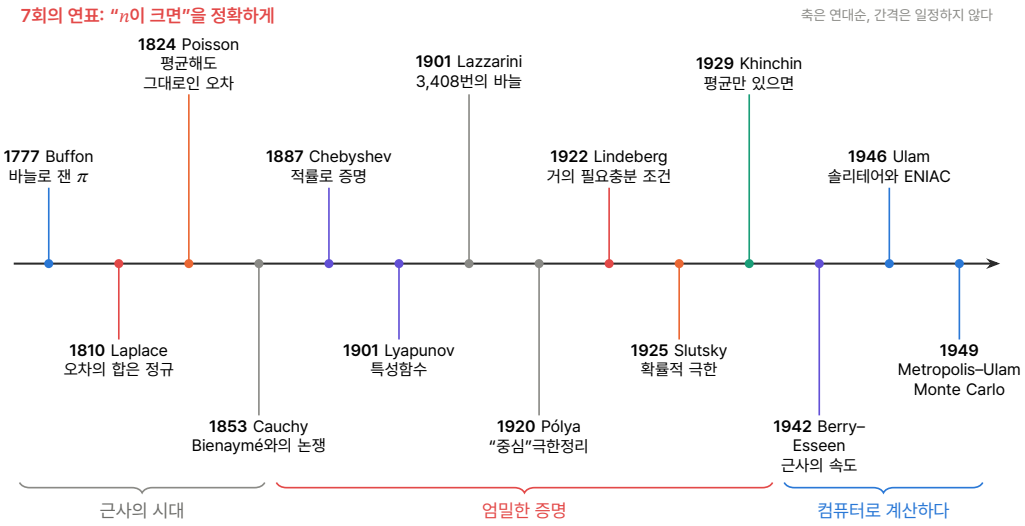
수리통계학 × 통계학의 역사

7회 강의노트

극한의 언어
수렴 · CLT · Monte Carlo

1777 → 1949

Buffon은 1777년 바늘을 마루에 던져 π를 재는 법을 보였고, Laplace는 1810년 작은 오차들의 합이 정규분포에 가까워진다는 것을 증명했다. 그 사이 Poisson은 아무리 많이 더해도 정규에 가까워지지 않는 분포를 찾아냈다. 이 노트는 "n이 크면"이라는 말을 확률수렴과 분포수렴으로 정확하게 만들고, 중심극한정리·Slutsky 정리·델타 방법을 증명한 뒤, 오차까지 말할 수 있는 시뮬레이션인 Monte Carlo 방법으로 이어간다.



다루는 것	확률수렴과 Khinchin의 대수의 법칙 · 분포수렴 · mgf 수렴 · 중심극한정리 · Slutsky · 델타 방법 · 다변량 CLT · Monte Carlo · 기각 표집 · 중요도 표집 · 분산이 무한할 때
원자료	Buffon(1777)의 바늘 · Lazzarini(1901)의 3,408번 · 6회의 범죄자 3,000명
노트북	`sessions/07_limits/07_limits.ipynb` — 절·정리 번호가 같다
워크북	`07_limits_workbook.pdf` (손으로 푸는 34문항) · 풀이판 `_workbook_sol.pdf`
선수 지식	1–6회 (특히 2회 Chebyshev·대수의 법칙·mgf, 5회 Cramér–Wold, 6회 t 분포)

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 ① ② ③으로 쓰고, 단계마다 오른쪽 끝에 근거(앞 정리의 번호나 도구)를 단다. 극한과 기댓값의 순서를 바꾸거나 적분가능성이 필요한 곳에는 **조건 확인** 단락이 붙는다. 읽을 때는 근거를 가리고 "왜 이 단계가 성립하는가"를 먼저 말해 본다. 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 [T1] 원소 추적, [T2] 대우, [T4] 부등식, [T5] 극한과 상극한, [T6] 적분의 순서, [T8] Taylor 전개를 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 1–6회에서 가져오는 것 2

0.2 표기 2

1 확률수렴 — Buffon의 바늘과 대수의 법칙 3

1.1 Buffon의 바늘 (1777) 3

1.2 확률수렴 4

1.3 평균만 있으면 — Khinchin의 대수의 법칙 5

2 분포수렴 8

2.1 정의와 예 8

2.2 두 수렴의 관계 9

2.3 mgf로 분포수렴 보이기 11

3 중심극한정리 13

3.1 mgf로 증명하기 13

3.2 더 넓은 정리와 속도 15

4 Slutsky의 정리와 델타 방법 18

4.1 Slutsky의 정리 (1925) 18

4.2 델타 방법 19

5 다변량으로 — 선형결합 하나면 된다 23

5.1 벡터의 수렴과 다변량 중심극한정리 23

5.2 다변량 델타 방법 24

6 계산으로 — Monte Carlo (1946 → 1949) 27

6.1 오차를 아는 시뮬레이션 27

6.2 원하는 분포에서 뽑기 — 기각 표집 . . . 28
6.3 오차를 줄이기 — 중요도 표집과 반대 변량 . . . 29

7 부숴보기 — 분산이 무한하고, 꼬리를 물을 때 32

7.1 평균은 있지만 분산이 없을 때 . . . 32
7.2 평균도 없을 때 . . . 33
7.3 가정이 성립해도 — 꼬리의 상대오차 . . . 34
7.4 기울기가 0이면 . . . 35

8 Beautiful insights — 평균을 내는 기계들 36

8.1 미니배치 경사의 잡음 . . . 36
8.2 MC dropout — 예측을 여러 번 평균하기 . . . 37
8.3 대각합을 무작위로 — Hutchinson 추정량 . . . 37

9 정리와 역사 노트 39

9.1 한 장 요약 . . . 39
9.2 Stigler의 명명 법칙 . . . 39
9.3 7 Pillars 지도 . . . 40
9.4 다음 회 — 추론의 기초: 표본 · 신뢰구간 · 유의성 검정 (1710 → 1937) . . . 40

A 연습문제 풀이 42

0

들어가며

> **여는 질문.** 사람의 키, 측정의 오차, 동전 백 번의 앞면 수, 효모 세포의 수(4회), 범죄자 넷의 평균 키(6회)는 모두 종 모양으로 흩어진다. 왜 그토록 많은 것이 같은 모양인가? 그리고 왜 어떤 것은 그렇지 않은가?

5회에서 de Moivre(1733)는 동전 던지기의 앞면 수가 정규곡선을 따른다는 것을 계산했다. Laplace(1810)는 이것을 **어떤** 작은 오차들의 합에도 넓혔다. 각 오차의 분포가 무엇이든, 독립인 오차를 많이 더하면 합은 정규분포에 가까워진다. 이것이 중심극한정리다. 그런데 Poisson(1824)은 이상한 분포를 하나 찾아냈다. 그 분포를 따르는 관측은 아무리 많이 평균해도 관측 하나보다 나아지지 않는다. 정리에는 조건이 있었다.

이 노트는 "n이 크면 대략 정규"라는 말을 정확하게 만든다. 확률변수의 수열이 "수렴한다"는 말에는 두 가지 뜻이 있다(1–2절). 중심극한정리는 mgf로 증명하고(3절), Slutsky의 정리와 델타 방법으로 표본평균의 함수까지 넓힌다(4절). 다변량으로 가면 8회의 χ^2 검정이 준비된다(5절). 그리고 이 극한 정리들이 Monte Carlo 시뮬레이션의 오차를 말해 준다(6절). Buffon(1777)의 바늘에서 Los Alamos(1949)의 컴퓨터까지, 난수로 계산하는 방법은 중심극한정리 덕분에 "얼마나 틀릴 수 있는지"를 함께 내놓는다. 7절은 분산이 무한한 분포에서 정리가 무너지는 모습을 보고, 8절은 미니배치 경사의 잡음과 MC dropout에서 같은 정리를 만난다.

학습 목표

1. 확률수렴과 분포수렴을 정의하고 둘의 관계를 증명한다. 평균만 있으면 성립하는 Khinchin의 대수의 법칙을 증명한다.
2. mgf로 중심극한정리를 증명하고, Berry–Esseen 한계로 근사의 속도를 말한다.
3. Slutsky의 정리와 델타 방법을 증명하고, 스튜던트화한 평균·로그 오즈·분산 안정화 변환에 쓴다.
4. 다변량 중심극한정리를 선형결합의 mgf로 증명하고 다항분포에 쓴다.
5. Monte Carlo 적분의 신뢰구간, 기각 표집, 중요도 표집을 증명하고 구현한다. 분산이 무한할 때 무엇이 달라지는지 시뮬레이션으로 본다.

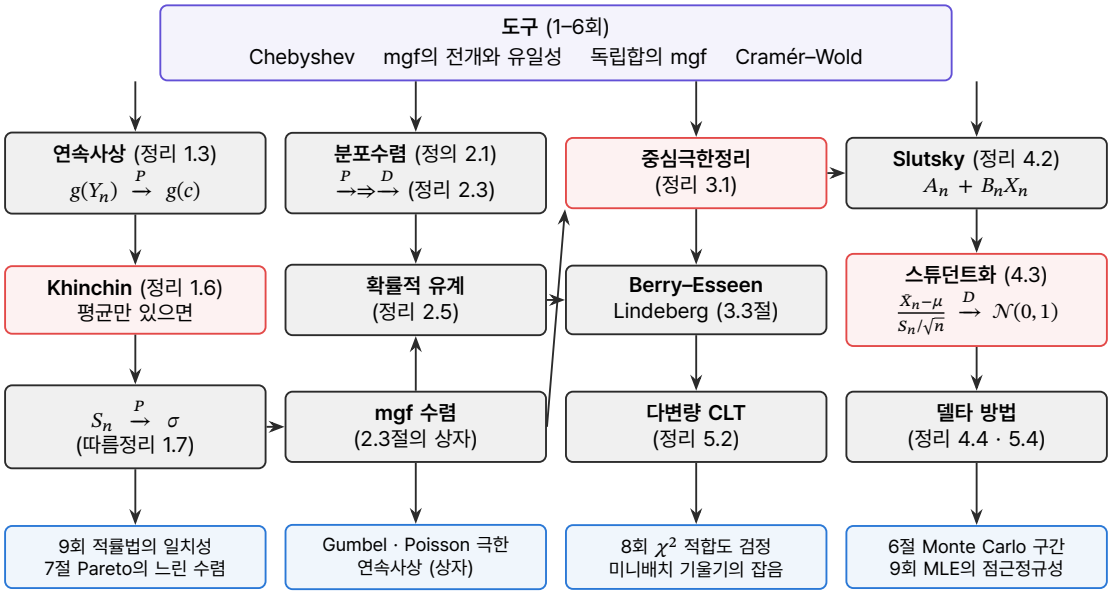


그림 0.1 이번 회차 정리들의 의존 관계. 화살표는 "증명에 쓰인다"는 뜻이다.

0.1 1–6회에서 가져오는 것

앞 회차	내용
2회 정의 6.1 · 정리 5.3 · 6.3	확률수렴, Chebyshev 부등식, 분산이 있을 때의 대수의 법칙
2회 정리 3.3 · 3.6 · 3.3절의 상자	mgf의 미분은 적률, 아핀변환과 독립합의 mgf, mgf가 같으면 분포가 같다
2회 정리 4.7 · 정리 8.1	확률적분변환(역함수 표집), Chernoff 한계
3회 정리 4.2 · 6.3	공분산의 계산, $\mathrm{Var}\,\bar{X} = \sigma^2/n$
4회 정리 2.3 · 3.7	이항의 Poisson 극한, 감마의 mgf
5회 따름정리 1.3 · 정리 3.8 · 3.10	적분형 de Moivre–Laplace, 다변량 정규의 선형변환, Cramér–Wold 판정
6회 정리 3.3 · 3.4 · 7절	$t(r) \to \mathcal{N}(0,1)$, Student의 t, 비정규 모집단에서의 포함확률

0.2 표기

$X_1, X_2, \ldots$ 가 독립이고 같은 분포를 따르면 "iid 수열"이라 하고, 평균 μ, 분산 σ^2, 표본평균 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$로 쓴다. Φ와 ϕ는 표준정규의 분포함수와 밀도, z_p는 위쪽 p 분위수다($z_{0.025} = 1.960$). 수렴의 기호는 확률수렴 $\xrightarrow{P}$, 분포수렴 $\xrightarrow{D}$이다. 함수 F의 **연속점**은 F가 연속인 점이다.

1

확률수렴 — Buffon의 바늘과 대수의 법칙

1.1 Buffon의 바늘 (1777)

> **역사** 박물학자 Buffon은 1733년 파리 과학 아카데미에서 이런 문제를 냈고, 풀이는 1777년 *Essai d'arithmétique morale*에 실었다. 폭 d의 널빤지를 깐 마루에 길이 $\ell \le d$인 바늘을 아무렇게나 던진다. 바늘이 널빤지의 이음매에 걸칠 확률은 얼마인가? 답에 π가 나온다. 그러면 거꾸로, 바늘을 많이 던져 π를 잴 수 있다. Laplace(1812)가 이것을 지적했고, 19세기에 여러 사람이 실제로 던졌다. 난수로 계산하는 방법, 곧 Monte Carlo의 가장 오래된 예다.

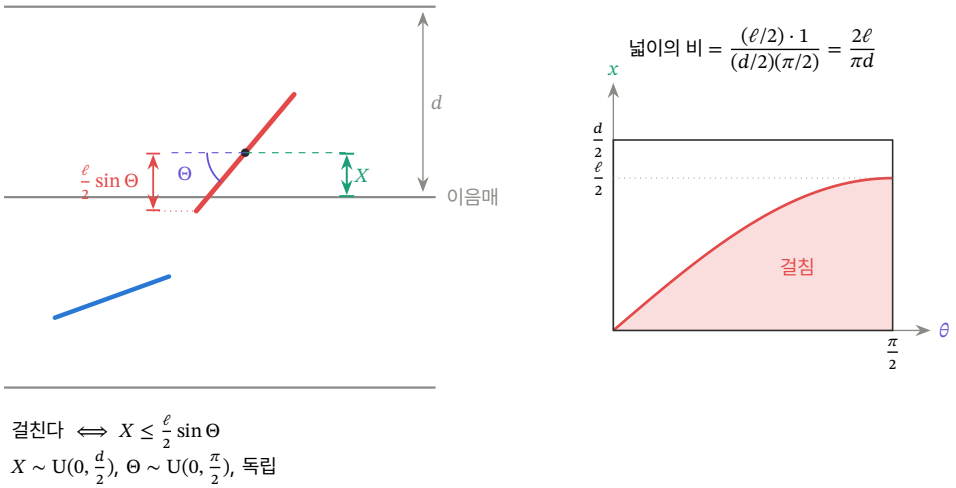


그림 1.1 바늘이 이음매에 걸치는 조건. 중심에서 가장 가까운 이음매까지의 거리 X가 바늘 반쪽의 수직 성분 $\frac{\ell}{2}\sin\Theta$보다 작으면 걸친다. 오른쪽: (Θ, X)의 직사각형에서 걸치는 영역(색칠)의 넓이 비가 $\frac{2\ell}{\pi d}$이다.

정리 1.1 (Buffon의 바늘) 바늘의 중심에서 가장 가까운 이음매까지의 거리 X와 바늘이 이음매와 이루는 예각 Θ가 독립이고 각각 $\mathrm{U}(0, \frac{d}{2})$, $\mathrm{U}(0, \frac{\pi}{2})$이면, $\ell \le d$일 때

$$P(\text{바늘이 이음매에 걸친다}) = \frac{2\ell}{\pi d}.$$

증명. **전략** 걸치는 사건을 (X, Θ)의 부등식 $X \le \frac{\ell}{2}\sin\Theta$로 쓰고, 독립에서 나오는 결합밀도를 그 영역 위에서 적분한다. 단계별 증명은 연습 1.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 43쪽에 있다.

n번 던져 K_n번 걸쳤다면 $\frac{K_n}{n} \approx \frac{2\ell}{\pi d}$이므로 π의 추정값은 $\hat{\pi}_n = \frac{2\ell n}{d K_n}$이다. 이것이 π로 "수렴"한다는 것은 무슨 뜻인가?

1.2 확률수렴

2회 정의 6.1에서 확률변수 수열이 상수로 확률수렴하는 것을 정의했다. 극한이 확률변수인 경우까지 넓힌다.

정의 1.2 (확률수렴) 같은 확률공간의 확률변수 $Y_1, Y_2, \dots$와 Y가 모든 $\varepsilon > 0$에 대해 $\lim_{n\to\infty} P\big(|Y_n - Y| \ge \varepsilon\big) = 0$을 만족하면 $Y_n \xrightarrow{P} Y$라 쓴다. Y가 상수 c이면 2회 정의 6.1이다.

정리 1.3 (연속사상 정리 (확률수렴))
(a) $Y_n \xrightarrow{P} c$이고 g가 c에서 연속이면 $g(Y_n) \xrightarrow{P} g(c)$.
(b) $Y_n \xrightarrow{P} a$, $W_n \xrightarrow{P} b$이고 $g : \mathbb{R}^2 \to \mathbb{R}$가 (a, b)에서 연속이면 $g(Y_n, W_n) \xrightarrow{P} g(a, b)$.

증명. **전략** g의 연속성에서 주어진 ε에 맞는 δ를 얻고, "g의 값이 멀다"는 사건을 "입력 가운데 하나가 멀다"는 사건으로 덮는다. (a)는 (b)의 특수한 경우다.

① $\varepsilon > 0$을 잡는다. g가 (a, b)에서 연속이므로 $\delta > 0$이 있어 $|y - a| < \delta$이고 $|w - b| < \delta$이면 $|g(y, w) - g(a, b)| < \varepsilon$이다. 연속의 정의 [T5]

② ①의 대우를 사건으로 쓰면 $\{|g(Y_n, W_n) - g(a, b)| \ge \varepsilon\} \subset \{|Y_n - a| \ge \delta\} \cup \{|W_n - b| \ge \delta\}$이다. 대우 [T2], 원소 추적 [T1]

③ 확률을 취한다.

$$P\big(|g(Y_n, W_n) - g(a, b)| \ge \varepsilon\big) \le P(|Y_n - a| \ge \delta) + P(|W_n - b| \ge \delta). \qquad \text{1회 정리 4.3(d)(g)}$$

④ 오른쪽의 두 항은 가정으로 $n \to \infty$에서 0으로 간다. ε은 임의였으므로 $g(Y_n, W_n) \xrightarrow{P} g(a, b)$이다. 정의 1.2

⑤ (a)는 $g(y,w)=g(y)$, $W_n \to b=0$으로 둔 (b)다. (b)

■

따름정리 1.4 (확률수렴의 사칙연산) $Y_n \xrightarrow{P} a$, $W_n \xrightarrow{P} b$이면 $Y_n+W_n \xrightarrow{P} a+b$, $Y_nW_n \xrightarrow{P} ab$이고, $b\neq 0$이면 $Y_n/W_n \xrightarrow{P} a/b$.

증명. $g(y,w)=y+w, yw, y/w$는 각각 (a,b)에서 연속이다(마지막은 $b\neq 0$일 때). ■

예제 1.5 (바늘로 잰 π) K_n/n은 걸칠 확률 $p=\frac{2\ell}{\pi d}$의 시행 n번의 성공 비율이므로 2회 정리 6.2로 $K_n/n \xrightarrow{P} p$이다. $g(x)=\frac{2\ell}{dx}$는 $p>0$에서 연속이므로 정리 1.3(a)로 $\hat{\pi}_n=g(K_n/n) \xrightarrow{P} g(p)=\pi$. 얼마나 빨리 다가가는지는 4절의 델타 방법이 답한다($\ell=d$, $n=1000$이면 표준편차 약 0.075).

1.3 평균만 있으면 — Khinchin의 대수의 법칙

2회 정리 6.3은 분산이 유한하다는 가정 아래 Chebyshev 부등식으로 증명했다. 분산이 무한한 분포는 드물지 않다(7절). Khinchin(1929)은 평균만 있으면 충분하다는 것을 보였다. 증명의 착상은 **잘라 내기**다. 큰 값을 잘라 분산이 있는 변수로 바꾸고, 잘라 낸 부분이 무시할 만함을 보인다.

정리 1.6 (Khinchin의 약한 대수의 법칙) $X_1, X_2, \ldots$가 iid이고 $E|X_1|<\infty$, $\mu=EX_1$이면 $\bar{X}_n \xrightarrow{P} \mu$.

증명. **전략** 큰 값을 잘라 분산이 있는 변수로 바꾼다(잘라 내기). 잘라 낸 부분이 확률로도 평균으로도 무시할 만함을 보이고, 남은 부분에 Chebyshev 부등식을 쓴다.

X를 X_1과 같은 분포의 변수로 쓰고 $E(|X|;A)=E\big(|X|\,1_A\big)$, $\tau(a)=E(|X|;|X|>a)$로 둔다.

① $a\to\infty$이면 $\tau(a)\to 0$이다. $E|X|<\infty$, 수렴하는 적분의 꼬리 [T5]

② n을 고정하고 $Y_i=X_i 1\{|X_i|\le n\}$, $T_n=\frac{1}{n}\sum_{i=1}^{n} Y_i$로 둔다. 잘라 내기

③ $\bar{X}_n \neq T_n$이면 어떤 i에서 $|X_i|>n$이다. $n\,1\{|X|>n\}\le |X|\,1\{|X|>n\}$이므로

$$P(\bar{X}_n\neq T_n)\le nP(|X|>n)\le E(|X|;|X|>n)=\tau(n). \quad \text{Boole 부등식, 단조성}$$

④ $ET_n-\mu=EY_1-EX=-E(X;|X|>n)$이므로 $|ET_n-\mu|\le\tau(n)$이다. 2회 정리 1.10

⑤ 아무 $M>0$에 대해 $X^2 1\{|X|\le n\}\le M|X|+n|X|1\{|X|>M\}$이다. $|X|\le M$이면 $X^2\le M|X|$이고, $M<|X|\le n$이면 $X^2\le n|X|$이기 때문이다. 경우 나누기 [T4]

⑥ Y_i들은 독립이고 유계이므로 분산이 있다.

$$\operatorname{Var} T_n = \frac{\operatorname{Var} Y_1}{n} \le \frac{E(X^2; |X| \le n)}{n} \le \frac{M\,E|X|}{n} + \tau(M). \qquad \text{2회 정리 2.8, ⑤}$$

⑦ n이 커서 $\tau(n) < \varepsilon/2$이면 ④로 $|ET_n - \mu| < \varepsilon/2$이고, $\{|\bar X_n - \mu| \ge \varepsilon\} \subset \{\bar X_n \ne T_n\} \cup \{|T_n - ET_n| \ge \varepsilon/2\}$이다. 따라서

$$P(|\bar X_n - \mu| \ge \varepsilon) \le \tau(n) + \frac{4}{\varepsilon^2}\Big(\frac{M\,E|X|}{n} + \tau(M)\Big). \qquad \text{③⑥, 2회 정리 5.3}$$

⑧ $n \to \infty$에서 오른쪽의 상극한은 $\frac{4}{\varepsilon^2}\tau(M)$ 이하다. M은 아무리 크게 잡아도 되므로 ①로 극한은 0이다. [T5]

조건 확인 쓴 가정은 $E|X| < \infty$ 하나다. 분산은 잘라 낸 Y_i에만 썼고, $|Y_i| \le n$이므로 모든 적률이 있다. 독립은 ⑥에서 T_n의 분산을 나눌 때 썼다. ∎

> **따름정리 1.7** (표본적률과 표본분산의 일치성) $X_1, X_2, \dots$가 iid이고 $E|X|^k < \infty$이면 $\frac{1}{n}\sum X_i^k \xrightarrow{P} EX^k$이다. 특히 $\sigma^2 < \infty$이면 $S_n^2 \xrightarrow{P} \sigma^2$, $S_n \xrightarrow{P} \sigma$.

증명. **전략** 적률은 iid 변수 X_i^k의 평균이므로 정리 1.6을 바로 쓴다. S_n^2은 두 표본적률의 연속함수로 쓰고 정리 1.3를 쓴다. 단계별 증명은 연습 1.7(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 43쪽에 있다.

따름정리의 가정은 분산의 유한성뿐이다. Chebyshev로 $S_n^2 \to \sigma^2$을 보이려면 $\operatorname{Var} S_n^2$, 곧 4차 적률이 필요했다(6회 연습 1.2). Khinchin 덕분에 적률을 하나 덜 요구한다. 9회에서 이것은 적률법 추정량의 **일치성**이 된다.

> **증명 없이 쓰는 사실** (거의 확실한 수렴과 강한 대수의 법칙) $P(\lim_n \bar X_n = \mu) = 1$, 곧 거의 모든 표본경로에서 $\bar X_n$이 μ로 가는 것을 **강한** 대수의 법칙이라 한다. iid이면 $E|X| < \infty$가 필요충분조건이다(Kolmogorov, 1933). 확률수렴보다 강한 이 수렴은 측도론의 도구(Borel–Cantelli 보조정리)가 필요해 이 과정에서는 다루지 않는다. 통계적 추론에 필요한 것은 대부분 확률수렴과 분포수렴이다.

연습문제

> **연습 1.1** (Buffon의 국수 ★★) 길이 ℓ인 바늘(길이에 제한 없음)이 이음매와 만나는 **횟수** N의 기댓값이 $\frac{2\ell}{\pi d}$임을 보여라. (힌트: 바늘을 길이 $\ell/m \le d$인 토막 m개로 나누면 N은 토막마다의 교차 횟수의 합이다. 기댓값의 선형성.) 휘어진 철사(국수)여도 같다(Barbier, 1860).
> 풀이 42쪽

연습 1.2 (연속이 아니면 ★) $Y_n = 1/n$ (상수)이고 $g(y) = 1\{y > 0\}$이면 $Y_n \xrightarrow{P} 0$이지만 $g(Y_n) \not\to g(0)$임을 보여라. 정리 1.3의 어느 가정이 깨졌나? 풀이 42쪽

연습 1.3 (기댓값은 따라오지 않는다 ★) $P(Y_n = n) = \frac{1}{n}$, $P(Y_n = 0) = 1 - \frac{1}{n}$이면 $Y_n \xrightarrow{P} 0$이지만 $EY_n = 1$임을 보여라. 확률수렴은 꼬리의 드문 큰 값을 보지 않는다. 풀이 42쪽

연습 1.4 (분산이 없는 표본 ★★) $t(2)$ 분포는 평균이 0이고 분산이 무한하다(6회 정리 3.3). 정리 1.6으로 $\bar{X}_n \xrightarrow{P} 0$임을 말하고, 노트북으로 $P(|\bar{X}_n| > 0.1)$을 $n = 100$과 10,000에서 구하라. 2회 정리 6.3은 왜 쓸 수 없나? 풀이 42쪽

연습 1.5 (잘라 내기의 부등식 ★★) 정리 1.6의 증명 ⑤의 부등식 $X^2 1\{|X| \le n\} \le M|X| + n|X|1\{|X| > M\}$을 $M \ge n$인 경우와 $M < n$인 경우로 나누어 확인하라. 풀이 42쪽

연습 1.6 (Buffon의 바늘 증명 ★ 유도형) 정리 1.1을 다음 순서로 증명하라.

(a) 바늘 반쪽이 이음매에 수직인 방향으로 차지하는 길이가 $\frac{\ell}{2}\sin\Theta$임을 그림으로 보이고, 걸치는 사건이 $\{X \le \frac{\ell}{2}\sin\Theta\}$임을 설명하라.

(b) X와 Θ가 독립이므로 (X, Θ)의 결합밀도가 직사각형 $(0, \frac{d}{2}) \times (0, \frac{\pi}{2})$ 위에서 상수 $\frac{4}{\pi d}$임을 보여라.

(c) $\ell \le d$이면 모든 θ에서 $\frac{\ell}{2}\sin\theta \le \frac{d}{2}$임을 확인하라. 이 조건이 없으면 (d)의 안쪽 적분은 어떻게 바뀌는가?

(d) 이중적분을 계산해 $\frac{2\ell}{\pi d}$를 얻어라.

풀이 43쪽

연습 1.7 (표본적률의 일치성 증명 ★ 유도형) 따름정리 1.7를 다음 순서로 증명하라.

(a) $X_1^k, X_2^k, \ldots$가 iid이고 $E|X_1^k| < \infty$임을 확인하고 정리 1.6을 써라.

(b) $S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum X_i^2 - \bar{X}_n^2\right)$임을 보여라.

(c) (a)와 따름정리 1.4로 $S_n^2 \xrightarrow{P} \sigma^2$을 얻어라. 상수 수열 $\frac{n}{n-1}$은 어떻게 처리하는가?

(d) 제곱근에 정리 1.3(a)를 써서 $S_n \xrightarrow{P} \sigma$를 얻어라. $\sigma^2 = 0$이면 어떻게 되는가?

풀이 43쪽

2

분포수렴

대수의 법칙은 $\bar{X}_n$이 μ **한 점**으로 모인다고 말한다. 그 점 주위에서 $\bar{X}_n$이 어떤 **모양**으로 흩어지는지는 말하지 않는다. 모양의 극한을 말하려면 분포함수의 극한이 필요하다.

2.1 정의와 예

> **정의 2.1** (분포수렴) X_n의 분포함수 F_n과 X의 분포함수 F가 F의 모든 연속점 x에서 $F_n(x) \to F(x)$를 만족하면 X_n이 X로 **분포수렴**한다고 하고 $X_n \xrightarrow{D} X$로 쓴다. X의 분포를 X_n의 **극한분포**라 한다.

분포수렴은 분포만의 성질이다. X_n과 X가 같은 확률공간에 있을 필요도 없다. "연속점"이라는 단서가 왜 필요한지는 첫 예가 보여 준다.

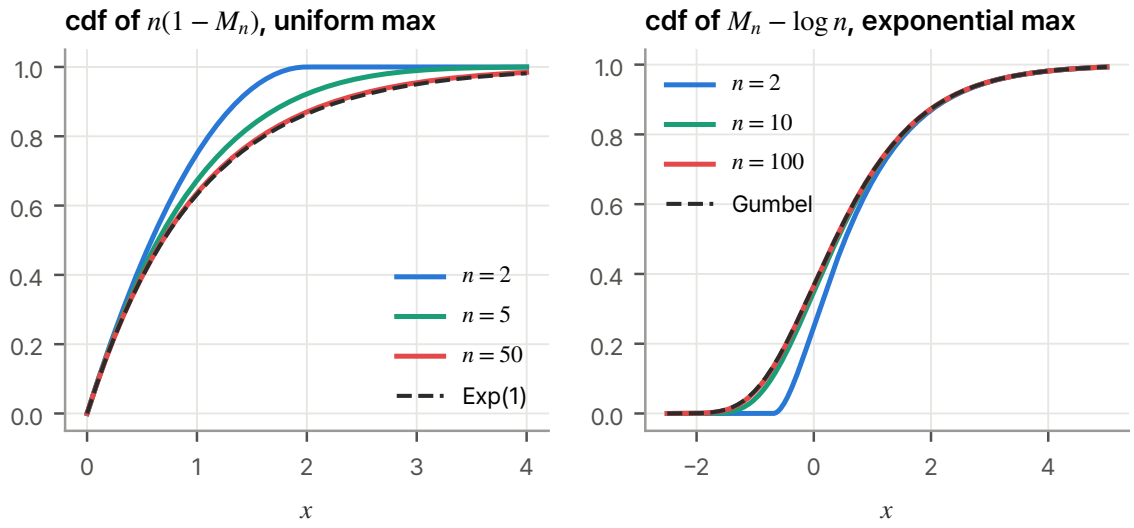


그림 2.1 왼쪽: $n(1-M_n)$의 분포함수($n=2,5,50$)가 Exp(1)의 분포함수(점선)로 다가간다. 오른쪽: Exp(1) 표본 n개의 최댓값에서 $\log n$을 뺀 값은 Gumbel 분포로 다가간다(연습 2.2).

예제 2.2 (분포수렴의 세 예)

(a) $X_n = \frac{1}{n}$ (상수)이면 $F_n(0)=0$이지만 극한 $X=0$의 $F(0)=1$이다. 그러나 0은 F의 불연속점이고, 다른 모든 x에서 $F_n(x)\to F(x)$이므로 $X_n \xrightarrow{D} 0$이다. 연속점의 단서가 없으면 이 당연한 수렴이 수렴이 아니게 된다.

(b) $U_1,\dots,U_n \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$의 최댓값 M_n에 대해 $x\ge 0$이면

$$P\big(n(1-M_n)>x\big)=P\Big(M_n<1-\frac{x}{n}\Big)=\Big(1-\frac{x}{n}\Big)^n \longrightarrow e^{-x}$$

이므로 $n(1-M_n)\xrightarrow{D}\mathrm{Exp}(1)$. 최댓값은 1에 $1/n$의 속도로 다가가고(3회 예제 3.5), 그 간격을 n배로 키우면 지수분포다. $n=5$에서 $P(n(1-M_n)>1)=0.328$, 극한은 $e^{-1}=0.368$이다.

(c) $X_n\sim \mathrm{b}(n,\lambda/n)$이면 4회 정리 2.3으로 모든 정수 k에서 $P(X_n=k)\to e^{-\lambda}\lambda^k/k!$이다. 분포함수는 유한 개의 확률의 합이므로 $X_n\xrightarrow{D}\mathrm{Poi}(\lambda)$.

2.2 두 수렴의 관계

정리 2.3 (확률수렴과 분포수렴)

(a) $X_n\xrightarrow{P}X$이면 $X_n\xrightarrow{D}X$.

(b) 상수 c에 대해 $X_n\xrightarrow{D}c$이면 $X_n\xrightarrow{P}c$.

증명. **전략** (a) X_n과 X가 ε 안에 있으면 두 사건 $\{X_n \le x\}$와 $\{X \le x \pm \varepsilon\}$이 서로 포함되므로, $F_n(x)$를 $F(x \pm \varepsilon)$ 사이에 끼운다. (b) 극한이 상수이면 연속점이 c 하나만 빼고 모두이므로, c의 양쪽에서 분포함수를 읽는다.

(a)

① $\varepsilon > 0$을 잡는다. $X_n \le x$이고 $|X_n - X| < \varepsilon$이면 $X < x + \varepsilon$이므로 $\{X_n \le x\} \subset \{X \le x + \varepsilon\} \cup \{|X_n - X| \ge \varepsilon\}$이다. 원소 추적 [T1]

② 같은 방법으로 $\{X \le x - \varepsilon\} \subset \{X_n \le x\} \cup \{|X_n - X| \ge \varepsilon\}$이다. 확률을 취하면

$$F(x-\varepsilon) - P(|X_n - X| \ge \varepsilon) \le F_n(x) \le F(x+\varepsilon) + P(|X_n - X| \ge \varepsilon).$$ 1회 정리 4.3(d)(g)

③ $n \to \infty$이면 $P(|X_n - X| \ge \varepsilon) \to 0$이므로

$$F(x-\varepsilon) \le \liminf_n F_n(x) \le \limsup_n F_n(x) \le F(x+\varepsilon).$$ 확률수렴, 상극한·하극한 [T5]

④ x가 F의 연속점이면 $\varepsilon \downarrow 0$에서 양 끝이 $F(x)$로 가므로 $\lim_n F_n(x) = F(x)$이다. 연속점의 정의

(b)

① 상수 c의 분포함수는 $x < c$에서 0, $x \ge c$에서 1이므로 c가 아닌 모든 점이 연속점이다. 정의 2.1

② $\{|X_n - c| \ge \varepsilon\} \subset \{X_n \le c - \varepsilon\} \cup \{X_n > c + \frac{\varepsilon}{2}\}$이므로

$$P(|X_n - c| \ge \varepsilon) \le F_n(c-\varepsilon) + 1 - F_n\big(c + \tfrac{\varepsilon}{2}\big).$$ [T1], Boole 부등식

③ $c - \varepsilon$과 $c + \frac{\varepsilon}{2}$는 연속점이므로 오른쪽은 $0 + 1 - 1 = 0$으로 간다. ①, 분포수렴

조건 확인 (b)에서 $c + \varepsilon$ 대신 $c + \frac{\varepsilon}{2}$를 쓴 것은 $P(X_n \ge c + \varepsilon)$을 분포함수로 쓰기 위해서다. $\{X_n \ge c + \varepsilon\} \subset \{X_n > c + \frac{\varepsilon}{2}\}$이고, 뒤 사건의 확률은 $1 - F_n(c + \frac{\varepsilon}{2})$이다. ■

> **가정 주의** (a)의 역은 극한이 상수가 아니면 거짓이다. $Z \sim \mathcal{N}(0,1)$이고 $X_n = -Z$이면 대칭으로 모든 n에서 X_n의 분포가 Z와 같으므로 $X_n \xrightarrow{D} Z$이지만, $|X_n - Z| = 2|Z|$이므로 X_n은 Z로 확률수렴하지 않는다. 분포수렴은 값이 가까워지는 것이 아니라 **분포가** 가까워지는 것이다.

분포수렴하는 수열은 "도망가지" 않는다. 이 성질이 4절의 증명에서 쓰인다.

정의 2.4 (확률적으로 유계) 모든 $\eta > 0$에 대해 B가 있어 모든 n에서 $P(|X_n| \le B) \ge 1-\eta$이면 $\{X_n\}$이 **확률적으로 유계**라 한다.

정리 2.5 (분포수렴하면 확률적으로 유계) $X_n \xrightarrow{D} X$이면 $\{X_n\}$은 확률적으로 유계다. 또 $\{Y_n\}$이 확률적으로 유계이고 $W_n \xrightarrow{P} 0$이면 $Y_n W_n \xrightarrow{P} 0$이다.

증명. **전략** 극한의 분포함수의 양 끝에서 연속점 $\pm B$를 골라 n이 클 때를 덮고, 처음 유한 개는 따로 덮는다. 둘째 주장은 $|Y_n|$이 큰 사건과 $|W_n|$이 큰 사건으로 나눈다. 단계별 증명은 연습 2.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 44쪽에 있다.

2.3 mgf로 분포수렴 보이기

분포함수를 직접 다루기 어려울 때, 2회의 mgf의 유일성(3.3절의 상자)이 극한으로 넓어진다.

증명 없이 쓰는 사실 (mgf의 수렴은 분포수렴을 준다) X_n의 mgf $M_n(t)$와 X의 mgf $M(t)$가 어떤 $h > 0$에 대해 $|t| < h$에서 존재하고 $M_n(t) \to M(t)$이면 $X_n \xrightarrow{D} X$이다(Curtiss, 1942). 증명은 유일성 정리와 같은 역변환 공식을 극한까지 끌고 간다. 같은 사실이 벡터의 mgf에도 성립한다.

예제 2.6 (mgf로 본 두 극한) (a) $\mathrm{b}(n, \lambda/n)$의 mgf는 $\left(1 + \frac{\lambda(e^t-1)}{n}\right)^n \to e^{\lambda(e^t-1)}$이고 이것은 $\mathrm{Poi}(\lambda)$의 mgf다. 예제 2.2(c)를 다시 얻는다. (b) 기하분포의 성공 확률이 $p = \lambda/n$이면 첫 성공까지의 시행 수를 n으로 나눈 값은 $\mathrm{Exp}(\lambda)$로 분포수렴한다(연습 2.3). 이산 시간의 기다림이 연속 시간의 기다림(4회 3절)이 된다.

증명 없이 쓰는 사실 (연속사상 정리 (분포수렴)) $X_n \xrightarrow{D} X$이고 g가 연속이면 $g(X_n) \xrightarrow{D} g(X)$이다(Mann과 Wald, 1943). 일반적인 증명에는 이 과정에서 다루지 않는 표현 정리가 필요하다. 이 노트에서 쓰는 경우($g(x) = x^2$, 아핀변환)는 연습 2.4처럼 분포함수로 직접 보일 수 있다.

연습문제

연습 2.1 (최솟값 ★) $U_1, \dots, U_n \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$의 최솟값 m_n에 대해 $nm_n \xrightarrow{D} \mathrm{Exp}(1)$임을 보여라. 풀이 43쪽

연습 2.2 (지수의 최댓값과 Gumbel ★★) $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathrm{Exp}(1)$의 최댓값 M_n에 대해 $P(M_n - \log n \le x) \to \exp(-e^{-x})$임을 보여라. 이 극한분포를 Gumbel 분포라 한다. $n = 10$과 100에서 $x = 0$의 값을 극한 e^{-1}과 비교하라. 풀이 43쪽

연습 2.3 (기다림의 극한 ★★) X_n이 성공 확률 $p_n = \lambda/n$인 시행에서 첫 성공까지의 시행 수이면 $X_n/n \xrightarrow{D} \mathrm{Exp}(\lambda)$임을 $P(X_n/n > x)$로 보여라. 풀이 43쪽

연습 2.4 (제곱의 분포수렴 ★★) $Z_n \xrightarrow{D} Z \sim \mathcal{N}(0,1)$이면 $Z_n^2 \xrightarrow{D} \chi^2(1)$임을 $P(Z_n^2 \le y) = F_n(\sqrt{y}) - P(Z_n < -\sqrt{y})$로 보여라. ($P(Z_n < -\sqrt{y})$를 $F_n(-\sqrt{y} \pm \varepsilon)$ 사이에 끼운다.) 풀이 44쪽

연습 2.5 (합은 따라오지 않는다 ★) $X_n \xrightarrow{D} Z$, $Y_n \xrightarrow{D} Z$ ($Z \sim \mathcal{N}(0,1)$)이어도 $X_n + Y_n$이 $\mathcal{N}(0,2)$로 분포수렴한다고 할 수 없음을 $Y_n = -X_n$으로 보여라. 4절의 Slutsky 정리가 한쪽을 **상수로의** 확률수렴으로 요구하는 이유다. 풀이 44쪽

연습 2.6 (확률적 유계성 증명 ★★ 유도형) 정리 2.5를 다음 순서로 증명하라. $\eta > 0$을 고정한다.

(a) 분포함수의 불연속점이 셀 수 있게 많다는 것(1회 연습 6.2)으로 $F(-B) < \frac{\eta}{4}$, $F(B) > 1 - \frac{\eta}{4}$인 F의 연속점 $\pm B$가 있음을 보여라.

(b) N이 있어 $n \ge N$이면 $P(|X_n| \le B) \ge F_n(B) - F_n(-B) \ge 1 - \eta$임을 보여라. 첫 부등식에서 $F_n(-B) \ge P(X_n < -B)$를 쓴다.

(c) $n < N$인 유한 개의 X_n도 덮이도록 B를 키울 수 있음을 1회 정리 6.3(b)로 보여라.

(d) 둘째 주장: $\{|Y_n W_n| \ge \varepsilon\} \subset \{|Y_n| > B\} \cup \{|W_n| \ge \varepsilon/B\}$를 보이고, $\limsup_n P(|Y_n W_n| \ge \varepsilon) \le \eta$에서 결론을 내라.

풀이 44쪽

3

중심극한정리

> **역사** de Moivre(1733)는 이항분포에서(5회 1절), Laplace(1810)는 독립인 오차의 일반적인 합에서 정규 극한을 얻었다. Laplace의 논증은 오늘의 특성함수와 같은 생각이었지만 엄밀하지 않았다. Poisson(1824)은 정리를 다듬으면서, 밀도가 $\frac{1}{\pi(1+x^2)}$인 오차는 아무리 평균해도 정규에 가까워지지 않음을 보였다. 엄밀한 증명은 러시아 학파가 했다. Chebyshev(1887)와 Markov(1898)는 적률로, Lyapunov(1901)는 특성함수로 증명했고, Lindeberg(1922)는 거의 필요충분인 조건을 찾았으며, Feller와 Lévy(1935)가 그 조건의 필요성을 보였다. "중심" 극한정리라는 이름은 Pólya(1920)가 확률론에서 **중심적인** 정리라는 뜻으로 붙였다.

3.1 mgf로 증명하기

정리 3.1 (중심극한정리) $X_1, X_2, \dots$ 가 iid이고 mgf가 어떤 $h > 0$에 대해 $|t| < h$에서 존재하며, 평균 μ, 분산 $\sigma^2 > 0$이면

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).$$

증명. **전략** 표준화한 합의 mgf의 로그를 $n\psi(t/\sqrt{n})$으로 쓰고, ψ를 0 근처에서 2차까지 Taylor 전개한다. 상수항과 1차 항은 0이고, 2차 항만 n을 곱해도 남는다. mgf의 극한이 정규의 mgf이면 2.3절의 상자로 끝난다.

① $Y_i = (X_i - \mu)/\sigma$로 두면 Y_i는 iid이고 평균 0, 분산 1이며, mgf $M_Y(s) = e^{-\mu s/\sigma} M_X(s/\sigma)$가 $|s| < \sigma h$에서 존재한다. 2회 정리 3.6

② $\psi = \log M_Y$는 $|s| < \sigma h$에서 무한 번 미분가능하고 $\psi(0) = 0$, $\psi'(0) = EY = 0$, $\psi''(0) = \operatorname{Var} Y = 1$이다. 2회 정리 3.3, 따름정리 3.4

③ $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$이므로

$$M_{Z_n}(t) = \left[M_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n, \qquad \log M_{Z_n}(t) = n\,\psi\left(\frac{t}{\sqrt{n}}\right).$$ 2회 정리 3.6, 3회 정리 3.4

④ $|t| < \sigma h$를 고정하고 $s = t/\sqrt{n}$으로 두면, Lagrange 나머지항이 있는 Taylor 정리로 0과 s 사이의 ξ_n이 있어

$$\psi(s) = \psi(0) + \psi'(0)s + \tfrac{1}{2}\psi''(\xi_n)s^2 = \tfrac{1}{2}\psi''(\xi_n)s^2.$$ [T8], ②

⑤ 따라서 $n\psi(t/\sqrt{n}) = \frac{t^2}{2}\psi''(\xi_n)$이다. $|\xi_n| \le |t|/\sqrt{n} \to 0$이고 ψ''는 0에서 연속이므로 이 값은 $\frac{t^2}{2}$로 간다. ②, [T5]

⑥ $M_{Z_n}(t) \to e^{t^2/2}$이 $|t| < \sigma h$의 모든 t에서 성립한다. 이것은 $\mathcal{N}(0,1)$의 mgf이므로 $Z_n \xrightarrow{D} \mathcal{N}(0,1)$이다. 2.3절의 상자, 5회 정리 1.8(ii)

조건 확인 ④에서 $s = t/\sqrt{n}$은 모든 $n \ge 1$에서 $|s| \le |t| < \sigma h$이므로 ψ가 정의된 구간 안에 있다. $\psi = \log M_Y$가 정의되려면 $M_Y > 0$이어야 하는데, $M_Y(s) = Ee^{sY}$는 양수의 기댓값이다. 2.3절의 상자는 mgf의 수렴을 0을 포함하는 열린 구간에서 요구하고, 여기서는 $(-\sigma h, \sigma h)$ 전체에서 성립한다. ∎

증명이 왜 **정규**인지를 말해 준다. ψ의 전개 $\psi(s) = \frac{s^2}{2} + \frac{\kappa_3}{6}s^3 + \frac{\kappa_4}{24}s^4 + \cdots$에서($\kappa_j$는 누율, κ_3은 왜도) 1차 항은 중심화로 사라지고, 3차 이상은 $n\psi(t/\sqrt{n})$에서

$$n\,\psi\!\left(\frac{t}{\sqrt{n}}\right) = \frac{t^2}{2} + \frac{\kappa_3 t^3}{6\sqrt{n}} + \frac{\kappa_4 t^4}{24n} + \cdots$$

처럼 $1/\sqrt{n}$, $1/n$의 속도로 사라진다(연습 3.1). 살아남는 것은 2차 항, 곧 분산뿐이고, 분산만으로 정해지는 분포가 정규다. 그리고 가장 늦게 사라지는 것은 왜도의 항이다. 6회 7절에서 치우친 모집단의 t-구간이 $1/\sqrt{n}$의 속도로만 나아진 이유가 이것이다.

따름정리 3.2 (정규근사의 표준 예) $n \to \infty$일 때

(a) $K_n \sim \mathrm{b}(n,p)$이면 $\frac{K_n - np}{\sqrt{npq}} \xrightarrow{D} \mathcal{N}(0,1)$ (de Moivre–Laplace, 5회 따름정리 1.3).

(b) $X_\lambda \sim \mathrm{Poi}(\lambda)$이면 $\lambda \to \infty$에서 $\frac{X_\lambda - \lambda}{\sqrt{\lambda}} \xrightarrow{D} \mathcal{N}(0,1)$.

(c) $V_r \sim \chi^2(r)$이면 $\frac{V_r - r}{\sqrt{2r}} \xrightarrow{D} \mathcal{N}(0,1)$.

증명. (a) K_n은 iid Bernoulli의 합. (c) V_r은 iid $\chi^2(1)$ r개의 합(4회 정리 3.9)이고 $\chi^2(1)$의 mgf는 $t < \frac{1}{2}$에서 존재한다. (b) 정수 $\lambda = m$이면 X_m은 iid Poi(1) m개의 합(4회 정리 2.4)이다. 정수가 아닌 λ는 mgf를 직접 쓴다: $\log M(t) = \lambda(e^{t/\sqrt{\lambda}} - 1) - \sqrt{\lambda}t = \frac{t^2}{2} + O(\lambda^{-1/2})$. ∎

예제 3.3 (근사는 얼마나 좋은가)

	정확한 값	정규근사	연속성 보정
$\mathrm{b}(100, \frac{1}{2})$, $P(K \le 45)$	0.1841	0.1587	0.1841
$\mathrm{Poi}(100)$, $P(X \ge 120)$	0.0282	0.0228	0.0256
$\chi^2(10)$, $P(V > r + 2\sqrt{2r})$	0.0410	0.0228	—
$\chi^2(100)$, $P(V > r + 2\sqrt{2r})$	0.0298	0.0228	—

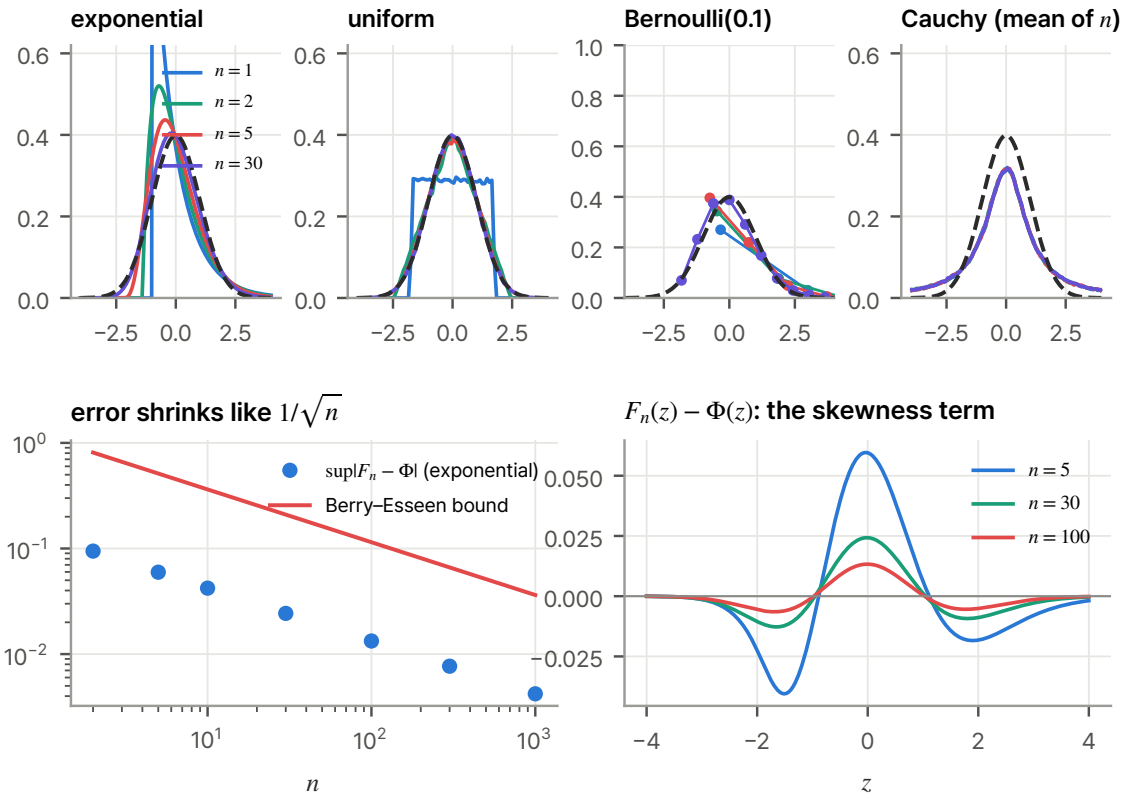


그림 3.1 표준화한 합 Z_n의 밀도. 위: 지수, 균등, Bernoulli(0.1), Cauchy를 $n = 1, 2, 5, 30$개 더했다(점선은 $\mathcal{N}(0,1)$). 대칭인 균등은 $n = 5$에서 이미 정규와 겹치고, 치우친 지수와 Bernoulli는 느리며, Cauchy는 n이 커져도 모양이 그대로다. 아래 왼쪽: 지수의 경우 $\sup_x |F_n(x) - \Phi(x)|$(점)와 Berry–Esseen 상한(선). 둘 다 $1/\sqrt{n}$의 기울기다. 아래 오른쪽: 오차 $F_n(z) - \Phi(z)$의 모양. 왜도 때문에 생기는 이 모양이 $1/\sqrt{n}$에 비례해 줄어든다.

> 대칭인 이항($p = \frac{1}{2}$)에서는 연속성 보정(정수 k를 구간 $[k - \frac{1}{2}, k + \frac{1}{2}]$로 보는 것)만 하면 넷째 자리까지 맞는다. 치우친 분포의 오른쪽 꼬리는 정규근사가 얕잡는다. $\chi^2(10)$에서 거의 두 배다. 8회의 χ^2 검정은 이 근사의 오차를 안고 쓴다.

3.2 더 넓은 정리와 속도

정리 3.1는 mgf를 가정했지만 결론에는 분산만 나온다. 실제로 가정은 분산까지 줄일 수 있고, 같은 분포라는 가정도 뺄 수 있다.

증명 없이 쓰는 사실 (분산만 있으면 — Lindeberg–Lévy, Lindeberg–Feller) (1) iid이고 $\sigma^2 < \infty$이면 mgf가 없어도 $Z_n \xrightarrow{D} \mathcal{N}(0,1)$이다. 증명은 mgf 대신 언제나 존재하는 특성함수 Ee^{itX}로 같은 테일러 전개를 한다(복소수가 필요하다). 이 노트는 이후 이 형태를 쓴다.
(2) 독립이지만 분포가 다르면, $s_n^2 = \sum_{i\le n}\sigma_i^2$에 대해 **Lyapunov 조건** $\frac{1}{s_n^3}\sum_{i\le n} E|X_i-\mu_i|^3 \to 0$이면 $\frac{\sum(X_i-\mu_i)}{s_n} \xrightarrow{D} \mathcal{N}(0,1)$이다(Lyapunov, 1901). 더 약한 Lindeberg 조건은 "어느 한 항도 전체 분산에서 무시할 수 없을 만큼 크지 않다"는 뜻이다. 작은 독립 효과가 많이 모이면 정규라는 직관의 정확한 형태다.

증명 없이 쓰는 사실 (Berry–Esseen 한계) iid이고 $\rho = E|X-\mu|^3 < \infty$이면 모든 n과 x에서

$$|P(Z_n \le x) - \Phi(x)| \le C\,\frac{\rho}{\sigma^3\sqrt{n}}, \qquad C \le 0.4748$$

이다(Berry 1941, Esseen 1942; 상수는 Shevtsova 2011). 정규근사의 오차는 $1/\sqrt{n}$의 속도로 줄고, 그 상수는 3차 절대적률이 정한다.

예제 3.4 (지수 합의 근사 오차) $\mathrm{Exp}(1)$이면 $\sigma = 1$이고 $\rho = E|X-1|^3 = \frac{12}{e} - 2 = 2.415$이다(연습 3.2). $\sum X_i \sim \Gamma(n,1)$이므로 오차를 정확히 잴 수 있다.

	$n=5$	$n=30$	$n=100$
실제 $\sup_x \lvert F_n(x)-\Phi(x)\rvert$	0.060	0.024	0.013
Berry–Esseen 상한	0.513	0.209	0.115

상한은 실제보다 9배쯤 크지만 속도 $1/\sqrt{n}$은 정확하다(그림 3.1 아래). $n=30$에서도 분포함수가 0.024만큼 어긋난다. 이 어긋남이 꼬리에 있으면 상대오차로는 크다(7절).

예제 3.5 (Poisson의 반례 — Cauchy 분포) X_i가 iid 표준 Cauchy(밀도 $\frac{1}{\pi(1+x^2)}$)이면 $\bar{X}_n$은 모든 n에서 다시 표준 Cauchy다(아래 상자). 평균을 내도 흩어짐이 줄지 않는다. $n=100$이어도 $P(|\bar{X}_{100}|>1) = \frac{1}{2}$로 관측 하나와 같다. 평균이 없으므로(2회 예제 1.8) Khinchin도 중심극한정리도 쓸 수 없다. Poisson은 1824년에 이것을 보였고, 분포의 이름은 30년 뒤 같은 사실을 다시 논한 Cauchy(1853)에게 갔다(9.2절).

증명 없이 쓰는 사실 (Cauchy의 평균은 Cauchy) $X_1,\dots,X_n$이 iid 표준 Cauchy이면 $\bar{X}_n$도 표준 Cauchy다. 특성함수 $Ee^{itX} = e^{-|t|}$로 한 줄에 증명된다: $\left(e^{-|t|/n}\right)^n = e^{-|t|}$. $n=2$는 합성곱(3회 정리 5.5)으로 직접 확인할 수 있다(연습 3.6).

연습문제

연습 3.1 (누율과 수렴 속도 ★★) $\psi(s) = \frac{s^2}{2} + \frac{\kappa_3}{6}s^3 + O(s^4)$이면 $n\psi(t/\sqrt{n}) - \frac{t^2}{2} = \frac{\kappa_3 t^3}{6\sqrt{n}} + O(\frac{1}{n})$임을 보여라. 지수분포의 κ_3을 $\psi(t) = -\log(1-t) - t$에서 구하라. 풀이 44쪽

연습 3.2 (지수분포의 3차 절대적률 ★) $X \sim \mathrm{Exp}(1)$이면 $E|X-1|^3 = \frac{12}{e} - 2$임을 보여라. (힌트: $\int_0^1 (1-x)^3 e^{-x}dx$와 $\int_1^\infty (x-1)^3 e^{-x}dx = \frac{6}{e}$로 나눈다.) 풀이 44쪽

연습 3.3 (확률이 다른 동전들 ★★) $X_i \sim \mathrm{Bernoulli}(p_i)$가 독립이고 $s_n^2 = \sum_{i\le n} p_i q_i \to \infty$이면 Lyapunov 조건이 성립함을 보여라. (힌트: $E|X_i - p_i|^3 \le p_i q_i$.) 따라서 확률이 제각각인 시행의 성공 수도 정규에 가깝다. 풀이 45쪽

연습 3.4 (균등 12개 ★) 옛 컴퓨터는 $U_1 + \cdots + U_{12} - 6$ ($U_i \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$)을 정규 난수로 썼다. 평균 0, 분산 1임을 보이고, 초과첨도가 $-1.2/12 = -0.1$임을 누율의 가법성으로 보여라. 이 난수의 결점은 무엇인가? 풀이 45쪽

연습 3.5 ($\chi^2(r)$의 정규 극한 ★★) $V_r \sim \chi^2(r)$의 mgf $(1-2t)^{-r/2}$으로 $\frac{V_r - r}{\sqrt{2r}}$의 mgf를 쓰고, $r \to \infty$에서 $e^{t^2/2}$로 감을 직접 보여라. (힌트: $-\frac{r}{2}\log(1-u) = \frac{r}{2}(u + \frac{u^2}{2} + O(u^3))$.) 풀이 45쪽

연습 3.6 (두 Cauchy의 평균 ★★★) X_1, X_2가 iid 표준 Cauchy이면 $X_1 + X_2$의 밀도가 $\frac{2}{\pi(4+z^2)}$임을 합성곱 $\int \frac{dx}{\pi^2(1+x^2)(1+(z-x)^2)}$으로 보이고, $\frac{X_1+X_2}{2}$가 다시 표준 Cauchy임을 결론지어라. (힌트: 부분분수 분해. 또는 노트북에서 수치적분으로 먼저 확인하라.) 풀이 45쪽

4

Slutsky의 정리와 델타 방법

중심극한정리의 Z_n에는 모르는 σ가 들어 있다. 실제로 쓰는 것은 σ 대신 S를 넣은 것이고, 관심 있는 양도 $\bar{X}$ 자체보다 그 함수(비율의 로그, π의 추정값)인 경우가 많다. 이 절의 두 정리가 그 간격을 메운다.

4.1 Slutsky의 정리 (1925)

보조정리 4.1 (확률적으로 무시할 수 있는 항) $Y_n \xrightarrow{D} Y$이고 $R_n \xrightarrow{P} 0$이면 $Y_n + R_n \xrightarrow{D} Y$.

증명. **전략** 정리 2.3(a)의 증명과 같은 끼우기다. $|R_n| < \varepsilon$인 사건 위에서 $Y_n + R_n$을 $Y_n \pm \varepsilon$ 사이에 둔다. 다른 점은 $x \pm \varepsilon$이 Y의 연속점이 되게 ε을 골라야 한다는 것이다. 단계별 증명은 연습 4.7(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 46쪽에 있다.

정리 4.2 (Slutsky의 정리) $X_n \xrightarrow{D} X$, $A_n \xrightarrow{P} a$, $B_n \xrightarrow{P} b$ (a, b는 상수)이면

$$A_n + B_n X_n \xrightarrow{D} a + bX.$$

증명. **전략** 확률변수인 계수 A_n, B_n을 극한의 상수 a, b로 바꾸고, 바꾸면서 생긴 차이 R_n이 확률적으로 0임을 보인다. 그러면 남는 것은 상수 계수의 아핀변환 $a + bX_n$이다.

① $A_n + B_n X_n = (a + bX_n) + R_n$, $R_n = (A_n - a) + (B_n - b)X_n$으로 쓴다. 더하고 빼기 [T4]

② $X_n \xrightarrow{D} X$이므로 $\{X_n\}$은 확률적으로 유계이고, $B_n - b \xrightarrow{P} 0$이므로 $(B_n - b)X_n \xrightarrow{P} 0$이다. 정리 2.5

③ $A_n - a \xrightarrow{P} 0$과 ②를 더하면 $R_n \xrightarrow{P} 0$이다. 따름정리 1.4

④ $b > 0$이면 $P(a + bX_n \le x) = F_n\left(\frac{x-a}{b}\right)$이고, x가 $a + bX$의 연속점인 것과 $\frac{x-a}{b}$가 X의 연속점인 것이 같으므로 $a + bX_n \xrightarrow{D} a + bX$이다. 정의 2.1

⑤ $b < 0$이면 $P(a + bX_n \le x) = 1 - P(X_n < y)$, $y = \frac{x-a}{b}$이다. y가 연속점이면 $P(X_n < y)$를 $F_n(y - \delta)$와 $F_n(y)$ 사이에 끼워 $F(y)$로 보내므로 같은 결론이 나온다. $b = 0$이면 $a + bX_n = a$는 상수다. 연습 2.4의 끼우기

⑥ ③④⑤에 보조정리 4.1을 쓰면 $A_n + B_n X_n \xrightarrow{D} a + bX$이다. 보조정리 4.1 ∎

> **Slutsky** 러시아의 경제학자이자 통계학자 Evgeny Slutsky는 1925년 *Metron*에 실은 논문에서 확률변수의 극한을 체계적으로 다뤘다. 2년 뒤(1927) 그는 무작위 충격의 이동평균이 경기순환처럼 보이는 주기를 만든다는 논문도 썼다(Slutsky 효과). 그의 정리는 Cramér의 교과서(1946)를 통해 통계학의 기본 도구가 되었다.

따름정리 4.3 (스튜던트화한 평균) $X_1, X_2, \dots$가 iid이고 $0 < \sigma^2 < \infty$이면

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).$$

따라서 $\bar{X}_n \pm z_{\alpha/2} S_n/\sqrt{n}$과 t-구간 $\bar{X}_n \pm t_{n-1,\alpha/2} S_n/\sqrt{n}$은 모집단의 분포와 상관없이 포함확률이 $1-\alpha$로 간다.

증명. $\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}} = \frac{\sigma}{S_n}\cdot Z_n$이다. 중심극한정리(3.3절의 상자 (1))로 $Z_n \xrightarrow{D} \mathcal{N}(0,1)$이고, 따름정리 1.7와 정리 1.3로 $\sigma/S_n \xrightarrow{P} 1$. Slutsky($a=0$, $B_n = \sigma/S_n$). t-구간은 $t_{n-1,\alpha/2} \to z_{\alpha/2}$(6회 정리 3.3(d))를 더한다. ∎

19세기의 천문학자들이 σ 자리에 s를 넣고 정규분포를 쓴 관행(6회 1절)이 이것으로 정당화된다. 단, "$n \to \infty$에서"다. 지수 모집단이면 1.96을 쓴 구간의 포함확률이 $n=30$에서 0.919, 200에서 0.945, 1000에서 0.948이다. 6회는 작은 n의 정확한 답을, 이 따름정리는 큰 n의 근사적 답을 준다.

4.2 델타 방법

정리 4.4 (델타 방법) $\sqrt{n}(Y_n - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$이고 g가 θ에서 미분가능하면

$$\sqrt{n}\big(g(Y_n) - g(\theta)\big) \xrightarrow{D} \mathcal{N}\big(0,\ g'(\theta)^2\sigma^2\big).$$

증명. **전략** 차분몫 $h(y) = \frac{g(y)-g(\theta)}{y-\theta}$을 확률변수 계수로 보면 $\sqrt{n}(g(Y_n)-g(\theta))$는 $h(Y_n)\cdot\sqrt{n}(Y_n-\theta)$이다. 계수가 $g'(\theta)$로 확률수렴함을 보이고 Slutsky를 쓴다.

① $h(y) = \frac{g(y)-g(\theta)}{y-\theta}$ $(y \neq \theta)$, $h(\theta) = g'(\theta)$로 두면 h는 θ에서 연속이다. 미분의 정의 [T5]

② $\sqrt{n}(Y_n - \theta)$는 확률적으로 유계이고 $\frac{1}{\sqrt{n}} \to 0$이므로 $Y_n - \theta = \frac{1}{\sqrt{n}} \cdot \sqrt{n}(Y_n - \theta) \xrightarrow{P} 0$이다. 정리 2.5

③ $h(Y_n) \xrightarrow{P} h(\theta) = g'(\theta)$이다. ①②, 정리 1.3(a)

④ 모든 y에서 $g(y) - g(\theta) = h(y)(y-\theta)$이므로(양변이 $y=\theta$에서 0)

$$\sqrt{n}\big(g(Y_n) - g(\theta)\big) = h(Y_n)\cdot\sqrt{n}(Y_n - \theta). \qquad ①$$

⑤ $A_n = 0$, $B_n = h(Y_n)$, $b = g'(\theta)$로 Slutsky를 쓰면 극한은 $g'(\theta)\cdot\mathcal{N}(0,\sigma^2) = \mathcal{N}(0, g'(\theta)^2\sigma^2)$이다. 정리 4.2, 5회 정리 3.8

■

이름은 $g(Y_n) - g(\theta) \approx g'(\theta)(Y_n-\theta)$, 곧 작은 변화 Δ의 선형 근사에서 왔다. 표준오차의 언어로 쓰면 $\mathrm{se}\big(g(Y)\big) \approx |g'(\theta)|\,\mathrm{se}(Y)$이다. $g'(\theta)=0$이면 극한이 0으로 퇴화하고 한 단계 더 전개해야 한다(연습 4.2).

예제 4.5 (로그 오즈) 성공 비율 $\hat p$에 대해 $\sqrt{n}(\hat p - p) \xrightarrow{D} \mathcal{N}(0, pq)$이고 $g(p) = \log\frac{p}{1-p}$이면 $g'(p) = \frac{1}{pq}$이므로

$$\sqrt{n}\Big(\log\frac{\hat p}{1-\hat p} - \log\frac{p}{1-p}\Big) \xrightarrow{D} \mathcal{N}\Big(0, \frac{1}{pq}\Big).$$

$p=0.2$, $n=50$이면 근사 분산은 $\frac{1}{npq} = 0.125$이고 시뮬레이션은 0.143이다($\hat p = 0$인 경우는 1.4×10^{-5}로 무시했다). 로그 오즈는 $(-\infty,\infty)$에서 움직이므로 $[0,1]$에 갇힌 $\hat p$보다 정규근사가 자연스럽고, 로지스틱 회귀(13회)와 오즈비의 구간(8회)이 이 척도를 쓴다.

예제 4.6 (Buffon의 π는 얼마나 정확한가) $\ell = d$이면 $p = \frac{2}{\pi}$, $\hat\pi_n = g(\hat p_n)$, $g(x) = \frac{2}{x}$이다. $g'(p) = -\frac{2}{p^2}$이므로

$$\sqrt{n}(\hat\pi_n - \pi) \xrightarrow{D} \mathcal{N}\Big(0,\ \frac{4}{p^4}\,p(1-p)\Big), \qquad \mathrm{sd}(\hat\pi_n) \approx \frac{2.37}{\sqrt{n}}.$$

$n=1000$이면 0.075(시뮬레이션 0.075)다. 소수 셋째 자리까지 맞히려면($\mathrm{sd}\approx 0.0005$) $n \approx (2.37/0.0005)^2 \approx 2.2\times10^7$번을 던져야 한다. $1/\sqrt{n}$의 정확도는 느리다.

가정 주의 이탈리아의 Lazzarini는 1901년 길이 2.5cm의 바늘을 3cm 간격의 선 위에 3,408번 던져 1,808번 걸쳤다고 보고했다. $\ell/d = \frac{5}{6}$이므로 $\hat\pi = \frac{2\cdot(5/6)\cdot 3408}{1808} = 3.1415929$, 오차 2.7×10^{-7}이다. 그런데 델타 방법의 표준편차는 0.051이다. 오차가 표준편차의 19만분의 1이다. 더 이상한 점이 있다. $\frac{355}{113} = 3.1415929$는 유명한 근삿값이고, $n = 3408 = 16\cdot213$, $K = 1808 = 16\cdot113$일 때만 정확히 나온다. 이 n에서 기대 교차 수는 1808.0이고 $P(K=1808) = 0.014$다. 이 노트의 연습 4.4은 "355/113이 나올 때까지 던지고 멈춘다"면 무슨 일이 생기는지 묻는다. Gridgeman(1960)과 Badger(1994)는 이 실험이 원하는 답에 맞춰 멈췄거나 꾸며졌다고 결론지었다. 좋은 추정은 오차의 크기와 함께 보고되고, 오차보다

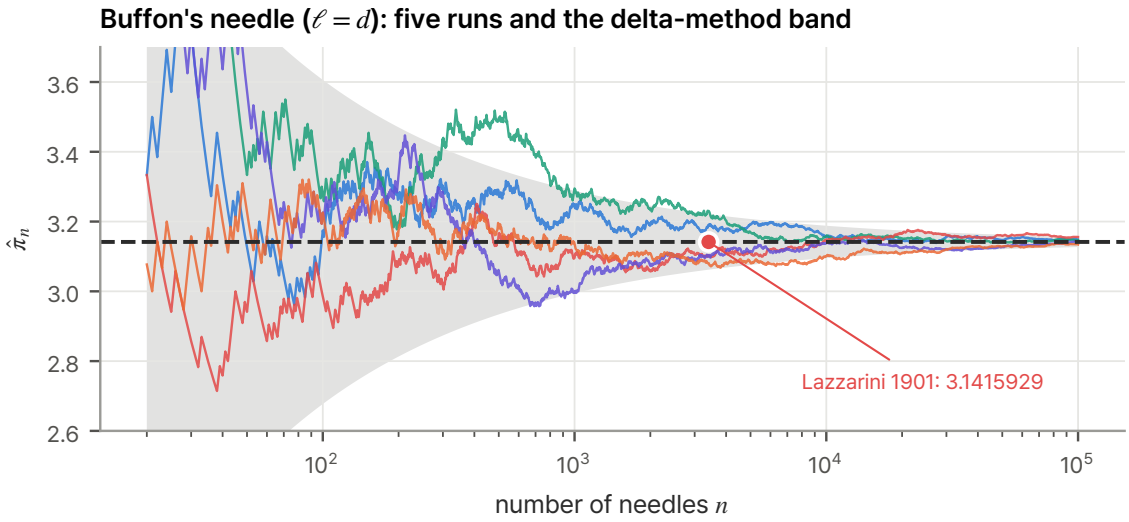


그림 4.1 바늘 던지기 다섯 번의 실행에서 $\hat{\pi}_n$의 경로(색 선)와 델타 방법의 $\pi \pm 1.96 \cdot 2.37/\sqrt{n}$ 띠(회색). 경로는 띠 안에서 $1/\sqrt{n}$의 속도로 좁아진다. 빨간 점은 Lazzarini(1901)가 보고한 $n = 3408$의 값으로, 띠의 폭에 비해 믿을 수 없을 만큼 π에 가깝다.

> 지나치게 좋은 결과는 의심해야 한다.

예제 4.7 (분산 안정화) $\hat{p}$의 분산 $\frac{pq}{n}$은 p에 의존한다. g를 $g'(p)^2 pq$가 상수가 되게 고르면 $g(\hat{p})$의 분산은 p와 상관없다. $g'(p) = \frac{1}{2\sqrt{pq}}$를 적분하면 $g(p) = \arcsin\sqrt{p}$이고, $\operatorname{Var} \arcsin\sqrt{\hat{p}} \approx \frac{1}{4n}$이다. $n = 100$에서 $4n \operatorname{Var}$를 시뮬레이션하면 $p = 0.05,\ 0.2,\ 0.5$에서 $1.14,\ 1.02,\ 1.01$이다. 같은 방법으로 Poisson은 $\sqrt{X}$의 분산이 $\frac{1}{4}$로 안정된다($\lambda = 2, 10, 100$에서 $0.39,\ 0.26,\ 0.25$). 분산이 평균에 따라 변하는 자료를 회귀(13회)나 분산분석(14회)에 넣기 전의 표준적인 변환이다.

연습문제

연습 4.1 (로그 평균 ★) $X_i \overset{\text{iid}}{\sim} \text{Exp}(1)$이면 $\sqrt{n}\log \bar{X}_n \xrightarrow{D} \mathcal{N}(0, 1)$임을 보여라. 풀이 45쪽

연습 4.2 (2차 델타 방법 ★★) $\sqrt{n}(Y_n - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$이고 $g'(\theta) = 0$, g''가 θ 근처에서 연속이면 $n\big(g(Y_n) - g(\theta)\big) \xrightarrow{D} \frac{1}{2} g''(\theta)\sigma^2 \chi^2(1)$임을 보여라(연습 2.4 사용). 예: $\mu = 0$이면 $n\bar{X}_n^2 \xrightarrow{D} \sigma^2 \chi^2(1)$. 풀이 45쪽

연습 4.3 (가장 좋은 바늘 ★★) $\ell \le d$에서 $r = \ell/d$, $p = \frac{2r}{\pi}$이면 $\hat{\pi}_n = \frac{2r}{\hat{p}_n}$의 점근분산이 $\frac{4r^2(1-p)}{p^3} = \frac{\pi^2(1-p)}{p}$임을 보이고, $r = 1$에서 가장 작음을 보여라. Lazzarini의 $r = \frac{5}{6}$은 $r = 1$보다 표준편차가 몇 배인가? 풀이 45쪽

연습 4.4 (원하는 답이 나올 때 멈추기 ★★) $r = \frac{5}{6}$으로 한 번씩 던지며 n이 213의 배수이고 $K_n/n = \frac{113}{213}$인 순간 멈춘다고 하자. 노트북으로 $n \le 5000$ 안에 멈출 확률을 시뮬레이션하라. 보고된 $\hat{\pi}$는 언제나 355/113이다. 이 실험의 오차를 0.051이라고 보고하는 것은 왜 틀렸나?

풀이 46쪽

연습 4.5 (두 변환 ★★) (a) $g(p) = \arcsin\sqrt{p}$이면 $g'(p)^2 pq = \frac{1}{4}$임을 보여라. (b) $X_\lambda \sim \mathrm{Poi}(\lambda)$이면 $\sqrt{\lambda}\cdot\frac{X_\lambda - \lambda}{\lambda}$의 극한과 델타 방법으로 $2(\sqrt{X_\lambda} - \sqrt{\lambda}) \xrightarrow{D} \mathcal{N}(0,1)$임을 보여라. 풀이 46쪽

연습 4.6 (Slutsky의 함정 ★) $X_n \xrightarrow{D} \mathcal{N}(0,1)$, $B_n \xrightarrow{D} 1$이면 $B_n X_n \xrightarrow{D} \mathcal{N}(0,1)$인가? B_n이 상수로 **분포**수렴하면 확률수렴이기도 함(정리 2.3(b))을 써서 답하라. 풀이 46쪽

연습 4.7 (무시할 수 있는 항 증명 ★★ 유도형) 보조정리 4.1을 다음 순서로 증명하라. F_Y를 Y의 분포함수, x를 그 연속점이라 하자.

(a) $\{Y_n + R_n \le x\} \subset \{Y_n \le x + \varepsilon\} \cup \{|R_n| \ge \varepsilon\}$와 $\{Y_n \le x - \varepsilon\} \subset \{Y_n + R_n \le x\} \cup \{|R_n| \ge \varepsilon\}$를 원소 추적으로 보여라.

(b) F_Y의 불연속점은 셀 수 있게 많으므로 $x \pm \varepsilon$이 모두 연속점인 ε을 얼마든지 작게 고를 수 있음을 설명하라.

(c) (a)(b)로 $F_Y(x-\varepsilon) \le \liminf_n P(Y_n + R_n \le x) \le \limsup_n P(Y_n + R_n \le x) \le F_Y(x+\varepsilon)$을 보여라.

(d) $\varepsilon \downarrow 0$으로 결론을 내라. 정리 2.3(a)의 증명과 무엇이 같고 무엇이 다른가?

풀이 46쪽

5

다변량으로 — 선형결합 하나면 된다

평균 하나가 아니라 여러 개를 함께 추정할 때가 많다. 다항분포의 범주별 비율, $(\bar{X}, \frac{1}{n}\sum X_i^2)$, 회귀계수의 벡터(13회). 벡터의 극한도 5회의 Cramér–Wold와 같은 생각, 곧 "모든 선형결합을 보면 된다"로 해결된다.

5.1 벡터의 수렴과 다변량 중심극한정리

정의 5.1 (벡터의 수렴) $\mathbb{R}^k$의 확률벡터 $\mathbf{X}_n$이 모든 성분에서 $\mathbf{X}$의 성분으로 확률수렴하면 $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$라 쓴다. 결합분포함수가 $\mathbf{X}$의 결합분포함수의 모든 연속점에서 수렴하면 $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$라 쓴다.

정리 1.3와 Slutsky의 정리는 벡터에서도 같은 증명으로 성립한다. 2.3절의 상자도 벡터의 mgf $M_{\mathbf{X}}(\mathbf{t}) = Ee^{\mathbf{t}'\mathbf{X}}$에 대해 성립한다.

정리 5.2 (다변량 중심극한정리) $\mathbf{X}_1, \mathbf{X}_2, \dots$가 $\mathbb{R}^k$의 iid 확률벡터이고 mgf가 $\mathbf{0}$의 근방에서 존재하며, 평균 $\boldsymbol{\mu}$, 공분산행렬 Σ이면

$$\sqrt{n}\,(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} \mathcal{N}_k(\mathbf{0}, \Sigma).$$

증명. **전략** 방향 $\mathbf{t}$를 하나 고정하면 벡터의 mgf는 1차원 변수 $Y_i = \mathbf{t}'(\mathbf{X}_i - \boldsymbol{\mu})$의 합의 mgf다. 정리 3.1의 증명 ④⑤를 $t = 1$에서 그대로 하고, 모든 $\mathbf{t}$에서 극한이 $\mathcal{N}_k(\mathbf{0}, \Sigma)$의 mgf임을 확인한다. 단계별 증명은 연습 5.5(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 47쪽에 있다.

증명은 1차원의 중심극한정리를 모든 방향 $\mathbf{t}$에 한 번씩 쓴 것이다. 벡터의 극한이 결합정규인 것은 5회 정리 3.10에서처럼 **모든 선형결합이 정규**이기 때문이다.

예제 5.3 (다항분포의 정규 극한) n번의 독립 시행에서 결과가 범주 j일 확률이 p_j $(j = 1, \dots, k)$이면, 시행 i의 결과를 원핫 벡터 $\mathbf{X}_i$(범주 j이면 j번째 성분만 1)로 쓸 수 있다. $E\mathbf{X}_i = \mathbf{p}$이고, $X_{ij}X_{il} = 0$ $(j \neq l)$, $X_{ij}^2 = X_{ij}$이므로

$$\operatorname{Cov}(\mathbf{X}_i) = \operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}' =: \Sigma_{\mathbf{p}}.$$

범주별 개수 $\mathbf{N} = \sum \mathbf{X}_i$는 다항분포이고(4회 정리 1.8의 공분산 $-np_jp_l$과 같다) 정리 5.2로

$$\sqrt{n}\,(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{D} \mathcal{N}_k(\mathbf{0},\ \Sigma_{\mathbf{p}}).$$

$\Sigma_{\mathbf{p}}\mathbf{1} = \mathbf{p} - \mathbf{p} = \mathbf{0}$이므로 극한은 특이 정규분포다(5회 명제 5.6). 비율의 합이 언제나 1이라는 제약 하나가 차원 하나를 없앤다(그림 5.1).

이 예제가 8회의 출발점이다. Pearson(1900)의 적합도 통계량 $X^2 = \sum_j \frac{(N_j - np_j)^2}{np_j}$은 $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p})$의 이차형식이고, 5회 정리 3.17(이차형식은 χ^2)과 정리 5.2를 합치면 $X^2 \xrightarrow{D} \chi^2(k-1)$이 나온다. 자유도 $k-1$은 위의 제약 하나 때문이다.

5.2 다변량 델타 방법

정리 5.4 (다변량 델타 방법) $\sqrt{n}(\mathbf{Y}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}_k(\mathbf{0}, \Sigma)$이고 $g : \mathbb{R}^k \to \mathbb{R}$가 $\boldsymbol{\theta}$에서 미분가능하며 기울기가 ∇g이면

$$\sqrt{n}\big(g(\mathbf{Y}_n) - g(\boldsymbol{\theta})\big) \xrightarrow{D} \mathcal{N}\big(0,\ \nabla g'\,\Sigma\,\nabla g\big).$$

증명. **전략** 미분가능성을 "1차 근사 + 나머지"로 쓴다. 1차 항은 정규벡터의 선형결합이고, 나머지는 확률적으로 유계인 양과 확률적으로 0인 양의 곱이다. 보조정리 4.1로 합친다. 단계별 증명은 연습 5.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 47쪽에 있다.

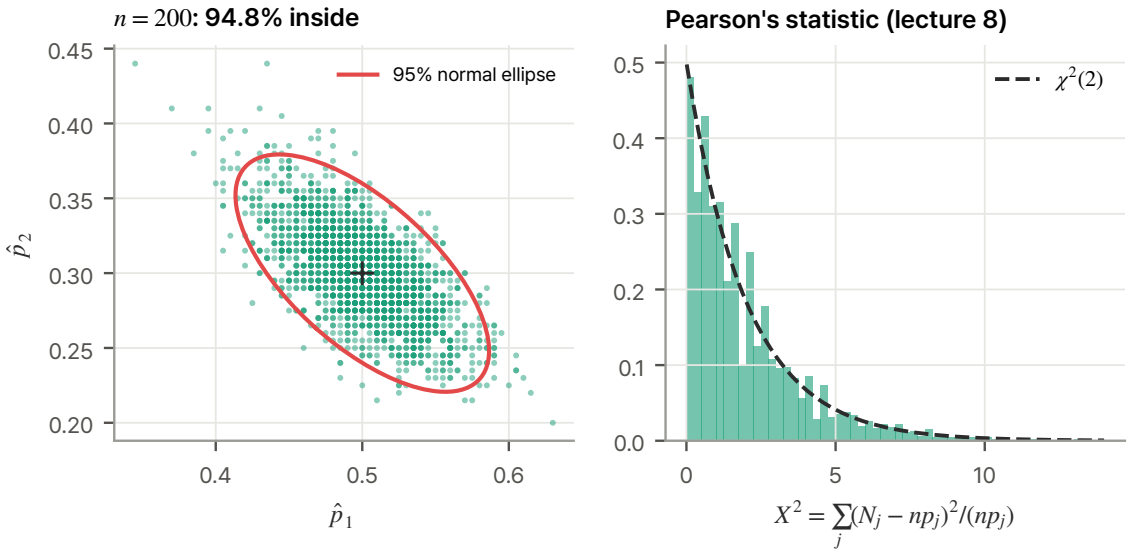


그림 5.1 왼쪽: 세 범주($\mathbf{p} = (0.5, 0.3, 0.2)$, $n = 200$)의 $(\hat{p}_1, \hat{p}_2)$ 2,000개와 정규 극한의 95% 타원. 오른쪽: Pearson 통계량 $\sum_j \frac{(N_j - np_j)^2}{np_j}$의 히스토그램과 $\chi^2(2)$ 밀도. 8회에서 증명한다.

예제 5.5 (표본분산의 점근분포) $\mathbf{Y}_n = \left(\bar{X}_n,\ \frac{1}{n}\sum X_i^2\right)$, $g(a,b) = b - a^2$이면 $g(\mathbf{Y}_n) = \frac{n-1}{n}S_n^2$이다. $\mu = 0$으로 두어도 되고(위치를 옮겨도 S^2는 그대로), 그러면 $\theta = (0, \sigma^2)$, $\nabla g = (-2a, 1) = (0, 1)$이므로 둘째 성분의 분산만 남는다:

$$\sqrt{n}\,(S_n^2 - \sigma^2) \xrightarrow{D} \mathcal{N}\big(0,\ \mathrm{Var}\,X^2\big) = \mathcal{N}\big(0,\ \mu_4 - \sigma^4\big).$$

($\frac{n-1}{n}$과의 차이는 Slutsky로 사라진다.) 정규이면 $\mu_4 - \sigma^4 = 2\sigma^4$으로 6회 따름정리 2.3과 맞고, 지수이면 $9 - 1 = 8$이다(시뮬레이션 7.98). 6회 7.4절에서 분산의 F 검정이 첨도에 민감했던 이유가 이 분산의 μ_4다.

연습문제

연습 5.1 (원핫 벡터의 공분산 ★) 예제 5.3의 $\mathrm{Cov}(\mathbf{X}_i) = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}'$를 성분별로 확인하고, $\Sigma_{\mathbf{p}}$의 계수가 $k-1$임을 보여라. (힌트: $\mathbf{1}$에 수직인 벡터 $\mathbf{v}$에 대해 $\mathbf{v}'\Sigma_{\mathbf{p}}\mathbf{v} = \sum p_j v_j^2 - (\sum p_j v_j)^2$.)

풀이 46쪽

연습 5.2 (비율 추정량 ★★) (X_i, Y_i)가 iid이고 $\mu_X \neq 0$이면 $R = \mu_Y/\mu_X$의 추정량 $\hat{R} = \bar{Y}/\bar{X}$에 대해 $\sqrt{n}(\hat{R} - R) \xrightarrow{D} \mathcal{N}\Big(0,\ \dfrac{\sigma_Y^2 - 2R\sigma_{XY} + R^2\sigma_X^2}{\mu_X^2}\Big)$임을 보여라.

풀이 47쪽

연습 5.3 (Fisher의 z ★) 이변량 정규 표본의 상관계수 r이 $\sqrt{n}(r-\rho) \xrightarrow{D} \mathcal{N}\big(0, (1-\rho^2)^2\big)$을 만족함을 받아들이자(13회). $z = \frac{1}{2}\log\frac{1+r}{1-r}$의 점근분산이 ρ와 상관없이 1임을 보여라(Fisher, 1921). 분산 안정화 변환이다.

풀이 47쪽

연습 5.4 (변동계수 ★★) $\mu > 0$이고 $EX^4 < \infty$인 iid 표본의 변동계수 $S_n/\bar{X}_n$의 점근분산을 다변량 델타 방법으로 구하라. 정규 모집단이면 $\frac{c^2}{2} + c^4$ $(c = \sigma/\mu)$임을 보여라.

풀이 47쪽

연습 5.5 (다변량 중심극한정리 증명 ★★ 유도형) 정리 5.2를 다음 순서로 증명하라.

(a) $\mathbf{t}$를 고정하고 $Y_i = \mathbf{t}'(\mathbf{X}_i - \boldsymbol{\mu})$로 두면 Y_i는 iid, 평균 0, 분산 $\mathbf{t}'\Sigma\mathbf{t}$임을 보여라(5회 보조정리 3.1).

(b) $M_{\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu})}(\mathbf{t}) = \big[M_Y(1/\sqrt{n})\big]^n$임을 보여라.

(c) $\mathbf{t}'\Sigma\mathbf{t} > 0$이면 정리 3.1의 증명 ④⑤를 $t = 1$에서 따라 (b)의 로그가 $\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}$로 감을 보여라. $\mathbf{t}'\Sigma\mathbf{t} = 0$이면 왜 처음부터 성립하는가?

(d) 극한이 $\mathcal{N}_k(\mathbf{0}, \Sigma)$의 mgf(5회 정의 3.4)임을 확인하고, 2.3절의 상자(벡터의 mgf)로 결론을 내라.

풀이 47쪽

연습 5.6 (다변량 델타 방법 증명 ★★ 유도형) 정리 5.4를 다음 순서로 증명하라.

(a) g가 $\boldsymbol{\theta}$에서 미분가능하다는 것을 $g(\mathbf{y}) - g(\boldsymbol{\theta}) = \nabla g'(\mathbf{y}-\boldsymbol{\theta}) + \|\mathbf{y}-\boldsymbol{\theta}\| r(\mathbf{y})$, $r(\boldsymbol{\theta})=0$, r은 $\boldsymbol{\theta}$에서 연속인 꼴로 써라.

(b) $\nabla g' \sqrt{n}(\mathbf{Y}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(0, \nabla g' \Sigma \nabla g)$임을 2.3절의 연속사상 상자(선형함수는 연속)와 5회 정리 3.8로 보여라.

(c) $\sqrt{n}\|\mathbf{Y}_n - \boldsymbol{\theta}\|$가 확률적으로 유계이고 $r(\mathbf{Y}_n) \xrightarrow{P} 0$임을 보여 둘째 항이 $\xrightarrow{P} 0$임을 보여라.

(d) 보조정리 4.1로 결론을 내라.

풀이 47쪽

6

계산으로 — Monte Carlo (1946 → 1949)

> **역사** 1946년 수학자 Stanisław Ulam은 병상에서 카드 놀이(Canfield 솔리테어)가 이길 확률을 계산하다가, 조합론으로 풀기보다 백 번 해 보고 세는 편이 빠르다는 것을 깨달았다. 그는 이 생각을 Los Alamos의 John von Neumann에게 말했고, 둘은 핵분열 장치 안의 중성자 확산을 난수로 흉내 내는 계획을 세웠다. 1948년 ENIAC이 그 계산을 했다. 동료 Nicholas Metropolis가 Ulam의 삼촌이 드나들던 도박장의 이름을 따 이 방법을 Monte Carlo라 불렀고, Metropolis와 Ulam(1949)이 *JASA*에 같은 이름의 논문을 냈다. 방법 자체는 Buffon의 바늘(1절)과 Student의 카드(6회)부터 있었다. 새로운 것은 컴퓨터, 그리고 결과의 오차를 함께 말하는 습관이었다.

6.1 오차를 아는 시뮬레이션

정리 6.1 (Monte Carlo 추정) $\theta = E\,h(X)$를 구하려고 X의 분포에서 iid 표본 $X_1, \dots, X_n$을 만들어 $\hat{\theta}_n = \frac{1}{n}\sum h(X_i)$로 둔다.

(a) $E|h(X)| < \infty$이면 $E\hat{\theta}_n = \theta$이고 $\hat{\theta}_n \xrightarrow{P} \theta$.

(b) $\sigma_h^2 = \mathrm{Var}\,h(X) < \infty$이고 s_h^2이 $h(X_i)$들의 표본분산이면, $\hat{\theta}_n \pm z_{\alpha/2}\, s_h/\sqrt{n}$의 포함확률은 $1-\alpha$로 간다.

증명. (a) 선형성과 정리 1.6. (b) 따름정리 4.3를 iid 변수 $h(X_i)$에 쓴다. ∎

정리는 Monte Carlo의 두 가지 성질을 말한다. 오차는 $\sigma_h/\sqrt{n}$으로 **차원과 상관없고**, 오차의 크기를 **같은 표본으로 어림할 수 있다**. 정확도를 한 자리 높이려면 n을 100배 해야 하므로, 실제의 기술은 σ_h를 줄이는 데 있다(6.3절).

예제 6.2 (Chebyshev 대신 중심극한정리) 2회 예제 6.4에서 $\int_0^1 e^{x^2}dx = 1.4627$을 균등 난수로 구할 때, 오차가 0.01 이상일 확률을 1% 이하로 하려면 Chebyshev로 $n \ge \frac{\sigma_h^2}{0.01^2 \cdot 0.01} = 2.25 \times 10^5$이면 충분했다($\sigma_h^2 = 0.2251$). 중심극한정리로는 $P(|\hat{\theta}_n - \theta| \ge 0.01) \approx P\big(|Z| \ge \frac{0.01\sqrt{n}}{\sigma_h}\big) = 0.01$에서

$$n \approx \left(\frac{z_{0.005}\,\sigma_h}{0.01}\right)^2 = \left(\frac{2.576 \times 0.474}{0.01}\right)^2 \approx 1.49 \times 10^4.$$

15분의 1이다. Chebyshev는 **모든** 분포에 맞는 보장이라 보수적이고, 중심극한정리는

이 분포에 맞는 근사다. 6회의 카드 실험에서 섞기 400번의 비율이 표준편차 $0.0077 \approx \sqrt{0.05 \cdot 0.95/750}$로 흩어진 것도 이 정리의 예다.

6.2 원하는 분포에서 뽑기 — 기각 표집

$h(X)$의 평균을 구하려면 X의 분포에서 뽑을 수 있어야 한다. 분포함수의 역함수를 알면 역함수 표집(2회 정리 4.7)을 쓰고, 정규처럼 역함수가 닫힌 꼴이 아니면 Box–Muller(3회 정리 5.4) 같은 변환을 찾는다. 밀도만 알 때 쓰는 일반적인 방법이 von Neumann(1951)의 기각 표집이다.

알고리즘 6.3 (기각 표집) 목표 밀도 f, 뽑기 쉬운 제안 밀도 g, 상수 M이 모든 x에서 $f(x) \le Mg(x)$를 만족한다.

1. $Y \sim g$와 $U \sim \mathrm{U}(0,1)$을 독립으로 뽑는다.
2. $U \le \dfrac{f(Y)}{Mg(Y)}$이면 Y를 받아들여 내놓고, 아니면 1로 돌아간다.

정리 6.4 (기각 표집의 정당성) 알고리즘 6.3이 내놓는 값의 밀도는 f이다. 한 번 시도에서 받아들일 확률은 $1/M$이고, 시도 횟수는 평균 M인 기하분포를 따른다.

증명. **전략** 한 번의 시도에서 "받아들이고 $Y \le x$"인 확률을 이중적분으로 계산하면 $F(x)/M$이 된다. 시도들이 독립이므로 내놓는 값의 분포는 받아들인 시도의 조건부분포이고, 시도 횟수는 기하분포다.

① 한 번의 시도에서 (Y,U)의 결합밀도는 $g(y)\cdot 1$ $(0<u<1)$이다. — 독립, 3회 정리 3.2

② 받아들이는 사건을 $A=\{U \le f(Y)/(Mg(Y))\}$라 하면

$$P(Y \le x,\ A) = \int_{-\infty}^{x} g(y)\,P\Big(U \le \frac{f(y)}{Mg(y)}\Big)dy \qquad \text{① , 반복적분 [T6]}$$

$$= \int_{-\infty}^{x} g(y)\frac{f(y)}{Mg(y)}\,dy = \frac{F(x)}{M}. \qquad 0 \le f/(Mg) \le 1$$

③ $x \to \infty$로 보내면 $P(A) = \dfrac{1}{M}$이다. — 1회 정리 4.7

④ k번째 시도에서 처음 받아들이고 그 값이 x 이하일 확률은 $(1-\frac{1}{M})^{k-1}\frac{F(x)}{M}$이다. — 시도들의 독립, ②③

⑤ k에 대해 더하면 내놓는 값이 x 이하일 확률은 $\frac{F(x)}{M}\sum_{k\ge 1}(1-\frac{1}{M})^{k-1} = F(x)$이다. — 가산가법성, 기하급수 [T5]

⑥ ④에서 $F(x)$를 1로 바꾸면 처음 받아들일 때까지의 시도 횟수가 성공 확률 $1/M$의 기하분

포이고 평균이 M임을 얻는다. 4회 1절

조건 확인 $g(y)=0$인 곳에서는 $f(y)\le Mg(y)=0$이므로 비 $f/(Mg)$를 정의하지 않아도 그 y는 적분에 기여하지 않는다. ②의 반복적분은 음이 아닌 함수의 적분이므로 Tonelli로 순서를 정할 수 있다 [T6]. $M\ge 1$인 것은 $1=\int f\le M\int g=M$에서 나온다. ■

예제 6.5 (지수 난수로 정규 난수 만들기) 반정규 밀도 $f(x)=\sqrt{2/\pi}\,e^{-x^2/2}$ $(x>0)$를 $g(x)=e^{-x}$로 덮는다. $\frac{f}{g}=\sqrt{2/\pi}\,e^{x-x^2/2}$은 $x=1$에서 최대이므로 $M=\sqrt{2e/\pi}=1.3155$이고, 받아들일 확률은 0.760이다. 받아들인 값에 무작위 부호를 붙이면 $\mathcal{N}(0,1)$이다. 지수 난수는 $-\log U$로 만든다(2회 정리 4.7). 받아들이는 조건은 $U\le e^{-(Y-1)^2/2}$로 간단해진다(연습 6.2).

6.3 오차를 줄이기 — 중요도 표집과 반대 변량

정리 6.6 (중요도 표집) $h(x)f(x)\neq 0$인 곳에서 $g(x)>0$이면, $Y_i\overset{\text{iid}}{\sim}g$에 대해

$$\hat\theta_n^{\mathrm{IS}}=\frac{1}{n}\sum_{i=1}^{n}h(Y_i)\,w(Y_i),\qquad w=\frac{f}{g}$$

는 $\theta=E_f h(X)$의 불편추정량이고 분산은 $\frac{1}{n}\big(E_g[h(Y)^2w(Y)^2]-\theta^2\big)$이다. $h\ge 0$이면 $g\propto hf$일 때 분산이 0이다.

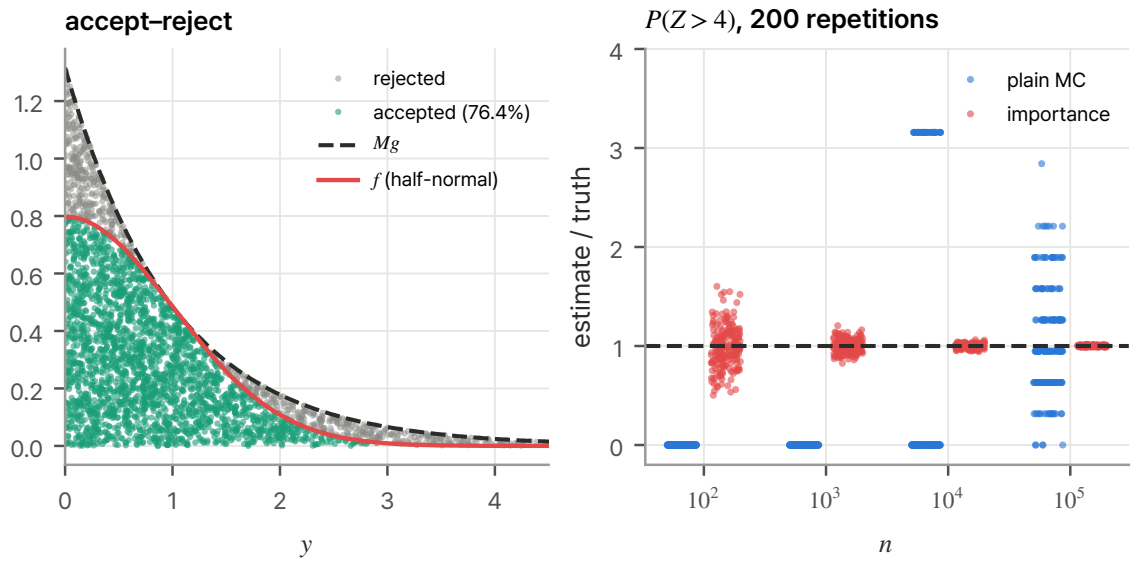


그림 6.1 왼쪽: 기각 표집. $(Y,UMg(Y))$를 Mg 아래에 고르게 뿌리고(회색) f 아래의 점만(청록) 남긴다. 남은 점의 가로 좌표가 f를 따른다. 오른쪽: $P(Z>4)$의 추정. 보통의 Monte Carlo(파랑)는 $n=10^4$에서 거의 언제나 0을 내놓고, 중요도 표집(빨강)은 같은 n에서 상대오차 2%로 맞힌다.

증명. **전략** 가중치 $w = f/g$가 g로 뽑은 표본을 f의 기댓값으로 되돌린다는 것을 적분 한 줄로 보이고, 분산은 iid 평균의 분산 공식으로 계산한다. 단계별 증명은 연습 6.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 48쪽에 있다.

분산이 0인 g는 θ를 알아야 만들 수 있으므로 실제로는 쓸 수 없다. 그러나 방향은 알려 준다. "hf가 큰 곳에서 많이 뽑고, 가중치로 바로잡아라."

예제 6.7 (드문 사건) $\theta = P(Z > 4) = 3.17 \times 10^{-5}$를 보통의 Monte Carlo로 구하면 $h = 1\{Z > 4\}$이고 상대 표준오차는 $\sqrt{(1-\theta)/(n\theta)}$다. $n = 10^4$이면 1.78, 곧 추정값이 대개 0이다. 상대오차를 10%로 하려면 $n \approx 3 \times 10^6$이 필요하다. 제안분포를 $g = \mathcal{N}(4, 1)$로 옮기면 $w(y) = \phi(y)/\phi(y-4) = e^{-4y+8}$이고, 한 표본의 분산이 4.5×10^{-9}로 보통 Monte Carlo의 $\theta(1-\theta) = 3.2 \times 10^{-5}$보다 약 7,000배 작다. $n = 10^4$에서 상대오차가 2%다. 금융의 극단 손실, 통신의 오류율, 희귀 변이의 검정처럼 작은 확률을 다루는 계산이 모두 이 방법을 쓴다.

예제 6.8 (반대 변량) U와 $1-U$는 같은 분포이고, h가 단조이면 $h(U)$와 $h(1-U)$는 음의 상관을 갖는다. 그래서 $\frac{1}{2}\big(h(U) + h(1-U)\big)$를 쓰면 같은 계산량에서 분산이 줄어든다. $h(x) = e^{x^2}$이면 난수 2개의 평균의 분산 $\sigma_h^2/2 = 0.1126$이 반대 변량에서는 0.0280으로 4.0배 작다(연습 6.5).

연습문제

연습 6.1 (원으로 잰 π ★) $U_1, U_2 \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$이면 $4 \cdot 1\{U_1^2 + U_2^2 \le 1\}$의 평균은 π이다. 분산을 구하고, $n = 10^4$에서 표준오차가 0.016임을 보여라. Buffon의 바늘($n = 10^4$에서 0.024)보다 나은가? 풀이 48쪽

연습 6.2 (받아들이는 조건 ★) 예제 6.5에서 $\dfrac{f(y)}{Mg(y)} = e^{-(y-1)^2/2}$임을 보여라. 풀이 48쪽

연습 6.3 (베타 난수 ★★) $\mathrm{Beta}(2.7, 6.3)$을 $g = \mathrm{U}(0,1)$로 기각 표집할 때 가장 작은 M은 밀도의 최댓값이다. 최빈값 $\frac{a-1}{a+b-2}$에서 $M = 2.67$, 받아들일 확률 0.375임을 확인하라. 4회 정리 4.2의 감마의 비를 쓰는 방법과 비교하면? 풀이 48쪽

연습 6.4 (두꺼운 꼬리가 필요하다 ★★) $f = \mathcal{N}(0,1)$, $h(x) = x^2$을 제안분포 $g = \mathcal{N}(0, \tau^2)$로 중요도 표집하면 $\tau^2 \le \frac{1}{2}$일 때 분산이 무한임을 보여라. 제안분포의 꼬리가 목표보다 얇으면 가중치 w가 폭발한다. 풀이 48쪽

연습 6.5 (반대 변량의 분산 ★★★) $\mathrm{Var}\,\frac{h(U)+h(1-U)}{2} = \frac{1}{2}\big(\sigma_h^2 + \mathrm{Cov}(h(U), h(1-U))\big)$임을 보이고, h가 증가함수이면 $\mathrm{Cov}(h(U), h(1-U)) \le 0$임을 보여라. (힌트: 독립인 U, V에 대해 $\big(h(U) - h(V)\big)\big(h(1-U) - h(1-V)\big) \ge 0$.) 풀이 48쪽

연습 6.6 (중요도 표집 증명 ★ 유도형) 정리 6.6를 다음 순서로 증명하라.

(a) $E_g[h(Y)w(Y)] = \int h\frac{f}{g}\,g = \int hf = \theta$를 보여라. $g(x) = 0$이고 $h(x)f(x) \neq 0$인 점이 있으면 어느 등호가 틀리는가?

(b) $\mathrm{Var}_g\big(h(Y)w(Y)\big) = E_g[h^2w^2] - \theta^2$을 써서 $\hat\theta_n^{\mathrm{IS}}$의 분산을 구하라(2회 정리 2.2, 2.8).

(c) $h \geq 0$, $\theta > 0$이고 $g = hf/\theta$이면 $h(Y)w(Y) = \theta$가 상수임을 보여 분산이 0임을 확인하라.

(d) (c)의 g가 실제로 쓸 수 없는 이유를 설명하라.

풀이 48쪽

7

부쉬보기 — 분산이 무한하고, 꼬리를 물을 때

중심극한정리의 가정은 단 하나, 유한한 분산이다. 이 가정을 깨면 무엇이 남고 무엇이 무너지는가? 그리고 가정이 성립해도 정리가 답하지 않는 질문은 무엇인가?

7.1 평균은 있지만 분산이 없을 때

밀도가 $f(x) = \alpha x^{-(\alpha+1)}$ $(x \geq 1)$인 Pareto 분포는 $\alpha > k$일 때만 k차 적률이 있다(연습 7.1). $\alpha = 1.5$이면 평균 $\mu = \frac{\alpha}{\alpha-1} = 3$은 있고 분산은 무한하다. 소득, 도시의 인구, 파일 크기, 보험 손해처럼 "소수가 전체의 큰 몫을 차지하는" 양이 이런 꼬리를 갖는다.

예제 7.1 (Pareto 표본의 평균) 정리 1.6으로 $\bar{X}_n \xrightarrow{P} 3$이다. 그러나 수렴은 느리고 기묘하다.

	$P(\lvert\bar{X}_n - 3\rvert < 0.1)$	$P(\bar{X}_n < 3)$	t-구간의 포함확률
$n = 100$	0.10	0.71	0.71
$n = 1{,}000$	—	—	0.76
$n = 10{,}000$	0.46	0.68	—

표본평균은 대개 참값보다 **작다**. 평균의 상당 부분을 드물게 나오는 아주 큰 값이 실어 나르는데, 대부분의 표본에는 그런 값이 없기 때문이다. 그리고 드물게 큰 값이 들어오면 평균이 크게 뛴다. 95% t-구간은 $n = 1000$에서도 76%만 맞고, 놓친 구간은 거의 모두 참값보다 아래에 있다.

무엇이 잘못되었나? $\sqrt{n}(\bar{X}_n - \mu)$의 "척도"가 맞지 않는다. 분산이 무한하면 합의 흩어짐은 $\sqrt{n}$보다 빨리 자란다.

> **증명 없이 쓰는 사실** (일반화된 중심극한정리 — 안정분포) 꼬리가 $P(|X|>x)\approx cx^{-\alpha}$ $(0<\alpha<2)$이면 적당한 상수 b_n에 대해 $\dfrac{\sum_{i\le n}X_i - b_n}{n^{1/\alpha}}$이 정규가 아닌 **$\alpha$-안정분포**로 분포수렴한다 (Lévy 1925, Gnedenko–Kolmogorov 1949). $\alpha=2$의 경계가 정규, $\alpha=1$의 대칭인 경우가 Cauchy다. 안정분포는 "같은 분포 둘을 더해 척도를 바꾸면 다시 같은 분포"가 되는 분포이고, 정규와 Cauchy 말고는 밀도가 닫힌 꼴이 아니다.

$\alpha=1.5$이면 합은 $n^{2/3}$의 척도로 흩어지므로 평균의 오차는 $n^{2/3}/n = n^{-1/3}$의 속도로만 준다. 정확도를 한 자리 높이려면 n을 100배가 아니라 1,000배 해야 한다. 그리고 극한분포는 오른쪽으로 크게 치우친 안정분포다(그림 7.1).

7.2 평균도 없을 때

Cauchy 분포는 평균이 없다. 예제 3.5처럼 $\bar{X}_n$은 모든 n에서 관측 하나와 같은 분포이므로 대수의 법칙조차 없다. 누적 평균의 경로는 수렴하지 않고 몇천 번째 관측에서도 크게 뛴다(그림 7.1 왼쪽). 그러나 중심을 추정할 방법이 없는 것은 아니다. 표본 **중앙값**은 Cauchy에서도 $\sqrt{n}$의 속도로 중심에 모이고, 점근분산은 $\dfrac{\pi^2}{4n}$이다($n=101$의 시뮬레이션에서 $n\,\mathrm{Var}=2.52$, 이론 2.47). 평균이 실패하는 곳에서 순위에 기반한 통계량이 살아남는다. 15회의 주제다.

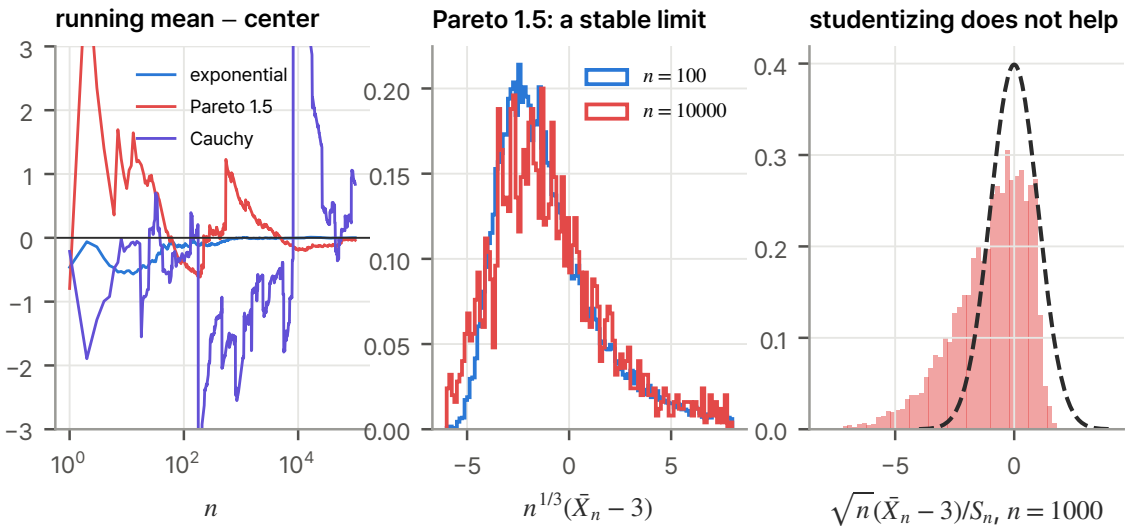


그림 7.1 왼쪽: 누적 평균의 경로. 지수(분산 유한)는 $1/\sqrt{n}$으로 안정되고, Pareto $\alpha=1.5$는 드문 큰 값마다 뛰며 천천히 내려앉고, Cauchy는 영원히 뛴다. 가운데: Pareto 표본의 $n^{1/3}(\bar{X}_n-3)$ ($n=100$, 10,000)은 같은 치우친 모양으로 모인다. 오른쪽: 같은 자료를 $\sqrt{n}(\bar{X}_n-3)/S_n$으로 표준화한 값은 표준정규(점선)와 거리가 멀다.

7.3 가정이 성립해도 — 꼬리의 상대오차

중심극한정리는 분포함수의 **절대** 오차가 0으로 간다고 말한다. 꼬리의 아주 작은 확률은 절대오차가 작아도 **상대**오차가 클 수 있다.

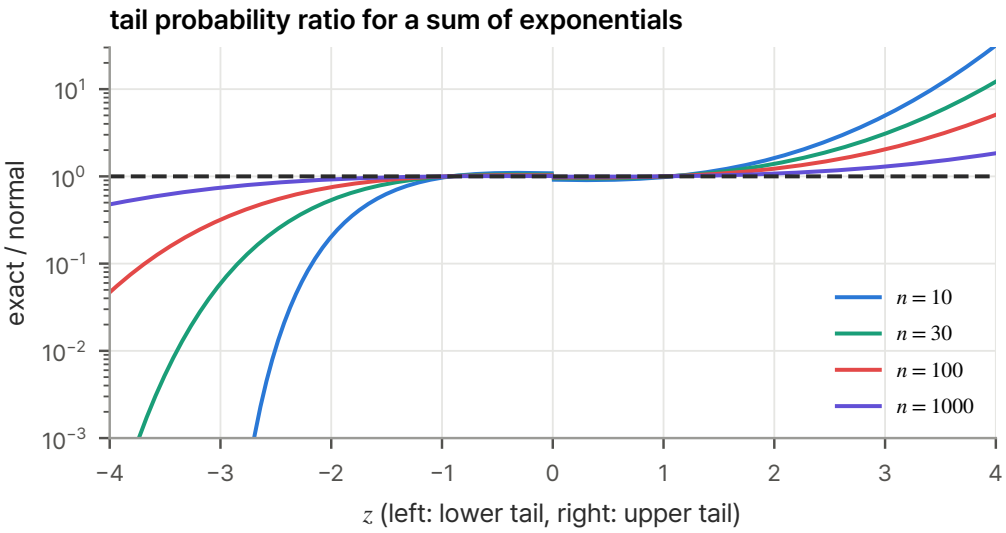


그림 7.2 Exp(1) n개의 표준화한 합에서 정확한 꼬리 확률과 정규근사의 비(오른쪽 $z > 0$: $P(Z_n > z)/(1 - \Phi(z))$, 왼쪽 $z < 0$: $P(Z_n < z)/\Phi(z)$). 정규근사는 오른쪽 꼬리를 얕잡고(비 > 1) 왼쪽 꼬리를 과장한다(비 < 1). n이 커지면 1로 가지만 꼬리가 깊을수록 느리다. 중심극한정리는 분포의 가운데에 대한 정리다.

예제 7.2 (지수 30개의 꼬리) Exp(1) 30개의 합 $S \sim \Gamma(30,1)$에서

$$P\big(S > 30 + 3\sqrt{30}\big) = 0.0042, \qquad P\big(S < 30 - 3\sqrt{30}\big) = 0.00008$$

인데 정규근사는 둘 다 0.00135다. 오른쪽 꼬리는 3.1배 얕잡고 왼쪽은 17배 과장한다. 분포함수의 오차는 기껏 0.024였다(예제 3.4). 꼬리의 확률을 물을 때는 Chernoff 한계(2회 정리 8.1)나 정확한 분포가 필요하다. 유전체 연구에서 $p < 5\times10^{-8}$ 같은 문턱을 정규근사로 계산하면 이 차이가 결론을 바꾼다.

7.4 기울기가 0이면

델타 방법은 $g'(\theta)\neq 0$일 때 $\sqrt{n}$의 속도로 정규 극한을 준다. $g'(\theta)=0$이면 1차 항이 사라져 극한이 0으로 퇴화한다. 예를 들어 $\mu=0$일 때 $\bar X_n^2$은 $\sqrt{n}$이 아니라 n의 속도로 0에 가고, $n\bar X_n^2 \xrightarrow{D} \sigma^2\chi^2(1)$이다(연습 4.2). 지수 표본에서 중심화한 $n\bar X^2$의 평균은 1.00, $P(n\bar X^2>3.84)=0.050$로 $\chi^2(1)$과 맞는다. 귀무가설 위에서 기울기가 0이 되는 이런 상황이 9회의 가능도비 검정(Wilks 정리)이 χ^2 분포를 갖는 이유다.

연습문제

연습 7.1 (Pareto의 적률 ★) $f(x)=\alpha x^{-(\alpha+1)}$ $(x\ge 1)$이면 $EX^k<\infty \iff k<\alpha$임을 보이고, $\alpha>1$이면 $EX=\frac{\alpha}{\alpha-1}$임을 보여라. 역함수 표집으로 $X=U^{-1/\alpha}$임을 보여라. 풀이 49쪽

연습 7.2 (중앙값은 왜 대개 작은가 ★★) Pareto $\alpha=1.5$의 중앙값은 $2^{1/\alpha}=1.59$로 평균 3보다 훨씬 작다. 이것과 "표본평균이 대개 참값보다 작다"는 관찰을 연결해 설명하라. 풀이 49쪽

연습 7.3 (Cauchy의 중앙값 ★★) 연속 밀도 f의 중앙값 m에서 $f(m)>0$이면 표본중앙값의 점근분산이 $\frac{1}{4nf(m)^2}$임을 받아들이자(8회). Cauchy에서 $\frac{\pi^2}{4n}$, 정규에서 $\frac{\pi\sigma^2}{2n}$임을 보이고, 정규에서 평균에 대한 중앙값의 효율 $\frac{2}{\pi}=0.64$를 구하라. 풀이 49쪽

연습 7.4 (꼬리의 비 ★) 노트북으로 $\Gamma(n,1)$의 정확한 꼬리 확률과 정규근사의 비를 $z=3$에서 구하라($n=30,300,3000$). 비가 1에 가까워지는 속도는? 풀이 49쪽

8

Beautiful insights — 평균을 내는 기계들

8.1 미니배치 경사의 잡음

신경망의 손실은 학습 자료 N개의 평균 $L(\theta)=\frac{1}{N}\sum_{i=1}^{N}\ell_i(\theta)$이다. 확률적 경사하강법(SGD)은 매 단계 B개를 뽑아 그 평균 기울기 $\hat{\mathbf{g}}=\frac{1}{B}\sum_{i\in\text{배치}}\nabla\ell_i$로 $\theta\leftarrow\theta-\eta\hat{\mathbf{g}}$를 한다. $\hat{\mathbf{g}}$는 N개의 기울기라는 유한 모집단에서 되돌려 넣지 않고 뽑은 표본평균이다.

명제 8.1 (미니배치 기울기의 평균과 공분산) $\mathbf{g}_1,\dots,\mathbf{g}_N$의 평균을 $\bar{\mathbf{g}}=\nabla L$, 공분산을 $\Sigma=\frac{1}{N}\sum(\mathbf{g}_i-\bar{\mathbf{g}})(\mathbf{g}_i-\bar{\mathbf{g}})'$라 하면, 크기 B의 단순무작위 배치에 대해

$$E\hat{\mathbf{g}}=\nabla L,\qquad \operatorname{Cov}(\hat{\mathbf{g}})=\frac{\Sigma}{B}\cdot\frac{N-B}{N-1}.$$

증명. I_i를 i가 배치에 들어가는 사건의 지시변수라 하면 $\hat{\mathbf{g}}=\frac{1}{B}\sum_i I_i\mathbf{g}_i$, $EI_i=\frac{B}{N}$이므로 평균이 나온다. $\sum I_i=B$가 상수이므로 $\mathbf{g}_i$를 $\tilde{\mathbf{g}}_i=\mathbf{g}_i-\bar{\mathbf{g}}$로 바꿔도 공분산은 같다. $\operatorname{Var}I_i=\frac{B}{N}(1-\frac{B}{N})$, $i\neq j$이면 $\operatorname{Cov}(I_i,I_j)=-\frac{\operatorname{Var}I_i}{N-1}$이다($\operatorname{Var}\sum I_i=0$에서). 3회 정리 4.2로

$$\operatorname{Cov}(\hat{\mathbf{g}})=\frac{\operatorname{Var}I_1}{B^2}\Big[\sum_i\tilde{\mathbf{g}}_i\tilde{\mathbf{g}}_i'-\frac{1}{N-1}\sum_{i\neq j}\tilde{\mathbf{g}}_i\tilde{\mathbf{g}}_j'\Big]=\frac{\operatorname{Var}I_1}{B^2}\Big[N\Sigma+\frac{N\Sigma}{N-1}\Big],$$

$\sum_{i\neq j}\tilde{\mathbf{g}}_i\tilde{\mathbf{g}}_j'=(\sum\tilde{\mathbf{g}}_i)(\sum\tilde{\mathbf{g}}_i)'-\sum\tilde{\mathbf{g}}_i\tilde{\mathbf{g}}_i'=-N\Sigma$이기 때문이다. 정리하면 결론이다. ∎

$\frac{N-B}{N-1}$은 4회 정리 1.5의 유한 모집단 수정 그대로다. $B\ll N$이면 무시되고, 다변량 중심극한정리로 $\hat{\mathbf{g}}\approx\mathcal{N}(\nabla L,\Sigma/B)$이다. 그러면 SGD의 한 단계는

$$\theta\leftarrow\theta-\eta\nabla L+\eta\boldsymbol{\xi},\qquad \boldsymbol{\xi}\approx\mathcal{N}(\mathbf{0},\Sigma/B)$$

곧 경사하강에 정규 잡음을 더한 것이다. 잡음의 세기를 정하는 것은 η와 B 따로가 아니라 대략 비 η/B다(Mandt, Hoffman, Blei 2017; Smith와 Le 2018). 배치를 k배로 키울 때 학습률도 k배로 키우라는 **선형 스케일링 규칙**(Goyal 외, 2017)이 여기서 나온다. 노트북의 선형회귀($N=2000$)에서 기울기 공분산의 대각합은 $B=8,64,512$에서 이론값 2.19, 0.266, 0.0256과 맞고, $B=512$에서는 유한 모집단 수정이 없으면 0.0344로 34% 틀린다.

가정 주의 정규 잡음이라는 근사는 개별 기울기의 분산이 유한하다는 가정 위에 있다. 깊은 신경망의 기울기 잡음이 꼬리가 두꺼운 안정분포에 가깝다는 관찰(Simsekli 외, 2019)과 보통의 배치 크기에서는 정규에 가깝다는 반론이 함께 있다. 7절의 그림이 둘 중 어느 쪽이냐에 따라 달라진다.

8.2 MC dropout — 예측을 여러 번 평균하기

드롭아웃은 학습 중에 뉴런을 확률 q로 끈다. Gal과 Ghahramani(2016)는 **예측할 때도** 드롭아웃을 켜고 같은 입력을 T번 통과시키자고 했다. 출력 $f_1(x), \dots, f_T(x)$는 무작위 마스크에 대한 iid 표본이므로 그 평균은 기대 예측 $E_m f(x; m)$의 Monte Carlo 추정이고, 표본표준편차는 예측의 흔들림, 곧 모형의 불확실성을 나타낸다. 이 해석이 근사적인 베이즈 추론이라는 논증은 16회에서 본다. 여기서 중요한 것은 정리 6.1다. 평균의 Monte Carlo 오차는 $s/\sqrt{T}$이므로 $T = 100$이면 불확실성의 10%가 계산 자체의 잡음이다. 불확실성의 크기를 말하려면 그것을 재는 계산의 오차부터 알아야 한다.

8.3 대각합을 무작위로 — Hutchinson 추정량

큰 행렬 A의 대각합 $\mathrm{tr}A$가 필요하지만 A를 만들 수는 없고 곱 $A\mathbf{z}$만 계산할 수 있는 경우가 많다(신경망의 헤시안, 연속 정규화 흐름의 야코비안, 점수 맞추기의 발산). Hutchinson(1989)의 방법은 무작위 벡터 하나로 대각합을 추정한다.

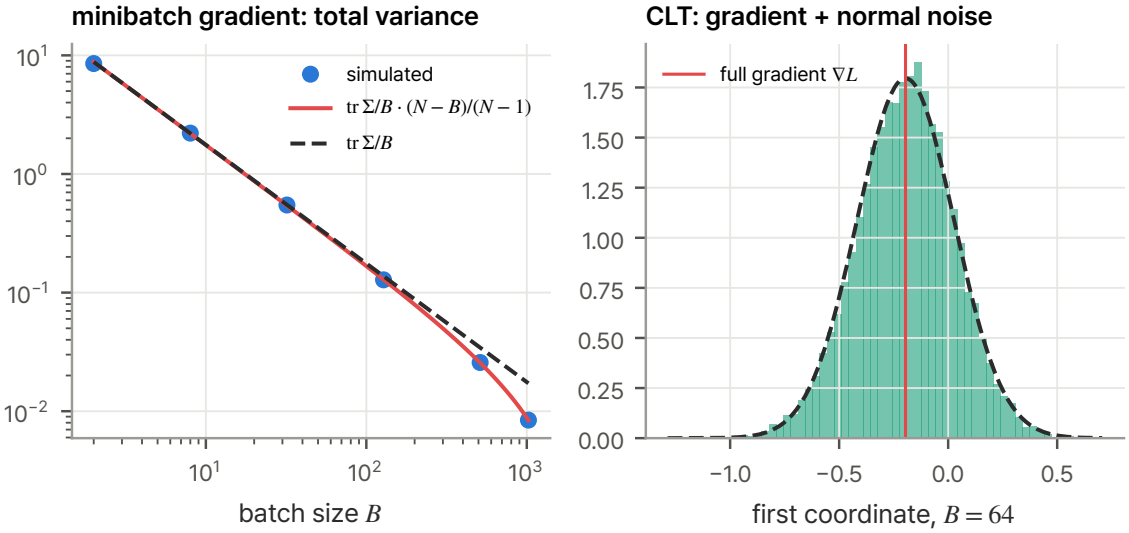


그림 8.1 왼쪽: 미니배치 기울기 공분산의 대각합(점)과 명제 8.1(실선). 점선은 유한 모집단 수정이 없는 $\mathrm{tr}\,\Sigma/B$. 오른쪽: 한 좌표의 미니배치 기울기($B = 64$)의 히스토그램과 정규 근사.

명제 8.2 (Hutchinson 추정량) $E\mathbf{z}=\mathbf{0}$, $E\mathbf{z}\mathbf{z}'=I$이면 $E\,\mathbf{z}'A\mathbf{z}=\mathrm{tr}A$이다. A가 대칭일 때

(a) z_i가 iid Rademacher(±1이 반반)이면 $\mathrm{Var}\;\mathbf{z}'A\mathbf{z}=2\big(\|A\|_F^2-\sum_i a_{ii}^2\big)$,

(b) $\mathbf{z}\sim\mathcal{N}(\mathbf{0},I)$이면 $\mathrm{Var}\;\mathbf{z}'A\mathbf{z}=2\|A\|_F^2$.

증명. $E\mathbf{z}'A\mathbf{z}=\sum_{i,j}a_{ij}Ez_iz_j=\sum_i a_{ii}$. (a) $z_i^2=1$이므로 $\mathbf{z}'A\mathbf{z}=\mathrm{tr}A+\sum_{i\neq j}a_{ij}z_iz_j$이고, $E(z_iz_jz_kz_l)$ $(i\neq j,\,k\neq l)$은 $\{k,l\}=\{i,j\}$일 때만 1이다. 따라서 분산은 $\sum_{i\neq j}a_{ij}(a_{ij}+a_{ji})=2\sum_{i\neq j}a_{ij}^2$. (b)는 연습 8.4. ∎

Rademacher가 대각 성분만큼 분산이 작다. m개의 탐침을 평균하면 분산이 $1/m$로 줄고, 정리 6.1로 오차 막대를 함께 낸다. 노트북의 30×30 대칭행렬에서 탐침 하나의 분산의 이론값은 Rademacher 872, 정규 916이고, 시뮬레이션이 이를 확인한다.

연습문제

연습 8.1 (지시변수의 공분산 ★★) 명제 8.1의 증명에서 $\mathrm{Cov}(I_i,I_j)=-\frac{B(N-B)}{N^2(N-1)}$을 $P(I_i=I_j=1)=\frac{B(B-1)}{N(N-1)}$로 직접 확인하라. 풀이 49쪽

연습 8.2 (선형 스케일링 ★) $B=256$, $\eta=0.1$로 학습하던 모형을 $B=8192$로 옮긴다. 잡음의 세기 η/B를 유지하는 학습률은? 한 에폭의 단계 수는 몇 배가 되나? 풀이 50쪽

연습 8.3 (몇 번 통과시키나 ★) MC dropout에서 예측 평균의 Monte Carlo 오차를 예측 표준편차의 5% 이하로 하려면 T는 얼마여야 하나? 풀이 50쪽

연습 8.4 (정규 탐침의 분산 ★★★) $\mathbf{z}\sim\mathcal{N}(\mathbf{0},I)$이고 A가 대칭이면 $A=Q\Lambda Q'$로 대각화해 $\mathbf{z}'A\mathbf{z}=\sum_i\lambda_iW_i^2$ $(W_i\overset{\text{iid}}{\sim}\mathcal{N}(0,1))$임을 보이고(5회 정리 3.8), $\mathrm{Var}=2\sum\lambda_i^2=2\|A\|_F^2$을 유도하라. 풀이 50쪽

연습 8.5 (탐침의 수 ★★) Rademacher 탐침 m개로 $\mathrm{tr}A$를 상대오차 1%(표준오차 기준)로 추정하려면 $m\geq\frac{2(\|A\|_F^2-\sum a_{ii}^2)}{(0.01\,\mathrm{tr}A)^2}$이어야 함을 보여라. $A=I_d$이면? A가 양의 정부호이고 고윳값이 고르면 m은 차원 d에 대해 어떻게 변하나? 풀이 50쪽

9

정리와 역사 노트

9.1 한 장 요약

정리	내용	조건
연속사상 (1.3)	$Y_n \xrightarrow{P} c$, g가 c에서 연속 $\Rightarrow g(Y_n) \xrightarrow{P} g(c)$	—
Khinchin (1.6)	$\bar{X}_n \xrightarrow{P} \mu$; 따라서 $S_n^2 \xrightarrow{P} \sigma^2$	iid, $E\lvert X\rvert < \infty$
두 수렴 (2.3, 2.5)	$\xrightarrow{P}\Rightarrow\xrightarrow{D}$; 상수로의 $\xrightarrow{D}\Rightarrow\xrightarrow{P}$; $\xrightarrow{D}\Rightarrow$ 확률적 유계	—
중심극한정리 (3.1)	$\sqrt{n}(\bar{X}_n-\mu)/\sigma \xrightarrow{D} \mathcal{N}(0,1)$, 오차 $\le 0.4748\,\rho/(\sigma^3\sqrt{n})$	iid, $\sigma^2<\infty$
Slutsky (4.2)	$A_n + B_nX_n \xrightarrow{D} a+bX$; $\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1)$	A_n, B_n은 상수로
델타 방법 (4.4, 5.4)	$\sqrt{n}(g(Y_n)-g(\theta)) \xrightarrow{D} \mathcal{N}(0, g'(\theta)^2\sigma^2)$, 벡터는 $\nabla g'\Sigma\nabla g$	g 미분가능
다변량 CLT (5.2)	$\sqrt{n}(\bar{\mathbf{X}}_n-\boldsymbol{\mu}) \xrightarrow{D} \mathcal{N}_k(\mathbf{0},\Sigma)$	모든 선형결합
Monte Carlo (6.1, 6.4, 6.6)	$\hat{\theta}_n \pm z\,s_h/\sqrt{n}$, 기각 표집(수락률 $1/M$), 중요도 가중치 f/g	$\operatorname{Var} h<\infty$

증명의 뼈대는 둘이다. 하나는 **mgf의 로그를 2차까지 전개**하는 것(중심극한정리, 다변량, χ^2 극한)이고, 다른 하나는 **작은 항을 따로 떼어 0으로 보내는** 것(Khinchin의 잘라 내기, Slutsky, 델타 방법)이다. 7절은 그 뼈대의 가정, 곧 유한한 분산을 깨면 척도가 $\sqrt{n}$에서 $n^{1/\alpha}$로 바뀌고 극한이 정규에서 안정분포로 바뀜을 보였다.

9.2 Stigler의 명명 법칙

> **Stigler's law of eponymy**
>
> - **Cauchy 분포**: Poisson(1824)이 평균을 내도 나아지지 않는 오차의 예로 먼저 썼다. Cauchy는 1853년 Bienaymé와의 논쟁에서 같은 분포를 다시 논했다. Stigler(1974)가 이 내력을 정리했다. 기하학에서는 17세기부터 "Agnesi의 마녀"라는 곡선으로 알려져 있었다.

- **Monte Carlo 방법**: 이름은 Metropolis(1949)가 붙였지만, Buffon(1777)의 바늘, Kelvin(1901)의 기체 운동 계산, Student(1908)의 카드, Fermi의 1930년대 중성자 계산이 모두 같은 방법이었다.
- **Slutsky의 정리**: 내용의 일부는 Cramér와 다른 이들이 독립적으로 썼고, 교과서의 이름이 굳은 것은 Cramér(1946)를 통해서다.

9.3 7 Pillars 지도

이번 회차는 둘째 기둥 **Information**의 수학적 핵심이다. Stigler(2016)는 이 기둥을 "정보는 관측 수의 제곱근으로 쌓인다"로 요약했다. 3회에서 $\mathrm{Var}\,\bar{X} = \sigma^2/n$로 그 크기를 보았다면, 이번 회차의 중심극한정리는 그 $\sqrt{n}$의 오차가 **어떤 모양**인지를 말한다. 그리고 7절은 이 규칙의 한계, 곧 분산이 무한하면 정보가 $\sqrt{n}$보다 느리게 쌓인다는 것을 보였다. 첫째 기둥 **Aggregation**(평균을 내면 개별 정보를 버리고도 더 많이 안다)도 대수의 법칙의 형태로 다시 나왔다.

9.4 다음 회 — 추론의 기초: 표본 · 신뢰구간 · 유의성 검정 (1710 → 1937)

이 노트의 극한 정리들로 이제 통계적 추론을 세울 수 있다. 8회는 Arbuthnot(1710)이 82년의 출생 기록으로 "우연인가 섭리인가"를 물은 최초의 유의성 검정에서 시작해, 신뢰구간(Neyman 1937)과 검정의 언어를 정의한다. 5절의 다항분포 극한은 Pearson(1900)의 χ^2 적합도 검정이 되고, 4절의 스튜던트화한 평균은 대표본 신뢰구간이 된다.

참고문헌

- Buffon, G.-L. L. *Essai d'arithmétique morale*. Supplément à l'Histoire naturelle 4, 1777.
- Laplace, P.-S. Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur application aux probabilités. *Mémoires de l'Académie des sciences de Paris* (1810).
- Lyapunov, A. M. Nouvelle forme du théorème sur la limite de probabilité. *Mémoires de l'Académie impériale des sciences de St.-Pétersbourg* 12 (1901).
- Lindeberg, J. W. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 15 (1922), 211–225.
- Pólya, G. Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Mathematische Zeitschrift* 8 (1920), 171–181.
- Slutsky, E. Über stochastische Asymptoten und Grenzwerte. *Metron* 5 (1925), 3–89.
- Khinchin, A. Ya. Sur la loi des grands nombres. *Comptes rendus* 189 (1929), 477–479.
- Curtiss, J. H. A note on the theory of moment generating functions. *Ann. Math. Statist.* 13 (1942), 430–433.
- Berry, A. C. The accuracy of the Gaussian approximation to the sum of independent variates. *Trans.*

AMS 49 (1941). · Esseen, C.-G. On the Liapounoff limit of error in the theory of probability. *Ark. Mat. Astr. Fys.* 28A (1942). · Shevtsova, I. On the absolute constants in the Berry–Esseen type inequalities. *arXiv:1111.6554* (2011).

- Gnedenko, B. V., Kolmogorov, A. N. *Limit Distributions for Sums of Independent Random Variables* (1949; 영역 1954).
- Metropolis, N., Ulam, S. The Monte Carlo method. *JASA* 44 (1949), 335–341. · Metropolis, N. The beginning of the Monte Carlo method. *Los Alamos Science* 15 (1987).
- von Neumann, J. Various techniques used in connection with random digits. *NBS Applied Mathematics Series* 12 (1951), 36–38.
- Gridgeman, N. T. Geometric probability and the number π. *Scripta Mathematica* 25 (1960). · Badger, L. Lazzarini's lucky approximation of π. *Mathematics Magazine* 67 (1994), 83–91.
- Stigler, S. M. Cauchy and the witch of Agnesi. *Biometrika* 61 (1974), 375–380.
- Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Comm. Statist. Simulation Comput.* 18 (1989), 1059–1076.
- Mandt, S., Hoffman, M. D., Blei, D. M. Stochastic gradient descent as approximate Bayesian inference. *JMLR* 18 (2017). · Smith, S. L., Le, Q. V. A Bayesian perspective on generalization and stochastic gradient descent. *ICLR* (2018). · Goyal, P. 외. Accurate, large minibatch SGD: training ImageNet in 1 hour. *arXiv:1706.02677* (2017). · Simsekli, U., Sagun, L., Gürbüzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. *ICML* (2019).
- Gal, Y., Ghahramani, Z. Dropout as a Bayesian approximation. *ICML* (2016).
- 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.

1절

연습 1.1 문제 6쪽

바늘을 길이 $\ell/m \le d$인 토막 m개로 나누면 $N = \sum_{j=1}^{m} N_j$이고 N_j는 토막 j의 교차 횟수다. 길이가 d 이하인 토막은 기껏 한 번 걸치므로 N_j는 지시변수이고 정리 1.1으로 $EN_j = \frac{2(\ell/m)}{\pi d}$. 토막들은 서로 의존하지만 기댓값의 선형성(2회 정리 1.10)은 독립을 요구하지 않으므로 $EN = m \cdot \frac{2\ell}{m\pi d} = \frac{2\ell}{\pi d}$. 휘어진 철사는 짧은 선분들로 어림하고 극한을 취한다.

연습 1.2 문제 7쪽

$|Y_n - 0| = \frac{1}{n} < \varepsilon$이 결국 성립하므로 $Y_n \xrightarrow{P} 0$. 그러나 $g(Y_n) = 1$이 모든 n에서 성립하고 $g(0) = 0$이다. g가 0에서 연속이 아니다(오른쪽 극한이 1).

연습 1.3 문제 7쪽

$P(|Y_n| \ge \varepsilon) = P(Y_n = n) = \frac{1}{n} \to 0$ $(\varepsilon \le n)$. $EY_n = n \cdot \frac{1}{n} = 1$. 확률 $\frac{1}{n}$로 나오는 값 n이 기댓값에는 끝까지 1만큼 기여한다.

연습 1.4 문제 7쪽

$t(2)$는 $E|X| < \infty$(6회 정리 3.3(b), $k = 1 < 2$)이고 대칭이므로 평균이 0이다. 정리 1.6으로 $\bar{X}_n \xrightarrow{P} 0$. 시뮬레이션은 $P(|\bar{X}_n| > 0.1) = 0.68$ $(n = 100)$, 0.018 $(n = 10{,}000)$. 분산이 무한하므로 Chebyshev 부등식의 오른쪽 $\sigma^2/(n\varepsilon^2)$이 무한대가 되어 2회 정리 6.3은 아무것도 말하지 못한다.

연습 1.5 문제 7쪽

$|X| > n$이면 왼쪽이 0이다. $|X| \le n$일 때, $|X| \le M$이면 $X^2 = |X| \cdot |X| \le M|X|$이고, $M < |X| \le n$이면 $X^2 \le n|X| = n|X|1\{|X| > M\}$이다. $M \ge n$이면 둘째 경우가 없고 첫째 경우만으로 충분하다.

어느 경우든 오른쪽의 두 항은 음이 아니다.

연습 1.6 문제 7쪽

(a) 바늘의 중심에서 바늘 끝까지는 길이 $\frac{\ell}{2}$이고, 바늘이 이음매와 이루는 예각이 Θ이면 그 끝이 이음매에 수직인 방향으로 $\frac{\ell}{2}\sin\Theta$만큼 떨어져 있다. 가장 가까운 이음매까지의 거리 X가 이 값 이하이면 바늘이 이음매를 넘고, 크면 넘지 않는다. 따라서 걸치는 사건은 $\{X \le \frac{\ell}{2}\sin\Theta\}$이다. (b) 독립이면 결합밀도는 주변밀도의 곱이다(3회 정리 3.2). X의 밀도는 $(0, \frac{d}{2})$에서 $\frac{2}{d}$, Θ의 밀도는 $(0, \frac{\pi}{2})$에서 $\frac{2}{\pi}$이므로 곱은 $\frac{4}{\pi d}$이다. (c) $\sin\theta \le 1$이므로 $\frac{\ell}{2}\sin\theta \le \frac{\ell}{2} \le \frac{d}{2}$이다. $\ell > d$이면 $\frac{\ell}{2}\sin\theta$가 $\frac{d}{2}$를 넘는 θ에서 안쪽 적분의 위끝이 $\min(\frac{\ell}{2}\sin\theta, \frac{d}{2})$가 되어 계산이 바뀐다. (d)

$$\int_0^{\pi/2}\int_0^{\frac{\ell}{2}\sin\theta} \frac{4}{\pi d}\,dx\,d\theta = \frac{4}{\pi d}\cdot\frac{\ell}{2}\int_0^{\pi/2}\sin\theta\,d\theta = \frac{4}{\pi d}\cdot\frac{\ell}{2}\cdot 1 = \frac{2\ell}{\pi d}.$$

연습 1.7 문제 7쪽

(a) X_i가 iid이면 그 함수 X_i^k도 iid이고(3회 정리 3.4), $E|X_1^k| = E|X_1|^k < \infty$이다. 정리 1.6으로 $\frac{1}{n}\sum X_i^k \xrightarrow{P} EX^k$이다. (b) $\sum(X_i - \bar X_n)^2 = \sum X_i^2 - 2\bar X_n \sum X_i + n\bar X_n^2 = \sum X_i^2 - n\bar X_n^2$이므로 $S_n^2 = \frac{1}{n-1}\sum(X_i-\bar X_n)^2 = \frac{n}{n-1}\big(\frac{1}{n}\sum X_i^2 - \bar X_n^2\big)$이다. (c) (a)로 $\frac{1}{n}\sum X_i^2 \xrightarrow{P} EX^2$, $\bar X_n \xrightarrow{P} \mu$이고, 따름정리 1.4로 $\frac{1}{n}\sum X_i^2 - \bar X_n^2 \xrightarrow{P} EX^2 - \mu^2 = \sigma^2$이다. 상수 수열 $\frac{n}{n-1} \to 1$은 확률변수 $\frac{n}{n-1}$으로 보면 1로 확률수렴하므로 다시 따름정리 1.4의 곱을 쓴다. (d) $\sqrt{\cdot}$는 $\sigma^2 > 0$에서 연속이므로 정리 1.3(a)로 $S_n \xrightarrow{P} \sigma$이다. $\sigma^2 = 0$이어도 $\sqrt{\cdot}$는 $[0,\infty)$의 0에서 (오른쪽으로) 연속이고 $S_n^2 \ge 0$이므로 결론은 같다. 이때 X는 상수이고 $S_n = 0$이다.

2절

연습 2.1 문제 12쪽

$$P(nm_n > x) = P(\text{모든 } U_i > \tfrac{x}{n}) = (1-\tfrac{x}{n})^n \to e^{-x}\ (x \ge 0).$$

연습 2.2 문제 12쪽

$$P(M_n \le x + \log n) = \big(1 - e^{-x-\log n}\big)^n = \big(1 - \frac{e^{-x}}{n}\big)^n \to \exp(-e^{-x})$$

(n이 커서 $x + \log n > 0$이면). $x = 0$에서 $n = 10$은 $0.9^{10} = 0.349$, $n = 100$은 0.366, 극한은 0.368이다.

연습 2.3 문제 12쪽

$P(X_n > k) = (1-p_n)^k$이므로 $P(X_n/n > x) = (1-\frac{\lambda}{n})^{\lfloor nx\rfloor}$. $\lfloor nx\rfloor = nx - \delta$ $(0 \le \delta < 1)$이고 $(1-\frac{\lambda}{n})^{-\delta} \to 1$이므로 극한은 $e^{-\lambda x}$, 곧 $\mathrm{Exp}(\lambda)$의 생존함수다.

연습 2.4 문제 12쪽

$y > 0$이고 $\pm\sqrt{y}$가 연속점(정규이므로 모든 점)이면 $P(Z_n^2 \le y) = P(-\sqrt{y} \le Z_n \le \sqrt{y}) = F_n(\sqrt{y}) - P(Z_n < -\sqrt{y})$. $F_n(-\sqrt{y}-\varepsilon) \le P(Z_n < -\sqrt{y}) \le F_n(-\sqrt{y})$이므로 극한은 $\Phi(-\sqrt{y})$ ($\varepsilon \to 0$, Φ 연속). 따라서 $P(Z_n^2 \le y) \to \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = P(Z^2 \le y)$이고 $Z^2 \sim \chi^2(1)$(4회 정리 3.9).

연습 2.5 문제 12쪽

$Y_n = -X_n$이면 대칭인 Z에 대해 $Y_n \xrightarrow{D} Z$이지만 $X_n + Y_n = 0$이므로 극한은 상수 0이지 $\mathcal{N}(0,2)$가 아니다. 분포수렴은 결합분포에 대해 아무것도 말하지 않는다.

연습 2.6 문제 12쪽

(a) $F(-B) \to 0$ ($B \to \infty$)이고 $F(B) \to 1$이므로(1회 정리 6.3(b)) 충분히 큰 B_0 이상에서 두 부등식이 성립한다. 불연속점은 셀 수 있게 많으므로 $[B_0, \infty)$에는 B와 $-B$가 모두 연속점인 B가 있다. 셀 수 있는 집합 둘(불연속점, 불연속점의 부호를 바꾼 것)의 합집합을 피하면 된다. (b) $\pm B$가 연속점이므로 $F_n(\pm B) \to F(\pm B)$이다. N이 있어 $n \ge N$이면 $F_n(B) > F(B) - \frac{\eta}{4} > 1 - \frac{\eta}{2}$, $F_n(-B) < F(-B) + \frac{\eta}{4} < \frac{\eta}{2}$이다. $P(|X_n| \le B) = F_n(B) - P(X_n < -B) \ge F_n(B) - F_n(-B) > 1 - \eta$이다. (c) $n = 1, \dots, N-1$ 각각에서 $P(|X_n| \le B_n) \ge 1 - \eta$인 B_n이 있다(1회 정리 6.3(b)를 X_n의 분포함수에 쓴다). $\max(B, B_1, \dots, B_{N-1})$이 모든 n을 덮는다. (d) $|Y_n| \le B$이고 $|W_n| < \varepsilon/B$이면 $|Y_n W_n| < \varepsilon$이므로 포함관계가 성립한다. 따라서 $P(|Y_n W_n| \ge \varepsilon) \le \eta + P(|W_n| \ge \varepsilon/B)$이고, 둘째 항이 0으로 가므로 상극한은 η 이하다. η가 임의이므로 극한은 0이다.

3절

연습 3.1 문제 17쪽

$n\psi(t/\sqrt{n}) = n\left[\frac{t^2}{2n} + \frac{\kappa_3 t^3}{6n^{3/2}} + O(n^{-2})\right] = \frac{t^2}{2} + \frac{\kappa_3 t^3}{6\sqrt{n}} + O(\frac{1}{n})$. 지수분포를 중심화하면 $\psi(t) = -\log(1-t) - t = \frac{t^2}{2} + \frac{t^3}{3} + \cdots$이므로 $\frac{\kappa_3}{6} = \frac{1}{3}$, $\kappa_3 = 2$. 왜도 2가 $1/\sqrt{n}$의 오차를 만든다.

연습 3.2 문제 17쪽

$\int_1^\infty (x-1)^3 e^{-x} dx = e^{-1}\int_0^\infty u^3 e^{-u} du = \frac{6}{e}$. 부분적분을 되풀이하면 $\int_0^1 (1-x)^3 e^{-x} dx = \frac{6}{e} - 2$ (계산: $\int_0^1 (1-x)^k e^{-x} dx = 1 - kI_{k-1}$, $I_0 = 1 - e^{-1}$에서 $I_1 = e^{-1}$, $I_2 = 1 - 2e^{-1}$, $I_3 = 1 - 3 + 6e^{-1}$). 합은 $\frac{12}{e} - 2 = 2.415$.

연습 3.3 문제 17쪽

$X_i - p_i$는 확률 p_i로 q_i, 확률 q_i로 $-p_i$이므로 $E|X_i-p_i|^3 = p_i q_i^3 + q_i p_i^3 = p_i q_i(p_i^2+q_i^2) \le p_i q_i$. 따라서 $\frac{1}{s_n^3}\sum E|X_i-p_i|^3 \le \frac{s_n^2}{s_n^3} = \frac{1}{s_n} \to 0$.

연습 3.4 문제 17쪽

$EU = \frac12$, $\operatorname{Var} U = \frac{1}{12}$이므로 합은 평균 6, 분산 1이다. 누율은 독립합에서 더해진다(로그 mgf가 더해지므로). U의 초과첨도가 $\kappa_4(U)/\sigma_U^4 = -1.2$이므로 합의 초과첨도는 $\frac{12\,\kappa_4(U)}{(12\sigma_U^2)^2} = \frac{1}{12}\cdot\frac{\kappa_4(U)}{\sigma_U^4} = -0.1$. 결점: 값이 $[-6,6]$ 밖으로 나가지 못하고 꼬리가 얇다($P(|\cdot|>3)=0.0021$, 정규 0.0027). 난수 12개를 쓰는 것도 비싸다.

연습 3.5 문제 17쪽

$M(t) = E\exp\!\big(t\frac{V_r - r}{\sqrt{2r}}\big) = e^{-tr/\sqrt{2r}}\big(1-\frac{2t}{\sqrt{2r}}\big)^{-r/2}$. $u = \frac{2t}{\sqrt{2r}}$로 두면 $\log M = -\frac{r}{2}u - \frac{r}{2}\log(1-u) = \frac{r}{2}\big(\frac{u^2}{2} + O(u^3)\big) = \frac{r}{4}\cdot\frac{4t^2}{2r} + O(r^{-1/2}) = \frac{t^2}{2} + O(r^{-1/2})$.

연습 3.6 문제 17쪽

$\frac{1}{(1+x^2)(1+(z-x)^2)} = \frac{Ax+B}{1+x^2} + \frac{C(x-z)+D}{1+(z-x)^2}$로 놓고 분자의 계수를 맞추면 ($z\neq 0$) $A = -C = \frac{2}{z(z^2+4)}$, $B = D = \frac{1}{z^2+4}$이고, $Ax/(1+x^2)$과 $C(x-z)/(1+(z-x)^2)$의 적분은 서로 상쇄된다(로그 항의 차가 $x\to\pm\infty$에서 0). 남는 것은 $\frac{1}{z^2+4}\big(\int\frac{dx}{1+x^2} + \int\frac{dx}{1+(z-x)^2}\big) = \frac{2\pi}{z^2+4}$. π^2으로 나누면 $f_{X_1+X_2}(z) = \frac{2}{\pi(4+z^2)}$, 척도 2의 Cauchy다. 2로 나누면(2회 예제 4.5) 밀도 $2f(2w) = \frac{4}{\pi(4+4w^2)} = \frac{1}{\pi(1+w^2)}$.

4절

연습 4.1 문제 21쪽

$\sqrt{n}(\bar X_n - 1) \xrightarrow{D} \mathcal{N}(0,1)$이고 $g=\log$, $g'(1)=1$. $\sqrt{n}(\log\bar X_n - \log 1) \xrightarrow{D} \mathcal{N}(0,1)$.

연습 4.2 문제 21쪽

테일러 정리로 $g(Y_n)-g(\theta) = \frac12 g''(\xi_n)(Y_n-\theta)^2$ (ξ_n은 Y_n과 θ 사이). $\xi_n \xrightarrow{P}\theta$이고 g''가 연속이므로 $g''(\xi_n)\xrightarrow{P} g''(\theta)$. $n(Y_n-\theta)^2 = \big(\sqrt{n}(Y_n-\theta)\big)^2 \xrightarrow{D} \sigma^2\chi^2(1)$(연습 2.4을 $\sqrt{n}(Y_n-\theta)/\sigma$에). Slutsky로 $n(g(Y_n)-g(\theta)) \xrightarrow{D} \frac12 g''(\theta)\sigma^2\chi^2(1)$. $g(x)=x^2$, $\theta=0$이면 $\frac12\cdot 2 = 1$.

연습 4.3 문제 21쪽

$g(x) = \frac{2r}{x}$, $g'(p) = -\frac{2r}{p^2}$이므로 점근분산은 $\frac{4r^2}{p^4}p(1-p) = \frac{4r^2(1-p)}{p^3}$. $2r = \pi p$를 넣으면 $\frac{\pi^2(1-p)}{p}$로,

p에 대해 감소하고 $p=\frac{2r}{\pi}$는 r에 대해 증가하므로 $r=1$이 최선이다. $r=1$: $\pi^2\cdot\frac{0.3634}{0.6366}=5.63$ (표준편차 $2.37/\sqrt{n}$), $r=\frac{5}{6}$: $\pi^2\cdot\frac{0.4695}{0.5305}=8.74$ ($2.96/\sqrt{n}$), 1.25배다.

연습 4.4 문제 22쪽

시뮬레이션(2만 번)에서 $n\le 5000$ 안에 멈출 확률은 약 0.34, 멈춘 시점의 중앙값은 $n\approx 1300$이다. 세 사람 가운데 한 명은 이 방법으로 소수 여섯째 자리까지 맞는 π를 "얻는다". 보고된 값은 멈춘 규칙 때문에 언제나 355/113이므로, 그 값의 분포는 고정된 n의 $\hat{\pi}_n$의 분포와 전혀 다르다. 표준오차 0.051은 "미리 정한 n번 던진다"는 실험의 성질이지 이 실험의 성질이 아니다. 자료를 보면서 멈출 때를 정하면 오차의 계산이 무효가 된다. 8회의 p-해킹과 같은 문제다.

연습 4.5 문제 22쪽

(a) $g'(p)=\frac{1}{2\sqrt{p}}\cdot\frac{1}{\sqrt{1-p}}$이므로 $g'(p)^2pq=\frac{1}{4}$. (b) 따름정리 3.2(b)로 $\frac{X_\lambda-\lambda}{\sqrt{\lambda}}\xrightarrow{D}\mathcal{N}(0,1)$. $\sqrt{X_\lambda}-\sqrt{\lambda}=\frac{X_\lambda-\lambda}{\sqrt{X_\lambda}+\sqrt{\lambda}}$이고 $\frac{\sqrt{X_\lambda}+\sqrt{\lambda}}{2\sqrt{\lambda}}\xrightarrow{P}1$(같은 따름정리로 $X_\lambda/\lambda\xrightarrow{P}1$)이므로 Slutsky로 $2(\sqrt{X_\lambda}-\sqrt{\lambda})\xrightarrow{D}\mathcal{N}(0,1)$.

연습 4.6 문제 22쪽

그렇다. 상수로의 분포수렴은 확률수렴이므로(정리 2.3(b)) $B_n\xrightarrow{P}1$이고 Slutsky를 쓸 수 있다. B_n의 극한이 상수가 아닌 확률변수이면 결론은 거짓일 수 있다(연습 2.5).

연습 4.7 문제 22쪽

(a) $Y_n+R_n\le x$이고 $|R_n|<\varepsilon$이면 $Y_n\le x-R_n<x+\varepsilon$이다. 따라서 $Y_n+R_n\le x$이면 $Y_n\le x+\varepsilon$이거나 $|R_n|\ge\varepsilon$이다. 마찬가지로 $Y_n\le x-\varepsilon$이고 $|R_n|<\varepsilon$이면 $Y_n+R_n<x$이다. (b) 불연속점의 집합 D는 셀 수 있으므로 $\{\varepsilon>0:x+\varepsilon\in D$ 또는 $x-\varepsilon\in D\}$도 셀 수 있다. 어떤 구간 $(0,\delta)$도 셀 수 없이 많은 점을 가지므로 그 안에 이 집합에 속하지 않는 ε이 있다. (c) (a)와 Boole 부등식으로 $P(Y_n\le x-\varepsilon)-P(|R_n|\ge\varepsilon)\le P(Y_n+R_n\le x)\le P(Y_n\le x+\varepsilon)+P(|R_n|\ge\varepsilon)$이다. $x\pm\varepsilon$이 연속점이므로 $P(Y_n\le x\pm\varepsilon)\to F_Y(x\pm\varepsilon)$이고 $P(|R_n|\ge\varepsilon)\to 0$이다. (d) x가 연속점이므로 (b)의 ε을 0으로 보내면 양 끝이 $F_Y(x)$로 간다. 끼우는 방식은 정리 2.3(a)와 같다. 다른 점은 거기서는 극한 $F(x\pm\varepsilon)$이 정확한 값이어서 ε을 아무렇게나 골라도 되었지만, 여기서는 $P(Y_n\le x\pm\varepsilon)$의 극한을 쓰므로 $x\pm\varepsilon$이 연속점이어야 한다는 것이다.

5절

연습 5.1 문제 25쪽

$EX_{ij}X_{il}=P$(범주 j이고 l) $=0$ $(j\ne l)$이므로 $\mathrm{Cov}=-p_jp_l$, $\mathrm{Var}\,X_{ij}=p_j-p_j^2$. 계수: $\Sigma_{\mathbf{p}}\mathbf{1}=\mathbf{0}$

이므로 계수는 $k-1$ 이하다. $\mathbf{v}\perp\mathbf{1}$이고 $\mathbf{v}\neq\mathbf{0}$이면 $\mathbf{v}'\Sigma_{\mathbf{p}}\mathbf{v}=\sum p_j v_j^2-(\sum p_j v_j)^2$은 확률 p_j로 값 v_j를 갖는 변수의 분산이다. v_j가 모두 같으면 $\sum v_j=0$에서 $\mathbf{v}=\mathbf{0}$이므로, 그렇지 않은 한 분산이 양수다($p_j>0$). 따라서 $\mathbf{1}^\perp$ 위에서 양의 정부호이고 계수는 $k-1$.

연습 5.2 문제 25쪽

$\mathbf{Y}_n=(\bar X,\bar Y)$, $g(a,b)=b/a$, $\nabla g=(-b/a^2,\ 1/a)=\frac{1}{\mu_X}(-R,\ 1)$. 정리 5.4로 분산은 $\frac{1}{\mu_X^2}(R^2\sigma_X^2-2R\sigma_{XY}+\sigma_Y^2)$. 노트북의 예($\mu_X=2$, $\mu_Y=3$, $\sigma_X=1$, $\sigma_Y=1.5$, $\rho=0.6$)에서 이론 0.450, 시뮬레이션 0.450.

연습 5.3 문제 25쪽

$g(r)=\frac{1}{2}\log\frac{1+r}{1-r}$, $g'(r)=\frac{1}{1-r^2}$이므로 점근분산은 $\frac{(1-\rho^2)^2}{(1-\rho^2)^2}=1$.

연습 5.4 문제 25쪽

$\mathbf{Y}_n=(\bar X,\frac{1}{n}\sum X_i^2)$, $\theta=(\mu,\mu^2+\sigma^2)$, $g(a,b)=\sqrt{b-a^2}/a$. 정규이면 더 쉽게 $(\bar X,S)$로 쓴다: 둘은 독립이고(6회 정리 2.2) $\sqrt{n}(S-\sigma)\xrightarrow{D}\mathcal{N}(0,\sigma^2/2)$(예제 5.5에 $\sqrt{\cdot}$의 델타 방법, $\frac{2\sigma^4}{(2\sigma)^2}=\frac{\sigma^2}{2}$). $g(m,s)=s/m$, $\nabla g=(-s/m^2,1/m)$이므로 분산은 $\frac{\sigma^2}{\mu^4}\sigma^2+\frac{1}{\mu^2}\frac{\sigma^2}{2}=c^4+\frac{c^2}{2}$.

연습 5.5 문제 25쪽

(a) Y_i는 iid 벡터 $\mathbf{X}_i$의 같은 함수이므로 iid이다. $EY_i=\mathbf{t}'(E\mathbf{X}_i-\boldsymbol{\mu})=0$이고 $\operatorname{Var}Y_i=\mathbf{t}'\Sigma\mathbf{t}$이다(5회 보조정리 3.1). (b) $\mathbf{t}'\sqrt{n}(\bar{\mathbf{X}}_n-\boldsymbol{\mu})=\frac{1}{\sqrt{n}}\sum_i Y_i$이므로 $M_{\sqrt{n}(\bar{\mathbf{X}}_n-\boldsymbol{\mu})}(\mathbf{t})=E\exp(\frac{1}{\sqrt{n}}\sum Y_i)=\prod_i Ee^{Y_i/\sqrt{n}}=\left[M_Y(1/\sqrt{n})\right]^n$이다. 둘째 등호에서 독립(3회 정리 3.4)을 썼다. $\mathbf{t}$가 $\mathbf{0}$ 근방에 있으면 $M_Y(s)=e^{-s\mathbf{t}'\boldsymbol{\mu}}M_{\mathbf{X}}(s\mathbf{t})$가 $|s|\le 1$에서 존재한다. (c) $\mathbf{t}'\Sigma\mathbf{t}>0$이면 $Y_i/\sqrt{\mathbf{t}'\Sigma\mathbf{t}}$는 평균 0, 분산 1이고, 정리 3.1의 증명 ④⑤를 $t=\sqrt{\mathbf{t}'\Sigma\mathbf{t}}$에서 쓰면 $n\log M_Y(1/\sqrt{n})\to\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}$이다. $\mathbf{t}'\Sigma\mathbf{t}=0$이면 $\operatorname{Var}Y_i=0$이므로 $Y_i=0$이고(2회 따름정리 5.6) mgf는 모든 n에서 $1=e^0$이다. (d) 극한 $\exp(\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t})$은 $\mathbf{0}$ 근방의 모든 $\mathbf{t}$에서 $\mathcal{N}_k(\mathbf{0},\Sigma)$의 mgf(5회 정의 3.4)와 같다. 2.3절의 상자는 벡터의 mgf에도 성립하므로 $\sqrt{n}(\bar{\mathbf{X}}_n-\boldsymbol{\mu})\xrightarrow{D}\mathcal{N}_k(\mathbf{0},\Sigma)$이다.

연습 5.6 문제 26쪽

(a) 미분가능성의 정의는 $\frac{g(\mathbf{y})-g(\theta)-\nabla g'(\mathbf{y}-\theta)}{\|\mathbf{y}-\theta\|}\to 0$ ($\mathbf{y}\to\theta$)이다. 이 몫을 $r(\mathbf{y})$ ($\mathbf{y}\neq\theta$), $r(\theta)=0$으로 두면 식이 성립하고 r은 θ에서 연속이다. (b) $\mathbf{V}_n=\sqrt{n}(\mathbf{Y}_n-\theta)\xrightarrow{D}\mathbf{V}\sim\mathcal{N}_k(\mathbf{0},\Sigma)$이다. $\mathbf{v}\mapsto\nabla g'\mathbf{v}$는 연속이므로 2.3절의 연속사상 상자로 $\nabla g'\mathbf{V}_n\xrightarrow{D}\nabla g'\mathbf{V}$이고, 5회 정리 3.8로 $\nabla g'\mathbf{V}\sim\mathcal{N}(0,\nabla g'\Sigma\nabla g)$이다. (c) $\|\mathbf{V}_n\|$은 같은 상자로 $\|\mathbf{V}\|$로 분포수렴하므로 확률적으로 유계다(정리 2.5). 그러면 $\mathbf{Y}_n-\theta=\mathbf{V}_n/\sqrt{n}\xrightarrow{P}\mathbf{0}$이고, r이 θ에서 연속이므로 $r(\mathbf{Y}_n)\xrightarrow{P}0$이다(정리 1.3의 벡터 판). 정리 2.5의 둘째 주장으로 $\|\mathbf{V}_n\|\,r(\mathbf{Y}_n)\xrightarrow{P}0$이다. (d) $\sqrt{n}(g(\mathbf{Y}_n)-g(\theta))=\nabla g'\mathbf{V}_n+\|\mathbf{V}_n\|\,r(\mathbf{Y}_n)$

이고, (b)(c)와 보조정리 4.1로 결론이 나온다.

6절

연습 6.1 문제 30쪽

$P(U_1^2+U_2^2 \le 1) = \frac{\pi}{4}$ 이므로 분산은 $16\cdot\frac{\pi}{4}(1-\frac{\pi}{4}) = 2.70$, $n = 10^4$ 에서 표준오차 $\sqrt{2.70}/100 = 0.016$. Buffon($\ell = d$)은 $2.37/100 = 0.024$ 이므로 원이 낫다. 같은 난수 수로 비교하면(바늘 하나에 난수 2개) 차이는 더 크다.

연습 6.2 문제 30쪽

$$\frac{f(y)}{Mg(y)} = \frac{\sqrt{2/\pi}e^{-y^2/2}}{\sqrt{2e/\pi}\,e^{-y}} = e^{-y^2/2+y-1/2} = e^{-(y-1)^2/2}.$$

연습 6.3 문제 30쪽

최빈값 $\frac{1.7}{7} = 0.243$ 에서 밀도는 2.67이므로 $M = 2.67$, 수락 확률 $1/M = 0.375$. 평균 2.67번 시도에 난수 5.3개. 4회 정리 4.2로는 $\Gamma(2.7)$과 $\Gamma(6.3)$ 난수 둘로 $\frac{G_1}{G_1+G_2}$를 만들면 기각이 없지만 감마 난수 자체가 기각 표집(Marsaglia–Tsang)을 쓴다. 모양 모수가 커지면 밀도가 뾰족해져 균등 제안의 M이 커지므로 감마의 비가 유리하다.

연습 6.4 문제 30쪽

$w = f/g$ 이고 $E_g[h^2w^2] = \int x^4\frac{f(x)^2}{g(x)}dx \propto \int x^4\exp\!\big(-x^2+\frac{x^2}{2\tau^2}\big)dx$. 지수의 계수 $-(1-\frac{1}{2\tau^2})$가 음이어야 적분이 유한하므로 $\tau^2 > \frac{1}{2}$이 필요하다. $\tau^2 \le \frac{1}{2}$이면 무한대다.

연습 6.5 문제 30쪽

$\operatorname{Var}\frac{A+B}{2} = \frac{1}{4}(\operatorname{Var}A + \operatorname{Var}B + 2\operatorname{Cov}(A,B)) = \frac{1}{2}(\sigma_h^2 + \operatorname{Cov}(A,B))$ ($A = h(U)$, $B = h(1-U)$, 같은 분포). U, V가 독립이고 h가 증가하면 $h(U)-h(V)$와 $h(1-U)-h(1-V)$는 부호가 반대이므로 곱의 기댓값 ≤ 0이다. 전개하면 $E[\cdot] = 2\big(E[h(U)h(1-U)] - Eh(U)\,Eh(1-U)\big) = 2\operatorname{Cov}(A,B)$ 이므로 $\operatorname{Cov}(A,B) \le 0$. e^{x^2}에서 분산은 0.1126에서 0.0280으로 줄었다.

연습 6.6 문제 31쪽

(a) $E_g[h(Y)w(Y)] = \int_{g>0} h(y)\frac{f(y)}{g(y)}g(y)\,dy = \int_{g>0} hf$ 이다. 가정으로 $g = 0$인 곳에서는 $hf = 0$ 이므로 이것은 $\int hf = \theta$이다. $g(x) = 0$이면서 $h(x)f(x) \neq 0$인 구간이 있으면 마지막 등호가 틀린다. 그 구간의 기여가 빠지기 때문이다. (b) $h(Y_i)w(Y_i)$들은 iid이므로(3회 정리 3.4) $\operatorname{Var}\hat\theta_n^{\mathrm{IS}} = \frac{1}{n}\operatorname{Var}_g(hw)$이다(2회 정리 2.8). $\operatorname{Var}_g(hw) = E_g[h^2w^2] - (E_g[hw])^2 = E_g[h^2w^2] - \theta^2$이다(2회 정리

2.2). (c) $g = hf/\theta$이면 $g \geq 0$이고 $\int g = 1$이다. $g(y) > 0$인 y에서 $h(y)w(y) = h(y)\frac{f(y)\theta}{h(y)f(y)} = \theta$이므로 hw는 상수이고 분산이 0이다. (d) $g = hf/\theta$를 쓰려면 정규화 상수 θ를 알아야 하는데, θ는 구하려던 값이다. 이 정리는 쓸 g를 주지 않고, g를 hf와 비슷한 모양으로 고르라는 방향을 준다.

7절

연습 7.1 문제 35쪽

$EX^k = \int_1^\infty \alpha x^{k-\alpha-1}dx$는 $k-\alpha-1 < -1$, 곧 $k < \alpha$일 때만 유한하고 그때 $\frac{\alpha}{\alpha-k}$이다. $k=1$이면 $\frac{\alpha}{\alpha-1}$. $F(x) = 1 - x^{-\alpha}$이므로 $F^{-1}(u) = (1-u)^{-1/\alpha}$이고 $1-U \sim U$이므로 $X = U^{-1/\alpha}$.

연습 7.2 문제 35쪽

분포의 절반이 1.59 아래에 있고 평균 3은 드문 큰 값이 끌어올린 것이다. 표본 n개의 평균은 그 큰 값을 **충분한 빈도로** 담아야 3 근처가 되는데, 큰 값일수록 드물어 대부분의 표본은 덜 담는다. 그래서 표본평균의 분포는 중앙값이 3보다 작고 오른쪽 꼬리가 긴 모양이며, 평균이 3이 되는 것은 그 긴 꼬리 덕분이다.

연습 7.3 문제 35쪽

Cauchy는 $f(0) = \frac{1}{\pi}$이므로 $\frac{1}{4n/\pi^2} = \frac{\pi^2}{4n}$. 정규는 $f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$이므로 $\frac{2\pi\sigma^2}{4n} = \frac{\pi\sigma^2}{2n}$. 평균의 분산 $\frac{\sigma^2}{n}$과의 비가 $\frac{2}{\pi} = 0.64$이다. 정규에서는 중앙값이 평균만큼 정확하려면 관측이 약 1.57배 필요하다.

연습 7.4 문제 35쪽

$z = 3$에서 오른쪽 꼬리의 비는 $n = 30, 300, 3000$에서 3.08, 1.56, 1.17이고 왼쪽은 0.06, 0.56, 0.85다. 로그 비가 대략 $\frac{\kappa_3 z^3}{6\sqrt{n}}$에 비례하므로(연습 3.1) n을 10배 하면 로그 비가 $\sqrt{10}$분의 1로 준다. 꼬리가 깊을수록(z가 클수록) 더 느리다.

8절

연습 8.1 문제 38쪽

$\mathrm{Cov}(I_i, I_j) = \frac{B(B-1)}{N(N-1)} - \frac{B^2}{N^2} = \frac{B[N(B-1)-B(N-1)]}{N^2(N-1)} = \frac{B(B-N)}{N^2(N-1)}$.

연습 8.2 문제 38쪽

B가 32배이므로 $\eta = 3.2$. 한 에폭의 단계 수는 $\frac{1}{32}$이다. 실제로는 처음 몇 에폭 동안 학습률을 서서히 올린다(warm-up). 큰 학습률에서는 $\eta\Sigma/B$의 정규 근사와 1차 전개가 함께 무너지기 때문이다.

연습 8.3 문제 38쪽

$s/\sqrt{T} \le 0.05s$에서 $T \ge 400$.

연습 8.4 문제 38쪽

$A = Q\Lambda Q'$이면 $\mathbf{z}'A\mathbf{z} = \mathbf{w}'\Lambda\mathbf{w}$, $\mathbf{w} = Q'\mathbf{z} \sim \mathcal{N}(\mathbf{0}, Q'Q) = \mathcal{N}(\mathbf{0}, I)$(5회 정리 3.8). 따라서 $\sum \lambda_i W_i^2$이고 W_i^2은 독립인 $\chi^2(1)$(분산 2)이므로 분산은 $2\sum \lambda_i^2 = 2\,\mathrm{tr}(A^2) = 2\|A\|_F^2$.

연습 8.5 문제 38쪽

m개 평균의 표준오차는 $\sqrt{2(\|A\|_F^2 - \sum a_{ii}^2)/m}$이므로 이것이 $0.01|\mathrm{tr}A|$ 이하인 조건이 식이다. $A = I_d$이면 비대각 성분이 없어 분산이 0이고 $m = 1$로 정확하다. 고윳값이 모두 λ 근처이면 $\mathrm{tr}A \approx d\lambda$, $\|A\|_F^2 \approx d\lambda^2$이므로 $m \gtrsim \frac{2d\lambda^2}{10^{-4}d^2\lambda^2} = \frac{2\times 10^4}{d}$로, 차원이 클수록 탐침이 **적게** 필요하다. 대각합은 많은 항의 합이어서 상대오차가 저절로 줄기 때문이다.