応用を目指す

数理統計学*

国友直人

東京大学経済学部・経済学研究科

2014 年 8 月 (1st Version)
2014 年 10 月 (2nd Version)
2014 年 12 月 (3rd Version)

*K-14-12, この原稿は朝倉書店「統計解析シリーズ」の中の 1 冊として出版される予定である。

目次

はしがき

この書物は朝倉書店「統計解析シリーズ」の中の基礎的な内容である「数理統計学」についての 1 冊として企画された。タイトルの鍵となる言葉として数理統計学と応用との橋渡しという意味を込めて「(応用を目指す) 数理統計学」という言葉を選んだ。数理統計学についてはしばらく前に標準的理論が確立した後に日本語でも立派な書籍がある中で本書を出版する意味について一言述べておこう。近年では日本の社会においても統計学・統計科学がかなりの注目を浴び、ビッグ・データ時代の先端的数理科学として勉学の意欲が高まっていると思われる。ただし、こうした動きは必ずしも数理統計学の標準的理論そのものの展開に主な注目が集まっているというわけではなく、数理統計的方法が様々な学問分野や社会・経済においてより広く応用できる可能性、現代の課題に向き合う有力な手段として注目されているのであろう。そうした中で数理統計学の研究内容も日々より高度に深化し続けているわけであるので、本書では数理統計学における新しい動き、現代社会・経済など様々な分野における応用の動きなどを配慮しつつも現代の数理統計学において基礎的内容と考えられる事項について述べることにした。

ここで簡単に本書の内容を説明しておこう。本書は内容を第 1 部「確率空間と確率分布」、第 2 部「数理統計の基礎」、第 3 部「数理統計の展開」という 3 部構成とした。第 1 部ではまず数理統計学の議論に必要な確率論をやや現代的側面を含めて数理統計学で必要な範囲に限り基礎的内容を説明した。第 1 章は確率測度と確率変数、第 2 章は確率分布と期待値、第 3 章は条件付期待値と独立性である。こうした標準的な内容に加えて第 4 章では確率変数とマルチンゲール、第 5 章では分布収束と中心極限定理についてやや立ち入って説明した。これらの内容は現代的な数理統計学やその応用において重要事項であり、様々な応用可能性を検討する上で有用と判断したからである。第 2 部は数理統計の基礎部分であるが、まず第 6 章では統計量と標本分布を説明

した。統計的推測論については第7章では推定論、第8章では検定論、第8章では統計的決定理論についてそれぞれその基礎的内容を説明した。第3部は展開部であるが、第10章で統計的関係の分析、第11章で近年での応用における統計解析につながる幾つかの話題を選んで説明した。

本書は「数理統計学とその応用」に関心のある学生を対象として書かれたが、より広く統計学・統計科学の理論と応用に関心のある関係者にとり何らかの役に立てば幸いである。数学や数理科学を専攻しているとは限らない多くの学生・院生が統計学・数理統計学を勉強、応用することを主な念頭においたので、本書では数理統計学の議論に利用する数理的内容についても部分的にではあるが付論などを含めて説明した。既に大学における基礎的数理を十分に習得済みの諸君はそうした内容を読み飛ばすことは可能である。ただし筆者は長年にわたり経済学部・経済学研究科の学生・院生の教育に関わっているので、大学においても十分に基礎数理について学ぶ機会がないままに (あるいは避けてきた) 応用分野を専攻している学生・院生諸氏にとって勉学の参考に有益、ではないかと考えて直観的説明などを交えて幾つかの内容を挿入した。むろん言及した議論の数理的細部を完全に理解には時間的余裕のある他の機会が必要であることを念頭に当面の勉学に資することを期待したい。

なお紙数の制約などで省略した幾つかの補足的な論点や練習問題の解答へのヒントなどは朝倉書店の HP
http://.....
などを通じて多くの読者にとり容易に手に入れることができることにした。また最後になるが。本書は著者が東京大学経済学部3・4年生を対象として行った講義「数理統計」の内容をもとに幾つかの項目を追加してまとめたことに言及しておこう。講義に参加し、質問やコメントを提供してくれた学生諸氏に感謝したい。また本書に掲載した幾つかの例の図の作成に助力してくれた元・現役の院生諸氏、三崎広海、

江原輩夫、来栖大輔、池田祐樹 (敬称略) の諸氏にはこの場を借りて感謝する。さらに滞りがちな著者に対して絶えず適切なプレッシャーを与えてくれた朝倉書店の川口達也氏のご努力に感謝する。

2014 年 12 月
国友直人

あとがき

本書では日本の大学の専門学部から大学院初級にかけての統計学・数理統計学の教育における基礎的内容について著者なりに一通りカバーしたつもりである。国際的には数理統計学の近年での研究・教育における展開は以前にも増して大きな広がりがある。国際的には専門の統計学科は通常の大学には存在し、幅広い統計学・数理統計学の教育が行うことが可能であるが、時関数が限られた日本における大学教育では数理統計学の基本を一通りカバーする必要があるので限られた分量の書籍では限界がある。本書では応用を念頭においているので近年での話題につながる有益と判断できる内容を幾つか挿入したが、事後的にはより多くの説明が望ましいことも少なくない。ここではそうした内容の多くについては本シリーズの書籍を含めて他の機会に譲ることとした。むろん近年に著しく展開している大部分の最新の議論も数理統計学の基本的内容の上に発展しているので本書のような教科書が存在することにもなにがしかの意義があると考える。

著者にとっては数理統計学には古くからのお馴染みの分野であるが、著者自身はどちらかというと数理統計学を利用した応用統計、特に経済・経営・金融に関わる経済統計学、計量経済学、計量ファイナンス、などを主な研究分野という事情から厳密な意味では数理科学としての数理統計学の専門家とは云えない。ただしスタンフォード大学 (経済学科・統計学科) の大学院生だった頃から現在に至るまで絶えず数理統計学の議論に親しんできた、と言う意味では数理統計学の研究・教育には深い関心がある。また東京大学経済学部において「数理統計」と言う講義を何回か担当することになったので、(応用に向けた) 数理統計学についてはまとめてみる良い機会となった。本書をまとめるにあたって、学部時代にお世話になった竹内啓先生、鈴木雪夫先生、大学院生の時の指導教官であり統計的多変量解析や時系列解析を中心として数理統計学の大家である T.W.Anderson 教授には特に感謝したい。また最後になるが休日中にも執筆の時間を費やすことに常に協力して

くれた家族にも感謝したい。

2014年12月
国友直人

Part-I：確率空間と確率分布

1　確率測度と確率変数

1.1 確率の源流

確率概念の源流としてはしばしば賭 (ギャンブル) の計算が挙げられる。確率論の歴史書では 17 世紀のフランスの哲学者 Pascal と数学者 Fermat の書簡集 [1] が著名であるが、場合の数に基づくサイコロ・ゲームでの「ルールの公平性」や分配問題が議論されている。例えばある賭博師は素人のお客を相手にうまく立ち回れたのか、また賭ゲームを中途で止めなければならないとき、掛け金の「公平な (fair)」分配」はどの様に考えるべきであろうか？この種の問題は東京の街角で見かける年末ジャンボやグリーン・ジャンボ宝くじの販売の宣伝ばかりでなく、倒産企業の価値の分配方式など現代の企業社会に通じる面も少なくない。古典的な賭ゲームで登場する確率は高校数学で出ている「組み合わせ確率」である。これは有限個の「同等に確からしい基本事象」より求められる。

統計学の分析対象として自然科学、工学や農学、医学、そして経済・経営・金融など社会科学をはじめ応用分野で登場する問題では事前に結果が確定できない現象の確からしさを「同等に確からしい基本事象」より確定できない方がむしろ一般的である。そこで多数回の試行が繰り返される事象の相対頻度およびその極限として直観的に想定できる数値を「経験確率」として説明する事も少なくない。後者の経験確率の難点は有限回の試行を見ただけでは確率そのものが一意に定まらないことである。その他、例えば明日の株価についての人々の予想を比率で表そうとして利用している確率は、主観 (subjective) 確率、あるいは個人 (personal) 確率 [2] と呼ばれているが、不確実性下における人々の行動を理解しようとする経済・経営の分野では特に重要な役割をはたしている。なお、主観確率に類似の概念として Keynes が提唱した論理確率 [3] なども挙げることができる。

無限個の事象

同等に確からしい二つの基本事象 $\{H\}$(表), $\{T\}$(裏) としてコイン投げを考えよう。いまコイン投げを非常に多く繰り返している状況を想定し、さらに事象 $A = \{$ いつ

[1] "Lettre de Monfieur Pascal $\grave{a}$ M. de Fermat"(1654).

[2] 例えば Savage, L. (1954) "The Foundations of Statistics,"Dover が代表的である。

[3] Keynes (1921) "A Treatise of Probability," in The Collected Writings of John Maynard Keynes, Cambridge University Press.

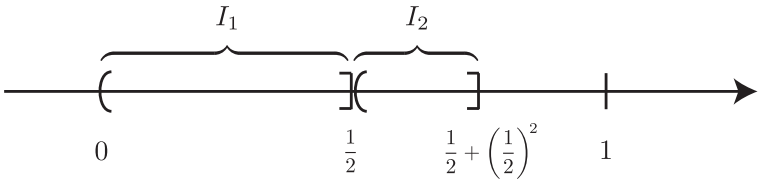

図 1-1:$(0,1]$ の分割

か表が出る } としよう。また $B_n = \{n$ 回目にはじめて表が出る事象 $\}$ とする。事象の系列 $B_1, B_2, \cdots$ は互いに背反なので $(B_i \cap B_j = \emptyset\ ,\ i \neq j, \emptyset$ は空集合) 高校数学の確率では集合 $A_n = \{n$ 回目までに表が出る事象 $\}$ の確率は

$$P(B_1) + P(B_2) + \cdots + P(B_n) = \frac{1}{2} + (\frac{1}{2})^2 + \cdots + (\frac{1}{2})^n = 1 - (\frac{1}{2})^n$$

となる。ここで繰り返しの回数を多くすると

$$P(A) = \lim_{n\to\infty}[1 - (\frac{1}{2})^n] = 1 \tag{1.1}$$

とするのが適当と考えられる。ここで集合 $A_n = \{n$ 回目までに表がでる事象 $\}$ とすると $A_1 \subset A_2 \subset \cdots$ より $A_n \longrightarrow A$ という集合上での確率を考察する必要が生じるわけである。

さて例で説明した「無限回のコイン投げ」におけるある事象 A は実際の話とは関係がない架空の事であろうか?例えば経済・経営での身近な例として、世界中の株価に関係するポートフォリオの価値が 24 時間以内に上下いずれかに変化する事象を考えよう。$24 \times \frac{1}{2}$ 時間以内に変化する可能性があり、$24 \times [\frac{1}{2} + (\frac{1}{2})^2 + \cdots + (\frac{1}{2})^n]$ $(n \geq 2)$ 時間以内に変化する可能性など、幾らでも考えられる。こうした 24 時間以内(実は任意の時間内)に価値が変化する事象に関連して確率やリスクを定義、あるいはリスクを計測し制御できることが社会的にも必要となっているのが現代の経済であろう。例えば新聞やテレビ・ニュースなどで報じられる近年の金融市場(外国為替・株式・国債など)の動きを見ると、「無限回のコイン投げ」における事象の確率を論じることは意味がある。サイコロ、トランプと言った説明に都合のよい例とは限らない実際の応用例では「基本事象が数えられないほど多く存在する」と見なすことによりかえって明快な分析が行えることも少なくない。ここで不確実性を表現する集合やその元が無限の可能性を考慮すると、無限の可能性を考慮した数理的な議論が必要となり組み合わせ数に基づく議論は数理的に破綻する。そこでコルモゴロフ [4] に始まる公理的確率 (axiomatic probability) の枠組みが必要となるのである。他方、現代的な確率論自体は洗練され厳密ではあるが、数理科学を専攻する専門家を除くとやや近づきがたい印象を与えている。ここでは様々な対象に対して統計分析を応

[4]Kolmogorov, A. (1933)「確率論の基礎概念」,(坂本訳) 筑摩書房.

用するときに生じる問題の解決に欠くことができない数理的議論の基礎を出来るだけ簡潔に要約してみよう。

1.2 確率空間

不確実性が付随する様々な事象からなる $\mathbf{\Omega}$(全事象とも呼ばれる)、その元 (基本事象)ω とおき、その上で確率測度 P を導入しよう。サイコロも目が基本事象なら確率測度は幾つかの基本事象の集合、例えば { 偶数の目 } と言った集合においても確率が定まる必要がある。そこで集合からなる集合族を考察する。ここで重要な役割を担うのは σ **加法族** (Ω 上の σ-field, $\mathcal{F}$ **シグマ集合体**) と呼ばれる集合族である。

定義 1-1: 次の 3 条件を満たすものを σ**-加法族**と呼ぶ。

1. $\mathcal{F} \ni \phi$,
2. $\mathcal{F} \ni A \Rightarrow \mathcal{F} \ni A^c$,
3. $\mathcal{F} \ni A_i \ (i = 1, 2, \cdots) \Rightarrow \mathcal{F} \ni \bigcup_{i=1}^{\infty} A_i$.

ここで ϕ は空集合、和集合 $A_1 \cup A_2 = \{\omega | \omega \in A_1 \text{ or } \omega \in A_2\}$ であるが、条件 3 では可算無限個までの和集合をとることが重要である。よく知られたその他の集合演算としては、A^c は A の補集合、積集合 $A_1 \cap A_2 = \{\omega | \omega \in A_1 \text{ and } \omega \in A_2\}$ などがあるが、集合の演算 $(A_1 \cap A_2)^c = A_1^c \cup A_2^c\}$ などもよく利用する。

例えば 6 個の目がでる可能性があるサイコロの目の例では、$A_1 =$ { 目が偶数 }、$A_2 =$ {1 以外の目 } などという集合を考える必要がある。ルーレットの例では角度として全事象 $\mathbf{\Omega} = [0, 2\pi)$ をとれば事象として任意の区間をとる必要があり、有限個では収まらない。よく利用される σ-集合族としては次の例を挙げておく。

例 1-1 : Borel Field (ボレル集合族) とは記号 $\mathcal{B}(\mathbf{R}^p)$ ={ 集合 $(a_1 < x_1 \leq b_1] \times \cdots \times (a_p < x_d \leq b_p]$ を含む $\mathbf{R}^p$ 上の最小の σ-加法族 } で表される。

例えば $\mathcal{B}(\mathbf{R}^1) = \mathcal{B}(\mathbf{R})$ とすると任意の区間 $(a, b]$ $(a < b)$ を含むが、さらに $(a, b] \cup (c, d]$,$(a, b] \cap (c, d]$ などの集合も含む。$p = 2$ のとき $\mathcal{B}(\mathbf{R}^2)$ は二次元領域 $(a_1, b_1] \times (a_2, b_2]$ $(a_1 < b_1, a_2 < b_2)$ を含むが、さらに $(a_1, b_1] \times (a_2, b_2] \cup (c_1, d_1] \times (c_2, d_»]$ などの集合を含む。

ここで任意の $x \in \mathbf{R}$ に対して $(-\infty, x] = \bigcap_{n \in \mathbf{N}} (-\infty, x + n^{-1})$ ($\mathbf{N}$ は自然数全体) なので、開集合全体から生成される (最小の)σ-集合族を意味する。ここで $\mathcal{F}_1$ が $\mathcal{F}_2$ より小さいとは $\mathcal{F}_1 \subset \mathcal{F}_2$ を意味する。"最小の..."とする意味はあまりにも大きな集合族を扱う可能性を排除するためである。$p = 1$ のときこれは 1 次元の区間及び区間の和集合を全て含むような集合全体を意味する。以下では標本空間 $\mathbf{\Omega}$ が与えられた

時の $\sigma-$ 集合族としては常に最小の $\sigma-$ 集合族をとりこの「最小の」という言葉は省略する。

次に σ-加法族上で確率測度を考えよう。現代的な確率論では次の 3 条件を持たす集合上の関数を確率測度 () と呼ぶ。

定義 1-2 : 次の 3 条件を満たすものを**確率測度** P(probability measure) と呼ぶ。

1. $\mathcal{F} \ni A \Rightarrow P(A) \geq 0$,
2. $P(\Omega) = 1$,
3. $\mathcal{F} \ni A_i \ (A_i \cap A_j = \phi) \Rightarrow P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

ここで古典的な組み合わせ確率との相違は最後の可算加法性に表れていることに注意する必要がある。組み合わせ確率では有限個の和 (有限加法性,finite additivity) を考察するので、実は 2 つの排反事象の和集合についての条件だけでよい。最初に言及した例でも明らかなように多くの応用問題でも可算加法性 (countable additivity) までは必要であり、有限加法性 (finite additivity) から可算加法性 (countable additivity) まで矛盾なく確率を考察できるのか、という論点は測度論 () における拡張定理 (補論 1-A を参照) と呼ばれる議論に重なる。結果は肯定的なので、応用上は有限の場合と無限の場合は矛盾無く扱える。

定義 1-3 : (、略して r.v.) は、任意の $B \in \mathcal{B}(\mathbf{R})$ に対し $\mathbf{X}^{-1}(B) \in \mathcal{F}$ となる ω の写像 (or 関数)$X(\omega)$ $(X : \Omega \mapsto \mathbf{R})$ である。

さらに p 個の確率変数 $X_i(\omega), i = 1, \cdots, p$ を並べて構成した (p 次元) 確率変数ベクトル $\mathbf{X}(\omega)$:

$$\mathbf{X}(\omega) = \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_p(\omega) \end{pmatrix} : \Omega \mapsto \mathbf{R}^p$$

とは写像の逆像が条件

$$\mathbf{X}^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{B}(\mathbf{R}^p)$$

を満足することで定よう。こうした定義した確率変数、確率変数ベクトルの定義に初めて接してその表現に戸惑う諸君も少なくないだろう。要するに（$p = 1$ のとき) 基本事象 ω の関数 $X(\omega)$ が実数値をとるとすると、対応する元の事象の集合 $X^{-1}(B) = \{\omega | X(\omega) \in \mathcal{B}(\mathbf{R})\}$ が $\mathcal{F}$ の元であり確率測度が定義できることを意味している。定義域が事象とその集合なのでそうした関数が常に考え得るか否かは自明でないこともあるが、応用上では集合の関数と考えて差し支えない。なお、測度論ではこの条件を満足するものを可測関数 (measurable function) という特別の名前

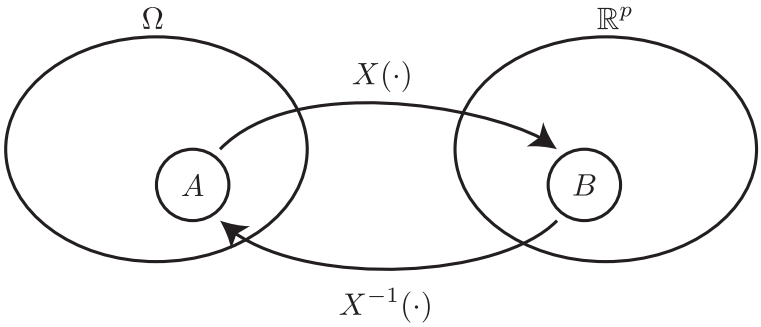

図 1-2:確率変数

をつけている。

例 1-2：サイコロの例では $\mathbf{\Omega} = \{\omega_1, \cdots, \omega_6\}$ とすると $X(\omega) = i$ (if $\omega = \omega_i$) とするのが自然である。

例 1-3：時刻 $t = 0$ から時刻 T までの倒産時刻を表現してみる。この例では集合 $\mathbf{\Omega} = [0, T]$ 上での確率変数を $X(\omega) = t/T$ $(0 \leq t \leq T)$ とすると $[0, 1]$ 上の値を確率変数の逆像上の確率測度と見なせる。

一般に

$$Q(B) \equiv P(\mathbf{X}^{-1}(B)) \quad \forall B \in \mathcal{B}(\mathbf{R}^p)$$

によりボレル集合族上に測度 $Q(\,\cdot\,)$ を定めよう。Q についての条件を調べてみると、確率測度の 3 条件を満たしていることが確認できる。したがって確率変数が定義されると元の事象まで遡ることなく $\mathbf{R}^p$ 上の確率を扱える事が分かる。

1.3 確率分布関数

定義 1-4 : **(累積)**(c.d.f.) は、任意の実数 $x \in \mathbf{R}$ に対する関数 $F(x) = P(\omega | X(\omega) \leq x)$ で定める。

確率分布関数 $F(\,\cdot\,)$ は性質

1. $\lim_{x \to -\infty} F(x) = 0,\ \lim_{x \to +\infty} F(x) = 1$,
2. $F(x)$ は i.e. $\lim_{y \downarrow x} F(y) = F(x)$,
3.

を満たす。逆にこれら 3 条件は分布関数の必要十分条件である。こうした事実は次のような方針で示すことができる。条件 (1) の前半は

$$\lim_{x \to -\infty} \{\omega | X(\omega) \leq x\} = \bigcap_{x \in \mathbf{R}} \{\omega | X(\omega) \leq x\} = \phi \,, \tag{1.2}$$

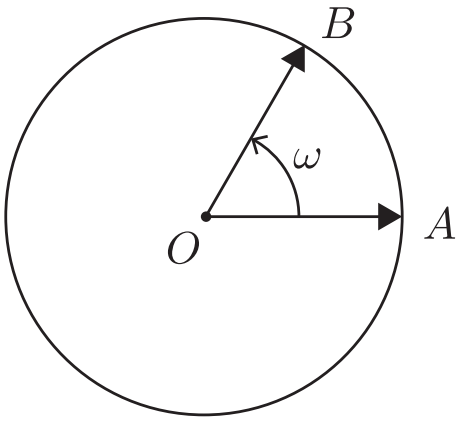

図 1-3:ルーレットの確率モデル

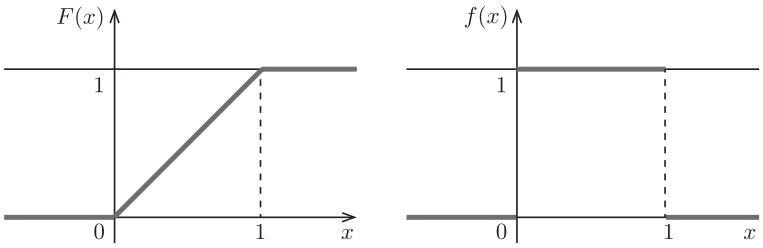

図 1-4:一様分布 ($X(\theta) = \theta/2\pi$)

を利用すればよい。条件 (2) は $y \downarrow x$ のとき (右からの極限の意味であり $y \to x + 0$ と表記することもある) のとき

$$F(y) - F(x) = P(\omega | x < X(\omega) \leq y) \to P(\phi) = 0 \tag{1.3}$$

より得られる。

統計学では様々な確率変数が登場するが、分類すると離散分布 (discrete distribution) と連続分布 (continuous distribution) が挙げられる。例えば (サイコロの目) により定義される確率分布は離散分布であるが、実数軸上で 6 個の点においてジャンプ (跳躍) がある。ルーレットの角度 θ より確率変数 $X(\theta) = \theta/(2\pi)$ と定めると確率変数 X の分布はジャンプを持たない連続分布となる。この確率分布関数はルーレットの角度が止まりうる弧の長さに比例する一様分布 (uniform distribution) であり、$F(x) = x$ $(0 \leq x \leq 1)$ となる。こうした連続分布の場合には連続点で微分可能であり、特に一様分布の場合には密度関数 (density function) は

$$f(x) = \frac{dF(x)}{dx} = 1 \ (0 \leq x \leq 1) \tag{1.4}$$

となる ($f(x) = 0$ (その他))。

応用上でしばしば登場する標準正規分布の密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{1.5}$$

で与えられる。この密度関数を $\phi(x)$ とすると標準正規分布の分布関数は

$$\Phi(x) = \int_{-\infty}^{x} \phi(z) dz \tag{1.6}$$

を表現する。
数理的には一般に離散分布の確率分布 $F_d(x)$, 密度関数を持つ分布 $F_c(x)$, の他に特異な確率分布 $F_s(x)$ を用いて

$$F(x) = \lambda_1 F_c(x) + \lambda_2 F_d(x) + \lambda_2 F_s(x) \tag{1.7}$$

と表現できる。(ただし $\lambda_i \geq 0, \lambda_1 + \lambda_2 + \lambda_3 = 1$ である。特異な分布は通常の応用では表れない。)

次に複数 p 個の確率変数により構成される確率変数ベクトル $\mathbf{X}(\cdot)$ から定められる多次元確率分布関数を導入しよう。

定義 1-5：**(同時) 確率分布関数** (c.d.f.) は、任意の実ベクトル $\mathbf{x} = (x_i) \in \mathbf{R}^p$ に対する関数

$$F(\mathbf{x}) = P(\omega | X_i(\omega) \leq x_i, i = 1, \cdots, p) \tag{1.8}$$

で定める。

この関数を p 次元同時確率分布関数 ($\mathbf{x}$ は $p \times 1$ 実ベクトル) と呼ぶ。多次元分布となる条件は各要素について条件 1 と条件 2 は例えば $F(x_1, \cdots, -\infty, \cdots, x_p) = 0$, $F(+\infty, \cdots, +\infty) = 1$ であるが、条件 3 は修正する必要があり、任意の $\mathbf{h} = (h_1, \cdots, h_p)' \in \mathbf{R}^p$ に対して $\lim_{\mathbf{h} \downarrow \mathbf{0}} F(\mathbf{x} + \mathbf{h}) = F(\mathbf{x})$ となる。例えば $p = 2$ なら条件として

$$\begin{aligned} \Delta_h F(\mathbf{x}) &= F((x_1 + h_1, x_2 + h_2)) - F(x_1 + h_1, x_2) - F(x_1, x_2 + h_2) \\ &\quad + F(x_1, x_2) \geq 0 \end{aligned}$$

が必要となる。

連続型の確率変数ベクトルの場合には多次元確率分布関数より密度関数 $f(\mathbf{x})$ は

$$f(\mathbf{x}) = \frac{\partial^p f(\mathbf{x})}{\partial x_1 \cdots \partial x_p} \tag{1.9}$$

により定める。

確率空間 (probability space)

ここで導入した 3 つの要素 $(\Omega, \mathcal{F}, P)$ を合わせて**確率空間**と呼ぶことにすると数学的には測度論の中の有限測度の議論に対応していることになる。経済学や工学などでは直観的に $\mathcal{F}$ を情報 (information) と呼ばれること少なくないが、応用からの要請として時間とともに情報が変化する場合を扱おうとするので、確率論は抽象的な測度論の議論そのものより内容は豊富になり応用も広がる。

ここで数理的にはこれからの議論は確率空間が完備確率空間 (complete probability space) である必要がある。これはある集合 $A \in \mathcal{F}$ が $P(A) = 0$ であるからといって別の集合 $B \subset A$ に対して $P(B) = 0$ となるとは限らない（完備とは限らない）からである。P の下では一般には確率 0 の集合の部分集合 $B \subset A = \{\omega \mid P(A) = 0\}$ が Ω に入っているとは限らなければ、確率 0 の部分集合を含んだ確率空間とすればよいが、これは確率空間が完備でなければ**完備化** (completion) すればよいことを意味する。

1-A 補論：拡張定理とその周辺

高校や大学における「教養の統計学」に登場する確率は有限加法性に基づいている。これを初等確率測度 (elementary probability)p と呼ぶと本章で定義した確率測度 (probability measure) との関係が気になる。応用上にも重要な数理的事実として測度論 (measure theory) では次のことが知られているが、関連する測度論的事項については例えば伊藤清 (1991) を参照されたい。ここで引用したのは定理 2.4 である。

定理 A-1(確率測度の拡張定理) : 集合 $\mathbf{\Omega}$ 上の $\sigma-$ 集合族 $\mathcal{F}$, $\mathcal{F}$ 上の初等確率測度 p、$\sigma(\mathcal{F})$ を最小の $\sigma-$ 集合族とする。このとき $\sigma(\mathcal{F})$ 上の確率測度まで拡張可能な必要十分条件は任意の互いに素な $\mathcal{F} \ni A_i$ $(A_i \cap A_j = \phi)$ に対し

$$p(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i) \ . \tag{1.10}$$

となることである。

お茶の時間 (Tea-Time)
高校や大学教養課程を超えて大学上級・大学院・研究における言語とは少しギャップがあると感じる学生は少なくない。著者もその一人であったので経験から、これか

ら確率論・統計学などを本格的に勉強したい諸君に参考文献を挙げておく。応用分野の研究を始めてだいぶ経ってから、 志賀浩二「ルベーグ積分 30 講（朝倉)」 を読んだところ、数理的に抽象化された集合・測度論の議論を直観的に説明しているので舞台裏をかいま見た気がした経験がある。伊藤清三「ルベーグ積分入門」（裳華房） は現在でも本格的な教科書だろう。また確率解析学の創始者による教科書 伊藤清「確率論」（岩波書店）1991 年、さらに確率論の基礎を巡る様々な問題のために Billingsley (1994)「Probability and Measure」John-Wiley を挙げておく。

2　確率分布と期待値演算

2.1 期待値

確率を応用する場では確率測度や確率分布と並んでより単純で縮約された情報である期待値 (expectation)、あるいは平均 (mean) と分散 (variance) と言った数値を利用することが多い。離散型確率変数 X の期待値は $P(X(\omega) = x_i) = p_i\ (> 0)$ のとき

$$\mathbf{E}[X] = \sum_i x_i p_i\ , \tag{2.1}$$

連続型確率変数 X の期待値は密度関数 $f(x)$ とすると

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{2.2}$$

と定義されることが多い。期待値は（確率分布の）平均値とも呼ばれ、分散や共分散などとともに基本的な量である。確率変数が離散型の場合と連続型の場合を同時に扱いつつ、確率変数列 $X_n(\omega)$ の期待値 $\mathbf{E}[X_n]$ の挙動を見通し良く議論するには高校〜大学初級で議論する積分（リーマン積分 (Riemann 積分) と呼ばれている）を拡張する必要がある。例えば確率変数列 X_n を $X_n = n^2$ (確率 $1/n$), $X_n = 0$ (確率 $1 - 1/n$) により定める。このとき $P(\omega | \lim_{n\to\infty} X_n = 0) = 1$, 他方、確率変数列の期待値は $\lim_{n\to\infty} \mathbf{E}[X_n] = +\infty$ となる。こうした不確実性下での収束の意味の説明は 4・5 章で議論する。そこでここでは図 2-1 で例示するようなルベーグ (Lebesgue) 積分の発想より期待値を説明する。図は ω が区間 $[0, 1]$ の任意の実数をとり、ω の関数、すなわち確率変数 $X(\omega)$ の値により面積（すなわち積分）を定める例示であるが、高校〜大学初級では変数 ω をとりうる値を細分して面積を定義していたことを思い出すとよい。

$(\Omega, \mathcal{F}, P)$ 上で 1 次元の確率変数 $X(\omega)$ の期待値 (すなわち確率分布の平均値)

$$\mathbf{E}[X(\omega)] = \int_\Omega X(\omega) P(d\omega) \tag{2.3}$$

を定めよう。(2.1) では $p_i = P(\omega | X = x_i) > 0$ となりうるが、連続分布では 1 点をとる確率は $P(\omega | X = x_i) = 0$ である。ここでは (i) 非負確率変数 X が有界 $(0 \le X \le M)$ の場合, (ii) 非負確率変数が有界とは限らない場合, (iii) 一般の確率変数の場合, という 3 段階で定義しよう。

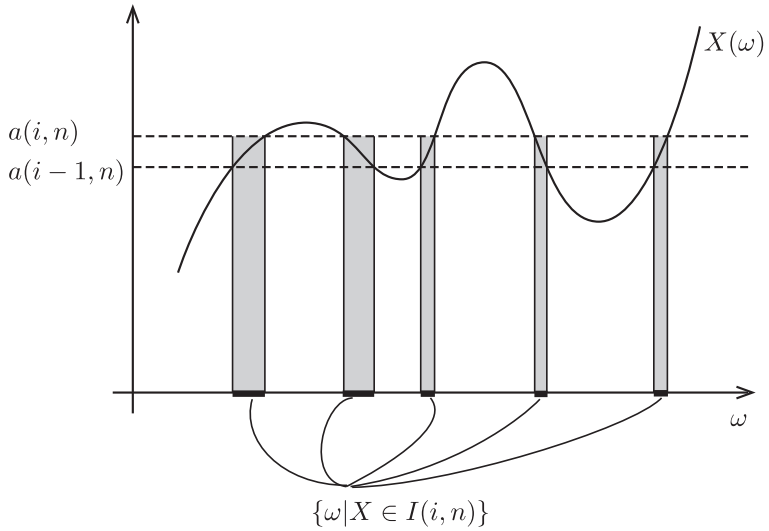

図 2-1:ルベーグ式期待値の直観的理解

(i) $X(\omega)$ のとる値域の区間 $[0, M]$ を $2^n M$ 等分し $a(i,n) = \frac{i}{2^n}(i = 1, \cdots, 2^n M - 1)$ とする $((a(0,n) = 0, a(2^n M, n) = M)$。なお高校で学ぶ初等確率では ω についての分割により直観的に平均値を定義するが、この方式では十分に正確に議論を展開するのが困難である。ここで区間 $I(i,n) = (a(i-1,n), a(i,n)]$ に対し、階段関数の和

$$\begin{aligned} S_n &= \sum_{i=1}^{2^n M} a(i,n) P(X \in I(i,n)) \\ &= \sum_{i=1}^{2^n M} a(i,n) \left[F(a(i,n)) - F(a(i-1,n)\right] \end{aligned}$$

を考察する。ここで S_n と S_{n+1} を比較すると、$0 \leq S_{n+1} \leq S_n$ となるので n について単調非増加列となる。(n についての分割は n が大きくなるとより細かくなる、つまり細分となっている。) したがって $n \to \infty$ のとき極限は必ず存在する。その極限 S を

$$\begin{aligned} \lim_{n\to\infty} S_n = S &= \int_\Omega X(\omega) P(dw) \\ &= \int_{\mathbf{R}} x dF(x) \end{aligned}$$

と表現する。第一の表現を確率測度 P についてのルベーグ (Lebesgue) 積分、第二の表現をスチィルチェス (Stiltjes) 積分と呼ばれる。(ii) 確率変数 X が非有界の場合には有界確率変数 X_M を $X_M = X$ $(X \leq M), X_M = M$ $(X > M)$ とおくと、期待値 $\mathbf{E}[X_M]$ は (i) により定義できる。ここで $\lim_{M\to\infty} \mathbf{E}[X_M] < \infty$ のとき、その極限によ

り $\mathbf{E}[X]$ を定義し、この極限が発散するとき、期待値は $+\infty$ とする。
(iii) 一般の場合には確率変数の期待値は (i) と (ii) に帰着して定める。確率変数 $X = X^+ - X^-$ とおき、X^+, X^- は $X^+ = X$ if $X \geq 0, X^+ = 0$ if $X < 0$; $X^- = -X$ if $X \leq 0, X^- = 0$ if $X > 0$ と定める。一般に確率変数 X の期待値は二つの期待値 $\mathbf{E}[X^+] < \infty, \mathbf{E}[X^-] < \infty$ の時にのみ期待値が有限値として定まる。有限値として期待値が定まるとき

$$\mathbf{E}[X] = \mathbf{E}[X^+] - \mathbf{E}[X^-] = \int_\Omega X(\omega) P(dw) \tag{2.4}$$

てあり積分可能 (integrable) と呼ぶ。絶対値関数は $|X| = X^+ + X^-$ であるので $\mathbf{E}[X]$ のと $\mathbf{E}[|X|] < \infty$ は同値となる。

例 2-1 : 確率変数 X の分布が離散的であるとき $p_i = P(\omega | X(\omega) = x_i) > 0$ $(i = 1, 2, ...)$ とおくと、

$$\mathbf{E}[X] = \int_{-\infty}^{+\infty} x dF(x) = \sum_i x_i p_i \tag{2.5}$$

となる。これは (単調非減少な) 確率分布関数 $F(x)$ は右連続であるから、例えば正の確率がある点 x_1 における左極限 $F(x_1-)$ と右極限 $F(x_1)$ の差 $dF(x_1) = F(x_1) - F(x_1-) = p_1\ (> 0)$ とできるからである。同様に $dF(x_i) = p_1\ (i = 1, \cdots)$, その他の点では $dF(x) = 0$ であるので期待値の定義から導かれる。
例 2-2 : 確率変数 X の確率分布が連続型で密度関数 $f(x)$ を持つ場合には $dF(x) = f(x)dx$ とおけば

$$\mathbf{E}[X] = \int_{-\infty}^{+\infty} x dF(x) = \int_{-\infty}^{+\infty} x f(x) dx \tag{2.6}$$

である。一様分布や正規分布なぢはの典型である。正規分布の場合には密度関数は母数 μ, σ を用いて

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (^\infty < x < +\infty) \tag{2.7}$$

で与えられる。期待値は

$$\begin{aligned} \mathbf{E}[X(\omega)] &= \mathbf{E}[\mu + (X(\omega) - \mu)] \\ &= \mu + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x - \mu) e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \mu + \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} \left[-e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right]_{-\infty}^{+\infty} \end{aligned}$$

という変形より右辺の[5]、第二項はゼロとなるので期待値 $\mathbf{E}[X] = \mu$ となる。

ここで説明したLebesgue積分の表現や用語をすぐに慣れなくても応用上はあまり気にする必要はないが、それは統計学入門書に説明されている期待値の数値を再考する必要がないからである。高校や大学初級の積分は (連続関数についての) Rieman式積分に基づいて説明されるが、ここで導入したのは Lebesgue 式積分の発想であり、「Rieman 積分可能なら Lebesgue 積分可能」という意味で一般に後者は前者を含む。ただし確率・統計を応用する上で表れる確率変数は ω の関数としては連続関数ではないことが一般的であるから数理的に整合的であるためには積分の定義を巡る説明が欠かせない。例えば連続型の確率分布は $F(x) = \int_{-\infty}^{x} f(z)dz$, すなわち (Lebesgue 測度に関する) 密度関数が存在すれば、$\int xdF(x) = \int xf(x)dx$ である。ここで Lebegue 測度とは直線上の長さ、$\mathbf{R}^2$ 上の面積等を一般化した測度 (measure) を意味するが、$\mu(\,\cdot\,)$ とすると積分は $\int xf(x)\mu(dx)$ と表現できる。離散型確率分布では和 $F(x) = \sum_{x_i \leq x} p_i$ と表現できるので、連続型・離散型を問わず期待値は

$$\mathbf{E}[X] = \int_{\mathbf{R}} xdF(x) \tag{2.7}$$

より、(2.1) と (2.2) のように別々に表現しないでよい。

2.2 期待値の一般化

次に確率分布の特性値を表現する為に確率変数 $X(\omega)$ の期待値を含む任意の関数 $g(\,\cdot\,)$ の期待値を定義する。

定義 2-1：確率変数 X は $\mathbf{\Omega} \mapsto \mathbf{R}$ の可測関数、$g(X)$ は $\mathbf{R} \mapsto \mathbf{R}$ の可測関数とする。このとき期待値を

$$\mathbf{E}[g(X)] = \int_{\Omega} g(\mathbf{X})P(d\omega) \tag{2.8}$$

で定める。

確率変数 $X(\omega)$ の期待値 $\mathbf{E}[X]$, 分散 $V[X] = \mathbf{E}[(X - \mathbf{E}(X))^2]$, (r 次モーメント) $\mathbf{E}[X^r]$ $(r = 1, 2, 3, 4, ..., k)$, 期待値周りの r 次積率 $\mathbf{E}[(X - \mathbf{X})^r]$ などが挙げられる。

さらに期待値は多次元の確率変数ベクトルに一般化できる。p 次元の確率変数 $\mathbf{X}(\omega) = (X_i(\omega))$ の (可測) 関数 $g(\mathbf{x})(\mathbf{x} \in \mathbf{R^p})$ に対して期待値

$$\mathbf{E}[g(\mathbf{X})] = \int_{\Omega} g(\mathbf{X})P(d\omega) \tag{2.9}$$

[5] ここで $e^{-x^2/2}$ の微分は $-xe^{-x^2/2}$, $e^{-x^2/2}$ の 2 回微分は $(x^2-1)e^{-x^2/2}$ であり 1 回微分すると x の 1 次関数、2 回微分すると x の二次関数が現われる。(一般には Hermite の多項式と呼ばれる。) ここでは指数関数 e^z の微分、$z = -x^2/2$ の微分を合成（合成関数の微分）を施した。

により同様に定める。具体例としては各要素の分散 $\sigma_{ii} = \mathbf{E}[(X_i - E(X_i))^2]$ や要素間の共分散 $\sigma_{ij} = \mathbf{E}[(X_i - E(X_i))(X_j - E(X_j))]$, 高次の積率 (モーメント) などが挙げられる。

例 2-3 : 期待値が存在しない確率分布の例としてはコーシー分布が挙げられる。密度関数は

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \quad (-\infty < x < \infty) \tag{2.10}$$

である。期待値を評価するために区間 $[-T_2, T_1]$ における積分を評価すると [6]

$$\int_{-T_2}^{T_1} \frac{1}{\pi}\frac{x}{1+x^2}dx = \frac{1}{2\pi}\left[\log(1+x^2)\right]_{-T_2}^{T_1}$$

となる。ここで $T_1, T_2 \to \infty$ のとき値は不定であるので、コーシー分布の期待値は存在しないことを意味する。同様に分散も存在しない。

ここで期待値は存在するが分散は存在しない確率分布を考えられるだろうか?積率 (モーメント) の非存在と分布関数や密度関数の形状はどの様に関係するので、コーシー分布の裾の形状が x^{-2} に近いことがこの問題の一つの鍵である。

2.3 期待値演算の利用

期待値を利用すると確率評価が可能となることがある。よく知られた例として期待値に関するマルコフ (Markov) の不等式を挙げておこう。

定理 2-1: 確率変数 $X(\omega)$, $g(x)$ を非負の実数値関数で $\mathbf{E}[g(X)] < \infty$ とする。このとき任意の正値 a に対し

$$\mathrm{P}\left(\omega | g(X(\omega)) \geq a\right) \leq \frac{\mathbf{E}[g(X)]}{a} \tag{2.11}$$

が成り立つ。

証明: 期待値について評価の範囲を制限すると次の展開より (2.10) が導かれる。

$$\begin{aligned}
\mathbf{E}[g(X)] &= \int_{-\infty}^{\infty} g(x)dF(x) \\
&\geq a\int_{\{x|g(x)\geq a\}} dF(x) \\
&= a\mathrm{P}\left(g(X) \geq a\right) .
\end{aligned}$$

[6] (合成) 関数 $\left[\log(1+x^2)\right]$ の微分が $2x/(1+x^2)$ となることを利用した。

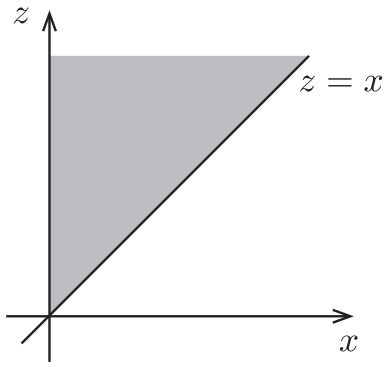

図 2-2: 積分範囲

Q.E.D.

ここで特に関数として $g(x) = (x - \mathbf{E}(X))^2$, $a = k^2V(X)$ と置けば、チェビシェフ (Tchebychev) の不等式

$$\mathrm{P}\left((X - \mathbf{E}(X))^2 < k^2V(X)\right) \geq 1 - \frac{1}{k^2}\,. \tag{2.12}$$

が得られる。
例えば k=3 とすると任意の確率分布について

$$\mathrm{P}\left(|X - \mathbf{E}(X)| < 3\sqrt{V(X)}\right) \geq \frac{8}{9}$$

が得られる。ここで $\sqrt{V(X)}$ は確率変数 X の と呼ばれているが、「ばらつきの指標」としてしばしば利用されている。この不等式は数値的な側面よりもより大数の法則などの分析に有効であることが挙げられる。

次に確率分布の と期待値の関係を考察しよう。確率変数 $X(\omega)$ に対し $|X|$ の分布関数を $F(z)$ とすると、期待値 $\int_0^\infty z df(z)$ が有限であれば $z = \int_0^z 1 \cdot dx$ より積分範囲 $((0 \leq x \leq z, 0 \leq z)$ は $(x \leq z, 0 \leq x)$ を意味する。したがって

$$\begin{aligned}
\mathbf{E}[|X|] &= \int_0^\infty \int_0^z dx dF(z) \\
&= \int_0^\infty \int_x^\infty dF(z) dx \\
&= \int_0^\infty \mathrm{P}\left(|X(\omega)| > x\right) dx\,.
\end{aligned}$$

この評価では 2 次元の平面上での積分範囲を図 2-2 として示しておく。これより期待値が存在する為には裾確率 $P(|X| > z)$ は $|z|$ が大きいとき十分なスピードで減衰する必要がある。同様にさらに高次のモーメント $\mathbf{E}[|X|^r]$ $(r \geq 2)$ が存在するには、

より急速に裾が減衰する必要があることも分かる。

確率分布として多用される正規分布は様々な性質を持つが分布の裾が急速に減衰する性質を持つ。標準正規確率分布 $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$, 密度関数 $\phi(z) = (1/\sqrt{2\pi})e^{-z^2/2}$ は $|x|$ が大きくなると (Mill の比と呼ばれている) 裾確率 $\bar{F}(x) = 1 - \Phi(x)$ と密度関数の比は $\bar{F}(x)/\phi(x) \sim x^{-1}$ と近似が可能である (一般に任意の $x > 0$ に対して不等式 $x^{-1}\phi(x) \geq \bar{F}(x) \geq (x + x^{-1})^{-1}\phi(x)$ が成立する)。裾確率は

$$\bar{F}(x) = 1 - F(x) \cong \frac{1}{x}\phi(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-\frac{1}{2}x^2} \tag{2.13}$$

となるので ($|x|$ の増加とともに) 急速に減衰する。ここで一般にある非負関数が密度関数である為には $|x|$ が大きくなるとゼロに収束する必要がある。この減衰するスピードが正規分布のように指数的でないときにはベキ関数的に減衰することが考えられる。

例 2-4 : 指数分布: 故障や病気などある現象が発生する時刻を確率変数とすると、確率分布として指数分布が用いられることがある。非負の母数 $\lambda > 0$ に対して非負値のみに対する密度関数は

$$f(x) = \lambda e^{-\lambda x} \ \ (x > 0) \tag{2.14}$$

で与えられる。確率分布関数は

$$F(x) = 1 - e^{-\lambda x} \ \ (x > 0) \tag{2.15}$$

となるので期待値は $\mathbf{E}[X] = 1/\lambda$, 分散は $\mathbf{V}[X] = 1/\lambda^2$ である。なお分布を特徴づける母数として $\beta = 1/\lambda$ をとることもある。

例 2-5 : パレート分布: 非負値のみをとる密度関数

$$f(x) = c\frac{1}{x^{\alpha+1}} \ \ (x > x_0 > 0, \alpha > 0) \tag{2.16}$$

で与えられる確率分布はパレート (Pareto) 分布と呼ばれる。ここで定数 $c = \alpha x_0^{\alpha}$ であるが、この分布は経済学では都市の大きさを示す分布[7]、所得や資産の分布がこうしたベキ法則 (Power Law) により表現されることがあるが、なぜパレート分布が有用なのかは興味深い話題である。

[7]例えば金本良嗣「都市経済学」(東洋経済) などに日本の実例がある。

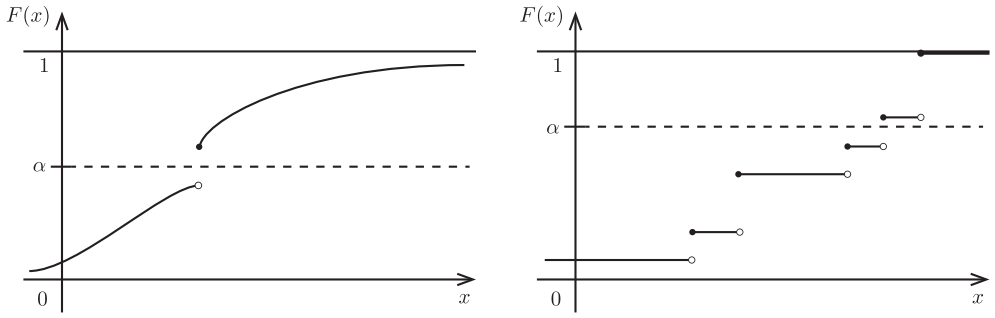

図 2-3: 分位点

ここで確率分布の積率 (moment) ではなくより直接に確率分布を特徴づける別の表現を導入する。

定義 2.2: 確率分布関数 $F(\cdot)$, 任意の α $(0 < \alpha < 1)$ に対して α 分位点 (quantile) とは

$$x_\alpha = \inf_x \{F(x) \geq \alpha\}$$

により定義する（100% 点とも呼ばれている)。

ここで確率分布に不連続点がある場合、離散分布の場合にはここでの分位点の定義が必ずしも自然ではないことを図 2-3 により例示しておく。伝統的な統計学の議論の中では $\alpha = 0.5$ となる中位数 (median), $\alpha = 0.25$ となる四分位点などが利用される。例えばコーシー分布のように期待値が定義できなくとも中位数や四分位点の差などにより確率分布の平均 (center) やばらつき (dispersion) は定義できる。

近年に盛んになっている統計的リスク管理の分野では確率分布の裾は重要な意味を持つ。11 章で再び言及するように例えば近年における内外の銀行業のリスク管理規制の中では金融資産の評価では「保有資産の損失分布 (loss distribution)」の定期的な確率評価が義務づけられている。ここでは分布の 1 パーセント分位点 (パーセンタイル)、5 パーセント分位点などが金融リスク管理上では **VaR**(Value-at-Risk) の基準とされている。

例 2-6：一般化パレート (generalized Pareto) 分布 ：確率分布関数の右裾を表現する為に（実数値をとる ξ を母数として) 右裾確率

$$\bar{F}(x) = 1 - F(x) = (1 + \xi x)^{-1/\xi} \ (\xi > 0, x > 0; \text{あるいは} \xi < 0, 0 < x < -\frac{1}{\xi}) \tag{2.17}$$

を当てはめることがある。これは一般化パレート分布であり 11 章で説明するように稀に起きる事象を数理的に説明するときにある種の妥当性がある。特に $\xi \to 0$ とすると指数分布を含む。

期待値ベクトル・共分散行列

p 次元確率変数ベクトル $\mathbf{X}(\omega) = (X_i(\omega))$ に対し、期待値ベクトルを定義する。多次元分布を扱う時には各要素の期待値を並べたベクトルは

$$\mathbf{E}[\mathbf{X}(\omega)] = \begin{pmatrix} \mathbf{E}[X_1(\omega)] \\ \vdots \\ \mathbf{E}[X_p(\omega)] \end{pmatrix} \tag{2.18}$$

であるのでこれを期待値ベクトルとする。(なお ω はしばしば省略する。) また確率変数ベクトルの第 i 要素と第 j 要素の共分散 (covariance) は

$$\begin{aligned} \sigma_{ij} &= \mathbf{E}[(X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))] \\ &= \int_{\mathbf{R}^p} (x_j - \mathbf{E}(X_i))(x_j - \mathbf{E}(X_j)) dF(x_1, \cdots, x_p) \\ &= \int_{\mathbf{R}^p} (x_j - \mathbf{E}(X_i))(x_j - \mathbf{E}(X_j)) f(x_1, \cdots, x_p) dx_1 \cdots dx_p \end{aligned}$$

により定義する。(ここで $f(x_1, \cdots, x_p)$ は同時密度関数であるが 3 章で説明される。) 一般の p 次元の確率分布が理解しにくければ $p = 2$ のときにより具体的に考えるとよい。) このとき分散共分散行列 (variance-covarince matrix, あるいは共分散行列) を $p \times p$ 行列 (しばしば $\mathbf{V}(\mathbf{X})$ と表記する)

$$\mathbf{\Sigma} = (\sigma_{ij}) = \mathbf{E}[(\mathbf{X}(\omega) - \mathbf{E}(\mathbf{X}))(\mathbf{X}(\omega) - \mathbf{E}(\mathbf{X}))^{'}] \tag{2.19}$$

により定義する。この行列の対角要素 σ_{ii} $(i = 1, \cdots, p)$ は第 i 要素の確率変数の分散 (variance)、$\sqrt{\sigma_{ii}}$ $(= \sigma_i; i = 1, \cdots, p)$ は第 i 要素の確率変数 X_j の標準偏差 (standard deviation) である。

定理 2-2: 2 つの確率変数 $X_i(\omega)$ $(i = 1, 2)$ が $\mathbf{E}[X_i^2] < \infty$ を満たすとする。このとき

$$|\mathbf{E}[X_i X_j]| \leq \sqrt{\mathbf{E}[X_i^2]\mathbf{E}[X_j^2]} \tag{2.20}$$

が成り立つ。

証明: 任意の実数 λ に対して

$$\mathbf{E}[(\lambda X_1 + X_2)^2] = \lambda^2 \mathbf{E}[X_1^2] + 2\lambda \mathbf{E}[X_1 X_2] + \mathbf{E}[X_2^2] \geq 0 \tag{2.21}$$

である。判別式を求めると

$$D = (\mathbf{E}[X_1 X_2])^2 - \mathbf{E}[X_1^2]\mathbf{E}[X_2^2] \leq 0$$

より不等式 (2.19) が得られる。 **Q.E.D.**

この不等式はコーシー・シュワルツ (Cauchy-Schwartz) の不等式と呼ばれている。等号は $D = (\mathbf{E}[X_i X_j])^2 - \mathbf{E}[X_i^2]\mathbf{E}[X_j^2] = 0$ の時に限り成立するが、この条件はある実数 λ が存在し

$$\mathrm{P}\left(\{\omega | \lambda X_i(\omega) + X_j(\omega) = 0\}\right) = 1 \tag{2.22}$$

すなわち確率 1 で二つの確率変数が比例関係にあることを意味する。なお上の結果より 2 つの確率変数 X_i, X_j のピアソン (Pearson) の相関係数 (単に相関係数 (correlation coefficient) と呼ぶことも多い) を

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} \tag{2.23}$$

により定義すると、不等式

$$-1 \leq \rho_{ij} \leq 1 \tag{2.24}$$

が得られる。

なお共分散がゼロの時には無相関と言われる。一般にゼロでない実数の組 a_i $(i = 1, \cdots, p)$ (ベクトル $\mathbf{a} = (a_1, \cdots, a_p)'$) に対して

$$\mathbf{V}[\sum_{i=1}^{p} a_i X_i] = \sum_{i,j=1}^{p} a_i a_j \sigma_{ij} = \mathbf{a}'\mathbf{\Sigma}\mathbf{a} \tag{2.25}$$

であるが、これは実ベクトル $\mathbf{a} = (a_1, \cdots, a_p)'$ に対する（線形代数の用語を用いると）二次形式として表現できる。分散は負の値をとらないことから行列 $\mathbf{\Sigma}$ は非負定符号行列、分散がゼロでなければ確率分布は退化していないので正定符号行列である。こうして共分散行列が実対称行列・非負定符号行列であるから線形代数の表現や議論が利用できる。

特に任意の i, j $(i \neq j)$ に対して互いに無相関であれば $\sigma_{ij} = 0$ $(i \neq j)$ であれば

$$\mathbf{V}[\sum_{i=1}^{p} a_i X_i] = \sum_{i=1}^{p} a_i^2 \sigma_{ii} \tag{2.26}$$

であるから、確率変数の線形和の分散は分散の線形和として簡単な形で表現できる。

2.4 確率分布の例

本節ではよく利用される基本的かつよく知られている幾つかの確率分布について言及しておくが、確率論や数理統計学の応用で重要な役割を演じる主な確率分布を理解しておく必要がある。

例 2-7 (二項分布, binomial distribution)： 離散分布の典型例として二項分布 B(n,p) がある。確率変数列 Z_i $(i = 1, \cdots, n)$ がそれぞれ独立に 1 と 0 を確率 $p, 1-p$ $(= q)$ でとる (n 回のベルヌイ試行) 和 $X_n = \sum_{i=1}^{n} Z_i$ とすると、確率変数 X_n がしたがう確率関数は

$$p(x) = {}_nC_x p^x (1-p)^{n-x} \ (x = 0, 1, \cdots, n) \tag{2.27}$$

で与えられる $(0 \leq p \leq 1)$。二項定理より

$$\sum_{x=0}^{n} p(x) = [p + (1-p)]^n = 1 \tag{2.28}$$

であるので、二項分布の期待値と分散は $\mathbf{E}[X_n] = np, \mathbf{V}[X_n] = npq$ (ただし $q = 1-p$) となる。

例 2-8 (ポワソン分布, Poisson distribution)： 事故や稀な事象の発生件数がしたがう離散分布の例としてポワソン分布 $P_o(\lambda)$ が知られている。確率変数 X がしたがう確率関数は

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \ , \ x = 0, 1, \cdots \tag{2.29}$$

で与えられる $(\lambda > 0)$。期待値と分散は $\mathbf{E}[X] = \mathbf{V}[X] = \lambda$ となる。二項分布 $B(n, p)$ において $np \to \lambda \ (> 0)$ のとき確率関数がポワソン分布に収束するが、この現象は小数の法則 () と呼ばれている。

例 2-9 (一様分布, uniform distribution)： 連続型確率変数 X のしたがう例として一様分布 $U(a, b)$ がある。一様分布の密度関数は

$$f(x) = \frac{1}{b-a} \ \ (a < x < b) \tag{2.30}$$

であり、期待値と分散は $\mathbf{E}[X] = (b-a)/2, \mathbf{V}[X] = (b-a)^2/12$ となる。

例 2-10 (正規分布)： 連続分布として統計的分析ではがもっともよく利用されている。正規分布は $N(\mu, \sigma^2)$ と表される。特に $\mu = 0, \sigma = 1$ のとき、標準正規分布 **(standard normal distribution)** と呼ぶが、密度関数は

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2.31}$$

で与えられる。この関数が密度関数となることは自明ではないが次のように確かめられる。

定理 **2-3:** 標準正規分布の密度関数について

$$\int_{-\infty}^{\infty} \phi(x)dx = 1 \tag{2.32}$$

が成り立つ。

証明: まず $I = \int_0^\infty e^{-x^2/2}dx$ とすると、関数の対称性より $I = \sqrt{\pi/2}$ を示せばよい。重積分

$$\begin{aligned} I^2 &= \int_0^\infty e^{-\frac{x^2}{2}}dx \int_0^\infty e^{-\frac{y^2}{2}}dy \\ &= \int_0^\infty \int_0^\infty e^{-\frac{x^2+y^2}{2}}dxdy \end{aligned}$$

を評価する。極座標変換により $x = r\cos\theta, y = r\sin\theta$ $(0 \le \theta \le \pi/2, r > 0)$ とすると、変数変換のヤコビアンは

$$\mathbf{J} = \left| \begin{array}{cc} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{array} \right|_+ = r(\cos^2\theta + \sin^2\theta) = r$$

となる。したがって, $dxdy = rdrd\theta$ を代入すると

$$\begin{aligned} I^2 &= \int_0^\infty \int_0^{\pi/2} re^{-\frac{r^2}{2}}drd\theta \\ &= \int_0^\infty re^{-\frac{r^2}{2}}dr \int_0^{\pi/2} d\theta = \frac{\pi}{2}\left[-e^{-\frac{r^2}{2}}\right]_0^\infty \\ &= \frac{\pi}{2} \end{aligned}$$

である。**Q.E.D.**

なおここでの積分は広義積分 $\lim_{M\to\infty,N\to\infty}\int_{-N}^{M}\phi(x)dx = 1$ の意味である。また証明の途中に積分の順序交換を行っているが、この操作は積分論における **Fubini** の定理により保証されるが説明は省略した。ここで変換のヤコビアンをあえて利用したが、数理統計学でよく利用される変数変換については章末でも説明したが２次元の図を利用してその意味を考察するとよい。

一般の正規分布として確率変数 X の密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.33}$$

と表現されるが、μ は期待値（または中心の位置)、σ^2 は分散 (または分布のスケール) を意味している。ここで変換により確率変数 $Z = (X-\mu)/\sigma$ とすると Z の確率分

布は **N(0,1)** となる。あるいは $\mathbf{E}[X] = \mathbf{E}[\mu+\sigma Z] = \mu$, $\mathbf{E}[(X-\mu)^2] = \mathbf{E}[\sigma^2 Z^2] = \sigma^2$ となることが分かる。

正規分布の分散は次のように直接に求めてもよい。被積分関数 $x^2 e^{-\frac{x^2}{2}} = (-x)[-xe^{-\frac{x^2}{2}}]$ と分解して微分 $[e^{-\frac{x^2}{2}}]' = -xe^{-\frac{x^2}{2}}$ であることを利用する。部分積分の公式 [8] を利用すると

$$\begin{aligned}\int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}} dx &= [(-x)e^{-\frac{x^2}{2}}]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-1)e^{-\frac{x^2}{2}} dx \\ &= \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}\end{aligned}$$

である。このことより $N(0,1)$ にしたがう確率変数 $\mathbf{X}$ の分散は $\mathbf{V}[X] = \mathbf{E}[X^2] = 1$ と確認できる。

例 **2-11** (多次元正規分布)： 確率変数 Z_i $(i = 1, \cdots, p)$ がそれぞれ **N(0,1)** にしたがい、共分散はゼロ ($\mathbf{E}[Z_i Z_j] = 0$ $(i \neq j)$) とする。ここで p 次元確率変数ベクトル $\mathbf{Z} = (Z_i)$ は同時密度関数

$$\begin{aligned}\phi(\mathbf{z}) &= \prod_{i=1}^{p} \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} \right] \\ &= (\frac{1}{\sqrt{2\pi}})^p e^{-\frac{\sum_{i=1}^p z_i^2}{2}}\end{aligned}$$

を持つとする。(互いに独立であれば **3** 章で説明するように同時密度関数は **1** 次元密度の積になる。) ここで線形変数 $\mathbf{y} = \boldsymbol{\mu} + \mathbf{Cz}$ とすると ($\mathbf{y} = (y_i), \mathbf{z} = (z_i), \boldsymbol{\mu} = (\mu_i)$ はそれぞれ $p \times 1$ ベクトル, $\mathbf{C} = (c_{ij})$ は $p \times p$ 行列)、確率変数ベクトル $\mathbf{Y}$ の分散共分散行列 ($p \times p$ 行列) は

$$V[\mathbf{Y}] = \mathbf{E}[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^{'}] = \mathbf{C}\mathbf{I}_p\mathbf{C}^{'} = \boldsymbol{\Sigma} \tag{2.34}$$

で与えられる。(ただし $\mathbf{I}_p$ は単位行列, $\mathbf{C}\mathbf{C}^{'} = \boldsymbol{\Sigma} = (\sigma_{ij})$ とおいた。) 変換のヤコビアンは $|\mathbf{C}| = |\boldsymbol{\Sigma}|^{1/2}$ であるから (行列式について転置行列について $|\mathbf{C}^{'}| = |\mathbf{C}|$ となることを利用すると、$|\boldsymbol{\Sigma}| = |\mathbf{C}\mathbf{C}^{'}| = |\mathbf{C}||\mathbf{C}^{'}| = |\mathbf{C}|^2$ より) $\mathbf{z} = \mathbf{C}^{-1}(\mathbf{y} - \boldsymbol{\mu})$, $d\mathbf{z} = |\mathbf{C}|_{+}^{-1} d\mathbf{y}$ を代入すると、$\mathbf{Y}$ の密度関数は

$$f(\mathbf{y}) = (\frac{1}{\sqrt{2\pi}})^p |\boldsymbol{\Sigma}|^{-1/2} e^{-\frac{(\mathbf{y}-\boldsymbol{\mu})^{'}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}} \tag{2.35}$$

で与えられる。ここで逆行列について $\boldsymbol{\Sigma}^{-1} = (\mathbf{C}\mathbf{C}^{'})^{-1} = \mathbf{C}^{'-1}\mathbf{C}^{-1}$ を利用した。また指数関数の所は二次形式であり和の記号により $\sum_{i,j=1}^{p}(y_i - \mu_i)\sigma^{ij}(y_j - \mu_j)$; , $\boldsymbol{\Sigma}^{-1} =$

[8] 一般に $\int_a^b f(x)g(x)dx = [f(x)G(x)]_a^b - \int_a^b f^{'}(x)G(x)dx$ である。ここで $G(x)^{'} = g(x)$ であるが、この公式は積の微分より導ける。

(σ^{ij}) と表しても良いが少し煩雑になる。記号が分かりにくい場合には $p=2$ の場合、すなわち二次元正規分布の密度関数を具体的に書いてみると良い。期待値ベクトルは $\boldsymbol{\mu}$, 共分散行列は $\boldsymbol{\Sigma}$ であるので、$\mathbf{p}$ 次元のは $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ と表現される。この方式では $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ を意味する。

特に $p=1$ であれば $Z \sim N(0,1)$ のとき $Y = cZ + \mu \sim N(\mu, \sigma^2), \sigma^2 = c^2$ なので 1 次元正規分布が導ける。

密度関数の変数変換の公式は一見すると複雑であるが、統計学における標本分布の導出に基本的役割を果たすので、簡単な例により操作に十分に慣れる必要がある。

例 **2-12** (対数正規分布) : 正値をとる確率変数 $X = e^Y$ として、その対数変換 $Y = \log X \sim N(\mu, \sigma^2)$ のとき $\mathbf{X}$ は対数正規分布にしたがうと呼ばれる。ヤコビアンは $|J| = 1/x$ なので $\frac{dy}{dx} = 1/x \ (x > 0)$, あるいは $\frac{dy}{dx} = 1/x \ (x > 0)$ があるので密度関数は変換 $x = e^y$ とすると Y が正規分布に従うので変数変換を用いて $f_X(x) = f_Y(x^{-1}(y))\frac{dy}{dx}$ より得られ

$$\phi(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} \tag{2.36}$$

で与えられる。

例 **2-11** では $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ である。(ここで $\mathbf{I}_p$ は単位行列である。) 線形変換 $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{CZ} \ (|\mathbf{C}| \neq 0)$ とすると $\sim N_p(\boldsymbol{\mu}, \mathbf{I}_p)$ であるが逆変換 $\mathbf{Z} = \mathbf{C}^{-1}(\mathbf{Y} - \boldsymbol{\mu})$ より

$$f_Y(\mathbf{y}) = f_Z(\mathbf{C}^{-1}(\mathbf{y} - \boldsymbol{\mu}))|\mathbf{C}|_+^{-1} \tag{2.37}$$

より例 **2.12** の密度関数が得られる。

例 **2-13** (χ^2**(1)** 分布) : 標準正規分布にしたがう確率変数 $\mathbf{X}\sim N(0,1)$ に対して変換された確率変数 $Y = X^2$ の確率分布は χ^2**(1)** (例 **2-16** の特殊ケース) と呼ばれる。この場合にはこれまでの例とは異なり変数変換は **1-1** でないことに注意する必要がある。

例 **2-14** (多項分布, **multinomial distribution**) : 離散分布である二項分布を多次元に一般化してカテゴリーの数を **k+1** とする。確率変数ベクトル $\mathbf{X} = (n_1, \cdots, n_{k+1})'$ の確率関数

$$P(\mathbf{X} = (n_1, \cdots, n_{k+1})') = \frac{n!}{n_1! \cdots n_{k+1}!} p_1^{n_1} \cdots p_{k+1}^{n_{k+1}} \tag{2.38}$$

で与えられるとき多項分布と呼ばれる。ここで $\sum_{i=1}^{k+1} p_1 = 1\ (p_i \geq 0), \sum_{i=1}^{k+1} n_i = n$ である。**k=1** なら二項分布、**k=3** なら三項分布である。

2.5 特性関数

確率変数 X の確率分布を F とする。確率分布の性質を調べる為に統計学ではよく積率母関数 (**moment generating function**) が利用される。この関数は任意の実数 θ に対して期待値 $M(\theta) = \mathbf{E}[e^{\theta X}]$ により定義する。例えば一階の微係数は

$$M^{(')}(\theta)|_{\theta=0} = \frac{dM(\theta)}{d\theta}|_{\theta=0} = \mathbf{E}[X] \tag{2.39}$$

となる。同様に任意の **k** 次積率が存在すれば,

$$M^{(k)}(\theta)|_{\theta=0} = \mathbf{E}[X^k]\ (k = 1, 2, \cdots) \tag{2.40}$$

で与えられるここで $M^{(k)}(\theta)$ は **k** 回の微係数 $\frac{d^k}{d\theta^k}M(\theta)$ を意味する。**k** 回微分可能で微係数が連続のとき $\mathbf{C}^k$ 級の関数と呼ぶが、積率関数は任意の分布関数に対して必ず存在すると言うわけではない。例えばコーシー分布をとると明らかである。

例 **2-**$12^{'}$ (対数正規分布) : 正値をとる確率変数 **X** として、その対数変換 $Y = \log X \sim N(\mu, \sigma^2)$ のとき **X** は対数正規分布にしたがうと云う。この確率分布の積率母関数は存在しないが期待値や分散は存在する。対数正規分布は経済・金融 (ファイナンス) などでは物価や資産価格を表現する分布としてよく用いられている。対数正規分布についても中心極限定理 (**central limit theorem**) は成り立つ。

一般に確率分布の性質を調べるには常にその存在が確認できる特性関数 (**charactereistic function**) を利用する必要がある。

定義 **2-3** : 確率変数 **X** が分布関数 **F** にしたがうとき特性関数 (**chracteristic function**) を

$$\psi(t; X) = \mathbf{E}[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} dF(x) \tag{2.41}$$

で定める。ただし虚数 i に対し $i^2 = -1, t \in \mathbf{R}$ である。

実数値をとる指数関数は

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots \tag{2.42}$$

と展開できる。そこで任意の実数 **x** と $i^2 = -1$(虚数) を代入すると実部・虚部はそれぞれ

$$\begin{aligned} e^{ix} &= 1 + \frac{ix}{1!} + \frac{(ix)^2}{2!} + \cdots + \frac{(ix)^n}{n!} + \cdots \\ &= [1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots] + i[\frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots] \\ &= \cos x + i \sin x \end{aligned}$$

となりが得られる。したがって特性関数は

$$\mathbf{E}[e^{itX}] = \mathbf{E}[\cos(tX)] + i\mathbf{E}[\sin(tX)] \tag{2.43}$$

と表現できるので特性関数は常に存在する [9]。

定理 2-4: 確率変数 **X** の特性関数 $\psi(\cdot)$ は次の性質を持つ。
(i) 任意の $t \in \mathbf{R}$ に対し $\psi(t)$ が存在する。
(ii)$\psi(0) = 1, |\psi(t)| \leq 1$。
(iii)$\psi(t)$ は連続関数。
(iv)$\psi(-t) = \bar{\psi}(t)$(ここで共役複素数は $\overline{a + ib} = a - ib$ で定める)。
(v)a,b を任意の実数とすると $\psi(t; aX + b) = e^{itb}\psi(at; X)$ となる。ここで $\psi(t; Y)$ は確率変数 Y の特性関数を意味する。
(vi)n 次モーメント $\mathbf{E}[X^n] = \alpha_n$ $(n \geq 1)$ が存在すれば

$$\psi^{(k)}(0) = i^k \alpha_k \ (k = 1, \cdots, n) \tag{2.44}$$

が成り立つ

(iii) の証明 ： 定義より

$$\begin{aligned} \psi(t + \Delta t) - \psi(t) &= \mathbf{E}\left[e^{i(t+\Delta t)x} - e^{itx}\right] \\ &= \mathbf{E}\left[e^{itx}(e^{\Delta tx} - 1)\right] \\ &= \int_{-\infty}^{\infty} e^{itx}(e^{i\Delta tx} - 1)dF(x) \end{aligned}$$

[9] ここで $e = 2.718...$ となる定数である。指数関数は微分方程式 $\frac{d}{dx}(e^x) = e^x$(初期条件を含む) を満たす連続関数である。さらに実数値をとる指数関数は複素数をとる関数まで考察すると様々な事項が統一的に議論できる。指数関数と三角関数の関係が典型的で例えばオイラーの公式より $e^{i\pi} = -1$ が導かれる。指数関数の微分 $\frac{d}{dx}[e^{ix}] = e^{ix} = i[\cos x + i \sin x] = -\sin x + i \cos x$ より三角関数の微分公式 $\frac{d}{dx}(\cos x) = -\sin x, \frac{d}{dx}(\sin x)) = \cos x$ が導かれる。指数関数や三角関数は解析関数と呼ばれる振る舞いの良い関数なので直観的計算が正当化される。興味のある諸君は HP(ホームページ) 補論で引用している説明など参考にされたい。

を評価すればよい。ここで関数 $|e^{itx}(e^{i\Delta tx}-1)| \leq 2$ は有界なので **Lebesgue** の収束定理[10] を利用すると

$$\lim_{\Delta t\to 0}|\psi(t+\Delta t)-\psi(t)| \leq \int_{-\infty}^{\infty} |\lim_{\Delta t\to 0} e^{itx}(e^{i\Delta tx}-1)|dF(x) = 0 \tag{2.45}$$

となる。

(vi) の証明 : $k=1$ のとき

$$\lim_{\Delta t\to 0}\frac{\psi(t+\Delta t)-\psi(t)}{\Delta t} = \lim_{\Delta t\to 0}\int_{-\infty}^{\infty}\frac{e^{i\Delta tx}-1}{\Delta t}e^{itx}dF(x) \tag{2.46}$$

である。ここで極限と積分の交換して

$$\lim_{\Delta t\to 0}\frac{e^{i\Delta tx}-1}{\Delta t} = ix$$

を利用する。(より正確には $|e^{ix}-(1+ix)| \leq \min 2|x|$ であるので **Lebesgue** の収束定理を用いる必要がある。) このことより

$$\lim_{\Delta t\to 0}\frac{\psi(\Delta t)-\psi(0)}{\Delta t} = i\ \mathbf{E}[X]$$

を得る。あとは **k** についての帰納法による。 **Q.E.D.**

例 2-15 (積率) : 原点回りの **r** 次積率を $\alpha_r = \mathbf{E}[X^r]$ $(r=1,2.\cdots)$、平均周りの **r** 次積率を $\mu_r = \mathbf{E}[(X-\mathbf{E}(X))^r]$ $(r=1,2.\cdots$、さらに対数特性関数が

$$\log\phi(t) = \sum_{j=1}^{r}\frac{(it)^j}{j!}\gamma_j + o(t^r)$$

と展開できるとするとその係数により **r** 次キュムラント γ_j が定まる。このとき例えば $\gamma_1=\alpha_1$, $\gamma_2=\alpha_2-\alpha_1^2=\mu_2$, $\gamma_3=\alpha_3-3\alpha_1\alpha_2+2\alpha_1^3=\mu_3$, $\gamma_4=\alpha_4-3\alpha_2^2-4\alpha_1\alpha_3+12\alpha_1^2\alpha_2-6\alpha_1^4=\mu_4-3\mu_2^2$ などの関係がある。

例 2-7′ (二項分布) : 二項分布の特性関数は

$$\begin{aligned}\psi(t) &= \sum_{x=0}^{n} e^{itx}{}_nC_x p^x(1-p)^{n-x}\\ &= [pe^{it}+(1-p)]^n\end{aligned}$$

で与えられる。性質 **(vi)** を用いると平均と分散は $\mathbf{E}[X]=np,\ \mathbf{V}[X]=np(1-p)$ となることが確認できる。

[10] 可積分関数列 g_n が g に収束するとき、もし適当な可積分関数 $G(x)$ が存在して $|g_n(x)| \leq G(x)$ のとき $\int_{-\infty}^{\infty}\lim_{n\to\infty} g_n(x)dF(x) = \lim_{n\to\infty}\int_{-\infty}^{\infty} g_n(x)dF(x)$ となる。

例 **2-8**′ (ポワソン分布)： ポワソン分布の特性関数は

$$\begin{aligned}\psi(t) &= \sum_{x=0}^{\infty} e^{itx} e^{-\lambda} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} e^{\lambda e^{it}} = \exp[\lambda(e^{it}-1)]\end{aligned}$$

で与えられる。平均と分散は $\mathbf{E}[X] = \mathbf{V}[X] = \lambda$ となることが確認できる。

例 **2-10**′ (正規分布)： 標準正規分布の特性関数は

$$\begin{aligned}\psi(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itz} e^{-\frac{z^2}{2}} dz \\ &= e^{-\frac{t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(z-it)^2}{2}} dz = e^{-\frac{t^2}{2}}\end{aligned}$$

で与えられる [11]。一般に $X \sim N(\mu, \sigma^2)$ のときには $X = \mu + \sigma Z$ $(Z \sim N(0,1))$ とすると

$$\begin{aligned}\psi(t) &= \mathbf{E}[e^{it\mu + \sigma itZ}] \\ &= \exp[i\mu t - \frac{\sigma^2 t^2}{2}]\end{aligned}$$

となる。これより平均と分散は $\mathbf{E}[X] = \mu,\ \mathbf{V}[X] = \sigma^2$ となることが確認できる。

例 **2-16** (指数分布とガンマ分布)： 指数分布の密度関数は

$$f(x) = \lambda e^{-\lambda x} \quad (x > 0) \tag{2.47}$$

であるが、その一般化としてガンマ分布がある。例えば正規分布から導かれる確率分布として χ^2- 分布が有用であるがこうした非負の確率変数がしたがう一般的な確率分布がガンマ分布があり確率密度関数は

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)} (\frac{x}{\beta})^{\alpha-1} e^{-\frac{x}{\beta}} \frac{1}{\beta} \quad (\alpha > 0, \beta > 0) \tag{2.48}$$

で与えられる。ここでガンマ関数は定積分

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \tag{2.49}$$

によって定義される。特に $\alpha = 1$ とするとガンマ分布は指数分布に一致する。また $\alpha = n/2, \beta = 2$ のとき自由度 **n** の χ^2(カイ二乗) 分布と呼ばれている。ガンマ分布の特性関数を評価すると

$$\psi(t) = (1 - it\beta)^{-\alpha} \tag{2.50}$$

[11] ここでの評価を数学的に正確にするには複素関数の積分を用いる必要がある (HP 上の数学補論を参照)。

となる。このことから $\mathbf{E}[X] = \alpha\beta, \mathbf{V}[X] = \alpha\beta^2$ となる。

補題 **2-5**： ガンマ関数は次の性質を持つ。
(i) $\Gamma(1) = 1$,
(ii) $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$,
(iii) 任意の整数 α に対し $\Gamma(\alpha + 1) = \alpha!$,
(iv) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$,
(v) ベータ関数

$$B(p, q) = \int_0^1 x^{p-1}(1 - x)^{q-1}dx \ \ (p > 0, q > 0) \tag{2.51}$$

に対して

$$\frac{\Gamma(p)\Gamma(q)}{\Gamma(p + q)} = B(p, q) \tag{2.52}$$

が成り立つ。

(ii) の証明 ： 部分積分 [12] を利用すると

$$\begin{aligned} \Gamma(\alpha + 1) &= \int_0^\infty x^\alpha e^{-x}dx \\ &= [x^\alpha(-e^{-x})]_0^\infty + \alpha\int_0^\infty x^{\alpha-1}e^{-x}dx \\ &= \alpha\Gamma(\alpha) \end{aligned}$$

となる。

例 **2-17** (ベータ分布)： 確率分布の台 (サポート,**support** と呼ばれる) が有界な確率分布としてはベータ分布が有用であり、特に **9** 章で説明するベイズ統計分析ではしばしば利用される。実数 $p > 0, q > 0, \beta > 0, 0 < x < \beta$ に対し、確率密度関数は

$$f(x|p, q) = \frac{1}{B(p, q)}(\frac{x}{\beta})^{p-1}(1 - \frac{x}{\beta})^{q-1}\frac{1}{\beta} \tag{2.53}$$

で与えられる。ここで $B(p, q) = \int_0^1 z^{p-1}(1 - z)^{q-1}$ (p, q は正実数) はである。若干の計算より

$$\mathbf{E}[X] = \frac{p}{p + q}\beta \ , \ \mathbf{V}[X] = \frac{pq}{(p + q)^2(p + q + 1)}\beta^2$$

となる。

[12] 二つの関数 f, g の積の微分は $[fg]' = f'g + fg'$ である。両辺を積分すれば $\int fg' = [fg] - \int f'd$ が得られる。ここでは $f(x) = x^\alpha, g(x) = -e^{-x}$ と置けばよい。

多次元の場合 ： 特性関数は多次元の場合に一般化される。確率変数ベクトル $\mathbf{X}$ が同時確率分布関数 $F(x_1, \cdots, x_p)$ にしたがう場合を考察しよう。なお一般に $\mathbf{p}$ 次元の場合についての議論の理解が難しければ $p = 2$ の場合、すなわち 2 次元の場合をより具体的に考えることがよい。$p = 2$ のとき例えば $\mathbf{X} = (X_1, X_2)^{'}$, $\mathbf{t} = (t_1, t_2)^{'}$ より $\mathbf{t}^{'}\mathbf{X} = t_1X_1 + t_2X_2$ である。

定義 **2-4** : p 次元確率変数 $\mathbf{X} = (X_j)$ が分布関数 $\mathbf{F}$ にしたがうとき特性関数 **(characteristic function)** を

$$\psi(t; \mathbf{X}) = \mathbf{E}[e^{i\mathbf{t}^{'}\mathbf{X}}] = \int_{\mathbf{R}^p} e^{i\mathbf{t}^{'}\mathbf{X}} dF(\mathbf{x}) \tag{2.54}$$

で定める。ただし $i^2 = -1, \mathbf{t}^{'}\mathbf{x} = \sum_{j=1}^{p} t_j x_j, \mathbf{t} = (t_j) \in \mathbf{R}^p$ である。

例 **2-14**$^{'}$ **(多項分布)** : **()** $M(n, p_1, \cdots, p_k)$ の確率関数は

$$p(x_1, \cdots, x_{k+1}) = \frac{n!}{x_1! \cdots x_{k+1}!} p_1^{x_1} \cdots p_{k+1}^{x_{k+1}} \tag{2.55}$$

で与えられる。($\sum_{i=1}^{k+1} x_i = n, \sum_{i=1}^{k+1} p_i = 1, p_i \geq 0$.)
確率変数ベクトル $\mathbf{X} = (X_1, \cdots, X_k)^{'}$ に対する特性関数は

$$\psi(t) = \left[\sum_{j=1}^{k} p_j e^{it_j} + p_{k+1} \right]^n \tag{2.56}$$

である。特性関数を偏微分すると

$$\frac{\partial \psi(\mathbf{t})}{\partial t_j}|_{\mathbf{t}=\mathbf{0}} = i\mathbf{E}[X_j]$$

となるので $\mathbf{E}[X_j] = np_j$ $(j = 1, \cdots, k)$ が得られる。同様に 2 回微分操作により $\mathbf{Cov}[X_i, X_j] = np_j(1 - p_j)$ $(i = j), \mathbf{Cov}[X_i, X_j] = -np_ip_j)$ $(i \neq j)$ となることが分かる。これらの表現は特に $p = 1$ の時には二項分布 $B(n, p)$ となることが確認できる。

例 **2-11**$^{'}$ **(**p 次元正規分布**)** : p 次元正規分布 $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ の特性関数は

$$\begin{aligned} \psi_Z(\mathbf{t}) &= (\frac{1}{\sqrt{2\pi}})^p \int_{\mathbf{R}^p} e^{i\mathbf{t}^{'}\mathbf{z}} e^{-\frac{\mathbf{z}^{'}\mathbf{z}}{2}} dz \\ &= \prod_{i=1}^{p} e^{-\frac{t_i^2}{2}} \\ &= \exp\left[-\frac{\mathbf{t}^{'}\mathbf{t}}{2} \right] \end{aligned}$$

で与えられる。次に一般に $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ のとき変換 $\mathbf{X} = \mathbf{CZ} + \boldsymbol{\mu}$ を利用するが、$p \times p$ 行列 $\boldsymbol{\Sigma}$ は $\mathbf{CC}' = \boldsymbol{\Sigma}$ を満たす行列を意味する。このとき特性関数は

$$\begin{aligned}
\psi_X(\mathbf{t}) &= \mathbf{E}[e^{i\mathbf{t}'\mathbf{X}}] \\
&= \mathbf{E}[e^{i\mathbf{t}'(\mathbf{CZ}+\boldsymbol{\mu})}] \\
&= e^{i\boldsymbol{\mu}'\mathbf{t}}\mathbf{E}[e^{i(\mathbf{C}'\mathbf{t})'\mathbf{Z}}] \\
&= \exp\left[i\boldsymbol{\mu}'\mathbf{t} - \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right]
\end{aligned}$$

となる。ここで $\psi_Z(\mathbf{C}'\mathbf{t}) = \mathbf{E}[e^{i(\mathbf{C}'\mathbf{t})'\mathbf{Z}}] = e^{-\frac{1}{2}\mathbf{t}'\mathbf{CC}'\mathbf{t}}$ および $\mathbf{CC}' = \boldsymbol{\Sigma}$ を用いた。したがって再び

$$\frac{\partial\psi(\mathbf{t})}{\partial t_j}|_{\mathbf{t}=\mathbf{0}} = i\mathbf{E}[X_j]$$

を用いると $\mathbf{E}[X_j] = \mu_j$ $(j = 1, \cdots, p)$ となることが分かる。同様に２回微分操作により $\mathbf{Cov}[X_i, X_j] = \sigma_{ij}$ となる。

次の結果は **Levy** の () と呼ばれているが、分布関数と (分布関数の **Fourier** 変換である) 特性関数が **1** 対 **1** の関係であることを示している。例えば中心極限定理は特性関数を用いて示されるが、確率分布より直接的に示すことはごく特殊な場合を除いて容易でないので有用なのである。ここでは導出の概略も示しておこう。

定理 **2-6:** **(1** 次元**)** 確率変数 X の分布関数 $F(\cdot)$, 特性関数 $\psi(t)$ とする。任意の実数 $x_1 < x_2$ に対して

$$P\left(X \in (x_1, x_2)\right) + \frac{1}{2}P(\{x_1\}) + \frac{1}{2}P(\{x_2\} = \lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-itx_1} - e^{-itx_2}}{it}\psi(t)dt \tag{2.57}$$

が成り立つ

連続型確率分布の場合には一点をとる確率はゼロ、すなわち $P(\{x_1\}) = P(\{x_2\}) = 0$ である。収束定理を用いると任意の $h > 0$ に対して

$$\begin{aligned}
\frac{1}{h}\left[F(x+h) - F(x)\right] &= \frac{1}{h}P(X \in (x, x+h]) \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}\left[\frac{1-e^{-ith}}{ith}\right]\psi(t)dt \\
&\to \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}\psi(t)dt \ \ (\text{as } h \to 0)
\end{aligned}$$

が得られる。$[1 - e^{ith}]/[ith] \to 1$ $(h \to 0)$ より左辺の極限は密度関数 $f(x)$ となるので、定理 **2-6** は非常に使いやすい形となる。

系 **2-7:** 特性関数 $\psi(t) \in L^1(-\infty, \infty)$ とする。(ここで L_1 は積分が存在して有限値として定まる、可積分を意味する。) このとき確率分布関数 F は微分可能であり

$$f(x) = F^{'}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt}\psi(t)dt \tag{2.58}$$

が成り立つ。

定理 **2-6** の証明 : 任意の実数 $x_1 < x_2$ に対して

$$\begin{aligned} & \frac{1}{2\pi}\int_{-T}^{T}\left[\frac{e^{-itx_1}-e^{-itx_2}}{it}\right]\psi(t)dt \\ = & \frac{1}{2\pi}\int_{-T}^{T}\left[\frac{e^{-itx_1}-e^{-itx_2}}{it}\right]\left[\int_{-\infty}^{\infty} e^{itx}dF(x)\right]dt \\ = & \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\int_{-T}^{T}\frac{e^{it(x-x_1)}-e^{it(x-x_2)}}{it}dt\right]dF(x) \end{aligned}$$

であるが、この操作は積分の順序交換に関する **Fubini** の定理から正当化できる。さらにカッコ内を $I(T)$ とすると $\cos$ 関数の性質 (原点について対称) より $\cos$ 関数項が消えて

$$I(T) = \frac{1}{\pi}\int_0^T \frac{\sin t(x-x_1)}{t}dt - \frac{1}{\pi}\int_0^T \frac{\sin t(x-x_2)}{t}dt \tag{2.59}$$

である。次に述べる補題 **2-8** を利用すると $I(T) = 0\ (x < x_1)$, $I(T) = \frac{1}{2}\ (x = x_1)$, $I(T) = 1\ (x_1 < x < x_2)$, $I(T) = \frac{1}{2}\ (x = x_2)$, $I(T) = 0\ (x > x_2)$ と評価できるので結果が得られる。 **Q.E.D.**

補題 **2-8** : 任意の正実数 a に対して

$$\left|\int_0^a \frac{\sin x}{x}dx - \frac{\pi}{2}\right| \leq \frac{2}{a} \tag{2.60}$$

が成立する。

補題 **2-8** の証明 : 任意の正実数 a に対して

$$\int_0^\infty\int_0^a e^{-xy}\sin x dxdy = \int_0^a \frac{\sin x}{x}dx \tag{2.61}$$

となることに注意する。次に

$$\begin{aligned} \int_0^\infty\int_0^a e^{-xy}e^{ix}dxdy &= \int_0^\infty\int_0^a e^{(i-x)y}dxdy \\ &= \int_0^\infty\left[\frac{e^{(i-y)x}}{i-y}\right]_0^a dy \\ &= \int_0^\infty \frac{e^{-yx}}{-1-y^2}\left[i\cos x - \sin x + y\cos x + iy\sin x\right]_0^a dy \end{aligned}$$

となる。右辺の虚数部分は

$$\frac{\pi}{2} - \cos a \int_0^\infty \frac{e^{-ay}}{1+y^2} dy - \sin a \int_0^\infty \frac{y e^{-ay}}{1+y^2} dy$$

となる。ここで $1/(1+y^2) \leq 1$ を利用すると、右辺の第 **2** 項・第 **3** 項はそれぞれ $1/a$ より小さくできる。 **Q.E.D.**

2-A 補論:変数変換の公式

変数変換の公式は大学教養課程・微積分において登場するが、二次元以上の場合は後半に登場する。その正確な証明にはかなり複雑な議論が必要となるが、ここでは応用上で必要な最小限の直観的説明にとどめる[13]。

まず 1 変数の変換を直観的に見てみよう。変換 $y = cz$ (c は定実数) では変数 z が **1** 単位増加すると y は $c > 0$ なら c 単位増加する。そこで形式的に $dy = |c|dz$ (あるいは $\frac{dy}{dz} = |c|$) と表現できるので絶対値の記号はこの場合は不要であるが $c < 0$ の場合は必要となる。) 確率変数 Z が密度関数 $f_Z(z)$ を持つとすると、逆関数 $z = y/c$ $(= z^{-1}(y))$ となるので確率変数 $Y = cZ$ の密度関数は $|c|dz = dy$ より形式的に $\int f_Z(z)dz = \int f_Z(\frac{y}{c})\frac{dy}{|c|}$ と表現できる。そこで Y がしたがう密度関数を

$$f_Y(y) = \quad f_Z(\frac{y}{c})\frac{1}{|c|}$$

とすれば $\int f_Y(y)dy = 1$ となるので整合的になる。ここで変換のヤコビアン $|J|_+ = |1/c|$ と呼ぼう。ここで確率は $P(Y \leq y) = P(cZ \leq y) = P(Z \leq y/c)$ とすると両辺を y で微分すると Y の密度関数として同一の形が得られる。

次に 2 次元の場合を考える。2 次元ベクトル $\mathbf{z} = (z_1, z_2)'$ より $\mathbf{R}^2 \to \mathbf{R}^2$ の $\mathbf{y} = (y_1, y_2)' = \mathbf{C}(z_1, z_2)'$ により面積がどの様に変化するかをみる。($\mathbf{C} = (c_{ij})$ は 2×2 行列である。) 単位ベクトル $\mathbf{e}_1 = (1, 0)'$ はベクトル $\mathbf{c}_1 = (c_{11}, c_{21})'$ に移り、単位ベクトル $\mathbf{c}_2 = \mathbf{e}_2 = (0, 1)'$ はベクトル $(c_{12}, c_{22})'$ に移る。したがって、二つの単位ベクトルで作られる領域が変換される領域の面積 $\mathbf{S}$ とすると幾何的に見ると

$$\begin{aligned} S^2 &= \|\mathbf{c}_1\|^2\|\mathbf{c}_2\|^2 \sin^2\theta \\ &= \|\mathbf{c}_1\|^2\|\mathbf{c}_2\|^2 - (\mathbf{c}_1, \mathbf{c}_2)^2 \\ &= (c_{11}^2 + c_{21}^2)(c_{12}^2 + c_{22}^2) - (c_{11}c_{12} + c_{21}c_{22})^2 \\ &= (c_{11}c_{22} - c_{12}c_{21})^2 = |\mathbf{C}|^2 \end{aligned}$$

である。($(\mathbf{c}_1, \mathbf{c}_2) = \|\mathbf{c}_1\|\|\mathbf{c}_2\| \cos\theta$ は内積である。) したがって、この変換による面積比は行列式 $\mathbf{S}$ で与えられる。

[13]詳しくは例えば杉浦光夫「解析入門 II」東京大学出版会, VII-4 節を参照。

一般に線形変換でない (1 対 1) 変換では事態はもう少し複雑となるが、直観的には 1 次元の形 $dy = |J|_+ dz$ と同様である。。直観的には各点において局所的に線形変換と見なすことが出来る場合（例えば $p = 2$ の場合）には変換 $\mathbf{y} = (y_1(\mathbf{z}), y_2(\mathbf{z}))'$ のヤコビアンは

$$|\mathbf{J}(z \to y)|_+ = \left| \begin{array}{cc} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{array} \right|_+ \tag{2.62}$$

により与えられる。（ここで $|\cdot|_+$ は行列式の正値を意味する。）この場合には直観的には

$$dy_1 dy_2 = |\mathbf{J}(z \to y)|_+ dz_1 dz_2 \tag{2.63}$$

と表現される。線形変換の例ではヤコビアンは一定値となり $dy_1 dy_2 = |\mathbf{C}|_+ dz_1 dz_2$ である。一般の場合の変換公式は一見すると複雑に見えるが重要なので次のようにまとめておく。

定理 2-A-1(変数変換の公式)：**A,B** を $\mathbf{R}^p$ の体積確定集合、写像 **c** は **A** を含む開集合 **U** で定義され、$\mathbf{R}^p$ の値をとる C^1 級写像で、**A** から **B** の上への一対一写像、かつすべての点 $\mathbf{y} \in U$ において $|\mathbf{J}(z \to y)| \neq 0$ とする $(\mathbf{y} = c(\mathbf{z}),\ \mathbf{J}(z \to y) = (\frac{\partial y_i}{\partial z_j})$ は $p \times p$ 行列である)。このとき $B = c(A)$ 上の実数値関数 **k** に対し、**(a)**$k(\mathbf{y})$ は **B** 上広義可積分、**(b)**$k(g(\mathbf{z}))|J(\mathbf{z})|$ は **A** 上広義可積分、は同値であり変換公式

$$\int_B k(\mathbf{y}) d\mathbf{y} = \int_A k(c(\mathbf{z}))|\mathbf{J}((z \to y))|_+ d\mathbf{z} \tag{2.64}$$

が成り立つ。

ここで例えば本文中の正規分布の積分計算では $z_1 = r, z_2 = \theta$ および $y_1 = x, y_2 = y$ とすると $dy_1 dy_2 = r dz_1 dz_2$ となることを確認せよ。

統計学ではしばしば連続型の確率分布を想定するために具体的な確率分布や確率計算には変数変換の操作が多用される。ここでは変数変換について有用な結果に言及しておく。抽象的なので実例を参考にして検討するとよい。

定理 2-A-2(変数変換における密度関数)：**A,B** を $\mathbf{R}^p$ の体積確定集合、**h** は **A** を含む開集合 **U** で定義され、$\mathbf{R}^p$ の値をとる C^1 級写像で、**A** から **B** の上への一対一写像、かつすべての点 $\mathbf{y} \in U$ において $|\mathbf{J}(z \to y)| \neq 0$ とする $(\mathbf{y} = h(\mathbf{z}),\ \mathbf{J}(y \to z) = (\frac{\partial z_i}{\partial y_j})$ は $p \times p$ 行列である)。このとき $B = h(A)$ 上で定義された確率変数ベクトル Z がしたがう連続型密度関数を $f_Z(z)$, Y がしたがう連続型密度関数を $f_Y(y)$ とする。この

とき

$$\int_A f_Z(\mathbf{z})d\mathbf{z} = \int_B f_Z(h^{-1}(\mathbf{y}))|\mathbf{J}((y \to z))|_+ d\mathbf{y} \tag{2.65}$$

が成り立つ。したがって

$$f_Y(\mathbf{y}) = f_Z(h^{-1}(\mathbf{y}))|\mathbf{J}((y \to z))|_+ \tag{2.66}$$

で与えられる。

3　条件付期待値と独立性

3.1 条件付確率と情報

複数の事象 A と B がある時、事象 B が与えられた時の事象 A の条件付確率 (**conditional probability**) は情報としての事象 B が与えられたときの事象 A のリスク評価値と見なすと分かりやすい。条件付確率は情報の有効な活用、情報の定義、情報の評価などは様々な実際的問題の分析に欠かせない重要な概念である。まずは初等確率の例により条件付確率を考察しよう。

例 **3.1:** コインを **3** 回投げた結果を基本事象、全事象を $\mathbf{\Omega} = (\omega_i)$ とする。基本事象は $\omega_1 = \{HHH\}$ などの集合であり全部で **8** 個あるが、不確実な事象としては集合 $A = \{$ 少なくとも 2 回表が出る $\}$, $B = \{$ 最初の結果が表 $\}$, $C = \{$ 最初の 2 回は表・裏の順番に出る $\}$ なども考えられる。同等に確からしい基本事象の中で事象 **A** に含まれる数から $P(A) = 4 \times (\frac{1}{2})^3 = \frac{1}{2}$ である。ここで予め情報 **B** が与えられたときの事象の不確実性の評価を考えよう。全事象は $\mathbf{\Omega}' = \{HHH, HHT, HTH, HTT\}$ であるので、情報 B が得られたという条件の下では

$$P(A|B) = 3 \times (\frac{1}{2})^2 = \frac{3}{4} \tag{3.1}$$

となる。したがって元の全事象に対して $\Omega' \subset \Omega$ とすると、事象 **B** が既知という情報の下で事象 **A** が起きる条件付確率は

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3 \times (\frac{1}{2})^3}{4 \times (\frac{1}{2})^3} \tag{3.2}$$

と定義することが合理的となる。さらに事象 **C** が情報として与えられた場合は全事象は $\mathbf{\Omega}'' = \{HTH, HTT\}$ より条件付確率は

$$P(A|C) = \frac{1}{2}$$

となる。このように条件 $P(A|C) = P(A)$, あるいは条件 $P(A \cap C) = P(A)P(C)$ が成り立つとき集合 **A,C** は互いに独立 (**independent**) と云う。すなわち集合が互いに独立な時には集合 **B** の情報は集合 **A** の確率評価とは無関係となることを意味する。

こうして事象の条件付確率と独立性を導入すると統計分析における情報の役割が明確になる。ここで情報とは標本空間 $\mathbf{\Omega}$ の上の集合から構成される集合族の要素、集合の分割として直観的に理解できる。一般に情報という言葉の意味はかなり曖昧に使われるが、集合と $\sigma-$ 集合族という言葉を用いると明確になる。(例えば Ω' 上の部分集合は Ω 上の集合より粗い、と言われる。) 次に条件付期待値 **(conditional expectation)** の概念を導入するが、これはマルコフ性やマルチンゲールをはじめとして時間とともに変動する現象を統計的に解析する際には必要不可欠な基礎である。条件付期待値は条件付確率と比較すると少し回りくどい数理的な議論が必要となる。まず例 **3.1** を一般化し、確率空間 $(\Omega, \mathcal{F}, P)$ 上で任意の集合 C を条件とする条件付確率を定義することから始めよう。

定義 3.1 ： 任意の集合 $A \in \mathcal{F}$ に対し事象 C が与えられた時の事象 A の条件付確率 (Conditional Probability) は

$$P(A|C) = \frac{P(A \cap C)}{P(C)} \quad (\text{ただし } P(C) > 0) \tag{3.3}$$

により定める。

ここで定義した条件付確率は確率測度である。次に事象による条件付けを確率変数により表現する為に、確率空間 $(\Omega, \mathcal{F}, P)$ 上で確率変数

$$X_1(\omega) \equiv \begin{cases} 1 & \textbf{if } \omega \in C \\ 0 & \textbf{if } \omega \notin C \end{cases}$$

を定める (ただし $1 > P(C) > 0$ とする)。同一の確率空間上にもう一つの確率変数 $X_2(\omega)$ が定義されているとき、$\mathbf{R}$ 上の任意の B に対して条件付確率 $P(X_2(\omega) \in B|C)$ が定まる。そこで X_2 の期待値を分解すると

$$\begin{aligned} \mathbf{E}[X_2] &= \int_\Omega X_2(\omega) P(d\omega) \\ &= \int_\Omega X_2(\omega) P(d\omega|C) P(C) + \int_\Omega X_2(\omega) P(d\omega|C^c) P(C^c) \\ &= \mathbf{E}[X_2|X_1 = 1] P(X_1 = 1) + \mathbf{E}[X_2|X_1 = 0] P(X_1 = 0) \end{aligned}$$

と書ける。ここで

$$Y(\omega) = \begin{cases} \mathbf{E}[X_2|X_1 = 1] & \textbf{if } X_1(\omega) = 1 \\ \mathbf{E}[X_2|X_1 = 0] & \textbf{if } X_1(\omega) = 0 \end{cases}$$

により確率変数 Y を定義すると、

$$\mathbf{E}[X_2] = \mathbf{E}_{X_1}[Y] = \mathbf{E}[\mathbf{E}(X_2|X_1)] \tag{3.4}$$

が成り立つ。右辺の期待値の記号 $\mathbf{E}_{X_1}[\,\cdot\,]$ は確率変数 X_1 についての期待値、$\mathbf{E}[X_2|X_1]$ は X_1 を条件とする X_2 の期待値、という意味である。この関係は期待値の繰り返し公式と呼ばれるが

$$\int_{\Omega} X_2(\omega)P(d\omega) = \int_{\Omega} Y(\omega)P(d\omega) \tag{3.5}$$

と表現してもよい。

同様に一般的に標本空間を分割して $\mathbf{\Omega} = \bigcup_{i=1}^{k} \mathbf{C}_i,\ P(\mathbf{C}_i) > 0,\ \mathbf{C}_i \cap \mathbf{C}_j = \phi\ (i \neq j)$ とする。このとき期待値の分解公式

$$\mathbf{E}[X_2] = \sum_{i=1}^{k} \mathbf{E}[X_2|\mathbf{C}_i]P(\mathbf{C}_i) \tag{3.6}$$

が成立する。ここで特に確率変数として $X_2(\omega) = 1\ (\text{if } \omega \in B), X_2(\omega) = 0\ (\text{if } \omega \notin B)$ とすると確率 $P(B)$ の分解が得られる。このようにして条件付確率の定義より得られる関係は **Bayes** の定理と呼ばれているが、特にベイズ統計学では基本的な役割を演じる。次の関係はベイズの公式と呼ばれるが、右辺と左辺における条件付けの集合が入れ替わっていることが重要である。

定理 3.1： 集合 $\mathbf{C}_i \cap \mathbf{C}_j = \phi, P(\mathbf{C}_i) > 0\ (i \neq j; i, j = 1, \cdots, n)$ とする。任意の集合 $\mathbf{B}\ (P(\mathbf{B}) > 0)$ に対して

$$P(\mathbf{C}_i|\mathbf{B}) = \frac{P(\mathbf{B}|\mathbf{C}_i)P(\mathbf{C}_i)}{\sum_{i=1}^{n} P(\mathbf{B}|\mathbf{C}_i)P(\mathbf{C}_i)}\ . \tag{3.7}$$

二つの確率変数 X_1, X_2 が離散型確率分布にしたがう場合、同時確率分布 **(joint probability distribution)**

$$F(x_1, x_2) = P(\omega|X_1(\omega) \leq x_1, X_2(\omega) \leq x_2) = \sum_{z_1 \leq x_1, z_2 \leq x_2} p_{x_1,x_2}(z_1, z_2) \tag{3.8}$$

(ただし同時確率関数 $p_{x_1,x_2}(x_1, x_2) = P(\omega|X_1(\omega) = x_1, X_2(\omega) = x_2)$**)**, 周辺確率分布 **(marginal probability distribution)**

$$F_i(x_i) = P(\omega|X_i(\omega) \leq x_i) = \sum_{z_i \leq x_i} p_{x_i}(z_i)\ (i = 1, 2), \tag{3.9}$$

(ただし周辺確率関数 $p_{x_i}(x_i) = P(\omega|X_i(\omega) = x_i)\ (i = 1, 2)$**)** より条件付確率関数を

$$p_{x_2|x_1}(x_2) = P(X_2(\omega) = x_2|X_1(\omega) = x_1) \tag{3.10}$$

によって定義しておこう。確率変数が離散分布にしたがう場合には条件付確率から条件付確率関数を表現することが可能である。

3.2 条件付期待値

一般的に確率空間 $(\Omega, \mathcal{F}, P)$ 上の二つの確率変数 $X_1(\omega), X_2(\omega)$ に対し $\forall \mathbf{x}_1 \in \mathbf{R}$ に対して $\mathbf{E}[X_2|\mathbf{X}_1 = \mathbf{x}_1]$ を定めよう。連続型確率分布にしたがう確率変数、例えば統計学の応用でもっともよく利用される正規分布などでは一点をとる確率 $P(\mathbf{X}_1 = \mathbf{x}_1) = 0$ となる。したがって条件付確率より直接的に条件付期待値を定めることには難点がある他方、応用上ではしばしば多数の確率変数の変動、時間とともに変動する確率変数の相互依存性 **(dependence)** を分析する際に一部分の情報 (X_1) が得られたという条件の元での $(X_2$ の$)$ 平均を考察する必要が生じる。

ここで応用上でよく利用される二次元分布関数が同時密度関数 $f(x_1, x_2)$, 周辺密度関数 $f_i(x_i)$ $(i = 1, 2)$ を持つ場合を考察しよう。すなわち任意の二次元の点 $(x_1, x_2)' \in \mathbf{R}^2$ に対して同時分布関数が

$$\begin{aligned} P(X_1(\omega) \leq x_1, X_2(\omega) \leq x_2) &= F(x_1, x_2) \\ &= \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(z_1, z_2) dz_1 dz_2 \ , \end{aligned}$$

と書けるとする。他方、X_1 の周辺分布関数より

$$\begin{aligned} P(X_1(\omega) \leq x_1) &= F_1(x_1) \\ &= \int_{-\infty}^{x_1} f_1(z_1) dz_1 \end{aligned}$$

である。こうした表現と同じことであるが念のために同時分布関数と周辺分布関数の両辺をそれぞれ微分すると同時密度関数と周辺密度関数はそれぞれ

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} \ , \ f_1(x_1) = \frac{\partial F_1(x_1)}{\partial x_1} \tag{3.11}$$

である。このとき同時分布より条件付分布関数と条件付密度関数を

$$F(x_2|x_1) = \begin{cases} \int_{-\infty}^{x_2} \frac{f(x_1, z)}{f_1(x_1)} dz & \textbf{if } f_1(x_1) > 0 \\ 0 & \textbf{if otherwise} \end{cases}$$

および

$$f(x_2|x_1) = \begin{cases} \frac{f(x_1, x_2)}{f_1(x_1)} & \textbf{if } f_1(x_1) > 0 \\ 0 & \textbf{if otherwise} \end{cases} \ ,$$

として $f(x_2|x_1)$ を条件付密度関数とおこう。ここで例示として二次元正規分布 (相関係数は **0.5**) の同時密度関数と条件付密度関数を図 **3-1** に例示しておくが、二つの密度関数の形状に注意しよう。

(図 3.1: 同時密度関数と条件付密度関数**)**

このとき

$$\mathbf{E}[X_2|X_1 = x_1] = \int_{-\infty}^{\infty} x_2 dF(x_2|x_1) = \int_{-\infty}^{\infty} x_2 f(x_2|x_1) dx_2 \tag{3.12}$$

となる。また確率変数 X_1 と X_2 の密度関数を $f_1(x_1), f_2(x_2)$ (これらは周辺密度関数、その分布関数をそれぞれ $F_1(x_1), F_2(x_2)$ とする) として、条件付期待値の繰り返し公式 $\mathbf{E}[X_2] = \mathbf{E}[\mathbf{E}[X_2|X_1]]$ を書き換えると

$$\begin{aligned}
\mathbf{E}[X_2] &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x_2 f(x_2|x_1) dx_2 \right] f_1(x_1) dx_1 \\
&= \int_{-\infty}^{\infty} x_2 \left[\int_{-\infty}^{\infty} f(x_2|x_1) f_1(x_1) dx_1 \right] dx_2 \\
&= \int_{-\infty}^{\infty} x_2 f_2(x_2) dx_2 \\
&= \int_{-\infty}^{\infty} x_2 dF_2(x_2)
\end{aligned}$$

となる。ここで周辺密度関数は条件付密度と同時密度関数より

$$\int_{-\infty}^{\infty} f(x_2|x_1) f_1(x_1) dx_1 = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 = f_2(x_2)$$

で与えられる。

一般に条件付期待値を定義するには、情報集合 $\mathcal{F}$ 上に確率変数が定まるとき、この $\mathcal{F}$ より小さな情報集合、すなわち部分 σ-集合族

$$\mathcal{F} \supset \mathcal{G} \equiv \{X_1^{-1}(B)|B \in \mathbf{B}(\mathbf{R})\} \tag{3.13}$$

を導入しよう。写像 $X_1(\cdot)$ の逆写像により定まる集合 $A = X^{-1}(B)$ は $A = \{\omega|X(\omega) \in B\}$ を意味するので $\mathcal{G}$ は $\mathcal{F}$ 全体ではなく確率変数 $\mathbf{X}_1(\omega)$ を決める情報全体を意味し、$\mathcal{G}$ はである。情報 $\mathcal{G}$ を知ることは確率変数が取る値 $X_(\omega)$ を知ることと同値になるので、X_1 が連続型の確率分布にしたがい一点をとる確率はゼロであるにも関わらず条件付確率との類推が可能となる [14]。そこで条件付期待値を次のように定義すると既に述べた $Y = \mathbf{E}[X_2|X_1]$ の一般化に対応する [15]。

[14] 例えば $\mathbf{\Omega} = \{$ サイコロの目 $\}$、$X_1(\omega) = 1$ (奇数の目), $X_1(\omega) = 2$ (偶数の目) とすると X_1 が含む情報は $\mathcal{F}$ の部分集合となる。

[15] こうした定義が意味を持つことは数理的に保障される (ラドン・ニコディウム (Radon-Nikodym) 定理による RN 微分の存在) ことについて HP(ホームページ) 補論で言及する。

定義 3.2 条件付期待値 (conditional expectation) : 確率空間 $(\Omega, \mathcal{F}, P)$ 上の確率変数 $X(\omega)$ が $\mathbf{E}[|X|] < +\infty$ $(i.e.$可積分$)$ のとき、$Y(\cdot)$: $\mathcal{G}$-可測の確率変数が存在し

$$\int_A X(\omega)P(d\omega) = \int_A Y(\omega)P(d\omega) \quad \forall A \subset \mathcal{G} \subset \mathcal{F} \tag{3.14}$$

となるとき $Y = \mathbf{E}[X|\mathcal{G}]$ を条件付期待値と呼ぶ。

特に確率変数 $X = X_2, X_1^{-1}(\omega) \in \mathcal{G} \subset \mathcal{F}, Y = \mathbf{E}[X_2|X_1]$ とすると条件付期待値の繰り返し公式

$$\mathbf{E}[X_2|X_1] = \mathbf{E}[X_2]$$

が一般に成立することが確認できる。

3.3 条件付分布の例

以上における条件付期待値の説明は二つの確率変数 (X_1, X_2) の同時分布を考える際に一方の確率変数の情報 $\{\omega|X_1 = x_1\}$ が与えられた時の他方の確率変数 X_2 の確率分布に関する議論であった。より一般に $\mathbf{p}$ 次元の確率変数 $\mathbf{X}$ を二つのベクトル (p_1 次元確率ベクトル $\mathbf{X}_1$ と p_2 次元確率ベクトル $\mathbf{X}_2(p = p_1 + p_2)$) に分割した場合にも p_2 次元確率変数ベクトル $\mathbf{X}_2$ の条件付期待値 $\mathbf{E}[\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1]$ について同様の議論が可能である。

例 3.2 (多次元正規分布) : p 次元確率変数ベクトル $\mathbf{X}$ が正規分布 $N_p[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$ にしたがうものとして、その密度関数 ($|\boldsymbol{\Sigma}| \neq 0$ を仮定する。($\boldsymbol{\Sigma}$ は正定符号行列であるが、任意のゼロでない実ベクトル $\mathbf{x}$ に対し $\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x} > 0$ を意味する。) すなわち

$$f(\mathbf{x}) = (\frac{1}{\sqrt{2\pi}})^p |\boldsymbol{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\} \tag{3.15}$$

で与えられるとする。期待値ベクトルと共分散行列は $\boldsymbol{\mu} = (\mu_i) = (\mathbf{E}(X_i))$, $\boldsymbol{\Sigma} = (\sigma_{ij}) = (\mathbf{E}[(X_i - \mu_i)(X_j - \mu_j)])$ である。

特に $p = 2$(二次元正規分布) の場合に $(X_1, X_2)'$ の期待値ベクトルを $\mu = (\mu_1, \mu_2)'$, 2×2 の分散共分散行列を

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

(X_1 と X_2 の相関係数 ρ(**(2.23)** の ρ_{12}) を用いて $\sigma_{11} = \sigma_1^2, \sigma_{22} = \sigma_2^2, \sigma_{12} = \rho\sigma_1\sigma_2$ と表すと行列式は $|\boldsymbol{\Sigma}| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_1^2\sigma_2^2(1 - \rho^2)$) とする。確率変数ベクトル $\mathbf{X} = (X_1, X_2)'$ がしたがう密度関数は

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}((\frac{x_1 - \mu_1}{\sigma_1})^2 + (\frac{x_2 - \mu_2}{\sigma_2})^2 - 2\rho(\frac{x_1 - \mu_1}{\sigma_1}(\frac{x_2 - \mu_2}{\sigma_2}))\right]$$

と表現できる。

ここで一般に $p\times 1$ の確率変数 $\mathbf{X}$ を p_1 確率ベクトルと p_2 確率ベクトル $(p=p_1+p_2)$ を要素として $\mathbf{X}=(\mathbf{X}_1^{'},\mathbf{X}_2^{'})^{'}$ と分割し、その分割に対応して $(p_1+p_2)\times 1$ ベクトル,$(p_1+p_2)\times(p_1+p_2)$ 行列

$$\mathbf{x}=\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix},\quad \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix},\ \boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}$$

としよう。次の補題 **3.2** および $\mathbf{X}_1$ の周辺分布が $\mathbf{X}_1\sim N_{p_1}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11})$ となることから、確率変数 $\mathbf{X}_2$ の条件付分布は

$$\mathbf{X}_2|\mathbf{X}_1=\boldsymbol{x}_1\sim N_{p_2}(\boldsymbol{\mu}_2+\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1),\boldsymbol{\Sigma}_{22.1}), \tag{3.16}$$

となる。ただし

$$\boldsymbol{\Sigma}_{22.1}=\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \tag{3.17}$$

は () であり、係数 $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$ を回帰係数と呼ぶ。

補題 **3.2**： $\boldsymbol{\Sigma}>0$ とする。このとき分割行列の逆行列は

$$\boldsymbol{\Sigma}^{-1}=\begin{bmatrix}\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{bmatrix}+\begin{bmatrix}-\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\\ \mathbf{I}_{p_2}\end{bmatrix}\boldsymbol{\Sigma}_{22.1}^{-1}\left[-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1},\mathbf{I}_{p_2}\right] \tag{3.18}$$

と表現できる。

証明は逆行列の定義 $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}=\mathbf{I}_p$ を確かめればよいが、$\mathbf{I}_p$ は $p\times p$ の単位行列 (対角要素は **1** でそれ以外はゼロの行列) を表す。そこで条件付分布 (正規分布) は次のように導出できる。$(p_1+p_2)\times(p_1+p_2)$ 行列

$$\mathbf{C}=\begin{bmatrix}\mathbf{I}_{p_1} & \mathbf{O}\\ -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{I}_{p_2}\end{bmatrix}$$

をとり計算すると

$$\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{'}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \mathbf{O}\\ \mathbf{O} & \boldsymbol{\Sigma}_{22.1}\end{bmatrix}$$

となる [16]。したがって、行列式について $|\mathbf{C}|=1$ (行列 $\mathbf{C}$ は三角行列ですべての対角成分は **1**) となるので

$$|\boldsymbol{\Sigma}|=|\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{'}|=|\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22.1}| \tag{3.19}$$

[16] ここで利用した線形変換 $\mathbf{C}$ は統計的には重要な意味がある。例えば $p=2,p_1=p_2=1$ のとき係数 $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$ は回帰係数と呼ばれている。

が得られる。次に補題 **3.2** を利用すると $f(\mathbf{x})$ の指数部分は

$$\begin{aligned}
& (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
= & \left[(\mathbf{x}_1-\boldsymbol{\mu}_1)', (\mathbf{x}_2-\boldsymbol{\mu}_2)'\right] \\
& \quad \times \left(\begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} + \begin{bmatrix} -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ \mathbf{I}_{p_2} \end{bmatrix} \boldsymbol{\Sigma}_{22.1}^{-1} \left[-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}, \mathbf{I}_{p_2}\right] \right) \begin{bmatrix} \mathbf{x}_1-\boldsymbol{\mu}_1 \\ \mathbf{x}_2-\boldsymbol{\mu}_2 \end{bmatrix} \\
= & (\mathbf{x}_1-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1) \\
& + \left[(\mathbf{x}_2-\boldsymbol{\mu}_2) - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\right]' \boldsymbol{\Sigma}_{22.1}^{-1} \left[(\mathbf{x}_2-\boldsymbol{\mu}_2) - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\right]
\end{aligned}$$

となる。したがって、$\mathbf{X}_1$ の周辺分布 $N_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ の密度関数を $f_1(\mathbf{x}_1)$ とすると、密度関数の比 $f(\mathbf{x})/f_1(\mathbf{x}_1)$ は

$$\begin{aligned}
(\frac{1}{\sqrt{2\pi}})^{p_2}|\boldsymbol{\Sigma}_{22.1}|^{-1/2} \exp\Big\{ & -\frac{1}{2}\left[\mathbf{x}_2 - \left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\right)\right]' \boldsymbol{\Sigma}_{22.1}^{-1} \\
& \times \left[\mathbf{x}_2 - \left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\right)\right]\Big\}
\end{aligned}$$

となる。すなわち確率変数の値 $\mathbf{X}_1 = \mathbf{x}_1$ を所与とするときの確率変数 $\mathbf{X}_2$ の条件付分布は正規分布である。

正規分布の拡張
特に平均ベクトル $\mathbf{0}$, 共分散行列 $\boldsymbol{\Sigma} = \mathbf{I}_p$ のとき、任意の直交行列 $\mathbf{U}$ に対して

$$\mathbf{UX} \stackrel{d}{=} \mathbf{X} \ \text{(分布が等しい)}$$

とき球面分布 (**spherical distribution**) と呼ばれる。ここで直交行列とは条件 $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}_p$ を満足する $p \times p$ 行列を意味する。$p = 2$ のとき幾何学的意味を考えると分かり易い。

また確率変数 $v \sim G$ に対して $\mathbf{X} \sim N_p(\mathbf{0}, (1/v)\boldsymbol{\Sigma})$ のとき正規混合 (**normal mixture**) 分布と呼ぶ。特に $mv \sim \chi^2(m)$ とすると自由度 m の **t** 分布が得られる。自由度が **1** ならコーシー (**Cauchy**) 分布である。
また確率 π $(0 < \pi < 1)$ と $1-\pi$ で共分散の異なる (正定数 a は例えば 3^2 などとする) 正規分布の混合である正規混合分布の密度関数は

$$f(\mathbf{x}) = \pi n_p(\mathbf{0}, \boldsymbol{\Sigma}) + (1-\pi) n_p(\mathbf{0} \quad , a\boldsymbol{\Sigma})$$

で与えられる。(ここで $n_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ は多次元正規分布 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ の密度関数を表すものとする。) この分布は正規分布よりも裾が厚い楕円分布 (**elliptically contoured distribution** の例となる。ここで楕円分布とは等高線 $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c$ (定数) が楕円となることにより定義される。楕円分布については例えば一部分の変数を条件としたとき条件付き分布も楕円分布になる。

こうした確率分布は正規分布の簡単な拡張であるが、正規分布より裾が厚くなる確率分布を比較的簡単に表現できるので実際に応用されることもある。例えば $\epsilon = 1-\pi$ を非常に小さくとると (例えば .01)、正常な正規分布のデータの中にごくわずかなはずれ値 (あるいは異常値) が混在している統計モデルが考えられる。

条件付期待値と回帰関数

二つの確率変数 $X_1(\omega), X_2(\omega)$ に対し条件付期待値 $\mathbf{E}[X_2|X_1 = x_1]$ は変数 x_1 の関数である。これを X_2 の X_1 への回帰関数と呼ぶ。確率変数 X_2 を X_1 の任意の関数 $g(X_1)$ で推定 (あるいは予測) するとき、回帰関数は平均二乗誤差を最小にする関数 g に一致する。ここで

$$\begin{aligned}
\mathbf{E}[(X_2 - g(X_1))^2] &= \mathbf{E}[(X_2 - \mathbf{E}[X_2|X_1]) + (\mathbf{E}[X_2|X_1] - g(X_1))]^2 \\
&= \mathbf{E}[(X_2 - \mathbf{E}[X_2|X_1])^2] \\
&\quad +2\mathbf{E}[\mathbf{E}[(X_2 - \mathbf{E}[X_2|X_1])|X_1](\mathbf{E}[X_2|X_1] - g(X_1))] \\
&\quad +\mathbf{E}[(\mathbf{E}[X_2|X_1] - g(X_1))]^2 \\
&= \mathbf{E}[(X_2 - \mathbf{E}[X_2|X_1])^2] + \mathbf{E}[(\mathbf{E}[X_2|X_1] - g(X_1))]^2
\end{aligned}$$

である。したがって、最小値は第２項がゼロとなる時に達成される。ここでの最小値に対応する $\mathbf{E}[(X_2 - \mathbf{E}[X_2|X_1])^2|X_1 = x_1]$ は条件 $X_1 = x_1$ の元での条件付分散と呼ぶ。正規分布の場合には回帰関数は線形関数となるが、係数 (行列) は回帰係数と呼ばれることが多い。

なお以上の議論は一般の次元 $p\ (= p_1 + p_2, p_1 \geq 1, p_2 \geq 1)$ の確率変数ベクトルについても成立する。

3.4 独立性

ここで独立性 (**independence**) を導入する。古典的な確率論や数理統計学の議論は独立性の場合を扱うことより発展したが、独立性は独立でない確率的な意味での従属性 (**dependence**) を分析する上でも重要な役割を果たす。

定義 3.3 : 確率空間 $(\mathbf{\Omega}, \mathcal{F}, P)$ 上の事象 $\mathbf{A}_i\ (i = 1, \cdots, N)$ が独立とは任意の $n\ (1 \leq n \leq N)$ に対して

$$P(\mathbf{A}_1 \cap \cdots \cap \mathbf{A}_n) = \prod_{i=1}^{n} P(\mathbf{A}_i) \tag{3.20}$$

が成り立つことである。

二つの事象の独立性 ($N = 2$) は確率が正の事象 $P(\mathbf{A}_2) > 0$ が成立するとき、条件付

確率に関する条件

$$P(\mathbf{A}_1|\mathbf{A}_2) = P(\mathbf{A}_1) \tag{3.21}$$

と同一である。定義 **3.3** では $P(A_1) = 0$ あるいは $P(A_2) = 0$ の場合も含んでいるが、独立性の解釈は条件付確率の方が直観的理解に沿っている。応用上の必要性より条件付期待値に関連して述べたように連続型の確率分布にしたがう確率変数の独立性まで定義しておく。

定義 **3.4** : 確率変数族 $\{X_\lambda\}_{\lambda\in\Lambda}$ の独立性とは、$\{X_\lambda\}$ が生成する σ-集合体 **(field)** の族 $\{\sigma(X_\lambda)\}_{\lambda\in\Lambda}$ が独立であることと定義する。この条件は任意の $\lambda_1, \cdots, \lambda_N \in \Lambda$($\Lambda$ は添字集合), 任意の $A_i \in \mathcal{F}_{\lambda_i}, (i = 1, \cdots, n; 1 \leq n \leq N)$ に対して

$$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i) \tag{3.22}$$

が成り立つことを意味する。ただし $\sigma(X_\lambda)$ は X_λ から生成される最小の σ-集合体を意味する.

例 **3.3** : 二つの確率変数 $X_1(\cdot)$ と $X_2(\cdot)$ に対して周辺分布 **(marginal distribution)** 関数を $F_1(x_1) = P(X_1(\omega) \leq x_1)$, $F_2(x_2) = P(X_2(\omega) \leq x_2)$, 同時分布 **(joint distribution)** 関数を $F(x_1, x_2) = P(X_1(\omega) \leq x_1, X_2(\omega) \leq x_2)$ とする。このとき確率変数 $X_1(\omega)$ と $X_2(\omega)$ が独立とは条件

$$F(x_1, x_2) = F_1(x_1)F_2(x_2) \tag{3.22}$$

で与えられる。同時密度関数 $f(x_1, x_2)$ を持つ場合には X_1, X_2 の周辺密度関数をそれぞれ $f_1(x_1), f_2(x_2)$ とすると、確率変数の独立性は条件

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} = f_1(x_1)f_2(x_2) \tag{3.23}$$

となる。

例 **3.4** : ２個以上の事象の独立性については若干の注意が必要である。例えば全事象 $\Omega = \{0, 1, 2, 3\}$, 事象 $A_i = \{o, i\}$ $(i = 1, 2, 3)$ に対して $P(\{0\}) = p > 0$, $P(\{A_i\}) = q > 0$ とする。このとき $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$ と $P(A_1 \cap A_2) = P(A_1)P(A_2)$ は自明な場合を除いて同時に成り立たない。例えば $p = .5, q = p^{1/3} - p \sim .293719$ としてみればよい。

例 **3.5** : 確率ベクトル $\mathbf{X} = (X_i)$ が **p** 次元正規分布 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, (ただし $p \geq 2, \boldsymbol{\Sigma} = (\sigma_{ij})$ にしたがうとき、任意の **i,j** について $\sigma_{ij} = 0$ $(i \neq j)$ ならば **p** 個の確率変数は互いに

独立となる。

定理 3.3： **1** 次元確率変数 X_1, X_2 が独立とする。
(i)$\mathbf{E}[X_j^2] < \infty$ $(j = 1, 2)$ のとき $\mathbf{E}[X_1X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2]$, $Cov(X_1, X_2) = 0$ である。
(ii) $h_1(X_1), h_2(X_2)$ を可測関数とすると $h_1(X_1)$ と $h_2(X_2)$ は独立となる。

証明 ： 独立性より

$$\begin{aligned}\mathbf{E}[X_1X_2] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2dF(x_1, x_2) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2dF(x_1)dF(x_2) \\ &= \int_{-\infty}^{\infty} x_1dF(x_1)\int_{-\infty}^{\infty} x_2dF(x_2) \\ &= \mathbf{E}[X_1]\mathbf{E}[X_2]\end{aligned}$$

である。また

$$\begin{aligned}Cov(X_1, X_2) &= \mathbf{E}[(X_1 - \mathbf{E}(X_1))(X_2 - \mathbf{E}(X_2))] \\ &= \mathbf{E}[X_1X_2 - X_1\mathbf{E}(X_2) - X_2\mathbf{E}(X_1) + \mathbf{E}(X_1)\mathbf{E}(X_2)] \\ &= \mathbf{E}[X_1X_2] - \mathbf{E}(X_1)\mathbf{E}(X_2)\end{aligned}$$

より **(i)** が得られる。**(ii)** は独立性の定義より導ける。 **Q.E.D.**

例 3.6： 確率変数 $X_i \sim N(0, 1)$ $(i = 1, 2)$ のとき相関係数 $\rho(X_1, X_2) = 0$ ならば X_i は互いに独立な確率変数、$X_j^2 \sim \chi^2(1)$ $(j = 1, 2)$ かつ $X_1^2 + X_2^2 \sim \chi^2(2)$ となる。ここで $\chi^2(2)$ は自由度 **1** の χ^2- 分布であるが、$Gamma(1, 2)$ 分布である。同様に $X_i \sim N(0, 1)$ $(i = 1, \cdots, n)$ のとき $\sum_{i=1}^n X_i^2 \sim \chi^2(n)$ となる。これはガンマ分布 $Gamma(\frac{n}{2}, 2)$ に対応する。

例 3.7： 確率変数 X_i $(i = 1, 2)$ の同時密度関数が

$$f(x_1, x_2) = \frac{1}{2} \ (-1 \leq x_1, x_2 \leq 0, 0 \leq x_1, x_2 \leq 1) \tag{3.27}$$

で与えられているとする。このとき確率変数 X_1^2 と X_2^2 は独立であるが、確率変数 X_1 と X_2 は独立でない。この例が定理 **3.3(ii)** と矛盾しないことは。$Y_j = X_j^2$ $(j = 1, 2)$ とすると $(X_1, X_2) = (\sqrt{Y_1}, \sqrt{Y_2})$ with$1/2$; $(X_1, X_2) = (-\sqrt{Y_1}, -\sqrt{Y_2})$ with$1/2$ により分かる。

定理 3.4： **1** 次元確率変数 X_1, X_2 が独立とする。X_i $(i = 1, 2)$ の (周辺) 確率分布

$F_i(x_i)$, (周辺) 密度関数 $f_i(x_i)$ とする。
(i) 和 $Y = X_1 + X_2$ の分布関数は

$$G(y) = \int_{-\infty}^{\infty} F_1(y - z_2) dF_2(z_2) \tag{3.28}$$

で与えられる。特に X_1, X_2 が密度関数を持つときには Y の密度関数は

$$g(y) = \int_{-\infty}^{\infty} f_1(y - z_2) f_2(z_2) dz_2 \tag{3.29}$$

で与えられる。
(ii) 和 $Y = X_1 + X_2$ の特性関数は

$$\psi(t; X_1 + X_2) = \psi(t; X_1)\psi(t; X_2) \tag{3.30}$$

で与えられる。

証明 : 和の分布関数は条件付分布により表現すると

$$\begin{aligned} G(y) &= \int_{-\infty}^{\infty} P(X_1 + X_2 \leq y | X_2 = z_2) dF_2(z_2) \\ &= \int_{-\infty}^{\infty} F_1(y - z_2) dF_2(z_2) \end{aligned}$$

より得られる。さらに

$$\begin{aligned} G(y) &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{y} f_1(z_1 - z_2) dz_1 \right] f_2(z_2) dz_2 \\ &= \int_{-\infty}^{y} \left[\int_{-\infty}^{\infty} f_1(z_1 - z_2) f_2(z_2) dz_2 \right] dz_1 \end{aligned}$$

という表現が得られる。両辺を **y** で微分すると密度関数の表現が得られる。**(ii)** は定理 **3.3(ii)** より期待値を計算すると得られる。 **Q.E.D.**

例 **3.8** (正規分布とポアソン分布の例) : 確率変数 $X_j \sim N(\mu_j, \sigma_j^2)$ $(j = 1, 2)$ とする。特性関数は $\psi(t; X_j) = \exp[it\mu_j - (1/2)t^2\sigma_j^2]$ であるから、和 $Y = X_1 + X_2$ の特性関数は $\psi(t; X_1 + X_2) = \exp[it(\mu_1 + \mu_2) - (1/2)t^2(\sigma_1^2 + \sigma_2^2)]$ である。したがって $Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2))$ となる。独立和の分布が元の分布型に一致する時、があると云う。
ポワソン分布の例では $X_j \sim P_o(\lambda_j)$ $(j = 1, 2)$ とする。特性関数は $\psi(t; X_j) = \exp[\lambda_j(e^{it} - 1)]$ であるから、和の特性関数は $\psi(t; X_1 + X_2) = \exp[(\lambda_1 + \lambda_2)(e^{it} - 1)]$ より $Y \sim P_o(\lambda_1 + \lambda_2)$ となる。

こうした例の特色は特性関数の形が指数関数系に属することによる。一般には分布が再生性を持つとは限らないので、畳込み (**convolution**) を用いて積分を評価する必要がある。

例 **3.9**(無相関と独立性) : 二つの確率変数 X, Y が独立であれば無相関であるが、逆は必ずしも成立しない。確率変数 X が原点に対称な密度関数 $f(x)$ を持つとき、確率変数 $Y = I(|X| \leq 1)$ とすると無相関になる。しかし明らかに X, Y は独立でない。

統計的分析ではしばしば独立な確率変数列の和の挙動を調べたいことが生じる。例えばクロスセクション・データをランダム標本の実現値と見なしてデータの平均により確率分布（母集団）の期待値を推定する問題が典型である。

独立性からの逸脱：二つの拡張方向

ここで独立な確率変数列を X_j $(j = 1, \cdots, n)$ とする。$n = 2$ の場合の議論より一般に有限個の独立な確率変数の和の分布、あるいは確率変数がベクトルの場合にも同様の結果を得ることが可能である。例えば上の定理より n 個の独立な確率変数 X_i $(i = 1, \cdots, n)$ に対し期待値、分散、特性関数はそれぞれ

$$\mathbf{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbf{E}[X_i] \,, \tag{3.30}$$

$$\mathbf{Var}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbf{Var}[X_i] \,, \tag{3.31}$$

$$\psi(t; \sum_{j=1}^{n} X_j) = \prod_{j=1}^{n} \psi(t; X_j) \tag{3.32}$$

が成り立つ。

ここで各確率変数の分布があらかじめ正規分布やポワソン分布となることが分かっている場合には確率変数の和の挙動の分析は比較的容易であるが、確率分布が分かっていない場合も実際の応用ではしばしば起きる。そこで独立な確率変数列 X_j $(j = 1, \cdots, n)$ より作られた和

$$M_n = \sum_{j=1}^{n} X_j \tag{3.33}$$

が $n \to \infty$ となるときの挙動を調べる必要が生じる。歴史的には大数の法則や中心極限定理は応用上での必要性から得られた。また時間的経過の中で得られる確率的現象、データの解析上では独立性の仮定が現実的でないことがある。

ここで独立性を緩める方向として二つの可能性がある。ここで議論の簡単化のために確率変数の期待値 $\mathbf{E}[X_j] = 0$ $(j = 1, \cdots, n)$ としておく。第一には、独立和は $M_n = M_{n-1} + X_n$ であるから $\mathbf{E}[M_n|M_{n-1}] = M_{n-1} + \mathbf{E}[X_n]$ より条件

$$\mathbf{E}[M_n|M_{n-1}, M_{n-2}, \cdots, M_1] = M_{n-1} \tag{3.34}$$

を満足することに注目する。この性質を満足する確率変数の系列（確率過程と呼ぶ）はマルチンゲール **(martingale)** と呼ばれているが、独立和の拡張としての一つの確率過程である。古典的な独立性の下での結果は、より本質的にはマルチンゲールについて成り立つ。

第二の方向は確率変数例 $\{X_j\}$ 自体に従属性を認めることが考えられる。互いに独立な確率変数列 $\{Z_i\}$, 定数 a(ただし $-1 < a < 1$) を用いて

$$X_i = aX_{i-1} + Z_i \ \ (i = 1, \cdots, n) \tag{3.35}$$

を満たす確率変数列 $\{X_i\}$ が考えられる。一般に確率変数列 $\{X_i\}$ の周辺確率分布が一定であれば定常性を持つと呼ぶが、この確率過程はマルコフ型の時系列モデルである。例えば $\{Z_i\}$ が互いに独立で $N(0, \sigma^2)$ にしたがう確率変数列とすると、応用上もよく利用される 1 次自己回帰モデルが得られる。この場合に $S_n = \sum_{j=1}^{n} X_j$ とすると、大数の法則や中心極限定理が成り立つが、$\mathbf{Cov}[X_j X_{j-k}] = a^{|k|}$ となるので分散の和公式は成立しない。こうした状況での統計的分析は統計学における統計的時系列解析や確率過程の解析、などでの主要なテーマとなっている。

4　確率変数の和とマルチンゲール

統計学の成立への一つの源流として生命表 (**life table**) の作成と生命保険の展開がしばしば言及される。個人の死亡する事象はそれぞれの関係者にとっては重大事であるが一見すると予測可能でない。しかしながら個々の死亡する確率を大量の観察データの平均をとると、「不特定多数の保険加入者が非常に多ければ」データから推定した死亡率が生命保険業を営むことが可能なようにほぼ実現する。こうした観察事実を説明、数理的に精緻化する理論として互いに独立である確率変数列の和に関する大数の法則や中心極限定理が考察され、古典的な確率論における中心的内容となった。本章では互いに独立な確率変数和の挙動をやや現代的なマルチンゲールを巡る視点より説明する。

4.1 大数の弱法則

確率変数列 $X_j\ (j=1,\cdots,n)$ より作り出す確率変数和 $S_n=\sum_{j=1}^n X_j$ について n が大きいときにその漸近挙動を考察する。期待値 $\mathbf{E}[S_n]=\mu(n)$, 分散 $\mathbf{V}[S_n]=\sigma^2(n)$ としてマルコフの不等式 (チェビシェフの不等式) を適用すると次の結果が得られる。

定理 **4.1**： $n\to\infty$ のとき $\sigma^2(n)/n^2\to 0$ とする。任意の正実数 $\epsilon>0$ に対し $n\longrightarrow\infty$ のとき

$$P\left(|\frac{S_n}{n}-\frac{\mu(n)}{n}|>\epsilon\right)\leq\frac{1}{\epsilon^2}\mathbf{V}(\frac{S_n}{n})=\frac{1}{\epsilon^2}\frac{\sigma^2(n)}{n^2}\longrightarrow 0 \tag{4.1}$$

となる。

この結果を

$$\frac{S_n}{n}-\frac{\mu(n)}{n}\xrightarrow{p}0 \tag{4.2}$$

と表すが、これは左辺が **0** に確率的に収束するという確率収束の意味である。確率変数の収束には実数列の収束とは異なり収束の意味から幾つかの収束概念があるが、定理 **4.1** は大数の弱法則 (**weak law of large numbers**) と呼ばれている。ここで「n が大きいときに確率変数列が確率収束する」をより正確に次のように定義する。

定義 **4.1** (確率収束)**:** 任意の正数 ϵ に対し $n\to\infty$ のとき

$$\lim_{n\to\infty}P(\omega||Y_n-Y|>\epsilon)=0 \tag{4.3}$$

となるとき確率変数列 Y_n は確率変数 Y に確率収束 (**convergence in probability**) すると云い、$Y_n\xrightarrow{p}Y$, あるいは $Y_n-Y\xrightarrow{p}0$ と表現する。

例 **4.1** : 互いに独立な確率変数列 X_j $(j = 1, \cdots, n)$ の期待値 $\mathbf{E}[X_j] = \mu$ (一定)、分散 $\mathbf{V}[X_j] = \sigma^2$ (一定) とする。このとき $\sigma^2(n) = n\sigma^2$ より

$$\frac{S_n}{n} \xrightarrow{p} \mu \tag{4.4}$$

となる。ここで期待値を最初から引いて確率変数列 $X_j^* = X_j - \mathbf{E}[X_j]$ $(j = 1, \cdots, n)$ についての和を M_n とすると $n \to \infty$ のとき

$$\frac{M_n}{n} \xrightarrow{p} 0 \tag{4.5}$$

である。

応用上では互いに相関のある確率変数列となる時系列や確率過程も重要である。互いに相関がある確率変数の和についても、相関があまり強くなければ大数の法則が成り立つ。ここでは相関がある確率変数列の典型例を挙げておく。

例 **4.2** : 互いに独立な確率変数列 $\{Z_j\}$, 定数 a(ただし $|a| < 1$) を用いて確率差分方程式

$$X_j = aX_{j-1} + Z_j \ \ (j = 1, \cdots, n) \tag{4.6}$$

を満たす確率変数列 $\{X_j\}$ を考える。時点 j において既に定まっている aX_{j-1} にノイズ U_j が加わり X_j が次々に定まることを意味する。こうした確率変数列を確率過程と呼ぶと、マルコフ型の 1 次自己回帰モデルである。例えば $\{Z_i\}$ が互いに独立に分散 σ^2 の確率分布 $N(0, \sigma^2)$ にしたがい、初期条件 $X_0 = 0$ とする (**1** 次自己回帰モデル)。$|a| < 1$ であれば X_j は $Cov(X_i, X_j) \sim a^{|i-j|}$ となるので上の定理より大数の法則は成立する。

ここで n を時間ととして互いに独立とは限らない確率変数列 $\{X_n\}$ の定常性 (**stationarity**) は次のように定義する。

定義 **4.2** (弱定常過程) : 確率変数列 $\{X_n\}$ が
(i) 任意の n に対して $\mathcal{E}(X_n) = \mu$,
(ii) 任意の m, n に対して $Cov(X_m, X_n) = \gamma(|n - m|)$
を満足するとき、弱定常過程 (**weakly stationary process**) と呼ぶ。関数 $\gamma(k)$ $(k = 1, 2, \cdots)$ を自己共分散関数 (autocovariance function) と呼ぶ。

例 **4.2** の続き : 確率変数列 $\{X_j\}$ が弱定常であれば両辺の期待値をとれば $\mathbf{E}(X_j) = a\mathbf{E}(X_{j-1})$ よりゼロとなる。係数について条件 $-1 < a < 1$ が満たされれば両辺の二乗の期待値より

$$\mathbf{E}(X_j^2) = a^2\mathbf{E}(X_{j-1}^2) + \sigma^2 \tag{4.7}$$

となるが定常性から $\gamma(0) = \mathbf{E}(X_j) = \mathbf{E}(X_{j-1}^2) = \sigma^2/(1-a^2)$ となる。共分散関数は $\gamma(k) = \mathbf{E}(X_j X_{j-k}) = a^k\gamma(0)$ となるので自己相関関数は $\rho(k) = a^k$ で与えられる。$|a| < 1$ であれば定理 4.1 の条件を満足する。

実数列 y_n が与えられたとき $y_n \to y$ に収束するとは、「任意の $\epsilon > 0$ に対し n_0 が存在して $n \geq n_0$ であれば $|y_n - y| < \epsilon$」と定められる。確率変数列 $Y_n(\omega)$ について確率の意味での収束 (convergence) には定義 4.1 の他に幾つかの異なる概念があるが、期待値の意味での収束は評価しやすいので応用上では役に立つ。

定義 4.3 (平均二乗収束): 確率変数列の分散が存在し

$$\lim_{n\longrightarrow\infty} \mathbf{E}[(Y_n - Y)^2] = 0 \tag{4.8}$$

となるとき確率変数列 Y_n は確率変数 Y へ平均二乗収束 (convergence in mean) すると云い、$l.i.m.\ Y_n \longrightarrow Y$ あるいは ($\xrightarrow{L^2}$) と記す。同様に p 次平均収束 $(p = 1, 2, \cdots)$ $\lim_{n\to\infty} \mathbf{E}[|Y_n - Y|^p] = 0$ も定義する。

例 4.3 : 独立な確率変数列 X_i $(i = 1, \cdots)$ の期待値 $\mathbf{E}[X_i] = \mu$, 分散 $\mathbf{V}[X_i] = \sigma^2$ とする。標本分散

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \xrightarrow{p} \sigma^2 \tag{4.9}$$

となることが期待される。ここで標本平均 $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ である。既に大数の弱法則より $\bar{X}_n \xrightarrow{p} \mu$ であるので

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{p} \sigma^2 \tag{4.10}$$

となる必要があるが、$\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X}_n - \mu)^2$ という関係を利用すればよい。なお $Y_i = (X_i - \mu)^2 - \sigma^2$ とするとマルコフの不等式を利用すると大数の弱法則が成り立つ十分条件としては、(少し強い条件であるが) 例えば $\mathbf{E}[X_i^4] < \infty$ とすればよいのである。

ここで定理 4.1 の証明より確率変数列 Y_n が Y に平均二乗収束すれば確率収束することがわかる。しかしながら、逆に確率収束するからといって平均二乗収束することにはならない。収束に関連して幾つかの例を挙げておく。

例 4.4 : 確率変数列 $\{Y_n\}$ を $Y_n = n^2$ (確率 n^{-1}), $Y_n = 0$ (確率 $1 - n^{-1}$) とする。このとき確率収束の定義より $Y_n \xrightarrow{p} 0$ である。期待値は $\mathbf{E}(Y_n) = n^2 \times n^{-1} + 0 \times (1 - n^{-1}) = n \longrightarrow \infty$ となりゼロに収束しない。$n \to \infty$ のとき $\mathbf{E}(Y_n^2)$ は発散する。

例 4.5： ルーレットの針が角度 $[0, 2\pi)$ 上のどこかに落ちる現象をイメージする。標本空間 $\mathbf{\Omega} = [0, 2\pi)$, 確率は弧の相対的な長さをとる。確率変数列を $X_1(\omega) = 1$ $(0 \leq \omega \leq 2\pi)$; $X_2(\omega) = 1$ $(0 \leq \omega \leq 2\pi \cdot \frac{1}{2}), 0$ (その他); $X_3(\omega) = 1$ $(2\pi \cdot \frac{1}{2} \leq \omega \leq 2\pi \cdot (\frac{1}{2} + \frac{1}{3})), 0$ (その他); などとして、各 n に対して X_n が 1 をとる定義域が $[0, 2\pi)$ を超えるときには ω より 2π(正整数) を引き定義域を 0 からの区間とする。(各 $X_n(\omega) = 1$ となる弧の 2π に対する相対的長さを $1/n$ とする。) このとき $0 < \epsilon < 1$ に対し

$$P(\omega||X_n| > \epsilon) = \frac{1}{n} \tag{4.11}$$

より $X_n \xrightarrow{p} 0$ だから確率変数列はゼロに確率収束する。ところが数列 $s_n = 2\pi(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n})$ は発散するので $P(\omega| \lim_{n\to\infty} X_n(\omega) = 0) = 0$ となる。すなわち確率収束、平均二乗収束は確率 1 の収束 (以下の定義 4.6 で述べる概念) とは異なり $P(\max_{1\leq j\leq n} X_j = 1) = 1$ なのである。

4.2 マルチンゲール

独立な確率変数の和を一般化し、時間の流れを $n = 1, 2, \cdots, \quad N$、情報が時間とともに変化する状況を考える。統計的分析では実際に時間とともに得られるデータを互いに独立な確率変数列の実現値、と見なすことが非現実的である場合が少なくないが、マルチンゲール理論は一つの解析手段を与えてくれる [17]

例えば期待値 $\mathbf{E}[X_i] = 0$ の互いに独立な確率変数列 $\{X_i\}$ に対し独立な確率変数の和を $M_n = \sum_{i=1}^{n} X_i$, 確率変数列 M_n より作られる σ-加法族を $\mathcal{F}_n = \mathcal{F}(M_n)$ とする。$\mathcal{F}_n$ は時点 n において利用可能な情報を意味するが、$M_n = M_{n-1} + X_n$ なので確率変数 M_{n-1} の情報も含むことから $\mathcal{F}_{n-1} \subset \mathcal{F}_n$ という増大列になる。このとき

$$\begin{aligned} \mathbf{E}[M_{n+1}|\mathcal{F}_n] &= \mathbf{E}[\sum_{i=1}^{n+1} X_i|\mathcal{F}_n] = \sum_{i=1}^{n} X_i + \mathbf{E}[X_{n+1}|\mathcal{F}_n] \\ &= M_n \ \ (a.s.) \end{aligned}$$

である。ここで almost surely(a.s.) とは確率 1 の事象を意味するがしばしば省略する。また条件付期待値は定義 3.2 より $\mathbf{E}[M_{n+1}|\mathcal{F}_n] = \mathbf{E}[M_{n+1}|M_n, M_{n-1}, \cdots, M_1]$ を意味する

一般に時点 0 において情報がないとみて $\mathcal{F}_0 = \{\phi, \mathbf{\Omega}\}$ とする。時点 1 では新たに情報が入り確率変数 M_1 の実現値が観察されるので、利用可能な情報を $\mathcal{F}_1$ とすると時点 0 では分からなかった可能性が 1 時点では分かるので $\mathcal{F}_0 \subset \mathcal{F}_1$ と表す。この操

[17] マルチンゲール理論は本シリーズ「確率過程の応用」で説明されるように経済やファイナンスと密接な関係がある。

作を繰り返して元の σ-加法族 $\mathcal{F}$ の () の増大列 $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$ をとる。つなわち任意の n に対して $\mathcal{F}_n \subset \mathcal{F}_{n+1}$, および $\mathcal{F}_\infty = \sigma(\bigcup_{n\geq 0} \mathcal{F}_n) \subset \mathcal{F}$ となる部分 σ 加法族の列を考察する。確率空間 $(\Omega, \mathcal{F}, P)$ 上に $\sigma-$ 加法族の列 $\{\mathcal{F}_n\}_{n\in\mathbf{N}}$ を加えた 4 つ組をと呼ぶとこの $(\Omega, \mathcal{F}, P, \{\mathcal{F}_n\}_{n\in N})$ が考察の対象である。ここで直観的解釈としては $\mathcal{F}_m = \sigma(M_0, M_1, \cdots, M_m)$ は確率変数列 $\{M_n\}$ の M_m までの情報 (information) を表す。

定義 4.4:(マルチンゲール) : 確率変数列 $\{M_n\}$ が $\mathcal{F}_n$-適合的であり ($\mathcal{F}_n$ で測れる (可測な) 確率変数 M_n を $\mathcal{F}_n-$ 適合的 (adapted) と呼ぶ) かつ可積分とする。このときマルチンゲール (martingale) とは

$$\mathbf{E}[M_n|\mathcal{F}_{n-1}] = M_{n-1} \ (a.s.) \tag{4.13}$$

を満足する確率変数列である

この定義では条件付期待値を用いていることが重要である。(無条件) 期待値を取ると $\mathbf{E}[\mathbf{E}[M_n|\mathcal{F}_{n-1}]] = \mathbf{E}[M_n] = \mathbf{E}[M_{n-1}]$ が得られるのでマルチンゲールの期待値は一定である。同様に優マルチンゲール (劣マルチンゲール) を $\mathbf{E}[M_n|\mathcal{F}_{n-1}] \leq M_{n-1}$ $(a.s.)$ $(\mathbf{E}[M_n|\mathcal{F}_{n-1}] \geq M_{n-1}$ $(a.s.))$ で定めると、確率系列の意味で増大列 (減少列) に対応する。ここで例によりマルチンゲールの有用性をを示そう。

例 4.6 (コイン投げ) : コインを何回も投げ、n 回投げるときの情報は $\sigma-$ 加法族の列 $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ である。表なら $X_n = 1$, 裏なら $X_n - 1$ とすると、情報にゆがみがないならば結果の確率変数の和 $M_n = \sum_{i=1}^n X_i$ はマルチンゲールとなる。

例 4.7 (公平なゲーム,Fair Game) : 何回も繰り返される (あるいは資産市場において (価格変動を伴う) 不確実な資産の売買) を考える。n 時点における価格を M_n $(n = 1, \cdots, N)$ として N_n は n 時点においてはじめて公示される (すなわち $\mathcal{F}_n-$ 適合的) とする。$n-1$ 時点から n 時点におけるゲームから予想される価格変化 $X_n = M_n - M_{n-1}$ とすると、確率変数列 $\{M_n\}$ を対象とするゲームにおいて「利得を得る戦略を立てる問題」、と解釈できる。このとき

$$\mathbf{E}[M_n - M_{n-1}|\mathcal{F}_{n-1}] = 0 \iff \text{公平なゲーム (Fair Game)} \tag{4.14}$$

を意味する。

例 4.8 (Random Walk) : しばしば (古典的) 確率論では互いに独立なベルヌイ試行 (二値をとる確率変数列) $X_j = 1$ (確率 1/2); $X_j = -1$ (確率 1/2) $(j = 1, \cdots, n)$ の和

$$M_n = X_0 + \sum_{j=1}^n X_j$$

を考える。この確率過程をランダム・ウオーク (random walk, 酔歩) と呼んでいる。ランダム・ウオークは X_0 を初期値とするマルチンゲールである。

確率変数列 $\{M_n\}$ から作り出される確率変数列 $X_n = M_n - M_{n-1}$ はマルチンゲール差分 (martingale difference) と呼ぶ。このとき $\mathbf{E}[X_n|\mathcal{F}_{n-1}] = 0$ である。なお一般には確率変数 $\{X_n\}$ は互いに無相関であっても独立な確率変数列とは限らない。また M_n の出発点を M_0 とするとマルチンゲール差分系列 $\{X_j\}$ により

$$M_n = M_0 + \sum_{j=1}^{n} X_j \tag{4.15}$$

と表現できる。

定義 4.5 (マルチンゲール変換) : マルチンゲール $\{M_n\}$ が $\mathcal{F}_n$-適合的、$\{C_i\}$ は $\mathcal{F}_{i-1}$-適合的 $(i \geq 1)$ とする。このときマルチンゲール変換は $Y_n = \sum_{i=1}^{n} C_i(M_i - M_{i-1})$ $(n \geq 1)$ で与えられる。ただし $M_0 = Y_0 = 0$ としておく。

初期値がゼロでなければ $Y_0 = M_0$ とすればよい。ここでマルチンゲール M_n で資産価格が表現される Fair Game において、時点 $i-1$ において C_i $(i = 1, \cdots, n)$ という戦略をとった時に時点 i で実現される利得 $C_i(M_i - M_{i-1})$ と解釈できる。時点 i におけるゲームの結果を見ないで時点 $i-1$ で予め戦略を立てるように制限しないと「ずる」,「まった」(将来の結果を見てから自分の態度を決める) という戦略が可能となる。

ここで $\forall m$ に対して $\{C_m\}$ が $\mathcal{F}_{m-1}$ 可測，$C_m(M_m - M_{m-1})$ が可積分としよう。このとき

$$\begin{aligned} \mathbf{E}[Y_n|\mathcal{F}_{n-1}] &= \sum_{i=1}^{n-1} C_i(M_i - M_{i-1}) + C_n\{\mathbf{E}[M_n|\mathcal{F}_{n-1}] - M_{n-1}\} \\ &= Y_{n-1} \end{aligned}$$

となる。したがってマルチンゲール変換により得られる確率変数列 Y_n は可積分性が成り立てば再びマルチンゲールとなる [18]。

[18] 時間間隔を例えば $[0,1]$ を n 等分して $t_{i-1} = (i-1)/n$ 時点に決定する (t_i 時点までとる) 戦略を C_{t_i}, マルチンゲール差分 $\Delta M_i = M_{t_i} - M_{t_{i-1}}$, $Y_n = \sum_{i=1}^{n} C_{t_i}\Delta M_i$ とする。$n \to \infty$ として得られるマルチンゲール変換は一定の仮定の下で伊藤積分 (Ito's Integral) と呼ばれる表現 $Y_t = \int_0^t C_u dM_u$ に対応する。適当な条件下でマルチンゲール変換がマルチンゲールとなることが伊藤理論がファイナンスで利用される一つの理由である。

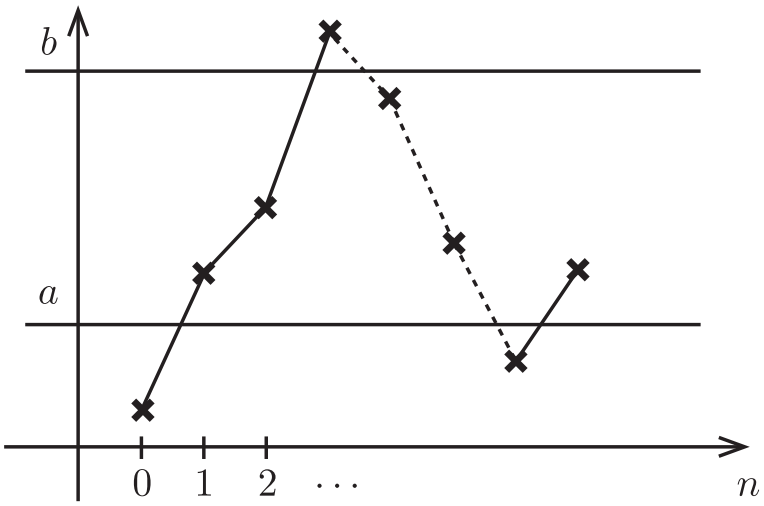

図 4-1: Buy at Low and Sell at High

次にマルチンゲール $\{M_n\}$ の系列が与えられたとき、十分に n が大きいときにこの確率変数列が収束する、すなわち
(C) $M_n \to M$ $a.s.$
となる条件を考察する。ここではマルチンゲール変換を利用して（経済的に意味のある）簡単な戦略を作ることにより問題に接近するが、仮定と結果の意味は明快と思われる。時間とともに変化するある資産価格がマルチンゲールにしたがうときの個人が次のように対応することを考えよう。
[**戦略**] 時点 k における公示される価格 M_k $(k = 1, \cdots, n)$, 2 つの実数を $a < b$ に対して次の条件を満たす戦略 $\{C_k\}$ を定める。条件 (i) 次に a 以下になるまで $C_k = 0$, 条件 (ii) a 以下になると b 以上になるまで $C_k = 1$ とする。この戦略を式で表現すると (i) $C_1 = I_{\{M_0<a\}}$, (ii) $\forall k \geq 2$ $C_k = I_{\{C_{k-1}=1\}} I_{\{M_{k-1}\leq b\}} + I_{\{C_{k-1}=0\}} I_{\{M_{k-1}<a\}}$ である。(ここで $I(A) = 1$ (A が真), $I(A) = 0$ (A が偽) という指標関数を意味する。) この戦略 $\{C_k\}$ からマルチンゲール変換 $Y_0 = 0, Y_n = \sum_{k=1}^{n} C_k(M_k - M_{k-1})$ を構成しよう。(この戦略から以下の結果を得るには横軸に n, 縦軸に $\{M_n, n \geq 1\}$ と 2 つの平行線 a, b を図 4.1 を見ることにより直観的に理解できる。M_n を時点 n の資産価格とすると「価格が a を下回れば買い、b を上回れば売り」を繰り返すときの Y_N は N 時点の損益である。不等式の最後の項は最後の買いの期間の損益に対応する。)

この戦略下では

$$U_N[a,b](\omega) = \{\forall n \leq N : M_n \text{下から上へ } [a,b] \text{ を上方に向かって渡る回数}\}$$

とすると不等式

$$Y_N(\omega) \geq (b-a)U_N[a,b](\omega) - [M_N(\omega) - a]^- \tag{4.16}$$

が成り立つ。(ただし確率変数 Z^- は $Z = Z^+ - Z^-$ $(Z^+ \geq 0, Z^- \geq 0)$ と分解した第 2 項である。) ここで両辺の期待値をとると $\{Y_n\}$ がマルチンゲールであるので

$$0 \geq (b-a)\mathbf{E}[U_N[a,b](\omega)] - \mathbf{E}[(M_N(\omega) - a)^-] \tag{4.17}$$

より Doob の不等式が得られる。

補題 4.2 (Doob の上渡回数定理): 確率変数列 $\{M_n\}$ を $\mathcal{F}_n$-マルチンゲールとする。このとき不等式

$$\mathbf{E}[(M_N - a)^-] \geq (b-a)\mathbf{E}[U_N[a,b]] \tag{4.18}$$

が成り立つ。

この補題は重要である。ここで確率変数列 $\{M_n\}$ が定義される事象の収束に関して

$$\begin{aligned} \{\omega \ : \ M_N(\omega) \text{ が収束 }\} &= \bigcup_{c \in \mathbf{R}} \{\omega \ : \ \lim_N M_N(\omega) = c\} \\ &= \bigcup_{c \in \mathbf{R}} \bigcap_{a<c<b} \{U_N[a,b] < +\infty\} \\ &= \bigcap_{\forall a<b} \{U_N[a,b] < +\infty\} \end{aligned}$$

となる関係に注目する。$N \longrightarrow \infty$ のとき $\mathbf{E}[U_N[a,b]] < \infty$ であれば $P(U_N[a,b] = \infty) = 0$ である。(すなわち N が大きくなっても区間をまたがる階数 $U_N[a,b]$ は有限なので、どの様な ω に対してもいつかは下から上に $[a,b]$ を横切ることがないことを意味する。) したがって次のマルチンゲール収束定理が得られる。これはマルチンゲールに関する重要な結果の一つである.

定理 4.3 (収束定理): 確率変数列 $\{M_n\}$ は $\mathcal{F}_n$-マルチンゲールであり，期待値についての条件

$$\sup_n \mathbf{E}[|M_n|] < +\infty \tag{4.19}$$

を仮定する．このとき可積分な確率変数 M が存在して

$$\lim_{n \to \infty} M_n(\omega) = M \ \ (a.s.) \tag{4.20}$$

となる.

なお例 4.4 やここで表れた収束概念は確率論ではもっとも強い概念であるが次のように定義される。

定義 4.6 (概収束): 確率空間 $(\mathbf{\Omega}, \mathcal{F}, P)$ において

$$P(\omega | \lim_{n\to\infty} Y_n(\omega) = Y(\omega)) = 1 \tag{4.21}$$

となるとき確率変数列 Y_n は確率変数 Y に概収束 (almost sure convergence) すると云い、$Y_n \longrightarrow Y$ $(a.s.)$ と表す。

4.3 大数の強法則

ここで互いに独立な確率変数列 X_i $(i = 1, \cdots)$ を規準化して $\mathbf{E}[X_i] = 0$ とする。確率変数列

$$M_n = \sum_{i=1}^{n} (\frac{1}{i}) X_i \tag{4.22}$$

により構成すると M_n はマルチンゲールになる。ここで例えば分散が有界であれば定理 4.3 の条件は満たされる。なお独立な確率変数和については分散についての条件は不要である。一般に大数の法則が成立する十分条件をさらに弱めることができる。ここでは $(\mathbf{E}[M_n])^2 \leq \mathbf{E}[M_n^2]$ および補題 4.6 を用いることより次の結果が得られる。

定理 4.4 : $\{X_i\}$ を互いに独立な確率変数列であり $\mathbf{E}(X_i) = 0,\ \mathbf{E}(X_i^2) = \sigma_i^2$ $(i = 1, \cdots, n)$ とする。分散についての条件

$$\sum_{i=1}^{\infty} \frac{1}{i^2} \sigma_i^2 < +\infty \tag{4.23}$$

が成り立てばマルチンゲール $n \longrightarrow \infty$ のとき $M_n = \sum_{i=1}^{n} (\frac{1}{i}) X_i$ は収束する。したがって $n \longrightarrow \infty$ のとき

$$\frac{1}{n} \sum_{k=1}^{n} X_k \to 0 \quad (a.s.) \tag{4.24}$$

となる。

系 4.5(大数の強法則) : 確率変数列 $\{X_i\}$ が互いに独立であって、期待値 $\mathbf{E}[X_i] = \mu$, 分散 $\mathbf{E}[X_i^2] = \sigma^2$ $(i = 1, \cdots, n)$ とする。$n \to \infty$ のとき

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \ (a.s.)$$

となる。

この定理 4.3〜定理 4.5 では収束において強い収束概念を用いていることが重要である。例えば慨収束すれば自動的に確率収束するが、逆は必ずしも成立しない。マルチンゲールの収束定理から得られたこの結果は大数の強法則 (strong law) と呼ばれるが、古典的議論ではかなり複雑な議論が必要である。

最後に大数の法則の証明に必要な Kronecker の補題を挙げておく。

補題 4.6 (Kronecker の補題)： $\{b_n\}$ を $\uparrow +\infty$ $(n \uparrow +\infty)$ となる正実数の増大列，和 $s_n = \sum_{i=1}^{n} x_i$ とする．

$$\sum_{k=1}^{\infty} \frac{x_k}{b_k} \text{ が収束 } \Rightarrow \frac{s_n}{b_n} \to 0 \tag{4.25}$$

となる．

補題 4.6 の証明： 部分和 $\{u_n\}$ を

$$u_n = \sum_{k=1}^{n} \frac{x_k}{b_k}$$

とすると仮定から $u_n \to u_\infty$ となる．ここで $u_n - u_{n-1} = x_n/b_n$ であるから $u_0 = 0$ として

$$\begin{aligned} \frac{s_n}{b_n} &= \frac{1}{b_n} \sum_{k=1}^{n} b_k (u_k - u_{k-1}) \\ &= u_n - \frac{1}{b_n} \sum_{k=2}^{n} (b_k - b_{k-1}) u_{k-1} \\ &\to u_\infty - u_\infty = 0 \quad (n \to \infty) \end{aligned}$$

より結論を得る。ただしこの導出では

$$\frac{1}{b_n} \sum_{k=2}^{n} (b_k - b_{k-1}) u_{k-1} \to u_\infty \quad (n \to \infty)$$

となることを用いたが、これは大きい N をとると $n \geq N$ に対して

$$\frac{1}{b_n} \sum_{k=1}^{n} u_{k-1}(b_k - b_{k-1}) = \frac{1}{b_n} \sum_{k=1}^{N} u_{k-1}(b_k - b_{k-1}) + \frac{1}{b_n} \sum_{k=N+1}^{n} u_{k-1}(b_k - b_{k-1})$$

と分解すると、右辺の第 1 項は $b_n \uparrow +\infty$ のとき 0 に収束し、第 2 項は u_∞ に収束する。このことは $k \geq N$ のとき $u_k > u_\infty - \epsilon\ (\epsilon > 0)$ とすると

$$\begin{aligned}\frac{1}{b_n}\sum_{k=N+1}^{n} u_{k-1}(b_k - b_{k-1}) &\geq (u_\infty - \epsilon)\frac{1}{b_n}\sum_{k=N+1}^{n}(b_k - b_{k-1}) \\ &= (u_\infty - \epsilon)\frac{[b_n - b_N]}{b_n} \to u_\infty - \epsilon\ ,\end{aligned}$$

また $k \geq N$ のとき $u_k < u_\infty + \epsilon\ (\epsilon > 0)$ とすれば確認できる。**Q.E.D.**

5　分布収束と中心極限定理

5.1 確率分布の収束の例

実数の収束とは異なり確率の意味での収束 (convergence) には幾つかの異なる概念がある。既に大数の弱法則は確率収束、大数の強法則は慨収束、に対応することを説明した。歴史的には大数の法則をより精緻化して、未知の確率分布をよく知られた分布（例えば正規分布）によって近似する方法の正当化の議論から分布収束が考えられた。

例 5.1： $\mathrm{B}(n,p)$ にしたがう確率変数 X_n としよう。確率変数 X_n は $[0,1]$ 上の短い区間 $[\frac{i-1}{n}, \frac{i}{n}]$ $(i = 1, \cdots, n)$ 上で 1 または 0 をとる独立なベルヌイ試行列 Z_j の和 $X_n = \sum_{j=1}^{n} Z_j$ と表現される。ここで短い区間において事象が起きる確率 p_n が小さいとして $np_n \to \lambda\ (>0)$ となる状況を考えよう。確率変数列 Z_j が互いに独立であるから確率変数 X_n の特性関数は

$$\begin{aligned} \psi_n(t; X_n) &= \mathbf{E}[e^{itX_n}] \\ &= \prod_{j=1}^{n} \mathbf{E}[e^{itZ_j}] \\ &= [p_n e^{it} + (1-p_n)]^n \end{aligned}$$

で与えられる。ここで n が大きいときには $p_n \sim \lambda/n$ より対数をとって評価すると

$$\begin{aligned} \log[\psi_n(t; X_n)] &= n \log[1 + p_n(e^{it} - 1)] \\ &\sim np_n(e^{it} - 1) \to \lambda(e^{it} - 1) \end{aligned}$$

となる。したがって、X_n の特性関数は $\psi_{X_n} \to \psi(t)$ よりポワソン分布 $(P_o(\lambda))$ の特性関数 $\psi(t)$ に収束する。

二項分布の確率評価を n が大きく $n\,p\ (= \lambda_n)$ が一定の場合にポワソン分布によりよく近似できること、また稀に起きる事象についての事象が生じる回数の分布としてポワソン分布で近似できること、などは古くから知られている。例 5.1 はこうした観察事実の数理的根拠を与えるので小数の法則 (law of small numbers) と呼ばれる。

例 5.2： 二項分布において確率 p が一定のとき n が大きいときには De Moivre は確

率変数 X_n の確率分布がある意味で正規分布で近似できることを見いだした。X_n の期待値は np, 分散は $np(1-p)$ であるから確率変数

$$Y_n = \frac{X_n - np}{\sqrt{np(1-p)}} \tag{5.1}$$

と規準化すると、確率変数 Y_n の期待値 0, 分散 1 となる。このとき Y_n の確率分布の挙動は特性関数により評価することができる。確率変数 Y_n の特性関数は

$$\begin{aligned}\psi_n(t;Y_n) &= \mathbf{E}[e^{it\frac{Y_n}{\sqrt{np(1-p)}}-it\frac{np}{\sqrt{np(1-p)}}}] \\ &= e^{-it\frac{np}{\sqrt{np(1-p)}}}\left[pe^{it\frac{1}{\sqrt{np(1-p)}}}+(1-p)\right]^n\end{aligned}$$

で与えられる。ここで対数をとると

$$\log\psi_n(t;Y_n) = -it\frac{np}{\sqrt{np(1-p)}} + n\log\left[1+p(e^{it\frac{1}{\sqrt{np(1-p)}}}-1)\right]$$

であるから、第二項は n が大きいとき

$$\begin{aligned}& n\log\left[1+p(e^{it\frac{1}{\sqrt{np(1-p)}}}-1)\right] \\ \sim\ & n\left[p(e^{it\frac{1}{\sqrt{np(1-p)}}}-1)-\frac{1}{2}p^2(e^{it\frac{1}{\sqrt{np(1-p)}}}-1)^2\right] \\ \sim\ & n\left[pit\frac{1}{\sqrt{np(1-p)}}+\frac{1}{2}p(it\frac{1}{\sqrt{np(1-p)}})^2-\frac{1}{2}p^2(it\frac{1}{\sqrt{np(1-p)}})^2\right]\end{aligned}$$

と評価できる。したがって第一項と第二項をあわせて $n\to\infty$ のとき、任意の t に対して

$$\psi_n(t;Y_n) \longrightarrow \psi(t) = e^{-\frac{1}{2}t^2} \tag{5.2}$$

となる。この $\psi(t)$ は標準正規分布の特性関数である。

定義 5.1 (分布収束)： 確率分布関数 F, F_n をそれぞれ確率変数 Y, Y_n の分布関数とする。F の任意の連続点 x において

$$F_n(x) \longrightarrow F(x) \quad (n\longrightarrow\infty) \tag{5.3}$$

となるとき，Y_n は Y に分布収束 (convergence in distribution) すると云い、$Y_n \xrightarrow{d} Y$ (あるいは $Y_n \xrightarrow{\mathcal{L}} Y$) と記す。

例 5.3： 数列 $x < x_n \to x$ のとき 1 点をとる確率変数列 $P(X_n = x_n) = 1$ とすると

$F(x)$ は不連続点になる。

例 5.4 : 確率変数 X, Y は互いに独立な確率変数で 2 点 1,0 を確率 1/2 をとるとする。$X_n = Y$ とすると $X_n \xrightarrow{d} Y$ であるが $P(|X_n - X| = 1) = \frac{1}{2}$ なので $X_n \xrightarrow{d} X$ ではない。確率変数列の確率収束は分布収束を意味するが確率変数列が分布収束しても確率収束するとは限らない。

分布関数の収束は確率の収束としては慨収束や確率収束に比べて弱い条件を課している。$n \to +\infty$ につれて確率分布の意味で収束するとき、そこで確率変数 X_n に対する確率分布を Q_n とすると確率変数 X に対する確率分布 Q に弱収束 (weak convergence) すると云い， $Y_n(t) \overset{w}{\Longrightarrow} Y(t)$ あるいは $Q_n \overset{w}{\Longrightarrow} Q$ と表す。

確率分布の収束を直接に調べるのは簡単でない場合が多い。そこで確率分布と特性関数は一対一対応していることに注目する。統計学ではしばしば確率変数列の特性関数を求め，その収束先を調べることにより確率分布の収束を調べることが行われる。ここで特性関数の収束先が分布関数の特性関数となるのが自明でなければ簡単な条件が必要となる。

補題 5.1 : F_n を分布関数列，φ_n を対応する特性関数列であって
(i) $\varphi_n(t) \to \varphi(t)$ $\forall t \in \mathbf{R}$,
(ii) $\varphi(\cdot) : t = 0$ で連続（あるいは (ii)$'$ $\varphi(t)$ は特性関数）とする。
このとき $F_n \xrightarrow{\mathcal{L}} F$ であって $\varphi(t)$ は F の特性関数となる。

補題 5.1 より例 5.1 と例 5.2 における特性関数の評価より確率分布のポアソン分布や正規分布への収束が正当化できる。前者は小数の法則、後者は中心極限定理 (CLT) に対応している。

5.2 中心極限定理

独立な確率変数列については古くから中心極限定理 (central limit theorem, 略して CLT) として研究され、統計的分析では広範に利用されている。ここで互いに独立に同一の確率分布にしたがう確率変数列 X_j $(j = 1, \cdots, n)$ の期待値 $\mathbf{E}[X_j] = \mu$, 分散 $\mathbf{V}[X_j] = \sigma^2$ として、規準化した標本平均

$$Y_n = \sum_{j=1}^{n} \left[\frac{X_j - \mu}{\sqrt{n}\sigma} \right] = \frac{\sqrt{n}}{\sigma} \left[\bar{X}_n - \mu \right] \tag{5.4}$$

とおく。確率変数 $X-\mu$ の特性関数を $\psi(t)$, Y_n の特性関数を $\psi_n(t;Y_n)$ として

$$\psi(\frac{t}{\sigma\sqrt{n}}) = \mathbf{E}\left[\exp(\frac{(X-\mu)it}{\sigma\sqrt{n}})\right] \tag{5.4}$$

の中の指数関数を展開 (Taylor 展開) すると、ある θ ($|\theta| \leq 1$) が存在して

$$\begin{aligned}
\psi_n(t;Y_n) &= \left[\psi(\frac{t}{\sqrt{n}\sigma})\right]^n \\
&= \left\{\mathbf{E}[1+\frac{1}{1!}\frac{it}{\sigma\sqrt{n}}(X-\mu)+\frac{1}{2!}\frac{(it)^2}{\sigma^2 n}(X-\mu)^2+\theta\frac{1}{3!}(\frac{it}{\sigma\sqrt{n}})^3(X-\mu)^3]\right\}^n \\
&= \left[1-\frac{1}{2}\frac{t^2}{n}+o(\frac{1}{n})\right]^n
\end{aligned}$$

となる。ここで $o(\ \cdot\ )$ は小さいオーダーを意味し、$M = o(1/n)$ とは $M_n/(1/n) \to 0$ as $n \to \infty$ である。簡単にこの条件を保証するには最後の項が有界と仮定すれば両辺を対数変換して $n \to \infty$ とすると次の結果が得られる。

定理 5.2 ： 互いに独立で同一の確率分布にしたがう有界な確率変数列 X_j $(j = 1, \cdots, n)$ の期待値 $\mathbf{E}[X_j] = \mu$, 分散 $\mathbf{V}[X_j] = \sigma^2$(一定) とする。規準化した標本平均 $Y_n = n^{-1/2}\sum_{j=1}^n (X_j - \mu)$ は $n \to \infty$ のとき

$$Y_n \xrightarrow{\mathcal{L}} N(0, \sigma^2) \tag{5.5}$$

となる。

上の結果は

$$\frac{1}{\sqrt{n}}\sum_{j=1}^n (X_j - \mu) \xrightarrow{\mathcal{L}} N(0, \sigma^2) \tag{5.6}$$

であるから、確率計算は

$$\sum_{j=1}^n X_j \sim N(n\mu, n\sigma^2)$$

と近似して行えばよいことを意味している。中心極限定理は極限の挙動についての理論であるが、ここで n が有限ならより正確な近似を考えることも可能である。標準正規分布の密度関数 $\phi(y)$, 適当に定数 c_2 を選び例えば $o(1/\sqrt{n})$ まで

$$P(Y_n \leq y) \sim \Phi(y) + \frac{1}{\sqrt{n}}[c_2(y^2-1)]\phi(y) \tag{5.7}$$

とするエッジワース展開が知られている。ここで i.i.d.(互いに独立に同一分布にしたがう) 確率変数列 X_i より基準化して $Y_n = \sum_{i=1}^{n}(X_i - \mu)/[\sigma\sqrt{n}]$ のとき γ_3 を $\{X_i\}$ の 3 次のキュムラントとすると $c_2 = (1/6)[-\gamma_3/\sigma^3]$,2 次エルミート (Hermite) 多項式 $h_2(y) = y^2 - 1$ により表現できるのである [19]。

例 5.5： 互いに独立に期待値 $\mathbf{E}[X_j] = \mu$, 分散 $\mathbf{V}[X_j] = \sigma^2$ の確率変数列の平均について定理 5.2 より

$$\frac{1}{\sqrt{n}}\left[\sum_{j=1}^{n}(\frac{X_j - \mu}{\sigma})\right] \tag{5.8}$$

は近似的に $N(0,1)$ となる。標本分散

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \xrightarrow{p} \sigma^2 \tag{5.9}$$

の挙動を考察すると、

$$\sqrt{n}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 - \sigma^2] \xrightarrow{d} N(0, (2+\kappa_4)\sigma^4) \tag{5.10}$$

となるが、ここで

$$\kappa_4 = \mathbf{E}[\frac{(X_i - \mu)^4}{\sigma^4}] - 3$$

で定義される kurtosis(尖度) である。(なお尖度は $\kappa_4 + 3$ で定義されることもある。) 同様に skewness(歪度) は $\kappa_3 = \mathbf{E}[(X_i - \mu)^3/\sigma^3]$ で定義する。特に正規分布の場合には $\kappa_4 = 0$ である。上の漸近分布を示すにはまず大数の弱法則より $\bar{X}_n \xrightarrow{p} \mu$ に注目する。さらに分散の評価より

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{p} 0 \tag{5.11}$$

であるので $\sqrt{n}[(1/n)\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 - \sigma^2] = \sqrt{n}[(1/n)\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2] - \sqrt{n}(\bar{X}_n - \mu)^2$ の右辺の第 2 項はゼロに確率収束する。そこで右辺の第一項に対して中心極限定理を用いると

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(X_i - \mu)^2 - \sigma^2\right] \xrightarrow{d} N(0, V[(X_i - \mu)^2])$$

となる。ここで右辺の分散は $V[(X_i - \mu)^2] = \mathbf{E}[(X_i - \mu)^4] - \sigma^4 = (2 + \kappa_4)\sigma^4)$ で与えられる。

[19] 例えば 竹内啓「確率分布の近似」(第 4 章, 教育出版 1975 年) に詳しい。

独立・同一分布に関する応用上の必要性から中心極限定理は様々な方向に拡張されている。例えば、独立変数列はマルチンゲール差分系列なので，何らかの条件の下でマルチンゲールに対して中心極限定理が成り立ち，正規分布による近似が意味を持つことが期待される。ここでが成り立たない例を示しておこう。

例 5.6： 互いに独立な確率変数列 $\{X_i\,, i = 1, \cdots, n\}$ を

$$X_i = \begin{cases} 1 & \text{確率 } \frac{1}{2} \\ -1 & \text{確率 } \frac{1}{2} \end{cases}$$

として，$\{X_i\}$ より確率変数列 $\{S_n\}$ を $S_n = \sum_{i=2}^{n} X_i$ により構成する．さらに，確率変数列 $\{W_i\}$ を $W_1 = 0$ であって $\forall n \geq 2$ については

$$W_n = \begin{cases} 1 & (X_1 = 1) \\ 2 & (X_1 = -1) \end{cases}$$

とおく．すると $n \to +\infty$ のとき中心極限定理より

$$\frac{S_n}{\sqrt{n}} \xrightarrow{\mathcal{L}} N(0,1) \tag{5.12}$$

となる。他方、確率変数列 $\{Y_n\}$ を $Y_n = \sum_{i=1}^{n} W_i Y_i$ により構成すると

$$\frac{Y_n}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i X_i \xrightarrow{\mathcal{L}} \frac{1}{2} N(0,1) + \frac{1}{2} N(0, 2^2)\,, \tag{5.13}$$

となる。ここで右辺は確率 $1/2$ で二つの確率分布 $N(0,1), N(0,2^2)$ にしたがう、すなわち混合分布を意味する。

次に応用上で有用な中心極限定理について述べておく。今までの規準化を変更して改めて各 n に依存する独立な三角型の確率変数列 $\{X_{n,k}\}$,$1 \leq k \leq n, n \geq 1$ および独立和の確率変数列 $\{Y_{n,k}\}$,$1 \leq k \leq n, n \geq 1$ を $X_{n,k} = Y_{n,k} - Y_{n,k-1}$ $(k = 1, \cdots, n)$ により構成しよう。各 X_j の分散を $\sigma_{n,k}^2 = E[X_{n,k}^2]$ としよう。このとき条件

$$\text{(I)} \qquad \sum_{k=1}^{n} \sigma_{n,k}^2 \longrightarrow \sigma^2 > 0 \text{ (一定値)}$$

$$\text{(II)} \qquad \sum_{k=1}^{n} \mathbf{E}[X_{n,k}^2 I_{\{|X_{n,k}| \geq \epsilon\}}] \longrightarrow 0$$

の後者はリンドバーグ (Lindberg) 条件と呼ばれている。この条件は要は個々の確率変数が分散の意味で全体への影響が小さく

$$\max_{1\leq k\leq n}\sigma_{n,k}^2 \longrightarrow 0 \tag{5.14}$$

を意味する。

また条件 (II) をより強く右辺の中をある δ (>0) を用いて $1\leq[|X_{n,k}|/\epsilon]^{\delta}$ で置き換えると

$$\text{(III)} \qquad \sum_{k=1}^{n}\mathbf{E}[|X_{n,k}|^{2+\delta}] \longrightarrow 0$$

が十分条件となる (特に $\delta=1$ のときリヤプノフ (Lyapunov) 条件と呼ばれる)。これらの条件の下で次の中心極限定理を得るが証明の概略は補論で説明しておこう。

定理 5.3： 独立な確率変数の和について分散 $\sigma_{n,k}^2$ について条件 (I) と Lindberg 条件 (II) を仮定する。$n\to\infty$ のとき

$$\sum_{k=1}^{n}X_{n,k}\xrightarrow{\mathcal{L}}N(0,\sigma^2) \tag{5.15}$$

となる。

5.3 不変原理

独立な確率変数の和に関する中心極限定理は数理統計学や応用問題においては様々な形で利用されている。中心極限定理は確率変数の和の分布が正規分布で近似できる根拠を与えている。近年では時間間隔が短い状況での分析、例えば金融市場の統計分析なども盛んに行われるようになった。中心極限定理の拡張として次に述べる不変原理 (invariance principle) による連続時間の確率過程として知られているブラウン運動が中心極限定理から自然に導かれることを主張している。

定理 5.4： 区間 $[0,1]$ 上を n 等分して各 i/n 時点で確率変数 Z_i の期待値 $\mathbf{E}[Z_i]=0,\mathbf{V}[Z_j]=\sigma^2/n$ とする。(微少な区間での確率変数の区間当たりの分散を $\sigma^2\times(1/n)$ とし。) $k(n)$ を n に依存させて各時刻 $(k(n)/n)=t(k_n)\in[0,1]$ 上の $k(n)$ 個の確率和 $X_n(t(k_n))=(\frac{1}{\sigma})\sum_{i=1}^{k(n)}Z_i$ とする。($[a]$ は a を超えない最大の整数を表す。) さらに $[0,1]$ 上の任意の時刻 t において補間により $X_n(t)$ が連続経路をとるようにしておく。$n\to\infty$ のとき $t(k_n)\to t\in[0,1]$ に対して

$$X_n(t(k_n))\xrightarrow{d}X(t)\ , \tag{5.16}$$

ただし $X(t) \sim N(0,t)$ となる確率過程であり、任意の $0 \leq s < t \leq 1$ に対し $X(s)$ と $X(t) - X(s)$ は独立な確率変数である。

この確率変数 $X(t)$ $(t \in [0,1])$ はブラウン運動 (Brownian Motion) と呼ばれているが時間軸上の経路は不規則な変動を示している。任意の t に対して正規分布 $N(0,t)$ にしたがい、t について連続な経路をとる確率変数列 $X(t)$ が存在するが、このブラウン運動は任意の ω に対して非有界変動なので数理的にその妥当性を示すのはかなり本格的な準備を必要とする。連続時間のブラウン運動上での数理は Kiyoshi Ito により開拓されたので Ito 解析と呼ばれている。伝統的には生物学や物理学で利用されていたが近年では経済学やファイナンスにおいて株価や資産価格の表現に用いられることも少なくない。非負値を取る現象には指数変換をとると $(Y(t) = e^{B(t)})$ 幾何ブラウン運動と呼ばれ、ァイナンス、保険学、計量経済学などで利用されている。

5.4 正規分布と無限分解可能分布

中心極限定理は独立確率変数の和の分布が正規分布にある意味で収束することを主張するものである。期待値 0, 分散 1 の独立な確率変数列 $\{X_j\}$ に対して

$$Y_n = \frac{1}{\sqrt{n}} [X_1 + \cdots + X_n] \xrightarrow{L} Y \tag{5.17}$$

とすると、$Y \sim N(0,1)$ となる。同様にして

$$Y_{2n} = \frac{1}{\sqrt{2n}} [X_1 + \cdots + X_n + X_{n+1} + \cdots + X_{2n}] \xrightarrow{L} \frac{1}{\sqrt{2}} Y^{(1)} + \frac{1}{\sqrt{2}} Y^{(2)} \tag{5.18}$$

ただし $Y^{(j)} \sim N(0,1)$ $(j = 1,2)$ は互いに独立な確率変数列である。ここで両辺の確率分布は等しいはずであるので特性関数は

$$\psi(t; Y_{2n}) = \psi(\frac{t}{\sqrt{2}}; Y_n)\psi(\frac{t}{\sqrt{2}}; Y_n) \tag{5.19}$$

を満たすはずである。すなわち和の確率変数列がある確率分布に収束することは極限分布はこのような特殊な条件を満たす確率分布である必要がある。それでは他の確率分布は極限分布として存在し得ないであろうか？例えば X_j がコーシー分布にしたがっている場合には

$$Y_n = \frac{1}{n} [X_1 + \cdots + X_n] \xrightarrow{L} Y \tag{5.20}$$

とすると Y は再びコーシー分布になるが、コーシー分布の特性関数が $\psi(t) = e^{-|t|}$ より導ける。正規分布やコーシー分布などを含む確率分布のクラスとして無限分解可

能分布や安定分布が知られているが、連続時間の確率過程モデルと関連している。

定義 5.2： 確率変数 X の分布関数が任意の n に対して独立で同一分布にしたがう確率変数 $X_{n,j}$ により

$$(X \text{ の分布}) = (X_{n,1} + \cdots X_{n,n} \text{の分布})$$

となるとき、無限分解可能分布 (infinitely divisible distribution) と呼ぶ。

正規分布, ガンマ分布, コーシー分布など無限分解可能な分布は多いが必ずしも確率分布が明示的に表現できるとは限らない。無限分解可能な分布の中でもさらに特殊なクラス[20] である安定分布という分布のクラスが金融確率過程の応用上などで議論されるようになっている。ここで指数 α $(0 < \alpha \leq 2)$ の安定分布とは特性関数が

$$\psi(t) = \exp\left[itc - d|t|^{\alpha}\left(1 + i\theta\frac{t}{|t|}\tan\frac{\pi}{2}\alpha\right)\right] \ (0 < \alpha < 1, 1 < \alpha < 2) \tag{5.21}$$

あるいは $\alpha = 1$ のとき

$$\psi(t) = \exp\left[itc - d|t|\left(1 + i\theta\frac{t}{|t|}\log|t|\right)\right] \tag{5.22}$$

により表現される。ただし c は実定数, d は実正定数, θ $(|\theta| \leq 1)$ は実定数である。なお、正規分布は $\alpha = 2$、コーシー分布は $\alpha = 1$ の安定分布であるが正規分布を除く安定分布は一般に二次の積率を持たない。無限分解可能な確率分布や安定分布は中心極限定理のある方向での拡張と考えられるが、正規分布に比較すると分散などは利用できないなどの制約を含め実際の統計分析での有用性については議論が分かれる。

5.A 数理的補論

ここでは分布収束に関する応用上で重要な追加事項について言及しておく。中心極限定理の証明、マルチンゲールの中心極限定理、時系列における中心極限定理などである。

定理 5.3 の証明の概略 ： 確率変数 $X_{n,k}$ の特性関数 $\psi_{n,k}(t)$ を領域 $\{|X_{n,k}| < \epsilon\}$ と $\{|X_{n,k}| \geq \epsilon\}$ に分けて評価する。特性関数 $\mathbf{E}[e^{itX_{n,k}}]$ の指数関数を展開 $(e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \cdots)$ を利用すると正確に評価が可能である。ここで任意の実数 x に対し不等式

[20] ある正定数 a_n と定数 b_n が存在して $(a_nX + b_n$の分布$) = (X_1^{(n)} + \cdots + X_n^{(n)}$の分布$)$ となる確率分布を安定分布 (stable distribution) と呼ばれるが、こうした議論については、例えば佐藤健一 (1990)「加法過程」(紀伊国屋) を参照されたい。

$|e^{ix}-(1+ix)| \leq \min\{x^2/2, 2|x|\}$ および $|e^{ix}-(1+ix-x^2/2)| \leq \min\{|x|^3/6, x^2\}$ である。次に適当に θ_i $(i=1,2)$ を選べば

$$\begin{aligned}\psi_{n,k}(t)-1 &= \mathbf{E}\left[1+itX_{n,k}-\frac{1}{2}t^2X_{n,k}^2+\frac{1}{6}\theta_1 tX_{n,k}^3||X_{n,k}|<\epsilon\right] \\ &\quad +\mathbf{E}\left[1+itX_{n,k}+\frac{1}{2}\theta_2 t^2X_{n,k}^2||X_{n,k}|\geq\epsilon\right]-1 \\ &= it\mathbf{E}[X_{n,k}]-\frac{1}{2}\mathbf{E}[X_{n,k}^2]+R_{n,k}\end{aligned}$$

と表現される。ここで残差 $R_{n,k}$ を評価する必要があるが、$\mathbf{E}[X_{n,k}]=0$ であるから条件の下で $n\to\infty$ のとき

$$\sum_{k=1}^{n}[\psi_{n,k}(t)-1] \to -\frac{1}{2\sigma^2}t^2 \tag{5.23}$$

となることを導ける。したがって、さらに独立性を用いると確率変数 $Y_n=\sum_{k=1}^n X_{n,k}$ の特性関数は

$$\psi_n(t)=\prod_{k=1}^{n}\psi_{n,k}(t)$$

となるので、対数変換を利用すると $\psi_{n,k}(t)-1\to 0$ より

$$\log\psi_n(t)=\sum_{k=1}^{n}\log\left[(\psi_{n,k}(t)-1)+1\right]\sim\sum_{k=1}^{n}[\psi_{n,k}(t)-1] \tag{5.24}$$

となることが分かる。 **Q.E.D.**

マルチンゲールの中心極限定理 :
次にマルチンゲール系列への中心極限定理に言及しておこう。任意の n に対して確率変数列 $M_{n,1}, M_{n,2}, \cdots, M_{n,n}$ を σ-集合体の列 $\mathcal{F}_{n,1}\subset\mathcal{F}_{n,2}\subset\cdots\subset\mathcal{F}_{n,n}$ に対するマルチンゲールとしよう．ただし $\mathcal{F}_{n0}=\{\phi,\Omega\}$ である。確率変数列 $X_{n,k}=M_{n,k}-M_{n,k-1}$ $(k=1,\cdots,n)$ をマルチンゲール差分系列、各項の条件付分散 $\sigma_{n,k}^2=\mathbf{E}[X_{n,k}^2|\mathcal{F}_{n,k-1}]$ とする。条件付き分散の条件および Lindeberg 型の条件

$$\text{(I)}^* \quad \sum_{k=1}^{n}\sigma_{n,k}^2 \xrightarrow{\mathcal{P}} \sigma^2>0 \text{ (一定値)}$$

$$\text{(II)}^* \quad \sum_{k=1}^{n}\mathbf{E}[X_{n,k}^2 I_{\{|X_{n,k}|\geq\epsilon\}}|\mathcal{F}_{n,k-1}]\to 0$$

を仮定する。このときマルチンゲールに関する中心極限定理が得られる。

定理 A.2： マルチンゲール差分系列 $\{X_{n,k}; 1 \leq k \leq n, n \in \mathbf{N}\}$ が条件 (I) と Lindberg 条件 (II) を満たすとする。$n \to +\infty$ のとき

$$\sum_{k=1}^{n} X_{n,k} \xrightarrow{\mathcal{L}} N(0, \sigma^2) \tag{5.25}$$

となる。ただし σ は一定値である.

なお定理 A.2 は定理 5.2 と定理 5.3 を含んでいるのでかなり一般的内容を含んでいる。独立確率変数和やマルチンゲールについての中心極限定理は数理統計学やその他の分野における近年の応用において様々な形で利用されている。

時系列の中心極限定理 ：
ここでは既に述べたごく簡単な 1 次自己回帰モデルを例示として用いて説明する。分散公式もより明示的に表現することが可能であるがここでは研究課題としておこう。

定理 A.3;： 確率変数列 $X_1, X_2, \cdots$ が例 4.5($X_j = aX_{j-1} + Z_j, |a| < 1$,$\{Z_j\}$ は互いに独立に同一分布 (i.i.d.) $N(0, \sigma^2)$ にしたがう確率変数列とする。$n \to +\infty$ のとき

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k \xrightarrow{\mathcal{L}d} N(0, \omega^2) \tag{5.26}$$

となる。ただし ω^2 は有限値であり

$$\omega^2 = \lim_{n \to \infty} \mathbf{V}[\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k] \tag{5.27}$$

である [21]。

[21] 例えば T.W.Anderson (1971), "The Statistical Analysis of Time Series," Wiley7.7 節に時系列における中心極限定理の説明がある。

Part-II：数理統計の基礎

6　統計量と標本分布

6.1 母集団・標本・統計的推測

統計学では分析対象を母集団、母集団からランダムに抽出された標本を確率変数、統計データを確率変数の実現値と見なすことにより確率論を利用している。こうした設定は例えば政府統計として総務省統計局などをはじめとして各統計部局が実施している標本調査、新聞やテレビ局などが行っているなどを例にすると分かりやすい。マスメディアによる世論調査の多くでは RDD(random digit dialing) と呼ばれる統計的方法を利用することが多い。これは登録された電話番号を母集団リストとして、乱数 (random number) をリストの記載番号に次々に割り付け, 標本を抽出する (サンプリング) という標本抽出法 (sampling method) が利用されている [22]。こうした世論調査では分析対象の母集団 (population) から無作為抽出 (ランダム・サンプリング) によって得られたデータをランダム（無作為）標本の実現値と見なせる根拠がある。

世論調査で典型的に質問されるのは「ある事柄を支持するか否か」であるので k れらの事象を 1 と 0 で表そう。調査の目的は母集団 $x_j = 0, 1 \ (j = 1, \cdots, N)$ における比率 $p = (1/N)\sum_{j=1}^{N} x_j$ (母平均) をデータより推定する問題と解釈できる。これは標本調査におけるランダムな操作により得られる大きさ n のランダム標本 (確率変数)$X_i \ (i = 1, \cdots, n)$ の平均 (標本平均) を $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ を $\hat{p}$ とすると、標本平均より母平均を推定することと解釈できる（ベルヌイ試行より $\mathbf{E}[X_i] = p$, $\mathbf{V}(X_i) = \mathbf{E}[(X - \mathbf{E}(X_i))^2] = p(1-p)$ となる)。この標本調査ではある標本が調査で一度選ばれると再び調査されないとしよう。すなわちこの標本調査が非復元抽出とする。標本から計算される標本平均を $\hat{p}_n$ で表すと次に説明する標本調査の議論より $\mathbf{V}[\hat{p}_n] = ((N-n)/(N-1))[p(1-p)/n]$ となり、$(N-n)/(N-1)$ は N が大きければほぼ 1 となる。また調査で得られる標本の大きさ n がある程度に大きければ正規分布で近似すると正規分布表より

$$\mathbf{P}\left(\omega||\hat{p}_n - p| \leq 1.96\sqrt{\frac{p(1-p)}{n}}\right) \sim .95 \tag{6.1}$$

[22]例えば日本経済新聞社の調査 (日経世論調査) では $U_n^* = 1000U_n$(U_n は $(0,1)$ 上で一様分布にしたがう一様乱数) を発生させ 0-9999 の数値を割り付けている。民間の世論調査については実施している各社の HP に説明がある。

となる。ここで標本数 n(例えば 1,000) とすると確率評価の右辺に表れる p は未知なのでデータより計算可能な $\hat{p}$ を利用すると右辺は $1.96\sqrt{\hat{p}(1-\hat{p})/n}$ とすることが合理化できるので、ほぼ信頼度 95% 程度の支持率評価が得られる。なおこの推定誤差の評価はあくまで一種の理想的状況下での確率計算であり、近年の標本調査論では回答拒否と質問内容の関係、電話による仕組み、電話による得られる有権者層と真の有権者の相違、等の非標本誤差という計量しにくい要素も考慮する必要があることが議論される。

社会や経済において重要かつ基本的な統計データとして政府統計・経済統計を挙げられる。例えば総務省統計局が毎月実施している家計調査など、各部局が調査を実施している政府統計の多くでは全体の母集団数 N に比べて標本調査数 n はかなり小さいことが一般的である。家計調査では調査世帯数は約 $9,000$ であるが、本来知りたい母集団 $4,000$ 万世帯 の情報として意味があるのだろうか？ここで無作為抽出法 (random sampling) に基づく標本調査を実施することで迅速性や調査コストの節約とともに数値の客観性と精度を保証してくれる。

有限母集団と無限母集団

数理統計学では標準的には無限母集団を想定することが多いが、まずは世論調査や政府統計における標本調査などで広く活用されている有限母集団の議論に言及しておこう。標本調査の母集団 $x_j = 1, 0$ $(j = 1, \cdots, N)$ における母平均 $\mu = (1/N)\sum_{i=1}^{N} x_i$ とする。この母集団からランダム (互いに独立に) に抽出した大きさ n $(< N)$ の標本 $X_i(\omega)$ から標本平均 $\bar{X}_n$ $(= (1/n)\sum_{i=1}^{n} X_i)$ の期待値を考える。N 個の母集団から n 個の標本を選ぶ組み合わせは $\binom{N}{n}$ $(= {}_NC_n)$ 通りであり、それぞれの組が選ばれる確率は $1/\binom{N}{n}$ $(= 1/{}_NC_n)$ なので「標本平均の期待値 (平均値) は母平均」となる結果

$$\mathbf{E}[\bar{X}_n] = \mu \tag{6.2}$$

が得られる。また標本 X_1 と標本 X_2 の共分散 $\mathbf{E}[(X_1 - \mu)(X_2 - \mu)] = \mathbf{E}[X_1 X_2] - \mu^2$ を求めると

$$\begin{aligned} \frac{1}{N}\frac{1}{N-1}\sum_{i\neq j} x_i x_j - \mu^2 &= \frac{1}{N}\frac{1}{N-1}[\sum_{i,j=1}^{N} x_i x_j - \sum_{i=1}^{N} x_i^2] - \mu^2 \\ &= -\frac{1}{N-1}[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2] \end{aligned}$$

と変形できる。ここで母分散を $\sigma^2 = (1/N)\sum_{i=1}^{N}(x_i - \mu)^2$ とおくと、標本平均の分散 $V(\bar{X}_n) = \mathbf{E}[(\bar{X}_n - \mu)^2]$ は

$$V(\bar{X}_n) = (\frac{1}{n})^2 \mathbf{E}[\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i\neq j}(X_i - \mu)(X_j - \mu)]$$

となる。右辺の各項の期待値を評価すると、第一項は $(1/n)^2 \times (n\sigma^2)$ であり、第二項は $(1/n)^2 \times n(n-1)(-1)(1/(N-1))\sigma^2$ であるので、整理すると

$$V(\bar{X}_n) = \frac{\sigma^2}{n}[1 - \frac{n-1}{N-1}] \tag{6.3}$$

となる。ここで第二項は $N \to \infty$ のとき 0 に収束するので有限母集団による効果、母集団補正項と呼ばれている。

社会や経済における標本調査では対象とする母集団の大きさが有限である。こうした有限標本という側面を明示的に考慮する統計的分析は標本抽出論、経済統計学、などの研究分野である。これに対して母集団の大きさが大きければ無限母集団と見なして分析を行う統計的方法も有力であり、数理統計学ではむしろ一般的に利用している。

母集団の大きさを無限とすると有限母集団からの標本の分析から生じる組み合わせ論的な煩雑さが無くなり、数理統計の分析が明解となることが多い。また分析対象である母集団の大きさが確定できない場合もあるので $N = \infty$ として分析する意味は小さくない。ここで無限母集団から標本を 1 つ選ぶと残された母集団の大きさはやはり無限であるから次に独立に標本を選ぶことが自然となる。こうした幾つかの理由から数理統計学の多くの議論では母集団の大きさは無限と見なし、母集団を確率分布で表現することが一般的に行われている。次に述べる母集団としてある正規分布を想定し、そこからランダムな標本を抽出することが典型例であるが、実際に得られるデータの統計分析では有限母集団と無限母集団の差 (現実社会と統計理論で想定する理想化された状況の差異) を意識することも大切である。

また統計分析ではパラメトリック接近、ノンパラメトリック接近、ベイズ接近、などの分析方法を理解しておくことも重要である。無限母集団を確率分布すると問題が明瞭になるが、さらに確率分布を小数の母数 (パラメター) により表現し、その母数を標本より推測する、という定式化はパラメトリック接近と呼ばれる。例えば 1 次元正規分布は期待値と分散で特長づけられるのでこの母数を未知として統計的推測を行うことが考えられる。他方、こうしたパラメトリック接近と現実に観察されるデータとの差を意識すると特定の確率分布を想定しない統計的方法も発展してきている。またパラメトリック統計では母数は未知ではあるが真値が存在する、と考える。これに対して母数についての確率分布を事前分布として定式化し、事前分布、標本、より母数の事後分布を分析するアプローチも有力であるので、このベイズ統計の基礎について 9 章で議論する。

6.2 標本空間・統計量

統計的推論を行うに際して実験や観察から得られるであろう情報の集合の全体を標本空間 (sample space) と呼ぶ。標本空間の上に $\sigma-$ 集合族および確率測度 P を導

入する。さらに母数 θ を導入し母数が動きうる空間を母数空間 $\theta \in \mathbf{\Theta}$ として、母数に依存する標本空間 $\mathbf{\Omega}$ 上の確率測度を $P(\ \cdot\ |\theta)$ とする。ここで母数 $\theta \neq \theta'$ なら二つの確率分布 $F(\cdot|\theta)$ と $F(\cdot|\theta')$ が異なることを仮定するが、この仮定を識別条件 (identification condition) と呼んでいる。

標本空間 $\mathbf{\Omega}$ が与えられるとき実験や観察により得られるデータの組 $\mathbf{x} = (x_1, \cdots, x_n)'$ を確率変数の組 $\mathbf{X} = (X_1, X_2, \cdots, X_n)'$ の実現値と見なす。本書では $\mathbf{x}$ を縦ベクトル、$(\ \cdots\ )'$ は縦ベクトルを横ベクトルに変換、転置操作を意味する。また説明の便宜上 $x_j\ (j = 1, \cdots, n)$ はスカラーを扱うがベクトルや行列でも良い。このとき $(X_1, X_2, \cdots, X_n)$ の関数として定まる確率変数を統計量 (statistic)、n を標本数 (sample size) と呼び例えば $T(X_1, \cdots, X_n)$ と記す。

例 6.1：観察で得られる n 次元データ $\mathbf{x} = (x_1, \cdots, x_n)'$ が確率変数 $\mathbf{X} = (X_1, \cdots, X_n)'$ の実現値で各確率変数は互いに独立に $N(\mu, \sigma^2)$ にしたがうとする。このとき母数 $\boldsymbol{\theta} = (\mu, \sigma^2)'$, 母数空間 $\mathbf{\Theta} = \{-\infty < \mu < \infty, 0 < \sigma < \infty\}$, 統計量として $T_1 = (1/n)\sum_{i=1}^n X_i\ (= \bar{X}_n)$, $T_2 = (1/(n-1))\sum_{i=1}^n (X_i - \bar{X}_n)^2$(しばしば s_n^2 と書く）などが挙げられる、それぞれ標本平均、標本 (不偏) 分散である。

例 6.2(t 検定の例)：よく数理統計学はゴセット (W. Gosset) がスチューデントの t 分布を正確に導いた頃から発展したと説明される。ビール会社ギネスの技師であった W.Gosset が Student というペンネームで書いた論文で t 分布が導入されたことに由来する逸話である。統計的品質管理の問題ではある母集団から独立に抽出された標本 $X_i\ (i = 1, \cdots, n)$ より期待値で表現される品質が変化していないか否かを評価する必要がよくある。母集団の確率分布が $N(\mu_0, \sigma^2)$ とすると標本平均 $\bar{X}_n$ は $N(\mu_0, \sigma^2)$ にしたがうので、σ^2 の推定量 $S^2 = (1/(n-1))\sum_{i=1}^n (X_i - \bar{X}_n)^2$ を用いて品質 μ_0 からの乖離 (ばらつき) を制御することが考えられる。[23] ゴセットは経験的に得られた数値と正規分布表の数値が整合的でないことに気がつき、母数 μ_0 を既知とした統計量

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2/(n-1)}} \tag{6.4}$$

の分布を正確に導いた (自由度 $n-1\ (= m)$ の t 分布) のである。ここで自由度 m の t 分布をその密度関数

$$f_m(t) = c_m \left[1 + \frac{t^2}{m}\right]^{-(m+1)/2} \tag{6.5}$$

[23] 品質管理 (Quality Control) では、正常な状態を $\mu_0 = 0$ とすると例えば標準正規分布表より 2(シグマ) や 3(シグマ) を管理限界としてランダムな抜き取り検査により事前に設定した管理限界を超えれば品質に異常が起きたと判断している。また近年の金融リスク管理では分位点に基づくリスク管理を行っている。

により定義しよう (c_m は $\int_{-\infty}^{\infty} f_n(t)dt = 1$ となる m に依存する定数)。次に示すように t 分布は互いに独立な確率変数 $N(0,1)$ と $\chi^2(m)$ 分布により

$$T = \frac{N(0,1)}{\sqrt{\chi^2(m)/m}} \tag{6.6}$$

と表現されるが、ここで $\chi^2(m)$ は自由度 m のカイ二乗 χ^2 分布を意味する。χ^2- 分布は $Gamma(m/2, 2)$ 分布であり、期待値は m、分散は $2m$ となる。ここで特に $m \to \infty$ のときには

$$f_m(t) \sim c\, e^{-\frac{t^2}{2}} \tag{6.7}$$

となるので正規分布に収束する。有限の自由度 m に対しては正規分布より少し裾が厚い確率分布であるので、リスク管理において正規分布を用いていたとすると評価が甘い基準となることが重要な意味を持つ。

なおゴセットが扱ったのは二組の 10 個からなるデータ・セットの平均値の比較、2 標本問題である。ここでは独立な確率変数の組 $X_i \sim N(\mu_0^{(1)}, \sigma^2)$, $Y_i \sim N(\mu_0^{(2)}, \sigma^2)$ $(i = 1, \cdots, n)$ とすると興味がある仮説は $\mathcal{H}_0:\ \mu = \mu_0^{(1)} - \mu_0^{(2)} = 0$ が妥当か否か、である。二つの確率変数の差 $Z_i = X_i - Y_i$ $(i = 1, \cdots, n)$ とするとゴセットの分析ではデータ数 $n = 10$ であったのでデータは非常に少ない標本であった。この場合には統計量 T がしたがう分布が正規分布ではなく t 分布を利用すると正確となることには大きな意味がある。小標本データを扱う場合には標本分布の正確な数理的議論、小標本理論 (small sample theory) の必要性が認識されたのである。

他方、社会や経済において遭遇するデータの中にはデータ数が非常に大きい場合が少なくない。T 統計量の場合には $\sqrt{n}(\bar{X}_n - \mu_0) \xrightarrow{d} N(0, \sigma^2)$, $\sum_{i=1}^n (X_i - \bar{X}_n)^2/(n-1) \xrightarrow{p} \sigma^2$ より $n \to \infty$ のとき $T \xrightarrow{d} N(0,1)$ となる。すなわち大数の法則や中心極限定理など大標本理論 (large sample theory)、漸近理論 (asymptotic theory) の適用が有用な場合も多い。

例 6.3: 観察で得られる n 個の p 次元データ $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ が p 次元確率変数 $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ の実現値、各確率変数は互いに独立に $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ にしたがうとする。このとき統計量として例えば

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\ , \tag{6.8}$$

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)' \tag{6.9}$$

が 1 次元データから類推できる。標本平均 $\bar{\mathbf{X}}_n$ は p 次元確率変数、標本共分散行列 $\mathbf{S} = (s_{ij})$ は $p \times p$ 行列の対称行列でありその第 (j,k) 要素は $s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)(X_{ki} - \bar{X}_k)$ である。t 統計量の拡張としてはホテリング (Hoteling) の T^2 統計量

$$T^2 = n(\bar{\mathbf{X}}_n - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}_0) \tag{6.10}$$

が知られている。ここで $p \times 1$ の母数ベクトル $\boldsymbol{\mu}_0 = (\mu_{0i})$ により複数の母平均を分析する際には基本的な役割を演じる。

6.3 t 統計量と F 統計量

正規分布 $N(\mu_0, \sigma^2)(\mu_0$ は既知とする) より独立に標本が得られるとき統計量の標本分布を考察しよう。互いに独立な確率変数列 X_j $(j = 1, \cdots, n)$ とすると標本平均 $\bar{X}_n$ は正規分布 $N(\mu_0, \sigma^2/n)$ にしたがう。ここで一般性を失うことなく問題を簡単化するために標本 X_j $(j = 1, \cdots, n)$ を変換し、$Z_j = (X_j - \mu_0)/\sigma$, $\bar{Z}_n = (1/n)\sum_{i=1}^{n} Z_j$ と置くと $Z_j \sim N(0, 1)$ となる。このとき T(t 統計量) は

$$T' = \frac{\sqrt{n}\bar{Z}_n}{\sqrt{\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2/(n-1)}} \tag{6.11}$$

の分布と同一である。標本分布の典型としてこの導出をやや詳しく見てみよう。
(i) 分子の $\sqrt{n}\bar{Z}_n \sim N(0, 1)$ である。
(ii) 分母と分子は統計的に独立となる。これは

$$Cov(\bar{Z}_n, Z_j - \bar{Z}_n) = 0 \ \ (j = 1, \cdots, n)$$

であり、$Z_j - \bar{Z}_n$ が正規分布にしたがうことより確認できる。
(iii) さらに $n \times 1$(タテ) ベクトル $\mathbf{Z} = (Z_1, \cdots, Z_n)'$, $\mathbf{1}_n = (1, \cdots, 1)'$, と置くと

$$\bar{Z}_n = \frac{\mathbf{1}_n'}{\mathbf{1}_n'\mathbf{1}_n}\mathbf{Z} \tag{6.12}$$

と書けることに注目する。分母の各要素は

$$\mathbf{Z} - (\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{Z} = \left[\mathbf{I}_n - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\right]\mathbf{Z} \tag{6.13}$$

である。$\mathbf{1}_n'\mathbf{1}_n = \sum_{j=1}^{n} 1^2 = n$ であるが、二乗和 $\sum_{i=1}^{n}(x_i - c)^2$ を c について最小化すると $c^* = \bar{x}_n$ となり、残差 $x_i - c^*$ が c^* に直交している。これは n 次元ベクトル $\mathbf{x}$ の $c\mathbf{1}$ への射影 (projection) と解釈され、$\mathbf{P}_n$ は線形代数では射影子と呼ばれる (付論 6-B を参照)。ここで $n \times n$ 行列

$$\mathbf{P}_n = \mathbf{I}_n - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)_n^{-1}\mathbf{1}_n' \tag{6.14}$$

と置くと、対称行列 $\mathbf{P}_n$ は $\mathbf{P}_n^2 = \mathbf{P}_n$, すなわちベキ等行列である。この行列の固有値は 1,0 であり、固有値 1 となる数は $\mathbf{tr}(\mathbf{P}_n) = n-1$、行列 $\mathbf{J}_n = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'$ もベキ等行列で階数は 1, $\mathbf{P}_n + \mathbf{J}_n = \mathbf{I}_n, \mathbf{P}_n\mathbf{J}_n = \mathbf{O}$ である。($r \times r$ 対称行列 $\mathbf{A} = (a_{ij})$ に対してトレースは $\mathrm{tr}(\mathbf{A}) = \sum_{k=1}^{r} a_{ii}$ で定義される。(ここで必要となる固有値と固有ベクトルについては付論 6-C を参照。) (行列 $\mathbf{J}_n = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'$ もベキ等行列で階数は 1, $\mathbf{P}_n + \mathbf{J}_n = \mathbf{I}_n, \mathbf{P}_n\mathbf{J}_n = \mathbf{O}$ である。) したがって、$n \times n$ の直交行列 $\mathbf{Q}$ がとれて $\mathbf{Z}^* = \mathbf{QZ} \sim N_n(\mathbf{0}, \mathbf{I}_n)$,

$$\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2 = \sum_{i=1}^{n-1} Z_i^{*2}\ ,\ Z_i^* \sim N(0,1)\ \ (i = 1, \cdots, n-1) \tag{6.15}$$

となることが分かる。
(iv) $Y_i = Z_i^{*2}$ とする。(ここで確率変数 Z と確率変数 Y の対応は 1 対 1 でないことに注意する。正規分布にしたがう Z は正値と非負値を取りうるが Y は非負値のみ取りうるので注意が必要である。) 変換 $v = u^2$ とすると $dv = 2udu = 2v^{1/2}du$ となることを考慮すると $P(Y \le y) = P(Z^2 \le y)$ は

$$\begin{aligned}\int_{-\sqrt{y}}^{\sqrt{y}} \frac{e^{-u^2/2}}{\sqrt{2\pi}}du &= 2\int_0^{\sqrt{y}} \frac{e^{-u^2/2}}{\sqrt{2\pi}}du \\ &= \frac{2}{\sqrt{2\pi}}\int_0^{y} e^{-v/2}\frac{1}{2\sqrt{v}}dv \\ &= \frac{1}{\sqrt{2\pi}}\int_0^{y} v^{1/2-1}e^{-v/2}dv\end{aligned}$$

となる。すなわち互いに独立な確率変数 Y_i の確率分布は $Y_i \sim Gamma(1/2, 2)$ となる。ここで Gamma 分布の特性関数の表現 $\psi(t) = [1 - it\beta]^{-\alpha}$ より再生性を持つので $\sum_{i=1}^{n-1} Y_i \sim Gamma((n-1)/2, 2)$ となる。
(v) 最後に変数変換 (付論 6-A-1 を参照) を利用する。変換 $U \sim N(0,1), V = \sum_{i=1}^{m} Y_i \sim \chi^2(m)(m = n-1)$ とすると変換 $(U, V) \to (T, V)$(ここで $T = U/\sqrt{V/m}$) のヤコビアンは $u = T\sqrt{v/m}, v = v$ より $\partial u/\partial T = \sqrt{v/m}$, $\partial u/\partial v = T/[2\sqrt{mv}]$, $\partial v/\partial v = 1$ となることが分かるので

$$|J|_+ = \left| \begin{array}{cc} [v/m]^{1/2} & t/[2\sqrt{mv}] \\ 0 & 1 \end{array} \right|_+ = v^{1/2}m^{-1/2} \tag{6.16}$$

である。(u,v) の同時密度関数は $g(u,v) = a_m e^{-u^2/2}v^{m/2-1}e^{-v/2}$,$a_m$ は定数であるから、ヤコビアンを乗じた (t, v) の同時密度関数は

$$f(t,v) = g(t\sqrt{\frac{v}{m}}, v)\sqrt{\frac{v}{m}} \tag{6.17}$$

で与えられる。したがって同時密度関数から T の周辺密度関数を求めるには (6.17) を変数 v で積分すればよいので

$$f(t) = \int_0^\infty f(t,v)dv = c_m \left[1 + \frac{t^2}{m}\right]^{-\frac{m+1}{2}} \tag{6.18}$$

となる。定数を評価すると $c_m = \Gamma((m+1)/2)/[\sqrt{\pi m}\Gamma(m/2)]$ となる。

定理 6.1： $U \sim N(0,1), V \sim \chi^2(m)$ が互いに独立のとき $T = U/\sqrt{V/m}$ の密度関数は $f(t)$ で与えられる (自由度 m の t 分布)。$0 < r < m$ に対し $\mathbf{E}[T^r]$ は存在し、特に $r = 1, 2$ のとき $\mathbf{E}[T] = 0$, $\mathbf{E}[T^2] = m/(m-2)$ となる。

この定理より統計量 T の分布は自由度 n-1($= m$) の t 分布にしたがうことが導けた。次にホテリング T^2 統計量については導出なしに結果のみを述べておく [24]。特に $p = 1$ のとき $T^2 \sim F(1, n-1)$ とかけるが $F(n_1, n_2)$ は自由度 n_1, n_2 の F 分布を意味し、その密度関数は定理 6.3 で与えられる。ここでは次の F 分布との関係を述べておく。

定理 6.2： $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ のとき $\mathbf{Y} = \sqrt{n}(\bar{X}_n - \boldsymbol{\mu}_0)$, $T^2 = \mathbf{Y}'\mathbf{S}^{-1}\mathbf{Y}$ とする。$[T^2/(n-1)][(n-p)/p] \sim F(p, n-p)$ となる。

F 分布と実験計画・分散分析入門

数理統計学の一つの源流はロザムステッド農事試験場である。フィッシャー (R.A. Fisher) は農事試験場で得られる限られた実験データを科学的に解釈するために様々な工夫を行ったが、現在では実験データの解析は医学・薬学ではもちろん経済学における政策評価 (policy evaluation) 問題においてすら重要である。ここでは典型的な事例として薬物の効果についての実験データの解析をイメージして考察すると現実性があるだろう。薬物の効果を測定する為に実験の対象を二つに分け、薬物を投与して数値を計測した群を**処理群** (treatment group)、そうでない (薬物を投与しないで数値を計測した) 群を**対照群** (controlled group) と呼び、二つの群で得られるデータを比較して有意性を調べる 2 標本問題の分析が基本である。こうした比較する二群データ、より一般化した二群以上のデータ分析の為に分散分析 (analysis of variance) と呼ばれる統計手法が開発された。

ここで 1 元配置の統計的モデル

$$X_{ij} = \mu + \mu_i + \epsilon_{ij} \ (i = 1, \cdots, m; j = 1, \cdots, n_i) \tag{6.19}$$

と表現してみよう。標本数 $n = \sum_{i=1}^m n_i$, $X_{ij} \sim N(\mu + \mu_i, \sigma^2)$ とすると仮説 $\mathcal{H}_0 : \mu_i = 0$ の元で i 群変動和

$$S_A = \sum_{i=1}^m n_i(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \ , \tag{6.20}$$

[24] 例えば Anderson (2003) がその系 5.2.1 などを参照。

誤差変動和

$$S_E = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2 \ , \tag{6.21}$$

とする。データの全変動和は

$$S = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2$$

であるが、

$$S = S_A + S_E \tag{6.22}$$

が成立する。ただし全平均 $\bar{\bar{X}} = (1/n)\sum_{i=1}^{m}\sum_{j=1}^{n_j} X_{ij}$, 群平均 $\bar{X}_{i\cdot} = (1/n_i)\sum_{j=1}^{n_i} X_{ij}$ である。このときF比は

$$F = [S_A/(m-1)]/[S_E/(n-m)] \tag{6.23}$$

により定義できる。ここで

$$Cov[X_{ij} - \bar{X}_{i\cdot}, \bar{X}_{i\cdot} - \bar{\bar{X}}] = 0$$

となるのでF比の分母と分子は独立となる。(ここでの議論は n 次元空間において二つのベクトルが直交しているということからも導けるが補論-B を参照されたい。一般の n 次元で考えにくければ n=2,3 で考えればよい。) ここで $n \times 1$ ベクトル

$$\mathbf{x} = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n_1} \\ x_{21} \\ \vdots \\ x_{mn_m} \end{bmatrix} \tag{6.24}$$

及び $n \times m$ 行列

$$\mathbf{K} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \\ \mathbf{0} & \vdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{1}_{n_m} \end{bmatrix}$$

さらに $\mathbf{J_n} = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'$ (= (任意の (i,j) 要素が n^{-1})) と定めると、全変動 $\mathbf{x}'[\mathbf{I}_n - \mathbf{J_n}]\mathbf{x}$ は

$$\mathbf{x}'[\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K} - \mathbf{J_n}]\mathbf{x} + \mathbf{x}'[\mathbf{I}_n - \mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}]\mathbf{x} \tag{6.25}$$

と分解される。特に $\mathbf{P}_n = \mathbf{I}_n - \mathbf{J_n}$, $\mathbf{P}_{n1} = \mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K} - \mathbf{J_n}$, $\mathbf{P}_{n2} = \mathbf{I}_n - \mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}$ とすると、前の議論と同様に $\mathbf{P}_n^2 = \mathbf{P}_n$, $\mathbf{P}_{n1}^2 = \mathbf{P}_{n1}$, $\mathbf{P}_{n2}^2 = \mathbf{P}_{n2}$, $\mathbf{P}_{n1}\mathbf{P}_{n2} = \mathbf{O}$, $\mathrm{tr}(\mathbf{P}_n) = n-1$, $\mathrm{tr}(\mathbf{P}_{n1}) = m-1$, $\mathrm{tr}(\mathbf{P}_{n2}) = n-m$ となる。したがって仮説 $\mathcal{H}_0$: $\mu_i = 0$ $(i = 1, \cdots, m)$ の下では $S_A/\sigma^2 \sim \chi^2(m-1)$, $S_E/\sigma^2 \sim \chi^2(n-m)$ で互いに独立となる。

一般の F 比について有用な結果を次にまとめておく。その導出では t 比の分布の導出と同様に (U, V) の同時分布 $(\chi^2(l), \chi^2(m)$ 密度の積) を用いるが、標準的には付論 6-A（付論 2-A も参照）で説明しているように変数変換 (F, V) を利用して求めればよい。

定理 6.3 : $U \sim \chi^2(l), V \sim \chi^2(m)$ が互いに独立であるとき非負の確率変数 $F = [U/l]/[V/m]$ がしたがう密度関数は

$$f(x) = c(l,m)x^{\frac{l}{2}-1}[1 + x\frac{l}{m}]^{-((l+m)/2)} \tag{6.26}$$

で与えられる (自由度 (l, m) の F 分布)。ここで定数は $B(\cdot,\cdot)$ をベータ関数として

$$c(l,m) = [(l/m)]^{l/2}[1/B(l/2, m/2)] \tag{6.27}$$

である。

ここで特に分母の自由度 m が大きいときには $V/m \xrightarrow{p} 1$ となることに注意する。しばしば応用では $m \to \infty$ のとき $lF \sim \chi^2(l)$ と近似できることが利用されている。

1 元配置の統計モデルは二元配置の統計モデル

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (i = 1, \cdots, m; j = 1, \cdots, l) \tag{6.28}$$

などに拡張されるが、因子数 (ここでは α と β なので 2) が増加すると表現が複雑になるほか、2 因子（α_i, β_j）が直接にもたらす効果 (主効果) の他に相互に関連して生じる効果 (交互作用) などをモデル化する必要が生じる。こうした議論を系統的に議論する研究分野は実験計画 (design of experiments) と呼ばれている。

6.4 経験分布・順序統計量・極値分布

確率分布 $F(x)$ にしたがう確率変数列 $X_1, X_2, \cdots, X_n$ が互いに独立なとき実現値 (縦) ベクトル $\mathbf{x} = (x_1, x_2, \cdots, x_n)'$ が得られる状況を想定する。任意の実数 x に対して統計量

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) \tag{6.29}$$

を経験分布関数 (empirical distribution function) と呼ぶ。ここで指標関数は $I(\omega) = 1$ (ω が成立), $I(\omega) = 0$ (ω が成立しない) により定義する。確率分布 $F(x)$ が分からないとき経験分布関数を用いて推測することが考えられる。1, 0 をとるベルヌイ試行からなる標本平均, 確率 $p = F(x)$ と見ると大数の法則と中心極限定理を用いれば n が大きいとき (任意の x に対して)

$$\sqrt{n}[F_n(x) - F(x)] \xrightarrow{d} N[0, F(x)(1 - F(x))] \tag{6.30}$$

となる。

確率変数列 $X_1, X_2, \cdots, X_n$ が互いに独立な確率変数とするとき大きさの順番に並べ替えた確率変数列 $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ を順序統計量 (ordered statistics) と云う。ここで確率分布 $F(x)$ にしたがう互いに独立な確率変数列の実現値 (縦) ベクトル $\mathbf{x} = (x_1, x_2, \cdots, x_n)'$ の順番を並び替えて $\mathbf{x}^*_{(n)} = (x_{(1)}, x_{(2)}, \cdots, x_{(n)})'$ が得られたとき、しばしば統計分析では想定する確率分布 $F(x)$ (例えば正規分布) が妥当か否かを検討する必要が生じる。ここで $F(x)$ がこのデータに妥当なデータであれば Q-Q プロットが利用できる。Q-Q プロットとは 2 次元グラフ

$$(F^{-1}(\frac{i - 0.5}{n}), x_{(i)}) \;\; i = 1, \cdots, n$$

で定義する。(なおここでは F^{-1} は逆関数、分位点を意味するが n 個のデータの i 番目を $(i - .5)/n$ と補正している。) 特に正規分布の逆関数 $F^{-1}(x) = \Phi^{-1}(x)$ をとる場合を正規 Q-Q プロットと呼ばれるが、正規分布が妥当であればグラフ上ではデータには直線的な関係があるはずなので正規分布の統計的検証として利用できる。

ここで真の確率分布 $F_0(x)$ としてその妥当性をデータから調べることを考えよう。例えば真の確率分布が正規分布であるとすると分布関数 $F_0(x)$ は x に依存しているので x に依存しないようなノンパラメトリック統計量が必要である。確率分布の検定問題ではしばしば Kolmogorov-Smirnov 統計量が利用されるが、この統計量は経験分布 F_n と F_0 の差を基準化して

$$\mathrm{KS}_n = \sup_x \sqrt{n}|F_n(x) - F_0(x)|$$

により与えられる。標本数 n が大きいとき KS_n の確率分布は

$$P(\mathrm{KS}_n \leq y) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2y^2} \;\; (y > 0) \tag{6.31}$$

となる [25]。

[25] 統計量の KS_n の極限は不変原理で説明した $[0, 1]$ 上の標準ブラウン運動 $X(t)$ により $KS = \sup_{0 \leq t \leq 1} |X(t) - tX(1)|$ と表現できることを利用して導ける (例えば Billingsley (1968), *Convergence in Probability measures*,Wiley,103 項) を参照。

次に順序統計量 $X_{(1)} = \min X_i, \cdots, X_{(k)}, \cdots, X_{(n)} = \max X_i$ の分布を考えよう。最大値に分布は

$$P\left(X_{(n)} \leq x\right) = \prod_{i=1}^{n} P\left(X_i \leq x\right) = F(x)^n \tag{6.32}$$

で与えられる。$X_{(k)}$ の分布は事象 $\{X_{(k)} \leq x\}$ の確率は互いに独立な n 個の確率変数 X_i の中で少なくとも k 個が x を超えない確率であるので

$$P\left(X_{(k)} \leq x\right) = \sum_{i=k}^{n} {}_nC_i [F(x)]^i [1-F(x)]^{n-i}$$

となる。分布関数を微分すれば

$$\begin{aligned} f_k(x) &= \sum_{h=k}^{n} {}_nC_h f(x)\{[F(x)]^{h-1}[1-F(x)]^{n-h} - [F(x)]^h[1-F(x)]^{n-h-1}\} \\ &= \sum_{h'=k-1}^{n} nf(x)_{n-1}C_{h'}[F(x)]^{h'}[1-F(x)]^{n'-k-1} \\ &\qquad - \sum_{h=k}^{n} nf(x)_{n-1}C_h[F(x)]^h[1-F(x)]^{n-h-1} \\ &= nf(x)_{n-1}C_{k-1}[F(x)]^{k-1}[1-F(x)]^{n-k} \\ &= \frac{1}{B(k, n-k+1)} f(x)[F(x)]^{k-1}[1-F(x)]^{n-k} \quad . \end{aligned}$$

となり密度関数が得られる。
例えば元の確率分布が一様分布であれば $F(x) = x, f(x) = 1\ (0 \leq x \leq 1)$ よりベータ関数の性質を用いると簡単な表現が得られ、順序統計量の期待値と二次積率は

$$\mathbf{E}(X_{(k)}) = \frac{k}{n+1}\ ,\ \mathbf{E}(X_{(k)}^2) = \frac{k(k+1)}{(n+1)(n+2)} \quad . \tag{6.33}$$

となる。また同様の考察により $X_{(k)}, X_{(h)} (k < h)$ の同時分布なども $X_{(h)}$ を条件とする条件付分布を利用することにより導くことができる。

ここで近年では損害保険、自然災害、極端な現象のリスク解析で重要となりつつある極値分布について言及しておこう。標本の大きさを示す順序統計量として最大値 $X_{(n)}$ がしたがう確率分布の分析が重要であるので、$M_n = X_{(n)}$ に対して実数列 a_n, b_n をとり $M_n^* = (M_n - b_n)/a_n$ と基準化しておく。代表的な例としては元の確率変数が指数分布にしたがう場合 $a_n = 1, b_n = \log n$ とすると $P(M_n^* \leq z) = P(M_n \leq a_n z + b_n)$ は $n \to \infty$ のとき

$$\begin{aligned} F^n(z + \log n) &= [1 - \exp[-z - \log n]]^n \\ &= \left[1 - n^{-1}\exp(-z)\right]^n \\ &\longrightarrow \exp[-e^{-z}] \end{aligned}$$

となる。右辺の確率分布は Gumbel 分布 (Type-I の極値分布) と呼ばれている。次に元の確率変数がコーシー分布にしたがう場合には

$$\begin{aligned} F^n(nz) &= \left[1 - \frac{1}{nz} \times nz(1 - F(nz))\right]^n \\ &\sim \left[1 - \frac{1}{n\pi z}\right]^n \\ &\longrightarrow \exp[-\frac{1}{\pi z}] \ . \end{aligned}$$

となり、右辺は Type-II の極値分布と呼ばれている。最後に元の確率分布が一様分布にしたがう場合は $a_n = 1/n, b_n = 1$ とおけば

$$\begin{aligned} F^n(n^{-1}z + 1) &= \left[1 + \frac{z}{n}\right]^n \\ &\longrightarrow \exp[z] \end{aligned}$$

となり、右辺は Type-III の極値分布と呼ばれる。

ここで述べた 3 種類の確率分布は極値分布と呼ばれているが、こうした例を一般することが統計的極値論 (statistical extreme value theory) の始まりである。標本数 n が大きいとき順序統計量の分布について次のような結果が成立する。

定理 6.4： 互いに独立で同一分布にしたがう確率変数列 $X_1, X_2, \cdots, X_n$ に対して $M_n = \max(X_1, \cdots, X_n)$ とする。このとき適当な実数列 a_n, b_n $(a_n > 0)$ が存在して

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) = F^n(a_n z + b_n) \tag{6.34}$$

が分布関数 $G(z)$ に収束するとする。このとき $G(z)$ は基準化のための定数を除き次の 3 種類に限られる。

(I) Gumbel 分布

$$G(z) = \exp\left(-e^{-z}\right) \quad -\infty < z < \infty \tag{6.35}$$

(II) Frechet 分布

$$G(z) = \exp\left(-z^{-\alpha}\right) \quad 0 < z < \infty \ , \alpha > 0 \ , G(z) = 0 \ (z \leq 0), \tag{6.36}$$

(III) Weibul 分布

$$G(z) = \exp\left\{-(-z)^{\alpha}\right\} \quad z \leq 0 \ , \alpha > 0 \ , G(z) = 1 \ (z \geq 0) \ . \tag{6.37}$$

ここで $F^n(a_n z + b_n)$ が $G(z)$ に収束するときには任意の実数 $t\ (\in (0,1])$ に対して $F^{[nt]}(a_{[nt]} z + b_{[nt]}) \to G(z)$ となることに注目しよう。

$$F^{[nt]}(a_n z + b_n) = [F^n(a_n z + b_n)]^{[nt]/n} \to G^t(z) \tag{6.38}$$

であるから実数列 $a_n/a_{[nt]} \to \alpha(t)$, $[b_n - b_{[nt]}]/a_{[nt]} \to \beta(t)$ をとれば極限分布は方程式

$$G^t(z) = G(\alpha(t)z + \beta(t)) \tag{6.39}$$

を満たす。したがって $G^{st}(z) = G(\alpha(st)z + \beta(st)) = [G^s(z)]^t$ である。これより任意の実数 s, t について

$$\alpha(st) = \alpha(s)\alpha(t)\ , \beta(st) = \alpha(s)\beta(t) + \beta(s) = \alpha(t)\beta(s) + \beta(t)\ , \tag{6.40}$$

となる関係が得られる。最初の式より実数 θ を用いて $\alpha(t) = t^{-\theta}$ となる必要があるので (a) $\theta = 0$, (b) $\theta > 0$, (c) $\theta < 0$ という 3 つの場合分けが必要となる。例えば (a) の場合を例にとると $\beta(st) = \beta(s) + \beta(t)$ となるので、その解はある実数 c を用いて $\beta(t) = -c \log t\ (t > 0)$ と表せば $[G(0)]^t = G(-c \log t)$ と云う表現が得られる。$u = -c \log t\ (c \neq 0)$ とおくと正実数 p を用いて $G(u) = [\exp(-p \exp(-u/c))]$ と云う形が導け、確率分布の性質よりグンベル分布の形が得られる。

ここで説明した極値統計量の漸近分布についての結果は初めて考察した二人の名前より Fisher-Tippet の定理 [26] と呼ばれている。なお定理 6.4 に表れる確率分布族は一般化極値分布 (generalized extreme value distribution) と呼ばれ、確率分布の位置・スケール・形状を表す母数 $\mu, \sigma\ (> 0), \xi$ を用いて

$$G(z) = \exp\left(-[1 + \xi(\frac{z-\mu}{\sigma})]^{-1/\xi}\right) \tag{6.41}$$

と表現することができる。ここでグンベル分布は $\xi\ (= 1/m) \to 0\ (m \to \infty)$ の極限と見なすことができ特に $\mu = 0, \sigma = 1$ としたのが標準 Gumbel 分布である。ここで念のために極限分布の形状を (変数軸は見やすいように調整した) 図 6-1 に示しておく。

[26] 統計的極値理論については本書の 11 章、あるいは国友直人・山本拓 監修 (2012),「21 世紀の統計科学」Vol-II に収録の渋谷政昭・高橋倫也「極値理論・信頼性・リスク管理」:自然・生物・健康の統計科学, 東京大学出版会, ＨＰ版 (http://www.cirje.e.u-tokyo.ac.jp/research/reports/R15ab.html) を参照。

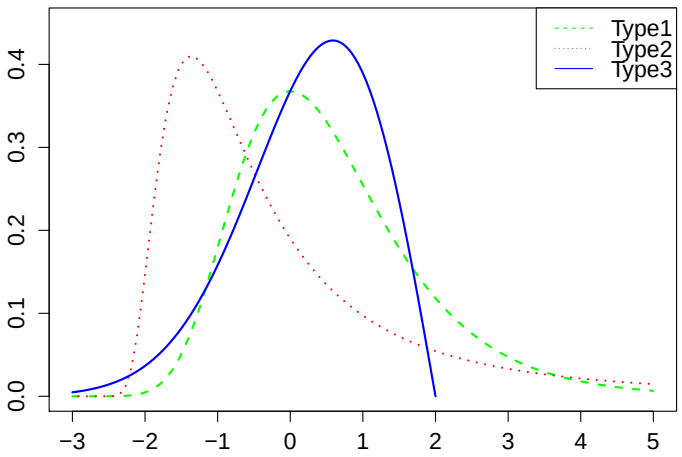

図 6-1：3 つのタイプの極値分布

付論 6-A：変数変換の公式 (再び)

数理統計学では変数変換は統計量の連続型分布を調べるための基本的な分析手段である。既に 2-A において 1 次元の変数変換と応用について言及したが、標本分布論の基礎として重要なので、再び 2 次元の変数変換および統計量のしたがう確率分布の導出について考察しよう。p 次元確率変数 $\mathbf{X} = (X_i)$ の同時分布関数 $F(\mathbf{x})$ を所与とするが、本文では $p = 2$ の場合を利用した。写像 $h(\,\cdot\,)$ により p 次元確率変数ベクトル $\mathbf{Y} = (Y_i) = h(\mathbf{X})$, $\mathbf{Y}$ の確率分布 $G(\mathbf{y})$ としよう。(p 次元の議論が分かりにくければ $p = 2$ の場合を考えればよい。) このとき

$$\begin{aligned} G(\mathbf{y}) &= P(\omega|\mathbf{Y} \leq \mathbf{y}) \\ &= P(\omega|h(\mathbf{X}) \leq \mathbf{y}) \\ &= P(\mathbf{X} \in h^{-1}((-\infty, y_1] \times \cdots \times (\infty, y_p)) \end{aligned}$$

と表現される。(ここで $h^{-1}(\cdot)$ は逆写像である。) 逆写像が単調なら右辺は $F(h^{-1}(\mathbf{y})$ となる。したがって次の関係が成立する。

定理 6-A-1： 条件 (I):$F(\mathbf{x})$ は連続な密度関数 $f(\mathbf{x})$ を持つ、条件 (II):写像 $\mathbf{y} = h(\mathbf{x})$ は一対一写像である、を仮定する。このとき $G(\mathbf{y})$ は連続微分可能で密度関数は

$$g(\mathbf{y}) = f(h^{-1}(\mathbf{y}))|\frac{\partial h^{-1}(\mathbf{y})}{\partial \mathbf{y}}|_+ \quad \mathbf{y} \in h(\mathbf{R}^p) \tag{6.42}$$

となる。ただし変換 $h^{-1}(\cdot)$ のヤコビアンは

$$\mathbf{J} = \left| (\frac{\partial h^i(\mathbf{y})}{\partial y_j})_{ij} \right|_+ \tag{6.43}$$

で与えられる。

例 6.4： t 統計量についての変換を再びまとめると、確率変数 $X_1 = U, X_2 = V$ および確率変数 $Y_1 = h_1(X_1, X_2) = T, Y_2 = h_2(X_1, X_2) = V$ とすると、$X_1 = Y_1\sqrt{Y_2/m}, X_2 = Y_2$ と書き直せる。このとき

$$|\mathbf{J}|_+ = \left| \begin{array}{cc} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{array} \right|_+ = \left| \begin{array}{cc} \sqrt{\frac{y_2}{m}} & (1/2)y_1/[y_2 m]^{1/2} \\ 0 & 1 \end{array} \right|_+ = \sqrt{\frac{y_2}{m}}$$

である。

例 6.5： F 統計量についての変換を取り上げよう。確率変数 $X_1 = U \sim \chi^2(l), X_2 = V \sim \chi^2(m)$ として、確率変数 $Y_1 = [U/l][V/m], Y_2 = V$ とすると、$X_1 = Y_1 Y_2(l/m), X_2 = Y_2$ と書き直せる。したがって変換のヤコビアンは

$$|J|_+ = \left| \begin{array}{cc} \frac{l}{m}Y_2 & \frac{l}{m}Y_1 \\ 0 & 1 \end{array} \right|_+ = \frac{l}{m}Y_2 \,. \tag{6.44}$$

となる。U と V の同時分布 $g_l(u)g_m(v)$ より $f(y_1, y_2) = g_l(y_1 y_2(l/m))g_m(y_2)(l/m)y_2$ として y_2 について区間 $[0, \infty)$ の範囲で積分すると F 分布の密度関数 (6.26) が導ける。

付論 6-B：射影 (projection) と最小二乗法

簡単な例により射影 (projection) の幾何的意味を説明する。ここで説明する議論は $n = 2, 3$ とした 2 次元・3 次元空間で考えればより直観的に理解できるが議論は一般の n 次元ユークリッド空間で成立する。n 次元 (縦) ベクトル $\mathbf{x} = (x_1, \cdots, x_n)'$ を n 次元 (縦) ベクトル $\mathbf{1}_n = (1, \cdots, 1)'$ へ垂線を下ろす操作は距離 $\|\mathbf{x} - c\mathbf{1}_n\|^2 = (\mathbf{x} - c\mathbf{1}_n)'(\mathbf{x} - c\mathbf{1}_n) = \sum_{i=1}^n (x_i - c)^2$ の変数 c についての最小化に対応する。

(図 6-B-1 を挿入)

この最小化問題の解は $c^* = (\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\mathbf{x}$ であるから、残差 (縦) ベクトル

$$\mathbf{e} = \mathbf{x} - c^*\mathbf{1}_n = \left[\mathbf{I}_n - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\right]\mathbf{x} \tag{6.45}$$

とすると、n 次元 (縦) ベクトル $\mathbf{e} = (e_i)$ とベクトル $\mathbf{1}_n$ は直交する、すなわち

$$\mathbf{e}'\mathbf{1}_n = \sum_{i=1}^n e_i = 0 \tag{6.46}$$

となる。したがって $\mathbf{J}_n = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'$, $\mathbf{P}_n = \mathbf{I}_n - \mathbf{J}_n$, とすると

$$\begin{aligned}\mathbf{x} &= \mathbf{J}_n\mathbf{x} + \mathbf{P}_n\mathbf{x} \\ &= \mathbf{x}_1 + \mathbf{x}_2\end{aligned}$$

と分解される。したがって関係 $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n [x_i - \bar{x}_n]^2 + n[\bar{x}]^2$ の両辺を単にベクトル $\mathbf{x} = (x_1, \cdots, x_n)'$ の二次形式で書き直すと、

$$\mathbf{x}'\mathbf{x} = \mathbf{x}'\mathbf{J}_n\mathbf{x} + \mathbf{x}'\mathbf{P}_n\mathbf{x} \tag{6.47}$$

は直角三角形に関するピタゴラスの定理を意味している。ただし $n \times n$ 行列 $\mathbf{J}_n, \mathbf{P}_n$ は条件 (i)$\mathbf{P}_n' = \mathbf{P}_n, \mathbf{J}_n' = \mathbf{J}_n$(対称行列), (ii)$\mathbf{P}_n^2 = \mathbf{P}_n, \mathbf{J}_n^2 = \mathbf{J}_n$(ベキ等行列) を満たしている。これらの条件を満たす行列は一般に射影行列 (projection matrix) と呼ばれているが、$n \times n$ ベキ等行列 $\mathbf{A}$ の固有値を λ とすると、$\mathbf{z} \neq \mathbf{0}$ に対して $\mathbf{Az} = \lambda\mathbf{z} = \mathbf{A}^2\mathbf{z} = \lambda^2\mathbf{z}$ より $(\lambda - \lambda^2)\mathbf{z} = \mathbf{0}$, つまり $\lambda = 0, 1$ となる。固有値の中で 1 となる固有値の数は $\mathbf{A} = (a_{ij})$ とすると $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$ である。

ここで一般に n 次元ベクトル $\mathbf{x}$ を r 個 $(1 \leq r \leq n)$ の線形独立な n 次元ベクトル $\mathbf{z}_j$ $(j = 1, \cdots, r)$ の線形和 $\sum_{j=1}^r c_j\mathbf{z}_j$ へ垂線を下ろす問題を考えよう。(縦) ベクトル $\mathbf{c} = (c_1, \cdots, c_r)'$, $n \times r$ 行列 $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_r)$ として

$$\|\mathbf{x} - \sum_{j=1}^r c_j\mathbf{z}_j\|^2 = \|\mathbf{x} - (\mathbf{z}_1, \cdots, \mathbf{z}_r)\begin{pmatrix} c_1 \\ \vdots \\ c_r \end{pmatrix}\|^2 \tag{6.48}$$

の最小化問題を考える。ベクトル $\mathbf{c}$ について最小化して例えば評価関数 $L(c_1, \cdots, c_r) = \sum_{i=1}[x_i - \sum_{j=1}^r c_j z_{ji}]^2$ を変数 c_j $(j = 1, \cdots, r$ について偏微分すれば正規方程式

$$\mathbf{Z}'\mathbf{Z}\mathbf{c}^* = \mathbf{Z}'\mathbf{x} \tag{6.49}$$

および

$$\mathbf{c}^* = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x} \tag{6.50}$$

が得られる。したがって垂線ベクトルは $\mathbf{Zc}^* = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}$, $[\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{x}$ により表現できる。ここで $\mathbf{J}_n = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, $\mathbf{P}_n = \mathbf{I}_n - \mathbf{J}_n = \mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ とすると $\mathbf{J}_n, \mathbf{P}_n$ は射影行列である。あるいは

$$\mathbf{I}_n = \mathbf{J}_n + \mathbf{P}_n \tag{6.51}$$

と言う単位行列の分解と理解できる。ここで 1 となる固有値は

$$\mathrm{rank}(\mathbf{J}_n) = \mathrm{rank}[\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'] = \mathrm{rank}[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}] = r \tag{6.52}$$

である。(ここで行列 $\mathbf{A}, \mathbf{B}$ の積が定義できれば $\mathrm{rank}(\mathbf{AB}) = \mathrm{rank}(\mathbf{BA})$ である。) 同様に $\mathrm{rank}(\mathbf{P}_n) = n - r$ となる。こうした解を求める問題は最小二乗法 (least squares method) と呼ばれ統計学では広く用いられている。

射影行列を巡る議論は次のようにまとめることができる。証明などは例えば 竹内啓「線形数学」(培風館) を参照されたい。

定理 A-2 : n 次元実ベクトル空間の任意の元 $\mathbf{x} \in \mathbf{R}^n$ とする。
(i) n 次元空間の部分空間 M, 直交補空間 $M^{\perp}$ とすると、n 次元ベクトル $\mathbf{x}$ は $\mathbf{x} = \mathbf{y} + \mathbf{z}$ $(\mathbf{y} \in M, \mathbf{z} \in \mathbf{M}^{\perp})$ と一意に分解される。
(ii) $n \times n$ 行列 $\mathbf{\Pi}$ が $\mathbf{\Pi}^2 = \mathbf{\Pi}, \mathbf{\Pi}' = \mathbf{\Pi}$ ($\mathbf{\Pi}$ は転置行列) ならばベクトル $\mathbf{y} = \mathbf{\Pi x}$, $\mathbf{z} = (\mathbf{I}_n - \mathbf{\Pi})\mathbf{x}$, は射影となる。
(iii) $\mathbf{M}$ への任意の射影は適当な $n \times m$ $(1 \leq m < n)$ 行列 $\mathbf{A}$ により
$\mathbf{\Pi} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ $(|\mathbf{A}'\mathbf{A}| \neq 0)$ と表現できる。

付論 6-C：固有値問題について

確率論や数理統計学では線形代数の基礎的な議論を多用する。数理統計学で特に有用な固有値問題は対称行列の場合なので一般の場合ほど複雑ではないが、応用上で重要な役割を果たすことが多いので基本事項に言及しておく。証明などは述べないが文献を挙げておく。

n 次 $(n \times n)$ 実正方行列 $\mathbf{A}$ の固有値 (characteristic roots, latent roots) λ(スカラー) と固有ベクトル (characteristic vectors, latent vectors) $\mathbf{x}$ $(\mathbf{x} \neq \mathbf{0})$ は方程式

$$\mathbf{Ax} = \lambda \mathbf{x} \tag{6.53}$$

の解として定義される。このとき $(\mathbf{A} - \lambda \mathbf{I}_n)\mathbf{x} = \mathbf{0}$ $(\mathbf{x} \neq \mathbf{0})$ より行列 $\mathbf{A} - \lambda \mathbf{I}_n$ は特異 (singular) なので方程式

$$\lambda(\mathbf{A}) = |\mathbf{A} - \lambda \mathbf{I}_n| = 0 \tag{6.54}$$

を満足する。この固有方程式は λ について n 次多項式なので $\lambda(\mathbf{A}) = (-1)^n \prod_{j=1}^{n} (\lambda - \lambda_j)$ である。

定理 A-3 : n 次実正方行列 $\mathbf{A} = (a_{ij})$ が対称行列 $(a_{ij} = a_{ji}\ \forall i, j)$ とする。
(i) すべての固有値は実数である。
(ii) すべての相異なる固有値 $\lambda_i \neq \lambda_j$ に対応する固有ベクトルを $\mathbf{x}_i$ $(i = 1, \cdots, p; p \leq n)$ とすると、$\mathbf{x}_i'\mathbf{x}_j = 0$ $(i \neq j)$ となる。(すなわち固有ベクトルは直交する。)

(iii) 適当に直交行列 $\mathbf{P}$(ただし直交行列は $\mathbf{PP}^{'} = \mathbf{P}^{'}\mathbf{P} = \mathbf{I}_n$ を満足する) がとれて

$$\mathbf{PAP}^{'} = \begin{bmatrix} \lambda_1 & 0 & \cdots & & \\ 0 & \lambda_2 & 0 & \cdots & \\ \vdots & & & & \vdots \\ 0 & \cdots & & \lambda_n & \end{bmatrix} = \mathbf{\Lambda} \, . \tag{6.55}$$

となる。

固有ベクトルは大きさは任意であるので (i.e. 定数倍してもよい) すべての固有値が互いに異なれば固有値 λ_j $(j = 1, \cdots, n)$ に対応する固有ベクトル $\mathbf{x}_j$ $(j = 1, \cdots, n)$ を用いて ($\|\mathbf{x}_i\| = 1, \mathbf{x}_i^{'}\mathbf{x}_j = 0$ $(i \neq j)$ とする) より $n \times n$ 行列 $\mathbf{P} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ とすると $\mathbf{AP} = \mathbf{P\Lambda}$ が得られる。したがって行列 $\mathbf{A}$ は

$$\mathbf{A} = \sum_{j=1}^{n} \lambda_j \mathbf{x}_j \mathbf{x}_j^{'} \tag{6.56}$$

と表現できる。特に対称行列の場合には固有値に重根があってもこのような表現が得られる。なお、この表現は行列 $\mathbf{A}$ のスペクトル分解と呼ばれるが、この名前が示すように固有値問題とは行列で表現される線形写像の性質の一般的分析とほぼ同一である。

任意のベクトル $\mathbf{y}$ $(\neq \mathbf{0})$ に対して $\mathbf{y}^{'}\mathbf{Ay} > 0$ となる実対称行列 $\mathbf{A}$ を正定符号 (positive definite) 行列と呼ぶ。また厳密な不等号を単なる不等号とすると非負定符号行列である。ある確率変数ベクトル $\mathbf{X} = (X_j)$ の分散共分散行列 $\mathbf{\Sigma} = (\sigma_{ij})$ は任意のベクトル $\mathbf{y} = (y_j)$ に対して

$$\mathbf{y}^{'}\mathbf{\Sigma y} = \sum_{j,k=1}^{n} y_j y_k \sigma_{jk} = \mathbf{V}[\sum_{j=1}^{n} y_j X_j] \geq 0 \tag{6.57}$$

となることより非負定符号行列である。等号は $\sum_{j=1}^{n} y_j X_j = 0$ の時にのみ成立するが、等号が成り立つとき確率変数は (確率 1 で) 退化していると云う。通常の統計分析では確率変数は退化していない (non-degenerate) ことが仮定される。

一般に $p \times p$ 正定符号 (実対称) 行列 $\mathbf{A}$ の固有値 (λ_i $(i = 1, \cdots, p)$ とする) は正である。この固有値を並べた対角行列 $\mathbf{\Lambda} = (\mathrm{diag}(\lambda_i))$ とすると、行列 $\mathbf{A}$ は直交行列 $\mathbf{P}$ により $\mathbf{A} = \mathbf{P\Lambda P}^{'}$ と分解できる。したがって、$\mathbf{\Lambda}^{1/2} = (\mathrm{diag}\sqrt{\lambda_i})$ とすると

$$\mathbf{A} = \mathbf{P\Lambda}^{1/2}\mathbf{P}^{'}\mathbf{P\Lambda}^{1/2}\mathbf{P}^{'} \tag{6.58}$$

となる。すなわち $\mathbf{A}^{1/2} = \mathbf{P\Lambda}^{1/2}\mathbf{P}^{'}$ と置けばよいことになる。

お茶の時間 (Tea-Time)

線型代数の教科書としては斉藤正彦「線型代数入門」(東京大学出版会) が定番であったが、固有値問題の説明は複雑である。竹内啓「線形数学」(培風館)、志賀浩二「固有値問題 30 講」(朝倉) などは個性的であり固有値問題の説明はより理解しやすいと思われる。むろん大学初級の講義のために書かれた教科書の多くには説明がある。

7　統計的推定論

7.1 統計量・推定量・推定値

数理統計学における理論はそれぞれ応用上の問題を解決することからまずその枠組みが考えられ、その後に数理的に一般化、抽象化されたものが多い。したがって多くの場合には抽象的な理論から学び始めるよりも、まずは歴史的展開を踏まえて理論と応用を学ぶ方がその間の事情や理論の意味を理解しやすい。

実験データ、観察データとして得られたn個の実数を小文字を使って(横)ベクトル $(x_1, x_2, \cdots, x_n)$ とする。さらに統計データそのものでなく、母集団から標本抽出により得られるn個の独立標本を確率変数と解釈するときには大文字の記号を使って(横)ベクトル $(X_1, X_2, \cdots, X_n)$ と表現する。これを標本 $(X_1, X_2, \cdots, X_n)$ として、標本の関数を**統計量** (statistic) と呼ぶ。標本の関数である統計量を使って母集団を表現している未知母数を推測するとき、統計量を**推定量** (estimator)、推定量を使って標本として実際に観測値として得られるデータを代入した統計量の値を**推定値** (estimate) と呼び区別する。

統計的推測の理論とは統計量の確率的性質についての議論が多いが、実際に得られる1組のデータは様々な偶然変動の結果としてある数値をとって観察される。むろん1回限りの観察ではただ特定の数値が観察されただけなので、例えば確率変数 $X \sim B(n,p)$ としたときのある観察値より求めた統計量が真の確率 p に一致することは期待できない。しかしこうした観察が仮に繰り返される状況ではすると平均的に(確率変数としての)統計量の挙動を分析できる。統計量がしたがう確率分布の性質を調べることは偶然変動の結果として「データがどのような確率法則により発生しているか」を科学的に分析することを意味する。

点推定

ここで母集団として確率分布を想定しよう。多くの応用上の問題では母集団としての確率分布についてはある程度の情報があるので、分布型についてはほぼ分かっていると見なすことが自然な状況がある。ここでは母集団としての確率分布(母集団分布)はその形を完全には分からずに確率分布の母数は未知であることを想定すると、こうした統計的モデルは**パラメトリック・モデル**と呼ばれる[27]。ここで未知母数を θ(ギリシャ文字のシータ)で表し、この母数について母集団から独立な標本抽

[27]確率分布の形を想定しないノン・パラメトリック分析、一部分の母数を想定するセミ・パラメトリック分析も重要である。

出により得られる標本 $(X_1, X_2, \cdots, X_n)$ から推測を行う。大きさ n の標本の関数

$$\hat{\theta}_n\left(X_1, X_2, \cdots, X_n\right)$$

よって母数 θ を推定するとき関数 $\hat{\theta}_n\left(X_1, X_2, \cdots, X_n\right)$ (シータ・ハット) を母数 θ の推定量とする。n 個の標本が観測値として値 $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ をとるとすると、これらの値を推定量に代入して得られる

$$\hat{\theta}_n\left(x_1, x_2, \cdots, x_n\right)$$

が推定値である。推定量 $\hat{\theta}_n = \hat{\theta}\left(X_1, X_2, \cdots, X_n\right)$ の方は統計量なので確率変数である。これに対して推定値 $\hat{\theta}_n\left(x_1, x_2, \cdots, x_n\right)$ の方は母数 θ がスカラーの場合には単なるデータの観測値で計算された数値となる。

例 7.1：平均 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ を考える。未知の母平均 μ の推定量としては標本平均

$$\hat{\mu}\ (= \bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7.1}$$

が自然である。(ここで記号 $\hat{\mu}$(ミュー・ハット) は未知母数 μ の推定量の意味とする。) 母分散 σ^2 の推定量としては標本不偏分散

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 \tag{7.2}$$

が最も良く用いられる。χ^2- 分布の性質から $\mathbf{E}[s_n^2] = \sigma^2$ となる。同様に 3 次積率・4 次積率 (moments) の推定量として平均周りの高次の標本積率

$$m_3 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^3\ , \tag{7.3}$$

$$m_4 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^4 \tag{7.4}$$

などが用いられる。

正規分布を仮定すると 3 次・4 次積率 $\mathbf{E}[(X-\mu)^3] = 0, \mathbf{E}[(X-\mu)^4] = 3\sigma^4$ となる。したがって歪度 (skewness)$\kappa_3 = \mathbf{E}[(X-\mu)^3]/\sigma^3 = 0$, 尖度 (kurtosis)$\kappa_4 = \mathbf{E}[(X-\mu)^4]/\sigma^4 - 3 = 0$ となる。(なお尖度を $\kappa_4^* = \mathbf{E}[(X-\mu)^4]/\sigma^4$ で定義することもある。) このことより母集団分布が正規分布であることの妥当性は統計量として

$$\begin{aligned} b_1 &= \frac{\sqrt{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^3}{[\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}]^3}\ , \\ b_2 &= \frac{n\sum_{i=1}^{n}(X_i - \bar{X}_n)^4}{[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2]^2} - 3\frac{(n-1)}{(n+1)} \end{aligned}$$

などが利用される。標本数 n が大きければ中心極限定理 (CLT) により近似的に

$$b_1 \overset{a}{\sim} N(0, \frac{6(n-2)}{(n+1)(n+2)})$$

$$b_2 \overset{a}{\sim} N(0, \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)})$$

となる [28]。

積率法 (Moment-Method, モーメント法)
一般に大きさ n の独立標本 $X_1, X_2, \cdots, X_n$ とする。n 個の標本から計算することのできる統計量、未知母数 θ の推定量としては一般には様々な方法が考えられる。

統計家カール・ピアソン (K. Pearson) は標本から計算される積率（標本積率）を母集団の積率に一致させる積率法 (method of moments) を提唱した。母集団分布の期待値 $\mu = \mathbf{E}(X)$ とすると標本平均 $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$, 母集団分布の分散 $\sigma^2 = \mathbf{E}[(X - \mathbf{E}(X))^2]$ とすると標本 (不偏) 分散 $s_n^2 = (1/(n-1))\sum_{i=1}^n (X_i - \bar{X}_n)^2$, を代入する推定法が考えられる。こうして母集団と標本の積率を対応させる推定法は統計的な推定の対象が母集団としての確率分布の積率であれば利用することが可能であるが、より一般的な状況への適用は何らかの工夫が必要となる。近年では積率法を一般化して一般化積率法 (Generalized Method of Moments, 略して GMM) と呼ばれる方法も利用されることが多くなっている。GMM については 10.3 節で再び言及する。

7.2 尤度関数と十分統計量

積率法では母集団についての情報としては確率分布の幾つかの積率のみを利用していることに注意しよう。パラメトリック・モデルの設定では母集団の確率分布が既知であるからその情報を利用することが考えられる。その情報が正しければより効率的な推定方法が得られることが期待できるのである。

母集団がしたがう確率関数を $p(x|\theta)$(密度関数なら $f(x|\theta)$), $X_1, X_2, \cdots, X_n$ をランダム・サンプリングによる n 個の独立標本とする。このとき確率変数 $X_1, X_2, \cdots, X_n$ は互いに独立なので、離散確率分布の場合には標本 $X_1, X_2, \cdots, X_n$ の同時確率関数は

$$p(x_1, \cdots, x_n|\theta) = \prod_{i=1}^n p(x_i|\theta) \tag{7.5}$$

[28] 例えば Kendall and Stuart,"The Advanced Theory of Statistics," Vol.1, Charles Griffin & Company Limited, p.325-326 に説明があるが高次の積率の存在条件などを仮定する必要がある。

あるいは連続分布の場合には同時密度関数は

$$f\left(x_1,\cdots,x_n|\theta\right)=\prod_{i=1}^{n}f(x_i|\theta) \tag{7.6}$$

と書ける。この関数を同時確率分布としてではなく母数 θ の関数とみて、母数 θ の関数

$$L_n(\theta|x_1,\cdots,x_n)=\prod_{i=1}^{n}p(x_i|\theta) \tag{7.7}$$

を尤度関数 (likelihood function) と呼ぶが、変数 x_i $(i=1,\cdots,n)$ を省略して $L_n(\theta)$ と表すことも多い。母集団分布として連続型分布を仮定するときにも同様に母集団の密度関数を $f(x|\theta)$ とすれば同時密度関数は $f\left(x_1,\cdots,x_n|\theta\right)=\prod_{i=1}^{n}f(x_i|\theta)$ となるので独立標本の場合には尤度関数は,

$$L_n(\theta|x_1,\cdots,x_n)=\prod_{i=1}^{n}f(x_i|\theta) \tag{7.8}$$

によって定められる。

例 7.2：成功と失敗の事象 (あるいはコインの表 (H) と裏 (T)) に対応する二つの値 (1,0) をとる確率変数をベルヌイ試行と呼び、1 をとる確率 $p(0<p<1)$ を未知の母数とする。確率関数は

$$p(x|p)=p^x\left(1-p\right)^{1-x},\quad x=0,1 \tag{7.9}$$

であり、互いに独立なベルヌイ試行 X_i $(i=1,\cdots,n)$ を n 回観測すれば n 個のデータ $x_1,\cdots,x_n$ にもとづく尤度関数は

$$\begin{aligned} L_n(p|x_1,\cdots,x_n) &= \prod_{i=1}^{n}p^{x_i}\left(1-p\right)^{1-x_i} \\ &= p^{\sum_{i=1}^{n}x_i}\left(1-p\right)^{n-\sum_{i=1}^{n}x_i} \end{aligned}$$

で与えられる。

母数 $\boldsymbol{\theta}$ はスカラーまたは有限個の要素のベクトルで母数空間はユークリッド空間、あるいはその一部を構成するのが典型的状況である。この場合に母数について観察データより統計的推測を行う形式はパラメトリック推測と呼ばれ標準的な統計的方法である。ここで統計量としてはなるべくその性質が分かりやすいものが選ばれる

が、パラメトリック推測では特になるべく情報の損失が少ないような統計的方法を考えるのが重要となる。そこで十分統計量 (sufficient statistic) を次のように定めよう。

定義 7.1： 任意の標本空間上の事象 A に対し統計量 $T(X_1, \cdots, X_n)$ の値を条件とするとき

$$P\left((x_1, \cdots, x_n) \in A | T(X_1, \cdots, X_n) = t, \theta\right) \tag{7.10}$$

が母数 θ に無関係となるとき T を十分統計量と呼ぶ。

このことから十分統計量の値 $T = t$ が分かればその他の情報は母数 θ についての統計的推測には不要であることを意味する。標本分布についての情報は尤度関数に集約されることは次のように正確に表現できる。

定理 7.1 (分解定理): 標本 $(X_1, \cdots, X_n)$ に対して同時密度関数 $f(x_1, \cdots, x_n)$ が存在して $\mathbf{x} = (x_1, \cdots, x_n)$ とする。このとき任意の統計量 T が十分統計量である為の必要十分条件は

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}) \tag{7.11}$$

と分解できることである。

このことは次のようにして確かめることができる。連続確率分布について T を十分統計量とすると、T の条件付分布が母数 θ に依存せず $f(\mathbf{x}|T = t) = h(\mathbf{x}; t)$ となる。このとき

$$f(\mathbf{x}|\theta) = f(\mathbf{x}|T = t)g(t|\theta) = h(\mathbf{x}; t)g(t|\theta)\ .$$

逆に $f(\mathbf{x}|\theta) = h(\mathbf{x}; t)g(t|\theta)$ とする。ここで $A_t = \{\mathbf{x}|T(\mathbf{x}) = t\}$ とすると

$$\int_{A_t} f(\mathbf{x}|\theta)d\mathbf{x} = g(T(\mathbf{x})|\theta) \int_{A_t} h(\mathbf{x}|\theta)d\mathbf{x}$$

より

$$f(\mathbf{x}|T = t(\mathbf{x})) = \frac{f(\mathbf{x}|\theta)}{\int_{A_t} f(\mathbf{x}|\theta)d\mathbf{x}} = \frac{h(\mathbf{x}, \mathbf{t})g(t|\theta)}{g(t|\theta)\int_{A_t} h(\mathbf{x}, t)d\mathbf{x}}$$

は母数 θ と独立となる。

ここでの議論は標本がベクトル $\mathbf{X}_i$ $(i = 1, \cdots, n)$ の場合や母数 (パラメター)$\boldsymbol{\theta} = (\theta_i)$ が有限次元のベクトルの場合には成り立つ。

例 7.3： 互いに独立な確率変数が $X_i \sim N(\mu, \sigma^2)$ $(i = 1, \cdots, n)$ とすると、同時密

度は

$$
\begin{aligned}
f(x_1,\cdots,x_n|\mu,\sigma^2) &= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x}_n)^2+n(\bar{x}_n-\mu)^2\right])
\end{aligned}
$$

となる。したがって十分統計量は$T_1=\sum_{i=1}^n X_i$ (あるいは$\bar{X}_n$), $T_2=\sum_{i=1}^n(X_i-\bar{X}_n)^2$ となる。この場合は母数$\boldsymbol{\theta}$は2個の母数μ,σ^2からなる2次元の確率変数ベクトルである。

例 7.4： 互いに独立な確率変数が$X_i\sim P_0(\lambda)$ $(i=1,\cdots,n)$ とすると周辺確率関数$f(x_i)=e^{-\lambda}\lambda^{x_i}/x_i!$ より、十分統計量は$T=\sum_{i=1}^n X_i$ となる。

例 7.5： 互いに独立な確率変数が$\mathbf{X}_i\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ $(i=1,\cdots,n)$ とすると、十分統計量は$T_1=\bar{\mathbf{X}}_n$, $\mathbf{A}_n=\sum_{i=1}^n(\mathbf{X}_i-\bar{\mathbf{X}}_n)(\mathbf{X}_i-\bar{\mathbf{X}}_n)'$ である。

例 7.6： 互いに独立に一様分布にしたがう確率変数を$X_i\sim U(0,\theta)$ $(i=1,\cdots,n)$ とすると、同時密度関数が

$$
f(x_1,\cdots,x_n|\theta)=(\frac{1}{\theta})^n I(0<x_1,\cdots,x_n<\theta) \tag{7.12}
$$

となるので順序統計量$X_{(n)}=\max_{1\le i\le n}X_i$が十分統計量となる。この場合には分布の台 (密度関数$f(x|\theta)>0$となる範囲が母数に依存するという非正則 (irregular) な統計モデルとなる。

十分統計量は情報損失をせずに統計的推測を行う為には基本的で重要な統計量である。統計分析において得られる標本そのものも定義7.1から十分統計量ではあるが、nが大きければ扱いが困難なのでなるべく数が少ない十分統計量を利用することが望ましい。最も数が少ない十分統計量は最少十分統計量 (minimum sufficient statistics) と呼ばれるが上で述べた統計量が例である[29]。

7.3 最尤推定量

積率法に対して、ロナルド・フィッシャー (R.A.Fisher) は最尤法 (maximum likelihood method）と呼ばれる別の推定法を提案した。この方法は様々な統計的推測の問題に適用可能なので応用範囲が広く，また結果として得られる推定量がいくつか

[29]例えば竹村彰通「現代数理統計学」が詳しく議論している。

の望ましい性質を持つことが知られるようになったので統計的分析では基本的な方法である。

離散型の母集団分布からの独立標本が得られる場合には母集団の確率関数 $p(x|\theta)$ より同時確率関数を母数 θ の関数と見た

$$L_n(\theta|x_1,\cdots,x_n)=\prod_{i=1}^{n}p(x_i|\theta) \tag{7.13}$$

を尤度（likelihood）関数と呼ぶが、変数 $x_1,\cdots,x_n$ を略して $L_n(\theta)$ と表すことも多い。最尤推定値（maximum likelihood estimate）とはこの尤度関数を最大にするような θ の値によって定義する。関数 $L_n(\theta)$ は変数 $x_1,x_2,\cdots,x_n$ の関数なので $L_n(\theta)$ を最大とする θ を $\hat{\theta}_n=\hat{\theta}\,(x_1,x_2,\cdots,x_n)$ と表現する。ここで変数 $x_1,x_2,\cdots,x_n$ を標本 $X_1,X_2,\cdots,X_n$ で置き換えて得られる推定量

$$\hat{\theta}_n=\hat{\theta}\,(X_1,X_2,\cdots,X_n)$$

を最尤推定量 (maximum likelihood estimator, 略して MLE) と呼ぶ。

連続型の母集団分布から独立標本が得られる場合には母集団分布の密度関数 $f(x|\theta)$ より同時密度関数は $f\,(x_1,\cdots,x_n|\theta)=\prod_{i=1}^{n}f(x_i|\theta)$ となるので尤度関数は

$$L_n(\theta|x_1,\cdots,x_n)=\prod_{i=1}^{n}f(x_i|\theta)$$

で定める。この関数を最大化し、変数 x_i $(i=1,\cdots,n)$ を確率変数 X_i $(i=1,\cdots,n)$ に置き換えれば最尤推定量が得られる。

例 7.2′：例 7-2 における尤度関数の対数をとることで得られる対数尤度関数 (log-likelihood) は

$$l_n(p)=\log L_n(p)=\sum_{i=1}^{n}x_i\log(p)+\left(n-\sum_{i=1}^{n}x_i\right)\log(1-p) \tag{7.14}$$

であるが、この対数尤度関数を変数 p で微分してゼロとおけば

$$\frac{\partial l_n}{\partial p}=\frac{\sum_{i=1}^{n}x_i}{p}+\frac{\sum_{i=1}^{n}x_i-n}{1-p}=0 \tag{7.15}$$

である。この方程式のように尤度関数や対数尤度関数を母数で微分してゼロと置いた方程式を**尤度方程式** (likelihood equation) と呼ぶ。この場合には微分して得られる解は確かに最大値であることを容易に確かめることができ、

$$\hat{p}_{ML}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

となる。したがって母数としての確率 p の最尤推定量

$$\hat{p}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7.16}$$

を得る。

例 7.3′：母集団分布を平均 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ とする。母集団の平均と分散は未知母数として推定法を構成したいが、互いに独立な n 個の標本 (確率変数) として観測値 $(x_1, \cdots, x_n)$ が得られるとしよう。尤度関数は

$$L_n(\mu, \sigma | x_1, \cdots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{7.17}$$

で与えられるので対数尤度関数は

$$l_n(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{7.18}$$

となる。まず母数 μ で微分すれば

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

より母平均の最尤推定値として

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{7.19}$$

が得られる。次に μ に $\hat{\mu}_{ML}$ を代入した尤度関数、すなわち集約尤度 (concentrated likelihood)（あるいはプロファイル尤度 (profile likelihood) と呼ばれることがある）を母数 σ で微分すれば

$$\frac{\partial l_n}{\partial \sigma} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_{ML})^2}{2\sigma^4} = 0$$

が得られる。この尤度方程式の根は

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{7.20}$$

となる。したがって平均と分散という二つの未知母数に対する最尤推定量はそれぞれ

$$\begin{aligned}\hat{\mu}_{ML} &= \frac{1}{n}\sum_{i=1}^{n} X_i \,(=\bar{X}_n)\,, \\ \hat{\sigma}^2_{ML} &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\end{aligned}$$

で与えられる。ここで得られた最尤推定量は確かに尤度関数を最大化している。分散の推定量は標本不偏分散と定数倍だけ異なっているが、この場合には尤度関数の最大化することを確かめることができる。

例 7.7：母集団分布が平均ベクトル $\boldsymbol{\mu}$, 分散共分散行列 $\boldsymbol{\Sigma}$ の正規分布 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 互いに独立な n 個の標本ベクトル $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ がとする。（各 $\mathbf{X}_i$ は $p \times 1$ の確率変数ベクトルである。）次の補題を利用すると最尤推定量は

$$\hat{\boldsymbol{\mu}}_{ML} = \bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \tag{7.21}$$

および

$$\hat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{n}\mathbf{A}_n \tag{7.22}$$

で与えられる。ただし変動和は

$$\mathbf{A}_n = \sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)' \tag{7.23}$$

である。

補題 7.2： 正定符号行列 $\boldsymbol{\Sigma}$ に対して正定符号行列 $\mathbf{A}_n$ の関数

$$g(\boldsymbol{\Sigma}) = -n\log|\boldsymbol{\Sigma}| - \mathrm{tr}[\boldsymbol{\Sigma}^{-1}\mathbf{A}_n] \tag{7.24}$$

を最大化する解は $\boldsymbol{\Sigma} = (1/n)\mathbf{A}_n$ となる。

証明 ： 対称行列 $\mathbf{C} = (c_{ij})$ に対してトレース (trace) を $\mathrm{tr}(\mathbf{C}) = \sum_i c_{ii}$ で定義する。ここでの証明の概略は三角行列 $\mathbf{T} = (t_{ij}), t_{ij} = 0\ (i < j)$ として変換 $\mathbf{A}_n^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{A}_n^{1/2} = \mathbf{H} = \mathbf{T}\mathbf{T}'$ がとれることを利用する。このとき未知母数の関数として $g(\boldsymbol{\Sigma})$ に比例する量として $g^*(\boldsymbol{\Sigma}) = n\log|\mathbf{T}^2| - \mathrm{tr}\mathbf{T}\mathbf{T}' = \sum_{i=1}^{p}[\log t_{ii}^2 - t_{ii}^2] - \sum_{i<j} t_{ij}^2$ とすると、この関数は $t_{ii}^2 = n, t_{ij} = 0\ (i < j)$ のとき最大値をとるので結果が得られる。 **Q.E.D.**

例 7.8：統計モデルが複雑になるにつれて尤度関数や尤度方程式は複雑になる。例えば一般化極値分布は母数として μ, σ, ξ を用いて

$$G(z) = \exp\left(-[1 + \xi(\frac{z-\mu}{\sigma})]^{-1/\xi}\right) \tag{7.25}$$

で与えられた。グンベル分布は $\xi \to 0$ $(-\infty < \xi < +\infty)$ の極限とみなせるが推定量の挙動の分析は $\xi \sim 0$ において注意する必要がある。

最尤法を使う場合には統計モデルが複雑になるにつれて尤度関数や尤度方程式が複雑になったり、例 7.8 のように統計量は明示的に表現されるとは限らないことに注意する必要がある。

7.4 小標本の標準理論

推定の基準

例えばベルヌイ試行の確率を未知母数 θ とするとき、未知母数 θ の推定量として n 個の標本から作った統計量候補の中からどの推定量を使ったらよいだろうか？ここで母数の点推定量を選択する基準を考察しよう。互いに独立な標本 $X_1, \cdots, X_n$(ここでは簡単化のために X_i はスカラーの確率変数とする) が得られるとき、推定量を母数の関数と見なせば $\hat{\theta}_n(X_1, \cdots, X_n)$ は母数 θ の回りで分布すると考えられる。したがって、推定量の良さを測る自然な尺度としては真の母数の回りで推定量がばらつく確率

$$P\left(|\hat{\theta}_n - \theta| < z\right) \tag{7.26}$$

を小さくすることが考えられる。しかしこの確率を z に依存し、様々な推定量について具体的に評価するのは一般には簡単ではない。そこで確率の代わりに推定量の真の母数の回りの 2 次積率

$$\mathrm{MSE}(\hat{\theta}_n) = \mathbf{E}\left[\hat{\theta}_n - \theta\right]^2 \tag{7.27}$$

を小さくする基準が考えられる。ここで MSE とは平均二乗誤差 (mean squared error) の略であり、推定量の選好の基準として使うときにはなるべくこの MSE が小さくなるように推定量を選ぶことが多くの場合には自然であり、ある推定量の MSE が別の推定量の MSE より小さい時により効率的 (efficient) と言う [30]。

[30] 例えば景気予測 (例えば GDP 成長率や物価) のエコノミストにとっては真の値 θ の周りに対称な損失関数を規準とすることが適切か否かは疑問であり、非対称な損失が妥当な場合もありうる。

この平均二乗誤差を推定量の期待値 $\mathbf{E}(\hat{\theta}_n)$ の回りで展開すると

$$\begin{aligned}\mathbf{E}\left[|\hat{\theta}_n-\theta|^2\right] &= \mathbf{E}\left[\left((\hat{\theta}_n-\mathbf{E}(\hat{\theta}_n))+(\mathbf{E}(\hat{\theta}_n)-\theta)\right)^2\right]\\ &= \mathbf{E}[\left(\hat{\theta}_n-\mathbf{E}(\hat{\theta}_n)\right)^2]+[\left(\mathbf{E}(\hat{\theta}_n)-\theta\right)^2]\\ &= \mathbf{V}\left(\hat{\theta}_n\right)+\left(\mathbf{E}(\hat{\theta}_n)-\theta\right)^2\end{aligned}$$

となる。右辺の第 1 項は推定量の分散であるが、第 2 項は推定量の期待値と真の母数 θ 差

$$B(\hat{\theta}_n)=\mathbf{E}(\hat{\theta}_n)-\theta \tag{7.28}$$

であり推定の偏り (bias, バイヤス) と呼ばれる）の二乗を示している。推定量が真の母数の回りに対称的に分布していれば推定のバイヤスはゼロとなることが期待できる。

例 7.9 : 例 7.1 では推定量 $T_1=\hat{X}_n$ の分布は二項分布の標本平均なので期待値 θ，分散は $(1/n)\theta(1-\theta)$ で与えられる。したがってこの推定量に偏りがないので、平均二乗誤差は

$$MSE(T_1)=\mathbf{E}\left(\hat{X}_n-\theta\right)^2=\frac{1}{n}\theta(1-\theta) \tag{7.29}$$

となる。同様に $T_2=\hat{X}_m$ $(m<n)$ も偏りがないので平均二乗誤差は

$$MSE(T_2)=\frac{1}{m}\theta(1-\theta) \tag{7.30}$$

となる。この例では推定量 T_1 の平均二乗誤差が T_2 の平均二乗誤差に比べてどんな母数 p をとっても一様に小さくなる。そこで誤差を平均的に小さくする意味では T_2 より T_1 が望ましいと云える。

例 7.10 : 例 7.9 において推定量 $T_3(X_1,\cdots,X_n)=1/2$ とすることは可能である。この推定法はデータからどの様な情報が得られたとしてもベルヌイ試行は公平であると見ると解釈できる。この推定量は確かに真の状態が $\theta=1/2$ であれば誤差はゼロとなり、こうした例は未知母数の推定では常に考えることができる。実際のデータとして有限個の観測値が得られる状況ではこうした推定法が誤りであることを示すことはできないという事実が重要であるが、未知母数の値 $(0\le\theta\le 1)$ が $\theta=1/2$ とは限らなければ明らかに不合理である。ここで未知母数 θ について何らかの意味で一様に悪くない性質 (一様性の推定基準と呼ぶ) を導入することによりこうした特定の状態についてのみ良い性質を持つ (言わば super 効率的な) 推定法を排除できる。

不偏性の基準

推定量の基準として不偏性 (unbiasedness) とは偏り (bias) のない性質であり，任意の母数 $\theta \in \mathbf{\Theta}$ のとる値に対して

$$\mathbf{E}(\hat{\theta}_n) = \theta \tag{7.31}$$

が成り立つことである。この性質を持つ推定量を不偏推定量 (unbiased estimator) と呼ぶ。

例 7.11 正規分布の例では標本平均 $\bar{X}_n$ は母平均 μ の不偏推定量であり、標本 (不偏) 分散 s_n^2 $(= (1/(n-1))\sum_{i=1}^n (X_i - \bar{X}_n)^2)$ は母分散 σ^2 の不偏推定量となる。ただし、最尤推定量 ((7.20)) の期待値は $\mathbf{E}\left(\hat{\sigma}_{ML}^2\right) = [(n-1)/n]\sigma^2$ となるので不偏推定量ではない。平均二乗誤差は $MSE(\hat{\sigma}_{ML}^2) < MSE(s_n^2)$ となる。

未知母数の推定量としての範囲として不偏推定量のクラスに限れば十分統計量に依存する推定量を用いる事が望ましいことを次のようにより示すことができる。まず統計量 $\phi(\mathbf{X})$ ((縦) 確率変数ベクトル $\mathbf{X} = (X_1, \cdots, X_n)'$ は θ の不偏推定量, $T(\mathbf{X})$ を十分統計量とする。このとき十分統計量で条件付した統計量 $\varphi(t) = \mathbf{E}[\phi(\mathbf{X})|T = t]$ とすると不等式

$$\begin{aligned}
\mathbf{E}[(\phi(\mathbf{X}) - \theta)^2] &= \mathbf{E}\left[(\phi(\mathbf{X}) - \varphi(T)) + (\varphi(T) - \theta)\right]^2 \\
&= \mathbf{E}[(\phi(\mathbf{X}) - \varphi(T))^2] + \mathbf{E}[(\varphi(T) - \theta)^2] \\
&\geq \mathbf{V}[\varphi(T)]
\end{aligned}$$

が成り立つ。これより次の定理が得られる。

定理 7-3 (Blackwell-Rao の定理)： 標本 $\mathbf{X} = (X_1, \cdots, X_n)'$ の各要素が互いに独立・同一分布にしたがうとき、$T(\mathbf{X})$ を十分統計量、統計量 $\phi(\mathbf{X})$ は母数 θ の不偏推定量とする。このとき、
(i) $\varphi(t) = \mathbf{E}[\phi(\mathbf{X})|T = t]$ は不偏推定量,
(ii)$Var[\varphi(T)|\theta] \leq Var[\phi(\mathbf{X})]$ となる。

例 7.12： 互いに独立な確率変数 $X_i \sim N(\theta, 1)$ $(i = 1, \cdots, n)$ のとき統計量 $\phi(\mathbf{X}) = X_1$ とする。十分統計量は $T = \frac{1}{n}[X_1 + \cdots + X_n]$ であるから二次元正規分布の性質より $Cov(X_1, T) = 1/n,\ Var(T) = 1/n$ となることを利用すると、十分統計量による X_1 の条件付期待値は

$$\mathbf{E}[X_1|T = t] = \theta + \frac{1/n}{1/n}(t - \theta) = t$$

で与えられる。

推定量が不偏性を持てば真の回りに偏りがないという意味では適切な推定量であるから不偏推定量のみを比較の対象とすることが考えられる。ここで推定量のクラスを不偏性を満たしている推定量のみに範囲を限定しておくと推定量の適切性、最適性の判定は比較的に容易となる。

ここで互いに独立な標本 $X_1, \cdots, X_n$ の同時密度関数が $f(\mathbf{x}|\theta)$ で与えられるとき ((縦) ベクトル $\mathbf{x} = (x_1, \cdots, x_n)'$)，任意の不偏推定量 $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X})$ $(\mathbf{X} = (X_1, \cdots, X_n))$ は任意の $\theta \in \mathbf{\Theta}$ について条件

$$\mathbf{E}\left(\hat{\theta}_n\right) = \theta \tag{7.32}$$

を満足する。この条件を θ について微分すると条件 $\frac{\partial}{\partial\theta}\mathbf{E}\left(\widehat{\theta}\right) = 1$ は

$$\begin{aligned}
\int \widehat{\theta}(x)\, \frac{\partial}{\partial\theta} f(\mathbf{x}|\,\theta)\, d\mathbf{x} &= \int \widehat{\theta}(\mathbf{x})\, \frac{\partial \log f(\mathbf{x}|\,\theta)}{\partial\theta} f(\mathbf{x}|\,\theta)\, d\mathbf{x} \\
&= \mathbf{E}\left[\widehat{\theta}(\mathbf{X})\, \frac{\partial \log f(\mathbf{X}|\,\theta)}{\partial\theta}\right] \\
&= Cov\left(\widehat{\theta}(\mathbf{X}), \frac{\partial \log f(\mathbf{X}|\,\theta)}{\partial\theta}\right) \leq \sqrt{V\left(\widehat{\theta}(\mathbf{X})\right)}\sqrt{I_n(\theta)}
\end{aligned}$$

となる。ただし途中の等号では

$$\mathbf{E}\left[\frac{\partial \log f(\mathbf{x}|\,\theta)}{\partial\theta}\right] = \frac{\partial}{\partial\theta}\left[\int f(\mathbf{x}|\,\theta)\, d\mathbf{x}\right] = 0 \tag{7.33}$$

となる条件を用いた。(なおここでは $\int f(\mathbf{x}|\,\theta)\, d\mathbf{x} = 1$ より求めたが、一般には期待値 (積分) と (母数についての) 微分が交換できる条件が成り立つとは限らないが後述の正則条件を用いた。) また最後の評価では Cauchy-Schwartz 不等式 $Cov(X,Y) \leq \sqrt{V(X)}\sqrt{V(Y)}$ を利用したが 2 章で説明したようにこの条件は相関係数の絶対値が 1 より小さいことと同値である。

特に n 個の標本が互いに独立な確率変数の場合には

$$\begin{aligned}
I_n(\theta) &= \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\sum_{i=1}^{n} \log f(X_i|\,\theta)\right)^2\right] \\
&= n\,\mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X_1|\,\theta)\right)^2\right] = n\, I_1(\theta)
\end{aligned}$$

より不偏推定量の分散は

$$V\left(\hat{\theta}_n\right) \geq \frac{1}{n\, I_1(\theta)}$$

となる。以上の結果を次のようにまとめておく。

定理 7-4： 同時分布について (次節で述べる) 正則条件を仮定する。任意の不偏推定量 $\hat{\theta}(X_1, \cdots, X_n)$ の分散は不等式

$$V(\hat{\theta}) \geq \frac{1}{n\, I_1(\theta)} \tag{7.34}$$

を満足する．ただし $I_1(\theta)$(あるいは $I_n(\theta)$) はフィッシャー情報量と呼ばれ

$$\begin{aligned} I_1(\theta) &= \mathbf{E}\left(\frac{\partial}{\partial\theta}\log f(x_1|\theta)\right)^2 \\ &= \mathbf{E}\left(-\frac{\partial^2}{\partial\theta^2}\log f(x_1|\theta)\right) \end{aligned}$$

で与えられる。

ここで上の導出の途中では結果が成り立つような幾つかの条件を用いた。次節で説明するが幾つかの数学的条件 (しばしば正則条件 (regularity condition) と呼ばれる) の下で一般的に最適な推定量の性質を導くことが可能なのである。

クラーメル (H. Cramer) とラオ (C.R. Rao) がほぼ同時に得たこの古典的なクラーメル・ラオ (CR) の不等式では左辺は推定量の分散なので右辺はその限界を示しているのでクラーメル・ラオ (CR) 下限と呼ばれている。したがってもし不偏推定量の中で右辺に一致する推定量が見つかれば最も良い推定量 (efficient estimator, 効率的推定量)、optimal estimator(最適推定量)) と呼ぶことが出来る。このような推定量が存在するとき推定量は一様最小不偏分散推定量 (uniformly minimum variance unbiased estimator, 略して UMVU) と呼ばれている。

例 7.13 : 母集団分布が平均 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ である場合を考える。仮に分散の母数 σ^2 の値を既知として母数を平均値 μ のみとして標本 1 個に対する情報量を計算すれば

$$\begin{aligned} I_1(\mu) &= \mathbf{E}\left(\tfrac{\partial}{\partial\mu}\log f(x|\mu)\right)^2 \\ &= \tfrac{1}{\sigma^4}\mathbf{E}\,(X-\mu)^2 = \tfrac{1}{\sigma^2} \end{aligned}$$

となる。このとき全体の標本からの推定量についてのクラメール・ラオの下限は $1/\left(nI_1(\theta)\right) = \sigma^2/n$ となり標本平均の分散に一致する。したがって標本平均 $\bar{X}$ は一様分散不偏推定量となる。

ここで CR 下限の導出より下限を達成するためには $\varphi(\mathbf{x}) = \hat{\theta}(\mathbf{x})$ $(\mathbf{x} = (x_1, \cdots, x_n)')$ とすると、CR 下限を導く不等式が等式となる同時密度関数についての条件

$$\varphi(\mathbf{x}) = \alpha(\theta)\frac{\partial \log f\,(\mathbf{x}|\theta)}{\partial\theta} + \beta(\theta)$$

を満足する必要がある。この条件を同時密度関数について解けば

$$f(\mathbf{x}|\theta) = B(\theta)h(\mathbf{x})\exp\left[A(\theta)\varphi(\mathbf{x})\right] \tag{7.35}$$

と表現できる。右辺に表れる確率分布型は指数分布族 (exponential family) と呼ばれている。統計学でよく利用する二項分布、ポワソン分布、正規分布、など多くの分布はこの分布族に入っている。

応用上の問題では母数がベクトルの場合がより一般的である。母数ベクトル $\boldsymbol{\theta} = (\theta_i)$ を $r \times 1$ ベクトルとすると、n 個の独立標本から得られるフィッシャー情報量は $r \times r$ 行列 $I_n(\boldsymbol{\theta}) = (I_{n,ij})$,

$$I_{n,ij} = \mathbf{E}\left[-\frac{\partial^2 \log f(\mathbf{X}|\theta)}{\partial\theta_i \partial\theta_j}\right] \tag{7.36}$$

により定義される。母数ベクトルについて $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ とすると CR-下限は行列の意味で同様に定義できるが、不偏推定量は $r \times 1$ ベクトル $\varphi(\mathbf{x})$ になり大小関係は行列の定符号行列の意味で

$$V[\varphi(\mathbf{X})] \geq I_n(\boldsymbol{\theta})^{-1} \tag{7.37}$$

となる。このとき第 i 要素 $\varphi_i(\mathbf{X})$ の分散は $I_n(\boldsymbol{\theta})^{ii}$ を $I_n(\boldsymbol{\theta})^{-1}$ の第 (i,i) 成分とすると

$$V[\varphi_i(\mathbf{X})] \geq I_n(\boldsymbol{\theta})^{ii} \tag{7.38}$$

となる。

こうした数理的表現とは別に実際に UMVU 不偏推定量を見つけ出すことはそれほど易しいことではない。例えば二項分布 $B(n,\theta)$ において $g(\theta) = \theta^2$ $(0 \leq \theta \leq 1)$ を推定する問題では $n = 1$ とすると明らかに $\hat{\theta} = \bar{X}_n^2$ は不偏ではない。また母数 $g(\theta) = \theta/(1-\theta)$ $(0 \leq \theta < 1)$ を推定する問題では不偏推定量は存在しない。仮に存在したとしてその推定量を $\varphi(\mathbf{X})$ とすると

$$\frac{\theta}{1-\theta} = \sum_{x=0}^{n} \varphi(x)_n C_x \theta^x (1-\theta)^{n-x} \tag{7.39}$$

となるはずである。ところが右辺は θ の n 次多項式であるからこの方程式の解は一般には存在しないのである。

バイアス補正

与えられた推定量が不偏ではない場合の少なくない。一般的に推定量のバイアスを減らす方法としてジャックナイフ (Jackknife) 法が知られている。この方法は次のよ

うに計算される。ある統計量 $T(X_1, \cdots, X_n)$ に対して $T_{-i}(X_1, \cdots, X_n)$ を第 i 標本を除いた統計量とする。そして $J_i(T) = nT(X_1, \cdots, X_n) - (n-1)T_{-i}(X_1, \cdots, X_n)$ を用いて

$$J(T) = \frac{1}{n}\sum_{i=1}^{n} J_i(T)$$

で統計量を定義する。例えば $X \sim B(1, \theta)$ に対し θ^2 の推定を考える。$T = \bar{X}_n^2$ としてジャックナイフ推定量を $J(T)$ とすると θ^2 の推定バイアスが小さくなることが確認できる。ジャッナイフ法は一度サンプリングしたデータより再びサンプリングして推定量を構成するという様々なリサンプリング (Resampling) 法の原型と見なすことができる。こうした resampling 法は高速計算機が利用可能になって急速に実用化されている。ただし、ここで構成した推定量は

$$J(T) = \bar{X}_n^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \tag{7.40}$$

となるので負値をとる確率が正となることの問題が生じ、この例では非負にとる必要がある。

7.5 統計的推定の漸近理論

一般的に最尤推定量が UMVU 推定量に一致する保証はない。実は UMVU 推定量を簡単に構成できる場合は母集団分布から独立に標本が得られる場合においてもかなり限られている。応用上で生じるより複雑な状況を設定すると小標本理論の展開はさらに困難となる。

しかしながら標本数 n がかなり大きい場合には「良い推定量」の議論を展開することは可能であり、標本数が大きいときの統計理論は漸近理論、大標本理論 (large sample theory) と呼ばれている。標本数が大きければ最尤推定量 $\hat{\theta}_{ML}$ の分散について一定の (正則条件と呼ばれる) 条件の下で近似的に

$$V\left[\sqrt{n}\left(\hat{\theta}_{ML} - \theta\right)\right] \cong \frac{1}{I_1(\theta)} \tag{7.41}$$

となることが知られている。このことは既に説明した中心極限定理を標本平均についての確率的法則と見るとその拡張された内容と解釈ができる。より実際的な観点からも最尤推定量（あるいは積率法などによる推定量など）が標本数がある程度多ければ近似的な意味でばらつきの少ない推定量であることが示せば、よい統計的な推定量となるので重要な意味がある。

最尤推定量の漸近的性質は次のように直観的に説明することができる。スカラー変数についての観測変数ベクトル $\mathbf{x} = (x_1, \cdots, x_n)'$, 対数尤度関数を $l_n(\theta) = \log L_n(\theta|\mathbf{x}) = \log \sum_{i=1}^n f(x_i|\theta)$ とすると、最尤推定量 $\hat{\theta}_{ML}$ は尤度方程式の解であるから、

$$\frac{\partial l_n(\theta|\mathbf{x})}{\partial \theta}|_{\hat{\theta}_{ML}} = 0 \tag{7.42}$$

である。ここで $\hat{\theta}_n$ が一致推定量 (consistent estimator) とは「任意の $\epsilon > 0$ に対して $n \to \infty$ のとき $P(\omega||\hat{\theta} - \theta| > \epsilon) \to 0$ となる」ことで定義しよう。最尤推定量が一致推定量であれば $\hat{\theta}_{ML} = \theta_0 + (\hat{\theta}_{ML} - \theta_0)$ として (意味を明確化するためにここで θ_0 は真の母数値とする)Taylor 展開すると

$$0 = \frac{\partial l_n(\theta_0|\mathbf{x})}{\partial \theta}|_{\theta_0} + \frac{\partial^2 l_n(\theta_0|\mathbf{x})}{\partial \theta^2}|_{\theta_0}(\hat{\theta}_{ML} - \theta_0) + o_p(1) \tag{7.43}$$

である。$o_p(1)$ は確率的に無視できる項を表現するが、正確には確率変数 Z_n に対して $Z_n/n^\alpha \xrightarrow{p} 0$ $(n \to \infty)$ のとき $Z_n = o_p(n^\alpha)$ と表現する。ここで漸近的には

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \frac{\frac{-\frac{1}{\sqrt{n}}\partial l_n(\theta|\mathbf{x})}{\partial \theta}|_{\theta_0}}{\frac{1}{n}\frac{\partial^2 l_n(\theta_0|\mathbf{x})}{\partial \theta^2}|_{\theta_0}} + o_p(1) \tag{7.44}$$

と表現できる。ここで分母の期待値は

$$\frac{1}{n}\mathbf{E}[-\frac{\partial^2 l_n(\theta|\mathbf{x})}{\partial \theta^2}|_{\theta_0}] = I_1(\theta_0)$$

であり、分子の期待値はゼロである (後に述べる正則条件の下で $\mathbf{E}[\frac{\partial l_n(\theta|\mathbf{x})}{\partial \theta}] = \int_{\mathbf{R}^n} \frac{1}{f(\mathbf{x}|\theta)}\frac{\partial f(\mathbf{x}|\theta)}{\partial \theta} f(\mathbf{x}|\theta)d\mathbf{x} = \frac{\partial}{\partial \theta}[\int_{\mathbf{R}^n} f(\mathbf{x}|\theta)d\mathbf{x}] = 0$ となるが、同時密度関数について $\int_{\mathbf{R}^n} f(\mathbf{x}|\theta)d\mathbf{x} = 1$ を利用した。) ここで分母は大数の法則よりフッシャー情報量に確率収束する。他方、分子は以下の例が示すように中心極限定理により $N(0, I_1(\theta_0))$ に収束する。したがって、次のような結果が得られる。

定理 7-5： (最尤推定量の漸近的性質) 独立標本 $X_1, \cdots, X_n$ が同一の分布にしたがい密度関数が $f(x|\theta)$ で与えられるとき，一定の正則条件 (regularity condition) の下で $n \to \infty$ につれて

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{\mathcal{L}} N(0, I_1(\theta_0)^{-1}) \tag{7.45}$$

となる。ここで $I_1(\theta)$ はフィッシャー情報量

$$I_1(\theta) = \mathbf{E}[-\frac{\partial^2 l_1}{\partial \theta^2}(\theta)] = \mathbf{E}[(\frac{\partial l_1}{\partial \theta}(\theta))^2] \tag{7.46}$$

で与えられる。

この結果より最尤推定量は多くの場合には漸近的に CR 下限を達成することが分かるが、最尤推定量の一致性と漸近正規性を導く十分条件を正確に表現しようとすると複雑に見えよう。なお、標準的な例では十分条件は満たされている場合が多いが例 7-6 に挙げた一様分布など確率分布の台 (密度関数が正の領域, サポート) が母数に依存するような例が反例となる。

例 7-14： 確率変数 $X_1, X_2, \cdots, X_n$ が互いに独立に正規分布 $N(0, \sigma^2)$ にしたがうときに母数 σ^2 の推定を考える [31]。尤度関数は

$$L_n = (\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2} \tag{7.47}$$

より対数尤度関数は

$$l_n = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2$$

となる。したがって母数で 2 回偏微分すると

$$\frac{1}{n}\frac{\partial^2 l_n}{\partial(\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{n}\frac{2}{2\sigma^6}\sum_{i=1}^n x_i^2$$

よりフィッシャー情報量は

$$I(\sigma^2) = \frac{1}{n}\mathbf{E}[-\frac{\partial^2 l_n(\sigma^2|\mathbf{x})}{\partial\sigma^2}] = \frac{1}{2\sigma^4}$$

となる。他方、 1 回の偏微分は

$$\frac{1}{\sqrt{n}}\frac{\partial l_n}{\partial\sigma^2} = \frac{1}{\sqrt{n}\sigma^4}\sum_{i=1}^n [x_i^2 - \sigma^2]$$

となる。スコア関数と呼ばれるこの項は期待値ゼロの独立和であるから中心極限定理を適用することができる。これより漸近的に $N(0, I_1(\sigma))$ にしたがうことがわかる。したがって

$$\sqrt{n}\left[\hat{\sigma}^2_{ML} - \sigma^2\right] \sim N(0, I_1(\sigma^2)^{-1}) \tag{7.49}$$

となる [32]。こうした最尤推定量に関する漸近的結果は母数 $\theta = (\theta_i)$ が有限次元 $(p\times 1)$ のベクトルの場合についても一定の正則条件の下で成立する。

[31] より一般に正規分布 $N(\mu, \sigma^2)$ のとき母数ベクトル $\boldsymbol{\theta} = (\mu, \sigma^2)$ についても同様の結果が得られるが導出はより複雑になる。

[32] この場合には最尤推定量は $\hat{\sigma}^2 = (1/n)\sum_{i=1}^n x_i^2$ より直接に漸近分布を求めることもできる。

定理 7-6： 独立ベクトル標本 $X_1, \cdots, X_n$ が同一の分布にしたがい密度関数が $f(\mathbf{x}|\boldsymbol{\theta})$ で与えられるとする。$r \times 1$ $(r \geq 1)$ 母数ベクトル $\boldsymbol{\theta} = (\theta_i)$ のとき一定の正則条件 (regularity condition) の下で $n \to \infty$ につれて

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, I_1(\boldsymbol{\theta}_0)^{-1}) \tag{7.50}$$

となる。ここで $I_1(\boldsymbol{\theta})$ はフィッシャー情報量

$$I_1(\boldsymbol{\theta}) = (\mathbf{E}[-\frac{\partial^2 l_1}{\partial\theta_i \partial_j \theta}(\boldsymbol{\theta})]) = (\mathbf{E}[(\frac{\partial l_1}{\partial \theta_i}(\theta)\frac{\partial l_1}{\partial \theta_j}(\theta))]) \tag{7.51}$$

で与えられる。(ただし $I_1(\boldsymbol{\theta})$ は $r \times r$ 正則行列とする。)

正則条件について

最尤推定法は応用上で用いられることが多いが、漸近論が成立するには一定の正則条件 (regularity conditions) が必要であり、条件を満たさない場合の議論は非正則問題と呼ばれている。密度関数の微係数が母数に依存せず母数について (数回の) 微係数が存在し、期待値 (積分) 操作と母数についての微分操作の交換可能な条件が必要であるが、典型的な非正則な例に言及しておこう。

例 7-15： 独立な確率変数列 X_i $(i = 1, \cdots, n)$ が指数分布にしたがい、密度関数

$$f(x|\theta) = e^{-(x-\theta)} \quad (x > \theta) \tag{7.52}$$

とする。このとき最尤推定量は $\hat{\theta}_n = \min_{1 \leq i \leq n} X_i$ である。ここで

$$\begin{aligned} P(\min X_i - \theta \leq z) &= 1 - P(\min X_i - \theta > z) \\ &= 1 - \textstyle\prod_{i=1}^n P(X_i - \theta > z) = 1 - e^{-nz} \end{aligned}$$

となるので $Z = n(\min_{1 \leq i \leq n} X_i - \theta)$ の分布は指数分布となる。
この例では $\frac{\partial^2 l_n(\theta)}{\partial \theta^2} = 0$, $\frac{\partial l_n(\theta)}{\partial \theta} = n$ となる。

この例のように極限分布への収束次数は常に $\sqrt{n}$ とは限らない。また期待値操作と微分の交換は一般には保障されないことに注意が必要である。例えば定理 7-5(母数がスカラーの場合), 定理 7-6(母数がベクトルの場合) での正則性の十分条件は (i) 確率分布の台 ($f(\mathbf{x}|\boldsymbol{\theta}$ が母数に依存しない)、(ii) 母数ベクトル $\boldsymbol{\theta} = (\theta_i)$ $(i = 1, \cdots, p)$ が真の母数値 $\boldsymbol{\theta}_0 = (\theta_{0i})$ の近くにおいて

$$\frac{\partial l_1(\theta)}{\partial \theta_i}, \frac{\partial^2 l_1(\theta)}{\partial \theta_i \partial \theta_j}, \frac{\partial^3 l_1(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k} \; 00$$

が存在する、(iii) フィッシャー情報量 $I(\boldsymbol{\theta}) = (I(\theta_i))$ が正則行列 (スカラーの場合は正値)、(iv) 可積分関数 $M(\boldsymbol{x})$ が存在して

$$\mathbf{E}[|\frac{\partial^3 l_1(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}|] < \mathbf{E}[M(\mathbf{x})] < \infty \tag{7.53}$$

である。

独立でない場合

最尤推定法が数理統計学で重視される一つの理由は尤度関数が想定できる状況では一般的に適用可能であることが挙げられる。例えば確率変数列 $X_1, X_2, \cdots, X_n$ が互いに独立とは限らない場合でも同時密度関数は初期条件 $X_0 = x_0$ を所与として条件付密度関数の積

$$L_n = f(x_n|x_{n-1}, \cdots, x_0, \theta) f(x_{n-1}|x_{n-2}, \cdots, x_0, \theta) \cdots f(x_1|x_0, \theta) \tag{7.54}$$

により表現することができる。ここで特に

$$f(x_n|x_{n-1}, x_{n-2}, \cdots, x_1, x_0, \theta) = f(x_n|x_{n-1}, x_0, \theta) \tag{7.55}$$

となるときマルコフ性があると云う。例えばデータ時系列を扱う場合には過去を所与とした現在の条件付分布が直前の過去値のみに依存する場合は独立標本の自然な拡張となるが、例えばマルチンゲールや自己回帰モデルなどの例は応用上で多用されている。初期値 x_0 を所与とすると対数尤度関数は

$$l_n(\theta) = \sum_{i=1}^{n} \log f(x_i|x_{i-1}, \theta) \tag{7.56}$$

と簡単化できるので多くの場合に最尤推定法の適用が可能となる。

例 7-16： 例として 1 次自己回帰モデル (AR(1) と書く)

$$X_j = aX_{j-1} + Z_j \ (j = 1, \cdots, n) \tag{7.57}$$

を考えよう。ここで Z_j は互いに独立な確率変数列 $Z_j \sim N(0, \sigma^2)$ であり、ゼロ時点 $j = 0$ における X_0 を所与とすると系列 Z_j $(j = 1, \cdots, n)$ および過去の X_{j-1} により各時点 j における確率変数 X_j が決まる統計モデルである。この場合は初期条件を所与とすると尤度関数は

$$L_n(\theta) = (\frac{1}{\sqrt{2\pi}})^{n/2} \exp[-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (X_j - aX_{j-1})^2] \tag{7.58}$$

となる。初期条件 X_0 が所与の下では最尤推定量は最小二乗推定量にほぼ一致し

$$\hat{a}_{LS} = \frac{\sum_{j=1}^{n} X_j X_{j-1}}{\sum_{j=1}^{n} X_{j-1}^2} \tag{7.59}$$

で与えられる。さらに

$$\sqrt{n}[\frac{\sum_{j=1}^{n} X_j X_{j-1}}{\sum_{j=1}^{n} X_{j-1}^2} - a] = \frac{\frac{1}{\sqrt{n}} \sum_{j=1}^{n} U_j X_{j-1}}{\frac{1}{n} \sum_{j=1}^{n} X_{j-1}^2} \tag{7.60}$$

と表現すると、この確率過程 X_j が定常過程であれば右辺の分母は $\mathbf{E}[X_1^2]$ に収束し、分子はマルチンゲールになるので中心極限定理より正規分布に収束する。したがってこの場合には漸近分布の分散はフィッシャー (Fisher) 情報量に一致する。

こうした分析対象の現象が時間の経過に依存し、変化していく時系列・確率過程の場合にも最尤推定法を適用できる。ただしこうした場合には小標本の推定理論が適用可能ではないので大標本の推定理論が必要となる場合が少なくないのである。

7.6 密度関数の推定問題

多くの統計的分析では確率分布が利用されているので標本データより確率分布を推定することは基本的な問題と考えられる。標本として独立な確率変数列の実現値が利用可能なとき確率分布が母数 θ に依存するとき $F(x|\theta)$、これに母数の推定量 $\hat{\theta}_n$ を代入すると $F(x|\hat{\theta}_n)$ により推定することができる。密度関数 $f(x|\theta)$ の推定についても同様に $f(x|\hat{\theta}_n)$ により推定することが考えられる。例えば正規分布の場合には標本平均 $\bar{x}_n$, 標本分散 $s_n^2 = (1/(n-1))\sum_{i=1}^n (x_i - \bar{x}_n)^2$ を利用して

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi s_n^2}} e^{\frac{1}{2s_n^2}(x-\bar{x}_n)^2} \tag{7.61}$$

とすることが考えられる。

実際の統計分析では確率分布の形が事前に正確に既知である状況は稀である。こうした場合には利用可能な統計データ $\mathbf{x} = (x_1, \cdots, x_n)'$ (n 次元 (縦) ベクトル) より確率分布を推定することが考えられる。こうした non-parametric(ノンパラメトリック) な分布関数の推定法としては経験分布関数 $F_n(x)$ を利用することが一般的である。任意の連続点 x に対して $n \to \infty$ のとき $F_n(x) \to F(x)$ となるのである種の妥当性があるが、有限の n に対しては階段関数になる。この関数は頻度分布に対応するが、確率分布としてなめらかな密度関数が妥当と考えられる状況も多い。そこでより滑らかな密度関数を推定する方法としてカーネル関数 $K(y)$ を利用して

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h} K(\frac{x - X_i}{h}) \tag{7.62}$$

とすることが考えられるが、この方法はカーネル推定法と呼ばれている。ここで h はバンド幅と呼ばれる正実数、でありカーネル関数 $K(y)$ としては条件 (i) $\int K(y)dy = 1$, (ii) $\int yK(y)dy = 0$, (iii) $\int y^2 K(y)dy = 1$ を満たす必要がある。例えば
(i) 標準正規分布の密度関数 $K_1(y) = (1/\sqrt{2\pi})\exp[(-1/2)y^2]$,
(ii)Epanechnikov 密度関数 $K_2(y) = (3/4\sqrt{5})(1 - y^2/5)$ $(|y| \le \sqrt{5})$

などがよく利用されている。こうしたカーネル推定量は密度関数の各点 x の周りの観測値をカーネル関数 $1/(nh)K((x-X_i)/h)$ を利用して平滑化 (smoothing) する推定法である。

密度関数は一般に連続点で定義されるが、統計データは有限個の実現値と見なすと、連続区間で定義されている密度関数の推定量の誤差評価には平均二乗積分誤差 (mean integrated squares error) を利用するのが一般的である。この誤差量は

$$\begin{aligned} MISE_f(\hat{f}) &= \int \mathbf{E}_f[\hat{f}_n(x) - f(x)]^2 dx \\ &= \int \mathbf{Var}_f[\hat{f}_n(x)]dx + \int [\mathbf{E}_f(\hat{f}_n(x)) - f(x)]^2 dx \end{aligned}$$

により定められる。このとき推定のバイアスは $(x-t)/h = y$ と置くと

$$\begin{aligned} \mathbf{E}_f[\hat{f}_n(x) - f(x)] &= \int [\frac{1}{h}K(\frac{x-t}{h})f(t)dt] - f(x) \\ &= \int K(y)[f(x-hy) - f(x)]dy \end{aligned}$$

となる。ここで推定量の評価では密度関数が変数 x の変化に伴い滑らかに変化すること (smooth な条件と呼ばれる) が仮定できればより正確な評価が可能となる。例えば関数の二回微分性を仮定すると $f(x-hy) - f(x) \sim (-hy)f^{'}(x) + [(-hy)^2/2]f^{''}(x)$ であるからバイアス項は

$$\int \mathbf{E}_f[\hat{f}_n(x) - f(x)] \sim \int y^2 K(y)dy \left[\frac{f^{''}(x)}{2}\right] h^2 + o(h^2)$$

となり、漸近分散項は $\mathbf{Var}(\hat{f}_n) = \mathbf{E}(\hat{f}_n^2) - [\mathbf{E}(\hat{f}_n)]^2$ であるが、この場合には第二項は相対的に小さいと評価できるので結局

$$\mathbf{Var}[\hat{f}_n(x)] \sim \frac{1}{nh}\int K(y)^2 f(x-hy)dy + o(\frac{1}{nh})$$

となる。したがって、バイアス項と漸近分散を合わせて評価すると、ある実定数 c_f が存在して

$$MISE_f(\hat{f}) = \int \mathbf{E}_f[\hat{f}_n(x) - f(x)]^2 dx \leq c_f[\frac{1}{nh} + h^4] \tag{7.63}$$

となる。この量を h について最小化する定数項を除いた n の次数の意味での最適な選択としては $h_n = n^{-1/5}$, $MISE_f(\hat{f}_n) = O(n^{-4/5})$ と云う条件が得られる。このように通常のパラメトリック・モデルの推定と異なりノンパラメトリック・モデルにおける誤差評価はより複雑となる。

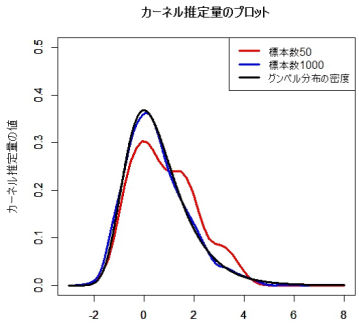

図 7-1: カーネル密度推定の例

例 7-17： 例として第一種の極値分布であるガンベル (Gumbel) 分布からランダムにデータを発生させ、正規カーネル関数を利用してバンド幅を最適にとり密度関数を推定した例を挙げておく。ここではガンベル分布にしたがう確率変数 X とすると $X = -\log Y$ となる Y は指数分布 $\mathrm{EX}(1)$ にしたがい、さらに $Y = -\log U$ として U は一様分布 $U(0,1)$ にしたがうことを利用すると、計算機上で一様乱数よりガンベル乱数を発生することができる。図 7-1 が示すようにこの場合にはデータ数があまり少なくなければ真の密度関数についての妥当な推定結果が得られる。

8　統計的検定論

8.1 仮説検定の発想

一回のコイン投げにおいて表の出る確率を p、結果に歪みがあるか否かについての実験を例に仮説検定問題を考察しよう。「公平なコイン投げ」とはベルヌイ試行についての仮説

$$H_0 : p = \frac{1}{2} \tag{8.1}$$

と解釈し、統計家は真の状態を知らないが実験を多数回できないので (大数の法則から真の確率を推定できるが) 少数回のコイン投げの観測値よりこの仮説の妥当性を検出したいとする。仮にコイン投げ 30 回の中の 20 回が表すると確率の点推定値は $\hat{p} = \frac{2}{3}$ となる。この値は仮説 $p = \frac{1}{2}$ から離れているので仮説 H_0 を捨てることは仮説「コインが歪んでいる」

$$H_1 : p > \frac{1}{2} \tag{8.2}$$

を受け入れると解釈できる。ここで仮説 H_0、あるいは仮説 H_1 を受け入れると言う統計的決定問題では第一の仮説では特定の確率値 $1/2$ に対し第二の仮説は仮説が成り立たないというより消極的主張なので仮説 H_0 を**帰無 (null) 仮説**，二番目の仮説を**対立 (alternative) 仮説**と呼ぶ。

30 回の中で 20 回表が出る確率は、表の出る回数 $X \sim B(n,p)$(二項分布) にしたがうので

$$ {}_{30}C_{20} \left(\frac{1}{2}\right)^{20} \left(1 - \frac{1}{2}\right)^{30-20} \cong .02798 \tag{8.3}$$

となる。同様に実験で 21 回表が出ることを考慮すると、同様に 22 回、23 回表がでる場合も考慮する必要があるであろう。ここで 20 回以上表がでる確率は

$$P = \sum_{i=20}^{30} \left(\frac{1}{2}\right)^{i} \left(1 - \frac{1}{2}\right)^{30-i}$$

となるが、二項分布を正規分布で近似すると $X \sim N(30 \times \frac{1}{2}, 30 \times \frac{1}{2} \times \frac{1}{2})$ より

$$P(X \geq 20) = P(\frac{X - 15}{\sqrt{30\left(\frac{1}{2}\right)^2}} \geq 1.826) \sim 0.034$$

となる。これらの確率は**確率値** ($P-$ 値) と呼ばれる。

この $P-$ 値は小さいので帰無仮説 H_0 を棄却して対立仮説 H_1 を受け入れるべきだろうか？帰無仮説が正しい場合に実際に観測された回数よりも多い回数が実現する確率が 3 %程度なのでこの数値は無視できないとする見方がある。他方、仮に帰無仮説が正しいとすると表の出る回数が現実に観測された 20 回を越える確率はほぼ 3 %と小さくもともとの帰無仮説は現実的ではないとの見方もある。論理的にはどちらの主張も正しく小さな値でも確率値が正の値である限り、たまたま 20 回表が出るという可能性を否定できない。しかし実際に実験する回数が限られている場合、(多くの) 統計家は後者の見方、統計的仮説検定 (hypothesis testing) の考え方を利用している。第 2 の観点に立つとき、どの程度の確率値により帰無仮説を否定して対立仮説を受け入れるべきかが統計的問題であるが、小さいと判断する基準値のことを**有意水準** (significance level)、危険率と呼ぶ。その具体的な数値は統計家や実証分析者の主観に依存するが、伝統的には 1 %,5 %,10 %などの数値を使うのが一般的である (有意水準はギリシャ文字の α(アルファ) で表記する) が、実験データから得られた確率値がこの α(アルファ) よりも小さければ帰無仮説 H_0 を棄却することになる。実験データから得られた確率値がこの α(アルファ) よりも大きければ、証拠が不十分であるとして帰無仮説 H_0 を棄却できないが、帰無仮説を受容すると云う。コインの数値例では確率値が 1 %と 5 %の間だったが有意水準 5 %では帰無仮説は棄却され、1 %では帰無仮説は棄却されずに受容される。このように、データから計算された確率値さえわかっていれば色々な有意水準で検定結果を求めることができるので確率値は統計データの有用な情報を含んでいる。

統計家の多くは帰無仮説は棄却されることに意味があるとする傾向がある。帰無仮説が受容される場合にはそのことは必ずしも仮説が”正しい”ことを積極的に主張してはいないと判断するからである。帰無仮説を否定するにはデータが不足していることもあり、より多くの標本により帰無仮説が棄却されることもある。他方、実証分析者は理論的な仮説を統計モデルの上で表現するとき、帰無仮説の検定で理論の検証を行うことがある。この場合には分析者は理論的な仮説がデータから棄却されずに受容されることを重視する傾向が生じる。

仮説検定の方法

母集団の確率分布の下で母数 (スカラー) を θ としよう。母数 θ がとる領域を大文字 Θ で表して母数空間 (parameter space)、帰無仮説はこの母数空間の一部分 Θ_0 として

$$H_0 : \theta \in \Theta_0 \tag{8.5}$$

で表す。対立仮説は母数空間 Θ の別の部分 Θ_1 により

$$H_1 : \theta \in \Theta_1 \tag{8.6}$$

で表す。通常は $\Theta = \Theta_0 \cup \Theta_1$ である。

標本を確率変数 (縦) ベクトル $\mathbf{X} = (X_1, \cdots, X_n)'$ とすると標本からの仮説検定問題は標本がどのような領域に入れば仮説を棄却すべきかの問題に帰着できる。この領域を棄却域 (rejection region) と呼び n に依存するので R_n で表そう。この領域 R_n が条件

$$P(\mathbf{X} \in \mathrm{R}_n|\theta) = \alpha, \quad \theta \in \Theta_0 \tag{8.7}$$

で決めれば、左辺の確率は帰無仮説が正しいときに標本が棄却域に入る確率を表す。このときには仮説が正しいにも関わらず帰無仮説を棄却するので誤った決定を下している。この誤りを犯す確率を第 1 種の過誤と呼ぶ。実数 $0 < \alpha < 1$ をあらかじめ指定してこの誤りをコントロールするが、伝統的にはこの値を 1 %,5 %,10 %などの値をとる。この第一種の過誤 100α %を固定して標本 $\mathbf{X} = (X_1, \cdots, X_n)$ から棄却域を決めてれば検定方式が定まる。

ここで統計的検定では第一種の過誤ばかりではない誤りがある。帰無仮説が正しくないときに仮説を受容する確率は

$$P(\mathbf{X} \in \mathrm{R}_n^c|\theta), \quad \theta \in \Theta_1 \tag{8.8}$$

と表せるので、これを第 2 種の過誤と呼ぶ。(ここで領域 R_n^c は領域 R_n の補集合を示す。) この第 2 種の過誤も誤りなのでこの確率を小さくすることが望ましい。このことは確率

$$\beta(\theta) = P(\mathbf{X} \in \mathrm{R}_n|\theta), \quad \theta \in \Theta_1 \tag{8.9}$$

大きくすることと同等である。この確率は対立仮説が正しいときに帰無仮説を棄却して対立仮説を受容する確率は検定の**検出力** (power) と呼ばれる。ここで左辺の $\beta(\theta)$ (ギリシャ文字のベータ) を未知母数 θ を変数とする。一般には検出力は未知母数に依存するので母数 θ の真の値に無関係に小さくすることはできない。

8.2 検定の標準理論

標本 $\mathbf{X} \in \mathrm{R}$ のとき関数 $\phi(\mathbf{X}) = 1$, 標本 $\mathbf{X} \in \mathrm{R}^c$ のとき関数 $\phi(\mathbf{X}) = 0$ により検定関数を定めよう。一般に離散分布の場合も考慮すれば棄却域 R_n の代わりに標本 $X = (X_1, \cdots, X_n)'$ に対して検定関数

$$\phi(X) = \phi(X_1, \cdots, X_n) = \begin{cases} 1 & X \in \mathrm{R}_{1n} \\ \gamma & X \in \mathrm{R}_{2n} \\ 0 & X \in \mathrm{R}_{3n} = (\mathrm{R}_{1n} \cup \mathrm{R}_{2n})^c \end{cases} \tag{8.10}$$

を導入しよう。ここで $\phi(\mathbf{X}) = 1$ は帰無仮説を棄却を意味するので領域 R_{1n} は棄却域、$\phi(X) = 0$ は帰無仮説を受容するので領域 $(\mathrm{R}_{1n} \cup \mathrm{R}_{2n})^c$ は受容域である。さらに $\phi(\mathbf{X}) = \gamma, 0 < \gamma < 1$ は確率 γ で帰無仮説を棄却することを意味する。この領域 R_{2n} は連続型分布の場合には必要ないが、離散型の場合に第 1 種の過誤を (先験的に固定する) ある水準 α にする為には必要である。例えば二項分布の例では確率関数はとびとびの値をとるが、領域 R_{2n} を使うと正確に第 1 種の過誤を固定できる。ここで 2 種類の過誤の確率は

$$P(\text{第 1 種の過誤}) = \mathbf{E}[\phi(\mathbf{X})|\theta_0] \tag{8.11}$$

および

$$P(\text{第 2 種の過誤}) = \mathbf{E}[1 - \phi(\mathbf{X})|\theta_1] = 1 - \mathbf{E}[\phi(\mathbf{X})|\theta_1] \tag{8.12}$$

である。検出力 (power function) は

$$\beta(\theta_1) = 1 - P(\text{第 2 種の過誤}) = \mathbf{E}[\phi(\mathbf{X})|\theta_1] \tag{8.13}$$

によって定義される。このとき第一種の過誤 $\alpha_1(\phi) = \mathbf{E}[\phi(\mathbf{X})|\theta_0]$, および第二種の過誤 $\alpha_2(\phi) = \mathbf{E}[1-\phi(\mathbf{X})|\theta_1]$ をまとめて $\alpha(\phi) = (\alpha_1(\phi), \alpha_2(\phi))$ としよう。$[0,1] \times [0,1]$ 内のこの二次元集合を L 集合と呼ぶと、第一種の過誤を 0 にとると第二種の過誤は 1 になる。また第一種の過誤を 1 にとると第二種の過誤は 1 になり、第一種の過誤と第二種の過誤を $1/2$ にすることができる。二つの過誤を $(0,1)$ にとり検定 ϕ_1, ϕ_2 が構成できるとすると二つの検定の線形結合 $\gamma_1\phi_1 + \gamma_2\phi_2$ $(\gamma_1 + \gamma_2 = 1, \gamma_1 \geq 0, \gamma_2 \geq 0)$ も検定方式になるので次のことが導かれる。

定理 8-1 ： 任意の検定関数 ϕ に対して第一種の過誤と第二種の過誤の集合 $L = \{\alpha(\phi)|(\alpha_1(\phi), \alpha_2(\phi))\}$ は次の性質を持つ。
(i) $(1,0) \in L$,$(0,1) \in L$ である。
(ii) L は点 $(1/2, 1/2)$ に対称であり $\alpha(\phi) \in L$ なら $\alpha(1-\phi) \in L$。
(iii) L は閉凸集合である。

もし 2 つの検定方式があってそれぞれの棄却域が R_{1n},R_{2n} で与えられそれぞれの検出力を $\beta_1(\theta)$,$\beta_2(\theta)$ とするとき対立仮説 $\theta \in \Theta_1$ がどのような値をとっても不等式

$$\beta_1(\theta) \geq \beta_2(\theta) \tag{8.14}$$

が成り立てば検定方式 1 は検定方式 2 よりも悪くないと考えられる。もし厳密な不等号がどこかで成立していれば常に検定方式 1 は検定方式 2 よりも良い方式となる。そこで一般に任意の検定方式に ϕ 対し任意の θ の値について

$$\beta^*(\theta) \geq \beta_\phi(\theta) \tag{8.15}$$

が成り立つよう検出力 β^* を持つ検定方式があれば一様最強力検定 (unifomly most powerful test, 略して UMP 検定) と呼ばれる。

ネイマン・ピアソンの補題

統計モデルにおける検定問題の帰無仮説が単純であれば UMP 検定が構成できる。ここで単純 (simple) 仮説は帰無仮説が母数の一点 $\theta = \theta_0$ となる場合である。これに対して帰無仮設が 1 点で表現できない仮説を複合仮説 (composite hypothesis) と呼ぶ。コイン投げの例では確率 p について帰無仮説 $H_0 : p = 1/2$ は単純仮説の例である。

ここで棄却域 R_n の代わりに標本 $\mathbf{X} = (X_1, \cdots, X_n)'$ に対して検定関数を

$$\phi(\mathbf{X}) = \phi(X_1, \cdots, X_n) = \begin{cases} 1 & X \in \mathrm{R}_{1n} \\ \gamma & X \in \mathrm{R}_{2n} \\ 0 & X \in \mathrm{R}_{3n} = (\mathrm{R}_{1n} \cup \mathrm{R}_{2n})^c \end{cases}$$

としよう。ここで $\phi(\mathbf{X}) = 1$ は帰無仮説を棄却を意味するので領域 R_{1n} は棄却域、$\phi(\mathbf{X}) = 0$ は帰無仮説を受容することを意味するので領域 $(\mathrm{R}_{1n} \cup \mathrm{R}_{2n})^c$ は受容域である。これに対して $\phi(\mathbf{X}) = \gamma, 0 < \gamma < 1$ は確率 γ で帰無仮説を棄却することを意味しています。この領域 R_{2n} は連続型分布の場合には必要ないが、離散型の場合に第 1 種の過誤を先験的に固定された水準の α にする為に必要であり、二項分布の例のように領域 R_{2n} により正確に第 1 種の過誤を固定できる。ここで 2 種類の過誤の確率は P(第 1 種の過誤) $= \mathbf{E}[\phi(\mathbf{X})|\theta_0]$ および P(第 2 種の過誤) $= \mathbf{E}[1 - \phi(\mathbf{X})|\theta_1] = 1 - \mathbf{E}[\phi(\mathbf{X})|\theta_1]$ となり、検出力は

$$\beta(\theta_1) = 1 - P(\text{第 2 種の過誤}) = \mathbf{E}[\phi(\mathbf{X})|\theta_1]$$

である。ネイマン (J.Neyman) とピアソン (K.Pearson) は次のように UMP 検定を求められることを示した。

定理 8-2 (ネイマン・ピアソン (Neyman-Peason) の基本補題) : $X_1, \cdots, X_n$ が密度関数 $f(x|\theta)$ となる母集団分布からの n 個のランダム標本、二つの単純仮説において帰無仮説 $H_0 : \theta = \theta_0$, 対立仮説 $H_1 : \theta = \theta_1$ とする。このとき正定数 c を定めて検定関数を $\mathbf{E}[\phi^*|\theta_0] = \alpha$ かつ

$$\phi^* = \begin{cases} 1 & \left(\prod_{i=1}^n f(x_i|\theta_1) > c \prod_{i=1}^n f(x_i|\theta_0) \right) \\ 0 & \left(\prod_{i=1}^n f(x_i|\theta_1) < c \prod_{i=1}^n f(x_i|\theta_0) \right) \end{cases} \tag{8.16}$$

とできれば、任意の検定関数 $\phi(\cdot)$ に対して

$$\mathbf{E}[\phi^*(X)|\theta_1] \geq \mathbf{E}[\phi(X)|\theta_1] \tag{8.17}$$

となる。

証明： 変数ベクトル $\mathbf{x} = (x_1, \cdots, x_n)'$ に対して同時密度関数を $f(\mathbf{x}|\,\theta) = \prod_{i=1}^{n} f(x_i|\,\theta)$ 検定関数

$$\phi(\mathbf{x}) = \begin{cases} 1 & (\text{もし})\mathbf{x} \in \mathrm{R}_n \quad (\text{棄却域}) \\ 0 & \qquad\qquad (\text{その他}) \end{cases}$$

とすると、検出力 (power) は

$$\begin{aligned} \beta(\theta_1) &= 1 - P(\text{第 2 種の過誤}) \\ &= 1 - \mathbf{E}[1 - \phi(\mathbf{x})\,|\,\theta_1] = \mathbf{E}[\phi(\mathbf{x})\,|\,\theta_1] \end{aligned}$$

で与えられる。ここでラグランジュ乗数 λ、ラグランジュ形式を

$$L(\theta_1) = \mathbf{E}[\phi(\mathbf{x})\,|\,\theta_1] - \lambda\,\mathbf{E}[\phi(\mathbf{x})\,|\,\theta_0] \tag{8.18}$$

とおくと、問題は制約条件

$$\mathbf{E}[\phi(\mathbf{x})\,|\,\theta_0] = \alpha \tag{8.19}$$

の下での検出力の最大化はラグランジュ形式

$$L(\theta_1) = \int \phi(\mathbf{x})\,[f(\mathbf{x}|\,\theta_1) - \lambda\, f(\mathbf{x}|\,\theta_0)]\,d\mathbf{x}$$

を最大化する問題となるここで検定関数を

$$\begin{aligned} \phi^*(\mathbf{x}) &= 1 \qquad \text{if } f(\mathbf{x}|\,\theta_1) - \lambda\, f(\mathbf{x}|\,\theta_0) > 0 \\ \phi^*(\mathbf{x}) &= 0 \qquad \text{if } f(\mathbf{x}|\,\theta_1) - \lambda\, f(\mathbf{x}|\,\theta_0) < 0 \end{aligned}$$

と置けば検定の検出力 (power) は最大化される。すなわち尤度比に基づく検定方式が求める解となる。 **Q.E.D.**

ここでは密度関数が存在する場合を議論したが、その他の場合（離散分布の場合や連続・離散の混合分布の場合）も検定方式をわずかに修正することにより結果は成立する。離散分布の場合には制約条件としての (第一種の過誤) を任意の水準とする為に同時確率関数 $p(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$ に対して検定関数として

$$\phi^* = \begin{cases} 1 & \left(\prod_{i=1}^{n} p(x_i|\theta_1) > c \prod_{i=1}^{n} p(x_i|\theta_0)\right) \\ \gamma & \left(\prod_{i=1}^{n} p(x_i|\theta_1) = c \prod_{i=1}^{n} p(x_i|\theta_0)\right) \\ 0 & \left(\prod_{i=1}^{n} p(x_i|\theta_1) < c \prod_{i=1}^{n} p(x_i|\theta_0)\right) \end{cases} \tag{8.20}$$

という検定関数を利用すれば、任意の $1 > \alpha > 0$ に対して P(第 1 種の過誤) $= \mathbf{E}[\phi(\mathbf{X})|\theta_0] = \alpha$ とできるのである。

例 8-1：分散が既知の正規母集団 $N(\mu, \sigma^2)$ から n 個の独立な標本 $X_1, \cdots, X_n$ が得られ、帰無仮説 $H_0 : \mu = \mu_0$ を対立仮説 $H_1 : \mu > \mu_0$ に対して検定することを考えよう。尤度関数は

$$L_n(\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\theta)^2}$$

で与えられる。仮説 H_0 と仮説 H_1 の下で尤度の比率を計算すれば

$$\begin{aligned} \frac{L_n(\mu_1)}{L_n(\mu_0)} &= \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu_0)^2}} \\ &= e^{\frac{1}{2\sigma^2}\left[2(\mu_1-\mu_0)\sum_{i=1}^n x_i + n\left(\mu_0^2-\mu_1^2\right)\right]} \end{aligned} \tag{8.21}$$

である。したがってネイマン・ピアソンの補題を用いると棄却域はこの尤度比 (likelihood ratio) が一定の数よりも大きい領域として与えられる。この尤度比は標本の関数としてみると $\sum_{i=1}^n x_i$ の関数なので棄却域は標本平均 $\bar{X}_n$ とある定数 c' を使って

$$(\mu_1 - \mu_0)\bar{X}_n \geq c'$$

と表現できる。したがって帰無仮説と対立仮説における母数について $\mu_1 > \mu_0$ が成り立てば最強力検定の棄却域はある定数 c を使って

$$\bar{X}_n \geq c$$

で与えられる。ここで定数 c は第 1 種の過誤が α となるように決める必要があるがので

$$P\left(\bar{X}_n \geq c|\mu = \mu_0\right) = P\left(\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma} \geq \frac{\sqrt{n}(c-\mu_0)}{\sigma}|\mu = \mu_0\right) = \alpha \tag{8.22}$$

と設定するが、左辺は帰無仮説の下で標準正規分布にしたがうので右側 100α パーセント点を $n(\alpha)$ とすれば

$$\frac{\sqrt{n}(c-\mu_0)}{\sigma} = n(\alpha)$$

を解くことにより定数 c を決めればよい。対立仮説が $H_1 : \mu_1 < \mu_0$ のときも同様な議論により最強力検定の棄却域は

$$\frac{\sqrt{n}(c-\mu_0)}{\sigma} = -n(\alpha)$$

により求められる。ここで棄却域は対立仮説の値には依存していないことに注意すると、帰無仮説と対立仮説における母数の関係が $\mu_1 > \mu_0$，あるいは $\mu_1 < \mu_0$ とな

る場合には上の棄却域による検定は一様最強力検定 (Uniformly Most Powerful Test, 略して UMP 検定) となる。

しかしながら対立仮説が $H_1 : \mu_1 \neq \mu_0$ となっている場合にはこの議論が成り立たない。仮に成り立つとするとそれぞれの片側検定と一致するはずであり矛盾が生じるので UMP 検定方式を構成するのができない。

不偏検定

例 8-1 で見たように単純な帰無仮説 $\mathcal{H}_0 : \theta = \theta_0$, 対立仮説 $\mathcal{H}_1 : \theta \neq \theta_0$ の場合には UMP 検定が存在しない。単純な帰無仮説 $\mathcal{H}_0 : \theta = \theta_0$ に対して対立仮説 $\mathcal{H}_1 : \theta > \theta_0$ の UMP 検定は実際に $\theta_1 < \theta_0$ であるときには検出力が α(第一種の過誤) よりも小さくなるので不合理な検定方式である。そこで検定方式として検出力が α よりも小さくならない不偏検定 (unbiased test) に限定することが考えられる。この条件は

$$\beta_\phi(\theta) \geq \alpha \ \ (\theta \neq \theta_0) \tag{8.23}$$

および

$$\beta_\phi(\theta_0) \leq \alpha \ \ (\theta = \theta_0) \tag{8.24}$$

と表現できる。検出力が母数 θ について微分可能であれば

$$\frac{\partial}{\partial\theta}\beta_\phi(\theta)|_{\theta=\theta_0} = 0 \tag{8.25}$$

を意味する。したがって、定理の証明を考慮すると制約条件付き最大化問題の制約条件が一つ増えたことを意味する。NP 補題の棄却域は適当に $\mathbf{x} = (x_1, \cdots, x_n)'$, 定数 c_1, c_2 として

$$f(\mathbf{x}|\theta_1) > c_1 f(\mathbf{x}|\theta_0) + c_2 \frac{\partial}{\partial\theta} f(\mathbf{x}|\theta_0) \tag{8.26}$$

と表現できる。この検定方式は UMPU(uniformly most powerful unbiased) 検定と呼ばれている。

UMPU 検定が簡単に構成できる例としては分散既知の正規分布の期待値の検定問題が知られている。例えば例 8-1 において $\mu_0 = 0$ とすれば UMPU 検定の棄却域はある定数 c を選び

$$|\bar{X}_n| > c$$

となり、直観的に妥当な検定方式が得られる。この場合には分散が未知であっても UMPU 検定を構成でき、t 検定に一致する。

二標本検定

母集団分布が二項分布や正規分布など基礎的分布の場合の平均や分散に関する仮説

検定問題については多くの統計学の教科書が説明している。ここでは二標本問題についてのみ言及しておく。

例 8-2 : 分散が未知で互いに独立な二つの正規母集団 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_1, \sigma_2^2)$ から n_1, n_2 個の独立な標本 $X_i\ (i = 1, \cdots, n_1), Y_i\ i = 1, \cdots, n_2)$ が得られるとき帰無仮説 $H_0 : \mu_1 = \mu_2$ を対立仮説 $H_1 : \mu_1 \neq \mu_2$ に対して検定を考える。標本平均 $\bar{X}_n, \bar{Y}_n$、標本不偏分散 s_1^2, s_2^2 とすると分散が等しい場合には標本平均の差 $\bar{X}_{n_1} - \bar{Y}_{n_2}$、プールしたデータより分散は $s^2 = (1/(n_1 + n_2 - 2))[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]$ により推定できる。したがって t 統計量は $t = (\bar{X}_{n_1} - \bar{Y}_{n_2})/[s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}]$ で与えられる。実際に二標本において分散が不均一な場合はベーレンス・フィッシャー (Behrens-Fisher) 問題と呼ばれているが、この統計量は帰無仮説の下で t 分布にしたがわない。そこでよりよく利用されるのがウエルチ (Welch) 統計量

$$t_W = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{8.27}$$

であり帰無分布は自由度

$$\nu = \frac{[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}]^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

の t 分布で近似できる。ウエルチ検定では棄却域を両側にそれぞれ有意水準 $\alpha/2$ をとれば $|t_W| > t_\nu(\alpha/2)$ の時に帰無仮説を棄却することでほぼ有意水準 α で検定が可能である。この方法は良い近似にもとづくが自由度は $n_1 - 1 \leq \nu \leq n_1 + n_2 - 2$ を満足している。

尤度比検定

仮説検定の様々な問題については NP 補題は重要な一般的原理を示唆している。大きさ n のランダムな標本に対して確率密度関数を $f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)\ (\mathbf{x} = (x_1, \cdots, x_n)')$ としよう。このとき棄却域の構成は帰無仮説と対立仮説の下での尤度関数の比、尤度比 (likelihood ratio) の不等式 [33]

$$LR_n(\theta_1, \theta_0) = \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} > c \tag{8.28}$$

に基づいている $(\mathbf{x} = (x_1, \cdots, x_n))$。対立仮説が正しければ左辺は大きくなると期待できるので尤度比は自然な検定方式を導くと考えられる。

帰無仮説と対立仮説が共に複合仮説の場合には上の検定方式をより一般化して分子と分母をそれぞれ母数空間 $\boldsymbol{\Theta}$ 上と帰無仮説の空間 $\boldsymbol{\Theta}_0$ 上で最大化した比の大きさ

$$LR_n = \frac{\max_{\theta \in \Theta} f(\mathbf{x}|\theta)}{\max_{\theta \in \Theta_0} f(\mathbf{x}|\theta)} > c \tag{8.29}$$

[33] 文献により分母・分子を逆に定義することがあるが、その場合には棄却域は zero 方向になる。

により棄却域を設定する検定方式を尤度比検定 (likelihood ratio) と呼ぶ。

例 8-3：再び例 8-1 の設定を用いて帰無仮説 $H_0 : \mu = \mu_0$, 対立仮説 $H_1 : \mu \neq \mu_0$ に対する検定を考える。帰無仮説の下で尤度関数を

$$\begin{aligned} L_n(\mu_0) &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}[\sum_{i=1}^n (x_i-\bar{x}_n)^2 + n(\bar{x}_n-\mu_0)^2]} \end{aligned}$$

と表現する。対立仮説の下で尤度関数を最大化すると、

$$L_n(\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x})^2}$$

で与えられる。したがって仮説 H_0 と仮説 H_1 の下で尤度比を計算すれば

$$LR_n = e^{\frac{1}{2\sigma^2} n\left(\bar{X}_n-\mu_0\right)^2}$$

となる。帰無仮説の下で $\sqrt{n}(\bar{X}_n - \mu_0) \sim N(0,1)$ となるので

$$2\log LR_n \sim \chi^2(1) \tag{8.30}$$

である ($\chi^2(1)$ は自由度 1 の χ^2- 分布である)。

さらにこの例から母集団分布が正規分布にしたがわない場合にも中心極限定理より n が大きいときに漸近的に $\sqrt{n}(\bar{X}_n - \mu_0) \sim N(0,1)$ にしたがうので帰無仮説の下で漸近的に $2\log LR \sim \chi^2(1)$ にしたがうことが分かる。この自由度は帰無仮説の動きうる範囲 (1 次元) に対応している。

母数や確率分布が多次元の場合も同様に尤度比検定を構成できる。多次元分布の例として $\mathbf{X}_i$ が p 次元正規分布にしたがう例を挙げておく。

例 8-4 : 正規母集団 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ から n 個の独立な標本 $\mathbf{X}_1, \cdots, \mathbf{X}_n$ が得られるとき、帰無仮説 $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ を対立仮説 $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ に対する検定を考える。尤度関数を

$$L_n(\mu) = \frac{1}{[(2\pi)^{p/2}|\boldsymbol{\Sigma}|]^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})} \tag{8.31}$$

と表現する。帰無仮説の下で母数 $\boldsymbol{\Sigma}$ について補題 7-2 を用いて最大化すると

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)'$$

となる。対立仮説の下で尤度関数を母数 $\boldsymbol{\Sigma}$ について最大化すると

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i - \bar{\mathbf{X}}_n)(\boldsymbol{X}_i - \bar{\mathbf{X}}_n)' = \frac{1}{n}\mathbf{A}_n$$

となる。そこで仮説 H_0 と仮説 H_1 の下で尤度比を計算すれば

$$LR_n = \left[\frac{|\hat{\mathbf{\Sigma}}_0|}{|\hat{\mathbf{\Sigma}}_1|}\right]^{\frac{n}{2}}$$

となる。この式を関係 $\boldsymbol{X}_i - \boldsymbol{\mu}_0 = (\boldsymbol{X}_i - \bar{\mathbf{X}}_n) + (\bar{\mathbf{X}}_n - \boldsymbol{\mu}_0)$ を利用して変形すると

$$[LR_n]^{\frac{2}{n}} = 1 + n(\bar{\mathbf{X}}_n - \boldsymbol{\mu}_0)' \mathbf{A}_n^{-1} (\bar{\mathbf{X}}_n - \boldsymbol{\mu}_0) \tag{8.32}$$

となる。ここで $p \times p$ 正則行列 $\mathbf{A}$,$p \times 1$ ベクトル $\mathbf{a}$ に対して $|\mathbf{A} + \mathbf{a}\mathbf{a}'| = |\mathbf{A}|[1 + \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}|$ が成り立つ (補題 3.2 の証明の類似の議論による) ことから $(1/n)\mathbf{A}_n \xrightarrow{p} \mathbf{\Sigma}$ および $\log[1+x] \sim x$ となる。したがって帰無仮説の下で漸近的に $2\ \log LR_n \sim \chi^2(p)$ となる。

例 $\mathbf{8-4}'$: 正規母集団 $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ から n 個の独立な標本 $\mathbf{X}_1, \cdots, \mathbf{X}_n$ が得られるとき、母数ベクトル $\boldsymbol{\mu}$ の一部分 r 次元ベクトル $\boldsymbol{\mu}_1$ $(1 \leq r \leq p)$ に対する帰無仮説 $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 (r \times 1)$ を対立仮説 $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_0$ に対して検定することが考えられる。例 8-3 の議論と同様に帰無仮説の下で漸近的に

$$2\ \log LR_n \sim \chi^2(r) \tag{8.33}$$

となるが、自由度 r は帰無仮説上で自由に動ける次元に対応する。

標準的なパラメトリック統計モデルの状況では一定の仮定の下で標本数が多いとき $(n \to \infty)$ 2 倍の対数尤度統計量 $2\log LR$ は $\mathbf{H}_0$ の下で漸近的に制約条件の数を自由度とする χ^2- 分布に漸近的にしたがう。このことを示すには尤度比統計量が真の母数の周りで近似的にどの様な挙動を示すかを分析する必要がある。例えば仮説が p 次元母数ベクトル $\boldsymbol{\theta} = (\theta_i)$ の一部分 $r\ < p)$ 次元母数ベクトル $\boldsymbol{\theta}_1 = (\theta_i)$ を用いて

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$$

としよう。制約条件を何も考慮しない対立仮説の下で、対数尤度関数 $l_n(\boldsymbol{\theta})$ を最尤推定量 $\hat{\boldsymbol{\theta}}_{ML} = (\hat{\theta}_i)$ の周りでテーラー (Talor) 展開して近似すると

$$\begin{aligned} l_n(\boldsymbol{\theta}) &\sim l_n(\hat{\theta}_{ML}) + \frac{1}{2}(\hat{\theta}_{ML} - \boldsymbol{\theta})' \left[(-\frac{\partial^2 l_n(\theta)}{\partial\theta_i \partial\theta_j})\right](\hat{\theta}_{ML} - \boldsymbol{\theta}) + o_p(1) \\ = \ & l_n(\hat{\theta}_{ML}) + \frac{1}{2}\sqrt{n}(\hat{\theta}_{ML} - \boldsymbol{\theta})' \left[\mathbf{I}_1(\boldsymbol{\theta})\right] \sqrt{n}(\hat{\theta}_{ML} - \boldsymbol{\theta}) + o_p(1) \end{aligned}$$

となる。ここで $\left[(-\frac{\partial^2 l_n(\theta)}{\partial\theta_i \partial\theta_j})\right]$ は $p \times p$ 行列, $p \times p$Fisher 情報行列

$$\mathbf{I}_1(\boldsymbol{\theta}) = \mathbf{E}\left[-\frac{1}{n}\frac{\partial^2 l_n(\theta)}{\partial\theta_i \partial\theta_j}\right]$$

をそれぞれ意味する。最尤推定の場合には 1 次微分項はゼロであるので二次の項まで考慮することが重要な点である。同様に帰無仮説の下で尤度関数を制約条件の下での最尤推定量 $\hat{\boldsymbol{\theta}}_{1.ML} = (\hat{\theta}_{1i})$ の周りでテーラー展開して近似すると

$$l_n(\boldsymbol{\theta}) \sim l_n(\hat{\theta}_{ML}) + \frac{1}{2}\sqrt{n}(\hat{\theta}_{1.ML} - \boldsymbol{\theta}_{10})' \left[-\frac{1}{n}\frac{\partial^2 l_n(\theta_1)}{\partial\theta_i \partial\theta_j} \right] \sqrt{n}(\hat{\theta}_{1.ML} - \boldsymbol{\theta}_{10}) + o_p(1)$$

となるが、$\left[-\frac{1}{n}\frac{\partial^2 l_n(\theta_1)}{\partial\theta_i \partial\theta_j} \right]$ は $r \times r$ 行列である。対数尤度比はそれどれの展開項の左辺と右辺第一項の差であるから補題 3-2 を Fisher 行列 $\mathbf{I}_1(\boldsymbol{\theta})$, 部分 Fisher 行列 $\mathbf{I}_{1,11}(\boldsymbol{\theta})(\mathbf{I}_1(\boldsymbol{\theta})$ の左上部分行列) に対して適用すると、二つの仮説の下で漸近的には最尤推定量の二次形式 (すなわち二次関数) で近似できる。ここで必要な議論は最尤推定量が漸近的に正規分布にしたがう条件にほぼ対応しているが、帰無仮説の下での展開 (制約条件のある場合)、対立仮説の下での展開 (つまり制約条件のない場合) を比較した尤度比統計量の漸近的挙動について次のようにまとめられる。

定理 8-3： (尤度比検定統計量の漸近的性質) 互いに独立な標本 $X_1, \cdots, X_n$ が同一の分布にしたがい密度関数が $f(\mathbf{x}|\boldsymbol{\theta})$ で与えられるとき，一定の正則条件 (regularity condition) の下で $n \to \infty$ につれて帰無仮説 H_0 の下で

$$2\log LR_n \xrightarrow{d} \chi^2(r) \tag{8.34}$$

である。ここで r は制約条件の数を意味する。

帰無仮説の下での分布が求められると標準的な場合の議論より棄却域を構成することができる。

8.3 区間推定

推定問題において未知母数の値を一点として推定する方法が点推定 (point estimation) であった。これに対して客観的信頼度の基準で未知母数が含まれる区間を標本から推定する方法は区間推定 (interval estimation) と呼ぶ。

例 8-5：ある家電メーカーの生産するテレビの寿命が未知の平均 μ, 標準偏差 (例えば)$\sigma_0 = 10$ で既知の正規分布 $N(\mu, \sigma_0^2)$ にしたがっているとする。 n 個の独立標本 $X_1, \cdots, X_n$ が利用可能とするとき信頼係数 99 %の信頼区間 (confidence interval) を求める。正規分布の母集団から独立に得られた標本平均

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

とすると期待値 $E(\bar{X}) = \mu$, 分散 $Var(\bar{X}) = \sigma_0^2/n$ の正規分布となる。そこで標本平均を基準化して**かなめの量** (pivot)

$$Z = \frac{\bar{X}_n - \mu}{\sigma_0/\sqrt{n}} \tag{8.35}$$

とおけば、この確率変数の分布は未知母数の値に依存せずに標準正規分布 $N(0,1)$ にしたがう。正規分布の数表を使えば

$$P(-2.58 \leq Z \leq 2.58) = .99$$

あるいは

$$P(-1.96 \leq Z \leq 1.96) = .95$$

となる。正規分布など基本的な確率分布の数表の利用法に慣れておくことは重要である。数値は統計学の入門的教科書、あるいは近年ではエクセルなどのソフトウエアで得られるが、正規分布表から求められたパーセント点 2.58 や 1.96 はもっともよく応用で使われる数値である。確率変数 Z の定義を代入して整理すれば

$$P\left(\bar{X}_n - 2.58\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 2.58\frac{\sigma_0}{\sqrt{n}}\right) = .99 \tag{8.36}$$

となる、例えば母数 μ の 99 %信頼区間は

$$[\bar{X}_n - 2.58\frac{\sigma_0}{\sqrt{n}}, \bar{X}_n + 2.58\frac{\sigma_0}{\sqrt{n}}] \tag{8.37}$$

で与えられる。このように得られる信頼区間は十分統計量である標本平均のみに依存しているので標本平均が母平均の情報を代表していると見なせる。また区間の長さは $5.16\sigma_0/\sqrt{n}$ であるので標本数 n が大きくなれば区間が狭くなりより精度が高く区間推定できる。

ここで母集団分布から大きさ n の独立標本 $X_1, \cdots, X_n$ が得られる時、区間の上限を示す統計量 $U(X_1, \cdots, X_n)$ と下限を示す統計量 $L(X_1, \cdots, X_n)$ を構成し、２つの統計量を使って確率

$$P(L(X_1, \cdots, X_n) \leq \theta \leq U(X_1, \cdots, X_n)) = 1 - \alpha$$

によって信頼係数 $100(1-\alpha)$ を定める。このとき区間

$$[L(X_1, \cdots, X_n), U(X_1, \cdots, X_n)] \tag{8.38}$$

を信頼区間 (confidence interval) と呼ぶ。この区間の上限と下限は確率変数であるから母集団の確率分布により変動し、標本の分布とともに区間は確率変数として変動

するランダム集合である。ここでデータ $x_1, \cdots, x_n$ から計算される $100(1-\alpha)$ パーセント**信頼区間** $[L(x_1, \cdots, x_n), U(x_1, \cdots, x_n)]$ はこのランダム集合のある実現と解釈できるが、信頼係数の意味については若干の注意が必要である。n 個の標本が実際に観測されるたびに 1 つの信頼区間が計算される。もし，真の母数 θ が未知、ある値をとるとすれば観測値から計算された信頼区間はこの θ の値を含むか否かのどちらかである。しかし真の母数の値は統計家には未知なので実際に計算された信頼区間が真の値を含んでいるか否かを判断出来ない。信頼係数 $100(1-\alpha)$ とはこうした信頼区間の観測を繰り返すことで何回も計算すれば標本分布からほぼ $100(1-\alpha)$ パーセントの割合で信頼区間が真の値を含むことを意味する。したがって，信頼区間

$$P\left(L(X_1, \cdots, X_n) \leq \theta \leq U(X_1, \cdots, X_n)\right) = 1-\alpha \tag{8.39}$$

を母数 θ がある範囲に入る確率とする解釈は古典的統計学の立場からは妥当ではない[34]。

例 8-6：前例で母集団の正規分布の分散が既知としたが多くの問題では分散を未知とする方が自然である。母分散が未知であっても変数 Z の分布は標準正規分布になることには変わらない。ただし、確率変数 Z は未知の母数 σ_0 を含むので変数 Z の式の中で未知の標準偏差 σ_0 の代わりに推定量

$$s_n = \hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}$$

を代入して求めた統計量

$$T = \frac{\bar{X}_n - \mu}{\hat{\sigma}/\sqrt{n}} \tag{8.40}$$

を t-統計量と呼ぶ。この統計量の分子 $\left(\bar{X}_n - \mu\right)/\left(\sigma_0/\sqrt{n}\right)$ は $N(0,1)$ にしたがい、分母を構成する $(n-1)\hat{\sigma}^2/\sigma_0^2)$ は分子とは独立に χ^2 分布にしたがう。したがって、統計量は自由度 $n-1$ の t-分布にしたがうので、数表から自由度 $n-1$ の t-分布の両側 100α %点を $t(\alpha/2, n-1)$ とすれば

$$P\left(-t(\frac{\alpha}{2}, n-1) \leq T \leq t(\frac{\alpha}{2}, n-1)\right) = 1-\alpha \tag{8.41}$$

となる。この場合には $100(1-\alpha)$ の信頼係数の信頼区間は

$$[\bar{X}_n - t(\frac{\alpha}{2}, n-1)\frac{s_n}{\sqrt{n}}, \bar{X}_n + t(\frac{\alpha}{2}, n-1)\frac{s_n}{\sqrt{n}}] \tag{8.42}$$

[34]信頼区間はベイズ統計学における事後分布にもとづく母数のベイズ信頼区間とは異なる。

で与えられる。この場合は標本から計算するのは標本平均 $\bar{X}_n$ と標本不偏分散 $\hat{\sigma}^2$ は母平均と母分散がともに未知なのでこの 2 つの母数の情報を標本平均と標本分散が代表していると解釈できる。この信頼区間の幅は $2 \times t(\alpha/2, n-1)s_n/\sqrt{n}$ であり、もし標本分散が大きければそれだけ大きくなり、例えば $n = 10$ のとき分散が既知とすると 95 %信頼区間の長さは $2 \times 1.96\sigma_0/\sqrt{10}$，分散が未知の場合には $2 \times 2.26S/\sqrt{10}$ となる。したがって後者の区間の方が長くなるが、直観的には標本分散はデータのばらつきを表現しているので，この量が大きければ母平均の信頼区間の精度が小さくなると解釈できる。

区間推定を応用する上では与えられた信頼係数に対して区間の幅が小さい方が望ましい。したがって最強力信頼区間とは与えられた情報の下ではそれ以上に区間幅を小さくとれない信頼領域を意味する。これまで説明したように $100(1-\alpha)$ の信頼係数の信頼区間は第一種の過誤が 100α) の検定問題における受容域 (acceptance region) に対応している。したがって最強力検定 (most powerful test) が存在する場合には最適な信頼区間、信頼域を構成することができる。

8.4 ブートストラップ法とリサンプリング法

統計的推定論では母集団分布 $F(x|\theta)$ に互いに独立にしたがう n 個の (それぞれスカラーの) 独立な確率変数 $X_1, \cdots, X_n$ より母数の推定量 $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$ を構成する方法を議論した。推定した結果の有意性や検定・信頼区間を構成するためには統計量のばらつきなどを推定する必要がある。例えば尤度関数が既知で最尤推定法が適用可能な場合には定理 7-5,7-6 を利用すると、フィッシャー (Fishe)r 情報量の推定量 $I_n(\hat{\boldsymbol{\theta}}_n)$, あるいは対数尤度関数の二回微分の推定量

$$\hat{I}_{n,ij}(\theta) = (-\partial^2 l_n(\hat{\theta})/\partial\theta_i\partial\theta_j)$$

を用いて推定量の分散・共分散、あるいは漸近的な意味での漸近分散・共分散を推定することは可能である。
それでは尤度関数が複雑であったり、尤度関数の利用が困難な場合、あるいは最尤推定ではない推定法を利用する場合に推定量のばらつきを推定する方法は可能であろうか？

エフロン (B. Efron) は実際に観察される統計データ $\mathbf{x} = (x_1, \cdots, x_n)'$ からのサンプリング (re-samplig) に基づくブートストラップ (bootstrap) 法と呼ばれる次の方法を提案した。(i) 観察データより経験分布関数 $F_n(x)$ (すなわち各観察値に $1/n$ の確率を付与した確率分布）を作成する。(ii) この確率分布関数よりサンプリングを繰り返し標本データ $\mathbf{x}^* = (\mathbf{x}_1^*, \cdots, \mathbf{x}_B^*)'$ を作成する。(各 $\mathbf{x}_i^*$ は n 次元データを意味する。)

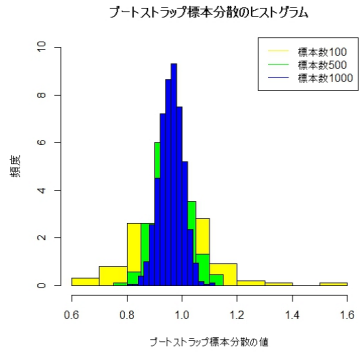

図 8-1: ブートストラップの例

この各ブートストラップ・データにもとづいて推定量を求め $\hat{\theta}_i^* = \hat{\theta}_i^*(\mathbf{x}_i^*)$ とする。このリサンプリング操作を B 回繰り返し、推定量 $\hat{\theta}_n$ の分散を

$$\hat{\sigma}^2(\hat{\theta}_n) = \frac{1}{B-1}\sum_{i=1}^{B}[\hat{\theta}_i^* - \hat{\theta}_n]^2 \tag{8.43}$$

により推定する。

例 8-7：例えば確率変数 X_i $(i = 1, \cdots, n)$ が正規分布 $N(\mu, \sigma^2)$ にしたがうとき標本平均の分散推定を取り上げてみよう。真の確率分布は事前には分からない状況を想定するが、シミュレーションにより真の分布から 1000 個の乱数を発生し、標本平均が真の期待値 $\mathbf{E}[X] = \theta$ の周りに分布するばらつき（真の値は 1 とした）を評価してみた例である。図 8-1 がその推定結果の一例を示しているが、リサンプリングの回数を十分に大きく取るとブートストラップ標本の分布は真の値の周りに分布し、妥当な結果が得られることが分かる。

例えば既に議論した簡単ではあるがより困難な場合として標本分散

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

の分散の評価を考えよう。この場合には漸近的な分散の評価式は既に登場しているが、高次の積率に依存していることに注意しよう。データ数が数がそれほど多くない場合にどのように評価したらよいかがここでの問題である。別の例としては標本分位点 $X_{([q\,n])}$ の分散評価の問題がある。n を標本数, $0 < q < 1$,$[y]$ は実数 y を超えない最大の整数として、例えば $q = 1/2$ は $X_{([q\,n])}$(もしくは $X_{([q\,n]+1)}$) は標本メディアン (Median) であるが母集団が連続型の確率変数の場合、順序統計量に基づく評価

より漸近的には $\sqrt{n}[X_{([n/2])} - x_{median}]$ は $N(0, 1/[4f(0)^2])$ にしたがうことが知られている。(ここで $f(x)$ は密度関数, x_{median} はメディアンであり密度がゼロ点で退化しないことを仮定した。)

こうした標本分散や標本分位点の分散推定、あるいは分散に基づく信頼区間の構成などの問題ではブートストラップが有力な手段である。このブートストラップ法は経験分布関数 F_n による確率分布関数 F の近似という観点から理論的正当化が可能である。

ブートストラップ法は今日では様々な形で統計分析で利用されている標準的な計算統計 (computational statistics) 法の一種である [35]。この方法は統計データが与えられたとき経験分布 F_n を元にしたリサンプリングにより情報を得ようとする統計的方法である、例えば推定のところで言及したジャックナイフ法などと通じるリサンプリング法と理解できる。

ブートストラップ法はかなり広範に適用可能であるが、十分に機能しない例として例 7-15 で言及した母数が端点となる場合を挙げることができる。こうした状況で利用可能なリサンプリング法としてはサブサンプリング (部分抽出) 法を挙げることができる [36]。ブートストラップ法では毎回経験分布から n 個のデータをリサンプリングすることになっている。サブサンプリング法では n 個のデータから b $(b \geq 1)$ 個のデータをサンプリングし、それぞれの組 (b 次元ベクトル) を $\mathbf{Y}_1, \cdots, \mathbf{Y}_{N_n}$ とする。この組み合わせは n から b を選ぶ組み合わせ数 N_n である。求めたい量は

$$J_n(\mathbf{x}, F) = P\left(\tau_n(\hat{\theta}_n - \theta) \leq x\right) \tag{8.44}$$

であるが、再標本の組 $\mathbf{Y}_i$ $(i = 1, \cdots, N_n)$ から得られる推定量 $\hat{\theta}_{b,i}$ とすると、$J_n(\mathbf{x}, F)$ の推定量は指標関数 $I(\cdot)$ を利用して

$$L_{n,b}(\mathbf{x}, F_n) = \frac{1}{N_n} \sum_{i=1}^{N_n} I\left(\tau_b(\hat{\theta}_{b,i} - \hat{\theta}_n) \leq x\right) \tag{8.45}$$

となる。ここで $\hat{\theta}_{b,i}$ は第 i 組の再標本からの推定値であるが、こうした推定が漸近的に意味を持つ条件は b を n に依存するように b_n にとり $\tau_n \to \infty, \tau_b \to \infty, b_n/n \to 0$ である。標本平均に代表される通常の状況では $\tau_n = \sqrt{n}$ であるから $b_n = n^\gamma$ $(0 < \gamma < 1/2)$ とすればよい。

---

[35] Efron,B.and R.Tibshirani (1993),"An Introduction to the Bootstrap," Chapman & Hall が基本的な文献である。身近な応用例として例えば国友・一場 (2006)「多期間リスク管理法と変額年金保険」日本統計学会誌,103-123 を挙げておく。

[36] Politis, D. , Romano, J.and and M.Wolf (1999), "Subsamplig,"Springer が基本的な文献である。

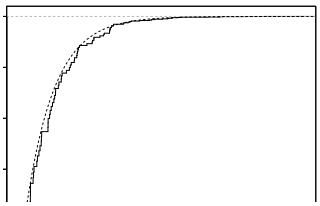

図 8-2: サブサンプリングの例

例 8-8：例えば例 7-15 の状況を取りあげると、この場合には $\tau_n = n$ であるから $b_n = n^\gamma$ $(0 < \gamma < 1)$ とすればよい。これをシミュレーションにより $\gamma = 1/2$ としてサブサンプリング法を実験してみると図 8-2 が示すように推定量の極限分布に近い確率分布を再現できることがわかる。したがってこの場合にはサブサンプリング法を利用して分散なども推定することができる。

ブートストラップ法やサブサンプリング法と云った計算機を用いた最近の計算統計の方法は経験分布 F_n は n が大きいとき真の分布 F に収束できるので、漸近的に近似できることからここで述べたリサンプリング法は統計理論からも正当化は可能である。

8.A 付論：局所漸近効率性と高次効率性

統計的推定論・検定論の展開において重要ではあるが技術的なかなり複雑で高度な議論が必要な問題について簡単に言及しておこう。互いに独立な標本 X_i $(i = 1, \cdots)$ が確率分布 $F(x|\theta)$ にしたがう場合、最尤推定量は漸近的には一定の条件下で下限 $I_1(\theta)^{-1}$ を達成する場合に効率的であることと説明した。漸近的に効率的な推定量 $\hat{\theta}_n$ とするとホッジス (Hodges) は次のような例を作成した。任意の実数 $\alpha > 0$ に対して

$$\hat{\theta}_n^* = \begin{cases} \hat{\theta}_n & \text{if } |\hat{\theta}_n| \geq n^{-1/4} \\ \alpha\hat{\theta}_n & \text{if } |\hat{\theta}_n| < n^{-1/4} \end{cases} \tag{8.46}$$

として新たな推定量を定義する。この推定量はもし $\theta = 0$ であれば漸近分散は $\alpha/I_n(\theta)$、$\theta \neq 0$ であれば $1/I_n(\theta)$ となるので $\hat{\theta}_n$ よりも局所的に効率的となる。

この例は不合理な推定量を与えていると見れるが、実は任意の効率的な推定量について適用可能である。こうした推定量を排除する一つの方法は (真の) 母数 $\theta = \theta_0$

について推定量がある種の滑らかな挙動を示す範囲に限定することである。母数の値 θ_0 として局所対立仮説 $H_{1l} : \theta = \theta_0 + \frac{h}{\sqrt{n}}$($h$ は任意の実数) の下で推定量 T_n

$$\sqrt{n}\left[T_n - \theta_0 - \frac{h}{\sqrt{n}}\right] \tag{8.47}$$

漸近的に正規分布 $N(0, \sigma^2(\theta_0))$ に一致するクラスを正則な推定量 (regular estimator) と呼ばれている。

こうした局所的に正則な推定量は LAN(local asymptotic normal) 推定量とも呼ばれるが、一定の仮定の下で最尤推定量はこうした正則な推定量の中で効率的となる[37]。漸近的に効率的な推定量は一般に一意には定まらない。すなわち最尤推定量以外の推定量、積率法をなどに基づく推定法が漸近的な意味で効率的となり漸近的に同等な推定量が複数ある場合も少なくない。したがって最尤推定量が何らかの意味で他の推定量よりも効率的ではないか、との疑問が生じる。この問題に対する解答を与えているのが高次漸近理論であり漸近分布を標本数 n により展開する漸近展開に基づく高次の漸近効率に関して一時期には活発に議論された[38]。そこでの主要な結論としては漸近的な意味で「バイアス補正」を漸近的に行うと最尤推定量は他の推定方法を高次の漸近分布の意味で優越するというものであるが、確率分布の漸近展開という手法を十分に利用する必要があるのでここでは言及するにとどめる。

[37]例えば Van der Vaart (1999),"Asymptotic Statistics," Cambridge University にく説明がある。

[38]例えば Akahira,M. and K. Takeuchi (1982),"Asymptotic Efficiency of Statistical Estimator,"Springer に詳しい説明がある。

9　統計的決定理論とベイズ推論

9.1 ベイズ推論

統計家を含め現実に直面する多くの場面では自然の状態について何らかの事前的知識を持っていることが多い。例えば過去の経験から何度も母数を推定した結果、母数について情報を持つ中で新たな情報を分析することがしばしば見られる。こうした状況においてさらに母数に関する事前的知識が確率分布として表現できる状況、すなわち母数 θ の事前情報が事前分布 $\pi(\theta)$ として利用できる場合を考察しよう。実は歴史的にはこの章で述べる逆確率 (inverse probability) による議論の方が事前分布を利用しない古典的統計学よりさらに溯るのであり、フィッシャー (R.A.Risher) は逆確率に代わる議論として最尤法を導入した。その意味ではここで説明するベイズ統計学は新古典派統計学と呼べるであろう。なお本章では表記の煩雑さを避け主に 1 次元母数空間を扱うが一般化は可能である。

ベイズ推論の例

ベイズ推論では事前分布 (prior distribution) の設定、標本分布の分析、事後分布 (posterior distribution) の導出が基本である。母数についての事前分布は過去の情報や統計的分析から得られる情報を表現し、母数の事後分布は事前分布と母数を所与とする標本データの情報を組み合わせた事後分布は母数についての全ての情報をまとめるものである。ここでは母数 θ がスカラー、標本分布が 1 次元確率変数で表現できる場合を考察するが多次元の場合も同様に分析できる。。

例 9-1： 二項分布 $X \sim B(n, \theta)$ に対し確率 θ $(0 \leq \theta \leq 1)$ の事前情報として、事前確率分布として

$$\pi(\theta) \sim \theta^{p-1}(1-\theta)^{q-1} \tag{9.1}$$

ベータ分布 $Beta(p, q)$ がよく用いられる。標本 $X = x(x$ はデータ$)$ が得られるとき (条件付分布についての)

$$\pi(\theta|\mathbf{x}) \sim \theta^{p-1}(1-\theta)^{q-1}\binom{n}{x}\theta^{x}(1-\theta)^{n-x} = \binom{n}{x}\theta^{p+x-1}(1-\theta)^{q+n-x-1} \tag{9.2}$$

より再びベータ分布 $Beta(p+x, q+n-x)$ となる。このように事前分布と事後分布が同一の分布族となるとき事前分布は自然共役 (natural conjugate) 分布と呼んでいる。標本データが得られた後の母数の推定量として事後分布の期待値を利用すると

$$\hat{\theta}_B = \mathbf{E}[\theta|X = x] = \frac{p+x}{p+q+n} \tag{9.3}$$

となる。

例 9-2 (ベータ分布)： ベータ分布は $[0,1]$ 上に値をとる確率 p に対する事前分布としては自然なクラスである。一般に実数 $p>0, q>0, \beta>0, 0<x<\beta=1$ に対し密度関数は

$$f(x|p,q)=\frac{1}{B(p,q)}x^{p-1}(1-x)^{q-1} \quad 0<x<1 \tag{9.4}$$

で与えられる。ここでベータ関数 $B(p,q)$ はガンマ関数と関係

$$B(p,q)=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \tag{9.5}$$

であるので、二項モデル $X\sim B(n,p)$ にしたがうデータが得られるとき、事後分布は $Beta(p+x, q+n-x)$ となる。その期待値 (事後平均) は関係

$$\begin{aligned}\int_0^1 x\cdot x^{p+x-1}(1-x)^{q+n-x-1}dx &= B(p+x+1+q+n-x)\\ &= \frac{\Gamma(p+1)\Gamma(q+n)}{\Gamma(p+1+q+n)}\end{aligned}$$

およびガンマ関数の性質 $\Gamma(y+1)=y\Gamma(y)$ を用いて評価できる。

事後分布の期待値をベイズ解と呼ぶと、標本数 n が大きければ p,q に対して X,n の値が相対的に大きくなり最尤推定量 $\hat{\theta}_{ML}=\frac{X}{n}$ に近いが n が大きくなければ事前分布が影響して値は異なる。この例では事前分布と事後分布の確率分布族が同一、となるので標本分布が二項分布の場合はベータ分布が自然共役分布である。別の例としては位置母数を推定する問題では標本分布が正規分布の場合には正規分布が自然共役分布となる。

例 9-3： 互いに独立な標本 X_i $(i=1,\cdots,n)$ が正規分布 $N(\theta,\sigma^2)$ にしたがうとする。簡単化のために σ^2 を既知として、θ の事前分布として $N(m,\tau^2)$ を想定しよう。この場合には θ の事後分布も正規分布になるが、正規分布の密度関数の積の表現より事後分布は正規分布 $N(m_*,\sigma_*^2)$ であって

$$m_*=\left[\frac{\sigma^2}{\sigma^2+n\tau^2}\right]m+\left[\frac{n\tau^2}{\sigma^2+n\tau^2}\right]\bar{X}_n\ ,\sigma_*^2=\frac{\sigma^2\tau^2}{\sigma^2+n\tau^2} \tag{9.6}$$

となる。したがって例えば母数の推定量として事後分布の期待値を利用すると

$$\hat{\theta}_B=\mathbf{E}[\theta|\bar{X}_n=x]=m_* \tag{9.7}$$

となる。この量は事前分布の期待値と標本平均の加重和となり事前情報が標本が得られると事前平均が修正されることを意味する。n が大きければ m_* はほぼ $\bar{x}_n$ に一致するが、事後分布の分散は n が大きいとき $\mathbf{E}[(\theta - \mathbf{E}(\theta))^2 | \bar{X}_n = x] \to 0$ である。事後分布の期待値は事前分布の期待値と標本平均の加重平均と表現され、事前情報が標本情報により改訂されることを意味しているが、データ数が大きくなると事前情報のウエイトは小さくなり標本情報により決定される。n が大きいときに事後分布を事後分布の確率極限 θ_0 の周りで評価すると

$$\begin{aligned} \frac{\theta - m_*}{\sqrt{\frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}}} &= \frac{1}{\sigma\sqrt{\frac{n\tau^2}{\sigma^2+n\tau^2}}}\sqrt{n}\left[(\theta-\theta_0)-(m_*-\theta_0)\right] \\ &\sim \frac{\sqrt{n}}{\sigma}[(\theta-\theta_0)-(\bar{x}_n-\theta_0)] \end{aligned}$$

となる。左辺は n が大きいとき $N(0,1)$ にしたがうので $\sqrt{n}(\theta-\theta_0)$ の事後分布は n が大きいとき期待値 $\sqrt{n}(\bar{x}_n-\theta_0)$ の正規分布 $N(\sqrt{n}(\bar{x}_n-\theta_0), \sigma^2)$ により近似できる。

一般に母数 θ の事前分布を $\pi(\theta)$, 互いに独立な標本の確率密度関数を $f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)$ としよう。事後分布の期待値を

$$M_n = \mathbf{E}[\theta | x_1, \cdots, x_n] \tag{9.8}$$

と表現すると、条件付期待値の繰り返しの公式より

$$\mathbf{E}[M_{n+1} | M_n, \cdots, M_1] = M_n \ (a.s.), \tag{9.9}$$

となる。したがって M_n はマルチンゲールである。このことから事後分布の期待値（事後平均）が存在するという条件の下ではマルチンゲールの収束定理により

$$M_n \longrightarrow \theta_0 \ \ (a.s.)$$

となる収束先 θ_0 が存在する。さらに事前分布に正規分布, 標本分布に正規分布を仮定したときの類似の議論より $I_1(\theta_0)$ を母数 θ_0 の Fisher 情報量とすると、互いに独立な標本が得られる場合には、一定の正則条件の下では事後分布 $\pi(\theta|x_1, \cdots, x_n)$ は θ_0 周りで近似的に正規分布

$$N\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n I_1(\theta_0)^{-1}\left(\frac{\partial \log f(x_i|\theta_0)}{\partial\theta}|_{\theta_0}\right), I_1(\theta_0)^{-1}\right] \tag{9.10}$$

にしたがうが、この種の命題は Bernstein-von-Mises 定理と呼ばれている。

ベイズ・リスク

母数の不確実性について、事前情報が事前分布として表現できるとき、ベイズ推論では事後分布がもっとも重要な情報を縮約していると考える。ここで事前情報が利用可能とき推定量の規準についてより一般的に考察してみよう。母数 θ と推定量 $d(\mathbf{x}) = \hat{\theta}_n(\mathbf{x})$ (標本として得られるデータ $\mathbf{x} = (x_1, \cdots, x_n)'$) との評価関数を損失関数 $l(d, \theta)$ とすると、平均二乗誤差 (MSE) とは $l(d, \theta) = |d - \theta|^2$ とすると二次リスク関数 $\mathrm{R}(d, \theta)$

$$\mathrm{R}(d, \theta) = \mathbf{E}[|d - \theta|^2] = \int_{\mathbf{x} \in \mathbf{R}^n} |d - \theta|^2 f(x_1, \cdots, x_n|\theta) d\mathbf{x} \tag{9.11}$$

と表現できる。より一般にはリスク関数 $R(d, \theta)$ は

$$\mathrm{R}(d, \theta) = \int_{\mathbf{x} \in \mathbf{R}^n} l(d, \theta) f(x_1, \cdots, x_n|\theta) d\mathbf{x} \tag{9.12}$$

である。ここで母数について事前分布が利用可能であれば θ の事前分布について期待値をとるとリスクを小さくする評価は

$$r(\pi, d) = \mathbf{E}_\theta[R(\theta, d)] \tag{9.13}$$

を最小化する問題と考えられる。このリスク最小化問題の解をベイズ解、リスクの値をベイズ・リスク (Bayes Risk) と呼ぶことしよう。ここで連続型分布を仮定して互いに独立な標本がしがたう密度関数 $f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta)$ ($\mathbf{x} = (x_1, \cdots .x_n)'$) とすると

$$\begin{aligned} r(d, \pi) &= \int \left[\int l(\mathbf{x}, \theta, d) f(\mathbf{x}|\theta) d\mathbf{x} \right] \pi(\theta) d\theta \\ &= \int \left[\int l(\mathbf{x}, \theta, d) \pi(\theta|\mathbf{x}) d\theta \right] h(\mathbf{x}) d\mathbf{x} \end{aligned}$$

と表現できる。ここで $\pi(\theta|\mathbf{x})$ は $\mathbf{X} = \mathbf{x}$ が与えられた下での θ の事後分布である。ここでベイズの定理の応用から

$$h(\mathbf{x}) = \int f(\mathbf{x}|\theta) \pi(\theta) d\theta = \int f(\mathbf{x}, \theta) d\theta$$

より変数ベクトル $\mathbf{x} = (x_1, \cdots, x_n)'$ の周辺密度が与えられる。標本 $\mathbf{X}$ の周辺密度関数、確率変数ベクトル $(\mathbf{X}, \theta)$ の同時密度関数 $f(\mathbf{x}, \theta)$ と表現したが、式変形において鍵となるのは条件付密度関数に関するベイズの公式である。ベイズ解は

$$r(d, \pi|\mathbf{x}) = \int l(\mathbf{x}, \theta, d) \pi(\theta|\mathbf{x}) d\theta \tag{9.14}$$

を最小化すれば良いので、ベイズ・アプローチでは事後分布を求めることが最も重要な問題となる。

例 9-4： ここで点推定問題では、損失関数として $l(d,\theta) = |d-\theta|^2$ とすると事後分布の標本を所与とする事後分布の期待値の規準が導かれる。他方、損失関数として $l(d,\theta) = |d-\theta|$ とすると事後分布の標本を所与とするメディアン (median) の規準が導かれる。このことは 1 次元母数空間として $\mathbf{\Theta} = [\theta_0, \theta_1]$ として

$$\begin{aligned}\mathbf{E}[|\theta - d||\mathbf{x}] &= \int_{\theta_0}^{d}(d-\theta)\pi(\theta|\mathbf{x})d\theta + \int_{d}^{\theta_1}(\theta-d)\pi(\theta|\mathbf{x})d\theta \\ &= d[\int_{\theta_0}^{d} - \int_{d}^{\theta_1}]\pi(\theta|\mathbf{x})d\theta - \int_{\theta_0}^{d}\theta\pi(\theta|\mathbf{x})d\theta + \int_{d}^{\theta_1}\theta\pi(\theta|\mathbf{x})d\theta\end{aligned}$$

を d について最小化すると条件 $[\int_{\theta_0}^{d} - \int_{d}^{\theta_1}]\pi(\theta|\mathbf{x})d\theta = 0$ が得られることより導ける。

例 9-5： 仮説検定問題では 0-1 損失関数を考えるのが妥当である。自然の状態を二つの仮説として表現するので統計家は二つの決定、帰無仮説か対立仮説のどちらかを選択し、誤りは第一種の過誤、第二種の過誤が存在している。前節では帰無仮説 $H_0 : \theta = \theta_0$ 対対立仮説 $H_1 : \theta = \theta_1$ が共に単純仮説の場合にはネイマン・ピアソン (NP) 補題により最適な検定方式を求めた。また前節の L 集合の説明では最強力検定は $[0,1],[1,0]$ を含む原点に対して凸集合で可能な検定方式が表現され、原点に近い限界が最強力検定に対応していた。母数 θ についての事前確率が利用可能なとき、原点に対して凸集合の場合には境界点で接する直線が存在するのでそれを $\pi_1\alpha_1 + \pi_2\alpha_2 = c(c$ は定数$)$ と表現することができる。これらの π_1 と π_2 はそれぞれ二つの仮説に対する事前確率に対応し、ベイズ解が最適解に対応している。

ベイズ計算統計

ベイズ分析ではデータ $\mathbf{x} = (x_1, \cdots, x_n)'$ が与えられた条件下で得られる事後分布 $\pi(\theta|x_1, \cdots, x_n)$ がもっとも重要な役割を演じる。従来は多くの統計的問題では事後分布は数理的には積分で表現はできても複雑なので具体的に計算することが困難と考えられていた。したがって教科書的な例や自然共役分布としてベイズの定理を用いて事後分布を明示的に求められる場合を除き、より複雑な事後分布の評価は簡単には出来なかった。実際の統計分析で必要となる確率モデルはかなり複雑となるので事後分布を積分計算により解析的に処理するには限界があったのである。

こうした計算上で生じる問題は計算機能力の飛躍的進歩と計算アルゴリズムの進歩により、計算機シミュレーションにより大量の数値計算が可能となり事後分布の計算が飛躍的に進歩した。例えば MH(メトロポリス・ヘイスティング) アルゴリズムなどに代表される MCMC(Markov-Cahin Monte-Carlo) と呼ばれている計算アル

ゴリズムが良い例である[39]。こうした計算アルゴリズはそれほど複雑でないが、繰り返し計算で事後分布を数値的に正確に評価することが可能となっている。

9.2 統計的決定問題

これまで議論してきた数理統計学の基礎理論はもともとそれぞれ極めて具体的な統計的問題を解決する理論として考案されたものである。統計的推定、統計的検定、統計的予測などの統計的推測問題は理論的には統計的決定問題として統一的に論じることができる。統計的決定理論と呼ばれている数理統計学の理論は元々は、フォン・ノイマン (John-von-Neumann)& オスカー・モルゲンシュタイン (O. Morgenstern) により始められたゲーム理論に触発され、数理統計家ワルド (A. Wald) により定式化されたものである。ワルドは統計家の直面する問題を自然対統計家のゲームとして理解すると、統計学理論の本質が統一的に理解できることを見いだし、数理統計学の基礎理論を構築した。

ここで自然の状態を表す母数 θ がとりうる範囲を母数空間 $\mathbf{\Theta}$, 実験や観察される確率変数 $\mathbf{X}$ のとりうる範囲を標本空間 $\boldsymbol{\mathcal{X}}$, 得られる情報の下での決定 d のとりうる範囲を決定空間 $\boldsymbol{\mathcal{D}}$ と呼ぼう。ここで決定は確率変数 $\mathbf{X}$ の関数 $d(\mathbf{X})$ である。確率変数 $\mathbf{X}$ がしたがう確率法則 $P(\cdot|\theta)$ とする。これは自然の状態が与えられた下で確率変数 $\mathbf{X}$ の確率分布が与えられる、と言う意味である。典型的な問題では標本空間は n 個の確率変数ベクトル $\mathbf{X} = (X_1, \cdots, X_n)'$, 統計データは実ベクトル $\mathbf{x} = (x_1, \cdots, x_n)'$ であり、これらがしたがう確率分布 $F(\mathbf{x}|\theta)$ が連続型の確率分布であれば確率密度関数 $f(\mathbf{x}|\theta)$, 離散型の確率分布であれば確率関数 $p(\mathbf{x}|\theta)$ である。

次に組 $\mathbf{x}, \theta, d$ の評価関数として効用関数を $u(\mathbf{x}, \theta, d)$ と表す。統計的問題では効用関数そのものではなく負の効用関数 $-u(\mathbf{x}, \theta, d)$ を損失関数 $l(\mathbf{x}, \theta, d)$ と表し、損失が小さいことが好ましい規準とすることが多い。結果に不確実性が存在する世界ではフォンノイマン・モルゲンシュタイン流の合理的な統計家を想定すると期待効用最大化、あるいは期待損失最小化、という規準が合理的と考えられる。そこで期待損失によりリスク関数

$$R(\theta, d) = \mathbf{E}_{X|\theta}[l(\mathbf{x}, \theta, d)] \tag{9.15}$$

を定義する。ここで期待値 $\mathbf{E}_{X|\theta}[\,\cdot\,]$ は確率分布 $F(\mathbf{x}|\theta)$ についての期待値を意味する

[39]国友・山本編「２１世紀の統計科学」Vol-III 収録の古澄秀男「マルコフ連鎖モンテカルロ法入門」を参照。

が、特に密度関数が存在する場合には

$$R(\theta, d) = \int_{\boldsymbol{\mathcal{X}}} l(\mathbf{x}, \theta, d) f(\mathbf{x}|\theta) d\mathbf{x} \tag{9.16}$$

となる。

ここである決定関数に対してリスク関数は未知母数 θ の関数としておくと、決定方式の良さを測る規準としては次のような許容性の概念が重要となる。

定義 9-1： 決定 d_i $(i = 1, 2)$ に対するリスク関数を $R(\theta, d_i)$ とする。
(i) 任意の θ に対して $R(\theta, d_1) \leq R(\theta, d_2)$ であり、ある θ_0 に対して $R(\theta, d_1) < R(\theta, d_2)$ となるならば決定 d_1 は d_2 に優越する $d_1 \succ d_2$ と言う。
(ii) 決定方式 d に優越する決定方式が存在しない場合には d は許容的 (admissible) と云う。決定方式 d を優越する決定方式が存在する場合には非許容的 (inadmissible) と云う。

定理 9-1： 事前分布 $\pi(\theta)$ に対するベイズ決定関数 d_π が一意である時、d_π は許容的である [40]。

証明の概略 ： 任意の θ に対してリスク関数 $R(\theta, d) \leq R(\theta, d_\pi)$ が成り立つとする。事前分布について期待値 $\mathbf{E}_\pi(\,\cdot\,)$ をとると

$$r(\pi, d) = \mathbf{E}_\pi[R(\theta, d)] \leq r(\pi, d_\pi) = \mathbf{E}_\pi[R(\theta, d_\pi)]$$

となる。ベイズ解は左辺を最小にする解である。仮定より一意であるから $d = d_\pi$ となり許容的となる。**Q**.**E**.**D**.

ベイズ解は理論的には許容的という良い性質がある。したがってある決定方式が許容的であるか否かはベイズ解となることを示すことで解決できる場合が少なくない。

ここでベイズ解と云う意味は事前分布が確率分布として表現できる場合でありこうした場合を proper 事前分布と呼んでいる。形式的には事前分布として $\pi(\theta) \sim$(一定値) という場合にも事後分布を求めることができる。例えば例 9-3 の期待値の事前分布として採用すれば範囲は $(-\infty, +\infty)$ なので improper な事前分布、あるいは散漫な (diffuse) 事前分布と呼ばれるが、これを事前知識を十分に持たない状況と解釈が可能である。こうした場合でも妥当な事後分布やベイズ決定関数が導かれる場合も少なくない。またこのことは一般にベイズ決定関数は許容的であるが、逆に許容的な決定関数はほぼベイズ決定関数であるという事実に関連しているのである。

事前分布の役割
ベイズ・アプローチについては事前分布の利用がもっとも論争的な論点である。科

[40] 例えば竹村彰通「現代統計学」328 項。

学的な統計分析に先だって様々な知見・経験を事前情報として利用することを全面的に否定する統計家や科学者は少ない。例えばパラメトリック統計では様々な確率分布を真の統計モデルとして想定することが多いが、「真の確率分布」が事前に分かっている場合は多くはないと言える。またある種の事前分布を想定した下での事後分布から得られる統計的知見は妥当なことが多いが、このことは統計的決定論の議論から裏付けられと考えられる。ただし、どのように事前情報を確率分布として表現し、どのような事前確率分布を用いて統計的推論を行うか、応用上の問題を解決する上でどのように事前分布 (prior distribution) を選択するかについては統計家の間ではなお論争的である [41]。

スタイン問題

統計理論家チャールズ・スタイン (C. Stein) は p 次元正規分布の平均の推定という極めて標準的問題を考察した。そして p 次元確率変数 $\mathbf{X}$ を原点に向かって縮小する統計量

$$d^S = \left[1 - \frac{p-2}{\|\mathbf{X}\|^2}\right]\mathbf{X} \tag{9.17}$$

を用いて次のような結果を導いた。ここでは単純化の為に分散共分散は単位行列, 標本 $n = 1$ として議論を進めるが、むろん $n \geq 1$ への一般化は可能である。

定理 9-2 : $\boldsymbol{\theta} = (\theta_i), \mathbf{d} = (d_i)$ に対して二次損失関数

$$l(\boldsymbol{\theta}, \mathbf{d}) = \|\mathbf{d} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{p}(d_i - \theta_i)^2 \ , \tag{9.18}$$

p 次元確率変数 $\mathbf{X}_i \sim N_p(\boldsymbol{\theta}, \sigma^2\mathbf{I}_p)$ $(i = 1, \cdots, n)$ とする。$p \geq 3$ のとき

$$\mathbf{E}[\sum_{i=1}^{p}(d_i^S - \theta_i)^2] < p\sigma^2 \tag{9.19}$$

となる。$\mathbf{X} = (X_i)$ の平均二乗誤差は $\mathbf{E}[\sum_{i=1}^{p}((X_i) - \theta_i)^2] = p\sigma^2$ より $p \geq 3$ のとき $\mathbf{X}$ は非許容的 (inadmissible) となる。

同様に議論より $n \quad \geq 1, p \geq 3$ のとき標本平均 $\bar{\mathbf{X}}_n = (1/n)\sum_{j=1}^{n}\mathbf{X}_j$ は非許容的になる。定理 9-2 は具体的にリスクを評価することで導けるが、スタインは正規分布に関するスタイン (Stein) の恒等式と呼ばれる等式を鍵として利用推定量のリスク関数を評価したのである。こうしたスタイン現象は特殊な例ではなく、多次元の統計分析では正規分布の期待値ベクトルばかりではなく分散・共分散の推定問題など他

[41] 例えば鈴木雪夫・国友直人 (1989) 「ベイズ統計学とその応用」東京大学出版会、を参照されたい。

の場合にも同様の結果が得られること分かり、許容性を巡る理論的展開が活発化している。スタインの縮小推定の議論は例えば小地域推定と呼ばれる政府統計の問題などでも応用されている[42]。

定理 9-2 の証明 : 一般性を失うことなく分散 $\sigma = 1$, 関数

$$d^S(\mathbf{X}) = [1 - g(\mathbf{x})]\mathbf{X}\ ,\ g(\mathbf{X}) = \frac{c}{\|\mathbf{X}\|^2}\mathbf{X} \tag{9.20}$$

とする。(c は定数であり例えば $c = p - 2$ とする。)

$$\begin{aligned}
\mathbf{E}[\|d(\mathbf{X}) - \boldsymbol{\theta}\|^2] &= \mathbf{E}[\|\mathbf{X} - \boldsymbol{\theta} - g(\mathbf{X})\|^2] \\
&= \mathbf{E}[\|\mathbf{X} - \boldsymbol{\theta}\|^2] + \mathbf{E}[\|g(\mathbf{X})\|^2] - 2\mathbf{E}[\sum_{i=1}^{p}(X_i - \theta_i)g_i(\mathbf{X})] \\
&= p + c^2\mathbf{E}[\frac{\|\mathbf{X}\|^2}{\|\mathbf{X}\|^4}] - 2\sum_{i=1}^{p}\mathbf{E}[(X_i - \theta_i)g_i(\mathbf{X})\|] \\
&= p + c^2\mathbf{E}[\frac{1}{\|\mathbf{X}\|^2}] - 2\sum_{i=1}^{p}\mathbf{E}[\frac{\partial}{\partial X_i}g_i(\mathbf{X})] \\
&= p + \mathbf{E}[\frac{1}{\|\mathbf{X}\|^2}][c^2 - 2c(p-2)] \\
&= p - (p-2)^2\mathbf{E}[\frac{1}{\|\mathbf{X}\|^2}] < p
\end{aligned}$$

となる。途中で次の補題および $g(\mathbf{X}) = (g_i(\mathbf{X}))$ に対し

$$\sum_{i=1}^{p}\frac{\partial}{\partial x_i}g_i(\mathbf{X}) = c\sum_{i=1}^{p}[\frac{1}{\|\mathbf{X}\|} - \frac{2X_i^2}{\|\mathbf{X}\|^4}] = c\frac{p-2}{\|\mathbf{X}\|^2}$$

となることを用いた。

補題 9-3(スタインの補題) : $Z \sim N(\mu, 1)$ のとき微分可能な関数 $g(z)$ に対して

$$\mathbf{E}\left[(Z - \mu)g(Z)\right] = \mathbf{E}[\frac{\partial g(Z)}{\partial Z}] \tag{9.21}$$

となる。

この補題は左辺が部分積分の公式[43] より

$$\begin{aligned}
&\int_{-\infty}^{\infty}(z-\mu)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-\mu)^2}g(z)dz \\
= &\left[-\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-\mu)^2}g(z)\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-\mu)^2}g^{'}(z)dz
\end{aligned}$$

[42]小地域統計の問題についての説明は、久保川達也「線形混合モデルの理論と応用」(国友・山本編「21 世紀の統計科学 Vol.III」収録) などにある。

[43]積の微分公式 $[fg]^{'} = f^{'}g + fg^{'}$ より $\int f^{'}g = [fg] - \int fg^{'}$ が導ける。

となることから右辺が導かれる。(**Q.E.D.**)

Part-III：数理統計の展開

10　統計的関係の推測

統計学の様々な応用分野では多数の変数の観測データを同時に統計的に分析することが求められることが多い。ここで利用可能な観測データの組の間にはしばしば確定的な関係ではなく、比例関係、反比例関係、循環的関係、などが誤差を伴う関係として理解すると有益な分析につながることが少なくない。統計的分析では観測される変数間の関係を確率的偶然変動を伴う統計的関係としてとらえることにより、統計的推定論、検定論、さらには統計的予測などにより意味ある結果を導く統計的方法が考案されている。統計的関係を巡る多くの統計的方法が既に存在し、多くの計算ソフトウエアに組み込まれている。統計的多変量解析や線形回帰分析など様々な統計的方法の数理的基礎を理解し、統一的に理解することが重要である。

ここで二つの変数 Y,Z のデータがほぼ

$$Y \sim e^{\alpha} z^{\beta} \tag{10.1}$$

(ここで α, β は母数とする) という統計的関係 (statistical relation) が想定できる場合がデータ分析の出発点としよう。ここでの関係は非線形的関係であるが、変数を対数変換することで線形関係の問題としてとらえることができる。むろん線形関係に帰着できないような関係もあるが、まずは線形関係の分析より出発することが有益な場合が多い。例えば変数変換 $\lambda \neq 0$ のとき

$$h(y, \lambda) = \frac{y^{\lambda} - 1}{\lambda} \tag{10.2}$$

とすると、対数変換は $\lambda \to 0$ という極限である。これがボックス・コックス (Box-Cox) 変換であるが、変換することで簡単な線形モデルによる統計分析に還元するアプローチは有用である。

ここでは変換を含めて既に適切に原データを処理した後に考え得る統計的関係を線形関係、観測値と理論的関係の差を観測誤差ととらえる線形回帰の問題を考察しよう。

10.1 最小二乗法と線形回帰モデル

線形回帰分析と最小二乗法 (least squares method) は経済分析など社会科学を含めてデータの統計分析ではよく用いられる方法である。この最小二乗法はしばしば

ガウス (K.F. Gauss) の方法と呼ばれるが[44]。誤差を伴って得られる観測データより観測誤差 (measurement errors) を (何らかの意味で) なるべく小さくして理論的な線形関係 (linear relationship) を計測することが主たる目的と考えられるが、観測誤差や変数誤差の問題は実は経済分析では無視できない場合も少なくない。例えば経済分析では所得、資産、資本ストック、土地価格の補足といったデータ収集上の問題、あるいは恒常所得や恒常消費といった実際に観察されない理論的な (真の) 変量に関する議論でも重要な役割を演じる。統計学や応用経済学でよく利用されている回帰 (regression)、回帰分析 (regression analysis) はもともと 19 世紀末〜20 世紀初頭にかけて行われた人間の形質遺伝に関わるデータ解析から考案された方法である[45]。生物統計学 (biometrics) から始まり近年では経済・経営・金融などを含む様々な分野での応用も盛んであり、回帰分析は様々な研究分野でごく常識的な統計的方法として定着している。

簡単な場合として二次元の単回帰 (simple regression) と呼ばれている場合を例として説明しよう。統計データとして n 組の被説明変数と説明変数の組 (二次元ベクトル) $\mathbf{x}_i^{'} = (y_i, z_i)^{'}\ i = 1, \cdots, n$ が与えられるとき、応用上でほぼ線形関係 $y = \beta_0 + \beta z_{(1)}$ によりデータを表現できることが重要な意味を持つことがある。むろん説明変数が 1 個であればデータの加減乗除により多くのことを表現することも可能であるが、実際の回帰分析では説明変数が複数考えられることより一般的であり、定数項 1 および $k-1$ $(k \geq 2)$ 個の説明変数 $z_{(j)}(j = 1, \cdots, k-1)$, 係数 β_j $(j = 0, 1, \cdots, k-1)$ を用いて線形関係を

$$y \sim \beta_0 + \beta_1 z_{(1)} + \cdots + \beta_{k-1} z_{(k-1)} \tag{10.3}$$

と表現して分析することが多いが、これを重回帰分析 (multiple regression) と呼ぶ。

最小二乗法

最小二乗法はある変数（被説明変数）を他の変数群（説明変数）による線形関係で説明する方法である。最小二乗法では n 組の誤差を説明変数である z 軸方向に測り、誤差の二乗和を最小化する方法であり、単回帰では観測データを (y_i, z_i) $(i = 1, \cdots, n)$ とすると、評価関数

$$L_2(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 z_i)^2 \tag{10.3}$$

を未知数 b_0 と b_1 について最小化することにより通常は解を求めることができる。こうした最小二乗法については様々な問題が存在する。例えば誤差の絶対値の和では

[44]正確には A. Legendre (1805)"Nouvelles Méthodes pour la détermination des Orbites Des Cométes" により導入された。K.F.Gauss が 1809 年に出版した本でオリジナリティを主張したので、科学論争に発展したようである、いずれの研究も当時の自然科学において重要であった天文学における観測データの解析に関してであったことは興味深い。

[45]例えば Stigler, S. (1986),"The History of Statistics," Harvard UP 等が詳しい。

なく誤差の二乗和を最小化することが合理的、すなわち誤差を点から直線への距離ではなく z 軸方向にとる必要があるか。一つの解釈は説明変数 z を与えた元での条件付き関係を推定しているという見方が標準的解釈である。そこで応用上での統計的関係の分析では対象の変数、例えば z と y について片方の変数をもう片方の変数で説明するというは発想には馴染まないものも少なくない。マクロ経済における消費と所得、教育学で議論する数学と英語の科目の関係、などが具体例として挙げて考えてみると良いであろう。

一般に線形回帰モデル

$$Y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j z_{ji} + U_i \ (i = 1, \cdots, n) \tag{10.4}$$

を考察しよう。ここで被説明変数の第 i 観測値 y_i, 第 j 番目の説明変数の第 i 観測値 z_{ji} $(i = 1, \cdots, n; j = 0, \cdots, k-1)$, $(z_{i0} = 1)$、係数 β_j $(j = 0, \cdots, k-1)$, 誤差項 U_i $(i = 1, \cdots, n)$ とする。(説明変数の観測値を並べて第 i 観測ベクトル $\mathbf{z}_i = (1, z_{1i}, \cdots, z_{k-1,i})'$ という表記を用いることもある。) (10.4) の表現は重回帰モデル、あるいは線形回帰モデルと呼ばれるが説明変数と誤差項は独立、あるいは説明変数を所与とするときに条件付期待値

$$(A1) \qquad \mathbf{E}[U_i|\mathbf{z}_i] = 0; (i = 1, \cdots, n)$$

を仮定する。条件 (A1) は例えば 6.3 節で説明した実験計画における分散分析モデルを例にとると分かりやすい。例えば説明変数 z が 0,1 の二値のみをとる説明変数 (ダミー変数と呼ぶ) とすると 1 元配置モデルは線形回帰モデルとして表現できる。この場合には説明変数は実験計画に置いてあらかじめ値を割り付けておく制御変数である。実験計画の設定で説明変数がよく制御されていれば誤差分散と共分散についての条件

$$(A2) \qquad \mathbf{E}[U_i U_j] = \sigma^2 \ (i = j), 0 \ (i \neq j)$$

を仮定しても差し支えなかろう。この線形回帰モデルは

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & \cdots & z_{k-1,1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_{1n} & \cdots & z_{k-1,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} U_1 \\ \vdots \\ U_n \end{bmatrix},$$

あるいは $n \times 1$ ベクトル $\mathbf{y} = (Y_i)$, $\mathbf{u} = (U_i)$, $n \times K$ 行列 $Z = (Z_{ij})$ を用いて

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{u} \tag{10.5}$$

と表現できる。ここで定数項 1 を含めて 2 個以上の説明変数がある場合には説明変数が互いに一時独立にとれる条件として

$$(A3) \qquad \text{「説明変数は固定された数値からの行列 Z で階数は k」}$$

を仮定する。様々な応用上では説明変数が制御変数でないことも少なくないがその場合には利用可能な $(k+1)\times 1$ 観測ベクトル $(y_i, \mathbf{z}_i^{'})^{'}$ は i $(i=1,\cdots,n)$ について互いに独立な確率変数列の実現値ベクトルと解釈する。
ここでベクトル $\mathbf{y}$ に対する距離 $\|\mathbf{y}\| = \sum_{j=1}^{n} y_i^2$, 損失関数を

$$\begin{aligned} L_2 &= \|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2 \\ &= \|(\mathbf{y} - \mathbf{Z}\mathbf{b}) + \mathbf{Z}(\mathbf{b} - \boldsymbol{\beta})\|^2 \end{aligned}$$

とする。ただし

$$\mathbf{b} = (\mathbf{Z}^{'}\mathbf{Z})^{-1}\mathbf{Z}^{'}\mathbf{y} \tag{10.6}$$

は係数ベクトルの最小二乗推定量 ($p\times 1$ ベクトル) である。行列 $\mathbf{Z}$ =(第 i 変数の j 番目の観測値) の階数を k する仮定 ((A3)) すると解は一意的に存在する。さらに $(\mathbf{y} - \mathbf{Z}\mathbf{b})^{'}\mathbf{Z}$ という性質を利用すると

$$L_2 = \|(\mathbf{y} - \mathbf{Z}\mathbf{b})\|^2 + \|\mathbf{Z}(\mathbf{b} - \boldsymbol{\beta})\|^2 \tag{10.7}$$

と分解できるのでベクトル $\mathbf{b} = (b_i)$ により誤差関数が最小化される。ここで射影行列を利用すると

$$\mathbf{y} = [\mathbf{Z}(\mathbf{Z}^{'}\mathbf{Z})^{-1}\mathbf{Z}^{'}]\mathbf{y} + [\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^{'}\mathbf{Z})^{-1}\mathbf{Z}^{'}]\mathbf{y} \tag{10.8}$$

という分解が得られ、$\hat{\mathbf{y}} = \mathbf{Z}\mathbf{B} = (\hat{y}_i)$ は理論値ベクトル、$\hat{\mathbf{u}} = [\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^{'}\mathbf{Z})^{-1}\mathbf{Z}^{'}]\mathbf{y} = (\hat{u}_i)$ は残差ベクトルに対応する。ここで観測値の変動和を理論値と残差により

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 + \sum_{i=1}^{n}\hat{u}_i^2 \tag{10.9}$$

と分解できる。ここで全変動 $(y_i - \bar{y})^2$ と回帰で説明できる回帰変動和 $\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2$ との比

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10.10}$$

により推定した回帰モデルの適合度 (決定係数) を定義しよう。また評価関数 L_1 を $\mathbf{b} = (B_I)$ により微分することで正規方程式

$$\mathbf{Z}^{'}\mathbf{Z}\mathbf{b} = \mathbf{Z}^{'}\mathbf{y}$$

の解としても理解できる。ここで (重) 回帰モデルを代入すると最小二乗推定量は

$$\mathbf{b} = (\mathbf{Z}^{'}\mathbf{Z})^{-1}\mathbf{Z}^{'}[\mathbf{Z}\boldsymbol{\beta} + \mathbf{u}] = \boldsymbol{\beta} + (\mathbf{Z}^{'}\mathbf{Z})^{-1}\mathbf{Z}^{'}\mathbf{u}$$

と表現される。右辺の第二項 (ベクトル) の各要素の期待値はゼロなので、誤差項に関する (A1),(A2) の下では最小二乗推定量の期待値は

$$\mathbf{E}[\mathbf{b}] = \boldsymbol{\beta}\ , \tag{10.11}$$

説明変数を (実験計画で典型的な分析者が設定できる) 制御変数とすると共分散行列は

$$\mathbf{E}[(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})'] = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1} \tag{10.12}$$

となる。推定量のクラスとして被説明変数の線形な範囲に限定すると最小二乗推定量はガウス・マルコフの定理と呼ばれる次の性質がある。

定理 10.1;: 線形回帰モデルにおいて (A1),(A2) および説明変数は (A3) を満足する固定された制御変数であることを仮定する。このとき係数推定量として不偏性を満たすように被説明変数の線形結合 $\mathbf{c}_j'\mathbf{y}$ $(j = 1, \cdots, k)$ をとると $\mathbf{b}$ の各要素の分散よりも小さくない。

この定理は最小二乗推定量は最良不偏推定量 (best linear unbiased estimator) であることを意味する。誤差項の分散 σ^2 の不偏推定量は

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i^2$$

で与えられるが、$n-k$ は自由度と呼んでいる。回帰係数の推定誤差の評価ではこの不偏推定量が用いられるが、説明変数が定数項の時のみの場合には $(k=1)$ 通常の不偏分散に対応している。
ここでさらに誤差項の確率分布について条件

$$(A4) \qquad \text{「誤差項は正規分布にしたがう」}$$

を仮定できれば、t 分布や F 分布を用いて統計的推定や統計的検定、さらに区間推定などの統計的推測論を適用することができる。

10.2 変数誤差と直交回帰

二次元空間上の n 個の点を $\mathbf{y}_i = (y_{1i}, y_{2i})'; (i = 1, \cdots, n)$ として直線を当てはめる問題では各点から直線上の点 $(a + b\xi_i, \xi_i)(\xi_i$ は未知) への距離の二乗和

$$L_d(a,b) = \sum_{i=1}^{n}(y_{1i} - a - b\xi_i)^2 + \sum_{i=1}^{n}(y_{2i} - \xi_i)^2 \tag{10.13}$$

で与えられる。そこで $L_d(a,b)$ を最小化するように係数 a,b を定める方法を考察しよう。この問題を図 10-1 に示しておくが線形回帰モデルと異なり二次元の各観測点 (x_i, y_i) $(i = 1, \cdots, n)$ より (最小二乗法のように x 軸へではなく) 直線上の点 (ξ_i, η_i) $(\eta_i = a + b\xi_i)$ への距離の二乗和を最小化する問題である。

(図 10-1：二次元データの直交回帰)

まず母数 a について最小化すると右辺の第二項より解は条件

$$a^* = \bar{y}_1 - b\bar{\xi}, \quad \bar{\xi}^* = \frac{1}{n}\sum_{i=1}^{n}\xi_i, \bar{y}_1 = \frac{1}{n}\sum_{i=1}^{n} y_{1i}$$

を満足する必要がある。次に直線上の点を表す ξ_i $(i = 1, \cdots, n)$ を母数として、関数 L_d を ξ_i について偏微分すると

$$\left(-\frac{1}{2}\right)\frac{\partial L_2}{\partial \xi_i} = y_{2i} - \xi_i^* + b(y_{1i} - a^* - b\bar{\xi}_i^*) = 0 \quad (i = 1, \cdots, n)$$

を得る。この式の両辺の平均をとれば条件 $\bar{\xi}^* = \bar{y}_2$ が導かれる。そこで既に得られた条件 $a^* = \bar{y}_1 - b\bar{\xi}$ を利用すると条件 $y_{2i} - \bar{y}_2 + b(y_{1i} - \bar{y}_1) - (\xi^* - \bar{\xi})(1 + b^2) = 0$ より

$$\xi_i^* - \bar{\xi}^* = \frac{y_{2i} - \bar{y}_2 + b(y_{1i} - \bar{y}_1)}{1 + b^2} \tag{10.14}$$

が得られる。この関係を $L_2(\cdot)$ に代入し整理すると、

$$\begin{aligned} L_2 &= \sum_{i=1}^{n}\left\{(y_{2i} - \bar{y}_2) - \frac{y_{2i} - \bar{y}_2 + b(y_{1i} - \bar{y}_1)}{1 + b^2}\right\}^2 + \sum_{i=1}^{n}\left\{y_{1i} - \bar{y}_1 - b\frac{y_{2i} - \bar{y}_2 + b(y_{1i} - \bar{y}_1)}{1 + b^2}\right\} \\ &= \sum_{i=1}^{n}\frac{[(y_{1i} - \bar{y}_1) - b(y_{2i} - \bar{y}_2)]^2}{1 + b^2} \end{aligned}$$

と整理される。(ここで最小二乗法では分子のみの最小化であることに注意する。) ここで分子の二乗を展開して整理すると

$$\begin{aligned} \frac{1}{n}L_2 &= \frac{s_{11} - 2bs_{12} + b^2 s_{22}}{1 + b^2} \\ &= \frac{(1, -b)\begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}\begin{pmatrix} 1 \\ -b \end{pmatrix}}{(1, -b)\begin{pmatrix} 1 \\ -b \end{pmatrix}} \end{aligned}$$

であるが、$s_{11} = (1/n)\sum_{i=1}^{n}(y_{1i} - \bar{y}_1)^2$, $s_{12} = (1/n)\sum_{i=1}^{n}(y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)$, $s_{22} = \frac{1}{n}\sum_{i=1}^{n}(y_{2i} - \bar{y}_2)^2$ とおいた。さらに母数 b について関数 L_2 を微分してゼロとおけば評価関数を最小化する解は

$$b^* = \frac{s_{11} - s_{22} + \sqrt{(s_{11} - s_{22})^2 + 4s_{12}^2}}{2s_{12}} \tag{10.15}$$

となる。

なおこの直交回帰の問題は線形代数における二次形式の比の最小化問題に帰着する。二次元ベクトル $\mathbf{c} = (1/\sqrt{1+b^2}, -b/\sqrt{1+b^2})$ と置くと、比の最小化問題は条件付最小化問題

$$\min_{\mathbf{c}} \mathbf{c}' \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \mathbf{c} \quad (\text{制約条件}) \quad \mathbf{c}'\mathbf{c} = 1$$

と同等である。この問題を条件付最適化問題でよく知られているように、ラグランジュ乗数 λ として関数 $L_2^* = \mathbf{c}'\mathbf{S}\mathbf{c} - \lambda(\mathbf{c}'\mathbf{c})$ を最小化する解 $\mathbf{c}$ は行列

$$\mathbf{S}_y = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = \sum_{j=1}^{n}(\mathbf{y}_j - \bar{\mathbf{y}}_n)(\mathbf{y}_j - \bar{\mathbf{y}}_n)' \tag{10.16}$$

の最小固有値に対応する固有ベクトルである。

ここで説明した議論は変数 $\mathbf{y}$ がベクトルの場合にもそのまま拡張することができる。スカラー観測値 y_{1i}, ベクトル観測値 $\mathbf{y}_{2i}$ と未知母数の系列 $\boldsymbol{\xi}_i$ の次元を $k = p - 1$ 次元とすると、同様の議論から評価関数

$$L_2 = \sum_{i=1}^{n} \frac{[(y_{1i} - \bar{y}_1) - \mathbf{b}'(\mathbf{y}_{2i} - \bar{\mathbf{y}}_2)]^2}{1 + \mathbf{b}'\mathbf{b}} \tag{10.17}$$

の最小化により得られるベクトル $\mathbf{b}$ が直交回帰の解であるが、これも固有値問題の解となっている。

10.3 線形関数関係と構造方程式

直交回帰の問題を観察可能な確率変数列 (Y_{1i}, Y_{2i}) $(i = 1, \cdots, n)$ が観測誤差 (V_{1i}, V_{2i}) を含み

$$Y_{1i} = \eta_i + V_{1i}\ ,\ Y_{2i} = \xi_i + V_{2i}; (i = 1, \cdots, n) \tag{10.18}$$

と表現する。確率的誤差 U, V をそれぞれ持つ変数間の期待値 $(\mathbf{E}[Y_{1i}] = \eta_i, \mathbf{E}[Y_{2i}] = \xi_i)$ の意味での**線形関数関係** (linear functional relationship)

$$\eta_i = \alpha + \beta\xi_i \ (i = 1, \cdots, n) \tag{10.19}$$

の推測問題と解釈することが可能である。ここで (V_{1i}, V_{2i}) は互いに独立な確率変数列で $\mathbf{E}[V_{1i}] = \mathbf{E}[V_{2i}] = \mathbf{E}[V_{1i}V_{2i}] = 0$, $\mathbf{E}[V_{1i}^2] = \mathbf{E}[V_{2i}^2] = \sigma^2$ とするのが標準的であるが、この種の統計モデルは一般に変数誤差 (errors-in-variables) モデルと呼ばれている。より一般に Y_{1i} をスカラー変数,$\mathbf{Y}_{2i}$ を $p-1$ $(p \geq 2)$ ベクトル変数としよう。観察可能な G 変数ベクトル $\mathbf{Y}_i = (Y_{1i}, \mathbf{Y}_{2i}')'$ の間に観察されない k-1 個の因子 $(k \leq p)\xi$ 因子により

$$\begin{bmatrix} Y_{1i} \\ \mathbf{Y}_{2i} \end{bmatrix} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\xi}_i + \begin{bmatrix} V_{1i} \\ \mathbf{V}_{2i} \end{bmatrix} \tag{10.20}$$

と表現すると因子分析モデルに対応する。ここで左辺の $p \times 1$ ベクトル $\mathbf{Y}_i$ は観測可能、$\boldsymbol{\xi}_i$ は観測不能な因子ベクトル、$\boldsymbol{\Lambda}$ は係数行列 (負荷行列と呼ばれる), $\mathbf{V}_i = (V_{1i}, \mathbf{V}_{2i}')$ は誤差ベクトルである。ここで特に $p = G = 2$ であり

$$\boldsymbol{\Lambda} = \begin{bmatrix} \beta \\ 1 \end{bmatrix}$$

とおけば変数誤差モデルの表現となっている。すなわち観察可能な次元よりも少ない次元 (今は 1) の観測できない因子 (unobserved factor) を推定する問題に対応している。この問題を一般的に考察すると、線形関係の分析を巡る様々な統計的方法を統一的に理解する道が開ける。

例 10.1 因子分析モデル

心理学や教育学などでは多次元の観測データよりデータの次元より少ない (観測不能な) 因子を推定する問題がしばしば考察されている。例えば英語、数学、国語、理科、社会 etc. などの各科目の背後に能力を示す (観測不能な) 因子が存在し、実際の科目の得点は「因子の線形結合プラス各人に起因するノイズ」などとしばしば解釈される。この場合には p 次元データより因子負荷行列 $\boldsymbol{\Lambda}$, $k-1$ $(k-1 \leq p)$ の因子,p 次元の個別ノイズを一意に定めることができないことが多いのでデータから統計的線形モデルを識別 (identified) する条件が必要となる。

例 10.2 需要と供給

応用経済学では市場で観察される価格と数量は需要関数と供給関数の均衡で決定するという考え方をよく用いる。ここで需要関数

$$q^d = \alpha_0 + \alpha_1 p + (\text{他の需要サイドの要因}, z_1) + u^d, \tag{10.21}$$

供給関数

$$q^s = \beta_0 + \beta_1 p + (\text{他の供給サイドの要因}, z_2) + u^s \tag{10.22}$$

とすると、供給量 $q = q^d = q^s$ と価格 p(u^d と u^s は期待値ゼロの需要と供給の攪乱項を表現する) により実際のデータの変動を解釈できるだろうか？この場合には p と q はいずれも説明変数、被説明変数の区別をするのが困難であり、均衡状態では $q = \pi_{q0} + \pi_{q1}z_1 + \pi_{q2}z_2 + v^q$, $p = \pi_{p0} + \pi_{p1}z_1 + \pi_{p2}z_2 + v^p$, と書ける。(ただし π_{qi}, π_{pi} は係数,v^q, v^p は期待値ゼロの確率的誤差項とする。) このとき係数間の関係より構造方程式としての需要関数は

$$\mathbf{E}[q] = \alpha_0 + \alpha_1\mathbf{E}[q] + \mathbf{E}[(\text{他の需要サイドの要因}, z_1)] \tag{10.23}$$

という線形関数関係と同等である。ここで母数 α_i $(i = 0, 1), \beta_i$ $(i = 0, 1)$ などを変数 q, p, z_1, z_2 の観測されるデータから推定する問題は線形関係の推定問題、構造方程式の推定問題と呼ばれている。

経験尤度と一般化積率法

一般に p 次元データ・ベクトル $(p = G + K)$ $\mathbf{x}_i = (\mathbf{y}_i^{'}, \mathbf{z}_i^{'})^{'}$ $(i = 1, \cdots, n)$ が利用可能なとき確率変数ベクトルについての関係 (構造方程式、あるいは推定方程式と呼ぶ)

$$\boldsymbol{\beta}^{'}\mathbf{y}_i - \boldsymbol{\gamma}\mathbf{z}_{1i} = U_i \ (i = 1, \cdots, n) \tag{10.24}$$

を推定する問題を考えよう。ここで被説明変数 G 次元ベクトル $\mathbf{y}_i$, 説明変数 K 次元ベクトル $\mathbf{z}_i$ として、$\mathbf{z}_{1i}$ は $\mathbf{z}_i$ の一部の $K_1 \times 1$ 変数ベクトル $(K_1 \leq K)$, U_i は誤差項で期待値 $\mathbf{E}(U_i) = 0$, 分散 $\mathbf{V}(U_i) = \sigma^2$ である。特に変数 y_{1i} の係数を基準化に利用して $\boldsymbol{\beta} = (1, -\boldsymbol{\beta}_2)$ すると構造方程式

$$y_{1i} = \boldsymbol{\beta}_2^{'}\mathbf{y}_{2i} + \boldsymbol{\gamma}\mathbf{z}_{1i} + U_i \tag{10.25}$$

と表現することができる。ここで変数群 $\mathbf{z}_i$ は説明変数ベクトル、外生変数ベクトル、操作変数ベクトルなどと呼ばれているが直交条件

$$\mathbf{E}(U_i\mathbf{z}_i) = \mathbf{0} \tag{10.26}$$

を満たす。こうした線形関係を推定する方法としては (i) 変数ベクトル $\mathbf{z}_i$ に方程式を回帰する。(ii) 回帰残差を係数ベクトル $\mathbf{c} = (\boldsymbol{\beta}^{'}, \boldsymbol{\gamma}^{'})^{'}$ について

$$L_2^* = \quad \mathbf{c}^{'}(\mathbf{Y}, \mathbf{Z}_1)^{'}\mathbf{Z}(\mathbf{Z}^{'}\mathbf{Z})^{-1}\mathbf{Z}^{'}(\mathbf{Y}, \mathbf{Z}_1)\mathbf{c} \tag{10.27}$$

を最小化することが考えられる。ここで $\mathbf{Y} = (y_{ij})$ は $n \times G$ 行列, $\mathbf{Z} = (z_{ij})$ は $n \times K$ 行列, $\mathbf{Z}_1 = (z_{1.ij})$ は $n \times K_1(z_{1,ij}$ は z_{ij} の一部からなる) 行列とした。この基準関数を最小化する推定法は二段階最小二乗法と呼ばれている。
他方、基準関数として

$$R_2^* = \frac{\mathbf{c}^{'}(\mathbf{Y}, \mathbf{Z}_1)^{'}\mathbf{Z}(\mathbf{Z}^{'}\mathbf{Z})^{-1}\mathbf{Z}^{'}(\mathbf{Y}, \mathbf{Z}_1)\mathbf{c}}{\mathbf{c}^{'}(\mathbf{Y}, \mathbf{Z}_1)^{'}(\mathbf{Y}, \mathbf{Z}_1)\mathbf{c}}, \tag{10.28}$$

の最小化問題となる。この解は最小分散比推定法、あるいは制限情報最尤法と呼ばれている。
ここで説明している問題を計量経済学では操作変数 (instrumental variables) を用いて解釈することがよく行われている。被説明変数ベクトル $\mathbf{y}_j^{'} = (y_{1j}, \mathbf{y}_{2j}^{'})$ と操作変数 (i.e. 説明変数) ベクトル $\mathbf{z}_j$ を用いた直交条件 (orthogonality condition)

$$\mathbf{E}[\mathbf{z}_j(y_{1j} - (\mathbf{y}_{2j}^{'}, \mathbf{z}_{1j}^{'})\boldsymbol{\theta})] = \mathbf{0} \tag{10.29}$$

を観測データ $(\mathbf{y}_j^{'}, \mathbf{z}_j^{'})$ より推定する問題と解釈できる。
ここで

$$\mathbf{g}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{z}_j(y_{1j} - \mathbf{x}_j^{'}\boldsymbol{\theta})$$

と置き、評価関数として二次形式

$$\mathbf{J}(\boldsymbol{\theta}, \mathbf{W}_n) = \mathbf{g}_n(\boldsymbol{\theta})^{'}\mathbf{W}_n\mathbf{g}_n(\boldsymbol{\theta}) \tag{10.30}$$

を最小化する $\boldsymbol{\theta}$ を一般化積率 (GMM, generalized method of moments) 推定量 $\hat{\boldsymbol{\theta}}_n$ と呼んでいる。ここで最小化の加重行列 $\mathbf{W}_n$ の選択としては、$\boldsymbol{\theta}$ の初期推定値 $\hat{\boldsymbol{\theta}}_n^{(0)}$ より残差系列を第 1 段階として求め, 次に $\tilde{u}_j = y_{1j} - \mathbf{x}_j^{'}\hat{\boldsymbol{\theta}}^{(0)}$ より

$$\mathbf{W}_n = \left[\frac{1}{n}\sum_{j=1}^{n} \tilde{u}_j^2\mathbf{z}_j\mathbf{z}_j^{'}\right]^{-1} \tag{10.31}$$

を用いて構成する方法が二段階 GMM(two-step GMM) 法と呼ばれている。
他方、ベクトル列 $\mathbf{y}_j$ $(j = 1, \cdots, n)$ がある確率分布 $F(\mathbf{y})$ から独立な標本, $P(\mathbf{y}_j = \mathbf{y}) = p_j$ $(j = 1, \cdots, n)$ であるとしよう。このときノンパラメトリック尤度関数は

$$L(F) = \prod_{j=1}^{n} p_j \tag{10.32}$$

と表現できる。この尤度関数を経験確率としての制約条件 $p_j \geq 0, \sum_{j=1}^{n} p_j = 1$ の下で最大化すると解は $p_j = 1/n$ $(j = 1, \cdots, n)$ である。$\boldsymbol{\theta}$ を母数ベクトル、構造方程式モデルに表れる変数 $\mathbf{y}_j, \mathbf{z}_j$ をまとめて $\mathbf{x}_j = (\mathbf{y}_j^{'}, \mathbf{z}_j^{'})^{'}$ とおき、期待値操作をデータ上の分布、すなわち経験分布で置き換えると

$$\sum_{j=1}^{n} p_j\ \mathbf{z}_i[y_{1j} - (\mathbf{y}_{2j}^{'}, \mathbf{z}_j^{'})\boldsymbol{\theta}] = 0 \tag{10.33}$$

という制約条件得られる。この制約条件および経験確率としての条件 $p_j \geq 0, \sum_{j=1}^{n} p_j = 1$ の下でノンパラメトリック尤度関数 (10.32) を最大化して母数 $\boldsymbol{\theta}$ を推定する方法は

経験尤度 (EL, empirical likelihood) 推定と呼ばれている。ここで特に $p_1 = \cdots = p_n$ とすると分散比の最小化に対応する[46]。なお以上の説明において制約条件が必ずしも線形である必要がないことに着目すると、GMM 推定法や経験尤度法は非線形方程式にもそのまま適用することができる。

このように回帰分析や直交回帰分析の問題は内容的には観測ベクトルにそれぞれ誤差を伴い観察される状況は、**変数誤差の問題** (errors-in-variables)、あるいは計量経済学における応用として内生変数間の関係を検討する**構造方程式** (structural equation)・**同時方程式** (simultaneous equation) の問題などに関わっているのである。こうした応用において現れる適切な確率モデルを定式化したり、その統計的推測理論を展開することは重要な統計的問題である。

なおここで簡単に述べた線形関係の統計的推測は折にふれて説明したように 2 次元データの解析にとどまることなく一般の次元や非線形問題への一般化が可能である。

10.4 主成分分析と多変量解析

一般に p 次元確率変数 $\mathbf{X}_j = (X_{ij})$ $(j = 1, \cdots, n)$ が与えられた時に、データに含まれる情報を縮約する様々な統計的方法が考案されている。多次元データの統計的分析は統計的多変量解析 (statistical multivariate analysis) と呼ばれている。特に統計的線形関係の分析では主成分分析 (principal components analysis), 因子分析 (factor analysis), 回帰分析 (regression analysis), 変数誤差分析 (errors-in-variables analysis) などが重要である。

p 次元確率変数 $\mathbf{X}$ の期待値ベクトル $\mathbf{E}[\mathbf{X}] = \boldsymbol{\mu}$ $(p \times 1)$, 共分散行列 $\mathbf{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{'}] = \boldsymbol{\Sigma}$ $(p \times p)$, 共分散行列が正則であることを仮定するが、一般の次元 p が分かりにくければ $p = 2$ の場合を考えればよい。このとき各変数の期待値周りの加重和 $\sum_{i=1}^{p} \alpha_i (X_i - \mu_i)$ の変動をなるべく上手く表現できる座標を求めよう。ばらつきは分散なので線形和の分散

$$\mathbf{V}[\boldsymbol{\alpha}^{'}(\mathbf{X} - \boldsymbol{\mu})] = \mathbf{E}[\boldsymbol{\alpha}^{'}(\mathbf{X} - \boldsymbol{\mu})]^2 \tag{10.34}$$

を最大化する係数ベクトル $\boldsymbol{\alpha} = (\alpha_i)$ を考えよう。(例えば $p = 2$ の場合には平面上の 2 次元データの解析である。2 次元のデータを素直に解析しようとするとき、線形変換とはもとの座標を回転することに対応するので分散最大化の意味は明確である。) 各係数を定数倍しても結果は変わらないので基準化の為に $\boldsymbol{\alpha}^{'}\boldsymbol{\alpha} = 1$ としてラグランジュ形式

$$L = \boldsymbol{\alpha}^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha} - l[\boldsymbol{\alpha}^{'}\boldsymbol{\alpha} - 1] \tag{10.35}$$

[46]例えば国友 (2011)「構造方程式モデルと計量経済学」3 章 (朝倉) を参照。

を最大化する。各要素で微分すると

$$\frac{1}{2}\frac{\partial L}{\partial \boldsymbol{\alpha}} = [\boldsymbol{\Sigma} - l\mathbf{I}_p]\,\boldsymbol{\alpha} = \mathbf{0}$$

を解く問題となる。ここで l_1 を固有方程式

$$|\boldsymbol{\Sigma} - l\mathbf{I}_p| = 0 \tag{10.36}$$

を満たす最大根とすると、最大根に対応する固有ベクトル $\mathbf{c}_{(1)}$ が求める解であるが、これを第一主成分ベクトルと呼んでいる。第一主成分を $U_1 = \mathbf{c}_{(1)}\mathbf{X}$ とおけば第一主成分と直交する (無相関な) 成分の中で分散最大化する成分を第二主成分、同様に第 i 主成分ベクトル $(i = 2, \cdots, p)$ も定義できる。また主成分分析により逆に共分散行列を表現すると $p \times p$ の非負定符号行列の分解 $\boldsymbol{\Sigma} = \sum_{i=1}^{p} l_i \mathbf{c}_{(i)}\mathbf{c}'_{(i)}$ に対応する。そこで第 i 成分の全体変動への寄与度は $l_i / \sum_{j=1}^{p} l_j$ となる。

実際に観察される多次元標本を $\mathbf{X}_j = (X_{ij})$ $(i = 1, \cdots, p; j = 1, \cdots, n)$ とすると多次元データにおける標本平均は $\bar{\mathbf{X}}_n = (1/n)\sum_{j=1}^{n} \mathbf{X}_j$, 標本積率行列は

$$\mathbf{A}_n = \sum_{j=1}^{n} (\mathbf{X}_j - \bar{\mathbf{X}}_n)(\mathbf{X}_j - \bar{\mathbf{X}}_n)' \tag{10.37}$$

となるので、共分散行列 $\boldsymbol{\Sigma}$ の推定量は $\mathbf{S}_n = [1/(n-1)]\mathbf{A}_n$ となる [47]。標本共分散推定量より固有方程式

$$[\mathbf{S}_n - \lambda\mathbf{I}_p]\mathbf{a} = \mathbf{0} \tag{10.38}$$

を解くことにより標本の第 i 主成分ベクトル $\mathbf{a}_{(i)}$ $(i = 1, \cdots, p)$ を求めることができる。

統計的推測を行うには $\bar{\mathbf{X}}_n$, $\mathbf{A}_n$ の標本分布を導く必要があるが、標本平均はともかく標本分散・共分散行列はそれほど容易ではない。確率変数 $X_j \sim N(0,1)$ のとき $(p = 1)$ $A_n = v \sim \chi^2(\nu)$ $(\nu = n-1)$ であるので密度関数は定数 $c(1,1,\nu)$ を用いて

$$f(v) = \frac{1}{c(1,1,\nu)} v^{\nu/2-1} e^{-v/2} \tag{10.39}$$

で与えられる。ここでは χ^2 分布についてのよく知られた積分計算より

$$c(1,1,\nu) = \int_0^\infty v^{\nu/2-1} e^{-v/2} dv = 2^{\nu/2}\Gamma(\frac{\nu}{2})$$

である。確率変数ベクトル $\mathbf{X}_j$ $(= (X_{ij}))$ $\sim N_p(\mathbf{0}, \mathbf{I}_p)$ のとき $\mathbf{A}_n = \mathbf{V} \sim \text{Wishart}(p, \mathbf{I}_p, \nu)$ にしたがう $(\nu = n-1)$。このウイッシャート (Wishart) 分布の密度関数は

$$f(\mathbf{V}) = \frac{1}{c(p, \mathbf{I}_p, \nu)} |\mathbf{V}|^{(\nu-p-1)/2} e^{(-1/2)\text{tr}(\mathbf{V})} \tag{10.40}$$

[47] 正規分布の下では最尤推定量である。不偏推定量は n の代わりに $n-1$ を採用する。

で与えられる。ここで対称行列 $\mathbf{V}$(任意の i, j について $v_{ij} = v_{ji}$) が正定符号であることを注意深く積分計算を行うと結局

$$c(p, \mathbf{I}_p, \nu) = 2^{\nu p/2} \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma[\frac{\nu + 1 - i}{2}]$$

となる。一般に $\mathbf{X}_j \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ のときには、標本積率行列 $\mathbf{A}_n$ がしたがうウイッシャート分布の密度関数は

$$f(\mathbf{V}) = \frac{1}{c(p, \mathbf{\Sigma}, \nu)} |\mathbf{V}|^{(\nu-p-1)/2} e^{(-1/2)\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{V})} \tag{10.41}$$

であるが、定数は $c(p, \mathbf{\Sigma}, \nu) = |\mathbf{\Sigma}|^{n/2} c(p, \mathbf{I}_p, \nu)$ である。さらに標本積率行列 $\mathbf{A}_n$ の固有値・固有ベクトルについてもその標本分布を調べることができる [48]。

こうした主成分 (principal components) は多次元データの次元縮約の方法として通常の経済の実証分析では時々しか登場しないが、特に教育学・心理学・社会学などを中心として幅広い分野の応用において広く用いられている。他方、多次元データの中から潜在的なより小さい次元（数）の因子 (factor) により説明しようとする試みも応用の中では重要であるが因子を巡る議論の中には混乱も見られる。例えばここで説明した**主成分モデル**は**統計的因子モデル** (statistical factor analysis) と関係はあるが同一でない。こうした主成分分析や因子分析などの方法は**統計的多変量解析** (statistical multivariate analysis) と呼ばれている。

10.5 非線形問題と分位点回帰

統計的関係の分析は変数間の関係が線形ではない非線形関係の分析にも拡張することが可能である。例えば非線形回帰モデルとは、ある関数 $g(\cdot)$ を用いて

$$y_i = g(z_{1i}, \cdots, z_{k-1,i}, \beta_1, \cdots, \beta_r) + u_i \ \ (i = 1, \cdots, n) \tag{10.42}$$

と表現される。ここで被説明変数の第 i 個目の観測値 y_i, 第 j 番目 $(j = 0, \cdots, k-1)$ の被説明変数の第 i 個目の観測値 z_{ji}, (z_{01} は定数)、母数 $\beta_j (j = 1, \cdots, r)$、誤差項 u_i $(i = 1, \cdots, n)$ とする。最小二乗法は線形回帰分析と同様に定義すれば、明示的に解を求めることが困難でも（解が存在すれば）計算機を用いれば最適解を求めることができる [49]。なお非線形回帰関係は様々な可能性を含んでいるので必ずしも複雑な分析がより深い意味をもつとは限らないことに注意を要する。

[48] 標本共分散行列の固有値や固有ベクトルに関する統計的分析については Anderson (2003, An Introduction to Multivariate Statistical Analysis, Wiley) が詳しい。

[49] 例えば Amemiya (1985),*Advanced Econometrics* に詳しい説明がある。

非線形問題は多岐に及ぶがここで分位点回帰問題に言及しておこう。被説明変数Yの変動を説明するリスク要因として幾つかの説明変数があるとき、説明変数ベクトル$\mathbf{Z} = (Z_1, \ldots, Z_p)'$が与えられた時の確率変数$Y$の条件付分布関数を$\mathrm{P}(Y \leq y|\mathbf{Z}) = F_Y(y|\mathbf{Z})$，条件付$\tau$分位点を$Q_\tau(Y|\mathbf{Z}) = \inf\{y|F_Y(y|\mathbf{Z}) \geq \tau\}$とする．このとき分位点回帰モデルは

$$\begin{aligned} Q_\tau(Y|\mathbf{Z}) &= \alpha(\tau) + \beta_1(\tau)Z_1 + \cdots + \beta_p(\tau)Z_p \\ &= \alpha(\tau) + \mathbf{Z}'\boldsymbol{\beta}(\tau) \end{aligned}$$

と表現される。ここで$\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \ldots, \beta_p(\tau))'$は定数項を表す母数$\alpha(\tau)$を除く未知母数ベクトルであるが、確率変数$U$を$U = Y - \{\alpha(\tau) + \mathbf{Z}'\boldsymbol{\beta}(\tau)\}$により定義すれば，分位点回帰モデルは

$$Y = \alpha(\tau) + \mathbf{Z}'\boldsymbol{\beta}(\tau) + U \tag{10.43}$$

と表現できる．この形式は統計的線形回帰モデルに類似しているが，右辺の母数の係数ベクトルがτに依存し，誤差項Uの分布関数を$F_U(u)$とすると

$$F_U(0) = \mathrm{P}\left(Y \leq \alpha(\tau) + \mathbf{Z}'\boldsymbol{\beta}(\tau)\right) = \tau \tag{10.44}$$

となるので意味が異なる。線形回帰モデルが被説明変数の期待値を説明変数により表現しているが、分位点回帰モデルでは被説明変数の確率分布の分位点を表現している。例えば近年の経済分析では被説明変数の期待値ではなく、所得や賃金が高い人々と低い人々との間で異なる説明変数の効果があることを検出する応用例などが重要である。こうした例でその汎用性が理解できると思われるが、分位点回帰法は多くの応用分野で注目されているのである。

ここで被説明変数Yと説明変数ベクトル$\mathbf{z}$について互いに独立なn個のデータの組$(y_i, \mathbf{z}_i)$ $(i = 1, \ldots, n)$が得られる状況を想定しよう。y_iは被説明変数，$\mathbf{z}_i = (z_{i1}, \cdots, z_{ip})$ $(i = 1, \cdots, n)$は説明ベクトルであるが、条件$n \geq p + 1$が成立し，被説明変数の条件付分布関数はLebesgue測度に関して絶対連続の場合のみを考察する。この仮定の下では密度関数が存在するので分位点関数は分布関数の逆関数として一意的に定義でき、分析は簡単化される。

ここで被説明変数Yと説明変数ベクトル$\mathbf{x}$のn組のデータより母数ベクトル$((p+1) \times 1)$ $(\alpha(\tau), \boldsymbol{\beta}(\tau)')'$を推定する問題を考える[50]損失関数として

$$\rho_\tau(u) = \tau(u)_+ + (1-\tau)(u)_- = \begin{cases} \tau u, & u \geq 0, \\ (\tau - 1)u, & u < 0 \end{cases}$$

[50]Koenker,R. and Bassett, G. (1978),"Regression Quantiles," *Econometrica*, 46, 33-50. Koenker, R. (2005),*Quantile Regression*, Cambridge UP に詳しい説明がある。

を用いるが,$(u)_+$，$(u)_-$ はそれぞれ u の正・負の部分を表す. 特に $\tau = 1/2$ のときには $\rho_{1/2}(u) = |u|/2$ となるので, 絶対値偏差の拡張になる。n 組のデータより評価基準

$$\min_{\{\alpha,\boldsymbol{\beta}\}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \mathbf{z}_i'\boldsymbol{\beta}) \tag{10.45}$$

を最小化する方法において解を $(\widehat{\alpha}(\tau), \widehat{\boldsymbol{\beta}}(\tau)')'$, 定数項を含む説明変数ベクトルのデータを $\mathbf{z}_i^* = (1, \mathbf{z}_i')'$ $(i = 1, \ldots, n)$ と表しておく。

分位点回帰モデルにおいて母係数ベクトルの推定がまず解決すべき統計的問題であるが、母係数ベクトルを $\boldsymbol{\delta}(\tau) = (\alpha(\tau), \boldsymbol{\beta}(\tau)')'$，推定量ベクトルを $\widehat{\boldsymbol{\delta}}(\tau) = (\widehat{\alpha}(\tau), \widehat{\boldsymbol{\beta}}(\tau)')'$ とする。このとき分位点回帰推定量は標本数 n が大きいときには一致性 (consistency) と漸近正規性 (asymptotic normality) を持つ。推定量の漸近的性質の分析は評価関数が非線形の場合には一般に複雑になるが，分位点回帰問題の場合には確率変数列 $(y_i, \mathbf{x}_i)$ $(i = 1, \cdots, n)$ について次のような条件

(A5) 確率変数列 $(y_i, \mathbf{z}_i)$ $(i = 1, \cdots, n)$ は互いに独立で同一分布 (i.i.d.) にしたがう.

(A6) $\mathbf{z}_i$ を所与とする u_i の条件付分布関数 $F_U(\cdot|\mathbf{z}_i)$ は ($\mathbf{z}_i$ に依存しない) 原点の近傍上で正の密度関数 $f_U(\cdot|\mathbf{z}_i)$ を持ち、この近傍上で $s \mapsto f_U(s|\mathbf{z}_i)$ は ($\mathbf{z}_i$ について一様に) 連続となる.

(A7) 行列 $\mathbf{C} = \mathbf{E}[\mathbf{z}_i^*\mathbf{z}_i^{*\prime}]$ は正定値行列となる.

(A8) 行列 $\mathbf{D} = \mathbf{E}[f_U(0|\mathbf{z}_i)\mathbf{z}_i^*\mathbf{z}_i^{*\prime}]$ は正定値行列となる.

を仮定しよう。これらの条件の下で分位点回帰推定量の漸近的性質について次のような結果が成り立つ。

定理 10.2： 分位点回帰推定量 $\widehat{\boldsymbol{\delta}}(\tau)$ について $n \to \infty$ のとき

(i) 条件 (A5)〜(A7) のもとで

$$\widehat{\boldsymbol{\delta}}(\tau) = \begin{bmatrix} \widehat{\alpha}(\tau) \\ \widehat{\boldsymbol{\beta}}(\tau) \end{bmatrix} \xrightarrow{p} \boldsymbol{\delta}(\tau) = \begin{bmatrix} \alpha(\tau) \\ \boldsymbol{\beta}(\tau) \end{bmatrix} \tag{10.46}$$

となる。

(ii) 条件 (A5)〜(A8) のもとで

$$\sqrt{n}\{\widehat{\boldsymbol{\delta}}(\tau) - \boldsymbol{\delta}(\tau)\} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \tau(1-\tau)\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}) \tag{10.47}$$

が成立する.

なお以上の説明では説明変数 $\mathbf{z}_i$ が確率的である場合を議論したが $\mathbf{z}_i$ が非確率的変数である場合にも同様の議論が可能である。特に条件

$$(A8)' \qquad \text{「正定値行列 C が存在し，} \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathbf{z}_i^*\mathbf{z}_i^{*'} = \mathbf{C} \text{ となる」}$$

の下では漸近分布は

$$\mathcal{N}(\mathbf{0}, \tau(1-\tau)\{f_U(0)\}^{-2}\mathbf{C}^{-1}) \tag{10.48}$$

と表現できる。なお実際に極限分布を利用する場合には原点における密度関数を推定することが必要となる [51]。

10.6 罰則法と情報量基準

線形回帰モデルは広範に応用されているが、実際のデータ解析では説明変数の選択問題が重要である [52]。データ分析では特定の被説明変数に対して多くの説明変数の候補があり得るのでどの変数を利用したらよいか様々な方法が提案されている。多くの場合には応用上で意味があると考えられる変数についての統計的有意性、回帰式の適合度などが利用されている。特に説明変数が多い場合にはこれらの基準に加えて幾つかの統計的基準が利用されている。

例えば Ridge 回帰法を取りあげてみよう。この Ridge 回帰は元々は説明変数行列が線形独立でない場合にも回帰分析を行うことができるという実用的な目的に為に提案されたものである。この方法は Ridge 回帰法は罰則付き最小二乗法として数理的にはある実数 t を指定したときの最小化問題

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \alpha - \mathbf{z}_i'\beta)^2 \quad \text{(制約条件)} \ \sum_{j=1}^{p}\beta_j^2 \le t, \quad t \ge 0$$

と定式化できる。これに対して Lasso (least absolute shrinkage and selection operator) 法 [53] では定数項以外のパラメータに絶対値 (L_1) ノルムの罰則をつけて最小二

[51] 例えば漸近分布の導出、漸近分散の推定法、損害保険データへの応用について、加藤賢悟・国友直人・増田智巳「Lasso 分位点回帰の理論と損害保険への応用」日本統計学会誌 (2009),121-149 が議論している。誤差分布に正規分布を仮定できれば例えば残差系列 $\hat{U}_i = Y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}(\tau)$ $(i = 1, \cdots, n)$ より $s(\tau) = [f_Y(F_Y^{-1}(\tau))]^{-1}$ $(= [f_U(0)]^{-1})$ の推定量として $\hat{s}_n(\tau) = [F_n^{-1}(\tau + h_n] - F_n^{-1}(\tau - h_n)]/[2h_n]$, $h_n = c_\tau n^{-1/3}(c_\tau = z_\alpha^{2/3}[1.5\phi(\Phi^{-1}(\tau))^2/(2\Phi^{-1}(\tau)^2 + 1)]^{1/3}$ (z_α は α $(= 1 - \tau)$ パーセント点) とすればよい。

[52] 回帰分析に関連する様々な数理統計的な問題については例えば佐和隆光 (1979)「回帰分析」朝倉書店が有用である。

[53] Tibshirani (1996), "Regression shrinkage and selection via lasso" *Journal of Royal Statistical Society*, 58, 267-288.

乗推定を適用する。すなわち Lasso 推定量は最小化問題

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \alpha - \mathbf{z}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \quad \lambda \geq 0 \tag{10.49}$$

の最適解として定義される。ここで λ はチューニング・パラメータと呼ばれているが、この式を λ 形式の Lasso と呼ぶ。λ 形式に対して t 形式の Lass を定義すると t 形式の Lasso 推定量とは最小化問題問題

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \alpha - \mathbf{z}_i'\boldsymbol{\beta})^2 \ \text{(制約条件)} \ \sum_{j=1}^{p} |\beta_j| \leq t, \quad t \geq 0$$

の最適解として定義される。なおこうした λ 形式と t 形式という二つの最小化問題は数学的には同値である。Lasso 回帰とは t あるいは λ をまず適当に定めて最小化問題を解き、次に t あるいは λ を変化させ統計的に有意になる説明変数を探索する方法であり、Lasso 選択ではある所まで各変数の係数推定値がゼロになることが特長であり、応用上では有効となることが少なくないことが知られている。

こうした罰則付き最小化による説明変数の選択方法は多くの応用問題において有用であるが、推定した回帰モデルのフィットの良さ (適合度) と未知母数の推定により生じるコストについてのある種のトレード・オフを統計的な手続きとして定式化している、と解釈することが出来る。例えば線形回帰モデルの場合には $k = n$ とすると推定誤差がゼロとすることが可能であるが、そうして推定した回帰モデルを用いて予測すると結果は不安定になる可能性が高い。この問題を解決する手段として多くの場合に情報量規準が用いられる。代表的な情報量規準として知られている赤池の情報量規準 AIC は元々は 11 章で説明する時系列モデルにおける 1 期先予測の誤差を小さくすると云う、予測誤差の観点からの説明変数の選択基準として考えられた。AIC は p 個の母数からなる母数ベクトル $\theta = (\theta_i)$, その最尤推定量 $\hat{\theta}_{ML}$, 尤度関数 $L_n(\theta)$ とすると

$$AIC = -2\log L(\hat{\boldsymbol{\theta}}_{ML}) + 2k \tag{10.50}$$

により定義される。この AIC 統計量を最小化する方法が AIC 規準であるが、第一項はデータへの統計モデルの良さを表し、第二項は統計モデルの良さ (すなわち単純さ) を表しているので AIC はこの二つの要素を組み合わせた規準であり、線形回帰モデルにおける説明変数の選択問題 (データとして利用可能な説明変数のリストから定数項を含めて k 個の適切な説明変数を選ぶ問題) を含め、尤度関数が与えられる場合の多くの統計的問題で有用である。

なお AIC 規準の第二項をデータ数 n に依存させ、母数の数に対する罰金 (ペナルティ) 項をより大きくとり $BIC = -2\log L(\hat{\boldsymbol{\theta}}_{ML}) + k\log(n)$ を最小化する規準はベイズ情報量規準 (あるいはシュワルツ規準) と呼ばれている。モデル選択の問題では AIC や BIC が一般的にしばしば用いられている。

11　広がる統計解析の世界

11.1 統計的時系列解析

数理統計学における標準的理論では互いに独立な標本の実現値として標本が得られることを想定して展開されることが多い。他方、工学・薬学・医学・経済・経営・金融をはじめとする応用の場で得られる時系列データではある母集団から互いに独立に得られる標本の実現値、と見なすことが非現実的な場合も少なくない。そこで実際の時系列データの統計分析と標準的理論のギャップを巡る統計的方法、統計的時系列分析 (statistical time series analysis) [54] が開発されている。

独立な確率変数列についての数理的理論は確率変数が定常過程 (stationary process) であれば軽微な修正で成り立つことが多い。一例としては 5 章で言及した時系列の中心極限定理 (定理 A-3) が典型的であるが、別の例としては 4 章で挙げた (離散) マルチンゲールを挙げることができる。独立確率変数の和はマルチンゲールであるからマルチンゲール差分は独立な確率変数列の一般化とみることができる。このことから独立確率変数列の和の挙動の理論は独立とは限らない確率変数列にもある範囲内で適用できることが分かる。こうした離散マルチンゲールにもとづく議論は近年の時系列データの分析では一つの基本的な道具を提供している。

時系列データの分析では 10 章で述べたような統計的関係の分析を適用することが重要である。例えば線形回帰モデルにおいて説明変数として被説明変数について有限個の過去の値をとると p 次自己回帰モデル (autoregressive model, 略して AR(p))

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j Y_{i-j} + U_i \ \ (i = 1, \cdots, n) \tag{11.1}$$

が得られる。ここで $\{U_i\}$ は互いに無相関な誤差項で $\mathbf{E}[U_i] = 0, \mathbf{E}[U_i^2] = \sigma^2$、$\beta_j$ $(j = 0, \cdots, p)$ は係数母数である。通常の回帰モデルとの相違は初期値 $Y_0, y_{-1}, \cdots, Y_{-(p-1)}$ が定まらないと確率変数列 $\{Y_i\}$ の分布が定まらないことである。ここでは簡単化のために初期値 $Y_0, Y_{-1}, \cdots, Y_{-(p-1)}$ がある値をとるときの条件付分布を考えるとしておこう。このとき確率変数列 (すなわち時系列) が定常的な挙動を示す為には方程式

$$|\lambda^p - \sum_{j=1}^{p} \lambda^{p-j} \beta_j| = 0 \tag{11.2}$$

[54] ここでは統計的時系列の初歩について例えば統計的時系列分析の基礎について山本拓「経済の時系列分析」(創文社)、沖本竜義「計量時系列分析」(朝倉書店) Anderson(1971), *The Statistical Analysis of Time Series*, Wiley, Brockwel and Davis (1991), *Time Series : Theory and Methods*, Springer、などを挙げておく。

の根の絶対値が 1 より小さくなる必要がある、さらに確率変数列 $\{Y_i\}$ が厳密に定常分布を持つには、初期値がとる確率分布を対応させる必要がある。線形時系列モデルとしては U_i を互いに無相関な誤差項 ($\mathbf{E}[U_i] = 0, \mathbf{E}[U_i^2] = \sigma^2$)、$\theta_j$ $(j = 0, \cdots, q)$ を係数母数として

$$Y_i = \theta_0 + \sum_{j=1}^{q} \theta_j U_{i-j} \quad (i = 1, \cdots, n) \tag{11.3}$$

と表現される確率モデルが q 次移動平均過程 (moving-average process, 略して MA(q)) と呼んでいる。
こうした線形時系列モデルの一般形として

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j Y_{i-j} + \sum_{j=1}^{q} \theta_j U_{i-j} \quad (i = 1, \cdots, n) \tag{11.4}$$

としたのが自己回帰移動平均 (autoregressive moving-average) 過程であり一般に ARMA(p,q) と表記している。

これらの統計的モデルを実際に観測される時系列データの解析に利用するには未知母数を推定する必要があるが、最小二乗推定法や最尤推定法などを適用することが考えられる。母数の統計的推定、検定、さらには利用可能な時系列データ $Y_i = y_i$ $(i = 1, \cdots, n)$ から将来値 Y_{n+h} $(h \geq)$ の予測の方法などが統計的時系列解析では標準的な方法として開発されている。現時点を t とすると予測問題とは過去・現在までに利用できる情報を用いて確率変数の実現値としての将来値を推定する問題である。例えば利用可能な情報が $Y_s = y_s$ $(s = 1, \cdots, t)$ で与えるとすると、3.3 節の議論を利用すると Y_{t+h} $(h \geq 1)$ の最適予測 $Y_{t+h|t}$ は条件付期待値

$$Y_{t+h|t} = \mathbf{E}[Y_{t+h}|Y_t, Y_{t-1}, \cdots, Y_1] \tag{11.5}$$

で与えられる。統計モデルである自己回帰移動平均モデルを利用するとこうした予測量は容易に構成することが可能となる。実際には次数 p, q などは未知であるので適切な次数を選ぶ必要があり、前章の (10.50) で説明した AIC 最小化基準を利用すればよいが、実は AR(p) モデルの次数選択問題において誤差分散の推定量 $\hat{\sigma}^2$ とするとき 1 期先予測誤差の MSE

$$FPE(p) = n \log \hat{\sigma}^2 + 2p$$

(ここで n はデータ数) の最小化基準とて提案されたものが AIC 最小化基準である。
一般に時系列データの場合には確率変数が互いに独立とはみなせないので小標本理論の展開は困難なので、大標本理論に基づく議論が必要となり分析はかなり複雑

になる。例えば 10 章のガウス・マルコフの定理はそのままでは成立しないので一致性および漸近正規性の議論が必要となる。なお線型モデルを更に発展して多次元時系列モデル、さらに様々な非線型モデルや定常性条件を満たさない非定常な時系列の解析に適用できる統計理論が統計的時系列解析の分野では展開している。時系列解析では自己回帰モデルや移動平均モデルと言った時間領域分析に加えて周波数領域の分析も発展しているがそこでは密度関数の推定で利用されたカーネル法を用いて時系列データに基づくスペクトル分析が有力な手段である。

実際に得られる時系列データ y_i $(i = 1, \cdots, n)$ が経済時系列のように定常過程の実現値とはみなしにくい場合、しばしば階差系列 $x_i = \Delta y_i = y_i - y_{i-1}$ $(i = 2, \cdots, n)$ あるいは季節階差系列 $x_i = \Delta_s y_i = y_i - y_{i-s}$ $(i = s+1, \cdots, n; s \geq 2)$ に ARMA(p,q) モデルを当てはめるて解析することが少なくない。階差操作を二回繰り返すと $x_i = \Delta^2 y_i = \Delta y_i - \Delta y_{i-1}$ $(i = 3, \cdots, n)$ となり $x_i = \Delta^d$ $(d = 2)$ に対応するが、季節階差では月次データ $s = 12$, 四半期データ $s = 4$ などが利用される。階差操作の逆は和分と呼ばれているので階差データに ARMA(p,q) をフィットさせることを自己回帰和分移動平均 (autoregressive integrated moving-average, 略して ARIMA(p,d,q)) モデルを適用していると云う。

11.2 統計的生存時間解析

生存時間解析

生存時間の統計的解析は生命表と呼ばれる人間の寿命の分析に端を発し、工学における機械や製品の故障・品質、すなわち統計的品質管理における統計的分析法として重要である。こうした機械の故障や寿命に関するリスク評価の研究分野は信頼性工学と呼ばれているが、その後、医学や薬学の問題に応用されるようになり、より広い生存時間解析 (survival analysis) として多くの統計的方法が応用されるようになっている。近年では更に経済・経営・金融などの応用分野においてさえも生存時間解析が利用されるようになり、例えば企業の倒産 (default) や社債の評価などで重要な役割を果たすようになっている。

生存関数

寿命を表す確率変数を $T(\omega)$, その確率分布関数を $F(t)$, 密度関数を $f(t)$ とする。ここで寿命や故障の時刻は非負であるから正値をとる分布を考える必要がある。生存関数 (survival function) は

$$S(t) = 1 - F(t) = P(T > t) \tag{11.6}$$

により定義する。ハザード関数 $\lambda(t)$ は

$$\begin{aligned}\lambda(t) &= \frac{f(t)}{1-F(t)} \\ &= \lim_{h\to 0}\frac{P(t<T\leq t+h|T>t)}{h}\end{aligned}$$

で与えられる。したがってハザード (危険) 率は時刻 t において生存しているという条件の下で期間 $(t, t+h]$ に寿命が来る確率の極限であるから

$$\int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{1-F(s)}ds = -\log S(t)$$

より

$$S(t) = e^{-\int_0^t \lambda(s)ds} \tag{11.7}$$

である。

例 11.1： ハザード率が時間に依存しない場合は指数分布であるが、密度関数は

$$f(t) = \lambda e^{-\lambda t} \tag{11.8}$$

で与えられる。Weibull 分布は

$$S(t) = e^{-(\lambda t)^{\alpha}} \tag{11.9}$$

により特長づけられる。($\alpha > 0, \lambda > 0$ である。)

例 11.2： Cox モデル、あるいは比例ハザードモデルと呼ばれているハザード関数の表現が有用である。この統計モデルではハザード関数が基準ハザード関数 $\lambda_0(t)$, 説明変数ベクトル $\mathbf{z}$, 係数ベクトル $\boldsymbol{\beta}$ とすると

$$\lambda(t) = \lambda_0(t)\exp[\mathbf{z}'\boldsymbol{\beta}] \tag{11.10}$$

で与えられる。観測データ $\mathbf{x}_i = (y_i, \mathbf{z}_i')'$ $(i = 1, \cdots, n)$, y_i は生存データ) より統計モデルを推定する方法が開発されている。

11.3 統計的極値解析

リスク管理と極端な事象

統計学の入門では確率変数 X がある範囲 $(a, b]$ に入る確率を計算する方法として正規分布を利用して

$$\mathbf{P}\left(a < X \leq b\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_a^b e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{11.11}$$

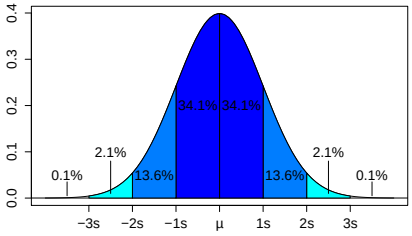

図 11-1:正規分布と正規確率

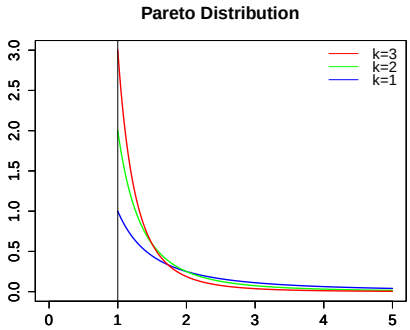

図 11-2:パレート分布と裾

と評価する統計的方法が説明されている。この種の計算では $[\mu - k\sigma, \mu + k\sigma]$ の確率は $k = 1, 2, 3$ に対し 68%,95%,99.9% などである。区間外の確率は $k = 5$ では 10^{-6},$k = 6$ では 10^{-23} などとなり裾確率は非常に小さいのが特徴である。例えば標本算術平均 $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ とすると中心極限定理

$$\frac{\sqrt{n}}{\sigma}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0,1) \tag{11.12}$$

により正規分布の利用を正当化することが多い。正規分布はよく知られているように裾確率は図のように急速に減衰する。正規分布と好対照な確率分布としてパレート分布の密度関数 $f(x) = k/x^{k+1}$ $(x \geq 1)$ を図 11-1.11-2 に示しておく。

近年の統計的リスク分析では突発的な現象や極端な現象 (extreme events) を分析することも重要である。例えば伝統的な損害保険の分野では風水害、地震、事故など稀な事象 (rare events) の分析が議論されている。また経済・金融でも銀行・保険会社では VaR(バリューアットリスク) と呼ばれる統計的リスク管理が業務上で必要である。この VaR 管理では通常は 1 日単位で金融機関が保有する価値の損失分布の左側 1%、5% の管理を問題とするが、実際にデータ上で 1% の確率事象が実現することは頻繁には起こらない。観察される日々のデータから稀な事象が分析対象であり、

正規分布の利用により問題が生じうる。統計学においては稀な現象の分析は統計的極値論 (statistical extreme value theory) と呼ばれているが、元々はオランダにおける 1950 年代の堤防の決壊により国土のほぼ 1/3 が大災害に見舞われ、堤防の修復の必要性などが契機となり理論が発展している。建設物の耐久性、ダムや河川管理の問題など工学分野を中心に幅広く応用可能性がある。裾確率の評価について既に 6 章で裾確率の評価に関連した 3 つのタイプの極値分布 (extreme value distribution) を議論した。極値分布は期待値の周りに左右対称で釣鐘型、裾が急速に減衰していく正規分布とはかなり異なる確率分布であり、これらの確率分布を利用すると裾確率の評価はかなり異なりうる。この確率分布は位置母数 μ, スケール母数 σ を含めて一般化極値分布 (GEVT, generalized extreme value distribution)

$$F(x) = \exp\left\{-[1+\xi(\frac{x-\mu}{\sigma})]^{-1/\xi}\right\}$$

であるが、スケール母数 ξ について $\xi \to 0$ のときグンベル分布が対応することについて既に言及した。こうした極値分布の利用は極値統計学 (statistical extreme value theory) の出発点である、中心極限定理は確率変数の平均の挙動についての確率法則であったが、極値分布は確率変数の最大値・最小値の挙動についての確率法則であり、極端な事象、稀に起きる事象のリスク分析に有用である [55]。

閾値極値論

3 つのタイプの極値分布を応用することは可能であるが、グループ別の最大値がデータとして得られる必要がある。例えば毎年繰り返される河川の氾濫、海岸に打ち寄せられる波浪など一部の自然現象ではこうしたデータが得られることはある。しかしながら、金融データのような明確な周期性の発生が期待できない場合には、グループ・データの最大値はあまり分析に適しているとは考えにくいのである閾値を超えたデータ、閾値極値の解析が考えられている [56]。

ここで確率変数 X が分布関数 F にしたがうとする。超過分布関数を

$$F_u(x) = \mathbf{P}\left(X \le x+u | X > u\right) \quad (x \ge 0) \tag{11.16}$$

により定める。このとき条件付確率より

$$F_u(x) = \frac{F(x+u)-F(u)}{1-F(u)} \tag{11.17}$$

[55] 統計的極値理論 (extreme value theory, EVT) については例えば Coles, S. (2001), "An Introduction to Statistical Modeling of Extreme Values,"Springer が実用的教科書、渋谷・高橋論文 (国友・山本編「21 世紀の統計科学 Vol-II」収録) などがある。

[56] 例えば、国友・山本編「２１世紀の統計科学 Vol-II」(東京大学出版会), 収録の渋谷・高橋論文を参照されたい。

である。仮に元の確率分布 $F(u)$ が一般化極値分布にしたがうとすれば、u が大きいとき裾確率を

$$1 - F(u) \sim [1 + \xi(\frac{u-\mu}{\sigma})]^{-1/\xi}$$

と表現できよう。同様に

$$1 - F(u+x) \sim [1 + \xi(\frac{u+x-\mu}{\sigma})]^{-1/\xi}$$

である。特に位置母数 $\mu = 0$ として条件付確率を適当な母数として $\sigma(u)$ $(= \sigma + \xi(u - \mu))$ とすると

$$P(X > u + x | X > u) \sim \quad \frac{[1 + \xi(\frac{u+x}{\sigma})]^{-1/\xi}}{[1 + \xi(\frac{u}{\sigma})]^{-1/\xi}} = \left[1 + \xi \frac{x}{\sigma(u)}\right]^{-1/\xi} \tag{11.18}$$

と表現できる。このことをまとめると Balkema-DeHaan の定理として次のような結果として知られている。

定理 11-1： $X_1, \cdots, X_n$ を独立で同一の確率変数列で分布を F とする。このとき基準化数列 $a_n > 0, b_n \in \mathbf{R}$ とある極値分布 $H_{\xi,\mu,\sigma}$, 右裾 x_F が存在すると仮定する。このときある非負関数 $\sigma(u)$ が存在して

$$\lim_{u \to x_F} |F_u(x) - G_{\xi,\sigma(u)}(x)| = 0 \tag{11.19}$$

であり極限分布は $\xi \neq 0$ のとき

$$G_{\xi,\beta}(x) = 1 - \left[1 + \xi \frac{x}{\sigma(u)}\right]^{-\frac{1}{\xi}} \quad (\xi \neq 0) \tag{11.20}$$

である。また $\xi \to 0$ のときには

$$G_{\xi,\beta}(x) = 1 - \exp\left[-\frac{x}{\sigma(u)}\right]$$

となる。

ここで分布関数 G は一般化パレート分布 (generalized Pareto distribution, 略して GPD) と呼ばれるが、確率分布の裾がパレート分布、すなわちべき関数的に減衰することを意味するからである。例えば指数分布に従う確率変数の場合には、$F(x) = 1 - e^{-x}$ $(0 \leq x)$ であるので

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)} = \frac{1 - e^{-(x+u)} - (1 - e^{-u})}{1 - (1 - e^{-u})} = 1 - e^{-x} \tag{11.21}$$

となるので $\xi \to 0$ の場合に対応する。また一様分布 $U(0,1)$ に従う確率変数の場合には、$F(x) = x$ $(0 \le x \le 1)$ より

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)} = \frac{x + u - u}{1 - u} = \frac{x}{1 - u} \tag{11.22}$$

となる。

分位点と母数の推定

一般化パレート (GPD) 分布 $G_{\xi,\sigma}(x)$ については $\mu = 0$ としておくと、$X \sim G_{\xi,\sigma}(x)$ のとき $F_u(x) = G_{\xi,\sigma+\xi u}(x)$ となることは有用である。また平均超過関数 (excess mean function) を $\mathbf{E}[X - u | X > u]$ により定義すると

$$\mathbf{E}[X - u | X > u] = \frac{\sigma + \xi u}{1 - \xi} \quad (0 \le \xi < 1) \tag{11.23}$$

より閾値 u の線形関数となるという性質がある。ここで仮に $F_u(x)$ に対する極限分布 $G_{\xi,\sigma(u)}(x)$ の $\sigma(u)$ $(= \sigma + \xi u)$ を一定値 β で近似できるとき極値分布を利用して高分位点の推定法として閾値 u を用いて

$$\frac{F(x+u) - F(u)}{1 - F(u)} \sim 1 - \left[1 + \xi \frac{x}{\beta}\right]^{-\frac{1}{\xi}}$$

が利用できる。すなわち

$$1 - F(x+u) \sim [1 - F(u)] \left[1 + \xi \frac{x}{\beta}\right]^{-\frac{1}{\xi}}$$

より、データ n 個の中で閾値 u を超える超過個数データが N_u 個であったとすると、裾確率を $1 - F(u)$ を N_u/n で推定すれば

$$F(x+u) \sim 1 - \frac{N_n}{n} \left[1 + \xi \frac{x}{\beta}\right]^{-\frac{1}{\xi}}$$

である。したがって $1 - p$ が小さいときに $F(x+u) = p$ を利用して

$$\hat{x}_p = -u + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{n}{N_u}(1-p) \right)^{-\hat{\xi}} - 1 \right] \tag{11.24}$$

より推定できる。

極値問題では裾指数 ξ の推定がもっとも重要な問題であるが、例えば一般化パレート分布を仮定して（極値分布が正しいと仮定して）最尤推定を行うことなどが考え

られるが、実際の数値計算上では$\xi = 0$の周辺に注意する必要がある。セミパラメトリック推定法としては例えば順序統計量$X_{(n)} \geq X_{(n-1)} \geq \cdots \geq X_{(1)}$を利用した

$$H_{n,k} = \frac{1}{k}\sum_{j=0}^{k-1} \log X_{(n-j)} - \log X_{(n-k)} \tag{11.25}$$

となるHill推定量が知られている。またPickand推定量は

$$P_{n,k} = \frac{1}{\log 2} \log \left[\frac{X_{(n-k)} - X_{(n-2k)}}{X_{(n-2k)} - X_{(n-4k)}} \right] \tag{11.26}$$

で与えられる。

Hill推定量については$k(n) \to \infty$かつ$k(n)/n \to 0$などの一定の正則条件の下で漸近的に一致性、さらに$N(0, \alpha^2)$への漸近正規性が成立する。ただし、Hill推定量のバイアスは小さくなく、また推定の際に利用する順序統計量の数kの選択が重要な問題であることが知られている。観測されるデータからkを選択する方法は幾つかの提案が共存している。

11.4 多次元分布と従属性

相関係数と従属性

例えば経済・金融の標準的分析として平均・分散アプローチが知られている。この方法は複数の資産価格のデータから計算される収益率を確率変数X_i $(i = 1, \cdots, p)$の実現値とみなし、これらの確率変数の線形和（ポートフォリオと呼ぶ）$Z = \sum_{i=1}^{p} w_i X_i$ $(\sum_{i=1}^{p} w_i = 1)$が変動するリスク分析を行う。個々の価格変動が与えられたとき期待値で表現される期待収益と変動のばらつきをコントロールすることが目的である。価格変動の平均とばらつき（分散）を同時に考慮し、期待値が一定の下で分散を最小化することが基本であり、数学的には期待値$\mathbf{E}(Z) = \mathbf{w}'\boldsymbol{\mu} =$ (一定) との制約条件の下で ($\mathbf{w} = (w_i)$とする)、分散

$$\mathbf{Var}(Z) = \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \tag{11.27}$$

を最小化する問題となる。この最小化問題の解$\{w_i\}$は二次計画問題として具体的に解くことができる。通常は確率変数ベクトル$\mathbf{X}$の期待値ベクトル$(\mathbf{E}(\mathbf{X}) = \boldsymbol{\mu})$、共分散行列

$$\boldsymbol{\Sigma} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] = (\sigma_{ij})$$

の正定値性と有限性が仮定される。

例 11-3： 経済・金融の分野では多次元正規分布が利用されることが少なくないが、正規分布は期待値と共分散行列のみにより特徴づけられる。多次元正規分布は

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c \text{ (一定)} \tag{11.28}$$

における密度関数が一定である。この性質を一般化すると楕円分布族 (**EC**, Elliptically Contoured Distribution) が得られるが、密度関数の存在を仮定すると $\mathbf{X} \sim \mathbf{EC}_p(\boldsymbol{\nu}, \boldsymbol{\Lambda})$ は [57]

$$f(\mathbf{x}) = |\boldsymbol{\Lambda}|^{-\frac{1}{2}} g((\mathbf{x}-\boldsymbol{\nu})'\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\nu})) \tag{11.29}$$

で与えられる。ここで **EC** 分布において変換 $\boldsymbol{\nu} \to \mathbf{0}$, $\mathbf{C}'\boldsymbol{\Lambda}^{-1}\mathbf{C} \to \mathbf{I}_p$ を施すと（変換 $\mathbf{Y} = \boldsymbol{\Lambda}^{-1/2}(\mathbf{X}-\boldsymbol{\nu})$）、密度関数は $g(\mathbf{y}'\mathbf{y})$ となるので SC (球面対称分布族、Spherically Contoured Distribution) とも呼ばれている。例としては正規分布の他に
(i) 多次元 t 分布 $g(\mathbf{y}) = c_{p,m}\left[1+\mathbf{y}'\mathbf{y}/m\right]^{-(m+p)/2}$ ($c_{p.m} = \Gamma(m+p/2)/[\Gamma(m/2)m^{p/2}\pi^{p/2}$),
(ii) 混合正規分布 $g(\mathbf{x}) = (1-\epsilon)n_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \epsilon n_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ $(c > 0)$
などある。

ここで特性関数を $\psi_X(\mathbf{t}) = E[\exp^{i\mathbf{t}'\mathbf{X}}]$ として変換 $\mathbf{Y}$ の密度関数は $\mathbf{y}'\mathbf{y}$ のみに依存し $h(\mathbf{y}'\mathbf{y})$ と表現できる。任意の直交変換 $\mathbf{Y}^* = \mathbf{QY}$ に対して密度関数は不変、したがって特性関数は $\mathbf{t}'\mathbf{t}$ の関数となる。このことより確率変数 $\mathbf{X}$ の特性関数は

$$\psi_X(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\nu}}\boldsymbol{\Psi}(\mathbf{t}'\boldsymbol{\Lambda}\mathbf{t})$$

と表現できる。したがって、ポートフォリオ $Z = \mathbf{w}'\mathbf{X}$ の分布の中心は $\mathbf{w}'\boldsymbol{\nu}$, ばらつき (dispersion) は $\mathbf{w}'\boldsymbol{\Lambda}\mathbf{w}$ により表現される。

例えばここで金融リスク管理問題では X を損失 (loss) を表す確率変数とするとき α%VaR を

$$\mathrm{VaR}_\alpha = \inf\{l \,:\, \mathbf{P}(X \geq l) \leq 1-\alpha\} \tag{11.30}$$

の評価が問題であり、実務界では例えば 95% や 99% などが利用されている。こうした分位点のリスク管理上でどのような多次元分布を用いたらよいかは自明ではないのである。

次に二つの確率変数の従属性を表現する尺度として既に登場した Pearson の相関係数 $\rho(X,Y)$ は

$$\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}} \tag{11.32}$$

[57] Anderson (2003) は関連する文献を含めてかなり詳しく議論しているが、例えば 4 次モーメント $E[(X_i-\mu_i)(X_j-\mu_j)(X_k-\mu_k)(X_l-\mu_l)]$ が比較的単純に表現できる。

により定義されるが、次のような問題が生じうることを述べておこう。$p=2$ の時に経済・金融などの分析では個別に対数変換を行い対数正規分布 $\log(X) \sim N(0,1)$, $\log(Y) \sim N(0,\sigma^2)$ を用ることがある。ここで $\sigma = 2, \rho(X,Y) = 0.7$ の場合を考えることは妥当であろうか？例えば $(X,Y) = (e^Z, e^{\sigma Z})$, $Z \sim N(0,1)$ ととすると相関係数のとる値の可能な範囲は $-0.09 \sim 0.666$ となるので、このような (X,Y) の 2 次元確率分布は存在しない。この例での重要な論点は Pearson の相関係数は変数変換するとその値は大きく変わりうることである。すなわち確率変数の従属性は注意深く検討する必要がある。

コピュラ (接合) 関数

楕円分布族とはかなり異なる多次元分布の扱いが生存時間解析と呼ばれている統計学の分野や生命保険などで議論されるようになり、特にコピュラ関数は近年では活発に研究、応用されている。

p 次元確率分布関数は $\mathbf{x} = (x_i) \in \mathbf{R}^p$ に対して確率

$$F(x_1, \cdots, x_p) = \mathbf{P}\left(X_1 \leq x_1, \cdots, X_p \leq x_p\right) \tag{11.33}$$

により定められる。他方、各成分 X_i の確率分布は周辺分布関数

$$F_i(x_i) = \mathbf{P}\left(X_i \leq x_i\right) \ (i = 1, \cdots, p) \tag{11.34}$$

で与えられる。このとき Sklar の定理として次のことが知られている。

定理 11-2：コピュラ関数を $C : [0,1]^p \to [0,1]$ とする。任意の確率分布関数に対してコピュラ関数

$$F(x_1, \cdots, x_p) = C(F_1(x_1), \cdots, F_p(x_p)) \tag{11.35}$$

が存在する。

ここでは [58] 連続分布で密度関数が存在する場合を考えると次のように理解できる。確率変数 X_i に対して、変換 $U_i = F(X_i)$ を施すとき $U_i \sim U([0,1])$ にしたがうことを利用すると、任意の $u_i \in [0,1]$ に対して

$$\mathbf{P}(U_i \leq u_i) = \mathbf{P}(F(X_i) \leq u_i) = \mathbf{P}(X_i \leq F^{-1}(u_i)) = F\left[F^{-1}(u_i)\right] = u_i$$

である。そこで $F(x_1, \cdots .x_p)$ は

$$\begin{aligned} F\left(F_1^{-1}(u_1), \cdots, F_p^{-1}(u_p)\right) &= \mathbf{P}\left(X_1 \leq F_1^{-1}(u_1), \cdots, X_p \leq F_p^{-1}(x_p)\right) \\ &= \mathbf{P}\left(F_1(x_1) \leq u_1, \cdots, F_p(x_p) \leq u_p\right) \\ &= \mathbf{P}\left(U_1 \leq u_1, \cdots, U_p \leq u_p\right) \end{aligned}$$

[58] 例えば国友・山本 (2008)Vol-III の塚原論文がコピュラ関数について説明している。

と表現できる。そこで右辺の最後の項を関数 $C(u_1, \cdots, u_p)$ とすればよい。
ここでコピュラ関数の例を幾つか挙げておこう。損害保険の分析などに使われることのあるクレイトン (Clayton) 族とは

$$C_\theta(u_1, u_2) = \left[\max\{u_1^{-\theta} + u_2^{-\theta} - 1, 0\}\right]^{-\frac{1}{\theta}} \tag{11.36}$$

で与えられる。ただし $\theta \in [-1, \infty)/\{0\}$ とする。
次にガンベル (Gumbel) 族とは

$$C_\theta(u_1, u_2) = \exp\{-\left[(-\log u_1)^\theta + (-\log u_2)^\theta\right]^{\frac{1}{\theta}} \tag{11.37}$$

で与えられる。ただし $\theta \geq 1$ とする。さらにガウス (Gauss) 族は

$$C_\theta(u_1, u_2) = \Phi_\theta\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right) \tag{11.38}$$

で与えられる。ただし Φ^{-1} は標準正規分布の逆関数、Φ_θ は相関係数が θ の 2 次元正規分布である。

ここでコピュラ関数の基本的性質について考えてみよう。$p = 2$ の場合には確率計算から等号 $C(u_1, u_2) - C(v_1, v_2) = C(u_1, u_2) - C(u_1, v_2) + C(u_1, v_2) - C(v_1, v_2)$ を用いて

$$|C(u_1, u_2) - C(v_1, v_2)| \leq |u_1 - v_1| + |u_2 - v_2|$$

となる。また

$$C(u_1, u_2) = \mathbf{P}\left(U_1 \leq u_1, U_2 \leq u_2\right) \leq \mathbf{P}\left(U_i \leq u_i\right) = u_i \ (i = 1, 2)$$

となる。さらに

$$\begin{aligned} C(u_1, u_2) &= \mathbf{P}\left(U_1 \leq u_1, U_2 \leq u_2\right) = 1 - \mathbf{P}\left(\{U_1 > u_1\} \cup \{U_2 > u_2\}\right) \\ &\geq 1 - \mathbf{P}\left(U_1 > u_1\right) - \mathbf{P}\left(U_2 > u_2\right) = 1 - \sum_{i=1}^{2}(1 - u_i) = u_1 + u_2 - 1 \end{aligned}$$

である。一般には次の命題が成り立つ。

定理 11-3：(i) 任意の $(u_1, \cdots, u_p), (v_1, \cdots, v_p) \in \mathbf{I}^p$ に対してコピュラ関数は

$$|C(u_1, \cdots, u_p) - C(v_1, \cdots, v_p)| \leq \sum_{i=1}^{p} |u_i - v_i| \tag{11.39}$$

を満足する。
(ii) コピュラ関数の上限と下限は

$$W = \max\{\sum_{i=1}^{p} u_i - p + 1, 0\} \leq C(u_1, \cdots, u_p) \leq \min\{u_1, \cdots, u_p\} = M \tag{11.40}$$

で与えられる。

ここで M はコピュラ関数であるが $p \geq 3$ のとき W はコピュラ関数とは限らない。一般に $C = M$ のときに共単調 (comonotone) と呼ばれることがある。この条件が成り立つと例えば $p = 2$ のとき二つの確率変数 X_i $(i = 1, 2)$ はある確率変数 Y と非減少関数 f_i により $X_i = f_i(Y)$ と表現できる。$p = 2$ のとき $C = W$ となるとき逆単調 (countermonotone) と呼ばれることがある。このようにコピュラ関数を導入することにより確率変数の従属性について分析が可能となる。

定理 11-4：分布関数 F を連続型とする。確率変数 $X_1, \cdots, X_p$ が互いに独立であることと

$$C(u_1, \cdots, u_p) = \prod_{i=1}^{p} u_i \tag{11.41}$$

は同等である。

このことは例えば $d = 2$ のときに

$$F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right) = u_1 u_2$$

は変数変換より

$$F(x_1, x_2) = F_1(x_1) F_2(x_2)$$

と同等となる。

ここで共分散をもう一度考察してみよう。分散の存在を仮定すると共分散を同時分布・周辺分布より表現できるのでその性質を考察してみよう。ここで任意の $a > b$ に対して指標関数 $I(\omega)$ により

$$a - b = \int_{-\infty}^{\infty} [I(b \leq x) - I(a \leq x)] dx$$

となることを利用すると、独立な二組の確率変数 $(X_1, Y_1), (X_2, Y_2)$ より

$$\begin{aligned} 2\mathbf{Cov}(X, Y) &= \mathbf{E}[(X_1 - X_2)(Y_1 - Y_2)] \\ &= \mathbf{E}[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ((I(X_1 \leq u) - I(X_2 \leq u))(I(Y_1 \leq v) - I(Y_2 \leq v))) du dw] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [P(X_1 \leq u, Y_1 \leq v) - P(X_2 \leq u) P(Y_2 \leq v)] du dw \end{aligned}$$

であるから

$$\mathbf{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F(v, w) - F_1(v) F_2(w)] \, dv dw \tag{11.42}$$

となる。この表現より共分散や Pearson の相関係数 $\rho(X,Y)$ は一般に変数変換に不変でないことが分かる。従属性についてはノンパラメトリック統計学において他の指標も議論されている。代表的な例としては n 組のデータ (x_i, y_i) $(i = 1, \cdots, n)$ より各ペアー $(x_i - x_j, y_i - y_j)$ $(i \neq j)$ において符合が一致すればスコア+1, 符合が異なればスコア-1 として

$$\tau = \frac{[(\text{正のスコア数}) - (\text{負のスコア数})]}{(\text{ペアの総数})}$$

が Kendall の τ と呼ばれている。ここで (X_i, Y_i) $(i = 1, 2)$ を二次元確率変数のコピーとすると、母集団における Kendall の τ は

$$\rho_\tau = \mathbf{P}\left((X_1 - X_2)(Y_1 - Y_2) > 0\right) - \mathbf{P}\left((X_1 - X_2)(Y_1 - Y_2) < 0\right) \tag{11.43}$$

と定義しよう。ここで例えばコピュラ関数の定義 $F(x_2, y_2) = C(F_1(x_2), F_2(Y_2))$ および

$$\begin{aligned} P(X_1 < X_2, Y_1 < Y_2) &= \mathbf{E}[P(X_1 < X_2, Y_1 < Y_2 | X_2, Y_2)] \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P(X_1 < x_2, Y_1 < x_2) dF(x_2, y_2) \end{aligned}$$

などの関係を利用すると、この相関係数は単調変換 (monotone transformation) について不変 (invariant) な次のような表現を持つことが分かる。

定理 11-5：Kendall's τ は

$$\rho_\tau = 4 \int \int_{I^2} C(u_1, u_2) dC(u_1, u_2) - 1 \tag{11.44}$$

と表現できる。ただし $I^2 = [0,1] \times [0,1]$ である。

ここで重要なことはこれらの相関係数はコピュラ関数によってのみ表現でき、コピュラ関数が 1 つの母数で決まる場合にはコピュラ母数と相関が対応することである。同様に Spearman の順位相関なども分析することができる。

例 11-5: アルキメデス型コピュラ ： コピュラ関数の中でも特にある単調減少凸関数 ϕ を用いて

$$C_\phi(u_1, u_2) = \phi^{-1}\left(\phi(u_1) + \phi(u_2)\right) \tag{11.45}$$

と表現できる $(\phi(0) = \infty, \phi(1) = 0$ とすると便利$)$ を満たすときアルキメデス型コピュラ関数と呼ばれる。アルキメデス型コピュラは多くの例を含んでいる。

統計的推測

例として挙げたコピュラ関数はいずれも一つの母数の関数として表現されていた。こうしたコピュラ関数はパラメトリック・コピュラ関数と呼ばれる。こうした場合ですら尤度関数を陽表的に解ける場合はほとんど無いので漸近理論を利用することが考えられる。n 個の互いに独立な標本から得られる尤度関数を

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i|\boldsymbol{\theta}) \tag{11.46}$$

とする。(ここで $f(\mathbf{x}_i|\boldsymbol{\theta}) = c(\mathbf{x}_i|\boldsymbol{\theta}) \prod_{j=1}^{p} f(x_j)$, $c(\mathbf{x}_i|\boldsymbol{\theta})$ はコピュラ密度関数である) 対数尤度関数を最大とする最尤推定量 $\hat{\boldsymbol{\theta}}_n$ $(r \times 1)$ が得られるが、一定の正則条件の下で Fisher 情報行列 $\mathbf{I}_r(\boldsymbol{\theta})$ を用いて

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{L} N_r\left(\mathbf{0}, \mathbf{I}_r^{-1}(\boldsymbol{\theta})\right) \tag{11.47}$$

が利用できる。このとき赤池の情報量基準は

$$AIC = -2\log L_n(\hat{\boldsymbol{\theta}}_n) + 2r \tag{11.48}$$

の最小化として与えられる。

ノンパラメトリック (セミ・パラメトリック) 法を用いることも考えられる。この場合には同時経験分布関数

$$F_n(x_1, \cdots, x_p) = \frac{1}{n}\sum_{i=1}^{n} 1\left(X_{i1} \leq x_1, \cdots, X_{ip} \leq x_p\right) \ (i = 1, \cdots, n) \tag{11.49}$$

及び周辺経験分布関数

$$F_{nj}(x_j) = \frac{1}{n}\sum_{i=1}^{n} 1\left(X_{ij} \leq x_j\right) \ (j = 1, \cdots, p) \tag{11.50}$$

を用いる。経験コピュラ (empirical coplula) は

$$C_n(u_1, \cdots, u_p) = F_n(F_{n1}^{-1}(u_1), \cdots, F_{np}^{-1}(u_p)) \tag{11.51}$$

により定義できる。ここで標本数 n が十分に大きいとき基準化された経験コピュラ関数 $\sqrt{n}\left[C_n(u_1, \cdots, u_p) - C(u_1, \cdots, u_p)\right]$ の挙動を調べればよいが、多次元経験分布の極限としての確率過程の分析が必要となるので複雑になる [59]。

裾従属性と渋谷の定理

例えば VaR で問題となるのは裾確率でありが、複数の確率分布の裾において従属性

[59] 例えば国友・山本 (2012, Vol-3) 収録の塚原論文を参照。

の扱いは重要な問題である。ここで Sibuya(1960) [60] が定式化した (上側) 裾従属性 (tail dependence) は

$$\lambda_U = \lim_{u \to 1} \mathbf{P}\left(X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)\right) \tag{11.55}$$

により定められる。同様に (下側) 裾従属性は

$$\lambda_L = \lim_{u \to 0} \mathbf{P}\left(X_2 < F_2^{-1}(u) | X_1 < F_1^{-1}(u)\right) \tag{11.56}$$

で定められる。もし分布が連続型であれば

$$\lambda_U = \lim_{u \to 1} \frac{C^*(u,u)}{1-u} \tag{11.57}$$

となるが、ここで $C^* = 1 - u_1 - u_2 + C(u_1, u_2)$ である。(同様に $\lambda_L = \lim_{u \to 0} \frac{C(u,u)}{u}$ である。)

ここで重要な例として 2 次元正規分布を考えてみよう。確率変数 $X = (X_1, X_2)'$ がしたがう (標準) 2 次元正規分布を $\mu_1 = \mu_2 = 0$, σ_{11} $(= \sigma_1^2) = \sigma_{22}$ $(= \sigma_2^2) = 1, \sigma_{12} = \rho$ としよう。標準正規分布関数 $\Phi(z)$ とすると、小さな s に対して裾確率 $P(1-s, 1-s) = P(X_1 > \Phi^{-1}(1-s), X_2 > \Phi^{-1}(1-s)) \leq P(X_1 + X_2 > 2\Phi^{-1}(1-s))$ となることに注目する。$X_1 + X_2$ は正規分布 $N(0, 2(1+\rho))$ にしたがうので $P(1-s, 1-s) \leq 1 - \Phi(\alpha\Phi^{-1}(1-s))$ $(\alpha = \sqrt{2/(1+\rho)})$ である。正規分布の裾確率は $s = 1 - \Phi(z) \sim [1/(\sqrt{2\pi})](1/z)e^{-z^2/2}$ と近似できるので ((2.13) を参照)

$$\frac{P(1-s, 1-s)}{s} \to 0 \ \ (s \to 0)$$

と評価できる。この結果を次のようにまとめておく。

定理 11-6： 二次元正規分布にしたがう確率変数

$$\mathbf{X} \sim N_2\left[\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right] \tag{11.58}$$

とする。Pearson の相関係数が $\rho < 1$ ならば $\lambda_U = 0$ となる。

この結果より極端な事象の同時依存性を正規分布を利用して分析する場合には問題が生じうると解釈できる。正規分布を利用すると裾確率は漸近的に独立性が自動的に成立してしまうので裾同時確率についての正確な評価に利用することは妥当でない場合が少なくないのである。渋谷の定理 (Shibuya(1960)) などを一つの契

[60] Sibuya, M. (1960),""Bivariate Extreme Statistics I," *Annals of Institute of Statistical Mathematics*, 195-210.

機に多変量極値の研究と応用が盛んになっている[61]。ここで基準化のために各確率変数 $M_{in} = \max_{1\leq j\leq n} Y_{ij,n}/n$ $(i = 1, 2)$ の周辺分布の極限として標準 Frechet 分布 $F(x) = \exp(-1/x)$ $(x > 0)$ を選んでおく。このとき二次元確率変数 $(Y_{1j,n}, Y_{2j,n})$ に対し $\mathbf{M}_n^* = (M_{1n}, M_{2n}) = (\max_{1\leq j\leq n} Y_{1j,n}/n, \max_{1\leq j\leq n} Y_{2j,n}/n)$ の極限分布は $n \to \infty$ のとき

$$G(x_1, x_2) = \exp\left(-V(x_1, x_2)\right) \tag{11.59}$$

および

$$V(x_1, x_2) = 2\int_0^1 \max\{\frac{w}{x_1}, \frac{1-w}{x_2}\}dH(w) \tag{11.60}$$

と表現できる。ただし H は $[0, 1]$ 上の分布関数で平均値 $\int_0^1 wdH(w) = 1/2$ を意味している。二次元極値分布の一例として $H(w) = 1/2$ $(w = 0, 1)$; $H(w) = 0$(それ以外) とすると $G(x_1, x_2) = \exp[-(x_1^{-1} + x_2^{-1})]$ となる。この形より実母数 α $(0 \leq \alpha \leq 1)$ を利用して一般化すると

$$G(x_1, x_2) = \exp\{-(x_1^{-1/\alpha} + x_2^{-1/\alpha})^\alpha\} \tag{11.61}$$

というロジスティック分布族が得られる。

[61]例えば L. de Haan and A. Ferreira (2006),*Extreme Value Theory*, Springer などが詳しい。

参考文献

本文で引用した個々の文献に加えてより一般的に統計学・数理統計学の理論と応用を勉強する為に幾つかの書籍を重要な参考文献として挙げておく。第 1 部の確率論の基礎事項については [1],[7],[15] を挙げておくが、いずれも数理科学に関心のある本格的な書籍である。測度論については 1 章・3 章で述べたように応用に関心のある立場からは [10] を勧めておく。第 2 部の数理統計学の基礎理論については著者は学部生時代に教科書として [8],[11] などで勉強したが、その後の日本語での良書としては [2],[8],[9],[12] などがあるが [9],[12] などがあり特に数理系の学生・院生向けの教科書である。数理統計学の標準的理論については [9] がかなり包括的に議論している。本書では数理統計学の諸分野についてはほんの少し議論しただけであるが、筆者の知っている分野についての書籍として、統計的時系列論について [13], 統計的多変量解析について [14], 統計的生存解析について [17], 統計的極値論について [16], 計量経済学について [4], [6] などを挙げておく。その他の応用の問題についての糸口を見つけるには [5] が収録している論文から始めるのが良いと思われる。

References

[1] 伊藤清 (1991),「確率論」, 岩波書店。

[2] 稲垣宣生 (2003),「数理統計学」, 改訂版, 裳華房。

[3] 国友直人・高橋明彦 (2003),「数理ファイナンスの基礎」, 東洋経済新報社。

[4] 国友直人 (2011),「構造方程式モデルと計量経済学」, 朝倉書店。

[5] 国友直人・山本拓 監修 (2012),「21 世紀の統計科学」Vol-I:社会・経済の統計科学, Vol-II:自然・生物・健康の統計科学, Vol-III:数理・計算の統計科学, 東京大学出版会, ＨＰ版（ＨＰ参照)。

[6] 国友直人 (2014)「計測誤差と統計学」日本統計学会誌 43-2,（HP 参照)。

[7] 舟木直久 (2004),「確率論」, 朝倉書店。

[8] 竹内啓 (1963), 「数理統計学」, 東洋経済新報社。

[9] 竹村彰通 (1991),「現代数理統計学」, 創文社。

[10] 志賀浩二 (1990),「ルベーグ積分 30 講」, 朝倉書店。

[11] 鈴木雪夫 (1975),「経済分析と確率・統計」, 東洋経済新報社。

[12] 吉田朋広 (2006)「数理統計学」, 共立出版。

[13] Anderson, T.W. (1971),"The Statistical Analysis of Time Series," Wiley.

[14] Anderson, T.W. (2003),"An Introduction to Multivariate Statistical Analysis," 3rd Edition, Wiley.

[15] Billingsley, P. (1994), "Probability and Measures," 3rd Edition, John-Wiley.

[16] Coles, S. (2001),"An Introduction to Statistical Modeling of Extreme Values," Springer.

[17] Miller, R. (1981),"Survival Analysis," Wiley.

[18] Van der Vaart (1998),"Asymptotic Statistics," Cambridge University Press.

練習問題

問題 1

1-1. (i) 集合 $A \setminus B = A \cap B^c$ と定義する。$A \setminus B = A \setminus (A \cap B)$ を示せ。
(ii) 集合 $A_i = \{x | x \in [1/i, 1]\}$ とするとき $\bigcup_{i=1}^{\infty} A_i$ および $\bigcap_{i=1}^{\infty} A_i$ を求めよ。

(iii) 集合 $A_i = \{x | x \in [a/i, b]\}$ $(0 < a \leq b)$, $B_i = [-a, b] \setminus A_i$ とするとき $\bigcup_{i=1}^{\infty} A_i$, $\bigcap_{i=1}^{\infty} A_i$, $\bigcup_{i=1}^{\infty} B_i$, $\bigcap_{i=1}^{\infty} B_i$ を求めよ。
$A \setminus B = (A \bigcup B) \bigcap B^c$ は集合差を表す。
1-2. $\mathcal{F}$ を $\mathbf{\Omega}$ 上の $\sigma-$ 加法族とする。加算個の事象 $A_1, A_, \cdots \in \mathcal{F}$ に対して「A_i が単調増加 $(A_1 \subset A_2 \subset \cdots)$ なら $P(\bigcup_{i=1}^{\infty} A_i) = \lim_{n \to \infty} P(A_n)$ となることを示せ。
1-3. 定義 1-3 の条件 3 を示せ。

問題 2

2-1. (i) 標準正規分布の分布関数 $\Phi(x)$, 密度関数 $\phi(x)$ とするとき

$$[x + x^{-1}]^{-1} \leq [1 - \Phi(x)]/[\phi(x)] \leq x^{-1}$$

を示せ。
(ii) X がパレート分布にしたがうとき期待値、分散などの性質を考察せよ。
2-2. X が正規分布、Y がポワソン分布にしたがうとき期待値 $\mathbf{E}(X)$、分散 $V(X)$、歪度 (skewness)、尖度 (kurtosis) などの性質を考察せよ。ただし歪度、尖度はそれぞれ $\kappa_3 = \mathbf{E}[(X - \mathbf{E}(X))^3]/(\sqrt{V(X)})^3$, $\kappa_4 = \mathbf{E}[(X - \mathbf{E}(X))^4]/(\sqrt{V(X)})^4$ により定義する。
2-3. 積率 $\mathbf{E}[|X|^r]$ $(r = 2, 3, 4)$ が有限であるための裾確率 $P(|X(\omega)| > x)$ の条件を考察せよ。
2-4. 微積分のロピタルの定理を用いて (2.13) を示せ。
2-5. 独立な確率変数列 X_i $(i = 1, \cdots, n)$ が $N(\mu, \sigma^2)$ にしたがうとき $Y = \sum_{i=1}^{n} X_i^2$ の特性関数と確率分布はどのように特長づけられるか？
2-6. $X \sim N(\mu, \sigma^2)$ のとき $Y = e^X$(対数正規分布) の期待値, 分散を求めよ。
2-7. 関数 $F_1(x) = \exp[-\exp(-x)]$ $(-\infty < x < \infty)$ および $F_2(x) = \exp[-x^{-1}]$ $(x > 0)$ はそれぞれ確率分布関数とみなせるか。

2-8. X が指数分布 $\mathrm{E}(\alpha)(\alpha > 0$ は母数) にしたがうとき、任意の $s > t > 0$ に対して $P(X > s|X > t) = P(X > s - t)$ となることを示せ。
2-9. $P(X > 0) = 1$ のとき

$$\mathbf{E}[\frac{1}{X}] \geq \frac{1}{\mathbf{E}[X]}$$

となることを示せ。不等号が成立する例を考えよ。
2-10. 確率変数 X がガンマ分布 (α, β) にしたがうとき $Y = X^{-1}$ の確率分布を求めよ。
2-11. 例 2-13 の説明を確かめよ。
2-12. X_i $(i = 1, 2) \sim N(0, 1)$ のとき変数変換を利用し $Y = X_1^2 + X_2^2$ の分布を求めよ。
2-13. $X|Y \sim B(Y, p)$, Y は $P_o(\lambda)$ にしたがうとするとき X の分布を求めよ（生態学の例では Y は eggs の数, $X|Y$ は生存数が典型例)。
2-14. $X \sim N(\mu, \sigma^2)$ のとき $Y = e^X$(対数正規分布) の期待値, 分散, 歪度, 尖度を求めよ。

問題 3

3-1. 定義 3.1 の条件付確率が確率測度であることを示せ.
3-2. (3.13) が $\sigma-$ 加法族であることを示せ。
3-3. 定理 3.1(ベイズの公式) を示せ。
3-4. 例 3.2 において $p = 2, p_1 = p_2 = 1$ のとき $\mathbf{\Sigma}$ の逆行列要素を評価して同時密度関数と条件付密度関数の表現を確認せよ。
3-5. 確率変数 X, Y について $Y \neq X$ のとき $\mathbf{E}[Y^2|X] = X^2, \mathbf{E}[Y|X] = X$ となることがあるか。
3-6. 正規分布にしたがう確率変数 X, Y について $\mathbf{E}[X|Y] = \mathbf{E}[Y|X]$ となることがあるか
3-7 三つの確率変数 X_1, X_2, X_3 について任意の二つのペアが独立なら三つの確率変数が独立と言えるか。(例 3.3 の例を参考にしなさい。
3-8. (3.35) で定められる確率変数列 $\{X_i\}$ の期待値、分散、(X_i, X_j) $(i \neq j)$ の共分散、相関を求めよ。(ただし初期条件 $X_0 = 0$ とせよ。)

問題 4

4-1. 期待値が一定である互いに独立な確率変数列 X_i $(i = 0, 1, 2, \cdots, n)$
(ただし $\mathbf{E}[X_i] = \mu, \mathbf{V}[X_i] = \sigma_i^2 < \infty$ とする) より作られるマルチンゲール (martingale) の例を挙げよ。

4-2. 互いに独立な確率変数列 $X_i \sim N(0,\sigma^2)$ $(i=1,2,\cdots,n)$ とする。(σ^2 は母数とする。) ある関数 σ_{1n},σ_{2n} をとり $Y_n=\sum_{i=1}^n X_i^2-\sigma_{1n}$, $W_n=\exp[\sum_{i=1}^n X_i-\sigma_{2n}]$ がマルチンゲールとなることがあるか？
4-3. マルチンゲール (martingale) について X_n が可積分かつ条件付期待値が $\mathbf{E}[X_{n+1}|X_n,X_{n-1},\cdots,X_1]=(1/n)(X_1+\cdots+X_n)$ のとき $Y_n=(1/n)(X_1+\cdots+X_n)$ はマルチンゲールとなるか。
4-4. マルチンゲール (martingale) $\{X_n\}$ が $X_n\to c$(一定値, 一定の確率変数) となる例、マルチンゲール $\{X_n\}$ が $X_n\to-\infty$ (あるいは $X_n\to\infty$) となる例を考えよ。
4-5. 例 4-3 における主張は確率変数列が X_i $(i=1,\cdots,n)$ が定常な (3.35) にしたがうときにも成立するか？(ただし定常条件は $|a|<1$ とする。)

問題 5

5-1. 自由度 n の t 分布は $n\to\infty$ のときに $N(0,1)$ に収束することを示せ。
5-2. 自由度 n の χ^2- 分布の自由度 $n\to+\infty$ のとき正規分布で近似できることを示せ。
5-3. 確率変数列 X_n が確率収束するが平均二乗収束するとは限らないことを示せ。
5-4. X_n がマルチンゲールのとき $n\to\infty$ のとき一定値に収束する例、発散する例を与えよ。
5-5. 積率母関数 $M_n(\theta)$, 特性関数 $\phi_n(t)$ とする。確率変数列 X_n がある確率変数 X に収束するとき M_n と ϕ_n はある関数 M と ϕ に収束すると云えるか。

問題 6

6-1. 有限母集団と無限母集団の場合の標本平均の分散公式の差（有限母集団修正）の導出を完成せよ。
6-2. (6.18) に表れる c_m を導出に表れる a_m より導け。
6-3. (6.23) の分母と分子が独立となることを考察せよ。
6-4. (6.25) の分解に表れる $\mathbf{P}_n,\mathbf{P}_{n1},\mathbf{P}_{n2}$ が射影行列であることを確認せよ。
6-5. $X_1,\cdots,X_n$ が独立に $N(\mu,\sigma^2)$ にしたがうとき確率変数 $Y=\sum_{i=1}^n X_i^2$ とする。$\mu=0$ のとき積率母関数・特性関数を求め Y がガンマ分布にしたがうことを示せ。さらに $\mu\neq0$ のときの Y のしたがう確率分布を考察せよ。
6-6. 互いに独立で同一のコーシー分布にしたがう場合の最大値の分布を求めよ。(ただし関数 $1/(1+x^2)$ の原始関数が $tan^{-1}x$ $(arctanx)$ となることを利用してよい。)
6-7. X_k $(k=1,\cdots,n)$ がパレート分布にしたがうとき、X_k $(k=1,\cdots,n)$ の最大値の確率分布を求めよ。$n\to\infty$ のとき最大値の分布の挙動を考察せよ。

問題 7

7-1. 母集団として正規分布を仮定するとき (例 7.1)、$\mathbf{E}[s_n^2] = \sigma^2$ となることを示せ。３次・４次積率、歪度・尖度も求めよ。
7-2. 例 7.4 と例 7.5 における十分統計量を導け。
7-3. 例 7.11 における不偏推定量 s_n^2 の不偏性と最尤推定量のバイアス、平均二乗誤差についての主張を示せ。
7-4. 独立な確率変数列 X_i $(i = 1, \cdots, n)$ が正規分布 $N(\mu, \sigma^2)$ にしたがうとき、μ が既知 (例えば 0) と未知の場合の σ^2 の UMVU を求めよ。
7-5. 互いに独立な確率変数列 X_i $(i = 1, \cdots, n)$ が母数 λ の指数分布にしたがうとする。母数 $\lambda(\lambda > 0)$ の推定量として $\hat{\lambda} = \frac{1}{m}\sum_{i=1}^{m} X_i (m$ は整数$, 0 < m \leq n)$ とするとき、この推定量の統計的意味での妥当性について議論せよ。
7-6. 互いに独立な確率変数列 X_i $(i = 1, \cdots, n)$ が母数 λ のポワソンにしたがうとする。(i) 母数 λ の UMVU を求めよ。(ii) 母数 $\lambda(\lambda > 0)$ の推定量として $\hat{\lambda} = \frac{1}{m}\sum_{i=1}^{m} X_i (m$ は整数$, 0 < m \leq n)$ とするとき、この推定量の妥当性と改善可能性について議論せよ。
7-7. 例 7.11 において $(n-1)s_n^2/\sigma^2$ が $\chi^2(n-1)$(ガンマ分布 $G(n/2, 2)$) にしたがうことを利用すると σ の不偏推定量が $s\sqrt{n-1}\Gamma((n-1)/2)/[\sqrt{2}\Gamma(n/2)]$ となる。
7-8. (i) 平均二乗積分誤差 (7.63) を h について最小化した最適な h^* を求めよ。
(ii) 漸近分散項 $\mathbf{Var}(\hat{f}_n)$ の第二項が n が大きいとき相対的に無視できることを示せ。

問題 8

8-1. $X_1, \cdots, X_n$ が独立に $N(0, \sigma^2)$ にしたがうとき帰無仮説 $H_0 : \sigma^2 \leq 1$, 対立仮説 $H_1 : \sigma^2 > 1$ に対する検定方式を与えよ。
8-2. 例 8-3 における尤度比検定統計量を導き、n が大きいとき帰無仮説および対立仮説における漸近分布を求めよ。
8-3. X が幾何分布 $p(x) = p(1-p)^x$ $(x = 0, 1, \cdots)$ にしたがうとき母数 p の推定方法と帰無仮説 $H_0 : p = p_0, H_1 : p > p_0$ の検定方法を考察せよ。母数 p についての十分統計量、Fisher 情報量を求めよ。
8-4. 互いに独立な確率変数列 X_i $(i = 1, \cdots, n)$ が母数 λ の指数分布にしたがうとする。帰無仮説 $\mathcal{H}_0 : \lambda = \lambda_0$ (既知の値) を対立仮説 $\mathcal{H}_1 : \lambda \neq \lambda_0$ に対して検定する問題を考察する。統計的に妥当と思われる検定方法を挙げその理由を述べよ。
8-5. 互いに独立な確率変数列 X_i $(i = 1, \cdots, n)$ が母数 λ のポワソンにしたがうとする。帰無仮説 $\mathcal{H}_0 : \lambda = \lambda_0$ (既知の値) を対立仮説 $\mathcal{H}_1 : \lambda \neq \lambda_0$ に対して検定することを考える。妥当と思われる検定方法を挙げその理由を述べよ。

8-6. Welch 検定において漸近的議論を用いて $N(0,1)$ による検定を利用することができるか否か考察せよ。

問題 9

9-1. 互いに独立な確率変数列 X_i $(i = 1, \cdots, n)$ が母数 λ のポワソンにしたがうとする。
帰無仮説 $\mathcal{H}_0 : \lambda = \lambda_0$ (既知の値) を対立仮説 $\mathcal{H}_1 : \lambda \neq \lambda_0$ に対して検定することを考える。妥当と思われる検定方法を挙げその理由を述べよ。
9-2. 標本が正規分布 $X \sim N(\mu, \sigma^2)$ から互いに独立に得られるとする ($\sigma^2 = \sigma_0^2$ 既知)。例 9-3 のように μ についての事前分布として正規分布を仮定できるとき事後分布を導出し μ についての推定量の性質を考察せよ。
9-3. 定理 8-1 の L 集合 (リスク集合) の凸性を示せ。さらに例 9-5 で述べられているベイズ解の説明を確認せよ。
9-4. 定理 9-2 の証明で用いた式変形を確認せよ。
9-5. 次に引用する文を読み数理統計学の観点からコメントせよ。
「いったい、なぜ人間は想起しやすさを過度に重視するのだろうか。... たった一度、されど一度。統計学が語るところの大数の法則に対して、小数の法則と呼んでもよい。大数の法則とは何度も実験を繰り返せば、観察される頻度はその確率に近づくことを表した法則である。統計学が何度も繰り返すことができる事柄を扱うのに対して、我々は繰り返しのきかない人生を生きている。繰り返しのきかない人生だからこそ、有限の経験に頼ってしまうのであろう。」
9-6. n 個の標本がポワソン分布 $X \sim P_o(\lambda)$ から独立に得られるとき、母数 λ についての事前分布がガンマ分布 $\Gamma(\alpha_0)$ と分かっているとする。(i) このとき標本が得られた時の母数 λ に関する事後分布を求め、合理的と考えられる規準に基づき λ の推定値を導け。得られた推定値を直観的に説明せよ。
(ii) 次に保守的な上司の損失関数として、推定値が真の値より大きいときの損失が推定値が真の値より小さいときの損失より大きいとき、すなわち

$$l(\lambda, \hat{\lambda}_n) \quad = 3(\hat{\lambda}_n - \lambda)_+ + (\hat{\lambda}_n - \lambda)_-$$

とする。(ただし関数 $|x|_+ = x$ $(x \geq 0); 0$ $(x < 0)$ および関数 $|x|_- = -x$ $(x < 0); 0$ $(x \geq 0)$ とした。) このとき母数 λ の推定値をどの様に上司に報告したら良いか？報告する推定値の意味を直観的に説明せよ。

問題 10

10-1. (10.1) 式の最小化により係数 b_0, b_1 が一意に定まらないことがあるか？
10-2. 単回帰 $k = 1$ のとき推定量を $b_0 = \sum_{j=1}^{n} a_j y_j, b_1 = \sum_{j=1}^{n} c_j y_j$ (a_j, c_j は実数列) としてガウス・マルコフの定理 (最小二乗推定量は最小分散不偏推定量) を示せ。
10-3. (10.12) を導け。更に $\mathbf{c}_j\mathbf{y}$ $(j = 1, \cdots, p)$ の期待値、分散・共分散を導き不偏性の条件を導き、定理 10.1 を示せ。
10-4. $k = 1$ のとき仮定 (A4) の下で統計量

$$T = \frac{b_1 - \beta_1}{\sqrt{a^{11}\hat{\sigma}^2}}$$

は自由度 $n - 2$ の t 分布にしたがうことを示せ。
10-5. 解 (10.15) および (10.17) を導け。
10-6. 10 章 2 節 (10.17) の最小化問題の解を求め、$\mathbf{b}$ がスカラーの場合に直交回帰の解に一致することを示せ。
10-7. (10.27) および (10.28) の最小化問題の解を求めよ。

問題 11

11-1. (11.22) を示せ。
11-2. (11.42) を示せ。
11-3. 本文の説明を利用して (11.43) を示せ。
11-4. (11.44) を示せ。
11-5. 二次自己回帰過程 (AR(2),(11.1) において $p = 2$) が定常的となる係数の条件を求めよ。ここで定常性は $\mathbf{E}(Y_t) = \mathbf{E}(Y_{t-1}) = \mathbf{E}(Y_{t-2}) = \cdots$, $\mathbf{E}(Y_t^2) = \mathbf{E}(Y_{t-1}^2) = \mathbf{E}(Y_{t-2}^2) = \cdots$ $\mathbf{E}(Y_tY_{t-1}) = \mathbf{E}(Y_{t-1}Y_{t-2}) = \mathbf{E}(Y_{t-2}Y_{t-3}) = \cdots$ などを意味する。1 次移動平均過程 ((11.2) において $q = 1$) のときは定常性の条件は何か。
11-6. AR(1),AR(2) に対し 1 期先予測・2 期先予測量を構成せよ。
11-7. Pearson 相関係数 ((11.32)) に関連して $(X_1, Y_1) = (e^Z, e^{\sigma Z})$,$(X_2, Y_2) = (e^Z, e^{-\sigma Z})$), $Z \sim N(0, 1)$ のとき (X_1, Y_1) および (X_2, Y_2) の相関係数がそれぞれ $e^{-\sigma}/\sqrt{(e-1)(e^{\sigma^2}-1)}$, $e^{\sigma}/\sqrt{(e-1)(e^{\sigma^2}-1)}$, となることを示し、(11.32) について本文で述べている主張を確かめよ。
11-8. $p = 2$ の場合に定理 11-3 を確かめよ。
11-9. (11.57) に続く二次元正規分布の裾確率評価を確かめよ。
11-10. 国際的な銀行業務に対する公的規制において当局は「銀行に対して損失分布 (その価値が将来に変動する資産を保有する場合には、損失が発生する可能性がある資産価値の) に対して対数値の変化額に正規分布を当てはめて 1% 分位点を管理する

（VaR（Value-at-Risk）基準と呼ばれる）」ことを推奨したことがある。この方法の適用について数理統計学的見地よりコメントし、妥当でないと考えるならば対案を示せ。