01

확률의 이해

01. 표본공간과 사건

■ 표본공간과 사건

- **확률실험(random experiment):** 여러 가지 가능한 결과 중 하나의 사건(event)을 발생시키는 활동.
- **표본공간(sample space):** 확률실험으로 발생할 수 있는 모든 가능한 결과들을 의미하며, 각각의 결과들은 상호 배타적(mutually exclusive)이어야 한다.
- 주사위를 던져 나오는 결과를 예로 들어보자. 여기서 나올 가능성이 있는 결과는 1, 2, 3, 4, 5, 6이며, 다음과 같이 표현할 수 있다.

$$S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$$

- 사건(event)은 표본공간의 부분집합이다. 주사위를 던져 홀수가 나오는 사건을 A, 5가 나오는 사건을 B라 하면 해당 사건은 다음과 같이 표현할 수 있다

$$A = \{1,\ 3,\ 5\}\ ,\ B = \{5\}$$

■ 표본공간과 사건

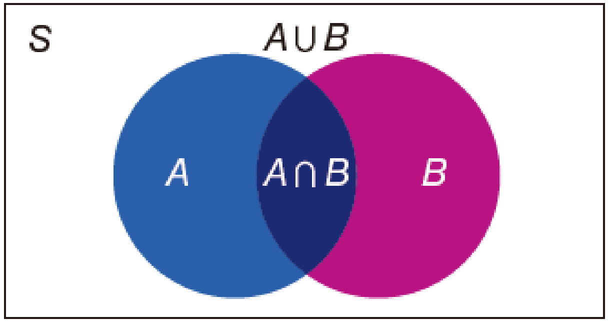
하나 더 알기 **상호 배타적(mutually exclusive)**

- 두 사건이 상호 배타적이라는 것은 두 사건 중 한 사건이 일어날 확률 [P(A or B)]가 두 사건이 각각 일어날 단순 확률의 합 [P(A)+P(B)]와 같다는 말이다. 또는 두 사건이 동시에 일어날 확률이 0이 되면 두 사건은 상호 배타적(mutually exclusive)이다. P(A or B)=P(A)+P(B)-P(A and B)에서 P(A and B)=0이므로 P(Aor B)=P(A)+P(B)가 된다.

01. 표본공간과 사건

사건의 연산

- **합사건(sum of events)**: 표본공간 S의 두 사건 A, B에 대하여 사건 A 또는 B 중에 적어도 하나의 사건이 일어나는 경우를 말하며, A∪B로 표현.



[합사건]

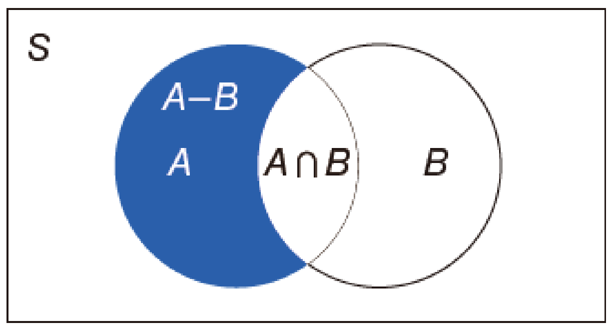
합사건의 계산

$$A \cup B = A + B - (A \cap B) \tag{4.3}$$

01. 표본공간과 사건

■ 사건의 연산

- **차사건(difference of events):** 표본공간 S의 두 사건 A, B에 대하여 사건 A만 발생하고 사건 B는 일어나지 않는 경우를 말하며, A-B라고 표현.



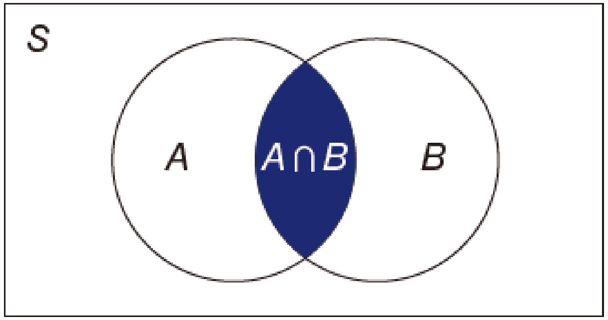
[차사건]

차사건의 계산

$$A - B = A - (A \cap B) = A \cap B^c \tag{4.4}$$

■ 사건의 연산

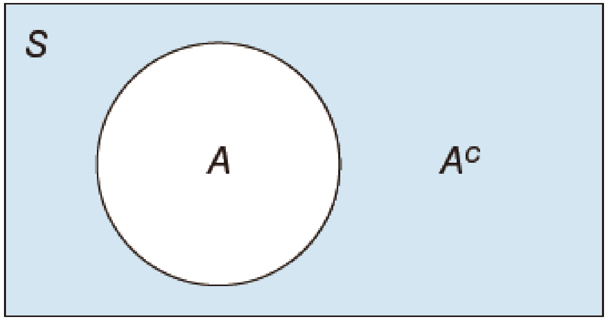
- **교사건(intersection of events):** 표본공간 S의 두 사건 A, B에 대하여 사건 A, B가 모두 일어나는 경우를 말하며, A∩B로 표현.



[교사건]

■ 사건의 연산

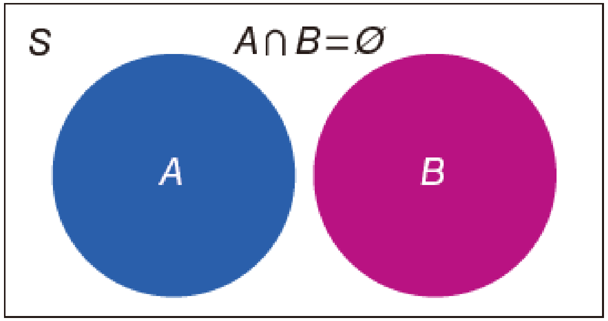
- **여사건(complementary events):** 표본공간 S에서 사건 A가 일어나지 않는 경우를 말하며, A^c 로 표현.



[여사건]

■ 사건의 연산

- **배반사건(exclusive events):** 두 사건이 동시에 발생하는 경우가 없는 사건으로, 한쪽 사건이 일어나면 다른 쪽 사건이 일어나지 않을 때의 두 사건을 의미.
- 사건 A와 사건 B의 교사건(A∩B)이 공집합(Ø)인 경우, 두 사건은 상호 배타적이라 표현할 수 있다.



[배반사건]

확률의 개념

- **확률(probability):** 여러 가지 사건이 나타날 수 있을 때 어떤 특정한 사건이 일어날 가능성을 수치로 나타낸 것.
- 확률은 0부터 1사이의 비율로 나타낸다.
- 표본공간(S)에서 $O_1,\ O_2,\ O_3, \cdots\cdots, O_k$ 의 k가지 사건이 있다고 하면, 다음과 같이 나타낼 수 있다.

$$S = \{O_1,\ O_2,\ O_3,\ \cdots\cdots, O_k\}$$

- 각각의 확률은 다음의 두 가지 조건을 만족해야 한다.

$$① \quad 0 \le P(O_i) \le 1,\ i = 1,\ 2,\ 3,\ \cdots\cdots,\ k$$

$$② \quad \sum_{i=1}^{k} P(O_i) = P(O_1) + P(O_2) + P(O_3) + \cdots\cdots + P(O_k) = 1$$

01. 표본공간과 사건

■ 사건에 확률을 부여하는 방법

- **고전적 확률(classical probability):** 확률실험(random experiment)의 결과가 언제나 동일하다는 가정으로 확률을 계산하며 이론적 확률(theoretical probability)이라고도 부른다.
- 예를 들어 주사위를 던져 숫자 5가 나오는 사건을 A라 할 때 A의 확률은 다음과 같이 나타낼 수 있다.

$$P(A) = \frac{1}{6}$$

고전적 확률

$$\text{고전적 확률} = \frac{\text{해당 사건의 결과 수}}{\text{전체 표본공간에서 가능한 모든 결과 수}} \tag{4.7}$$

■ 사건에 확률을 부여하는 방법

- **상대도수 확률(relative frequency probability):** 해당 사건이 발생한 횟수를 실제 관측치의 총 횟수로 나누어 계산하는 것으로, 경험적 확률(empirical probability)이라 부른다.
- 이론적인 예측이 아닌 실제 실험을 무한히 반복했을 때 사건이 일어날 확률이 특정 값에 가까워지는 경우 그 값을 상대도수 확률이라고 한다.

사건에 확률을 부여하는 방법

[표 4-1] 상대도수 확률: 주사위를 60,000번 던진 결과

시행 횟수	결과	사건 발생 횟수	상대도수	고전적 확률
60,000번	1	9,273	0.1546	0.1667
	2	11,234	0.1872	0.1667
	3	8,931	0.1489	0.1667
	4	9,931	0.1655	0.1667
	5	9,317	0.1553	0.1667
	6	11,314	0.1886	0.1667
합계		60,000	1.0000	1.0000

- 주사위를 60,000번 던졌을 때 숫자 5가 나타날 확률을 상대도수 확률로 계산하면 다음과 같이 계산할 수 있다.

$$\frac{9,317}{60,000} = 0.1553$$

01. 표본공간과 사건

■ 사건에 확률을 부여하는 방법

- **주관적 확률(subjective probability):** 어떤 사건이 일어날 것이라고 각 개인이 느끼는 확률을 의미.

02

확률의 구분

02. 확률의 구분

결합확률

- **결합확률(joint probability):** 두 개 이상의 사건이 동시에 일어날 확률.
- 사건 A, B의 결합확률은 $A \cap B$의 확률을 지칭하며, 다음과 같이 표기한다.

결합확률

$$P(A \cap B) \tag{4.9}$$

[표 4-2] 경영경제통계학 수강생의 소속 학과와 성별의 결합확률표

학과 \ 성별	여학생(G)	남학생(G^c)
경영학과(A)	34명, $P(A \cap G) = 0.298$	45명, $P(A \cap G^c) = 0.395$
타학과(A^c)	22명, $P(A^c \cap G) = 0.193$	13명, $P(A^c \cap G^c) = 0.114$

■ 결합확률

- 경영경제통계학 수업에서 수강생이 경영학과 학생인 경우를 사건 A, 여학생의 경우를 사건 G라고 가정하자. [표 4-2]는 예시의 각 경우의 수와 결합확률을 나타낸 표로 결합확률표라고 부른다.

[표 4-2] 경영경제통계학 수강생의 소속 학과와 성별의 결합확률표

학과 \ 성별	여학생(G)	남학생(G^c)
경영학과(A)	34명, $P(A \cap G) = 0.298$	45명, $P(A \cap G^c) = 0.395$
타학과(A^c)	22명, $P(A^c \cap G) = 0.193$	13명, $P(A^c \cap G^c) = 0.114$

02. 확률의 구분

■ 조건부확률

- 사건 B가 일어났다는 가정하에 사건 A가 일어날 확률을 특정 조건하에서의 확률이라는 의미로 **조건부확률(conditional probability)**이라 부른다. 조건부확률은 다음과 같이 표기하고 'B 조건부 확률 A'라고 읽는다.

조건부확률

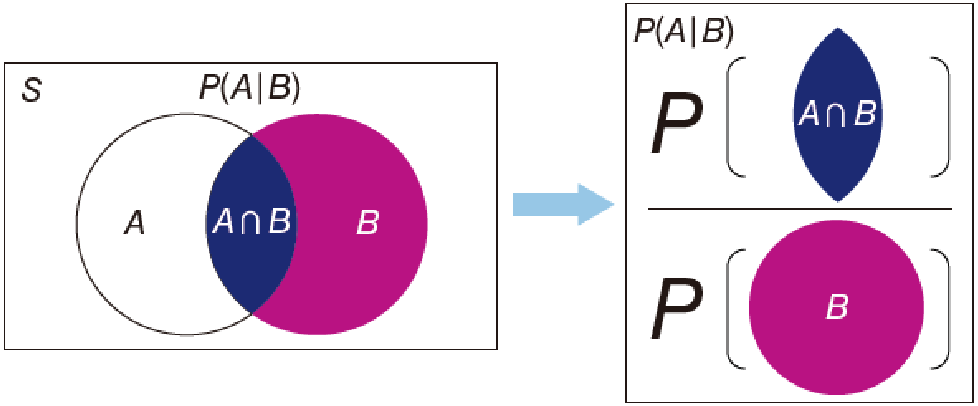
$$P(A \mid B) \tag{4.11}$$

02. 확률의 구분

■ 조건부확률

- 사건 B가 발생하면 B가 새로운 표본공간이 되고, 그 상황에서 A가 일어나는 사건은 A∩B가 된다. 그렇다면 P(A|B)는 다음과 같이 계산할 수 있다.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



[조건부확률: P(A | B)]

■ 조건부확률

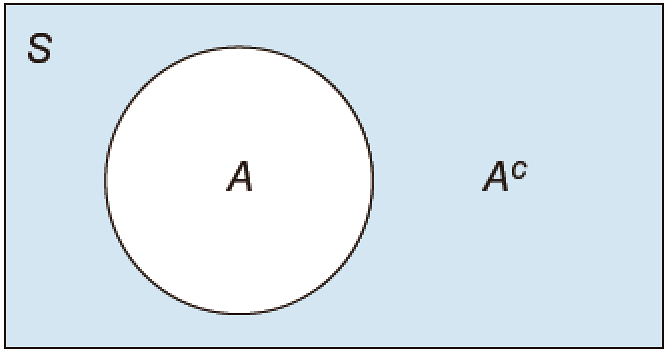
[표 4-4] 경영경제통계학 수강생의 소속 학과와 성별의 조건부확률

성별 / 학과	여학생(G)	남학생(G^c)	합계	조건부확률
경영학과(A)	34명, $P(A \cap G) = 0.298$	45명, $P(A \cap G^c) = 0.395$	79명, $P(A) = 0.693$	$P(G \mid A) = \frac{P(G \cap A)}{P(A)} = \frac{34}{79}$ $P(G^c \mid A) = \frac{P(G^c \cap A)}{P(A)} = \frac{45}{79}$
타학과(A^c)	22명, $P(A^c \cap G) = 0.193$	13명, $P(A^c \cap G^c) = 0.114$	35명, $P(A^c) = 0.307$	$P(G \mid A^c) = \frac{P(G \cap A^c)}{P(A^c)} = \frac{22}{35}$ $P(G^c \mid A^c) = \frac{P(G^c \cap A^c)}{P(A^c)} = \frac{13}{35}$
합계	56명, $P(G) = 0.491$	58명, $P(G^c) = 0.509$	114명, $P(S) = 1$	
조건부확률	$P(A \mid G) = \frac{P(A \cap G)}{P(G)} = \frac{34}{56}$ $P(A^c \mid G) = \frac{P(A^c \cap G)}{P(G)} = \frac{22}{56}$	$P(A \mid G^c) = \frac{P(A \cap G^c)}{P(G^c)} = \frac{45}{58}$ $P(A^c \mid G^c) = \frac{P(A^c \cap G^c)}{P(G^c)} = \frac{13}{58}$		

03

확률의 법칙

■ 여사건 법칙



[여사건의 확률 : $P(A^c)$]

여사건

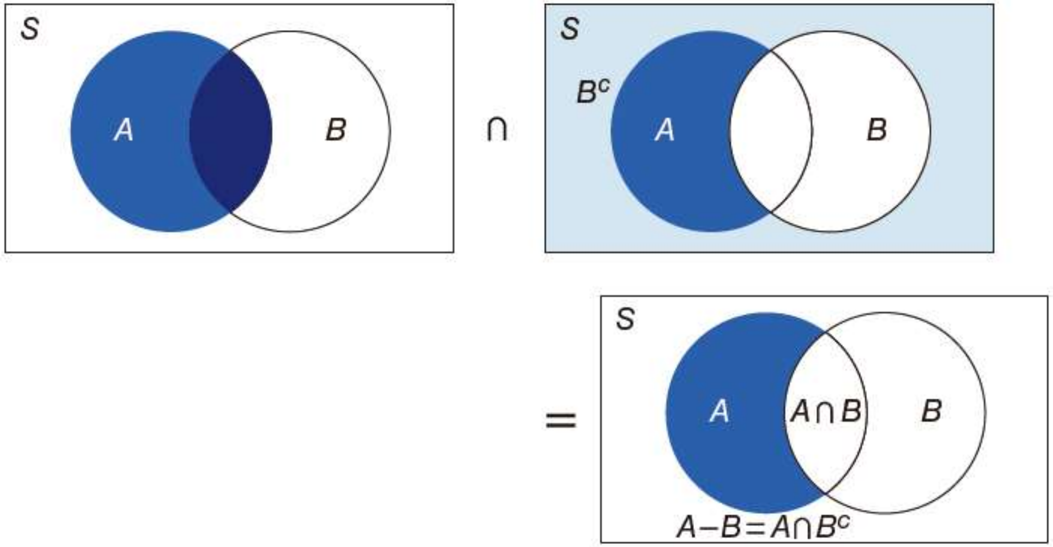
$$P(A^c) = 1 - P(A) \quad (4.13)$$

$$\boxed{\begin{aligned} P(A) \cap P(A^c) &= 0 \\ P(A) \cup P(A^c) &= 1 \end{aligned}}$$

03. 확률의 법칙

■ 여사건 법칙

- 이전 절에서 확인한 차사건의 확률 P(A-B)를 여사건과 교사건의 확률로 표현하면 $P(A \cap B^c)$이고, 벤다이어그램에서 확인할 수 있다.



[차사건의 확률 : $P(A-B) = P(A \cap B^c)$]

■ 곱셈 법칙

- **곱셈 법칙(multiplication rule):** 조건부확률에서 파생된다. 사건 A와 사건 B 두 사건이 있을 때 다음 공식이 성립된다.

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

- 여기서 P(A∩B)의 확률을 얻으면 아래의 곱셈 법칙이 완성된다.

확률의 곱셈 법칙(multiplication rule)

$$P(A\cap B)=\mathrm{P}(B|A)\cdot P(A)$$
$$P(A\cap B)=\mathrm{P}(A|B)\cdot P(B)$$
$$P(A\cap B)=\mathrm{P}(B|A)\cdot P(A)=\mathrm{P}(A|B)\cdot P(B) \quad (4.16)$$

03. 확률의 법칙

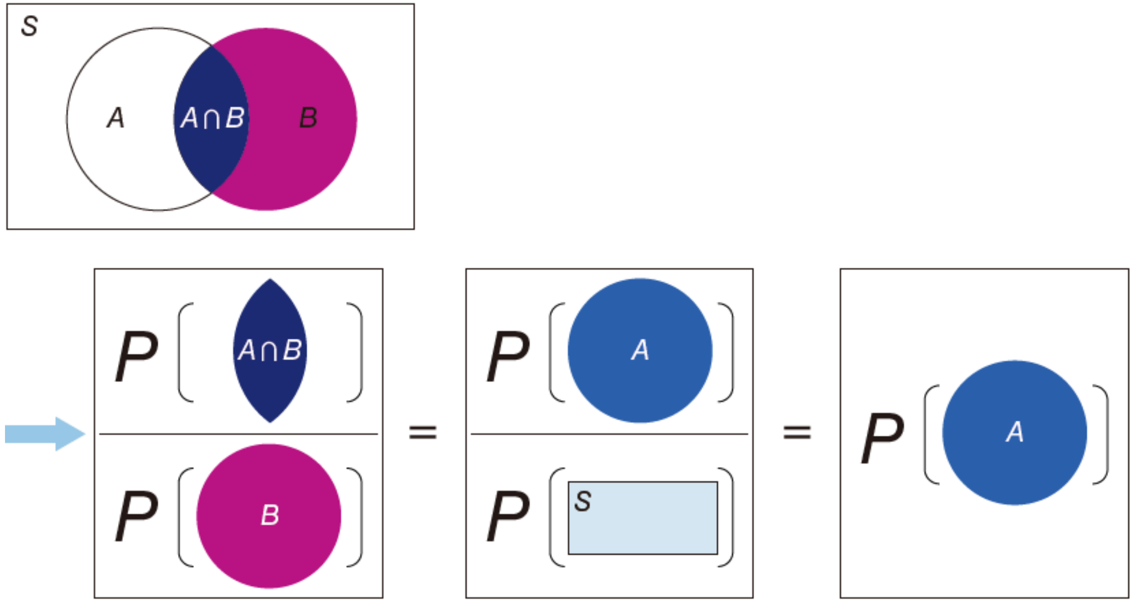
■ 곱셈 법칙 – 독립사건에서의 곱셈 법칙

- 먼저, A와 B가 독립사건이면 $P(A \cap B) = P(A) \cdot P(B)$가 성립된다. 확률의 곱셈법칙에서 $P(A \cap B) = P(B \mid A) \cdot P(A) = P(A \mid B) \cdot P(B)$임을 확인하였다. 독립사건인 경우 $P(A \mid B) = P(A),, P(B \mid A) = P(B)$를 유추할 수 있다. 즉, $P(A \mid B)$가 확률 A와 같으므로 B가 일어났다는 가정하에 A가 일어날 확률이나 전체에서 A가 일어날 확률이 같다.
- 즉, A가 일어났다는 가정하에 B가 일어날 확률이나 A가 일어나지 않았다는 가정하에 B가 일어날 확률이나 마지막으로 전체 모사건에서 B가 일어날 확률이 모두 같다는 의미이다.

$$P(B \mid A) = P\left(B \mid A^c\right) = P(B)$$

03. 확률의 법칙

곱셈 법칙 – 독립사건에서의 곱셈 법칙



[독립사건의 이해: P (A| B) = P (A)]

■ 곱셈 법칙 – 독립사건에서의 곱셈 법칙

- 여기서 A와 B가 서로 독립사건이라는 의미는 사건 B가 일어났다는 가정하에 사건 A가 일어날 확률 $\frac{P(A \cap B)}{P(B)}$와 전체에서 A가 일어날 확률인 P(A)가 같음을 확인하였고, 다음의 수식들을 유추할 수 있다.

A와 B가 서로 독립사건일 경우 성립하는 계산식

$$P(A \cap B) = P(A) \cdot P(B)$$
$$P(A \mid B) = P(A),\ P\left(A \mid B^c\right) = P(A) \tag{4.18}$$
$$P(B \mid A) = P(B),\ P\left(B \mid A^c\right) = P(B)$$

곱셈 법칙 – 종속사건에서의 곱셈 법칙

- 종속사건은 독립사건의 반대 개념으로 서로의 사건이 각각 다른 사건의 발생 확률에 영향을 준다는 의미이다.

*A*와 *B*가 서로 종속사건일 경우 성립하는 계산식

$$P(A \cap B) \neq P(A) \cdot P(B)$$
$$P(A \mid B) \neq P(A),\ P(A \mid B^c) \neq P(A) \qquad (4.19)$$
$$P(B \mid A) \neq P(B),\ P(B \mid A^c) \neq P(B)$$

■ 곱셈 법칙 – 종속사건에서의 곱셈 법칙

하나 더 알기 전혀 다른 개념인 독립사건과 배반사건

- 배반사건은 A와 B의 교사건의 확률이 0인 사건을 말하며, $P(A \cap B) \neq P(A) \cdot P(B)$로 표현되기 때문에 $P(A|B)=0$ 이다. 사건 B가 발생했다는 가정하에서는 사건 A가 일어날 수 없기 때문이다. 따라서 당연히 $P(A|B) \neq P(A)$가 성립한다.

03. 확률의 법칙

■ 덧셈 법칙

- 확률의 덧셈 법칙(addition rule)은 합사건과 동일한 논리로 이해한다. A와 B의 합사건의 확률은 확률 A와 확률 B의 합에서 A와 B의 교사건의 확률을 빼서 계산한다.

확률의 덧셈 법칙

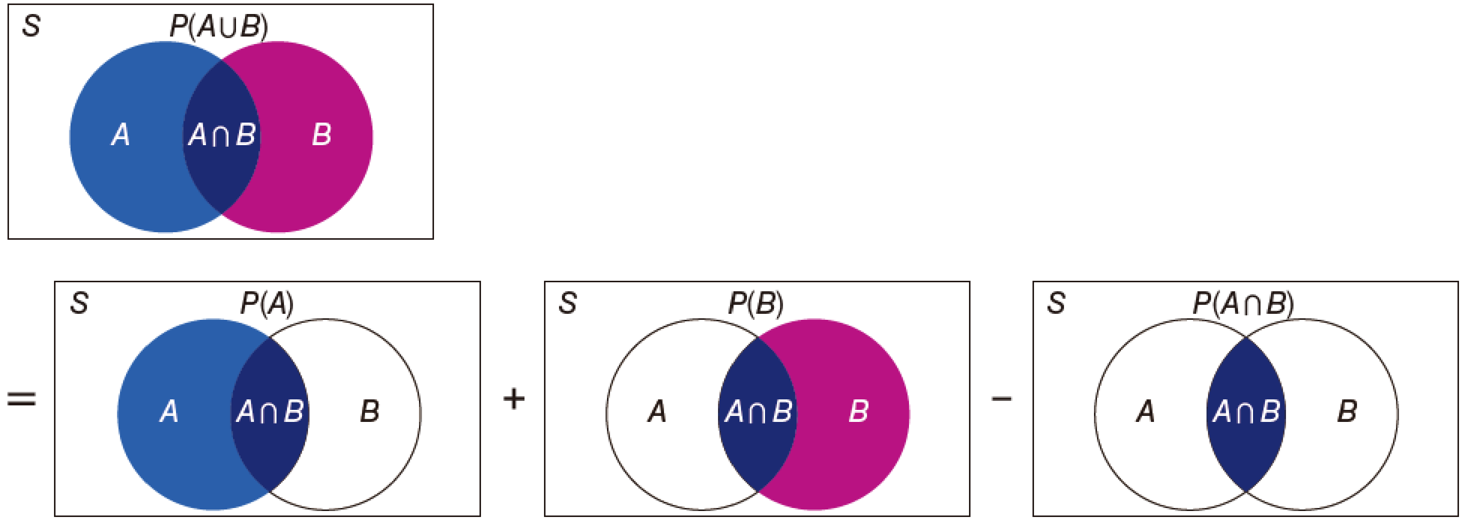
$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.20}$$

- 이는 단순히 확률 A와 확률 B를 더하면 A와 B의 교사건을 두 번 더하는 결과가 나타나 해당 교사건을 빼준 것이다. 그런데 만약 A와 B가 배반사건이라면, A와 B의 교사건이 없으므로 단순히 확률 A와 확률 B를 더하면 된다.

$$P(A \cup B) = P(A) + P(B)$$

03. 확률의 법칙

■ 덧셈 법칙



[확률의 덧셈 법칙]

03. 확률의 법칙

■ 베이즈 정리 및 활용

- **사전확률(prior probability):** 특정 사건이 일어나기 전의 확률로 관심을 갖는 사건이 일어날 가능성.
- **사후확률(posterior probability):** 관심사건의 증거에 대한 조건부확률.

■ 베이즈 정리 및 활용

예제 4-2 사전확률과 사후확률

- 스마트폰을 제조하는 S사는 서로 다른 두 개의 공장에서 스마트폰을 생산한다. 각 공장을 S_1 과 S_2 라고 하자. S_1 의 생산량은 65%, S_2 의 생산량은 35%이다. 통계적으로 어떤 스마트폰이 S_1 에서 생산될 확률은 $P(S_1) = 65\%$, S_2에서 생산될 확률은 $P(S_2) = 35\%$ 이다. 각 공장의 근무 환경이나 근무자의 숙련도 차이 등으로 불량률에 차이가 존재한다. S_1 과 S_2 에서 생산된 제품의 불량률은 각각 0.5%, 0.7%이다. 스마트폰이 불량일 사건을 A라 정의하면, $P(A|S_1) = 0.5\%$, $P(A|S_2) = 0.7\%$ 인 것이다.

① 사용자가 구입한 스마트폰이 불량품일 확률 P(A)는?

② 사용자가 구입한 스마트폰이 불량이라면 S_1 이나 S_2 중 어느 공장에서 생산된 것일까? ($P(S_1|A)$와 $P(S_2|A)$ 비교)

예제 4-2 사전확률과 사후확률

풀이

① 사전확률은 $P(S_1)$과 $P(S_2)$ 이고, 사후확률은 $P(S_1|A)$ 또는 $P(S_2|A)$ 이다. S_1 과 S_2 는 서로 배반사건이며, 두 사건의 합사건이 전체사건이다.

$$P(S_1 \cap S_2) = 0,\ P(S_1 \cup S_2) = 1$$

따라서 제품이 불량일 확률 P(A)는 S_1 에서 생산하여 불량일 확률 $P(A \cap S_1)$ 과 S_2 에서의 불량확률 $P(A \cap S_2)$ 를 합하여 계산할 수 있다.

$$P(A) = P(A \cap S_1) + P(A \cap S_2) = P(S_1)P(A|S_1) + P(S_2)P(A|S_2)$$
$$= 65\% \cdot 0.5\% + 35\% \cdot 0.7\% = 0.57\%$$

예제 4-2 **사전확률과 사후확률**

풀이

② 사후확률 $P(S_1|A)$와 $P(S_2|A)$계산은 조건부확률과 교사건의 교환법칙 S_1 활 S_2하여 다음과 같이 답을 얻을 수 있다.

$$P(S_1|A)=\frac{P(S_1\cap A)}{P(A)}=\frac{P(A\cap S_1)}{P(A)}=\frac{P(S_1)P(A|S_1)}{P(A)}=\frac{65\%\cdot 0.5\%}{0.57\%}=57.02\%$$

$$P(S_2|A)=\frac{P(S_2\cap A)}{P(A)}=\frac{P(A\cap S_2)}{P(A)}=\frac{P(S_2)P(A|S_2)}{P(A)}=\frac{35\%\cdot 0.7\%}{0.57\%}=42.98\%$$

$P(S_1|A)$와 $P(S_2|A)$의 합이 전체 확률인 1이 된다.

여기서 $P(S_1)>P(S_1|A)$ 이고 $P(S_2)<P(S_2|A)$ 임을 확인하면 S_1의 불량률이 S_2의 불량률에 비해 더 작으므로 불량 제품일 경우 S_1에서 제조되었을 확률 $P(S_1|A)$가 전체 사건 중 S_1의 생산품일 확률 $P(S_1)$에 비해 더 낮다는 결론이 나온다.

■ 베이즈 정리 및 활용

- [예제 4-2]를 기준으로 S_1 과 S_2 가 전체 사건의 분할일 경우 임의의 사건 A에 대하여 다음의 식이 성립한다.

$$P(S_1 \mid A) = \frac{P(S_1)P(A \mid S_1)}{P(S_1)P(A \mid S_1) + P(S_2)P(A \mid S_2)}$$

- 사후확률 $P(S_1 \mid A)$를 계산하는 방법을 정리한 것이 베이즈 정리이다.

베이즈 정리(Bayes theorem)

$S_1, S_2, S_3, \cdots, S_{n-1}, S_n$의 n개의 사건이 전체 사건의 분할일 경우 임의의 사건 A에 대하여 다음의 식이 성립한다.

$$P(S_k \mid A) = \frac{P(S_k)P(A \mid S_k)}{P(S_1)P(A \mid S_1) + P(S_2)P(A \mid S_2) + \cdots + P(S_n)P(A \mid S_n)} \quad (4.23)$$

■ 베이즈 정리 및 활용

예제 4-3 베이즈 정리

- 10,000명 중 1명에게 발병하는 희귀병이 있다. 희귀병 환자가 진단기를 통하여 양성 반응을 보일 확률은 99.9%이고, 해당 병과 상관없이 건강한 사람이 해당 진단기를 통하여 양성반응을 보일 확률은 1%이다. 그렇다면 만약 어떤 개인이 본 진단기를 통하여 양성반응 결과가 나왔을 경우 진짜로 자신이 희귀병에 걸렸을 확률을 계산해보자.

① 임의의 개인을 대상으로 의료진단기의 결과가 양성인 경우 진짜 희귀병에 걸렸을 확률을 계산하시오.

② 의사의 전문적 진단으로 희귀병에 걸렸을 확률이 40% 정도임을 확인하였다. 이 경우 의료진단기 결과가 양성인 경우 진짜 희귀병에 걸렸을 확률을 계산하시오.

예제 4-3 베이즈 정리

풀이

① 희귀병에 걸린 사건 A의 확률은 다음과 같다.

$$P(A) = \frac{1}{10,000}$$

희귀병 환자가 진단기의 양성반응(P)을 보일 확률 $P(P|A)$는 99.9%, $P(P|A^c)$는 1%이다. 본 문제에서 알고자 하는 바는 진단기의 결과가 양성반응(P)을 보였을 때 진짜로 희귀병에 걸렸을 확률인 $P(P|A)$이며, 이는 베이즈 정리를 이용하여 다음과 같이 계산할 수 있다.

$$P(A) = 0.0001,\ P\left(A^c\right) = 0.9999,\ P(P|A) = 0.999,\ P\left(P|A^c\right) = 0.01$$

$$\begin{aligned} P(A|P) &= \frac{P(A \cap P)}{P(P)} = \frac{P(A)P(P|A)}{P(P \cap A) + P\left(P \cap A^c\right)} = \frac{P(A)P(P|A)}{P(A)P(P|A) + P\left(A^c\right)P\left(P|A^c\right)} \\ &= \frac{0.0001 \cdot 0.999}{0.0001 \cdot 0.999 + 0.9999 \cdot 0.01} = 0.009892 = 0.9892\% \end{aligned}$$

예제 4-3 베이즈 정리

풀이

② 희귀병에 걸렸을 사건 A의 확률 $P(A)$는 의사의 진단 결과 40% 확률이었고, 희귀병 환자가 진단기의 양성반응(P)을 보일 확률 $P(P|A)$ 는 99.9%, $P(P|A^c)$ 는 1%이다.

이 문제에서 알고자 하는 바는 진단기의 결과가 양성반응(P)을 보였을 때 진짜로 희귀병에 걸렸을 확률인 $P(A|P)$ 이며, 베이즈 정리를 이용하여 계산한다.

$$P(A)=0.4,\ P(A^c)=0.6,\ P(P|A)=0.999,\ P(P|A^c)=0.01$$

$$P(A|P)=\frac{P(A\cap P)}{P(P)}=\frac{P(A)P(P|A)}{P(P\cap A)+P(P\cap A^c)}=\frac{P(A)P(P|A)}{P(A)P(P|A)+P(A^c)P(P|A^c)}$$

$$=\frac{0.4\cdot 0.999}{0.4\cdot 0.999+0.6\cdot 0.01}=0.9852=98.52\%$$

예제 4-3 **베이즈 정리**

풀이

- 이 경우 진단기의 양성 결과는 98.52%의 희귀병 확률을 나타낸다. 이를 기반으로 본 진단기는 관심을 갖는 사건의 확률을 증폭시키는 역할을 한다고 볼 수 있다. (1)의 결과도 1%에 미치지 못하는 확률이 나타나지만 본래의 확률인 0.01%보다는 훨씬 더 높은 결과를 보인다. (2)의 결과도 40%의 사전확률이 진단기의 결과로 98.52%로 증폭된다.
- (1)의 답이 우리에게 시사하는 바가 있다. 이와 같이 만 명 중 한 명이 걸릴 수 있는 희귀병의 경우 일반 대중을 대상으로 진단을 하지는 않는다. 일반 대중을 상대로 검사를 실시하는 경우 질병에 걸릴 확률이 너무 낮아서 진단기의 결과 양성이 나왔을 경우 진짜로 병에 걸렸을 가능성도 원래의 확률보다는 높다 하더라도 실효성이 없을 만큼 낮은 경우가 많다. 반면 의사의 진단을 통하여 질병에 걸렸을 가능성이 어느 정도 있을 때는 해당 진단기의 결과가 매우 유용하다. 이것이 의료진단기가 주로 병원에만 있고 길거리에서 만인을 대상으로 진단기를 사용하지 않는 이유이다.

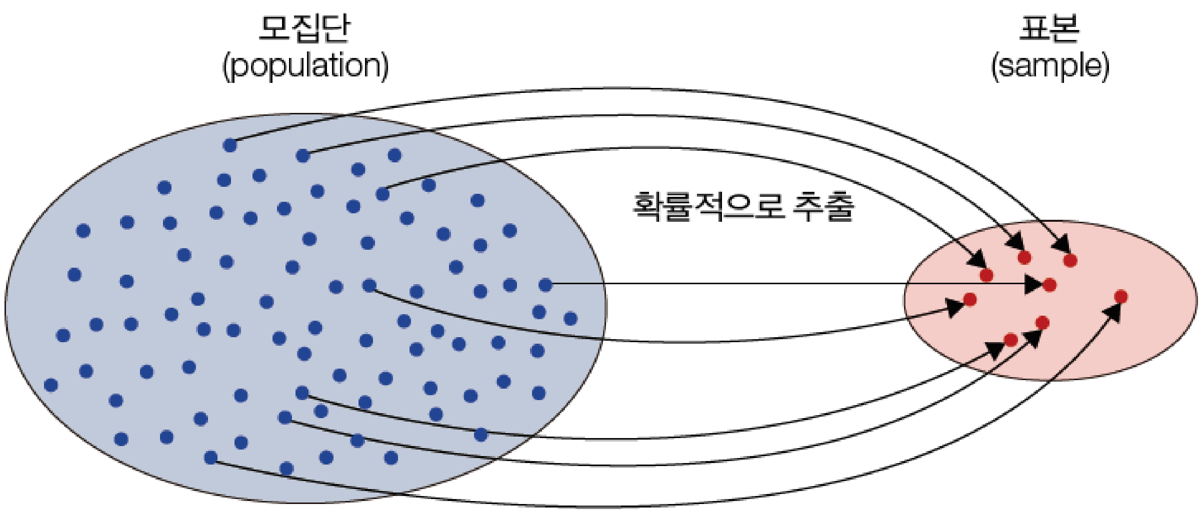
01

표본추출

01. 표본추출

■ 모집단과 표본의 관계

- **모집단(population):** 관심의 대상이 되는 집단의 모든 관측값 또는 측정값들의 집합
- **표본(sample):** 관심의 대상이 되는 모집단으로부터 추출된 값들의 집합. 즉, 표본은 모집단의 부분집합.



[모집단과 표본의 관계]

01. 표본추출

표본추출의 필요성

- 모집단의 평균이나 분산 등의 특성을 알고 싶은 상황에서 모집단에 대한 정보를 빠르게 얻고자 한다면 표본이 필요하다.

① 모집단 전체를 조사하기 위해서는 많은 비용이 소요된다.

② 시간적 제약이 있을 수 있다.

③ 애초에 모집단의 정보를 수집하는 것이 불가능한 경우도 있다.

④ 분석을 통하여 피실험대상품의 파손이 전제될 경우 모집단에 대한 분석은 의미가 없어진다.

01. 표본추출

■ 표본추출의 방법

- 표본추출은 크게 무작위추출과 유의추출로 나눌 수 있다.
- **무작위추출(random sampling):** 추출자의 의도를 완전히 배제하여 임의로 추출하는 경우.
- **유의 추출(purposive selection):** 표본을 주관적으로 선택하여 추출하는 방법.
- 무작위추출에는 다음과 같은 방법이 있다.

  ① 단순무작위추출법: 모든 표본에 대하여 동일한 가능성 하에서의 추출

  ② 계통추출법: 임의로 하나의 표본을 추출한 후 그 다음 매 n번째 표본을 추출

  ③ 층화추출법: 모집단을 몇 개의 그룹으로 나누어 각 그룹별로 무작위 추출

  ④ 다단계추출법: 2단계 이상의 추출 단계를 설정하여 무작위 추출

■ 표본추출의 방법

하나 더 알기 표본분포와 표본통계량분포

- 표본분포는 추출한 표본 자체에 대한 분포로 표본의 평균, 분산, 표준편차 등의 통계량을 관찰할 수 있다. 이번 장에서 다루게 될 표본통계량분포는 크기가 n인 표본을 모든 가능한 경우로 추출하였을 때 관찰되는 통계량(평균 등)에 대한 분포를 의미한다. 따라서 표본통계량분포를 산출하기 위해서는 여러 번의 표본추출이 있어야 가능하다. 다음 절에서부터 표본의 평균, 비율, 분산에 대한 분포를 고려할 것이며, 이를 표본평균분포, 표본비율분포, 표본분산분포라 한다.

02

표본평균분포

02. 표본평균분포

■ 표본평균분포의 평균, 분산, 표준편차

- **표본평균분포(sampling distribution of the sample mean):** 특정 크기의 모든 가능한 표본들로부터 계산된 표본평균들의 분포를 의미.

[표8-1] 수강생 6명의 중간시험성적

수강생 명단	중간시험성적
주연	95
민환	76
영진	84
여준	89
지원	91
현미	81

02. 표본평균분포

■ 표본평균분포의 평균, 분산, 표준편차

- 모집단의 평균

$$\frac{95+76+84+89+91+81}{6}=86.0$$

- 모집단의 분산

$$\frac{1}{6}\left\{(95-86)^2+(76-86)^2+(84-86)^2+(89-86)^2+(91-86)^2+(81-86)^2\right\}=40.6667$$

- 표준편차

$$\sqrt{40.6667}=6.3770$$

표본평균분포의 평균, 분산, 표준편차

[표8-2] 수강생 6명의 중간고사 성적을 대상으로 크기 2의 표본에 대한 중간시험성적 평균분포

수강생 명단	중간성적	평균	수강생 명단	중간성적	평균
(주연, 주연)	(95, 95)	95.0	(여준, 주연)	(89, 95)	92.0
(주연, 민환)	(95, 76)	85.5	(여준, 민환)	(89, 76)	82.5
(주연, 영진)	(95, 84)	89.5	(여준, 영진)	(89, 84)	86.5
(주연, 여준)	(95, 89)	92.0	(여준, 여준)	(89, 89)	89.0
(주연, 지원)	(95, 91)	93.0	(여준, 지원)	(89, 91)	90.0
(주연, 현미)	(95, 81)	88.0	(여준, 현미)	(89, 81)	85.0
(민환, 주연)	(76, 95)	85.5	(지원, 주연)	(91, 95)	93.0
(민환, 민환)	(76, 76)	76.0	(지원, 민환)	(91, 76)	83.5
(민환, 영진)	(76, 84)	80.0	(지원, 영진)	(91, 84)	87.5
(민환, 여준)	(76, 89)	82.5	(지원, 여준)	(91, 89)	90.0
(민환, 지원)	(76, 91)	83.5	(지원, 지원)	(91, 91)	91.0
(민환, 현미)	(76, 81)	78.5	(지원, 현미)	(91, 81)	86.0
(영진, 주연)	(84, 95)	89.5	(현미, 주연)	(81, 95)	88.0
(영진, 민환)	(84, 76)	80.0	(현미, 민환)	(81, 76)	78.5
(영진, 영진)	(84, 84)	84.0	(현미, 영진)	(81, 84)	82.5
(영진, 여준)	(84, 89)	86.5	(현미, 여준)	(81, 89)	85.0
(영진, 지원)	(84,91)	87.5	(현미, 지원)	(81, 91)	86.0
(영진, 현미)	(84,81)	82.5	(현미, 현미)	(81, 81)	81.0

02. 표본평균분포

■ 표본평균분포의 평균, 분산, 표준편차

- 표본평균의 평균

$$\frac{1}{36}(95.0+85.5+89.5+92.0+93.0+88.0+85.5+76.0+80.0+82.5+83.5$$
$$+78.5+89.5+80.0+84.0+86.5+87.5+82.5+92.0+82.5+86.5+89.0$$
$$+90.0+85.0+93.0+83.5+87.5+90.0+91.0+86.0+88.0+78.5+82.5$$
$$+85.0+86.0+81.0)=86.0$$

- 표본평균의 분산

$$\frac{1}{36}\{(95.0-86)^2+(85.5-86)^2+(89.5-86)^2+\cdots\cdots\cdot+(85.0-86)^2$$
$$+(86.0-86)^2+(81.0-86)^2\}=20.3333$$

- 표준편차

$$\sqrt{20.3333}=4.5092$$

02. 표본평균분포

■ 표본평균분포의 평균, 분산, 표준편차

- 모집단의 평균, 분산, 표준편차를 각각 다음과 같이 나타낸다.

$$\mu_X = E(X),\ \sigma_X{}^2 = Var(X),\ \sigma_X = \sigma(X)$$

- 표본평균의 평균, 분산, 표준편차는 다음과 같이 나타낸다.

$$\mu_{\overline{X}} = E(\overline{X}),\ \sigma_{\overline{X}}{}^2 = Var(\overline{X}),\ \sigma_{\overline{X}} = \sigma(\overline{X})$$

표본평균분포의 평균, 분산, 표준편차

- [표 8-3]에서 확인할 수 있는 바는 모집단의 평균과 표본평균분포의 평균이 동일하다는 것이다. 또한 분산은 모집단의 분산을 표본의 크기로 나누어 준 값과 동일하다.

[표8-3] 모집단과 크기가 2인 표본평균분포의 통계량 비교

모집단의 통계량		표본평균분포의 통계량	
평균 $\mu_X = E(X)$	86.0	평균 $\mu_{\overline{X}} = E(\overline{X})$	86.0
분산 $\sigma_X^2 = Var(X)$	40.6667	분산 $\sigma_{\overline{X}}^2 = Var(\overline{X})$	20.3333
표준편차 $\sigma_X = \sigma(X)$	6.3770	표준편차 $\sigma_{\overline{X}} = \sigma(\overline{X})$	4.5092

02. 표본평균분포

■ 표본평균분포의 평균, 분산, 표준편차

- 표본평균분포의 분산 $\sigma_{\overline{X}}^{\ 2} = 20.3333$ 은 모분산을 표본의 크기 2로 나누어 준 다음 수식과 같다.

$$\sigma_{\overline{X}}^{\ 2} = \frac{\sigma_X^{\ 2}}{2} = \frac{40.6667}{2} = 20.3333$$

- 표본평균분포의 표준편차 $\sigma_{\overline{X}} = 4.5092$ 는 모표준편차를 2로 나누어 준 다음 수식과 같다.

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{2}} = \frac{6.3770}{\sqrt{2}} = 4.5092$$

표본평균분포의 평균, 분산, 표준편차(표본의 크기 n, 모집단과 비교)

$$\mu_{\overline{X}} = \mu_X, \ \ \sigma_{\overline{X}}^{\ 2} = \frac{\sigma_X^{\ 2}}{n}, \ \ \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \tag{8.9}$$

표본평균분포의 평균, 분산, 표준편차

- 표본의 크기가 n인 일반적인 상황에서 표본평균분포의 통계량을 모집단의 통계량으로 표현한 것이다.

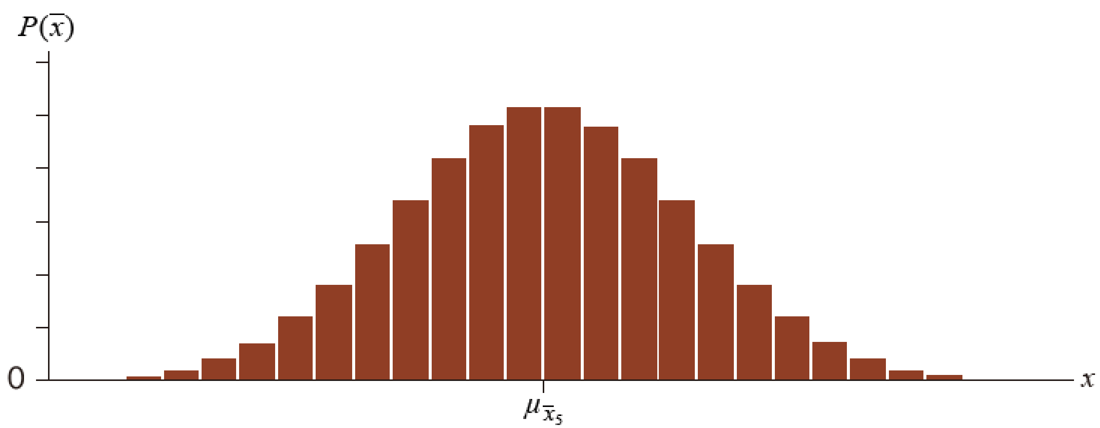
표본평균분포의 평균, 분산, 표준편차(표본의 크기 n, 모집단과 비교)

$$\mu_{\overline{X}} = \mu_X,\ \ \sigma_{\overline{X}}^{\ 2} = \frac{\sigma_X^{\ 2}}{n},\ \ \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \tag{8.9}$$

02. 표본평균분포

표본평균분포의 평균, 분산, 표준편차

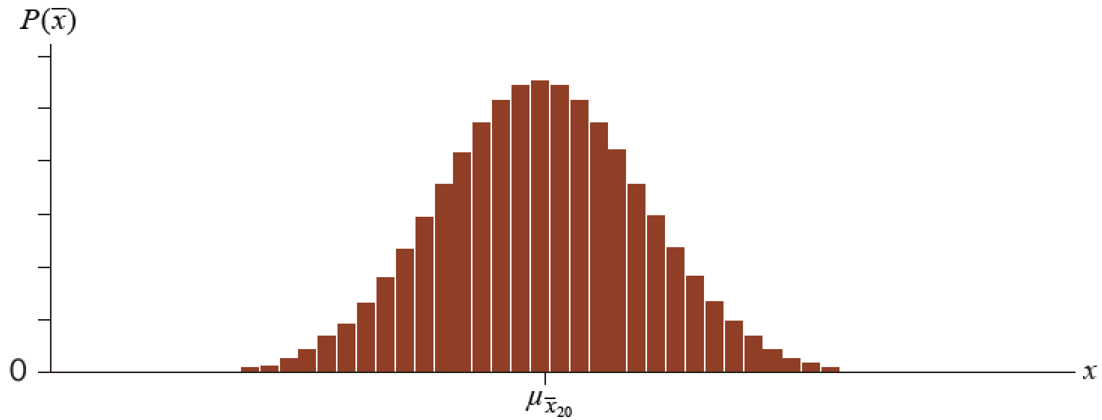
- 표본평균분포의 평균은 모집단의 평균과 일치한다. 분산은 n의 크기에 따라 달라지며, n이 커지면 커질수록 분산은 작아짐을 확인할 수 있다. 따라서 n이 커짐에 따라 표본평균분포는 그림과 같이 변화되는 추이를 보인다.



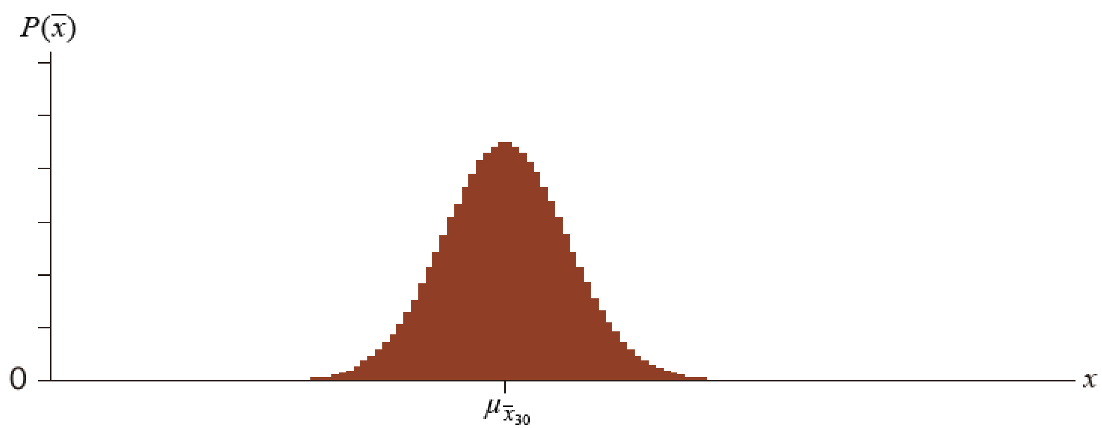
(a) $n = 5$일 때 표본평균 $\overline{X}$의 분포

02. 표본평균분포

■ 표본평균분포의 평균, 분산, 표준편차



(b) n = 20일 때 표본평균 $\overline{X}$의 분포



(c) n = 30일 때 표본평균 $\overline{X}$의 분포

[n의 크기에 따른 표본평균분포의 형태 비교]

02. 표본평균분포

■ 표본의 크기에 따른 표본평균분포의 통계량

- n의 크기에 따른 표본평균분포의 표기 방법을 알아보자. 표본의 크기가 2이면 다음과 같이 나타낸다.

$$\overline{X_2} = \frac{X_1 + X_2}{2}$$

- 유사하게 표본의 크기가 n인 경우 표본평균의 확률변수는 다음과 같다.

표본평균의 확률변수(표본의 크기 n)

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{8.11}$$

02. 표본평균분포

■ 표본의 크기에 따른 표본평균분포의 통계량

[표8-4] 6명 수강생의 중간시험성적에 대한 표본의 크기에 따른 표본평균분포의 통계량 비교

표본 크기	표본의 크기 $n=1$	표본의 크기 $n=2$	표본의 크기 n
확률변수	$\overline{X}=X$	$\overline{X}_2=\dfrac{X_1+X_2}{2}$	$\overline{X}_n=\dfrac{X_1+X_2+\cdots+X_n}{n}$
평균	$\mu_X=86.0$	$\mu_{\overline{X}_2}=86.0$	$\mu_{\overline{X}_n}=86.0$
분산	$\sigma_X{}^2=40.6667$	$\sigma_{\overline{X}_2}{}^2=\dfrac{40.6667}{2}=20.3333$	$\sigma_{\overline{X}_n}{}^2=\dfrac{\sigma_X{}^2}{n}=\dfrac{40.6667}{n}$
표준편차	$\sigma_X=6.3770$	$\sigma_{\overline{X}_2}=\dfrac{6.3770}{\sqrt{2}}=4.5092$	$\sigma_{\overline{X}_n}=\dfrac{\sigma_X}{\sqrt{n}}=\dfrac{6.3770}{\sqrt{n}}$

■ 중심극한정리

- **중심극한정리(central limit theorem):** 동일한 분포를 가지는 분포들의 평균은 그 개수가 많아지면서 언제나 정규분포로 수렴한다는 것.

중심극한정리(central limit theorem)

표본의 크기가 n인 표본평균의 확률변수 $\overline{X_n}$는 다음과 같이 정의된다.

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{8.12}$$

여기서 n이 충분히 크면 $\overline{X_n}$는 정규분포에 가까워진다.

① 모집단이 정규분포를 따를 경우, 표본의 크기 n에 상관없이 $\overline{X_n}$는 정규분포를 따른다.

② 모집단이 정규분포를 따르지 않는 경우, 표본의 크기 n이 충분히 크면 $\overline{X_n}$는 근사적으로 정규분포를 따른다.

■ 중심극한정리

- 표본평균분포에 대하여 모집단이 정규분포이든 정규분포를 따르지 않든 표본의 수를 늘리면 정규분포를 따른다고 가정하여 문제를 해결할 수 있다. 즉, n이 충분히 클 때 표본평균의 확률변수 $\overline{X_n}$ 는 평균이 μ , 표준편차가 $\frac{\sigma_X}{\sqrt{n}}$ 인 정규분포를 따른다. 이를 기호로 나타내면 다음과 같다.

$$\overline{X_n} \sim N\left(\mu, \left(\frac{\sigma_X}{\sqrt{n}}\right)^2\right)$$

- 또한 표준화 과정을 통하여 새로운 확률변수 $\frac{\overline{X_n}-\mu}{\frac{\sigma_X}{\sqrt{n}}}$ 는 평균이 0, 표준편차가 1인 표준정규분포를 따른다.

$$Z = \frac{\overline{X_n}-\mu}{\frac{\sigma_X}{\sqrt{n}}} \sim N\left(0, 1^2\right)$$

■ 중심극한정리

예제 8-1 표본평균분포 I

어떤 커피머신은 매번 에스프레소 30ml를 추출한다. 해당 커피머신이 한 번에 추출하는 에스프레소의 양은 평균 30ml, 표준편차 2ml인 정규분포를 따른다. 커피를 매우 좋아하는 소비자 A는 매번 4번의 샷을 추출하여 아메리카노, 라떼 등으로 커피를 마신다. 4번의 추출된 에스프레소 양의 평균인 확률변수 $\overline{X}$에 대하여 다음의 물음에 답하시오.

(1) 확률변수 $\overline{X}$의 평균, 분산, 표준편차를 계산하시오.

(2) 4번에 걸쳐 추출된 에스프레소 양의 평균이 31ml 이상일 확률을 계산하시오.

■ 중심극한정리

예제 8-1 표본평균분포 I

➡ 풀이

(1) 표본평균분포는 모집단의 평균이 동일하므로 30ml이고, 표본이 4인 표본평균분포이므로 다음과 같이 분산과 표준편차를 계산할 수 있다.

$$\mu_{\overline{X}} = E\left(\overline{X}\right) = 30,\ \ \sigma_{\overline{X}}{}^{2} = \frac{2^{2}}{4} = 1,\ \ \sigma_{\overline{X}} = \frac{2}{\sqrt{4}} = 1$$

∴ 표본평균분포의 평균 30, 분산 1, 표준편차 1

예제 8-1 표본평균분포 I

풀이

(2) 4번에 걸쳐 추출된 에스프레소 양의 평균 확률분포 $\overline{X}$는 $\frac{X_1+X_2+X_3+X_4}{4}$ 이고, 해당확 률은 P($\overline{X}$ > 31) 이다. 모집단의 확률분포 X가 정규분포를 따르므로 중심극한정리에 의하여 $\overline{X}$ 도 정규분포를 따른다. 따라서 정규분포의 확률을 계산하는 방식에 따라 표준화하여 다음과 같이 확률을 계산할 수 있다.

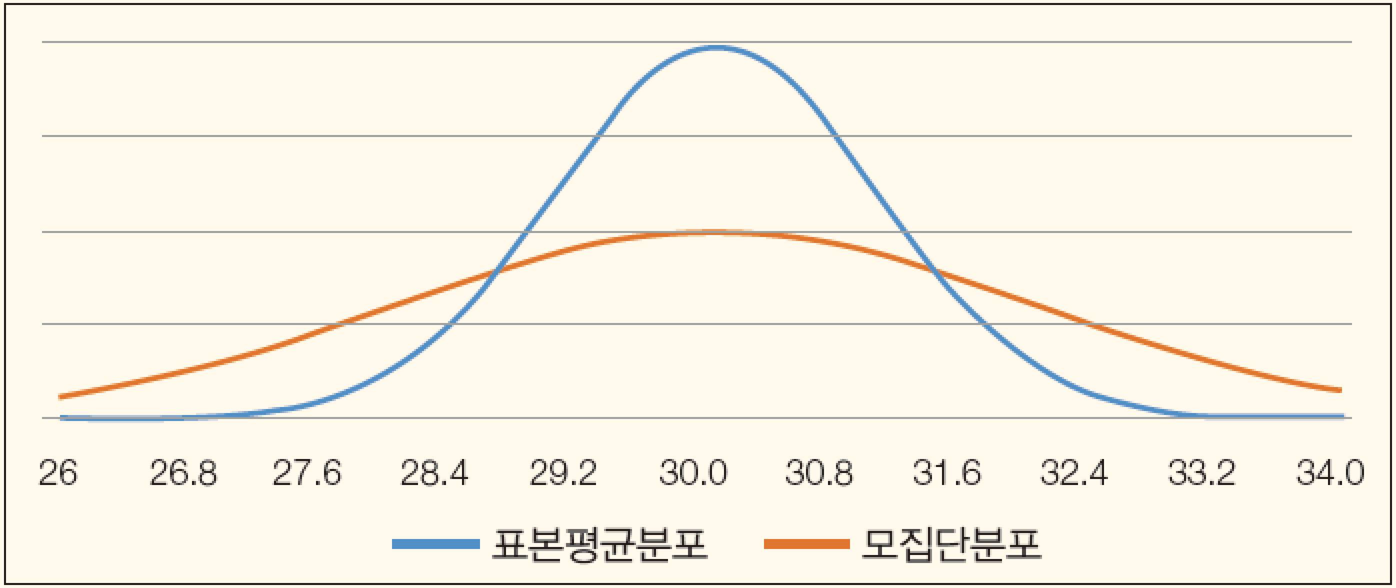
$$P(\overline{X}>31)=P\left(\frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{X}}}>\frac{31-\mu_{\overline{X}}}{\sigma_{\overline{X}}}\right)=P\left(\frac{\overline{X}-30}{1}>\frac{31-30}{1}\right)=P(Z>1)=1-P(Z\le 1)$$

표준정규분포의 확률분포표에서 P(Z ≤ 1) = 0.8413임을 확인할 수 있으며, 이 문제에서 원하는 확률은 1 - P(Z ≤ 1) = 0.1587임을 알 수 있다.

∴ 4번에 걸쳐 추출된 에스프레소 양의 평균이 31ml 이상일 확률은 15.87%이다.

예제 8-1 표본평균분포 I

풀이



모집단 $X \sim N(30, 2^2)$과 표본평균 $\overline{X} \sim N(30, 1^2)$의 각 정규분포의 형태 비교

■ 중심극한정리

예제 8-2 표본평균분포 II

- A 특별시의 주택보유자가 부담해야 하는 보유세는 평균 60만 원, 표준편차는 9만 원이고, 세금의 특성상 해당 확률분포는 정규분포를 따르지 않는다(일반적으로 세금은 적게 내는 사람의 수가 다수이고 많이 내는 사람이 소수이다). A 특별시의 주택보유자 중 36명의 표본을 추출하여 보유세의 평균값이 62만 원 이상일 확률을 계산하시오(단, 36명의 표본은 표본평균분포의 정규성을 확보할 만큼 충분하다).

중심극한정리

예제 8-2 표본평균분포 II

풀이

모집단인 A 특별시 주택보유자의 보유세에 대한 확률변수를 X, 36명의 표본평균분포에 대한 확률변수를 X라고 하자. 문제에서 묻는 바는 $P(\overline{X} \geq 62)$ 인 확률이다. 우선 X의 평균과 표준편차값을 다음과 같이 계산할 수 있다.

$$E(\overline{X}) = 60$$

$$\sigma_{\overline{X}} = \frac{9}{\sqrt{n}} = \frac{9}{\sqrt{36}} = 1.5000$$

$P(\overline{X} \geq 62)$의 계산을 위하여 X를 표준화하여 표준정규분포표를 이용한다.

$$P(\overline{X} > 62) = P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} > \frac{62 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) = P\left(\frac{\overline{X} - 60}{1.5} > \frac{62 - 60}{1.5}\right) = P(Z > 1.33) = 1 - P(Z \leq 1.33)$$

02. 표본평균분포

중심극한정리

예제 8-2 표본평균분포 II

풀이

표준정규분포의 확률분포표에서 $P(Z \le 1.33) = 0.9082$임을 확인할 수 있으며,
이 문제에서 원하는 확률은 $1 - P(Z \le 1.33) = 0.0918$임을 알 수 있다.

∴ 표본 36명의 보유세 평균이 62만 원 이상일 확률은 9.18%이다.