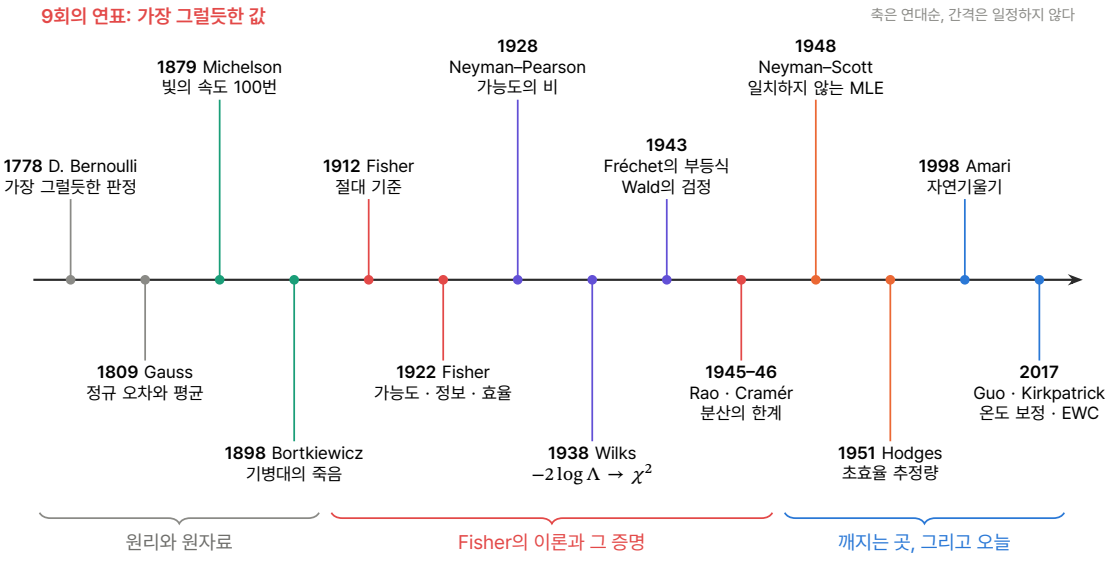
수리통계학 × 통계학의 역사

9회 강의노트

최대가능도
가장 그럴듯한 값의 이론

1778 → 1951

Daniel Bernoulli는 1778년 서로 어긋나는 관측들에서 “가장 그럴듯한” 값을 물었다. Fisher는 1922년 그 생각을 가능도라는 이름으로 이론의 중심에 놓았다. 이 노트는 최대가능도 추정량을 정의하고, Fisher 정보와 Cramér–Rao 한계, 최대가능도 추정량의 점근정규성, 가능도비 검정을 증명한 뒤, 그 이론이 무너지는 Hodges의 초효율 추정량을 본다.



다루는 것	가능도와 최대가능도 추정량 · 불변성 · 일치성(Gibbs 부등식) · 점수와 Fisher 정보 · Cramér–Rao 한계 · 점근정규성 · Newton–Raphson과 Fisher 채점 · Wilks · Wald · 점수 검정 · 초효율
원자료	von Bortkiewicz(1898)의 프로이센 기병대 · Michelson(1879)의 빛의 속도 · The Federalist의 may
노트북	`sessions/09_mle/09_mle.ipynb` — 절·정리 번호가 같다
워크북	`09_mle_workbook.pdf` (손으로 푸는 문제) · 풀이판 `_workbook_sol.pdf`
선수 지식	1–8회 (특히 2회 Gibbs 부등식, 3회 Cauchy–Schwarz, 7회 중심극한정리·Slutsky·델타 방법, 8회 검정과 구간)

이 노트를 쓰는 법

이 노트는 다른 책 없이 읽도록 썼다. 쓰이는 정의와 정리는 모두 이 노트나 앞 회차의 노트에서 증명한다. 이 과정의 범위를 넘어 증명하지 않는 사실은 **증명 없이 쓰는 사실** 상자에 따로 밝힌다.

각 절은 **역사 → 정의·정리 → 증명 → 그림 → 연습문제** 순서다. 핵심 정리의 증명은 번호 붙은 단계 ①②③으로 쓰고, 단계마다 오른쪽 끝에 근거(앞 정리의 번호나 앞 회차의 정리)를 단다. 읽을 때는 근거를 가리고 "왜 이 단계가 성립하는가"를 먼저 말해 본다. 미분과 적분의 순서를 바꾸거나 정칙조건을 쓰는 곳에는 **조건 확인** 단락이 붙는다. 나머지 정리는 **증명 전략**만 적고, 증명을 소문항으로 나눈 유도형 연습으로 돌린다. 이 연습은 건너뛰지 않고 손으로 쓴다.

근거에 붙은 [T1]–[T8]은 별책 **증명 도구 상자**의 도구 번호다. 이 회차에서는 [T2](귀류, 식별가능성), [T4](Jensen), [T6](적분 기호 아래의 미분), [T8](Taylor 전개, o_P와 O_P)를 쓴다.

연습문제는 소절 끝에 모여 있다. 난이도는 ★ 계산·확인, ★★ 증명·응용, ★★★ 도전이다. 모든 문제의 풀이가 부록 A에 있고, 문제 끝의 **풀이 ○○쪽**을 누르면 그 쪽으로 간다. 손으로 푸는 문제는 따로 **워크북**(계산 드릴, 빈칸 증명, 증명 문제, 퀄형)에 모았다.

상자의 뜻

> **정의** 대상을 처음 정하는 문장. 왼쪽 선이 보라다.

> **정리** 증명할 주장(정리·보조정리·따름정리·명제). 왼쪽 선이 빨강이다.

> **예제** 숫자를 넣어 끝까지 계산해 보는 곳. 왼쪽 선이 파랑이다.

> **가정 주의** 빼면 정리가 무너지는 가정. 왼쪽 선이 주황이다.

> **증명 없이 쓰는 사실** (예) 이 과정에서 증명하지 않고 쓰는 결과. 무엇이 참인지, 왜 그럴듯한지, 어디서 증명되는지를 밝힌다.

> **역사** 그 개념이 태어난 사람과 문제. 회색 선 옆의 작은 글씨다.

> **연습** 소절 끝의 연습문제. 가는 청록 선이고, 끝에 풀이가 있는 쪽 번호가 붙는다.

색의 뜻 수식과 그림에서 보라는 모수, 청록은 데이터·확률변수, 파랑은 통계량, 빨강은 핵심 결론이다. 노트북과 같은 색이다.

차례

0 들어가며 1

0.1 1–8회에서 가져오는 것 2

0.2 표기 2

1 가능도와 최대가능도 추정량 3

1.1 정의 3

1.2 왜 참값으로 가는가 — Gibbs 부등식 4

2 Fisher 정보와 Cramér–Rao 한계 8

2.1 점수와 정보 8

2.2 Cramér–Rao 한계 9

3 최대가능도 추정량의 점근정규성 12

4 계산 — Newton–Raphson과 Fisher 채점 16

4.1 두 반복법 16

4.2 가능도가 봉우리를 여럿 가질 때 17

4.3 적률법 대 최대가능도 17

5 가능도로 검정하기 — Wilks, Wald, Rao 20

5.1 세 검정 20

5.2 구간으로 뒤집기 22

6 데이터로 — 기병대, 빛의 속도, 그리고 “may” 24

6.1 프로이센 기병대 (1875–1894) 24

6.2 Michelson의 빛의 속도 (1879) 24

6.3 The Federalist의 “may” — 음이항의 MLE 25

7 부숴보기 — 초효율과 너무 많은 모수 27

7.1 Hodges의 초효율 추정량 (1951) . . . 27

7.2 받침이 모수에 의존하면 . . . 27

7.3 모수가 자료와 함께 늘어나면 — Neyman–Scott (1948) . . . 28

8 Beautiful insights — 손실함수 속의 가능도 30

8.1 교차엔트로피는 음의 로그가능도다 . . . 30

8.2 Fisher 정보는 곡률이고, 곡률은 거리다 . . . 30

8.3 온도 하나의 최대가능도 — 신경망의 보정 . . . 31

9 정리와 역사 노트 32

9.1 한 장 요약 . . . 32

9.2 Stigler의 명명 법칙 . . . 32

9.3 7 Pillars 지도 . . . 33

9.4 다음 회 — 여러 모수와 숨은 변수: EM (1894 → 1977) . . . 33

A 연습문제 풀이 35

0

들어가며

> **여는 질문.** 같은 별의 위치를 다섯 번 쟀는데 값이 서로 어긋난다. 그 가운데 "가장 그럴듯한" 참값은 무엇인가? 평균인가, 중앙값인가, 아니면 관측의 오차가 어떤 모양인지에 따라 달라지는가?

Daniel Bernoulli는 1778년 이 질문에 처음으로 확률의 답을 주었다. 오차의 밀도를 정하면, 참값이 θ일 때 지금의 관측들이 나올 "확률"은 밀도들의 곱이다. 그 곱을 가장 크게 하는 θ를 고르자는 것이다. 그는 반원 모양의 오차 밀도를 골랐고, 계산이 너무 어려워 실제로 쓰이지는 않았다. Euler는 같은 논문에 붙인 논평에서 원리 자체를 의심했다. Gauss(1809)는 정규 오차에서 이 원리가 평균을 준다는 것을 보였고(5회), Fisher는 1912년부터 1922년까지 이것을 "가능도"라는 이름으로 추정 이론 전체의 중심에 놓았다.

이 노트는 그 이론을 정리한다. 가능도와 최대가능도 추정량을 정의하고 그것이 참값으로 가는 이유를 Gibbs 부등식으로 보인다(1절). 추정의 정밀도에는 Fisher 정보로 정해지는 넘을 수 없는 한계가 있고(2절, Cramér–Rao), 최대가능도 추정량은 표본이 크면 그 한계에 닿는다(3절). 4절은 가능도를 수치로 최대화하는 방법을, 5절은 가능도로 만드는 세 가지 검정(가능도비, Wald, 점수)을 다룬다. 6절은 세 원자료에 쓰고, 7절은 "한계보다 나은" 추정량이라는 역설(Hodges 1951)을, 8절은 신경망의 교차엔트로피 손실과 자연기울기에서 같은 수학을 만난다.

학습 목표

1. 가능도를 정의하고, 여러 모형에서 최대가능도 추정량을 구하며, 불변성과 일치성을 증명한다.
2. 점수와 Fisher 정보의 두 표현이 같음을 보이고, Cauchy–Schwarz로 Cramér–Rao 한계를 증명한다.
3. 테일러 전개·중심극한정리·Slutsky로 최대가능도 추정량의 점근정규성을 증명한다.
4. Newton–Raphson과 Fisher 채점법을 구현하고, 가능도가 여러 봉우리를 가질 때를 본다.
5. Wilks 정리를 증명하고, 가능도비·Wald·점수 검정의 관계를 기하로 이해한다.

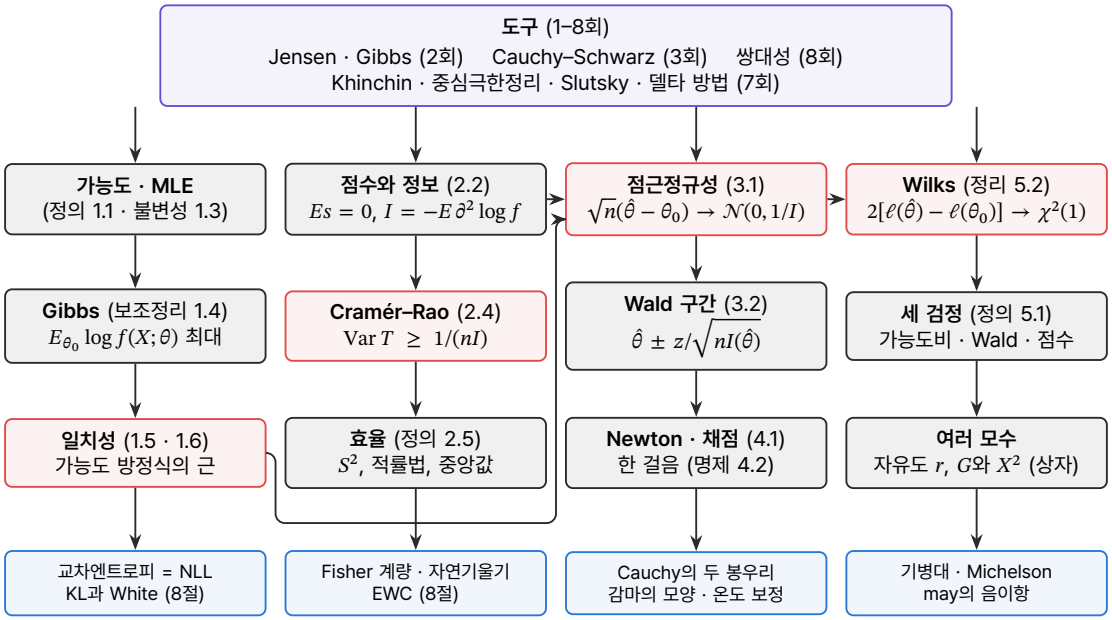


그림 0.1 이번 회차 정리들의 의존 관계. 화살표는 "증명에 쓰인다"는 뜻이다.

0.1 1–8회에서 가져오는 것

앞 회차	내용
2회 정리 5.9 · 8.3	Jensen 부등식, Gibbs 부등식 ($D(p\|q) \geq 0$)
3회 정리 4.4	Cauchy–Schwarz 부등식
5회 정리 1.5 · 6회 따름정리 2.3	평균이 가능도를 최대로 하면 정규(Gauss), $\mathrm{Var}\, S^2 = 2\sigma^4/(n-1)$
7회 정리 1.6 · 3.1 · 4.2 · 4.4	대수의 법칙(Khinchin), 중심극한정리, Slutsky, 델타 방법
8회 정리 2.4 · 3.5 · 5.1	대표본 구간, 구간과 검정의 쌍대성, Pearson의 χ^2

0.2 표기

$X_1, \ldots, X_n$이 밀도(또는 확률질량함수) $f(x;\theta)$에서 나온 iid 표본이고 참값은 θ_0이다. 로그가능도의 미분은 ℓ', ℓ''로 쓰고, 한 관측의 로그밀도의 미분은 $\frac{\partial}{\partial\theta}\log f(x;\theta)$로 쓴다. "정칙조건"은 2.1절의 상자에 모아 둔다.

1

가능도와 최대가능도 추정량

> **역사** Daniel Bernoulli의 1778년 논문(**서로 어긋나는 여러 관측의 가장 그럴듯한 판정**)의 문제의식은 이랬다. 천문학자들은 관측들의 평균을 취하는데, 모든 관측에 같은 무게를 주는 것이 옳은가? 크게 벗어난 관측은 덜 믿어야 하지 않는가? 그는 오차의 밀도로 반원 $\sqrt{r^2-x^2}$을 골라 그 곱을 최대로 하는 값을 찾았고, 그 값이 평균보다 덜 벗어난 관측에 무게를 준다는 것을 보였다. 같은 해 Euler는 논평에서 확률을 곱하는 원리를 비판했다. 130여 년 뒤 Fisher는 케임브리지를 졸업하던 1912년에 같은 원리를 다시 제안했고, 1921년 "가능도"(likelihood)라는 이름을 붙였다. 그는 가능도가 확률이 아니라는 점을 강조했다. θ에 대한 확률분포가 아니라, 자료를 고정한 θ의 함수다.

1.1 정의

정의 1.1 (가능도와 최대가능도 추정량) 관측값 $\mathbf{x}=(x_1,\dots,x_n)$이 주어질 때 θ의 함수

$$L(\theta)=L(\theta;\mathbf{x})=\prod_{i=1}^{n} f(x_i;\theta), \qquad \ell(\theta)=\log L(\theta)=\sum_{i=1}^{n}\log f(x_i;\theta)$$

를 **가능도**와 **로그가능도**라 한다. L을 Ω 위에서 최대로 하는 $\hat\theta=\hat\theta(\mathbf{x})$를 **최대가능도 추정량**(MLE)이라 한다.

이산이면 $L(\theta)$는 "θ일 때 지금의 자료가 나올 확률"이다. 연속이면 밀도이므로 확률이 아니지만, 관측을 작은 구간으로 읽으면 확률에 비례한다. 로그는 단조이므로 L과 ℓ의 최대점은 같고, 곱이 합이 되어 미분이 쉽다. 내부의 최대점은 **가능도 방정식** $\ell'(\theta)=0$을 푼다.

예제 1.2 (다섯 가지 MLE)

(a) **Bernoulli.** $\ell(p)=k\log p+(n-k)\log(1-p)$, $\ell'(p)=\frac{k}{p}-\frac{n-k}{1-p}=0$에서 $\hat p=k/n$.

(b) **Poisson.** $\ell(\lambda)=\sum x_i\log\lambda-n\lambda-\sum\log x_i!$에서 $\hat\lambda=\bar x$. 프로이센 기병대 280군단–해(4회)에서 $\hat\lambda=196/280=0.70$.

(c) **정규.** $\ell(\mu,\sigma^2)=-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2$. μ에 대해 미분하면 $\hat\mu=\bar x$, 이것을 넣고 σ^2에 대해 미분하면 $\hat\sigma^2=\frac{1}{n}\sum(x_i-\bar x)^2$. MLE는 n으로 나눈다. 불편이 아니다(6회 정리 1.1).

(d) **균등** $\mathrm{U}(0,\theta)$. $L(\theta)=\theta^{-n}\mathbb{1}\{\theta\ge x_{(n)}\}$은 $\theta\ge x_{(n)}$에서 감소하므로 $\hat\theta=x_{(n)}$. 최대점이 경계에 있어 가능도 방정식이 아니다. 8회의 독일 전차 문제와 같은 구조다.

(e) **Laplace 위치.** $\ell(\theta) = -\sum |x_i - \theta| +$ 상수를 최대로 하는 것은 표본중앙값이다(5회 연습 1.4).

정리 1.3 (불변성) $\hat\theta$가 θ의 MLE이면, 어떤 함수 g에 대해서도 $g(\hat\theta)$는 $\eta = g(\theta)$의 MLE이다. 여기서 η의 가능도는 $L^*(\eta) = \sup_{\theta : g(\theta)=\eta} L(\theta)$로 정한다.

증명. **전략** $\hat\eta = g(\hat\theta)$에서 L^*의 값을 아래와 위에서 누른다. $\hat\theta$는 $g(\theta) = \hat\eta$인 θ 가운데 하나이고, L^*는 어느 η에서도 L의 전체 최댓값을 넘지 않는다. 단계별 증명은 연습 1.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 36쪽에 있다.

정규 표준편차의 MLE는 $\hat\sigma = \sqrt{\hat\sigma^2}$이고, Poisson에서 "한 해에 아무도 죽지 않을 확률" $e^{-\lambda}$의 MLE는 $e^{-\bar x} = 0.497$이다. 불편성은 이런 변환에서 보존되지 않지만(Jensen) 최대가능도는 보존된다.

1.2 왜 참값으로 가는가 — Gibbs 부등식

로그가능도의 평균 $\frac{1}{n}\ell(\theta)$는 대수의 법칙으로 $E_{\theta_0} \log f(X;\theta)$로 간다. 그 극한이 참값 θ_0에서 최대라는 것이 일치성의 열쇠다.

보조정리 1.4 (기대 로그가능도는 참값에서 최대) $\theta \neq \theta_0$이면 $f(\cdot;\theta) \neq f(\cdot;\theta_0)$라 하자(**식별가능성**). 기댓값이 존재하면

$$E_{\theta_0} \log f(X;\theta) < E_{\theta_0} \log f(X;\theta_0) \qquad (\theta \neq \theta_0).$$

증명. **전략** 두 밀도의 비의 로그에 Jensen 부등식을 써서 기댓값과 로그의 순서를 바꾸고, 밀도의 적분이 1을 넘지 않음을 쓴다. 등호의 경우를 따로 보아 식별가능성으로 엄격한 부등식을 얻는다.

① $A = \{x : f(x;\theta_0) > 0\}$으로 두고 $Y = f(X;\theta)/f(X;\theta_0)$라 하자. $P_{\theta_0}(X \in A) = 1$이므로 Y는 확률 1로 정의되고 $Y \geq 0$이다. 받침의 정의

② log는 오목하므로

$$E_{\theta_0} \log Y \leq \log E_{\theta_0} Y.$$ Jensen, 2회 정리 5.9 [T4]

③ 오른쪽의 기댓값은

$$E_{\theta_0} Y = \int_A \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)\,dx = \int_A f(x;\theta)\,dx \leq 1.$$ 기댓값의 정의

④ ②③과 log의 단조성으로 $E_{\theta_0} \log Y \leq \log 1 = 0$, 곧 $E_{\theta_0} \log f(X;\theta) \leq E_{\theta_0} \log f(X;\theta_0)$이다. 기댓값의 선형성

⑤ 등호이면 ②와 ③이 모두 등호다. log가 순오목이므로 ②의 등호는 Y가 확률 1로 상수 c일 때뿐이고, ③의 등호와 $\log c = 0$에서 $c = 1$이다. 그러면 A 위에서 $f(x;\theta) = f(x;\theta_0)$이고 $\int_A f(x;\theta)dx = 1$이므로 A 밖에서는 $f(x;\theta) = 0$이다. Jensen의 등호 조건

⑥ ⑤는 두 밀도가 같다는 뜻이므로 식별가능성에 어긋난다. 따라서 $\theta \neq \theta_0$이면 부등식은 엄격하다. 식별가능성 [T2]

조건 확인 ④에서 $E\log Y = E\log f(X;\theta) - E\log f(X;\theta_0)$로 쓰는 데 두 기댓값의 존재를 쓴다. Jensen 부등식 ②는 $EY \le 1$로 유한하므로 $E\log Y = -\infty$인 경우에도 성립한다. "두 밀도가 같다"는 확률 0인 집합을 빼고 같다는 뜻이다. 이산이면 적분이 합으로 바뀌고, 2회 정리 8.3(Gibbs 부등식) 그대로다. ∎

두 기댓값의 차 $D(\theta_0\|\theta) = E_{\theta_0}\log\frac{f(X;\theta_0)}{f(X;\theta)} \ge 0$이 Kullback–Leibler 발산이다(5회 명제 6.2). MLE는 "자료의 분포에서 KL 발산으로 가장 가까운 모형"을 고르는 것이다(8절).

> **정리 1.5** (참값의 가능도가 이긴다) 식별가능하고 위 기댓값들이 유한하면, 고정된 $\theta \neq \theta_0$에 대해 $\lim_{n\to\infty} P_{\theta_0}\big(L(\theta_0;\mathbf{X}) > L(\theta;\mathbf{X})\big) = 1$.

증명. **전략** 로그가능도의 차를 n으로 나누면 iid 확률변수의 평균이다. 대수의 법칙(7회 정리 1.6)으로 그 평균은 보조정리 1.4의 음수 $-D(\theta_0\|\theta)$로 확률수렴하고, 음수로 가는 평균이 0 이상일 확률은 0으로 간다. 단계별 증명은 연습 1.7(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 36쪽에 있다.

그림 1.1의 왼쪽이 이 수렴이다. 표본이 커질수록 평균 로그가능도가 기대 로그가능도에 붙고, 그 꼭대기가 참값 0.7로 모인다. 오른쪽은 기병대 자료를 군단 1개, 5개, 14개로 늘려 가며 그린 상대가능도 $L(\lambda)/L(\hat\lambda)$다. 관측이 쌓일수록 가능도가 $\hat\lambda$ 둘레로 좁아진다. 그 폭이 $1/\sqrt{nI}$에 비례한다는 것이 3절의 내용이다.

> **정리 1.6** (가능도 방정식의 일치하는 근) θ가 실수이고 정리 1.5의 조건에 더해 θ_0가 모수공간의 내점이며 $\ell(\theta)$가 θ에 대해 미분가능하면, 확률이 1로 가는 사건 위에서 가능도 방정식의 근 $\hat\theta_n$이 있어 $\hat\theta_n \xrightarrow{P} \theta_0$이다.

증명. **전략** θ_0의 양옆 두 점보다 θ_0의 가능도가 높은 사건 위에서는 그 사이에 극대점이 있고, 그곳에서 미분이 0이다. 그 사건의 확률이 1로 감을 정리 1.5로 보인다.

① $a > 0$을 $[\theta_0 - a, \theta_0 + a]$가 모수공간 안에 들도록 작게 잡는다. θ_0는 내점

② $S_n = \{\ell(\theta_0) > \ell(\theta_0 - a)\} \cap \{\ell(\theta_0) > \ell(\theta_0 + a)\}$로 두면 $P(S_n^c)$는 두 사건의 여사건의 확률의 합 이하이고, 그 둘이 0으로 가므로 $P(S_n) \to 1$이다. 정리 1.5를 두 점에, Boole 부등식

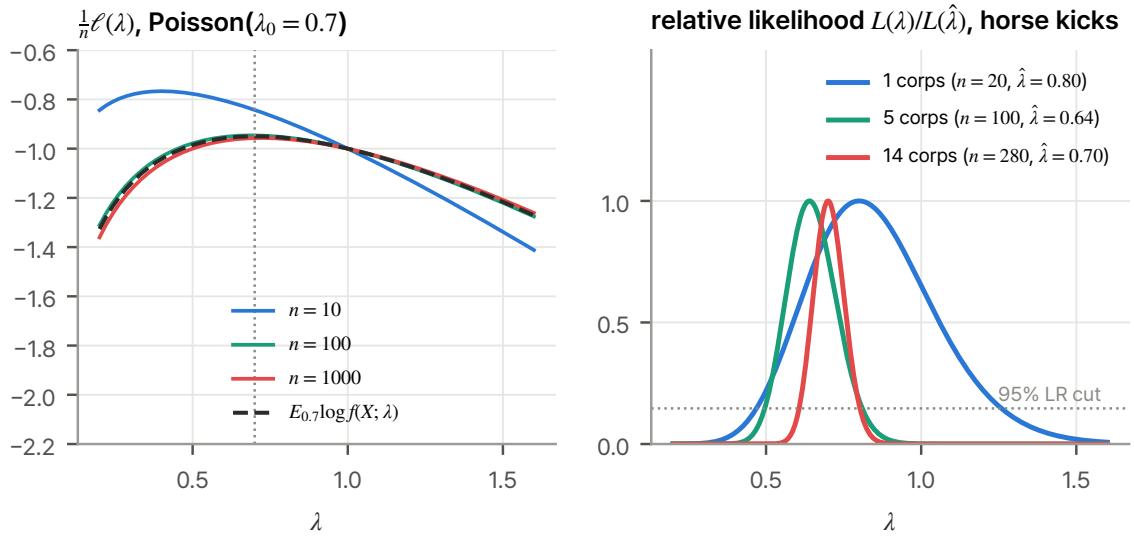


그림 1.1 왼쪽: Poisson($\lambda_0 = 0.7$) 표본의 평균 로그가능도(λ에 의존하지 않는 항은 뺐다)와 그 극한 $E_{0.7}\log f(X;\lambda)$(점선). 극한은 참값에서 최대다(보조정리 1.4). 오른쪽: 기병대 자료의 상대가능도. 점선 아래로 떨어지는 λ는 95% 가능도비 구간(5절) 밖이다.

③ S_n 위에서 연속함수 ℓ은 닫힌 구간 $[\theta_0 - a, \theta_0 + a]$에서 최댓값을 갖는다. 끝점의 값이 $\ell(\theta_0)$보다 작으므로 최대점 t는 열린 구간 안에 있다. 미분가능하면 연속, 최대·최소 정리

④ 안쪽의 최대점에서 미분가능한 함수의 미분은 0이므로 $\ell'(t) = 0$이다. Fermat의 정리

⑤ 가능도 방정식의 근 가운데 θ_0에 가장 가까운 것을 $\hat\theta_n$이라 하자(근이 없으면 아무 값). S_n 위에서는 ④의 근 t가 있으므로 $|\hat\theta_n - \theta_0| \le |t - \theta_0| < a$이다. ③④

⑥ $P(|\hat\theta_n - \theta_0| < a) \ge P(S_n) \to 1$이다. ⑤의 $\hat\theta_n$은 a에 의존하지 않으므로 이것이 충분히 작은 모든 $a > 0$에서 성립하고, $\hat\theta_n \xrightarrow{P} \theta_0$이다. 확률수렴의 정의

조건 확인 가장 가까운 근이 있으려면 근의 집합이 닫혀 있어야 한다. ℓ'가 연속이면(2.1절의 정칙조건 (R3)이 그렇다) 근의 집합은 닫힌 집합이다. ⑤의 $\hat\theta_n$은 모르는 θ_0를 쓰므로 통계량이 아니다. 정리는 "일치하는 근이 있다"는 존재의 주장이다. ■

정리는 "가능도 방정식의 **어떤** 근"이 일치한다고 말한다. 근이 하나뿐이면 그것이 MLE이고 일치한다. 근이 여럿이면(4절의 Cauchy) 가장 큰 극대점이 MLE이며, 일반적인 일치성은 Wald(1949)의 더 강한 조건이 필요하다.

연습문제

연습 1.1 (지수분포 ★) Exp (평균 θ)의 MLE가 $\bar{X}$임을 보이고, 비율 $\lambda = 1/\theta$의 MLE $1/\bar{X}$가 불편이 아님을 보여라($E(1/\bar{X}) = \frac{n}{n-1}\lambda$). 풀이 35쪽

연습 1.2 (두 모수 지수 ★★) 밀도 $\frac{1}{\beta}e^{-(x-\mu)/\beta}$ $(x \geq \mu)$의 (μ, β)의 MLE가 $(X_{(1)},\ \bar{X} - X_{(1)})$임을 보여라. 풀이 35쪽

연습 1.3 (다항분포 ★★) 다항분포의 p_j의 MLE가 N_j/n임을 Lagrange 승수로 보여라. 이것으로 8회 따름정리 5.2에서 독립 모형의 MLE가 $\hat{p}_{i\cdot}\hat{p}_{\cdot j}$임을 설명하라. 풀이 35쪽

연습 1.4 (기병대의 확률 ★) 기병대 자료에서 한 군단–해에 죽은 병사가 없을 확률의 MLE를 구하고 관측 비율 144/280과 비교하라. 풀이 35쪽

연습 1.5 (균등분포의 KL ★★) $\mathrm{U}(0,\theta_0)$와 $\mathrm{U}(0,\theta)$의 $D(\theta_0\|\theta)$를 구하라. $\theta < \theta_0$이면 무한대인 이유를 설명하고, 이것이 MLE가 $x_{(n)}$보다 작아질 수 없는 이유와 같음을 보여라. 풀이 35쪽

연습 1.6 (불변성의 증명 ★ 유도형) 정리 1.3를 다음 순서로 증명하라. $\hat{\eta} = g(\hat{\theta})$이다.

(a) $L^*(\hat{\eta}) \geq L(\hat{\theta})$임을 보여라. 상한 sup의 정의 가운데 어느 부분을 쓰는가?

(b) 모든 η에서 $L^*(\eta) \leq L(\hat{\theta})$임을 보여라.

(c) (a)와 (b)로 $L^*(\hat{\eta}) = \sup_{\eta} L^*(\eta)$를 결론지어라.

(d) g가 일대일이면 $L^*(\eta) = L(g^{-1}(\eta))$임을 확인하라. $g(\theta) = \theta^2$, $\theta \in \mathbb{R}$처럼 일대일이 아니면 왜 L^*를 상한으로 정의해야 하는지 설명하라.

풀이 36쪽

연습 1.7 (참값의 가능도가 이긴다: 증명 ★★ 유도형) $Y_i = \log\frac{f(X_i;\theta)}{f(X_i;\theta_0)}$로 두고 정리 1.5를 다음 순서로 증명하라.

(a) $\{L(\theta_0;\mathbf{X}) > L(\theta;\mathbf{X})\} = \{\bar{Y}_n < 0\}$임을 보여라.

(b) Y_i들이 iid이고 $EY_i = -D(\theta_0\|\theta) < 0$임을 보조정리 1.4로 보이고, Khinchin의 대수의 법칙(7회 정리 1.6)으로 $\bar{Y}_n \xrightarrow{P} -D$를 얻어라.

(c) $D > 0$이면 $P(\bar{Y}_n \geq 0) \leq P(|\bar{Y}_n + D| \geq D)$임을 보이고 결론을 내려라.

(d) $\operatorname{Var} Y_i = \sigma^2 < \infty$이면 $P(\bar{Y}_n \geq 0) \leq \frac{\sigma^2}{nD^2}$임을 Chebyshev 부등식(2회 정리 5.3)으로 보여라. 확률이 1로 가는 속도에 대해 무엇을 말하는가?

풀이 36쪽

2

Fisher 정보와 Cramér–Rao 한계

2.1 점수와 정보

증명 없이 쓰는 사실 (정칙조건) 이 절과 다음 절은 다음을 가정한다. (R1) 받침 $\{x : f(x;\theta) > 0\}$이 θ에 의존하지 않는다. (R2) 참값 θ_0는 모수공간의 내점이다. (R3) $f(x;\theta)$는 θ에 대해 세 번 미분가능하다. (R4) $\int f(x;\theta)\,dx = 1$을 θ에 대해 적분 기호 안에서 두 번 미분할 수 있다. (R5) θ_0의 근방에서 $\left|\frac{\partial^3}{\partial\theta^3}\log f(x;\theta)\right| \le M(x)$이고 $E_{\theta_0}M(X) < \infty$. (R4)는 지배수렴 정리로 정당화되며, 지수족(11회)을 비롯한 보통의 모형이 만족한다. 균등분포 $\mathrm{U}(0,\theta)$는 (R1)을 깬다.

정의 2.1 (점수와 Fisher 정보) 한 관측의 **점수**는 $s(x;\theta) = \frac{\partial}{\partial\theta}\log f(x;\theta)$이고, X의 **Fisher 정보**는 $I(\theta) = E_\theta\big[s(X;\theta)^2\big]$이다.

정리 2.2 (점수의 평균과 정보의 두 표현) 정칙조건 아래

$$E_\theta\, s(X;\theta) = 0, \qquad I(\theta) = \mathrm{Var}_\theta\, s(X;\theta) = -E_\theta\Big[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\Big].$$

iid 표본 n개의 정보는 $nI(\theta)$이다.

증명. **전략** 항등식 $\int f(x;\theta)\,dx = 1$을 θ에 대해 두 번 미분하고, 미분과 적분의 순서를 바꾼다(R4). 한 번 미분하면 점수의 평균이, 두 번 미분하면 정보의 두 표현이 나온다.

① 받침 위에서 $\frac{\partial f}{\partial\theta} = \frac{\partial \log f}{\partial\theta} f = s f$이다. — 로그의 미분

② $1 = \int f\,dx$를 미분하면

$$0 = \frac{d}{d\theta}\int f\,dx = \int \frac{\partial f}{\partial\theta}\,dx = \int s f\,dx = E_\theta\, s(X;\theta). \qquad \text{(R4) [T6], ①}$$

③ 곱의 미분과 ①로 $\frac{\partial}{\partial\theta}(s f) = \frac{\partial s}{\partial\theta} f + s\frac{\partial f}{\partial\theta} = \big(\frac{\partial^2 \log f}{\partial\theta^2} + s^2\big) f$이다. — 곱의 미분

④ ②의 $0 = \int s f\,dx$를 다시 미분하면

$$0 = \int \frac{\partial}{\partial\theta}(s f)\,dx = E_\theta\Big[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\Big] + E_\theta\big[s^2\big]. \qquad \text{(R4), ③}$$

⑤ $E\,s = 0$이므로 $E\,s^2 = \mathrm{Var}\,s$이고, ④로 $I(\theta) = \mathrm{Var}\,s = -E\frac{\partial^2}{\partial\theta^2}\log f$이다. — ②④

⑥ 표본의 점수는 $\ell'(\theta) = \sum_i s(X_i;\theta)$이고 항들이 독립이므로 $\operatorname{Var}\ell'(\theta) = nI(\theta)$이다.

2회 정리 2.8

조건 확인 ②와 ④에서 바꾼 순서가 정칙조건 (R4)다. 충분조건은 θ의 근방에서 $|\partial f/\partial\theta|$와 $|\partial(sf)/\partial\theta|$가 θ에 무관한 적분가능한 함수로 눌리는 것이다(도구 상자 T6의 적분 기호 아래의 미분). (R1)도 여기서 쓰인다. 받침이 θ에 의존하면 적분의 끝이 움직여 경계항이 생긴다. $\mathrm{U}(0,\theta)$이면 $\frac{d}{d\theta}\int_0^\theta \frac{1}{\theta}\,dx = 0$인데 $\int_0^\theta \frac{\partial}{\partial\theta}\frac{1}{\theta}\,dx = -\frac{1}{\theta}$이다. ∎

두 표현은 두 가지 그림을 준다. $\operatorname{Var} s$는 "점수가 자료에 따라 얼마나 흔들리는가", 곧 자료가 θ에 대해 얼마나 민감한가이고, $-E\ell''$는 "로그가능도의 꼭대기가 얼마나 뾰족한가"다. 뾰족할수록 θ를 정확히 짚는다.

예제 2.3 (정보의 예)

모형	점수 $s(x;\theta)$	$I(\theta)$	$1/(nI)$
Bernoulli(p)	$\frac{x}{p} - \frac{1-x}{1-p}$	$\frac{1}{p(1-p)}$	$\frac{p(1-p)}{n}$
Poisson(λ)	$\frac{x}{\lambda} - 1$	$\frac{1}{\lambda}$	$\frac{\lambda}{n}$
$\mathcal{N}(\mu,\sigma^2)$, μ	$\frac{x-\mu}{\sigma^2}$	$\frac{1}{\sigma^2}$	$\frac{\sigma^2}{n}$
$\mathcal{N}(\mu,\sigma^2)$, σ^2	$-\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}$	$\frac{1}{2\sigma^4}$	$\frac{2\sigma^4}{n}$
Cauchy 위치 θ	$\frac{2(x-\theta)}{1+(x-\theta)^2}$	$\frac{1}{2}$	$\frac{2}{n}$

마지막 열이 다음 정리의 한계다. 처음 셋은 표본평균의 분산과 같다.

2.2 Cramér–Rao 한계

정리 2.4 (Cramér–Rao 부등식) 정칙조건 아래 통계량 $T = T(\mathbf{X})$의 평균이 $E_\theta T = k(\theta)$이고 분산이 유한하면

$$\operatorname{Var}_\theta T \;\geq\; \frac{k'(\theta)^2}{nI(\theta)}.$$

특히 T가 θ의 불편추정량이면 $\operatorname{Var}_\theta T \geq \dfrac{1}{nI(\theta)}$이다.

증명. **전략** T와 표본의 점수 $\ell'(\theta)$의 공분산이 $k'(\theta)$임을 보이고, 공분산에 Cauchy–Schwarz 부등식을 쓴다.

① $k(\theta) = E_\theta T = \int T(\mathbf{x})\,L(\theta;\mathbf{x})\,d\mathbf{x}$이다. 기댓값의 정의

② 미분과 적분의 순서를 바꾸고 $\frac{\partial L}{\partial\theta} = \ell'(\theta)L$을 쓰면

$$k'(\theta) = \int T(\mathbf{x})\frac{\partial L}{\partial\theta}\,d\mathbf{x} = \int T(\mathbf{x})\,\ell'(\theta)\,L\,d\mathbf{x} = E_\theta\big[T\,\ell'(\theta)\big]. \qquad \text{(R4) [T6]}$$

③ $E_\theta\ell'(\theta) = 0$이므로 $E_\theta[T\ell'(\theta)] = \mathrm{Cov}\big(T, \ell'(\theta)\big) = \mathrm{Cov}\big(T - k(\theta), \ell'(\theta)\big)$이다. 정리 2.2

④ $k'(\theta)^2 = \mathrm{Cov}(T-k, \ell')^2 \le \mathrm{Var}\,T \cdot \mathrm{Var}\,\ell'(\theta)$이다. Cauchy–Schwarz, 3회 정리 4.4

⑤ $\mathrm{Var}\,\ell'(\theta) = nI(\theta) > 0$으로 나누면 부등식이다. 불편이면 $k(\theta)=\theta$, $k'(\theta)=1$이다. 정리 2.2

조건 확인 ②에서 순서를 바꾸는 함수는 L이 아니라 $T\cdot L$이다. 엄밀하게는 θ의 근방에서 $|T(\mathbf{x})\,\partial L/\partial\theta|$가 적분가능한 함수로 눌려야 한다(T6). (R4)는 $T\equiv 1$인 경우만 말하므로 이것은 T에 대한 추가 가정이다. 받침이 θ에 의존하면 ②가 깨진다(연습 2.4). ■

증명은 Cauchy–Schwarz 한 줄이다. 등호는 $T - k(\theta)$가 점수 $\ell'(\theta)$의 상수배일 때뿐이다. 그런 T가 있으면 한계를 정확히 이루는 **효율적인** 추정량이고, 11회에서 그런 모형이 지수족뿐임을 본다.

정의 2.5 (효율) θ의 불편추정량 T의 **효율**은 $\dfrac{1/(nI(\theta))}{\mathrm{Var}_\theta\,T} \le 1$이고, 효율이 1이면 **효율적**이라 한다. 두 추정량의 효율의 비를 **상대효율**이라 한다.

예제 2.6 (한계에 닿는 것과 닿지 않는 것) (a) Poisson의 $\bar X$는 $\mathrm{Var} = \lambda/n = 1/(nI)$이므로 효율적이다. Bernoulli의 $\hat p$, 정규 평균의 $\bar X$도 그렇다. (b) 정규 분산의 불편추정량 S^2은 $\mathrm{Var}\,S^2 = \frac{2\sigma^4}{n-1}$(6회 따름정리 2.3)로 한계 $\frac{2\sigma^4}{n}$보다 크다. 효율은 $\frac{n-1}{n}$이다. (c) 한계는 **불편** 추정량에 대한 것이다. MLE $\hat\sigma^2$(n으로 나눈 것)는 치우쳐 있지만 평균제곱오차가 $\frac{(2n-1)\sigma^4}{n^2}$로 S^2의 $\frac{2\sigma^4}{n-1}$보다 작다. $n=10$이면 $0.19\sigma^4$와 $0.22\sigma^4$이다(6회 연습 1.4). 평균제곱오차는 분산과 치우침의 제곱의 합이고, 치우침을 조금 허락하면 분산을 더 줄일 수 있다. 16회의 수축 추정이 이 생각을 극단까지 밀고 간다.

연습문제

연습 2.1 (정보의 계산 ★) 예제 2.3의 표에서 Poisson, 정규 분산, Cauchy의 정보를 두 표현으로 각각 계산해 같음을 확인하라. (Cauchy: $\int \frac{4u^2}{(1+u^2)^2}\cdot\frac{du}{\pi(1+u^2)} = \frac12$.) 풀이 36쪽

연습 2.2 (모수를 바꾸면 ★★) $\eta = g(\theta)$로 모수를 바꾸면 정보가 $I^*(\eta) = I(\theta)/g'(\theta)^2$임을 보여라. Bernoulli의 로그 오즈 $\eta = \log\frac{p}{1-p}$의 정보는 $p(1-p)$이다. 7회 예제 4.5의 델타 방법과 같은 답인가? 풀이 36쪽

연습 2.3 (지수 비율의 한계 ★★) Exp (비율 λ)에서 λ의 불편추정량 $\frac{n-1}{n\bar{X}}$의 분산이 $\frac{\lambda^2}{n-2}$임을 보이고($\sum X_i \sim \Gamma(n, 1/\lambda)$, 4회 연습 3.4), Cramér–Rao 한계 $\frac{\lambda^2}{n}$과 비교하라. 풀이 36쪽

연습 2.4 (균등분포에서는 ★★) $\mathrm{U}(0,\theta)$에서 $\frac{n+1}{n}X_{(n)}$은 불편이고 분산이 $\frac{\theta^2}{n(n+2)}$임을 보여라. "정보" $E[s^2] = \frac{1}{\theta^2}$로 만든 한계 $\frac{\theta^2}{n}$보다 훨씬 작다. 정리 2.4의 증명이 어디서 무너지는가? 풀이 37쪽

3

최대가능도 추정량의 점근정규성

> **역사** Edgeworth(1908)는 가능도를 최대로 하는 추정량이 큰 표본에서 가장 작은 분산을 가진다는 것을 거의 증명했지만, 그의 글은 읽히지 않았다. Fisher(1922)는 **통계학의 수학적 기초**에서 일치성, 효율, 충분성이라는 말을 만들고 최대가능도 추정량이 "효율적"이라고 주장했다. 엄밀한 증명은 Cramér(1946)의 교과서가 정리했다. 그 증명의 뼈대가 아래의 정리다.

정리 3.1 (MLE의 점근정규성) 정칙조건 아래 $0 < I(\theta_0) < \infty$이고 $\hat{\theta}_n$이 가능도 방정식의 일치하는 근(정리 1.6)이면

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}\!\left(0,\ \frac{1}{I(\theta_0)}\right).$$

증명. **전략** 가능도 방정식 $\ell'(\hat{\theta}_n) = 0$을 θ_0 둘레에서 Taylor 전개해 $\sqrt{n}(\hat{\theta}_n - \theta_0)$를 분자와 분모의 비로 쓴다 [T8]. 분자는 중심극한정리로, 분모는 대수의 법칙으로 다루고 Slutsky 정리로 합친다.

① $\hat{\theta}_n$과 θ_0 사이의 θ_n^*가 있어

$$0 = \ell'(\hat{\theta}_n) = \ell'(\theta_0) + (\hat{\theta}_n - \theta_0)\ell''(\theta_0) + \tfrac{1}{2}(\hat{\theta}_n - \theta_0)^2 \ell'''(\theta_n^*). \qquad \text{Taylor, (R3) [T8]}$$

② 정리하면

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{n^{-1/2}\,\ell'(\theta_0)}{-n^{-1}\ell''(\theta_0) - \frac{1}{2}(\hat{\theta}_n - \theta_0)\,n^{-1}\ell'''(\theta_n^*)}. \qquad (1)$$

③ 분자: $\ell'(\theta_0) = \sum_i s(X_i;\theta_0)$는 평균 0, 분산 $I(\theta_0)$인 iid의 합이므로 $n^{-1/2}\ell'(\theta_0) \xrightarrow{D} \mathcal{N}(0, I(\theta_0))$이다. 정리 2.2, 중심극한정리(7회 정리 3.1)

④ 분모의 첫 항: $-n^{-1}\ell''(\theta_0)$는 평균이 $I(\theta_0)$인 iid의 평균이므로 $\xrightarrow{P} I(\theta_0)$이다. 정리 2.2, Khinchin(7회 정리 1.6)

⑤ 분모의 둘째 항: θ_n^*가 (R5)의 근방 안에 있는 사건 위에서 $|n^{-1}\ell'''(\theta_n^*)| \le n^{-1}\sum_i M(X_i) \xrightarrow{P} EM < \infty$이다. 여기에 $\hat{\theta}_n - \theta_0 \xrightarrow{P} 0$을 곱하므로 둘째 항은 $\xrightarrow{P} 0$이다. (R5), 정리 1.6, 7회 정리 2.5 [T8]

⑥ 분모는 $\xrightarrow{P} I(\theta_0) > 0$이고, 극한은 $\frac{1}{I(\theta_0)}\mathcal{N}(0, I(\theta_0)) = \mathcal{N}(0, 1/I(\theta_0))$이다. Slutsky, 7회 정리 4.2

조건 확인 ①의 전개는 ℓ이 세 번 미분가능해야 하고(R3), ⑤는 θ_n^*가 (R5)의 근방에 있어야 쓸 수 있다. $|\theta_n^* - \theta_0| \le |\hat{\theta}_n - \theta_0|$이므로 일치성(정리 1.6)이 그 사건의 확률이 1로 감을 준다. ②에서

분모가 0인 사건의 확률은 ④⑤로 0으로 가므로 극한에 영향이 없다. ③은 $0 < I(\theta_0) < \infty$를 쓴다. ∎

정리는 MLE의 분산이 근사적으로 $\frac{1}{nI(\theta_0)}$, 곧 Cramér–Rao 한계라고 말한다. 큰 표본에서 MLE보다 정확한 (불편에 가까운) 추정량은 없다. 이것을 **점근 효율**이라 한다. 그 대가로 정칙조건이 필요하고, 7절에서 그것이 깨지는 경우를 본다.

정규 근사가 얼마나 빨리 맞는지는 모형에 따라 다르다(그림 3.1). 기병대와 같은 $n = 280$, $\lambda_0 = 0.7$의 Poisson에서는 거의 정규다. 지수분포의 비율 $\hat\lambda = 1/\bar X$는 $n = 10$에서 오른쪽으로 크게 치우쳐 있다. $1/x$가 볼록해서 $E\hat\lambda = \frac{n}{n-1}\lambda$로 위로 치우치기 때문이다(연습 1.1).

따름정리 3.2 (대표본 구간과 함수)

(a) I가 연속이면 $\hat\theta_n \pm z_{\alpha/2}/\sqrt{nI(\hat\theta_n)}$은 근사 $100(1-\alpha)\%$ 구간이다. $nI(\hat\theta_n)$ 대신 관측 정보 $-\ell''(\hat\theta_n)$를 써도 된다.

(b) g가 미분가능하고 $g'(\theta_0) \neq 0$이면 $\sqrt{n}\big(g(\hat\theta_n) - g(\theta_0)\big) \xrightarrow{D} \mathcal{N}\big(0,\ g'(\theta_0)^2 / I(\theta_0)\big)$.

증명. **전략** (a)는 $I(\hat\theta_n) \xrightarrow{P} I(\theta_0)$를 Slutsky 정리로 정리 3.1에 붙이고 8회 정리 2.4를 쓴다. 관측 정보는 정리 3.1의 증명 ④⑤를 $\hat\theta_n$에서 되풀이한다. (b)는 델타 방법이다 [T8]. 단계별 증명은 연습 3.5(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 37쪽에 있다.

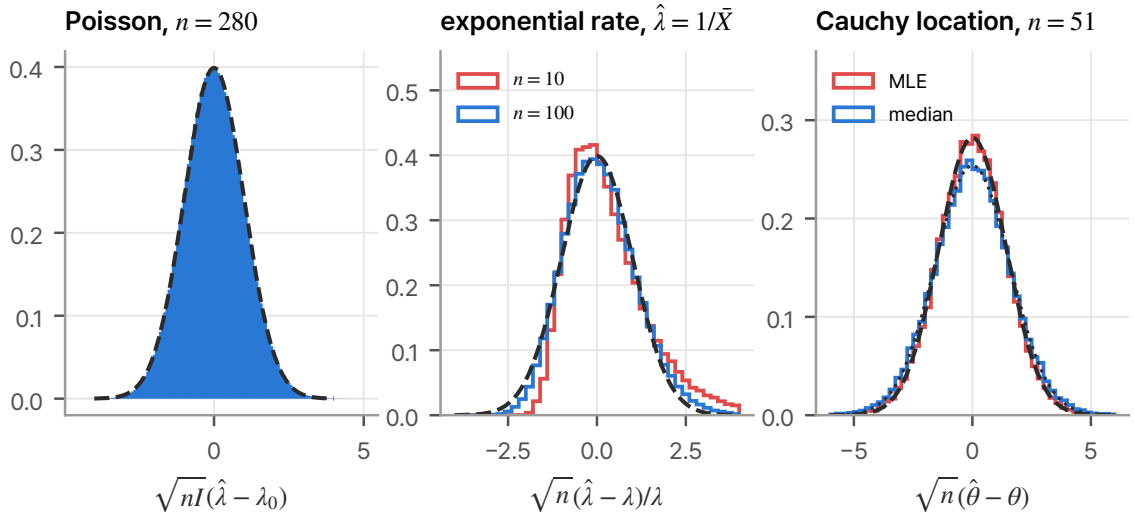


그림 3.1 MLE의 표본분포와 정리 3.1의 정규 극한(점선). 왼쪽: Poisson, $n = 280$ (정확한 분포; 막대 높이는 밀도로 맞췄다). 가운데: 지수분포의 비율, $n = 10$과 100. 오른쪽: Cauchy 위치, $n = 51$의 MLE와 표본중앙값, 그리고 극한 $\mathcal{N}(0, 2)$(점선)와 $\mathcal{N}(0, \pi^2/4)$(점).

예제 3.3 (Cauchy 위치의 MLE와 중앙값) 표준 Cauchy 위치모형에서 $I = \frac{1}{2}$이므로 MLE의 점근분산은 $\frac{2}{n}$이다. 표본중앙값은 $\frac{\pi^2}{4n} = \frac{2.47}{n}$(8회 예제 1.5)이므로 중앙값의 효율은 $\frac{2}{2.47} = 0.81$이다. $n = 51$에서 $n\mathrm{Var}$는 MLE 2.11(시뮬레이션), 중앙값 2.62(순서통계량의 밀도로 정확히 계산)로 극한값 2와 2.47보다 조금 크지만, 비는 0.81 그대로다(그림 3.1). Cauchy에서 평균은 아예 쓸 수 없지만(7회) MLE는 중앙값보다도 낫다. 가능도는 "먼 관측의 무게를 줄이는" 방법을 모형에서 스스로 찾는다. 점수 $\frac{2u}{1+u^2}$가 $|u|$가 크면 0으로 가기 때문이다. Bernoulli(1778)가 원했던 것이 이것이었다.

증명 없이 쓰는 사실 (여러 모수) $\theta \in \mathbb{R}^p$이면 Fisher 정보는 행렬 $I(\theta)_{jk} = E[s_j s_k] = -E\left[\frac{\partial^2 \log f}{\partial\theta_j \partial\theta_k}\right]$이고, 같은 정칙조건 아래 $\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{D} \mathcal{N}_p\left(\mathbf{0}, I(\theta_0)^{-1}\right)$이다. 증명은 위와 같고 테일러 전개를 벡터로 한다. 정규 (μ, σ^2)이면 $I = \mathrm{diag}\left(\frac{1}{\sigma^2}, \frac{1}{2\sigma^4}\right)$이고, 감마$(\alpha, \beta)$이면 $I = \left(\begin{smallmatrix} \psi'(\alpha) & 1/\beta \\ 1/\beta & \alpha/\beta^2 \end{smallmatrix}\right)$이다($\psi'$는 trigamma 함수). 다른 모수를 모를 때 한 모수의 점근분산은 I^{-1}의 대각성분이고, 이것은 다른 모수를 알 때의 $1/I_{jj}$보다 크거나 같다.

연습문제

연습 3.1 (Poisson의 구간 ★) 기병대 자료($n = 280$, $\hat\lambda = 0.70$)로 따름정리 3.2의 95% 구간을 구하라. $e^{-\lambda}$의 구간은? 풀이 37쪽

연습 3.2 (관측 정보 ★★) Cauchy 위치모형에서 곡률 $J(\theta) = -\ell''(\theta) = \sum \frac{2(1-u_i^2)}{(1+u_i^2)^2}$ ($u_i = x_i - \theta$)는 θ에 따라 음수가 될 수 있음을 보여라(예제 4.3의 골짜기). 극대점인 $\hat\theta$에서는 왜 음수일 수 없는가? 참값에서 $J(\theta_0)$의 기댓값은 $\frac{n}{2}$이다. 풀이 37쪽

연습 3.3 (감마의 모양 모수 ★★★) 감마(α, β)의 정보 행렬을 유도하고 β를 모를 때 $\hat\alpha$의 점근분산이 $\frac{\alpha}{n(\alpha\psi'(\alpha) - 1)}$임을 보여라. $\alpha = 2$이면 $\frac{6.90}{n}$이다($\psi'(2) = \frac{\pi^2}{6} - 1$). 풀이 37쪽

연습 3.4 (정규의 두 모수 ★★) $\mathcal{N}(\mu, \sigma^2)$의 정보 행렬이 대각임을 보이고, 그것이 6회의 $\bar X \perp S^2$와 어떻게 어울리는지 말하라. 풀이 37쪽

연습 3.5 (대표본 구간과 함수: 증명 ★★ 유도형) 따름정리 3.2를 다음 순서로 증명하라.

(a) I가 연속이면 $I(\hat\theta_n) \xrightarrow{P} I(\theta_0)$임을 연속사상 정리로 보이고, Slutsky 정리로 $\sqrt{nI(\hat\theta_n)}\,(\hat\theta_n - \theta_0) \xrightarrow{D} \mathcal{N}(0,1)$을 얻어라.

(b) (a)에서 구간이 θ_0를 포함할 확률이 $1-\alpha$로 감을 보여라(8회 정리 2.4).

(c) 평균값 정리로 $\ell''(\hat\theta_n) = \ell''(\theta_0) + (\hat\theta_n - \theta_0)\ell'''(\tilde\theta_n)$으로 쓰고, 정리 3.1의 증명 ④ ⑤를 되풀이해 $-n^{-1}\ell''(\hat\theta_n) \xrightarrow{P} I(\theta_0)$를 보여라.

(d) 따름정리의 (b)를 델타 방법(7회 정리 4.4)으로 보이고, 정리 1.3로 $g(\hat\theta_n)$이 $g(\theta)$의 MLE임을 확인하라.

풀이 37쪽

4

계산 — Newton–Raphson과 Fisher 채점

닫힌 꼴의 MLE는 드물다. Cauchy 위치, 감마의 모양, 로지스틱 회귀(13회)의 가능도 방정식은 수치로 풀어야 한다.

4.1 두 반복법

알고리즘 4.1 (Newton–Raphson과 Fisher 채점) 시작값 $\theta^{(0)}$에서 $t = 0, 1, 2, \dots$에 대해

$$\text{Newton–Raphson:}\quad \theta^{(t+1)} = \theta^{(t)} - \frac{\ell'(\theta^{(t)})}{\ell''(\theta^{(t)})}, \qquad \text{Fisher 채점:}\quad \theta^{(t+1)} = \theta^{(t)} + \frac{\ell'(\theta^{(t)})}{nI(\theta^{(t)})}$$

를 $|\theta^{(t+1)} - \theta^{(t)}|$가 충분히 작아질 때까지 되풀이한다.

Newton–Raphson은 로그가능도를 지금의 점에서 2차식으로 어림해 그 꼭대기로 옮겨 간다. 참값 근처에서는 오차가 매 단계 제곱으로 줄지만(이차 수렴), $\ell'' > 0$인 곳(골짜기)에서 시작하면 반대로 간다. Fisher 채점은 관측된 곡률 $-\ell''$ 대신 그 기댓값 nI를 쓴다. $nI > 0$이므로 언제나 오르막 방향으로 가고, 지수족에서는 두 방법이 같다. 다변수이면 나눗셈이 정보 행렬의 역행렬 곱이 된다.

명제 4.2 (한 걸음 추정량) $\tilde{\theta}_n$이 $\sqrt{n}(\tilde{\theta}_n - \theta_0)$가 확률적으로 유계인 추정량(예: 표본중앙값)이면, 그것에서 Fisher 채점을 한 번만 한 $\hat{\theta}_n^{(1)} = \tilde{\theta}_n + \frac{\ell'(\tilde{\theta}_n)}{nI(\tilde{\theta}_n)}$은 MLE와 같은 극한분포 $\mathcal{N}(0, 1/I(\theta_0))$를 갖는다.

증명. **전략** $\ell'(\tilde{\theta}_n)$을 θ_0 둘레에서 평균값 정리로 전개해 $\sqrt{n}(\hat{\theta}_n^{(1)} - \theta_0)$를 두 항으로 나눈다. 첫 항은 "확률적으로 유계 $\times$ 0으로 가는 것"이고 [T8], 둘째 항은 정리 3.1의 분자를 I로 나눈 것이다. 단계별 증명은 연습 4.4(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 38쪽에 있다.

좋은 시작점에서 한 걸음이면 충분하다. Cauchy에서 중앙값으로 시작한 한 걸음 추정량의 $n\mathrm{Var}$는 $n = 51$에서 2.13으로 MLE의 2.11과 거의 같고, 출발점인 중앙값의 2.62보다 훨씬 작다.

4.2 가능도가 봉우리를 여럿 가질 때

예제 4.3 (Cauchy 표본의 두 봉우리) 표본 $(-4.2,\ -2.8,\ 0.3,\ 0.8,\ 1.1,\ 9.5,\ 12.1)$의 Cauchy 위치 로그가능도는 $\theta = 0.62$와 $\theta = 9.31$에서 두 극대를 갖는다(그림 4.1). 오른쪽의 두 관측이 작은 봉우리를 만들고, 두 봉우리 사이 $\theta = 7.53$에 골짜기(극소)가 있다. $[-6, 15]$에서 고르게 고른 시작점 211개로 Newton–Raphson을 돌리면 전역 최대 0.62에 닿는 것은 15개뿐이고, 그 가운데 14개는 중앙값 0.8 근처의 [0, 1.3]에서 출발했다. 82개는 골짜기의 극소점 7.53에 멈추고(Newton은 $\ell' = 0$인 점을 찾을 뿐 극대와 극소를 구별하지 못한다), 100개는 발산하거나 수렴하지 않는다. Fisher 채점은 $nI > 0$ 때문에 언제나 오르막으로 가므로 모든 시작점이 자기가 속한 봉우리(7.53의 왼쪽이면 0.62, 오른쪽이면 9.31)에 닿는다. 시작점을 중앙값으로 잡고, 여러 시작점에서 돌려 가장 높은 봉우리를 고르는 것이 표준적인 처방이다. Cauchy 위치 가능도의 극대는 n개까지 생길 수 있다(Barnett, 1966).

4.3 적률법 대 최대가능도

예제 4.4 (감마의 모양) 감마(α, β)에서 적률법(4회 6절)은 $\hat{\alpha}_{\mathrm{MM}} = \bar{x}^2/s^2$이고, MLE는 β를 소거한 방정식 $\log\alpha - \psi(\alpha) = \log\bar{x} - \overline{\log x}$를 Newton–Raphson으로 푼다($\psi$는 digamma 함수). $\alpha = 2$에서 점근분산은 적률법 $\frac{2\alpha(\alpha+1)}{n} = \frac{12}{n}$, MLE $\frac{6.90}{n}$이므로(연습 3.3) 적률법의 효율은 0.58이다. 시뮬레이션에서 $n\mathrm{Var}$는 $n = 200$이면 12.1과 7.3, $n = 1000$이면 12.0과 7.0으로 극한값에 다가간다. Pearson의 적률법은 계산이 쉽지만 정보의 40%를 버린다. 4회에서 미뤄 둔 비교의 답이다.

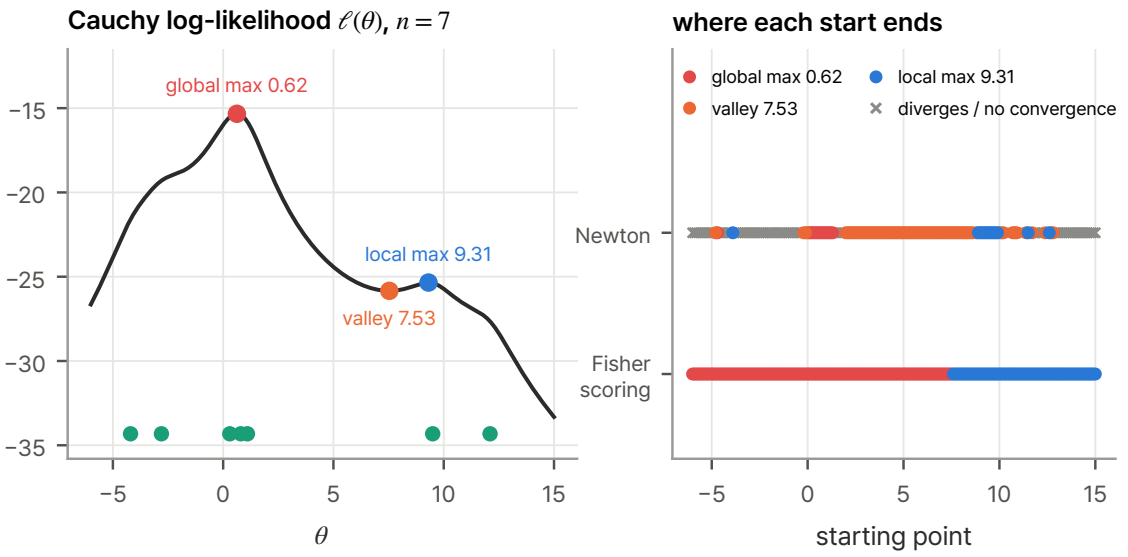


그림 4.1 왼쪽: 일곱 개의 Cauchy 관측(점)과 로그가능도. 봉우리 둘(0.62, 9.31)과 골짜기 하나(7.53)가 있다. 오른쪽: 시작점에 따른 도착점. Newton–Raphson(위)은 중앙값 근처에서만 전역 최대에 닿고 넓은 범위에서 골짜기에 멈추거나 발산한다(회색 ×). Fisher 채점(아래)은 언제나 시작점이 속한 봉우리에 닿는다.

연습문제

연습 4.1 (Poisson의 Newton ★) Poisson에서 Newton–Raphson의 한 걸음이 $\theta^{(1)} = 2\theta^{(0)} - \frac{(\theta^{(0)})^2}{\bar{x}}$ 임을 보이고, 시작값이 $2\bar{x}$보다 크면 음수로 튕겨 나감을 보여라. Fisher 채점은 한 걸음에 $\bar{x}$에 닿는다. 풀이 38쪽

연습 4.2 (이차 수렴 ★★) ℓ'가 세 번 미분가능하고 $\ell''(\hat\theta) \neq 0$이면 Newton–Raphson의 오차 $e_t = \theta^{(t)} - \hat\theta$가 $e_{t+1} \approx \frac{\ell'''(\hat\theta)}{2\ell''(\hat\theta)} e_t^2$을 만족함을 보여라. 풀이 38쪽

연습 4.3 (적률법의 분산 ★★★) 감마$(\alpha, 1)$에서 $\hat\alpha_{\mathrm{MM}} = \bar{X}^2/S^2$의 점근분산이 $\frac{2\alpha(\alpha+1)}{n}$ 임을 다변량 델타 방법(7회 정리 5.4)으로 보여라. (감마의 3차, 4차 중심적률은 2α, $3\alpha^2 + 6\alpha$.) 풀이 38쪽

연습 4.4 (한 걸음 추정량: 증명 ★★ 유도형) 명제 4.2을 다음 순서로 증명하라.

(a) 평균값 정리로 $\ell'(\tilde\theta_n) = \ell'(\theta_0) + (\tilde\theta_n - \theta_0)\ell''(\theta_n^*)$인 θ_n^*가 $\tilde\theta_n$과 θ_0 사이에 있음을 보여라.

(b) (a)를 $\hat\theta_n^{(1)}$의 정의에 넣어 다음을 보여라.

$$\sqrt{n}(\hat\theta_n^{(1)} - \theta_0) = \sqrt{n}(\tilde\theta_n - \theta_0)\Big(1 - \frac{-n^{-1}\ell''(\theta_n^*)}{I(\tilde\theta_n)}\Big) + \frac{n^{-1/2}\ell'(\theta_0)}{I(\tilde\theta_n)}$$

(c) $-n^{-1}\ell''(\theta_n^*) \xrightarrow{P} I(\theta_0)$와 $I(\tilde\theta_n) \xrightarrow{P} I(\theta_0)$를 보여라($I$는 연속). 정리 3.1의 증명 가운데 어느 단계를 되풀이하는가?

(d) 첫 항이 $\xrightarrow{P} 0$이고(7회 정리 2.5) 둘째 항이 $\mathcal{N}(0, 1/I(\theta_0))$로 감을 보여 결론을 내려라.

(e) 시작점이 $n^{1/4}(\tilde{\theta}_n - \theta_0)$만 확률적으로 유계이면 (d)의 어디가 무너지는가?

풀이 38쪽

5

가능도로 검정하기 — Wilks, Wald, Rao

> **역사** Neyman과 Pearson(1928)은 두 가설의 가능도의 비로 검정을 만들자고 제안했다. Wilks(1938)는 그 비의 로그에 -2를 곱하면 큰 표본에서 χ^2 분포를 따른다는 것을 보였다. Wald(1943)는 추정값과 귀무가설 값의 거리로, Rao(1948)는 귀무가설에서의 점수로 검정을 만들었다. 세 검정은 큰 표본에서 같은 결론을 내지만, 작은 표본과 계산의 편의에서 다르다.

5.1 세 검정

$H_0 : \theta = \theta_0$를 검정한다. 로그가능도 곡선 하나에서 세 가지 거리를 잴 수 있다(그림 5.1).

정의 5.1 (가능도비 · Wald · 점수 통계량)

$$\Lambda_{\mathrm{LR}} = 2\big[\ell(\hat\theta) - \ell(\theta_0)\big], \qquad W = nI(\hat\theta)\,(\hat\theta - \theta_0)^2, \qquad R = \frac{\ell'(\theta_0)^2}{nI(\theta_0)}.$$

가능도비는 두 점의 **높이**의 차, Wald는 두 점의 **가로** 거리(를 곡률로 잰 것), 점수는 귀무가설 점에서의 **기울기**다.

정리 5.2 (Wilks의 정리) 정리 3.1의 조건 아래 $H_0 : \theta = \theta_0$가 참이면

$$\Lambda_{\mathrm{LR}} = 2\big[\ell(\hat\theta_n) - \ell(\theta_0)\big] \xrightarrow{D} \chi^2(1),$$

그리고 W와 R도 $\chi^2(1)$로 분포수렴하며 세 통계량의 차는 $\xrightarrow{P} 0$이다.

증명. **전략** 이번에는 ℓ을 $\hat\theta_n$ 둘레에서 Taylor 전개한다 [T8]. $\ell'(\hat\theta_n) = 0$이므로 1차항이 사라지고, 가능도비는 "곡률 × 거리의 제곱"이 된다. 세 통계량을 모두 $I(\theta_0)\,n(\hat\theta_n - \theta_0)^2$으로 바꿔 비교한다.

① $\hat\theta_n$과 θ_0 사이의 θ_n^{**}가 있어

$$\ell(\theta_0) = \ell(\hat\theta_n) + \tfrac{1}{2}(\theta_0 - \hat\theta_n)^2 \ell''(\theta_n^{**}).$$ Taylor, $\ell'(\hat\theta_n) = 0$ [T8]

② 따라서

$$\Lambda_{\mathrm{LR}} = \big[-n^{-1}\ell''(\theta_n^{**})\big] \cdot \big[\sqrt{n}(\hat\theta_n - \theta_0)\big]^2.$$ ①

③ 첫 인수는 $\xrightarrow{P} I(\theta_0)$이다. 정리 3.1의 증명 ④ ⑤

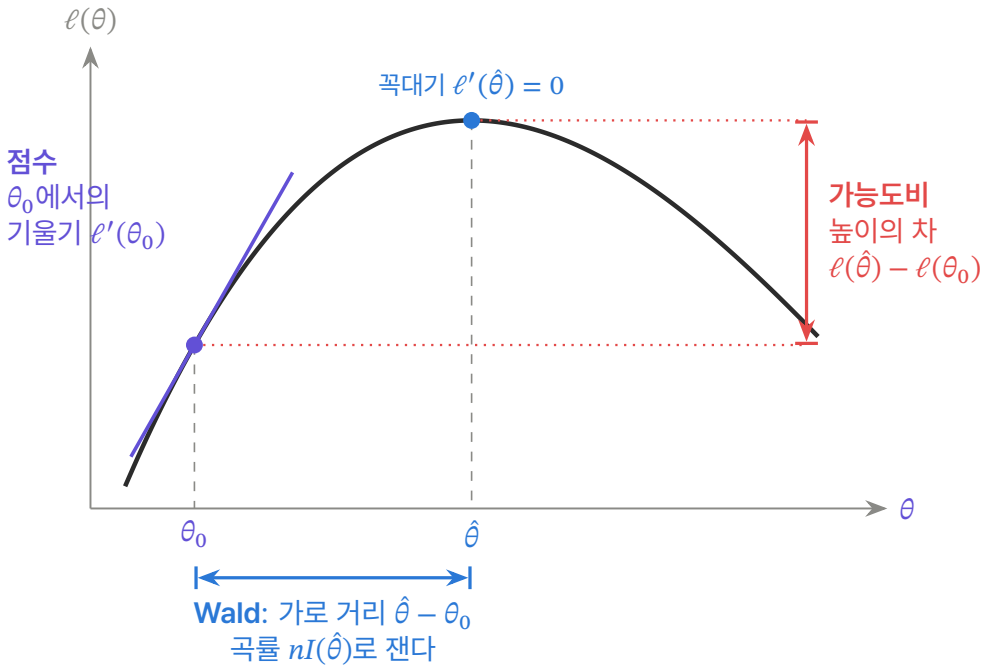


그림 5.1 세 검정의 기하. 로그가능도 $\ell(\theta)$에서 가능도비는 꼭대기와 θ_0의 높이 차, Wald는 $\hat\theta$와 θ_0의 가로 거리, 점수는 θ_0에서의 접선의 기울기를 잰다. 로그가능도가 이차식이면 셋은 같다.

④ 둘째 인수는 $\mathcal{N}(0, 1/I(\theta_0))$의 제곱으로, 곧 $\frac{1}{I(\theta_0)}\chi^2(1)$로 분포수렴한다.

정리 3.1, 7회 연습 2.4

⑤ Slutsky 정리로 $\Lambda_{\mathrm{LR}} \xrightarrow{D} I(\theta_0)\cdot\frac{1}{I(\theta_0)}\chi^2(1) = \chi^2(1)$이다. 7회 정리 4.2

⑥ $W = I(\hat\theta_n)\cdot n(\hat\theta_n-\theta_0)^2$이고 $I(\hat\theta_n)\xrightarrow{P} I(\theta_0)$이므로 같은 극한이다. $R = \big(n^{-1/2}\ell'(\theta_0)\big)^2/I(\theta_0)$은 정리 3.1의 증명 ③으로 $\chi^2(1)$이다. 연속사상, Slutsky

⑦ 정리 3.1의 증명 ①에서 $n^{-1/2}\ell'(\theta_0) = \sqrt{n}(\hat\theta_n-\theta_0)\big(I(\theta_0)+o_P(1)\big)$이다. 그러므로 세 통계량은 모두 $I(\theta_0)\,n(\hat\theta_n-\theta_0)^2 + o_P(1)$이고, 서로의 차는 $\xrightarrow{P} 0$이다. $O_P\cdot o_P = o_P$ [T8]

조건 확인 ①은 $\hat\theta_n$이 가능도 방정식의 근이어야 쓸 수 있다. θ_0가 모수공간의 경계에 있으면 $\hat\theta_n$이 경계에 붙어 $\ell'(\hat\theta_n)\neq 0$일 수 있고, 극한도 $\chi^2(1)$이 아니다(6.3절의 *may*, 연습 6.3). ⑦에서 $n(\hat\theta_n-\theta_0)^2$이 확률적으로 유계라는 것은 ④에서 나온다. ■

7회 7.4절에서 $g'(\theta)=0$이면 극한이 χ^2가 된다고 했다. 로그가능도의 꼭대기에서 기울기가 0이기 때문에 가능도비가 정규가 아니라 χ^2인 것이다.

극한은 얼마나 빨리 오는가? Poisson이면 $\sum X_i \sim \mathrm{Poi}(n\lambda_0)$이므로 세 검정의 실제 크기를 시뮬레이션 없이 정확히 계산할 수 있다(그림 5.2). $\lambda_0 = 0.7$, 공칭 크기 0.05일 때 $n=10$이면 가능도비 0.057, Wald 0.087, 점수 0.034이고, $n=280$이면 0.050, 0.051, 0.049다. Wald는 분산을 $\hat\lambda$로 추정하므로, $\hat\lambda$가 작게 나온 표본에서 통계량이 부풀어 너무 자주 기각한다. 점수 검정은 H_0의 분산

λ_0을 쓰므로 그런 일이 없다.

> **증명 없이 쓰는 사실** (여러 모수의 Wilks) $\theta \in \mathbb{R}^p$이고 H_0가 r개의 독립인 제약(예: $\theta_1 = \cdots = \theta_r = 0$)이면, 정칙조건 아래 $2[\ell(\hat\theta) - \ell(\hat\theta_0)] \xrightarrow{D} \chi^2(r)$이다($\hat\theta_0$는 H_0 아래의 MLE). 자유도는 "자유로운 모수의 수의 차이"다. 8회의 Fisher의 자유도와 같은 셈이다. 다항분포에서 H_0의 점수 통계량은 정확히 Pearson의 X^2이고, 가능도비는 $G = 2\sum O \log \frac{O}{E}$다(연습 5.2).

5.2 구간으로 뒤집기

8회 정리 3.5의 쌍대성으로 세 검정은 세 구간을 준다. Wald 구간은 $\hat\theta \pm z/\sqrt{nI(\hat\theta)}$(따름정리 3.2), 가능도비 구간은 $\{\theta : 2[\ell(\hat\theta) - \ell(\theta)] \le \chi^2_{1,\alpha}\}$, 점수 구간은 $\{\theta : R(\theta) \le \chi^2_{1,\alpha}\}$이다. 비율의 점수 구간이 8회의 Wilson 구간이다. Wald 구간만 언제나 $\hat\theta$에 대해 대칭이고, 가능도비 구간은 모수를 바꿔도(예: λ와 $\log\lambda$) 같은 구간을 준다는 장점이 있다(연습 5.3).

> **예제 5.3** (기병대의 세 검정) 14개 군단 280군단–해에서 $\hat\lambda = 0.70$이다. Fisher가 쓴 10개 군단의 값 $\lambda_0 = 0.61$을 검정하면(정보 $nI(\lambda) = n/\lambda$)

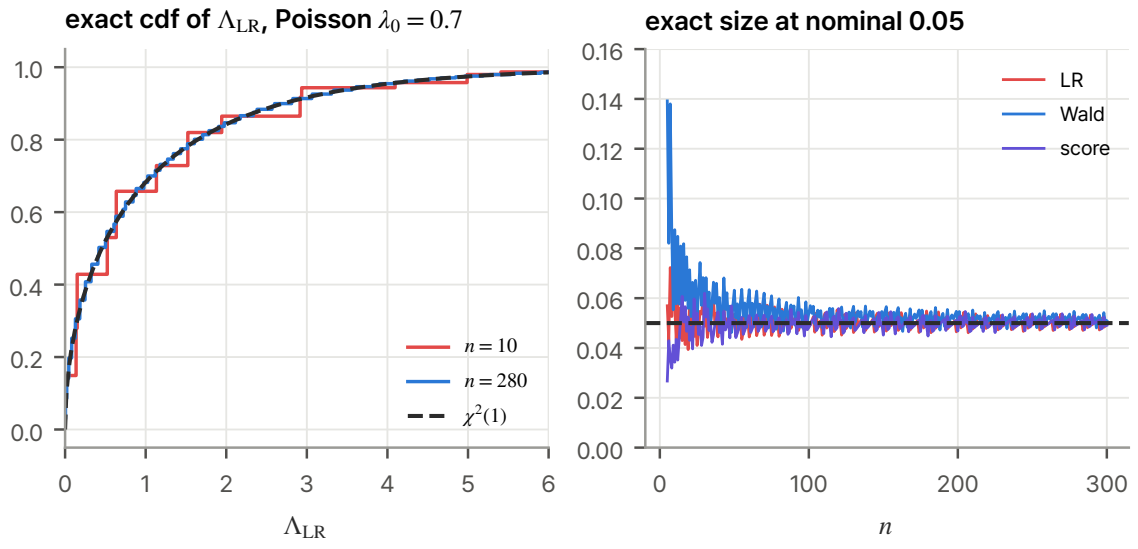


그림 5.2 왼쪽: $\lambda_0 = 0.7$인 Poisson에서 Λ_{LR}의 정확한 분포함수와 $\chi^2(1)$(점선). 오른쪽: 공칭 0.05인 세 검정의 정확한 크기. 이산분포라서 톱니 모양이고, n이 커지면 모두 0.05로 모인다.

	통계량	p	95% 구간
Wald	$1.80^2 = 3.24$	0.072	$[0.602,\ 0.798]$
가능도비	3.55	0.060	$[0.607,\ 0.803]$
점수	$1.93^2 = 3.72$	0.054	$[0.609,\ 0.805]$

$n = 280$에서도 세 p-값이 0.05 근처에서 갈린다. 경계에 가까운 결론은 검정의 선택에 민감하다. Poisson의 로그가능도는 오른쪽으로 꼬리가 길어서 Wald 구간(대칭)이 두 구간보다 왼쪽에 있다.

연습문제

연습 5.1 (Poisson의 세 통계량 ★) Poisson에서 $\Lambda_{\mathrm{LR}} = 2n\big[\bar{x}\log\frac{\bar{x}}{\lambda_0} - (\bar{x}-\lambda_0)\big]$, $W = \frac{n(\bar{x}-\lambda_0)^2}{\bar{x}}$, $R = \frac{n(\bar{x}-\lambda_0)^2}{\lambda_0}$ 임을 보이고 예제 5.3의 수치를 확인하라. 풀이 39쪽

연습 5.2 (G 통계량 ★★) 다항분포에서 $H_0: \mathbf{p} = \mathbf{p}^0$의 가능도비가 $G = 2\sum_j N_j \log\frac{N_j}{np_j^0}$ 임을 보이고, $\log(1+u) \approx u - \frac{u^2}{2}$로 $G \approx X^2$ 임을 보여라. 점수 통계량이 정확히 X^2 임도 보여라. 풀이 39쪽

연습 5.3 (가능도비 구간의 불변성 ★★) $\eta = g(\theta)$ (g 단조)로 모수를 바꾸면 가능도비 구간은 정확히 g로 옮겨지지만 Wald 구간은 그렇지 않음을 보여라. 기병대 자료에서 $\log\lambda$의 Wald 구간을 지수로 되돌려 λ의 Wald 구간과 비교하라. 풀이 39쪽

연습 5.4 (Wilks의 시뮬레이션 ★) Poi(2), $n = 50$에서 Λ_{LR}을 10만 번 시뮬레이션해 평균이 1에 가깝고 $P(\Lambda > 3.84) \approx 0.05$임을 노트북으로 확인하라. 풀이 39쪽

6

데이터로 — 기병대, 빛의 속도, 그리고 “may”

6.1 프로이센 기병대 (1875–1894)

예제 6.1 (하나의 λ인가, 군단마다 다른가) 4회 예제 6.2에서 14개 군단의 20년 합계가 Poisson보다 흩어졌다($\sum \frac{(T_c-\bar{T})^2}{\bar{T}} = 27.3$, 자유도 13). 이것을 가능도비로 다시 묻는다. H_0: 모든 군단이 같은 λ (모수 1개). H_1: 군단마다 λ_c (모수 14개). H_1의 MLE는 $\hat{\lambda}_c = T_c/20$, H_0의 MLE는 $\hat{\lambda} = 0.70$이고

$$G = 2\sum_{c}\Big[T_c \log \frac{T_c}{20\hat{\lambda}} - (T_c - 20\hat{\lambda})\Big] = 26.1,$$

자유도 $14-1=13$, $p = 0.016$이다. 점수 검정에 해당하는 Pearson의 27.3 ($p = 0.011$)과 거의 같다. 군단 사이에 실제로 차이가 있다는 증거이고, Fisher가 일부 군단을 뺀 이유를 뒷받침한다(4회 6절).

6.2 Michelson의 빛의 속도 (1879)

예제 6.2 (정밀하지만 정확하지 않다) Albert Michelson은 1879년 회전 거울로 빛의 속도를 100번 쟀다(`HistData::Michelson`, 단위는 km/s에서 299,000을 뺀 값). 정규모형의 MLE는 $\hat{\mu} = \bar{x} = 852.4$, $\hat{\sigma} = 78.6$이고(n으로 나눔), μ의 95% 가능도비 구간은 $[836.8,\ 868.0]$이다(t-구간 $[836.7,\ 868.1]$과 거의 같다). 오늘의 값(공기 중으로 환산)은 734.5로 구간에서 $\frac{852.4-734.5}{7.9} = 14.9$ 표준오차나 떨어져 있다. 측정 사이의 흩어짐은 작았지만, 모든 측정에 같은 방향의 오차(장치나 보정의 계통 오차)가 있었다. 가능도와 표준오차는 **주어진 모형 안에서의** 불확실성만 잰다. 모형이 “측정 = 참값 + 평균 0인 오차”를 가정하는 한, 평균이 0이 아닌 오차는 보이지 않는다.

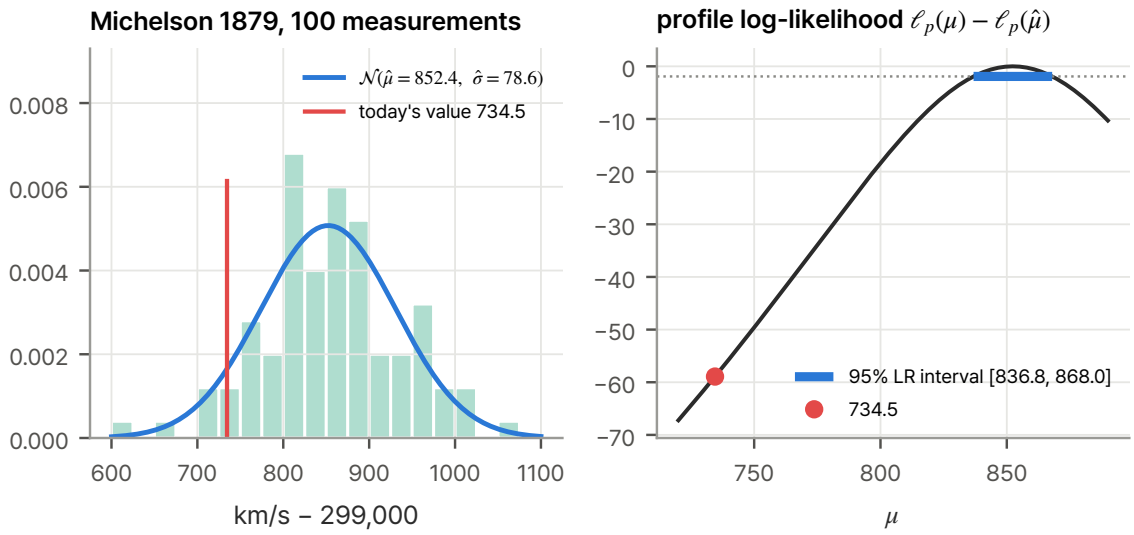


그림 6.1 왼쪽: Michelson의 100번의 측정과 정규 MLE 적합. 빨간 선은 오늘의 값 734.5. 오른쪽: μ의 로그가능도(정규, σ를 MLE로 프로파일)와 가능도비 95% 구간(가로선). 구간은 좁지만 참값을 한참 비켜간다.

6.3 The Federalist의 "may" — 음이항의 MLE

예제 6.3 (적률법과 최대가능도, 그리고 경계) 4회 예제 6.3에서 Madison의 글 262조각의 *may* 개수에 음이항을 적률법으로 맞췄다($\hat r = 1.23$, $\hat p = 0.651$). 최대가능도로 맞추면 $\hat r = 1.19$, $\hat p = 0.644$이다(두 모수를 수치로 최대화; 평균은 두 방법 모두 $\bar x = 0.656$으로 같다). 음이항은 $r \to \infty$에서 Poisson이 되므로 Poisson 대 음이항의 가능도비 $2[\ell_{\text{NB}} - \ell_{\text{Poi}}] = 23.3$으로 과산포를 검정할 수 있다. 다만 귀무가설이 모수공간의 **경계**($1/r = 0$)에 있어 Wilks의 정칙조건(R2)이 깨지고, 극한분포는 $\chi^2(1)$이 아니라 $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$이다(Self–Liang, 1987). 어느 쪽으로 읽든 23.3은 압도적이다.

연습문제

연습 6.1 (군단의 가능도비 ★) 예제 6.1의 G를 유도하라. H_1의 로그가능도에서 H_0의 로그가능도를 빼면 $\sum T_c \log \frac{\hat\lambda_c}{\hat\lambda}$만 남는 이유는? 풀이 39쪽

연습 6.2 (Michelson의 t-모형 ★★) Michelson 자료에 $t(\nu = 4)$ 위치–척도 모형을 MLE로 맞추면 위치의 추정값이 정규의 평균과 얼마나 다른가(노트북)? 6회 8.4절의 가중치 해석으로 설명하라. 그래도 참값과의 차이는 왜 해결되지 않는가? 풀이 40쪽

연습 6.3 (경계에서의 가능도비 ★★★) $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, $H_0 : \mu = 0$ 대 $H_1 : \mu \geq 0$이면 $\hat{\mu} = \max(\bar{X}, 0)$이고 $2[\ell(\hat{\mu}) - \ell(0)] = n\bar{X}^2\, 1\{\bar{X} > 0\}$임을 보여라. 그 분포가 $\frac{1}{2}$의 확률로 0, $\frac{1}{2}$의 확률로 $\chi^2(1)$임을 결론지어라.

풀이 40쪽

7

부쉬보기 — 초효율과 너무 많은 모수

7.1 Hodges의 초효율 추정량 (1951)

Cramér–Rao 한계는 **불편**추정량의 것이고, MLE의 점근 효율은 "어떤 추정량도 점근분산이 $1/I$보다 작을 수 없다" 처럼 들린다. Joseph Hodges는 1951년 그렇지 않다는 예를 만들었다(Le Cam 1953이 발표했다).

예제 7.1 (한 점에서만 이기는 추정량) $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$이고

$$T_n = \begin{cases} \bar{X}_n, & |\bar{X}_n| \geq n^{-1/4}, \\ 0, & |\bar{X}_n| < n^{-1/4} \end{cases}$$

라 하자. $\theta \neq 0$이면 $|\bar{X}_n| \to |\theta| > 0$이므로 결국 $T_n = \bar{X}_n$이고 $\sqrt{n}(T_n - \theta) \xrightarrow{D} \mathcal{N}(0,1)$, 곧 MLE와 같다. 그런데 $\theta = 0$이면 $P(|\bar{X}_n| < n^{-1/4}) = P(|Z| < n^{1/4}) \to 1$이므로 $T_n = 0$이 되어 $\sqrt{n}\,T_n \xrightarrow{P} 0$, 점근분산이 0이다. 한계 $1/I = 1$보다 작다.

대가는 어디에 있는가? θ가 0에 가깝지만 0이 아니면, 예를 들어 $\theta = n^{-1/4}$ 근처이면 T_n은 자주 0으로 잘못 끌려가 큰 오차를 낸다. 평균제곱오차를 n배 한 위험 $nE_\theta(T_n - \theta)^2$의 최댓값은 $n = 100$에서 5.6, $n = 10{,}000$에서 69로, n과 함께 끝없이 커진다(그림 7.1). 한 점에서 이긴 만큼 그 근처에서 크게 진다. Le Cam(1953)은 이런 초효율이 일어나는 점의 집합이 측도 0임을 보였고, Hájek(1970)의 수렴 정리와 국소 점근 최소최대 정리는 "근처의 모든 θ를 함께 볼 때" MLE가 최선임을 정확히 말한다. 수축 추정(16회)은 이 생각을 다른 방향에서 다시 만난다.

7.2 받침이 모수에 의존하면

$\mathrm{U}(0, \theta)$의 MLE $X_{(n)}$은 (R1)을 깬다. 그 결과는 나쁘지 않고 오히려 좋다. $n(\theta - X_{(n)}) \xrightarrow{D}$ Exp(평균 θ)로(7회 예제 2.2(b)), 오차가 $1/\sqrt{n}$이 아니라 $1/n$의 속도로 준다. 정규 극한도, Cramér–Rao 한계도, Wilks의 χ^2도 이런 "비정칙" 모형에는 맞지 않는다. 8회의 독일 전차 문제와 15회의 부트스트랩의 실패가 같은 구조다.

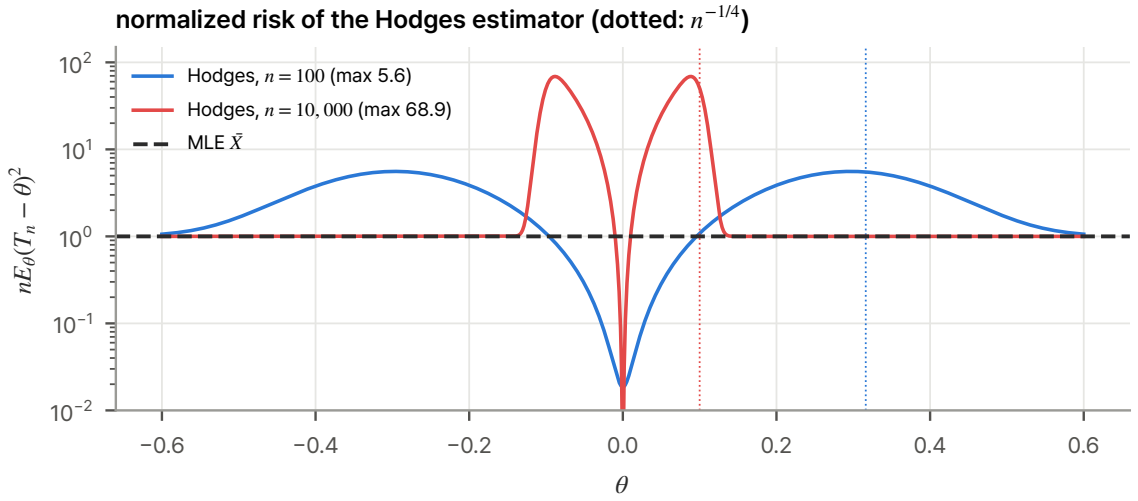


그림 7.1 Hodges 추정량의 정규화한 위험 $nE_\theta(T_n-\theta)^2$ ($n=100$, 10,000). 점선은 MLE의 위험 1.0에서는 MLE를 이기지만 $|\theta| \approx n^{-1/4}$ 근처에서 크게 지고, 그 봉우리는 n이 클수록 높아진다.

7.3 모수가 자료와 함께 늘어나면 — Neyman–Scott (1948)

예제 7.2 (일치하지 않는 MLE) n명에게서 두 번씩 측정한다: $X_{i1}, X_{i2} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \dots, n$. 사람마다 평균 μ_i가 다르고 측정 오차 σ^2는 같다. 모수는 $n+1$개이고, MLE는 $\hat{\mu}_i = \bar{X}_i$,

$$\hat{\sigma}^2 = \frac{1}{2n}\sum_{i}\sum_{j=1}^{2}(X_{ij}-\bar{X}_i)^2 = \frac{1}{4n}\sum_{i}(X_{i1}-X_{i2})^2 \xrightarrow{P} \frac{\sigma^2}{2}.$$

$n \to \infty$에서 MLE가 참값의 절반으로 간다. 각 μ_i가 자료 두 개로만 추정되어 정보가 쌓이지 않기 때문이다. 정리 3.1은 모수의 수가 고정되어 있다고 가정했다. 10회는 각 μ_i를 보이지 않는 확률변수로 바꿔 EM으로 이 문제("귀찮은 모수"가 많은 경우)를 풀고, 16회의 계층 모형이 같은 생각을 Bayes의 언어로 다시 쓴다.

연습문제

연습 7.1 (Hodges의 위험 ★★) $\theta = 0$에서 $nE_0T_n^2 = E\big[Z^2; |Z| \ge n^{1/4}\big]$ ($Z \sim \mathcal{N}(0,1)$)임을 보이고 이것이 0으로 감을 보여라. $\theta_n = c\,n^{-1/4}$ ($0 < c < 1$)이면 $nE(T_n-\theta_n)^2 \to \infty$임을 보여라.

풀이 40쪽

연습 7.2 (균등분포의 MLE의 속도 ★) $P\big(n(\theta - X_{(n)}) > x\big) = (1 - \frac{x}{n\theta})^n \to e^{-x/\theta}$를 확인하고, MLE를 $\frac{n+1}{n}X_{(n)}$으로 고치면 치우침이 사라짐을 보여라.

풀이 40쪽

연습 7.3 (Neyman–Scott의 고침 ★★) 예제 7.2에서 $D_i = X_{i1} - X_{i2} \sim \mathcal{N}(0, 2\sigma^2)$만 쓰면 μ_i가 사라진다. D_i들의 가능도로 만든 MLE $\frac{1}{2n}\sum D_i^2$가 일치함을 보여라. (조건부 또는 주변 가능도의 생각이다.) 풀이 40쪽

8

Beautiful insights — 손실함수 속의 가능도

8.1 교차엔트로피는 음의 로그가능도다

분류 신경망은 입력 x마다 범주 $k = 1, \dots, K$의 확률 $q_k(x;\theta)$(softmax)를 내놓고, 정답 y에 대해 손실 $-\log q_y(x;\theta)$를 최소화한다. 이것을 **교차엔트로피**라 부르지만, 학습 자료 전체의 합

$$-\sum_{i=1}^{n} \log q_{y_i}(x_i;\theta) = -\ell(\theta)$$

은 범주형(다항) 모형의 음의 로그가능도 그대로다. 교차엔트로피를 최소화하는 것은 최대가능도 추정이다. 회귀의 제곱오차 손실은 정규 오차의 음의 로그가능도이고(분산 고정), 절댓값 손실은 Laplace 오차의 그것이다(1절 예제 1.2(e)). 손실함수를 고르는 것은 오차의 분포를 고르는 것이다.

자료의 경험분포 $\hat{P}_n$에 대해 $-\frac{1}{n}\ell(\theta) = E_{\hat{P}_n}[-\log f(X;\theta)]$이고, 이것은 $D(\hat{P}_n \| f_\theta)$와 θ에 의존하지 않는 항(경험분포의 엔트로피)의 합이다. 그래서 MLE는 경험분포에서 KL 발산으로 가장 가까운 모형을 고른다(보조정리 1.4). 모형이 틀렸을 때도 MLE는 "참 분포에 KL로 가장 가까운 모형"으로 간다(White, 1982). Michelson의 예(6절)처럼 그것이 참값이라는 보장은 없다.

8.2 Fisher 정보는 곡률이고, 곡률은 거리다

두 가까운 모수의 KL 발산을 전개하면

$$D(\theta \,\|\, \theta + \delta) = \tfrac{1}{2} I(\theta)\, \delta^2 + O(\delta^3)$$

이다(연습 8.1). Fisher 정보는 모수공간에서 "분포가 얼마나 달라지는가"를 재는 거리(계량)다. 여러 모수이면 $\tfrac{1}{2}\delta' I(\theta)\delta$이다.

이 사실이 두 가지 기계학습 방법의 바탕이다.

- **자연기울기** (Amari 1998). 보통의 경사하강은 모수공간의 유클리드 거리로 "가장 가파른" 방향을 찾는다. 같은 분포를 다른 방식으로 모수화하면 방향이 바뀐다. 자연기울기 $I(\theta)^{-1}\nabla\ell$은 분포 사이의 거리(KL)로 가장 가파른 방향이어서 모수화에 무관하다. 4절의 Fisher 채점이 바로 음의 로그가능도의 자연기울기 방법이다. K-FAC(Martens–Grosse 2015) 같은 신경망 최적화기가 이 행렬을 근사한다.

- **탄성 가중치 고정** (EWC; Kirkpatrick 외 2017). 과제 A를 배운 신경망이 과제 B를 배우며 A를 잊지 않게 하려면, A에 중요한 가중치를 덜 움직이게 해야 한다. "중요함"의 척도가 A의 자료로 잰 Fisher 정보의 대각성분 F_j이고, 손실에 $\frac{\lambda}{2}\sum_j F_j(\theta_j-\theta^*_{A,j})^2$을 더한다. 정리 3.1이 말하는 MLE 근처의 정규 근사(16회의 Laplace 근사)를 과제 A의 사후분포로 쓰는 것이다.

8.3 온도 하나의 최대가능도 — 신경망의 보정

현대의 깊은 분류기는 정확하지만 확신이 지나치다. 예측 확률 0.99인 사례의 실제 정답률이 그보다 낮다(Guo 외 2017). 처방은 놀랍게 단순하다. 출력 직전의 점수(logit) $\mathbf{z}$를 하나의 온도 $T>0$로 나눠 $\mathrm{softmax}(\mathbf{z}/T)$로 쓰고, T를 **검증 자료**의 음의 로그가능도를 최소화하도록, 곧 MLE로 고른다. 모수 하나의 MLE이므로 4절의 Newton 몇 걸음이면 된다. 다만 $a=1/T$로 쓰면 로그가능도가 오목하지만 Newton의 첫걸음이 넘칠 수 있으므로, 로그가능도가 늘 때까지 걸음을 반으로 줄이는 안전장치를 단다(노트북). 정확도는 그대로이고(순위가 바뀌지 않는다) 확률의 보정만 좋아진다. 8회 8.1절의 PIT 진단으로 그 효과를 확인할 수 있다.

연습문제

연습 8.1 (KL과 Fisher 정보 ★★) $D(\theta\|\theta+\delta)=E_\theta[\log f(X;\theta)-\log f(X;\theta+\delta)]$를 δ에 대해 2차까지 전개해 $\frac{1}{2}I(\theta)\delta^2$임을 보여라(정리 2.2). Poisson이면 $D(\lambda\|\lambda')=\lambda\log\frac{\lambda}{\lambda'}-\lambda+\lambda'$이다. 이것을 직접 전개해 확인하라. 풀이 40쪽

연습 8.2 (제곱오차와 정규 ★) $Y_i=m(x_i;\theta)+\varepsilon_i$, $\varepsilon_i\overset{\text{iid}}{\sim}\mathcal{N}(0,\sigma^2)$이면 θ의 MLE가 제곱오차의 합을 최소화하는 값과 같음을 보여라. σ^2의 MLE는? 풀이 41쪽

연습 8.3 (온도의 가능도 방정식 ★★) logit $\mathbf{z}_i$와 정답 y_i에 대해 $\ell(T)=\sum_i\big[z_{i,y_i}/T-\log\sum_k e^{z_{ik}/T}\big]$이다. $\ell'(T)=0$이 "정답의 logit의 평균 = 모형이 예측하는 logit의 기댓값의 평균"으로 쓰임을 보여라. 풀이 41쪽

9

정리와 역사 노트

9.1 한 장 요약

정리	내용	조건
불변성 (1.3)	$g(\hat\theta)$는 $g(\theta)$의 MLE	—
일치성 (1.4–1.6)	기대 로그가능도는 참값에서 최대(Gibbs), 가능도 방정식의 일치하는 근	식별가능
정보 (2.2)	$Es = 0, I = \operatorname{Var} s = -E\,\partial^2 \log f$, 표본은 nI	정칙조건
Cramér–Rao (2.4)	$\operatorname{Var} T \geq k'(\theta)^2/(nI)$	정칙조건
점근정규성 (3.1, 3.2)	$\sqrt{n}(\hat\theta - \theta_0) \xrightarrow{D} \mathcal{N}(0, 1/I)$, 구간 $\hat\theta \pm z/\sqrt{nI(\hat\theta)}$	정칙조건
한 걸음 (4.2)	$\sqrt{n}$-일치 시작점에서 채점 한 번이면 효율적	—
Wilks (5.2)	$2[\ell(\hat\theta) - \ell(\theta_0)] \xrightarrow{D} \chi^2(1)$, Wald · 점수와 점근 동치	정칙조건, 내점

이번 회차의 모든 점근 결과는 하나의 계산, 곧 **점수의 테일러 전개**에서 나왔다. 분자는 중심극한정리로 정규가 되고, 분모(곡률)는 대수의 법칙으로 정보가 되고, Slutsky가 둘을 합친다. 7절은 그 계산의 가정(정칙조건, 고정된 모수의 수, 한 점이 아니라 근방에서의 비교)을 하나씩 깨 보았다.

9.2 Stigler의 명명 법칙

> **Stigler's law of eponymy**
>
> - **Cramér–Rao 부등식**: Rao(1945)와 Cramér(1946)보다 먼저 Fréchet(1943)가 증명했고, Aitken과 Silverstone(1942), Darmois(1945)도 같은 결과를 냈다. 그래서 프랑스 문헌에서는 Fréchet–Darmois–Cramér–Rao 부등식이라 부르기도 한다.
> - **최대가능도**: Fisher(1912, 1922)의 이름으로 기억되지만 원리 자체는 Lambert(1760), Daniel Bernoulli(1778), Gauss(1809)에게 있었다. Fisher의 공헌은 원리가 아니라 그 원리가 **왜** 좋은지(일치성, 효율, 정보)를 말한 것이다.
> - **Fisher 정보**: Edgeworth(1908)가 같은 양으로 추정량의 오차를 계산했다.

9.3 7 Pillars 지도

이번 회차는 셋째 기둥 **Likelihood**의 중심이다. 8회가 Arbuthnot의 "가설 아래에서의 확률"로 이 기둥을 시작했다면, Fisher는 같은 확률을 모수의 함수로 읽어 추정과 검정을 하나의 이론으로 묶었다. 둘째 기둥 **Information**도 다시 나왔다. 3회의 $\sqrt{n}$ 규칙이 Fisher 정보 nI라는 정확한 형태를 얻었다. 정보는 관측 수에 비례해 쌓이고, 오차는 그 제곱근으로 준다.

9.4 다음 회 — 여러 모수와 숨은 변수: EM (1894 → 1977)

이번 회차의 예들은 가능도를 직접 최대화할 수 있었다. 그런데 자료의 일부가 보이지 않으면 (혼합분포에서 각 관측이 어느 성분에서 왔는지, 4회의 Pearson의 게) 가능도가 합의 로그가 되어 다루기 어렵다. 10회는 Pearson(1894)이 적률법으로 손으로 풀었던 혼합분포에서 출발해, 보이지 않는 것을 기댓값으로 채우고 최대화하기를 되풀이하는 EM 알고리즘(Dempster–Laird–Rubin 1977)을 증명하고, 7절의 Neyman–Scott 문제를 다시 본다.

참고문헌

- Bernoulli, D. Diiudicatio maxime probabilis plurium observationum discrepantium (1778). Kendall, M. G. 영역과 Euler의 논평, *Biometrika* 48 (1961), 1–18.
- Edgeworth, F. Y. On the probable errors of frequency-constants. *JRSS* 71 (1908).
- Fisher, R. A. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* 41 (1912), 155–160. · On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. A* 222 (1922), 309–368.
- Rao, C. R. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* 37 (1945), 81–91. · Large sample tests of statistical hypotheses concerning several parameters. *Proc. Camb. Phil. Soc.* 44 (1948), 50–57.
- Cramér, H. *Mathematical Methods of Statistics*. Princeton, 1946.
- Neyman, J., Pearson, E. S. On the use and interpretation of certain test criteria. *Biometrika* 20A (1928). · Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* 9 (1938), 60–62. · Wald, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. AMS* 54 (1943), 426–482.
- Wald, A. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* 20 (1949), 595–601.
- Le Cam, L. On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.* 1 (1953), 277–330. · Hájek, J. A characterization of limiting distributions of regular estimates. *Z. Wahrsch.* 14 (1970), 323–330.
- Neyman, J., Scott, E. L. Consistent estimates based on partially consistent observations. *Econometrica*

16 (1948), 1–32.

- Barnett, V. D. Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika* 53 (1966), 151–165.
- Self, S. G., Liang, K.-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *JASA* 82 (1987), 605–610. · White, H. Maximum likelihood estimation of misspecified models. *Econometrica* 50 (1982), 1–25.
- Amari, S. Natural gradient works efficiently in learning. *Neural Computation* 10 (1998), 251–276. · Martens, J., Grosse, R. Optimizing neural networks with Kronecker-factored approximate curvature. *ICML* (2015).
- Kirkpatrick, J. 외. Overcoming catastrophic forgetting in neural networks. *PNAS* 114 (2017), 3521–3526. · Guo, C. 외. On calibration of modern neural networks. *ICML* (2017).
- Michelson, A. A. Experimental determination of the velocity of light. *Astronomical Papers of the American Ephemeris* 1 (1880), 109–145. · 자료: `HistData`의 `Michelson`; `vcd`의 `VonBort`, `Federalist`.
- Stigler, S. M. *The Seven Pillars of Statistical Wisdom*. Harvard, 2016.

A

연습문제 풀이

각 풀이의 머리에 있는 문제 ○○쪽을 누르면 문제로 돌아간다. 계산 문제의 수치는 노트북에서 같은 계산으로 확인했다. 유도형 연습의 풀이는 본문에서 연습으로 돌린 정리의 완성된 증명이다.

1절

연습 1.1 문제 7쪽

$\ell(\theta) = -n\log\theta - \frac{1}{\theta}\sum x_i$, $\ell'(\theta) = -\frac{n}{\theta} + \frac{\sum x_i}{\theta^2} = 0$에서 $\hat\theta = \bar x$. 불변성으로 $\hat\lambda = 1/\bar X$. $\sum X_i \sim \Gamma(n, 1/\lambda)$이므로 $E\frac{1}{\sum X_i} = \frac{\lambda}{n-1}$(4회 연습 3.4)이고 $E\frac{1}{\bar X} = \frac{n\lambda}{n-1}$.

연습 1.2 문제 7쪽

$L(\mu,\beta) = \beta^{-n}\exp\!\big(-\frac{1}{\beta}\sum(x_i-\mu)\big)1\{\mu \le x_{(1)}\}$. β를 고정하면 L은 μ에 대해 증가하므로 $\hat\mu = x_{(1)}$. 이것을 넣고 β로 미분하면 $\hat\beta = \frac{1}{n}\sum(x_i - x_{(1)}) = \bar x - x_{(1)}$.

연습 1.3 문제 7쪽

$\ell = \sum_j N_j \log p_j$를 $\sum p_j = 1$ 아래에서 최대화한다. $\frac{\partial}{\partial p_j}[\ell - \lambda(\sum p_j - 1)] = \frac{N_j}{p_j} - \lambda = 0$에서 $p_j = N_j/\lambda$이고 합이 1이므로 $\lambda = n$. 독립 모형 $p_{ij} = p_{i\cdot}p_{\cdot j}$이면 $\ell = \sum_i N_{i\cdot}\log p_{i\cdot} + \sum_j N_{\cdot j}\log p_{\cdot j}$로 행과 열이 따로 떨어지므로 각각 다항의 MLE $N_{i\cdot}/n$, $N_{\cdot j}/n$이다.

연습 1.4 문제 7쪽

$P(X=0) = e^{-\lambda}$의 MLE는 불변성으로 $e^{-0.70} = 0.497$. 관측 비율은 $144/280 = 0.514$로 기대도수 139와 관측 144의 차이다.

연습 1.5 문제 7쪽

$\theta \ge \theta_0$이면 $D(\theta_0\|\theta) = \int_0^{\theta_0}\frac{1}{\theta_0}\log\frac{1/\theta_0}{1/\theta}dx = \log\frac{\theta}{\theta_0}$. $\theta < \theta_0$이면 (θ,θ_0)에서 $f(x;\theta) = 0$이지만 $f(x;\theta_0) > 0$이므로 $\log\frac{f(x;\theta_0)}{f(x;\theta)} = \infty$인 곳에 양의 확률이 있어 $D = \infty$. 자료에서도 같다. $\theta < x_{(n)}$이면 관측 $x_{(n)}$의 밀도가 0이어서 $L(\theta) = 0$이다.

연습 1.6 문제 7쪽

(a) $g(\hat\theta)=\hat\eta$이므로 $\hat\theta$는 $\{\theta : g(\theta)=\hat\eta\}$의 원소다. 상한은 그 집합의 모든 원소에서의 값 이상이므로 $L^*(\hat\eta)=\sup_{g(\theta)=\hat\eta}L(\theta)\ge L(\hat\theta)$이다. 상한이 "상계" 라는 성질만 썼다. (b) $\{\theta : g(\theta)=\eta\}$는 모수공간의 부분집합이므로 그 위의 상한은 전체 위의 상한 $\sup_\theta L(\theta)=L(\hat\theta)$를 넘지 않는다. (c) (b)를 $\eta=\hat\eta$에 쓰고 (a)와 합치면 $L^*(\hat\eta)=L(\hat\theta)$이고, (b)로 이것은 모든 $L^*(\eta)$ 이상이다. 따라서 $\hat\eta$는 L^*의 최대점이다. (d) g가 일대일이면 $\{\theta : g(\theta)=\eta\}=\{g^{-1}(\eta)\}$이므로 상한은 그 한 점의 값이다. $g(\theta)=\theta^2$이면 $\eta>0$에 $\theta=\pm\sqrt{\eta}$ 두 값이 대응하므로 "η의 가능도" 를 하나로 정하려면 둘 중 큰 값을 골라야 한다. 상한이 그 선택이다.

연습 1.7 문제 7쪽

(a) log는 증가하므로 $L(\theta_0)>L(\theta)\iff \ell(\theta)-\ell(\theta_0)<0\iff \frac{1}{n}\sum_i Y_i<0$이다. (b) Y_i는 iid인 X_i의 같은 함수이므로 iid이다. $EY_i=E_{\theta_0}\log f(X;\theta)-E_{\theta_0}\log f(X;\theta_0)=-D(\theta_0\|\theta)$이고, 보조정리 1.4로 $D>0$이다. Khinchin의 대수의 법칙은 평균의 존재만 요구하므로 $\bar Y_n\xrightarrow{P}-D$이다. (c) $\bar Y_n\ge 0$이면 $\bar Y_n+D\ge D$이므로 $\{\bar Y_n\ge 0\}\subset\{|\bar Y_n+D|\ge D\}$이다. 오른쪽의 확률은 확률수렴의 정의로 0으로 가므로 $P(L(\theta_0)>L(\theta))=1-P(\bar Y_n\ge 0)\to 1$이다. (d) Chebyshev 부등식으로 $P(|\bar Y_n+D|\ge D)\le \frac{\operatorname{Var}\bar Y_n}{D^2}=\frac{\sigma^2}{nD^2}$이다. 틀린 값이 이길 확률은 적어도 $1/n$의 속도로 줄고, 그 상수는 KL 발산의 제곱에 반비례한다. 참값과 가까운 θ일수록 D가 작아 더 많은 자료가 필요하다.

2절

연습 2.1 문제 10쪽

Poisson: $\operatorname{Var}(\frac{X}{\lambda}-1)=\frac{\lambda}{\lambda^2}$, $-E(-\frac{X}{\lambda^2})=\frac{1}{\lambda}$. 정규 분산: $\operatorname{Var}s=\frac{\operatorname{Var}(X-\mu)^2}{4\sigma^8}=\frac{2\sigma^4}{4\sigma^8}$, $-E\big[\frac{1}{2\sigma^4}-\frac{(X-\mu)^2}{\sigma^6}\big]=\frac{1}{2\sigma^4}$. Cauchy: $Es^2=\frac{4}{\pi}\int\frac{u^2}{(1+u^2)^3}du=\frac{4}{\pi}\cdot\frac{\pi}{8}=\frac{1}{2}$이고, $-E\frac{\partial^2}{\partial\theta^2}\log f=E\frac{2(1-u^2)}{(1+u^2)^2}=\frac{2}{\pi}\big(\int\frac{du}{(1+u^2)^3}-\int\frac{u^2du}{(1+u^2)^3}\big)=\frac{2}{\pi}\big(\frac{3\pi}{8}-\frac{\pi}{8}\big)=\frac{1}{2}$.

연습 2.2 문제 10쪽

연쇄법칙으로 $\frac{\partial}{\partial\eta}\log f=s(\theta)\frac{d\theta}{d\eta}=\frac{s(\theta)}{g'(\theta)}$이므로 $I^*(\eta)=\frac{I(\theta)}{g'(\theta)^2}$. 로그 오즈이면 $g'(p)=\frac{1}{p(1-p)}$이므로 $I^*=\frac{1}{p(1-p)}\cdot p^2(1-p)^2=p(1-p)$, $\hat\eta$의 점근분산 $\frac{1}{np(1-p)}$는 7회 예제 4.5와 같다.

연습 2.3 문제 11쪽

$S=\sum X_i\sim\Gamma(n,1/\lambda)$에서 $E\frac{1}{S}=\frac{\lambda}{n-1}$, $E\frac{1}{S^2}=\frac{\lambda^2}{(n-1)(n-2)}$이므로 $\operatorname{Var}\frac{n-1}{S}=(n-1)^2\big[\frac{\lambda^2}{(n-1)(n-2)}-$

$\frac{\lambda^2}{(n-1)^2}\big] = \frac{\lambda^2}{n-2}$. 비율의 정보는 $I = \frac{1}{\lambda^2}$ 이므로 한계는 $\frac{\lambda^2}{n}$ 이고 효율은 $\frac{n-2}{n}$ 이다.

연습 2.4 문제 11쪽

$EX_{(n)} = \frac{n\theta}{n+1}$, $\operatorname{Var} X_{(n)} = \frac{n\theta^2}{(n+1)^2(n+2)}$ 이므로 $\operatorname{Var}\frac{n+1}{n}X_{(n)} = \frac{\theta^2}{n(n+2)}$. 증명이 무너지는 곳은 $\int_0^\theta f(x;\theta)dx = 1$을 미분하는 단계다. 적분의 위끝이 θ여서 경계항 $f(\theta;\theta) = \frac{1}{\theta}$가 생기고, 점수의 평균이 $-\frac{1}{\theta} \neq 0$이다.

3절

연습 3.1 문제 14쪽

$0.70 \pm 1.96\sqrt{0.70/280} = [0.602,\ 0.798]$. $e^{-\lambda}$는 단조이므로 끝점을 옮기면 $[e^{-0.798}, e^{-0.602}] = [0.450,\ 0.548]$. 델타 방법으로는 $0.497 \pm 1.96 \cdot 0.497 \cdot 0.050 = [0.448,\ 0.545]$로 거의 같다.

연습 3.2 문제 14쪽

$|u_i| > 1$인 관측은 음의 항을 준다. θ가 관측들에서 멀리 떨어져 대부분의 $|u_i|$가 1보다 크면 합이 음수가 된다. 예제 4.3의 골짜기 $\theta = 7.53$에서는 $J = -0.48$이다. 한편 $\hat\theta$가 극대점이면 2계 필요조건으로 $\ell''(\hat\theta) \le 0$, 곧 $J(\hat\theta) \ge 0$이다. 그래서 관측 정보 $J(\hat\theta)$로 만든 구간(따름정리 3.2(a))은 엄격한 극대점($J(\hat\theta) > 0$)이면 언제나 정의된다. 기댓값은 한 관측에 $E\frac{2(1-U^2)}{(1+U^2)^2} = \frac{1}{2}$ (연습 2.1의 계산).

연습 3.3 문제 14쪽

$\log f = -\log\Gamma(\alpha) - \alpha\log\beta + (\alpha-1)\log x - \frac{x}{\beta}$에서 $\frac{\partial^2}{\partial\alpha^2} = -\psi'(\alpha)$, $\frac{\partial^2}{\partial\alpha\partial\beta} = -\frac{1}{\beta}$, $\frac{\partial^2}{\partial\beta^2} = \frac{\alpha}{\beta^2} - \frac{2x}{\beta^3}$이고 $EX = \alpha\beta$이므로 $I = \left(\begin{smallmatrix}\psi'(\alpha) & 1/\beta \\ 1/\beta & \alpha/\beta^2\end{smallmatrix}\right)$. $(I^{-1})_{11} = \frac{\alpha/\beta^2}{\psi'(\alpha)\alpha/\beta^2 - 1/\beta^2} = \frac{\alpha}{\alpha\psi'(\alpha)-1}$. $\alpha = 2$이면 $\psi'(2) = 0.6449$이므로 $\frac{2}{0.2899} = 6.90$.

연습 3.4 문제 14쪽

$\frac{\partial^2 \log f}{\partial\mu\partial\sigma^2} = -\frac{X-\mu}{\sigma^4}$의 기댓값이 0이므로 비대각 성분이 0이다. 극한에서 $\hat\mu$와 $\hat\sigma^2$은 무상관인 결합정규, 곧 독립이다(5회 정리 B). 정규 표본에서는 이것이 극한이 아니라 정확히 성립한다는 것이 6회의 Student 정리다.

연습 3.5 문제 14쪽

(a) $g(t) = \sqrt{I(t)}$는 θ_0에서 연속이고 $\hat\theta_n \xrightarrow{P} \theta_0$이므로 연속사상 정리(7회)로 $\sqrt{I(\hat\theta_n)} \xrightarrow{P} \sqrt{I(\theta_0)}$

이다. 정리 3.1로 $\sqrt{nI(\theta_0)}(\hat\theta_n-\theta_0)\xrightarrow{D}\mathcal{N}(0,1)$이고, Slutsky 정리로 여기에 $\sqrt{I(\hat\theta_n)/I(\theta_0)}\xrightarrow{P}1$을 곱해도 극한은 같다. (b) 구간이 θ_0를 포함하는 사건은 $\{|\sqrt{nI(\hat\theta_n)}(\hat\theta_n-\theta_0)|\le z_{\alpha/2}\}$이다. (a)와 극한분포의 연속성으로 그 확률은 $P(|Z|\le z_{\alpha/2})=1-\alpha$로 간다(8회 정리 2.4). (c) 평균값 정리로 $\ell''(\hat\theta_n)=\ell''(\theta_0)+(\hat\theta_n-\theta_0)\ell'''(\tilde\theta_n)$이고 $\tilde\theta_n$은 $\hat\theta_n$과 θ_0 사이에 있다. 정리 3.1의 증명 ④로 $-n^{-1}\ell''(\theta_0)\xrightarrow{P}I(\theta_0)$이고, ⑤와 같이 $|n^{-1}\ell'''(\tilde\theta_n)|\le n^{-1}\sum M(X_i)$가 확률적으로 유계이며 $\hat\theta_n-\theta_0\xrightarrow{P}0$이므로 둘째 항은 $\xrightarrow{P}0$이다. (d) 델타 방법(7회 정리 4.4)을 정리 3.1의 극한 $\mathcal{N}(0,1/I(\theta_0))$에 쓰면 $\sqrt{n}(g(\hat\theta_n)-g(\theta_0))\xrightarrow{D}\mathcal{N}(0,g'(\theta_0)^2/I(\theta_0))$이다. 정리 1.3로 $g(\hat\theta_n)$은 $g(\theta)$의 MLE이다.

4절

연습 4.1 문제 18쪽

$\ell'(\theta)=n(\frac{\bar x}{\theta}-1)$, $\ell''(\theta)=-\frac{n\bar x}{\theta^2}$이므로 $\theta^{(1)}=\theta^{(0)}+(\frac{\bar x}{\theta^{(0)}}-1)\frac{(\theta^{(0)})^2}{\bar x}=2\theta^{(0)}-\frac{(\theta^{(0)})^2}{\bar x}$. $\theta^{(0)}>2\bar x$이면 음수다. Fisher 채점은 $nI=n/\theta$이므로 $\theta^{(1)}=\theta^{(0)}+(\frac{\bar x}{\theta^{(0)}}-1)\theta^{(0)}=\bar x$.

연습 4.2 문제 18쪽

$e=e_t$로 쓰면 $\ell'(\theta_t)=\ell''(\hat\theta)e+\frac12\ell'''(\hat\theta)e^2+O(e^3)$, $\ell''(\theta_t)=\ell''(\hat\theta)+\ell'''(\hat\theta)e+O(e^2)$이므로 $\frac{\ell'(\theta_t)}{\ell''(\theta_t)}=e+\frac{\ell'''}{2\ell''}e^2-\frac{\ell'''}{\ell''}e^2+O(e^3)$. $e_{t+1}=e_t-\frac{\ell'}{\ell''}=\frac{\ell'''(\hat\theta)}{2\ell''(\hat\theta)}e_t^2+O(e_t^3)$.

연습 4.3 문제 18쪽

$\Gamma(\alpha,1)$에서 $\operatorname{Var}X=\alpha$, $\mu_3=2\alpha$, $\mu_4=3\alpha^2+6\alpha$이므로 $(\bar X,S^2)$의 점근 공분산은 $\left(\begin{smallmatrix}\alpha & 2\alpha\\ 2\alpha & \mu_4-\alpha^2\end{smallmatrix}\right)=\left(\begin{smallmatrix}\alpha & 2\alpha\\ 2\alpha & 2\alpha^2+6\alpha\end{smallmatrix}\right)$(7회 정리 5.2와 예제 5.5). $g(a,b)=a^2/b$의 기울기는 (α,α)에서 $(2,-1)$이므로 분산은 $4\alpha-2\cdot2\cdot2\alpha+2\alpha^2+6\alpha=2\alpha(\alpha+1)$.

연습 4.4 문제 18쪽

(a) ℓ'가 미분가능하므로 $\tilde\theta_n$과 θ_0 사이에서 평균값 정리를 쓴다. (b) $\hat\theta_n^{(1)}-\theta_0=(\tilde\theta_n-\theta_0)+\frac{\ell'(\theta_0)+(\tilde\theta_n-\theta_0)\ell''(\theta_n^*)}{nI(\tilde\theta_n)}$이다. $\tilde\theta_n-\theta_0$의 항을 묶고 $\sqrt{n}$을 곱하면 주어진 식이다. (c) $|\theta_n^*-\theta_0|\le|\tilde\theta_n-\theta_0|\xrightarrow{P}0$이므로 정리 3.1의 증명 ④ ⑤(평균값 정리로 $\ell''(\theta_n^*)$를 $\ell''(\theta_0)$와 ℓ''' 항으로 나눈다)로 $-n^{-1}\ell''(\theta_n^*)\xrightarrow{P}I(\theta_0)$이다. I가 연속이므로 연속사상 정리로 $I(\tilde\theta_n)\xrightarrow{P}I(\theta_0)$이다. (d) (c)로 괄호 안은 $1-\frac{I(\theta_0)+o_P(1)}{I(\theta_0)+o_P(1)}\xrightarrow{P}0$이고, $\sqrt{n}(\tilde\theta_n-\theta_0)$는 확률적으로 유계이므로 곱은 $\xrightarrow{P}0$이다(7회 정리 2.5). 둘째 항은 정리 3.1의 증명 ③과 Slutsky 정리로 $\frac{1}{I(\theta_0)}\mathcal{N}(0,I(\theta_0))=\mathcal{N}(0,1/I(\theta_0))$이다. 다시 Slutsky로 합친다. (e) 그러면 첫 항은 $n^{1/4}(\tilde\theta_n-\theta_0)\cdot n^{1/4}\cdot o_P(1)$이고, $n^{1/4}$배를 이

기려면 괄호가 $n^{-1/4}$보다 빨리 0으로 가야 한다. 괄호는 대략 $\tilde{\theta}_n - \theta_0$의 크기로 가므로 곱은 $O_P(1)\cdot n^{1/4}\cdot O_P(n^{-1/4}) = O_P(1)$이 되어 0으로 가지 않을 수 있다. 시작점이 $\sqrt{n}$의 속도로 맞아야 한 걸음으로 충분하다.

5절

연습 5.1 문제 23쪽

$\ell(\lambda) = n\bar{x}\log\lambda - n\lambda$이므로 $\Lambda = 2n[\bar{x}\log\frac{\bar{x}}{\lambda_0} - (\bar{x}-\lambda_0)]$. $I = 1/\lambda$이므로 $W = \frac{n}{\bar{x}}(\bar{x}-\lambda_0)^2$, $\ell'(\lambda_0) = n(\frac{\bar{x}}{\lambda_0} - 1)$이므로 $R = \frac{n^2(\bar{x}-\lambda_0)^2/\lambda_0^2}{n/\lambda_0} = \frac{n(\bar{x}-\lambda_0)^2}{\lambda_0}$. $n = 280$, $\bar{x} = 0.7$, $\lambda_0 = 0.61$이면 $\Lambda = 3.55$, $W = 3.24$, $R = 3.72$.

연습 5.2 문제 23쪽

H_1의 MLE는 $\hat{p}_j = N_j/n$이므로 $\Lambda = 2\sum N_j \log\frac{N_j/n}{p_j^0} = G$. $N_j = E_j(1+u_j)$로 쓰면 $\sum E_j u_j = 0$이고 $G = 2\sum E_j(1+u_j)\log(1+u_j) \approx 2\sum E_j(1+u_j)(u_j - \frac{u_j^2}{2}) \approx \sum E_j u_j^2 = X^2$. 점수 검정은 $\ell'(\mathbf{p}^0)$와 정보 행렬로 만든 이차형식이고, 제약 $\sum p_j = 1$을 반영해 계산하면 $\sum \frac{(N_j - np_j^0)^2}{np_j^0}$, 곧 Pearson의 X^2가 정확히 나온다.

연습 5.3 문제 23쪽

가능도비 구간 $\{\theta : \ell(\theta) \ge \ell(\hat{\theta}) - \frac{1}{2}\chi^2_{1,\alpha}\}$은 가능도의 높이로만 정해지므로 $\eta = g(\theta)$로 바꿔도 같은 점들의 집합, 곧 g로 옮긴 구간이다. Wald 구간은 $\hat{\eta} \pm z/\sqrt{nI^*}$로 새 척도에서 대칭이므로 되돌리면 비대칭이 된다. 기병대: $\log\lambda$의 Wald 구간 $\log 0.7 \pm 1.96/\sqrt{280\cdot 0.7}$을 되돌리면 $[0.609,\ 0.805]$로, λ의 Wald $[0.602,\ 0.798]$보다 가능도비 $[0.607,\ 0.803]$에 가깝다.

연습 5.4 문제 23쪽

10만 번의 시뮬레이션에서 평균은 1.00, $P(\Lambda > 3.84)$는 0.05를 조금 넘는다. $\sum X_i \sim \text{Poi}(100)$이므로 시뮬레이션 없이 정확히 계산할 수도 있다(5.1절). 평균 1.002, 크기 0.0517이다.

6절

연습 6.1 문제 25쪽

$\ell_1 - \ell_0 = \sum_c[T_c\log\hat{\lambda}_c - 20\hat{\lambda}_c] - \sum_c[T_c\log\hat{\lambda} - 20\hat{\lambda}] = \sum_c T_c\log\frac{\hat{\lambda}_c}{\hat{\lambda}} - \left(\sum T_c - 280\hat{\lambda}\right)$이고 괄호는

0이다. $G = 2(\ell_1 - \ell_0) = 2\sum T_c \log \frac{T_c}{20\hat{\lambda}} = 26.1$ ($T_c = 0$인 항은 0).

연습 6.2 문제 25쪽

$t(4)$ 위치–척도 MLE는 위치 850.9, 척도 63.9이다. 정규의 평균 852.4와 1.5밖에 다르지 않다. 큰 잔차(620이나 1070 같은)는 무게 $\frac{\nu+1}{\nu+r^2}$이 작아(620의 무게는 0.29) 영향이 줄지만, 이 자료에서는 그런 관측이 적다. 참값과의 차이 118은 모든 관측에 공통인 계통 오차이므로 어떤 위치 추정량도 고치지 못한다.

연습 6.3 문제 26쪽

$\ell(\mu) = -\frac{n}{2}(\bar{x}-\mu)^2$+상수를 $\mu \geq 0$에서 최대화하면 $\hat{\mu} = \max(\bar{x}, 0)$. $\bar{x} > 0$이면 $2[\ell(\bar{x})-\ell(0)] = n\bar{x}^2$, $\bar{x} \leq 0$이면 0이다. $\sqrt{n}\bar{X} \sim \mathcal{N}(0,1)$이므로 확률 $\frac{1}{2}$로 0, 확률 $\frac{1}{2}$로 $Z^2 \mid Z > 0 \sim \chi^2(1)$이다.

7절

연습 7.1 문제 28쪽

$\theta = 0$이면 $\sqrt{n}\bar{X} = Z$이고 $T_n = \bar{X}1\{|Z| \geq n^{1/4}\}$이므로 $nET_n^2 = E[Z^2; |Z| \geq n^{1/4}] \to 0$. $\theta_n = cn^{-1/4}$이면 $\sqrt{n}\bar{X} = \sqrt{n}\theta_n + Z = cn^{1/4} + Z$이고 $|\bar{X}| < n^{-1/4} \iff |c + Zn^{-1/4}| < 1$의 확률이 1로 가므로 대부분 $T_n = 0$이고 $nE(T_n - \theta_n)^2 \approx n\theta_n^2 = c^2\sqrt{n} \to \infty$.

연습 7.2 문제 28쪽

$P(n(\theta - X_{(n)}) > x) = P(X_{(n)} < \theta - \frac{x}{n}) = (1 - \frac{x}{n\theta})^n \to e^{-x/\theta}$. $EX_{(n)} = \frac{n\theta}{n+1}$이므로 $\frac{n+1}{n}X_{(n)}$은 불편이다.

연습 7.3 문제 29쪽

$D_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 2\sigma^2)$의 가능도는 μ_i를 포함하지 않는다. $2\sigma^2$의 MLE는 $\frac{1}{n}\sum D_i^2$이므로 $\hat{\sigma}^2 = \frac{1}{2n}\sum D_i^2 \xrightarrow{P} \frac{ED^2}{2} = \sigma^2$. 예제의 MLE의 두 배다. 귀찮은 모수를 없앤 가능도로 추정하면 일치한다.

8절

연습 8.1 문제 31쪽

$\log f(X;\theta+\delta) = \log f(X;\theta) + \delta s + \frac{\delta^2}{2}\frac{\partial^2 \log f}{\partial\theta^2} + O(\delta^3)$이므로 $D = -\delta Es - \frac{\delta^2}{2}E\frac{\partial^2 \log f}{\partial\theta^2} + O(\delta^3) =$

$\frac{\delta^2}{2} I(\theta) + O(\delta^3)$. Poisson: $\lambda \log \frac{\lambda}{\lambda+\delta} - \lambda + \lambda + \delta = -\lambda \log(1 + \frac{\delta}{\lambda}) + \delta = \frac{\delta^2}{2\lambda} + O(\delta^3)$이고 $I = 1/\lambda$.

연습 8.2 문제 31쪽

$\ell = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum (Y_i - m(x_i;\theta))^2$이므로 σ^2에 관계없이 θ의 최대점은 제곱합의 최소점이다. $\hat{\sigma}^2 = \frac{1}{n}\sum (Y_i - m(x_i;\hat{\theta}))^2$.

연습 8.3 문제 31쪽

$\frac{d}{dT}\left[\frac{z_{iy}}{T} - \log \sum_k e^{z_{ik}/T}\right] = -\frac{z_{iy}}{T^2} + \frac{1}{T^2}\sum_k p_{ik}(T) z_{ik}$ ($p_{ik}(T)$는 온도 T의 softmax 확률). 합을 0으로 놓으면 $\sum_i z_{i,y_i} = \sum_i \sum_k p_{ik}(T) z_{ik}$.