확률, 확률변수, 기대치

CE730 금융통계

조남운

Namun Cho/ mailto:namun@snu.ac.kr

주제

- 확률 모형과 확률 법칙
- 확률변수
- 기대값
- 경우의 수

불확실성
Uncertainty

- Karl Pearson의 동전던지기 실험
  - 24,000번 던짐
  - 앞면이 12,012번 나옴
- 동전의 앞면이 나올 상대빈도수: 12012/24000=0.5005
- 1/2에 가까운 값
- 우리가 직접 해본다면?
  - 12012번 나오지는 않겠지만 12000번과 크게 다르지는 않은 값일 것임

불확실성
Uncertainty

- Karl Pearson의 동전던지기 실험
  - 24,000번 던짐
  - 앞면이 12,012번 나옴
- 동전의 앞면이 나올 상대빈도수: 12012/24000=0.5005
- 1/2에 가까운 값
- 우리가 직접 해본다면?
  - 12012번 나오지는 않겠지만 12000번과 크게 다르지는 않은 값일 것임

확률
Probability

- 불확실성을 객관적으로 다루기 위한 개념
- 철학적으로 명확한 합의가 있는 개념은 아님
  - 심지어 객관성 마저도!
- 여기에서는 빈도론적 확률 개념을 채용
  - 측정을 다수 반복할 때 수렴하는 상대빈도
  - 예: 수만번의 동전던지기 실험 (상대빈도는 0.5 근처에서 형성되고 시행수가 많아질수록 0.5에 가까워짐)

다시, 불확실성

- 우리는 동전 던지기의 결과 앞면이 나올 확률이 1/2 라는 것을 알고 (혹은 믿고) 있음
- H를 앞면 (Head)이 나오는 사건이라고 지칭해보자
  - 뒷면은 T (Tail)
- 하지만 지금 던질 동전이 앞면이 나올 것인지 뒷면이 나올 것인지는 모름
  - 하지만 앞면이나 뒷면이 나올 사건이 비슷한 빈도수로 나올 것이라는 것은 추측할 수 있음
  - $P(H) = 1/2$

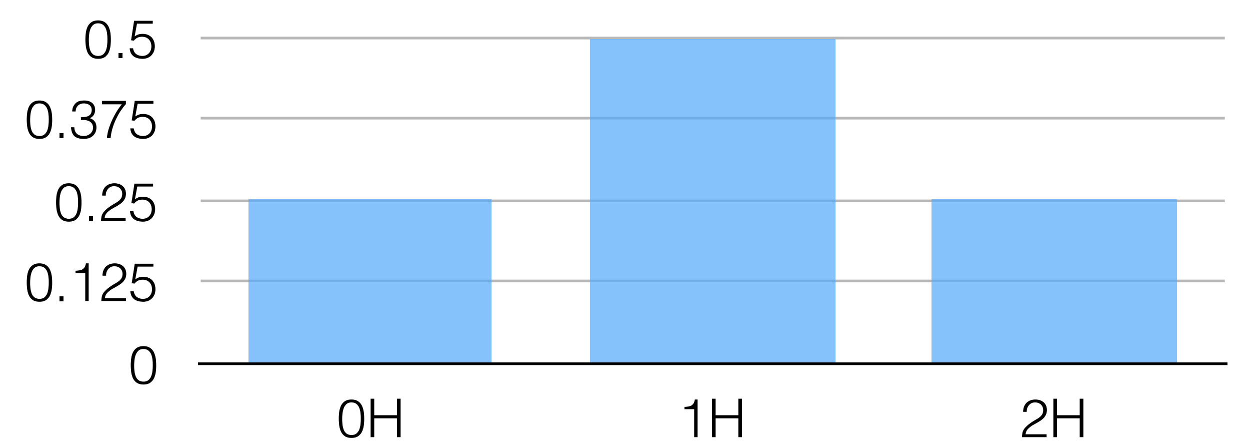
다시, 불확실성

- 우리는 동전 던지기의 결과 앞면이 나올 확률이 1/2 라는 것을 알고 (혹은 믿고) 있음
- H를 앞면 (Head)이 나오는 사건이라고 지칭해보자
  - 뒷면은 T (Tail)
- 하지만 지금 던질 동전이 앞면이 나올 것인지 뒷면이 나올 것인지는 모름
  - 하지만 앞면이나 뒷면이 나올 사건이 비슷한 빈도수로 나올 것이라는 것은 추측할 수 있음
  - P(H) = 1/2

두 번 던진다면?

- 순서에는 관심이 없이 오로지 H가 몇 번 나오는지에만 관심이 있다면 경우의 수는 총 3가지임
    - H가 2번 나올 경우
    - H가 1번 나올 경우
    - H가 0번 나올 경우
- 각각의 경우의 수는
    - 1/4, 1/2, 1/4

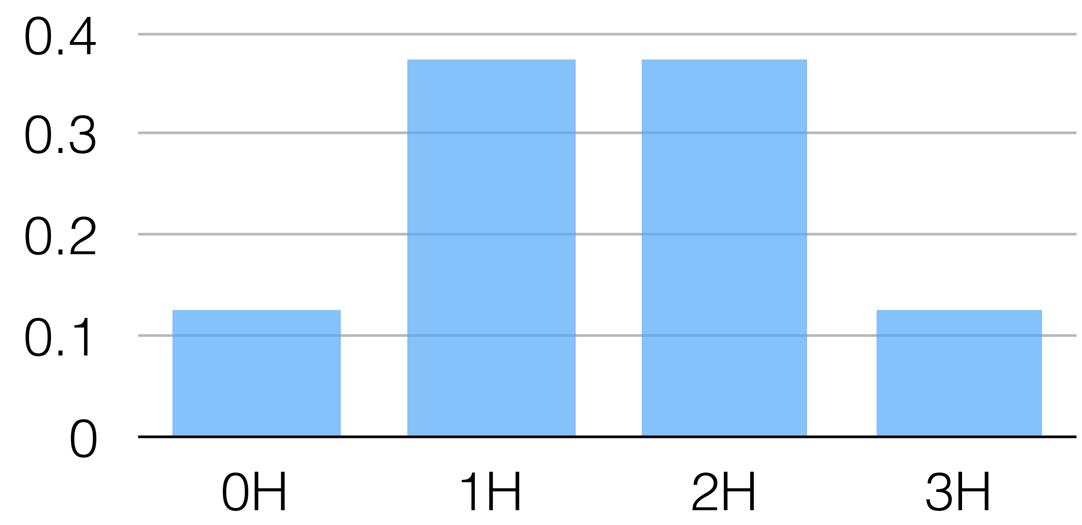
Events	확률
HH	1/4
HT	1/4
TH	1/4
TT	1/4



세 번 던진다면?

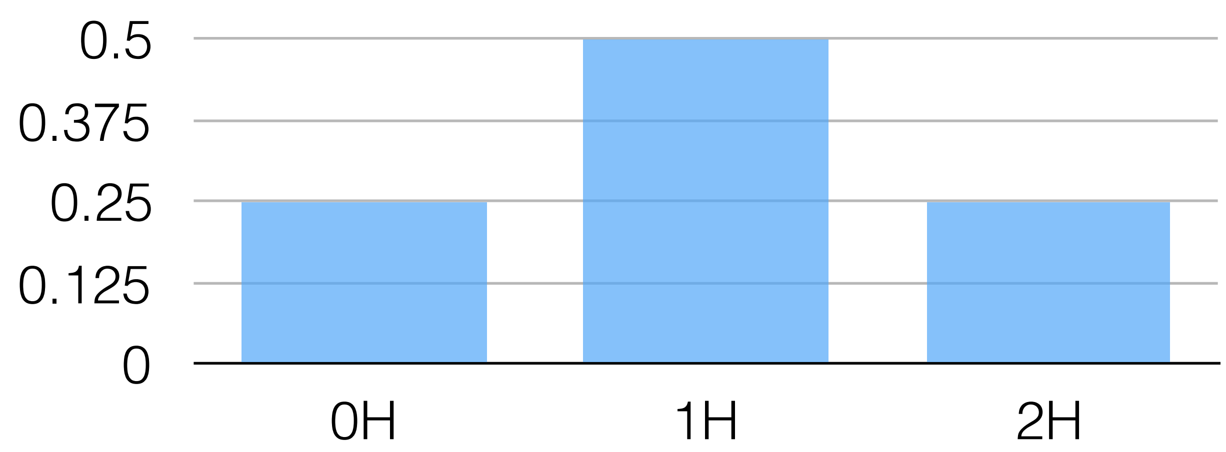
Events	확률
0H	0.125
1H	0.375
2H	0.375
3H	0.125

- n번 던진다면 어떻게 될까?

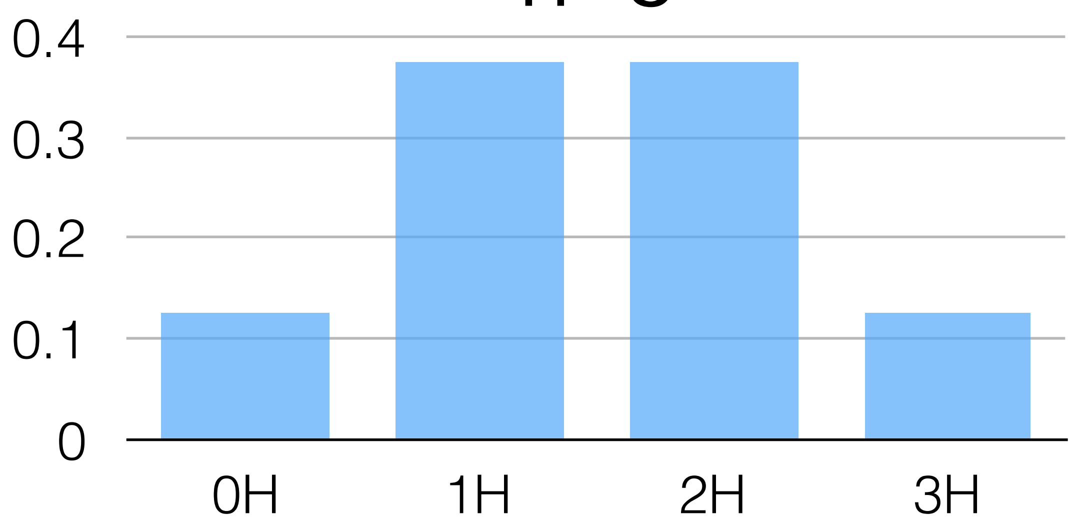


10,000번 동전 던지기

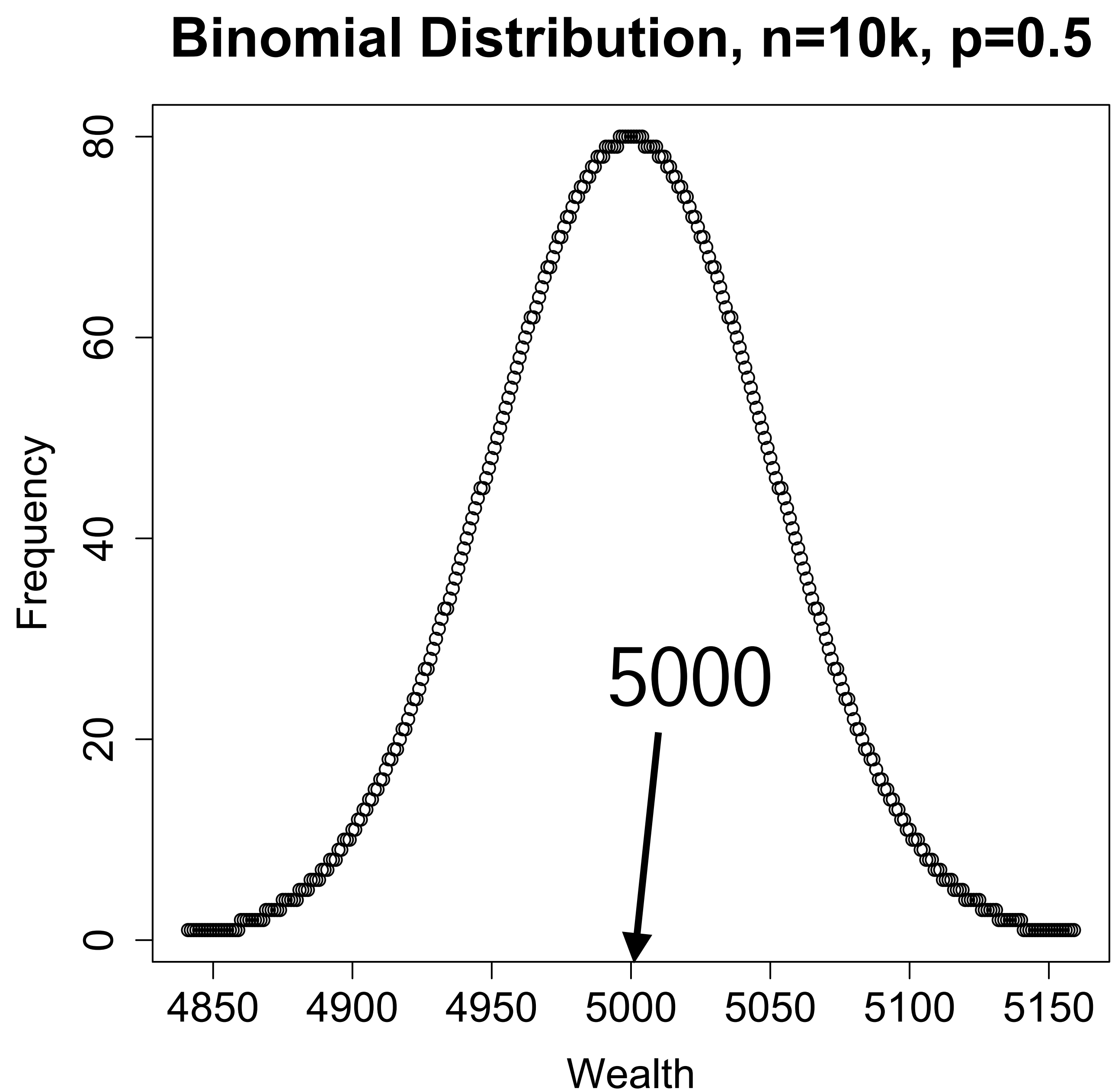
n=2



n=3



# of Head



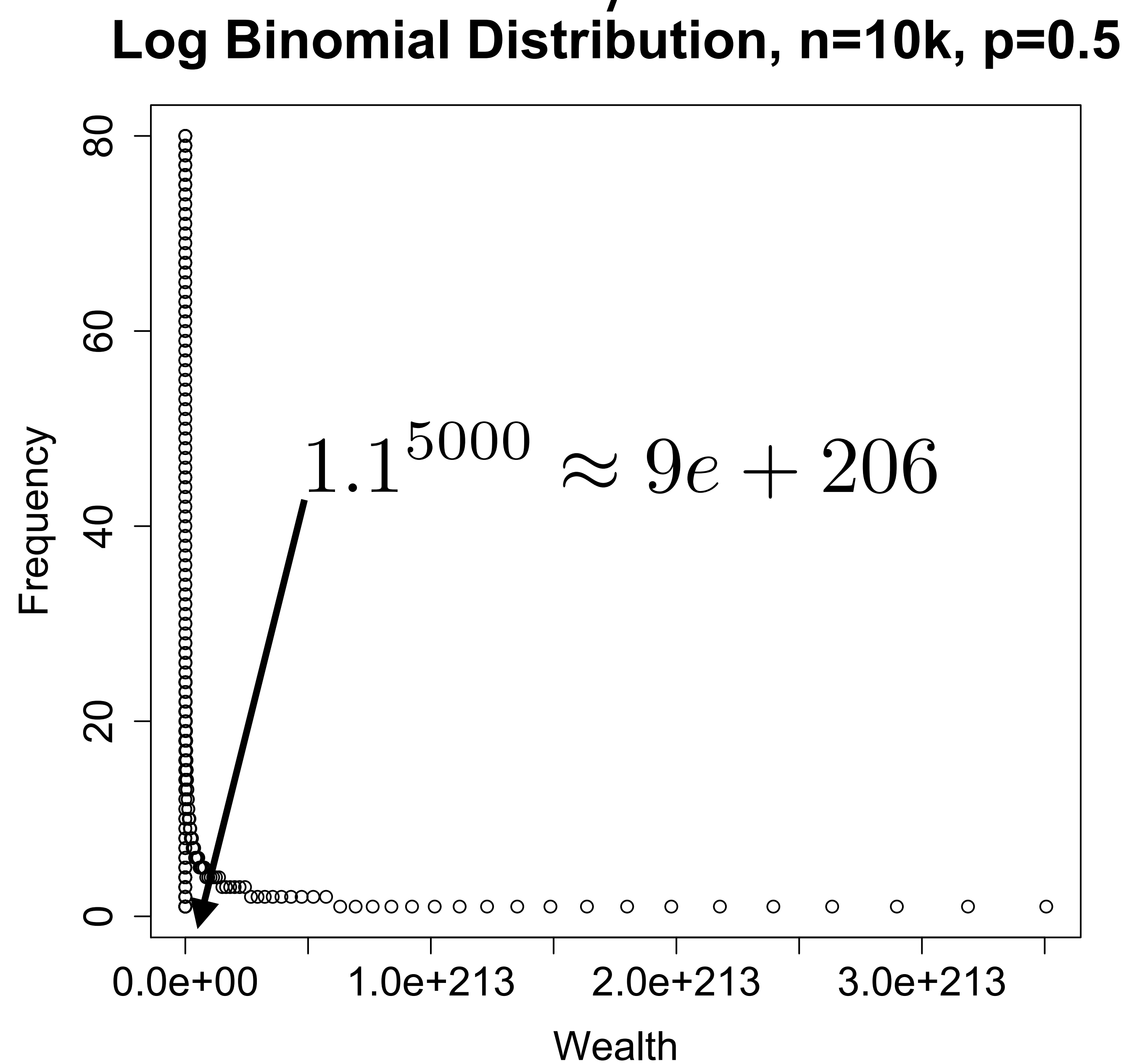
n=10,000

Namun Cho/ mailto:namun@snu.ac.kr

동전던지기: 지수증가의 경우

- 10,000명에게 초기 자금을 동일하게 부여
- 만일 앞면이 나올 때마다 1 점씩 부과한다면 분포는 앞에서와 동일
- 하지만 앞면이 나올 때마다 그 사람의 자금에 10%씩을 부과한다면
- 매우 불평등한 부의 분배가 나타남
- 현대의 부의 분포와 유사한 패턴

Payoff



확률 모형과 확률 법칙

Namun Cho/ mailto:namun@snu.ac.kr

확률 모형: 구성 요소 Probability Model

- 표본공간 Sample Space
  - 어떤 실험에서 일어날 수 있는 모든 경우의 집합
  - 경우의 수가 셀 수 있는가 (countable)에 따라 이산형, 연속형으로 분류
- 확률
  - 각 모든 경우에 대한 확률
- $P(A) \in [0,1]$: A가 일어날 확률

이산형 표본공간: 동전던지기 실험 (1회)

- 실험내용:
  - 동전을 한 번 던지는 사건
- 표본공간 S

$$S = \{H, T\}$$

- 각 경우의 확률
  - P(H) = p
  - P(T) = 1-p

$$p \in [0,1]$$

이산형 표본공간: 동전던지기 실험 (1회)

- 실험내용:
  - 동전을 한 번 던지는 사건
- 표본공간 S

$$S = \{H, T\}$$

- 각 경우의 확률
  - P(H) = p
  - P(T) = 1-p

$$p \in [0,1]$$

이산형 표본공간: 동전던지기 실험 (2회)

- 실험내용:
  - 동전을 두 번 던지는 사건
- 표본공간 S

$$S = \{HH, HT, TH, TT\}$$

- 각 경우의 확률
-

$$P(HH) = p_1, p(HT) = p_2, p(TH) = p_3, p(TT) = p_4$$

$$p_i > 0, \quad \sum_i p_i = 1, \quad i = 1,2,3,4$$

이산형 표본공간: 동전던지기 실험 (2회)

- 실험내용:
  - 동전을 두 번 던지는 사건
- 표본공간 S

$$S = \{HH, HT, TH, TT\}$$

- 각 경우의 확률

- $$P(HH) = p_1, p(HT) = p_2, p(TH) = p_3, p(TT) = p_4$$

$$p_i > 0, \quad \sum_i p_i = 1, \quad i = 1,2,3,4$$

연속형 표본공간: 한국 20대 남성의 키

$$S = \{x \,|\, 0.5m \le x \le 2.5m\}$$

$$P(x_i) = 0, \quad \forall x_i \in S$$

- 연속형 표본공간에서 단일 수치의 확률은 0
- 이 경우 확률은 구간에서만 의미가 있음

$$P(0.5m \le x \le 2.5m) = 1$$

https://m.blog.naver.com/ferozen2su/140141511231

예제 4.1

- A, B사가 합병논의중
  - 합병되면 두 기업 주식가치는 높아지는 상황 (우측표)
  - 현재 이에 대한 기대로 주가가 상승한 상황
- 합병성사확률이 아래와 같을 때 의사결정
  - Case1: 50%
  - Case2: 80%

	합병전	현재가	합병후
A사	5000	5400	5500
B사	2000	3000	4000

풀이

- Case2 (80%)
  - 합병될 경우 (80%)
    - A: 5400 ⟶ 5500
    - B: 3000 ⟶ 4000
  - 합병되지 않을 경우 (20%)
    - A: 5400 ⟶ 5000
    - B: 3000 ⟶ 2000
- A사의 손익 기대값
  - 100*0.8 + (-400)*0.2=0
  - 무차별: 현재 가격은 기대값을 정확히 반영하고 있음 ⇒ 사도 되고 안사도 됨
- B사의 손익 기대값, Case 1도 직접 해볼 것

	합병전	현재가	합병후
A사	5000	5400	5500
B사	2000	3000	4000

주관적 확률
Belief

- 현실에서 이러한 합병 가능성에 대한 확률 평가는 투자자마다 다를 것임
  - 이러한 "주관적" 확률을 믿음(belief)이라고 함
  - 예제4.1에서는 합병확률을 50%로 믿는 투자자들에게는 A사 주식이 고평가 되어 있다고 느낄 것
    - A사 주식을 판매하려 할 것
    - 이러한 행동은 A사 주식의 가격을 낮추는 방향의 영향을 줌
  - 반면 80%로 보는 투자자들에게는 B사의 주식이 저평가되어 있다고 느낄 것
    - B사 주식을 구매하려고 할 것
    - 이러한 행동은 B사의 주식의 가격을 높이는 방향의 영향을 줌

시장의 정보처리기능

- 결국 시장은 이러한 주관적 확률을 모으는 기능을 수행
- 믿음의 가중평균
  - 가중치: 자금의 양
    - 더 많은 자금을 거래에 사용하는 참가자의 확률은 더 큰 영향을 미칠 것이기 때문

시장 정보처리 기능의 효율성

- 한편, 불확실한 사건의 결론이 나게 될 경우 (예제 4.1), 즉 합병이 되었다면
  - 합병확률을 더 정확히 평가한 사람은 더 높은 수익을 얻게 됨 ⇒ 자금의 양이 많아짐 ⇒ 가중치가 높아짐
- 이러한 반복 거래에서 더 성공적인 예측을 지속적으로 거둔 사람의 비중은 시장에서 강해지게 됨 ⇒ 시장의 가중평균 예측은 더 정확한 방향으로 조정됨

임의, 사건
Random, Event

- 임의 Random
  - 결과는 불확실하지만 그 불확실성을 설명할 수 있는 법칙이 존재하는 실험
- 사건 Event
  - 표본공간 (sample space)의 부분집합

사건간의 관계

- 상호배반사건 Mutually Exclusive Event
  - 서로 동시에 일어날 수 없는 사건들
  - 예: H와 T, HH와 HT 등
- 독립사건 Independent Event
  - 각 사건의 확률이 서로 다른 사건의 확률과 아무런 관련이 없는 사건
  - 예: 동전1의 동전던지기결과 A 와 동전2의 동전던지기결과 B

확률법칙

- A, B ∈ S: Event
- S: Sample Space
- A∩B=∅ implies A and B are disjoint events
- 확률법칙 (LP1)

$$0 \le P(A) \le 1$$

- 확률법칙 (LP2)

$$A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B)$$

- 확률법칙 (LP3)

$$P(S) = 1$$

확률관련 표현들

- 비조건부 확률 Unconditional Probability
  - 사건 A가 일어날 확률: $P(A)$
- 조건부 확률 Conditional Probability
  - 사건 B가 발생한다는 조건 하에서 사건 A가 발생할 확률: $P(A|B)$
- 합 확률 Union Probability
  - 사건 A 혹은 사건 B가 일어날 확률: $P(A \cup B)$
- 결합확률 Joint Probability
  - 사건 A와 B가 동시에 일어날 확률: $P(A \cap B)$

확률 법칙의 확장

- 집합 연산의 정의와 확률법칙 (LP1-3)을 이용하여 확률법칙을 확장할 수 있음

$P(A^{\complement}) = 1 - P(A)$

Proof.

$$S = A \cup A^{\complement} \qquad \text{(Definition of Complement)}$$

$$A \cap A^{\complement} = \emptyset \qquad \text{(Mutually Exclusive)}$$

$$P(A \cup A^{\complement}) = P(A) + P(A^{\complement}) \qquad \text{(LP2)}$$

$$\therefore P(A^{\complement}) = 1 - P(A)$$

$\square$

$P(A \setminus B) = P(A \cap B^{\complement}) = P(A) - P(A \cap B)$

Proof.

$$A = (A \cap B) \cup (A \cap B^{\complement})$$

$$(A \cap B) \cap (A \cap B^{\complement}) = \emptyset \qquad \text{(Mutually Exclusive)}$$

$$P(A) = P(A \cap B) + P(A \cap B^{\complement}) \qquad \text{(LP2)}$$

$$\therefore P(A \setminus B) = P(A \cap B^{\complement}) = P(A) - P(A \cap B)$$

$\square$

$$B \subset A \quad \Rightarrow \quad P(A \setminus B) = P(A) - P(B)$$

General Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof.

$$
\begin{aligned}
&A \cup B = B \cup (A \cap B^{\complement}) \\
&B \cap (A \cap B^{\complement}) = \emptyset && \text{(Mutually Exclusive)} \\
&P(B \cup (A \cap B^{\complement})) = P(B) + P(A \cap B^{\complement}) && \text{(LP2)} \\
&\therefore P(A \cup B) = P(A) + P(B) - P(A \cap B) \\
&\qquad (\because P(A \setminus B) = P(A \cap B^{\complement}) = P(A) - P(A \cap B))
\end{aligned}
$$

$\square$

예제 4.2

- 생명보험사의 판단문제
  - 보험금 지급을 위해
- Event W:
  - 부인 사망
- Event H:
  - 남편 사망
- Event W and H
  - 동시 사망 (사망이 서로 무관할 경우를 전제)
- Q: 최소 1명이 사망할 확률
  - W or H

	Probability
W	0.02
H	0.007
W and H	0.0007
W or H	?

확률의 합법칙

확률의 곱법칙 Generalized Multiplication Rule

- 예제 4.3
- 경기후퇴가 발생할 확률 (event R)
  - 18%
- 경기후퇴(R)로 인해 장기채권 수익이 떨어질(D) 확률
  - P(R|D)= 83%
- Q: P(R and D) = ?

$$P(RD) = P(R)P(D|R)$$

$$P(A \cap B) = P(AB)$$

$$= P(A)P(B|A) = P(B)P(A|B)$$

조건부확률
Conditional Probability

- Event B가 일어난 상황을 전제로 했을 때 A가 발생할 확률
- Event B가 주어진 상황에서 A의 확률
- 만일 A와 B가 상호배제적일 때는 0
  - 동시에 일어날 수 없으므로
- A와 B가 독립 (서로 무관)일 때에는 P(A|B) = P(A)
  - P(A and B) = P(A)P(B) 이므로

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

독립사건의 성질

- 아래 사건이 독립일 때
  - A and B
  - B and C
  - C and A
- A,B,C가 서로 독립이 아닐 수 있음
  - 반례로 증명:
    - S={a,b,c,d}, A={a,b}, B={a,d}, C={b,d}
    - P(a)=P(b)=P(c)=P(d)=1/4

독립 사건의 확률

- “독립사건” A1 … An 이 동시에 일어날 확률은
  - 각 사건의 확률의 곱과 같음

$$P(A_1 \cap A_2 \cap \cdots A_n) = \prod_{i=1}^{n} P(A_i)$$

임의실험과 독립성 Random Experiment and Independence

- 임의 실험이 서로 독립적으로 반복시행할 경우 각 시행으로 만들어지는 표본 공간은 서로 독립
  - 서로의 원소가 완전히 다름을 의미
  - 따라서 서로 다른 표본공간에 속한 사건들 역시 서로 독립임

동전 던지기 n회

- 동전 던지기 n회
  - 각 회차의 동전던지기에 의한 표본공간은 서로 독립: S1, S2, … Sn
  - S1 = S2 = … = Sn = {H,T}
    - 같은 집합인데 독립? YES!
- 가령 동전던지기 2회 (n=2)의 표본공간 S는
  - S = S1 × S2 = {HH, HT, TH, TT}
  - 가령 TT에서 T∈ S1, T∈ S2 로 엄밀히는 서로 다른 의미임

조건부확률의 확장

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$$

$$(X = A \cap B)$$

- 이 정리는 임의의 사건 A,B,C에 대해서 성립함

$$\begin{aligned} P(A \cap B \cap C) &= P(X \cap C) \\ &= P(C|X)P(X) \\ &= P(C|A \cap B)P(A)P(B|A) \end{aligned}$$

예제 4.4

- 기업 수익 추정 문제
  - EPS는 이벤트 이름을 각 가격으로 부여
- 매출호조(Event G): 80%
  - 매출호조일 경우 EPS 3000 가능성: 90%
    - Event 3000|G
  - 매출호조일 경우 EPS 2000 가능성: 10%
- 매출저조(Event B): 20%
  - 매출저조일 경우 EPS 2000 가능성: 40%
  - 매출저조일 경우 EPS 1000 가능성: 60%
- P(1000), P(2000), P(3000) = ?

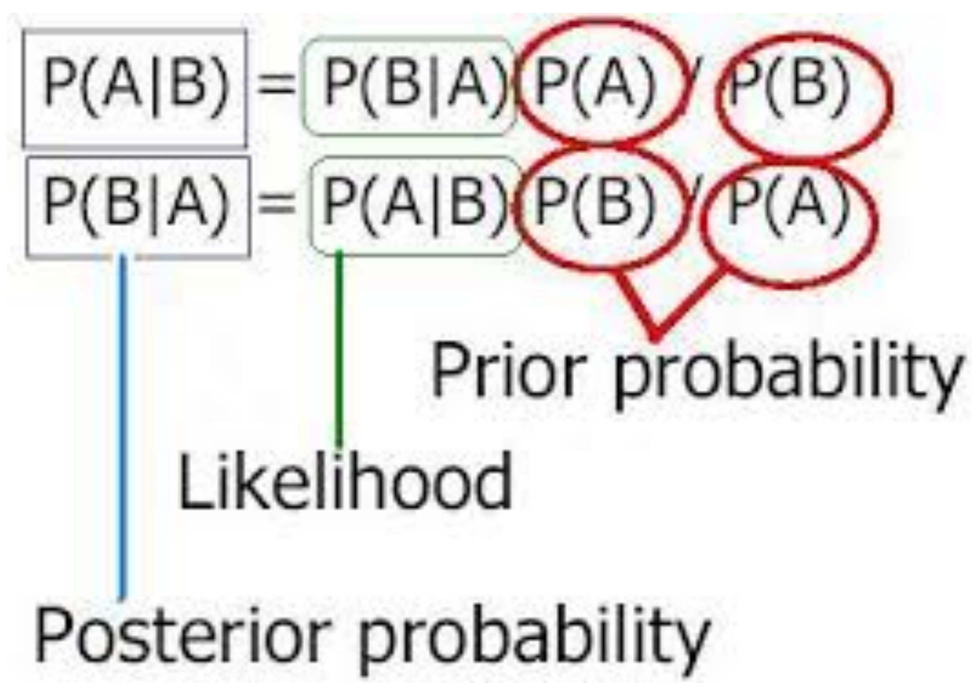
Solution

- P(G and 1000) = P(1000|G)P(G)
  - P(1000|G) = 0 이므로 P(G and 1000)=0
- P(B and 1000) = P(1000|B)P(B)
- P(G and 2000) = P(2000|G)P(G)
- …
- P(B and 3000) = P(3000|B)P(B)

베이즈 정리
Bayes' Theorem

- Prior Probability
  - 현재 조건에서 알 수 있는 확률
  - 참값으로부터 편차(오차) 존재
- Posterior Probability
  - 추가적인 정보를 통해 보정한 더 나은 추측 확률
- 베이즈정리: 확률의 개선을 이루는 역할을 수행

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$



Namun Cho/ mailto:namun@snu.ac.kr

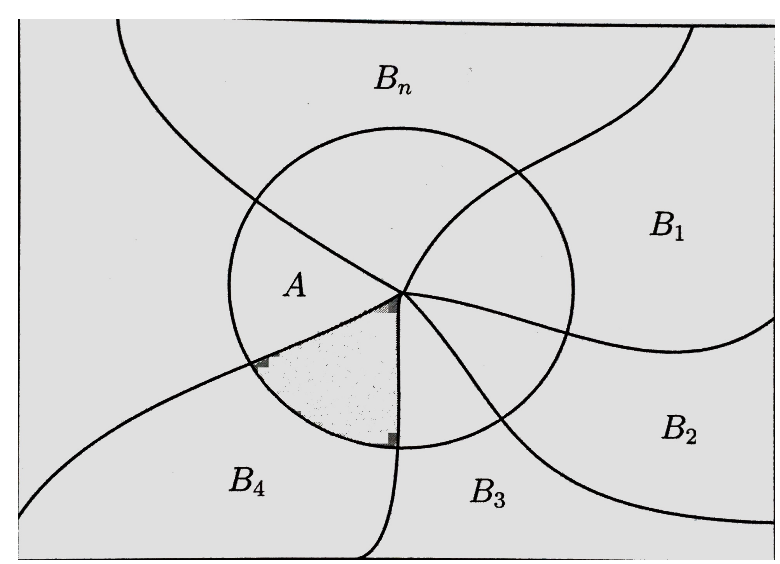
일반화

$$P(B_k | A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(B_k)P(A | B_k)}{\sum_{i=1}^{n} P(B_i)P(A | B_i)}$$

- 사전확률: P(Bk), P(A)
- 알고 있는 정보
  - P(Bi), ∀ i (사전정보)
  - P(A|Bi) ∀ i
    - 실험으로 알게 되는 정보
- 실험으로 알고자 하는 정보
  - P(Bk|A)

Partition

$$S = B_1 \cup B_2 \cup \cdots \cup B_n$$
$$B_i \cap B_j = \emptyset \quad \forall i \neq j$$



예제 4.5

- 서랍속 약을 먹었는데 [배가 아픔]
- 서랍속 약은 90%가 [아스피린], 10%가 [독약]
- 의사가 알고있는 사실
  - [독약]을 먹었을 때 [배가 아플] 확률 99%
  - [아스피린] 먹었을 때 [배가 아플] 확률 2%
- 이 사람이 독약을 먹었을 확률 = ?
  - P(독약|배가아픔) = ?

Solution

- P(독약|배가아픔) =
  - 분자: P(배가아픔|독약)P(독약)
  - 분모: P(배가아픔) = P(배가아픔|독약)P(독약) + P(배가아픔|아스피린)P(아스피린)
- 환자는 의사에게 서랍속 약의 구성 비율을 알려줌으로써 (베이즈 정리에 의해) 더 나은 독약 먹었을 확률을 구할 수 있게 됨

Ex 4.6

- 경제는 성장기조[EU]이거나 침체기조[ED]
  - 통계적으로 P(EU)=3/4, P(ED)=1/4
  - 통계 집계후 판단하므로 당장 알 수 없음
- 통계적으로 아래와 같은 사실이 알려져 있음
  - EU일때 주가상승(MU)일 확률 80%
  - ED일때 주가하락(MD)할 확률 70%
  - 주가는 즉각 관찰됨
- Q: 주가상승이 관찰됨 ⇒ 성장기조일 확률은?

Solution

- 알고자 하는 것: $P(EU|MU)$
- 분자: $P(EU)P(MU|EU)$
- 분모: $P(MU)=P(EU)P(MU|EU)+P(ED)P(MU|ED)$

Ex 4.7

- 기업이 대출 받고 1년 이내 [부도]날 확률 2%
- [부도]난 기업 중 95%가 신용평가기관에 의해 [부도위험]기업으로 분류되어 있음
- [부도비발생]기업 중 3%는 신용평가기관에 의해 [부도위험]기업으로 잘못 분류되어 있음
- Q: [부도위험]기업으로 분류되어 있는 기업이 실제로 부도날 확률은?
  - 풀이 전에 먼저 직관적인 답의 대략적인 범위를 생각해볼 것

Solution

- 알고자 하는 것: P(부도발생|부도위험)
  - P(부도발생) = 0.02
  - P(부도비발생) = 1-P(부도)
  - P(부도위험|부도발생) = 0.95
  - P(무위험|부도발생) = 0.05
  - P(부도위험|부도비발생) = 0.03
  - P(무위험|부도비발생) = 0.97
- 분자: P(부도발생*부도위험) = P(부도발생)P(부도위험|부도발생)
- 분모: P(부도위험) = P(부도발생)P(부도위험|부도발생)+P(부도비발생)P(부도위험|부도비발생)
- 0.3925

확률변수
Random Variable

확률변수의 의미

- 엄밀히 보자면, 확률"변수"라는 말은 부정확한 용어
- 확률변수는 "함수"(변수 사이의 관계)임
  - input: 표본공간의 partition
    - 즉, 상호 배제적 사건의 집합 (합집합=표본공간)
  - output: 사건의 집합의 index

$$\sum_x P(X = x) = 1$$

예: 동전던지기2회에서 나온 앞면의 횟수

- S={HH,HT,TH,TT}
- P(HH)=⋯ =P(TT) = 1/4
- 앞면의 횟수를 확률변수 X라고 한다면
- 이산적 확률변수

x	Event	P(X=x)
0	{TT}	1/4
1	{HT,TH}	1/2
2	{HH}	1/4

예: 1m 철봉에서 관찰된 흠집의 위치

- X: 왼쪽 기준으로부터 측정한 흠집 위치
- 통계적으로 다음과 같은 관계가 관찰되었다면:

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} 2x\,dx$$

- P(0≤X≤1)=1
- 연속형 확률변수
- 2x: 확률밀도함수 (pdf:probability density function)

확률변수를 "안다"는 것의 의미

- 일반적 변수 X를 안다는 것은 그것의 값을 아는 것으로 충분
- 확률변수 X를 안다는 것은 X의 모든 input(x) 과 각 input의 확률(P(X=x))을 알아야 함
  - 확률분포함수를 알아야 함
- 불확실한 사건을 다루기 위해 고안된 개념
- 확률변수의 이름이 명확할 때에는 P(X=x)를 축약하여 P(x)로 표기하기도 함

임의실험의 2회 반복시행

- 둘 중 하나의 결과{H,T}가 발생하는 임의 실험
  - P(H) = p
- X를 H가 나오는 횟수로 규정
- X의 이산분포함수는 오른쪽과 같이 계산할 수 있음
- 모집단: 무한히 많은 X들(2회반복실험)의 집합
- 표본: 이들 중 일부
  - X1, X2, … , Xn
  - 임의표본(Random Sample) 혹은 i.i.d. (identically independent distribution)

x	P(x)
0	$(1-p)^2$
1	$2p(1-p)$
2	p^2

Namun Cho/ mailto:namun@snu.ac.kr

n=3 case

- “동전2번던지기” 실험을 3번 수행하여 얻은 표본
- 하첨자는 서로 다른 3개 실험의 구분을 위해 붙임
- 이들은 서로 독립이며 같은 확률 분포를 가지고 있음 (i.i.d.)
- 따라서 세 번의 실험에서 나올 수 있는 모든 경우들에 대한 확률은 각 1,2,3번 실험 결과의 확률의 곱

$$X_1, X_2, X_3$$

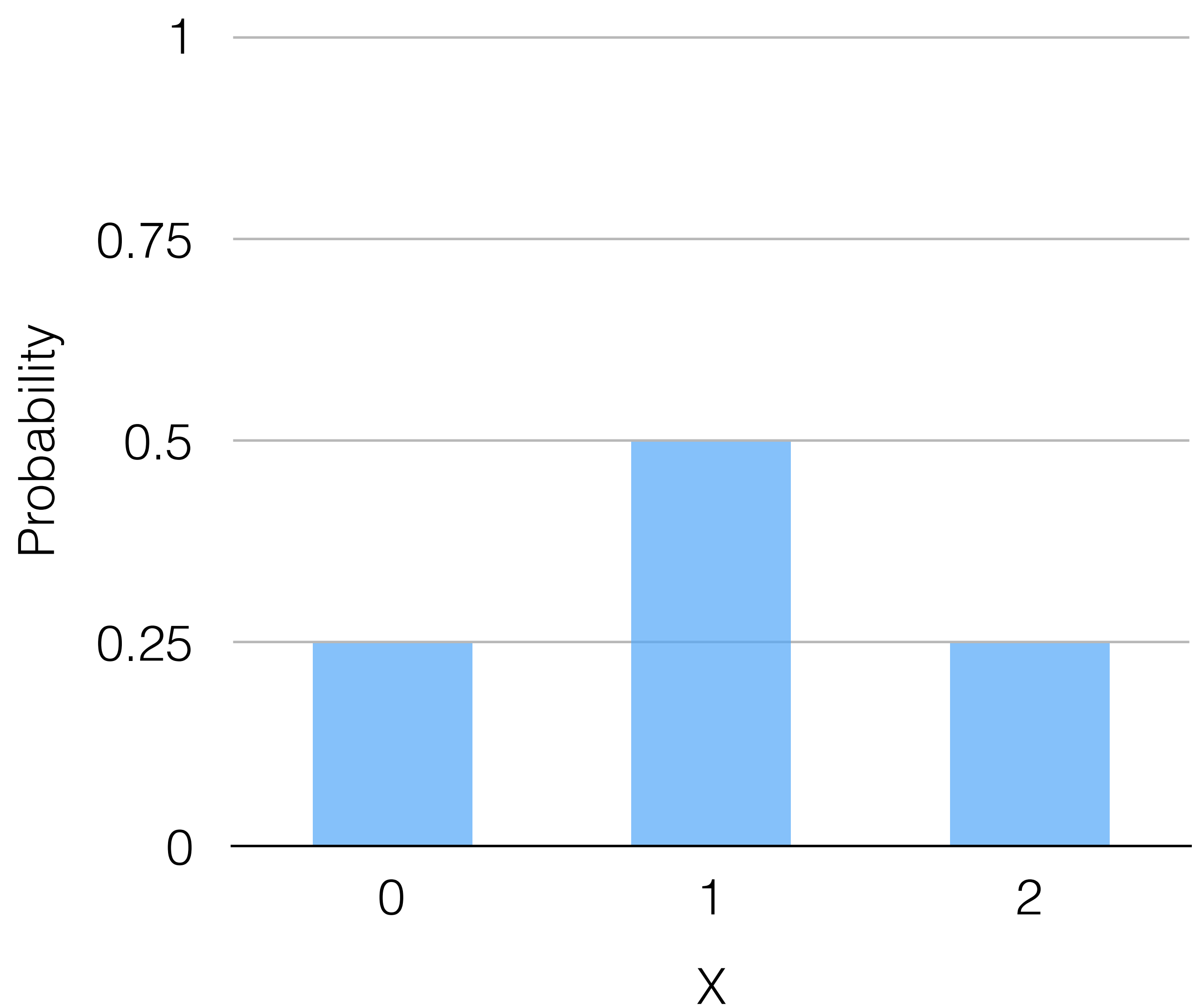
확률변수의 덧셈

- X1의 분포
- X1+X2의 분포
- X1+X2+X3의 분포
- 표본평균의 분포
  - (X1+X2+X3)/3
- 표본평균 자체가 확률변수
  - 표본평균은 통계량

$x_1 + x_2 + x_3$	$P(x_1 + x_2 + x_3)$
0	$(1-p)^6$
1	$6p(1-p)^5$
2	$\ldots$
3	$20p^3(1-p)^3$
4	$\ldots$
5	$\ldots$
6	p^6

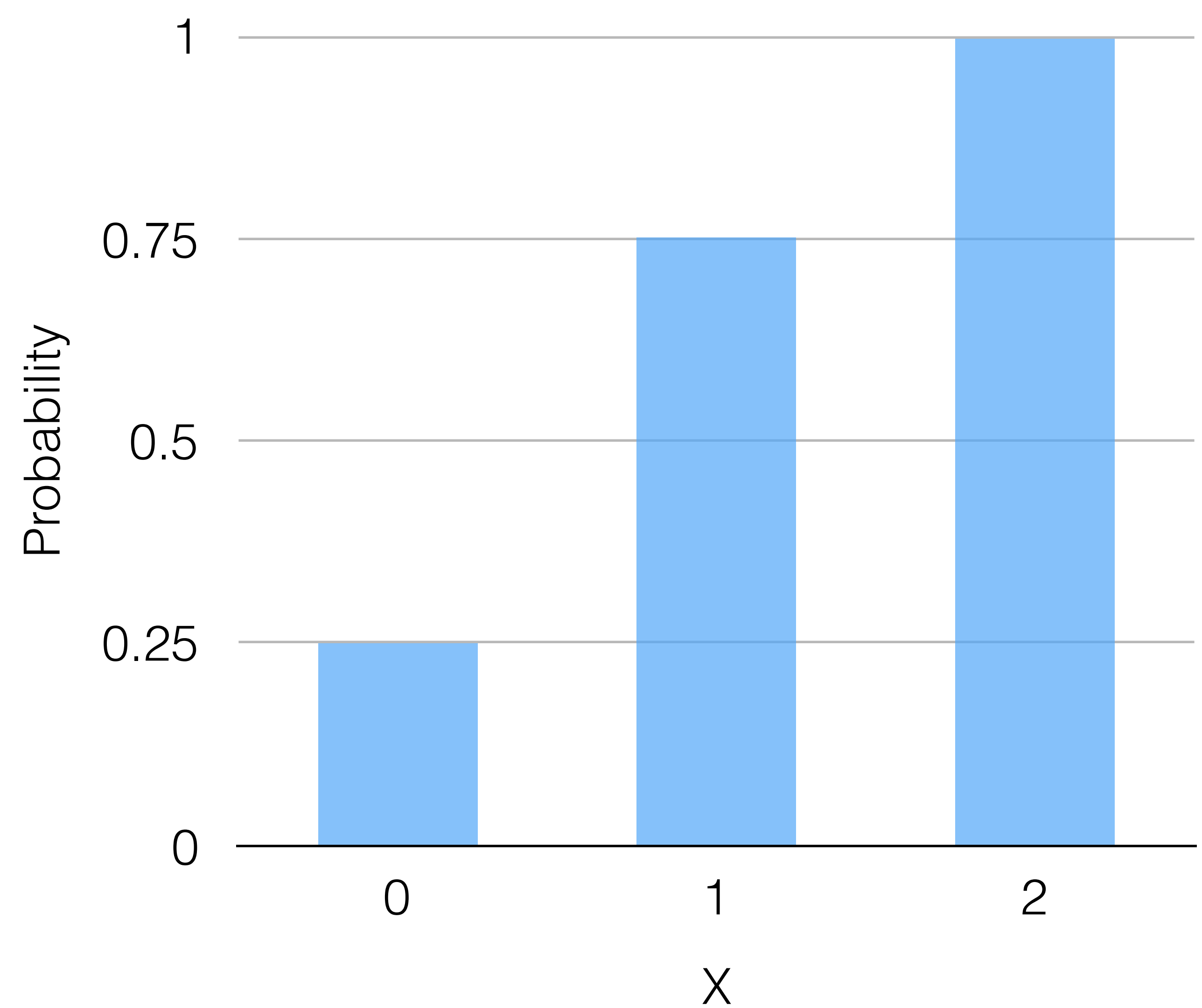
확률질량함수
PMF: Probability Mass Function

- PMF
  - f(x)로 표기
  - P(X=x)
  - p=1/2 일 때



누적분포함수
CDF: Cumulative Distribution Function

- CDF
  - 분포함수라고도 함
  - F(x)로 표기
  - $P(X \leq x)$



확률밀도함수 PDF: Probability Density Function

- 연속형 확률변수에서는 모든 단일값에 대한 확률이 0임
  - P(X=x) = 0
- 따라서 충분히 작은 dx만큼의 구간에 대해 확률을 정의
  - 이렇게 정의한 f(x)를 확률밀도함수라고 명명
- 이때 누적분포함수는 F(x)로 명명

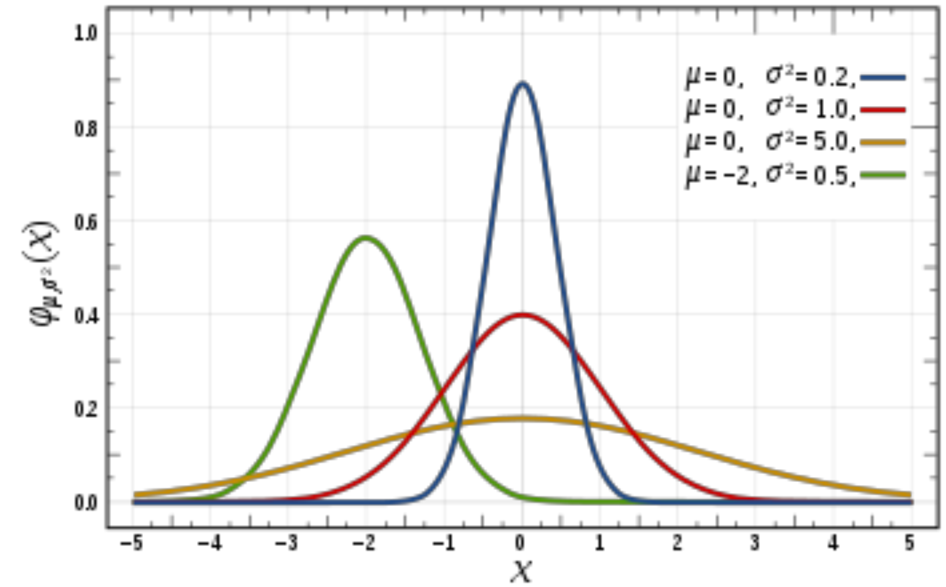
$$P(x < X \leq x + dx) = f(x)dx$$

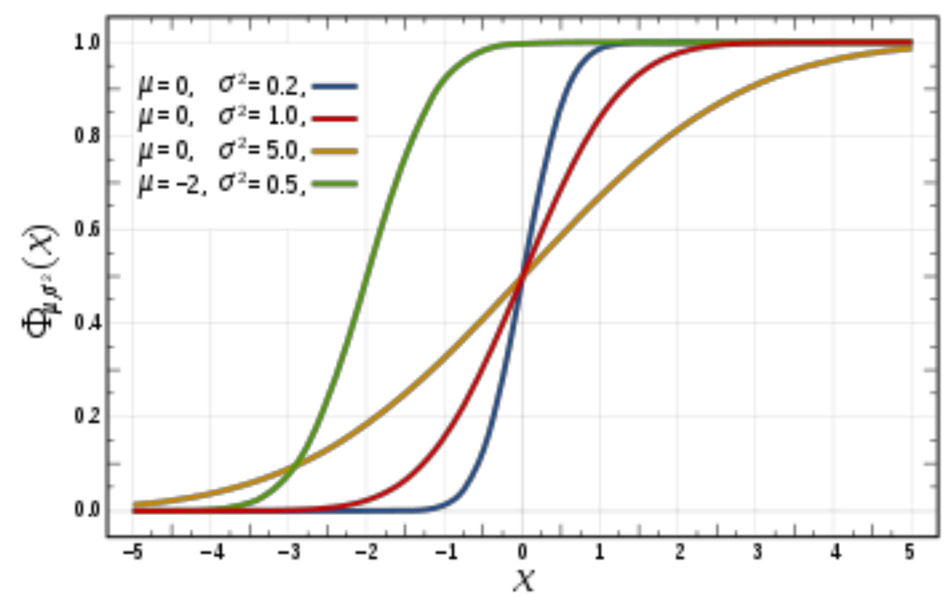
$$F(x) := P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

정규분포의 pdf, cdf

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- μ : X의 평균
- σ : X의 분산
- 가장 중요한 분포
  - 중심극한정리





https://en.wikipedia.org/wiki/Normal_distribution

기대값
Expected Value

Namun Cho/ mailto:namun@snu.ac.kr

기대값의 정의

- 확률변수 X의 중심성을 나타내는 함수
- E(X)∈실수집합
  - input: 확률변수
  - output: “하나의”실수값

$$E(X) := \sum_{i=1}^{n} x_i P(X = x_i)$$

$$E(X) := \int_{-\infty}^{\infty} x f(x) dx$$

기대값의 성질

- 기대값의 정의로부터 도출 가능함
- a,b,c: 상수 (constants)
- X,Y: 확률변수
- g(·): 양함수 (explicit function)

E1

$$E(c) = c$$

E2

$$E(g(X)) = \begin{cases} \sum_{i=1}^{n} g(x_i)P(X = x_i) \\ \int_{-\infty}^{\infty} g(x)f(x)dx \end{cases}$$

E3

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

E2: $g(X) = X^2$ 의 예

x	$P(x)$	x^2	$x^2\ P(x)$
-1	1/6	1	0.1667
0	1/4	0	0.0000
1	1/6	1	0.1667
2	1/24	2	0.0833
3	3/8	3	1.1250
		$E(X^2)$	1.5417

x^2	$P(x^2)$
-1	0
0	1/4
1	2/6
2	1/24
3	3/8
$E(X^2)$	1.5417

Namun Cho/ mailto:namun@snu.ac.kr

분산과 공분산 Variance and Covariance

- 확률변수의 변동성 측도
- σ^2 으로 표기하기도 함
- 기대값으로 정의되므로 분산 공분산 역시 실수(real number)임
- 공분산은 두 확률변수 사이의 결합적 변동성 측도

$$Var(X) := E(X - E(X))^2$$
$$= E(X^2) - [E(X)]^2$$

$$Cov(X, Y) := E[(X - E(X))(Y - E(Y))]$$
$$= E(XY) - E(X)E(Y)$$

(공)분산의 성질

- 각 성질은 기대값의 정의와 (공)분산의 정의에 의해 증명 가능
  - E1, E2, E3

V1 $$Var(c) = 0$$

V2 $$Var(cX) = c^2 Var(X)$$

V3 $$Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$$

예제 4.8
포트폴리오의 기대수익률과 분산

기업 Xi	주식구성비 wi	기대수익률 Ri
X (X1)	50%	8%
Y (X2)	10%	12%
Z (X3)	40%	16%

	(공)분산
XX	0.2
YY	0.3
ZZ	0.25
XY	0.15
YZ	0.10
ZX	0.12

Namun Cho/ mailto:namun@snu.ac.kr

기대수익률 E(Rp)

기업 Xi	주식구성비 wi	기대수익률 Ri	wi*Ri
X (X1)	50%	8%	4%
Y (X2)	10%	12%	1.2%
Z (X3)	40%	16%	6.4%
		E(Rp)	11.6%

$$R_p := \sum_i^n w_i R_i$$

$$E(Rp) = E\left(\sum_i^n w_i R_i\right)$$

$$= \sum_i^n w_i E(R_i)$$

(by E3)

분산 Var(Rp)

기업 Xi	주식구성비 wi
X (X1)	50%
Y (X2)	10%
Z (X3)	40%

$$R_p := \sum_i^n w_i R_i$$

	(공)분산	wi*wj	wi*wj*Cov(Ri,Rj)
XX	0.2	25%	0.05
YY	0.3	1%	0.003
ZZ	0.25	16%	0.04
XY	0.15	5%	0.015
YZ	0.10	4%	0.008
ZX	0.12	20%	0.048
		Var(Rp	0.164

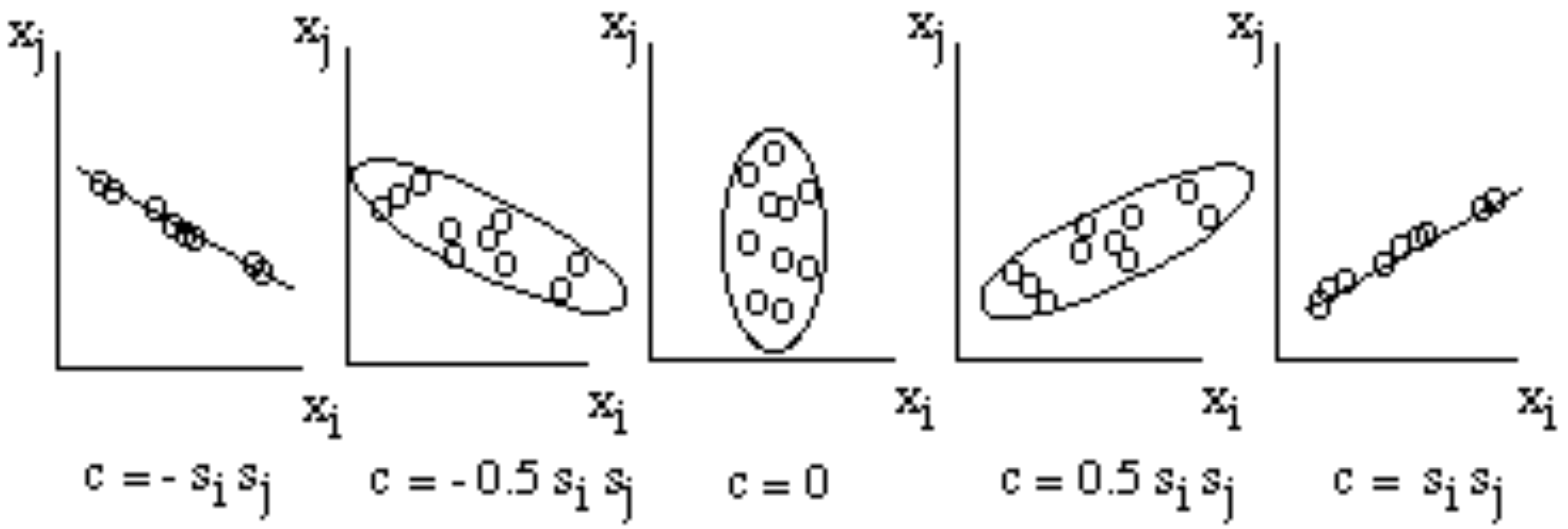
$$Var(R_p) = Var\left(\sum_i^n w_i R_i\right)$$

$$= \sum_i^n w_i^2 Var(R_i) + 2\sum_{i<j} w_i w_j Cov(R_i, R_j)$$

공분산에는 2가 곱해짐에 유의

Namun Cho/ mailto:namun@snu.ac.kr

공분산의 의미



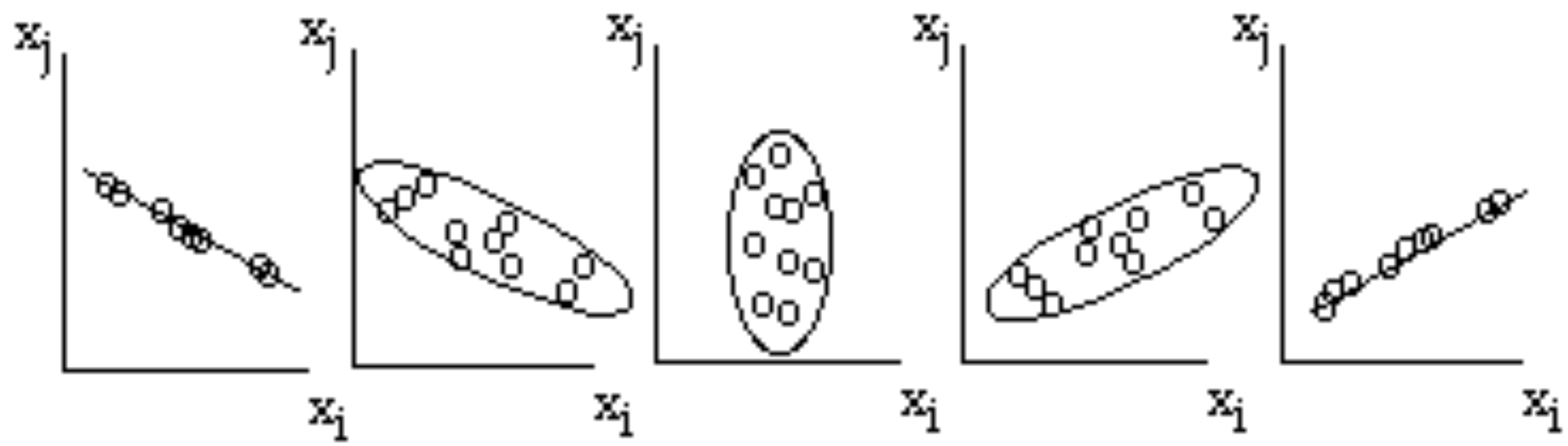
https://www.cs.princeton.edu/courses/archive/fall08/cos436/Duda/PR_Mahal/cov.htm

- 양의 공분산: 정의 상관관계
- 음의 공분산: 음의 상관관계
- Scale 문제 존재
  - 각 변수의 표준편차의 크기에 영향을 받음 ⇒ 상관만 보기 위해서는 Scale 효과를 제거할 필요가 있음 ⇒ 상관계수 (correlation coefficient)

상관계수
Correlation Coefficient

$$\rho_{XY} := \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

$$= \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$



$\rho = -1$ $\rho = 0$ $\rho = 1$

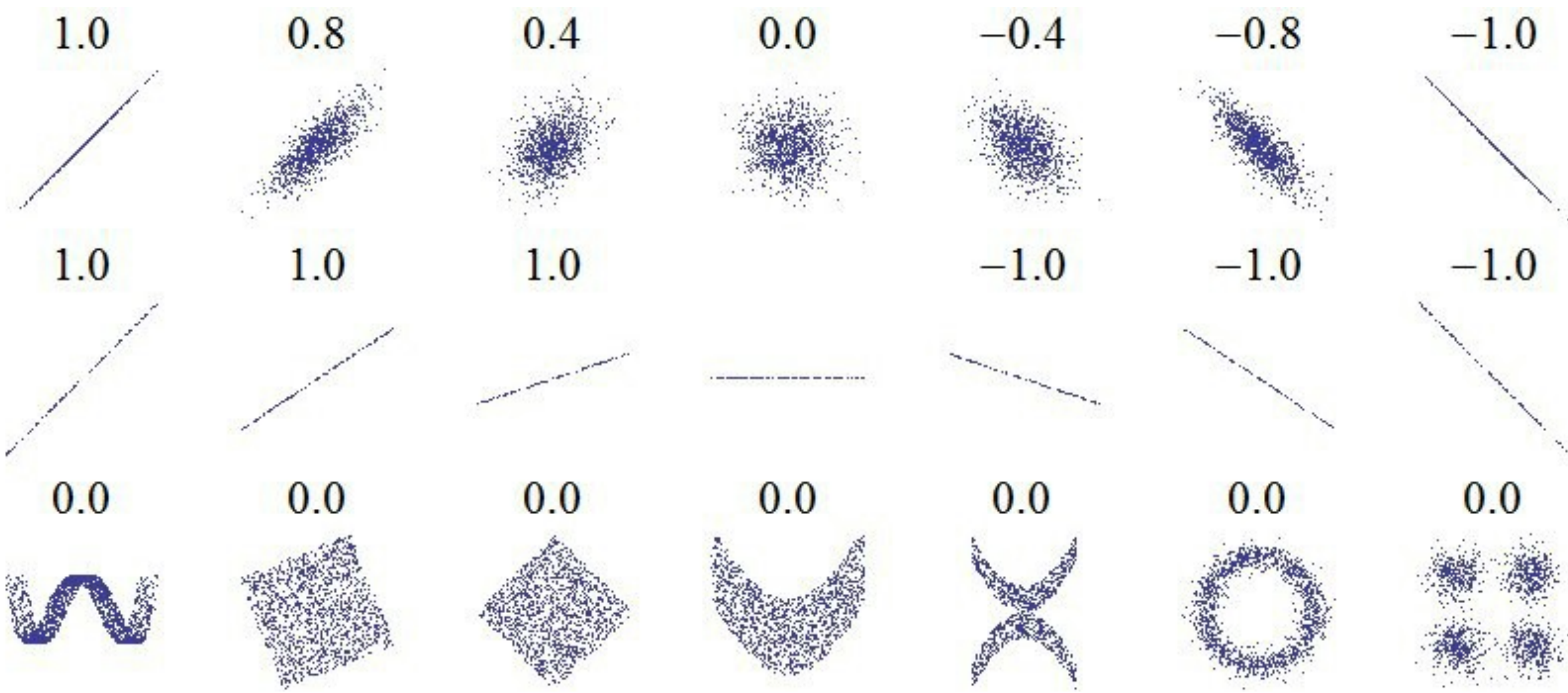
$\rho = -0.5$ $\rho = 0.5$

- 공분산을 각 변수의 표준편차로 나누어 scale effect를 제거함

Namun Cho/ mailto:namun@snu.ac.kr

ρ 와 다양한 분포

- 주의: 기울기와는 무관함



https://stats.stackexchange.com/questions/194636/is-it-correct-to-use-correlation-coefficient-in-this-case

Namun Cho/ mailto:namun@snu.ac.kr

포트폴리오의 리스크와 상관계수

- 지배원리 (Dominance Principle)
  - "기대수익률이 동일한 포트폴리오라면 분산이 적은 쪽이 더 낫다"
- 두 개의 금융자산 A,B 으로 구성된 포트폴리오 p의 위험 분산효과를 분석해보자

포트폴리오

주식	기대수익률 $E(R_i)$	표준편차 σ_i^2	투자비율 w_i
A	20%	0.25	50%
B	15%	0.17	50%

기대수익률

$$E(Rp) = E\left(\sum_{i \in \{A,B\}} w_i R_i\right)$$
$$= \sum_{i \in \{A,B\}} w_i E(R_i)$$

주식	기대수익률	표준편차	투자비율	wi*E(Ri)
A	20%	0.25	50%	10%
B	15%	0.17	50%	7.5%
			E(Rp)	17.5%

분산

$$Var(R_p) = Var\left(\sum_{i\in\{A,B\}} w_i R_i\right)$$

$$= \sum_{i\in\{A,B\}} w_i^2 Var(R_i) + 2\sum_{i<j} w_i w_j Cov(R_i, R_j), \quad i, j \in \{A, B\}$$

$$= \sum_i \sum_j w_i w_j Cov(R_i, R_j)$$

$$Cov(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$ 이므로

$$Var(R_p) = \sum_{i\in\{A,B\}} w_i^2 Var(R_i) + 2\sum_{i<j} w_i w_j \sigma_A \sigma_B \rho_{AB}, \quad i, j \in \{A, B\}$$

상관계수별 표준편차

$$\sigma_p = \sqrt{Var(R_p)} = \sqrt{\sum_{i \in \{A,B\}} w_i^2 Var(R_i) + 2 \sum_{i \neq j \in \{A,B\}} w_i w_j \sigma_A \sigma_B \rho_{AB}}$$

- ρ = 1 ⇒ 21%
  - 분산의 평균과 같음 (위험분산효과가 없음)
  - 분산투자이득 0%p
- σ = 0 ⇒ 15% (분산투자이득: 6%p)
- σ = -1 ⇒ 4% (분산투자이득: 17%p)
  - 위험분산효과 (변동성을 줄이는 효과)가 가장 큼

분산투자의 이득
Gain for Diversification

- ρ < 1 인 주식들에 나누어 투자할 경우 언제나 리스크는 감소함
- Negative Risk 뿐만 아니라 Positive Risk도 감소

최소분산포트폴리오 Minimum Value Portfolio

- 투자자의 선택 문제
- 투자자는 자신의 포트폴리오 구성 wi 를 바꿀 수 있음
- 따라서 wi 를 조절하여 자신의 목적을 달성할 수 있음
  - 최소분산포트폴리오: 분산을 최소로 하는 포트폴리오
  - 목적 설정에 따라 다른 구성도 가능
    - 예: 기대수익을 최대로 하는 포트폴리오 등

2기업 주식투자 문제

$$Var(R_p) = \sum_{i \in \{A,B\}} w_i^2 Var(R_i) + 2 \sum_{i<j} w_i w_j \sigma_A \sigma_B \rho_{AB}, \quad i, j \in \{A, B\}$$

- 이제 더이상 wA, wB는 0.5가 아님
    - 조절하여 분산을 최소로 하는 투자비율 찾기
    - 나머지 값은 시장에 의해 주어진 값: 상수취급
- wB = 1-wA 이므로 대입하여 재계산

$$\sigma_p^2 = w_A^2 \sigma_A^2 + (1 - w_A)^2 \sigma_B^2 + 2 w_A (1 - w_A) \sigma_{AB}$$

$$\arg\min_{w_A} \sigma_p^2$$

극소화 1,2계조건

$$\arg\min_{x} f(x)$$

- 이 문제는 최적화 (Optimization) 문제임
  - 목적함수 f(x)값을 가장 작게 만드는 x를 찾기
- 수리적으로 다음의 사실이 알려져 있음
  - x의 집합 내부 (경계가 아닌 값)에 해 x* 가 있다면 다음의 조건이 성립한다
  - $f'(x^*) = 0$ (1계조건: FOC: First Order Condition)
  - $f''(x^*) > 0$ (2계조건: SOC: Second Order Condition)
  - 이 명제의 역은 참이 아님

Solution

$$\arg\min_{w_A} \sigma_p^2$$

$$\sigma_p^2 = w_A^2\sigma_A^2 + (1-w_A)^2\sigma_B^2 + 2w_A(1-w_A)\sigma_{AB}$$

$$\frac{\partial \sigma_p^2}{\partial w_A} = 2w_A\sigma_A^2 - 2(1-w_A)\sigma_B^2 + 2(1-w_A)\sigma_{AB} - 2w_A\sigma_{AB} = 0$$

(FOC)

$$2\left(\sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}\right) w_A - 2\left(\sigma_B^2 - \sigma_{AB}\right) = 0$$

$$\frac{\partial^2 \sigma_p^2}{\partial^2 w_A} = 2\sigma_A^2 + 2\sigma_B^2 - 4\sigma_{AB} = 2(\sigma_A + \sigma_B)^2 > 0$$

(SOC)

$$\therefore w_A^* = \frac{\sigma_B^2 - \sigma_{AB}}{\sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}} = \frac{\sigma_B^2 - \sigma_A\sigma_B\rho_{AB}}{\sigma_A^2 + \sigma_B^2 - 2\sigma_A\sigma_B\rho_{AB}}, \quad w_B^* = 1 - w_A^*$$

n개의 동일 가중 포트폴리오의 수익률과 분산 구하기

- 동일 가중 포트폴리오 equally weighted portfolio
  - n개의 금융자산 비중이 1/n으로 동일한 자산 포트폴리오
  - 개별 금융자산의 수익률과 분산은 각각 Ri와 **σ**i로 표기하자

$$w_i = 1/n \quad \forall i$$

수익률

$$E(Rp) = E\left(\sum_i w_i R_i\right) = \frac{1}{n}\sum_i R_i$$

- 기대수익률은 모든 주식 수익률의 평균이 됨

분산

$$Var(R_p) = Var\left(\sum_i^n w_i R_i\right) = \sum_i \sum_j w_i w_j Cov(R_i, R_j)$$

$$= \frac{1}{n^2}\sum_i \sum_j Cov(R_i, R_j) = \frac{1}{n}\underbrace{\sum_i \frac{\sigma_i^2}{n}}_{\bar{\sigma}_i} + \frac{n-1}{n}\underbrace{\sum_i \sum_{j \neq i} \frac{\sigma_{ij}}{n(n-1)}}_{\bar{\sigma}_{ij}}$$

$$\lim_{n\to\infty} Var(R_p) = \lim_{n\to\infty}\left[\underbrace{\frac{1}{n}}_{\to 0}\bar{\sigma}_i + \underbrace{\frac{n-1}{n}}_{\to 1}\bar{\sigma}_{ij}\right]$$

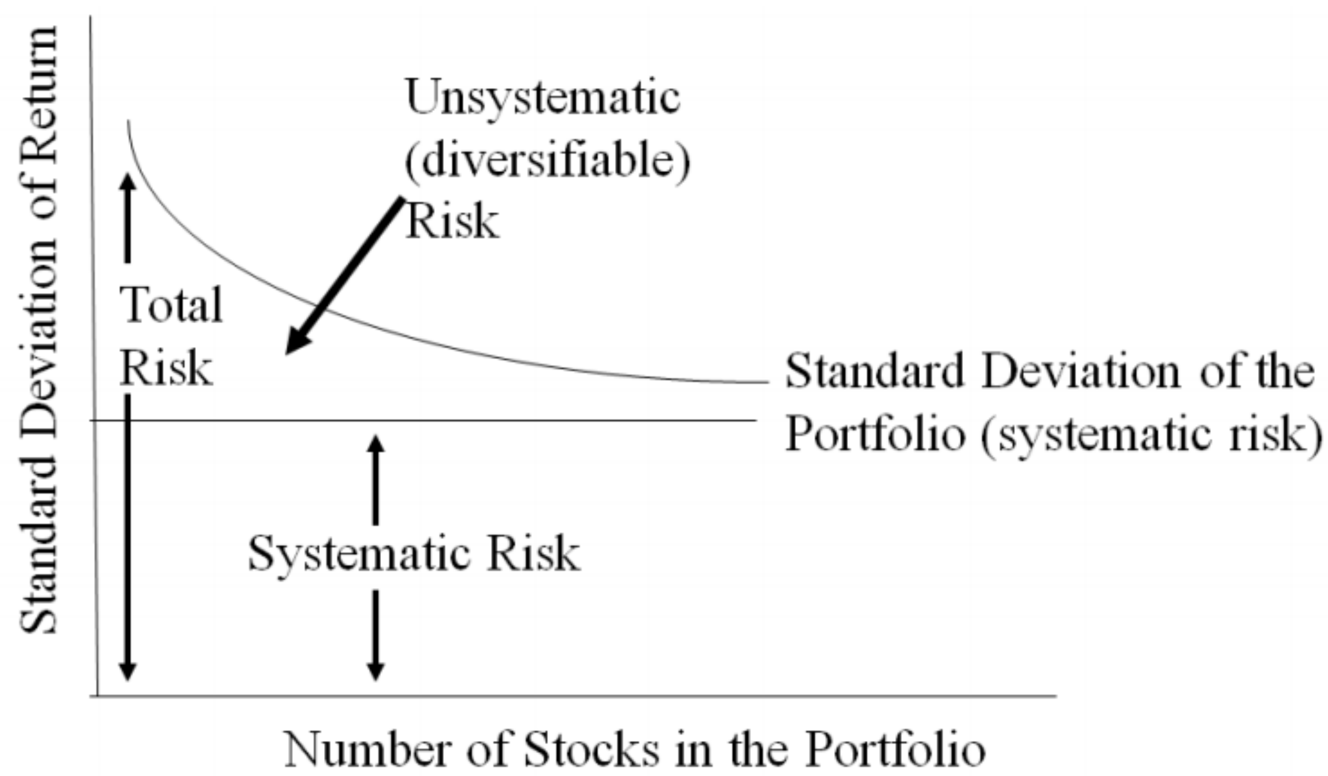
$$= \bar{\sigma}_{ij}$$

완전분산 포트폴리오

Completely Diversified Portfolio

완전분산포트폴리오의 의미

- 분산투자로 제거할 수 없는 위험을 의미
- 체계적 위험
- 분산투자로 제거할 수 있는 위험: 비체계적 위험



https://seekingalpha.com/article/4006697-many-stocks-make-well-diversified-portfolio

결합확률분포: 예
Joint Probability Distribution

- 두 확률변수의 동시 발생
  - 일종의 연산
  - X와 Y의 결합
- 예:
  - P(X=6 and Y=2) = 2/12

Y→ X↓	1	2
2	0	3/12
4	2/12	3/12
6	1/12	2/12
8	0	1/12

주변확률분포 Marginal Probability Distribution

- 결합확률분포로부터 구한 단일 확률변수의 확률분포

Y→ X↓	1	2	P(X)
2	0.00	0.25	0.25
4	0.17	0.25	0.42
6	0.08	0.17	0.25
8	0.00	0.08	0.08
P(Y)	0.25	0.75	

연습

- 다음 값들을 구해보라
  - E(X)
  - E(Y)
  - Var(X)
  - Var(Y)
  - Cov(X,Y)

Y→ X↓	1	2	P(X)
2	0.00	0.25	0.25
4	0.17	0.25	0.42
6	0.08	0.17	0.25
8	0.00	0.08	0.08
P(Y)	0.25	0.75	

일반화: 세 확률변수 X,Y,Z

$$P(X=x, Y=y) = \sum_{z} P(X=x, Y=y, Z=z)$$

$$P(X=x) = \sum_{y} \sum_{z} P(X=x, Y=y, Z=z)$$

$$\sum_{x} \sum_{y} \sum_{z} P(X=x, Y=y, Z=z) = 1$$

$$f(x, y) = \int_{z} f(x, y, z) dz$$

$$f(x) = \int_{y} \int_{z} f(x, y, z) dz dy \qquad \int_{x} \int_{y} \int_{z} f(x, y, z) dz dy dx = 1$$

기대값과 독립

$$E(XY) = E(X)E(Y)$$

$$E(XY) := \sum_{i}^{n} \sum_{j}^{n} x_i y_j P(X = x_i, Y = y_j)$$

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$ ∵ X, Y가 독립

$$E(XY) = \sum_{i}^{n} x_i P(X = x_i) \cdot y_j \sum_{j}^{n} P(Y = y_j) = E(X)E(Y)$$

Namun Cho/ mailto:namun@snu.ac.kr

독립사건의 Covariance

- 독립사건의 Covariance = 0

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$
$$= E(X)E(Y) - E(X)E(Y) = 0$$

표본 추정량

- 통계학의 주요 목표:
  - 모수들에 대한 E(X), Var(X), Cov(X,Y)들을 표본으로부터 추정하기
- 실제 모수값은 X, Y의 분포, 그리고 X와 Y의 결합분포를 정확히 알고 있어야 할 수 있음
- 표본으로부터 구할 수 있는 해당 모수들에 대한 자연스러운 추정량은 표본으로부터 구한 E, Var, Cov 들일 것임
- 모수대신 추정치를 쓸 경우 데이터 하나를 사용하므로 n 대신 n-1 로 나누는 것

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

$$Var(X) = E(X - E(X))^2$$

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

$$Cov(X, Y) = E(X - E(X))E(Y - E(Y))$$

$$\Rightarrow \frac{1}{n-1}\sum_{i}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

평균에 대한 불편추정량

Proof.

표본평균 $\bar{X}$ 가 불편 (unbiased) 추정량이기 위해서는 아래의 식이 성립해야 한다 :

$$E(\bar{X} - X) = 0, \quad \text{or} \quad \bar{X} = E(X) = \mu$$

따라서 $E(\bar{X}) = \mu$ 임을 증명하면 된다.

$$E(\bar{X}) = E(\frac{i}{n} x_i) = \frac{1}{n} \sum_i^n E(x_i)$$

$$= \frac{1}{n} \sum_i^n E(X) = E(X) = \mu$$

$\square$

분산에 대한 불편추정량

Proof.

$$Var(X) := E((X - E(X))^2) = E((X - \mu)^2) = E(X^2 - 2\mu X + \mu^2)$$

$$= E(X^2) - 2\mu E(X) + \mu^2$$

모분산 $Var(X)$를 σ^2 이라고 하고 표본 $\hat{X}$ 의 분산을 $Var(\hat{X})$라고 하면 표본의 분산은 다음과 같이 기술할 수 있다

$$Var(\hat{X}) = E(\hat{X}^2) - \bar{X}^2$$

$$= E\left(\frac{\sum_i^n x_i^2}{n}\right) - \left(\frac{\sum_i^n x_i}{n}\right)^2$$

분산에 대한 불편추정량 (계속)

Proof.

$$= \frac{\sum_i^n E(x_i^2)}{n} - \frac{1}{n^2} E\left(\sum_i^n x_i^2 + 2\sum_i \sum_{j\neq i} x_i x_j\right)$$

$$= \frac{n-1}{n^2}(\sum_i^n E(x_i^2)) - \frac{2}{n^2}\left(\sum_i \sum_{j\neq i} E(x_i x_j)\right) \tag{1}$$

임의추출이므로 $x_i, x_j (j \neq i)$는 서로 독립이므로

$$E(x_i x_j) = E(x_i)E(x_j), \quad i \neq j$$

$$E(x_i^2) = E(X^2) = \sigma^2 + \mu^2$$

$$E(x_i) = E(X) = \mu$$

분산에 대한 불편추정량 (계속)

Proof.

위 세 식을 식(1)에 대입하면:

$$\frac{n-1}{n^2}n(\sigma^2+\mu^2)-\frac{1}{n^2}\mu^2(n^2-n)=\frac{n-1}{n}\sigma^2$$

$$\therefore E(Var(\hat{X}))=\frac{n-1}{n}\sigma^2$$

$$\sigma^2=\frac{n}{n-1}E(Var(\hat{X}))=E\left(\frac{n}{n-1}Var(\hat{X})\right)=\frac{n}{n-1}\left(\frac{\sum_i^n(x_i-\bar{X})^2}{n}\right)$$

$$=\frac{1}{n-1}\sum_i^n(x_i-\bar{X})^2$$

□

n-1로 나누는 이유에 대한 직관적 설명

- $E(X)=\mu$ 이지만 $E(X^2) = \mu^2 + \sigma^2$ 라는 사실에서 기인
- 즉, 표본의 제곱의 기대값은 모평균의 제곱이 아니라 모평균의 제곱에 모분산을 더한 값(더 큰 값)이기 때문에 n이 아닌 n-1 로 나누어야 분포성을 설명할 수 있다는 말임

공분산행렬
Covariance Matrix

- 각 변수의 분산 조합을 행렬로 나타낸 것
- 가령 예제 4.10에서 X, Y, Z의 공분산행렬에서 가령
- Cov(X,Y) = 1/(3-1) * [(21-8)(6-9)+(-12-8)(9-9)+(15-8)(12-9)]= 30
- Cov(X,X) = Var(X)
- 대칭행렬이므로 삼각행렬만으로 충분

%	X	Y	Z
Year 1	21	6	5
Year 2	-12	9	4
Year 3	15	12	6
Mean	8	9	5

%	X	Y	Z
X	Cov(X,X)	Cov(X,Y)	Cov(X,Z)
Y	Cov(Y,X)	Cov(Y,Y)	Cov(Y,Z)
Z	Cov(Z,X)	Cov(Z,Y)	Cov(Z,Z)

경우의 수

경우의 수에 대한 기본원칙

- 확률론에서는 경우의 수에 대한 이해가 필요함
- 본 절에서는 이에 대한 몇 가지 원칙을 확인하고자 함

카드 늘어놓기

- 3장의 서로 다른 카드 (J,Q,K)를 배열하는 모든 경우의 수는?
- 첫번째 자리에 놓을 수 있는 경우의 수:3
  - 두번째 자리에 놓을 수 있는 경우의 수: 2
    - 세번째 자리에 놓을 수 있는 경우의 수:1
- 결합하면 3*2*1=6
- 일반화: 서로 다른 n개의 것을 늘어놓는 경우의 수는 n!

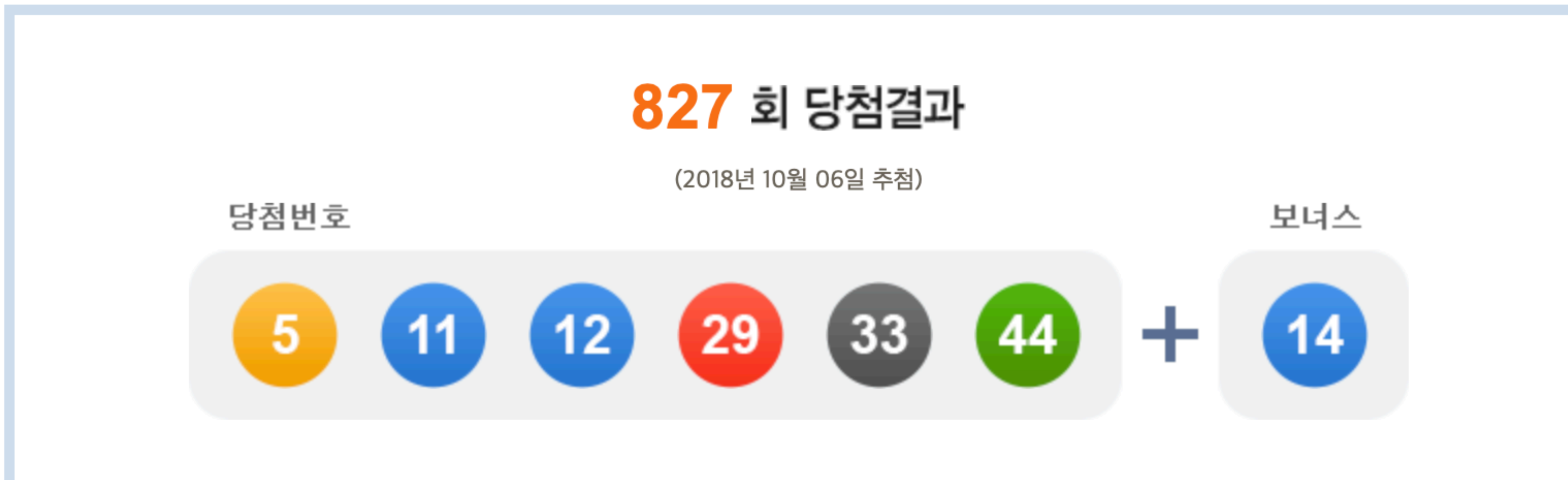
	1st	2nd	3rd
1	J	Q	K
2	Q	J	K
3	Q	K	J
4	J	K	Q
5	K	J	Q
6	J	Q	J

$$n! := \prod_{i=n}^{1} i$$

n종류의 물건 r(<n)개를 늘어놓는 경우의 수

- 로또는 45개의 숫자 중 순서 관계 없이 6개를 고르는 방식의 복권
- 순서가 의미있을 경우:

$$nPr := \frac{n!}{(n-r)!}$$



순위	등위별 총 당첨금액	당첨게임 수	1게임당 당첨금액	당첨기준	비고
1등	18,097,505,625원	15	1,206,500,375원	당첨번호 **6개** 숫자일치	1등 자동8 수동7
2등	3,016,250,958원	46	65,570,673원	당첨번호 **5개** 숫자일치 +보너스 숫자일치	
3등	3,016,252,562원	2,113	1,427,474원	당첨번호**5개** 숫자일치	
4등	5,308,200,000원	106,164	50,000원	당첨번호 **4개** 숫자일치	
5등	8,806,855,000원	1,761,371	5,000원	당첨번호 **3개** 숫자일치	

Namun Cho/ mailto:namun@snu.ac.kr

로또의 확률

$$nPr := \frac{n!}{(n-r)!}$$

- n=45
- r=6
- 로또의 룰은 순서를 의미없게 생각하므로 6! 경우의 수만큼을 나눠야 함
- 1등 당첨확률은 약 1/800만

$$_{45}P_6 = \frac{45!}{(45-6)!} = 5864443200$$

$$_nC_r = \binom{n}{k} := \frac{n!}{r!(n-r)!}$$

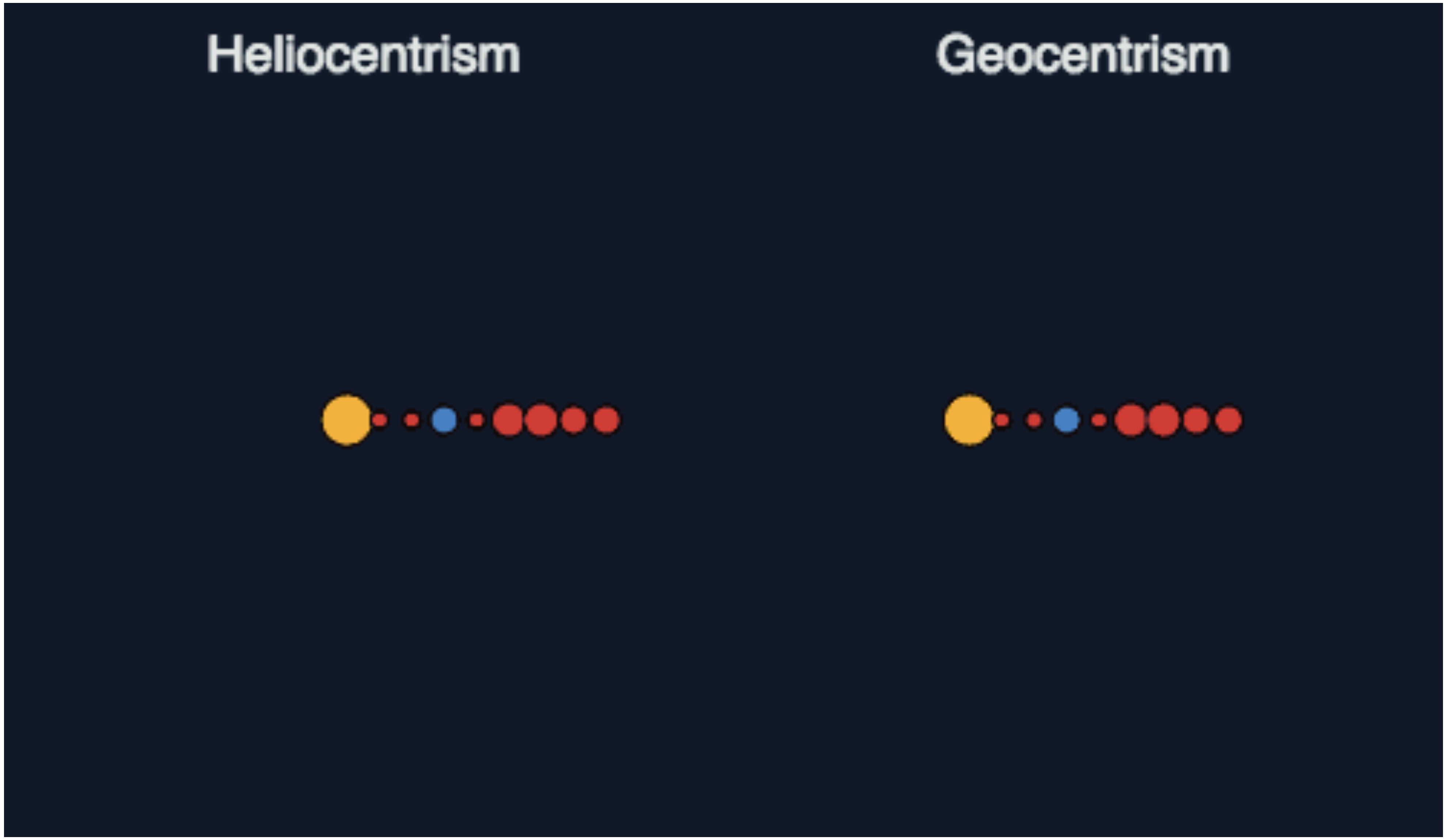
$$\binom{45}{6} = \frac{_{45}P_6}{6!} = 8145060$$

Namun Cho/ mailto:namun@snu.ac.kr

Next Topics

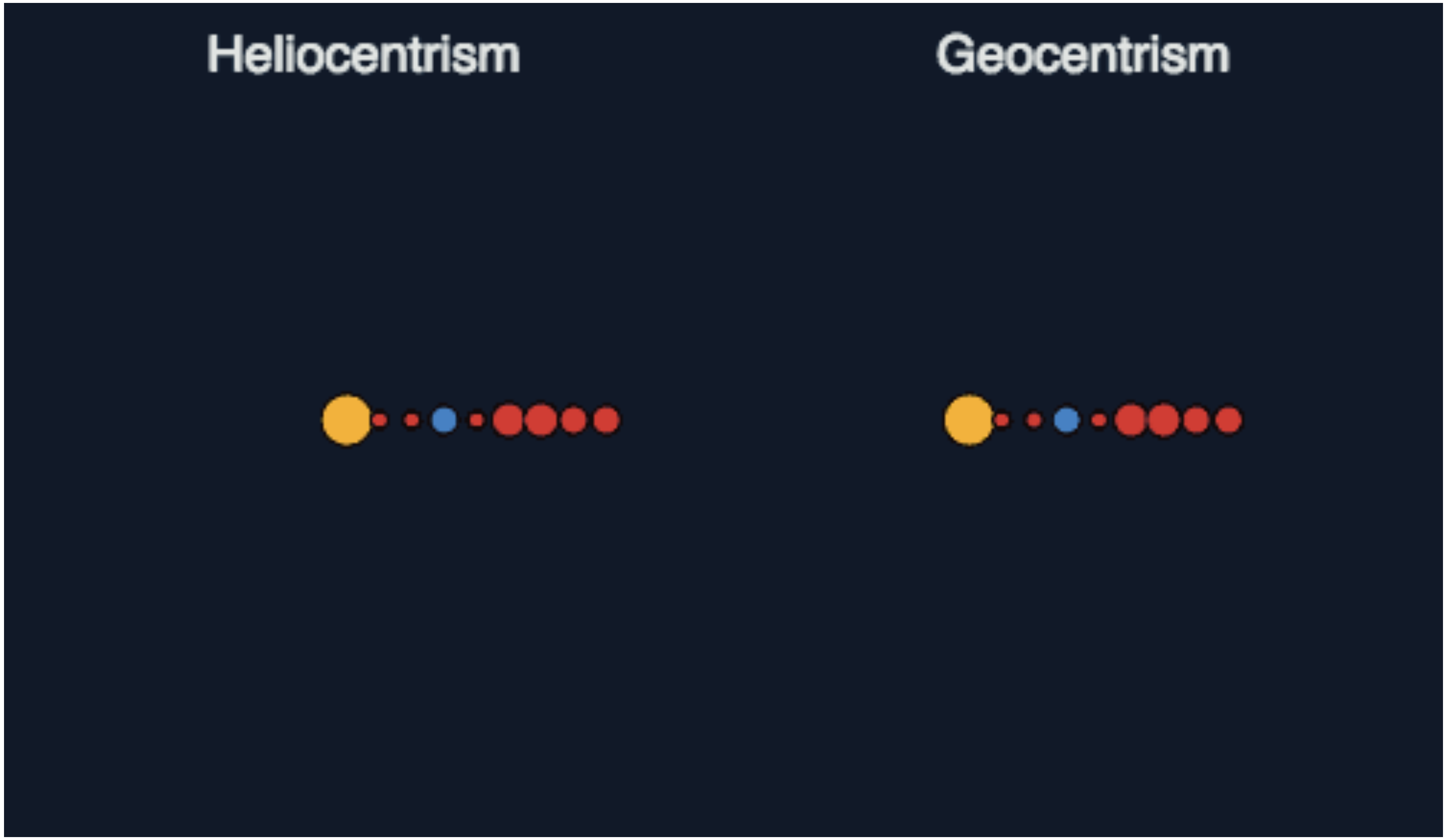
- 분포이론과 중심극한정리
  - 주교재 5장
- 금융데이터의 측정과 확률분포
  - 보조교재 4장

수고하셨습니다!



Namun Cho/ mailto:namun@snu.ac.kr

수고하셨습니다!



Namun Cho/ mailto:namun@snu.ac.kr