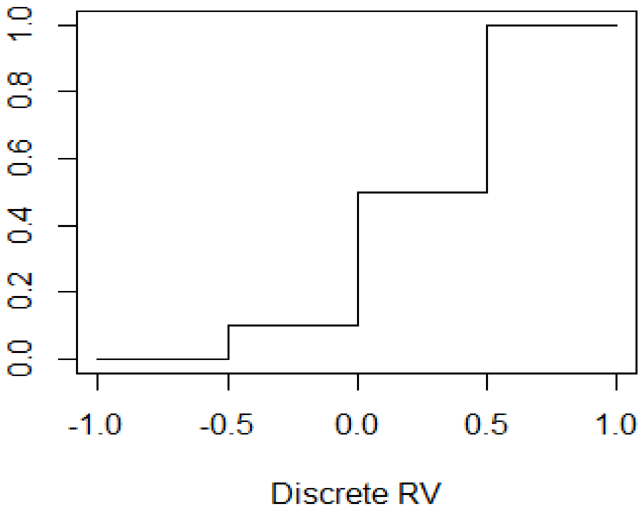
- **CDF의 특징**
  - (1) $F_X(-\infty) = 0$, and $F_X(\infty) = 1$
  - (2) If $b \geq a$, $F_X(b) \geq F_X(a)$: 단조 비 감소.
  - (3) For $b \geq a$, $F_X(b) - F_X(a) = \mathrm{Pr}[a < x \leq b]$.

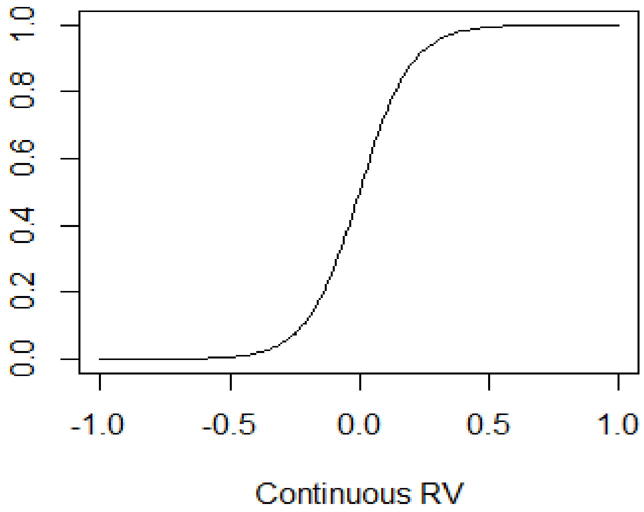
누적 분포 함수

- X: 동전 세 개를 던졌을 때 나온 앞면의 개수
- $S_X = \{0, 1, 2, 3\}$
- $F_X(x) = \begin{cases} 0, & x < 0 \\ 1/8, & 0 \le x < 1 \\ 4/8, & 1 \le x < 2 \\ 7/8, & 2 \le x < 3 \\ 1, & 3 \le x \end{cases}$



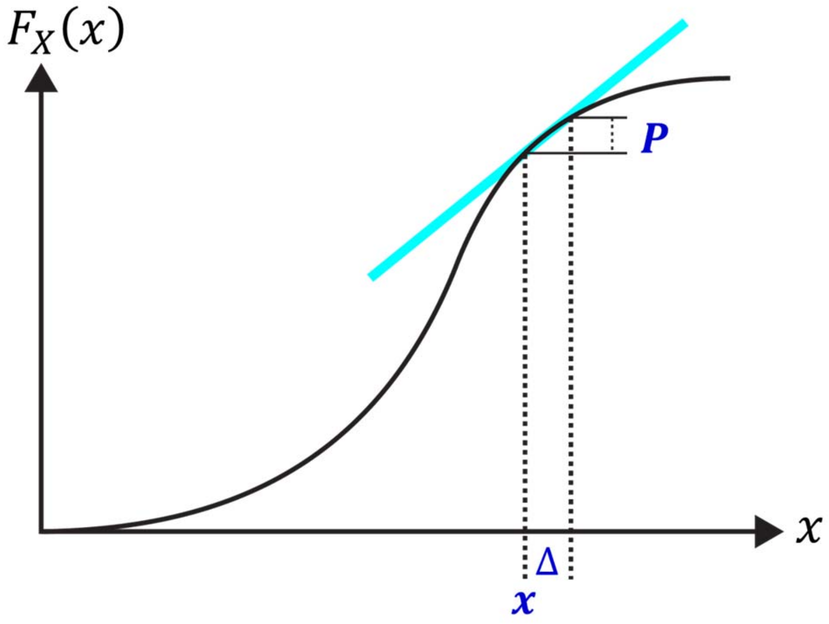
- 이산 및 연속 랜덤 변수의 **CDFs**

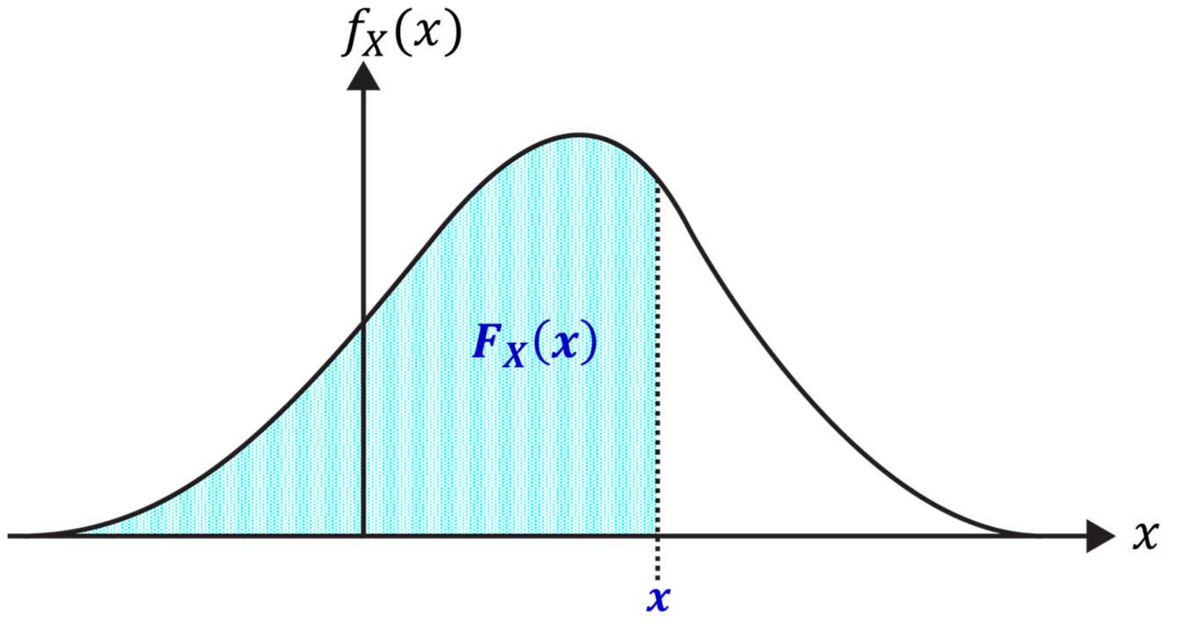




확률 분포 함수

- **PDF (확률 분포 함수; probability density function)**
  - 연속 랜덤 변수의 분포 함수
  - 연속 랜덤 변수에 대해, 확률 질량 ($Pr\,[X = x]$) 은 항상 0이지만, 확률 분포는 0이 아니다
  - $\boldsymbol{f_X(x) = \frac{\mathrm{Pr}[x < X \leq x + \Delta]}{\Delta} = \frac{F_X(x+\Delta) - F_X(x)}{\Delta} = \frac{dF_X(x)}{dx}}$
  - **PDF** 는 **CDF**의 미분이다





PMF vs. PDF

	PMF	PDF
①	$P_X(x) \geq 0$	$f_X(x) \geq 0$
②	$P_X(x) \leq 1$	$f_X(x) \leq 1$ (?)
③	$F_X(x) = \sum_{u \leq x} P_X(u)$	$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$
④	$\sum_{-\infty}^{\infty} P_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
⑤	$\Pr[a < X \leq b] = \sum_{x \in (a,b]} P_X(x)$ $= F_X(b) - F_X(a)$	$\Pr[a < X \leq b] = \int_{a}^{b} f_X(x)\, dx$ $= F_X(b) - F_X(a)$

기대값

- **기대값 (Expectation): 랜덤 변수의 값을 나타내는 고정된 값**

- **이산 랜덤 변수:** $\boldsymbol{E[X] = \mu_X = \sum_{x \in S_X} x P_X(x)}$
- **연속 랜덤 변수:** $\boldsymbol{E[X] = \mu_X = \int_{-\infty}^{\infty} x f_X(x) dx}$

- **기대값의 성질**
  - (1) $\mathrm{E}[x] = x$: 랜덤 변수가 아닌 변수 혹은 상수의 기대값은 그 변수이거나 그 값 자체이다.
  - (2) $\mathrm{E}\big[X - \mathrm{E}[X]\big] = 0$
  - (3) $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$: 선형성 (linearity).
  - (4) 일반적으로, $\mathrm{E}[g(X)] = \sum g(x) P_X(x)$ or $\int_{-\infty}^{\infty} g(x) f_X(x) dx$
  - (5) 일반적으로, $\mathrm{E}[g(X)] \neq g(\mathrm{E}[X])$. 예시. $\mathrm{E}[X^2] \neq \mathrm{E}[X]^2$. g()가 선형 함수라면 어떠한가?

표본 평균

- **표본 평균 (Sample Mean): 반복 실험으로 구한 랜덤 표본의 평균**

- **예시: 주사위 굴리기**
  - 5번의 실험으로부터, 당신은 6, 4, 1, 4, 2 를 얻었다
  - 표본 평균은 3.4이고, 이것은 다른 실험에선 달라질 수 있다
  - 기대값은 3.5이고, 고정된 값이다.

- **표본 평균은 더 많은 실험을 할수록 기대값에 가까워진다**
  - 큰 수의 법칙

분산

- **분산 (Variance): 랜덤 변수가 얼마나 많이 변할 수 있는지를 나타내는 고정된 값**
  - $\mathbf{Var}[\boldsymbol{X}] = \boldsymbol{\sigma_X^2} = \boldsymbol{E}\left[(\boldsymbol{X} - \boldsymbol{E}[\boldsymbol{X}])^2\right] = \boldsymbol{E}\left[\boldsymbol{X}^2\right] - \boldsymbol{E}[\boldsymbol{X}]^2$
  - 이산 랜덤 변수: $\mathbf{Var}[\boldsymbol{X}] = \sum_{x \in S_X} (\boldsymbol{x} - \boldsymbol{\mu_X})^2 \boldsymbol{P_X}(\boldsymbol{x})$
  - 연속 랜덤 변수: $\mathbf{Var}[\boldsymbol{X}] = \int_{-\infty}^{\infty} (\boldsymbol{x} - \boldsymbol{\mu_X})^2 \boldsymbol{f_X}(\boldsymbol{x}) \boldsymbol{dx}$

- **표준 편차 (Standard deviation)**
  - $\mathrm{SD}[X] = \sigma_X = \sqrt{\mathrm{Var}[X]}$

- **분산의 성질**
  - $\mathrm{Var}[X] \geq 0$: 분산은 0보다 작지않다
  - $\mathrm{Var}[x] = 0$: 상수의 분산
  - $\mathrm{Var}[aX+b] = a^2\mathrm{Var}[X]$: 오프셋 (offset, b)에 영향을 받지 않는다

분산

- **Example**

세 명의 고객이 스마트폰 가게에 들어가는 실험에서, 관측값은 N이고 N은 구매된 핸드폰의 개수이다. N의 PMF는 다음과 같다.

$$P_N(n) = \begin{cases} (4-n)/10 & n = 0, 1, 2, 3 \\ 0 & \text{otherwise.} \end{cases} \qquad (3.86)$$

다음을 계산하시오.

(a) 기대값 E[N]
(b) 2차 모멘트 $E[N^2]$
(c) 분산 Var[N]
(d) 표준 편차 σ_N

Yates, Probability and Stochastic Process 3rd, Wiley, 2014

요약

- **랜덤 변수는 확률 분포에 의해 정의되는 랜덤 값의 요약이다**
  - CDF: $F_X(x) = \Pr[X \leq x]$
  - PMF: $P_X(x) = \Pr[X = x]$
  - PDF: $f_X(x) = dF_X(x)/dx$
- **기대값: 랜덤 변수의 값을 나타냄**
  - $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$
- **분산: 랜덤 변수가 얼마나 변하는지를 나타냄**
  - $Var[X] = E[(X - E[X])^2]$

많이 사용하는 확률 분포

이산 균등 분포 (Discrete Uniform Distribution)

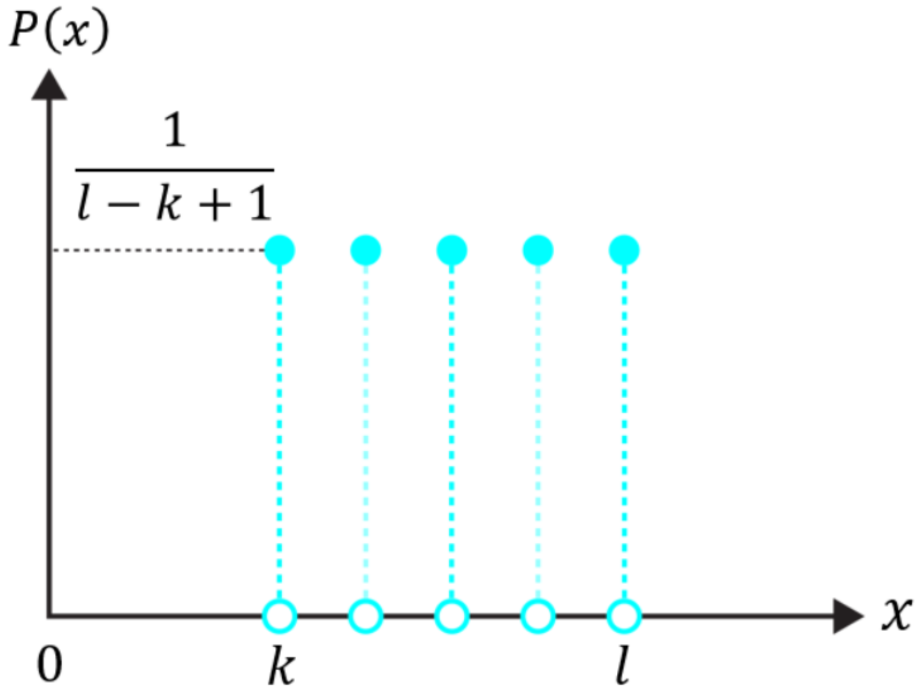
- $X \sim Unif(k, l)$; **균등 랜덤 변수**

- $P_X(x) = \begin{cases} \frac{1}{(l-k+1)}, & x = k, k+1, \cdots, l-1, l \\ 0, & otherwise \end{cases}$

- $E[X] = (k + l)\,/2\,;\;\; Var[X] = \boldsymbol{(}l - k\boldsymbol{)}\;\boldsymbol{(}l - k\; + 2)/12$

- **모든 표본 값은 동일한 확률을 갖는다**

- **예시: 이상적인 주사위 굴리기의 결과 ~ Unif(1,6)**



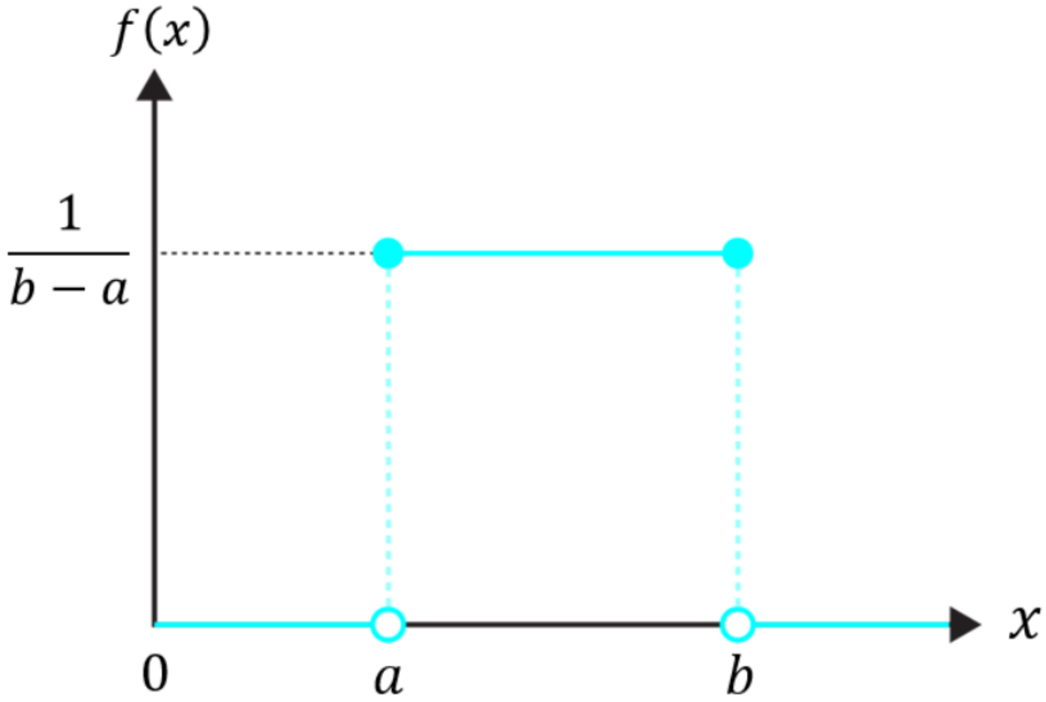
연속 균등 분포

- $X \sim Unif(a, b)$; 균등 랜덤 변수

- $f_X(x) = \begin{cases} \frac{1}{(b-a)}, & a \le x < b \\ 0, & \text{otherwise} \end{cases}$

- $E[X] = (a+b)/2\,, Var[X] = (b-a)^2/2$

- 모든 표본 값은 동일한 확률 밀도를 갖는다



베르누이 (Bernoulli) 분포

- $X \sim Bern(p)$**; 베르누이 랜덤 변수**

- $$P[X] = \begin{cases} p, & x=1 \\ 1-p, & x=0 \\ 0, & \text{otherwise} \end{cases}$$

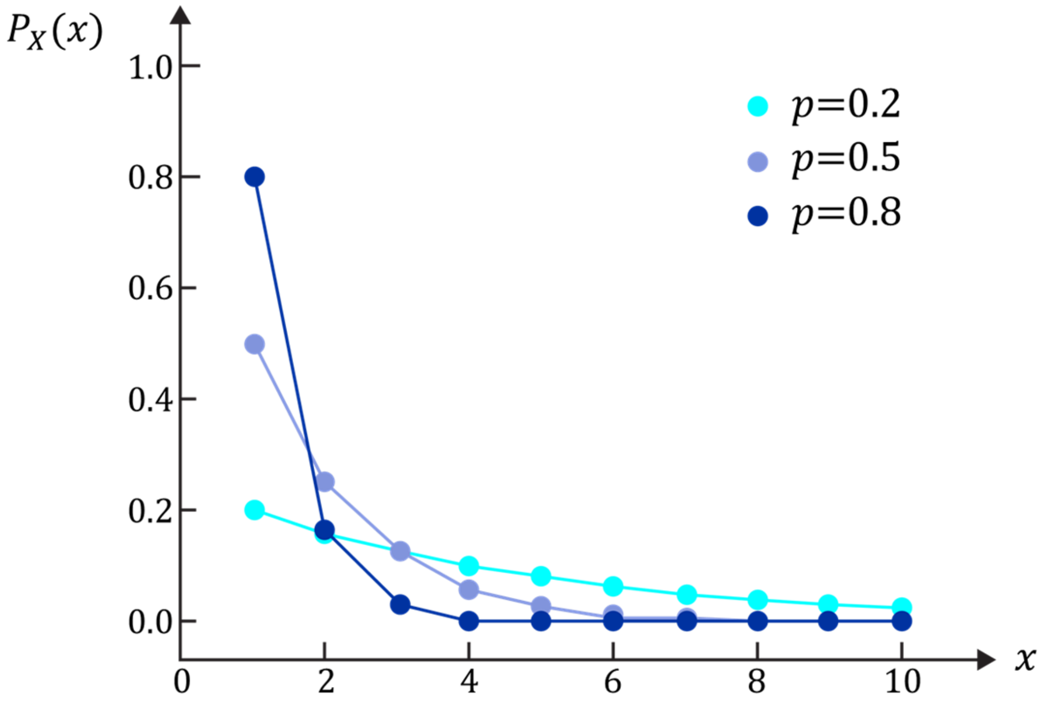
- $E[X] = p$, $Var[X] = p(1-p)$

- **이진 랜덤 결과**
  - 동전 던지기, 성공/실패, 네/아니오
  - 이진 분류를 많이 사용하기 때문에 중요함 (예시. 나이를 대신한 old / young)

기하분포(Geometric Distribution)

- $X \sim Geo(p)$; **기하 랜덤변수**
- $P(x) = p(1-p)^{x-1}$ **for** x=1, 2, 3, ⋯
- $E[X] = 1/p,\ \ \mathrm{Var}[X] = (1-p)/p^2$

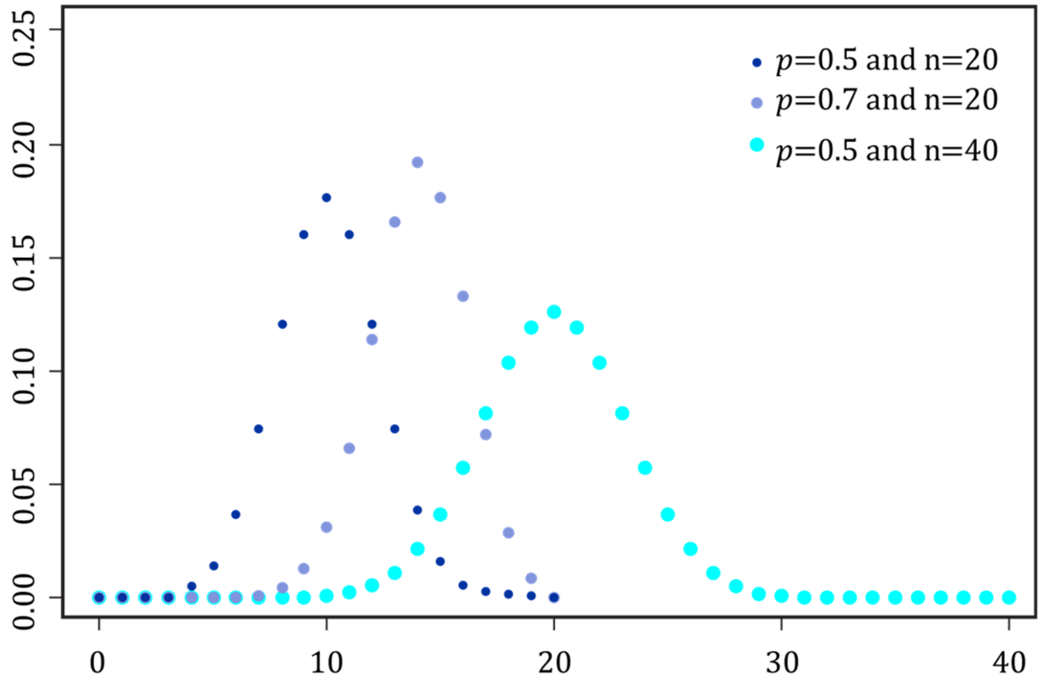
- **첫번째 성공할 때의 베르누이 횟수에 대한 분포**
  - 앞면이 나올 확률이 1/3일때 세번의 시도에 처음 앞면이 나올 확률
  - $X \sim Geo(\frac{1}{3});\ \ \Pr[X=3] = \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3}$



이항분포(Binomial Distribution)

- $X \sim B(n,p)$; **이한 랜덤변수**
- $P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$ **for** x=0, 1, 2, 3, ⋯, n
- $E[X] = np$, $\mathrm{Var}[X] = np(1-p)$

- **베르누이 시행에서 n번 성공할 확률**
  - 전송 실패확률이 0.1일때, 10번중에 2번 전송 실패할 확률 p = 0.19
  - Sum of *n* Bernoulli RVs ~ B(n,p)

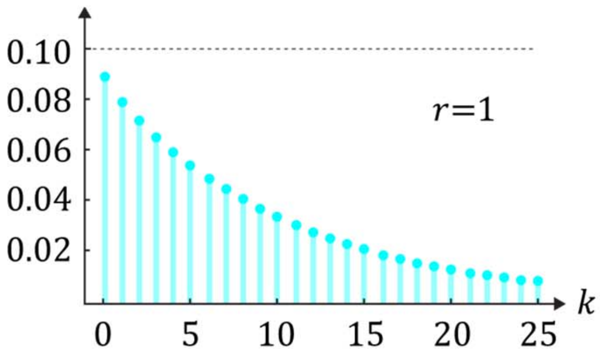


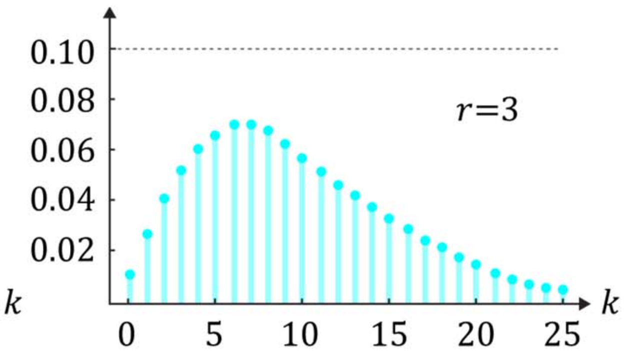
음이항 분포(Negative Binomial Distribution)

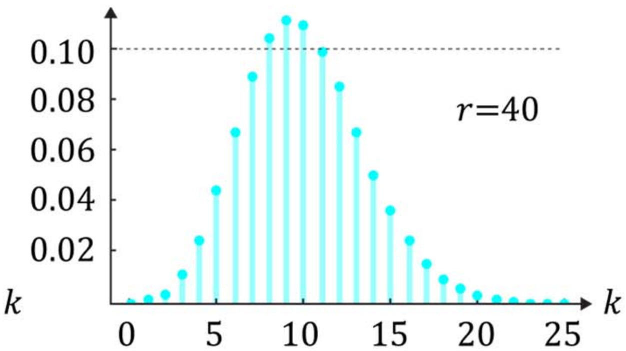
- X ~ $NB(r,p)$
- $P(x) = \binom{x+r-1}{x} p^x (1-p)^r$ **for** x**=0, 1, 2, ⋯**
- $E[X] = rp(1-p)$; $\mathrm{Var}[X] = rp(1-p)^2$

- **r번 실패할 때까지, 성공한 횟수에 대한 분포**

 − X ~ NB(1,p), then X+1 ~ Geo(1-p)



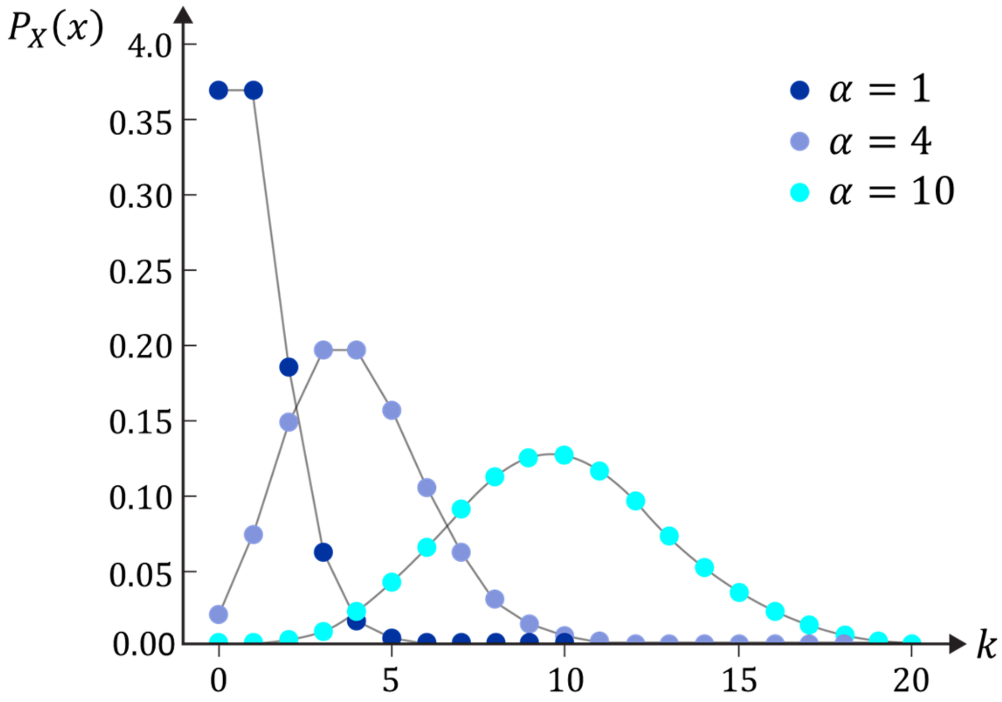




푸아송 분포(Poisson Distribution)

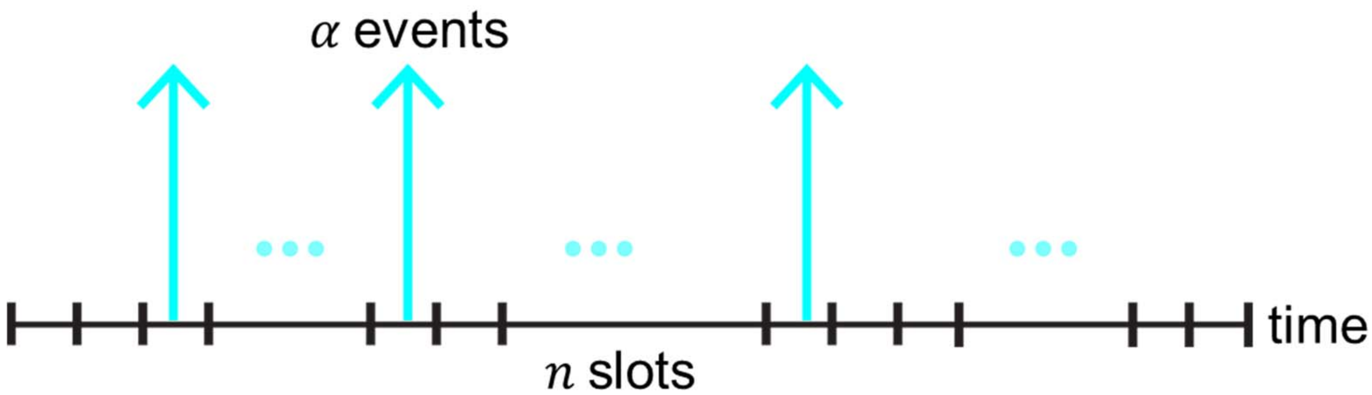
- $X \sim Poi(\alpha)$; **푸아송 랜덤변수**
- $P(x) = \begin{cases} \alpha^x e^{-\alpha}/x!, & x = 0,1,2,\cdots \\ 0\,, & \text{otherwise} \end{cases}$
- $E[X] = \alpha,\ \mathrm{Var}[X] = \alpha$

- **평균적으로 event가 α번 발생할때, x 번 event가 발생할 확률에 대한 분포**
  - 손님이 매장에 방문하는 경우
  - 웹사이트에 접속 요청
  - 무선통신에 패킷 수신(Packet reception)
  - 유전에서 변이



푸아송 분포(Poisson Distribution)

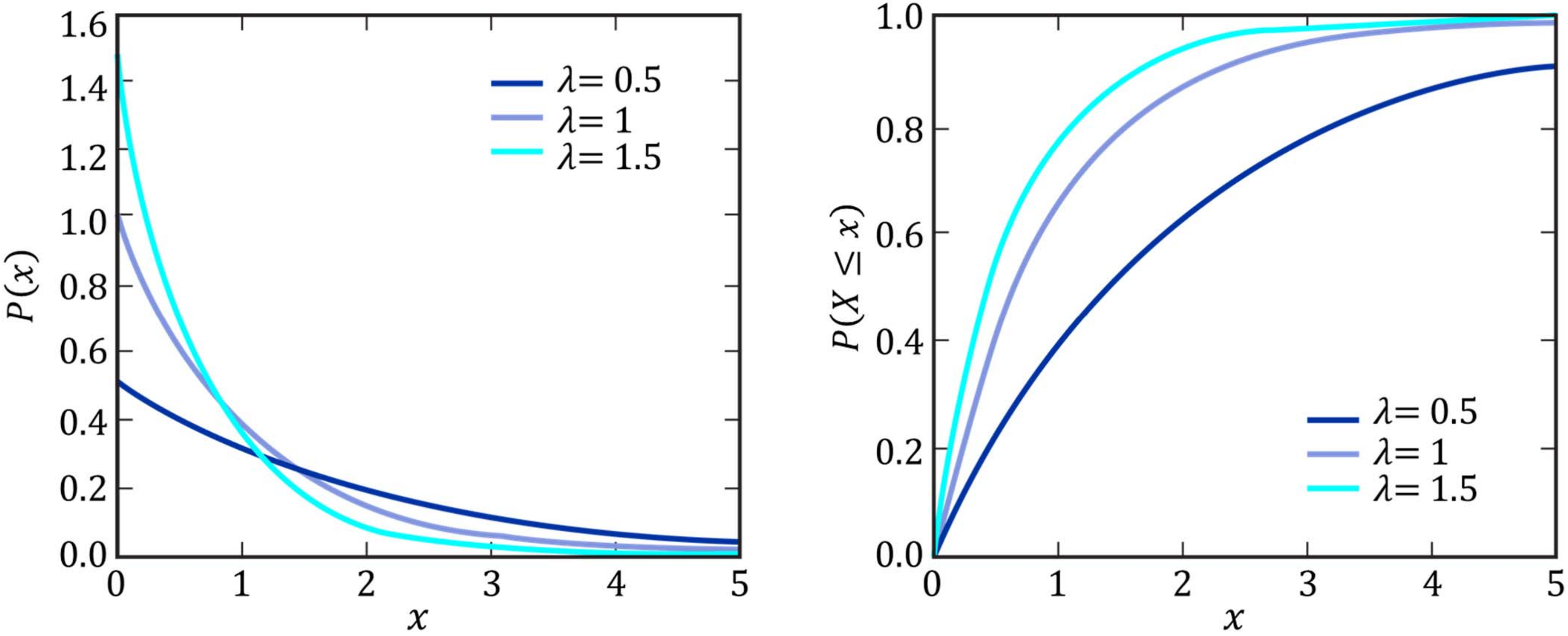
- **푸아송 분포를 이벤트의 발생횟수로 의미하는 이유**
  - n 단위 시간동안 이벤트가 평균적으로 α번 발생
  - 단위시간동안 이벤트가 발생할 확률 ~α/n
  - X: 실제로 이벤트가 발생한 횟수, $X \sim B(n, \alpha/n)$
  - When $n \to \infty$, $B(n, \alpha/n) \to Poi(\alpha)$



- **푸아송 랜덤변수 vs. 푸아송 랜덤과정**
  - α: 특정시점에 이벤트가 평균적으로 발생할 확률; $X \sim Poi(\alpha)$ → 푸아송 랜덤 변수
  - αt: 시간에 따라서 이벤트가 발생할 확률 t; $X(t) \sim Poi(\alpha t)$ → 푸아송 랜던 과정

지수분포(Exponential Distribution)

- $X \sim Exp(\lambda)$; 지수 랜덤변수
- $f_X(x) = \lambda e^{-\lambda x}$ **for** $x \geq 0$ → $F_X(x) = 1 - e^{-\lambda x}$
- $E[X] = \frac{1}{\lambda}$; $\text{Var}[X] = \frac{1}{\lambda^2}$
- **단위 시간동안 λ 번의 이벤트가 발생할 때, 첫 이벤트가 발생할 때까지 걸리는 시간의 분포**



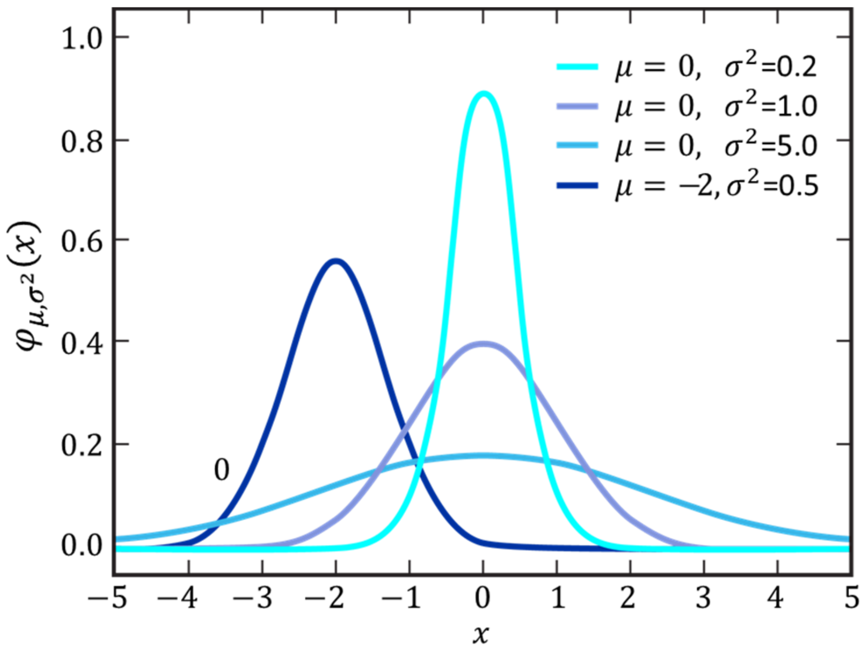
지수분포(Exponential Distribution)

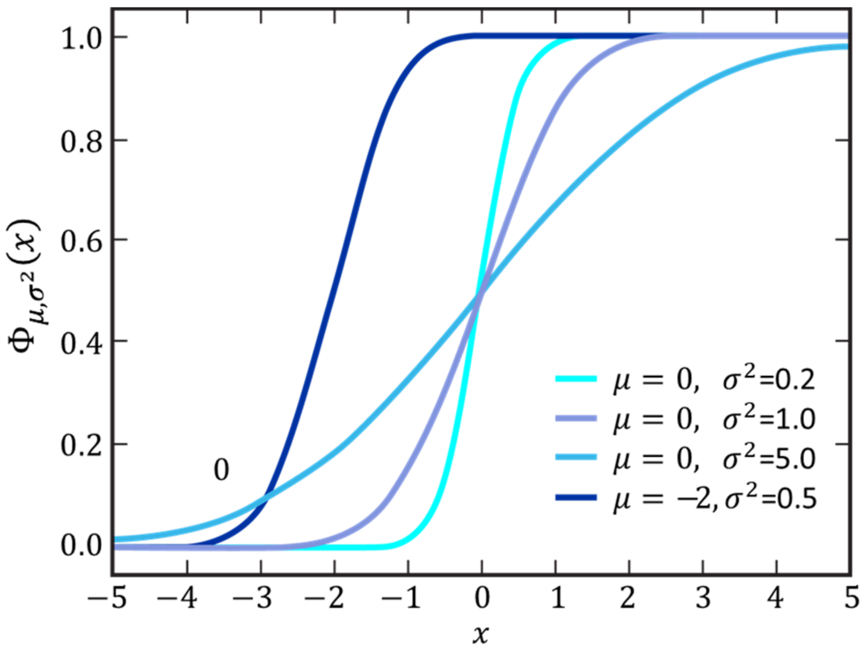
- **왜 지수분포가 event가 발생할때까지의 시간을 의미하는지?**
  - $X \sim Exp(\lambda)$
  - Let K be a discrete RV s.t. $\Pr[K = k] = \Pr[k-1 < X \leq k]$
  - Then, $P_K(k) = F_X(k) - F_X(k-1) = (1-e^{-\lambda})(e^{-\lambda})^{k-1}$; $K \sim Geo\left(1-e^{-\lambda}\right)$
  - 기하분포: 첫번째 성공까지 걸리는 시간; 지수분포: 이벤트에 대해 기다리는 시간
- **망각성질(Memoryless)**
  - $X \sim Exp(\lambda)$ and $\Pr[X > x] = 1 - F_X(x) = e^{-\lambda x}$.
  - For $x_2 > x_1$, $\Pr[X > x_2 | X > x_1] = \frac{\Pr[\,X>x_1, X>x_2]}{\Pr[X>x_1]} = \frac{\Pr[\,X>x_2]}{\Pr[X>x_1]} = e^{-\lambda(x_2-x_1)} = \Pr[X > x_2 - x_1]$.
  - 버스를 기달리는 것이 지수분포라고 가정하자. 이 경우에 버스를 기다린지 5분 이전 (Pr[X>10|X>5]) 에 버스가 올 시간의 확률과 버스를 기다린지 5분 이후(Pr[X>5])에 버스가 올 시간의 확률은 동일하다. 즉, 이전에 기다린 시간은 영향을 주지 않는다.

정규 분포(Normal Distribution)

- $X \sim N(\mu, \sigma^2)$; **가우시안 랜덤변수(normal random variable)**
- $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$
- $E[X] = \mu$; $\mathrm{Var}[X] = \sigma^2$
- **표준정규분포**
  - $Z \sim N(0, 1)$, $\Phi(x) = \Pr[Z \le x]$
  - $\sigma Z + \mu \sim N(\mu, \sigma^2) \leftrightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$

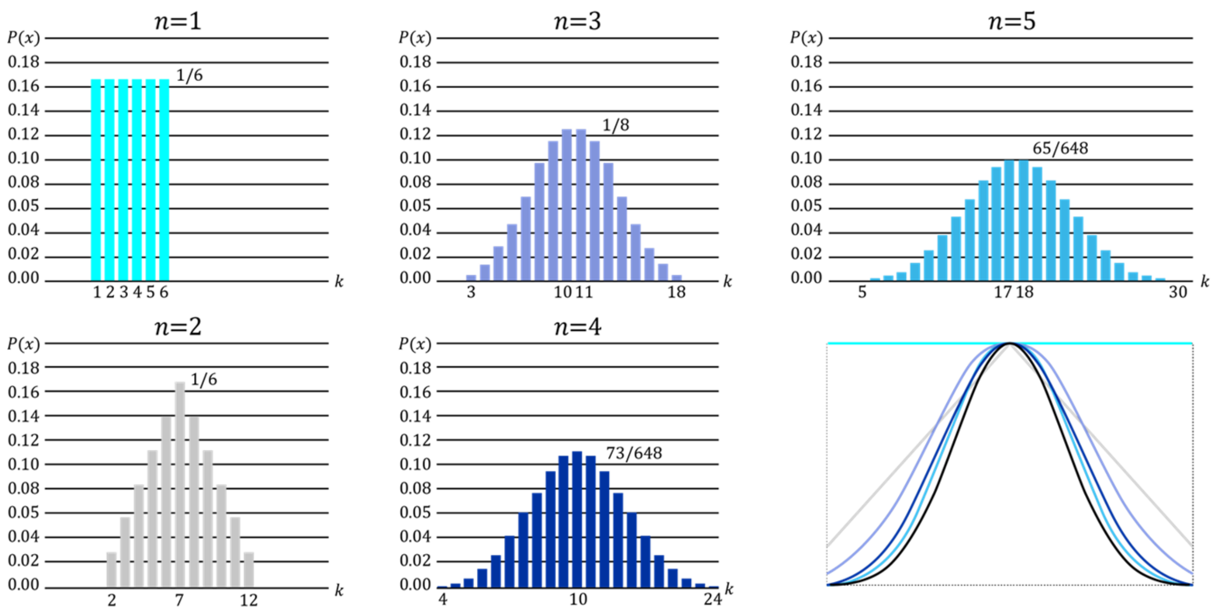
Quiz
어떤 데이터의 분포가 평균 1로 주어지는 정규분포를 정확히 따른다고 할때, 이 데이터의 값이 2.64보다 클 확률은 0.05이다. 이때, 이 데이터의 값이 -0.64 보다 클 확률은 얼마인가?





가우시안 분포(Normal Distribution)

- **중심극한정리(Central limit theorem)**
  - $Y = X_1 + X_2 + \cdots + X_n$ 는 동일하고 독립적인 분포를 가진다
  - $n \to \infty$, $Y \to$ 정규분포를 가지게 된다.
- **실제의 현상은 보통 매우 작고 독립적인 사건의 합으로 이루어진다. 그렇기 때문에 정규분포는 실제 상황에서 매우 중요한 분포이다.**
- **예시: 주사위의 결과:** $K = X_1 + X_2 + \cdots + X_n$



총정리

- 이산확률변수

	PMF	Exp	Var	Description
$Bern(p)$	$\begin{cases} p, & x=1 \\ 1-p, & x=0 \end{cases}$	p	$p(1-p)$	성공 또는 실패
$Geo(p)$	$p(1-p)^{x-1}$	$1/p$	$(1-p)/p^2$	베르누이 시행에서 첫 성공할때까지 시행횟수
$B(n,p)$	$\binom{n}{k} p^x(1-p)^{n-x}$	np	$np(1-p)$	베르누이 시행에서 n번중에 성공횟수
$NB(r,p)$	$\binom{x+r-1}{x} p^x(1-p)^r$	$rp(1-p)$	$pr(1-p)^2$	베르누이 시행에서 r번 실패했을때 성공횟수
$Poi(\alpha)$	$\alpha^x e^{-\alpha}/\ x!,$	α	α	이벤트의 횟수
$Unif(k,l)$	$\frac{1}{l-k+1}$	$\frac{k+l}{2}$	$\frac{(l-k)\ (l+k+2)}{12}$	동일분포

총정리

- 연속확률변수

	PMF	Exp	Var	Description
$Unif(a,b)$	$\begin{cases} \frac{1}{(b-a)}, a \le x < b \\ 0, \text{otherwise} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	동일분포
$Exp(\lambda)$	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$	이벤트가 발생할때까지 시간
$N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	중심극한정리

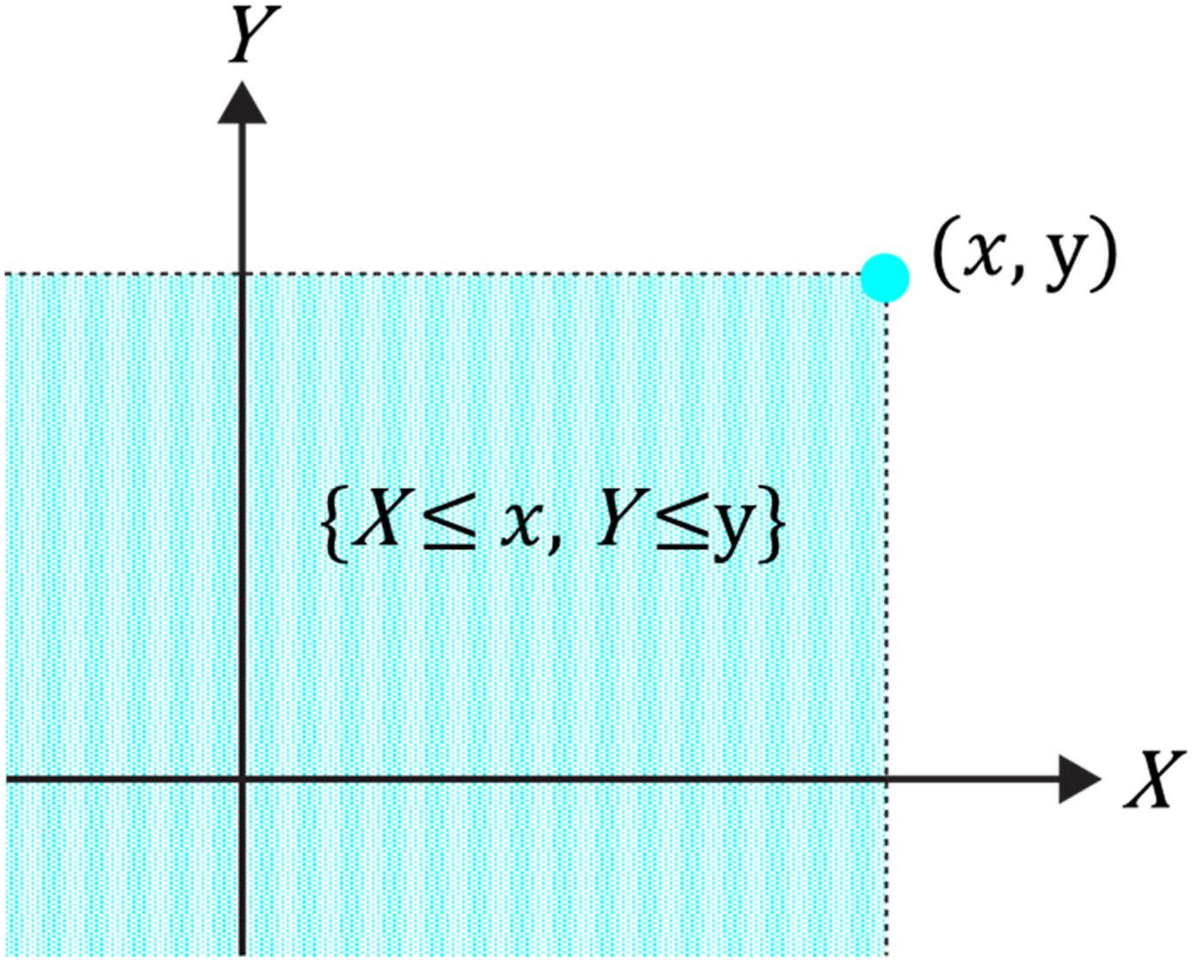
결합 확률분포

결합확률분포(Joint Probability Distribution)

- 2개의 연관된 확률 변수 간의 결합분포확률
  - X: 생일에 대한 확률(1~365). (균등분포?)
  - Y: 초등학교 1학년 시험성적000 (0~100). (정규분포?)
  - 각각의 분포도 흥미롭지만 2개의 연관 분포에 관심이 있다
    - E.g. Pr[Y>50|X<30] vs. Pr[Y>50|X>330]
  - X와 Y가 관계가 없는 독립적인 분포라고 가정해보자
    - Pr[Y>50|X<30] = Pr[Y>50] = Pr[Y>50|X>330]
  - 일반적으로 아래와 같이 표현할 수 있다.
    - $\Pr[Y > 50|X < 30] = \frac{\Pr[X<30, Y>50]}{\Pr[X<30]} = \frac{\Pr[X<30, Y>50]}{\Pr[X<30, Y<\infty]}$
    - $\Pr[X \le x, Y \le y]$ 위와 같은 확률분포를 구할 수 있다면, 우리가 알고 싶은 모든것을 계산할 수 있다.

결합누적분포함수(Joint Cumulative Distribution Function)

- **결합분포는 X와 Y라는 랜덤변수 2개의 관계를 $\mathrm{Pr}[X, Y]$ 과 같이 표현한다.**
- **Joint CDF**
  - $\boldsymbol{F_{X,Y}(x, y) = \mathrm{Pr}[X \leq x, Y \leq y]}$



결합누적분포의 특성

- $0 \leq F_{X,Y}(x,y) \leq 1$
- **When** $x_1 \leq x_2$ **and** $y_1 \leq y_2$**,** $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$
- $F_{X,Y}(x,\infty) = \Pr[X < x, Y < \infty] = F_X(x)$**; marginal cdf**
- $F_{X,Y}(\infty,\infty) = 1$ **,** $F_{X,Y}(-\infty,-\infty) = 0$ **,** $F_{X,Y}(-\infty,\infty) =?$ **,** $F_{X,Y}(\infty,-\infty) =?$
- $\Pr[x_1 < X \leq x_2, y_1 < Y \leq y_2] = F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - F_{X,Y}(x_2, y_1) + F_{X,Y}(x_1, y_1)$

결합 PMF 와 PDF

이산 랜덤변수	연속 랜덤변수
$P_{X,Y}(x,y) = \Pr[X = x, Y = y]$	$f_{X,Y}(x,y) = \Pr[x < X \leq x + dx, y < Y \leq y + dy]/dxdy$
$P_X(x) = \sum_{y \in S_Y} P_{X,Y}(x,y)$ $P_Y(x) = \sum_{x \in S_X} P_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$ $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$
$\sum_{x \in S_X} \sum_{y \in S_Y} P_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dxdy = 1$
$F_{X,Y}(x,y) = \sum_{u \leq x} \sum_{v \leq y} P_{X,Y}(u,v)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)dudv$
	$f_{X,Y}(x,y) = \partial^2 F_{X,Y}(x,y)/\partial x \partial y$
$\Pr[A] = \sum_{(x,y) \in A} P_{X,Y}(x,y)$	$\Pr[A] = \iint_A f_{X,Y}(x,y)dxdy$

두 랜덤 변수의 함수의 평균

- **하나의 랜덤 변수에 대해서,** $\mathrm{E}[g(X)] = \sum g(x)P_X(x)$ **또는** $\mathrm{E}[g(X)] = \int g(x)f_X(x)dx$**.**
- **두 랜덤 변수에 대해서,**
  - $\mathbf{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Y})] = \sum \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ 또는 $\mathbf{E}[\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Y})] = \int \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{f}_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{dxdy}$

- **예시**
  - $\mathrm{E}[X] = \int_x \int_y x f_{X,Y}(x,y)dydx = \int_x x\left(\int_y f_{X,Y}(x,y)dy\right)dx = \int_x x f_X(x)dx = \mathrm{E}[X]$
  - $\mathrm{E}[X+Y] = \iint (x+y)f_{X,Y}(x,y)dxdy = \iint x f_{X,Y}(x,y)dxdy + \iint y f_{X,Y}(x,y)dxdy = \mathrm{E}[X] + \mathrm{E}[Y]$
  - 일반적으로, $\mathrm{E}[a_1 g_1(X,Y) + \cdots + a_n g_n(X,Y)] = a_1\mathrm{E}[g_1(X,Y)] + \cdots + a_n\mathrm{E}[g_n(X,Y)]$
  - $g(X,Y) = (X + Y - \mu_X - \mu_Y)^2$

$$\begin{aligned} E[g(X,Y)] = \mathrm{Var}[X+Y] &= \mathrm{E}[(X+Y-\mu_X-\mu_Y)^2] = \mathrm{E}[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)] \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y] \end{aligned}$$

공분산

- **공분산**
  - 두 랜덤 변수 간의 관계를 간단하게 나타내는 값
  - 두 변수 간에 함께 변하는 정도를 나타낸다.

$$\mathrm{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)]$$
$$= E[XY] - E[X]E[Y]$$
$$\sigma_{X,Y} = r_{X,Y} - \mu_X \mu_Y$$

  - 상관도: $E[XY]$
  - 상관되어 있지 않음: When $\mathrm{Cov}[X, Y]$ = 0
- **평균, 분산, 공분산**
  - E[*X*]: the value representing *X*.
  - Var[*X*]: how much *X* varies.
  - Cov[*X*,*Y*]: how much *X* and *Y* vary together.

공분산

- **공분산의 성질**
  - $\mathrm{Cov}[aX + b, cY + d] = ac\mathrm{Cov}[X, Y]$
  - $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$
- **상관계수**
  - 공분산을 분산으로 스케일한다. 스케일에서 자유로움

$$\rho_{X,Y} = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}}$$

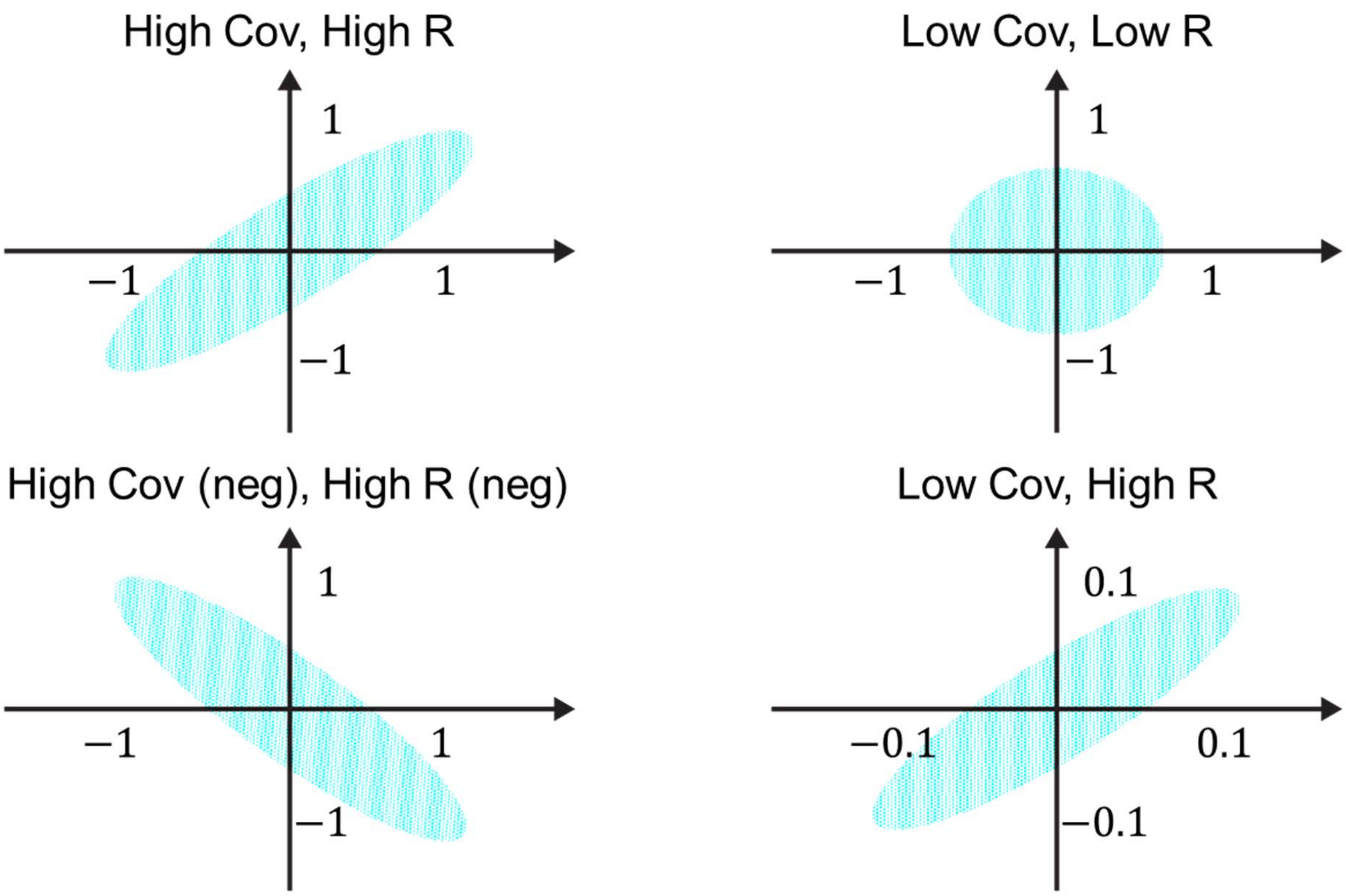
  - -1 과 1 사이로 제한됨
- **공분산 vs. 상관계수**
  - 상관계수는 분산에 상대적인 공분산이다.
  - $\mathrm{Cov}[X,Y_1] > \mathrm{Cov}\,[X,Y_2]$ → Var[Y]에 의존적이기 때문에 직관적이지 않다.

    만약 $Y_2 = 2Y_1$이면, $\mathrm{Cov}[X,Y_1] = \mathrm{Cov}\,[X,Y_2]/2$.
  - $\rho_{X,Y_1} > \rho_{X,Y_2}$ → 직관적이다, X_1 은Y_2 보다 Y_1과 상관도가 높다.

    $Y_2 = 2Y_1$이더라도, $\rho_{X,Y_1} = \rho_{X,Y_2}$.

공분산

- 상관계수는 스케일에 자유롭지만, 공분산은 그렇지 않다.



공분산

- **공분산 행렬**

$$\mathbf{C}_{X,Y} = \mathrm{E}\left[\left(\begin{bmatrix} X \\ Y \end{bmatrix} - \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}\right)\left(\begin{bmatrix} X \\ Y \end{bmatrix} - \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}\right)^T\right] = \begin{bmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{bmatrix}$$

  - 1) 대칭이다 ($\mathbf{C}^T = \mathbf{C}$), 2) 양의 준 정부호 행렬(semi-positive definite)이다. ($\mathbf{x}^T\mathbf{C}\mathbf{x} \geq 0$ for any $\mathbf{x}$), 3) $\rho_{X,Y}^2 = 1$ 가 아니라면, 가역행렬이다.

- **예시**

  - Cov[Height(cm), Weight(kg)] vs. Cov[Height(m), Weight(g)]? 상관계수는 어떨까?

  - The joint probability density function of random variables X and Y is

$$f_{X,Y}(x,y) = \begin{cases} xy & 0 \leq x \leq 1, 0 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

    Find the following quantities.
    (1) $E[X]$ and $\mathrm{Var}[X]$
    (2) $E[Y]$ and $\mathrm{Var}[Y]$
    (3) The correlation $r_{X,Y} = E[XY]$
    (4) The covariance $\mathrm{Cov}[X,Y]$
    (5) The correlation coefficient $\rho_{X,Y}$

Yates, Probability and Stochastic Process 3rd, Wiley, 2014

요약

- 두 랜덤 변수 $\boldsymbol{X}$ 와 $\boldsymbol{Y}$에 대해,
  - 공동 CDF: $F_{X,Y}(x,y) = \Pr[X \leq x, Y \leq y]$
  - 공동 PMF: $P_{X,Y}(x,y) = \Pr[X = x, Y = y]$
  - 공동 PDF: $f_{X,Y}(x,y) = \frac{\Pr[x<X\leq x+dx, y<Y\leq y+dy]}{dxdy} = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$
- **공분산:** $\mathrm{Cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$; **X 와 Y** 가 얼마나 함께 변하는지
- **상관계수:** $\rho_{X,Y} = \frac{Cov[X,X]}{\sqrt{Var[X]Var[Y]}}$; **-1 과 1 사이로 스케일됨**

조건부 확률 분포

조건부 확률 분포

- **B일 때, A의 조건부 확률: Pr[A|B] = Pr[AB]/Pr[B].**
  - 조건부 확률은 사건 B가 이미 발생하고 난 뒤의 확률을 이야기한다.
- **조건부 랜덤 변수의 조건부 분포 X|Y=y**
  - Y가 특정한 값인 y로 특정지어졌을 때, 랜덤 사건 X를 나타냄
  - 조건부 cdf/pmf/pdf 는 Y=y일 때의 X의 확률 분포를 나타낸다.

- **조건부 분포**
  - 조건부 PMF: $P_{X|Y}(x|y) = \Pr[X = x|Y = y] = \frac{\Pr[X=x,Y=y]}{\Pr[Y=y]} = \frac{P_{X,Y}(x,y)}{P_Y(y)}$
    - $P_{X,Y}(x,y) = P_{X|Y}(x|y)P_Y(y) = P_{Y|X}(y|x)P_X(x)$
  - 조건부 PDF: $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
    - $f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$

베이즈 정리

- **이산형 분포에서,**

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} = \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)} = \frac{P_{Y|X}(y|x)P_X(x)}{\sum_u P_{Y|X}(y|u)P_X(u)}$$

$$\uparrow$$

$$P_Y(y) = \sum_u P_{X,Y}(u,y) = \sum_u P_{Y|X}(y|u)P_X(u)$$

- **연속형 분포에서,**

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|u)f_X(u)\,du}$$

- **베이즈 정리는 <u>얻기 쉬운 정보로부터 얻기 어려운 정보를 얻어내기 위해 사용되어진다.</u>**

예시

- X 가 pmf 는 $P_X(0.2)=0.6$, $P_X(0.5)=0.3$, and $P_X(0.8)=0.1$ 를 만족하는 이산형 랜덤 변수라 하자. $Y \sim Bern(X)$ 일 때, $Pr[X=0.5|Y=1]$의 값은?

- 전체 인구의 10%가 겪고있는 질병에 대하여 실제로 이 질병이 걸렸는지 아닌지 검사를 한다고 하자. 실제로 질병에 걸린 사람이 이 검사에서 양성으로 나올 확률은 90%이고, 실제로 질병에 걸리지 않은 사람이 음성으로 나올 확률 또한 90%이다. 어떤 환자가 이 검사에서 양성 판정을 받았을 때, 실제로 이 환자가 질병에 걸렸을 확률은 얼마인가?

평균과 분산

- **평균**
  - $E[X|Y=y]=\sum_x xP_{X|Y}(x,y)\ or\ \int_x\ xf_{X|Y}(x,y)dx$: *y에 대한 함수*.
- **랜덤 변수의 함수의 평균**
  - $E[g(X,Y)|Y=y]=\sum_x g(x,y)P_{X|Y}(x,y)\ or\ \int_x\ g(x,y)f_{X|Y}(x,y)dx$
- **분산**
  - $\mathrm{Var}[X|Y=y]=\mathrm{E}[(X-\mathrm{E}[X|Y=y])^2|Y=y]=\mathrm{E}[X^2|Y=y]-\mathrm{E}[X|Y=y]^2$

- **랜덤 변수의 함수로써 나타내는 평균**
  - $E[X|Y=y]=g(y)$: 변수 y에 관한 함수, 고정된 값
  - $E[X|Y]=g(Y)$: 랜덤 변수 Y에 관한 함수, 랜덤한 값
    - $\mathrm{E}\big[\mathrm{E}[X|Y]\big]=\mathrm{E}[X],\quad \mathrm{E}\big[\mathrm{E}[g(X)|Y]\big]=\mathrm{E}[g(X)]$
    - $\mathrm{E}_Y\big[\mathrm{Var}_X[X|Y]\big]=\mathrm{Var}[X]$? 아니다. 사실, $\mathrm{Var}[X]=\mathrm{E}_Y\big[\mathrm{Var}_X[X|Y]\big]+\mathrm{Var}_Y\big[\mathrm{E}_X[X|Y]\big]$ 이고, 이를 the law of total variance 라 한다.

독립 (Independency)

- **두 랜덤 변수가 독립이면 아래와 동치이다.**
  - $P_{X,Y}(x,y) = P_X(x)P_Y(y)$, or $f_{X,Y}(x,y) = f_X(x)f_Y(y)$
- **독립인 X and Y 는**
  - $P_{X|Y}(x|y) = P_X(x)$ → $E[X|Y] = E[X]$
  - $E[XY] = E[X]E[Y]$ → 일반적으로, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$
  - $\mathrm{Cov}[X,Y] = 0$ → $Var[X+Y] = Var[X] + Var[Y]$: 상관되어있지않음.
- **독립 → 상관되지않음, 역은 항상 성립하지 않는다.**

- **독립은 아래와 같이 해석될 수 있다.**
  - 관계없음
  - 하나의 랜덤 변수를 아는 것이 다른 하나의 변수를 추정하는데 도움이 될 수 없다.

이변량 정규 분포

- **X and Y 가 이변량 정규 분포이면**

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{bmatrix}\right) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- $\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$ **and** $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{bmatrix}$ **이면, 결합 확률 분포는 다음과 같다.**

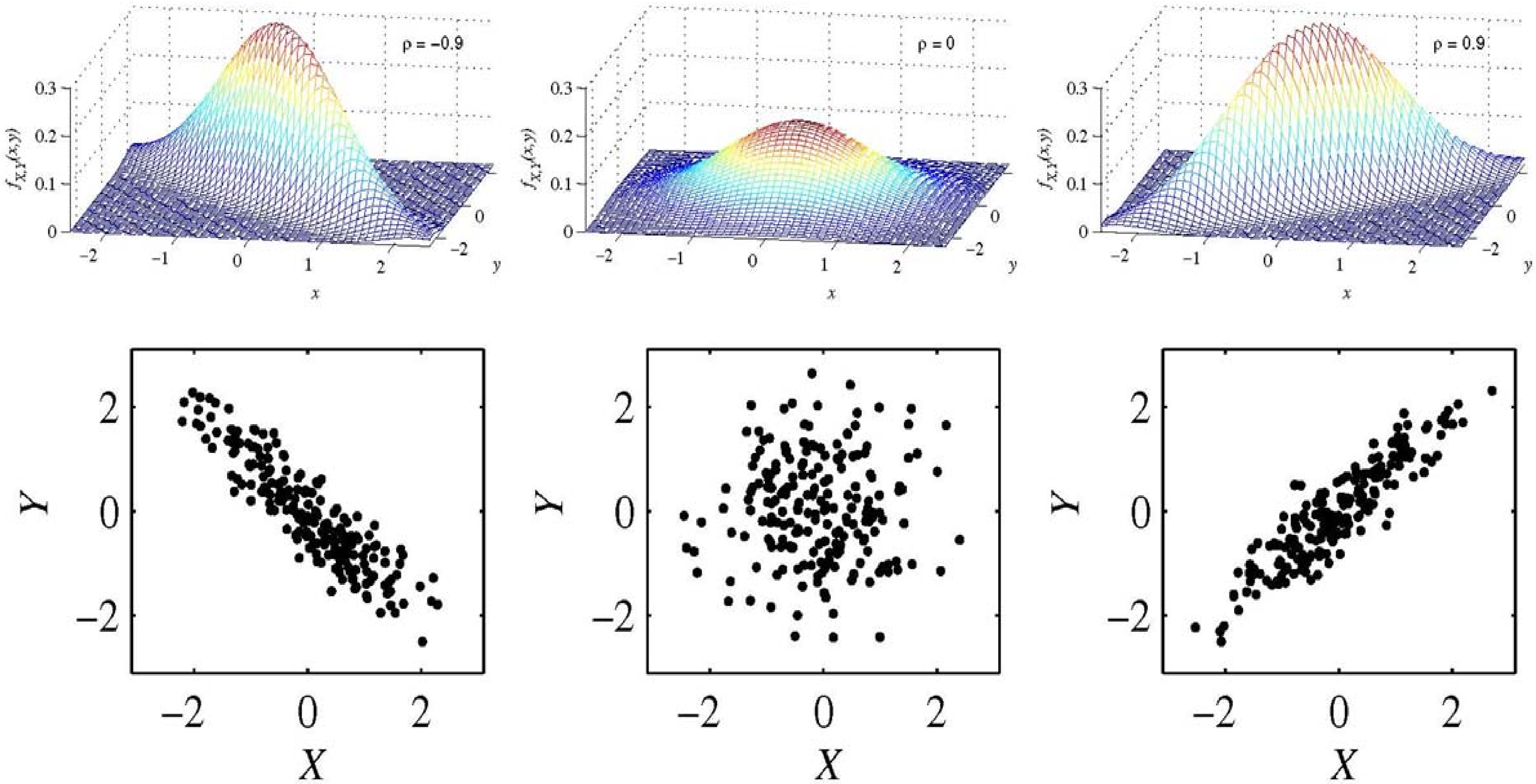
$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}\left(\begin{bmatrix} x \\ y \end{bmatrix} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\begin{bmatrix} x \\ y \end{bmatrix} - \boldsymbol{\mu}\right)\right] \\ &= \frac{1}{2\pi\sigma_X\sigma_Y\left(1-\rho_{X,Y}^2\right)^{1/2}} exp\left[-\frac{1}{2(1-\rho_{X,Y}^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho_{X,Y}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right] \end{aligned}$$

- **일변량 정규분포의 확률 분포 함수와의 유사성을 확인해보자.**

$$f_X(x) = \frac{1}{\sqrt{2\pi}(\sigma_X^2)^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_X)\cdot\frac{1}{\sigma_X^2}\cdot(x-\mu_X\right)\right]$$

이변량 정규 분포

- **시각화**



Yates, Probability and Stochastic Process 3rd, Wiley, 2014

이변량 정규 분포

- **When** $\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{bmatrix}\right) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,
  - (1) $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$: marginal 또한 정규분포를 따른다.
  - (2) $X|Y \sim \mathrm{N}\left(\mu_X + \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y - \mu_Y), \sigma_X^2(1 - \rho_{X,Y}^2)\right)$: 조건부 변수 또한 정규분포를 따른다.
  - (3) $X \perp Y \leftrightarrow \sigma_{X,Y} = 0$: 상관도 0이면 서로 독립인 것과 같다.
  - (4) $aX + bY \sim N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{X,Y}\right)$: 선형 조합 또한 정규 분포를 따른다.

- **Example**

A person's white blood cell (WBC) count W (measured in thousands of cells per microliter of blood) and body temperature T (in degrees Celsius) can be modeled as bivariate Gaussian random variables such that W is Gaussian $(7, 2)$ and T is Gaussian $(37, 1)$. To determine whether a person is sick, first the person's temperature T is measured. If $T > 38$, then the person's WBC count is measured. If $W > 10$, the person is declared ill (event I).

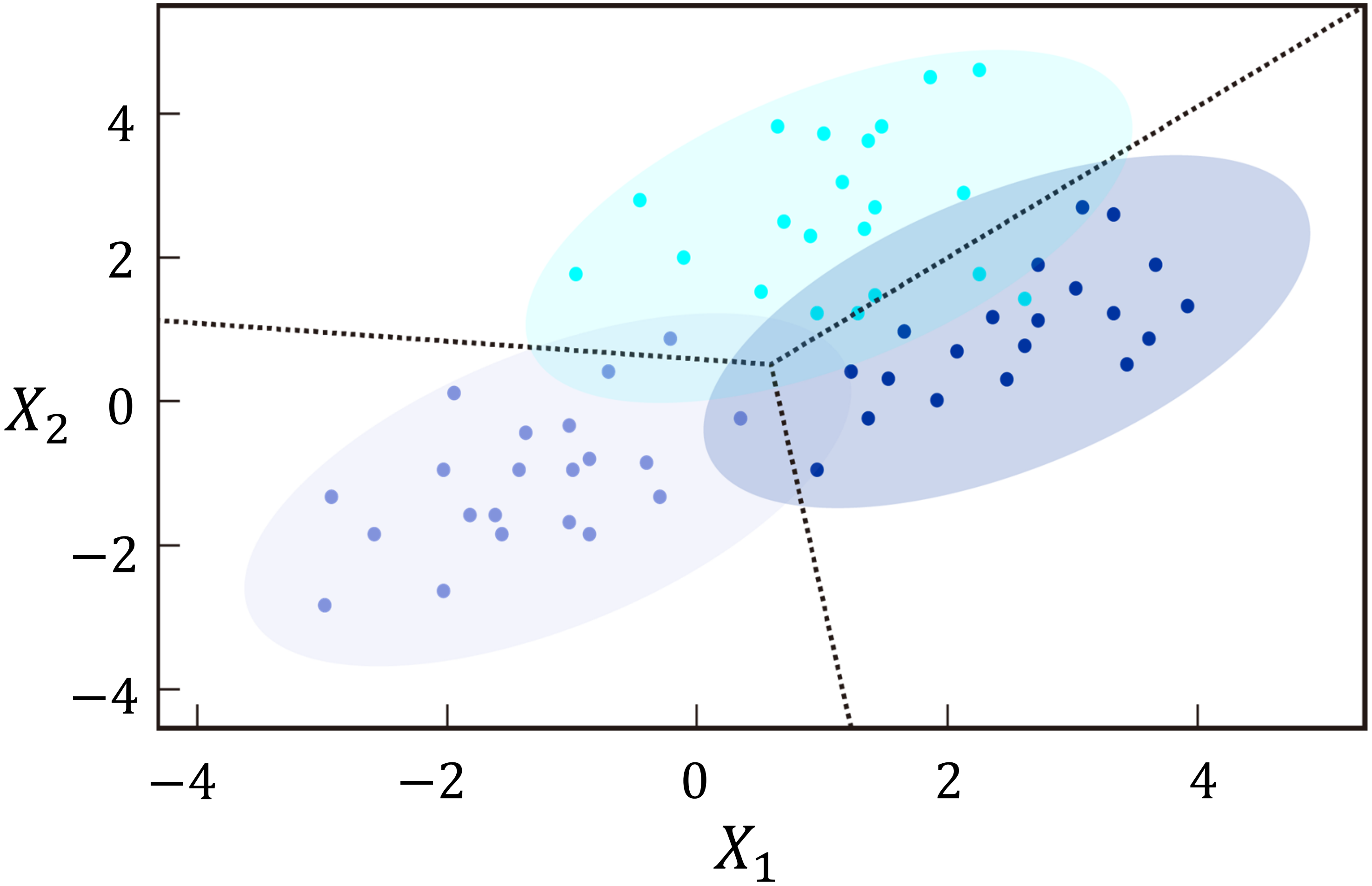
(a) Suppose W and T are uncorrelated. What is $P[I]$? Hint: Draw a tree diagram for the experiment.

(b) Now suppose W and T have correlation coefficient $\rho = 1/\sqrt{2}$. Find the conditional probability $P[I|T = t]$ that a person is declared ill given that the person's temperature is $T = t$.

Yates, Probability and Stochastic Process 3rd, Wiley, 2014

이변량 정규 분포

- **결합정규분포 in LDA**
  - LDA (Linear Discriminant Analysis) 는 분류에 있어서 널리 쓰이는 기계학습 기법 중에 하나이다.
  - 각 클래스에서의 샘플들은 각각의 평균을 가진 결합 정규 분포에서 나온 것으로 취급된다.



Summary

- **조건부 확률 변수 $X|Y$**
  - $P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$ 또는 $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
- **Bayes' 정리**
  - $P_{X|Y}(x|y) = \frac{P_{Y|X}(y|x)P_X(x)}{\sum_u P_{Y|X}(y|u)P_X(u)}$ 또는 $f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|u)f_X(u)du}$
- **독립**
  - $X \perp Y \leftrightarrow P_{X,Y}(x,y) = P_X(x)P_Y(y)$, or $f_{X,Y}(x,y) = f_X(x)f_Y(y)$
  - $X \perp Y \rightarrow \sigma_{X,Y} = 0$, 하지만 $X \perp Y \nleftarrow \sigma_{X,Y} = 0$ 이 일반적이다
- 이변량 정규분포**(Bivariate Gaussian distribution)**
  - $\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{bmatrix}\right) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - when $f_{X,Y}(x,y) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}\left(\begin{bmatrix} x \\ y \end{bmatrix} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\begin{bmatrix} x \\ y \end{bmatrix} - \boldsymbol{\mu}\right)\right]$

다변수 확률 분포

무작위 벡터(Random vector)

- ***X*: 하나의 무작위 변수, 단 변량 분포**
- ***X*, *Y*: 두 개의 무작위 변수, 2변수 분포**
- **X_1, X_2, …, X_n: 여러 개의 무작위 변수, 다변량 분포**
- **무작위 벡터(Random vector)**
  - 유한 개의 확률 변수의 벡터; $\mathbf{X} = [X_1, X_2, \cdots, X_n]^T$
  - 임의의 벡터는 하나의 확률 변수와 두 개의 확률 변수의 일반화입니다.

- 벡터 주석
  - $\mathbf{x} = \mathbf{y} \ \rightarrow x_1 = y_1, x_2 = y_2, \dots, x_n = y_n$
  - $\mathbf{x} < \mathbf{y} \ \rightarrow x_1 < y_1, x_2 < y_2, \dots, x_n < y_n$

다 변수 확률 분포

길이가 n인 무작위 벡터 $\mathbf{X} = [X_1, X_2, \cdots, X_n]^T$ and a vector $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$

- 다 변수 **CDF**
  - $F_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = F_{\mathbf{X}}(\mathbf{x}) = \Pr[\mathbf{X} \le \mathbf{x}] = \Pr[X_1 \le x_1, X_2 \le x_2, \cdots, X_n \le x_n]$
- 다 변수 **PMF**
  - $P_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = P_{\mathbf{X}}(\mathbf{x}) = \Pr[\mathbf{X} = \mathbf{x}] = \Pr[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n]$
  - $F_{\mathbf{X}}(\mathbf{x}) = \sum_{-\infty}^{x_1} \cdots \sum_{-\infty}^{x_n} P_{\mathbf{X}}(x_1, \cdots, x_n) = \sum_{-\infty}^{\mathbf{x}} P_{\mathbf{X}}(\mathbf{u})$
- 다 변수 **PDF**
  - $f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = f_{\mathbf{X}}(\mathbf{x}) = \frac{\Pr[\mathbf{x}<\mathbf{X}\le\mathbf{x}+d\mathbf{x}]}{dx_1 dx_2 \cdots dx_n} = \frac{\Pr[x_1<X\le x_1+dx_1,\cdots, x_n<X_n\le x_n+dx_n]}{dx_1 dx_2 \cdots dx_n} = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \cdots \partial x_n}$
  - $F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(u_1, \cdots, u_n)\, du_1 \cdots du_n = \int_{-\infty}^{\mathbf{x}} f_{\mathbf{X}}(\mathbf{u}) d\mathbf{u}$
- $\mathbf{X} = [(X_1, \cdots, X_k), (X_{k+1}, \cdots, X_n)]^T = \left[\mathbf{X}_\mathbf{a}^T, \mathbf{X}_\mathbf{b}^T\right]^T$ **일때,**
  - **한계 분포(Marginal distribution):** $F_{\mathbf{X}_\mathbf{a}}(\mathbf{x}_\mathbf{a}) = \int_{-\infty}^{\infty} F_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}_\mathbf{b}$
  - **조건부 PDF:** $\mathbf{X}_\mathbf{a}|\mathbf{X}_\mathbf{b} \sim f_{X_1,\cdots,X_k|X_{k+1},\cdots,X_n}(x_1, \cdots, x_k | x_{k+1}, \cdots, x_n) = f_{\mathbf{X}_\mathbf{a}|\mathbf{X}_\mathbf{b}}(\mathbf{x}_\mathbf{a}|\mathbf{x}_\mathbf{b}) = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{X}_\mathbf{b}}(\mathbf{x}_\mathbf{b})}$

독립

- **랜덤변수 X_1, X_2, …, X_n 가 독립이기 위한 필요충분 조건**
  - $P_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = P_{X_1}(x_1)P_{X_2}(x_2)\cdots P_{X_n}(x_n)$, or
  - $f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n)$

- **독립적이고 동일하게 분산된(iid) 확률 변수**
  - X_i's are iid 과 '서로 독립적이며 동일한 분포를 갖는 경우'는 서로 필요충분 조건입니다, i.e.
  - $P_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} P_X(x_i)$, or $f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f_X(x_i)$

- **iid 무작위 변수의 예제**
  - N개의 동전을 뒤집는 것은 *n* iid Bernoulli 확률 변수로 모델링 할수 있다
  - N 분자의 운동은 또한 iid확률로 모델링 될 수 있다

기대값, 분산 및 공분산

- **길이가 n인 랜덤 벡터** $\mathbf{X} = [X_1, X_2, \cdots, X_n]^T$**,**

- **기대값 벡터:** $\mathrm{E}[\mathbf{X}] = \boldsymbol{\mu}_{\mathbf{X}} = \left[\mathrm{E}[X_1], \mathrm{E}[X_2], \cdots, \mathrm{E}[X_n]\right]^{\mathrm{T}}$**.**

- **공분산 행렬:** $\mathbf{C}_{\mathbf{X}} = \mathrm{E}[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T] = \begin{pmatrix} \sigma_{X_1}^2 & \cdots & \sigma_{X_1,X_n} \\ \vdots & \ddots & \vdots \\ \sigma_{X_1,X_n} & \cdots & \sigma_{X_n}^2 \end{pmatrix}$

- **공분산 벡터에 대한 속성**
  - (1) $(\mathbf{C}_{\mathbf{X}})_{ij} = \sigma_{X_i,X_j}$
  - (2) $\mathbf{C}_{\mathbf{X}} = \mathrm{E}[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{X}}^T$
  - (3) $\mathbf{C}_{\mathbf{X}}^T = \mathbf{C}_{\mathbf{X}}$: 대칭
  - (4) For any vector x, $\mathbf{x}\mathbf{C}_{\mathbf{X}}\mathbf{x}^{\boldsymbol{T}} \geq 0$: semi-positive definite
  - (5) $\det(\mathbf{C}_{\mathbf{X}}) \geq 0$

많은 랜덤 변수의 합

- $W_n = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$ 라 하면
  - $\mathrm{E}[W_n] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n] = \sum_{i=1}^{n} \mathrm{E}[X_i]$
  - $$\begin{aligned}\mathrm{Var}[W_n] &= \mathrm{E}[(\textstyle\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i)^2] = \mathrm{E}[(\sum_{i=1}^{n}(X_i - \mu_i))^2] \\ &= \mathrm{E}\left[\textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \mu_i)\left(X_j - \mu_j\right)\right] \\ &= \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}\left[X_i, X_j\right] = \sum_{i=j} \mathrm{Cov}\left[X_i, X_j\right] + \sum_{i \neq j} \mathrm{Cov}\left[X_i, X_j\right] \\ &= \textstyle\sum_{i=1}^{n} \mathrm{Var}[X_i] + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{Cov}\left[X_i, X_j\right]\end{aligned}$$
- **X_i's 가 독립이면,** $\mathrm{Var}[W_n] = \sum_{i=1}^{n} \mathrm{Var}[X_i]$**.**
- **X_i's 가 iid이면,** $\mathrm{Var}[W_n] = n\mathrm{Var}[X]$**.**

- **특별한 경우**
  - (1) $X_i \sim \mathrm{Poi}(\lambda_i)$ and X_i's are indep. → $\sum_{i=1}^{n} X_i \sim \mathrm{Poi}(\sum_{i=1}^{n} \lambda_i)$
  - (2) $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ and X_i's are indep. → $\sum_{i=1}^{n} X_i \sim \mathrm{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$

iid 랜덤 변수의 평균

- $X_i \sim X$ 인 X_i**'s가 iid 무작위 변수 일 때,** $W_n = \sum_{i=1}^{n} \frac{X_i}{n}$ **라 하자**
- $\mathrm{E}[W_n] = \sum_{i=1}^{n} \mathrm{E}\left[\frac{X_i}{n}\right] = \mathrm{E}[X]$
  - $\mathrm{Var}[W_n] = \sum_{i=1}^{n} \mathrm{Var}\left[\frac{X_i}{n}\right] = \sum_{i=1}^{n} \frac{1}{n^2} \mathrm{Var}[X_i] = \frac{1}{n} \mathrm{Var}[X]$
  - N이 클 때 Wn의 분산은 0에 가까워 진다. 이 의미는 E[X]를 평균을 통해 매우 정확히 측정할 수 있다는 것을 의미한다.
- **많은 수의 법칙: For iid** X_i**'s,** $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathrm{E}[X]$**.**
- **예제**
  - (1) 우리는 동전의 앞면이 나올 확률 p가 균등하게 나오지 않는 동전을 가지고 있습니다. 하지만, p를 알지 못합니다. 우리는 이제 동전을 n번 던지면서 p를 예측해야 합니다. 각각의 동전 던짐은 iid입니까? 10회에 비해 20회 뒤집을 때 p를 얼마나 정확하게 예측할 수 있습니까?
  - (2) 두 개의 설문조사 회사가 대통령 지지율에 대해 보고합니다. A는 100명의 사람들을, B는 500명의 사람들을 조사합니다. 어떤 회사가 얼만큼 더 정확하게 조사합니까?

중심 극한 정리

- X_i**들은** $\mathrm{E}[X] = \mu$ **과** $\mathrm{Var}[X] = \sigma^2$**에 iid이다.**
  - $\mathrm{E}[\sum_{i=1}^{n} X_i] = n\mu$, $\mathrm{Var}[\sum_{i=1}^{n} X_i] = n\sigma^2$
  - $Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}$, $\mathrm{E}[Z_n] = 0$,$\mathrm{Var}[Z_n] = 1$.
- **중심 극한 정리<Central Limit Theorem (CLT)>**
  - n이 무한대에 가까워 지면, X의 분포에 상관없이 $\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z)$ 또는 $\lim_{n\to\infty} Z_n \sim N(0,1)$이다.
  - CLT는 확률 분포의 왕인 보편적 분포를 보편적으로 만듭니다.
- **Example**
  - 예를 들어, 물의 온도는 매번 측정하는 온도가 다르기 때문에 무작위적인 변수 T로 나타낼 수 있습니다. 물리학에서 물의 온도는 물 분자의 활동에 의해 결정됩니다. X_i 는 온도에 대한 하나의 물 분자의 활성을 설명하는 무작위 변수라 하자. 그러면, T 는 많은 분자와 많은 분자 X_i's의 합으로 표현될 수 있습니다. $T = X_1 + X_2 + \cdots + X_n$
  - X_i 의 분포는 매우 어렵고 양자 역학에 따라 복잡합니다. 그러나, n이 매우 큰 경우(물 분자의 수와 같이), X_i 분포에 관계없이 T는 정규 분포를 따른다. 이것은 우리가 중심 극한 정리라고 부르고 있는 것으로 수학적으로 증명되어 있다.

정리

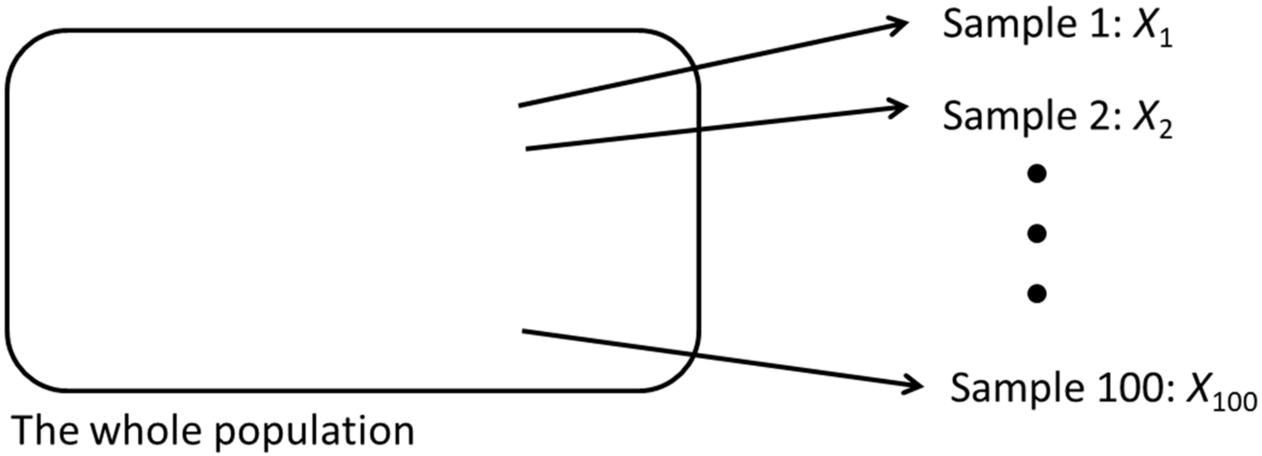
- **다 변수 cdf/pdf/pmf 는 이변수의 확장입니다.**
- **Iid RVs: X_1, X_2, …, X_n 는 서로 독립적이고 같은 분포를 갖는다.**
- **무작위 벡터는 cdf/pdf/pmf에 의해 완전히 정의되지만, 개대 벡터와 공분산 행렬에 의해 간단히 특징지어 질 수 있다.**
- **Gaussian random vector (GRV)는 이 변수량 가우시아 확률 변수의 일반화입니다.**
- **X 와 Y 가 독립일 때,** $\boldsymbol{f_{X+Y}(w) = (f_X * f_Y)(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w-x)dx}$
- $W_n = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$ 일 때,
- $\mathrm{E}[W_n] = \sum_{i=1}^{n} \mathrm{E}[X_i]$ **이고** $\mathrm{Var}[W_n] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{Cov}\left[X_i, X_j\right]$.
- **X_i's 가 iid일 때,** $\mathrm{E}[W_n] = n\mathrm{E}[X]$ **이고** $\mathrm{Var}[W_n] = n\mathrm{Var}[X]$.
- **큰 수의 법칙: iid X_i's에 대해,** $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathrm{E}[X]$.
- **중심 극한 정리:** X_i**'s 가 iid 이며** $\mathrm{E}[X] = \mu$ **이고** $\mathrm{Var}[X] = \sigma^2$ 일때, $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1)$

통계

랜덤 샘플링

- **한국 성인의 평균 신장(>= 18세)를 알고 싶다고 할때....**
  - 모든 성인의 신장을 측정하고 평균을 취하거나,
  - 무작위로 n명(~100명)의 성인을 선택하고 평균을 취함

- **임의 추출 (Random Sampling)**

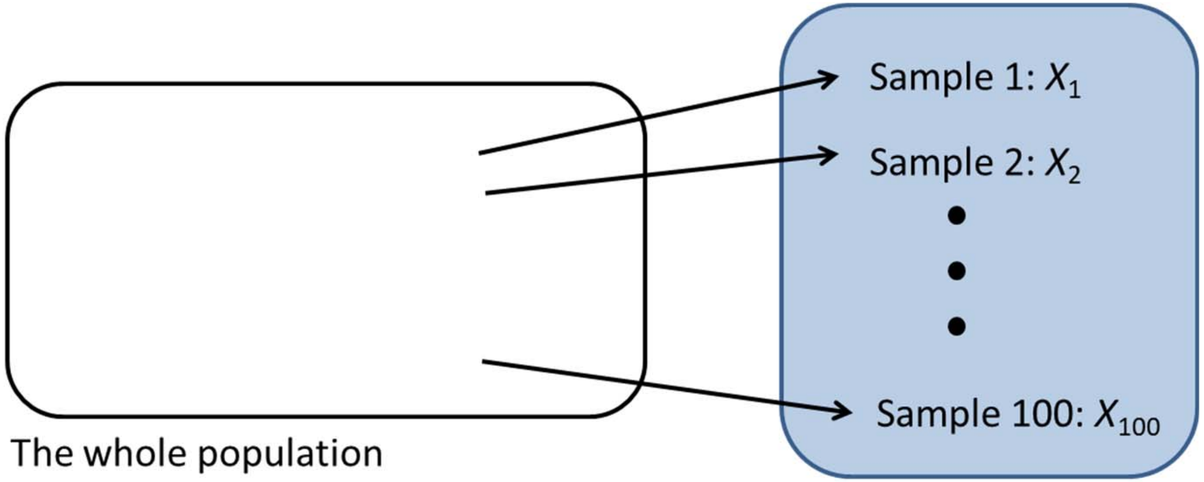


Quiz
2개의 당첨 티켓과 8개의 꽝 티켓이 상자에 있다. 10명의 사람이 순서대로 티켓을 뽑아서 갖는다. 뽑힌 티켓은 다시 상자에 넣지 않는다. 이때 두 번째 뽑는 사람과 다섯 번째 뽑는 사람 중에서 당첨 티켓을 뽑을 확률이 높은 사람은 누구인가?

  - Xi 들은 동일한 분포를 갖는가? 독립적인가?
  - 임의표본 (Random Sample): 같은 모집단으로부터 임의로 추출된 데이터

통계

- **통계는 과학적 방법으로 관찰되 무작위 표본에서 전체 인구에 대한 무언가를 말하기 위한 도구**
- **예제**
  - 전체 확률은 *N*(0,1)를 따른다.
  - 5개의 무작위 샘플: 1.40, 0.44, -1.90, -1.05, -0.04
  - 관찰된 샘플의 평균은 -0.23이지만, 실제 평균은 0.
  - 관찰 된 -0.23이 실제로 0을 나타낼 수 있습니까?
- **통계는 관찰 된 무작위 표본으로부터 계산된 값입니다.**
  - $T = g(X_1, X_2, \ldots, X_n)$



표본 평균

- **실제 평균(모집단 평균)**
  - $E[X]$: 모집단의 평균
- **표본 평균**
  - 관찰 된 샘플의 평균
  - 진정한 의미를 나타내지만, 평등하지는 않다.
  - n이 클 때, 정규 분포에 가깝다.

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

- **표본 평균의 기대값 및 분산**
  - $E[\bar{X}] = E[X] = \mu$
  - $\mathrm{Var}[\bar{X}] = \mathrm{Var}[X]/n = \sigma^2/n$

표본 분산

- **실제 분산 (모집단 분산)**
  - $\mathrm{Var}[X]$: 전체 모집단의 분산

- **표본 분산**
  - 관찰된 샘플의 분산
  - 실제 분산을 나타내지만, 동일하지는 않다.

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

- 표본 분산의 평균 및 분산
  - $E[S^2] = \mathrm{Var}[X] = \sigma^2$; n-1로 나누는 이유이다.
  - $\mathrm{Var}[S^2] =?$

통계에서의 핵심 확률 분포

- **모집단 분포**
  - 모집단의 모든 요소의 빈도 분포

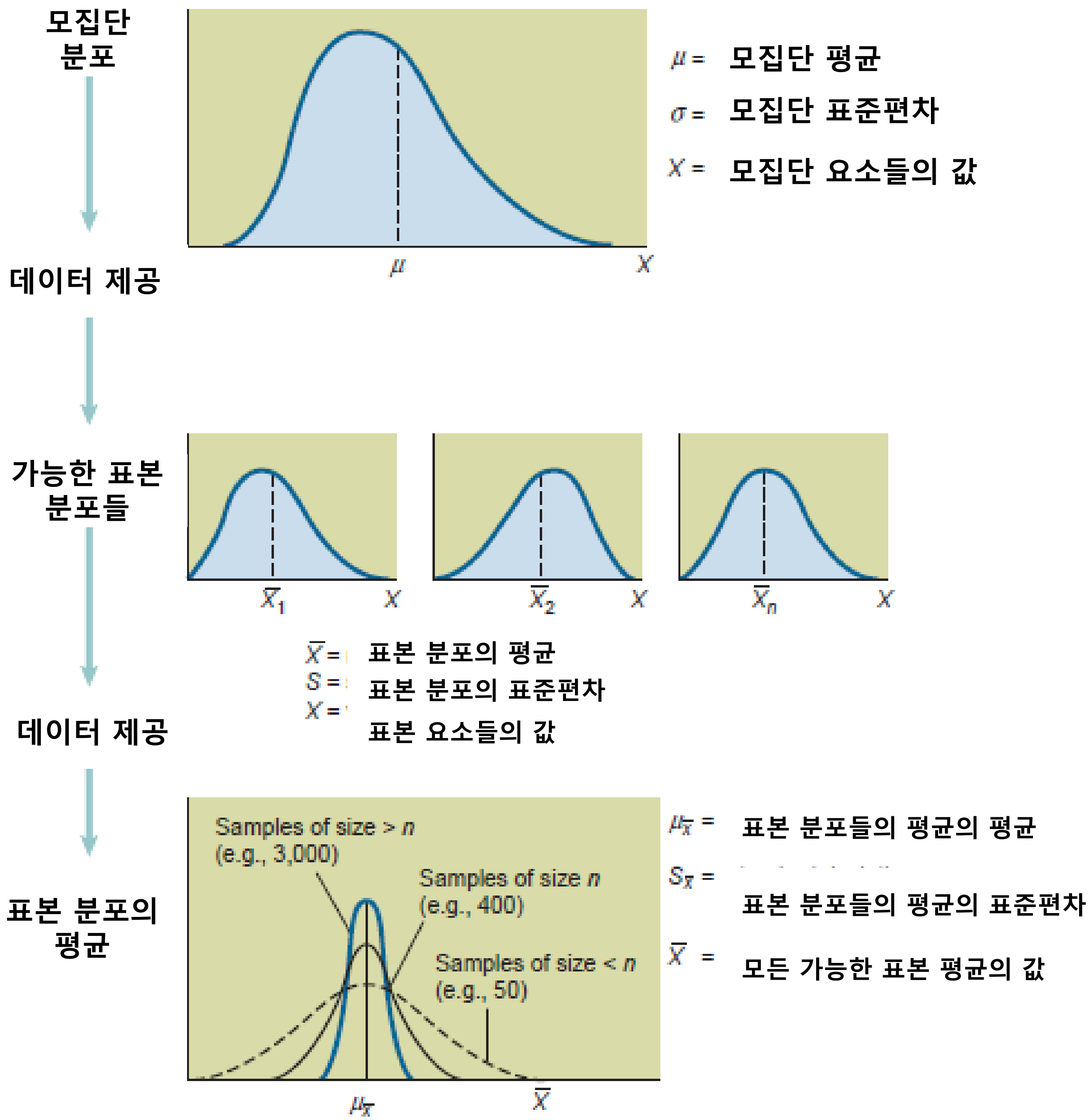
- **샘플 분포**
  - 표본의 모든 요소의 빈도 분포

- **표본 평균이 표본 분포**
  - 주어진 표본에서 많은 표본 평균의 빈도 분포
  - 사실상 모든 모집단에서 추출한 표본 분포는 CTL에 의해 평균이 μ, 표준 편차가 $\sigma/\sqrt{n}$ (표준 오차라고도 함)과 동일한 정규 분포에 접근합니다.

  - **표준 오차**: 표본 분포의 표준 편차

Key Distribution

- 요약



표본 평균의 신뢰구간

- **표본 평균이 모집단 평균에 얼마나 가까운지에 대한 신뢰도**
  - 신뢰구간은 모든 통계에 적용될 수 있는 일반적인 용어이지만, 우리는 그 중 표본 평균에 집중 할 것임.

- **신뢰구간은 표본 평균이 실제 평균에 얼마나 더 가까운지를 과학적으로 나타냄.**

- **표본크기(n)과 표본 표준편차(σ)에 따라 달라짐.**
  - 모집단의 표준편차를 모르는 경우, 표본 표준편차 S를 사용함.
  - 추정된 표준편차는 $S/\sqrt{n}$ 로 나타냄.

표본 평균의 신뢰구간

- **표본 평균 $\bar{X}$ 는 실제 평균 μ의 일반적인 추정치임.**
  - $E[\bar{X}] = \mu$: 불편의/비편향적
  - $Var[\bar{X}] = E[(\bar{X} - \mu)^2] = \sigma^2/n$: 이 또한 MSE임.
  - $\bar{X}$의 추정 분산은 S^2/n로 표현됨.
  - 표준 편차는 $S/\sqrt{n}$로 표현됨.

- **(*p* x 100)% 신뢰구간**

$$p = \Pr\left[\bar{X} - k\frac{S}{\sqrt{n}} < \mu < \bar{X} + k\frac{S}{\sqrt{n}}\right] = \Pr\left[-k < \frac{\bar{X} - \mu}{S/\sqrt{n}} < k\right]$$

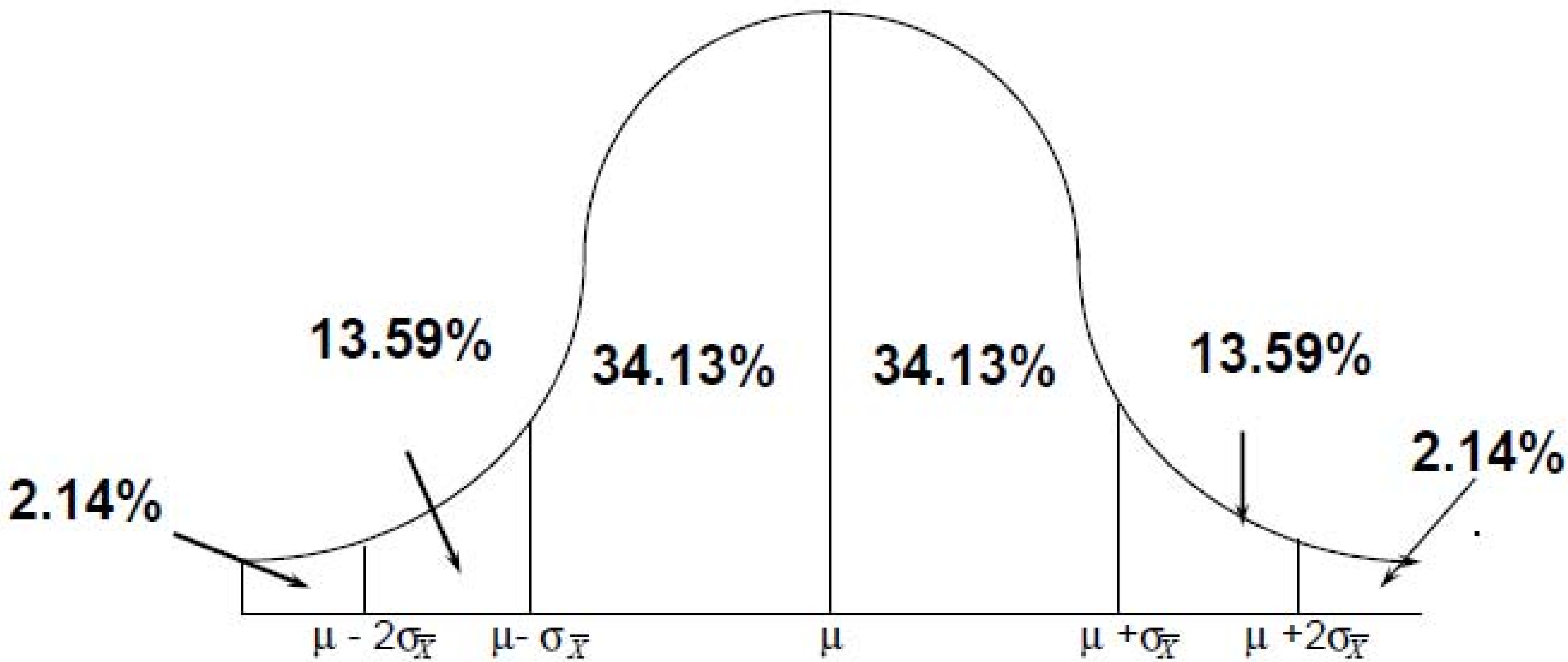
  - 실제 평균이 신뢰구간 안에 있을 확률은 p.
  - *n*이 충분히 크다면 S→σ 이고 $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ ~ N(0,1) 이라 할 수 있다.

표본 평균의 신뢰구간

- **(px100)% 신뢰구간의 다른 관점**

$$p = \Pr\left[\bar{X} - k\frac{S}{\sqrt{n}} < \mu < \bar{X} + k\frac{S}{\sqrt{n}}\right] = \Pr\left[\mu - k\frac{S}{\sqrt{n}} < \bar{X} < \mu + k\frac{S}{\sqrt{n}}\right]$$

Distribution of the sample mean



예제

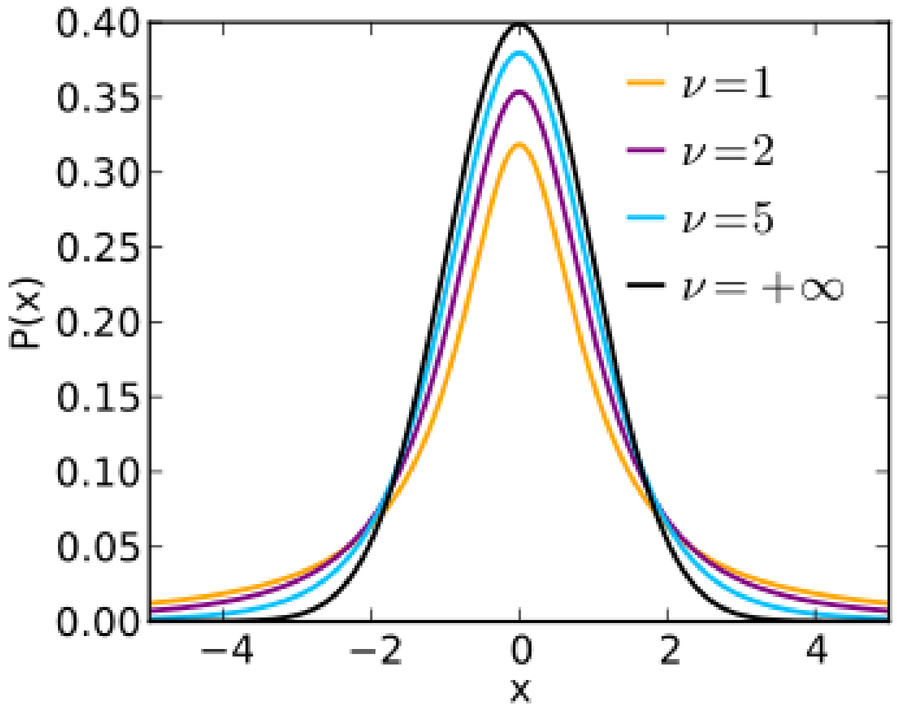
- **임의의 표본** $1.40, 0.44, -1.90, -1.05, -0.04$
  - $\bar{X} = -0.23,\ S = 1.28,\ S/\sqrt{n} = 0.57$
  - 실제 평균이 -0.80 과 0.24 사이일 확률은 68%
  - 95% 신뢰구간은 -1.35 ~ 0.88

- **표본 개수와 신뢰구간 사이의 관계**

n	95% confidence interval:		interval width:
5	$475 \pm 1.96\left(100/\sqrt{5}\right)$	= 387.35 to 562.65	175.30
30	$475 \pm 1.96\left(100/\sqrt{30}\right)$	= 439.22 to 510.78	71.56
100	$475 \pm 1.96\left(100/\sqrt{100}\right)$	= 455.40 to 494.60	39.20
500	$475 \pm 1.96\left(100/\sqrt{500}\right)$	= 466.23 to 483.76	17.53

T-검정

- n이 충분히 크지 않다면, $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ 이 N(0,1)을 따르지 않는다.
- 전체 모집단이 정규분포를 따른다 가정하면, ***t*** 는 ***n*–1**의 자유도를 갖는 스튜던트 t-분포를 따른다.
- 전체 모집단이 정규분포를 따르지 않더라도 t-분포는 ***t*값**에 대한 좋은 근사치이다.
- $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ 이 t-통계량이다.



$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

https://en.wikipedia.org/wiki/Student%27s_t-distribution

요약

- **통계학은 과학적 방법으로 관찰 된 임의의 표본으로부터, 전체 모집단에 대해 '무엇인가'를 말하기 위한 도구이다.**
  - 여기서 '무엇인가'는 보통 평균을 의미함

- **임의의 표본 추출: 동일한 모집단에서 독립적이고, 동일한 분포를 가지는 표본을 임의로 선택함.**
- **표본 평균:** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- **신뢰 구간: 추정치가 실제 값에 얼마나 가까운지**

- **T-통계량**
  - $t = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim T(n-1)$
  - $\left(\Pr\left[-k < \frac{\bar{X}-\mu}{S/\sqrt{n}} < k\right] \times 100\right)\%$ 신뢰구간은 $\left(\bar{X} - k\frac{S}{\sqrt{n}}, \bar{X} + k\frac{S}{\sqrt{n}}\right)$
    - 실제 평균이 구간 안에 있을 확률

가설 검정

- **다음을 테스트하길 원한다고 가정,**
  - 소금물이 어는 온도가 0인지에 대해 테스트하길 원한다.
- **우리는 모집단 측정 값, 예를 들어 소금물 온도의 모집단 평균과 같은 추정치가 필요함.**
- **모집단 특징징은 관측되지 않음.**
- **따라서 우리는 몇몇 표본 관측치가 필요함.**
- **표본 통계에 기초하여, 모집단 측정이 필요한 테스트를 충족하는지 하지 못하는지 판단 해야함.**

가설 검정

- **전체 모집단에 대해 가설을 하나 세울 수 있음.**
- **임의의 관측된 표본에 대해 우리가 세운 가설에 대한 테스트를 수행할 수 있다.**

- **예시: 소금물이 어는 온도**
  - 우리의 가설: 소금물 어는 온도는 0이 아니다.
  - 전체 모집단: 소금물이 언다는 모든 현상
  - 임의의 표본: 소금물 어는 점에 대한 반복된 실험.

- **가설 검증: 가설의 통계적 검증**
  - **귀무가설 (Null Hypothesis)** H_0: $\mu = 0$, 소금물의 어는 점은 0도이다.
    - **귀무분포 (Null Distribution)**: 귀무가설이 참일 때 나타나는 전체 모집단의 분포.
  - **대립 가설 (Alternative Hypothesis)** H_A: $\mu \neq 0$, 어는점이 0도가 아니다.

- **귀무가설에 대한 증거를 바탕으로 귀무가설을 기각하거나 기각하지 않음.**

가설 검증의 단계

H_0 과 H_1 을 공식화

적절한 통계량 (z, t, F, 카이-제곱 등)를 선택

유의 수준을 선택

데이터 수집 및 검정 통계량 계산

검정 통계량과 관련된 확률 결정(p-값)

유의 수준과 비교

검정 통계량의 중요 값 결정

계산된 검정 통계량이 기각 영역에 속하는지 확인

귀무 가설 H_0를 기각 혹은 기각하지 않음

유의 수준

- 1종 오류는 실제 결과가 사실이지만 표본 결과가 귀무가설을 기각하도록 유도할 때 발생함.

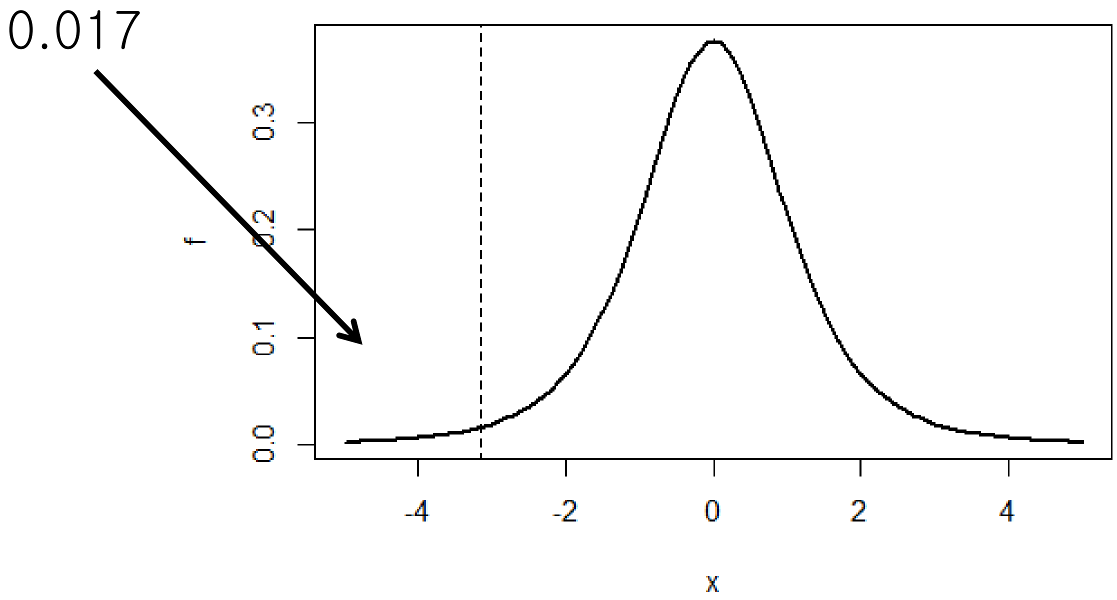
- 1종 오류의 확률 (α) 을 유의 수준이라고 함.

- 일반적으로 1%, 5%, 10% 유의수준을 사용함.

- 가설검정에 사용되는 유의 수준 (α) 가 높을수록, 실제 결과가 참이지만 귀무 가설을 기각할 확률이 높아짐.

예제

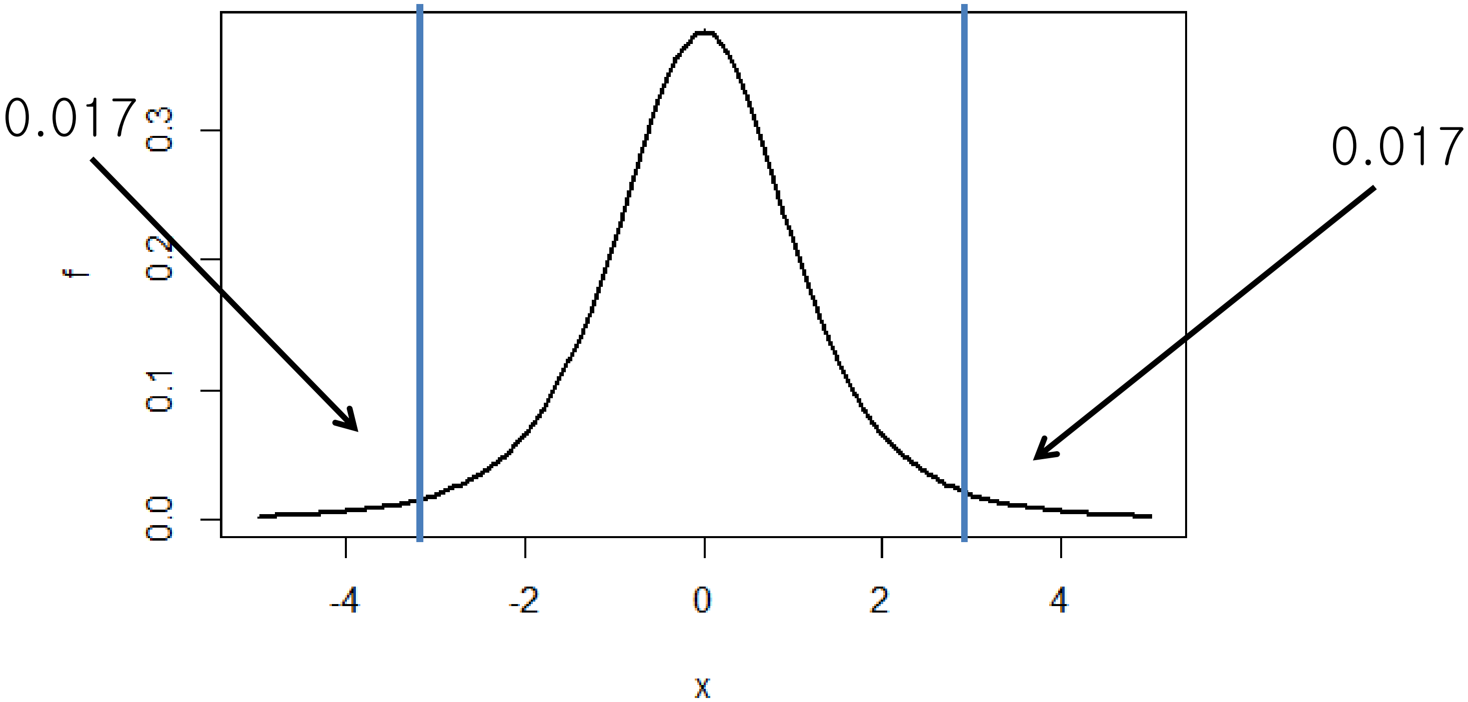
- **관측된 소금물 동결온도**
  - $X = (-0.31, -0.67, -0.61, -2.07, -1.31), \bar{X} = -0.99, S = 0.70.$
  - $\bar{X} < 0.$ $\mu \neq 0$ 이거나 $\mu = 0$ 인동안의 무작위 효과 때문이 아닐까?
  - $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$.를 사용.
- ***H_0*: $\mu = 0$ vs. *H_A*: $\mu \neq 0$.**
- ***t* ~ *T*(4)의 귀무분포, t 통계량은 –3.15.**



- **t 통계량 -3.15에서의 확률은 0.017.**

P-값

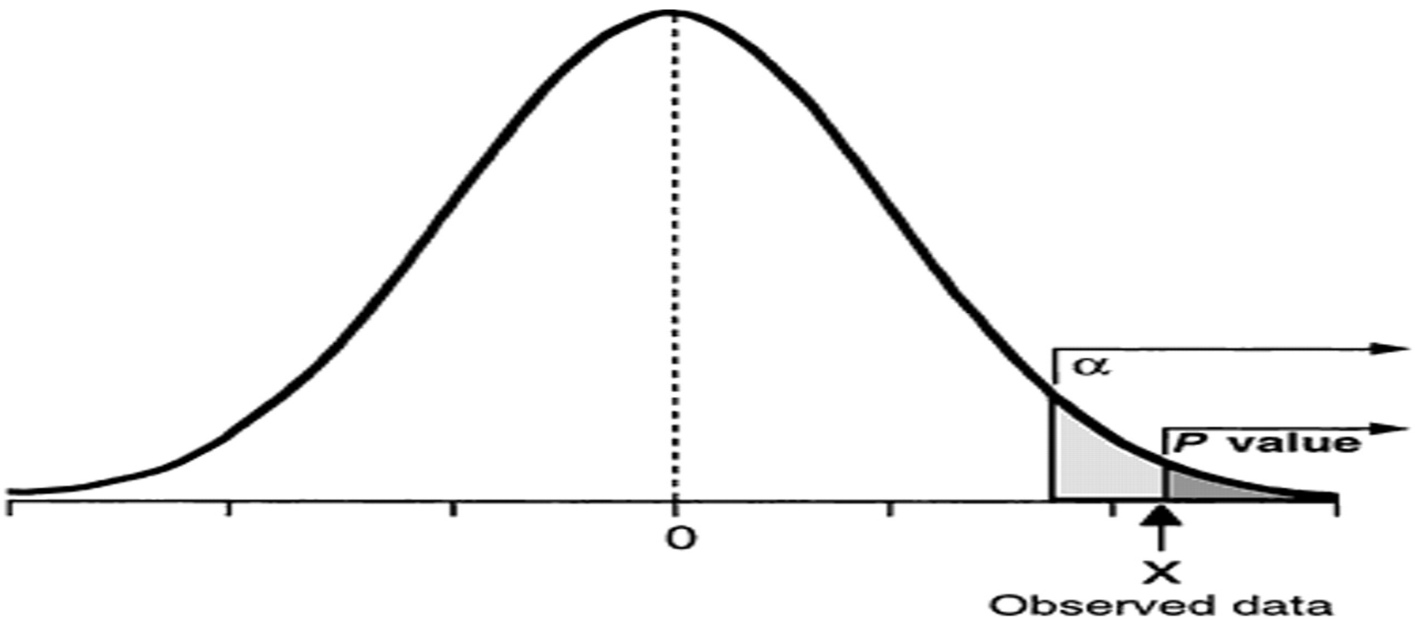
- **P-값: 관측된 통계량을 초과할 확률(대립가설에 대해)은 귀무 분포에서 비롯됨.**



- **이 예제에서 p = 0.034, < -3.15 와 > 3.15 모두 H_A: $\mu \neq 0$.에 멀리 떨어져 있기 때문.**

유의수준

- **유의수준**
  - 귀무가설을 기각하거나 기각하지 않을 p-값의 임계 값.
  - 전통적으로 1%와 5%를 사용.
- **이 예제에서 , p=0.034 혹은 3.4%는 두가지 가능한 결론이 존재.**
  - 5% 유의수준에서 소금물의 어는점은 0과 유의한 차이를 보이고, H_0 를 기각할 수 있다.
  - 1% 유의수준에서 소금물의 어는점은 0과 유의한 차이를 보이지 않고, H_0 를 기각할 수 없다.



T-검정

- T-통계량을 사용한 예제. (스튜던트) t-검정이라 불림.
- R을 통해 간단하게 수행할 수 있음.

```
> x = c(-0.31,-0.67,-0.61,-2.07,-1.31);
> t.test(x)

 One Sample t-test

data:  x
t = -3.1608, df = 4, p-value = 0.03416
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.8671291 -0.1208709
sample estimates:
mean of x
   -0.994
```

- One-sample test: 평균이 특정한 값과 다른지 여부를 검정.
- Two-sample test: 두 표본 그룹의 평균 차이를 검정.

Appendix

References

- Probability and Stochastic Processes: A Friendly Introduction to Electrical and Computer Engineers (3rd edition), Yates and Goodman, Wiley
- Probability, Statistics, and Random Processes for Electrical Engineering (3rd edition), Leon-Garcia, Pearson International Edition.
- An Introduction to Statistical Learning with Applications in R, James, Witten, Hastie, Tibshirani, Springer
- Pattern Recognition and Machine Learning, Bishop, Springer

Lecturer

- **Junhee Seok, PhD**
  - Assistant Professor, Electrical Engineering, Korea University
  - Director of Mirae Asset AI Fintech Research Center
  - Education
    - BS, Electrical Engineering, KAIST, 2001
    - PhD, Electrical Engineering, Stanford University, 2011
  - Professional Experiences
    - Postdoctoral Fellow, Statistics, Stanford University
    - Assistant Professor, HBMI, Northwestern University
  - Research Area
    - Big data analytics, Machine Learning, AI
    - Biomedicine, Finance, Climate, IoT, Materials, and etc.