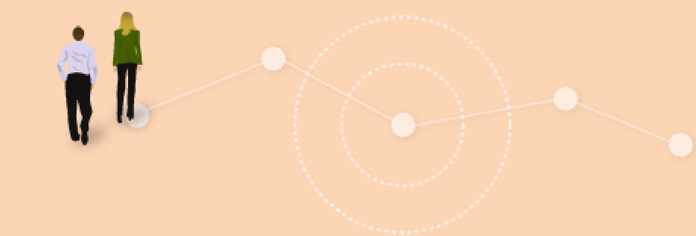
4 장

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

확률

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

→ $n(A) = 3$

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

확률

■ 확률이 가지는 조건

❶ 확률은 0 ~ 1의 값을 가진다.

❷ 모든 사건에 대한 확률의 합은 1이다.

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

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

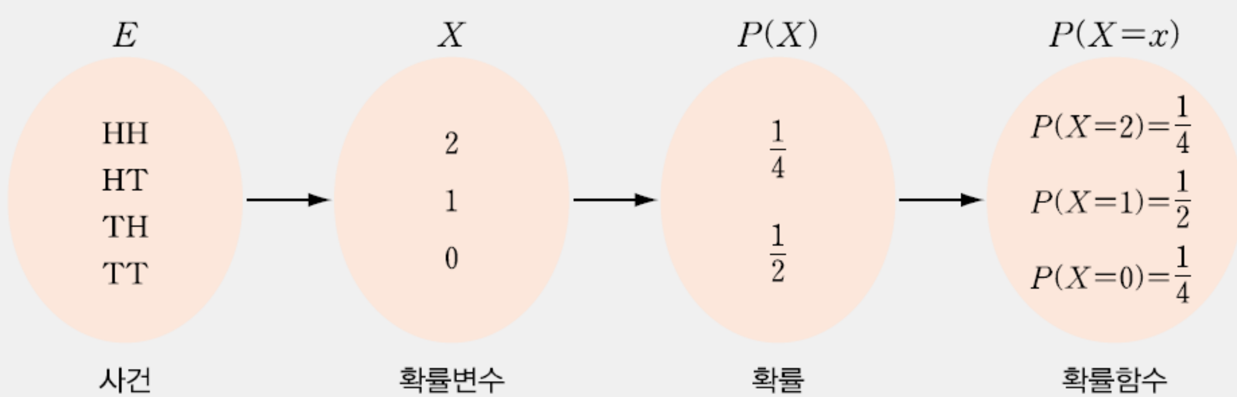
- **이산 확률변수**(discrete random variable) : 수집된 데이터의 확률변수 중 셀 수 있는 특정한 값들로 구성되거나 일정한 범위로 나타나는 확률변수
- **이산 확률변수**(discrete random variable) : 연속형이거나 무한한 경우와 같이 셀 수 없는 확률변수

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

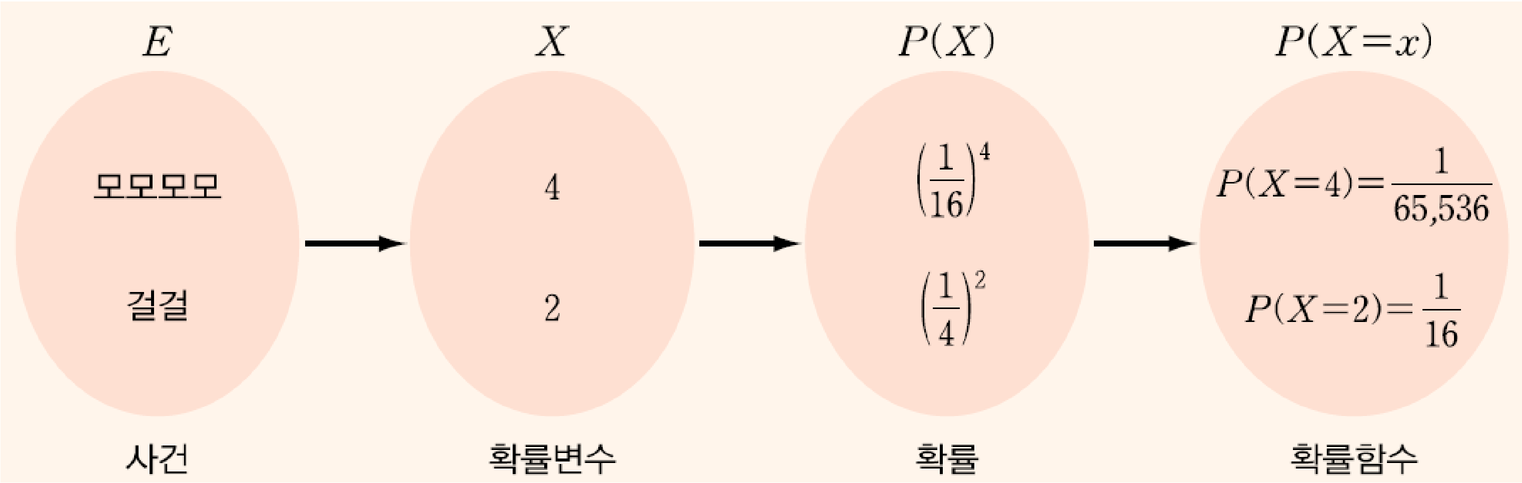
말 4개를 가지고 윷놀이를 할 때, 말을 최소한으로 움직이면서 가장 빨리 이기는 경우의 확률변수에 대해 설명하라. 또한 이에 대한 확률과 확률함수를 구하라.

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

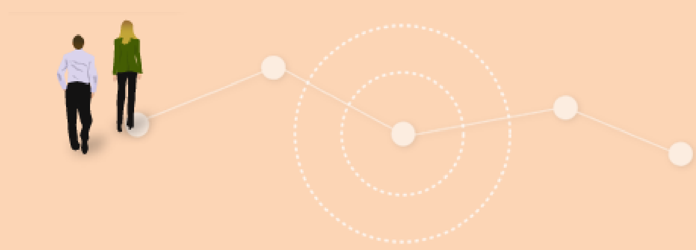
∴ '모'가 연속 4회 : $P(4)=\left(\frac{1}{16}\right)^4$

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



02 확률변수의 기대값과 분산

:: Keywords 확률변수의 기대값 | 확률변수의 분산 | 확률변수의 표준편차



확률변수의 기대값

■ 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 분산

- **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 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 기대값과 분산

예제 4-3 확률변수의 기대값과 분산

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, 자금의 한계로 인해 단 한 명의 선수만을 영입할 수 있다. 성적에 따른 인센티브와 각 선수들을 영입했을 때 얻을 수 있는 성적에 대한 확률을 기반으로 기대값과 확률변수의 분산을 계산하고 누구를 영입할 것인지 판단하라.

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)

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

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

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

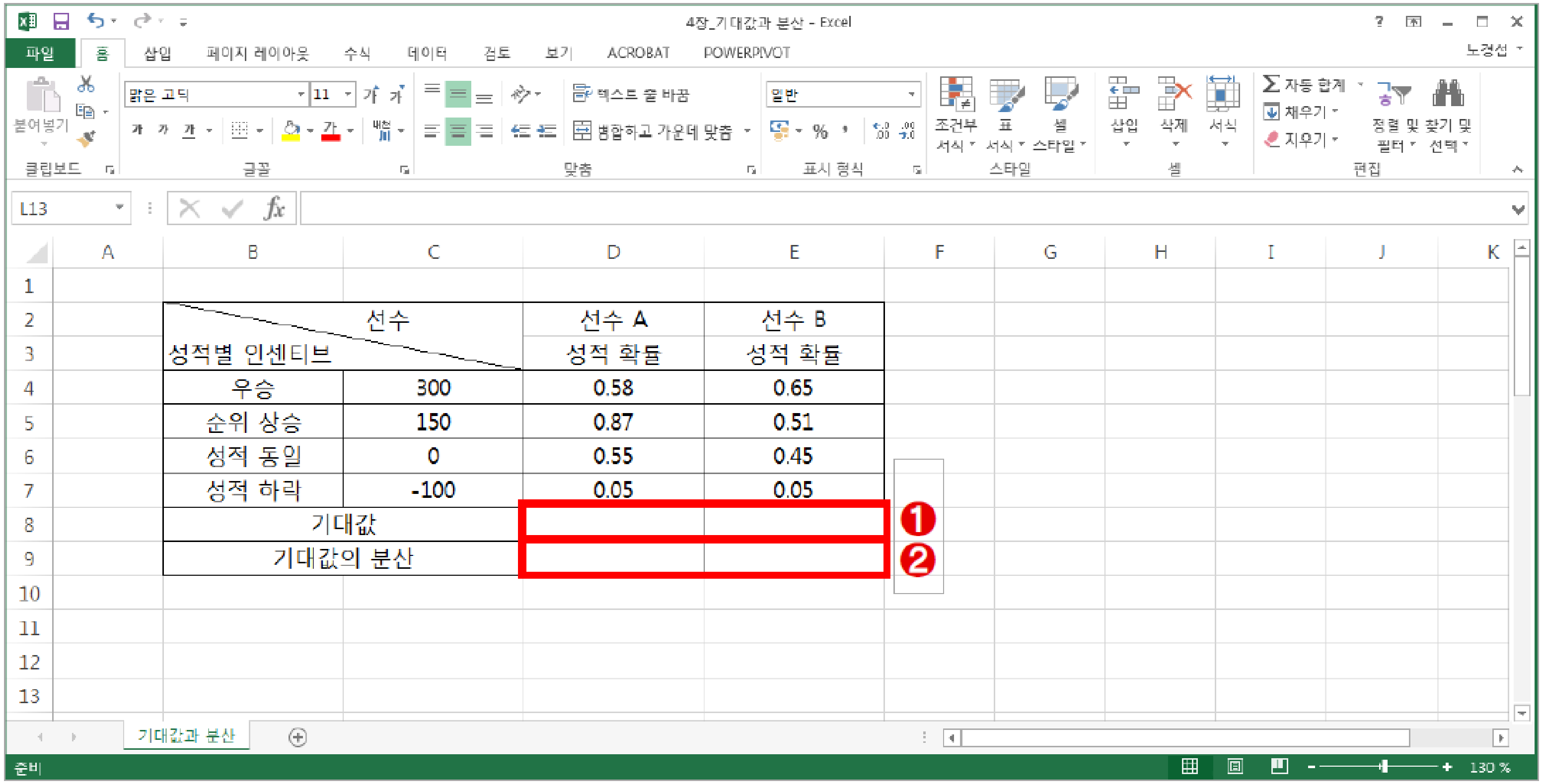
$$\begin{aligned}\text{B의 분산} &= (300 - 266.5)^2 \times 0.65 + (150 - 266.5)^2 \times 0.51 \\ &\quad + (0 - 266.5)^2 \times 0.45 + [(-100) - 266.5)^2] \times 0.05 \\ &= 46,327.435\end{aligned}$$

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)



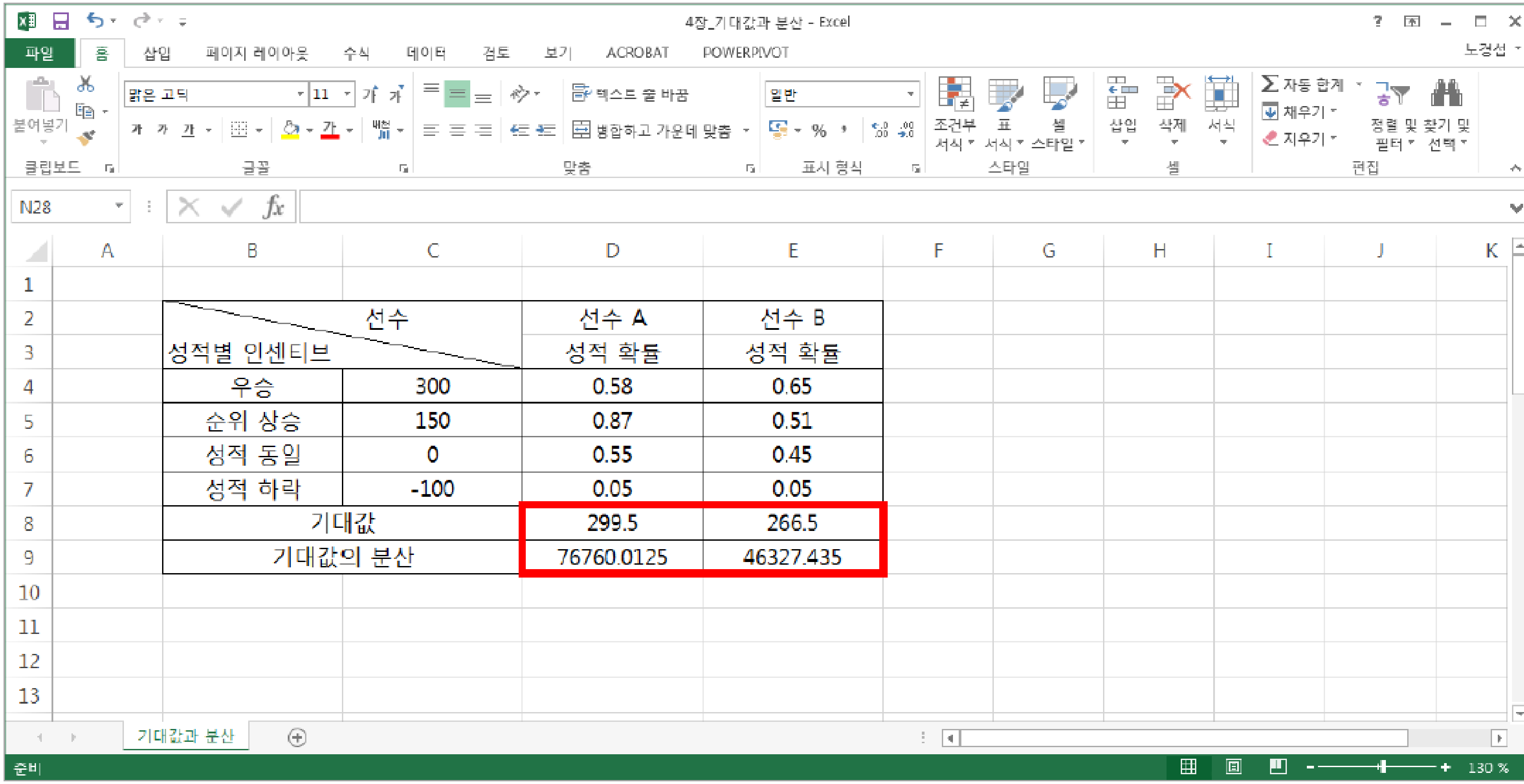
선수 / 성적별 인센티브		선수 A 성적 확률	선수 B 성적 확률
우승	300	0.58	0.65
순위 상승	150	0.87	0.51
성적 동일	0	0.55	0.45
성적 하락	-100	0.05	0.05
기대값			
기대값의 분산			

❶
❷

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