



이화여자대학교
EWha WOMANS UNIVERSITY

확률 및 통계학

■ 확률













송수민

soominsong@ewha.ac.kr

확률의 정의

- **확률** = 전체 모집단 중 어떤 특정한 사건이 나타날 비율
 = 특정사건의 수 / 전체 모집단의 수 (교재 31페이지 그림 2.1 참고)
- 확률은 항상 0과 1사이의 값을 가지며, 모든 구간의 확률을 합하면 1

전세계 인구수 : 69억명 (2010년도 기준 추정)

	국가명	인구(명)	기준년도	대륙	수도
1	 중국	1,338,612,968	2010	아시아	베이징
2	 인도	1,156,897,766	2010	아시아	뉴델리
3	 미국	307,212,123	2010	북아메리카	워싱턴
4	 인도네시아	240,271,522	2010	아시아	자카르타
5	 브라질	198,739,269	2010	남아메리카	브라질리아
6	 파키스탄	174,578,558	2010	아시아	이슬라마바드
7	 방글라데시	156,050,883	2010	아시아	다카
8	 나이지리아	149,229,090	2010	아프리카	아부자
9	 러시아	140,041,247	2010	유럽	모스크바
10	 일본	127,078,679	2010	아시아	도쿄
25	 대한민국	48,508,972	2010	아시아	서울
50	 북한	22,665,345	2010	아시아	평양

전세계 사람 중
 무작위로 1명을 뽑았을 때
 그 사람이 브라질 사람일 확률
 = $198,739,269 / 69\text{억}$
 = 0.029

배반사건, 합집합의 확률

1000명의 성인 남자를 랜덤하게 뽑아 체형과 소득수준을 조사하였다.

표2.2 성인 남자의 체형과 소득수준

체형	소득수준			
	낮음	보통	높음	합계
마른형	100	150	50	300
보통형	50	350	70	470
비만형	120	60	50	280
합계	270	560	170	1000

- (1) 임의로 뽑은 성인남자가 보통체형이고 소득수준이 높음일 확률
- (2) 임의로 뽑은 성인남자가 보통체형이거나 소득수준이 높음일 확률
- (3) 임의로 뽑은 성인남자가 마른체형이거나 비만형일 확률

조건부 확률(1)

(교과서 38페이지 예 2.4)

성인남자의 체형과 소득수준 데이터

- (1) 보통체형의 성인남자 중 한 명을 임의로 뽑았을 때 소득수준이 높음일 확률
- (2) 소득수준이 높음인 성인남자 중 한 명을 임의로 뽑았을 때 체형이 보통체형일 확률

조건부 확률(2)

● 조건부 확률의 활용 - 연관성 분석(Association Rules)

- 장바구니 분석이라고도 하며, 하나의 사건에 포함되어 있는 둘 이상의 변수들의 상호 관련성을 발견하는 분석법
- 연관성 분석의 평가기준

지지도 (발생비율) (Support)	<ul style="list-style-type: none">▪ 전체 거래 건수(또는 고객) 중 A와 B를 함께 구매한 거래 건수 비율(또는 고객 수 비율)▪ $\text{Support} = (A \cap B) \div \text{전체}$▪ Support가 5%라는 것은 전체 거래건수 중에서 5% 거래가 A와 B를 함께 구매한 것을 의미
신뢰도 (연관성) (Confidence)	<ul style="list-style-type: none">▪ A를 구매한 거래 건수 중 B를 함께 구매한 거래 건수 비율로 상품 A와 B의 연관성 평가함▪ $\text{Confidence} = (A \cap B) \div A$▪ Confidence가 7%라는 것은 A를 산 영수증 건수에서 7%가 B를 동시에 샀다는 것을 의미
향상도 (유용성) (Lift)	<ul style="list-style-type: none">▪ 전체 거래 중 B가 포함되어 있는 거래 비율 대비 A를 구매한 거래 중 B를 동시에 구매한 거래의 비율로 연관규칙의 유용성을 평가함▪ $\text{Lift} = ((A \cap B) \div A) / (B \div \text{전체})$▪ Lift가 1보다 크면 클수록 양의 상관관계가 많으므로 유용한 연관규칙임을 의미

사건의 독립성

성인남자의 체형과 소득수준 데이터

- (1) 성인남자 중 두 명을 임의로 뽑았을 때, 소득수준이 한 명은 낮음이고, 한 명은 높을 일 확률
- (2) 성인 남자 중 두 명을 임의로 뽑았을 때, 두 명 모두 체형이 마른형일 확률

주사위를 두 번 던졌다.

- (1) 첫 번째는 짝수가 나오고, 두 번째에는 1이나 5가 나올 확률
- (2) 두 번 모두 홀수가 나올 확률

베이즈 공식(1)

(교과서 38페이지 예 2.5, 예 2.6)

베이즈 공식(2)

어느 마을에 공장이 생겼다.
공장으로 인해 대기오염이 발생할
확률은 **0.6**이다. 대기오염이 있을 때
마을주민의 호흡기 질환 발생율은 **0.8**,
대기 오염이 없을 때는 **0.1**이다.

- 마을 주민 중 한 명이 호흡기 질환에 걸렸을 때 원인이 대기오염일 확률

보험회사는 사고 위험률을 기준으로
낮음, 보통, 높음으로 나눈다.
각 부류의 사람은 각각 **20%, 50%, 30%**이며, 1년간 사고를 낼 확률은
0.05, 0.15, 0.30이다.

- 1년간 사고를 내지 않은 사람이 사고 위험률이 높음일 확률

베르누이 시행(이항실험)

- 베르누이 시행(이항실험)
 - 각 시행에서 가능한 결과는 성공과 실패 두 가지
 - 각 시행은 독립적
 - 각 시행에서 성공 확률이 일정
- 성공 확률이 p 인 베르누이 시행을 n 번 반복했을 때 성공횟수 X 에 대한 확률

$$- P(X = k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- 교과서 46페이지 예 2.7

A공장 생산품의 불량률이 **5%**라 한다.
이 공장에서 생산되는 제품 중 열개를 랜덤추출 하였다.

- (1) 불량품이 8개일 확률
- (2) 불량품이 8개 이상일 확률