PSH

수리통계학 강의록

Mathematical Statistics Lecture Note

June 15, 2019

Seoul National University

함께했던 날들과,
함께할 날들을 위해.
2018년의 끝자락, 관악에서.

Acronyms

WLOG 일반성을 잃지 않고 (Without Loss Of Generality)
TFAE 다음은 서로 동치이다. (The Followings Are Equivalent.)
ow. 그렇지 않으면 (OtherWise)
ae. 거의 어디서나 (Almost Everywhere)
LUBP Least Upper Bound Property
GLBP Greatest Lower Bound Property
MSP Monotone Sequence Property
MCT 단조수렴정리 (Monotone Convergence Theorem)
DCT Lebesgue의 지배수렴정리 (Lebesgue's Dominated Convergence Theorem)
FTC 미적분학의 기본정리 (Funcdamental Theorem of Calculus)
MVT 평균값 정리 (Mean Value Theorem)
IVT 중간값 정리 (Intermediate Value Theorem)
INFT 역함수 정리 (INverse Function Theorem)
rv. 확률변수 (Random Variable) 혹은 확률벡터 (Random Vector)
PDF 확률밀도함수 (Probability Density Function)
PMF 확률질량함수 (Probability Mass Function)
CDF 누적분포함수 (Cumulative Distribution Function)
MGF 적률생성함수 (Moment Generating Function)
CF 특성함수 (Characteristic Function)
PGF 확률생성함수 (Probability Generating Function)
iid. Independent and Identically Distributed rv.
io. Infinitely Often
as. 거의 확실하게 (Almost Surely)

Contents

Part I

Probability and Measure

Chapter 1

Measure Theory

Chapter 2

Probability Theory

Abstract 확률론은 20세기 들어 급격하게 발전한 분야이다. 애초에 확률이라는 개념이 수학에 편입된 것이 그리 오래되지 않았다. 이는 Descart의 연역주의의 영향이 진하게 남아있던 근대 유럽의 수학에서 불확실성을 다루기를 꺼려했기 때문이다. 오죽했으면 "거의 확실한 것은 거의 확실히 거짓이다."라고까지 했을까. 하지만 도박 문제(de Méré's problem)와 같이 불확실성을 계량하여 다루어야 할 필요성은 조금씩 늘어갔고, 이러한 현실적 요구에 확률은 Pascal, Fermat, Lagrange 등의 기라성같은 수학자들에 의해 조금씩 건드려지기 시작했다. 이때까지만 하더라도 확률이 무엇인지에 대한 수학자들의 생각은 '어떤 사건이 발생할 가능성' 정도였다. 이러한 확률의 의미가 직관적으로 분명하였기에 이에 의문을 제기하는 사람도 없었고, 그럴 필요도 느끼지 못했다. 그러나 미적분학에서 극한의 개념이 그러하였듯, 확률에 대한 연구가 계속될수록 미묘한 잡음이 발생하기 시작했고, 이는 확률의 개념에 대한 엄밀한 수학적 접근이 필요함을 암시했다. 결국 '확률은 무엇인가?'라는 질문의 답을 찾기 위한 긴 여정이 시작되었고, Laplace가 확률에 해석학을 끼얹은 것을 시작으로 Kolomogorov가 그의 명저 *Grundbegriffe der Wahrscheinlichkeitsrechnung*(영어: *Foundations of The Theory of Probability*)에서 측도론으로 확률을 정의하면서 그 여정은 일단락되게 된다. 본 장에서는 그 여정의 끝에서 수학자들이 괘뚫어본 확률의 본질에 대해 살펴보도록 하자.

2.1 Probability Spaces

단도직입적으로 말하면, 확률은 측도의 특별한 한 종류에 불과하다. 곧 확률은 일종의 넓이나 부피와 같은 개념으로 무언가를 재는 역할을 한다. 현실적인 의미를 생각하면 '가능성'을 잰다고도 할 수 있겠다.

Definition 2.1 측도공간 $(\Omega, \mathscr{F}, \mathbb{P})$에 대해 $\mathbb{P}(\Omega) = 1$이면 이때의 유한 측도 $\mathbb{P}$를 **확률측도 (probability measure)**라 하고, 유한 측도공간 $(\Omega, \mathscr{F}, \mathbb{P})$를 **확률공간 (probability space)**이라 한다. 나아가 집합 Ω를 **표본공간 (sample space)**이라 하고, σ−대수 $\mathscr{F}$에 속하는 임의의 집합 E를 **사건 (event)**이라 하여 $\mathbb{P}(E)$의 값을 사건 E의 **확률 (probability)**이라 한다.

확률의 본질이 측도라는 위의 정의는 나름 설득력이 있다. 그렇다면 이걸로 다 된 것일까? 아쉽게도 이제부터 해야 할 일이 태산이다. 일단 위의 정의를 받아들이기로 했다면, 지금까지 우리가 배웠던 확률의 대한 모든 내용들을 측도론의 언어로 다시 써야 한다. 곧 확률론을 다루는 본 장의 내용은 기본적으로 '번역 작업' 으로, 다행히 대부분의 경우 이 번역 작업은 크게 어렵지 않을 것이다. 이는 측도론이 확률의 내용들을 형식화하기에 좋은 이론이라서이기도 하지만, 앞서 우리가 측도론을 배우며 이 순간을 위해 조금씩 준비해 둔 것들이 꽤 많기 때문이다.

본격적으로 시작하기에 앞서, 맥락상 다소 뜬금없기는 하지만, 우리의 확률에 대한 인식의 근간을 이루는 equally likely outcome model을 한 번은 언급하고 지나가는 것이 좋을 것 같다. 고등학교에서 경우의 수를 세는 문제로 흔히 접하는 equally likely outcome model은 표본공간으로 항상 유한집합 Ω를 가지고, 이의 모든 부분집합은 사건으로 간주된다. 나아가 한원소 집합인 사건은 특별히 **근원사건 (elementary event)**이라 불리며 각 근원사건의 확률은 정확히 $1/|\Omega|$로 주어진다. (이렇게 각 근원사건의 확률이 같으므로 'equally likely' 이다. 고등학교에서는 흔히 '같은 정도로 확실하다' 로 번역한다.) 이러한 setting에서 우리는 임의의 사건 E의 확률을 $|E|/|\Omega|$로 정하고, 여기서의 $|E|$를 구하기 위해 여태껏 경우의 수를 열심히 계산하여 왔다. 이상의 내용을 측도론의 언어로 담백하게 번역하면 equally likely outcome model은 '표본공간 Ω에서의 셈측도 #에 대해 $(\Omega, \mathscr{P}(\Omega), \#/|\Omega|)$로 주어진 확률공간' 이 된다. 곧 측도론으로 바라보면 equally likely outcome model도 측도공간의 특별한 한 종류에 불과하다.

보통 초급 확률론 교재에서는 고등학교에서와 마찬가지로 이 equally likely outcome model에 집중하여 경우의 수를 계산하는 fancy한 trick을 소개하는 데 많은 분량을 할애하곤 하지만, 여기서는 이에 대한 논의는 Sheldon Ross의 A First Course in Probability를 참고문헌으로 실어두는 것으로 대신한다. 우리는 일반적인 이론의 전개에 보다 관심이 있기에 확률공간의 한 예시에 불과한 equally likely outcome model에 대해서는 그다지 관심이 없고, 곧 이 책에서 equally likely outcome model이 다시 등장하는 일은 (아마) 없을 것이다. 또한, 일반적으로 표본공간 위의 σ−대수 $\mathscr{F}$가 모든 한원소 집합을 포함한다는 보장이 없으므로 근원사건이라는 개념도 그다지 쓸모가 없을 것이다.

다시 원래의 이야기로 돌아와, 본격적으로 번역 작업을 시작해보자. 우선 확률의 기본적인 성질 정도는 측도의 성질들로부터 거의 자명하게 얻어진다.

Theorem 2.2 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 사건 E, F에 대해 다음이 성립한다.

i. $\mathbb{P}(\emptyset) = 0$.
ii. (σ−가법성) 서로소인 사건열 $\{E_i\}$에 대해 $\mathbb{P}(\bigsqcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$이다.
iii. $0 \le \mathbb{P}(E) \le 1$.
iv. $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$.
v. (단조성) 만약 $E \subseteq F$이면 $\mathbb{P}(E) \le \mathbb{P}(F)$이다.

Proof 이는 정의와 측도의 기본적인 성질로부터 자명하다. □

Theorem 2.3 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에 대해 다음이 성립한다.

i. (포함배제의 원리) 사건 $E_1, \cdots, E_l$에 대해

$$\mathbb{P}\left(\bigcup_{i=1}^{l} E_i\right) = \sum_{i=1}^{l} (-1)^{i-1} \sum_{1 \le j_1 < \cdots < j_i \le l} \mathbb{P}\left(\bigcap_{k=1}^{i} E_{j_k}\right)$$

이다.
ii. (σ−반가법성) 사건열 $\{E_i\}$에 대해 $\mathbb{P}(\bigcup_{i=1}^{\infty} E_i) \le \sum_{i=1}^{\infty} \mathbb{P}(E_i)$이다.

Proof i는 측도론의 포함배제의 원리를 $(\Omega, \mathscr{F}, \mathbb{P})$에 적용한 결과이고, ii는 측도의 σ−가법성이 σ−반가법성을 함의한다는 점에서 자명하다. □

Theorem 2.4 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 사건열 $\{E_i\}$에 대해

$$\mathbb{P}(\liminf_{i\to\infty} E_i) \le \liminf_{i\to\infty} \mathbb{P}(E_i) \le \limsup_{i\to\infty} \mathbb{P}(E_i) \le \mathbb{P}(\limsup_{i\to\infty} E_i)$$

가 성립한다. 특별히, $E_i \to E$이면 $\mathbb{P}(E_i) \to \mathbb{P}(E)$이다.

Proof 이는 정리 ? ? 로부터 자명하다. □

위의 정리의 단서를 흔히 **확률측도의 연속성 (continuity of probability measure)**이라 한다. 한편, 확률론에서는 사건열 $\{E_i\}$에 대해 이의 상극한 $\limsup_{i\to\infty} E_i$를 간단히 E_i io. 라 쓰기도 한다. 여기서 io.는 infinitely often의 줄임말로 곧 $\omega \in E_i$ io.는 $\omega \in \Omega$가 사건 $E_1, E_2, \cdots$에 무한히 많이 속한다는 것인데, 이는 $\limsup_{i\to\infty} E_i = \bigcap_{j=1}^{\infty} \bigcup_{i=j}^{\infty} E_i$임을 생각해 보면 나름 make sense하는 표기법이다.

Definition 2.5 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에 대해 영집합인 사건을 **영사건 (null event)**이라 한다.

확률측도도 측도이기에 '($\mathbb{P}$−) 거의 어디서나' 라는 개념이 자주 쓰이는데, 확률론에서는 이를 **($\mathbb{P}$−) 거의 확실하게 (($\mathbb{P}$−)almost surely)**라 하기도 한다. 이는 영사건이 확률이 0인

사건이므로 어떤 성질이 $\mathbb{P}$−거의 어디서나 성립하면 곧 1의 확률로 성립하게 되기 때문에 생겨난 관례이다.

이와 관련하여 '확률이 0인 사건' 과 '불가능한 사건' 은 서로 다르다는 것에 주의할 필요가 있다. 확률인 0인 사건은 표본공간 위의 σ−대수 $\mathscr{F}$에 속하는 사건이지만 그 확률이 0일 뿐이고, 불가능한 사건은 애초에 $\mathscr{F}$에 속하지 않아 그 이름과는 달리 엄밀히는 사건이 아니다. 간단한 예시를 위해 $[0,1]$에서 임의로 점 하나를 택하는 상황을 생각해보자. 이를 확률공간으로 형식화한다면, 표본공간은 $\Omega=[0,1]$이고 $\mathscr{F}=\mathscr{B}_1|_\Omega$, 확률측도는 $\mathbb{P}=(\mu_1)_\mathscr{F}$ 정도로 둘 수 있을 것이다. 그렇다면 $\mathbb{P}\{0.5\}=0$이므로 정확히 0.5를 뽑는 사건은 확률이 0인 사건이지만, 그렇다고 이가 일어나는 것 자체가 불가능한 것은 아니다. 반대로, 2를 뽑는 사건은 애초에 발생이 불가능한 사건으로 이 경우에 $\{2\}\notin\mathscr{F}$이므로 이는 엄밀하게는 사건이 아니어서 확률의 부여가 불가능하다. 이와 비슷하게, '확률이 1인 사건' 과 '항상 발생하는 사건' 도 서로 다른다.

이어서, 조건부확률을 도입하고 그 성질을 측도론으로 보이자. 다만, 이후에 조건부기댓값과 조건부분포를 엄밀히 도입하기 위해서는 조건부확률의 개념을 격변에 준할 정도로 일반화시켜야 하는데, 이에 하나의 절을 오롯히 할애해야 할 정도의 논의가 필요하므로 여기에서는 아주 간단하게만 다루도록 한다.

Definition 2.6 확률공간 $(\Omega,\mathscr{F},\mathbb{P})$와 영사건이 아닌 사건 E에 대해 **사건 E에 대한 조건부확률 (conditional probability under event E)**을 $\mathbb{P}(\cdot|E):\mathscr{F}\to\mathbb{R}_0^+$로 쓰고 $\mathbb{P}(\cdot|E):F\mapsto\mathbb{P}(F\cap E)/\mathbb{P}(E)$로 정의한다.

Proposition 2.7 확률공간 $(\Omega,\mathscr{F},\mathbb{P})$와 영사건이 아닌 사건 E에 대한 조건부확률 $\mathbb{P}(\cdot|E)$는 확률측도이다. 따라서 $(\Omega,\mathscr{F},\mathbb{P}(\cdot|E))$는 확률공간을 이룬다.

Proof 우선 $\mathbb{P}(\emptyset|E)=\mathbb{P}(\emptyset)/\mathbb{P}(E)=0,\mathbb{P}(\Omega|E)=\mathbb{P}(E)/\mathbb{P}(E)=1$임은 분명하고, 임의의 서로소인 사건열 $\{E_i\}$에 대해

$$\begin{aligned}\mathbb{P}\left(\bigsqcup_{i=1}^{\infty}E_i|E\right)&=\frac{\mathbb{P}(\bigsqcup_{i=1}^{\infty}E_i\cap E)}{\mathbb{P}(E)}\\&=\frac{\mathbb{P}(\bigsqcup_{i=1}^{\infty}(E_i\cap E))}{\mathbb{P}(E)}\\&=\sum_{i=1}^{\infty}\frac{\mathbb{P}(E_i\cap E)}{\mathbb{P}(E)}\\&=\sum_{i=1}^{\infty}\mathbb{P}(E_i|E)\end{aligned}$$

이므로 $\mathbb{P}(\cdot|E)$가 확률측도임을 안다. □

Theorem 2.8 확률공간 $(\Omega,\mathscr{F},\mathbb{P})$에 대해 다음이 성립한다.

i. 사건 E와 영사건이 아닌 사건 F에 대해 $\mathbb{P}(E\cap F)=\mathbb{P}(E|F)\mathbb{P}(F)$이다.

ii. (전확률 공식) 서로소인 가산개의 사건 $E_1, E_2, \cdots$에 대해 각 E_i가 영사건이 아니고 $\bigsqcup_{i=1}^{k} E_i = \Omega$이면 임의의 사건 E에 대해 $\mathbb{P}(E)=\sum_{i=1}^{k}\mathbb{P}(E|E_i)\mathbb{P}(E_i)$이다. (여기서 k는 유한할 수도 있고, ∞일 수도 있다.)

Proof i. 이는 조건부확률의 정의로부터 자명하다.

ii. i로부터 $\mathbb{P}(E)=\mathbb{P}(E\cap\bigsqcup_{i=1}^{k}E_i)=\mathbb{P}(\bigsqcup_{i=1}^{k}(E\cap E_i))=\sum_{i=1}^{k}\mathbb{P}(E\cap E_i)=\sum_{i=1}^{k}\mathbb{P}(E|E_i)\mathbb{P}(E_i)$이다. □

Theorem 2.9 (Bayes) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 서로소인 가산개의 사건 $E_1, E_2, \cdots$에 대해 각 E_i가 영사건이 아니고 $\bigsqcup_{i=1}^{k} E_i = \Omega$이면 임의의 사건 E에 대해

$$\mathbb{P}(E_1|E)=\frac{\mathbb{P}(E|E_1)\mathbb{P}(E_1)}{\sum_{i=1}^{k}\mathbb{P}(E|E_i)\mathbb{P}(E_i)}$$

이다. (여기서 k는 유한할 수도 있고, ∞일 수도 있다.)

Proof 전확률 공식으로부터 $\mathbb{P}(E_1|E)=\mathbb{P}(E\cap E_1)/\mathbb{P}(E)=\mathbb{P}(E|E_1)\mathbb{P}(E_1)/\sum_{i=1}^{k}\mathbb{P}(E|E_i)\mathbb{P}(E_i)$가 자명하다. □

비록 증명은 간단하지만 Bayes의 정리는 실험적으로 구하는 것이 불가능한 조건부확률을 구하게 해준다는 엄청난 실용성과 함의를 지닌다. 쉽고 즐거운 예시를 위해 유리가 예나에게 휴대전화로 0x2661(UTF-16 ♡)과 0x2665(UTF-16 ♥) 중 하나의 정보를 임의로 전송하였는데, 송수신의 과정에서 정보의 일부가 손상되어 결과적으로 예나는 0x2663(UTF-16 ♣)을 수신한 상황을 생각해보자. 이를 복원하기 위해 예나는 E를 0x2663을 수신하는 사건, F, G를 각각 0x2661과 0x2665를 전송하는 사건이라 두고 두 조건부확률 $\mathbb{P}(F|E)$와 $\mathbb{P}(G|E)$를 비교하여 전자가 더 크다면 0x2661로 복원하고 후자가 더 크다면 0x2665로 복원하며, 만약 같다면 재전송을 요청하기로 하였다. 이는 훌륭한 통계적 사고방식이지만 $\mathbb{P}(F|E)$와 $\mathbb{P}(G|E)$는 일의 선후가 뒤바뀐 확률이라 실험적으로 구할 수가 없다는 치명적인 문제가 있다. 예나가 자신의 휴대전화로 자신에게 0x2661과 0x2665를 수 회 전송하는 실험을 통해 근사하게나마 구할 수 있는 확률은 $\mathbb{P}(E|F)$와 $\mathbb{P}(E|G)$ 뿐이다. 여기서 Bayes의 정리는 $\mathbb{P}(F|E)$와 $\mathbb{P}(G|E)$를 $\mathbb{P}(E|F)$와 $\mathbb{P}(E|G)$의 조합으로써 계산할 수 있도록 하여 문제를 해결하는 결정적인 역할을 한다. 따라서 예나가 실험적으로 구한 $\mathbb{P}(E|F)$와 $\mathbb{P}(E|G)$의 근사치가 만약 $1/200, 1/300$이었다면 Bayes의 정리로부터 $\mathbb{P}(F|E)\approx 3/5, \mathbb{P}(G|E)\approx 2/5$가 되어 ♣를 ♡로 복원할 수 있다. 요컨대, Bayes의 정리는 선후관계나 인과관계를 역전시켜주는 마법의 공식이다.

이러한 Bayes 정리는 이후 통계학에 큰 지각변동을 일으켜 이를 기초로 하는 Bayesian이라는 독자적인 학파가 구성되기에 이르렀고, 오늘날 기계학습과 같은 분야에서 요긴하게

쓰이는 모양이다. (이와 구분하여 기존의 통계학 학파를 빈도주의라 한다.) 물론, 이런 학파는 어디까지나 확률의 해석에 대한 차이로 구분되는 것이지 확률의 측도론적 정의나 접근방식으로 구분되는 것은 아니기에 빈도주의와 Bayesian의 구분이 본 장에서는 필요하지 않지만, 이 책에서는 특별한 언급이 없는 이상 빈도주의의 관점에서 확률을 바라본다. 여기에는 통계학 교양 정도를 들은 수준에서는 빈도주의의 관점이 Bayesian의 관점보다 조금 더 친숙하다는 것을 빼면 다른 그럴싸한 이유는 없다. 만약 자신이 Bayesian이라면 그들의 방식대로 해석하면 그만이고, 당연히 Bayesian에게도 측도론적인 엄밀한 확률론은 훌륭한 이론의 토대가 될 것이다.

이번 절에서 마지막으로 다룰 것은 바로 독립에 관한 내용인데, 앞서 확률의 기본적인 성질과 기본적인 조건부확률을 큰 어려움 없이 도입할 수 있었던 것과는 달리, 독립성을 측도론의 언어로 번역해 내는 것은 살짝 어렵다. 이는 독립이라는 개념이 측도론에서는 그 느낌조차 찾아보기 힘든, 완전히 새로운 개념이기 때문이다. 우리는 측도론에서 적분을 정의할 때와 비슷하게 독립을 그 정의를 점차 일반화시키는 방법으로 도입할 것이다. 우선 가장 기본적인 독립의 정의로 시작한다.

Definition 2.10 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 사건 $E_1, \cdots, E_k$를 생각하자. 만약 각 $l \leq k$와 임의의 서로다른 $i_1, \cdots, i_l \leq k$에 대해 $\mathbb{P}(\bigcap_{j=1}^{l} E_{i_j}) = \prod_{j=1}^{l} \mathbb{P}(E_{i_j})$가 성립하면 이때의 사건 $E_1, \cdots, E_k$를 **(서로) 독립 ((mutually) independent)**이라 한다. 한편, 만약 위의 성질이 $l = 2$에 대해서만 만족되면, 즉 임의의 서로다른 $i, j \leq k$에 대해서 $\mathbb{P}(E_i \cap E_j) = \mathbb{P}(E_i)\mathbb{P}(E_j)$가 성립하는 것에 그치면 이때의 사건 $E_1, \cdots, E_k$를 **pairwise 독립 (− independent)**이라 한다.

서로 독립인 것과 pairwise 독립이 다르다는 사실은 잘 알려진 사실이다. 흔해빠진 주사위와 동전 예시를 피하기 위해 해석개론을 수강신청한 수지와 이제훈이 같이 밤새 과제를 하게 된 것을 계기로 서로 어느정도 이성으로서 호감을 가지게 된, 이른바 '썸 탄다' 불리우는 상황을 생각하자. 수업을 듣던 어느날, 수지와 이제훈의 교재가 실수로 서로 바뀌어 다음 수업시간 전에 만나 책을 다시 바꾸기로 하였는데, 둘은 이를 앞두고 책 사이에 고백편지를 살짝 끼워넣을까 고민하고 있다고 하자. 여기서 사건 E, F를 각각 수지와 이제훈이 고백편지를 끼워넣는 사건이라 하고, 수지나 이제훈이 고백편지를 넣을 확률은 $1/2$로 같으며 이는 서로 독립이라 하자. 이제 사건 G를 어느 한쪽만 고백편지를 받는 사건이라 하면 E, F, G는 pairwise 독립이지만 서로 독립은 아니다. (직접 계산해보자.)

이제 유한개의 사건 사이의 독립을 무한개의 사건 사이의 독립으로 확장하는 것은 어렵지 않다.

Definition 2.11 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 사건의 모임 $\mathscr{C}$를 생각하자. 만약 임의의 사건 $E_1, \cdots, E_k \in \mathscr{C}$가 서로 독립이면 이때의 집합 $\mathscr{C}$를 **독립 (independent)**이라 한다.

다음으로, 유한개의 사건의 모임 사이의 독립을 정의한다.

Definition 2.12 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 사건의 모임 $\mathscr{C}_1, \cdots, \mathscr{C}_k$를 생각하자. 만약 각 $i \leq k$에 대해 임의로 택한 사건 $E_i \in \mathscr{C}_i$가 서로 독립이면 이때의 집합족 $\mathscr{C}_1, \cdots, \mathscr{C}_k$를 **독립 (independent)**이라 한다.

마지막으로 이를 무한개의 사건의 모임 사이의 독립으로까지 확장하면 독립의 가장 일반적인 정의를 얻는다.

Definition 2.13 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 사건의 모임의 모임 Γ를 생각하자. 만약 임의의 사건의 모임 $\mathscr{C}_1, \cdots, \mathscr{C}_k \in \Gamma$가 서로 독립이면 이때의 집합족 Γ를 **독립 (independent)**이라 한다.

이렇게까지 일반적인 형태의 독립성을 고려하는 이유는 다음 정리 때문이다.

Theorem 2.14 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 독립인 사건의 모임의 모임 $\{\mathscr{C}_\alpha\}$에 대해 각 $\mathscr{C}_\alpha$가 $\pi-$system이라 하면 $\{\sigma(\mathscr{C}_\alpha)\}$는 독립이다.

Proof 집합족 $\{\mathscr{C}_\alpha\}$에서 임의로 유한개의 원소를 택하여 이를 $\mathscr{C}_1, \cdots, \mathscr{C}_k$라 하고 이들이 생성하는 $\sigma-$대수 $\sigma(\mathscr{C}_1), \cdots, \sigma(\mathscr{C}_k)$가 서로 독립임을 보이면 증명은 충분하다. 이를 위해 사건 $E_2 \in \mathscr{C}_2, \cdots, E_k \in \mathscr{C}_k$를 임의로 택하고 집합족 $\mathscr{L} = \{F \in \mathscr{F} : F, E_2, \cdots, E_k$가 서로 독립$\}$을 생각하면 주어진 조건으로부터 $\mathscr{C}_1 \subseteq \mathscr{L}$임은 분명하다. 나아가 $\Omega \in \mathscr{L}$ 또한 분명하고, 임의의 $F \in \mathscr{L}$에 대해 $F, E_2, \cdots, E_k$가 서로 독립이면 $F^c, E_2, \cdots, E_k$도 서로 독립임을 쉽게 보일 수 있으므로 $F^c \in \mathscr{L}$이다. 비슷한 방법으로 $\mathscr{L}$에 속하는 임의의 서로소인 사건열 $\{F_i\}$에 대해 $\bigsqcup_{i=1}^\infty F_i, E_2, \cdots, E_k$가 서로 독립임을 보일 수 있으므로 $\bigsqcup_{i=1}^\infty F_i \in \mathscr{L}$도 성립한다. 이로부터 $\mathscr{L}$은 $\lambda-$system이 되어 $\mathscr{C}_1$이 $\pi-$system이라는 사실과 Dynkin의 π-λ 정리로부터 $\sigma(\mathscr{C}_1) = \lambda(\mathscr{C}_1) \subseteq \mathscr{L}$이고, 곧 $\sigma(\mathscr{C}_1), \mathscr{C}_2, \cdots, \mathscr{C}_k$는 서로 독립이다. 이제 이를 $k-1$번 반복하면 $\sigma(\mathscr{C}_1), \cdots, \sigma(\mathscr{C}_k)$가 서로 독립임을 보일 수 있고, 증명이 끝난다. □

위의 정리는 서로 독립인 사건의 모임들이 생성하는 $\sigma-$대수도 서로 독립임을 함의하는데, 잠시 이 결과의 의미에 대해 생각해보자. 독립성은 기본적으로 '정보'에 대한 이야기이다. 가장 기본적인 형태로 두 사건 E, F가 서로 독립인 경우를 살펴보면, 이는 사건 E의 발생여부에 대한 정보가 주어지더라도 사건 F의 발생여부에 대한 정보는 일체 추론해 낼 수 없음을 의미한다. 이는 2개 이상의 사건, 나아가 무한개의 사건의 독립에 대해서도 마찬가지이다. 예컨대 앞서 든 수지와 이제훈의 예시에서 E, F, G는 서로 독립이 아니었는데, 이는 수지가 고백편지를 넣는 사건 E와 이제훈이 고백편지를 넣는 사건 F의 발생여부에 대한 정보가 사건 G의 발생여부에 대한 정보를 100% 함의하기 때문이다.

사건의 모임 사이의 독립도 이와 비슷하게 이해할 수 있다. 앞서 사건들 사이의 독립에서 각 사건은 그 사건의 발생여부에 대한 이진 정보를 의미했다. 그렇다면 사건의 모임은 그

모임에 속한 각 사건들의 발생여부에 대한 이진 정보의 조합으로 이루어진 보다 풍성한 정보로 이해하는 것이 자연스러울 것이다. 따라서 사건의 모임 사이의 독립은 각 사건의 모임이 함의하는 정보가 서로 독립적이라는, 즉 어느 하나를 안다고 해서 다른 하나를 일체 추론해내지 못한다는 것을 의미한다. 이러한 해석의 관점에서 보면, 위의 정리의 결과는 어떤 사건의 모임들이 함의하는 정보가 서로 추론이 불가능하다면, 각 모임을 그가 생성하는 σ−대수로 확장하여 훨씬 더 많은 정보로 얻어도, 여전히 서로 추론이 불가능함을 의미한다. 뭔가 자명한 듯 자명하지 않은 이 결과를 보다 효율적으로 쓰기 위해 따름정리 하나를 소개한다.

Corollary 2.15 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 가산개의 무한한 행과 열을 가지는 사건의 배열

$$\begin{matrix} E_{11} & E_{12} & \cdots \\ E_{21} & E_{22} & \cdots \\ \vdots & \vdots & \ddots \end{matrix}$$

에 대해 $\{E_{ij}\}_{ij}$가 독립이라 하고, 각 $i \in \mathbb{N}$에 대해 집합족 $\mathscr{G}_i$를 $\{E_{ij}\}_j$가 생성하는 σ−대수라 하면 $\{\mathscr{G}_i\}$는 독립이다. 한편, 행이나 열의 개수가 유한한 경우에도 같은 결과가 성립하며, 나아가 각 행의 열의 개수가 달라도 가산개이기만 하면 여전히 같은 결과가 성립한다.

Proof 각 $i \in \mathbb{N}$에 대해 $\{E_{ij}\}_j$의 모든 유한 교집합의 모임 $\mathscr{P}_i$를 생각하면 $\mathscr{P}_i$는 π−system이고 $\sigma(\mathscr{P}_i) = \mathscr{G}_i$임이 거의 분명하다. 따라서 $\{\mathscr{P}_i\}$가 서로 독립이라는 사실만 보이면 앞선 정리로부터 $\{\mathscr{G}_i\}$가 독립이 되어 증명이 끝난다. 이를 위해 $\{\mathscr{P}_i\}$에서 임의로 유한개의 원소를 택하여 이를 $\mathscr{P}_1, \cdots, \mathscr{P}_k$라 하고 다시 임의로 사건 $F_1 \in \mathscr{P}_1, \cdots, F_k \in \mathscr{P}_k$를 택하면 각 $i \leq k$에 대해 $\mathscr{P}_i$의 구성으로부터 적당한 사건 $E_{i1}, \cdots, E_{il_i}$가 존재하여 $F_i = \bigcap_{j=1}^{l_i} E_{ij}$이다. 그렇다면 $\{E_{ij}\}_{ij}$가 독립이라는 사실로부터 $\mathbb{P}(\bigcap_{i=1}^{k} F_i) = \mathbb{P}(\bigcap_{i=1}^{k}\bigcap_{j=1}^{l_i} E_{ij}) = \prod_{i=1}^{k}\prod_{j=1}^{l_i}\mathbb{P}(E_{ij}) = \prod_{i=1}^{k}\mathbb{P}(\bigcap_{j=1}^{l_i} E_{ij}) = \prod_{i=1}^{k}\mathbb{P}(F_i)$가 되어 $\{\mathscr{P}_i\}$가 서로 독립임을 알고, 증명은 이로써 충분하다. □

이 따름정리를 적당히 응용하면 독립성에 대한 진부한 연습문제들, 예컨대 사건 E, F, G, H가 서로 독립이라면 $E \cap F$와 $G \setminus H$가 독립임을 보이라는 식의 문제들을 아주 깔끔하게 해결할 수 있다. 예시로 든 문제의 경우 배열 $EF//GH$를 생각하면 그만이다. 한편, 앞서 사건의 모임을 그에 포함된 각 사건의 발생여부로 구성된 정보의 집합으로 해석하였는데, 이는 독립의 경우에만 한정되는 해석이 아니어서 특히 그 모임이 σ−대수 $\mathscr{G}$인 경우 모든 사건의 집합 $\mathscr{F}$를 '전체 정보' 로, $\mathscr{G}$는 이의 '부분 정보' 로 해석하는 것이 유용한 경우가 많다.

나중을 위해 측도론에서 잠시 등장했던 Borel-Cantelli의 정리를 조금 보강하는 것으로 이번 절을 마친다.

Theorem 2.16 (Borel-Cantelli) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$와 사건열 $\{E_i\}$에 대해 다음이 성립한다.

i. 만약 $\sum_{i=1}^{\infty} \mathbb{P}(E_i) < \infty$이면 $\mathbb{P}(E_i \text{ io.}) = 0$이다.
ii. 만약 $\sum_{i=1}^{\infty} \mathbb{P}(E_i) = \infty$이고 $\{E_i\}$가 독립이면 $\mathbb{P}(E_i \text{ io.}) = 1$이다.

Proof i. 이는 측도론에서 배운 Borel-Cantelli의 정리로부터 자명하다.

ii. 우선 $\{E_i\}$가 독립이므로 따름정리 2.15로부터 $\{E_i^c\}$도 독립이다. 이제 임의의 $\varepsilon > 0$과 임의의 $j \in \mathbb{N}$를 택하면 $\sum_{i=j}^{\infty} \mathbb{P}(E_i) = \infty$이므로 $\lim_{k\to\infty} \exp(-\sum_{i=j}^{k} \mathbb{P}(E_i)) = 0$이다. 이로부터 적당한 $k_0 \in \mathbb{N}$가 존재하여 $k_0 \geq j$이고 $\exp(-\sum_{i=j}^{k_0} \mathbb{P}(E_i)) < \varepsilon$이므로 $\mathbb{P}(\bigcap_{i=j}^{k_0} E_i^c) = \prod_{i=j}^{k_0} \mathbb{P}(E_i^c) = \prod_{i=j}^{k_0} [1 - \mathbb{P}(E_i)] \leq \prod_{i=j}^{k_0} \exp(-\mathbb{P}(E_i)) = \exp(-\sum_{i=j}^{k_0} \mathbb{P}(E_i)) < \varepsilon$이다. 이는 곧 $\mathbb{P}(\bigcap_{i=j}^{\infty} E_i^c) \leq \mathbb{P}(\bigcap_{i=j}^{k_0} E_i^c) < \varepsilon$임을 뜻하므로 $\mathbb{P}(\bigcap_{i=j}^{\infty} E_i^c) = 0$임을 알고, 이는 다시 $\mathbb{P}(E_i \text{ io.}) = \mathbb{P}((\liminf_{i\to\infty} E_i^c)^c) = 1 - \mathbb{P}(\liminf_{i\to\infty} E_i^c) \geq 1 - \mathbb{P}(\bigcap_{i=j}^{\infty} E_i^c) = 1$에서 $\mathbb{P}(E_i \text{ io.}) = 1$임을 뜻한다. □

2.2 Random Variables and Random Vectors

앞선 절에서 사건 그 자체에 관련된 확률의 내용들을 측도론의 틀에 맞추어 열심히 옮겨 놓았으니, 이번 절에서는 확률변수라는 개념을 추가하여 보다 내용을 풍성하게 만들어보도록 하자. 확률변수를 통해 우리는 일일히 사건을 정의하지 않고도 확률의 여러 내용들을 보다 더 편리하게 사용할 수 있다.

Definition 2.17 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에 대해 가측함수 $X : \Omega \to \mathbb{R}^n$를 **($n$차원) 확률벡터 (($n$ dimensional) random vector)**라 하고, 특별히 $n = 1$이면 **확률변수 (random variable)**라 한다.

확률변수는 본질적으로 가측함수이기에 확률변수의 기본적인 성질들이 가측함수의 성질들로부터 자명하게 성립하는 것이 전혀 이상하지 않다.

Proposition 2.18 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 함수 $X : \Omega \to \mathbb{R}^n$에 대해 X가 rv.일 필요충분조건은 $X_1, \cdots, X_n$이 모두 rv.인 것이다.

Proof 이는 가측함수의 성분도 가측함수라는 점에서 자명하다. □

Theorem 2.19 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^m$와 Borel 함수 $g : \mathbb{R}^m \to \mathbb{R}^n$에 대해 합성 $g \circ X : \Omega \to \mathbb{R}^n$도 rv.이다.

Proof 이는 Borel 함수와 가측함수의 합성은 가측이라는 점에서 자명하다. □

비록 정의상 rv.는 엄연한 함수이지만 그 이름에서도 잘 드러나듯이 확률론에서는 이를 마치 변수처럼 생각하고 사용하는 경우가 많다. 이는 이론적인 이유에서라기보다 실생활의 응용에서 이렇게 생각하는 편이 조금 더 직관적으로 편하기 때문이다. 이러한 우리의 인식은 rv.와 관련된 여러 표기상의 관례에 잘 나타나는데, 위의 정리에서의 합성 $g \circ X$를 마치 g에 X라는 변수를 대입한 것으로 생각하여 $g(X)$로 쓰는 관례가 대표적인 예시이다. 그러나 이런 표기법은 어디까지나 관례일 뿐, rv.가 변수가 아닌 함수라는 사실은 항상 염두에 두고 있어야 한다.

이어서, 측도론에서 FTC를 일반화하는 과정에서 잠시 스쳐 지나갔던 pushfowarding이 다시 등장한다. 비록 측도론에서의 pushfowarding은 도구 역할에 그쳤지만, 확률론에서의 pushfowarding은 빼놓을 수 없는 핵심적인 개념이다.

Definition 2.20 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$에 대해 pushfoward 측도 $X_*\mathbb{P}$를 rv. X의 **분포 (distribution)**라 하고 $\mathbf{P}_X$로 쓴다. 특별히, $n \geq 2$인 경우 $\mathbf{P}_X$를 rv. $X_1, \cdots, X_n$의 **결합분포 (joint distribution)**라 하고 $\mathbf{P}_{X_1,\cdots,X_n}$으로 쓰기도 한다.

교양 통계학에서 배운 PDF나 CDF와 같이 확률론에는 방금 정의한 분포와 쉽게 혼동될 법한 개념들이 많고, 실제로 각자의 정의도 서로 긴밀히 연결되어 있다. 여기에 한술 더 떠서 문헌마다 조금씩 용어를 다르게 쓰는 바람에 혼란이 가중되는 부분이 없지 않지만, 대부분 논의의 맥락으로 적당히 구별할 수 있다. 아무튼 구태여 혼란을 초래할 필요는 없기에, 이 책에서는 용어를 최대한 잘 구별하여 사용하였다.

Proposition 2.21 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$에 대해 $(\mathbb{R}^n, \mathscr{B}_n, \mathbf{P}_X)$는 확률공간을 이룬다.

Proof 먼저 $\mathscr{B}_n \subseteq X_*\mathscr{F}$임을 보이기 위해 임의의 $A \in \mathscr{B}_n$를 택하면 $X^{-1}(A) \in \mathscr{F}$이므로 $A \in X_*\mathscr{A}$에서 $\mathscr{B}_n \subseteq X_*\mathscr{F}$이다. 따라서 $\mathbf{P}_X$가 확률측도임을 보이면 충분한데, 이는 $\mathbf{P}_X(\mathbb{R}^n) = \mathbb{P}(X^{-1}(\mathbb{R}^n)) = \mathbb{P}(\Omega) = 1$에서 쉽게 알 수 있고, 곧 증명이 끝난다. □

위의 명제는 분포를 통해 서로다른 확률공간에 정의된 확률측도를 각각 다루는 대신 이들을 pushfowarding하여 $\mathscr{B}_n$에서 정의된 확률측도로 일관되게 다룰 수 있음을 함의한다. 그리고 생각해보면, 이가 곧 사건을 직접 정의하고 사용하는 대신 rv.를 도입하여 사용하는 이유이다. 무엇이 될 지 모르는 확률공간 대신 우리가 잘 알고있는 실수공간에서의 확률측도를 다루는 것이 훨씬 편하다. 한편, 위의 정리의 역이 성립한다는 것도 꽤나 흥미로운 사실이다.

Theorem 2.22 Borel σ−대수 $\mathscr{B}_n$ 위의 확률측도 μ에 대해 적당한 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$가 존재하여 $\mu = \mathbf{P}_X$이다. 특별히, 이때 $(\Omega, \mathscr{F}) = (\mathbb{R}^n, \mathscr{B}_n)$이도록 잡을 수 있다.

PROOF 거의 자명하다. 함수 $X:\mathbb{R}^n\to\mathbb{R}^n$를 항등함수로 두면 이는 확률공간 $(\mathbb{R}^n,\mathscr{B}_n,\mu)$에서 정의된 rv.이고, 임의의 사건 E에 대해 $\mathbf{P}_X(E)=\mu(X^{-1}(E))=\mu(E)$에서 $\mu=\mathbf{P}_X$이다. □

Theorem 2.23 확률공간 $(\Omega,\mathscr{F},\mathbb{P})$에서 정의된 rv. $X:\Omega\to\mathbb{R}^m$와 Borel 함수 $g:\mathbb{R}^m\to\mathbb{R}^n$에 대해 $\mathbf{P}_{g(X)}=\mathbf{P}_X\circ g^{-1}$이다.

PROOF 임의의 $A\in\mathscr{B}_n$에 대해 $\mathbf{P}_{g(X)}(A)=\mathbb{P}((g\circ X)^{-1}(A))=\mathbb{P}(X^{-1}(g^{-1}(A)))=\mathbf{P}_X(g^{-1}(A))=(\mathbf{P}_X\circ g^{-1})(A)$이므로 $\mathbf{P}_{g(X)}=\mathbf{P}_X\circ g^{-1}$이다. □

자연스러운 다음 순서는 고등학교 시절부터 들어와 이름만은 익숙한 이산확률변수와 연속확률변수를 엄밀하게 측도론의 언어로 정의하는 것이다. 물론, 고등학교나 교양 통계학에서 각각을 정의하지 않는 것은 아니지만, 그 정의가 뭔가 어색하고 작위적이라는 느낌을 지우기 힘든데, 아래의 측도론적인 정의는 더할 나위 없이 깔끔하고 명쾌하다.

Definition 2.24 확률공간 $(\Omega,\mathscr{F},\mathbb{P})$에서 정의된 rv. $X:\Omega\to\mathbb{R}^n$에 대해 $\mathbf{P}_X$가 이산측도이면 이때의 rv. X를 **이산확률벡터 (discrete rv.)**라 한다. 또한, 만약 $\mathbf{P}_X$가 μ_n에 대해 절대연속이거나 특이연속이면 이때의 rv. X를 각각 **연속확률벡터 (continuous rv.)** 혹은 **특이확률벡터 (singular rv.)**라 한다. 특별히, $n=1$인 경우 이산확률벡터, 연속확률벡터, 특이확률벡터를 각각 **이산확률변수**, **연속확률변수**, **특이확률변수**라 한다.

Proposition 2.25 확률공간 $(\Omega,\mathscr{F},\mathbb{P})$에서 정의된 rv. $X:\Omega\to\mathbb{R}^n$에 대해 이는 이산확률벡터, 연속확률벡터, 특이확률벡터의 정의 중에서 두 개의 이상을 동시에 만족시킬 수 없다.

PROOF 모순을 유도하기 위해 X가 이산확률벡터인 동시에 연속확률벡터라고 하자. 그렇다면 $\mathbf{P}_X$는 가산 지지집합 $A\in\mathscr{B}_n$를 가지는데, $\mu_n(A)=0$에서 $\mathbf{P}_X(A)=0$의 모순이 발생한다. 이번에는 X가 이산확률벡터인 동시에 특이확률벡터라 하자. 그렇다면 이건과 같이 $\mathbf{P}_X$는 가산 지지집합 $A\in\mathscr{B}_n$를 가지는데, 이의 가산개의 원소를 $x_1,x_2,\cdots$와 같이 나열하면 $\mathbf{P}_X(A)=\sum_{i=1}^{k}\mathbf{P}_X\{x_i\}=0$에서 모순이 발생한다. (여기서 k는 유한할 수도 있고, ∞일 수도 있다.) 마지막으로 X가 연속확률벡터인 동시에 특이확률벡터라 하면 $\mathbf{P}_X\ll\mu_n$이고 $\mathbf{P}_X\perp\mu_n$이므로 $\mathbf{P}_X=0$의 모순이 발생하고, 증명은 이로써 충분하다. □

일반적으로, 어떤 rv.가 이산인지, 연속인지, singular인지는 확률측도 $\mathbb{P}$와 rv. X 모두에 의해 결정되는 것이지, 이 중 어느 하나에 의해 일방적으로 결정되는 것이 아니다. 즉, 둘 중 어느 하나만 보고서 X가 이산인지, 연속인지, singular인지는 알 수 없다. 이로 말미암아 우리가 rv.에 대해 당연하게 생각하던 사실들에 미묘한 혼란이 생겨나게 된다.

우선 X가 이산확률벡터이지만 그 치역은 가산이 아닐 수 있다. 물론, 대부분의 응용에서는 이산확률변수의 치역도 가산으로 주어지지만, 이는 우연의 일치 그 이상도 이하도 아니다. 극단적인 예시로 가측공간 $(\mathbb{R},\mathscr{B}_1)$ 위의 측도 $\mathbb{P}$를

$$\mathbb{P} : A \mapsto \begin{cases} 1 & \text{when } 0 \in A \\ 0 & \text{ow.} \end{cases}$$

으로 잡아 확률공간 $(\mathbb{R}, \mathscr{B}_1, \mathbb{P})$를 구성하고 rv. $X : \mathbb{R} \to \mathbb{R}$를 항등함수로 두면 명백히 X는 모든 실수를 그 함숫값으로 가지고 심지어 연속이지만, 분포 $\mathbf{P}_X$가 한원소 집합 $\{0\}$을 지지집합으로 가지므로 X는 이산확률벡터이다. 다만, 정의로부터 분포가 가산 지지집합을 가지므로 X가 그 가산개의 값을 제외한 나머지 값을 가질 확률이 0이 되어 '사실상' 치역이 가산이라고 생각할 수는 있다. 하지만 영집합과 공집합이 비슷하지만 완전히 같지는 않은 것처럼 이 경우에도 '사실상' 치역이 가산인 것과 치역이 정말 가산인 것은 구분해야 할 것이다.

비슷하게, X가 연속확률벡터이지만 함수로서 연속이 아닐 수 있다. 애초에 표본공간 Ω에 위상구조가 존재한다는 보장이 없으므로 연속성을 논할 수조차 없다. 그렇다고 표본공간에 위상구조가 적당히 정의되어 있고, 이에 대해 X가 연속이라고 해서 이가 연속확률변수냐 하면 이것 또한 아니다. 앞서 든 예시를 생각해보면 표본공간이 $\mathbb{R}$이고 이 위에 표준위상을 잡더라도 X가 연속이지만 연속확률변수가 아닐 수 있다. 다만, 정의로부터 연속확률벡터와 특이확률벡터는 point mass를 가질 수 없으므로 확률이 표본공간의 어느 한 점에 집중되어 있지 않고 전체에 고르게 퍼져 있으니, 이런 의미에서 '연속' 이라 생각할 수는 있다.

한편, 위의 정의에서 고등학교나 교양 통계학에서는 들어보지 못한 특이확률벡터라는 새로운 종류의 rv.가 등장했다. 다른 두 종류의 rv.는 고등학교 수준에서도 접할 수 있는 반면, 이제서야 특이확률벡터를 도입하는 것에는 그럴만한 이유가 있다. 우선 특이확률벡터는 다분히 이론적인 필요에 의한 rv.로 실생활의 응용에서는 거의 쓸모가 없다. 또한, 이산확률변수나 연속확률변수의 경우 기댓값이나 분산과 같은 개념의 도입과 계산이 쉬운 반면, 특이확률변수의 경우 이에 상당한 이론적 뒷받침이 필요하다. 이런 이유에서 이 책에서도 특이확률변수가 구체적인 예시로 주어지는 것은 이후에 배울 Cantor 분포 하나뿐이다. 그렇다면 이런 단점에도 불구하고 특이확률벡터를 도입해야 할 이론적인 필요가 대체 무엇인가? 다음 정리는 이 질문에 대한 답이자 측도론의 에필로그에서 예고한 이번 절의 클라이막스이다.

Theorem 2.26 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$ 위의 rv. $X : \Omega \to \mathbb{R}^n$에 대해 다음의 조건

i. Rv. $X_{\text{ac}} : \mathbb{R}^n \to \mathbb{R}^n$는 확률공간 $(\mathbb{R}^n, \mathscr{B}_n, \mathbb{P}_{\text{ac}})$에서 정의된 연속확률벡터이다.
ii. Rv. $X_{\text{pp}} : \mathbb{R}^n \to \mathbb{R}^n$는 확률공간 $(\mathbb{R}^n, \mathscr{B}_n, \mathbb{P}_{\text{pp}})$에서 정의된 이산확률벡터이다.
iii. Rv. $X_{\text{sc}} : \mathbb{R}^n \to \mathbb{R}^n$는 확률공간 $(\mathbb{R}^n, \mathscr{B}_n, \mathbb{P}_{\text{sc}})$에서 정의된 특이확률벡터이다.

를 만족하는 적당한 $\mathscr{B}_n$ 위의 확률측도 $\mathbb{P}_{\text{ac}}, \mathbb{P}_{\text{pp}}, \mathbb{P}_{\text{sc}}$와 rv. $X_{\text{ac}}, X_{\text{pp}}, X_{\text{sc}}$가 존재하여 $\alpha+\beta+\gamma=1$인 적당한 $\alpha, \beta, \gamma \geq 0$에 대해 $\mathbf{P}_X = \alpha\mathbf{P}_{X_{\text{ac}}} + \beta\mathbf{P}_{X_{\text{pp}}} + \gamma\mathbf{P}_{X_{\text{sc}}}$이다.

PROOF 정리 2.21로부터 $(\mathbb{R}^n, \mathscr{B}_n, \mathbf{P}_X)$가 확률공간이므로 Lebesgue의 분해정리로부터 $\mathbf{P}_X$는 절대연속성분 $(\mathbf{P}_X)_{\text{ac}}$, 순수 점 성분 $(\mathbf{P}_X)_{\text{pp}}$, 특이연속성분 $(\mathbf{P}_X)_{\text{sc}}$에 대해 $\mathbf{P}_X = (\mathbf{P}_X)_{\text{ac}} + (\mathbf{P}_X)_{\text{pp}} + (\mathbf{P}_X)_{\text{cs}}$와 같이 분해된다. 또한, $\mathbf{P}_X(\mathbb{R}^n) = 1$이므로 $\alpha = (\mathbf{P}_X)_{\text{ac}}(\mathbb{R}^n), \beta = (\mathbf{P}_X)_{\text{pp}}(\mathbb{R}^n), \gamma = (\mathbf{P}_X)_{\text{cs}}(\mathbb{R}^n)$라 하면 α, β, γ는 모두 유한하고 $\alpha + \beta + \gamma = 1$이다. 이제 α, β, γ가 모두 0이 아닌 특별한 경우를 생각해보자. 그렇다면 $\mathbb{P}_1 := (\mathbf{P}_X)_{\text{ac}}/\alpha, \mathbb{P}_2 := (\mathbf{P}_X)_{\text{pp}}/\beta, \mathbb{P}_3 := (\mathbf{P}_X)_{\text{cs}}/\gamma$가 모두 $\mathscr{B}_n$ 위의 확률측도이므로 $(\mathbf{P}_X)_{\text{ac}}, (\mathbf{P}_X)_{\text{pp}}, (\mathbf{P}_X)_{\text{cs}}$의 성질과 정리 2.22로부터 적당한 확률공간 $(\mathbb{R}^n, \mathscr{B}_n, \mathbb{P}_{\text{ac}})$에서 정의된 연속확률벡터 $X_{\text{ac}} : \mathbb{R}^n \to \mathbb{R}^n$, 적당한 확률공간 $(\mathbb{R}^n, \mathscr{B}_n, \mathbb{P}_{\text{pp}})$에서 정의된 이산확률벡터 $X_{\text{pp}} : \mathbb{R}^n \to \mathbb{R}^n$, 적당한 확률공간 $(\mathbb{R}^n, \mathscr{B}_n, \mathbb{P}_{\text{sc}})$에서 정의된 특이확률벡터 $X_{\text{cs}} : \mathbb{R}^n \to \mathbb{R}^n$가 존재하여 $\mathbb{P}_1 = \mathbf{P}_{X_{\text{ac}}}, \mathbb{P}_2 = \mathbf{P}_{X_{\text{pp}}}, \mathbb{P}_3 = \mathbf{P}_{X_{\text{cs}}}$이고, 곧 $\mathbf{P}_X = (\mathbf{P}_X)_{\text{ac}} + (\mathbf{P}_X)_{\text{pp}} + (\mathbf{P}_X)_{\text{cs}} = \alpha\mathbb{P}_1 + \beta\mathbb{P}_2 + \gamma\mathbb{P}_3 = \alpha\mathbf{P}_{X_{\text{ac}}} + \beta\mathbf{P}_{X_{\text{pp}}} + \gamma\mathbf{P}_{X_{\text{cs}}}$이다. 한편, α, β, γ 중에 일부가 0인 경우에 대해서도 이와 비슷하게 하면 된다. □

연속확률벡터, 이산확률벡터, 특이확률벡터의 세 가지 분류가 서로 배타적인 관계인 것은 맞지만 그렇다고 임의의 rv.가 반드시 이 셋 중 하나에 속하는 것은 아니다. 즉, rv. 중에는 연속도, 이산도, singular도 아닌 골치아픈 것들이 존재한다. (이런 rv.를 흔히 **mixed type**이라 부르며 특이확률벡터에 비할 바는 아니지만 그 실용성은 많이 떨어지는 편이다.) 이런 상황에서 위의 정리는 임의의 rv.에 대해 비록 이가 mixed type이더라도 그 분포는 적당한 연속확률벡터, 이산확률벡터, 특이확률벡터의 분포의 합으로 분해할 수 있다는 놀라운 결과를 함의한다. 곧 연속확률벡터, 이산확률벡터, 특이확률벡터의 세 가지 분류는 rv.의 공간의 기저와 비슷한 역할을 하며, 이 세 가지 rv.를 정의한 순간 사실상 모든 rv.의 분류를 끝마친 것과 다름없다.

이러한 이유로 이론 전개에 있어서는 mixed type rv.를 고려할 필요 없이 연속확률벡터, 이산확률벡터, 특이확률벡터의 세 가지 rv.만 생각하면 되고, mixed type rv.는 이 세 종류의 rv.의 성질들을 적당히 섞어 가질 뿐이다. 이러니 고등학교 시절부터 연속확률벡터, 이산확률벡터의 두 가지 종류에만 지대한 관심을 가진 것이 너무나 당연하다. 이론적으로 다루기 힘든 특이확률벡터를 제외하면 이 둘을 다룸으로써 우리는 고등학교때부터 우리도 모르는 사이에 사실상 온갖 종류의 rv.를 모두 다루고 있던 셈이다!

이제 클라이막스의 여운을 뒤로 하고, CDF를 살펴볼 순서이다. 흔히 교양 통계학에서는 PDF를 배운 뒤 CDF를 배우므로 PDF의 개념을 도입하지도 않고 CDF를 정의하는 것이 의아할 수 있다. 하지만, 적어도 이론적으로는 CDF가 PDF보다 더 기본적인 개념이기에 이를 먼저 도입한다.

Definition 2.27 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$에 대해 rv. X의 **(누적) 분포함수 ((cumulative) distribution function)**를 $F_X : \mathbb{R}^n \to \mathbb{R}$로 쓰고 $F_X : x \mapsto \mathbf{P}_X(\prod_{i=1}^n (-\infty, x_i])$로 정의한다. 특별히, $n \geq 2$인 경우 F_X를 rv. $X_1, \cdots, X_n$의 **결합 (누**

적) 분포함수 (joint (cumulative) distribution function)라 하고 $F_{X_1,\cdots,X_n}$ 으로 쓰기도 한다.

CDF의 기본적인 성질은 정리 ? ? 로부터 대부분 자명하게 유도된다.

Theorem 2.28 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$에 대해 다음이 성립한다.

i. $0 \le F_X \le 1$.
ii. 임의의 유계인 semi-open box $B \subseteq \mathbb{R}^n$에 대해 $\Delta_B F_X = \mathbf{P}_X(B) \ge 0$이다.
iii. CDF F_X는 각 변수에 대해 증가한다.
iv. CDF F_X는 오른쪽 연속이다.
v. 각 $i \le n$에 대해 $\lim_{x_i \to -\infty} F_X(x) = 0$이다. [1]
vi. $\lim_{x_1,\cdots,x_n \to \infty} F_X(x) = 1$. [2]
vii. 임의의 $x \in \mathbb{R}^n$에 대해 $B_x = \prod_{i=1}^n (-\infty, x_i]$라 하면 $F_X(x-) = \mathbf{P}_X(B_x^\circ)$이고 $F_X(x) - F_X(x-) = \mathbf{P}_X(\partial B_x)$이다. [3]

Proof i – vi. 이는 CDF의 정의와 정리 ? ? 로부터 자명하다.

vii. 임의의 $x \in \mathbb{R}^n$를 택하여 $B = \prod_{i=1}^n (-\infty, x_i]$라 하고, 집합열 $\{B_j\}$를 $B_j := \prod_{i=1}^n (-\infty, x_i - 1/j]$로 두면 이는 $\mathscr{S}_n$에 속하는 증가하는 집합열로서 $B_j \uparrow \prod_{i=1}^n (-\infty, x_i) = B^\circ$이다. 따라서 $F_X(x - \mathbf{1}/j) = \mathbf{P}_X(B_j) \uparrow \mathbf{P}_X(B^\circ)$이므로 임의의 $\varepsilon > 0$을 택하면 적당한 $j_0 \in \mathbb{N}$가 존재하여 $\mathbf{P}_X(B^\circ) - \mathbf{P}_X(B_{j_0}) < \varepsilon$이다. 이제 $\delta = 1/j_0$라 하면 $||x - y|| < \delta$이고 $x > y$인 모든 $y \in \mathbb{R}^n$에 대해 $B_{j_0} \subseteq \prod_{i=1}^n (-\infty, y_i) \subseteq B^\circ$에서 $\mathbf{P}_X(B^\circ) - \varepsilon < \mathbf{P}_X(B_{j_0}) \le F_X(y) = \mathbf{P}_X(\prod_{i=1}^n (-\infty, y_i]) \le \mathbf{P}_X(B^\circ)$이므로 $|F_X(y) - \mathbf{P}_X(B^\circ)| < \varepsilon$가 되어 $F_X(x-) = \mathbf{P}_X(B_x^\circ)$임을 안다. 이제 $F_X(x) - F_X(x-) = \mathbf{P}_X(B) - \mathbf{P}_X(B^\circ) = \mathbf{P}_X(\partial B)$임은 자명하다. □

위의 정리에서 $n = 1$인 경우 vii는 $F_X(x) - F_X(x-) = \mathbf{P}_X\{x\}$가 되어 이 경우에는 CDF를 분포의 point mass를 구하는 용도로 사용할 수 있다. 한편, 위의 정리의 역 비슷한 정리도 성립한다.

Theorem 2.29 함수 $F : \mathbb{R}^n \to [0, 1]$에 대해 이가

i. 함수 F는 오른쪽 연속이고 각 변수에 대해 증가한다.
ii. 유계인 semi-open box $B \subseteq \mathbb{R}^n$에 대해 $\Delta_B F \ge 0$이다.
iii. 각 $i \le n$에 대해 $\lim_{x_i \to -\infty} F(x) = 0$이고 $\lim_{x_1,\cdots,x_n \to \infty} F_X(x) = 1$이다.

를 만족하면 적당한 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$가 존재하여 $F = F_X$이다. 특별히, 이때 $(\Omega, \mathscr{F}) = (\mathbb{R}^n, \mathscr{B}_n)$이도록 잡을 수 있다.

Proof 정리 ? ? 로부터 적당한 $\mathscr{B}_n$ 위의 측도 μ가 존재하여 임의의 $x \in \mathbb{R}^n$에 대해 $F(x) = \mu(\prod_{i=1}^n (-\infty, x_i])$이다. 그런데 정리 ? ? 의 vii와 주어진 조건으로부터 $\mu(\mathbb{R}^n) = 1$이 되어

μ는 확률측도이고, 곧 정리 2.22로부터 적당한 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$ 위의 n차원 rv. X가 존재하여 $\mu = \mathbf{P}_X$이다. 이상으로부터 임의의 $x \in \mathbb{R}^n$에 대해 $F(x) = \mu(\prod_{i=1}^n (-\infty, x_i]) = \mathbf{P}_X(\prod_{i=1}^n (-\infty, x_i]) = F_X(x)$가 성립한다. 한편, 이때 $(\Omega, \mathscr{F}) = (\mathbb{R}^n, \mathscr{B}_n)$이도록 잡을 수 있음은 정리 2.22로부터 자명하다. □

곧 우리는 CDF가 될 수 있는 함수를 위의 정리의 조건 i – iii으로 완벽히 characterize할 수 있다. 한편, CDF의 연속성에 대한 다음 정리들도 꽤나 흥미롭다.

Theorem 2.30 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$와 한 점 $x_0 \in \mathbb{R}^n$ 대해 $B_{x_0} = \prod_{i=1}^n (-\infty, x_0^i]$라 하면 TFAE.

i. CDF F_X가 x_0에서 연속이다.
ii. $F_X(x_0) = \mathbf{P}_X(B_{x_0}) = \mathbf{P}_X(B_{x_0}^\circ)$.
iii. $\mathbf{P}_X(\partial B_{x_0}) = 0$.

Proof i $\Rightarrow$ ii. 집합열 $\{A_j\}$를 $A_j = \prod_{i=1}^n (-\infty, x_0^i - 1/j]$로 두면 이는 $B_{x_0}^\circ$로 수렴하는 증가하는 집합열이므로 $F_X(x_0 - \mathbf{1}/j) = \mathbf{P}_X(A_j) \uparrow \mathbf{P}_X(B_{x_0}^\circ)$이다. 한편, 가정으로부터 F_X가 x_0에서 연속이므로 $\mathbf{P}_X(A_j) = F_X(x_0 - \mathbf{1}/j) \uparrow F_X(x_0) = \mathbf{P}_X(B_{x_0})$가 되어 $F_X(x_0) = \mathbf{P}_X(B_{x_0}) = \mathbf{P}_X(B_{x_0}^\circ)$임을 안다.

ii $\Rightarrow$ iii. 이는 $\mathbf{P}_X(\partial B_{x_0}) = \mathbf{P}_X(B_{x_0}) - \mathbf{P}_X(B_{x_0}^\circ) = 0$에서 자명하다.

iii $\Rightarrow$ i. 0으로 수렴하는 $\mathbb{R}^n$에 속하는 임의의 수열 $\{h_j\}$를 택하여 각 $j \in \mathbb{N}$에 대해 $a_j = \min_{i=1}^n h_j^i$, $b_j = \max_{i=1}^n h_j^i$라 하면 $F_X(x_0 + a_j\mathbf{1}) \le F_X(x_0 + h_j) \le F_X(x_0 + b_j\mathbf{1})$이고 $a_j, b_j \to 0$이다. 따라서 함수 $f : \mathbb{R} \to [0,1]$를 $f : x \mapsto F_X(x_0 + x\mathbf{1})$로 두고 이가 0에서 연속임을 보이는 것으로 증명은 충분한데, CDF의 성질로부터 $\lim_{x \downarrow 0} f(x) = f(0)$임은 이미 알고 있으므로 $\lim_{x \uparrow 0} f(x) = f(0)$이라는 사실만 보이면 된다. 이를 위해 $x_j \uparrow 0$인 임의의 실수열 $\{x_j\}$를 택하여 집합열 $\{A_j\}$를 $A_j = \prod_{i=1}^n (-\infty, x_0^i + x_j^i]$로 두면 이는 $B_{x_0}^\circ$로 수렴하는 증가하는 집합열이다. 그렇다면 가정으로부터 $f(x_j) = F_X(x_0 + x_j\mathbf{1}) = \mathbf{P}_X(A_j) \uparrow \mathbf{P}_X(B_{x_0}^\circ) = \mathbf{P}_X(B_{x_0}) = F_X(x_0) = f(0)$이고, 증명이 끝난다. □

Lemma 2.31 단조함수 $f : \mathbb{R} \to \mathbb{R}$는 가산개의 불연속점만을 가진다.

Proof 우선 f가 증가함수라 하고 집합 $A = \{x \in \mathbb{R} : f\text{가 } x\text{에서 불연속}\}$를 생각하자. 그렇다면 임의의 $x \in A$에 대해 $f(x-) < f(x+)$이므로 $f(x-) < p_x < f(x+)$인 적당한 $p_x \in \mathbb{Q}$를 택할 수 있고, 이로써 함수 $g : A \to \mathbb{Q}$를 $g : x \mapsto p_x$로 두자. 한편, 임의의 $x, y \in A$에 대해 $x < y$인데 $f(y-) < f(x+)$이면 $z = (x+y)/2$에 대해 $f(z) \le f(y-) < f(x+) \le f(z)$의 모순이 발생하므로 $(f(x-), f(x+))$와 $(f(y-), f(y+))$는 서로소가 되어 g는 단사이고, 곧 A는 가산이다. □

Theorem 2.32 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$에 대해 F_X의 연속점의 집합은 조밀하다. 나아가, $n = 1$인 경우 F_X가 가산개의 불연속점만을 가진다.

PROOF 임의의 $x_0 \in \mathbb{R}^n$를 택하여 함수 $f : \mathbb{R} \to [0,1]$를 $f : h \mapsto F_X(x_0 + h\mathbf{1})$로 두자. 그렇다면 f는 증가함수가 되어 위의 보조정리로부터 가산개의 불연속점만을 가지고, 곧 0으로 수렴하면서 각 점에서 f가 연속인 실수열 $\{h_i\}$를 적당히 택할 수 있다. 이는 각 $i \in \mathbb{N}$에 대해 $\mathbf{P}_X(\prod_{i=1}^n (-\infty, x_0^i + h_i)) = F_X((x_0 + h_i\mathbf{1})-) = f(h_i-) = f(h_i) = F_X(x_0 + h_i\mathbf{1})$임을 함의하므로 정리 2.30로부터 F_X는 $x_0 + h_i\mathbf{1}$에서 연속이고, $x_0 + h_i\mathbf{1} \to x_0$임은 자명하므로 F_X의 연속점의 집합이 조밀함을 안다. 한편, $n = 1$인 경우에는 F_X가 증가함수이므로 다시 위의 보조정리로부터 가산개의 불연속점만을 가짐이 자명하다. □

이제 PDF를 도입하는 것으로 이번 절을 마무리하자. 측도론으로 PDF를 도입하는데 핵심적인 역할을 하는 것은 다름아닌 Radon-Nikodým 도함수의 개념이다.

Definition 2.33 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$에 대해 적당한 $\mathscr{B}_n$ 위의 σ-유한 측도 μ가 존재하여 $\mathbf{P}_X \ll \mu$라 하자. 이때 Radon-Nikodým 도함수 $d\mathbf{P}_X/d\mu$를 X의 μ에 대한 **(확률) 밀도함수 ((probability) density function)**라 하고 f_X로 쓴다. 특별히, $n \geq 2$인 경우 f_X를 rv. $X_1, \cdots, X_n$의 **결합 (확률) 밀도함수 (joint (probability) density function)**라 하고 $f_{X_1,\cdots,X_n}$으로 쓰기도 한다.

이렇게 Radon-Nikodým 도함수로 PDF를 정의하면서 이번에도 우리가 PDF에 대해 당연하게 생각하던 사실들에 미묘한 혼란이 생겨나게 된다. 우선, Radon-Nikodým 도함수가 유일하기는 하지만 그 유일성이 'μ-거의 어디서나 같은 함수를 하나로 볼 때' 성립하는 것이므로 엄밀히는 유일하지 않다. 즉, μ-영집합에서 조금씩 다른 무한히 많은 함수들이 모두 하나의 rv.의 PDF가 될 수 있으므로 일반적으로 특정 점에서의 PDF의 값을 물어보는 것은 의미가 없다. 그러나 Radon-Nikodým 도함수를 추상적인 도구로만 사용했던 측도론에서와 달리 확률론에서는 PDF를 구체적인 함수로 다루어야 할 필요가 있으므로 혼란을 피하기 위해 PDF가 될 수 있는 무한히 많은, μ-거의 어디서나 같은 함수 중 특정한 하나를 PDF의 **version**이라 한다. 물론, 표기상의 편의를 이유로 논의의 대상이 PDF인지, PDF의 version인지를 명시적으로 밝히지 않는 경우가 대부분이지만, L^p 공간에서 동치류로서의 f와 구체적인 함수 f를 논의의 맥락으로 큰 혼란 없이 구분하였듯이 이 둘 또한 구분에 큰 어려움은 없을 것이다.

다음으로, 지금까지는 연속확률벡터와 이산확률벡터에 대해서만 PDF를 생각했지만 위의 정의에서 볼 수 있듯이 임의의 rv.에 대해서도 $\mathbf{P}_X \ll \mu$인 σ-유한 측도 μ만 잘 잡아주면 얼마든지 X의 PDF를 생각할 수 있다. 한편, 연속확률벡터, 이산확률벡터, 특이확률벡터의 세 가지 rv.만 생각해도 충분하다는 사실에 놀라워했던 것이 불과 몇 페이지 전인데 굳이 이렇게 지나칠 정도로 일반적으로 PDF를 도입하는 것이 의아하게 느껴질 수 있다. 그러나

이는 mixed-type rv.의 PDF를 직접 다루기 위함이 아니라 연속확률벡터와 이산확률벡터의 PDF를 동시에 다루기 위함이다. 고등학교나 교양 통계학에서 연속확률벡터와 이산확률벡터의 기댓값이나 분산 같은 성질들을 논할 때, 각각 경우를 나누어 전자는 적분으로, 후자는 합으로 접근했던 것을 기억할 것이다. 우리는 위에서 일반적으로 정의된 PDF와 측도론의 에필로그에서 소개한 셈측도를 이용해 드디어 이런 번거로움에서 벗어날 수 있다.

한편, 정의에서 볼 수 있듯이 $\mathbf{P}_X \ll \mu$ 인 $\sigma-$유한 측도 μ 를 무엇으로 택하는지에 따라 그 PDF가 달라지므로 μ 를 반드시 명시해 주어야 하는데, 우리가 주로 다룰 연속확률벡터와 이산확률벡터의 경우 이에 대한 관례가 있다. 먼저 연속확률벡터 $X : \Omega \to \mathbb{R}^n$ 의 경우 그 정의상 $\mathbf{P}_X \ll \mu_n$ 이 항상 성립하므로 특별한 언급이 없는 이상 연속확률벡터의 PDF는 μ_n 에 대한 PDF로 생각한다. 이산확률벡터 $X : \Omega \to \mathbb{R}^n$ 의 경우는 조금 복잡한데, 우선 정의상 $\mathbf{P}_X$ 는 가산 지지집합 $A \in \mathscr{B}_n$ 를 가진다. 물론, 이때의 가산 지지집합은 유일하지 않지만 임의의 $x \in A$ 에 대해 $\mathbf{P}_X\{x\} > 0$ 이라는 조건을 추가하면 유일하게 주어진다. 그렇다면 $(\mathbf{P}_X, \mathscr{B}_n, \#)$ 에 대해 $\mathbf{P}_X \ll \#_A$ 임이 자명하고, 이때 $\#_A$ 가 $\sigma-$유한임도 자명하므로 특별한 언급이 없는 이상 이산확률벡터의 PDF는 $\#_A$ 에 대한 PDF로 생각한다. 나아가, 이산확률벡터의 경우 PDF의 version으로 특별한 언급이 없는 이상 $\mathbb{R}^n \setminus A$ 에서 0인 version을 택하는 관례도 있다. 이런 관례를 따르면 이때의 PDF가 우리가 기존에 알던 PMF가 된다는 것을 잠시 후에 알게 될 것이다.

이제 PDF의 기본적인 성질들을 보자. 이는 대부분 적분의 성질들로부터 거의 자명하게 얻어진다.

Theorem 2.34 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$ 에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$ 에 대해 적당한 $\mathscr{B}_n$ 위의 $\sigma-$유한 측도 μ 가 존재하여 $\mathbf{P}_X \ll \mu$ 라 하자. 그렇다면 μ 에 대한 X 의 PDF f_X 에 대해 다음이 성립한다.

i. $f_X \geq 0$ ($\mu-$ae.).
ii. $\int_{\mathbb{R}^n} f_X \, d\mu = 1$.
iii. 임의의 $A \in \mathscr{B}_n$ 에 대해 $\int_A f_X \, d\mu = \mathbf{P}_X(A) = \mathbb{P}\{X \in A\}$ 이다.

Proof iii. 이는 $\int_A f_X \, d\mu = \int_A (d\mathbf{P}_X / d\mu) \, d\mu = \int_A d\mathbf{P}_X = \mathbf{P}_X(A)$ 에서 자명하다.

i. 만약 $f_X \geq 0$ ($\mu-$ae.) 가 아니라면 집합 $A = \{x \in \mathbb{R}^n : -f_X(x) > 0\}$ 이 양의 측도를 가지므로 iii과 따름정리 ? ? 의 iii으로부터 $\mathbf{P}_X(A) = \int_A f_X \, d\mu < 0$ 의 모순이 발생한다. 따라서 $f_X \geq 0$ ($\mu-$ae.) 이어야 한다.

ii. iii으로부터 $\int_{\mathbb{R}^n} f_X \, d\mu = \mathbf{P}_X(\mathbb{R}^n) = 1$ 이므로 자명하다. □

특별히, 이산확률벡터의 경우 위의 정리는 다음 따름정리와 같이 합의 형태로 표현된다.

Corollary 2.35 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 이산확률벡터 $X : \Omega \to \mathbb{R}^n$에 대해 $A \in \mathscr{B}_n$를 $\mathbf{P}_X$의 가산 지지집합이라 하고 임의의 $x \in A$에 대해 $\mathbf{P}_X\{x\} > 0$이라 하면 다음이 성립한다.

i. 임의의 $x \in \mathbb{R}^n$에 대해 $f_X(x) = \mathbf{P}_X\{x\}$이고, 따라서 $0 \le f_X \le 1$이다.
ii. $\sum_{x\in\mathbb{R}^n} f_X(x) = 1$.
iii. 임의의 $B \in \mathscr{B}_n$에 대해 $\sum_{x\in B} f_X(x) = \mathbf{P}_X(B) = \mathbb{P}\{X \in B\}$이다.

Proof 표기의 편의를 위해 $\mathscr{A} = \mathscr{B}_n|_A$라 하면 $\#_A|_{\mathscr{A}} = \#|_{\mathscr{A}}$는 가측공간 $(A, \mathscr{A})$ 위의 셈측도이다.

i. 만약 $x \in A$이면 정리 ? ? 의 iii으로부터 $\mathbf{P}_X\{x\} = \int_{\{x\}} f_X \, d\#_A = \int_{\{x\}} f_X|_A \, d\#|_{\mathscr{A}} = f_X(x)$이고 $x \notin A$이면 f_X의 version을 택하는 관례로부터 $\mathbf{P}_X\{x\} = 0 = f_X(x)$이다.

ii. 위의 정리와 셈측도의 성질 그리고 f_X의 version을 택하는 관례와 정리 ? ? 의 iii으로부터 $1 = \mathbf{P}_X(A) = \int_A f_X \, d\#_A = \int_A f_X|_A \, d\#|_{\mathscr{A}} = \sum_{x\in A} f_X(x) = \sum_{x\in\mathbb{R}^n} f_X(x)$이다.

iii. 위의 정리와 셈측도의 성질 그리고 f_X의 version을 택하는 관례와 정리 ? ? 의 iii으로부터 $\mathbf{P}_X(B) = \mathbf{P}_X(A \cap B) = \int_{A\cap B} f_X \, d\#_A = \int_{A\cap B} f_X|_A \, d\#|_{\mathscr{A}} = \sum_{x\in A\cap B} f_X(x) = \sum_{x\in B} f_X(x)$이다. □

다음 정리는 연속확률벡터의 CDF와 PDF의 관계에 대한 정리로서 연속확률벡터의 CDF와 PDF를 계산하는 데 유용하게 사용된다.

Theorem 2.36 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 연속확률벡터 $X : \Omega \to \mathbb{R}^n$에 대해 다음이 성립한다.

i. $f_X(x) = F_X^{(s)}(x)$.
ii. $F_X(x) = \int_{\prod_{i=1}^n (-\infty, x_i]} f_X \, d\mu_n$.

특별히, $x_0 \in \mathbb{R}^n$의 근방에서 $\partial^n F_X / \partial x_1 \cdots \partial x_n$이 존재하고 이가 x_0에서 연속이면 $f_X(x_0) = (\partial^n F_X / \partial x_1 \cdots \partial x_n)(x_0)$이다. 이때, 편미분의 순서는 중요하지 않다.

Proof 이는 따름정리 ? ? 와 정리 2.34의 iii으로부터 자명하다 □

비슷하게, 다음 정리는 어떤 rv.가 연속확률벡터인지를 판단할 때 유용하다.

Theorem 2.37 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$에 대해 TFAE.

i. Rv. X는 연속확률벡터이다.
ii. CDF F_X가 절대연속이다.
iii. 적분가능한 Borel 함수 $f : \mathbb{R}^n \to \overline{\mathbb{R}}$가 존재하여 $F_X(x) = \int_{\prod_{i=1}^n (-\infty, x_i]} f \, d\mu_n$이다.

PROOF i $\Leftrightarrow$ ii. 이는 정리 ? ? 로부터 자명하다.

i $\Rightarrow$ iii. 이는 $f = f_X$ 로 두고 정리 2. 36의 ii를 생각하면 자명하다.

iii $\Rightarrow$ ii. 함수 f가 적분가능하므로 거의 어디서나 유한하고, 따라서 WLOG, 필요다하면 영집합에서의 f의 값을 0으로 바꾸어 $|f| < \infty$라 하자. 이제 집합열 $\{A_j\}$를 $A_j = |f|^{-1}((j,\infty))$로 두면 이는 $\emptyset$으로 수렴하는 감소하는 집합열이 되어 $\int_{A_j} |f|\,d\lambda_n \to \int_{\emptyset} |f|\,d\lambda_n = 0$이다. 이로부터 임의의 $\varepsilon > 0$을 택하면 적당한 $j_0 \in \mathbb{N}$가 존재하여 $\int_{A_{j_0}} |f|\,d\lambda_n < \varepsilon/2$이고, $\mu_n(A) < \varepsilon/2j_0 =: \delta$인 임의의 $A \in \mathscr{B}_n$에 대해 $\int_A |f|\,d\mu_n \le \int_{A\setminus A_{j_0}} |f|\,d\mu_n + \int_{A_{j_0}} |f|\,d\mu_n < \mu_n(A\setminus A_{j_0})j_0 + \varepsilon/2 < \varepsilon$이다. 따라서 임의의 유계이고 서로소인 $B_1, \cdots, B_l \in \mathscr{S}_n$에 대해 $\sum_{k=1}^l \mu_n(B_k) < \delta$이면 $\sum_{k=1}^l \Delta_{B_k} F_X = \sum_{k=1}^l \int_{B_k} f\,d\mu_n = \int_{\bigsqcup_{k=1}^l B_k} f\,d\mu < \varepsilon$이 되어 F_X가 절대연속임을 안다. □

2.3 Expectation

이번 절에서는 rv.의 대표적인 통계량 중 하나인 기댓값과 분산을 엄밀하게 도입하고, 이와 관련된 부등식을 증명하는 것을 목표로 한다. 통계학의 큰 연구주제 중 하나는 '정보의 단순화' 라 할 것이다. 당연히, 이때 단순화의 정도와 이에 따른 정보의 손실은 서로 trade-off의 관계에 있어서 단순화를 많이 하면 할수록 정보의 전달이 용이하지만 그만큼 손실되는 정보도 많아지고, 역으로 손실되는 정보의 양을 줄이면 줄일수록 단순화의 정도가 감소하여 전달이 어려워진다. 이런 상황에서 우리는 정보의 손실을 최소한으로 유지하면서도 정보를 단순하게 전달할 수 있는 좋은 방법을 찾고자 하는데, 이번 절에서 알아볼 기댓값은 어떤 분포의 정보를 단 하나의 통계량으로 단순화하여 전달할 수 있는 좋은 방법 중 하나이다. (이와 같이 분포의 정보를 요약하는 통계량을 그 분포의 대푯값이라 하며 기댓값 외에도 중앙값, 최빈값 등이 있다. 각 대푯값은 나름의 장단점이 있기에 어떤 상황에서 무엇을 써야 하는지를 잘 판단하여 사용해야 한다.)

기댓값은 확률변수가 가질 것으로 가장 기대할 수 있는 값이라 해석할 수 있다. 기댓값은 분포의 여러 대푯값 중에서 가장 널리 이용되는 통계량으로 추정이나 관련 가설의 검정 등에 있어 많은 장점을 지니지만, outlier와 같은 극단치에 민감하여 때로는 분포에 대해 다소 오인의 가능성이 있는 정보를 전달할 수도 있다는 단점을 가진다. 이러한 기댓값은 기본적으로 확률변수의 적분으로 정의된다.

Definition 2.38 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}$에 대해 만약 X가 적분가능하다면 $\int_\Omega X\,d\mathbb{P}$를 X의 **기댓값 (expectation)** 혹은 **평균 (mean)**이라 하고 $\mathbf{E}(X), \mu_X$ 혹은 간단히 μ로 쓴다.

기댓값이 정의되기 위해서는 확률공간과 그 위에서 정의된 rv.만 있으면 충분하다는 점에 주목하기 바란다. 고등학교나 교양 통계학에서는 기댓값이 정의되기 위해서는 rv.가 반드시 연속이거나 이산이어서 PDF나 PMF가 있어야 했다. 그러나 위의 정의는 일반적인 mixed-type rv.에 대해서도 기댓값을 정의할 수 있도록 해준다. 한편, 정의에서 볼 수 있듯이 기댓값은 확률측도의 적분에 불과하므로 이에 관한 기본적인 성질들은 대부분 적분의 성질들로부터 자명하게 유도된다.

Proposition 2.39 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}$에 대해 $\mathbf{E}(X)$가 존재할 필요충분조건은 $\mathbf{E}(|X|)$가 존재하는 것이다.

Proof 이는 X가 적분가능할 필요충분조건이 $|X|$가 적분가능할 것이라는 점에서 자명하다. □

Theorem 2.40 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X, Y : \Omega \to \mathbb{R}$에 대해 $\mathbf{E}(X), \mathbf{E}(Y)$가 모두 존재한다면 다음이 성립한다.

i. (선형성) 임의의 $a, b \in \mathbb{R}$에 대해 $\mathbf{E}(aX+bY)$가 존재하고 $\mathbf{E}(aX+bY) = a\mathbf{E}(X)+b\mathbf{E}(Y)$이다.
ii. 만약 $X = Y$ (as.)라면 $\mathbf{E}(X) = \mathbf{E}(Y)$이다.
iii. 만약 $X \le Y$ (as.)라면 $\mathbf{E}(X) \le \mathbf{E}(Y)$이다.
iv. $\mathbf{E}(1) = 1$.
v. $\inf_{\omega\in\Omega} X(\omega) \le \mathbf{E}(X) \le \sup_{\omega\in\Omega} X(\omega)$.
vi. $|\mathbf{E}(X)| \le \mathbf{E}(|X|)$.

Proof i – iii, vi. 이는 적분의 성질로부터 자명하다.
iv. 이는 $\mathbf{E}(1) = \int_\Omega d\mathbb{P} = \mathbb{P}(\Omega) = 1$에서 분명하다.
v. 이는 iii과 iv로부터 자명하다. □

기댓값의 선형성으로부터 $X - \mu_X$의 기댓값은 항상 0이다. 이렇게 원래의 rv.에서 그 기댓값을 빼는 것을 **centering**이라 하며, 기댓값이 서로다른 분포를 비교할 때에 자주 사용되는 방법이다. 한편, 다양한 수렴정리들도 동일하게 성립한다.

Theorem 2.41 (Monotone convergence theorem) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 음이 아닌 rv. $X_i : \Omega \to \mathbb{R}_0^+$의 열 $\{X_i\}$와 음이 아닌 rv. $X : \Omega \to \mathbb{R}_0^+$에 대해 $X_i \uparrow X$ (as.)이고 $\mathbf{E}(X)$가 존재한다고 하면 모든 $i \in \mathbb{N}$에 대해 $\mathbf{E}(X_i)$도 존재하고 $\mathbf{E}(X_i) \uparrow \mathbf{E}(X)$이다.

Proof 모든 $i \in \mathbb{N}$에 대해 $0 \le X_i \le X$이므로 $\mathbf{E}(X_i)$가 존재한다. 이제 정리는 측도론에서 배운 MCT에서 자명하다. □

Theorem 2.42 (Lebesgue's dominated convergence theorem) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X_i : \Omega \to \mathbb{R}$의 열 $\{X_i\}$와 rv. $X : \Omega \to \mathbb{R}$에 대해 $X_i \to X$ (as.)라 하자. 또한 어떤 rv. $Y : \Omega \to \mathbb{R}_0^+$가 존재하여 $\mathbf{E}(Y)$가 존재하고 모든 $i \in \mathbb{N}$에 대해 $|X_i| \leq Y$이면 $\mathbf{E}(X_i)$와 $\mathbf{E}(X)$가 모두 존재하고 $\mathbf{E}(X_i) \to \mathbf{E}(X)$이다.

Proof 이는 측도론에서 배운 DCT에서 자명하다. □

Corollary 2.43 (Bounded convergence theorem) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X_i : \Omega \to \mathbb{R}$의 열 $\{X_i\}$와 rv. $X : \Omega \to \mathbb{R}$에 대해 $X_i \to X$ (as.)라 하자. 또한 $\{X_i\}$가 균등하게 유계라 하자. 즉, 어떤 $M > 0$이 존재하여 모든 $i \in \mathbb{N}$에 대해 $|X_i| \leq M$이라 하자. 그렇다면 $\mathbf{E}(X_i)$와 $\mathbf{E}(X)$가 모두 존재하고 $\mathbf{E}(X_i) \to \mathbf{E}(X)$이다.

Proof 이는 측도론에서 배운 유계수렴정리로부터 자명하다. □

비록 기댓값의 정의는 충분히 일반적이지만, 이가 기댓값을 계산하는 구체적인 방법을 알려주지는 않는다. 결국 기댓값을 계산하기 위해서는 우리가 여태껏 해왔던 것처럼 PDF의 도움을 받아야 한다.

Theorem 2.44 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$와 Borel 함수 $g : \mathbb{R}^n \to \mathbb{R}$에 대해 $\mathbf{E}(g(X))$가 존재한다면 (혹은 g가 $\mathbf{P}_X$−적분가능하다면) $\mathbf{E}(g(X)) = \int_{\mathbb{R}^n} g \, d\mathbf{P}_X$이고, 이때 g는 $\mathbf{P}_X$−적분가능하다 (혹은 $\mathbf{E}(g(X))$가 존재한다).

Proof 기댓값 $\mathbf{E}(g(X))$가 존재한다면 $\mathbf{E}(g(X)) = \int_\Omega g \circ X \, d\mathbb{P} = \int_{\mathbb{R}^n} g \, d\mathbf{P}_X < \infty$이므로 이는 자명하다. 한편, $\mathbf{P}_X$−적분가능한 g에 대해서도 비슷하게 하면 된다. □

Theorem 2.45 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$와 Borel 함수 $g : \mathbb{R}^n \to \mathbb{R}$에 대해 적당한 $\mathscr{B}_n$ 위의 σ−유한 측도 μ가 존재하여 $\mathbf{P}_X \ll \mu$라 하자. 만약 $\mathbf{E}(g(X))$가 존재한다면 (혹은 $f_X g$는 μ−적분가능하다면) $\mathbf{E}(g(X)) = \int_{\mathbb{R}^n} f_X g \, d\mu$이고, 이때 $f_X g$는 μ−적분가능하다 (혹은 $\mathbf{E}(g(X))$가 존재한다).

Proof 기댓값 $\mathbf{E}(g(X))$가 존재한다면 $\mathbf{E}(g(X)) = \int_{\mathbb{R}^n} g \, d\mathbf{P}_X = \int_{\mathbb{R}^n} f_X g \, d\mu < \infty$이므로 이는 자명하다. 한편, $f_X g$가 μ−적분가능한 경우에 대해서도 비슷하게 하면 된다. □

특별히, 이산확률벡터의 경우 위의 정리는 다음 따름정리와 같이 합의 형태로 표현된다.

Corollary 2.46 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 이산확률벡터 $X : \Omega \to \mathbb{R}^n$와 Borel 함수 $g : \mathbb{R}^n \to \mathbb{R}$에 대해 $\mathbf{E}(g(X))$가 존재한다면 $\mathbf{E}(g(X)) = \sum_{x \in \mathbb{R}^n} f_X(x) g(x)$이다.

Proof 가정으로부터 $\mathbf{P}_X$의 가산 지지집합 $A \in \mathscr{B}_n$가 존재하여 WLOG, 필요하다면 몇몇 원소를 제거하여 임의의 $x \in A$에 대해 $\mathbf{P}_X\{x\} > 0$이라 해도 된다. 표기의 편의를 위해

$\mathscr{A} = \mathscr{B}_n|_A$ 라 하면 $\#_A|_{\mathscr{A}} = \#|_{\mathscr{A}}$ 는 가측공간 $(A, \mathscr{A})$ 위의 셈측도이므로 위의 정리와 f_X 의 version을 택하는 관례 그리고 정리 ? ? 의 iii으로부터 $\mathbf{E}(g(X)) = \int_{\mathbb{R}^n} f_X g \, d\# = \int_A f_X g \, d\#_A = \int_A (f_X g)|_A \, d\#|_{\mathscr{A}} = \sum_{x\in A} f_X(x)g(x) = \sum_{x\in\mathbb{R}^n} f_X(x)g(x)$ 이다. □

다음으로 넘어가기 전에, 기댓값에 대해 한 가지 유념해야 할 점이 있다. 기댓값의 정의에는 이가 마치 어떤 확률변수의 특성인 것처럼 쓰여 있지만 사실 기댓값은 확률변수의 특성이 아닌, 분포의 특성이어서 서로 다른 확률공간에서 정의된 서로다른 확률변수라도 그 분포가 같다면 기댓값도 같다. 실제로 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$ 에서 정의된 rv. $X : \Omega \to \mathbb{R}^n$ 와 확률공간 $(\Omega', \mathscr{F}', \mathbb{P}')$ 에서 정의된 rv. $Y : \Omega' \to \mathbb{R}^n$ 에 대해 만약 $\mathbf{P}_X = \mathbf{P}_Y$ 라면 정리 2. 44 로부터 $\mathbf{E}(X) = \int_{\mathbb{R}^n} x \, d\mathbf{P}_X(x) = \mathbf{E}(Y)$ 이다. 이러한 사실은 기댓값이 분포의 정보를 요약하여 전달하는 통계량이라는 점에서 어떻게 보면 당연한 것이다.

기댓값에 이어 살펴볼 통계량은 분산이다. 기댓값이 분포의 정보를 요약하여 전달하는 좋은 통계량인 것은 사실이지만, 이가 분포의 정보를 완벽하게 전달하지는 못한다. 이때 발생하는 정보의 손실이 크게 중요하지 않은 경우도 있지만, 때로는 기댓값만으로는 부족한 경우도 있다. 이에 기댓값만을 전달할 때 손실되는 정보를 보충할 목적으로 보조적인 통계량을 제공할 수 있는데, 보통 분포가 기댓값으로부터 퍼져있는 정도를 의미하는 분산이나 표준편차를 제공한다. (이와 같이 분포가 퍼져있는 정도를 의미하는 통계량을 그 분포의 산포도라 하며 분산과 표준편차 이외에도 평균절대편차 (MAD; Mean Absolute Deviation), 사분위수범위 (IQR; InterQuartile Range) 등이 있다.) 이러한 분산은 기본적으로 기댓값으로 정의된다.

Definition 2.47 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$ 에서 정의된 rv. $X : \Omega \to \mathbb{R}$ 에 대해 만약 $\mathbf{E}((X - \mu_X)^2)$ 이 존재한다면 이를 X 의 **분산 (variance)** 이라 하고 $\mathbf{Var}(X)$, σ_X^2 혹은 간단히 σ^2 으로 쓴다. 나아가, $\mathbf{Var}(X)$ 가 존재한다면 $\sqrt{\mathbf{Var}(X)}$ 를 X 의 **표준편차 (standard deviation)** 라 하고 $\mathbf{Sd}(X)$, σ_X 혹은 간단히 σ 로 쓴다.

Proposition 2.48 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$ 에서 정의된 rv. $X : \Omega \to \mathbb{R}$ 에 대해 $\mathbf{Var}(X)$ 가 존재할 필요충분조건은 $\mathbf{E}(X^2)$ 이 존재하는 것이다.

Proof 만약 $\mathbf{Var}(X)$ 가 존재한다면 정의로부터 $\mathbf{E}(X)$ 와 $\mathbf{E}((X - \mu_X)^2)$ 가 존재하고, 곧 $X^2 = (X - \mu_X)^2 + 2\mu_X X - \mu_X^2$ 에서 $\mathbf{E}(X^2)$ 이 존재함을 안다. 역으로 $\mathbf{E}(X^2)$ 이 존재한다면 $|X| \le X^2 \mathbf{1}_{\{|X|\ge 1\}} + \mathbf{1}_{\{|X|<1\}} \le X^2 + 1$ 에서 $\mathbf{E}(X)$ 가 존재함을 알고, 곧 $(X - \mu_X)^2 = X^2 - 2\mu_X X + \mu_X^2$ 에서 $\mathbf{Var}(X) = \mathbf{E}((X - \mu_X)^2)$ 이 존재한다. □

분산의 기본적인 성질들은 모두 기댓값의 성질들에 기인하다. 특히 다음 정리의 i는 간단한 공식이지만 분산을 실제로 계산할 때 유용하게 사용되는 공식이다.

Theorem 2.49 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$ 에서 정의된 rv. $X : \Omega \to \mathbb{R}$ 에 대해 $\mathbf{Var}(X)$ 가 존재한다면 다음이 성립한다.

i. $\mathbf{Var}(X) = \mathbf{E}(X^2) - \mu_X^2$.

ii. 임의의 $a, b \in \mathbb{R}$에 대해 $\mathbf{Var}(aX+b)$가 존재하고 $\mathbf{Var}(aX+b) = a^2\mathbf{Var}(X)$이다.

PROOF i. 이는

$$\begin{aligned}\mathbf{Var}(X) &= \mathbf{E}((X-\mu_X)^2) \\ &= \mathbf{E}(X^2 - 2\mu_X X + \mu_X^2) \\ &= \mathbf{E}(X^2) - 2\mu_X\mathbf{E}(X) + \mu_X^2 \\ &= \mathbf{E}(X^2) - \mu_X^2\end{aligned}$$

에서 자명하다.

ii. 우선 $\mathbf{Var}(X)$가 존재하므로 $\mathbf{E}(X)$와 $\mathbf{E}(X^2)$이 존재하여 $\mathbf{E}((aX+b)^2) = \mathbf{E}(a^2X^2 + 2abX + b^2)$가 존재하여 $\mathbf{Var}(aX+b)$도 존재한다. 이제

$$\begin{aligned}\mathbf{Var}(aX+b) &= \mathbf{E}([aX+b-\mathbf{E}(aX+b)]^2) \\ &= \mathbf{E}(a^2[X-\mathbf{E}(X)]^2) \\ &= a^2\mathbf{E}((X-\mu_X)^2) \\ &= a^2\mathbf{Var}(X)\end{aligned}$$

에서 정리는 자명하다. □

이제 기댓값에 관한 다양한 부등식을 알아보는 것으로 이번 절을 마무리하자. 이러한 부등식은 이후 확률론의 다양한 정리를 증명할 때에 약방의 감초와 같이 사용될 것이다.

Theorem 2.50 (Jensen's inequality) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X, Y : \Omega \to \mathbb{R}$와 함수 $g : \mathbb{R} \to \mathbb{R}$에 대해 적당한 구간 $I \subseteq \mathbb{R}$가 존재하여 g가 I에서 볼록하고 (혹은 오목하고) $X \in I$ (as.) 라 하자. 만약 $\mathbf{E}(X)$와 $\mathbf{E}(g(X))$가 존재한다면 $g(\mathbf{E}(X)) \le \mathbf{E}(g(X))$이다 (혹은 $g(\mathbf{E}(X)) \ge \mathbf{E}(g(X))$이다).

PROOF 먼저 I가 열린구간인 경우를 생각하여 $I = (a, b)$로 두면 가정과 따름정리 ? ? 의 iii으로부터 $\mathbf{E}(X) \in I$이다. 한편, g가 I에서 볼록하므로 $(\mathbf{E}(X), g(\mathbf{E}(X)))$를 지나는 supporting line $l : y = mx + n$이 존재하여 $mX + n \le g(X)$ (as.)이고, 따라서 $g(\mathbf{E}(X)) = m\mathbf{E}(X) + n = \mathbf{E}(mX+n) \le \mathbf{E}(g(X))$이다.

이제 I가 닫힌구간인 경우를 생각하여 $I = [a, b]$로 두면 가정으로부터 $\mathbf{E}(X) \in I$인데, 만약 $\mathbf{E}(X) \in I^\circ$라면 이전과 같이 하여 $g(\mathbf{E}(X)) \le \mathbf{E}(g(X))$임을 보일 수 있으므로 $\mathbf{E}(X)$가 I의 양 끝점인 경우만 고려하면 된다. 간결한 논의를 위해, $\mathbf{E}(X) = b$라 하자. (반대로 $\mathbf{E}(X) = a$인 경우에도 이와 비슷하게 하면 된다.) 그렇다면 정리 ? ? 의 ii로부터 $X = b$ (as.)이므로 $g(X) = g(b)$ (as.)에서 $g(\mathbf{E}(X)) = g(b) = \mathbf{E}(g(X))$이고, 증명이 끝난다.

구간 I가 한쪽 끝점만 포함하는 경우에도 I가 닫힌구간인 경우와 비슷하게 하면 같은 결론을 얻고, g가 I에서 오목한 경우에는 $-g$를 생각하면 되므로 증명은 이로써 충분하다. □

Theorem 2.51 (Hölder's inequality) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X, Y : \Omega \to \mathbb{R}$와 $1/p + 1/q = 1$인 $p, q > 1$에 대해 $\mathbf{E}(|X|^p)$와 $\mathbf{E}(|Y|^q)$가 존재한다면 $\mathbf{E}(|XY|)$도 존재하고 $\mathbf{E}(|XY|) \leq [\mathbf{E}(|X|^p)]^{1/p}[\mathbf{E}(|Y|^q)]^{1/q}$이다.

Proof 이는 측도론에서 배운 Hölder의 부등식으로부터 자명하다. □

Theorem 2.52 (Minkowski's inequality) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X, Y : \Omega \to \mathbb{R}$와 $p \geq 1$에 대해 $\mathbf{E}(|X|^p)$와 $\mathbf{E}(|Y|^p)$가 존재한다면 $\mathbf{E}(|X+Y|^p)$도 존재하고 $[\mathbf{E}(|X+Y|^p)]^{1/p} \leq [\mathbf{E}(|X|^p)]^{1/p} + [\mathbf{E}(|Y|^p)]^{1/p}$이다.

Proof 이는 측도론에서 배운 Minkowski의 부등식으로부터 자명하다. □

Corollary 2.53 (Cauchy-Schwarz's inequality) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X, Y : \Omega \to \mathbb{R}$에 대해 $\mathbf{E}(X^2)$과 $\mathbf{E}(Y^2)$이 존재한다면 $\mathbf{E}(|XY|)$도 존재하고 $\mathbf{E}(|XY|) \leq \sqrt{\mathbf{E}(X^2)\mathbf{E}(Y^2)}$이다.

Proof 이는 위의 Hölder의 부등식에서 $p = q = 2$인 특수한 경우이다. □

Theorem 2.54 (Liapounov's inequality) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}$와 $0 < p < q$인 $p, q \in \mathbb{R}$에 대해 $\mathbf{E}(|X|^q)$가 존재한다면 $\mathbf{E}(|X|^p)$도 존재하고 $[\mathbf{E}(|X|^p)]^{1/p} \leq [\mathbf{E}(|X|^q)]^{1/q}$이다.

Proof 우선 $|X|^p \leq |X|^q \mathbf{1}_{\{|X|\geq 1\}} + \mathbf{1}_{\{|X|<1\}} \leq |X|^q + 1$에서 $\mathbf{E}(|X|^p)$이 존재함을 안다. 이제 함수 $g : \mathbb{R} \to \mathbb{R}$를 $g : x \mapsto x^{q/p}\mathbf{1}_{\mathbb{R}^+}(x)$로 두면 이는 볼록함수이므로 Jensen의 부등식으로부터 $[\mathbf{E}(|X|^p)]^{q/p} = g(\mathbf{E}(|X|^p)) \leq \mathbf{E}(g(|X|^p)) = \mathbf{E}(|X|^q)$이고, 곧 증명이 끝난다. □

아래의 두 부등식은 증명의 도구로 쓰일 뿐만 아니라 어떤 확률분포가 특정 값을 얼마 이상 벗어날 확률에 대한 상계로서의 가치를 지닌다. 이를 두고 아래의 두 부등식이 'tail probability를 bounding한다' 고 한다.

Theorem 2.55 (Markov's inequality) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}$와 $p > 0$에 대해 $\mathbf{E}(|X|^p)$이 존재한다면 임의의 $x > 0$에 대해 $\mathbb{P}\{|X| \geq x\} \leq \mathbf{E}(|X|^p)/x^p$이다.

Proof 이는 측도론에서 배운 Markov의 부등식으로부터 자명하다. □

Corollary 2.56 (Chebyshev's inequality) 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X : \Omega \to \mathbb{R}$에 대해 $\mathbf{Var}(X)$가 존재한다면 임의의 $x > 0$에 대해 $\mathbb{P}\{|X - \mu_X| \geq x\} \leq \mathbf{Var}(X)/x^2$이다.

PROOF 이는 rv. $X-\mu_X$에 $p=2$인 경우의 Markov의 부등식을 적용한 결과이다. □

다음은 Chebyshev의 부등식의 기초적인 응용으로, 분산이 0이라는 것의 의미를 잘 나태나준다.

Proposition 2.57 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X:\Omega\to\mathbb{R}$에 대해 $\mathbf{Var}(X)=0$이면 $\mathbf{E}(X)$가 존재하고 $X=\mathbf{E}(X)$ (as.)이다.

PROOF 우선 $\mathbf{E}(X)$가 존재함은 자명하다. 이제 사건열 $\{E_i\}$를 $E_i=\{|X-\mu_X|>1/i\}$로 두면 이는 $\{X\neq\mu_X\}$로 수렴하는 사건열로 $\mathbb{P}\{X=\mu_X\}=1-\mathbb{P}\{X\neq\mu_X\}=1-\lim_{i\to\infty}\mathbb{P}(E_i)$이다. 한편, 각 $i\in\mathbb{N}$에 대해 Chebyshev의 부등식으로부터 $\mathbb{P}(E_i)\leq\mathbf{Var}(X)/i^2=0$이므로 $\mathbb{P}(E_i)\to 0$에서 $\mathbb{P}\{X=\mu_X\}=1$임을 안다. □

2.4 Moments

Definition 2.58 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X:\Omega\to\mathbb{R}$와 $n\in\mathbb{N}$에 대해 다음을 정의한다.

i. 만약 $\mathbf{E}(X^n)$이 존재한다면 이를 X의 **n차 적률 (nth moment)**이라 하고 $\mu_{n,X}$ 혹은 간단히 μ_n으로 쓴다.
ii. 만약 $\mathbf{E}((X-\mu_X)^n)$이 존재한다면 이를 X의 **n차 중심화된 적률 (nth centeral moment)**이라 하고 $\tau_{n,X}$ 혹은 간단히 τ_n으로 쓴다.
iii. 만약 $\mathbf{E}([(X-\mu_X)/\sigma_X]^n)$이 존재한다면 이를 X의 **n차 표준화된 적률 (nth standardized moment)**이라 하고 $\kappa_{n,X}$ 혹은 간단히 κ_n으로 쓴다.
iv. 만약 $\mathbf{E}(X^{\underline{n}})$이 존재한다면 이를 X의 **n차 (하향) 계승적률 (nth (falling) factorial moment)**이라 하고 $\mu_{\underline{n},X}$ 혹은 간단히 $\mu_{\underline{n}}$으로 쓴다.
v. 만약 $\mathbf{E}(X^{\overline{n}})$이 존재한다면 이를 X의 **n차 (상향) 계승적률 (nth (rising) factorial moment)**이라 하고 $\mu_{\overline{n},X}$ 혹은 간단히 $\mu_{\overline{n}}$으로 쓴다.

Definition 2.59 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X:\Omega\to\mathbb{R}$에 대해 만약 X의 3차 표준화된 적률이 존재한다면 이를 특별히 X의 **왜도 (skewness)**라 하고 $\mathbf{Skew}(X)$, τ_X 혹은 간단히 τ로 쓴다. 비슷하게, X의 4차 표준화된 적률이 존재한다면 $\kappa_{4,X}-3$을 특별히 X의 **첨도 (kurtosis)**라 하고 $\mathbf{Kurt}(X)$, κ_X 혹은 간단히 κ로 쓴다.

Theorem 2.60 확률공간 $(\Omega, \mathscr{F}, \mathbb{P})$에서 정의된 rv. $X:\Omega\to\mathbb{R}$에 대해 $\mathbf{Var}(X)$가 존재한다면 다음이 성립한다.

i. $\mathbf{Var}(X)=\mathbf{E}(X^2)-\mu_X^2$.

ii. 임의의 $a, b \in \mathbb{R}$에 대해 $a \neq 0$이면 $\mathbf{Skew}(aX+b)$가 존재하고 $\mathbf{Skew}(aX+b) = \mathbf{Skew}(X)/a^3$이다.

iii. 임의의 $a, b \in \mathbb{R}$에 대해 $a \neq 0$이면 $\mathbf{Kurt}(aX+b)$가 존재하고 $\mathbf{Kurt}(aX+b) = \mathbf{Kurt}(X)/a^4$이다.

Notes

1 여기서 $\lim_{x_i \to -\infty} F_X(x) = 0$은 임의의 $\cdots, x_{i-1}, x_{i+1}, \cdots \in \mathbb{R}$와 임의의 $\varepsilon > 0$에 대해 적당한 $M \in \mathbb{R}$이 존재하여 $x_i < M$인 임의의 $x_i \in \mathbb{R}$에 대해 $F_X(x) < \varepsilon$이라는 의미이다.

2 여기서 $\lim_{x_1, \cdots, x_n \to \infty} F_X(x) = 1$은 임의의 $\varepsilon > 0$에 대해 적당한 $M \in \mathbb{R}$이 존재하여 $x \geq M\mathbf{1}$인 임의의 $x \in \mathbb{R}^n$에 대해 $|F_\mu(x) - 1| < \varepsilon$이라는 의미이다.

3 집합 $A \subseteq \mathbb{R}^m$에서 정의된 함수 $f : A \to \mathbb{R}^n$와 한 점 $x_0 \in \mathrm{acc}A$에 대해 극한 $\lim_{x \uparrow x_0} f(x)$가 존재하는 경우 그 값을 $f(x_0-)$로 쓴다. 한편, 여기서 $\lim_{x \uparrow x_0} f(x) = f(x_0-)$는 임의의 $\varepsilon > 0$에 대해 적당한 $\delta > 0$가 존재하여 $||x - x_0|| < \delta$이고 $x < x_0$인 임의의 $x \in A$에 대해 $||f(x) - f(x_0-)|| < \varepsilon$이라는 의미이다. 이제 $f(x_0+)$도 비슷하게 정의된다.

4