数理統計学 概説

Part 1 | 確率論・統計学の基礎

Takahiro INOUE

やること

数理統計の一部の話題をピックアップして紹介します. 具体的には,
以下の内容をやります.

Part 1 | 確率論・統計学の基礎

確率変数や確率分布, 期待値といった事柄について, 順番に定義から
確認していきます.

Part 2 | 統計的推定

パラメータの点推定を扱います. 不偏推定量, 最尤推定量, Fisher情報量,
Cramér–Raoの下限などの話題に触れます.

Part 3 | Bayes統計

事前分布, 事後分布, 予測分布などの話を, 例をとおして解説します.

数理統計の分野には, 他にも統計的検定, 区間推定, 回帰分析などの話題があります.
統計検定を受験するのであれば, このあたりの話題も勉強すると良いと思います.

今日の内容

Section 1 | 確率と確率変数

確率空間, 確率変数, 確率分布, 分布関数, 確率密度関数, 変数変換,
条件付き確率, Bayesの定理, 独立

Section 2 | 期待値

期待値, モーメント, 共分散, 確率母関数, 特性関数, モーメント母関数

Section 3 | 統計量と極限定理

統計量, 標本分布, 収束性, 大数の法則, 中心極限定理

確率論をきちんとやると, 測度論やLebesgue積分の知識が必要になります.

後々の定義・定理の理解に必要なところは, できるだけ噛み砕いて話そうと思います.
その他は, 今後もし確率論の本を開くことがあった際に死なないぐらいに,
雰囲気だけ味わうようにします.

Section 1

確率と確率変数

事象と確率

「ランダムさ」を測る

試行 (trial): 不確実性を伴う実験, 観察, 調査などの総称

ex) サイコロ投げ, 内閣支持率の調査, 株価の変動, ……

試行の結果のランダム性を定量的に評価したい ➔ 「**確率**」

「確率」は日常用語としても馴染み深い

ex) 「明日の降水確率は30%」「勝つ確率は0.1%」など

➔ では, 「確率」は**数学的には**どう定義される?

古典的な定義では不十分

Laplaceによる古典的な確率の定義:

<u>Def.</u> (組合せ確率)

試行によって起こりうる結果を**根元事象** (elementary event) という.

Ω を根元事象の集合とし, Ω の部分集合 $A \subseteq \Omega$ を**事象** (event) という.

各根元事象が<u>同様に確からしく起こる</u>とき, 事象 A の**確率** (probability) を

$\Pr(A) \coloneqq |A|/|\Omega|$ で定める.

この定義では, 例えば, 以下のようなときに**確率を計算できない**!

- 根元事象の起こり方が同様に確からしくないとき
- Ω も A も無限集合のとき

➔ 以降では, **公理的に確率を定義する**ために必要な
二つの土台を作っていく.

土台1: 標本空間

標本空間 (sample space) Ω:

試行によって起こりうる結果に対応する「何か」を全て集めたもの.

各 $\omega \in \Omega$ を**標本** (sample) という.

ex) サイコロ投げ: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$.
ω_i を「i の目が出る」という結果に対応させる.

1000人に対する内閣支持率の調査: $\Omega = \{\omega_+, \omega_-\}^{1000}$.
ω_+ を「支持」, ω_- を「不支持」という結果に対応させる.

イメージ:
標本空間 Ω は「パラレルワールドの集合」. 各パラレルワールド ω では,
試行の結果が確定している.

➔ 標本 $\omega \in \Omega$ が確定すると, 試行の結果が確定する.

標本空間は何でも良い

実は, Ω の具体的な作り方はあまり重要ではない.

何らかの方法で起こりうる結果に対応するなら, **何でも良い**.

ex) サイコロ投げの例

- $\Omega = \{1, 2, 3, 4, 5, 6\}$ と要素を整数にしておいてもよい.
- $\Omega = \mathbb{R}$ としてもよい. 適当に $\mathbb{R} = \bigsqcup_{i=1}^{6} A_i$ と分割し, $\omega \in A_i$ のときに「i の目が出る」という結果に対応させればよい.

$\sqcup$: 共通部分がないときに, 和集合の記号 $\cup$ の代わりに書くことがある.

この状況をパラレルワールド的に解釈すると,

- Ω: あらゆる試行の結果が確定したパラレルワールドの集まり
- A_i: サイコロ投げの試行で i の目が出るパラレルワールドの集まり

➔ もう一つの土台は, 標本空間 Ω をもとにして作る.

「事象」とは何だろう

もう一つの土台は, **「事象」の集まり**.

「事象」: 試行の結果を集めたもので,「確率」を計算できるもの.
「指定の結果が発生するパラレルワールドの寄せ集め」

ex) サイコロ投げで「確率」を計算できるとうれしい「事象」
「偶数の目が出る」,「素数の目が出ない」,
「1の目 か 5の目が出る」,「2の倍数 かつ 3の倍数の目が出る」, ……

➔ **「事象」に課すべき条件**:

1. 「何かが起きる」は「事象」になる.
2. 「ある事象が起こらない」も「事象」になる.
3. 「与えられた事象のうち, いずれかの事象が起きる」も「事象」になる.

➔ この条件を数学的にきちんと書き直してみよう.

土台2: σ-加法族

Def. (σ-加法族, 可測空間)

↓ **冪集合** (power set) 2^Ω: Ω の部分集合全体.

集合 Ω に対し, Ω の部分集合の族 $\mathcal{F} \subseteq 2^\Omega$ が **σ-加法族** (σ-field) であるとは,

※「集合の集合」の場合に「集合の『族』」ということが多い.

1. [全体集合は可測集合]
$$\Omega \in \mathcal{F}$$

2. [補集合は可測集合]
$$A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F} \ \ (A^c = \Omega \setminus A)$$

3. [可算個の和集合は可測集合] **可算個**:「自然数で対象を番号付けできる」ということ.
$$A_i \in \mathcal{F}\ (i = 1, 2, \dots) \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

が成り立つものをいう. 組 $(\Omega, \mathcal{F})$ を**可測空間** (measurable space) といい,

$F \in \mathcal{F}$ を**可測集合** (measurable set) という.

➔ σ-加法族には, どんな集合が含まれるだろうか?
とくに, **共通部分や有限個の和集合**などは可測になるだろうか?

σ-加法族に含まれている集合

以下に示す集合も, σ-加法族に含まれていることが示せる:

Prop.

集合 Ω の部分集合からなる σ-加法族 $\mathcal{F}$ に対し, 次が成立する.

(1) [空集合は可測集合]

$$\emptyset \in \mathcal{F}.$$

(2) [可算個の共通部分は可測集合]

$$A_i \in \mathcal{F}\ (i = 1, 2, \dots) \Longrightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}.$$

(3) [有限個の和集合, 共通部分は可測集合]

$$A_i \in \mathcal{F}\ (i = 1, 2, \dots, n) \Longrightarrow \bigcup_{i=1}^{n} A_i, \bigcap_{i=1}^{n} A_i \in \mathcal{F}.$$

➔ 標本空間 Ω に対して作ったσ-加法族の要素を**事象** (event) という.
事象の補集合を**余事象**, 和集合を**和事象**, 共通部分を**積事象**という.

cf. 命題の証明

(1) 定義1より $\Omega \in \mathcal{F}$ なので, 定義2より $\emptyset = \Omega^c \in \mathcal{F}$.

(2) $A_i \in \mathcal{F}\ (i = 1, 2, \dots)$ のとき, 定義2より $A_i^c \in \mathcal{F}\ (i = 1, 2, \dots)$.

さらに, 定義3とde Morganの法則を利用すると,

$$\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c \in \mathcal{F}.$$

よって, 定義2より $\bigcap_{i=1}^{\infty} A_i = \left(\left(\bigcap_{i=1}^{\infty} A_i\right)^c\right)^c \in \mathcal{F}$.

(3) $A_i \coloneqq \emptyset\ (i = n + 1,\ n + 2, \dots)$ と定めると, (1)より $A_i \in \mathcal{F}\ (i = 1, 2, \dots)$.

よって定義3より,

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

共通部分についても同様に示せる. ■

サイコロ投げの例

$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$ とする.

ex. 1)

$\mathcal{F}_1 = \{\emptyset, \Omega\}$ とすると, $\mathcal{F}_1$ は σ-加法族. 自明な (trivial) σ-加法族という.
「何かが起こる」「何も起こらない」という事象しか考えられない.

ex. 2)

$\mathcal{F}_2 = \{\emptyset, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}, \Omega\}$ とすると, $\mathcal{F}_2$ は σ-加法族.
「奇数の目が出る」という事象と「偶数の目が出る」という事象を
考えられるようになった (**サイコロ投げに関する「情報」が増えた**).

ex. 3)

$\mathcal{F}_3 = 2^{\Omega}$ とすると, $\mathcal{F}_3$ は σ-加法族. サイコロ投げに関する
あらゆる事象を考えられる.

➔ 用意する σ-加法族によって, 「確率」を考えられる事象が変わる.
σ-加法族は**試行について知っている「情報」を表している**と思える.

「確率」はどう定義するべきか?

「確率」を定義するための二つの土台が準備できた!
事象が起こる「確率」を定義しよう.

➔ **「確率」に課すべき条件**:

0. 「確率」は非負値をとるようにしたい.
1. 確実に起こる事象の「確率」を基準として, その他の事象の「確率」を割合として評価できるようにしたい.
2. 同時には起こり得ない (**排反**な) いくつかの事象のうちのいずれかが起こる「確率」は, それぞれの事象が起こる「確率」の和になってほしい.

➔ この条件をきちんと書き直してみる.

確率空間

Def. (確率, 確率空間)

集合 Ω の σ-加法族 $\mathcal{F}$ が与えられたとする. 関数 $\Pr : \mathcal{F} \to [0, \infty]$ は,

次の1, 2を満たすとき**確率** (probability) という:

1. [全事象の確率は1]

$$\Pr(\Omega) = 1.$$

2. [**σ-加法性**: 可算個の排反事象の和事象の確率は, 各事象の確率の和]

$$A_i \in \mathcal{F}\ (i = 1, 2, \ldots), A_i \cap A_j = \emptyset\ (i \neq j) \Longrightarrow \Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i).$$

組 $(\Omega, \mathcal{F}, \Pr)$ を**確率空間** (probability space) という.

※ 定義2のみを満たす $\mu: \mathcal{F} \to [0, \infty]$ を $\mathcal{F}$ 上の**測度** (measure) という.

➔ こうして定義した確率の満たす基本的な性質を見よう.

確率の性質

Prop. (確率の基本性質)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ と事象 $A, A_i \in \mathcal{F}$ $(i = 1, 2, \dots)$ に対して次が成り立つ.

(1) [空事象の確率] $\Pr(\emptyset) = 0.$

(2) [余事象の確率] $\Pr(A^c) = 1 - \Pr(A).$

(3) [確率の値域] $0 \le \Pr(A) \le 1.$

(4) [包除原理] $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2).$

(5) [**単調性**] $A_1 \subseteq A_2 \Longrightarrow \Pr(A_1) \le \Pr(A_2).$

(6) [**連続性1**] $A_1 \subseteq A_2 \subseteq \cdots \Longrightarrow \Pr(\bigcup_{n=1}^{\infty} A_n) = \lim_{n \to \infty} \Pr(A_n).$

(6′) [**連続性2**] $A_1 \supseteq A_2 \supseteq \cdots \Longrightarrow \Pr(\bigcap_{n=1}^{\infty} A_n) = \lim_{n \to \infty} \Pr(A_n).$

(7) [**劣加法性**] $\Pr(\bigcup_{n=1}^{\infty} A_n) \le \sum_{n=1}^{\infty} \Pr(A_n).$

➔ これらは定義を使って証明できる.
図形の面積と同様の性質があることがわかる.

cf. 基本性質の証明1

(1) $A = A \sqcup \emptyset$ と定義2より, $\Pr(A) = \Pr(A) + \Pr(\emptyset)$. これより, $\Pr(\emptyset) = 0$.

(2) $\Omega = A \sqcup A^{\mathrm{c}}$ と定義2より, $\Pr(\Omega) = \Pr(A) + \Pr(A^{\mathrm{c}})$.

定義1より $\Pr(A^{\mathrm{c}}) = 1 - \Pr(A)$.

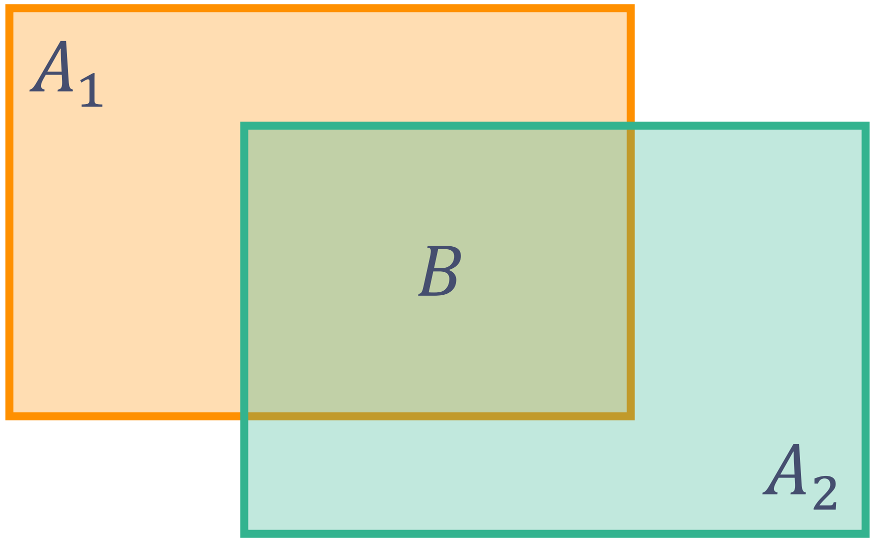
(3) 定義より $\Pr(A) \geq 0, \Pr(A^{\mathrm{c}}) \geq 0$. (2) の結果を使うと, $0 \leq \Pr(A) \leq 1$.

(4) $B \coloneqq A_1 \cap A_2$ とする. $A_1 \cup A_2 = (A_1 \setminus B) \sqcup (A_2 \setminus B) \sqcup B$ と定義2より,

$\Pr(A_1 \cup A_2) = \Pr(A_1 \setminus B) + \Pr(A_2 \setminus B) + \Pr(B)$.

ここで, $(A_i \setminus B) \sqcup B = A_i$ なので $\Pr(A_i \setminus B) + \Pr(B) = \Pr(A_i)$ を使い,

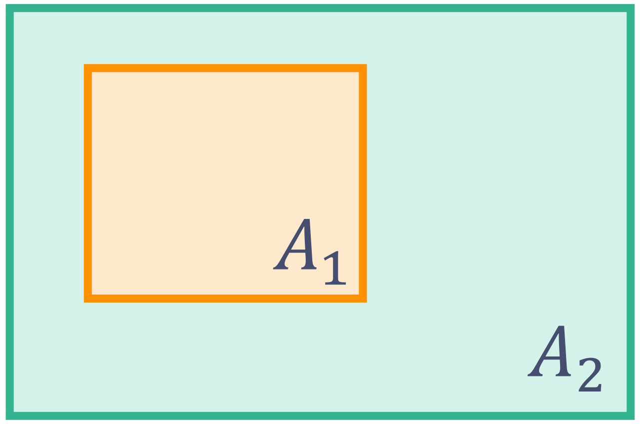
$\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(B)$.



cf. 基本性質の証明2

(5) $A_2 = (A_2 \setminus A_1) \sqcup A_1$ より,

$\Pr(A_2) = \Pr(A_2 \setminus A_1) + \Pr(A_1) \geq \Pr(A_1)$.



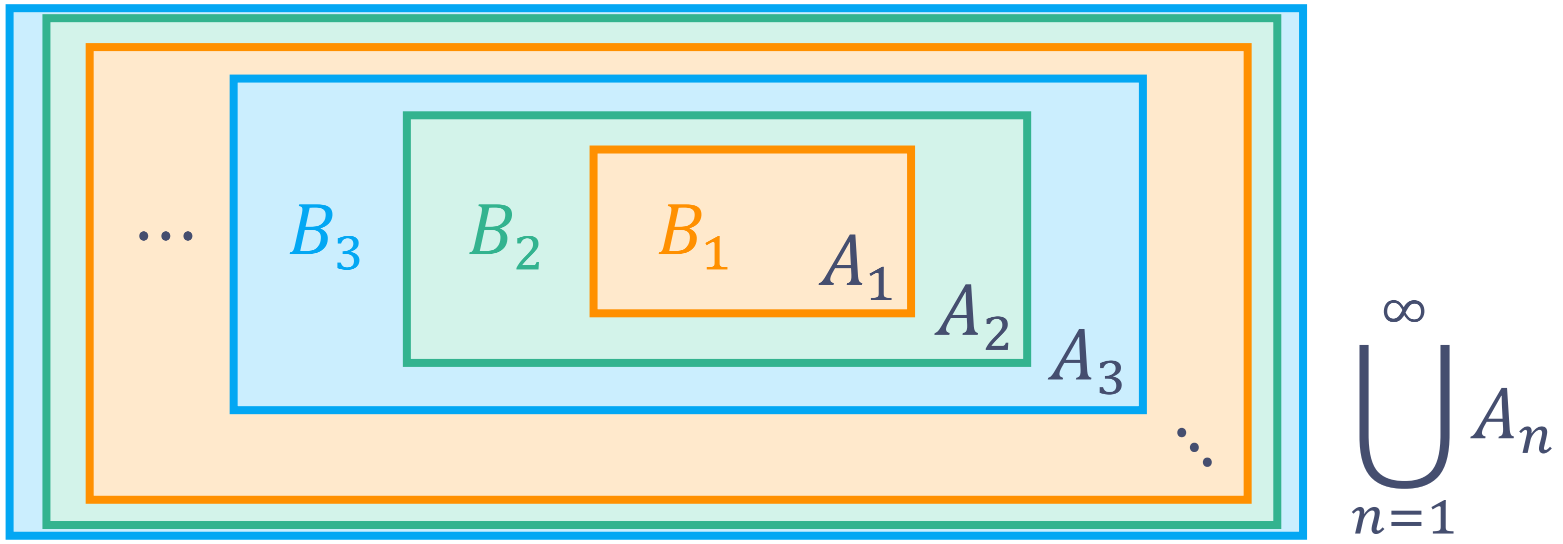
(6) $B_1 \coloneqq A_1, B_i \coloneqq A_i \setminus A_{i-1}$ $(i = 2, 3, \dots)$ と定義すると, $B_i \cap B_j = \emptyset$.

よって, 定義2より $\Pr(\bigcup_{n=1}^{\infty} A_n) = \Pr(\bigcup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \Pr(B_n)$.

また, $\Pr(B_i) = \Pr(A_i) - \Pr(A_{i-1})$ $(i = 2, 3, \dots)$ より,

$\Pr(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \Pr(B_n) = \lim_{n\to\infty} \sum_{i=1}^{n} \Pr(B_i) = \lim_{n\to\infty} \Pr(A_n)$.

(6′) (6)で, 補集合を考えればよい.



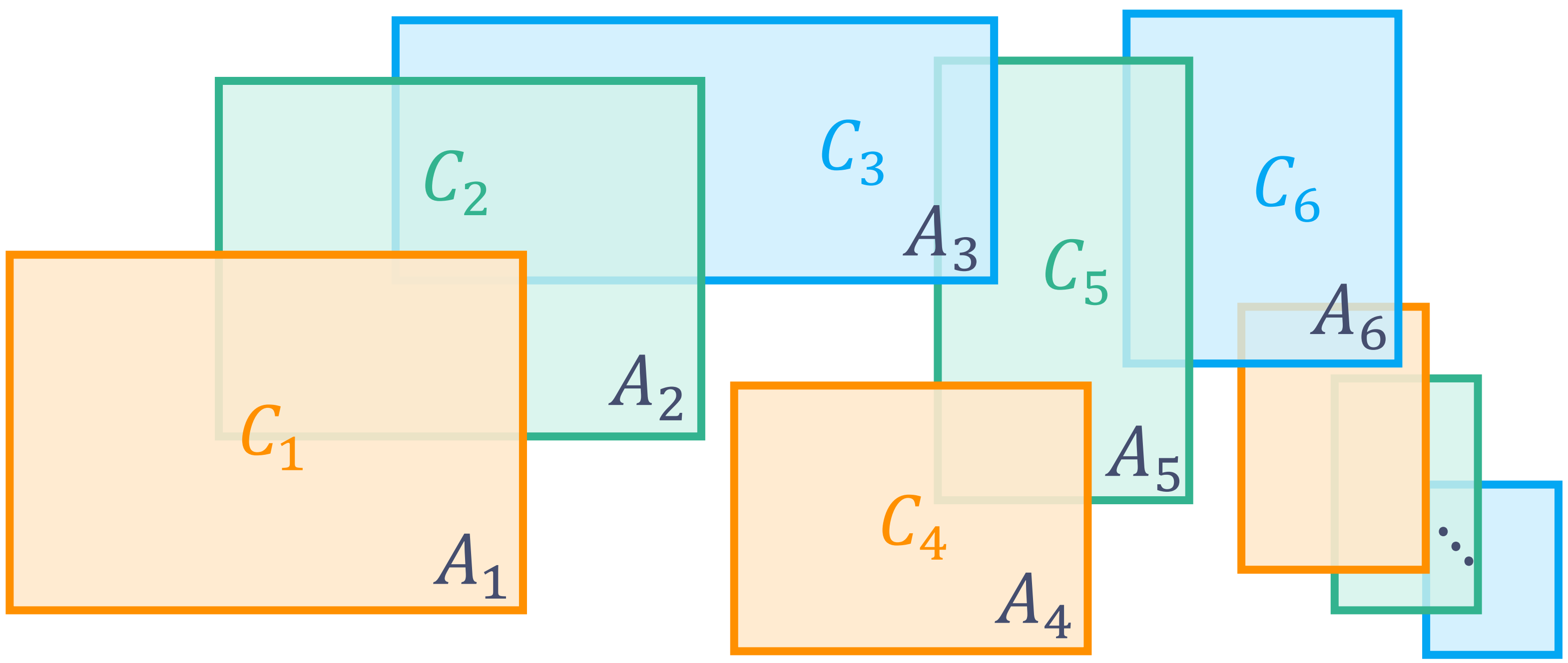
cf. 基本性質の証明3

(7) $C_1 \coloneqq A_1, C_i \coloneqq A_i \setminus \left(\bigcup_{j=1}^{i-1} A_j\right)$ $(i = 2, 3, \dots)$ と定義すると, $C_i \cap C_j = \emptyset$.

よって, 定義2より $\Pr(\bigcup_{n=1}^{\infty} A_n) = \Pr(\bigcup_{n=1}^{\infty} C_n) = \sum_{n=1}^{\infty} \Pr(C_n)$.

また, $\Pr(C_i) = \Pr(A_i) - \Pr\left(\bigcup_{j=1}^{i-1} A_j\right) \leq \Pr(A_i)$ $(i = 2, 3, \dots)$ より,

$\Pr(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \Pr(C_n) \leq \sum_{n=1}^{\infty} \Pr(A_n)$. ■



(C_i は $\bigcup_{j=1}^{i-1} A_j$ と比べて新しく増えた部分)

σ-加法族をうまく選ぶ

集合 Ω に対し, Ω の全ての部分集合を含む 2^Ω はσ-加法族.

常に可測空間 $(\Omega, 2^\Omega)$ に確率を与えるようにすれば良いのでは?

➔ 実は, Ω が非可算無限の濃度をもつときに, 2^Ω では σ-加法族としては大きすぎて, **「適切な確率」を与えられない部分集合が存在する**!
(例は省略. 例えば, 高橋『確率論』の9章に例がある.)

- 考えたい事象が全て含まれていて,
- 全ての事象の確率が定義できる

ような σ-加法族が必要.

➔ 考えたい事象が全て入るような **「手頃なサイズ」の σ-加法族**を作ろう.

最小の σ-加法族

Def. (最小の σ-加法族)

集合 Ω の部分集合族 $\mathcal{A}$ に対し, σ-加法族 $\mathcal{F}$ が次の1, 2を満たすとき,

$\mathcal{F} = \sigma(\mathcal{A})$ と書き, $\mathcal{A}$ **を含む最小のσ-加法族**という:

1. $\mathcal{A} \subseteq \mathcal{F}$.
2. $\mathcal{A}$ を含む任意の σ-加法族 $\mathcal{G}$ に対し, $\mathcal{F} \subseteq \mathcal{G}$.

➔ 実は, $\Sigma(\mathcal{A}) \coloneqq \{\mathcal{A}$ を含む σ–加法族$\}$ とすると,

$$\sigma(\mathcal{A}) = \bigcap_{\mathcal{G} \in \Sigma(\mathcal{A})} \mathcal{G}$$

となり, **いつでも一意的に存在する**! (証明は略)

➔ 無事, 手頃なサイズの σ-加法族が作れた. 実は, $\sigma(\mathcal{A})$ に含まれる集合には (ある条件のもとで) **うまく確率が定義できる**!

cf. Hopfの拡張定理

σ-加法族の定義1, 2に加えて (定義3の代わりに)

3′. $A_i \in \mathcal{F}\ (i = 1, \dots, n) \Longrightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{F}$

が成り立つ集合族のことを**有限加法族**という.

また, $\mathcal{F}$ を有限加法族として, 確率の定義1に加えて (定義2の代わりに)

2′. $A_i \in \mathcal{F}\ (i = 1, \dots, n), A_i \cap A_j = \emptyset\ (i \neq j) \Longrightarrow \mathrm{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathrm{P}(A_i)$

が成り立つ $\mathrm{P}: \mathcal{F} \to [0, \infty]$ のことを**有限加法的確率**という.

Thm. (Hopfの拡張定理)　※一般の測度の場合は「Carathéodoryの拡張定理」ともいう.

集合 Ω に対して, 有限加法族 $\mathcal{A} \subseteq 2^{\Omega}$ と有限加法的確率 $\mathrm{P}^*: \mathcal{A} \to [0, 1]$

を考える. このとき, 次の(1), (2)は同値:

(1) $\sigma(\mathcal{A})$ 上の確率 P が一意的に存在して, $\forall A \in \mathcal{A}, \mathrm{P}(A) = \mathrm{P}^*(A)$.

(2) P^* は $\mathcal{A}$ 上で σ-加法性を満たす.

➔ この定理により, $\mathcal{A}$ が有限加法族なら, $A \in \mathcal{A}$ にうまく確率を定めれば
$\sigma(\mathcal{A})$ の全ての集合にきちんと確率が定義できる!

Borel集合族

確率論では, とくに $\mathbb{R}^d$ を可測空間にして扱うことが多い.

➔ 汎用的な σ-加法族を用意しておこう!

1次元の区間の集合を $I = \{(a, b] \mid -\infty \le a \le b \le \infty\}$ とする
(ただし, $b = \infty$ のときは $(a, b] = (a, \infty)$ と解釈).

そして, d-次元区間の集合

$$\mathcal{A}_d \coloneqq \{(a_1, b_1] \times \cdots \times (a_d, b_d] \mid (a_i, b_i] \in I, i = 1, \ldots, d\}$$

を考える.

<u>Def.</u> (Borel集合族)

上記の $\mathcal{A}_d$ に対し, $\sigma(\mathcal{A}_d)$ を $\mathbb{R}^d$ の**Borel集合族** (Borel field) といい,

$\mathcal{B}_d \coloneqq \sigma(\mathcal{A}_d)$ と書く.

➔ こうして $(\mathbb{R}^d, \mathcal{B}_d)$ が**可測空間として扱えるようになった**!

cf. Borel集合族の中身

$\mathcal{B}_d$ の中身は具体的に書き表すのは難しい.

➔ 一見, $\mathcal{B}_d$ は得体の知れないものに思えるが,
$\mathcal{B}_d$ は**我々が思いつくような $\mathbb{R}^d$ の部分集合はほとんど含む**!

ex) Borel集合の例

- d-次元開区間/閉区間 $(a_i, b_i)\times\cdots\times(a_d, b_d)$, $[a_i, b_i]\times\cdots\times[a_d, b_d]$
- 1点集合 $\{\boldsymbol{x}\}$ $\left(\boldsymbol{x} \in \mathbb{R}^d\right)$
- 有理点 $\mathbb{Q}^d$
- $\mathbb{R}^d$ の開集合, 閉集合 (開集合/閉集合については, 位相空間の本を参照)
- Borel集合の可算個の和集合/共通部分や補集合 ($\because$ $\mathcal{B}_d$ は σ-加法族)

➔ 大体の集合はBorel集合になっていると考えればよいので,
ひとまずBorel集合族の中身は**あまり気にしなくても問題ない**.

ここまでのまとめ

- **確率空間** $(\Omega, \mathcal{F}, \Pr)$ は, 標本空間 Ω, σ-加法族 $\mathcal{F}$, 確率 $\Pr$ の三つ組.
  - Ω: 「試行後のパラレルワールドの寄せ集め」
  - $\mathcal{F}$: 事象 (「所定の結果が起こるパラレルワールドの集合」) の集合
  - $\Pr$: 事象の起こりやすさを測る関数
- $\mathbb{R}^d$ に対しては, **Borel集合族** $\mathcal{B}_d$ を σ-加法族として可測空間 $\left(\mathbb{R}^d, \mathcal{B}_d\right)$ を考えることが多い.

確率変数と確率分布

標本は確率的に変わる

標本空間 Ω の中身: 試行によって起こりうる結果に対応する「何か」.

➔ **標本 $\omega \in \Omega$ が背後で (確率的に) 選ばれて, その結果が (何らかの形で) 観測される**と考える.

ex) サイコロ投げ $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$.

現実に見えているのは, ω_i のような「得体の知れないもの」ではなく, サイコロの目 $i \in \mathbb{Z}$.

➔ $X(\omega_i) = i$ という標本からサイコロの目への関数 X があり, 観測されたサイコロの目 $X = i$ (**実現値**: realization) をとおして, 背後では標本 ω_i が発生していたのだと考える.

➔ 確率的に変わっているのは $\omega \in \Omega$ の方.
どの標本が選ばれたかを知るために,「適当な関数」を考えて観測する.

確率変数はある種の関数だ

観測を与えるための関数 $\boldsymbol{X}: \Omega \to \mathbb{R}^d$ が**確率変数**!

➔ 確率変数がランダムな値をとるのは, **選ばれた標本が変わっているから**.

ただし, どのような関数でも良いわけではなく, 「実現値に関する事象」の確率を考えようとしたときに, **不都合のないように**しておきたい.

ex) サイコロ投げ $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$. サイコロの目 $X(\omega_i) = i$.

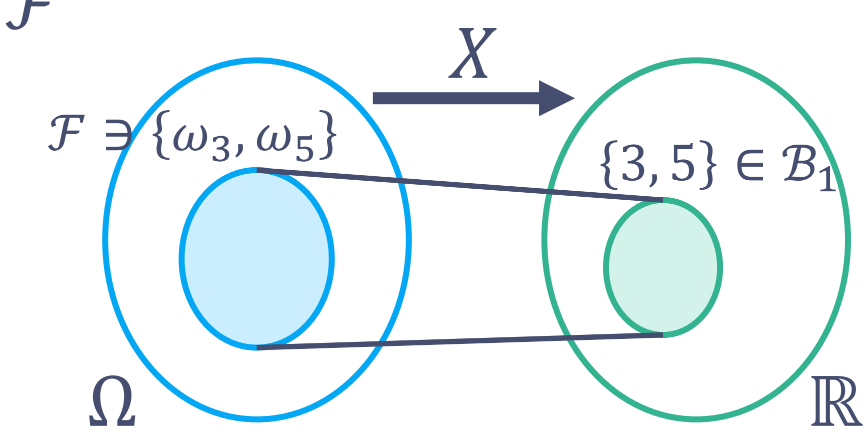
$X \in \{3, 5\}$ となる (「サイコロの目が3か5になる」) 確率を考える際, $\{\omega_3, \omega_5\}$ の確率を考えることになる.

➔ $\{3, 5\} \in \mathcal{B}_1$ の X による**逆像**が事象でないと確率を計算できない!

$$\{\omega_3, \omega_5\} = X^{-1}(\{3, 5\}) \in \mathcal{F}$$

でないと困る!

※ 関数 $f: A \to B$ に対し, $S \subseteq B$ の f による**逆像**とは, f で写した先の元が S に入るような A の部分集合 $f^{-1}(S) \coloneqq \{a \in A | f(a) \in S\}$.



➔ この条件を踏まえて, 確率変数をきちんと定義しよう.

確率変数の定義

Def. (確率変数)

可測空間 $(\Omega, \mathcal{F})$ に対して, 関数 $\boldsymbol{X}: \Omega \to \mathbb{R}^d$ が

↓ $\boldsymbol{X}$ を変数らしく書くときの表記.

任意のBorel集合 $B \in \mathcal{B}_d$ について $\{\boldsymbol{X} \in B\} \coloneqq \boldsymbol{X}^{-1}(B) \in \mathcal{F}$

を満たすとき, $\boldsymbol{X}$ を d-次元**確率変数** (random variable) という.

※ 確率変数の定義には, 確率は必要ないことに注意.

実は, **確率変数は Ω 上の可測関数**:

Def. (可測写像)

※ 写像のうち, 終域が $\mathbb{R}^d$ のような数の空間のとき, 関数と呼ぶことが多い.

可測空間 $(X_1, \mathcal{F}_1), (X_2, \mathcal{F}_2)$ に対して, 写像 $f: X_1 \to X_2$ が

任意の $S \in \mathcal{F}_2$ について $f^{-1}(S) \in \mathcal{F}_1$

を満たすとき, f を**可測写像** (measurable map) という.

➔ $\{\boldsymbol{X} \in B\} \in \mathcal{F}$ とは, 「$\boldsymbol{X}$ についての条件は (基本的に) 事象になる」ということ. **安心して X に関する事象の確率を計算できる**!

新しい確率変数をつくる

確率変数同士を四則演算したり, 関数で変換したりした結果がきちんと

確率変数になるかどうかは, 次の命題が教えてくれる.

Prop.

可測空間 $(\Omega, \mathcal{F})$ における d-次元確率変数 $\boldsymbol{X}, \boldsymbol{Y}$, 1次元確率変数 Z と,

$(\mathbb{R}^d, \mathcal{B}_d)$ から $(\mathbb{R}^p, \mathcal{B}_p)$ への可測関数 $f\colon \mathbb{R}^d \to \mathbb{R}^p$ に対して, 次が成立.

(1) $a, b \in \mathbb{R}$ とすると, $a\boldsymbol{X} + b\boldsymbol{Y}$ は d-次元確率変数.

(2) $Z \cdot \boldsymbol{X}, \boldsymbol{X}/Z$ は d-次元確率変数 (ただし, $\boldsymbol{X}/Z$ については $Z \neq 0$).

(3) $f(\boldsymbol{X})$ は p-次元確率変数.

成立は, 直観的には明らか (確率変数をいじってもランダムさは残るはず).
もちろん, これらはきちんと証明できる (が, ここでは認める).

➔ ひとまず, **確率変数の演算はあまり気にせずやれば良い**.

cf. 可測関数ってどんなもの?

結論から言えば, 大体の関数が可測関数になる!

とくに, 次の命題が成立している:

Prop.

可測空間 $(\Omega, \mathcal{F})$ から $\left(\mathbb{R}^d, \mathcal{B}_d\right)$ への (区分的) 連続関数は可測.

前ページの命題を合わせると, 多くの関数が可測になることがわかる.

例えば $\mathbb{R} \to \mathbb{R}$ の関数では, f を連続関数とすると

$f^+(x) \coloneqq \max(f(x), 0)$, $f^-(x) \coloneqq \max(-f(x), 0)$, $|f|(x) \coloneqq f^+(x) + f^-(x)$,

も連続なので可測になる.

➔ 可測関数についても**あまり気にしなくて良い**.

確率変数から確率空間をつくる

確率空間 $(\Omega, \mathcal{F}, \mathrm{Pr})$ 上の確率変数 $\boldsymbol{X}: \Omega \to \mathbb{R}^d$ を考えよう.

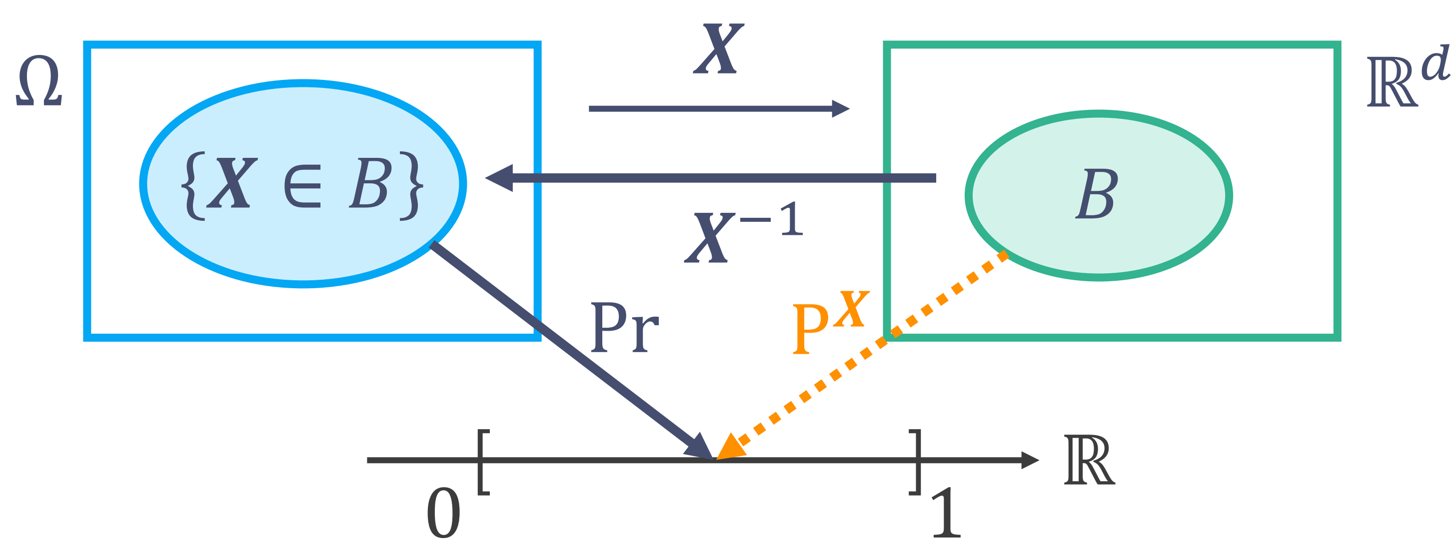
$\boldsymbol{X}$ は確率変数なので, 任意の $B \in \mathcal{B}_d$ に対して $\{\boldsymbol{X} \in B\}$ は事象. そこで,

$$\mathrm{P}^{\boldsymbol{X}}(B) \coloneqq \mathrm{Pr}(\boldsymbol{X} \in B), \qquad B \in \mathcal{B}_d$$

と定義すると, $\mathrm{P}^{\boldsymbol{X}}: \mathcal{B}_d \to [0, \infty]$ は**確率になる**ことが示せる (やってみよう).

Def. (確率分布)

確率空間 $(\Omega, \mathcal{F}, \mathrm{Pr})$ 上の確率変数 $\boldsymbol{X}: \Omega \to \mathbb{R}^d$ に対して, 上で定義された確率 $\mathrm{P}^{\boldsymbol{X}}: \mathcal{B}_d \to [0, \infty]$ を $\boldsymbol{X}$ の**確率分布** (probability distribution) と呼ぶ.



➔ こうして, **$\boldsymbol{X}$ の終域の可測空間は確率空間 $\left(\mathbb{R}^d, \mathcal{B}_d, \mathrm{P}^{\boldsymbol{X}}\right)$ になった**!

具体的な標本空間は要らない

ある試行に対する標本空間からつくった確率空間 $(\Omega, \mathcal{F}, \mathrm{Pr})$ を扱ってきた.
➔ 実際には, 具体的な Ω を考えて $(\Omega, \mathcal{F}, \mathrm{Pr})$ を構成することは**ほぼない**!

むしろ興味があるのは, 現象を確率変数 $\boldsymbol{X}: \Omega \to \mathbb{R}^d$ で観測したときの
$\boldsymbol{X}$ の値の定まり方 (i.e., $\boldsymbol{X}$ の確率分布).

ex) 公平なコインを投げたときのコインの面: $\Omega = \{\mathrm{H}, \mathrm{T}\}$

$$X(\omega) = \begin{cases} 1 & (\omega = \mathrm{H}) \\ 0 & (\omega = \mathrm{T}) \end{cases}$$

赤と白の玉が50個ずつ入った袋から玉を取り出す: $\Omega' = \{\text{赤}, \text{白}\}$

$$Y(\omega) = \begin{cases} 1 & (\omega = \text{赤}) \\ 0 & (\omega = \text{白}) \end{cases}$$

$\mathrm{P}^X = \mathrm{P}^Y$ となる. あとは, 確率空間 $(\mathbb{R}, \mathcal{B}_1, \mathrm{P}^X)$ を調べればよい.
もとの確率空間 $(\Omega, \mathcal{F}, \mathrm{Pr})$, $(\Omega', \mathcal{F}', \mathrm{Pr}')$ はもう必要ない!

➔ ひとたび確率変数の確率分布がわかってしまえば,
具体的な試行のことは忘れてしまえば良い!

確率分布を $\mathbb{R}^d$ 上の関数で表現する

d-次元確率変数 $\boldsymbol{X}$ の確率分布は $\mathcal{B}_d$ 上の関数 $\mathrm{P}^{\boldsymbol{X}}\colon \mathcal{B}_d \to [0,\infty]$.

➔ 実際は $\mathrm{P}^{\boldsymbol{X}}$ に対応する $\mathbb{R}^d$ 上の関数を利用することが多い.

Def. (分布関数)

d-次元確率変数 $\boldsymbol{X} = (X_1, \dots, X_d)$ の確率分布 $\mathrm{P}^{\boldsymbol{X}}$ が与えられたとする.

$\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ に対して関数 $F_{\boldsymbol{X}}\colon \mathbb{R}^d \to [0,1]$ を

$$F_{\boldsymbol{X}}(\boldsymbol{x}) \coloneqq \mathrm{P}^{\boldsymbol{X}}\big((-\infty, x_1]\times\cdots\times(-\infty, x_d]\big)$$

で定義する. $F_{\boldsymbol{X}}$ を $\boldsymbol{X}$ の**分布関数** (distribution function) という.

また, 各 X_i $(i = 1, \dots, d)$ に対する分布関数

$$F_{X_i}(x_i) = \mathrm{P}^{\boldsymbol{X}}\big((-\infty,\infty]\times\cdots\times(-\infty, x_i]\times\cdots\times(-\infty,\infty]\big)$$

を X_i の**周辺分布関数** (marginal distribution function) という.

➔ 分布関数の性質を (1次元の場合で) 見てみよう.

分布関数の性質 (1次元バージョン)

Prop. (分布関数の性質)

1次元確率変数 X の確率分布 P^X と分布関数 F_X について, 次が成立する:

(1) [単調性] 任意の $x \le y$ に対し, $F_X(x) \le F(y)$.

(2) [極限] $F_X(\infty) \coloneqq \lim_{x\to\infty} F_X(x) = 1,\ F_X(-\infty) = \lim_{x\to-\infty} F_X(x) = 0.$

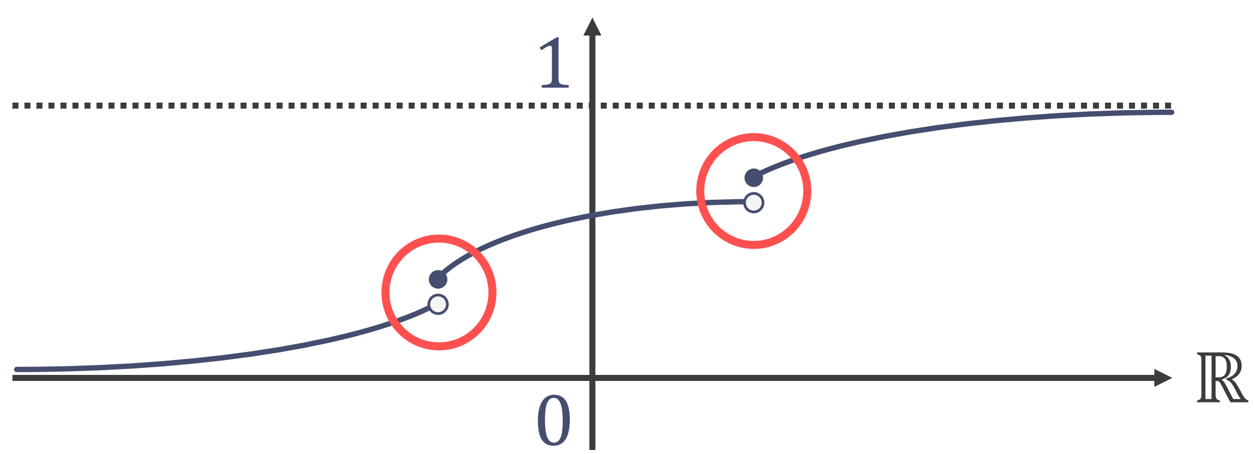
(3) [右連続性] $\forall x \in \mathbb{R},\ F_X(x+) = F_X(x).$

(4) [一点の確率] $\forall x \in \mathbb{R},\ \mathrm{P}^X(\{x\}) = F_X(x) - F_X(x-).$

$f(x+) \coloneqq \lim_{y \searrow x} f(y)$

$f(x-) \coloneqq \lim_{y \nearrow x} f(y)$

(5) [不連続点] F_X の不連続点の数は高々可算個.



➔ とくに, (1)-(3) の性質が**分布関数の本質**! (1)-(3) を満たす関数があれば, **その関数を分布関数にもつ確率変数が作れる**!

分布関数と確率は1対1対応する

Prop.

前の命題の (1)-(3) を満たす関数 $F: \mathbb{R} \to [0,1]$ が与えられたとする.

このとき, 次が成立する:

(a) $F(x) = \mathrm{P}_F\big((-\infty, x]\big)$ をみたす $(\mathbb{R}, \mathcal{B}_1)$ 上の確率 P_F が一意的に定まる.

(b) 適当な確率空間とその上の確率変数 X をとって, X の分布関数が F となるようにできる.

前の命題と併せて, **分布関数と確率は1対1に対応する**! (多次元でも同様.)

このようにして, d-次元確率変数 $\boldsymbol{X}$ に無関係に可測空間 $\left(\mathbb{R}^d, \mathcal{B}_d\right)$ に F から定義される確率 $\mathrm{P}_F: \mathcal{B}_d \to [0,1]$ を**確率分布**という.

※「X の分布をこのように定める」などと言うときの「分布」は, 分布関数 (あるいはそれに準じるもの) を設定して得られた確率分布だと理解できる.

➔ 確率分布にはどのようなものがあるのかを, しばらく見ていこう.

確率分布のタイプ

代表的な確率分布は,「離散型」「連続型」の2タイプに分かれる.

Def. (確率分布のタイプ)

$\boldsymbol{X}$ を確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の d-次元確率変数, $\mathrm{P}^{\boldsymbol{X}}$ を $\boldsymbol{X}$ の分布とする.

(1) 高々可算な集合 $A = \{\boldsymbol{a}_i \in \mathbb{R}^d \mid i = 1, 2, \dots\}$ があって,

$\Pr(\boldsymbol{X} \in A) = 1$ となるとき, $\mathrm{P}^{\boldsymbol{X}}$ は**離散型** (discrete type) の分布という.

(2) $\boldsymbol{X}$ の分布関数 F が $\mathbb{R}^d$ で連続なとき,

$\mathrm{P}^{\boldsymbol{X}}$ は**連続型** (continuous type) の分布という.

※ これ以外に「特異型」というのもある. 応用上は上記二つで十分なので, とくに気にしなくても良い.

以下では, 確率変数 $\boldsymbol{X}$ が分布 P に従うとき,「$\boldsymbol{X} \sim \mathrm{P}$」と表す.

※「確率変数が〇〇型分布に従う」ということを,「確率変数は〇〇型である」ともいう.

➔ 離散型から見ていこう.

離散型分布の確率関数

離散型分布を指定する際, 分布関数の代わりに確率関数をよく使う.

Def. (確率関数)

分布関数 $F\colon \mathbb{R}^d \to [0,1]$ から定まる分布 $\mathrm{P}_F\colon \mathcal{B}_d \to [0,1]$ は離散型とし,

高々可算な集合 $A = \{\boldsymbol{a}_i \in \mathbb{R}^d \mid i = 1, 2, \dots\}$ に対して, $\mathrm{P}_F(A) = 1$ だとする.

このとき, 分布関数は $p(\boldsymbol{a}) \coloneqq \mathrm{P}_F(\{\boldsymbol{a}\})$ とすると

$$F(\boldsymbol{x}) = \sum_{i:\boldsymbol{a}_i \le \boldsymbol{x}} p(\boldsymbol{a}_i)$$

と表せる. この $p\colon A \to [0,1]$ を**確率関数** (probability function) という.

とくに, 分布 P_F は確率関数を利用して次で表せる:

$$\mathrm{P}_F(B) = \mathrm{P}_F(B \cap A) = \sum_{\boldsymbol{a} \in B \cap A} \mathrm{P}_F(\{\boldsymbol{a}\}) = \sum_{\boldsymbol{a} \in B \cap A} p(\boldsymbol{a})$$

➔ 確率関数の値 $p(\boldsymbol{a}_i)$ は **$\boldsymbol{a}_i \in A$ が出る確率**を表す.

離散型分布の例1

ex. 1) 二項分布

X がパラメータ $n \in \mathbb{Z}_{>0}, p \in (0,1)$ の**二項分布** (binomial distribution)

に従うとき, $X \sim \mathrm{Bin}(n,p)$ と書く.

確率関数: $p(k) = \binom{n}{k} p^k (1-p)^{n-k},\ \ k = 0, 1, \dots, n$　　$\binom{n}{k}$: 二項係数

➔ 成功確率 p のBernoulli試行を n 回行ったときの成功回数の分布

↓ある種の極限 (Poissonの少数の法則): 分布収束の話題で詳述.

ex. 2) Poisson分布

X が強度 $\lambda > 0$ の**Poisson分布** (Poisson distribution)

に従うとき, $X \sim \mathrm{Poi}(\lambda)$ と書く.

確率関数: $p(k) = \mathrm{e}^{-\lambda} \dfrac{\lambda^k}{k!},\ \ k = 0, 1, 2, \dots$

➔ 「一定期間に平均 λ 回起きる事象」が起こる回数の分布

離散型分布の例2

ex. 3) 幾何分布

X がパラメータ $p \in (0,1)$ の**幾何分布** (geometric distribution)

に従うとき, $X \sim \mathrm{Ge}(p)$ と書く.　　※「無記憶性」をもつ. 独立性の話題で詳述.

確率関数: $p(k) = (1-p)^k\, p,\ \ k = 0, 1, 2, \dots$

➔ 成功確率 p のBernoulli試行が初めて成功するまでの失敗回数の分布

↓一般化

ex. 4) 負の二項分布

X がパラメータ $r \in \mathbb{Z}_{>0}, p \in (0,1)$ の**負の二項分布** (negative binomial distribution) に従うとき, $X \sim \mathrm{NB}(r,p)$ と書く.

確率関数: $p(k) = \binom{r+k-1}{k} p^r (1-p)^k,\ \ k = 0, 1, 2, \dots$

➔ 成功確率 p のBernoulli試行が r 回成功するまでの失敗回数の分布

連続型分布の確率密度関数

連続型分布を指定する際, 分布関数の代わりに確率密度関数をよく使う.

Def. (確率密度関数)

分布関数 $F\colon \mathbb{R}^d \to [0,1]$ から定まる分布 $\mathrm{P}_F\colon \mathcal{B}_d \to [0,1]$ は連続型とする.

分布関数が, ある非負関数 $f\colon \mathbb{R}^d \to [0,\infty)$ で

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$$

となるものを用いて

$$F(\boldsymbol{x}) = \int_{(-\infty,x_1]\times\cdots\times(-\infty,x_d]} f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$$

と書けるとき, P_F を**絶対連続型** (absolutely continuous type) の分布,

f を分布 P_F の**確率密度関数** (probability density function: pdf) という.

➔ 確率関数とは異なり, 確率密度関数の値 $f(\boldsymbol{x})$ は **$\boldsymbol{x}$ での確率ではない**!
とくに, $f(\boldsymbol{x})$ は**非負値であること以外に値の制限はない**ことに注意.

連続型分布での1点の確率

確率変数 $\boldsymbol{X}$ が連続型の分布関数 F に対して $\boldsymbol{X} \sim \mathrm{P}_F$ であるとする.

連続型分布では $\boldsymbol{X}$ の分布関数 F は連続だった.

これより,

$$\mathrm{P}_F(\{\boldsymbol{x}\}) = \Pr(\boldsymbol{X} = \boldsymbol{x}) = F(\boldsymbol{x}) - \lim_{\boldsymbol{h} \searrow \boldsymbol{0}} F(\boldsymbol{x} - \boldsymbol{h}) = F(\boldsymbol{x}) - F(\boldsymbol{x}) = 0.$$

➔ $\boldsymbol{X} = \boldsymbol{x}$ となる確率は0!

ただし, **「確率が0」とは, 「絶対に起こらない」ということではない**!

直観的には不思議な気はするが, 数学的にはそう定めざるをえない.

1点からなる図形の面積が0であるのと同様の話.

➔ では, 確率密度関数の値 $f(\boldsymbol{x})$ は何を表しているのか?

確率の密度

実は, 次の命題が成り立つ.

Prop.

d-次元確率変数 $\boldsymbol{X}$ が連続型分布関数 F に対して $\boldsymbol{X} \sim \mathrm{P}_F$ であるとする.

このとき, 任意の $B \in \mathcal{B}_d$ に対して次が成立する:

$$\mathrm{P}_F(B) = \Pr(\boldsymbol{X} \in B) = \int_B f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,.$$

➔ $\boldsymbol{x} = (x_1, \ldots, x_d)$ として, $B_{\boldsymbol{x}} \in \mathcal{B}_d$ を, 十分小さい $\delta_i > 0$ をとって

$$B_{\boldsymbol{x}} = [x_1, x_1 + \delta_1] \times \cdots \times [x_d, x_d + \delta_d] \supseteq \{\boldsymbol{x}\}$$

としてみる.

$\boldsymbol{\varepsilon} \geq \boldsymbol{0}$ を十分小さくとると $f(\boldsymbol{x} + \boldsymbol{\varepsilon}) \approx f(\boldsymbol{x})$ なので, ←Taylor展開を考えてみるとよい

$$\mathrm{P}_F(B_{\boldsymbol{x}}) = \int_{B_{\boldsymbol{x}}} f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \approx \int_{B_{\boldsymbol{x}}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{z} = f(\boldsymbol{x}) \cdot \delta_1 \cdots \delta_d = f(\boldsymbol{x})\,\mathrm{vol}(B_{\boldsymbol{x}}).$$

※ f は「Lebesgue測度 vol に関する確率測度 P_F のRadon-Nikodym微分」というもの.

➔ **$f(x)$ は X が x に十分近い値をとりやすいかどうかの指標**と考えられる.
$\boldsymbol{X}$ が近くにくる確率が高い $\boldsymbol{x}$ ほど $f(\boldsymbol{x})$ が大きい, i.e., 確率密度が高い.

確率密度関数の求め方

確率密度関数が分かれば, 分布関数は積分すれば求まる.

逆に分布関数が求まれば, 確率密度関数は求まる?

➔ 「積分の逆演算は微分」!

Prop.

分布関数 $F: \mathbb{R}^d \to [0,1]$ から定まる確率分布は絶対連続型であるとし,

確率密度関数を $f: \mathbb{R}^d \to [0,\infty)$ とする. このとき, 微分可能な点で

$$f(\boldsymbol{x}) = \frac{\partial^d}{\partial x_1 \cdots \partial x_d} F(\boldsymbol{x})$$

が成立する.

➔ 諸計算には確率密度関数が便利なので,
以降は主にこちらを利用することになる.

連続型分布の例1

ex. 1) 一様分布

X が $(a,b) \subseteq \mathbb{R}$ 上の**一様分布** (uniform distribution)

に従うとき, $X \sim \mathrm{U}(a,b)$ と書く.

確率密度関数: $f(x) = \frac{1}{b-a}\, 1_{(a,b)}(x)$

$1_A(x) = \begin{cases} 1 \ (x \in A) \\ 0 \ (x \notin A) \end{cases}$: 定義関数

➔ (a,b) の区間内の点を「どこでも等しい確率でとる」分布

ex. 2) 正規分布

X が平均 $\mu \in \mathbb{R}$, 分散 $\sigma^2 > 0$ の**正規分布** (normal distribution)

に従うとき, $X \sim \mathcal{N}(\mu, \sigma^2)$ と書く.

確率密度関数: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

➔ 測定誤差が従う分布. 統計の基本となる大事な分布. (cf. 中心極限定理)

連続型分布の例2

ex. 3) 指数分布

X がパラメータ $\lambda > 0$ の**指数分布** (exponential distribution)

に従うとき, $X \sim \mathrm{Exp}(\lambda)$ と書く.　　※「無記憶性」をもつ. 独立性の話題で詳述.

確率密度関数: $f(x) = \lambda \mathrm{e}^{-\lambda x}\, 1_{(0,\infty)}(x)$

➔ 生存時間が従う分布

↓一般化

ex. 4) ガンマ分布

X がパラメータ $\alpha, \beta > 0$ の**ガンマ分布** (gamma distribution)

に従うとき, $X \sim \mathrm{Ga}(\alpha, \beta)$ と書く.

確率密度関数: $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x}\, 1_{(0,\infty)}(x)$　　$\Gamma(x) = \int_0^\infty z^{x-1}\mathrm{e}^{-z}\, \mathrm{d}z$: ガンマ関数

➔ χ^2 分布を含み, 統計学では重要な分布

確率変数の変数変換

確率変数 $\boldsymbol{X}$ を, 全単射な可測関数 $\boldsymbol{g}: \mathbb{R}^d \to \mathbb{R}^d$ で変数変換してみよう.

変数変換後の確率変数 $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ の確率 (密度) 関数は,
$\boldsymbol{X}$ の確率 (密度) 関数とどのような関係にあるだろうか?

$\boldsymbol{X}$ が**離散型分布に従うときは簡単**:

$\boldsymbol{X}$ のとりうる値の集合を $A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots\}$ とし, $\boldsymbol{X}$ の確率関数を $p_{\boldsymbol{X}}$ とすると,

$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ の確率関数 $p_{\boldsymbol{Y}}$ は $\boldsymbol{y} \in \{\boldsymbol{g}(\boldsymbol{a}) | \boldsymbol{a} \in A\}$ に対して

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = p_{\boldsymbol{X}}\left(\boldsymbol{g}^{-1}(\boldsymbol{x})\right)$$

で求まる.

➔ では**絶対連続型のとき**, 確率密度関数はどのような関係になるだろうか?

変数変換公式の導出

変数変換のための全単射な可測関数 $\boldsymbol{g}: \mathbb{R}^d \to \mathbb{R}^d$ が
1階の偏導関数をもつとして, $\boldsymbol{X}$ の確率密度関数を $f_{\boldsymbol{X}}$ としよう.

$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ の分布関数 $F_{\boldsymbol{Y}}$ を計算してみる. $A = \prod_{i=1}^{d}(-\infty, y_i]$ とすると,

$$
\begin{aligned}
&F_{\boldsymbol{Y}}(\boldsymbol{y}) \\
&= \Pr(\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X}) \in A) && \leftarrow \text{分布関数の定義} \\
&= \Pr\left(\boldsymbol{X} \in \boldsymbol{g}^{-1}(A)\right) && \leftarrow \text{逆像の定義より, 一般に } f(x) \in S \Leftrightarrow x \in f^{-1}(S) \\
&= \int_{\boldsymbol{g}^{-1}(A)} f_{\boldsymbol{X}}(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} && \leftarrow \text{確率密度関数の積分で確率が求まる (p.44の命題)}
\end{aligned}
$$

となるが, $\boldsymbol{y}' = \boldsymbol{g}(\boldsymbol{z})$ の置換積分をすることで

$$F_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_A f_{\boldsymbol{X}}\left(\boldsymbol{g}^{-1}(\boldsymbol{y}')\right)\left|\det J_{\boldsymbol{g}^{-1}}(\boldsymbol{y}')\right| \mathrm{d}\boldsymbol{y}'$$

を得る. ここで, $J_{\boldsymbol{f}}(\boldsymbol{y})$ は $\boldsymbol{y}$ における関数 $\boldsymbol{f}$ の**Jacobi行列**を表し,

$$J_{\boldsymbol{f}}(\boldsymbol{y}) = \begin{pmatrix} \frac{\partial f_1}{\partial y_1}(\boldsymbol{y}) & \cdots & \frac{\partial f_1}{\partial y_d}(\boldsymbol{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_d}{\partial y_1}(\boldsymbol{y}) & \cdots & \frac{\partial f_d}{\partial y_d}(\boldsymbol{y}) \end{pmatrix}.$$ これより $f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left(\boldsymbol{g}^{-1}(\boldsymbol{y})\right)\left|\det J_{\boldsymbol{g}^{-1}}(\boldsymbol{y})\right|$.

↓ cf. 逆写像定理

とくに, Jacobi行列 $J_{g^{-1}}(\boldsymbol{y})$ については, $J_{g^{-1}}(\boldsymbol{y}) = \left(J_{g}(\boldsymbol{x})\right)^{-1}$ の関係がある.

よって, $J_{g^{-1}}(\boldsymbol{y}) J_{g}(\boldsymbol{x}) = I_d$ (単位行列) なので, $\det J_{g^{-1}}(\boldsymbol{y}) \cdot \det J_{g}(\boldsymbol{x}) = 1$.

こうして, 以下の変数変換公式が得られた:

Thm. (確率変数の変数変換)

$\boldsymbol{X}$ は絶対連続型分布に従う d-次元確率変数で, 確率密度関数を $f_{\boldsymbol{X}}$ とする.

1階の偏導関数をもつ全単射な可測関数 $\boldsymbol{g}\colon \mathbb{R}^d \to \mathbb{R}^d$ に対して, $\boldsymbol{Y} = g(\boldsymbol{X})$ の確率密度関数 $f_{\boldsymbol{Y}}$ は,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\left|\det J_{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{y})\right)\right|} f_{\boldsymbol{X}}\left(\boldsymbol{g}^{-1}(\boldsymbol{y})\right).$$

※ $J_{g}(\boldsymbol{x}) = \left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\right)$ とも書く. $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ より, $J_{g}(\boldsymbol{x}) = \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)$ とも書ける. また, $J_{g^{-1}}(\boldsymbol{y}) = \left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right)$ なので, $f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left(\boldsymbol{g}^{-1}(\boldsymbol{y})\right)\left|\det\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right)\right|$ と書ける. これを記号的に "$f_{\boldsymbol{Y}}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$" と覚えると良い.

➔ とてもよく使う基本的な公式. ぜひ使えるようにしよう.

「ほとんど確実に」

以降でよく現れる文言について定義しておく.

Def. (「ほとんど確実に」)

確率空間 $(\Omega, \mathcal{F}, \mathrm{Pr})$ において, 標本 $\omega \in \Omega$ に関する命題 $\mathcal{P}(\omega)$ が

$$\mathrm{Pr}(\{\omega \in \Omega \mid \mathcal{P}(\omega)\text{が真である}\}) = 1$$

のとき, 「命題 $\mathcal{P}$ は**ほとんど確実に** (almost surely) 成り立つ」という.

ex) $X \sim \mathcal{N}(0, 1)$ とすると,

$$\mathrm{Pr}(X^2 > 0) = \int_{\{x \neq 0\}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\, \mathrm{d}x = 1$$

ほとんど確実に $X^2 > 0$ である. これを記号で「$X^2 > 0,\ a.s.$」とも書く.

もし $X = 0$ となるなら $X^2 = 0$ だが, $\mathrm{Pr}(X = 0) = 0$. よって,

$$X^2(\omega) > 0, \forall \omega \in \Omega \setminus \{X = 0\}.$$

このように, **確率 0 の事象を除けば常に** $X^2(\omega) > 0$ が成り立つ.

➔ 「ほとんど確実に」とは **「ほとんどすべての標本 $\omega \in \Omega$ で」** の意味.
通常, **確率変数に関する等式や不等式はa.s.の意味で述べられる**.

ここまでのまとめ

- **確率変数**は, 可測空間 $(\Omega, \mathcal{F})$ から $\left(\mathbb{R}^d, \mathcal{B}_d\right)$ への可測関数.
- **確率分布**は, 確率変数をとおして, $\left(\mathbb{R}^d, \mathcal{B}_d\right)$ に誘導された確率.
- 確率分布と分布関数は1対1に対応する.
- **確率分布の分布関数の形による大別**:
  - **離散型分布**: 分布関数は確率関数で表せる.
    - **確率関数**: 各値を取る確率.
  - **連続型分布**: 分布関数は (絶対連続型なら) 確率密度関数で表せる.
    - **確率密度関数**: 各値の近傍の値の取りやすさ.
    - **変数変換の公式**: 記号的には "$f_Y(y)\,\mathrm{d}y = f_X(x)\,\mathrm{d}x$"
- 「**ほとんど確実に**」とは, 「確率0の事象に含まれる標本を除いたすべての標本で」の意味.
  - 確率変数に関する等式・不等式は, 通常a.s.の意味で述べられる.

条件付き確率

複数の要素が絡んだ試行を扱う

ここでは,「複合的な」試行を具体的に扱ってみよう.

ex) AとBの二つの箱に入っていた球を混ぜ, その中から一つ取り出す.

箱A | **赤球**10個, **白球**5個.　　**箱B** | **赤球**3個, **白球**5個.

標本空間: $\Omega = \{\omega_{a0}, \omega_{a1}, \omega_{b0}, \omega_{b1}\}$. ($a, b$ が箱**A**, **B**に, $0, 1$ が**赤**, **白**に対応.)

σ-加法族: $\mathcal{F} = 2^{\Omega}$ (任意の事象を考えられるようにしておく.)

確率: どの球も「同様に確からしく」取り出されるとして,

$$\Pr(\{\omega_{a0}\}) = \frac{10}{23}, \Pr(\{\omega_{a1}\}) = \frac{5}{23}, \Pr(\{\omega_{b0}\}) = \frac{3}{23}, \Pr(\{\omega_{b1}\}) = \frac{5}{23}$$

とする (他の事象の確率は, σ-加法性が成り立つように定める).

➔ $(\Omega, \mathcal{F}, \Pr)$ は確率空間.

選んだ球が**箱A**の球である事象を A, **赤**である事象を R と書くと,

$$\Pr(A) = \Pr(\{\omega_{a0}, \omega_{a1}\}) = \frac{15}{23}, \Pr(R) = \Pr(\{\omega_{a0}, \omega_{b0}\}) = \frac{13}{23},$$

$$\Pr(R \cap A) = \Pr(\{\omega_{a0}\}) = \frac{10}{23}.$$

状況が確定したときの確率

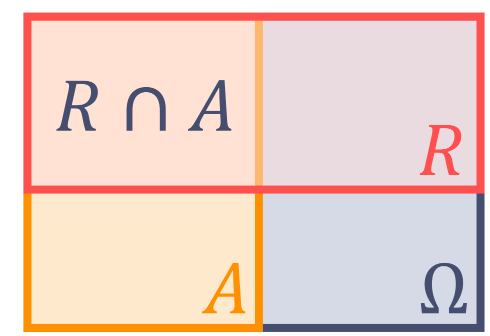
今の例で, 「取り出した球は**箱A**の球だった」(つまり, 事象 A が起こった) という状況が判明したとしよう.

↓「R given A」と読めるとカッコいい.

このもとで, 取り出した球が**赤**である確率を $\Pr(R|A)$ と書くことにすると,

$$\Pr(R|A) = \frac{10}{15}.$$

箱A: **赤球**10個, **白球**5個.
箱B: **赤球**3個, **白球**5個.



ところで, これは次のようにも書ける:

$$\Pr(R|A) = \frac{10/23}{15/23} = \frac{\Pr(R \cap A)}{\Pr(A)}.$$

➔ **すでに起こった事象 A を改めて全事象とみて**, 確率を再計算する!

$\Omega_A = A = \{\omega_{a0}, \omega_{a1}\}$, $\mathcal{F}_A = 2^{\Omega_A}$ として, 新たに $\Pr(\cdot\,|A) : \mathcal{F}_A \to [0, \infty]$ を

$$\Pr(\{\omega_{a0}\}|A) = \frac{10}{15}, \Pr(\{\omega_{a1}\}|A) = \frac{5}{15}$$

などと定める. **新たな確率空間 $(\Omega_A, \mathcal{F}_A, \Pr(\cdot\,|A))$** を考えることができる!

➔ 実は, この $\Pr(\cdot\,|A)$ は**もとの $(\Omega, \mathcal{F})$ に対する確率にもなる**!
きちんと定義してみよう.

条件付き確率

Def. (条件付き確率)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ を考える. $\Pr(B) > 0$ なる $B \in \mathcal{F}$ が与えられたとき,

$$\Pr(A|B) \coloneqq \frac{\Pr(A \cap B)}{\Pr(B)}, A \in \mathcal{F}$$

と定める. この $\Pr(\cdot\,|B) : \mathcal{F} \to [0, \infty]$ を事象 B のもとでの**条件付き確率**

(conditional probability) という.

Prop.

上で定義された $\Pr(\cdot\,|B) : \mathcal{F} \to [0, \infty]$ は, **可測空間 $(\Omega, \mathcal{F})$ における確率**.

pf.

$\Pr(\Omega|B) = \frac{\Pr(\Omega \cap B)}{\Pr(B)} = \frac{\Pr(B)}{\Pr(B)} = 1$. さらに, 排反な $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ をとると,

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i \,\middle|\, B\right) = \frac{\Pr\left(\bigcup_{i=1}^{\infty} A_i \cap B\right)}{\Pr(B)} \underset{\downarrow \Pr \text{ の } \sigma\text{-加法性}}{=} \frac{\sum_{i=1}^{\infty} \Pr(A_i \cap B)}{\Pr(B)} \underset{\downarrow \Pr(\cdot\,|B) \text{ の定義}}{=} \sum_{i=1}^{\infty} \Pr(A_i|B).$$

■

➔ 条件付き確率の性質をいくつか見てみよう.

乗法定理

条件付き確率の定義から直ちに以下の乗法定理が成立することがわかる.

Prop. (乗法定理)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ を考える. $A, B \in \mathcal{F}$ は $\Pr(A), \Pr(B) > 0$ とする.

このとき,

$$\Pr(A \cap B) = \underbrace{\Pr(B)}_{\downarrow B\text{ が起き}} \underbrace{\Pr(A|B)}_{\downarrow\text{その上で }A\text{ が起こる}} = \Pr(A)\Pr(B|A)\,.$$

とくに, $A_i \in \mathcal{F}$ $(i = 1, \ldots, n)$ で $\Pr(A_1 \cap \cdots \cap A_{n-1}) > 0$ のとき,

乗法定理を繰り返して利用すると,

$$\begin{aligned}
&\Pr(A_1 \cap \cdots \cap A_n) \\
&= \Pr(A_1 \cap \cdots \cap A_{n-1}) \Pr(A_n | A_1 \cap \cdots \cap A_{n-1}) \\
&= \cdots \\
&= \Pr(A_1) \Pr(A_2|A_1) \Pr(A_3|A_1 \cap A_2) \times \cdots \times \Pr(A_n | A_1 \cap \cdots \cap A_{n-1})\,.
\end{aligned}$$

➔ 当たり前ではあるものの, **頻繁に使う公式**!

全確率の公式

次の公式も基本的:

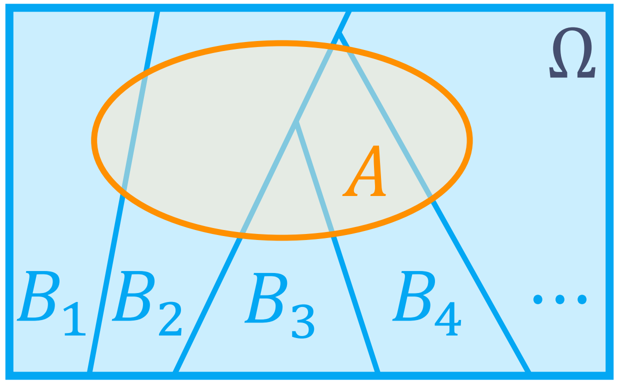
Prop. (全確率の公式)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ を考える. $A, B_i \in \mathcal{F}$ $(i = 1, 2 \ldots)$ とし,

$\Pr(B_i) > 0$ $(i = 1, 2 \ldots)$ かつ互いに排反, i.e., $B_i \cap B_j = \emptyset$ $(i \neq j)$,

さらに, $\bigcup_{i=1}^{\infty} B_i = \Omega$ であるとする. このとき,

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr(B_i) \Pr(A|B_i).$$



pf.

$A = A \cap \Omega = A \cap \left(\bigcup_{i=1}^{\infty} B_i\right) = \bigcup_{i=1}^{\infty}(A \cap B_i)$ で, $A \cap B_i$ たちは互いに排反.

よって,

$$\Pr(A) = \Pr\left(\bigcup_{i=1}^{\infty}(A \cap B_i)\right) \overset{\downarrow \Pr \text{ の } \sigma\text{-加法性}}{=} \sum_{i=1}^{\infty} \Pr(A \cap B_i) \overset{\downarrow \text{乗法定理}}{=} \sum_{i=1}^{\infty} \Pr(B_i) \Pr(A|B_i). \quad \blacksquare$$

➔ A の起こる確率を**各 B_i が起こったときそれぞれで場合分けして分割している**!

Bayesの定理

Bayes推論において基本的なのが次の定理:

Thm. (Bayesの定理)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ を考える. $A, B_i \in \mathcal{F}$ $(i = 1, 2 \ldots)$ とし, $\Pr(A) > 0$,

$\Pr(B_i) > 0$ $(i = 1, 2 \ldots)$ かつ互いに排反, i.e., $B_i \cap B_j = \emptyset$ $(i \neq j)$,

さらに, $\bigcup_{i=1}^{\infty} B_i = \Omega$ であるとする. このとき,

$$\Pr(B_i|A) = \frac{\Pr(B_i)\Pr(A|B_i)}{\sum_{j=1}^{\infty}\Pr(B_j)\Pr(A|B_j)}.$$

pf.

乗法定理から $\Pr(A \cap B_i) = \Pr(A)\Pr(B_i|A) = \Pr(B_i)\Pr(A|B_i)$. よって,

$\Pr(B_i|A) = \frac{1}{\Pr(A)}\Pr(B_i)\Pr(A|B_i)$. $\Pr(A)$ に全確率の公式を利用すればよい. ■

➔ 「**結果 A が起きた原因が事象 B_i である確率**」が求められる!

Bayesの定理の利用法

ex) 感染者問題

C: 自分が病気 C に罹っている事象. $\Pr(C) = p \in (0,1)$.

P / N: 病気 C の検査で陽性反応が出る / 出ない.

$$\Pr(P|C) = \frac{99}{100}, \Pr(N|C) = \frac{1}{100}, \qquad \Pr(P|C^c) = \frac{2}{100}, \Pr(N|C^c) = \frac{98}{100}$$

この検査を受けて陽性だったとき, 本当に病気 C にかかっている確率は

$$\Pr(C|P) = \frac{\Pr(C)\Pr(P|C)}{\Pr(C)\Pr(P|C) + \Pr(C^c)\Pr(P|C^c)} = \frac{99p}{97p+2} > p = \Pr(C).$$

➔ 結果 P という情報を得ることで, 原因 C に対する確率が更新された (**Bayes更新**された) と見ることができる!

更新前の $\Pr(C)$ を**事前確率** (prior probability),
更新後の $\Pr(C|P)$ を**事後確率** (posterior probability) という.

➔ **同様な考え方はBayes推論でもでてくる**.

事象に関する条件付き分布

確率空間 $(\Omega, \mathcal{F}, \Pr)$ に対し, 確率変数 $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ が与えられたとする.

事象 A が $\Pr(A) > 0$ のとき, 分布関数

$$F_{\boldsymbol{X}|A}(\boldsymbol{x}) \coloneqq \Pr(\boldsymbol{X} \leq \boldsymbol{x} | A) = \frac{\Pr(\{\boldsymbol{X} \leq \boldsymbol{x}\} \cap A)}{\Pr(A)}$$

を定義することができる (**条件付き分布関数**).

すると, $F_{\boldsymbol{X}|A}$ に対応する確率分布 $\mathrm{P}_{F_{\boldsymbol{X}|A}}$ が一意的に存在する.

Def. (事象に関する条件付き分布)

上で定義される分布関数 $F_{\boldsymbol{X}|A}$ に対応する確率分布 $\mathrm{P}_{F_{\boldsymbol{X}|A}}$ を,

事象 A に関する $\boldsymbol{X}$ の**条件付き分布** (conditional distribution) という.

➔ これを利用して, 離散型確率変数の実現値が与えられた下での
条件付き分布を定義しよう.

条件付き確率関数

確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数 $\boldsymbol{X}, \boldsymbol{Y}$ が離散型分布に従うとし,

$\boldsymbol{Y} = \boldsymbol{y}_0$ の下での $\boldsymbol{X}$ の**条件付き分布関数**を次で定義する:

$$F_{X|Y}(\boldsymbol{x}|\boldsymbol{y}_0) \coloneqq \begin{cases} F_{X|\{Y=y_0\}}(\boldsymbol{x}) = \dfrac{\Pr(\boldsymbol{X} \leq \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}_0)}{\Pr(\boldsymbol{Y} = \boldsymbol{y}_0)} & (\Pr(\boldsymbol{Y} = \boldsymbol{y}_0) > 0) \\ 0 & (\Pr(\boldsymbol{Y} = \boldsymbol{y}_0) = 0) \end{cases}$$

ここで, $\boldsymbol{X}$ の取りうる値の集合を $A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots\}$ とする.

$\Pr(\boldsymbol{Y} = \boldsymbol{y}_0) > 0$ のときは, $\boldsymbol{X}, \boldsymbol{Y}$ の確率関数を p_X, p_Y とし, $(\boldsymbol{X}, \boldsymbol{Y})$ の

同時確率関数を $p_{X,Y}$ とすると,

$$F_{X|Y}(\boldsymbol{x}|\boldsymbol{y}_0) = \sum_{i:\boldsymbol{a}_i \leq \boldsymbol{x}} \frac{\Pr(\boldsymbol{X} = \boldsymbol{a}_i, \boldsymbol{Y} = \boldsymbol{y}_0)}{\Pr(\boldsymbol{Y} = \boldsymbol{y}_0)} = \sum_{i:\boldsymbol{a}_i \leq \boldsymbol{x}} \frac{p_{X,Y}(\boldsymbol{a}_i, \boldsymbol{y}_0)}{p_Y(\boldsymbol{y}_0)}.$$

➔ 対応する確率分布は離散型.

$\dfrac{p_{X,Y}(\boldsymbol{a}_i, \boldsymbol{y}_0)}{p_Y(\boldsymbol{y}_0)}$ の部分を「条件付き確率関数」と定めることができる!

条件付き分布 (離散型)

Def. (条件付き確率関数)

離散型確率変数ベクトル $\boldsymbol{X}, \boldsymbol{Y}$ の確率関数を $p_{\boldsymbol{X}}, p_{\boldsymbol{Y}}$ とし, $(\boldsymbol{X}, \boldsymbol{Y})$ の同時確率関数を $p_{\boldsymbol{X},\boldsymbol{Y}}$ とする. $\boldsymbol{X}$ の取りうる値の集合を A とすると,

$$p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}_0) := \begin{cases} \dfrac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}_0)}{p_{\boldsymbol{Y}}(\boldsymbol{y}_0)} & (p_{\boldsymbol{Y}}(\boldsymbol{y}_0) > 0) \\ 0 & (p_{\boldsymbol{Y}}(\boldsymbol{y}_0) = 0) \end{cases} \qquad (\boldsymbol{x} \in A)$$

と定めた $p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}_0)$ を $\boldsymbol{X}$ の $\boldsymbol{Y} = \boldsymbol{y}_0$ の下での**条件付き確率関数**という.

条件付き確率関数を使うと, $\boldsymbol{Y} = \boldsymbol{y}_0$ の下での $\boldsymbol{X}$ の**条件付き分布** $\mathrm{P}_{\boldsymbol{X}|\boldsymbol{Y}}$ は

$$\mathrm{P}_{\boldsymbol{X}|\boldsymbol{Y}}: \mathcal{B}_d \ni B \mapsto \Pr(\boldsymbol{X} \in B | \boldsymbol{Y} = \boldsymbol{y}_0) = \sum_{\boldsymbol{x} \in B \cap A} p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}_0) \in [0, \infty]$$

と表すことができる (A は $\boldsymbol{X}$ の取りうる値の集合).

➔ 今度は, 絶対連続型確率変数の実現値が与えられた下での条件付き分布を定義しよう.

絶対連続型のときは別アプローチで

確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数 $\boldsymbol{X}, \boldsymbol{Y}$ が絶対連続型分布に従うとする.

$\Pr(\boldsymbol{Y} = \boldsymbol{y}_0) = 0$ であったから, 離散型分布と同様の流れでは

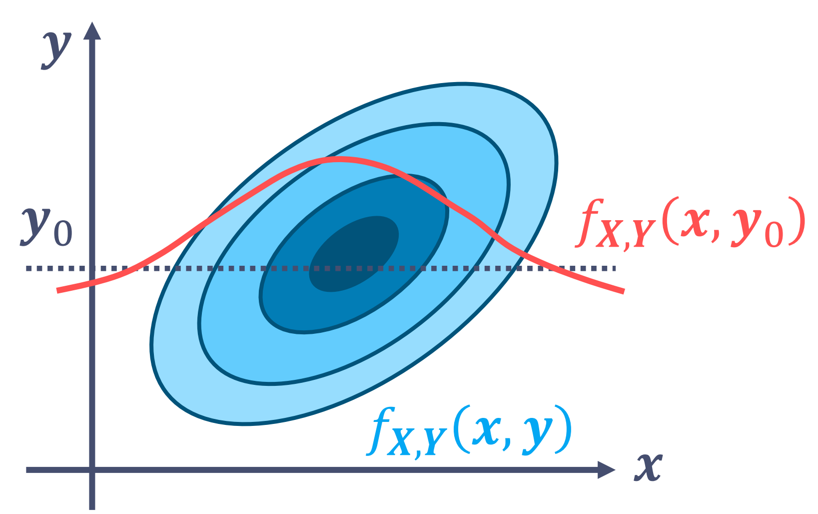
条件付き確率密度関数を定義できない...... 別アプローチで定義しよう.

$\boldsymbol{Y} = \boldsymbol{y}_0$ が確定したとき, $\boldsymbol{X}$ はどのような値を取る確率が高いだろうか?

「確率密度関数の値: 各値の近傍の値の取りやすさ」だった.

➔ 同時確率密度関数 $f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})$ で $\boldsymbol{y} = \boldsymbol{y}_0$ とした関数 $g(\boldsymbol{x}) \coloneqq f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}_0)$ が **大きい値を取りやすい**はず.

$g(x)$ **は確率密度関数のようなものと思えそう**.



➔ $g(\boldsymbol{x})$ は非負. あとは全範囲での積分が 1になるように調整しよう.

条件付き確率密度関数

定数 $c > 0$ を使って, 条件付き確率密度関数が $g(\boldsymbol{x})/c$ と書けるとすると,

$$\int_{\mathbb{R}^d} g(\boldsymbol{x})/c \ \mathrm{d}\boldsymbol{x} = 1.$$

$g(\boldsymbol{x}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}_0)$ を代入すると $c = f_{\boldsymbol{Y}}(\boldsymbol{y}_0)$ がわかる.

Def. (条件付き確率密度関数)

絶対連続型確率変数ベクトル $\boldsymbol{X}, \boldsymbol{Y}$ の確率密度関数を $f_{\boldsymbol{X}}, f_{\boldsymbol{Y}}$ とし, $(\boldsymbol{X}, \boldsymbol{Y})$ の同時確率密度関数を $f_{\boldsymbol{X},\boldsymbol{Y}}$ とする. このとき,

$$f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}_0) := \begin{cases} \dfrac{f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}_0)}{f_{\boldsymbol{Y}}(\boldsymbol{y}_0)} & (f_{\boldsymbol{Y}}(\boldsymbol{y}_0) > 0) \\ 0 & (f_{\boldsymbol{Y}}(\boldsymbol{y}_0) = 0) \end{cases}$$

と定め, $f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}_0)$ を $\boldsymbol{X}$ の $\boldsymbol{Y} = \boldsymbol{y}_0$ の下での**条件付き確率密度関数**という.

※ この条件付き確率密度関数を積分することできちんと条件付き確率を与えるかどうかは, 本来証明が必要なもの. 議論が煩雑になってしまうので, 認めることにする.

➔ $f_{\boldsymbol{Y}}(\boldsymbol{y}_0)$ は**確率密度関数にするための調整項**.

条件付き分布 (絶対連続型)

$\boldsymbol{X}$ の $\boldsymbol{Y} = \boldsymbol{y}_0$ の下での条件付き確率密度関数が定義できたので,

$\boldsymbol{X}$ の $\boldsymbol{Y} = \boldsymbol{y}_0$ の下での**条件付き分布関数**も

$$F_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) := \int_{\prod_{i=1}^{d}(-\infty,x_i]} f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{z}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{z}$$

と定義することができる.

これに対応する確率分布を $\boldsymbol{X}$ の $\boldsymbol{Y} = \boldsymbol{y}_0$ の下での**条件付き分布** $\mathrm{P}_{\boldsymbol{X}|\boldsymbol{Y}}$ という.

$\mathrm{P}_{\boldsymbol{X}|\boldsymbol{Y}}$ は条件付き確率密度関数を利用すると

$$\mathrm{P}_{\boldsymbol{X}|\boldsymbol{Y}}: \mathcal{B}_d \ni B \mapsto \Pr(\boldsymbol{X} \in B|\boldsymbol{Y} = \boldsymbol{y}_0) = \int_B f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}_0)\, \mathrm{d}\boldsymbol{x} \in [0, \infty]$$

と表せる.

➔ 確率変数の実現値の下での条件付き分布がうまく定義できた.

条件付き確率 (密度) 関数の性質

条件付き確率で成り立った乗法定理, 全確率の公式, Bayesの定理は,

実は, **条件付き確率 (密度) 関数に対しても成立する**!

Prop. (条件付き確率 (密度) 関数の性質)

$d_{\boldsymbol{X}}$-次元確率変数 $\boldsymbol{X}$ と $d_{\boldsymbol{Y}}$-次元確率変数 $\boldsymbol{Y}$ の確率 (密度) 関数を $f_{\boldsymbol{X}}, f_{\boldsymbol{Y}}$,

$(\boldsymbol{X}, \boldsymbol{Y})$ の同時確率 (密度) 関数を $f_{\boldsymbol{X},\boldsymbol{Y}}$ とする. このとき, 次が成立する:

(1) [乗法定理] $f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = f_{\boldsymbol{Y}}(\boldsymbol{y}) f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})$.

(2) [全確率の公式] $f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{\mathbb{R}^{d_{\boldsymbol{Y}}}} f_{\boldsymbol{Y}}(\boldsymbol{y}) f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$.

(3) [Bayesの定理]

$$f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})}{\int_{\mathbb{R}^{d_{\boldsymbol{X}}}} f_{\boldsymbol{X}}(\boldsymbol{x}') f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}')\, \mathrm{d}\boldsymbol{x}'}.$$

➔ これらは, Bayes推論の際に頻繁に用いる.

ここまでのまとめ

- **条件付き確率**:
  ある事象が起こった上で, 改めて起こった事象を全事象とみなすことで得られる可測空間 $(\Omega, \mathcal{F})$ 上の確率.

- **条件付き分布**:
  確率変数 $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ の値が確定した上で, その事象を全事象とみなすことで得られる可測空間 $\left(\mathbb{R}^d, \mathcal{B}_d\right)$ 上の確率.

- 条件付き確率や条件付き確率 (密度) 関数については,

  乗法定理, 全確率の公式, Bayesの定理

  が成立.

独立性

二つの事象の独立性

事象 A, B が $\Pr(A), \Pr(B) > 0$ であり $\Pr(A|B) = \Pr(A)$ とする.

➔ 事象 A は**事象 B の結果から「影響」を受けていない**と解釈できる!
さらにこのとき $\Pr(B|A) = \Pr(B)$ でもあるので (なぜか考えてみよう),
事象 B も**事象 A の結果から「影響」を受けていない**!

とくに, $\Pr(A \cap B) = \Pr(A)\Pr(B)$ が成り立っている.

Def. (二つの事象の独立性)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ を考える. 事象 $A, B \in \mathcal{F}$ に対して,

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

が成り立つとき, 事象 A, B は**独立** (independent) であるという.

※ 定義では, $\Pr(A) = 0$ や $\Pr(B) = 0$ の場合も許していることに注意.
なお, $\Pr(A) = 0$ とすると, $A \cap B \subseteq A$ なので, 確率の単調性から $0 \leq \Pr(A \cap B) \leq \Pr(A) = 0$ となり,
$\Pr(A \cap B) = 0$. よって, 確率0の事象は, あらゆる事象と独立.

➔ 二つの事象の積事象の確率は, **独立ならば計算が容易になる**!
三つ以上の事象の独立性も定義しよう.

複数の事象の独立性

Def. (複数の事象の独立性)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ を考える. 事象列 $\mathcal{A} = \{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ が与えられたとする.

(1) 任意の $i, j \in \mathbb{Z}_{>0}$ $(i \neq j)$ に対して,

$$\Pr\left(A_i \cap A_j\right) = \Pr(A_i) \Pr\left(A_j\right)$$

が成り立つとき, $\mathcal{A}$ は**組ごとに独立** (pairwise independent) という.

(2) 任意の $n \in \mathbb{Z}_{>0}$ に対して, 任意の異なる $i_1, \ldots, i_n \in \mathbb{Z}_{>0}$ をとり,

$$\Pr\left(A_{i_1} \cap \cdots \cap A_{i_n}\right) = \prod_{k=1}^{n} \Pr\left(A_{i_k}\right)$$

が成り立つとき, $\mathcal{A}$ は**独立** (independent) という.

$\mathcal{A} = \{A_i\}_{i=1}^{m}$ として, $\Pr(A_1 \cap \cdots \cap A_m) = \prod_{i=1}^{m} \Pr(A_i)$ だけでは
$\mathcal{A}$ の独立性は言えないことに注意!

➔ 二つの事象の独立性よりも複雑に思える……
具体例をとおして, もう少し詳しく見てみよう.

独立にならない例

ex) 一つの (公平な) サイコロを2回投げる試行.

A: 2回目に1, 2, 5のいずれかが出る. $\Pr(A) = \frac{1}{2}$

B: 2回目に4以上が出る. $\Pr(B) = \frac{1}{2}$

C: 2回の目の数の和が9になる. $\Pr(C) = \frac{4}{36}$

$A \cap B \cap C$: 1回目に4, 2回目に5が出る. $\Pr(A \cap B \cap C) = \frac{1}{36}$

このとき $\Pr(A \cap B \cap C) = \Pr(A)\Pr(B)\Pr(C)$ だが, $\{A, B, C\}$ は**独立でない**!

$$\Pr(A \cap B) = \frac{1}{6}, \Pr(B \cap C) = \frac{3}{36}, \Pr(A \cap C) = \frac{1}{36}$$

➔ $\{A, B, C\}$ は**組ごとに独立ですらない**!

➔ **事象列 $\mathcal{A}$ から選んだあらゆる組合せの事象が独立になって初めて $\mathcal{A}$ が独立だと言える**.

「組ごとに独立」だが「独立」でない例

ex) パリティビット.

0, 1のいずれかからなる長さ3の文字列をつくる.

このとき, 1, 2文字目は「同様に確からしく」ランダムに選び,

3文字目は文字列全体で1の数が偶数個になるように選ぶ.

あり得るパターン: $\{000, 101, 011, 110\}$

A_1: 1文字目が1,　A_2: 2文字目が1,　A_3: 3文字目が1.

$$\Pr(A_1) = \Pr(A_2) = \Pr(A_3) = \frac{1}{2}$$

$$\Pr(A_1 \cap A_2) = \Pr(A_2 \cap A_3) = \Pr(A_1 \cap A_3) = \frac{1}{4}$$

$$\Pr(A_1 \cap A_2 \cap A_3) = 0 \neq \Pr(A_1)\Pr(A_2)\Pr(A_3)$$

➔ 事象の組 $\{A_1, A_2, A_3\}$ は組ごとに独立だが独立ではない.

➔ **「独立」のほうが「組ごとに独立」よりも強い概念**.

確率変数の独立性

確率変数の独立性も, 事象の独立性と同様に定義される.

<u>Def.</u> (確率変数の独立性)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ と d-次元確率変数列 $\mathcal{X} = \{\boldsymbol{X}_i : \Omega \to \mathbb{R}^d\}_{i=1}^{\infty}$ を考える.

(1) 任意の $i, j \in \mathbb{Z}_{>0}$ $(i \neq j)$ と, 任意の $B_i, B_j \in \mathcal{B}_d$ に対して,

$$\Pr\left(\boldsymbol{X}_i \in B_i, \boldsymbol{X}_j \in B_j\right) = \Pr(\boldsymbol{X}_i \in B_i) \Pr\left(\boldsymbol{X}_j \in B_j\right)$$

が成り立つとき, $\mathcal{X}$ は**組ごとに独立** (pairwise independent) という.

(2) 任意の $n \in \mathbb{Z}_{>0}$ に対して, 任意の異なる $i_1, \dots, i_n \in \mathbb{Z}_{>0}$ と

任意の $B_{i_1}, \dots, B_{i_n} \in \mathcal{B}_d$ をとり,

$$\Pr\left(\boldsymbol{X}_{i_1} \in B_{i_1}, \dots, \boldsymbol{X}_{i_n} \in B_{i_n}\right) = \prod_{k=1}^{n} \Pr\left(\boldsymbol{X}_{i_k} \in B_{i_k}\right)$$

が成り立つとき, $\mathcal{X}$ は**独立** (independent) という.

➔ 確率変数の独立性を定義どおり確認するのは煩雑.

確率変数の独立性の判定1

実は, 有限個の確率変数列の独立性はもう少し単純に判定できる!

Prop. (確率変数の独立性の判定1)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ と d-次元確率変数列 $\mathcal{X} = \{\boldsymbol{X}_i : \Omega \to \mathbb{R}^d\}_{i=1}^N$ を考える.

$\mathcal{X}$ が独立である

$\Leftrightarrow$ 任意の $B_1, \ldots, B_N \in \mathcal{B}_d$ に対し,

$$\Pr(\boldsymbol{X}_1 \in B_1, \ldots, \boldsymbol{X}_N \in B_N) = \prod_{i=1}^N \Pr(\boldsymbol{X}_i \in B_i).$$

とはいえ, 任意のBorel集合に関して調べる必要があるので,
これでも判定はまだ面倒……

➔ 有限個の確率変数列の独立性については, **さらに簡単な判定法がある**!

cf. 命題の証明

pf.

十分性 (⇒) は明らか. 必要性 (⇐) を示す.

任意の $n \in \{1, \dots, N\}$ に対し, 任意の異なる $i_1, \dots, i_n \in \{1, \dots, N\}$ と

$B_{i_1}, \dots, B_{i_n} \in \mathcal{B}_d$ をとる.

$k = i_1, \dots, i_n$ 以外の B_k については, すべて $B_k = \mathbb{R}^d \in \mathcal{B}_d$ としておけば,

$\Pr\left(X_k \in \mathbb{R}^d\right) = 1$ であるから,

$$\Pr\left(\boldsymbol{X}_{i_1} \in B_{i_1}, \dots, \boldsymbol{X}_{i_n} \in B_{i_n}\right) = \Pr(\boldsymbol{X}_1 \in B_1, \dots, \boldsymbol{X}_N \in B_N) = \prod_{k=1}^{n} \Pr\left(\boldsymbol{X}_{i_k} \in B_{i_k}\right).$$

■

確率変数の独立性の判定2

Prop. (確率変数の独立性の判定2)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ と d-次元確率変数列 $\mathcal{X} = \{\boldsymbol{X}_i\colon \Omega \to \mathbb{R}^d\}_{i=1}^N$ を考える.

各 $\boldsymbol{X}_i$ $(i = 1, \dots, N)$ の分布関数をそれぞれ F_i とし, 確率 (密度) 関数 f_i を持つとする. また, Nd-次元確率変数ベクトル $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)^\top$ の同時分布を $F_{\boldsymbol{X}}$ とし, これが確率 (密度) 関数 $f_{\boldsymbol{X}}$ を持つとする.

このとき, 以下の (1)–(3) は同値:

(1) $\mathcal{X}$ が独立である.

(2) 任意の $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$ に対し, $F_{\boldsymbol{X}}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) = \prod_{i=1}^N F_i(\boldsymbol{x}_i)$.

(3) 任意の $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$ に対し, $f_{\boldsymbol{X}}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) = \prod_{i=1}^N f_i(\boldsymbol{x}_i)$.

➔ **分布関数や確率 (密度) 関数を見ることで独立性を判定できる**!

独立に同一分布に従う

統計学では, 得られるデータがすべてある分布から発生していると考えて, その分布に関する未知量を推定する. そのため, 確率変数列が**独立ですべて同じ分布に従っている場合**が特に重要になってくる.

Def. (独立に同一分布に従う)

確率変数列 $\mathcal{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\}$ (有限個でも可算無限個でも良い) が独立であって, 各 $\boldsymbol{X}_i$ がすべて同じ分布に従うとき,

$\mathcal{X}$ は**独立に同一分布に従う** (independently identically distributed) といい,

これを「$\mathcal{X}$ はi.i.d.」などと略する.

➔ 後の議論では, i.i.d.の状況が頻繁に現れる.

計算練習: Box-Muller変換

Box-Muller変換: 標準正規分布に従う乱数を作る方法.

$U_1, U_2 \sim \mathrm{U}(0,1)$ で, U_1 と U_2 は独立とする.

$R = \sqrt{-2\log U_1}\,, \Theta = 2\pi U_2$ とおき, $X = R\cos\Theta\,, Y = R\sin\Theta$ と変数変換.

$$x(u_1, u_2) = \sqrt{-2\log u_1}\cos 2\pi u_2\,, y(u_1, u_2) = \sqrt{-2\log u_1}\sin 2\pi u_2$$

とおくと, $r(x,y) = \sqrt{x^2+y^2}, \theta(x,y) = \arctan\frac{y}{x}$ なので,

$$u_1(x,y) = \mathrm{e}^{-\frac{1}{2}(x^2+y^2)}, u_2(x,y) = \frac{1}{2\pi}\arctan\frac{y}{x}.$$

変数変換の公式:

$$f_Y(\boldsymbol{y}) = \frac{1}{\left|\det J_g\left(\boldsymbol{g}^{-1}(\boldsymbol{y})\right)\right|} f_X\left(\boldsymbol{g}^{-1}(\boldsymbol{y})\right)$$

変数変換のJacobianは

$$\det\left(\frac{\partial(x,y)}{\partial(u_1,u_2)}\right) = \det\begin{pmatrix} \frac{1}{u_1 r}\cos\theta & -2\pi r\sin\theta \\ \frac{1}{u_1 r}\sin\theta & 2\pi r\cos\theta \end{pmatrix} = \frac{2\pi}{u_1}(>0).$$

変数変換の公式より $f_{X,Y}(x,y) = \frac{u_1}{2\pi} = \frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}(x^2+y^2)} = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2}\,\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}y^2}.$

➔ X, Y は独立で, それぞれ標準正規分布に従う!

ここまでのまとめ

- **独立性**: 各事象 / 確率変数が他に影響を及ぼさない状況.
  - 事象の独立性の確認 ➔ 定義通り
  - 有限個の確率変数の独立性の確認
    ➔ 同時分布関数 / 同時確率 (密度) 関数がそれぞれの
    周辺分布関数 / 周辺確率 (密度) 関数の積の形に分解できるか?
- **独立同分布に従う (i.i.d.)**: 確率変数列が独立で, 従う分布が同一.

Section 2

期待値

期待値の定義

確率変数を要約しよう

確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数 $X\colon \Omega \to \mathbb{R}$ を考えよう
(しばらくの間は1次元の確率変数を考えることにする).

X のランダムな挙動は, 確率 $\Pr$ (したがって, X の確率分布) が決めている.

➔ 確率 $\Pr$ が分かれば X の挙動は完全につかめる.

……とはいえ, 「結局のところ X の取る値は大体どのぐらいになるのか」
が分かった方が, X の挙動を**より直観的に**理解できそう.

➔ 「**X が大体どのような値をとると期待できるか**」を表す値を
定めてみよう.

離散型確率変数の期待値

まずは, 確率変数が**離散型**の場合を考えてみよう.

Def. (離散型確率変数の期待値)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ と1次元の離散型確率変数 $X\colon \Omega \to \mathbb{R}$ が与えられたとし, X の取りうる値の集合を $A = \{a_i \in \mathbb{R} | i = 1, 2, \ldots\}$ とし, X の確率関数を p_X とする. このとき, a_i たちの $p_X(a_i)$ による重み付き平均

$$\mathbb{E}[X] \coloneqq \sum_{i=1}^{\infty} a_i \cdot p_X(a_i)$$

で定まる値を確率変数 X の**期待値** (expectation) という.

$\mu = \mathbb{E}[X]$ とおくと $0 = \sum_{i=1}^{\infty} (a_i - \mu) \cdot p_X(a_i)$ となっているので, "**重心**".

μ からの a_i の距離↑ ↑ a_i の重み

➔ では, X が**連続型**の場合は?

連続型確率変数を「近似」する

連続型確率変数では, $\Pr(X = x) = 0$.

➔ 離散型確率変数の期待値のように,

$$\sum_{x \in \mathbb{R}} x \Pr(X = x)$$

ではうまく定義できない……

➔ 離散型確率変数の期待値の定義を利用できるよう,
連続型確率変数を**離散型確率変数で「近似」する**ことはできないか?

つまり,

連続型確率変数 X をだんだん精度良く「近似」するような
離散型確率変数列 $\{X_n\}_{n=1}^{\infty}$ をつくって,

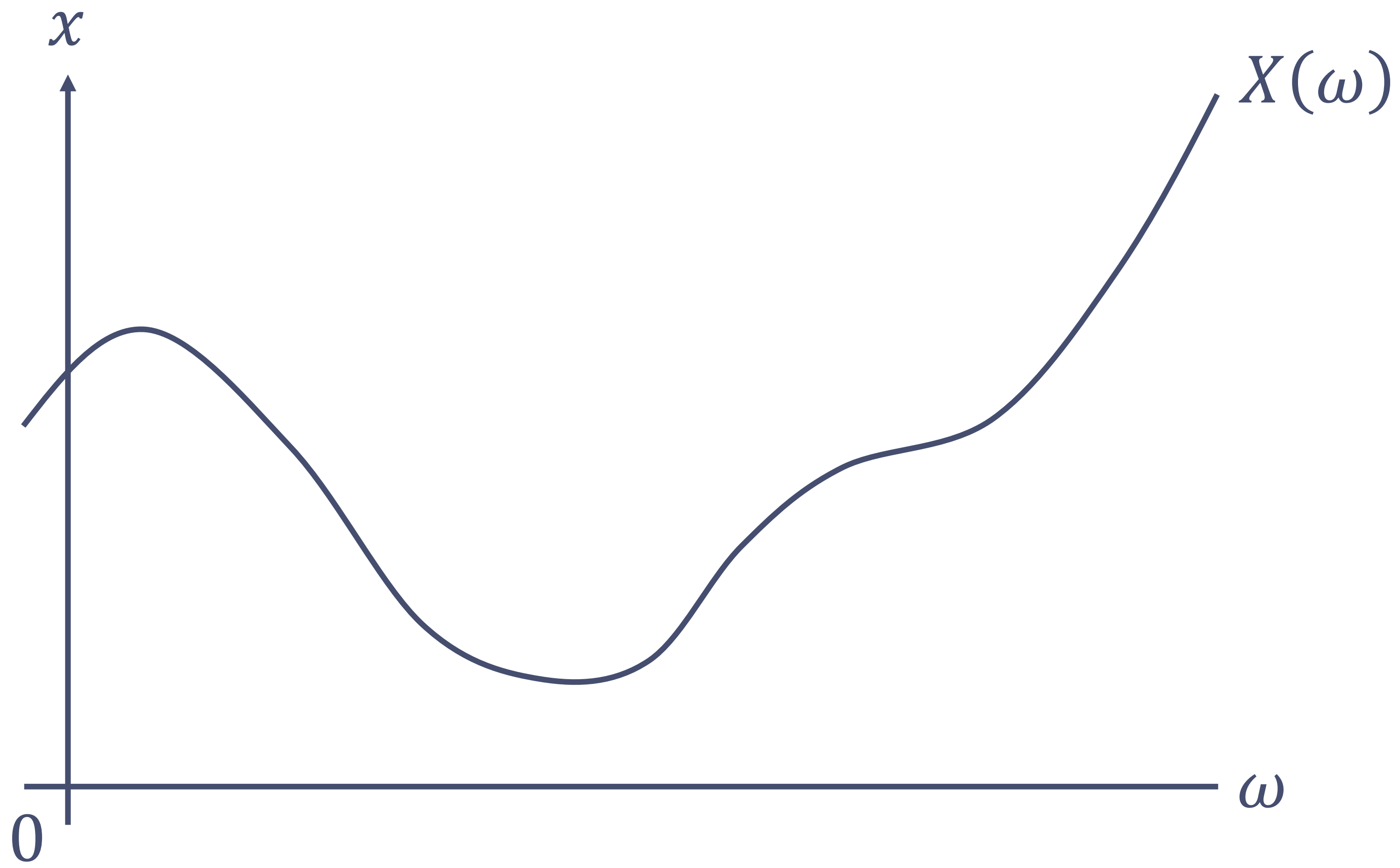
$$\mathbb{E}[X] \coloneqq \lim_{n \to \infty} \mathbb{E}[X_n]$$

のように定義できないだろうか?

➔ どんな離散型確率変数列 $\{X_n\}_{n=1}^{\infty}$ を作ればよいだろうか?

値域を分割しよう

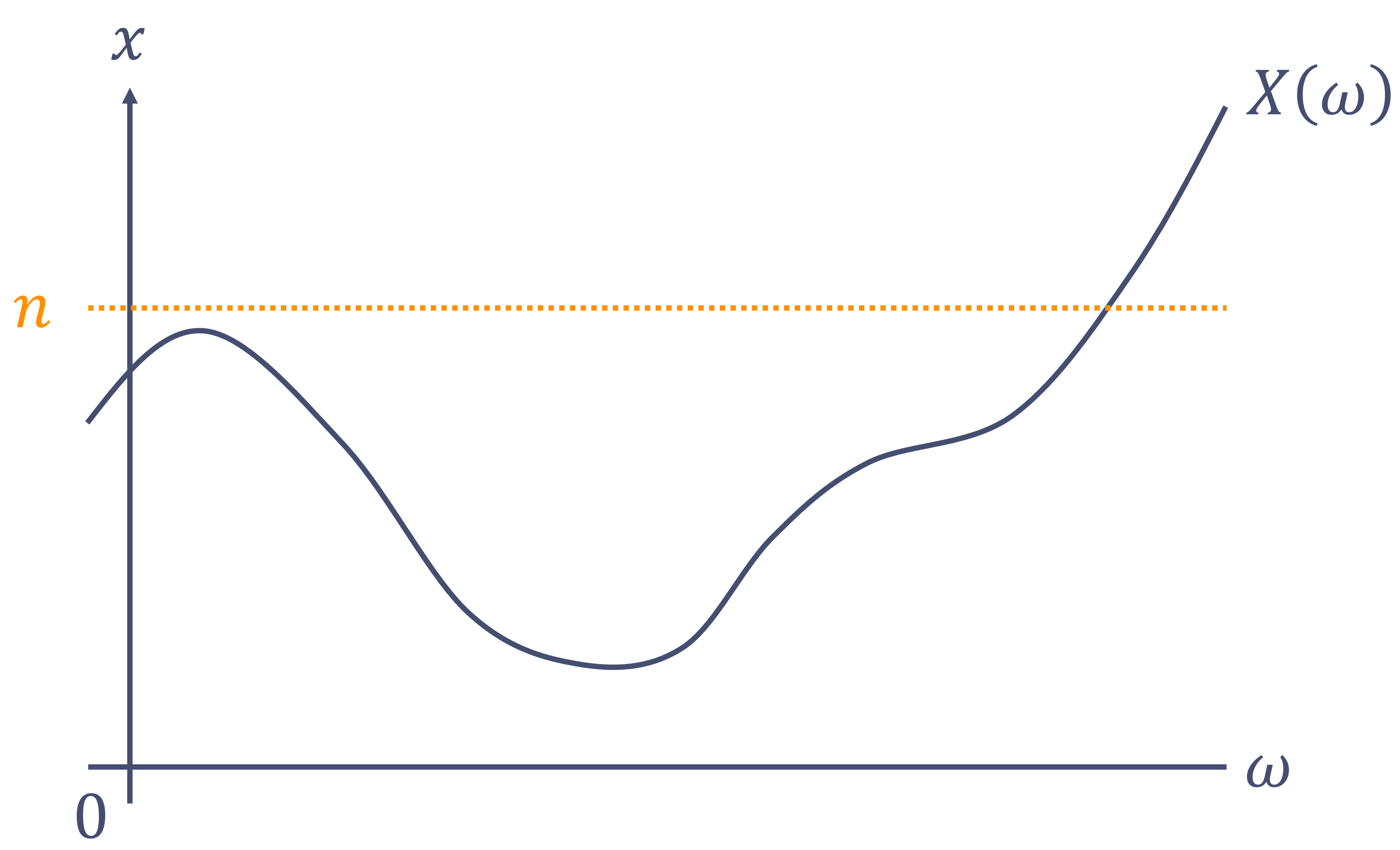
まずは X を連続型の**非負値**確率変数とする.
この X に対して, 次のとおり離散型確率変数列 $\{X_n\}_{n=1}^{\infty}$ を作ろう:



値域を分割しよう

まずは X を連続型の**非負値**確率変数とする.
この X に対して, 次のとおり離散型確率変数列 $\{X_n\}_{n=1}^{\infty}$ を作ろう:

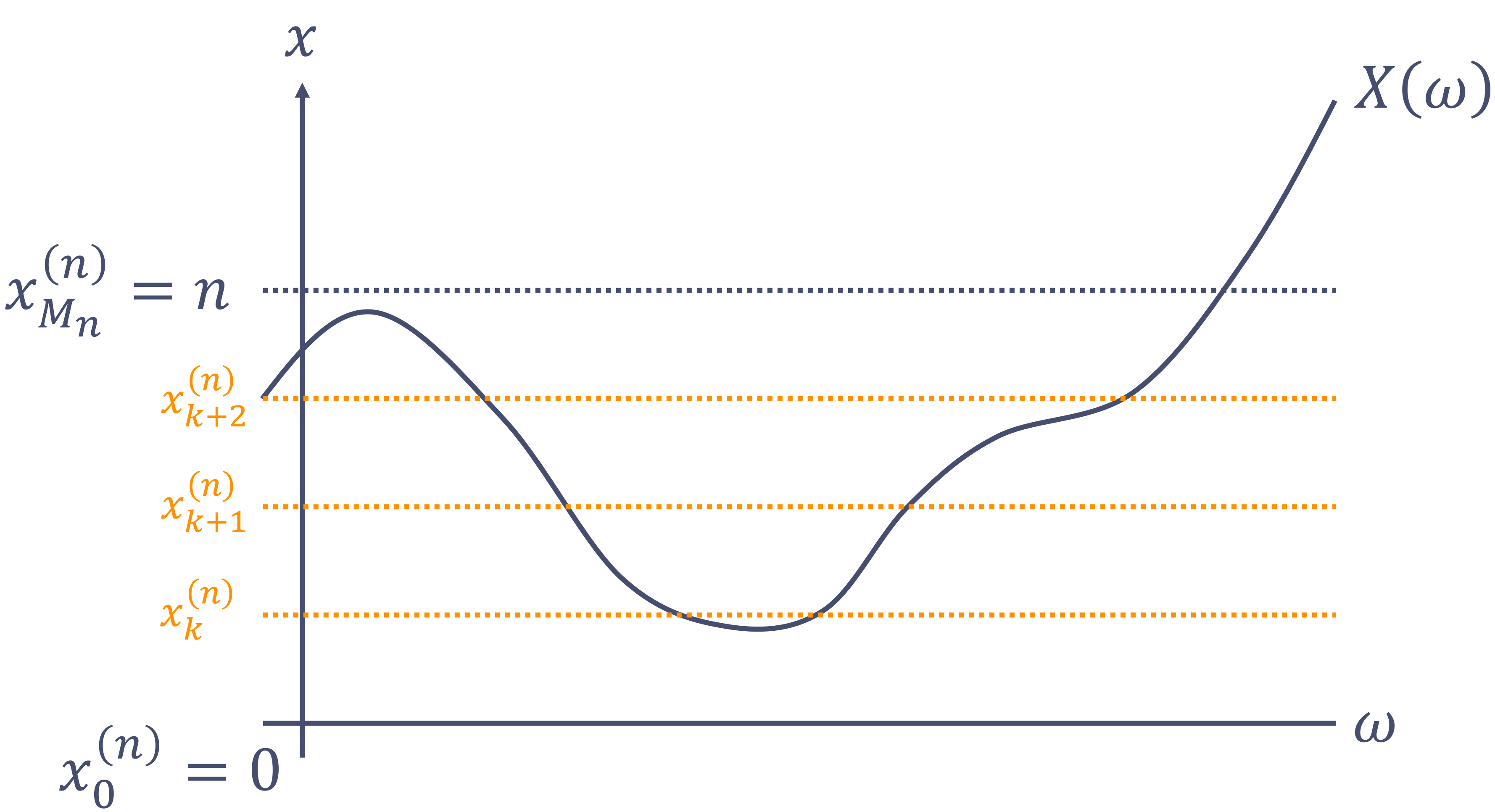
1. $X(\omega)$ の**値域**を $0 \leq X(\omega) < n$ の部分と $X(\omega) \geq n$ の部分に分ける.



値域を分割しよう

まずは X を連続型の**非負値**確率変数とする.
この X に対して, 次のとおり離散型確率変数列 $\{X_n\}_{n=1}^{\infty}$ を作ろう:

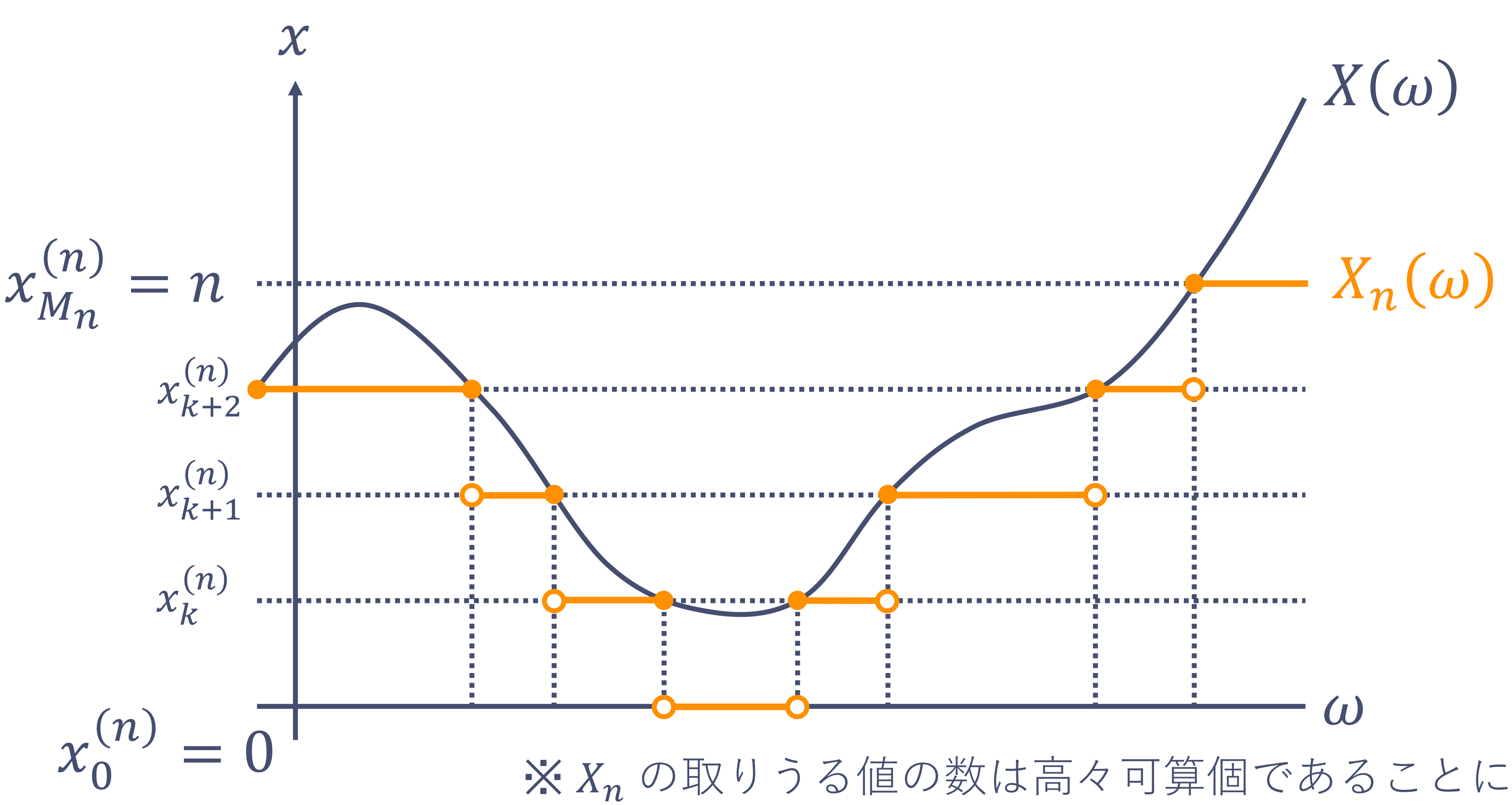
2. $0 \leq X(\omega) < n$ **の部分**をさらに $n \cdot 2^n$ 等分する. ($M_n \coloneqq n \cdot 2^n$ とし,
$k = 0, 1, \ldots, M_n$ に対して, $x_k^{(n)} \coloneqq k \cdot 2^{-n}$ かつ, $x_{M_n+1}^{(n)} \coloneqq \infty$ と書く.)



値域を分割しよう

まずは X を連続型の**非負値**確率変数とする.
この X に対して, 次のとおり離散型確率変数列 $\{X_n\}_{n=1}^{\infty}$ を作ろう:

$$3.\ X_n(\omega) \coloneqq \begin{cases} x_k^{(n)} & \left(x_k^{(n)} \leq X(\omega) < x_{k+1}^{(n)} \text{ のとき } (k = 0, \dots, M_n - 1)\right) \\ n & \left(n = x_{M_n}^{(n)} \leq X(\omega) < x_{M_n+1}^{(n)} = \infty \text{ のとき}\right) \end{cases} \text{とする.}$$



※ X_n の取りうる値の数は高々可算個であることに注意.

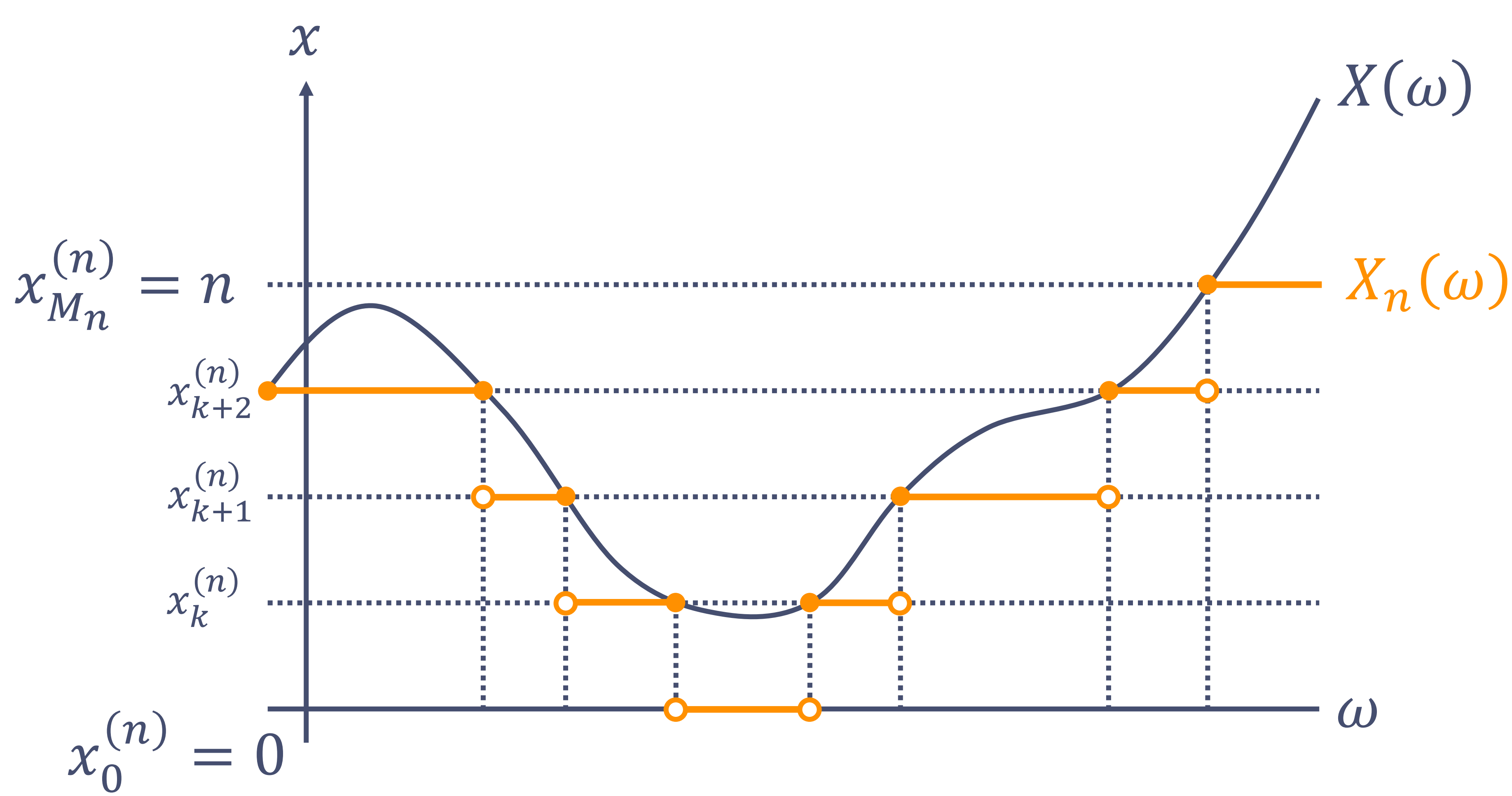
X_n の性質

n を一つ固定する. $k = 0, 1, \ldots, M_n$ に対しては, $x_k^{(n)} = k \cdot 2^{-n}$ だったので

$$x_{2k}^{(n+1)} = 2k \cdot 2^{-(n+1)} = k \cdot 2^{-n} = x_k^{(n)}.$$

➔ n のときの分割は $n+1$ **のときも維持される** (分割が細かくなっただけ).

➔ X_n たちの作り方から, **すべての** $\omega \in \Omega$ **で** $X_n(\omega) \le X_{n+1}(\omega)$ が成り立つ!
さらに, **すべての** $\omega \in \Omega$ **で** $X_n(\omega) \le X(\omega)$ も成り立つ!



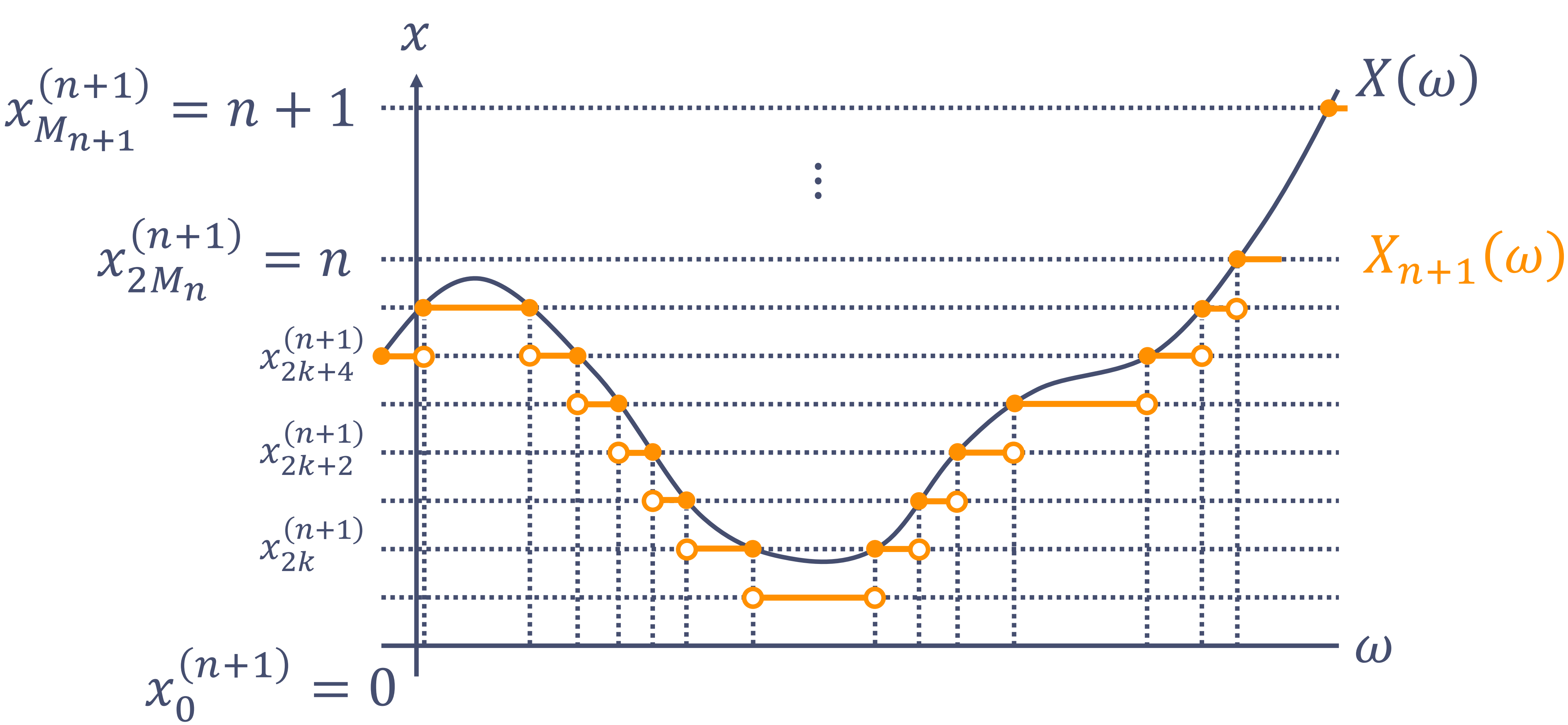
X_n の性質

n を一つ固定する. $k = 0, 1, \ldots, M_n$ に対しては, $x_k^{(n)} = k \cdot 2^{-n}$ だったので

$$x_{2k}^{(n+1)} = 2k \cdot 2^{-(n+1)} = k \cdot 2^{-n} = x_k^{(n)}.$$

➔ n のときの分割は $n+1$ **のときも維持される** (分割が細かくなっただけ).

➔ X_n たちの作り方から, **すべての** $\omega \in \Omega$ **で** $X_n(\omega) \leq X_{n+1}(\omega)$ が成り立つ!
さらに, **すべての** $\omega \in \Omega$ **で** $X_n(\omega) \leq X(\omega)$ も成り立つ!



X_n は X の近似列

いま, $\omega \in \Omega$ **を固定した実数列** $\{X_n(\omega)\}_{n=1}^{\infty} \subseteq \mathbb{R}$ を考えてみよう.

1. すべての n で $X_n(\omega) \leq X_{n+1}(\omega) \Rightarrow$ この数列は**単調増加**.
2. すべての n で $X_n(\omega) \leq X(\omega)$

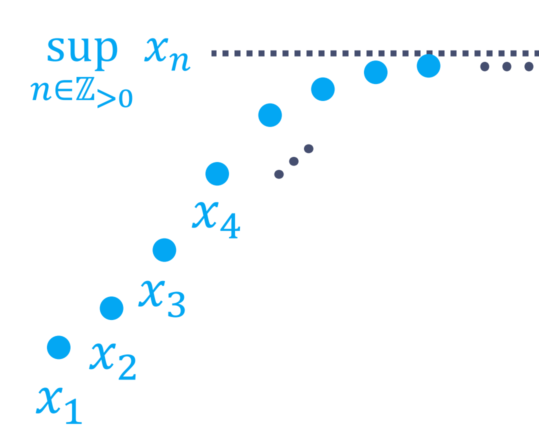
任意の n に対して $X_n(\omega) \leq u$
↓ をみたすような最小の u のこと.

$\Rightarrow$ この数列は上に有界. とくに, $X(\omega)$ は $X_n(\omega)$ の**上限**

$$X(\omega) = \sup_{n \in \mathbb{Z}_{>0}} X_n(\omega).$$

すると, **実数の連続性** ← 詳しくは微積分学の教科書を参照.

「単調増加する実数列は, その上限に収束する」



より,

$$X(\omega) = \lim_{n \to \infty} X_n(\omega), \qquad \forall \omega \in \Omega$$

つまり, **X_n が Ω 上で X に各点収束する**ことがわかる!

➔ X_n は n が増大するごとに X を精度良く近似するようになる!
上の1, 2が成り立つ離散型確率変数の列 $\{X_n\}$ を**近似単関数列**という.

連続型確率変数の期待値

連続型の非負値確率変数 X の近似単関数列 $\{X_n\}_{n=1}^{\infty}$ をとる.

各 X_n は離散型で, 取りうる値は $x_k^{(n)}$ $(k = 0, \ldots, M_n)$.

➔ X_n の期待値は, X_n の確率関数を p_n とすると

$$\mathbb{E}[X_n] = \sum_{k=0}^{M_n} x_k^{(n)} p_n\left(x_k^{(n)}\right).$$

ここで, $\{\mathbb{E}[X_n]\}_{n=1}^{\infty}$ は単調増加 (証明は次ページ).

よって, 実数の連続性より

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \sup_{n\in\mathbb{Z}_{>0}} \mathbb{E}[X_n].$$

➔ $\lim_{n\to\infty} \mathbb{E}[X_n]$ が (∞ となるときも含めて) きちんと定まったので,
これを $\mathbb{E}[X]$ と定めれば良い!
しかも, これは **X が離散型のときの定義と一致する**!

cf. $\{\mathbb{E}[X_n]\}$ が単調増加である理由

$$
\begin{aligned}
\mathbb{E}[X_{n+1}] &= \sum_{k=0}^{M_{n+1}} x_k^{(n+1)} p_{n+1}\left(x_k^{(n+1)}\right) && \leftarrow \text{定義} \\
&= \sum_{k=0}^{2M_n} x_k^{(n+1)} p_{n+1}\left(x_k^{(n+1)}\right) + \sum_{k=2M_n+1}^{M_{n+1}} x_k^{(n+1)} p_{n+1}\left(x_k^{(n+1)}\right) && \leftarrow \text{和の分割} \\
&\geq \sum_{k=0}^{2M_n} x_k^{(n+1)} p_{n+1}\left(x_k^{(n+1)}\right) && \leftarrow x_k^{(n+1)} p_{n+1}\left(x_k^{(n+1)}\right) \geq 0 \text{ となり第2項が非負なので} \\
&= \sum_{k=0}^{M_n} \left(x_{2k}^{(n+1)} p_{n+1}\left(x_{2k}^{(n+1)}\right) + x_{2k+1}^{(n+1)} p_{n+1}\left(x_{2k+1}^{(n+1)}\right)\right) && \leftarrow \text{2項ずつの和に書き換え} \\
&> \sum_{k=0}^{M_n} x_{2k}^{(n+1)} \left(p_{n+1}\left(x_{2k}^{(n+1)}\right) + p_{n+1}\left(x_{2k+1}^{(n+1)}\right)\right) && \leftarrow x_{2k+1}^{(n+1)} > x_{2k}^{(n+1)} \\
&= \sum_{k=0}^{M_n} x_k^{(n)} p_n\left(x_k^{(n)}\right) && \leftarrow x_{2k}^{(n+1)} = x_k^{(n)} \text{ と } p_{n+1}\left(x_{2k}^{(n+1)}\right) + p_{n+1}\left(x_{2k+1}^{(n+1)}\right) = p_n\left(x_k^{(n)}\right) \text{ より} \\
&= \mathbb{E}[X_n]. && \blacksquare
\end{aligned}
$$

期待値の定義

Def. (期待値)

確率空間 $(\Omega, \mathcal{F}, \Pr)$ と1次元の確率変数 X が与えられたとする.

(1) X が非負値の場合は, X の近似単関数列 $\{X_n\}_{n=1}^{\infty}$ によって定まる

$$\mathbb{E}[X] \coloneqq \lim_{n\to\infty} \mathbb{E}[X_n]$$

を X の**期待値** (expectation) という.

(2) 一般の X の場合,

$$X = X_+ - X_-, \qquad X_+ \coloneqq \max\{X, 0\}, \qquad X_- \coloneqq \max\{-X, 0\}$$

と非負値確率変数の差に分解し, $\mathbb{E}[X_+], \mathbb{E}[X_-]$ のいずれかが有限のとき,

$$\mathbb{E}[X] \coloneqq \mathbb{E}[X_+] - \mathbb{E}[X_-]$$

で X の期待値を定める. とくに, $\mathbb{E}[X_+], \mathbb{E}[X_-] < \infty$ のとき,

X は**可積分** (integrable) という.

➔ 期待値が定義できた! しかし, **定義通り期待値を計算するのは大変**......

cf. 近似単関数列について

確率変数 X の近似単関数列 $\{X_n\}_{n=1}^{\infty}$ を具体的に構成して期待値を定義した.

近似単関数列は, 各 $\omega \in \Omega$ で

1. すべての n で $X_n(\omega) \leq X_{n+1}(\omega)$.
2. すべての n で $X_n(\omega) \leq X(\omega)$.

が成り立ってさえいればよいので, 先程構成したものでなくても良い.

近似単関数列のとり方によらず, 期待値は同じ値になるのか?
(このことを「期待値はwell-definedか?」と表現することが多い.)

➔ 実は, **どのような近似単関数列をとっても, 同じ値に収束する**!

➔ 期待値はwell-defined.

ここまでのまとめ

- **期待値**: 確率変数が平均的にどの値をとるかを表す値.
- **離散型確率変数の期待値**: 取りうる値たちの確率による重み付き平均
- **連続型確率変数の期待値**:
  近似単関数列をとって, 離散型確率変数の期待値の極限を使って定義.
  定義通り計算するのは大変.
  - 近似単関数列の作り方: 確率変数の値域を細かく分割していく.

期待値の計算と性質

期待値の定義を書き直す

定義通り期待値を計算するのは大変. より簡単な計算法を考えていこう.

再び, X を**非負値**確率変数とし, X の確率分布を P とする.

先程具体的に構成した近似単関数列 $\{X_n\}_{n=1}^{\infty}$ を使うと,

$$\mathbb{E}[X_n] = \sum_{k=0}^{M_n} x_k^{(n)} p_n\left(x_k^{(n)}\right).$$

X_n の作り方から,

$$p_n\left(x_k^{(n)}\right) = \Pr\left(X_n = x_k^{(n)}\right) = \Pr\left(x_k^{(n)} \leq X < x_{k+1}^{(n)}\right) = \mathrm{P}\left(\left[x_k^{(n)}, x_{k+1}^{(n)}\right)\right)$$

が成り立つので,

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n] = \lim_{n\to\infty}\left(\sum_{k=0}^{M_n} x_k^{(n)} \mathrm{P}\left(\left[x_k^{(n)}, x_{k+1}^{(n)}\right)\right)\right).$$

$n \to \infty$ のとき $M_n \to \infty$ で, 区間 $\left[x_k^{(n)}, x_{k+1}^{(n)}\right)$ はだんだん狭くなっていく.

➔ 「$\mathbb{E}[X] = \int_0^\infty x\, \mathrm{P}(\mathrm{d}x)$」のような**積分**が定義できそう?

新たな積分の定義に向けて

もう少し一般化して, 可測関数 $g\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ があったときに, $\mathbb{E}[g(X)]$ がどう書けるかを考えてみよう.

$Y = g(X)$ として, Y の確率分布を Q とすると

$$\mathbb{E}[g(X)] = \lim_{n\to\infty}\left(\sum_{k=0}^{M_n} y_k^{(n)}\mathrm{Q}\left(\left[y_k^{(n)}, y_{k+1}^{(n)}\right)\right)\right)$$

と表せる ($y_k^{(n)}$ たちは, p.88の $x_k^{(n)}$ と同様の定義). ところで,

$$\begin{aligned}\mathrm{Q}\big([a,b)\big) &= \Pr\big(Y = g(X) \in [a,b)\big)\\ &= \Pr\left(X \in g^{-1}\big([a,b)\big)\right) = \mathrm{P}\left(g^{-1}\big([a,b)\big)\right)\end{aligned}$$

なので,

$$\mathbb{E}[g(X)] = \lim_{n\to\infty}\left(\sum_{k=0}^{M_n} y_k^{(n)}\mathrm{P}\left(g^{-1}\left(\left[y_k^{(n)}, y_{k+1}^{(n)}\right)\right)\right)\right).$$

➔ 「$\mathbb{E}[g(X)] = \int_0^\infty y\,\mathrm{P}\big(g^{-1}(\mathrm{d}y)\big)$」と書いても良さそう? さらに, 形式的に $x = g^{-1}(y)$ とすると, $\mathbb{E}[g(X)] = \int_0^\infty g(x)\,\mathrm{P}(\mathrm{d}x)$? きちんと定義してみよう.

Lebesgue–Stieltjes積分

Def. (Lebesgue–Stieltjes積分)

与えられた確率分布 P に対し, 可測関数 $g\colon \mathbb{R} \to \mathbb{R}$ を $g = g_+ - g_-$ と分解し,

$$L_\pm \coloneqq \lim_{n\to\infty}\left(\sum_{k=0}^{M_n} x_k^{(n)} \mathrm{P}\left(g_\pm^{-1}\left(\left[x_k^{(n)}, x_{k+1}^{(n)}\right)\right)\right)\right) \text{ (複号同順)}$$

と定める. このとき, L_+, L_- のいずれかが有限なら,

$$\int_{\mathbb{R}} g(x)\,\mathrm{P}(\mathrm{d}x) \coloneqq L_+ - L_-$$

と書いて, これを P による g の**Lebesgue–Stieltjes積分**(ルベーグ スティルチェス)という.

$L_+ < \infty$ かつ $L_- < \infty$ のとき, g は **P-可積分** (P-integrable) という.

また, $A \in \mathcal{B}_1$ のとき, $\int_A g(x)\,\mathrm{P}(\mathrm{d}x) \coloneqq \int_{\mathbb{R}} g(x) 1_A(x)\,\mathrm{P}(\mathrm{d}x)$ と定め,

とくに $A = (a, b]$ **のとき** $\int_a^b g(x)\,\mathrm{P}(\mathrm{d}x) \coloneqq \int_{(a,b]} g(x)\,\mathrm{P}(\mathrm{d}x)$ と書く.

※ この積分は, 確率分布 P の代わりに一般の測度 μ を使っても定義できる.
とくに, $\mu\big((a,b]\big) = b - a$ となる測度のとき, **Lebesgue積分**と呼ばれる.
Lebesgue積分のとき, $\mu(\mathrm{d}x)$ を単に $\mathrm{d}x$ と書くことが多い.

期待値をより計算しやすく

特に, 確率変数 X (非負値でなくても良い) が $X \sim \mathrm{P}$ のとき $\mathbb{E}[g(X)]$ は

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\,\mathrm{P}(\mathrm{d}x)\,.$$

(X に関する期待値をとることを明示する場合, $\mathbb{E}_X[g(X)]$ とも書く. さらに,

$X \sim \mathrm{P}$ を明示する場合, $\mathbb{E}_{X\sim\mathrm{P}}[g(X)]$ と書くこともある.)

➔ 期待値をLebesgue–Stieltjes積分で書き直すことができた!
(まだ書き直しただけ.)

X が $\{a_1, a_2, \ldots\}$ の値をとる**離散型確率変数** (確率関数: p_X) なら,
定義通り計算すると

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\,\mathrm{P}(\mathrm{d}x) = \sum_{i=1}^{\infty} g(a_i)\cdot p_X(a_i)\,.$$

X が連続型の場合, 定義通り計算するしかない……?

➔ 実は, **絶対連続型の場合は, より計算しやすい方法がある**!

絶対連続型分布での期待値計算

確率密度関数を使うと, 期待値が計算できる! (証明は略.)

Prop. (絶対連続型分布についての期待値計算)

1次元確率変数 $X \sim \mathrm{P}$ は絶対連続型で, その確率密度関数を f とする.

可測関数 $g\colon \mathbb{R} \to \mathbb{R}$ が P-可積分であるとき,

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\,\mathrm{P}(\mathrm{d}x) = \int_{\mathbb{R}} g(x) f(x)\,\mathrm{d}x\,.$$

この定理は d-次元確率変数 $\boldsymbol{X}$ と可測関数 $g\colon \mathbb{R}^d \to \mathbb{R}$ に対しても成り立つ.

右辺は, (大抵の場合は) **通常のRiemann積分として計算して問題ない**!

➔ これなら計算できそう. 期待値の計算方法についてまとめておこう.

期待値計算のまとめ

期待値の計算方法をまとめると以下のとおり:

Thm. (期待値の計算)

d-次元確率変数 $\boldsymbol{X} \sim \mathrm{P}$ は**離散型** ($\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots\}$ をとり, 確率関数: $p_{\boldsymbol{X}}$) か, **絶対連続型** (確率密度関数: f). P-可積分な可測関数 $g\colon \mathbb{R}^d \to \mathbb{R}$ に対し,

$$\mathbb{E}[g(\boldsymbol{X})] = \int_{\mathbb{R}^d} g(\boldsymbol{x})\,\mathrm{P}(\mathrm{d}\boldsymbol{x}) = \begin{cases} \displaystyle\sum_{i=1}^{\infty} g(\boldsymbol{a}_i) \cdot p_{\boldsymbol{X}}(\boldsymbol{a}_i) & (\boldsymbol{X}\text{: 離散型}) \\ \displaystyle\int_{\mathbb{R}^d} g(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} & (\boldsymbol{X}\text{: 絶対連続型}) \end{cases}$$

➔ Lebesgue–Stieltjes積分による表記は,「**離散型と絶対連続型の期待値をまとめて表現するための記号**」と思っておけば十分.

cf. L-S積分の積分区間について

$$\int_a^b g(x)\,\mathrm{P}(\mathrm{d}x) \coloneqq \int_{(a,b]} g(x)\,\mathrm{P}(\mathrm{d}x)$$

と定義されていた. $\int_{[a,b]} g(x)\,\mathrm{P}(\mathrm{d}x)$ はどうなるだろうか?

$$\int_{[a,b]} g(x)\,\mathrm{P}(\mathrm{d}x) = \int_{\mathbb{R}} g(x)1_{[a,b]}(x)\,\mathrm{P}(\mathrm{d}x)$$

← 定義.

$$= \int_{\mathbb{R}} g(x)\left(1_{\{a\}}(x) + 1_{(a,b]}(x)\right)\,\mathrm{P}(\mathrm{d}x)$$

← $[a,b] = \{a\} \cup (a,b]$ (非交差和)
また, 一般に集合 A, B に対し,
$1_{A\cup B}(x)$
$= 1_A(x) + 1_B(x) - 1_{A\cap B}(x)$.

$$= \int_{\mathbb{R}} g(x)1_{\{a\}}(x)\,\mathrm{P}(\mathrm{d}x) + \int_a^b g(x)\,\mathrm{P}(\mathrm{d}x)$$

← 積分の線形性

$$= \mathbb{E}\left[g(X)1_{\{a\}}(X)\right] + \int_a^b g(x)\,\mathrm{P}(\mathrm{d}x)$$

← 期待値による書き換え

$$= g(a)\mathrm{P}(\{a\}) + \int_a^b g(x)\,\mathrm{P}(\mathrm{d}x)\,.$$

← $g(X)1_{\{a\}}(X)$ は
確率 $\mathrm{P}(\{a\})$ で $g(a)$,
確率 $1 - \mathrm{P}(\{a\})$ で 0 をとる
確率変数.

➔ P が連続型なら $\int_{[a,b]} g(x)\,\mathrm{P}(\mathrm{d}x) = \int_a^b g(x)\,\mathrm{P}(\mathrm{d}x)$ だが,
そうでない場合は $g(a)\mathrm{P}(\{a\})$ **の項が必要**になる.

cf. Riemann積分との関連

Def. (Riemann-Stieltjes積分)

※ この積分は, 分布関数 F の代わりに
一般の有界変動関数を使っても定義できる.

与えられた分布関数 F と関数 $g\colon [a,b] \to \mathbb{R}$ に対し,

分割 $\Delta\colon a = x_0 < x_1 < \cdots < x_n = b$ をとって, $|\Delta| \coloneqq \max\limits_{1 \le k \le n} |x_k - x_{k-1}|$ とする.

$R_+ \coloneqq \lim\limits_{|\Delta| \to 0} \sum_{k=1}^{n} \sup\limits_{x \in [x_{k-1}, x_k]} g(x) \left(F(x_k) - F(x_{k-1}) \right)$ と

$R_- \coloneqq \lim\limits_{|\Delta| \to 0} \sum_{k=1}^{n} \inf\limits_{x \in [x_{k-1}, x_k]} g(x) \left(F(x_k) - F(x_{k-1}) \right)$ が一致するならば,

g は **F-Riemann可積分**といい, $\int_{[a,b]} g(x)\, \mathrm{d}F(x) \coloneqq R_+ (= R_-)$ と定める.

これを F による g の**Riemann-Stieltjes積分**という.

P_F を F に対応する確率分布とすると, 実は
「g: F-**Riemann可積分** $\Rightarrow$ g: P_F-**可積分**」が成立し, **積分値が一致する**.

逆は必ずしも正しくない.

➔ **Lebesgue-Stieltjes積分のほうが, 多くの関数を積分できる**!
(広義Riemann-Stieltjes積分の場合はもう少し条件が必要.)

cf. 「可積分」について

「確率変数 X が可積分」というのは, Lebesgue–Stieltjes積分での関数の可積分性と何か関係があるのか?

$\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ としよう. 実は, 標本空間 Ω 上のLebesgue–Stieltjes積分も定義できて, $g(\boldsymbol{X})$ の期待値は

$$\mathbb{E}[g(\boldsymbol{X})] \coloneqq \int_\Omega g\big(\boldsymbol{X}(\omega)\big)\,\mathrm{Pr}(\mathrm{d}\omega)$$

と (本来は) 定義される.

($\boldsymbol{X}$ の確率分布を $\mathrm{P}_{\boldsymbol{X}}$ とすると, $\boldsymbol{X}$ による変数変換

$$\int_\Omega g\big(\boldsymbol{X}(\omega)\big)\,\mathrm{Pr}(\mathrm{d}\omega) = \int_{\mathbb{R}^d} g(\boldsymbol{x})\,\mathrm{P}_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x})$$

が成り立つことが示せるので, はじめから右辺の形で期待値を定義した.)

とくに, 1次元確率変数 X に対しては, $\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{Pr}(\mathrm{d}\omega)$.

➔ $X\colon \Omega \to \mathbb{R}$ が可積分というのは, **X が Pr-可積分**ということ.

期待値の基本性質

以下の性質は, 期待値の定義を用いて示すことができる. (証明は略.)

Prop. (期待値の基本性質)

1次元の確率変数 X, Y は可積分とする.

(1) [定数の期待値] $X = 1, a.s.$ ならば, $\mathbb{E}[X] = 1$.

(2) [**線形性**] 任意の $a, b \in \mathbb{R}$ に対し, $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

(3) [**単調性**] $X \leq Y, a.s.$ ならば, $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

(4) [期待値0のa.s.非負の確率変数]
$X \geq 0, a.s.$ かつ $\mathbb{E}[X] = 0$ ならば, $X = 0, a.s.$

(5) [期待値の絶対値は絶対値の期待値以下]
$|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.

➔ 具体的な計算をする際には, とくに線形性が重要.

独立な確率変数と期待値

確率変数たちの積の期待値についても, 有用な性質がある:

Prop. (独立な確率変数の積の期待値)

1次元確率変数 $X_1, \dots, X_n$ は**独立**で, Borel可測関数 $g_i\colon \mathbb{R} \to \mathbb{R}$ $(i = 1, \dots, n)$ に対して $g_i(X_i)$ は可積分とする. このとき, $g_1(X_1), \dots, g_n(X_n)$ も独立で,

$$\mathbb{E}\left[\prod_{i=1}^{n} g_i(X_i)\right] = \prod_{i=1}^{n} \mathbb{E}[g_i(X_i)]$$

が成立する.

注意:

- **独立でないと使えない**!
- この逆「確率変数の積の期待値が期待値の積 ⇒ 独立」は **一般に成立しない**! (cf. 清水『統計学への確率論, その先へ』)

➔ 後で頻繁に利用する性質. 独立性の確認を忘れないように.

cf. 命題の証明

g_i はBorel可測だから, 任意の $A_i \in \mathcal{B}_1$ に対して $g_i^{-1}(A_i) \in \mathcal{B}_1$.
$X_1, \dots, X_n$ は独立なので,

$$\begin{aligned}
\Pr(g_1(X_1) \in A_1, \dots, g_n(X_n) \in A_n) &= \Pr\left(X_1 \in g_1^{-1}(A_1), \dots, X_n \in g_n^{-1}(A_n)\right) \\
&= \prod_{i=1}^{n} \Pr\left(X_i \in g_i^{-1}(A_i)\right) \\
&= \prod_{i=1}^{n} \Pr(g_i(X_i) \in A_i).
\end{aligned}$$

よって, $g_1(X_1), \dots, g_n(X_n)$ も独立.

$\boldsymbol{X} = (X_1, \dots, X_n)$ の分布関数を $F_{\boldsymbol{X}}$ とし, 各 X_i の分布関数を F_i とする.
X_i の独立性から $F_{\boldsymbol{X}}(x_1, \dots, x_n) = \prod_{i=1}^{n} F_i(x_i)$. よって,

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{n} g_i(X_i)\right] &= \int_{\mathbb{R}^n} \prod_{i=1}^{n} g_i(x_i)\, F_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x}) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^{n} g_i(x_i)\, F_1(\mathrm{d}x_1) \dots F_n(\mathrm{d}x_n) \\
&= \prod_{i=1}^{n} \int_{\mathbb{R}} g_i(x_i) F_i(\mathrm{d}x_i) = \prod_{i=1}^{n} \mathbb{E}[g_i(X_i)].
\end{aligned}$$

■

Markovの不等式

確率変数の値が大きくなる確率を評価するのに, 次の不等式が利用できる.

Thm. (**Markovの不等式**)

d-次元確率変数 $\boldsymbol{X}$ と, 非負値可測関数 $h: \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ に対して,

$h(\boldsymbol{X})$ が可積分なら, 任意の $\varepsilon > 0$ に対して

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

ε が大きいと, 不等式の右辺は小さくなっていく.

➔ 「$h(\boldsymbol{X})$ の値が大きくなる確率は, 値が大きくなるほど小さくなる」

➔ 実は, この不等式はもう少し一般化できる.

Chebyshevの不等式

Markovの不等式を一般化したのが, 次のChebyshevの不等式.

Thm. (**Chebyshevの不等式**)

d-次元確率変数 $\boldsymbol{X}$ と, $\mathbb{R}_{>0}$ 上の正値非減少可測関数 $\nu\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ に対して, $\nu(\boldsymbol{X})$ が可積分なら, 任意の $\varepsilon > 0$ に対して

$$\Pr(\|\boldsymbol{X}\| > \varepsilon) \leq \frac{\mathbb{E}[\nu(\|\boldsymbol{X}\|)]}{\nu(\varepsilon)}.$$

この式で, $\boldsymbol{X}$ を非負値な $h(\boldsymbol{X})$ で置き換えて, $\nu(x) \coloneqq |x|\ (x \geq \varepsilon)$ とすれば,

Markovの不等式が得られる.

→ **確率値の上からの評価をする際に便利な式.** 後述の「確率収束」を示す際によく利用される.

Jensenの不等式

次のJensenの不等式は, 期待値の評価をする際に有用:

<u>Thm.</u> (**Jensenの不等式**)

$g: \mathbb{R} \to \mathbb{R}$ は凸関数とする, i.e., 任意の $x, y \in \mathbb{R}$ と任意の $\theta \in (0,1)$ で,

$$g(\theta x + (1-\theta)y) \leq \theta g(x) + (1-\theta)g(y)$$

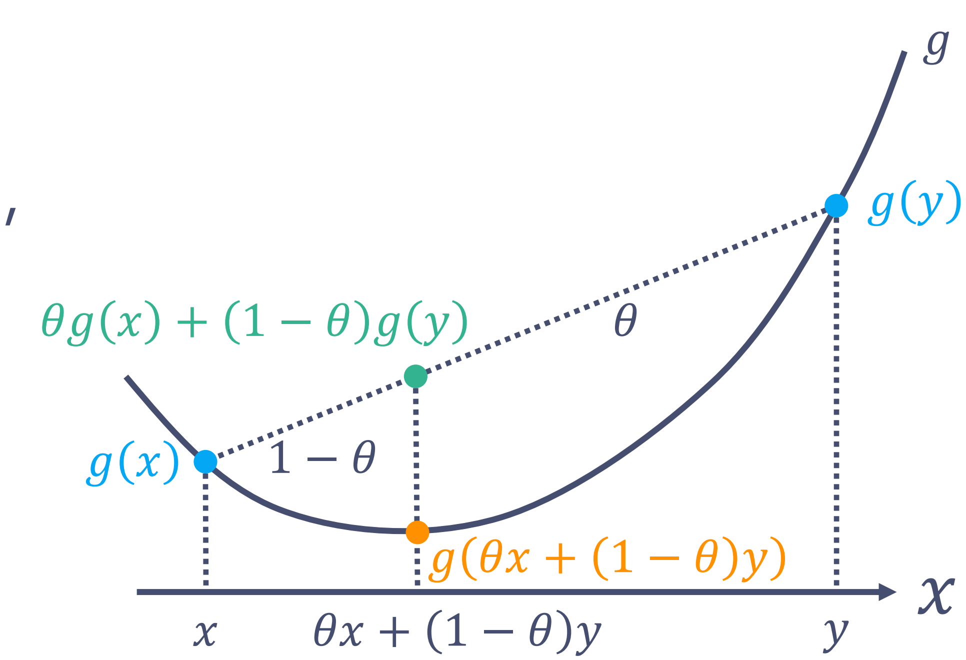
が成立するとする. このとき, $X, g(X)$ が可積分なら, $g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$.

等号成立は, X が定数, a.s. のときに限る.

➔ 「EMアルゴリズム」の収束性を言うときや,

「KLダイバージェンス」の非負性を言う

ときなど, 用途は色々.



Cauchy–Schwarzの不等式

最後に, よく見かける不等式を紹介しておこう.

Thm. (**Cauchy–Schwarzの不等式**)

1次元確率変数 X, Y は, ともに X^2, Y^2 が可積分であるとする. このとき,

$|\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$. 等号成立は $Y = kX\ (k \in \mathbb{R})$ となるときに限る.

※ Cauchy–Schwarzの不等式は, **Hölderの不等式**の特別な場合.

「X^2, Y^2 **が可積分な確率変数に対しては共分散が存在する**」ことを保証するための不等式になっている.

また, 後述の「相関係数」の値域を示すのにも利用される.

➔ 以上の不等式は, 証明などでよく現れる.
(完璧に覚えずとも, 思い出せるようにはしておくと良い.)

条件付き期待値

本来, 条件付き期待値を定義するには, 「Radon-Nikodymの定理」などの道具が必要になってきて, かなりややこしい.

確率変数が離散型/絶対連続型であるような特別な場合について,
結果のみ示しておく:

Thm. (条件付き期待値の計算)

d-次元確率変数 $\boldsymbol{X}$, d'-次元確率変数 $\boldsymbol{Y}$ は**離散型** ($\boldsymbol{X}$ は $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots\}$ をとる) か, **絶対連続型**. 可積分な可測関数 $g\colon \mathbb{R}^d \to \mathbb{R}$ に対し, $Y = y$ の下での**条件付き期待値** $\mathbb{E}[g(\boldsymbol{X}) | \boldsymbol{Y} = \boldsymbol{y}]$ は, 以下で計算できる:

$$\mathbb{E}[g(\boldsymbol{X}) | \boldsymbol{Y} = \boldsymbol{y}] = \begin{cases} \displaystyle\sum_{i=1}^{\infty} g(\boldsymbol{a}_i) \cdot p_{X|Y}(\boldsymbol{a}_i | \boldsymbol{y}) & (\boldsymbol{X}\text{: 離散型}) \\ \displaystyle\int_{\mathbb{R}^d} g(\boldsymbol{x}) f_{X|Y}(\boldsymbol{x} | \boldsymbol{y})\, \mathrm{d}\boldsymbol{x} & (\boldsymbol{X}\text{: 絶対連続型}) \end{cases}$$

➔ 期待値についての性質は, **条件付き期待値に対しても成立する**!

ここまでのまとめ

- 期待値はLebesgue-Stieltjes積分で書ける.
    - 具体的な計算では, 以下の公式を使えばよく, 通常のRiemann積分で計算すれば問題ない.
    - **期待値の計算**:

$$\mathbb{E}[g(\boldsymbol{X})] = \int_{\mathbb{R}^d} g(\boldsymbol{x})\,\mathrm{P}(\mathrm{d}\boldsymbol{x}) = \begin{cases} \displaystyle\sum_{i=1}^{\infty} g(\boldsymbol{a}_i)\cdot p_{\boldsymbol{X}}(\boldsymbol{a}_i) & (\boldsymbol{X}\text{: 離散型}) \\ \displaystyle\int_{\mathbb{R}^d} g(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} & (\boldsymbol{X}\text{: 絶対連続型}) \end{cases}$$

- **期待値の基本性質**:
    - 線形性, 単調性
    - Markovの不等式, Chebyshevの不等式, Jensenの不等式, Cauchy-Schwarzの不等式.
- **条件付き期待値**も同様の計算ができ, 同様の性質が成立.

モーメント・共分散

分布の特徴量

期待値を利用することで, 分布の「特徴量」を定義することができる:

Def. (モーメント, 平均, 分散, 標準偏差)

1次元確率変数 X と正整数 $k \in \mathbb{Z}_{>0}$ に対して,

$$\mu_k \coloneqq \mathbb{E}[X^k], \qquad \alpha_k \coloneqq \mathbb{E}[(X-\mu_1)^k]$$

とする. μ_k を原点周りの k 次の**モーメント** (moment) といい, とくに μ_1 を**平均** (mean) という. α_k を平均周りの k 次のモーメントといい, とくに

$$\mathbb{V}[X] \coloneqq \sigma^2 \coloneqq \alpha_2 = \mathbb{E}[(X-\mu_1)^2]$$

を**分散** (variance) という. σ を**標準偏差** (standard deviation) という.

平均は分布の「中心」(密度関数の重心) を表し,
分散はその「中心」からのばらつきを表す量になっている.

他にも, **歪度** (skewness) α_3/σ^3, **尖度** (kurtosis) α_4/σ^4 などの量もあり,
それぞれ分布の歪み具合や尖り具合を表す.

※ 尖度を $\frac{\alpha_4}{\sigma^4} - 3$ で定めることもある.

➔ モーメントを見ることで, **分布の様子を知ることができる**.

分散の基本性質

分散について, 以下の性質は基本的:

<u>Prop.</u> (分散の基本性質)

1次元確率変数 X は, 平均 μ, 分散 σ^2 をもつ (i.e., $\mu, \sigma^2 < \infty$) とする.

(1) [**分散公式**] $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

(2) [分散は非負] $\mathbb{V}[X] \geq 0$.

(3) [線形変換後の分散] 定数 $a, b \in \mathbb{R}$ に対して, $\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$.

<u>pf.</u>

(1) 期待値の線形性から

$$\mathbb{V}[X] = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2$$
$$= \mathbb{E}[X^2] - \mu^2.$$

(2) $(X - \mu)^2 \geq 0$ なので, 期待値の単調性から成立.

(3) $\mathbb{V}[aX + b] = \mathbb{E}\left[\left(aX + b - (a\mu + b)\right)^2\right] = \mathbb{E}[a^2(X-\mu)^2] = a^2\mathbb{V}[X]$. ■

確率変数の正規化

統計では平均が0, 分散が1となるように確率変数を**正規化** (normalization) することがよくある. (標準化: standardization ともいう.)

Prop. (確率変数の正規化)

1次元確率変数 X は, 平均 μ, 分散 σ^2 をもつ (i.e., $\mu, \sigma^2 < \infty$) とする.

$$Z \coloneqq \frac{X-\mu}{\sigma}$$

とおくと, $\mathbb{E}[Z] = 0$, $\mathbb{V}[Z] = 1$ となる.

pf.

期待値の線形性より, $\mathbb{E}[Z] = \frac{\mathbb{E}[X]-\mu}{\sigma} = 0$.

また, 前ページの性質 (3) より $\mathbb{V}[Z] = \frac{1}{\sigma^2}\mathbb{V}[X] = 1$. ■

低次のモーメントの存在

分散が非負より, $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ なので $|\mathbb{E}[X]| \leq \sqrt{\mathbb{E}[X^2]}$.
➔ $\mathbb{E}[X^2] < \infty \Longrightarrow \mathbb{E}[X] < \infty$ (**2次モーメントが存在すれば平均も存在する**).

実は, 一般に次が成立する:

Prop.

1次元確率変数 X と, $1 \leq p < q$ を満たす実数 $p, q \in \mathbb{R}$ に対し,

$$|\mathbb{E}[X^q]| < \infty \Longrightarrow |\mathbb{E}[X^p]| < \infty.$$

pf.

$$\begin{aligned} |\mathbb{E}[X^p]| &\leq \mathbb{E}[|X^p|] = \mathbb{E}[|X|^p] \\ &= \mathbb{E}\left[|X|^p 1_{\{|X| \leq 1\}}\right] + \mathbb{E}\left[|X|^p 1_{\{|X| > 1\}}\right] \\ &\leq 1 + \mathbb{E}\left[|X|^q 1_{\{|X| > 1\}}\right] \\ &\leq 1 + \mathbb{E}[|X|^q]. \end{aligned}$$

期待値の定義より $|\mathbb{E}[X^q]| = |\mathbb{E}[(X^q)_+] - \mathbb{E}[(X^q)_-]| < \infty$ なので,
$\mathbb{E}[|X|^q] = \mathbb{E}[|X^q|] = \mathbb{E}[(X^q)_+] + \mathbb{E}[(X^q)_-] < \infty$. よって, $|\mathbb{E}[X^p]| < \infty$. ■

➔ とくに $p, q \in \mathbb{Z}_{>0}$ の場合は, 「**高次のモーメントが存在すれば, それより低次のモーメントも存在する**」ということを述べている.

階乗モーメント

とくに離散型の確率変数に対しては, 次の「階乗モーメント」が有用なことがある:

Def. (階乗モーメント)

1次元確率変数 X と正整数 $k \in \mathbb{Z}_{>0}$ に対して,

$$\mathbb{E}[X(X-1)\cdots(X-k+1)]$$

を k 次の**階乗モーメント** (factorial moment) という.

これを利用すると, 分散が次のように表せる:

Prop.

1次元確率変数 X は, 平均 μ, 分散 σ^2 をもつ (i.e., $\mu, \sigma^2 < \infty$) とする.

このとき, $\mathbb{V}[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2$.

pf.

$\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ と $\mathbb{E}[X(X-1)] = \mathbb{E}[X^2] - \mathbb{E}[X]$ より. ■

➔ これを利用して, 具体的な分布で平均と分散を計算してみよう.

例: Poisson分布の平均

パラメータ $\lambda > 0$ のPoisson分布 (確率関数 $p(x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, x \in \mathbb{Z}_{\geq 0}$) に従う確率変数 X の平均を求めてみよう.

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}$$
← 離散型確率変数に対する期待値の計算.

$$= \mathrm{e}^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!}$$
← $\mathrm{e}^{-\lambda}$ は x によらない. $x = 0$ の項は 0 なので, $x = 1$ からの和でよい.
分母の x と約分.

$$= \mathrm{e}^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^{y+1}}{y!}$$
← 和の変数変換: $y = x - 1$ とおく.

$$= \lambda \mathrm{e}^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}$$
← λ を一つだけ和の外へ.

$$= \lambda \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda}$$
← 指数関数 e^{λ} の定義.

$$= \lambda.$$

➔ パラメータ $\lambda > 0$ のPoisson分布の平均は λ.

例: Poisson分布の分散

2次の階乗モーメントを使って確率変数 X の分散を求めてみよう.

$$
\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} && \leftarrow \text{離散型確率変数に対する期待値の計算.} \\
&= \mathrm{e}^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^x}{(x-2)!} && \leftarrow \mathrm{e}^{-\lambda}\text{ は } x \text{ によらない. } x=0,1 \text{ の項は } 0 \text{ なので, } x=2 \text{ からの和でよい. 約分をする.} \\
&= \mathrm{e}^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^{y+2}}{y!} && \leftarrow \text{和の変数変換: } y = x-2 \text{ とおく.} \\
&= \lambda^2\mathrm{e}^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} && \leftarrow \lambda \text{ を二つ分だけ和の外へ.} \\
&= \lambda^2\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} && \leftarrow \text{指数関数 } \mathrm{e}^{\lambda} \text{ の定義.} \\
&= \lambda^2.
\end{aligned}
$$

よって, $\mathbb{V}[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

➔ パラメータ $\lambda > 0$ のPoisson分布の分散も λ.
$\mathbb{E}[X^2]$ を計算するより**楽に分散が計算できる**.

確率変数の共分散と相関係数

二つの確率変数が独立でないとき, それらの関係を捉えるのに役立つのが共分散・相関係数.

Def. (共分散, 相関係数)

1次元確率変数 X, Y の平均をそれぞれ μ_X, μ_Y とする.

$$\mathrm{Cov}[X, Y] \coloneqq \mathbb{E}_{(X,Y)}[(X - \mu_X)(Y - \mu_Y)]$$

を X, Y の**共分散** (covariance) といい,

$$\rho[X, Y] \coloneqq \begin{cases} \dfrac{\mathrm{Cov}[X, Y]}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} & (\mathbb{V}[X]\mathbb{V}[Y] \neq 0) \\ 0 & (\mathbb{V}[X]\mathbb{V}[Y] = 0) \end{cases}$$

を X, Y の (Pearsonの) **相関係数** (correlation) という.

とくに, $\rho[X, Y] \neq 0$ のとき X, Y には**相関がある**といい, $\rho[X, Y] = 0$ のときは**無相関**という.

➔ まずは, 共分散の性質を見ていこう.

共分散の基本性質

Prop. (共分散の基本性質)

1次元確率変数 X, Y, Z について, 次が成立する:

(1) [**共分散公式**]　$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

(2) [可換性]　$\mathrm{Cov}[X,Y] = \mathrm{Cov}[Y,X]$.

(3) [線形変換]
実数 $a,b,c,d \in \mathbb{R}$ に対し, $\mathrm{Cov}[aX+b, cY+d] = ac\,\mathrm{Cov}[X,Y]$.

(4) [線形和の共分散]
実数 $a,b \in \mathbb{R}$ に対し, $\mathrm{Cov}[aX+bY, Z] = a\,\mathrm{Cov}[X,Z] + b\,\mathrm{Cov}[Y,Z]$.

(5) [共分散と分散の関係]
$\mathrm{Cov}[X,X] = \mathbb{V}[X]$.

pf. 略 (いずれも定義通り計算すれば示せる).

➔ **分散と似たような性質が成り立っている**!

確率変数の線形和の分散

共分散の現れる重要な公式を紹介しておく:

Prop. (確率変数の線形和の分散)

1次元確率変数 $X_1, \dots, X_n$ と, 実数 $a_1, \dots, a_n \in \mathbb{R}$ に対し,

$$\mathbb{V}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \mathbb{V}[X_i] + 2\sum_{i<j} a_i a_j \operatorname{Cov}[X_i, X_j].$$

pf. $\mu_i = \mathbb{E}[X_i]$ とすると, $\mathbb{E}[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i \mu_i$.

$$\mathbb{V}\left[\sum_{i=1}^{n} a_i X_i\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} a_i \mu_i\right)^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} a_i (X_i - \mu_i)\right)^2\right]$$

$$= \sum_{i=1}^{n} a_i^2 \mathbb{E}[(X_i - \mu_i)^2] + 2\sum_{i<j} a_i a_j \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

$$= \sum_{i=1}^{n} a_i^2 \mathbb{V}[X_i] + 2\sum_{i<j} a_i a_j \operatorname{Cov}[X_i, X_j]. \quad \blacksquare$$

独立な確率変数の和の分散

この命題で $a_1 = \cdots = a_n = 1$ とし, 確率変数が互いに無相関の場合が重要:

Cor. (独立な確率変数の和の分散)

1次元確率変数 $X_1, \dots, X_n$ が**互いに無相関**のとき,

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i].$$

この結果は, 統計的推定の際にも利用される.

➔ 続いて, 相関係数の性質を見ていこう.

相関係数の値域

Prop. (相関係数の値域)

1次元確率変数 X, Y について, $\mathbb{V}[X], \mathbb{V}[Y] < \infty$ ならば, $\rho[X, Y]$ は存在して,

$|\rho[X, Y]| \leq 1$. 等号成立は $Y = aX + \mathbb{E}[Y] - a\mathbb{E}[X], a \neq 0$ となるとき.

pf.

X, Y の平均を μ_X, μ_Y とする. Cauchy–Schwarzの不等式から,
$|\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]|^2 \leq \mathbb{E}[(X - \mu_X)^2]\mathbb{E}[(Y - \mu_Y)^2]$, i.e.,

$$|\mathrm{Cov}[X, Y]|^2 \leq \mathbb{V}[X]\mathbb{V}[Y].$$

等号成立は $Y - \mu_Y = a(X - \mu_X)$ となる実数 $a \in \mathbb{R}$ が存在するとき.
$\mathbb{V}[X], \mathbb{V}[Y] < \infty$ ならば $\mathrm{Cov}[X, Y] < \infty$ である.

$\mathbb{V}[X]\mathbb{V}[Y] = 0$ のときは, $\rho[X, Y] = 0$.
$\mathbb{V}[X]\mathbb{V}[Y] \neq 0$ のときは両辺 $\mathbb{V}[X]\mathbb{V}[Y]$ で割って $|\rho[X, Y]|^2 \leq 1$ である.
また, $\mathbb{V}[X]\mathbb{V}[Y] \neq 0$ より $X - \mu_X \neq 0, Y - \mu_Y \neq 0$. これより, 等号成立は
$Y - \mu_Y = a(X - \mu_X)$ となる実数 $a \neq 0$ が存在するとき. ■

➔ 相関係数の値にはどのような意味があるのだろうか?

相関係数は線形関係の度合いを見るもの

Prop. (相関係数の意味合い)

1次元確率変数 X, Y について, $\mathbb{V}[X] > 0$ とする.

$Y = aX + b\ (a, b \in \mathbb{R}, a \neq 0)$ という線形関係がある $\Leftrightarrow |\rho[X, Y]| = 1$.

とくに, $a > 0 \Leftrightarrow \rho[X, Y] = 1$; $a < 0 \Leftrightarrow \rho[X, Y] = -1$.

pf. $a > 0$ とする. ($a < 0$ でも同様.)

($\Rightarrow$) X の平均を μ_X とする. $Y = aX + b$ のとき, $\mathbb{V}[Y] = a^2\mathbb{V}[X]$. さらに,
$\mathrm{Cov}[X, Y] = \mathbb{E}\big[(X - \mu_X)\big(aX + b - (a\mu_X + b)\big)\big] = a\mathbb{V}[X]$.
よって, $\rho[X, Y] = \dfrac{a\mathbb{V}[X]}{\sqrt{\mathbb{V}[X]}\cdot a\sqrt{\mathbb{V}[X]}} = 1$.

($\Leftarrow$) $\rho[X, Y] = 1$ より, 前ページの等号成立条件から $Y = aX + b$ となる
$a, b \in \mathbb{R}, a \neq 0$ が存在. とくに, $\rho[X, Y] > 0$ より $\mathrm{Cov}[X, Y] = a\mathbb{V}[X] > 0$.
よって, $a > 0$. ■

→ 相関係数の値は, **確率変数間の線形関係の程度を示すもの**!
(相関係数は $X - \mu_X, Y - \mu_Y$ のなす角の余弦を表しているとも見れる.)

独立な確率変数の相関係数

独立な確率変数は, もちろん線形関係もないので無相関になる:

Prop. (独立な確率変数の相関係数は0)

1次元確率変数 X, Y について, X, Y が独立ならば, X, Y は無相関.

pf.

X, Y が独立なので, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. ゆえに,

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0. \quad \blacksquare$$

この逆は必ずしも成り立たない!

たとえば, $X \sim \mathcal{N}(0,1)$ で $Y = X^2$ とする (明らかに独立でない).

$\mathbb{E}[Y] = \mathbb{E}[X^2] = 1$ であり, $\mathbb{E}[X^3] = 0$ である (計算するとわかる).

$\mathrm{Cov}[X, Y] = \mathbb{E}[X^3] - \mathbb{E}[X] = 0$ となるので, X, Y は無相関.

➔ 相関係数で見れるのは, **線形関係のみ**であることに注意.

確率変数ベクトルの平均と分散共分散行列

以下, 確率変数からなる任意の行列 $Z = (Z_{i,j})$ に対しては,
$\mathbb{E}[Z] \coloneqq (\mathbb{E}[Z_{i,j}])$ と, 行列の各成分ごとに期待値 $\mathbb{E}$ を作用させるものとする.

Def. (確率変数ベクトルの平均と分散共分散行列)

p-次元確率変数 $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ と q-次元確率変数 $\boldsymbol{Y} = (Y_1, \dots, Y_q)^\top$ に対して, p-次元ベクトル $\boldsymbol{\mu}_{\boldsymbol{X}} \coloneqq \mathbb{E}[\boldsymbol{X}]$ を $\boldsymbol{X}$ の**平均ベクトル** (mean vector), $p \times q$ 行列

$$\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}] \coloneqq \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})^\top] = \begin{pmatrix} \mathrm{Cov}[X_1, Y_1] & \cdots & \mathrm{Cov}[X_1, Y_q] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_p, Y_1] & \cdots & \mathrm{Cov}[X_p, Y_q] \end{pmatrix}$$

を $\boldsymbol{X}, \boldsymbol{Y}$ の**分散共分散行列** (variance–covariance matrix) という.

とくに, $\mathbb{V}[\boldsymbol{X}] \coloneqq \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}]$ を**分散行列**ともいう.

➔ これについて, 性質を見ておこう.

平均ベクトルと共分散行列の性質

Prop. (平均ベクトルと分散共分散行列の性質)

p-次元確率変数 $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ と q-次元確率変数 $\boldsymbol{Y} = (Y_1, \ldots, Y_q)^\top$, および定数成分の $k \times p$ 行列 A と $l \times q$ 行列 B に対し, 以下が成立:

(1) $\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^\top]^\top = \mathbb{E}[\boldsymbol{Y}\boldsymbol{X}^\top]$.

(2) $\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Y}]^\top$.

(3) $\mathbb{E}[A\boldsymbol{X}] = A\mathbb{E}[\boldsymbol{X}]$.

(4) $\mathrm{Cov}[A\boldsymbol{X}, B\boldsymbol{Y}] = A\,\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}]\,B^\top$.

(5) $\mathbb{V}[\boldsymbol{X}]$ は非負定値対称行列 (i.e., 固有値がすべて非負の対称行列).
とくに, $\boldsymbol{X}$ のどの成分も定数でないならば, $\mathbb{V}[\boldsymbol{X}]$ は正定値対称行列 (i.e., 固有値がすべて正の対称行列).

pf. 略 (練習問題! cf. 清水『統計学への確率論, その先へ』)

ここまでのまとめ

- **モーメント**: 分布の特徴量.
- **分散**, **共分散**に関連した公式: 大体似ているので覚えやすい.
- **確率変数の和の分散公式**: 互いに無相関な場合は分散の和でよい.
- **相関係数**: 二つの確率変数の線形関係を評価するための値.

分布を特徴づける関数

確率母関数

1次元の離散型確率変数の分布の**完全な情報**を持った関数を考えてみよう.

Def. (確率母関数)

1次元の離散型確率変数 X のとりうる値の集合は $\mathbb{Z}_{\geq 0}$ とし, 確率関数を p_X とする. このとき, $|t| \leq 1$ に対して定義された関数

$$P_X(t) \coloneqq \mathbb{E}[t^X] = \sum_{x=0}^{\infty} p_X(x)\, t^x$$

を X の**確率母関数** (probability generating function) という.

$|t| \leq 1$ なので, $|\mathbb{E}[t^X]| \leq \mathbb{E}[|t|^X] \leq \mathbb{E}[1] = 1 < \infty$ となり, **常に存在する**.

➔ 確率母関数 P_X には, **確率関数 p_X のすべての値の情報が含まれている**.
逆に, 確率関数 p_X のすべての値は確率母関数 P_X から復元できる?

確率母関数から確率関数を復元する

Prop. (確率関数の復元)

$\mathbb{Z}_{\geq 0}$ -値の確率変数 X の確率母関数 P_X は, 十分小さな $\varepsilon > 0$ に対して $(-\varepsilon, \varepsilon)$ で無限階微分可能で,

$$p_X(x) = \frac{P_X^{(x)}(0)}{x!}, \qquad x = 0, 1, \dots .$$

pf. (Sketch)

↓これができるかどうかの確認が要る.

$x = 0$ のときは自明. $P_X(t) = \sum_{x=0}^{\infty} p_X(x)\, t^x$ は**項別微分が可能**で,

$\frac{\mathrm{d}}{\mathrm{d}t} P_X(t) = \sum_{x=0}^{\infty} p_X(x) \frac{\mathrm{d}}{\mathrm{d}t}(t^x) = \sum_{x=1}^{\infty} x\, p_X(x)\, t^{x-1}$ $(= \mathbb{E}[X t^{X-1}])$ が成立.

$t = 0$ とすると $P_X^{(1)}(0) = 1 \cdot p_X(1)$. 同様に,

$$\frac{\mathrm{d}^x}{\mathrm{d}t^x} P_X(t) = \sum_{y=0}^{\infty} p_X(y) \frac{\mathrm{d}}{\mathrm{d}t}(t^y) = \sum_{y=x}^{\infty} y(y-1)\cdots(y-x+1) p_X(y)\, t^{y-x}$$
$$(= \mathbb{E}[X(X-1)\cdots(X-x+1) t^{X-x}])$$

となり, $t = 0$ とすると $P_X^{(x)}(0) = x! \cdot p_X(x)$ が成り立つ. ■

➔ これより, **確率母関数と確率関数が1対1に対応する**ことがわかった!
さらに, **階乗モーメント** $P_X^{(x)}(1) = \mathbb{E}[X(X-1)\cdots(X-x+1)]$ **もわかる**.

特性関数

確率母関数で, 離散型確率変数の確率関数や階乗モーメントが計算できた.

➔ **連続型**確率変数でも, 同様のものは考えられる!

Def. (特性関数)

d-次元実確率変数 $\boldsymbol{X}$ の確率分布を $\mathrm{P}_{\boldsymbol{X}}$ とする. $\boldsymbol{t} \in \mathbb{R}^d$ に対して,

$$\phi_X(\boldsymbol{t}) := \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\boldsymbol{X}^\top \boldsymbol{t}}\right] = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\boldsymbol{x}^\top \boldsymbol{t}}\ \mathrm{P}_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x})$$

を X の**特性関数** (characteristic function) という.

$\left|\mathbb{E}\left[\mathrm{e}^{\mathrm{i}\boldsymbol{X}^\top \boldsymbol{t}}\right]\right| \leq \mathbb{E}\left[\left|\mathrm{e}^{\mathrm{i}\boldsymbol{X}^\top \boldsymbol{t}}\right|\right] \leq \mathbb{E}[1] = 1 < \infty$ なので, 特性関数は**常に存在する**.

とくに, $\boldsymbol{X}$ が絶対連続型 (密度関数: $f_{\boldsymbol{X}}$) のとき, $\phi_X(\boldsymbol{t}) = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\boldsymbol{x}^\top \boldsymbol{t}} f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$.

➔ これは $f_{\boldsymbol{X}}(\boldsymbol{x})$ の**Fourier変換**!

➔ Fourier変換した関数は, 逆Fourier変換することで元の関数に戻った.
特性関数 $\phi_{\boldsymbol{X}}$ が分かれば, 確率分布 $\mathrm{P}_{\boldsymbol{X}}$ もわかる?

確率分布と特性関数の1対1対応

(予想通り) 特性関数 ϕ_X が分かれば, 確率分布 P_X も計算できる!

Thm. (Lévy(レヴィ)の反転公式, 1次元版)

1次元確率変数 X の確率分布を P_X, 分布関数を F_X, 特性関数を ϕ_X とする.

分布関数 F_X の連続点 a, b $(a < b)$ に対して,

$$\mathrm{P}_X\big((a,b)\big) = F_X(b) - F_X(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{\mathrm{e}^{-ita} - \mathrm{e}^{-itb}}{\mathrm{i}t} \phi_X(t)\,\mathrm{d}t$$

が成立する.

※ 多次元版のLévyの反転公式も成立する. cf. 西尾『確率論』

➔ **確率分布 P_X と特性関数 ϕ_X が1対1に対応している**ことが分かった!

➔ この公式を使って確率を計算することはほとんどない.
大事なのは, **確率分布と特性関数に1対1対応がある**という事実.

特性関数でモーメント計算

確率母関数で階乗モーメントを計算できたように,
特性関数でもモーメントの計算ができる!

Prop. (特性関数とモーメント)

1次元確率変数 X が, 正整数 $n \in \mathbb{Z}_{>0}$ に対して $\mathbb{E}[|X|^n] < \infty$ であれば,

X の特性関数 ϕ_X は $\mathbb{R}$ 上 n 階連続的微分可能で,

$$\mathbb{E}[X^n] = (-\mathrm{i})^n \phi_X^{(n)}(0) < \infty.$$

pf. (Sketch)

↓これができるかどうかの確認が要る.

期待値と微分の順序交換をすると,

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi_X(t) = \mathbb{E}\left[\frac{\partial^n}{\partial t^n}\mathrm{e}^{\mathrm{i}Xt}\right] = \mathbb{E}\left[(\mathrm{i}X)^n \mathrm{e}^{\mathrm{i}Xt}\right] = \mathrm{i}^n \mathbb{E}\left[X^n \mathrm{e}^{\mathrm{i}Xt}\right]$$

なので, $t = 0$ を代入して両辺に $(-\mathrm{i})^n$ をかけると $(-\mathrm{i})^n \phi_X^{(n)}(0) = \mathbb{E}[X^n]$. ■

➔ 特性関数でもモーメントが計算できた.
ただ, 複素数が出てくるので少し扱いづらい……

モーメント母関数

実は, **複素数を利用しないでも同様に計算できることがある**!

Def. (モーメント母関数)

d-次元実確率変数 $\boldsymbol{X}$ の確率分布を $\mathrm{P}_{\boldsymbol{X}}$ とする. $\boldsymbol{t} \in \mathbb{R}^d$ に対して,

$$m_X(\boldsymbol{t}) \coloneqq \mathbb{E}\left[\mathrm{e}^{\boldsymbol{X}^\top \boldsymbol{t}}\right] = \int_{\mathbb{R}^d} \mathrm{e}^{\boldsymbol{x}^\top \boldsymbol{t}}\ \mathrm{P}_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x})$$

を X の**モーメント母関数** (moment generating function) という.

モーメント母関数は**常に存在するとは限らない**ことに注意.

とくに, $\boldsymbol{X}$ が絶対連続型 (密度関数: $f_{\boldsymbol{X}}$) のとき, $m_X(\boldsymbol{t}) = \int_{\mathbb{R}^d} \mathrm{e}^{\boldsymbol{x}^\top \boldsymbol{t}} f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$.

➔ これは $f_{\boldsymbol{X}}(\boldsymbol{x})$ の**Laplace変換**!
(やはり, **モーメント母関数も確率分布と1対1に対応する**!)

➔ モーメント母関数を使って, モーメントを計算する公式を見よう.

モーメント母関数でモーメント計算

Prop. (モーメント母関数とモーメント)

1次元確率変数 X に対し, ある $\varepsilon > 0$ をとって, $|t| \leq \varepsilon$ のすべての $t \in \mathbb{R}$ で X のモーメント母関数 $m_X(t)$ が存在するとき, m_X は $t = 0$ の近傍で無限階連続的微分可能で,

$$\mathbb{E}[X^n] = m_X^{(n)}(0) < \infty.$$

pf. (Sketch)

↓これができるかどうかの確認が要る.

期待値と微分の順序交換をすると,

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} m_X(t) = \mathbb{E}\left[\frac{\partial^n}{\partial t^n} \mathrm{e}^{Xt}\right] = \mathbb{E}[X^n \mathrm{e}^{Xt}]$$

なので, $t = 0$ を代入して $m_X^{(n)}(0) = \mathbb{E}[X^n]$. ■

➔ **複素数を使うことなく, モーメントの計算ができる**!

➔ ところで, モーメント母関数と特性関数は定義が似ている.
何か関係はあるだろうか?

特性関数をモーメント母関数から求める

実は, 適当な条件のもとで, **モーメント母関数から特性関数を求められる**!

Prop. (モーメント母関数と特性関数)

1次元確率変数 X のモーメント母関数を m_X, 特性関数を ϕ_X とする.

ある $\varepsilon > 0$ に対して $|t| < \varepsilon$ でモーメント母関数 $m_X(t)$ が存在し,

ある関数 $g: \mathbb{R} \to \mathbb{R}$ で $m_X(t) = g(t)$ と表されたとする. この g をそのまま

複素数に拡張した $g(z)$ $(z \in \mathbb{C})$ を考えたとき, g が虚軸を含む領域

$$D_\varepsilon = \{z \in \mathbb{C} \mid |\mathrm{Re}(z)| < \varepsilon\}$$

で正則なら, $\phi_X(t) = g(\mathrm{i}t)$ が成立する, i.e.,

$m_X(t)$ の t を形式的に $\mathrm{i}t$ に変えたものが $\phi_X(t)$ である.

➔ 実用上は, 特性関数を得る際にモーメント母関数を経由すれば問題ない.

例: 正規分布に従う変数の特性関数

$X \sim \mathcal{N}(0,1)$ として, X のモーメント母関数/特性関数などを求めてみよう.

モーメント母関数は,

$$\begin{aligned} m_X(t) = \mathbb{E}[\mathrm{e}^{Xt}] &= \int_{\mathbb{R}} \mathrm{e}^{xt} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} + xt\right) \mathrm{d}x \\ &= \exp\left(\frac{t^2}{2}\right) \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2}\right) \mathrm{d}x \\ &= \exp\left(\frac{t^2}{2}\right). \end{aligned}$$

よって, 特性関数は $\phi_X(t) = \exp\left(\frac{(\mathrm{i}t)^2}{2}\right) = \exp\left(-\frac{t^2}{2}\right)$.

さらに, $m_X'(t) = t\exp\left(\frac{t^2}{2}\right)$, $m_X''(t) = (1+t^2)\exp\left(\frac{t^2}{2}\right)$ より,

$\mathbb{E}[X] = m_X'(0) = 0$, $\mathbb{E}[X^2] = m_X''(0) = 1$, $\mathbb{V}[X] = 1$.

独立性と特性関数

確率変数の独立性を判定するのに, 特性関数を利用することができる.

Thm. (確率変数の独立性の判定3: Kac(カッツ)の定理)

d-次元確率変数列 $\mathcal{X} = \{\boldsymbol{X}_i\}_{i=1}^N$ を考える.

各 $\boldsymbol{X}_i$ $(i = 1, \dots, N)$ の特性関数をそれぞれ ϕ_i とし, Nd-次元確率変数

ベクトル $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_N)^\top$ の特性関数を $\phi_{\boldsymbol{X}}$ とする.

このとき, $\mathcal{X}$ が独立であることと,

任意の $\boldsymbol{t}_1, \dots, \boldsymbol{t}_N \in \mathbb{R}^d$ に対し, $\phi_{\boldsymbol{X}}(\boldsymbol{t}_1, \dots, \boldsymbol{t}_N) = \prod_{i=1}^N \phi_i(\boldsymbol{t}_i)$ であることは同値.

➔ **特性関数の引数の方も完全に分離されている**点に注意.

独立な確率変数列の和

統計では, i.i.d.確率変数列 $X_1, \dots, X_n$ の平均を知るときに, 標本平均

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

の従う分布を考えることで平均を推定する.

独立な確率変数の和の分布は, どうなるのだろうか?

Prop. (独立な確率変数の和の分布)

d-次元確率変数 $\boldsymbol{X}, \boldsymbol{Y}$ は独立で, それぞれ確率分布 $\mathrm{P}_{\boldsymbol{X}}, \mathrm{P}_{\boldsymbol{Y}}$ に従うとする.

このとき, $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{Y}$ の確率分布 $\mathrm{P}_{\boldsymbol{Z}}$ は

$$\mathrm{P}_{\boldsymbol{Z}}(A) = \int_{\mathbb{R}^d} \mathrm{P}_{\boldsymbol{X}}(A - \boldsymbol{y})\, \mathrm{P}_{\boldsymbol{Y}}(\mathrm{d}\boldsymbol{y}), \qquad A \in \mathcal{B}_d$$

をみたす $(A - \boldsymbol{y} \coloneqq \{\boldsymbol{z} - \boldsymbol{y} \mid \boldsymbol{z} \in A\})$. この分布を $\mathrm{P}_{\boldsymbol{Z}} = \mathrm{P}_{\boldsymbol{X}} * \mathrm{P}_{\boldsymbol{Y}}$ と表し,

分布 $\mathrm{P}_{\boldsymbol{X}}, \mathrm{P}_{\boldsymbol{Y}}$ の**畳み込み** (convolution) という. とくに, $\mathrm{P}_{\boldsymbol{X}} * \mathrm{P}_{\boldsymbol{Y}} = \mathrm{P}_{\boldsymbol{Y}} * \mathrm{P}_{\boldsymbol{X}}$.

畳み込みをもう少し具体的に

実用上は, 以下を利用すればよい:

Prop. (独立な確率変数の和の確率関数, 確率密度関数)

d-次元確率変数 $\boldsymbol{X}, \boldsymbol{Y}$ は独立で, $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{Y}$ とする.

(1) $\boldsymbol{X}, \boldsymbol{Y}$ が離散型で, 確率関数をそれぞれ $p_{\boldsymbol{X}}, p_{\boldsymbol{Y}}$ とする. $\boldsymbol{Z}$ の確率関数 $p_{\boldsymbol{Z}}$ は

$$p_{\boldsymbol{Z}}(\boldsymbol{z}) = p_{\boldsymbol{X}} * p_{\boldsymbol{Y}}(\boldsymbol{z}) = \sum_{\boldsymbol{y}} p_{\boldsymbol{X}}(\boldsymbol{z} - \boldsymbol{y})\, p_{\boldsymbol{Y}}(\boldsymbol{y})\,.$$

(2) $\boldsymbol{X}, \boldsymbol{Y}$ が絶対連続型で, 確率密度関数をそれぞれ $f_{\boldsymbol{X}}, f_{\boldsymbol{Y}}$ とする.

$\boldsymbol{Z}$ の確率密度関数 $f_{\boldsymbol{Z}}$ は

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{\boldsymbol{X}} * f_{\boldsymbol{Y}}(\boldsymbol{z}) = \int_{\mathbb{R}^d} f_{\boldsymbol{X}}(\boldsymbol{z} - \boldsymbol{y})\, f_{\boldsymbol{Y}}(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}\,.$$

➔ 導出の仕方を1次元の場合で確認してみよう.

1次元の場合の命題の証明

(2) のみ示す. (離散型もほぼ同様.)

変数変換 $Z = X + Y, T = Y$ を考える.

$X = Z - T, Y = T$ より, 変換のJacobianは $\det \frac{\partial(x,y)}{\partial(z,t)} = 1$.

変数変換の公式より,

$$f_{Z,T}(z,t) = f_{X,Y}(z-t,t) \left| \det \frac{\partial(x,y)}{\partial(z,t)} \right| = f_X(z-t) f_Y(t).$$

t を周辺化すれば,

$$f_Z(z) = \int_{\mathbb{R}} f_{Z,T}(z,t)\, \mathrm{d}t = \int_{\mathbb{R}} f_X(z-t) f_Y(t)\, \mathrm{d}t\,. \qquad \blacksquare$$

➔ 独立な確率変数の和の分布を求めることができた.
ただし, 積分は面倒. もう少し楽な方法はないだろうか?

特性関数などを利用する方法

Prop. (独立な確率変数の和の特性関数/モーメント母関数)

d-次元確率変数 $\boldsymbol{X}, \boldsymbol{Y}$ は独立で, $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{Y}$ とする.

$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ の特性関数を $\phi_{\boldsymbol{X}}, \phi_{\boldsymbol{Y}}, \phi_{\boldsymbol{Z}}$ とすると,

$$\phi_{\boldsymbol{Z}}(\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\boldsymbol{t}) \cdot \phi_{\boldsymbol{Y}}(\boldsymbol{t}), \qquad \boldsymbol{t} \in \mathbb{R}^d$$

が成立する. また, $\boldsymbol{X}, \boldsymbol{Y}$ のモーメント母関数 $m_{\boldsymbol{X}}, m_{\boldsymbol{Y}}$ が存在すれば, $\boldsymbol{Z}$ のモーメント母関数 $m_{\boldsymbol{Z}}$ も存在し,

$$m_{\boldsymbol{Z}}(\boldsymbol{t}) = m_{\boldsymbol{X}}(\boldsymbol{t}) \cdot m_{\boldsymbol{Y}}(\boldsymbol{t}), \qquad \boldsymbol{t} \in \mathbb{R}^d$$

が成立する.

pf.

※ Kacの定理と見比べてみよう. 「$\phi_{\boldsymbol{Z}}(\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\boldsymbol{t}) \cdot \phi_{\boldsymbol{Y}}(\boldsymbol{t})$」が成り立ったとしてもそこから $\boldsymbol{X}, \boldsymbol{Y}$ の独立性は言えないことに注意.

独立性より, $\phi_{\boldsymbol{Z}}(\boldsymbol{t}) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\boldsymbol{Z}^\top \boldsymbol{t}}\right] = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\boldsymbol{X}^\top \boldsymbol{t}}\right]\mathbb{E}\left[\mathrm{e}^{\mathrm{i}\boldsymbol{Y}^\top \boldsymbol{t}}\right] = \phi_{\boldsymbol{X}}(\boldsymbol{t}) \cdot \phi_{\boldsymbol{Y}}(\boldsymbol{t})$.

(モーメント母関数も同様.) ■

→ **特性関数/モーメント母関数が分かっている分布なら**, これで十分.
(特性関数/モーメント母関数が確率分布と1対1に対応するから.)

例: 正規分布に従う独立変数の和の分布

$X, Y \sim \mathcal{N}(0,1)$ (X, Y は独立) として, $Z = X + Y$ の分布を求めてみよう.

畳み込みで計算すると,

$$f_Z(z) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-t)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$$
$$= \frac{1}{\sqrt{2\pi \cdot 2}} \exp\left(-\frac{z^2}{4}\right) \int_{\mathbb{R}} \frac{\sqrt{2}}{\sqrt{2\pi}} \exp\left(-\left(t-\frac{z}{2}\right)^2\right) \mathrm{d}t$$
$$= \frac{1}{\sqrt{2\pi \cdot 2}} \exp\left(-\frac{z^2}{4}\right).$$

よって, $Z \sim \mathcal{N}(0,2)$.

モーメント母関数で計算すると,

$$m_Z(t) = m_X(t) m_Y(t) = \exp(t^2)$$

となり, やはり $\mathcal{N}(0,2)$ のモーメント母関数と一致する (確認せよ).

➔ 独立に正規分布に従う確率変数の和の分布も正規分布になっていることが分かった. このようなことは他の分布でも起こりうる.

分布の再生性

Prop. (分布の再生性)

確率変数の和について, 次の関係が成立する. ただし, 左辺はそれぞれの分布に従う独立な二つの確率変数の和を意味し, 右辺はその和の分布を意味するものとする:

(1) [正規分布] $\mathcal{N}(\mu_X, \sigma_X^2) + \mathcal{N}(\mu_Y, \sigma_Y^2) \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

(2) [二項分布] $\mathrm{Bin}(m, p) + \mathrm{Bin}(n, p) \sim \mathrm{Bin}(m + n, p)$.

(3) [Poisson分布] $\mathrm{Poi}(\lambda_1) + \mathrm{Poi}(\lambda_2) \sim \mathrm{Poi}(\lambda_1 + \lambda_2)$.

(4) [ガンマ分布] $\mathrm{Ga}(\alpha_1, \beta) + \mathrm{Ga}(\alpha_2, \beta) \sim \mathrm{Ga}(\alpha_1 + \alpha_2, \beta)$.

(5) [カイ2乗分布] $\chi_m^2 + \chi_n^2 \sim \chi_{m+n}^2$.

➔ 他にも, 再生性を持つ分布はいくつかある.

ここまでのまとめ

- 分布を特徴づける関数: 分布と1対1に対応する.
  - **確率母関数**: 離散型確率変数に対して定義. いつでも存在.
  - **特性関数**: 一般の確率変数に対して定義. いつでも存在.
  - **モーメント母関数**: 一般の確率変数に対して定義. 存在しないことも.
- 上記の関数を利用すると, 分布の平均, 分散といったモーメントが計算できる.
- **独立な確率変数の和の分布**: 畳み込み.
  - 特性関数/モーメント母関数を使うと, 簡便に計算できることも.
  - **再生性**を持つ分布もある.

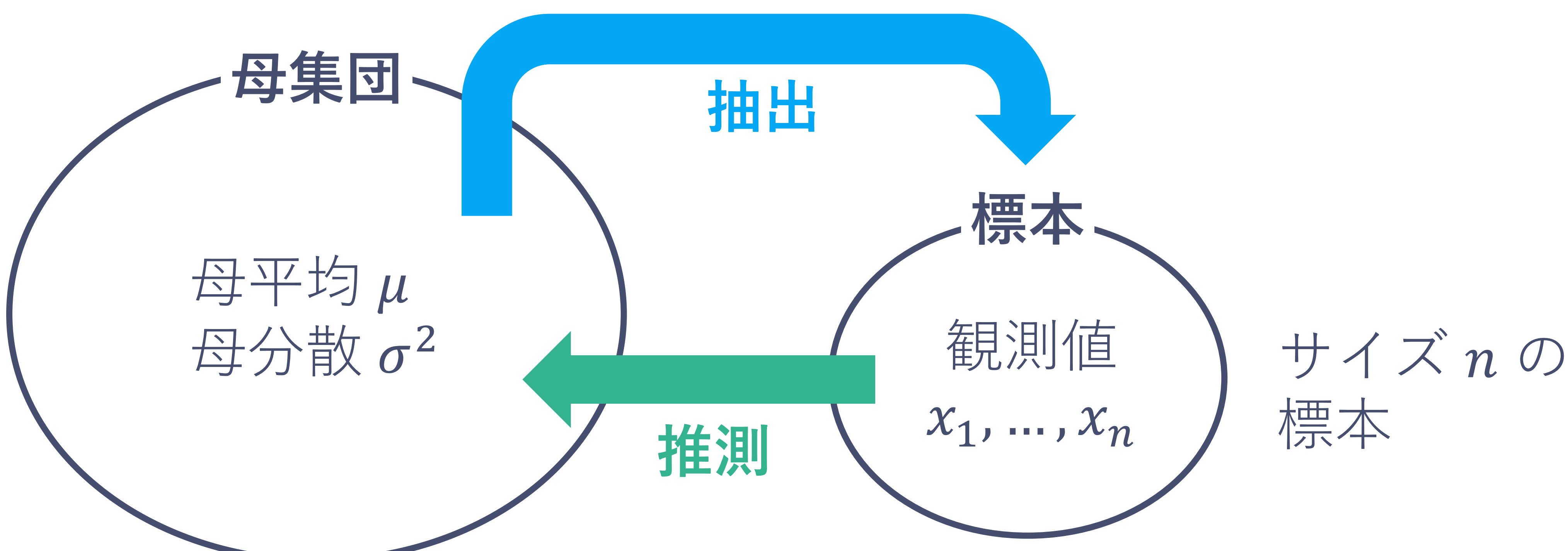
Section 3
統計量と極限定理

統計量と標本分布

統計の言葉遣い

推測統計 (inferential statistics): 「全体の情報を部分の情報から帰納する」

- **母集団** (population): 「情報」を知りたい集団.「全体」
- **母数** (population parameter): 母集団の「情報」
  - 母平均 μ, 母分散 σ^2 など
- **標本** (sample): 母集団から**抽出** (sampling) してくるデータ.「部分」.
  - **サンプルサイズ** (sample size): 抽出されたデータの数



母集団に統計的モデルを仮定しよう

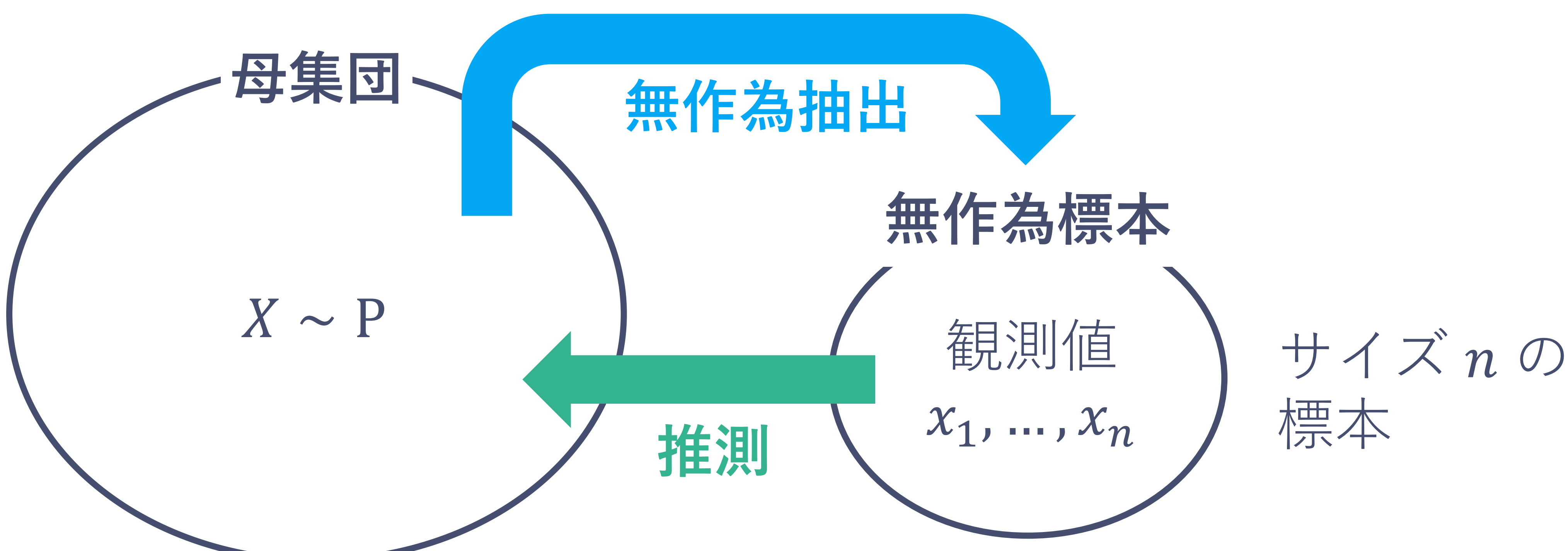
推測統計では通常, 母集団に**統計的モデル**を設定:

- 母集団: **分布 P に従う** (P を**母集団分布**という)
- 標本: 分布 P からの**確率変数** (しばらくは1次元で考える)
- 観測値: **標本 X の実現値**

標本 $X_1, \ldots, X_n$ が分布 P から**無作為抽出** (random sampling) された

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{P}$$

の状況を考えることが多い (得られた標本を**無作為標本**という).



母数を推定するには

無作為標本 $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{P}$ が得られたら, これらを**組み合わせて**母数を推定したい.

統計量 (statistics): $X_1, \dots, X_n$ の実数値関数. **確率変数**.

ex) **標本平均**: 母平均 μ の推定量

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

標本分散: 母分散 σ^2 の推定量

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

標本分布 (sampling distribution): 統計量の従う確率分布.

➔ 統計量の「信頼性」を与えるためには, **標本分布がわかる**必要がある.

標本平均の平均と分散

平均 μ, 分散 σ^2 の母集団分布からの無作為標本 $X_1, \dots, X_n$ を考えよう ($\mathbb{E}[X_i] = \mu, \mathbb{V}[X_i] = \sigma^2$).

標本平均 $\bar{X}$ の平均と分散は (X_i たちが独立であることに注意すると)

$$\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu,$$

$$\mathbb{V}[\bar{X}] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

標本平均の平均は, 母平均と一致する (**不偏性**).

また, $\mathbb{V}[\bar{X}] = \mathbb{E}[(\bar{X} - \mu)^2]$ なので, $\bar{X}$ の μ からの推定誤差は平均 σ^2/n.
➔ n が大きくなるにつれて誤差が小さくなる (cf. 大数の法則).

➔ 標本平均の平均と分散は求まるが, (一般の場合は) **その分布は不明**.

標本分散の平均と不偏分散

一方, 同じ状況で

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] &= \sum_{i=1}^{n}\mathbb{E}[(X_i - \mu + \mu - \bar{X})^2] \\
&= \sum_{i=1}^{n}\mathbb{E}[(X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2] \\
&= \sum_{i=1}^{n}\mathbb{E}[(X_i - \mu)^2] - 2\sum_{i=1}^{n}\mathbb{E}[(X_i - \mu)(\bar{X} - \mu)] + n\mathbb{E}[(\bar{X} - \mu)^2] \\
&= n\sigma^2 - 2n\mathbb{E}\left[(\bar{X} - \mu)\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right] + n\mathbb{E}[(\bar{X} - \mu)^2] \\
&= n\sigma^2 - n\mathbb{E}[(\bar{X} - \mu)^2] = (n-1)\sigma^2.
\end{aligned}
$$

$\mathbb{E}[S^2] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n-1}{n}\sigma^2$. 平均が**母分散と一致しない**.

➔ n で割る代わりに $n-1$ で割った $V^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ なら, $\mathbb{E}[V^2] = \sigma^2$.

V^2 を**不偏分散** (unbiased variance) という. 一般の場合は, **V^2 の分布は不明**.

正規母集団からのサンプリング

一般の母集団分布 P では**標本分布が不明**:

統計量 $T(X_1, \dots, X_n)$ の分布関数 $F_T(t)$ は

$A_t \coloneqq \{\boldsymbol{x} \in \mathbb{R}^n | T(\boldsymbol{x}) \le t\}$ とすると,

$$F_T(t) = \Pr(T \le t) = \int_{A_t} \mathrm{P}(\mathrm{d}x_1) \cdots \mathrm{P}(\mathrm{d}x_n)$$

右辺の積分計算は一般に困難.

ただし, **母集団分布が正規分布であるときはうまく計算できる**!

(このような母集団を**正規母集団**と呼ぶ.)

➔ 以下しばらくは正規母集団からの無作為標本 $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ を扱い, 標本平均, 不偏分散といった統計量の分布を求めてみよう.

標本平均の分布: 母平均に関する推定

母平均についての推定に利用される標本平均 $\bar{X}$ の分布を求めよう.

$\bar{X}$ の特性関数 $\phi_{\bar{X}}(t)$ は

$$\begin{aligned}\phi_{\bar{X}}(t) &= \prod_{i=1}^{n} \phi_{X_i}\left(\frac{t}{n}\right) & \leftarrow X_i \text{ たちの独立性} \\ &= \prod_{i=1}^{n} \exp\left(\frac{\mu i t}{n} - \frac{\sigma^2 t^2}{2n^2}\right) & \leftarrow X_i \sim \mathcal{N}(\mu, \sigma^2) \text{ のとき}, \phi_{X_i}(t) = \exp\left(\mu i t - \frac{\sigma^2 t^2}{2}\right) \\ &= \exp\left(\mu i t - \frac{\sigma^2 t^2}{2n}\right)\end{aligned}$$

これは, $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ **の特性関数と一致**! つまり $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ がわかった.

とくに, 正規化を行うと,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

➔ σ^2 がわかっている状況では, μ に関する推定は**正規分布**を利用すれば良い.

不偏分散の分布: 母分散に関する推定

母分散の推定に利用される不偏分散 V^2 の分布について見ていこう.

Lem. (標本平均と不偏分散の独立性)

正規母集団からの無作為標本 $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ について, 標本平均

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ と不偏分散 $V^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ は独立.

pf. (略. cf. 蓑谷『数理統計ハンドブック』)

cf.) V^2 の中に $\bar{X}$ が現れるのに, $\bar{X}$ と独立になることに注意.

ところで, 不偏分散を導出した際の計算から

$$(n-1)V^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

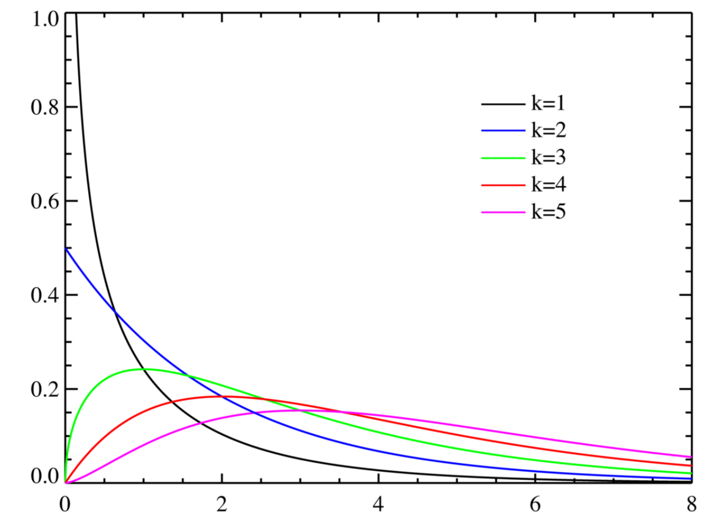
で, 両辺 σ^2 で割ると, $\frac{(n-1)V^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$.

$\frac{X_i-\mu}{\sigma}$ も $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ も $\mathcal{N}(0,1)$ に従う確率変数で, 独立.

➔ $\frac{(n-1)V^2}{\sigma^2}$ の従う分布は何だろう?

カイ2乗分布

Def. (カイ2乗分布)



(Wikipediaから引用)

$k \in \mathbb{Z}_{>0}$ とする. 確率密度関数が $x > 0$ に対し

$$f_X(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)} \left(\frac{1}{2}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1} \mathrm{e}^{-\frac{x}{2}}$$

で与えられる X は**自由度 k のカイ二乗分布** (chi-square distribution) に従うといい, $X \sim \chi_k^2$ と書く. ($\Gamma(x) = \int_0^\infty t^{x-1} \mathrm{e}^{-x}\, \mathrm{d}x$ はガンマ関数.)

自由度 k のカイ二乗分布に従う $X \sim \chi_k^2$ に対し,

$\mathbb{E}[X] = k$, $\mathbb{V}[X] = 2k$, $\phi_X(t) = (1 - 2\mathrm{i}t)^{-\frac{k}{2}}$ (演習問題).

Prop.

確率変数 $Z \sim \mathcal{N}(0, 1)$ に対し, $Z^2 \sim \chi_1^2$ である.

pf. 特性関数を計算すれば良い (演習問題). ■

不偏分散の分布はカイ二乗分布!

カイ二乗分布には再生性があったので, 以下が成立:

Prop.

確率変数 $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ に対し, $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$.

さて, 確率変数 $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ を用いて

$$U \coloneqq \frac{(n-1)V^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2 - \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

と表せた. 第1項 $X \coloneqq \sum_{i=1}^{n} Z_i^2$ は χ_n^2 に, 第2項 $Y \coloneqq \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ は χ_1^2 に従う. $X = U + Y$ と変形すると U と Y が独立なので $\phi_X(t) = \phi_U(t)\phi_Y(t)$.

これより, $\phi_U(t) = \phi_X(t)\big(\phi_Y(t)\big)^{-1} = (1-2\mathrm{i}t)^{-\frac{n-1}{2}}$.

よって, $U = \frac{(n-1)V^2}{\sigma^2} \sim \chi_{n-1}^2$.

➔ σ^2 に関する推定は**自由度 $n-1$ のカイ二乗分布**を利用すれば良い.

標本分散と不偏分散の分散

先程計算しなかった, 標本分散と不偏分散の分散を計算してみよう.

まず, 自由度 $n-1$ のカイ二乗分布の分散から $\mathbb{V}\left[\frac{(n-1)V^2}{\sigma^2}\right] = 2(n-1)$.

分散の性質より, $\mathbb{V}\left[\frac{(n-1)V^2}{\sigma^2}\right] = \frac{(n-1)^2}{\sigma^4}\mathbb{V}[V^2]$. よって, $\mathbb{V}[V^2] = \frac{2\sigma^4}{n-1}$.

また, $S^2 = \frac{n-1}{n}V^2$ なので, $\mathbb{V}\left[\frac{(n-1)V^2}{\sigma^2}\right] = \mathbb{V}\left[\frac{nS^2}{\sigma^2}\right] = \frac{n^2}{\sigma^4}\mathbb{V}[S^2]$.

よって, $\mathbb{V}[S^2] = \frac{2(n-1)\sigma^4}{n^2}$.

とくに $\mathbb{V}[S^2] < \mathbb{V}[V^2]$ であるから,

標本分散で母分散を推定する際のズレは不偏分散よりも小さい.

(ただし, $\mathbb{E}[S^2] = \sigma^2 - \frac{1}{n}\sigma^2$ だったから, 母分散を平均的に過小推定する.)

➔ 母分散 σ^2 の推定量として S^2, V^2 のどちらを利用するかは,
以上のことを踏まえて判断すると良い.

母分散が未知のときの母平均の推定

母分散 σ^2 が既知のときは,

$$Z \coloneqq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

となるのだったが, 一般には σ^2 は未知のことが多い.

➔ 上式で σ の代わりに, その**不偏推定量** $V = \sqrt{V^2}$ を利用した

$$T \coloneqq \frac{\bar{X}-\mu}{V/\sqrt{n}}$$

はどんな分布に従うだろうか?

$U \coloneqq \frac{(n-1)V^2}{\sigma^2}$ とおくと, $U \sim \chi^2_{n-1}$ で, Z と U は独立だった. Z, U を用いると

$T = \frac{Z}{\sqrt{U/(n-1)}}$ と書き直せる.

➔ T の従う分布には特別な名前がついている.

Studentのt-分布

Def. (Studentのt-分布)

Z, U を独立な確率変数とし, $Z \sim \mathcal{N}(0,1), U \sim \chi_m^2$ とする. このとき, 統計量

$$T := \frac{Z}{\sqrt{U/m}}$$

は**自由度 m のStudentのt-分布** (Student's t-distribution) に従うといい,

$T \sim t_m$ と表す.

cf.) "Student" はt-分布の発見者 Gossetのペンネーム

確率密度関数は次のとおり:

Prop. (Studentのt-分布の確率密度関数)

$T \sim t_m$ のとき, T の確率密度関数は $t \in \mathbb{R}$ に対して以下で与えられる:

$$f_T(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \frac{1}{\sqrt{\pi m}} \frac{1}{\left(1 + \frac{t^2}{m}\right)^{\frac{m+1}{2}}}$$

pf. 計算を頑張るとできる (演習問題). ■

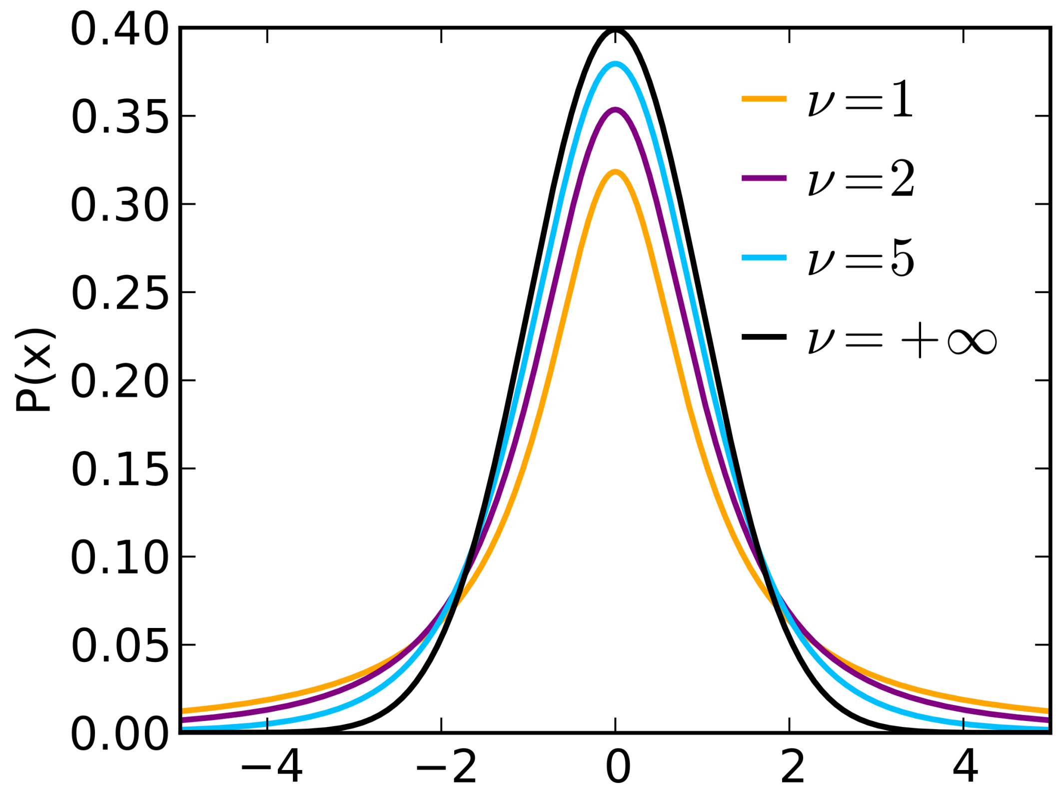
t-統計量はt-分布に従う

先程の $T := \frac{\bar{X}-\mu}{V/\sqrt{n}} = \frac{Z}{\sqrt{U/(n-1)}}$ に戻ろう.

$Z \sim \mathcal{N}(0,1), U \sim \chi^2_{n-1}$ で, Z, U は独立だったので, $T \sim t_{n-1}$.

➔ これを利用することで, **母分散 σ^2 が未知の場合でも**,

母平均 μ についての推定ができる! T を**t-統計量** (t-statistics) という.



自由度 ν のt-分布の確率密度関数 (Wikipediaから引用)

➔ 標準正規分布に似ているが, 微妙に異なった分布になっている.

SnedecorのF-分布

以下のF-分布も, 統計でよく現れる.

Def. (Snedecor(スネデカー)のF-分布)

S, T を独立な確率変数とし, $S \sim \chi_m^2, T \sim \chi_n^2$ とする. このとき,

$$F := \frac{S/m}{T/n}$$

は**自由度 (m, n) のSnedecorのF-分布** (Snedecor's F-distribution) に従うといい, $F \sim F_{m,n}$ と表す.

Prop. (SnedecorのF-分布の確率密度関数)

$F \sim F_{m,n}$ のとき, F の確率密度関数は $t > 0$ に対して以下で与えられる:

$$f_F(t) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}} t^{\frac{m}{2}-1}}{\left(1 + \left(\frac{m}{n}\right)t\right)^{\frac{m+n}{2}}}$$

pf. 計算を頑張るとできる (演習問題). ■

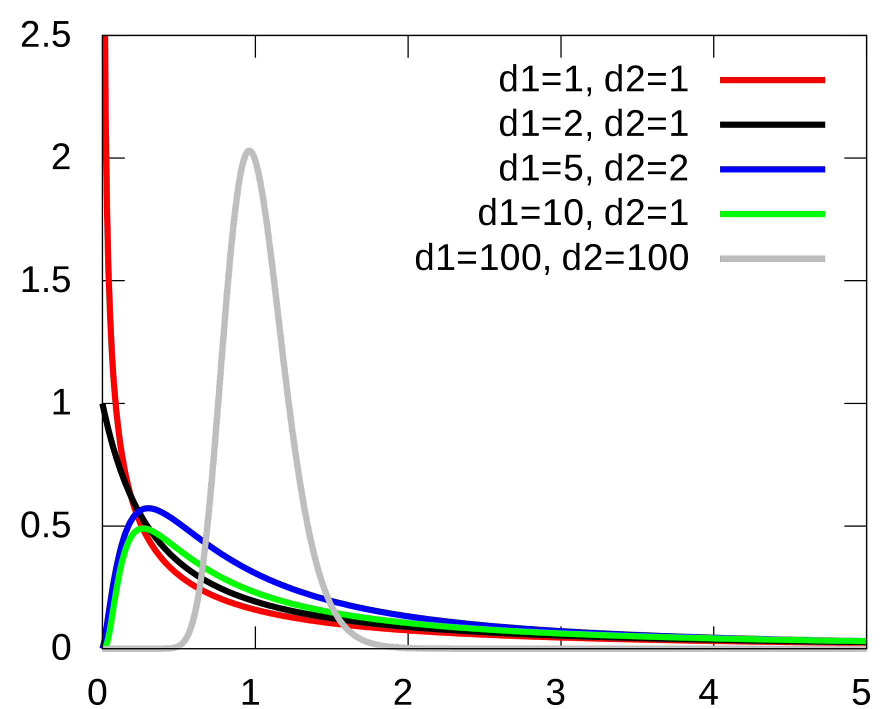
F-統計量はF-分布に従う

$X_1, \ldots, X_{n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2), Y_1, \ldots, Y_{n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ としよう. X_i と Y_j は独立.

X_i, Y_j たちの不偏分散を V_1^2, V_2^2 とし, $S \coloneqq \frac{(n_1-1)V_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$,

$T \coloneqq \frac{(n_2-1)V_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$ とすると, S, T は独立なので,

$F \coloneqq \frac{S/(n_1-1)}{T/(n_2-1)} = \frac{V_1^2/\sigma_1^2}{V_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$ となる. F を**F-統計量** (F-statistics) という.



自由度 (d_1, d_2) のF-分布の確率密度関数 (Wikipediaから引用)

➔ 等分散 $\sigma_1^2 = \sigma_2^2$ かどうかの検定などに利用される分布.

ここまでのまとめ

- **推測統計**: 母集団に統計的モデルを設定し, 母数を推定する.
- **統計量**: 母数を含まない標本の実数値関数
- **標本分布**: 統計量の従う分布
- **標本平均の分布**: $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. $Z \coloneqq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$
- **不偏分散の分布**: $U \coloneqq \frac{(n-1)V^2}{\sigma^2} \sim \chi^2_{n-1}$.
- 母分散の代わりに不偏分散を利用すると $T \coloneqq \frac{\bar{X}-\mu}{V/\sqrt{n}} \sim t_{n-1}$
- 分散の同等性の検定にはSnedecorのF-分布が利用される.
  $F \coloneqq \frac{V_1^2/\sigma_1^2}{V_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$.

確率変数列と確率分布の収束

一般の分布からのサンプリング

母集団分布は, 一般には正規分布とは限らない

➔ **正確な標本分布を導くのが困難**

漸近理論 (asymptotic theory) / **大標本理論** (large sample theory):

サンプルサイズ n が十分大きいときの**標本分布の近似**を考えたい

近似的には母集団分布を**正規分布であるとみなせることが多い**!

(cf. 中心極限定理)

➔ $n \to \infty$ での確率変数や確率分布の振る舞いを議論できるよう,
収束性に関する概念について見ていこう.

確率変数列の「収束」とは?

確率変数列 $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ とは, 「$(\Omega, \mathcal{F})$ 上の**可測関数**の列」のことだった.

この「収束」の意味は, 色々考えることができる.

確率変数列の収束として考えやすいのは, $\omega \in \Omega$ **を固定するごとの収束**
(関数列の各点収束).

➔ 確率論では, 確率が0になるような ω の集合 (零集合) を除いて議論するほうが便利 (cf. “almost surely”).

$\boldsymbol{X}$ を確率変数として, 「$\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}$ となる確率」が1であれば,

確率0の集合を除いて $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ が $\boldsymbol{X}$ に収束すると言えそう.

➔ そもそも, 「$\lim_{n\to\infty} \boldsymbol{X}_n$」は確率変数になってくれるのだろうか?

上・下極限は確率変数

Def. (上極限・下極限)

$(\Omega, \mathcal{F})$ 上の1次元確率変数列 $\{X_n\}$ の**上極限** (limit superior),

下極限 (limit inferior) をそれぞれ

$$\limsup_{n\to\infty} X_n(\omega) \coloneqq \lim_{n\to\infty}\left(\sup_{k\geq n} X_k(\omega)\right), \liminf_{n\to\infty} X_n(\omega) \coloneqq \lim_{n\to\infty}\left(\inf_{k\geq n} X_k(\omega)\right)$$

で定める.

↑ ω を固定した数列 $\{X_n(\omega)\}$ を考えたときの,
n 以降の上限・下限の n に関する極限.

Prop.

$(\Omega, \mathcal{F})$ 上の1次元確率変数列 $\{X_n\}$ に対し,

$$\sup_{n\in\mathbb{Z}_{>0}} X_n, \inf_{n\in\mathbb{Z}_{>0}} X_n, \limsup_{n\to\infty} X_n, \liminf_{n\to\infty} X_n$$

はすべて確率変数になる (i.e., $\mathcal{F}$-可測である).

pf. 略.

➔ 上極限と下極限を使って, 確率変数列の極限を定義しよう.

確率変数列の極限

Def. (極限)

$(\Omega, \mathcal{F}, \Pr)$ 上の d 次元確率変数列 $\{\boldsymbol{X}_n\}$ を考える. $\boldsymbol{X}_n$ の各成分 $X_n^{(k)}$ に対し,

$$X_*^{(k)} \coloneqq \liminf_{n\to\infty} X_n^{(k)} = \limsup_{n\to\infty} X_n^{(k)}, \mathrm{a.s.} \quad (k = 1, \dots, d)$$

となるとき, 写像 $\boldsymbol{X}_* = \left(X_*^{(1)}, \dots, X_*^{(d)}\right) : \Omega \to \mathbb{R}^d$ を $\lim_{n\to\infty} \boldsymbol{X}_n$ と書く. $\lim_{n\to\infty} \boldsymbol{X}_n$ は $(\Omega, \mathcal{F}, \Pr)$ 上の d 次元確率変数で, $\{\boldsymbol{X}_n\}$ の**極限** (limit) という.

➔ こうして $\lim_{n\to\infty} \boldsymbol{X}_n$ が確率変数であることが確かめられた.
収束性を一つ定義しよう.

概収束

Def. (概収束)

$(\Omega, \mathcal{F}, \Pr)$ 上の d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元確率変数 $\boldsymbol{X}$ に対して,

$$\Pr\left(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}\right) = 1$$

となるとき, $\boldsymbol{X}_n$ は $\boldsymbol{X}$ に**概収束** (almost-sure convergence) するといい,

$$\boldsymbol{X}_n \xrightarrow{\text{a.s.}} \boldsymbol{X} \ (n \to \infty)$$

と書く.

$\boldsymbol{X}_n \xrightarrow{\text{a.s.}} \boldsymbol{X} \ (n \to \infty)$ とは $\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}$, a.s. ということ, i.e.,

ほとんどすべての $\omega \in \Omega$ を固定してできる**点列 $\{\boldsymbol{X}_n(\omega)\}_{n=1}^{\infty}$ の収束**

を意味する.

➔ さらに, 概収束よりも「弱い」収束も定義しよう.

cf. 概収束のための十分条件

$\boldsymbol{X}_n \xrightarrow{\text{a.s.}} \boldsymbol{X}\ (n \to \infty)$ のための十分条件が知られている:

Prop. (概収束の十分条件)

$(\Omega, \mathcal{F}, \Pr)$ 上の d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元確率変数 $\boldsymbol{X}$ を考える.

任意の $\varepsilon > 0$ に対し,

$$\sum_{n=1}^{\infty} \Pr(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) < \infty$$

ならば, $\boldsymbol{X}_n \xrightarrow{\text{a.s.}} \boldsymbol{X}\ (n \to \infty)$.

この証明には, **Borel-Cantelliの補題**が利用される.

この補題自体も, 概収束を示すのに有用な補題.

確率収束

Def. (確率収束)

$(\Omega, \mathcal{F}, \Pr)$ 上の d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元確率変数 $\boldsymbol{X}$ について,

任意の $\varepsilon > 0$ に対して

$$\lim_{n\to\infty} \Pr(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) = 0$$

となるとき, $\boldsymbol{X}_n$ は $\boldsymbol{X}$ に**確率収束** (convergence in probability) するといい,

$$\boldsymbol{X}_n \xrightarrow{\mathrm{p}} \boldsymbol{X} \ (n \to \infty)$$

と書く.

「どんなに (小さい) $\varepsilon > 0$ をとったとしても, $\boldsymbol{X}_n$ と $\boldsymbol{X}$ の差が ε 以上になる (i.e., 外れ値をとる) **確率**は0に近づく.」ということ.

確率の収束であって, **可測関数 X_n 自体の収束を述べてはいない**.

➔ 概収束と確率収束の違いをもう少し詳しく見てみよう.

概収束と確率収束の違い

$\Omega = [0,1), \mathcal{F} = \{I \cap \Omega \mid I \in \mathcal{B}_1\}, \Pr = U[0,1)$ (一様分布) として,
確率空間 $(\Omega, \mathcal{F}, \Pr)$ を考える. Ω 上の確率変数を次で定義しよう:

$$X_{n,k} \coloneqq 1_{A_{n,k}}(\omega), \left(A_{n,k} \coloneqq [(k-1)2^{-n}, k2^{-n}); n = 1,2,\ldots;\ k = 1,\ldots,2^n\right)$$

さらに, $Y_1 = X_{1,1}, Y_2 = X_{1,2}, Y_3 = X_{2,1}, \ldots$ などと定めて $\{Y_m\}_{m=1}^{\infty}$ を作る.

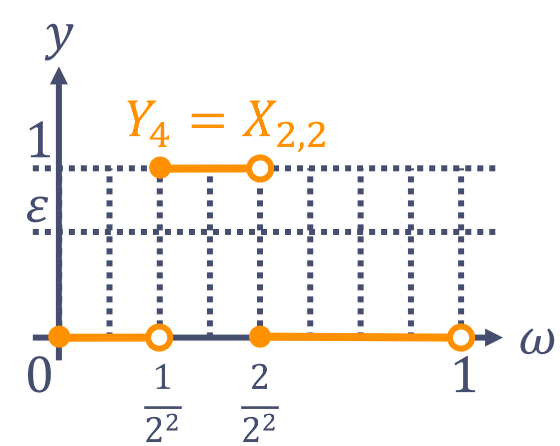
$\omega \in \Omega$ を任意に固定すると, ω を含む $A_{n,k}$ は無限個あるので

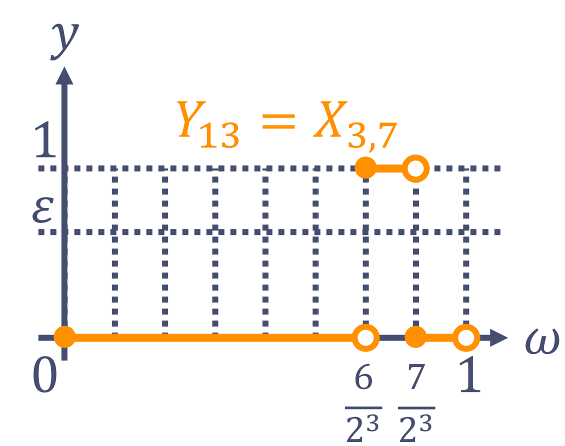
$$\liminf_{m\to\infty} Y_m = 0 \neq 1 = \limsup_{m\to\infty} Y_m \,.$$

となり, Y_m **は概収束しない**. しかし, 任意の $\varepsilon > 0$ に対して,

$\Pr(|Y_m - 0| > \varepsilon) = 2^{-n} 1_{\{0<\varepsilon<1\}} \to 0 \ (m \to \infty)$ (ただし $Y_m = X_{n,k}$)

となるので Y_m **は0に確率収束する**.





➔ **概収束より確率収束のほうが弱い収束概念**.

cf. 確率収束のための必要十分条件

Prop. (確率収束のための必要十分条件)

$x \wedge y \coloneqq \min\{x, y\}$ とする. d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元確率変数 $\boldsymbol{X}$ に対し, 次が成立: $\boldsymbol{X}_n \xrightarrow{\mathrm{p}} \boldsymbol{X} \Leftrightarrow \mathbb{E}[\|\boldsymbol{X}_n - \boldsymbol{X}\| \wedge 1] \xrightarrow{n\to\infty} 0.$

pf. $Y_n \coloneqq \|\boldsymbol{X}_n - \boldsymbol{X}\| \wedge 1$ とする. $Y_n \leq 1$ である. $\varepsilon > 0$ をとる.

(⇒) $\boldsymbol{X}_n \xrightarrow{\mathrm{p}} \boldsymbol{X}$ より, 十分大きな n をとると, $\Pr\left(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \frac{\varepsilon}{2}\right) < \frac{\varepsilon}{2}$ とできる.
$1 = 1_{\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon/2\}} + 1_{\{\|\boldsymbol{X}_n - \boldsymbol{X}\| \leq \varepsilon/2\}}$ より,

$$\begin{aligned}
\mathbb{E}[Y_n] &= \mathbb{E}\left[Y_n 1_{\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon/2\}}\right] + \mathbb{E}\left[Y_n 1_{\{\|\boldsymbol{X}_n - \boldsymbol{X}\| \leq \varepsilon/2\}}\right] \\
&\leq \mathbb{E}\left[1 \cdot 1_{\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon/2\}}\right] + \mathbb{E}\left[\frac{\varepsilon}{2} 1_{\{\|\boldsymbol{X}_n - \boldsymbol{X}\| \leq \varepsilon/2\}}\right] \\
&= \Pr\left(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2} \Pr\left(\|\boldsymbol{X}_n - \boldsymbol{X}\| \leq \frac{\varepsilon}{2}\right) < \varepsilon.
\end{aligned}$$

(⇐)
$$\begin{aligned}
&\Pr(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) = \mathbb{E}\left[1_{\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon\}}\right] = (\varepsilon \wedge 1)^{-1} \mathbb{E}\left[(\varepsilon \wedge 1) 1_{\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon\}}\right] \\
&\leq (\varepsilon \wedge 1)^{-1} \mathbb{E}\left[(\|\boldsymbol{X}_n - \boldsymbol{X}\| \wedge 1) 1_{\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon\}}\right] \leq (\varepsilon \wedge 1)^{-1} \mathbb{E}[Y_n] \xrightarrow{n\to\infty} 0.
\end{aligned}$$

■

p次平均収束

Def. (p次平均収束)

$(\Omega, \mathcal{F}, \Pr)$ 上の d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元確率変数 $\boldsymbol{X}$ について,

各 $n \in \mathbb{Z}_{>0}$ に対して $\mathbb{E}[\|\boldsymbol{X}_n\|^p] < \infty$ とする ($p \geq 1$).

$$\lim_{n\to\infty} \mathbb{E}[\|\boldsymbol{X}_n - \boldsymbol{X}\|^p] = 0$$

となるとき, $\boldsymbol{X}_n$ は $\boldsymbol{X}$ に p **次平均収束** (convergence in L^p) するといい,

$$\boldsymbol{X}_n \xrightarrow{L^p} \boldsymbol{X} \ (n \to \infty)$$

と書く.

$\mathbb{E}[\|\boldsymbol{X}_n - \boldsymbol{X}\|^p]$ は $\boldsymbol{X}_n$ と $\boldsymbol{X}$ の「距離」と見ることができる.

(cf. L^p 空間のノルム)

➔ p 次平均収束は, **$\boldsymbol{X}_n$ と $\boldsymbol{X}$ の距離の収束**を意味している.
最後にもう一つ収束の概念を定義する.

分布収束

Def. (分布収束)

$(\Omega_n, \mathcal{F}_n, \mathrm{Pr}_n)$ 上の d 次元確率変数 $\boldsymbol{X}_n$ の列 $\{\boldsymbol{X}_n\}$ と $(\Omega, \mathcal{F}, \mathrm{Pr})$ 上の d 次元確率変数 $\boldsymbol{X}$ について, $\mathbb{R}^d$ 上の任意の有界連続関数 $f: \mathbb{R}^d \to \mathbb{R}$ に対し

$$\lim_{n\to\infty} \mathbb{E}[f(\boldsymbol{X}_n)] = \mathbb{E}[f(\boldsymbol{X})]$$

のとき, $\boldsymbol{X}_n$ は $\boldsymbol{X}$ に**分布収束** (convergence in distribution) するといい,

$$\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X} \ (n \to \infty)$$

と書く.

$\boldsymbol{X}_n$ と $\boldsymbol{X}$ の分布をそれぞれ $\mathrm{P}_n = \mathrm{Pr} \circ \boldsymbol{X}_n^{-1}, \mathrm{P} = \mathrm{Pr} \circ \boldsymbol{X}^{-1}$ と書くと, $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X} \ (n \to \infty)$ とは $\lim_{n\to\infty} \int_{\mathbb{R}^d} f(\boldsymbol{x})\, \mathrm{P}_n(\mathrm{d}\boldsymbol{x}) = \int_{\mathbb{R}^d} f(\boldsymbol{x})\, \mathrm{P}(\mathrm{d}\boldsymbol{x})$ ということ.

文字通り, **確率分布 P_n が P へと収束する**ことを表す. (cf. 分布の弱収束)

$\boldsymbol{X}_n$ の定義される確率空間が n ごとに異なっていても良いことに注意.

➔ P_n が P に分布収束するなら, **n が十分大きいときには P_n は P の近似**.
分布収束は, 数理統計学では重要な収束概念. 同値な条件がいくつかある.

ポルトマントーの補題

Thm. (ポルトマントーの補題: the Portmanteau lemma)

d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元確率変数 $\boldsymbol{X}$ に対し, $\boldsymbol{X}_n$ と $\boldsymbol{X}$ の分布を

それぞれ P_n, P, 分布関数を F_n, F と書く. $n \to \infty$ のとき, 次の (1)-(7) は同値:

(1) $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X}$, i.e.,任意の有界連続関数 f に対し, $\mathbb{E}[f(\boldsymbol{X}_n)] \to \mathbb{E}[f(\boldsymbol{X})]$.

(2) 任意の有界**Lipschitz連続**(リプシッツ)関数 f に対し, $\mathbb{E}[f(\boldsymbol{X}_n)] \to \mathbb{E}[f(\boldsymbol{X})]$.

(3) 任意の**非負値**連続関数 f に対し, $\liminf_{n\to\infty} \mathbb{E}[f(\boldsymbol{X}_n)] \geq \mathbb{E}[f(\boldsymbol{X})]$.

(4) 任意の開集合 $G \subseteq \mathbb{R}^d$ に対し, $\liminf_{n\to\infty} \mathrm{P}_n(G) \geq \mathrm{P}(G)$.

(5) 任意の閉集合 $K \subseteq \mathbb{R}^d$ に対し, $\limsup_{n\to\infty} \mathrm{P}_n(K) \leq \mathrm{P}(K)$.

(6) $F(\boldsymbol{x})$ の任意の**連続点** $\boldsymbol{x} \in \mathbb{R}^d$ において $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ (各点収束).

(7) 境界 ∂D が $\mathrm{P}(\partial D) = 0$ となる任意の $D \in \mathcal{B}_d$ で $\mathrm{P}_n(D) \to \mathrm{P}(D)$.

➜ 分布収束の証明をする際は, どれを示しても良い. (6) がわかりやすい.

Lévyの連続性定理

さらに, 特性関数の収束について次の結果が知られている:

Thm. (Lévyの連続性定理)

↓特性関数の代わりにモーメント母関数としても成立.

d 次元確率変数列 $\{\boldsymbol{X}_n\}$ に対し, $\boldsymbol{X}_n$ の特性関数を ϕ_n と書く.

(1) d 次元確率変数 $\boldsymbol{X}$ に対し, $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X}$ となるならば,

任意の $\boldsymbol{t} \in \mathbb{R}^d$ に対して $\lim_{n\to\infty} \phi_n(\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\boldsymbol{t})$ が成り立つ.

(2) 関数 ϕ: $\mathbb{R}^d \to \mathbb{C}$ に対し, 任意の $\boldsymbol{t} \in \mathbb{R}^d$ について $\lim_{n\to\infty} \phi_n(\boldsymbol{t}) = \phi(\boldsymbol{t})$ で,

かつ **ϕ が $\boldsymbol{t} = \boldsymbol{0}$ で連続である**ならば, ϕ はある確率変数 $\boldsymbol{X}$ の特性関数

で $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X}$ が成り立つ.

※ (2) で ϕ が $\boldsymbol{t} = \boldsymbol{0}$ で連続の条件は, ϕ が特性関数になるために外せない.
(cf. Bochnerの定理: 特性関数の特徴づけ)

➔ 特性関数が収束することがわかれば分布収束が示せるし, 逆も成立する.
分布収束を示すのに便利な定理!

Glivenkoの定理

Lévyの連続性定理の系として, 次のGlivenko(グリベンコ)の定理が成立する:

Cor. (Glivenkoの定理)

d 次元確率変数 $\{\boldsymbol{X}_n\}, \boldsymbol{X}$ に対し, $\boldsymbol{X}_n, \boldsymbol{X}$ の特性関数をそれぞれ ϕ_n, ϕ と書く.

任意の $\boldsymbol{t} \in \mathbb{R}^d$ に対して $\lim_{n\to\infty} \phi_n(\boldsymbol{t}) = \phi(\boldsymbol{t})$ ならば $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X}$.

(この逆の成立はLévyの連続性定理の(1)で述べられている.)

Lévyの連続性定理の(2)で, ϕ がはじめから確率変数 $\boldsymbol{X}$ の特性関数

であることがわかっている状況に対応する.

➔ **分布収束先の分布がわかっていて特性関数が計算できる場合**は,
連続性定理をこの形で利用すればよい.

Poissonの少数法則

次の事実を, Glivenkoの定理を使って示そう:

Thm. (Poissonの少数法則)

1次元確率変数列 $\{X_n\}$ は $X_n \sim \mathrm{Bin}(n, p)$ とする. また, $X \sim \mathrm{Poi}(\lambda)$ とする.

$np = \lambda$ のもとで, $n \to \infty, p \to 0$ とすると, X_n は X に分布収束する.

pf.

X_n の特性関数は, $\phi_n(t) = \left(p\mathrm{e}^{\mathrm{i}t} + 1 - p\right)^n$. $p = \lambda/n$ と

$$a_n \to a \Longrightarrow \lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = \mathrm{e}^a$$

であることを利用すると,

$$\phi_n(t) = \left(1 + \frac{\left(\mathrm{e}^{\mathrm{i}t} - 1\right)\lambda}{n}\right)^n \xrightarrow{n\to\infty} \exp\left(\left(\mathrm{e}^{\mathrm{i}t} - 1\right)\lambda\right) = \phi_X(t)$$

となり, Glivenkoの定理より $X_n \xrightarrow{\mathrm{d}} X$. ■

Cramér–Woldの方法

多次元確率変数の同時分布収束を示すのは一般に難しいが, 1次元の分布収束に帰着させる**Cramér–Woldの方法**(クラメール ウォルド) (Cramér–Wold device) がLévyの連続性定理から示せる.

Cor. (Cramér–Woldの方法)

d 次元確率変数 $\{\boldsymbol{X}_n\}, \boldsymbol{X}$ に対し, 次は同値:

(1) $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X}$ $(n \to \infty)$.

(2) 任意の $\boldsymbol{t} \in \mathbb{R}^d$ に対し, $\boldsymbol{t}^\top \boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{t}^\top \boldsymbol{X}$ $(n \to \infty)$.

pf. $\boldsymbol{X}_n, \boldsymbol{X}$ の特性関数をそれぞれ ϕ_n, ϕ とする. $s \in \mathbb{R}, \boldsymbol{t} \in \mathbb{R}^d$ として,

(1) $\Leftrightarrow \lim_{n\to\infty} \phi_n(s\boldsymbol{t}) = \phi(s\boldsymbol{t})$ (∵ Lévyの連続性定理)

$\Leftrightarrow \lim_{n\to\infty} \mathbb{E}\left[\exp\left(\mathrm{i}s(\boldsymbol{t}^\top \boldsymbol{X}_n)\right)\right] = \mathbb{E}\left[\exp\left(\mathrm{i}s(\boldsymbol{t}^\top \boldsymbol{X})\right)\right] \Leftrightarrow$ (2) (∵ Lévyの連続性定理). ■

➔ 多次元版の中心極限定理の証明などに利用される.

cf. Schefféの有用収束定理

シェッフェ
Schefféの有用収束定理より, **密度関数の収束からも分布収束がいえる**.

Lem. (Schefféの有用収束定理)

d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元確率変数 $\boldsymbol{X}$ に対し, $\boldsymbol{X}_n, \boldsymbol{X}$ が絶対連続型で

それぞれ分布関数 $F_{\boldsymbol{X}_n}, F_{\boldsymbol{X}}$, 分布 P_n, P と確率密度関数 p_n, p をもつとする.

このとき, $p_n \to p$, a.s. ならば, $\boldsymbol{X}_n, \boldsymbol{X}$ 間の全変動距離は

$$d_{\mathrm{TV}}(\boldsymbol{X}_n, \boldsymbol{X}) \coloneqq \sup_{A \in \mathcal{B}_d} |\mathrm{P}_n(\boldsymbol{X}_n \in A) - \mathrm{P}(\boldsymbol{X} \in A)| \to 0. \ (n \to \infty)$$

とくに, $\left|F_{\boldsymbol{X}_n}(\boldsymbol{x}) - F_{\boldsymbol{X}}(\boldsymbol{x})\right| \leq d_{\mathrm{TV}}(\boldsymbol{X}_n, \boldsymbol{X})$ なので, $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X}$.

pf. 略 (cf. Gut, "Probability – A Graduate Course" など.)

逆は成り立たない, i.e., $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X}$ であっても $p_n \to p$, a.s. とは限らない.

cf.) $d_{\mathrm{TV}}(\boldsymbol{X}_n, \boldsymbol{X}) \to 0$ となる収束は**全変動収束** (total variation convergence) という収束.

各収束の強弱

各収束の強弱について, 次が成立する:

Prop. (各収束の強弱)

d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元確率変数 $\boldsymbol{X}$ に対し,

(1) $\boldsymbol{X}_n \xrightarrow{\text{a.s.}} \boldsymbol{X}\ (n \to \infty)$ ならば $\boldsymbol{X}_n \xrightarrow{\text{p}} \boldsymbol{X}\ (n \to \infty)$.

(2) $\boldsymbol{X}_n \xrightarrow{L^p} \boldsymbol{X}\ (n \to \infty)$ ならば $\boldsymbol{X}_n \xrightarrow{\text{p}} \boldsymbol{X}\ (n \to \infty)$.

(3) $\boldsymbol{X}_n \xrightarrow{\text{p}} \boldsymbol{X}\ (n \to \infty)$ ならば $\boldsymbol{X}_n \xrightarrow{\text{d}} \boldsymbol{X}\ (n \to \infty)$.

いずれも**逆は必ずしも成り立たない**. まとめると次のとおり:

概収束 $\searrow$

確率収束 $\Longrightarrow$ **分布収束**

$\boldsymbol{p}$ **次平均収束** $\nearrow$

cf. 確率収束するが p 次平均収束しない例

$\Omega = [0,1), \mathcal{F} = \{I \cap \Omega \mid I \in \mathcal{B}_1\}, \Pr = U[0,1)$ (一様分布) として,
確率空間 $(\Omega, \mathcal{F}, \Pr)$ を考える. Ω 上の確率変数を次で定義しよう:

$$\text{定数 } q > 0 \text{ に対し, } X_n(\omega) \coloneqq n 1_{(0, n^{-q})}(\omega).$$

任意の $\varepsilon > 0$ に対し, $\Pr(|X_n| > \varepsilon) = n^{-q} \to 0 \ (n \to \infty)$ となり $X_n \xrightarrow{\mathrm{p}} 0$.

ところが, $p > q$ に対し, $\mathbb{E}[|X_n|^p] = n^{p-q} \to \infty \ (n \to \infty)$ となり,

X_n は0に p 次平均収束しない.

cf. p 次平均収束するが概収束しない例

確率変数を

$$X_n := \begin{cases} 0 & \text{確率 } \dfrac{1}{n} \\ 1 & \text{確率 } 1 - \dfrac{1}{n} \end{cases}$$

と定義する.

任意の $p \geq 1$ に対し,

$$\begin{aligned} \mathbb{E}[|X - 1|^p] &= \frac{1}{n}|0 - 1|^p + \left(1 - \frac{1}{n}\right)|1 - 1|^p \\ &= \frac{1}{n} \to 0 \ (n \to \infty) \end{aligned}$$

なので, $X_n \xrightarrow{L^p} 1$ である. ところが, X_n は1に概収束しないことが

Borel–Cantelliの補題から示せる.

cf. 分布収束するが確率収束しない例

確率変数を

$$X_{2k+1}(\omega) := \begin{cases} 1 & \omega \in \left[0, \frac{1}{2}\right] \\ 0 & \omega \in \left(\frac{1}{2}, 1\right] \end{cases}, X_{2k}(\omega) := \begin{cases} 0 & \omega \in \left[0, \frac{1}{2}\right] \\ 1 & \omega \in \left(\frac{1}{2}, 1\right] \end{cases}$$

と定義する.

任意の n に対し, $\Pr(X_n = 0) = \Pr(X_n = 1) = \frac{1}{2}$ なので, X_n の分布は

すべて同じ, i.e., $X_n \xrightarrow{\mathrm{d}} X_1$.

ところが, $n = 2k$ のとき, $\Pr(|X_{2k} - X_1| \geq 1) = 1$ なので

X_n は X_1 に確率収束しない.

cf. 収束性についてのその他の関係

一般には,「確率収束 ⟹ 分布収束」の逆は成立しないが, 次のように
特殊な場合には逆が成立する:

Prop.

d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元**定数ベクトル** $\boldsymbol{c}$ に対し,

$\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{c}\ (n \to \infty)$ ならば $\boldsymbol{X}_n \xrightarrow{\mathrm{p}} \boldsymbol{c}\ (n \to \infty)$.

また, 概収束と p 次平均収束についても次の関係がある:

Prop.

$p \geq 1$ とする. d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と d 次元確率変数 $\boldsymbol{X}$ は p 次可積分

とし, $\boldsymbol{X}_n \xrightarrow{\mathrm{a.s.}} \boldsymbol{X}$ とする. このとき,

$$\boldsymbol{X}_n \xrightarrow{L^p} \boldsymbol{X} \Leftrightarrow \mathbb{E}[\|\boldsymbol{X}_n\|^p] \to \mathbb{E}[\|\boldsymbol{X}\|^p]$$

さらに, 一様可積分の条件のもとでは, 概収束 ⟹ 1次平均収束もいえる.
(Vitaliの収束定理)

Slutskyの定理

d 次元確率変数列 $\{\boldsymbol{X}_n\}$ と与えられた写像 $\boldsymbol{f}\colon \mathbb{R}^d \to \mathbb{R}^\ell$ に対して

$\boldsymbol{f}(\boldsymbol{X}_n)$ の $n \to \infty$ での振る舞いを調べるのに, 次の**Slutskyの定理**(スラツキー)が役立つ:

Thm. (Slutskyの定理)

d 次元確率変数 $\{\boldsymbol{X}_n\}, \{\boldsymbol{Y}_n\}, \boldsymbol{X}$ と k 次元確率変数 $\{\boldsymbol{Z}_n\}, \boldsymbol{Z}$ を考える.

(1) $\boldsymbol{f}\colon \mathbb{R}^d \to \mathbb{R}^\ell$ に対し, $C(\boldsymbol{f})$ を $\boldsymbol{f}$ の連続点の集合とする.

$\boldsymbol{X}_n \xrightarrow{*} \boldsymbol{X}$ かつ $\Pr\big(\boldsymbol{X} \in C(\boldsymbol{f})\big) = 1$ ならば, $\boldsymbol{f}(\boldsymbol{X}_n) \xrightarrow{*} \boldsymbol{f}(\boldsymbol{X})$. $(* = \mathrm{a.s.}, \mathrm{p}, \mathrm{d})$ ↓ L^p では成り立たない.

(2) $\boldsymbol{X}_n \xrightarrow{*} \boldsymbol{X}$ かつ $\boldsymbol{X}_n - \boldsymbol{Y}_n \xrightarrow{\mathrm{p}} 0$ ならば, $\boldsymbol{Y}_n \xrightarrow{*} \boldsymbol{X}$. $(* = \mathrm{p}, \mathrm{d})$ ← すべて a.s. の収束にしたものも成立.

(3) $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X}$ かつ $\boldsymbol{Z}_n \xrightarrow{\mathrm{d}} \boldsymbol{c}$ (**定数ベクトル**) ならば, $\begin{pmatrix} \boldsymbol{X}_n \\ \boldsymbol{Z}_n \end{pmatrix} \xrightarrow{\mathrm{d}} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{c} \end{pmatrix}$.

$\boldsymbol{X}_n \xrightarrow{*} \boldsymbol{X}$ かつ $\boldsymbol{Z}_n \xrightarrow{*} \boldsymbol{Z}$ ならば $\begin{pmatrix} \boldsymbol{X}_n \\ \boldsymbol{Z}_n \end{pmatrix} \xrightarrow{*} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Z} \end{pmatrix}$. $(* = \mathrm{a.s.}, L^p, \mathrm{p})$

cf.) (1) は**連続写像定理**ともいう.

➔ 利用方法を見てみよう.

Slutskyの定理の利用例

ex) $X_n \xrightarrow{\mathrm{d}} X \sim \mathcal{N}(0,1)$ とする.

$f(x) = 1/x$ は $x = 0$ 以外で連続. $\Pr(X \neq 0) = 1$ なので, $1/X_n \xrightarrow{\mathrm{d}} 1/X$.

変数変換を考えると, $Z = 1/X$ の確率密度関数は,

$$g(z) = \frac{1}{\sqrt{2\pi}\, z^2} \exp\left(-\frac{1}{2z^2}\right) 1_{\{z \neq 0\}}(z).$$

さらに, Slutskyの定理の(1)(3)を組み合わせると次が言える:

Cor. (Slutskyの定理の系)

d 次元確率変数 $\{\boldsymbol{X}_n\}, \boldsymbol{X}$ と k 次元確率変数 $\{\boldsymbol{Y}_n\}, \boldsymbol{Y}$ を考える.

$\boldsymbol{f}\colon \mathbb{R}^{d+k} \to \mathbb{R}^{\ell}$ に対し, $C(\boldsymbol{f})$ を $\boldsymbol{f}$ の連続点の集合とし, $\boldsymbol{c} \in \mathbb{R}^k$ とする.

(1) $\boldsymbol{X}_n \xrightarrow{\mathrm{d}} \boldsymbol{X}, \boldsymbol{Y}_n \xrightarrow{\mathrm{d}} \boldsymbol{c}$ で $\Pr\left(\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{c}\end{pmatrix} \in C(\boldsymbol{f})\right) = 1$ ならば, $\boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{Y}_n) \xrightarrow{\mathrm{d}} \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{c})$.

(2) $\boldsymbol{X}_n \xrightarrow{\mathrm{p}} \boldsymbol{X}, \boldsymbol{Y}_n \xrightarrow{\mathrm{p}} \boldsymbol{Y}$ で $\Pr\left(\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix} \in C(\boldsymbol{f})\right) = 1$ ならば, $\boldsymbol{f}(\boldsymbol{X}_n, \boldsymbol{Y}_n) \xrightarrow{\mathrm{p}} \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{Y})$.

※ (2) はすべて a.s. の収束にしたものも成立.

デルタ法

デルタ法 (delta method) は, 関数で変換した後のパラメータの推定に関して有用.

Thm. (デルタ法)

1次元確率変数 $\{X_n\}, X$ と定数 $\theta \in \mathbb{R}$,

単調増加し発散する数列 $\{a_n\}_{n=1}^{\infty}$ ($a_n \nearrow \infty$) に対し, $a_n(X_n - \theta) \xrightarrow{\mathrm{d}} X$ とする.

関数 $g \colon \mathbb{R} \to \mathbb{R}$ について, θ で連続的微分可能で $g'(\theta) \neq 0$ とする. このとき,

$$a_n\big(g(X_n) - g(\theta)\big) \xrightarrow{\mathrm{d}} g'(\theta)X.$$

とくに, $\sqrt{n}(X_n - \mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2)$ のときは,

$$\sqrt{n}\big(g(X_n) - g(\mu)\big) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2\big(g'(\mu)\big)^2\right).$$

pf. 略. (Slutskyの定理を利用する.)

cf.) $g'(\theta) = 0$ のときは「2次のデルタ法」を利用する.

➔ デルタ法の利用方法を見てみよう.

デルタ法の利用法

$\sqrt{n}(X_n - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2)$ かつ, $\sigma^2 = \sigma^2(\theta)$ (未知パラメータに依存)

という状況を考える. $\boldsymbol{\theta}$ の推定をする際の分散 $\boldsymbol{\sigma^2(\theta)}$ が $\boldsymbol{\theta}$ に依存しており,

これを何らかの方法で推定する必要がある.

➔ **θ の推定にさらなる誤差が生じてしまう**.

$\sqrt{n}\big(g(X_n) - g(\theta)\big) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2(\theta)\big(g'(\theta)\big)^2\right)$ なので,

$\sigma^2(\theta)\big(g'(\theta)\big)^2 = c$ (定数) となるように g を選ぶと,

$g(\theta)$ の分散が θ に依存しなくなる!

これを**分散安定化変換** (variance stabilization) という.

ここまでのまとめ

- **概収束**: 確率変数がa.s.で各点収束
- **確率収束**: 「確率変数が収束先から離れる確率」が0に収束
- **p 次平均収束**: 確率変数間の「距離」が0に収束
- **分布収束**: 確率変数の従う分布が収束
  - ポルトマントーの補題, Lévyの連続性定理を利用して示せる.
- 収束性の強弱関係: 概収束/p 次平均収束 $\Rightarrow$ 確率収束 $\Rightarrow$ 分布収束
- **Slutskyの定理**: 変数変換後の収束先に関する性質.
- **デルタ法**: 推定誤差を抑える分散安定化変換に利用.

極限定理

大数の法則

n 個の標本 $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ の標本平均 $\overline{\boldsymbol{X}}_n$ を求めることで母平均 $\boldsymbol{\mu}$ を推定することの妥当性を与えるのが, 次の**大数の法則**.

<u>**Thm.**</u> (大数の弱/強法則: Weak/Strong Law of Large Numbers)

d 次元確率変数列 $\{\boldsymbol{X}_n\}$ はi.i.d.で $\overline{\boldsymbol{X}}_n := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$ とする.

(1) [**弱法則**] $\mathbb{E}[\|\boldsymbol{X}_1\|] < \infty$ ならば, $\overline{\boldsymbol{X}}_n \xrightarrow{\mathrm{p}} \boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_1]$ $(n \to \infty)$.

(2) $\mathbb{E}[\|\boldsymbol{X}_1\|^2] < \infty$ ならば, $\overline{\boldsymbol{X}}_n \xrightarrow{L^2} \boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_1]$ $(n \to \infty)$.

(3) [**強法則**] $\mathbb{E}[\|\boldsymbol{X}_1\|] < \infty$ かつ $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_1] \Leftrightarrow \overline{\boldsymbol{X}}_n \xrightarrow{\mathrm{a.s.}} \boldsymbol{\mu}$ $(n \to \infty)$.

<u>pf.</u> 略. (Fergusonを参照)

「$n \to \infty$ で**標本平均が母平均に (ある意味で) 収束する** (**一致性**をもつ)」

➔ 十分大きな n では, 標本平均 $\overline{\boldsymbol{X}}_n$ が母平均 $\boldsymbol{\mu}$ の**「良い」推定量**になる!

標本分散の一致性

n 個の標本 $X_1, \dots, X_n$ の標本分散 $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ を求めることで母分散 σ^2 を推定するのも, **大数の法則を考えれば妥当**.

$M_2 \coloneqq \frac{1}{n}\sum_{i=1}^n X_i^2$ (2次のモーメントの推定量) とすると,

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i^2 - 2\bar{X}_n X_i + \bar{X}_n{}^2\right) = \frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{2}{n}\bar{X}_n \sum_{i=1}^n X_i + \bar{X}_n{}^2$$
$$= M_2 - 2\bar{X}_n{}^2 + \bar{X}_n{}^2 = M_2 - \bar{X}_n{}^2.$$

大数の強法則より, $\bar{X}_n \xrightarrow{\text{a.s.}} \mu = \mathbb{E}[X_1]$ かつ $M_2 \xrightarrow{\text{a.s.}} \mathbb{E}[X_1^2]$.

Slutskyの定理 (の系) より,

$$S_n^2 \xrightarrow{\text{a.s.}} \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \sigma^2$$

となって, 標本分散 S_n^2 は $n \to \infty$ で σ^2 に概収束する (**一致性をもつ**).

(なお, 再びSlutskyの定理より, $(\bar{X}_n, S_n^2) \xrightarrow{\text{a.s.}} (\mu, \sigma^2)$ と同時概収束する.)

➔ 十分大きな n では, 標本分散 S_n^2 が母分散 σ^2 の **「良い」推定量**になる!

経験分布関数の一致性

n 個の標本が $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{P}$ とし, P の分布関数を F とする.

任意の $x \in \mathbb{R}$ に対し, 分布関数の値 $F(x) = \Pr(X \leq x)$ は不明なのが普通.

$F(x)$ の推定量として, 次の**経験分布関数** (empirical distribution function)

$$F_n(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i) \quad \leftarrow x \text{ 以下に出現する観測値の割合}$$

を利用する (これは各 x を固定するごとに確率変数, i.e., ω の可測関数).

$$\mathbb{E}\left[1_{(-\infty,x]}(X_i)\right] = \int_{\mathbb{R}} 1_{(-\infty,x]}(x')\,\mathrm{P}(\mathrm{d}x') = \int_{-\infty}^{x} \mathrm{P}(\mathrm{d}x') = \Pr(X_i \leq x) = F(x)$$

となり, $\left\{1_{(-\infty,x]}(X_i)\right\}$ はi.i.d.となることがいえる.

大数の強法則より $F_n(x) \xrightarrow{\text{a.s.}} F(x)$ が各 $x \in \mathbb{R}$ に対して成立 (**一致性をもつ**).

➔ 十分大きな n では, 各 x に対して $F_n(x)$ が $F(x)$ の**「良い」推定量**になる!
　では, $F_n(x)$ は「関数としては」$F(x)$ のよい推定になるのだろうか?

Glivenko–Cantelliの定理

実は,「各 x に対して $F_n(x) \xrightarrow{\text{a.s.}} F(x)$」よりも強い結果が成立する:

グリベンコ　カンテリ

Thm. (Glivenko–Cantelliの定理)

n 個の標本が $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{P}$ とし, P の分布関数を F とする.

経験分布関数を $F_n(x)$ とすると,

$$\mathrm{KS}_n \coloneqq \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$$

が成り立つ.

コルモゴロフ　スミルノフ

cf.) KS_n は**Kolmogorov–Smirnov統計量**と呼ばれ, 標本が F に対応する分布に従うかを統計的に検定するのに利用される.

➔ $F_n(x)$ は x **に関して一様に** $F(x)$ に概収束することがわかるので, n が十分大きければ**経験分布関数により分布関数を近似できる**!

モンテカルロ法

大数の法則を利用することで, 積分の数値計算もできる:

Cor. (モンテカルロ法)

n 個の d 次元標本が $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \mathrm{P}$ とし, P は絶対連続型で確率密度関数 p を持つとする. $\mathbb{E}[|f(\boldsymbol{X}_1)|] < \infty$ ならば, 関数 $f\colon \mathbb{R}^d \to \mathbb{R}$ に対し, 以下が成立:

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i) \xrightarrow{\text{a.s.}} \mathbb{E}[f(\boldsymbol{X}_1)] = \int_{\mathbb{R}^d} f(\boldsymbol{x}) p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \ \ (n \to \infty)$$

ex) $I = \int_a^b f(x)\,\mathrm{d}x \ \ (-\infty < a < b < \infty)$ の計算 (f: 与えられた関数)

$I = \int_a^b (b-a) f(x) \cdot \frac{1}{b-a}\,\mathrm{d}x$ で, $\frac{1}{b-a}$ は $U(a,b)$ の密度関数. よって,

$$I = \mathbb{E}[(b-a)f(X)], \qquad X \sim U(a,b)$$

なので, $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} U(a,b)$ として $\frac{1}{n}\sum_{i=1}^{n}(b-a)f(X_i)$ で I を推定できる.

大数の弱法則の主張について

大数の弱法則: $\mathbb{E}[\|\boldsymbol{X}_1\|] < \infty$ ならば, $\overline{\boldsymbol{X}}_n \xrightarrow{\mathrm{p}} \boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_1]$ $(n \to \infty)$.

$\overline{\boldsymbol{X}}_n \xrightarrow{\mathrm{p}} \boldsymbol{\mu} \Leftrightarrow \forall \varepsilon > 0, \lim_{n\to\infty} \Pr(\|\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\| > \varepsilon) = 0$ だった.

これをもう少し詳しく書いてみると,

↓ε と δ に依存して決まる

任意の $\varepsilon, \delta > 0$ に対してある $N \in \mathbb{Z}_{>0}$ が存在し, $n \geq N$ のとき
$\overline{\boldsymbol{X}}_n$ の $\boldsymbol{\mu}$ からの誤差 $\|\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\|$ が ε より大きくなる確率は δ 未満

という主張になっている.

➔ 「$n \to \infty$ で $\overline{\boldsymbol{X}}_n$ が $\boldsymbol{\mu}$ に収束する」ということは**述べていない**!

大数の強法則の主張について

大数の強法則: $\mathbb{E}[\|X_1\|] < \infty$ かつ $\boldsymbol{\mu} = \mathbb{E}[X_1] \Leftrightarrow \overline{X}_n \xrightarrow{\text{a.s.}} \boldsymbol{\mu}\ (n \to \infty)$.

$\overline{X}_n \xrightarrow{\text{a.s.}} \boldsymbol{\mu} \Leftrightarrow \Pr\left(\lim_{n\to\infty} \overline{X}_n = \boldsymbol{\mu}\right) = 1$ だった.

これをもう少し詳しく書いてみると,

↓ω と ε に依存して決まる

ほとんどすべての $\omega \in \Omega$ で, 任意の $\varepsilon > 0$ に対してある $N \in \mathbb{Z}_{>0}$ が存在し,
$n \geq N$ のとき $\overline{X}_n(\omega)$ の $\boldsymbol{\mu}$ からの誤差 $\|\overline{X}_n(\omega) - \boldsymbol{\mu}\|$ が ε 未満

という主張になっている.

➔ 「$n \to \infty$ で $\overline{X}_n$ が $\boldsymbol{\mu}$ に収束する」ことを述べている! 大数の弱法則と強法則は, 主張の見た目は似ていても**異なったことを述べている**のに注意.

重複対数の法則

実は, 大数の法則での収束の様子をより精密に述べることができる:

Thm. (重複対数の法則: Law of Iterated Logarithm)

1次元確率変数列 $\{X_n\}_{n=1}^{\infty}$ はi.i.d.で, $\mu := \mathbb{E}[X_1], \sigma^2 := \mathbb{V}[X_1] < \infty$ とする.

(1) $\limsup_{n\to\infty} \frac{\bar{X}_n - \mu}{\sqrt{2n^{-1}\log\log n}} = \sigma, a.s.$ $\left(\liminf_{n\to\infty} \frac{\bar{X}_n - \mu}{\sqrt{2n^{-1}\log\log n}} = -\sigma, a.s.\right)$

(2) 定数 $a, \tau < \infty$ が $\limsup_{n\to\infty} \frac{\bar{X}_n - a}{\sqrt{2n^{-1}\log\log n}} = \tau$ ならば, $a = \mu, \tau^2 = \sigma^2$.

pf. 略 (cf. Chow and Teicher, 1988 や Billingsley,1995)

(1) は「ほとんどすべての ω で, 任意の $\varepsilon > 0$ に対してある $N \in \mathbb{Z}_{>0}$ があり, $n \geq N$ のとき $\left|\sup_{k\geq n} \frac{\bar{X}_k(\omega) - \mu}{\sqrt{2k^{-1}\log\log k}} - \sigma\right| < \varepsilon$」という内容.

➔ μ からの変動 $\bar{X}_n - \mu$ の大きさの上下限を示す. とくに, **$\bar{X}_n$ の μ への収束スピードが $\sqrt{2n^{-1}\log\log n}$ と同程度**であることがわかる.

cf. Cramérの定理

大数の弱法則によれば, 任意の $\varepsilon > 0$ に対して $\lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| > \varepsilon) = 0$.

➔ $\Pr(|\bar{X}_n - \mu| > \varepsilon)$ は一体どの程度の速さで0に収束するだろうか?

Thm. (Cramérの定理)

n 個の標本 $X_1, \ldots, X_n$ はi.i.d.とし, すべての $t \in \mathbb{R}$ に対して

モーメント母関数 $M(t) = \mathbb{E}[e^{X_1 t}]$ が存在するとする. $\mu \coloneqq \mathbb{E}[X_1]$ とする.

X_1 の**大偏差収束率関数** (rate function) $I(x)$ を $\log M(t)$ のLegendre変換

$$I(x) \coloneqq \sup_{t\in\mathbb{R}}(tx - \log M(t))$$

で与えると, すべての $x > \mu$ と $n \in \mathbb{Z}_{>0}$ に対して以下が成立:

$$\Pr(\bar{X}_n \geq x) \leq \exp\bigl(-nI(x)\bigr).$$

特に, $\Pr(\bar{X}_n - \mu \geq \varepsilon) \leq \exp\bigl(-nI(\mu + \varepsilon)\bigr)$ となり, 指数関数的に0に近づくことがわかる. 同様に, $\Pr(\bar{X}_n - \mu \leq -\varepsilon)$ も指数関数的に0に近づく.

➔ 標本平均 $\bar{X}_n$ が μ から外れる確率は, n が大きくなるにつれて **指数関数的に0に近づく**. (cf. **大偏差原理**)

大数の法則の収束の様子を調べる

大数の強法則により, $\bar{X}_n \xrightarrow{\text{a.s.}} \mu = \mathbb{E}[X_1]$ がわかった. この収束の様子について
もう少し詳しく考えてみよう.

発散する実数列 $\{r_n\}_{n=1}^{\infty}, r_n \to \infty$ を考える. まず, 以下がわかる:

- $r_n(\bar{X}_n - \mu) \xrightarrow{\text{a.s.}} 0$ なら, $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$ となる速さは $r_n^{-1} \to 0$ よりも**速い**.
- $r_n(\bar{X}_n - \mu) \xrightarrow{\text{a.s.}} \infty$ なら, $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$ となる速さは $r_n^{-1} \to 0$ よりも**遅い**.

r_n^{-1} の収束スピードとの比較以外にも収束の様子を調べられないか, i.e.,

r_n をうまく選ぶことで, $r_n(\bar{X}_n - \mu) \xrightarrow{\text{a.s.}} 0$ や $r_n(\bar{X}_n - \mu) \xrightarrow{\text{a.s.}} \infty$ とならない
非自明な極限が現れないだろうか?

➔ 実は, r_n をうまく選ぶと, $r_n(\bar{X}_n - \mu)$ は**分布収束する**!
　このときの r_n を $\bar{X}_n$ の**収束率** (rate of convergence) という.

➔ では, $\bar{X}_n$ の収束率はどうなるだろうか?

中心極限定理

$\bar{X}_n$ の収束率が $\sqrt{n}$ であることを示すのが, **中心極限定理**!

Thm. (多次元版の中心極限定理: Central Limit Theorem)

d 次元確率変数列 $\{\boldsymbol{X}_n\}$ はi.i.d.で平均 $\boldsymbol{\mu} \coloneqq \mathbb{E}[\boldsymbol{X}_1]$ と有限な分散行列

$\Sigma \coloneqq \mathbb{V}[\boldsymbol{X}_1]$ をもつとする. $\bar{\boldsymbol{X}}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$ とする. このとき,

$$\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{\mathrm{d}} \mathcal{N}(\boldsymbol{0}, \Sigma), \qquad n \to \infty.$$

pf. 略 (cf. Ferguson)

「十分大きな n では, $\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})$ の分布を $\mathcal{N}(\boldsymbol{0}, \Sigma)$ **で近似できる**」
ということを述べている. 統計学では, 特に重要になる定理!

➔ 以下, この定理についてもう少し詳しく見ていく.

cf. Lindeberg-Fellerの定理

実は, 独立だが同分布ではない確率変数列への中心極限定理の拡張がある:

Thm. (Lindeberg-Fellerの定理)
リンデベルグ　フェラー

1次元確率変数列 $\{X_n\}_{n=1}^{\infty}$ は独立であるとし, $\mu_i \coloneqq \mathbb{E}[X_i], \sigma_i^2 \coloneqq \mathbb{V}[X_i] < \infty$ とする. $S_n \coloneqq \sum_{i=1}^{n} X_i$, $B_n^2 \coloneqq \sum_{i=1}^{n} \sigma_i^2$ とする. このとき, **Lindeberg条件**:

任意の $\varepsilon > 0$ に対して $\frac{1}{B_n^2}\sum_{i=1}^{n} \mathbb{E}\left[(X_i - \mu_i)^2\, 1_{\{|X_i - \mu_i| \geq \varepsilon B_n\}}\right] \to 0 \ (n \to \infty)$

が成立することと,

$$\text{(i)} \lim_{n\to\infty} \max_{1\leq i\leq n} \frac{\sigma_i^2}{B_n^2} = 0 \text{ かつ (ii) } B_n^{-1}(S_n - \mathbb{E}[S_n]) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

が成立することは同値.

Lindeberg条件は「$B_n^{-1}\sum_{i=1}^{n}(X_i - \mu_i)$ の各項 $B_n^{-1}|X_i - \mu_i|$ が一様に小さい」ことを表す. (i) は「X_i の分散 σ_i^2 の S_n の分散 B_n^2 への寄与が小さい」ことを表す. (cf. 蓑谷『数理統計ハンドブック』)

Berry-Esseenの定理

中心極限定理により, $\sqrt{n}(\bar{X}_n - \mu)$ の分布を $\mathcal{N}(0, \sigma^2)$ で近似できることがわかった.

この近似の誤差に関する評価を与えるのが, 次の**Berry-Esseenの定理**(ベリー エシーン).

<u>**Thm.**</u> (Berry-Esseenの定理)

1次元確率変数列 $\{X_n\}_{n=1}^{\infty}$ はi.i.d.で, $\mu \coloneqq \mathbb{E}[X_1], \sigma^2 \coloneqq \mathbb{V}[X_1]$, および3次の平均周りの絶対モーメント $\rho = \mathbb{E}[|X_1 - \mu|^3] < \infty$ をもつとする. F_n, Φ をそれぞれ $\sqrt{n}(\bar{X}_n - \mu)/\sigma, \mathcal{N}(0,1)$ の分布関数とすると, 任意の $n \in \mathbb{Z}_{>0}$ に対して

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{c\rho}{\sqrt{n}\sigma^3}$$

が成立する. c は定数で, $0.4097 \leq c \leq 0.4784$ である.

➔ とくに, **分布収束の収束スピードが $\frac{1}{\sqrt{n}}$ と同程度**であることもわかる.

標本平均の漸近正規性

母平均が μ, 母分散が σ^2 の母集団分布を P とし, P からの n 個の (1次元) ランダムサンプルを $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{P}$ としよう.

標本平均 $\bar{X}_n$ の従う分布を厳密に計算するのは困難 (cf. p.160)

n が十分大きいものとして, 近似分布を求めてみよう. 中心極限定理から

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2) \ (n \to \infty).$$

➔ n が十分大きいとき, $\sqrt{n}(\bar{X}_n - \mu)$ は**漸近的に** $\mathcal{N}(0, \sigma^2)$ に従う. これを

$$\sqrt{n}(\bar{X}_n - \mu) \overset{\text{asym.}}{\sim} \mathcal{N}(0, \sigma^2)$$

← 分布収束することの別表記でしかない.

と書こう. (右辺の分布を**漸近分布** (asymptotic distribution) といい, $\bar{X}_n$ は**漸近正規性** (asymptotic normality) をもつという.)

とくに, 正規分布の性質より, $\bar{X}_n \overset{\text{asym.}}{\sim} \mu + \frac{1}{\sqrt{n}} \mathcal{N}(0, \sigma^2) = \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

➔ 標本平均は n **が十分大きいときに正規分布に従うと近似できる**.

t-統計量の漸近正規性

ところで, $\bar{X}_n \overset{\text{asym.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ がわかっても, σ^2 が不明な場合があった.
t-統計量のときのように, σ^2 を不偏分散 V^2 で推定しても良いだろうか?

$V^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\bar{X}_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right).$

大数の法則から $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{d}} \mathbb{E}[X_1^2], \bar{X}_n \xrightarrow{\text{d}} \mu$ なので, Slutskyの定理から

$$V^2 \xrightarrow{\text{d}} 1 \cdot (\mathbb{E}[X_1^2] - \mu^2) = \sigma^2.$$

$\bar{X}_n \overset{\text{asym.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ より, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0,1)$. 再びSlutskyの定理から,

$$T = \frac{\bar{X}_n - \mu}{V/\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

t-統計量が漸近正規性をもつことがわかった. とくに,

$$\bar{X}_n \overset{\text{asym.}}{\sim} \mathcal{N}\left(\mu, \frac{V^2}{n}\right).$$

➔ 標本平均の漸近分布の分散を**推定値で代用するのは妥当**.

de Moivre-Laplaceの定理

Bernoulli分布から n 個のランダムサンプル $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(1, p)$ を得たとする. $S_n \coloneqq \sum_{i=1}^{n} X_i = n\bar{X}_n$ の近似分布を求めてみよう.

(とくに, $S_n \sim \mathrm{Bin}(n, p)$ で, $\mathbb{E}[S_n] = np, \mathbb{V}[S_n] = np(1-p)$ である.)

$\mathbb{E}[X_1] = p, \mathbb{V}[X_1] = p(1-p)$ であり, 中心極限定理より,

$$\sqrt{n}(\bar{X}_n - p) \overset{\text{asym.}}{\sim} \mathcal{N}\big(0, p(1-p)\big).$$

これより,

$$S_n - np = n(\bar{X}_n - p) \overset{\text{asym.}}{\sim} \sqrt{n}\mathcal{N}\big(0, p(1-p)\big) = \mathcal{N}\big(0, np(1-p)\big)$$

となり,

$$S_n \overset{\text{asym.}}{\sim} \mathcal{N}\big(np, np(1-p)\big) = \mathcal{N}(\mathbb{E}[S_n], \mathbb{V}[S_n])$$

であることがわかる. (**de Moivre(ド・モアブル)-Laplace(ラプラス)の定理**)

➔ 二項分布に従う変数は, n **が十分大きいときに正規分布に従うと近似できる**.

ここまでのまとめ

- **大数の弱法則**: 標本平均は母平均に確率収束.
  - 「母平均からの誤差が大きくなる確率が小さくなっていく」
- **大数の強法則**: 標本平均は母平均に概収束.
  - 「母平均からの誤差が小さくなっていく」
- **Glivenko-Cantelliの定理**:
  経験分布関数は, 真の分布関数に一様に概収束.
- **重複対数の法則**: 標本平均が母平均に概収束する速さは $\sqrt{\frac{\log\log n}{n}}$ と同程度.
- **中心極限定理**:
  標本平均と母平均との差に $\sqrt{n}$ をかけたものは, n が十分大きいときは正規分布で近似できる.
- **Berry-Esseenの定理**: 中心極限定理の分布収束の速さは $\frac{1}{\sqrt{n}}$ と同程度.

参考文献

参考文献1

[1] 清水『統計学への確率論, その先へ』, 内田老鶴圃, 2019.

[5] [6] などの一般的な確率論の本と比べると, かなり丁寧に書かれていてわかりやすい. Lebesgue積分などの話についても丁寧に書かれている. 今回, 説明を考える際に大いに参考にした.

[2] 平岡, 堀『プログラミングのための確率統計』, オーム社, 2009.

初学者向け. わかりやすい例や喩えも多く, とっつきやすい. イメージを掴むためのはじめの一冊に良い.

[3] 高橋『確率論』, 朝倉書店, 2008.

初学者向け. [2] よりは数学書らしい本. 離散確率の場合からはじまり, 公理的確率論の入門までカバーされている. 説明が丁寧なので, 数学書としてはかなり読みやすい.

[4] 成田『確率解析への誘い』, 共立出版, 2016.

内容は [3] より難しいものの, 解析学の基本事項から丁寧に書かれているので読みやすい. きちんと確率論を勉強する際には良い. 確率過程や確率微分方程式の話も載っている.

[5] 舟木『確率論』, 朝倉書店, 2004.

[6] 熊谷『確率論』, 共立出版, 2003.

確率論の入門的なところを一通り終えたら, こうした少し難しめの本に挑戦してみるとよい.
[5] は比較的読みやすい.

参考文献2

[7] 久保川『現代数理統計学の基礎』, 共立出版, 2017.

確率論にあまり深入りすることなく, 数理統計全般についてコンパクトにまとまっている読みやすい本.
統計検定1級の勉強にはちょうど良い. 全体の流れを考えるときに参考にした.

[8] 竹村『現代数理統計学』, 創文社, 1991.

かなり説明が丁寧に書かれている本で, とてもわかりやすい. 一読の価値あり.
現在は絶版しているが, 新装改訂版が学術図書出版社から出ており, 演習問題の解答もついた.

[9] 吉田『数理統計学』, 朝倉書店, 2006.

測度論の部分がきちんと書いてある. 測度論にある程度なれておかないと, 読むのは難しい.

[10] 蓑谷『数理統計ハンドブック』, みみずく舎, 2009.

辞書的な本. 読み通すような本ではないが, いろいろと載っている.

[11] Ferguson『必携 統計的大標本論 その基礎理論と演習』, 共立出版, 2017.

幅広い話題が取り扱われているが, 中級者向け. 演習本としても良い.

参考文献3

[12] 尾畑『確率モデル要論 –確率論の基礎からマルコフ連鎖へ–』, 牧野書店, 2012.

サブタイトルにある通り, 前半は確率論の基礎的な部分, 後半はMarkov連鎖に関する話題が取り扱われている. 少々マニアックな話題も多少載っていて面白い.

[13] 黒木『数理統計学 –統計的推論の基礎–』, 共立出版, 2020.

適度に厳密な理論展開で読みやすい. 数理統計の基本的な部分はこれでカバーできるので, 数理統計のはじめの一冊に良い.

[14] van der Vaart, “Asymptotic Statistics”, Cambridge University Press, 1998.

[15] DasGupta, “Probability for Statistics and Machine Learning”, Springer, 2011.

[16] Gut, “Probability: A Graduate Course”, Springer, 2013.

これらの本も定理の主張などで参考にした.