1. 통계학이란

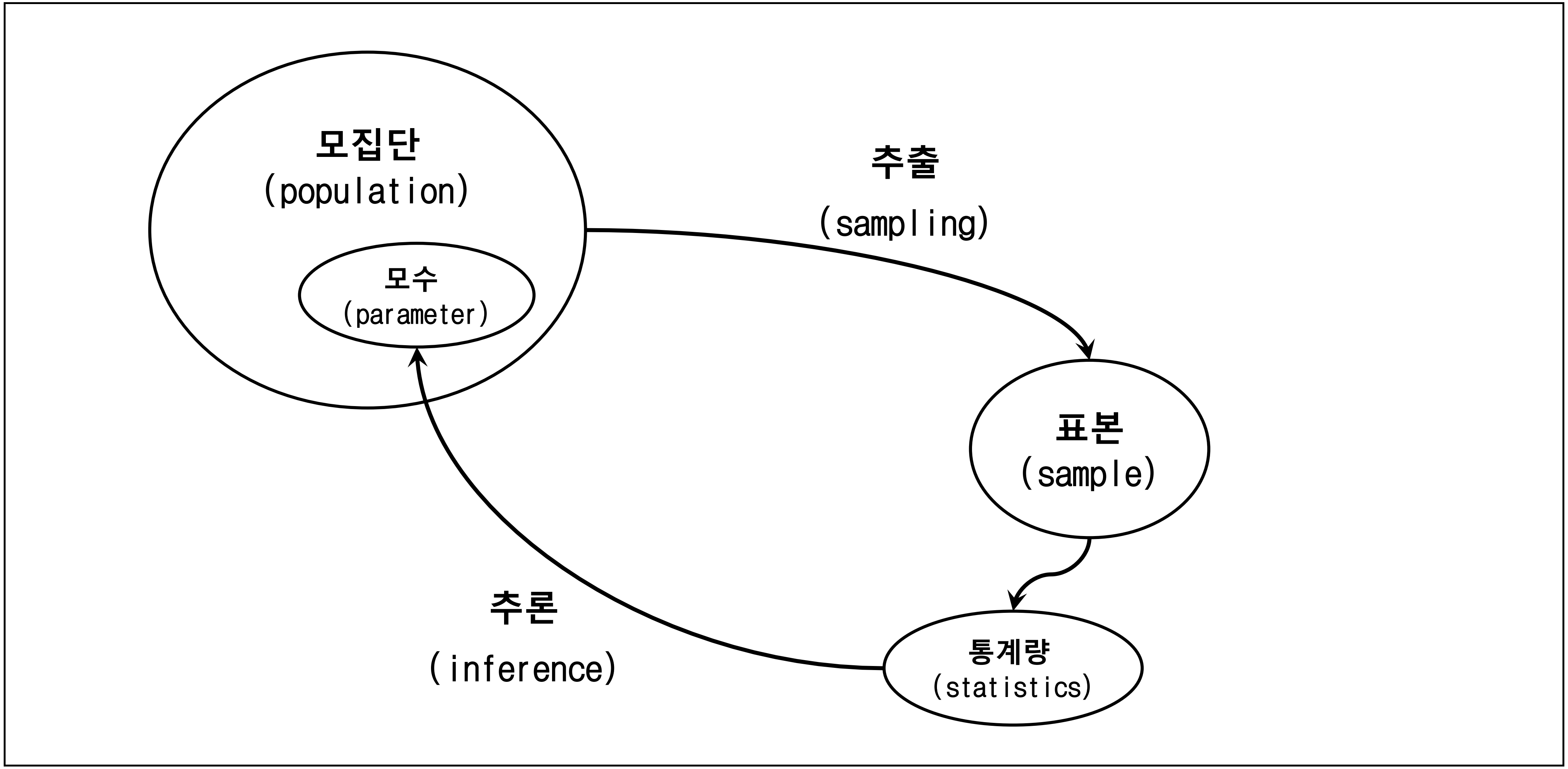
통계학의 정의 및 개요

- **통계학이란?**

  과학적인 방법에 의해 자료를 수집하고 목적에 맞는 적절한 방법에 의해 정리, 분석하는 학문.

  - **기술통계학 (descriptive statistics)**
    - 자료의 특성을 표, 그림, 통계량 등을 사용하여 쉽게 파악할 수 있도록 정리, 요약하는 방법을 다루는 분야.
  - **추론통계학 (inferential statistics)**
    - 관찰 불가능한 모집단의 여러 특성에 대해 관찰 가능한 표본 자료의 정보를 분석하여 과학적으로 추론하는 방법을 다루는 분야.
    - 합리적인 의사결정을 위한 근거를 제공하는 것이 목적.

통계학의 정의 및 개요



모집단
(population)
모수
(parameter)
추출
(sampling)
표본
(sample)
통계량
(statistics)
추론
(inference)

자료의 종류

- **통계학에서의 자료의 구분**
  - **질적 자료(qualitative data)**
    - 자료 값이 양적인 의미를 가지지 않는 자료. 주로 문자로 표현되며, 사칙연산이 불가능함.
    - 관측결과가 몇 개의 범주 또는 항목의 형태로 나타나는 자료인 경우 **범주형 자료** (categorical data)라고도 함.

    - **명목 자료**(nominal data) : 순위의 개념이 없다.
      예) 혈액형, 성별, 직업

    - **순서 자료**(ordinal data) : 순위의 개념을 갖는다.
      예) 학점, 선호도, 옷 사이즈

    - 경우에 따라 질적 자료가 숫자로 표현되기도 함.
      예) 교육수준 : 중졸 -> 1, 고졸 -> 2, 대졸 -> 3

자료의 종류

- **양적 자료(quantitative data)** 또는 **수치형 자료(numerical data)**
  - 자료 자체가 숫자로 표현되며 숫자 자체가 양을 나타냄. 사칙연산이 가능함.

  - **연속형 자료**(continuous data)
    예) 시간, 길이, 온도, 무게

  - **이산형 자료**(discrete data) 또는 계수형 자료(counting data)
    예) 교통사고 건수, 고객의 수, 불량품의 수

  - 경우에 따라 양적 자료는 범주화가 가능하다.
    예) 시험성적 : 90~100 -> 수, 80~89 -> 우, 70~79 -> 미, 60~69 -> 양, 0~59 -> 가

2. 확률

확률의 기본 개념과 규칙, 성질

- **확률모형과 확률에 관한 기본 개념**
  - 확률실험 (random experiment)
    - 시행을 반복할 때마다 나오는 결과가 우연에 의존하여 매번 달라지는 현상 또는 실험
  - 표본공간 (sample space)
    - 확률 실험에서의 모든 관찰 가능한 결과의 집합, S로 표기

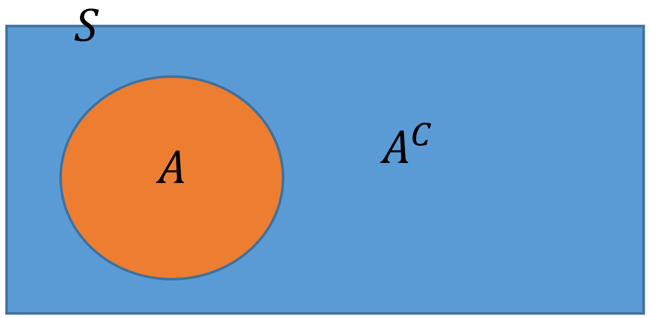
      - 동전 1개를 던지는 실험 : $S = \{H, T\}$

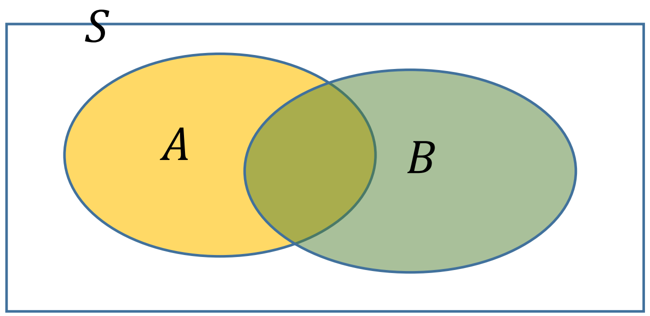
      - 주사위 1개를 던지는 실험 : $S = \{1, 2, 3, 4, 5, 6\}$

      - 내일 A 기업의 주가의 수익률 예측 : $S = \{R \mid -30\% \le R \le 30\%\}$

확률의 기본 개념과 규칙, 성질

- 사건 (event)
  - 표본공간의 임의의 부분집합, A, B 등으로 표기
    - $P[A^C]$: 사건 A 를 제외한 나머지 사건의 확률
    - $P[A \cap B]$: 사건 A 와 사건 B가 동시에 발생할 확률
    - $P[A \cup B]$: 사건 A 또는 사건 B가 발생할 확률





확률의 기본 개념과 규칙, 성질

- **확률의 정의**
  - 고전적 접근 (P. Laplace)
    - n개의 실험결과로 구성된 표본공간에서 각 실험결과가 일어날 가능성이 같은 경우,
    - m $(m \leq n)$ 개의 실험 결과로 구성된 사건 A 의 확률을 아래와 같이 정의함.

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

  - 상대적 비율에 의한 접근 (Richard Von Mises)
    - n 번의 반복된 실험 중 어떤 사건 A가 발생한 횟수를 m이라고 할 때,
      사건 A의 상대빈도는 $\frac{m}{n}$으로 구해짐.
    - 이 실험의 반복 횟수 n을 무한히 증가했을 때, 사건 A의 상대빈도가 수렴하는 값을 사건 A 의 확률로 정의하고자 함.

확률의 기본 개념과 규칙, 성질

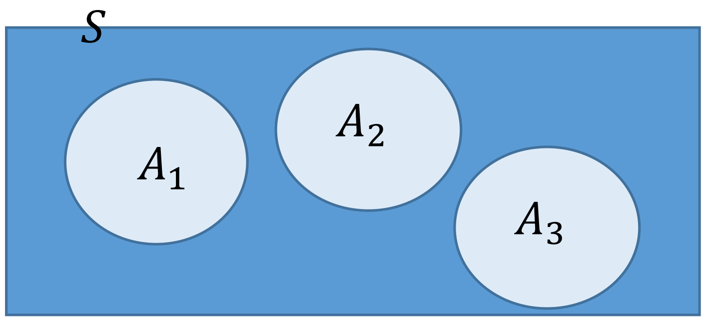
- **확률의 공리적 정의**
  - 확률의 공리(A. N. Kolmogorov)

    (1) 임의의 사건 A에 대하여 $P(A) \geq 0$

    (2) $P(S) = 1$

    (3) 표본공간 S에 정의된 서로 상호배반인 사건 $A_1, A_2, \ldots$에 대하여

    $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$가 성립

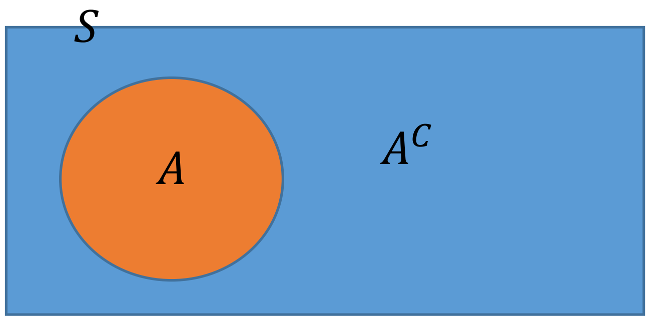


  - 공리적 접근방식

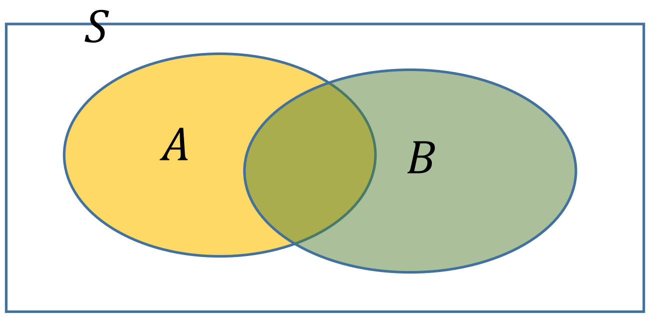
    표본공간을 정의역으로 하며, 위 세가지 공리를 만족하는 함수를 확률로 정의.

확률의 기본 개념과 규칙, 성질

- **확률의 규칙**
  - 여사건의 확률
    - $P[A^C] = 1 - P[A]$



  - 합사건의 확률
    - $P[A \cup B] = P[A] + P[B] - P[A \cap B]$



확률의 기본 개념과 규칙, 성질

- 예제

  A 은행 개인고객 중 예금이 5% 이상 증가한 고객 비율은 20%이다. A 은행 개인고객 중 가계대출이 5% 이상 증가한 고객 비율은 30%이다. A 은행 고객 중 예금이 5% 이상 증가하면서 가계대출이 5% 이상 증가한 고객 비율은 10%이다.

  - A 은행에서 임의로 1명을 뽑았을 때 그 고객이 가계대출이 5%이상 증가하지 않은 고객일 확률을 구하여라.

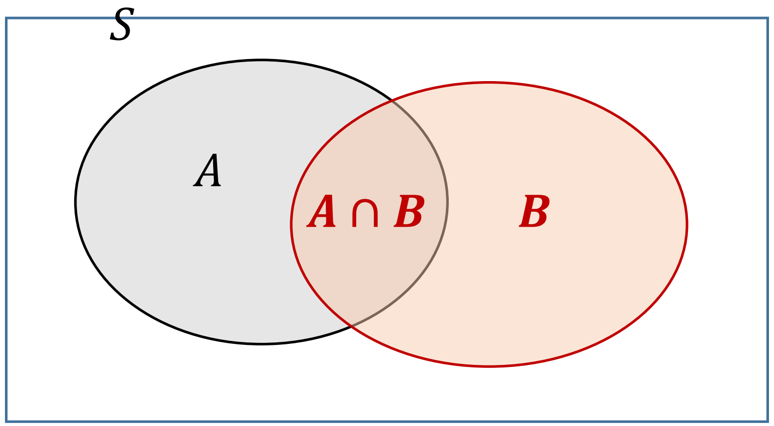
  - A 은행에서 임의로 1명을 뽑았을 때 그 고객이 예금이 5% 이상 증가하거나, 가계대출이 5% 이상 증가한 고객일 확률을 구하여라.

조건부 확률과 독립

- **조건부 확률**
  - 조건부 확률의 정의

사건 A와 B가 표본공간 S 상에 정의되어 있으며 $P(B) > 0$라고 가정.
이 때 B가 일어났다는 가정 하의 사건 A가 일어날 조건부확률은 다음과 같이 정의됨.

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



조건부 확률과 독립

- 예제
  - 한 개의 주사위를 던질 때 결과가 2 이상인 사건을 A, 결과가 4 미만인 사건을 B라 할 때, $P[A|B]$, $P[B|A]$를 구하여라.

조건부 확률과 독립

- 예제

어느 은행의 기업고객 200개를 대상으로 신용등급과 영업이익을 조사한 자료가 다음과 같다. 신용등급은 "우수"와 "불량"으로 구분되고, 영업이익은 "흑자"와 "적자"로 구분된다.

		신용등급		계
		우수	불량	
영업이익	흑자	70	10	80
	적자	30	90	120
계		100	100	200

- 임의로 선택된 기업의 영업이익이 "흑자"일 확률을 구하여라

- 임의로 선택된 기업의 신용등급이 "우수"일 때, 이 기업의 영업이익이 "흑자"일 확률을 구하여라.

조건부 확률과 독립

- **통계적 독립**
  - 독립 사건의 정의

    두 사건 A와 B가 다음 중 하나를 만족시키면 서로 독립이라고 함. (단, $P(A) > 0$, $P(B) > 0$)

    - $P(A|B) = P(A)$
    - $P(A \cap B) = P(A) \cdot P(B)$
    - $P(B|A) = P(B)$

조건부 확률과 독립

- 예제

  한국종합주가지수가 전거래일에 비해 상승할 확률이 0.3이고, 일별 상승여부가 서로 독립적이라고 한다.

  - 한국종합주가지수가 3일 연속을 상승할 확률을 구하시오.

  - 한국종합주가지수가 2일 연속으로 하락하지 않을 확률을 구하시오.

조건부 확률과 독립

- 예제

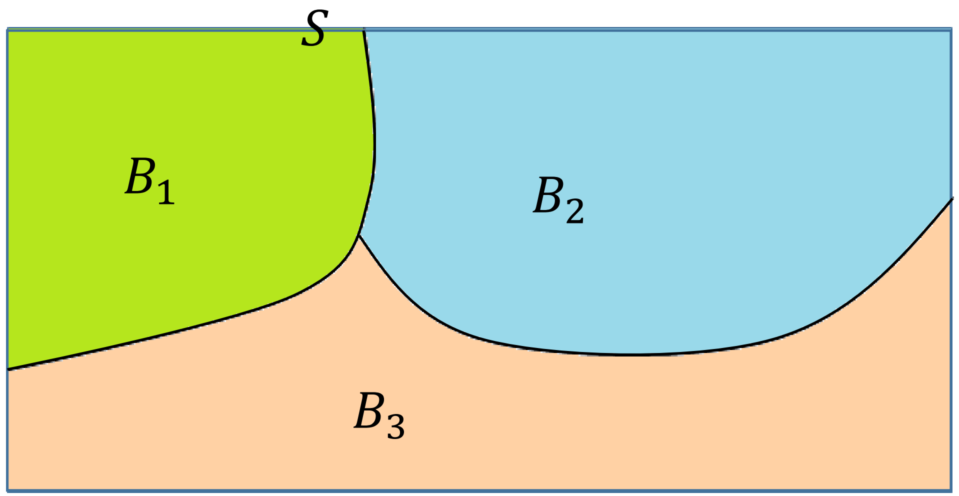
  어느 은행에서 고객 300명을 대상으로 성별과 그 은행에서 최근에 출시한 금융상품에 가입했는지 여부를 다음과 같이 조사하였다.

		상품 가입 여부		계
		가입	가입하지 않음	
성별	남성	100	50	150
	여성	100	50	150
계		200	100	300

- ◆ 고객이 남성인지 여부와 금융상품에 가입했는지 여부는 서로 독립인가?

베이즈 정리

- **표본공간의 분할**
  - $B_1, \ldots, B_k$가 다음 조건을 만족하면 표본 공간 S의 분할이라고 함.
    - 서로 다른 i, j에 대해 $B_i \cap B_j = \emptyset$: 상호배반
    - $B_1 \cup B_2 \cup \cdots \cup B_k = S$
  - $k = 3$인 경우

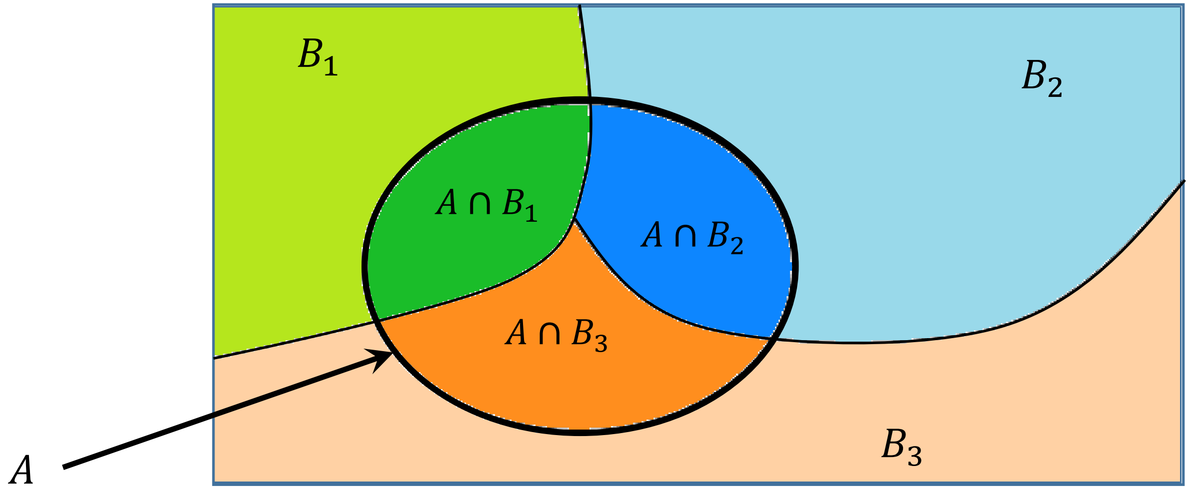


베이즈 정리

- **전확률공식**
  - 사건 $B_1, B_2, \ldots, B_k$는 상호배반이며, $B_1 \cup \cdots \cup B_k = S$라고 하자.
  - 이 때 S에서 정의되는 임의의 사건 A에 대하여 다음이 성립.

$$\begin{aligned} P(A) &= P(A \cap B_1) + \cdots + P(A \cap B_{\boldsymbol{k}}) \\ &= P(B_1)P(A|B_1) + \cdots + P(B_k)P(A|B_k) \end{aligned}$$

- $k = 3$인 경우 : $P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)$

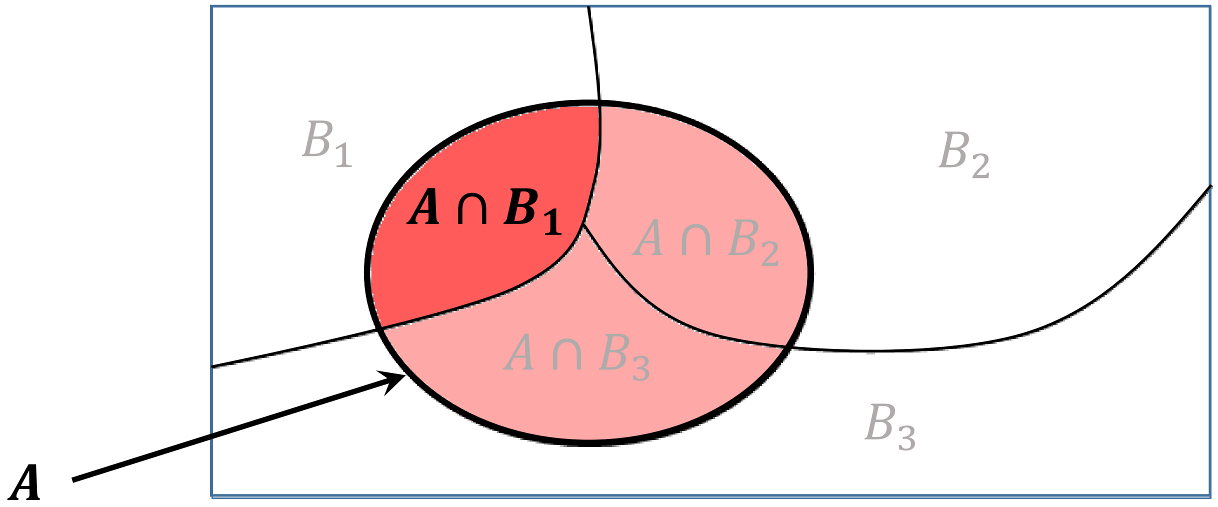


베이즈 정리

- **베이즈 정리**
  - 사건 $B_1, B_2, \ldots, B_k$는 상호배반이며, $B_1 \cup \cdots \cup B_k = S$라고 하자.
  - 이 때 사건 A가 일어났다는 조건 하에서 사건 B_i가 일어날 확률은 다음과 같다.

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{P(B_1)P(A|B_1) + \cdots + P(B_k)P(A|B_k)}$$

◆ $k = 3$인 경우 : $P(B_1|A) = \frac{P(A \cap B_1)}{P(A)} = \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1)+P(B_2)P(A|B_2)+P(B_3)P(A|B_3)}$



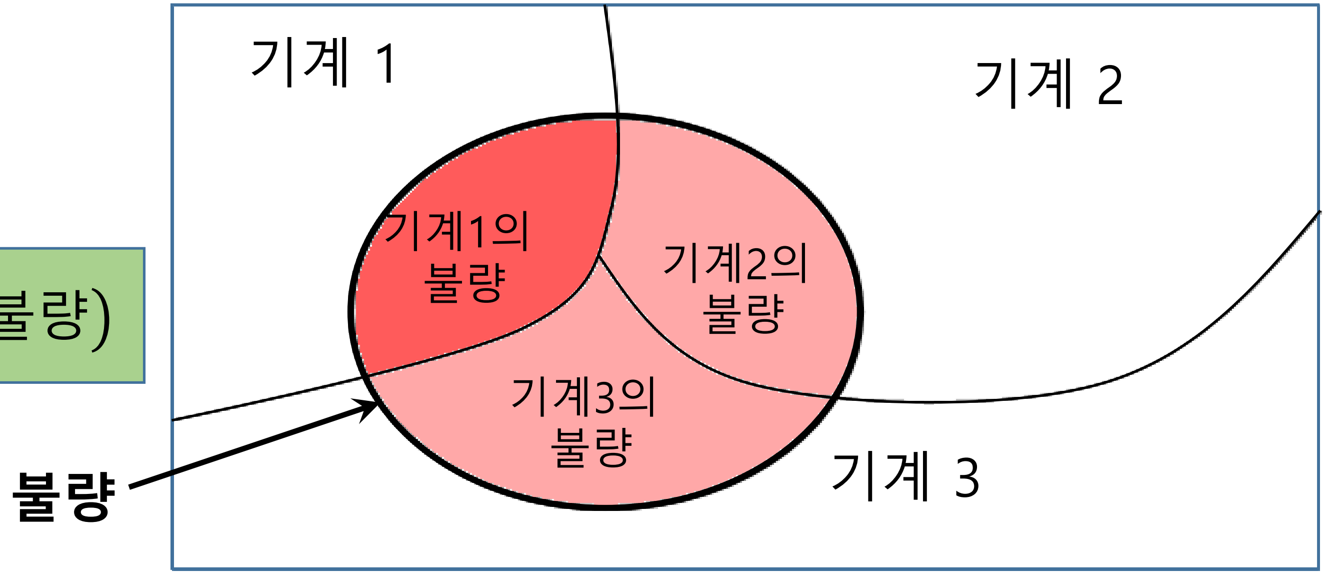
베이즈 정리

- 베이즈 정리 활용
  - $B_1, B_2, \ldots, B_k$ 으로 분할된 사건의 각 확률을 알고,
  - 각 B_i를 전제로 했을 때의 사건 A가 발생할 조건부 확률을 알 때,
  - 사건 A를 전제로 한 각 B_i의 조건부 확률을 구하기 위한 정리

$P(\text{기계}1)$
$P(\text{기계}2)$
$P(\text{기계}3)$

$P(\text{불량}|\text{기계}1)$
$P(\text{불량}|\text{기계}2)$
$P(\text{불량}|\text{기계}3)$

$P(\text{기계}1|\text{불량})$



베이즈 정리

- 예제
  - 보험회사는 사람들을 사고 위험률이 낮은 사람, 보통인 사람, 높은 사람의 세 부류로 나누는데 이들이 차지하는 비율은 각각 20%, 50%, 30%이다. 또한, 그들의 자료에 의하면 각 부류의 사람들이 1년 동안 사고를 낼 확률은 각각 0.05, 0.15, 0.30이다. 한 보험 가입자가 1년 동안 사고를 일으키지 않았다면, 그가 사고 위험률이 낮은 그룹에 속하는 사람일 확률을 구하여라.

베이즈 정리

- 예제
  - A 산업의 기업 중 5%가 파산한다고 한다. B 은행에서 A 산업 기업들의 파산 가능성을 알아보기 위해서 신용평가기관 평가결과를 조사한 결과, 최근 1년간 파산한 기업 중 불량으로 판정된 기업은 95%이고, 최근 1년간 파산하지 않은 기업 중 90%는 우량으로 판정된 것으로 나타났다. 어떤 기업의 신용평가 결과가 불량으로 판정될 때, 이 기업이 파산될 확률을 구하시오.