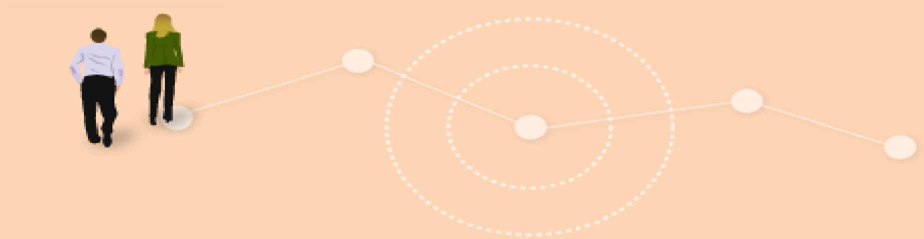
Chapter 04

확률과 통계 Probability and Statistics

목 차

01 확률과 의사결정

:: Keywords 확률 | 확률변수 | 확률함수



확률과 의사결정

■ 통계의 목적

표본으로부터 모수를 추정하기 위함

■ 추정의 이유

❶ 모집단을 대상으로 하는 조사가 불가능하거나

❷ 시간과 비용 등의 물리적 한계 때문

확률과 의사결정

월드컵과 치킨집

[그림 4-1] **월드컵과 치킨**

월드컵 기간에 판매되는 치킨 양은 평소 판매량을 훨씬 뛰어넘는다. 치킨집 점주는 월드컵이라는 특수한 이벤트에 대비하여 재료를 얼마나 준비해야 할까?

평소 치킨이 50마리 판매된다고 가정했을 때, 월드컵 기간에 100마리를 준비하는 경우와 120마리 혹은 140마리를 준비하는 경우 중 어느 쪽이 더 많은 수익을 창출할 수 있을까? 많이 준비해서 모두 팔 수 있으면 다행이겠지만, 다 팔리지 않는다면 남은 재료는 모두 손실로 연결될 것이다.

통계는 이런 상황에서 해결책을 제시한다. 과거 월드컵 기간의 판매량과 월드컵 기간 전후의 판매량을 현재와 비교하여 예를 들어 '135마리'라는 구체적인 수치를 도출할 수 있을 것이다. 그런데 막상 월드컵 기간에 치킨을 팔아보니 135마리가 아니었다면 이 예측은 무의미한 걸까? 그렇지 않다. 135라는 수치를 제시하면서, 이 수치가 맞을 확률을 같이 나타내면 된다. 135마리를 기준으로 132, 133, 134, 136, 137마리일 확률 정보가 같이 제시된다면, 점주는 이 데이터를 바탕으로 최적의 의사결정을 통해 손실을 최소화할 수 있다.

확률론

아무리 정교하게 분석된 통계자료일지라도 100% 맞을 수는 없기 때문에,

→ 그 결과를 확률(probability)과 함께 표현

일정 조건 하에서 동일한 실험을 지속적으로 N회 반복했을 때,

사건 A가 n번 발생할 확률은

$$P(A) = \frac{n(A)}{N}$$

Ex. 동전을 두 번 던질 때 앞면이 나올 확률

풀이) (앞면이 두 번 모두 나오는 경우) = 1

(앞면이 한 번만 나오는 경우) = 2

(앞면이 두 번 모두 나오지 않는 경우) = 1

→ $n(A) = 3$

$$P(A) = \frac{n(A)}{N} = \frac{3}{4}$$

■ 확률이 가지는 조건

❶ 확률은 0 ~ 1의 값을 가진다.

❷ 모든 사건에 대한 확률의 합은 1이다.

$$\sum_{i=1}^{n} P(E_i) = 1 \quad (E : \text{사건(Event)},\ i : \text{시행 횟수},\ P : \text{확률})$$

통계적 확률

통계적 확률은 기본적인 확률의 개념과 같다. 다만 반복적인 실행을 n번 해서 사건 A가 일어난 횟수를 r이라 했을 때, n을 충분히 크게 한다면 상대도수로 나타나는 $\frac{r}{n}$은 일정한 확률값 p로 근사하게 된다. 이때 p를 사건 A가 발생할 **통계적 확률** 혹은 **경험적 확률**이라 한다.

■ 확률의 덧셈법칙 (addition rule of probability)

서로 다른 사건 A와 B가 발생할 때, A 혹은 B가 일어날 확률

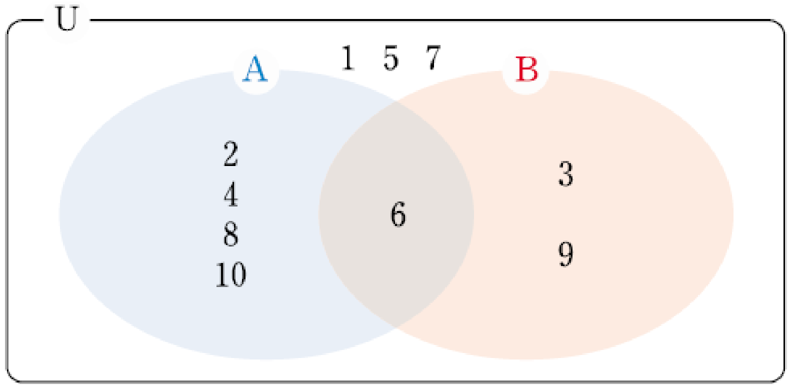
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

cf. 배반사건이라면, 중복되는 부분이 없이 $P(A \cap B)=0$이므로

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

확률

Ex. 1부터 10까지 숫자가 적힌 공 10개가 들어 있는 주머니 안에서 공 1개를 뽑을 때 2의 배수가 나오는 사건을 A, 3의 배수가 나오는 사건을 B라 하면, 두 사건이 발생하는 경우는 다음과 같다.

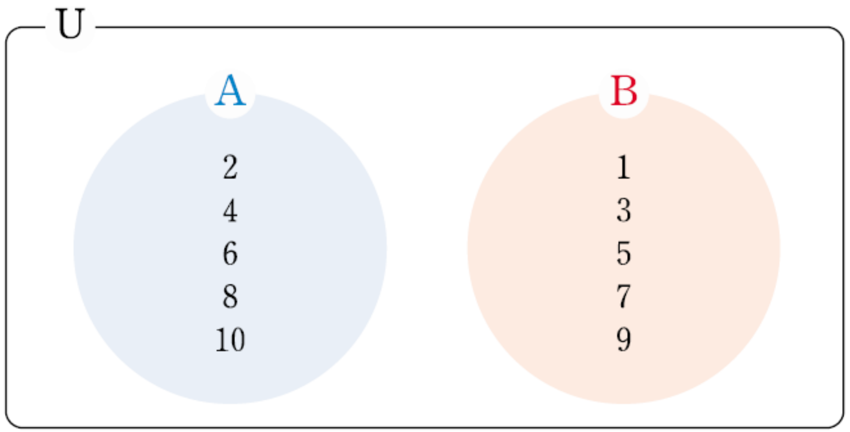


사건 A 혹은 B가 발생할 확률은 A의 확률과 B의 확률을 더한 후 공통으로 발생하는 사건인 6을 빼 주어야 함.

$$
\begin{aligned}
P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\
&= \frac{5}{10} + \frac{3}{10} - \frac{1}{10} = \frac{7}{10}
\end{aligned}
$$

확률

Ex. 짝수가 나오는 사건을 A, 홀수가 나오는 사건을 B라 하면, 두 사건이 발생하는 경우는 다음과 같이 $P(A \cap B) = 0$ 이다.



따라서 사건 A 혹은 B가 발생할 확률은

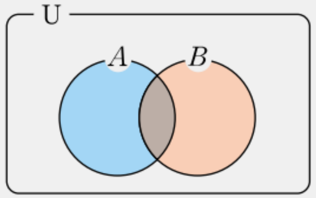
$$P(A \cup B) = P(A) + P(B) = \frac{5}{10} + \frac{5}{10} = 1$$

확률

Note 확률 사건

- **합사건($A \cup B$)**

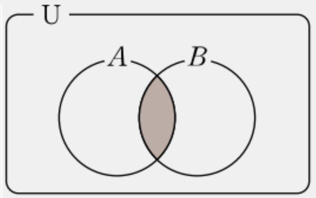
  사건 A와 B에 대해 A가 발생하거나 B가 발생하는 사건을 A와 B의 합사건이라 한다.



[그림 4-4] 합사건

- **곱사건($A \cap B$)**

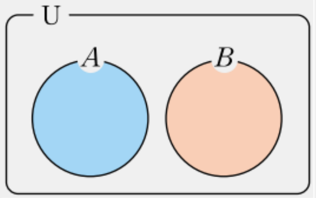
  사건 A와 B에 대해 A와 B가 동시에 발생하는 사건을 A와 B의 곱사건이라 한다.



[그림 4-5] 곱사건

- **배반사건($A \cap B = \varnothing$)**

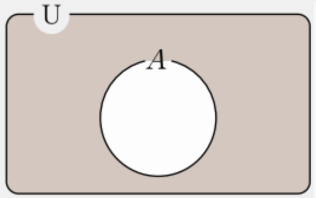
  사건 A와 B에 대해 A나 B 중 어느 하나의 사건이 발생하면 다른 사건이 발생하지 않는 사건을 배반사건이라 한다.



[그림 4-6] 배반사건

- **여사건(A^c)**

  사건 A에 대해 A가 발생하지 않는 사건을 A의 여사건이라 한다.



[그림 4-7] 여사건

예제 4-1 주사위 던지기의 확률 계산

주사위를 두 번 던져서 나오는 결과의 합이 10 이상이 될 확률을 구하라.

확률 (예제 풀이)

주사위를 두 번 던져서 합이 10 이상이 나오는 경우는

$$(4,\ 6),\ (6,\ 4),\ (5,\ 5),\ (5,\ 6),\ (6,\ 5),\ (6,\ 6)$$

합이 10이 되는 경우의 확률

(4, 6)인 경우 : $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$

(5, 5)인 경우 : $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$

(6, 4)인 경우 : $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$

합이 11이 되는 경우의 확률

(5, 6)인 경우 : $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$

(6, 5)인 경우 : $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$

합이 12가 되는 경우의 확률

(6, 6)인 경우 : $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$

$\therefore$ 모든 확률을 더하면 $\frac{6}{36}=\frac{1}{6}$

■ 조건부 확률과 확률의 곱셈법칙(multiplication rule of probability)

확률의 덧셈법칙에서는 사건 A와 B가 발생하는 과정에 순서라는 개념이 없지만, 만약 A가 발생한 상황 하에 B가 발생할 확률을 구한다면 확률은 $P(B|A)$로 표시하고, 조건부 확률(conditional probability)이라고 한다.

조건부 확률은 $P(B|A)=\dfrac{P(B\cap A)}{P(A)}$ 으로 구하며 $P(B\cap A)=P(A)\cdot P(B|A)$으로 표현되며, 이를 확률의 곱셈법칙이라 한다.

예를 들어, OO학과의 남녀 학생 100명 중 남학생의 수는 55명이고 이 중 통계학에 관심이 있는 남학생의 수가 20명이라고 하자. 이때 남학생을 선택하는 사건을 A라 하고 남학생 중 통계학에 관심이 있는 학생을 선택하는 사건을 B라 했을 때, 남학생을 선택할 확률 $P(A)=\dfrac{55}{100}$이고, 남학생 중 통계학에 관심이 있는 학생을 선택할 확률 $P(B|A)=\dfrac{P(A\cap B)}{P(A)}=\dfrac{20}{55}$이다. 확률의 곱셈법칙을 적용하면, $P(A\cap B)=P(A)\cdot P(B|A)=\dfrac{55}{100}\times\dfrac{20}{55}=\dfrac{1}{5}$이다.

예제 4-2 주머니 속 공 선택의 확률 계산하기(확률의 곱셈법칙)

주머니 속에 파란 공 5개와 노란 공 7개가 들어 있다. 공을 두 번 꺼낼 때, 모두 파란 공인 경우의 복원추출 확률과 비복원추출 확률을 구하라.[1]

확률

• 복원추출

$$\text{공의 총 개수} = 12,\ \text{파란 공} = 5,\ \text{노란 공} = 7$$

파란 공을 뽑고 난 후 다시 파란 공을 뽑는 조건부 확률이므로 다음 식을 적용한다.

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

곱셈법칙에 의해 계산하면 다음과 같다.

$$P(A \cap B) = P(A) \cdot P(B|A) = \frac{5}{12} \times \frac{5}{12} = \frac{25}{144} = 0.174\ (\text{소수점 4자리 반올림})$$

• 비복원추출

$$P(A \cap B) = P(A) \cdot P(B|A) = \frac{5}{12} \times \frac{4}{11} = \frac{20}{132} = 0.152\ (\text{소수점 4자리 반올림})$$

확률변수와 확률함수

■ 확률변수(random variable)

실험의 결과(사건)에 실수값을 대응시키고 그 값에 확률을 부여한 것

■ 확률함수(probability function)

반복적으로 어떤 실험을 할 때, 각각의 실험 결과가 어떨지는 그 순간에는 알 수 없지만, 실험을 다 마친 후에는 어떤 결과가 몇 번씩 발생했는지를 총체적으로 살펴볼 수 있는데, 이 결과의 수에 확률이 부여된 것

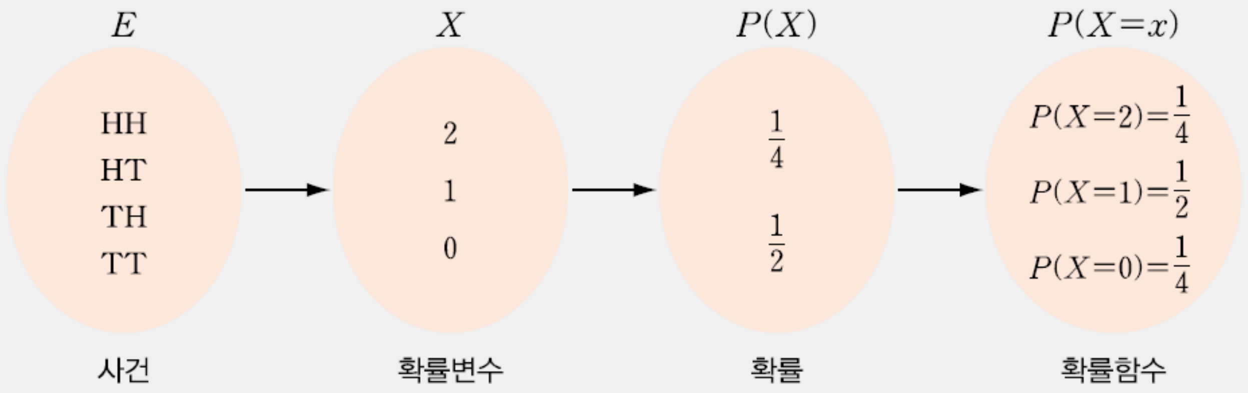
- **이산 확률변수**(discrete random variable) : 수집된 데이터의 확률변수 중 셀 수 있는 특정한 값들로 구성되거나 일정한 범위로 나타나는 확률변수
- **연속 확률변수**(continuous random variable) : 연속형이거나 무한한 경우와 같이 셀 수 없는 확률변수

확률변수와 확률함수

Note **사건, 확률변수, 확률, 확률함수의 관계**

동전을 두 번 던질 때 앞면이 나오는 경우를 기준으로 표본, 확률변수, 확률, 확률함수의 관계를 살펴보자. 동전을 두 번 던졌을 때 발생되는 사건은 HH, HT, TH, TT로 총 4가지이다. 앞면이 몇 번 나올 것인가를 기준으로 사건을 분류하면, 확률변수는 HH는 앞면이 2번이므로 2, HT와 TH는 앞면이 1번이므로 1, TT는 앞면이 나오지 않으므로 0이 된다.

확률변수가 2인 경우는 H가 나올 확률 $\frac{1}{2}$이 모두 나와야하므로 $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$, 확률변수가 1인 경우는 두 가지이므로 각각의 확률 $\frac{1}{4}$을 더해 $\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$이 된다. 확률변수가 0인 경우는 T라는 확률 $\frac{1}{2}$이 모두 나와야하므로 $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$이 된다. 확률함수로 표현하면 $P(X=2)=\frac{1}{4}$, $P(X=1)=\frac{1}{2}$, $P(X=0)=\frac{1}{4}$로 표현할 수 있다. 이를 정리하면 [그림 4-4]와 같다.



[그림 4-4] 동전 던지기의 사건, 확률변수, 확률, 확률함수의 관계

확률변수와 확률함수

예제 4-2 윷놀이의 확률변수 계산

말 4개를 가지고 윷놀이를 할 때, 말을 최소한으로 움직이면서 가장 빨리 이기는 경우의 확률변수에 대해 설명하라. 또한 이에 대한 확률과 확률함수를 구하라.

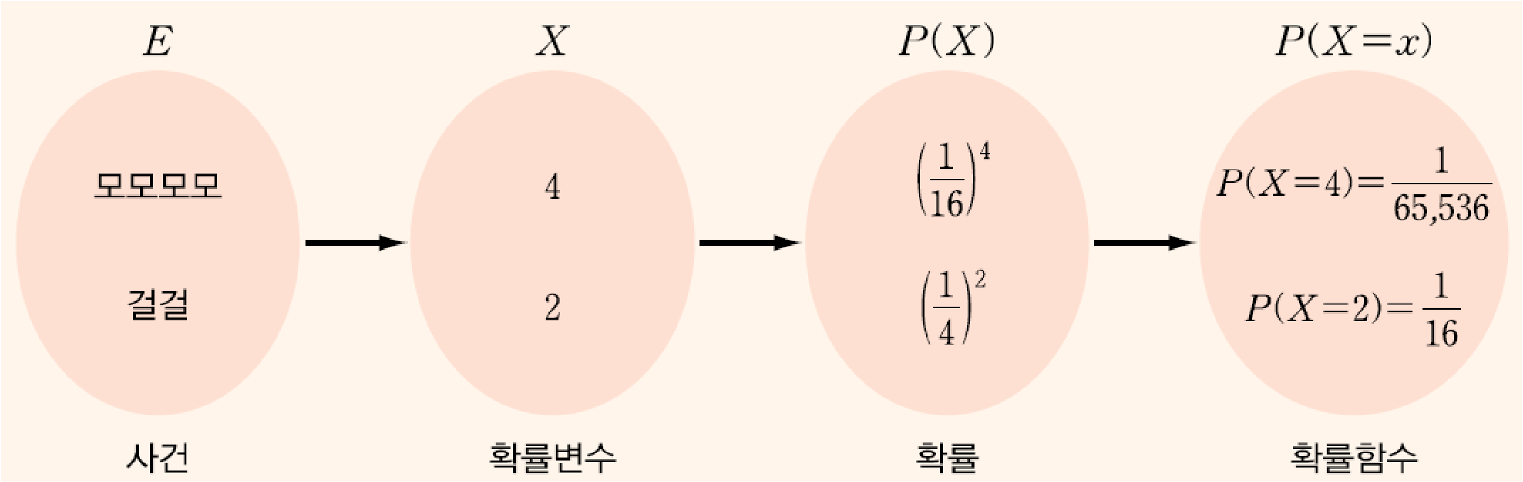
확률변수와 확률함수 (예제 풀이)

윷놀이에서 가장 빨리 돌아서 나오는 경우 : 모, 걸, 걸

말이 4개 : 가장 빨리 이기는 경우 → '모'가 4번 나온 후에 '걸'이 2번

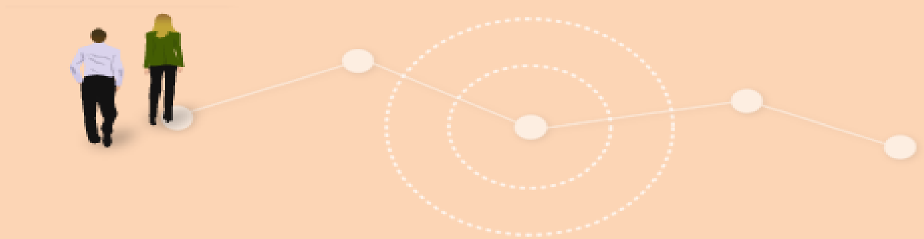
∴ '모'가 연속 4회 : $P(4) = \left(\frac{1}{16}\right)^4$

'걸'이 연속 2회 : $P(2) = \left(\frac{4}{16}\right)^2$



02 확률변수의 평균과 분산

:: Keywords 확률변수의 평균(기대값) | 확률변수의 분산 | 확률변수의 표준편차



확률변수의 기대값

룰렛! 나는 이 경기에서 이길 수 있을까?

카지노에서 접할 수 있는 룰렛은 회전판을 돌려서 자신이 배팅한 숫자나 색깔에 볼이 들어가면 이기는 게임이다. 일단 게임이 시작되면 어떤 경우라도 볼이 숫자에 들어가지 않는 경우는 없다. 반드시 결과는 발생하며 오직 그 결과를 맞히는 사람만이 경기의 승자가 된다.

결국 다른 도박도 마찬가지지만 룰렛도 딜러와 경기자 간의 확률 싸움이다. 통계학을 위한 확률에 입문하는 우리가 카지노에서 이길 확률은 어느 정도일까?[5]

[그림 4-10] 룰렛의 확률

확률변수의 평균

■ 기대값 (expected value)

어떤 사건에 대해 그 사건이 벌어질 확률을 곱해서 전체 사건에 대해 합한 값

확률 = $\sum_{i=1}^{n} P(E_i) = 1$ 이므로, 기대값 $E(X) = \sum x P(x)$

→ 기대값이란

사건에서 발생하는 해당 값과 그 사건이 발생할 확률을 곱해서 모두 더한 값

Ex. 주사위를 던졌을 때의 기대값

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6}$$

$$= \frac{1+2+3+4+5+6}{6} = 3.5$$

확률변수의 기대값

Note 기대값의 성질

a가 상수, X와 Y가 확률변수일 때 다음이 성립한다.

(1) $E(a) = a$

(2) $E(aX) = aE(X)$

(3) $E(aX \pm b) = aE(X) \pm b$

(4) $E(aX \pm bY) = aE(X) \pm bE(Y)$

(5) $E(XY) = E(X)E(Y)$, X와 Y는 확률적으로 독립

확률변수의 분산과 표준편차

■ **확률변수의 분산은 기대값의 특성을 나타내는 값으로, 확률변수들이 기대값으로부터 벗어나는 정도를 나타낸다.**

→ (평균으로부터 산포되어 있는 정도를 분산이라 한 것과 같이,)
확률에서 분산은 기대값과 어느 정도 차이가 있는지를 나타낸다.

$$Var(X) = E(X-\mu)^2 = \sum (X-\mu)^2 P(X)$$

확률변수의 분산과 표준편차

X	x_1	x_2	x_3	x_4	x_5	x_6	합
결과	1	2	3	4	5	6	21

평균(기대값) $m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1+2+3+4+5+6}{6} = 3.5$

X	x_1	x_2	x_3	x_4	x_5	x_6	합
$P(x_i)$[6]	$P(x_1)$	$P(x_2)$	$P(x_3)$	$P(x_4)$	$P(x_5)$	$P(x_6)$	1

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \times (x_i\text{의 상대도수})$$

$$= \sum_{i=1}^{n}(x_i - m)^2 \times \frac{x_i\text{의 상대도수}}{n}$$

$$= \sum_{i=1}^{n}(x_i - m)^2 P(x_i)$$

확률변수의 기대값

Note 분산의 성질

a와 b가 상수, X와 Y가 확률변수일 때 다음이 성립한다.

(1) $Var(a) = 0$

(2) $Var(aX) = a^2 Var(X)$

(3) $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

$Var(X+Y) = Var(X) + Var(Y)$, X와 Y는 확률적으로 독립

(4) $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X, Y)$

$Var(X-Y) = Var(X) + Var(Y)$, X와 Y는 확률적으로 독립

확률변수의 표준편차

표준편차는 분산(σ^2)의 제곱근이므로

$$\sigma = \sqrt{E(X-\mu)^2} = \sqrt{\sum (X-\mu)^2 P(X)}$$

→ 표준편차를 확인하는 이유는

분산이 측정치와 평균의 차의 제곱을 모두 더한 값이라 평균과 상당한 차이가 나기 때문

확률변수의 표준편차

예제 4-4 확률변수의 평균과 분산을 이용한 예

1. 확률변수 X의 평균이 3이고 분산이 5일 때, 확률변수 X^2의 평균을 구하라.
2. $E(X)=4$일 때, $E(3X-1)$, $E(-2X+1)$을 구하라.
3. $V(X)=3$일 때, $V(2X-1)$을 구하라.

풀이

1. $V(X)=E(X^2)-[E(X)]^2$에 대입하면, $5=E(X^2)-3^2$, $\therefore E(X^2)=$ 14
2. $E(3X-1)=3E(X)-1=11$, $E(-2X+1)=-2E(X)+1=$ -7
3. $V(2X-1)=2^2\,V(X)=12$

확률변수의 기대값과 분산

예제 4-5 실사례에서 확률변수의 평균과 분산 활용하기

준비파일 | 4장_기대값과 분산.xlsx

영국의 프리미어리그에서 우승을 노리는 N팀이 이번에 FA가 되는 선수 A와 B를 스카우트하려고 한다. A와 B는 경기 스타일과 포지션이 겹치지 않아 모두 영입하면 좋겠지만, 팀 사정상 선수 한 명만 영입할 수 있다. 성적에 따른 인센티브와 각 선수를 영입했을 때 얻을 수 있는 성적에 대한 확률을 기반으로 평균과 확률변수의 분산을 계산하고 누구를 영입할 것인지 판단하라.

[표 4-3] 선수 A와 선수 B 영입 후 예상 확률

성적별 인센티브 \ 선수		선수 A 성적 확률	선수 B 성적 확률
우승	300	0.58	0.65
순위 상승	150	0.87	0.51
성적 동일	0	0.55	0.45
성적 하락	−100	0.05	0.05

확률변수의 기대값과 분산 (예제 풀이)

$$\text{A의 기댓값} = 300 \times 0.58 + 150 \times 0.87 + 0 \times 0.55 + (-100) \times 0.05 = 299.5$$

$$\text{B의 기대값} = 300 \times 0.65 + 150 \times 0.51 + 0 \times 0.45 + (-100) \times 0.05 = 266.5$$

$$\begin{aligned}\text{A의 분산} &= (300 - 299.5)^2 \times 0.58 + (150 - 299.5)^2 \times 0.87 \\ &\quad + (0 - 299.5)^2 \times 0.55 + [(-100) - 299.5)^2] \times 0.05 \\ &= 76,760.0125\end{aligned}$$

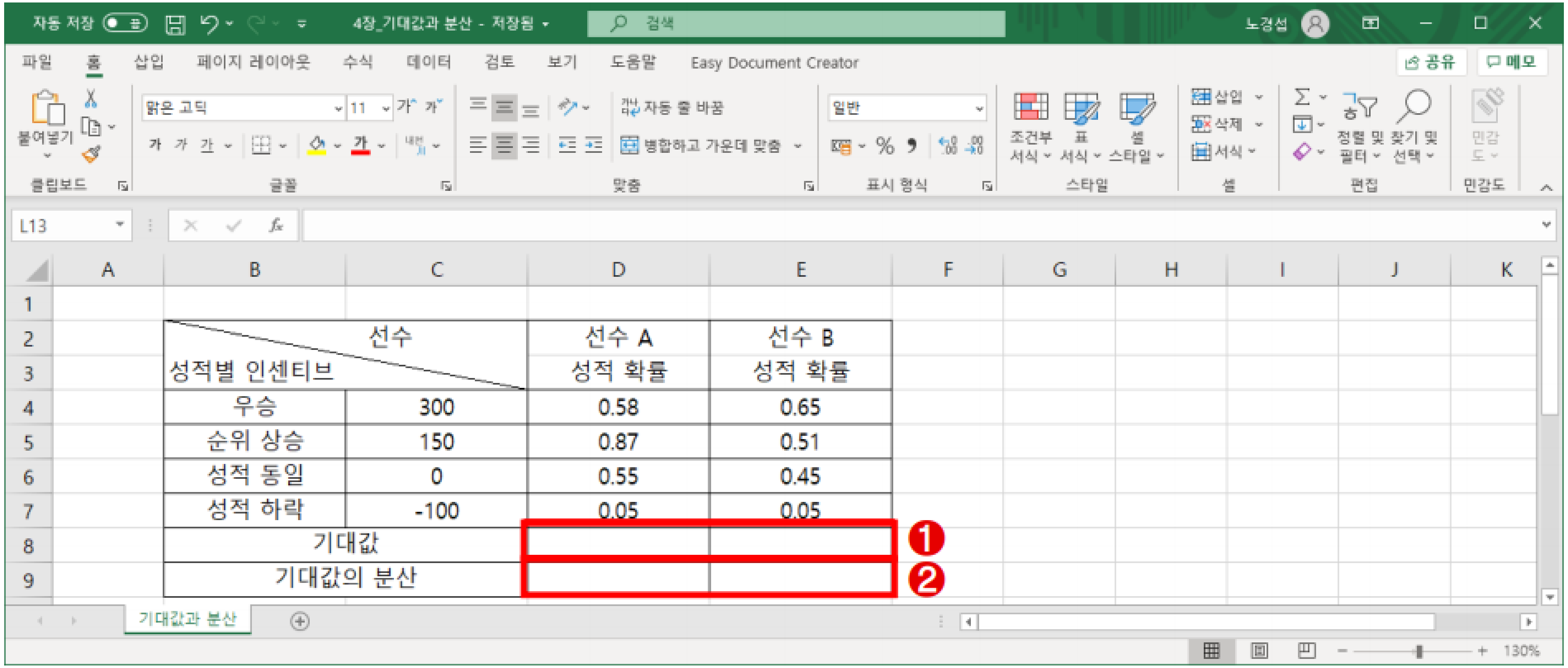
$$\begin{aligned}\text{B의 분산} &= (300 - 266.5)^2 \times 0.65 + (150 - 266.5)^2 \times 0.51 \\ &\quad + (0 - 266.5)^2 \times 0.45 + [(-100) - 266.5)^2] \times 0.05 \\ &= 46,327.435\end{aligned}$$

결론

기대값과 분산이 서로 다른 결과가 나왔는데, 우승 확률은 조금 떨어지더라도 기대값이 높은 선수 A를 선택할 것인지, 혹은 기대값이 적더라도 조금이라도 안정적이면서 우승 확률이 높은 선수 B를 선택할 것인지를 결정한다.

→ N팀은 우승을 노리는 팀이라고 했다. 그러므로 A보다는 안정적인 분산을 가지고 있으면서도 우승의 성공 확률이 높은 B를 선택하게 될 것이다.

확률변수의 기대값과 분산 (예제 Excel 풀이)

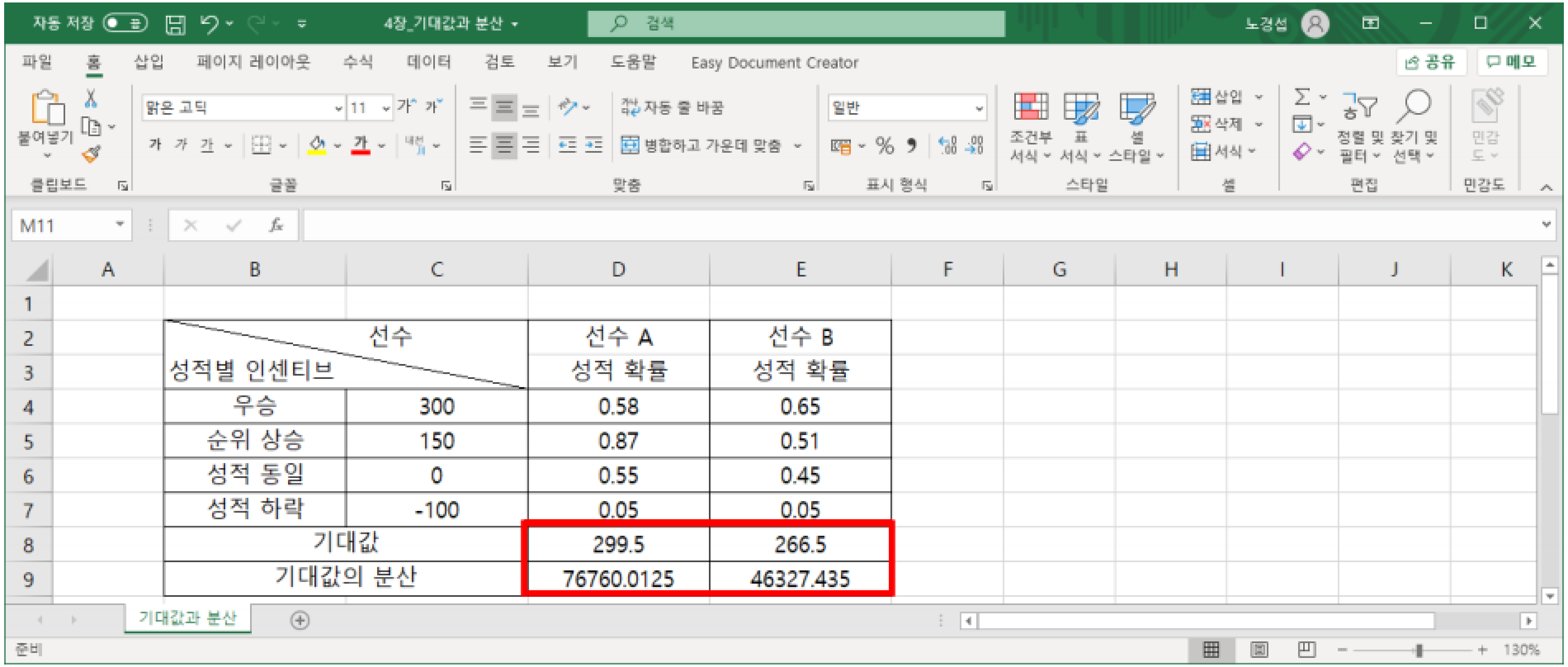


	A	B	C	D	E
1					
2		선수		선수 A	선수 B
3		성적별 인센티브		성적 확률	성적 확률
4		우승	300	0.58	0.65
5		순위 상승	150	0.87	0.51
6		성적 동일	0	0.55	0.45
7		성적 하락	-100	0.05	0.05
8		기대값			
9		기대값의 분산			

D8셀 : 기대값을 구하는 계산식'=$C$4*D4+$C$5*D5+$C$6* D6+$C$7*D7'
E8셀 : 기대값을 구하는 계산식'=$C$4*E4+$C$5*E5+$C$6*E6+$C$7*E7'

분산 D9셀 : '=($C$4-D8)^2*D4+($C$5-D8)^2* D5+($C$6-D8)^2*D6+($C$7-D8)^2*D7'
분산 E9셀 : '=($C$4-E8)^2*E4+($C$5-E8)^2* E5+($C$6-E8)^2*E6+($C$7-E8)^2*E7'

확률변수의 기대값과 분산 (예제 Excel 풀이 완성)



선수 / 성적별 인센티브		선수 A 성적 확률	선수 B 성적 확률
우승	300	0.58	0.65
순위 상승	150	0.87	0.51
성적 동일	0	0.55	0.45
성적 하락	-100	0.05	0.05
기대값		299.5	266.5
기대값의 분산		76760.0125	46327.435

Q&A

통계학, 제대로 시작하자!

통계의 쓰임을 이해하고,
실제로 활용할 수 있어야 한다.