目次

第 1 章

縦割りの統計学

推測統計学は，観測データの背後に確率モデルを想定する仮定から始まる．

1.1 記述 vs 推測

1.1.1 二項対立の構造

記述統計学と推測統計学との手法の違いは，データの背後に確率論的な構造を考えるかどうかという点にある．データを所与のものと思い，データに対する理解を深める手法と，データを確率変数の実現値と捉えて予測を目指す手法とである．

これはアルゴリズム vs 推論という二項対立でもある．汎関数の計算・実装と，推定である．現在は前者の黄金期であり，推論法に進化を要求する．アルゴリズムは統計学の外に触発される，神経回路網，サポートベクトルマシン，ブースティングなど．そしてこれを受け止めるために数理がさらに豊かになる．

1.1.2 回帰の例

線型回帰の最小二乗法は古典的なアルゴリズムである．その推定の正確性を評価する推論法には，標準誤差などがある（95% で入る振れ幅）．近年のアルゴリズムには lowess (locally weighted scatterplot smoother) がある．これには ± 2 倍のブートストラップ標準誤差などで推論される．ブートストラップ法は解析的な方法ではない（標準誤差のように式を持たない）が，それゆえにいかなる複雑なアルゴリズムにも適用できることが有用になる，計算量にものを言わせる脳筋手法である．

1.1.3 仮説検定

推定量の評価以外にもう一つ，推論法が要請される分野は，仮説検定である．2 標本 t-統計量 $t=\dfrac{\overline{x}-\overline{y}}{\widehat{\mathrm{sd}}}$ は，ステューデントの t 分布が帰無分布となる．

1.2 頻度派

20C から統計学の営みが開始され，卓上計算機でも実行可能なアルゴリズムによる理論が完成した．これを古典理論，または，頻度派的な推論という．

1.3 臨床 vs 疫学

経済にミクロとマクロがあるように，生物統計学にも臨床統計と疫学統計がある．これらをまとめて生物統計学と呼ぶようだ．

1.4 因果推論

物理学実験手法は確率された数理科学であるが，その他の自然科学・社会科学に回転させるには，因果推論の技法が必要になる．物理現象をモデルするのには Euclid 空間などが用いられたが，社会現象・生物現象をモデル化するには，確率や因果といった概念の形式化も必要となる．Newton の「力」のようなものである．

第 2 章

ノンパラメトリック推定

Lehmann[2] が当時の実用書で読みやすい．基本的に分布の正規近似と極限定理を用いて，どのように検定を構成すれば良いかを記述した，応用数学者の真髄が現れている書籍である．

2.1 歴史

歴史 2.1.1 (漸近論)**.** 生存解析の分野で，60s に Wilcoxon 統計量をはじめとしてノンパラメトリックな手法が発展し，パラメトリックな手法に比べて遜色ないことが明らかになった．40s のジャックナイフ法からその萌芽があったが，80s に統計計算として再興した．以来，漸近理論は数理統計学の中心になっている．主要な手法は分布の近似を標本数に依存した関数で近似する Edgeworth 展開，コーニッシュ・フィッシャー展開や，統計量を変換することで近似を改良する分散安定化変換や正規化変換がある．

歴史 2.1.2 (前園さんの研究)**.** Wilcoxon 統計量の性質を研究していると，U-統計量というクラスにつながり，この Edgeworth 展開を考えている．U-統計量は条件付き期待値＝射影の方法で漸近正規性がすぐに示せる，線型な統計汎関数である．70s には独立確率変数の和の収束レートの理論を U-統計量に一般化する理論が「ベリー・エシーン限界」を見つけ，80s には Edgeworth 展開が考えられて標本数の逆数までの近似が得られた．Lai and Wang (93) は「漸近 U-統計量」に理論を一般化した．このクラスはほとんどの独立同一観測から構成される統計量が満たす．

2.2 Wilcoxon の順位和検定

2.2.1 基本的な設定

N 人から，処置群 n 人を同様に確からしく選ぶ．全被験者 N 人の中で，効果を定量化して，順位をつける．順位の付け方は，処置群の中で $S_1 < \cdots < S_n$ となり (添字 $1,\cdots,n$ を並べ替える)．対照群の中でも $R_1 < \cdots < R_m$ を満たすようにつけたとすると，$\{S_1,\cdots,S_n,R_1,\cdots,R_m\} = [N]$ を満たす．すると，順位の和

$$W_s := S_1 + \cdots + S_n$$

が十分に小さいとき，処置効果がないという仮説 H を棄却することが考え得る．

このとき，W_s は独立ではない確率変数の和であるが，これも漸近正規になる．

2.2.2 2 標本問題化

与えられた母集団から，上述の被験者 N 人を抽出することを考慮に入れる．そしてこの母集団から任意に抽出した 1 人が，処置を受けたときの反応 Y と，対照として用いられたときの反応 X とを比較したい．母集団からの抽出の確らしさを仮定すれば，処置効果がないということは X と Y の分布が等しいということを意味する．ここにさらに，「母集団は十分に大きいので，X と Y の間の相関は無視できる」という仮定も置いたとき，このモデルを**母集団モデル**といい，i.i.d. 列 $X_1,\cdots,X_m \sim F$ と $Y_1,\cdots,Y_n \sim G$ の間で $H : G = F$ を検定する．これを **2 標本問題**という．

定理 2.2.1 . 帰無仮説 $H : F = G$ の下で，$(S_1,\cdots,S_n)$ は $[N]$ から $(s_1,\cdots,s_n)$ の選び方 $\binom{N}{n}$ 通りの上の一様分布に従う．

注 2.2.2. このとき，分布 F には一才の仮定を置いてないので，どのような場合でも使える．この観点から，順位検定は**ノンパラメトリックな検定**であるという．

2.2.3　統計量の分布

Hajek (1961) による理論の特殊化になる．

記法 2.2.3. 母集団の列 $\Pi_1, \Pi_2, \cdots$ は，$\Pi_n := \{v_{n1}, \cdots, v_{nn}\}$ からなるとする．N 番目の母集団から，大きさ $n := n(N)$ の標本を抽出し，値を $A_{N_1}, \cdots, A_{N_n}$ とする．標本平均と母平均とをそれぞれ

$$S_N := \sum_{i=1}^n A_{N_i}, \quad v_N := \frac{\sum_{i=1}^N v_{Ni}}{N}$$

とおく．S_N の各項は**有限標本からの抽出であるが故の重属性**があるので，中心極限定理が直接には適用できない．

定理 2.2.4 (Hajek 1961 の特殊化). 標準化された変数 $S_N^* := \dfrac{S_N - E[S_N]}{\sqrt{\mathrm{Var}[S_N]}}$ が漸近的に $N(0,1)$ に従うための十分条件は，次の 2 条件が与える：

(1) $n, m := N - n \to \infty$.
(2) $\dfrac{\max_{i\in[N]}(v_{Ni} - v_N)^2}{\sum_{j=1}^N (v_{Nj} - v_N)^2} \max\left(\dfrac{m}{n}, \dfrac{n}{m}\right) \to 0.$

系 2.2.5 . 条件 (2) を $N \to \infty$ のとき，$\dfrac{\max_{i\in[N]}(v_{Ni} - v_N)^2}{\sum_{j=1}^N (v_{Nj} - v_N)^2/N}$ が有界である，という条件に弱めても，定理は成り立つ．

例 2.2.6. Π_N は d_N 個の 1 と $N - d_N$ 個の 0 からなるとすると，確率変数 S_N は超幾何分布に従う確率変数である．すなわち，定理は標準化された超幾何確率変数は $N(0,1)$ に漸近することが含意されている．

例 2.2.7. $\Pi_N := [N]$ のとき，S_N は Wilcoxon 統計量になる．

2.3　U-統計量

Hoeffding (48) によるアプローチで，対立仮説の下での順位統計量の漸近分布を考える．

記法 2.3.1. $\varphi : \mathbb{R}^2 \to \mathbb{R}$ を任意の実関数とし，これに対して

$$U := \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \varphi(X_i, Y_j)$$

とおく．なるほど，U 統計量は，母集団モデル／2 標本問題において自然に現れるのか．

要諦 2.3.2. 基本的な Hajek (61) の方針は，$S = \sum_{i=1} a_i(X_i) + \sum_{j=1} b_j(Y_j)$ という，核 a_i, b_j が定める線型な統計汎関数の和と漸近同等であることを示すにあたって，a_i, b_j を「射影（2 乗和を最小にするもの）」により構成することである．

定理 2.3.3 . $m \leqslant n$ かつ $m/n \xrightarrow{m,n\to\infty} \lambda \in \mathbb{R}$ とする．このとき，$T := \sqrt{m}(U - \theta)$ は漸近正規で，$N(0, \sigma^2), \sigma^2 = \sigma_{10}^2 + \lambda\sigma_{01}^2$ に従う．

これを単一標本で考えることとする．

定理 2.3.4 . $X_1, \cdots, X_N$ は独立同分布に従うとし，

$$U := \frac{1}{\binom{N}{2}} \sum_{i<j} \varphi(X_i, X_j)$$

とする．このとき，$T := \sqrt{N}(U - \theta)$ は漸近正規で，$N(0, 4\sigma_1^2)$ に従う．

第 3 章

臨床統計

臨床試験に関する統計解析では，adaptation についての統計的研究が大事になる．また，因果推論がほとんど全ての問題設定である．さらに，Woodcock (2005) が指摘しているように，Bayes 流のアプローチは，一般に利用される手法に比べて，時間・予算・人的資源・サンプルサイズを節約しつつ適切な情報を得ることができるため，医薬品開発の分野で関心が高まっている．Temple (2005) は，Bayes 流のアプローチが用いられていないにも拘らず FDA の審査官が Bayes 流の思考プロセスを採用していることを指摘している．人間共同体の全体を実験としているのである！

clinical test と社会実験と政策評価と心理学実験とは全て相似形であるはず．そもそも全ての「社会科学」的な行為が，因果推論を基本言語として統合されるのかもしれない．初めから物理学が厳密な実験科学であったが，メディアが揃うことで起こる，数理科学による回転である．なら，どこにポジションを取るか？

3.1 治験とは何か

定義 3.1.1 (clinical research, case-history research, observational study, cohort / prospective / follow-up study, case-control / retrospective study, double blind test).

(1) ヒトを対象にした医学研究のことを**臨床研究**，それ以外を**非臨床研究**という．
(2) 臨床研究は，設計のない**症例報告**から，コホート研究＝追跡研究＝前向き研究や患者-対照研究＝後ろ向き研究などの**観察研究**，通常の診療行為以外の介入を行う**介入研究**または**臨床試験**という．新薬開発における介入研究を**治験**という．
(3) 抗がん剤を除いた治験は，3 段階に分けられる．第 I 相は安全性の検査，第 II 相は主に第 III 相の設計のための予備調査をする．第 III 相にて，効果の定量的な評価のために，盲検などの本格的なデザインがなされる．
(4) 二重盲検とは，患者だけでなく医者も，処置群か対照群か判別がつかないように設計された試験をいう．

注 3.1.2. 疫学的には，仮説生成のために患者-対照研究を行い，仮設検定のためにコホート研究を行う．症例対照研究は，マッチングなどにより因果効果の測定をする観察研究である．

3.2 FDA draft guidance for adaptive design

FDA Guidance document "Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry"を 2019 年 12 月に公開している．adaptation とは修正または変更である．これを ad hoc にするのではなく，by design で行う，という時代の進化をいう．しかし，adaptation は研究者やスポンサーにとっては魅力的であろうと，規制当局との衝突が起こるため，法学的な視点も必要となる．

歴史 3.2.1. FDA Modernization Act (97) では，"adequate and well-controlled clinical trials"によって被験薬の効果を認めることとしている．これは，試験目的，解析方法，デザイン，患者選択，患者割り当て，試験参加者，反応の評価，効果の評価の 8 要件を満たす protocol(実施計画書) によって認められる．デザインに組み込まれたプロトコル逸脱を，adaptive という．

定義 3.2.2 (adaptive design). a clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial.

注 3.2.3. "prospectively planned"により，バイアスが発生しない．

3.3 研究倫理

生命倫理学は 1960 年の米国から始まった．大きく分けて医療倫理と研究倫理がある．しかし，診療と研究の別とはなんだろうか？ヒトゲノムは個人情報であるか？

3.4 生存時間解析

Cox 72 は臨床医学の原著論文の中で最も頻繁に応用される統計文献となった．それは，これらの分野（医薬学・生物学・公衆衛生・疫学）では，「イベントの発生にまで掛かる時間」が主要な研究対象となるためである．なお，工学では信頼性分析，経済学では継続時間分析，社会学ではイベント履歴分析という．

歴史 3.4.1.

(1) 生存時間を推定するための最も古典的な方法として，生命表 (life table) は Halley 1656-1742 が発明した．58 の Kaplan-Meier 推定量も本質的に変わらず，この時点までその歴史が続く．
(2) 1960 半ばに群間比較の手法が取り入れられた．Wilcoxon の順序情報を利用するノンパラメトリック検定を，打ち切りがある場合に拡張した一般化 Wilcoxon 法 (Gehan 65) と log-rank 法 (Mantel 66) など．
(3) 60 後半から 70s にかけて，臨床研究への応用が盛んになり，共変量のモデル化が問題になり，Cox によりセミパラメトリックモデルが始まった．これは生存時間分布に全く仮定を置かないことになる．
(4) 1980s に確立過程論（特にマルチンゲール）が追いついて，Cox 回帰の理論的な正当化がなされた．

信頼性工学では，Weibull 分布を主にしたパラメトリックな手法が用いられ，臨床統計はログランク検定などのノンパラか Cox 回帰を使う．

記法 3.4.2. $X:\Omega\to\mathbb{R}_+$ を発生時刻とする．

定義 3.4.3 (survival function, hazard function / ratio, expected remaining lifetime). X は確率密度関数 $f:\mathbb{R}_+\to[0,1]$ を持つとする．$f(x)$ とは，年齢 x で死亡する確率である．

(1) $S(x):=P[X>x]$ により定まる $S:\mathbb{R}_+\to[0,1]$ を**生存関数**という．累積分布関数 F に対して $S=1-F$ が成り立つ．
(2) $h(x):=\dfrac{f(x)}{S(x)}$ により定まる $h:\mathbb{R}_+\to\mathbb{R}_+$ を**危険関数**または**（確率過程の）強度関数**という．時刻 x まで生存している
(3) $\mathrm{mrl}(x):=E[X-x|X>x]$ により定まる $\mathrm{mrl}:\mathbb{R}_+\to\mathbb{R}_+$ を**平均余命関数**という．
(4) $H(x):=\displaystyle\int_0^x h(u)du=-\log S(x)$ により定まる $H:\mathbb{R}_*\to\mathbb{R}_+$ を**累積危険関数**という．

命題 3.4.4 .

$$S(x)=e^{-H(x)}=\exp\left(-\int_0^x h(u)du\right).$$

要諦 3.4.5. 多くの分布族が f のモデルに使われる．Weibull 分布は Rosen and Rammler (33) が「粉末状炭の細かさを決定する法則」を記述するのに使用し，Weibull (39,51) が次に物質の寿命の解析に使用した．

3.5 打ち切りと切断

これらの研究対象として基本的な欠測データは，打ち切り (censoring) と切断である．尤度を用いた古典的なアプローチと，Aalen 75 による計数過程によるアプローチがある．こちらは，確率過程の方法論で，確率積分・連続時間マルチンゲールとの合わせ技になる．

定義 3.5.1 (counting process). 非負整数値の単調増加過程 $N:\mathbb{R}_+\to\mathbb{N}$ を**計数過程**という．Poisson 過程は計数過程である．

3.6 Cox モデル

また，集団が一様でない場合（多くの臨床試験，コホート調査，観測調査がそうである）は，共変量による線形回帰と誤差項を考えることになる．

記法 3.6.1. 2 つ以上のグループのイベント発生までの時間を比較し，共変量も調整することを考える．データ $(T_j,\delta_j,Z_j(t))$ $(j\in[n])$ を集める．$T_j\in\mathbb{R}_+$ は調査時間，$\delta_i\in 2$ はその期間内にイベントが発生したか，$Z_j\in\mathbb{R}^p$ は経時共変量とする．

模型 3.6.2. Cox モデルは，ハザード関数 $h(t|Z)$ を，乗法的関係・比例関係としてモデルする：

$$h(t|Z)=h_0(t)c(\beta^\top Z(t))\qquad(h_0\in\mathcal{L}(\mathbb{R}_+),c\in\mathcal{L}(\mathbb{R}),\beta\in\mathbb{R}^p).$$

これは，共変量 Z の効果についてのみ，$\beta\in\mathbb{R}^p$ というパラメトリックな仮定を置いている．ただし c は既知の関数として固定する．$c=\exp$ など．スケール h_0 については何の仮定もおかない局外母数とする．

第 4 章

疫学統計

臨床試験に関する統計学がミクロ生物学だとしたら，疫学とは人類と疾患要因（動物・菌類・ウイルスかもしれないし，環境や生活習慣かもしれない）との系を対象とするマクロ生物学である．僕が本当にやりたいことはここにあるかもしれない．
予防医療に関するデータ活用や，人間・生物を取り巻く生態系をデザインしたいとは思う．

4.1 公衆衛生

医師法第 1 条には「医師は，医療及び保健指導を掌ることによって公衆衛生の向上及び増進に寄与し，以て国民の健康な生活を確保するものとする」とある．公衆衛生とは，マクロで見た人類共同体の well-being をいう (absense of disease では足りない).

> Public health has been defined as "the science and art of preventing disease, prolonging life and promoting health through the organized efforts and informed choices of society, organizations, public and private, communities and individuals". Analyzing the determinants of health of a population and the threats it faces is the basis for public health.

疫学とは「ある人間集団を対象として，人間の健康と人間に生じる異常の原因を，宿主・病因・環境・行動のそれぞれの面から研究する分野」である．

4.2 実験計画法

歴史 4.2.1. 元々は Fisher が 20s に農学試験をモデルに理論化した．注目したい要因（これを**因子**という）による効果を**要因効果**，それ以外を**実験誤差**と呼ぶ．各因子（変数）のカテゴリ数（定義域の濃度）を**水準 (level)** という．要因効果と実験誤差とに分解して解析する手法を**分散分析**という．

公理 4.2.2 . 実験計画の 3 原則は以下の通りで，(3) が物理学にはない，特有のものである．

(1) 局所管理科：因子以外の要因を一定にする．
(2) 同条件でなるべく反復する．
(3) 無作為化，または，介入ができない場合は共変量調整

4.3 ブロックデザイン

edge が任意個数の頂点と結合できるグラフをハイパーグラフといい，ブロックデザインはハイパーグラフである．

複数の因子があるとき，どのブロックでどの因子を共通とするかの定め方が，「均一」になるデザインを「ブロック」という．これは特殊なハイパーグラフを構成する問題になる．その後分散分析がなされ，処置効果が推定される．最も効率よく処置効果が推定できるようなブロックデザインが問題になり，これを最適計画と呼ぶ．

第 5 章

計算統計学

統計的な手法を用いて，次元の呪いや爆発的な計算量を乗り越える手法をいう．統計学には計算機科学者の参入もあり，全く違う統計モデル観がある．生物統計学とは世界観が違う．特に隠れ変数がその萌芽で，ベイズ法というのは隠れ変数の極みとも思えるのではないか（推定量を推定分布としてしまう）．

(1) 統計量は解析的な計算でも求められるが，リサンプリングによる実験を手元で繰り返すことによって近似計算も可能である（ブートストラップ Efron 79）．
(2) 確率計算は時折，高次元の積分計算を必要とする．する．超一様分布列が用いられることもある．

5.1 リサンプリング法

歴史 5.1.1. ブートストラップは計算機を援用したリサンプリング法である．リサンプリング法自体の歴史は古く，1960s に Simon が社会学にて用いている．数式が増えてな統計学専攻生のシミュレーションの手段として使われていた．70s に計算機が成熟する前に，モンテカルロ法などの有効性が議論されていた．

歴史 5.1.2 (Bootstrap の正当化)**.** 70s で計算環境の整備によりブートストラップ法が提案され，その後 70s 後半に，これを正当化する 2 つの理論が生まれた．

(1) Huber 81 による頑健統計学の完成で，統計的汎関数の理論が広く知られた．
(2) Bhattacharya and Ghosh 78 の Edgeworth 展開理論も成熟した．

ノンパラメトリック最尤法と解釈できる．名前は"by one's own bootstraps"（自力で行う）から．なんの自力かというと電力供給を計算機にすれば良いだけ，という意味である．

議論 5.1.3. いろいろな滑らかさが十分条件になるが，例えば Tukey が普及させた jackknife 法 (Quenouille) は非復元リサンプリング法であるが，推定量 $\widehat{\theta}$ の経験分布関数に対する滑らかさが必要になる．ブートストラップを簡単に言えば，$H(F)$ が滑らかで，何かの期待値として表されているとき，近似可能である，という原理である．F を経験分布関数 $\mathbb{F}_n$ で推定するときは特にノンパラメトリックブートストラップという．

定義 5.1.4 . $T := T(X_1,\cdots,X_n)$ を θ の推定量とし，X_i を除いた $n-1$ 個のデータから定まる推定量を T_i とする．$J^i := nT-(n-1)T_i$ として，

$$J(T) := \frac{1}{n}\sum_{i=1}^n J^i$$

を **jackknife statistic** という．

定理 5.1.5 . T が一致推定量でなく，さらに $\exists_{\theta_1\in\mathbb{R}}\ E[T]=\theta+\dfrac{\theta_1}{n}$ を満たすならば，$J(T)$ は一致推定量である．

5.2 超一様分布列

歴史 5.2.1 (low-discrepancy sequences)**.** Hammersley and Handscomb "Monte Carlo Methods"(1964) では「準乱数 (quasi-random numbers)」と，疑似乱数と区別して呼ばれていた．

一様分布論は，空間内の点列の漸近分布を議論するエルゴード理論の一部であるが，これが"irregularities of distribution"なる分野に変化した．分布の不規則性＝一様分布からのズレの尺度を"discrepancy"と呼ぶ．

数値積分では 5,6 次元を超えると，「最後の手法 (last resort)」としてのモンテカルロ法しか残らないことが知られているが，収束が非常に遅い．1960s にすでに，モンテカルロ法の問題点を克服するために，「準乱数」が用いられていた．特に，ソ連での水爆開発でのモンテカルロ計算の高速化に研究されていた．Riemann 積分の代表点を「一様に」取ると，うまく積分の値に収束するかもしれない．特に MBS(morgage-backed securities) の価格計算では 360 次元の数値積分が登場するが，このような例でも著しい高速化が可能になった．

5.3　平均場近似

歴史 5.3.1. 多体問題を 1 体問題に帰着される手法として平均場近似が考えられていた (Weiss 07)．統計手法では，大自由度統計モデルに関する積分（平均）の計算において，特殊な分布属では計算の困難さが下がり，真の分布が近似できる．

統計計算が容易になるグラフ構造に関して，汎用的なアルゴリズムの形で初めて明確に議論したのが Pearl(88) である，これをグラフによる因果推論の議論と同時に発表したため，混乱されやすいが，全く独立の 2 つの議論である．

5.4　EM 法

議論 5.4.1. 統計モデルの変数は，観測変数／顕在変数と潜在変数／隠れ変数の 22 種類があり，音声認識での隠れマルコフモデルや，時系列解析の状態空間モデル，因子分析モデルなどの CS 的な統計モデルでは隠れ変数を持つモデルが多い．そのようなモデルのパラメータに対する最尤推定を行うアルゴリズムを EM 法 (Dempster, Larid, Rubin 77) である．

ただし，E ステップに計算困難性があり，例えばモンテカルロ法で近似できるが，以前計算量は大きい．変分近似を採用する手法を**一般化 EM 法**という．

5.5　変分ベイズ法

歴史 5.5.1. ベイズ推定は最尤推定と異なり，未学習データの予測値ではなく予測分布を求める．少ないデータ数に対して最尤推定よりも高い汎化能力を持つ利点はあるが，計算困難性が強い．特に積分計算について，従来は MCMC 法が用いられたが，ここに変分近似ほうを用いる手法は**変分ベイズ法**と呼ばれる．

5.6　マルコフ連鎖モンテカルロ法

歴史 5.6.1. 統計物理で 1950s に導入され（動的モンテカルロ法とも呼ばれる)，統計科学の分野に流入したのは 1990s で，これがベイズ統計のきっかけとなった．計算の困難さがアルゴリズムの発展によって除去されたことによる影響が大きい．また変数の離散・連続の別を撤廃したことが実質的には大きい．

基本的には種々の離散・連続分布から，解析的性質がよくわかっているか／どれほど高次元かに依らずサンプルを得るための手法である．

5.7　逐次モンテカルロ法

歴史 5.7.1. 分野によっては particle filter / Monte Carlo filter や，ロボティクスでは condensation と呼ばれる手法の総称．遺伝的アルゴリズムに似た面もあるが，状態空間モデルに基づく統計を背景にしている．

第 6 章

計量経済学と多変量・時系列解析

6.1 歴史

(1) 現代的な数理統計学の理論体系が確立したのは 1920s の Fisher, Neyman, E. S. Pearson らによる．
(2) この枠組み（相関分析など）に沿って，30s からは自然な形で多変量に拡張された．Fisher 自身も多くの貢献をしたが，Hotelling, Wilks, Wishart らの仕事が基礎になっている．
(3) 多変量推測統計の Anderson による標準的教科書"An Introduction to Multivariate Statistical Analysis"は 1958 年に刊行された．この頃には多変量正規分布と線型モデルに基づく推測理論の基礎は確立していた．
(4) 1970s 以降の計算機時代では，今日ではほとんどのデータは多変量データであり，取り立てて多変量分析と言わなくても，新たな統計手法は押し並べて多変量データを扱う手法となっている．今日の統計パッケージの中にある探索的データ解析や射影追跡などは，正規分布を前提としない，非線形的手法である．

[1] の序文．

例 6.1.1. 回帰分析，分散分析，主成分分析，判別分析，因子分析，分割表，グラフィカルモデル．

6.2 多変量解析の考え方

多変量解析においては，連続分布としては多変量正規分布以外に扱いやすい分布が少ない．ところが，実際のデータ解析の場面では多変量正規分布の仮定の妥当性が疑われる場面が多く，記述統計的な手法と推測統計的な手法の間に大きなギャップがあるのが現状．[1]

6.3 用語と記法

行列やベクトルを言葉として用いるのが特徴であり，経済学はこの Anderson らの慣習を強く受け継いでいる．

定義 6.3.1 (data matrix, sample size, dimension).

(1) 多変量データを**データ行列**とも呼ぶ．一般には行を個体，列を変数とする．それぞれの添字を t と i,j と使い分ける．
(2) $n\times m$ データ行列の行数 n を**標本の大きさ**といい，列数 m を標本の**次元**という．†1
(3) 値が数ではない変数を**質的変数**という．これを量的変数に変換する手法を**数量化**という．

注 6.3.2 (spatial statistics, random field)**.** データ行列の古典的な枠組みに入らない多変量データには，地理的なデータなどがある．これには**空間統計**や**確率場**の手法がある．また，質問表に分岐がある場合は，特定の回答をした者以外の値が**欠測値**となる．

†1 高次元データを，なんらかの形で 2 次元空間に写して解釈する記述統計的手法を，次元の縮約という．

6.4　経験論 vs 計量経済学

構造方程式モデルの前夜まで繋ぐ歴史がある．

歴史 6.4.1 (empirics を乗り越えた偉人たち：Slutsky, Koopmans と Cowles 委員会の時代)**.** 経済学では，Sir William Petty が 17 世紀イギリスにおいて経済データを記録して（作り出して）以来、経験的／実証的な研究がずっと重要な役割を果たしてきた．[†2]

(1) 帰納的アプローチ・記述統計学：データを慎重に検討することで経済学的な洞察が得られる．
(2) 理論的アプローチ・推測統計学：既存の統計理論の検証が大事（ちょうど実験物理学のように）．

Léon Walras 34-10 (仏) は経済分析に数学的手法を積極的に活用し，一般均衡理論を最初に定式化した．つまり，「需要曲線」を初めて定立した．これは，凸解析や不動点定理などでかなりの分析が可能な数学的に優れた構造を持ち，例えば均衡価格の存在が示せる．そこで弟子の Henry Moore は，統計的な検証をしようと試みる．さらにその弟子が同様の研究を 1930s まで続け，Pigou なども同様の研究をしている．その後すぐにアメリカで Wesley Clair Mitchell 74-48 を中心に，ビジネスサイクルの研究が National Bureau of Economic Research で始まった．その中心的な手法が帰納的アプローチであった．Schumpeter や Pigou は理論研究を志向したが，その統計手法は徹底されたものではなかった．NBER の研究は特に Eugene Slutsky 1880-1948 や **Tjalling Koopmans** 10-86 によって批判され，Cowles 財団を立ち上げ，これが数理経済学・計量経済学の先駆けとなった．Kantorovich とノーベル記念賞を受賞した．

歴史 6.4.2 (Haavelmo による確率論的アプローチによる Cawles 財団の加速)**.** Tinbergen はこの方向からビジネスサイクル理論に切り込み，Keynes と戦った．Tinbergen を乗り越える形で，計量経済学は一気に「確率論的アプローチ」の色を深めた，特に Haavelmo 11-．これに Cawles 委員会も応え，教科書なども整備されていった．

ある意味で、ホーヴェルモーの研究は昔の計量経済学者たちが事実上主張してきたことではあった。でも、それを確固たるものにしたのは ホーヴェルモーだ。クープマンス主導のコウルズ委員会 と、古参の NBER 制度学派たち（Rutlege Vining が筆頭) との間の手法論争は、1947 年から 1949 年まで続き、その結果として確率的理論研究が経験論的経済学での支配的な理論として確立した。

歴史 6.4.3 (Cowles 委員会の超克：Lucas の時系列)**.** Herman Wold は，連立方程式系の大規模モデルの「同時手法」を批難し，代わりに再帰的手法を支持した．**Robert E. Lucas** 37- (1976) が最も有名な批判を発表した。大規模モデルの構造パラメータはとりわけ政策に用いられるときには一定だと想定されている。でもこれは新マクロ経済学理論の本流になりつつあった理論の主張とは一致しないのだ、と彼は主張したのだった。 別の一連の論文で ルーカスは新しい計量経済学の方法論を述べた。それが**時系列の計量経済学**だ。

6.5　構造方程式モデル

構造方程式・同時方程式 (simultaneous equiation) は，なんと，数理統計学の「推定方程式」なる用語と同義である！

歴史 6.5.1 (同時方程式モデル)**.** 29 世界恐慌で，統計的な手法の変革が迫られた．特に A. Cowles がコウルズ委員会を創始し，現在の Cowles 財団となっている．この委員会のモノグラフ Koopmans (50), Hood and Koopmans (53) が同時方程式モデルの始まりである．Klein (50) がこれをマクロ計量モデルに応用した．

同時方程式は 2 本などに見えても，実際は無数に存在し，見かけ上は全ての方程式を同時に満たす．（満たすべき方程式の線型空間でも生成しているのか？）この全体を構造型または構造方程式と呼ぶ．

[†2] 労働価値説を初めて唱え、政治算術派の先駆となったことから、古典派経済学と統計学の始祖ともいわれる。

6.6　状態空間モデル

観測方程式と推移方程式の組みをいい，推移方程式により変化していく観測されない変数（**状態変数**）を推定することが問題になる．将来の値を予測する推定問題に加えて，現在利用可能な変数から現在の状態変数を推定するのが**濾過問題 (filtering)**，過去の状態変数を推定するのが**平滑問題 (smoothing)** である．

6.7　プロビットモデル

割り当てなど，説明変数がダミー変数になることはよくあるが，結果変数がダミー変数になるとき，制限被説明変数モデル (limiteddependent variable model) という．ここで，条件付き確率 $P[Y_i = 1|X_i] = \Phi(\alpha + \beta X_i)$ を考えるとき，Φ が正規ならば**プロビットモデル**，ロジスティックモデルならロジットモデルである．Tobin (58) は Tobit モデルも開発した．

第 7 章

フィルタリング理論と工学

7.1 歴史

歴史 7.1.1. Kalman (60), Kalman and Bucy (61) による．70s から経済にも応用された．

第 8 章

頑健統計

ある時期，正規分布から外れた状態の下でもロバストであるような推定量についての研究が盛んに行われたことがあった．その結論としては，位置母数に関する限り，正規分布を想定した推定方法は，正規分布からのズレが著しく大きくない限り十分効率が良い＝ロバストである．また，正規分布からのズレが大きい時も，標本数が著しく小さくない限り，そのことが検出できることがわかる．

8.1 歴史

歴史 8.1.1. 1920s に Fisher が統計モデルの量を倍にしたが，それらは仮定に強く依存するものであった．そこで，「分布に仮定を置かなくても使えるモデル」が志向されたのと同様に，パラメトリックモデルの表現力と両立させるために，セミパラメトリックモデルとその上の頑健統計が志向されるのは来たる研究テーマとして当然の流れであろう．

Huber (64) が初めてこの問題に挑み，「真の分布」を含むようなモデルとその近傍を考察し，そのようなモデルのすべてでうまく働く推定量のクラスを確定した．引き続き複数のアプローチを開発し，Robust Statistics (81) にまとめた．一方 Hampel (68, 71, 74) は影響関数による手法を開発した．正則関数との類比でいえば，微分が影響関数で，特異点が breakdown point である．これは Huber (64) より簡略化され，さらに適応範囲も広い（最尤法と尤度比検定が使えるすべてのモデルに使える）数理手法と言える．

8.2 議論

母数 θ を観測した値 X_i から推定することを考えると，$X_i = \theta + \epsilon_i$ と表せる．

例えば誤差分布に正規分布を仮定することは，よく管理された実験を表す．もっと裾が長いことの方が考ええる．このときに，正規分布に対する精度が落ちても，頑健な統計量を考えたい．M 推定量の改良として，最小化する目的関数を

$$\sum \rho(X_i - \theta) \quad \rho(u) = u^2 1_{|u| \leqslant k} + (2k|u| - k^2) 1_{|u|>k}$$

とすることによってこれを与えたのが Huber である．

8.3 von Mises 解析

歴史 8.3.1. von Mises (47) で意識されたのは"Volterra derivative"である．これは 60s の頑健統計の勃興，特に Hampel の博士論文 (68) まで待って，「影響曲線」なる概念に進化した．1980s に，Frechet 微分，Hadamard 微分，コンパクト微分の言葉で基礎づけがなされた．([6] 全文)．

> The purpose of these notes is to provide von Mises' theory with a rig- orous mathematical framework which is sufficiently straightforward so that it can be applied routinely with little more effort than is required for the calculation of the influence curve. The approach presented here is based on the Hadamard derivative and is applicable to diverse forms of sta- tistical functionals.[6]

第 9 章

高次元統計学

9.1 lasso 推定量

定義 9.1.1 (least absolute shrinkage and selection operator).

第 10 章

極値統計学

Emil Julius Gumbel 91-66 は Heidelberg 数理統計学教授の座を追われたユダヤ系数学者で，最終的には Fisher と極値統計学を築いた．統計学の中でも特殊な分野と見做されるかもしれないが「この外れ値を採用すべきか否か？」という問題を数理化する非常に基本的な試みである．その開拓者はなんと Frechet (27) と Fisher (28) であった．

第 11 章

時系列解析

確率変数列 $X_1, \cdots, X_n$ の実現値と見做し得るデータに対する統計解析を多変量解析と呼ぶならば，確率過程の実現値と見做し得るデータに対する統計解析を時系列解析という．そこで，

11.1　歴史

(1) 1940s より Wiener や Kolmogorov により弱定常過程のスペクトル解析と予測の理論が確立される．特に連続時間の確率過程の理論の精緻な発展の原動力となった．
(2) 1976 年刊行の George Box と Jenkins による"Time Series Analysis, Forecasting and Control"が，自己回帰移動平均 (ARMA) モデルと呼ばれる線型モデルの構築と予測のための標準的な手続きが確立した．これ以降は ARMA モデルの限界を乗り越えることを念頭に，非線形モデルなどさまざまな手法が提案されることになる．

通常の統計学は主に独立標本に対する議論であるが，時系列解析は時間軸の間にも従属関係がある状況での統計解析であり，その点で一般化であると捉えられる．

11.1.1　パラメータ付け

まず，時系列データを定量的に特徴付ける (パラメータ付ける) 手法を考える．

(1) 時間領域法：自己相関など．
(2) 周波数領域法：スペクトル解析など，

定義 11.1.1 (autocorrelation function). 確率過程の共分散関数の概念を自己相関または自己共分散といい，相関係数版の概念を**自己相関関数**という．

注 11.1.2. しかし，実際のデータから推定しにくい，視覚化しにくいなどの難点がある．

11.1.2　推定法

等間隔標本 $(X_t)_{t\in\mathbb{Z}}$ の i.i.d. を仮定すると Glivenko-Cantelli の補題から一致推定が可能であるが，一般には他の仮定を必要とする．

11.2　基本的性質

i.i.d. 以外に等間隔標本 $(X_t)_{t\in\mathbb{Z}}$ にどんな仮定をおけるかを考える．

11.2.1　エルゴード性

記法 11.2.1. はじめの n 回の観測からの標本平均を $\widetilde{x} = \widetilde{x}_n := n^{-1} \sum_{t=1}^{n} X_t$ とする.

定義 11.2.2 (EPCL: ergodic property with a constant limit).

$$\exists_{\mu \in \mathbb{R}}\ P\left[\lim_{n\to\infty} \widetilde{x} = \mu\right] = 1.$$

11.3　定常過程

11.4　スペクトル解析

11.5　リード・ラグ効果と CCK 理論

定義 11.5.1 .

(1) 2 つの時系列がタイムラグを持って相関する現象を lead-lag 効果という.
(2)

要諦 11.5.2. 現状の取引の流動性では，ms の世界で存在する．結果，80s までは分のスケールで観測されていた現象だが，現在は高頻度データと呼ばれる分野となってしまった．そこで，古いモデルが使えなくなってしまったこの現象に対して，Hoffmann, Rosenbaum, and Yoshida (2013) が新しいモデルと推定法を提案した.

第 12 章

乱数

サイコロを使う，国勢調査報告書の中の数字を抜き出す．擬似乱数という計算機援用による方法が von Neumann によって始められた．これは「一様性」を満たせば良いので，「予測不可能かどうか」を厳密な意味で満たしている必要はない．

12.1 一様乱数の生成法

Lehmer (1948) による線形合同法 (linear congruential method) が 40 年に渡って使われてきた．が，高次元化や，大量の乱数を作る実験には不向きである．そこで現代的には，M-系列を用いる方法などがある．

模型 12.1.1. よく選択された a, M について，

$$X_n \equiv aX_{n-1} + c \mod M$$

を用いて非負正数列 (X_n) を生成する．

要諦 12.1.2. 高次元になればなるほど規則的にバラバラになってしまう．これを「多次元疎結晶構造」という．そこで，高次の項を追加することも考えられる．

定義 12.1.3 . 次の 3 条件を満たす独立な Rademacher 確率変数列 $(X_n)_{n\in[T]}$ を**擬似ランダム系列**という．

(1) $\displaystyle\lim_{T\to\infty}\frac{|\{n\in[T]\mid X_n=1\}|}{|\{n\in[T]\mid X_n=-1\}|}=1.$

(2) $\displaystyle\lim_{T\to\infty}\frac{|\{n\in[T]\mid X_n=X_{n+1}=\cdots=X_{n+k+1}\}|}{|\{n\in[T]\mid X_n=X_{n+1}=\cdots=X_{n+k}\}|}=1.$

(3) $\displaystyle\lim_{T\to\infty}\frac{1}{T-s}\sum_{t=1}^{T-s}X_tX_{t+s}=1_{s=0}.$

要諦 12.1.4. 自己相関のない Rademacher 確率変数を，位相をずらして各桁とすれば良い．

12.2 一様乱数の変換

$X := F^{-1}(U)$ とすれば，所望の分布が得られる．これを逆関数法という．

記法 12.2.1. 分布関数 F に対して，$F^{-1}(y) = \inf\{x\in\mathbb{R}\mid F(x)\geqslant y\}$ と定める．この値を quantiles という．

12.3　統計的検定

例えば合同法に対するスペクトル検定によって，乱数の精度を検証できるが，1 周期全体の帯域的性質はわかっても，局所的な性質は検定に頼って泥臭く検証していくことになる．これは結局適合度検定であり，χ^2-検定や Kolmogorov-Smirnov 検定となる．K-S 検定では

$$K_n := \max(K_n^+, K_n^-), \qquad \left(K_n^+ := \sqrt{n} \max_{x \in \mathbb{R}} (F_n(x) - F(x)), K_n^- := \sqrt{n} \max_{x \in \mathbb{R}} (F(x) - F_n(x)) \right).$$

なる統計量を用いる．これは可測性が怪しいはず．

第 13 章

数理物理学

13.1　浸透理論

percolation 理論は，ランダム系の統計的考察をする数理工学の分野と，spin glass (磁性体のスピンがアモルファスのように乱雑なまま固まった物質のこと) における秩序の拡散を記述する．狭義には，相転移の最も簡単なモデルをいう．

第 14 章

参考文献

参考文献

[1]『統計科学のフロンティア１　統計学の基礎Ｉ　線型モデルからの出発』（第 I 部「多変量解析入門」（竹村彰道），第 II 部「時系列解析入門」（谷口正信））
[2] E. L. Lehmann - Nonparametrics: Statistical Methods Based on Ranks
[3] Shein-Chung Chow, Mark Chang - Adaptive Design Methods in Clinical Trials
[4] シリーズ生命倫理学『医学研究』
[5] John P. Klein and Melvin L. Moeschberger - Survival Analysis
[6] Luisa T. Fernholz. (1983).