데이터로 배우는 통계학

자연과학대학 통계학과
장원철 교수

죽은 베이즈가 살아있는 실종자를 찾는다

1. 베이즈 정리란?

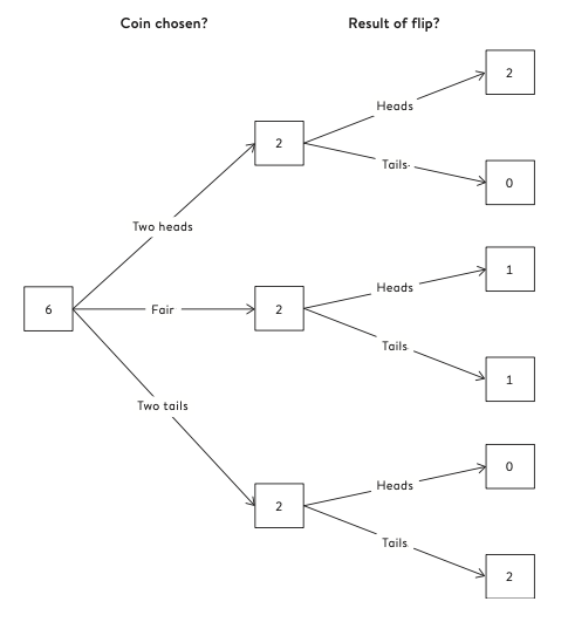
베이즈 정리(Bayes Theorem)란?

[토마스 베이즈]
(Wikipedia)

- 사전확률(prior): 어떤 특정 사건에 관한 선험적 믿음
- 가능도(likelihood): 주어진 자료를 관측할 확률
- 사후확률(posterior): 자료를 추가하여 사전확률을 업데이트한 확률
- $\text{posterior} \propto \text{prior} \times \text{likelihood}$
- 베이즈 정리:

$$\Pr(\mathrm{B}|\mathrm{A}) = \frac{\Pr(\mathrm{A}|\mathrm{B}) \cdot \Pr(\mathrm{B})}{\Pr(\mathrm{A})}$$

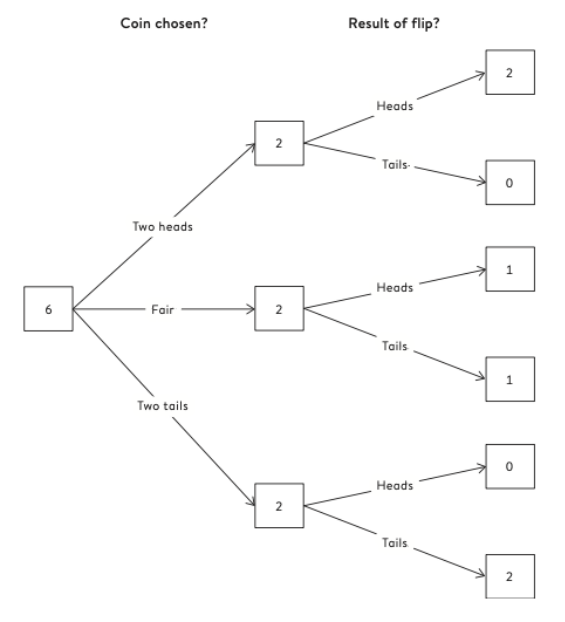
동전을 3개 던지자



[3번 동전 던지기의 기대돗수나무]
(The Art of Statistics, p309)

- 주머니에 동전이 3개 있다. 하나는 앞면만 있는 동전이고 다른 하나는 정상적으로 앞면과 뒷면이 다 있고 마지막 동전은 두면 모두 뒷면이다.
- 동전을 던졌을 때 앞면이 나왔다면 그 동전의 다른 면이 앞면일 확률은?

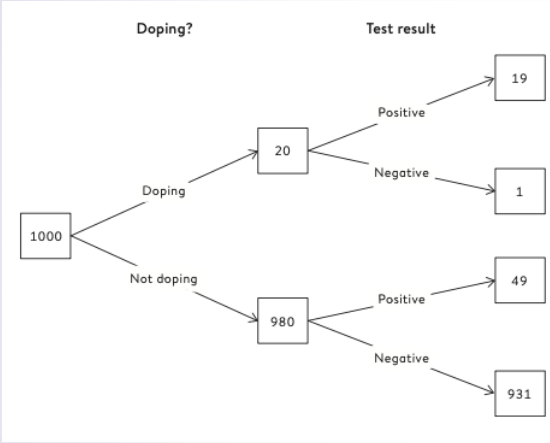
동전을 3번 던지자



[3번 동전 던지기의 기대돗수나무]
(The Art of Statistics, p309)

- 이 문제에 대한 답을 구하기 위해 기대돗수나무를 활용하자.
- 동전을 던졌을 때 앞면이 나오는 경우는 총 3번이며 이 중 2번은 양쪽 모두 앞면인 동전에서 나온다!

도핑테스트 결과는 얼마나 믿을 수 있나?

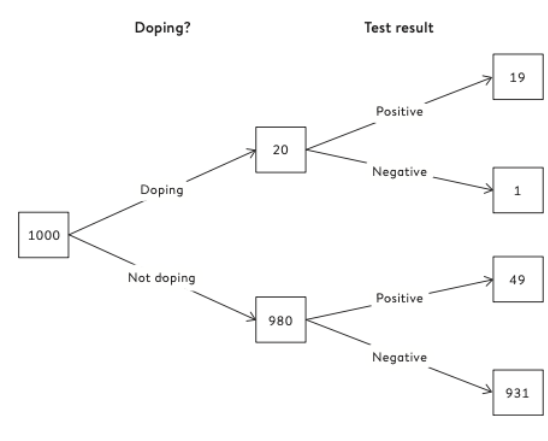


[도핑테스트 기대돗수나무]
(The Art of Statistics, p311)

- 운동경기에서 금지약물 복용 여부를 알아내기 위해 실시하는 도핑테스트의 경우 정확도가 95% 정도라고 알려져 있다. 즉 특이도와 민감도 모두 95%이다.

- 선수 50명당 1명꼴로 금지약물을 복용하고 있다고 하자. 만약 한 선수가 도핑테스트 결과가 양성이 나왔을 경우 실제 그 선수가 약물을 복용하고 있었을 확률은 얼마인가?

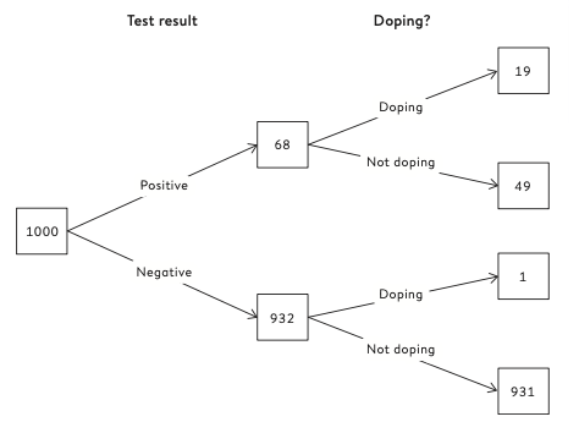
도핑테스트 결과는 얼마나 믿을 수 있나?



[도핑테스트 기대돗수나무]
(The Art of Statistics, p311)

- 1,000명의 선수에 대한 기대돗수나무를 이용하면 검사 결과가 양성인 경우가 19+49=68명이고 이중 실제로 약물을 복용한 비율은 19/68 =28%이다.

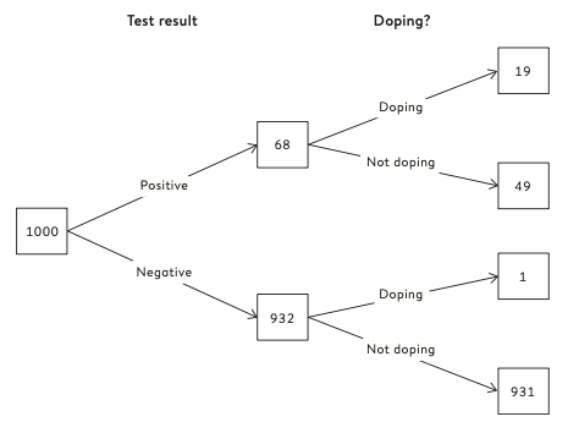
도핑테스트 결과는 얼마나 믿을 수 있나?



[도핑테스트 역기대돗수나무]
(The Art of Statistics, p309)

- 앞의 결과를 좀 더 이해하기 쉽게 표현하기 위해 왼쪽 그림과 같은 역기대돗수나무를 생각해보자.
- 이 경우 검사 결과가 먼저 나오고 그 이후에 약물복용 여부를 표시하였다.

도핑테스트 결과는 얼마나 믿을 수 있나?



[도핑테스트 역기대돗수나무]
(The Art of Statistics, p309)

- 원래 기대돗수나무는 인과적 기간순서(약물복용 후 약물검사)를 바탕으로 작성되었지만 역기대돗수나무는 사실을 알게 되는 순서(약물검사 결과 후 약물복용야부 판단)에 기초한 것이다.
- 이러한 순서 바꾸기는 베이즈 정리를 통해서 가능하다.

조건부 확률의 이해

- 우리는 $\Pr(A|B) = \Pr(B|A)$로 오해하는 경우가 많다.
- P-value은 경우 귀무가설이 참인 경우 우리가 관측된 검정통계량의 값과 같거나 더 극단적인 갑을 가질 확률로 정의하지만 많은 사람들이 관측된 검정통계량의 값을 기반으로 귀무가설이 참일 확률로 오해한다.
- 검사의 오류에서도 마찬가지로 무죄 추정의 원칙에서 이러한 증거를 모을 확률을 이러한 증거를 고려했을 때 피고가 무죄일 확률로 잘못 이해한다.

오즈(odds)와 가능도비(likelihood ratios)

- **앞의 도핑테스트 예제에서 우리가 관심이 있는 사항은 도핑테스트 결과가 양성인 경우 실제 금지약물을 복용한 비율(=19/68)이다.**
- **위의 값을 계산하기 위해 3가지 정보를 이용하였다.**
  - 전체 선수 중 약물을 복용한 비율: 1/50 또는 기대돗수나무 기준 20/1000
  - 민감도(약물을 복용한 선수 중 약물검사결과가 양성인 비율): 0.95 또는 19/20
  - 1- 특이도(약물을 복용하지 않은 선수 중 약물검사결과가 양성인 비율): 0.05 또는 49/980

오즈(odds)와 가능도비(likelihood ratios)

- 어떤 사건이 일어날 오즈는 (사건이 일어날 확률)/(1- 사건이 일어날 확률)로 정의된다.
- 앞의 도핑테스트에서 약물을 복용한 선수의 오즈는 1/49이다.
- 가능도비는 두 개의 가설 중 어느 가설이 맞는지 여부를 제시한 것으로 생각하면 된다. 예를 들어 법정시스템에서 가능도비는 (피고가 유죄일 경우 이러한 증거를 수집할 확률)/(피고가 무죄일 경우 이러한 증거를 수집할 확률)로 생각하면 된다. 여기서 우리가 귀무가설에 해당하는 경우를 분모에 집어넣는다고 생각하면 된다.
- 도핑테스트에서 가능도비는 (민감도)/(1-특이도) = 0.95/0.05 = 19로 생각할 수 있다.

베이즈 정리

- 베이즈 정리를 이용하여 사전오즈(prior odds)와 가능도비를 바탕으로 사후오즈(posterior odds)는 다음과 같이 계산할 수 있다.
- $\text{prior odds} \times \text{likelihood ratios} = \text{posterior odds}$
- 도핑테스트에서 사후오즈(검사결과가 양성일 때 선수들이 약물을 복용했을 오즈)는 1/49×19=19/49이다. 따라서 사후확률(검사결과가 양성인데 약물을 복용했을 확률)은 19/(19+49) = 19/68 = 0.28로 주어진다.

오늘의 강의 요점

- **베이즈 정리**
  - Prior
  - Likelihood
  - Posterior

출처

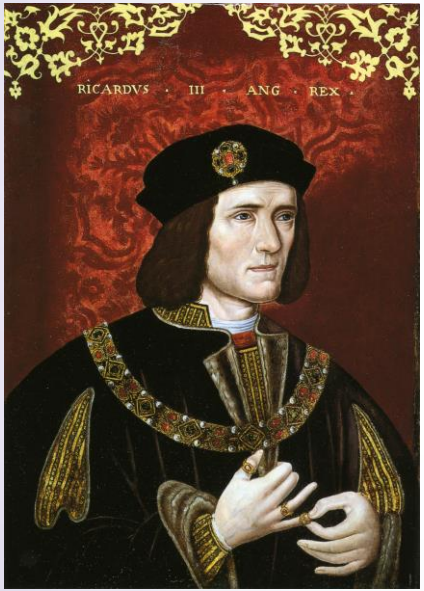
#1 Wikipedia https://en.wikipedia.org/wiki/Thomas_Bayes

#2~4 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

죽은 베이즈가 살아있는 실종자를 찾는다

2. 리처드 3세의 유해는 발견되었는가?

리처드 3세가 누구야?



[리처드 3세의 초상화]
(Wikipedia)

- 리처드 3세(1452-1485)는 잉글랜드 요크 왕가 최후의 왕으로 장미전쟁 중 보즈워스필드 전투에서 전사하였다.
- 이후 장미전쟁에서 승리한 튜더왕가의 옹호자인 셰익스피어가 “리처드 3세”라는 희극에서 리처드 3세를 사악한 꼽추로 묘사하였다.
- 리처드 3세의 시신은 훼손된 후에 레스터에 있는 그레이프리어스 수도원으로 옮겨졌다고 전해지는데 이후 이곳에 주차장이 들어섰다.

리처드 3세의 유골은 발견되었는가?

- **2012년 8월 25일 고고학자들이 이 주차장에서 리처드 3세의 유해발굴을 시작했는데 몇 시간 후에 첫 번째 해골을 발견했다. 이 해골이 리처드 3세의 유골이 맞을까?**
- **이 질문에 답변하기 위해서는 다음과 같은 가정이 성립하여야 한다.**

1. 리처드 3세가 정말 그레이프라이어스 수도원에 매장되었다.
2. 리처드 3세의 시신은 지난 527년 동안 파헤쳐져서 옮겨지지 않았다.
3. 발견된 첫 번째 해골이 우연히 리처드 3세의 유골이었다.

리처드 3세의 유골은 발견되었는가?

- 위의 가정 중 첫 번째와 두 번째 가정이 맞을 확률을 각각 50%라고 하자.
- 그리고 최대 100명의 다른 시신이 이 주차장에 묻혔다고 가정하자.
- 이 경우 3가지 가정이 모두 맞을 확률은 1/2×1/2×1/100=1/400로 상당히 낮다!
- 리처드 3세가 수도원의 성가대석에 묻혀있다고 전해지기 때문에 실제 고고학자들은 어디를 파야 할지 확신이 있었기 때문에 위의 가정이 맞을 확률을 1/40로 간주하였다.

리처드 3세의 유골은 발견되었는가?

리처드 3세의 유골임을 증빙하는 자료는 다음과 같다.

1. 고고학자들이 방사성 탄소연대 측정을 실시한 결과 해골은 95%의 확률로 1456년에서 1530년 사이의 것으로 추정되었다.

2. 그 해골의 주인은 30세가량의 남성으로 척추측만증을 앓았으며 사후에 시신이 훼손되었다는 사실도 밝혀졌다.

3. 또한 유전자 분석을 통해 리처드 3세의 근친 후손들과 모계 쪽 mDNA를 공유하고 있었다. 하지만 남성 Y 염색체의 경우 친척 관계를 뒷받침하지 못했는데 아버지를 잘못 알고 있는 경우가 종종 있기 때문에 부계가 단절된 것으로 설명할 수 있다.

가능도비의 계산

- 각각의 증거들에 대해서 가능도비는 (해골이 리처드 3세의 것일 때 이런 증거를 수집할 확률)/(해골이 리처드 3세의 것이 아닐 때 이런 증거를 수집할 확률)로 정의할 수 있으며 고고학자들은 가능도비를 다음과 같이 추산하였다.

증거	1456~1530년	해골의 나이와 성별	척추 측만증	시신 훼손	mDNA 일치	Y염색체 불일치	증거 종합
가능도비	1.8	5.3	212	42	478	0.16	6,500,000
영국 법정에서 표현방식	약한 뒷받침	약한 뒷받침	적당히 강한 뒷받침	적당한 뒷받침	적당히 강한 뒷받침	약한 반대 증거	극히 강한 뒷받침

[리처드 3세 유골 가능성에 대한 각각 증거의 가능도비]
(The Art of Statistics, p319)

리처드 3세의 유골은 발견되었는가?

- 베이즈 정리를 사용하여 사후오즈비를 계산하면 1/399×6,500,000=16,290으로 나온다.
- 고고학자들의 경우 사전오즈를 1/39로 사용하였기 때문에 사후오즈는 166,667이었고 이에 해당하는 사후확률(유골이 리처드 3세의 것일 확률)은 0.999994라는 확률이 구해진다.
- 영국의 법정에서는 가능도비를 이용하여 개별 증거의 경중을 논할 수 있는데 각각의 가능도비를 곱해서 전체 증거(즉 증거들의 결합)에 관한 가능도비를 제공하는 것은 허용하지 않는다.

영국법정에서 가능도비에 관한 표현

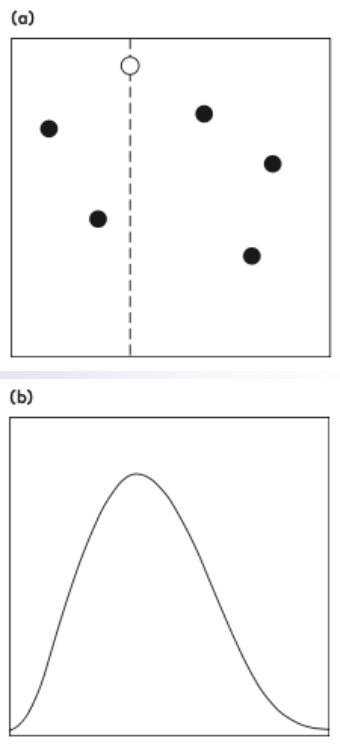
가능도비	1~10	10~100	100~1,000	1,000~10,000	10,000~100,000	100,000 이상
법정표현	약한 뒷받침	적당한 뒷받침	적당히 강한 뒷받침	강한 뒷받침	매우 강한 뒷받침	극히 강한 뒷받침

[영국법정에서 가능도비 표현방식]
(The Art of Statistics, p320)

베이즈 추론

- 베이즈 정리를 소개한 토마스 베이즈의 1763년 논문에는 다음과 같은 문제에 대한 해답을 고려하였다.
- 흰 공을 당구대 위에 무작위로 던지고 그 위치에서 수직선을 표시한 후 빨간 공 5개를 당구대 위에 던진다고 생각해 보자. 빨간 공의 정확한 위치는 알려주지 않고 단지 수직선 기준으로 왼쪽에 2개, 오른쪽에 3개가 떨어졌다고만 알려줄 경우 선의 위치는 어디인가?
- 이 문제에 대한 베이즈의 답은 3/7이다!

베이즈 추론



[베이즈 당구장]
(The Art of Statistics, p325)

- 왼쪽 그림에서 (a)는 공들의 위치를 나타내면 (b)는 수직선의 위치에 대한 사후확률분포를 나타낸다.
- 직관적인 위치 추정치는 2/5 (정확히 얘기하면 수직선 왼쪽에 떨어지는 빨간 공의 개수가 이항분포를 따른다는 가정을 하고 이 분포의 평균)이다.
- 여기서 흰 공의 위치에 대한 사전확률은 1/2로 볼 수 있다.

베이즈 추론

- 베이즈는 수직선의 위치는 (왼쪽에 놓인 빨간 공의 개수+1)/(전체 빨간 공의 개수+2)로 추정하였다.
- 즉 빨간 공을 던지기 전에는 수직선의 위치가 ½이 될 것이라는 것이 합리적 추론이며 이후 데이터를 통해서 이 위치에 대한 추론이 업데이트 되는 것이다.
- 만약 빨간 공 5개 모두 수직선의 오른쪽에 놓일 경우 수직선의 위치에 관한 추정치는 0/5이 아니라 1/7로 추정된다. 결론적으로 베이즈 추정치의 경우 데이터로 부터 얻어진 추정치 0/5가 사전확률 1/2쪽으로 움직이는 것을 알 수 있으며 이러한 현상을 shrinkage라고 한다.

베이즈 추론

- 만약 빨간 공의 개수가 엄청나게 많아질 경우 분자와 분모에 각각 더한 1과 2는 별 영향을 주지 못한다.
- 베이즈 추론의 경우 앞의 베이즈 정리를 보다 일반화한 방법으로 **사전확률분포**를 가능도를 이용하여 **사후확률분포**를 계산하는 방식이다.
- 앞의 그림에서 사후확률분포를 계산하면 하얀 공의 위치를 포함할 95% 확률구간은 (0.12, 0.78)임을 알 수 있다.
- 이 예제에서 사전확률분포는 베타분포를 사용하였고 사후확률분포 역시 베타분포이다.

오늘의 강의 요점

- **베이즈 추론**
  - Shrinkage
  - 사전확률분포
  - 사후확률분포

출처

#1 Wikipedia https://bit.ly/2JyJBDA

#2~4 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

죽은 베이즈가 살아있는 실종자를 찾는다

3. 네이트 실버와 선거 예측

죽은 베이즈가 선거분석을 살린다

- 위의 제목은 2020년에 치러진 21대 국회의원 선거 예측과 출구조사에 관여한 6명의 서울대 교수(장원철, 김용대, 박원호, 박종희, 이우주, 황승식)가 기획한 온라인 학술행사에서 한겨레 신문과 공동으로 주요 격전지 선거 예측을 담당한 박종희 교수의 발표 제목이다.
- 박종희 교수는 다수준 모형(multi-level modeling) 또는 계층적 모형(hierarchical modeling)을 이용하여 여러 가지 여론조사의 결과를 합쳐서 하나의 예측 결과를 제시하는 메타분석을 사용하였다.
- 이 메타분석의 주요 원리는 다수준 회귀분석(multi-level regression)과 사후층화(post-stratification)이다.

다수준 회귀와 사후층화

- 선거 예측에서 다수준 회귀와 사후층화의 기본 아이디어는 모든 유권자를 작은 단위로 쪼개는 것이다. 이 단위 안에 속한 사람들은 동질적이라는 가정(예를 들면 사는 지역, 나이, 성별, 이전투표 행태등이 유사하다)을 한다.
- 이러한 단위에 속하는 사람들의 숫자는 인구통계 데이터를 이용하여 추정할 수 있고 각 단위 안에서 사람들의 투표 성향은 동일하다고 간주한다.

다수준 회귀와 사후층화

- 여기서 각 단위 안의 투표성향을 예측하기 위해 회귀분석을 실시할 수 있는데 이 경우 굉장히 많은 (즉 단위의 숫자만큼) 회귀분석식이 결과로 제시된다.
- 또 다른 극단의 형태로는 모든 사람들을 다 하나의 집단 안에 있다고 간주하고 단 하나의 회귀분석만 시행하는 것이다.(위의 관점에서 얘기하면 단위 별로 실시한 회귀분석식이 사실은 모두 동일하다는 것이다.)

네이트 실버

- 네이트 실버는 미국의 선거 예측 전문가로 fivethirty-eight.com이라는 예측 전문 사이트를 운영하고 있다. 네이트 실버의 예측 방법 역시 여러 개의 여론 선거 결과를 하나로 묶는 다수준 회귀모형에 기반을 두고 있다.
- 2008년과 2012년 미국 대선과 상원의원 선거에서 상당히 높은 예측력을 보여주었으나 2016년 트럼프 당선을 예측하지 못해서 비난을 받았다.
- 하지만 선거 예측의 정확성은 당선 유무가 아니라 득표율을 얼마나 정확히 예측하는지로 결정할 수 있으며 실제 트럼프 당선을 예측했다고 주장하는 사람들 중 정확한 예측치를 제시한 사람은 아무도 없었다.

베이즈 인자(Bayes Factor)와 베이지안 가설검정

- 기존의 가설검정은 대립가설을 증명하거나 하지 못하는 경우만 결론으로 제시할 수 있다. 즉 귀무가설이 맞는지 여부에 대해서는 아무런 얘기를 할 수 없다.
- 베이지안 가설검정에서는 귀무가설이 참인지 여부에 대해서 결론을 내릴 수 있다.
- 베이즈 인자(Bayes Factor)는 2개의 가설에 대해서 개개의 증거가 지지하는 정도를 표현한 것으로 가능도비와 동일하게 간주할 수 있다.

베이즈 인자(Bayes Factor)와 베이지안 가설검정

- 베이즈 인자와 가능도비의 차이점은 가능도에 들어있는 모수(parameter)들에 대해서 이 모수의 사전분포를 이용하여 평균을 계산한다(정확히 예기하면 사전분포에 대해서 적분을 한다)는 점이다. 따라서 사전분포가 베이즈 인자 계산에 중요한 역할을 한다.

도핑테스트 결과는 얼마나 믿을 수 있나?

- 베이지안 학파의 저명한 2명의 교수 로버트 카스와 애드리언 래프터리는 베이즈 인자의 값에 대해 다음과 같은 가이드라인을 제시했다.

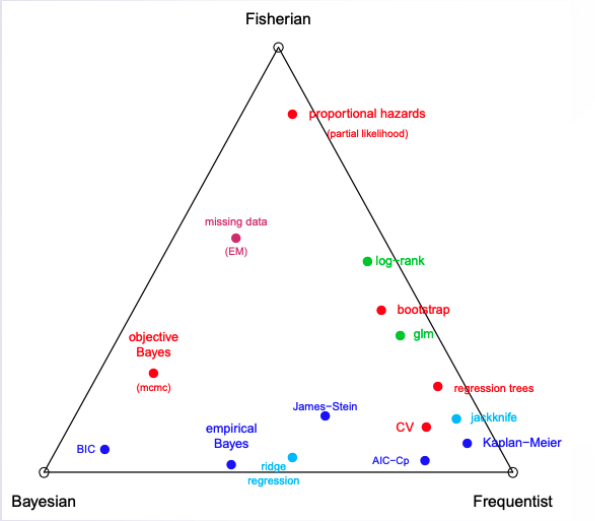
베이즈 인자	1~3	3~20	20~150	150 이상
대립가설을 지지하는 정도	간단히 한 번 언급할 정도	긍정적	강한	아주 강한

[베이즈 인자의 값에 따른 대립가설을 지지하는 정도]
(The Art of Statistics, p333)

통계학 분야의 3대 진영

- 빈도주의자(네이먼-피어슨): 가설검정 시 제1종의 오류와 제2종의 오류를 통제하고 대립가설 선정한다. 신뢰구간 바탕의 추론과 같은 모집단에서 반복적으로 표본추출을 한다는 가정이 필수이다.
- 피셔리언: 통계학의 아버지로 불리는 R. A. Fisher의 이름에서 따왔으며 유의성검정을 바탕으로 기능도 기반의 추론을 강조하는 그룹이다. 유의성 검정은 우리가 배운 가설검정에서 대립가설에 관한 설정없이 p-value를 계산하여 귀무가설의 기각 여부만 판단한다.
- 베이지안

베이지언(Bayesian), 피셔리언(Fisherian), 빈도주의자(Frequentist)



[1950년대부터 1990년대 사이 발표된 통계학 분야 15주제에 걸친 Bayesian, Frequentist, Fisherian의 영향]
(Computer Age Statistical Inference, p265)

- 왼쪽 그림은 1950년대부터 90년대까지 개발된 통계학 분야의 15개 주요 주제에 관해서 3개 진영(베이지언, 피셔리언, 빈도주의자)의 상대적 영향력을 나타낸다.
- 여기서 색깔은 컴퓨팅의 중요도를 의미하며 빨강색의 경우 핵심적, 보라색은 매우 중요, 녹색은 중요, 하늘색은 덜 중요, 청색은 중요하지 않음을 나타낸다.

오늘의 강의 요점

- **선거예측**
  - 다수준 회귀
  - 사후층화
- **베이즈 인자와 가설검정**
- **베이지안, 피셔리안, 빈도주의자**

출처

#1 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

#2 Efron and Hastice, (2016), Computer Age Statistical Inference, Cambridge University Press

죽은 베이즈가 살아있는 실종자를 찾는다

Lab 11. 베이지안 계층모형

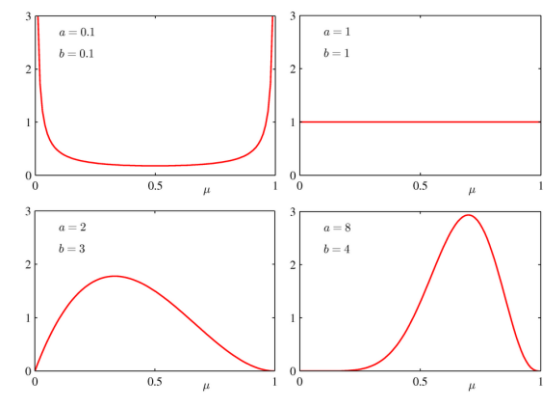
베이지안 계층모형

- 서울시 관악구에서 감염성 질환(예: 코로나19)의 유병률을 조사하고자 한다. 20명의 표본을 뽑아서 감염 여부를 조사한 결과 아무도 감염이 된 사람이 없었다.
- 다른 구의 유병률을 살펴본 결과 일반적으로 5%에서 최대 20%까지 분포하고 있으며 서울시 전체의 유병률은 10%이다. 그렇다면 위의 조사 결과만을 바탕으로 관악구의 유병률을 0%라고 추정하는 것은 합리적일까?
- 이 경우 베이지안 계층모형을 이용하여 유병률을 추정할 수 있다.

관악구의 유병률은?

- 가능도(likelihood): 먼저 우리가 알고자 하는 질병의 유병률을 θ라고 하면 20명 중 감염성 질환에 걸린 사람의 숫자는 이항분포(20, θ)를 따른다고 가정할 수 있다.
- 사전분포(prior): 유병률은 0.05에서 0.20 사이에 분포되어 있고 전체 평균이 0.10이다. 이렇게 0과 1 사이의 값을 가지는 확률분포로 베타분포가 있다. 확률의 사전 분포로 베타분포가 많이 사용된다.
- 사후분포(posterior): 이 경우 베이즈 법칙을 사용하여 사후분포를 구할 경우 사후분포 역시 베타분포를 따른다!

베타(Beta)분포

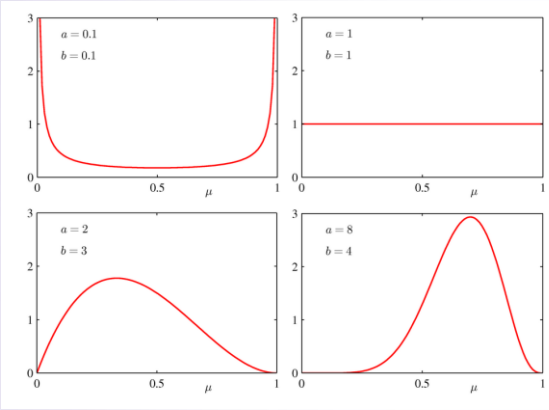


[다양한 베타분포]
(Pattern Recognition and Machine Learning, p72)

- 베타분포는 두 개의 모수 a,b 를 가지고 있으며 이 모수들의 값에 따라 옆의 그림과 같이 다양한 분포 형태를 가지게 된다. 베타분포의 평균은 $a/(a+b)$ 이다.

기타 출처 #1

베타(Beta)분포



[다양한 베타분포]
(Pattern Recognition and Machine Learning, p72)

- 유병률의 예제의 경우 각 구의 유병율의 분포가 왼쪽 하단 그림과 비슷한 형태를 가질 것이라고 가정할 수 있다.
- 이때 평균이 0.10이고 유병률의 범위가 0.05에서 0.20이라는 점을 착안해서 a, b의 값을 정할 수 있다.

기타 출처 #1

사전분포

- 우리는 유병률 θ의 사전분포가 베타분포(a,b)를 따르고 $a=2$, $b=20$이라고 가정한다.
- 이 경우 기댓값은 0.09, 최빈값(mode)는 0.05이며 유병률이 0.05와 0.20 사이에 있을 확률은 0.66이다.

사후분포

- 사후분포 역시 베타분포를 따른다. 이 경우 사후분포의 첫 번째 모수값은 2(사전분포에서 a값)+0(관악구 검사대상 중 감염병 환자 수)=2 이고 두번째 모수 값은 20(사전분포에서 b값)+20(관악구 검사대상 중 감염병에 걸리지 않은 숫자)=40이다.

- 즉 정리하면

  → 사전분포: $\theta \sim Beta(a,b)$

  → 가능도: $y|\theta \sim$ 이항분포 (n,θ)

  → 사후분포: $\theta|y \sim Beta(a+y, b+n-y)$

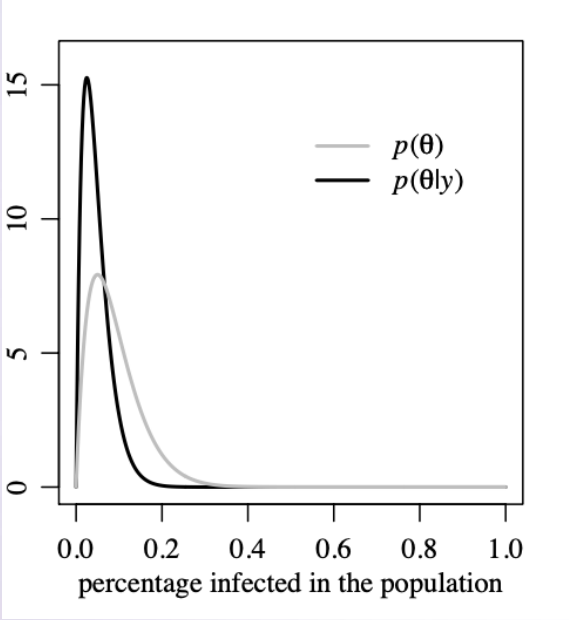
사후분포

- 사후분포의 평균$(0.048 = (a+y)/(a+b+n))$은 사전분포의 평균 $(0.09 = a/(a+b))$와 가능도 (이항분포)의 평균$(0 = y/n)$의 가중평균임을 알 수 있다.

$$\begin{aligned}\frac{a+y}{a+b+n} &= \frac{n}{a+b+n}\frac{y}{n} + \frac{a+b}{a+b+n}\frac{a}{a+b} \\ &= \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0 \qquad \left(w = a+b, \theta_0 = \frac{a}{a+b}\right)\end{aligned}$$

- 여기서 w는 degree of confidence라고 하며 이 값이 클수록 사전분포의 기댓값에 가까워짐을 알 수 있다.

그래서 유병률의 추정치는?



[사전분포와 사후분포]
(A First Course in Bayesian Statistical Methods, p4)

- 이 경우 사후분포의 평균은 0.048이며 최빈값은 0.025이다.
- 즉 관악구의 유병률에 관한 추정은 0보다는 0.048로 하는 것이 합리적으로 보인다.
- 왼쪽 그림은 사전분포와 사후분포의 변화를 보여준다.

기타 출처 #2

호세 이글레시아스의 시즌 타율은?

[호세 이글레시아스의 디트로이트 타이거 시절]
(Wikipedia)

- 호세 이글레시아스는 LA 에인절스 소속의 메이저리거 선수이다.
- 2013년 주전 첫 해의 첫 달 타율은 0.450(20타수 9안타)로 아주 좋았다.

사진 출처 #3

호세 이글레시아스의 시즌 타율은?

[호세 이글레시아스의 디트로이트 타이거 시절]
(Wikipedia)

- 참고로 코로나로 단축된 2020년도 타율도 0.373으로 좋은 편이지만 개인 통산타율은 0.278이다.
- 만약 2013년의 첫 달 성적으로 시즌 타율을 예측한다면 0.450이 될 것이고 95% 신뢰구간은 (0.229, 0.672)이다.

사진 출처 #3

호세 이글레시아스의 시즌 타율은?

- 앞의 경우와 마찬가지로 베이지안 계층모형을 사용할 수 있다.
- 이 경우 사전분포는 메이저리그 전체 선수중 시즌당 500타수 이상 기록한 선수들의 직전 3년간의 타율분포를 사용하였다. (평균 0.275, 표준편차 0.027)
- 가능도의 경우 평균이 0.450, 표준편차가 0.111인 분포를 따른다.

호세 이글레시아스의 시즌 타율은?

- 앞에서와 비슷한 방법으로 사후분포의 평균과 표준편차를 구하면 각각 0.285, 0.00069가 나온다. 이 경우는 가능도에 사용된 데이터가 훨씬 적기 때문에 사후분포의 평균은 사전분포의 평균에 훨씬 가깝다!
- 호세 이글레시아스의 2013년 시즌 최종타율은 0.303이고 4월을 제외할 경우 0.293이었다.

오늘의 강의 요점

- **베이지안 계층모형**
  - 유병률
  - 타율예측

출처

#1 Bishop, (2006), Pattern Recognition and Machine Learning, Springer

#2 P. D. Hoff, (2009), A First Course in Bayesian Statistical Methods, Springer

#3 Wikipedia https://en.wikipedia.org/wiki/Jos%C3%A9_Iglesias_(baseball)