Preface

확률론은 통계학을 공부하는 데 있어 굉장히 중요한 과목이다. 그러므로 열심히 공부해야 한다.

덤으로 극단값 이론의 기초도 수록하였다.

최대한 제가 이해할 수 있는 수준의 내용으로 구성하였으므로, 그러므로 기초 레벨에 해당이 된다.

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

Part I

Intro

1 Introduction

1.1 Probability Theory

- Probability models: random experiment를 묘사하는데 목적이 있음
- Random experiment: 무작위성이 있어 미래에 일어날 결과물을 정확하게 예측할 수 없는 실험
- **Probability space**: 확률론의 기초가 됨, 확률공간의 키가 되는 아이디어는 **stabilization of the relative frequencies**임

우리가 random experiement를 **독립**적으로, 반복적으로 수행한다고 하고 어떤 특정한 **사건(event)** A가 일어나는지 아닌지를 기록한다고 하자. $f_n(A)$를 처음 n개의 독립시행에서 A사건이 일어난 횟수라고 하고, $r_n(A) = f_n(A)/n$이라고 하자. 그러면 이 relative frequency $r_n(A)$는 $n \to \infty$일 때 다음과 같다고 생각하는 것이다(stabilization).

$$r_n(A) \overset{n\to\infty}{\longrightarrow} \text{ some real number.}$$

Part II

Probability Theory

2 The Elements of Proability Theory

2.1 Probability Triples

다음은 콜모고로프 가 정리한 수리적 기반의 확률론이다.

Q. 왜 probability triple이 필요한가? Single도 아니고 double도 아니고 왜 triple이어야 하는가?

- **Sample space** Ω (표본공간): 이것은 any non-empty set이면 된다. 예를 들어 uniform distribution일 때 $\Omega = [0, 1]$이 있다.
- $\mathcal{F}$: σ-algebra 또는 σ-field: 이것은 Ω의 subset들의 collection으로 $\emptyset, \Omega$ 등을 포함한다.
- **Probability** P: a mapping from $\mathcal{F}$ to $[0, 1]$ with
  - $P(\emptyset) = 0$
  - $P(\Omega) = 1$
  - P is countably additive, $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$

2.2 Field and σ-field

Definition 2.1 (Field)**.** The class $\mathcal{A}$ of subsets of Ω is called a **field** if it contains Ω and is closed under the formulation of complements and finite unions, that is if:

1. $\Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \implies A^c \in \mathcal{A}$
3. $A_1, A_2 \in \mathcal{A} \implies A_1 \cup A_2 \in \mathcal{A}$

Definition 2.2 (σ-field)**.** The class $\mathcal{F}$ of subsets of Ω is called a σ-**field** if it is a field and if it is closed under the formulation of countable unions, that is if:

4. $A_1, A_2, ... \in \mathcal{F} \implies \cup_{n=1}^{\infty} A_n \in \mathcal{F}$

- Recall that the elements of any field or σ-field are called **random events** (or simply **events**).

2.3 $\pi - \lambda$ System

Some intuition for $\pi - \lambda$ is that you can take a finite non π-system such as $S = \{\{1,2\},\{2,3\}\}$, and this is not enought to guarantee uniqueness on the σ-algebra generated by S, which includes sets like $\{2\},\{1,2,3\}$. But, at least in the countable case, you can use the π-system property to do disjointification/partitioning on Ω, which finished the proof.

Lemma 2.1 (σ-algebra and π-λ system)**.** A family of sets is a σ-algebra iff it is both π and λ.

2.4 Probabilities

Definition 2.3 (Probability)**.**

- Let Ω be any set and $\mathcal{A}$ be a field of its subsets. We say that P is a **probability** on the measurable space $(\Omega, \mathcal{A})$ if P is defined for all events $A \in \mathcal{A}$ and satisfies the following axioms.

1. $P(A) \geq 0$ for each $A \in \mathcal{A}$; $P(\Omega) = 1$
2. P is **finitely additive**. That is, for any finite number of pairwise disjoint events $A_1, \dots, A_n \in \mathcal{A}$ we have

$$P\Big(\cup_{i=1}^{n} A_i \Big) = \sum_{i=1}^{n} P(A_i).$$

3. P is continuous at $\emptyset$. That is, for any events $A_1, A_2, \dots, \mathcal{A}$ such that $A_{n+1} \subset A_n$ and $\cap_{n=1}^{\infty} A_n = \emptyset$, it is true that

$$\lim_{n \to \infty} P(A_n) = 0.$$

Note that conditions 2 and 3 are equivalent to the next one 4.

4. P is σ-additive (countably additive), that is

$$P\Big(\cup_{n=1}^{\infty} A_n \Big) = \sum_{n=1}^{\infty} P(A_n)$$

for any events $A_1, A_2, \dots \in \mathcal{A}$ which are pairwise disjoint.

Example 2.1 (A probability measure which is additive but not σ-additive)**.** Let Ω be the set of all rational numbers r of the unit interval $[0, 1]$ and $\mathcal{F}_1$ the class of the subsets of Ω of the form $[a, b]$, $(a, b]$, (a, b) or $[a, b)$ where a and b are rational numbers. Denote by $\mathcal{F}_2$ the class of all finite sums of disjoint sets of $\mathcal{F}_1$. Then $\mathcal{F}_2$ is a field. Let us define the probability measure P as follows:

$$P(A) = b - a, \quad \text{if } A \in \mathcal{F}_1,$$
$$P(B) = \sum_{i=1}^{n} P(A_i), \quad \text{if } B \in \mathcal{F}_2, \text{ that is, } B = \sum_{i=1}^{n} A_i, A_i \in \mathcal{F}_1.$$

Consider two disjoint sets of $\mathcal{F}_2$ say

$$B = \sum_{i=1}^{n} A_i \quad \text{and } B' = \sum_{j=1}^{m} A'_j,$$

where $A_i, A'_j \in \mathcal{F}_1$ and all A_i, A'_j are disjoint. Then $B + B' = \sum_{k=1}^{m+n} C_k$ where either $C_k = A_i$ for some $i = 1, \dots, n$, or $C_k = A'_j$ for some $j = 1, \dots, m$. Moreover,

$$P(B + B') = P\Big(\sum_k C_k\Big) = \sum_k P(C_k) = \sum_{i,j} (P(A_i) + P(A'_j))$$
$$= P(A_i) + \sum_j P(A'_j) = P(B) + P(B').$$

and hence P is an additive measure.

Obviously every one-point set $\{r\} \in \mathcal{F}_2$ and $P(\{r\}) = 0$. Since Ω is a countable set and $\Omega = \sum_{i=1}^{\infty} \{r_i\}$, we get

$$P(\Omega) = 1 \neq 0 = \sum_{i=1}^{\infty} P(\{r_i\}).$$

This contradiction shows that P is not σ-additive.

3 Random Variables

3.1 Random Variables

Definition 3.1 (Random Variables)**.** Given a probability triple $(\Omega, \mathcal{F}, P)$, a **random variable** is a function X from Ω to $\mathbb{R}$, such that

$$\{\omega \in \Omega; X(\omega) \leq x\} \in \mathcal{F}, \quad x \in \mathbb{R}.$$

Q. Random variable을 정의하는데 왜 inverse image를 쓰는가?

Commonly a probability measure P is added to $(\Omega, \mathcal{F})$. Then sets like $\{X \in A\} := \{\omega \in \Omega | X(\omega) \in A\}$ can $= X^{-1}(A)$ be **measured** if they belong to $\mathcal{F}$. 예를 들면 $X : \Omega \to \mathbb{R}$이 확률변수일 때 $X < 1$일 확률을 구하려면 $X^{-1}(-\infty, 1)$이 가측이어야 할 것이다.

3.2 Radon-nikodym derivative

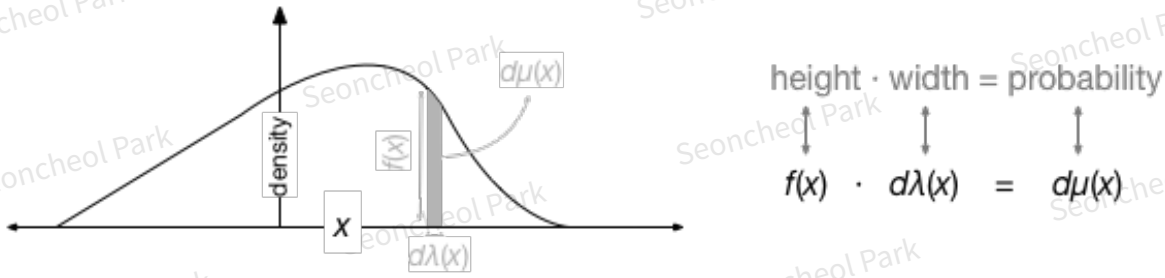

Figure 3.1: Change of measures.

확률측도는 volume element의 일반화라고 볼 수 있다.

- $\mu(x)$: probability measure, interval이나 set of points들을 인풋으로 받고 area/volume에 해당하는 확률(양수)을 아웃풋으로 주는 함수다.

- $\lambda(x)$: reference measure. We often take $\lambda(x)$ as the Lebesgue measure which is essentially just a uniform function over the sample space.

The reference measure $\lambda(x)$ is essentially just a meter-stick that allows us to express the probability measure as a simple function $f(x)$. That is, we represent the probability measure $\mu(x)$ as $f(x)$ by comparing the probability measure to some specified reference measure $\lambda(x)$. This is essentially the intuition that is given by the Radon-Nikodym derivative

$$f(x) = \frac{d\mu(x)}{d\lambda(x)}$$

or equivalently

height = area / width.

Note that we can also represent the same idea by

$$\mu(A) = \int_{A \in X} f(x) d\lambda(x),$$

where $\mu(A)$ is the sum of the probability of events in the set A which is itself a subset of the entire sample space X. Note that when $A = X$ then the integral must equal 1 by definition of probability.

라돈-니코딤 정리는 조건부 확률에 응용된다고 함.

3.3 Integration

3.4 리만-스틸체스 적분

종종 헷갈리는 표현이 기댓값을 다음과 같이 분포함수를 이용해 표현하는 경우가 있다.

$$E(X) = \int x dF(x).$$

우리가 알고 있는 정적분은 x 축을 따라가며 함수값 f(x)가 만드는 면적을 계산한다.

$$\int_a^b f(x) dx.$$

위 식을 더 확장하면 x 대신 임의의 곡선 $g(x)$를 적분 변수로 두고 $f(x)$를 단순하게 정적분 할 수도 있다.

$$\int_{x=a}^{b} f(x)dg(x).$$

여기서 $dg(x)$는 $g(x)$의 미분소(differential)로, $g(x)$의 움직임을 결정하는 x는 단조 증가하거나 감소한다. 위와 같이 리만 적분을 일반화한 정적분을 **리만-스틸체스 적분(Riemann-Stieltjes Integral)**이라 한다. 리만 적분의 정의를 이용해 리만-스틸체스의 적분을 표현할 수도 있다.

$$\int_{x=a}^{b} f(x)dg(x) = \lim_{N\to\infty} \sum_{n=0}^{N-1} f(t_n)[g(x_{n+1}) - g(x_n)].$$

여기서 x_n은 정적분을 위해 구간 $[a, b]$를 나눈 점, t_n은 닫힌 세부공간 $[x_n, x_{n+1}]$사이에 있는 임의점이다.

3.5 리만 적분과 르베그 적분

여기는 Confused when changing from Lebesgue Integral to Riemann Integral 에 올라왔던 내용을 살펴보기로 한다. 여기서 질문자는 리만 적분을 어떻게 르베그 적분으로 바꾸는지에 대해 관심이 있다.

다음과 같이 확률공간 $(\Omega, \mathcal{F}, P)$에서 정의된 음이 아닌 확률변수 X가 지수분포를 따른다고 하자.

$$P(X < x) = 1 - e^{-\lambda x}.$$

한편, 르베그 적분으로 X의 기댓값을 쓰면 다음과 같다.

$$E[X] = \int_{\{\omega | X(\omega) \geq 0\}} X(\omega)dP(\omega).$$

여기서 질문자는 이것을 리만 적분으로 어떻게 바꾸냐

$$E[X] = \int_{0}^{\infty} x\lambda e^{-\lambda x}dx$$

를 물어보고 있다.

답변은 이것이 적분의 문제가 아닌 변수변환의 문제라고 한다.

By definition, given $X : \Omega \to \mathbb{R}$ a random variable, $E[X] = \int_\Omega X$. X defines a measure $\tilde{m}$ in $\mathbb{R}$, called the **push-forward**, by $\tilde{m}(A) = P(X^{-1}(A))$. By definition, this measure is invariant under X, and hence

$$\int_{\mathbb{R}} f d\tilde{m} = \int_\Omega f \circ X dP.$$

The equality follows from the usual arguments (prove for characteristics, simple functions, then use convergence. Recall that $1_A \circ X = 1_{X^{-1}(A)}$).

Let h be the density of X. We then have, by definition of density, that $\tilde{m}(A) = P(X^{-1}(A)) = \int_A h dm$ for any $A \in \mathcal{B}(\mathbb{R})$, where m is the Lebesgue measure. By **change of variables**, we have

$$\int_{\mathbb{R}} f d\tilde{m} = \int_{\mathbb{R}} f \cdot h dm.$$

Combining these equations,

$$\int_{\mathbb{R}} f \cdot h dm = \int_\Omega f \circ X dP.$$

Taking $f = \text{Id}$ yields

$$\int_{\mathbb{R}} x h(x) dx = \int_\Omega X dP = E[X].$$

Taking $f = \text{Id} \cdot \mathbf{1}_I$, where I is some interval (for example, $(0, +\infty)$ as in your case), we have

$$\int_I x h(x) dx = \int_{X^{-1}} X dP,$$

recalling again that $\mathbf{1}_A \circ X = \mathbf{1}_{X^{-1}(A)}$. Since $P(X < 0)$ in your case is 0, this last integral is actually equal to the integral over the whole space, and hence to $E[X]$, which gives your equality.

Definition 3.2 (Integrable Random Variable)**.** Gut (2014) 의 53쪽에 따르면, $E|X| < \infty$인 경우 random varible X가 integrable 하다고 부른다.

Example 3.1. Given a probability measure P and sample space Ω, it is true that

$$\int_\Omega dP = 1.$$

Because

$$\int_{\Omega} dP = P(\Omega) = 1.$$

More generally

$$\int_{A} dP = \int_{\Omega} 1_A dP = P(A), \quad A \in \mathcal{F}.$$

Definition 3.3 ($\mathcal{L}^p$)**.** 다음과 같은 확률공간 $(\Omega, \mathcal{F}, P)$를 생각하자. $p > 1$에 대해, 확률변수 X가 $E|X|^p < \infty$이면 $X \in \mathcal{L}^p$라고 하며 다음과 같은 놈 $\|X_p\| = (E|X|^p)^{\frac{1}{p}}$를 정의할 수 있다.

4 Probability Inequalities

4.1 왜 concentration inequality가 필요한가?

- 출처: Concentration Inequalities
- High-Dimensional Probability 책에 있는 동전 던지기 예제 생각
- i번째 동전던지기: 앞면이 나오면 1, 뒷면이 나오면 0인 Bernoulli random variable로 간주 가능
- N번 던졌을 때 나온 앞면의 수: $S_N = \sum_i X_i$
- de Moivre-Laplace theorem (Binomial의 CLT)

$$Z_N \xrightarrow{D} \mathcal{N}(0,1)$$

이때

$$Z_N = \frac{S_N - N_p}{\sqrt{Np(1-p)}}$$

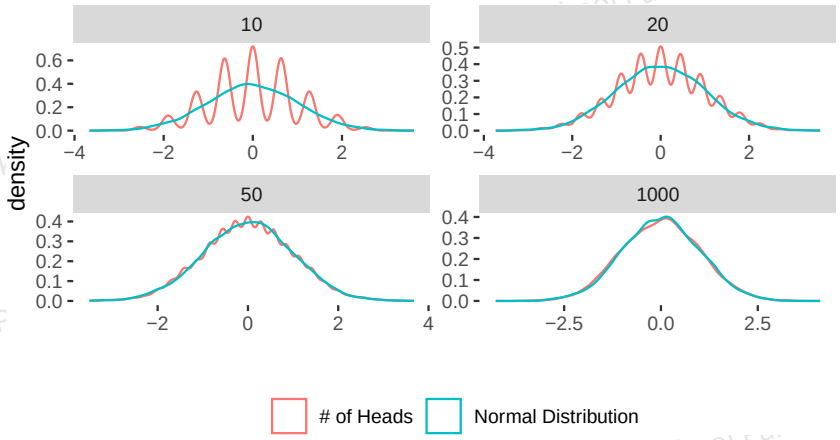

Figure 4.1: Figure: CLT 묘사.

Q. N번 시행 시 $\frac{3}{4}$ 이상 앞면이 나올 확률을 구하고 싶다.

- Gaussian density는 exponential decay하는데, Z_N 이 분포수렴하는 속도는 훨씬 느림
- CLT의 quantitative version인 Berry-Essen CLT를 보면

$$|P\{Z_n \geq t\} - P\{Z \geq t\}| \leq \frac{C}{\sqrt{N}}$$

이때 C는 상수이며, convergence의 order가 $\frac{1}{\sqrt{N}}$ 임을 (아래 그림에 녹색으로 표시) 확인 가능

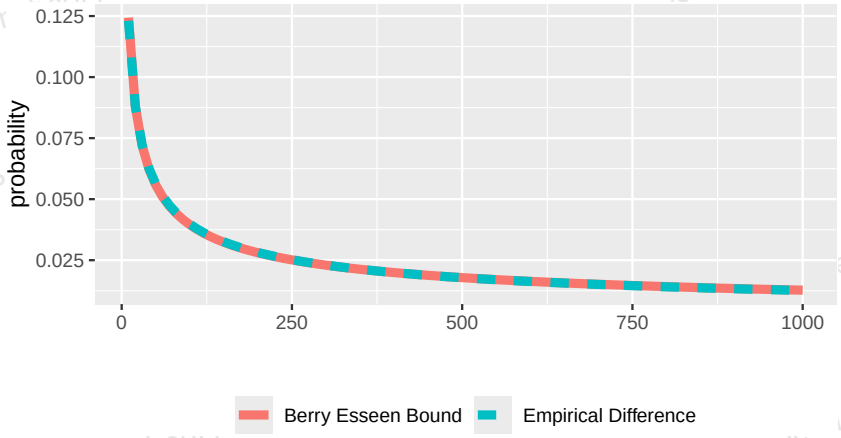

Figure 4.2: Figure: Berry-Essen bound와 empirical difference.

4.2 Markov inequality

Theorem 4.1 (Markov inequality)**.** 음이 아는 확률변수 X에 대해

$$P\{X \geq t\} \leq \frac{E[X]}{t}$$

Proof

확률공간 (Ω, Σ, P)을 생각하자.

$$EX = \int XdP \geq \int_{\{X \geq t\}} XdP \geq t \int_{\{X \geq t\}} dP \geq t \cdot P\{X \geq t\}$$

> **Remark**
>
> - 마르코프 bound는 매우 약한 (즉 true probabilty로의 수렴이 느린) bound
> - 그러나 X에 대한 제약조건이 없음 (기댓값 계산 필요, 음이 아닌 확률변수)

4.3 Chebyshev inequality

Theorem 4.2 (Chebyshev inequality). 어떤 확률변수 X에 대해

$$P\{|X - E(X)| \geq t\} \leq \frac{\text{Var}(X)}{t^2}$$

> **Proof**
>
> $|X - E(X)| \geq t$를 제곱한 후 마르코프 부등식을 적용
>
> $$P\{|X - E(X)|^2 \geq t^2\} \leq \frac{E[(X - E(X))]^2}{t^2} = \frac{\text{Var}(X)}{t^2}$$

> **Remark**
>
> - 체비세프 부등식을 쓰려면 분산이 정의되어야 함

4.4 Hoeffding's Inequality

- (드디어) $\sum_i X_i$에 대한 exponential bound를 줌
- 그러나 독립 가정이 필요
- 단순한 케이스로 먼저 $X_1, \dots, X_N$이 symmetric Bernoulli라고 하자. 이는 즉 반반의 확률로 1 또는 -1을 갖는 확률변수

Theorem 4.3 (Symmetric Bernoulli에서의 Hoeffding's inequality). $X_1, \dots, X_N$이 symmetric Bernoulli 확률변수라고 하자. 어떤 $t \geq 0$에 대해 $a \in \mathbb{R}^n$이 존재해

$$P\{\sum_{i=1}^{N} a_i X_i \geq t\} \leq \exp\left(-\frac{t^2}{2\|a\|^2}\right)$$

Proof

마르코프 부등식을 적용하면 다음과 같다.

$$P\{\sum_{i=1}^{N} a_i X_i \geq t\} = P\{\exp(\lambda \sum_{i=1}^{N} a_i X_i) \geq e^{\lambda t}\} \leq e^{-\lambda t} E\{\exp(\lambda \sum_{i=1}^{N} a_i X_i)\}$$

독립성에 의해 다음과 같다.

$$E\{\exp(\lambda \sum_{i=1}^{N} a_i X_i)\} = E\{\prod_{i=1}^{N} \exp(\lambda a_i X_i)\} = \prod_{i=1}^{N} E\{\exp(\lambda a_i X_i)\}$$

X_i를 $1/2$의 확률로 -1과 1을 갖는 확률변수라고 제한했으므로, 위의 기댓값을 쉽게 구할 수 있다.

$$E\{\exp(\lambda a_i X_i)\} = \frac{e^{\lambda a_i} + e^{-\lambda a_i}}{2} \leq e^{\lambda^2 a_i^2/2}$$

지수함수의 테일러 급수 전개를 이용하면

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}, \quad \frac{e^x + e^{-x}}{2} = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!}, \quad e^{x^2/2} = \sum_{k=0}^{\infty} \frac{x^{2k}}{2^k k!}, \quad \Longrightarrow \quad \frac{e^x + e^{-x}}{2} \leq e^{x^2/2}.$$

$\|a\|^2 = 1$이라 두고 위의 결과를 대입해보자.

$$P\{\sum_{i=1}^{N} a_i X_i \geq t\} \leq e^{-\lambda t}(\prod_{i=1}^{N} e^{\lambda^2 a_i^2/2}) \leq e^{-\lambda t}(e^{\lambda^2 \sum_{i=1}^{N} a_i^2/2}) = e^{-\lambda t}(e^{\lambda^2/2}) = e^{\lambda^2/2 - \lambda t}.$$

위의 부등식은 모든 λ에 대해 성립하고, $\lambda = t$일 때 최소화된다. 따라서

$$P\{\sum_{i=1}^{N} a_i X_i \geq t\} \leq e^{-t^2/2}.$$

따라서, homogeneity에 의해 $\|a\| = 1$을 가정하면 다음과 같다.

$$P\{\sum_{i=1}^{N} \frac{a_i}{\|a\|} X_i \geq \frac{t}{\|a\|}\} \leq e^{-\frac{t^2}{2\|a\|^2}}.$$

X_i가 1 또는 0을 갖는 베르누이 확률변수라고 할 때, $Y_i = 2(X_i - \frac{1}{2})$로 놓으면 Y_i는 symmetric