Ch6. 통계학 (Statistics)

목차

Part 1

확률변수의 성질

기댓값 (Expectation)

- 확률변수 X의 **기댓값**(평균)은 PDF $f_X(x)$를 이용해 계산함

$$E[X] = \mu = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

- **이산형** 확률변수의 경우: $E[X] = \sum_x x \cdot P(X = x)$
- 기댓값은 **선형성**(Linearity)을 가짐
  - $E[aX + bY] = aE[X] + bE[Y]$
- 면접 핵심: 다양한 분포에 대해 기댓값을 직접 유도할 수 있어야 함

분산과 표준편차

- **분산**: 확률변수가 기댓값으로부터 얼마나 퍼져 있는지 측정함

$$\mathrm{Var}(X) = E\left[(X - E[X])^2\right] = E[X^2] - (E[X])^2$$

- 분산은 항상 **비음수** (≥ 0)
- **표준편차**: 분산의 제곱근

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

- 유용한 성질들:
  - $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$
  - $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$

공분산과 상관계수

공분산 (Covariance)

- 두 변수의 **선형 관계 방향과 크기**를 측정함

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$
$$= E[XY] - E[X]E[Y]$$

- 범위: $(-\infty, +\infty)$
- 단위가 X와 Y의 단위에 의존함

상관계수 (Correlation)

- 공분산을 **정규화**하여 단위 무관하게 만듦

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}$$

- 범위: $[-1, +1]$
- $|\rho| = 1$이면 완전 선형관계
- **무상관** $\neq$ **독립** (중요!)

유도 예시: 균등분포 $U(a, b)$

- $f_X(x) = \frac{1}{b-a}$ (구간 $[a, b]$에서)

기댓값 유도:

$$E[X] = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{a+b}{2}$$

분산 유도:

$$E[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a}\, dx = \frac{a^2+ab+b^2}{3}$$

$$\mathrm{Var}(X) = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

- 면접에서 **지수분포, 이항분포** 등에 대해서도 동일한 유도 요구됨

Part 2

큰 수의 법칙 & 중심극한정리

큰 수의 법칙 (LLN)

정의

- i.i.d. 확률변수를 충분히 많이 샘플링하면, **표본평균**은 모평균에 수렴함

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow{n\to\infty} \mu$$

- 카지노가 개별 게임에서 지더라도 장기적으로 이익을 보는 이유
- 동전을 5번 연속 앞면이 나와도, n이 커지면 비율은 0.5로 수렴함

면접 포인트

- CLT와 구분할 것: LLN은 **수렴값**, CLT는 **수렴분포**
- A/B 테스트에서 충분한 샘플 크기가 필요한 이론적 근거

중심극한정리 (CLT)

- **어떤 분포의** 확률변수든, 충분히 많이 반복 샘플링하면 **표본평균의 분포**는 정규분포에 근사함

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- 표준화하면:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- 원래 분포가 이항, 포아송, 지수 등 무엇이든 적용 가능함
- 가설검정의 **이론적 토대**: Z-test를 사용할 수 있는 근거
- **면접 핵심**: "표본 크기가 충분히 크면" → CLT를 떠올려야 함

CLT가 왜 중요한가?

A/B 테스트, 가설검정, 신뢰구간 — 통계학의 거의 모든 실무 도구가 CLT에 기반함. 분포의 형태와 무관하게 정규분포를 가정할 수 있게 해주는 정리임.

Part 3

가설검정 체계

가설검정 3단계 절차

❶ **귀무가설(H_0)과 대립가설(H_1) 수립**— H_0를 기각하거나, 기각에 실패하거나 (채택이 아님!)

❷ **검정통계량 계산 및 p-value 산출**— 귀무가설 하에서 관측값의 확률을 계산함

❸ **유의수준(α)과 비교**— 보통$\alpha = 0.05$, p-value$< \alpha$이면H_0기각

단측검정 vs 양측검정

단측검정 (One-tailed)

- $H_0 : \mu = \mu_0$
- $H_1 : \mu < \mu_0$ 또는 $H_1 : \mu > \mu_0$
- 한 방향의 차이만 검정함
- 예: "새 기능이 전환율을 **높이는지**" 검정

양측검정 (Two-tailed)

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$
- 양쪽 방향의 차이를 모두 검정함
- 예: "전환율이 **달라졌는지**" 검정
- 유의수준이 양쪽으로 나뉨 → $\alpha/2$

A/B 테스트와 가설검정의 연결

- A/B 테스트는 가설검정의 **가장 대표적인 실무 응용**임
- **구조**: 통제군(A)과 처리군(B)에 서로 다른 버전을 노출

요소	A/B 테스트에서의 역할
H_0	두 그룹의 핵심 지표가 동일함
H_1	처리군에서 유의미한 차이가 발생함
검정통계량	두 그룹 평균 차이의 Z-score
유의수준 α	사전에 결정 (보통 0.05)
표본 크기	검정력 분석으로 사전 계산 필요

- 그룹 간 **동질성 확보**가 핵심: 연령, 성별, 위치, 디바이스 등 균형 필요
- Uber Eats 예시: 이메일 캠페인이 전환율을 높이는지 검증

Part 4

검정통계량의 종류

Z-test vs t-test vs Chi-squared

Z-test

- 검정통계량이 **정규분포**를 따른다고 가정
- 모분산(σ^2)이 **알려져 있을 때**
- 표본 크기가 **클 때** (CLT 적용)

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

t-test

- **t-분포** 사용 (정규분포보다 꼬리가 두꺼움)
- 모분산이 **미지**일 때 표본분산 s^2 사용
- 자유도: $n - 1$

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

Chi-squared (χ^2)

- **적합도 검정**, 범주형 변수 독립성 검정
- 관측값과 기댓값 비교

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Z-test vs t-test: 언제 무엇을 쓰는가?

조건	사용할 검정
모분산이 **알려져 있음**	Z-test
모분산이 **미지**이고 표본 크기 작음	t-test
표본 크기 $n > 200$	어느 것이든 (t-분포 $\approx$ 정규분포)
원래 모집단이 정규분포	Z-test 또는 t-test 직접 사용 가능
모집단 비정규 + $n > 30$	CLT 적용 후 Z-test

- t-분포는 자유도가 커질수록 정규분포에 수렴함
- **모비율 검정**: $np_0 > 10$이고 $n(1-p_0) > 10$일 때 Z-test 사용 가능
- 면접에서 "왜 이 검정을 선택했는지" 설명할 수 있어야 함

Part 5

모비율 검정과 A/B 테스팅

모비율의 가설검정

- 모비율 p는 베르누이 확률변수의 합으로 표현 가능함
- CLT에 의해 표본비율은 정규분포에 근사함

가설 설정:

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : p \neq p_0$$

검정통계량:

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

- A/B 테스트의 **핵심 공식**: 통제군과 처리군의 전환율 차이를 검정할 때 사용
- 면접에서 "어떤 검정통계량을 쓸 것인지"와 "왜 유효한지 (CLT)" 반드시 언급해야 함

A/B 테스트의 흔한 함정들

- **그룹 불균형**: 인구통계, 디바이스, 위치 등의 차원에서 균형이 깨지면 결과가 왜곡됨
- **실험 기간 부족**: 주간 계절성이 있는 지표를 2일만 측정하면 위험함
  - 충분한 검정력(power)과 유의수준 확보 후에도 바로 종료하지 말 것
- **다중검정 문제**: 변형(variant)이 많아지면 우연에 의한 유의미한 결과 증가
  - 1,000개 버튼 색상 테스트 → 이론적으로 가능하나, 표본 크기 급증 + 교호작용 문제
- **Novelty Effect**: 새로운 것에 대한 일시적 반응으로 효과가 과대평가됨
- **교호작용**: 동시 실행 중인 다른 실험과의 상호작용 미고려

Part 6

p-value와 신뢰구간

p-value와 신뢰구간의 올바른 해석

p-value

- H_0가 참이라는 가정 하에 관측된 검정통계량 **이상으로** 극단적인 값을 얻을 확률
- $p < \alpha \rightarrow H_0$ 기각
- $p \geq \alpha \rightarrow H_0$를 기각하지 못함 (참이라는 뜻이 아님!)
- **오해 금지**: "H_0가 참일 확률"이 아님

신뢰구간 (CI)

- 동일 실험을 반복하면, $(1-\alpha)\%$ 의 CI가 모수의 참값을 포함함

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- 95% CI에 0이 포함되면 → H_0 기각 불가
- **모비율의 CI**:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Part 7

제1종/제2종 오류와 검정력

제1종 오류와 제2종 오류

	H_0 참	H_0 거짓
H_0 기각	제1종 오류 (α) — 거짓 양성	올바른 결정 (Power = $1-\beta$)
H_0 미기각	올바른 결정 ($1-\alpha$)	제2종 오류 (β) — 거짓 음성

- α **(유의수준)**: 참인 H_0를 기각할 확률 → 보통 0.05
- β: 거짓인 H_0를 기각하지 못할 확률
- **검정력 (Power)** = $1-\beta$: 거짓인 H_0를 올바르게 기각할 확률 → 보통 0.8 이상 목표
- α를 줄이면 β가 증가하는 **트레이드오프** 존재

검정력 분석과 다중검정 보정

검정력 분석 (Power Analysis):

- 원하는 검정력 수준에 필요한 **최소 표본 크기** 결정에 사용함
- 세 가지 요소: 유의수준(α), 효과크기(effect size), 표본크기(n)
- 표본이 클수록 → 검정력이 높아짐

다중검정 보정 (Bonferroni Correction):

- 100개의 가설검정을 $\alpha = 0.05$로 수행하면, 우연에 의해 **5개**가 유의미해짐
- **보정**: $\alpha_{\text{new}} = \alpha / m$ (m: 검정 횟수)
  - 예: $0.05/100 = 0.0005$
- 제1종 오류를 잘 통제하지만, **제2종 오류가 증가**할 수 있음 (보수적)
- A/B 테스트에서 여러 지표를 동시에 검정할 때 필수 고려사항임

Part 8

MLE와 MAP

최대우도추정 (MLE)

- 관측 데이터가 가장 **그럴듯한(likely)** 모수 θ를 찾음

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \; L(\theta) = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i \mid \theta)$$

- 로그를 취하면 곱셈이 덧셈으로 변환됨 (계산 편의):

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

- 미분하여 0으로 놓고 θ 를 풀면 됨
- 예: 지수분포의 MLE → $\hat{\lambda} = \frac{n}{\sum x_i}$

MAP: 사전분포를 포함한 추정

MAP (Maximum A Posteriori)

- MLE에 **사전분포(prior)** $g(\theta)$를 추가한 것

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \; g(\theta) \cdot \prod_{i=1}^{n} f(x_i \mid \theta)$$

- 로그 형태:

$$\sum_{i=1}^{n} \log f(x_i \mid \theta) + \log g(\theta)$$

MLE vs MAP 비교

	MLE	MAP
사전분포	사용하지 않음	사용함
관점	빈도론적	베이지안
특수 관계	균등 사전분포의 MAP = MLE	—
데이터 많을 때	MLE ≈ MAP	—

Part 9

면접 문제 패턴별 정리

40문항을 패턴으로 분류하기

패턴	문항 번호	핵심 기법
CLT 응용	6.1, 6.15, 6.19, 6.30	CLT로 정규근사 → 검정/추정
가설검정 설계	6.5, 6.6, 6.10, 6.11, 6.15, 6.19	H_0/H_1 설정 + 검정통계량 선택
A/B 테스트	6.3, 6.7, 6.8, 6.10	오류, 검정력, 다중검정
기댓값 계산 (조건부/재귀)	6.12, 6.13, 6.14, 6.20, 6.21, 6.22, 6.24, 6.40	조건부 기댓값 + 기하분포
분산/공분산 조작	6.4, 6.17, 6.23, 6.37, 6.38	공식 유도 + 독립 vs 무상관
MLE/MAP	6.25, 6.29, 6.34	우도함수 → 로그 → 미분
분포 변환	6.18, 6.26, 6.32, 6.35, 6.39	CDF/PDF 변환, MGF

패턴 A

CLT 응용 문제

CLT 응용: 핵심 접근법

CLT를 언제 쓰는가?

- "표본 크기가 충분히 크면" → CLT 적용
- 이항, 포아송, 베르누이 등 **어떤 분포든** 표본평균이 정규분포를 따름

#6.1 (Uber): CLT 설명 + 실제 활용

- 정의: $\bar{X}_n \sim N(\mu, \sigma^2/n)$ — n이 충분히 클 때
- 실무: Uber 신기능이 예약 수를 늘리는지 검증할 때 각 예약이 베르누이 → CLT로 정규근사

#6.15 (DE Shaw): 동전 1000번 중 550번 앞면 — 편향된 동전인가?

- $H_0 : p = 0.5 \to \mu = np = 500, \sigma = \sqrt{np(1-p)} = \sqrt{250} \approx 15.8$
- $Z = \frac{550-500}{15.8} = 3.16 \to p < 0.001 \to$ **편향 가능성 높음**

CLT 응용: 편향 감지와 정규근사

#6.19 (Morgan Stanley): 60% 앞면 동전 감지에 필요한 횟수

- 95% CI: $\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}$
- $p = 0.6$으로 놓고 CI 하한을 0.5로 설정:

$$0.6 - 1.96\sqrt{\frac{0.6 \times 0.4}{n}} = 0.5$$

- 풀면 $n = 93$회 → 생각보다 적은 횟수로 감지 가능

#6.30 (Uber): 베르누이 생성기로 정규분포 시뮬레이션

- n개 베르누이 시행의 합 → CLT에 의해 정규근사
- 표준화: $x = \frac{\sum x_i - np}{\sqrt{np(1-p)}}$
- $np > 10, n(1-p) > 10$ 조건 필요 (최소 $n = 20, p = 0.5$)

패턴 B

가설검정 설계 문제

가설검정 설계: 핵심 접근법

#6.5 (Facebook): 10번 던져 1번 앞면 — 공정한 동전?

- $H_0 : p = 0.5, H_1 : p < 0.5$ (단측검정)
- 주의: $n = 100$이므로 **CLT 적용 불가** → 정확 이항 검정 사용
- $P(1 \text{ 이하 앞면} \mid p = 0.5) = \frac{C(10,0)+C(10,1)}{2^{10}} = \frac{11}{1024} \approx 0.0107$
- $p\text{-value} = 0.0107 < 0.05 \rightarrow H_0$ 기각

#6.11 (Google): 동전 던지기에서 앞면 비율의 신뢰구간 유도

- n이 충분히 크면 CLT 적용: $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$
- 95% CI: $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

검정 유형 선택과 다중검정

#6.9 (Facebook): Z-test vs t-test 선택 기준

상황	선택
모분산 알려짐	Z-test
모분산 미지 + 소표본	t-test
대표본 ($n > 200$)	Z-test (t-분포 ≈ 정규분포)
모비율 검정 ($np_0 > 10$)	Z-test

#6.10 (Amazon): 수백 개 가설의 동시 검정

- $\alpha = 0.05$로 100개 검정 → 우연히 **5개** 유의미
- Bonferroni: $\alpha_{\text{new}} = 0.05/100 = 0.0005$
- 제1종 오류는 잘 통제되나, 보수적이어서 제2종 오류 증가 가능

패턴 C

A/B 테스트 관련 문제

A/B 테스트 면접 체크리스트

#6.3 함정 (Pitfalls)

- 그룹 불균형
- 실험 기간 부족
- 다중검정 미보정
- 결과 귀인 어려움

#6.7 Type I/II 오류

- Type I (α): 거짓 양성
- Type II (β): 거짓 음성
- 트레이드오프 존재
- α와 β 동시에 줄이기 불가

#6.8 검정력 (Power)

- $1 - \beta$: 진짜 효과를 감지할 확률
- 최소 0.8 이상 권장
- 표본 크기 ↑ → 검정력 ↑
- 효과크기 + α로 사전 계산

#6.6 비전문가 설명

- 가설검정: 두 그룹 비교 실험
- p-value: "우연일 확률" (단순화)
- CI: "참값이 이 범위에 있을 가능성"
- α: 허용 오류 한계

패턴 D

기댓값 계산 문제 (조건부/재귀)

조건부 기댓값: 재귀적 접근법

핵심 기법: 첫 번째 결과에 조건부로 분해

#6.12 (Two Sigma): 연속 앞면 2번까지 필요한 횟수

- $E[X]$를 첫 번째 결과에 조건부로 분해:
  - $E[X] = \frac{1}{2}(1 + E[X|H]) + \frac{1}{2}(1 + E[X|T])$
- $E[X|T] = E[X]$ (꼬리면 처음부터 다시)
- $E[X|H]$를 다시 분해: $E[X|H] = \frac{1}{2}(1 + 0) + \frac{1}{2}(1 + E[X]) = 1 + \frac{1}{2}E[X]$
- 대입 후 정리: $E[X] = 6$

면접 팁: 이 "재귀적 조건부 기댓값" 패턴은 매우 자주 출제됨

기하분포 기반 기댓값 계산

#6.13 (Citadel): 주사위 6면 모두 보기까지 기대 횟수

- k개의 면을 본 후, 새로운 면이 나올 확률: $p = \frac{6-k}{6}$
- 기하분포의 기댓값: $E = 1/p$

$$E[X] = \frac{6}{6} + \frac{6}{5} + \frac{6}{4} + \frac{6}{3} + \frac{6}{2} + \frac{6}{1} = 14.7$$

#6.14 (Akuna Capital): 연속 5가 2번 나올 때까지 기대 횟수

- #6.12와 동일한 재귀 패턴, 성공 확률 $1/6$
- $E[X|Y] = \frac{1}{6}(1) + \frac{5}{6}(1 + E[X])$
- 풀면: $E[X] = 42$

지시변수를 활용한 기댓값

#6.20 (Uber): n개 중 복원추출 n번 → 서로 다른 값의 기대 수

- $X_i = 1$ (값 i가 추출됨)인 지시변수 정의
- $P(X_i = 0) = \left(\frac{n-1}{n}\right)^n$ → $P(X_i = 1) = 1 - \left(\frac{n-1}{n}\right)^n$
- 기댓값의 선형성: $E\left[\sum X_i\right] = n\left[1 - \left(\frac{n-1}{n}\right)^n\right]$
- n이 크면 $\left(\frac{n-1}{n}\right)^n \approx 1/e$ → 약 $n(1 - 1/e) \approx 0.632n$

#6.24 (Citadel): 첫 번째 에이스까지 기대 카드 수

- 52장 중 에이스 4장이 5개 구간을 나눔 → 비에이스 48장이 균등 배분
- 첫 에이스 전 카드: $48/5 = 9.6$ → 에이스 포함 **10.6장**

복잡한 기댓값: 국수와 주사위

#6.21 (Goldman Sachs): 100개 국수 고리의 기대 수

- 양 끝 $2n$개에서 2개를 선택, 같은 국수일 확률: $\frac{n}{\binom{2n}{2}} = \frac{1}{2n-1}$
- 재귀: $E[X_n] = \frac{1}{2n-1}(1 + E[X_{n-1}]) + \frac{2n-2}{2n-1}E[X_{n-1}]$
- 패턴: $E[X_{100}] = 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{199} \approx 3.3$

#6.22 (Morgan Stanley): 두 주사위 최대값의 기댓값

- $Y = \max(X_1, X_2)$, 분할표로 계산:

$$E[Y] = \frac{1}{36}\sum_{i=1}^{6} i \cdot (2i-1) = \frac{161}{36} \approx 4.47$$

- 핵심: $P(Y = i) = P(Y \le i) - P(Y \le i-1) = (i/6)^2 - ((i-1)/6)^2$

패턴 E

분산/공분산 조작 문제

분산 조작과 독립 vs 무상관

분산의 선형결합 (#6.17)

- $\mathrm{Var}(aX + bY)$를 구하라

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)$$

- 독립이면 $\mathrm{Cov} = 0$
- 그렇지 않으면 공분산에 따라 범위가 달라짐

독립 vs 무상관 (#6.37)

- **독립**: $P(X, Y) = P(X)P(Y)$ — 모든 x, y에 대해
- **무상관**: $\mathrm{Cov}(X, Y) = 0$
- **독립 → 무상관** (항상 참)
- **무상관 → 독립** (거짓!)
- 반례: $X \in \{-1, 0, 1\}$ 균등, $Y = \mathbf{1}_{X=0}$

공분산 유도: X와 X^2

#6.38 (Citadel): $X \sim U(-1, 1), Y = X^2$의 공분산

$$\text{Cov}(X, X^2) = E[X^3] - E[X] \cdot E[X^2]$$

- $E[X] = 0$ (대칭 분포)
- $E[X^3] = \int_{-1}^{1} x^3 \cdot \frac{1}{2} dx = 0$ (기함수)
- 따라서 $\text{Cov}(X, X^2) = 0 - 0 = 0$

핵심 교훈: 공분산이 0이어도 **의존관계는 존재**할 수 있음

→ $Y = X^2$이므로 Y는 X에 완전히 결정되지만, 공분산은 0

패턴 F

MLE/MAP 유도 문제

MLE 유도 예시

#6.25 (Spotify): $U(a, b)$의 MLE

- 우도함수: $L(a,b) = \prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n}$
- $(b-a)$를 최소화하면 L이 최대화됨
- 따라서: $\hat{a} = \min(x_1, \ldots, x_n), \hat{b} = \max(x_1, \ldots, x_n)$

#6.34 (Tesla): 지수분포의 MLE

- $f(x|\lambda) = \lambda e^{-\lambda x}$ → 로그우도:

$$\log L(\lambda) = n \log \lambda - \lambda \sum x_i$$

- $\frac{d}{d\lambda} = \frac{n}{\lambda} - \sum x_i = 0$
- $\hat{\lambda} = \frac{n}{\sum x_i}$ (표본평균의 역수)

MLE vs MAP: 핵심 비교 (#6.29)

로그 형태 비교:

$$\text{MLE}: \quad \max_{\theta} \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

$$\text{MAP}: \quad \max_{\theta} \left[\sum_{i=1}^{n} \log f(x_i \mid \theta) + \log g(\theta) \right]$$

- 차이는 오직 **사전분포 항** $\log g(\theta)$ 하나임
- $g(\theta)$가 균등분포(무정보 사전분포)이면 → MLE = MAP
- 데이터가 많아지면 우도 항이 사전분포를 압도함 → MLE ≈ MAP
- MAP는 **정규화(regularization)** 역할을 함
  - 가우시안 사전분포 → L2 정규화와 동일
  - 라플라스 사전분포 → L1 정규화와 동일

패턴 G

분포 변환과 고급 문제

CDF 변환과 MGF

#6.32 (Goldman Sachs): $Y = F(X)$의 분포

- F는 X의 CDF → $Y = F(X)$의 분포는?
- $F_Y(y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y$
- 따라서 $Y \sim U(0,1)$ — CDF 변환의 핵심 결과

#6.35 (Citadel): $\log X \sim N(0,1)$일 때 $E[X]$

- $Y = \log X \sim N(0,1)$이므로 $X = e^Y$
- MGF 활용: $M_Y(s) = E[e^{sY}] = e^{s^2/2}$ (표준정규의 MGF)
- $E[X] = E[e^Y] = M_Y(1) = e^{1/2} = \sqrt{e}$

고급 기댓값과 샘플링

#6.26 (Google): 단조증가 수열의 기대 길이

- $U(0,1)$에서 i.i.d. 추출, 단조증가인 동안 계속 추출
- $P(X_i = 1) = 1/i!$ (i개 원소의 순서가 증가 수열일 확률)
- $E[\text{길이}] = \sum_{i=1}^{\infty} \frac{1}{i!} = e - 1 \approx 1.718$

#6.40 (Two Sigma): 합이 1을 넘을 때까지의 기대 추출 횟수

- $m(t) = E[N_t]$ → 재귀: $m(t) = 1 + \int_0^t m(t-x)dx$
- 미분하면: $m'(t) = m(t), m(0) = 1$
- 해: $m(t) = e^t$ → $m(1) = e \approx 2.718$

원에서 균일 샘플링 (#6.39)

잘못된 방법: 반지름 r을 $U(0, R)$에서 추출 → 중심에 점이 밀집됨

올바른 방법 (역변환법):

1. 반지름 r에서 점의 수는 원둘레 $2\pi r$에 비례 → $f(r) = \frac{2r}{R^2}$
2. CDF: $F(r) = \frac{r^2}{R^2}$ → 역변환: $r = R\sqrt{y}, y \sim U(0, 1)$
3. 각도: $\theta \sim U(0, 2\pi)$
4. 좌표 변환: $x = r\cos\theta, y = r\sin\theta$

핵심: 면적 요소 $dA = r\,dr\,d\theta$이므로 단순 균등 추출이 아닌 **제곱근 보정** 필요

추정량의 성질 (#6.28)

비편향(Unbiased): $E[\hat{\theta}] = \theta$

일치(Consistent): $n \to \infty$일 때 $\hat{\theta} \to \theta$

예시	비편향	일치
첫 번째 표본 X_1	O (평균 추정량)	X (표본 수에 무관)
$S^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$	X ($\frac{n-1}{n}\sigma^2$로 편향)	O ($n \to \infty$에서 편향 소멸)
$s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$	O (Bessel 보정)	O

- **Bessel 보정**: 분모를 $n-1$로 → $E[s^2] = \sigma^2$ (비편향)
- 면접 팁: "비편향이지만 일치적이지 않은 추정량" 예시를 즉시 제시할 수 있어야 함

주사위 게임 비교 (#6.27)

Game 1: 주사위 2개 → 곱 = $X_1 \cdot X_2$ → $E = E[X_1] \times E[X_2] = 3.5^2 = 12.25$

Game 2: 주사위 1개 → 제곱 = X^2 → $E[X^2] = ?$

핵심 통찰 (분산 활용):

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 \geq 0$$

$$\therefore \; E[X^2] \geq (E[X])^2$$

- $E[X^2] = \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) = \frac{91}{6} \approx 15.17$
- Game 2가 **항상** Game 1보다 기댓값이 높음
- 차이 = $\mathrm{Var}(X) = 15.17 - 12.25 = 2.92$

블렌딩된 평균과 표준편차 (#6.36)

두 하위 그룹의 평균/표준편차를 합치는 방법:

결합 평균:

$$\bar{\mu} = \frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2}$$

결합 표준편차 (Bessel 보정 포함):

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + n_1(\mu_1 - \bar{\mu})^2 + n_2(\mu_2 - \bar{\mu})^2}{n_1 + n_2 - 1}}$$

K개 하위그룹으로 확장:

$$\bar{\mu} = \frac{\sum n_i\mu_i}{\sum n_i}, \quad s = \sqrt{\frac{\sum(n_i - 1)s_i^2 + \sum n_i(\mu_i - \bar{\mu})^2}{\sum n_i - 1}}$$

기하분포 기댓값 유도 (#6.31)

기하분포 PMF: $f(k) = (1-p)^{k-1}p$

$$E[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\sum_{k=1}^{\infty} k(1-p)^{k-1}$$

급수 전개:

$$\sum_{k=1}^{\infty} k(1-p)^{k-1} = \sum_{k=1}^{\infty}(1-p)^{k-1} + (1-p)\sum_{k=1}^{\infty}(1-p)^{k-1} + \cdots$$

$$= \frac{1}{p} + \frac{1-p}{p} + \frac{(1-p)^2}{p} + \cdots = \frac{1}{p} \cdot \frac{1}{1-(1-p)} = \frac{1}{p^2}$$

$$\therefore E[X] = p \cdot \frac{1}{p^2} = \frac{1}{p}$$

MGF 유도: 정규분포 (#6.33)

MGF 정의: $M_X(s) = E[e^{sX}]$

표준정규 $N(0,1)$의 MGF:

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}(x^2 - 2sx)} dx$$

완전제곱식 완성: $x^2 - 2sx = (x-s)^2 - s^2$

$$M_X(s) = e^{s^2/2} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}(x-s)^2} dx}_{=1 \text{ (정규분포 PDF 적분)}} = e^{s^2/2}$$

일반 정규분포 $X = \sigma Y + \mu$:

$$M_X(s) = e^{\mu s + \sigma^2 s^2/2}$$

면접 문제 난이도별 분포

난이도	문항 수	출제 기업 예시	핵심 주제
Easy (9)	6.1~6.9	Uber, Facebook, Twitter, Lyft	CLT 설명, CI/p-value, Type I/II, Z vs t
Medium (16)	6.10~6.25	Google, Two Sigma, Citadel, DE Shaw	조건부 기댓값, 분산 조작, MLE, CI 유도
Hard (15)	6.26~6.40	Google, Facebook, Goldman Sachs, Tesla	MGF, CDF 변환, 추정량 성질, 분포 시뮬레이션

- Easy: **개념 설명** 중심 — 비전문가에게 설명할 수 있는가?
- Medium: **계산 + 유도** — 공식을 적용하고 변형할 수 있는가?
- Hard: **수학적 깊이** — 증명, 변환, 구성(construction)이 가능한가?

통계 면접 4대 전략

1. 분포 식별

- 문제에서 어떤 분포가 관련되는지 파악
- 이항, 기하, 정규, 균등, 지수
- CLT 적용 가능 여부 판단

2. 도구 선택

- Z-test vs t-test vs χ^2
- MLE vs MAP
- 정확검정 vs 근사검정
- 표본 크기에 따라 결정

3. 공식 유도

- 기댓값: 정의에서 출발
- 분산: $E[X^2] - (E[X])^2$
- 조건부 기댓값으로 분해
- 지시변수 + 선형성 활용

4. 실무 연결

- A/B 테스트로 연결
- 검정력, 표본크기 계산
- 다중검정 보정
- 비즈니스 의사결정 프레임

핵심 공식 요약

$$E[X] = \int x f(x) dx \,/\, \mathrm{Var}(X) = E[X^2] -$$

$$(E[X])^2 \,/\, Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

$$\mathrm{CI} : \bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \,/\, \hat{\theta}_{\mathrm{MLE}} =$$

$$\arg\max \sum \log f(x_i|\theta)$$

$$\text{Bonferroni: } \alpha_{\mathrm{new}} = \alpha/m \,/\, \text{Power} = 1 - \beta$$