理論・推論・信頼性

現代統計学

統計理論を体系化する4巻シリーズ

巻1

確率と統計的推論の原理

データサイエンス実務者のための包括的テキスト

2026年3月14日

目次

巻1の読み方

直観的理解

巻1は、全4巻の土台である。 ここで扱うのは「確率をどう厳密に書くか」「推定量や検定をどう正当化するか」 「大標本でなぜ正規近似が効くのか」という、 統計学を道具箱ではなく*原理*として理解するための骨格である。
構成は、大学院の理論統計コア科目で繰り返し採られる 「確率の基礎 → 推定 → 検定 → 漸近理論」 の流れを土台にしつつ、 近年のデータサイエンス向けテキストが重視する 「例・演習・コードで往復する学び方」を重ねてある。 読む順番に迷ったときは、本ガイドの「三つの読み進め方」から 自分の目的に合うものを選べばよい。

この巻で身につける見方

- 数式を「記号の並び」としてではなく、 「何を主張しているか」を日本語で説明できるようにする。
- 推定・検定・信頼区間を別々の手法として暗記するのではなく、 確率論・損失・漸近近似という共通原理で結びつけて理解する。
- コード例は答えを出すためではなく、 有限標本で何が起きるかを観察し、 理論の適用条件を確かめるために使う。

この巻からつながる現代的テーマ

2026年時点の統計学教育では、 確率・推定・検定・漸近理論はそれ自体で完結する基礎ではなく、 因果推論、統計的学習、再標本化、確率過程、生成モデルへ進むための共通言語として扱われることが多い。 本巻は、その接続点を理解するための下地を与える。

因果推論 無作為化、共変量調整、反実仮想を理解するには、 推定量・検定・信頼区間の性質を土台から押さえる必要がある。

統計的学習 回帰、正則化、交差検証、分類器の評価も、 結局は損失、汎化誤差、標本変動をどう捉えるかという問題に帰着する。

再標本化とノンパラメトリック法 ブートストラップや順列検定は、 本巻で学ぶ漸近理論や検定理論を モデル仮定の弱い場面へ拡張するための自然な次の一歩である。

生成モデルと現代AI 生成モデル、系列モデル、拡散モデルを学ぶときも、 確率モデル、尤度、サンプリング、不確実性評価の考え方が基礎になる。

三つの読み進め方

最短ルート まず全体像をつかみたい読者向けである。 第1章は §1.1〜§1.3 を中心に読み、第2章は大数の法則・中心極限定理・デルタ法を優先する。 その後、 第3章では MLE と Cramér–Rao 下界、 第4章では検定と信頼区間の双対性、 第5章では損失・リスク・James–Stein 推定量、 第6章では MLE の漸近理論と三大検定の漸近等価性を読む。

標準ルート 章を順に読み、各章末の演習から 「理論を1問、計算・実装を1問」選んで解く読み方である。 実務者が最も理解を定着させやすい。 数式の読み下しとシミュレーションを往復することで、 漸近理論の主張が有限標本でどこまで信用できるかが見えてくる。

理論重視ルート 証明まで追いたい読者向けである。 付録A を先にざっと見て測度論の語彙を確認し、 各章の定義・定理・演習を飛ばさずに進める。 とくに第2章と第6章は、 巻2以降のベイズ・ノンパラメトリック・高次元理論の前提になる。

章のつながり

章	中心となる問い	読み終えた直後の実務的な効き方
第1章	確率を厳密に扱うには、なぜ測度と可測性が必要か。	条件付き期待値、分布、確率不等式をひとつの言語で説明できるようになる。後続章のすべての前提である。
第2章	「標本が増えると真値に近づく」とは何を意味するか。	一致性、正規近似、デルタ法の使いどころが明確になる。 実務でよく使う Wald 型近似の前提もここで理解する。
第3章	統計モデルをどう置き、推定量の良さをどう測るか。	MLE、モーメント法、ロバスト推定を比較するときに、 十分統計量やフィッシャー情報量を軸に整理できる。
第4章	検定・信頼区間・多重比較はどうつながっているか。	p値は判断材料、検出力は性能指標、FDR は多重比較で制御する誤り率として整理し、 検定手続きをどう設計するかを一つの視点で理解できる。
第5章	推定量や検定の「良さ」を何で定義すべきか。	損失関数が変わると最適な手続きも変わることを理解し、 縮小推定や経験ベイズを正則化の祖先として見られるようになる。

章	中心となる問い	読み終えた直後の実務的な効き方
第6章	有限標本で解けない問題を、大標本近似でどう扱うか。	MLE、Wald 検定、スコア検定、サンドイッチ推定量の関係を整理し、「近似で済ませてよい場面」と「シミュレーションで確認すべき場面」を切り分けられる。

本ガイドの役割は、巻全体の地図を先に与えることである。 各章末の「次章への橋渡し」や「参考文献ノート」は、 いま読んでいる章から次へ進むための局所的な案内として使い分けるとよい。

巻1で頻出する記号

記号	意味
$(\Omega, \mathcal{F}, \mathrm{P})$	確率空間。Ω は起こりうる結果全体、$\mathcal{F}$ は確率を割り当てられる事象の族、P は確率測度。
$X, Y, X_1, \ldots, X_n$	確率変数。観測データは通常 $X_1, \ldots, X_n$ と書く。
$F_X(x)$	確率変数 X の分布関数。$F_X(x) = \mathrm{P}(X \le x)$。
$f(x \mid \theta)$	パラメータ θ の下での密度関数または確率質量関数。
$\mathbb{E}[X], \mathrm{Var}(X)$	期待値と分散。平均的な位置とばらつきを表す。
$\hat{\theta}_n$	標本サイズ n に基づく推定量。n が大きくなると添字が重要になる。
i.i.d.	独立同一分布。各標本が同じ分布に従い、互いに独立であること。
$\xrightarrow{p}, \xrightarrow{d}, \xrightarrow{\text{a.s.}}$	それぞれ確率収束・分布収束・概収束。漸近理論の基本言語。
$\ell_n(\theta)$	対数尤度。多くの場合、MLE はこれを最大化して求める。
$\mathcal{I}(\theta)$	フィッシャー情報量（または情報行列）。推定の難しさと精度を測る。
$R(\theta, \delta)$	決定規則 δ のリスク関数。損失の期待値として定義される。
$o_p(1), O_p(1)$	確率的 little-o / big-O。漸近展開で誤差の大きさを記述するときに使う。

コードと再現性

本書のコード例は、最短のライブラリ呼び出しよりも 「何を確かめるための計算か」が読めることを優先している。 とくに以下の4点を一貫した約束として採用する。

- 乱数シードを明示し、毎回同じ図や数値の傾向を再現できるようにする。
- 理論上の主張を1回の出力で済ませず、 反復シミュレーションでばらつきや近似誤差も確認する。
- R と Python は競合する言語としてではなく、 「古典統計に強い R」と「機械学習接続に強い Python」を 使い分ける道具として扱う。
- 理論とコードが食い違うときは、 まず仮定（独立性、有限分散、モデルの正しさ）が成り立っているかを点検する。

実務では「動いたコード」より 「繰り返しても結論が大きくぶれないコード」の方が重要である。 そのため本書では、数値実験を 予測可能性・計算可能性・安定性を確かめる手段として使う。

併読ガイド

理論を厚くしたいとき 古典的推測理論は *Statistical Inference* と *Asymptotic Statistics* を併読すると見通しがよい。

実務との接点を増やしたいとき 近年のデータサイエンス向けテキストである *Probability and Statistics for Data Science* や *Veridical Data Science* は、 本巻の理論が分析実務でどこに効くかを補助線として与えてくれる。

巻2以降への接続を意識したいとき 本ガイドで全体の地図を押さえたうえで、 各章末の「次章への橋渡し」と「参考文献ノート」を併用すると、 巻の中の接続と巻2以降への伸び方を追いやすい。 とくに第5章と第6章は、 巻2のベイズ・計算統計、巻3の高次元理論への入口になる。

第I部

確率論の基礎

第1章

測度論的確率論

問いと学習目標

この章で答える問い

- なぜ確率を厳密に扱うために、集合論的な枠組み（σ-加法族・測度）が必要なのか？
- 「確率変数」とは何を「変数」にしているのか？　関数としてどう定義されるのか？
- 期待値はなぜリーマン積分ではなくルベーグ積分で定義するのか？
- 条件付き期待値はなぜ「確率変数」になるのか？
- 指数型分布族とは何か、なぜ統計学で中心的な役割を果たすのか？

読み終えたらできるようになること

1. 確率空間 $(\Omega, \mathcal{F}, \mathrm{P})$ の各構成要素を説明し、具体例を構成できる。
2. 確率変数・分布関数・密度関数の関係を述べ、変数変換を実行できる。
3. 期待値・分散を測度論的に定義し、基本性質を証明できる。
4. 条件付き期待値を定義から説明し、tower property を使った計算ができる。
5. 主要な確率分布を指数型分布族の枠組みで整理できる。
6. マルコフ・チェビシェフ・イェンセンの不等式を使い分けられる。

直観的理解

確率論は「不確実性を数学で扱う」ための言語である。 サイコロの目から金融市場の変動、遺伝子の発現まで、 あらゆるランダムな現象を統一的に記述する枠組みが測度論的確率論である。 学部レベルの確率論が「道具の使い方」を教えるものだとすれば、 本章は「道具がなぜ正しく動くのか」を理解するための基盤を築く。
本章のロードマップは次のとおりである。 まず **確率空間**（§1.1）で 確率を厳密に定義するための土台を据える。 次に**確率変数と分布**（§1.2）で ランダムな量を関数として定式化し、 **期待値と条件付き期待値**（§1.3）で 平均的な振る舞いを積分として

捉える。**特性関数**（§1.4）で分布を フーリエ変換の言葉に翻訳し、**主要な確率分布族**（§1.5）で 統計学の道具箱を整備する。**独立性**（§1.6）と **確率不等式**（§1.7）は、 後章のすべての理論を支える基盤となる。

1.1　確率空間

1.1.1　なぜ σ-加法族が必要か

学部の確率論では、標本空間 Ω のすべての部分集合に 確率を割り当てることができた。 たとえばサイコロを1回振る実験では $\Omega = \{1,2,3,4,5,6\}$ であり、 Ω の部分集合は $2^6 = 64$ 個しかないから、 すべての部分集合に確率を定めることに困難はない。

しかし $\Omega = [0,1]$ のように標本空間が非可算集合であるとき、 Ω のすべての部分集合（冪集合 2^Ω）に対して、 「長さの自然な一般化」を保ったまま確率を割り当てることは一般にできない。

ここで**冪集合**（power set）2^Ω とは、 Ω のすべての部分集合を要素とする集合族のことである。

注意 1.1.1. 選択公理を仮定すると、ルベーグ測度に関して $[0,1]$ の「測れない」部分集合（ヴィタリ集合） が存在することが知られている（Vitali, 1905）。 すなわち、$[0,1]$ 上のすべての部分集合に対して 「長さの自然な一般化」を矛盾なく定義することはできない。 これが σ-加法族という 「確率を測れる事象のみを集めた枠組み」を導入する根本的な理由である。

このような事情から、確率を割り当てる事象の範囲を適切に制限する必要がある。 その制限を数学的に表したものが σ-加法族である。

1.1.2　確率空間の定義

確率論のすべての構成は、**確率空間**という三つ組の上に成り立つ。

定義 1.1.2 (σ-加法族)**.** Ω の部分集合の族 $\mathcal{F} \subset 2^\Omega$ が **σ-加法族**（σ-algebra）であるとは、 以下の3条件を満たすことをいう：

(i) $\Omega \in \mathcal{F}$（全体集合を含む）。

(ii) $A \in \mathcal{F}$ ならば $A^c \in \mathcal{F}$（補集合について閉じている）。

(iii) $A_1, A_2, \ldots \in \mathcal{F}$ ならば $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$（可算和について閉じている）。

読み下し

σ-加法族とは、「確率を測ることが許される事象の集まり」である。 条件 (i)〜(iii) は、 「何かが起きる」こと全体、「起きない」こと、「少なくとも1つが起きる」ことに 確率を割り当てられることを保証する。 有限個の和だけでなく可算無限個の和についても閉じている点が、 単なる加法族との違いであり、極限操作を可能にする核心的な条件である。

定義 1.1.3 (確率空間)**.** 三つ組 $(\Omega, \mathcal{F}, \mathrm{P})$ を**確率空間**と呼ぶ。ここで

(i) Ω は**標本空間**（sample space）であり、起こりうるすべての結果の集合である。

(ii) $\mathcal{F}$ は Ω 上の σ-加法族であり、「確率を測れる」事象の集まりである。

(iii) $\mathrm{P}\colon \mathcal{F} \to [0,1]$ は**確率測度**（probability measure）である。

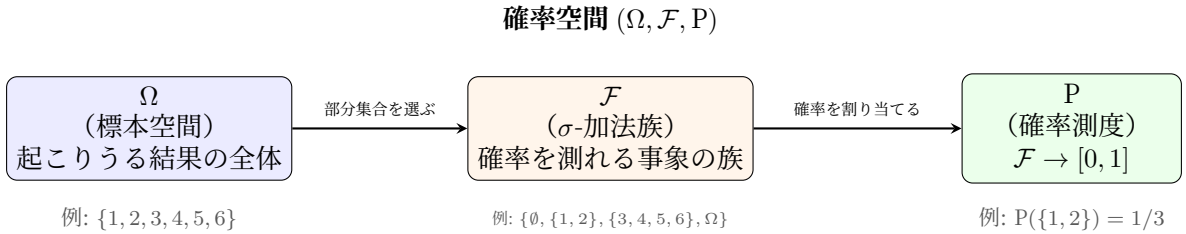


図 1.1: 確率空間の三つ組。標本空間 Ω の部分集合のうち σ-加法族 $\mathcal{F}$ に属するものだけに、確率測度 P が確率を割り当てる。

定義 1.1.4 (確率測度)**.** $(\Omega, \mathcal{F})$ 上の関数 $\mathrm{P}\colon \mathcal{F} \to [0,1]$ が確率測度であるとは：

(i) $\mathrm{P}(\Omega) = 1$（正規化条件）。

(ii) 互いに素な事象の列 $\{A_n\}_{n=1}^{\infty}$（すなわち $i \neq j$ のとき $A_i \cap A_j = \emptyset$）に対して、

$$\mathrm{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathrm{P}(A_n)$$

が成り立つ（σ-加法性 / 完全加法性）。

読み下し

確率測度の2つの条件は、「何かは必ず起きる（全確率は1）」ことと、「互いに排他的な事象の確率は足し合わせられる」ことを述べている。σ-加法性は有限和だけでなく可算無限和でも成り立つことを要求しており、これにより極限と確率の交換が可能になる。

1.1.3　具体例

例 1.1.5 (コイン投げ)**.** 公正なコインを1回投げる実験を考える。$\Omega = \{H, T\}$、$\mathcal{F} = 2^{\Omega} = \{\emptyset, \{H\}, \{T\}, \Omega\}$、$\mathrm{P}(\{H\}) = \mathrm{P}(\{T\}) = 1/2$ とすれば、$(\Omega, \mathcal{F}, \mathrm{P})$ は確率空間となる。ここでは Ω が有限集合なので、冪集合全体を σ-加法族にとってよい。

例 1.1.6 (ボレル σ-加法族)**.** $\Omega = \mathbb{R}$ のとき、すべての開区間 (a, b) を含む最小の σ-加法族を **ボレル σ-加法族** $\mathcal{B}(\mathbb{R})$ と呼ぶ。連続分布を扱う際の標準的な σ-加法族である。「最小の」とは、開区間を含むすべての σ-加法族の共通部分をとったものという意味であり、このような構成が σ-加法族になることは σ-加法族の定義から確かめられる。$\mathcal{B}(\mathbb{R})$ は開集合・閉集合・可算和・可算積をすべて含むが、$\mathbb{R}$ のすべての部分集合を含むわけではない。

1.1.4　確率測度の基本性質

命題 1.1.7 (確率測度の性質)**.** *確率空間* $(\Omega, \mathcal{F}, \mathrm{P})$ *において、以下が成り立つ：*

(i) $\mathrm{P}(\emptyset) = 0$。

(ii) $A \subset B$ *ならば* $\mathrm{P}(A) \leq \mathrm{P}(B)$*（単調性）*。

(iii) $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$*（余事象）*。

(iv) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$*（包除原理）*。

(v) $A_1 \subset A_2 \subset \cdots$ *で* $A = \bigcup_{n=1}^{\infty} A_n$ *のとき、* $\mathrm{P}(A_n) \to \mathrm{P}(A)$*（下からの連続性）*。

(vi) $A_1 \supset A_2 \supset \cdots$ *で* $A = \bigcap_{n=1}^{\infty} A_n$ *のとき、* $\mathrm{P}(A_n) \to \mathrm{P}(A)$*（上からの連続性）*。

読み下し

これらの性質はすべて σ-加法性から導かれる。とりわけ (v)(vi) の連続性は確率測度が極限操作と相性がよいことを示しており、第2章で学ぶ収束理論の出発点となる。

Proof. **証明の方針**: いずれの性質も、事象を互いに素な事象に分解し、σ-加法性を適用するという共通の戦略で示せる。

(i) Ω と $\emptyset$ は互いに素であり、$\Omega = \Omega \cup \emptyset$ より $1 = \mathrm{P}(\Omega) = \mathrm{P}(\Omega) + \mathrm{P}(\emptyset)$ から $\mathrm{P}(\emptyset) = 0$。

(ii) $B = A \cup (B \setminus A)$ と互いに素に分解すると、$\mathrm{P}(B) = \mathrm{P}(A) + \mathrm{P}(B \setminus A) \geq \mathrm{P}(A)$。

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

(iv) $A \cup B = A \cup (B \setminus (A \cap B))$ と $B = (A \cap B) \cup (B \setminus (A \cap B))$ から導かれる。

(v) $B_1 = A_1$、$B_n = A_n \setminus A_{n-1}$（$n \geq 2$）とおくと $\{B_n\}$ は互いに素で $\bigcup_{n=1}^{\infty} B_n = A$ であるから、σ-加法性より $\mathrm{P}(A) = \sum_{n=1}^{\infty} \mathrm{P}(B_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mathrm{P}(B_n) = \lim_{N\to\infty} \mathrm{P}(A_N)$。

(vi) $C_n = A_1 \setminus A_n$ とおくと $C_n \subset C_{n+1} \subset \cdots$ で $\bigcup C_n = A_1 \setminus A$ である。(v) より $\mathrm{P}(A_1 \setminus A) = \lim_n \mathrm{P}(A_1 \setminus A_n)$。また $A_1 = A_n \sqcup (A_1 \setminus A_n)$ および $A_1 = A \sqcup (A_1 \setminus A)$ であるから、$\mathrm{P}(A_1 \setminus A_n) = \mathrm{P}(A_1) - \mathrm{P}(A_n)$, $\mathrm{P}(A_1 \setminus A) = \mathrm{P}(A_1) - \mathrm{P}(A)$ である。したがって $\mathrm{P}(A_1) - \mathrm{P}(A) = \lim_n (\mathrm{P}(A_1) - \mathrm{P}(A_n))$ より $\lim_n \mathrm{P}(A_n) = \mathrm{P}(A)$。 □

1.2　確率変数と分布関数

確率空間の上で「数値を返すランダムな量」を扱うために、確率変数を定義する。

定義 1.2.1 (確率変数)**.** 確率空間 $(\Omega, \mathcal{F}, \mathrm{P})$ 上の関数 $X \colon \Omega \to \mathbb{R}$ が **確率変数**（random variable）であるとは、X が $\mathcal{F}/\mathcal{B}(\mathbb{R})$-**可測**であること、すなわち、すべてのボレル集合 $B \in \mathcal{B}(\mathbb{R})$ に対して

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

が成り立つことをいう。

読み下し

確率変数とは、ランダムな結果 ω を実数値に対応させる「翻訳装置」である。可測性の条件は、「X がある範囲に入る確率を計算できる」ことを保証している。たとえば「$X \le 3$ となる確率は？」と問うとき、$\{\omega : X(\omega) \le 3\}$ が $\mathcal{F}$ に属していなければ P を適用できないが、可測性はこれを保証する。

注意 1.2.2. 「確率変数」という名称は歴史的なものであるが、数学的には確率変数は*関数*（Ω から $\mathbb{R}$ への写像）であって「変数」ではない。学部で慣れ親しんだ「X は正規分布に従う確率変数」という表現は、実際には「関数 $X\colon \Omega \to \mathbb{R}$ の像の分布が 正規分布に従う」ことを意味している。

1.2.1　分布関数

確率変数 X の確率的な振る舞いは、その**分布**（$\mathbb{R}$ 上の確率測度 $P_X = \mathrm{P} \circ X^{-1}$）で 完全に特徴づけられる。分布を表現する最も基本的な方法が分布関数である。

定義 1.2.3 (累積分布関数)**.** 確率変数 X の**累積分布関数**（CDF）を

$$F_X(x) = \mathrm{P}(X \le x), \quad x \in \mathbb{R}$$

と定義する。

読み下し

$F_X(x)$ は「X の値が x 以下になる確率」を表す。x を $-\infty$ から ∞ まで動かすと、確率が 0 から 1 まで 単調に増加する関数が得られる。

定理 1.2.4 (分布関数の特徴づけ)**.** *関数 $F\colon \mathbb{R} \to [0,1]$ がある確率変数の累積分布関数であるための 必要十分条件は：*

(i) *F は単調非減少。*

(ii) *F は右連続：すべての x で $\lim_{h \downarrow 0} F(x+h) = F(x)$。*

(iii) $\lim_{x \to -\infty} F(x) = 0$、$\lim_{x \to \infty} F(x) = 1$。

読み下し

この定理は、(i)〜(iii) を満たす関数を1つ与えれば、それに対応する確率分布がただ1つ存在することを保証する。逆に、任意の確率分布の CDF は必ずこの3条件を満たす。

1.2.2　離散分布と連続分布

定義 1.2.5 (離散分布)**.** 確率変数 X が**離散分布**に従うとは、高々可算個の点 $\{x_1, x_2, \dots\}$ が存在して $\sum_k \mathrm{P}(X = x_k) = 1$ が成り立つことをいう。このとき $p(x_k) = \mathrm{P}(X = x_k)$ を**確率質量関数**（PMF）と呼ぶ。

定義 1.2.6 (連続分布)**.** 確率変数 X が**連続分布**に従うとは、 非負可測関数 f_X が存在して

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt, \quad x \in \mathbb{R}$$

が成り立つことをいう。f_X を**確率密度関数**（PDF）と呼ぶ。

読み下し

離散分布では各点に正の確率が「塊」として乗っており、 連続分布では確率が滑らかに「塗り広げ」られている。 連続分布では個々の点の確率は $\mathrm{P}(X = x) = 0$ であり、確率は区間の「面積」（密度関数の積分）としてのみ意味を持つ。

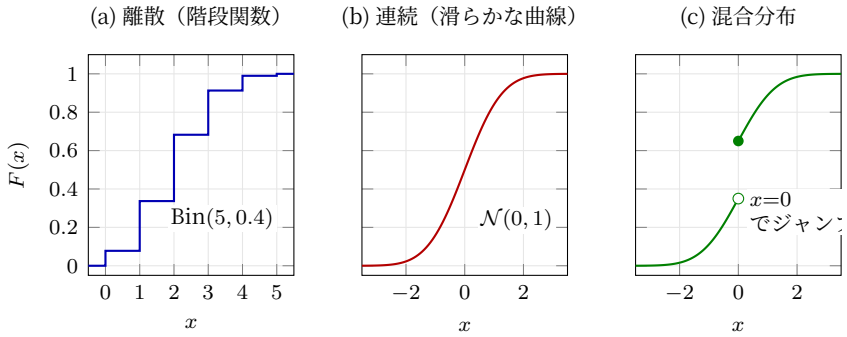


図 1.2: 累積分布関数 (CDF) の3つの型。 (a) 離散分布の CDF は階段関数で、各台の点でジャンプする。 (b) 連続分布の CDF は滑らかな曲線で、ジャンプを持たない。 (c) 混合分布の CDF は連続部分と離散的なジャンプの両方を持つ。 いずれも右連続・単調非減少・$F(-\infty) = 0$, $F(\infty) = 1$ を満たす。

1.2.3　確率変数の変換

確率変数 X に関数を施して新しい確率変数 $Y = g(X)$ を作ることは実務上頻繁に現れる。 Y の分布を求める一般的な方法を述べる。

定理 1.2.7 (変数変換公式)**.** *X が密度関数 f_X を持つ連続確率変数とし、 $g\colon \mathbb{R} \to \mathbb{R}$ が単調で微分可能な関数とする。 $Y = g(X)$ の密度関数は*

$$f_Y(y) = f_X\bigl(g^{-1}(y)\bigr) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

で与えられる。

読み下し

変数変換では、もとの密度に逆関数のヤコビアン（微分の絶対値）を掛ける。 直観的には、g が区間を引き伸ばす場所では密度が薄まり、 縮める場所では密度が濃くなる —— その補正がヤコビアンの役割である。

例 1.2.8 (対数変換)**.** $X \sim \mathrm{Exp}(1)$（密度 $f_X(x) = e^{-x}$, $x > 0$）とし、$Y = -\log X$ とおく。$g(x) = -\log x$ は $(0,\infty)$ 上で単調減少、$g^{-1}(y) = e^{-y}$、$\left|\frac{d}{dy} e^{-y}\right| = e^{-y}$ であるから、

$$f_Y(y) = e^{-e^{-y}} \cdot e^{-y} = \exp(-y - e^{-y}), \quad y \in \mathbb{R}.$$

これはガンベル分布（Gumbel distribution）の密度関数である。

1.3　期待値と条件付き期待値

1.3.1　期待値の定義

学部の確率論では、離散確率変数の期待値を $\mathbb{E}[X] = \sum_k x_k p(x_k)$ と、連続確率変数の期待値を $\mathbb{E}[X] = \int x f_X(x)\,dx$ と別々に定義した。しかし、離散でも連続でもない混合分布や、より一般的な確率変数を統一的に扱うために、**ルベーグ積分**による定義が必要になる。

定義 1.3.1 (期待値)**.** 確率変数 X の**期待値**を、ルベーグ積分により

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,d\mathrm{P}(\omega)$$

と定義する。$\mathbb{E}[|X|] < \infty$ のとき X は**可積分**であるといい、このとき期待値が well-defined である。

読み下し

期待値とは「確率で重みづけた平均値」である。ルベーグ積分による定義 $\int_\Omega X\,d\mathrm{P}$ は、離散分布のときは $\sum_k x_k \mathrm{P}(X = x_k)$ に、連続分布のときは $\int x f_X(x)\,dx$ に一致する。つまりルベーグ積分は、これまでの2つの公式を統一する書き方である。ルベーグ積分の構成（単関数による近似）の詳細は付録Aを参照されたい。

命題 1.3.2 (期待値の性質)**.** *可積分な確率変数 X, Y と定数 $a, b \in \mathbb{R}$ に対して：*

(i) **線形性***:* $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$。

(ii) **単調性***: $X \le Y$ がほとんど確実に（a.s.[1]）成り立つならば* $\mathbb{E}[X] \le \mathbb{E}[Y]$。

(iii) **三角不等式***:* $|\mathbb{E}[X]| \le \mathbb{E}[|X|]$。

定義 1.3.3 (分散と共分散)**.** 確率変数 X の**分散**を

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

と定義する（ただし $\mathbb{E}[X^2] < \infty$ を仮定する）。

読み下し

分散は「平均からのずれの二乗」の期待値であり、分布の「散らばり具合」を測る。

[1] a.s. は almost surely（ほとんど確実に）の略。$X \le Y$ a.s. とは $\mathrm{P}(\{\omega : X(\omega) \le Y(\omega)\}) = 1$ を意味する。

第2の等号 $\mathbb{E}[X^2] - (\mathbb{E}[X])^2$ は計算上便利な公式であり、「二乗の期待値から期待値の二乗を引く」と覚えるとよい。

X, Y の**共分散**を

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

と定義する。

読み下し

共分散は2つの確率変数が「一緒に動く度合い」を測る。$\mathrm{Cov}(X,Y) > 0$ なら X が大きいとき Y も大きい傾向があり、$\mathrm{Cov}(X,Y) < 0$ なら逆の傾向がある。$\mathrm{Cov}(X,Y) = 0$ のとき X と Y は**無相関**という。独立ならば無相関だが、逆は一般には成り立たない（反例は §1.6 で述べる）。

1.3.2　条件付き期待値

条件付き期待値は確率論の中でも特に重要かつ抽象度の高い概念である。まず離散的な場合の復習から出発し、一般的な定義へと進もう。

離散的な条件付き期待値の復習.　離散確率変数 Y が値 y をとるという条件のもとでの X の条件付き期待値は

$$\mathbb{E}[X \mid Y = y] = \sum_{x} x \cdot \mathrm{P}(X = x \mid Y = y)$$

で定義された。ここで y を動かして $\mathbb{E}[X \mid Y] = g(Y)$ と書けば、これは Y の関数、すなわちそれ自体が確率変数になる。

なぜ一般化が必要か.　連続確率変数では $\mathrm{P}(Y = y) = 0$ であるから、上の定義をそのまま使うことはできない。また、Y という1つの確率変数ではなく、「σ-加法族 $\mathcal{G}$ が表す情報が与えられたとき」という、より一般的な条件付けを扱いたい。そこで以下の測度論的な定義を導入する。

定義 1.3.4 (条件付き期待値)**.** 可積分確率変数 X と $\mathcal{F}$ の部分 σ-加法族 $\mathcal{G}$ に対して、$\mathbb{E}[X \mid \mathcal{G}]$ とは、以下の2条件を満たす（a.s. で一意な）確率変数 Z をいう：

(i) Z は $\mathcal{G}$-可測（Z は $\mathcal{G}$ の情報のみで決まる）。

(ii) すべての $G \in \mathcal{G}$ に対して $\int_G Z \, d\mathrm{P} = \int_G X \, d\mathrm{P}$。

このような Z の存在と a.s. での一意性は、ラドン・ニコディムの定理（付録A）によって保証される。

読み下し

条件付き期待値 $\mathbb{E}[X \mid \mathcal{G}]$ は、「$\mathcal{G}$ が表す情報だけを使って X を最もよく予測したもの」である。条件 (i) は「答は手持ちの情報のみで決まる」こと、条件 (ii) は「$\mathcal{G}$ のどの事象で区切って平均をとっても、X そのもので平均をとった結果と一致する」ことを要求する。
機械学習の文脈では、$\mathbb{E}[Y \mid \boldsymbol{X}]$ は 入力 $\boldsymbol{X}$ が与えられたときの目的変数 Y の最適予測（二乗誤差最小化の意味で）に他ならない。

例 1.3.5 (離散的な条件付き期待値)**.** $\Omega = \{1,2,3,4\}$、各結果が等確率 $1/4$ で起きるとする。$X(\omega) = \omega$、$\mathcal{G} = \sigma(\{1,2\},\{3,4\})$（すなわち ω が前半か後半かだけがわかる情報）とおくと、

$$\mathbb{E}[X \mid \mathcal{G}](\omega) = \begin{cases} \frac{1+2}{2} = 1.5 & \omega \in \{1,2\} \\ \frac{3+4}{2} = 3.5 & \omega \in \{3,4\} \end{cases}$$

であり、これは確かに $\mathcal{G}$-可測な確率変数であり、$\{1,2\}$ 上でも $\{3,4\}$ 上でも X との積分が一致する。

定理 1.3.6 (条件付き期待値の性質)**.** *可積分確率変数 X, Y と部分 σ-加法族 $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$ に対して：*

(i) **線形性***:* $\mathbb{E}[aX + bY \mid \mathcal{G}] = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]$。

(ii) **繰り返しの法則**（*tower property*）*:* $\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}] \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}]$。

(iii) **全期待値の法則***:* $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$。

(iv) **既知量の括り出し***: Y が $\mathcal{G}$-可測で XY が可積分ならば* $\mathbb{E}[XY \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}]$。

(v) **独立な場合***: X が $\mathcal{G}$ と独立ならば* $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$。

読み下し

(ii) の tower property は「粗い情報で予測した結果を、さらに粗い情報で予測しても、最初から粗い情報だけで予測した結果と同じ」ということを述べている。(iii) は (ii) で $\mathcal{G} = \{\emptyset, \Omega\}$（情報なし）とした特殊ケースである。(v) は「$X$ の振る舞いについて $\mathcal{G}$ が何の情報も持たないなら、条件付けは無意味であり、無条件の期待値に一致する」ことを意味する。

実務ポイント

条件付き期待値の tower property は、統計学と機械学習の至る所で現れる。たとえば：

- **全分散の公式**: $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X])$。これは Y の変動を「X を知っても残る変動」と「X で説明できる変動」に分解する。
- **反復期待値**: ベイズ推論で事前分布→事後分布の計算を 段階的に行う際の理論的根拠となる（巻2第1章）。

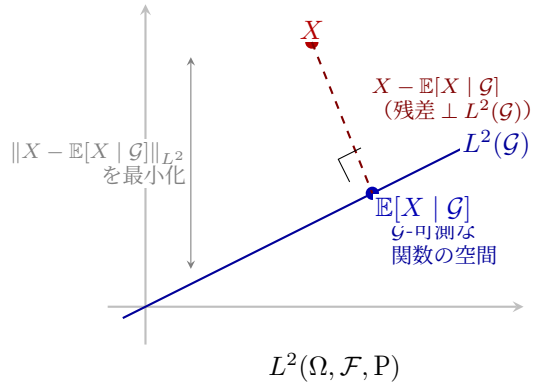


図 1.3: 条件付き期待値の L^2 射影としての解釈。$\mathbb{E}[X \mid \mathcal{G}]$ は確率変数 X を $\mathcal{G}$-可測な関数の閉部分空間 $L^2(\mathcal{G})$ へ直交射影したものであり、二乗誤差 $\mathbb{E}[(X-Z)^2]$ を最小にする $\mathcal{G}$-可測な Z に一致する。残差 $X - \mathbb{E}[X \mid \mathcal{G}]$ は $L^2(\mathcal{G})$ と直交する。

1.4　特性関数とモーメント母関数

分布を特徴づける強力な道具として、モーメント母関数と特性関数がある。まずモーメント母関数から始めよう。

定義 1.4.1 (モーメント母関数)**.** 確率変数 X の**モーメント母関数**（MGF）を

$$M_X(t) = \mathbb{E}[e^{tX}], \quad t \in \mathbb{R}$$

と定義する。ただし 0 を含む開区間で有限値をとるとき「存在する」という。

読み下し

$e^{tX} = \sum_{k=0}^{\infty} (tX)^k/k!$ と展開すると $M_X(t) = \sum_{k=0}^{\infty} t^k \mathbb{E}[X^k]/k!$ となるから、M_X の n 階微分を $t=0$ で評価すれば $M_X^{(n)}(0) = \mathbb{E}[X^n]$ が得られる。これが「モーメント**母**関数」と呼ばれる理由である。

モーメント母関数は強力だが、存在しない場合がある（たとえばコーシー分布 t_1 はモーメント母関数を持たない）。常に存在する代替手段が特性関数である。

定義 1.4.2 (特性関数)**.** 確率変数 X の**特性関数**（characteristic function）を

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)], \quad t \in \mathbb{R}$$

と定義する。ここで $i = \sqrt{-1}$ は虚数単位である。$\left|e^{itX}\right| = 1$ であるから、特性関数はすべての確率変数に対して存在する。

読み下し

特性関数は確率分布のフーリエ変換に他ならない。モーメント母関数の指数 tX を itX に置き換えただけであるが、$\left|e^{itX}\right| \le 1$ が常に成り立つため、積分の絶対収束が無条件に保証される。

定理 1.4.3 (一意性定理)**.** *2つの確率変数 X と Y が同一の分布に従うための必要十分条件は $\varphi_X(t) = \varphi_Y(t)$ がすべての $t \in \mathbb{R}$ で成り立つことである。*

読み下し

この定理は、特性関数が分布を完全に決定することを述べている。 2つの確率変数の特性関数が一致すれば、それらの分布は同じである。 中心極限定理の証明（第2章）では、 特性関数の収束から分布の収束を導くためにこの定理を用いる。

定理 1.4.4 (反転公式)**.** *特性関数 φ_X が可積分 $\int_{-\infty}^{\infty} |\varphi_X(t)|\, dt < \infty$ のとき、X は密度関数*

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t)\, dt$$

を持つ。

注意 1.4.5. 反転公式の可積分条件は、正規分布やガンマ分布では満たされるが、 すべての分布で成り立つわけではない。 たとえば一様分布 $\mathrm{Unif}(0,1)$ の特性関数 $\varphi(t) = (e^{it} - 1)/(it)$ は $t \to \infty$ で $O(1/t)$ の減衰しか持たず、 $\int_{-\infty}^{\infty} |\varphi(t)|\, dt$ が発散するためこの条件を満たさない。 同様に、離散分布の特性関数は周期的であるため可積分ではない。

1.5　主要な確率分布族

1.5.1　指数型分布族

統計学で最も頻繁に現れる分布の多くは、 **指数型分布族**という共通の構造を持つ。 一般的な定義に入る前に、具体例を通じてこの構造を見てみよう。

例 1.5.1 (正規分布の指数型表現)**.** $X \sim \mathcal{N}(\mu, \sigma^2)$ の密度関数は

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{\mu^2}{2\sigma^2} - \log \sigma \right). \end{aligned}$$

ここで自然パラメータ $\boldsymbol{\eta} = (\mu/\sigma^2,\ -1/(2\sigma^2))^\top$、 十分統計量 $\boldsymbol{T}(x) = (x,\ x^2)^\top$ と読み取れる。

例 1.5.2 (ポアソン分布の指数型表現)**.** $X \sim \mathrm{Pois}(\lambda)$ の確率質量関数は

$$p(x \mid \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} = \frac{1}{x!} \exp(x \log \lambda - \lambda).$$

自然パラメータは $\eta = \log \lambda$、十分統計量は $T(x) = x$、 対数分配関数は $A(\eta) = e^\eta$（$= \lambda$）である。

これらの例に共通するパターンを一般化する。

定義 1.5.3 (指数型分布族)**.** パラメータ $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$ を持つ分布族が **指数型分布族**（exponential family）に属するとは、密度関数（または質量関数）が

$$f(x \mid \boldsymbol{\theta}) = h(x) \exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{T}(x) - A(\boldsymbol{\theta})\Big)$$

と書けることをいう。ここで

- $\boldsymbol{\eta}(\boldsymbol{\theta}) \in \mathbb{R}^k$：**自然パラメータ**（natural parameter）。$\boldsymbol{\theta}$ を自然パラメータに変換する写像。
- $\boldsymbol{T}(x) \in \mathbb{R}^k$：**十分統計量**（sufficient statistic）。データ x から必要な情報を抽出する関数。
- $A(\boldsymbol{\theta})$：**対数分配関数**（log-partition function）。正規化定数の対数。
- $h(x) \geq 0$：基底測度。パラメータに依存しない x の関数。

ベクトル $\boldsymbol{\theta}, \boldsymbol{\eta}$ 等の太字はベクトル値であることを示す。

読み下し

指数型分布族の密度は「指数の中身がパラメータとデータの内積」という構造を持つ。この構造のおかげで、十分統計量・最尤推定・共役事前分布といった統計的概念が自然かつ統一的に扱える。

命題 1.5.4 (対数分配関数とモーメント)**.** *自然パラメータ表現 $f(x \mid \boldsymbol{\eta}) = h(x)\exp(\boldsymbol{\eta}^\top \boldsymbol{T}(x) - A(\boldsymbol{\eta}))$ において：*

(i) $\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\eta}}[\boldsymbol{T}(X)]$。

(ii) $\nabla^2_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{T}(X))$。

(iii) $A(\boldsymbol{\eta})$ *は凸関数である。*

ここで $\nabla_{\boldsymbol{\eta}}$ は自然パラメータに関する勾配（偏微分ベクトル）、$\nabla^2_{\boldsymbol{\eta}}$ はヘッセ行列（二階偏微分の行列）を表す。

読み下し

(i) は「対数分配関数を微分するだけで十分統計量の期待値が得られる」という非常に便利な性質である。(ii) より共分散行列が得られ、(iii) より A が凸であることから最尤推定の最適化が扱いやすくなる。

Proof. **証明の方針**: $\int f(x \mid \boldsymbol{\eta})\, dx = 1$ の両辺を $\boldsymbol{\eta}$ で微分する。積分記号下の微分の正当化には優収束定理（付録A）を用いる。

(i) $\int h(x)\exp(\boldsymbol{\eta}^\top \boldsymbol{T}(x) - A(\boldsymbol{\eta}))\, dx = 1$ の両辺を η_j で偏微分すると

$$\int h(x)(T_j(x) - \partial_j A(\boldsymbol{\eta}))\exp(\boldsymbol{\eta}^\top \boldsymbol{T}(x) - A(\boldsymbol{\eta}))\, dx = 0.$$

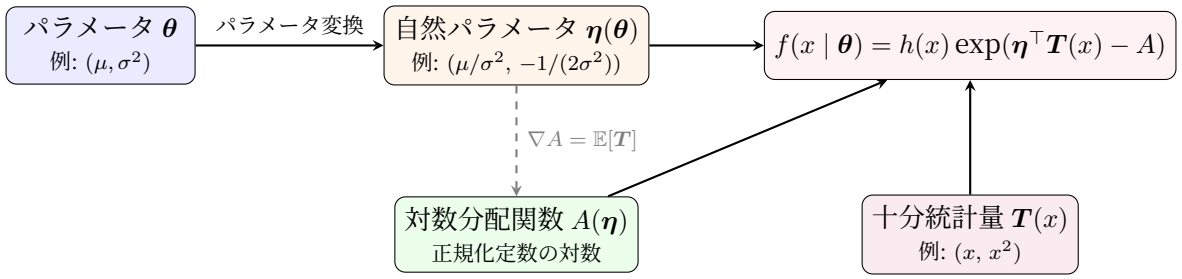
すなわち $\mathbb{E}_{\boldsymbol{\eta}}[T_j(X)] - \partial_j A(\boldsymbol{\eta}) = 0$。

(ii) (i) の等式をさらに η_k で微分すれば、$\partial_j \partial_k A = \mathbb{E}[T_j T_k] - \mathbb{E}[T_j]\mathbb{E}[T_k] = \mathrm{Cov}(T_j, T_k)$ が得られる。

(iii) (ii) よりヘッセ行列が共分散行列に等しく、 共分散行列は半正定値であるから A は凸である。 □

実務ポイント

指数型分布族は統計学の中心的構成要素である。 正規分布、ポアソン分布、二項分布、ガンマ分布、ベータ分布など、 実務で最もよく用いる分布の多くがこの族に属する。 一般化線形モデル（巻2第3章）では応答分布として指数型分布族を仮定し、 ベイズ推論（巻2第1章）では共役事前分布が自然に導かれる。
一方、指数型分布族に属さない重要な分布も存在する。 たとえばコーシー分布 t_1 は期待値が存在せず、 混合分布（たとえば正規混合モデル）は一般に指数型分布族に属さない。



∇A: 十分統計量の期待値　$\nabla^2 A$: 十分統計量の共分散（$\geq 0 \Rightarrow A$ は凸）

図 1.4: 指数型分布族の構造。 パラメータ $\boldsymbol{\theta}$ は自然パラメータ $\boldsymbol{\eta}$ に変換され、 データ x の情報は十分統計量 $\boldsymbol{T}(x)$ に集約される。 対数分配関数 A は正規化を担うと同時に、 その微分がモーメント（期待値・共分散）を与える。

1.5.2　主要な離散分布

以下では主要な離散分布を列挙する。いずれも指数型分布族に属する （幾何分布と負の二項分布を含む）。

1. **ベルヌーイ分布** $\mathrm{Bern}(p)$: $\mathrm{P}(X=1)=p$, $\mathrm{P}(X=0)=1-p$。 成功／失敗の二値試行をモデル化する最も基本的な分布。 $\mathbb{E}[X]=p$, $\mathrm{Var}(X)=p(1-p)$。

2. **二項分布** $\mathrm{Bin}(n,p)$: $\mathrm{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$、$k=0,1,\dots,n$。 n 回の独立なベルヌーイ試行での成功回数の分布。 $\mathbb{E}[X]=np$, $\mathrm{Var}(X)=np(1-p)$。

3. **ポアソン分布** $\mathrm{Pois}(\lambda)$: $\mathrm{P}(X=k)=e^{-\lambda}\lambda^k/k!$、$k=0,1,2,\dots$。 単位時間あたりの発生回数のモデルとして広く用いられる。 $\mathbb{E}[X]=\mathrm{Var}(X)=\lambda$。

4. **幾何分布** $\mathrm{Geom}(p)$: $\mathrm{P}(X=k)=(1-p)^{k-1}p$、$k=1,2,\dots$。 最初の成功までの試行回数。 $\mathbb{E}[X]=1/p$, $\mathrm{Var}(X)=(1-p)/p^2$。

5. **負の二項分布** $\mathrm{NB}(r,p)$: r 回目の成功までの失敗回数の分布。 $\mathbb{E}[X]=r(1-p)/p$, $\mathrm{Var}(X)=r(1-p)/p^2$。

1.5.3　主要な連続分布

1. **一様分布** $\mathrm{Unif}(a,b)$: $f(x)=1/(b-a)$、$a<x<b$。 区間上で「偏りなく」ランダムに値をとる分布。 $\mathbb{E}[X]=(a+b)/2$, $\mathrm{Var}(X)=(b-a)^2/12$。

2. **正規分布** $\mathcal{N}(\mu,\sigma^2)$: $f(x)=(2\pi\sigma^2)^{-1/2}\exp(-(x-\mu)^2/(2\sigma^2))$。 中心極限定理の帰結として、 多くの独立な微小な変動の和は正規分布に近づく。 $\mathbb{E}[X]=\mu$, $\mathrm{Var}(X)=\sigma^2$。

3. **指数分布** $\mathrm{Exp}(\lambda)$: $f(x)=\lambda e^{-\lambda x}$、$x>0$。 待ち時間のモデル。無記憶性 $\mathrm{P}(X>s+t\mid X>s)=\mathrm{P}(X>t)$ を持つ。 $\mathbb{E}[X]=1/\lambda$, $\mathrm{Var}(X)=1/\lambda^2$。

4. **ガンマ分布** $\mathrm{Gamma}(\alpha,\beta)$: $f(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$、$x>0$。 指数分布の一般化。$\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}e^{-t}\,dt$ はガンマ関数である。 $\mathbb{E}[X]=\alpha/\beta$, $\mathrm{Var}(X)=\alpha/\beta^2$。

5. **ベータ分布** $\mathrm{Beta}(\alpha,\beta)$: $f(x)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$、$0<x<1$。 $B(\alpha,\beta)=\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ はベータ関数である。 確率や割合のモデリングに用いられ、 ベルヌーイ分布の共役事前分布でもある。 $\mathbb{E}[X]=\alpha/(\alpha+\beta)$, $\mathrm{Var}(X)=\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$。

6. **カイ二乗分布** χ^2_n: $\mathrm{Gamma}(n/2,1/2)$ の特殊ケース。 $Z_1,\dots,Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ のとき $\sum_{i=1}^n Z_i^2\sim\chi^2_n$。 検定統計量の分布として中心的な役割を果たす（巻1第4章）。

7. **t 分布** t_n: $Z\sim\mathcal{N}(0,1)$, $V\sim\chi^2_n$ が独立のとき $T=Z/\sqrt{V/n}\sim t_n$。 自由度 n が大きいとき正規分布に近づくが、 裾が重い（自由度 $n=1$ はコーシー分布に一致する）。

8. **F 分布** $F_{m,n}$: $U\sim\chi^2_m$, $V\sim\chi^2_n$ が独立のとき $(U/m)/(V/n)\sim F_{m,n}$。 分散分析や回帰分析の検定で用いられる（巻1第4章、巻2第3章）。

1.5.4　多変量正規分布

定義 1.5.5 (多変量正規分布)**.** d 次元確率ベクトル $\boldsymbol{X}=(X_1,\dots,X_d)^\top$ が **多変量正規分布** $\mathcal{N}_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$ に従うとは、 $\boldsymbol{\Sigma}$ が正定値のとき密度関数が

$$f(\boldsymbol{x})=(2\pi)^{-d/2}\,|\boldsymbol{\Sigma}|^{-1/2}\exp\!\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

で与えられることをいう。 ここで $\boldsymbol{\mu}\in\mathbb{R}^d$ は平均ベクトル、 $\boldsymbol{\Sigma}\in\mathbb{R}^{d\times d}$ は共分散行列、 $|\boldsymbol{\Sigma}|$ は $\boldsymbol{\Sigma}$ の行列式である。

読み下し

多変量正規分布は1次元の正規分布を d 次元に拡張したものである。 密度関数の指数部分 $(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ は**マハラノビス距離**の二乗であり、 $\boldsymbol{\mu}$ を中心とする楕円体上で等高線をなす。 $\boldsymbol{\Sigma}$ が対角行列のとき各成分は独立であり、 1次元正規分布の直積に分解される。

命題 1.5.6 (多変量正規分布の基本性質)**.** $\boldsymbol{X}\sim\mathcal{N}_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$ のとき：

(i) 任意の $\boldsymbol{a}\in\mathbb{R}^d$ に対して $\boldsymbol{a}^\top\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{a}^\top\boldsymbol{\mu},\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a})$（線形結合は1次元正規）。

(ii) *行列 $\boldsymbol{B} \in \mathbb{R}^{m\times d}$ に対して $\boldsymbol{BX} \sim \mathcal{N}_m(\boldsymbol{B\mu}, \boldsymbol{B\Sigma B}^\top)$（アフィン変換で正規性が保たれる）。*

(iii) *多変量正規分布では、無相関と独立が同値である。*

1.6　独立性

直観的には、独立とは「一方の結果が他方の結果に影響しない」ことである。 まず2つの事象の場合に定義し、次に一般の場合に拡張する。

定義 1.6.1 (2事象の独立性)**.** 事象 $A, B \in \mathcal{F}$ が**独立**であるとは、

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$$

が成り立つことをいう。

読み下し

独立性の定義は「A と B が同時に起きる確率が、 それぞれの確率の積に等しい」ことを要求する。 条件付き確率 $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$ を用いると、 独立性は $\mathrm{P}(A \mid B) = \mathrm{P}(A)$（$B$ が起きたという情報が A の確率を変えない）と読み替えられる。

一般の n 個の事象に拡張する際は、 ペアワイズ独立（任意の2つが独立）では不十分であることに注意が必要である。

定義 1.6.2 (事象の独立性（一般）)**.** 事象 $A_1, \ldots, A_n$ が**（相互に）独立**であるとは、 任意の添字の部分集合 $S \subset \{1, \ldots, n\}$（$|S| \geq 2$）に対して

$$\mathrm{P}\left(\bigcap_{i\in S} A_i\right) = \prod_{i\in S} \mathrm{P}(A_i)$$

が成り立つことをいう。

注意 1.6.3. $n = 3$ の場合を例にとると、3つの事象 A_1, A_2, A_3 の相互独立性には

$$\begin{aligned}
&\mathrm{P}(A_1 \cap A_2) = \mathrm{P}(A_1)\mathrm{P}(A_2),\\
&\mathrm{P}(A_1 \cap A_3) = \mathrm{P}(A_1)\mathrm{P}(A_3),\\
&\mathrm{P}(A_2 \cap A_3) = \mathrm{P}(A_2)\mathrm{P}(A_3),\\
&\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3)
\end{aligned}$$

の4つの等式すべてが必要である。 上3つ（ペアワイズ独立）だけでは最後の等式は従わない。

定義 1.6.4 (確率変数の独立性)**.** 確率変数 $X_1, \ldots, X_n$ が独立であるとは、 任意のボレル集合 $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$ に対して

$$\mathrm{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} \mathrm{P}(X_i \in B_i)$$

が成り立つことをいう。 さらにすべてが同一の分布に従うとき、 **独立同一分布**（independent and identically distributed, i.i.d.） であるといい、$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$ と書く。

読み下し

確率変数の独立性は「任意の事象の組み合わせについて 確率が積に分解する」ことを要求する。 等価な条件として、同時分布関数が周辺分布関数の積に分解すること、 すなわち $F_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$ がすべての $(x_1,\dots,x_n)$ で成り立つこと、がある。 密度関数が存在する場合は同時密度が各周辺密度の積に分解することも同値である。

注意 1.6.5 (無相関と独立の違い). 無相関性（$\mathrm{Cov}(X,Y) = 0$）は独立性の必要条件ではあるが十分条件ではない。 たとえば $X \sim \mathcal{N}(0,1)$ とし $Y = X^2$ とおくと、 $\mathrm{Cov}(X,Y) = \mathbb{E}[X^3] = 0$（正規分布の奇数次モーメントは 0） だが X と Y は明らかに独立ではない （$X = 2$ と知れば $Y = 4$ が確定する）。

ただし、命題1.5.6(iii) で見たように、 多変量正規分布に限っては無相関と独立が同値である。 この例外こそ、正規分布が際立って扱いやすい理由の1つである。

注意 1.6.6 (条件付き独立への入口). 統計モデルでは、独立性はしばしば **条件付き独立**の形で現れる。 3つの確率変数 X, Y, Z について、 任意のボレル集合 A, B に対し

$$\mathrm{P}(X \in A, Y \in B \mid Z) = \mathrm{P}(X \in A \mid Z)\mathrm{P}(Y \in B \mid Z)$$

が a.s. で成り立つとき、 X と Y は Z を与えたもとで条件付き独立であるという。

これは「共変量や潜在変数を固定すると依存がほどける」という状況を表す概念であり、回帰、因果推論、グラフィカルモデルではこの見方が繰り返し現れる。 本巻ではまず独立性そのものを基礎として押さえ、 後続巻では「何を条件に入れると独立が近づくか」を考える。

1.7　確率不等式

確率論・統計学では、確率変数の正確な分布が わからなくても何らかの上界が欲しい場面が頻繁にある。 本節で紹介する不等式群はそのための基本的な道具であり、 第2章の大数の法則から、 巻4の高次元統計に至るまで繰り返し用いられる。

定理 1.7.1 (マルコフの不等式). *非負確率変数 $X \geq 0$ と $a > 0$ に対して*

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

読み下し

「非負の確率変数が閾値 a 以上になる確率は、 期待値を a で割った値で上から抑えられる」。 仮定は $X \geq 0$ だけという非常に弱い条件だが、 その分だけ上界は緩い（タイトではない）。

Proof. $X \geq a \cdot \mathbf{1}\{X \geq a\}$ に注意する（$\mathbf{1}\{\cdot\}$ は指示関数）。 両辺の期待値をとると $\mathbb{E}[X] \geq a \cdot \mathbb{E}[\mathbf{1}\{X \geq a\}] = a \cdot \mathrm{P}(X \geq a)$。 □

定理 1.7.2 (チェビシェフの不等式)**.** $\mathbb{E}[X^2] < \infty$ *のとき、* $\varepsilon > 0$ *に対して*

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

読み下し

「確率変数が平均から ε 以上離れる確率は、 分散を ε^2 で割った値以下」。 これはマルコフの不等式を $(X - \mathbb{E}[X])^2$ に適用した系であり、 分散の情報を使うぶんだけマルコフよりタイトな上界を与える。

Proof. $Y = (X - \mathbb{E}[X])^2$ は非負確率変数であり、 $\mathrm{P}(|X - \mathbb{E}[X]| \geq \varepsilon) = \mathrm{P}(Y \geq \varepsilon^2)$ にマルコフの不等式を適用すると $\mathrm{P}(Y \geq \varepsilon^2) \leq \mathbb{E}[Y]/\varepsilon^2 = \mathrm{Var}(X)/\varepsilon^2$。 □

定理 1.7.3 (イェンセンの不等式)**.** φ *が凸関数で* $\mathbb{E}[|X|] < \infty$ *ならば*

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

φ *が凹関数ならば不等号が逆転する。*

読み下し

「凸関数の期待値は、期待値の凸関数以上」。 たとえば $\varphi(x) = x^2$（凸）ならば $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$ となり、 $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$ が直ちに従う。 $\varphi(x) = \log x$（凹）ならば $\log(\mathbb{E}[X]) \geq \mathbb{E}[\log X]$ となり、 これは情報理論でKLダイバージェンスの非負性を示す際に用いられる。

定理 1.7.4 (ヘルダーの不等式)**.** $p, q > 1$、$1/p + 1/q = 1$ *のとき*

$$\mathbb{E}[|XY|] \leq (\mathbb{E}[|X|^p])^{1/p} \cdot (\mathbb{E}[|Y|^q])^{1/q}.$$

$p = q = 2$ *の場合は***コーシー・シュワルツの不等式***：*

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}$$

となる。

読み下し

ヘルダーの不等式は2つの確率変数の積の期待値を、 各々の L^p ノルム （$\|X\|_p = (\mathbb{E}[|X|^p])^{1/p}$： $|X|^p$ の期待値の p 乗根） の積で上から抑える。ここで L^p ノルムとは、 p 乗の期待値が有限な確率変数の空間における「大きさ」の尺度である。 コーシー・シュワルツの不等式は 相関係数が $[-1, 1]$ に収まることの証明に直接用いられる。

実務ポイント

確率不等式の実務での使い分け：

- **マルコフ**: 期待値しかわからないとき。最も粗いが最も汎用的。
- **チェビシェフ**: 分散までわかるとき。オンラインアルゴリズムの 信頼区間の設計などで使われる。
- **イェンセン**: 凸性を利用した理論的な導出。KLダイバージェンス、 相互情報量、EMアルゴリズムの導出の基盤。
- **コーシー・シュワルツ**: 2変量の関係の上界。相関の有界性の証明。

第2章では、これらに加えて **ヘフディングの不等式**（より高精度な集中不等式）を学ぶ。

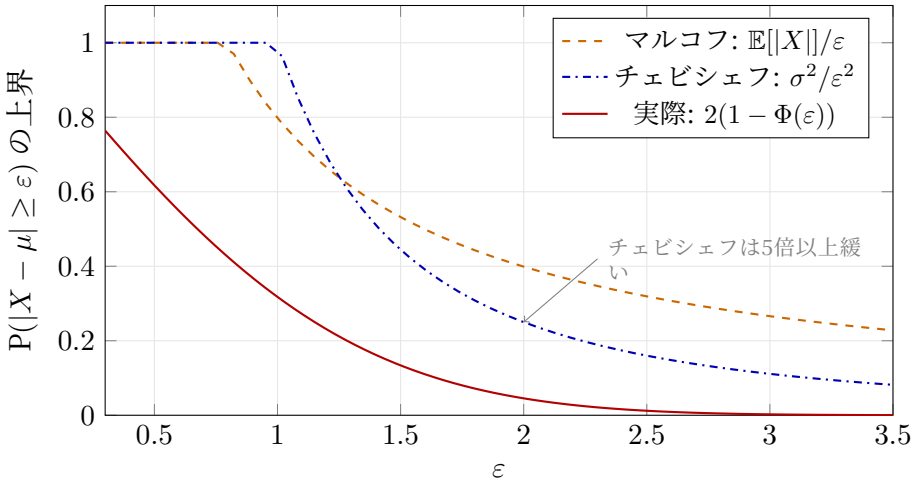


図 1.5: $X \sim \mathcal{N}(0,1)$ に対する確率不等式の比較。 マルコフの上界（$\mathbb{E}[|X|]/\varepsilon$、ただし $\mathbb{E}[|X|] = \sqrt{2/\pi}$）、 チェビシェフの上界（$\sigma^2/\varepsilon^2 = 1/\varepsilon^2$）、 実際の確率 $2(1-\Phi(\varepsilon))$ を重ねた。 分布の情報を使うほどタイトな上界が得られるが、 いずれも分布非依存の汎用性とのトレードオフである。

1.8　コード例：分布の可視化と確率不等式の検証

本節では、主要な確率分布の形状と確率不等式のタイト度を Pythonコードで確認する。

コード例: 主要な分布の密度関数

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

x = np.linspace(-4, 4, 400)

fig, axes = plt.subplots(1, 3, figsize=(12, 3))
# 正規分布
for mu, s in [(0,1), (0,0.5), (1,1)]:
    axes[0].plot(x, stats.norm.pdf(x, mu, s),
```

```
                  label=f'N({mu},{s**2})')
axes[0].set_title('Normal'); axes[0].legend()

# ガンマ分布
xp = np.linspace(0, 10, 400)
for a, b in [(1,1), (2,1), (5,1)]:
    axes[1].plot(xp, stats.gamma.pdf(xp, a, scale=1/b),
                  label=f'Gamma({a},{b})')
axes[1].set_title('Gamma'); axes[1].legend()

# ベータ分布
xb = np.linspace(0, 1, 400)
for a, b in [(0.5,0.5), (2,5), (5,2)]:
    axes[2].plot(xb, stats.beta.pdf(xb, a, b),
                  label=f'Beta({a},{b})')
axes[2].set_title('Beta'); axes[2].legend()

plt.tight_layout(); plt.savefig('distributions.pdf')
```

コード例: チェビシェフの不等式のタイト度

```
import numpy as np
import matplotlib.pyplot as plt

# チェビシェフの上界と実際の確率を比較する
rng = np.random.default_rng(42)
X = rng.standard_normal(100000)  # N(0,1)
mu, sigma2 = 0, 1

epsilons = np.linspace(0.5, 3.0, 20)
actual = [np.mean(np.abs(X - mu) >= eps) for eps in epsilons]
chebyshev = [sigma2 / eps**2 for eps in epsilons]

plt.figure(figsize=(6, 3.5))
plt.plot(epsilons, actual, 'o-', label='actual P(|X-mu|>=eps)')
plt.plot(epsilons, chebyshev, 's--', label='Chebyshev upper bound')
plt.xlabel('epsilon'); plt.ylabel('Probability')
plt.legend(); plt.title('Chebyshev inequality: tightness')
plt.tight_layout(); plt.savefig('chebyshev.pdf')
```

読み下し

上のコードを実行すると、チェビシェフの不等式の上界が 実際の確率よりもかなり大きい（緩い）ことが視覚的にわかる。 たとえば $\varepsilon = 2$ のとき、 $\mathcal{N}(0,1)$ の実際の確率は $\mathrm{P}(|X| \geq 2) \approx 0.046$ だが、 チェビシェフの上界は $1/4 = 0.25$ であり、5倍以上の差がある。 よりタイトな上界が必要な場合は、 分布の情報をさらに活用する集中不等式（第2章）を用いる。

実務ポイント

Monte Carlo 法は、 乱数を発生させて確率や積分を経験平均で近似する方法の総称である。 たとえば事象 A の確率は、独立な標本 $X_1, \dots, X_m$ を用いて

$$\hat{p}_m = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{X_i \in A\}$$

と近似できる。第2章で学ぶ大数の法則は $\hat{p}_m$ が $\mathrm{P}(X \in A)$ に近づく理由を与え、 中心極限定理は近似誤差の大きさを評価する基盤を与える。 現代の統計計算で重要な重要度サンプリング、MCMC、逐次 Monte Carlo も、 この同じ原理の延長にある。

主要な結果

重要結果

本章の核心的な結論：

1. **確率空間**$(\Omega, \mathcal{F}, \mathrm{P})$は、 非可算な標本空間でも確率を矛盾なく扱うための最小限の枠組みである。 「何に確率を割り当てるか」は σ-加法族が決める。
2. **確率変数**は可測関数として定義され、 分布関数・密度関数・変数変換はこの言語の上で統一的に記述できる。
3. **期待値と条件付き期待値**はルベーグ積分の言葉で定義される。 とくに条件付き期待値は、平均を取り直す操作ではなく、 与えられた情報に基づく最良予測として理解できる。
4. **特性関数と主要分布族**を通じて、 分布の比較・和の分布・指数型分布族による整理が可能になる。 これは後続章の極限定理と推定論の共通基盤である。
5. **独立性と確率不等式**は、 収束理論・検定・漸近近似・統計計算を支える基本道具である。 とくにマルコフ・チェビシェフ・イェンセンの不等式は、 情報が限られた状況での安全側評価を与える。 条件付き独立と Monte Carlo 法は、この基礎の自然な延長にある。

1.9　演習問題

理論問題

演習問題 1.1. $(\Omega, \mathcal{F}, \mathrm{P})$ を確率空間、$\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ とする。 **Borel–Cantelliの第一補題**を証明せよ：$\sum_{n=1}^{\infty} \mathrm{P}(A_n) < \infty$ ならば $\mathrm{P}(\limsup_{n\to\infty} A_n) = 0$。 ここで $\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$ は「無限回起きる事象」を意味する。

ヒント: 確率測度の上からの連続性（命題1.1.7(vi)）を用いよ。

演習問題 1.2. $X \sim \mathcal{N}(\mu, \sigma^2)$ の特性関数が $\varphi_X(t) = \exp(i\mu t - \sigma^2 t^2/2)$ であることを示せ。

ヒント: まず $\mu = 0, \sigma = 1$ の場合に $\mathbb{E}[e^{itX}] = \int e^{itx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$ を平方完成で計算し、一般の場合は $X = \mu + \sigma Z$（$Z \sim \mathcal{N}(0,1)$）を用いよ。

演習問題 1.3. $X_1, \dots, X_n$ が独立で $X_i \sim \mathrm{Pois}(\lambda_i)$ のとき、$S = X_1 + \cdots + X_n \sim \mathrm{Pois}(\lambda_1 + \cdots + \lambda_n)$ であることを特性関数を用いて示せ。

演習問題 1.4. 条件付き期待値の繰り返しの法則 $\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}] \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}]$（$\mathcal{G} \subset \mathcal{H}$）を定義1.3.4の2条件から証明せよ。

演習問題 1.5. ベータ分布 $\mathrm{Beta}(\alpha, \beta)$ の密度

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \qquad 0 < x < 1$$

を考える。

(a) $\int_0^1 f(x)\, dx = 1$ を確認せよ。

(b) $\mathbb{E}[X]$ と $\mathrm{Var}(X)$ を求めよ。

ヒント: ベータ関数の関係式 $B(\alpha+1,\beta) = \frac{\alpha}{\alpha+\beta} B(\alpha,\beta)$, $B(\alpha+2,\beta) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} B(\alpha,\beta)$ を用いよ。

計算・実装問題

数値実験を含む問題では、(1) 設定（分布、パラメータ、反復回数、乱数シード）、(2) 作成した図表または表、(3) 比較指標、(4) 結果から読める一言考察、の4点を答えに含めよ。

演習問題 1.6. $X \sim \mathcal{N}(\mu, \sigma^2)$ のモーメント母関数が $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$ であることを示し、$M_X'(0)$ と $M_X''(0)$ から $\mathbb{E}[X]$ と $\mathbb{E}[X^2]$ を求めることで $\mathrm{Var}(X) = \sigma^2$ を確認せよ。

演習問題 1.7. $X \sim \mathrm{Gamma}(\alpha, \beta)$ に対して変数変換 $Y = 1/X$ の密度関数を求めよ。$\alpha = 3$, $\beta = 2$ の場合に R または Python で Y のヒストグラムと理論密度を重ねてプロットし、結果を確認せよ。

演習問題 1.8. $X \sim \mathrm{Exp}(\lambda)$（$\lambda = 1$）とする。$n = 10000$ 個のサンプルを生成し、各 $\varepsilon \in \{0.5, 1.0, 1.5, 2.0\}$ に対して

(a) マルコフの不等式の上界 $\mathbb{E}[|X - \mathbb{E}[X]|]/\varepsilon$

(b) チェビシェフの不等式の上界 $\mathrm{Var}(X)/\varepsilon^2$

(c) シミュレーションによる実際の確率 $\hat{\mathrm{P}}(|X - \mathbb{E}[X]| \geq \varepsilon)$

を比較する表を作成せよ。どの ε で不等式が最もタイトか考察せよ。

演習問題 1.9. 以下の分布を指数型分布族の標準形に書き直し、自然パラメータ η、十分統計量 $T(x)$、対数分配関数 $A(\eta)$、基底測度 $h(x)$ をそれぞれ特定せよ：

(a) 二項分布 $\mathrm{Bin}(n, p)$（n は既知）

(b) 指数分布 $\mathrm{Exp}(\lambda)$

(c) ガンマ分布 Gamma(α, β)（α は既知）

演習問題 1.10. $\boldsymbol{X} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$ ただし $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ とする。 $\rho = 0, 0.5, 0.9$ の各場合について Python でサンプルを生成し、 散布図をプロットせよ。 $\rho = 0$ のとき散布図が円形に、 $\rho \to 1$ のとき直線的になることを確認せよ。

略解の指針

ここでは解答の骨格だけを示す。証明や計算の細部は自分で埋めることを前提とする。

- **演習1.1** 使う道具: 上からの連続性。 最初の1手: $B_n = \bigcup_{k=n}^{\infty} A_k$ と置くと $B_n \downarrow \limsup A_n$ である。 途中の要点: $\mathrm{P}(B_n) \le \sum_{k=n}^{\infty} \mathrm{P}(A_k)$ が成り立ち、右辺は $n \to \infty$ で 0 に行く。 最終形: $\mathrm{P}(\limsup A_n) = \lim_n \mathrm{P}(B_n) = 0$。

- **演習1.2** 使う道具: 平方完成とアフィン変換。 最初の1手: まず $Z \sim \mathcal{N}(0,1)$ について $\mathbb{E}[e^{itZ}]$ を積分で計算する。 途中の要点: 指数部を $-(z - it)^2/2 - t^2/2$ と平方完成すると、積分値は $e^{-t^2/2}$ になる。 最終形: $X = \mu + \sigma Z$ を用いて $\varphi_X(t) = e^{i\mu t - \sigma^2 t^2/2}$ を得る。

- **演習1.3** 使う道具: 独立性と特性関数の積。 最初の1手: $\varphi_{X_i}(t) = \exp\{\lambda_i(e^{it} - 1)\}$ を書く。 途中の要点: 独立性より $\varphi_S(t) = \prod_i \varphi_{X_i}(t)$ であり、指数が和にまとまる。 最終形: $\varphi_S(t) = \exp\{(\lambda_1 + \cdots + \lambda_n)(e^{it} - 1)\}$ なので $S \sim \mathrm{Pois}(\lambda_1 + \cdots + \lambda_n)$。

- **演習1.4** 使う道具: 条件付き期待値の定義。 最初の1手: $\mathbb{E}[X \mid \mathcal{H}]$ が $\mathcal{H}$-可測であり、$\mathcal{G} \subset \mathcal{H}$ だから $\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}] \mid \mathcal{G}]$ は $\mathcal{G}$-可測である。 途中の要点: 任意の $G \in \mathcal{G}$ に対し $\int_G \mathbb{E}[\mathbb{E}[X \mid \mathcal{H}] \mid \mathcal{G}]\, d\mathrm{P} = \int_G \mathbb{E}[X \mid \mathcal{H}]\, d\mathrm{P} = \int_G X\, d\mathrm{P}$ を示す。 最終形: 定義の2条件を満たすので $\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}] \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}]$。

- **演習1.5** 使う道具: ベータ関数の漸化式。 最初の1手: 正規化は $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = B(\alpha, \beta)$ を使えば終わる。 途中の要点: $\mathbb{E}[X] = B(\alpha+1, \beta)/B(\alpha, \beta)$, $\mathbb{E}[X^2] = B(\alpha+2, \beta)/B(\alpha, \beta)$ と書く。 最終形: $\mathbb{E}[X] = \alpha/(\alpha+\beta)$, $\mathrm{Var}(X) = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$。

- **演習1.6** 使う道具: 正規分布の指数積分。 最初の1手: $X = \mu + \sigma Z$ と置いて $M_X(t) = e^{\mu t}\mathbb{E}[e^{\sigma t Z}]$ に分ける。 途中の要点: 標準正規の mgf は $\mathbb{E}[e^{sZ}] = e^{s^2/2}$ である。 最終形: $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$, $M_X'(0) = \mu$, $M_X''(0) = \mu^2 + \sigma^2$ から $\mathrm{Var}(X) = \sigma^2$。

- **演習1.7** 使う道具: 変数変換。 最初の1手: $y = 1/x$, $x = 1/y$, $|dx/dy| = y^{-2}$ を使う。 途中の要点: 密度は $f_Y(y) = f_X(1/y) y^{-2}$ によって与えられる。 最終形: $f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} e^{-\beta/y}$, $y > 0$。 数値実験ではヒストグラムと理論密度の重ね描きで確認する。

- **演習1.8** 使う道具: マルコフ、チェビシェフ、経験確率。 最初の1手: Exp(1) では $\mathbb{E}[X] = 1$, $\mathrm{Var}(X) = 1$ に加え、 $\mathbb{E}|X - 1| = 2/e$ を計算する。 途中の要点: したがって上界は $(2/e)/\varepsilon$ と $1/\varepsilon^2$ になる。 最終形: どちらの上界も安全側だが、ε が大きくなるほど実際の確率との差は縮まりやすい。

- **演習1.9** 使う道具: 指数型分布族の標準形への書き換え。 最初の1手: 二項分布は $\eta = \log\{p/(1-p)\}$ を用いて $\exp\{\eta x - n\log(1 + e^\eta)\}\binom{n}{x}$ と書く。 途中の要点: 指数分布は

$\eta = -\lambda < 0$, ガンマ分布は $\eta = -\beta < 0$ と置けばよい。 最終形: それぞれ $T(x)$, $A(\eta)$, $h(x)$ を読み取り、 自然パラメータ空間が $\eta < 0$ のような制約を持つことも確認する。

- **演習1.10** 使う道具: 共分散行列の幾何学的解釈。 最初の1手: 各 ρ について平均 0、共分散 Σ の2次元正規標本を生成する。 途中の要点: 固有値は $1+\rho$, $1-\rho$ なので、ρ が大きいほど楕円が細くなる。 最終形: $\rho = 0$ ではほぼ円形、$\rho = 0.9$ では対角線方向に細長い散布図になる。

次章への橋渡し

本章では確率空間と確率変数の基盤を築き、 期待値・条件付き期待値・主要な分布族・確率不等式という 確率論の基本的な道具を整備した。

しかし統計学では、1つの確率変数ではなく *確率変数の列* $X_1, X_2, \ldots$ の振る舞いが本質的に重要である。 標本サイズ n を大きくしたとき、 標本平均 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ は 真の期待値 $\mathbb{E}[X]$ に「近づく」はずだ —— だが、 ランダムな量が確定的な値に「近づく」とは正確には何を意味するのか？

次章（第2章「収束理論と極限定理」）では、 概収束・確率収束・L^p 収束・分布収束という4つの収束概念を導入し、 それぞれの関係を明らかにする。 そして**大数の法則**と**中心極限定理**—— 統計学の二本柱ともいうべき極限定理——を証明する。 本章で学んだチェビシェフの不等式と特性関数が、 これらの証明で決定的な役割を果たす。

参考文献ノート

確率論の測度論的基礎については Durrett (2019) *Probability: Theory and Examples* が構成から証明まで自己完結的な標準的教科書であり、 Williams (1991) *Probability with Martingales* は 簡潔で読みやすい入門書として定評がある。 統計学の視点から確率論をまとめたものとしては Keener (2010) *Theoretical Statistics* の第1章、 Schervish (1995) *Theory of Statistics* の付録B が有用である。 指数型分布族の理論は Brown (1986) *Fundamentals of Statistical Exponential Families* が 決定的な参考文献であり、 本書の後の章（巻2第1章、巻2第3章）でもこの文献を参照する。 日本語の教科書では 吉田朋広 (2006)『数理統計学』が 測度論的確率論から統計的推測まで体系的に扱っている。 より実務寄りの補助線としては、 Fernandez-Granda (2025) *Probability and Statistics for Data Science* が 具体例・演習・コードを通じて確率論と統計的推論を接続している。

第2章

収束理論と極限定理

問いと学習目標

この章で答える問い

- 「標本平均が母平均に近づく」とは、数学的に何を意味するのか？
- 確率的な収束にはどのような種類があり、それぞれどう違うのか？
- なぜ統計量の分布は標本サイズが大きくなると正規分布に近づくのか？
- 正規近似はどれくらい信頼できるのか？その誤差はどう評価できるか？

読み終えたらできるようになること

1. 概収束・確率収束・L^p収束・分布収束を定義し、その含意関係を説明できる。
2. 大数の法則（弱・強）が保証する内容を述べ、標本平均の一致性を示せる。
3. 中心極限定理を用いて標本平均の近似分布を導出できる。
4. デルタ法を用いて変換された統計量の漸近分布を求められる。
5. スラツキーの補題を使い、分散未知の場合の推論を正当化できる。

直観的理解

統計学は有限のデータから無限のパターンを推測する学問である。「標本が大きくなると推定量は真の値に近づく」――この直観は正しいが、「近づく」の意味を厳密にしなければ、その上に推論を組み立てることはできない。
本章では、確率変数の列が「収束する」ことの4つの定式化を導入し、それぞれの強弱関係を明らかにする。その上で、統計学の二大柱である大数の法則と中心極限定理を証明し、デルタ法やスラツキーの補題といった漸近論の実用道具を整備する。本章の内容は、第3章以降のすべての推定・検定・信頼区間の理論的基盤となる。

2.1　収束の4つの概念

確率変数列 $\{X_n\}_{n=1}^{\infty}$ と確率変数 X がすべて同一の確率空間 $(\Omega, \mathcal{F}, \mathrm{P})$ 上で定義されているとする。「X_n が X に近づく」ことを定式化する方法は複数あり、それぞれ異なる観点から「近さ」を測る。ここでは4つの収束概念を導入する。

2.1.1　概収束

定義 2.1.1 (概収束)**.** $X_n \xrightarrow{\text{a.s.}} X$ とは、

$$\mathrm{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}) = 1$$

が成り立つことをいう。

読み下し

概収束（almost sure convergence）は最も強い収束概念の一つである。「確率1の事象の上で、各 ω ごとに実数列 $X_n(\omega)$ が $X(\omega)$ に収束する」ことを要求する。「ほぼ確実に、すべての標本パスが収束する」と読む。

2.1.2　確率収束

定義 2.1.2 (確率収束)**.** $X_n \xrightarrow{p} X$ とは、すべての $\varepsilon > 0$ に対して

$$\lim_{n \to \infty} \mathrm{P}(|X_n - X| > \varepsilon) = 0$$

が成り立つことをいう。

読み下し

確率収束は「X_n が X から ε 以上離れる確率が、n を大きくすればいくらでも小さくできる」ことを意味する。概収束との違いは、各 ω ごとの収束は要求せず、確率的に「大きなずれが稀になる」ことだけを要求する点にある。

例 2.1.3 (概収束と確率収束の直観的な違い)**.** 射撃場の比喩で考えよう。射手が練習を重ねるにつれ、的の中心に近づくとする。

概収束は「ほぼ確実に、ある回数以降はずっと的の中心付近に当たり続ける」こと。**確率収束**は「n 回目の射撃が大外れする確率は n とともに減る」こと。

確率収束では、どれだけ上達しても稀に大外れする可能性が（全体として）残りうるが、概収束ではそのような大外れはほぼ確実に有限回で終わる。

2.1.3　L^p 収束

定義 2.1.4 (L^p 収束)**.** $p \geq 1$ に対して $X_n \xrightarrow{L^p} X$ とは、

$$\lim_{n \to \infty} \mathbb{E}[|X_n - X|^p] = 0$$

が成り立つことをいう。$p = 2$ の場合を**平均二乗収束**と呼ぶ。

読み下し

L^p 収束は「X_n と X の差の p 乗の期待値がゼロに近づく」ことを意味する。$p = 2$ の場合は、差の分散と差の平均の二乗の和がゼロに近づくことに相当する。L^p 収束は、大きなずれに対して概収束・確率収束よりも厳しいペナルティを課す。

2.1.4　法則収束（分布収束）

定義 2.1.5 (法則収束（分布収束）). $X_n \xrightarrow{d} X$ とは、F_X の連続点であるすべての x において

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

が成り立つことをいう。

読み下し

法則収束は最も弱い収束概念であり、確率変数の値そのものではなく「分布の形」が近づくことだけを要求する。X_n と X は異なる確率空間上で定義されていてもよい。中心極限定理の主張はこの意味での収束である。

2.1.5　収束概念の関係

4つの収束概念の間には、以下の含意関係が成り立つ。

定理 2.1.6 (収束概念の関係). *以下の含意が成り立つ：*

$$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$

$$X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$

逆は一般には成り立たない。ただし $X_n \xrightarrow{d} c$（c は定数）ならば $X_n \xrightarrow{p} c$ が成り立つ。

読み下し

図2.1に示した通り、収束の強弱関係は「概収束と L^p 収束がそれぞれ確率収束を含意し、確率収束は分布収束を含意する」とまとめられる。分布収束が最も弱く、概収束と L^p 収束が強い。ただし概収束と L^p 収束の間には直接の含意関係がない（例2.1.9と例2.1.8がそれぞれ反例を示す）。
統計学で最もよく使うのは確率収束（推定量の一致性）と分布収束（漸近分布の導出）である。

概収束 $\implies$ 確率収束の証明. 背理法で示す。$X_n \xrightarrow{\text{a.s.}} X$ を仮定し、$\varepsilon > 0$ を固定する。

事象 $A_n = \{|X_n - X| > \varepsilon\}$ を考え、$B_N = \bigcup_{n\geq N} A_n$ とおくと $B_N \downarrow B_\infty = \bigcap_{N=1}^{\infty} B_N$。概収束の仮定より $\mathrm{P}(B_\infty) = 0$ である。確率測度の上からの連続性（第1章、命題1.1.7 (vi)）から $\mathrm{P}(A_n) \leq \mathrm{P}(B_n) \to \mathrm{P}(B_\infty) = 0$。 □

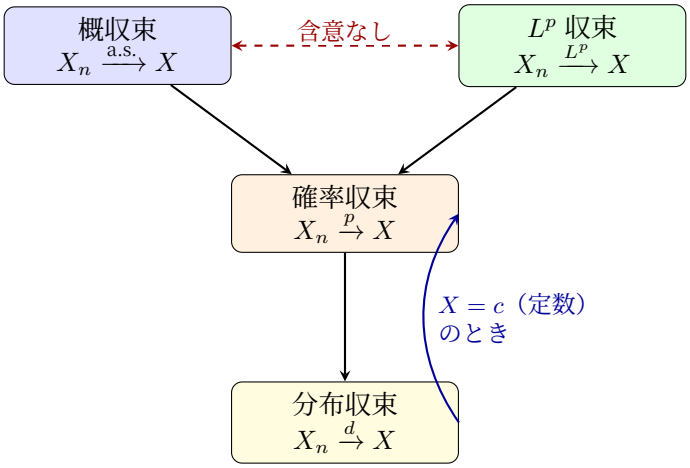


図 2.1: 収束概念の含意関係。実線矢印は含意を、赤い破線は直接の含意関係がないことを示す。 分布収束先が定数の場合に限り、確率収束が従う（青い矢印）。

ここで、逆が成り立たないことを具体例で確認しよう。

例 2.1.7 (確率収束するが概収束しない例：巡回する指示関数)**.** $\Omega = [0,1]$、P をルベーグ測度とする。 自然数 n を $n = 2^k + j$（$0 \le j < 2^k$）と書き、 $X_n = \mathbf{1}_{[j/2^k,(j+1)/2^k]}$ と定義する。

具体的に最初の数項を書き出すと：

$$X_1 = \mathbf{1}_{[0,1]}, \quad X_2 = \mathbf{1}_{[0,1/2]}, \quad X_3 = \mathbf{1}_{[1/2,1]},$$
$$X_4 = \mathbf{1}_{[0,1/4]}, \quad X_5 = \mathbf{1}_{[1/4,1/2]}, \quad X_6 = \mathbf{1}_{[1/2,3/4]}, \quad X_7 = \mathbf{1}_{[3/4,1]}, \quad \dots$$

各 X_n の台の幅は $2^{-k} \to 0$ であるから $\mathrm{P}(|X_n| > \varepsilon) \le 2^{-k} \to 0$ となり、$X_n \xrightarrow{p} 0$。

しかし各 $\omega \in [0,1]$ に対し、ω を含む区間 $[j/2^k, (j+1)/2^k]$ は無限個存在するため、$X_n(\omega) = 1$ となる n が無限に現れ、 $X_n(\omega) \to 0$ とはならない。よって $X_n \xrightarrow{\text{a.s.}} 0$ ではない。

例 2.1.8 (L^2 収束するが概収束しない例)**.** 上と同じ構成 $X_n = \mathbf{1}_{[j/2^k,(j+1)/2^k]}$ を考えると、$\mathbb{E}[X_n^2] = 2^{-k} \to 0$ であるから $X_n \xrightarrow{L^2} 0$ でもある。 しかし概収束しないことは変わらない。

例 2.1.9 (概収束するが L^1 収束しない例)**.** $\Omega = [0,1]$、P をルベーグ測度とし、 $X_n = n \cdot \mathbf{1}_{[0,1/n]}$ と定義する。

各 $\omega > 0$ に対し、$n > 1/\omega$ ならば $X_n(\omega) = 0$ であるから $X_n \xrightarrow{\text{a.s.}} 0$。しかし $\mathbb{E}[|X_n|] = n \cdot (1/n) = 1 \not\to 0$ であるから $X_n \xrightarrow{L^1} 0$ ではない。

これは裾の重い分布（大きな値を低確率でとる）が L^p 収束を妨げる例である。

実務ポイント

推定量の一致性（consistency）は通常、確率収束の意味で定義される：$\hat{\theta}_n \xrightarrow{p} \theta_0$。 漸近分布の導出には分布収束を使う：$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$。 概収束は確率的保証として最も強いが、 多くの統計的応用では確率収束で十分であることが多い。

2.2　Borel–Cantelliの補題と概収束

概収束を示すための強力な道具が Borel–Cantelli の補題である。第1章の演習問題1.9で第一補題を証明したが、ここではその結果を概収束の判定に応用する。

定理 2.2.1 (Borel–Cantelliの第一補題)**.** *事象列 $\{A_n\}_{n=1}^{\infty}$ に対して、$\sum_{n=1}^{\infty} \mathrm{P}(A_n) < \infty$ ならば*

$$\mathrm{P}\left(\limsup_{n\to\infty} A_n\right) = 0.$$

読み下し

$\limsup_{n\to\infty} A_n = \bigcap_{N=1}^{\infty} \bigcup_{n\geq N} A_n$ は「A_n が無限回起こる」事象（A_n infinitely often, 略して A_n i.o.）である。第一補題は「各事象の確率の総和が有限ならば、それらの事象が無限回起こる確率はゼロ」と読む。

定理 2.2.2 (Borel–Cantelliの第二補題)**.** *事象列 $\{A_n\}_{n=1}^{\infty}$ が**独立**で $\sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty$ ならば* $\mathrm{P}(\limsup_{n\to\infty} A_n) = 1$。

証明の方針. 各 N に対して $\mathrm{P}\left(\bigcap_{n=N}^{M} A_n^c\right) = \prod_{n=N}^{M}(1-\mathrm{P}(A_n))$ と独立性から書け、$1-x \leq e^{-x}$ を用いて $\prod(1-\mathrm{P}(A_n)) \leq \exp(-\sum \mathrm{P}(A_n)) \to 0$（$M\to\infty$）を示す。 □

命題 2.2.3 (概収束の Borel–Cantelli 判定法)**.** *$X_n \xrightarrow{a.s.} X$ であるための十分条件は、すべての $\varepsilon > 0$ に対して $\sum_{n=1}^{\infty} \mathrm{P}(|X_n - X| > \varepsilon) < \infty$ が成り立つことである。*

読み下し

概収束を示すには、$|X_n - X| > \varepsilon$ となる事象の確率を各 n について上から評価し、その総和が有限であることを Borel–Cantelli の第一補題で示せばよい。この方法は強大数の法則の証明の鍵となる。

例 2.2.4 (Borel–Cantelli による概収束の確認)**.** $X_1, X_2, \ldots$ が i.i.d. で $\mathbb{E}[X_1^4] < \infty$ とし、$\mu = \mathbb{E}[X_1]$ とする。チェビシェフの不等式の4次版を用いると

$$\mathrm{P}(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\mathbb{E}[(\bar{X}_n - \mu)^4]}{\varepsilon^4}.$$

独立性から $\mathbb{E}[(\bar{X}_n - \mu)^4] = O(n^{-2})$ と評価できるので、$\sum_{n=1}^{\infty} \mathrm{P}(|\bar{X}_n - \mu| > \varepsilon) < \infty$。Borel–Cantelli の第一補題より $\bar{X}_n \xrightarrow{a.s.} \mu$ が従う。

ただしこの証明は4次モーメントの存在を仮定しており、Kolmogorov の強大数の法則（1次モーメントのみで十分）よりも強い仮定を課している。

2.3　大数の法則

大数の法則は、標本平均が母平均に収束することを保証する定理群である。「弱」は確率収束、「強」は概収束の意味での収束を主張する。

定理 2.3.1 (弱大数の法則（WLLN）)**.** $X_1, X_2, \dots$ が *i.i.d.* で $\mathbb{E}[|X_1|] < \infty$ のとき、 $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ とおくと

$$\bar{X}_n \xrightarrow{p} \mathbb{E}[X_1].$$

読み下し

弱大数の法則は「標本平均 $\bar{X}_n$ は、標本サイズ n を大きくすると 母平均 $\mathbb{E}[X_1]$ に確率収束する」と読む。 つまり、$\bar{X}_n$ が母平均から大きくずれる確率は n とともにゼロに近づく。

証明（$\mathrm{Var}(X_1) < \infty$ の場合）. **証明の方針**：チェビシェフの不等式を使い、 $\bar{X}_n$ と μ のずれの確率を分散で上から押さえる。 i.i.d. の仮定から $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$ となることが鍵である。

$\mu = \mathbb{E}[X_1]$、$\sigma^2 = \mathrm{Var}(X_1) < \infty$ とおく。 i.i.d. より $\mathbb{E}[\bar{X}_n] = \mu$ かつ $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$。 チェビシェフの不等式（定理1.7.2）より、$\varepsilon > 0$ に対して

$$\mathrm{P}(\left|\bar{X}_n - \mu\right| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad (n \to \infty).$$

□

注意 2.3.2. 上の証明は分散の有限性を仮定したが、 $\mathbb{E}[|X_1|] < \infty$ のみの仮定でも WLLN は成立する。 この場合は特性関数を用いた証明が必要となる。

定理 2.3.3 (強大数の法則（SLLN, Kolmogorov）)**.** $X_1, X_2, \dots$ が *i.i.d.* で $\mathbb{E}[|X_1|] < \infty$ のとき、

$$\bar{X}_n \xrightarrow{a.s.} \mathbb{E}[X_1].$$

読み下し

強大数の法則は弱大数の法則よりも強い主張である。 「確率が高い」だけでなく「ほぼ確実に」収束することを保証する。 直観的には、$\bar{X}_n$ の標本パスは（確率1で）ある n 以降 母平均の任意の近傍にとどまり続ける。

証明の概略（4次モーメント有限の場合）. **証明の方針**：4次モーメントの評価と Borel–Cantelli の補題を組み合わせる。 $\sum_n \mathrm{P}(\left|\bar{X}_n - \mu\right| > \varepsilon) < \infty$ を示せば 命題2.2.3から概収束が従う。

$Y_i = X_i - \mu$ とおくと $\mathbb{E}[Y_i] = 0$ であり、

$$\mathbb{E}\left[(\bar{X}_n - \mu)^4\right] = \frac{1}{n^4}\mathbb{E}\left[\left(\sum_{i=1}^{n} Y_i\right)^4\right].$$

展開すると、独立性と $\mathbb{E}[Y_i] = 0$ から $\mathbb{E}[Y_i Y_j] = 0$（$i \neq j$）であり、 $\mathbb{E}[Y_i^2 Y_j^2] = \sigma^4$（$i \neq j$）、$\mathbb{E}[Y_i^4] = \mu_4$ と書ける。 整理すると

$$\mathbb{E}\left[(\bar{X}_n - \mu)^4\right] = \frac{n\mu_4 + 3n(n-1)\sigma^4}{n^4} = O(n^{-2}).$$

マルコフの不等式より $\mathrm{P}(|\bar{X}_n - \mu| > \varepsilon) \le C/n^2$。$\sum_{n=1}^{\infty} n^{-2} < \infty$ であるから Borel–Cantelli の第一補題より $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$。

注意：完全な証明（$\mathbb{E}[|X_1|] < \infty$ のみの仮定）は Kolmogorov の切断法（truncation method）を用い、X_i を $X_i^{(n)} = X_i \mathbf{1}\{|X_i| \le n\}$ で近似する。詳細は付録Aまたは Durrett (2019), Theorem 2.4.1 を参照。 □

例 2.3.4 (大数の法則の数値的確認)**.** $X_1, X_2, \ldots \sim \mathrm{Exp}(1)$（平均 $\mu = 1$）とする。$n = 10, 100, 1000, 10000$ に対する $\bar{X}_n$ の振る舞いを見ると：

n	10	100	1,000	10,000
$\bar{X}_n$ の典型的な値	0.7–1.3	0.9–1.1	0.97–1.03	0.99–1.01
$\sigma/\sqrt{n}$（標準誤差）	0.316	0.100	0.032	0.010

標準誤差が $1/\sqrt{n}$ の速度で減少し、$\bar{X}_n$ が $\mu = 1$ の周囲に集中していく。

コード例：大数の法則のシミュレーション（Python）

```
import numpy as np
import matplotlib.pyplot as plt

rng = np.random.default_rng(42)
n_max = 5000
X = rng.exponential(scale=1.0, size=n_max)  # Exp(1)

# 累積平均の計算
cumulative_mean = np.cumsum(X) / np.arange(1, n_max + 1)

plt.figure(figsize=(8, 4))
plt.plot(cumulative_mean, linewidth=0.8)
plt.axhline(y=1.0, color='r', linestyle='--', label=r'$\mu = 1$')
plt.xlabel('標本サイズ $n$')
plt.ylabel(r'$\bar{X}_n$')
plt.title('大数の法則：標本平均の収束')
plt.legend()
plt.tight_layout()
plt.show()
```

実務ポイント

大数の法則はデータサイエンスの多くの手法の理論的基盤である：

- **モンテカルロ法**：$\mathbb{E}[g(X)]$ の推定量 $n^{-1}\sum g(X_i)$ の一致性は SLLN が保証する。
- **交差検証**：テスト誤差の平均が汎化誤差に収束する根拠。
- **経験分布関数**：$\hat{F}_n(x) = n^{-1}\sum \mathbf{1}\{X_i \le x\}$ が真の分布関数に概収束する（Glivenko–Cantelli の定理）のも SLLN の系である。

ただし大数の法則は「最終的に」収束することを保証するだけで、有限標本でどれくらいの精度が得られるかは中心極限定理や集中不等式で評価する必要がある。

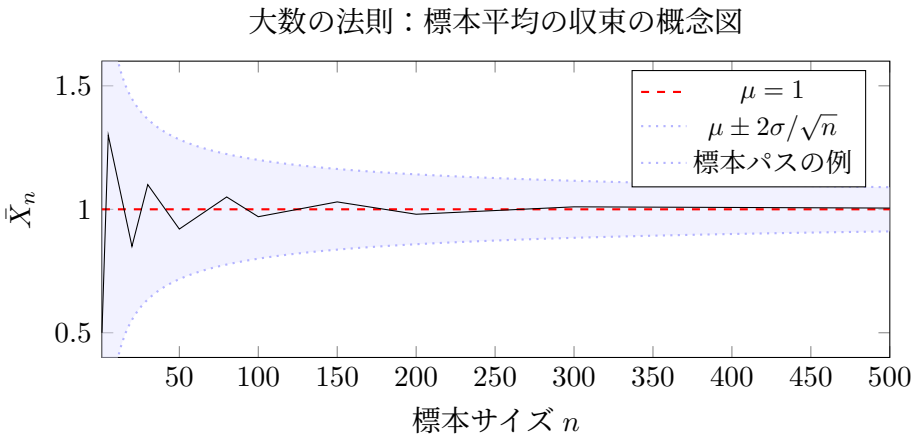


図 2.2: 大数の法則の概念図。Exp(1) からの標本平均 $\bar{X}_n$ は n の増加に伴い母平均 $\mu = 1$ に収束する。 青い帯は $\mu \pm 2\sigma/\sqrt{n}$ の範囲であり、$O(1/\sqrt{n})$ で縮小する。

2.4　中心極限定理

大数の法則は $\bar{X}_n \to \mu$ を保証するが、 $\bar{X}_n$ が μ のまわりにどのように分布するか——すなわち 「誤差の大きさと形状」——は教えてくれない。 中心極限定理（CLT）がこの問いに答える。

定理 2.4.1 (古典的中心極限定理（Lindeberg–Lévy）)**.** *$X_1, X_2, \ldots$ が i.i.d. で* $\mathbb{E}[X_1] = \mu$、$\mathrm{Var}(X_1) = \sigma^2 \in (0, \infty)$ のとき、

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

読み下し

中心極限定理は「i.i.d. 確率変数の標準化された標本平均は、 元の分布が何であっても（分散が有限であれば）標準正規分布に近づく」と読む。 $\sqrt{n}$ のスケーリングは本質的であり、標本平均の誤差が $O(1/\sqrt{n})$ であることを意味する。

証明（特性関数を用いる方法）. **証明の方針**： 標準化された確率変数の和の特性関数を計算し、 それが標準正規分布の特性関数 $e^{-t^2/2}$ に収束することを示す。 Lévy の連続性定理 （特性関数の各点収束は分布収束と同値）で結論を得る。

$Y_i = (X_i - \mu)/\sigma$ とおく。$\mathbb{E}[Y_i] = 0$、$\mathrm{Var}(Y_i) = 1$ であり、 Y_i の特性関数は $\varphi_Y(t) = 1 - t^2/2 + o(t^2)$ $(t \to 0)$ 。

$S_n = n^{-1/2} \sum_{i=1}^{n} Y_i$ の特性関数は 独立性より $\varphi_{S_n}(t) = [\varphi_Y(t/\sqrt{n})]^n$。

$\varphi_Y(t/\sqrt{n}) = 1 - t^2/(2n) + o(1/n)$ であるから、

$$\varphi_{S_n}(t) = \left[1 - \frac{t^2}{2n} + o(1/n)\right]^n \to e^{-t^2/2}.$$

最後のステップでは $(1 + a_n/n)^n \to e^a$ $(a_n \to a)$ を用いた。 右辺は $\mathcal{N}(0, 1)$ の特性関数であり、 Lévy の連続性定理（定理1.4.3の系）から $S_n \xrightarrow{d} \mathcal{N}(0, 1)$。 □

例 2.4.2 (CLT の数値的確認：コイン投げ)**.** $X_i \sim \mathrm{Bern}(1/2)$（$\mu = 1/2,\, \sigma^2 = 1/4$）のとき、CLT は $\sqrt{n}(\bar{X}_n - 1/2)/(1/2) \xrightarrow{d} \mathcal{N}(0,1)$ を予測する。

$n = 10$ のとき、$S_n = \sum X_i \sim \mathrm{Bin}(10, 1/2)$ の分布を $\mathcal{N}(5, 2.5)$ と比較すると既にかなり良い近似が得られる。

$n = 100$ では二項分布のヒストグラムと正規密度はほぼ区別がつかない。 ただし $X_i \sim \mathrm{Bern}(0.01)$ のような強い歪みがある場合、 $n = 100$ でも正規近似は不正確であり、$n = 1000$ 以上が必要となる。

コード例：中心極限定理の視覚化（Python）

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

rng = np.random.default_rng(42)
n_values = [1, 5, 30, 100]
n_sim = 10000

fig, axes = plt.subplots(1, 4, figsize=(14, 3))
for ax, n in zip(axes, n_values):
    # Exp(1) から標本平均を n_sim 回シミュレーション
    samples = rng.exponential(scale=1.0, size=(n_sim, n))
    means = samples.mean(axis=1)
    standardized = (means - 1.0) / (1.0 / np.sqrt(n))

    ax.hist(standardized, bins=40, density=True, alpha=0.7, label='実験')
    x = np.linspace(-4, 4, 200)
    ax.plot(x, stats.norm.pdf(x), 'r-', label=r'$\mathcal{N}(0,1)$')
    ax.set_title(f'$n = {n}$')
    ax.set_xlim(-4, 4)
    ax.legend(fontsize=8)

plt.suptitle('CLT: 指数分布の標本平均の標準化', y=1.02)
plt.tight_layout()
plt.show()
```

2.4.1　CLTの拡張：独立だが同一分布でない場合

古典的 CLT は i.i.d. を仮定するが、 実際のデータ解析では各観測の分散が異なる状況（層別サンプリング、 回帰分析の残差など）がしばしば現れる。 Lindeberg–Feller の CLT は、独立だが同一分布でなくてもよい場合に CLT が成立するための条件を与える。

独立な確率変数 $X_1, X_2, \ldots$ で $\mathbb{E}[X_i] = 0$、$\mathrm{Var}(X_i) = \sigma_i^2 < \infty$ とし、 $s_n^2 = \sum_{i=1}^n \sigma_i^2$ とする。 まず $n = 2$ の簡単な場合で Lindeberg 条件の意味を確認しよう。

例 2.4.3 (Lindeberg 条件の $n = 2$ での意味)**.** X_1, X_2 が独立、$\mathbb{E}[X_i] = 0$、$\sigma_1^2 = 1$, $\sigma_2^2 = 1$ とすると $s_2^2 = 2$。 Lindeberg 条件は $\varepsilon > 0$ に対して

$$\frac{1}{2}\left(\mathbb{E}[X_1^2 \mathbf{1}\{|X_1| > \varepsilon\sqrt{2}\}] + \mathbb{E}[X_2^2 \mathbf{1}\{|X_2| > \varepsilon\sqrt{2}\}]\right) \to 0$$

を要求する。これは「各 X_i が全体の標準偏差 s_n に比べて 極端に大きな値を取る寄与が無視できる」ことを意味する。

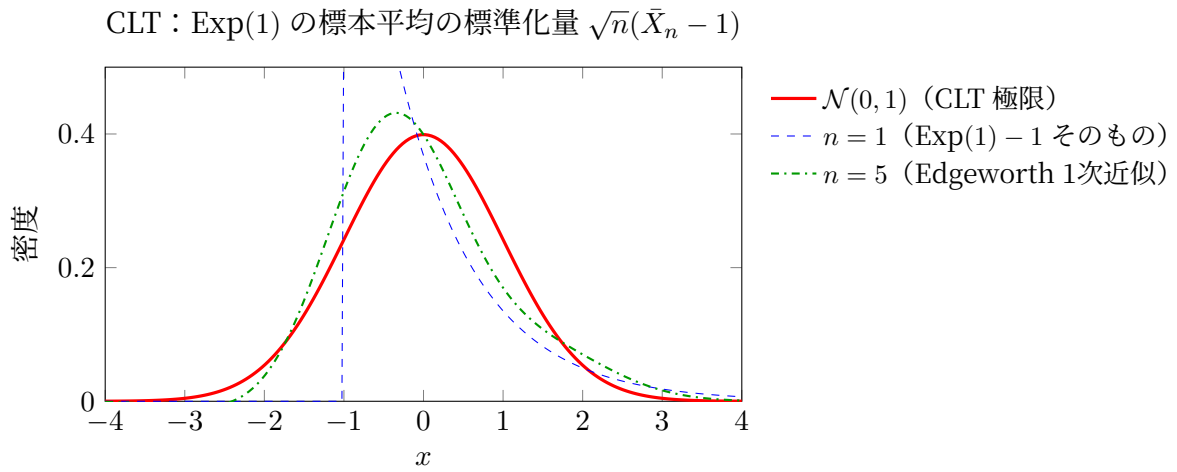


図 2.3: CLT による正規近似の改善。$X_i \sim \mathrm{Exp}(1)$ の標準化標本平均について、$n = 1$（元の指数分布、強い右歪み）から n の増加とともに $\mathcal{N}(0,1)$ に近づく様子を示す。$n = 5$ の曲線は Edgeworth 展開の1次近似であり、正規密度に歪度の補正項を加えたものである。

一般の場合の定理を述べる。

定理 2.4.4 (Lindeberg–Fellerの中心極限定理)**.** *独立な確率変数 $X_1, X_2, \ldots$ が $\mathbb{E}[X_i] = 0$、$\mathrm{Var}(X_i) = \sigma_i^2 < \infty$ を満たし、$s_n^2 = \sum_{i=1}^n \sigma_i^2$ とする。* **Lindeberg条件***：すべての $\varepsilon > 0$ に対して*

$$L_n(\varepsilon) := \frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\bigl[X_i^2 \mathbf{1}\{|X_i| > \varepsilon s_n\}\bigr] \to 0 \quad (n \to \infty)$$

が成り立てば、$s_n^{-1} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(0,1)$。

読み下し

Lindeberg 条件は「どの個別の確率変数も、全体の変動 s_n に対して 支配的に大きな寄与をしない」ことを要求する。 $L_n(\varepsilon)$ は「$|X_i|$ が εs_n を超える部分の 二乗モーメントの総和が、全分散 s_n^2 に比べて無視できる」ことを測っている。 i.i.d. の場合は各項が $\sigma^2 \cdot \mathbb{E}[(X_1/\sigma)^2 \mathbf{1}\{|X_1| > \varepsilon\sigma\sqrt{n}\}]$ となり、 被積分関数が 0 に収束するので優収束定理より Lindeberg 条件は自動的に満たされる。 したがって古典的 CLT は Lindeberg–Feller の特殊ケースである。

証明の方針. まず、各 X_i/s_n の特性関数を2次までテイラー展開する。 次に、

$$\sum_{i=1}^{n} \log \varphi_{X_i}(t/s_n)$$

を評価する。 Lindeberg 条件は、テイラー展開の剰余項に対して

$$\bigl|\varphi_{X_i}(t/s_n) - 1 + t^2\sigma_i^2/(2s_n^2)\bigr| \le \frac{t^2}{s_n^2}\,\mathbb{E}\bigl[X_i^2 \mathbf{1}\{|X_i| > \varepsilon s_n\}\bigr]$$

という一様評価を与えるために使われる。 これにより $\sum \log \varphi_{X_i}(t/s_n) \to -t^2/2$ が示され、

Lévy の連続性定理から結論が従う。 完全な証明は Billingsley (1995), Theorem 27.2 を参照。 □

実務ポイント

Lindeberg–Feller CLT は、異なるグループからのデータを混合して 平均を計算する場面（例：層別サンプリング）で重要となる。 各グループの分散が異なっていても、「一つのグループが全体の変動を支配しない」 限り、標準化された平均は正規近似が正当化される。

2.4.2　多変量中心極限定理

定理 2.4.5 (多変量中心極限定理)**.** $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ *が* $\mathbb{R}^d$ *値の i.i.d. 確率ベクトルで* $\mathbb{E}[\boldsymbol{X}_1] = \boldsymbol{\mu}$、$\mathrm{Cov}(\boldsymbol{X}_1) = \boldsymbol{\Sigma}$ *のとき、*

$$\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

読み下し

多変量 CLT は「各成分の標本平均だけでなく、成分間の相関構造も含めて 同時に正規分布に収束する」と読む。 回帰分析における係数ベクトルの同時推論（同時信頼領域）の基盤となる。

2.4.3　Berry–Esseenの定理：正規近似の精度

CLT は $n \to \infty$ での収束を保証するが、 有限の n で正規近似がどの程度正確かは教えてくれない。 Berry–Esseen の定理がこの問いに答える。

定理 2.4.6 (Berry–Esseenの定理)**.** $X_1, X_2, \dots$ *が i.i.d. で* $\mathbb{E}[X_1] = 0$、$\mathbb{E}[X_1^2] = \sigma^2 > 0$、$\mathbb{E}[|X_1|^3] = \rho < \infty$ *のとき、*

$$\sup_{x \in \mathbb{R}} \left| \mathrm{P}\left(\frac{\bar{X}_n}{\sigma/\sqrt{n}} \le x \right) - \Phi(x) \right| \le \frac{C\rho}{\sigma^3 \sqrt{n}},$$

ここで $C \le 0.4748$*（Shevtsova, 2011）、*Φ *は標準正規分布の分布関数。*

読み下し

Berry–Esseen の定理は「CLT の正規近似の誤差は $O(1/\sqrt{n})$ で減少し、 その定数は第3モーメント（歪度に関連）で決まる」と読む。 歪みの大きい分布ほど正規近似の精度が悪くなることを定量的に示している。

例 2.4.7 (Berry–Esseen 評価の数値例)**.** $X_i \sim \mathrm{Exp}(1) - 1$（平均0、分散1）の場合を考える。 $Y = X_i + 1 \sim \mathrm{Exp}(1)$ とおくと $\rho = \mathbb{E}[|Y-1|^3] = \int_0^1 (1-y)^3 e^{-y}\,dy + \int_1^\infty (y-1)^3 e^{-y}\,dy$。 第2項は $t = y-1$ と置換すると $e^{-1}\Gamma(4) = 6e^{-1}$。 第1項は部分積分の繰り返しで $6e^{-1} - 2$ と求まる。 よって $\rho = 12e^{-1} - 2 \approx 2.41$ であり、 上界は $C\rho/(\sigma^3\sqrt{n}) \approx 1.144/\sqrt{n}$ となる。

n	Berry–Esseen 上界	実際の最大誤差（数値計算）	比率
10	0.362	0.033	11.0
100	0.114	0.010	11.4
1,000	0.036	0.003	12.0

上界は実際の誤差の11–12倍程度であり、かなり保守的である。これは Berry–Esseen の定理がすべての分布に対する一様な上界を与えるためであり、個別の分布に対してはタイトではない。ただし収束速度が $O(1/\sqrt{n})$ であることは正しく反映されている。対称分布（例：$\mathcal{N}(0,1)$）では誤差は $O(1/n)$ に改善されることが知られている。

実務ポイント

「$n \geq 30$ なら正規近似が使える」という経験則は、Berry–Esseen の観点からは不十分である。歪度の大きい分布（対数正規分布、ポアソン分布で λ が小さい場合など）では $n = 100$ でも正規近似が不正確なことがある。実務では、正規近似に依存する前に歪度（$\gamma_1 = \rho/\sigma^3$）を確認し、必要に応じてブートストラップ法や正確法（exact method）を検討すべきである。

2.5　デルタ法

CLT で得られた漸近正規性を、統計量の変換に「伝搬」させる道具がデルタ法である。

定理 2.5.1 (デルタ法)**.** $\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ *とする。関数 g が θ で微分可能かつ $g'(\theta) \neq 0$ ならば、*

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2 [g'(\theta)]^2).$$

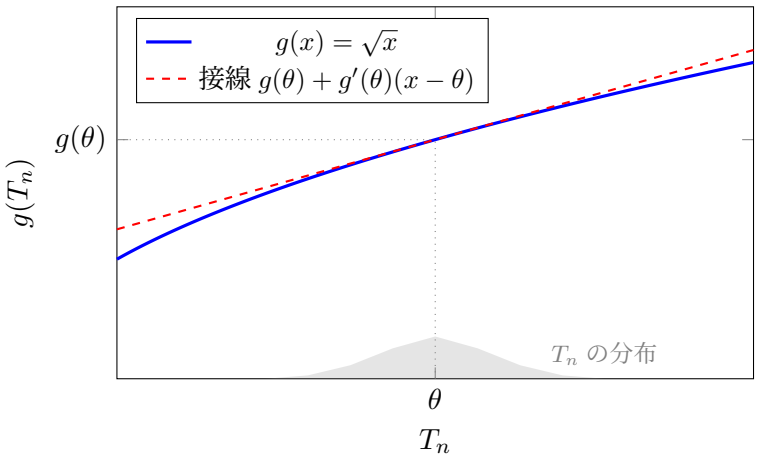


図 2.4: デルタ法の幾何学的直観。T_n が θ の近傍に集中しているとき、非線形関数 g を接線（1次テイラー近似）で置き換えられる。$g(T_n)$ の漸近分散が $[g'(\theta)]^2\sigma^2$ となるのは、接線の傾き $g'(\theta)$ が分布の「伸縮率」を決めるためである。

読み下し

デルタ法は「漸近正規な統計量 T_n を滑らかな関数 g で変換しても 漸近正規性は保たれ、漸近分散は $g'(\theta)^2$ 倍になる」と読む。 図2.4に示す通り、T_n が θ の近くに集中しているので g をその近傍で1次のテイラー近似で置き換えられるというのが核心である。

注意 2.5.2 (o_p, O_p 記法). 確率変数列 $Y_n = o_p(1)$ とは $Y_n \xrightarrow{p} 0$ のことである。 より一般に $Y_n = o_p(a_n)$ は $Y_n/a_n \xrightarrow{p} 0$ を意味する。 同様に $Y_n = O_p(a_n)$ は 任意の $\varepsilon > 0$ に対して $\sup_n \mathrm{P}(|Y_n/a_n| > M) < \varepsilon$ となる $M > 0$ が存在することを意味する（確率的有界性）。 これらの記法は漸近論の議論を簡潔にするための標準的な道具であり、 詳細は van der Vaart (1998) を参照。

Proof. **証明の方針**：テイラー展開で $g(T_n)$ を線形化し、 スラツキーの補題で残余項を処理する。

テイラー展開より $g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + o(|T_n - \theta|)$。 $T_n \xrightarrow{p} \theta$（定理2.1.6と仮定から）であるから、 剰余項を $\sqrt{n}$ 倍すると $o_p(1)$ となり

$$\sqrt{n}(g(T_n) - g(\theta)) = g'(\theta) \cdot \sqrt{n}(T_n - \theta) + o_p(1).$$

スラツキーの補題（定理2.6.3）から $\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} g'(\theta) \cdot \mathcal{N}(0, \sigma^2) = \mathcal{N}(0, \sigma^2 [g'(\theta)]^2)$。 □

例 2.5.3 (分散安定化変換). $X_1, \ldots, X_n \sim \mathrm{Pois}(\lambda)$ のとき、 CLT より $\sqrt{n}(\bar{X}_n - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda)$。 漸近分散が λ に依存する——これは信頼区間の構成に不便である。

$g(x) = \sqrt{x}$ とおくとデルタ法より

$$\sqrt{n}(\sqrt{\bar{X}_n} - \sqrt{\lambda}) \xrightarrow{d} \mathcal{N}(0, 1/4).$$

漸近分散が λ に依存しなくなり、標準誤差 $1/(2\sqrt{n})$ を用いて λ の近似的な推論（信頼区間の構成など）が容易になる。

例えば $n = 100$, $\bar{X}_n = 4.2$ のとき、$\sqrt{\bar{X}_n}$ の標準誤差は $1/(2\sqrt{100}) = 0.05$ である。 これを用いると、$\sqrt{\lambda}$ の 95% 信頼区間は $2.049 \pm 1.96 \times 0.05 = 2.049 \pm 0.098$ となり、 λ の近似95%信頼区間は $(2.049 - 0.098)^2 \approx 3.81$ から $(2.049 + 0.098)^2 \approx 4.61$ と求められる。

例 2.5.4 (ログオッズ変換（ロジット変換）). 成功確率 p の推定量を標本比率 $\hat{p}_n$ とする。 CLT より $\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p))$。

$g(x) = \log(x/(1-x))$（ロジット関数）とすると $g'(x) = 1/(x(1-x))$ であるからデルタ法より

$$\sqrt{n}(\mathrm{logit}(\hat{p}_n) - \mathrm{logit}(p)) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{p(1-p)}\right).$$

ロジスティック回帰の係数の漸近分布導出に直結する重要な結果である。

例 2.5.5 (Fisher の z 変換（相関係数）). 標本相関係数 $\hat{\rho}$ が ρ のまわりで漸近正規であるとき、 $g(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$（Fisher の z 変換）を適用すると $g'(\rho) = 1/(1-\rho^2)$ であるからデルタ

法より

$$\sqrt{n-3}\,(g(\hat{\rho})-g(\rho)) \approx \mathcal{N}(0,1).$$

漸近分散が ρ に依存しなくなるため、 $z=g(\hat{\rho})$ を用いた信頼区間の構成が実務で広く使われている。

定理 2.5.6 (多変量デルタ法)**.** *$\sqrt{n}(\boldsymbol{T}_n-\boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_d(\mathbf{0},\boldsymbol{\Sigma})$ のとき、 $g\colon \mathbb{R}^d \to \mathbb{R}^k$ が $\boldsymbol{\theta}$ で微分可能で ヤコビ行列 $\boldsymbol{J}=\nabla g(\boldsymbol{\theta})^\top$ の階数が k ならば、*

$$\sqrt{n}(g(\boldsymbol{T}_n)-g(\boldsymbol{\theta})) \xrightarrow{d} \mathcal{N}_k(\mathbf{0},\boldsymbol{J}\boldsymbol{\Sigma}\boldsymbol{J}^\top).$$

読み下し

多変量デルタ法の漸近分散 $\boldsymbol{J}\boldsymbol{\Sigma}\boldsymbol{J}^\top$ は、 元の共分散行列 $\boldsymbol{\Sigma}$ をヤコビ行列 $\boldsymbol{J}$ で「変換」したものである。 最尤推定量 $\hat{\boldsymbol{\theta}}$ の漸近正規性と組み合わせると、 $g(\hat{\boldsymbol{\theta}})$ の漸近分布が自動的に得られる。

実務ポイント

デルタ法は実務で極めて頻繁に使われる：

- **分散安定化変換**：漸近分散がパラメータに依存しないようにする。
- **対数変換による信頼区間**：ハザード比やオッズ比の信頼区間は 対数スケールで構成し、指数関数で戻す。
- **疫学指標**：リスク比（RR）、オッズ比（OR）の標準誤差導出。

ただし $g'(\theta)=0$ のとき（例：$\theta=0$ での $g(x)=x^2$）、 通常のデルタ法は適用できず、2次デルタ法が必要となる。

2.6　連続写像定理とスラツキーの補題

収束を「保存」し「組み合わせる」ための2つの基本道具を述べる。

定理 2.6.1 (連続写像定理（CMT）)**.** *g が連続ならば：*

(i) $X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X)$。

(ii) $X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X)$。

(iii) $X_n \xrightarrow{a.s.} X \implies g(X_n) \xrightarrow{a.s.} g(X)$。

読み下し

連続写像定理は「連続な変換は収束を保存する」と読む。 分布収束・確率収束・概収束のすべてで成り立つ。 直観的には、X_n が X に近いとき、g の連続性から $g(X_n)$ も $g(X)$ に近いのは自然である。

例 2.6.2 (連続写像定理の応用)**.** CLT より $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} Z \sim \mathcal{N}(0,1)$ のとき：

(i) $g(x) = x^2$ は連続であるから $Z_n^2 \xrightarrow{d} Z^2 \sim \chi^2(1)$。 これが1標本のカイ二乗検定統計量の漸近分布の出発点となる。

(ii) $g(x) = |x|$ は連続であるから $|Z_n| \xrightarrow{d} |Z|$ （半正規分布に従う）。

(iii) $\boldsymbol{Z}_n \xrightarrow{d} \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{I})$ のとき $\|\boldsymbol{Z}_n\|^2 \xrightarrow{d} \chi^2(d)$。 多変量の場合の Wald 検定統計量の根拠となる。

定理 2.6.3 (スラツキーの補題)**.** *$X_n \xrightarrow{d} X$ かつ $Y_n \xrightarrow{p} c$（c は定数）ならば：*

(i) $X_n + Y_n \xrightarrow{d} X + c$。

(ii) $Y_n X_n \xrightarrow{d} cX$。

(iii) $c \neq 0$ のとき $X_n / Y_n \xrightarrow{d} X/c$。

読み下し

スラツキーの補題は「分布収束する列に、確率収束する列を足したり掛けたり 割ったりしても、分布収束が保たれる」と読む。 重要な注意点として、$X_n \xrightarrow{d} X$ と $Y_n \xrightarrow{d} Y$ だけでは $X_n + Y_n$ の極限分布は一般に決まらない（同時分布の情報が必要）。 スラツキーの補題が使えるのは、一方が定数に確率収束する場合に限られる。

例 2.6.4 (分散未知の場合の t 統計量)**.** $X_1, \dots, X_n$ が i.i.d. で $\mathbb{E}[X_1] = \mu$、$\mathrm{Var}(X_1) = \sigma^2$ とする。

CLT より $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0,1)$。 一方、標本標準偏差 $S_n = \sqrt{(n-1)^{-1}\sum (X_i - \bar{X}_n)^2}$ は SLLN より $S_n \xrightarrow{p} \sigma$（$S_n^2 \xrightarrow{p} \sigma^2$ と連続写像定理）。

スラツキーの補題（(iii)の形）より

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma}{S_n/\sigma} \xrightarrow{d} \frac{\mathcal{N}(0,1)}{1} = \mathcal{N}(0,1).$$

これにより、σ が未知でも S_n で置き換えた推論が（大標本で）正当化される。

正規母集団の場合の正確な結果は異なることに注意： $X_i \sim \mathcal{N}(\mu, \sigma^2)$ のとき T_n は 自由度 $n-1$ の t 分布に*正確に*従う（n が有限でも成り立つ）。 t 分布は $n \to \infty$ で $\mathcal{N}(0,1)$ に収束するので、両者は整合する。

実務ポイント

スラツキーの補題は統計学で最も頻繁に使われる漸近的ツールの一つである。 典型的なパターンは以下の通り：

1. CLT で「既知パラメータ版」の漸近分布を得る。
2. 未知パラメータの一致推定量を構成する（大数の法則を使う）。
3. スラツキーの補題で推定量への置き換えを正当化する。

この三段論法は、第3章以降で 最尤推定量の漸近推論を展開する際に繰り返し現れる。

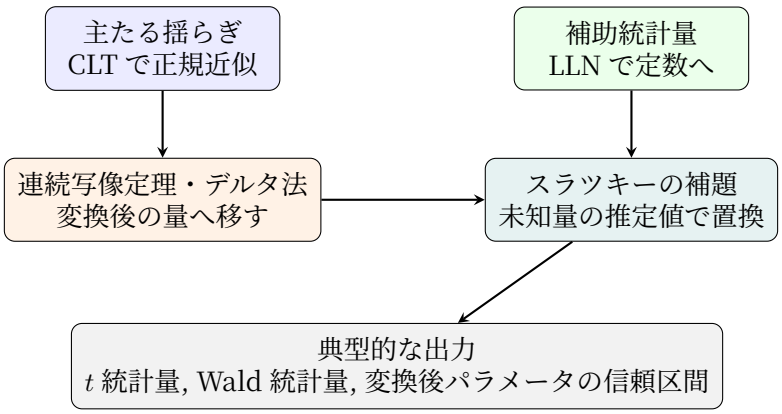


図 2.5: 収束定理を組み合わせる標準レシピ。主たる統計量の揺らぎは CLT で捉え、補助量は LLN で定数化し、必要な変換を連続写像定理やデルタ法で処理した後に、スラツキーの補題で推定値への置き換えを正当化する。

主要な結果

重要結果

本章の核心的な結論：

1. **概収束・確率収束・L^p収束・分布収束**は それぞれ異なる「近づき方」を表す。強い収束から弱い収束への含意関係を意識することが重要である。
2. **大数の法則**は標本平均の一致性を保証し、 推定量が真の値に近づくための理論的土台を与える。
3. **中心極限定理**は $\sqrt{n}(\bar{X}_n - \mu)$ を正規分布で近似できることを示し、 大標本推論の出発点を与える。 Berry–Esseen の定理は、その近似誤差が有限標本でどれほど残るかを評価する。
4. **デルタ法・連続写像定理・スラツキーの補題**により、 標本平均以外の統計量にも漸近分布を機械的に押し広げられる。
5. 本章で整備した収束の道具立ては、 次章以降の最尤推定量・検定統計量・信頼区間の漸近理論を正当化する共通言語である。

2.7　演習問題

理論問題

演習問題 2.1 (確率収束の一意性)**.** $X_n \xrightarrow{p} X$ かつ $X_n \xrightarrow{p} Y$ ならば $\mathrm{P}(X = Y) = 1$ であることを示せ。 ヒント：三角不等式 $|X - Y| \le |X_n - X| + |X_n - Y|$ を使え。

演習問題 2.2 (Borel–Cantelli の第二補題)**.** $\{A_n\}$ が独立で $\sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty$ ならば $\mathrm{P}(\limsup A_n) = 1$ であることを示せ。 ヒント：$1 - x \le e^{-x}$ を用いて $\mathrm{P}(\bigcap_{n=N}^{M} A_n^c)$ を評価せよ。

演習問題 2.3 (標本分散の概収束)**.** $X_1, \dots, X_n$ が i.i.d. で $\mathbb{E}[X_1^4] < \infty$ のとき、$S_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ が $\sigma^2 = \mathrm{Var}(X_1)$ に概収束することを示せ。 *ヒント*：$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum X_i^2 - \bar{X}_n^2\right)$ と書き、各項に SLLN を適用せよ。

演習問題 2.4 (L^p 収束と確率収束の関係)**.** $X_n \xrightarrow{L^p} X$（$p \geq 1$）ならば $X_n \xrightarrow{p} X$ であることを、マルコフの不等式を用いて示せ。 また、逆が一般には成り立たないことを反例で示せ。

演習問題 2.5 (極値分布)**.** $X_1, X_2, \dots$ が i.i.d. $\mathrm{Exp}(1)$ のとき、$M_n = \max(X_1, \dots, X_n) - \log n$ の極限分布を求めよ。 *ヒント*：$\mathrm{P}(M_n \leq x) = (1 - e^{-(x+\log n)})^n$ を評価せよ。

計算・実装問題

数値実験を含む問題では、(1) 設定（分布、パラメータ、反復回数、乱数シード）、(2) 作成した図表または表、(3) 比較指標、(4) 結果から読める一言考察、の4点を答えに含めよ。

演習問題 2.6 (ロジット変換の漸近分布)**.** $\hat{p}_n$ を成功確率 p の二項分布の標本比率とする。 デルタ法を用いて $\log(\hat{p}_n/(1-\hat{p}_n))$ の漸近分布を求めよ。 $p = 0.3$, $n = 200$ のとき、漸近的な95%信頼区間を数値で計算せよ。

演習問題 2.7 (t 統計量の分布のシミュレーション)**.** $X_1, \dots, X_n$ が i.i.d. $\mathcal{N}(\mu, \sigma^2)$ のとき、$\bar{X}_n$ と S_n^2 が独立であることを示し、$T = \sqrt{n}(\bar{X}_n - \mu)/S_n$ が自由度 $n-1$ の t 分布に従うことを導け。 さらに $n = 5, 30, 100$ でシミュレーションを行い、t 分布と正規分布を比較せよ。

演習問題 2.8 (Berry–Esseen の数値検証)**.** $X_i \sim \mathrm{Pois}(1) - 1$（平均0、分散1）に対して、$n = 10, 50, 100, 500$ のそれぞれについて：(a) $\bar{X}_n$ の標準化量の経験分布関数を10,000回のシミュレーションで求めよ。 (b) 標準正規分布の分布関数との最大乖離（Kolmogorov–Smirnov 統計量）を計算せよ。 (c) Berry–Esseen の上界と比較し、収束速度が $O(1/\sqrt{n})$ であることを確認せよ。

演習問題 2.9 (モンテカルロ積分の精度)**.** $I = \int_0^1 e^{-x^2}\,dx$ の値をモンテカルロ法で推定する。(a) $U_1, \dots, U_n \sim \mathrm{Unif}(0,1)$ を用いて $\hat{I}_n = n^{-1}\sum e^{-U_i^2}$ と推定するとき、SLLN と CLT を用いて $\hat{I}_n$ の一致性と漸近的な95%信頼区間を導出せよ。 (b) Python または R で $n = 1000$ として数値実験を行い、真の値（数値積分で計算）と比較せよ。

演習問題 2.10 (Fisher の z 変換の実装)**.** 二変量正規分布 $\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$（$\rho = 0.5$）から $n = 50$ の標本を1,000回生成し、各回で標本相関係数 $\hat{\rho}$ を計算せよ。 (a) $\hat{\rho}$ のヒストグラムと、デルタ法で予測される漸近正規分布を重ねて描画せよ。 (b) Fisher の z 変換 $g(\hat{\rho})$ のヒストグラムと $\mathcal{N}(g(\rho), 1/(n-3))$ を比較せよ。 (c) どちらの正規近似がより正確か評価せよ。

略解の指針

ここでは解法の骨組みと到達点だけを示す。証明の細部や数値確認は自分で埋めること。

- **演習2.1** 使う道具: 三角不等式。 最初の1手: 任意の $\varepsilon > 0$ に対して $\{|X - Y| > \varepsilon\} \subset \{|X_n - X| > \varepsilon/2\} \cup \{|X_n - Y| > \varepsilon/2\}$ と書く。 途中の要点: 右辺の確率は仮定より 0 に収束する。 最終形: $\mathrm{P}(|X - Y| > \varepsilon) = 0$ が全 $\varepsilon > 0$ で成り立つので $\mathrm{P}(X = Y) = 1$。

- **演習2.2** 使う道具: 独立性と指数評価。 最初の1手: $\mathrm{P}(\bigcap_{n=N}^{M} A_n^c) = \prod_{n=N}^{M}(1-\mathrm{P}(A_n))$ を書く。 途中の要点: $1-x \le e^{-x}$ により右辺は $\exp\{-\sum_{n=N}^{M}\mathrm{P}(A_n)\}$ 以下となり、和が発散するので 0 に行く。 最終形: $\mathrm{P}(\bigcup_{n\ge N} A_n) = 1$ が全 N で成り立つから $\mathrm{P}(\limsup A_n) = 1$。

- **演習2.3** 使う道具: 強法則。 最初の1手: $S_n^2 = \frac{n}{n-1}\{n^{-1}\sum X_i^2 - \bar{X}_n^2\}$ に分解する。 途中の要点: $\mathbb{E}[X_1^4] < \infty$ から $\mathbb{E}[X_1^2] < \infty$ なので $n^{-1}\sum X_i^2 \to \mathbb{E}[X_1^2]$, $\bar{X}_n \to \mathbb{E}[X_1]$ が概収束する。 最終形: $S_n^2 \to \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \sigma^2$ a.s.

- **演習2.4** 使う道具: マルコフの不等式。 最初の1手: $\mathrm{P}(|X_n - X| > \varepsilon) \le \mathbb{E}[|X_n - X|^p]/\varepsilon^p$ と書く。 途中の要点: L^p 収束なら右辺は 0 に行く。 最終形: 逆向きの反例は $X_n = n$ with prob. $1/n$, 0 otherwise とすれば、 $X_n \xrightarrow{p} 0$ だが $\mathbb{E}|X_n| = 1$ で L^1 収束しない。

- **演習2.5** 使う道具: 最大値の分布関数。 最初の1手: $\mathrm{P}(M_n \le x) = \mathrm{P}(\max_i X_i \le x + \log n)$ を独立性で積に分解する。 途中の要点: $(1 - e^{-x}/n)^n \to e^{-e^{-x}}$ を使う。 最終形: $M_n = \max_i X_i - \log n$ は Gumbel 分布 $\mathrm{P}(M_n \le x) \to e^{-e^{-x}}$ に収束する。

- **演習2.6** 使う道具: デルタ法。 最初の1手: $g(p) = \log\{p/(1-p)\}$ と置いて $g'(p) = 1/\{p(1-p)\}$ を計算する。 途中の要点: $\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p))$ をデルタ法に通す。 最終形: $\sqrt{n}\{g(\hat{p}_n) - g(p)\} \xrightarrow{d} \mathcal{N}(0, 1/\{p(1-p)\})$。 $p = 0.3$, $n = 200$ では近似標準誤差は $\sqrt{1/(200 \cdot 0.3 \cdot 0.7)}$ である。

- **演習2.7** 使う道具: 正規標本の直交分解。 最初の1手: 標本平均と残差ベクトルを直交成分に分ける。 途中の要点: Cochran の定理から $\bar{X}_n$ と $(n-1)S_n^2/\sigma^2$ は独立で、後者は χ^2_{n-1} に従う。 最終形: $T = \sqrt{n}(\bar{X}_n - \mu)/S_n$ は t_{n-1} に従い、 シミュレーションでは n が大きいほど正規分布に近づく。

- **演習2.8** 使う道具: Berry–Esseen の上界。 最初の1手: $X_i = \mathrm{Pois}(1) - 1$ では平均 0、分散 1、三次絶対モーメント $\rho_3 = \mathbb{E}|X_i|^3$ が有限であることを確認する。 途中の要点: 上界は一般に $C\rho_3/\sqrt{n}$ の形で与えられる。 最終形: 経験的な KS 統計量も $O(n^{-1/2})$ の速さで減り、有限標本では上界はかなり保守的だが収束次数は一致する。

- **演習2.9** 使う道具: SLLN と CLT。 最初の1手: $Y_i = e^{-U_i^2}$ と置けば $\hat{I}_n = n^{-1}\sum Y_i$ は標本平均である。 途中の要点: $\hat{I}_n \to I$ a.s., かつ $\sqrt{n}(\hat{I}_n - I) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(Y_1))$ を使う。 最終形: 95% CI は $\hat{I}_n \pm 1.96\,\widehat{\mathrm{sd}}(Y)/\sqrt{n}$。 数値実験では真値はおよそ 0.7468 である。

- **演習2.10** 使う道具: 相関係数のデルタ法。 最初の1手: $\hat{\rho}$ 自体の漸近分散は $(1-\rho^2)^2/n$ である。 途中の要点: $g(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$ の導関数は $1/(1-\rho^2)$ なので、 $g(\hat{\rho})$ の分散はほぼ $1/n$ に平準化される。 最終形: z 変換後のヒストグラムの方が対称で、 $\mathcal{N}(g(\rho), 1/(n-3))$ が元の $\hat{\rho}$ よりよく当てはまる。

次章への橋渡し

本章で確率変数の列が「収束する」ことの意味を厳密にし、 大数の法則で推定量の一致性、中心極限定理で漸近分布を得る道具を整えた。 しかし、これらはあくまで「標本平均」という特定の統計量についての結果である。

次の第3章では、データの生成過程を統計モデルとして定式化し、 パラメータを推定する一般的な枠組み（最尤推定、モーメント法、M推定量）を導入する。 十分統計量によるデータの縮約、Cramér–Rao の下界による推定量の最適性など、 統計的推論の核心に踏み込む。本章で整備した収束の道具立ては、推定量の一致性の証明や 漸近分散の評価において繰り返し使われることになる。

参考文献

Durrett, R. (2019). *Probability: Theory and Examples* (5th ed.). Cambridge University Press. ——測度論的確率論に基づく収束理論の標準的参考書。強大数の法則の完全証明はここを参照。

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). Wiley. ——分布収束と弱収束の理論を詳細に展開。Portmanteau 定理の証明を含む。

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. —— o_p, O_p 記法と漸近展開の方法論。デルタ法の一般化も詳しい。

Casella, G. & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Duxbury Press. ——統計推論の観点からの収束理論の導入。スラツキーの補題の応用例が豊富。

Wasserman, L. (2004). *All of Statistics*. Springer. ——簡潔で実用的な収束理論の概説。データサイエンスとの接点が明確。

Shevtsova, I. (2011). On the absolute constants in the Berry–Esseen type inequalities for identically distributed summands. *arXiv:1111.6554*. ——Berry–Esseen 定数の最良の上界。

第II部

統計的推論の理論

第3章

統計モデルと推定

問いと学習目標

この章で答える問い

- データの生成過程を数学的に記述する「統計モデル」とは何か？
- データ全体を保持しなくても、パラメータの推定に必要な情報を 損失なく凝縮できるか？
- パラメータを推定する方法にはどのようなものがあり、 それぞれどのような性質を持つか？
- 推定量の「良さ」をどう定量化し、最適な推定量の限界はどこにあるか？

読み終えたらできるようになること

1. 統計モデル・パラメータ空間・識別可能性を定義し、具体例で説明できる。
2. 十分統計量・最小十分統計量・完備統計量の概念と相互関係を理解する。
3. 指数型分布族の統計的性質（十分性・完備性・フィッシャー情報量との関係）を導出できる。
4. 最尤推定量・モーメント法推定量・M推定量を構成し、その性質を比較できる。
5. Cramér–Rao下界を用いて推定量の最適性を評価できる。

直観的理解

データから有用な結論を導くには、 「データの生成過程を数学的に記述する」ことと「その記述に基づいてパラメータを推定する」ことが不可分に結びついている。
本章の流れは次の通りである。 まず、統計モデルを定義し、データからパラメータに関する情報を 損失なく凝縮する**十分統計量**の概念を導入する。 次に、**指数型分布族**がこの凝縮において特に優れた構造を持つことを見る。 その上で、**フィッシャー情報量**によって 「データがパラメータについてどれだけの情報を運ぶか」を定量化し、 **最尤推定法・モーメント法・M推定法** という具体的な推定手法を学ぶ。 最後に、Cramér–Raoの下界によって「推定の限界」を定量化する。

検定と信頼集合は第4章で、 漸近理論の詳細は第6章で展開する。

3.1　統計モデルの定式化

定義 3.1.1 (統計モデル)**.** データ $\boldsymbol{x} = (x_1, \dots, x_n)$ が取りうる値の集合を $\mathcal{X}$ とする。 **統計モデル**とは、密度関数（または質量関数）の族

$$\mathcal{P} = \{ f(\boldsymbol{x} \mid \theta) : \theta \in \Theta \}$$

をいう。Θ を**パラメータ空間**と呼ぶ。 $\Theta \subset \mathbb{R}^d$ と有限次元であるとき**パラメトリックモデル**、 無限次元であるとき**ノンパラメトリックモデル**と呼ぶ。

読み下し

統計モデルとは、「データがどのような確率的な仕組みで生成されたか」を パラメータ θ で索引づけた候補の集合である。 ちょうど地図帳が地域を索引で引けるように、統計モデルはデータの背後にありうる分布を θ で引けるようにする。 モデルを設定して初めて、「データから θ を推定する」という問いが 明確に定式化される。

定義 3.1.2 (識別可能性)**.** 統計モデル $\{ f(\cdot \mid \theta) : \theta \in \Theta \}$ が**識別可能**であるとは、 $\theta_1 \neq \theta_2 \implies f(\cdot \mid \theta_1) \neq f(\cdot \mid \theta_2)$ （a.e.）が成り立つことをいう。

読み下し

識別可能性は「異なるパラメータ値が異なる分布を生む」ことを要求する。 もし $\theta_1 \neq \theta_2$ でも同じ分布が生じるなら、 いくらデータを集めても両者を区別できず、推定問題が意味をなさない。

例 3.1.3 (識別不可能なモデル)**.** 混合正規分布 $f(x \mid p, \mu_1, \mu_2) = p\,\varphi(x - \mu_1) + (1-p)\,\varphi(x - \mu_2)$（$\varphi$ は標準正規密度）は、 (p, μ_1, μ_2) と $(1-p, \mu_2, \mu_1)$ が同じ分布を生むため、 パラメータの順序に関して識別不可能である。 実務では成分に順序制約 $\mu_1 \leq \mu_2$ を課して識別可能性を確保する。

3.2　十分統計量とファクトリゼーション定理

定義 3.2.1 (十分統計量)**.** 統計量 $T = T(\boldsymbol{X})$ が θ に対して**十分**であるとは、 T が与えられたときの $\boldsymbol{X}$ の条件付き分布が θ に依存しないこと、 すなわち

$$\mathrm{P}(\boldsymbol{X} \in A \mid T = t, \theta) = \mathrm{P}(\boldsymbol{X} \in A \mid T = t) \quad \text{for all } A, t, \theta$$

が成り立つことをいう。

読み下し

十分統計量とは、パラメータ θ についての情報をデータ全体から 「損失なく」凝縮し

たものである。T が十分統計量であれば、T さえ知っていれば 元データ $\boldsymbol{X}$ を覗き返す必要がない。 日常のアナロジーでいえば、果汁を搾った後の果物の殻には もはや味に関する情報が残っていない——十分統計量は「完全な果汁」である。

十分統計量の判定には、以下のファクトリゼーション定理が決定的に重要である。

定理 3.2.2 (Neyman–Fisherのファクトリゼーション定理)**.** *統計量 $T(\boldsymbol{X})$ が θ に対して十分であるための必要十分条件は、 関数 g と h が存在して*

$$f(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x}), \theta) \cdot h(\boldsymbol{x})$$

と分解できることである。

読み下し

この定理は「密度関数がデータ $\boldsymbol{x}$ を見るとき、 θ が絡む部分は $T(\boldsymbol{x})$ だけを通じて作用する」と述べている。 $h(\boldsymbol{x})$ は θ と無関係な部分であり、 パラメータの推論には寄与しない。

証明の概略（離散の場合）. **方針**：十分性の定義を条件付き確率の計算に翻訳し、 $g \cdot h$ という積の構造がちょうど θ を消す仕組みを見る。

($\Longleftarrow$) $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x} \mid T = t, \theta) = \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}, T = t \mid \theta) / \mathrm{P}(T = t \mid \theta)$。

$T(\boldsymbol{x}) = t$ のとき 分子 $= g(t,\theta)h(\boldsymbol{x})$、 分母 $= \sum_{\boldsymbol{y}:T(\boldsymbol{y})=t} g(t,\theta)h(\boldsymbol{y}) = g(t,\theta)\sum_{\boldsymbol{y}:T(\boldsymbol{y})=t} h(\boldsymbol{y})$。

よって $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x} \mid T = t, \theta) = h(\boldsymbol{x}) / \sum_{\boldsymbol{y}:T(\boldsymbol{y})=t} h(\boldsymbol{y})$ で θ に依存しない。 ここが核心である——$g(t,\theta)$ が分子と分母の両方に現れるため約分される。

($\Longrightarrow$) は $g(t,\theta) = \mathrm{P}(T = t \mid \theta)$、 $h(\boldsymbol{x}) = \mathrm{P}(\boldsymbol{X} = \boldsymbol{x} \mid T = T(\boldsymbol{x}))$ とおけばよい。 □

例 3.2.3 (正規分布の十分統計量)**.** $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$（$\mu, \sigma^2$ ともに未知）のとき、

$$f(\boldsymbol{x} \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right).$$

$\sum x_i$ と $\sum x_i^2$ の関数として書けるので、 $T(\boldsymbol{X}) = (\sum X_i, \sum X_i^2)$ は (μ, σ^2) に対する十分統計量である。

実務的に言えば、$n = 10{,}000$ 個のデータを持っていても、 この2つの数値さえ記録しておけば μ と σ^2 の推論には十分なのである。

3.3　最小十分統計量と完備統計量

十分統計量は一意ではない——例えばデータ全体 $\boldsymbol{X}$ 自身も 自明な十分統計量である。では、情報を最大限に凝縮した 「最も小さい」十分統計量は何か？

3.3.1　最小十分統計量

定義 3.3.1 (最小十分統計量)**.** 十分統計量 T が**最小十分**であるとは、 他の任意の十分統計量 U に対して $T = g(U)$ となる関数 g が存在することをいう。

読み下し

最小十分統計量とは、十分統計量の中で最もデータを圧縮したものである。 他のどんな十分統計量からも計算できるが、 逆に最小十分統計量から他の十分統計量を復元することは一般にはできない。 情報理論的に言えば、パラメータに関する情報を保ったまま データ圧縮を限界まで進めた結果である。

定理 3.3.2 (Lehmann–Scheffé の判定法)**.** *$f(\boldsymbol{x} \mid \theta)/f(\boldsymbol{y} \mid \theta)$ が θ に依存しないことと $T(\boldsymbol{x}) = T(\boldsymbol{y})$ が同値であるならば、 T は最小十分統計量である。*

読み下し

この判定法の直観は明快である。 尤度の比 $f(\boldsymbol{x} \mid \theta)/f(\boldsymbol{y} \mid \theta)$ が θ に依存しないとは、データ $\boldsymbol{x}$ と $\boldsymbol{y}$ が θ に関して 「同じ情報を持っている」ことを意味する。 最小十分統計量は、まさにこの「情報の等価性」で データ点をグループ化したものなのである。

3.3.2　完備統計量

定義 3.3.3 (完備性)**.** 十分統計量 T が**完備**であるとは、 すべての $\theta \in \Theta$ に対して $\mathbb{E}_\theta[g(T)] = 0$ ならば $\mathrm{P}_\theta(g(T) = 0) = 1$ がすべての θ で成り立つことをいう。

読み下し

完備性は、「T の関数で期待値がゼロになるのは 自明なもの（ほぼ確実にゼロの関数）だけ」という条件である。 言い換えれば、T は θ に関して「冗長な情報を含まない」。 この性質は一見抽象的だが、後述の Lehmann–Scheffé 定理により、 最小分散不偏推定量の一意性を保証するという 実践的に非常に重要な帰結を持つ。

定理 3.3.4 (Basu の定理)**.** *T が完備十分統計量で、V が補助統計量（V の分布が θ に依存しない）ならば、 T と V は独立である。*

読み下し

Basuの定理は「パラメータについてすべての情報を持つ統計量（完備十分）と、 パラメータについて何の情報も持たない統計量（補助）は独立である」 と述べている。 直観的には、両者の間に依存関係があるとすれば、 補助統計量が間接的にパラメータの情報を運ぶことになり矛盾するからである。

例 3.3.5 (正規分布における完備十分統計量と補助統計量)**.** $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$（$\sigma^2$ 既知）のとき、 $\bar{X}$ は μ に対する完備十分統計量であり、 $S^2 = (n-1)^{-1}\sum(X_i - \bar{X})^2$ は補助統計量である。 Basuの定理より $\bar{X}$ と S^2 は独立。

この独立性は t 検定の理論的基盤となる重要な事実である （第4章参照）。

3.4　指数型分布族の統計的性質

第1章で導入した 指数型分布族（定義1.5.3）は、 十分統計量と完備性に関して非常に優れた構造を持つ。 ここではその統計的性質をまとめる。

定理 3.4.1 (指数型分布族と十分統計量)**.** *自然パラメータ形式* $f(x \mid \boldsymbol{\eta}) = h(x)\exp(\boldsymbol{\eta}^\top \boldsymbol{T}(x) - A(\boldsymbol{\eta}))$ *のとき、* $\boldsymbol{T}(\boldsymbol{X}) = \sum_{i=1}^{n} \boldsymbol{T}(X_i)$ *は* $\boldsymbol{\eta}$ *に対する十分統計量である。*

読み下し

ファクトリゼーション定理（定理3.2.2）を適用すれば直ちにわかる。 指数型分布族の密度は $h(x)$（θ と無関係）と $\exp(\boldsymbol{\eta}^\top \boldsymbol{T}(x) - A(\boldsymbol{\eta}))$（$\boldsymbol{T}(x)$ と $\boldsymbol{\eta}$ の関数） に分解される。 したがって $\boldsymbol{T}$ は自動的に十分統計量となる。 これが指数型分布族の「計算上の恩恵」の根源である。

定理 3.4.2 (完備性)**.** *自然パラメータ空間* $\mathcal{H} = \{\boldsymbol{\eta} : A(\boldsymbol{\eta}) < \infty\}$ *が* $\mathbb{R}^k$ *の開集合を含むとき、* $\boldsymbol{T}$ *は完備十分統計量である。*

定理 3.4.3 (フィッシャー情報量と対数分配関数)**.** *指数型分布族において*

$$\mathcal{I}(\boldsymbol{\eta}) = \nabla^2 A(\boldsymbol{\eta}) = \mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{T}(X)).$$

読み下し

この結果は、指数型分布族においてフィッシャー情報量が 対数分配関数 $A(\boldsymbol{\eta})$ のヘッシアンに一致することを述べている。 A は凸関数（命題1.5.4）であるから、 $\mathcal{I}(\boldsymbol{\eta})$ は半正定値であり、情報量が非負であることと整合する。 対数分配関数を2回微分するだけでフィッシャー情報量が得られるという事実は、 一般化線形モデル（巻2）やベイズ推論（巻2第7章）での 計算を大幅に簡素化する。

実務ポイント

指数型分布族は統計学の中心的構成要素である。 正規分布、ポアソン分布、二項分布、ガンマ分布、ベータ分布など、 実務で最もよく用いる分布の多くがこの族に属する。 「十分統計量が自然に得られる」「完備性が保証される」 「フィッシャー情報量が対数分配関数から計算できる」という 三つの性質が同時に成り立つことが、指数型分布族の威力である。

3.5　フィッシャー情報量

推定の良し悪しを論じるには、 「データがパラメータについてどれだけの情報を運ぶか」を定量化する必要がある。 その役割を担うのがフィッシャー情報量である。

3.5.1　スコア関数

定義 3.5.1 (スコア関数)**.** 対数尤度 $\ell(\theta; X) = \log f(X \mid \theta)$ の θ に関する微分

$$s(\theta; X) = \frac{\partial}{\partial \theta} \ell(\theta; X)$$

を**スコア関数**と呼ぶ。

読み下し

スコア関数 $s(\theta; X)$ は「対数尤度の θ に関する傾き」であり、 データ X がパラメータの値をどの方向に引っ張るかを示す。 真のパラメータにおいてスコアの期待値がゼロになるという 次の命題は、最尤推定の理論的基盤である。

命題 3.5.2 (スコア関数の期待値)**.** *正則条件*[1] *の下で* $\mathbb{E}_\theta[s(\theta; X)] = 0$。

Proof. $\int f(x \mid \theta)\, dx = 1$ の両辺を θ で微分し、 微分と積分の交換が正当化されることを用いると、 $\int \frac{\partial}{\partial \theta} f(x \mid \theta)\, dx = 0$。 $\frac{\partial}{\partial \theta} f = f \cdot \frac{\partial}{\partial \theta} \log f = f \cdot s$ であるから、 $\mathbb{E}_\theta[s(\theta; X)] = 0$。 □

3.5.2　フィッシャー情報量の定義

定義 3.5.3 (フィッシャー情報量)**. フィッシャー情報量**を

$$\mathcal{I}(\theta) = \mathbb{E}_\theta[s(\theta; X)^2] = \mathrm{Var}_\theta(s(\theta; X)) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta^2} \ell(\theta; X)\right]$$

と定義する。2番目の等号はスコアの期待値が0であることから、 3番目の等号は正則条件の下で成立する。

読み下し

フィッシャー情報量はスコアの分散として定義される。 直観的には、スコアのばらつきが大きいほど 「データが θ の値を強く示唆する」ことを意味し、推定が容易になる。 3番目の等式は「対数尤度の曲率が大きいほど情報量が大きい」と読める。 対数尤度が鋭いピークを持てば θ の推定が精密にでき、 なだらかなピークなら推定が粗くなる、という直観と整合する。

定義 3.5.4 (フィッシャー情報行列)**.** パラメータ $\boldsymbol{\theta} \in \mathbb{R}^d$ の場合、**フィッシャー情報行列**の (j, k) 成分を

$$[\mathcal{I}(\boldsymbol{\theta})]_{jk} = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial \ell}{\partial \theta_j} \cdot \frac{\partial \ell}{\partial \theta_k}\right] = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}\right]$$

[1] 本章で繰り返し用いる「正則条件」（regularity conditions）とは、 以下の条件群を指す。厳密な定式化と検証は第6章で行う。

1. **支台のパラメータ非依存性**： $\{x : f(x \mid \theta) > 0\}$ が θ に依存しない。
2. **微分と積分の交換可能性**： $\frac{\partial}{\partial \theta} \int f(x \mid \theta)\, dx = \int \frac{\partial}{\partial \theta} f(x \mid \theta)\, dx$ が成立する（2次の微分についても同様）。
3. **識別可能性**： $\theta_1 \neq \theta_2 \implies f(\cdot \mid \theta_1) \neq f(\cdot \mid \theta_2)$。
4. **フィッシャー情報量の正則性**： $0 < \mathcal{I}(\theta) < \infty$。
5. **パラメータ空間の内点条件**： 真のパラメータ θ_0 が Θ の内点である。

一様分布 $\mathrm{Unif}(0, \theta)$ のように支台がパラメータに依存する場合、 条件(1)が破れるため Cramér–Rao 下界は適用できない （演習3.7参照）。

と定義する。

読み下し

スカラーパラメータ θ のフィッシャー情報量 $\mathcal{I}(\theta)$ は「1つの方向の推定精度」を表すスカラー値であった。パラメータが $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^\top$ のベクトルになると、各成分の推定精度だけでなく成分間の相互作用も重要になる。フィッシャー情報行列の (j,k) 成分 $[\mathcal{I}(\boldsymbol{\theta})]_{jk}$ は、θ_j と θ_k の同時推定における情報の結びつきを定量化する。対角成分 $[\mathcal{I}]_{jj}$ は θ_j 単独の情報量であり、非対角成分はパラメータ間の推定が互いにどれだけ「干渉」するかを表す。

命題 3.5.5 (n 個の i.i.d. 観測のフィッシャー情報量)**.** *$X_1, \dots, X_n$ が i.i.d. のとき、全体のフィッシャー情報量は $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$ である。*

読み下し

データが増えるほど情報量は線形に増大する。標本サイズを2倍にすれば、パラメータについての情報量も2倍になる。この加法性は、i.i.d. の仮定から自然に従う。

例 3.5.6 (ベルヌーイ分布のフィッシャー情報量)**.** $X \sim \mathrm{Bern}(p)$ のとき、$\ell(p;x) = x\log p + (1-x)\log(1-p)$、$s(p;x) = x/p - (1-x)/(1-p)$、$\mathcal{I}(p) = 1/(p(1-p))$。

p が 0 または 1 に近いほど情報量が大きい。これは直観と合う——コインがほぼ「表しか出ない」場合、1回投げるだけでそのことが高い確信度でわかるからである。逆に $p = 1/2$ のとき情報量は最小値 4 をとり、推定が最も難しい。

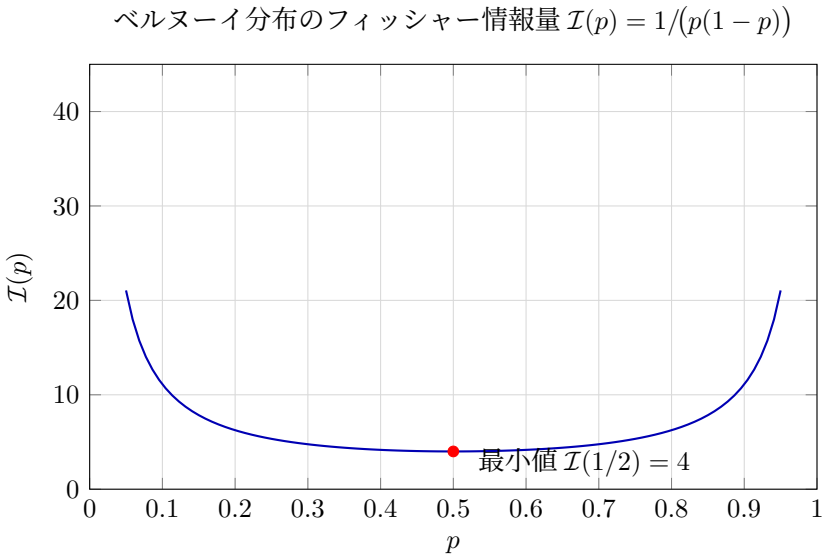


図 3.1: ベルヌーイ分布のフィッシャー情報量。p が 0 または 1 に近いほど情報量は増大し（1回の観測で p を高精度に推定できる）、$p = 1/2$ で最小値 4 をとる（推定が最も困難）。

3.6　最尤推定量（MLE）の構成と性質

3.6.1　尤度関数

定義 3.6.1 (尤度関数)**.** 観測データ $\boldsymbol{x} = (x_1, \dots, x_n)$ が得られたとき、θ の関数として

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

を**尤度関数**（likelihood function）と呼ぶ。対数をとった

$$\ell(\theta; \boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

を**対数尤度関数**と呼ぶ。

読み下し

尤度関数はデータを固定してパラメータを変数と見た関数であり、密度関数のパラメータとデータの役割を入れ替えたものである。「このパラメータ値のもとで、手元のデータが出現する尤もらしさはどの程度か」を測る。尤度は確率ではないことに注意——積分して1になる保証はない。

注意 3.6.2 (尤度原理)**.** **尤度原理**によれば、データから得られる θ に関する推論は尤度関数のみを通じて行うべきである。この原理を採用するかどうかは統計学の哲学的立場に依存するが、最尤推定法やベイズ推論は尤度原理と整合的である。

3.6.2　最尤推定量の定義

定義 3.6.3 (最尤推定量)**.** 最尤推定量（maximum likelihood estimator, MLE）$\hat{\theta}_n$ は

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} L(\theta; \boldsymbol{x})$$

と定義される。通常は対数尤度関数 $\ell(\theta) = \log L(\theta; \boldsymbol{x})$ を最大化する。

直観的理解

最尤推定の基本的な考え方は、「観測されたデータが最も起こりやすくなるようなパラメータ値を選ぶ」というものである。たとえば10回コインを投げて7回表が出たとき、「$p = 0.7$ のコインが最もこの結果を生みやすい」と考える。この原理は数学的には単純だが、実用的には非常に強力であり、漸近的に最適な性質を持つ（第6章で詳述）。

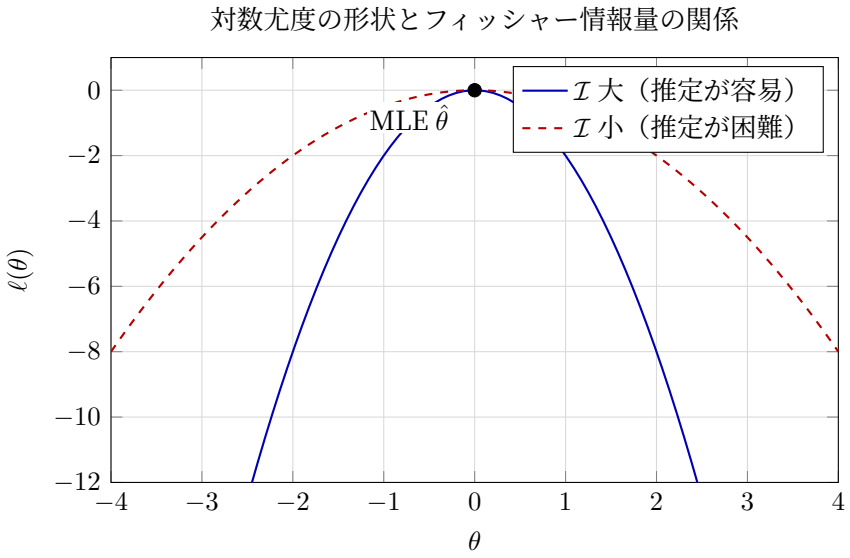


図 3.2: 対数尤度の曲率とフィッシャー情報量。 フィッシャー情報量 $\mathcal{I}(\theta)$ は対数尤度の曲率 $-\partial^2\ell/\partial\theta^2$ の期待値であり、 曲率が大きいほど（実線）ピークが鋭く推定が精密になる。 曲率が小さいと（破線）ピークがなだらかで推定の不確実性が大きい。

3.6.3　MLEの計算

対数尤度が微分可能な場合、MLEはスコア方程式

$$s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = 0$$

の解として見つかる。 多次元パラメータの場合、 $\nabla_\theta \ell(\theta) = \mathbf{0}$ の解を求める。

読み下し

スコア方程式の解がMLEとなるのは、対数尤度が凹関数の場合に限られる。 複数の極大値がある場合は、全極値を調べてグローバルな最大値を選ぶ必要がある。 指数型分布族の場合、対数尤度は凹であるため （$A(\boldsymbol{\eta})$ が凸であることから従う）、 スコア方程式の解は一意であり計算が容易である。

例 3.6.4 (正規分布のMLE). $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$（独立）のとき、対数尤度は

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

スコア方程式を解くと

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \tag{3.1}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \tag{3.2}$$

これより

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \quad \hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\hat{\mu}_n$ は不偏だが、$\hat{\sigma}^2_n$ は σ^2 の不偏推定量ではない（$\mathbb{E}[\hat{\sigma}^2_n] = (n-1)\sigma^2/n$）。MLEは不偏性を必ずしも保証しないという重要な事実の一例である。

3.6.4　MLEの存在性と不変性

定理 3.6.5 (MLEの存在性)**.** *次の条件の下で、MLEは存在する：*

1. *パラメータ空間 Θ がコンパクト集合*
2. *$\theta \mapsto \ell(\theta; \boldsymbol{x})$ が各 $\boldsymbol{x}$ に対して連続*

Proof. Θ がコンパクトで $\ell(\theta;\boldsymbol{x})$ が連続であるため、最大値定理（Weierstrass）より最大値が存在する。 □

定理 3.6.6 (MLEの不変性)**.** *$\hat{\theta}_n$ が θ のMLEであり、$\eta = g(\theta)$ がパラメータの関数であるとき、$\hat{\eta}_n = g(\hat{\theta}_n)$ は η のMLEである。*

読み下し

MLEの不変性は実用上便利な性質である。例えば、正規分布の分散 σ^2 のMLEが $\hat{\sigma}^2$ であれば、標準偏差 σ のMLEは $\sqrt{\hat{\sigma}^2}$ となる。わざわざ尤度を最大化し直す必要がない。

3.6.5　MLEの漸近性質の概要

MLEの漸近理論は現代統計学の基礎である。ここではその主要な結果を述べるにとどめ、厳密な条件と証明は第6章に委ねる。

定理 3.6.7 (MLEの一致性と漸近正規性)**.** *正則条件の下で、MLE $\hat{\theta}_n$ は*

1. $\hat{\theta}_n \xrightarrow{p} \theta_0$　　*（一致性：標本サイズが大きくなるにつれ、推定量は真の値に確率的に近づく）*
2. $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0)^{-1})$　　*（漸近正規性：真の値まわりの変動は正規分布で近似できる）*

を満たす。すなわち、大標本の下でMLEはCramér–Rao下界（定理3.10.5）を 漸近的に達成する。

読み下し

この結果は、MLEが漸近的に「最良の」推定量であることを示している。フィッシャー情報量 $\mathcal{I}(\theta_0)$ が大きいほど 漸近分散 $\mathcal{I}(\theta_0)^{-1}$ は小さくなり、推定が精密になる。漸近有効性の厳密な定式化と、正則条件の詳細、局所漸近正規性（LAN）の枠組みでの議論は第6章で展開する。

実務ポイント

漸近理論の結果は大標本近似であり、有限標本では近似の精度が問題となる。 特に以下の場合は注意が必要である：

- 分布が正規分布から大きく外れている場合（歪みや重い裾）
- パラメータが空間の境界付近にある場合（例：$p \approx 0$ のベルヌーイ分布）
- 標本サイズがパラメータ数に比して小さい場合

これらの場合には、ブートストラップやシミュレーションによる検証が推奨される。

実務ポイント

推定量の精度は、推定式そのものだけでなく「どの情報を事前に使うか」にも依存する。 たとえばランダム化比較試験では、予後的なベースライン共変量を あらかじめ指定して治療効果の推定に組み込むと、 比較可能性を保ったまま分散を下げられることが多い。 これは観察研究での交絡調整とは役割が異なり、 ランダム化で確保された設計の上で効率を改善する発想である。 実務では、どの共変量を使うかだけでなく、 何を推定対象（estimand）とし、どの主推定量で推定し、 どの感度分析で確かめるかを計画段階でそろえておくことが重要である。

3.7　モーメント法

モーメント法（method of moments）は、 MLEとは異なるアプローチで推定量を構成する古典的な方法である。

定義 3.7.1 (モーメント法推定量)**.** 母集団の k 番目のモーメントを $\mu_k = \mathbb{E}[X^k]$ とし、 標本モーメントを $\hat{\mu}_k = n^{-1}\sum_{i=1}^{n} X_i^k$ とするとき、 モーメント法推定量は

$$\mu_k(\hat{\theta}_n) = \hat{\mu}_k, \quad k = 1, \dots, p$$

を満たす $\hat{\theta}_n$ である。

読み下し

モーメント法の考え方は単純である——「母集団のモーメントとパラメータの関係式に、 標本モーメントを代入して解く」。 例えば、正規分布 $\mathcal{N}(\mu, \sigma^2)$ では $\mathbb{E}[X] = \mu$, $\mathbb{E}[X^2] = \mu^2 + \sigma^2$ であるから、 $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = n^{-1}\sum X_i^2 - \bar{X}^2$ が得られる。

直観的理解

モーメント法は直感的で実装が簡単であるが、 一般にはMLEほど効率的ではない。 その理由は、モーメント法がデータの要約統計量（モーメント）のみを使うのに対し、MLEは尤度関数全体の形状を利用するからである。 ただし、MLEの計算が困難な場合

にモーメント法は有力な代替手段となる。 また、モーメント法推定量をMLEの計算の初期値として用いることも多い。

例 3.7.2 (ガンマ分布のモーメント法)**.** Gamma(α, β) 分布の場合、$\mathbb{E}[X] = \alpha/\beta$, $\mathrm{Var}(X) = \alpha/\beta^2$ である。 モーメント法では

$$\bar{X} = \frac{\hat{\alpha}}{\hat{\beta}}, \quad S^2 = \frac{\hat{\alpha}}{\hat{\beta}^2} \tag{3.3}$$

これより

$$\hat{\alpha} = \frac{\bar{X}^2}{S^2}, \quad \hat{\beta} = \frac{\bar{X}}{S^2}$$

実務ポイント

データが重い裾を持つ分布（例：t 分布で自由度が小さい場合）から来ているとき、 高次モーメントの標本推定は不安定になる。 このような場合、モーメント法は性能が劣化する。 一方で、1次・2次のモーメントのみを用いるモーメント法推定量は 比較的安定であり、ロバストな推定の出発点として有用である。

3.8　M推定量

M推定量は、最尤推定量をより一般的な形で拡張した推定量のクラスである。 この枠組みにより、ロバスト推定や正則化推定など多様な推定法を 統一的に扱うことができる。

定義 3.8.1 (M推定量)**.** 損失関数 $\rho : \mathbb{R} \times \Theta \to \mathbb{R}$ に対して、 M推定量 $\hat{\theta}_n$ は

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \rho(X_i, \theta)$$

と定義される。同値的に、$\psi = \partial\rho/\partial\theta$ として 推定方程式

$$\sum_{i=1}^{n} \psi(X_i; \hat{\theta}_n) = \mathbf{0}$$

の解として特徴づけられる。 最尤推定の場合、$\rho(x, \theta) = -\log f(x \mid \theta)$ （すなわち ψ はスコア関数）である。

読み下し

M推定量の「M」は maximum（最大化）に由来する（歴史的には minimum でもある）。 MLEをスコア方程式の解として見たとき、 スコア関数を他の ψ に置き換えることで 異なる性質を持つ推定量が得られる。 特に重要なのは、ψ を有界にすることで外れ値の影響を制限するロバスト推定への応用である。

例 3.8.2 (Huber推定量)**.** 位置パラメータ θ のロバスト推定のため、 Huber関数を用いたM推定量を考える：

$$\psi_k(x;\theta) = \begin{cases} x-\theta & \text{if } |x-\theta| \le k \\ k\cdot \operatorname{sgn}(x-\theta) & \text{if } |x-\theta| > k \end{cases}$$

ここで $\operatorname{sgn}(u)$ は符号関数（$u>0$ のとき $+1$、$u<0$ のとき -1、$u=0$ のとき 0）、 $k>0$ は調節パラメータである。 $|x-\theta|\le k$ の範囲では通常の平均のように振る舞い（効率性）、 $|x-\theta|>k$ の範囲では影響を定数 k で打ち切る（ロバスト性）。 $k\to\infty$ で標本平均、$k\to 0$ で標本中央値に帰着する。

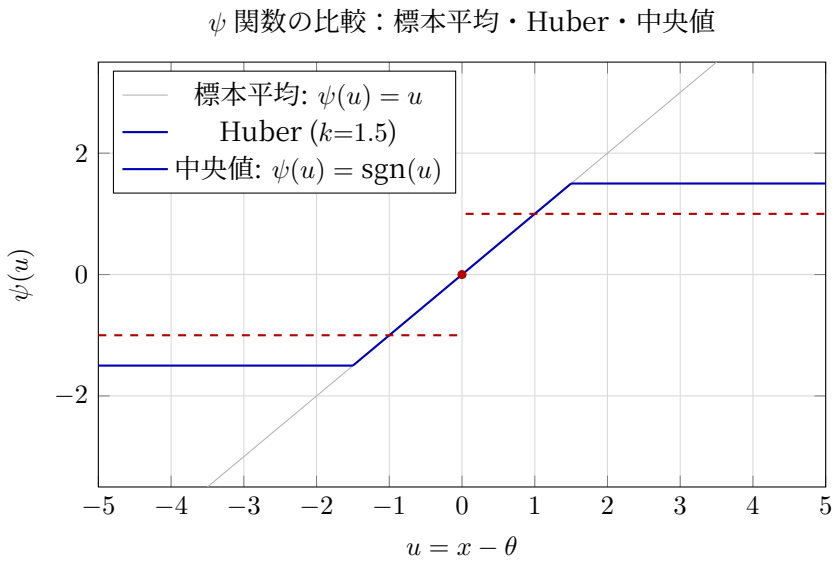


図 3.3: 3種の ψ 関数の比較。 標本平均の $\psi(u)=u$ は外れ値の影響を無制限に受ける。 Huber関数は $|u|\le k$ では恒等関数（効率性）、 $|u|>k$ では定数（ロバスト性）として両者を折衷する。 中央値の $\psi(u)=\operatorname{sgn}(u)$ は最もロバストだが正規分布の下で効率が低い。

例 3.8.3 (分位点回帰のM推定量)**.** τ-分位点 q_τ の推定は、非対称絶対値損失

$$\rho_\tau(u) = u(\tau - \mathbf{1}\{u<0\})$$

の最小化として定式化される。 これもM推定量の枠組みに収まり、 漸近理論が統一的に適用できる。 分位点回帰は金融リスク管理における VaR（Value at Risk）の推定や、 賃金分布の下位・上位層に異なる効果を持つ政策の分析など、 「平均」では捉えられない分布の特定位置での関係を明らかにしたい場面で広く用いられる。

定理 3.8.4 (M推定量の漸近正規性)**.** *正則条件の下で、M推定量 $\hat\theta_n$ は*

$$\sqrt{n}(\hat\theta_n-\theta_0) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0},\; H^{-1}\Sigma (H^{-1})^{\top}\Big)$$

を満たす。ここで $H=\mathbb{E}[\nabla_\theta \psi(X,\theta_0)]$、 $\Sigma=\mathbb{E}[\psi(X,\theta_0)\psi(X,\theta_0)^{\top}]$ である。$H^{-1}\Sigma(H^{-1})^{\top}$ を **サンドイッチ分散** *と呼ぶ。*

読み下し

サンドイッチ分散 $H^{-1}\Sigma(H^{-1})^\top$ の構造を直観的に理解しよう。 $\Sigma = \mathbb{E}[\psi\psi^\top]$ は推定方程式の「ノイズの大きさ」を表し、 $H = \mathbb{E}[\nabla\psi]$ は推定方程式の「感度」（θ を少し動かしたとき ψ がどれだけ変化するか）を表す。 H^{-1} で両側から挟む操作は、 ノイズ Σ を感度 H のスケールに変換することに対応する。 感度が高い（H が大きい）ほど、同じノイズでもパラメータの推定精度は高くなる。
MLEの場合は $H = -\mathcal{I}(\theta_0)$、 $\Sigma = \mathcal{I}(\theta_0)$ であるから サンドイッチ分散は $\mathcal{I}(\theta_0)^{-1}$ に簡約化される。 一般のM推定量ではこの簡約化が起きないため、 H と Σ を別々に推定する必要がある。 この「サンドイッチ標準誤差」は、モデルが正しく特定されていない場合でも 有効な標準誤差を与えるため、 計量経済学やロバスト統計学で広く用いられる。 詳細は第6章で扱う。

実務ポイント

表3.1に 本章で学んだ3つの推定法の特徴を整理する。 手法選択の出発点として活用してほしい。

表 3.1: 推定法の特徴比較

	MLE	モーメント法	M推定量
漸近効率性	最良（CR下界達成）	一般に劣る	ψ の選択に依存
計算容易性	モデル依存（反復法要）	高い（閉じた式多い）	中程度（反復法要）
ロバスト性	低い（外れ値に敏感）	低い	高い（ψ が有界）
不偏性	一般に不偏でない	近似的に不偏	一般に不偏でない
適用場面	モデルが正しく特定されている場合	初期推定値の計算、探索的分析	外れ値やモデルの誤特定が懸念される場合

実務では、まずモーメント法で初期推定値を求め、 それをMLEの数値最適化の初期値に用いるのが定番の手順である。 外れ値の存在が疑われるデータに対しては、 Huber型のM推定量が効率性とロバスト性のバランスに優れた選択となる。
なお、M推定量の実用上の注意点として以下がある：

- **計算コスト**：M推定量は一般に反復法（IRLS: 反復再重み付き最小二乗法など）で求めるため、 MLEと同程度の計算負荷がかかる。高次元では1反復あたりの計算量も $O(nd)$ に増大する。
- **初期値依存性**：ψ 関数が非凸の場合、反復法は初期値に依存して局所解に収束しうる。 実務ではモーメント法や中央値で初期推定値を得てから反復を開始するのが安全な手順である。

3.9　U統計量

U統計量（U-statistic）は、 対称な関数に基づいた不偏推定量の一般的なクラスであり、 母集団の汎関数を不偏に推定するための標準的な道具である。

定義 3.9.1 (U統計量)**.** パラメータ $\theta = \mathbb{E}[h(X_1, \ldots, X_r)]$ を 対称カーネル $h : \mathbb{R}^r \to \mathbb{R}$ で表した

とき、θ のU統計量は

$$U_n = \binom{n}{r}^{-1} \sum_{1 \le i_1 < \cdots < i_r \le n} h(X_{i_1}, \ldots, X_{i_r})$$

と定義される。r を**次数**（order）と呼ぶ。

読み下し

U統計量は「データの r 個の組み合わせすべてにカーネル h を適用して平均する」という構成である。これにより $\mathbb{E}[U_n] = \mathbb{E}[h(X_1, \ldots, X_r)] = \theta$ が自動的に成り立ち、不偏推定量が系統的に構成できる。

例 3.9.2 (U統計量としての標本分散)**.** *母集団分散 $\sigma^2 = \mathbb{E}[(X_1 - X_2)^2/2]$ を推定する場合、カーネル $h(x_1, x_2) = (x_1 - x_2)^2/2$（次数 $r = 2$）に対するU統計量は*

$$U_n = \binom{n}{2}^{-1} \sum_{i<j} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = s^2$$

となり、Bessel補正付きの不偏分散に一致する。

例 3.9.3 (Wilcoxon順位和統計量)**.** *二標本問題で $\theta = \mathrm{P}(X > Y)$ を推定するU統計量はカーネル $h(x, y) = \mathbf{1}\{x > y\}$ に対応し、$U = (n_1 n_2)^{-1} \sum_{i,j} \mathbf{1}\{X_i > Y_j\}$ はWilcoxon順位和検定の基礎となる統計量である。*

定理 3.9.4 (U統計量の漸近正規性)**.** *次数 r のU統計量 U_n を考える。*

$$\sigma_1^2 = \mathrm{Var}(\mathbb{E}[h(X_1, \ldots, X_r) \mid X_1]) > 0$$

とおくと、

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, r^2 \sigma_1^2).$$

読み下し

漸近分散が σ_1^2（カーネルの第1引数に関する条件付き期待値の分散）のみで決まることが注目に値する。次数 r の全組み合わせを使うが、漸近的には各データ点の「1次寄与」のみが支配的となるのである。この性質はHoeffdingの射影（projection）として知られる。

コード例：U統計量としてのWilcoxon統計量（R）

```
set.seed(789)
x <- rnorm(30, mean = 0.5)  # 処理群
y <- rnorm(25, mean = 0)    # 対照群

# U統計量を定義から直接計算
u_stat <- mean(outer(x, y, FUN = function(a, b) as.numeric(a > b)))
cat("U統計量 (定義から):", round(u_stat, 4), "\n")

# R組み込みのWilcoxon検定と比較
wtest <- wilcox.test(x, y)
u_builtin <- wtest$statistic / (length(x) * length(y))
```

```
cat("U統計量 (wilcox.test):", round(u_builtin, 4), "\n")
# 両者は一致する
```

3.10　不偏性とCramér–Raoの下界

推定量を構成する方法をいくつか学んだ。では、推定量の「良さ」をどう評価すればよいか？本節では不偏性を中心とした評価基準と、推定量の分散の理論的下限を与える Cramér–Rao の下界を導く。

3.10.1　不偏推定量と効率性

定義 3.10.1 (不偏推定量)**.** 推定量 $\hat{\theta}_n$ が**不偏**であるとは、

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

がすべての $\theta \in \Theta$ で成立することである。

読み下し

不偏性は「平均的には真の値に一致する」という性質である。ただし、不偏性だけでは推定量の良さは保証されない。分散が大きければ、個々の推定値は真の値から大きく外れうる。不偏推定量の中で分散が最小のものが最も望ましい。

定義 3.10.2 (効率性)**.** 二つの不偏推定量 $\hat{\theta}_1, \hat{\theta}_2$ について、すべての θ で $\mathrm{Var}(\hat{\theta}_1) \le \mathrm{Var}(\hat{\theta}_2)$ が成立するとき、$\hat{\theta}_1$ は $\hat{\theta}_2$ より**効率的な推定量**であるという。

3.10.2　Rao–Blackwell定理

十分統計量を用いて推定量を改善する方法を与える。

定理 3.10.3 (Rao–Blackwell)**.** *$T = T(\boldsymbol{X})$ が十分統計量で、$\hat{\theta}$ が θ_0 の不偏推定量であるとき、*

$$\hat{\theta}^* = \mathbb{E}[\hat{\theta} \mid T]$$

は θ_0 の不偏推定量であり、かつすべての θ に対して

$$\mathrm{Var}_\theta(\hat{\theta}^*) \le \mathrm{Var}_\theta(\hat{\theta})$$

が成立する。等号は $\hat{\theta}$ がすでに T の関数であるとき、かつそのときに限り成立する。

Proof. **方針**：条件付き期待値の基本性質（繰り返しの法則と分散分解公式）だけで示せる。

$\hat{\theta}^* = g(T)$ と書ける。不偏性は

$$\mathbb{E}[\hat{\theta}^*] = \mathbb{E}[\mathbb{E}[\hat{\theta} \mid T]] = \mathbb{E}[\hat{\theta}] = \theta_0$$

による（繰り返しの法則）。分散の不等式は条件付き分散の公式

$$\mathrm{Var}(\hat{\theta}) = \underbrace{\mathbb{E}[\mathrm{Var}(\hat{\theta} \mid T)]}_{\geq 0} + \underbrace{\mathrm{Var}(\mathbb{E}[\hat{\theta} \mid T])}_{=\mathrm{Var}(\hat{\theta}^*)} \geq \mathrm{Var}(\hat{\theta}^*)$$

から従う。第1項は非負であり、$\hat{\theta}$ が T で条件づけても なおばらつく分だけ分散が大きいことを示している。 □

読み下し

Rao–Blackwell定理は、任意の不偏推定量を十分統計量で条件づけることで 分散を減らせることを保証する。 直観的には、十分統計量が持つ「余計な情報」を使って 推定量のノイズを取り除く操作に相当する。

3.10.3　Lehmann–Scheffé定理

定理 3.10.4 (Lehmann–Scheffé). *$T = T(\boldsymbol{X})$ が θ の完備十分統計量であり、$\hat{\theta}$ が θ_0 の不偏推定量であるとき、$\hat{\theta}^* = \mathbb{E}[\hat{\theta} \mid T]$ は θ_0 の一意な最小分散不偏推定量（UMVUE）である。*

読み下し

完備十分統計量 T の関数として表される不偏推定量は 一意であり（完備性による）、かつRao–Blackwell定理により分散が最小である。 したがって、完備十分統計量さえ見つかれば、 UMVUEの構成は $\mathbb{E}[\hat{\theta} \mid T]$ の計算に帰着される。 指数型分布族の多くは完備十分統計量を持つため （定理3.4.2）、 これらの分布に対するUMVUEは系統的に求められる。

3.10.4　Cramér–Raoの下界

定理 3.10.5 (Cramér–Rao下界). *正則条件の下で、θ_0 の任意の不偏推定量 $\hat{\theta}_n$ に対して*

$$\mathrm{Var}(\hat{\theta}_n) \geq \frac{1}{n\mathcal{I}(\theta_0)}$$

が成立する。ここで$\mathcal{I}(\theta)$ は 単一観測のフィッシャー情報量（定義3.5.3）である。

読み下し

Cramér–Rao下界は 「推定量の分散はフィッシャー情報量の逆数以下には下がらない」 と述べている。 情報量が大きいほど下界は小さくなり、推定が精密にできる。 逆に情報量が小さいと、どんな不偏推定量を用いても分散を ある水準以下には下げられない—— これは推定問題の本質的な困難さの指標である。

Proof. **方針**：不偏性の条件をパラメータで微分し、 Cauchy–Schwarzの不等式を適用する。 鍵は「推定量とスコア関数の共分散が1になる」という事実である。

不偏性の条件 $\mathbb{E}[\hat{\theta}_n] = \theta_0$ をパラメータで微分すると （微分と期待値の順序交換を仮定）

$$\frac{\partial}{\partial \theta_0} \mathbb{E}[\hat{\theta}_n] = 1$$

左辺を計算すると、 $\mathbb{E}[\hat{\theta}_n] = \int \hat{\theta}_n \prod_i f(x_i \mid \theta)\, d\boldsymbol{x}$ を θ で微分して

$$\mathrm{Cov}\left(\hat{\theta}_n,\ \frac{\partial \log L}{\partial \theta_0} \right) = 1$$

が得られる（スコアの期待値が0であることを使った）。

この等式は証明の核心であり、次のように読める：「不偏推定量は、スコア関数（対数尤度の傾き）と 共分散がちょうど1でなければならない。 つまり、不偏であるためには推定量がデータの尤度の変化に 一定の感度で応答する必要がある。」

Cauchy–Schwarzの不等式を適用すると

$$1 = \left| \mathrm{Cov}\left(\hat{\theta}_n,\ \frac{\partial \log L}{\partial \theta_0} \right) \right|^2 \le \mathrm{Var}(\hat{\theta}_n) \cdot \mathrm{Var}\left(\frac{\partial \log L}{\partial \theta_0} \right) = \mathrm{Var}(\hat{\theta}_n) \cdot n\mathcal{I}(\theta_0)$$

よって

$$\mathrm{Var}(\hat{\theta}_n) \ge \frac{1}{n\mathcal{I}(\theta_0)}$$

□

例 3.10.6 (正規分布の平均のCR下界)**.** $\mathcal{N}(\mu, \sigma^2)$ 分布で σ^2 既知の場合、フィッシャー情報量は $\mathcal{I}(\mu) = 1/\sigma^2$ である。 したがって、μ の任意の不偏推定量は

$$\mathrm{Var}(\hat{\mu}) \ge \frac{\sigma^2}{n}$$

を満たす。標本平均 $\bar{X}$ は $\mathrm{Var}(\bar{X}) = \sigma^2/n$ であり、 この下界を達成する**効率的な推定量**である。

例 3.10.7 (ベルヌーイ分布のCR下界)**.** $X_1, \dots, X_n \sim \mathrm{Bern}(p)$ のとき、 $\mathcal{I}(p) = 1/(p(1-p))$ であるから、 p の任意の不偏推定量の分散は $p(1-p)/n$ 以上である。 標本比率 $\hat{p} = \bar{X}$ は $\mathrm{Var}(\hat{p}) = p(1-p)/n$ であり、 CR下界を達成する。

実務ポイント

CR下界は不偏推定量に対する下界であるが、 バイアスを許容すれば分散をさらに下げられる場合がある。 推定量のMSE（平均二乗誤差）は $\mathrm{MSE} = \mathrm{Var} + \mathrm{Bias}^2$ と分解されるため、 わずかなバイアスと引き換えに分散を大幅に下げれば MSEは改善される。この考え方は、正則化推定（巻2）や James–Stein推定量（第5章）の理論的基盤となる。

3.11　実践：推定量の有限標本比較

本節では、各種推定量（MLE、モーメント法、ロバストM推定量）の 有限標本における振る舞いをシミュレーションで比較する。

3.11.1　正規分布の平均推定：CR下界の検証

コード例：CR下界の数値的検証（R）

```
set.seed(42)
n <- 30; mu <- 5; sigma <- 2
B <- 10000

# 標本平均のシミュレーション
xbar <- replicate(B, mean(rnorm(n, mu, sigma)))

# 理論値との比較
cat("標本分散:", var(xbar), "\n")
cat("CR下界 (sigma^2/n):", sigma^2 / n, "\n")
# 両者はほぼ一致する（標本平均は効率的推定量）
```

3.11.2　ガンマ分布：MLEとモーメント法の効率比較

コード例：ガンマ分布の推定量比較（Python）

```
import numpy as np
from scipy.stats import gamma

rng = np.random.default_rng(123)
alpha_true, beta_true = 2.0, 1.5
n, B = 50, 5000

mse_mom = np.zeros(B)
mse_mle = np.zeros(B)

for b in range(B):
    x = gamma.rvs(alpha_true, scale=1/beta_true, size=n, random_state=rng)

    # モーメント法
    xbar, s2 = x.mean(), x.var(ddof=1)
    alpha_mom = xbar**2 / s2
    mse_mom[b] = (alpha_mom - alpha_true)**2

    # MLE（scipy.stats.gamma.fit を利用）
    # 注: scipy の gamma は (shape=alpha, loc, scale) のパラメータ化を用いる。
    # 統計学の慣例では rate beta = 1/scale であるため、
    # floc=0 で位置パラメータを固定し shape のみを比較する。
    alpha_mle, _, scale_mle = gamma.fit(x, floc=0)
    mse_mle[b] = (alpha_mle - alpha_true)**2

print(f"モーメント法 MSE: {mse_mom.mean():.4f}")
print(f"MLE MSE:          {mse_mle.mean():.4f}")
print(f"効率比 (MoM/MLE): {mse_mom.mean()/mse_mle.mean():.2f}")
```

実務ポイント

ガンマ分布の形状パラメータ α の推定では、 MLEがモーメント法より一貫して効率的であることが シミュレーションで確認できる。 ただし、外れ値が存在する場合はHuber 型のM推定量が 両者より安定した推定を与えることが多い。

3.11.3　ロバスト推定：外れ値のある場合

コード例：外れ値がある場合の位置推定（R）

```
library(MASS)
set.seed(456)
n <- 100; mu <- 0; sigma <- 2

# 10%の外れ値を混合した汚染正規分布
x <- c(rnorm(90, mu, sigma), rnorm(10, mu + 10, sigma))

# 三つの推定量を比較
est_mean   <- mean(x)                # 標本平均
est_median <- median(x)              # 中央値
est_huber  <- huber(x)$mu            # Huber推定量

cat("標本平均:", round(est_mean, 3), "\n")
cat("中央値:  ", round(est_median, 3), "\n")
cat("Huber:   ", round(est_huber, 3), "\n")
cat("真の値:  ", mu, "\n")
# 外れ値がある場合、標本平均は大きくバイアスされるが、
# 中央値とHuber推定量はロバストに真の値に近い推定を与える
```

主要な結果

重要結果

本章の核心的な結論：

1. **統計モデル**を置くことで、 「何を未知パラメータとみなすか」「どの情報がデータに含まれているか」を明確にできる。
2. **十分統計量・最小十分統計量・完備統計量**は、 データを損失なく縮約し、不偏推定量の最適化を可能にする。 Rao–Blackwell 定理と Lehmann–Scheffé 定理はその代表的帰結である。
3. **指数型分布族**は、 十分性・完備性・フィッシャー情報量がきれいに結びつく中心的なクラスである。
4. **最尤推定量・モーメント法・M推定量・U統計量**は それぞれ計算容易性、効率性、ロバスト性、対称統計量の理論という別々の強みを持つ。 さらに、設計段階で利用する共変量情報も推定精度を左右する。
5. **Cramér–Rao下界**は推定精度の限界を与えるが、 境界問題や非正則モデルでは

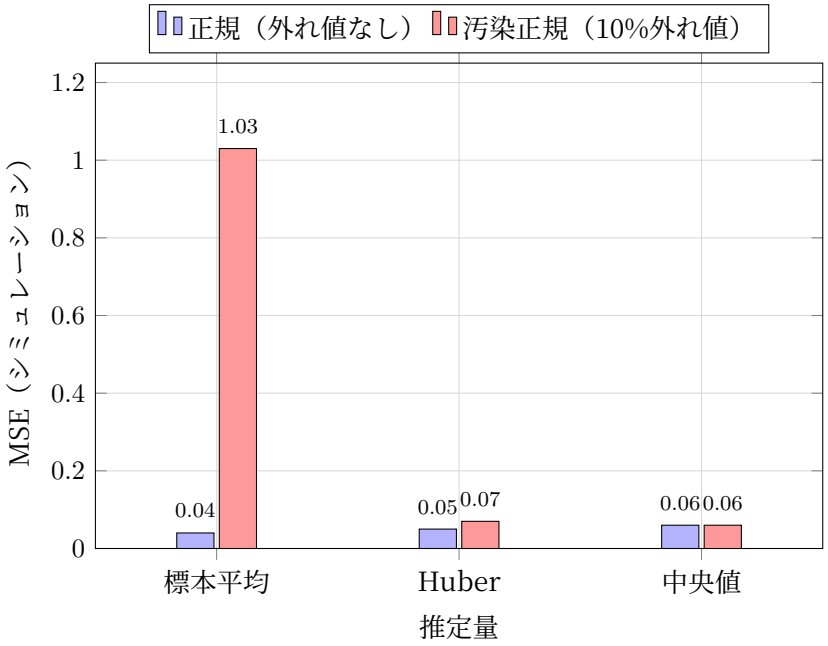


図 3.4: 外れ値の有無による推定量のMSE比較（位置パラメータ推定、$n = 100$）。 正規分布の下では標本平均が最も効率的だが、 10%の外れ値が混入するとMSEが急増する。 Huber推定量と中央値はロバストにMSEが低く保たれる。 数値は上記コード例に基づく典型的な結果を示す。

> そのまま使えない。 したがって、有限標本での構造理解と漸近理論の両方が必要になる。

3.12　演習問題

基礎理論

演習問題 3.1. $X_1, \dots, X_n$ が i.i.d. $\mathrm{Unif}(0, \theta)$ のとき、 $T = X_{(n)} = \max(X_1, \dots, X_n)$ が θ に対する十分統計量であることを示せ。 この統計量は最小十分統計量か？

演習問題 3.2. ポアソン分布 $\mathrm{Pois}(\lambda)$ からの i.i.d. 標本 $X_1, \dots, X_n$ に対して $T = \sum_{i=1}^{n} X_i$ が完備十分統計量であることを示せ。

演習問題 3.3. $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ のフィッシャー情報行列を $\boldsymbol{\theta} = (\mu, \sigma^2)$ について計算せよ。

演習問題 3.4. Basuの定理を用いて次を示せ： $X_1, \dots, X_n \sim \mathrm{Exp}(\lambda)$ のとき、 $\bar{X}$ と $(X_1/\sum X_i, \dots, X_n/\sum X_i)$ は独立である。

演習問題 3.5. 対数正規分布 $\mathrm{LogNormal}(\mu, \sigma^2)$ の密度を2次元指数型分布族の形に書き、 自然パラメータ、十分統計量、自然パラメータ空間を求めよ。

推定問題

演習問題 3.6. 指数分布 $\mathrm{Exp}(\lambda)$ から n 個の独立な標本 $X_1, \dots, X_n$ を得た。

1. 最尤推定量 $\hat{\lambda}_n$ を求めよ。
2. $\hat{\lambda}_n$ の期待値と分散を求めよ。
3. フィッシャー情報量 $\mathcal{I}(\lambda)$ を計算し、Cramér–Rao下界を示せ。

演習問題 3.7. 一様分布 $\mathrm{Unif}(0, \theta)$ から n 個の標本を得た。

1. 最尤推定量 $\hat{\theta}_n = \max(X_1, \dots, X_n)$ の分布を求めよ。
2. $\hat{\theta}_n$ は不偏か？もし不偏でなければ、不偏推定量を構成せよ。
3. モーメント法による推定量を求めよ。
4. CR下界がこの問題に適用できない理由を説明せよ。（ヒント：正則条件の何が破れるか）

演習問題 3.8. ポアソン分布 $\mathrm{Pois}(\lambda)$ から n 個の標本を得た。

1. モーメント法とMLEを比較せよ。
2. 分散 $\mathrm{Var}(X) = \lambda$ を推定する場合、どの推定量が優れているか、分散を計算して比較せよ。

演習問題 3.9. ベイズ推定量 $\hat{\theta}_B = \mathbb{E}[\theta \mid \boldsymbol{X}]$（事前分布 $\theta \sim \mathrm{Beta}(a, b)$）について：

1. ベルヌーイ試行 $\mathrm{Bern}(p)$ の場合、p のベイズ推定量を求めよ。
2. 大標本の極限でこのベイズ推定量はMLE $\hat{p}_n = \bar{X}$ に収束するか調べよ。

計算・実装問題

数値実験を含む問題では、(1) 設定（分布、パラメータ、反復回数、乱数シード）、(2) 作成した図表または表、(3) 比較指標、(4) 結果から読める一言考察、の4点を答えに含めよ。

演習問題 3.10. 次の手順に従いシミュレーション研究を設計せよ：

1. $\mathcal{N}(0, 1)$ から $n = 50$ 個の標本を1000回抽出する。
2. 各抽出に対して標本平均 $\bar{X}$ と標本中央値 $\mathrm{med}(X)$ を計算する。
3. これら2つの推定量の標本分布を比較し、MSEを計算せよ。
4. 正規分布のもとでは標本平均が効率的であるが、t_3 分布のもとではどちらが優れるか、同様のシミュレーションで調べよ。

略解の指針

ここでは、各問題をどの定理につなげればよいかと、到達点だけを整理する。

- **演習3.1** 使う道具: 因子分解定理と最小十分性の判定。 最初の1手: 同時密度を $\theta^{-n}\mathbf{1}(0 < x_{(1)}, x_{(n)} < \theta)$ と書き、 $x_{(n)}$ のみで θ 依存が決まることを見る。 途中の要点: 比 $f_\theta(x)/f_\theta(y)$ が θ に依らないための条件は $x_{(n)} = y_{(n)}$ である。 最終形: $X_{(n)}$ は十分かつ最小十分統計量である。

- **演習3.2** 使う道具: ポアソン和の分布とべき級数の一意性。 最初の1手: $T = \sum_i X_i \sim \mathrm{Pois}(n\lambda)$ であり、同時密度は T を通じて因子化される。 途中の要点: $\mathbb{E}_\lambda[g(T)] = 0$ が全 λ で成り立つなら、 $\sum_t g(t)(n\lambda)^t/t! \equiv 0$ だから各係数が 0。 最終形: T は完備十分統計量である。

- **演習3.3** 使う道具: スコアと二階微分。 最初の1手: 対数尤度を μ と σ^2 で偏微分する。 途中の要点: 交差項の期待値は 0 になり、行列は対角化する。 最終形: $\mathcal{I}_n(\mu, \sigma^2) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix}$。

- **演習3.4** 使う道具: Basu の定理。 最初の1手: $T = \sum_i X_i$ が λ に対して完備十分であることを確認する。 途中の要点: 比率ベクトル $(X_i/T)_i$ の分布は尺度変換で λ に依らないので補助統計量である。 最終形: Basu の定理により $\bar{X} = T/n$ と比率ベクトルは独立である。

- **演習3.5** 使う道具: 指数型分布族への書き換え。 最初の1手: 対数正規密度の指数部を $\eta_1 \log x + \eta_2 (\log x)^2 - A(\eta)$ の形に整理する。 途中の要点: $\eta_1 = \mu/\sigma^2$, $\eta_2 = -1/(2\sigma^2)$, $T(x) = (\log x, (\log x)^2)$, $h(x) = 1/x$ と読める。 最終形: 自然パラメータ空間は $\eta_2 < 0$ を満たす半平面である。

- **演習3.6** 使う道具: ガンマ分布の逆モーメント。 最初の1手: 対数尤度から $\hat{\lambda}_n = 1/\bar{X} = n/\sum_i X_i$ を得る。 途中の要点: $\sum_i X_i \sim \mathrm{Gamma}(n, \lambda)$ を使って $\mathbb{E}[\hat{\lambda}_n] = n\lambda/(n-1)$, $\mathrm{Var}(\hat{\lambda}_n) = n^2\lambda^2/((n-1)^2(n-2))$ を計算する。 最終形: 1標本あたりのフィッシャー情報量は $1/\lambda^2$, CR 下界は λ^2/n である。

- **演習3.7** 使う道具: 最大値の分布。 最初の1手: $\mathrm{P}(X_{(n)} \le t) = (t/\theta)^n$ for $0 < t < \theta$ を出す。 途中の要点: $\mathbb{E}[X_{(n)}] = n\theta/(n+1)$ だから $(n+1)X_{(n)}/n$ が不偏推定量になる。 最終形: モーメント法推定量は $2\bar{X}$ であり、 CR 下界が使えない原因は support が θ に依存することである。

- **演習3.8** 使う道具: モーメント法と MLE の比較。 最初の1手: ポアソンでは $\mathbb{E}[X] = \lambda$ なのでモーメント法も MLE も $\bar{X}$ になる。 途中の要点: $\mathrm{Var}(\bar{X}) = \lambda/n$ であり、 分散 $\mathrm{Var}(X) = \lambda$ の推定でも同じ推定量に帰着する。 最終形: この問題では両者は一致し、優劣はつかない。

- **演習3.9** 使う道具: Beta-Bernoulli 共役。 最初の1手: $S = \sum_i X_i$ とおけば事後分布は $\mathrm{Beta}(a+S, b+n-S)$ である。 途中の要点: 事後平均は $(a+S)/(a+b+n)$ であり、 $n \to \infty$ で $S/n \to p$ なら $\bar{X}$ に近づく。 最終形: ベイズ推定量は $(a+S)/(a+b+n)$, 大標本では MLE と一致する。

- **演習3.10** 使う道具: 経験 MSE 比較。 最初の1手: 正規分布と t_3 分布で同じ実験を回し、 各回の $(\hat{\theta}-\theta)^2$ を平均する。 途中の要点: 正規分布では平均が効率的、重尾分布では中央値のロバスト性が効いてくる。 最終形: 分布形が変わると最適推定量も変わることを数値で確認する問題である。

次章への橋渡し

本章では、統計モデルの設定からパラメータの推定方法、 そして推定量の最適性まで一貫して扱った。 推定量が「どれだけ良いか」は フィッシャー情報量とCramér–Rao下界で定量化できることを見た。

しかし、推定はデータ分析の一側面にすぎない。 実務ではしばしば「あるパラメータ値は妥当か？」という**検定**や、 「パラメータはどの範囲にあるか？」という**信頼区間**の構成が求められる。 第4章では、Neyman–Pearson理論に基づく 最適な検定の構成と、検定と信頼集合の間の深い双対性を明らかにする。 本章で導入したフィッシャー情報量・十分統計量・尤度関数は、 そこでの理論の根幹をなす道具として再び登場する。

参考文献ノート

統計モデルと推定の古典的な教科書として Casella and Berger (2002), Lehmann and Casella (1998), Schervish (1995) がある。 十分統計量と完備統計量の理論は Lehmann and Scheffé (1950, 1955) に遡る。 フィッシャー情報量の概念は Fisher (1925) に起源を持つ。

MLEの漸近理論については Wald (1949) の先駆的業績があり、 現代的な扱いは van der Vaart (1998) が標準的である。 M推定量の一般理論は Huber (1964, 1981) によって確立された。 U統計量は Hoeffding (1948) が導入し、 その漸近理論は Serfling (1980) に詳しい。 ランダム化比較試験における共変量調整の実務整理としては FDA (2023) と EMA (2015) が有用であり、 estimand と感度分析の枠組みは ICH E9(R1) (2019) が標準的な参照点である。

第4章

検定・信頼集合・多重比較

問いと学習目標

この章で答える問い

- 仮説検定において「最適」とは何を意味するか？
- 尤度比検定・Wald検定・スコア検定はどう使い分けるか？
- 検定と信頼区間はどのような数学的関係にあるか？
- 数千の仮説を同時に検定するとき、何が問題になり、どう対処するか？

読み終えたらできるようになること

1. Neyman–Pearson補題を用いて単純仮説に対する最強力検定を構成できる
2. 単調尤度比の性質を用いて複合仮説に対するUMP検定を識別できる
3. 尤度比・Wald・スコアの三大検定の特徴を理解し、場面に応じて使い分けられる
4. 検定と信頼集合の双対性を理解し、検定の棄却域を反転して信頼区間を構成できる
5. ピボット量・漸近理論・プロファイル尤度による信頼区間を構成できる
6. BH法によるFDR制御を理解し実装できる
7. 順列検定の原理と適用条件を説明できる

直観的理解

検定と信頼集合は数学的に**表裏一体**の関係にある。 水準αの検定が与えられたとき、その検定で棄却されないパラメータ値の集合が$100(1-\alpha)\%$信頼集合となる。 逆に、信頼集合が与えられれば、「パラメータ値がその集合に含まれるか否か」で検定を構成できる。
本章では、まずNeyman–Pearson理論に基づく検定の最適性から出発する。 単純仮説ではNP補題が最強力検定を与え、 単調尤度比の性質によって片側複合仮説にも拡張される。 しかし、両側検定では一般にUMP検定が存在しない。 この限界を受けて、

大標本で広く使える三大検定手法（尤度比・Wald・スコア）を導入する。 続いて、検定と信頼区間の双対性を明示的に示し、 ピボット量・漸近的方法による信頼区間の構成、 同時信頼区間、多重検定とFDR制御、順列検定までを統一的に扱う。

4.1　Neyman–Pearson理論

4.1.1　仮説検定の基本設定

定義 4.1.1 (仮説検定の枠組み)**.** データ$\mathbf{X} = (X_1, \ldots, X_n)$がパラメータ$\theta \in \Theta$を持つ分布族$\{P_\theta : \theta \in \Theta\}$に従うとする。 パラメータ空間$\Theta$を$\Theta_0$と$\Theta_1 = \Theta \setminus \Theta_0$に分割し、 **帰無仮説**$H_0 : \theta \in \Theta_0$と**対立仮説**$H_1 : \theta \in \Theta_1$を設定する。 **検定**とは、観測値$\mathbf{X}$に基づいて$H_0$を棄却するか否かを決める規則であり、 **棄却域**$R \subset \mathcal{X}$を指定して$\mathbf{X} \in R$なら棄却する。

読み下し

仮説検定とは、「データが帰無仮説と矛盾するか否か」を判定する手続きである。 棄却域Rは「H_0と矛盾するデータの集合」を数学的に定めたものである。

定義 4.1.2 (第一種誤りと第二種誤り)**.**　・**第一種誤り**（偽陽性）：H_0が真であるのにH_0を棄却する

・**第二種誤り**（偽陰性）：H_1が真であるのにH_0を採択する

検定の**有意水準**αとは、第一種誤りの確率の上限であり、

$$\sup_{\theta \in \Theta_0} \mathrm{P}_\theta(\mathbf{X} \in R) \leq \alpha \tag{4.1}$$

を満たす検定を「水準αの検定」と呼ぶ。

読み下し

有意水準αの条件式 $\sup_{\theta\in\Theta_0} \mathrm{P}_\theta(\mathbf{X} \in R) \leq \alpha$は 次のように読める：「帰無仮説が正しいようなあらゆるパラメータ値θを考えたとき、 棄却域にデータが入る確率は、最悪の場合でもα以下でなければならない。」 sup（上限）を取るのは、Θ_0が複合仮説（複数のパラメータ値を含む）の場合に、 どの$\theta \in \Theta_0$に対しても第一種誤りがαを超えないことを保証するためである。

直観的理解

仮説検定には根本的なトレードオフがある。 第一種誤りの確率を下げれば検出力（正しくH_0を棄却する能力）は低下し、 その逆も然りである。 Neyman–Pearsonの枠組みでは、αを固定してその制約の下で検出力を最大化する。 これは「無罪推定」の法的原則に類似する。 有意水準αは「無実の人を誤って有罪にする確率」の上限であり、この制約下で「真犯人を見逃さない確率」を最大にしたい。

4.1.2 検出力関数

定義 4.1.3 (検出力関数)**.** 棄却域Rを持つ検定の**検出力関数**（power function）は、

$$\beta(\theta) = \mathrm{P}_\theta(\mathbf{X} \in R) \tag{4.2}$$

で定義される。 $\theta \in \Theta_0$では$\beta(\theta) \le \alpha$（第一種誤りの制御）、 $\theta \in \Theta_1$では$\beta(\theta)$が大きいほど良い検定である。

読み下し

検出力関数$\beta(\theta)$は、「パラメータの真の値がθであるとき、 検定がH_0を棄却する確率」をθの関数として描いたものである。 理想的には、Θ_0上でα以下に抑えつつ、Θ_1上でできるだけ1に近づけたい。

例 4.1.4 (正規分布の片側検定の検出力)**.** $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$で、 $H_0 : \mu \le 0$, $H_1 : \mu > 0$に対して$\bar{X} > z_\alpha / \sqrt{n}$で棄却する検定を考える。 検出力関数は

$$\beta(\mu) = \mathrm{P}_\mu\left(\bar{X} > \frac{z_\alpha}{\sqrt{n}}\right) = 1 - \Phi\left(z_\alpha - \sqrt{n}\,\mu\right) \tag{4.3}$$

μが大きいほど（効果が強いほど）検出力は高く、 nが大きいほど（標本が多いほど）検出力は高い。 $\mu = 0$（H_0の境界）では$\beta(0) = \alpha$となる。

4.1.3 p値

定義 4.1.5 (p値)**.** 検定統計量$T(\mathbf{X})$に基づく検定の**p値**とは、 帰無仮説の下で、観測されたのと同等以上に極端な値が得られる確率である：

$$p = \sup_{\theta \in \Theta_0} \mathrm{P}_\theta(T(\mathbf{X}) \ge T(\mathbf{x}_{\text{obs}})) \tag{4.4}$$

読み下し

p値は「帰無仮説を仮定したとき、観測データと同等かそれ以上に帰無仮説に反する データが偶然得られる確率」を表す。 p値が小さいほど、データは帰無仮説と矛盾する。 $p \le \alpha$のときH_0を棄却する規則は、ちょうど水準αの検定に対応する。

定理 4.1.6 (p値の分布)**.** *H_0が単純仮説（$\Theta_0 = \{\theta_0\}$）であり、 検定統計量Tが連続分布を持つとき、p値は帰無仮説の下で一様分布に従う：*

$$p \sim \mathrm{Unif}(0,1) \quad \textit{under } H_0 \tag{4.5}$$

Proof. F_0をH_0の下でのTの累積分布関数とし、$p = 1 - F_0(T)$とする。 F_0が連続であるとき、$U = F_0(T) \sim \mathrm{Unif}(0,1)$（確率積分変換）より、 $p = 1 - U \sim \mathrm{Unif}(0,1)$が従う。 □

実務ポイント

p値の正しい解釈には注意が必要である。

- p値は「H_0が正しい確率」ではない。 p値はH_0を仮定した下での条件付き確率であり、H_0の事後確率ではない。
- $p < 0.05$は「効果が大きい」を意味しない。 大標本では極めて小さな効果量でも統計的に有意になりうる。
- $p > 0.05$は「効果がない」を意味しない。 検出力が低い場合、実質的に重要な効果を見逃す可能性がある。
- しきい値通過だけで結論を固定しない。 検定統計量、効果量、信頼区間、分析計画を合わせて示す方が解釈は安定する。

効果量と信頼区間を合わせて報告することが、現代の統計的実践では推奨される。

4.1.4　Neyman–Pearson補題

定理 4.1.7 (Neyman–Pearson補題)**.** *単純仮説$H_0 : \theta = \theta_0$と$H_1 : \theta = \theta_1$に対する検定において、 尤度比検定*

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \dfrac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} > k \\ 0 & \text{if } \dfrac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} < k \end{cases} \tag{4.6}$$

（ここで$k \geq 0$は$\mathbb{E}_{\theta_0}[\phi(\mathbf{X})] = \alpha$を満たすように選ぶ）は、 同じ有意水準$\alpha$を持つすべての検定の中で最も大きい検出力を有する。 すなわち、$\mathbb{E}_{\theta_0}[\phi'(\mathbf{X})] \leq \alpha$を満たす任意の検定$\phi'$に対して、

$$\mathbb{E}_{\theta_1}[\phi(\mathbf{X})] \geq \mathbb{E}_{\theta_1}[\phi'(\mathbf{X})] \tag{4.7}$$

が成立する。

読み下し

NP補題は「二つの単純仮説の間で最も効率よく判別する検定は、 尤度比——すなわちH_1の下でデータが得られる尤もらしさとH_0の下での尤もらしさの比——が 閾値を超えたら棄却するものだ」と述べている。 直観的には、尤度比が大きいデータほどH_1を支持する証拠が強いので、 それらを優先的に棄却域に入れるのが最も効率的である。

証明. **方針**： NP検定ϕの棄却域Rと別の検定ϕ'の棄却域R'を比較する。 Rに含まれてR'に含まれない部分では尤度比が大きく、 R'に含まれてRに含まれない部分では尤度比が小さいことを利用する。

$R = \{\mathbf{x} : L(\theta_1; \mathbf{x}) > k\, L(\theta_0; \mathbf{x})\}$とし、 ϕ'を$\mathbb{E}_{\theta_0}[\phi'(\mathbf{X})] \leq \alpha$を満たす任意の検定とする。 検出力の差は

$$\mathbb{E}_{\theta_1}[\phi - \phi'] = \int (\phi(\mathbf{x}) - \phi'(\mathbf{x}))\, f(\mathbf{x}; \theta_1)\, d\mathbf{x} \tag{4.8}$$

R上では$f(\mathbf{x};\theta_1) \geq k\, f(\mathbf{x};\theta_0)$であり、 R^c上では$f(\mathbf{x};\theta_1) \leq k\, f(\mathbf{x};\theta_0)$である。 よって、$\phi - \phi' \geq 0$の領域では$f(\mathbf{x};\theta_1) \geq k f(\mathbf{x};\theta_0)$、 $\phi - \phi' \leq 0$の領域では$f(\mathbf{x};\theta_1) \leq k f(\mathbf{x};\theta_0)$であるから、

$$\mathbb{E}_{\theta_1}[\phi - \phi'] \geq k \int (\phi(\mathbf{x}) - \phi'(\mathbf{x}))\, f(\mathbf{x};\theta_0)\, d\mathbf{x} \tag{4.9}$$

$$= k\,(\mathbb{E}_{\theta_0}[\phi] - \mathbb{E}_{\theta_0}[\phi']) \tag{4.10}$$

$$= k\,(\alpha - \mathbb{E}_{\theta_0}[\phi']) \geq 0 \tag{4.11}$$

最後の不等式は$\mathbb{E}_{\theta_0}[\phi'] \leq \alpha = \mathbb{E}_{\theta_0}[\phi]$による。 □

例 4.1.8. $X_1, \ldots, X_n$i.i.d.$\mathcal{N}(\mu, 1)$で$H_0 : \mu = 0$と$H_1 : \mu = 1$の場合を考える。 尤度比は

$$\frac{L(1;\mathbf{X})}{L(0;\mathbf{X})} = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(X_i-1)^2/2}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-X_i^2/2}} = \exp\left(n\bar{X} - \frac{n}{2}\right) \tag{4.12}$$

これは$\bar{X}$の単調増加関数であるから、 最適検定は$\bar{X} > c$の形となる（cは$\mathrm{P}_0(\bar{X} > c) = \alpha$で決まる）。 すなわち$c = z_\alpha/\sqrt{n}$であり、片側$z$検定に対応する。

4.1.5 UMP検定と単調尤度比

NP補題は二つの単純仮説に対する最適検定を与えた。 では、複合仮説——たとえば$H_0 : \mu \leq 0$, $H_1 : \mu > 0$——に対して最適な検定は存在するだろうか？

定義 4.1.9 (一様最強力検定)**.** 検定ϕが$H_0 : \theta \in \Theta_0$に対する水準$\alpha$の **一様最強力**（uniformly most powerful, UMP）検定であるとは、 $\mathbb{E}_\theta[\phi'(\mathbf{X})] \leq \alpha$（$\forall \theta \in \Theta_0$）を満たす 任意の検定$\phi'$に対して、

$$\mathbb{E}_\theta[\phi(\mathbf{X})] \geq \mathbb{E}_\theta[\phi'(\mathbf{X})], \quad \forall \theta \in \Theta_1 \tag{4.13}$$

が成立することをいう。

読み下し

UMP検定は、対立仮説空間Θ_1のすべての点で 他のどの水準α検定よりも検出力が高い。 これは非常に強い要求であり、一般には存在しない。 しかし、片側検定で分布族が「単調尤度比」の性質を持つ場合には存在する。

定義 4.1.10 (単調尤度比)**.** 分布族$\{f_\theta : \theta \in \Theta \subset \mathbb{R}\}$が 統計量$T(\mathbf{x})$について**単調尤度比**（monotone likelihood ratio, MLR）を持つとは、 $\theta_1 > \theta_0$なる任意のθ_0, θ_1に対して、 尤度比$f_{\theta_1}(\mathbf{x})/f_{\theta_0}(\mathbf{x})$が $T(\mathbf{x})$の単調非減少関数であることをいう。

直観的理解

MLR性は「Tが大きいデータほど大きいθを支持する」という 自然な性質を数学的に定式化したものである。 指数型分布族は自然パラメータに関してMLRを持つため、 正規分布、指数分布、ポアソン分布などの重要な分布族に適用できる。

定理 4.1.11 (Karlin–Rubin の定理)**.** *分布族*$\{f_\theta\}$*が*$T(\mathbf{X})$*についてMLRを持つとする。* $H_0 : \theta \le \theta_0, H_1 : \theta > \theta_0$*を検定するとき、*

$$\phi(\mathbf{x}) = \begin{cases} 1 & if\, T(\mathbf{x}) > c \\ 0 & if\, T(\mathbf{x}) < c \end{cases} \tag{4.14}$$

（c*は*$\mathbb{E}_{\theta_0}[\phi(\mathbf{X})] = \alpha$*で定める） はUMP水準*$\alpha$*検定である。*

読み下し

Karlin–Rubinの定理は次のように読める：「分布族がMLRを持つなら、片側仮説$H_0 : \theta \le \theta_0$に対して、 十分統計量$T$が閾値$c$を超えたら棄却する検定がUMP（一様最強力）である。」 重要な点は、閾値cが対立仮説の具体的な値θ_1に依存しないことである。 cは$\mathbb{E}_{\theta_0}[\phi] = \alpha$の条件のみで決まるため、 Θ_1内のどのθ_1に対しても同一の検定がNP最適となる。

証明の方針. 任意の$\theta_1 > \theta_0$を固定すると、 $H_0' : \theta = \theta_0, H_1' : \theta = \theta_1$に対するNP検定は、 MLR性より$T(\mathbf{x}) > c$の形となる。 しかもこの$c$は$\theta_1$の選び方に依存しない （$\mathbb{E}_{\theta_0}[\phi] = \alpha$のみで決まる）。 よって同一の検定が$\Theta_1$内のすべての$\theta_1$に対して NP最適であり、UMP性が成立する。 さらに$\theta < \theta_0$では検出力関数が単調であるため $\beta(\theta) \le \beta(\theta_0) = \alpha$が保証される。 □

例 4.1.12 (指数分布のUMP検定)**.** $X_1, \dots, X_n$i.i.d.Exp(λ)（レートλ、平均$1/\lambda$）で $H_0 : \lambda \ge \lambda_0, H_1 : \lambda < \lambda_0$を考える。

λが小さい（レートが低い）ほど各観測値は大きくなる傾向があるため、 十分統計量$T = \sum_{i=1}^n X_i$はλの減少に対して増大する。 指数分布族は自然パラメータ$\eta = -\lambda$について Tに関するMLRを持つ。 H_1は$\lambda < \lambda_0$、すなわち$\eta > -\lambda_0$に対応するから、 Karlin–Rubinの定理より$T > c$で棄却する検定がUMPとなる。

棄却限界cは、H_0の境界$\lambda = \lambda_0$において $2\lambda_0 T \sim \chi^2_{2n}$が成り立つことを利用し、 $\mathrm{P}_{\lambda_0}(T > c) = \alpha$から定める。

実務ポイント

両側検定$H_0 : \theta = \theta_0, H_1 : \theta \ne \theta_0$に対しては、 一般にUMP検定は**存在しない**。 直観的には、$\theta < \theta_0$の検出と$\theta > \theta_0$の検出を 同時に最適化することは不可能だからである。 この制約が、次節で導入する尤度比検定・Wald検定・スコア検定といった 漸近的に最適な検定手法の動機となる。

4.2　尤度比検定、Wald検定、スコア検定

複合仮説や両側検定では一般にUMP検定が存在しないため、 大標本で良い性質を持つ検定手法が求められる。 本節では、三つの代表的な漸近検定を導入する。 いずれも正則条件下で漸近的に同等であるが、 有限標本での振る舞いと計算上の特徴は異なる。

4.2.1　一般化尤度比検定

定義 4.2.1 (一般化尤度比検定統計量)**.** 複合仮説$H_0 : \theta \in \Theta_0$と$H_1 : \theta \in \Theta_1$に対して、 一般化尤度比検定統計量は

$$\Lambda = \frac{\max_{\theta \in \Theta_0} L(\theta; \mathbf{X})}{\max_{\theta \in \Theta} L(\theta; \mathbf{X})} \tag{4.15}$$

で定義される。$0 \leq \Lambda \leq 1$であり、 Λが小さいほどH_0に対する証拠が強い。

読み下し

一般化尤度比Λは、「帰無仮説のもとで最も尤もらしいパラメータ値での尤度」と「全パラメータ空間で最も尤もらしい値での尤度」の比である。 この比が小さいとは、帰無仮説に制限すると データのフィットが大幅に悪化することを意味し、H_0の棄却を示唆する証拠が強い。

直観的理解

尤度比検定は、NP補題の自然な拡張と見なせる。 NP補題では単純仮説の尤度の比を取ったが、 複合仮説ではそれぞれの仮説の下で「最も有利な」パラメータ値を選んで比を取る。 この検定の強力さは、パラメータ空間の幾何学的構造を活用していることに由来する。

例 4.2.2. $X_1, \ldots, X_n$i.i.d.$\mathcal{N}(\mu, \sigma^2)$で、 $H_0 : \mu = \mu_0$（σ^2は未知）に対する検定を考える。Θ_0の下でのMLE：$\hat{\mu} = \mu_0$、 $\hat{\sigma}_0^2 = \frac{1}{n}\sum_i (X_i - \mu_0)^2$。 Θ全体でのMLE：$\hat{\mu} = \bar{X}$、 $\hat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$。 尤度比は

$$\Lambda = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right)^{n/2} = \left(\frac{1}{1 + t^2/(n-1)} \right)^{n/2} \tag{4.16}$$

ここで$t = \sqrt{n}(\bar{X} - \mu_0)/S$は$t$統計量である。 Λが小さい$\Leftrightarrow$ $|t|$が大きいので、 尤度比検定は通常のt検定と同等である。

4.2.2　Wald検定

定義 4.2.3 (Wald検定統計量)**.** $\hat{\theta}_n$をMLEとするとき、Wald検定統計量は

$$W = n(\hat{\theta}_n - \theta_0)^\top \mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \tag{4.17}$$

で定義される。帰無仮説$H_0 : \theta = \theta_0$の下で、$W \xrightarrow{d} \chi^2_p$である。 したがって、水準$\alpha$の漸近棄却域は$W > \chi^2_{p,\alpha}$で与えられる。 ここで$\chi^2_{p,\alpha}$は$\chi^2_p$分布の上側$\alpha$分位点である。

読み下し

Wald統計量は「MLEが帰無仮説の値θ_0からどれだけ離れているか」を、 フィッシャー情報量で標準化して測ったものである。 フィッシャー情報量はパラメータ推定の「精度」を表すから、 Wは「推定精度で測った$\hat{\theta}$とθ_0の距離の二乗」にnをかけたものと読める。 1次元の場合、$W = (\hat{\theta}_n - \theta_0)^2 / \widehat{\mathrm{Var}}(\hat{\theta}_n)$ となり、馴染みのあるz検定の二乗に帰着する。

直観的理解

Wald検定は計算が簡単で、MLEとその標準誤差さえあれば直ちに実行できる。多くの統計ソフトウェアが回帰係数の検定にWald検定を既定で使用する。しかし、制約なしMLE$\hat{\theta}_n$のみを必要とするため、制約付き最適化を避けられるという利点がある一方、パラメータが境界付近にある場合には性能が劣化する。

4.2.3　スコア検定（Rao検定）

定義 4.2.4 (スコア検定統計量)**.** スコア検定統計量（Rao統計量）は

$$S = \frac{1}{n}\left(\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\theta_0}\right)^{\top} \mathcal{I}(\theta_0)^{-1} \left(\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\theta_0}\right) \tag{4.18}$$

で定義される。ここで$\ell(\theta) = \log L(\theta; \mathbf{X})$は対数尤度、$\mathcal{I}(\theta_0)$は1標本あたりのフィッシャー情報量である。帰無仮説の下で$S \xrightarrow{d} \chi^2_p$。

読み下し

スコア検定は「帰無仮説の値θ_0における対数尤度の傾き（スコア関数）が どれだけゼロから離れているか」を測る。もしθ_0が真の値であれば、MLEの近くにいるのでスコアは小さいはずである。スコアが大きいことは、θ_0が真の値から遠いことを示唆する。
スコア検定の利点は、**制約なしのMLEを計算する必要がない**ことである。帰無仮説のパラメータ値θ_0でスコアとフィッシャー情報量を計算するだけでよい。これは、制約なしの最適化が困難な場合（複雑な非線形モデル等）に有利である。

4.2.4　三つの検定の関係

定理 4.2.5 (Wilks の定理と三つの検定の漸近等価性)**.** *正則条件の下で、帰無仮説$H_0 : \theta = \theta_0$が真であるとき、大標本では次が成立する：*

1. *尤度比統計量：$-2\log \Lambda \xrightarrow{d} \chi^2_p$*
2. *Wald統計量：$W \xrightarrow{d} \chi^2_p$*
3. *スコア統計量：$S \xrightarrow{d} \chi^2_p$*

さらに、三つの統計量の差は$o_p(1)$であり、漸近的に同じ検出力を持つ。

証明の方針. 対数尤度$\ell(\theta)$をθ_0の周りでTaylor展開する。MLE$\hat{\theta}_n$ではスコアがゼロ：$0 = \frac{\partial \ell}{\partial \theta}|_{\hat{\theta}_n}$。1次近似すると、

$$0 \approx \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_0} + \left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta_0} (\hat{\theta}_n - \theta_0) \tag{4.19}$$

表 4.1: 三大検定の使い分け

検定	必要な計算	強み	注意点
尤度比検定	制約付き・制約なしの両方の MLE	変換に不変で、境界付近でも比較的安定	最適化を2回解く必要があり、計算コストが高い
Wald検定	制約なし MLE と標準誤差	実装が最も簡単で、回帰係数の個別検定に向く	再パラメータ化に弱く、境界付近や小標本で不安定になりやすい
スコア検定	帰無仮説の値 θ_0 でのスコアと情報量	制約なし最適化が不要で、変数追加や複雑モデルで有利	帰無仮説から離れた領域での挙動は尤度比検定ほど直感的でない

大数の法則より $\frac{1}{n}\frac{\partial^2 \ell}{\partial \theta^2}\big|_{\theta_0} \xrightarrow{p} -\mathcal{I}(\theta_0)$であるから、

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \mathcal{I}(\theta_0)^{-1} \cdot \frac{1}{\sqrt{n}} \frac{\partial \ell}{\partial \theta}\bigg|_{\theta_0} \tag{4.20}$$

この関係を用いると、Wald統計量とスコア統計量が漸近的に一致することが示される。 さらに対数尤度を2次まで展開すれば、 $-2\log\Lambda \approx W \approx S$が得られる。 □

重要結果

三つの検定の実践的比較： 漸近的に等価であるが、有限標本では計算量・変換不変性・境界付近での安定性が異なる。 使い分けは表4.1に整理した。 Wald検定は計算が最も簡単であるが、パラメータが境界付近にある場合 （例：確率$p \approx 0$）にカバレッジが劣化する。 尤度比検定はパラメータ変換に対して不変であり、多くの場合に最も信頼性が高い。

例 4.2.6. ロジスティック回帰で$H_0 : \beta_j = 0$をテストする場合、 三つの検定統計量を比較できる。 大標本では同じ結論に至るが、小標本ではWald検定が過度に自由（liberal）に、 尤度比検定が最も正確な有意水準を維持する傾向がある。

コード例：R による三大検定の比較

```
# R: 三つの検定統計量の比較
data(iris)
d <- subset(iris, Species != "versicolor")
d$y <- as.integer(d$Species == "virginica")
fit <- glm(y ~ Sepal.Length, family = binomial, data = d)
fit0 <- glm(y ~ 1, family = binomial, data = d)

# Wald検定（summary の z 統計量の二乗）
cs <- summary(fit)$coefficients
wald <- (cs[2, 1] / cs[2, 2])^2

# 尤度比検定
lr <- as.numeric(2 * (logLik(fit) - logLik(fit0)))
```

```
# スコア検定（anova の Chisq 統計量）
sc <- anova(fit0, fit, test = "Rao")[2, "Rao"]

cat(sprintf("Wald = %.2f, LR = %.2f, Score = %.2f\n",
            wald, lr, sc))
```

4.3　検定と信頼集合の双対性

ここまでは「検定」に焦点を当ててきたが、実は検定と信頼集合は同一のコインの裏表である。本節では、この双対性を定理として明示的に示し、前節の三大検定手法から対応する信頼区間が自然に得られることを見る。

4.3.1　双対性定理

定理 4.3.1 (検定と信頼集合の双対性)**.** *各$\theta_0 \in \Theta$に対して、帰無仮説$H_0 : \theta = \theta_0$の水準α検定ϕ_{θ_0}（棄却域R_{θ_0}）が与えられたとする。このとき、*

$$C(\mathbf{X}) = \{\theta_0 \in \Theta : \mathbf{X} \notin R_{\theta_0}\} \tag{4.21}$$

はθの$100(1-\alpha)\%$信頼集合である。すなわち、

$$\mathrm{P}_{\theta_0}(\theta_0 \in C(\mathbf{X})) \geq 1 - \alpha, \quad \forall \theta_0 \in \Theta \tag{4.22}$$

逆に、$C(\mathbf{X})$が$100(1-\alpha)\%$信頼集合であるとき、$R_{\theta_0} = \{\mathbf{X} : \theta_0 \notin C(\mathbf{X})\}$は$H_0 : \theta = \theta_0$の水準$\alpha$検定の棄却域を定義する。

読み下し

双対性定理は次のように読める：**信頼集合とは、検定で棄却されないパラメータ値の集合である。**すなわち、信頼区間$C(\mathbf{X})$に含まれるθ_0とは、「もし$\theta = \theta_0$を帰無仮説として検定したら棄却されなかった」値の全体に他ならない。この双対性により、検定の最適性（例：UMP性）は信頼集合の最適性（最も短い信頼区間）と直結する。

Proof. ϕ_{θ_0}が水準αの検定であるとは、$\mathrm{P}_{\theta_0}(\mathbf{X} \in R_{\theta_0}) \leq \alpha$を意味する。$C(\mathbf{X})$の定義より、

$$\theta_0 \in C(\mathbf{X}) \iff \mathbf{X} \notin R_{\theta_0} \tag{4.23}$$

したがって、

$$\mathrm{P}_{\theta_0}(\theta_0 \in C(\mathbf{X})) = \mathrm{P}_{\theta_0}(\mathbf{X} \notin R_{\theta_0}) = 1 - \mathrm{P}_{\theta_0}(\mathbf{X} \in R_{\theta_0}) \geq 1 - \alpha \tag{4.24}$$

逆方向も同様の議論で成立する。 □

4.3.2　三大検定に対応する信頼区間

定理 4.3.1の双対性を用いれば、 第 4.2節の三大検定手法から、 それぞれ対応する信頼区間が自然に構成される。

定義 4.3.2 (尤度比信頼区間)**.** 尤度比検定統計量（定義 4.2.1）から得られる信頼集合は、

$$C_{\mathrm{LR}}(\mathbf{X}) = \left\{ \theta_0 : -2\log \frac{L(\theta_0; \mathbf{X})}{\max_\theta L(\theta; \mathbf{X})} \le \chi^2_{p,\alpha} \right\} \tag{4.25}$$

すなわち、尤度比検定で棄却されないθ_0の全体である。

定義 4.3.3 (Wald信頼区間)**.** Wald検定統計量（定義 4.2.3）から得られる信頼集合は、

$$C_{\mathrm{W}}(\mathbf{X}) = \left\{ \theta_0 : n(\hat\theta_n - \theta_0)^\top \mathcal{I}(\hat\theta_n)(\hat\theta_n - \theta_0) \le \chi^2_{p,\alpha} \right\} \tag{4.26}$$

1次元パラメータの場合、楕円体が区間に退化し、

$$\hat\theta_n \pm z_{\alpha/2} \frac{1}{\sqrt{n\mathcal{I}(\hat\theta_n)}} \tag{4.27}$$

となる。これは実務で最も頻繁に使われる信頼区間の形式である。

定義 4.3.4 (スコア信頼区間)**.** スコア検定統計量（定義 4.2.4）から得られる信頼集合は、

$$C_{\mathrm{S}}(\mathbf{X}) = \left\{ \theta_0 : \frac{1}{n} \left(\frac{\partial \ell}{\partial \theta}\Big|_{\theta=\theta_0} \right)^\top \mathcal{I}(\theta_0)^{-1} \left(\frac{\partial \ell}{\partial \theta}\Big|_{\theta=\theta_0} \right) \le \chi^2_{p,\alpha} \right\} \tag{4.28}$$

重要結果

三つの信頼区間の比較： 定理 4.2.5により、 大標本では三つの信頼区間は漸近的に一致する。 しかし、有限標本では以下の特徴がある：

- **尤度比区間**：パラメータ変換に対して不変であり、 多くの場合に最良のカバレッジを持つ。
- **Wald区間**：計算が最も簡単だが、 パラメータが境界付近にある場合にカバレッジが劣化する。
- **スコア区間**：帰無仮説の値θ_0で評価するため、 Wald区間よりカバレッジが安定する。

例 4.3.5 (正規分布における双対性)**.** $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$（$\sigma^2$既知）において、 $H_0 : \mu = \mu_0$の水準α検定は

$$\text{reject } H_0 \text{ if } \left| \frac{\bar X - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2} \tag{4.29}$$

この検定の棄却域を反転すると、棄却されないμ_0の集合は

$$C(\mathbf{X}) = \left\{ \mu_0 : \left| \frac{\bar X - \mu_0}{\sigma/\sqrt{n}} \right| \le z_{\alpha/2} \right\} = \left[\bar X - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar X + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right] \tag{4.30}$$

これはまさに平均μの$100(1-\alpha)\%$信頼区間であり、 検定と信頼区間の双対性を具体的に示している。

4.4　ピボット量に基づく信頼区間

前節では検定から信頼区間を構成する一般論を見た。 本節では、ピボット量という特殊な構造を持つ統計量を用いて、 有限標本で正確な信頼区間を直接構成する方法を紹介する。

4.4.1　ピボット量の概念と構成

定義 4.4.1 (ピボット量)**.** θを未知母数、$\mathbf{X} = (X_1, \ldots, X_n)$を観測データとする。 統計量$Q(\mathbf{X}, \theta)$が**ピボット量**（pivot quantity）であるとは、 その分布がθに依存しないことをいう。

読み下し

ピボット量とは、「データとパラメータの両方を含むが、 どのパラメータ値のもとでも同じ確率分布に従う統計量」である。 例えば$Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$は、 μとσの値によらず標準正規分布に従うのでピボット量である。 ピボット量の分布は既知であるため、分位点を用いて直接信頼区間を構成できる。

定理 4.4.2 (ピボット量からの信頼区間構成)**.** *$Q(\mathbf{X}, \theta)$がピボット量で、 $\mathrm{P}(a < Q(\mathbf{X}, \theta) < b) = 1 - \alpha$が成立するとき、 $Q(\mathbf{X}, \theta) \in (a, b)$を$\theta$について逆解すると、 θの$100(1-\alpha)\%$信頼区間$I(\mathbf{X})$が得られる：*

$$\mathrm{P}(\theta \in I(\mathbf{X})) = 1 - \alpha$$

Proof. $Q(\mathbf{X}, \theta)$がピボット量なのでその分布はθに依存しない。 したがって$\mathrm{P}(a < Q(\mathbf{X}, \theta) < b) = 1 - \alpha$は 全ての$\theta$について成立する。 $Q(\mathbf{X}, \theta) \in (a, b) \Leftrightarrow \theta \in I(\mathbf{X})$ となるように逆解すれば、 $\mathrm{P}(\theta \in I(\mathbf{X})) = 1 - \alpha$が得られる。 □

読み下し

ピボット量による信頼区間の構成は、 第 4.3節の双対性の特殊な場合と見なすこともできる。 ピボット量$Q(\mathbf{X}, \theta)$から、 $H_0 : \theta = \theta_0$に対して$|Q(\mathbf{X}, \theta_0)| > c$で棄却する検定を構成し、 棄却されない$\theta_0$の集合として信頼区間が得られる。 ピボット量が利用できる場合、漸近的手法と異なり **有限標本で正確な**カバレッジが保証される点が大きな利点である。

4.4.2　正規母集団における信頼区間

例 4.4.3 (平均値の信頼区間（分散既知）)**.** $X_1, \ldots, X_n$i.i.d.$\mathcal{N}(\mu, \sigma^2)$で、$\sigma^2$は既知とする。 ピボット量は

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

標準正規分布の上側$\alpha/2$分位点を$z_{\alpha/2}$とすると、

$$\mathrm{P}\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

μについて逆解すれば、

$$I(\mathbf{X}) = \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

例 4.4.4 (平均値の信頼区間（分散未知）)**.** 分散が未知の場合、σを標本標準偏差Sに置き換えるとピボット量はt分布に従う：

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

ここで$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$は不偏分散。 t_{n-1}の上側$\alpha/2$分位点を$t_{n-1,\alpha/2}$とすると、

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

例 4.4.5 (分散の信頼区間)**.** $X_1,\dots,X_n$ i.i.d. $\mathcal{N}(\mu,\sigma^2)$ （μは未知）において、ピボット量

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

を用いると、σ^2の$100(1-\alpha)\%$信頼区間は

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}},\ \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$

実務ポイント

分散未知の場合にt分布を使用することは、小標本で重要である。 nが大きい場合、$t_{n-1,\alpha/2} \approx z_{\alpha/2}$となり、 正規分布に基づく区間と近似する。 実務では$n \geq 30$程度で近似が十分正確であることが多い。

4.5　漸近的信頼区間とプロファイル尤度

ピボット量は有限標本で正確な信頼区間を与えるが、 利用できるのは正規分布などの特定の分布族に限られる。 本節では、漸近理論に基づく信頼区間の構成法を扱う。 第 4.3節で導入した三つの信頼区間（定義 4.3.3–4.3.4） の漸近的な性質を掘り下げ、プロファイル尤度による方法も紹介する。

4.5.1　Wald信頼区間の漸近的性質

定義 4.3.3のWald信頼区間は漸近的に正確である：

定理 4.5.1 (Wald信頼区間のカバレッジ)**.** *MLEの漸近正規性* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$ *が成立するとき、Wald信頼区間のカバレッジは*

$$\mathrm{P}(\theta_0 \in C_{\mathrm{W}}(\mathbf{X})) \to 1 - \alpha \quad (n \to \infty)$$

実務ポイント

Wald区間は計算が簡単で、推定量と標準誤差のみを必要とする。 しかし、**パラメータが境界に近い場合**（例：確率パラメータが0や1に近い場合）には、 カバレッジが名目水準$1-\alpha$を大きく下回ることがある。 有限標本では、スコア区間やWilson区間の方が優れることが多い。

4.5.2　スコア信頼区間の漸近的性質

実務ポイント

スコア信頼区間（定義 4.3.4）は、 Wald区間と同じ漸近的カバレッジ$1-\alpha$を持つ。 有限標本では、ベルヌーイ比率に対するWilson区間のように Wald区間より良いカバレッジを示す例が多いが、 これは一般理論として常に成り立つわけではない。 とくに、パラメータが空間の境界に近い場合や 分布の歪みが強い場合には、スコア区間の方が安定することが多い。

例 4.5.2 (ベルヌーイ分布：Wald区間 vs スコア区間)**.** $X_1, \ldots, X_n$i.i.d.Bern(p)に対して、Wald信頼区間は$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$であり、 スコア信頼区間は

$$\left| \frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}} \right| \leq z_{\alpha/2}$$

をpについて逆解して得られる（Wilson区間とも呼ばれる）。

$p = 0.05, n = 50$の場合にシミュレーションすると、 Wald区間のカバレッジは約0.88（名目0.95を大幅に下回る）であるのに対し、 スコア区間のカバレッジは約0.94で名目値に近い。 これは$\hat{p}$の分布が$p \approx 0$で強い非対称性を持つことによる。

4.5.3　尤度比信頼区間

定理 4.5.3 (Wilksの定理)**.** *正則条件下で、真のパラメータを*θ_0*とするとき、*

$$-2\log \frac{L(\theta_0; \mathbf{X})}{L(\hat{\theta}_n; \mathbf{X})} \xrightarrow{d} \chi^2_p \quad (n \to \infty) \tag{4.31}$$

ここで$p = \dim(\theta)$*。 より一般に、*r*個の制約を持つ複合帰無仮説では*χ^2_r*に従い、 水準*α*の棄却限界値は*$\chi^2_{r,\alpha}$*となる。*

読み下し

Wilksの定理は、尤度比検定統計量の漸近分布を与える。 これにより、尤度比信頼区間（定義 4.3.2）は $\{-2\log\Lambda(\theta)\le\chi^2_{p,\alpha}\}$として構成できる。 尤度比区間の利点は、パラメータの再パラメータ化に対して不変であることである。

4.5.4　プロファイル尤度

定義 4.5.4 (プロファイル尤度)**.** パラメータ$\boldsymbol{\theta}=(\theta_1,\boldsymbol{\theta}_{-1})$の分割に対して、 プロファイル尤度は

$$L_p(\theta_1)=\max_{\boldsymbol{\theta}_{-1}} L(\theta_1,\boldsymbol{\theta}_{-1};\mathbf{X})$$

対応する対数プロファイル尤度は$\ell_p(\theta_1)=\log L_p(\theta_1)$である。

直観的理解

プロファイル尤度は、関心のあるパラメータθ_1に焦点を当て、 妨害パラメータ$\boldsymbol{\theta}_{-1}$を「最も有利に」選ぶことで除去する方法である。 各θ_1の値に対して、残りのパラメータを最適化してから尤度を評価する。 これは射影のアナロジーで理解できる： 高次元パラメータ空間をθ_1軸に「射影」して、 θ_1のみの関数に落とし込むのである。

定理 4.5.5 (プロファイル尤度に基づく信頼区間)**.** *相対プロファイル尤度を*

$$\hat\theta_1=\arg\max_{\vartheta_1} L_p(\vartheta_1),\qquad R(\theta_1)=\frac{L_p(\theta_1)}{L_p(\hat\theta_1)}$$

と定義すると、Wilksの定理より$100(1-\alpha)\%$*信頼区間は*

$$I(\mathbf{X})=\{\theta_1:-2\log R(\theta_1)\le\chi^2_{1,\alpha}\}$$

例 4.5.6 (単回帰のプロファイル尤度)**.** 線形回帰モデル $Y_i=\beta_0+\beta_1X_i+\varepsilon_i$を考え、 誤差は$\varepsilon_i$i.i.d.$\mathcal{N}(0,\sigma^2)$とする。 関心が$\beta_1$にあるとき、 β_0とσ^2を最適化して除去した プロファイル尤度から得られる信頼区間は、 通常の線形正規モデルでも、 σ^2を未知としてプロファイルした場合には 尤度比信頼区間に対応し、Wald区間と有限標本で一般には一致しない。 大標本では両者は漸近的に一致するが、 非線形モデルや境界に近い問題では、 プロファイル尤度区間の利点がより明瞭に現れる。

重要結果

信頼区間の構成法の選択指針：

1. **ピボット量が利用可能な場合**（正規分布の平均・分散など）： 有限標本で正確な区間が得られるため、これを優先する。
2. **一般のパラメトリックモデルの場合**： 尤度比区間が変換不変性により最も信頼性が高い。 計算の容易さが求められる場合はWald区間を用いるが、 パラメータ境界付近ではスコア区間に切り替える。

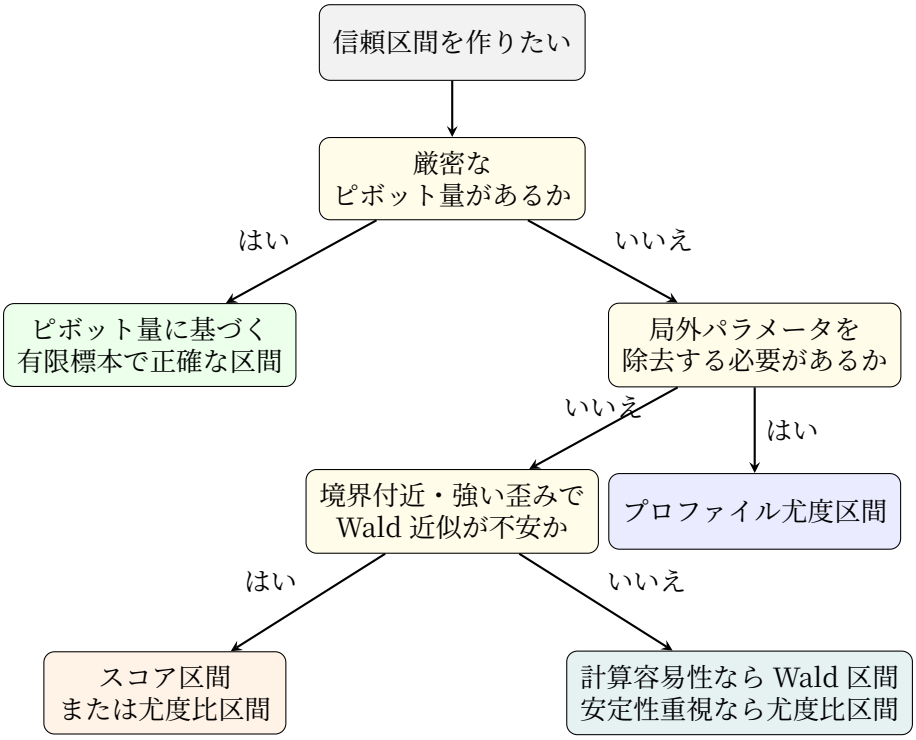


図 4.1: 信頼区間の構成法マップ。まず厳密なピボット量の有無を確認し、無理なら局外パラメータ、境界・歪み、計算コストの順に判断する。

3. **妨害パラメータがある場合**：プロファイル尤度を用いる。

いずれの方法も、定理 4.3.1の双対性を通じて 対応する検定と結びついている。

4.6　同時信頼区間

複数のパラメータについて同時に信頼区間を構成する場合、 各区間を個別に$1-\alpha$で構成しても、 **全体**として$1-\alpha$のカバレッジは保証されない。 同時推定には専用の方法が必要である。

4.6.1　Bonferroni法

定義 4.6.1 (Bonferroni同時信頼区間)**.** m個の母数関数$h_1(\boldsymbol{\theta}), \ldots, h_m(\boldsymbol{\theta})$に対して、 各関数について信頼度$1-\alpha/m$の区間$I_j$を構成する。 すると、全ての関数が同時に対応する値を含む確率は

$$\mathrm{P}(h_j(\boldsymbol{\theta}) \in I_j,\ j = 1, \ldots, m) \geq 1 - \alpha$$

読み下し

Bonferroni法は、所望の同時カバレッジ確率$1-\alpha$を m個の区間に均等に配分する。 各区間の信頼度を$1-\alpha/m$とすることで、 全体的な誤り率をαに抑制する。 これは確率の加法性（$\mathrm{P}(\bigcup A_i) \leq \sum \mathrm{P}(A_i)$）に基づく。

定理 4.6.2 (Bonferroni不等式)**.** *事象$A_1, \dots, A_m$（独立性は不要）に対して、*

$$\mathrm{P}\left(\bigcap_{j=1}^{m} A_j\right) \geq 1 - \sum_{j=1}^{m} \mathrm{P}(A_j^c)$$

各$\mathrm{P}(A_j^c) = \alpha/m$と設定すれば、 $\mathrm{P}\left(\bigcap_j A_j\right) \geq 1 - \alpha$。

読み下し

Bonferroni不等式の式 $\mathrm{P}(\bigcap_j A_j) \geq 1 - \sum_j \mathrm{P}(A_j^c)$は次のように読める： 「$m$個の事象がすべて起こる確率は、 各事象が起こらない確率の合計を1から引いた値以上である。」 これは補事象の和集合に関する $\mathrm{P}(\bigcup_j A_j^c) \leq \sum_j \mathrm{P}(A_j^c)$（Booleの不等式）から直ちに従う。 各$\mathrm{P}(A_j^c) = \alpha/m$と設定すれば合計が$\alpha$となり、 全体の確率が$1-\alpha$以上に制御される。

直観的理解

Bonferroni法はmが小さい場合（$m \leq 5$程度）に効率的だが、 mが大きいと非常に保守的になる。 例えば$m = 100$で$\alpha = 0.05$なら、 各区間の信頼度は$1 - 0.05/100 = 99.95\%$となり、 極めて広い区間になってしまう。

4.6.2　Scheffé法

定義 4.6.3 (Scheffé法による同時信頼区間)**.** k個の平均$\mu_1, \dots, \mu_k$に対し、 条件$\sum c_j = 0$を満たす**全ての**線形対比 $\psi = \sum c_j \mu_j$について、 同時$100(1-\alpha)\%$信頼区間を与える方法である。 対比ψの推定量を$\hat{\psi}$、 その推定分散を$\widehat{\mathrm{Var}}(\hat{\psi})$とすると、 Scheffé法の同時信頼区間は

$$\hat{\psi} \pm \sqrt{(k-1)\, F_{k-1, N-k, \alpha}} \cdot \sqrt{\widehat{\mathrm{Var}}(\hat{\psi})}$$

で与えられる。 対比は共通シフト $\boldsymbol{\mu} \mapsto \boldsymbol{\mu} + a\mathbf{1}$ （$\mathbf{1} = (1, \dots, 1)^\top$）で不変なので、 対応する信頼領域は平均ベクトル全体の有界な楕円体ではなく、 $\sum c_j = 0$を満たす$(k-1)$次元の対比空間で考える。 ここで$F_{k-1, N-k, \alpha}$はF分布の上側α分位点であり、$k-1$は対比空間の自由度である。

実務ポイント

Scheffé法は、**事前に指定されていない**対比に対しても 制御された誤り率を保証する。 データを見た後に関心のある対比を選択する場合に有用である。 一方、少数の事前指定された対比にはBonferroni法の方が検出力が高い。 全ペア比較（$\mu_i - \mu_j$）に特化する場合は、 Tukey HSD法がScheffé法より効率的である。

例 4.6.4 (分散分析における同時推定)**.** 一元配置分散分析で、処理群$i = 1, \dots, k$の平均を$\mu_1, \dots, \mu_k$とする。 各群のサンプルサイズをn_i、プール標準誤差を$\hat{\sigma}$とするとき、 Scheffé

法により、全ての対比$\sum c_j \mu_j$（$\sum c_j = 0$）に対して同時$100(1-\alpha)\%$信頼区間は

$$\sum_{j=1}^{k} c_j \bar{y}_{j\cdot} \pm \sqrt{(k-1) F_{k-1, N-k, \alpha}} \cdot \hat{\sigma} \sqrt{\sum_{j=1}^{k} \frac{c_j^2}{n_j}}$$

ここで$N = \sum n_i$。

4.7　多重検定とFDR制御

4.7.1　多重検定問題

ゲノミクスや神経科学では、数千〜数百万の仮説を同時に検定する。各検定の有意水準を$\alpha = 0.05$に設定すると、帰無仮説が全て真であっても約5%が偽陽性として棄却される。$m = 10{,}000$個の検定なら、期待的に500個が偽陽性となる。この問題に対処するための枠組みを導入する。

定義 4.7.1 (FWER)**.** FWER（familywise error rate、族全体の第一種誤り率）は、**少なくとも一つの**真の帰無仮説を棄却する確率：

$$\mathrm{FWER} = \mathrm{P}\left(\bigcup_{i \in \mathcal{H}_0} \{\text{reject } H_i\} \right) \tag{4.32}$$

ここで$\mathcal{H}_0$は真の帰無仮説のインデックス集合。

定義 4.7.2 (FDR)**.** FDR（false discovery rate、偽発見率）は、棄却された仮説の中で偽の発見の期待割合：

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{\max(R, 1)} \right] \tag{4.33}$$

ここでVは偽の棄却数（偽発見数）、Rは全棄却数である。

読み下し

FWERは「一つでも間違いを犯す確率」を制御するのに対し、FDRは「間違いの*割合*」を制御する。FDRはFWERより緩い基準であるため、同じ誤り率のもとでFDR制御の方が多くの発見を許容する。大規模検定ではFDR制御が標準的である。

直観的理解

身近な比喩で言えば、FWERは「一つでも不良品があれば出荷停止」、FDRは「不良品率を5%以下に抑える」という品質管理に対応する。1万個の製品を検査するとき、前者は極めて保守的になるが、後者は現実的な基準を与える。

4.7.2　FWER制御法

定義 4.7.3 (Bonferroni修正)**.** 各検定の有意水準をα/mに設定する。FWERを確実にα以下に制御するが、mが大きいと非常に保守的になる。

定義 4.7.4 (Holm法)**.** Bonferroni法より検出力が高い改良版（逐次棄却法）。 p値を昇順に$p_{(1)} \le \cdots \le p_{(m)}$と並べ：

1. $k=1$から順に$p_{(k)} \le \alpha/(m-k+1)$を確認
2. 初めて条件を満たさないk^*で停止し、$H_{(1)},\dots,H_{(k^*-1)}$を棄却

Holm法はBonferroni法を常に支配する（同じかより多くの仮説を棄却する）。

4.7.3　Benjamini–Hochberg法

定理 4.7.5 (Benjamini–Hochberg法)**.** *p値を$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$と順序付け、$H_{(1)},H_{(2)},\dots,H_{(m)}$をそれぞれの仮説とする。次のアルゴリズムにより、FDRを有意水準α以下に制御できる：*

1. *最大のkを見つけよ：$p_{(k)} \le \frac{k}{m}\alpha$*
2. *$H_{(1)},\dots,H_{(k)}$を棄却する*
3. *そのようなkが存在しなければ、すべての帰無仮説を採択する*

p値が独立であるか、正の依存性（PRDS条件）を満たすとき、 $\mathrm{FDR} \le \frac{m_0}{m}\alpha \le \alpha$ *（m_0は真の帰無仮説の数）。*

証明の方針. m個の仮説のうちm_0個が真の帰無仮説であるとする。 鍵となるのは、BH閾値$k\alpha/m$は順位kに比例して増加するため、 「本物の発見」が多く混じると閾値が緩くなり、偽発見の割合が自動的に制御されるという点である。

形式的には、真の帰無仮説$i \in \mathcal{H}_0$について、 $p_i \sim \mathrm{Unif}(0,1)$であることを利用して$\mathbb{E}[V/\max(R,1)]$を上から評価する。 独立性の下では、各真の帰無仮説iが棄却される条件付き確率を 直接計算でき、$\sum_{i\in\mathcal{H}_0}(1/m)\cdot\alpha = (m_0/m)\alpha$ が導かれる。 完全な証明はBenjamini and Hochberg (1995)を、正の依存性（PRDS）下での有効性の証明についてはBenjamini and Yekutieli (2001)を参照。 □

例 4.7.6. マイクロアレイ実験で$m = 10{,}000$個の遺伝子を検定し、 p値が得られたとしよう。$\alpha = 0.05$でFDRを制御したい。 BH閾値は$p_{(k)} \le k/(10{,}000)\times 0.05 = k/200{,}000$である。

例えば$p_{(100)} = 0.0004$, $p_{(101)} = 0.0006$の場合、 $100/200{,}000 = 0.0005$であるから$p_{(100)} \le 0.0005$は満たされるが、 $p_{(101)} > 101/200{,}000 = 0.000505$なので$k=100$で停止する。 100個の仮説を棄却する。 ここでBH法が保証するのは

$$\mathbb{E}\left[\frac{V}{\max(R,1)}\right] \le 0.05$$

という反復実験平均での偽発見割合の制御であり、 今回の100個の棄却の中に偽発見が約5個あると 直接解釈してよいわけではない。 Bonferroni修正なら各$p < 0.05/10{,}000 = 5\times10^{-6}$が必要で、 棄却される仮説数は大幅に減少する。

実務ポイント

実務では以下の指針が有用である：

表 4.2: 多重性調整の選択ガイド

場面	主に守りたい量	推奨される方法	典型例
少数の主要仮説を確認したい	少なくとも一つの偽陽性を避ける（FWER）	Holm 法, Bonferroni 修正	主要評価項目が固定された臨床試験, 事前登録された比較
多数の候補から有望な信号を拾いたい	偽発見の割合（FDR）	BH 法	ゲノム解析, 特徴探索, 大規模スクリーニング
複数の係数や対比を区間で一括報告したい	全区間の同時被覆	Bonferroni 同時区間, Scheffé 法	分散分析の対比, 複数係数の一括報告

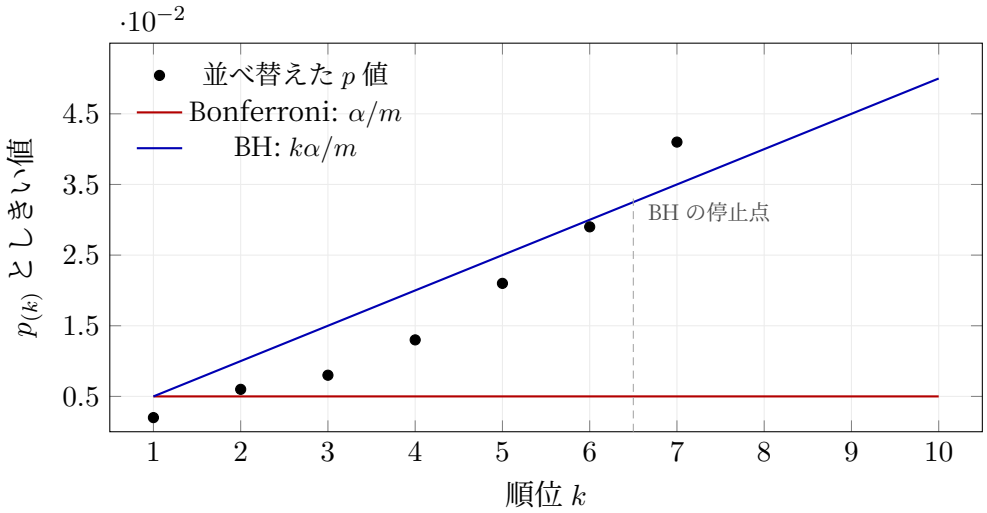


図 4.2: Bonferroni 修正と BH 法のしきい値の違い。ここでは $\alpha = 0.05$, $m = 10$ の模式例で、Bonferroni は1件のみ棄却する一方、BH 法は順位に応じてしきい値を緩めるため6件まで棄却できる。

- 探索的研究（仮説生成）：FDR制御（BH法）が適切
- 確認的研究（事前登録された少数の仮説）：FWER制御（Holm法）が適切
- 「Bonferroni法 $\subset$ Holm法 $\subset$ BH法」 の順に棄却する仮説が増える

確認的解析では、主要評価項目、検定順序、多重性調整法を事前に固定し、 試験全体の第一種過誤を守ることが重要である。

4.8　順列検定

これまでの検定手法は、パラメトリックモデルの仮定や漸近理論に依拠していた。 本節では、分布仮定に依存しないノンパラメトリック検定法である順列検定を導入する。

4.8.1　基本原理

定義 4.8.1 (順列検定)**.** 与えられた検定統計量$T(\mathbf{X})$に対して、帰無仮説の下ではデータの置換がすべて等確率で起こると仮定する。観測値の検定統計量$T(\mathbf{x}_{\text{obs}})$が、置換分布の上位$\alpha$の割合（または絶対値が上位$\alpha$）にある場合、帰無仮説を棄却する。

p値は

$$p = \frac{\#\{\pi : |T(\mathbf{X}_\pi)| \geq |T(\mathbf{x}_{\text{obs}})|\}}{N_{\text{perm}}} \tag{4.34}$$

ここでπはデータの置換、N_{perm}は全置換数。

読み下し

順列検定の原理は次のように読める：「もし帰無仮説が正しければ、データのラベル（例：処理群/対照群の割り当て）は無意味であり、ラベルをランダムに入れ替えても統計量の分布は変わらないはずだ。実際に入れ替えて統計量の分布を作り、観測値がその分布の中で極端な位置にあるかを確認する。」

例 4.8.2. 二標本問題で$\mathbf{X} = (X_1, \ldots, X_{n_1})$と$\mathbf{Y} = (Y_1, \ldots, Y_{n_2})$が与えられたとき、$H_0$: 両群は同一分布に従う。検定統計量$T = \bar{X} - \bar{Y}$として、全データを混合し$n_1$個と$n_2$個にランダム分割する。

$$p = \frac{\#\{\text{置換で } |T_{\text{perm}}| \geq |T_{\text{obs}}|\}}{\binom{n_1+n_2}{n_1}} \tag{4.35}$$

4.8.2　順列検定の有効性

定理 4.8.3 (順列検定の正確性)**.** *帰無仮説の下でデータが**交換可能**（exchangeable）であるとき、順列検定は正確な有意水準αを持つ。すなわち、帰無仮説が真であるとき$\mathrm{P}(p \leq \alpha) \leq \alpha$。*

Proof. 交換可能性の下では、$\mathbf{X}$の全ての置換が同じ確率を持つ。したがって観測されたTは、全置換のもとでのTの値の中で一様にランダムな順位を占める。N_{perm}個の置換のうち、Tが上位αに入る確率は$\leq \alpha$である（離散性による正確なαの保証）。 □

実務ポイント

順列検定の実装上の注意点：

1. **計算量**：nが大きいと$\binom{n}{k}$は爆発的に増大する。通常、$B = 10{,}000$〜$100{,}000$回のモンテカルロ近似を用いる。この場合のp値の標準誤差は約$\sqrt{p(1-p)/B}$。

2. **何を置換するか**：単回帰や単純な二群比較以外（共変量がある場合等）では、単にYを置換すると共変量との相関構造が壊れる。回帰の文脈では、縮小モデル（帰無仮説下のモデル）の残差を置換し、それを縮小モデルの予測値に加えた「疑似的なデータ」で検定を行う**Freedman–Lane法**（Freedman and Lane, 1983）が、共変量を適切に調整しつつ検定を行う標準的な手法である。

3. **パラメトリック検定との関係**：分布仮定が正しい場合、パラメトリック検定の方が検出力が高い。順列検定は仮定のロバスト性が求められる場面で選択する。

指標	主に答える問い	単独では足りない点
p値	帰無仮説の下で、観測値以上に極端な結果がどれほど出にくいか。	効果の大きさや実務的重要性はわからない。
信頼区間	どの程度の効果量がデータと整合的で、どれほど精度よく推定できたか。	事前に定めた検定計画や多重性の問題までは表現しない。
効果量	差がどの程度大きいか、実務的にどれほど意味がありそうか。	不確実性や標本変動の大きさは単独ではわからない。

表 4.3: p値、信頼区間、効果量の役割の違い。確認的解析では、この三つを併せて読むと解釈が安定する。

4.9　実践：検出力シミュレーション、カバレッジ検証、FDR制御

4.9.1　検出力の計算と標本サイズ設計

定義 4.9.1 (効果量)**.** 効果量は、検定される仮説の実務的な重要性を測定する無次元量である：

- Cohen の d（Cohen, 1988）：二つの平均の差をプール標準偏差で正規化した量。$d = 0.2$（小）、0.5（中）、0.8（大）が目安。
- 相関係数r：-1から1の間の値
- オッズ比OR：2値アウトカムの効果指標

例 4.9.2 (標本サイズの設計公式)**.** $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$で $H_0 : \mu = 0$, $H_1 : \mu = \delta > 0$（片側検定）の検出力は

$$1 - \beta = \Phi\big(\sqrt{n}\,\delta - z_\alpha\big) \tag{4.36}$$

所定の検出力$1 - \beta$を達成するために必要な標本サイズは

$$n = \frac{(z_\alpha + z_\beta)^2}{\delta^2} \tag{4.37}$$

ここでz_αは標準正規分布の上側α分位点である。

読み下し

効果量δが小さいほど、所定の検出力を達成するために 大きな標本サイズが必要である。 例えば$\alpha = 0.05$, $1 - \beta = 0.80$で$\delta = 0.5$なら$n \approx 25$、 $\delta = 0.2$なら$n \approx 155$である。この計算は実験設計の段階で**データ収集前に**行うべきである。

実務ポイント

確認的解析では、少なくとも

1. 検定統計量とp値
2. 点推定と信頼区間
3. 効果量
4. その解析が事前に主要解析として定められていたか

を一緒に示すと、統計的有意性と実務的含意を混同しにくい。

4.9.2　コード例

コード例：R による検出力シミュレーション

```
# === R: 検出力シミュレーション ===
set.seed(42)
n <- 30; delta <- 0.5; alpha <- 0.05; B <- 10000

rejections <- replicate(B, {
  x <- rnorm(n, mean = delta, sd = 1)
  t.test(x, mu = 0, alternative = "two.sided")$p.value < alpha
})
power_sim <- mean(rejections)
power_theory <- power.t.test(
  n = n,
  delta = delta,
  sd = 1,
  sig.level = alpha,
  type = "one.sample",
  alternative = "two.sided"
)$power
cat(sprintf("シミュレーション: %.3f, 理論値: %.3f\n",
            power_sim, power_theory))
```

コード例：Python による検出力シミュレーション

```
# === Python: 検出力シミュレーション ===
import numpy as np
from scipy import stats

rng = np.random.default_rng(42)
n, delta, alpha, B = 30, 0.5, 0.05, 10_000

rejections = np.array([
    stats.ttest_1samp(
        rng.normal(delta, 1, n),
        0,
        alternative="two-sided",
    ).pvalue < alpha
    for _ in range(B)
])
power_sim = rejections.mean()
t_crit = stats.t.ppf(1 - alpha / 2, df=n - 1)
ncp = np.sqrt(n) * delta
power_theory = (
```

```
    stats.nct.sf(t_crit, df=n - 1, nc=ncp)
    + stats.nct.cdf(-t_crit, df=n - 1, nc=ncp)
)
print(f"シミュレーション: {power_sim:.3f}, "
      f"理論値: {power_theory:.3f}")
```

コード例：R による Wald 区間とスコア区間の比較

```
# === R: Wald区間 vs スコア区間のカバレッジ比較 ===
set.seed(123)
n <- 50; p_true <- 0.05; B <- 10000

counts <- rbinom(B, n, p_true)

cover_wald <- vapply(counts, function(x) {
  phat <- x / n
  se <- sqrt(phat * (1 - phat) / n)
  (phat - 1.96 * se <= p_true) & (p_true <= phat + 1.96 * se)
}, logical(1))

cover_score <- vapply(counts, function(x) {
  ci <- prop.test(x, n, conf.level = 0.95, correct = FALSE)$conf.int
  (ci[1] <= p_true) & (p_true <= ci[2])
}, logical(1))

cat(sprintf("Wald: %.3f, Score(Wilson): %.3f (名目値: 0.950)\n",
            mean(cover_wald), mean(cover_score)))
# パラメータが境界付近(p=0.05)ではスコア区間が大幅に優れる
```

コード例：Python によるカバレッジ比較

```
# === Python: カバレッジ比較 ===
import numpy as np
from statsmodels.stats.proportion import (
    proportion_confint)

rng = np.random.default_rng(123)
n, p_true, B = 50, 0.05, 10_000

cover_wald = np.zeros(B, dtype=bool)
cover_wilson = np.zeros(B, dtype=bool)
for i in range(B):
    x = rng.binomial(n, p_true)
    lo_w, hi_w = proportion_confint(x, n, method='normal')
    lo_s, hi_s = proportion_confint(x, n, method='wilson')
    cover_wald[i] = (lo_w <= p_true <= hi_w)
    cover_wilson[i] = (lo_s <= p_true <= hi_s)

print(f"Wald: {cover_wald.mean():.3f}, "
      f"Score(Wilson): {cover_wilson.mean():.3f}")
```

コード例：R による FDR 制御の検証

```
# === R: FDR制御の検証 ===
set.seed(456)
m <- 1000; m0 <- 900; m1 <- m - m0; alpha <- 0.05
B <- 2000
```

```
fdp_bh <- fdp_holm <- fdp_bonf <- numeric(B)
rej_bh_n <- rej_holm_n <- rej_bonf_n <- integer(B)

for (b in seq_len(B)) {
  p_null <- runif(m0)
  p_alt  <- rbeta(m1, shape1 = 1, shape2 = 10)
  p_values  <- c(p_null, p_alt)
  true_null <- c(rep(TRUE, m0), rep(FALSE, m1))

  rej_bh   <- p.adjust(p_values, "BH") < alpha
  rej_holm <- p.adjust(p_values, "holm") < alpha
  rej_bonf <- p.adjust(p_values, "bonferroni") < alpha

  rej_bh_n[b] <- sum(rej_bh)
  rej_holm_n[b] <- sum(rej_holm)
  rej_bonf_n[b] <- sum(rej_bonf)

  fdp_bh[b] <- if (any(rej_bh)) mean(true_null[rej_bh]) else 0
  fdp_holm[b] <- if (any(rej_holm)) mean(true_null[rej_holm]) else 0
  fdp_bonf[b] <- if (any(rej_bonf)) mean(true_null[rej_bonf]) else 0
}

cat(sprintf("BH:   平均棄却 %.1f, 推定FDR %.3f\n",
    mean(rej_bh_n), mean(fdp_bh)))
cat(sprintf("Holm: 平均棄却 %.1f, 推定FDR %.3f\n",
    mean(rej_holm_n), mean(fdp_holm)))
cat(sprintf("Bonf: 平均棄却 %.1f, 推定FDR %.3f\n",
    mean(rej_bonf_n), mean(fdp_bonf)))
```

コード例：R による順列検定

```
# === R: 順列検定 ===
set.seed(789)
x <- rnorm(20, mean = 0)
y <- rnorm(20, mean = 0.5)
T_obs <- mean(x) - mean(y)

combined <- c(x, y)
B <- 10000
T_perm <- replicate(B, {
  idx <- sample.int(length(combined))
  mean(combined[idx[1:20]]) - mean(combined[idx[21:40]])
})

# モンテカルロ近似では +1 補正で p = 0 を避ける
p_perm <- (sum(abs(T_perm) >= abs(T_obs)) + 1) / (B + 1)
p_t    <- t.test(x, y)$p.value
cat(sprintf("順列検定 p = %.4f, t検定 p = %.4f\n",
            p_perm, p_t))
```

主要な結果

重要結果

本章の核心的な結論：

1. **Neyman–Pearson 理論**は、 単純仮説の間では尤度比検定が最強力であることを示す。 ただし両側複合仮説では一般に一様最強力検定は存在しない。
2. **尤度比・Wald・スコアの三大検定**は 大標本で漸近的に等価だが、 有限標本では計算量、変換不変性、境界付近での安定性が異なる。
3. **信頼集合は、対応する検定で棄却されないパラメータ値の集合**である。 この双対性により、検定と信頼区間は同じ原理から構成できる。
4. **信頼区間の作り方**には、 ピボット量に基づく厳密法、Wald 型近似、スコア区間、尤度比区間、 プロファイル尤度区間という使い分けがある。
5. **多重検定と順列検定**は、 単一の p 値だけでは足りない実務上の問題に答える。 FWER と FDR のどちらを守るか、また分布仮定をどこまで置けるかが設計上の分岐点になる。

4.10　演習問題

数値実験を含む問題では、(1) 設定（分布、パラメータ、反復回数、乱数シード）、 (2) 作成した図表または表、(3) 比較指標、(4) 結果から読める一言考察、 の4点を答えに含めよ。

検定問題

演習問題 4.1. Poisson分布Pois(λ)からn個の標本を観測する。 $H_0 : \lambda = \lambda_0 = 1$, $H_1 : \lambda = \lambda_1 = 2$を検定する。

1. Neyman–Pearson補題により、最適な検定統計量を導出せよ。
2. $\alpha = 0.05$のとき、棄却域を決定せよ。
3. サンプルサイズ$n = 100$のとき、検出力$1 - \beta$を計算せよ。

演習問題 4.2. 正規分布$\mathcal{N}(\mu, \sigma^2)$で、$H_0 : \sigma^2 = 1$, $H_1 : \sigma^2 = 4$を検定する場合：

1. 尤度比検定統計量を導出せよ。
2. この統計量が従う分布を大標本で求めよ。
3. Wald検定およびスコア検定と比較せよ。

演習問題 4.3. 線形回帰$y = \beta_0 + \beta_1 x + \epsilon$で$H_0 : \beta_1 = 0$を検定する場合：

1. Wald検定の統計量を導出せよ。

2. $\epsilon \sim \mathcal{N}(0, \sigma^2)$の下で統計量が従う分布を求めよ。

3. σ^2が未知の場合と既知の場合で統計量にどのような違いが生じるか説明せよ。

4. 尤度比検定とスコア検定の形式を述べよ。

信頼区間と双対性

演習問題 4.4. 定理 4.3.1の双対性を用いて、 $\mathcal{N}(\mu, \sigma^2)$（$\sigma^2$既知）における$\mu$の検定

$$\text{reject } H_0 : \mu = \mu_0 \text{ if } \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

から信頼区間$\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$を導出せよ。 また、$\sigma^2$が未知の場合にこの手順を繰り返し、 t分布に基づく信頼区間を導出せよ。

演習問題 4.5. 指数分布Exp(λ)（平均$1/\lambda$）からn個の独立標本を取得した。 ピボット量$Q = 2n\lambda\bar{X}$がχ^2_{2n}分布に従うことを 示し、λの95%信頼区間を構成せよ。

演習問題 4.6. 均一分布Unif$(0, \theta)$からn個の独立標本$X_1, \ldots, X_n$を取得した。 最大値$X_{(n)} = \max(X_1, \ldots, X_n)$の分布を導出し、 これを用いて$\theta$のピボット量を構成し、信頼区間を求めよ。

演習問題 4.7. ベルヌーイ試行$X_1, \ldots, X_n$i.i.d.Bern(p)に対して、 Wald信頼区間とスコア信頼区間（Wilson区間）を導出し、 $p = 0.05$, $n = 50$, モンテカルロ反復$B = 10{,}000$の場合について、 各区間のカバレッジ確率と平均区間長をシミュレーションで比較せよ。

演習問題 4.8. 線形回帰$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ （ε_ii.i.d.$\mathcal{N}(0, \sigma^2)$）において、 プロファイル尤度に基づくβ_1の信頼区間が 古典的なWald区間と一致することを示せ。

同時推論・多重検定

演習問題 4.9. k群の平均$\mu_1, \ldots, \mu_k$に対して、 全ての対比$\sum c_j \mu_j$（$\sum c_j = 0$）について 同時信頼度95%のScheffé信頼区間を構成したい。 Bonferroni法と比較して、 Scheffé法がいつより効率的か、非効率かを論述せよ。

演習問題 4.10. $m = 1000$個の遺伝子の発現量を検定し、p値が得られた。

1. Bonferroni修正で有意水準$\alpha = 0.05$を適用した場合、 各検定の有意水準はいくつか？

2. Benjamini–Hochberg法を適用し、$\alpha = 0.05$でFDRを制御する場合、 棄却する仮説の個数は？ （p値の第k位数が$p_{(k)} = 0.001 \times k/1000$と仮定）

3. 両方法の検出力を比較せよ。

演習問題 4.11. 二標本t検定における順列検定の実装：

1. $n_1 = n_2 = 20$の場合、可能な置換の総数はいくつか？

2. モンテカルロ近似で$B = 10{,}000$回の置換をサンプルした場合、 真のp値が0.05なら、推定されたp値の標準誤差はどのくらいか？

3. $X_i \sim \mathcal{N}(0,1), Y_j \sim \mathcal{N}(\Delta,1)$ と $X_i \sim t_3, Y_j \sim t_3 + \Delta$ の二つの設定を考える。$\Delta = 0$と$\Delta = 0.5$の両方について、名目5%の第一種過誤率と検出力を パラメトリックt検定と順列検定で比較するシミュレーション研究を設計せよ。

演習問題 4.12. 複数の対比検定で$\{c_1, c_2, c_3\}$の三つの対比を同時に検定する。

1. FWER制御のため各対比の有意水準をBonferroni法で設定せよ。
2. Holm法を適用して検出力を改善せよ。
3. FDR制御とFWER制御の違いを説明し、どのような場面で各々が適切か議論せよ。

略解の指針

ここでは、どの原理を使って解答を組み立てるかを明示する。数値計算の細部は自分で回収すること。

- **演習4.1** 使う道具: Neyman–Pearson 補題。最初の1手: 尤度比は $S = \sum_i X_i$ の単調増加関数になることを示す。途中の要点: H_0 の下で $S \sim \mathrm{Pois}(n)$ なので、棄却域は $S \geq c_\alpha$ の形で決まり、c_α は $\mathrm{P}_{H_0}(S \geq c_\alpha) \leq 0.05$ を満たす最小値である。最終形: 検出力は H_1 の下の $\mathrm{Pois}(2n)$ 分布で同じ棄却域に入る確率になる。
- **演習4.2** 使う道具: 尤度比、Wald、スコア。最初の1手: 尤度を μ で最大化すると、残る統計量は残差平方和に依存する。途中の要点: 尤度比統計量は $\hat{\sigma}^2$ の関数で書け、帰無仮説の近傍では二次展開により Wald・スコアと同じ χ^2_1 極限に落ちる。最終形: 有限標本では形が違っても、大標本では三者は漸近的に一致する。
- **演習4.3** 使う道具: 線形回帰の OLS 理論。最初の1手: $H_0 : \beta_1 = 0$ に対する Wald 統計量は $\hat{\beta}_1 / \widehat{\mathrm{se}}(\hat{\beta}_1)$ である。途中の要点: σ^2 既知なら正規分布、未知なら t_{n-2} 分布に従う。最終形: 尤度比検定とスコア検定も同じ 1 次元制約に対する検定であり、大標本では χ^2_1 にそろう。
- **演習4.4** 使う道具: 検定と信頼区間の双対性。最初の1手: 棄却されない μ_0 の集合を不等式でそのまま書き下す。途中の要点: 知られている分散では z 統計量、未知分散では t 統計量に置き換わる。最終形: 反転で得る集合が $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$, あるいは $\bar{X} \pm t_{\alpha/2,n-1} S/\sqrt{n}$ になる。
- **演習4.5** 使う道具: ガンマ分布とカイ二乗分布の関係。最初の1手: $\sum_i X_i \sim \mathrm{Gamma}(n, \lambda)$ を使って $2\lambda \sum_i X_i$ の分布を求める。途中の要点: $Q = 2n\lambda\bar{X}$ は χ^2_{2n} に従うので、その分位点を逆に解けばよい。最終形: 95% CI は $[\chi^2_{2n,0.025}/(2n\bar{X}),\ \chi^2_{2n,0.975}/(2n\bar{X})]$。
- **演習4.6** 使う道具: 最大値の exact pivot。最初の1手: $X_{(n)}/\theta$ の分布関数 $\mathrm{P}(X_{(n)}/\theta \leq u) = u^n$ を求める。途中の要点: 分位点は $u_\alpha = \alpha^{1/n}$ の形で書ける。最終形: 二側信頼区間は $[X_{(n)}/(1-\alpha/2)^{1/n},\ X_{(n)}/(\alpha/2)^{1/n}]$ の形になる。
- **演習4.7** 使う道具: Wald 区間と Wilson 区間。最初の1手: Wald は正規近似をそのまま反転し、Wilson はスコア検定を反転する。途中の要点: $p = 0.05$, $n = 50$ では境界付近なので Wald は被覆を落としやすい。最終形: シミュレーションでは Wilson 区間の方が被覆が安定し、平均区間長とのトレードオフを確認する。

- **演習4.8** 使う道具: 正規線形モデルのプロファイル尤度。 最初の1手: β_1 を固定して残りのパラメータを最尤化する。 途中の要点: 正規線形モデルではプロファイル対数尤度は $\hat{\beta}_1$ のまわりで二次関数になり、標準化すると同じ判定条件に落ちる。 最終形: したがって得られる区間は Wald 型の区間と同じ形に帰着する。

- **演習4.9** 使う道具: 同時信頼区間の対象集合。 最初の1手: Bonferroni は有限個の対比集合、Scheffé は全対比を守る方法だと整理する。 途中の要点: 守る対象が広い分だけ Scheffé は保守的になりやすい。 最終形: 少数の事前指定対比なら Bonferroni が有利で、全対比を一括で扱うなら Scheffé が自然である。

- **演習4.10** 使う道具: Bonferroni と BH。 最初の1手: Bonferroni の閾値は $0.05/1000 = 5 \times 10^{-5}$ である。 途中の要点: 仮定された列では $p_{(k)} = 0.001k/1000 \le 0.05k/1000$ が全 k で成り立つ。 最終形: BH では 1000 個すべて棄却でき、Bonferroni よりはるかに検出力が高い。

- **演習4.11** 使う道具: 組合せと Monte Carlo 誤差。 最初の1手: 置換総数は $\binom{40}{20}$ である。 途中の要点: Monte Carlo による p 値推定の標準誤差は $\sqrt{p(1-p)/B}$ だから、$p = 0.05$, $B = 10{,}000$ では約 0.0022 になる。 最終形: 正規分布では両法の差は小さく、 t_3 のような重尾では順列検定の頑健性が見えやすい。

- **演習4.12** 使う道具: Bonferroni, Holm, FDR。 最初の1手: Bonferroni では各対比を $\alpha/3$ で判定する。 途中の要点: Holm は p 値を小さい順に並べる step-down 法で、FWER を保ったまま少し緩くできる。 最終形: FWER は「1つでも誤棄却しない」を重視し、FDR は「誤棄却の割合」を重視する。 確認的解析では前者、探索的解析では後者が自然である。

次章への橋渡し

本章では、仮説検定と信頼区間の理論を統一的に扱い、 多重検定やノンパラメトリック手法にまで展開した。 これらの手法はいずれも「リスクの制御」という共通の課題に取り組んでいるが、 その最適化を体系的に論じる枠組みはまだ導入していない。

次章（第 5章「決定理論」）では、 推定量や検定を**損失関数とリスク関数**の言葉で評価する 統一的な枠組みを構築する。 本章で「UMP検定」として現れた最適性の概念は、 決定理論の言葉では「許容性」「ミニマックス性」「ベイズ性」として再定式化され、 推定・検定・予測を包含するより広い視点から 統計的手続きの良し悪しを論じることが可能になる。

参考文献ノート

Neyman–Pearson理論の原典はNeyman and Pearson (1933)。 本章の内容の大部分は Casella and Berger (2002, *Statistical Inference*, 2nd ed.) および Lehmann and Romano (2005, *Testing Statistical Hypotheses*, 3rd ed.) に基づく。 三大検定の比較はBuse (1982, ”The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note”) の優れたサーベイが参考になる。 p値の解釈については Wasserstein and Lazar (2016, ”The ASA’s Statement on p-Values: Context, Process, and Purpose”) が必読である。

FDR制御の原典はBenjamini and Hochberg (1995)。 多重検定の包括的な解説はEfron (2010, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*) が優れている。 順列検定の理論はLehmann and Romano (2005, Ch. 5)およびGood (2005, *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd ed.) を参照。 プロファイル尤度についてはDavison (2003, *Statistical Models*) が 丁寧な解説を与えている。 報告実務の観点では、 ASA声明に加えてFDAの multiple endpoints guidance と EMAの multiplicity guideline が、 p値・信頼区間・多重性管理の役割分担を簡潔に整理している。

第5章

決定理論

問いと学習目標

この章で答える問い

- 推定量や検定の「よさ」は、何を基準に定義すればよいのか？
- すべてのパラメータで一様に最良の手続きがないとき、何を指針に選べばよいのか？
- なぜ 3 次元以上では、標本平均より縮小推定の方がよいことがあるのか？

読み終えたらできるようになること

1. 問題に応じて損失関数を選び、リスク関数を導出できる。
2. ベイズ推定量・ミニマックス推定量・許容性の違いを説明できる。
3. James–Stein 推定量と SURE の考え方を、正則化や経験ベイズと結びつけて説明できる。
4. 実務で「平均的に強い手続き」と「最悪ケースに強い手続き」を使い分けられる。

直観的理解

推定量にはさまざまな選び方がある——最尤推定量、モーメント推定量、 ベイズ推定量など。だが「どの推定量が*最*もよいのか」を議論するには、 まず「よさ」を数学的に定義する枠組みが必要である。 決定理論はまさにその枠組みを提供する。
損失関数で「間違いのコスト」を定量化し、 リスク関数で「推定量の平均的な性能」を測る。 すると、すべてのθに対して最良の推定量は一般に存在しない という基本的な事実に直面する。 この困難を乗り越えるために、ベイズ基準（事前分布を仮定して平均リスクを最小化）、 ミニマックス基準（最悪ケースを最小化）、 許容性（他に支配されない最小限の合理性）という三つの最適性概念が登場する。
本章のクライマックスは Stein のパラドックスである。 3次元以上の正規平均の同時推定で、最尤推定量が許容でないという 驚くべき結果は、縮小推定・正則化・経験ベイズへとつながる 現代統計学の重要な出発点となる。

実務ポイント

本章の抽象語彙は、そのまま実務の判断基準になる。 たとえば A/B テストでは「平均的な見逃しを減らす」のか 「最悪ケースの損失を抑える」のかで最適な手続きが変わる。 また、縮小推定は 「多くの係数を同時に推定するとき、少しバイアスを入れても分散を大きく下げた方が得」 という正則化の基本発想を、最もきれいな形で示す。

5.1　損失関数とリスク関数

第3章では推定量の構成法を学んだが、 複数の推定量をどう比較すればよいかは未解決であった。 決定理論はこの問いに答える枠組みである。

5.1.1　決定問題の定式化

定義 5.1.1 (統計的決定問題)**.** 統計的決定問題は以下の四つ組$(\Theta, \mathcal{A}, L, P_\theta)$で定式化される：

(i) **パラメータ空間**Θ： 真の未知状態$\theta \in \Theta$が属する集合。

(ii) **決定空間**$\mathcal{A}$： 統計家が取りうる行動$a \in \mathcal{A}$の集合。 点推定では$\mathcal{A} = \Theta$、検定では$\mathcal{A} = \{0, 1\}$。

(iii) **損失関数**$L\colon \Theta \times \mathcal{A} \to [0, \infty)$： 真のパラメータが$\theta$のとき行動$a$を取った場合の損失。

(iv) **統計モデル**$\{P_\theta : \theta \in \Theta\}$： 観測$\boldsymbol{X}$が従う分布族。

読み下し

決定問題とは「θがわからない状態でデータ$\boldsymbol{X}$を観測し、 行動aを選ばなければならない」という状況の数学的モデル化である。 点推定は$a = \hat{\theta}$、検定は$a \in \{\text{棄却}, \text{採択}\}$、信頼区間は$a = C(\boldsymbol{X}) \subset \Theta$として統一的に扱える。

定義 5.1.2 (決定規則)**.** **決定規則**（または推定量）δとは、 観測データ$\boldsymbol{X}$を行動に写す関数$\delta\colon \mathcal{X} \to \mathcal{A}$である。 すべての決定規則の集まりを$\mathcal{D}$と書く。

定義 5.1.3 (リスク関数)**.** 決定規則δの**リスク関数**は、 固定されたθの下での期待損失として定義される：

$$R(\theta, \delta) = \mathbb{E}_\theta\big[L(\theta, \delta(\boldsymbol{X}))\big] = \int L(\theta, \delta(\boldsymbol{x}))\, f(\boldsymbol{x} \mid \theta)\, d\boldsymbol{x} \tag{5.1}$$

読み下し

リスク関数$R(\theta, \delta)$は「推定量δが、 θが真の値であるときに平均してどれだけ間違えるか」を測る。 θごとに値が変わるため、 リスク関数はθの関数として描くことができる。 すべてのθでリスクが最小の推定量が存在すれば理想的だが、 一般にはそのような推定量は存在しない。

5.1.2　標準的な損失関数

定義 5.1.4 (二乗誤差損失)**.**

$$L(\theta, a) = (\theta - a)^2$$

対応するリスクは平均二乗誤差$\mathrm{MSE}(\delta) = \mathbb{E}_\theta[(\theta - \delta(\boldsymbol{X}))^2]$であり、 第3章のバイアス-分散分解$\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Var}$と直結する。

定義 5.1.5 (絶対誤差損失)**.**

$$L(\theta, a) = |\theta - a|$$

二乗誤差損失より外れ値に頑健である。対応するリスクは$\mathbb{E}_\theta[|\theta - \delta(\boldsymbol{X})|]$。

定義 5.1.6 (0-1損失)**.**

$$L(\theta, a) = \begin{cases} 0 & a = \theta \\ 1 & a \neq \theta \end{cases}$$

分類や検定の文脈で自然に現れる。対応するリスクは誤り確率$\mathrm{P}_\theta(\delta(\boldsymbol{X}) \neq \theta)$。

例 5.1.7 (0-1損失と仮説検定)**.** 検定問題$H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$を 決定空間$\mathcal{A} = \{0, 1\}$（0=採択、1=棄却）の決定問題とみなすと、 損失関数は

$$L(\theta, a) = \begin{cases} a & \theta \in \Theta_0 \quad \text{（第一種の過誤：偽の棄却）} \\ 1 - a & \theta \in \Theta_1 \quad \text{（第二種の過誤：棄却し損ない）} \end{cases}$$

に対応する。 単純仮説$H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$の下で、 事前分布$\pi(\theta_0) = \pi_0, \pi(\theta_1) = 1 - \pi_0$のベイズリスクを最小化する 決定規則は尤度比検定$\mathbf{1}(f(\boldsymbol{x}; \theta_1)/f(\boldsymbol{x}; \theta_0) > c)$の形になる。 特に、第一種の過誤確率を$\alpha$以下に制約した下でのベイズ最適検定は、 第4章の Neyman–Pearson 補題が与える最強力検定と一致する。 すなわち、**NP検定は0-1損失の下での決定理論的に最適な検定である**。

定義 5.1.8 (重み付き損失)**.** 多次元パラメータ$\boldsymbol{\theta} \in \mathbb{R}^p$に対して、

$$L(\boldsymbol{\theta}, \boldsymbol{a}) = (\boldsymbol{\theta} - \boldsymbol{a})^\top W (\boldsymbol{\theta} - \boldsymbol{a})$$

ここでWは正定値行列。$W = I_p$のとき通常の二乗誤差損失に帰着する。

実務ポイント

損失関数の選択は問題の文脈に依存する。 医療診断では偽陰性（病気を見逃す）と偽陽性（健康な人を病気と判定）の コストが大きく異なるため、非対称損失$L(\theta, a) = w_1 \cdot \mathbf{1}(a < \theta) \cdot (\theta - a) + w_2 \cdot \mathbf{1}(a > \theta) \cdot (a - \theta)$ が適切なことがある。 金融リスク管理では下方リスクのみを罰するPinball損失が用いられ、 これは分位点回帰（quantile regression）に対応する。 損失関数を変えると最適な推定量も変わるため、 分析の目的に合った損失の選択が決定理論の出発点となる。

例 5.1.9 (リスク関数の比較)**.** $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$に対して、 二乗誤差損失の下で二つの推定量を比較する。

- 標本平均$\delta_1(\boldsymbol{X}) = \bar{X}$： $R(\theta, \delta_1) = \mathbb{E}_\theta[(\theta - \bar{X})^2] = 1/n$（$\theta$に依存しない）。
- 定数推定量$\delta_2(\boldsymbol{X}) = 0$： $R(\theta, \delta_2) = \theta^2$。

δ_1はθによらず一定のリスクを持つ。 δ_2は$\theta = 0$付近ではリスクが0だが、$|\theta| > 1/\sqrt{n}$ではδ_1に劣る。 どちらが「よい」かはθの値に依存し、 一方が他方を一様に支配しているわけではない。 この状況が決定理論の各種最適性基準を必要とする理由である。

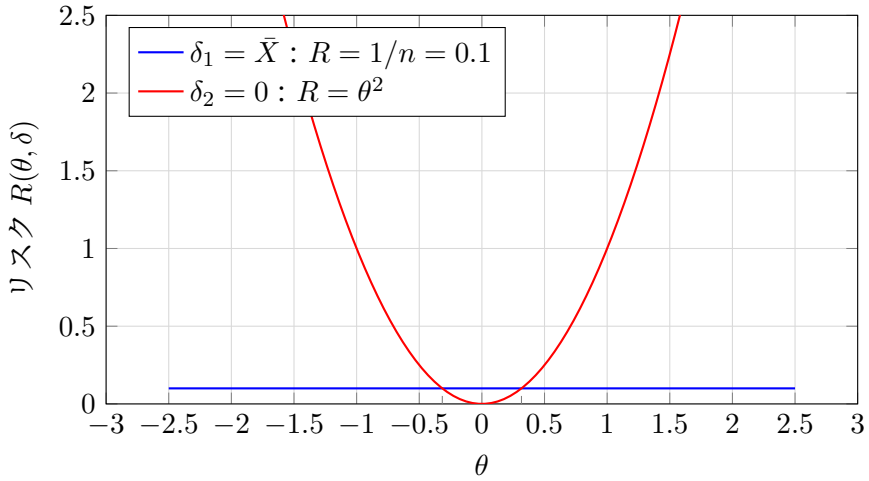


図 5.1: リスク関数の比較（$n = 10$、二乗誤差損失）。 標本平均δ_1はθに依存しない定数リスク$1/n$を持つ（青線）。 定数推定量$\delta_2 = 0$は$\theta = 0$付近で優れるが、 $|\theta| > 1/\sqrt{n} \approx 0.316$で劣る（赤線）。 どちらも他方を一様に支配しておらず、最適な推定量はθに依存する。

5.2　ベイズリスクとベイズ推定量

リスク関数はθの関数であるため、推定量の優劣がθに依存してしまう。 ベイズ的アプローチでは、θに事前分布πを置くことで リスクを単一の数値に集約し、この問題を解決する。

5.2.1　ベイズリスク

定義 5.2.1 (ベイズリスク)**.** 事前分布$\pi(\theta)$の下で、 決定規則δの**ベイズリスク**は

$$r_\pi(\delta) = \int_\Theta R(\theta, \delta)\, \pi(\theta)\, d\theta = \mathbb{E}_\pi\big[R(\theta, \delta)\big] = \mathbb{E}\big[L(\theta, \delta(\boldsymbol{X}))\big] \tag{5.2}$$

右辺の期待値はθと$\boldsymbol{X}$の同時分布$\pi(\theta) f(\boldsymbol{x} \mid \theta)$に関する。

読み下し

ベイズリスクは、リスク関数$R(\theta, \delta)$を事前分布πで重み付け平均したものである。「θの値がわからない」という不確実性を、πを通じた平均で処理する。 事前分布の選択はベイズ推定量を大きく左右するが、 データが十分にあれば事前の影響は薄まる（第6章で詳述）。

定義 5.2.2 (ベイズ推定量)**.** 事前分布πの下で、ベイズリスクを最小化する決定規則δ_π^*を**ベイズ推定量**と呼ぶ：

$$\delta_\pi^* = \arg\min_{\delta \in \mathcal{D}} r_\pi(\delta)$$

5.2.2　損失関数ごとのベイズ推定量

ベイズ推定量の計算は、反復期待値の公式により「各$\boldsymbol{x}$ごとに事後損失を最小化する」問題に帰着する。

定理 5.2.3 (ベイズ推定量の点ごとの最適化)**.** *ベイズリスクは次のように分解できる：*

$$r_\pi(\delta) = \mathbb{E}_{\boldsymbol{X}}\Big[\mathbb{E}_{\theta|\boldsymbol{X}}\big[L(\theta, \delta(\boldsymbol{X})) \mid \boldsymbol{X}\big]\Big]$$

したがって、ベイズ推定量は各$\boldsymbol{x}$で事後期待損失を最小化する：

$$\delta_\pi^*(\boldsymbol{x}) = \arg\min_{a \in \mathcal{A}} \mathbb{E}\big[L(\theta, a) \mid \boldsymbol{X} = \boldsymbol{x}\big]$$

Proof. 反復期待値の公式$\mathbb{E}[g(\theta, \boldsymbol{X})] = \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[g(\theta, \boldsymbol{X}) \mid \boldsymbol{X}]]$を$g(\theta, \boldsymbol{x}) = L(\theta, \delta(\boldsymbol{x}))$に適用する。外側の期待値は非負な被積分関数の積分なので、各$\boldsymbol{x}$で内側の期待値を最小化すれば全体が最小化される。 □

この定理から、損失関数ごとのベイズ推定量が直ちに得られる。

系 5.2.4 (損失関数別のベイズ推定量)**.** *事後分布$\pi(\theta \mid \boldsymbol{X})$が与えられたとき：*

(i) **二乗誤差損失**$L(\theta, a) = (\theta - a)^2 \implies \delta_\pi^*(\boldsymbol{X}) = \mathbb{E}[\theta \mid \boldsymbol{X}]$*（事後平均）*

(ii) **絶対誤差損失**$L(\theta, a) = |\theta - a| \implies \delta_\pi^*(\boldsymbol{X}) = \mathrm{median}(\theta \mid \boldsymbol{X})$*（事後中央値）*

(iii) **0-1損失**$L(\theta, a) = \mathbf{1}(\theta \neq a) \implies \delta_\pi^*(\boldsymbol{X}) = \mathrm{mode}(\theta \mid \boldsymbol{X})$*（事後最頻値 / MAP推定量）*

Proof. (i) $\frac{d}{da}\mathbb{E}[(\theta - a)^2 \mid \boldsymbol{X}] = -2\mathbb{E}[\theta - a \mid \boldsymbol{X}] = 0$より$a = \mathbb{E}[\theta \mid \boldsymbol{X}]$。二階微分は$2 > 0$なので最小値。(ii)は条件付き中央値が$\mathbb{E}[|\theta - a| \mid \boldsymbol{X}]$を最小化する性質、(iii)は離散の場合に最大確率の点が0-1損失を最小化することによる。 □

例 5.2.5 (正規-正規共役モデル)**.** $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$（$\sigma^2$既知）、事前分布$\mu \sim \mathcal{N}(\mu_0, \tau^2)$とする。事後分布は

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}\left(\frac{\sigma^2 \mu_0 + n\tau^2 \bar{X}}{\sigma^2 + n\tau^2},\ \frac{\sigma^2 \tau^2}{\sigma^2 + n\tau^2}\right)$$

読み下し

事後平均は$w \cdot \mu_0 + (1 - w) \cdot \bar{X}$と書ける。ここで$w = \sigma^2/(\sigma^2 + n\tau^2)$は事前情報への重みであり、データが増える（$n$が大きくなる）につれて$w \to 0$となり、ベイズ推定量は最尤推定量$\bar{X}$に収束する。事前分散$\tau^2$が大きいほど事前の影響が弱く、$\tau^2 \to \infty$の極限では$\delta_\pi^*(\boldsymbol{X}) \to \bar{X}$となる。

二乗誤差損失の下で、ベイズ推定量（事後平均）は

$$\delta_\pi^*(\boldsymbol{X}) = \frac{\sigma^2}{\sigma^2 + n\tau^2}\,\mu_0 + \frac{n\tau^2}{\sigma^2 + n\tau^2}\,\bar{X}$$

事前平均μ_0と最尤推定量$\bar{X}$の**精度重み付き平均**である。

5.2.3　事前分布の選択

定義 5.2.6 (Jeffreys事前分布)**.** **Jeffreys事前分布**は

$$\pi_J(\theta) \propto \sqrt{\det\bigl(\mathcal{I}(\theta)\bigr)}$$

ここで$\mathcal{I}(\theta)$はフィッシャー情報行列 （第3章、定義参照）。 この事前分布はパラメータの再パラメータ化に対して不変であるという 優れた性質を持つ。

例 5.2.7 (Jeffreys事前分布：ベルヌーイ分布)**.** $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(p)$に対して、 フィッシャー情報量は$\mathcal{I}(p) = 1/[p(1-p)]$であるから、

$$\pi_J(p) \propto \bigl[p(1-p)\bigr]^{-1/2}$$

これは$\mathrm{Beta}(1/2, 1/2)$分布にほかならない。 事後分布は$p \mid \boldsymbol{X} \sim \mathrm{Beta}(1/2 + \sum X_i,\, 1/2 + n - \sum X_i)$ となり、対応するベイズ推定量（二乗誤差損失）は $\delta_\pi^* = (\sum X_i + 1/2)/(n+1)$である。最尤推定量$\hat{p} = \bar{X}$との差は$O(1/n)$であり、 大標本では一致する。

実務ポイント

無情報事前分布（Jeffreys、Laplace $\pi(\theta) \propto 1$など）は 「データに語らせる」ためのツールとして有用だが、万能ではない。 高次元パラメータでは無情報事前分布が不適切な事後分布を もたらすことがある。 また、$\pi(\theta) \propto 1$は$\theta \in \mathbb{R}$では非正則 （全区間で積分が発散）であり、 事後分布が正則になるかの確認が不可欠である。 実務では、弱情報事前分布（weakly informative prior）—— データ尺度を考慮した適度に広い正則事前分布——が推奨される場面が多い。

5.3　ミニマックス推定

ベイズ基準は事前分布の選択に依存する。 事前知識がないか、最悪ケースに対する保証が必要な状況では、 ミニマックス基準が有用である。

5.3.1　ミニマックス原理

定義 5.3.1 (ミニマックスリスクとミニマックス推定量)**.** 決定規則δの**最大リスク**を

$$\sup_{\theta \in \Theta} R(\theta, \delta)$$

で定める。 **ミニマックスリスク**は

$$R_* = \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta)$$

この値を（漸近的に）達成するδ^*を**ミニマックス推定量**という。

読み下し

ミニマックス推定量は「自然」が最も意地悪なパラメータ値θを選んでも、 リスクが最も小さくなる推定量である。 これはゲーム理論的な解釈を持つ： 統計家が推定量δを選び、自然がθを選ぶ二人ゼロ和ゲームにおいて、 ミニマックス推定量は統計家の最適戦略に対応する。

直観的理解

ミニマックスとベイズの関係は、ゲーム理論のミニマックス定理を 想起すると理解しやすい。 統計家が「推定量δを選ぶ」、自然が「パラメータθを選ぶ」 という二人ゲームを考える。 ベイズ推定量は自然が事前分布πに従ってθを選ぶときの最適応答であり、 ミニマックス推定量は自然が最悪のθを選ぶときの最適応答である。 自然の「最悪の戦略」が事前分布π^*（最小好意的事前分布）として 表現できるとき、二つのアプローチが合流する。

5.3.2　ミニマックスとベイズの関係

定理 5.3.2 (ミニマックス-ベイズの等価性)**.** *事前分布πに対するベイズ推定量δ_πが*

$$\sup_{\theta \in \Theta} R(\theta, \delta_\pi) = r_\pi(\delta_\pi) = \int R(\theta, \delta_\pi)\, \pi(\theta)\, d\theta \tag{5.3}$$

を満たすならば、以下が成立する：

(i) δ_π*はミニマックス推定量である。*

(ii) π*は***最小好意的事前分布***（least favorable prior）である。 すなわち、すべての事前分布π'に対して$r_\pi(\delta_\pi) \geq r_{\pi'}(\delta_{\pi'})$。*

Proof. 任意の決定規則δに対して、

$$\sup_\theta R(\theta, \delta) \geq \int R(\theta, \delta)\, \pi(\theta)\, d\theta = r_\pi(\delta) \geq r_\pi(\delta_\pi)$$

第一の不等号は上限が平均以上であること、 第二はδ_πがベイズ推定量であること（r_πを最小化）による。 仮定(5.3)より $r_\pi(\delta_\pi) = \sup_\theta R(\theta, \delta_\pi)$なので、 $\sup_\theta R(\theta, \delta) \geq \sup_\theta R(\theta, \delta_\pi)$。 δは任意だったからδ_πはミニマックス。 □

読み下し

定理5.3.2の条件(5.3)は、 ベイズ推定量のリスク関数が定数r_πに等しいこと、 すなわちリスクがθに依存しないことを意味する。 リスクが一定であれば最大値は平均値に等しくなるから、 条件が自動的に満たされる。 したがって、ミニマックス推定量の候補を見つけるには、 「リスク関数が定数になるような事前分布」を探せばよい。

例 5.3.3 (正規分布の平均のミニマックス推定)**.** $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$で、 二乗誤差損失を考える。標本平均$\bar{X}$のリスクは

$$R(\mu, \bar{X}) = \mathbb{E}_\mu[(\mu - \bar{X})^2] = \frac{1}{n}$$

μに依存しない定数リスクである。 事前分布$\mu \sim \mathcal{N}(0, M^2)$の下でのベイズ推定量 $\delta_\pi = nM^2\bar{X}/(1+nM^2)$は$M \to \infty$で$\bar{X}$に収束し、 その最大リスクも$1/n$に収束する。 定理5.3.2の条件を（極限で）満たすため、 $\bar{X}$はミニマックス推定量である。

5.4　許容性と完備クラス定理

5.4.1　許容性の概念

定義 5.4.1 (支配と許容性)**.** 決定規則δ_1がδ_2を**支配**（dominate）するとは、

$$R(\theta, \delta_1) \le R(\theta, \delta_2) \quad \text{for all } \theta \in \Theta$$

かつ少なくとも一つのθ_0で$R(\theta_0, \delta_1) < R(\theta_0, \delta_2)$ が成立することをいう。

決定規則δが**許容**（admissible）であるとは、 δを支配する決定規則が存在しないことをいう。

読み下し

許容性は「最小限の合理性」を保証する概念である。 許容でない推定量は、すべてのθで同等以上（かつあるθでより良い） 性能の推定量が存在するので、使い続ける理由がない。 ただし、許容であることだけでは推定量の「良さ」の保証としては弱い。 例えば、定数推定量$\delta(\boldsymbol{X}) = c$は二乗誤差損失の下で $\theta = c$付近では優れたリスクを持ち、 場合によっては許容であり得る（支配する推定量が見つからない）が、 実際の分析で使うことは稀である。 許容性は必要条件であり、十分条件ではない。

5.4.2　ベイズ推定量と許容性

ベイズ推定量と許容性の間には深い関係がある。

定理 5.4.2 (一意なベイズ推定量の許容性)**.** *事前分布πに対するベイズ推定量δ_πが存在し、 かつ唯一のベイズ推定量であるならば、δ_πは許容である。*

Proof. 背理法で示す。δ_πが許容でないと仮定すると、 あるδ'がδ_πを支配する。 すなわち、すべてのθで$R(\theta, \delta') \le R(\theta, \delta_\pi)$ かつ少なくとも一つのθ_0で$R(\theta_0, \delta') < R(\theta_0, \delta_\pi)$である。 この不

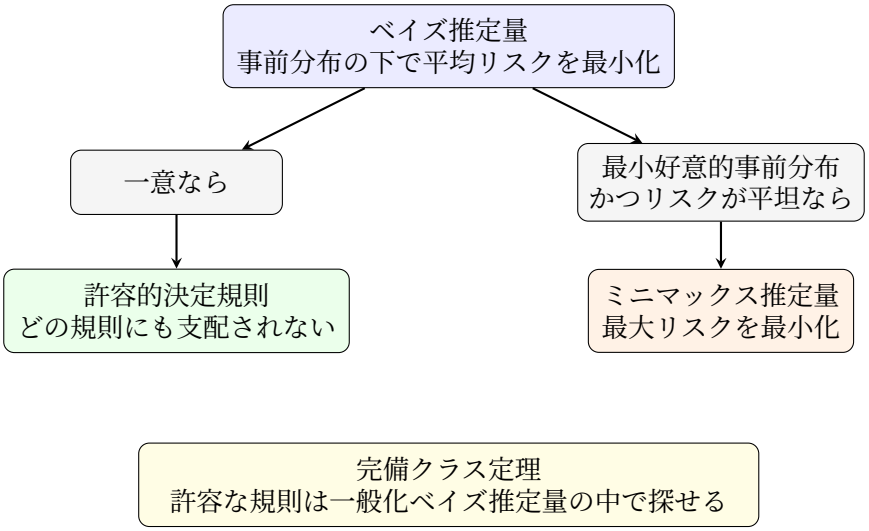


図 5.2: 決定理論の主要概念の関係。矢印は本文で示した十分条件を表す。完備クラス定理は、許容な規則を探す空間を一般化ベイズ推定量へ絞れることを述べる。

等式を事前分布πで積分すると、

$$r_\pi(\delta') = \int R(\theta, \delta')\,\pi(d\theta) \le \int R(\theta, \delta_\pi)\,\pi(d\theta) = r_\pi(\delta_\pi)$$

を得る。一方、δ_πはベイズ推定量なので、 $r_\pi(\delta_\pi) \le r_\pi(\delta')$でもある。 したがって$r_\pi(\delta') = r_\pi(\delta_\pi)$であり、 δ'もベイズ推定量になる。 これはδ_πが唯一のベイズ推定量であることに矛盾する。 □

定理 5.4.3 (完備クラス定理)**.** *パラメータ空間Θが有限でリスク集合が閉であるとき、 すべての許容的決定規則はある事前分布πに対するベイズ推定量である。 すなわち、ベイズ推定量のクラスは***完備クラス***（complete class）である：*

$$\{\text{許容的決定規則}\} \subseteq \{\text{ベイズ推定量}\}$$

Θが無限（例えば$\Theta = \mathbb{R}$）の場合、許容的決定規則は ベイズ推定量またはその極限（一般化ベイズ推定量）として表現される：

$$\{\text{許容的決定規則}\} \subseteq \overline{\{\text{ベイズ推定量}\}}$$

ここで閉包はリスク関数の収束に関してとる。

読み下し

完備クラス定理は「良い推定量を探すならベイズ推定量の中を探せばよい」 ということを保証する、頻度主義とベイズの橋渡しとなる深い結果である。 パラメータ空間が有限の場合の証明は 超平面分離定理（凸集合の分離定理）に基づく。 無限の場合はさらに技術的であり、 Wald (1950)、Le Cam (1955) による一般的な理論が基盤となる。

例 5.4.4 (正規平均の推定における許容性)**.** $X \sim \mathcal{N}(\mu, 1)$の単一観測に対して、二乗誤差損失を考える。

- $\delta_1(X) = X$（最尤推定量）：$R(\mu, \delta_1) = 1$（全てのμで定数）。事前$\mu \sim \mathcal{N}(0, M^2)$で$M \to \infty$の極限ベイズ推定量。1次元では許容。
- $\delta_2(X) = cX$（$0 < c < 1$）：$R(\mu, \delta_2) = (1-c)^2\mu^2 + c^2$。$\mu = 0$付近でリスクが低いが$|\mu|$が大きいと劣化。$c = \tau^2/(1+\tau^2)$とおけば事前$\mu \sim \mathcal{N}(0, \tau^2)$のベイズ推定量であり、許容。
- $\delta_3(X) = 0$（定数推定量）：$R(\mu, \delta_3) = \mu^2$。$\mu = 0$でリスクが0であるため、これを支配する推定量$\tilde{\delta}$が存在するなら$R(0, \tilde{\delta}) = 0$でなければならない。したがって$\tilde{\delta}(X) = 0$が$P_{\mu=0}$-ほとんど確実に成り立つ。正規分布族は互いに絶対連続なので、これは全てのμに対して$\tilde{\delta}(X) = 0$を意味し、厳密な改善は起こせない。ゆえにδ_3も許容である。

5.5　James–Stein 推定量と縮小推定

決定理論の最も驚くべき帰結が、Stein のパラドックスである。

5.5.1　Stein のパラドックス

定理 5.5.1 (Stein のパラドックス)**.** $\boldsymbol{X} = (X_1, \dots, X_p) \sim \mathcal{N}(\boldsymbol{\mu}, I_p)$（$p \geq 3$）で、*二乗誤差損失*

$$L(\boldsymbol{\mu}, \boldsymbol{a}) = \|\boldsymbol{\mu} - \boldsymbol{a}\|^2$$

を考える。この結果は James & Stein (1961) による。最尤推定量$\hat{\boldsymbol{\mu}} = \boldsymbol{X}$のリスクは

$$R(\boldsymbol{\mu}, \boldsymbol{X}) = \mathbb{E}\big[\|\boldsymbol{\mu} - \boldsymbol{X}\|^2\big] = p$$

一方、**James–Stein 推定量**

$$\delta^{\mathrm{JS}}(\boldsymbol{X}) = \left(1 - \frac{p-2}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X} \tag{5.4}$$

のリスクは

$$R(\boldsymbol{\mu}, \delta^{\mathrm{JS}}) = p - (p-2)^2\,\mathbb{E}\left[\frac{1}{\|\boldsymbol{X}\|^2}\right] < p \tag{5.5}$$

すべての$\boldsymbol{\mu} \in \mathbb{R}^p$に対して成立する。したがって、$p \geq 3$のとき*最尤推定量$\boldsymbol{X}$は*許容でない。

読み下し

James–Stein 推定量の式(5.4)は「観測値$\boldsymbol{X}$に縮小係数$1 - (p-2)/\|\boldsymbol{X}\|^2$を掛ける」と読む。$\|\boldsymbol{X}\|^2$が大きい（データが原点から遠い）ときは縮小係数が1に近く、推定量は$\boldsymbol{X}$とほとんど変わらない。$\|\boldsymbol{X}\|^2$が小さい（データが原点に近い）ときは強く縮小される。分子の$p-2$は次元による補正であり、$p \geq 3$のときのみ縮小の利得がバイアスのコストを上回る。リスクの式(5.5)は、改善幅$(p-2)^2\mathbb{E}[1/\|\boldsymbol{X}\|^2]$が常に正であるため全ての$\boldsymbol{\mu}$でリスクが減少することを示している。

直観的理解

このパラドックスが驚くべき理由を考えよう。 X_1は東京の気温、X_2は為替レート、X_3は打率の推定値だとする。 これらは互いに無関係なのに、3つを同時に推定するとき、 各変数を個別に推定するより全体を原点方向に「縮める」方が 平均二乗誤差が改善される。
直観的な理解：p次元正規分布から1標本を観測すると、 $\|\boldsymbol{X}\|^2$は$\|\boldsymbol{\mu}\|^2$を系統的に過大評価する （$\mathbb{E}[\|\boldsymbol{X}\|^2] = \|\boldsymbol{\mu}\|^2 + p$）。 James–Stein 推定量は$\boldsymbol{X}$を原点に向かって縮小することで、 この過大評価を補正する。 $p \geq 3$のとき、この補正の利得がバイアスのコストを上回る。
なお、$p = 1, 2$のときには$\boldsymbol{X}$は許容であり、 パラドックスは$p \geq 3$でのみ成立する。

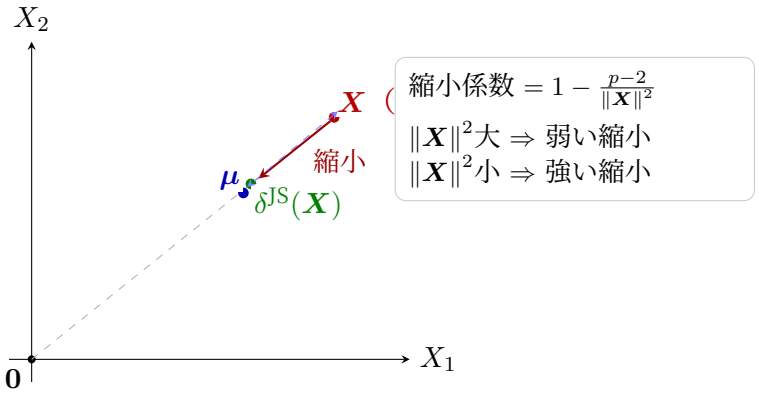


図 5.3: James–Stein推定量の幾何学的解釈（2次元で図示）。 観測値$\boldsymbol{X}$を原点方向に縮小係数$1 - (p-2)/\|\boldsymbol{X}\|^2$だけ 引き寄せる。縮小によりバイアスが生じるが、 $p \geq 3$では分散の減少がバイアスのコストを上回り、 総合的なリスク（二乗誤差）が改善される。

この証明を与えるために、 まず強力な汎用ツールであるStein の不偏リスク推定を導入する。

5.5.2　Stein の補題と不偏リスク推定（SURE）

補題 5.5.2 (Stein の補題)**.** *$Z \sim \mathcal{N}(\mu, 1)$で、$g\colon \mathbb{R} \to \mathbb{R}$が$\mathbb{E}[|g'(Z)|] < \infty$を満たす弱微分可能な関数であるとき、*

$$\mathbb{E}\big[(Z-\mu)\, g(Z)\big] = \mathbb{E}\big[g'(Z)\big] \tag{5.6}$$

Proof. $Z - \mu \sim \mathcal{N}(0,1)$であるから、$\mathbb{E}[(Z-\mu)g(Z)]$を標準正規密度$\varphi$を用いて書く：

$$\begin{aligned}\mathbb{E}[(Z-\mu)g(Z)] &= \int_{-\infty}^{\infty} (z-\mu)\, g(z)\, \varphi(z-\mu)\, dz \\ &= \int_{-\infty}^{\infty} g(z)\, \big[(z-\mu)\varphi(z-\mu)\big]\, dz\end{aligned}$$

$(z-\mu)\varphi(z-\mu) = -\varphi'(z-\mu)$に注意して部分積分を行う：

$$= \int_{-\infty}^{\infty} g(z)\,[-\varphi'(z-\mu)]\,dz = -\big[g(z)\varphi(z-\mu)\big]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} g'(z)\,\varphi(z-\mu)\,dz$$

境界項はφの急減少性により消え、$= \mathbb{E}[g'(Z)]$を得る。 □

定理 5.5.3 (Stein の不偏リスク推定（SURE）)**.** *$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, I_p)$に対して、推定量$\hat{\boldsymbol{\mu}} = \boldsymbol{X} + \boldsymbol{g}(\boldsymbol{X})$（$\boldsymbol{g}\colon \mathbb{R}^p \to \mathbb{R}^p$は弱微分可能）のリスクは*

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \mathbb{E}\big[\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2\big] = p + \mathbb{E}\big[\|\boldsymbol{g}(\boldsymbol{X})\|^2 + 2\,\mathrm{div}\,\boldsymbol{g}(\boldsymbol{X})\big] \tag{5.7}$$

ここで$\mathrm{div}\,\boldsymbol{g}(\boldsymbol{X}) = \sum_{i=1}^{p} \frac{\partial g_i}{\partial X_i}(\boldsymbol{X})$はダイバージェンスである。

特に、

$$\mathrm{SURE}(\hat{\boldsymbol{\mu}}) = \|\boldsymbol{g}(\boldsymbol{X})\|^2 + 2\,\mathrm{div}\,\boldsymbol{g}(\boldsymbol{X}) + p$$

*は$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$の**不偏推定量**であり、$\boldsymbol{\mu}$を含まないため、データだけからリスクを推定できる。*

Proof. $\hat{\boldsymbol{\mu}} = \boldsymbol{X} + \boldsymbol{g}(\boldsymbol{X})$、$\boldsymbol{\mu} - \hat{\boldsymbol{\mu}} = (\boldsymbol{\mu} - \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{X})$より、

$$\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2 = \|\boldsymbol{\mu} - \boldsymbol{X}\|^2 + 2(\boldsymbol{X} - \boldsymbol{\mu})^{\top}\boldsymbol{g}(\boldsymbol{X}) + \|\boldsymbol{g}(\boldsymbol{X})\|^2$$

期待値を取ると$\mathbb{E}[\|\boldsymbol{\mu} - \boldsymbol{X}\|^2] = p$。クロス項にStein の補題（補題5.5.2の多次元版）を適用すると

$$\mathbb{E}\big[(X_i - \mu_i) g_i(\boldsymbol{X})\big] = \mathbb{E}\left[\frac{\partial g_i}{\partial X_i}(\boldsymbol{X})\right]$$

各iについて足し合わせて $\mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu})^{\top}\boldsymbol{g}(\boldsymbol{X})] = \mathbb{E}[\mathrm{div}\,\boldsymbol{g}(\boldsymbol{X})]$。よって

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = p + 2\,\mathbb{E}[\mathrm{div}\,\boldsymbol{g}(\boldsymbol{X})] + \mathbb{E}[\|\boldsymbol{g}(\boldsymbol{X})\|^2]$$

これが式(5.7)である。 □

読み下し

SURE公式(5.7)が革命的なのは、推定量のリスク$\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$を**未知の$\boldsymbol{\mu}$を使わずに**推定できることである。$\boldsymbol{g}$と$\mathrm{div}\,\boldsymbol{g}$はデータ$\boldsymbol{X}$のみから計算可能であり、リスクの不偏推定量をデータのみで構成できる。これにより、推定量のクラス内でSUREを最小化するような「データ駆動型のチューニング」が可能となる。

5.5.3　James–Stein推定量のリスク解析

定理5.5.1の証明. James–Stein推定量(5.4)を $\hat{\boldsymbol{\mu}} = \boldsymbol{X} + \boldsymbol{g}(\boldsymbol{X})$の形に書く：

$$\boldsymbol{g}(\boldsymbol{X}) = -\frac{p-2}{\|\boldsymbol{X}\|^2}\,\boldsymbol{X}$$

各成分は$g_i(\boldsymbol{X}) = -(p-2)X_i/\|\boldsymbol{X}\|^2$。 ダイバージェンスを計算する。$X_i$について偏微分すると

$$\frac{\partial g_i}{\partial X_i} = -(p-2)\cdot\frac{\|\boldsymbol{X}\|^2 - 2X_i^2}{\|\boldsymbol{X}\|^4}$$

$i = 1,\dots,p$について足し合わせると

$$\operatorname{div}\boldsymbol{g}(\boldsymbol{X}) = -(p-2)\cdot\frac{p\|\boldsymbol{X}\|^2 - 2\|\boldsymbol{X}\|^2}{\|\boldsymbol{X}\|^4} = -\frac{(p-2)^2}{\|\boldsymbol{X}\|^2}$$

（$\sum X_i^2 = \|\boldsymbol{X}\|^2$を使った。）

読み下し

ダイバージェンスの結果$\operatorname{div}\boldsymbol{g} = -(p-2)^2/\|\boldsymbol{X}\|^2$は、 商の微分則$\partial(X_i/\|\boldsymbol{X}\|^2)/\partial X_i = (1/\|\boldsymbol{X}\|^2) - 2X_i^2/\|\boldsymbol{X}\|^4$ をp個足し合わせて得られる。 $\sum_i(1/\|\boldsymbol{X}\|^2) = p/\|\boldsymbol{X}\|^2$、$\sum_i 2X_i^2/\|\boldsymbol{X}\|^4 = 2/\|\boldsymbol{X}\|^2$であるから $(p-2)/\|\boldsymbol{X}\|^2$となり、$-(p-2)$を掛けて結論を得る。 この**負の**ダイバージェンスが、 SURE公式においてリスクをpから引き下げる役割を果たす。

$\|\boldsymbol{g}(\boldsymbol{X})\|^2 = (p-2)^2/\|\boldsymbol{X}\|^2$であるから、 SURE公式(5.7)より

$$\begin{aligned} R(\boldsymbol{\mu},\delta^{\mathrm{JS}}) &= p + \mathbb{E}\left[\frac{(p-2)^2}{\|\boldsymbol{X}\|^2} - 2\frac{(p-2)^2}{\|\boldsymbol{X}\|^2}\right] \\ &= p - (p-2)^2\,\mathbb{E}\left[\frac{1}{\|\boldsymbol{X}\|^2}\right] \end{aligned}$$

$\|\boldsymbol{X}\|^2 \sim \chi^2_p(\|\boldsymbol{\mu}\|^2)$（非心カイ二乗分布、非心度$\|\boldsymbol{\mu}\|^2$） は正の確率変数なので$\mathbb{E}[1/\|\boldsymbol{X}\|^2] > 0$。 $p \geq 3$のとき$(p-2)^2 > 0$であるから、 $R(\boldsymbol{\mu},\delta^{\mathrm{JS}}) < p = R(\boldsymbol{\mu},\boldsymbol{X})$ がすべての$\boldsymbol{\mu}$で成立する。 □

定義 5.5.4 (正値部分James–Stein推定量)**.** James–Stein推定量(5.4)は縮小係数が負になることがある （$\|\boldsymbol{X}\|^2 < p-2$のとき推定が原点の反対側に飛ぶ）。 **正値部分**James–Stein推定量はこれを修正したものである：

$$\delta^{\mathrm{JS}+}(\boldsymbol{X}) = \left(1 - \frac{p-2}{\|\boldsymbol{X}\|^2}\right)^{+}\boldsymbol{X} \tag{5.8}$$

ここで$(x)^+ = \max(x,0)$。 この推定量はJames–Stein推定量をさらに支配する （$R(\boldsymbol{\mu},\delta^{\mathrm{JS}+}) \leq R(\boldsymbol{\mu},\delta^{\mathrm{JS}})$） 。

例 5.5.5 (数値例：$p=5$の場合のリスク比較)**.** $\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{\mu},I_5)$で$\boldsymbol{\mu} = (1,1,1,1,1)^\top$ （$\|\boldsymbol{\mu}\|^2 = 5$）の場合を考える。

$\|\boldsymbol{X}\|^2 \sim \chi^2_5(5)$（非心カイ二乗分布）に対して、 $\mathbb{E}[1/\|\boldsymbol{X}\|^2]$の正確な値は 非心カイ二乗分布のモーメント公式から計算できる。 数値的には$\mathbb{E}[1/\|\boldsymbol{X}\|^2] \approx 0.149$であるから、

$$R(\boldsymbol{\mu},\delta^{\mathrm{JS}}) \approx 5 - 9\times 0.149 \approx 3.66$$

最尤推定量のリスク5に対して約27%のリスク削減である。

$\|\boldsymbol{\mu}\|^2$が大きくなると $\mathbb{E}[1/\|\boldsymbol{X}\|^2] \to 0$となるため改善幅は縮小するが、すべての$\boldsymbol{\mu}$で改善は正であり続ける。

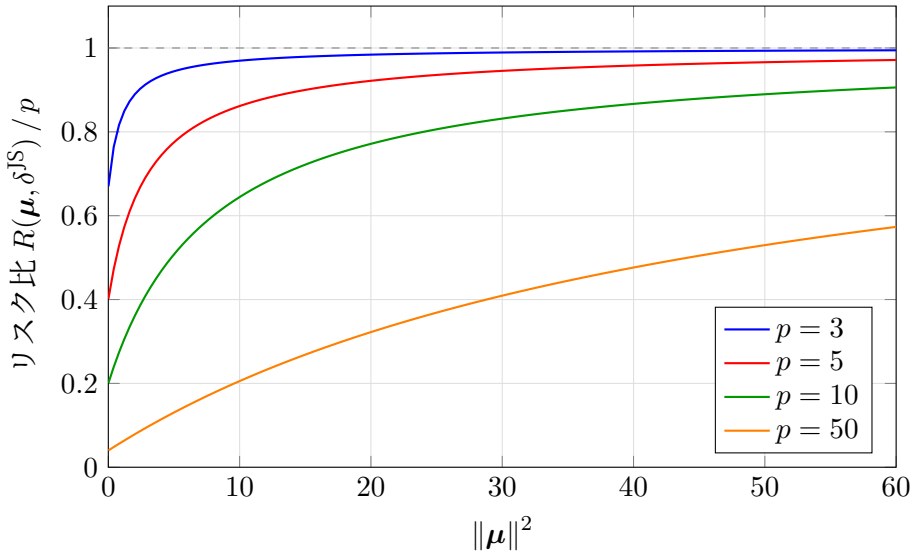


図 5.4: James–Stein推定量のリスク比$R(\boldsymbol{\mu}, \delta^{\mathrm{JS}})/p$と$\|\boldsymbol{\mu}\|^2$の関係。次元$p$が大きいほど$\|\boldsymbol{\mu}\|^2 = 0$付近での改善幅が大きい。$\|\boldsymbol{\mu}\|^2 \to \infty$ではリスク比は1に近づくが、すべての$\boldsymbol{\mu}$で1未満であり続ける（James–Stein推定量による一様な支配）。近似式$R/p \approx 1 - (p-2)^2/[p(p + \|\boldsymbol{\mu}\|^2 - 2)]$を用いた。

5.5.4　経験ベイズとの関係

James–Stein推定量は経験ベイズの視点から自然に理解できる。

$X_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1)$、$\mu_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2)$という階層モデルを考える。このとき$\mu_i$のベイズ推定量（事後平均）は

$$\delta_{\tau^2}(X_i) = \frac{\tau^2}{1+\tau^2} X_i = \left(1 - \frac{1}{1+\tau^2}\right) X_i$$

τ^2は未知だが、周辺分布$X_i \sim \mathcal{N}(0, 1+\tau^2)$から推定できる。$\sum X_i^2 \sim (1+\tau^2)\chi^2_p$の期待値は$p(1+\tau^2)$であるから、$1/(1+\tau^2)$の自然な推定量は$(p-2)/\|\boldsymbol{X}\|^2$（不偏性のための補正で$p$ではなく$p-2$を用いる）。代入すると

$$\hat{\delta}(X_i) = \left(1 - \frac{p-2}{\|\boldsymbol{X}\|^2}\right) X_i$$

これはまさにJames–Stein推定量(5.4)である。

読み下し

James–Stein推定量の「縮小」は、暗黙の階層モデルにおいて 個別の推定値をグローバルな構造に向かって引き寄せる操作と解釈できる。データから事前分布のハイパーパラメータを推定して ベイズ推定量を構成する——これがRobbins (1956) の **経験ベイズ**の基本思想であり、James–Stein推定量はその具体例である。

この視点は、第6章以降で学ぶ正則化（Ridge回帰、LASSO）の理論的基盤ともなる。Ridge回帰の推定量$\hat{\boldsymbol{\beta}}_\lambda = (X^\top X + \lambda I)^{-1} X^\top \boldsymbol{y}$は、事前$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1} I)$のベイズ推定量であり、James–Stein推定量と同じ「原点方向への縮小」の原理に基づく。

5.5.5　縮小推定の実務的意義

実務ポイント

James–Stein推定量はそのまま実務で使うことは稀である。$\|\boldsymbol{X}\|^2 \approx 0$のとき推定が不安定になり、正値部分(5.8)を使っても原点への過度の縮小が起こり得る。また、原点が特別な意味を持たない場合には不自然である。
実務では、同じ「縮小」の原理がより洗練された形で実装される：

- **Ridge回帰**（巻2 第10章）：回帰係数をゼロに向かって縮小。λは交差検証で選択。
- **LASSO**（巻2 第10章）：スパースな縮小。一部の係数を正確にゼロにする。
- **階層ベイズ**（巻2 第7章）：経験ベイズをさらに発展させ、ハイパーパラメータにも事前分布を置く。
- **SURE最小化**：推定量のパラメトリッククラス内でSURE(5.7)を 最小化して最適な縮小度を選択する（ウェーブレット閾値選択などで活用）。

Stein のパラドックスの核心的メッセージは：**高次元推定では、個別の推定を信頼するよりも、全体の構造を活用した縮小・正則化が本質的に必要である。**

5.6　最適性基準の選択指針

本章では三つの最適性基準——ベイズ、ミニマックス、許容性——を導入した。実際の問題でどの基準を用いるべきかの指針をまとめる。

実務ポイント

最適性基準の選択ガイド

基準	適する状況	注意点
ベイズ	事前情報が利用可能。複数パラメータの同時推定（階層モデル）。小標本で特に有効	事前分布の選択に結果が依存する。非正則事前分布では事後分布の正則性の確認が必要
ミニマックス	最悪ケースの保証が必要。敵対的環境（ゲーム理論的状況）。事前情報が不確実または存在しない	保守的になりがち。実務での最悪ケースは理論上の最悪と異なることがある
許容性	推定量の妥当性の最低基準として。候補の選定段階で許容でない推定量を排除	許容であっても実用的に良いとは限らない（定数推定量も許容でありうる）。必要条件として使用

実務的な手順：

1. まず**許容性**で候補を絞る（完備クラス定理により、ベイズ推定量のクラス内で探す）
2. 信頼できる事前情報があれば**ベイズ基準**で選択する
3. 最悪ケース保証が必要なら**ミニマックス基準**を適用する
4. 多くの場合、ベイズ推定量は良好なミニマックス性質も持つ（定理5.3.2参照）

実務ポイント

ベイズと頻度主義は、互いに排他的な陣営というより、 同じ問題を異なる角度から評価する二つの言語とみると理解しやすい。 ベイズは事前情報と損失を明示して個別問題の意思決定を最適化する立場であり、 頻度主義は反復標本での誤差確率や被覆確率を通じて 手続きの長期性能を保証する立場である。 現代実務では両者は役割分担で併用されることが多く、 事前情報の統合や逐次的な設計更新ではベイズ的整理が有効であり、 主要解析や感度分析の透明な報告では頻度主義の整理が依然として重要である。

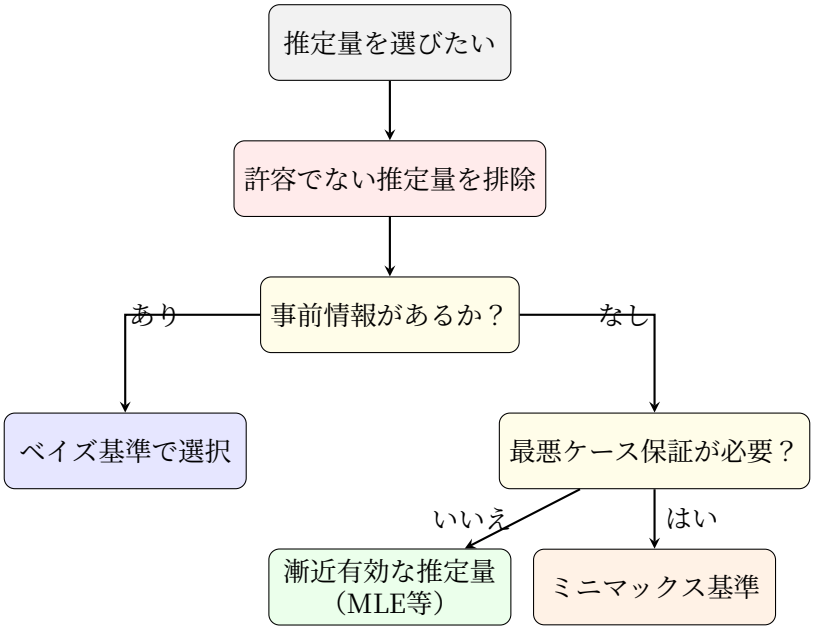


図 5.5: 最適性基準の選択フローチャート。 まず許容でない推定量を排除し（完備クラス定理が指針）、 事前情報の有無と最悪ケース保証の要否に応じて基準を選択する。

5.7　実践：各種決定規則のリスク比較

James–Stein推定量のリスク改善をシミュレーションで確認する。

コード例：James–Stein推定量のリスク比較（R）

```
# James--Stein推定量と最尤推定量のリスク比較
set.seed(42)
p <- 10          # 次元
n_sim <- 5000    # シミュレーション回数
mu <- rep(1, p) # 真のパラメータ

mse_ml <- mse_js <- mse_jsplus <- numeric(n_sim)
for (i in seq_len(n_sim)) {
  X <- rnorm(p, mean = mu, sd = 1)
  norm_sq <- sum(X^2)

  # 最尤推定量
  ml <- X
  mse_ml[i] <- sum((ml - mu)^2)

  # James--Stein推定量
  shrink <- 1 - (p - 2) / norm_sq
  js <- shrink * X
  mse_js[i] <- sum((js - mu)^2)

  # 正値部分James--Stein推定量
  shrink_pos <- max(0, shrink)
  jsplus <- shrink_pos * X
  mse_jsplus[i] <- sum((jsplus - mu)^2)
}
```

```
cat("理論リスク（MLE）:", p, "\n")
cat("推定リスク（MLE）:", mean(mse_ml), "\n")
cat("推定リスク（JS）: ", mean(mse_js), "\n")
cat("推定リスク（JS+）:", mean(mse_jsplus), "\n")
cat("リスク削減率（JS+）:",
    round(100 * (1 - mean(mse_jsplus) / mean(mse_ml)), 1), "%\n")
```

コード例：SUREによるリスク推定（Python）

```
import numpy as np

def james_stein_risk_simulation(p, mu, n_sim=5000, seed=42):
    """James--Stein推定量のリスクをシミュレーションとSUREで比較"""
    if p < 3:
        raise ValueError("James--Stein推定量は p >= 3 を仮定する")

    rng = np.random.default_rng(seed)
    mu = np.asarray(mu, dtype=float)

    mse_ml = np.zeros(n_sim)
    mse_js = np.zeros(n_sim)
    sure_js = np.zeros(n_sim)

    for i in range(n_sim):
        X = rng.normal(loc=mu, scale=1.0)
        norm_sq = np.sum(X**2)

        # 最尤推定量
        mse_ml[i] = np.sum((X - mu)**2)

        # James--Stein推定量
        shrink = 1 - (p - 2) / norm_sq
        js = shrink * X
        mse_js[i] = np.sum((js - mu)**2)

        # SURE（通常のJames--Stein推定量に対するリスク推定）
        # g(X) = -(p-2)/||X||^2 * X
        # div g = -(p-2)^2/||X||^2
        # SURE = ||g||^2 + 2 * div g + p
        g = -((p - 2) / norm_sq) * X
        g_norm_sq = np.sum(g**2)
        div_g = -(p - 2)**2 / norm_sq
        sure_js[i] = g_norm_sq + 2 * div_g + p

    print(f"次元 p = {p}, ||mu||^2 = {np.sum(mu**2):.1f}")
    print(f"  MLE リスク（シミュレーション）: {mse_ml.mean():.3f}")
    print(f"  JS  リスク（シミュレーション）: {mse_js.mean():.3f}")
    print(f"  JS  リスク（SURE推定）:         {sure_js.mean():.3f}")
    print(f"  リスク削減率: {100*(1 - mse_js.mean()/mse_ml.mean()):.1f}%")

# 実行例
james_stein_risk_simulation(p=10, mu=[1]*10)
james_stein_risk_simulation(p=10, mu=[5]*10)  # ||mu||^2が大きい場合
```

主要な結果のまとめ

重要結果

本章の核心的な結論：

1. **すべてのθで最良の推定量は存在しない**。 推定量の比較には、損失関数の選択と最適性基準 （ベイズ、ミニマックス、許容性）の指定が不可欠。
2. **ベイズ推定量**は事前分布の下で平均リスクを最小化し、 損失関数に応じて事後平均・事後中央値・事後最頻値として具体化される。
3. **ミニマックス推定量**は最悪ケースのリスクを最小化する。 「リスクがθに依存しないベイズ推定量」はミニマックスでもある （定理5.3.2）。
4. **完備クラス定理**により、 許容推定量はベイズ推定量（またはその極限）の中にある。 頻度主義とベイズのアプローチは、許容性を通じて深くつながる。 現代実務では、事前情報の統合と長期性能の保証を役割分担で使い分ける。
5. **Stein のパラドックス**： $p \geq 3$次元の正規平均推定で最尤推定量は許容でない。 SUREを通じて証明され、 **高次元推定では縮小・正則化が本質的に有利** という現代統計学の基本原理を示す。 経験ベイズ、Ridge回帰、LASSOはすべてこの原理の実装である。

次章への橋渡し

本章では推定量の有限標本での最適性を論じた。 しかし、多くの問題では最適な推定量を閉じた形で求めることが難しい。 そこで、$n \to \infty$の下で推定量の性質を調べる **漸近理論**が重要となる。 次章（第6章）では、 最尤推定量の一致性と漸近正規性、 その漸近分散がフィッシャー情報量の逆数に一致すること（漸近有効性）、 さらにLAN理論と経験過程の基礎を扱う。 漸近理論は、決定理論で確立した最適性の概念を 「大標本での近似的な最適性」へと拡張する枠組みである。 第4章で現れた Wald・尤度比・スコアの三つの検定が、 なぜ大標本でそろっていくのかも次章で整理される。

演習問題

標準問題

演習問題 5.1. $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\theta)$（期待値$\theta$）に対して、 二乗誤差損失を考える。

(a) 標本平均$\bar{X}$のリスク関数$R(\theta, \bar{X})$を計算せよ。

(b) $c\bar{X}$の形の推定量の中でリスク$R(\theta, c\bar{X})$を最小化する c^*を求めよ。$c^* = 1$（つまり$\bar{X}$が最適）とならないことを示せ。

(c) $c^*\bar{X}$は$\bar{X}$を支配するか？ すなわち、$c^*\bar{X}$は全てのθで$\bar{X}$以下のリスクを持つか？

演習問題 5.2. 正規分布$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$（$\sigma^2$既知） に対して、無情報事前分布$\pi(\mu) \propto 1$の下で以下を示せ：

(a) 事後分布$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\bar{X}, \sigma^2/n)$を導出せよ。

(b) 対応するベイズ推定量（二乗誤差損失）は$\bar{X}$であることを示せ。

(c) $n \to \infty$でベイズリスクの漸近挙動を分析せよ。

演習問題 5.3. ベルヌーイ分布$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(p)$に対して、 Jeffreys事前分布$\pi(p) \propto [p(1-p)]^{-1/2}$の下で以下を行え：

(a) pのベイズ推定量（二乗誤差損失）を導出せよ。

(b) 最尤推定量$\hat{p} = \bar{X}$との差をnの関数として表せ。

(c) $p = 0.01$, $n = 20$のとき、 ベイズ推定量と最尤推定量のリスクを数値的に比較せよ。

演習問題 5.4. $X \sim \mathrm{Bin}(n, p)$で$p \in [0,1]$を二乗誤差損失の下で推定する。

(a) ベイズ推定量$\delta_\pi(X) = (X + a)/(n + a + b)$が 事前$p \sim \mathrm{Beta}(a, b)$に対応することを示せ。

(b) $a = b = \sqrt{n}/2$のとき、 リスク関数$R(p, \delta_\pi)$がpに依存しない（定数リスク） となることを確認し、 定理5.3.2によりδ_πがミニマックスであることを示せ。

発展問題

演習問題 5.5. SURE公式(5.7)を用いて以下を示せ。

$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, I_p)$に対して、 推定量$\hat{\boldsymbol{\mu}}_c = c\boldsymbol{X}$（$c \in [0,1]$）のSUREを計算し、 SUREを最小化する$c^*$が

$$c^* = 1 - \frac{p}{\|\boldsymbol{X}\|^2}$$

ではなく$c^* = \max\left(0,\, 1 - p/\|\boldsymbol{X}\|^2\right)$ と正値部分を取る必要がある理由を説明せよ。 また、この推定量とJames–Stein推定量(5.4)を比較せよ。 *ヒント:* まず固定したcに対するSUREをcの二次式として求めよ。 その上で、制約$c \in [0,1]$の下で最小化すると、 無制約最小値が負になる領域では境界$c = 0$が選ばれることを確認せよ。

演習問題 5.6. Stein の補題（補題5.5.2）の多次元版を述べ、 定理5.5.1の証明における ダイバージェンス$\operatorname{div} \boldsymbol{g}(\boldsymbol{X}) = -(p-2)^2/\|\boldsymbol{X}\|^2$ の導出を詳細に行え（$\partial g_i/\partial X_i$を明示的に計算せよ）。 *ヒント:* $\boldsymbol{g}(\boldsymbol{x}) = -(p-2)\boldsymbol{x}/\|\boldsymbol{x}\|^2$ と置き、 各成分 $g_i(\boldsymbol{x}) = -(p-2)x_i/\|\boldsymbol{x}\|^2$ を偏微分して総和を取れ。

演習問題 5.7. $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, I_p)$に対して、 原点ではなく一般の点$\boldsymbol{\nu} \in \mathbb{R}^p$に向かって縮小する James–Stein推定量

$$\delta^{\mathrm{JS}}_{\boldsymbol{\nu}}(\boldsymbol{X}) = \boldsymbol{\nu} + \left(1 - \frac{p-2}{\|\boldsymbol{X} - \boldsymbol{\nu}\|^2}\right)(\boldsymbol{X} - \boldsymbol{\nu})$$

が、$p \geq 3$のとき最尤推定量$\boldsymbol{X}$を支配することを SURE公式を用いて示せ。 *ヒント:* $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{\nu}$ と置くと、 原点へ縮小する通常の James–Stein 推定量の議論に帰着できる。

計算・実装問題

数値実験を含む問題では、(1) 設定（分布、パラメータ、反復回数、乱数シード）、(2) 作成した図表または表、(3) 比較指標、(4) 結果から読める一言考察、 の4点を答えに含めよ。

演習問題 5.8.（数値実験）$p = 3, 5, 10, 50$の各次元について、 $\|\boldsymbol{\mu}\|^2 = 0, 1, 5, 10, 50, 100$の複数の設定で 最尤推定量、James–Stein推定量（正値部分）、 正則化推定量$c\boldsymbol{X}$（cはSUREで選択）のリスクを シミュレーション（10,000回）で比較せよ。 各設定では、例えば $\boldsymbol{\mu} = (\sqrt{\|\boldsymbol{\mu}\|^2}, 0, \ldots, 0)$ と置いてよい。 結果を表またはグラフにまとめ、 以下を考察せよ：

(a) pが大きくなるほどJames–Stein推定量の改善幅は増加するか？

(b) $\|\boldsymbol{\mu}\|^2$が増加すると改善幅はどう変化するか？

(c) SURE推定リスクと実際のリスクの一致度はどの程度か？

演習問題 5.9.（数値実験）$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$（$n = 5$）に対して、 以下の3つの推定量のリスク関数をシミュレーション（10,000回）で描画し、比較せよ：

(a) 最尤推定量$\hat{\mu} = \bar{X}$

(b) 事前$\mu \sim \mathcal{N}(0, 1)$のベイズ推定量$\delta_1 = n\bar{X}/(n+1)$

(c) 事前$\mu \sim \mathcal{N}(0, 10^2)$のベイズ推定量$\delta_2 = 100n\bar{X}/(100n+1) \approx \bar{X}$

$\mu \in [-5, 5]$の範囲でリスク関数$R(\mu, \delta)$を描画し、 事前分散の選択がリスクのプロファイルにどのような影響を与えるかを考察せよ。 あわせて、最尤推定量のリスクがμに依らず一定であることを図でも確認し、 定理5.3.2で述べた 「定数リスクがミニマックス性と結びつく」という見方を振り返れ。

演習問題 5.10.（数値実験）$X_1, \ldots, X_{20} \overset{\text{i.i.d.}}{\sim} t_3(\mu)$（自由度3の$t$分布、位置パラメータ$\mu$）に対して、 以下の3つの推定量を比較せよ：

(a) 標本平均$\bar{X}$

(b) 標本中央値

(c) 10%トリム平均（上下各10%を除外した平均）

二乗誤差損失と絶対誤差損失のそれぞれについて リスクをシミュレーション（10,000回）で推定し、 $\mu = 0, 1, 3$の場合の結果を表にまとめよ。 損失関数の選択が推定量の評価にどう影響するかを考察し、 外れ値が生じうる場面でどの推定量が望ましいかを議論せよ。

略解の指針

ここでは解法の流れも一段深く書く。証明の細部や数値計算は自分で埋めることを前提とする。

- **演習5.1** 使う道具: バイアス・分散分解。 最初の1手: $\mathbb{E}[\bar{X}] = \theta$, $\mathrm{Var}(\bar{X}) = \theta^2/n$ を先に書く。 途中の要点: $R(\theta, \bar{X}) = \theta^2/n$, $R(\theta, c\bar{X}) = \theta^2\{c^2/n + (c-1)^2\}$ となるので、 c について微分すると最適値は $c^* = n/(n+1)$ である。 最終形: 最小リスクは $\theta^2/(n+1)$ で、$R(\theta, \bar{X}) - R(\theta, c^*\bar{X}) = \theta^2/\{n(n+1)\} > 0$。 したがって $c^*\bar{X}$ は $\bar{X}$ を一様に支配する。

- **演習5.2** 使う道具: 正規尤度と平坦事前の共役計算。 最初の1手: 尤度を μ について平方完成する。 途中の要点: 事後密度は $\exp\{-n(\mu-\bar{X})^2/(2\sigma^2)\}$ に比例するので、 平坦事前の下でも正規核がそのまま残る。 最終形: 事後分布は $\mathcal{N}(\bar{X}, \sigma^2/n)$, 二乗誤差損失のベイズ推定量は事後平均 $\bar{X}$ である。 点ごとのリスクも事後分散も σ^2/n のオーダーで、 $n\to\infty$ では $O(n^{-1})$ で消える。

- **演習5.3** 使う道具: Beta-Bernoulli 共役。 最初の1手: $S=\sum X_i$ と置いて事後分布を $\mathrm{Beta}(S+1/2, n-S+1/2)$ と書く。 途中の要点: ベイズ推定量は事後平均 $(S+1/2)/(n+1)$ であり、 MLE $\bar{X}=S/n$ との差は $(1/2-\bar{X})/(n+1)$ に整理できる。 最終形: リスクは $R(p,\delta_J)=\{np(1-p)+(1/2-p)^2\}/(n+1)^2$。 $p=0.01$, $n=20$ では $R(p,\delta_J)\approx 9.93\times10^{-4}$, $R(p,\bar{X})=p(1-p)/n\approx 4.95\times10^{-4}$ で、 この設定では MLE の方が小さな二乗リスクを持つ。

- **演習5.4** 使う道具: 共役事前とリスク計算。 最初の1手: $\delta_\pi(X)=(X+a)/(n+a+b)$ のバイアスと分散を分ける。 途中の要点: バイアスは $\{a(1-p)-bp\}/(n+a+b)$, 分散は $np(1-p)/(n+a+b)^2$ である。 $a=b=\sqrt{n}/2$ を代入すると $p(1-p)$ と $(1-2p)^2/4$ が合わさって $1/4$ に潰れる。 最終形: $R(p,\delta_\pi)=1/\{4(\sqrt{n}+1)^2\}$ と p に依らない定数リスクになる。 したがって定理5.3.2から δ_π はミニマックスである。

- **演習5.5** 使う道具: SURE 公式。 最初の1手: $g(\boldsymbol{X})=(c-1)\boldsymbol{X}$ と置いて $\mathrm{SURE}(c)=p+(c-1)^2\|\boldsymbol{X}\|^2+2(c-1)p$ を得る。 途中の要点: これは c の上に凸な二次式なので、 無制約最小値は $1-p/\|\boldsymbol{X}\|^2$ になる。 ただし $\|\boldsymbol{X}\|^2<p$ の領域ではこれが負になり、 $c\in[0,1]$ の制約の下では境界 $c=0$ が選ばれる。 最終形: $c^*=\max(0,1-p/\|\boldsymbol{X}\|^2)$。 James–Stein 推定量の縮小係数 $1-(p-2)/\|\boldsymbol{X}\|^2$ と比べると、 こちらの方がしきい値が大きく、やや強く縮小する。

- **演習5.6** 使う道具: 多次元版 Stein の補題。 最初の1手: $g_i(\boldsymbol{x})=-(p-2)x_i/\|\boldsymbol{x}\|^2$ を各成分ごとに偏微分する。 途中の要点: $\partial g_i/\partial x_i=-(p-2)\{\|\boldsymbol{x}\|^2-2x_i^2\}/\|\boldsymbol{x}\|^4$ となるので、総和を取ると $-(p-2)\{p\|\boldsymbol{x}\|^2-2\|\boldsymbol{x}\|^2\}/\|\boldsymbol{x}\|^4$ に整理できる。 最終形: $\sum_i \partial g_i/\partial x_i=-(p-2)^2/\|\boldsymbol{x}\|^2$。 これを Stein の補題に代入すると、James–Stein のリスク改善項が現れる。

- **演習5.7** 使う道具: 平行移動による帰着。 最初の1手: $\boldsymbol{Y}=\boldsymbol{X}-\boldsymbol{\nu}$ と置いて $\boldsymbol{Y}\sim\mathcal{N}(\boldsymbol{\mu}-\boldsymbol{\nu}, I_p)$ を見る。 途中の要点: 二乗誤差損失は平行移動で不変なので、 $\left\|\delta^{\mathrm{JS}}_{\boldsymbol{\nu}}(\boldsymbol{X})-\boldsymbol{\mu}\right\|^2=\left\|\delta^{\mathrm{JS}}(\boldsymbol{Y})-(\boldsymbol{\mu}-\boldsymbol{\nu})\right\|^2$ と書ける。 また SURE のダイバージェンス計算も $\boldsymbol{Y}$ に対してそのまま使える。 最終形: 原点へ縮小する通常の James–Stein 推定量の支配性が $\boldsymbol{\mu}-\boldsymbol{\nu}$ の問題へ移るので、 $\delta^{\mathrm{JS}}_{\boldsymbol{\nu}}$ も $\boldsymbol{X}$ を支配する。

- **演習5.8** 使う道具: モンテカルロによるリスク推定。 最初の1手: 次元 p と $\|\boldsymbol{\mu}\|^2$ の格子を固定し、各点で平均二乗誤差を推定する。 途中の要点: $\boldsymbol{\mu}=(\sqrt{\|\boldsymbol{\mu}\|^2},0,\dots,0)$ と置けば 方向を固定したまま信号強度だけを動かせる。 各設定で $\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2$ の平均と、 SURE が返す推定リスクの平均を並べて比較する。 最終形: 改善幅は一般に p が大きいほど見えやすく、 $\|\boldsymbol{\mu}\|^2$ が大きくなると縮小の利得は小さくなる。 SURE は平均的には実リスクをよく追うが、 単発の実現値ではばらつきが残る。

- **演習5.9** 使う道具: バイアス・分散分解。 最初の1手: 正規事前分散を τ^2 とすると、 推定量は $\delta_\tau=a_\tau\bar{X}$, $a_\tau=n\tau^2/(n\tau^2+1)$ と書ける。 途中の要点: $n=5$ では $a_1=5/6$,

$a_{100} = 500/501$ であり、 前者は目に見える縮小、後者はほぼ MLE である。 理論リスク $R(\mu, \delta_\tau) = a_\tau^2/n + (1-a_\tau)^2\mu^2$ を同時に描くとシミュレーション結果の形を説明しやすい。 最終形: 強い縮小は $\mu \approx 0$ で有利だが、 $|\mu|$ が大きくなるとバイアスが支配的になる。 拡散事前のベイズ推定量はほぼ一定リスクの MLE に重なる。

- **演習5.10** 使う道具: 損失関数ごとの経験リスク比較。 最初の1手: 各推定量について二乗誤差と絶対誤差を別々に平均する。 途中の要点: $t_3(\mu)$ は対称だが重尾なので、 標本平均は少数の大きな外れ値に引かれやすい。 一方、中央値とトリム平均は中心位置には素直だが、完全に対称な場面では少し分散を犠牲にする。 最終形: 二乗誤差ではトリム平均や中央値が標本平均より安定しやすく、 絶対誤差では母集団中央値に対応する標本中央値が最も有利になりやすい。 損失関数を変えると「よい推定量」の順位も変わることが確認点である。

参考文献ノート

最初に読むなら、 Berger か Casella and Berger で古典的な背骨を押さえたのち、 Efron や FDA (2026), Barter and Yu (2024) に進むと、 本章の考え方が大規模推論や現代実務にどう伸びるかを追いやすい。

- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer. ——決定理論の包括的な教科書。許容性と完備クラス定理の詳細な証明を含む。
- Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd ed. Duxbury. ——第7章が決定理論を扱う。本章と同等の範囲をカバー。
- Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer. ——第4–5章でベイズ推定とミニマックス推定を厳密に展開。
- James, W. and Stein, C. (1961). "Estimation with quadratic loss." *Proceedings of the Fourth Berkeley Symposium*, 1, 361–379. ——Stein のパラドックスの原論文。
- Stein, C. M. (1981). "Estimation of the mean of a multivariate normal distribution." *Annals of Statistics*, 9, 1135–1151. ——SURE（不偏リスク推定）を導入した論文。
- Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press. ——経験ベイズの現代的展開。James–Steinと大規模推論の関係を詳述。
- Efron, B. and Morris, C. (1973). "Stein's estimation rule and its competitors: an empirical Bayes approach." *Journal of the American Statistical Association*, 68, 117–130. ——James–Stein推定量の経験ベイズ的解釈を与えた古典的論文。
- FDA (2026). *Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products*. ——ベイズ法が規制実務でも、適応的設計や主要推論を含む現在進行形の方法であることを示す文書。

- Barter, R. and Yu, B. (2024). *Veridical Data Science*. MIT Press. ——予測可能性・計算可能性・安定性という観点から、 実務で「信頼できる手続き」をどう選ぶかを論じる。 決定理論の思想を現代のデータ分析に引き寄せて読む補助線として有益である。

第6章

漸近理論

問いと学習目標

この章で答える問い

- 最尤推定量は標本サイズが大きくなると真の値に近づくか。どのような速度と分布で近づくか。
- 尤度比検定・Wald 検定・スコア検定は、大標本でどのような意味で等価になるのか。
- M推定量や経験分布関数のような、より一般の統計量にも同じ考え方を広げられるか。
- 漸近近似は便利だが、どのような場面で有限標本シミュレーションによる確認が必要か。

読み終えたらできるようになること

1. MLE の一致性・漸近正規性・漸近有効性を、証明の骨格とともに説明できる。
2. デルタ法とスラツキー型の議論を使って、新しい統計量の漸近分布を導出できる。
3. 三大検定の漸近等価性を理解し、近似の使い分けを判断できる。
4. M推定量、サンドイッチ推定量、経験過程、LAN 条件の役割を大まかに整理できる。
5. 有限標本では何を追加で検証すべきかを、シミュレーションで確認できる。

直観的理解

統計的推論では、有限個のデータから母集団の性質を推測する。 しかし、推定量の正確な分布を求めることは多くの場合困難である。 漸近理論は、「標本サイズnが十分に大きいとき、推定量や検定統計量はどのように振る舞うか」を解明する。

実務ポイント

漸近理論は強力だが、 「n が大きいから大丈夫」という合言葉ではない。 離散デー

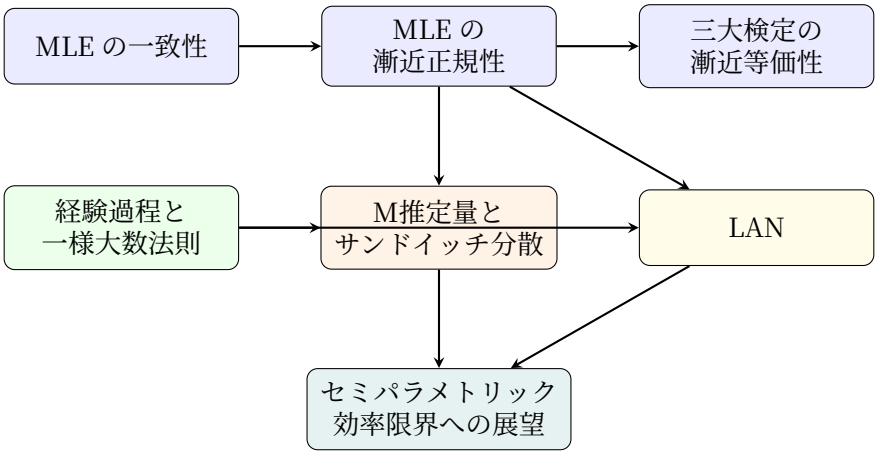


図 6.1: 本章の論理地図。MLE の極限定理を起点に、三大検定・M推定量・経験過程・LAN へと話が広がり、最後にセミパラメトリック理論へ接続する。

表 6.1: 漸近理論の主要道具とその出口

道具	本章で押さえる役割	後続章・実務での出口
MLE の漸近正規性	推定誤差の $1/\sqrt{n}$ スケールと情報量による分散を与える	Wald 区間, 尤度比検定, GLM の標準誤差, ラプラス近似の基礎になる
M推定量とサンドイッチ分散	モデル誤特定下でも使える一般的な正規近似を与える	ロバスト回帰, 一般化推定方程式, 因果推論のロバスト標準誤差へ接続する
経験過程と一様大数法則	関数全体の揺らぎと一様収束を扱う	Kolmogorov–Smirnov 検定, ノンパラメトリック推論, 統計的学習理論の基盤になる
LAN と効率限界	局所的に見た最適推定の下界を与える	セミパラメトリック効率, 影響関数, DML の理論的土台になる

タ、強い歪み、重い裾、高次元、境界に近いパラメータでは、 正規近似が遅かったり壊れたりする。 そのため本章では、定理そのものだけでなく、 Berry–Esseen 型の誤差評価やシミュレーション検証を合わせて読むことが重要である。

6.1　MLEの漸近理論

第3章では最尤推定量（MLE）の構成法と有限標本での性質を議論した。 本節では、標本サイズ$n \to \infty$の下でのMLEの振る舞いを三つの段階――一致性、漸近正規性、漸近有効性――で明らかにする。

6.1.1　一致性

定義 6.1.1 (一致推定量)**.** 推定量列$\{\hat{\theta}_n\}$がθ_0の**一致推定量**（consistent estimator）であるとは、

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \quad (n \to \infty)$$

が成り立つことをいう。すなわち、任意の$\varepsilon > 0$に対して

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\hat{\theta}_n - \theta_0\right| > \varepsilon\right) = 0$$

である。

読み下し

一致性の式$\hat{\theta}_n \xrightarrow{p} \theta_0$は、「標本サイズ$n$が大きくなるにつれ、推定量$\hat{\theta}_n$は 真のパラメータ$\theta_0$に確率的に近づく」と読む。 これは推定法の最低限の合理性を保証する： データを増やせば増やすほど、推定は正確になる。

なぜMLEが一致推定量であるかを直観的に理解するために、まず$n = 2$の小さな例から始めよう。

例 6.1.2 (ベルヌーイ分布のMLE)**.** $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(p_0)$とする。 MLEは$\hat{p}_n = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$である。

$n = 2$の場合、$\hat{p}_2$は$\{0, 0.5, 1\}$の3値しか取れず、p_0からのずれは大きい。 しかし大数の法則（第2章）により$\bar{X} \xrightarrow{p} p_0$であるから、 nが大きくなるにつれ$\hat{p}_n$はp_0に集中する。

一般のMLEについても、次の定理により一致性が保証される。

定理 6.1.3 (MLEの一致性)**.** *以下の正則条件を満たすとき、最尤推定量$\hat{\theta}_n$はθ_0の一致推定量である：*

1. *パラメータ空間Θはコンパクト*
2. *対数尤度$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta)$は$\theta$に関して連続（概確実に）*
3. **識別可能性**：*すべての$\theta \neq \theta_0$に対して$\mathbb{E}_{\theta_0}[\log f(X_1;\theta)] < \mathbb{E}_{\theta_0}[\log f(X_1;\theta_0)]$*
4. **一様収束**：$\sup_{\theta\in\Theta} |\ell_n(\theta) - \ell(\theta)| \xrightarrow{p} 0$、 ここで$\ell(\theta) = \mathbb{E}_{\theta_0}[\log f(X_1;\theta)]$

読み下し

条件3は、Kullback-Leibler情報量$\mathrm{KL}(f_{\theta_0} \| f_\theta) > 0$（$\theta \neq \theta_0$）と同値である。 「尤度を最大にするパラメータは真のパラメータに限る」という意味であり、 モデルが識別可能であることの帰結である。
条件4は、有限標本の対数尤度がその期待値に一様に近づくことを要求する。 パラメータ空間がコンパクトならば、大数の法則と連続性から通常成り立つ。

証明の方針. 証明は「最大化する対象が極限と入れ替えられる」ことに基づく。

ステップ1（極限の最大化）： 条件3（KL情報量の正値性）により、極限関数$\ell(\theta) = \mathbb{E}_{\theta_0}[\log f(X_1;\theta)]$は$\theta = \theta_0$で一意的に最大化される。

ステップ2（一様収束による近似）：条件4により$\ell_n(\theta) \approx \ell(\theta)$が$\theta$について一様に成立する。したがって$\ell_n$の最大点$\hat{\theta}_n$は$\ell$の最大点$\theta_0$に近くなければならない。

ステップ3（形式化）：任意の$\varepsilon > 0$に対して、Θのコンパクト性とℓの連続性から$\sup_{|\theta-\theta_0|\geq\varepsilon} \ell(\theta) < \ell(\theta_0)$が成り立つ。一様収束と合わせると、十分大きな$n$では$\ell_n(\hat{\theta}_n) > \ell_n(\theta)$（$|\theta-\theta_0| \geq \varepsilon$のすべての$\theta$に対して）が高確率で成り立ち、$\left|\hat{\theta}_n - \theta_0\right| < \varepsilon$を得る。　□

直観的理解

一致性の証明の核心は単純である。対数尤度の期待値$\ell(\theta)$は、KL情報量の性質によりθ_0で最大になる。大数の法則が「期待値への収束」を保証するので、経験的な対数尤度$\ell_n(\theta)$もθ_0の近くで最大となる。つまり、「KL情報量が真のパラメータを見分ける力を持ち、大数の法則がその力を実現する」のである。

6.1.2　漸近正規性

一致性は「近づく」ことを保証するが、「どのくらいの精度で近づくか」は教えてくれない。漸近正規性は、MLEの分布が大標本で正規分布に近づくことを示し、精度を定量化する。

まず、漸近正規性の鍵となるフィッシャー情報量を定義する。

定義 6.1.4 (フィッシャー情報量)**.** スコア関数$s(\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta)$に対して、**フィッシャー情報量**は

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[s(\theta)^2\right] = \mathrm{Var}_\theta[s(\theta)] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

で定義される。多次元パラメータ$\theta \in \mathbb{R}^p$の場合は$p \times p$の**フィッシャー情報行列**となり、(i,j)成分は$[\mathcal{I}(\theta)]_{ij} = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X;\theta)\right]$である。

読み下し

フィッシャー情報量$\mathcal{I}(\theta)$は、「データ1個あたり、パラメータθについてどれだけの情報を持つか」を測る量である。$\mathcal{I}(\theta)$が大きいほど、対数尤度の曲率が急であり、MLEの推定精度が高くなる。等号$\mathbb{E}[s(\theta)^2] = -\mathbb{E}[\partial^2\log f/\partial\theta^2]$は、「スコアのばらつきの大きさ」と「対数尤度の曲がり具合」が等しいことを述べている。

定理 6.1.5 (MLEの漸近正規性)**.** *正則条件の下で、最尤推定量$\hat{\theta}_n$は次のように分布収束する：*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0,\ \mathcal{I}(\theta_0)^{-1}\right)$$

多次元パラメータ$\theta \in \mathbb{R}^p$の場合は$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}_p(0,\ \mathcal{I}(\theta_0)^{-1})$である。

読み下し

この式は「MLEの推定誤差$\hat{\theta}_n - \theta_0$を$\sqrt{n}$倍で拡大すると、分散$\mathcal{I}(\theta_0)^{-1}$の正規分布に収束する」と読む。言い換えれば、MLEの推定誤差はおよそ$1/\sqrt{n}$のオーダーであり、その散らばりはフィッシャー情報量の逆数で決まる。情報量が大きい（対数尤度の曲

率が急な）問題ほど、推定は精密になる。

証明の方針（1次元の場合）. 証明は、対数尤度のTaylor展開と中心極限定理の組み合わせによる。

ステップ1（一階条件）：$\hat{\theta}_n$は平均対数尤度$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i;\theta)$を最大化するため、$\ell_n'(\hat{\theta}_n) = 0$を満たす。

ステップ2（Taylor展開）：θ_0の周りで一階条件を展開する：

$$0 = \ell_n'(\hat{\theta}_n) \approx \ell_n'(\theta_0) + \ell_n''(\theta_0)\left(\hat{\theta}_n - \theta_0\right) \tag{6.1}$$

整理すると、

$$\hat{\theta}_n - \theta_0 \approx -\left[\ell_n''(\theta_0)\right]^{-1}\ell_n'(\theta_0)$$

ステップ3（各項の漸近挙動）：

- スコアの平均：$\sqrt{n}\,\ell_n'(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i;\theta_0)$。$\mathbb{E}[s(X_i;\theta_0)] = 0$であるから、中心極限定理（第2章）により$\sqrt{n}\,\ell_n'(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0))$。
- ヘッセ行列の平均：$\ell_n''(\theta_0) = \frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta_0)$。大数の法則により$\ell_n''(\theta_0) \xrightarrow{p} -\mathcal{I}(\theta_0)$。

ステップ4（Slutskyの定理による結合）：ステップ2の両辺に$\sqrt{n}$を掛け、ステップ3の結果を代入すると、

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \mathcal{I}(\theta_0)^{-1}\cdot\sqrt{n}\,\ell_n'(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$$

が得られる。最後の等号ではSlutskyの定理（第2章）を用いた。 □

例 6.1.6 (正規分布のMLE). $X_1,\dots,X_n$ i.i.d. $\mathcal{N}(\mu,\sigma^2)$に対して、MLEは$\hat{\mu}_n = \bar{X}$、$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n(X_i - \bar{X})^2$である。

フィッシャー情報行列は

$$\mathcal{I}(\mu,\sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$$

したがって、漸近分布は

$$\sqrt{n}\begin{pmatrix}\hat{\mu}_n - \mu \\ \hat{\sigma}_n^2 - \sigma^2\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\ \begin{pmatrix}\sigma^2 & 0 \\ 0 & 2\sigma^4\end{pmatrix}\right)$$

$\hat{\mu}_n = \bar{X}$の場合、漸近分布$\mathcal{N}(\mu,\sigma^2/n)$は実は正確な有限標本分布と一致する。しかし$\hat{\sigma}_n^2$の場合、正確な分布は$\sigma^2\chi^2_{n-1}/n$であり、漸近正規近似は$n$が小さいときには精度が落ちる。

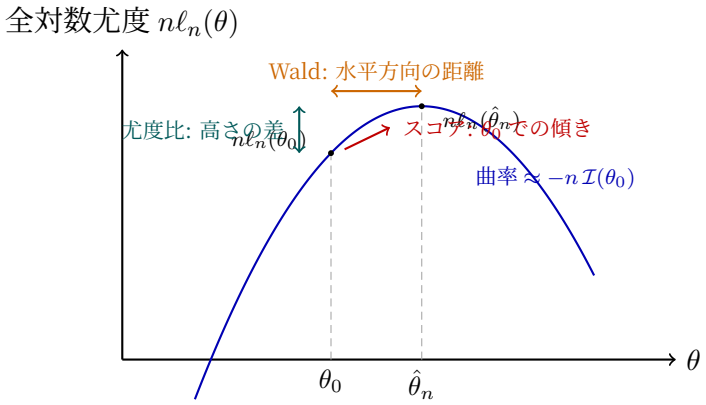


図 6.2: 対数尤度の局所二次近似の見取り図。描かれているのは全対数尤度 $n\ell_n(\theta)$ である。Wald 検定は θ_0 と $\hat{\theta}_n$ の距離、尤度比検定は対数尤度の高さの差、スコア検定は θ_0 での傾きに対応する。三大検定と LAN は同じ局所二次近似から生まれる。

実務ポイント

漸近正規性により、大標本ではMLEの近似的な信頼区間を

$$\hat{\theta}_n \pm z_{\alpha/2}\sqrt{\frac{1}{n\hat{\mathcal{I}}_n}}$$

と構成できる（$\hat{\mathcal{I}}_n$は情報量の推定値、$z_{\alpha/2}$は標準正規分布の分位点）。 これがWald型信頼区間であり、第4章で述べたWald検定と表裏一体の関係にある。
ただし、「大標本」がどの程度のnを意味するかは問題に依存する。 裾の重い分布、境界付近のパラメータ（例：$p \approx 0$のベルヌーイ）、 曲率の強い問題では、漸近近似の収束が遅い。 有限標本での性能は本章末の数値実験で検証する。

6.1.3　漸近有効性

MLEの漸近分散が$\mathcal{I}(\theta_0)^{-1}$であることを見たが、 これはどの程度「良い」値なのか。Cramér–Rao不等式がその答えを与える。

定理 6.1.7 (Cramér–Rao不等式（情報不等式）). *正則条件の下で、θ_0の任意の不偏推定量$\tilde{\theta}_n$に対して*

$$\mathrm{Var}_{\theta_0}(\tilde{\theta}_n) \geq \frac{1}{n\,\mathcal{I}(\theta_0)}$$

*が成り立つ。この下界$1/(n\,\mathcal{I}(\theta_0))$を***Cramér–Rao下界***と呼ぶ。*

読み下し

Cramér–Rao不等式は「どれほど巧みに推定量を設計しても、 不偏推定量の分散は$1/(n\,\mathcal{I}(\theta_0))$より小さくできない」と読む。 フィッシャー情報量が大きいほど下界は小さく、精密な推定が可能になる。

定義 6.1.8 (漸近有効性)**.** 推定量$\hat{\theta}_n$が**漸近有効**であるとは、

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0,\ \mathcal{I}(\theta_0)^{-1})$$

が成り立つことをいう。すなわち、漸近分散がCramér–Rao下界に一致する。

読み下し

漸近有効性は、「大標本の意味で、これ以上精度の良い（正則な）推定量は存在しない」ことを意味する。 定理6.1.5と合わせると、 **MLEは漸近有効である**。 これが、統計学においてMLEが標準的な推定法として広く用いられる理論的根拠である。

直観的理解

MLEの漸近理論は三つの主張から成る：

1. **一致性**：標本を増やせば真の値に近づく（基本的な合理性）
2. **漸近正規性**：推定誤差は正規分布で近似でき、$1/\sqrt{n}$の速度で縮小する
3. **漸近有効性**：大標本で達成可能な最小の分散を実現する（最適性）

ただし、これらはすべて「$n \to \infty$」の結果であることに注意が必要である。 有限標本では、MLEが最適とは限らない——第5章のJames–Stein推定量はその好例であった。

6.1.4　デルタ法の応用

実務では、パラメータそのものではなく、その関数$g(\theta)$に興味がある場合が多い。 第2章で導入したデルタ法を漸近正規性と組み合わせると、 変換パラメータの漸近分布が直ちに得られる。

定理 6.1.9 (デルタ法（再掲）)**.** $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$*が成り立ち、* g*が*θ_0*で微分可能で*$g'(\theta_0) \neq 0$*ならば、*

$$\sqrt{n}\Big(g(\hat{\theta}_n) - g(\theta_0)\Big) \xrightarrow{d} \mathcal{N}\big(0,\ [g'(\theta_0)]^2\,\sigma^2\big)$$

読み下し

デルタ法は「推定量に滑らかな変換gを施しても漸近正規性は保たれ、 漸近分散はgの導関数で調整される」と読む。 MLEと組み合わせれば、$g(\hat{\theta}_n)$の漸近分散は$[g'(\theta_0)]^2/[n\,\mathcal{I}(\theta_0)]$となる。

例 6.1.10 (指数分布の平均のMLE)**.** $X_1, \dots, X_n$ i.i.d. $\mathrm{Exp}(\lambda)$（rate parameterization、密度$\lambda e^{-\lambda x}$）とする。 MLEは$\hat{\lambda}_n = 1/\bar{X}$、フィッシャー情報量は$\mathcal{I}(\lambda) = 1/\lambda^2$、 漸近分布は$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda^2)$である。

平均$\mu = 1/\lambda$に興味がある場合、$g(\lambda) = 1/\lambda$として$g'(\lambda) = -1/\lambda^2$であるから、

$$\sqrt{n}\left(\frac{1}{\hat{\lambda}_n} - \frac{1}{\lambda}\right) = \sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{1}{\lambda^2}\right) = \mathcal{N}(0,\ \mu^2)$$

これは中心極限定理から直接得られる結果と一致する（$\mathrm{Var}(X_i) = 1/\lambda^2 = \mu^2$）。

例 6.1.11 (対数オッズ比の漸近分布)**.** $X_1, \dots, X_n$i.i.d.Bern(p)に対して、対数オッズ$\eta = \log \frac{p}{1-p}$の推定を考える。$g(p) = \log \frac{p}{1-p}$とすると$g'(p) = \frac{1}{p(1-p)}$であり、MLEの漸近分散は$1/\mathcal{I}(p) = p(1-p)$であるから、デルタ法より

$$\sqrt{n}(\hat{\eta}_n - \eta) \xrightarrow{d} \mathcal{N}\big(0,\ [g'(p)]^2 \cdot p(1-p)\big) = \mathcal{N}\left(0,\ \frac{1}{[p(1-p)]^2} \cdot p(1-p)\right) = \mathcal{N}\left(0,\ \frac{1}{p(1-p)}\right)$$

これは、対数オッズのスケールでの信頼区間構成の理論的根拠となる。

6.2　三大検定の漸近等価性

第4章で導入した尤度比検定、Wald検定、スコア検定は、有限標本では一般に異なる結論を与える。しかし大標本では、これらは漸近的に等価になる。本節では、MLEの漸近正規性を用いてこの等価性を示す。

定理 6.2.1 (三大検定の漸近等価性)**.** *帰無仮説$H_0 : \theta = \theta_0$（1次元の場合）の下で、以下の三統計量はいずれも漸近的にχ^2_1分布に従い、互いに$o_p(1)$の差しか持たない：*

1. **尤度比統計量**：$2n\big[\ell_n(\hat{\theta}_n) - \ell_n(\theta_0)\big]$
2. **Wald統計量**：$n(\hat{\theta}_n - \theta_0)^2\, \mathcal{I}(\hat{\theta}_n)$
3. **スコア統計量**：$n\big[\ell_n'(\theta_0)\big]^2 / \mathcal{I}(\theta_0)$

証明の骨格. **鍵となる考え方**：対数尤度をθ_0の周りで2次までTaylor展開し、MLEの漸近正規性と情報行列の一致推定量を代入すると、三つの統計量が同じ2次形式に帰着する。

$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i;\theta)$として、$\hat{\theta}_n$の周りで展開すると

$$\ell_n(\theta_0) \approx \ell_n(\hat{\theta}_n) + \frac{1}{2}\ell_n''(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)^2$$

（$\hat{\theta}_n$で一階微分は0）。$-\ell_n''(\hat{\theta}_n) \approx \mathcal{I}(\hat{\theta}_n)$（大数の法則）を用いると、

$$2n[\ell_n(\hat{\theta}_n) - \ell_n(\theta_0)] \approx n(\hat{\theta}_n - \theta_0)^2\, \mathcal{I}(\hat{\theta}_n) = W$$

が得られる。スコア統計量も同様に、式(6.1)の関係からWと漸近的に一致する。 □

実務ポイント

三つの検定は漸近的に等価だが、有限標本では振る舞いが異なる。実務での使い分けの指針：

- **尤度比検定**：最も安定。制約なし・制約ありの両方のMLEが必要
- **Wald検定**：制約なしのMLEのみで計算可能。パラメータ化に依存する弱点がある
- **スコア検定**：帰無仮説下のMLEのみで計算可能。モデルが複雑な場合に有利

小標本ではWald検定が他より保守的（棄却しにくい）になりやすいことが知られている。

6.3　M推定量の漸近理論

MLEは対数尤度を最大化する推定量であるが、これを一般の目的関数に拡張したのがM推定量（Maximum-likelihood type estimator）である。ロバスト推定や分位点回帰など、多くの重要な推定法がこの枠組みに含まれる。

6.3.1　M推定量の定義と一致性

定義 6.3.1 (M推定量)**.** 目的関数$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta)$を最小化する推定量

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \; M_n(\theta)$$

を**M推定量**と呼ぶ。

読み下し

M推定量は「n個のデータそれぞれについて損失$m(X_i, \theta)$を計算し、その平均を最小にするθを推定値とする」と読む。$m(x,\theta) = -\log f(x;\theta)$（負の対数尤度）とすれば最尤推定量、$m(x,\theta) = (x-\theta)^2$とすれば標本平均、$m(x,\theta) = |x-\theta|$とすれば標本中央値が得られる。

定理 6.3.2 (M推定量の一致性)**.** *次の条件の下で*$\hat{\theta}_n \xrightarrow{p} \theta_0$*が成り立つ：*

1. $m(x,\theta)$*が*θ*に関して連続（概確実に）*
2. *極限目的関数*$M(\theta) = \mathbb{E}[m(X,\theta)]$*が*$\theta_0$*で一意的に最小化される*
3. *一様収束：*$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$

読み下し

証明の構造はMLE一致性（定理6.1.3）と全く同じである。一様大数法則により$M_n \approx M$が成り立ち、Mの一意最小点がθ_0であるから、M_nの最小点$\hat{\theta}_n$もθ_0に近づく。

6.3.2　M推定量の漸近正規性とサンドイッチ推定量

定理 6.3.3 (M推定量の漸近正規性)**.** $m(x,\theta)$*が*θ*について2回微分可能で、追加の正則条件が満たされるとき、*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\Big(0,\; H^{-1}\Sigma\, H^{-\top}\Big)$$

ここで

$$H = \mathbb{E}\big[\nabla_\theta^2 m(X,\theta_0)\big] \qquad \text{（ヘッセ行列の期待値）} \tag{6.2}$$

$$\Sigma = \mathbb{E}\Big[\nabla_\theta m(X,\theta_0)\,\nabla_\theta m(X,\theta_0)^\top\Big] \qquad \text{（勾配の外積の期待値）} \tag{6.3}$$

行列$V = H^{-1}\Sigma H^{-\top}$*を***サンドイッチ分散行列***と呼ぶ。*

読み下し

サンドイッチ分散行列$V = H^{-1}\Sigma H^{-\top}$は、ヘッセ行列$H$が「パン」、勾配の外積$\Sigma$が「具」の役割を果たすことからこの名がある。
MLEの特殊な場合、$H = \Sigma = \mathcal{I}(\theta_0)$が成り立つため、$V = \mathcal{I}(\theta_0)^{-1}$に簡略化される（情報行列等式）。しかしモデルが誤指定されている場合やロバスト推定量の場合は$H \neq \Sigma$となり、サンドイッチ推定量が必要になる。

証明の方針. MLE漸近正規性の証明（定理6.1.5）と同様のTaylor展開による。一階条件$\nabla_\theta M_n(\hat\theta_n) = 0$を$\theta_0$の周りで展開すると

$$0 \approx \nabla_\theta M_n(\theta_0) + \nabla_\theta^2 M_n(\theta_0)\,(\hat\theta_n - \theta_0)$$

整理して$\sqrt{n}$を掛けると、

$$\sqrt{n}(\hat\theta_n - \theta_0) \approx -\big[\nabla_\theta^2 M_n(\theta_0)\big]^{-1}\sqrt{n}\,\nabla_\theta M_n(\theta_0)$$

$\nabla_\theta^2 M_n(\theta_0) \xrightarrow{p} H$（大数の法則）、$\sqrt{n}\,\nabla_\theta M_n(\theta_0) \xrightarrow{d} \mathcal{N}(0,\Sigma)$（中心極限定理）を用いれば結論が得られる。 □

例 6.3.4 (Huberのロバスト推定量)**.** $Y_i = \mu + \varepsilon_i$で$\varepsilon_i$に外れ値が含まれる場合、Huber損失

$$\rho_k(u) = \begin{cases} u^2/2 & \text{if } |u| \le k \\ k\,|u| - k^2/2 & \text{if } |u| > k \end{cases}$$

に基づくM推定量$\hat\mu_H = \arg\min_\mu \sum_i \rho_k(Y_i - \mu)$は、二乗損失（平均）と絶対損失（中央値）の中間的な性質を持ち、外れ値に対して頑健である。

この推定量の漸近分散は$V = H^{-1}\Sigma H^{-\top}$の形で与えられ、正規分布からのずれが大きいほど、MLEに対する相対効率が向上する。

実務ポイント

サンドイッチ分散行列$V = H^{-1}\Sigma H^{-\top}$の推定量は、$H$と$\Sigma$をそれぞれ標本類似量で置き換えて得られる：

$$\hat V = \hat H_n^{-1}\,\hat\Sigma_n\,\hat H_n^{-\top}, \quad \hat H_n = \frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 m(X_i,\hat\theta_n), \quad \hat\Sigma_n = \frac{1}{n}\sum_{i=1}^n \nabla_\theta m(X_i,\hat\theta_n)\,\nabla_\theta m(X_i,\hat\theta_n)^\top$$

これは回帰分析におけるHuber–White標準誤差（HC標準誤差）として広く使われている。 モデルの正しさに依存しない標準誤差推定を提供するため、 誤差分布の仮定に自信がない場合のデフォルトの選択肢である。

6.4　経験過程の基礎

ここまでの漸近理論は、パラメトリックモデルにおける「特定のパラメータの推定量」に焦点を当てていた。 経験過程の理論は、視点を「関数全体の収束」に広げる。 これはノンパラメトリック推論や高次元統計の理論的基盤となる。

6.4.1　経験分布関数とGlivenko–Cantelli定理

定義 6.4.1 (経験分布関数)**.** $X_1, \dots, X_n$を累積分布関数Fを持つi.i.d.確率変数列とするとき、**経験分布関数**は

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \le x)$$

で定義される。各xについて、$F_n(x)$は$X_i \le x$である観測値の割合である。

読み下し

経験分布関数$F_n(x)$は、 「n個のデータのうちx以下のものの割合」を表す。 $\mathbf{1}(X_i \le x)$は$X_i \le x$のとき1、それ以外のとき0をとる指示関数である。 各xを固定すると、$F_n(x)$はBern$(F(x))$の標本平均であるから、 大数の法則により$F_n(x) \xrightarrow{p} F(x)$が成り立つ。 問題は、この収束が$x$について一様に成り立つかどうかである。

定理 6.4.2 (Glivenko–Cantelli定理)**.**

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0 \quad (n \to \infty)$$

*すなわち、経験分布関数は真の分布関数に***一様に***概収束する。*

読み下し

Glivenko–Cantelli定理は「統計学の基本定理」とも呼ばれる。 $\sup_x$を取っているため、F_nはFのどこでも一様に近づく。 これにより、F_nから導かれる多くの統計量（中央値、四分位数、Kolmogorov–Smirnov検定統計量など）の 一致性を保証できる。

証明の方針. 各xについて$F_n(x) \xrightarrow{p} F(x)$は大数の法則から直ちに従う。 一様性の証明は、$\mathbb{R}$を有限個の区間$[x_0, x_1], [x_1, x_2], \dots, [x_{K-1}, x_K]$に分割し、 Fの変動が各区間でε以下になるようにKを選ぶ。 各分割点で大数の法則を適用し、分割点間の変動はF_nの単調性で制御する。 $\varepsilon \to 0$と$n \to \infty$の極限を適切に取ることで結論が得られる。 □

6.4.2　経験過程とDonskerの定理

Glivenko–Cantelli定理は$F_n - F \to 0$を述べるが、 収束の速さ（偏差の大きさ）については何も教えない。 経験過程はこの問題を扱う。

定義 6.4.3 (経験過程)**.**

$$\mathbb{G}_n(x) = \sqrt{n}\big[F_n(x) - F(x)\big]$$

これは、経験分布と真の分布の偏差を$\sqrt{n}$で拡大した確率過程である。

読み下し

$\mathbb{G}_n(x)$は「経験分布が真の分布からどの程度ずれているかを、 $\sqrt{n}$倍に拡大して観察する」ものと読む。 各xを固定すると$\mathbb{G}_n(x) = \sqrt{n}[\bar{Z}_x - F(x)]$（$Z_{i,x} = \mathbf{1}(X_i \le x)$）であり、 中心極限定理から$\mathbb{G}_n(x) \xrightarrow{d} \mathcal{N}(0, F(x)(1-F(x)))$が成り立つ。 Donskerの定理は、この収束が関数空間での収束に拡張されることを述べる。

定理 6.4.4 (Donskerの定理)**.**

$$\mathbb{G}_n \xrightarrow{d} \mathbb{G} \qquad \textit{（Skorokhod空間}D[-\infty,\infty]\textit{での弱収束）}$$

*ここで$\mathbb{G}$は***ブラウン橋***（Brownian bridge）であり、平均0のガウス過程で*

$$\mathrm{Cov}(\mathbb{G}(x), \mathbb{G}(y)) = F(x \wedge y) - F(x)F(y)$$

を満たす。ここで$x \wedge y = \min(x, y)$である。

読み下し

Donskerの定理は、経験過程$\mathbb{G}_n$全体がブラウン橋$\mathbb{G}$に弱収束することを述べる。 中心極限定理が「$\sqrt{n}(\bar{X} - \mu)$が正規分布に収束する」点の結果であるのに対し、 Donskerの定理は関数としての収束——**関数中心極限定理**——を与える。
この結果から、$\sup_x |\mathbb{G}_n(x)|$のような経験過程の汎関数の分布も ブラウン橋の対応する汎関数の分布に収束し、 Kolmogorov–Smirnov検定などの理論的基礎が得られる。

直観的理解

Glivenko–Cantelli定理とDonskerの定理の関係は、 大数の法則と中心極限定理の関係に対応する。 対応関係は表6.2の通りである。 つまり、経験過程の理論は、古典的な極限定理を関数空間に持ち上げたものである。

例 6.4.5 (Kolmogorov–Smirnov検定)**.** 帰無仮説$H_0 : F = F_0$の下で、検定統計量

$$D_n = \sup_x |F_n(x) - F_0(x)| = \frac{1}{\sqrt{n}} \sup_x |\mathbb{G}_n(x)|$$

を用いる。Donskerの定理により$\sqrt{n}\, D_n \xrightarrow{d} \sup_x |\mathbb{G}(x)|$であり、 右辺の分布はKolmogorov分布と呼ばれ、F_0の形に依存しない（分布自由性）。 これがKolmogorov–Smirnov検定の

表 6.2: 古典的極限定理と経験過程の対応

観点	点の結果	関数の結果	実務での役割
平均的な収束	大数の法則	Glivenko–Cantelli 定理	経験分布や目的関数が一様に安定することを保証する
$\sqrt{n}$ スケールの揺らぎ	中心極限定理	Donsker の定理	Kolmogorov–Smirnov 検定や汎関数の近似分布を与える

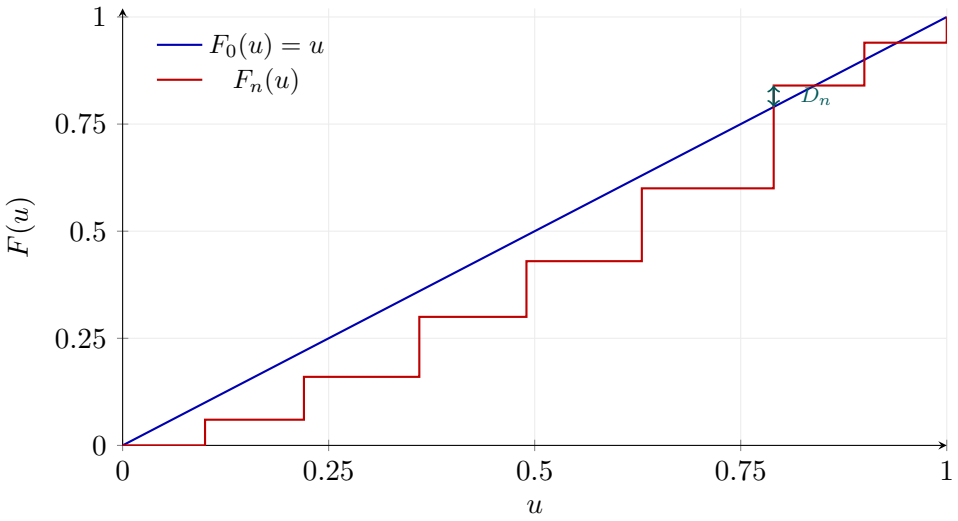


図 6.3: Kolmogorov–Smirnov 統計量は、帰無分布 F_0 と経験分布 F_n の最大縦距離として読める。確率積分変換により帰無分布を一様分布へ移せるため、極限分布が F_0 に依存しないことが見えやすい。

理論的根拠である。

例 6.4.6 (標本中央値の漸近分布)**.** Fが中央値mにおいて連続微分可能で$f(m) = F'(m) > 0$とする。 経験過程の理論から、標本中央値$\hat{m}_n$の漸近分布は

$$\sqrt{n}(\hat{m}_n - m) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{1}{4f(m)^2}\right)$$

で与えられる。正規分布$\mathcal{N}(\mu, \sigma^2)$の場合、$f(\mu) = 1/(\sigma\sqrt{2\pi})$であるから $1/(4f(\mu)^2) = \pi\sigma^2/2$となり、$\mathrm{Var}(\hat{m}_n) \approx \pi\sigma^2/(2n)$である。 標本平均の分散$\sigma^2/n$と比較すると$\pi/2 \approx 1.57$倍大きい。すなわち、正規分布では中央値はMLEの約64%の効率しか持たない。 しかし、裾の重い分布（例：コーシー分布）では中央値の方が効率的になりうる。

6.4.3　一様大数法則とVC次元

Glivenko–Cantelli定理を指示関数$\mathbf{1}(X \le x)$の族から一般の関数族に拡張したのが、 一様大数法則である。

定理 6.4.7 (一様大数法則)**.** *関数族$\mathcal{F} = \{f_\theta : \theta \in \Theta\}$が***Glivenko–Cantelliクラス***であるとき、*

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i) - \mathbb{E}[f_\theta(X)]\right| \xrightarrow{p} 0$$

特に、VC次元が有限な関数族はGlivenko–Cantelliクラスである。

読み下し

一様大数法則は「関数族$\mathcal{F}$のすべてのf_θについて、経験平均$\frac{1}{n}\sum_i f_\theta(X_i)$が期待値$\mathbb{E}[f_\theta(X)]$に同時に近づく」と読む。「同時に」（一様に）という点が通常の大数の法則との違いであり、これによりM推定量の一致性（定理6.3.2の条件3）が保証される。

定義 6.4.8 (VC次元)**.** 集合族$\mathcal{C} = \{C_\theta : \theta \in \Theta\}$の**VC（Vapnik–Chervonenkis）次元**とは、$\mathcal{C}$によって**粉砕**（shatter）できる点集合の最大サイズである。

点集合$\{x_1, \dots, x_d\}$が$\mathcal{C}$によって粉砕されるとは、$\{x_1, \dots, x_d\}$の任意の部分集合Aに対して、ある$C_\theta \in \mathcal{C}$が存在して$C_\theta \cap \{x_1, \dots, x_d\} = A$が成り立つことをいう。

例 6.4.9 (半平面のVC次元)**.** $\mathbb{R}^2$における半平面$\{(x_1, x_2) : a_0 + a_1 x_1 + a_2 x_2 \geq 0\}$の族のVC次元は3である。

$d = 3$**が可能**：一般の位置にある3点は、任意の部分集合を半平面で分離できる（$2^3 = 8$通りすべて実現可能）。

$d = 4$**が不可能**：一般の位置にある4点で、凸包の内部にある1点だけを選ぶ分割は半平面では実現できない（Radonの定理）。

読み下し

VC次元は関数族の「表現力」の尺度である。VC次元が有限であれば、一様大数法則が成り立ち、経験的な最適化が理論的に正当化される。この概念は第6章の漸近理論の文脈では基礎的な道具であるが、巻3（第17章）の統計的学習理論において汎化誤差の解析に本格的に活用される。

6.5　局所漸近正規性

MLEの漸近理論は、個々のモデルにおける推定量の性質を述べるものであった。局所漸近正規性（Local Asymptotic Normality, LAN）は、視点をさらに抽象化し、統計的実験そのものの漸近的構造を明らかにする。これにより、「なぜMLEが漸近有効なのか」に対するより深い理解が得られる。

6.5.1　LAN条件

定義 6.5.1 (局所漸近正規性（LAN）)**.** パラメトリックモデル$\{P_\theta : \theta \in \Theta\}$が$\theta_0$で**局所漸近正規**（LAN）であるとは、局所パラメータ$t \in \mathbb{R}^p$に対して対数尤度比が

$$\log \frac{L_n(\theta_0 + n^{-1/2}t)}{L_n(\theta_0)} = t^\top \Delta_n - \frac{1}{2} t^\top \mathcal{I}(\theta_0)\, t + o_p(1)$$

と表せることをいう。ここで$\Delta_n \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0))$は**中心統計量**と呼ばれ、

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(X_i; \theta_0)$$

（スコア関数の正規化和）である。

読み下し

LAN条件は「対数尤度比をθ_0の$1/\sqrt{n}$近傍で見ると、tに関する2次形式$t^\top \Delta_n - \frac{1}{2} t^\top \mathcal{I}(\theta_0)\, t$に近づく」と読む。言い換えると、局所的には統計的実験が

$$\mathcal{N}\big(\mathcal{I}(\theta_0)^{-1} t,\ \mathcal{I}(\theta_0)^{-1}\big)$$

からtを推定する問題に漸近的に帰着する。

例 6.5.2 (i.i.d.サンプルのLAN)**.** $X_1, \dots, X_n$ i.i.d. $f(x;\theta)$の場合を確認する。$\theta_0 + n^{-1/2} t$でのTaylor展開（多次元パラメータ$t \in \mathbb{R}^p$の場合）：

$$\log \frac{f(X_i; \theta_0 + n^{-1/2} t)}{f(X_i; \theta_0)} \approx \frac{1}{\sqrt{n}} t^\top s(X_i; \theta_0) - \frac{1}{2n} t^\top \mathcal{I}(\theta_0) t + O(n^{-3/2})$$

n個を合算すると

$$\log \frac{L_n(\theta_0 + n^{-1/2} t)}{L_n(\theta_0)} \approx t^\top \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(X_i; \theta_0)}_{\Delta_n} - \frac{1}{2} t^\top \mathcal{I}(\theta_0) t$$

中心極限定理により$\Delta_n \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0))$であるから、LAN条件が成立する。

6.5.2　漸近ミニマックス定理と畳み込み定理

LAN構造から、推定量の漸近的な限界に関する深い結果が得られる。

定理 6.5.3 (漸近ミニマックス定理（Hájek–Le Camの不等式）)**.** *LAN条件の下で、ボウル型の損失関数Lに対して、任意の推定量列$\hat{\theta}_n$の最大リスクは*

$$\sup_{|t| \le M} \mathbb{E}_{\theta_0 + n^{-1/2} t}\big[L(t, \sqrt{n}(\hat{\theta}_n - \theta_0 - n^{-1/2} t))\big] \ge \mathbb{E}[L(0, Z)]$$

を満たす。ここで$Z \sim \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$である。MLEはこの下界を漸近的に達成する。

読み下し

この定理は「局所的に最善を尽くす推定量の最大リスクは、フィッシャー情報量で決まる正規分布のリスク以上である」と読む。MLEがこの下界を達成するということは、局所的にも大域的にも（漸近的に）最適であることを意味する。

表 6.3: 漸近近似が遅い・壊れる典型場面

状況	何が壊れやすいか	この巻での対処
境界・非正則モデル	χ^2 近似や $\sqrt{n}$ 収束が崩れ、Wald 型近似が不安定になる	尤度比・スコアの方が安定しやすい。必要なら厳密法やシミュレーションで補う
強い歪み・重い裾	正規近似の収束が遅く、有限標本で分布が大きく歪む	Berry–Esseen 型評価と章末の数値実験で近似誤差を点検する
離散データ・稀少事象	Wald 区間が過小被覆しやすく、標準誤差も過小評価されやすい	第4章のスコア区間や尤度比区間を優先する
モデル誤特定・分散不均一	フィッシャー情報量ベースの標準誤差が楽観的になる	M推定量のサンドイッチ分散でロバストに補正する
高次元（p が固定でない）	固定次元の古典漸近理論そのものが適用不能になる	巻3 の高次元理論と正則化の枠組みに切り替える

定理 6.5.4 (畳み込み定理（Convolution theorem）). *LAN条件の下で、任意の正則な推定量列T_nの漸近分布は*

$$\sqrt{n}(T_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}) * W$$

の形に書ける（$*$は畳み込み）。ここでWは何らかの分布である。Wが点質量のとき（つまりT_nの漸近分布が正規分布そのもののとき）、T_nは漸近有効である。

読み下し

畳み込み定理は「正則な推定量の漸近分布は、 最適な正規成分$\mathcal{N}(0, \mathcal{I}^{-1})$に余計なノイズ$W$を足したもの以上に広がる」と読む。 MLEは$W = 0$（点質量）を達成するため漸近有効であり、 他の推定量はノイズWの分だけMLEに劣る。

実務ポイント

LAN理論は「なぜMLEが良いのか」に対する最も深い回答を与えるが、 以下の限界も理解しておくべきである：

- LAN条件は**正則**モデルで成り立つ。 パラメータが境界にある場合（例：$\theta \geq 0$の下で$\theta_0 = 0$）、 混合モデル、変化点検出などの非正則モデルではLAN条件が成り立たず、 MLEの収束速度が$\sqrt{n}$より遅くなったり、漸近分布が非正規になったりする。
- LAN理論は漸近的（$n \to \infty$）な結果である。 有限標本では、第5章で見たJames–Stein推定量のように、 縮小推定量がMLEを改善できる場合がある。

6.6　セミパラメトリック効率限界への展望

本章のここまでの議論では、モデルがパラメトリック（有限次元のパラメータで完全に指定される）であることを前提としてきた。しかし実務では、誤差分布を特定しない回帰モデルや、有限次元の関心パラメータと無限次元の局外パラメータを含むモデルが重要である。これらは**セミパラメトリックモデル**と呼ばれる。

セミパラメトリックモデルでは、関心パラメータの推定における効率限界が、パラメトリックモデルのCramér–Rao下界とは異なる形で現れる。鍵となる概念は以下の通りである：

- **接線空間**（tangent space）：パラメトリック部分空間と局外パラメータの摂動方向を含む関数空間
- **効率的影響関数**：接線空間への射影により得られる最適な影響関数
- **セミパラメトリック効率限界**：効率的影響関数の分散で与えられる漸近分散の下界

直観的には、局外パラメータ（例：誤差分布の形状）が未知であることにより、推定に使える情報の一部が失われる。セミパラメトリック効率限界は、この情報の損失を正確に定量化する。

これらの概念の厳密な定式化と、二重ロバスト推定やターゲット推定（TMLE）への応用は、巻3の第14章で詳しく展開する。本章で確立したLAN理論とM推定量の漸近理論が、その基盤となる。

実践：漸近近似の精度検証

漸近理論の結果は$n \to \infty$の極限であるが、実務では「自分のデータのnで漸近近似はどの程度正確か」を知る必要がある。以下のシミュレーションで、有限標本分布と漸近分布を比較する。

コード例：Rによる漸近近似の精度検証

```
# R: MLEの漸近正規性の有限標本精度を検証
set.seed(42)
n_values <- c(10, 30, 100, 500)
n_sim <- 5000
true_lambda <- 2  # 指数分布のrateパラメータ

par(mfrow = c(2, 2))
for (n in n_values) {
  # MLE: lambda_hat = 1/Xbar の標準化量を繰り返し計算
  z_stats <- replicate(n_sim, {
    x <- rexp(n, rate = true_lambda)
    lambda_hat <- 1 / mean(x)
    # 漸近分布: sqrt(n)(lambda_hat - lambda) ~ N(0, lambda^2)
    sqrt(n) * (lambda_hat - true_lambda) / true_lambda
  })
  # ヒストグラムと標準正規密度の重ね描き
  hist(z_stats, breaks = 50, prob = TRUE, col = "lightblue",
       main = paste("n =", n),
       xlab = "標準化MLE", xlim = c(-4, 4))
```

```
  curve(dnorm(x), add = TRUE, col = "red", lwd = 2)
  legend("topright", "N(0,1)", col = "red", lwd = 2, cex = 0.8)
}
```

コード例：Pythonによる経験過程とDonskerの定理の可視化

```
# Python: 経験過程の可視化とKolmogorov--Smirnov検定
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

rng = np.random.default_rng(42)
n_values = [50, 200, 1000]
fig, axes = plt.subplots(1, 3, figsize=(15, 4))

for ax, n in zip(axes, n_values):
    X = rng.standard_normal(n)
    x_grid = np.linspace(-3, 3, 500)

    # 経験分布関数
    F_emp = np.array([np.mean(X <= x) for x in x_grid])
    F_true = stats.norm.cdf(x_grid)

    # 経験過程 G_n(x) = sqrt(n) * (F_n(x) - F(x))
    G_n = np.sqrt(n) * (F_emp - F_true)

    ax.plot(x_grid, G_n, 'b-', alpha=0.8, label=r'$\mathbb{G}_n(x)$')
    ax.axhline(y=0, color='gray', linestyle='--', alpha=0.5)
    ax.set_title(f'n = {n}')
    ax.set_xlabel('x')
    ax.set_ylim(-2.5, 2.5)
    ax.legend()

    # K-S検定
    ks_stat, p_val = stats.kstest(X, 'norm')
    ax.text(0.02, 0.98, f'KS = {ks_stat:.3f}\np = {p_val:.3f}',
            transform=ax.transAxes, va='top', fontsize=9)

plt.suptitle('経験過程とブラウン橋への収束', fontsize=13)
plt.tight_layout()
plt.savefig('empirical_process.pdf')
plt.show()
```

コード例：Rによるサンドイッチ推定量の比較

```
# R: モデル誤指定下でのサンドイッチ推定量
library(sandwich)  # vcovHC()

set.seed(42)
n <- 200

# データ生成: 分散不均一モデル
x <- runif(n, 1, 5)
y <- 1 + 2 * x + rnorm(n, sd = 0.5 * x)  # 分散が x に比例

fit <- lm(y ~ x)

# 通常の標準誤差（等分散仮定）
```

```
se_ols <- summary(fit)$coefficients[, "Std. Error"]

# 有限標本では HC3 が安定しやすい
se_hc <- sqrt(diag(vcovHC(fit, type = "HC3")))

cat("通常の標準誤差:", round(se_ols, 4), "\n")
cat("HC3 標準誤差:  ", round(se_hc, 4), "\n")
cat("→ ロバスト標準誤差は常に大きいとは限らないため、係数ごとに比較する\n")
```

主要な結果

重要結果

本章の漸近理論の核心をまとめる：

1. **MLEの一致性**（定理6.1.3）：KL情報量の正値性と一様大数法則により、MLEは真のパラメータに確率収束する。
2. **MLEの漸近正規性**（定理6.1.5）：$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$。推定誤差は$1/\sqrt{n}$の速度で縮小し、正規分布で近似できる。
3. **漸近有効性**：MLEはCramér–Rao下界を漸近的に達成し、正則推定量の中で最適である。
4. **三大検定の漸近等価性**（定理6.2.1）：尤度比検定・Wald検定・スコア検定は大標本で同じχ^2分布に従う。
5. **M推定量の統一的枠組み**（定理6.3.3）：MLE・ロバスト推定・分位点回帰など広いクラスの推定量を サンドイッチ分散行列$V = H^{-1}\Sigma H^{-\top}$で統一的に扱える。
6. **Glivenko–Cantelli定理とDonskerの定理**（定理6.4.2、6.4.4）：経験分布は真の分布に一様収束し、偏差$\sqrt{n}(F_n - F)$はブラウン橋に弱収束する。
7. **局所漸近正規性（LAN）**（定義6.5.1）：正則モデルの対数尤度比は局所的にガウス的な2次形式で近似でき、 これが漸近有効性の深い根拠となる。

章末演習問題

標準問題

演習問題 6.1（ポアソン分布のMLEの漸近性質）．$X_1, \ldots, X_n$ i.i.d. $\mathrm{Pois}(\lambda)$に対して、MLEは$\hat{\lambda}_n = \bar{X}$である。以下を示せ：

1. 一致性：$\hat{\lambda}_n \xrightarrow{p} \lambda$
2. フィッシャー情報量：$\mathcal{I}(\lambda) = 1/\lambda$
3. 漸近正規性：$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda)$
4. 漸近有効性を確認せよ

演習問題 6.2 (デルタ法の応用)**.** 指数分布$X_1, \dots, X_n$i.i.d.Exp(λ)（rate parameter）のMLEは$\hat{\lambda}_n = 1/\bar{X}$である。 デルタ法を用いて、$g(\lambda) = e^{-\lambda}$（成功確率の変換）の推定量$g(\hat{\lambda}_n)$の漸近分布を導出し、 $\lambda = 1$、$n = 50$の場合の近似的な95%信頼区間を求めよ。

演習問題 6.3 (三大検定の数値比較)**.** $X_1, \dots, X_n$i.i.d.Pois(λ)に対する$H_0 : \lambda = 1$を考える。

1. 尤度比統計量、Wald統計量、スコア統計量をそれぞれ$\bar{X}$とnで表せ
2. $n = 20$、$\bar{X} = 1.5$の場合に三つのp値を数値的に計算し、差を比較せよ
3. シミュレーション（各nで10,000回、$n = 10, 50, 200$）により、 帰無仮説下での三統計量の経験分布を ヒストグラムまたはQ-Qプロットでχ^2_1分布と比較せよ

発展問題

演習問題 6.4 (GlivenkoとDonskerの違い)**.** 1. Glivenko–Cantelli定理が「大数の法則の関数版」、Donskerの定理が「中心極限定理の関数版」であることを、 数学的な対応関係を明示して説明せよ

2. $\mathbb{G}_n(x) = \sqrt{n}[F_n(x) - F(x)]$のブラウン橋への収束が、 Kolmogorov–Smirnov検定にどのように応用されるかを述べよ
3. 標準正規分布からの$n = 100$の標本で経験過程を計算し、 理論的なブラウン橋の分散$F(x)(1 - F(x))$と比較するシミュレーションを実装せよ

演習問題 6.5 (M推定量とサンドイッチ推定量)**.** 絶対損失$m(x, \theta) = |x - \theta|$に基づくM推定量（標本中央値の一般化）を考える。

1. $X_1, \dots, X_n$i.i.d.Fのとき、このM推定量の漸近分布を サンドイッチ形式$H^{-1}\Sigma H^{-1}$を用いて導出せよ （ヒント：$\rho'(u) = \operatorname{sgn}(u)$の微分の取り扱いに注意）
2. 二乗損失に基づくM推定量（標本平均）と比較して、 正規分布およびt分布（自由度3）の場合の漸近相対効率を計算せよ

ヒント: 推定方程式 $\sum_{i=1}^{n} \operatorname{sgn}(X_i - \theta) = 0$を用い、 真の中央値$\theta_0$の近傍で $\Sigma = \operatorname{Var}(\operatorname{sgn}(X - \theta_0))$, $H = 2f(\theta_0)$ となることを使え。

演習問題 6.6 (LAN条件の検証)**.** $X_1, \dots, X_n$i.i.d.$\mathcal{N}(\mu, 1)$とする。

1. $\theta_0 = \mu_0$のもとで対数尤度比 $\log\{L_n(\mu_0 + n^{-1/2}t)/L_n(\mu_0)\}$を計算し、 LAN展開$t\,\Delta_n - \frac{1}{2}t^2\,\mathcal{I}(\mu_0)$の形になることを直接確認せよ
2. 畳み込み定理（定理6.5.4）の意味を、この具体例で説明せよ

ヒント: 二乗完成ではなく、 $\sum_{i=1}^{n}\{(X_i - \mu_0 - t/\sqrt{n})^2 - (X_i - \mu_0)^2\}$ を そのまま展開すると、$\Delta_n = n^{-1/2}\sum_{i=1}^{n}(X_i - \mu_0)$ と $\mathcal{I}(\mu_0) = 1$ が現れる。

計算・実装問題

数値実験を含む問題では、(1) 設定（分布、パラメータ、反復回数、乱数シード）、 (2) 作成した図表または表、(3) 比較指標、(4) 結果から読める一言考察、 の4点を答えに含めよ。

演習問題 6.7 (漸近近似の精度：有限標本シミュレーション)**.** $X_1, \dots, X_n$i.i.d.Bern(p)に対するMLEの漸近正規近似の精度を、 $p = 0.05$（偏った場合）と$p = 0.5$（対称な場合）について比較せよ。 各設定ではモンテカルロ反復$B = 10{,}000$を用いよ。

1. $n = 20, 50, 100, 500$について、 $\sqrt{n}(\hat{p}_n - p)/\sqrt{p(1-p)}$の経験分布をヒストグラムで描き、$\mathcal{N}(0,1)$と重ねよ
2. $p = 0.05$の場合に漸近近似が遅いことの理由を、 Bern(p)の歪度$\gamma_1 = (1-2p)/\sqrt{p(1-p)}$を用いて説明せよ
3. 第4章で導入した ウィルソン信頼区間とWald信頼区間のカバレッジ確率を$n = 30$で比較し、 有限標本での漸近近似の限界を議論せよ

略解の指針

ここでは使う定理だけでなく、途中で確認すべき式も書く。証明の細部や作図は自分で埋めること。

- **演習6.1** 使う道具: 大数の法則、中心極限定理、フィッシャー情報量。 最初の1手: $\hat{\lambda}_n = \bar{X}$ に対して $\mathbb{E}[X_i] = \lambda$, $\mathrm{Var}(X_i) = \lambda$ を使う。 途中の要点: 一致性は大数の法則、 漸近正規性は $\sqrt{n}(\bar{X} - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda)$ を与える CLT から出る。 また対数尤度の二階微分の期待値より $\mathcal{I}(\lambda) = 1/\lambda$ を計算する。 最終形: $\hat{\lambda}_n \xrightarrow{p} \lambda$, $\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda)$, $\mathcal{I}(\lambda) = 1/\lambda$ なので漸近分散は $\mathcal{I}(\lambda)^{-1}$ に一致する。

- **演習6.2** 使う道具: 二重のデルタ法。 最初の1手: まず $\bar{X}$ の CLT を $h(x) = 1/x$ に通して $\hat{\lambda}_n = h(\bar{X})$ の漸近分布を出す。 途中の要点: $h'(1/\lambda) = -\lambda^2$, $g'(\lambda) = -e^{-\lambda}$ なので、 二回デルタ法を使うと分散係数は $\lambda^2 e^{-2\lambda}$ に落ちる。 最終形: $g(\hat{\lambda}_n)$ は $\mathcal{N}(e^{-\lambda}, \lambda^2 e^{-2\lambda}/n)$ に漸近する。 $\lambda = 1$, $n = 50$ では標準誤差は $e^{-1}/\sqrt{50} \approx 0.052$ で、 近似 95% 区間はおよそ $[0.266, 0.470]$ になる。

- **演習6.3** 使う道具: 三大検定の公式。 最初の1手: 尤度比は $2n\{\bar{X}\log\bar{X} - (\bar{X}-1)\}$, Wald は $n(\bar{X}-1)^2/\bar{X}$, スコアは $n(\bar{X}-1)^2$ と書く。 途中の要点: $n = 20$, $\bar{X} = 1.5$ では おおよそ LR $= 4.33$, Wald $= 3.33$, Score $= 5.00$。 これを χ^2_1 に当てると p 値はそれぞれ 約 0.037, 0.068, 0.025 で、同じデータでも結論が少しずれる。 最終形: n を増やすと三統計量の経験分布はどれも χ^2_1 に近づくが、 有限標本ではスコア、Wald、尤度比の順に近似の癖が異なることが確認点である。

- **演習6.4** 使う道具: 経験分布関数の一様収束と弱収束。 最初の1手: $F_n - F$ と $\sqrt{n}(F_n - F)$ を別々のスケールで見比べる。 途中の要点: Glivenko–Cantelli では $\sup_x |F_n(x) - F(x)| \to 0$ a.s. を見るのに対し、 Donsker では過程 $\mathbb{G}_n(x) = \sqrt{n}\{F_n(x) - F(x)\}$ の分布極限を見る。 シミュレーションでは点ごとの分散が $F(x)(1-F(x))$ になることを重ね描きで確認する。 最終形: Glivenko–Cantelli は一様大数法則、 Donsker はブラウン橋への弱収束であり、Kolmogorov–Smirnov 統計量は後者から導かれる。

- **演習6.5** 使う道具: M 推定のサンドイッチ分散。 最初の1手: 中央値の推定方程式 $\sum_{i=1}^{n}\operatorname{sgn}(X_i-\theta)=0$ を使い、 $\Sigma=1, H=2f(\theta_0)$ を確認する。 途中の要点: したがってサンドイッチ分散は $H^{-1}\Sigma H^{-1}=1/\{4f(\theta_0)^2\}$ になる。 正規分布では $f(0)=1/\sqrt{2\pi}$ だから中央値の漸近分散は $\pi/2$、 標本平均の分散 1 に対する相対効率は $2/\pi$ である。 t_3 では $f(0)=2/(\pi\sqrt{3})$ なので中央値の漸近分散は $3\pi^2/16$ となり、 標本平均の分散 3 より小さい。 最終形: $\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{d}\mathcal{N}(0,1/\{4f(\theta_0)^2\})$。 正規分布では中央値の相対効率は $2/\pi$, t_3 では中央値の方が標本平均より効率的になる。

- **演習6.6** 使う道具: 対数尤度比の直接展開。 最初の1手: $\mu_0+t/\sqrt{n}$ を代入した二乗和の差を展開する。 途中の要点: 交差項だけが $tn^{-1/2}\sum_{i=1}^{n}(X_i-\mu_0)$ を生み、 二次項は $-t^2/2$ にまとまる。 したがって中心統計量は $\Delta_n=n^{-1/2}\sum_{i=1}^{n}(X_i-\mu_0)$ で、 その分散は $\mathcal{I}(\mu_0)=1$ である。 最終形: 対数尤度比は $t\Delta_n-t^2/2$, $\Delta_n=n^{-1/2}\sum_{i=1}^{n}(X_i-\mu_0)$ となり、 このモデルでは標本平均が畳み込み定理の下限 1 を達成する。

- **演習6.7** 使う道具: 標準化ヒストグラムと被覆確率比較。 最初の1手: $p=0.05$ と $p=0.5$ を分けて、 標準化統計量のヒストグラムと区間被覆を別々に集計する。 途中の要点: 歪度は $\gamma_1=(1-2p)/\sqrt{p(1-p)}$ なので、 $p=0.05$ では約 4.13, $p=0.5$ では 0 になる。 この差が正規近似の速さの差として表れる。 カバレッジ比較では第4章の結果とつなげて、 Wald が境界付近で下振れしやすいことを確認する。 最終形: $p=0.05$ では漸近近似がかなり遅く、 Wald 区間は被覆が崩れやすい一方、Wilson 区間はかなり安定である。

次巻への橋渡し：漸近理論の先にあるもの

本章で確立した漸近理論は、巻1（推測理論の基礎）を締めくくるものであると同時に、 巻2以降のすべての章の理論的基盤となる。

実務ポイント

本章の先で理論を使い分けるときは、次の地図を持っておくと見通しがよい。

- **古典的漸近近似**： 次元が固定され、モデルが正則で、統計量が滑らかなときの第一選択である。 MLEの漸近正規性、Wald・尤度比・スコアのχ^2近似はこの型に属する。
- **再標本化**： 解析的に分布を出しにくい統計量や、区間推定の被覆精度を点検したい場面では、 bootstrap が有力な補助線になる。経験過程の理論はその正当化の共通言語である。
- **非漸近・高次元理論**： pがnとともに増える場合や有限標本保証が重要な場合は、集中不等式と確率過程の上界が主役になる。

- **巻2・第7章**（ベイズ推論）： ベイズ推定量の漸近性質（Bernstein–von Misesの定理）は、 LAN理論の直接的な帰結として理解できる。 事後分布は大標本で正規分布に近づき、MLEと一致する。

- **巻2・第8章**（計算統計）： ブートストラップ法は、解析的な分布近似が難しい統計量に対して 経験分布から近似分布を構成する方法である。 Donskerの定理と経験過程の理論が、その漸近正当化の基盤を提供する。

- **巻2・第11章**（ノンパラメトリック推論）：カーネル密度推定やノンパラメトリック回帰の収束速度は、本章のパラメトリックな$\sqrt{n}$レートとは異なり、バイアスと分散のトレードオフから$n^{-2s/(2s+d)}$のレートとなる。

- **巻3・第13章**（高次元統計）：次元pがnと同程度に大きい場合、古典的な漸近理論（p固定で$n \to \infty$）は破綻する。非漸近的な集中不等式と確率過程の上界が、新たな理論的枠組みを提供する。

- **巻3・第14章**（セミパラメトリック理論）：本章で予告したセミパラメトリック効率限界の理論は、LAN理論と経験過程を基盤として構築される。

巻2では、ベイズ推論と計算統計から始めて、漸近理論を実際のモデリングと計算に応用する段階に進む。

参考文献ノート

最初に読む順としては、van der Vaart (1998) で古典的漸近理論の骨格を押さえ、Efron and Tibshirani (1993) で再標本化へ、Vershynin (2026) で高次元確率へ進むと、本章末の地図に対応づけやすい。

教科書

- **van der Vaart**, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press. — 本章の標準的な参考文献。LAN理論とセミパラメトリック理論の定番。

- **Lehmann**, E.L. and **Casella**, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer. — MLE漸近理論の古典的な取り扱い。

- **van der Vaart**, A.W. and **Wellner**, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer. — 経験過程の包括的な理論。Donskerクラスの一般論。

- **Le Cam**, L. and **Yang**, G.L. (2000). *Asymptotics in Statistics*, 2nd ed. Springer. — LAN理論とLe Camの漸近決定理論。

発展的文献

- **Huber**, P.J. and **Ronchetti**, E.M. (2009). *Robust Statistics*, 2nd ed. Wiley. — M推定量とロバスト推定の理論。

- **Efron**, B. and **Tibshirani**, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall. — 再標本化法の古典的入門。巻2第8章の計算統計へ進む読者に有用。

- **Kosorok**, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer. — 経験過程からセミパラメトリック理論への橋渡し。本章から巻3第14章へ進む読者に推奨。

- **Fernandez-Granda**, C. (2025). *Probability and Statistics for Data Science*. Cambridge University Press. — 厳密さを保ちながらも、例題・演習・コードを通じて 漸近近似をデータサイエンスの文脈に接続する近年の総合テキスト。
- **Vershynin**, R. (2026). *High-Dimensional Probability*, 2nd ed. Cambridge University Press. — 非漸近的 concentration と高次元統計の標準的入口。巻3第13章へ進む前の地図として有用。

付録

付録A

測度論の補足

問いと学習目標

この付録で答える問い

- 統計学の定理の証明で「積分と極限の交換」が使われるが、いつ交換が許されるのか？
- 密度関数はなぜ存在するのか？　尤度比はどのような意味で「測度の微分」なのか？
- 二重積分の順序交換（Fubini の定理）はいつ正当化されるか？
- 条件付き期待値はなぜ「L^2 射影」として理解できるのか？

読み終えたらできるようになること

1. 単調収束定理・Fatou の補題・優収束定理の仮定と主張を述べ、使い分けられる。
2. 具体的な統計量の計算で優収束定理や Fubini の定理を正しく適用できる。
3. Radon–Nikodym 定理を用いて密度関数・尤度比の存在を説明できる。
4. L^2 空間の射影定理が条件付き期待値の最適性を保証する仕組みを理解できる。
5. 積分と微分の交換が正当化される条件（正則条件）を判定できる。

直観的理解

本付録は、本巻の各章で用いる測度論的な道具立てを自己完結的にまとめたものである。 確率空間・確率変数・期待値・独立性といった基本概念は第1章で すでに導入済みであるから、ここでは本文の議論を支える 「裏方の定理」——積分と極限の交換、密度関数の存在保証、 積分順序の交換、微分と積分の交換——に焦点を当てる。
各定理には「統計学のどの場面で使うか」を明示した。 証明を読み飛ばしても、定理の主張と適用場面を把握すれば 本文の理解に支障はない。

A.1　測度空間と積分

第1章では確率測度 P を定義したが、 統計学の裏側では「確率に限らない一般の測度」が重要な役割を果たす。 例えば密度関数は、確率測度のルベーグ測度に対する Radon–Nikodym 導関数として定義される（第A.5節）。

定義 A.1.1 (測度)**.** 可測空間 $(X, \mathcal{F})$ 上の**測度**とは、 関数 $\mu\colon \mathcal{F} \to [0, \infty]$ で次の2条件を満たすものをいう：

(i) $\mu(\emptyset) = 0$。

(ii) 互いに素な $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ に対して $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$ （σ-加法性）。

$\mu(X) < \infty$ のとき**有限測度**、$\mu(X) = 1$ のとき確率測度である。 $X = \bigcup_n A_n$（$\mu(A_n) < \infty$）と書けるとき σ-**有限測度**と呼ぶ。

読み下し

確率測度は「全体が1」という制約付きの測度であり、 一般の測度は「長さ」「面積」「体積」の数学的抽象化である。 $\mathbb{R}$ 上のルベーグ測度は区間 $[a, b]$ に長さ $b - a$ を対応させる測度であり、 連続分布の密度関数を定義する際の基準となる。

定義 A.1.2 (ルベーグ積分の構成)**.** 測度空間 $(X, \mathcal{F}, \mu)$ 上の非負可測関数 f に対して、 ルベーグ積分を以下で定義する：

$$\int_X f \, d\mu = \sup\left\{ \int_X s \, d\mu : s \text{ は単関数で } 0 \le s \le f \right\}.$$

ここで**単関数**（simple function）とは $s = \sum_{k=1}^{N} a_k \mathbf{1}\{A_k\}$（$a_k \ge 0$, $A_k \in \mathcal{F}$）の形の関数で、 その積分は $\int s \, d\mu = \sum_{k=1}^{N} a_k \mu(A_k)$ と定義される。

一般の可測関数 f に対しては、$f^+ = \max(f, 0)$、$f^- = \max(-f, 0)$ として $\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu$ と定義する。 ただし、少なくとも一方が有限であるときに限り well-defined である。

読み下し

ルベーグ積分は「関数を下から単関数で近似し、その上限をとる」構成である。 リーマン積分が x 軸を分割するのに対し、ルベーグ積分は y 軸（値域）を分割する。 この構成により、極限操作との相性がよくなり、 以下の収束定理が成り立つ。

A.2　積分と極限の交換：収束定理

統計学では「n を大きくしたとき何が起こるか」を論じる場面が多く、 「積分（＝期待値）と極限の交換」を正当化する収束定理が不可欠である。 本節の3つの定理は互いに密接に関連しており、 単調収束定理 $\Rightarrow$ Fatou の補題 $\Rightarrow$ 優収束定理 の順に導出される。

定理 A.2.1 (単調収束定理（MCT）)**.** *非負可測関数の列 $\{f_n\}$ が単調増加（$f_n \le f_{n+1}$ a.e.）し、 $f_n \uparrow f$ a.e. ならば*

$$\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu.$$

読み下し

「非負で単調に増える関数列ならば、積分と極限を自由に交換してよい」 という定理である。 非負性と単調性の2条件が揃えば、追加の仮定なしに交換が正当化される。 統計学では、正の項の無限和と積分の交換（例：分配関数の計算、 ベイズ推論の周辺尤度計算）の正当化に用いられる。

例 A.2.2 (MCT の応用：期待値の尾部確率公式（離散版）)**.** 非負の整数値をとる確率変数 X に対して $\mathbb{E}[X] = \sum_{n=0}^{\infty} \mathrm{P}(X > n)$ を示す（重積分の順序交換である Tonelli の定理を、計数測度に対して適用した結果と見なせる）。

$f_N = \sum_{n=0}^{N} \mathbf{1}\{X > n\}$ は非負で単調増加であり、$f_N \uparrow \sum_{n=0}^{\infty} \mathbf{1}\{X > n\}$。 MCT より

$$\mathbb{E}\left[\sum_{n=0}^{\infty} \mathbf{1}\{X > n\}\right] = \lim_{N\to\infty} \sum_{n=0}^{N} \mathrm{P}(X > n) = \sum_{n=0}^{\infty} \mathrm{P}(X > n).$$

$X \sim \mathrm{Pois}(\lambda)$ で検証すると、$\mathbb{E}[X] = \lambda$ であり、一方 $\sum_{n=0}^{\infty} \mathrm{P}(X > n) = \sum_{n=0}^{\infty} (1 - \sum_{k=0}^{n} e^{-\lambda}\lambda^k/k!)$ も λ に一致する。

定理 A.2.3 (Fatou の補題)**.** *非負可測関数の列 $\{f_n\}$ に対して*

$$\int \liminf_{n\to\infty} f_n \, d\mu \le \liminf_{n\to\infty} \int f_n \, d\mu.$$

読み下し

「非負関数の列に対して、積分の下極限は、下極限の積分*以上*である」。 等号が成り立つとは限らない。 逆向きの不等式（上からの評価）には優収束定理（定理A.2.5）のような追加条件が必要になる。

例 A.2.4 (Fatou の補題で等号が成り立たない例)**.** $f_n = n \cdot \mathbf{1}\{[0, 1/n]\}$（ルベーグ測度上）とする。 各点で $f_n(x) \to 0$（$x > 0$ で）であるから $\int \liminf f_n \, dx = \int 0 \, dx = 0$。 しかし $\int f_n \, dx = n \cdot (1/n) = 1$ であるから $\liminf \int f_n \, dx = 1$。 不等式 $0 \le 1$ は成り立つが、等号は成立しない。

この「質量が原点に集中する」現象が、 Fatou の補題が一般に等号を保証しない理由である。 優収束定理の「優関数 g による上からの抑え」は、 この種の質量の逸散を防ぐ役割を果たす。

実務ポイント

Fatou の補題は、漸近理論でリスク関数の下界を導くときに使われる。 例えば、推定量の列 $\{\hat{\theta}_n\}$ のリスク $R_n = \mathbb{E}[L(\theta, \hat{\theta}_n)]$ の $\liminf$ を評価する際、 損失関数の非負性を利用して Fatou の補題を適用できる。 ミニマックス理論（第5章）での 下界の議論に

も現れる。

定理 A.2.5 (優収束定理（DCT）)**.** *可測関数の列 $\{f_n\}$ が $f_n \to f$ a.e. で、可積分関数 g（$\int g\,d\mu < \infty$）により $|f_n| \le g$ a.e. がすべての n で成り立つとき*

$$\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu.$$

証明の概略. $g + f_n \ge 0$ と $g - f_n \ge 0$ にそれぞれ Fatou の補題を適用する。

$$\int (g+f)\,d\mu \le \liminf \int (g+f_n)\,d\mu = \int g\,d\mu + \liminf \int f_n\,d\mu, \tag{A.1}$$

$$\int (g-f)\,d\mu \le \liminf \int (g-f_n)\,d\mu = \int g\,d\mu - \limsup \int f_n\,d\mu. \tag{A.2}$$

$\int g\,d\mu < \infty$ を用いて整理すると、$\int f\,d\mu \le \liminf \int f_n\,d\mu$ と $\limsup \int f_n\,d\mu \le \int f\,d\mu$ が得られ、合わせて $\lim \int f_n\,d\mu = \int f\,d\mu$。 □

例 A.2.6 (DCT の応用：指数分布のモーメントのパラメータ微分)**.** $X \sim \mathrm{Exp}(\theta)$（$\theta > 0$）のとき、$\mathbb{E}_\theta[X^2] = \int_0^\infty x^2 \theta e^{-\theta x}\,dx = 2/\theta^2$ であるが、この式を θ で微分してよいか？

$f(x,\theta) = x^2\theta e^{-\theta x}$ の θ 微分は

$$\frac{\partial f}{\partial \theta}(x,\theta) = x^2 e^{-\theta x}(1-\theta x).$$

$\theta \in [\theta_0 - \delta, \theta_0 + \delta] \subset (0,\infty)$ 上で $\theta_1 = \theta_0 - \delta > 0$ とすると、$\left|\frac{\partial f}{\partial \theta}\right| \le x^2 e^{-\theta_1 x}(1 + \theta_0 x + \delta x)$ であり、右辺は x の多項式と指数減衰の積で可積分な**優関数**を与える。

定理A.6.1（DCT に基づく微分と積分の交換）より

$$\frac{d}{d\theta}\mathbb{E}_\theta[X^2] = \int_0^\infty x^2 e^{-\theta x}(1-\theta x)\,dx = -\frac{4}{\theta^3}.$$

直接計算 $\frac{d}{d\theta}(2/\theta^2) = -4/\theta^3$ と一致する。

重要結果

優収束定理は統計学で最も頻繁に使われる測度論の定理である。本巻では以下の場面で（多くは暗黙に）適用される：

- **スコア関数の期待値**（第3章）：$\int f(x \mid \theta)\,dx = 1$ のパラメータ微分で 微分と積分を交換する際の正当化。
- **Fisher 情報量の2つの表現の同値性**（第3章）：対数尤度の2階微分と積分の交換。
- **MLE の一致性**（第6章）：対数尤度関数の一様収束の証明。
- **MLE の漸近正規性**（第6章）：スコア関数の中心極限定理への帰着。

実務ポイント

3つの収束定理の使い分けは以下の通りである：

- **MCT**：関数列が非負で単調増加ならば、無条件で交換可能。 級数と積分の交換（$\sum \int = \int \sum$）に特に便利。
- **Fatou**：非負関数に対して「不等式」のみが得られる。 上界が欲しいときではなく、下界が欲しいときに使う。
- **DCT**：可積分な優関数 g が見つかれば、非単調・符号付きでも等号が得られる。 統計学で最も汎用的。

A.3　Borel–Cantelli の補題

定理 A.3.1 (Borel–Cantelli の第一補題)**.** *事象の列* $\{A_n\}_{n=1}^{\infty}$ *が* $\sum_{n=1}^{\infty} \mathrm{P}(A_n) < \infty$ *を満たすならば*

$$\mathrm{P}\left(\limsup_{n\to\infty} A_n\right) = 0.$$

読み下し

$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$ は「A_n が無限回起こる事象」を表す。 第一補題は「確率の和が有限ならば、事象は有限回しか起こらない（a.s.）」と述べている。

Proof. 任意の N に対して $\mathrm{P}(\bigcup_{k=N}^{\infty} A_k) \le \sum_{k=N}^{\infty} \mathrm{P}(A_k) \to 0$ $(N \to \infty)$。 確率測度の上からの連続性より $\mathrm{P}(\limsup A_n) = \lim_{N\to\infty} \mathrm{P}(\bigcup_{k=N}^{\infty} A_k) = 0$。 □

定理 A.3.2 (Borel–Cantelli の第二補題)**.** *事象の列* $\{A_n\}$ *が* **独立** *で* $\sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty$ *ならば*

$$\mathrm{P}\left(\limsup_{n\to\infty} A_n\right) = 1.$$

証明の概略. 独立性より $\mathrm{P}\left(\bigcap_{k=N}^{M} A_k^c\right) = \prod_{k=N}^{M} (1 - \mathrm{P}(A_k))$。 不等式 $1 - x \le e^{-x}$ を用いると $\prod_{k=N}^{M} (1 - \mathrm{P}(A_k)) \le \exp\left(-\sum_{k=N}^{M} \mathrm{P}(A_k)\right) \to 0$ （$M \to \infty$）。よって $\mathrm{P}(\bigcup_{k=N}^{\infty} A_k) = 1$ が全ての N で成り立ち、 $\mathrm{P}(\limsup A_n) = 1$。 □

実務ポイント

Borel–Cantelli の補題は、推定量の**概収束**（a.s. convergence）を 示す際の基本ツールである。 例えば強大数の法則（第2章、定理2.3.3）の証明では、 $\mathrm{P}(|\bar{X}_n - \mu| > \varepsilon)$ の和が有限であることを示し、 第一補題を適用して $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$ を導く。

A.4　積測度と Fubini の定理

定義 A.4.1 (積 σ-加法族と積測度)**.** 可測空間 $(X,\mathcal{F})$ と $(Y,\mathcal{G})$ に対して、**積 σ-加法族** $\mathcal{F}\otimes\mathcal{G}$ は $\{A\times B : A\in\mathcal{F},\, B\in\mathcal{G}\}$ を含む最小の σ-加法族である。

σ-有限測度 μ（X 上）と ν（Y 上）に対して、$(\mu\otimes\nu)(A\times B)=\mu(A)\cdot\nu(B)$ を満たす $\mathcal{F}\otimes\mathcal{G}$ 上の一意な測度 $\mu\otimes\nu$ が存在する。これを**積測度**と呼ぶ。

読み下し

確率変数 $X_1,\dots,X_n$ が独立であるとは、同時分布が各周辺分布の積測度であることに他ならない。i.i.d. 標本 $X_1,\dots,X_n$ の確率空間は、共通の分布に対応する測度の n 重積として構成される。

定理 A.4.2 (Tonelli の定理)**.** *σ-有限測度空間 $(X,\mathcal{F},\mu)$ と $(Y,\mathcal{G},\nu)$ 上の* **非負** *可測関数 $f\colon X\times Y\to[0,\infty]$ に対して*

$$\int_{X\times Y} f\,d(\mu\otimes\nu)=\int_X\left(\int_Y f(x,y)\,d\nu(y)\right)d\mu(x)=\int_Y\left(\int_X f(x,y)\,d\mu(x)\right)d\nu(y).$$

読み下し

Tonelli の定理は「非負ならば積分の順序を自由に交換できる」と述べている。可積分性の確認なしに使えるため、まず Tonelli で $\int|f|\,d(\mu\otimes\nu)$ の有限性を確認し、次に Fubini を適用する、という手順が標準的である。

定理 A.4.3 (Fubini の定理)**.** *σ-有限測度空間上の* **可積分** *関数 $f\in L^1(\mu\otimes\nu)$ に対して*

$$\int_{X\times Y} f(x,y)\,d(\mu\otimes\nu)=\int_X\left(\int_Y f(x,y)\,d\nu(y)\right)d\mu(x)=\int_Y\left(\int_X f(x,y)\,d\mu(x)\right)d\nu(y).$$

すなわち、積分の順序交換が正当化される。

例 A.4.4 (Tonelli の定理：$\mathbb{E}[X]=\int_0^\infty \mathrm{P}(X>t)\,dt$)**.** 非負確率変数 X に対して、$\mathbb{E}[X]$ を尾部確率の積分として表す **尾部確率公式**を導出する。

$f(x,t)=\mathbf{1}\{t<x\}$ とおく。これは非負可測であるから、Tonelli の定理を確率測度 P（x の積分）とルベーグ測度（t の積分）に適用して

$$\int_0^\infty\int_\Omega \mathbf{1}\{t<X(\omega)\}\,d\mathrm{P}(\omega)\,dt=\int_\Omega\int_0^\infty \mathbf{1}\{t<X(\omega)\}\,dt\,d\mathrm{P}(\omega).$$

左辺 $=\int_0^\infty \mathrm{P}(X>t)\,dt$、右辺 $=\int_\Omega X(\omega)\,d\mathrm{P}(\omega)=\mathbb{E}[X]$。

具体例：$X\sim\mathrm{Exp}(\lambda)$（$\lambda=2$）のとき $\mathrm{P}(X>t)=e^{-2t}$ であるから

$$\int_0^\infty e^{-2t}\,dt=\frac{1}{2}=\mathbb{E}[X].\quad\checkmark$$

この公式は、独立な確率変数の和の期待値計算や、生存時間解析における平均余命の表現に頻出する。

重要結果

Fubini の定理は統計学で以下の場面に現れる：

- **周辺分布の導出**：同時密度 $f_{X,Y}(x,y)$ から $f_X(x) = \int f_{X,Y}(x,y)\,dy$ を得る操作。
- **ベイズリスクの計算**（第5章）：$r_\pi(\delta) = \int_\Theta \int_{\mathcal{X}} L(\theta,\delta(\boldsymbol{x}))f(\boldsymbol{x}\mid\theta)\,d\boldsymbol{x}\,\pi(\theta)\,d\theta$ の積分順序交換。
- **期待値の計算の簡略化**：非負確率変数 X に対する公式 $\mathbb{E}[X] = \int_0^\infty \mathrm{P}(X > t)\,dt$ は Tonelli の定理から導かれる（例A.4.4）。

A.5　Radon–Nikodym 定理と測度の絶対連続性

定義 A.5.1 (絶対連続性と特異性)**.** σ-有限測度 μ, ν が $(X,\mathcal{F})$ 上に与えられているとする。

(i) ν が μ に対して**絶対連続**（$\nu \ll \mu$）であるとは、$\mu(A) = 0 \implies \nu(A) = 0$ が成り立つことをいう。

(ii) μ と ν が**互いに特異**（$\mu \perp \nu$）であるとは、$A \cap B = \emptyset$, $A \cup B = X$ なる集合が存在して $\mu(B) = 0$ かつ $\nu(A) = 0$ が成り立つことをいう。

読み下し

直観的には、$\nu \ll \mu$ は「μ で測って大きさゼロの集合は ν でも大きさゼロ」ということであり、ν は μ に比べて「支持の範囲が狭い」関係にある。連続分布の確率測度はルベーグ測度に対して絶対連続であり、離散分布の確率測度は数え上げ測度に対して絶対連続である。

定理 A.5.2 (Radon–Nikodym 定理)**.** *μ, ν を $(X,\mathcal{F})$ 上の σ-有限測度とし、$\nu \ll \mu$ とする。このとき、非負可測関数 f が a.e. 一意に存在して*

$$\nu(A) = \int_A f\,d\mu, \quad \forall A \in \mathcal{F}.$$

*この f を***Radon–Nikodym 導関数** $\frac{d\nu}{d\mu}$ *と書く。*

読み下し

Radon–Nikodym 定理は「一つの測度を別の測度で割り算できる」ことを保証する。この「商」が密度関数に他ならない。確率測度 P がルベーグ測度 λ に対して絶対連続ならば、$dP/d\lambda = f$ が確率密度関数である。

例 A.5.3 (正規分布の密度を R–N 導関数として理解する)**.** P を $\mathcal{N}(\mu,\sigma^2)$ の確率測度、λ をルベーグ測度とする。$P \ll \lambda$（$\lambda(A) = 0$ ならば $P(A) = 0$）であるから、R–N 定理により

$$\frac{dP}{d\lambda}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

が存在する。これが「確率密度関数」の正体である。

さらに、2つの正規分布 $P_0 = \mathcal{N}(0,1)$ と $P_1 = \mathcal{N}(\mu_1, 1)$ について、$P_1 \ll P_0$（同じ台を持つ）であるから、尤度比は

$$\frac{dP_1}{dP_0}(x) = \frac{dP_1/d\lambda}{dP_0/d\lambda}(x) = \exp\left(\mu_1 x - \frac{\mu_1^2}{2}\right).$$

これは第4章のNeyman–Pearson 補題で用いる尤度比そのものである。

重要結果

Radon–Nikodym 定理は本巻の多くの概念を支える基盤である：

- **密度関数の存在**：確率測度 P のルベーグ測度に対する密度 $f(x) = dP/d\lambda$ の存在が保証される。
- **尤度比**（第4章）：2つのモデル P_{θ_0} と P_{θ_1} の尤度比 $L(\theta_1)/L(\theta_0) = dP_{\theta_1}/dP_{\theta_0}$ は Radon–Nikodym 導関数である。
- **条件付き期待値の存在**（第1章、定義1.3.4）：条件付き期待値 $\mathbb{E}[X \mid \mathcal{G}]$ の存在は Radon–Nikodym 定理から導かれる。
- **十分統計量**（第3章）：ファクトリゼーション定理の一般的な証明（連続の場合）には Radon–Nikodym 定理が必要である。

定理 A.5.4 (Lebesgue 分解定理)**.** *σ-有限測度 μ, ν に対して、一意な分解*

$$\nu = \nu_{ac} + \nu_s, \quad \nu_{ac} \ll \mu, \quad \nu_s \perp \mu$$

が存在する。

読み下し

任意の測度は、基準測度に対して「密度を持つ部分」と「全く支持が重ならない部分」に一意に分解できる。混合分布（連続成分と離散成分を持つ分布）はこの分解の統計的な典型例である。例えば、確率 p で 0（点質量）、確率 $1-p$ で Exp(λ)（連続成分）に従う混合分布は、ルベーグ測度に対して ν_{ac}（密度 $(1-p)\lambda e^{-\lambda x}$ の部分）と ν_s（原点での点質量 $p\delta_0$ の部分）に分解される。

A.6　積分と微分の交換

統計的推論では、パラメータ θ に依存する確率密度 $f(x \mid \theta)$ の期待値をパラメータで微分する操作が頻出する（スコア関数の期待値、Fisher 情報量の計算など）。この操作は**積分と微分の交換**に他ならない。

定理 A.6.1 (パラメータ微分と積分の交換)**.** *$\Theta \subset \mathbb{R}$ を開区間、$f(x, \theta)$ を可測関数とし、以下の条件を仮定する：*

(i) *各 $\theta \in \Theta$ に対して $\int f(x,\theta)\,d\mu(x)$ が存在する。*

(ii) *ほとんどすべての x に対して $\theta \mapsto f(x,\theta)$ は微分可能。*

(iii) *可積分関数 $g(x)$ が存在して、すべての $\theta \in \Theta$ に対して $\left|\frac{\partial}{\partial\theta} f(x,\theta)\right| \le g(x)$ a.e.*

このとき

$$\frac{d}{d\theta}\int f(x,\theta)\,d\mu(x) = \int \frac{\partial}{\partial\theta} f(x,\theta)\,d\mu(x).$$

読み下し

この定理は「θ で微分する操作と x で積分する操作を交換してよい」と述べている。条件 (iii) が鍵であり、「θ を少し動かしたときの f の変化率が、θ に依存しない可積分関数 $g(x)$ で抑えられる」ことを要求する。これはまさに優収束定理の条件であり、証明は差分商に DCT を適用することで得られる。

Proof. $\theta_n \to \theta_0$ なる任意の列をとる。差分商 $h_n(x) = \frac{f(x,\theta_n)-f(x,\theta_0)}{\theta_n-\theta_0}$ に対して、平均値の定理より $|h_n(x)| = \left|\frac{\partial f}{\partial\theta}(x,\tilde\theta_n)\right| \le g(x)$（ある $\tilde\theta_n$ が θ_0 と θ_n の間に存在）。仮定 (ii) より $h_n(x) \to \frac{\partial f}{\partial\theta}(x,\theta_0)$ a.e. であるから、優収束定理（定理A.2.5）を適用して

$$\lim_{n\to\infty}\int h_n(x)\,d\mu(x) = \int \frac{\partial f}{\partial\theta}(x,\theta_0)\,d\mu(x).$$

□

例 A.6.2 (スコア関数の期待値がゼロであることの証明)**.** 密度 $f(x\mid\theta)$ を持つ統計モデルにおいて、$\int f(x\mid\theta)\,dx = 1$ の両辺を θ で微分する。

条件 (iii) が成り立つ（これが「正則条件」の一つ）と仮定すると、定理A.6.1より

$$0 = \frac{d}{d\theta}\int f(x\mid\theta)\,dx = \int \frac{\partial}{\partial\theta} f(x\mid\theta)\,dx = \int \frac{\partial \log f}{\partial\theta}\cdot f(x\mid\theta)\,dx = \mathbb{E}_\theta[s(\theta;X)].$$

これが命題3.5.2（スコア関数の期待値がゼロ）の証明の核心である。ここで $s(\theta;X) = \frac{\partial}{\partial\theta}\log f(X\mid\theta)$ はスコア関数。

重要結果

この定理は統計学の**正則条件**の核心部分である。具体的には：

- **スコア関数の期待値がゼロ**（命題3.5.2）：$\int f(x\mid\theta)\,dx = 1$ の両辺を θ で微分する操作で使用。条件 (iii) が正則条件の一つに対応する。
- **Fisher 情報量の2つの表現の同値性**（定義3.5.3）：$\mathcal{I}(\theta) = \mathbb{E}[s(\theta)^2] = -\mathbb{E}[\ell''(\theta)]$ の導出で、2階微分と積分の交換にも同じ定理を適用する。
- **Cramér–Rao 下界の証明**（定理3.10.5）：不偏性条件 $\mathbb{E}_\theta[\hat\theta] = \theta$ のパラメータ微分。

注意 A.6.3 (多次元への拡張)*.* パラメータ $\boldsymbol{\theta} \in \mathbb{R}^d$ の場合も、各偏微分 $\partial/\partial\theta_j$ に対して同様の条件を課せば、勾配 $\nabla_{\boldsymbol{\theta}}$ と積分の交換が正当化される。Fisher 情報行列（定義3.5.4）の計算ではこの拡張が用いられる。

A.7　L^p 空間

条件付き期待値の存在証明（次節）には、「関数の空間に距離や角度の構造を入れ、その空間上で最良近似（射影）を見つける」という考え方が必要になる。本節では、そのための道具立てを整備する。

まず L^p 空間を定義し、そこに距離（ノルム）が入ることを確認する。次に L^2 の場合に内積が定義でき、「角度」や「直交」が意味を持つ空間になることを見る。

定義 A.7.1 (L^p 空間)**.** 測度空間 $(X, \mathcal{F}, \mu)$ と $1 \le p < \infty$ に対して

$$L^p(\mu) = \left\{ f : X \to \mathbb{R} \text{ は可測}, \ \|f\|_{L^p} := \left(\int |f|^p \, d\mu \right)^{1/p} < \infty \right\}$$

と定義する。$p = \infty$ の場合は $\|f\|_{L^\infty} = \operatorname{ess\,sup} |f|$。

読み下し

L^p 空間は「p 乗可積分な関数の集合」である。確率論では $\mathbb{E}[|X|^p] < \infty$ を「$X \in L^p$」と書く。L^1 は「期待値が有限」、L^2 は「分散が有限」に対応する。$\|f\|_{L^p}$ は関数 f の「大きさ」を測る尺度（ノルム）であり、p を大きくするほど大きな値を重視する。

定理 A.7.2 (Hölder の不等式)**.** $p, q > 1$ で $1/p + 1/q = 1$ のとき

$$\int |fg| \, d\mu \le \|f\|_{L^p} \cdot \|g\|_{L^q} .$$

$p = q = 2$ の場合は Cauchy–Schwarz の不等式となる。

読み下し

Hölder の不等式は「2つの関数の積の積分は、各関数の L^p ノルムの積で上から抑えられる」と読む。$p = q = 2$ の場合の Cauchy–Schwarz は特に重要で、「共分散は各変数の標準偏差の積以下」（$|\mathrm{Cov}(X, Y)| \le \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$）と同値であり、相関係数が $[-1, 1]$ に収まることの根拠でもある。

定理 A.7.3 (Minkowski の不等式)**.** $1 \le p \le \infty$ のとき

$$\|f + g\|_{L^p} \le \|f\|_{L^p} + \|g\|_{L^p} .$$

すなわち、$\|\cdot\|_{L^p}$ は三角不等式を満たし、ノルムである。

読み下し

Minkowski の不等式は L^p のノルムが「三角不等式を満たす」ことを保証する。これにより L^p は距離空間となり、関数間の「近さ」を定量的に測れるようになる。直観的には「2つの関数の和の大きさは、各関数の大きさの和を超えない」という自然な性質である。

L^p 空間にはもう一つ重要な性質——**完備性**——がある。 完備性とは「コーシー列が必ず極限を持つ」性質であり、 収束先が空間の外に逃げないことを保証する。

定理 A.7.4 (Riesz–Fischer の定理（完備性）)**.** *$L^p(\mu)$（$1 \le p \le \infty$）は完備なノルム空間（**バナッハ空間**）である。*

読み下し

完備性は「L^p の中で収束しそうな列は、実際に L^p の中で収束する」 ことを意味する。 これは以下の射影定理を使うための必須条件である。 有限次元のユークリッド空間 $\mathbb{R}^d$ は自動的に完備であるが、 無限次元の関数空間では完備性の証明が必要になる。

$p = 2$ の場合、L^2 にはさらに豊かな構造が入る。

定義 A.7.5 (L^2 空間の内積)**.** $L^2(\mu)$ 上の**内積**を

$$\langle f, g \rangle = \int f\, g \, d\mu$$

と定義する。この内積からノルム $\|f\|_{L^2} = \sqrt{\langle f, f \rangle}$ が復元される。 内積が定義された完備なノルム空間を**ヒルベルト空間**と呼ぶ。

読み下し

内積は「2つの関数がどの程度似ているか」を測る量であり、 $\mathbb{R}^d$ でのベクトルの内積 $\boldsymbol{a} \cdot \boldsymbol{b} = \sum_i a_i b_i$ の 関数空間版である。 $\langle f, g \rangle = 0$ のとき f と g は**直交する**という—— ちょうど $\mathbb{R}^d$ のベクトルが直角に交わるのと同じ幾何学的意味を持つ。 確率変数で言えば、 $\langle X - \mathbb{E}[X],\, Y - \mathbb{E}[Y] \rangle = \mathrm{Cov}(X, Y) = 0$ は無相関を意味する。
L^2 がヒルベルト空間であること（完備性 + 内積の存在）が、 次の射影定理を使うための正確な前提条件である。

定理 A.7.6 (L^2 空間の射影定理)**.** *L^2 の**閉部分空間** M[1] に対して、任意の $f \in L^2$ は一意に*

$$f = \hat{f} + r, \quad \hat{f} \in M,\ r \perp M$$

*と分解できる。ここで $\hat{f}$ は M への**直交射影**であり、 $\left\| f - \hat{f} \right\|_{L^2} = \min_{g \in M} \|f - g\|_{L^2}$ を達成する。*

読み下し

射影定理は「L^2 の元 f を閉部分空間 M に最も近い点 $\hat{f}$ で 近似でき、残差 $r = f - \hat{f}$ は M に直交する」と述べている。 $\mathbb{R}^3$ のベクトルを平面に射影する操作の無限次元版である。
$\hat{f}$ が最良近似であることは直交性から従う： $g \in M$ を任意にとると $\|f - g\|^2 = \left\| (f - \hat{f}) + (\hat{f} - g) \right\|^2 = \|r\|^2 + \left\| \hat{f} - g \right\|^2 \ge \|r\|^2 = \left\| f - \hat{f} \right\|^2$ （ピタゴラスの定理の

[1] 閉部分空間とは、L^2 の線形部分空間で、 M 内のコーシー列の極限が再び M に属する（M が閉じている）もの。 例えば $M = L^2(\mathcal{G})$（$\mathcal{G}$-可測な L^2 関数全体）は閉部分空間である。

L^2 版）。

重要結果

L^2 射影の統計学的応用：

- **条件付き期待値**（定義1.3.4）：$\mathbb{E}[X \mid \mathcal{G}]$ は X の $L^2(\mathcal{G})$（$\mathcal{G}$-可測な L^2 関数の空間）への直交射影である。これが「与えられた情報の下での最良予測」であることの根拠。
- **最小二乗推定**：線形回帰における最小二乗推定量は、応答ベクトルの列空間への直交射影として理解できる（巻2参照）。
- **条件付き分散の分解**：$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid \mathcal{G})] + \mathrm{Var}(\mathbb{E}[X \mid \mathcal{G}])$ はピタゴラスの定理の L^2 版に対応する。

例 A.7.7 (条件付き期待値は L^2 射影である)**.** $X \in L^2(\mathrm{P})$ を確率変数、$\mathcal{G}$ を部分 σ-加法族とする。$M = L^2(\mathcal{G}) = \{Y \in L^2 : Y \text{ は } \mathcal{G}\text{-可測}\}$ は L^2 の閉部分空間である。

射影定理により、X の M への射影 $\hat{X}$ は $\mathbb{E}[(X - \hat{X})Y] = 0$（すべての $Y \in M$ に対して）を満たす。$\hat{X} = \mathbb{E}[X \mid \mathcal{G}]$ と同一視すると、この直交条件は「予測残差 $X - \mathbb{E}[X \mid \mathcal{G}]$ が $\mathcal{G}$-可測な 任意の関数と無相関である」ことを意味する。

具体例：$X \sim \mathcal{N}(0,1)$, $Y = X + Z$（$Z \sim \mathcal{N}(0,1)$, 独立）のとき、$\mathbb{E}[X \mid Y] = Y/2$ は X の $L^2(\sigma(Y))$ への射影であり、$X - Y/2 \perp g(Y)$（すべての g）が成り立つ。$\mathrm{Var}(X) = 1 = \mathrm{Var}(Y/2) + \mathrm{Var}(X - Y/2) = 1/2 + 1/2$ はピタゴラスの定理に対応する。

A.8　条件付き期待値の存在と正則条件付き確率

第1章（定義1.3.4）では 条件付き期待値を定義し、その性質を述べた。ここではその**存在**を Radon–Nikodym 定理から導出する。

定理 A.8.1 (条件付き期待値の存在)**.** *$(\Omega, \mathcal{F}, \mathrm{P})$ を確率空間、X を可積分確率変数（$\mathbb{E}[|X|] < \infty$）、$\mathcal{G} \subset \mathcal{F}$ を部分 σ-加法族とする。このとき、以下の2条件を満たす $\mathcal{G}$-可測確率変数 Y が a.s. で一意に存在する：*

(i) Y *は* $\mathcal{G}$*-可測。*

(ii) *すべての* $G \in \mathcal{G}$ *に対して* $\int_G Y \, d\mathrm{P} = \int_G X \, d\mathrm{P}$。

読み下し

条件 (i) は「Y は情報 $\mathcal{G}$ のみに基づいて計算できる」ことを要求し、条件 (ii) は「$\mathcal{G}$ で区別できるどの事象上でも、Y の平均と X の平均が一致する」ことを要求する。この2条件が条件付き期待値 $\mathbb{E}[X \mid \mathcal{G}]$ の定義そのものであり、「利用可能な情報 $\mathcal{G}$ の下で X の平均値を最もよく近似する量」である。存在が自明でないのは、$\mathcal{G}$ が一般の σ-加法族であるときに このような Y が構成できるかという点にある。

証明の概略. $\mathcal{G}$ 上の測度 $\nu(G) = \int_G X\,d\mathrm{P}$（$X \geq 0$ の場合）は $\mathrm{P}|_{\mathcal{G}}$ に対して絶対連続である。 Radon–Nikodym 定理（定理A.5.2）を $(\Omega, \mathcal{G}, \mathrm{P}|_{\mathcal{G}})$ 上で適用すると、 $\mathcal{G}$-可測関数 $Y = d\nu/d(\mathrm{P}|_{\mathcal{G}})$ が存在して $\nu(G) = \int_G Y\,d\mathrm{P}$ がすべての $G \in \mathcal{G}$ で成り立つ。 一般の X は $X = X^+ - X^-$ に分解して適用する。 □

直観的理解

条件付き期待値の存在は、本巻で最も深い存在定理の一つである。 その証明の論理を整理しよう：

1. X から $\mathcal{G}$ 上の測度 ν を構成する （$\nu(G) = \int_G X\,d\mathrm{P}$）。
2. $\nu \ll \mathrm{P}|_{\mathcal{G}}$ であるから、 Radon–Nikodym 定理が適用でき、 密度 $Y = d\nu/d(\mathrm{P}|_{\mathcal{G}})$ が存在する。
3. この Y が条件付き期待値 $\mathbb{E}[X \mid \mathcal{G}]$ である。

つまり、条件付き期待値とは 「X による測度変換の、制限された情報 $\mathcal{G}$ に対する密度」 なのである。

定義 A.8.2 (正則条件付き確率)**.** $(\Omega, \mathcal{F}, \mathrm{P})$ を確率空間、 $\mathcal{G} \subset \mathcal{F}$ を部分 σ-加法族とする。 写像 $\mathrm{P}(\cdot \mid \mathcal{G})\colon \mathcal{F} \times \Omega \to [0,1]$ が **正則条件付き確率**であるとは：

(i) 各 $A \in \mathcal{F}$ に対して $\omega \mapsto \mathrm{P}(A \mid \mathcal{G})(\omega)$ は $\mathbb{E}[\mathbf{1}\{A\} \mid \mathcal{G}]$ の一つの版。

(ii) a.s. な ω に対して $A \mapsto \mathrm{P}(A \mid \mathcal{G})(\omega)$ は $(\Omega, \mathcal{F})$ 上の確率測度。

定理 A.8.3 (正則条件付き確率の存在)**.** *Ω がポーランド空間（完備可分距離空間）であり、$\mathcal{F}$ がそのボレル σ-加法族であるとき、 任意の部分 σ-加法族 $\mathcal{G}$ に対して正則条件付き確率が存在する。*

読み下し

正則条件付き確率は、ベイズ統計学の厳密な基礎を提供する。 事後分布 $\pi(\theta \mid \boldsymbol{x})$ が 「θ に関する確率測度」として well-defined であるためには、 正則条件付き確率の存在が必要である。 $\mathbb{R}^d$ 上の分布は常にこの条件を満たすため、 通常の統計的応用では問題にならないが、 理論的整合性の保証として重要である。

A.9　演習問題

演習問題 A.1 (DCT の適用条件の確認)**.** $f_n(x) = nx\,e^{-nx^2}$（$x \geq 0$）に対して、以下を示せ：

1. 各 $x > 0$ で $f_n(x) \to 0$（$n \to \infty$）。
2. $\int_0^\infty f_n(x)\,dx = 1/2$（すべての n）。
3. したがって $\lim \int f_n \neq \int \lim f_n$ であり、 積分と極限の交換が成り立たない。
4. DCT の仮定の何が満たされないかを特定せよ。 *ヒント*：$|f_n| \leq g$ を満たす可積分な g は存在するか？

演習問題 A.2 (Fubini を用いた期待値の計算)**.** 非負確率変数 X に対して、尾部確率公式 $\mathbb{E}[X^2] = \int_0^\infty 2t\,\mathrm{P}(X > t)\,dt$ を Tonelli の定理を用いて証明せよ。

ヒント：$X^2 = \int_0^\infty 2t\,\mathbf{1}\{X > t\}\,dt$ を示してから Tonelli を適用せよ。

演習問題 A.3 (Radon–Nikodym 導関数の計算)**.** P_0 を Exp(1)（密度 e^{-x}、$x \geq 0$）、 P_1 を Exp(2)（密度 $2e^{-2x}$、$x \geq 0$）の確率測度とする。

1. $P_1 \ll P_0$ であることを確認せよ。
2. dP_1/dP_0 を求めよ。
3. この尤度比が Neyman–Pearson 検定 $H_0 : \lambda = 1$ vs. $H_1 : \lambda = 2$ の 検定統計量に一致することを説明せよ。

演習問題 A.4 (条件付き期待値と L^2 射影)**.** X, Z が独立で $X, Z \sim \mathcal{N}(0,1)$、$Y = 2X + Z$ とする。

1. $\mathbb{E}[X \mid Y]$ を $L^2(\sigma(Y))$ への射影として求めよ。 ヒント：$\mathbb{E}[X \mid Y] = aY$ の形を仮定し、直交条件 $\mathbb{E}[(X - aY)Y] = 0$ から a を決定せよ。
2. $\mathrm{Var}(X) = \mathrm{Var}(\mathbb{E}[X \mid Y]) + \mathbb{E}[\mathrm{Var}(X \mid Y)]$ を検証せよ （ピタゴラスの定理の確認）。

演習問題 A.5 (優収束定理の応用：ベルヌーイ分布の Fisher 情報量)**.** $X \sim \mathrm{Bern}(p)$ のとき、Fisher 情報量の等式 $\mathcal{I}(p) = \mathbb{E}[s(p;X)^2] = -\mathbb{E}[\ell''(p;X)]$ が成り立つことを、以下の手順で確認せよ。

1. 対数尤度 $\ell(p;x) = x\log p + (1-x)\log(1-p)$ のスコア関数 $s(p;x)$ を求めよ。
2. $\mathbb{E}[s(p;X)^2]$ を直接計算せよ。
3. $\ell''(p;x)$ を求め、$-\mathbb{E}[\ell''(p;X)]$ と比較せよ。
4. この等式の成立に、定理A.6.1の 条件がどのように関わるかを説明せよ。

略解の指針

ここでは付録の演習についても、使う定理と途中の要点を明示する。細部の証明は自分で埋めることを前提とする。

- **演習A.1** 使う道具: 変数変換と優収束定理の対偶的な見方。 最初の1手: 各 $x > 0$ で $nxe^{-nx^2} \to 0$ を確認し、 $u = nx^2$ の変数変換で $\int_0^\infty f_n(x)\,dx$ を計算する。 途中の要点: $\int_0^\infty nxe^{-nx^2}\,dx = \frac{1}{2}\int_0^\infty e^{-u}\,du = \frac{1}{2}$。 したがって点wise limit は 0 でも積分値は消えない。もし可積分な支配関数 g が存在すれば、 定理A.2.5 により $\int f_n \to 0$ になるはずなので矛盾する。 最終形: 極限と積分が交換できない原因は、 $|f_n| \leq g$ を満たす可積分な支配関数が存在しないことである。
- **演習A.2** 使う道具: Tonelli の定理。 最初の1手: 固定した実数 $x \geq 0$ に対して $x^2 = \int_0^\infty 2t\,\mathbf{1}\{x > t\}\,dt$ を示す。 途中の要点: 被積分関数 $(\omega, t) \mapsto 2t\,\mathbf{1}\{X(\omega) > t\}$ は非負なので、 Tonelli により積分順序を交換できる。 その後は $\mathbb{E}[\mathbf{1}\{X > t\}] = \mathrm{P}(X > t)$ を使うだけでよい。 最終形: $\mathbb{E}[X^2] = \int_0^\infty 2t\,\mathrm{P}(X > t)\,dt$。

- **演習A.3** 使う道具: Radon–Nikodym 導関数の密度比表示。 最初の1手: 両測度がルベーグ測度に関して密度 $f_0(x) = e^{-x}\mathbf{1}\{x \ge 0\}$, $f_1(x) = 2e^{-2x}\mathbf{1}\{x \ge 0\}$ を持つことを書く。 途中の要点: $f_0(x) > 0$ on $[0, \infty)$ なので $P_1 \ll P_0$ であり、 $dP_1/dP_0 = f_1/f_0 = 2e^{-x}$ が成り立つ。 この尤度比は x の単調減少関数なので、 Neyman–Pearson 検定では小さい x が H_1 を支持する。 最終形: $\frac{dP_1}{dP_0}(x) = 2e^{-x}\mathbf{1}\{x \ge 0\}$。 1標本の指数分布検定では、これがそのまま尤度比統計量になる。

- **演習A.4** 使う道具: L^2 射影と共分散計算。 最初の1手: $\mathbb{E}[X \mid Y] = aY$ と置き、 直交条件 $\mathbb{E}[(X - aY)Y] = 0$ から a を決める。 途中の要点: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, 2X + Z) = 2$, $\mathrm{Var}(Y) = \mathrm{Var}(2X + Z) = 5$ だから $a = 2/5$。 したがって $\mathbb{E}[X \mid Y] = (2/5)Y$。 さらに $\mathrm{Var}(\mathbb{E}[X \mid Y]) = (2/5)^2 \mathrm{Var}(Y) = 4/5$ なので、 $\mathbb{E}[\mathrm{Var}(X \mid Y)] = 1 - 4/5 = 1/5$ になる。 最終形: $\mathbb{E}[X \mid Y] = (2/5)Y$, $\mathrm{Var}(X) = 1 = \mathrm{Var}(\mathbb{E}[X \mid Y]) + \mathbb{E}[\mathrm{Var}(X \mid Y)] = 4/5 + 1/5$。

- **演習A.5** 使う道具: スコア関数の直接計算と正則条件。 最初の1手: $\ell(p; x) = x \log p + (1 - x) \log(1 - p)$ を微分して $s(p; x) = x/p - (1 - x)/(1 - p)$ を得る。 途中の要点: $s(p; x) = \{x - p\}/\{p(1-p)\}$ なので $\mathbb{E}[s(p; X)^2] = \mathrm{Var}(X)/\{p^2(1-p)^2\} = 1/\{p(1-p)\}$。 また $\ell''(p; x) = -x/p^2 - (1 - x)/(1 - p)^2$ だから $-\mathbb{E}[\ell''(p; X)] = 1/\{p(1 - p)\}$ と一致する。 有限標本空間では期待値は有限和なので、 定理A.6.1 の正則条件は 微分と総和の交換を正当化する形で満たされる。 最終形: $\mathcal{I}(p) = \mathbb{E}[s(p; X)^2] = -\mathbb{E}[\ell''(p; X)] = 1/\{p(1 - p)\}$。

A.10　本付録のまとめと本文への対応

以下の表は、本付録の各定理が本巻のどの章で （明示的にまたは暗黙に）使用されるかを示す。

定理	主な使用箇所	目的
単調収束定理	第1章（期待値の構成）	非負関数の積分と極限の交換
Fatou の補題	第5章（リスクの下界）	非負関数の積分の下界評価
優収束定理	第3章, 第6章	微分と積分の交換の正当化
Borel–Cantelli	第2章（強大数の法則）	概収束の証明
Fubini の定理	第3章, 第5章	積分順序の交換
Radon–Nikodym	第3章, 第4章	密度関数・尤度比の存在
微分と積分の交換	第3章, 第6章	正則条件、Fisher 情報量
L^2 射影定理	第1章（条件付き期待値）	最良予測の最適性

表 A.1: 測度論の定理と本巻での使用箇所

参考文献

Durrett, R. (2019). *Probability: Theory and Examples* (5th ed.). Cambridge University Press. ——測度論的確率論の標準的教科書。本付録の各定理の完全な証明を含む。

Billingsley, P. (2012). *Probability and Measure* (Anniversary ed.). Wiley. ——測度論と確率論を一体的に扱う古典。Fubini の定理と条件付き期待値の厳密な議論に優れる。

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press. ——コンパクトで読みやすい測度論的確率論の入門書。条件付き期待値と L^2 理論の解説が明快。 Doob のマルチンゲール理論を用いて、強大数の法則や Lévy's upward convergence theorem などの主要な結果を証明している。

Schilling, R.L. (2017). *Measures, Integrals and Martingales* (2nd ed.). Cambridge University Press. ——ルベーグ積分から始める測度論の教科書。演習問題が豊富で自習に適する。

索引