1

확률 probability

1. 개념

- 확률은 미래에 발생할 사건에 대한 믿음에 대한 측정값이다.
- 확률 개념은 물리, 화학, 사회 과학 등에서 발생하는 관심 현상의 측정값을 불확실성에 의해 예측할 수 없는 경우 사용 (예) 1분간 당신 맥박 수, 다리가 무너지기 전 최대 하중 등
- 이런 상황을 랜덤(RANDOM)이라 한다. 랜덤상황이더라도 이 사건에 대한 상대 빈도(relative frequency) 정보가 있다면 예측이 가능

2. 정의

- 확률은 관심 사건이 일어날 가능성 (chance or likelihood)을 숫자로 표현한 것
- 확률의 0 (일어날 가능성 없음)과 1(항상 일어남) 사이의 값(0% TO 100%)

3. 확률 측정 Probability measure

1)상대 빈도 relative frequency

동전을 던지는 경우 {앞 면이 나올 사건}에 관심이 있어 실험을 한다고 하자. 10번을 던지니 6번이 앞 면이었다면 상대빈도는 0.6이다. 계속 100번 던지니 52번이 앞 면이 나왔다면 상대 빈도는 0.52이다. 1000번을 던지니 515번이 앞면이었다면 상대 빈도는 0.515이다.

- $P(A) = \lim_{n\to\infty} \frac{f}{n}$: 관심사건 A 확률, n=실험 횟수, f=사건 A가 발생한 횟수
- 무한히 많은 시행 후에는 관심 사건의 나타날 가능성을 예상-확률은 무한 실험 후에 관심사건의 횟수

동전 던지기 역사

- Count Buffon (1707-1788): 4040번 동전 던지기 실험 앞면 2048회, P(앞면)= 0.5069
- Karl Pearson (1900): 24, 000 던지기 앞면 12,012, P(앞면)=0.5005
- John Kerrich : 10,000 던지기, 앞면 5067 heads, P(앞면)=0.5067.

- 상대 빈도 개념의 확률을 정의 예 : 공정 생산 제품의 불량률(확률)에 대한 모형을 위해서는 제품 검사(확률실험)를 통하여 검사 제품 개수 중 불량품의 개수(상대 빈도)를 계산하면 된다.

2)Laplace 확률

- 표본공간의 각 원소들이 일어날 가능성이 같다고(equally likely) 가정하여 확률을 정의하는 것을 Laplace 확률(고전적 정의)이라 한다.
- 주사위를 던지는 실험에서 짝수가 나올 확률은 3/6=1/2으로 정의한다. 표본공간의 원소 개수는 6개이고 짝수 사건의 원소는 3개이므로 짝수가 발생할 확률은 0.5이다.
- 고전적 정의의 가정은 각 원소(주사위 눈금)가 나타날 확률과 동일(1/6)하다는 것이다.
- 일반적으로 Laplace 확률 정의는 표본공간의 개수가 유한이고 원소 모두를 알고 있을 때 사용하는데 대부분 이 정의에 의해 확률 모형이 정의된다.

4.확률 공리 Axiom

표본공간 S인 실험에서 임의의 사건 A에 대해 아래 조건을 만족하는 P(A)를 A의 확률(probability)이라고 정의하고 이를 확률의 공리(axiom)라 한다.

○ 공리1: P(A)≥0 모든 사건의 확률 값은 0보다 크거나 같다.

○ 공리2: P(S)=1 표본공간의 확률 값은 1이다.

○ 공리3: 만약 가 상호 배반적 사건이라면( ),

확률을 어떻게 정의하더라도 위의 조건만 만족하면 확률로 정의될 수 있다. 예를 들어 동전을 던지는 실험에서 앞이 나오는 사건의 확률을 2/3, 뒤가 나오는 사건의 확률을 1/3로 정의하면 이것 역시 확률로 정의될 수 있다. 그러나 이런 확률의 정의는 객관성이 없다. 이런 확률 정의를 주관적 확률(subjective)이라 한다.

5. 확률 계산

이산형(표본공간의 원소의 수가 한정적일 경우, 헤아릴 수 있는 경우) 확률실험에서 사건 A에 대한 확률 계산 방법으로 sample-point 방법이 있다.

연속형인 경우에는 관심 구간의 적분 값

sample-point 방법에 의해 사건 A의 확률을 정의하는 순서

I) 실험을 정의하고 표본공간을 정의한다. 표본공간의 원소 개수를 카운트한다. N개

II) 표본공간의 각 원소에 확률 공리를 만족하도록 하여 적절한 확률 값을 할당한다. 라플라스 동등성 가정에 의해 각 원소의 활당 확률은 $1/N$

III) 사건A에 대한 원소를 정의한다. 원소의 개수를 n이라 하자.

IV) 사건 A의 원소들의 확률을 더해 사건 A의 확률로 정의한다. $P(A) = \frac{n}{N}$

본 강의의 예제문제는 Keller 9판 “Statistics for Business and economics” 예제를 활용하였음

5대의 컴퓨터 중 2대가 고장이다. 2대를 임의 선택했을 때 모두 고장이 없을 확률을 구하시오

1) 곱의 법칙

r번의 실험, 각 실험의 결과 수가 $n_1, n_2, \ldots, n_r$이라 하면 실험 전체 결과 수는 $n_1 * n_2 * \ldots * n_r$

컴퓨터 비번을 만들려고 한다. 총 7자리 중 첫 2자리는 소문자, 3번째는 대문자, 남은 4자리는 0~9까지 숫자로 나열하여 만들 때 총 비번 개수는?

2) 반복있고 순서 고려

실험의 결과 수가 n이고 실험을 r번 반복할 때 나타날 수 있는 결과(경우) 수 = n^r

4자리로 숫자로 구성된 비밀번호을 잊었다. 반복이 가능하다면, 최악의 경우 몇 번이나 시도해야 맞출 수 있나?

3) 반복없이 나열

실험의 결과 수가 n이고 이를 나열하여 만들 수 있는 총 결과 수 = $n! = n*(n-1)*\ldots*2*1$

4자리로 숫자로 구성된 비밀번호을 잊었다. 첫 번호에 0을 사용할 수 없고 반복이 불가능하다면, 최악의 경우 몇 번이나 시도해야 맞출 수 있나?

4) 반복없고 순서 고려

n개의 서로 다른 원소들 중 r개를 뽑아 순서대로 배열할 경우 발생하는 총 결과 수 $nPr = \frac{n!}{(n-r)!}$

4자리로 숫자로 구성된 비밀번호을 잊었다. 반복이 불가능하다면, 최악의 경우 몇 번이나 시도해야 맞출 수 있나?

주머니에 6개의 칩이 있거 칩에는 E, E, P, P, P, R이 각각 적혀있다. (1) 복원 추출(with replacement) (2) 비복원 추출(without replacement)로 하나씩 차례로 6개를 뽑아 영어 단어 PEPPER를 만들 확률을 구하시오.

3명의 여자와 3명의 남자가 일렬의 의자에 무작위로 앉는다. (1) 여자 3명이 나란히 앉을 확률을 구하시오. (2) 남녀가 번갈아 앉을 확률을 구하시오. 만약 원탁에 앉는다면 확률은 어떻게 되는가?

5) 반복, 순서 모두 없는 경우

n개의 서로 다른 원소들 중 r개를 뽑아 순서없이 배열할 경우 발생하는 총 결과 수 $nCr = \frac{n!}{(n-r)!r!}$

5명의 대학원생과 3명의 학부생이 공무원 시험에 응시하였다. 여기서 4명을 선발할 경우 대학원 생이 3명 포함되어 있을 확률을 구하시오.

주머니에 30개 red 공과 70개 green 공이 있다. 주머니에서 공을 뽑아 색을 확인하고 뽑은 공을 넣은 후 다시 공을 뽑아 색을 확인한다. (복원 추출) 20번 시행 했을 때 red 공이 정확하게 5개 추출될 확률을 구하시오. 만약 비복원인 경우 red 공의 개수가 정확하게 5개 뽑힐 확률을 구하시오.

6.이변량 확률

1)결합확률 joint prob. $P(AB) = P(A \cap B)$

두 개의 서로 다른 확률실험의 사건들의 교집합 확률을 결합확률이라 한다.

H대 학생들 500명 대상으로 색명여부, 성별을 조사한 결과이다.

	남자(G1)	여자(G2)	합계
색맹(C1)	50	20	70
정상(C2)	250	180	430
합계	300	200	500

H대 학생 중 한 명을 임의 추출했을 때 남자이고 색맹일 확률 $P(G1 \cap C1) = \frac{50}{500} = 0.1$

2)주변확률 marginal prob. $P(A)$ or $P(B)$

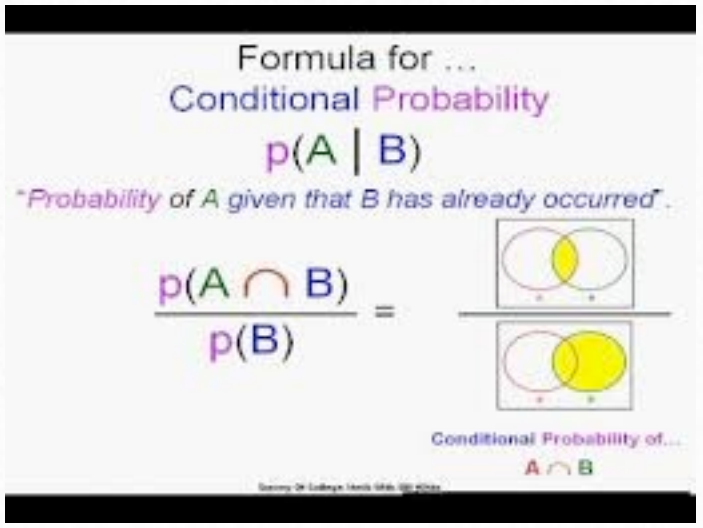
결합 확률실험에서 하나의 실험에만 관심

위의 조사에서 H대 학생 중 한 명을 선택했을 때 색맹일 확률 $P(C1) = 70/500$

3) 조건부 확률 conditional $P(A \mid B) = \frac{P(AB)}{P(B)}$

임의 실험 사건 결과가 주어졌을 때 다른 실험 사건 확률, 사건 B가 주어졌을 때(발생 했을 때) 사건 A의 조건부 확률

표본공간이 S에서 B로 줄어든다.



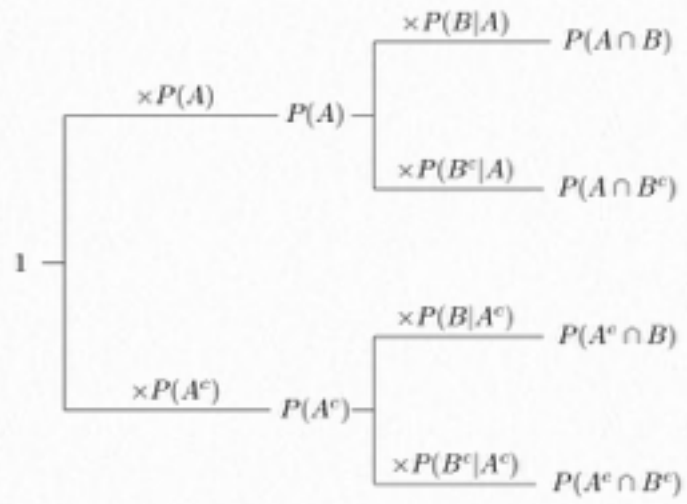
4) 독립 independence

[정의] 두 사건 A, B에 대하여 $P(AB) = P(A)P(B)$ (교집합의 확률이 각 사건의 주변확률 곱과 같으면), 두 사건은 서로 독립이다.

[정리] 두 사건이 독립이면 조건부 확률과 주변확률은 동일하다. $P(B \mid A) = P(B)$

확률계산 법칙

(1) Tree Diagram



(2) Addition Rule

$P(A \cup B) = P(A) + P(B) - P(AB)$

(3) Chain Rule, Multiplcative Rule

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

재범 가능성과 학력을 조사한 결과 표이다.

	재범	범죄 않음
대졸	10	30
고졸	27	33

사건 A={대졸}, 사건 B={재범}으로 정의할 때

(1)$P(A)$ (2)$P(B)$ (3)P(AB)
(4)P(A ∪ B) (5)P(B|$\overline{A}$)

공정한 주사위를 두 번 던질 때 X=1번 주사위 눈금, Y=2번 주사위 눈금, X+Y=7가 주어진 경우 P(X=4 or Y=4) 구하시오.

라디오 수신 채널이 양호할 확률이 0.8, 불량일 확률이 0.2이다. 수신 채널이 양호하다면 트랜스미션 오류 확률은 0.1이고, 채널이 불량이라면 미션 오류 확률이 0.3이다. 트랜스미션 오류가 발생하지 않은 경우 채널이 정상이었을 확률을 구하시오.

7. 베이즈 정리 Bayes Theorem

1)전확률법칙 total prob. rule

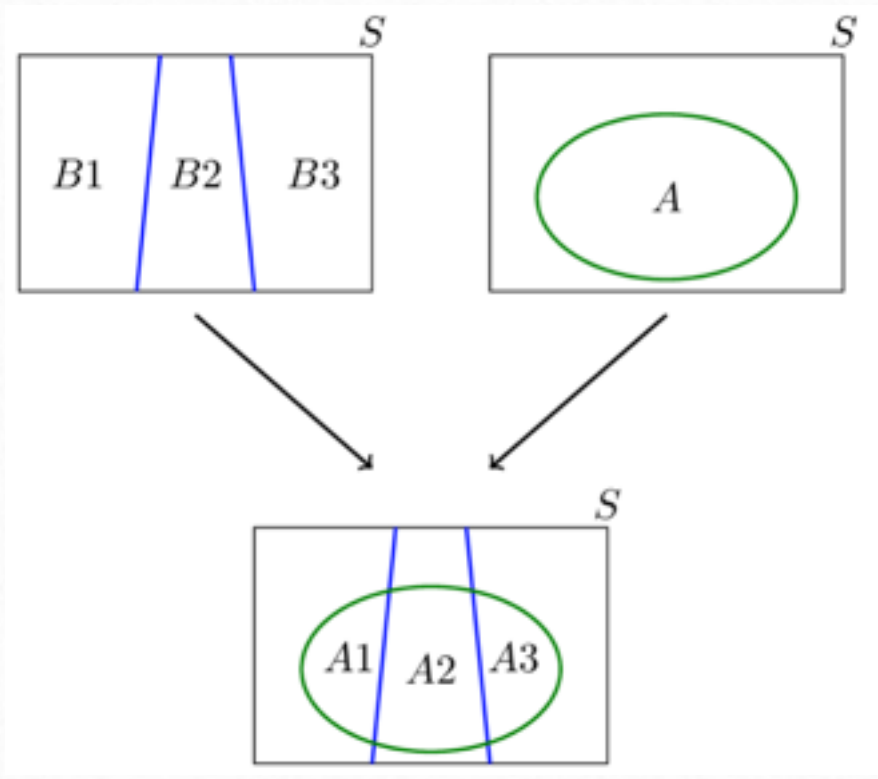
사건의 확률을 다른 사건들의 합(union)이나 곱(intersection)을 이용하여 표현하여 각각을 계산하기도 한다. 이런 방법을 event-composition 방법

P(A)를 구하는 방법으로 A를 mutually exclusive($B_i \cap B_j = \phi$) and exhaustive 사건(각 사건은 서로 배반이고 합집합은 S와 동일) B_1, B_2, B_3 사건에 의해 구할 수 있다. 이를 일반화 한 것을 전확률 법칙(law of total probability)이라 한다.

$$P(A) = P(AB_1) + P(AB_2) + P(AB_3)$$

$$P(A_1) = P(AB_1), P(A_2) = P(AB_2), P(A_3) = P(AB_3)$$

(일반화) $P(A) = \sum_i^k P(AB_i)$, { B_i} = 상호 exclusive and exhaustive (배타적이고 전체)

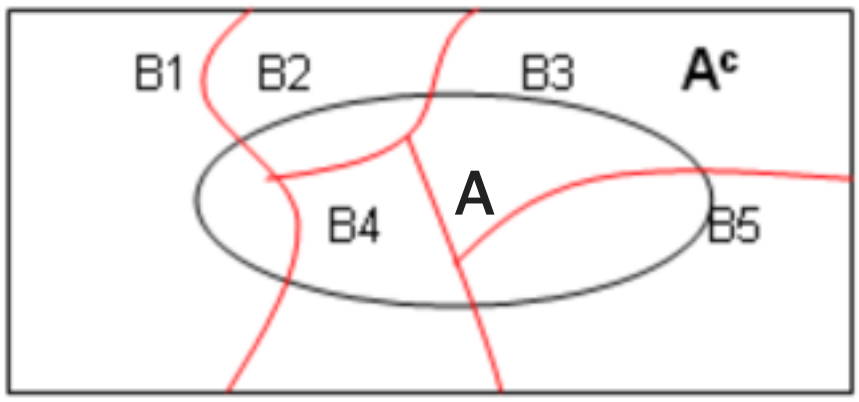


2)베이즈 정리

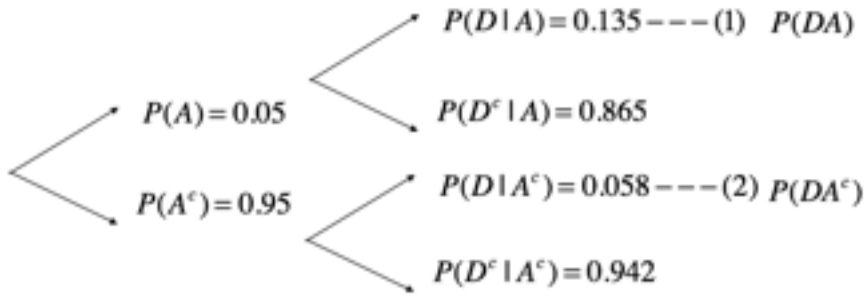
영국 철학자 Thomas Bayes의 이름에서 유래된 것으로 임의의 사건의 확률(Bi)을 새로운 사건 A에 의해 update 하는 것이다. 이런 개념은 관측치($x_1, x_2, \ldots, x_n$)에 의해 모수의 사전 분포함수($\pi(\theta)$)를 update한 사후 분포함수($\pi(\theta \mid x_1, x_2, \ldots, x_n)$)를 구하는데 사용한다.

만약 표본공간 { B_i} = 상호 exclusive and exhaustive (배타적이고 전체) $\sum_i^k B_i = S$이면,

$$P(B_j \mid A) = \frac{P(B_j)P(A \mid B_j)}{P(A)} = \frac{P(B_j)P(A \mid B_j)}{\sum_{i=1}^{k} P(B_i)P(A \mid B_i)}$$



제품 생산 공장이 5개 있고 각 공장에서는 동일한 양의 제품을 생산한다. A 공장 불량률은 5%이고, 나머지 공장은 불량률이 2%이라고 하자. 어느 공장에서 왔는지 모르는 박스를 하나 뜯어 그 중 3개를 검사하였더니 그 중 한 개가 불량이었다. 그 제품이 A 공장에서 생산되었을 확률을 구하시오.

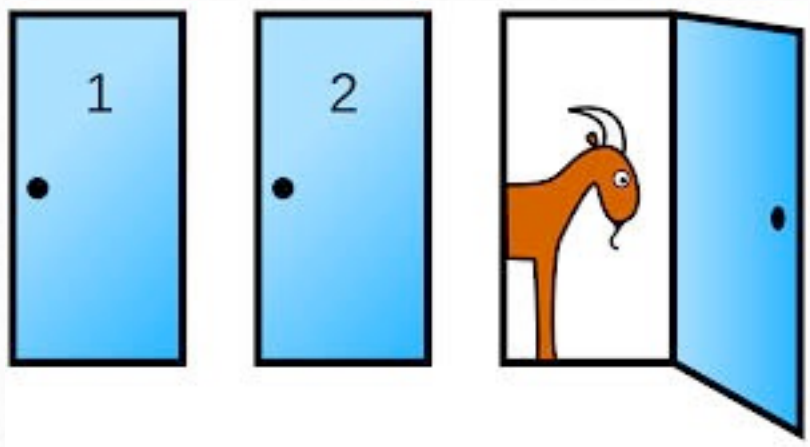


이메일에서 스팸 필터하는 방법은 단어를 체크하는 것이다. 기존 조사를 통하여 다음을 알고 있다고 하자. 이메일 50% 는 스팸이다. 스팸메일 1%는 “refinance" 단어가 있다. Non-스팸메일의 0.001%만이 “refinance" 단어를 가지고 있다. 지금 막 도착한 메일에 “refinance" 단어가 있다. 이 메일이 스팸일 확률을 구하시오.

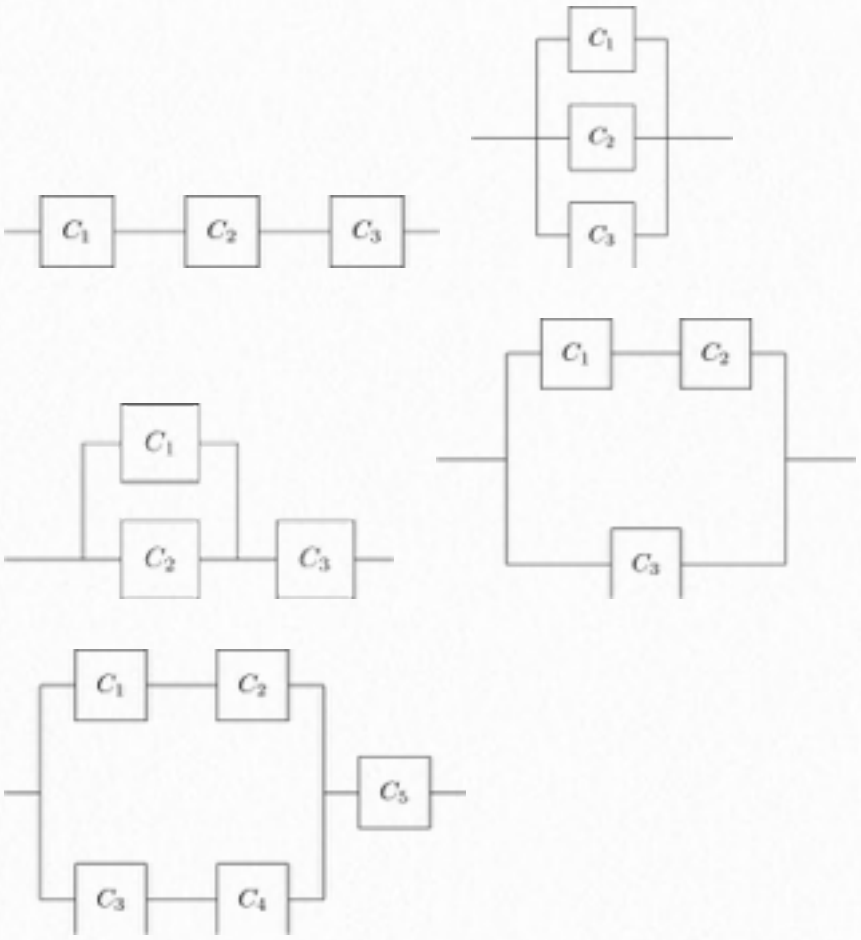
Graduate Management Admission Test (GMAT) MBA 진학을 위한 필수 시험이다. (200 to 800점).

GMAT 성적이 650점 이상이어야 지원이 가능하다. GMAT 650점 이상인 지원자 중 52%가 preparatory 코스를 수강했다. 반면, 650점 미만 지원자는 23%만 preparatory 코스를 수강 하였다. 나는 GMAT 650점 이상 받을 확률이 10%이다. preparatory 코스 수강료는 $500이다. 만약 코스를 수강하고 GMAT 시험을 보았을 때 성적이 650점 이상일 확률이 현재보다 2배 이상이면 코스를 수강하려고 한다. 나는 코스를 수강해야 하나?

당신은 게임 쇼에 게스트이다. 무대에 커튼이 3개 있으며, 2개 커튼에는 당나귀, 다른 한 커튼 뒤에는 스포츠 카가 있다. 쇼 호스트 몬티는 당신에게 커튼 하나를 임의로 택하게 한다. 당신이 1을 택하였다면 커튼 3을 열어 그 곳에 당나귀가 있음을 보여준다. 그리고 당신에게 첫 선택을 바꿀 것인지 묻는다. 바꾼다면 당신의 선택은 2번이 되고 그렇지 않으면 1번이 된다.
(1) 바꾸지 않을 경우 당신이 스포츠 카를 받을 확률?
(2) 바꾸는 경우 당신이 스포츠 카를 받을 확률?



각 회로가 작동할 확률을 $P(C_i) = 0.1 * i$ 이라 가정할 때 다음 회로가 작동할 확률?



현재 상황 노아웃에 주자 1루 상황이다. 타자 A에게 맡긴다면 이전 데이터에 근거하면 득점 확률이 0.39이다. 번트를 하게 하는 것이 득점 확률을 높일 수 있다면 번트를 하게 할 것이다. **감독을 과연 타자 A에게 번트를 하게 할 것인가?**

타자 A의 이전 번트 결과를 정리하면 다음과 같다. 다른 상황은 발생하지 않는다.

1) P(번트 성공1=원아웃 주자 2루)=0.75

2) P(번트 실패1=원아웃 주자 1루)=0.1

3) P(번트 실패2=투아웃 주자 없음)=0.1

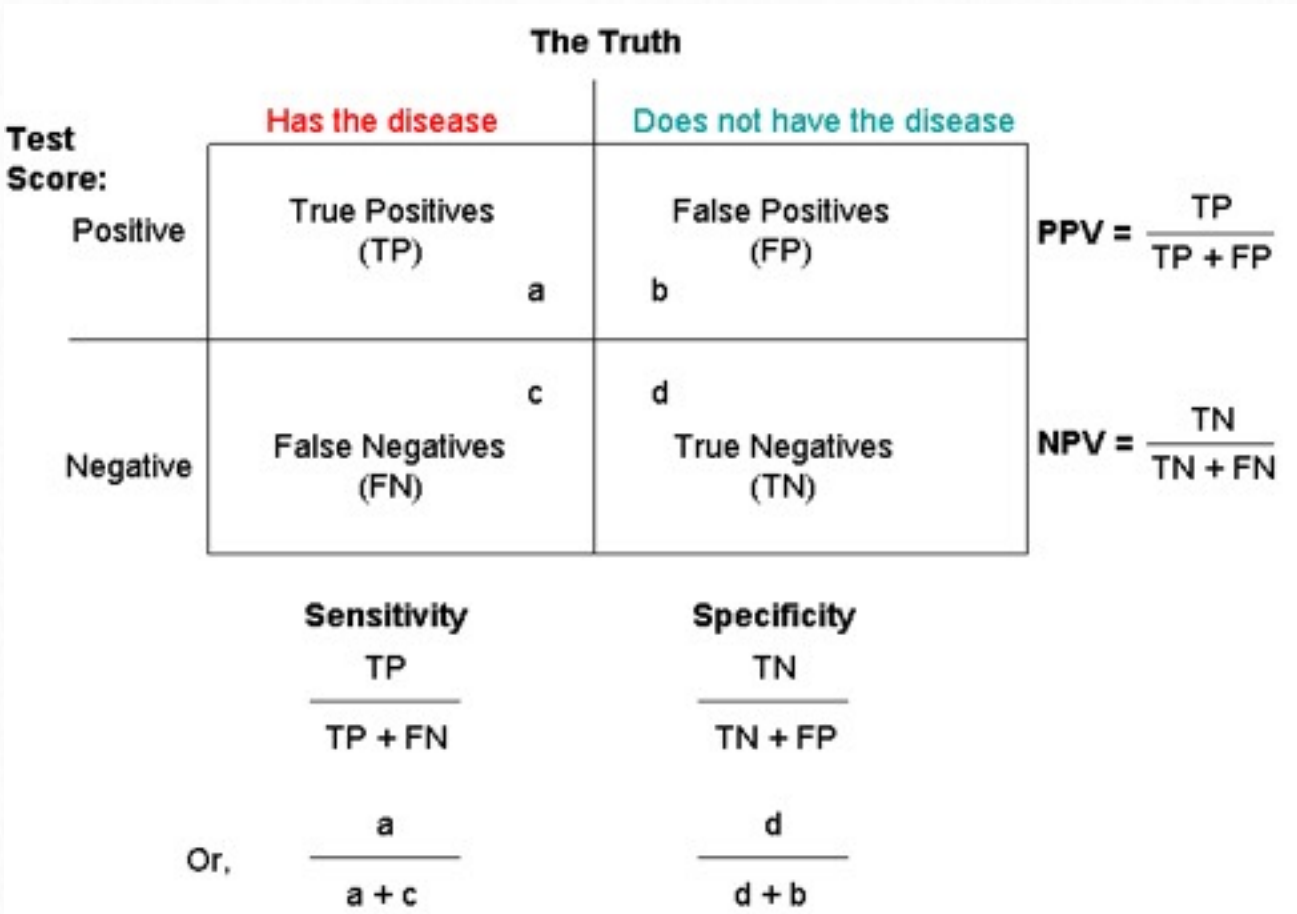
4) P(번트 성공2=노아웃 주자 1,2루)=0.05

(Part 2) 주자가 1루에 있다. 1루 주자의 도루 성공 가능성은 68%이다. (1) 노아웃 no out (2) one out (3) two outs 상황에서 도루하는 것이 득점 확률을 높이는지 하지 않는 것이 높이는지 결정해 보시오.

2015년 한 해 동안 SK팀 경기를 분석한 결과 아래 상황에서 득점할 확률에 대하여 다음 데이터를 얻었다.

베이스 현황	노아웃	원아웃	투아웃
주자 없음	0.26	0.16	0.07
주자 1루	0.39	0.26	0.13
주자 2루	0.57	0.42	0.24
주자 3루	0.72	0.55	0.28
주자 1,2루	0.59	0.45	0.24
주자 1,3루	0.76	0.61	0.37
주자 2,3루	0.83	0.74	0.37
만루 loaded	0.81	0.67	0.43

8.베이즈 정리 (활용)



◉ 민감도 : 양성(positive) 환자를 검사하여 양성이라고 진단할 확률 = TP/(TP+FN)

◉ 특이도 : 음성(negative) 환자를 검사하여 음성이라고 진단할 확률 = TN/(TN+FP)

◉ 양성예측도 positive predictive value : 유병(양성)이라 진단한 사람 중 유병인 사람의 확률

$$PPV = \frac{TP}{TP + FP}$$

◉ 음성예측도 negative predictive value : 건강(음성)이라 진단한 사람 중 건강한 사람의 확률

$$NPV = \frac{TN}{TN + FN}$$

◉ 검사 정확율 Accuracy = (TP+TN)/(TP+FN+FP+TN)

암을 진단하는 PSA 테스트(비용 $50)는 정확도가 낮아 false positive=0.135, false negative=0.3 이다. 이를 보완하기 위하여 조직검사(비용 $1,000)를 병행한다. 조직검사 정확도는 100%라고 가정하자. 다음은 각 연령별 암 발생 비율이다.

연령군	발생비율 (사전확률)
40대	0.01
50대	0.022
60대	0.046
70대	0.079

각 연령군에서 PSA 테스트 결과, 암으로 진단할 확률, 암이 아닌 것으로 진단할 확률을 구하고 해석하시오.

연령군	P(C|PT) (사후확률)	P(NC|PT)
40대		
50대		
60대		
70대		

인구 1000명인 경우 각 연령군의 조직검사까지 갔을 때 PSA 결과 PT 확률과 암으로 판명되는 사람 수, 그리고 암 환자 1인당 비용을 구하시오.

연령군	P(PT)	암 진단(명)	일인당 비용
40대			
50대			
60대			
70대			

PSA 검사의 민감도, 특이도, Positive Predictive Value (양성 예측률), Negative Predicted Value (음성 예측률), 그리고 정확도 accuracy를 구하시오.

(false positive=0.5, false negative=0.5), (false positive=0.7, false negative=0.3)일 경우 반복하시오.