제 1장 서론

통계학 : 자료의 수집과정을 설계하고, 자료를 요약하고 해석하여 결론을 이끌어 내거나 일반화하는 전체적인 원리와 방법론

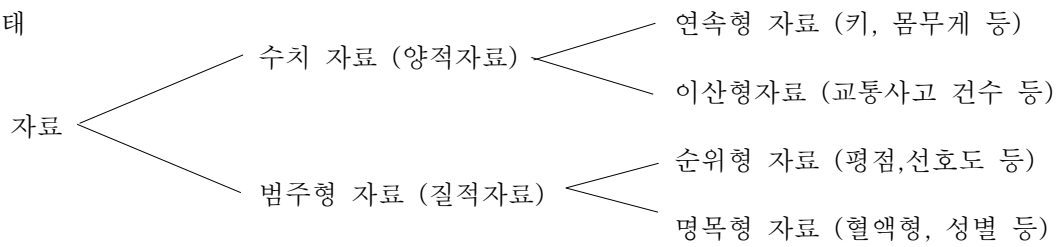
모집단 (Population) : 얻고자하는 정보와 관련있는 모든 개체로부터 얻을 수 있는 모든 관측값들의 집합

표본 (Sample) : 모집단의 일부분으로, 원하는 정보를 얻기 위해 수행한 관측과정을 통하여 실제로 얻어진 관측결과의 집합

통계학의 목표 : 1. 관측값으로부터 효율적인 추론을 할 수 있도록 표본추출의 과정과 범위를 설계
2. 표본에 포함되어 있는 정보를 분석하여 모집단에 관하여 추론을 한다. 이때 추론에 수반되는 불확실성도 측정한다.

제 2장 표와 그림을 통한 자료의 요약

1. 자료의 형태



2. 범주형 자료의 요약 & 이산형 자료의 요약

	도수분포표	원형그래프	막대그래프	파레토그림
용어	도수:관측값의 개수 상대도수:도수/전체개수			
특징	범주와 그에 대응하는 도수, 상대도수로 나열된 표	원을 상대도수에 비례하여 중심각을 나누어 조각을 나눈 형태	도수의 크기를 막대로 그려 나타냄. 상대도수도 포함가능	상대도수 큰 순서로 범주를 왼쪽부터 오른쪽으로 배열, 누적상대도수를 각 범주의 막대 위 중앙에 표시하고 점들을 연결한 그림.
장점		전체에서 각 범주가 차지하는 비율을 파악하기 용이	각 범주 간 도수 비교가 용이하다.	문제 파악의 수단으로 자주 사용 상대도수 증가정도, 큰 도수의 범주들이 차지하는 비율 파악가능
단점		범주 간 도수비교 및 도수크기의 차이 파악 어려움	전체에서 차지하는 비율 파악은 원형그래프가 용이	범주의 순서가 의미가 있는 자료에는 유용하지 않음

*원형그래프의 각도는 $360^\circ \times$ (상대도수) 로 얻어진다.
*도수분포표의 상대도수의 합은 1임을 기억하자.
*이산형은 범주형 자료의 기법을 이용하여 분석가능하다.

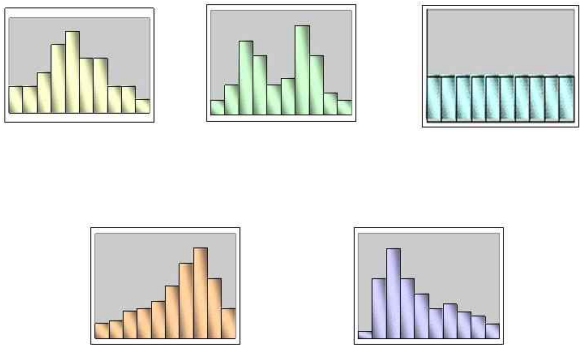
3. 연속형 자료의 요약

	연속형 자료를 위한 도수분포표의 작성방법
1. 자료의 범위	자료에서 최댓값과 최솟값을 찾아 범위 (range:최댓값-최솟값) 를 구한다.
2. 계급구간의 폭	계급의 개수가 5개에서 15개 정도가 되도록 대략 정한다. 계급구간의 폭=자료의 범위/구간의 개수 보다 조금 큰 값 으로한다.
3. 계급구간	모든 관측값을 포함하도록 각 계급구간의 경계점을 구한다.
4. 도수	각 계급구간에 속하는 관측값의 개수를 세어 계급의 도수를 구한다.
5. 상대도수	계급의 도수/전체 관측값의 개수

*계급구간의 폭이 너무 크면 자료가 간략히 요약되지만 많은 정보를 잃어버림
*계급구간의 폭이 너무 작으면 경향 파악이 어려워 자료의 요약으로서 도수분포표의 의미가 약해짐
*계급의 개수를 꼭 5~15개 안에서 정할 필요는 없음
*전체 계급구간의 시작값은 자료의 범위가 전체 계급구간의 중간에 오도록 정함.
*동일하지 않은 계급구간의 폭을 사용함으로써 도수분포표에 더 많은 정보를 포함시키고 명확하게 할 수 있음

점도표	히스토그램	도수다각형	줄기-잎 그림
관측값의 개수가 상대적으로 적은 경우에 이용	상대도수 막대그래프와 같으나 x축이 구간으로 이루어진 그래프.	히스토그램 각 막대의 상단의 중앙점을 직선연결한 그래프.	수직선을 기준으로 왼쪽이 줄기, 오른쪽을 잎으로 하여 관측값을 보여줌.

*점도표는 정보의 손실이 없다.
*히스토그램의 막대의 높이는 항상 상대도수를 계급구간의 폭으로 나눈 값으로 사용하는 것이 바람직하다.
*도수다각형은 분포의 모양을 파악하는데 용이함, 특히 두 개 이상의 분포 파악에 유리함.
*줄기-잎 그림은 분포를 파악하는 데에 용이하면서도 개개의 관측값에 대한 정보를 잃어버리는 히스토그램과 도수 다각형의 단점을 보완함.
*취하는 관측값의 수가 적으면 범주형 자료요약, 수가 많으면 연속형 자료요약.
** 분포의 모양



*순서대로 종모양, 이봉형, 균일형, 오른쪽으로 편중, 왼쪽으로 편중.
*이봉형의 경우 : 상이한 집단의 자료들이 섞여 있을 때
ex) 남녀 구별하지 않은 몸무게 자료, 서울 강남과 강원도 삼척의 아파트 가격 자료

제 3장 수치를 통한 연속형 자료의 요약

-도표나 그림을 통한 자료의 요약은 작성자의 주관적 판단에 따라 달라짐.
-따라서 이론적 근거를 제시할 수 있는 수치로 자료를 요약하여 대략적인 분포상태를 파악하기 위함.

1. 중심위치의 측도

- 평균 (Mean) : $\bar{x} = \frac{\text{모든 관측값의 합계}}{\text{총 자료의 개수}} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

→ 표본평균($\bar{x}$)은 관측값들의 무게중심에 해당 & 극단적으로 아주 큰 값이나 작은값에 영향을 많이 받음
→ 모평균 : μ 로 표기한다.

- 중앙값 (Median) : 관축값을 크기 순서로 배열하면

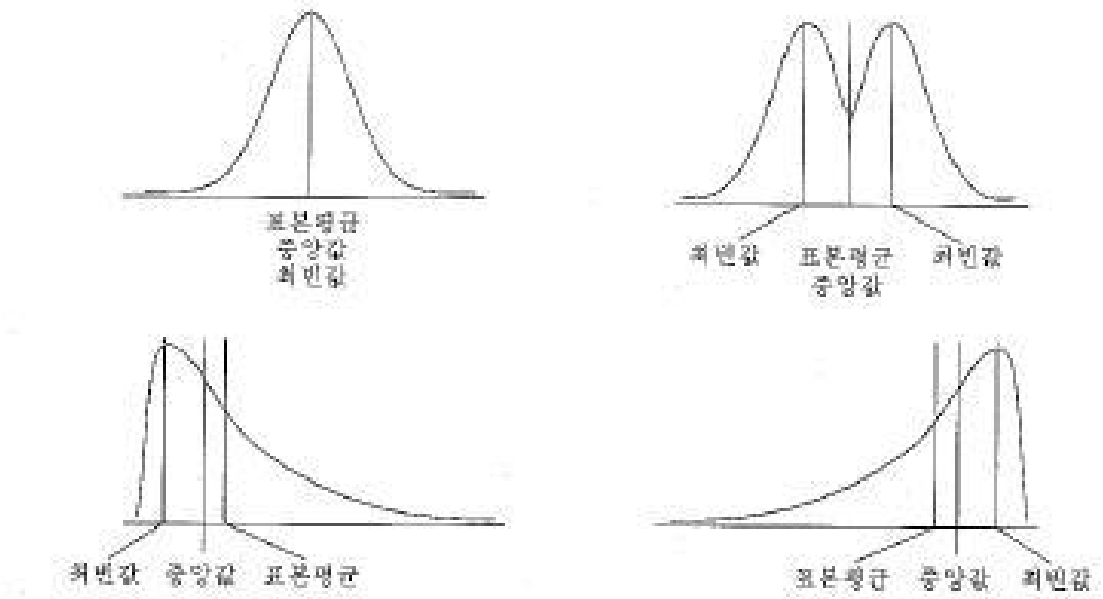
  (i) 자료의 개수 n 이 홀수 : $\frac{(n+1)}{2}$ 번째 관측값

  (ii) 자료의 개수 n 이 짝수 : $\frac{n}{2}$ 번째 관측값과 $\frac{n}{2}+1$ 번째 관측값의 평균

→ 중앙값은 관측값들의 변화에 영향을 받지 않는다.

- 최빈값 (Mode) : 관측값 중에서 가장 자주 나오는 값.

→ 이산형, 범주형 자료에 사용, 연속형 자료엔 부적절. (연속형의 최빈값 : 최대의 도수를 갖는 계급구간의 중앙값)
→ 이산, 범주형에서 대푯값으로 쓰임. 단봉형 분포일 때만 대푯값으로 유효함.



[표본평균, 중앙값, 최빈값의 비교]

2. 퍼진 정도의 측도

-자료의 퍼짐 정도에 대한 측도, 즉 관측값들이 '중심위치' 로부터 얼마나 떨어져 있는지 나타내는 측도

- 편차 (Deviation) : 각 관측값과 표본평균의 차이, 즉 $(x_i - \bar{x})$
  이 때 표본평균이 중심위치의 측도로 사용됨

→ Problem : $\sum_i (x_i - \bar{x}) = 0$ since $\bar{x}$ is the center of gravity. so we consider a square of deviation.

- 분산 (Variance) : $s^2 = \frac{\text{편차의 제곱합}}{n-1} = \frac{1}{n-1}\sum(x_i - \bar{x})^2$
  $= \frac{1}{n-1}(\sum x_i^2 - n\overline{x^2})$
  $= \frac{1}{n-1}(\sum x_i^2 - \frac{1}{n}(\sum x_i)^2)$

→ 표본분산 : s^2 , 모분산 : σ^2 , 표본개수 : n , 모집단 개수 : N

- 표준편차 (Standard Deviation) : $s = +\sqrt{s^2}$

→ 표본표준편차 : s , 모표준편차 : σ

- 범위 (Range) : (관측값 중에서 최댓값) - (관측값 중에서 최소값)

→ 쉽고 빠르게 구할 수 있으나 이상점에 영향을 받고, 자료의 수에 상관없이 같게 나옴

- 제 $100\times p$ 백분위수 구하는 방법

1. 관측값을 작은 순서로 배열한다.
2. 관측값의 개수 n 에 백분율 p를 곱한다.

⇒ i) 만약 $n\times p$ 가 정수 : $n\times p$번째 작은 관측값과 $n\times p+1$번째 작은 관측값의 평균이 백분위수
ii) 만약 $n\times p$ 가 정수 X : $n\times p$의 정수부분에서 1을 더한 값 m을 구한 후, m번째 작은 관측값이 백분위수

→ 제 $100\times p$ 백분위수 (the $100\times p$ -th percentile) : 관측값이 np개 이상 s.t. 관측값 ≤ 백분위수
& 관측값이 $n(1-p)$개 이상 s.t. 백분위수 ≤ 관측값

- 사분위수 (Quartile) :
  -제 1사분위수 : $Q_1 =$ 제 25 백분위수
  -제 2사분위수 : $Q_2 =$ 제 50 백분위수 = 중앙값
  -제 3사분위수 : $Q_3 =$ 제 75 백분위수

→ *** 사분위수 범위 : $IQR = Q_3 - Q_1$ (이상점에 영향받지 않음)

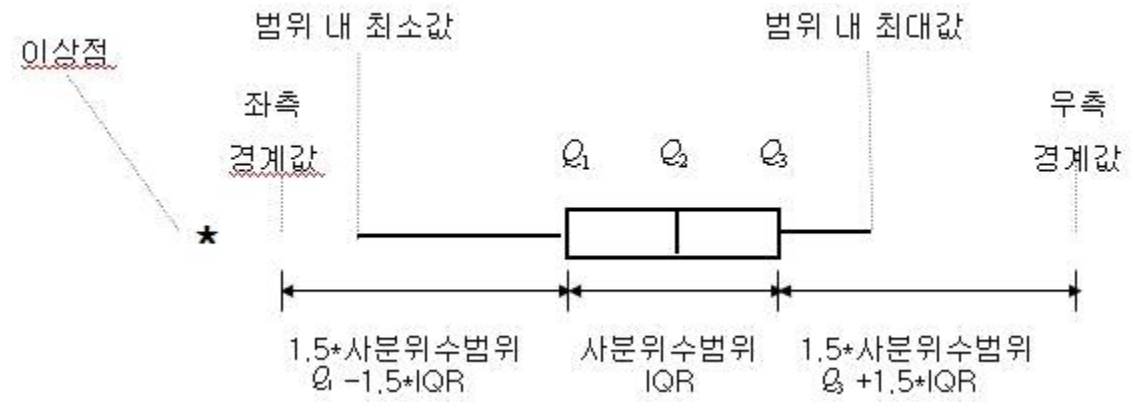
→ *** 표준편차와 사분위수범위의 비교 (둘 다 퍼진 정도를 측정하기 위한 수치)
표준편차 : 중심위치의 측도가 표본평균일 경우 측도로 선택
사분위수범위 : 중심위치의 측도가 중앙값일 경우 측도로 선택
+(표본평균과 중앙값의 특징이 표준편차와 사분위수범위의 특징이 됨)

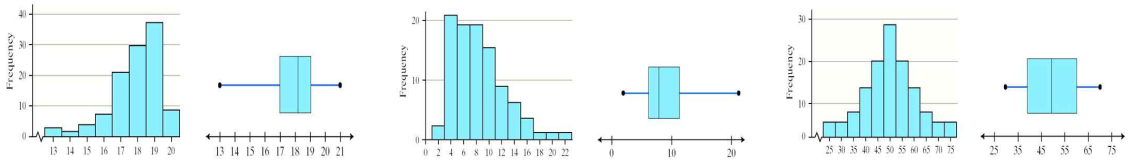
- 변동계수 (Coefficient of Variation) : $CV = \frac{\text{표준편차}}{\text{표본평균}} \times 100$

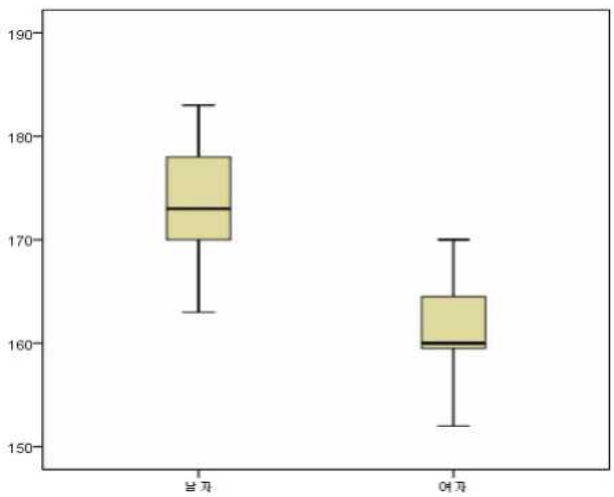
→ 두 개 이상의 분포를 비교할 때 상대적으로 퍼진 정도를 나타내는 수치. (ex : 단위, 중심위치가 다를 때 사용)

3. 상자 그림 (Box plot / Box-whisker plot)

- 상자 그림의 작성과정
  1. 사분위수를 결정한다.
  2. Q_1과 Q_3을 네모난 상자로 연결하고, 중앙값(Q_2)의 위치에 수직선을 긋는다.
  3. $IQR = Q_3 - Q_1$ 을 계산한다.
  4. 상자 양끝에서 $1.5 \times IQR$ 크기 내에 있는 최솟값과 최댓값을 Q_1, Q_3 으로부터 각각 선으로 연결한다.
  5. 양 경계를 벗어나는 자료값들을 * 로 표시하고, 이 점들을 이상점이라고 한다.







	남 자	여 자
Q_1	170	159.25
Q_2	173	160
Q_3	178	164.75
IQR	8	5.5
최소값	163	152
최대값	183	170
$Q_1 - 1.5IQR$	158	151
$Q_3 + 1.5IQR$	190	173

제 4장 두 변수 자료의 요약

1. 두 범주형 변수의 요약

- 분할표 (Contingency table) : 두 변수가 모두 범주형에 속하는 경우 도수분포표를 2차원으로 확장한 형태

→ 한 변수에 대한 범주는 왼쪽에, 또 다른 변수에 대한 범주는 위쪽에 표시하고, 두 범주들이 교차하는 칸(cell) 마다 각 변수의 범주를 동시에 갖는 관측값들의 수를 그 칸의 도수로 기록하면 된다.

<두 변수 자료의 분할표>

	찬성	미결정	반대	합계
남자	112	36	28	176
여자	84	68	72	224
합계	196	104	100	400

	찬성	미결정	반대	합계
남자	0.28	0.09	0.07	0.44
여자	0.21	0.17	0.18	0.56
합계	0.49	0.26	0.25	1.00

- 남자이면서 찬성한 사람의 비율은 전체 중의 28%이다.

■ 남녀별 지지비율 비교

	찬성	미결정	반대	합계
남자	0.636	0.205	0.159	1.00
여자	0.375	0.304	0.321	1.00

2. 그림을 통한 두 연속형 변수의 요약

- 산점도 (Scatter diagram) : 변수 x와 y를 정하여 직교좌표계의 순서쌍으로 표현한 그래프

→ 두 변수 간의 관계를 파악할 때 쓰임. (ex : 관측값들이 직선 or 곡선의 띠를 형성하는지, 관계없이 흩어졌는지)

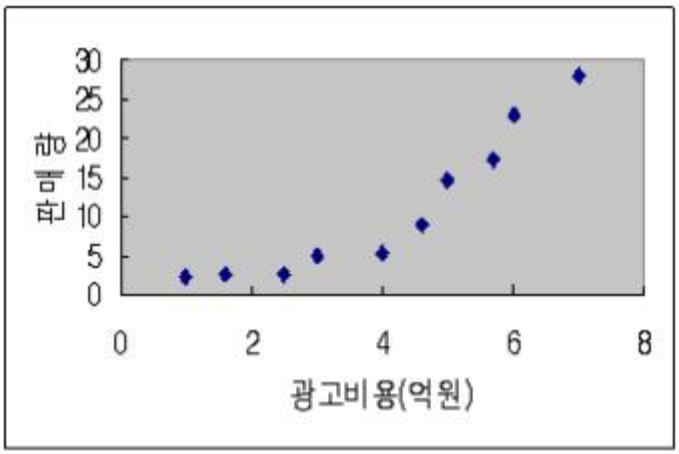


그림 : 광고비용과 판매량의 산점도

3. 수치를 통한 두 연속형 변수의 요약

- 상관계수 (Correlation coefficient) : 두 변수 (x,y)에 대하여 관측값 n개의 순서쌍이 $(x_1,y_1),\cdots,(x_n,y_n)$일 때

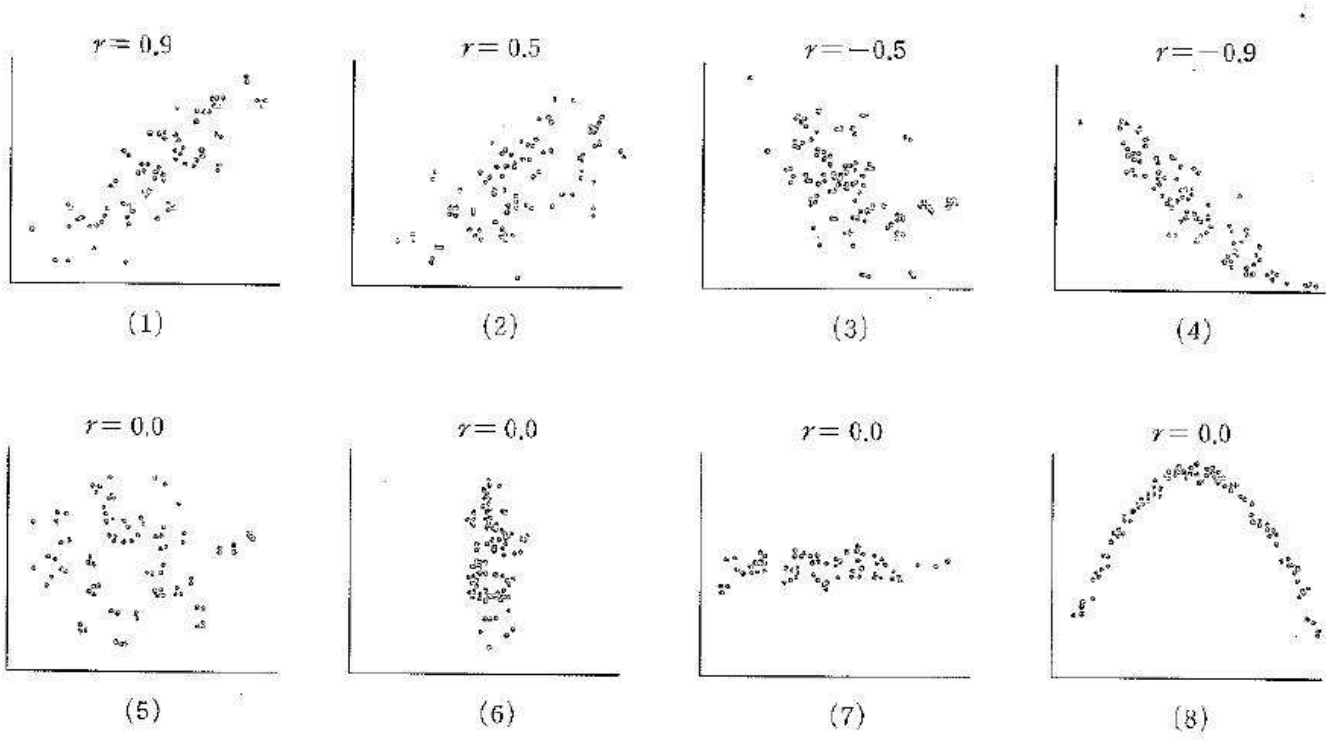
$$\text{(표본)상관계수 } r=\frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \quad \text{s.t. } S_{xx}=x\text{의 편차의 제곱합},\ S_{yy}=y\text{의 편차의 제곱합},\ S_{xy}=(x\text{편차}\times y\text{편차})\text{의 제곱합}$$

→ ***산점도에서 선형적 관계(linear relationship)의 연관성 척도

- 표본상관계수의 특징
  1) $-1<r<1$
  2) $r>0 \Rightarrow$ 양의 기울기 방향으로 띠를 형성, $x<y \Rightarrow f(x)<f(y)$
     $r<0 \Rightarrow$ 음의 기울기 방향으로 띠를 형성, $x<y \Rightarrow f(x)>f(y)$
     $r=1 \Rightarrow$ 모든 점이 정확히 양의 기울기를 갖는 직선 위에 위치
     $r=-1 \Rightarrow$ 모든 점이 정확히 음의 기울기를 갖는 직선 위에 위치
     $r\approx 0 \Rightarrow$ 두 변수 사이에 선형관계가 존재하지 않음
  3) 표본상관계수의 단위는 없다.

→ ***주의사항 : 큰 상관계수의 값이 항상 두 변수 사이의 어떤 '인과관계'를 의미하지 않음. '연관성'이 높은 것임.

ex) 살인사건발생건수 x 와 종교집회의횟수 y 에 대한 산점도의 상관계수가 +1에 가깝다고 할 때
인과관계가 성립된다면 살인사건의 발생을 줄이기 위해 종교집회를 줄이면 됨 ⇒ 잘못된 판단
사실은 제 3의 변수가 존재 : '도시주민의 수' - 살인사건&종교집회가 주민 수多 : 높음, 주민 수小 : 낮음.
이 때 주민의 수 같은 변수를 잠재변수(lucking variable)라고 함.



[상관계수와 산포도의 관계]

제 5장 확률

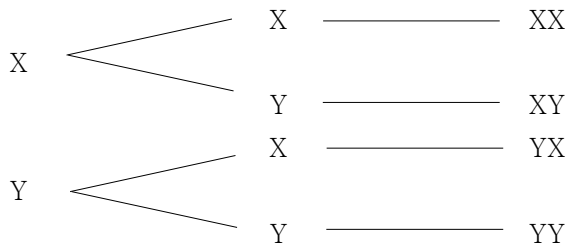
-표본으로부터 모집단을 파악하는 것과 같은 통계적 추론은 확률이론을 기초로 한다. 이러한 확률이론은 불확실성을 구체적으로 수치화하는 작업으로써 통계적 추론에서 아주 유용한 방법이다.

1. 사건의 확률

- 표본공간 (Sample space:Ω) : 한 실험에서 나올 수 있는 모든 결과들의 모임
- 근원사건 (Elementary outcomes:$\omega_1, \omega_2, \cdots$) : 표본공간을 구성하는 개개의 결과
- 사건 (Event:$A, B, \cdots$) : 표본공간의 부분집합으로 어떤 특성을 갖는 결과들의 모임 (즉, 근원사건들의 집합)

→ 포함관계 : $\omega_1, \omega_2, \cdots \in A \subset \Omega$

example) 두 차례 대선에서 여당후보 X나 야당후보 Y 중 하나를 선택하였다고 하자.
사건 A를 두 선거에서 여당후보를 한 번 선택한 사건, 사건 B를 두번 다 야당만 선택하는 사건이라 하면 표본공간과 각 사건을 표본공간의 구성원(근원사건)을 이용하여 나타내어 보아라.



첫 번째 선거 두 번째 선거 가능한 결과들

표본공간 $\Omega = \{XX, XY, YX, YY\}$, 사건 $A = \{XY, YX\}$, 사건 $B = \{YY\}$.

- 사건의 확률 : 동일한 조건하에서 한 가지 실험을 반복할 때 전체 실험 횟수에서 그 사건이 일어나리라고 예상되는 횟수의 비율을 말하고, 사건을 A라고 하면 사건 A의 확률은 $P(A)$로 표시한다.

→ 표본공간이 $\Omega = \{x : 0 \le x \le 1\}$ 와 같은 경우는 원소의 개수를 셀 수 없으므로 이 경우를 연속표본공간이라 함

- 확률의 법칙 : (1) 모든 사건 A에 대하여 $0 \le P(A) \le 1$
  (2) $P(A) = \sum_{\omega_i \in A} P(\omega_i)$
  (3) $P(\Omega) = \sum_{\omega_i \in \Omega} P(\omega_i) = 1$

2. 확률의 계산

- 규칙 1. (균일 확률) : 표본공간 Ω가 k개의 원소로 이루어져 있고 각 근원사건이 일어날 가능성이 동일(equally likely)하다고 하자. 이때 근원사건 중 하나가 일어날 확률은 $1/k$ 로 주어진다. 또 사건 A가 m개의 근원사건으로 이루어져 있다면 사건 A가 일어날 확률은 다음과 같이 주어진다.

$$P(A)=\frac{m}{k}=\frac{A\text{에 속하는 근원사건의 개수}}{\Omega\text{에 속하는 근원사건의 개수}}$$

→ ex) 주사위를 한 번 던질 때 1의 눈이 나올 확률 : 각 눈이 나올 확률이 동일함

- 규칙 2. (상대도수 수렴치로서의 확률) :동일한 실험을 N회 반복할 때 사건 A의 상대도수는 다음과 같다.

$$r_N(A)=\frac{N\text{번의 시행 중 } A\text{가 일어난 횟수}}{N}$$

여기서 N이 증가하면 상대도수가 일정한 값으로 수렴할 때 그 값으로 사건 A가 일어날 확률 $P(A)$를 추정한다.

→ ex) 새로 개발된 상품이 시장에서 성공을 할 확률 : 많은 사람을 대상으로 시장조사를 해야함

→ 증가함이 무한일 경우 : $\lim_{N\to\infty}\frac{f_N(A)}{N}=P(A)$, $f_N(A)$: N번 반복 시 사건 A가 일어나는 횟수

3. 확률 법칙

- 배반사건 : $A\cap B=\varnothing$
- 합사건의 확률법칙 : $P(A\cup B)=P(A)+P(B)-P(A\cap B)$
- 여사건의 확률법칙 : $P(A^c)=1-P(A)$

4. 조건부확률과 독립성

- 사건 B가 주어졌을 때 사건 A의 조건부확률은 $P(A|B)=\frac{P(A\cap B)}{P(B)}$ (단, $P(B)>0$)

→ $P(A\cap B)=P(A|B)\cdot P(B)$

- $P(A\cap B)=P(A)P(B)\Leftrightarrow$ 사건 A와 B가 서로 독립 $\Leftrightarrow$ $P(A|B)=P(A)$ 또는 $P(B|A)=P(B)$

→'독립' 과 '배반'은 다름. (독립일 경우 $P(A\cap B)=P(A)P(B)$, 배반일 경우 $P(A\cap B)=0$)

제 6장 확률분포

1. 확률변수

- 확률변수 (Random variable) : 각각의 근원사건들에 실숫값을 대응시키는 함수, $X, Y, \cdots$ 등으로 표시한다.

→ 유한 or 가산무한 : 이산확률변수 (discrete RV), 비가산무한(연속적 구간 내) : 연속확률변수 (Continuous RV)

2. 이산확률분포

- 확률분포 (Probability distribution) : 확률변수가 갖는 값들과 그에 대응하는 확률값을 나타내는 것으로 나열된 표나 수식으로 표현된다. 보통은 확률변수 X의 분포라 한다.

→ def. 확률변수 X 가 값 x를 가질 때 $P(X=x)=f(x)$: 확률질량함수 (probability mass function)

→ def. 확률변수 X 가 값 x보다 작거나 같을 때 $P(X \le x) = \sum_{X \le x} f(x)$: 누적 확률분포 함수

example)

[이산확률변수의 확률분포]

X가 취하는 값 (x)	확률 $f(x)$
x_1	$f(x_1)$
x_2	$f(x_2)$
⋮	⋮
x_k	$f(x_k)$
합계	1

- 이산 확률질량함수의 조건 : (1) 모든 x_i 에 대하여 $0 \le f(x_i) \le 1$
  (2) $\sum_{\forall x_i} f(x_i) = 1$

3. 이산확률분포의 기댓값(평균) 과 표준편차

- 확률변수 X의 기댓값(평균) : $E(X) = \mu = \sum x_i \cdot f(x_i)$

- 확률변수 X의 분산 : $Var(X) = \sigma^2 = E(X-\mu)^2 = \sum_i (x_i-\mu)^2 f(x_i)$
- 확률변수 X의 표준편차 : $sd(X) = \sigma = + \sqrt{Var(X)}$

→ ***분산의 계산식 : $Var(X) = E(X^2) - (E(X))^2$

3. 두 확률변수의 결합분포

- 결합확률분포 (Joint Probability Distribution) : X가 취하는 값과 Y가 취하는 값의 각 쌍에 대응하는 확률

→ *** X, Y가 이산일 때 $f(x_i, y_j) = P[X = x_i,\ Y = y_j]$ for all $1 \le i \le m, 1 \le j \le n$

→ $P[X+Y=3]$, $P[X>Y]$ 와 같은 경우는 $Z = X+Y$, $Z=(X,Y)$ s.t. $X>Y$ 와 같이 생각하자.

- X, Y 의 주변확률분포 (Marginal Probability Distribution) :

$$f_X(x_i) = P[X = x_i] = \sum_{j=1}^{n} f(x_i, y_j)$$

$$f_Y(y_j) = P[Y = y_j] = \sum_{i=1}^{m} f(x_i, y_j)$$

X \ Y	y_1	y_2	$\cdots$	y_n	합 계
x_1	$f(x_1,y_1)$	$f(x_1,y_2)$		$f(x_1,y_n)$	$\sum_{j=1}^{n} f(x_1,y_j) = f_X(x_1)$
x_2	$f(x_2,y_1)$	$f(x_2,y_2)$		$f(x_2,y_n)$	$\sum_{j=1}^{n} f(x_2,y_j) = f_X(x_2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
x_m	$f(x_m,y_1)$	$f(x_m,y_2)$		$f(x_m,y_n)$	$\sum_{j=1}^{n} f(x_m,y_j) = f_X(x_m)$
합 계	$\sum_{i=1}^{m} f(x_i,y_1)$			$\sum_{i=1}^{m} f(x_i,y_n)$	1

- 결합확률분포의 기댓값 :

$$E[XY] = \sum_{i}^{m}\sum_{j}^{n} x_i y_j f(x_i, y_j)$$

$$E[X] = \sum_{i}^{m}\sum_{j}^{n} x_i f(x_i, y_j) = \sum_{i}^{m} x_i f_X(x_i) = \mu_X$$

$$E[aX+bY] = aE(X) + bE(Y)$$

$g(X,Y)$: 함수 $\Rightarrow$ $E[g(X,Y)] = \sum_{i}^{m}\sum_{j}^{n} g(x_i, y_j) f(x_i, y_j)$

4. 공분산과 상관계수

- X와 Y의 공분산 (Covariance) : $Cov(X,Y) = E(X-\mu_X)E(Y-\mu_Y) = E(XY)-\mu_X\mu_Y$

→ 두 개의 확률변수 X와 Y가 상호 어떤 관계를 가지며 변화하는가를 나타내주는 척도

→ $Cov(aX,bY) = abCov(X,Y)$, $Cov(X,X) = Var(X)$, $Var(X\pm Y) = Var(X)+Var(Y)\pm 2Cov(X,Y)$

- X와 Y의 상관계수 (Correlation Coefficient) : $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$

→ $-1 \le Corr(X,Y) \le 1$, 정확한 선형관계 $Y = aX+b$ 가 성립 $\Rightarrow$ $Corr(X,Y) = \pm 1$

→ $Corr(aX,bY) = \frac{ab}{|ab|}Corr(X,Y)$

5. 두 확률변수의 독립성

- X와 Y가 독립 $\Leftrightarrow$ $f(x_i,y_j) = f_X(x_i)f_Y(y_j)$ for all i,j

→ 두 변수가 독립 $\Rightarrow$ $E(XY) = E(X)E(Y)$ $\Rightarrow$ $Cov(X,Y) = 0$, $Corr(X,Y) = 0$. '역은 성립하지 않음'

제 7장 이항분포와 그에 관련된 분포들

-이산확률분포 중 가장 기본적이고 실제로 많이 적용되는 이항분포와 그에 관계된 분포들에 대해 알아본다.
-시행(trial) : 매번 반복되는 추출(실험)

- 베르누이 시행 (Bernoulli trial) : (1) 각 시행은 성공(S) 실패(F)의 두 결과만을 갖는다.
  (2) 각 시행에서
  성공할 확률은 $P(S)=p$, 실패할 확률은 $P(F)=q=1-p$ 로
  그 값이 일정하다. (상수이다)
  (3) 각 시행은 서로 독립으로 각 시행의 결과가
  다른 시행의 결과에 영향을 미치지 않는다. (독립이다)

→ 베르누이 시행은 복원추출이다. ex) 비복원추출 : 두 번째 결과는 첫 번째 결과에 영향을 미치므로 베르누이 X.

- 이항분포 (Binomial distribution) : 베르누이 시행을 n번 시행하였을 때 그 중 성공의 횟수를 확률변수 X로 정의한다면 이 확률변수 X는 이항분포를 따른다고 말한다. 즉,
  $X \sim Bin(n,p)$, $x=0,1,2\cdots,n$
  n : 베르누이 시행의 반복횟수
  p : 각 시행에서 성공할 확률 $P(S)$
  X : n번 시행 중 성공의 횟수

→ 이항분포의 확률질량함수 : $f(x)=P(X=x)=\binom{n}{x}p^x q^{n-x}$

→ 이항분포의 누적분포함수 : $F(x)=P(X\le x)=\sum_{k=0}^{x}\binom{n}{k}p^k q^{n-k}$

→ 이항분포의 기댓값, 분산, 표준편차 : $E(X)=np$, $Var(X)=npq$, $sd(X)=\sqrt{npq}$

- 초기하분포 (Hypergeometric distribution) : 유한한 모집단의 구성원소가 두 가지의 범주로 분류될 때 비복원방법으로 추출되는 표본 중 어느 한 범주에 속하는 수를 X라 할 때, 확률변수 X의 확률분포 $X \sim HYP(N,D,n)$. 즉,
  N : 모집단의 크기
  n : 표본의 크기
  D : 모집단 내에서 범주 A에 속하는 구성원소의 수
  X : 표본 내에서 범주 A에 속하는 구성원소의 수

→ 초기하분포의 확률질량함수 : $f(x)=P(X=x)=\dfrac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$, $x=0,1,2,\cdots,n$

→ 초기하분포의 기댓값, 분산 : $E(X)=np$, $Var(X)=npq\dfrac{N-n}{N-1}$, 여기서 $p=\dfrac{D}{N}$, $\dfrac{N-n}{N-1}$은 유한모집단 수정요인

→ 초기하분포는 비복원추출시행이다.
→ *** 모집단에 비해 표본의 크기가 상대적으로 작으면 ($n<0.05N$) 이항분포와 동일하다.
∵ 베르누이 독립성 위반 미약해짐

- 포아송 분포 (Poisson distribution) : 확률변수 X가 평균(모수)이 λ인 포아송분포를 따른다고 하면

  확률질량함수는 $P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}$, $x=0,1,\cdots$ 이다.

→ 발생가능성이 그다지 크지 않은 현상에 대한 발생건수에 대한 확률에 관심을 갖는 경우

→ 확률변수 X : 어느 구간 내에서 발생한 사건의 수

→ $X \sim Poi(\lambda)$ 일 때, 평균 $\mu=\lambda$, 분산 $\sigma^2=\lambda$

- 포아송분포이기 위한 3가지 가정

1. 주어진 구간에서 사건의 평균 발생횟수의 확률분포는 구간의 시작점에는 관계가 없고 구간의 길이에만 영향을 받는다.
2. 한 순간에 2회 이상의 사건이 발생할 확률은 거의 0에 가깝다.
3. 한구간에서 발생한 사건의 횟수는 겹치지 않는 다른 구간에서 발생하는 사건의 수에 영향을 받지 않는다.

- 포아송분포의 확산성

기본단위 (면적, 길이, 시간) 당 관심있는 사건의 발생 수가 평균이 λ인 포아송분포를 따른다면
t 단위당 발생하는 사건의 수는 평균이 $t\lambda$인 포아송분포

example) 어느 지역의 지하수가 특별한 종류의 박테리아를 1ml 당 두 마리 꼴로 포함하고 있다 하자. 만일 2ml의 지하수 물을 꺼내어 접시에 담았을 때 적어도 한 마리의 박테리아가 접시에서 검출될 확률은?

X : 접시에서 검출된 박테리아의 수

$X \sim Poi(4)$, $t\lambda=2\times2=4$, $P(X\geq1)=1-P(X=0)=1-e^{-4}$

제 8장 정규분포

-연속확률분포들 중에서 대부분의 통계학 이론의 기본이 되는 정규분포에 관하여 알아본다.
-정규분포는 여러 분야의 학자들에 의해 실제의 여러 종류의 자료를 설명하는 데 정규분포가 좋은 분포임이 밝혀짐

1. 연속확률변수와 확률분포

- 연속확률변수 (continuous random variable) : 확률변수 X의 가능한 값들이 몸무게, 키, 제품의 수명, 기차의 도착시간 등과 같이 셀 수 없이 무한히 많을 때 이를 연속확률변수라고 한다.

- X의 확률밀도함수 (probability density function) $f(x)$: 1) 모든 x값들에 대하여 $f(x) \geq 0$

  2) $P(a \leq X \leq b) = \int_a^b f(x)dx$

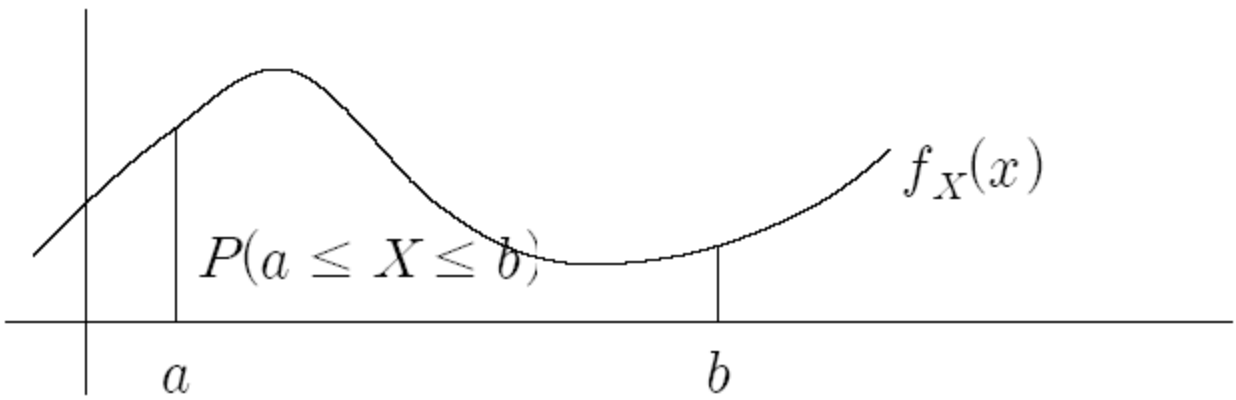
  3) $P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$

→ X의 누적분포함수 : $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$, $(F'(x) = f(x))$

- 균일 분포 (Uniform distribution) : 확률변수 X가 어느 구간 (a,b)에서 정의되고, 그 구간에서 확률밀도함수가 똑같은 높이로 일정한 확률분포

  확률밀도함수 : $f(x) = \frac{1}{b-a}$, $a \leq X \leq b$

  평균 : $E(X) = \frac{a+b}{2}$, 분산 : $Var(X) = \frac{(b-a)^2}{12}$



구간 [a, b]에서의 확률

2. 정규분포

- 정규분포 (Normal Distribution) : 확률변수 X 가 평균 μ 와 분산 σ^2 을 갖는 정규분포를 따른다고 하자.

  즉, $X \sim N(\mu, \sigma^2)$ 이면 정규분포의 확률밀도함수 (p.d.f) 는

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \ , \ -\infty < x < \infty$$

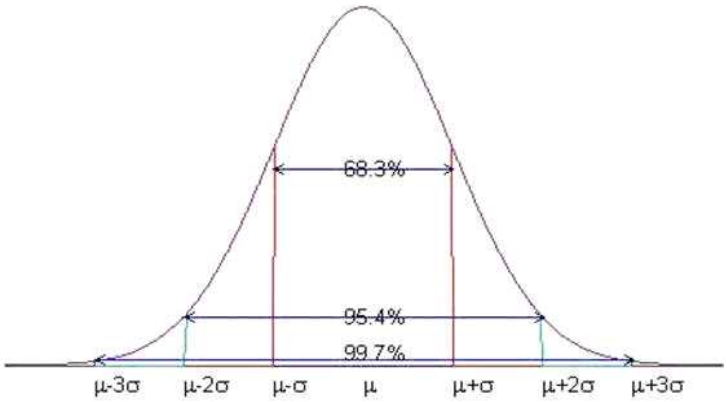
→ 정규분포의 p.d.f 는 종모양(bell shape) 이다

→ p.d.f 아래쪽의 총 면적은 항상 1이다.

→ 평균 μ는 p.d.f 의 중앙에 위치해 있고 μ에 대해 좌우대칭이며 μ에서 p.d.f 높이가 가장 크다.

→ 평균=중위수=최빈값

→ *** 정규분포의 모양과 위치는 평균 μ와 표준편차 σ에 따라 달라진다.



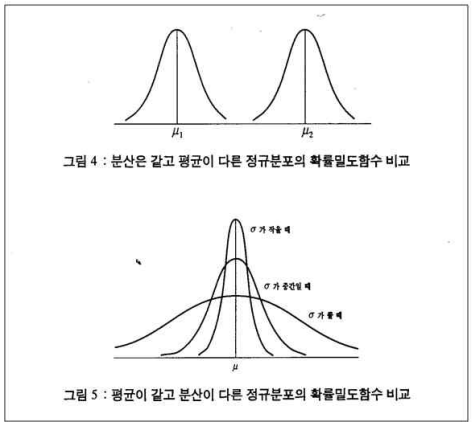


그림 4 : 분산은 같고 평균이 다른 정규분포의 확률밀도함수 비교

그림 5 : 평균이 같고 분산이 다른 정규분포의 확률밀도함수 비교

- 평균이 다르고, 분산이 같은 경우 ➔ 수평이동
- 평균이 같고, 분산이 다른 경우 ➔ 분포의 폭이 변한다.

- 표준정규분포 (Standard Normal Distribution) : $Z \sim N(0,1)$ 인 분포를 말한다.

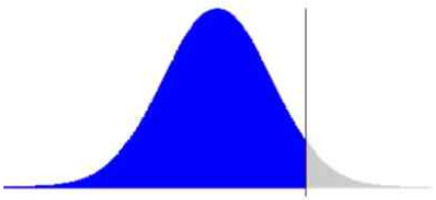
  확률밀도함수는 $\phi(z) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}z^2)$

  누적분포함수는 $\Phi(z) = P(Z \le z)$

*** 표준정규분포를 따르는 Z의 확률 계산

$$P(a \le Z \le b) = P(Z \le b) - P(Z \le a)$$
$$= \Phi(b) - \Phi(a)$$

$P[Z \le z]$: Cumulative distribution (누적확률)

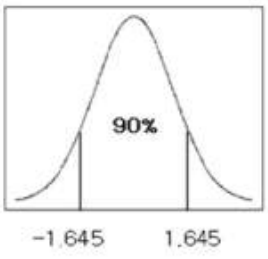


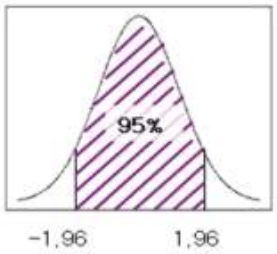
ex) $P[Z \le 1.98] = 0.9761$

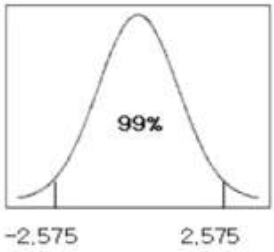
$P[Z \le -0.03] = 0.4880$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998

*** 기억하기







$P(-1.645 \le Z \le 1.645) = 0.90$ $P(-1.96 \le Z \le 1.96) = 0.95$ $P(-2.575 \le Z \le 2.575) = 0.99$

$P(Z \le -1.645) = 0.05$ $P(Z \le -1.96) = 0.025$ $P(Z \le -2.575) = 0.005$

*** 정규분포일 때 확률 계산

- 표준정규확률변수 (standard normal random variable) : 확률변수 X가 $N(\mu, \sigma^2)$일 때 표준화된 확률변수 $Z = \frac{X-\mu}{\sigma}$ 는 정규분포 $N(0,1)$을 따른다.

→ $X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$: '표준화'

→ $P(X \le x) = P(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}) = P(Z \le \frac{x-\mu}{\sigma})$

- 정규분포의 성질

1) $X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$: 표준화 가능

2) a,b가 상수이고, $X \sim N(\mu, \sigma^2) \Rightarrow aX+b \sim N(a\mu+b, a^2\sigma^2)$

3) $X \sim N(\mu_X, \sigma_X^2)$ 와 $Y \sim N(\mu_Y, \sigma_Y^2)$ 가 서로 독립 $\Rightarrow X+Y \sim N(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$

$X-Y \sim N(\mu_X-\mu_Y, \sigma_X^2+\sigma_Y^2)$