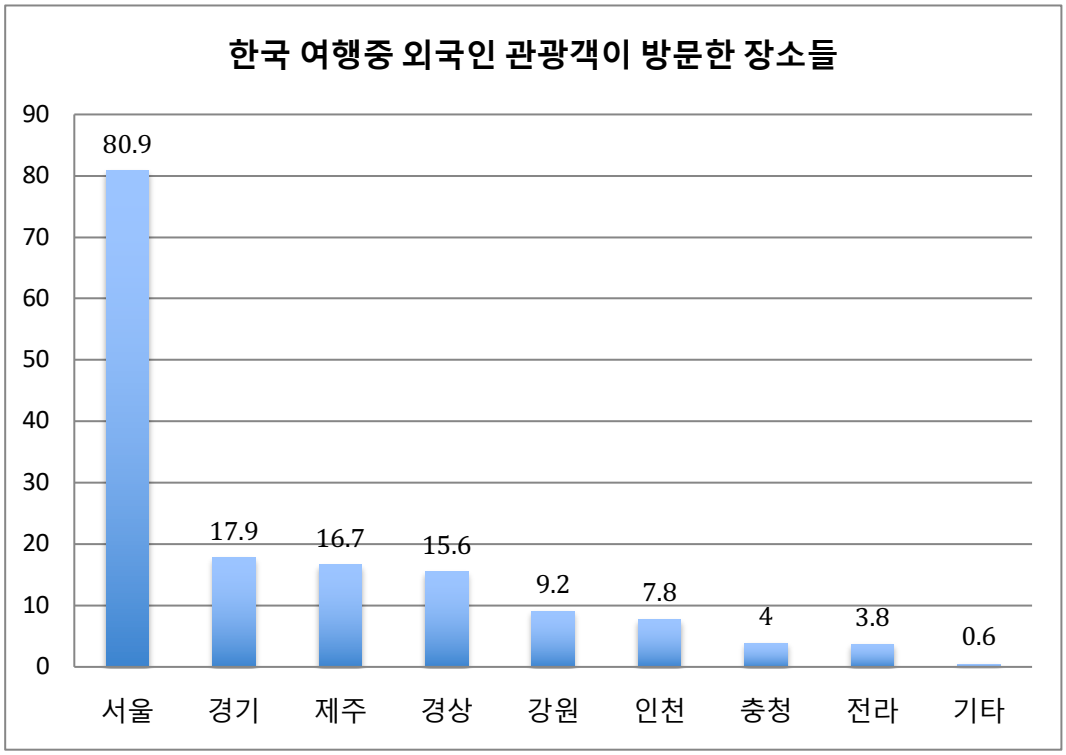
Chapter 4 Probability

한국을 찾는 외국인 수는 계속 증가하고 있으며 특히 중국 관광객 증가는 예전에 가장 많은 일본인 관광객 수를 넘어섰습니다. 하지만 문화체육관광부 2013 년 조사에 따르면 80.9%의 관광객들은 서울을 방문하지만 지방의 경우는 낮게 나타납니다.



제주도의 경우는 중국(35.1%), 싱가포르(21.3%), 말레시아(27%) 관관객들을 제외하고는 방문률이 10%도 안됩니다. 이렇게 지방이 낮은 이유를 보면 방한 목적과 방한 기간 활동에서 쇼핑이 가장 높기 때문입니다. 그래서 인기있는 장소들은 명동, 동대문, 시내 면세점, 백화점과 같은 지역으로 나타나고 만족도에서 치안 4.24 점/5.0 점 다음으로 높은 4.21 점으로 나타납니다. 하지만 가장 불편함을 느끼는 언어 소통은 외국 관광객을 유치하기 위해 중요한 요건으로 작용합니다.

이러한 통계를 위해서 모든 관광객에게 설문조사를 하지 못하므로 설문 조사에 응답한 관광객으로 부터 얻은 정보만으로 모든 외국 관광객의 생각을 읽어야 합니다. 다시 말해,

설문 조사에 응한 사람들(sample)이 모든 관광객(population)의 생각을 잘 대변해 줘야 하는데, 이러한 해석을 돕기위해서 반드시 필요한 것이 확률(probability)와 표본 추출(sampling)방법 입니다. 이번 장에서 부터 6 장까지는 확률에 대한 설명을 하고 7 장에서 표본 추출에 대한 설명을 하도록 하겠습니다.

외국인 관광객의 조사내용처럼 다른 많은 조사 내용을 보시면 많이 사용하는 정보값으로 몇 %라는 확률을 사용합니다. 이 값을 통해서 알지 못했던 사실도 알 수 있게도 하고 앞으로 발생될 사건에 대한 예측을 할 수 있는 값으로 사용됩니다.

확률은 어떤 사건이 발생할 것이라는 예측 값으로 0 과 1 사이에 숫자나 백분율로 0%에서 100%로 나타냅니다. 예를 들어 일기 예보에서 비올 확률 70%와 같이 사용합니다. 그런데, 비가 오지 않을 확률이 30%이기 때문에 비가 반드시 온다고 확실할 수 없습니다. 만일 인간이 신이라면 이런 확률은 존재하지 않고 0%와 100%만 존재 하겠지만 정확하게 알기에는 어려움이 많기 때문에 이와 같이 확신을 확률로 말을 합니다.

이러한 확률값은 듣는 것에 따라 느낌도 달라지고, 확률이 있어도 무시하는 경우도 많이 있습니다. 예를 들어 어떤 질병이 있는데 이 병은 만명 중 한명이 걸린다고 하는 것과 0.01%가 걸릴 확률이 있다고 하면 같은 내용이지만 우리가 느끼기에는 만명 중 한명이 질병에 대해서 더욱 심각하게 듣게 됩니다. 확률을 무시하는 경우로 대표적인 것이 로또 입니다. 로또는 당첨확률이 매우 낮다하더라도 무시하고 로또를 구매합니다.

역사적으로 볼 때 확률의 발달이 된것은 약 1500 년대에 도박에 확률을 적용하면서 였습니다. 그 전에는 확률은 신의 영역이라 생각하여 인간이 다루지 않았다고 합니다. 그러다 보니 확률을 말할 때 가장 많이 사용하는 도박과 관련된 예제가 동전 던저기, 주사위 던지기, 카드놀이가 많이 활용 되고 있습니다.

우리가 확률을 필요로 하는 이유는 random 요소 때문입니다. 설명을 위해 예측으로 random 요소를 설명하겠습니다. X 와 Y 로 구성된 좌표에 점을 하나 찍고 선을 이 점을 통과한다고 하면 방향을 예측하는다는 것은 불가능 합니다. 하지만 선이 지나갈 다른 점이 하나를 아고 있다면 방향을 예측할 수 있습니다.

하지만 전체적인 방향이 직선이면 이 예측은 맞게지만 만약 이 선이 곡선이 된다면 다음 방향에 대해서 예측하는 것은 확률적으로 처음보다는 높지만 정확하게는 맞추지는 못합니다. 우리가 신이라면 100% 확실한 방향을 제시할 수 있겠지만 그렇지 못하기 때문에 알기 어렵습니다. 주식 차트가 대표적인 예입니다. 주식 가격이 직선으로만 움직인다면 모든 사람들은 많은 돈을 벌겠지만, 가격을 결정하는 여러 요인 즉 random 요인들은 다음날 가격을 예측하기 어렵게 만듭니다.

우리가 관찰하고 측정하는 모든 것에 random 요소가 포함됩니다. 방 온도를 측정을 할 때 계측기가 25 도라고 말하지만 이 값은 24.98 ~ 25.02 와 같은 일정 범위 내에서 계속 변합니다. 이런한 것을 noise 때문이라고 말하는데 이것이 random 요소 입니다.

이번 장에서는 확률에 대한 가장 기본적인 이론들을 배울것이고 다음 장부터는 셀 수 있는 데이타로 부터의 확률과 측정값으로 부터 확률을 계산하는 법을 배울 것입니다.

1. 확률 개론

용어	설명
Experiment(실험)	측정할 데이타를 모으는 과정으로 예를 들어 동전던지기, 주사위 던기기, 카드 뽑기와 같은 것을 말합니다.
Sample Space(값영역) 표기: Ω	실험에서 나올 수 있는 모든 값들입니다. 동전은 앞/뒤, 주사위는 1, 2, 3, 4, 5, 6 입니다.
Event(사건)	Sample Space 의 부분으로 실험으로 부터 나타난 값중 어떤 조건을 만족하는 값들로 예를 들어 주사위를 던전 홀수가 나오는 사건에 값은 1,3,5 입니다.

확률(probability)은 다시 classical 확률, empirical 확률, subjective 확률 3 가지로 분류될 수 있습니다.

1.1. Classical Probability

Sample space 를 전부 알고 있을 때 Event 가 발생될 확률로 실험으로 나올 수 있는 값들은 모두 같은 확률이라고 가정합니다.

$$P(A) = \frac{\text{Event A 에 해당하는 개수}}{\text{Sample Space 로 부터 나올 수 있는 개수}} \tag{4.1}$$

예를 들어 한개의 주사위에서 Sample space 는 1,2,3,4,5,6 이고 event 는 주사위를 던저 나온 값이 짝수 경우 2,4,6 이므로 1/2 이 됩니다. 즉 주사위를 던저 짝수가 나올 확률은 50%가 됩니다.

만일 3 개의 동전을 던질 경우 앞면을 H, 뒷면을 T 로 표시하여 나올 수 있는 Sample space 는 HHH, HHT, HTH, HTT, THH, THT, TTH, TTT 로 8 가지가 됩니다. 여기서 event 를

동전 앞면이 한번 나오는 경우로 했다면 HTT, TTH 2 번만 있기 때문에 확률은 2/8 즉 앞면이 한번 나올 확률은 25%가 됩니다.
여기서 중요한 내용은 분모와 분자의 단위가 같기 때문에 probability 는 단위에 의존하지 않습니다. 따라서 확률값은 서로 다른 내용이라도 비교가 가능합니다.

1.2. Empirical Probability

Classical Probability 는 sample space 를 전부 알아서 모든 sample space 를 포함하여 확률을 계산한다고 했지만, 실제로 일상 생활에서 발생하는 일들은 sample space 전체를 알지는 못합니다.
예를 들어 어떤 영화에 대해 선호도를 알고 싶을 때 모든 사람들에게 물어 보지 못하고 영화를 관람한 몇 명을 대표로 하여 영화 선호도가 조사하여 선호도를 조사 할 수 있습니다.
다른 예는 선거날 출구 조사를 통해서 각 후보자의 지지율을 알아 보는 것입니다.
Empirical probability 는 이렇게 실험을 통해서 총 관측된 것중 측정을 원하는 사건이 발생되는 회수를 나눈 값입니다.

$$P(A) = \frac{\text{Event A 에 해당하는 개수}}{\text{총 관찰된 수}} \tag{4.2}$$

예를 들어 식품을 만드는 업체가 신제품의 소비자 반응을 얻기 위해서 길거리에서 시음회를 하고 설문응답자에게 선물을 주겠다고 하였을 때 1000 명의 참가자 중에 신제품이 좋다고 말한 응답자가 800 명이명 다른 사람이 좋다고 할 확률이 0.8 이 됩니다.

다음으로 돈전 던지기와 같이 명확하게 sample space 를 결정 지을 수 있지만 그렇지 않고 10 번 던졌 앞면이 한번 나온것이 2 개 있다 5 개 있다와 같이하여 20%, 50%라고 확률을 말하기도 합니다. 하지만 이러한 경우 동전 던지는 회수가 많아지면 얻으려고 하는 확률값은 classical probability 으로 실제 계산되는 값에 근접하는게 되는데 이러한 방식으로 확률을 값을 측정하는 방식을 Monte Carole 절차라 합니다.

1.3. Subjective Probability

경험과 직감에서 나오는 확률로 예를 들어 주식 가격이 오를 확률이 70%와 같이 미래를 예측하는데 사용되는 확률입니다. 즉 미래를 말하다 보니 관찰은 있을 수 없고 sample space 역시 없기 때문에 사건에 대한 확률을 classical probability 나 empirical probability 로 알 수 없습니다.

1.4. 기본 속성

- 사건 A 가 반드시 발생할 경우 $P(A) = 1$
- 사건 A 가 반드시 발생하지 않을 경우 $P(A) = 0$
- 사건이 발생하였을 때 P(A)의 범위는 0 에서 1 사이 입니다: $0 \leq P(A) \leq 1$
- Sample space 에서 모든 각각의 사건이 발생될 확률의 합은 1 입니다.
- 시건 A 외 다른 모든 사건들 A′을 complement 라고 하고 확률은 $P(A') = 1 - P(A)$

다음은 2012 년 시간대별 개방식 톨게이트에 입구 출구 교통량입니다.

시간대	차량수	확률(%)	오전(A)/오후(A′)
00~04	16,565,878	3.72	P(A) = 0.39
04~08	53,272,145	11.97	
08~12	103,364,881	23.23	
12~16	99,346,012	22.33	P(A') = 0.61
16~20	109,019,962	24.50	
20~24	63,344,273	14.24	

출처: 한국 도로공사(단위 대)

여기서 확률값은 Empirical probability 로 관찰된 정보를 갖고 확률을 측정한 것입니다. 여기서 sample space 는 00~04, 04~08 인 시간대이고, 단독 사건들의 확률을 합하면 1 이 됩니다.

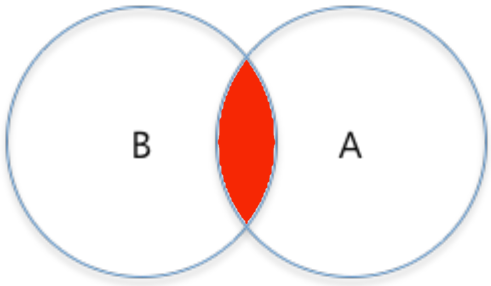
Complement 에 대해서는 만일 사건(A)을 0~12 시까지 확률은 38.9%로 이 사건에 complement(A′)는 12-24 시 P(A′)는 1-0.389 으로 0.61 입니다. 이와 같은 데이타로 총 차량

수만 알면 대략 시간대별 몇 대의 차가 이동하는 지를 예측할 수 있는데 이 때 부터는 확률은 subjective probability 가 됩니다.

2. 여러개의 Event 들에 대한 확률

두개 이상의 숫자가 있으면 일반적으로 수학에서는 +, −,×,÷과 같은 사칙연산을 합니다. Event 에 대한 확률 역시 이러한 연산 작업을 할 수 있게 됩니다.

2.1. And 연산: Intersection (Joint Probability)



두개의 Event 가 동시에 발생 할 확률을 말하는 것으로 예를 들어 고속도로 통행량의 통계를 입구와 출구를 분리를 했을 경우를 보겠습니다. 계산의 편리를 위해서 단위를 백만대로 하겠습니다.

시간대	입구 차량수	출구 차량수	합
00~04	7	9	16
04~08	30	23	53
08~12	54	49	103
12~16	50	50	100
16~20	52	57	109
20~24	29	35	64
합	222	223	445

출처: 한국 도로공사(단위 백만대)

사건 A 로 04~08 시 사이에 입/출구 통과할 차량 확률은 53/445 로 0.12 입니다. 그리고 사건 B 는 입구 차량수의 확률로 222/445 로 약 0.5 가 됩니다. 그럼 두 사건 A 와 B 가 동시에 발생할 확률은 04~08 시 사이에 입구를 통과할 차량 확률은 30/445 로 0.067 이 됩니다. 즉 1000 대의 차가 총 다닌다면 67 대가 04~08 사이에 입구를 통과할 것이라는 예측을 할 수 있습니다.

위의 예와 같이 가로인 시간대별 통행량과 세로인 입/출구 통행량을에 교차지점을 계산하는 것을 두 event 의 intersection 이라고 하고 이 사건의 확률을 joint probability 라고 합니다. 그리고 각각의 두 사건에 대한 확률 P(A)와 P(B)을 marginal probability 라고 합니다. 이러한 관계를 수식으로 다음과 같이 표현 합니다.

$$P(A\ and\ B) = P(A \cap B) \tag{4.3}$$

2.2. Or 연산: Union

두개의 Event 가 모두 발생될 확률을 구하는 것입니다. 예를 들어 통행 차량 예에서 사건 A 를 04~08 시에 입출구 차량수와 사건 B 로 00~12 시 사이에 입구 차량수에 해당하는 확률을 구한다고 한다면 다음의 영역과 같습니다.

시간대	입구 차량수	출구 차량수	합
00~04	7	9	16
04~08	30	23	53
08~12	54	49	103
12~16	50	50	100
16~20	52	57	109
20~24	29	35	64
합	222	223	445

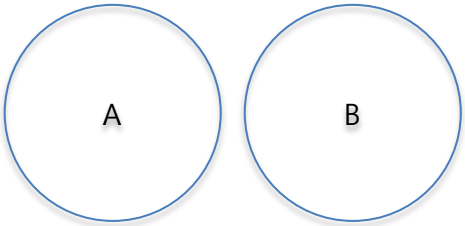
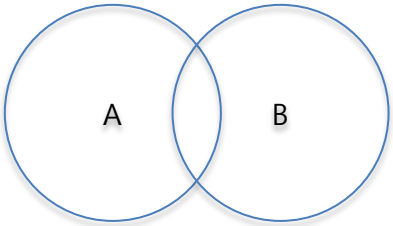
출처: 한국 도로공사(단위 백만대)

확률은 (7+30+54+23)/445 로 0.2562 가 됩니다.

$$P(A\ or\ B) = P(A \cup B) \tag{4.4}$$

Intersection 은 사칙연산 중 곱셈과 유사하고 Or 는 덧셈과 유사합니다. 하지만 단순히 두개의 값을 곱하거나 덧셈을 하면 원하는 결과를 얻지 못합니다. 그럼 이런 연산을 확률에서 어떻게 구현해야 되는지 알아 보겠습니다.

2.3. 덧셈

	
$P(A) + P(B)$: mutually exclusive	$P(A) + P(B) - P(A \cap B)$

두 사건 A, B 에 확률을 합을 할 경우 단순 덧셈이 가능할 때는 이 두 사건에 공통된것이 없어야 합니다. 예를 들어 교통량에서 04~08 시에 입출구 차량수는 16~20 시에 입출구 차량수와 관련이 없기 때문에 두 사건이 발생할 확률은 P(A) + P(B)와 같이 더하기만 하면 됩니다. 이렇게 두 사건에 공통점이 없는 것을 mutually exclusive 되어 있다고 합니다.

하지만 만일 두 사건이 공통된 지점이 있다면 단순히 더 했을 때 문제가 발생합니다. 예를 들어 04~08 시 사이에 입출구 차량에 대한 확률과 00~12 사이에 입구 차량에 대한 확률을 04~08 시에 입구 차량 수와 겹치게 됩니다.

시간대	입구 차량수	출구 차량수	합
00~04	7	9	16
04~08	30	23	53
08~12	54	49	103
12~16	50	50	100
16~20	52	57	109
20~24	29	35	64
합	222	223	445

수식으로 다시 점검을 하면 다음과 같습니다.

	설명	확률
P(A)	04~08 시 입출구 차량	(30+23)/445 = 0.12

P(B)	00~12 시 입구 차량	(7+30+54)/445 = 0.204

만일 단순히 P(A) + P(B)를 하게 되면 0.324 로 앞에서 구했던 0.256 보다 훨씬 큰 값이 나타납니다 이유는 P(A)와 P(B)의 joint probability 가 존재하여 P(A)도 포함되고 P(B)에도 포함되어 있는 부분이 있기 때문입니다. 즉 30 이라는 숫자가 양쪽에 있어서 실제로 P(A)+P(B)의 계산은 (30+23 + 7+30+54)/445 같이 된 것입니다. 따라서 정확한 계산을 위해서 joint probability 를 더한 값에서 한번 빼줘야 합니다. 그럼 결과는 0.324 – 0.067 로 0.257 가 됩니다.

$$P(A\ or\ B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.4}$$

이 수식에서 두 사건이 mutual exclusive 일 경우는 joint probability 가 0 이므로 P(A)+P(B)가 됩니다.

즉, marginal probability 들의 합은 union 보다 크거나 같습니다. 이것을 Boole's inequality 라고 하고 Event 가 A_1 에서 부터 A_n 까지 있을 때 다음과 같습니다.

$$P(A_1 \cup A_2 \cdots \cup A_n) \leq \sum_{i=1}^{n} P(A_i)$$

2.4. 곱셈

곱셈은 Conditional probability 과 독립성에 관한 이해가 필요합니다. Conditional probability 는 확률을 얻기 위한 특별한 조건하에 확률을 보는 것으로 즉 데이타를 보는 관점에 따라서 확률을 달라지게 됩니다. 예를 들어 커피 가게에서 제품별 이벤트전과 후에 판매되는 물량을 측정 했을 경우를 보겠습니다.

제품	아메리카노	라떼	카라멜마키아도	총합
이벤트전	102	75	34	211
이벤트기간	98	163	82	343
이벤트후	104	130	103	337

총합	304	368	219	891

카라멜 마키아도의 판매량에 대한 확률 P(A)는 219/891 로 0.2458 이 됩니다. 그리고 이벤트후 총 판매량에 대한 확률 P(B)는 337/891 로 0.378 이 됩니다. 이벤트 후 카라멜 마키아도 판매량의 확률은 103/891 로 0.1156 이 됩니다.
그럼 이벤트 후 총 판매량에만 보았을 때 이벤트후 카라멜 마키아도 판매량에 대한 확률은 수식 4.5 와 같이 계산이 됩니다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4.5}$$

위의 수식으로 계산을 하게 된다면

$$P(A|B) = \frac{\frac{103}{891}}{\frac{337}{891}} = \frac{103}{337} = 0.3056$$

이벤트 후 카라멜 마키아도의 판매량이 확률인 $P(A \cap B)$를 전체 판매량의 관점에서 보는 것이 아니라 이벤트 후에서 비율을 알아내는 것입니다.

그런데 역으로 카라멜 마키아도의 관점에서 이벤트 후의 카라멜 마키아도의 판매량의 확률은 보면 다른 결과가 나타납니다.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{4.6}$$

수식 4.6 을 적용하게 되면 확률은 103/219 = 0.4703 이 됩니다.

이 결과가 의미하는 것은 어느 관점에서 결과를 해석하는가에 따라서 다른 결과가 나타난다는 것입니다. 이벤트 후의 사건으로만 보았을 때 카라멜 마키아도의 비율을 약 30%를 차지하지만 카라멜 마키아도만 보았을 때는 이벤트 후 판매량이 전체 카라멜 마키아도 판매량의 약 47%로 높게 나타납니다.

다른 두 커피음료들에 대해 같은 방식을 적용해서 보면 아메리카노의 관점에서 이벤트 후의 판매량은 104/304 로 약 34%, 라떼의 경우 130/368 로 약 35%로 카라멜 마키아도의 이벤트 후 판매량 비중 증가 보다 낮게 나타나는 것을 알 수 있습니다.

이와같이확률을 sample space 전체에서 보고 계산하는 것이 아니라 어떤 event 안에서 해당 event 가 발생될 확률을 얻는것을 conditional probability 이라고 합니다. 이것은 데이타를 보는 관점에 따라서 해석이 달라질 수 있기 때문에 이 확률을 통해서 여러 방면으로 데이타를 해석 할 수 있습니다. 다음은 conditional probability 가 어떻게 자료 해석에 영향을 미칠 수 있는가를 보여주는 예제 입니다.

2008 년 근로 복지 공단은 S 전자 반도체 공장에서 일어난 백혈병은 산업 재해로 볼 수 없다고 결론을 내면서 근거 자료로 산업안전보건연구원의 역할 조사를 제출 했습니다.

"2007 년까지 지난 10 년간 전체 반도체 산업 종사자 22 만 9683 명을 대상으로 조사한 결과 여성의 경우 암 발병 비율이 1.31 배 높을 뿐이어서 통계적 의미가 없고 남성은 오히려 일반인보다 낮은 수준을 보이고 있다"

이 내용에 맞도록 가상의 자료를 만들면 다음과 같습니다.

	여성 암환자수	정상인 수	합계
A: 3 라인 공정	8	497	500
B: 다른 반도체 공정	280	99,700	100,000
C: 반도체 외 일반인 sample	220	99,795	100,000
합계	508	199,992	200,500

국가암정보센터(http://www.cancer.go.kr/mbs/cancer/)의 암 발생률 조사에 따르면 2007 년 여성 연령표준발생률은 10 만명당 253.4 명을 근거로 총 조사 대상 200,500 명을 대상으로 했을 때 508 명이 암 발생환자가 됩니다. 반도체 산업 여성 종사자의 암 발병 비율이 일반인

보다 1.31 배 높다는 주장으로 508 명을 일반인(x)과 반도체 공정에 종사(1.31x)하는 사람을 분류를 하면

$$508 = 1.31x + x$$
$$x = 220$$

반도체 공정에서 일하는 여성의 총 암 환자수는 288 명이 됩니다. 관점에 따른 여성 암 환자수의 비율은 다음과 같습니다.

	전체 암환자관점	각 group 별 관점
A: 3 라인 공정	1.57%	1.6%
B: 다른 반도체 공정	55.12%	0.28%
C: 반도체 외 일반인 sample	43.31%	0.22%
합	100%	0.2534%

전체 암환자 관점에서 보면 A 에 암환자 비율은 매우 낮게 나타납니다. 하지만 각 group 별로 암 발생률을 보면 A 에 암 발생 비율이 일반 발생률이 6.3 배나 높게 나타납니다. 만일 A 에 암 발생률이 다른 반도체 공정에 사람과 동일한 비율로 암이 발생된다면 1.43 명이 되어야 합니다. 이것만 보도라도 A 에 암 발생률의 심각함을 알 수 있습니다. 비록 가상 데이타이지만 만일 이렇다면 매우 3 라인 공정은 매우 심각합니다.

Conditional probability 을 계산하기 위해서 우선 전체를 보았을 때와 조건속에서 보았을 때의 차이점을 알아야 합니다. P(A|B)에서 P(A)값은 조건이 없이 전체로 보았을 때의 확률이고 이것을 prior probability 라고 합니다. 그리고 P(B)라는 조건하에서 P(A)를 구하는 P(A|B)를 posterior probability 라고 합니다.

Conditional probability 로 부터 확률의 곱셈방법을 알 수 있습니다.

$$P(A\ and\ B) = P(A \cap B) = P(B)P(A|B) = P(A)P(B|A) \tag{4.7}$$

Or 와 같이 단순히 P(A)P(B)가 아니라 conditional probability 를 사용해야만 원하는 결과를 얻을 수 있음을 알 수 있습니다. Or 에서 mutual exclusive 처럼 intersection 으로 0 으로 만들어 그냥 두개의 marginal probability 를 단순히 합하게 만드는 것 처럼 두 가지 event 가 서로 독립되어 있다면 P(B|A)는 P(B)가 되고 P(A|B)는 P(A)가 됩니다. 그래서 P(A and B)는 P(A)P(B)와 같이 됩니다. 이와 같은 방식으로 우리는 두 사건의 독립성 테스트를

할 수 있습니다. 예를 들어 이벤트로 인한 커피 판매량에 대한 이벤트와 커피종류의 관계는 독립적이지 않습니다. 이것을 확인해 보도록 하겠습니다.

카라멜 마키아도의 판매량에 대한 확률 P(A)와 이벤트후 총 판매량에 대한 확률 P(B)라고 했을 때

P(A) = 0.2458
P(B) = 0.378
P(A)P(B) = 0.0929
P(A and B) = P(A|B)P(B) = 103/891 = 0.1156

두개의 사건이 독립이 경우로 식당에서 판매하는 두 제품다 5000 원인데 둘 다 500 원 올렸을 때 하루 판매량을 측정한 결과 입니다.

	제품 A	제품 B	합(확률)
가격 인상전	20(0.267)	30(0.4)	50(0.667)
가격 인상후	10(0.133)	15(0.2)	25(0.333)
합(확률)	30(0.4)	45(0.6)	75

보시는 것과 같이 가격 인상이 10%증가로 판매량이 반으로 줄었습니다. 그럼 제품과 가격인상에 대한 의존성 측정을 위해 사건 A 를 제품 B 판매량, 사건 B 를 가격 인상후 판매량 이라 했을 때 두 사건의 marginal probability 값들은

$$P(A) = \frac{45}{75} = 0.6, \;\; P(B) = \frac{25}{75} = 0.33$$

$$P(A \; and \; B) = \frac{15}{75} = 0.2$$

P(A)P(B)를 보면 0.6x0.33 으로 0.2 가 됩니다. 이와 같이 다른 사건들도 보시면 joint probability 는 간단하게 해당 marginal probability 들의 곱인것을 아실 수 있습니다.

또한 P(A|B)=P(A)와 같고 P(B|A)=P(B)와 같음을 알 수 있습니다. 위의 예제에서 P(A|B)는 가격 인상 후에서만 본 제품 B 의 판매량의 확률은 15/25 로 P(A)인 0.6 과 같습니다. 즉 가격 인상 후에서 본 판매량의 비중은 제품의 전체 판매량의 비중과 같다는 것으로 가격 변동이

특정 제품에 판매량 비중에 관계가 없으므로 고객이 선호하는 음식이 없다는 것을 알 수 있습니다.

일반적으로 독립적인 사건이란 서로가 영향을 미치지 않는 것으로 주사위를 던질때 첫번재 던지 주사위와 두번째 던지 주사위는 서로 영향을 미치지 않습니다. 그래서 만일 두개의 주사위를 던졌을 때 첫번째 3 이 나올 확률 1/6 과 두번째 6 이 나올 확률이 1/6 이면 첫번째 3 이 나오고 두번째 6 이 나올 확률은 단순히 두 값을 곱하기만 하면 되어 1/36 이 됩니다.

확률에서 곱셈은 중요합니다. 특히 모든 시스템은 원인과 결과로 값을 생성하기 때문에 최초 원인이 발생하고 결과가 나올 확률을 P(0)라고 한다면 두번째 발생되는 확률은 반드시 첫번째 나타난 값의 영향으로 나타나므로 이 값이 나타날 확률 P(0 and 1)은 P(0)P(1|0)과 같습니다. 그래서 여러 시간이 지난 후 확률값은 P(0)P(1|0)P(2|1) ...과 같이 됩니다. 그런데 동전 던지기나 주사위 던지기 처럼 전 결과에 상관없이 값이 발생되는 것은 상호 독립이기 때문에 P(0)P(1)P(2)..와 같습니다.

마지막으로 mutually exclusive 와 independent 의 차이점에 대해서 다시 알아보겠습니다. 두 사건에 대해서 mutually exclusive 한 것은 공통된 것이 없는 것이고 independent 는 서로 연관이 없다는 점에서 다릅니다. 하지만 두 사건이 동시에 mutually exclusive 하고 independent 하지는 못합니다. 이것은 수식으로 본다면 간단합니다. 예를 들어 식당에 제품들의 가격 상승후 판매량에서 가격 상승전 사건을 P(A)라하고 가격 상승 후 사건을 P(B)라고 했을 때 두 사건은 동시에 존재 할 수 없기 때문에 mutually exclusive 입니다. 하지만 independent 를 위해서 P(A and B)는 반드시 P(A)P(B)와 같아야 되는데 공통점이 없기 때문에 P(A and B)는 0 이고 P(A)는 0.667, P(B)는 0.337 로 이를 곱하면 0 이 나오지 않습니다. 따라서 독립적이라고 말할 수 없습니다.

jMath

커피 가계에서 제품별 이벤트전과 후에 판매되는 물량에 대한 jMath 를 활용하는 방법은 소개하겠습니다.

제품	아메리카노	라떼	카라멜마키아도
이벤트전	102	75	34

이벤트기간	98	163	82
이벤트후	104	130	103

각 제품별 이벤트 전/기간/후에 대한 확률을 계산하기 위해서는 column 별 합을 계산을 한 후에 element 에 해당 column 합을 다음과 같이 나누어야 합니다.

$$\begin{bmatrix}102 & 75 & 34\\ 98 & 163 & 82\\ 104 & 130 & 103\end{bmatrix} \cdot/ \begin{bmatrix}304 & 368 & 219\\ 304 & 368 & 219\\ 304 & 368 & 219\end{bmatrix}$$

여기서 "./"는 두 Matrix 를 element 별로 나눈다는 뜻입니다.

이러한 계산을 하기 위한 code 는 다음과 같습니다.

```
 > var data = jMath('102 75 34; 98 163 82; 104 130 103');
 > console.log( data['./']( data.sum(1).repeatRow(2) ).toString() );

0.3355263157894737  0.20380434782608695 0.1552511415525114
0.3223684210526316  0.4429347826086957  0.3744292237442922
0.34210526315789475 0.3532608695652174  0.4703196347031963
```

변수 data 는 위의 테이블을 그대로 Matrix 로 표현한 값으로 column 별 합을 계산하기 위해서 sum 함수를 호출할 때 방향 입력값 1 로 했습니다. 결과는

```
> data.sum(1).toString()
[304   368   219]
```

이 값을 모든 element 에 해당 column 의 합으로 나뉘어야 하는데 이를 위해서 필요한 함수가 jMath.prototype.repeatRow(n)입니다. 이 함수는 matrix 를 row 방향으로 n 번 반복하는 함수입니다.

$$[304 \quad 368 \quad 219] \rightarrow \text{repeatRow(2)} \rightarrow \begin{bmatrix}304 & 368 & 219\\ 304 & 368 & 219\\ 304 & 368 & 219\end{bmatrix}$$

다음 이벤트 전/기간/후에 관점에서 확률 계산은 row 별 합을 구하고 element 에 해당 row 합을 다음과 같이 나누면 됩니다.

$$\begin{bmatrix}102 & 75 & 34\\ 98 & 163 & 82\\ 104 & 130 & 103\end{bmatrix} \cdot/ \begin{bmatrix}211 & 211 & 211\\ 343 & 343 & 343\\ 337 & 337 & 337\end{bmatrix}$$

이러한 계산을 하기 위한 code 는 다음과 같습니다.

```
 > var data = jMath('102 75 34; 98 163 82; 104 130 103');
```

```
> console.log( data['./']( data.sum(2).repeatColum(2) ).toString() );

0.4834123222748815  0.35545023696682465 0.16113744075829384
0.2857142857142857  0.4752186588921283  0.239067055393586
0.3086053412462908  0.3857566765578635  0.3056379821958457
```

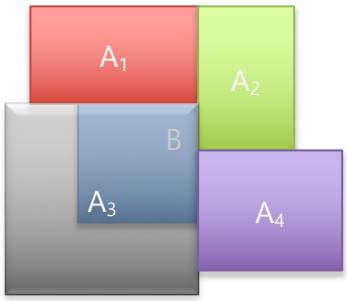
여기서 jMath.prototype. repeatColumn(n)은 matrix 를 column 방향으로 n 번 반복을 합니다.

$$\begin{bmatrix}211\\343\\337\end{bmatrix} \rightarrow \text{repeatColumn(2)} \rightarrow \begin{bmatrix}211 & 211 & 211\\343 & 343 & 343\\337 & 337 & 337\end{bmatrix}$$

3. Bayes' Theorem

Conditional probability 로 보는 관점에 따른 확률의 변화를 보았습니다. 보는 관점을 뒤집었을 때 확률을 얻는 Bayes' Theorem 을 소개 하겠습니다.

$$P(A_m|\,B) = \frac{P(A_m \cap B)}{P(B)} = \frac{P(A_m)P(B|A_m)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \tag{4.8}$$



수식을 그림으로 설명을 하겠습니다. Event B 는 Event A1, A2, A3, A4 의 교집합의 합과 같습니다. 그래서

$$\begin{aligned} P(B) = {} & P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) + P(A_4 \cap B) \\ & = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) + P(A_4)P(B|A_4) \\ & = \sum_{i=1}^{4} P(A_i)P(B|A_i) \end{aligned}$$

수식 4.8 이 의미하는 것은 B 를 형성하는 A_i 의 B 에 차이하는 비율을 얻는 것입니다. 그림에서 A_3 가 B 에서 찾이하는 비율이 가장 크게 나타납니다.

수식 4.8 로 부터 용어를 설명하면 $P(A_m)$는 prior probability, $P(B|A_m)$는 likelihood, 마지막으로 $P(A_m|B)$는 posterior 입니다. Bayes' Theorem 을 통해서 데이타를 이해하는 방법에 중요성을 알게 됩니다. 예를 들어 병원에서 질병을 진단하는 새로운 방법론을 만들고 제대로 판단하는가를 알아 보려고 할 때 다음과 같이 결과를 얻었습니다.

	질병있음	질병없음	
Positive	300 명	1000 명	1300 명
Negative	10 명	300 명	310 명
	310 명	1300 명	1610 명

여기서 positive 는 양성으로 질병이 있다고 진단을 내린 것이고 negative 는 음성으로 질병이 없다고 내린 것입니다. 검사를 위해서 미리 알고 있는 정보(Prior probability)는 실제 질병이 있는 사람과 질병이 없는 사람에 대한 정보이고 질병이 있는 사람들 관점에서 양성 반응이 나타난 정보와 질병이 없는 사람이 양상 반응이 나타난 정보(likelihood)를 알아 낼 수 있습니다. 이 결과로 부터 알고 싶은 내용은 실험을 해서 얻아야만 하는 양성 반응을 나타낸 사람들에 관점에서 실제 질병이 있을 확률인 posterior probabilty 를 얻는 것입니다.

D_+는 양성 반응 D_-는 음성 반응, T_+질병이 있을 경우, T_-질병이 없는 경우로 이들의 확률을 보시면 다음과 같습니다. 먼저 prior prababilty 를 보면 실험전에 실제 질병이 있는 환자와 없는 환자를 구분할 수 있기 때문에 확률을 계산할 수 있습니다.

$$P(T_+) = \frac{310}{1610} = 0.1925, \quad P(T_-) = \frac{1300}{1610} = 0.8075$$

likelihood 를 조사해 보면 질병이 있는 환자가 양성 반응을 나타낼 경우는

$$P(D_+|T_+) = \frac{300}{310} = 0.9677$$

이것만 보면 진단 결과가 잘 나타난것 처럼 보입니다. 하지만 질병이 없는 사람들이 양성 반응을 얻는 경우를 보면

$$P(D_+|T_-) = \frac{1000}{1300} = 0.7692$$

높게 나타납니다.

이 결과를 같고 양성 반응을 보인 사람들 중에 실제 질병이 있을 확률의 계산 공식은 다음과 같습니다.

$$P(T_+|\, D_+) = \frac{P(T_+)P(D_+|T_+)}{P(T_+)P(D_+|T_+) + \ P(T_-)P(D_+|T_-)}$$

이 수식을 적용하여 계산하면

$$P(T_+|D_+) = \frac{0.1925 \times 0.9677}{0.1925 \times 0.9677 + 0.8075 \times 0.7692} = 0.231$$

즉 0.231 의 정확성은 새로운 진단 방법이 효과적이라고 말할 수 없습니다. 결과 만일 테스트를 한다고 질병있는 환자만 새로운 진단법으로 하여 결과를 내렸다면 심각한 문제를 알지 못하고 오진을 할 확률이 매우 높습니다.

이 계산의 의미는 D_+인 양성으로 진단된 사람들 중 진짜 질병이 있는 사람의 비율로 그림으로 표시한다면

왼쪽의 300 은 양성반응을 받은 질병이 있는 사람의 수, 오른쪽의 1000 은 양성반응을 받은 질병이 없는 사람의 수로 이 두 수의 합이 양성 반응으로 나타난 총 사람의 수이고 그 중 질병이 있는 사람 비율은 300/13000 으로 0.231 이 됩니다.

4. 개수 세기

확률에서 대부분은 어떤 특정 사건(event)가 발생 될 수 있는 총개를 세는 것입니다. 간단한 개수 세는 방법으로 simple event 들이 서로에게 영향이 없이 이루어 질때 각각의 simple event 가 발생될 수 있는 수를 곱하면 됩니다. 예를 들어 라면을 판매를 하는데 3 가지 종류의 라면에 치즈, 떡, 만두, 해산물 중 하나를 추가하고 크기를 일반, 곱배기로 하면 총 만들수 있는 메뉴의 개수는 3x4x2 로 24 가지가 됩니다. 하지만 이렇게 총 개수를 세는 방법은 상황에 따라 달라집니다.

4.1. Permutations

숫자 1,2,3 을 한번만 사용해서 만들수 있는 두자리 숫자의 개수는

12, 13, 21, 23, 31, 32

총 여섯가지의 경우가 가능합니다. 다른 예로 1 부터 100 까지 적혀 있는 공을 5 개를 뽑을 때 한번 뽑은 공을 다시 넣지 않는다면 나올 수 있는 경우의 수는 처음에는 100 개, 다음은 99 개, 다음은 98 개과 같이 점차 각 simple event 에 총 경우의 수는 한개 씩 줄어들게 됩니다. 이러한 방식으로 숫자를 세는 것을 permutations 이라고 합니다.

$$nPx = \frac{n!}{(n-x)!} \tag{4.9}$$

여기서 n!은 n x (n-1) x (n-2) … x 2 x 1 과 같습니다.

3 자리 숫자로 2 자리 숫자를 만드는 경우는 3!/1!로 3x2 인 6 이 됩니다. 100 개의 공으로부터 5 개씩 뽑는 경우는 100!/95!가 됩니다. 즉, 100x99x98x97x96 가지의 다른 조합의 공을 빼 낼 수 있습니다. 4.9 의 수식을 다른 방식으로 해석을 하면 100!에서 불필요한 95!를 제거하여 하는 것입니다.

기획사들은 유명 가수들의 음반 출시일이나 영화 개봉날짜가 겹치게 되면 전쟁터에서 싸우는 것과 같은 비유를 합니다. 만일 이 싸움에서 지게 되면 매출에 타격이 크기 때문에 너무 어려운 상대와 같은 날짜에 출시하는 것은 피하는게 좋을 것입니다. 그럼 음반사들이 총 n 개의 음반을 1 월에 출시하려고 할때 날짜가 겹치지 않을 확률을 구해 보겠습니다.
1 월은 총 31 일이 있어 n 개의 음반이 출시 될 수 있는 날짜의 조합은 31^n 개입니다. 예를 들어 2 개의 음반을 출시하는 모든 경우는 한 음반은 1 일날 출시되고 다른 음반은 1~31 일까지로 31 가지의 경우가 있고, 다른 경우로 한 음반이 2 일날 출시되면 다른 음반은 1~31 일까지 다시 31 가지의 경우가 있어, 이를 총 합하면 31x31 가지가 됩니다.

그럼 n 개의 음반이 모두 다른날 출시 되기 위해서는 첫 번째 음반은 31 일 중 아무때나 가능하고 두 번째 음반은 첫 번째 음반 출시 날짜를 뺀 30 일 중 아무 때나 가능합니다. 그리고 다음 음반은 29 일 중 아무때나 가능합니다. 이러한 방식으로 확률을 구하면 결과는 다음 수식과 같습니다.

$$\frac{{}_{31}P_n}{31^n}$$

음반 개수	확률
5	0.7122
10	0.1964
15	0.0167

여기서 중요한 사실중 하나는 만일 음반 A,B,C,D,E 가 있는데 날짜만 보았을 때 5 개의 출시되는 날짜는 동일한데 순서가 A,B,C,D,E 와 E,D,C,B,A 와 같이 있으면 다른 것으로 판단 계산된 결과 입니다. 이러한 순서가 중요하지 않는 경우는 다음에 배울 Combination 을 통해서 처리할 수 있습니다.

jMath 는 permutation 을 계산하는 함수는 없고 jMath.prototype.factorial(x)만 존재를 합니다. 음반 예제에서 31 일 동안 5 개가 모두 다른 날에 출시가 될 확률을 jMath 를 이용하여 계산을 하면

```
> jMath.factorial(31)/jMath.factorial(31-5)/Math.pow(31,5)
0.7121873785219827
```

4.2. Combinations

Permutations 는 원하는 개수만 뽑아 전체가 나올 수 있는 개수를 세는 것과 같지만 combination 과 다른점은 선택되는 순서가 중요하다는 것입니다.

숫자 1,2,3 에서 한번만 사용해서 두자리 숫자를 만드는데 숫자의 순서가 중요하지 않는다면 permuation 에서 얻은 숫자에서 절반이 빠지게 됩니다.

12, 13, ~~21~~, 23, ~~31, 32~~

다른 예를 들어서 로또의 경우 45 개의 숫자가 적힌 공을 6 번을 뽑지만 나온 순서는 중요하지 않습니다. 이것을 수식으로 다음과 같이 표시 합니다.

$$nCx = \binom{n}{x} = \frac{n!}{(n-x)!\,x!} = \frac{nPx}{x!} \tag{4.10}$$

여기서 분모의 x!는 순서로 인한 영향을 제거해 줍니다.
숫자 1,2,3 의 예를 적용하면 3!/(1!2!) = (3x2)/2 로 3 이 됩니다. 즉 분모값 2 는 순서에 의한 중복된 개수를 제거하기 위한 값이 됩니다.
로또의 예를 적용하면 45!/(39!6!)으로 8,145,060 경우가 나타납니다. 이로 부터 1 등 당첨될 확률을 구하면 0.00000012277 로 매우 작음을 알 수 있습니다.

jMath.prototype.nchoosek(n,x)로 combination 을 계산할 수 있습니다. 로또예를 적용하면

```
> jMath.nchoosek(45,6)
8145060
```

Combination 을 보시면 다음과 같은 특성이 있습니다.

$$\binom{n}{0} = \binom{n}{n} = \frac{n!}{n!\,0!} = 1$$

$$\sum_{x=0}^{n} \binom{n}{x} = 2^n$$

예를 들어 3 개의 요소가 있으면 다음과 같은 조합이 가능합니다.

0	0	0	${}_3C_0$
0	0	1	${}_3C_1$
0	1	0	
1	0	0	
0	1	1	${}_3C_2$
1	0	1	
1	1	0	
1	1	1	${}_3C_3$

0 은 없는 경우이고 1 은 존재할 경우로 한개도 없는 상황은 ${}_3C_0$ 개가 있고 1 개는 ${}_3C_1$, 2 개는 ${}_3C_2$, 3 개는 ${}_3C_3$ 과 같습니다. 즉 모든 조합의 합은 8 개로 2^3 과 같습니다.

Combination 의 다른 특성으로

$$\binom{n}{x} = \binom{n-1}{x} + \binom{n-1}{x-1} \tag{4.11}$$

이 수식에 대한 해석을 로또로 해석을 하면 첫번째 수식은 45 개의 공들 중에서 한개는 무조건 선택하지 않겠다고 하고 6 개를 뽑는 경우의 수이고 두번째 수식은 45 개의 공들 중에서 한개는 무조건 선택이 되어 있고 나머지 5 개를 뽑는 경우 입니다. 즉 한개의 공은 선택되었을 때와 그렇지 않을 경우의 합이 되므로 총 45 개중 6 개를 뽑는 경우의 수와 같게 됩니다.

$$\binom{n}{x} = \frac{n}{x}\binom{n-1}{x-1} = \binom{n}{n-x} \tag{4.12}$$

n!를 계산 하는 방식은 gamma function 으로 구할 수 있습니다.

$$(x-1)! = \Gamma(x) = \int_0^\infty e^{-t}t^{x-1}\,dt \tag{4.13}$$

마지막으로 n 개로 부터 x_1, x_2, x_3, x_4 씩 각각 선출될 개수는 다음과 같습니다. 여기서 $x_1+x_2+x_3+x_4=n$ 일 경우

$$\binom{n}{x_1}\binom{n-x_1}{x_2}\binom{n-x_1-x_2}{x_3}\binom{n-x_1-x_2-x_3}{x_4}$$
$$= \frac{n!}{x_1!\,(n-x_1)!}\frac{(n-x_1)!}{x_2!\,(n-x_1-x_2)!}\frac{(n-x_1-x_2)!}{x_3!\,(n-x_1-x_2-x_3)!}\frac{x_4!}{x_4!\,0!} = \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!}$$

예를 들어 20 명이 수용 인원이 6 명, 7 명, 2 명, 5 명인 4 개의 방으로 나뉘어 질 경우 총 경우의 수는

$$\frac{20!}{6!\,7!\,2!\,5!} \approx 2.7935 \times 10^9$$

4.3 조합과 비즈니스

A~Z, 0 ~ 9, 공백으로 문장을 만들 경우의 수는 같은 글자가 반복되도 상관이 없고 갖 글자 선택은 전에 어떤것을 선택했는가에 상관이 없는 즉 독립적이기 때문에 모든 위치에 선택될 수 있는 총 글자의 개수는 37 개가 됩니다. 이러한 조합으로 100 페이지를 글자로 채운다고 했을 때 한 페이지단 200 자씩 입력이 된다면 총 만들 수 있는 책의의 개수는

$$37^{100\times200}$$

이 많은 책 중에는 반드시 Novel 문학상을 받을 수 있는 책도 있을 것이고, 재미 있는 소설도 있을 것입니다. 하지만 문제는 이 수 많은 책에서 어떻게 이러한 책을 찾는가 입니다. 모든 책을 하나씩 읽는다면 인간이 죽을 때 까지 다 검색하지는 못할 것입니다. 그래서 알고리즘이 필요하고 가장 좋은 방법은 진화를 이용한 방식입니다.

이러한 적용은 비즈니스 모델에서도 찾아 볼 수 있습니다. 예를 들어 음식점을 개업을 했을 경우 홍보(A,B,C 방식), 직원관리 방식(A,B,C 방식), 가계 운영 방식(A,B,C 방식)이 있을 경우 가능한 조합은 3x3x3 으로 총 27 가지가 됩니다. 이 중에서 가장 최상의 방식을 찾아서

운영을 하면 성공 확률이 높게 되는데 이러한 것을 찾는 것 역시 하나하나 확인하면서 하기에는 시간과 비용이 많이 발생하기 때문에 책 예제와 같은 알고리즘이 필요합니다.