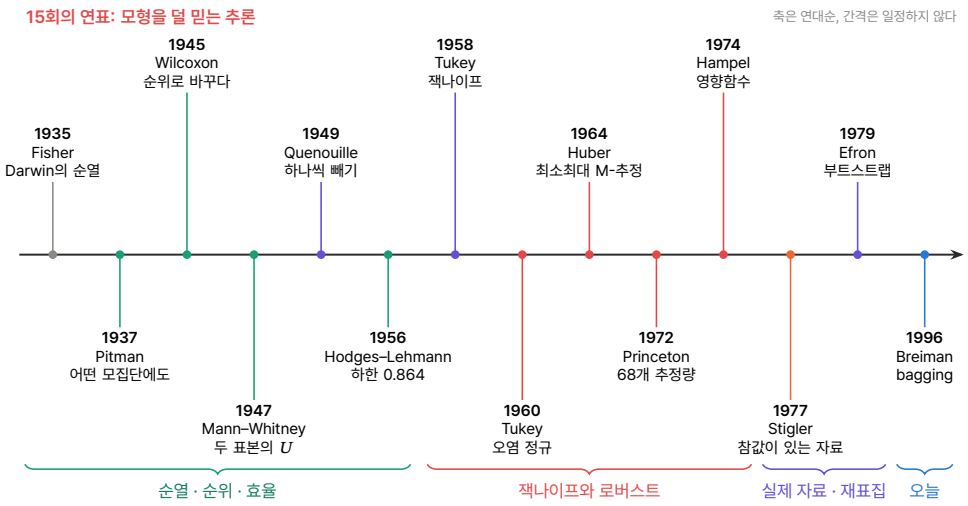
수리통계학 × 통계학의 역사

15회 강의노트

가정이 틀리면
순위 · 로버스트 · 부트스트랩

1935 → 1979

정규분포는 계산을 쉽게 하지만, 실제 측정에는 큰 오차가 섞이고 분포의 꼴은 아무도 모른다. 20세기 중반의 통계학은 그 두려움에 세 가지로 답했다. 값 대신 순위를 쓰는 검정, 큰 오차 하나에 무너지지 않는 추정, 그리고 분포를 자료 자신으로 바꾸는 부트스트랩이다. 이 노트는 세 답의 정리를 증명하고, 참값이 알려진 18–19세기의 측정으로 그것들이 실제로 무엇을 고쳤고 무엇을 고치지 못했는지 본다.



다루는 것	순열검정과 교환가능성 · Wilcoxon의 순위합·부호순위, Mann–Whitney · Hodges–Lehmann 추정 · Pitman 효율과 하한 0.864 · M-추정과 Huber · 영향함수와 붕괴점 · Stigler(1977)의 재현 · 경험분포와 plug-in · 부트스트랩과 BCa · IRLS 로버스트 회귀 · bagging과 Huber 손실
원자료	Darwin의 옥수수(1876) · Cavendish의 지구 밀도(1798) · Michelson의 빛의 속도(1879) · Newcomb의 통과 시간(1882) · 질산 공장의 stack loss(1965)
노트북	`sessions/15_robust_bootstrap/15_robust_bootstrap.ipynb` — 절·정리 번호가 같다
워크북	`15_robust_bootstrap_workbook.pdf` (손으로 푸는 35문항) · 풀이판 `_workbook_sol.pdf`
선수 지식	1–14회 (특히 7회 중심극한정리·델타 방법, 8회 순서통계량·순위, 9회 MLE의 점근정규성, 13회 최소제곱, 14회 무작위화 검정)

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 정의)를 단다. 읽을 때는 근거를 가리고 "왜 이 단계가 성립하는가"를 먼저 말해 본다. 극한과 적분의 교환이나 이 과정에서 증명하지 않는 부분은 **조건 확인** 단락에 따로 밝힌다. 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, [T3] 귀납, [T4] 부등식, [T5] 극한, [T6] 극한과 적분의 교환, [T7] 변수변환, [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 1–14회에서 가져오는 것 2

0.2 표기 3

1 순열과 순위 (1935–1956) 4

1.1 교환가능성과 순열검정 4

1.2 순위합: Wilcoxon과 Mann–Whitney 5

1.3 부호순위와 Hodges–Lehmann 추정 7

2 효율 — 순위로 바꾸면 얼마나 잃는가 9

2.1 Pitman 효율 9

2.2 하한 0.864 11

3 로버스트 추정 (1960–1974) 13

3.1 M-추정 13

3.2 영향함수와 붕괴점 15

4 참값이 있는 자료로 — Stigler (1977) 18

5 부트스트랩 (1949–1979) 20

5.1 plug-in과 부트스트랩 20

5.2 더 나은 구간: BCa 22

5.3 잭나이프 23

6 원자료로 — 회귀도 로버스트하게 25

7 부숴보기 — 분포를 덜 믿어도 깨지는 곳 28

7.1 부트스트랩은 최댓값에서 실패한다 28

7.2 순위합은 중앙값의 검정이 아니다 . . . 28
7.3 로버스트 추정량은 무엇을 추정하는가 . . . 29

8 오늘 — bagging과 Huber 손실 30

8.1 Bagging과 random forest . . . 30
8.2 Huber 손실과 smooth L1 . . . 30

9 정리와 역사 노트 33

9.1 한 장 요약 . . . 33
9.2 Stigler의 명명 법칙 . . . 34
9.3 7 Pillars 지도 . . . 34
9.4 다음 회 — 베이즈의 귀환: MCMC와 Stein의 역설 (1763 → 1990) . . . 34

A 연습문제 풀이 36

0

들어가며

> **여는 질문.** Cavendish(1798)는 지구의 밀도를, Michelson(1879)은 빛의 속도를, Newcomb(1882)은 빛이 7.4 km를 왕복하는 시간을 쟀다. 오늘 우리는 참값을 안다. 그들의 평균은 얼마나 틀렸나? 큰 오차에 덜 흔들리는 다른 추정량을 썼다면 더 맞았을까? 그리고 정규분포를 믿지 않을 때, 그 추정량의 오차는 무엇으로 재는가?

13회의 t와 F, 14회의 분산분석은 오차가 정규분포를 따른다고 가정했다. 정규분포는 계산을 아름답게 만든다. 그러나 1950년대의 통계학자들은 그 가정이 얼마나 위험한지 하나씩 드러냈다. 측정에는 가끔 엉뚱하게 큰 오차가 섞인다. 11회에서 본 Newcomb의 측정 66개 가운데 두 개가 그랬다. 평균은 그런 오차 하나에 끌려가고, 표준편차는 더 크게 끌려간다. 분포의 꼴을 모르면 t 분포의 임계값도 근거를 잃는다.

통계학은 세 가지로 답했다. 첫째는 **순위**다. Frank Wilcoxon(1945)은 관측값 대신 순위를 쓰면 귀무가설 아래의 분포가 자료의 분포와 무관해진다는 것을 보였다. 둘째는 **로버스트 추정**이다. John Tukey(1960)는 정규분포에 아주 조금만 다른 분포가 섞여도 평균이 최선의 자리를 잃는다는 것을 보였고, Peter Huber(1964)는 그런 섞임에 가장 강한 추정량을 찾았다. 셋째는 **부트스트랩**이다. Bradley Efron(1979)은 모르는 분포를 자료의 경험분포로 바꾸고 컴퓨터로 다시 뽑아, 거의 모든 추정량의 표준오차와 신뢰구간을 구했다.

이 노트는 그 순서를 따른다. 1절은 순열검정과 Wilcoxon의 순위검정, 2절은 그 효율(Pitman, Hodges–Lehmann의 0.864)을 증명한다. 3절은 M-추정, 영향함수, 붕괴점을, 4절은 참값이 알려진 역사적 측정으로 로버스트 추정량을 시험한 Stigler(1977)를 재현한다. 5절은 부트스트랩, 6절은 로버스트 회귀, 7절은 부트스트랩과 순위검정이 실패하는 곳, 8절은 bagging과 신경망의 Huber 손실이다.

학습 목표

1. 교환가능성에서 순열검정의 정확성을 보이고, Wilcoxon 순위합·부호순위 통계량의 귀무분포와 적률을 증명한다.
2. Pitman 효율 $12\sigma^2(\int f^2)^2$을 유도하고, Hodges–Lehmann의 하한 108/125를 증명한다.
3. M-추정량의 점근분산, 영향함수, 붕괴점을 계산하고 Huber 추정량을 구현한다.

4. 경험분포와 plug-in 원리로 부트스트랩을 정의하고, 평균에서 그 일치성을 증명하며 BCa 구간을 계산한다.
5. 참값이 알려진 측정으로 추정량들을 비교하고, 부트스트랩이 실패하는 경우와 bagging·Huber 손실 속의 같은 생각을 안다.

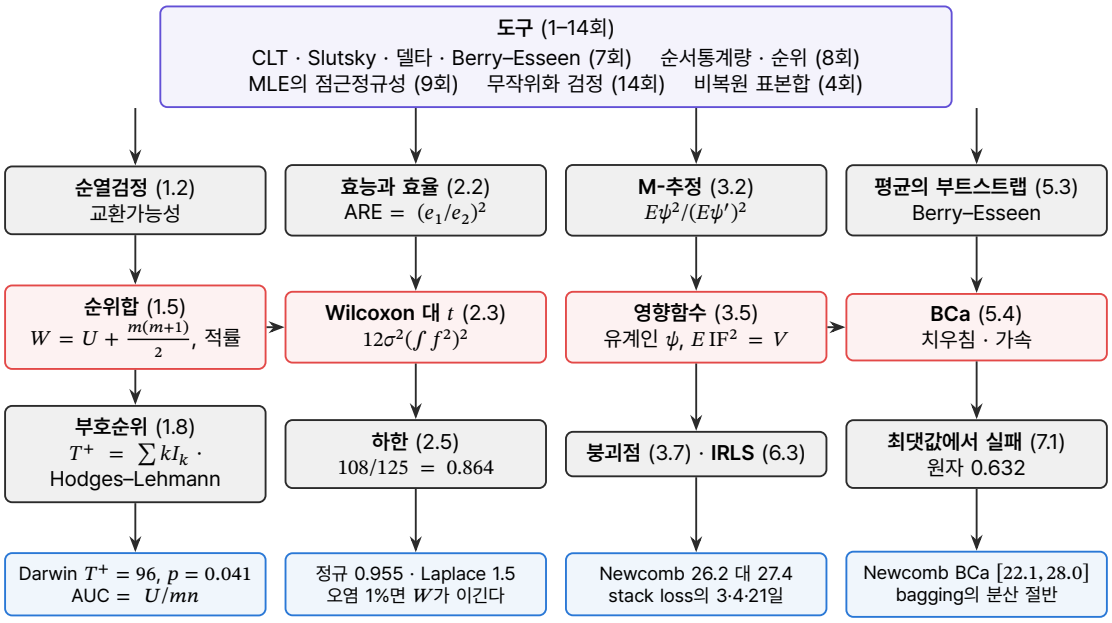


그림 0.1 이번 회차 정리들의 의존 관계. 화살표는 "증명에 쓰인다"는 뜻이다.

0.1 1–14회에서 가져오는 것

앞 회차	내용
3회 · 11회 6절	Michelson·Cavendish의 편향, Newcomb의 큰 오차 두 개
7회 정리 3.1 · 4.4, 3절	중심극한정리, 델타 방법, Berry–Esseen 한계
8회 정리 1.4 · 보조정리 8.1	표본분위수의 점근정규성, 교환가능한 순위
9회 정리 3.1	MLE의 점근정규성 (M-추정으로 일반화한다)
12회 8절	ROC의 AUC는 Mann–Whitney 통계량
14회 정리 1.7 · 예제 1.9	무작위화 검정의 정확성, Darwin의 옥수수

0.2 표기

표본 $X_1, \dots, X_n$은 분포함수 F, 밀도 f에서 온다. **경험분포함수**는 $\hat{F}_n(x) = \frac{1}{n}\#\{i : X_i \le x\}$, 곧 각 관측에 확률 $1/n$을 주는 분포다. 순서통계량은 $X_{(1)} \le \dots \le X_{(n)}$이다. 추정량을 분포의 함수 $T(F)$로 보면(평균 $\int x\,dF$, 중앙값 $F^{-1}(1/2)$) 표본의 추정량은 $T(\hat{F}_n)$이다. 이것을 **plug-in** 추정량이라 한다.

1

순열과 순위 (1935–1956)

> **역사** 14회에서 Fisher(1935)는 Darwin의 옥수수에 무작위화 검정을 썼다. 1937년 E. J. G. Pitman은 같은 생각을 두 표본, 상관, 분산분석으로 넓히고 "어떤 모집단에서 온 표본에도 쓸 수 있는 유의성 검정"이라 불렀다. 그러나 순열을 모두 세는 계산은 손으로 할 수 없었다. 1945년 American Cyanamid의 화학자 Frank Wilcoxon은 한 쪽짜리 논문에서 관측값을 순위로 바꾸면 순열분포가 자료와 무관한 표 하나로 정해진다는 것을 보였다. 1947년 Mann과 Whitney가 두 표본의 경우를 정리했고, Hodges와 Lehmann(1956, 1963)은 이 검정들이 정규분포에서도 거의 손해가 없음을 증명하고 거기서 추정량을 만들었다.

1.1 교환가능성과 순열검정

정의 1.1 (교환가능) 확률벡터 $(Z_1, \dots, Z_N)$의 결합분포가 성분의 순서를 어떻게 바꿔도 같으면 **교환가능**하다고 한다. 독립이고 같은 분포를 따르면 교환가능하다.

정리 1.2 (두 표본 순열검정) $X_1, \dots, X_m \overset{\text{iid}}{\sim} F$, $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} G$가 독립이고 $H_0 : F = G$라 하자. 합친 표본 $\mathbf{Z} = (X_1, \dots, X_m, Y_1, \dots, Y_n)$의 성분을 m개와 n개로 나누는 $\binom{N}{m}$가지 방법 $(N = m + n)$ 위에서 통계량 T의 분포를 만들고, 관측된 나눔의 T가 그 분포에서 차지하는 위쪽 꼬리 확률을 p-값으로 쓰면, H_0 아래 $P(p \le \alpha) \le \alpha$이다.

증명. **전략** 합친 표본의 값들(순서 없는 다중집합)을 조건으로 둔다. 그러면 귀무가설 아래 나눔은 $\binom{N}{m}$가지 위에서 균등하고, 14회 정리 1.7의 상황이 된다. 조건부 한계의 기댓값을 취해 무조건 한계를 얻는다.

① H_0 아래 $\mathbf{Z}$의 N개 성분은 iid이므로 교환가능하다. — 정의 1.1

② 합친 표본의 값들의 다중집합을 $\mathcal{V}$라 하고, 그 값들에 크기 순서로 이름 $1, \dots, N$을 붙인다. 순열로 섞어도 $\mathcal{V}$는 그대로이므로, $\mathcal{V}$가 주어졌을 때 처음 m자리에 오는 이름의 집합은 $\binom{N}{m}$가지 부분집합 위에서 균등하다. — ①, 순열에 불변인 사건 [T1]

③ 조건부로 이것은 배정이 유한집합 위에서 균등하게 뽑히는 14회 정리 1.7의 상황이다. 따라서 $P(p \le \alpha \mid \mathcal{V}) \le \alpha$이다. — 14회 정리 1.7

④ $P(p \le \alpha) = E\big[P(p \le \alpha \mid \mathcal{V})\big] \le \alpha$이다. 반복기댓값, 3회 정리 2.5

■

14회의 무작위화 검정은 확률의 출처가 실험자의 동전이었다. 여기서는 출처가 표본추출과 교환가능성이다. 계산은 같다.

1.2 순위합: Wilcoxon과 Mann–Whitney

보조정리 1.3 (순위는 균등한 순열이다) $Z_1, \dots, Z_N$이 교환가능하고 같은 값이 나올 확률이 0이면, 순위 벡터 $(R_1, \dots, R_N)$ (R_i는 Z_i보다 작거나 같은 성분의 수)는 $\{1, \dots, N\}$의 $N!$개 순열 위에서 균등분포다.

증명. **전략** 순열로 섞은 벡터는 원래 벡터와 같은 분포를 갖고, 그 순위 벡터는 원래 순위 벡터를 같은 순열로 섞은 것이다. 그래서 순위 벡터의 분포는 섞기에 불변이고, 섞기에 불변인 순열 위의 분포는 균등분포뿐이다.

① 동점의 확률이 0이므로 확률 1로 $\mathbf{R} = (R_1, \dots, R_N)$은 $\{1, \dots, N\}$의 순열이다. 가정

② 순열 π에 대해 $\mathbf{Z}^\pi = (Z_{\pi(1)}, \dots, Z_{\pi(N)})$의 순위 벡터는 $(R_{\pi(1)}, \dots, R_{\pi(N)}) = \mathbf{R} \circ \pi$이다. $Z_{\pi(i)}$보다 작거나 같은 성분의 수는 섞기 전과 같기 때문이다. 순위의 정의 [T1]

③ $\mathbf{Z}^\pi$와 $\mathbf{Z}$는 같은 분포이므로 모든 순열 $\mathbf{r}$에 대해 $P(\mathbf{R} = \mathbf{r}) = P(\mathbf{R} \circ \pi = \mathbf{r}) = P(\mathbf{R} = \mathbf{r} \circ \pi^{-1})$이다. 교환가능, ②

④ 순열 $\mathbf{r}, \mathbf{s}$가 주어지면 $\mathbf{s} = \mathbf{r} \circ \pi^{-1}$인 순열 π가 있으므로 $P(\mathbf{R} = \mathbf{r}) = P(\mathbf{R} = \mathbf{s})$이다. 확률은 모두 $1/N!$이다. 순열은 합성에 대해 군 [T1]

(8회 보조정리 8.1은 이 균등분포의 한 성분의 주변분포다.) ■

정의 1.4 (Wilcoxon 순위합과 Mann–Whitney) 합친 표본에서의 순위로 $W = \sum_{i=1}^{m} R(X_i)$를 **Wilcoxon 순위합**, $U = \#\{(i, j) : X_i > Y_j\}$를 **Mann–Whitney 통계량**이라 한다.

정리 1.5 (순위합의 귀무분포) $F = G$가 연속이면

(a) $W = U + m(m+1)/2$이다.

(b) W의 분포는 F와 무관하고, $EW = \dfrac{m(N+1)}{2}$, $\operatorname{Var} W = \dfrac{mn(N+1)}{12}$이며, EW를 중심으로 대칭이다.

(c) 일반적으로(귀무가설과 무관하게) $U/(mn)$은 $P(X > Y)$의 불편추정량이다.

증명. **전략** (a)는 X의 순위를 "자기보다 작은 X의 수 + 자기보다 작은 Y의 수 + 1"로 쪼개 센다. (b)는 보조정리 1.3로 W를 $\{1,\dots,N\}$에서 비복원으로 뽑은 m개의 합으로 바꾼다. (c)는 지시변수의 합의 기댓값이다.

(a)

① X 표본에서 k번째로 작은 값 $X_{(k)}$의 합친 순위는 $(k-1)+\#\{j : Y_j < X_{(k)}\}+1$이다.

순위의 정의, 동점 없음

② $k=1,\dots,m$에 대해 더하면 $W=\sum_{k=1}^{m} k+\sum_{k}\#\{j : Y_j<X_{(k)}\}=\frac{m(m+1)}{2}+U$이다.

정의 1.4

(b)

① $(R(X_1),\dots,R(X_m))$은 $\{1,\dots,N\}$에서 m개를 비복원으로 뽑은 것과 같은 분포이고, 이 분포는 F와 무관하다.

보조정리 1.3의 주변분포

② 모집단 $\{1,\dots,N\}$의 평균은 $\frac{N+1}{2}$, 분산(분모 N)은 $\frac{N^2-1}{12}$이다.

$\sum k^2=\frac{N(N+1)(2N+1)}{6}$ [T3]

③ 비복원 표본합의 공식으로

$$EW=m\cdot\frac{N+1}{2},\qquad \operatorname{Var}W=m\cdot\frac{N^2-1}{12}\cdot\frac{N-m}{N-1}=\frac{mn(N+1)}{12}.$$

4회 정리 1.5, 14회 보조정리 1.3

④ 순위 r을 $N+1-r$로 바꾸는 대응은 순열 위의 균등분포를 보존하고 W를 $m(N+1)-W$로 보낸다. 따라서 둘은 같은 분포이고 W는 EW에 대해 대칭이다. ①

(c)

① $U=\sum_{i,j}\mathbf{1}\{X_i>Y_j\}$이므로 $EU=\sum_{i,j}P(X_i>Y_j)=mn\,P(X>Y)$이다.

기댓값의 선형성, 2회

■

$U/(mn)$은 12회 8절의 AUC, 곧 "양성 하나와 음성 하나를 뽑았을 때 양성의 점수가 더 클 확률"의 추정량이다. 분류기의 AUC를 보고하는 것은 두 점수 분포의 Mann–Whitney 통계량을 보고하는 것이다. 정확한 귀무분포는 W의 값마다 그 값을 갖는 부분집합의 수를 세는 점화식으로 계산한다(연습 1.2). 표본이 크면 정규근사를 쓴다.

예제 1.6 (Michelson의 다섯 묶음) Michelson의 1879년 측정 100개는 시간 순서대로 20개씩 다섯 묶음으로 나뉜다(`MichelsonSets`). 첫 묶음(평균 852.5)과 마지막 묶음(839.5)을 Wilcoxon–Mann–Whitney로 비교하면 $U = 225.5$, $U/(mn) = 0.564$, 양쪽 $p = 0.50$이다(t 검정은 0.56). 다섯 묶음 전체의 Kruskal–Wallis 순위검정(분산분석의 순위판)은 $p = 0.93$이다. 묶음 사이에는 차이가 없다. 3회에서 본 117 km/s의 편향은 실험 내내 같은 크기로 있었다.

1.3 부호순위와 Hodges–Lehmann 추정

정의 1.7 (Wilcoxon 부호순위) 짝 차이 $d_1, \dots, d_n$에서 $|d_i|$들의 순위를 R_i^+라 하고 $T^+ = \sum_{d_i>0} R_i^+$를 **부호순위 통계량**이라 한다.

정리 1.8 (부호순위의 귀무분포) d_i들이 독립이고 각각 0에 대해 대칭인 연속분포를 따르면 $T^+ = \sum_{k=1}^{n} k\, I_k$이고 $I_1, \dots, I_n$은 독립인 Bernoulli(1/2)이다. 따라서

$$E T^+ = \frac{n(n+1)}{4}, \qquad \operatorname{Var} T^+ = \frac{n(n+1)(2n+1)}{24}.$$

증명. **전략** 대칭이면 각 차이의 부호는 절댓값과 독립인 공정한 동전이다. 절댓값들이 순위를 정하고, 부호들이 그 순위를 더할지 말지를 정한다.

① d_i와 $-d_i$가 같은 분포이므로 모든 $B \subset (0,\infty)$에서 $P(d_i > 0, |d_i| \in B) = P(d_i < 0, |d_i| \in B) = \frac{1}{2} P(|d_i| \in B)$이다. 곧 부호 $\operatorname{sign}(d_i)$는 $|d_i|$와 독립이고 ± 1이 반반이다.

대칭, $P(d_i = 0) = 0$ [T1]

② d_i들이 독립이므로 부호 n개와 절댓값 n개는 서로 독립이고, 부호들은 iid다. ①, 독립

③ I_k를 "절댓값이 k번째로 작은 차이가 양수"의 지시변수라 하자. 어느 차이가 k번째인지는 절댓값들이 정하므로, 절댓값들이 주어졌을 때 $I_1, \dots, I_n$은 독립인 Bernoulli(1/2)이다. 이 조건부 분포가 절댓값과 무관하므로 무조건으로도 그렇다. ②

④ $T^+ = \sum_{d_i>0} R_i^+ = \sum_{k=1}^{n} k\, I_k$이다. 정의 1.7

⑤ 따라서

$$E T^+ = \frac{1}{2}\sum_{k=1}^{n} k = \frac{n(n+1)}{4}, \qquad \operatorname{Var} T^+ = \frac{1}{4}\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{24}.$$

독립합의 분산, 2회 정리 2.8

■

정의 1.9 (Hodges–Lehmann 추정량) 짝 차이의 **Walsh 평균** $(d_i+d_j)/2$ ($i \le j$, 모두 $n(n+1)/2$개)의 중앙값을 Hodges–Lehmann 추정량이라 한다. 두 표본에서는 모든 $Y_j - X_i$의 중앙값이다.

T^+는 $(d_i + d_j)/2 > 0$인 Walsh 평균의 개수와 같다(연습 1.3). 그래서 "$d_i - \theta$의 부호순위 검정이 기각하지 않는 θ"들의 집합(신뢰구간)은 정렬한 Walsh 평균의 양 끝을 잘라 얻고, 그 중심이 Hodges–Lehmann 추정량이다. 검정을 뒤집어 추정량을 만든 것이다.

예제 1.10 (Darwin의 옥수수, 세 번째) Darwin의 15쌍(14회 예제 1.9)에서 양수인 차이의 순위합은 $T^+ = 96$ (가능한 최대 120)이고, 정확한 양쪽 p-값은 0.041이다. 무작위화 검정 0.0527, t 검정 0.0497과 비슷하다. Walsh 평균 120개의 중앙값인 Hodges–Lehmann 추정은 3.13인치(평균 2.62, 중앙값 3.0)이고, $P(T^+ \le 25) = 0.024$를 써서 Walsh 평균의 26번째와 95번째 값으로 만든 95.2% 구간은 $[0.50,\ 5.19]$다. 음수인 두 차이(-8.4, -6.0)가 평균을 끌어내리지만, 부호순위는 그것들을 크기 대신 절댓값의 순위 14와 10으로만 센다($T^- = 24$).

연습문제

연습 1.1 (작은 표의 귀무분포 ★) $m = 2$, $n = 3$이면 $\binom{5}{2} = 10$가지 순위 집합의 W를 모두 써서 귀무분포를 만들고, $EW = 6$, $\mathrm{Var}\, W = 3$을 확인하라. 풀이 36쪽

연습 1.2 (점화식 ★★) $N_{m,n}(w)$를 $\{1, \dots, m+n\}$에서 고른 m개의 합이 w인 경우의 수라 하면 $N_{m,n}(w) = N_{m-1,n}(w - m - n) + N_{m,n-1}(w)$임을 보여라(가장 큰 순위 $m+n$이 X의 것인지 아닌지로 나눈다). 풀이 36쪽

연습 1.3 (Walsh 평균 ★★) 동점이 없으면 $T^+ = \#\{(i,j) : i \le j,\ d_i + d_j > 0\}$임을 보여라. ($|d_i|$의 순위가 k인 양수 d_i는 자기 자신을 포함해 절댓값이 자기 이하인 k개의 d_j와 짝을 이룰 때 합이 양수다.) 풀이 36쪽

연습 1.4 (AUC의 분산 ★) H_0 아래 $U/(mn)$의 분산이 $\frac{N+1}{12mn}$임을 정리 1.5로 보여라. 양성과 음성이 각각 100개인 분류기의 AUC가 0.5에서 얼마나 멀어야 5%에서 "우연보다 낫다"고 할 수 있는가? 풀이 36쪽

2

효율 — 순위로 바꾸면 얼마나 잃는가

> **역사** 순위 검정이 처음 나왔을 때 많은 통계학자는 그것을 "빠르지만 거친" 방법으로 여겼다. 값을 버리고 순서만 쓰니 정보를 잃는다는 것이다. 1948년 Pitman은 콜럼비아 대학 강의록에서 두 검정을 비교하는 기준을 내놓았다. 같은 대립가설에 같은 검정력을 얻는 데 필요한 표본 크기의 비다. Hodges와 Lehmann(1956)은 그 기준으로 Wilcoxon 검정이 정규분포에서조차 t 검정의 95.5%의 효율을 갖고, 어떤 분포에서도 86.4%보다 못하지 않으며, 꼬리가 두꺼우면 얼마든지 나을 수 있음을 증명했다. "빠르고 거친" 방법이 사실상 공짜였던 것이다.

2.1 Pitman 효율

정의 2.1 (점근상대효율) 같은 가설을 검정하는 두 검정이 있고, 대립가설 θ가 0에 다가갈 때 같은 크기와 같은 검정력에 필요한 표본 크기를 $N_1(\theta)$, $N_2(\theta)$라 하자. $\lim_{\theta\to 0} N_2/N_1$을 검정 1의 검정 2에 대한 **Pitman 점근상대효율**(ARE)이라 한다. ARE가 1.5면 검정 1은 같은 일을 표본 2/3로 한다.

증명 없이 쓰는 사실 (국소 대립가설 아래의 점근정규성) $\theta_N = h/\sqrt{N}$일 때 $\sqrt{N}(T_N - \mu(0))/\sigma(0) \xrightarrow{D} \mathcal{N}\big(h\,\mu'(0)/\sigma(0),\ 1\big)$이 성립한다고 하자. 두 표본 평균의 차에서는 중심극한정리(7회 정리 3.1)로 바로 나오고, Mann–Whitney 통계량에서는 U-통계량의 사영(Hoeffding 1948)과 대립가설의 "근접성"(Le Cam)이 필요하다. 이 과정에서는 그것을 증명 없이 쓴다.

명제 2.2 (효능과 효율) 위의 조건 아래 크기 α의 한쪽 검정의 검정력은 $1 - \Phi\big(z_\alpha - h\,e\big)$로 수렴한다. $e = \mu'(0)/\sigma(0)$을 **효능**이라 한다. 두 검정의 ARE는 효능의 비의 제곱 $(e_1/e_2)^2$이다.

증명. **전략** 기각역을 표준화한 통계량으로 쓰고 위 상자의 극한분포를 넣어 검정력의 극한을 얻는다. 두 검정의 검정력이 같아지는 조건을 표본 크기에 대해 푼다. 단계별 증명은 연습 2.5(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 37쪽에 있다.

정리 2.3 (Wilcoxon 대 t) $X \sim F$, $Y \sim F(\cdot - \theta)$, F의 밀도 f와 분산 σ^2이 있고, $m/N \to \lambda \in (0,1)$이라 하자. Wilcoxon–Mann–Whitney 검정의 두 표본 t 검정에 대한 ARE는

$$\mathrm{ARE}(W, t) = 12\,\sigma^2 \Big(\int f(x)^2\,dx \Big)^2 .$$

증명. **전략** 명제 2.2로 ARE는 효능의 비의 제곱이다. 두 통계량마다 평균의 θ에 대한 도함수와 귀무가설 아래의 표준편차($\sqrt{N}$을 곱한 것)를 구한다.

① t: $T = \bar{Y} - \bar{X}$의 평균은 θ이고 분산은 $\sigma^2(\frac{1}{m} + \frac{1}{n}) = \frac{\sigma^2}{N\lambda(1-\lambda)}$이다. 따라서 $\mu'(0) = 1$, $\sigma(0) = \sigma/\sqrt{\lambda(1-\lambda)}$, $e_t = \sqrt{\lambda(1-\lambda)}/\sigma$이다. — 2회 정리 2.8, $m/N \to \lambda$

② Wilcoxon: $T = U/(mn)$의 평균은 $p(\theta) = P(Y > X) = \int F(x+\theta) f(x)\,dx$이다. — 정리 1.5(c), 반복기댓값

③ $p'(0) = \int f(x)^2\,dx$이다. — 적분 기호 아래 미분 [T6]

④ 귀무가설 아래 $\mathrm{Var}\,T = \frac{N+1}{12mn} \approx \frac{1}{12N\lambda(1-\lambda)}$이므로 $\sigma(0) = 1/\sqrt{12\lambda(1-\lambda)}$, $e_W = \sqrt{12\lambda(1-\lambda)} \int f^2$이다. — 정리 1.5(b), 연습 1.4

⑤ $(e_W/e_t)^2 = 12\,\sigma^2 \big(\int f^2 \big)^2$이다. $\lambda(1-\lambda)$는 지워진다. — 명제 2.2

조건 확인 ③에서 $\frac{F(x+\theta)-F(x)}{\theta}$는 f의 구간 $[x, x+\theta]$ 위의 평균이다. f가 유계($\le M$)이면 차분몫과 $f(x)$의 곱은 적분가능한 $Mf(x)$로 눌리고, f의 연속점에서 $f(x)^2$로 수렴하므로 지배수렴으로 미분과 적분을 바꿀 수 있다 [T6]. 명제 2.2는 위의 상자(국소 대립가설 아래의 점근정규성)를 가정한다. Mann–Whitney 통계량에 대한 그 사실은 이 과정에서 증명하지 않는다. ∎

예제 2.4 (분포에 따른 효율)

분포	포물선	균등	정규	로지스틱	t_5	Laplace	t_3
$\mathrm{ARE}(W,t)$	0.864	1	$3/\pi = 0.955$	$\pi^2/9 = 1.097$	1.24	1.5	1.90

정규분포에 표준편차가 3배인 정규를 ε만큼 섞으면(Tukey의 오염 정규) ARE는 $\varepsilon = 0.01$에서 1.01, 0.05에서 1.20, 0.10에서 1.37이다. 오염이 1%만 되어도 Wilcoxon이 t를 앞선다. Cauchy처럼 분산이 없으면 t 검정은 근거를 잃고 ARE는 무한대다.

2.2 하한 0.864

정리 2.5 (Hodges–Lehmann (1956)) 분산이 유한한 모든 밀도 f에 대해 $\mathrm{ARE}(W,t) \geq 108/125 = 0.864$이다. 등호는 포물선 밀도 $f(x) \propto (5\sigma^2 - x^2)_+$일 때다.

증명. **전략** ARE는 위치와 척도에 무관하므로 평균 0, 분산 1인 밀도에서 $\int f^2$의 최솟값을 찾으면 된다. 후보 h와의 차의 제곱 $(f-h)^2 \geq 0$을 적분해 $\int f^2$를 아래에서 누른다 [T4].

① f를 $\frac{1}{s}f(\frac{x-a}{s})$로 바꾸면 σ^2은 s^2배, $\int f^2$은 $\frac{1}{s}$배가 되어 $12\sigma^2(\int f^2)^2$은 그대로다. 그래서 평균 0, 분산 1로 놓는다. 치환 [T7]

② $h(x) = c\,(5 - x^2)_+$, $c = \frac{3}{20\sqrt{5}}$로 두면 $\int h = 1$, $\int x^2 h = 1$이다. $a = 5c$, $b = c$로 쓰면 받침 위에서 $h = a - bx^2$이다. 직접 적분, 연습 2.2

③ $0 \leq \int (f-h)^2 = \int f^2 - 2\int fh + \int h^2$이므로 $\int f^2 \geq 2\int fh - \int h^2$이다. 전개 [T4]

④ 모든 x에서 $h(x) \geq a - bx^2$이고(받침 밖에서 오른쪽이 음수) $f \geq 0$이므로 $\int fh \geq \int f\,(a - bx^2) = a - b$이다. $\int f = 1$, $\int x^2 f = 1$

⑤ h의 받침 위에서 $h = a - bx^2$이므로 $\int h^2 = \int h\,(a - bx^2) = a - b = 4c = \frac{3\sqrt{5}}{25}$이다. ②

⑥ ③–⑤로 $\int f^2 \geq 2(a-b) - (a-b) = \frac{3\sqrt{5}}{25}$이고, $12(\int f^2)^2 \geq 12 \cdot \frac{45}{625} = \frac{108}{125}$이다. 등호는 $\int (f-h)^2 = 0$, 곧 $f = h$일 때다. ③의 등호 조건

■

정리 2.5는 "보험료"를 알려 준다. Wilcoxon을 쓰면 최악의 경우에도 표본의 13.6%만큼 손해고, 정규에서는 4.5%이며, 꼬리가 두꺼우면 큰 이득이다. Chernoff와 Savage(1958)는 순위 대신 정규분포의 분위수(정규 점수)를 쓰는 검정이 모든 분포에서 t 이상의 효율을 가짐을 보였다.

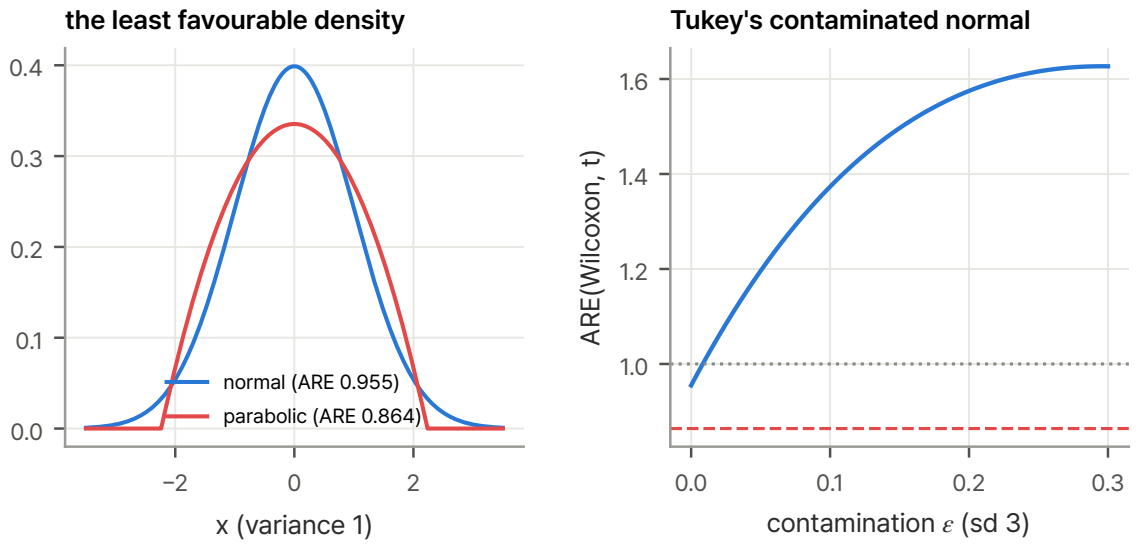


그림 2.1 왼쪽: 하한을 이루는 포물선 밀도와 같은 분산의 정규밀도. 오른쪽: 오염 비율 ε (표준편차 3배의 정규를 섞음)에 따른 Wilcoxon의 t에 대한 효율(정리 2.3). 회색 점선은 1, 빨간 점선은 하한 0.864다. $\varepsilon \approx 0.01$에서 1을 넘는다.

연습문제

연습 2.1 (세 분포의 효율 ★) 정규에서 $\int \phi^2 = 1/(2\sqrt{\pi})$, Laplace $\frac{1}{2}e^{-|x|}$에서 $\sigma^2 = 2$, $\int f^2 = 1/4$, 로지스틱에서 $\sigma^2 = \pi^2/3$, $\int f^2 = 1/6$임을 보이고 표의 $3/\pi$, 1.5, $\pi^2/9$를 얻어라. 풀이 37쪽

연습 2.2 (포물선 밀도 ★★) 정리 2.5의 h에서 $\int h = 1$, $\int x^2 h = 1$, $\int h^2 = 3\sqrt{5}/25$를 직접 계산하라. 풀이 37쪽

연습 2.3 (효능에서 표본 크기로 ★) 정규 자료에서 t 검정이 100쌍으로 얻는 검정력을 Wilcoxon으로 얻으려면 대략 몇 쌍이 필요한가? Laplace 자료에서는? 풀이 37쪽

연습 2.4 (부호검정 ★★) 한 표본 위치 문제에서 양수의 개수로 하는 부호검정의 효능이 $2f(0)$, t 검정의 효능이 $1/\sigma$임을 보이고 ARE $4\sigma^2 f(0)^2$을 구하라. 정규분포에서는 $2/\pi = 0.637$이다. (8회 정리 1.4의 표본 중앙값의 분산과 비교하라.) 풀이 37쪽

연습 2.5 (효능과 효율의 증명 ★★ 유도형) 명제 2.2를 다음 순서로 증명하라. $\theta = h/\sqrt{N}$이다.

(a) 크기 α의 한쪽 검정의 기각역이 $Z_N = \sqrt{N}(T_N - \mu(0))/\sigma(0) > z_\alpha$임을 적고, 위 상자의 가정으로 $Z_N \xrightarrow{D} \mathcal{N}(he, 1)$임을 확인하라.

(b) 극한분포가 연속이므로 $P(Z_N > z_\alpha) \to 1 - \Phi(z_\alpha - he)$임을 보여라. 분포수렴의 정의에서 어느 점의 연속성을 쓰는가?

(c) 같은 고정된 θ에서 $h = \theta\sqrt{N}$이다. 두 검정의 극한 검정력이 같으려면 $\theta\sqrt{N_1}e_1 = \theta\sqrt{N_2}e_2$이어야 함을 보이고 $N_2/N_1 = (e_1/e_2)^2$을 얻어라.

풀이 37쪽

3

로버스트 추정 (1960–1974)

> **역사** 1960년 Tukey는 정규분포에 표준편차가 세 배인 정규를 조금 섞은 분포에서 여러 추정량을 비교했다. 섞는 비율이 1%만 되어도 표본평균은 최선의 자리를 잃었다(11회 7절의 $\varepsilon^* = 0.0018$은 그 극단이다). 1964년 취리히의 Peter Huber는 **위치모수의 로버스트 추정**에서 질문을 뒤집었다. 오염의 비율만 알고 꼴은 모를 때, 최악의 경우의 분산이 가장 작은 추정량은 무엇인가? 답은 가운데서는 평균처럼, 바깥에서는 중앙값처럼 행동하는 추정량이었다. Frank Hampel(1968, 1974)은 관측 하나가 추정량을 얼마나 움직이는지 재는 영향함수와, 추정량이 무너지기 전에 견딜 수 있는 오염의 비율인 붕괴점을 정의했다. 1972년 Princeton의 대규모 모의실험은 68개의 추정량을 비교했고, Robert Hogg(1974)는 자료의 꼬리 모양을 보고 추정량을 고르는 적응적 방법을 내놓았다.

3.1 M-추정

정의 3.1 (M-추정량) 함수 ψ와 척도 $s>0$에 대해 $\sum_{i=1}^{n}\psi\big((X_i-\theta)/s\big)=0$의 해 $\hat\theta$를 **M-추정량**이라 한다. $\psi(u)=u$이면 평균, $\psi(u)=\operatorname{sign}(u)$이면 중앙값이고,

$$\psi_k(u)=\max\big(-k,\ \min(k,u)\big)$$

이면 **Huber 추정량**이다. 척도는 보통 중앙절대편차를 0.6745로 나눈 값(정규분포에서 σ의 일치추정량)으로 미리 정한다.

Huber 추정량은 표준화한 잔차가 $\pm k$ 안이면 평균처럼 값을 그대로 쓰고, 밖이면 $\pm k$로 잘라 쓴다. MLE는 $\psi=-(\log f)'$ 인 M-추정량이다(9회). 그래서 Huber 추정량은 가운데는 정규, 꼬리는 Laplace인 밀도의 MLE이기도 하다.

정리 3.2 (M-추정량의 점근정규성) 척도를 1로 알고, F가 0에 대해 대칭이며, ψ가 홀함수이고 단조증가하며 거의 어디서나 미분가능하고, $E\psi(X)^2<\infty$, $0<E\psi'(X)<\infty$라 하자. 그러면 $\theta=0$에서

$$\sqrt{n}\,\hat\theta \xrightarrow{D} \mathcal{N}\big(0,\ V(\psi,F)\big), \qquad V(\psi,F)=\frac{E\psi(X)^2}{\big(E\psi'(X)\big)^2}.$$

증명. **전략** 9회 정리 3.1의 증명에서 점수함수 자리에 ψ를 둔다. $\Psi_n(t)=\frac{1}{n}\sum_i\psi(X_i-t)$의 영점을 0 둘레에서 전개해 $\sqrt{n}\,\hat\theta$를 분자와 분모의 비로 쓰고 [T8], 분자는 중심극한정리로, 분모는 대수의 법칙으로 다룬 뒤 Slutsky 정리로 합친다.

① ψ가 단조증가이므로 $\Psi_n(t)$는 t에 대해 감소하고, $\hat\theta$는 그 영점이다. 정의 3.1

② $\lambda(t)=E\psi(X-t)$로 두면 $\lambda(0)=0$이고(ψ가 홀함수, F가 대칭) $\lambda'(0)=-E\psi'(X)<0$이다. 그래서 작은 $\varepsilon>0$에서 $\lambda(\varepsilon)<0<\lambda(-\varepsilon)$이다. 대칭, 적분 기호 아래 미분 [T6]

③ $\Psi_n(\pm\varepsilon)\xrightarrow{P}\lambda(\pm\varepsilon)$이므로 $P\big(\Psi_n(\varepsilon)<0<\Psi_n(-\varepsilon)\big)\to 1$이다. 그 사건 위에서 $|\hat\theta|<\varepsilon$이므로 $\hat\theta\xrightarrow{P}0$이다. Khinchin(7회 정리 1.6), ①

④ $A_n=\frac{1}{n}\sum_i\psi'(X_i)$로 두면, 0과 $\hat\theta$ 사이의 ξ_n이 있어

$$0=\Psi_n(\hat\theta)=\Psi_n(0)-\hat\theta A_n+\tfrac{1}{2}\hat\theta^2 B_n, \qquad B_n=\tfrac{1}{n}\textstyle\sum_i\psi''(X_i-\xi_n).$$ Taylor [T8]

⑤ 정리하면

$$\sqrt{n}\,\hat\theta=\frac{\frac{1}{\sqrt{n}}\sum_i\psi(X_i)}{A_n-\frac{1}{2}\hat\theta B_n}.$$ ④

⑥ 분자: $\psi(X_i)$는 평균 $0(\lambda(0)=0)$, 분산 $E\psi(X)^2$인 iid이므로 $\frac{1}{\sqrt{n}}\sum\psi(X_i)\xrightarrow{D}\mathcal{N}(0,E\psi^2)$이다. 중심극한정리, 7회 정리 3.1

⑦ 분모: $A_n\xrightarrow{P}E\psi'(X)>0$이고, $|\frac{1}{2}\hat\theta B_n|\le\frac{1}{2}|\hat\theta|\sup|\psi''|\xrightarrow{P}0$이다. Khinchin(7회 정리 1.6), ③

⑧ 극한은 $\frac{1}{E\psi'}\mathcal{N}(0,E\psi^2)=\mathcal{N}\big(0,V(\psi,F)\big)$이다. Slutsky, 7회 정리 4.2

조건 확인 ②의 미분은 ψ가 거의 어디서나 미분가능하고 $E\psi'(X)<\infty$인 데서 지배수렴으로 정당화된다 [T6]. ④와 ⑦은 ψ가 두 번 미분가능하고 ψ''가 유계라고 쓴 것이다. 9회 정리 3.1의 ⑤와 같은 자리다. Huber의 ψ_k는 $\pm k$에서 꺾이므로 이 조건을 만족하지 않는다. 그때는 ψ의 단조성으로 나머지를 다룬다(Huber 1964). 이 부분은 증명 없이 쓴다. ∎

예제 3.3 (오염 정규에서의 점근분산) $(1-\varepsilon)\mathcal{N}(0,1)+\varepsilon\mathcal{N}(0,9)$에서 nV를 정확히 적분하면

ε	평균	중앙값	10% 절사평균	Huber ($k=1.345$)
0	1.000	1.571	1.060	1.053
0.05	1.400	1.681	1.169	1.169
0.10	1.800	1.803	1.296	1.299

이다. 정규분포에서 Huber 추정량은 평균의 95% 효율을 갖고(1/1.053), 오염이 5%면 평균보다 17% 낫다. 중앙값은 정규에서 36%를 잃지만 오염이 커지면 평균을 따라잡는다. $k=1.345$는 정규에서 효율이 정확히 0.95가 되도록 고른 값이다.

증명 없이 쓰는 사실 (Huber의 최소최대 정리 (1964)) $F=(1-\varepsilon)\Phi+\varepsilon H$ (H는 아무 대칭 분포) 꼴의 모든 분포 가운데 최악의 점근분산을 가장 작게 하는 M-추정량은 ψ_k이고, k는 $\frac{2\phi(k)}{k}-2\Phi(-k)=\frac{\varepsilon}{1-\varepsilon}$으로 정해진다. $\varepsilon=0.01,\ 0.05,\ 0.10$이면 $k=1.945,\ 1.398,\ 1.140$이다. 최악의 분포는 가운데가 정규이고 꼬리가 지수적으로 줄어드는 분포다.

3.2 영향함수와 붕괴점

정의 3.4 (영향함수) 추정량을 분포의 함수 $T(F)$로 쓸 때, 점 x에 질량 t를 더한 분포 $F_t=(1-t)F+t\delta_x$에 대해

$$\mathrm{IF}(x;T,F)=\lim_{t\downarrow 0}\frac{T(F_t)-T(F)}{t}$$

를 **영향함수**라 한다. 관측 하나가 x에 더해질 때 추정량이 움직이는 양의 n배쯤이다.

명제 3.5 (M-추정량의 영향함수) 정리 3.2의 조건에서 $\mathrm{IF}(x;T,F)=\psi(x-T(F))/E\psi'(X-T(F))$이다. 따라서 평균의 영향함수는 $x-\mu$로 유계가 아니고, 중앙값은 $\mathrm{sign}(x-m)/(2f(m))$, Huber 추정량은 $\psi_k(x-\theta)/(2\Phi(k)-1)$로 유계다. 그리고 $E\,\mathrm{IF}(X)^2=V(\psi,F)$이다.

증명. **전략** $T_t=T(F_t)$가 만족하는 추정 방정식을 t로 미분하고 $t=0$을 넣는다(음함수 미분). 마지막 식은 영향함수의 정의를 정리 3.2의 V에 넣어 비교한다. 단계별 증명은 연습 3.5(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 38쪽에 있다.

영향함수는 3절의 모든 이야기를 한 그림에 담는다(그림 3.1). 평균은 이상치 하나에 끝없이 끌려가고, Huber 추정량과 중앙값은 일정한 양 이상 끌려가지 않는다. 그리고 점근분산은 영향함수의 제곱의 평균이다. 이상치에 둔감할수록, 정규 자료에서는 조금씩 정보를 버린다.

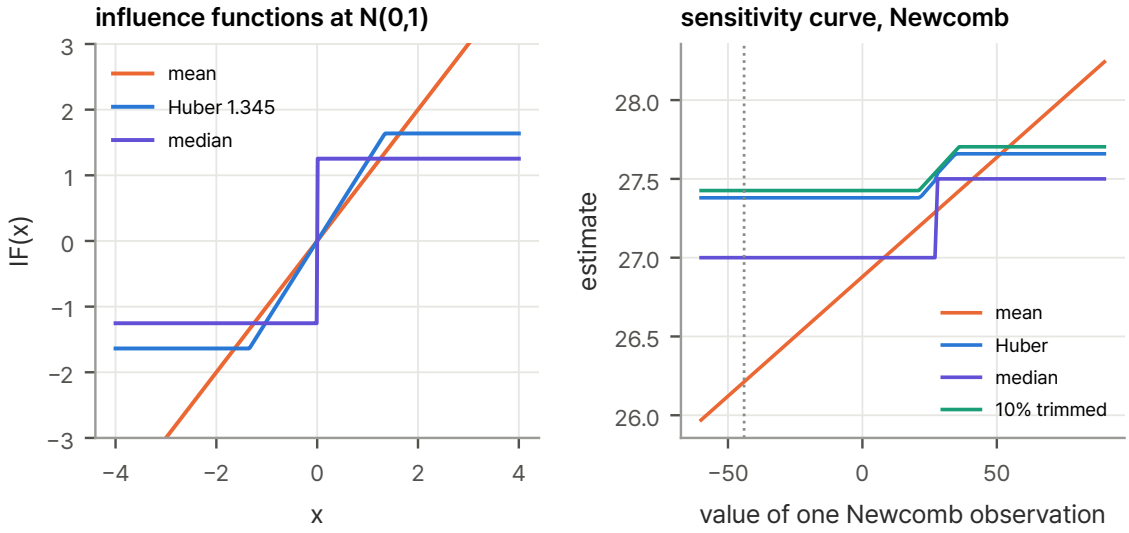


그림 3.1 왼쪽: 표준정규에서의 영향함수. 평균은 직선(유계가 아님), Huber는 $\pm k$에서 잘린 직선, 중앙값은 계단이다. 오른쪽: Newcomb의 자료에서 가장 큰 오차(-44, 점선)를 x로 바꿔 넣을 때 네 추정량이 움직이는 모양(민감도 곡선). 평균만 끝없이 따라간다.

정의 3.6 (붕괴점) 표본 n개 가운데 k개를 마음대로 바꿔 추정량을 원하는 만큼 멀리 보낼 수 있는 가장 작은 k를 k^*라 할 때 k^*/n을 **유한 표본 붕괴점**이라 한다.

명제 3.7 (세 추정량의 붕괴점) 표본평균의 붕괴점은 $1/n$, 표본중앙값은 $\lfloor n/2 \rfloor / n$ (곧 $1/2$ 근처), 양쪽에서 $\lfloor \alpha n \rfloor$개씩 자른 절사평균은 $(\lfloor \alpha n \rfloor + 1)/n$이다.

증명. **전략** 추정량마다 두 가지를 보인다. k^*개를 바꾸면 추정량을 끝없이 보낼 수 있다는 것, 그보다 적게 바꾸면 추정량이 원래 표본의 최솟값과 최댓값 사이에 묶인다는 것이다. 단계별 증명은 연습 3.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 38쪽에 있다.

예제 3.8 (Newcomb의 66번) Newcomb의 통과 시간(24,800나노초에서 뺀 값)에는 -44와 -2의 두 큰 오차가 있다. 여러 추정량은

$$\text{평균 } 26.21, \quad \text{중앙값 } 27.0, \quad \text{10\% 절사평균 } 27.43,$$
$$\text{Huber } 27.38, \quad \text{Hodges–Lehmann } 27.5$$

이다. 로버스트 추정량들은 서로 0.5 안에서 일치하고 평균만 1 넘게 떨어져 있다. 부트스트랩 표준오차(5절)는 평균이 1.32, 중앙값과 절사평균이 0.68로, 로버스트 추정량이 두 배 정밀하다. 그러나 오늘의 값으로 옮긴 참값은 33.02다(Stigler 1977). 어느 추정량도 5 넘게 빗나갔다. 4절의 주제다.

연습문제

연습 3.1 (Huber 추정량은 MLE다 ★) $f(u) \propto \exp(-\rho_k(u))$, $\rho_k(u) = u^2/2$ ($|u| \le k$), $k|u| - k^2/2$ ($|u| > k$)의 위치모수 MLE가 Huber 추정량임을 보여라. 풀이 37쪽

연습 3.2 (정규에서의 Huber 효율 ★★) $\mathcal{N}(0,1)$에서 $E\psi_k' = 2\Phi(k) - 1$, $E\psi_k^2 = 2\Phi(k) - 1 - 2k\phi(k) + 2k^2\Phi(-k)$임을 보이고, $k = 1.345$에서 효율 $1/V = 0.95$를 확인하라. 풀이 38쪽

연습 3.3 (중앙값의 영향함수 ★★) $T(F) = F^{-1}(1/2)$에 대해 F_t의 중앙값을 직접 계산해 $\mathrm{IF}(x) = \mathrm{sign}(x-m)/(2f(m))$을 보이고, $E\,\mathrm{IF}^2$이 8회 정리 1.4의 점근분산 $1/(4f(m)^2)$임을 확인하라. 풀이 38쪽

연습 3.4 (붕괴점 ★) $n = 66$인 Newcomb의 자료에서 평균, 중앙값, 10% 절사평균의 붕괴점을 구하라. 큰 오차가 몇 개까지면 10% 절사평균이 그것을 모두 버리는가? 풀이 38쪽

연습 3.5 (M-추정량의 영향함수: 증명 ★★ 유도형) 명제 3.5를 다음 순서로 증명하라. $T = T(F)$, $T_t = T(F_t)$로 쓴다.

(a) T_t가 $(1-t)\int \psi(y - T_t)\,dF(y) + t\,\psi(x - T_t) = 0$을 만족함을 M-추정량의 정의에서 보여라.

(b) (a)를 t로 미분하고 $t = 0$을 넣어 $T' = \psi(x-T)/E\psi'(X-T)$를 얻어라. $\int \psi(y-T)\,dF = 0$을 어디서 쓰는가? (미분과 적분의 교환은 [T6]의 조건으로 확인한다.)

(c) 중앙값($\psi = \mathrm{sign}$)에서는 ψ'이 없다. $E\psi'(X-t)$ 대신 $-\frac{d}{dt}E\,\mathrm{sign}(X-t) = 2f(t)$를 쓰면 $\mathrm{IF} = \mathrm{sign}(x-m)/(2f(m))$이 됨을 보여라. Huber 추정량의 분모 $2\Phi(k) - 1$도 구하라 ($F = \Phi$).

(d) $E\,\mathrm{IF}(X)^2 = V(\psi, F)$임을 보여라.

풀이 38쪽

연습 3.6 (세 추정량의 붕괴점: 증명 ★★ 유도형) 명제 3.7을 증명하라. 표본 $x_1, \dots, x_n$을 고정하고 k개를 마음대로 바꾼 표본을 y라 한다.

(a) 평균: $k = 1$로 $\bar{y}$를 원하는 만큼 멀리 보낼 수 있음을 보여라.

(b) 중앙값(n이 홀수, $n = 2r-1$): $k \le r-1$이면 y의 중앙값이 원래 표본의 최솟값과 최댓값 사이에 있음을 보여라. (바뀌지 않은 값이 r개 이상 남는다. 중앙값보다 작거나 같은 값이 r개, 크거나 같은 값이 r개다.) $k = r$이면 끝없이 보낼 수 있음을 보여라.

(c) 절사평균: 양쪽에서 $g = \lfloor \alpha n \rfloor$개씩 자를 때, $k \le g$이면 남는 값이 모두 원래 표본의 최솟값과 최댓값 사이에 있음을, $k = g + 1$이면 그렇지 않음을 보여라.

풀이 38쪽

4

참값이 있는 자료로 — Stigler (1977)

> **역사** 로버스트 추정량들은 대부분 시뮬레이션으로 비교되었다. 오염 정규, t 분포, Cauchy 같은 "두꺼운 꼬리"의 표본을 만들어 분산을 잰 것이다. 1977년 Stephen Stigler는 **로버스트 추정량은 실제 자료를 다루는가?**에서 다른 길을 택했다. 18–19세기의 과학자들이 오늘 참값을 아는 양을 잰 기록 24개를 모았다. Cavendish의 지구 밀도, Michelson과 Newcomb의 빛의 속도, Short(1763)의 태양 시차 같은 것이다. 그리고 11개의 추정량을 각 자료에 써서 참값과의 거리를 쟀다. 결론은 겸손했다. 약하게 자른 평균(10% 절사평균)이 전체적으로 가장 나았고, 표본평균도 크게 뒤지지 않았으며, 극단적으로 로버스트한 추정량은 오히려 나빴다. 그리고 어느 추정량이든 오차의 대부분은 측정의 체계적 편향이었다.

예제 4.1 (일곱 개의 자료) Cavendish(29개, 참값 5.513), Newcomb(66개, 33.02), Michelson(100개 전체와 20개씩의 다섯 묶음, 734.5)에 여섯 추정량을 쓰면 다음과 같다. 마지막 열은 평균의 오차를 평균의 표준오차로 나눈 것이다.

자료	n	평균	중앙값	10% 절사	25% 절사	Huber	HL	(평균−참)/SE
Cavendish	29	5.482	5.470	5.484	5.479	5.483	5.480	−0.8
Newcomb	66	26.21	27.00	27.43	27.27	27.38	27.50	−5.1
Michelson	100	852.4	850.0	852.3	849.0	851.5	850.0	14.9
묶음 1	20	852.5	850.0	853.1	853.0	852.9	852.5	7.3
묶음 2	20	854.5	850.0	859.4	857.0	859.4	860.0	5.8
묶음 3	20	858.5	850.0	854.4	848.0	853.1	852.5	8.0
묶음 4	20	857.0	855.0	856.3	851.0	855.0	855.0	5.7
묶음 5	20	839.5	840.0	837.5	841.0	836.8	840.0	7.1

표를 세 갈래로 읽을 수 있다. 첫째, Cavendish에서는 모든 추정량이 0.02 안에서 같고 참값과의 차이도 표준오차 안이다. 큰 오차가 없으면 무엇을 써도 같다. 둘째, Newcomb에서는 두 큰 오차 때문에 로버스트 추정량들이 평균보다 참값에 1 이상 가깝다. 3절의 이론이 말한 그대로다. 셋째, Michelson에서는 추정량들의 차이가 몇 km/s인데 참값과의 거리는 100 km/s를 넘는다. 표준오차의 6–15배다. 이 오차는 분포의 꼬리가 아니라 장치의 보정에서 왔다. 어떤 추정량도 체계적 편향을 고치지 못한다(3회 7절).

가정 주의 로버스트 추정량은 "분포가 정규에서 조금 벗어난" 경우를 지킨다. 모든 관측이 같은 방향으로 틀린 편향은 자료 안에서 보이지 않는다. 그것을 잡으려면 다른 방법으로 잰 측정과 비교해야 한다. Michelson 자신이 그 일을 했다. 그의 1879년 값은 1882년, 1924년, 1926년의 측정으로 차례로 고쳐졌다.

연습문제

연습 4.1 (오차의 분해 ★) Michelson 100개의 평균 852.4, 표준편차 79.0으로 평균제곱오차 $(\bar{X} - 734.5)^2$를 편향의 제곱과 분산 s^2/n으로 나누어 각각의 비율을 구하라. 로버스트 추정량이 줄일 수 있는 것은 어느 쪽인가? 풀이 39쪽

연습 4.2 (Newcomb의 두 오차를 빼면 ★) −44와 −2를 빼면 평균과 표준편차가 어떻게 되는가? 그 평균을 표 속의 로버스트 추정량들과 비교하라(노트북). 풀이 39쪽

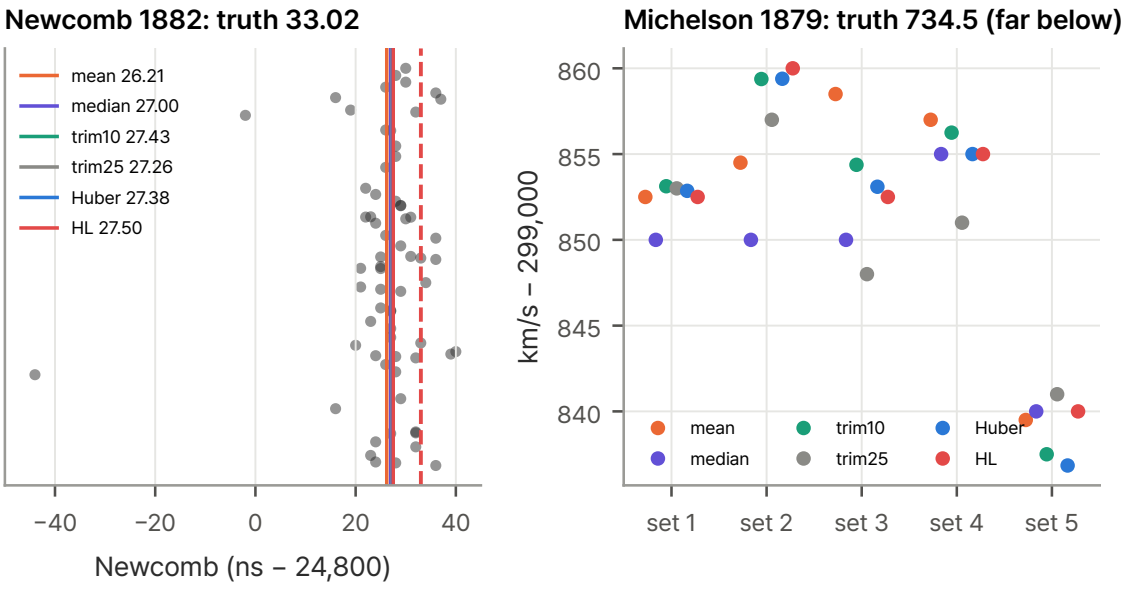


그림 4.1 참값이 알려진 측정. 왼쪽: Newcomb의 66번(점)과 여섯 추정량(세로선). 두 큰 오차(−44, −2)가 평균만 끌어내린다. 빨간 점선은 참값 33.02. 오른쪽: Michelson의 다섯 묶음에서 여섯 추정량의 값(참값 734.5는 그림 아래 멀리 있다). 추정량 사이의 차이는 편향에 비하면 작다.

5

부트스트랩 (1949–1979)

> **역사** 로버스트 추정량에는 곤란한 점이 하나 있었다. 표준오차의 공식이 분포에 의존한다는 것이다. 정리 3.2의 $V(\psi, F)$에는 모르는 F가 들어 있다. Maurice Quenouille(1949)는 관측을 하나씩 빼고 다시 계산해 치우침을 줄이는 방법을 냈고, Tukey(1958)는 그것으로 분산까지 재며 "잭나이프"(무엇에든 쓰는 주머니칼)라 불렀다. 1979년 Bradley Efron은 *Annals of Statistics*에서 더 단순하고 더 일반적인 생각을 내놓았다. 모르는 F를 경험분포 $\hat{F}_n$으로 바꾸고, $\hat{F}_n$에서 표본을 다시 뽑아 추정량의 분포를 컴퓨터로 만든다. 이름은 "자기 장화 끈을 잡아당겨 스스로를 들어 올린다"는 허풍에서 왔다.

5.1 plug-in과 부트스트랩

정의 5.1 (부트스트랩) 자료 $X_1, \dots, X_n$이 주어졌을 때 $\hat{F}_n$에서 뽑은 크기 n의 표본, 곧 원래 자료에서 복원추출한 $X_1^*, \dots, X_n^*$를 **부트스트랩 표본**이라 한다. 통계량 $\hat{\theta} = T(X_1, \dots, X_n)$의 **부트스트랩 분포**는 자료를 고정했을 때 $\hat{\theta}^* = T(X_1^*, \dots, X_n^*)$의 분포다. 원리는 한 줄이다.

$$F \text{ 에서의 } \hat{\theta} - \theta \text{ 의 분포} \;\approx\; \hat{F}_n \text{ 에서의 } \hat{\theta}^* - \hat{\theta} \text{ 의 분포.}$$

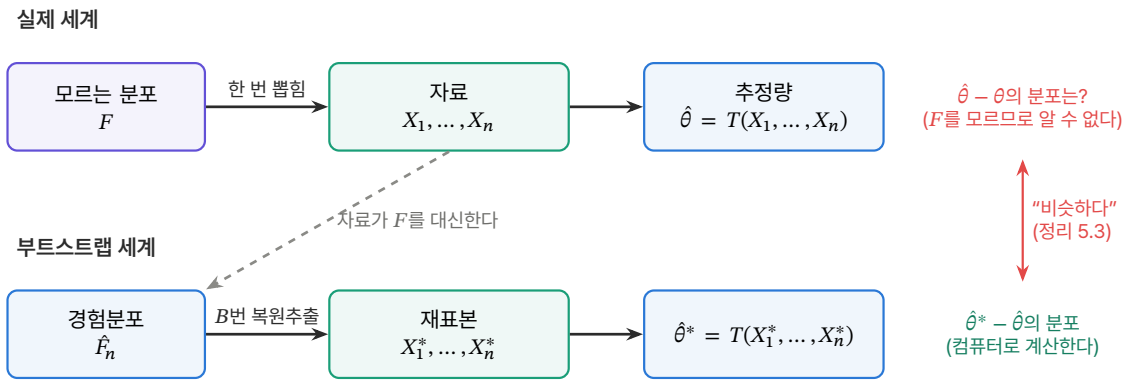


그림 5.1 부트스트랩의 두 세계. 실제 세계에서는 모르는 F에서 자료가 한 번 뽑히고, 추정량의 오차 분포를 알 수 없다. 부트스트랩 세계에서는 자료의 경험분포 $\hat{F}_n$이 F의 자리에 서고, 거기서는 표본을 몇 번이든 다시 뽑을 수 있다. 두 세계의 오차 분포가 비슷하다는 것이 정리 5.3의 내용이다.

알고리즘 5.2 (몬테카를로 부트스트랩)

1. $b = 1, \dots, B$에 대해 자료에서 n개를 복원추출해 $\hat\theta^*_b$를 계산한다.
2. 표준오차: $\widehat{\mathrm{SE}} = \sqrt{\frac{1}{B-1}\sum(\hat\theta^*_b - \bar\theta^*)^2}$.
3. 백분위 구간: $\hat\theta^*_b$들의 $\alpha/2$, $1-\alpha/2$ 분위수.

$\hat F_n$에서의 분포는 원리적으로 정확히 계산할 수 있는 양이다(n^n가지 표본). 몬테카를로는 그것을 근사할 뿐이다(7회). B가 수천이면 그 오차는 무시할 만하다.

정리 5.3 (평균의 부트스트랩) $X_1, X_2, \dots \overset{\text{iid}}{\sim} F$이고 $E|X|^3 < \infty$, $\sigma^2 = \mathrm{Var}\,X > 0$이라 하자. 자료를 고정한 조건부 확률 P^*에서

(a) $E^*\bar X^* = \bar X$, $\mathrm{Var}^*\bar X^* = \hat\sigma^2/n$, 여기서 $\hat\sigma^2 = \frac{1}{n}\sum(X_i - \bar X)^2$이다.

(b) 거의 확실하게 $\sup_x \left|P^*\left(\sqrt{n}(\bar X^* - \bar X) \le x\right) - P\left(\sqrt{n}(\bar X - \mu) \le x\right)\right| \to 0$이다.

증명. **전략** 자료를 고정하면 부트스트랩 표본은 $\hat F_n$에서 뽑은 iid 표본이다. (a)는 그 평균과 분산이다. (b)는 두 분포함수를 모두 정규분포함수와 비교한다. 부트스트랩 쪽은 $\hat F_n$에 쓴 Berry–Esseen 한계로, 실제 쪽은 F에 쓴 같은 한계로 누르고, 가운데를 삼각부등식으로 잇는다 [T4].

(a)

① P^*에서 X_i^*는 $\hat F_n$에서 독립으로 뽑히므로 $E^*X_i^* = \frac{1}{n}\sum X_j = \bar X$, $\mathrm{Var}^* X_i^* = \frac{1}{n}\sum(X_j - \bar X)^2 = \hat\sigma^2$이다. 정의 5.1

② $E^*\bar X^* = \bar X$, $\mathrm{Var}^*\bar X^* = \hat\sigma^2/n$이다. 선형성, 독립합의 분산(2회 정리 2.8)

(b)

① $\hat\rho = \frac{1}{n}\sum|X_i - \bar X|^3$으로 두면, 조건부로 iid인 X_i^*에 Berry–Esseen 한계를 써서

$$\sup_x \left|P^*\left(\sqrt{n}(\bar X^* - \bar X) \le x\right) - \Phi(x/\hat\sigma)\right| \le \frac{C\hat\rho}{\hat\sigma^3\sqrt{n}}.$$

7회 3절의 상자, $\hat F_n$에

② $\hat\sigma \to \sigma > 0$이고 $\hat\rho \to E|X-\mu|^3 < \infty$이므로 ①의 오른쪽은 0으로 간다. 대수의 법칙, 연속사상 [T5]

③ 같은 한계를 F에 쓰면 $\sup_x|P\left(\sqrt{n}(\bar X - \mu) \le x\right) - \Phi(x/\sigma)| \le C\rho/(\sigma^3\sqrt{n}) \to 0$이다. 7회 3절의 상자

④ $\sup_x|\Phi(x/\hat\sigma) - \Phi(x/\sigma)| \to 0$이다. 평균값 정리로 이 차는 $|x|\,|\frac{1}{\hat\sigma} - \frac{1}{\sigma}|\,\phi(\xi)$이고, $|x|\phi(\xi)$는 $\hat\sigma$가 σ 근처에 있는 동안 x에 대해 유계다. 평균값 정리 [T8]

⑤ ①–④를 삼각부등식으로 이으면 결론이 나온다. [T4]

조건 확인 ②에서 "거의 확실하게"라고 쓴 수렴은 강한 대수의 법칙이 필요하다. 이 과정은 그

것을 진술만 했다(7회 1절). 약한 대수의 법칙(7회 정리 1.6)만 쓰면 같은 논증으로 ⑤의 상한이 확률수렴으로 0으로 간다는, 조금 약한 결론이 나온다. $\hat{\rho}$의 수렴은 $|X_i - \bar{X}|^3$을 $|X_i - \mu|^3$과 $\bar{X} - \mu$의 다항식으로 눌러 얻는다($E|X|^3 < \infty$를 쓴다). ■

부트스트랩 표준오차 $\hat{\sigma}/\sqrt{n}$은 보통의 공식 $s/\sqrt{n}$과 $\sqrt{(n-1)/n}$배 다를 뿐이다. 평균에서는 새것이 없다. 부트스트랩의 값은 공식이 없는 추정량, 곧 중앙값, 절사평균, Huber 추정량, 상관계수, 회귀계수의 비 같은 것에서 같은 계산이 그대로 통한다는 데 있다.

5.2 더 나은 구간: BCa

백분위 구간은 $\hat{\theta}^*$의 분포가 $\hat{\theta}$를 중심으로 대칭이고 치우침이 없을 때 잘 맞는다. Efron(1987)의 BCa(치우침 수정·가속) 구간은 두 가지를 고친다.

정의 5.4 (BCa 구간) $\hat{z}_0 = \Phi^{-1}\big(\#\{\hat{\theta}^*_b < \hat{\theta}\}/B\big)$ (치우침), 잭나이프 값 $\hat{\theta}_{(i)}$ (관측 i를 뺀 추정)과 그 평균 $\hat{\theta}_{(\cdot)}$으로

$$\hat{a} = \frac{\sum(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\big[\sum(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\big]^{3/2}} \quad \text{(가속)}, \qquad \alpha_j = \Phi\Big(\hat{z}_0 + \frac{\hat{z}_0 + z_{(j)}}{1 - \hat{a}(\hat{z}_0 + z_{(j)})}\Big)$$

로 두고, 부트스트랩 분포의 α_1, α_2 분위수를 구간으로 쓴다($z_{(1)} = z_{\alpha/2}$의 아래쪽, $z_{(2)}$는 위쪽 분위수).

증명 없이 쓰는 사실 (BCa의 정확도 (Efron 1987, Hall 1988)) 매끄러운 통계량에서 BCa 구간의 포함확률 오차는 $O(1/n)$이고, 백분위 구간과 정규근사 구간은 $O(1/\sqrt{n})$이다. 변환불변이어서, θ의 구간을 $g(\theta)$로 옮기면 $g(\theta)$의 BCa 구간이 된다.

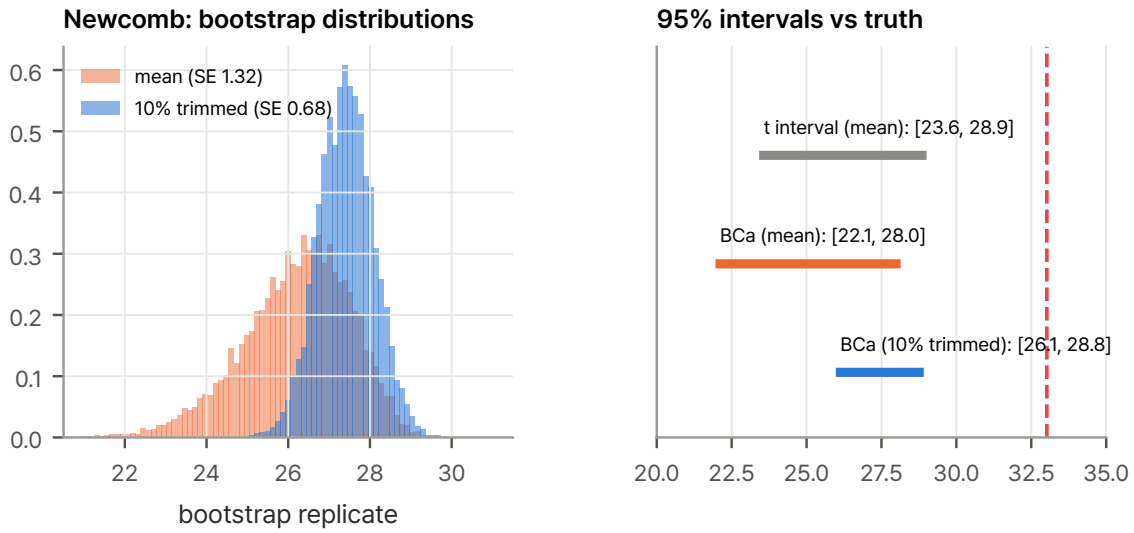


그림 5.2 Newcomb의 부트스트랩 분포. 왼쪽: 평균(주황)과 10% 절사평균(파랑)의 부트스트랩 분포. 평균의 분포는 넓고 왼쪽 꼬리가 길다. 오른쪽: 세 가지 95% 구간(t, 평균의 BCa, 절사평균의 BCa)과 참값 33.02(빨간 점선).

예제 5.5 (Newcomb와 Cavendish의 부트스트랩) $B = 20{,}000$으로 Newcomb의 66번에서 표준오차를 구하면 평균 1.32 (공식 $s/\sqrt{n} = 1.32$), 중앙값 0.68, 10% 절사평균 0.68이다. 평균의 95% 구간은 t로 $[23.6,\ 28.9]$, BCa로 $[22.1,\ 28.0]$이다. BCa는 두 큰 음의 오차가 만든 왼쪽 꼬리를 반영해 아래로 늘어난다. 절사평균의 BCa 구간은 $[26.1,\ 28.8]$로 절반쯤 좁다. 어느 구간도 참값 33.02를 포함하지 않는다. Cavendish의 평균에서는 부트스트랩 표준오차 0.038이 공식 0.038과 같고, BCa 구간 $[5.41,\ 5.56]$이 t 구간과 거의 같다(그림 5.2).

5.3 잭나이프

정의 5.6 (잭나이프 표준오차) $\hat\theta_{(i)}$를 관측 i를 뺀 추정, $\hat\theta_{(\cdot)} = \frac{1}{n}\sum \hat\theta_{(i)}$라 할 때 $\widehat{\mathrm{SE}}_{\mathrm{jack}} = \sqrt{\frac{n-1}{n}\sum(\hat\theta_{(i)} - \hat\theta_{(\cdot)})^2}$이다.

잭나이프는 n번의 계산으로 끝나고 매끄러운 통계량에서는 부트스트랩과 비슷한 답을 준다. 그러나 중앙값처럼 매끄럽지 않은 통계량에서는 일치하지 않는다(연습 5.2). 부트스트랩이 더 일반적인 이유다.

연습문제

연습 5.1 (평균의 잭나이프 ★) $\hat\theta = \bar X$이면 $\widehat{\mathrm{SE}}_{\mathrm{jack}} = s/\sqrt{n}$으로 보통의 공식과 정확히 같음을 보여라. 풀이 39쪽

연습 5.2 (중앙값의 잭나이프 ★★) $n = 2m$이 짝수이면 관측 하나를 뺀 중앙값은 $X_{(m)}$ 또는 $X_{(m+1)}$의 두 값뿐임을 보이고, 잭나이프 표준오차가 $X_{(m+1)} - X_{(m)}$ 하나에만 의존해 일치추정량이 될 수 없음을 설명하라. 풀이 39쪽

연습 5.3 (부트스트랩 표본의 다양성 ★★) 크기 n의 부트스트랩 표본에 원래 관측 i가 한 번도 들어가지 않을 확률이 $(1-1/n)^n \to e^{-1} = 0.368$임을 보이고, 부트스트랩 표본에 서로 다른 관측이 평균 $n(1-(1-1/n)^n) \approx 0.632n$개 있음을 보여라. (8절의 out-of-bag이 이것이다.) 풀이 39쪽

연습 5.4 (정확한 부트스트랩 ★★) $n = 3$, 자료 $(1, 2, 6)$일 때 서로 다른 부트스트랩 표본(다중집합)이 $\binom{5}{3} = 10$개임을 보이고, 각각의 확률(다항계수/27)로 중앙값의 부트스트랩 분포를 정확히 계산하라. 풀이 39쪽

6

원자료로 — 회귀도 로버스트하게

13회의 최소제곱은 회귀의 평균이다. 잔차 하나가 크면 계수 전체가 그쪽으로 끌려간다. 위치모수의 M-추정을 회귀로 옮기면 로버스트 회귀가 된다.

정의 6.1 (회귀의 M-추정) $\sum_{i=1}^{n}\psi\big((y_i-\mathbf{x}_i'\mathbf{b})/s\big)\,\mathbf{x}_i=\mathbf{0}$의 해 $\hat{\boldsymbol{\beta}}$을 회귀의 M-추정량이라 한다. $\psi(u)=u$이면 최소제곱의 정규방정식(13회 정리 2.2)이고, $\psi=\psi_k$이면 Huber 회귀다.

알고리즘 6.2 (반복 가중 최소제곱 (IRLS))

1. 최소제곱으로 $\mathbf{b}^{(0)}$을 구한다.
2. 잔차 $r_i=y_i-\mathbf{x}_i'\mathbf{b}^{(t)}$와 척도 $s=\mathrm{MAD}(r)/0.6745$를 계산하고, 무게 $w_i=\psi(r_i/s)/(r_i/s)$ (Huber이면 $\min(1,\,k/|r_i/s|)$)를 정한다.
3. 가중 최소제곱 $\mathbf{b}^{(t+1)}=(X'WX)^{-1}X'W\mathbf{y}$ ($W=\mathrm{diag}(w_i)$)를 풀고, 바뀌지 않을 때까지 되풀이한다.

명제 6.3 (IRLS의 고정점) IRLS가 수렴하면 그 극한(척도 s를 고정했을 때)은 회귀의 M-추정 방정식을 푼다.

증명. 고정점에서 $X'W(\mathbf{y}-X\mathbf{b})=\mathbf{0}$, 곧 $\sum_i w_i(y_i-\mathbf{x}_i'\mathbf{b})\mathbf{x}_i=\mathbf{0}$이다. $w_ir_i=\psi(r_i/s)\,s$이므로 이것은 $s\sum\psi(r_i/s)\mathbf{x}_i=\mathbf{0}$와 같다. ∎

IRLS는 13회 따름정리 3.7의 가중 최소제곱을 되풀이하는 것이다. 큰 잔차를 낸 관측의 무게를 줄이고 다시 맞춘다. 무게 w_i 자체가 진단이다. 어느 관측을 얼마나 믿지 않았는지 알려 준다.

예제 6.4 (질산 공장의 stack loss) Brownlee(1965)의 자료는 암모니아를 산화해 질산을 만드는 공장의 21일 동안의 운전 기록이다. 반응되지 못하고 굴뚝으로 빠져나간 암모니아의 양(stack loss)을 공기 흐름, 냉각수 온도, 산 농도로 회귀한다. 세 방법의 계수(절편, 공기, 냉각수, 산)는

	절편	공기 흐름	냉각수 온도	산 농도
최소제곱	−39.92	0.716	1.295	−0.152
Huber ($k = 1.345$)	−41.05	0.827	0.939	−0.129
최소절댓값	−39.69	0.832	0.574	−0.061

이다. Huber 회귀는 3, 4, 21일째의 무게를 0.82, 0.53, 0.38로 줄였다. 이 자료를 분석한 많은 연구가 이상치로 지목한 날들이다. 냉각수 온도의 계수가 1.30에서 0.94로 줄어든 것은 그 몇 날이 최소제곱의 기울기를 끌어올렸다는 뜻이다(그림 6.1).

가정 주의 Huber 회귀는 y 방향의 이상치에 강하지만, $\mathbf{x}$ 방향으로 동떨어진 지렛대 점(13회 명제 2.7)에는 여전히 약하다. 지렛대가 1에 가까운 점은 잔차가 작게 나오므로 무게가 줄지 않는다. 그런 점까지 막으려면 최소중앙제곱(Rousseeuw 1984)처럼 붕괴점이 높은 추정이 필요하다.

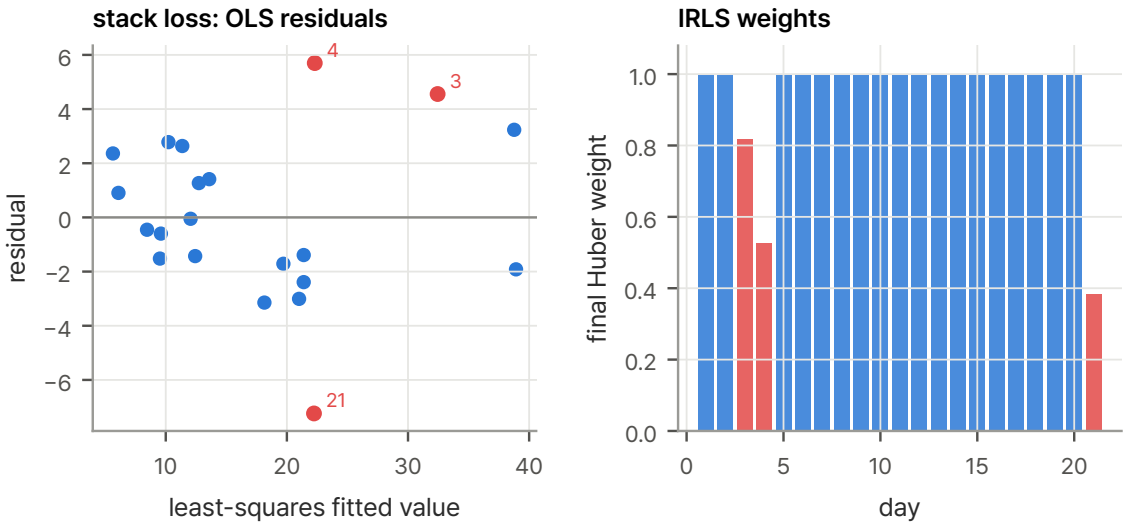


그림 6.1 Stack loss 자료. 왼쪽: 최소제곱 적합값과 잔차. 3, 4, 21일째(빨강)가 크게 벗어난다. 오른쪽: Huber IRLS의 마지막 무게. 큰 잔차를 낸 날일수록 무게가 작다.

연습문제

연습 6.1 (최소절댓값도 IRLS ★★) $\psi = \text{sign}$이면 IRLS의 무게가 $w_i = 1/|r_i|$임을 보이고, 고정점이 13회 1절의 최소절댓값 문제의 정상점임을 설명하라. 잔차가 0에 가까우면 어떤 문제가 생기는가? 풀이 40쪽

연습 6.2 (무게의 해석 ★) Huber의 무게 $w_i = \min(1, k/|u_i|)$에서 무게가 0.38인 관측의 표준화 잔차 $|u_i|$는 얼마인가($k = 1.345$)? 풀이 40쪽

7

부쉬보기 — 분포를 덜 믿어도 깨지는 곳

7.1 부트스트랩은 최댓값에서 실패한다

8회의 독일 전차 문제는 일련번호의 최댓값 $X_{(n)}$으로 전차의 수 θ를 추정했다. 부트스트랩으로 그 분포를 흉내 낼 수 있을까?

정리 7.1 (최댓값의 부트스트랩) $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathrm{U}(0,\theta)$이면 $n(\theta - X_{(n)})/\theta \xrightarrow{D} \mathrm{Exp}(1)$로 연속분포에 수렴한다. 그러나 부트스트랩 분포는

$$P^*\!\left(X^*_{(n)} = X_{(n)}\right) = 1 - \left(1 - \frac{1}{n}\right)^n \;\to\; 1 - e^{-1} = 0.632$$

의 원자를 0에 가지므로, $n(X_{(n)} - X^*_{(n)})/X_{(n)}$의 부트스트랩 분포는 $\mathrm{Exp}(1)$로 수렴하지 않는다.

증명. **전략** 실제 분포는 최댓값의 분포함수를 직접 계산해 극한을 구한다. 부트스트랩 분포는 0에 있는 원자의 확률을 계산하고, 연속분포로 수렴하는 분포함수가 한 점에서 큰 점프를 가질 수 없음을 쓴다. 단계별 증명은 연습 7.4(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 40쪽에 있다.

부트스트랩이 통하려면 통계량이 F에 대해 "매끄럽게" 의존해야 한다(평균처럼). 최댓값은 분포의 끝 하나에만 의존하고, 경험분포는 그 끝을 언제나 자료의 최댓값에 둔다. $n = 50$의 균등 표본으로 부트스트랩을 2만 번 하면 63.6%가 원래 최댓값을 그대로 준다. 대안은 n보다 작은 m개만 뽑는 "m-out-of-n" 부트스트랩이나, 문제의 구조를 쓰는 모수적 방법(8회의 피벗 $X_{(n)}/\theta$)이다.

7.2 순위합은 중앙값의 검정이 아니다

Wilcoxon–Mann–Whitney 검정은 흔히 "두 중앙값이 같은가"의 검정이라 소개된다. 정리 1.5이 쓴 귀무가설은 $F = G$, 곧 분포 전체가 같다는 것이다.

예제 7.2 (분산만 다른 두 집단) 두 집단이 모두 0을 중심으로 대칭이지만(중앙값이 같다) 크기 10의 집단은 표준편차 4, 크기 40의 집단은 표준편차 1이라 하자. 5% 검정을 2만 번 하면 실제 크기는 Mann–Whitney 0.147, 합동분산 t 0.293, Welch t 0.051이다. 표준편차를 바꾸면(작은 집단이 조용하면) 0.007, 0.00025, 0.049가 된다. 순위검정이 "분포를 가정하지 않는다"는 것은 귀무가설 $F = G$ 아래의 분포가 F와 무관하다는 뜻이지, 어떤 차이에도 둔감하다는 뜻이 아니다. $P(X > Y) = 1/2$이어도 퍼짐이 다르면 U의 분산이 정리의 값과 달라진다.

7.3 로버스트 추정량은 무엇을 추정하는가

예제 7.3 (한쪽으로만 섞이면) $0.9\,\mathcal{N}(0,1) + 0.1\,\mathcal{N}(5,1)$에서 평균의 극한은 0.5, 중앙값은 0.140, Huber 추정량($k = 1.345$, 척도 1)은 0.182다. 오염이 대칭이 아니면 세 추정량은 서로 다른 값으로 간다. "참값"이 0인지(오염을 버려야 할 측정 실수로 볼 때), 0.5인지(모든 관측이 같은 모집단일 때)는 통계학이 아니라 문제가 정한다. 로버스트 추정은 "대부분의 자료가 말하는 값"을 추정하고, 평균은 "모든 자료의 무게중심"을 추정한다.

연습문제

연습 7.1 (원자의 크기 ★) $n = 10,\ 50,\ 1000$에서 $P^*(X^*_{(n)} = X_{(n)})$을 계산하라. 이 확률이 연속분포에서는 0이어야 하는 이유는? 풀이 40쪽

연습 7.2 (Cauchy의 부트스트랩 ★★) Cauchy 표본에서 평균의 부트스트랩 표준오차가 무엇을 재는지 생각해 보라. 자료가 늘어날수록 그 값은 어떻게 되는가? (7회 3절: Cauchy의 평균은 Cauchy다. 정리 5.3은 어느 가정을 쓰는가?) 풀이 40쪽

연습 7.3 (두 중앙값을 비교하려면 ★★) 두 집단의 중앙값의 차를 비교하는 구간을 만드는 방법 두 가지(부트스트랩, 8회 정리 2.6의 분위수 구간을 각 집단에)를 설명하고, 예제 7.2의 상황에서 어느 쪽이 크기를 지키겠는지 말하라. 풀이 40쪽

연습 7.4 (최댓값의 부트스트랩: 증명 ★★ 유도형) 정리 7.1를 다음 순서로 증명하라.

(a) $0 \le x \le n$이면 $P\big(n(\theta - X_{(n)})/\theta > x\big) = (1 - x/n)^n$ 임을 보이고, $n \to \infty$에서 e^{-x}로 감을 보여라 [T5].

(b) 자료를 고정하면 $X^*_{(n)} = X_{(n)}$은 "n번의 복원추출 가운데 적어도 한 번 최댓값이 뽑힌다"와 같다. 그 확률이 $1 - (1 - 1/n)^n$ 임을 보여라.

(c) $V_n^* = n(X_{(n)} - X^*_{(n)})/X_{(n)}$의 부트스트랩 분포함수를 G_n이라 하면 $G_n(0) \ge 1 - (1 - 1/n)^n$이다. 만약 G_n이 Exp(1)의 분포함수 G로 모든 점에서 수렴한다면 $G(0) = 0$과 모순임을 보여라. (G는 0에서 연속이다.) 풀이 40쪽

8

오늘 — bagging과 Huber 손실

8.1 Bagging과 random forest

Leo Breiman(1996)은 부트스트랩을 추정량의 오차를 재는 데가 아니라 예측기를 **만드는** 데 썼다. 부트스트랩 표본마다 예측기를 학습하고 그 예측을 평균한다. **bagging**(bootstrap aggregating)이다.

명제 8.1 (평균한 예측기의 분산) 예측값 $\hat f_1, \dots, \hat f_B$가 같은 분산 σ^2과 쌍마다 같은 상관 ρ를 가지면 $\operatorname{Var}\left(\frac{1}{B}\sum \hat f_b\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$이다.

증명. $\frac{1}{B^2}\left(B\sigma^2 + B(B-1)\rho\sigma^2\right)$를 정리한다(3회 8절). ∎

부트스트랩 표본들은 같은 자료에서 왔으므로 ρ가 0이 아니다. 그래서 B를 늘려도 분산은 $\rho\sigma^2$ 밑으로 내려가지 않는다. Breiman(2001)의 random forest는 나무마다 쓸 수 있는 변수를 무작위로 제한해 ρ를 줄인다. 이득이 큰 것은 결정 나무처럼 자료가 조금 바뀌면 예측이 크게 바뀌는 "불안정한" 예측기다.

예제 8.2 (결정 그루터기) $x \sim \mathrm{U}(-1,1)$, $y = \mathbf{1}\{x > 0\} + \mathcal{N}(0,1)$인 자료 100개에 분기 하나짜리 나무(그루터기)를 맞추고 $x_0 = 0.05$에서 예측한다. 자료를 200번 새로 만들면 그루터기 예측의 분산은 0.131, 부트스트랩 50번으로 bagging한 예측의 분산은 0.068로 절반쯤이 된다. 대신 평균은 0.86에서 0.77로 참값 1에서 조금 더 멀어진다. 분기점 근처에서 bagging은 계단을 매끄러운 경사로 바꾼다. 분산이 크게 줄고 치우침이 조금 는다.

부트스트랩 표본마다 뽑히지 않은 약 36.8%의 관측(연습 5.3)은 그 나무의 검증 자료로 쓸 수 있다. 이것이 **out-of-bag** 오차이고, 교차검증 없이 일반화 오차를 추정한다.

8.2 Huber 손실과 smooth L1

신경망의 회귀 손실로 널리 쓰이는 **Huber 손실**과 객체 검출의 **smooth L1** 손실(Girshick 2015)은 Huber(1964)의 ρ_k 그대로다.

명제 8.3 (Huber 손실의 세 얼굴) $\rho_k(u) = u^2/2\ (|u| \le k),\ k|u| - k^2/2\ (|u| > k)$에 대해

(a) $\rho_k' = \psi_k$이다. 경사하강의 한 걸음은 잔차를 $\pm k$로 자른 것을 쓴다.

(b) $\rho_k(u) = \min_z\{\frac{1}{2}(u-z)^2 + k|z|\}$이다. 큰 잔차를 "이상치 몫" z와 "정상 몫" $u-z$로 나누어, 이상치 몫에는 절댓값 벌점을 준 것이다.

(c) smooth L1 손실 $\frac{1}{2\beta}u^2\ (|u| < \beta)$, $|u| - \beta/2$ (그 밖)은 $\rho_\beta(u)/\beta$다.

증명. **전략** (a)는 구간마다 미분하고 $|u| = k$에서 양쪽 도함수가 같음을 확인한다. (b)는 z에 대한 최소화를 z의 부호로 나눠 푼다. (c)는 두 식을 구간마다 비교한다. 단계별 증명은 연습 8.3(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 41쪽에 있다.

(b)의 z는 16회에서 볼 lasso와 같은 연성 임계값이다. 로버스트 회귀는 "관측마다 이상치 몫이 있을 수 있고, 그 몫은 드물다(희소하다)"고 가정한 lasso와 같다. 경사 자르기(gradient clipping)도 같은 생각이다. 한 미니배치의 큰 기울기가 학습을 끌고 가지 못하게 자른다.

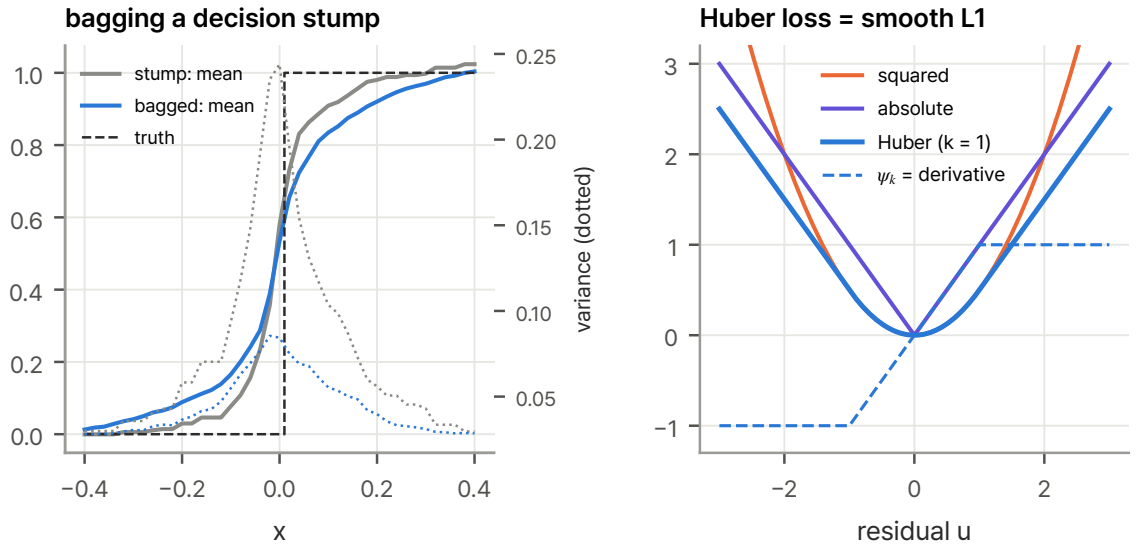


그림 8.1 왼쪽: 그루터기 예측(회색)과 bagging 예측(파랑)의 x에 대한 평균 곡선과 분산. bagging은 계단을 경사로 바꾸고 분기점 근처의 분산을 줄인다. 오른쪽: 제곱 손실, 절댓값 손실, Huber 손실($k = 1$)과 그 도함수. Huber의 도함수는 $\pm k$에서 잘린다.

연습문제

연습 8.1 (상관이 남기는 분산 ★) $\rho = 0.4$, $\sigma^2 = 1$이면 $B = 1,\ 10,\ 100,\ \infty$에서 명제 8.1의 분산은 얼마인가?

풀이 41쪽

연습 8.2 (연성 임계값 ★★) 명제 8.3(b)의 최소화를 $z > 0$, $z < 0$, $z = 0$의 경우로 나눠 풀어라.

풀이 41쪽

연습 8.3 (Huber 손실의 세 얼굴: 증명 ★ 유도형) 명제 8.3를 증명하라.

(a) $|u| < k$와 $|u| > k$에서 ρ_k'을 구하고, $u = \pm k$에서 왼쪽과 오른쪽 도함수가 같아 ρ_k가 미분가능함을 보여라. $\rho_k' = \psi_k$를 확인하라.

(b) 연습 8.2의 결과 $z^* = \operatorname{sign}(u)\max(|u| - k, 0)$를 $\frac{1}{2}(u - z)^2 + k|z|$에 넣어 $\rho_k(u)$를 얻어라.

(c) $\beta\,\rho_\beta(u)$가 아니라 $\rho_\beta(u)/\beta$가 smooth L1 손실임을 $|u| < \beta$와 $|u| \geq \beta$에서 각각 확인하라.

풀이 41쪽

9

정리와 역사 노트

9.1 한 장 요약

정리	내용	조건
순열검정 (1.2)	나눔 위의 분포로 만든 p-값은 정확하다	$F = G$ (교환가능)
순위합 (1.5)	$W = U + m(m+1)/2$, $EW = m(N+1)/2$, $\operatorname{Var} W = mn(N+1)/12$	$F = G$ 연속
부호순위 (1.8)	$T^+ = \sum k I_k$, $ET^+ = n(n+1)/4$	0에 대칭
Wilcoxon 대 t (2.3)	ARE $= 12\sigma^2(\int f^2)^2$, 정규에서 $3/\pi$	위치 이동
Hodges–Lehmann (2.5)	ARE $\geq 108/125 = 0.864$	분산 유한
M-추정 (3.2)	$\sqrt{n}\hat{\theta} \to \mathcal{N}(0, E\psi^2/(E\psi')^2)$	대칭, 단조 ψ
영향함수 (3.5)	$\psi(x-\theta)/E\psi'$, $E\,\mathrm{IF}^2 = V$	—
붕괴점 (3.7)	평균 $1/n$, 중앙값 $1/2$, 절사평균 α	—
평균의 부트스트랩 (5.3)	조건부 분포가 표본분포에 균등하게 가까워진다	$E\lvert X\rvert^3 < \infty$
IRLS (6.3)	고정점 = 회귀의 M-추정 방정식	—
최댓값의 부트스트랩 (7.1)	원자 $1-(1-1/n)^n \to 0.632$, 실패	—
Huber 손실 (8.3)	$\rho_k' = \psi_k$, 연성 임계값의 최소	—

세 답은 모두 “모형을 덜 믿는” 방법이었다. 순위검정은 귀무가설 아래의 분포를 모형 없이 계산했고, 로버스트 추정은 모형이 조금 틀려도 크게 틀리지 않는 추정량을 골랐으며, 부트스트랩은 모형 대신 자료를 분포로 썼다. 그리고 셋 다 대가를 알려 주는 정리가 있었다. 0.864, 95%, $O(1/n)$이다. 그러나 4절의 Michelson이 보여 주듯, 어느 방법도 모든 관측에 공통인 편향은 고치지 못한다.

9.2 Stigler의 명명 법칙

> 이름과 발견
>
> - **Wilcoxon–Mann–Whitney**: 순위합 검정은 독일의 교육심리학자 Deuchler(1914) 등이 먼저 썼다(Kruskal 1957의 조사). Wilcoxon(1945)과 Mann–Whitney(1947)가 퍼뜨렸다.
> - **Pitman 효율**: Pitman의 1948년 강의록은 출판되지 않았고, Noether(1955)가 일반형으로 정리했다.
> - **잭나이프**: Quenouille(1949)가 치우침을 줄이려고 만들었고, 이름과 분산 추정은 Tukey(1958)의 것이다.
> - **부트스트랩**: 자료에서 다시 뽑는 생각은 Simon(1969)의 교육용 "재표집"에도 있었다. Efron(1979)이 이론과 이름을 주었다.
> - **Huber 손실**: 기계학습에서는 smooth L1이라는 이름으로 다시 발견되었다.

9.3 7 Pillars 지도

이번 회차는 일곱째 기둥 **Residual**(잔차)과 넷째 기둥 **Intercomparison**(자료 안의 비교)을 칠한다. 순열검정과 부트스트랩은 외부의 기준(정규분포표) 없이 자료 자신의 변동으로 자료를 판단한다. Stigler가 Intercomparison이라 부른 생각의 정점이다. 로버스트 회귀의 무게와 영향함수는 잔차를 진단의 도구로 바꾼다.

9.4 다음 회 — 베이즈의 귀환: MCMC와 Stein의 역설 (1763 → 1990)

이 과정은 1회 Bayes의 당구대에서 시작했다. 16회는 그리로 돌아간다. Laplace는 파리의 출생 기록으로 남아가 더 많이 태어날 확률을 계산했고, Fisher는 사전분포를 거부했으며, 베이즈 추론은 Bletchley Park의 암호 해독과 1950년대의 몬테카를로 속에서 살아남았다. Metropolis(1953)와 Gibbs 표집이 계산의 벽을 허물었고, James와 Stein(1961)은 정규 평균 세 개 이상을 동시에 추정할 때 표본평균이 최선이 아님을 보였다. 모수가 확률변수가 되는 순간, 13회의 ridge와 이번 회차의 Huber 손실은 사전분포가 된다.

참고문헌

- Pitman, E. J. G. Significance tests which may be applied to samples from any populations. *JRSS Suppl.* 4 (1937), 119–130.
- Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin* 1 (1945), 80–83. · Mann, H. B., Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the

other. *Ann. Math. Statist.* 18 (1947), 50–60.

- Hodges, J. L., Lehmann, E. L. The efficiency of some nonparametric competitors of the t-test. *Ann. Math. Statist.* 27 (1956), 324–335. · Estimates of location based on rank tests. *Ann. Math. Statist.* 34 (1963), 598–611. · Chernoff, H., Savage, I. R. Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* 29 (1958), 972–994.
- Tukey, J. W. A survey of sampling from contaminated distributions. In Olkin (ed.), *Contributions to Probability and Statistics*. Stanford, 1960. · Huber, P. J. Robust estimation of a location parameter. *Ann. Math. Statist.* 35 (1964), 73–101.
- Hampel, F. R. The influence curve and its role in robust estimation. *JASA* 69 (1974), 383–393. · Andrews, D. F. et al. *Robust Estimates of Location*. Princeton, 1972. · Hogg, R. V. Adaptive robust procedures. *JASA* 69 (1974), 909–923.
- Stigler, S. M. Do robust estimators deal with real data? *Ann. Statist.* 5 (1977), 1055–1098.
- Quenouille, M. H. Approximate tests of correlation in time series. *JRSS B* 11 (1949), 68–84. · Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7 (1979), 1–26. · Better bootstrap confidence intervals. *JASA* 82 (1987), 171–185.
- Brownlee, K. A. *Statistical Theory and Methodology in Science and Engineering*, 2nd ed. Wiley, 1965. · Rousseeuw, P. J. Least median of squares regression. *JASA* 79 (1984), 871–880.
- Breiman, L. Bagging predictors. *Machine Learning* 24 (1996), 123–140. · Random forests. *Machine Learning* 45 (2001), 5–32. · Girshick, R. Fast R-CNN. *ICCV* (2015).
- 자료: HistData `ZeaMays`, `Cavendish`, `Michelson`, `MichelsonSets`; MASS `newcomb`; datasets `stackloss`.

A

연습문제 풀이

각 풀이의 머리에 있는 문제 ○○쪽을 누르면 문제로 돌아간다. 계산 문제의 수치는 노트북에서 같은 계산으로 확인했다.

1절

연습 1.1 문제 8쪽

$\{1, \dots, 5\}$에서 두 개를 고른 합은 3, 4, 5, 6, 5, 6, 7, 7, 8, 9이므로 W는 3, 4, 8, 9에서 1/10, 5, 6, 7에서 2/10의 확률이다. $EW = 60/10 = 6$, $EW^2 = 390/10 = 39$이므로 $\operatorname{Var} W = 3$. 정리의 식 $m(N+1)/2 = 6$, $mn(N+1)/12 = 3$과 같다.

연습 1.2 문제 8쪽

순위 $m+n$(가장 큰 값)이 X의 것이면 나머지 $m-1$개의 X 순위는 $\{1, \dots, m+n-1\}$에서 합 $w-m-n$을 이루고, Y의 것이면 m개의 X 순위가 $\{1, \dots, m+n-1\}$에서 합 w를 이룬다. 두 경우는 겹치지 않는다.

연습 1.3 문제 8쪽

$d_i + d_j > 0$이려면 두 수 가운데 절댓값이 큰 쪽이 양수여야 한다. 절댓값의 순위가 k인 양수 d_i는 절댓값이 자기 이하인 k개의 d_j(자기 포함)와 짝을 이룰 때 합이 양수이고, 그 짝들에서 절댓값이 큰 쪽은 d_i다. 따라서 합이 양수인 짝의 수는 양수 d_i들의 순위의 합 T^+다.

연습 1.4 문제 8쪽

$U = W - m(m+1)/2$이므로 $\operatorname{Var}(U/mn) = \dfrac{mn(N+1)/12}{(mn)^2} = \dfrac{N+1}{12mn}$. $m = n = 100$이면 $\sqrt{201/120000} = 0.041$이고, $1.96 \times 0.041 = 0.080$이므로 AUC가 0.58보다 커야(또는 0.42보다 작아야) 5%에서 우연과 구별된다.

2절

연습 2.1 문제 12쪽

정규: $\int \phi^2 = \frac{1}{2\pi}\int e^{-x^2}dx = \frac{1}{2\sqrt{\pi}}$, $12\cdot\frac{1}{4\pi} = 3/\pi$. Laplace: $\int x^2\frac{1}{2}e^{-|x|} = 2$, $\int \frac{1}{4}e^{-2|x|} = \frac{1}{4}$, $12\cdot 2\cdot\frac{1}{16} = 1.5$. 로지스틱: $\sigma^2 = \pi^2/3$, $\int f^2 = \int_0^1 u(1-u)\,du = 1/6$ ($u = F(x)$, $f = u(1-u)$로 치환), $12\cdot\frac{\pi^2}{3}\cdot\frac{1}{36} = \pi^2/9$.

연습 2.2 문제 12쪽

$\int_{-\sqrt{5}}^{\sqrt{5}}(5-x^2)\,dx = \frac{20\sqrt{5}}{3}$이므로 $c = \frac{3}{20\sqrt{5}}$. $c\int x^2(5-x^2)\,dx = c\cdot\frac{20\sqrt{5}}{3} = 1$. $\int h^2 = c\int h(x)(5-x^2)\,dx$ (받침 위)$= c(5-1) = 4c = \frac{3}{5\sqrt{5}} = \frac{3\sqrt{5}}{25}$.

연습 2.3 문제 12쪽

정규에서는 $100/0.955 \approx 105$쌍, Laplace에서는 $100/1.5 \approx 67$쌍이다.

연습 2.4 문제 12쪽

$S = \frac{1}{n}\#\{X_i > 0\}$의 평균은 $P(X>0) = 1-F(-\theta)$이므로 $\mu'(0) = f(0)$, 귀무가설 아래 표준편차는 $1/(2\sqrt{n})$이어서 효능은 $2f(0)$. t의 효능은 $1/\sigma$이므로 ARE $= 4\sigma^2 f(0)^2$. 정규에서 $4\cdot\frac{1}{2\pi} = \frac{2}{\pi}$. 이것은 평균과 표본 중앙값의 분산의 비 $\sigma^2/(1/(4f(0)^2))$와 같다(8회 정리 1.4).

연습 2.5 문제 12쪽

(a) 귀무가설 아래 $Z_N = \sqrt{N}(T_N - \mu(0))/\sigma(0) \xrightarrow{D} \mathcal{N}(0,1)$이므로 크기 α의 한쪽 검정은 $Z_N > z_\alpha$에서 기각한다. $\theta_N = h/\sqrt{N}$ 아래에서는 상자의 가정으로 $Z_N \xrightarrow{D} \mathcal{N}(h\mu'(0)/\sigma(0), 1) = \mathcal{N}(he, 1)$이다. (b) $\mathcal{N}(he,1)$의 분포함수는 연속이므로 분포수렴의 정의가 모든 점에서, 특히 z_α에서 성립한다(7회 정의 2.1). 따라서 $P(Z_N > z_\alpha) = 1 - P(Z_N \le z_\alpha) \to 1 - \Phi(z_\alpha - he)$이다. (c) 고정된 작은 θ에서 표본 크기 N_j를 쓰면 $h_j = \theta\sqrt{N_j}$이고 극한 검정력은 $1-\Phi(z_\alpha - \theta\sqrt{N_j}e_j)$이다. Φ는 순증가이므로 두 검정력이 같은 것은 $\theta\sqrt{N_1}e_1 = \theta\sqrt{N_2}e_2$일 때이고, 곧 $N_2/N_1 = (e_1/e_2)^2$이다.

3절

연습 3.1 문제 17쪽

$-\log L(\theta) = \sum \rho_k(x_i - \theta) +$ 상수이고 ρ_k는 볼록하며 $\rho_k' = \psi_k$이다. 최소점의 조건은 $\sum \psi_k(x_i - \theta) = 0$이다.

연습 3.2 문제 17쪽

$\psi_k' = \mathbf{1}\{|u| < k\}$이므로 $E\psi_k' = 2\Phi(k) - 1$. $E\psi_k^2 = \int_{-k}^{k} u^2\phi + 2k^2\Phi(-k)$이고 $\int_{-k}^{k} u^2\phi = 2\Phi(k) - 1 - 2k\phi(k)$ (부분적분). $k = 1.345$에서 $E\psi_k' = 0.8214$, $E\psi_k^2 = 0.7102$, $V = 1.0526$, 효율 0.950.

연습 3.3 문제 17쪽

$x > m$이면 F_t의 중앙값 m_t는 $(1-t)F(m_t) = \frac{1}{2}$를 풀므로 $F(m_t) = \frac{1}{2(1-t)} \approx \frac{1}{2} + \frac{t}{2}$, 곧 $m_t \approx m + \frac{t}{2f(m)}$. $x < m$이면 $(1-t)F(m_t) + t = \frac{1}{2}$에서 $m_t \approx m - \frac{t}{2f(m)}$. $\mathrm{IF}^2 = 1/(4f(m)^2)$로 상수이므로 평균도 같다.

연습 3.4 문제 17쪽

평균 1/66, 중앙값 33/66 = 0.5. 10% 절사평균은 양쪽에서 $\lfloor 6.6 \rfloor = 6$개씩 자르므로 붕괴점은 $7/66 = 0.106$이고, 한쪽의 큰 오차가 6개까지는 모두 잘려 나간다. Newcomb의 두 오차는 넉넉히 버려진다.

연습 3.5 문제 17쪽

(a) $T(G)$는 $\int \psi(y - T(G))\, dG(y) = 0$의 해로 정의된다. $G = F_t = (1-t)F + t\delta_x$를 넣으면 $\int \psi\, dF_t = (1-t)\int \psi(y - T_t)\, dF(y) + t\,\psi(x - T_t)$이다. (b) (a)의 식을 t로 미분하면

$$-\int \psi(y - T_t)\, dF(y) - (1-t)\, T_t' \int \psi'(y - T_t)\, dF(y) + \psi(x - T_t) - t\, T_t' \psi'(x - T_t) = 0$$

이다. $t = 0$을 넣으면 $T_0 = T$이고 첫 항은 T의 정의로 0이므로 $T' = \psi(x - T)/E\psi'(X - T)$이다. 둘째 항의 미분과 적분의 교환은 ψ'이 유계이면 지배수렴으로 정당하다([T6]). (c) $\psi = \mathrm{sign}$이면 $E\,\mathrm{sign}(X - t) = 1 - 2F(t)$이므로 $-\frac{d}{dt} E\,\mathrm{sign}(X - t) = 2f(t)$이고, (b)의 분모 자리에 이것을 넣으면 $\mathrm{IF} = \mathrm{sign}(x - m)/(2f(m))$이다. Huber는 $\psi_k' = \mathbf{1}\{|u| < k\}$이므로 $E\psi_k'(X) = P(|X| < k) = 2\Phi(k) - 1$이다. (d) $E\,\mathrm{IF}(X)^2 = E\psi(X - T)^2/(E\psi')^2 = V(\psi, F)$이다(정리 3.2의 V).

연습 3.6 문제 17쪽

(a) x_1을 $y_1 = x_1 + c$로 바꾸면 $\bar{y} = \bar{x} + c/n$이고 c는 마음대로다. 따라서 $k^* = 1$이다. (b) $n = 2r - 1$이면 중앙값은 r번째 순서통계량이다. $k \le r - 1$개만 바꾸면 바뀌지 않은 원래 값이 $n - k \ge r$개 남는다. 이 값들은 모두 $\max x$ 이하이므로 y에서 $\max x$ 이하인 값이 r개 이상이고, $y_{(r)} \le \max x$이다. 같은 방법으로 $y_{(r)} \ge \min x$이다. $k = r$이면 r개를 모두 아주 큰 값 c로 바꿔 $y_{(r)} = c$로 만들 수 있다. 붕괴점은 $r/n = \lceil n/2 \rceil / n$이다. (c) $k \le g$이면 $\min x$보다 작은 y값은 기껏 $k \le g$개이므로 $y_{(g+1)} \ge \min x$이고, 같은 방법으로 $y_{(n-g)} \le \max x$이다. 남는 값 $y_{(g+1)}, \dots, y_{(n-g)}$는 모두 $[\min x, \max x]$에 있다. $k = g + 1$이면 $g + 1$개를 큰 값 c로 바꿔 $y_{(n-g)} = c$로 만들 수 있고 절사평균은 $c/(n - 2g)$ 이상으로 커진다. 붕괴점은 $(g+1)/n$이다.

4절

연습 4.1 문제 19쪽

편향의 제곱 $117.9^2 = 13{,}900$, 분산 $79.0^2/100 = 62.4$이므로 평균제곱오차의 99.6%가 편향이다. 로버스트 추정량이 줄일 수 있는 것은 분산 쪽뿐이고, 그것도 이 자료에서는 크지 않다.

연습 4.2 문제 19쪽

두 값을 빼면 64개의 평균은 $(66\times 26.21 + 44 + 2)/64 = 27.75$, 표준편차는 5.08이다(11회 예제 6.2의 계산과 같다). 로버스트 추정량들(27.0–27.5)과 가깝다. 이상치를 "골라 빼는" 것과 로버스트 추정량은 이 자료에서 거의 같은 답을 준다.

5절

연습 5.1 문제 24쪽

$\hat\theta_{(i)} = (n\bar X - X_i)/(n-1)$이므로 $\hat\theta_{(i)} - \hat\theta_{(\cdot)} = -(X_i - \bar X)/(n-1)$. $\frac{n-1}{n}\sum \frac{(X_i-\bar X)^2}{(n-1)^2} = \frac{s^2}{n}$.

연습 5.2 문제 24쪽

$X_{(m)}$ 이하의 관측을 빼면 남은 $2m-1$개의 중앙값은 $X_{(m+1)}$이고, $X_{(m+1)}$ 이상을 빼면 $X_{(m)}$이다. 각각 m번이므로 $\widehat{\mathrm{SE}}^2_{\mathrm{jack}} = \frac{n-1}{n}\cdot n\cdot (d/2)^2$, $d = X_{(m+1)} - X_{(m)}$이다. d는 가운데 이웃한 두 순서통계량의 간격으로, n을 곱해도 상수로 가지 않는 무작위한 양(대략 지수분포)이다. 그래서 잭나이프 분산과 참 분산의 비가 1로 가지 않는다.

연습 5.3 문제 24쪽

n번의 복원추출이 모두 관측 i를 피할 확률은 $(1-1/n)^n \to e^{-1}$. 서로 다른 관측의 수는 지시변수 n개의 합이므로 평균은 $n(1-(1-1/n)^n) \approx 0.632n$.

연습 5.4 문제 24쪽

$(1,2,6)$의 개수 (a,b,c), $a+b+c=3$인 다중집합이 10개이고 확률은 $\frac{3!}{a!b!c!}/27$이다. 중앙값은 1이 두 번 이상이면 1 ($\frac{1+3+3}{27} = \frac{7}{27}$), 6이 두 번 이상이면 6 ($\frac{7}{27}$), 그 밖에는 2 ($\frac{13}{27}$)이다.

6절

연습 6.1 문제 27쪽

$\psi = \text{sign}$이면 $w_i = \text{sign}(r_i)/r_i = 1/|r_i|$이고 고정점의 조건은 $\sum \text{sign}(r_i)\mathbf{x}_i = \mathbf{0}$, 곧 $\sum |r_i|$의 (부분) 기울기가 0인 조건이다. 잔차가 0에 가까우면 무게가 무한대로 가므로 $\max(|r_i|, \delta)$로 막는다.

연습 6.2 문제 27쪽

$k/|u_i| = 0.38$이므로 $|u_i| = 1.345/0.38 = 3.5$ (척도 단위).

7절

연습 7.1 문제 29쪽

$1 - 0.9^{10} = 0.651$, $1 - 0.98^{50} = 0.636$, $1 - 0.999^{1000} = 0.632$. 연속분포는 한 점에 확률을 주지 않으므로, 부트스트랩 분포의 큰 원자는 참 분포를 흉내 내지 못한다는 표시다.

연습 7.2 문제 29쪽

부트스트랩 표준오차 $\hat{\sigma}/\sqrt{n}$은 가장 큰 $|X_i|$ 하나에 지배된다. Cauchy에서 $\max|X_i|$는 n에 비례해 커지므로 $\hat{\sigma} \approx \max|X_i|/\sqrt{n}$은 $\sqrt{n}$처럼 커지고, $\hat{\sigma}/\sqrt{n}$은 줄지 않은 채 표본마다 크게 달라지는 무작위한 양으로 남는다(노트북: $n = 100$에서 10,000까지 중앙값이 1 근처). 평균의 참 분포는 n과 무관한 Cauchy이므로 "줄지 않는다"는 것은 맞지만, 부트스트랩 분포는 어떤 고정된 분포로도 수렴하지 않는다. 정리 5.3은 $E|X|^3 < \infty$(적어도 분산이 유한함)를 썼다.

연습 7.3 문제 29쪽

(1) 두 집단에서 각각 복원추출해 중앙값의 차의 부트스트랩 분포를 만든다. 중앙값은 분포의 매끄러운 함수이므로(밀도가 양수이면) 부트스트랩이 일치한다. (2) 각 집단의 중앙값 구간(8회 정리 2.6)을 만들어 차의 구간을 보수적으로 합친다. 둘 다 "중앙값이 같다"를 귀무가설로 삼으므로 분산이 달라도 크기를 대체로 지킨다. 순위합 검정은 $F = G$를 검정하므로 그렇지 않다.

연습 7.4 문제 29쪽

(a) $X_{(n)} \le s$는 모든 $X_i \le s$와 같으므로 $P(X_{(n)} \le s) = (s/\theta)^n$이다. $s = \theta(1 - x/n)$을 넣으면 $P\big(n(\theta - X_{(n)})/\theta \ge x\big) = (1 - x/n)^n$이고(연속이므로 $\ge$와 $>$가 같다), $\log(1 - x/n)^n = n\log(1 - x/n) \to -x$이므로 e^{-x}로 간다([T5], [T8]). (b) 복원추출 n번이 모두 최댓값이 아닌

관측을 뽑을 확률은 독립이므로 $(1-1/n)^n$이다. 그 여사건이 $X^*_{(n)} = X_{(n)}$이다(연속분포이므로 자료에 같은 값은 확률 1로 없다). (c) $V_n^* \geq 0$이고 $V_n^* = 0$은 $X^*_{(n)} = X_{(n)}$과 같으므로 $G_n(0) = 1-(1-1/n)^n \to 1-e^{-1} = 0.632$이다. Exp(1)의 분포함수는 0에서 연속이고 $G(0)=0$이므로, 분포수렴이라면 $G_n(0) \to 0$이어야 한다(7회 정의 2.1). 모순이다.

8절

연습 8.1 문제 32쪽

$0.4 + 0.6/B$: $B=1$이면 1, 10이면 0.46, 100이면 0.406, ∞이면 0.4.

연습 8.2 문제 32쪽

$z>0$이면 미분해 $z = u-k$이고 $u>k$일 때만 양수다. $z<0$이면 $z=u+k$이고 $u<-k$일 때만 음수다. 나머지 경우($|u|\leq k$)는 $z=0$이 최소다. 값은 $|u|\leq k$에서 $u^2/2$, $|u|>k$에서 $k^2/2 + k(|u|-k) = k|u| - k^2/2$.

연습 8.3 문제 32쪽

(a) $|u|<k$에서 $\rho_k'(u)=u$, $u>k$에서 k, $u<-k$에서 $-k$이다. $u=k$에서 왼쪽 도함수는 k, 오른쪽 도함수도 k이고 $u=-k$에서도 둘 다 $-k$이므로 ρ_k는 미분가능하고 $\rho_k' = \psi_k$이다. (b) z^*를 넣는다. $|u|\leq k$이면 $z^*=0$이고 값은 $u^2/2$다. $u>k$이면 $z^*=u-k$이고 값은 $\frac{1}{2}k^2 + k(u-k) = ku - \frac{k^2}{2}$다. $u<-k$도 대칭이다. (c) $|u|<\beta$이면 $\rho_\beta(u)/\beta = \frac{u^2}{2\beta}$이고, $|u|\geq\beta$이면 $\rho_\beta(u)/\beta = (\beta|u|-\beta^2/2)/\beta = |u| - \beta/2$이다. 둘 다 smooth L1의 식이다. $\beta\rho_\beta$는 첫 구간에서 $\beta u^2/2$가 되어 맞지 않는다.