수리통계학 × 통계학의 역사

15회 워크북 — 풀이판

가정이 틀리면
손으로 푸는 35문항

학생판과 같은 문제에 빈칸의 답과 완전한 풀이를 붙였다. 학생판으로 먼저 풀고, 힌트와 최종답으로도 막힐 때 이 판을 본다.

부	하는 일	문항	예상 시간
A 계산 드릴	순위 · 효율 · M-추정 · 잭나이프 · 부트스트랩을 손으로 계산한다	14	2시간 16분
B 빈칸 증명	노트의 단계형 증명을 책을 덮고 다시 쓴다	7	2시간 50분
C 증명 문제	변형, 일반화, 반례 만들기	10	2시간 47분
D 퀄형	여러 절을 잇는 종합 문제. 제한 시간 안에	4	2시간 20분
	합계	35	약 10시간

노트 `15_robust_bootstrap_notes.pdf` — 문항의 "정리 3.2" 같은 번호는 노트의 번호다

도구 `00_toolkit_notes.pdf` — [T1] 원소 추적, [T3] 귀납, [T4] 부등식, [T5] 극한, [T6] 극한과 적분의 교환, [T7] 변수변환, [T8] Taylor

검산 노트북 `15_robust_bootstrap.ipynb`의 "워크북 검산" 절: `check("A2c", "11/35")`

이 워크북을 쓰는 법

1. **노트를 덮고 푼다.** 문제의 머리 오른쪽에 쓸 도구, 관련 정리 번호, 예상 시간이 있다. 예상 시간의 두 배가 지나도 진척이 없으면 그때 노트를 편다.
2. **막히면 3단으로 본다.** 뒤의 **힌트와 최종답**에서 힌트를 먼저 본다. 다 푼 뒤에는 최종답과 맞춰 본다. 그래도 어긋나거나 증명이 막히면 풀이판을 본다.
3. **B부는 순서대로 푼다.** 같은 정리를 여러 번 만난다. 처음에는 근거만 빈칸이고(①), 다음에는 단계의 이름만 있고(②), 마지막에는 전략만 있다(③). 앞 단계의 답을 보면서 다음 단계를 쓰지 않는다.
4. **계산은 검산한다.** 최종답이 수나 식인 문항은 노트북의 `check()`에 넣으면 맞는지만 알려 준다. 정답은 보여 주지 않는다. 표를 써서 구한 소수 답은 표의 정밀도만큼 맞으면 맞다고 본다.
5. **증명은 근거를 단다.** 등호와 부등호마다 어느 정의, 정리, 앞 회차의 결과를 썼는지 오른쪽에 적는다. 노트의 단계형 증명과 같은 모양이다.

표시 ★ 정의를 바로 쓰는 문제, ★★ 두세 단계를 잇는 문제, ★★★ 퀄에 나오는 길이의 문제. [T1]–[T8]은 증명 도구 상자의 도구 번호다.

A

계산 드릴

순위, 정확한 귀무분포, 효율, M-추정, 잭나이프와 부트스트랩을 손으로 계산한다. 답이 수나식으로 떨어지므로 다 풀면 뒤의 최종답이나 노트북의 `check()`로 맞춰 본다. 표준정규 분위수가 필요한 곳은 표를 쓴다.

A1 순위합과 Mann–Whitney ★

정의 1.4, 정리 1.5(a) · 6분

$X = (1.8,\ 3.1,\ 4.6)$, $Y = (2.2,\ 2.9,\ 5.0,\ 5.7)$이다. (a) 합친 표본에서의 순위로 Wilcoxon 순위합 W를 구하라. (b) Mann–Whitney 통계량 $U = \#\{(i,j) : X_i > Y_j\}$를 직접 세고 $W = U + m(m+1)/2$를 확인하라.

풀이 정렬하면 $1.8_X,\ 2.2_Y,\ 2.9_Y,\ 3.1_X,\ 4.6_X,\ 5.0_Y,\ 5.7_Y$이므로 X의 순위는 $1, 4, 5$이고 (a) $W = 10$이다. (b) 1.8보다 작은 Y는 없고, 3.1과 4.6보다 작은 Y는 각각 두 개다. $U = 0 + 2 + 2 = 4$이고 $U + 3 \cdot 4/2 = 4 + 6 = 10 = W$이다.

최종답 (a) $W = 10$ (b) $U = 4$

A2 작은 표본의 정확한 귀무분포 ★★

정리 1.5(b) · 12분

$m = 3$, $n = 4$이고 $F = G$가 연속이다. (a) EW, (b) $\operatorname{Var} W$를 정리의 공식으로 구하라. (c) $\{1, \dots, 7\}$에서 고른 세 수의 합이 10 이하인 경우를 모두 세어 $P(W \le 10)$을 구하라.

풀이 $N = 7$. (a) $EW = 3 \cdot 8/2 = 12$. (b) $\operatorname{Var} W = 3 \cdot 4 \cdot 8/12 = 8$. (c) 합 6: $\{1,2,3\}$. 합 7: $\{1,2,4\}$. 합 8: $\{1,2,5\}, \{1,3,4\}$. 합 9: $\{1,2,6\}, \{1,3,5\}, \{2,3,4\}$. 합 10: $\{1,2,7\}, \{1,3,6\}, \{1,4,5\}, \{2,3,5\}$. 모두 $1+1+2+3+4 = 11$개이고, 보조정리 1.3으로 35개가 같은 확률이므로 $P(W \le 10) = 11/35 \approx 0.314$이다. 대칭으로 $P(W \ge 14)$도 같다.

최종답 (a) $EW = 12$ (b) $\operatorname{Var} W = 8$ (c) $\frac{11}{35} \approx 0.314$

A3 부호순위의 정확한 꼬리 확률 ★★

정리 1.8 · 12분

짝 차이가 $d = (1.2,\ -0.4,\ 2.5,\ 0.9,\ -1.7,\ 3.1)$이다. (a) T^+를 구하라. (b) 귀무가설 아래 $\operatorname{Var} T^+$를 구하라. (c) 정리 1.8의 표현 $T^+ = \sum k I_k$를 써서 정확한 $P(T^+ \ge 16)$을 구하라.

풀이 (a) 절댓값의 순위는 $0.4{:}1,\ 0.9{:}2,\ 1.2{:}3,\ 1.7{:}4,\ 2.5{:}5,\ 3.1{:}6$이고 양수의 순위는 $3, 5, 2, 6$이므로 $T^+ = 16$이다. (b) $\operatorname{Var} T^+ = 6 \cdot 7 \cdot 13/24 = 91/4 = 22.75$. (c) $T^+ \ge 16$은 양수가 아닌 순위들의 합이 $21 - T^+ \le 5$인 것과 같다. 합이 5 이하인 부분집합은 $\varnothing, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}$의 10개이므로 $P(T^+ \ge 16) = 10/64 = 5/32 \approx 0.156$이다.

최종답 (a) $T^+ = 16$ (b) $\frac{91}{4} = 22.75$ (c) $\frac{5}{32} \approx 0.156$

A4 Hodges–Lehmann 추정 ★

정의 1.9 · 8분

짝 차이 $d = (-1,\ 2,\ 3,\ 7)$에서 Walsh 평균 $(d_i + d_j)/2$ $(i \le j)$를 모두 적고 Hodges–Lehmann 추정값을 구하라. 양수인 Walsh 평균의 수가 T^+와 같음도 확인하라.

풀이 Walsh 평균을 정렬하면 $-1,\ 0.5,\ 1,\ 2,\ 2.5,\ 3,\ 3,\ 4.5,\ 5,\ 7$이다. 다섯째와 여섯째의 평균 $(2.5 + 3)/2 = 2.75$가 Hodges–Lehmann 추정값이다. 양수는 9개이고, 절댓값의 순위는 $1{:}1, 2{:}2, 3{:}3, 7{:}4$이므로 $T^+ = 2 + 3 + 4 = 9$로 같다(노트 연습 1.3).

최종답 2.75

A5 두 표본 순열검정 ★

정리 1.2 · 8분

$X = (5, 7, 9)$, $Y = (1, 2, 4)$이다. $T = \bar{X} - \bar{Y}$로 한쪽 순열검정을 할 때 정확한 p-값을 구하라.

풀이 s를 X 자리에 온 세 값의 합이라 하면 $\bar{X} - \bar{Y} = s/3 - (28 - s)/3 = (2s - 28)/3$이고, 관측값은 $s = 21$, $T = 14/3$이다. 여섯 값 $\{1, 2, 4, 5, 7, 9\}$에서 셋을 고른 $\binom{6}{3} = 20$가지 가운데 합이 21 이상인 것은 $\{5, 7, 9\}$ 하나뿐이다. 정리 1.2로 20가지가 같은 확률이므로 $p = 1/20 = 0.05$이다.

최종답 $\frac{1}{20}$

A6 두 분포에서의 효율 ★

정리 2.3 · 10분

ARE$(W,t) = 12\sigma^2(\int f^2)^2$을 (a) 지수분포 $f(x)=e^{-x}$ $(x>0)$와 (b) 삼각분포 $f(x)=1-|x|$ $(|x|<1)$에서 구하라.

풀이 (a) $\sigma^2=1$, $\int f^2=\frac{1}{2}$이므로 $12\cdot 1\cdot\frac{1}{4}=3$이다. 꼬리가 한쪽으로 긴 분포에서 Wilcoxon은 t의 세 배 효율이다. (b) $\sigma^2=2(\frac{1}{3}-\frac{1}{4})=\frac{1}{6}$, $\int f^2=2\cdot\frac{1}{3}=\frac{2}{3}$이므로 $12\cdot\frac{1}{6}\cdot\frac{4}{9}=\frac{8}{9}\approx 0.889$이다. 하한 0.864(정리 2.5)에 가깝다. 삼각분포는 포물선 밀도처럼 꼬리가 없다.

최종답 (a) 3 (b) $\frac{8}{9}\approx 0.889$

A7 부호검정의 효율 ★

명제 2.2, 노트 연습 2.4 · 8분

부호검정의 t 검정에 대한 ARE는 $4\sigma^2 f(0)^2$이다. (a) 표준 Laplace 밀도 $\frac{1}{2}e^{-|x|}$와 (b) 표준 로지스틱 밀도 $f(x)=e^{-x}/(1+e^{-x})^2$ $(\sigma^2=\pi^2/3)$에서 구하라.

풀이 (a) $\sigma^2=2$, $f(0)=\frac{1}{2}$이므로 $4\cdot 2\cdot\frac{1}{4}=2$이다. Laplace에서는 표본 중앙값이 MLE이고, 부호검정이 t의 두 배 효율이다. (b) $f(0)=\frac{1}{4}$이므로 $4\cdot\frac{\pi^2}{3}\cdot\frac{1}{16}=\frac{\pi^2}{12}\approx 0.822$이다.

최종답 (a) 2 (b) $\frac{\pi^2}{12}\approx 0.822$

A8 Huber 추정량을 손으로 ★★

정의 3.1, 알고리즘 6.2 · 12분

자료 (1, 2, 3, 4, 20), $k=1.5$, 척도 $s=1$로 고정한다. (a) $\sum_i \psi_k(x_i-\theta)=0$의 해를 구하라. (b) 평균 $\theta^{(0)}=6$에서 시작해 무게 $w_i=\min(1,\,k/|x_i-\theta^{(0)}|)$로 가중평균 한 번 $(\theta^{(1)}=\sum w_i x_i/\sum w_i)$을 계산하라.

풀이 (a) θ가 3 근처이면 $x=1$의 잔차는 -1.5로, $x=20$의 잔차는 1.5로 잘리고 나머지 셋은 그대로다. $-1.5+(2-\theta)+(3-\theta)+(4-\theta)+1.5=9-3\theta=0$에서 $\theta=3$이다. 확인: 잔차 $-2,-1,0,1,17$에서 $|-2|,|17|>1.5$이고 나머지는 1.5 이하다. 중앙값도 3, 평균은 6이다. (b) 잔차 $-5,-4,-3,-2,14$에서 무게는 $\frac{3}{10},\frac{3}{8},\frac{1}{2},\frac{3}{4},\frac{3}{28}$이다. $\sum w_i x_i=\frac{3}{10}+\frac{3}{4}+\frac{3}{2}+3+\frac{15}{7}$, $\sum w_i=\frac{3}{10}+\frac{3}{8}+\frac{1}{2}+\frac{3}{4}+\frac{3}{28}$이고, 분모 280으로 통분하면 $\theta^{(1)}=\frac{2154}{569}\approx 3.786$이다. 한 걸음에 6에서 3.79로 내려왔다.

최종답 (a) $\hat\theta=3$ (b) $\theta^{(1)}=\frac{2154}{569}\approx 3.786$

A9 영향함수의 값 ★

명제 3.5 · 8분

$F=\mathcal{N}(0,1)$에서 $x=3$의 영향함수 값을 (a) Huber 추정량($k=1.345$), (b) 평균, (c) 중앙값에 대해 구하라. $\Phi(1.345)=0.9107$을 쓴다.

풀이 (a) $\psi_k(3)/(2\Phi(k)-1)=1.345/0.8214\approx 1.637$. (b) $x-\mu=3$. (c) $\text{sign}(3)/(2\phi(0))=\sqrt{2\pi}/2=\sqrt{\pi/2}\approx 1.253$. x를 더 키우면 (a)와 (c)는 그대로이고 (b)만 커진다.

최종답 (a) 1.637 (b) 3 (c) $\sqrt{\pi/2}\approx 1.253$

A10 잭나이프의 치우침 수정 ★★

정의 5.6 · 10분

자료 $(1,3,4,8)$에서 $\hat\theta=\bar X^2$으로 μ^2을 추정한다. 관측 i를 뺀 값 $\hat\theta_{(i)}$와 그 평균 $\hat\theta_{(\cdot)}$으로 (a) 잭나이프 치우침 추정 $(n-1)(\hat\theta_{(\cdot)}-\hat\theta)$와 (b) 수정한 추정량 $\hat\theta-(n-1)(\hat\theta_{(\cdot)}-\hat\theta)$를 구하라. (b)가 $\bar X^2-s^2/n$과 같은지 확인하라.

풀이 $\bar X=4$, $\hat\theta=16$. 하나를 뺀 평균은 5, $\frac{13}{3}$, 4, $\frac{8}{3}$이고 제곱은 25, $\frac{169}{9}$, 16, $\frac{64}{9}$이다. 평균은 $\hat\theta_{(\cdot)}=\frac14\left(41+\frac{233}{9}\right)=\frac{602}{36}=\frac{301}{18}$이다. (a) $3\left(\frac{301}{18}-16\right)=3\cdot\frac{13}{18}=\frac{13}{6}$. (b) $16-\frac{13}{6}=\frac{83}{6}\approx 13.83$. $s^2=\frac{9+1+0+16}{3}=\frac{26}{3}$이므로 $\bar X^2-s^2/n=16-\frac{26}{12}=\frac{83}{6}$으로 같다. $E\bar X^2=\mu^2+\sigma^2/n$이므로 이것은 μ^2의 불편추정량이다.

최종답 (a) $\frac{13}{6}$ (b) $\frac{83}{6}\approx 13.83$

A11 정확한 부트스트랩 분포 ★★

정리 5.3(a) · 10분

자료 $(0,3,9)$에서 크기 3의 부트스트랩 표본의 평균 $\bar X^*$를 생각한다. (a) $\text{Var}^*\bar X^*$를 정리 5.3(a)로 구하라. (b) $3^3=27$가지 순서 있는 표본을 세어 $P^*(\bar X^*>\bar X)$를 구하라.

풀이 $\bar X=4$, $\hat\sigma^2=\frac{16+1+25}{3}=14$. (a) $\text{Var}^*\bar X^*=14/3$. (b) 합이 12보다 커야 한다. 9가 두 번 이상이면 언제나 크다: 두 번 $3\cdot 2=6$가지, 세 번 1가지. 9가 한 번이면 나머지 둘의 합이 3보다 커야 하므로 $(3,3)$뿐이고 순서까지 3가지다. 9가 없으면 최대 9로 작다. 모두 $6+1+3=10$가지이므로 $10/27\approx 0.370$이다. 분포가 오른쪽으로 치우쳐 평균보다 큰 쪽의 확률이 반이 안 된다.

최종답 (a) $\frac{14}{3}$ (b) $\frac{10}{27}$

A12 부트스트랩 최댓값이 둘째로 큰 값일 확률 ★★

정리 7.1 · 8분

서로 다른 값 $n = 5$개에서 부트스트랩 표본을 뽑는다. $X^*_{(n)} = X_{(n-1)}$, 곧 부트스트랩 표본의 최댓값이 원래 자료의 둘째로 큰 값일 확률을 구하라.

풀이 $P^*(X^*_{(n)} \le X_{(n-1)}) = (1 - \frac{1}{n})^n$이고 $P^*(X^*_{(n)} \le X_{(n-2)}) = (1 - \frac{2}{n})^n$이므로 $(\frac{4}{5})^5 - (\frac{3}{5})^5 = \frac{1024-243}{3125} = \frac{781}{3125} \approx 0.250$이다. $n \to \infty$이면 $e^{-1} - e^{-2} = 0.233$으로 간다. 부트스트랩 최댓값의 분포는 자료의 맨 위 몇 값에 고정된 원자들로 남는다.

최종답 $\frac{781}{3125} \approx 0.250$

A13 BCa의 수정된 분위수 ★★

정의 5.4 · 12분

$B = 1000$개의 부트스트랩 값 가운데 560개가 $\hat\theta$보다 작았고, 잭나이프로 구한 가속은 $\hat a = 0.05$이다. (a) $\hat z_0$, (b) 95% 구간의 아래쪽 수준 α_1 ($z_{(1)} = -1.960$), (c) 위쪽 수준 α_2 ($z_{(2)} = 1.960$)를 구하라. $\Phi^{-1}(0.56) = 0.151$을 쓴다.

풀이 (a) $\hat z_0 = \Phi^{-1}(0.56) \approx 0.151$. (b) $\hat z_0 + z_{(1)} = -1.809$, 분모 $1 + 0.05 \cdot 1.809 = 1.090$이므로 $\alpha_1 = \Phi(0.151 - 1.659) = \Phi(-1.508) \approx 0.0658$. (c) $\hat z_0 + z_{(2)} = 2.111$, 분모 $1 - 0.106 = 0.894$이므로 $\alpha_2 = \Phi(0.151 + 2.360) = \Phi(2.511) \approx 0.9940$. 구간은 부트스트랩 분포의 6.6%와 99.4% 분위수다. 백분위 구간(2.5%, 97.5%)보다 위로 옮겨 간다.

최종답 (a) $\hat z_0 \approx 0.151$ (b) $\alpha_1 \approx 0.0658$ (c) $\alpha_2 \approx 0.9940$

A14 Laplace에서의 Huber 점근분산 ★★

정리 3.2 · 12분

표준 Laplace 밀도 $\frac{1}{2}e^{-|x|}$에서 $k = 1$인 Huber 추정량의 $V(\psi_1, F) = E\psi_1^2 / (E\psi_1')^2$을 구하라. 같은 분포에서 평균의 V는 2, 중앙값의 V는 1이다.

풀이 $E\psi_1' = P(|X| < 1) = 1 - e^{-1}$. $\int_0^1 x^2 e^{-x} dx = 2 - 5e^{-1}$이고 $P(|X| > 1) = e^{-1}$이므로 $E\psi_1^2 = 2 - 4e^{-1}$이다. $V = \frac{2 - 4/e}{(1 - 1/e)^2} \approx \frac{0.5285}{0.3996} \approx 1.323$이다. 평균(2)보다 34% 작고, MLE인 중앙값(1)보다 32% 크다.

최종답 $\frac{2 - 4/e}{(1 - 1/e)^2} \approx 1.323$

B

빈칸 증명

노트의 단계형 증명을 책을 덮고 다시 쓴다. 같은 정리를 여러 번 만나며 도움이 줄어든다. ① 단계는 주어지고 근거만 빈칸이다. ② 단계의 이름만 주어진다. ③ 전략만 주어진다. 근거에는 정리나 정의의 번호, 앞 회차의 정리, 도구 번호를 쓴다.

B1 순위합의 귀무분포, 근거 채우기 ★

① · 정리 1.5(a)(b) · 15분

$F = G$가 연속일 때 $W = U + \frac{m(m+1)}{2}$, $EW = \frac{m(N+1)}{2}$, $\operatorname{Var} W = \frac{mn(N+1)}{12}$이다. 각 단계의 근거를 써라.

① $X_{(k)}$의 합친 순위는 $(k-1) + \#\{j : Y_j < X_{(k)}\} + 1$이다. — 순위의 정의, 동점의 확률 0

② k에 대해 더하면 $W = \frac{m(m+1)}{2} + U$이다. — $\sum_{k=1}^{m} k$, U의 정의

③ $(R(X_1), \dots, R(X_m))$은 $\{1, \dots, N\}$에서 m개를 비복원으로 뽑은 것과 같은 분포이다. — 보조정리 1.3의 주변분포

④ 모집단 $\{1, \dots, N\}$의 평균은 $\frac{N+1}{2}$, 분산은 $\frac{N^2-1}{12}$이다. — $\sum k^2$의 공식 [T3]

⑤ $\operatorname{Var} W = m \cdot \frac{N^2-1}{12} \cdot \frac{N-m}{N-1} = \frac{mn(N+1)}{12}$이다. — 비복원 표본합, 4회 정리 1.5

⑥ W와 $m(N+1) - W$는 같은 분포다. — $r \mapsto N+1-r$이 균등분포를 보존

풀이 빈칸의 빨간 글씨가 근거다. 노트 정리 1.5의 증명과 같다. 다섯째 단계의 $\frac{N-m}{N-1}$은 유한 모집단 수정이다. $N = m + n$이므로 $(N^2-1)\frac{N-m}{N-1} = (N+1)n$이다.

B2 M-추정량의 점근정규성, 근거 채우기 ★★

① · 정리 3.2 · 20분

정리 3.2의 조건(대칭, 홀함수이고 단조증가하는 ψ, ψ'' 유계)에서 $\sqrt{n}\,\hat\theta \xrightarrow{D} \mathcal{N}(0, V)$이다. 근거를 채워라. $\Psi_n(t) = \frac{1}{n}\sum \psi(X_i - t)$, $\lambda(t) = E\psi(X - t)$.

① Ψ_n은 감소하고 $\hat\theta$는 그 영점이다. — ψ의 단조성, 정의 3.1

② $\lambda(0) = 0$이고 $\lambda'(0) = -E\psi'(X) < 0$이다. — 대칭·홀함수, 적분 기호 아래 미분 [T6]

③ $\hat\theta \xrightarrow{P} 0$이다. — $\Psi_n(\pm\varepsilon) \xrightarrow{P} \lambda(\pm\varepsilon)$, Khinchin

④ $0 = \Psi_n(0) - \hat\theta A_n + \frac{1}{2}\hat\theta^2 B_n$이다. — Taylor 전개 [T8]

⑤ $\sqrt{n}\,\hat\theta = \dfrac{n^{-1/2}\sum \psi(X_i)}{A_n - \frac{1}{2}\hat\theta B_n}$이다. — ④를 정리

⑥ 분자 $\xrightarrow{D} \mathcal{N}(0, E\psi^2)$이다. — $E\psi(X) = 0$, CLT 7회 정리 3.1

⑦ 분모 $\xrightarrow{P} E\psi'(X)$이다. — Khinchin, $|B_n| \le \sup|\psi''|$, ③

⑧ $\sqrt{n}\,\hat\theta \xrightarrow{D} \mathcal{N}(0, E\psi^2/(E\psi')^2)$이다. — Slutsky, 7회 정리 4.2

풀이 빈칸의 빨간 글씨가 근거다. 노트 정리 3.2의 증명과 같다. ③의 논증: 감소함수 Ψ_n이 ε에서 음수, $-\varepsilon$에서 양수이면 영점은 $(-\varepsilon, \varepsilon)$ 안에 있다. 그 사건의 확률이 1로 간다.

B3 하한 108/125, 근거 채우기 ★★

① · 정리 2.5 · 15분

분산이 유한한 모든 밀도에서 $12\sigma^2(\int f^2)^2 \ge \frac{108}{125}$이다. 근거를 채워라. $h(x) = c(5 - x^2)_+$, $c = \frac{3}{20\sqrt{5}}$, $a = 5c$, $b = c$.

① 평균 0, 분산 1로 놓아도 된다. — 위치·척도 변환에 불변 [T7]

② $\int h = 1$, $\int x^2 h = 1$이다. — 직접 적분

③ $\int f^2 \ge 2\int fh - \int h^2$이다. — $\int(f - h)^2 \ge 0$ [T4]

④ $\int fh \ge \int f(a - bx^2) = a - b$이다. — $h \ge a - bx^2$, $f \ge 0$, $\int x^2 f = 1$

⑤ $\int h^2 = a - b = \frac{3\sqrt{5}}{25}$이다. — 받침 위에서 $h = a - bx^2$

⑥ $12(\int f^2)^2 \ge 12 \cdot \frac{45}{625} = \frac{108}{125}$이다. — ③–⑤

풀이 빈칸의 빨간 글씨가 근거다. 노트 정리 2.5의 증명과 같다. 등호는 $\int(f - h)^2 = 0$, 곧 $f = h$일 때다.

B4 순위합의 귀무분포, 단계의 이름만 ★★

② · 정리 1.5 · 25분

정리 1.5의 (a)(b)(c)를 증명하라. 단계의 이름만 주어진다. 각 단계 아래에 식과 근거를 써라.

① **(a)** $X_{(k)}$**의 순위를 세 부분으로 쪼갠다**

② k**에 대해 더한다**

③ **(b)** X **쪽 순위를 비복원 표본으로 본다**

④ **모집단** $\{1,\dots,N\}$**의 평균과 분산**

⑤ **비복원 표본합의 평균과 분산**

⑥ **대칭:** $r \mapsto N+1-r$

⑦ **(c)** U**를 지시변수의 합으로 쓰고 기댓값을 취한다**

풀이

① $X_{(k)}$의 합친 순위 $= (k-1) + \#\{Y_j < X_{(k)}\} + 1.$ — 순위의 정의

② $W = \frac{m(m+1)}{2} + U.$ — U의 정의

③ $(R(X_i))_{i\le m}$은 $\{1,\dots,N\}$의 비복원 표본. — 보조정리 1.3

④ 평균 $\frac{N+1}{2}$, 분산 $\frac{N^2-1}{12}$. — [T3]

⑤ $EW = \frac{m(N+1)}{2}$, $\operatorname{Var} W = \frac{mn(N+1)}{12}$. — 4회 정리 1.5

⑥ $W \overset{d}{=} m(N+1) - W.$ — 균등분포의 보존

⑦ $EU = \sum_{i,j} P(X_i > Y_j) = mn\,P(X > Y).$ — 선형성

B5 M-추정량의 점근정규성, 단계의 이름만 ★★

② · 정리 3.2 · 30분

정리 3.2를 증명하라(ψ''가 유계라고 가정해도 된다).

① **영점** $\hat\theta$**와 감소하는** Ψ_n

② $\lambda(t) = E\psi(X-t)$**의 0 근처 모양**

③ **일치성** $\hat\theta \xrightarrow{P} 0$

④ **영점의 Taylor 전개**

⑤ **비로 쓰기**

⑥ **분자: 중심극한정리**

⑦ **분모: 대수의 법칙과 나머지**

⑧ **Slutsky로 합치기**

풀이 노트 정리 3.2의 단계형 증명과 같다. 단계별 근거는 B2의 풀이에 있다. 9회 정리 3.1과 다른 점은 둘이다. 일치성을 가능도가 아니라 Ψ_n의 단조성으로 얻고, 분자의 분산 $E\psi^2$과 분모의 극한 $E\psi'$이 같지 않아 극한분산이 $1/I$ 대신 샌드위치 꼴 $E\psi^2/(E\psi')^2$이 된다. ψ가 점수함수이면 두 표현이 같아져(9회 정리 2.2) $1/I$로 돌아간다.

B6 순위합의 귀무분포, 전략만 ★★

③ · 정리 1.5 · 30분

정리 1.5의 (a)(b)(c)를 증명하라.

전략 (a)는 X의 순위를 "작은 X의 수 + 작은 Y의 수 + 1"로 센다. (b)는 순위 벡터의 균등성으로 W를 비복원 표본합으로 바꾼다. (c)는 지시변수의 합이다.

풀이 B4의 풀이와 같다. 이 증명에서 분포 F가 쓰인 곳은 "동점의 확률이 0"과 "교환가능"뿐이다. 그래서 W의 귀무분포는 F와 무관한 표 하나로 정해진다. (c)만은 귀무가설을 쓰지 않으므로 대립가설 아래에서도 $U/(mn)$은 $P(X > Y)$의 불편추정량이다.

B7 하한 108/125, 전략만 ★★★

③ · 정리 2.5 · 35분

분산이 유한한 모든 밀도 f에서 $\mathrm{ARE}(W, t) = 12\sigma^2(\int f^2)^2 \geq \frac{108}{125}$임을 증명하고, 등호가 성립하는 밀도를 밝혀라.

전략 위치와 척도를 정규화하고, 포물선 밀도 h와의 차의 제곱을 적분해 $\int f^2$를 아래에서 누른다.

풀이 노트 정리 2.5의 단계형 증명과 같다(근거는 B3의 풀이). $h = c(r^2 - x^2)_+$ 꼴에서 $\int h = \frac{4}{3}cr^3 = 1$, $\int x^2 h = \frac{4}{15}cr^5 = 1$을 풀면 $r^2 = 5$, $c = \frac{3}{4r^3} = \frac{3}{20\sqrt{5}}$이다. 이 증명은 "$\int f^2$를 분산 1의 조건에서 최소로 하는 밀도"를 찾은 것이고, 답은 라그랑주 승수법으로 얻는 $(a - bx^2)_+$ 꼴이다.

C

증명 문제

노트의 정리를 변형하고, 일반화하고, 반례로 경계를 확인한다. 증명은 노트의 단계형 모양으로 쓴다. 한 줄에 한 단계, 줄마다 근거.

C1 순열 p-값의 정확한 분포 ★★

정리 1.2 · 15분

정리 1.2의 상황에서 $M = \binom{N}{m}$이라 하자. 합친 표본의 값이 주어졌을 때 M가지 나눔의 T 값이 모두 다르면, 귀무가설 아래 $k = 1, \dots, M$에서 $P(p \le k/M) = k/M$임을 보여라. p의 가장 작은 값은 무엇이고, 그것이 크기 $\alpha < 1/M$의 검정에 대해 무엇을 뜻하는가?

풀이 값의 다중집합 $\mathcal{V}$가 주어지면 관측된 나눔은 M가지 위에서 균등하다(정리 1.2의 증명 ②). T 값이 모두 다르면 M가지를 T가 큰 순서로 늘어놓을 수 있고, j번째로 큰 나눔이 관측되면 $p = j/M$이다. 따라서 조건부로 $P(p = j/M \mid \mathcal{V}) = 1/M$이고 $P(p \le k/M \mid \mathcal{V}) = k/M$이다. 이 값이 $\mathcal{V}$와 무관하므로 무조건으로도 같다. p의 최솟값은 $1/M$이므로 $\alpha < 1/M$이면 검정은 결코 기각하지 않는다. A5의 $m = n = 3$에서는 $M = 20$이라 5%보다 작은 p-값은 나올 수 없다.

C2 지시변수로 구하는 $\operatorname{Var} U$ ★★

정리 1.5(b), 3회 정리 4.2(c) · 20분

$U = \sum_{i,j} I_{ij}$, $I_{ij} = \mathbf{1}\{X_i > Y_j\}$이다. 귀무가설($F = G$ 연속) 아래 다음을 보이고 $\operatorname{Var} U = \frac{mn(N+1)}{12}$을 순위를 쓰지 않고 다시 얻어라.

(a) $\operatorname{Var} I_{ij} = \frac{1}{4}$.

(b) $j \neq k$이면 $\operatorname{Cov}(I_{ij}, I_{ik}) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$. $i \neq l$이면 $\operatorname{Cov}(I_{ij}, I_{lj}) = \frac{1}{12}$.

(c) 첨자가 하나도 겹치지 않으면 공분산은 0이다.

풀이 (a) $P(X > Y) = \frac{1}{2}$이므로 $\frac{1}{2} \cdot \frac{1}{2}$. (b) $P(X_i > Y_j, X_i > Y_k)$는 세 iid 연속변수 가운데 X_i가 최대일 확률 $\frac{1}{3}$이다(보조정리 1.3). 같은 방법으로 $P(X_i > Y_j, X_l > Y_j) = \frac{1}{3}$. (c) 독립이다. 분산을 공분산으로 전개하면(3회 정리 4.2(c))

$$\operatorname{Var} U = mn \cdot \frac{1}{4} + mn(n-1) \cdot \frac{1}{12} + mn(m-1) \cdot \frac{1}{12} = \frac{mn}{12}\big(3 + n - 1 + m - 1\big) = \frac{mn(N+1)}{12}.$$

정리 1.5(a)로 $\operatorname{Var} W = \operatorname{Var} U$이므로 정리 1.5(b)와 같다.

C3 Hodges–Lehmann 추정량의 붕괴점 ★★★

정의 1.9, 3.6 · 25분

한 표본의 Hodges–Lehmann 추정량은 Walsh 평균 $M = n(n+1)/2$개의 중앙값이다.

(a) k개의 관측을 바꾸면 영향을 받는 Walsh 평균의 수가 $M - \frac{(n-k)(n-k+1)}{2}$임을 보여라.

(b) 영향을 받는 Walsh 평균이 $M/2$보다 많아야 추정량을 끝없이 보낼 수 있다. 붕괴점 k^*/n이 $n \to \infty$에서 $1 - 1/\sqrt{2} \approx 0.293$으로 감을 보여라.

(c) 평균(0), 중앙값($\frac{1}{2}$)과 비교하라.

풀이 (a) Walsh 평균 $(d_i + d_j)/2$가 바뀐 관측을 하나도 쓰지 않으려면 i, j 모두 바뀌지 않은 $n-k$개 안에 있어야 하고, 그런 쌍$(i \le j)$은 $\frac{(n-k)(n-k+1)}{2}$개다. (b) 영향을 받은 값을 모두 $+\infty$ 쪽으로 보내면, 그 수가 $M/2$ 이하일 때 중앙값은 영향받지 않은 Walsh 평균들의 범위 안에 머물고, $M/2$보다 많으면 끝없이 커진다. 조건 $M - \frac{(n-k)(n-k+1)}{2} > \frac{M}{2}$는 $(n-k)(n-k+1) < \frac{n(n+1)}{2}$와 같다. 양변을 n^2으로 나누고 $n \to \infty$이면 $(1-k/n)^2 < \frac{1}{2}$, 곧 $k/n > 1 - 1/\sqrt{2}$이다. 따라서 $k^*/n \to 1 - 1/\sqrt{2} \approx 0.293$이다. (c) 평균은 관측 하나로 무너지고(붕괴점 $1/n \to 0$), 중앙값은 절반까지 견딘다. Hodges–Lehmann은 그 사이이고, 정규에서 효율은 $3/\pi = 0.955$로 중앙값의 $2/\pi = 0.637$보다 훨씬 높다.

최종답 (c) $1 - \frac{1}{\sqrt{2}} \approx 0.293$

C4 부호검정의 효율은 0에서 무한대까지 ★★

명제 2.2, 노트 연습 2.4 · 15분

부호검정의 t 검정에 대한 ARE $4\sigma^2 f(0)^2$에 대해

(a) ARE가 0인 대칭 밀도를 만들어라.

(b) t_ν 분포$(\nu > 2,\ \sigma^2 = \nu/(\nu-2))$에서 $\nu \downarrow 2$이면 ARE가 무한대로 감을 보여라.

(c) t_3에서 ARE를 구하라. t_3의 $f(0) = \frac{2}{\pi\sqrt{3}}$이다.

풀이 (a) $(-2,-1) \cup (1,2)$ 위의 균등밀도는 대칭이고 $f(0) = 0$이므로 ARE $= 0$이다. 중앙값 근처에 관측이 없으면 부호의 개수는 위치의 작은 이동을 거의 보지 못한다. (b) $f_\nu(0) = \Gamma(\frac{\nu+1}{2})/(\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2}))$는 $\nu \to 2$에서 $\Gamma(\frac{3}{2})/\sqrt{2\pi} = \frac{1}{2\sqrt{2}} > 0$으로 가고, $\sigma^2 = \nu/(\nu-2) \to \infty$이므로 ARE $\to \infty$이다. (c) $4 \cdot 3 \cdot \frac{4}{3\pi^2} = \frac{16}{\pi^2} \approx 1.62$. 반례 (a)는 "ARE의 하한이 있다"는 정리 2.5가 부호검정에는 없음을 보여 준다.

최종답 (c) $\frac{16}{\pi^2} \approx 1.62$

C5 M-추정량의 이동 등변성과 유일성 ★

정의 3.1 · 12분

척도 s를 고정한 위치 M-추정량에서 ψ가 연속이고 순증가하는 홀함수라 하자.

(a) 방정식 $\sum \psi((x_i-\theta)/s)=0$의 해가 하나뿐임을 보여라.

(b) 모든 x_i에 c를 더하면 해도 c만큼 움직임을 보여라.

(c) 자료가 어떤 점 c에 대해 대칭이면(곧 $\{x_i-c\}=\{c-x_i\}$) 해는 c임을 보여라.

풀이 (a) g는 순감소하는 연속함수이고, $\theta<\min x_i$이면 모든 항이 양수, $\theta>\max x_i$이면 모든 항이 음수다. 중간값 정리로 영점이 있고, 순감소이므로 하나뿐이다. (b) $\sum\psi((x_i+c-(\theta+c))/s)=g(\theta)$이므로 θ가 원래 방정식의 해이면 $\theta+c$가 새 방정식의 해이고, (a)로 유일하다. (c) $g(c)=\sum\psi((x_i-c)/s)$에서 x_i-c들의 집합이 부호를 바꿔도 같으므로, 홀함수 ψ의 값들도 짝을 지어 지워지고 $g(c)=0$이다. Huber의 ψ_k는 순증가가 아니지만($|u|>k$에서 평평하다) 가운데 구간에 관측이 하나라도 있으면 같은 논증이 된다.

C6 분산 추정량의 잭나이프 ★★

정의 5.6 · 15분

$\hat\theta=\hat\sigma^2=\frac{1}{n}\sum(X_i-\bar X)^2$이다. 잭나이프로 치우침을 고친 추정량 $n\hat\theta-(n-1)\hat\theta_{(\cdot)}$가 $s^2=\frac{1}{n-1}\sum(X_i-\bar X)^2$과 정확히 같음을 보여라.

풀이 분산을 쌍의 차로 쓰면 $\hat\sigma^2=\frac{1}{n^2}\sum_{j<l}(X_j-X_l)^2$이다. 관측 i를 빼면 $\hat\theta_{(i)}=\frac{1}{(n-1)^2}\sum_{j<l;\ j,l\neq i}(X_j-X_l)^2$이다. 쌍 (j,l)은 $i\notin\{j,l\}$인 $n-2$개의 i에서 나오므로 $\sum_i\hat\theta_{(i)}=\frac{n-2}{(n-1)^2}S$, $S=\sum_{j<l}(X_j-X_l)^2$이다. 그러면

$$n\hat\theta-(n-1)\hat\theta_{(\cdot)}=\frac{S}{n}-\frac{(n-1)(n-2)}{n(n-1)^2}S=\frac{S}{n}\Big(1-\frac{n-2}{n-1}\Big)=\frac{S}{n(n-1)}=s^2.$$

마지막은 $s^2=\frac{n}{n-1}\hat\sigma^2=\frac{S}{n(n-1)}$이다. 잭나이프는 치우침이 정확히 $1/n$ 꼴인 추정량의 치우침을 완전히 지운다.

C7 부트스트랩은 왜도를 본다 ★★

정리 5.3 · 15분

(a) $E^*\big[(\bar X^*-\bar X)^3\big]=\hat\mu_3/n^2$, $\hat\mu_3=\frac{1}{n}\sum(X_i-\bar X)^3$임을 보여라. (b) 자료 $(0,3,9)$에서 이 값을 구하고, A11의 27가지 표본으로 확인하라. (c) 정규근사 $\mathcal{N}(\bar X,\hat\sigma^2/n)$은 이 3차 적률을 얼마로 주는가?

풀이 (a) $\bar X^*-\bar X=\frac{1}{n}\sum D_i$이고 D_i는 P^*에서 iid, $E^*D_i=0$, $E^*D_i^3=\hat\mu_3$이다. $(\sum D_i)^3$의 기댓값에서 $E^*D_i^2D_j=E^*D_i^2\,E^*D_j=0$ $(i\neq j)$, $E^*D_iD_jD_k=0$ (모두 다를 때)이므로 $E^*(\sum D_i)^3=n\hat\mu_3$이고, n^3으로 나누면 $\hat\mu_3/n^2$이다. (b) 편차 $-4,-1,5$의 세제곱의 평균은 $(-64-1+125)/3=20$이므로 $20/9$이다. 27가지의 $(\bar X^*-4)^3$의 평균도 $20/9$이다. (c) 정규분

포는 대칭이므로 0이다. 부트스트랩 분포는 자료의 치우침을 $1/\sqrt{n}$의 차수로 담고, 이것이 BCa가 정규근사보다 정확한 이유의 한 부분이다(정의 5.4 아래 상자).

최종답 (b) $\frac{20}{9}$

C8 IRLS는 목적함수를 줄인다 ★★★

명제 6.3, 명제 8.3 · 25분

위치 문제에서 Huber 목적함수 $L(\theta)=\sum_i \rho_k(x_i-\theta)$를 생각한다(척도 1).

(a) $g(t)=\rho_k(\sqrt{t})$ $(t\ge 0)$가 오목함수임을 보여라.

(b) 오목성으로 $\rho_k(u)\le \rho_k(u_0)+\frac{w_0}{2}(u^2-u_0^2)$, $w_0=\psi_k(u_0)/u_0$을 보여라.

(c) IRLS의 한 걸음 $\theta^{(1)}=\sum w_i x_i/\sum w_i$ $(w_i=\psi_k(r_i)/r_i,\ r_i=x_i-\theta^{(0)})$이 $L(\theta^{(1)})\le L(\theta^{(0)})$을 만족함을 보여라.

풀이 (a) $t\le k^2$에서 $g(t)=t/2$, $t>k^2$에서 $g(t)=k\sqrt{t}-k^2/2$이다. 도함수는 $\frac{1}{2}$에서 $\frac{k}{2\sqrt{t}}$로 $t=k^2$에서 이어지며 감소하므로 g는 오목하다. (b) 오목함수는 접선 아래에 있으므로 $g(t)\le g(t_0)+g'(t_0)(t-t_0)$이다. $t=u^2$, $t_0=u_0^2$을 넣고 $g'(u_0^2)=\psi_k(u_0)/(2u_0)=w_0/2$를 확인한다($|u_0|\le k$이면 $\frac{1}{2}$, 아니면 $\frac{k}{2|u_0|}$). (c) $Q(\theta)=\sum_i\big[\rho_k(r_i)+\frac{w_i}{2}\big((x_i-\theta)^2-r_i^2\big)\big]$로 두면 (b)로 모든 θ에서 $L(\theta)\le Q(\theta)$이고 $Q(\theta^{(0)})=L(\theta^{(0)})$이다. Q는 가중 제곱합이므로 최소점이 $\theta^{(1)}=\sum w_i x_i/\sum w_i$이다. 따라서 $L(\theta^{(1)})\le Q(\theta^{(1)})\le Q(\theta^{(0)})=L(\theta^{(0)})$이다. 이것이 MM(위로 누르고 최소화) 알고리즘이고, A8의 한 걸음이 목적함수를 줄인 이유다.

C9 독립이 아니어도 교환가능하면 ★★

정의 1.1, 정리 1.2 · 15분

$Z_i=\eta+\varepsilon_i$ $(i=1,\dots,N)$이고 $\eta\sim\mathcal{N}(0,\tau^2)$, $\varepsilon_i\overset{\text{iid}}{\sim}\mathcal{N}(0,1)$이 서로 독립이다.

(a) Z_i들은 독립이 아님을 보여라($\mathrm{Cov}(Z_1,Z_2)$).

(b) $(Z_1,\dots,Z_N)$이 교환가능함을 보여라.

(c) 처음 m개를 X, 나머지를 Y로 둔 순열검정이 여전히 정확함을 보여라. 정리 1.2의 증명 가운데 어느 단계가 독립을 쓰지 않았는가?

풀이 (a) $\mathrm{Cov}(Z_1,Z_2)=\mathrm{Var}\,\eta=\tau^2>0$. (b) 결합밀도는 $\int\prod_i \phi(z_i-e)\,\frac{1}{\tau}\phi(e/\tau)\,de$이고, 곱 $\prod_i\phi(z_i-e)$는 z_i의 순서를 바꿔도 같으므로 결합밀도가 순열에 불변이다. (c) 정리 1.2의 증명에서 ①만 iid를 썼고, 그것도 교환가능성을 얻는 데만 썼다. ②–④는 교환가능성만 쓰므로 그대로 성립한다. 같은 블록(가족, 학교) 안의 관측처럼 공통 효과로 상관된 자료에서도, 처치가 무작위로 배정되었다면 순열검정은 정확하다.

C10 순위 통계량은 단조변환에 불변이다 ★

정의 1.4 · 10분

h가 순증가함수이면 X_i, Y_j를 $h(X_i)$, $h(Y_j)$로 바꿔도 W와 U가 그대로임을 보여라. 두 표본 t 통계량은 그렇지 않음을 $X=(1,2)$, $Y=(3,10)$과 $h(x)=\log x$로 보여라.

풀이 h가 순증가이므로 모든 쌍의 대소가 보존된다. U는 대소의 개수이므로 그대로이고, 정리 1.5(a)로 W도 그대로다. 원래 자료에서 $\bar{Y}-\bar{X}=5$, 합동분산 $s_p^2=(0.5+24.5)/2=12.5$이므로 $t=5/\sqrt{12.5}=\sqrt{2}\approx 1.41$이다. 로그를 취하면 평균의 차는 $\frac{1}{2}(\log 3+\log 10)-\frac{1}{2}\log 2\approx 1.354$이고 $t\approx 1.95$로 바뀐다. 두 경우 모두 X의 순위가 $1,2$이므로 $W=3$, $U=0$이다. 순위 검정의 결론은 자료를 어떤 단위나 단조변환으로 재든 같다.

D

퀄형

여러 절과 앞 회차의 도구를 잇는 종합 문제다. 머리의 제한 시간을 재고, 노트를 덮고 푼다. 표준정규 분위수와 분포함수 값은 표를 쓴다.

D1 정규 이동 모형에서 두 검정의 국소 검정력 ★★★

명제 2.2, 정리 2.3 · 제한 35분

$X_1,\dots,X_m \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $Y_1,\dots,Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta,1)$이고 $m=n=50$이다. 크기 5%의 한쪽 검정으로 $H_0:\theta=0$ 대 $\theta>0$을 검정한다.

(a) $p(\theta)=P(Y>X)$를 Φ로 쓰고 $\theta=0.5$에서 값을 구하라.

(b) 명제 2.2와 정리 2.3의 증명에 나온 효능 e_W로 Wilcoxon 검정의 $\theta=0.5$에서의 근사 검정력을 구하라($N=100$, $\lambda=\frac{1}{2}$).

(c) 같은 방법으로 t 검정의 근사 검정력을 구하고, 두 검정력의 차이를 ARE $3/\pi$와 연결해 설명하라.

풀이 (a) $Y-X\sim\mathcal{N}(\theta,2)$이므로 $p(\theta)=\Phi(\theta/\sqrt{2})$이고, $\theta=0.5$에서 $\Phi(0.354)\approx 0.638$이다. (b) $e_W=\sqrt{3}\cdot\frac{1}{2\sqrt{\pi}}\approx 0.4886$, $h=0.5\cdot 10=5$이므로 검정력 $\approx 1-\Phi(1.645-2.443)=\Phi(0.798)\approx 0.788$이다. (c) $e_t=\sqrt{\lambda(1-\lambda)}/\sigma=0.5$이므로 $1-\Phi(1.645-2.5)=\Phi(0.855)\approx 0.804$이다. $(e_W/e_t)^2=3/\pi=0.955$이므로 Wilcoxon은 t가 100개로 하는 일을 약 105개로 한다. 같은 100개에서는 검정력이 0.016 낮다. 정규 자료에서 순위로 바꾼 값이 이것이다.

최종답 (a) $\Phi(0.354)\approx 0.638$ (b) ≈ 0.788 (c) ≈ 0.804

D2 대립가설 아래의 Mann–Whitney 분산 ★★★

정리 1.5(c), 2회 정리 2.8 · 제한 40분

$X\sim\text{Exp}(1)$, $Y\sim\text{Exp}(2)$ (평균 $\frac{1}{2}$)이고 모든 관측이 독립이다. $p=P(X>Y)$, $q_1=P(X>Y_1, X>Y_2)$, $q_2=P(X_1>Y, X_2>Y)$로 둔다.

(a) p를 구하라.

(b) q_1을 구하라.

(c) q_2를 구하라.

(d) C2의 방법으로 일반적으로 $\operatorname{Var}U=mn\big[p(1-p)+(n-1)(q_1-p^2)+(m-1)(q_2-p^2)\big]$임을 보이고, $m=n=10$에서 $\operatorname{Var}\big(U/(mn)\big)$을 구하라. 귀무가설 아래의 값 $\frac{N+1}{12mn}$과 비교하라.

풀이 (a) $p = Ee^{-Y} = \int_0^\infty e^{-y} 2e^{-2y} dy = \frac{2}{3}$. (b) $q_1 = E(1-e^{-2X})^2 = 1 - 2Ee^{-2X} + Ee^{-4X} = 1 - \frac{2}{3} + \frac{1}{5} = \frac{8}{15}$. (c) $\min(X_1, X_2) \sim \text{Exp}(2)$이고 $Y \sim \text{Exp}(2)$이므로 $q_2 = \frac{1}{2}$. (d) $U = \sum I_{ij}$의 분산은 대각 항 $mn\,p(1-p)$, 같은 i를 공유하는 $mn(n-1)$개의 공분산 $q_1 - p^2$, 같은 j를 공유하는 $mn(m-1)$개의 공분산 $q_2 - p^2$, 나머지 0의 합이다. $m = n = 10$이면

$$\text{Var}\,\frac{U}{100} = \frac{1}{100}\Big[\frac{2}{9} + 9\cdot\frac{4}{45} + 9\cdot\frac{1}{18}\Big] = \frac{1}{100}\cdot\frac{137}{90} = \frac{137}{9000} \approx 0.0152.$$

귀무가설 아래 값은 $\frac{21}{1200} = 0.0175$이다. 대립가설 아래에서는 분산이 달라지므로, 검정력을 계산할 때 귀무가설의 분산을 그대로 쓰면 안 된다.

최종답 (a) $p = \frac{2}{3}$ (b) $q_1 = \frac{8}{15}$ (c) $q_2 = \frac{1}{2}$ (d) $\frac{137}{9000} \approx 0.0152$

D3 치우친 자료의 부트스트랩 구간 ★★★

정리 5.3, 알고리즘 5.2 · 제한 30분

자료가 $(0, 0, 0, 0, 10)$이다. K를 부트스트랩 표본에 10이 뽑힌 횟수라 하자.

(a) $\bar{X}^* = 2K$이고 $K \sim \text{b}(5, \frac{1}{5})$임을 보이고 $P^*(\bar{X}^* = 0)$을 구하라.

(b) $\text{Var}^*\,\bar{X}^*$를 두 가지로 구하라(이항분포와 정리 5.3(a)).

(c) 95% 백분위 구간의 위 끝(97.5% 분위수)을 구하라. 아래 끝은 무엇인가?

(d) t 구간 $\bar{X} \pm t_{0.975,4}\, s/\sqrt{5}$ ($t_{0.975,4} = 2.776$)와 비교하고, 어느 쪽이 이 자료에 맞는지 논하라.

풀이 (a) 표본 평균은 $10K/5 = 2K$이고, 다섯 번의 독립 복원추출에서 10이 뽑힐 확률은 각각 $\frac{1}{5}$이다. $P^*(K=0) = (0.8)^5 = \frac{1024}{3125} \approx 0.328$. (b) $4\,\text{Var}\,K = 4\cdot 5\cdot 0.2\cdot 0.8 = 3.2$. 정리 5.3(a)로는 $\hat{\sigma}^2 = \frac{1}{5}(4\cdot 4 + 64) = 16$이므로 $16/5 = 3.2$로 같다. (c) 누적확률은 $K=2$까지 0.942, $K=3$까지 0.993이므로 97.5% 분위수는 $K=3$, 곧 $\bar{X}^* = 6$이다. 아래 끝은 $\bar{X}^* = 0$이다(확률 0.328의 원자). (d) $s^2 = \frac{4\cdot 4 + 64}{4} = 20$, $s/\sqrt{5} = 2$이므로 t 구간은 $2 \pm 5.55 = [-3.55,\ 7.55]$이다. 자료가 음수일 수 없다면 t 구간의 아래 절반은 의미가 없고, 백분위 구간 $[0, 6]$은 자료의 범위를 지킨다. 그러나 $n = 5$에서 부트스트랩 분포는 다섯 값의 원자들로 이루어져 구간의 포함확률을 보장하지 못한다. 정리 5.3은 $n \to \infty$의 결과다.

최종답 (a) $\frac{1024}{3125} \approx 0.328$ (b) $\frac{16}{5}$ (c) 6

D4 순열분포의 평균과 분산, 그리고 정규근사 ★★★

정리 1.2, 4회 정리 1.5 · 제한 35분

합친 값 $\mathbf{Z}=(Z_1,\dots,Z_N)$을 고정하고, 그 가운데 m개를 무작위로 골라 X로, 나머지 n개를 Y로 둔다. $\tau^2=\frac{1}{N}\sum(Z_i-\bar Z)^2$이다.

(a) 순열분포에서 $E(\bar X-\bar Y)=0$임을 보여라.

(b) $\mathrm{Var}(\bar X-\bar Y)=\dfrac{N^2\tau^2}{mn(N-1)}$임을 보이고, A5의 자료 $\{1,2,4,5,7,9\}$, $m=n=3$에서 값을 구하라.

(c) 정규근사 $P\big(\mathcal{N}(0,\mathrm{Var})\ge 14/3\big)$을 구하고 A5의 정확한 p-값 0.05와 비교하라.

풀이 (a) X 쪽 합 $S=\sum_{i\in X}Z_i$는 비복원 표본합이므로 $ES=m\bar Z$이다. $\bar X-\bar Y=\frac{S}{m}-\frac{N\bar Z-S}{n}=\frac{N}{n}\big(\frac{S}{m}-\bar Z\big)$이므로 평균은 0이다. (b) $\mathrm{Var}\,S=m\tau^2\frac{N-m}{N-1}=m\tau^2\frac{n}{N-1}$이므로

$$\mathrm{Var}(\bar X-\bar Y)=\frac{N^2}{n^2m^2}\cdot\frac{mn\tau^2}{N-1}=\frac{N^2\tau^2}{mn(N-1)}.$$

자료에서 $\bar Z=\frac{14}{3}$, 편차의 제곱합 $\frac{408}{9}$이므로 $\tau^2=\frac{68}{9}$이고 $\mathrm{Var}=\frac{36\cdot 68/9}{9\cdot 5}=\frac{272}{45}\approx 6.04$이다. (c) $z=\frac{14/3}{\sqrt{6.044}}=1.898$이고 $1-\Phi(1.898)\approx 0.029$이다. 정확한 값 0.05보다 작다. 순열분포는 20개의 원자로 된 이산분포이고 가장 바깥 원자가 확률 1/20을 갖는데, 정규근사는 그것을 연속 꼬리로 펴서 작게 준다. N이 크면 둘은 가까워지고(순열 중심극한정리), 그때의 정규근사가 합동분산 z 검정이다.

최종답 (b) $\frac{272}{45}\approx 6.04$ (c) ≈ 0.029