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．

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 Wilcoxon の順位和検定

2.1.1 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．

2.1.2 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 2.1.1. 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 2.1.2. 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．

2.1.3 統計量の分布

Hajek (1961) による理論の特殊化になる．

記法 2.1.3. 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 2.1.4 (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 2.1.5. 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 2.1.6. 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 2.1.7. $\Pi_N := [N]$ のとき，S_N は Wilcoxon 統計量になる．

2.2　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 2.2.1. $\varphi : \mathbb{R}^2 \to \mathbb{R}$ 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 2.2.2. 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 2.2.3. $m \leqslant n$ 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 2.2.4. $X_1, \cdots, X_N$ 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

多変量解析

3.1 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 3.1.1. 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．

3.2 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]

3.3 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 3.3.1 (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 3.3.2 (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．

第 4 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？

4.1 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 4.1.1. 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 要件からなる．

定義 4.1.2 (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 4.1.3. "prospectively planned"により，バイアスが発生しない．

4.2 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？

4.3　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 4.3.1.

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

定義 4.3.3 (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 4.3.4.

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

要諦 4.3.5. 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．

4.4　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 4.4.1 (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．

4.5　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 4.5.1. 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 4.5.2. 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 5 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．

5.1　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．

第 6 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，

6.1 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．

6.1.1 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 6.1.1 (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 6.1.2. 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．

6.1.2 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．

6.2 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．

6.2.1　エルゴード性

記法 6.2.1. はじめの n 回の観測からの標本平均を $\tilde{x} = \tilde{x}_n := n^{-1} \sum_{t=1}^{n} X_t$ とする．

定義 6.2.2 (EPCL: ergodic property with a constant limit)**.**

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

6.3　定常過程

6.4　スペクトル解析

6.5　リード・ラグ効果と CCK 理論

定義 6.5.1.

(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 6.5.2. 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 7 章

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 8 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