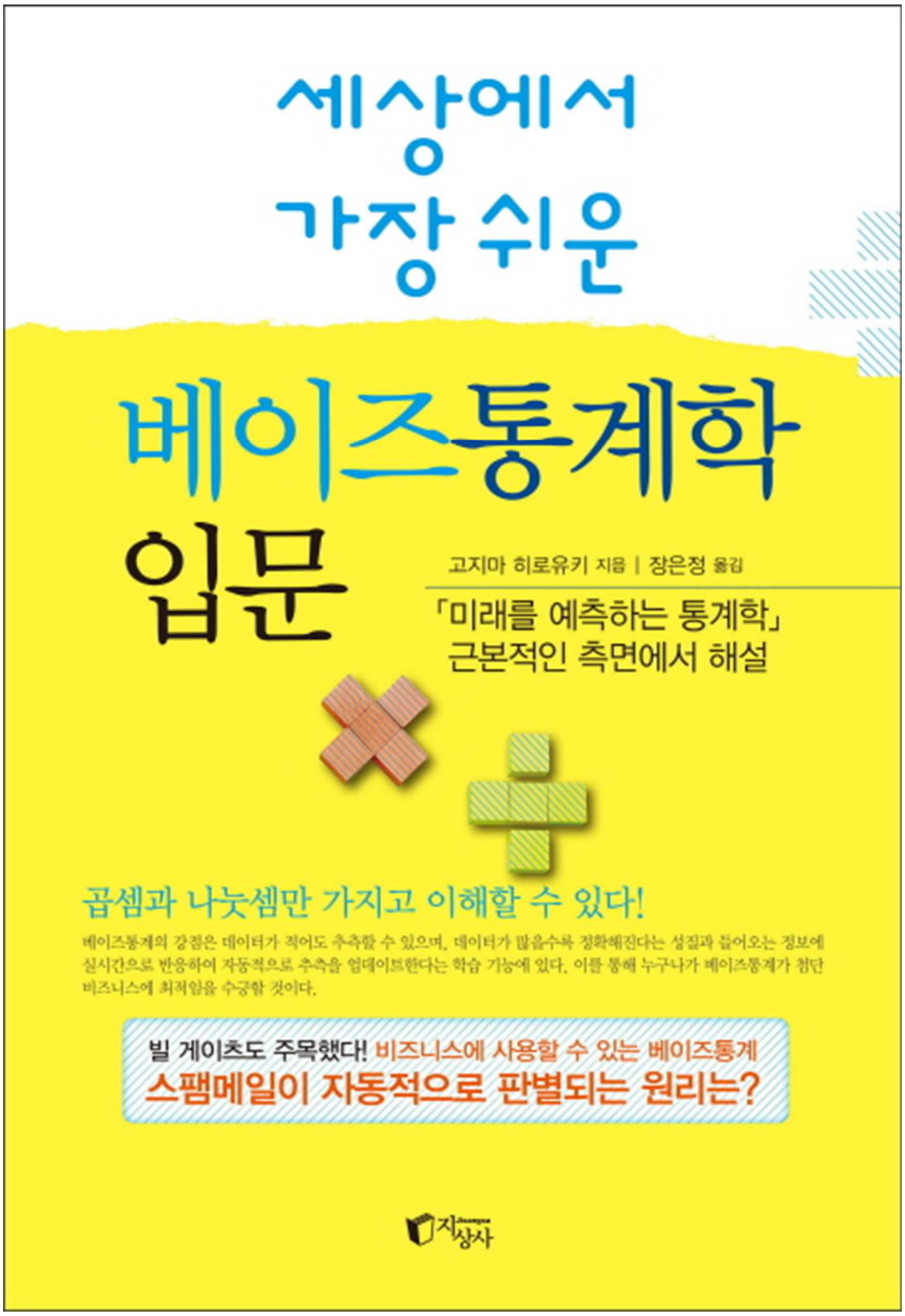
세상에서 가장 쉬운 베이즈 통계학 입문



제0강 사칙연산만으로 이해하는 베이즈통계학
이 책의 특장
0-1 예비지식이 전무한 상태에서도 실제 활용할 수 있는 수준까지 도달할 수 있다.
0-2 면적도와 산수, 이 두 가지로 해결한다
0-3 빌 게이츠도 주목했다! 비즈니스에 사용할 수 있는 베이즈통계
0-4 베이즈통계는 인간의 심리에 의존한다
0-5 빈칸 채우기 형식의 간단한 연습문제는 독학에 최적이다

제1부

속성! 베이즈통계학의 에센스를 이해한다

제1강 정보를 얻으면 확률이 바뀐다

'베이즈 추정'의 기본적인 사용 방법

1-1 베이즈추정으로 '쇼핑족'과 '아이쇼핑족'을 판별한다

점원은 손님이 쇼핑족인지 아이쇼핑족인지 판별해야 한다.

1-2 [1단계] 경험에서 '사전확률'을 설정한다

추측을 위해 가장 먼저 해야 할 일은 손님의 두 가지 타입-'쇼핑족'과 '아이쇼핑족'-에 대한 그 비율이 각각 몇 인지 수치를 배정하는 것이다. 이 타입에 대한 확률(비율)을 '사전확률'
'사전'이란 '어떤 정보가 들어 오기 전'을 의미.
'어떤 정보'란 손님이 말을 거는 행동 등 추가적인 상황
'말을 걸었다'는 정보에 의해 손님의 타입에 대해 추측을 개정
'사전확률'이라는 것은 '말을 건다. 걸지 않는다'는 행동의 관측이 이루어지기 전의 상태를 말함

경험상 손님중 '쇼핑족'의 비율은 20%= 0.2, 당연히 '아이쇼핑족'은 80%=0.8. 이것이 타입에 대한 '**사전분포**'라고 부른다. 이를 그림으로 나타내면 도표1-1과 같다. 각 직사각형의 면적의 비율은 0.2:0.8. 면적이 베이즈 확률을 다루는 중요한 역할

도표1-1 사전분포로 직사각형을 분할한다
0.2 0.8

A 쇼핑족	B 아이쇼핑족

위의 그림은 '둘로 분기된 세계'이다. A 세계는 쇼핑족와 B 세계는 아이쇼핑족. 철학에서는 이러한 관점을 '가능세계'라 부른다. 면적의 비율을 0.1과 0.4로 하지 않는 이유는 '확률은 전부 더해서 1이 되도록 설정한다'는 수학의 약속에 근거한 것. 이것을 '**정규화 조건**'이라고 한다.

1-3 [2단계] 타입별로 '말 거는'행동을 하는 '조건부 확률'을 설정한다

쇼핑족과 아이쇼핑족 각각의 손님이 어느 정도의 확률로 점원에게 '**말 걸기**' 행동을 하는 가를 설정한다
사전확률과 달리 이 확률은 경험에 의해 설정해야 한다. '**타입의 차이에 의거한 행동의 확률**'은 어떠한 경험, 실증, 실험에 기반한 수치가 필요하다.

도표1-2 행동에 대한 조건부 확률

타입	말을 걸 확률	말을 걸지 않을 확률	합계
쇼핑족	0.9	0.1	1.0
아이쇼핑족	0.3	0.7	1.0
합계	1.2	0.8	

표의 가로방향은 정규화조건이 충족(0.9+0.1=1.0, 0.3+0.7=1.0) 그러나 세로방향은 정규화조건이 성립되지 않음(0.9+0.3=1.2, 0.1+0.7=0.8) 표의 가로방향은 특정 타입의 손님에 대해 일어날 수 있는 두 가지의 귀결을 나타낸다. 그러나 세로방향은 각각 다른 타입의 사람에 대한 행동을 나타내는 것이지 행동 전체(여기선 두 가지의 배타적인 행동)를 아우르는 확률적 사건이 아니므로 반드시 1.0이 아니어도 된다.

위의 확률은 '**조건부 확률**'이다. 즉 '**타입을 한정한 경우 각 행동의 확률**'을 의미한다. 타입이 행동의 원인이고 그 결과인 각 행동의 확률은 '**원인을 알고 있을 때의 결과의 확률**'

도표1-1의 두 개의 세계는 다음과 같이 다시 나눌 수 있다.

도표1-3 네 개로 분기된 세계

쇼핑족	아이쇼핑족
0.2	0.8

말을 건다 0.9	A 쇼핑족이 말을 건다	C 아이쇼핑족이 말을 건다	말을 건다 0.3
말을 걸지 않는다 0.1	B 쇼핑족이 말을 걸지 않는다	D 아이쇼핑족이 말을 걸지 않는다	말을 걸지 않는다 0.7

도표1-4 네 개로 분기한 세계 각각의 확률

A 0.2*0.9=0.18	C 0.8*0.3=0.24
B 0.2*0.1=0.02	D 0.8*0.7=0.56

A, B, C, D 네 개의 세계의 확률을 더하면 1.0이 된다

A + B + C + D = 0.18 + 0.02 + 0.24 + 0.56 = 1.0

1-4 [3단계] 관측한 행동에서 '가능성이 사라진 세계'를 제거한다

이제 손님이 말을 걸었다. 말을 걸었다는 사건이 관측되었으므로 '**말을 걸지 않는다**'라는 세계는 사라진다.

도표1-5 정보에 따라 가능성이 한정된다

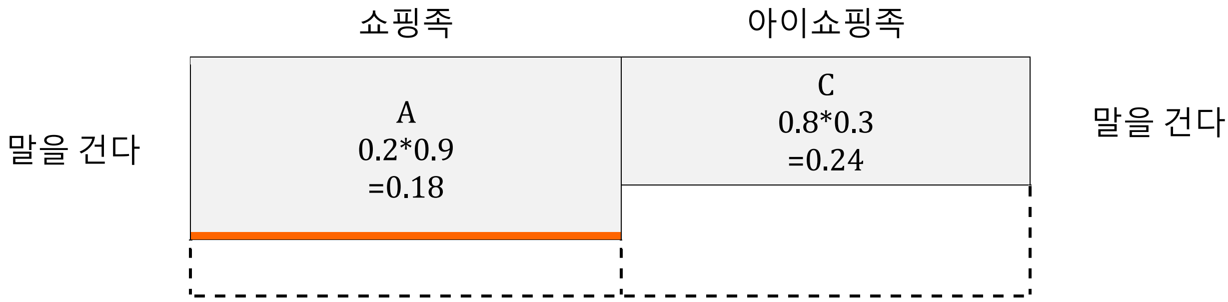
	0.2	0.8	
0.9	A 0.2*0.9	C 0.8*0.3	0.3
0.1	~~B~~ ~~0.2*0.1~~	~~D~~ ~~0.8*0.7~~	0.7

B와 D의 가능성이 사라진다.

1-5 [4단계] 쇼핑족의 베이즈 역확률을 구한다

이제 남은 세계는 쇼핑족&말걸기 세계(A)와 아이쇼핑족&말걸기 세계(C)이며 손님은 두 개의 세계중 하나에 속하게 된다.

도표1-7 말을 거지 않는 세계의 소멸

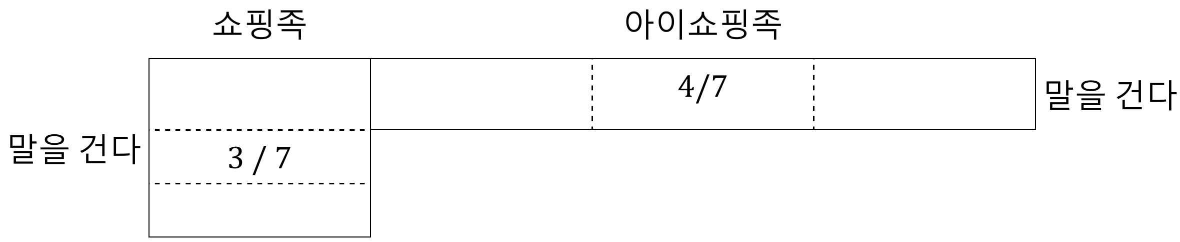


행동의 관측에 따라 가능성이 두 가지로 좁혀졌지 때문에 각각의 확률(직사각형의 면적)을 더해도 1이 되지 않는다. 따라서 비례관계를 유지한 채 정규화조건을 만족하도록 계산한다.

(왼쪽 직사각형의 면적) : (오른쪽 직사각형의 면적)
= 0.18 : 0.24 = 3 : 4

합계 3+4=7로 나누어 계산하면
(왼쪽 직사각형의 면적) : (오른쪽 직사각형의 면적) = 3 : 4 = 3/7 : 4/7

도표1-8 정규화 조건을 회복시켜 사후확률을 구한다



말을 걸어온 손님이 쇼핑족일 확률을 3/7이라고 추정할 수 있다. 이 확률을 **베이즈 역확률** 또는 **사후확률**이라 한다.

손님의 타입이 두 종류이며 각 타입이 '말을 건다' 또는 '걸지 않는다'의 두 가지 행동을 확률적으로 선택한다. 타입이라는 원인으로부터 행동의 결과가 일어나는 것이다. 그러나 관측된 결과로부터 타입이라는 원인을 거슬러 올러간다. 분석의 흐름은 결과 ⇨ 원인 으로 이어지는 것이라 **사후확률**을 **역확률**이라고 한다. 역확률에서 '역'이라는 말은 이를 가리킨다.

1-6 베이즈 추정의 프로세스 정리

도표1-9 손님의 타입에 대한 베이즈 추정의 프로세스

1. 타입에 대한 사전확률의 설정 ⇨ 쇼핑족과 아이쇼핑족이 있다
2. 각 타입의 행동에 대한 조건부 확률의 설정 ⇨ 각각 말을 걸 확률은?
3. 행동의 관측 ⇨ 말을 걸어왔다
4. 일어나지 않을 가능성의 소거 ⇨ '말을 걸지 않는다'를 지운다
5. 타인에 대한 확률의 정규화 ⇨ 더해서 1이 되도록 만든다
6. 사후확률(베이즈 역확률) ⇨ 말을 거는 손님이 살 확률이 달라졌다

도표1-10 손님의 타입에 대한 베이즈 갱신

1. 쇼핑족일 사전확률 = 0.2
2. '말 걸기'를 관측
3. 타입 '쇼핑족'의 사후확률 = 3/7 = 0.428…

눈앞의 손님이 '쇼핑족' 타입일 확률은 아무 것도 관측되지 않았을 때 0.2(사전확률)이지만 '말 걸기'가 관측된 이후 그 정보가 갱신되어 약 0.43(=4/7, 사후확률)이 된다. '쇼핑족'이라고 완전히 확신할 수 없지만 그 가능성은 두 배 높아진다. 이것을 '**베이즈 갱신**'이라고 한다.
그리고 이 모두 과정을 '**베이즈 추정**'이라고 한다. 베이즈 추정이란 '**사전확률을 행동의 관찰(정보)에 의거해 사후확률로 베이즈 갱신하는 것**'이다.

제1강의 정리

1. 각 타입 '쇼핑족'과 '아이쇼핑족'의 확률을 산정한다(사전확률)
2. '쇼핑족'타입이 '말을 건다', '말을 걸지 않는다'는 행동을 얼마만큼의 확률로 취하는가. '아이쇼핑족'타입이 '말을 건다', '말을 걸지 않는다'는 행동을 얼마만큼의 확률로 취하는가를 설정한다(조건부확률). 이때 경험이나 데이터가 필요하다
3. '말을 건다'는 행동이 관측되었기 때문에 '말을 걸지 않는다'가 속한 세계를 소거한다
4. '쇼핑족&말을 건다'의 확률과 '아이쇼핑족&말을 건다'의 확률그룹이 정규화조건을 충족하도록 한다. 즉 비례관계를 유지하여 더해서 1이 되도록 만든다.
5. 정규화조건을 회복한 타입 '쇼핑족'의 확률이 , '말을 거는' 행동이 관찰된 타입 '쇼핑족'의 사후확률이 된다
6. 사전확률이 행동을 관찰함으로써 사후확률로 갱신된다. 이것을 베이즈 갱신이라 한다

/ 연습문제

제2강 베이즈 추정은 때로 직감에 크게 반한다①

객관적인 데이터를 사용할 때 주의할 점

2-1 암에 걸려 있을 확률을 계산한다

객관적인 데이터를 얻기 쉬운 사례에 대한 베이즈 추정

연관되어 보이지만 잘 구분해야 하는 별개의 확률
- 특정 암에 대한 간이 검사 확률
- 실제 특정 암에 걸릴 확률

사전확률
- 0.1% - 특정 암에 걸릴 확률
- 95% - 간이검사가 정확할 확률 (양성을 양성으로 판별하는 확률)

문제설정
어느 특정 암에 걸릴 확률이 0.1%(0.001)이라고 하자. 이 암에 결렸는지를 진단하는 간이검사가 있는데, 이 암에 걸려 있는 사람은 95%(0.95)의 확률로 양성진단을 받는 다고 한다. 한편 건강한 사람이 양성으로 오진을 받을 확률은 2%(0.02)이다. 그렇다면 이 검사에[서 양성이라고 진단받았을 때 당신이 이 암에 걸려 있을 확률은 얼마나 될까?

2-2 의료데이터를 근거로 '사전확률'을 설정한다

이 예제에서는 사전확률을 경험이 아닌 객관적인 역학 데이터를 얻을 수 있다.
사전확률이란 '**각 타입에 대한 정보를 얻기 전의 존재확률**'
두 가지의 타입
- 암에 걸려 있는 사람 - 0.01%
- 건강한 사람 - 99.9%

도표2-1 암의 이환율에 따른 사전분포

0.001	0.999
암	건강

간이 검사를 받기 전 두 개의 가능한 세계로 나눌 수 있다. 좌측은 암에 걸려 있는 세계, 오른쪽은 건강한 세계. 당신은 두 가지 가능세계중 어느 한쪽에 속하는 데, **아무런 개인적인 정보가 없는 현재로서는 좌측의 세계에 속할 확률이 0.001, 우측의 세계에 속할 확률이 0.999로 추측**된다.

2-3 검사의 정밀도를 근거로 '조건부 확률'을 설정한다

각 타입별로 **특정한 정보(검사결과로서 양성/음성)**를 초래하는 조건부 확률을 설정해야 한다. 검사결과로서 양성/음성이 정보에 해당한다.

도표2-2 검사정밀도에 따른 조건부 확률

타입	양성일 확률	음성일 확률
암에 걸린 환자	0.95	0.05
건강한 사람	0.02	0.98

위의 확률은 타입을 한정한 경우에 각 검사결과의 조건부 확률이다. **타입을 결과의 원인으로 잡는다면 원인(암인가 건강한가)을 알고 있을 때의 결과(양성/음성)의 확률이라고 볼 수 있다.**

2개로 나뉜 세계를 정보(양성/음성)에 따라 각각 다시 2개로 나누면 도표2-3과 같다.

도표2-3 네 개로 분기된 세계

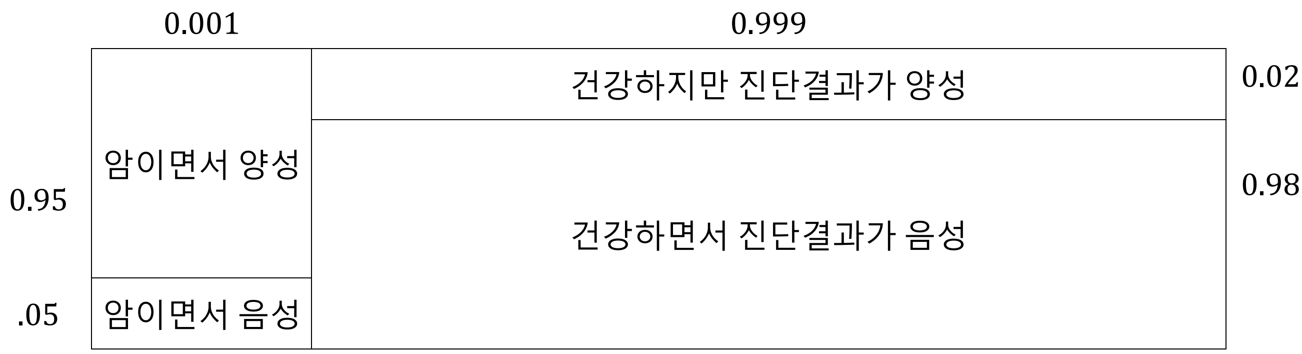


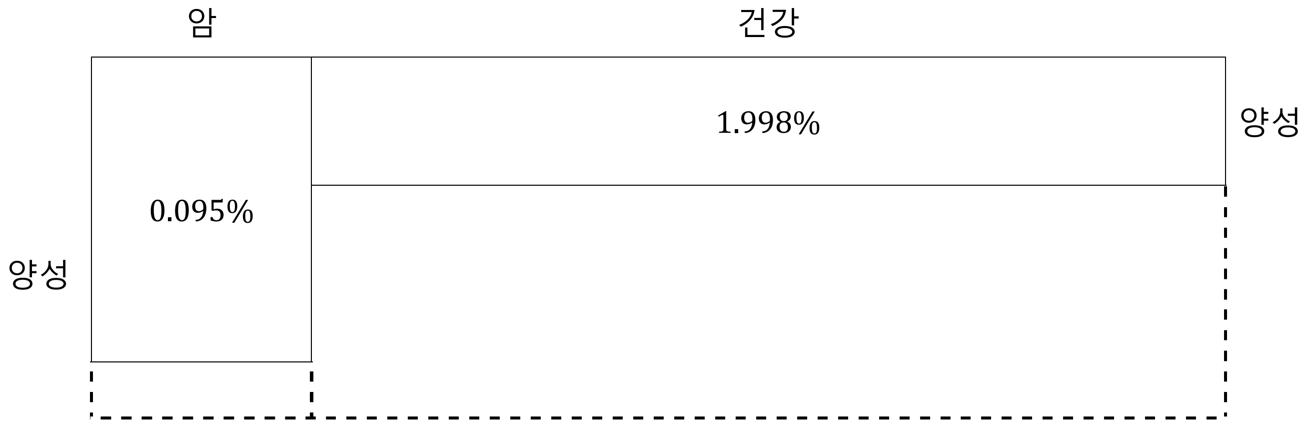
도표2-4 네 개로 분기된 세계 각각의 확률

0.005% (0.001*0.05)	

2-4 검사결과가 양성이므로 '일어날 가능성이 없는 세계'를 소거한다

현재 당신의 **검사결과 양성판정**을 받았다. 즉 당신의 몸속에서 일어나는 일에 관한 **정보를 한 가지 관측**한 셈이다. 이 정보로 인해 음성인 세계는 사라진다.

도표2-5 정보에 따라 가능성이 한정된다



2-5 당신이 암일 것이라는 '베이즈 역확률'을 구한다

양성을 관측하였으므로 **암&양성** 또는 **건강&양성**인 두 개의 세계만 남는다. 가능성이 4개에서 2개로 줄어들어 확률(직사각형의 면적)을 더해도 1이 되지 않는다. 정규화 조건을 복구하기 위해 비례관계를 유지한 상태로 '더해서 1이 되도록'한다.

(왼쪽 직사각형의 면적) : (오른쪽 직사각형의 면적) = 0.095 : 1.998

0.095 + 1.998 = 2.093

(왼쪽 직사각형의 면적) : (오른쪽 직사각형의 면적) = 0.095/2.093 : 1.998/2.093 = 0.0454 : 0.9546

0.0454 + 0.9546 =1.0 (정규화 조건 만족)

이 결과로 양성이라는 검사결과를 받았을 때 , 당신이 암에 걸려 있을 사후확률은 **4.5%**임을 알 수 있다.

2-6 베이즈 추정의 프로세스 정리

도표2-7 암 이환율의 베이즈 추정 프로세스

1. 암인가 건강한 가에 대한 사전확률을 설정(역학 데이터 이용)
2. 검사의 정밀도에 대한 조건부 확률을 설정(치료 데이터 이용)

3. 검사결과를 관측-양성이라는 정보 얻음
4. 음성일 가능성 소거
5. 암/건강에 대한 확률의 정규화
6. 암일 사후확률(베이즈 역확률)

도표2-8 암 검사에 대한 베이즈 갱신

1. 암일 사전확률 = 0.001
2. 검사로 양성 관측
3. 암일 사후확률 = 0.045

아무런 관측이 없을 때 암에 걸릴 확률은 0.001(사전확률)이지만 양성이 관측된 후 그 정보를 근거로 수치가 갱신되어 0.045(사후확률)가 되었다. 즉 확률이 **0.1%에서 4.5%**(45배)로 올라갔다.

제2강의 정리

1. 타입 '암', '건강'의 사전확률을 설정한다(역학데이터를 이용한다)
2. 암 검사의 감도를 설정한다. 즉 암인 사람의 양성/음성의 조건부 확률과 건강한 사람의 양성/음성의 조건부 확률을 설정한다(치료데이터 활용)
3. '양성'이 관측되었기 때문에 '음성'의 세계를 소거한다
4. '암&양성'의 확률과 '건강&양성'일 확률 값에 대해 정규화 조건을 복구시킨다(비례관계를 유지한 상태로 더해서 1이 되도록 한다)
5. 정규화조건이 복구된 '암&양성'의 수치가 검사결과 양성이 나온 사람이 실제로 암일 사후확률(베이즈 역확률)이다
6. 사전확률이 검사결과를 관측함으로써 사후확률로 갱신된다(베이즈 갱신)

/ 연습문제

제3강 주관적인 숫자여도 추정이 가능하다

곤란한 상황에서 쓰는 '이유 불충분의 원리'

3-1 초콜릿을 준 그녀의 마음을 추정한다

지금까지의 베이즈 추정의 수순
(사전확률) ⇨(조건부 확률) ⇨(관측에 의한 정보의 입수) ⇨(사후확률)
1,2장은 객관적인 데이터를 참고하여 사전확률을 설정. 그러나 객관적인 사전데이터가 없어도 추정가능. 사전확률을 주관적으로 설정하여 추정을 실시할 수 있다

문제설정
당신이 남성이라고 가정하자. 특정여성 동료가 자신에게 호감을 가지고 있는 지 알고 싶은 상황이다. 그런 와중에 당신은 밸런타인데이에 그녀로부터 초콜릿을 받았다. 이때 그녀가 당신을 진지하게 생각하고 있을 확률이 얼마라고 추정해야 할까?

3-2 주관적으로 당신을 마음에 두고 있는 가에 대한 '사전확률'을 설정한다

이 문제의 특수성
사전확률을 객관적인 통계데이터를 이용해 얻을 수 없다
사전확률
어떤 정보가 들어오기 전 각 타입('진심'과 '논외')에 대한 비율

이유불충분의 원리
특정 여성의 마음에 대한 추측이라 사전확률을 구할 데이터가 없다. 따라서 이유불충분의 원리를 채용. '진심'과 '논외'에 대한 사전확률근거가 없어 일단 대등하게 설정(사전확률을 0.5와 05로 설정)

도표3-1 이유불충분의 원리에 따른 사전분포

0.5	0.5
진심	논외

3-3 어떻게든 데이터를 입수하여 '조건부 확률'을 설정한다

타입별로 조건부확률을 설정해야 한다.

통계적으로 평균 그녀들은 관심있는 상대에게 42.5% 확률로, 논외인 상대에게 22%의 확률로 초콜릿을 준다.

도표3-2 직장여성이 초콜릿을 줄 조건부 확률

타입	초콜릿을 줄 확률	초콜릿을 주지 않을 확률
진심	0.4	0.6
논외	0.2	0.8

계산의 간편함을 위해 끝수를 버림

위의 확률은 ‘타입을 특정한 경우 각 행동의 확률’이다. 원인(진심/논외)을 알고 있을 때의 결과(준다/주지 않는다)의 확률이라 할 수 있다.

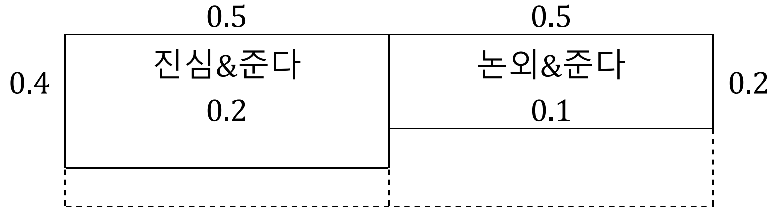
앞서 설정한 두 개의 세계(진심/논외)를 다시 각각 둘로 나누어 다음과 같이 네 개로 나눈다

도표3-3 네 개로 나뉜 세계 각각의 확률

3-4 초콜릿을 받았으므로 '일어난 가능성이 없는 세계'를 소거한다

운 좋게도 동료여성으로부터 초콜릿을 받았다. 그래서 '주지 않는다'는 세계는 사라진다.

도표3-4 정보에 따라 가능성이 한정된다

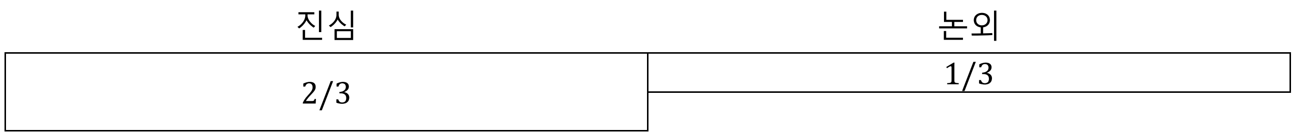


동료여성의 행동을 관측함으로써 가능세계는 4개에서 2개로 좁혀지고 비례관계를 유지한 채 모든 확률이 합쳐 1이 되도록 정규화 조건을 회복시킨다

(왼쪽 직사각형의 면적) : (오른쪽 직사각형의 면적) = 0.2 : 0.1 = 2: 1

(왼쪽 직사각형의 면적) : (오른쪽 직사각형의 면적) = 2: 1 = 2/3 : 1/3

도표3-5 정규화 조건을 이용해 사후확률을 구한다



이 결과로부터 그녀의 진심일 확률은 2/3 = 66%가 된다

3-5 베이즈 추정의 프로세스 정리

도표3-6 '진심' '논외'의 베이즈 추정 프로세스

1.	진심/논외에 대한 사전확률을 설정(데이터를 얻을 수 없으므로 '이유 불충분의 원리'에 따라 반반으로 설정하다)
2.	동료 여성의 행동에 대한 조건부 확률을 설정(조사 데이터를 이용)
3.	행동을 관측
4.	가능성의 소거
5.	타입에 대한 확률의 정규화
6.	진심일 사후확률(베이즈 역확률)

도표3-7 그녀의 마음에 대한 베이즈 갱신

1.	당신이 진심일 확률 = 0.5
2.	초콜릿을 받았다
3.	당신이 진심일 사후확률 = 2/3 = 약 0.66

이유불충분 상황이라도 사전확률을 반반으로 하는 게 지나치다면 진심 0.4, 논외 0.6으로 설정

3-6 '신념의 정도'에도 베이즈 추정을 사용할 수 있다

확률

중고등학교에서 배운 확률은 객관적인 확률-누가 주사위를 던지든 확률은 1/6. 주사위를 던져 1이 나올 가능성의 정도- 모든 사람의 공통된 판다

그러나 이번 강의 확률은 동료 여성이 당신을 마음에 두고 있을 확률에서 그 확률은 **믿음의 정도**와 같은 것이다. 사람이 마음으로 생각하는 수치라고 해석하는 확률을 **주관확률**이라 부른다.

제3강의 정리

1. 타입에 대한 사전확률을 설정한다(데이터를 얻을 수 없으므로 이유불충분 원리를 채용하여 반반으로 설정한다)
2. 행동에 대한 조건부 확률을 설정한다(조사 데이터를 활용)
3. 얻은 행동의 정보로부터 일어날 수 없는 가능성을 소거한다
4. 남의 세계의 확률값은 비례관계를 유지한 채 더해서 1이 되도록 정규화조건을 복구시킨다
5. 타입에 대한 사후확률(베이즈 역확률)을 얻을 수 있다

6. 사전확률이 행동을 관찰함에 따라 사후확률로 변경된다(베이즈 갱신)
7. 여기서 다룬 확률은 '주관 확률'이다

/ 연습문제

제4강 '확률의 확률'을 사용하여 추정의 폭을 넓힌다

4-1 첫째는 여자아이다. 그렇다면 둘째는 남아일까 여아일까?

문제설정
어떤 부부의 첫째 아이가 여아였다고 치자. 이때 그 부부에게서 태어날 둘째 아이가 여아일 확률은 몇일까?

특정부부에 대한 출산에 관한 추정이 가능한 이유는 베이즈 추정의 '느슨한' 특성 덕분인데, 느슨함이란 사전확률이라는 불가사의한 것을 설정한다는 것.

4-2 '확률의 확률'을 사전확률로 설정한다

여아가 태어날 확률을 p라고 한다. 남아 또는 여아이므로 당연히 확률p는 0.5라고 생각할테지만 그건 인류 전체로 통계를 낸 경우이고 이번은 특정 부부에 대한 것이라 확률p는 반드시 0.5는 아니다. '그 부부에게 여아가 태어날 확률 p는 0~1을 만족하는 연속적인 수이다. 그리고 p는 부부의 타입을 나타내는 수이다.

타입p는 세 개의 확률중 어느 곳에 속하는 지 알 수 없어 추축의 대상이다. 세 개의 세계중 어느 하나가 나머지 둘보다 높거나 낮다고 확신할 수 없으므로 '이유불충분의 원리'를 적용해 사전확률을 모두 1/3으로 설정한다.

타입p를 0.6, 0.5, 0.4 3개의 값으로 설정한다. 이 부부에게서 태어날 아이가 여아일 확률 p를 0.6, 0.5, 0.4의 세 값으로 결정했으므로 이 부부는 이 세가지의 타입중 어느 하나에 해당한다고 가정한다. 이 부부가 어느 타입에 속하는 지에 대한 통계적인 데이터가 전혀 없으므로 '이유 불충분의 원리'를 채용한다. 세 타입에 확률을 각각 1/3씩 설정하는 것이다.

도표4-1 이유불충분 원리에 따른 사전분포

1/3	1/3	1/3
P=0.4	P=0.5	P=0.6

P 자체도 확률이므로 'P=0.4'일 사전확률인 1/3은 '확률의 확률'이다. 사전확률 1/3은 세 가지로 설정되어 있는 타입의 확률 p의 값 중 어느 것이 진실인가에 대한 가능성'을 나타내 주는 수치. 즉 사전확률은 그 부부가 어느 가능 세계에 속해 있는 가에 대한 확률을 나타내며 확률p는 각 가능세계에서 그 부부가 여아를 낳을 확률을 나타낸다. 즉 전혀 다른 종류의 확률이다.

4-3 여아가 태어날 확률을 그대로 조건부 확률로 사용한다

이번 단계에선 타입별로 특정행동을 초래하는 조건부확률을 설정해야 함. 이번에는 타입 그 자체가 그 조건부 확률이다.

도표4-2 그 부부가 여아/남아를 낳을 조건부 확률

타입	여아를 낳을 확률	남아를 낳을 확률
P=0.4	0.4	0.6
P=0.5	0.5	0.5
P=0.6	0.6	0.4

여지껏 조건부확률은 원인이 특정되어 있을 때 결과의 확률을 구한 것. 여기서
원인은

- ‘여아를 낳기 쉽다’
- ’남아를 낳기 쉽다’이며

결과는

- ‘여아가 태어난다’
- ’남아가 태어난다’이다.

도표4-3 여섯 개로 분기된 세계

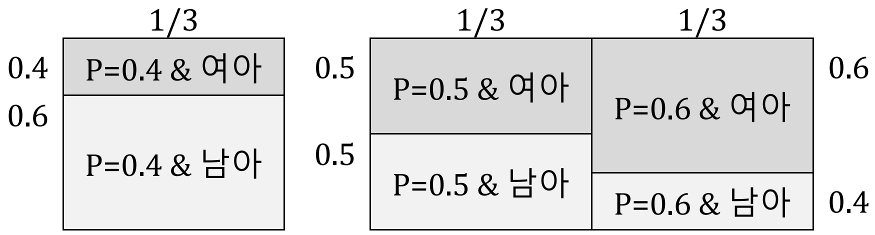


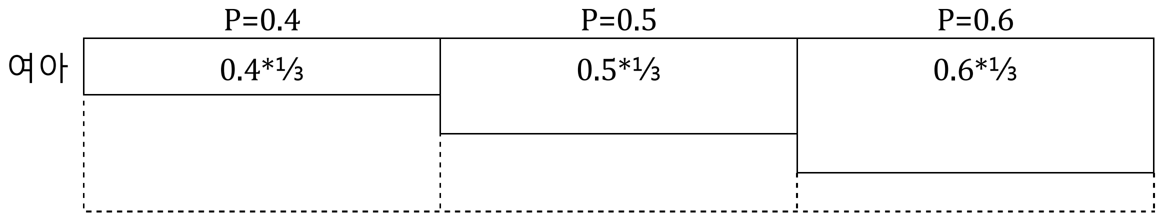
도표4-4 여섯 가지 세계의 확률

	P=0.4	P=0.5	P=0.6
여아	0.4*⅓	0.5*⅓	0.6*⅓
남아	0.6*⅓	0.5*⅓	0.4*⅓

4-4 첫째 아이가 여아였기 때문에 ‘일어날 가능성이 없는 세계’를 소거한다

이 특정 부부는 이미 첫째가 여아라는 결과를 갖고 있어 첫째가 남아라는 세계는 사라진다

도표4-5 정보에 따라 가능성이 한정된다



이부부의 첫째가 여아라는 결과가 관측되었기 때문에 가능성은 6개→3개로 좁혀진다. 세 개의 비례관계를 유지한 채 확률의 합이 1이 되도록 정규화 조건을 회복시킨다.

=0.4*⅓ : 0.5*⅓ : 0.6*⅓
= 0.4 : 0.5 : 0.6
= 4 : 5 : 6
= 4/15 : 5/15 : 6/15

(확률 p=0.4의 사후확률) = 4/15 = 0.27
(확률 p=0.5의 사후확률) = 5/15 = 0.33
(확률 p=0.6의 사후확률) = 6/15 = 0.4

4-5 베이즈 추정의 프로세스 정리

도표4-6 부부의 타입에 대한 베이즈 추정 프로세스

1. 여아가 태어날 확률p에 타입을 세 가지로 정하고 사전확률을 설정(이유불충분의 원리를 이용해 대등하게 설정)
2. 여아가 태어날 조건부 확률을 설정(타입p가 그대로 조건부 확률이 된다)
3. 첫째가 여아였다는 사실을 관측
4. 가능성을 소거
5. 타입p의 확률을 정규화
6. 타입p의 사후확률(베이즈 역확률)

도표4-7 부부의 타입에 대한 베이즈 갱신

1. 타입p=0.4의 사전확률=0.33
   타입p=0.5의 사전확률=0.33
   타입p=0.6의 사전확률=0.33
3. 첫째 아이는 여아였다

6. 타입p=0.4의 사전확률=0.27
   타입p=0.5의 사전확률=0.33
   타입p=0.6의 사전확률=0.40

처음에 여아가 태어나기 전 3가지 타입의 가능성을 대등하게 보고 확률을 0.33씩 정하였다. 그러나 첫째 아이가 여아라는 정보에 의해 사후확률의 비율이 달라진다. 여아가 태어났다는 정보를 얻은 후 여자아이를 낳기 쉬운 부부라고 추정결과가 달라진다.

이번 예에는 객관확률과 주관확률이 혼재

- 객관확률 – 타입을 나타내는 확률 p
- 주관확률 – 사전확률과 사후확률(추정자의 마음에 의거한 것)

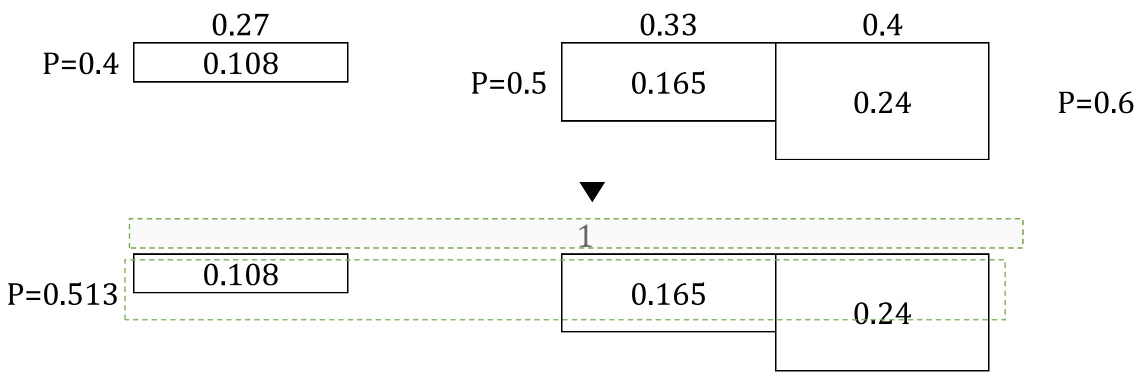
4-6 다음에 여아가 태어날 확률을 구하려면 기대치를 사용한다

(타입p=0.4의 사전확률)=0.27
(타입p=0.5의 사전확률)=0.33
(타입p=0.6의 사전확률)=0.40

이 부부로부터 태어날 둘째 아이가 여아일 확률을 하나의 수치로 구하려면 '평균치'를 사용한다

(p의 기대치) = 0.4*0.27 + 0.5*0.33 + 0.6*0.4 = 0.513

도표4-8 타입의 평균치를 계산한다



제4강의 정리

1. 타입을 확률로 설정하고 그에 대한 사전확률을 설정한다(데이터를 얻을 수 없으므로 이유 불충분 원리를 채용하여 대등하게 설정한다) 사전확률은 확률의 확률이 된다
2. 조건부 확률을 설정한다(이것은 타입의 확률 그 자체로 설정하면 된다)
3. 얻은 정보(첫째 아이가 여아였다)로부터 일어나지 않을 가능성을 소거한다

4. 남의 세계의 확률값에 대한 정규화 조건을 복구한다
5. 타입에 대한 사후확률(베이즈 역확률)을 얻을 수 있다
6. 사전확률이 얻은 정보에 의해 사후확률로 갱신된다(베이즈 갱신)
7. 사전확률과 사후확률은 모두 주관확률이다
8. 각 타입(확률로 표현되고 있는)의 확률이 나왔으므로 그것을 평균화(기대치를 구한다)하여 타입의 평균치를 구할 수 있다. 그것이 다음에 태어날 아이가 여아일 확률이 된다.

/ 연습문제

column 베이즈는 어떤 사람이었을까?

제5강 추론의 프로세스에서 부각되는 베이즈 추정의 특징

5-1 알고 보면 표준통계학보다 오랜 역사를 지닌 베이즈 통계학

지금의 통계학-학교에서 배운 통계학은 네이만-피어슨이 완성-19세기 말~20세기초에 완성
베이즈 통계학은 18세기 완성

5-2 추론이란 무엇인가?

추론 - 명확하지 않은 사항에 대해 몇가지 증거를 바탕으로 추리하여 그 사실을 밝혀내려는 행위

예제

- 명확하지 않은 사항
  눈 앞에 단지가 두 개 있고 둘중 하나가 A단지 또는 B단지라는 것 알지만 겉모습으론 알기 어려움
- 지식
  A 단지에는 열개의 흰색 공이 있고 B단지에는 열개의 검정색 공이 있다
- 추측을 위한 증거
  아무 단지에서 공을 꺼냈더니 검정색이다
- 그렇다면 이 단지는 A일까 아니면 B일까?

5-3 논리적 추론의 프로세스

지식에 따른 사실관계를 요약하면

- 사실1 - A 혹은 B
- 사실2 - A라면 흰 공
- 사실3 - B라면 검은 공
- 사실4 - 검은 공(흰 공이 아니다)

누구나 쉽게 꺼낸 단지가 B임을 알 수 있지만 수학적인 증명(논리적인 연역)에 의한 추론이 필요
대표적인 증명방법 -자연연역
결과가 A 라고 가정한다. 가정과 사실2로부터 흰 공으로 결론이 나지만 사실4에서 검은 공이므로 가정 A 는 부정되고 B로 결론

5-4 확률적 추론의 프로세스

단지 A 에는 흰 공9개+검은 공1개
단지 B 에는 흰 공2개+검은 공8개
공을 1개 꺼냈더니 검은 공. 그렇다면 어느 단지에서 공을 꺼낸 것인가?

위의 문제는 앞서의 논리적 추론이 통하지 않음. 다시 요약하면

- 사실1 - A 혹은 B
- 사실2 - A라면 대체로 흰 공
- 사실3 - B라면 대체로 검은 공
- 사실4 - 검은 공(흰 공이 아니다)

결론은 '대체로 B일 것이다'
위의 '대체로'라는 해석에 표준통계학과 베이즈통계학의 입장이 다름

- 표준통계학 추정에서 '대체로 B일 것이다'를 리스크는 있지만 B로 결론짓자라는 의미
- 베이즈통계학에서 A와 B 모두 가능하지만 B의 가능성이 훨씬 클 것이다라는 입장

제5강의 정리

1. 논리적 추론(자연연역)은 논리학의 연역법에 따라 엄밀하게 결론을 도출하는 것
2. 지식으로서 가지고 있는 사실에 불확실한 부분이 있을 때는 확률적 추론이 된다
3. 확률적 추론에서는 '대체로 **다'라는 결론이 나온다
4. 확률적 추론에는 표준통계학 추정과 베이즈 추정의 두 가지 방법이 있다
5. 표준통계학추정에서는 일정한 리스크를 감수하고 '**다'와 같은 형식으로 하나의 결론을 내린다
6. 베이즈 추정에서는 '양쪽 다 가능하지만 **의 가능성이 더 높다'라는 형식으로 양다리에 걸친 결론을 도출한다.

/ 연습문제

제6강 명쾌하고 엄밀하지만 쓸 데가 한정된 네이만-피어슨식 추정

6-1 네이만-피어슨 식 추정으로 단지 문제를 풀어보자

앞서 네 가지로 정리한 지식에서

- 사실1 - A 혹은 B
- 사실2 - A라면 대체로 흰 공
- 사실3 - B라면 대체로 검은 공
- 사실4 - 검은 공(흰 공이 아니다)

사실2와 3의 '대체로'라는 말 때문에 논리적인 추론을 사용할 수 없었다. 그런데 '대체로'라는 말을 확률적 수치로 바꾸고 이 일정기준을 만족하면 '잘못된 판단' 리스크를 감수한다면 가능하다. 네이만-피어슨 통계학에선 확률10%를 판단을 그르칠 리스크이며 이 방법으로 계속 추정해나가면 10% 확률로 잘못된 결론을 내리게 된다.

6-2 가설검정의 프로세스

네이만-피어슨의 추론방법이 가설검정. 대략의 가설검정 수순

1. 검정하려는 가설 A를 세운다. 이 가설 A를 '귀무가설'이라 한다
2. A가 아닌 경우 결론지을 B를 준비한다. 이 가설B를 '대립가설'이라 부른다
3. A가 옳다는 가정하에, 작은 확률 α 로 밖에 관측되지 않는 현상X를 생각한다
4. 현상X가 관측되었는 가 확인한다
5. 현상X가 관측된 경우 귀무가설A가 틀렸다고 판단. 귀무가설A를 기각하고 대립가설B를 채택한다
6. 현상X가 관측되지 않는 경우 귀무가설A를 기각할 수 없으며 귀무가설A를 채택한다

위의 프로세스를 요약하면, A가 옳을 경우 α라는 낮은 확률로 밖에 일어나지 않는 현상이 실제로 관측되었을 때 , A가 원래 잘못된 것이라고 판단하여 가설A를 버린다. 관측되지 않았을 때는 버릴 이유가 없으므로 유지한다. 여기서 α를 유의수준이라 한다.

6-3 가설검정에서는 판단을 내리지 않는 사례도 있다

가설검정은 명쾌한 방법, 널리 사용됨. 핵심은 유의수준 α를 얼마로 정하는 가가 중요. 유의수준 α는 거의 관측되지 않는 현상의 확률. 보통은 5% 또는 1%로 설정한다(이런 설정 수치도 과학적 근거는 없다)

제6강의 정리

1. 표준 확률적 추론은 네이만-피어슨 통계학에 따른 것이다
2. 먼저 귀무가설과 대립가설을 설정한다
3. 유의수준α를 설정한다. 보통은 α=0.05 혹은 α=0.01
4. 귀무가설하에서 유의수준α이하로만 관측되는 현상X에 주목한다
5. 현상X가 관측되었다면 귀무가설을 기각하고 대립가설을 채택한다
6. 현상X가 관측되지 않았다면 귀무가설을 채택한다
7. 가설검정은 유의수준α의 확률로 잘못될 수 있는 리스크를 가지고 있다

제7강 베이즈 추정은 적은 양의 정보로 그럴듯한 결론을 이끌어낸다

네이만-피어슨 식 추정과 다른 점

7-1 베이즈 추정으로 단지 문제를 푼다

베이즈 추정에선 유의수준과 같은 개념을 필요하지 않다

7-2 단지A와 단지B를 타입으로 설정한다

문제설정

> 눈앞에 단지가 하나 있는 데, 단지A나 B중 하나임을 알고 있지만 겉으로 봐서는 어느 쪽인지 알 수가 없다. 단지 A에는 흰 공 아홉 개와 검은 공 한 개가 들어 있고, 단지B에는 흰 공 두개와 검은 공 여덟 개가 들어 있다는 정보를 가지고 있다. 이때 단지에서 공을 한 개 꺼냈더니 검은 공이었다. 눈앞에 있는 단지는 어느 것일까?

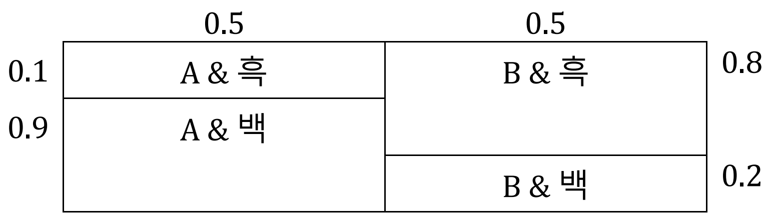
- 타입의 설정 : 단지A와 B
- 사전확률의 설정 : 이유불충분의 원리, 각각 0.5로 설정

도표7-1 이유불충분의 원리에 따른 사전분포

0.5	0.5
단지 A	단지 B

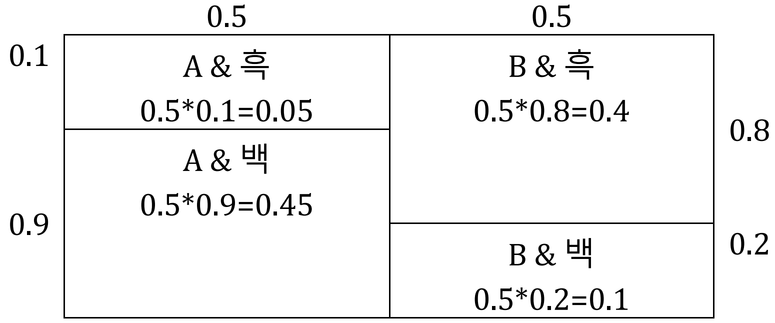
타입에 의존하여 검은 공/흰 공이 나올 조건부 확률설정
- 단지A인 경우 검은 공일 조건부 확률 0.1
- 단지A인 경우 흰 공일 조건부 확률 0.9
- 단지B인 경우 검은 공일 조건부 확률 0.8
- 단지B인 경우 흰 공일 조건부 확률 0.2

도표7-2 조건부 확률의 설정



네 가지의 가능세계(직사각형)에 각각의 확률을 계산

도표7-3 네 가지 가능성에 대한 확률계산



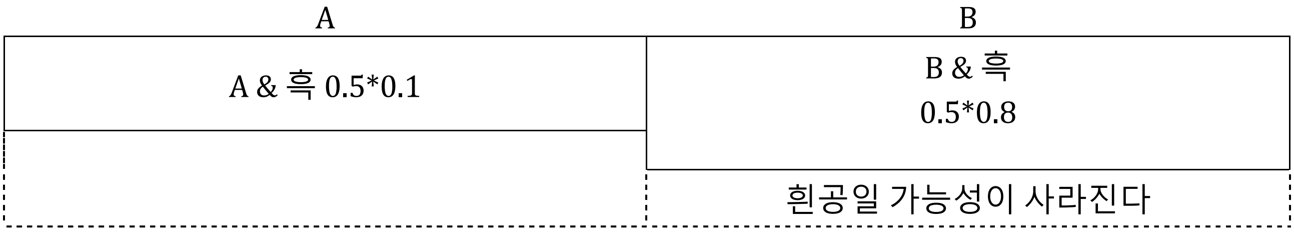
관측된 공이 검정이므로 흰 공이 속한 세계를 소거

검은 공이 관측된 세계에서 각 확률을 정규화

(단지가 A일 사후확률) : (단지가 B일 사후확률)
= 0.5*0.1 : 0.5*0.8
= 1 : 8
=1/9 : 8/9

단지B라고 판단하는 것이 타당

도표7-4 2가지 가능성이 소멸



7-3 베이즈 추정은 어떤 환경에서도 '일단' 추정을 내린다

유의수준 설정이 없어 어느 환경에서도 일단 추정이 가능. 네이만-피어선식은 어느 한 쪽으로 판단을 내리는 것인 반면 베이즈는 양쪽의 가능성을 남겨두고 그 가능성의 비율관계를 제시

수치를 계산하는 것은 통계 전문가의 일이고, 베이즈의 추정을 두고 판단을 내리는 것은 사장의 재량 그래서 **베이즈 추정은 사장의 확률**이라고 함

7-4 베이즈 추정과 네이만-피어슨 식 추정에서 서로 다른 '리스크'의 의미

- 네이만-피어슨 식 추정에서 유의수준은 리스크의 지표-5%로 설정하여 계속 가설검정을 반복하면 5%의 확률로 잘못된 결론을 내린다
- 베이즈 추정에서는 사후확률이 리스크

- 베이즈 추정이 유의수준이 필요없는 것은 사전확률 덕분
- 사전확률은 주관적
- 항상 자의성이 있으면 책임은 통계전문가의 판단으로 남는다

도표7-5 단지에 대한 베이즈 갱신

A일 사전확률 = 0.5, B일 사전확률 =0.5
▼
검은 공이 관측되었다
▼
A일 사후확률=1/9=0.11, B일 사후확률=8/9=0.89

7-5 논리적인 관점에서 본 베이즈 추정의 프로세스

- 사실1 - A 혹은 B
- 사실2 - A라면 대체로 흰 공
- 사실3 - B라면 대체로 검은 공
- 사실4 - 검은 공(흰 공이 아니다)

위의 사실에서 베이즈 추정이 추리하는 방식

- 사실2에서 문제의 단지가 A라고 가정하면, 관측된 공이 검은 공 또는 흰 공 모두 가능하지만 흰 공이 많아 '대체로' 흰 공이라는 결론
- 사실3에서 문제의 단지가 B라고 가정하면, 검은 공 또는 흰 공 모두 가능하지만 대체로 검은 공이라는 결론 도출
- 사실4에서 A의 흰 공과 B의 흰 공이 소거 되고 검은 공만 남음. A의 검은 공은 B 의 검은 공에 비해 가능성이 낮아 B라는 결론

제7강의 정리

1. 단지가 A인가 B인가를 타입으로 설정한다
2. 이유불충분의 원리에 따라 A의 사전확률을 0.5, B의 사전확률을 0.5로 설정한다
3. A에 든 검은 공의 조건부 확률을 0.1, 흰 공의 조건부 확률을 0.9로 설정하고, B에 든 검은 공의 조건부 확률을 0.8, 흰 공의 조건부 확률을 0.2로 설정한다
4. 관측된 공이 검은 공이라는 사실에 따라 흰 공일 가능성을 소거한다
5. 검은 공의 확률에 대해 정규화 조건을 복구한다
6. A일 사후확률과 B일 사후확률이 구해지고, '대체로 B일 것이다'라는 결론을 내린다

/ 연습문제

제8강 베이즈 추정은 '최우원리'에 근거해 있다

베이즈통계학과 네이만-피어슨 통계학의 접점

8-1 베이즈통계학과 네이만-피어슨 통계학의 공통점

베이즈통계학에서는 네이만-피어슨 통계학에서 설정하지 않는 사전확률이라는 것을 도입했다

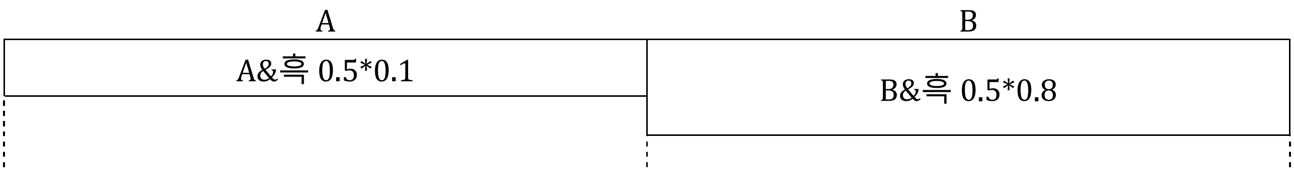
8-2 여러 학문에서 두루 쓰이는 '최우원리'

- 베이즈통계학과 네이만-피어슨 통계학의 공통된 발상- 최우원리
- '최우원리'란 '세상에서 일어나는 일은 일어날 확률이 큰 것이다'라는 원리

8-3 베이즈 추정은 최우원리에 근거하고 있다

- 단지 A에서 흰 공이 관측될 확률이 매우 크다.
- 단지 B에서 검은 공이 관측될 확률이 매우 크다.
- 그리고 검은 공이 관측되어 '단지는 B일 것이다'라는 판정을 내렸다
- 이것은 결과의 확률을 가장 높이는 원인을 선택하고 있기 때문

도표7-4 2가지 가능성이 소멸



앞서 사후확률 도출에서 검은 공이 관측되어 흰 공의 세계, 2가지가 소멸되고 B의 사후확률이 A보다 훨씬 커서 결론은 '단지는 B일 것이다'로 내렸다. 이는 검은 공이라는 현상이 관측될 확률을 크게 만드는 원인 B가 선택된 것과 마찬가지이다. 즉 최우원리가 적용되는 순간이다.

- 사후확률은 (사전확률)*(조건부 확률)에 비례

사전확률이 크거나 조건부 확률이 큰 원인으로 선택되기 쉽다. 이것이 최우원리이다

8-4 네이만-피어슨 통계학도 최우원리에 근거하고 있다

표준통계학에서는 추정이 아니라 '통계적 추정을 입증'하는 데 최우원리를 도입하고 있다.

통계적 추정 입증- 통계학에서 추정을 할 때 '왜 그렇게 생각하는 가', '그렇게 생각하는 것이 어떤 이점을 가져다 주는가'를 설명하는 것.

N회 관측하여 x회 일어날 경우 최우추정량이 x/N이 된다. 최우원리는 평균과 연결되어 있다.

제8강의 정리

1. 최우원리란 관측된 현상이 일어날 확률이 가장 커지는 원인을 채용하는 원리이다
2. 베이즈통계학의 사전확률은 최우원리를 응용한 한 가지 형태라고 볼 수 있다
3. 표준통계학의 점추정에서는 관측된 현상의 확률을 최대한으로 하는 함수를 추정치로서 채용한다. 이것도 최우원리를 응용한 것이다
4. 표준통계학과 베이즈통계학에는 최우원리라는 공통된 사상이 내재해있다

/ 연습문제

제9강 베이즈 추정은 때로 직감에 크게 반한다②

몬티 홀 문제와 세 죄수 문제

9-1 베이즈 역확률의 패러독스

사전확률에 주관성이 결부된다 ⇨ 수학과 철학의 경계선상의 이론 ⇨ 특수한 설정하에 베이즈추정은 상식에 반하는 결과 도출

9-2 패러독스#1 몬티 홀의 문제

몬티 홀 문제

당신은 세 커튼 A, B, C 앞에 서 있다. 셋중 어느 하나의 커튼 뒤에 자동차가 상품으로 숨겨져 있다. 당신은 셋중 하나를 골라 자동차가 나오면 그것을 상품으로 받게 된다. 이때 당신인 커튼 A 를 고르자 사회자는 선택하지 않은 커튼중 B 를 열어 보이면 '여기에는 자동차가 없습니다'라고 말한다. 그리고 '남은 커튼은 당신이 선택한 A 와 나머지 C, 이렇게 두 가지입니다. 당신은 지금이라도 커튼을 바꿔 선택할 수 있는데, 어떻게 하시겠습니까?' 라고 묻는다.

해답은 '선택을 바꾸는 것이다'. C 커튼에 자동차가 있을 확률은 A 보다 크기 때문이다. 그러나 이에 대한 반론도 있다. 어차피 자동차가 숨겨져 있을 가능성이 있는 커튼은 두 개중 하나이므로 확률은 반반이다. 어느 것을 선택하든 확률은 달라지지 않는다.

9-3 패러독스#1 세 죄수문제

세 죄수 문제

죄수 알란, 버나드, 찰스가 있다. 셋중 둘은 처형되고 한 명은 석방된다. 그러나 누가 석방될지는 모른다. 이때 알란은 간수에게 말을 건다. '세 명중 두 명은 처형될테니 나를 제외하고 누가 처형될지 가르쳐 줘도 나쁠 것 없다' 간수는 일리가 있다고 판단되어 '버나드가 처형된다'라고 알려준다. 그러자 알란은 기뻐했다. 그는 다음과 같이 생각했기 때문이다. '아무 것도 모르는 상태에서 내가 석방될 확률은 1/3 이다. 그러나 버나드가 처형된다는 것을 안 이상 , 나와 찰스중 하나는 처형되고, 하나는 석방된다. 따라서 내가 석방될 확률은 1/2 로 높아졌다고

패러독스#1 과 #2 는 같은 구조의 문제. 알란, 버나드, 찰스는 커튼 A, B, C 이고 석방은 자동차이다. 간수가 버나드가 처형될 것이라고 알려준 것은, 사회자가 커튼 B 을 열어 자동차가 없음을 보여준 것에 해당된다. 그리고 커튼 A 에 자동차가 있다는 것이 알란이 석방되는 것에 해당된다.

알란의 논리가 이상한 점은, 간수가 알란이 아닌 사람이 처형될 것으로 알려주었다는 것만으로 알란이 석방될 확률이 1/3 ⇨ 1/2 로 높아진다는 점이다.

버나드가 아니라 찰스가 처형된다고 바꾸어도 결과는 같다. 그렇다면 누가 처형될지 묻거나 가르쳐줄 필요도 없으며, 알란은 자신이 석방된 확률이 1/2 이라고 추정하면 되기 때문이다.

패러독스 #1 과 #2 는 동전의 앞/뒷면 관계이다. 한 쪽에 수긍하지 못하면 다른 쪽에 수긍해야 한다.

9-4 본질적으로 같은 문제이다

두 문제에서 공통된 포인트 - 정보의 입수에 따라 확률이 달라진다
‘이전의 예에서는 정보에 의해 확률이 달라지는 것이 핵심’인데, 이는 사전확률과 사후확률로 대표되는 베이즈 추정의 원리.
패러독스 #1 과 #2 는 정보에 따른 확률의 변화가 여러 사람들의 직관에 반하는 예이다.

몬티 홀 문제에서 사회자가 커튼 B 를 보여준 후 확률(커튼 A 에 자동차가 있을 확률)이 달라지는가(1), 그대로인가(2), 두 가지의 생각 방법

1. 사고법 #1 - A 와 B 의 확률 모두 ⅓ -> ½
2. 사고법 #2 - A 의 확률 ⅓ -> ⅓, C 의 확률 ⅓ -> ⅔ , 선택을 바꾸는 게 유리하다

패러독스 #1 과 #2 는 같은 문제임에도 패러독스 #1 에서는 사고법#2 가 맞고 패러독스 #2 문제에서는 사고법#1 이 맞다는 역설적인 판단을 한다.

많은 문헌에서 사고법#2 가 옳다고 판단. ‘선택자 자신에 대한 확률은 달라지지 않으며 선택자가 관여하지 않는 측의 확률이 변화한다’

여기서 다루는 확률은 주관적이며 정답은 존재하지 않는다. 왜냐면 당신이 이미 선택한 시점에서 자동차나 복권 1 등은 결정된 것이며, 변화하는 것은 ‘당신의 주관적인 추측’이기 때문이다.
주관이므로 정답은 유일하다고 단정할 수 없다.

9-5 베이즈 추정으로 패러독스에 접근한다

- 타입과 사전확률 설정
  - A - 커튼 A 에 자동차
  - B - 커튼 B 에 자동차
  - C - 커튼 C 에 자동차

사전확률은 모두 ⅓ 으로 설정

도표 9-1 이유불충분 원리에 따른 사전분포

1/3	1/3	1/3
A	B	C

조건부확률 설정

당신이 A 를 선택하면 사회자가 B 와 C 중 어느 것을 열 것인가에 대한 조건부확률
만일 A 에 자동차가 있다면 사회자는 B 와 C 를 대등한 확률 ½ 확률로 연다. 만일 B 에 자동차가 있다면 1.0 의 확률로 C 를 연다. 만일 C 에 자동차가 있다면 1.0 의 확률로 B 를 연다.

조건부확률을 도입하면 네 개로 분기된 세계가 된다

도표 9-2 조건부 확률의 설정

1/3	1/3	1/3

<table>
<tr><td>1/2</td><td>A&B 열기</td><td rowspan="2">B&C 열기</td><td rowspan="2">C&B 열기</td><td rowspan="2">1</td></tr>
<tr><td>1/2</td><td>A&C 열기</td></tr>
</table>

- A&B 열기 - A 선택 & A 자동차 & B 열기
- A&C 열기 - A 선택 & A 자동차 & C 열기
- B&C 열기 - A 선택 & B 자동차 & C 열기
- C&B 열기 - A 선택 & C 자동차 & B 열기

이때 당신은 사회자가 B 를 연 행위(B 열기)를 보고 B 에 자동차가 없음을 알게 된다. '커튼 B 에 자동차 있음'이라는 것이 없어지므로 이와 관련된 세계 B&C 열기는 소멸된다. 그리고 A&C 열기도 소멸된다. A&C 열기는 'A 선택 & B 자동차 & C 열기'이다. 따라서 남은 그림은 도표 9-3 이다.

도표 9-3 일어날 가능성이 없는 세계의 소거

A		C
A&B 열기 1/3 * 1/2		C&B 열기 1/3*1

정규화에 따라 사후확률을 구하면
(A 일 사후확률) : (C 일 사후확률)
= ⅓ * ½ : ⅓ *1
= 1 : 2
= ⅓ : ⅔

위의 해법은 세 죄수 문제에도 똑같이 적용할 수 있어 알란이 석방될 확률은 ⅓ , 찰스가 석방될 확률은 ⅔ 이다.

결국 '사회자나 간수가 질문자에 관련된 정보를 주지 않았으므로 질문자에 대한 사후확률은 바뀌지 않는다'

9-6 모델의 설정 자체로 결론이 달라진다

사회자가 연 커튼에 대한 조건부 확률의 설정에 자의성이 있다

참가자가 A 를 선택하고 A 에 자동차가 있을 때 사회자가 B 와 C 둘 중 고민하지 않고 B 를 열기로 미리 정해두었다면 도표 9-2 는 도표 9-4 로 바뀐다

도표 9-4 조건부 확률의 설정

	1/3	1/3	1/3	
1	☺ A&B 열기 🚐	☺ B&C 열기	☺ C&B 열기	1

도표 9-5 일어날 가능성이 없는 세계의 소거

A		C
A&B 열기		C&B 열기

1/3 * 1	1/3 * 1

A 와 C 의 사후확률은 ½ 이 된다.
이 결론은 사고법#1 의 결론과 같다.

사회자가 C 을 열기로 미리 정했다고 해도 결론은 동일하다. 그렇다면 B 와 C 를 대등하게 다루어야 한다. 그러나 이것은 이유불충분의 원리를 조건부확률에 까지 확장하는 것으로 통상의 베이즈 추론에서 벗어난 것이다.

확률적 추론이라는 것은 어디까지나 확률현상의 원리를 어떻게 상상할 것인가 하는 '주관'에 의존하므로 모델을 어떻게 설정하느냐에 따라 결론이 달라진다.

제9강의 정리

1. 몬티 홀 문제와 세 죄수 문제는 같은 것을 다른 형식으로 기술한 것이다
2. 한쪽이 이상하다고 생각한다면 다른 쪽도 받아 들여야 한다
3. 두 문제 모두 베이즈 추정으로 풀이하여 답을 구할 수 있다
4. 결론 모델의 설정(확률 현상을 어떻게 상상할 것인가)에 의존하여 도출되기 때문에 정답이란 존재하지 않는다

/ 연습문제

column '속설'에 대한 두 가지 법칙

제10강 복수의 정보를 얻었을 때의 추정

'독립시행 확률의 승법공식'을 사용한다

10-1 복수의 정보를 바탕으로 베이즈 추정을 실시한다

추정이란 일반적으로 복수의 정보로부터 이루어지는 법
복수의 정보를 얻었을 때 추정을 알아본다

10-2 두 종류의 시행을 조합하려면

- 시행 – 여러 개의 **현상**의 귀결 가능성이 있어 각각의 가능성에 확률을 할 수 있을 때의 그 현상
- 시행 – 주사위를 던져 나온 눈을 **확인하는 것**
- 시행 – 내일의 날씨가 맑음, 구름, 비, 눈의 네 가지 귀결중 **무엇이 될 것인가를 보는 것**

두 종류의 시행을 묶어 또 다른 시행으로 보고 확률을 생각해보기

제1시행 – 동전을 던져 앞뒤를 귀결로 하는 확률현상
제2시행 – 주사위를 던져 나올 눈을 귀결로 하는 확률현상
제3시행 – 제1시행의 귀결 '앞면' + 제2시행의 귀결 '4' = '앞면&4' ⇨ 직적시행
위 직적시행의 귀결 도표10-1과 같이 12개

도표10-1 두 가지 시행의 묶음을 만든다

동전 던지기 시행의 귀결

앞
뒤

주사위 던지기 시행의 결과

1	2	3	4	5	6

▼

직적시행의 귀결

앞&1	앞&2	앞&3	앞&4	앞&5	앞&6
뒤&1	뒤&2	뒤&3	뒤&4	뒤&5	뒤&6

10-3 독립적 직적시행의 확률은 곱셈으로 구할 수 있다

시행의 독립성 ⇨ 두 개의 시행이 독립 ⇨ 한쪽 시행의 귀결이 다른 쪽 시행의 귀결에 영향을 주지 않음(반대의 경우는 종속시행)

도표10-2 독립시행에서의 면적

6개 모두 동일한 면적

앞&1	앞&2	앞&3	앞&4	앞&5	앞&6	2개는
뒤&1	뒤&2	뒤&3	뒤&4	뒤&5	뒤&6	같은 면적

6개 모두 동일한 면적

- 격자모양에 늘어선 직사각형 12개의 면적은 모두 같다
- 묶음의 확률은 각 확률의 곱이 된다

10-4 독립시행의 확률에 승법공식

동전과 주사위의 앞/뒤, 1~6 모두는 같은 확률로 일반적으론 같은 확률은 아니다.

- 제1시행의 귀결 : a, b, c, d
- 제2시행의 귀결 : x, y, z
- 그러나 각각 일어날 확률이 같다고 단정할 수 없다

도표10-3 두 가지 시행이 독립적인 경우의 직적시행

	a확률	b확률	c확률	d확률
x확률	a&x	b&x	c&x	d&x
y확률	a&y	b&y	c&y	d&y
z확률	a&z	b&z	c&z	d&z

a확률	b확률	c확률	d확률
귀결a	귀결b	귀결c	귀결d

x확률	귀결x
y확률	귀결y
z확률	귀결z

10-3의 첫 그림에서 두 개의 시행의 결합에 따른 직적시행의 12개의 확률을 보여주고 있음

10-3의 두 번째 그림은 직적시행중 하나의 행을 빼어 본 것으로 어느 행을 빼도 a, b, c, d 확률의 비례관계가 동일하다.

(a확률) : (b확률) : (c확률) : (d확률)

10-3의 세 번째 그림은 직적시행중 하나의 열을 빼어 본 것으로 어느 열을 빼도 x, y, z 확률의 비례관계가 동일하다.

(x확률) : (y확률) : (z확률)

독립시행 확률의 승법공식

(a&x확률) = (a&x직사각형 면적)= (a확률)* (x확률)

(b&z확률) = (b&z직사각형 면적)= (b확률)* (z확률)

제10강의 정리

1. 두 가지 시행을 묶은 직적시행은 직사각형을 격자모양에 분할하여 그림으로 표시한다
2. 두 가지 시행이 독립되어 있다는 것은 직관적으로 생각할 때 한 쪽의 귀결이 다른 쪽 귀결이 일어나는 데에 영향을 주지 않는 것이다
3. 두 가지 시행이 독립되어 있을 때 다음과 같은 확률의 승법 공식이 성립한다

   (제1시행의 귀결이 a이고 , 제2시행의 귀결이 x)일 확률

   =(a확률) * (x확률)

/ 연습문제

제11강 복수의 정보를 얻었을 때의 추정②

스팸메일 필터의 예

11-1 스팸메일 필터는 베이즈 추정에 기원하다

- 확률적 추론(통계적 추정, 베이즈 추정)에는 복수의 정보가 사용됨
- 정보가 많은 수록 신빙성이 높아짐
- 복수의 정보를 사용하는 베이즈 추정방식 – 확률의 승법공식

11-2 필터에 '사전확률'을 설정한다

- 이유불충분의 원리를 적용하여 스팸메일/ 일반메일의 사전확률을 각각 0.5로 설정

11-3 스캔할 글자나 문구와 그 '조건부 확률'을 설정한다

1차 검출포인트 - 메일 내용중 URL 링크여부

(확고한 관계)

관측	추정
스팸메일 -> URL링크가 있다	URL링크가 있다-> 스팸메일
일반메일 -> URL링크가 없다	URL링크가 없다-> 일반메일

(현실적 관계)

관측	추정
스팸메일 -> URL링크가 있다	URL링크가 있다-> 대체로 스팸메일
일반메일 -> URL링크가 없다	URL링크가 없다-> 대체로 일반메일

'**대체로**'를 수치로 평가하는 것이 베이즈 추정의 역할

스팸메일과 일반메일에 어느 정도의 비율로 URL이 있는 가를 설정(도표11-2,11-3)

도표11-2 링크가 걸려 있을 조건부 확률

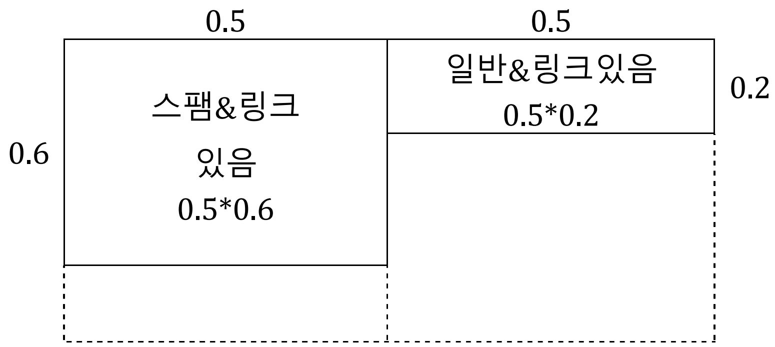
타입	URL 있을 확률	URL 없을 확률
스팸	0.6	0.4
일반	0.2	0.8

도표11-3 4개로 분기된 세계

11-4 스캔결과 스팸메일의 '베이즈 역확률'이 구해진다

(관측) 스캔한 결과 '링크'가 있었다 ⇨ 링크없음 세계 2개가 없어짐

도표11-4 가능세계가 2개로 한정된다



(스팸메일일 사후확률) : (일반메일일 사후확률)
= 0.5*0.6 : 0.5*0.2
= 0.6 : 0.2
= 3: 1
= 3/4 : 1/4

이로부터 필터는 다음과 같이 판정

(링크가 있다는 조건에서 스팸일 사후확률)= 3/4 = 0.75

도표11-5 스캔 전과 스캔 후

스캔 전 스팸일 확률 0.5 – 스캔(링크가 있음을 확인) → 스캔 후 스팸일 확률 0.75

11-5 두 번째 정보로 인해 세계는 여덟 개로 나뉜다

도표11-6 링크가 붙어 있을 조건부 확률

타입	‘만남’ 단어가 있을 확률	‘만남’ 단어가 없을 확률
스팸	0.4	0.6
메일	0.05	0.95

링크여부와 만남여부에 따라 세계는 8개로 분기

도표11-7 링크가 걸려 있을 조건부 확률

4개의 가능세계		
	‘만남’있음	‘만남’없음
‘링크’있음	‘링크’있음&‘만남’있음	‘링크’있음’&‘만남’없음
‘링크’없음	‘링크’없음&‘만남’있음	‘링크’없음’&‘만남’없음

스팸일 경우		
	‘만남’있음	‘만남’없음
‘링크’있음	0.6*0.4	0.6*0.6
‘링크’없음	0.4*0.4	0.4*0.6

일반일 경우		
	‘만남’있음	‘만남’없음
‘링크’있음	0.2*0.05	0.2*0.95
‘링크’없음	0.8*0.05	0.8*0.95

도표11-8 여덟 개로 분기된 세계의 확률

<table>
<tr><th></th><th>스팸(0.5)</th><th>일반(0.5)</th><th></th></tr>
<tr><td></td><td rowspan="3">0.5*0.6*0.4</td><td>0.5*0.2*0.05</td><td>링크있음&만남있음</td></tr>
<tr><td>링크있음&만남있음</td><td>0.5*0.2*0.95</td><td>링크있음*만남없음</td></tr>
<tr><td></td><td>0.5*0.8*0.05</td><td>링크없음&만남있음</td></tr>
<tr><td>링크있음&만남없음</td><td>0.5*0.6*0.6</td><td rowspan="2">0.5*0.8*0.95</td><td></td></tr>
<tr><td>링크없음&만남있음</td><td>0.5*0.4*0.4</td><td>링크없음&만남없음</td></tr>
</table>

링크없음&만남없음	0.5*0.4*0.6	

11-6 두 가지 정보로부터 ‘일어날 가능성이 없는 세계’를 소거한다

링크와 만남, 두 개가 모두 검출되는 경우 스팸메일일 확률을 계산해보자
이제 도표11-8의 8개의 세계중 2가지만 남고 나머지 6개는 사라진다.
도표11-9 스캔에 의해 가능성은 두 가지만 남는다

	스팸(0.5)	일반(0.5)	
링크있음&만남있음	0.5*0.6*0.4	0.5*0.2*0.05	링크있음&만남있음

(스팸메일일 사후확률) : (일반메일일 사후확률)
= 0.5*0.6*0.4 : 0.5*0.2*0.05
= 0.24 : 0.01
= 24 : 1
= 24/25 : 1/25

링크있음과 만남있음이라는 정보하에서 사후확률의 값은 다음과 같다
(스팸메일일 사후확률)
= 24/25 = 0.96

두 가지 정보를 이용한 베이즈 추정을 도식으로 나타내면 다음과 같다

도표11-10 스캔 전과 2회 스캔 후

스팸일 사전확률 : 0.5	스캔 →	스팸일 사후확률 : 0.75	스캔 →	스팸일 사후확률 : 0.95
	↑ 링크가 있는 지 확인		↑ 만남이 있는 지 확인	

제11강의 정리

1. 두 가지 정보를 사용한 베이즈 추정은 기본적으로 같은 방법이다
2. 사전확률의 타입을 두 개로 설정한 경우 두 개의 정보를 사용하면 세계는 8개로 분기된다

3. 8개의 세계 각각의 확률은 확률의 승법공식을 이용해서 구한다
4. 1개의 정보를 사용할 때보다 두 개의 정보를 사용할 때 스팸메일의 판정도가 높아진다

/ 연습문제

제12강 베이즈 추정에서는 정보를 순차적으로 사용할 수 있다

>>'축차합리성'

12-1 베이즈 추정에서는 이전 정보를 잊어도 앞뒤가 들어 맞는다

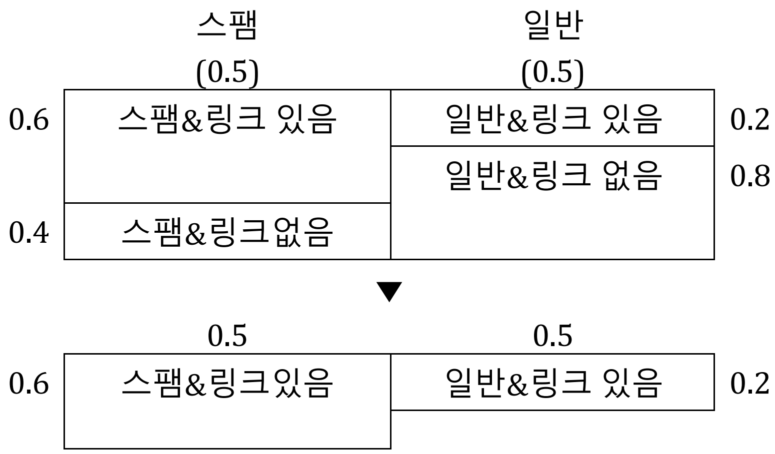
앞서 두 가지 정보를 바탕으로 사후확률을 계산. 요약하면 도표12-1과 같다.

도표12-1 두 가지 정보에 의한 베이즈 추정

타입에 대한 사전확률	정보1 →	타입에 대한 사후확률1	정보2 →	타입에 대한 사후확률2

- 축차추정 : 연속해서 들어오는 정보에 대한 연속적인 추정
- 교묘한 성질 : 정보1에서 타입에 대한 확률을 개정하면 정보2를 사용할 때는 앞의 정보1은 잊어도 된다 → 축차합리성

도표12-2 정보1단계의 정보에 따른 베이즈 추정



정보1이라는 전제하에

(스팸일 사후확률) : (일반일 사후확률)

= 0.3 : 0.1 = 0.75 : 0.25

12-2 정보1로부터 얻은 사후확률을 '사전확률'로 설정한다

- 재미있는 발상 : 지금 구한 사후확률을 타입에 대한 사전확률로 재설정

도표12-3 정보1로부터 얻은 사후확률을 '사전확률'로 설정

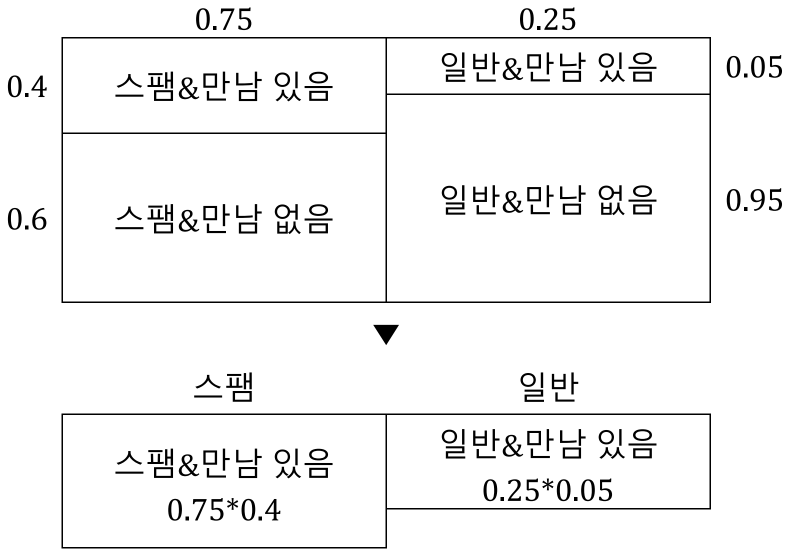
0.75	0.25
스팸	일반

- 이유불문하고 이제 스팸메일일 사전확률은 0.75, 일반메일일 사전확률은 0.25
- 이런 발상이 가능한 이유 : 본래 사전확률은 근거없이 설정되어 있는 값

12-3 정보2를 사용하여 베이즈 갱신을 한다

앞서 재설정된 사전확률을 바탕으로 제2의 정보, ‘만남’을 갖고 추정을 시도

도표12-4 정보2를 사용한 베이즈 추정에 의한 사후확률

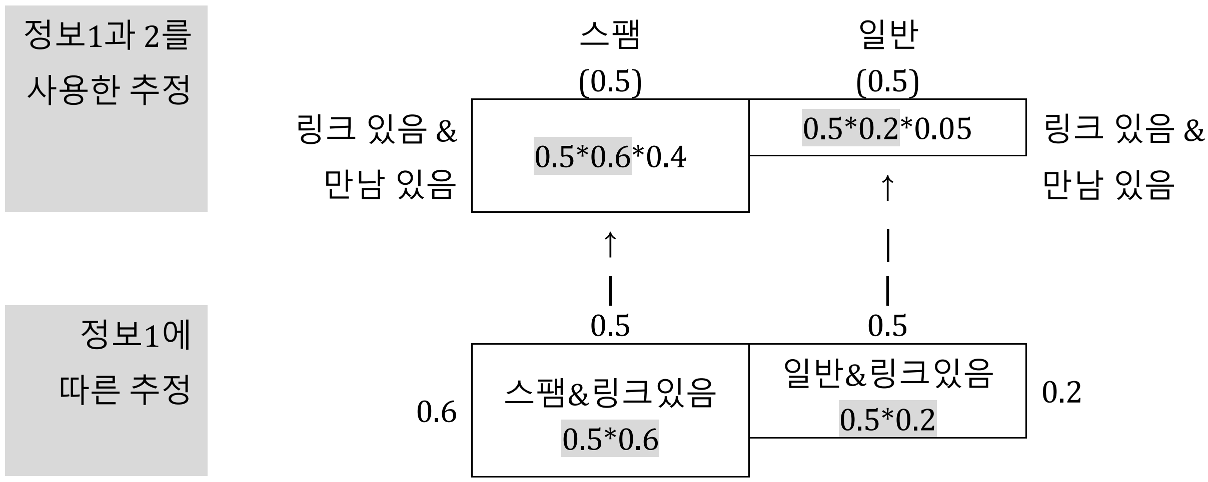


실제 ‘만남’이 검출되어 ‘만남 없음’이 있는 2개의 세계는 사라짐.

(스팸메일일 사후확률) : (일반 메일일 사후확률)
= 0.75*0.4 : 0.25*0.05
= 24 : 1
= 24/25 : 1/25

두 가지 정보(정보1 & 정보2)를 사용한 베이즈 추정의 사후확률과 동일→이유는 도표12-5

도표12-5 두 가지 정보에 따른 개정과 축차적 개정이 일치하는 이유



- 상단 그림 : 두 개의 정보를 이용해 사후확률을 한번에 구함
- 하단 그림 : 정보1로부터 각 타입의 확률을 개정하여 얻은 사후확률

위의 그림중 0.5*0.6 와 0.5*0.2 는 상단그림과 하단그림에서 중복된 것이다. 아래 그림의 비례관계에 정보2를 사용하여 베이즈 추정을 실시하면(도표12-4) 상단의 그림이 나온다.

정보1로부터 수정한 사후확률을 사전확률로 사용하고 여기에 정보2를 결합하여 구한 사후확률과 정보1과 정보2를 한 번에 사용해서 구한 사후확률이 일치한다.

12-4 베이즈 추정은 인간다운 추정이다

축차합리성 : 두 가지 정보를 한꺼번에 사용하여 추정한 결과와 첫 번째 정보를 사용하여 추정하고 , 그 추정결과를 사전확률로 두고 두 번째 정보를 사용하여 추정한 결과가 완전히 일치하는 성질

도표12-6 축차합리성

타입에 대한 사전확률	→	정보1+정보2	→	타입에 대한 사후확률
				↑
타입에 대한 사전확률	정보1 →	타입에 대한 사후확률		같다
잊다 ↓		↓		↓
		타입에 대한 사전확률	정보2 →	타입에 대한 사후확률

축차합리성이 성립한다는 것은 정보를 한번에 이용하지 않고 축차적으로 (순서대로) 이용해도 같은 결과를 얻을 수 있음을 의미. 즉 이전에 사용한 정보는 잊어도 된다
이는 베이즈 추정의 유용성을 보여주는 대목임
학습기능 : 한번 사용한 정보는 버려도 현재의 추정에 반영. 베이즈추정으로 개정된 타입에 대한 사후확률은 모든 정보를 활용한 내용이 됨. 정보를 입수하면 자동으로 똑똑해지는 인간다운 기능(정보→인상의 개정→정보의 망각을 반복하며 평가를 확고히 함)

제12강의 정리

1. 두 가지 정보를 한꺼번에 사용해서 구한 사후확률과 첫 번째 정보로 얻은 사후확률을 사전확률로 재설정하여 두 번째 정보를 이용해 개정한 사후확률은 항상 일치한다
2. 1의 성질을 축차합리성이라 부른다
3. 축차합리성은 학습기능의 일종으로 간주할 수 있다

4. 베이즈 추정에서 일단 추축에 사용한 정보는 버려도 문제되지 않는다

/ 연습문제

제13강 베이즈 추정은 정보를 얻을수록 더 정확해진다

13-1 ‘적당적당한’ 추측에서 더 ‘정확한’ 추정으로 만들려면

사장의 확률 : 대략으로 사전확률을 정한다(이유불충분의 원리로 사전에 아무 정보가 없어 대등한 사전확률을 설정) 그러나 정보가 적어도 추정이 가능하다는 게 장점. 또 일단 사후확률에 반영한 정보는 버려도 됨(베이즈 추정의 학습기능)

또 다른 베이즈 추정의 학습기능#2 : 정보가 많을수록 더 정확한 추정을 한다(도표13-1)

도표13-1 정보가 많으면 많을수록 더 정확한 추정이 이루어진다

정보없음		적은 정보		많은 정보
주관적인 추정	→	그 나름의 추정	→	거의 정확한 추정

13-2 단지 문제에서 공을 두 개 꺼낸다

문제설정

눈앞에 단지가 하나 있는 데, 단지A나 B중 하나임을 알고 있지만 겉으로 봐서는 어느 쪽인지 알 수가 없다. 단지 A에는 흰 공 아홉 개와 검은 공 한 개가 들어 있고, 단지B에는 흰 공 두개와 검은 공 여덟 개가 들어 있다는 정보를 가지고 있다.

7강과 달리 이번에는 맨 처음 꺼낸 공을 다시 단지에 넣고 새로 공을 한 개 뽑은 경우의 추정. 즉 첫 번째 공의 색상과 두 번째 공의 색상이라는 두 가지 정보를 사용한다. 두 번째 공이 검정색인 경우, 흰색인 경우를 추정.

이유불충분의 원리에 따라 타입 A와 타입B를 모두 0.5로 설정
첫 번째 공이 검은 공, 두 번째 공이 흰 공 인 경우 흑&백으로 표기(●○)

(흑&백(●○)일 확률) = (흑일 확률) * (백일 확률) ← 확률의 승법공식

두 개의 공이 모두 단지A에서 나온 것으로 추정하면

(흑&백(●○)일 확률) = (흑일 확률) * (백일 확률) = 0.1 * 0.9 = 0.09

두 개의 공이 모두 단지B에서 나온 것으로 추정하면

(흑&백(●○)일 확률) = (흑일 확률) * (백일 확률) = 0.8 * 0.2 = 0.16
이런 식으로 8개의 세계가 나온다

도표13-2 두 가지 정보로 인해 세계는 여덟 개로 나뉜다

	A(0.5)	B(0.5)	
흑&흑	0.5*0.1*0.1	0.5*0.8*0.8	흑&흑
흑&백	0.5*0.1*0.9		
백&흑	0.5*0.9*0.1		
백&백	0.5*0.9*0.9	0.5*0.8*0.2	흑&백
		0.5*0.2*0.8	백&흑
		0.5*0.2*0.2	백&백

13-3 두 번째도 검은 공이었을 때의 추정

두 번째 공이 검은 색, 즉 흑&흑이라면 나머지 세계는 사라진다

도표13-3 두 번째도 검은 공이었을 때의 추정

	A	B	
흑&흑	0.5*0.1*0.1	0.5*0.8*0.8	흑&흑

정규화를 하여 구한 사후확률은 다음과 같다.

(흑&흑일 때 A일 사후확률) : (흑&흑일 때 B일 사후확률)
= 0.5*0.1*0.1 : 0.5*0.8*0.8
= 1 : 64
= 1/65 : 64/65

단지가 B일 확률은 약 98%(=64/65)로 높아진다

도표13-4 검은 공을 두 번 꺼냈을 때의 추정

B일 사전확률: 0.5	검은 공 →	B일 사후확률: 8/9(0.89)	검은 공 →	B일 사후확률: 64/65(0.98)

첫 번째 공이 검은 색일 때 단지가 B일 사후확률은 0.89로 상승, 다시 두 번째에도 검은 공이 나오면 사후확률이 0.98로 상승

13-4 두 번째가 흰 공이었을 때의 추정

두 번째가 흰공이라면 도표13-2의 여덟 개의 세계중 흑&백의 세계만 남는다

도표13-5 두 번째가 흰 공이었을 때의 추정

	A	B
흑&백	0.5*0.1*0.9	

	0.5*0.8*0.2	흑&백

정규화를 하여 구한 사후확률은 다음과 같다.

(흑&백일 때 A일 사후확률) : (흑&백일 때 B일 사후확률)
= 0.5*0.1*0.9 : 0.5*0.8*0.2
= 9 : 16
= 9/25 : 16/25
= 0.36 : 0.64

도표13-6 첫 번째에 검은 공, 두 번째에 흰 공을 뽑았을 때의 추정

B일 사전확률: 0.5	검은 공 →	B일 사후확률: 8/9(0.89)	흰 공 →	B일 사후확률: 16/25(0.64)

첫 번째 공이 검은 색이어서 B일 확률은 높아졌으나 두 번째 공이 흰 색이 나와 0.64로 낮아져 B일 거라는 추측이 낮아졌다

13-5 최신 관측결과에 따라 결론이 달라진다

검은 공이 나오면 B의 사후확률이 높아지고, 다시 흰 공이 나오면 A의 사후확률이 높아진다. 매우 자연스러운 추정

도표13-7 정보로부터 추정결과가 어느 쪽으로 기우는가

사전	검은 공		검은 공		흰 공		
a=b=0.5	→	a↓, b↑	→	a↓, b↑	→	a↑, b↓	→

베이즈 추정의 축차합리성에 따라 (n+1)번째 사후확률을 계산하려면 n번째 사후확률에 정보가 전부 반영되어 n번째의 사후확률을 사전확률로 설정하고 n번째가 검은 공이라는 정보를 이용해 베이즈 추정을 한다.

n+1번째 공을 관측한 후 사후확률을 a'와 b'라고 하면

(n+1번째가 검은 공일 때의 A의 사후확률) : (n+1번째가 검은 공일 때의 B의 사후확률)
= a' : b'
= a*0.1 : b*0.8
= a : 8b

도표13-8 n+1번째가 검은 공이었을 때의 변화

	A (확률a)	B (확률b)	
n+1번째가 흑	a*0.1	b*0.8	n+1번째가 흑

13-6 여러 번 관측할수록 추측은 진실에 가까워진다

n+1번째가 검은 공이었을 때 사후확률의 비례관계는

a : b → a : 8b

왜 B의 사후확률이 A의 8배가 될까? A에서 검은 공이 관측될 확률이 0.1, B에서 검은 공이 관측될 확률이 0.8로 8배 크기 때문이다. 눈앞의 단지가 B라면 관측이 반복될수록 검은 공을 꺼내는 횟수가 많아진다.

도표13-9 검은 공의 관측횟수와 사후확률과 발생확률

검은 공 횟수	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	18	20
사후확률b	$8.62*1/10^{14}$	$3.00*1/10^{12}$	$1.10*1/10^{10}$	$4.00*1/10^{9}$	$1.40*1/10^{7}$	$5.22*1/10^{6}$	0.0002	0.007	0.1957	0.898	0.9968	0.9999	1	1	1	1	1	1	1	1	1
발생확률	$1.05*1/10^{14}$	$8.00*1/10^{13}$	$1.05*1/10^{11}$	$8.00*1/10^{10}$	$1.30*1/10^{8}$	$1.66*1/10^{7}$	$2.00*1/10^{6}$	0.00001	0.00009	0.0005	0.002	0.0074	0.0222	0.0545	0.109	0.1746	0.2182	0.2054	0.1369	0.0576	0.0115

도표13-9는 공을 20번 관측하였을 때 검은 공이 나온 횟수에 대응하여 '단지가 B일 사후확률'이 얼마 되는 가를 보여준다.

제13강의 정리

1. 베이즈 추정은 정보에 따라 판단이 흔들리는 상태를 묘사한다
2. 검은 공이 관측되면 검은 공이 많은 단지 쪽으로 판단이 기울고, 흰 공이 관측되면 흰 공이 많은 단지로 판단이 기운다
3. 베이즈 추정에서는 정보가 대량 있으면 올바른 결론을 내릴 수 있다

/ 연습문제

column 베이즈 역확률을 복권시킨 학자들

제2부

완전독학! ‘확률론’에서 ‘정규분포에 따른 추정’까지

제14강 '확률'은 '면적'과 동일한 성질을 지닌다

>>확률론의 기본

14-1 복잡한 베이즈 추정에는 확률기호가 필요하다

복잡한 베이즈 추정,연속형 사전분포 등을 사용하려면 확률기호가 필수

14-2 확률은 함수의 형태로 기술한다

확률=수학개념=사건에 0~1사이의 수치를 하나 대응시키는 것

{사건}→{수치}

확률모델 : 사건을 골라 그것에 대한 수치를 할당하는 것
예를 들어

확률모델 – 내일날씨
4개의 사건 – {맑음, 흐림, 비, 눈}
정규화조건- SUM(4개 사건 수치)=1
확률모델 – 맑음 : 0.3, 흐림 : 0.4, 비 : 0.2, 눈 : 0.1을 할당하는 것

근원사상 - 4개의 사건 = 더 이상 분해할 수 없는 가장 근본이 되는 사건
근원사상을 조합하여 사건을 만들 수 있음

{근원사상: 비, 눈} = [사건: 우산을 사용한다]

집합 {비, 눈}을 사상(事象), 근원사상(맑음, 흐림, 비, 눈)도 {맑음}, {흐림}, {비},{ 눈}처럼 집합표시({})를 하면 사상이 됨

근원사상→{근원사상}→[사상]→사건

확률 – 확률모델에서 사상A가 일어날 확률은 p(A)로 기술, 앞서 예에서 근원사상에 대해 다음과 같이 쓸 수 있다.

p({맑음})=0.3, p({흐림})=0.4 , p({비})=0.2, p({눈})=0.1

근원사상이 아닌 사상에 대한 확률은 그 사상을 구성하는 근원사상의 합으로 정의

p(우산을 사용한다)=p({비,눈})= p({비})+ p({눈})=0.2+0.1=0.3

우산을 사용한다는 현상이 일어날 확률은 0.3, 정리하면

확률p:[사상]→[수치], [수치]=p(사상)

또 다른 예(주사위를 던져 나오는 예)

근원사상 = {눈1, 눈2, 눈3, 눈4, 눈5, 눈6} 또는 {1,2,3,4,5,6}

‘짝수’ 사상 = {2,4,6}, ‘4이하’사상 = {1,2,3,4}

사상에 대한 확률을 정하려면 먼저 근원사상에 대해 확률을 설정한다

p({1})=1/6, p({2})=1/6, p({3})=1/6, p({4})=1/6, p({5})=1/6, p({6})=1/6

‘짝수’ 사상 p(짝수)=p({2,4,6})=1/6+1/6+1/6=1/2

‘4이하’사상 p(4이하)= p({1,2,3,4})=1/6+1/6+1/6+1/6=2/3

짝수라는 사상은 E, 4이하라는 사상을 F라고 하면 다음과 같이 표현

p(E)=1/2, p(F)=2/3

14-3 확률은 면적과 동일한 성질을 지닌다

도표14-1은 근원사상, 사상, 확률 정의에서 확률은 면적과 같은 성질을 가진다는 사실을 보여준다. 사상F=’4이하’의 확률, p(F)는 직사각형1~4까지의 면적임

도표14-1 확률모델의 면적도

1	2	3	4	5	6

각 직사각형의 면적은 1/6

1	2	3	4	5	6
사상F					

p(F)=□부분의 면적=2/3

확률의 가법법칙(加法法則)

사상A와 사상B는 공통의 근원사상(겹치는 부분)이 없을 때 ‘사상A or 사상B’의 확률은 ‘A확률+B확률’이다.

p(A or B)=p(A) + p(B)

도표14-2 확률의 가법법칙

1	2	3	4	5	6

A=2이하, B=5이상, A or B={1,2,5,6}

p(A)=왼쪽 □의 면적, p(B)=오른쪽 □의 면적, p(A또는B)= 좌우 □의 면적

14-4 베이즈 추정의 사전확률을 확률기호로 나타내면?

2강의 ‘암’과 ‘건강’ 에서 확률모델의 근원사상의 집합은 {암, 건강}, 여기서 할당한 사전확률은

p(암)=0.001, p(건강)=0.999

도표14-3 암 이환율에 따른 사전분포

0.001	0.999
{암}	

p(암)=0.001	{건강} p(건강)=0.999

4강에서 소개한 둘째아이가 여아일 확률에서 근원사상으로 확률로 설정({0.4},{0.5},{0.6})
0.4라는 것은 이 부부에게서 여자아이가 태어날 확률이 0.4라는 사건을 의미

도표14-4 어느 부부에게 태어날 아이가 여아일 확률에 대한 사전분포

1/3	1/3	1/3
{0.4} p({0.4})=1/3	{0.5} p({0.5})=1/3	{0.6} p({0.6})=1/3

p({0.4})=1/3 에서 0.4도 확률이고, 1/3도 확률이라 헷갈리지만 확률0.4는 어느 부부에게서 여아가 태어날 확률이 0.4라는 **근원사상(사건)**을 의미하고 1/3은 그 근원사상을 어느 정도의 가능성으로 보는 가에 대한 **신념의 정도**를 의미

14-5 &로 연결된 사상을 확률기호로 나타내면?

두 가지 확률현상을 합체한 경우 &로 연결한 사상을 만드는 데, 이를 **직적시행**이라 한다
도표14-5 동전 던지기와 주사위 던지기의 직적시행

동전 던지기 시행의 귀결
앞
뒤

주사위 던지기 시행의 결과					
1	2	3	4	5	6

▼

직적시행의 귀결					
앞&1	앞&2	앞&3	앞&4	앞&5	앞&6
뒤&1	뒤&2	뒤&3	뒤&4	뒤&5	뒤&6

각 칸은 (동전 던지기의 귀결) & (주사위 던지기의 귀결)과 같은 형태로 &로 연결된 귀결로 채운다. 이것이 직적시행이라는 확률모델에서의 근원사상이 된다. 다음의 12개가 근원사상

앞&1, 앞&2, 앞&3, 앞&4, 앞&5, 앞&6
뒤&1, 뒤&2, 뒤&3, 뒤&4, 뒤&5, 뒤&6

동전 던지기의 '앞'이라는 사상은 다음과 같다

'앞'={ 앞&1, 앞&2, 앞&3, 앞&4, 앞&5, 앞&6 }

주사위 던지기의 '2'라는 사상은 다음과 같다

'2'={ 앞&2, 뒤&2 }

사상'앞'과 사상'2'가 모두 일어나는 사상은 { 앞&2 }

도표14-6 직적공간에서 원시행의 사상

	↓ 사상 '앞'과 '2'가 둘 다 일어남				사상 '앞'
앞&1	앞&2	앞&3	앞&4	앞&5	앞&6
뒤&1	뒤&2	뒤&3	뒤&4	뒤&5	뒤&6
	사상 '2'				

동전 던지기와 주사위 던지기는 독립시행이므로 12개의 모든 근원사상에 대해

p(동전 던지기의 귀결 **&** 주사위던지기의 귀결)

= p(동전 던지기의 귀결) * p(주사위던지기의 귀결)

우변의 곱셈으로 좌변의 확률이 정해진다

p({뒤&4})= p({뒤})* p({4})=1/2 * 1/6=1/12

확률의 가법법칙을 사용하면

p({앞})=p({앞&1,앞&2,앞&3,앞&4,앞&5,앞&6})

= p({앞&1})+ p({앞&2})+ p({앞&3})+ p({앞&4})+ p({앞&5})+ p({앞&6})
= 1/2*6
= 1/2

제14강의 정리

1. 확률모델은 근원사상(事象), 확률에 의해 구성된다
2. 근원사상이란 이 이상 분해할 수 없는 근본적인 사건을 말한다
3. 사상은 근원사상을 몇 가지 모아 놓은 집합을 말한다
4. 근원사상에 대해서 그 확률은 p({e})로 표시한다
5. 예를 들어 근원사상 e, f, g로 구성된 사상{e, f, g}의 확률은 p({e, f, g})=p({e})+ p({f})+ p({g})로 정의된다
6. 확률의 가법법칙이란 A와 B가 중첩되지 않는 사살일 때, p(A or B) = p(A) + p(B)가 성립되는 것이다
7. 두 가지 확률현상을 결합하여 만드는 직적시행은 a&b와 같은 근원사상으로 이루어져 있으며, 이 확률은 통상 승법법칙이 성립되도록 정의(독립시행으로 가정된다)되므로 다음과 같이 곱셈으로 계산할 수 있다. p({a&b}) = p({a}) * p({b})

/ 연습문제

제15강 정보를 얻은 후 확률의 표시법

>>‘조건부 확률’의 기본적인 성질

15-1 ‘조건부 확률’을 사용하여 ‘베이즈 역확률’을 나타내려면

지금까지 배운 것은 ‘정보를 얻었을 때 확률이 변화한다’ → 변화하는 확률을 기술하는 것 → 조건부확률. 15강에서는 조건부 확률을 사용하여 베이즈 역확률을 표현하는 공식을 보여줌

15-2 ‘조건부 확률’이란 부분을 전체로 간주하여 수치를 수정하는 것

★주사위 던지기 예 : 주사위 눈이 짝수가 나올 확률

사상(E) = ‘주사위 눈이 짝수다’ = E = {2, 4, 6}

사상E의 확률 = p(E) =3/6 = 1/2

★이제 제3자가 몰래 결과를 보고 ‘6은 아니다’라는 정보를 준다.

조건부확률 → ‘6이 아니다’라는 정보를 얻은 후 짝수일 확률

사상(F)=’6이 아니다’=F={1, 2, 3, 4, 5}

p(E|F) = 사상 F가 일어났다는 정보하에 사상E의 확률

도표15-1 조건부 확률의 사고법

색칠한 부분이 사상E

1	2	3	4	5	6

아무런 정보가 없을 때 p(E)=3/6=1/2

▼

사상 F하는 정보를 입수

사상F

1	2	3	4	5	~~6~~

2와 4에서 사상E와 사상F의 중첩

전체가 F={1,2,3,4,5}로

변화한다. F와 중첩된 부분에서 사상E는

전체의 5분의 2를 차지하므로

p(E|F)=2/5

사상F로 인해 두 가지가 변화

#1 사상F가 전체가 되었으므로 사실 F의 확률이 1로 설정되어야 한다

#2 사상F가 전체세계로 한정되므로 사상E와 사상F의 공통부분에 한정하여 확률 생각

사상F가 일어났다는 정보하에서 E의 조건부확률은 , F를 전체로 생각하여 'E와 F의 중첩'이 F안에 차지하는 비율 (E와 F가 중첩된 면적) ÷ (F의 면적) 이 된다

p(E|F) = (E와 F가 중첩된 면적) ÷ (F의 면적) = p(E와 F가 중첩) ÷ p(F)

p(E|F) = p({2, 4}) / p({1, 2, 3, 4, 5}) = 2/6 ÷ 5/6 = 2/5

조건부 확률이란 얻은 정보인 사상을 전체로 재설정하고, 가능성이 사라진 근원사상을 소멸시켜서 비율을 새로이 정하는 것

조건부 확률의 공식

사상 B라는 정보를 얻었을 때 사상A의 조건부 확률 p(A|B)는 다음 식으로 정의된다

p(A|B) = p(A와 B의 중첩) ÷ p(B)

15-3 타입이 부여된 확률='조건부 확률'

조건부 확률을 사용할 때

1. 타입별로 데이터의 확률을 설정할 때
2. 사후확률을 계산할 때

7강과 13강의 단지 속 색깔공의 사례

문제설정

눈앞에 단지가 하나 있는 데, 단지A나 B중 하나임을 알고 있지만 겉으로 봐서는 어느 쪽인지 알 수가 없다. 단지 A에는 흰 공 아홉 개와 검은 공 한 개가 들어 있고, 단지B에는 흰 공 두개와 검은 공 여덟 개가 들어 있다는 정보를 가지고 있다. 이때 단지에서 공을 한 개 꺼냈더니 검은 공이었다. 눈앞에 있는 단지는 어느 것일까?

사상으로 위의 문제를 풀면

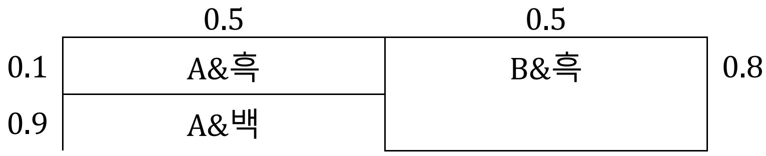
근원사상의 집합 = {A&흑, A&백, B&흑, B&백}

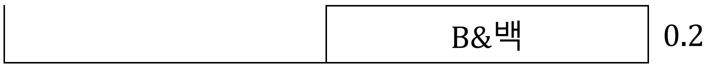
7강과 13강에서 설명한 '단지 A에서 검은 공이 나올 확률이 0.1' → 조건부 확률

위의 조건부 확률을 식으로 쓰면

p(흑|A) = 0.1

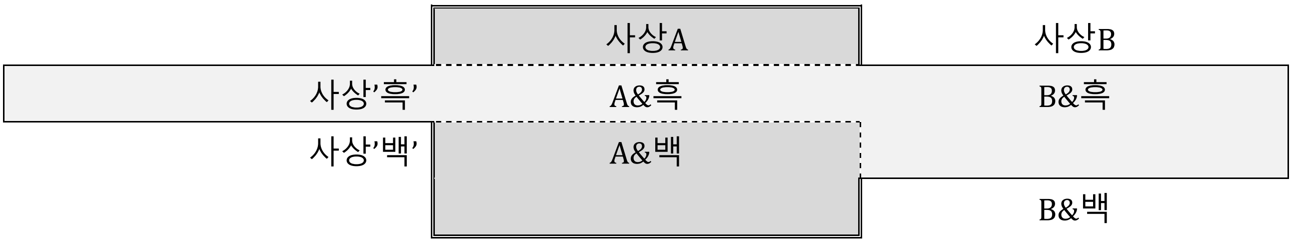
도표15-2 조건부 확률의 설정





7강에서 A&흑의 확률을 0.5*0.1로 계산하였는데, 다시 계산하면 다음과 같다.

도표15-3 A& 혹은 사상 ‘A’와 사상 ‘흑’의 중첩



직적시행에서 사상 A = {A&흑, A&백}, 사상 흑 = {A&흑, B&흑}

따라서 사상A와 흑의 중첩은 ={A&흑}

조건부 확률의 정의로 바꾸면

p(흑|A) =p(사상 A와 사상 흑의 중첩) ÷ p(A) = p(A&흑) ÷ p(A)

p(A&흑) = p(A) * p(흑|A) …(1)

p(A&흑) = p(A) * p(흑|A) = 0.5 * 0.1 = 0.05 …(2)

→ 확률이 직사각형의 면적

& 사상의 확률법칙

p(타입정보) = p(타입) * p(정보|타입)

15-4 조건부 확률의 공식을 통해 사후확률을 이해한다

단지의 예에서 베이즈 추정은 ‘검은 공이었다는 사실로부터 단지B일 확률’을 추정하였다

→ 결과로부터 원인을 계산 → p(B|흑)를 계산하는 것

p(B|흑) = p(B & 흑) ÷ p(흑) …(3)

(3)의 p(B & 흑)은 (1)과 (2)의 식에서 p(A&흑)을 구한 것과 같이

p(B & 흑) = p(B) * p(흑|B) …(4)

p(B & 흑) = 0.5*0.8 …(5)

(3)의 p(흑)에서 흑이라는 사상은

흑 = { A&흑, B&흑}

p(흑) = p(A & 흑)+ p(B & 흑)

p(흑) = p(A)*p(흑|A) + p(B)*p(흑|B) …(6)

(4)와 (6)을 (3)에 대입하면

$$\mathrm{p(B|흑)} = \frac{p(B)p(\text{흑}|B)}{p(A)p(\text{흑}|A)+p(B)p(\text{흑}|B)} \quad \cdots(7)$$

이것이 **베이즈 공식**이다

p(B|흑) = 0.5*0.8÷{0.5*0.1+0.5*0.8}=8/9

(7)의 우변에서 p(A)와 p(B)는 타입의 사전확률, p(흑|A)와 p(흑|B)는 원인에서 결과는 낳는 확률.

지금까지의 계산은 면적도를 활용한 방법을 수식화한 것

도표15-4 베이즈 역확률의 식

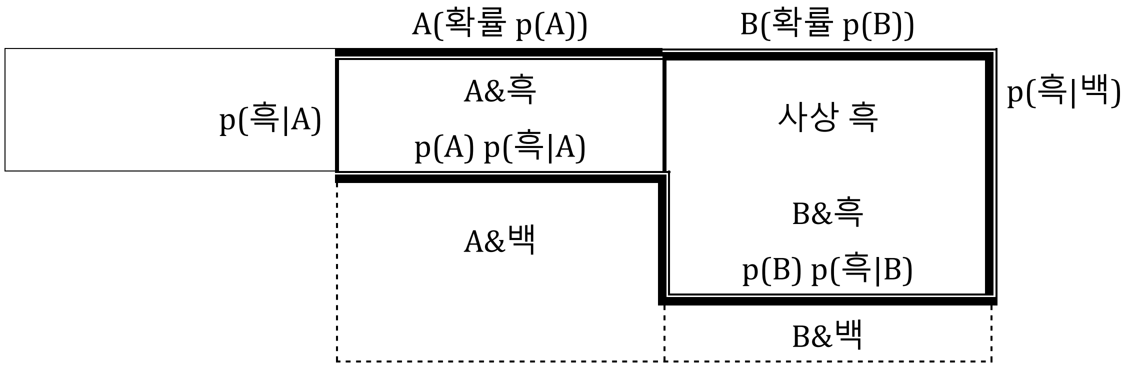


도표15-4 에서 흑이라는 정보하에서 다음과 같은 비례관계가 성립

(A의 사후확률) : (B의 사후확률) = (A&흑의 면적) : (B&흑의 면적)

이것을 조건부 확률로 기술하면

p(A)p(흑|A) : p(B)p(흑|B) …(8)

(8)식의 좌우계산은 직사각형의 가로세로 길이인 확률을 곱한 것

p(A)p(흑|A) : p(B)p(흑|B)

$$= \frac{p(A)p(\text{흑}|A)}{p(A)p(\text{흑}|A)+p(B)p(\text{흑}|B)} : \frac{p(B)p(\text{흑}|B)}{p(A)p(\text{흑}|A)+p(B)p(\text{흑}|B)}$$

$$(\text{B일 사후확률}) = \frac{p(B)p(\text{흑}|B)}{p(A)p(\text{흑}|A)+p(B)p(\text{흑}|B)} \cdots(9)$$

(9)는 (7)식과 같은 것이다. 흑이라는 정보를 얻은 사실하에서 B의 조건부 확률이라는 것은 (A&흑의 직사각형+B&흑의 직사각형) 면적에서 (B&흑인 직사각형)이 차지하는 비율이다.

제15강의 정리

1. 조건부 확률이란 정보가 들어와서 근원사상이 줄어든 세계에 비례관계를 부여하는 것이다
2. 사상B라는 정보하에서 사상A의 조건부 확률 p(A|B)는 다음의 식으로 정의된다
   p(A|B) = p(A와 B의 중첩) / p(B)
3. 베이즈 추정에서는 조건부 확률의 공식2를 두 단계로 사용한다

4. 첫 번째 단계는 타입&정보의 확률을 구하는 것. 즉
   p(타입&정보) = p(타입) / p(정보|타입)
5. 두 번째 단계는 사후확률을 구하는 것. 그것은 주어진 데이터하에서 p(타입&정보)의 비례관계를 4를 이용해서 계산하여 정규화 조건을 충족시켜주면 된다

/ 연습문제

제16강 더 범용적인 추정을 위한 ‘확률분포도’

16-1 실용레벨로 나아가기 위해 필요한 ‘확률분포도’와 ‘기대치’

조금 더 복잡한 설정의 추정, 범용적인 추정 → **확률분포도**와 **기대치**에 대한 이해가 필요

16-2 ‘동일한 확률’형 확률모델을 생각하다

확률모델은 근원사상과 그에 대한 확률분할로 정의
근원사상 = 각 확률이 같은 정도로 발생하는 상태

★동전 던지기 예

근원사상의 집합 : { 앞, 뒤 }

확률의 분할 : p({앞}) = 1/2, p({뒤}) = 1/2

★주사위 던지기 예

근원사상의 집합 : { 1,2,3,4,5,6 }

확률의 분할 : p({1}) = 1/5, p({2}) = 1/6, p({3}) = 1/6, p({4}) = 1/6, p({5}) = 1/6, p({6}) = 1/6

도표16-1 동전과 주사위에서 ‘각 확률이 같은 정도로 발생하는 상태’

동전의 확률모델 면적도

앞	뒤

각 직사각형의 면적은 1/2

주사위의 확률모델 면적도

1	2	3	4	5	6

각 직사각형의 면적은 1/6

이번에는 룰렛의 확률모델을 생각해보자. 근원사상은 1~36까지의 정수로

$$\{1, 2, 3, \ldots, 35, 36\}$$

확률의 분할은 다음과 같이 설정할 수 있다

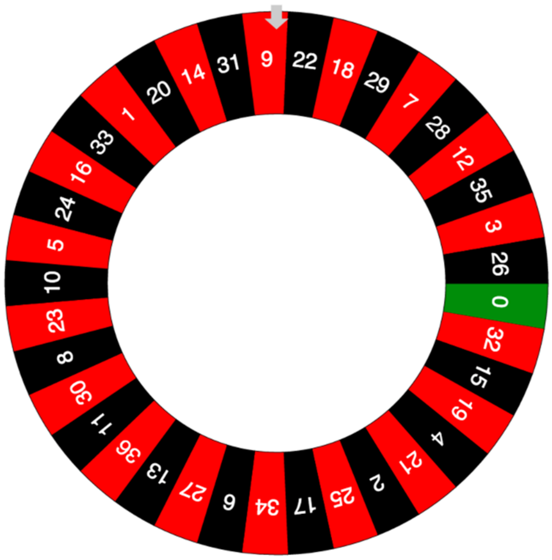
$$p(\{x\})=1/36\ (x=1, 2, 3, \ldots 36)$$

이 모델에서 $1 \le x \le k$를 만족하는 정수 x가 될 확률, $p(1 \le x \le k)$은 $1 \le x \le k$가 전체 36분의 k라는 비율을 차지한다

$$p(1 \le x \le k) = k/36$$

k=36이라고 해보자. x가 1~36사이의 어느 숫자이든 $1 \le x \le k$를 만족하므로 확률은 1.0이 된다

도표16-2 룰렛에서 '각 확률이 같은 정도로 발생하는 상태'



룰렛의 확률모델 면적도

1	2	…	35	36

각 직사각형의 면적은 1/36

16-3 '각 확률이 같은 정도로 발생하는 상태'모델을 연속화한 '균등분포'

룰렛 확률모델은 1~36의 정수를 각 확률이 같은 정도로 발생하는 상태(즉 사상)으로 설정한 것인데(이산적), 이것은 연속 무한 개의 근원사상으로 확장한 것이 **균등분포**라는 확률모델이다.

룰렛의 원주에 $0 \le x \le 1$사이의 무수히 많은 수의 x가 그려져 있다고 상상해보자. 이것은 이 책에서는 [0, 1]-룰렛모델로 부른다. $0 \le x \le 1$에서 랜덤으로 하나(사상)를 선택하였다고 하자(여기서 x는 근원사상이 된다)

그러나 확률분할에 주의해야 한다. 주사위나 동전던지기 처럼(이산적 분포) 각 확률이 같은 정도로 발생하는 상태의 확률로 분할할 수 없다. 다만 정규화조건에 따라 전체의 확률은 1이다. 사상{x}에 동일한 확률 a를 배정하려면 $0 \le x \le 1$사이에는 무수히 많은 x가 있어

$$(0 \le x \le 1\text{이 되는 모든 수 } x\text{에 대한 } \{x\}\text{의 확률의 합}) = (\text{무한 개의 } a\text{의 합}) = 1$$

이 되어야 한다.

아무리 작은 수 a라도 무수하 많다면 1을 넘을 것이다. 그래서 $a=0$이어야 한다. 그런데 $a=0$이면 두 가지의 문제가 생긴다

첫 번째 문제 : 무한 개의 0을 더해 1이 된다는 것

두 번째 문제 : $0\le x\le 1$ 이 되는 각x에 대해 그 확률 $p(\{x\})=0$이라면 $0\le x\le 0.5$가 되는 x가 선택될 확률은 어떻게 계산해야 하는 가?

그래서 다음과 같이 확률을 설정한다

[0, 1]-룰렛 모델의 확률설정

[0, 1]-룰렛모델에서는 $0<t\le 1$을 만족하는 각 t에 대해 '0이상 t미만의 수'의 집합을 기본사상으로 한다. 즉

$E=\{0\le x<t$를 만족하는 $x\}$

가 기본사상이다. 그리고 이 사상E에 대한 확률을

$p(E)=t$

로 배정한다. 이후 이 사상 E를 $\{0\le x<t\}$, 그 확률 $p(E)$를 $p(0\le x<t)$로 약기한다

- $t=0.5$이라면 사상$\{0\le x<0.5\}$는 '0 이상 0.5미만의 수가 선택된다'는 의미의 사상. '0 이상 1이하의 수'에 비해 절반이므로 확률을 0.5로 할당
- $t=0.7$이라면 마찬가지고 70%이므로 0.7로 설정

도표16-3 [0, 1]-룰렛의 확률

[0,1]-룰렛의 확률

E 0≤x<0.5	

직사각형의 면적은 0.5 → $p(0\le x<0.5)=0.5$

E 0≤x<0.7	

직사각형의 면적은 0.7 → $p(0\le x<0.7)=0.7$

16-4 [0, 1]-룰렛모델의 일반사상의 확률

확률의 가법법칙에 따라 [0, 1]-룰렛모델의 사상의 확률을 계산할 수 있다

★$0.5\le x<0.7$ 범위의 수 x가 선택된다는 사상 $\{0.5\le x<0.7\}$의 확률

$0\le x<0.5$와 $0.5\le x<0.7$ 범위를 합치면 $0\le x<0.7$이 된다. 확률의 가법법칙에 따라

$p(0\le x<0.5) + p(0.5\le x<0.7) = p(0\le x<0.7)$

$p(0.5\le x<0.7) = 0.7 - 0.5 = 0.2$

확률은 $(0.5\le x<0.7)$의 폭이 0.2이다

도표16-4 [0, 1]-룰렛모델의 일반사상

	0.5≤x<0.7	

직사각형의 면적은 0.2 → p(0.5≤x<0.7) = 0.2

[0, 1]-룰렛모델은 0≤x≤1의 범위의 수에서 랜덤으로 선택되는 모델인데 길이가 1인 예이다.

예를 들어 2≤x<5의 범위의 수에서 랜덤으로 선택되는 균등분포를 생각해보자.
도표16-5 [2, 5]-룰렛의 확률

[2, 5]-룰렛의 면적도

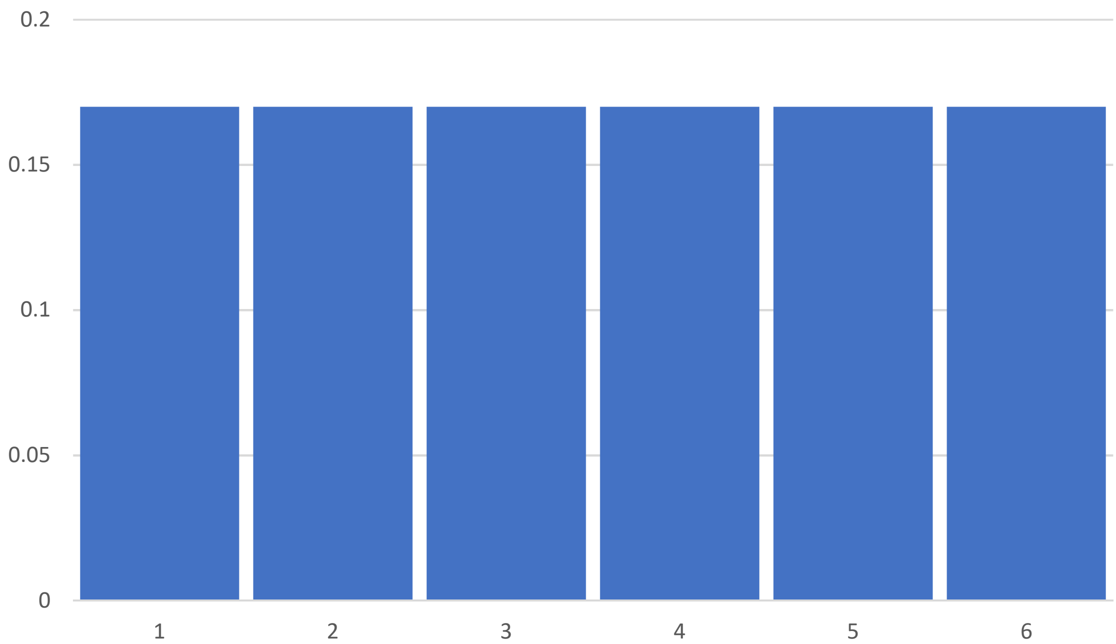
2≤x<5	

근원사상은 그림과 같이 {2≤x<t}와 같은 사상(단 t는 2<t≤5를 충족한다). 전체 길이가 3이라는 점에 의거해 사상{2≤x<t}의 길이는 t-2이므로 p(2≤x<t)=(t-2)/3으로 설정된다(사상 구간의 길이÷3이 된다)

16-5 복잡한 확률모델을 그림으로 나타낼 수 있는 '확률분포도'

연속형 확률모델을 그림으로 나타내려면 직사각형 면적도 대신 다른 방법 사용 → 확률분포
확률분포도 가로축 : 사상의 수치, 세로축 : 확률밀도(이산형의 확률)

주사위의 확률분포도를 그리면 가로축은 1~6의 사상, 세로축은 확률 1/6이 된다
도표16-6 주사위의 확률분포도



예를 들어 2≤x≤4의 눈이 나올 확률은 2~4까지의 세 막대의 높이 합계를 구하면 된다

$p(2 \le x \le 4) = p(\{2, 3, 4\}) = 1/6 + 1/6 + 1/6 = 1/2$

이번에는 연속형인 균등분포 [0, 1]-룰렛 모델의 확률분포도를 그려보자. 여섯 개의 막대가 무수히 가느다란 막대(거의 실선보다 얇은)로 이루어진 확률분포도를 연상하기 바란다. 확률분포도의 가로축에는 $0 \le x \le 1$이 되는 수 x가 줄지어 있다. 세로 축의 높이가 1이다. 이때 높이1은 각 x가 선택될 확률이 아니다. 균등분포와 같은 연속형 확률모델의 경우 확률을 높이가 아닌 면적으로 나타낸다. 면적으로 생각한다면 CD는 선이므로 면적은 0이 된다.

도표16-7 균등분포의 확률분포도

근원사상 $\{0.5 \le x < 0.7\}$의 확률은 도표16-8의 색칠된 부분의 직사각형의 면적이 된다 가로 0.2, 세로 1이므로 면적은 0.2*0.1=0.2

도표16-8 연속형 확률분포도에서는 확률은 면적으로 표시된다

★확률밀도와 확률의 관계
속도에는 거리가 없다. 분속10m에서 m은 거리가 아니다. 분속은 순간스피드로 1분간 분속10m를 유지하면 거리 10m가 나온다는 얘기이다. 분속은 시간과 곱하여야 비로소 거리가 생긴다. 확률밀도는 분속과 같은 개념이다.

분속 → 확률밀도(세로축)
시간 → 구간의 폭(가로축)
거리 → 확률

제16강의 정리

1. 동전과 주사위는 각 수가 '각 확률이 같은 정도로 발생하는 상태'로 설정되는 확률모델이다
2. $0 \le x \le 1$의 수가 '각 확률이 같은 정도로 발생하는 상태'로 설정되는 것이 [0, 1]-룰렛모델
3. [0, 1]-룰렛모델은 균등분포의 확률모델이며 사상$\{0 \le x < t\}$라는 폭을 가진 구간을 기본으로 생각한다
4. 사상$\{0 \le x < t\}$의 확률 p(사상$\{0 \le x < t\}$)는 폭 t로 설정된다
5. 확률분포란 가로축에 수치, 세로축에 확률을 설정한 것이다. 연속형의 경우 세로축은 확률이 아니라 확률밀도를 나타낸다

6. 균등분포의 확률분포도는 수평한 직선(선분)이 된다. 사상의 확률은 직사각형의 면적이 된다
7. 균등분포에서 (확률)=(확률밀도) * (구간의 길이)

/ 연습문제

제17강 두 가지 숫자로 성격이 정해지는 ‘베타분포’

베이즈 추정에서 자주 사용하는 베타분포를 알아본다

17-1 베이즈 추정에 자주 사용되는 연속형 분포 ‘베타 분포’

지금까지의 베이즈 추정은 사전분포를 위한 타입의 설정이 유한개. 가령 4 강에서 여아를 낳을 확률은 0.4, 0.5, 0.6 세 가지로 설정했지만 확률 p 는 0≤p≤1 을 만족하는 임의의 p 이며 따라서 타입은 무한이 된다. 그래서 사전확률분포는 연속형 확률분포이어야 한다.

17-2 베타분포는 어떤 분포인가?

확률분포에서 가로축 x 는 근원사상의 근간이 되는 수치, 세로축 y 는 확률밀도(확률밀도란 가로축 구간의 길이를 곱하면 확률로 전환되는 양)

★베타분포

$y=(정수)*x^{\alpha-1}(1-x)^{\beta-1}$ $(0\leq x\leq 1)$ …(1)

여기서 α와 β는 1 이상의 자연수로 베타분포의 종류를 특정해주는 값
α와 β가 작을수록 베타분포 그래프는 단순, 클수록 복잡
(정수)는 정규화 조건(전 사상의 확률이 1 이라는 조건)을 만족하기 위한 수치

★α=1, β=1 인 경우

$y=(정수)*x^0(1-x)^0 = (정수)*1*1=(정수)$ $(0\leq x\leq 1)$

y=(정수)는 가로축 x 와 평행으로 [0, 1]-룰렛 모델과 같다. 정규화조건 1 을 만족해야 하므로 (정수)=1 이어야 한다.

$y=1$ $(0\leq x\leq 1)$ …(2)

★α=2, β=1 인 경우

$y=(정수)*x^1(1-x)^0$ $(0\leq x\leq 1)$

$y=(정수)x$ $(0\leq x\leq 1)$ …(3)

여기서 (정수)=2 가 된다. 이유는 나중에…

★α=1, β=2 인 경우

$y=(정수)*x^0(1-x)^1$ $(0\leq x\leq 1)$

$y=(정수)(1-x)$ $(0\leq x\leq 1)$ …(4)

여기서 (정수)=2 가 된다. 이유는 나중에…

★α=2, β=2 인 경우

$y=(정수)*x^1(1-x)^1$ $(0\leq x\leq 1)$

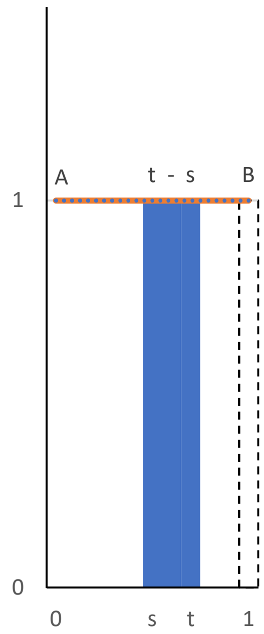
$y=(정수)x(1-x)$ $(0\leq x\leq 1)$ …(5)

여기서 (정수)=6 가 된다. 이유는 나중에…

17-3 α=1,β=1의 예는 [0, 1]-룰렛

α=1, β=1 인 경우 [0, 1] - 룰렛 모델(균등분포중 하나)이다. 따라서 베타분포이기도 하다

도표 17-1 α=1,β=1 인 베타분포의 확률분포도

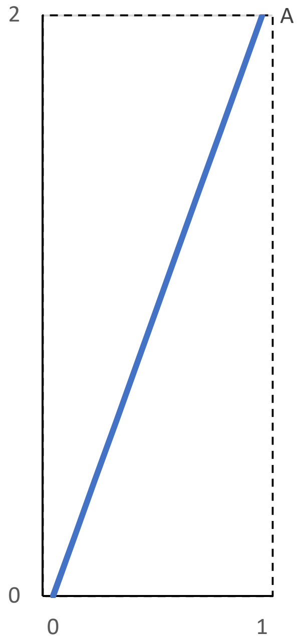


17-4 α=2, β=1의 예

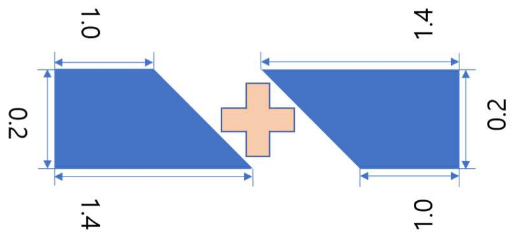
α=2, β=1 인 경우 베타분포는 1 차함수, y=(정수)x (0≤x≤1) …(3)
여기서 (정수)=2 인 이유는 다음과 같다.
확률분포도에서 확률은 면적이다. 전체사상의 확률 p(0≤x≤1)은 삼각형 OAB 의 면적이다. 정규화 조건에 따라 삼각형 OAB 면적은 1 이다. 삼각형 면적=(밑변)*(높이)÷ 2 인데, (밑변)=1 이므로 (높이)=2 가 되어야 면적이 1 이 된다. 따라서 (정수)=2 이다.

y=2x (0≤x≤1) …(6)

도표 17-2 α=2, β=1 인 베타분포의 확률분포도



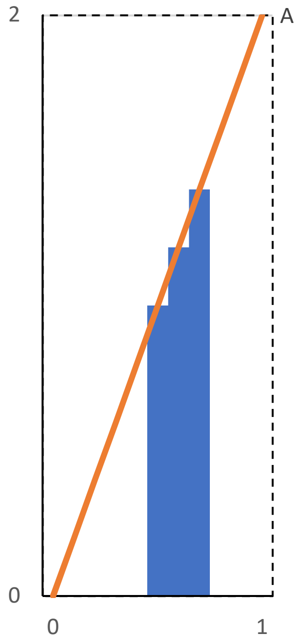
★ 사상 {0.5≤x<0.7}의 확률 p(0.5≤x<0.7)은 얼마인가?
확률 p(0.5≤x<0.7)는 도표 17- 3 의 색칠한 사다리꼴의 면적이다. 사다리꼴을 하나 더 복사하여 상하로 뒤집어 이어 붙이면 긴 직사각형이 된다. 이 직사각형의 면적을 구한 뒤 절반으로 나누면 사다리꼴의 면적이 된다.



사다리꼴 면적=(가로)*(세로)÷2 이다.
(가로)는 가로축 x 의 너비 0.7- 0.5=0.2 이다
(세로)는 사다리꼴을 양변의 높이 1.4+1.0 을 더해 사다리꼴 두개로 만든 직사각형의 높이가 된다.
그래서 확률은 (1+1.4)*(0.2)÷2=0.24 이다.

p(0.5≤x<0.7)=0.24

도표 17-3 베타분포 y=2x 에서의 확률



17-5 α=1,β=2의 예

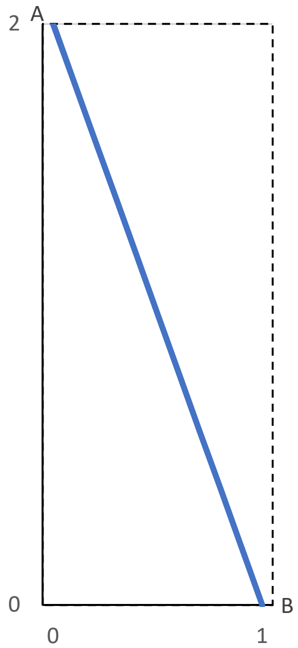
이 경우에도 베타분포는 1 차함수이다.

y=(정수)(1-x)　　　(0≤x≤1) …(4)

앞서의 경우와 마찬가지로 전체사상의 확률 p(0≤x≤1)은 삼각형 OAB 의 면적은 1 이어야 한다.
삼각형의 면적을 구하는 것이므로 (정수)=2 이어야 한다.

y=2(1-x)　　　(0≤x≤1) …(7)

도표 17-4 α=1, β=2 인 베타분포의 확률분포도



17-6 α=2, β=2의 예

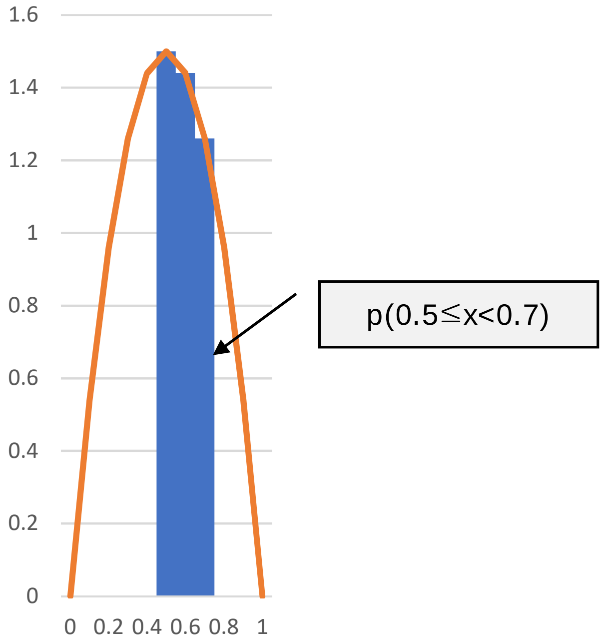
이 경우 베타분포는 2 차 함수이다.

$y=(\text{정수})x(1-x) \qquad (0 \leq x \leq 1) \cdots (5)$

이번 경우에는 2 차 함수라서 적분을 해야 한다. 전체사상의 확률 즉 면적=1 이라는 조건하에 적분을 하면 (정수)=6 이어야 한다.

$y=6x(1-x) \qquad (0 \leq x \leq 1) \cdots (8)$

도표 17-5 α=2, β=2 인 베타분포의 확률분포도

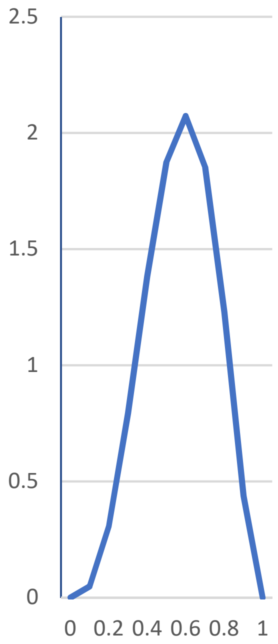


17-7 베타분포 α, β가 커지면 복잡해진다

α=4, β=3 인 베타분포는

$y=60x^3(10x)^2 \qquad (0 \leq x \leq 1) \cdots (9)$

도표 17-5 α=4, β=3 인 베타분포의 확률분포도



제17강의 정리

1. 베타분포는 x 의 거듭제곱과 (1-x)의 거듭제곱을 곱한 형태다
2. x 의 0 제곱과 (1-x)의 0 제곱의 경우 균등분포와 일치한다
3. x 의 1 제곱과 (1-x)의 0 제곱의 경우, x 의 0 제곱과 (1-x)의 1 제곱의 경우 확률분포도는 선분이 된다
4. x 의 1 제곱과 (1-x)의 1 제곱의 경우 확률분포도는 포물선이 된다
5. 정수는 정규화 조건(전체면적이 1)으로부터 정해진다

/ 연습문제

제18강 확률분포의 성격을 결정짓는 '기대치'

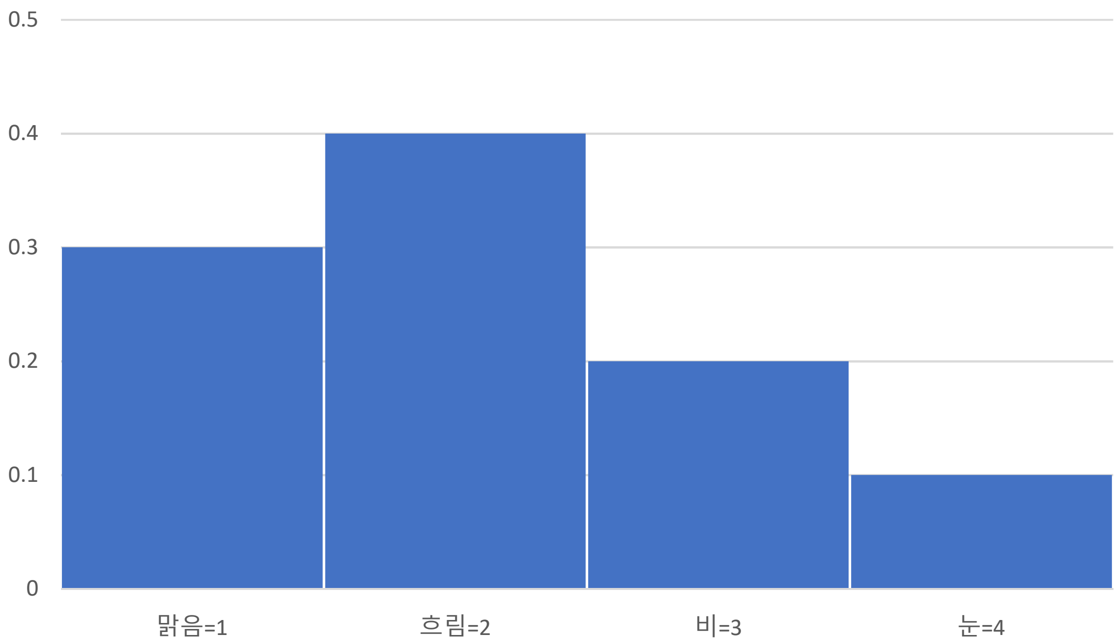
18-1 확률분포를 하나의 수치로 대표하려면

2 강의 결과는 암일 사후확률이 4.5%, 아닐 확률이 95.5%로 각각의 가능성을 표시하였다. 반면에 4 강에서 여아가 태어날 확률 0.4, 0.5, 0.6 의 사후확률을 각각 27%, 33%, 40%로 결과를 냈다. 그러나 4 강의 결론으로는 '부부에게 다음에 여아가 태어날 확률'에 대한 답이 되질 못한다. 하나의 숫자로 답해야 한다. 그것이 '**기대치**'이다.

18-2 기대치의 계산방법

14 강에서 다룬 날씨확률모델에서 근원사상은 {맑음, 흐림, 비, 눈} 이며, 각각의 확률이 p({맑음})=0.3, p({흐림})=0.4, p({비})=0.2, p({눈})=0.1 이었다. 기대치를 계산하기 위해 근원사상에 숫자를 배정한다.
맑음→1, 흐림→2, 비→3, 눈→4

도표 18-1 날씨의 확률분포도



기대치는 다음과 같이 계산한다
(확률분포의 기대치) = (수치) * (그 수치가 나올 확률) 의 합계
(날씨 확률분포의 기대치) = 1*0.3+2*0.4+3*0.2+4*0.1 = 2.1

- 기대치 2.1 은 흐림과 비 사이의 중간에 위치한다.
- (수치) * (그 수치가 나올 확률)는 중요도를 매긴다는 의미 → 가중평균

18-3 장기적으로 볼 때 기대치는 현실을 적중시킨다

하루가 아니라 만일 N 일간 날씨를 기록하였다면 p({맑음})=0.3, p({흐림})=0.4, p({비})=0.2, p({눈})=0.1 확률에서
맑은 날은 0.3N, 흐린 날은 0.4N, 비가 오는 날은 0.2N, 눈이 오는 날은 0.1N 일로 예상된다.

1*0.3N + 2*0.4N + 3*0.2N + 4*0.1N = 2.1

이는 앞서 구한 기대치 2.1 에 관측한 일수인 N 을 곱한 것이다
매일의 기대치를 합계해 나가면 장기적으로 실제 점수의 합계와 거의 같은 값이 된다.

18-4 기대치 확률분포도를 '야지로베에'로 간주했을 때의 지점

기대치는 아래 그림과 같은 장난감 '야지로베에'로 설명할 수 있다. 확률분포를 '야지로베에'로 보면 기대치는 균형을 이루는 지점이다.

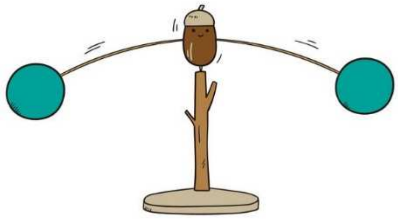


도표 18-2 기대치는 야지로베에의 지점

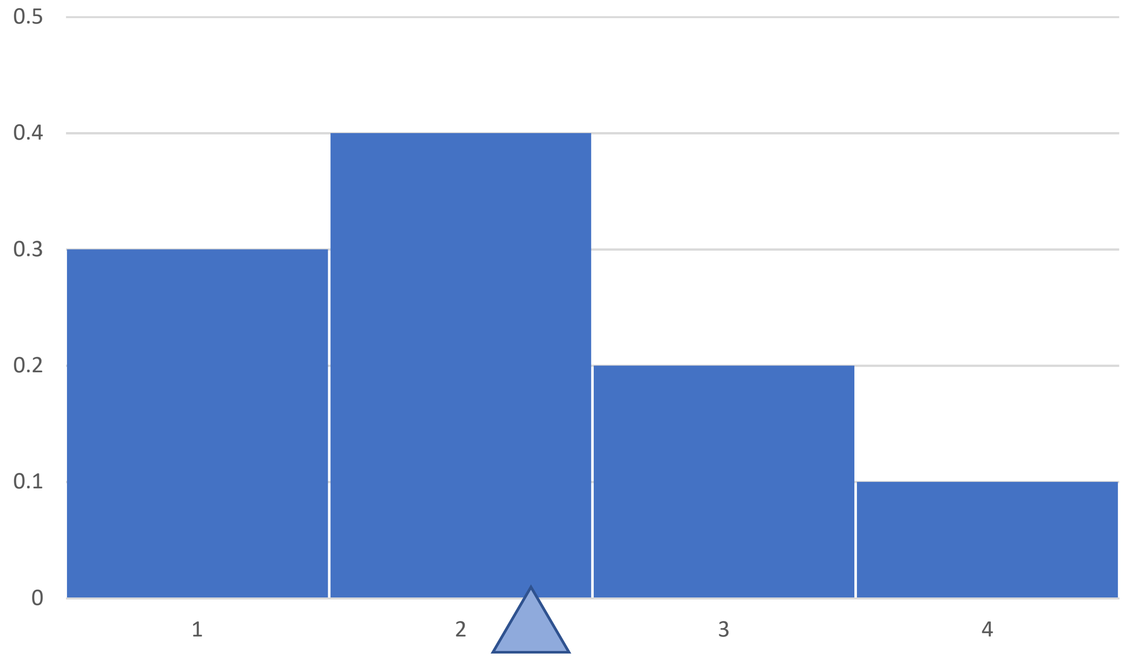
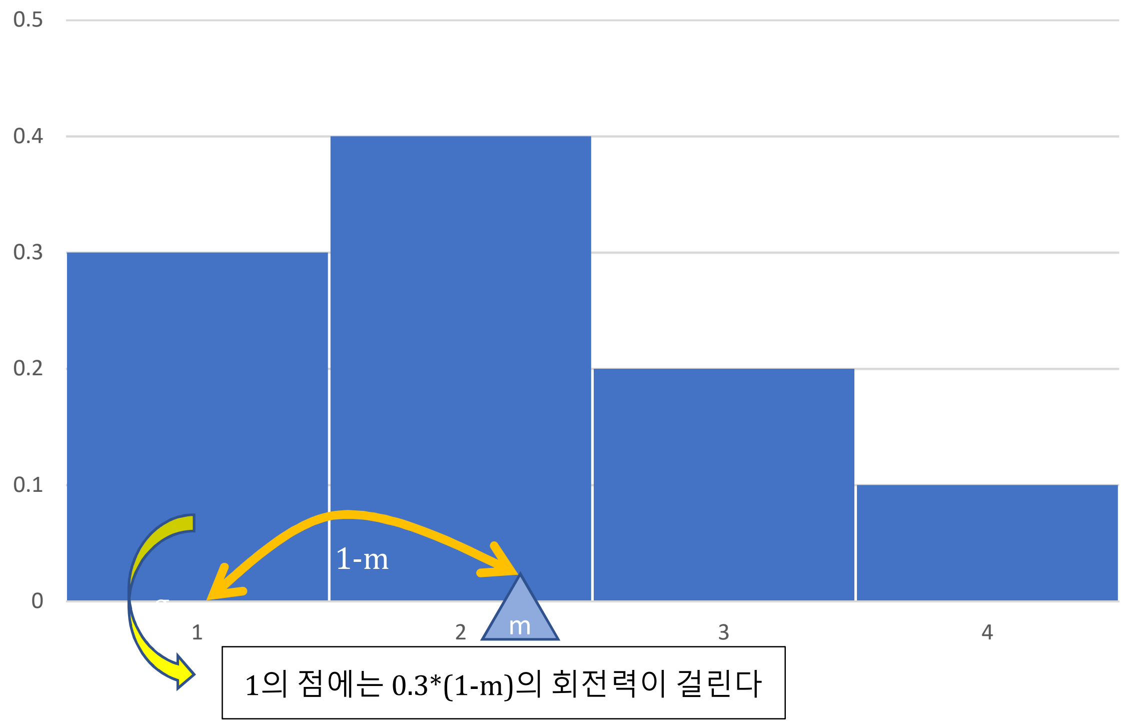


도표 18-3 야지로베에 걸린 회전력



m 은 균형을 이루기 위한 지점
0.3*(1-m) + 0.4*(2-m) + 0.2*(3-m) + 0.1*(4-m) = 0

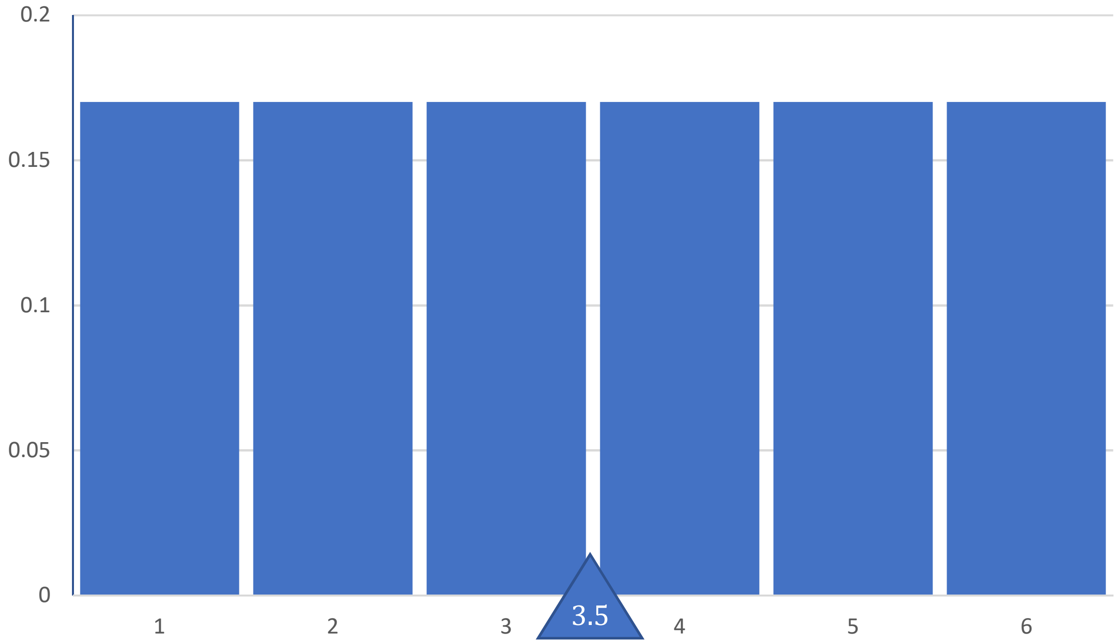
(x 의 기대치) = m

18-5 주사위와 여아의 사례에서 기대치를 구한다

주사위의 근원사상은 {1, 2, 3, 4, 5, 6}, 확률은 p({1})=1/6, p({2})=1/6, p({3})=1/6, p({4})=1/6, p({5})=1/6, p({6})=1/6 인데, 기대치를 계산하면

(주사위의 기대치) = 1*(1/6) + 2*(1/6) + 3*(1/6) + 4*(1/6) + 5*(1/6) + 6*(1/6) = 3.5

도표 18-4 주사위의 기대치



4 강의 여야가 태어날 확률의 경우 x=0.4, 0.5, 0.6 이고 확률분포가 0.27, 0.33, 0.4 였다. 기대치는 다음과 같다.

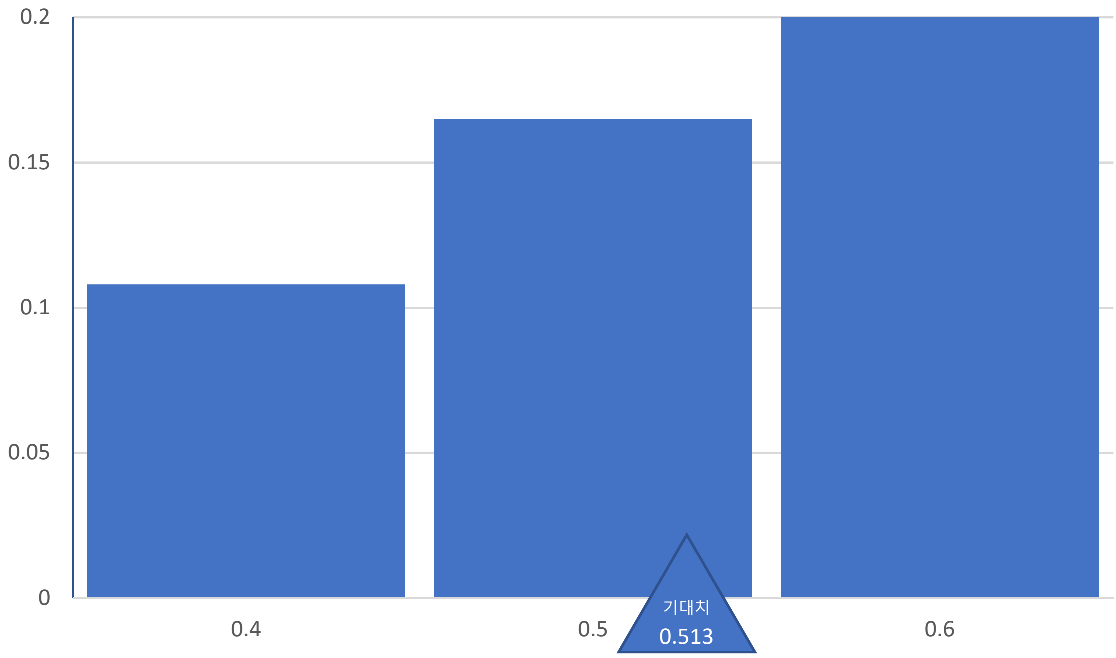
(x 의 기대치) = 0.4*0.27 + 0.5*0.33 + 0.6*0.4 = 0.513

기대치는 확률분포를 대표하는 수치로서

(첫째 아이가 여아였던 부부에게 둘째도 여아가 태어날 확률) = 0.513

베이즈 추정에서 다음에 여아가 태어날 확률이 아무런 정보가 없을 때(0.5) 보다 조금 크게 나온다.

도표 18-5 한 부부의 둘째 아이가 여아일 확률의 기대치



18-6 베타분포에서의 기대치를 구한다

연속형 확률분포에서 연속 무한개의 수치에 대해 확률밀도가 주어져 수치로 형태를 파악하기 어렵다. 기대치를 계산하려면 적분이 필요하다. 베타분포를 예를 들어 살펴보면

$y=(\text{정수})*x^{\alpha-1}(1-x)^{\beta-1}\ (0 \le x \le 1)$

x 는 사상의 바탕이 되는 수치, y 는 확률밀도이다. 베타확률분포의 기대치는 다음과 같다.

(베타분포의 기대치) = $\alpha / (\alpha+\beta)$

도표 18-6 a=1, b=1 인 베타분포의 기대치

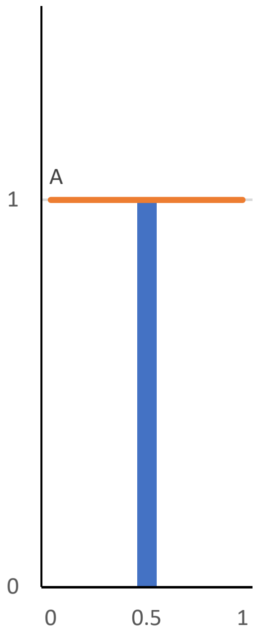


도표 18-7 a=2, b=1 인 베타분포의 확률분포도

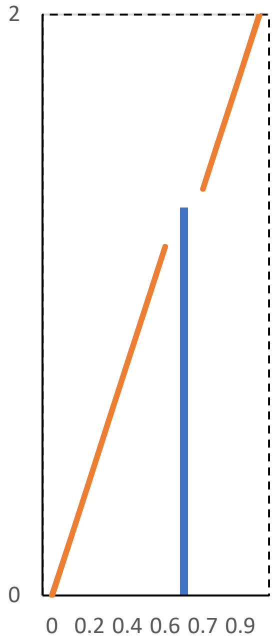


도표 18-8 a=1, b=2 인 베타분포의 기대치

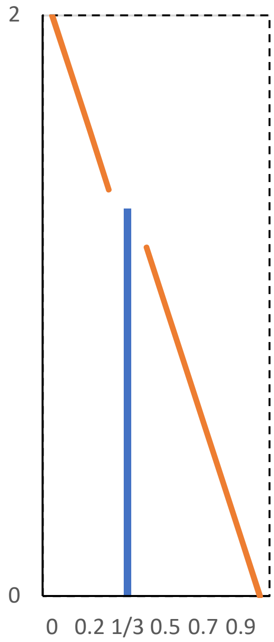


도표 18-9 a=2, b=2 인 베타분포의 기대치

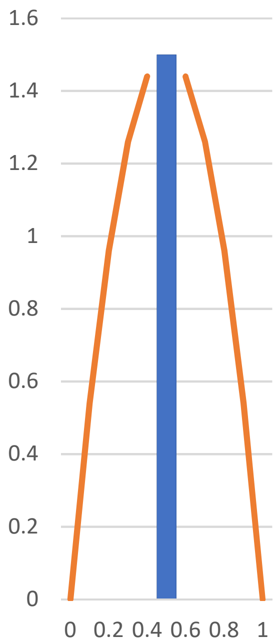
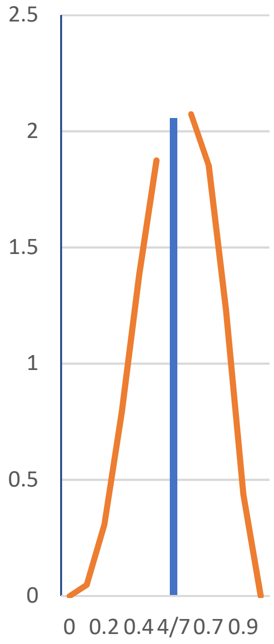


도표 18-10 a=4, b=3 인 베타분포의 확률분포도



제18강의 정리

1. 기대치는 확률분포를 하나의 수치로 대표하는 값이다
2. 기대치는 (수치)*(그 수치가 나올 확률)의 합계 로 계산된다
3. 기대치는 많은 횟수의 합계라는 의미에사 현실을 적중시킨다. 즉 N 의 값이 충분히 클 때 (N 번으로 실현한 수치의 합계)=(기대치의 N 배)가 성립한다.
4. 기대치는 확률분포도에서 야지로베에의 균형이 이루어 지는 지점이 된다
5. a,b 를 설정하는 베타분포의 기대치는 a/(a+b)

/ 연습문제

column 주관확률이란 어떤 확률인가?

제19강 확률분포도를 사용한 고도의 추정①

>>'베타분포'의 경우

19-1 여아의 사례를 더 정확하게 추정한다

베타분포를 사용한 베이즈 추정
4강의 여아가 태어날 확률은 0.4, 0.5, 0.6 세 개의 타입만 설정했지만 0~1 이하의 모든 수치를 타입으로 다루어야 한다 → 연속형 확률분포 필요 → 베타분포 사용

19-2 균등분포를 사전분포로 설정하여 추정한다

x : 어느 부부에게서 여아가 태어날 확률(타입) , $0 \le x \le 1$
타입이 3개일 땐 각 x에 수치를 사전확률로 충분. 그러나 이번엔 x가 연속 무한 개이므로 일일이 확률을 설정할 순 없고 확률밀도를 설정한다. 타입에 대한 가능성의 설정을 확률밀도로 하는 경우 이것을 '**사전분포**'라 부른다(이산형 타입의 사전확률에 해당)

사전분포로 균등분포를 사용 → 부부의 타입이 어떤 x이건 대등하다(각 확률이 같은 정도로 발생하는 상태)고 가정

타입x(x는 어느 부부에게 여아가 태어날 확률)에 대한 사전분포는,

$$y=1 \ (0 \le x \le 1)$$

이는 모든 타입x의 가능성이 확률밀도를 1이라는 의미이다(확률밀도는 확률이 아니다. 확률밀도는 x에 대한 폭을 곱하여 면적화했을 때 비로소 확률이 된다)

도표19-1 타입이 균등분포인 경우

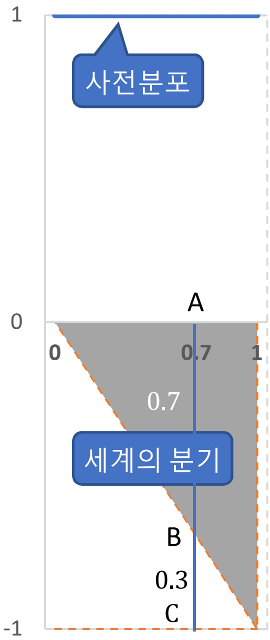


도표19-1에서 x축의 윗 부분은 사전분포이다. x축의 아래 직사각형부분은 도표4-3에 해당하는 것이다. 도표4-3에서는 3개의 세계가 모두 6개의 세계로 분기되었는데, 도표19-1에서는 수직선으로 그린 선분으로 분기되어 있다. 도표19-2 타입이 유한에서 무한 개로 변하는 모습이다.

도표4-3 여섯 개로 분기된 세계

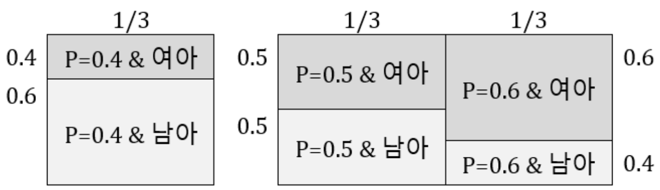


도표19-2 유한에서 무한으로

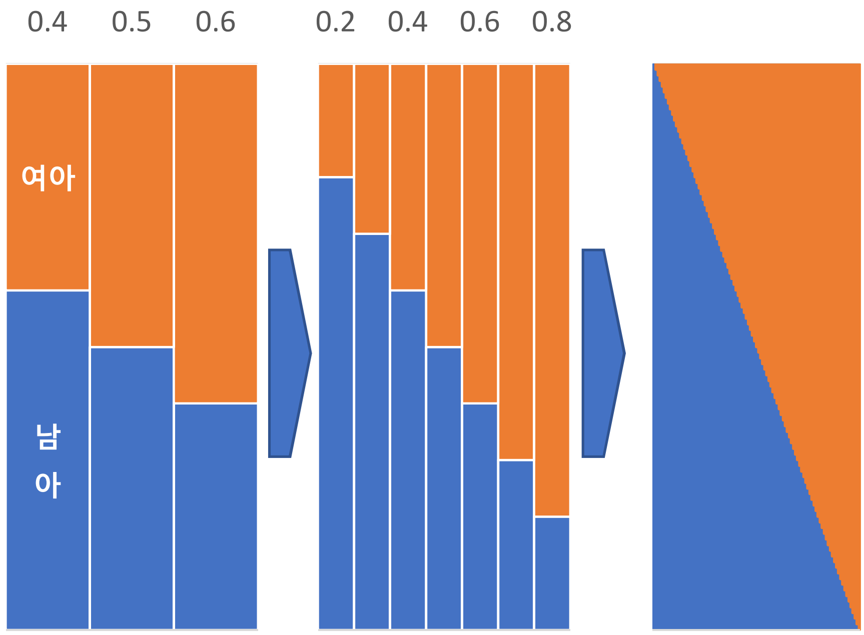
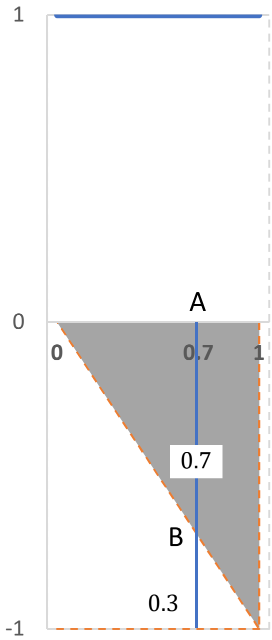


도표19-1에서 부부의 타입이 0.7(x=0.7)이라고 하자(이 부부에게 태어날 아이가 여아일 확률이 0.7이라는 가능세계) 이 부부의 첫째 아이가 여아(라는 가능세계)일 확률밀도는 0.7이다(선분AB) 반대로 남아일 확률밀도는 0.3(선분BC)

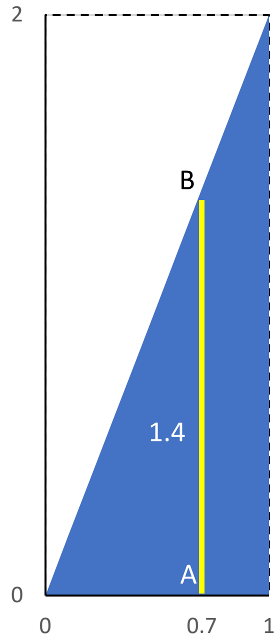
(선분AB의 길이)=(타입 x=0.7일 확률밀도) * (타입 x=0.7하에서 여아가 태어날 확률)
= (x=0.7일 때의 y값) * p(여아|x=0.7)
= 1* 0.7 = 0.7

이제 이 부부의 첫째 아이가 여아였다는 정보를 얻었다. 그러면 남아가 태어나는 가능세계를 사라진다. 도표19-1의 연한 색의 삼각형은 제거한다

도표19-3 남아가 태어났다는 가능세계가 사라진다



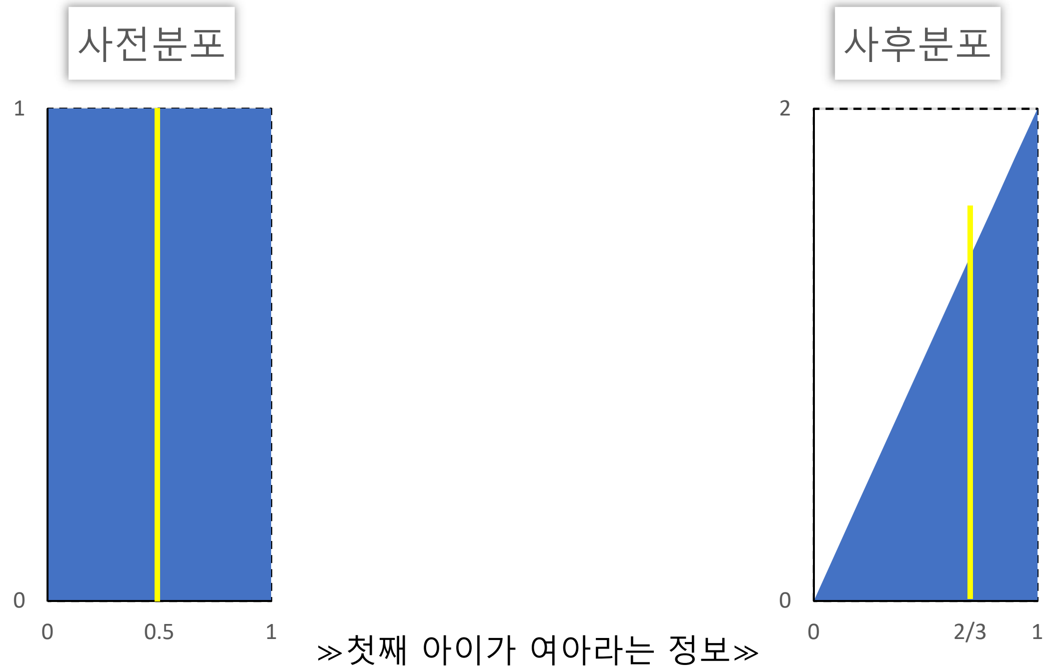
위 아래를 뒤집어 두배 늘린다



남아 가능세계가 사라지면서 정규화 조건(전 사상의 확률이 1)이 무너졌으므로(면적이 0.5로 줄었다) 다시 1의 면적이 되도록 각 선분의 비례관계를 유지하면서 확률밀도를 변경(1→2)한다(도표19-3의 오른쪽 그림) 이 오른쪽 그림은 베타분포인 α=2, β=1의 경우이다.

이 삼각형은 '부부의 첫째 아이가 여아였다'는 정보하에서 부부의 타입 x에 관한 사후분포가 된다(사후확률이 아니라 사후분포이다. 이 분포도가 확률밀도를 나타낸 것이기 때문이다)

도표19-4 사전분포와 사후분포



이 부부에게서 첫째 아이가 태어나기 전의 타입x에 대한 사전분포는 균등분포였다. 그러나 첫째 아이가 여아였다는 정보을 얻으므로써 타입x에 관한 사후확률분포는 z=2x라는 베타분포로 개정되었다. 이것은 타입x의 사후확률밀도가 x값이 클수록 커진다는 걸 의미한다.

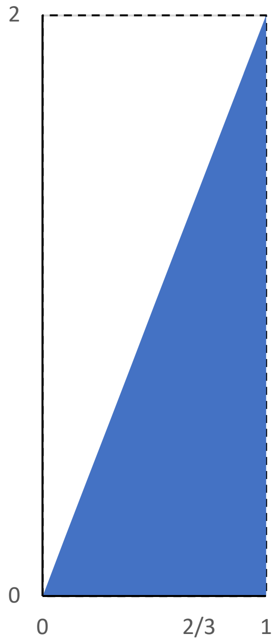
19-3 둘째도 여아였을 때의 추정

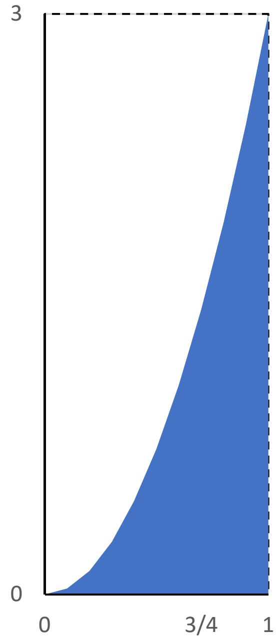
{생략}

도표19-5 사전분포와 사후분포

{생략}

도표19-6 둘째도 여아였을 때의 사후분포





≫둘째도 여아라는 정보 ≫

19-4 균등분포가 아닌 사전분포를 설정하여 추정한다

여아가 태어날 확률의 사전분포를 균등분포로 보는 것은 타당하지 않음. 타입0,타입1이 타입0.5과 대등하다고 생각하지 않는다. 0.5 주변의 타입이 일어나기 쉽고 0.5에서 먼 타입이 일어나기 어렵다는 게 바람직

사전분포를 α=2, β=2인 베타분포로 설정하면

$y=6x(1-x) \quad (0 \le x \le 1)$

이 사전분포는 타입이 0.5에서 멀수록 그 확률밀도는 낮아진다. 이때 타입 x인 부부에게 여아가 태어날 확률은

p((타입x)&(여아))

= p(타입x)*p(여아|x)

$=6x(1-x)*x$

$=6x^2(1-x)$

정규화 조건을 실행하면 사후분포(베타분포)는

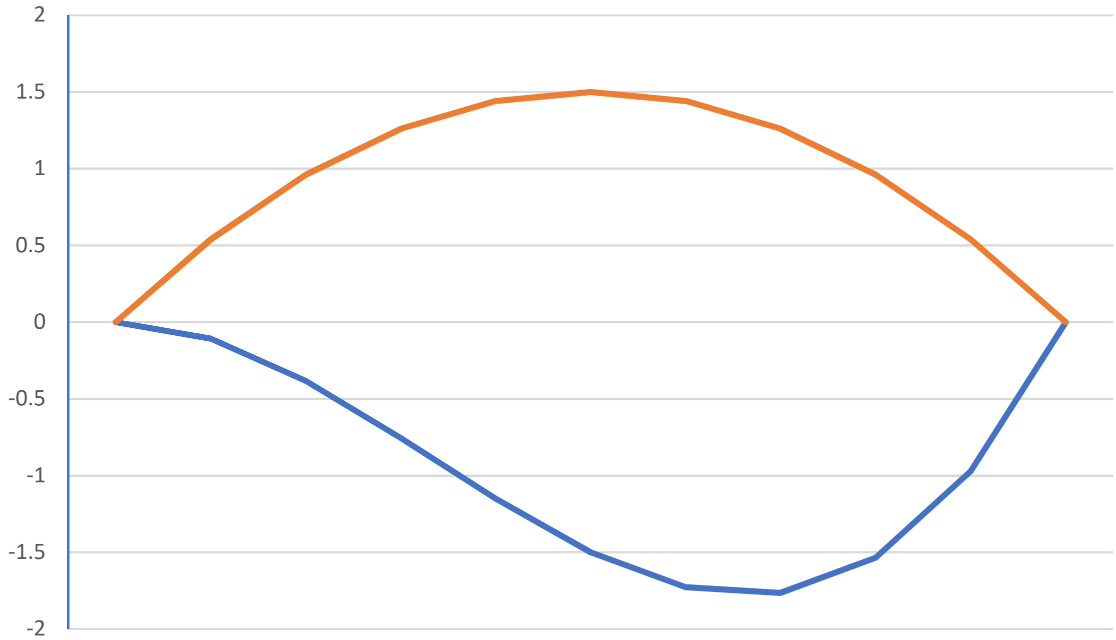
$z=12x^2(1-x)$

여아가 태어날 확률은 베타분포의 기대치는

$$\frac{\alpha}{\alpha+\beta}=\frac{3}{3+2}=\frac{3}{5}$$

이것은 균등분포를 사전분포로 했을 때보다 여아가 태어날 확률의 어림값이 0.5에 조금 더 가까운 수치 → 무난한 수치

도표19-7 균등하지 않은 베타분포로 이루어진 사전분포



19-5 베타분포를 사전분포에 사용하는 이유

베이즈 추정에서 사전분포로 베타분포를 사용하는 이유→사후분포도 베타분포가 되어야 편리

사후분포=사전분포→공액사전분포

베이즈 추정에서는 추정하고 싶은 확률모델의 공액사전분포를 사전분포로 사용하는 것이 통례

첫 번째 이유: 사전분포와 사후분포가 같은 분포를 따르면 계산이 간편

두 번째 이유: 철학적으로 사전분포≠사후분포 는 이상

제19강의 정리

1. ‘부부의 첫 아이가 여아일 때 다음 아이가 여아일 확률 x는?’ 을 추정하려는 경우, 타입을 0≤x≤1로 설정한다
2. 타입x의 사전분포를 균등분포로 설정하면 사후분포는 베타분포가 된다
3. 세계의 분기는 p(타입x)*x와 p(타입x)*(1-x)로 계산한다
4. 타입x의 확률분포가 아니라 타입 그 자체를 추정하려고 할 때는 베타분포의 기대치를 사용한다
5. 공액사전분포란 사전분포와 사후분포가 같은 분포를 따르게 하는 사전분포를 말한다
6. ‘태어날 아이가 여아인가 남아인가’에 대한 추정의 공액사전분포는 베타분포다

/ 연습문제

제20강 동전 던지기나 천체 관측에서 관찰되는

>>'정규분포'

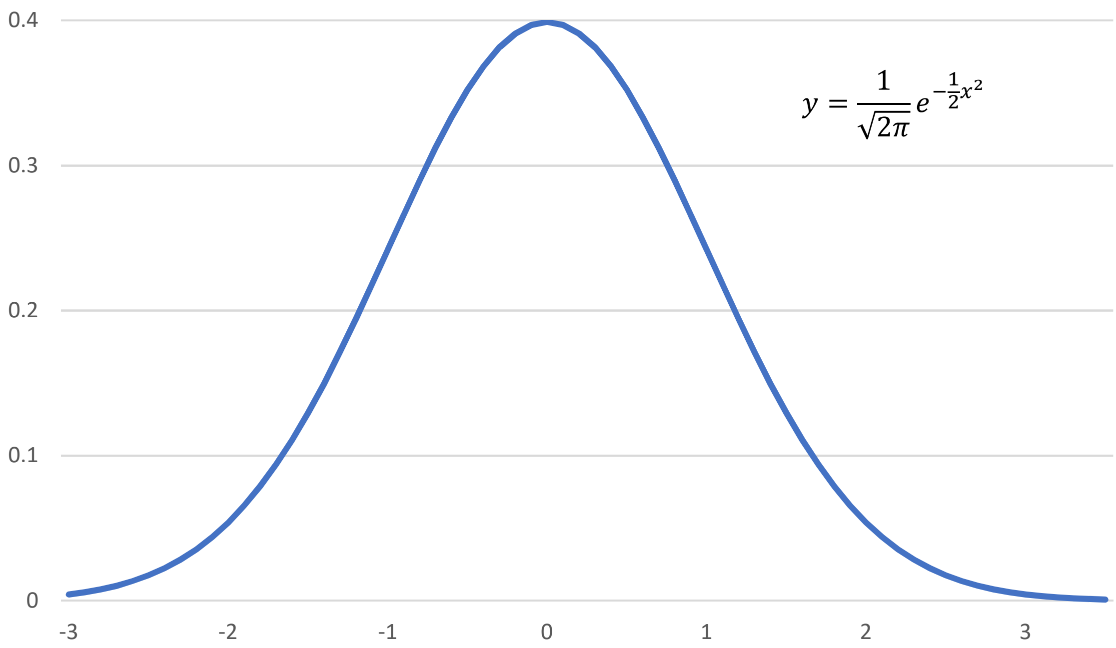
20-1 통계학의 주역인 '정규분포'

정규분포가 범용되는 이유

첫째 : 편리한 수학적 조작

둘째 : 자연계, 사회에 자주 출현하는 확률분포 - 다양한 생물종의 키 데이터, 체내 조성물(혈액 등)의 분포, 전파 노이즈, 주식수익률

도표20-1 표준정규분포



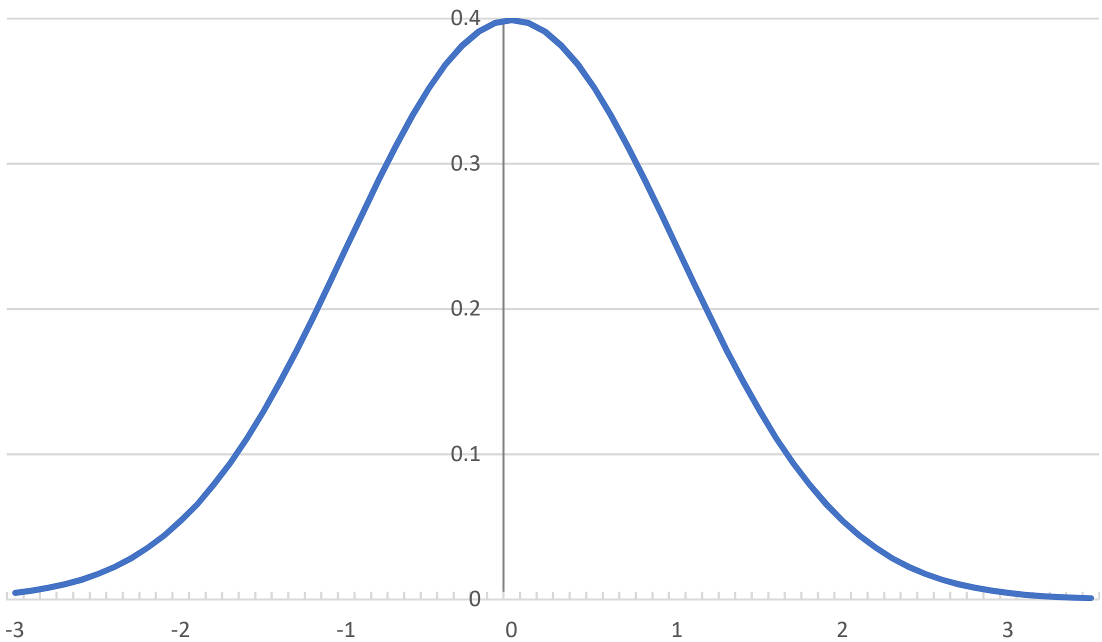
20-2 종 모양을 띤 정규분포

표준정규분포 ~ 정규분포의 대표선수격, 가로축x는 타입을 나타내는 수치, 세로축y는 그것이 출현할 확률밀도

- y축(x=0)을 축으로 좌우대칭형이다
- 종 모양을 띠며, 가장 높은 장소는 x=0인 부분이다
- 확률밀도는 아무리 큰 양수x이거나, 아무리 작은 음수 x여도 0은 되지 못한다
- x≥2 부분에서 그래프는 급격히 낮아진다. 마찬가지로 x≤-2인 부분에서도 그래프가 급격히 낮아진다

정규분포는 연속형 확률분포라서 높이 y는 확률이 아니라 확률밀도를 나타내므로 '폭을 가진 부분의 면적이 확률이 된다' -1≤x≤1을 충족하는 x가 관측될 확률은 약 0.6826이다.

도표20-2 표준정규분포의 확률



20-3 정규분포는 μ와 σ로 인해 특정된다

표준정규분포에서 일반정규분포를 얻는 방법

- 단계1: y축을 중심으로 좌우를 σ배로 늘린다. 정규화조건(전체면적 1)을 만족하기 이해 각 부분의 높이는 $\frac{1}{\sigma}$이 된다
- 단계2: 산꼭대기 x좌표는 μ가 되는 부분까지 옆으로 평행이동한다

μ와 σ의 역할

μ는 확률분포의 평균치, σ는 표준편차로 분포의 흩어짐, 퍼짐의 정도. 관측치의 평균치에서의 오차, 흔들림의 정도를 표현하는 지표

일반정규분포는 μ와 σ로 특징지울 수 있다. μ=3와 σ=2인 예제

1단계(도표20-3의 상단) – 표준정규분포 : x=0인 부분에서 꼭대기, 퍼짐은 1

2단계(도표20-3의 아래) – 표준정규분포를 좌우로 2배 확대하고 전체면적 1을 유지하기 위해 높이는 1/2로 변환(이때 σ=2이지만 μ는 아직 0이다)

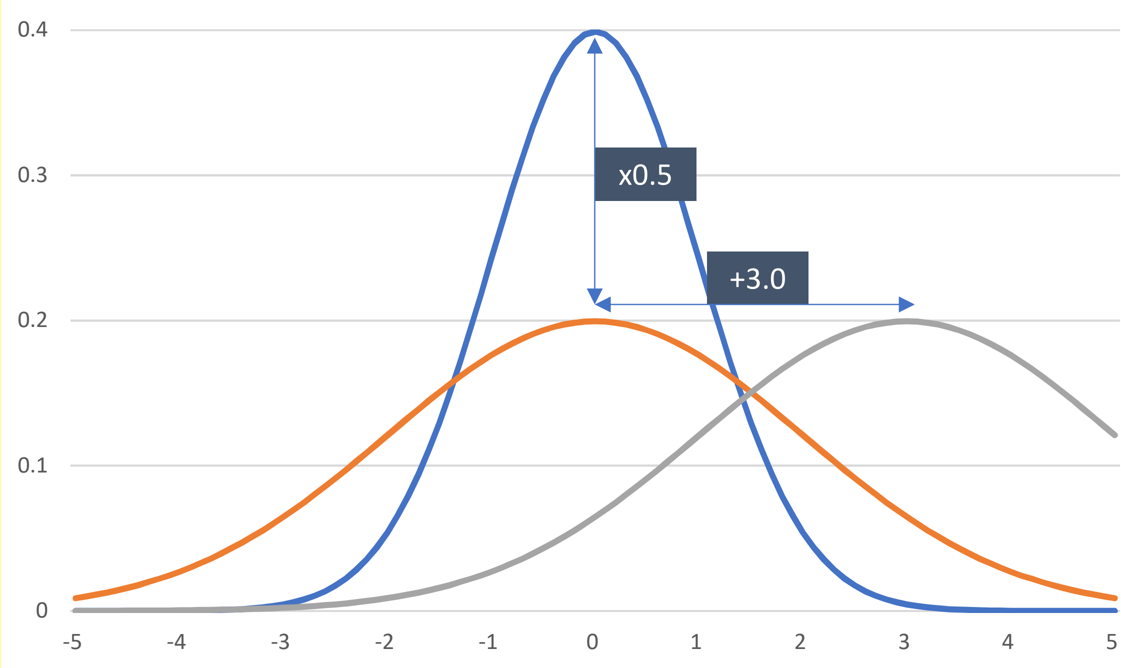
3단계(도표20-3의 오른쪽) – 2단계의 그래프를 오른쪽으로 3칸 평행이동

일반정규분포의 성질

- 정규분포는 평균μ와 표준편차σ을 주면 하나로 특정된다
- 평균μ는 분포의 평균치
- 표준편차σ는 분포의 표준편차

- 표준정규분포는 평균μ=0, 표준편차σ=1인 경우이다. 평균μ와 표준편차σ인 정규분포의 분포도는 표준정규분포의 분포도를, 면적이 바뀌지 않도록 좌우로 σ배, y방향으로 1/σ배로 늘려서 x방향으로 μ만큼 평행이동시킨 것이다

도표20-3 일반정규분포를 만드는 방법



20-4 일반정규분포의 확률은 표준정규분포의 형태로 되돌려 생각한다

표준정규분포를 이용하여 일반정규분포의 확률 구하기

μ=3, σ=2인 정규분포에서 1≤x≤5의 범위x가 관측될 확률을 구해보자.

μ=3, σ=2인 정규분포는 왼쪽으로 3칸(-3칸 평행이동) 이동하고 분포의 폭은 1/2로 줄이면 μ=3-3=0, σ=2*(1/2)=1인 표준정규분포가 된다. 변수 x를 z=(x-3)/2로 변형하면 변수z는 표준정규분포를 따르는 변수가 된다.

1≤x≤5 → 1-3=**-2≤x-3≤2**=5-3

이를 다시 2로 나누면

-2÷2≤(x-3) ÷2≤2÷2

-1≤z≤1을 얻을 수 있다.

$$\mathrm{p}(1 \leq x \leq 5) = p\left(-1 \leq \frac{x-3}{2} \leq 1\right) = p(-1 \leq z \leq 1) \cong 0.6826$$

20-5 정규분포에서 복수의 관측치의 평균은 정규분포

정규분포의 관측치 평균의 성질

평균μ, 표준편차σ인 정규분포에 따라 관측된 수치를 n개 관측하고 그 평균을 $\bar{x}$라고 하면

$\bar{x}$ =(n개의 관측치의 합) ÷ n

$\bar{x}$도 정규분포에 따라 평균과 표준편차는 다음과 같이 주어진다

평균 μ, 표준편차$\frac{\sigma}{\sqrt{n}}$

정규분포는 평균을 취해도 정규분포 그대로라는 것이다.

제20강의 정리

1. 정규분포는 자연이나 사회에서 자주 관측되는 확률분포다
2. 정규분포는 평균 μ와 표준편차 σ를 결정하면 하나로 특정된다
3. 평균 μ는 그래프의 정점의 위치를 나타내고 , 표준편차 σ는 산의 퍼진 정도를 나타낸다
4. 표준정규분포는 모든 정규분포의 중요한 근본이 된다 이것은 μ=0, σ=1 일 때를 말한다
5. 평균 μ, 표준편차 σ인 정규분포를 확률분포로 가지는 변수 x를 $z = \frac{x-\mu}{\sigma}$로 변수변환하면 변수z는 표준 정규분포를 확률분포로 가지는 변수가 된다
6. 평균μ , 표준편차 σ인 정규분포에 따라 수치를 n개 관측하고 그 평균치를 $\bar{x}$라 하면 $\bar{x}$는 평균μ, 표준편차$\frac{\sigma}{\sqrt{n}}$ 의 정규분포에 따른다

/ 연습문제

제21강 확률분포도를 사용한 고도의 추정②

>>'정규분포'의 경우

21-1 정규분포를 사전분포로 설정하여 추정한다

정규분포를 사용한 베이즈 추정
사전분포로 정규분포를 사용하는 상황

- 다루는 확률 모델이 정규분포로 주어져 있다 → 사전분포와 모델의 확률분포를 동일한 족으로 만들려는 발상 → 공액사전분포
- 상정된 타입에 대 특정 타입의 주변에 몰려 있을 법하며, 그로부터 떨어진 타입은 거의 없어 보인다 → '있을 법한 타입'이 어딘 가에 집중해 있다는 사전 선입견에 기인한다. 가령 일본성인여성 키를 100~200cm로 균등하게 가능성을 두고 모델을 설정하는 것은 타당하지 못하다. 평균160이므로 대개 그 주변일 가능성이 크고, 180/140 일 가능성은 상당히 낮다. 키에 대한 사전분포를 설정한다면 160 주변으로 두텁고, 평균에서 멀 수록 얇아 지게 하는 것이 적절 → 정규분포

21-2 정밀도가 나쁜 온도계로 욕조의 온도를 추정한다

지금껏 베이즈 추정에선 사전확률과 각 타입으로부터 얻은 정보를 &로 연결해 사상을 만들어 확률을 계산

- 암&양성, 건강&음성
- 진심&안준다, 논외&준다

정규분포를 공액사전분포로 하는 경우 '~&~' 형태를 띤 사건의 확률을 구하는 것이다.

정밀도가 낮은 온도계로 뜨거운 물의 온도를 잰다
욕조물을 적정 온도 42℃로 데우려고 한다. 다 데워졌을 즈음 온도를 쟀다. 사용한 온도계 정밀도가 낮아 계측된 온도x는 평균 (θ), 표준편차 2℃인 정규분포의 확률분포를 따른다.
지금 40℃이라면 실제온도는 얼마?

21-3 정규분포에 의한 베이즈 추정의 단계

단계1: 사전분포를 정규분포로 설정한다
(추정)하려는 것은 실제 온도θ, (정보)는 지금 40℃라는 관측결과, 적정온도 42℃로 데우려고 했으므로 평균이 42℃인 정규분포로 설정(공액사전확률), 표준편차는 3℃

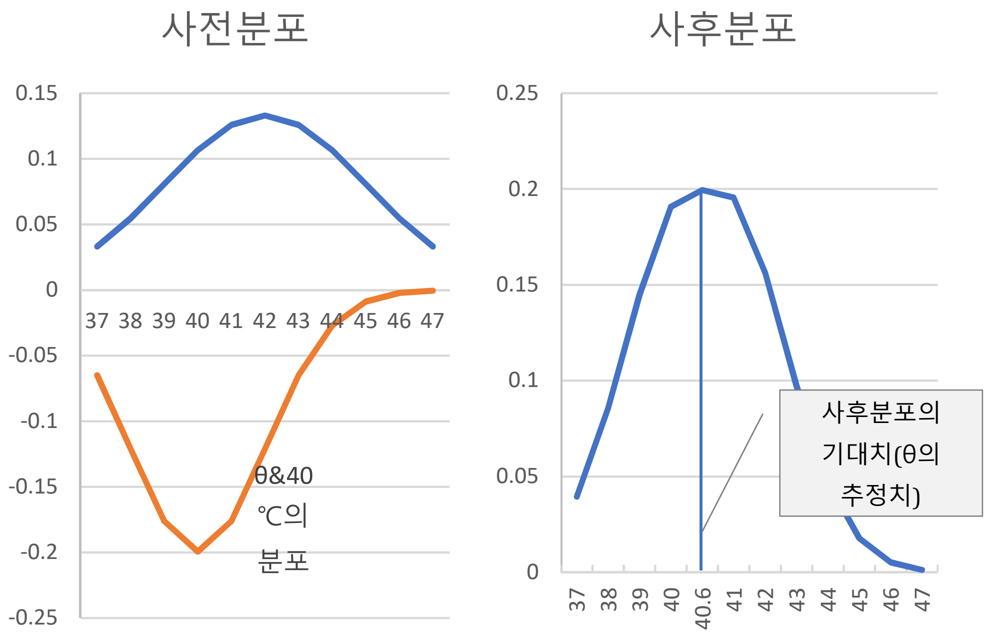
사전분포의 설정타입θ는 평균42, 표준편차3인 정규분포를 따른다

단계2: 타입θ하에서 40℃라는 온도가 계측될 확률밀도를 함수로 구한다

확률밀도를 계산→타입&정보=(실제의 온도θ)&(계측된 온도x)→그러나 (1)수 많은 조합이 난관 (2) 타입&정보의 확률은 '조건부 확률'인데, 계산이 복잡→이 책은 부득이 다음과 같이 처리

- 근원사상 : (실제온도θ)&(계측된 온도x) 조합중 'θ&40'에 대한 확률분포만 다룸
- 근원사상 'θ&40'의 분포가 정규분포가 되는 것과 평균/표준편차 계산은 결론만 제시

도표21-1 정규분포를 사용한 베이즈 추정



상단의 그래프는 θ의 사전분포(평균42, 표준편차3인 정규분포)
하단의 그래프는 타입이 θ(실제 온도θ)일 때 40℃로 계측된 확률밀도의 그래프

단계3: 사전분포를 구하고 그 분포의 기대치를 계산한다

정규화 조건이 충족되도록 비례관계를 수정하면 다음 결과가 나옴

사후분포 근원사상 'θ&40'의 분포를 정규화조건에 맞춰 비례관계를 조정하면 40℃라는 정보하에 각 θ의 사후분포를 구할 수 있다. 이 사후분포는 θ에 대한 정규분포. 이정규분포의 평균은 다음의 계산으로

$$\theta\text{의 사후분포의 기대치}=\frac{\frac{1}{3^2}*42+\frac{1}{2^2}*40}{\frac{1}{3^2}+\frac{1}{2^2}}\fallingdotseq 40.6$$

21-4 사후분포는 무엇을 의미하는 가

욕조물의 온도가 사전에는 평균42℃, 표준편차3인 정규분포를 따른다. 하나로 대표할 값이 기대치(=평균치)42℃라고 판단했다. 정밀도가 낮은 온도계로 계측을 하니 40℃라는 정보를 얻고 이를 이용해 θ에 대한 사후분포를 얻을 수 있다(도표21-1의 오른쪽 정규분포) 이 확률분포의 기대치는 40.6℃(욕조물의 온도추정치)

도표21-2 온도계의 계측 결과로부터 정보를 개정한다

사전의 선입견 θ=42℃	⇒	정보(계측결과) x=40℃	⇒	사후의 개정치 θ=40.6℃

처음에 42℃라는 예상, 그러나 계측결과 40℃를 참고하여 개정이 이루어짐. 개정치는 40에 가까움→사전분포 표준편차가 3, 온도계의 표준편차2라서 상대적으로 오차가 작은 온도계가 사전분포에 큰 영향을 미친다.

21-5 정규분포에 의한 베이즈 추정의 공식

정규분포에 따른 베이즈 추정 공식

추정하려는 θ에 대한 사전분포를 평균μ_0, 표준편차σ_0인 정규분포로 설정하고 관측하는 정보x가 평균θ, 표준편차σ인 정규분포를 따른다고 하자. 단 μ_0, σ_0, σ는 구체적인 값을 알고 있다. 즉 정보x에 대해 조건부 확률밀도 p(x|θ)는 평균 θ, 표준편차σ인 정규분포라고 한다.

(i) 정보를 1회만 관측한 경우의 공식
관측된 값을 x라고 하면,
(x의 관측하에 θ의 사후확률)p(θ|x)는 θ에 대한 정규분포가 된다
정규분포 p(θ|x)의 평균(기대치)은 $\frac{\frac{1}{\sigma_0^2}\mu_0+\frac{2}{\sigma^2}x}{\frac{1}{\sigma_0^2}+\frac{2}{\sigma^2}}$

(ii) 정보를 n회 관측한 경우의 공식
관측된 n개의 값의 평균치(관측치의 합계 ÷ n)를 $\bar{x}$라고 하면,
($\bar{x}$의 관측하에서 θ의 사후분포) p(θ| $\bar{x}$)는 θ에 관한 정규분포가 된다
정규분포 p(θ| $\bar{x}$)의 평균(기대치)은 $\frac{\frac{1}{\sigma_0^2}\mu_0+\frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}$

사후분포는 정규분포로 그 평균은 다음과 같이 계산된다. 다음은 (i)를 말로 풀어보는 것
(사전분포의 평균) ÷ (사전분포의 분산) + (관측치) ÷ (정보x의 분산)
그 값을
(사전분포의 분산의 역수) + (정보x의 분산의 역수)

21-3절의 계산을 다시 해보면
(사전분포의 평균) ÷ (사전분포의 분산) =42 ÷ 3^2
(관측치) ÷ (정보x의 분산) = 40 ÷ 2^2
(사전분포의 분산의 역수) = $1/3^2$
(정보x의 분산의 역수) = $1/2^2$

21-6 온도를 2회 쟀을 때의 베이즈 추정

욕조물 데우기 추정에서 2회 계측하는 경우 (ii)의 공식을 이용한다. 21-2절의 문제를 다음과 같이 바꾸어 보자

정밀도가 낮은 온도계로 물의 온도를 2회 잰다 욕조물을 적정온도 42℃로 데우려고 한다. 데워졌을 것이라 판단되어 온도를 쟀다. 온도계가 정밀도가 낮아 계측된 온도x는 실제 평균θ, 표준편차2℃인 정규분포를 따르는 확률분포이다. 지금 온도는 첫 번째가 40℃, 두 번째가 41℃. 실제 온도는 몇 도일까?

이 경우 두 번의 측정치의 평균은

$$\bar{x} = \frac{40+41}{2} = 40.5$$

가 된다. 공식(ii)를 적용하여(n=2임을 유의) 정규분포 p(θ| $\bar{x} = 40.5$)의 평균은 다음과 같이 계산된다.

$$\frac{\frac{1}{\sigma_0^2}\mu_0+\frac{2}{\sigma^2}x}{\frac{1}{\sigma_0^2}+\frac{2}{\sigma^2}} = \frac{\frac{1}{9}\times 42+\frac{2}{4}\times 40.5}{\frac{1}{9}+\frac{2}{4}} ≒ 40.77$$

제21강의 정리

1. 타입이 θ이고 정보가 x인 베이즈 추정에서 정보x의 확률분포 p(x|θ)가, θ를 평균으로 하는 정규분포인 경우 θ의 공액사전분포로서 정규분포를 설정한다
2. 1의 경우 사후분포 p(x|θ)도 정규분포가 된다
3. θ의 사전 분포를 평균 μ_0, 표준편차 σ_0의 정규분포로 설정하고, 관측하는 정보x가 평균θ, 표준편차 σ인 정규분포를 따른다고 하자. 단 μ_0, σ_0 ,σ는 구체적으로 알고 있다고 치자. 이때 관측된 값 x하에서 θ의 사후분포는 정규분포이며, 그 평균은

$$\frac{\frac{1}{\sigma_0^2} * \mu_0 + \frac{1}{\sigma^2} * x}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}$$

4. 복수 관측할 경우는, 관측된 n개의 값의 평균치(관측치의 합계÷n)를 라 하면, 관측된 $\bar{x}$하에서 θ의 사후분포는 정규분포이며 그 평균은

$$\frac{\frac{1}{\sigma_0^2} * \mu_0 + \frac{n}{\sigma^2} * \bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

/ 연습문제

보강▶ 베타분포의 적분계산

마치며

연습문제 해답