Chapter 5: Probability

목차

Part 1

기초 개념 (Basics)

조건부 확률 (Conditional Probability)

- **정의**: 사건 B가 발생했을 때 사건 A가 발생할 확률
- 공식: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- **독립 사건**: $P(A|B) = P(A)$ 이면 A와 B는 독립
  - B를 알아도 A의 확률에 영향 없음
- **조건부 독립**: 사건 C가 주어졌을 때 A, B가 독립
  - $P(A \cap B|C) = P(A|C) \cdot P(B|C)$
- 인터뷰 힌트: **"~가 주어졌을 때(given that)"** 표현이 나오면 조건부 확률 적용

베이즈 정리 (Bayes' Rule)

- **핵심 공식:**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- **용어 정리:**
  - $P(A)$: **사전확률(Prior)** — 사전 지식
  - $P(B|A)$: **우도(Likelihood)** — 데이터가 가설을 지지하는 정도
  - $P(A|B)$: **사후확률(Posterior)** — 업데이트된 믿음
- **면접 단골:** 희귀 질병 진단, 스팸 필터, 불량품 검출
- ML에서 핵심: 데이터가 주어졌을 때 최적의 조건부 분포를 찾는 것

베이즈 정리 — 희귀 질병 예제 (Q5.6)

- **문제 (Amazon)**: 유병률 0.1%, 검사 정확도 98%, 오진율 1%. 양성일 때 실제 환자일 확률?

$$P(\text{질병}|\text{양성}) = \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.01 \times 0.999} \approx 8.93\%$$

- **놀라운 결과**: 98% 정확도인데도 실제 확률은 ~9%에 불과
- **교훈**: 기저율(base rate)을 무시하면 큰 오류 발생
  - 유병률이 낮을수록 양성예측도(PPV)도 낮아짐
- **인터뷰 팁**: 직관과 수학이 충돌하는 문제 — 반드시 베이즈로 풀 것

전체 확률의 법칙 (Law of Total Probability)

- **상호배타적 사건** $B_1, B_2, \ldots, B_n$ 이 전체 표본공간을 구성할 때:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$$

- **직관**: 사건 A를 여러 시나리오로 **분해**하여 계산
- **활용 예시**:
  - 고객 세그먼트별 구매 확률 계산
  - "결과의 트리(tree of outcomes)" 문제에서 필수
- **인터뷰 팁**: 여러 경로/시나리오가 있는 문제 → 전체 확률 법칙 + 조건부 분해

순열과 조합 (Counting)

순열 (Permutation)

- **순서가 중요**할 때 사용
- $P(n, k) = \frac{n!}{(n-k)!}$
- 예: 비밀번호 생성, 순위 배정
- "5명이 3석에 앉는 경우의 수?"
- $P(5, 3) = 60$

조합 (Combination)

- **순서가 무관**할 때 사용
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- 예: 레스토랑 3곳 선택, 카드 뽑기
- "52장에서 5장 뽑는 경우?"
- $\binom{52}{5} = 2{,}598{,}960$

Part 2

확률변수 (Random Variables)

확률변수의 기초

- **확률변수**: 확률분포가 부여된 수량
- 이산(Discrete): 셀 수 있는 범위 → **확률질량함수**(PMF) $f_X(x)$
- 연속(Continuous): 셀 수 없는 범위 → **확률밀도함수**(PDF) $f_X(x)$
- **필수 조건** (두 경우 모두):
    - 비음수: $f_X(x) \geq 0$
    - 이산: $\sum_x f_X(x) = 1$
    - 연속: $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
- 연속의 경우 특정 값 x의 확률은 0 — 구간의 "밀도"만 측정 가능

누적분포함수 (CDF)

- **정의**: $F_X(x) = P(X \le x)$
- **이산**: $F_X(x) = \sum_{k \le x} P(X = k)$ — PMF의 누적합
- **연속**: $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$ — PDF의 적분
- **특성**:
  - 비음수이며 단조 증가
  - $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to \infty} F_X(x) = 1$
- **인터뷰 팁**: 확률변수를 평가할 때 PDF와 CDF를 모두 파악해야 함
  - "X가 특정 구간에 들 확률?" → CDF 활용

결합/주변/조건부 분포

- **결합분포 (Joint)**: 두 확률변수 X, Y를 동시에 분석
  - $\iint f_{X,Y}(x,y)\,dx\,dy = 1$
- **주변분포 (Marginal)**: 한 변수를 적분하여 제거
  - $f_X(x) = \int f_{X,Y}(x,y)\,dy$
- **조건부분포 (Conditional)**: 한 변수가 주어졌을 때 다른 변수의 분포
  - $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ — 베이즈 정리의 확장
- **인터뷰 팁**: 2개 이상의 확률변수 문제 → 결합분포로 사고할 것
  - 매우 기술적인 라운드에서 출제됨

Part 3

확률분포 (Probability Distributions)

이항분포 (Binomial Distribution)

정의 및 공식

- n번 독립 시행, 성공 확률 p일 때 성공 횟수 k의 확률
- PMF: $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$
- **평균**: $\mu = np$
- **분산**: $\sigma^2 = np(1-p)$

언제 사용하는가?

- 동전 던지기 (n번 중 앞면 횟수)
- 사용자 가입률 (n명 중 가입자 수)
- **이진 결과**의 성공 횟수 계산
- 각 시행이 **독립**이고 성공/실패만 있을 때
- 핵심: 고정된 횟수 + 이진 결과 + 독립

포아송분포 (Poisson Distribution)

정의 및 공식

- **고정된 구간** 내에서 사건 발생 횟수
- 사건 발생률(rate): λ
- PMF: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- **평균**: $\mu = \lambda$
- **분산**: $\sigma^2 = \lambda$ (평균 = 분산!)

언제 사용하는가?

- 웹사이트 시간당 방문 수
- 직물 1㎡당 결함 수
- 콜센터 분당 전화 수
- **연속 구간**에서 발생률 λ로 사건이 발생할 때
- 핵심: 연속 구간 + 발생률 + 횟수

이항분포 vs 포아송분포 — 구분 기준

기준	이항분포	포아송분포
시행 횟수	고정 (n번)	구간으로 정의
결과	성공/실패 (이진)	발생 횟수 (0, 1, 2, ...)
파라미터	n, p	λ (발생률)
평균	np	λ
분산	$np(1-p)$	λ
예시	동전 10번 중 앞면 수	1시간 동안 전화 수
핵심 질문	"n번 중 몇 번 성공?"	"구간 내 몇 번 발생?"

> n이 크고 p가 작으면 이항 $\approx$ 포아송 ($\lambda = np$)

Part 4

연속 확률분포

연속 분포 비교 개요

균일분포 (Uniform)

- $f(x) = \frac{1}{b-a}, a \le x \le b$
- $\mu = \frac{a+b}{2}$
- $\sigma^2 = \frac{(b-a)^2}{12}$
- 랜덤 샘플링, 난수 생성

지수분포 (Exponential)

- $f(x) = \lambda e^{-\lambda x}, x \ge 0$
- $\mu = \frac{1}{\lambda}$
- $\sigma^2 = \frac{1}{\lambda^2}$
- 대기시간, 고장 간격

정규분포 (Normal)

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- 평균 μ, 분산 σ^2
- CLT 기반, 자연 현상

균일분포 (Uniform Distribution)

- **정의**: 구간 $[a, b]$에서 모든 값이 동일한 확률을 가짐
- PDF: $f(x) = \frac{1}{b-a}$ (구간 내), 0 (구간 외)
- CDF: $F(x) = \frac{x-a}{b-a}$
- **평균**: $\mu = \frac{a+b}{2}$, **분산**: $\sigma^2 = \frac{(b-a)^2}{12}$
- **활용 예시**:
    - 난수 생성 (Random Number Generation)
    - 가설 검정 (p-value 계산)
    - 시뮬레이션의 기초 분포

지수분포 (Exponential Distribution)

- **정의**: 포아송 과정에서 사건 간 **대기 시간**의 분포
- PDF: $f(x) = \lambda e^{-\lambda x}, x \geq 0$
- CDF: $F(x) = 1 - e^{-\lambda x}$
- **평균**: $\mu = \frac{1}{\lambda}$, **분산**: $\sigma^2 = \frac{1}{\lambda^2}$
- **핵심 특성 — 무기억성(Memorylessness)**:
  - $P(X > s + t \mid X > s) = P(X > t)$
  - 이미 s만큼 기다렸어도 추가 대기 확률은 변하지 않음
- **활용**: 고객 구매까지 대기시간, 신용 부도 시점, 장비 고장 간격
- **인터뷰 팁**: 무기억성 문제가 자주 출제됨

정규분포 (Normal Distribution)

- PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- 평균 μ, 분산 σ^2으로 완전히 결정됨
- **중요한 이유 2가지:**
  - 자연 현상과의 적합성 (키, 몸무게, 시험 점수 등)
  - **중심극한정리(CLT)**: 표본 크기가 크면 표본 평균은 정규분포에 수렴
- **68-95-99.7 규칙**: $\mu \pm 1\sigma$ (68%), $\mu \pm 2\sigma$ (95%), $\mu \pm 3\sigma$ (99.7%)
- 표준정규분포: $Z = \frac{X-\mu}{\sigma}$ $(\mu = 0, \sigma = 1)$
- **면접 필수**: PDF를 기억하고 CLT와 연결하여 설명할 수 있어야 함

확률분포 선택 가이드

상황	분포	핵심 파라미터
n번 시행 중 성공 횟수	**이항**	n, p
구간 내 사건 발생 횟수	**포아송**	λ
구간 내 균등 확률	**균일**	a, b
사건 간 대기 시간	**지수**	λ
종모양 곡선, CLT 적용	**정규**	μ, σ^2

> 분포를 외우는 것보다 **"이 상황에 어떤 분포가 맞는가?"** 를 판단하는 능력이 핵심

Part 5

마르코프 체인 (Markov Chains)

마르코프 체인 기본 개념

- **정의**: 유한 상태 집합에서 다음 상태가 **현재 상태에만** 의존하는 확률 과정
- **마르코프 성질**: 현재 상태가 주어지면 과거와 미래는 조건부 독립
  - $P(X_{t+1}|X_t, X_{t-1}, \ldots) = P(X_{t+1}|X_t)$
- **전이행렬(Transition Matrix) P**:

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mn} \end{pmatrix}$$

- p_{ij}: 상태 i에서 상태 j로 전이할 확률
- 각 행의 합 = 1

마르코프 체인 — 상태 분류와 정상분포

- **상태 분류:**
  - **재귀 상태(Recurrent):** 진입하면 반드시 다시 돌아옴
  - **일시 상태(Transient):** 떠나면 영원히 돌아오지 않을 확률 > 0
- **정상분포(Stationary Distribution) π:**
  - $\pi = \pi P$ 를 만족하는 확률 벡터
  - 전이를 반복해도 분포가 변하지 않음 — **장기 안정 상태**
- **활용 예시:**
  - 사용자 상태 모델링 (신규 / 활성 / 이탈)
  - 웹페이지 방문 패턴 (PageRank)
- **인터뷰 팁:** 여러 상태 간 전이 + 장기 행동 문제 → 마르코프 체인

Part 6

인터뷰 문제 패턴 및 풀이

인터뷰 문제 접근 4단계

1. 분류

문제 유형 파악: 조건부 확률? 분포? 조합? 마르코프?

2. 모델링

확률변수, 사건, 분포 정의. "given that" → 베이즈

3. 계산

공식 적용. 대칭성/재귀 구조 활용. 복잡하면 분해

4. 검증

극단값 체크, 확률 합=1 확인, 직관과 대조

Easy: 조합 + 기초 확률 패턴

Q5.1 (Google): 7전 4선승제에서 7경기까지 가는 확률?

- 6경기 후 3:3이어야 함 → $P = \binom{6}{3} \cdot \left(\frac{1}{2}\right)^6 = \frac{20}{64} = \frac{5}{16}$

Q5.3 (Uber): 주사위 3번 — 순증가 순서 확률?

- 세 수가 모두 다를 확률: $1 \times \frac{5}{6} \times \frac{4}{6} = \frac{5}{9}$
- 그중 순증가는 $\frac{1}{3!} = \frac{1}{6}$ → 답: $\frac{5}{9} \times \frac{1}{6} = \frac{5}{54}$

Q5.5 (JP Morgan): (0,0,0)→(3,3,3) 경로 수?

- 9보 중 각 방향 3보씩 → $\frac{9!}{3!3!3!} = 1680$

Easy: 베이즈 정리 활용 패턴

Q5.7 (Facebook): 공정/불공정 동전, 5번 뒤면 → 불공정 동전일 확률?

- $P(U|5T) = \frac{P(5T|U) \cdot P(U)}{P(5T|U) \cdot P(U) + P(5T|F) \cdot P(F)}$
- $= \frac{1 \times 0.5}{1 \times 0.5 + (1/32) \times 0.5} \approx 97\%$

Q5.9 (Microsoft): 3명 각각 1/3 거짓말, 모두 "비 온다" → 실제 비 올 확률?

- 세 명 모두 진실할 확률: $(2/3)^3$, 모두 거짓말: $(1/3)^3$
- 베이즈 적용 → $P(R|YYY) = \frac{8}{11}$

> **패턴**: 사전정보 + 새 관측 → 사후확률 업데이트 = 베이즈

Easy: 재귀적 사고 패턴

Q5.8 (Goldman Sachs): A, B 교대 동전, 앞면이면 승리. A가 이길 확률?

- A가 앞면(확률 p) → A 승리
- A, B 모두 뒷면(확률 $(1-p)^2$) → 원점 회귀 → $P(A)$로 재귀
- $P(A) = p + (1-p)^2 \cdot P(A)$ → $P(A) = \frac{1}{2-p}$

Q5.16 (Optiver): 테니스 듀스, 1번 선수 승률 60%

- 2점 후 게임 종료 또는 원점 회귀
- $p = 0.6^2 + 2(0.6)(0.4) \cdot p$ → $p \approx 0.692$

> **패턴**: "원점 회귀" 구조 → 재귀 방정식으로 풀기

Easy: 조건부 확률의 함정

Q5.14 (D.E. Shaw): 두 아이 중 하나가 남자. 둘 다 남자일 확률?

- 직관적 답: 1/2 (잘못됨!)
- 전체 경우: BB, BG, GB, GG → 하나가 남자 → BB, BG, GB
- $P(\text{둘 다 남자}|\text{하나가 남자}) = \frac{1}{3}$

Q5.15 (JP Morgan): 8서랍 중 1/2 확률로 편지 넣음. 7개 비었을 때 8번째에 있을 확률?

- 베이즈: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{1}{9}$

> **패턴**: 표본 공간을 정확히 정의해야 함 — 직관에 속지 말 것

Easy: 전략적 기댓값 패턴

Q5.12 (JP Morgan): 주사위 최대 2번, 중간 멈춤 가능. 참가비?

- 1차 기대값: 3.5
- 전략: 1~3 나오면 재시도(기대값 3.5), 4~6이면 유지(기대값 5)
- $E = \frac{1}{2} \times 3.5 + \frac{1}{2} \times 5 = 4.25$
- **참가비 상한: $4.25**

Q5.4 (Zenefits): 100장 카드에서 2장, 한 장이 다른 장의 정확히 2배일 확률?

- 전체: $\binom{100}{2} = 4950$, 조건 충족 쌍: (1,2), (2,4), ..., (50,100) = 50개
- $P = \frac{50}{4950} \approx 0.01$

Medium: 공정성 생성과 기하학

Q5.20 (Airbnb): 불공정 동전으로 공정한 결과 만들기?

- 한 번으로는 불가능
- 두 번 던지기: HT와 TH는 확률이 동일 ($p(1-p) = (1-p)p$)
- HH, TT는 무시하고 재시도
- **HT = 앞면, TH = 뒷면**으로 정의 → 공정한 결과

Q5.10 (Bloomberg): 원에서 두 현(chord)이 교차할 확률?

- 원 위 4점 선택 → 현 2개 구성 방법: 3가지
- 교차하는 배치: 1가지 → $P = \frac{1}{3}$

Medium: 수열 경쟁과 마르코프

Q5.23 (Lyft): H, T 무작위 수열에서 HHT vs HTT — 먼저 나올 확률?

- 첫 H 이후: 다음이 H(확률 1/2) → HHT가 반드시 먼저
- 다음이 T(확률 1/4 TT → HTT 먼저, 확률 1/4 TH → 원점 회귀)
- $p = \frac{1}{2} + \frac{1}{4}p$ → $p = \frac{2}{3}$

Q5.25 (Facebook): 정삼각형 꼭짓점의 3개미, 랜덤 방향 이동. 충돌 없을 확률?

- 충돌 없으려면 모두 같은 방향 (시계 or 반시계)
- $P = \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}$
- k각형 확장: $P = 2 \times \left(\frac{1}{2}\right)^k = \frac{1}{2^{k-1}}$

Medium: 재귀 관계식 패턴

Q5.26 (Robinhood): 편향 동전 n번, 앞면 총수가 짝수일 확률의 점화식?

- P_n: n번 후 짝수 개 앞면 확률
- 전체 확률 법칙 적용:
    - $P_n = (1-p) \cdot P_{n-1} + p \cdot (1 - P_{n-1})$
- **해석**: 뒷면이면 짝수 유지, 앞면이면 짝수↔️홀수 전환

Q5.27 (Citadel): Alice(승률 p) vs Bob, 2점 차로 승리. Bob 승리 확률?

- 한 라운드 후 점수 차이 0 또는 결정
- $P(B^*) = q^2 + 2pq \cdot P(B^*) + 0$ → $P(B^*) = \frac{q^2}{1-2pq}$ (단, $q = 1 - p$)

Hard: 분포와 고급 조합론

Q5.30 (Spotify): 주사위 n번, 최대값이 정확히 r일 확률?

- 모든 값 $\leq r$: $P(B_r) = (r/6)^n$
- 모든 값 $\leq r-1$: $P(B_{r-1}) = ((r-1)/6)^n$
- $P(\max = r) = \left(\frac{r}{6}\right)^n - \left(\frac{r-1}{6}\right)^n$

Q5.33 (Quora): N개 정규분포 표본 중 k개가 값 Y보다 클 확률?

- 각 표본이 Y를 초과할 확률: $p = 1 - \Phi\left(\frac{Y-\mu}{\sigma}\right)$
- k개가 초과 → **이항분포**: $\binom{N}{k} p^k (1-p)^{N-k}$

Hard: 재귀와 극한의 문제

Q5.31 (Goldman Sachs): 아메바 — 매분 1/4 확률로 사망/유지/2분열/3분열. 멸종 확률?

- p = 멸종 확률, 각 경우의 멸종 확률:
  - 사망: 1, 유지: p, 2분열: p^2, 3분열: p^3
- $p = \frac{1}{4}(1 + p + p^2 + p^3)$ → $p = \sqrt{2} - 1 \approx 0.414$

Q5.32 (Lyft): n번 동전, k개 앞면. 첫 번째가 앞면이었을 확률?

- 조건부 확률: $P(\text{첫 H}|k\text{개 H}) = \frac{k}{n}$
- 직관: n개 위치 중 k개가 H — 대칭성으로 첫 위치가 H일 확률 = k/n

Hard: 기하학적 확률

Q5.22 (Goldman Sachs): 길이 1 막대를 랜덤으로 3등분 → 삼각형 가능 확률?

- 두 절단점이 같은 쪽: 삼각형 불가 (확률 1/2)
- 다른 쪽이라도 한 조각 > 1/2: 삼각형 불가 (확률 1/4)
- $P(\text{삼각형}) = 1 - \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

Q5.34 (Akuna Capital): 단위원 위 3점 → 삼각형이 중심 포함할 확률?

- 두 번째 점 각도 θ $(0 < \theta < \pi)$, 세 번째 점이 조건 만족 확률: $\frac{\theta}{2\pi}$
- 전체: $\frac{1}{\pi} \int_0^\pi \frac{\theta}{2\pi} \, d\theta = \frac{1}{4}$

Hard: 색상 공 문제의 우아한 풀이

Q5.35 (Citadel): 빨간 공 r개 + 흰 공 w개. 한 색이 소진될 때까지 뽑기. 흰 공 먼저 소진될 확률?

- **복잡한 풀이**: $w + r - 1$번째까지 뽑기에서 조합 계산
  - $P = \frac{r}{w+r}$
- **우아한 풀이**: 마지막 남은 공의 색만 보면 됨
  - 마지막 공이 빨간색 = 흰 공이 먼저 소진
  - $P = \frac{r}{w+r}$ — 무작위 추출에서 마지막이 빨간색일 확률과 동일!

> **교훈**: 복잡한 문제도 관점을 바꾸면 한 줄로 풀림

Part 7

핵심 요약

인터뷰 핵심 도구 6가지

1. 베이즈 정리

"given that" → 즉시 베이즈. 사전확률 + 우도로 사후확률 계산

2. 전체 확률 법칙

여러 시나리오로 분해. 고객 세그먼트, 결과 트리 문제

3. 순열/조합

순서 중요 → 순열, 무관 → 조합. 카드/좌석/경로 문제

4. 분포 선택

이진+고정횟수→이항, 발생률→포아송, 대기시간→지수, CLT→정규

5. 재귀적 사고

"원점 회귀" 패턴 발견 → 방정식 세우기. 게임/시리즈 문제

6. 마르코프 체인

다중 상태 전이 + 장기 행동 → 전이행렬 + 정상분포

면접에서 자주 하는 실수 TOP 5

❶ **기저율 무시**: 베이즈 문제에서 사전확률($P(A)$)을 빠뜨리고 우도만으로 판단

❷ **표본공간 오류**: "두 아이 중 하나가 남자" 같은 문제에서 조건부 표본공간 정의 실패

❸ **분포 혼동**: 이항분포와 포아송분포의 적용 상황을 구분하지 못함

❹ **독립성 가정 누락**: 결합확률 계산 시 독립 여부 확인 없이 곱셈 적용

❺ **재귀 구조 미인식**: 반복 게임에서 원점 회귀 패턴을 놓치고 무한 경우를 열거

분포별 면접 빈출 질문 유형

분포	빈출 질문	핵심 키워드
이항	동전/A/B테스트 성공 횟수	"n번 중", "성공률 p"
포아송	시간/공간당 발생 횟수	"시간당", "비율 λ"
지수	다음 사건까지 대기시간	"얼마나 기다려야", "무기억성"
정규	표본 평균, 신뢰구간	"CLT", "분포 수렴"
기하	첫 성공까지 시행 횟수	"처음으로 성공할 때까지"
마르코프	상태 전이, 장기 비율	"상태", "전이", "장기적으로"

확률 면접의 황금률

"모든 확률 문제는 **표본공간 정의 → 사건 분해 → 공식 적용 → 검증**의 4단계로 풀 수 있다. 직관에 속지 말고 구조적으로 접근하라."