確率論・統計学（工学）

@Metachick_2021

May 22, 2025

準備：記号の定義

以降、以下の記号は断りなく使用する。

- $\mathbb{N}$：自然数の集合
- $\mathbb{Z}$：整数全体の集合
- $\mathbb{Q}$：有理数全体の集合
- $\mathbb{R}$：実数全体の集合
- $\mathbb{C}$：複素数全体の集合
- $X_{>k}$：集合 X の要素のうち、k より大きいもの全体の集合
- $X \setminus Y$：X と Y の差集合

注意点

- スライドは定義や定理などの事実と、大まかなポイントのみを載せています
- 必ずしもすべての用語がこのスライド内で定義されているとは限りません、あくまで要点のみです
- 細かいポイントなどは口頭で説明していくので、適宜メモを活用してください
- 厳密性とわかりやすさのバランスは半々くらいを目標としました
- ミスがあれば教えてください

目次

確率論の基礎

「まずは、お堅い話をします！」

- 少し真面目に、公理的に確率論を構成していきます。
- ここが理解できなくても、困ることはあまりないので気にせず先へ…。

確率の公理

公理（Колмого́ров）

集合 Ω とその部分集合族 $\mathcal{F}$ に対して、関数 $P : \mathcal{F} \to \mathbb{R}$ が以下の 3 条件を満たすとき、組 $(\Omega, \mathcal{F}, P)$ を確率空間という。

1. $\forall A \in \mathcal{F}, 0 \leq P(A)$
2. $P(\Omega) = 1$
3. 互いに素な事象列 $\{A_i\}_{i=1}^{\infty}$ に対して、$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

- Ω：標本空間、試行で起こりうる事象全体。（ex：$\{$表, 裏$\}$ など）
- $\mathcal{F}$：事象の族、Ω が可算集合の場合には、冪集合とすればよい。Ω が非可算集合の場合には、Ω 上の完全加法族を採用する。（ex：$\{\{\}, \{$表$\}, \{$裏$\}, \{$表, 裏$\}\}$ など）
- P：確率、$\mathcal{F}$ 上の関数であってその起こりやすさを割り当てるもの。

確率の基礎的な命題

命題（基本的な事実）

確率空間 $(\Omega, \mathcal{F}, P)$ に対して、以下が成り立つ。

1. $P(\{\}) = 0$
2. $A \subset B \implies P(A) \leq P(B)$
3. $P(A^c) = 1 - P(A)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$（加法定理）

証明は事象の分割や公理から直接得られる。

条件付き確率とベイズの定理

定義（条件付き確率）

事象 B が起こったときの事象 A の条件付き確率 $P(A|B)$ を以下で定める：

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- 表記が紛らわしい…。$A|B \to A/B \to \frac{A}{B}$ みたいなイメージ？
- 定義から直ちに $P(B \mid A) = \frac{P(A|B)\,P(B)}{P(A)}, \quad P(A) > 0.$ が得られる。（ベイズの定理）
- 事象 A, B が独立とは，$P(A \cap B) = P(A)P(B)$ をいう。このとき $P(A \mid B) = P(A)$ となる。

確率 0 は不可能ではない

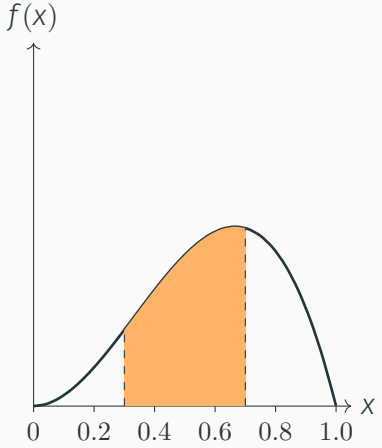
「$[0, 1]$ からランダムに実数を選ぶとき、0.5 が選ばれる確率は？」

- P(選んだ実数 $= 0.500000\cdots$) と P(選んだ実数 $= 0.500001\cdots$) は区別する。
- 総和が 1 になるように各標本 $\omega \in \Omega$ に確率を割り当てようとすると、高々可算個の標本にしか正の確率を割り当てることができない。
- そこで、確率を個々の標本に割り当てるのではなく、標本の集合に割り当てる。
- P(選んだ実数 $= 0.500000\cdots$) を考えることは不可能だが、
  $P(0.5 \leq$ 身長 $\leq 0.5001)$ を考えることならできるのではないだろうか。

確率 0 は不可能ではない

$$P\left(a \leq X \leq b\right) = \int_a^b f(x)\,\mathrm{d}x$$

- 面積がそのまま確率になる
- この考え方を用いて、連続の確率を考える。
- この関数 f を確率密度関数という。



確率 0 は不可能ではない

確率空間 $(\Omega, \mathcal{F}, P)$ について、Ω が非可算集合であるときには、$\mathcal{F}$ は Ω 上の σ-代数を採用する。例えば、$\mathcal{F}$ は開区間、閉区間、半開区間の可算合併、可算交叉で書けるもの全体などと定義される[1]。前頁を踏まえて考えると…

- 事象 ω に対して直接確率を割り当てるのではなく、ω を一度数値 x に対応させて、x の条件と確率を結びつけるという考え方のほうが普遍的であると思い至る。
- もう少し正確な言葉運びをすれば、「まず関数 $X : \Omega \to \mathbb{R}$ によって ω を数値 x に写し，その逆像として得られる事象 $X \in B$ に確率を割り当てる方が普遍的である」

[1] 正確に述べるのなら、$[0, 1]$ 上の開区間全体に対して、補集合と可算合併、可算交叉を繰り返して得られる最小の σ-代数：ボレル σ-代数 $\mathcal{B}([0, 1])$ など

確率変数

確率変数

確率空間 $(\Omega, \mathcal{F}, P)$ 上で，標本点 $\omega \in \Omega$ に数値を対応させる関数 $X : \Omega \to \mathbb{R}$ を確率変数と呼ぶ。

- Ω が可算集合の場合：例えば、$\Omega = \{\text{表}, \text{裏}\}$ について、$X(\text{表}) = 1, X(\text{裏}) = 0$ と割り当てればよい。このとき、確率は $P(\text{表}) = P(\{\omega \mid X(\omega) = 1\})$ などと表すことができる。
- Ω が非可算集合の場合：例えば、$\Omega = [0, 1]$ について、$X(\omega) = \omega$ と割り当てる。確率を $P(\{\omega | 0 \leq X(\omega) \leq 0.5\})$ などとすればよい。

確率分布（累積分布関数）

確率変数が取りうる値や、そのルールを表現する方法を確率分布といい、主に分布関数と質量/密度関数の二つが存在する。

定義（累積分布関数）

確率変数 X の累積分布関数 $F : \mathbb{R} \to [0, 1]$ を以下で定める。

$$F(x) = P(X \leq x)$$

- F は広義単調増加関数で，右連続かつ $\lim_{x \to -\infty} F(x) = 0,\ \lim_{x \to +\infty} F(x) = 1$
- 離散分布では階段状のステップ関数，連続分布では滑らかな関数となる。

確率分布（確率質量関数）

定義（確率質量関数, PMF）

離散型確率変数 X の確率質量関数 $p : \mathbb{R} \to [0, 1]$ を以下で定める。

$$p(x) = P(X = x)$$

- 「可算集合の時には、そのまま確率を割り当てる」という発想を体現している
- $\forall x \in \mathbb{R}, p(x) > 0$
- $\sum_{x \in \mathcal{X}} p(x) = 1$ （ただし、$\mathcal{X} = \{x \mid p(x) > 0\}$）

確率分布（確率密度関数）

定義（確率密度関数, PDF）

連続型確率変数 X の確率密度関数 $f : \mathbb{R} \to [0, \infty)$ を以下で定める。

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \land \int_{-\infty}^{\infty} f(x)\,dx = 1$$

- 「非可算集合の時には、”領域”に確率を割り当てる」という発想を体現している
- 確率密度関数は累積分布関数の導関数：$f(x) = F'(x)$（微分可能な場合）

演習問題

1. 命題（基本的な事実）の 1〜4 について、それぞれ証明を与えよ。
2. 外見が同じ壺 A,B があり、A には白い球が 49 つ、黒い球が 51 つ、B には白い球が 50 つ、黒い球が 40 つ入っている。このとき、片方の壺から取り出した球が白い球がであったとき、その壺が A である確率を求めよ。
3. 確率質量関数 $f(x)$ が $f(x) = 1/n \quad (x = 1, 2, \cdots, n)$ であるような確率分布を n 点に関する一様分布という。X, Y が n 点に関する一様分布に従うとき、$X + Y$ の確率質量関数および累積分布関数を求めよ。
4. 確率密度関数 $f(x)$ が $f(x) = \frac{1}{b-a}$ であるような確率分布を区間 $[a, b]$ 上の一様分布という。X が区間 $[0, 1]$ 上の一様分布に従うとき、e^X の確率密度関数および累積分布関数を求めよ。

期待値と母関数

期待値

定義（期待値）

確率変数 X の期待値（または平均値）$E[X]$ は以下で定義される。

- X が離散型の場合（確率質量関数 $p(x)$ をもつ）：

$$E[X] = \sum_{x} x \cdot p(x)$$

- X が連続型の場合（確率密度関数 $f(x)$ をもつ）：

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

- μ や $\bar{X}$ 、あるいは $E(X)$ などとも表記される。
- 線形性：$E[aX + bY] = aE[X] + bE[Y]$ が成り立つ。

分散

定義（分散、標準偏差）

確率変数 X の分散 $V[X]$ 、標準偏差 σ は以下で定義される。

$$V[x] = E[(x-\mu)^2], \qquad \sigma = \sqrt{V[x]}$$

- σ^2 や $V(X)$ などとも表記される。
- $V[x] = E[X^2] - E[X]^2$ が成り立つ。計算が便利。
- 確率変数 X を $\frac{X-\mu}{\sigma}$ に変換する（平均、分散を 1 に変換する）ことを標準化という。

積率

定義（積率/モーメント）

確率変数 X の α まわりの n 次積率は以下で定義される。

$$E[(X-\alpha)^n]$$

- 原点周りの積率を μ_n 、期待値周りの積率を μ'_n と表記することが多い。
- また、$E[(\frac{X-\mu}{\sigma})^n]$ を標準化積率という。
- 3 次標準化積率を歪度、4 次標準化積率を尖度という。

積率母関数

定義（積率母関数）

確率変数 X の原点まわりの n 次モーメントを μ_n とする。このとき、以下で定まる関数 $M_X(t)$ を積率母関数という。

$$M_X(t) = E[(e^{tX})]$$

- 収束性の観点から、必ずしも積率母関数は存在するとは限らない。
- 両側 Laplace 変換の結果により、少なくとも有界なら問題ないことがわかる。
- Taylor 展開を考えれば、$M_X^{(n)}(0) = \mu_n$ となる。
- 積率母関数さえ分かれば、n 次積率が計算可能！

積率母関数の一意性定理

定理（一意性定理）

確率変数 X, Y の積率母関数をそれぞれ $M_X(t), M_Y(t)$ とする。このとき、
$M_X(t) = M_Y(t)$ が成り立つとき、X と Y の分布は等しい。

- 積率は分布の形に関する情報を持っている
- 積率母関数は全ての積率に関する情報を持っているので、分布が一意的に決定できる
- 様々な証明で活躍する

Марковの不等式

定理（Марков）

確率変数 X と、正の定数 a に関して以下の不等式が成り立つ。

$$P(|X| \geq a) \leq \frac{E[|X|]}{a}$$

連続確率変数に対しての証明を与えよう：

$$\begin{aligned} E[|X|] &\geq \int_{-\infty}^{\infty} |x| f(x) dx \\ &\geq \int_{-\infty}^{-a} |x| f(x) dx + \int_{a}^{\infty} |x| f(x) dx \\ &\geq a \left(\int_{-\infty}^{-a} f(x) + \int_{a}^{\infty} f(x) dx \right) \\ &= aP(X \leq -a) + aP(X \geq a) = aP(|X| \geq a) \end{aligned}$$

Чебышёвの不等式

定理（Чебышёв）

確率変数 X の期待値を μ、分散を σ^2 とする。このとき、以下の不等式が成り立つ。

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- $P(|X-\mu| < k\sigma) = 1 - P(|X-\mu| \geq k\sigma) \geq 1 - \frac{1}{k^2}$ が直ちに得られる。
- 期待値と分散さえ分かっていれば適用可能！
- 証明はМарковの不等式において、X を $(X-\mu)^2$ に、a を $(k\sigma)^2$ とすればよい。
- 実際、分散の定義に注意して
  $P(|X-\mu| \geq k\sigma) = P(|X-\mu|^2 \geq (k\sigma)^2) \leq \frac{E[(X-\mu)^2]}{k^2\sigma^2} = \frac{1}{k^2}$ となる。

演習問題

1. 2 個のさいころを投げて、出た目を X, Y で表す。$X - Y$ の期待値、分散を求めよ。
2. 確率変数 X の分散が 0 ならば、X は定数であることを示せ。
3. $E[X] = 10, V[X] = 0.25$ なる確率変数 X について、$9 \leq X \leq 11$ となる確率が 0.75 以上となることを示せ。
4. $E[X] = 10, V[X] = 4$ なる確率変数 X について、$9 \leq X \leq 11$ となる確率に Ч е б ы ш ё в の不等式を適用しても自明な下界しか得られないことを確認せよ。

色々な確率分布

確率分布まとめ

「色々な確率分布を見てみよう！」

離散型確率分布

- 一様分布
- 二項分布
- 超幾何分布
- ポアソン分布

連続型確率分布

- 一様分布
- 正規分布
- 指数分布
- ガンマ分布
- ベータ分布

一様分布（離散型）

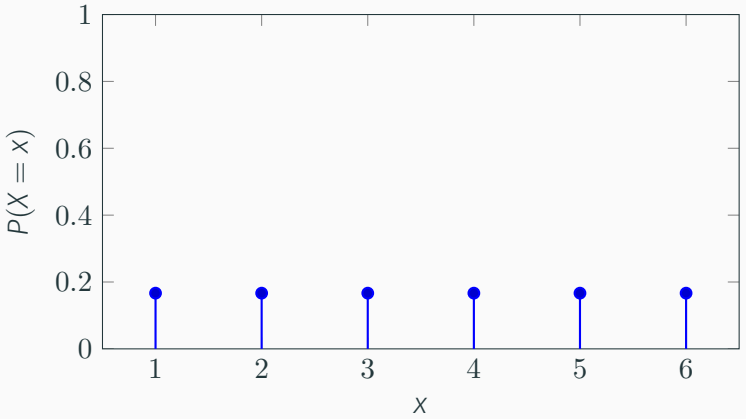
定義（離離散型様分布）

確率質量関数が以下で与えられるような確率分布を一様分布という。

$$f(x) = \frac{1}{n} \quad (x = 1, 2, \cdots, n)$$

- 気持ち：n 個の事象が等確率で起こる場合の分布
- 期待値：$\frac{n+1}{2}$
- 分散：$\frac{n^2-1}{12}$
- 積率母関数：$\frac{e^t - e^{t(n+1)}}{n(1-e^t)}$

離散型一様分布 (1〜6)



二項分布

定義（二項分布）

確率質量関数が以下で与えられるような確率分布を二項分布という。

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} (x = 1, 2, \cdots, n)$$

- 気持ち：確率 p で成功し、$1-p$ で失敗する試行を同じ条件で n 回繰り返したときに、x 回成功する確率
- 表記：$Bi(n, p)$
- 期待値：np
- 分散：$np(1-p)$
- 積率母関数：$(e^t p + 1 - p)^n$

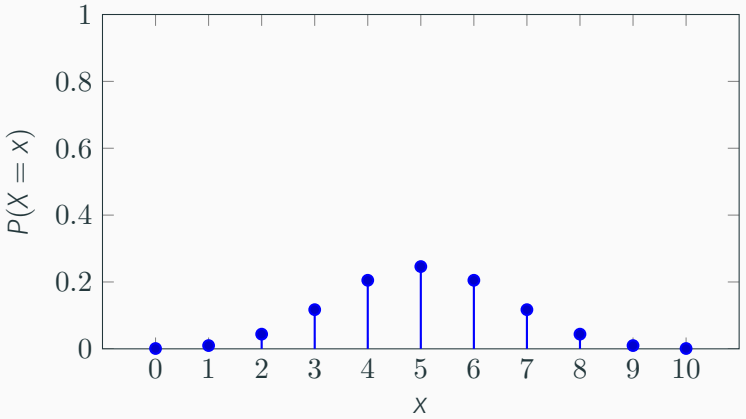
二項分布の期待値

確率 p で成功し、$1-p$ で失敗する試行において、i 回目に成功したら 1 、失敗したら 0 をとるような確率変数を X_i とする。

$$
\begin{aligned}
E[X] &= E[X_1 + X_2 + \cdots + X_n] \\
&= E[X_1] + E[X_2] + \cdots + E[X_n] \\
&= p + p + \cdots + p \\
&= np
\end{aligned}
$$

分散についても同様。また、定義から愚直に計算することや積率母関数から導くこともももちろん可能。積率母関数の導出は、二項定理を素朴に用いるだけ。

二項分布 (n=10, p=0.5)



超幾何分布

定義（超幾何分布）

確率質量関数が以下で与えられるような確率分布を超幾何分布という。

$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} (x = 1, 2, \cdots, n)$$

- 気持ち：M 個の A と $N-M$ 個の B から n 回の非復元抽出をした際に A が x 回得られる確率
- 期待値：$\frac{nM}{N}$
- 分散：$\frac{nM(N-M)(N-n)}{N^2(N-1)}$
- 積率母関数：初等的な表示がない（超幾何級数を用いれば可能）
- 二項分布との関係：超幾何級数において、$N \to \infty$ とし、$\frac{M}{N} \to p$ とすれば、$Bi(n, p)$ に収束する。

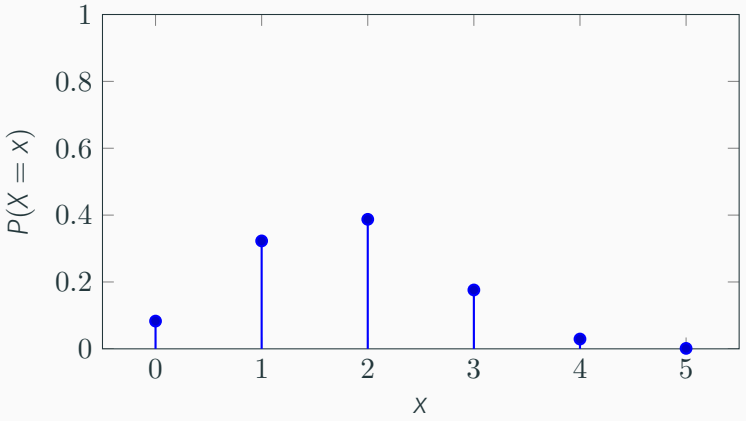
超幾何級数の期待値

i 回目に選んだものが A なら 1 、B なら 0 をとるような確率変数を X_i とする。

$$
\begin{aligned}
E[X] &= E[X_1 + X_2 + \cdots + X_n] \\
&= E[X_1] + E[X_2] + \cdots + E[X_n] \\
&= \frac{M}{N} + \frac{M}{N} + \cdots + \frac{M}{N} \\
&= \frac{nM}{N}
\end{aligned}
$$

独立でなくても、期待値の線型性は利用できる。

超幾何分布 (N=20, M=7, n=5)



捕獲再捕獲法

問題設定：捕獲再捕獲法

ある個体群から無作為に M 体を捕獲し、標識を付けて戻す。その後、無作為に n 体を再び捕獲したところ、x 体に標識が付いていた。このとき、個体群の個体数 N はどのように推定されるだろうか？

- x は (N, M, n) の超幾何分布に従う。
- ここから、超幾何分布と照らし合わせていくことで N を逆算することが可能になる。
- 雑な近似をするのなら、$N = \frac{Mn}{x}$ となる。

ポアソン分布

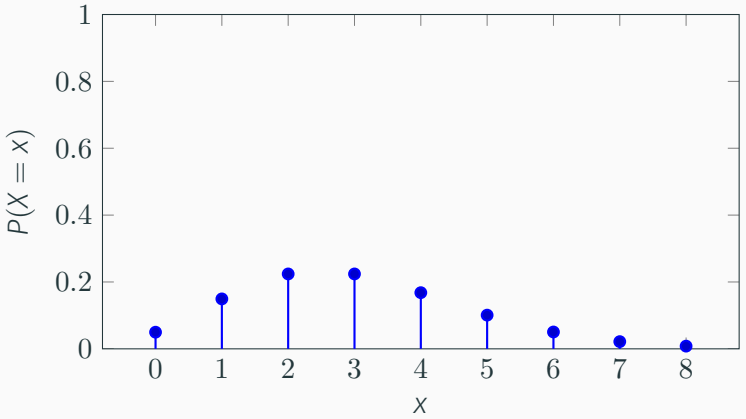
定義（ポアソン分布）

確率質量関数が以下で与えられるような確率分布をポアソン分布という。

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

- 気持ち：単位時間あたりの発生期待値が λ 回であるような現象が、単位時間に x 回発生する確率
- 表記：$Po(\lambda)$
- 期待値：λ
- 分散：λ
- 積率母関数：$e^{\lambda(e^t-1)}$
- 二項分布との関係：$\lambda = np$ を保ったまま、$n \to \infty,\ p \to 0$ の $Bi(n,p)$ を考えると、ポアソン分布に収束する。

ポアソン分布 ($\lambda = 3$)



一様分布（連続型）

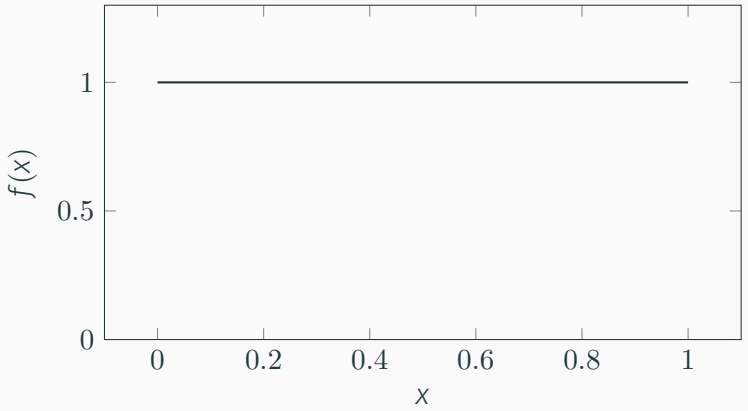
定義（連続型一様分布）

確率密度関数が以下で与えられるような確率分布を区間 $[a, b]$ 上の一様分布という。

$$f(x) = \frac{1}{b-a}$$

- 気持ち：区間 $[a, b]$ 上から一様に点を選ぶ場合の確率
- 期待値：$\frac{a+b}{2}$
- 分散：$\frac{(a-b)^2}{12}$
- 積率母関数：$\frac{e^{tb}-e^{ta}}{(b-a)t}$

連続型一様分布 (0 – 1)



正規分布

定義（正規分布）

確率密度関数が以下で与えられるような確率分布を正規分布という。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 気持ち：独立同分布の確率の和が従う分布の吸引的不動点（中心極限定理）
- 表記：$N(\mu, \sigma^2)$
- 期待値：μ
- 分散：σ
- 積率母関数：$e^{\mu t + \frac{\sigma^2 t^2}{2}}$

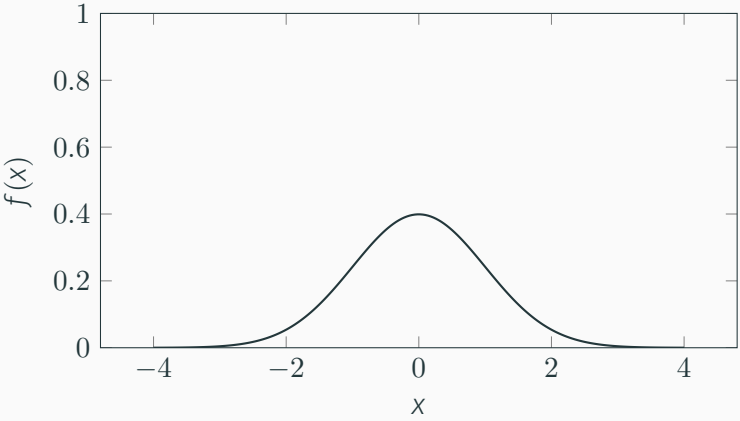
「正規分布の基本の式は e^{-x^2} である」

$$e^{-x^2} \xrightarrow{\text{正規化}} \frac{1}{\int_{\infty}^{\infty} e^{-x^2}} e^{-x^2} = \frac{1}{\sqrt{\pi}} e^{-x^2} \xrightarrow{\text{平均を}\mu\text{に}} \frac{1}{\sqrt{\pi}} e^{-(x-\mu)^2} \xrightarrow{\text{分散を}\sigma^2\text{に}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

正規分布において、以下の値は覚えておくと便利である。

- $P(|X-\mu| \leq \sigma) \simeq 0.683$
- $P(|X-\mu| \leq 2\sigma) \simeq 0.954$
- $P(|X-\mu| \leq 3\sigma) \simeq 0.997$

正規分布 ($\mu = 0,\ \sigma = 1$)



指数分布

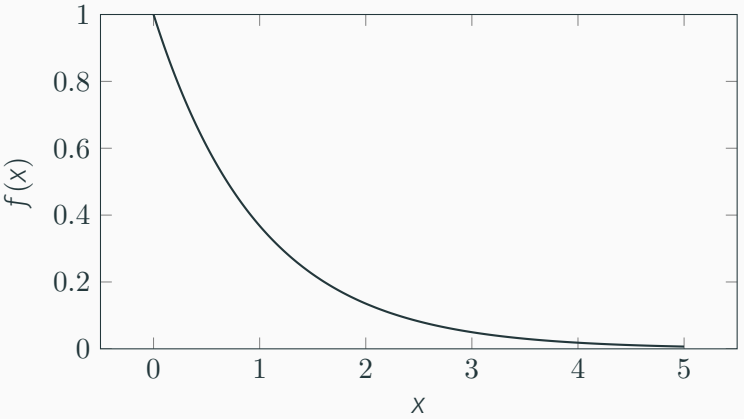
定義（指数分布）

確率密度関数が以下で与えられるような確率分布を指数分布という。

$$f(x) = \lambda e^{-\lambda x}$$

- 気持ち：単位時間当たりの期待発生回数が λ 回の事象が発生間隔 x で発生する確率
- 表記：$Ex(\lambda)$
- 期待値：$\frac{1}{\lambda}$
- 分散：$\frac{1}{\lambda^2}$
- 積率母関数：$(1 - \frac{t}{\lambda})^{-1}$

指数分布 ($\lambda = 1$)



分布表の見方

- 離散型の確率分布の表：一般的には確率質量関数
- 連続型の確率分布の表：一般的には累積分布関数

例えば、以下の（かなり雑な）標準正規分布表は標準正規分布に従う確率変数 X について、$P(X \leq z)$ を表している。

標準正規分布 ($\mu = 0, \sigma = 1$)

z	-1.0	-0.5	0.0	0.5	1.0
$\Phi(z)$	0.1587	0.3085	0.5000	0.6915	0.8413

演習問題

以下の問題を解くにあたって、次頁の表を利用しても構わない。

1. とあるコールセンターでは 1 時間あたり平均 3 回電話が鳴る。このとき、1 時間に 4 回以上電話が鳴る確率を求めよ。また、ある時刻で電話が鳴ってから、2 時間ずっと電話がならない確率を求めよ。
2. 13 個の赤玉、7 個の白玉が入った袋がある。この袋から 5 個の球を袋に戻さずに連続して取りだすとき、取り出した球のうち 4 個が赤球となる確率を求めよ。
3. λ に関するポアソン分布の確率質量関数 f と指数分布の累積分布関数 F について、$f(0) + F(1)$ の値は何か。
4. あるテストを受験した生徒 A の偏差値が 60 であった。受験した生徒の得点が正規分布に従うと仮定したとき、生徒 A は上位何%に位置すると考えられるか？

確率分布表まとめ

二項分布 (n=10, p=0.5)

x	0	1	2	3	4	5	6	7	8	9	10
$P(X=x)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

超幾何分布 (N=20, K=7, n=5)

x	0	1	2	3	4	5
$P(X=x)$	0.083	0.323	0.387	0.176	0.029	0.001

ポアソン分布 ($\lambda = 3$)

x	0	1	2	3	4	5
$P(X=x)$	0.050	0.149	0.224	0.224	0.168	0.101

指数分布 ($\lambda = 3$)

x	0	1	2	3	4	5
$F(x)$	0.000	0.950	0.998	0.9999	1.000	1.000

標準正規分布 ($\mu = 0, \sigma = 1$)

z	-1.0	-0.5	0.0	0.5	1.0
$\Phi(z)$	0.1587	0.3085	0.5000	0.6915	0.8413

複数の確率分布

中心極限定理への準備

まずは、以下の概念を理解しよう：

- 確率分布の畳み込み
- 複数の確率変数の取扱い

確率変数の和の分布

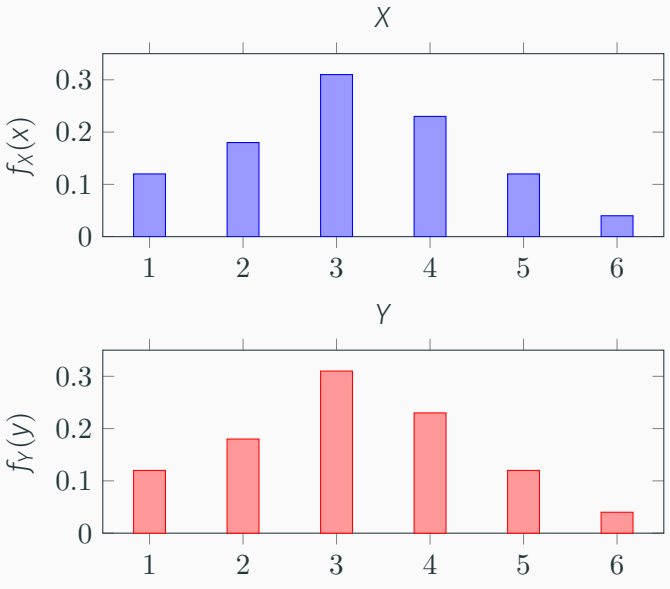
定理（確率変数の和の分布 - 離散型）

確率変数 X, Y がそれぞれ確率質量関数 $f(x), g(x)$ で定まる分布に従うとき、$X + Y$ の確率分布は以下で与えられる。

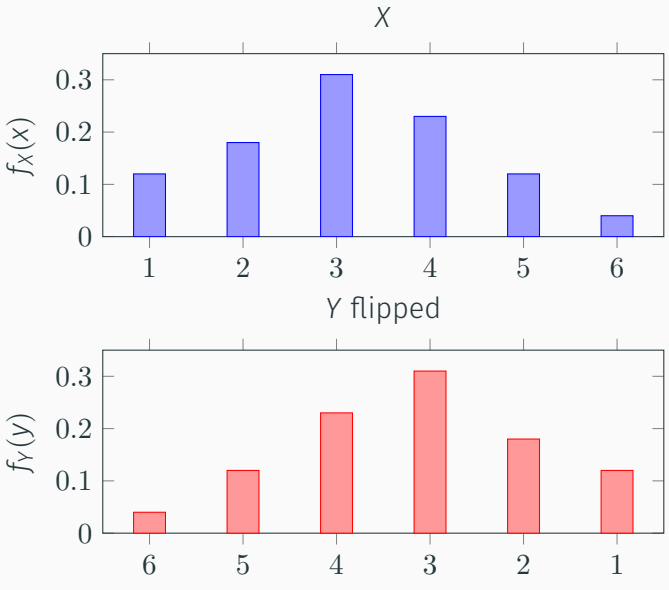
$$h(z) = \sum_x f(x)g(z - x)$$

- この操作を畳み込みという。
- 確率分布の表をひっくり返して、足しあわせるイメージ。（を参照）
- 次頁で不均一な（立方体）サイコロの和が 7 になるケースを考えよう。

Step 1: 分布図を上下に並べる

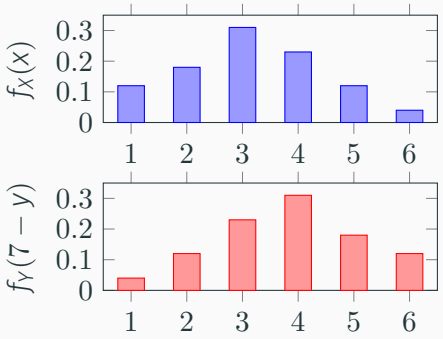


Step 2: 下段だけ反転

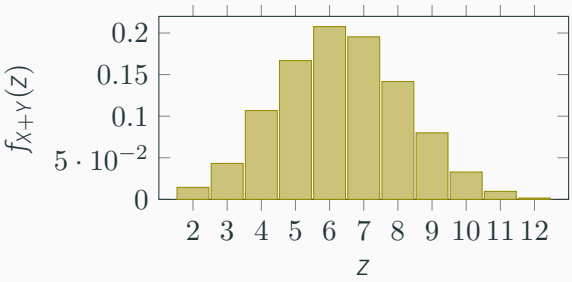


Step 3: これらの積の和をとれば、右のグラフの 7 の値が得られる

左：Step 2 の分布図



右：$X+Y$ の分布 $f_{X+Y}(z)$



確率変数の和の分布

定理（確率変数の和の分布 - 連続型）

確率変数 X, Y がそれぞれ確率密度関数 $f(x), g(x)$ で定まる分布に従うとき、$X + Y$ の確率分布は以下で与えられる。

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z - x)$$

- この操作も畳み込みという。$f * g$ などと表記される。
- 確率分布をひっくり返して足し合わせることは変わらず行われる
- さっきの分布図を確率密度関数のグラフに変えればよい

様々な確率分布の畳み込み

いくらかの確率分布には再生性が確認される[2]。下二つは二項分布の極限であったことを思い出せば、再生性の本質は二項分布にあると考えられる。

- $Bi(n,p) * Bi(m,p) = Bi(n+m,p)$
- $Po(\lambda) * Po(\mu) = Po(\lambda+\mu)$
- $N(\mu_1,\sigma_1^2) * N(\mu_2,\sigma_2^2) = N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$

[2] もちろん、再生性を持たないような確率分布がほとんどである。

同時確率分布

定義（同時分布）

2 つ以上の確率変数 (X, Y) の同時確率質量関数 または同時確率密度関数 をそれぞれ以下で定める。

$$p_{X,Y}(x, y) = P(X = x,\, Y = y), \quad f_{X,Y}(x, y) \quad \text{(連続型)}$$

- 正規化条件：
  - 離散型：$\displaystyle\sum_{x}\sum_{y} p_{X,Y}(x, y) = 1$
  - 連続型：$\displaystyle\iint_{\mathbb{R}^2} f_{X,Y}(x, y)\, dx\, dy = 1$
- 確率の取得：任意の集合 $A \subseteq \mathbb{R}^2$ に対し

$$P\big((X, Y) \in A\big) = \sum_{(x,y)\in A} p_{X,Y}(x, y) \quad \text{または} \quad \iint_A f_{X,Y}(x, y)\, dx\, dy.$$

周辺確率分布

定義（周辺分布）

同時分布から 1 変数だけの分布を取り出したものを周辺分布関数といい、以下で定める。

$$p_X(x) = \sum_y p_{X,Y}(x,y), \quad f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

で得られる。

- 同様に Y の周辺分布は

$$p_Y(y) = \sum_x p_{X,Y}(x,y), \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

- 周辺化によって他の変数の「影響」を無視できる。

共分散と相関係数

定義（共分散）

2 つの確率変数 X, Y の共分散を以下で定める。

$$C[X, Y] = E\big[(X - E[X])(Y - E[Y])\big]$$

定義（相関係数）

2 つの確率変数 X, Y の相関係数を以下で定める。

$$\rho_{X,Y} = \frac{C[X, Y]}{\sqrt{V[X]\, V[Y]}}, \quad -1 \leq \rho_{X,Y} \leq 1$$

- $C[X, Y] > 0$ は正の共変動、$C[X, Y] < 0$ は負の共変動を示す。
- 相関係数は単位に依存せず線形関係の強さを表す。

分散共分散行列

定義（分散共分散行列）

ベクトル確率変数 $\mathbf{X} = (X_1, \ldots, X_n)^\top$ について、以下を分散共分散行列という。

$$\Sigma = V[\mathbf{X}] = \Big[C[X_i, X_j]\Big]_{i,j=1}^{n}$$

- 対称行列：$\Sigma = \Sigma^\top$
- 半正定値：任意の非零列ベクトル $\mathbf{a}$ に対し $\mathbf{a}^\top \Sigma \mathbf{a} \geq 0$
- 対角要素が各変数の分散を表す：$\Sigma_{ii} = V[X_i]$

確率変数の独立性

定義（独立性）

確率変数 X, Y が独立とは、任意の x, y に対し以下が成り立つことをいう。

$$P(X \le x,\ Y \le y) = P(X \le x)\, P(Y \le y)$$

- $p_{X,Y}(x, y) = p_X(x) p_Y(y)$ や $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ と同値
- 独立ならば $C[X, Y] = 0$ となるが，逆は一般には成り立たない。
- 複数変数の場合は「すべての組み合わせで同様に分解できる」ことが必要。

演習問題

1. 離散型確率変数 X, Y の同時確率質量関数が

$$p_{X,Y}(x,y) = \begin{cases} \dfrac{x+y}{12}, & x \in \{1,2\},\ y \in \{1,2\}, \\ 0, & \text{それ以外} \end{cases}$$

のとき、周辺分布 $p_X(x)$, $p_Y(y)$ 、$P(X = 2 \text{かつ} Y \leq 1)$。また、$X$ と Y が独立か判定せよ。

2. 連続型確率変数 (X, Y) の同時確率密度関数が

$$f_{X,Y}(x,y) = \begin{cases} 2(x+y), & 0 \leq x \leq 1,\ 0 \leq y \leq 1, \\ 0, & \text{それ以外} \end{cases}$$

のとき、周辺密度 $f_X(x)$, $f_Y(y)$ 、期待値 $E[X]$, $E[Y]$ 、共分散 $\mathrm{Cov}(X,Y)$, 相関係数 $\rho_{X,Y}$ を求めよ。

演習問題

3. X, Y が区間 $[0, 1]$ 上の一様分布に従うとき、$X + Y$ の確率密度関数および累積分布関数を求めよ。
4. 確率変数 (X_1, X_2, X_3) の分散共分散行列

$$\Sigma = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 9 & 2 \\ 0 & 2 & 16 \end{pmatrix}$$

   が与えられているとき、$V[X_1 + X_2]$ と $V[2X_1 - X_3]$ を求めよ。
5. 共分散がゼロかつ独立でない例を示せ。

中心極限定理

大数の弱法則

定理（大数の弱法則）

独立で同じ分布に従う確率変数[3] $X_1, X_2, \cdots, X_n$ に対して、その分散が有限ならば任意の ε に対して

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \leq \varepsilon\right) = 1$$

- 証明は、Ч е б ы ш ё в の不等式を用いれば良い。
- 十分に大きい標本数を取れば、その平均値を真の平均値とみなして良いことを主張している。

[3] しばしば独立同分布や、i.i.d. などと表記される。

大数の強法則

定理（大数の強法則）

独立同分布な確率変数 $X_1, X_2, \cdots, X_n$ に対して、その平均値が存在すれば、

$$P\left(\lim_{n\to\infty}\left|\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right|=0\right)=1$$

- 証明は専門的であるので、省略。（別で PDF を作成するかも）
- 十分に大きい標本数を取れば、その平均値を真の平均値とみなして良いことを主張しているという点では弱法則と一緒。

強法則と弱法則の違い

強法則には誤差 ε が出てこないので、より強いと考えられる。

- 弱法則：
  十分大きな n で「平均 $\bar{X}$ と母平均 μ の差が ε 未満になる確率」が 1 に収束する
- 強法則：
  「十分大きな n で平均 $\bar{X}$ と母平均 μ の差が 0 になる確率」が 1 に収束する。

上のような収束を確率収束といい、下のような収束を概収束という。

確率分布の収束

定義（分布収束、確率分布、概収束）

- 分布収束 (in distribution)：確率変数列 $\{X_n\}$ が X に分布収束するとは、

$$X_n \xrightarrow{d} X \iff \lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad (\text{累積分布関数} F_X \text{ の連続点で})$$

- 確率収束 (in probability)：確率変数列 $\{X_n\}$ が X に確率収束するとは、

$$X_n \xrightarrow{P} X \iff \forall \varepsilon > 0,\ \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$$

- 概収束 (almost sure)：確率変数列 $\{X_n\}$ が X に概収束するとは、

$$X_n \xrightarrow{a.s.} X \iff P\left(\left\{\omega \middle| \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

より一般的な表記を採用したが、先ほどと言っていることは一緒である。

中心極限定理

定理（中心極限定理）

独立同分布な確率変数 $X_1, X_2, \cdots, X_n$ に対して、$S_n = X_1 + X_2 + \cdots + X_n$ と定める。平均値 $E[X_n] = \mu$ と分散 $V[X_n] = \sigma^2$ が有限ならば以下で与えられる Z_n は標準正規分布に分布収束する。

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

- 大雑把にいえば、独立同分布な確率変数を足し合わせたら、$N(n\mu, n\sigma^2)$ に収束することを主張している。
- 畳み込みという演算において、正規分布が吸引的不動点であることを意味している。
- この定理によって、いたるところに正規分布（に近似される分布）が登場するのである。

中心極限定理の証明（積率母関数を用いる方法）

証明を与えよう[4]

- 収束定理[5]より、$M_Z(t)$ が $e^{\frac{t^2}{2}}$ に各点収束することを示せばよい。
- すなわち、各 t に対して、$\lim_{n\to\infty} M_Z(t) = e^{\frac{t^2}{2}}$ を示せばよい。
- $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ に対する積率母関数

$$M_{Z_n}(t) = E[e^{tZ_n}] = e^{-tn\mu/(\sigma\sqrt{n})}\left(M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n.$$

[4]積率母関数は全ての確率分布に対して定義されるわけではないので、本来は $\varphi_X(t) = E[e^{itX}]$ で定義される特性関数を用いる必要がある。

[5]詳細は附録に。

中心極限定理の証明（積率母関数を用いる方法）

- $M_X(u)$ のテイラー展開（$u \to 0$）

$$M_X(u) = 1 + \mu u + \frac{\sigma^2}{2} u^2 + o(u^2).$$

- よって

$$\begin{aligned}
\log M_{Z_n}(t) &= -\frac{t\sqrt{n}\mu}{\sigma} + n \log\left(1 + \mu \frac{t}{\sigma\sqrt{n}} + \frac{\sigma^2}{2} \frac{t^2}{\sigma^2 n} + o(\tfrac{1}{n})\right) \\
&= -\frac{t\sqrt{n}\mu}{\sigma} + n\left(\mu \frac{t}{\sigma\sqrt{n}} + \frac{t^2}{2n} + o(\tfrac{1}{n})\right) \\
&= \frac{t^2}{2} + n\, o(\tfrac{1}{n}) \xrightarrow[n\to\infty]{} \frac{t^2}{2}.
\end{aligned}$$

- 以上より、

$$M_{Z_n}(t) \to e^{t^2/2}$$

演習問題

1. 表、裏が出る確率がそれぞれ $\frac{1}{2}, \frac{1}{2}$ のコインを 10000 回投げて、表が 4900 回以上 5100 回以下の確率を求めよ。
2. 表、裏が出る確率がそれぞれ $\frac{9}{10}, \frac{1}{10}$ のコインを 10000 回投げて、表が 8010 回以上 9090 回以下の確率を求めよ。
3. 大数の法則を用いて、以下を説明せよ。

$$\lim_{n\to\infty} \frac{1}{2^n} \int_{[-1,1]^n} \frac{e^{x_1} + e^{x_2} + \cdots + e^{x_n}}{x_1^4 + x_2^4 + \cdots + x_n^4} dx_1 dx_2 \cdots dx_n = \frac{5}{2}(e - e^{-1})$$

4. $\lambda = 1$ の指数分布において、中心極限定理を用いて以下を説明せよ。
（γ 分布に関する知識が必要）

$$\lim_{n\to\infty} \frac{1}{n!} \int_0^n x^n e^{-x} dx = \frac{1}{2}$$

正規分布表 (累積分布関数 $\Phi(z)$)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621

統計学の基礎

データとその種類

定義（データ）- JIS X 0001「情報処理用語一基本用語」による

情報の表現であって，伝達，解釈又は処理に適するように形式化され，再度情報として解釈できるもの。

データには種類があり、上二つは質的データ、下二つは量的データと呼ばれている。

- 名義尺度：ラベル付けのみ
- 順序尺度：順序が定義できるような尺度
- 間隔尺度：和と差に意味があるような尺度
- 比尺度：比にも意味があるような尺度

基本的な統計量

基本的な統計量として、平均、中央値、最頻値、分散などが挙げられる。このうち、平均や分散に関しては以前に記述した通りである。

定義（中央値、最頻値）

データを小さい順に並べた際の中央の値を中央値という。データの個数が偶数の場合には中央二つの値の平均を指す。また、度数が最大となるような階級の階級値を最頻値という。

- 階級は、単にデータを幾らかの区分に分けたものである。
- 一般的には n 個のデータに対して、$1 + \log_2 n$ 個の階級を用意する。

相関係数

Pearson の相関係数に関しては、以前に定義した通り。

定義（Spearman の相関係数）

データ x_i, y_i に対して、その順位を R_{xi}, R_{yi} とする。このとき、Spearman の相関係数 r_s は以下で定まる。

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (R_{xi} - R_{yi})^2$$

- なんで 1 から引くの？ → 順位が全部一致するときに相関係数が 1 になるように
- なんで $\frac{6}{n^3-n}$ がかかってるの？ → 相関係数が -1 から 1 になるようにスケールしてる。つまり、$\sum_{i=1}^{n}(R_{xi} - R_{yi})^2$ の最大値。
- 順序尺度に関しても計算可能

偏相関係数

定義（偏相関係数）

データ x, y, z に関して、データ z を第三の変数とした偏相関係数を以下で定める。

$$r_{x,y\cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

- 3 つ以上の変数があるときに、第三の変数の影響を除いた相関係数
- 疑似相関などの判定に用いることができる

回帰

アルゴリズム（最小二乗法）

データ x, y が与えられたときに、以下を最小にするような A, B を用いて $y = Ax + B$ と近似する手法を最小二乗法という。

$$\sum_{i=1}^{n} (y_i - (Ax_i + B))^2$$

- $A = \frac{C[X,Y]}{V[X]}$, $B = \mu_Y - A\mu_X$ となる。
- 導出は、A, B に関して偏微分するなどして愚直に。
- 個人的には、以下の形が一番覚えやすい：

$$\frac{Y - \mu_Y}{X - \mu_X} = \frac{\sigma_Y}{\sigma_X}$$

演習問題

覚えてますか〜？？

1. スピアマンの相関係数の式を述べよ
2. 最小二乗法の式を述べよ
3. データの種類を 4 つ挙げよ

統計的推定

母集団と標本

定義（母集団、標本）

分析したい集団を母集団といい、母集団から分析のために選び出された要素を標本という。

- 標本から母集団についての推測を行うことを統計的推測という
- 母集団が何らかの確率分布に従い、かつその分布がパラメーターに従うとき[6]、それを母数という。
- 標本から求めた統計量を推定量という。母数 θ に対して推定量を $\hat{\theta}$ とあらわす。推定量の具体的な値を推定値という。
- 例えば、母平均 θ の推定量 $\hat{\theta}$ に関する推定値は 2.13 であるなどと表現する。

[6]例えば、ポアソン分布なら λ などがパラメーターに該当する

推定の種類

「統計的推定には、大きく二つの種類がある」

点推定：母数をピンポイントで推定

- 母平均 θ として尤もらしいのはこの値だ！
- モーメント法、最尤法など

区間推定：母数が（確率 p で）入っているであろう区間を推定

- この区間は 90%の確率で母平均 θ を含んでいるだろう！
- 正規分布や t 分布を用いる方法が主流

積率法

アルゴリズム（積率法）

母数の k 次積率 μ_k を標本の k 次積率 $\hat{\mu}_k$ に置き換えて推定を行う方法

- 単に、母数の積率を推定量で代用するというだけ
- 最も単純な方法のひとつ
- ここから、積率母関数の一意性定理より、尤もらしい分布も推定できる
- 例えば、正規分布なら平均と分散をそれぞれ標本平均・標本分散で推定

最尤法

定義（尤度）

標本 $\{x_1, \ldots, x_n\}$ が得られたとき、そのデータの尤度を以下で定める。

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) \quad \left(\text{離散分布なら確率質量関数, 連続分布なら確率密度関数}\right)$$

- ある標本があったときに、母数 θ に対してどれくらい尤もらしい結果かを示す値のこと
- 例えば、表が出る確率が θ であるコインを 100 回投げた場合に 55 回表が出た。このときの尤度 $L(\theta)$ は $L(\theta) = \theta^{55}(1-\theta)^{45}$ となる。[7]

[7] ところで、これは定数倍すれば、$\binom{55}{100}\theta^{55}(1-\theta)^{45}$ となる。

アルゴリズム（最尤法）

尤度 $L(\theta)$ を最大化するような母数 $\hat{\theta}_{\mathrm{MLE}}$ を推定量とする：

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta).$$

- 通常は対数を取った 対数尤度 $\ell(\theta) = \log L(\theta)$ を最大化する
- これは単に計算を簡単にするための工夫である
- 方程式 $\frac{\partial}{\partial\theta}\ell(\theta) = 0$ を解く
- 一般に解が存在、一意とは限らないので注意

推定量の評価

推定量の評価には、主に三つの尺度が存在する。不偏性、一致性、有効性の 3 つである。厳密な定義は次頁で述べることとして、まずはお気持ちを述べよう。

- 不偏性：推定量の平均値が真の母数と一致する
- 一致性：標本サイズを増やすと推定量が真の母数に近づく
- 有効性：不偏推定量の中で分散が最小、または漸近的に最小分散を達成する

不偏性

定義（不偏性）

推定量 $\hat{\theta}_n$ が母数 θ に対して不偏であるとは、以下が成り立つことをいう。

$$E[\hat{\theta}_n] = \theta$$

- 母分散を推定する際に、通常の分散を用いると不偏性が達成されない。
- そこで、不偏分散 $\frac{n}{n-1}\sigma^2$ を採用する。

一致性

定義（一致性）

推定量 $\hat{\theta}_n$ が母数 θ に対して一致性を持つとは、以下が成り立つことをいう。

$$\hat{\theta}_n \xrightarrow{p} \theta \quad (n \to \infty)$$

- 上の収束は、確率収束であった。
- 独立同分布な標本ならば大数の法則によって一致性は保証されている。

有効性

定義（有効性）

推定量 $\hat{\theta_n}$ が有効性を持つとは、推定量 $\hat{\theta}_n$ の分散が不偏推定量の中で最小であるか、漸近的にはクレイマー–ラオ下限を達成する場合をいう。

$$\mathrm{Var}(\hat{\theta}_n) \geq \frac{1}{I(\theta)}, \quad \text{等号成立なら有効推定量}$$

ここで $I(\theta)$ はフィッシャー情報量である.

- 不偏性、一致性を備えたような 2 つの推定量の比較に用いる尺度
- 分散が小さいような推定量を有効と見做す

区間推定

アルゴリズム（区間推定）

1. 推定したい母数 θ の点推定量 $\hat{\theta}$ を決める。
2. $\hat{\theta}$ の標本分布を近似し、標準誤差 $\mathrm{SE}(\hat{\theta}) = \frac{\sigma}{\sqrt{n}}$ を求める。
3. 信頼水準 $1-\alpha$ を設定し、対応する分位点を取得する。
   （例えば正規分布なら $z_{1-\alpha/2}$）
4. 信頼区間を以下の形で構成する。

$$\left[\hat{\theta} - z_{1-\alpha/2}\,\mathrm{SE}(\hat{\theta})\,,\ \hat{\theta} + z_{1-\alpha/2}\,\mathrm{SE}(\hat{\theta})\right]$$

- 点推定量の不確実性を区間で表現する
- 「母数がこの区間に入る確率が $1-\alpha$」ではなく、「繰り返し抽出で区間の $1-\alpha$ が真の値を含む」
- 分布や母数の既知・未知に応じて，正規分布や t 分布を用いる

正規分布にしたがう母集団の例

正規分布にしたがう母集団の母平均 μ の区間推定を実際に行う：

- 母分散 σ^2 が既知の場合、標本平均 $\bar{X}$ は $N(\mu,\ \sigma^2/n)$ に従う。
- 信頼区間は

$$\left[\bar{X} - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\,,\ \bar{X} + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right]$$

  として構成する。
- 例：$n = 25$，$\bar{X} = 10$，$\sigma = 2$，信頼水準 95%なら $\bar{X} \pm 1.96\frac{2}{5} = [9.216, 10.784]$

カイ二乗分布

定義（カイ二乗分布）

自由度 ν のカイ二乗分布は、独立な標準正規変数 $Z_1, \dots, Z_\nu \sim N(0,1)$ を用いて

$$X = \sum_{i=1}^{\nu} Z_i^2$$

と定義される分布である。密度関数は

$$f_\nu(x) = \frac{1}{2^{\nu/2}\,\Gamma(\nu/2)} x^{\frac{\nu}{2}-1} e^{-x/2}, \quad x > 0$$

であり、自由度が小さいほど裾が厚く、自由度が大きくなると正規分布に近づく。

- 標本分散を用いた検定や区間推定で登場する。
- t 分布や F 分布の構成要素としても重要。

t 分布

一般に、母分散は未知である。そこで、t 分布を用いる

定義（t 分布）

自由度 ν の t 分布は、独立な標準正規分布 $Z \sim N(0,1)$ と自由度 ν の χ^2 分布 $U \sim \chi^2_\nu$ を用いて

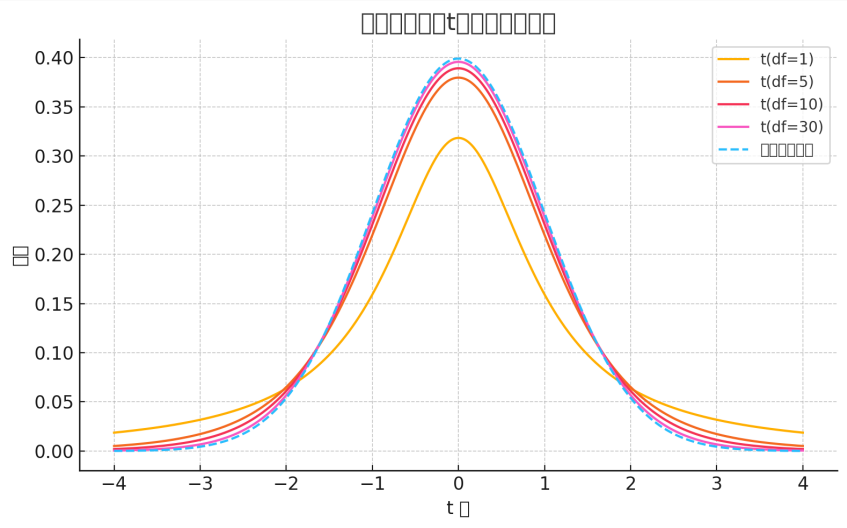
$$T = \frac{Z}{\sqrt{U/\nu}}$$

と定義される分布である。密度関数は

$$f_\nu(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

であり、裾が正規分布より厚い形状を持つ

t 分布の形状


□□□□□□□t□□□□□□□□
t(df=1)
t(df=5)
t(df=10)
t(df=30)
□□□□□□□
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
□□
−4
−3
−2
−1
0
1
2
3
4
t □

t 分布を用いた区間推定の例

t 分布を用いた区間推定を実際に行う：

- 母分散未知の正規母集団について、標本平均 $\bar{X}$ と標本分散 S^2 を求める。
- 統計量 $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ は自由度 $n-1$ の t 分布に従う。
- 信頼区間は

$$\left[\bar{X} - t_{n-1,\,1-\alpha/2}\,\frac{S}{\sqrt{n}},\ \bar{X} + t_{n-1,\,1-\alpha/2}\,\frac{S}{\sqrt{n}}\right]$$

  として構成する。
- 例：$n = 16$，$\bar{X} = 5$，$S = 1.5$，95%信頼区間では $\bar{X} \pm 2.12\frac{1.5}{4} = [4.205, 5.795]$。

小サンプルが不安定な理由

- データ点が少ないと、たまたま取れた極端な値の影響を受けやすい
- 標本分散 S^2 の自由度が小さいほど、χ^2 分布上での変動（推定ノイズ）が大きくなる
- 結果として、標準化統計量 $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ の分布の裾が厚くなり、正規分布では近似できない

t 分布による補正イメージ

- 分母に未知の S を用いることで増えた「ばらつき」を裾の厚い分布でカバー
- 自由度 $n-1$ が小さいほど裾が厚くなり、小サンプルの不確実性を反映
- 自由度が増えると裾は細くなり、最終的には標準正規分布に収束
- これにより、有限サンプルでも指定した信頼水準を正しく保てる

なぜ t 分布を用いるのか？

- 母分散未知＋小サンプルでの「標本分散のぶれ」を無視すると、信頼区間のカバー率が落ちる
- t 分布の厚い裾は、そのぶれ分の余裕を自然に広い区間として反映
- 信頼区間を $\left[\bar{X} \pm t_{n-1,1-\alpha/2}\,\frac{S}{\sqrt{n}}\right]$ とすることで、真のカバー率を維持
- 大サンプル時は $t_{n-1,1-\alpha/2} \approx z_{1-\alpha/2}$ となり、正規近似に移行

統計的検定

仮説検定

「仮説検定：データからある仮説の正しさを判断する方法」

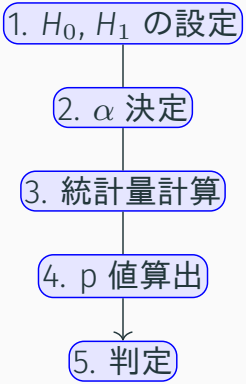
- 仮説を仮定する
- その仮説のもとで、データの事象はどの程度の確率で起こっているのか計算する
- その確率が珍しいものかを判定する

仮説検定の論理

「仮説検定は間違いを示すことしかできない」

- 反証したい事柄を帰無仮説、帰無仮説の否定を対立仮説という。
- 帰無仮説が正しいとして、その現象が極端すぎるかどうか検証する。
- 極端すぎれば、帰無仮説が否定され、すなわち対立仮説が採択される。
- 極端すぎなければ何も言えない。
- 「極端すぎる」ということばを「矛盾」と読み替えると背理法になる

仮説検定の手順



仮説検定の具体例

問題：コインの例

コインを 100 回投げたときに 63 回表が出た。このコインの表が出る確率は 0.5 であるか？

- 結論は「コインは公平でない」or「コインは公平か公平でないか分からない」
- 有意水準を 0.05(5%) とする。
- 帰無仮説 H_0：表の確率が 0.5 である。対立仮説 H_1：表の確率が 0.5 より大きい。
- 帰無仮説が正しいと仮定する。このとき、コインが 63 回以上出る確率、すなわち p 値は 0.6%である。
- これは、(事前に決めておいた) 有意水準よりも明らかに小さいので珍しすぎる現象である。
- よって、コインは公平でない。

仮説検定

問題：コインの例

コインを 100 回投げたときに 63 回表が出た。このコインの表が出る確率は 0.5 であるか？

- 有意水準が 0.005 としていた場合、帰無仮説は棄却されない。
- つまり、有意水準の決め方次第で結果が変わる。
- 仮説検定は有意水準の決め方という主観的な判断に依存する。
- それゆえに、当然事前に設定しておかなければならない。

片側検定と両側検定

片側検定

- 帰無仮説 H_0: $\theta = \theta_0$
- 対立仮説 H_1: $\theta > \theta_0$ または $\theta < \theta_0$
- p 値: 片側の極端値累積確率
- 例: コインの例で「表の確率 > 0.5」の検定

両側検定

- 帰無仮説 H_0: $\theta = \theta_0$
- 対立仮説 H_1: $\theta \neq \theta_0$
- p 値: 両側の極端値累積確率の和
- 例: コインの例で「表の確率 $\neq 0.5$」の検定

第一種の誤り (Type I Error)

- 帰無仮説が真であるのに棄却してしまう誤り
- 発生確率: 有意水準 α（例: 0.05）
- 例: 公平なコインなのに「公平でない」と判断
- 被る不利益: 偽陽性、不必要な追加検証、リソース浪費

第二種の誤り (Type II Error)

- 帰無仮説が偽であるのに棄却できない誤り
- 発生確率: β (検出力は $1-\beta$)
- 例: 不公平なコインなのに「公平かどうか分からない」と判断
- 被る不利益: 偽陰性、真の効果を見落としてしまう
- 第一種の誤りの発生確率とはトレードオフの関係

検定の種類

- パラメトリック検定：母集団がどんな分布に従うか分かっているような検定。特に、正規母集団を前提としている。
- ノンパラメトリック検定：母集団が従う分布が不明な検定

二項検定 (Binomial Test)

定義（二項検定）

二項分布を用い、成功確率 p が仮定値 p_0 と等しいかを検定する方法。

- 帰無仮説: $H_0 : p = p_0$ ／ 対立仮説: $H_1 : p \neq p_0$（片側／両側）
- 統計量: 観測成功回数 k の確率分布
- p 値: 帰無仮説が正しい場合に観測した成功回数 k 以上（または以下）が起こる確率
- 例: コイン投げ $n = 20$ 回で表が $k = 15$ 回以上出る確率

z 検定 (Z-test)

定義（z 検定）

母分散 σ^2 が既知の場合に母平均 μ を検定する方法。

- 帰無仮説: $H_0 : \mu = \mu_0$ ／ 対立仮説: $H_1 : \mu \neq \mu_0$
- 統計量: $Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1)$
- p 値: Z が観測値以上（絶対値）になる確率
- 例: 標本平均が平均値 μ_0 と異なるほど極端な値となる確率

t 検定 (t-test)

定義（t 検定）

母分散が未知の場合に母平均 μ を検定する方法。

- 帰無仮説: $H_0 : \mu = \mu_0$ ／ 対立仮説: $H_1 : \mu \neq \mu_0$
- 統計量: $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$
- p 値: T が観測値以上（絶対値）になる確率
- 例: 少数サンプルでも平均の差を検定可能

二標本 t 検定 (Two-sample t-test)

定義（二標本 t 検定）

2 つの独立したサンプルの母平均が等しいかを検定する方法。

- 帰無仮説: $H_0 : \mu_1 = \mu_2$ ／ 対立仮説: $H_1 : \mu_1 \neq \mu_2$
- 統計量 (等分散の場合): $T = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \sim t_{n_1+n_2-2}$
- p 値: T が観測値以上（絶対値）になる確率
- 変分散の場合は Welch の t 検定を使用

Welch の検定と通常の二標本 t 検定の違い

定義（Welch の t 検定）

2 群の母分散が等しくない場合に使用する、二標本 t 検定の拡張版。

- 通常の二標本 t 検定は「母分散が等しい」ことを前提にプール分散 s_p^2 を使う
- Welch の検定は分散が異なることを許容し、別々の標本分散と自由度を用いる
- Welch 統計量:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- 自由度:

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- 分散が等しいか不明なときは Welch を使うのが安全

対応のある二標本 t 検定 (Paired t-test)

定義（対応のある二標本 t 検定）

同じ対象に対して 2 条件で測定したときの差を検定する方法。

- 各ペア (x_i, y_i) の差 $d_i = x_i - y_i$ を求める
- 帰無仮説: H_0: 差の平均 $\mu_d = 0$
- 統計量:

$$T = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

- 例: 同一人物の前後比較（施術前後など）
- 対応がある =「差」に注目して一標本 t 検定を行うこと

母分散検定 (Variance Test)

定義（分散検定）

正規分布を前提に、母分散 σ^2 が仮定値 σ_0^2 と異なるかを検定する方法。

- 統計量: $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$
- p 値: χ^2 が観測値以上（または以下）になる確率
- 例: 生産プロセスのばらつき検定
- 授業でのカイ二乗検定はこれを指している

F 検定 (F-test)

定義（F 検定）

2 つの母分散が等しいかを比較する方法。

- 統計量: $F = s_1^2 / s_2^2 \sim F_{n_1-1, n_2-1}$
- p 値: F が観測値以上（または以下）になる確率
- 例: グループ間の分散比較

適合度のカイ二乗検定

定義（適合度検定）

観測した度数分布が，あらかじめ想定した理論的分布 $P = (p_1, \ldots, p_k)$ に適合しているかをカイ二乗分布を用いて検証する。

- 統計量：

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \quad E_i = Np_i$$

- 自由度：$\nu = k - 1$
- p 値：$P(\chi^2_\nu \geq \chi^2_{obs})$
- 条件：全ての期待度数 $E_i \geq 5$ が望ましい

例: サイコロ 60 回の出目 ([8,9,12,11,10,10])

- $E_i = 60/6 = 10$, $\chi^2 = 1.00$, $\nu = 5$
- 棄却点 $\chi^2_{0.05,5} \approx 11.07$, $1.00 < 11.07$ より H_0 を棄却せず

独立性のカイ二乗検定

定義（独立性検定）

属性 A, B の独立性を検証する。

- 統計量：

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{\left(n\, p_{i,j} - n\, p_i\, p_j\right)^2}{n\, p_i\, p_j}$$

- 帰無仮説 H_0: すべての i, j に対し $P(A_i \cap B_j) = P(A_i)P(B_j)$
- 自由度：$\nu = (r-1)(c-1)$
- 棄却規準:
  - $\chi^2 > \chi^2_\alpha\big((r-1)(c-1)\big)$ のとき H_0 を棄却
  - $\chi^2 \leq \chi^2_\alpha\big((r-1)(c-1)\big)$ のとき 棄却しない
- パラメータ:
  - n: 標本数、$p_{i,j}$: 属性が (i, j) となる確率、p_i: カテゴリ i の周辺確率

独立性検定の例 (薬と副作用)

目的

薬の種類 (A vs B) と副作用の有無 (有 vs 無) が統計的に独立かどうか検証する。

1. データ:

	有	無	合計
A	20	30	50
B	15	35	50
合計	35	65	100

独立性検定の例 (薬と副作用)

目的

薬の種類 (A vs B) と副作用の有無 (有 vs 無) が統計的に独立かどうか検証する。

1. 帰無仮説 H_0: 薬の種類と副作用は独立
2. 期待度数:

$$E_{ij} = N \times P(A_i)P(B_j), \quad E_{A,\text{有}} = 100 \times \frac{50}{100} \times \frac{35}{100} = 17.5$$

3. χ^2 統計量の計算:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx 0.71$$

4. 自由度 $\nu = (2-1)(2-1) = 1$, $\chi^2_{0.05,1} = 3.84$
5. 判断: $0.71 < 3.84$ → H_0 を棄却せず → 独立と判断

附録

附録 1: 積率母関数による分布収束定理

定理

確率変数列 X_n の積率母関数 $M_{X_n}(t)$ が存在[8]、各 t に対し $M_{X_n}(t) \to M(t)$ と収束し、$M(t)$ がある分布の積率母関数であれば、X_n はその分布に分布収束する。

- 証明は Lévy の定理（特性関数版）と積率母関数の一意性から従う。
- 各点収束と分布収束を結び付ける定理！

[8] ある近傍 $|t| < \varepsilon$ で一様に定義され

附録 1: 積率母関数による分布収束定理

定理（Lévy の収束定理）

特性関数列 $\varphi_{X_n}(t)$ が任意の t で収束し、極限関数 $\varphi(t)$ が連続であれば、確率変数列 X_n は分布収束し、その極限分布の特性関数は $\varphi(t)$ である。

1. $\varphi_{X_n}(t) \to \varphi(t)$ から、φ は特性関数として適切。（連続性・非退化性を満足）
2. $\varphi_{X_n}(0) = 1$ かつ連続性から、$\{X_n\}$ は緊密である（Прохоров の定理）。
3. 任意の部分列 X_{n_k} には更に収束部分列が取り出せ、極限分布 F が存在。
4. 部分収束先の特性関数は限りなく $\varphi_{X_n}(t) \to \varphi(t)$ に一致する。
5. 特性関数の一意性より、全列の分布関数が一意的に F に収束する。

※ う〜ん、あってんのかな？ これ。測度論の言葉を翻訳したけど、ごめん自信ない。専門書を参照してください。