44.006 — University of Seoul, Spring 2026

Lecture 1: Stochastic Convergence

March, 2026

Lecturer: *Jong-June Jeon* *Scribe*: *Jong-June Jeon*

1 대수의 법칙과 중심 극한 정리

독자 여러분은 독립적인 동전 던지기 실험을 언젠가 들어 본 적이 있을 것이다. 앞면이 나올 확률이 p인 동전을 반복해서 던지는 실험이다. 그때 각 시행이 서로 영향을 받지 않는 독립적인 실험을 무한히 반복하는 과정을 생각해보자. 자연스럽게 동전의 앞면이 나올 상대빈도가 p 로 수렴한다는 것을 여러분은 자연스럽게 받아들일 것이다. 우리는 이 수렴과정을 좀 더 엄밀하게 분석하기 위해서 확률변수 도입하여 수식으로 표현한다.

$X_i \in \{0, 1\}$ 은 i번째 시행에서 동전이 앞 면이 나왔을 때 1, 뒷면이 나왔을 때 0를 가지는 확률변수라 하자. n번째까지 실험을 반복 했을 때 동전이 앞 면이 나온 수 의 총합은 $S_n = \sum_{i=1}^{n} X_i$ 일 것이고, 앞면이 나온 횟수의 비, 상대빈도는 S_n/n 이다. S_n/n은 확률변수 $\{X_i\}_{i=1}^{n}$의 함수이므로 여전히 랜덤한 값을 가지는 확률변수다. 상대빈도의 정의 상 이 값이 0에서 1까지의 값을 가진 다는 것을 쉽게 알 수 있다. 앞서 언급한 상대 빈도가 S_n/n 으로 표시 되고, 대수의 법칙에 의해

$$S_n/n \to p, \text{ as } n \to \infty \tag{1}$$

라 표기해도 쉽게 받아 들일 것이다. 그런데 여기서 생각해 봐야 할 일이 있다.

식 (1)의 좌항은 랜덤한 값이고 우항은 랜덤하지 않은 값이다. 극한이라 하면 n이 커지면서 좌항이 우항으로 무한히 가까이 가야 함을 의미하는데, 랜덤한 값이 랜덤하지 않은 상수로 가까이 간다함은 어떤 의미일까? 실험 할 때마다 다르게 나올 수 있는 이 불확실성을 가진 값이 고정된 실수 p에 가까이 간다는 것을 정의하기 위해서 확률적 수렴(in probability convergence)의 개념이 등장한다.

확률적 수렴에서는 어떤 사건의 열들을 먼저 생각한후 그 사건들의 각 확률이 어떻게 되는지를 생각한다. 예를들어 n번째 시행에서 상대 빈도가 p보다 $\epsilon > 0$보다 가까운 사건 A_n을 생각하고, 이 사건이 일어날 확률을 각각 $P(A_n)$ 이라 표시하는 것이다. 이 확률값들은 확실히 실수다. 확률적 수렴은 이 실수값이 1로 가까이 가는 것을 의미한다. 여기서는 우리가 상대 빈도가 p보다 $\epsilon > 0$만큼 떨어져 있는 사건 B_n을 정의했으므로 확률값들이 0으로 가까이간다고 함으로써 동전의 상대빈도가 p로 수렴함을 확률적 수렴으로써 정의하게 된다. 중요한 점은 먼저 사건의 열들을 생각하고 다음 각 사건에 대응 되는 확률 값들 즉 실수 열을 생각 한 후 해당 실수 열의 값이 0로 간다는 것으로 정의하였고 이는 확률값의

극한으로 랜덤한 값의 수렴을 이용했다는 것이다.

$$P(A_n) = P(|S_n/n - 1/2| > \epsilon) \to 0,$$

$n \to \infty$.
한편 우리는 동전던지기 실험에서중심 극한 정리에 대한 이야기도 알고 있다. 역사적으로 중심극한정리의 출발점이 된 de Moivre-Laplace의 정리부터 살펴보는 것이 자연스럽다. de Moivre의 정리는 동전을 여러 번 던질 때, 앞면이 나온 횟수의 분포가 시행 횟수가 증가함에 따라 점점 종 모양의 곡선으로 가까워진다는 사실을 설명한다. 여기서 분포가 가까워진다는 말은 패턴이 가까워진다고 흔히 설명하기도 한다. 그러면 패턴이 가까워진다고 의미를 좀더 자세히 살표보자.
de Moivre의 정리는 동전을 던지는 독립적인 실험을 같은조건에서 반복하는 독립 이항 시행에서 S_n을 평균 $\mu = np$ 중심으로 적절히 이동시키고 분산 $\sigma^2 = npq$의 크기에 맞게 스케일을 조정한 확률변수 $\frac{S_n - np}{\sqrt{npq}}$ 의 분포에 대한 수렴성을 보인다. 이 확률변수에 대한 누적 분포 함수 F_n이라 표시하자. x위에서 F_n의 함수값은

$$F_n(x) = P(\frac{S_n - np}{\sqrt{npq}} \leq x)$$

로 주어진다. 여기서 분포에 대한 수렴이란 이 누적분포함수 F_n이 어떤 누적 분포함수 F로 점별 수렴을 의미한다. 후에 조금 더 엄밀하게 논의하겠지만 대충 아래와 같은 수렴을 의미한다.

$$F_n(x) \to F(x) \text{ as} n \to \infty \text{ for all } x$$

특별히 de Moivre의 정리에서 F_n이 수렴하는 F가 표준정규분포의 누적분포함수에 해당한다. 흥미로운 점은 위에서 정의한 확률변수 $\frac{S_n - np}{\sqrt{npq}}$ 이 대수의 법칙 에서 살펴본 예와 같이 어떤 상수항으로 수렴 하는 것이 아니라 분포의 수렴은 '어떤 패턴을 가진 확률변수 수렴' 한다는 것이다. 우리는 분포의 수렴이 실수의 수렴과 다른 개념을 가지고 있으며 사실상 누적 분포 함수열이 어떤 누적 분포 함수로의 수렴으로 정의 됨을 확인 할 수 있다. 즉, 분포수렴은 우리가 익숙하게 받아들여 온 실수값의 수렴과는 구별되는 개념으로 패턴들이 어떤 고정된 패턴으로 수렴을 의미한다. 앞으로는 패턴이라는 용어대신 확률분포 혹은 분포라는 용어를 사용하겠다.
중심극한정리(Central Limit Theorem)를 분포 수렴의 관점에서 살펴보자. 분산이 유한한 동일, 독립인 확률변수 열 $\{X_i\}$를 생각하자. 확률변수 X_i의 평균이 μ, 분산이 σ^2일 때 확률변수 $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ 의 누적분포 함수를 F_n이라 하겠다.

$$F_n(x) = P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq x\right) = P\left(\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \leq x\right)$$

이 때 중심극한정리의 결과로 다음 사실을 알 수 있다.

- F_n 은 확률변수 $\sqrt{n}\left(\frac{\bar{X}_n-\mu}{\sigma}\right)$의 누적분포함수다.
- 누적분포함수 F가 존재하여 $F_n \to F$ as $n \to \infty$ 가 성립한다.
- X_i의 패턴(분포)와 상관없이 F는 항상 표준정규분포의 누적분포함수다.

확실히, 세 번째 사실은 흥미롭다. X_i의 분포에 따라 F_n은 바뀔 것이다. 예를 들면 X_i 가 0 또는 1을 가지는 베르누이 분포일때, X_i가 0을 포함한 양의 정수 값을 가지는 포아송 분포 일때, X_i가 0에서 1사이의 실수값을 가지는 균등분포 이렇게 세가지 경우만 생각해보자. 이 때 각 경우에 따라 F_n은 다르게 주어질 것인데, 수렴하는 F는 그것에 관계없이 항상 같다는 것은 신기하다. 이런 특징으로 중심극한정리는 기술의 발명이 아니라 자연법칙의 발견이라고 한다. 중심 극한 정리는 데이터의 미지(unknown)의 분포를 사용하지 않고도 모평균에 대한 추론을 가능 하게 해 준다.
여기서 분포에 대한 수렴을 조금 더 자세히 들여다 보자. $M_n = \frac{S_n-n\mu}{\sqrt{n}\sigma}$ 라 놓자. F_n은 확률변수 M_n의 누적분포함수다. 앞서 분포의 수렴은 누적 분포 함수의 수렴으로 정의하였다. 누적분포함수는 다음 집합에 대한 확률 값을 대응 시키는 함수며 함수 값은 다음과 같이 정해진다.

$$F_n(x) = P(M_n \in (-\infty, x])$$

그리고 표준정규분포를 따르는 확률변수 Z를 도입하자. 집합 $(-\infty, x]$에 대한 확률변수 M_n의 확률값 즉 분포를 P_n으로 표시하자. $P_n((-\infty, x])$의 값은 $P(M_n \in (-\infty, x])$로 정해진다. 여기서 집합 B에 대한 (실수위의 보렐집합) $P_n(B)$의 값은 $P(M_n \in B)$로 정해지나 이 값은 F_n 누적분포함수로 구할 수 있음이 알려져 있다. 다시말해 F_n을 아는 것과 P_n을 아는 것은 같다. 한편 표준정규분포를 따르는 확률변수 Z를 도입하자. Z의 분포를 P_Z라고 두겠다. 집합 B에 대해 $P_Z(B)$의 값은 $P(Z \in B)$로 정해진다. 이 분포 $P_Z(B)$ 역시 표준 정규분포의 누적 분포 분수로 부터 완벽히 알 수 있다. 그래서 분포 수렴은 누적분포 함수의 수렴으로 처음에 정의 했지만 일반적인 경우 확률변수의 분포 함 수렴으로 나타낸다. 그래서 분포 수렴은 사실상 확률 측도의 수렴이며

$$P_n(B) \to P_Z(B) \text{ for } B \text{ as } n \to \infty$$

로 정의할 수 있다. 더 확실히 해야할 부분은 어떤 집합 B에 대해서 위해서 언급한 수렴이 성립 해야 하는 것 하는 것이다. 과연 모든집합에 대해서 위 수렴이 성립 해야 하는 것일까? 이 부분은 분포으로의 정의를 압력 하게 다루는 과정에서 더 자세히 살펴보도록 하겠다
이렇게 문포의 수량은 다음과 같이

$$P_n \to P_Z \text{ as } n \to \infty$$

로 표시 한다. 그리고 P_n은 확률변수 M_n의 분포, P_Z는 확률변수 Z의 분포이므로 다음과 같이 표기

하기도 한다.

$$M_n \to_d Z \text{ as } n \to \infty$$

여기서 작은 결론을 내 보자. 우리는 대 수의 법칙과중심 극한 정리를 이용해서 두가지 다른 확률적 수렴을 살펴 보았다. 대수의 법칙을 통해 얻어진 홍 일적 수렴은 랜덤한 확률변수 가 상수로 수렴하는 상황을 나타내고 있다. 반면 중심 극한 성 리는 랜덤한 확률변수가 여전히 랜덤한 개체로 수렴 즉, 분포 수렴을 보여 주고 있다.
표기상에서는 $S_n/n \to p, \text{ as } n \to \infty$ 그리고 $M_n \to_d Z \text{ as } n \to \infty$ 로 서로 유사해 보이지만, 수렴성을 정의하는 방법은 상이하다.
- P와 P_Z의 차이가 궁금한 사람은 질문!

2 분포 수렴

먼저 확률 변수의 분포 수렴의 정의를 살펴 보자.

Definition 2.1 (Convergence in distribution). Let $X_1, X_2, \ldots$ be a sequence of random variables with cumulative distribution functions F_n, and let X be a random variable with cumulative distribution function F.
We say that X_n converges in distribution to X, and write

$$X_n \xrightarrow{d} X,$$

if

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \text{for all } x \text{ at which } F \text{ is continuous.}$$

정의에서 흥미로운 점은 바로 'x at which F is continuous'라는 것이다. 왜 모든 x가 아니라 수렴한 분포에서 연속적인 점, 근방에서 확률이 부드럽게 변하는 점을 대상으로로 제약조 건을 주었을까 하는 점이다. 예를 살펴보면서 그 이후에 대해서 알아 보도록 하자.

Example 2.2 (점프가 있는 극한분포). 극한 확률변수 X를 다음과 같이 정의하자.

$$\mathbb{P}(X = 0) = \frac{1}{2}, \qquad X \sim \frac{1}{2} \cdot \text{Unif}(1, 2).$$

즉, X의 누적분포함수 F는 $x = 0$에서 크기 $\frac{1}{2}$의 점프를 갖는다. 확률변수열 $\{X_n\}$을 다음과 같이 정의한다. 다음 분포수렴의 정의를 살펴보겠다.

- n이 홀수일 때,

$$X_n = \begin{cases} \frac{1}{n}, & \textit{w.p. } \frac{1}{2}, \\ U(1,2), & \textit{w.p. } \frac{1}{2}, \end{cases}$$

- n이 짝수일 때,

$$X_n = \begin{cases} -\frac{1}{n}, & \textit{w.p. } \frac{1}{2}, \\ U(1,2), & \textit{w.p. } \frac{1}{2}. \end{cases}$$

그러면 임의의 $x \neq 0$에 대해

$$\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x),$$

즉 $F_n(x) \to F(x)$가 성립한다. 따라서

$$X_n \xrightarrow{d} X.$$

그러나 점프가 존재하는 점 $x = 0$에서는

$$F_n(0) = \begin{cases} 0, & n \text{ 홀수}, \\ \frac{1}{2}, & n \text{ 짝수}, \end{cases}$$

로서 $\{F_n(0)\}$는 수렴하지 않는다.

이 예는 분포수렴에서 누적분포함수의 점별 수렴을 모든 점이 아니라 극한 분포 F가 연속인 점에서만 요구해야 함을 보여준다. 왜냐하면, 극한 분포에서 확률값의 연속성이 없는 부분에서는 F_n을 통한 확률의 수렴을 보장하는 것은 매우 강한 조건이며 분포의 수렴에서는 이러한 점에서 확률값의 수렴을 제외시켰다. 이것이 바로 분포 수렴의 조건에서 'x at which F is continuous'로 사건집합을 제한한 이유다.

Definition 2.3 (분포수렴). 확률변수열 $\{X_n\}$이 확률변수 X로 분포수렴(*convergence in distribution*)한다고 함은, X_n과 X가 유도하는 확률측도를 각각

$$P_n(A) = \mathbb{P}(X_n \in A), \qquad P(A) = \mathbb{P}(X \in A)$$

라고 할 때, 임의의 보렐 집합 $A \subset \mathbb{R}$ 중 $P(\partial A) = 0$ 를 만족하는 모든 A에 대하여 $\lim_{n\to\infty} P_n(A) = P(A)$ 가 성립하는 것을 말한다. 이때 이를 $X_n \xrightarrow{d} X$ 로 쓴다.

여기서 ∂A 는 집합 A의 경계를 의미하며 $P(\partial A) = 0$ 를 만족하는 모든 A라는 뜻은 극한 분포에서 점프가 없는 집합만을 고려 하겠다는 뜻이다. 분포의 수렴은 확률측도 열 $\{P_n\}$의 수렴이라고도 말하며, 극한확률을 단순히 P로 표기 하기도 한다.

Remark 2.4. 확률변수 열 $\{X_n\}$이 어떤 상수 $c \in \mathbb{R}$로 확률수렴한다고 하자. 즉,$X_n \xrightarrow{p} c$. 이는 정의에 따라 다음을 의미한다: $\forall \varepsilon > 0, \qquad \mathbb{P}(|X_n - c| > \varepsilon) \to 0 \quad (n \to \infty)$. 한편, 상수 c는 확률변수로 생각할 수 있으며, 그 분포는 한 점에 집중된 확률측도, 즉 Dirac measure δ_c로 주어진다:

$$\delta_c(B) = \begin{cases} 1, & c \in B, \\ 0, & c \notin B, \end{cases} \qquad B \in \mathcal{B}(\mathbb{R}).$$

X_n의 분포를 P_{X_n}이라 하면,

$$X_n \xrightarrow{d} c \quad \Longleftrightarrow \quad P_{X_n} \Rightarrow \delta_c.$$

즉, 확률변수 열이 상수로 확률수렴한다는 것은 그 분포가 점 c에 집중된 하나의 점질량(point mass)으로 수렴한다는 것과 같다.

3 Portmanteau Lemma

분포의 수렴은 먼저 사건 집합을 고정하고 각 분포의 확률값(실수)가 극한분포의 확률값으로 수렴하는지 확인해야 한다. 이러한 수렴성은 분포 수렴의 정의에 명시된 모든 집합에 대해서 확인 해야 하며 이 과정을 기술하는 것이 편리하지 않을 수 있다. Portmanteau Lemma은 이 와 같은 분포수렴을 적분의 수렴으로 표현할 수 있는 방법을 제시 해 주며, 이 과정은 이후에 있을 여러 가지 해석적인 설명을 쉽게 만들어 준다. 즉, Portmanteau Lemma는 분포 수렴을 편리하게 기술 할 수 있는 도구를 제공 하는 것이다.

Theorem 3.1 (Portmanteau Lemma)**.** 확률변수열 $\{X_n\}$과 확률변수 X가 유도하는 확률측도를 각각 P_n과 P라 하자. 다음 조건들은 서로 동치이다.

1. 임의의 보렐 집합 $A \subset \mathbb{R}$ 중 $P(\partial A) = 0$ 를 만족하는 모든 A에 대하여 $\lim_{n\to\infty} P_n(A) = P(A)$.
2. 모든 유계이고 연속인 함수 $f : \mathbb{R} \to \mathbb{R}$에 대하여 $\lim_{n\to\infty} \int f \, dP_n = \int f \, dP$.
3. 임의의 닫힌 집합 $F \subset \mathbb{R}$에 대하여 $\limsup_{n\to\infty} P_n(F) \leq P(F)$.
4. 임의의 열린 집합 $G \subset \mathbb{R}$에 대하여 $\liminf_{n\to\infty} P_n(G) \geq P(G)$.

Portmanteau Lemma 는 분포수렴을 적분의 수렴으로 문제를 재표현 할 수 있는 도구를 제공한다. 적분의 수렴은 해석학에서 다양한 성질이 밝혀져 있어 그것들을 활용하기 쉬우며, 분포 수렴에서 다양한 사건 집합을 고려 하는 대신 유계 연속함수를 생각하는 것이 분석에서 더 편리할 때가 많다.
한편 3, 4 부등식에서 각각 닫힌 집합과 열린 집합을 고려 해야 되는 이유를 살펴보자. 확률측도열 $\{P_n\}$ 을 $P_n := \delta_{1/n},\ P := \delta_0$ 로 정의하자. 그러면 자명하게 $P_n \Rightarrow P$ 가 성립한다. 그러나 닫힌

집합 $F = \{0\}$ 에 대해 $P_n(F) = 0 \quad (\forall n)$, 이므로 $\liminf_{n\to\infty} P_n(F) = 0 < 1 = P(F)$. 따라서 $\liminf_{n\to\infty} P_n(F) \geq P(F) \quad (\forall F \text{ closed})$ 는 약수렴이 성립하더라도 일반적으로 성립하지 않는다. 다음으로 역방향을 보자. 왜 $\liminf_{n\to\infty} P_n(F) \geq P(F)$ (모든 닫힌집합 F) 만으로는 약수렴을 보장하지 않는가. 약수렴(분포수렴) $P_n \Rightarrow P$ 에서 집합확률의 수렴을 논할 때 핵심적인 장애물은 경계 (*boundary*) 에 존재하는 확률질량이다. 실제로 임의의 Borel 집합 A 에 대해 일반적으로

$$P(A^\circ) \;\leq\; P(A) \;\leq\; P(\overline{A}), \qquad P(\overline{A}) = P(A^\circ) + P(\partial A)$$

가 성립한다. 특히

$$P(\partial A) > 0$$

이면 A의 내부와 폐포(혹은 A 자체) 사이에서 확률이 불연속적으로 점프한다.

이제 닫힌집합 F에 대해 다음과 같은 상황을 생각하자:

$$P(\partial F) > 0, \qquad P_n(\partial F) = 0 \;\; (\forall n).$$

즉, 극한 측도 P는 F의 경계 ∂F에 양의 확률질량을 가지지만, 각 P_n은 경계에 확률을 전혀 두지 않는다고 하자. 이때 F는 닫힌집합이므로 $F = \overline{F}$ 이고 $F = F^\circ \cup \partial F$ (서로소 합) 이므로

$$P(F) = P(F^\circ) + P(\partial F).$$

반면 $P_n(\partial F) = 0$ 이면

$$P_n(F) = P_n(F^\circ \cup \partial F) = P_n(F^\circ) + P_n(\partial F) = P_n(F^\circ)$$

가 되어, 각 n에서 F의 확률과 F°의 확률이 동일하다. 즉, P_n의 입장에서는 F와 F° 사이의 "경계에서의 점프"가 전혀 관측되지 않는다.

만약 어떤 방식으로든 $P_n(F)$의 값들이 $P_n(F^\circ)$ 와 같아 극한이 $P(F^\circ)$ 수준에 머문다면 극한 측도 P가 경계에 갖는 양의 질량 $P(\partial F)$ 만큼의 증가분은 P_n에서 끝까지 반영되지 않는다. 그 결과 $P(F) > P(F^\circ)$ 이지만,

$$P_n(F) = P_n(F^\circ) \;\not\to\; P(F)$$

가 될 수 있다. 다시 말해 $\liminf_{n\to\infty} P_n(F) \geq P(F)$ 와 같은 부등식만으로는 경계에서의 확률점프를 통제할 수 없으므로, 약수렴 $P_n \Rightarrow P$를 결론낼 수 없다. Lemma의 증명은 textbook 을 참조하길 바란다.

4 점근적 크기 비교: o, O, o_p, O_p

이 절에서는 실수 수열에서의 점근적 표기 o, O와 확률변수 열에서의 확률적 점근 표기 o_p, O_p를 정의하고, 이러한 조건들이 실제로 만족되는지를 확인하는 방법을 설명한다. 이후 Prohorov 정리를 통해 분포수렴과 O_p의 관계를 설명하고, Slutsky 정리를 이용하여 확률변수의 대수적 연산 하에서 이러한 성질들이 어떻게 유지되는지를 살펴본다.

4.1 실수 수열에서의 o, O

$\{a_n\}$, $\{b_n\}$을 실수 수열이라 하자. 두 수열의 크기를 비교하는 것은 많은 경우 극한에서 그 값의 비를 비교하는 경우가 많다. 특히 각 수열의 극한이 동시에 0으로 가거나 혹은 무한대로 발산하는 경우 어떤 수열이 더 빨리 0으로 수렴하거나 혹은 더 빠르게 무한대로 발산하는지 관심이 있는 경우가 있다. 예를 들면 $S_n = \sum_{i=1}^{n} a_i$ 가 수렴하는 값인지를 확인하고 싶을때 $\{a_n\}$이 $1/n$보다 빠르게 0으로 수렴하는지 혹은 그렇지 않은지를 판정해야만 한다. 우리는 두 수열비를 계산하고 그것에 극한을 조사한다. $a_n/(1/n)$ 값이 0으로 간다는 뜻은 a_n이 $1/n$보다 빨리 0으로 수렴한다는 뜻이며 이를 $a_n = o(1/n)$ as $n \to \infty$로 표시한다. 이러한 표기는 0으로 수렴 하는 수혈 뿐만 아니라 발산 하는 수열의 속도를 표기하는 데도 사용된다. $a_n = o(n)$이라는 것은 a_n/n 이 0으로 간다는 뜻이고 a_n의 발산 속도가 n보다 느림을 의미한다. 한편 O의 표기는 수렴의 같거나 빠를때 또는 발산 속도가 같거나 느릴때를 나타내기 위해 사용한다. 실제로 수렴속도가 같음을 나타낼때는 Θ 표기를 사용하거나 $\asymp$기호를 사용한다($a_n = \Theta(b_n)$ 혹은 $a_n \asymp b_n$).

Definition 4.1 ($o(\cdot)$). $a_n = o(b_n)$ as $n \to \infty$라 함은 $\frac{a_n}{b_n} \to 0$ as $n \to \infty$ 임을 의미한다.

Definition 4.2 ($O(\cdot)$). $a_n = O(b_n)$ as $n \to \infty$ 라 함은 어떤 상수 $C \geq 0$와 충분히 큰 n에 대해 $|a_n| \leq C b_n$ 가 성립함을 의미한다.

a_n이 0으로 수렴하는 수렴의 경우로 한정해서 말하면 다음과 같은 little o, big O 에 대한 설명이 가능할 것이다.

- $a_n = o(b_n)$: a_n은 극한에서 b_n에 비해 무시할 수 있을 정도로 작다.
- $a_n = O(b_n)$: a_n은 b_n과 같은 차수(order)를 가지거나 혹은 무시할 수 있을 정도로 작다.

두 개 이상의 실수열이 주어진 경우에 우리는 각 실수열의 원소들이 어떤 연산을 통해서 새로 정의된 경우 그곳에 점근적 크기를 비교 할 수 있다. $c \in \mathbb{R} \setminus \{0\}$ 에 대해

$$a_n = O(b_n) \Rightarrow ca_n = O(b_n), \qquad a_n = o(b_n) \Rightarrow ca_n = o(b_n).$$

Addition

$$a_n = O(b_n),\ c_n = O(b_n) \Rightarrow a_n + c_n = O(b_n),$$
$$a_n = O(b_n),\ c_n = o(b_n) \Rightarrow a_n + c_n = O(b_n),$$
$$a_n = o(b_n),\ c_n = o(b_n) \Rightarrow a_n + c_n = o(b_n).$$

Multiplication

$$a_n = O(b_n),\ c_n = O(d_n) \Rightarrow a_n c_n = O(b_n d_n),$$
$$a_n = o(b_n),\ c_n = O(1) \Rightarrow a_n c_n = o(b_n),$$
$$a_n = o(b_n),\ c_n = o(d_n) \Rightarrow a_n c_n = o(b_n d_n).$$

Dominating order

$a_n \to 0$, $b_n \to 0$ 인 두 실수열에 대해

$$a_n = o(b_n)$$

이면 b_n 이 a_n 을 점근적으로 지배한다 (asymptotically dominates)고 말한다.
이 경우 b_n 을 *dominating order*라 한다.

지배 관계는 다음 성질들을 갖는다:

- $a_n = o(b_n) \ \Rightarrow\ a_n + b_n \asymp b_n$
- $a_n = o(b_n) \ \Rightarrow\ \max\{|a_n|, |b_n|\} \asymp |b_n|$
- $a_n = o(b_n),\ c_n = O(1) \ \Rightarrow\ c_n a_n = o(b_n)$

따라서 서로 다른 차수의 항들이 결합된 경우, 전체 점근적 크기는 dominating order에 의해 결정된다.

Example 4.3. $a_n = o(1/n^2) + o(\log n/n)$ 로 계산되었으면 *dominating term* 은 큰쪽 $o(\log n/n)$이므로 $a_n = o(\log n/n)$ 다.

4.2 확률변수 열에서의 o_p, O_p

이제 $\{X_n\}$을 확률변수 열이라 하자. 실수열의 수렴 속도를 표현 하는 것처럼 확률변수에도 확률 수렴의 속도를 표현 하는 방법이 있다. 실수열의 수렴과 비교 하였을때 확률변수 열의 수렴을 정의하는 개념적 차이는 확률 변수가 랜덤한 값을 가지기 때문에 해당 수렴 여부를 확률로 표현한다는 것이다. 어떤 양의 실수열 a_n에 대해서 $X_n/a_n \xrightarrow{p} 0$ as $n \to \infty$는 확률수렴은

$$P(|X_n/a_n - 0| \geq \epsilon) \to 0 \text{ as } n \to \infty$$

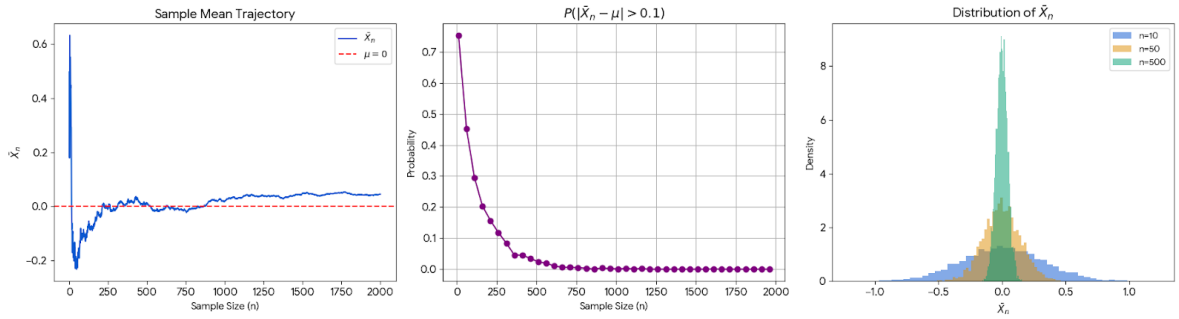

Figure 1: 표본평균이 모형균으로 수렴하는 것을 시각화한 결과. 표본의 크기를 1에서 2000까지 증가시킴.

다. 이는 $P(|X_n - 0| \geq a_n\epsilon) \to 0$ as $n \to \infty$와 같으며 X_n이 0의 $a_n\epsilon$ 근방에 확률이 몰려 있음을 의미한다.

Definition 4.4 ($o_p(\cdot)$). 확률변수 열 X_n이 $X_n = o_p(1)$ 이라 함은 $X_n \xrightarrow{p} 0$ 임을 의미한다. 더 일반적으로,

$$X_n = o_p(a_n) \quad \Longleftrightarrow \quad \frac{X_n}{a_n} \xrightarrow{p} 0.$$

만약 $X_n = \mu + o_p(1/n)$ 이라고 하면 $X_n - \mu = o_p(1/n)$ 이면

$$P(|X_n - \mu| \geq \epsilon/n) \to 0 \text{ as } n \to \infty$$

이고, 이는 X_n이 μ의 $1/n$ 근방에 확률이 몰려 있음을 의미한다.

Example 4.5. 랜덤 표본의 표본평균의 수렴속도는?
이어서 확률변수 가 확률적으로 값을가 질 수 있는 상한의 속도를 표시 하는 방법을 살펴 보겠다. 실수에서 $O(\cdot)$는 실수의 속도를 제한 하는 용도로 사용 되었다. 마찬가지로 확률변수 열에 대해서도 동일한 논리가 적용 된다.

Definition 4.6 ($O_p(\cdot)$). 확률변수 열 X_n이 $X_n = O_p(1)$ 이라 함은 임의의 $\varepsilon > 0$에 대해 어떤 $M > 0$, N이 존재하여 $P(|X_n| > M) < \varepsilon$ for all $n \geq N$ 가 성립함을 의미한다. 일반적으로,$X_n = O_p(a_n) \quad \Longleftrightarrow \quad \frac{X_n}{a_n} = O_p(1)$.

실수열에서 유계(Bounded)가 값이 특정 상수를 넘지 못함을 의미한다면 $O_p(1)$은 확률변수 열 $\{X_n\}$의 확률적 유계를 의미하는 것으로 확률변수가 무한대로 발산하지 않고 특정 범위 내에 머무를 확률이 1에 가깝다는 것을 의미한다. $X_n = O_p(1)$은 다음과 같은 특징을 가진다.

- n이 커지더라도 확률분포의 질량(mass)이 양 끝으로 흩어지지 않고, 특정 유한한 구간 내에 안정적으로 유지된다.
- $X_n = O_p(1)$이라는 선언은 해당 확률변수가 n의 증가에 따라 통제 불능 상태로 커지지 않음을 보장한다.

단일 확률변수는 확률적으로 유계다. 하지만 확률변수 열에서 각 확률변수의 기댓값 유한성($E|X_n| < \infty$)이 $X_n = O_p(1)$를 의미하지는 않는다.

Example 4.7 (반례: 기댓값은 유한하나 유계는 아닌 경우)**.** 확률변수 X_n이 다음과 같다고 가정하자.

$$P(X_n = n) = 1$$

이 경우 $E[X_n] = n$으로 모든 n에 대해 유한하지만, $n \to \infty$일 때 X_n은 무한대로 발산하므로 $O_p(1)$이 아닙니다.

Definition 4.8 (확률분포 열의 Tightness)**.** 확률측도들의 열(sequence of probability measures)을 $\{P_n\}_{n\geq 1}$ 이라 하자. 여기서 각 P_n은 어떤 위상공간 S 위에서 정의된 확률측도이다. 확률측도 열 $\{P_n\}_{n\geq 1}$이 *tight*하다는 것은 임의의 $\varepsilon > 0$에 대하여 콤팩트 집합 $K_\varepsilon \subset S$가 존재하여

$$\inf_n P_n(K) > 1 - \varepsilon$$

을 만족하는 것을 말한다.

참고로 실수값을 가지는 확률변수의 경우에는 두 정의가 동치다. 확률 벡터의 경우에도 노름을 도입하기 때문에 두 정의는 동치다. 차이점은 $O_p(1)$을 생각하기 위해서는 확률 벡터의 크기 라는 개념을 도입해야 한다. 예를 들면 $\{|X_n| \leq M\}$ 과 같이 사건을 생각해야한다. 하지만 확률변수를 명시적으로 표시하지 않고 그것에 대응되는 확률분포 열 $\{P_n\}$만을 이용해서 확률적 유계를 tightness로 정의할 수 있다. (metric 공간에서 bounded 와 totally bounded 의 차이를 확인해보자) 확률분포 열의 tightness 는 위상에 대한 개념만을 요구하기 때문에 좀 더 일반적이라 할 수 있다. 만약 X_n 이 닫힌 집합위에 정의된 함수인 경우에 샘플 공간위의 metric 을 통해 위상[1]을 정의할 수 있으며 그 매트릭 공간에서 compact set을 생각하고 확률분포 열의 tightness을 정의할 수 있다. 특별히 확률분포 열의 tight하기 위해서는 샘플 공간을 덮을 수 있는 셀 수 있는 조밀집합 (countable dense set)의 존재 여부가 중요한 역할을 한다. 다음 예제는 확률변수 열이 확률적으로 유계가 되기 위해서 가져야 하는 충분조건 하나를 소개한다.

Example 4.9. 확률변수열 $\{X_n\}_{n\geq 1}$에 대하여, $\sup_{n\geq 1} \mathbb{E}[|X_n|^p] < \infty$ 라고 하자. 단 $p > 0$. 이때 $X_n = O_p(1)$ 임을 보일 수 있다. 임의의 $M > 0$ 에 대해 *Markov* 부등식을 적용하면 $\mathbb{P}(|X_n| > M) \leq$

[1]열린집합의 모임으로 정의를 찾아보길 바람

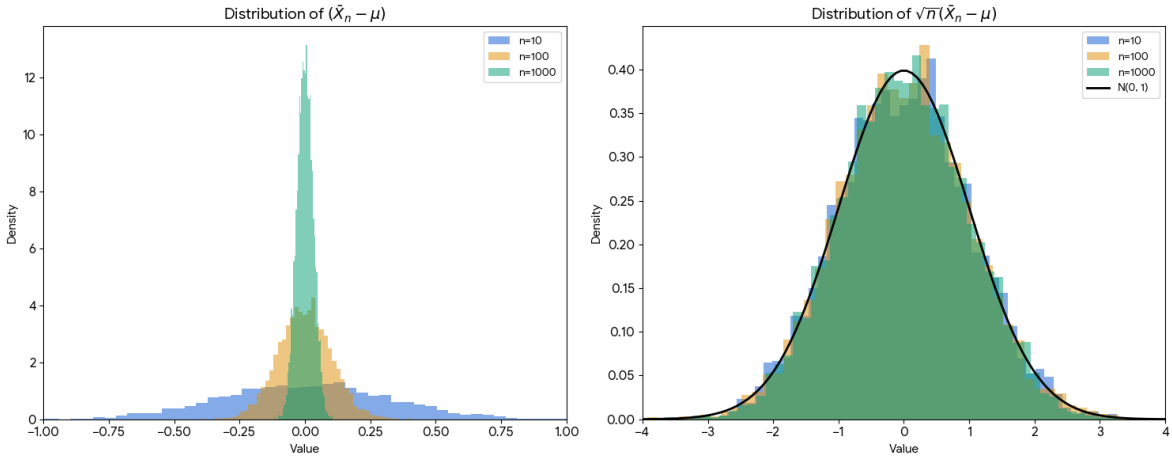

Figure 2: $(\bar{X} - \mu)$ (좌)와 $\sqrt{n}(\bar{X} - \mu)$ (우)의 히스토그램 비교

$\frac{\mathbb{E}[|X_n|^p]}{M^p} \leq \frac{C}{M^p}$. 이제 $\varepsilon > 0$ 를 임의로 주면, $M = (C/\varepsilon)^{1/p}$ 로 택할 때 $\sup_{n\geq 1} \mathbb{P}(|X_n| > M) \leq \frac{C}{M^p} = \varepsilon$. 따라서 정의에 의해 $X_n = O_p(1)$ 이다.

다음으로 o_p, O_p의 대수적 성질 및 지배 원리에 대해서 알아보자. 확률적 근사에서 가장 중요한 것은 복잡한 수열 중 어떤 항이 점근적으로 '지배적(Dominant)'이며, 어떤 항이 '무시 가능한지(Negligible)'를 판단하는 것이다. 표준화된 표본 평균을 생각할 때, 모표준편차 σ 대신 표본표준편차 s_n을 사용하는 경우를 생각하보자.

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \cdot \frac{\sigma}{s_n}$$

여기서 일치성(Consistency)에 의해 $s_n \xrightarrow{p} \sigma$이므로, $\frac{\sigma}{S_n} = 1 + o_p(1)$로 표현할 수 있다. 결과적으로 전체 통계량은 다음과 같이 전개할 수 있음이 알려져 있다.

$$\underbrace{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}}_{O_p(1)} \cdot \underbrace{(1 + o_p(1))}_{1+\text{무시 가능}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} + o_p(1)$$

따라서 s_n이 포함된 식이라 할지라도, 점근적으로는 우리가 잘 아는 표준정규분포를 따르는 항이 지배하게 된다. 이는 점근적 성질을 유도해야 할때 어떤 부분이 지배항지 확인하는 작업이 매우 중요함을 보여준다. 아래는 확률적 수렴항에 대한 연산에 대한 결과를 정리한 것이다.

- 곱셈 규칙 (Product Rule): 서로 다른 차수의 곱은 차수의 곱으로 나타난다.
  - $O_p(a_n)O_p(b_n) = O_p(a_n b_n)$
  - $o_p(a_n)O_p(b_n) = o_p(a_n b_n)$

– $o_p(a_n)o_p(b_n) = o_p(a_n b_n)$

- 흡수 법칙 (Absorption Law): $o_p(a_n)$은 $O_p(a_n)$의 부분집합이다. 즉, 수렴하는 것은 유계인 것에 흡수된다.

$$O_p(a_n) + o_p(a_n) = O_p(a_n)$$

- 지배 원리 (Dominance Principle): 서로 다른 차수의 수열이 합해질 때, 차수가 낮은(더 천천히 발산하거나 더 빨리 수렴하는) 항은 차수가 높은 항에 지배된다.

$$O_p(n^a) + O_p(n^b) = O_p(n^a), \quad \text{if } a > b$$

Example 4.10 (차수 정리에 대한 예시).

1. *Sample Mean*의 분산 추정: 표본 평균 $\bar{X}_n$에 대해 $\bar{X}_n - \mu = O_p(n^{-1/2})$임을 알고 있다. 이때 $(\bar{X}_n - \mu)^2$은 어떻게 되는가?

$$(\bar{X}_n - \mu)^2 = O_p(n^{-1/2}) \cdot O_p(n^{-1/2}) = O_p(n^{-1})$$

   즉, 오차의 제곱은 원래 오차보다 훨씬 빠른 속도로 *0*으로 수렴하므로, *1*차 근사식에서는 $o_p(n^{-1/2})$로 취급되어 사라진다.

2. *Taylor* 전개에서의 활용: 함수 g에 대해 $g(X_n)$을 θ 주변에서 전개하면 다음과 같다:

$$g(X_n) = g(\theta) + g'(\theta)(X_n - \theta) + \frac{1}{2}g''(\tilde{\theta})(X_n - \theta)^2$$

   만약 $\sqrt{n}(X_n - \theta) = O_p(1)$, 즉 $(X_n - \theta) = O_p(n^{-1/2})$라면:

$$g(X_n) = g(\theta) + \underbrace{g'(\theta)(X_n - \theta)}_{O_p(n^{-1/2})} + \underbrace{R_n}_{O_p(n^{-1})}$$

   여기서 $R_n = O_p(n^{-1})$은 $o_p(n^{-1/2})$이므로[2], $\sqrt{n}$을 곱했을 때 $o_p(1)$이 되어 사라진다. 이 지배 원리가 델타 방법의 정당성을 부여한다.

4.3 Prohorov 정리와 O_p의 관계

Theorem 4.11 (Prohorov). 확률변수 열 $\{X_n\}$에 대해 다음은 동치이다.

[2] $R_n/n^{-1/2} = O_p(n^{-1/2})$이고 $O_p(n^{-1/2})$는 $n \to \infty$일 때 $o_p(1)$이다.

1. $\{X_n\}$의 분포가 *tight* 하다.

2. 모든 부분열에 대해 분포수렴하는 부분부분열이 존재한다.

Prohorov 정리가 의미하는 것은 (2)이면 (1)이라는 사실을 다시 살펴보자. 만약 X_n이 분포수렴한다고 가정하면 (2)의 결과는 자명하며 (1)을 함의한다. 예를들어 CLT를 이용하여 $X_n \xrightarrow{d} N(0,1)$을 증명하였다면 Prohorov 정리를 통해 $X_n = O_p(1)$임을 알 수 있다. 즉, X_n이 분포수렴한다는 사실은 $X_n = O_p(1)$임을 의미하며 $X_n = O_p(1)$는 X_n의 분포수렴에 대한 필요조건이라는 것이다. 즉, $O_p(1)$라는 것은 분포열의 분포수렴을 논할 수 있는 최소한의 조건을 제공한다. 이것은 우리가 어떤 추정량 $\hat{\theta}_n$을 얻었을 때, $\sqrt{n}(\hat{\theta}_n - \theta)$의 분포 수렴을 생각하기 위해서는 최소한 $\sqrt{n}(\hat{\theta}_n - \theta) = O_p(1)$이 성립해야함을 의미한다.

Example 4.12. 어떤 확률변수열의 확률적 수렴의 차수 o_p, O_p를 어떻게 보이는가?

$o_p(1)$를 보이는 방법

- *Markov* 부등식, *Chebyshev* 부등식을 이용해 $\mathbf{Pr}(|X_n| > \varepsilon) \to 0$ 을 직접 보인다.
- 평균제곱수렴: $E[X_n^2] \to 0 \;\Rightarrow\; X_n = o_p(1)$.

$O_p(1)$를 보이는 방법

- $\sup_n E|X_n|^p < \infty$ (어떤 $p > 0$) 이면 *Markov* 부등식으로 $O_p(1)$.
- 분포수렴을 이용하는 방법 (*Prohorov* 정리).

4.4 Slutsky 정리와 대수적 연산

두 개의 분포열 $\{X_n\}$, $\{Y_n\}$에 대해서 $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$ as $n \to \infty$를 가정해보자. 우리는 분포수렴의 정의에서 $\{X_n\}$에 대응하는 분포열 $\{P_n\}$과 $\{Q_n\}$이 각각 확률분포 P와 Q로 수렴한다는 것을 알고 있다. 그렇다면 $X_n + Y_n$이라는 확률변수열이 분포수렴한다면 수렴하는 분포는 무엇이될까? 자명하게도 $P+Q$는 아닐 것이다. 왜냐하면 $P+Q$는 확률분포가 아니다. 다음 우리의 예상은 $X+Y$의 확률분포라 생각할 수 있을것인데 일반적으로 $X_n + Y_n \xrightarrow{d} X + Y$ as $n \to \infty$는 성립하지 않는다. (사실상 이 명제가 성립하기 위해서는 분포열에 대한 특별한 가정이 필요하다.)

Theorem 4.13 (Slutsky)**.** 만약

$$X_n \xrightarrow{d} X, \qquad Y_n \xrightarrow{p} c$$

이면,

$$X_n + Y_n \xrightarrow{d} X + c, \quad X_n Y_n \xrightarrow{d} Xc.$$

Slutsky 정리는 다음과 같은 상황에 유용하게 사용된다. $T_n = S_n + R_n$ 이라고 하자. 만약 $\sqrt{n}S_n \xrightarrow{d} Z$ 고 $\sqrt{n}R_n \xrightarrow{p} 0$ 임을 안다면, $\sqrt{n}T_n \xrightarrow{d} Z$ 임을 Slutsky 정리를 통해서 안다.
Slutsky 정리는 점근적 전개(asymptotic expansion)에서

$$\text{leading term} + o_p(\text{scale})$$

와 같은 형태의 표현으로 극한에서의 분포계산을 가능하게 해준다. 이는 M-estimator, asymptotic normality, influence function 분석의 핵심 도구이다. 다음으로 조금더 일반적인 형태의 continuous mapping theorem 을 살펴보자.

Theorem 4.14 (Continuous Mapping Theorem)**.** 확률변수열 X_n과 확률변수 X가 $X_n \xrightarrow{p} X$ 를 만족하고, 함수 $g : \mathbb{R} \to \mathbb{R}$가 X의 값에서 연속이라 하자. 그러면 $g(X_n) \xrightarrow{p} g(X)$ 가 성립한다.

Theorem 4.15 (Continuous Mapping Theorem for Weak Convergence)**.** 확률변수열 X_n이 $X_n \xrightarrow{d} X$ 를 만족하고, 함수 g가 X의 분포가 질량을 두는 점들에서 연속이면 $g(X_n) \xrightarrow{d} g(X)$ 가 성립한다.

Example 4.16. (중심극한정리) 서로 독립이고 동일분포인 확률변수 $X_1, \ldots, X_n$이 $\mathbb{E}[X_i] = \mu, \quad \text{Var}(X_i) = \sigma^2 < \infty$ 를 만족한다고 하자. 표본평균을 $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ 라 하면 중심극한정리에 의해 $\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0,1)$ 가 성립한다. 그러나 실제로는 모분산 σ^2를 알 수 없는 경우 대신 표본분산 $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ 을 이용하여 σ를 추정한다. 대수의 법칙에 의해 $S_n^2 \xrightarrow{p} \sigma^2$ 이므로 $\frac{\sigma}{S_n} \xrightarrow{p} 1$. *Slutsky* 정리를 적용하면,

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{S_n}\right) = \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \cdot \left(\frac{\sigma}{S_n}\right) \xrightarrow{d} \mathcal{N}(0,1) \cdot 1 = \mathcal{N}(0,1).$$

즉 모분산 σ^2를 모르더라도 일관된 분산추정량 S_n^2를 대입하면 정규극한분포는 변하지 않는다. 이는 중심극한정리의 실용적 형태이며, *Studentized statistic*의 근거가 된다.

Example 4.17. (*U-statistic* 분석에 *CLT*와 *Slutsky* 정리의 적용)
$i.i.d.$ 표본 $X_1, \ldots, X_n$에 대해 $\mu = \mathbb{E}[X_1], \quad \sigma^2 = \text{Var}(X_1) < \infty, \quad \mathbb{E}[X_1^4] < \infty$ 라 하자. 분산의 대표적인 *U-statistic* 추정량으로 $U_n := \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \frac{(X_i - X_j)^2}{2}$ 를 정의한다. U_n은 $n(n-1)/2$ 개의 *dependent* 항의 합으로 이루어져있어 독립항의 합에 대한 점근적근사방법인 *CLT*를 직접 적용하기 어렵다. *U-statistic*의 점근적 분석의 일반적인 접근은 *dependent* 항의 합을 점근적인 성질이 같은 독립항과 나머지 항으로 분해하고, 분해한 독립항에 *CLT*를 적용하여 원래 *U-statistic*의 성질을 밝힌다. 이때

$$\sqrt{n}(U_n - \theta) = \sqrt{n}(\textit{CLT term}) + o_p(1)$$

이 되는데 $\sqrt{n}$(*CLT term*) $\xrightarrow{d} N$ 이고 $o_p(1)$는 *0*으로 확률수렴하는 항이라 결국 $\sqrt{n}(U_n - \theta) \xrightarrow{d} N + 0$ 이라는 논리를 적용한다.
좀 더 구체적으로 살펴보면 대칭 커널 $h(x, y) = \frac{(x-y)^2}{2}$ 에 대한 *2*차 *U-statistic*[3]이며

$$\theta = \mathbb{E}[h(X_1, X_2)] = \frac{1}{2}\mathbb{E}[(X_1 - X_2)^2] = \sigma^2$$

이므로 U_n은 σ^2의 불편추정량이다. *Hájek decomposition*에 의해 $U_n - \theta = \frac{2}{n}\sum_{i=1}^{n} \psi(X_i) + R_n$, 임이 알려져 있다. 여기서 $\psi(x) = \mathbb{E}[h(x, X_2)] - \theta, \qquad R_n = o_p(n^{-1/2})$ 이다. 따라서

$$\sqrt{n}(U_n - \theta) = \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) + o_p(1).$$

이제 중심극한정리에 의해

$$\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) \xrightarrow{d} \mathcal{N}\big(0,\, 4\,\mathrm{Var}(\psi(X_1))\big).$$

마지막으로 *Slutsky* 정리에 의해 $\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}\big(0,\, 4\,\mathrm{Var}(\psi(X_1))\big)$. 한편 $\mathrm{Var}(\psi(X_1))$는 계산가능한 항으로 $(E(X_1 - \mu)^4 - \sigma^4)/4$로 알려져 있다.

5 Stochastic Convergence of Random Vector

이 절에서는 $\mathbb{R}^k$-값 확률벡터의 확률적 수렴 개념들을 정리한다. 이러한 개념들은 점근적 전개, 중심극한정리의 확장, 그리고 이후 다룰 Slutsky 정리와 Delta method의 기초를 이룬다.

5.1 확률수렴 (Convergence in Probability)

확률벡터 $X_n, X \in \mathbb{R}^k$에 대하여 X_n이 X로 확률수렴한다고 함은, 임의의 $\varepsilon > 0$에 대해

$$\mathbb{P}\big(\|X_n - X\| > \varepsilon\big) \to 0 \quad \text{as } n \to \infty$$

가 성립하는 것을 말하며, 이를

$$X_n \xrightarrow{p} X$$

로 표기한다. 여기서 $\|\cdot\|$는 $\mathbb{R}^k$ 위의 임의의 노름이며, 유한 차원에서는 노름의 선택과 무관하다.

[3] U-Statistics 의 점근적 분석을 참고하여라

5.2 분포수렴 (Convergence in Distribution)

확률벡터 $X_n, X \in \mathbb{R}^k$에 대해 X_n이 X로 분포수렴한다고 함은, 임의의 bounded continuous 함수 $f : \mathbb{R}^k \to \mathbb{R}$에 대해

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$$

가 성립하는 것을 말하며, 이를

$$X_n \xrightarrow{d} X$$

로 표기한다. 이는 X_n의 분포가 X의 분포로 약수렴함을 의미한다.

확률수렴은 분포수렴을 함의하지만, 그 역은 일반적으로 성립하지 않는다.

5.3 Stochastic Convergence of Random Vectors

벡터값 확률변수 $X_n \in \mathbb{R}^m$와 양수 확률변수 a_n에 대하여 $X_n = o_p(a_n)$ 이라 함은

$$P(\|X_n\| > a_n\epsilon) \xrightarrow{p} 0$$

을 의미한다. 벡터값 확률변수 $X_n \in \mathbb{R}^m$와 $a_n > 0$ a.s. 인 확률변수에 대하여

$$X_n = o_p(a_n) \iff \exists\, Y_n \text{ such that } X_n = a_n Y_n, \quad Y_n \xrightarrow{p} 0.$$

같은 맥락에서 벡터값 확률변수 $X_n \in \mathbb{R}^m$와 $a_n > 0$ a.s. 인 확률변수에 대하여

$$X_n = O_p(a_n) \iff \exists\, Y_n \text{ such that } X_n = a_n Y_n, \qquad Y_n = O_p(1).$$

6 Weak Convergence of Random Vectors

이 절에서는 $\mathbb{R}^k$-값 확률벡터의 분포수렴을 다루기 위한 핵심 도구들을 다룬다. 특히 characteristic function, Cramér–Wold device, 그리고 multivariate central limit theorem을 소개한다.

6.1 Characteristic Function

확률벡터 $X \in \mathbb{R}^k$의 *characteristic function*은

$$\varphi_X(t) = \mathbb{E}\big[e^{it^\top X}\big], \qquad t \in \mathbb{R}^k$$

로 정의된다. Characteristic function은 항상 존재하며, 확률분포를 유일하게 결정한다.

Theorem 6.1 (Lévy Continuity Theorem). 확률벡터열 $X_n, X \in \mathbb{R}^k$에 대하여 $X_n \xrightarrow{d} X$ 일 필요충분조건은

$$\varphi_{X_n}(t) \to \varphi_X(t) \quad \forall t \in \mathbb{R}^k$$

가 성립하는 것이다.

따라서 characteristic function은 다변량 분포수렴을 증명하는 가장 기본적인 도구가 된다.

6.2 Cramér–Wold Device

다변량 분포수렴은 모든 방향으로의 1차원 사영을 통해 판별할 수 있다. Cramér–Wold device는 이를 공식화한 정리다.

Theorem 6.2 (Cramér–Wold Device). 확률벡터열 $X_n, X \in \mathbb{R}^k$에 대하여 다음은 동치이다.

1. $X_n \xrightarrow{d} X$.
2. 모든 $a \in \mathbb{R}^k$에 대해
$$a^\top X_n \xrightarrow{d} a^\top X.$$

이 결과에 대한 직관은 characteristic function 의 수렴을 통해 얻을 수 있다. 즉 다변량 분포수렴은 모든 선형결합의 1차원 분포수렴으로 환원될 수 있다. Cramér–Wold device 의 결과는 Lévy Continuity Theorem의 직접적인 결과라고 볼 수 있는데
X_n과 X에 대한 임의의 방향 $a \in \mathbb{R}^k$ 로 사영을 $Y_n = a^\top X_n/t$ 과 $Y = a^\top X/t$ 이라 놓으면

$$\mathbb{E}\big[e^{i(t^\top X_n)}\big] = \mathbb{E}\big[e^{itY_n}\big] \to \mathbb{E}\big[e^{itY}\big] = \mathbb{E}\big[e^{i(a^\top X)}\big]$$

6.3 nonindentical sum and CLT

1차원 Lindeberg–Feller CLT 독립인 확률변수 수열 $Y_{n1}, Y_{n2}, \ldots, Y_{nn}$에 대하여 $\mathbb{E}[Y_{ni}] = 0$, $\mathrm{Var}(Y_{ni}) = \sigma_{ni}^2$이라 하자. $s_n^2 = \sum_{i=1}^n \sigma_{ni}^2$일 때, 임의의 $\epsilon > 0$에 대하여 다음의 **Lindeberg 조건**이 만족되면:

$$\frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E}\left[Y_{ni}^2 \cdot \mathbb{I}(|Y_{ni}| > \epsilon s_n)\right] \to 0 \quad \text{as } n \to \infty$$

다음의 분포 수렴이 성립한다:

$$\frac{\sum_{i=1}^n Y_{ni}}{s_n} \xrightarrow{d} \mathcal{N}(0,1)$$

*의미: 개별 변수의 변동성이 전체 합의 변동성에 비해 충분히 작아, 어느 하나의 변수가 합의 분포를 지배하지 못함을 보장한다.

6.4 Multivariate Central Limit Theorem

이제 위 도구들을 이용하여 다변량 중심극한정리를 서술한다.

Theorem 6.3 (Multivariate CLT). $i.i.d.$ 확률벡터 $X_1, X_2, \cdots \in \mathbb{R}^k$가

$$\mathbb{E}[X_1] = \mu, \qquad \mathrm{Cov}(X_1) = \Sigma$$

를 만족한다고 하자. 그러면

$$\sqrt{n}\,(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

여기서

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

이다.

증명 스케치. 임의의 $a \in \mathbb{R}^k$에 대해

$$a^\top \sqrt{n}(\overline{X}_n - \mu) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} a^\top X_i - a^\top \mu\right).$$

이는 1차원 중심극한정리에 의해

$$\xrightarrow{d} \mathcal{N}(0, a^\top \Sigma a).$$

따라서 Cramér–Wold device에 의해

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

□

다변량 CLT는 이후 M-estimator의 점근정규성, Delta method, 그리고 고차원 추론의 기본 도구로 사용된다.

실제 통계적 추론(예: 가중 회귀 분석이나 불균형 설계)에서는 관측값들이 독립적이지만 서로 다른 분포를 갖는 경우가 빈번하다. 이를 위해 $i.i.d.$ 가정을 완화한 다변량 CLT를 살펴본다.

Theorem 6.4 (Multivariate Lindeberg–Feller CLT). 독립인 확률벡터 수열 $X_{n1}, X_{n2}, \ldots, X_{nn} \in \mathbb{R}^k$에 대하여 $\mathbb{E}[X_{ni}] = \mu_{ni}$이고 $\mathrm{Cov}(X_{ni}) = \Sigma_{ni}$라 하자. $\Sigma_n = \sum_{i=1}^{n} \Sigma_{ni}$가 가역(*invertible*)이고, 임의

의 $\epsilon > 0$에 대하여 다음의 ***Lindeberg* 조건**이 만족된다고 가정하자:

$$\frac{1}{\lambda_{\min}(\Sigma_n)} \sum_{i=1}^{n} \mathbb{E}\left[\|X_{ni} - \mu_{ni}\|^2 \cdot \mathbb{I}(\|X_{ni} - \mu_{ni}\| > \epsilon\sqrt{\lambda_{\min}(\Sigma_n)})\right] \to 0, \quad (n \to \infty)$$

여기서 $\lambda_{\min}$는 행렬의 최소 고유값이다. 그러면

$$\Sigma_n^{-1/2} \sum_{i=1}^{n} (X_{ni} - \mu_{ni}) \xrightarrow{d} \mathcal{N}(0, I_k)$$

가 성립한다.

다변량 확률벡터 $X_{n1}, \ldots, X_{nn} \in \mathbb{R}^k$의 점근적 정규성을 증명하기 위해, 임의의 고정된 벡터 $a \in \mathbb{R}^k$를 이용한 선형 조합(Projection)을 고려한다. 여기서는 모든 방향 a에 대한 정사영한 변수의 1차원 Lindeberg 조건이 성립함을 보인다. 이는 Cramér–Wold device에 의해 다변량분포 수렴을 의미한다.

1. $Y_{ni} = a^\top (X_{ni} - \mu_{ni})$로 정의하면, Y_{ni}는 1차원 확률변수이며 $s_n^2(a) := \mathrm{Var}(Y_{ni}) = a^\top \Sigma_{ni} a$이다.

2. 코시-슈바르츠에 의해 $Y_{ni}^2 \le \|a\|^2 \|X_{ni} - \mu_{ni}\|^2$.

3. 임의의 방향 a에 대한 1차원 Lindeberg 합 $L_n(a)$는 다변량 Lindeberg 조건에 의해 다음과 같이 지배된다.

$$\begin{aligned}
L_n(a) &:= \frac{1}{s_n^2(a)} \sum_{i=1}^{n} \mathbb{E}\left[Y_{ni}^2 \cdot \mathbb{I}\left(|Y_{ni}| > \epsilon s_n(a)\right)\right] \\
&\le \frac{1}{\|a\|^2 \lambda_{\min}(\Sigma_n)} \sum_{i=1}^{n} \mathbb{E}\left[\|a\|^2 \|X_{ni} - \mu_{ni}\|^2 \cdot \mathbb{I}\left(|Y_{ni}| > \epsilon s_n(a)\right)\right] \\
&\le \frac{\|a\|^2}{\|a\|^2 \lambda_{\min}(\Sigma_n)} \sum_{i=1}^{n} \mathbb{E}\left[\|X_{ni} - \mu_{ni}\|^2 \cdot \mathbb{I}\left(\|X_{ni} - \mu_{ni}\| > \epsilon\sqrt{\lambda_{\min}(\Sigma_n)}\right)\right] \\
&= \frac{1}{\lambda_{\min}(\Sigma_n)} \sum_{i=1}^{n} \mathbb{E}\left[\|X_{ni} - \mu_{ni}\|^2 \cdot \mathbb{I}\left(\|X_{ni} - \mu_{ni}\| > \epsilon\sqrt{\lambda_{\min}(\Sigma_n)}\right)\right] \xrightarrow{n\to\infty} 0
\end{aligned}$$

이로써 모든 방향에 대해 1차원 CLT가 성립하며, Cramér–Wold Device에 의해 다변량 CLT 증명이 완결된다.

Remark 6.5. Σ_n의 고윳값이 균일하게 발산한다고 가정할 때($\lambda_{\min}(\Sigma_n) \asymp \lambda_{\max}(\Sigma_n)$), 다변량 조건에서 $\lambda_{\max}$ 대신 $\lambda_{\min}$을 기준으로 보더라도, 전체 분산에 대한 개별 항의 기여도가 점근적으로 사라진다는 본질은 동일하다.

Lyapunov 조건 (충분조건) 실무적으로 Lindeberg 조건보다 검증이 쉬운 **Lyapunov 조건**이 자주 사용된다. 어떤 $\delta > 0$에 대하여

$$\frac{\sum_{i=1}^{n} \mathbb{E}[\|X_{ni} - \mu_{ni}\|^{2+\delta}]}{(\sum_{i=1}^{n} \lambda_{\min}(\Sigma_{ni}))^{1+\delta/2}} \to 0$$

이 만족되면 Lindeberg 조건이 성립하며, 따라서 다변량 CLT가 적용된다. 이는 고차 적률(higher-order moments)의 존재가 점근적 정규성을 보장함을 보여준다.

이러한 Non-$i.i.d.$ CLT는 회귀 분석의 계수 추정량 $\hat{\beta}$이 $X^\top X$의 구조에 따라 어떻게 점근 정규성을 갖는지 설명하는 핵심 근거가 된다.

7 참고사항

7.1 기본 확률측도와 분포(법칙)에 대한 확률의 구분

확률론에서 가장 기본이 되는 객체는 확률공간 $(\Omega, \mathcal{F}, \mathbb{P})$ 이다. 여기서 $\mathbb{P} : \mathcal{F} \to [0, 1]$ 는 표본공간(sample space) Ω 위에 정의된 확률측도(probability measure)다. $\mathbb{P}(A)$는 사건 $A \in \mathcal{F}$에 대해 Ω 위에서 정의된 확률이다. 여기서 $w \in \Omega$는 숫자로 표현 될 필요는 없으며, Ω 위해 연산이 정의 될 필요도 없다. 표본 공간 위에서 오직 필요한 것은 집합 연산 뿐이다. 한편 확률의 공리적 정의를 만족시키기 위한 조건으로 확률 함수의 정의역을 시그마 대수 σ-algebra로 제한하였다. (Ω의 모든 부분집합이 아니다) 확률 변수는 대수적 연산이 정의되지 않은 표본 공간 위의 원소들을 다루는데 효과적인 도구를 제공한다. 확률 변수를 통해 표본공간 위에 표본들은 우리가 수학적 도구로써 쉽게 다를 수 있는 실수 공간 위의 원소로 변환 된다. 실공간 위에서는 연산을 정의할 수 있고 거리(distance), 각도(angle)와 같은 유용한 수학적 개념들을 도입할 수 있어 확률을 다루는데 훨씬 편리해 진다.

다음으로 확률 변수가 표본을 실수로 옮기는 함수로서 가져야만 하는 특별한 성질을 살펴보겠다. 확률변수 $X : \Omega \to \mathbb{R}$가 주어졌다고 하자. 이때 X가 Ω 위의 확률을 실수공간 $\mathbb{R}$로 옮김과 동시에 $\mathcal{F}$의 시그마대수 구조를 보존할 수 있도록, X에 특별한 제약조건을 준다. 이 제약을 X의 measurability 라고 한다. 뿐만 아니라 $f(X)$와 같이 확률변수의 변환을 도입할 때, measurable 함수의 개념을 사용하는 것은 변환한 함수 $f(X)$ 여전히 시그마 대수 구조를 보존하도록 만들어 주기 위함이다. 표본을 X를 통해 실수 위로 옮긴후 $A \in \mathcal{F}$은 $B = \{X(w) : w \in A\} \subset \mathbb{R}$ 가 될 것이고 실수의 부분집합의 확률값은 이미 앞서 정의하였던 $\mathbb{P}(A)$가 되어야 할 것이다. 확률 변수를 통해 정의된 확률을 확률 변수의 확률분포 라고 부르고 그 확률 값은 다음과 같이 정의한다.

$$P_X(B) := \mathbb{P}(A).$$

위에서 논의한 확률은 표본 공간 위해서 정의된 확률과 확률 변수를 통해 실수 위해서 정의 된 확률,

	표본공간 확률	분포(실수공간) 확률
공간	Ω	$\mathbb{R}$
시그마대수	$\mathcal{F}$	$\mathcal{B}(\mathbb{R})$
확률측도	$\mathbb{P}$	P_X
사건의 예	$A \subset \Omega$	$B \subset \mathbb{R}$
확률값	$\mathbb{P}(A)$	$P_X(B) = \mathbb{P}(X \in B)$
사건의 연결	—	$X^{-1}(B) \in \mathcal{F}$

Table 1: $(\Omega, \mathcal{F}, \mathbb{P})$, $(\mathbb{R}, \mathcal{B}, P_X)$의 비교

두 가지 방법으로 표현 될 수 있음을 확인 하였다. 일반적으로 우리는 표본을 실수 위해서 표현하고 그것들의 다양한 연산을 함께 생각한다. 다시 말해 실제로 우리가 편리하게 다루어야 할 대상은 사실상 표본 공간 위에서 확률이 아니라 확률 변수를 통해 실수 위해서 정의된 확률임을 유추 해 볼 수 있다. 그래서 실수 위에서 다양한 패턴 혹은 사건을 표현할 수 있는 보렐-시그마 대수 $\mathcal{B}$ (열린집합으로 생성된 최소의 시그마 대수)를 생각하고 그 역상을 이용해서 확률값을 대응시킨다. 즉 $B \in \mathcal{B}$에 대해서

$$P_X(B) := \mathbb{P}(\{X^{-1}(B)\}).$$

여기서 확률변수가 표본공간의 시그마 대수 구조를 보존해야하므도, for $\forall B \in \mathcal{B}\ X^{-1}(B) \in \mathcal{F}$ 를 요구한다. 이런 조건을 확률변수 X가 $\mathcal{F}-$measurable 하다고 말한다. 따라서 확률변수는 다음과 같은 구조를 가진 measurable 함수로 정의한다. $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. 결론적으로 확률론에서 근본적인 확률은 표본공간 위의 $\mathbb{P}$이지만, 실제로 계산하고 분석하는 대상은 대부분 확률변수이기 때문에 우리는 주로 분포 P_X, 실수공간 위의 확률을 다루게 된다.

분포는 흔히 다음과 같은 기호로도 표현되므로 아래의 다양한 표현방법을 익히는 것이 좋다.
(확률측도의 표현)

$$\mu_X := \mathbb{P}_X, \qquad \mathcal{L}(X) := \mathbb{P}_X.$$

따라서 다음은 모두 같은 의미를 가진다:

$$\mathbb{P}(X \in B) = \mathbb{P}_X(B) = \mu_X(B) = \mathcal{L}(X)(B).$$

(기대값의 표현) 일반적인 측도위에서 정의된 적분 (abtract integral)을 이용하면 measurable 함수에 대한 적분을 자연스럽게 정의할 수 있다. 확률변수 X에 대한 기대값은 표본공간 Ω에서 적분하는 대신,

X의 분포 $\mathbb{P}_X$에 대해 적분하여 표현할 수 있다:

$$\mathbb{E}[f(X)] = \int_{\Omega} f(X(\omega))d\mathbb{P}(\omega) = \int_{\mathbb{R}} f(x)dP_X(x).$$

(Radon-Nykodym Derivatives) Radon–Nikodym 정리는 일반적인 측정가능공간 measurable space $(\Omega, \mathcal{F})$ 위에서 정의된 두 측도 $\mathbb{P}, \mathbb{Q}$ 사이의 관계를 다룬다. Radon–Nikodym derivative는 "한 측도가 다른 측도에 대해 얼마나 연속적으로 분포하는가" 를 측정하는 도구이다. 만약 $\mathbb{Q}$가 $\mathbb{P}$에 대해 절대연속이고 두 측도가 σ-finite 라고 하면 measurable function f가 존재해서

$$\mathbb{Q}(A) = \int_A f(w)d\mathbb{P}(w) \text{ for all } A \in \mathcal{F}$$

여기서 f를 $d\mathbb{Q}/d\mathbb{P}$라 표현하고 $d\mathbb{Q}/d\mathbb{P} : \Omega \mapsto \mathbb{R}$ 그리고 $\mathcal{F} - measurable$이다. 여기서 유의하여 받아들여야 하는 점은 $\mathbb{P}, \mathbb{Q}$ 둘 다 확률측도일 필요는 없다는 것이다.
이는 분포함수에서 Radon-Nykodym Derivatives를 다루게 되면, measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ 위에 정의한 두 측도 P, Q에 대해 정의할 수 있다. 특별히 P가 르벡측도 (유한하지는 않지만 σ-finite) 그리고 Q가 연속형 확률분포라고 하면 비음인 Borel 가측 함수가 f가 존재해서 모든 보렐집합 A에 대해서

$$\mathbb{Q}(A) = \int_A f(x)d\mathbb{P}(x) \text{ for all } A \in \mathcal{B}(\mathbb{R})$$

이 성립한다. 여기서 f를 확률측도 Q의 (르벡측도에 대한) pdf라고 한다. 그리고 우리는 f 가 특별한 경우의 Radon-Nykodym Derivatives $d\mathbb{Q}/d\mathbb{P}$임을 알 수 있다.

8 분포간의 거리

Definition 8.1 (Total Variation Distance)**.** 측도공간 $(\mathcal{X}, \mathcal{F})$ 위의 두 확률측도 P, Q에 대해 total variation distance는 두 분포가 사건(event) 위에서 얼마나 다르게 행동하는지를 측정하는 거리이다. 두 확률측도 P, Q에 대해 total variation distance는

$$d_{\text{TV}}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

로 정의된다.

모든 measurable set A 중에서 확률 차이가 가장 크게 나는 사건을 택했을 때의 최대값이다. 또 다른

동치 표현으로는, 함수 f가 항상 $|f(x)| \le 1$을 만족할 때

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2} \sup_{\|f\|_\infty \le 1} \left| \int f\, dP - \int f\, dQ \right|$$

로도 쓸 수 있다.

Proposition 8.2. 어떤 측도 μ에 대해 $P, Q \ll \mu$이고, *Radon–Nikodym derivative* (확률밀도함수) $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$가 존재한다고 하자. 그러면 *total variation distance*는 다음과 같이 적분 형태로 주어진다:

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2} \int |p - q|\, d\mu.$$

특히 $\mathbb{R}^d$에서 르벡측도에 대한 *pdf*가 존재하면 $d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\int_{\mathbb{R}^d} |p(x) - q(x)|\, dx$가 된다.

Proposition 8.3. $d_{\mathrm{TV}}(P_n, P) \to 0$이면 $P_n \Rightarrow P$이다.

Proof. TV distance의 동치 표현에 의해

$$d_{\mathrm{TV}}(P_n, P) = \frac{1}{2} \sup_{\|f\|_\infty \le 1} \left| \int f\, dP_n - \int f\, dP \right|.$$

따라서 임의의 (bounded measurable) f에 대해

$$\left| \int f\, dP_n - \int f\, dP \right| \le 2\|f\|_\infty\, d_{\mathrm{TV}}(P_n, P).$$

즉, $d_{\mathrm{TV}}(P_n, P) \to 0$이면 모든 bounded measurable 함수에 대해 적분이 수렴한다. 특히 모든 bounded continuous f에 대해서도 성립하므로 $P_n \Rightarrow P$이다. □

따라서 TV 수렴은 분포수렴보다 강한 수렴이다.

Theorem 8.4 (Scheffé's lemma)**.** 밀도함수 $p_n = \frac{dP_n}{d\mu}$, $p = \frac{dP}{d\mu}$가 존재한다고 하자.
또한 다음을 가정하자:

- $p_n(x) \to p(x)$가 μ-*a.e.* 성립한다.
- $\int p_n\, d\mu = 1$이고 $\int p\, d\mu = 1$이다.

그러면 다음이 성립한다:

$$\int |p_n - p| \, d\mu \;\to\; 0.$$

즉,

$$d_{\mathrm{TV}}(P_n, P) = \frac{1}{2} \int |p_n - p| \, d\mu \;\to\; 0.$$

따라서 P_n은 P로 *total variation distance*에서 수렴하며, 특히 분포수렴 $P_n \Rightarrow P$도 성립한다.

이 정리는 확률 매도 함수의 수렴이 분포의 수렴보다 훨씬 더 강한 가정임을 나타낸다 왜냐하면 total variation distance의 수렴이 본질적으로 분포 수렴보다 엄격한 조건이며, 확률 밀 도 함수의 수렴이 total variation distance의 수렴을 함의하기 때문이다.