確率・統計入門

森 立平

2025-12-10

目次

前書き

これは確率論と統計学の入門書である。確率論を数学的に取り扱うには通常は測度論とルベーグ積分を用いる。本書では測度論を学ぶ前の数学専攻の学生を対象に確率論と統計学の基礎を解説する。測度論とルベーグ積分を省略するため、しばしば積分と極限の交換などの等式を証明なしに用いる。後で測度論を学んだ後にぜひ振り返って欲しい。

第 I 部

確率論

第 1 章

はじめに

1.1 なぜ確率論と統計学を学ぶか？

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

1.2 本書の構成

1.3 その他の参考文献

第 2 章

集合論

2.1 集合

集合は対象の集まりである。例えば

$$A = \{1, 2, 3\}, \qquad B = \{\text{赤}, \text{青}, \text{黄}, \text{緑}\}$$

のように表す。集合を構成するものを**要素**もしくは**元**という。

- $x \in A$ で要素 x は集合 A に含まれる
- $x \notin A$ で要素 x は集合 A に含まれない

を表す。

$|A|$ で集合 A の要素数を表すことにする。要素数が無限の集合を考えることもできる。例えば

$$A = \{n \in \mathbb{N} \mid n \text{ は偶数}\}, \qquad B = \{x \in \mathbb{R} \mid x \text{ は無理数}\}$$

という集合は無限集合の例となる。何も要素を持たない集合 (要素数が零の集合) を**空集合**といい、$\emptyset$ で表す。

2.2 集合の関係

集合 A の要素がすべて集合 B に含まれるとき、A を B の**部分集合** という。部分集合に関連する集合関係について記号を以下のように定義する。

$$A \subseteq B \overset{\text{def}}{\iff} \forall x\,(x \in A \implies x \in B) \qquad (A \text{ は } B \text{ の部分集合})$$

$$A \supseteq B \overset{\text{def}}{\iff} B \subseteq A \qquad (A \text{ は } B \text{ の上位集合})$$

$$A = B \overset{\text{def}}{\iff} (A \subseteq B \land A \supseteq B) \qquad (A \text{ と } B \text{ は等しい})$$

$$A \neq B \overset{\text{def}}{\iff} \neg(A = B) \qquad (A \text{ と } B \text{ は等しくない})$$

また、

$$A \subsetneq B \overset{\text{def}}{\iff} (A \subseteq B \land A \neq B) \qquad (A \text{ は } B \text{ の真部分集合})$$
$$A \supsetneq B \overset{\text{def}}{\iff} B \subsetneq A \qquad (A \text{ は } B \text{ の真上位集合})$$

とする。部分集合や上位集合のような集合間の関係を**包含関係**という。集合 A, B について、$A \subseteq B$ と $B \subseteq A$ のどちらかが成り立つとき、「A と B の間に包含関係が成り立つ」という。

包含関係は以下の三条件を満たす。

- (反射律)　$A \subseteq A$.
- (反対称律)　$(A \subseteq B \land B \subseteq A) \iff A = B$.
- (推移律)　$(A \subseteq B \land B \subseteq C) \implies A \subseteq C$.

2.3　集合の演算

複数の集合から新しい集合を作る演算がある。

$$A \cup B := \{x \mid x \in A \text{ **または** } x \in B\} \qquad \textbf{(和集合)}$$
$$A \cap B := \{x \mid x \in A \text{ **かつ** } x \in B\} \qquad \textbf{(積集合)}$$
$$A \setminus B := \{x \mid x \in A \text{ **かつ** } x \notin B\} \qquad \textbf{(差集合)}$$

これらの演算は以下の法則を満たす。

- (交換法則)　$A \cup B = B \cup A, \quad A \cap B = B \cap A$.
- (結合法則)　$(A \cup B) \cup C = A \cup (B \cup C), \quad (A \cap B) \cap C = A \cap (B \cap C)$.
- (分配法則)　$A \cup (B \cap C) = (A \cup B) \cap (A \cup C), \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
- (冪等法則)　$A \cup A = A, \quad A \cap A = A$.
- (吸収法則)　$A \cup (A \cap B) = A, \quad A \cap (A \cup B) = A$.
- $A \cup \emptyset = A, \quad A \cap \emptyset = \emptyset$.

和集合と積集合は結合法則満たすことから、括弧を使わずに $A \cup B \cup C$ と表すことができる。集合 A と B が $A \cap B = \emptyset$ を満たすとき、**確率論文脈では、「A と B は排反である」**(集合論文脈では「互いに素である」)という。

2.4　補集合

全体の集合 Ω というのが文脈上存在する場合は、

$$A^{\mathrm{c}} := \Omega \setminus A \qquad \textbf{(補集合)}$$

と定義する。

補集合に関連して以下の法則が成り立つ。

- $\Omega^{\mathrm{c}} = \emptyset, \quad \emptyset^{\mathrm{c}} = \Omega.$
- $A \cup \Omega = \Omega, \quad A \cap \Omega = A.$
- $A \cup A^{\mathrm{c}} = \Omega, \quad A \cap A^{\mathrm{c}} = \emptyset.$
- (二重補集合の法則) $\quad (A^{\mathrm{c}})^{\mathrm{c}} = A.$
- (ド・モルガンの法則) $\quad (A \cup B)^{\mathrm{c}} = A^{\mathrm{c}} \cap B^{\mathrm{c}}, \quad (A \cap B)^{\mathrm{c}} = A^{\mathrm{c}} \cup B^{\mathrm{c}}$

2.5 集合族

集合 Ω について 2^{Ω} を Ω のすべての部分集合からなる集合を表す。

$$2^{\Omega} \coloneqq \{A \subseteq \Omega\}.$$

これを Ω の**冪集合**という。例えば $\Omega = \{1, 2, 3\}$ のとき、

$$2^{\Omega} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\}$$

である。Ω が有限集合のとき、$|2^{\Omega}| = 2^{|\Omega|}$ が成り立つ。また、2^{Ω} の部分集合を Ω **上の (部分) 集合族**と呼ぶ。集合 Λ の各 $\lambda \in \Lambda$ に対して集合 $A_{\lambda} \subseteq \Omega$ が存在するとき、集合族

$$\{A_{\lambda} \subseteq \Omega \mid \lambda \in \Lambda\} \subseteq 2^{\Omega}$$

を**添字集合 Λ で添字付けられた Ω 上の集合族**という。添字集合が有限集合の場合は集合族全体の和集合や積集合は二つの集合の和集合と積集合の定義を繰り返し用いることで定義できる。それらは以下のように表す。

$$\bigcup_{\lambda \in \Lambda} A_{\lambda} \quad , \quad \bigcap_{\lambda \in \Lambda} A_{\lambda}.$$

添字集合 Λ が無限集合の場合は和集合と積集合を以下で定義する。

$$\bigcup_{\lambda \in \Lambda} A_{\lambda} \coloneqq \{x \in \Omega \mid \exists \lambda \in \Lambda, \quad x \in A_{\lambda}\}$$
$$\bigcap_{\lambda \in \Lambda} A_{\lambda} \coloneqq \{x \in \Omega \mid \forall \lambda \in \Lambda, \quad x \in A_{\lambda}\}$$

この定義は Λ が有限集合の場合も正しいものである。この場合もド・モルガンの法則は成り立つ。つまり、

$$\left(\bigcup_{\lambda \in \Lambda} A_{\lambda}\right)^{\mathrm{c}} = \bigcap_{\lambda \in \Lambda} A_{\lambda}^{\mathrm{c}}$$
$$\left(\bigcap_{\lambda \in \Lambda} A_{\lambda}\right)^{\mathrm{c}} = \bigcup_{\lambda \in \Lambda} A_{\lambda}^{\mathrm{c}}$$

が成り立つ。

証明: $x \in \Omega$ について、

$$\begin{aligned}
x \in \left(\bigcup_{\lambda \in \Lambda} A_\lambda \right)^{\mathrm{c}} &\iff x \notin \left(\bigcup_{\lambda \in \Lambda} A_\lambda \right) \\
&\iff \neg \left(\exists \lambda \in \Lambda, \quad x \in A_\lambda \right) \\
&\iff \forall \lambda \in \Lambda, \quad x \notin A_\lambda \\
&\iff \forall \lambda \in \Lambda, \quad x \in A_\lambda^{\mathrm{c}} \\
&\iff x \in \bigcap_{\lambda \in \Lambda} A_\lambda^{\mathrm{c}}.
\end{aligned}$$

第 3 章

確率空間

3.1　確率論を数学的に定式化するには

確率は身近に現れる (感じられる) ものであるが、それを数学的に定式化することは自明な問題ではない。実際に確率論には複数の数学的定式化が存在する。その中で圧倒的に一般的なのが**測度論的確率論**と呼ばれる定式化である。測度というのは集合の「面積」のようなものであり、確率を測度と捉えるのが測度論的確率論である。これは多くの人間の直感にも自然なものであろう。

まず、確率を考える集合について考えよう。例えばコインを投げて表もしくは裏が出る確率を考えたいときは

$$\Omega = \{\mathrm{H},\, \mathrm{T}\}$$

という集合になる。また、明日の天気の確率を考えたいときは

$$\Omega = \{\text{” 晴”},\, \text{” 雨”},\, \text{” 雪”}\}$$

という集合になるだろう。この Ω の部分集合に確率を与える関数 $P\colon 2^{\Omega} \to \mathbb{R}$ を定義しよう。偏りのないコインの場合は以下のようになる。

$$P(\emptyset) = 0, \qquad P(\{\mathrm{H}\}) = \frac{1}{2}, \qquad P(\{\mathrm{T}\}) = \frac{1}{2}, \qquad P(\{\mathrm{H},\, \mathrm{T}\}) = 1.$$

また、天気の場合は例えば以下のようになる。

$$\begin{aligned} P(\emptyset) &= 0, & P(\{\text{晴}\}) &= 0.7, & P(\{\text{雨}\}) &= 0.2, & P(\{\text{雪}\}) &= 0.1 \\ P(\{\text{晴}, \text{雨}\}) &= 0.9, & P(\{\text{雨}, \text{雪}\}) &= 0.3, & P(\{\text{雪}, \text{晴}\}) &= 0.8, & P(\{\text{晴}, \text{雨}, \text{雪}\}) &= 1. \end{aligned}$$

このように集合 Ω の部分集合に確率を与えることを考える。Ω の要素一つずつに確率を与えれば十分であるようにも思えるが、Ω が連続的な場合には Ω の一つの要素の確率は 0 になってしまうことが多い。例えば明日の正午の気温が $10\pi\,^{\circ}\mathrm{C}$ になる確率は 0 であろう。そのため、Ω の要素ではなく部分集合に確率を与えることにする。そのために測度論が適しているわけである。

3.2　有限集合上の確率空間

確率を考える集合を Ω とする。この Ω のことを**標本空間**という。また、Ω の部分集合のことを**事象**という。そして、事象に確率を与える関数 $P\colon 2^{\Omega} \to \mathbb{R}_{\geq 0}$ を**確率測度**という。確率測度は以下の条件を満たす。

1. $P(\Omega) = 1$.
2. $\forall A,\, B \subseteq \Omega, \quad A \cap B = \emptyset \implies P(A \cup B) = P(A) + P(B)$.

一つ目の条件は全体の確率が 1 であるという意味の条件である。二つ目の条件は排反な二つの事象の和集合の確率はそれぞれの事象の確率の和であるという意味の条件である。この二つ目の条件を**有限加法性**という。例えば

$$P(\{\text{” 晴”}, \text{” 雨”}\}) = P(\{\text{” 晴”}\}) + P(\{\text{” 雨”}\})$$

という等式は「晴れもしくは雨になる確率 $=$ 晴れになる確率 $+$ 雨になる確率」という意味の等式になる。よって有限加法性が自然な条件であることが分かるだろう。また、これらのことから、P は各要素 $\omega \in \Omega$ に対する確率 $P(\{\omega\})$ から一意に定まることが分かる。この標本空間と確率測度のペア $(\Omega,\, P)$ を**確率空間**という。

3.3　可算無限集合上の確率

標本空間 Ω が可算無限集合のときも、同様に確率測度を定義することもできるが、ここではより強い以下の条件を考える。

1. $P(\Omega) = 1$.
2. $\forall (A_n \subseteq \Omega)_{n\geq 0}, \quad \forall i \neq j,\, A_i \cap A_j = \emptyset \implies P\Big(\bigcup_{n\geq 0} A_n\Big) = \sum_{n\geq 0} P(A_n)$.

ここで二つ目の条件における $_{n\geq 0}$ という添字において n は非負整数全体をわたる。今後もこの記法を用いる。この二つ目の条件を**完全加法性**もしくは **σ-加法性**という。右辺の無限和において $P(A_n) \geq 0$ なので、無限和 $\sum_{n\geq 0} P(A_n)$ が存在するときこれは絶対収束する。よって、この無限和は事象列の並び順に依存しないことが分かる。完全加法性ではなく有限加法性だけを使って確率論を構築する試みもあるが、確率測度の連続性などの性質が失なわれるため標準的な確率論では完全加法性を課す。

3.4　すべての部分集合を可測にはできない

より一般に Ω が非可算無限集合の場合を考えよう。例えば $\Omega = [0, 1)$ の場合が考えられる。このとき、すべての $A \subseteq \Omega$ に確率を与えることができるのだろうか？ そうすると、ごく自然な性質を持つような確率測度が**存在しない**ことが、選択公理を認めると示される。

定理 3.1 (非可測集合の存在)**.** $\Omega = [0, 1)$ とする。また、集合の平行移動を

$$A + x \coloneqq \{a + x - \lfloor a + x \rfloor \mid a \in A\}$$

と定義する。このとき、

$$\forall x \in \Omega,\ A \subseteq \Omega, \quad P(A+x) = P(A) \qquad \textbf{(平行移動不変性)}$$

を満たす確率測度 $P\colon 2^{[0,1)} \to \mathbb{R}_{\geq 0}$ は存在しない。

証明. Ω 上の同値関係を $x \sim y \overset{\text{def}}{\iff} x - y \in \mathbb{Q}$ と定義する。選択公理より、この同値関係の同値類から一つずつ要素を含む集合 V が存在する (Vitali 集合)。任意の $x \in [0,1)$ について、ある $z \in V$ が**唯一**存在して $x \sim z$ である。よって、任意の $x \in [0,1)$ について、ある $z \in V$ と $q \in \mathbb{Q} \cap [0,1)$ が**唯一**存在して $x = z + q - \lfloor z + q \rfloor$ であることから

$$[0,1) = \bigcup_{q \in \mathbb{Q} \cap [0,1)} (V+q)$$

であり、右辺は互いに排反である。よって条件を満たす P が存在すると仮定すると、

$$\begin{aligned} 1 = P([0,1)) &= P\left(\bigcup_{q \in \mathbb{Q} \cap [0,1)} (V+q) \right) \\ &= \sum_{q \in \mathbb{Q} \cap [0,1)} P(V+q) \qquad \textbf{(完全加法性)} \\ &= \sum_{q \in \mathbb{Q} \cap [0,1)} P(V) \qquad \textbf{(平行移動不変性)} \end{aligned}$$

ここで $P(V)$ をどのように定めても、それを無限回足して 1 にすることはできない。よって P は存在しない。 □

3.5 確率論の公理

定理 3.1 より、

1. 選択公理
2. 確率測度の完全加法性
3. $[0,1)$ 上の平行移動不変な確率測度
4. Ω のすべての部分集合に確率を与える

のどれかを諦めないといけない。標準的な確率論では 4 を諦める。以下に確率空間の厳密な定義を述べる。

定義 3.1 (完全加法族)**.** Ω を集合とする。Ω 上の集合族 $\mathcal{F} \subseteq 2^{\Omega}$ が以下を満たすとする。

1. $\Omega \in \mathcal{F}$.
2. $\forall A \in \mathcal{F}, \quad A^{\mathrm{c}} = \Omega \setminus A \in \mathcal{F}$.
3. $\forall (A_n \in \mathcal{F})_{n \geq 0}, \quad \bigcup_{n \geq 0} A_n \in \mathcal{F}$.

このとき、$\mathcal{F}$ を Ω 上の完全加法族もしくは σ-加法族という。

定義 3.2 (確率空間)**.** Ω を集合とし、$\mathcal{F} \subseteq 2^{\Omega}$ を Ω 上の完全加法族とする。

また、$P\colon \mathcal{F} \to \mathbb{R}_{\geq 0}$ が以下を満たすとする。

1. $P(\Omega) = 1$.
2. $\forall (A_n \in \mathcal{F})_{n\geq 0}, \quad \forall i \neq j,\, A_i \cap A_j = \emptyset \implies P\left(\bigcup_{n\geq 0} A_n\right) = \sum_{n\geq 0} P(A_n)$.

このとき、$(\Omega, \mathcal{F}, P)$ を確率空間という。また、Ω を標本空間、$\mathcal{F}$ を事象集合、P を確率測度という。

このように事象集合 $\mathcal{F}$ の元についてのみ確率が与えられる。

> **ノート**
>
> ペア $(\Omega, \mathcal{F})$ を可測空間という。確率空間 $(\Omega, \mathcal{F}, P)$ から $P(\Omega) = 1$ の条件を除いたものが一般の測度空間である。

定理 3.1 の証明では選択公理を用いたが、実際に選択公理を認めないと定理 3.1 が成立しないことが分かっている。以降、Ω が非可算無限集合である場合も含めて、$\mathcal{F} = 2^{\Omega}$ だと思うことにする。これは正しくない場合もあるのだが、選択公理を使わない限り矛盾は導かれないので、問題になることはほとんどない。よって、以降は (Ω, P) を確率空間とする。

$\Omega = [0, 1)$ で任意の $0 \leq a < b \leq 1$ について $P([a, b)) = b - a$ であるような確率空間はとても基本的なものである。今後この性質を満たす確率空間が存在すると仮定して話を進める。

3.6 確率の性質

補題 3.1 (確率のいくつかの性質)**.** 確率空間 (Ω, P) と任意の $A, B \subseteq \Omega$ について以下が成り立つ。

1. $P(A^{\mathrm{c}}) = 1 - P(A)$.
2. $B \subseteq A \implies P(B) \leq P(A)$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
4. $P(A \cup B) \leq P(A) + P(B)$ *(ブールの不等式、union bound).*

証明.

1. $$1 = P(\Omega) = P(A \cup A^{\mathrm{c}}) = P(A) + P(A^{\mathrm{c}}).$$

2. $$P(A) = P(B \cup (A \setminus B)) = P(B) + P(A \setminus B) \geq P(B).$$

3. $$P(A \cup B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A)$$
   $$P(B) = P((B \setminus A) \cup (A \cap B)) = P(B \setminus A) + P(A \cap B)$$
   より、$P(B \setminus A)$ を消去することで得られる。

4. 3 より自明

□

補題 3.2 (ユニオンバウンド)**.** 確率空間 (Ω, P) と $(A_n \subseteq \Omega)_{n\geq 0}$ について、

$$P\left(\bigcup_{n\geq 0} A_n\right) \leq \sum_{n\geq 0} P(A_n).$$

証明. $B_0 \coloneqq A_0$, $B_n \coloneqq A_n \setminus \bigcup_{k=0}^{n-1} A_k$ とおくと、

$$\begin{aligned} P\left(\bigcup_{n\geq 0} A_n\right) &= P\left(\bigcup_{n\geq 0} B_n\right) \\ &= \sum_{n\geq 0} P(B_n) \\ &\leq \sum_{n\geq 0} P(A_n)\,. \end{aligned}$$

□

定理 3.2 (確率測度の連続性)**.** 確率空間 (Ω, P) と事象列 $A_0 \subseteq A_1 \subseteq \cdots \subseteq \Omega$ について

$$P\left(\bigcup_{n\geq 0} A_n\right) = \lim_{n\to\infty} P(A_n).$$

また、事象列 $\Omega \supseteq A_0 \supseteq A_1 \supseteq \cdots$ について

$$P\left(\bigcap_{n\geq 0} A_n\right) = \lim_{n\to\infty} P(A_n).$$

証明. 事象列 $A_0 \subseteq A_1 \subseteq \cdots \subseteq \Omega$ について考える。$B_0 \coloneqq A_0$, $n \geq 1$ について $B_n \coloneqq A_n \setminus A_{n-1}$ とおく。このとき、$i \neq j$ について $B_i \cap B_j = \emptyset$。また $k \geq 0$ について、$\bigcup_{n=0}^{k} B_n = \bigcup_{n=0}^{k} A_n = A_k$ が成り立つ.

$$\begin{aligned} P\left(\bigcup_{n\geq 0} A_n\right) = P\left(\bigcup_{n\geq 0} B_n\right) &= \sum_{n\geq 0} P(B_n) \\ &= \lim_{k\to\infty} \sum_{n=0}^{k} P(B_n) = \lim_{k\to\infty} P\left(\bigcup_{n=0}^{k} B_n\right) = \lim_{k\to\infty} P(A_k). \end{aligned}$$

事象列 $\Omega \supseteq A_0 \supseteq A_1 \supseteq \cdots$ について考える。このとき、$A'_n \coloneqq A_n^{\mathrm{c}}$ とおくと、$A'_0 \subseteq A'_1 \subseteq \cdots \subseteq \Omega$ を満たす。よって、

$$P\left(\bigcup_{n\geq 0} A'_n\right) = \lim_{n\to\infty} P(A'_n)$$

である。ド・モルガンの公式より、

$$
\begin{aligned}
P\left(\left(\bigcap_{n\geq 0} A_n\right)^{\mathrm{c}}\right) &= \lim_{n\to\infty} P(A_n^{\mathrm{c}}) \\
\iff 1 - P\left(\bigcap_{n\geq 0} A_n\right) &= \lim_{n\to\infty} (1 - P(A_n)) \\
\iff P\left(\bigcap_{n\geq 0} A_n\right) &= \lim_{n\to\infty} P(A_n).
\end{aligned}
$$

□

第 4 章

確率変数と確率分布

4.1　確率変数

確率空間の上の様々な部分集合の確率を調べたい。そのためには確率変数を導入すると便利である。

定義 4.1 (確率変数)**.** 確率空間 (Ω, P) について、関数 $X\colon \Omega \to \mathbb{R}$ を**確率変数**という。

また、任意の $A \subseteq \mathbb{R}$ について

$$\Pr(X \in A) := P(\{\omega \in \Omega \mid X(\omega) \in A\})$$

と定義する。さらに、任意の $a \in \mathbb{R}$ について

$$\Pr(X \geq a) := P(X \in \{x \in \mathbb{R} \mid x \geq a\}) = P(\{\omega \in \Omega \mid X(\omega) \geq a\})$$

と定義する。同様に $\Pr(X > a)$, $\Pr(X \leq a)$, $\Pr(X < a)$, $\Pr(X = a)$ なども定義される。

> **ノート**
>
> 本当の確率論では確率変数は「確率空間から位相空間への写像で開集合の逆像が可測集合になるような関数」として定義される。また、文献によっては確率変数は「確率空間から可測空間への写像で可測集合の逆像が可測集合になるような関数」として定義されることもある。開集合族が生成する完全加法族をボレル集合族と呼ぶが、この方法で位相空間から可測空間を作ることができる。そうすると前者の定義は「確率空間から位相空間への写像でボレル集合の逆像が可測集合になるような関数」と等価なので、後者の定義に含まれる。後者の定義の中の可測空間が位相空間由来のものに限定したのが前者の定義であると言える。

任意の関数 $f\colon \mathbb{R} \to \mathbb{R}$ について $f \circ X$ は確率変数である。このとき、$\Pr(f \circ X \in A)$ と書く代わりに $\Pr(f(X) \in A)$ と書く。

> **ノート**
>
> 本当の確率論の言葉で述べると任意のボレル関数 (開集合 (ボレル集合としても等価) の逆像がボレル集合となる関数) f について $f \circ X$ は確率変数である。

共通の確率空間 (Ω, P) 上の確率変数 X, Y についてその和 $X+Y$ や積 XY も確率変数である。

例 4.1 (コイン投げ)**.** コインを 2 回独立に投げる場合の確率空間 (Ω, P) を以下で定義する。

- $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$.
- $P(A) = \frac{|A|}{4}$.

確率変数 X_1 と X_2 を

$$\begin{array}{llll} X_1(\mathrm{HH}) = 1 & X_1(\mathrm{HT}) = 1 & X_1(\mathrm{TH}) = 0 & X_1(\mathrm{TT}) = 0 \\ X_2(\mathrm{HH}) = 1 & X_2(\mathrm{HT}) = 0 & X_2(\mathrm{TH}) = 1 & X_2(\mathrm{TT}) = 0 \end{array}$$

とすると、$k \in \{1, 2\}$ について、X_k は k 回目のコイン投げが表だったときに 1、裏だったときに 0 になる。また、$X = X_1 + X_2$ は表が出た回数を表す確率変数となる。表が一回以上出る確率は以下のように表せる。

$$\Pr(X \geq 1) = P(\{\omega \in \Omega \mid X(\omega) \geq 1\}) = P(\{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}\}) = \frac{3}{4}.$$

4.2 確率分布

定義 4.2 (累積分布関数)**.** 確率変数 X について**累積分布関数** $F_X \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ は以下で定義される。

$$F_X(x) := \Pr(X \leq x) \qquad \forall x \in \mathbb{R}.$$

累積分布関数は定義より単調非減少関数であることが分かる。累積分布関数は連続とは限らないが右連続であることは以下のように確認できる。

$$\begin{aligned} \lim_{n\to\infty} \Pr\left(X \leq x + \frac{1}{n}\right) &= \lim_{n\to\infty} P\left(\left\{\omega \in \Omega \mid X(\omega) \leq x + \frac{1}{n}\right\}\right) \\ &= P\left(\bigcap_{n=0}^{\infty} \left\{\omega \in \Omega \mid X(\omega) \leq x + \frac{1}{n}\right\}\right) \\ &= P(\{\omega \in \Omega \mid X(\omega) \leq x\}) \\ &= \Pr\left(X \leq x\right). \end{aligned}$$

累積分布関数 $F_X \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ は以下の性質を持つ。

1. F_X は単調非減少。
2. F_X は右連続。

3. $\lim_{x\to\infty} F_X(x) = 1$.
4. $\lim_{x\to-\infty} F_X(x) = 0$.

確率変数 X の像が高々可算集合のとき、X を**離散型確率変数**という。離散型確率変数でない確率変数を**連続型確率変数**という。

X が離散型確率変数のとき、A を X の像の部分集合とすると

$$\begin{aligned}\Pr(X \in A) &= P(\{\omega \in \Omega \mid X(\omega) \in A\}) \\ &= P\left(\bigcup_{x\in A} \{\omega \in \Omega \mid X(\omega) = x\}\right) \\ &= \sum_{x\in A} P(\{\omega \in \Omega \mid X(\omega) = x\}) \\ &= \sum_{x\in A} \Pr(X = x)\end{aligned}$$

であるので、

$$f_X(x) \coloneqq \Pr(X = x) \qquad \forall x \in \mathbb{R}$$

という関数を用いて

$$\Pr(X \in A) = \sum_{x\in A} f_X(x)$$

と表せる。この $f_X(x)$ を X の**確率質量関数**という。

X が連続型確率変数のとき、

$$\Pr(X \in A) = \int_A f_X(x)\mathrm{d}x \qquad \forall A \subseteq \mathbb{R}$$

を満たす $f_X\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ が存在するとき、この $f_X(x)$ を X の**確率密度関数**という。しかし、確率密度関数は常に存在するとは限らない。

例 4.2. $\Omega = [0, 1)$ とし、$P\colon 2^\Omega \to \mathbb{R}_{\geq 0}$ を $P([a, b)) = b - a$ を満たす確率測度とする。確率空間 (Ω, P) 上の確率変数 $X\colon [0, 1) \to \mathbb{R}$ を

$$X(\omega) = \begin{cases} 2\omega & \text{if } \omega < \frac{1}{2} \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

と定義する。

この時 X の像は $[0, 1)$ であるので X は連続型確率変数である。

また、$\Pr(X = \frac{1}{2}) = P([\frac{1}{2}, 1) \cup \{\frac{1}{4}\}) = \frac{1}{2}$ である。よって X の確率密度関数は存在しない。

連続確率変数 X が確率密度関数を持つとき、

$$F_X(x) = \int_{-\infty}^{x} f_X(z)\mathrm{d}z$$

という関係が成り立つ。よって累積分布関数が微分可能なときは

$$f_X(x) = \frac{\mathrm{d}F_X}{\mathrm{d}x}$$

とおいても構わない。確率密度関数はその積分値にのみ意味を持つ。そのため確率密度関数は一意には定まらない。

例 4.2 の確率変数 X の累積分布関数は以下のようになる。

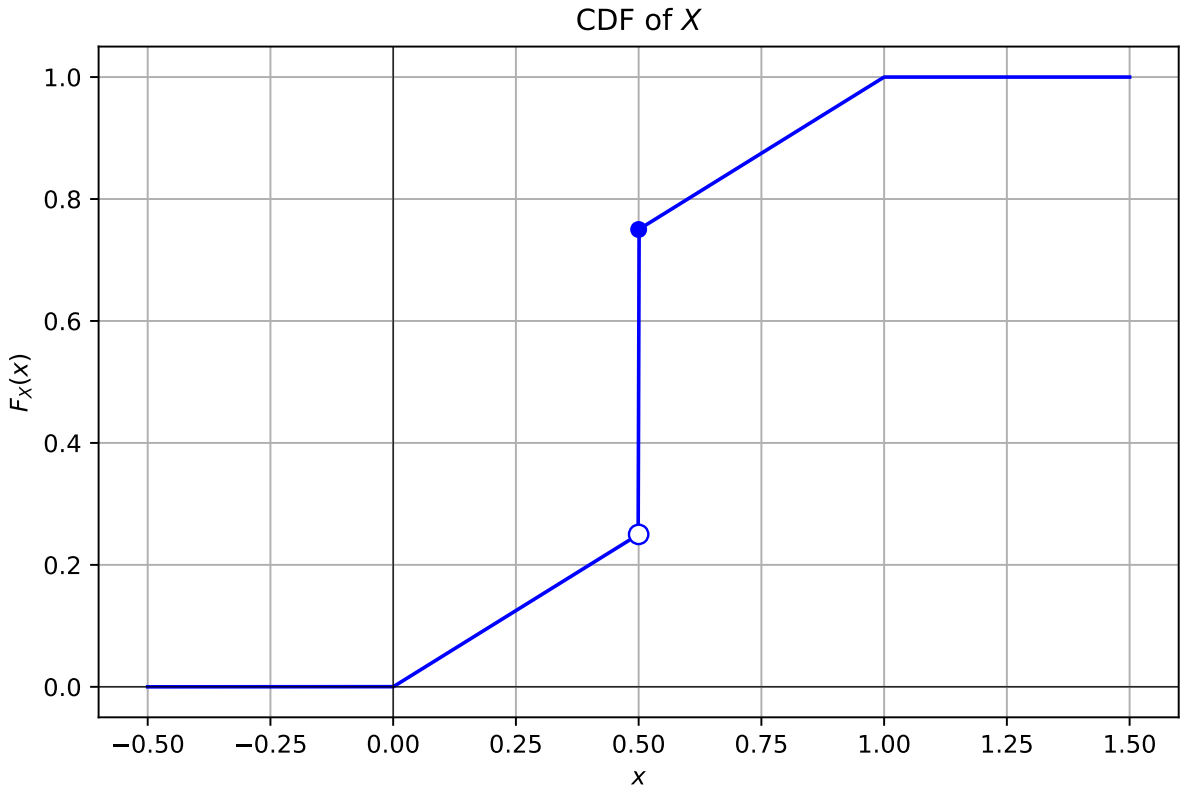

図 4.1: X の累積分布関数

4.3 離散確率分布の例

4.3.1 様々な離散確率分布

- 二項分布 $\mathrm{Binom}(n,p)$: 表が出る確率が p のコインを n 回独立に投げたとき、表が出る回数の分布

$$\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}.$$

- 幾何分布: 表が出るまでに投げるコインの回数の分布

$$\Pr(X=k)=(1-p)^{k-1}p.$$

- 超幾何分布: 袋の中に N 個のボールがあって、そのうち K 個が当たりとし、n 個引いたときの当たりの個数の分布

$$\Pr(X=k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

- ポアソン分布:

$$\Pr(X=k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}.$$

4.3.2 二項分布

二項分布の確率質量関数は

$$f_X(k)=\binom{n}{k}p^k(1-p)^{n-k}$$

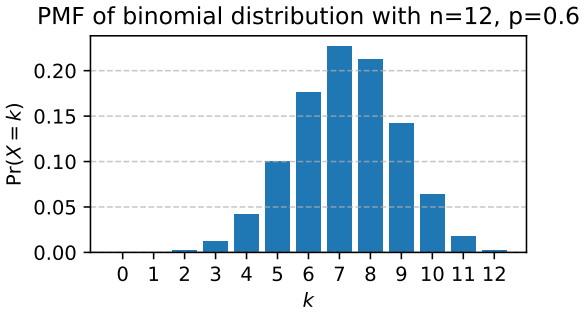

図 4.2: 二項分布の確率質量関数

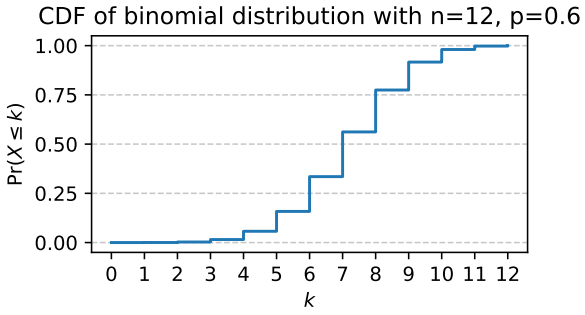

図 4.3: 二項分布の累積分布関数

4.3.3 幾何分布

$$f_X(k)=(1-p)^{k-1}p,\qquad \forall k\geq 1$$

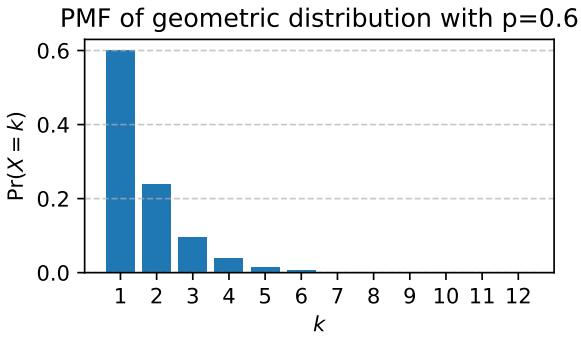

図 4.4: 幾何分布の確率質量関数

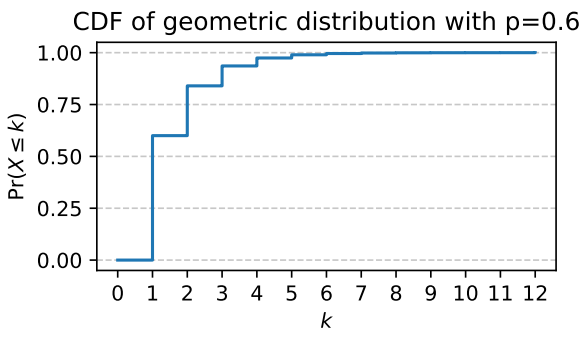

図 4.5: 幾何分布の累積分布関数

4.3.4 超幾何分布

$$f_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \qquad \forall k \in \{0, \dots, \min\{K, n\}\}$$

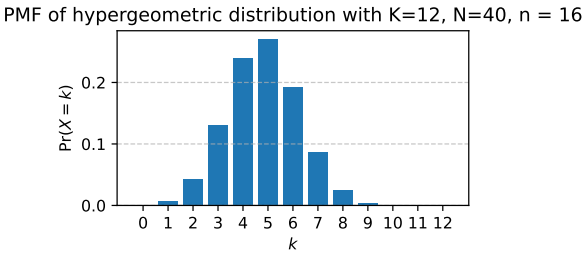

図 4.6: 超幾何分布の確率質量関数

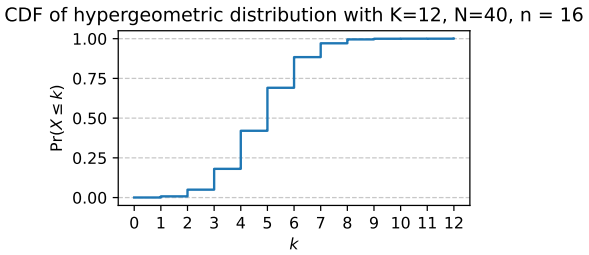

図 4.7: 超幾何分布の累積分布関数

4.3.5 ポアソン分布

$$f_X(k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \qquad \forall k \geq 0$$

4.3.6 確率空間は何か？

確率変数が従う確率分布のみに注目した議論をする場合、確率空間を陽に考えないことがしばしばある。

- 二項分布: n 回のコイン投げ、もしくは無限回のコイン投げ。
- 幾何分布: 無限回のコイン投げ。

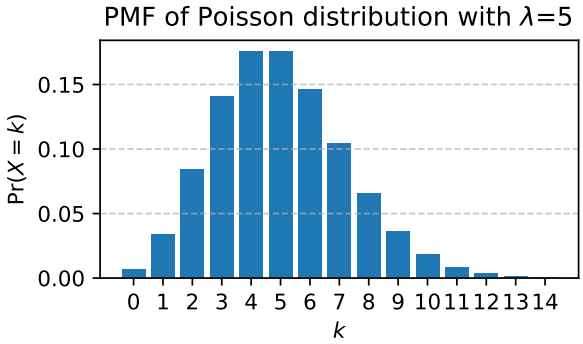

図 4.8: ポアソン分布の確率質量関数

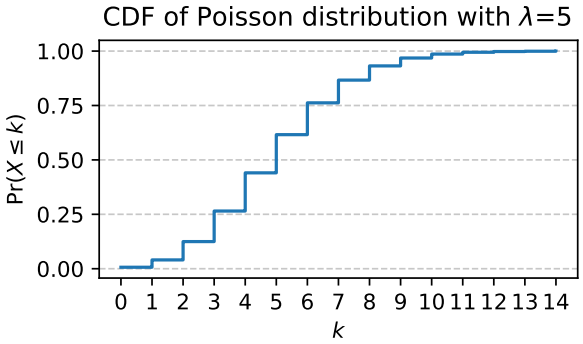

図 4.9: ポアソン分布の累積分布関数

- 超幾何分布: K 個の当たりと $N-K$ 個のはずれのボールの $\binom{N}{K}$ 通りの並び順。
- ポアソン分布: ？

4.4 連続確率分布の例

- 一様分布
- 正規分布
- 指数分布

4.4.1 一様分布

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}, \qquad F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b} & \text{if } x \in [a,b] \\ 1 & \text{otherwise.} \end{cases}$$

4.4.2 正規分布 (ガウス分布)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(z-\mu)^2}{2\sigma^2}} \mathrm{d}z$$

特に $\mu = 0$, $\sigma = 1$ のとき**標準正規分布**という。

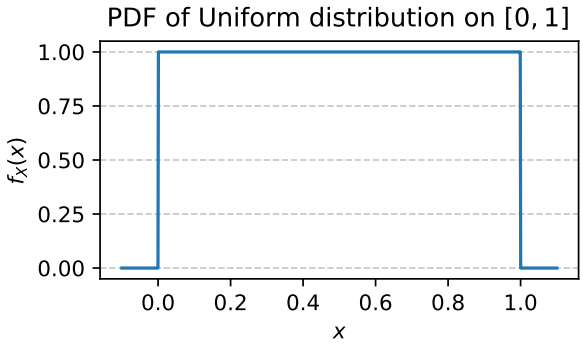

図 4.10: 一様分布の確率密度関数

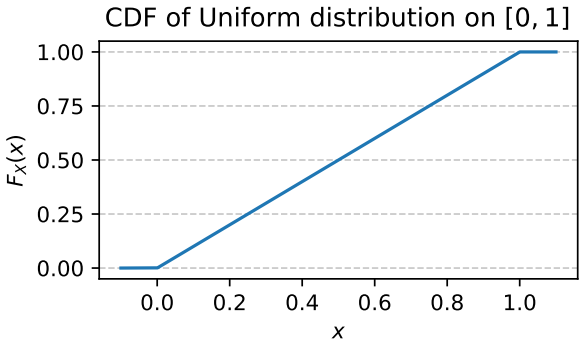

図 4.11: 一様分布の累積分布関数

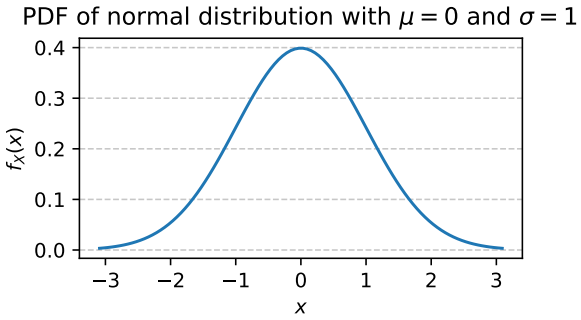

図 4.12: 標準正規分布の確率密度関数

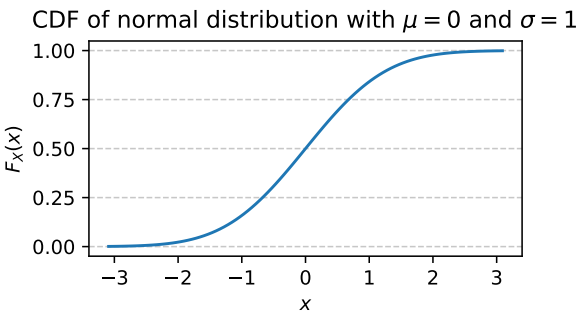

図 4.13: 標準正規分布の累積分布関数

4.4.3 指数分布

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda \mathrm{e}^{-\lambda x} & \text{otherwise} \end{cases}, \qquad F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - \mathrm{e}^{-\lambda x} & \text{otherwise} \end{cases}$$

4.5 確率密度関数

補題 4.1. 確率変数 X について、

$$f_{X+a}(x) = f_X(x-a) \qquad \forall a \in \mathbb{R}$$
$$f_{aX}(x) = \frac{1}{|a|} f_X(x/a) \qquad \forall a \in \mathbb{R}_{\neq 0}$$

はそれぞれ $X + a$ と aX の確率密度関数になる。

証明. 関数 $f_Z \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ が

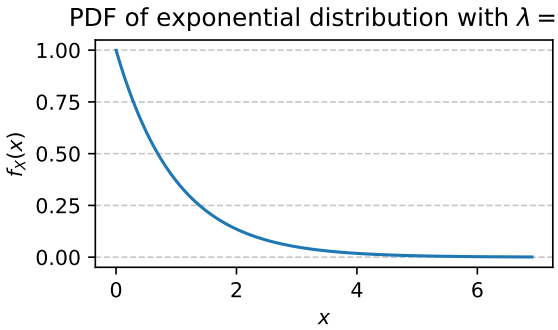

図 4.14: 指数分布の確率密度関数

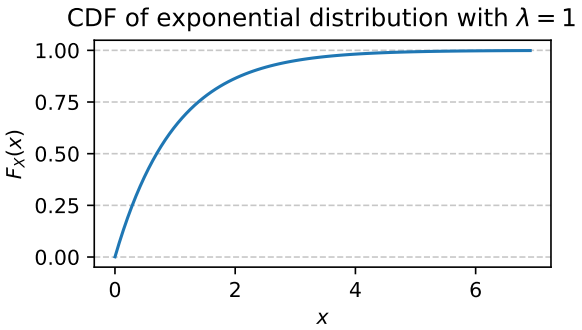

図 4.15: 指数分布の累積分布関数

$$\Pr(Z \leq x) = \int_{-\infty}^{x} f_Z(z)\mathrm{d}z$$

を満たすとき、f_Z は確率変数 Z の確率密度関数となる。

任意の $a \in \mathbb{R}$ について、

$$\begin{aligned}\Pr(X + a \leq x) &= \Pr(X \leq x - a) \\ &= \int_{-\infty}^{x-a} f_X(z)\mathrm{d}z \\ &= \int_{-\infty}^{x} f_X(z' - a)\mathrm{d}z' \qquad (z' = z + a).\end{aligned}$$

任意の $a > 0$ について、

$$\begin{aligned}\Pr(aX \leq x) &= \Pr(X \leq x/a) \\ &= \int_{-\infty}^{x/a} f_X(z)\mathrm{d}z \\ &= \int_{-\infty}^{x} \frac{1}{a} f_X(z'/a)\mathrm{d}z' \qquad (z' = az).\end{aligned}$$

任意の $a < 0$ について、

$$\begin{aligned}\Pr(aX \leq x) &= \Pr(X \geq x/a) \\ &= \int_{x/a}^{\infty} f_X(z)\mathrm{d}z \\ &= \int_{x}^{-\infty} \frac{1}{a} f_X(z'/a)\mathrm{d}z' \qquad (z' = az) \\ &= -\int_{-\infty}^{x} \frac{1}{a} f_X(z'/a)\mathrm{d}z'.\end{aligned}$$

□

補題 4.2. $J \subseteq \mathbb{R}$ を有界とは限らない区間とし、$g\colon J \to \mathbb{R}$ を J の内点で**微分可能**で $g'(x) > 0$ と する。確率変数 X が $\Pr(X \in J) = 1$ を満たし確率密度関数を持つとき、

$$f_{g(X)}(x) = \begin{cases} \frac{1}{g'(g^{-1}(x))} f_X(g^{-1}(x)) & \textit{if } x \in \mathrm{Image}(g) \\ 0 & \textit{otherwise} \end{cases}$$

は $g(X)$ の確率密度関数になる。

証明. $\Pr(g(X) \in \mathrm{Image}(g)) = 1$ なので、$x \notin \mathrm{Image}(g)$ について $f_{g(X)}(x) = 0$ とおいてよい。

任意の $x \in \mathrm{Image}(g)$ について

$$\begin{aligned} \Pr(g(X) \leq x) &= \Pr(X \leq g^{-1}(x)) \\ &= \int_{-\infty}^{g^{-1}(x)} f_X(z)\mathrm{d}z \\ &= \int_{\inf J}^{g^{-1}(x)} f_X(z)\mathrm{d}z \\ &= \int_{\inf \mathrm{Image}(g)}^{x} f_X(g^{-1}(z')) \frac{1}{g'(g^{-1}(z'))}\mathrm{d}z' \qquad (z' = g(z)) \\ &= \int_{-\infty}^{x} f_{g(X)}(z')\mathrm{d}z'. \end{aligned}$$

□

第 5 章

複数の確率変数

5.1 事象の条件付き確率と独立性

定義 5.1 (条件付き確率)**.** 確率空間 (Ω, P) の事象 $A, B \subseteq \Omega$ について $P(B) > 0$ のとき、B **における** A **の条件付き確率**は以下で定義される。

$$P(A \mid B) \coloneqq \frac{P(A \cap B)}{P(B)}.$$

二つの事象 $A, B \subseteq \Omega$ を考える文脈では $P(A \cap B)$ を**同時確率**、$P(A)$, $P(B)$ を**周辺確率**という。

定義 5.2 (事象の独立性)**.** 確率空間 (Ω, P) の事象 $A, B \subseteq \Omega$ について

$$P(A \cap B) = P(A)P(B)$$

を満たすとき、**事象** A **と** B **は独立である**という。

例 5.1 (二回のコイン投げ)**.** 標本空間を $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$ とし、確率測度を $P(A) = \frac{|A|}{4} \quad \forall A \subseteq \Omega$ とする。このとき、$A = \{\mathrm{HH}, \mathrm{HT}\}$, $B = \{\mathrm{HH}, \mathrm{TH}\}$ とおくと、

$$P(A \cap B) = \frac{1}{4}, \qquad P(A) = P(B) = \frac{1}{2}$$

より $P(A \cap B) = P(A)P(B)$ を満たすことが分かる。よって事象 A, B は独立である。

事象 A と B が独立であり、$P(B) > 0$ であるとき、$P(A \mid B) = P(A)$ である。

5.2 確率変数の条件付き確率と独立性

事象は確率変数を通じて表すことが多い。そのため確率変数を用いた条件付き確率も定義する。

定義 5.3 (確率変数)**.** 確率空間 (Ω, P) 上の確率変数 X_1, X_2 について同時確率を

$$\Pr(X_1 \in A,\, X_2 \in B) \coloneqq P(\{\omega \in \Omega \mid X_1(\omega) \in A\} \cap \{\omega \in \Omega \mid X_2(\omega) \in B\}\}) \quad \forall A, B \subseteq \mathbb{R}$$

と定義する。確率変数が三つ以上の場合も同様に定義する。また $\Pr(X_2 \in B) > 0$ のとき、条件付き確率は以下で定義する。

$$\begin{aligned}\Pr(X_1 \in A \mid X_2 \in B) &:= P(\{\omega \in \Omega \mid X_1(\omega) \in A\} \mid \{\omega \in \Omega \mid X_2(\omega) \in B\}) \\ &= \frac{\Pr(X_1 \in A, X_2 \in B)}{\Pr(X_2 \in B)} \qquad \forall A, B \subseteq \mathbb{R}.\end{aligned}$$

任意の $A, B \subseteq \mathbb{R}$ について、

$$\Pr(X_1 \in A,\, X_2 \in B) = \Pr(X_1 \in A)\Pr(X_2 \in B)$$

を満たすとき、**確率変数 X_1 と X_2 は独立である**という。

補題 5.1. 確率空間 (Ω, P) 上の確率変数 X_1, X_2 が独立であるとする。このとき任意の関数 $f_1, f_2 \colon \mathbb{R} \to \mathbb{R}$ について、$f_1(X_1), f_2(X_2)$ は独立である。

証明.

$$\begin{aligned}\Pr(f_1(X_1) \in A,\, f_2(X_2) \in B) &= \Pr(X_1 \in f_1^{-1}(A),\, X_2 \in f_2^{-1}(B)) \\ &= \Pr(X_1 \in f_1^{-1}(A))\Pr(X_2 \in f_2^{-1}(B)) \\ &= \Pr(f_1(X_1) \in A)\Pr(f_2(X_2) \in B).\end{aligned}$$

□

二つ以上の確率変数の累積分布関数を

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) := \Pr(X_1 \le x_1, \dots, X_n \le x_n)$$

と定義する。

5.3 離散型確率変数と確率質量関数

定義 5.4. 確率空間 (Ω, P) 上の離散型確率変数 X_1, X_2 について、**同時確率質量関数**を

$$f_{X_1, X_2}(x_1, x_2) := \Pr(X_1 = x_1,\, X_2 = x_2)$$

と定義する。

同時確率質量関数からそれぞれの確率変数の確率質量関数が得られる。

$$\begin{aligned}f_{X_1}(x_1) &= \sum_{x_2} f_{X_1, X_2}(x_1, x_2) \\ f_{X_2}(x_2) &= \sum_{x_1} f_{X_1, X_2}(x_1, x_2)\end{aligned}$$

それぞれの確率変数の確率質量関数を**周辺質量関数**と呼ぶ。同時確率質量関数から周辺質量関数を計算する操作のことを**周辺化**という。

定義 5.5. 確率空間 (Ω, P) 上の離散型確率変数 X_1, X_2 について、**条件付き確率質量関数**を

$$f_{X_1|X_2}(x_1 \mid x_2) \coloneqq \Pr(X_1 = x_1 \mid X_2 = x_2)$$

と定義する。

条件付き確率の定義より

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1|X_2}(x_1 \mid x_2) f_{X_2}(x_2)$$

が成り立つ。

補題 5.2. 確率空間 (Ω, P) 上の**離散型**確率変数 X_1, X_2 について、X_1 と X_2 が独立 $\iff$

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \qquad \forall x_1 \in \mathrm{Image}(X_1), x_2 \in \mathrm{Image}(X_2).$$

証明. $\Longrightarrow$ は自明。$\Longleftarrow$ を示す。

$$\begin{aligned}
\Pr(X_1 \in A, X_2 \in B) &= \sum_{x_1 \in A} \sum_{x_2 \in B} \Pr(X_1 = x_1, X_2 = x_2) \\
&= \sum_{x_1 \in A} \sum_{x_2 \in B} \Pr(X_1 = x_1) \Pr(X_2 = x_2) \\
&= \Pr(X_1 \in A) \Pr(X_2 \in B)
\end{aligned}$$

□

例 5.2 (二回のコイン投げ)**.** 標本空間を $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$ とし、確率測度を $P(A) = \frac{|A|}{4} \quad \forall A \subseteq \Omega$ とする。

$$\begin{aligned}
X_1(\mathrm{HH}) = X_1(\mathrm{HT}) = 1, \qquad & X_1(\mathrm{TH}) = X_1(\mathrm{TT}) = 0 \\
X_2(\mathrm{HH}) = X_2(\mathrm{TH}) = 1, \qquad & X_2(\mathrm{HT}) = X_2(\mathrm{TT}) = 0
\end{aligned}$$

と定義する。このとき、

$$\begin{aligned}
f_{X_1, X_2}(x_1, x_2) &= \frac{1}{4} \qquad \forall x_1, x_2 \in \{0, 1\} \\
f_{X_1}(x_1) = f_{X_2}(x_2) &= \frac{1}{2} \qquad \forall x_1, x_2 \in \{0, 1\}
\end{aligned}$$

より $f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ を満たすことが分かる。よって確率変数 X_1, X_2 は独立である。

5.4 連続型確率変数と確率密度関数

確率空間 (Ω, P) 上の連続型確率変数 X_1, X_2 について、**同時確率密度関数**を

$$\Pr(X_1 \in A, X_2 \in B) = \int_A \left(\int_B f_{X_1, X_2}(x_1, x_2) \mathrm{d}x_2 \right) \mathrm{d}x_1$$

を満たすものと定義する。X_1 と X_2 が確率密度関数を持つ場合でも X_1 と X_2 の同時確率密度関数が存在するとは限らない。例えば $X_1 = X_2$ の場合がその例である。逆に X_1 と X_2 が同時確率密度関数を持つとき、それぞれの確率密度関数は

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2)\mathrm{d}x_2$$
$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2)\mathrm{d}x_1$$

で得られる。この操作を確率密度関数の**周辺化**という。同時確率密度関数を持つ確率変数 X_1, X_2 が独立であるとき、

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

が成り立つ。

5.5 三つ以上の独立確率変数

三つ以上の確率変数についても同時確率質量関数、同時確率密度関数を同様に定義する。独立性についても同様に定義する。

定義 5.6. 確率空間 (Ω, P) 上の確率変数 $X_1, X_2, \dots, X_n$ が**独立** $\overset{\text{def}}{\iff}$

$$\Pr(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{k=1}^{n} \Pr(X_k \in A_k) \qquad \forall A_1, \dots, A_n \subseteq \mathbb{R}.$$

定義 5.7. 確率空間 (Ω, P) 上の確率変数 $X_1, X_2, \dots, X_n$ が**互いに独立** $\overset{\text{def}}{\iff}$ 任意の $1 \leq i < j \leq n$ について、X_i と X_j が独立。

確率変数 $X_1, \dots, X_n$ が独立であるとき、それらは互いに独立であることは以下から分かる。

$$\begin{aligned}\Pr(X_1 \in A,\, X_2 \in B) &= \Pr(X_1 \in A,\, X_2 \in B,\, X_3 \in \mathbb{R}, \dots, X_n \in \mathbb{R}) \\ &= \Pr(X_1 \in A) \Pr(X_2 \in B) \prod_{k=3}^{n} \Pr(X_k \in \mathbb{R}) \\ &= \Pr(X_1 \in A) \Pr(X_2 \in B).\end{aligned}$$

一方で確率変数 $X_1, \dots, X_n$ が互いに独立であっても、それらが独立であるとは限らない。例えば、離散型確率変数 $X_1, \dots, X_n$ を

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \begin{cases} \frac{1}{2^{n-1}} & \text{if } \sum_{k=1}^{n} x_k \text{ is even} \\ 0 & \text{otherwise} \end{cases} \qquad \forall x_1, \dots, x_n \in \{0, 1\}$$

と定義する。このとき、確率変数 X_n を周辺化すると

$$\begin{aligned}f_{X_1, \dots, X_{n-1}}(x_1, \dots, x_{n-1}) &= f_{X_1, \dots, X_n}(x_1, \dots, x_{n-1}, 0) + f_{X_1, \dots, X_n}(x_1, \dots, x_{n-1}, 1) \\ &= \frac{1}{2^{n-1}}\end{aligned}$$

となる。つまり、$X_1, \dots, X_{n-1}$ は $\{0,1\}^{n-1}$ 上の一様分布に従う。この確率分布は $X_1, \dots, X_n$ について対称なので、どの確率変数を周辺化しても一様分布に従う。一様分布は独立なので、$n \geq 3$ のとき、どの二つの確率変数も独立である。

一方で、$n \geq 2$ のとき、これらの確率変数の周辺確率は一様である。つまり、

$$f_{X_k}(0) = f_{X_k}(1) = \frac{1}{2} \qquad \forall k = 1, 2, \dots, n.$$

しかし、

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_{k=1}^{n} f_{X_k}(x_k) \qquad \forall x_1, \dots, x_n \in \{0,1\}$$

は成り立たないので独立ではない。

5.6 独立な離散型確率変数の和

離散型確率変数 X_1 と X_2 が独立であるとする。このとき、$X_1 + X_2$ の確率質量関数は

$$f_{X_1+X_2}(x) = \sum_z f_{X_1}(z) f_{X_2}(x - z)$$

で与えられる。これを確率質量関数の**畳み込み**という。

例 5.3. 二項分布の $N = 1$ の場合をベルヌーイ分布 $\mathrm{Ber}(p)$ と呼ぶ。つまり、$X \sim \mathrm{Ber}(p) \overset{\text{def}}{\iff}$

$$\Pr(X = 0) = 1 - p, \qquad \Pr(X = 1) = p$$

である。確率変数 X_1, X_2 が独立で $\mathrm{Ber}(p)$ に従うとする。このとき、$X_1 + X_2$ は $\mathrm{Binom}(2, p)$ に従う。

$$\begin{aligned}
\Pr(X_1 + X_2 = 0) &= \Pr(X_1 = 0,\, X_2 = 0) \\
&= \Pr(X_1 = 0) \Pr(X_2 = 0) = (1-p)^2 \\
\Pr(X_1 + X_2 = 1) &= \Pr(X_1 = 0,\, X_2 = 1) + \Pr(X_1 = 1,\, X_2 = 0) \\
&= \Pr(X_1 = 0) \Pr(X_2 = 1) + \Pr(X_1 = 1) \Pr(X_2 = 0) \\
&= 2p(1-p) \\
\Pr(X_1 + X_2 = 2) &= \Pr(X_1 = 1,\, X_2 = 1) \\
&= \Pr(X_1 = 1) \Pr(X_2 = 1) = p^2
\end{aligned}$$

と計算できる。よって

$$\Pr(X_1 + X_2 = k) = \binom{2}{k} p^k (1-p)^{2-k}$$

が成り立ち $X_1 + X_2 \sim \mathrm{Binom}(2, p)$ であることが分かる。一般に、独立確率変数 X_1, X_2 について $X_1 \sim$

$\mathrm{Binom}(n,p)$, $X_2 \sim \mathrm{Binom}(m,p)$ のとき、$X_1 + X_2 \sim \mathrm{Binom}(n+m,p)$ である。

$$\begin{aligned}
\Pr(X_1 + X_2 = k) &= \sum_{\ell \geq 0} \Pr(X_1 = \ell) \Pr(X_2 = k - \ell) \\
&= \sum_{\ell \geq 0} \binom{n}{\ell} p^{\ell} (1-p)^{n-\ell} \binom{m}{k-\ell} p^{k-\ell} (1-p)^{m-k+\ell} \\
&= \left(\sum_{\ell \geq 0} \binom{n}{\ell} \binom{m}{k-\ell} \right) p^k (1-p)^{n+m-k} \\
&= \binom{n+m}{k} p^k (1-p)^{n+m-k}.
\end{aligned}$$

よって、$X_1, \ldots, X_n$ が独立で $\mathrm{Ber}(p)$ に従うとき、$X_1 + \cdots + X_n$ は $\mathrm{Binom}(n,p)$ に従う。

例 5.4. 確率変数 X_1, X_2 が独立で $X_1 \sim \mathrm{Poisson}(\lambda_1)$, $X_2 \sim \mathrm{Poisson}(\lambda_2)$ とする。このとき、

$$\begin{aligned}
f_{X_1+X_2}(x) &= \sum_{z=0}^{x} f_{X_1}(z) f_{X_2}(x-z) \\
&= \sum_{z=0}^{x} \frac{\lambda_1^z}{z!} \mathrm{e}^{-\lambda_1} \frac{\lambda_2^{x-z}}{(x-z)!} \mathrm{e}^{-\lambda_2} \\
&= \frac{1}{x!} \mathrm{e}^{-(\lambda_1+\lambda_2)} \sum_{z=0}^{x} \binom{x}{z} \lambda_1^z \lambda_2^{x-z} \\
&= \frac{1}{x!} \mathrm{e}^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^x
\end{aligned}$$

が成り立つ。よって $X_1 + X_2 \sim \mathrm{Poisson}(\lambda_1 + \lambda_2)$ である。独立なポアソン分布の和はポアソン分布に従う。

これらの例のように、確率分布の族 (集合) が畳み込みに閉じているとき、確率分布の族は**再生性**を持つという。二項分布の場合はパラメータ p を固定したときに再生性を持つ。

5.7 独立な連続型確率変数の和

連続確率変数 X_1 と X_2 が独立で密度関数を持つとき、

$$\begin{aligned}
\Pr(X_1 + X_2 \in A) &= \int\int_{y+z \in A} f_{X_1}(y) f_{X_2}(z) \mathrm{d}y \mathrm{d}z \\
&= \int\int_{x \in A} f_{X_1}(y) f_{X_2}(x-y) \mathrm{d}y \mathrm{d}x \qquad (x = y + z) \\
&= \int_A \left(\int_{-\infty}^{\infty} f_{X_1}(y) f_{X_2}(x-y) \mathrm{d}y \right) \mathrm{d}x \qquad (x = y + z)
\end{aligned}$$

と表せるので、$X_1 + X_2$ は確率密度関数を持ち

$$f_{X_1+X_2}(x) = \int_{-\infty}^{\infty} f_{X_1}(z) f_{X_2}(x-z) \mathrm{d}z$$

とすることができる。確率変数 X_1, X_2 が独立で $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ とする。このとき、

$$
\begin{aligned}
f_{X_1+X_2}(x) &= \int_{-\infty}^{\infty} f_{X_1}(z) f_{X_2}(x-z) \mathrm{d}z \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \mathrm{e}^{-\frac{(z-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} \mathrm{e}^{-\frac{(x-z-\mu_2)^2}{2\sigma_2^2}} \mathrm{d}z \\
&= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{\sigma_2^2(z-\mu_1)^2+\sigma_1^2(x-z-\mu_2)^2}{2\sigma_1^2\sigma_2^2}} \mathrm{d}z \\
&= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{(\sigma_1^2+\sigma_2^2)\left(z-\frac{\sigma_2^2\mu_1+\sigma_1^2(x-\mu_2)}{\sigma_1^2+\sigma_2^2}\right)^2+\sigma_2^2\mu_1^2+\sigma_1^2(x-\mu_2)^2-\frac{(\sigma_2^2\mu_1+\sigma_1^2(x-\mu_2))^2}{\sigma_1^2+\sigma_2^2}}{2\sigma_1^2\sigma_2^2}} \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \mathrm{e}^{-\frac{\sigma_2^2\mu_1^2+\sigma_1^2(x-\mu_2)^2-\frac{(\sigma_2^2\mu_1+\sigma_1^2(x-\mu_2))^2}{\sigma_1^2+\sigma_2^2}}{2\sigma_1^2\sigma_2^2}} \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \mathrm{e}^{-\frac{(x-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}}
\end{aligned}
$$

が成り立つので $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ であることが分かる。よって正規分布は再生性を持つ。

第 6 章

期待値、分散、モーメント

6.1 期待値

定義 6.1 (期待値)**.** 離散型確率変数 X の**期待値**は

$$\begin{aligned}\mathbb{E}[X] &:= \sum_{x \in \mathrm{Image}(X)} x \Pr(X = x) \\ &= \sum_{x \in \mathrm{Image}(X)} x f_X(x)\end{aligned}$$

と定義される。ここで、**右辺の和が絶対収束しない場合は (適当な順番で和を取って収束したとしても) 期待値は定義されない**。

連続型確率変数 X が確率密度関数 f_X を持つとき、その**期待値**は

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x$$

と定義される。ただし、広義積分で上記の積分が存在する場合でも

$$\int_{-\infty}^{\infty} |x| f_X(x) \mathrm{d}x$$

が存在しない場合は期待値は定義されない。

連続型確率変数の期待値に関する様々な証明はルベーグ積分の知識を必要とするので本書では扱わない。以下、証明はすべて離散確率変数の場合に限って与える。

補題 6.1. 確率変数 X, Y について

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

である。また、X と Y が独立のとき、

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$

である。

証明.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_z z f_{X+Y}(z) \\
&= \sum_z z \sum_x f_{X,Y}(x, z-x) \\
&= \sum_{x,y} (x+y) f_{X,Y}(x,y) \qquad (y = z-x) \\
&= \sum_{x,y} x f_{X,Y}(x,y) + \sum_{x,y} y f_{X,Y}(x,y) \\
&= \sum_x x f_X(x) + \sum_y y f_Y(y) \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

X と Y が独立のとき、

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_z z f_{XY}(z) \\
&= \sum_z z \sum_{x\neq 0} f_{X,Y}(x, z/x) \\
&= \sum_z z \sum_{x\neq 0} f_X(x) f_Y(z/x) \\
&= \sum_{x\neq 0,\, y} xy f_X(x) f_Y(y) \qquad (y = z/x) \\
&= \sum_{x,y} xy f_X(x) f_Y(y) \\
&= \left(\sum_x x f_X(x)\right)\left(\sum_y y f_Y(y)\right) \\
&= \mathbb{E}[X]\,\mathbb{E}[Y]
\end{aligned}$$

□

補題 6.2 (Law of the unconscious statistician (LOTUS))**.** 任意の関数 $g\colon \mathbb{R} \to \mathbb{R}$ について、

1. X が離散型確率変数のとき、

$$\mathbb{E}[g(X)] = \sum_x g(x) f_X(x).$$

2. X が連続型確率変数で確率密度関数を持つとき、

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x.$$

証明. X を離散型確率変数とする。

$$\begin{aligned}
\mathbb{E}[g(X)] &= \sum_x x \Pr(g(X) = x) \\
&= \sum_x xP(\{\omega \in \Omega \mid g(X(\omega)) = x\}) \\
&= \sum_x xP\left(\bigcup_{y \in \mathrm{Image}(X):\ g(y)=x} \{\omega \in \Omega \mid X(\omega) = y\}\right) \\
&= \sum_x \sum_{y \in \mathrm{Image}(X):\ g(y)=x} xP(\{\omega \in \Omega \mid X(\omega) = y\}) \\
&= \sum_{y \in \mathrm{Image}(X)} g(y) f_X(y).
\end{aligned}$$

□

命題 6.1 (期待値の性質)**.** 任意の確率変数 X と $a \in \mathbb{R}$ について

$$\begin{aligned}
\mathbb{E}[X + a] &= \mathbb{E}[X] + a \\
\mathbb{E}[aX] &= a\mathbb{E}[X].
\end{aligned}$$

定理 6.1 (マルコフの不等式)**.** 任意の**非負**確率変数 X と $a > 0$ について

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

証明.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \mathrm{Image}(X)} f_X(x)x \\
&= \sum_{\mathrm{Image}(X):\ x \geq a} f_X(x)x + \sum_{\mathrm{Image}(X):\ x<a} f_X(x)x \\
&\geq \sum_{\mathrm{Image}(X):\ x \geq a} f_X(x)x \qquad (\Pr(X \geq 0) = 1) \\
&\geq \sum_{\mathrm{Image}(X):\ x \geq a} f_X(x)a \\
&= \Pr(X \geq a)a.
\end{aligned}$$

□

6.2 分散

定義 6.2 (分散)**.** 確率変数 X が期待値を持つとき、その**分散**を

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

と定義する。また、分散の平方根を**標準偏差**という。

確率変数 X が分散を持つとき、

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\end{aligned}$$

である。分散は定義より非負の値を取る。

命題 6.2 (分散の性質)**.** 任意の確率変数 X と $a \in \mathbb{R}$ について

$$\begin{aligned}\mathrm{Var}[X+a] &= \mathrm{Var}[X] \\ \mathrm{Var}[aX] &= a^2\mathrm{Var}[X].\end{aligned}$$

分散は直感的には期待値からのはずれ具合を表す値である。

定理 6.2 (チェビシェフの不等式)**.** 任意の確率変数 X と $a > 0$ について

$$\Pr(|X-\mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}.$$

証明.

$$\begin{aligned}\Pr(|X-\mathbb{E}[X]| \geq a) &= \Pr((X-\mathbb{E}[X])^2 \geq a^2) \\ &\leq \frac{\mathbb{E}\left[(X-\mathbb{E}[X])^2\right]}{a^2} = \frac{\mathrm{Var}[X]}{a^2}.\end{aligned}$$

□

補題 6.3 (互いに独立な確率変数の和)**.** 確率変数 $X_1, X_2, \ldots, X_n$ が互いに独立のとき

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n].$$

証明.

$$\begin{aligned}\mathrm{Var}[X_1 + \cdots + X_n] &= \mathbb{E}\left[((X_1 + \cdots + X_n) - \mathbb{E}[X_1 + \cdots + X_n])^2\right] \\ &= \mathbb{E}\left[((X_1 - \mathbb{E}[X_1]) + \cdots + (X_n - \mathbb{E}[X_n]))^2\right] \\ &= \mathbb{E}\left[\sum_i (X_i - \mathbb{E}[X_i])^2 + 2\sum_{i<j}(X_i - \mathbb{E}[X_i])\left(X_j - \mathbb{E}\left[X_j\right]\right)\right] \\ &= \sum_i \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right] + 2\sum_{i<j}\mathbb{E}\left[(X_i - \mathbb{E}[X_i])\left(X_j - \mathbb{E}\left[X_j\right]\right)\right] \\ &= \sum_i \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right] + 2\sum_{i<j}\mathbb{E}[X_i - \mathbb{E}[X_i]]\,\mathbb{E}\left[X_j - \mathbb{E}\left[X_j\right]\right] \\ &= \sum_i \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right] = \sum_i \mathrm{Var}[X_i].\end{aligned}$$

□

例 6.1. 互いに独立な確率変数 $X_1, \dots, X_n$ のそれぞれが確率変数 X と同分布であるとし、

$$Y \coloneqq \frac{1}{n}(X_1 + \cdots + X_n)$$

と定義する。このとき、

$$\mathbb{E}[Y] = \mathbb{E}[X]$$
$$\mathrm{Var}[Y] = \frac{1}{n}\mathrm{Var}[X]$$

である。互いに独立な確率変数の平均を取ると期待値は変わらず、分散は小さくなる。

例 6.2. 独立確率変数 $X_1, \dots, X_n \sim \mathrm{Ber}(1/2)$ について

$$Y_S \coloneqq \sum_{i \in S} X_i \mod 2 \qquad \forall S \subseteq \{1, 2, \dots, n\}$$

と定義すると、これらは互いに独立である。また、$S \neq \emptyset$ について $Y_S \sim \mathrm{Ber}(1/2)$ である。任意の関数 $g\colon \{0,1\} \to \mathbb{R}$ について、

$$Y \coloneqq \frac{1}{2^n - 1} \sum_{S \subseteq \{1,\dots,n\}\colon\ S \neq \emptyset} g(Y_S)$$

と定義すると、$X \sim \mathrm{Ber}(1/2)$ について、

$$\mathbb{E}[Y] = \mathbb{E}[g(X)]$$
$$\mathrm{Var}[Y] = \frac{1}{2^n - 1}\mathrm{Var}[g(X)].$$

6.3 共分散

定義 6.3 (共分散)**.** 確率変数 X, Y が期待値を持つとき、その**共分散**を

$$\begin{aligned}\mathrm{Cov}[X,\, Y] &\coloneqq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

と定義する。共分散がゼロである確率変数のペアを**無相関**であるという。

定義より、$\mathrm{Cov}[X,\, X] = \mathrm{Var}[X]$ であることが分かる。

命題 6.3. 独立確率変数 X, Y は無相関である。

逆に無相関であっても独立とは限らない。

例 6.3. 確率変数 X を

$$f_X(0) = f_X(+1) = f_X(-1) = \frac{1}{3}$$

を満たすものとし、$Y = X^2$ とする。このとき、

$$\begin{aligned}
\mathrm{Cov}\left[X,\, Y\right] &= \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \\
&= \mathbb{E}\left[X^3\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[X^2\right] \\
&= \mathbb{E}\left[X\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[X^2\right] \\
&= \mathbb{E}\left[X\right]\left(1 - \mathbb{E}\left[X^2\right]\right) \\
&= 0
\end{aligned}$$

なので、X と Y は無相関である。一方で

$$f_Y(0) = \frac{1}{3} \qquad f_Y(1) = \frac{2}{3}$$
$$f_{X,Y}(0,0) = \frac{1}{3} \qquad f_{X,Y}(1,1) = \frac{1}{3} \qquad f_{X,Y}(-1,1) = \frac{1}{3}$$

なので X と Y は独立ではない。

共分散は正の値も負の値も取り得る。大雑把に言うと、

- X と Y の共分散が正 $\iff$ X が大きいとき Y も大きい
- X と Y の共分散が負 $\iff$ X が大きいとき Y は小さい

という意味になる。

補題 6.4. 任意の確率変数 $X_1, \ldots, X_n$ について

$$\mathrm{Var}\left[\sum_i X_i\right] = \sum_i \mathrm{Var}\left[X_i\right] + 2\sum_{i<j} \mathrm{Cov}\left[X_i,\, X_j\right].$$

証明. 補題 6.3 の証明参照。 □

6.4 モーメントとモーメント母関数

定義 6.4 (モーメント)**.** 確率変数 X と正の整数 $n \geq 1$ について、

$$\mu_n(X) := \mathbb{E}\left[X^n\right]$$

を X の n 次モーメントという。

定義 6.5 (モーメント母関数 (積率母関数))**.** 確率変数 X について、

$$M_X(t) := \mathbb{E}\left[\mathrm{e}^{tX}\right] \qquad t \in \mathbb{R}$$

を X の**モーメント母関数**という。すべての $t \in \mathbb{R}$ で $M_X(t)$ が存在しない場合もある。また、

$$K_X(t) \coloneqq \log M_X(t)$$

を X の**キュムラント母関数**という。

今後は以下の補題を認めることにする。証明にはルベーグ積分の知識が必要である。

定理 6.3. 確率変数 X について、ある $\epsilon > 0$ が存在し、モーメント母関数 $M_X(t)$ が $t \in (-\epsilon, \epsilon)$ **で存在するとき、**

$$\begin{aligned} M_X(t) &= \sum_{n \geq 0} \frac{\mathbb{E}\left[X^n\right]}{n!} t^n \qquad \forall t \in (-\epsilon, \epsilon) \\ \mu_n(X) &= \left. \frac{\mathrm{d}^n M_X(t)}{\mathrm{d}t^n} \right|_{t=0}. \end{aligned}$$

証明. 前半の証明を与える。離散型確率変数 X について、

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[\mathrm{e}^{tX}\right] \\ &= \sum_x \mathrm{e}^{tx} f_X(x) \\ &= \sum_x \left(\sum_{n \geq 0} \frac{(tx)^n}{n!} \right) f_X(x) \end{aligned}$$

である。ここで、任意の $t \in (-\epsilon, \epsilon)$ について

$$\begin{aligned} &\sum_x \sum_{n \geq 0} \left| \frac{(tx)^n}{n!} f_X(x) \right| \\ &= \sum_x \sum_{n \geq 0} \frac{|tx|^n}{n!} f_X(x) \\ &= \sum_x \mathrm{e}^{|tx|} f_X(x) \\ &\leq \sum_x (\mathrm{e}^{tx} + \mathrm{e}^{-tx}) f_X(x) \\ &= \mathbb{E}\left[\mathrm{e}^{tX}\right] + \mathbb{E}\left[\mathrm{e}^{-tX}\right] \\ &= M_X(t) + M_X(-t) < \infty \end{aligned}$$

よって、無限和

$$\sum_x \sum_{n \geq 0} \frac{(tx)^n}{n!} f_X(x)$$

は任意の $t \in (-\epsilon, \epsilon)$ について**絶対収束する**。そのため、和の順序を変えても収束値は変化しない。よって、

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[\mathrm{e}^{tX}\right] \\ &= \sum_x \left(\sum_{n\geq 0} \frac{(tx)^n}{n!} \right) f_X(x) \\ &= \sum_{n\geq 0} \sum_x \frac{(tx)^n}{n!} f_X(x) \\ &= \sum_{n\geq 0} \frac{\sum_x x^n f_X(x)}{n!} t^n \\ &= \sum_{n\geq 0} \frac{\mathbb{E}\left[X^n\right]}{n!} t^n \qquad \forall t \in (-\epsilon, \epsilon). \end{aligned}$$

□

系 6.1. 確率変数 X について、ある $\epsilon > 0$ が存在し、モーメント母関数 $M_X(t)$ が $t \in (-\epsilon, \epsilon)$ **で存在するとき**、

$$\begin{aligned} \left.\frac{\mathrm{d}K_X(t)}{\mathrm{d}t}\right|_{t=0} &= \mathbb{E}\left[X\right] \\ \left.\frac{\mathrm{d}^2K_X(t)}{\mathrm{d}t^2}\right|_{t=0} &= \mathrm{Var}\left[X\right]. \end{aligned}$$

証明.

$$\begin{aligned} \left.\frac{\mathrm{d}K_X(t)}{\mathrm{d}t}\right|_{t=0} &= \left.\frac{M'_X(t)}{M_X(t)}\right|_{t=0} = \mathbb{E}\left[X\right] \\ \left.\frac{\mathrm{d}^2K_X(t)}{\mathrm{d}t^2}\right|_{t=0} &= \left.\frac{M''_X(t)M_X(t) - M'_X(t)^2}{M_X(t)^2}\right|_{t=0} = M''_X(0) - M'_X(0)^2 = \mathrm{Var}\left[X\right]. \end{aligned}$$

□

また、重要度は低くなるが、

$$\begin{aligned} \left.\frac{\mathrm{d}^3K_X(t)}{\mathrm{d}t^3}\right|_{t=0} &= \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^3\right] \\ \left.\frac{\mathrm{d}^4K_X(t)}{\mathrm{d}t^4}\right|_{t=0} &= \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^4\right] - 3\mathrm{Var}\left[X\right]^2 \end{aligned}$$

が成り立つ。一般に

$$\kappa_n(X) := \left.\frac{\mathrm{d}^nK_X(t)}{\mathrm{d}t^n}\right|_{t=0}$$

を X の n **次キュムラント** と呼ぶ。

定理 6.4. 確率変数 X と Y のモーメント母関数が 0 を含む開区間 $(-\epsilon, \epsilon)$ で存在し、それらが等しいとき、X の分布と Y の分布は等しい。

証明. $\mathrm{Image}(X)$ と $\mathrm{Image}(Y)$ が有限の場合に限って証明を与える (この場合はモーメント母関数は $\mathbb{R}$ 全体で存在するのだが)。

$$\{x_0, \dots, x_{N-1}\} \coloneqq \mathrm{Image}(X) \cup \mathrm{Image}(Y)$$

とする。

$$M_X(t) = \sum_{k=0}^{N-1} f_X(x_k) \mathrm{e}^{t x_k}$$
$$M_Y(t) = \sum_{k=0}^{N-1} f_Y(x_k) \mathrm{e}^{t x_k}$$

なので、

$$0 = M_X(t) - M_Y(t) = \sum_{k=0}^{N-1} (f_X(x_k) - f_Y(x_k)) \mathrm{e}^{t x_k} \qquad \forall t \in (-\epsilon, \epsilon)$$

各 $k \in \{0, 1, \dots, N-1\}$ について、$t_k \coloneqq \epsilon \frac{k}{N}$ とおくと、

$$\sum_{k=0}^{N-1} (f_X(x_k) - f_Y(x_k)) \mathrm{e}^{t_\ell x_k} = 0 \qquad \forall \ell \in \{0, 1, \dots, N-1\} \tag{6.1}$$

である。ここで、$N \times N$ 実行列 V を

$$V_{\ell k} = \mathrm{e}^{t_\ell x_k} = \mathrm{e}^{\frac{\epsilon x_k}{N} \ell} \qquad \forall k, \ell \in \{0, 1, \dots, N-1\}$$

とおく。この行列 V は Vandermonde 行列の転置であり正則なので、

$$\sum_{k=0}^{N-1} V_{\ell k} g_k = 0 \qquad \forall \ell \in \{0, 1, \dots, N-1\}$$
$$\implies g_k = 0 \qquad \forall k \in \{0, 1, \dots, N-1\}$$

よって、

$$f_X(x_k) = f_Y(x_k) \qquad \forall k \in \{0, 1, \dots, N-1\}$$

である。 □

定理 6.4 より、モーメント母関数には確率変数の分布のすべての情報が含まれていると言える。ただし、モーメント母関数は原点まわりで存在しないこともあるので、分布の情報をすべて含む関数としては**特性関数**

$$\varphi_X(t) \coloneqq \mathbb{E}\left[\mathrm{e}^{itX}\right] \qquad \forall t \in \mathbb{R}$$

の方が優秀である。特性関数は常に存在する。一方でモーメント母関数は確率の集中を示す文脈では中心的な役割を果たす。

第 II 部

統計学

第 7 章

ベイズ推定

統計的推論とは現実の推定問題を確率論に基づきモデル化し、誤り確率を最小化するように推論する方法論である。統計的推論には大きく分けて二種類の流派がある。

- ベイズ主義: 推論する対象の分布 (事前分布) を仮定する。
- 頻度主義: 推論する対象の分布 (事前分布) を仮定しない。

例えば統計的推論は以下のような問題に適用されている。

7.1 最大事後確率推定

データが取り得る値の集合を $\mathcal{X}$ とし、分布のパラメータの取り得る値の集合を Θ とする。簡単のため、$\mathcal{X}$ と Θ は高々可算集合とする。データ $x \in \mathcal{X}$ からパラメータ $\theta \in \Theta$ を推定する問題を考える。このとき、x と θ が何かしらの確率分布に従っていると仮定する。パラメータ θ に対する x の確率質量関数を $p(x \mid \theta)$ と表す。また、パラメータ θ の確率質量関数を $\pi(\theta)$ と表す。つまり、パラメータ $\theta \in \Theta$ とデータ $x \in \mathcal{X}$ が選ばれる確率は

$$\pi(\theta)p(x \mid \theta)$$

である。また、

$$p(x) = \sum_{\theta \in \Theta} \pi(\theta)p(x \mid \theta), \qquad p(\theta \mid x) := \frac{\pi(\theta)p(x \mid \theta)}{p(x)}$$

と定義する。ベイズ推定の文脈では

- $\pi(\theta)$: 事前確率
- $p(x \mid \theta)$: 尤度
- $p(\theta \mid x)$: 事後確率

と呼ぶ。

得られたデータ $x \in \mathcal{X}$ からパラメータ $\theta \in \Theta$ を推定する関数 $\hat{\theta}\colon \mathcal{X} \to \Theta$ を**推定量 (estimator)** もしくは推定関数という。推定量 $\hat{\theta}$ の誤り確率を

$$P_{\mathrm{err}}(\hat{\theta}) := \sum_{\theta \in \Theta} \pi(\theta) \sum_{x \in \mathcal{X}} p(x \mid \theta)\, \mathbb{I}\{\hat{\theta}(x) \neq \theta\}$$

と定義する。このとき、

$$\begin{aligned}
P_{\mathrm{err}}(\hat{\theta}) &= \sum_{\theta \in \Theta} \pi(\theta) \sum_{x \in \mathcal{X}} p(x \mid \theta)\, \mathbb{I}\{\hat{\theta}(x) \neq \theta\} \\
&= \sum_{\theta \in \Theta} \pi(\theta) \sum_{x \in \mathcal{X}} p(x \mid \theta)(1 - \mathbb{I}\{\hat{\theta}(x) = \theta\}) \\
&= 1 - \sum_{\theta \in \Theta} \pi(\theta) \sum_{x \in \mathcal{X}} p(x \mid \theta)\, \mathbb{I}\{\hat{\theta}(x) = \theta\} \\
&= 1 - \sum_{x \in \mathcal{X}} \sum_{\theta \in \Theta} \pi(\theta) p(x \mid \theta)\, \mathbb{I}\{\hat{\theta}(x) = \theta\} \\
&= 1 - \sum_{x \in \mathcal{X}} \pi\left(\hat{\theta}(x)\right) p\left(x \mid \hat{\theta}(x)\right) \\
&\geq 1 - \sum_{x \in \mathcal{X}} \max_{\theta \in \Theta} \pi\left(\theta\right) p\left(x \mid \theta\right)
\end{aligned}$$

と下から抑えることができ、

$$\begin{aligned}
\hat{\theta}_{\mathrm{MAP}}(x) &:= \arg\max_{\theta \in \Theta} p\left(\theta \mid x\right) \\
&= \arg\max_{\theta \in \Theta} p\left(\theta \mid x\right) p\left(x\right) \\
&= \arg\max_{\theta \in \Theta} \pi\left(\theta\right) p\left(x \mid \theta\right)
\end{aligned}$$

という推定量により等号が達成される。この推定量を最大事後確率 (maximum a posteriori; MAP) 推定量と呼ぶ。

7.2 最尤推定

MAP 推定は誤り確率を最小化する推定方法であるが、事前確率 $\pi(\theta)$ を仮定しないと用いることができない。一方で尤度を最大化する推定量

$$\hat{\theta}_{\mathrm{ML}}(x) := \arg\max_{\theta \in \Theta} p\left(x \mid \theta\right)$$

を最尤推定 (maximam a priori; ML) 量という。Θ が有限集合で、事前確率が一様分布 $\pi(\theta) = \frac{1}{|\Theta|}$ のとき、最尤推定は最大事後確率推定と一致する。

7.3 全変動距離

特にパラメータが二値である場合を考える。この章では $\Theta = \{0, 1\}$ とする。また、

$$p^{(0)}(x) := p(x \mid 0) \qquad\qquad p^{(1)}(x) := p(x \mid 1)$$

とする。このとき、

$$\begin{aligned} P_{\mathrm{err}}(\hat{\theta}_{\mathrm{MAP}}) &= 1 - \sum_{x \in \mathcal{X}} \max_{\theta \in \{0,1\}} \pi\left(\theta\right) p\left(x \mid \theta\right) \\ &= 1 - \sum_{x \in \mathcal{X}} \max\left\{\pi(0) p^{(0)}\left(x\right), \pi(1) p^{(1)}\left(x\right)\right\} \end{aligned}$$

である。ここで、

$$\begin{aligned} &\max\{a, b\} - \min\{a, b\} = |\, a - b \,| \\ &\max\{a, b\} + \min\{a, b\} = a + b \end{aligned} \qquad \forall a, b \in \mathbb{R}$$

であるので、

$$\max\{a, b\} = \frac{1}{2}(a + b + |\, a - b \,|) \qquad \forall a, b \in \mathbb{R}.$$

よって、

$$\max\left\{\pi(0) p^{(0)}\left(x\right), \pi(1) p^{(1)}\left(x\right)\right\} = \frac{1}{2}\left(p(x) + \left|\, \pi\left(0\right) p^{(0)}\left(x\right) - \pi\left(1\right) p^{(1)}\left(x\right) \,\right|\right)$$

よって、

$$\begin{aligned} P_{\mathrm{err}}(\hat{\theta}_{\mathrm{MAP}}) &= 1 - \sum_{x \in \mathcal{X}} \max\left\{\pi(0) p^{(0)}\left(x\right), \pi(1) p^{(1)}\left(x\right)\right\} \\ &= 1 - \sum_{x \in \mathcal{X}} \frac{1}{2}\left(p(x) + \left|\, \pi\left(0\right) p^{(0)}\left(x\right) - \pi\left(1\right) p^{(1)}\left(x\right) \,\right|\right) \\ &= 1 - \frac{1}{2}\left(1 + \sum_{x \in \mathcal{X}} \left|\, \pi\left(0\right) p^{(0)}\left(x\right) - \pi\left(1\right) p^{(1)}\left(x\right) \,\right|\right) \\ &= \frac{1}{2}\left(1 - \sum_{x \in \mathcal{X}} \left|\, \pi\left(0\right) p^{(0)}\left(x\right) - \pi\left(1\right) p^{(1)}\left(x\right) \,\right|\right) \end{aligned}$$

定義 7.1 (全変動距離)**.** 高々可算集合 $\mathcal{X}$ 上の関数 f について、

$$\|f\|_1 \coloneqq \sum_{x \in \mathcal{X}} |\, f(x) \,|$$

と定義する。また、$\mathcal{X}$ 上の確率質量関数 $p^{(0)}, p^{(1)}$ について、

$$d_{\mathrm{TV}}(p^{(0)}, p^{(1)}) \coloneqq \frac{1}{2}\left\|p^{(0)} - p^{(1)}\right\|_1$$

を $p^{(0)}$ と $p^{(1)}$ の**全変動距離**という。

これらの記法を用いると、

$$P_{\mathrm{err}}(\hat{\theta}_{\mathrm{MAP}}) = \frac{1}{2}\left(1 - \left\|\pi(0) p^{(0)} - \pi(1) p^{(1)}\right\|_1\right)$$

と表せる。また、$\pi(0) = \pi(1) = 1/2$ のとき、

$$P_{\mathrm{err}}(\hat{\theta}_{\mathrm{MAP}}) = \frac{1}{2}\left(1 - d_{\mathrm{TV}}(p^{(0)}, p^{(1)})\right)$$

である。

7.4 損失関数

パラメータが取り得る値の集合が実数の部分集合 $\Theta \subseteq \mathbb{R}$ であると仮定する。真のパラメータとその推定値の間の「誤差」を表す関数 $L\colon \Theta \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ を**損失関数**と呼ぶ。また、**期待損失** $R\colon \Theta \times (\mathcal{X} \to \mathbb{R}) \to \mathbb{R}_{\geq 0}$ を

$$\begin{aligned} R(\theta, \hat{\theta}) &:= \mathbb{E}\left[L(\theta, \hat{\theta}(X)) \mid \theta\right] \\ &= \sum_{x \in \mathcal{X}} L(\theta, \hat{\theta}(x)) p(x \mid \theta) \end{aligned}$$

と定義する。また、**ベイズリスク** $\rho\colon \mathcal{P}(\Theta) \times (\mathcal{X} \to \mathbb{R}) \to \mathbb{R}_{\geq 0}$ を

$$\begin{aligned} \rho(\pi, \hat{\theta}) &:= \mathbb{E}\left[L(\theta, \hat{\theta}(X))\right] \\ &= \sum_{\theta \in \Theta} \sum_{x \in \mathcal{X}} L(\theta, \hat{\theta}(x)) p(x \mid \theta) \pi(\theta) \end{aligned}$$

と定義する。$\Theta \subseteq \mathbb{R}$ が非可算無限集合の場合は $\pi(\theta)$ を確率密度関数とし、

$$\begin{aligned} \rho(\pi, \hat{\theta}) &:= \mathbb{E}\left[L(\theta, \hat{\theta}(X))\right] \\ &= \int \sum_{x \in \mathcal{X}} L(\theta, \hat{\theta}(x)) p(x \mid \theta) \pi(\theta) \mathrm{d}\theta \end{aligned}$$

と定義する。このとき、$p(x \mid \theta)$ は条件付き確率というより、$\theta \in \Theta$ というパラメータを持った確率質量関数 ($\theta \in \Theta$ から定まる確率質量関数) と理解すれば十分である。

例えば Θ が高々可算集合で、

$$L(\theta, \theta') = \mathbb{I}\{\theta \neq \theta'\} \qquad \forall \theta, \theta' \in \Theta$$

と定義すると、ベイズリスク $\rho(\pi, \hat{\theta})$ は推定量 $\hat{\theta}$ の誤り確率 $P_{\mathrm{err}}(\hat{\theta})$ である。

その他の重要な損失関数の例として二乗誤差がある。

$$L(\theta, \theta') = (\theta - \theta')^2.$$

損失関数 L を定めたときに、ベイズリスクを最小化する推定量

$$\begin{aligned} & \hat{\theta} = \arg\min_{\hat{\theta}} \rho(\pi, \hat{\theta}) \\ \iff & \hat{\theta}(x) = \arg\min_{\theta' \in \Theta} \sum_{\theta \in \Theta} L(\theta, \theta') p(x \mid \theta) \pi(\theta) \qquad \forall x \in \mathcal{X} \\ \iff & \hat{\theta}(x) = \arg\min_{\theta' \in \Theta} \sum_{\theta \in \Theta} L(\theta, \theta') p(\theta \mid x) p(x) \qquad \forall x \in \mathcal{X} \\ \iff & \hat{\theta}(x) = \arg\min_{\theta' \in \Theta} \sum_{\theta \in \Theta} L(\theta, \theta') p(\theta \mid x) \qquad \forall x \in \mathcal{X} \end{aligned}$$

ここで、

$$\mathbb{E}\left[L(\theta, \theta') \mid x\right] = \sum_{\theta \in \Theta} L(\theta, \theta') p(\theta \mid x) \qquad \forall x \in \mathcal{X}$$

を損失関数の事後平均という。各 $x \in \mathcal{X}$ について、$\theta' = \hat{\theta}(x)$ が損失関数の事後平均を最小化するとき、ベイズリスクを最小化する。

損失関数が $L(\theta, \theta')$ が各 $\theta \in \Theta$ を固定したときに θ' について凸関数であるとき、事後平均 $\mathbb{E}\left[L(\theta, \theta') \mid x\right]$ も θ' について凸関数となる (凸関数の非負倍は凸関数であり、凸関数の和は凸関数なので)。さらに、$L(\theta, \theta')$ が θ' について微分可能なとき、

$$\frac{\partial \mathbb{E}\left[L(\theta, \theta') \mid x\right]}{\partial \theta'} = \sum_{\theta \in \Theta} \frac{\partial L(\theta, \theta')}{\partial \theta'} p(\theta \mid x) = 0 \qquad \forall x \in \mathcal{X}$$

を満たす θ' を $\hat{\theta}(x)$ として選択するのが最適である。二乗誤差 $L(\theta, \theta') = (\theta - \theta')^2$ のとき、この条件は

$$0 = \sum_{\theta} 2(\theta' - \theta) p(\theta \mid x) = 2\left(\theta' - \sum_{\theta} \theta p(\theta \mid x)\right)$$

となり、

$$\hat{\theta}(x) = \sum_{\theta} \theta p(\theta \mid x)$$

とするのが最適であることが分かる。この右辺の値をパラメータ θ の事後平均という。

第 8 章

仮説検定

8.1 事前分布を仮定しない推定問題

ベイズ推定では推定したいパラメータ $\theta \in \Theta$ の事前分布 $\pi(\theta)$ を既知として仮定した。しかし、現実の問題ではこの事前分布を適切に仮定する方法がない場合もある。例えば

- 開発中の薬に効果があるかないか
- ある患者が病気かどうか
- サイコロに偏りがあるかどうか

といった問題について事前分布をどのように仮定するのが適切か不明瞭である。そのような状況でデータ $x \in \mathcal{X}$ から パラメータ $\theta \in \Theta$ を推定する問題を考える。尤度 $p(x \mid \theta)$ は既知とする。

8.2 伝統的な仮説検定

パラメータの集合を二つの部分集合 Θ_0 と Θ_1 に分割する。つまり、$\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$ である。そしてパラメータが Θ_0 に属するか Θ_1 に属するかを知りたいとする。このとき二つの命題

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

を仮説という。その二つの仮説のうちの**通常成り立っていると考える方**を H_0 とし (対応するパラメータ集合は Θ_0) **帰無仮説**と呼ぶ。また、そうでない方を H_1 とし (対応するパラメータ集合は Θ_1) **対立仮説**と呼ぶ。帰無仮説の例として

- 開発中の薬に効果はない
- ある患者が病気ではない
- サイコロに偏りはない

などがある。それらに対応する対立仮説はそれぞれ

- 開発中の薬に効果がある

- ある患者が病気である
- サイコロに偏りがある

となる。仮説検定の考え方では帰無仮説を棄却するかしないかを決める。帰無仮説を棄却した場合、対立仮説を正しいと考え、帰無仮説を棄却しなかった場合は何も言えないと結論づける。

8.3 単純仮説

仮説 Θ_0 と Θ_1 がそれぞれ一元集合であるとき、H_0 と H_1 を**単純仮説**という。$\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$ として、$p_0(x) \coloneqq p(x \mid \theta_0)$ と $p_1(x) \coloneqq p(x \mid \theta_1)$ とする。データから仮説を推定する関数 $E\colon \mathcal{X} \to [0,1]$ を**検定関数**という。各 $x \in \mathcal{X}$ について、$E(x)$ は**帰無仮説を棄却する確率**とする。この検定関数について二種類の誤り確率を

$$\alpha_E \coloneqq \mathbb{E}\left[E(X) \mid \theta_0\right] = \sum_x E(x)p(x \mid \theta_0)$$
$$\beta_E \coloneqq 1 - \mathbb{E}\left[E(X) \mid \theta_1\right] = 1 - \sum_x E(x)p(x \mid \theta_1)$$

と定義する。このとき、α_E は帰無仮説が正しいときに帰無仮説を棄却する確率であり、**第一種誤り確率**もしくは**有意水準**という。また、β_E は対立仮説が正しいときに帰無仮説を棄却しない確率であり、**第二種誤り確率**という。第一種誤り確率だけを小さくしたければ $E(x) = 0$ とすればよいし、第二種誤り確率だけを小さくしたければ $E(x) = 1$ とすればよい。

8.4 最強力検定

定義 8.1. 検定関数 $E\colon \mathcal{X} \to [0,1]$ が有意水準 $\alpha \in [0,1]$ の**最強力検定** $\overset{\text{def}}{\iff}$ $\alpha_E = \alpha$ であり、任意の $F\colon \mathcal{X} \to [0,1]$ について、$\alpha_F \leq \alpha$ ならば $\beta_F \geq \beta$ が成り立つ。

実現可能な誤り確率 (α, β) の集合

$$C = \{(\alpha_E, \beta_E) \mid E\colon \mathcal{X} \to [0,1]\}$$

について考える。自明に $(1,0)$ と $(0,1)$ は実現可能である。

この集合 C は凸集合である。

$$\alpha_{pE+(1-p)F} = p\alpha_E + (1-p)\alpha_F$$
$$\beta_{pE+(1-p)F} = p\beta_E + (1-p)\beta_F$$

であることから、

$$\left(\alpha_{pE+(1-p)F}, \quad \beta_{pE+(1-p)F}\right) = p\left(\alpha_E, \quad \beta_E\right) + (1-p)\left(\alpha_F, \quad \beta_F\right)$$

が確認できる。また、検定結果を反転させた検定関数 $1 - E(x)$ を考えると、

$$\left(\alpha_{1-E}, \quad \beta_{1-E}\right) = (1, \quad 1) - \left(\alpha_E, \quad \beta_E\right)$$

である。

まとめると、実現可能な (α, β) の集合 C は

1. $(1,0), (0,1) \in C$
2. C は 凸集合
3. $(\alpha, \beta) \in C \iff (1-\alpha, 1-\beta) \in C$

を満たす。

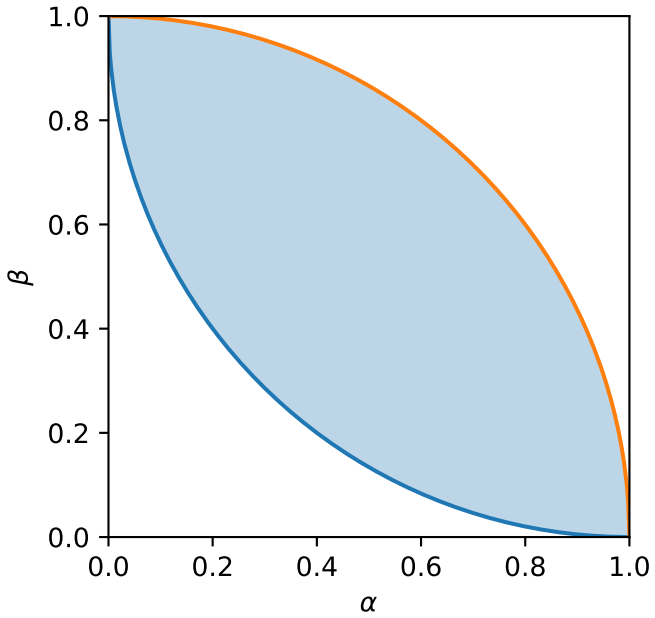

図 8.1: 実現可能な (α, β) の集合の例

各 $\alpha \in [0,1]$ について、β を最小化するのが最強力検定であるので、この図の下のカーブが最強力検定で実現される (α, β) となる。

8.5 尤度比検定

ベイズ推定の枠組みでは MAP 推定量が誤り確率を最小化する推定量であった。この MAP 推定量は

$$E_{\mathrm{MAP}}(x) = \begin{cases} 0 & \text{if } \frac{p(x|\theta_0)}{p(x|\theta_1)} \geq \frac{\pi(\theta_1)}{\pi(\theta_0)} \\ 1 & \text{otherwise} \end{cases}$$

と表すことができる。

一般的に $\eta > 0$, $\kappa \in [0,1]$ について

$$E(x) = \begin{cases} 0 & \text{if } \frac{p(x|\theta_0)}{p(x|\theta_1)} > \eta \\ 1 & \text{if } \frac{p(x|\theta_0)}{p(x|\theta_1)} < \eta \\ \kappa & \text{otherwise} \end{cases}$$

という形の検定関数を尤度比検定という。

補題 8.1 (ネイマン・ピアソンの補題)**.** 任意の尤度比検定は最強力検定である *(*逆も成り立つ*)*。

証明. 尤度比検定 E における尤度比の閾値を $\eta > 0$ とする。任意の $F\colon \mathcal{X} \to [0,1]$ について

$$\begin{aligned} & (F(x) - E(x))(p_0(x) - \eta p_1(x)) \geq 0 \qquad \forall x \in \mathcal{X} \\ \iff & F(x)p_0(x) - E(x)p_0(x) - \eta F(x)p_1(x) + \eta E(x)p_1(x) \geq 0 \qquad \forall x \in \mathcal{X} \\ \implies & \alpha_F - \alpha_E - \eta(1-\beta_F) + \eta(1-\beta_E) \geq 0 \\ \iff & (\alpha_F - \alpha_E) + \eta(\beta_F - \beta_E) \geq 0. \end{aligned}$$

よって $\alpha_F \leq \alpha_E$ ならば $\beta_F \geq \beta_E$ である。 □

データ $\mathcal{X}$ が連続な場合は確率質量関数の代わりに確率密度関数を考える。

例 8.1. コインを持っており、表が出る確率は p_0 か p_1 のどちらかである。コインを独立に N 回投げて表が出る確率が p_0 か p_1 かを推定したい。$\mathcal{X} = \{0,1\}^N$ とし、0 は裏、1 は表に対応するものとする。このとき尤度は

$$p(\mathbf{x} \mid p_k) = \prod_{i=1}^{N} p_k^{x_i}(1-p_k)^{1-x_i} \qquad \text{for } k \in \{0,1\}$$

である。このとき尤度比は

$$\begin{aligned} \frac{p(\mathbf{x} \mid p_0)}{p(\mathbf{x} \mid p_1)} &= \frac{\prod_i p_0^{x_i}(1-p_0)^{1-x_i}}{\prod_i p_1^{x_i}(1-p_1)^{1-x_i}} \\ &= \left(\frac{p_0}{p_1}\right)^{\sum_i x_i} \left(\frac{1-p_0}{1-p_1}\right)^{N-\sum_i x_i} \end{aligned}$$

である。よって尤度比は表が出た回数 $T(\mathbf{x}) = \sum_i x_i$ から定まる。なので $p_1 > p_0$ とすると、表が出た回数が多いときに $E(\mathbf{x}) = 1$ とすることになる。

第 III 部

漸近論

第 9 章

大数の法則と集中不等式

9.1 大数の弱法則

表が出る確率が 1/2 のコインを 100 回独立に投げたときに表が出る回数は大体 50 回くらいになるだろう。それを一般的な形で述べたものが大数の法則である。

定理 9.1 (大数の弱法則 (分散有限を仮定))**.** 確率変数 X が**分散を持つ**とする。確率変数 $X_1, \dots, X_N$ が独立同分布で X と同じ分布に従うとする。このとき、任意の $\epsilon > 0$ について

$$\lim_{N\to\infty} \Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}[X]\right| \geq \epsilon\right) = 0$$

が成り立つ。

証明.

$$\begin{aligned}
\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{N} X_i - \mathbb{E}[X]\right| \geq \epsilon\right) &= \Pr\left(\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}[X]\right)^2 \geq \epsilon^2\right) \\
&= \Pr\left(\left(\sum_{i=1}^{N} X_i - N\mathbb{E}[X]\right)^2 \geq \epsilon^2 N^2\right) \\
&\leq \frac{N\mathrm{Var}[X]}{\epsilon^2 N^2} = \frac{\mathrm{Var}[X]}{\epsilon^2 N} \longrightarrow 0.
\end{aligned}$$

□

9.2 チェルノフ上界

上記の大数の弱法則の証明では確率が 0 に収束するスピートは $O(1/N)$ であった。より詳しく確率が 0 にいくスピードを解析しよう。

補題 9.1 (チェルノフ上界)**.** 任意の確率変数 X と $a \in \mathbb{R}$ について、

$$\Pr(X \geq a) \leq \frac{M_X(t)}{\mathrm{e}^{at}} = \mathrm{e}^{K_X(t)-at} \qquad \forall t \geq 0$$
$$\Pr(X \leq a) \leq \frac{M_X(t)}{\mathrm{e}^{at}} = \mathrm{e}^{K_X(t)-at} \qquad \forall t \leq 0.$$

証明. $t = 0$ のときは不等式の右辺は 1 となるので、不等式は自明に成り立つ。任意の $t > 0$ について、

$$\begin{aligned}\Pr(X \geq a) &= \Pr\left(\mathrm{e}^{tX} \geq \mathrm{e}^{ta}\right) \\ &\leq \frac{M_X(t)}{\mathrm{e}^{at}} \qquad (\text{マルコフの不等式}).\end{aligned}$$

が成り立つ。もう一つの不等式も同様に示すことができる。 □

マルコフの不等式は非負の確率変数にしか適用できないが、チェルノフ上界は任意の確率変数に適用できる。

補題 9.2 (確率変数の和に対するチェルノフ上界)**.** 任意の確率変数 X と $a \in \mathbb{R}$ について、

$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N} X_i \geq a\right) \leq \mathrm{e}^{-(at-K_X(t))N} \qquad \forall t \geq 0$$
$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N} X_i \leq a\right) \leq \mathrm{e}^{-(at-K_X(t))N} \qquad \forall t \leq 0.$$

証明.

$$\begin{aligned}\Pr\left(\frac{1}{N}\sum_{i=1}^{N} X_i \geq a\right) &= \Pr\left(\sum_{i=1}^{N} X_i \geq aN\right) \\ &\leq \mathrm{e}^{K_{\sum_i X_i}(t)-atN} \qquad (\text{チェルノフ上界}) \\ &= \mathrm{e}^{(K_X(t)-at)N} \qquad \left(K_{\sum_i X_i}(t) = \sum_i K_{X_i}(t) = NK_X(t)\right).\end{aligned}$$

□

このようにチェルノフ上界を使うと N について指数関数の上界が得られる。係数 $at - K_X(t)$ が正であれば、確率は指数関数的に小さいことになる。ここで、最適な t を選ぶことで、この係数 $at - K_X(t)$ を最大化することを考える。

9.3 キュムラント母関数の性質

キュムラント母関数

$$K_X(t) = \log M_X(t) = \log \mathbb{E}\left[\mathrm{e}^{tX}\right]$$

の性質を改めて考えよう。発散する場合は $+\infty$ に値を取るとみなして $K_X \colon \mathbb{R} \to (-\infty, +\infty]$ と考えることにする。

まず、$K_X(0) = 0$ である。

ある $t > 0$ について、$M_X(t) < +\infty$ と仮定すると、任意の $s \in (0, t)$ について

$$\begin{aligned} M_X(s) &= \mathbb{E}\left[\mathrm{e}^{sX}\right] \\ &= \mathbb{E}\left[\mathrm{e}^{sX}\mathbb{1}_{\{X\geq 0\}}\right] + \mathbb{E}\left[\mathrm{e}^{sX}\mathbb{1}_{\{X<0\}}\right] \\ &\leq \mathbb{E}\left[\mathrm{e}^{tX}\mathbb{1}_{\{X\geq 0\}}\right] + 1 \\ &\leq M_X(t) + 1 < +\infty \end{aligned}$$

である。同様に、ある $t < 0$ について、$M_X(t) < +\infty$ と仮定すると、任意の $s \in (t, 0)$ について $M_X(s) < +\infty$ である。よって、$M_X(t)$ **や** $K_X(t)$ **が有限となる範囲は 0 を含む区間**となる。ここでいう区間とは一般的に空集合、もしくは $a < b$ について、

$$[a,\, a] \quad (a,\, b) \quad [a,\, b) \quad (a,\, b] \quad [a,\, b] \quad (a,\, +\infty) \quad [a,\, +\infty) \quad (-\infty,\, b) \quad (-\infty,\, b] \quad (-\infty,\, +\infty)$$

のいずれかの形の集合を指す。この区間を

$$\mathrm{dom}(K_X) \coloneqq \{t \in \mathbb{R} \mid K_X(t) < +\infty\}$$

と表す。区間は 1 次元の凸集合と一言で理解できる。

証明はしないが、キュムラント母関数 $K_X(t)$ は $\mathrm{dom}(K_X)$ の**内点で何回でも微分可能であり、無限和や積分を取る前に微分しても構わない**。

$$\begin{aligned} \frac{\mathrm{d}K_X(t)}{\mathrm{d}t} &= \frac{\mathbb{E}\left[X\mathrm{e}^{tX}\right]}{\mathbb{E}\left[\mathrm{e}^{tX}\right]} \\ \frac{\mathrm{d}^2K_X(t)}{\mathrm{d}t^2} &= \frac{\mathbb{E}\left[X^2\mathrm{e}^{tX}\right]\mathbb{E}\left[\mathrm{e}^{tX}\right] - \mathbb{E}\left[X\mathrm{e}^{tX}\right]^2}{\mathbb{E}\left[\mathrm{e}^{tX}\right]^2} \\ &= \frac{\mathbb{E}\left[X^2\mathrm{e}^{tX}\right]}{\mathbb{E}\left[\mathrm{e}^{tX}\right]} - \left(\frac{\mathbb{E}\left[X\mathrm{e}^{tX}\right]}{\mathbb{E}\left[\mathrm{e}^{tX}\right]}\right)^2 \\ &= \frac{\mathbb{E}\left[\left(X - \frac{\mathbb{E}[X\mathrm{e}^{tX}]}{\mathbb{E}[\mathrm{e}^{tX}]}\right)^2 \mathrm{e}^{tX}\right]}{\mathbb{E}\left[\mathrm{e}^{tX}\right]} \geq 0 \end{aligned}$$

ここで確率変数 Z_t を導入し、確率質量関数

$$f_{Z_t}(x) = \frac{f_X(x)\mathrm{e}^{tx}}{\sum_x f_X(x)\mathrm{e}^{tx}}$$

を持つものとすると、

$$\begin{aligned} \frac{\mathrm{d}K_X(t)}{\mathrm{d}t} &= \mathbb{E}\left[Z_t\right] \\ \frac{\mathrm{d}^2K_X(t)}{\mathrm{d}t^2} &= \mathrm{Var}\left[Z_t\right] \end{aligned}$$

であることが分かる。また、X が決定的 ($\Pr(X = \mathbb{E}[X]) = 1$) でない限り、$K_X$ は $\mathrm{dom}(K_X)$ で狭義凸である。

よって X のキュムラント母関数 $K_X(t)$ が原点付近で存在すると仮定すると、$K_X(t)$ は

- 0 を含む区間で定義され、
- 原点を通り、
- 凸関数で、
- 原点の傾きは $\mathbb{E}[X]$

であることが分かる。

> **ノート**
>
> 関数 $f\colon \mathbb{R} \to (-\infty, +\infty]$ が下半連続であるとは、任意の $\alpha \in \mathbb{R}$ について $\{t \in \mathbb{R} \mid f(t) \leq \alpha\}$ が閉集合であることをいう。キュムラント母関数は下半連続の凸関数である。

また、キュムラント母関数の簡単な性質として以下が成り立つ。任意の独立確率変数 X と Y と $a \in \mathbb{R}$ について

$$\begin{aligned} K_{X+a}(t) &= \log \mathbb{E}\left[\mathrm{e}^{t(X+a)}\right] = \log\left(\mathbb{E}\left[\mathrm{e}^{tX}\right] \cdot \mathrm{e}^{ta}\right) = K_X(t) + at \\ K_{aX}(t) &= \log \mathbb{E}\left[\mathrm{e}^{t(aX)}\right] = K_X(at) \\ K_{X+Y}(t) &= \log \mathbb{E}\left[\mathrm{e}^{t(X+Y)}\right] = \log\left(\mathbb{E}\left[\mathrm{e}^{tX}\right]\mathbb{E}\left[\mathrm{e}^{tY}\right]\right) = K_X(t) + K_Y(t). \end{aligned}$$

9.4 キュムラント母関数の例

9.5 ルジャンドル変換

チェルノフ上界に現れる係数 $at - K_X(t)$ の最大化はルジャンドル変換を用いて $K_X^*(a)$ と表せる。ルジャンドル変換を定義する際は $+\infty$ という値を許して $(-\infty, +\infty]$ を値域として考えると都合がよい。このように $+\infty$ を値として許した場合にも凸性を通常の関数と同じように定義する。一般に区間上で定義された凸関数を実数全体に拡張し、元の定義域の外で $+\infty$ を取ることにするとやはり凸関数になる。また、$f\colon \mathbb{R} \to (-\infty, +\infty]$ の**有効領域**を

$$\mathrm{dom}(f) := \{x \in \mathbb{R} \mid f(x) < +\infty\}$$

と定義する。凸関数の有効領域は区間になる。

定義 9.1 (ルジャンドル変換)**.**

恒等的に $+\infty$ ではない**凸関数** $f\colon \mathbb{R} \to (-\infty, +\infty]$ のルジャンドル変換 $f^*\colon \mathbb{R} \to (-\infty, +\infty]$ を以下で定義する。

$$f^*(a) := \sup_{t \in \mathbb{R}} \{at - f(t)\}.$$

ルジャンドル変換において f は凸関数なので、$at - f(t)$ は凹関数 (上に凸の関数) になる。簡単のため f が $\mathrm{dom}(f)$ の内点で微分可能であると仮定しよう (キュムラント母関数はこの仮定は満たす)。このとき、

- $at - f(t)$ の微分が 0 になる点、つまり $f'(t_a) = a$ を満たす $t_a \in \mathrm{dom}(f)^\circ$ が存在するとき、t_a で $at - f(t)$ は最大化される。

- そのような $t_a \in \mathrm{dom}(f)^\circ$ が存在しないときは、$\mathrm{dom}(f)$ の端への極限で sup が達成される。

凸関数は $\mathrm{dom}(f)$ の上では接線の集合で表すことができる。接線 $ax+b$ は傾き a と切片 b のペアで表すことができる。傾き a を持つ f の接線は

$$a(x-t_a)+f(t_a) = ax-(at_a-f(t_a)) = ax-f^*(a)$$

この傾き a から接線 $ax+b$ の切片の -1 倍である $-b$ への関数が f^* である。

例えば f が 0 で微分可能であるとき

$$f^*(f'(0)) = 0$$

である。

また、ルジャンドル変換 f^* は凸関数である。

補題 9.3. 凸関数 $f\colon \mathbb{R} \to (-\infty, +\infty]$ のルジャンドル変換 $f^*\colon \mathbb{R} \to (-\infty, +\infty]$ は凸関数である。

証明. 任意の $a_1, a_2 \in \mathbb{R}$ と $\lambda \in [0,1]$ について

$$\begin{aligned} f^*(\lambda a_1 + (1-\lambda)a_2) &= \sup_{t\in\mathbb{R}} \{(\lambda a_1 + (1-\lambda)a_2)t - f(t)\} \\ &= \sup_{t\in\mathbb{R}} \{\lambda(a_1 t - f(t)) + (1-\lambda)(a_2 t - f(t))\} \\ &\le \sup_{t\in\mathbb{R}} \{\lambda(a_1 t - f(t))\} + \sup_{t\in\mathbb{R}} \{(1-\lambda)(a_2 t - f(t))\} \\ &\le \lambda f^*(a_1) + (1-\lambda) f^*(a_2). \end{aligned}$$

□

ルジャンドル変換を凸でない関数 f についても同様に定義した場合でも、ルジャンドル変換 f^* は同様に凸関数となる。

> **ノート**
>
> またルジャンドル変換 f^* は下半連続である。ルジャンドル変換は凸で下半連続な関数を凸で下半連続な関数に写す。また、凸で下半連続な関数 f について $f = f^{**}$ である。

9.6 クラメールの定理

補題 9.4 (イェンセンの不等式)**.** 任意の凸関数 $f\colon \mathbb{R} \to \mathbb{R}$ と確率変数 X について

$$\mathbb{E}[f(X)] \ge f(\mathbb{E}[X]).$$

X の像が有限のときの証明. 確率変数 X は $k = 1, 2, \dots, m$ について確率 p_k で値 a_k をとると仮定する。m についての帰納法で示す。$m = 1$ のときは明らかに成り立つ。X の像のサイズが m 未満のときにイェンセンの不等式が成り立つと仮定すると、

$$\begin{aligned}
\mathbb{E}\left[f(X)\right] &= \sum_{k=1}^{m} p_k f(a_k) = \left(\sum_{k=1}^{m-1} p_k\right) \sum_{k=1}^{m-1} \frac{p_k}{\sum_{\ell=1}^{m-1} p_\ell} f(a_k) + p_m f(a_m) \\
&\geq \left(\sum_{k=1}^{m-1} p_k\right) f\left(\sum_{k=1}^{m-1} \frac{p_k}{\sum_{\ell=1}^{m-1} p_\ell} a_k\right) + p_m f(a_m) \quad \text{(帰納法の仮定)} \\
&\geq f\left(\left(\sum_{k=1}^{m-1} p_k\right) \sum_{k=1}^{m-1} \frac{p_k}{\sum_{\ell=1}^{m-1} p_\ell} a_k + p_m a_m\right) \quad (f \text{ の凸性}) \\
&= f\left(\sum_{k=1}^{m} p_k a_k\right) = f(\mathbb{E}\left[X\right]).
\end{aligned}$$

□

補題 9.2 の右辺を t について最小化することを考えよう。

補題 9.5. レート関数 $I_X \colon \mathbb{R} \to [0, +\infty]$ を

$$I_X(a) := K_X^*(a) = \sup_{t \in \mathbb{R}} \{at - K_X(t)\}$$

と定義する。このとき、

1. $\mathrm{dom}(K_X) = \{0\}$ のとき、$I_X(a) = 0$.
2. ある $\epsilon > 0$ が存在して $K_X(\epsilon) < +\infty$ のとき、$\mathbb{E}[X] < +\infty$ であり、

$$\sup_{t \geq 0} \{at - K_X(t)\} = \begin{cases} I_X(a) & \textit{if } a > \mathbb{E}\left[X\right] \\ 0 & \textit{otherwise.} \end{cases}$$

3. ある $\epsilon > 0$ が存在して $K_X(-\epsilon) < +\infty$ のとき、$\mathbb{E}[X] > -\infty$ であり、

$$\sup_{t \leq 0} \{at - K_X(t)\} = \begin{cases} I_X(a) & \textit{if } a < \mathbb{E}\left[X\right] \\ 0 & \textit{otherwise.} \end{cases}$$

証明. 1 は自明。2 を示す。ある $\epsilon > 0$ について、$K_X(\epsilon) < +\infty$ と仮定する。一般に、

$$\mathrm{e}^{tx} \geq tx + 1 \qquad \forall t, x \in \mathbb{R}$$

より、$M_X(t) \geq t\mathbb{E}\left[X\right] + 1$ である。よって、

$$\mathbb{E}\left[X\right] \leq \frac{M_X(\epsilon) - 1}{\epsilon} < +\infty$$

である。また、イェンセンの不等式より、

$$K_X(t) = \log \mathbb{E}\left[\mathrm{e}^{tX}\right] \geq \mathbb{E}\left[\log \mathrm{e}^{tX}\right] = t\mathbb{E}\left[X\right] \qquad \forall t \in \mathbb{R}$$

である。よって、

$$I_X(\mathbb{E}[X]) = \sup_{t\in\mathbb{R}} \{t\mathbb{E}[X] - K_X(t)\} = 0$$

である。任意の $a > \mathbb{E}[X]$ と $t < 0$ について、

$$ta - K_X(t) \leq t\mathbb{E}[X] - K_X(t) \leq 0$$

であるので、任意の $a > \mathbb{E}[X]$ について

$$I_X(a) := \sup_{t\in\mathbb{R}} \{at - K_X(t)\} = \sup_{t\geq 0} \{at - K_X(t)\}$$

また、$\sup_{t\geq 0} \{at - K_X(t)\}$ は a について単調なので、任意の $a < \mathbb{E}[X]$ について

$$\sup_{t\geq 0} \{at - K_X(t)\} = 0$$

である。

3 は 2 と同様に示せる。

□

よってチェルノフ上界を最適化することで確率の指数的な上界を得る。

定理 9.2 (最適化されたチェルノフ上界)**.**

$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N} X \geq a\right) \leq \mathrm{e}^{-I_X(a)N} \qquad \forall a > \mathbb{E}[X]$$
$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N} X \leq a\right) \leq \mathrm{e}^{-I_X(a)N} \qquad \forall a < \mathbb{E}[X].$$

この指数は漸近的に最適である。証明は少し難しいので紹介しない。

定理 9.3 (クラメールの定理)**.**

$$\lim_{N\to\infty} \frac{1}{N} \log \Pr\left(\frac{1}{N}\sum_{i=1}^{N} X \geq a\right) = -I_X(a) \qquad \forall a > \mathbb{E}[X]$$
$$\lim_{N\to\infty} \frac{1}{N} \log \Pr\left(\frac{1}{N}\sum_{i=1}^{N} X \leq a\right) = -I_X(a) \qquad \forall a < \mathbb{E}[X].$$

第 10 章

正規分布と中心極限定理

10.1 中心極限定理

大数の法則により $\frac{1}{N}\sum_{k=1}^{N} X_k$ は期待値周辺に集中することが分かる。また、クラメールの定理によりその集中のスピードも精密に評価することができる。一方で期待値周辺の挙動を

$$\frac{1}{\sqrt{N}}\left(\sum_{k=1}^{N} X_k - N\mathbb{E}\left[X\right]\right)$$

より詳しく見る。

定理 10.1 (中心極限定理)**.** 分散を持つ確率変数 X とその i

$$\frac{1}{\sqrt{N\mathrm{Var}\left[X\right]}}\left(\sum_{k=1}^{N} X_k - N\mathbb{E}\left[X\right]\right)$$

10.2 特性関数

定義 10.1. 確率変数 X について、特性関数 $\varphi_X \colon \mathbb{R} \to \mathbb{C}$ を以下で定義する。

$$\varphi_X(t) := \mathbb{E}\left[\mathrm{e}^{itX}\right].$$

定理 10.2. 確率変数 X, Y について

$$\varphi_X = \varphi_Y \iff F_X = F_Y.$$

確率変数の像が有限の場合の証明. $\mathrm{Image}(X)$ と $\mathrm{Image}(Y)$ が有限の場合に限って証明を与える。

$$\{x_0, \dots, x_{N-1}\} := \mathrm{Image}(X) \cup \mathrm{Image}(Y)$$

とする。

$$\varphi_X(t) = \sum_{k=0}^{N-1} f_X(x_k)\mathrm{e}^{itx_k}, \qquad \varphi_Y(t) = \sum_{k=0}^{N-1} f_Y(x_k)\mathrm{e}^{itx_k}$$

なので、

$$0 = \varphi_X(t) - \varphi_Y(t) = \sum_{k=0}^{N-1} (f_X(x_k) - f_Y(x_k))\mathrm{e}^{itx_k} \qquad \forall t \in \mathbb{R}.$$

よって、

$$\sum_{k=0}^{N-1} (f_X(x_k) - f_Y(x_k))\mathrm{e}^{i\ell x_k} = 0 \qquad \forall \ell \in \{0, 1, \dots, N-1\} \tag{10.1}$$

である。ここで、$N \times N$ 実行列 V を

$$V_{\ell k} = \mathrm{e}^{i\ell x_k} \qquad \forall k, \ell \in \{0, 1, \dots, N-1\}$$

とおく。この行列 V は Vandermonde 行列の転置であり正則なので、

$$\begin{aligned} &\sum_{k=0}^{N-1} V_{\ell k} g_k = 0 \qquad \forall \ell \in \{0, 1, \dots, N-1\} \\ \Longrightarrow\ & g_k = 0 \qquad \forall k \in \{0, 1, \dots, N-1\} \end{aligned}$$

よって、

$$f_X(x_k) = f_Y(x_k) \qquad \forall k \in \{0, 1, \dots, N-1\}$$

である。 □

10.3 特性関数の応用

10.4 連続性定理

10.5 中心極限定理の証明

第 11 章

サノフの定理、KL ダイバージェンス

参考文献

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.