目次

第 1 章

因果推論

統計的因果推論の研究は次の 2 つに分類できる．

(1) 因果グラフを所与として，因果関係を定式化する実証分野．Rubin による因果モデル RCM と Pearl による構造的因果モデル SCM とがある．

(2) 因果グラフを未知として，因果グラフの構成算譜を定式化する理論分野．これを**統計的因果探索**という．

これは数理科学の拡張の過渡期における重要な瞬間であり，物理学の意味から「実験」を「計算」に拡張して汎神化する過程である．この過程を経た後に，学問のあり方は大きく変化する．このフェーズでは数学も変化が必要である．私が憧れたのはこの未来感であり，未来を見せた啓蒙に対する数学の責任であり，属人化される知である．

まさに私の目前で，相関関係を超えた因果関係の数理構造が定式化されつつある．この営みに乗らないはずがない．事事無礙の法界を写すことが数理の営みであるのならば，因果推論は莫大な霊性源とならないはずがない．

1.1 歴史

歴史 1.1.1 (調査観察データに対する因果推論の潮流).

(1) Donald Rubin の因果モデルは手法としては「欠測データ解析の方法論」として結実した．EM アルゴリズムと多重代入法など．その後の MCMC 法を用いたベイズ統計学の発展にも寄与し，統計学のパラダイムに大きな変化をもたらした．

(2) もう一つの流れが，モデル仮定をなるべくおかずに推定することを目指すセミパラメトリック推定法である．Huber の M-推定，Liang & Zeger の一般化推定方程式など．医学データなど，経時的多変量データや，共変量が多い場合の解析に強い．

(3) 共変量と結果変数の回帰関係を仮定せずに，共変量の影響を削除する傾向スコアが頻繁に利用されるようになったが，これも Rubin のセミパラメトリック解析の一種である．

(4) さらに，Pearl 流の手法としては「構造方程式モデル」と「ベイジアンネットワーク」とに代表される．Wright のパス解析のノンパラメトリックモデルへの拡張と言える．[†1]

歴史 1.1.2 (セミパラメトリックモデル). 医学や経済学での調査観察研究では共変量を非常に多く考える必要があるため，線型の回帰モデルではデータの説明力が低く，またカーネル法などのノンパラメトリック法ではいわゆる「次元の呪い」の問題が完全には解決できない．CS 的に後者を解決する突破口ももちろんある．

1.2 枠組み

「潜在」の接頭語は，結果変数の確率変数化をいう．

結果変数は，割り当てや共変量などが定める可測空間 $(\Omega, \mathcal{F}, P)$ 上の可測関数であるが，これを測るための情報 $\mathcal{F}$ を人類は持ち得ない，という形式化が，Rubin 因果モデルである．そこで，人類に可測な範囲への射影を取ることは（条件付き）期待値を取ることとなり，これを平均処置効果という．

記法 1.2.1 (outcomes that would have been observed under some intervention).

[†1] 構造的因果モデルについて．黒木 学, 小林 史明

(1) A をどのような処置を行ったかを表す独立変数とする．
(2) Y^a によって，$A = a$ とした場合の独立変数 Y の実現値とする．
(3) T-次元の場合，多重指数記法 $\overline{A}_T := (A_1, \cdots, A_T)$ を採用すると，$Y^{\overline{a}_T}$ で表す．
(4) $G : \mathcal{X} \to \mathcal{A}$ を確率的な介入 (stochastic intervention) とする．これは，共変量 $X : \Omega \to \mathcal{X}$ の値に対して，$A = 1$ を分配する確率 $g(x)$ を表す確率密度関数である．Y^G で，この処置を行った場合の反応を表す．

1.2.1　汎関数推定

記法 1.2.2. 真の分布 $\mathbb{P}$ に対して，$\mathbb{P} \in \mathcal{P}$ を満たす統計モデル $\mathcal{P}$ を立てる．これ自体を推定するのではなく，パラメータなど，汎関数 $\psi(\mathbb{P})$ の推定を考える．

例 1.2.3. 汎関数の推定量 $\widehat{\psi}$ を立てるのに，次のような手法が考えられる．

(1) パラメトリックベイズ，MLE（最尤推定法）
(2) ノンパラメトリックな最尤推定法，plug-in.
(3) ノンパラメトリック影響関数に基づいた手法．double machine learning や targeted leaerning ともいう．

注 1.2.4. 統計学的な数理的課題と，因果論的な話題はいつだって区別するべきである．「どうやって」推定するか，と，「何を」推定するべきか．反事実標本上の確率分布を $\mathbb{P}^*$ としたとき，$\psi^*(\mathbb{P}^*) = \psi(\mathbb{P})$ と標本分布 $\mathbb{P}$ に翻訳するところまでが因果推論における形式論で，その後起こることは純粋な汎関数推定の問題であり，数理的な状況に対処する段階とを峻別するべきである．

例 1.2.5 (因果推論以外の文脈で生じる汎関数推定)**.**

(1) integrated squared density : $\psi = \int p(x)^2 dx$.
(2) entropy : $\psi = -\int p(x) \log p(x) dx$.
(3) support size : $\psi = \Sigma_x 1_{\{p(x)>0\}}$.
(4) mutual information : $\psi = \int\int p(x,y) \log \dfrac{p(x,y)}{p(x)p(y)} dxdy$.

1.2.2　潜在反応モデル

公理 1.2.6 (ATE)**.** 次を仮定する．

(1) Positivity : $\exists_{\delta>0}\, \mathbb{P}[\mathbb{P}[A = a|X] \geqslant \delta] = 1$.
(2) Consistency : $A = a \Rightarrow Y = Y^a$.
(3) Ignorability : $A \perp\!\!\!\perp Y^a | X$.

このとき，

$$E[Y|X, A = a] = E[Y^a|X, A = a] = E[Y^a|X]$$

より，$E[Y^a]$ という仮想上の量を，$E[E[Y|X, A = a]]$ という求めることが可能な量によって推定出来る．

> 反事実，潜在反応などの用語が出来た．$E[y_1] - E[y_0]$ をはじめに定義したのは Rubin 1974 で，Rosenbaum & Rubin 1983 は「平均処置効果」と言い直した．
> このモデルの上に立脚した統計的因果推論には，Rubin 流の RCMs(Rubin's Causal Models) と Pearl 流の構造的因果モデル (SCMs) との 2 つがある．Judea Pearl はベイジアンネットワークの理論を体系的に整備した人工知能研究者である．グラフィカルモデルがよく用いられ，実際 Judea Pearl の論文はすべてグラフ用語でも記述されている（この表記の重要性は強調されるが，必ずしも理論的に不可分ではない）．

公理 1.2.7 (counterfactual model / approach)**.** ある行為をした場合による事実と，反事実の乖離を**因果関係**という．この定立を**反事実モデル**という．

定義 1.2.8 (potential outcome, Fundamental Problem of Causal Inference, causal effect, average treatment effect)**.**

(1) 独立変数の値域の濃度と同数の，仮想的な従属変数を**潜在的な結果変数**という．「仮想的な」というのは，独立変数は 1 つの値にしか実現しないので，1 つを除いて欠測することが約束されていることをいう．このことを**因果推論の根本問題** (Holland 1986) という．
(2) 潜在的な結果変数の差 $y_1 - y_0$ を**因果効果**という．一方で，潜在的な結果変数は 1 つを除いて欠測することが約束されているため，畢竟**平均処置効果** $E[y_1 - y_0] = E[y_1] - E[y_0]$ を考えることとなる．

注 1.2.9. もし群別が無作為ならば，独立変数と（潜在的な）従属変数はすべて独立であるから，欠測を無視して，観測された各群の平均値の差で平均処置効果を不偏推定出来る．

定義 1.2.10 (avarage treatment effecti on the treated, avarage treatment effect on the untreated, quantile treatment effect)**.**

(1) $TET := E[y_1 - y_0 | z = 1]$ を**処置群での平均介入効果**という．政策評価など，こちらのほうが重要視される場面もある．
(2) $TEU := E[y_1 - y_0 | z = 0]$ を**対照群での因果効果**という．
(3) $Q_\alpha(a)$ を，変数 a の $100 \times (1 - \alpha)$ パーセント分位点とする．$Q_\alpha(y_1) - Q_\alpha(y_0)$ を**分位点での因果効果**という．所得の分布など，期待値でない特性値に関心がある場面もある．

定義 1.2.11 (intermediate variable, moderator variable, covariate / control variable / confounding factor)**.** 独立変数，従属変数の他に，

(1) 独立変数の関数で，従属変数がその関数であるような変数を，**中間変数**または媒介変数という．これを介して起こる因果効果を間接効果という．[†2]
(2) 独立変数がこの関数となっているような変数を，**調整変数**という．この変数が定める同値類で，従属変数と独立変数との関係が大きく変わるような変数である．
(3) 従属変数と独立変数が，いずれもこの関数であるような変数を**共変量**や統制変数という．医学分野では交絡要因ともいう．

注 1.2.12. これらの峻別，特に中間変数を共変量と誤認することは因果効果を過小評価することに繋がる．

1.2.3　区間推定の営み

空想上の量 ψ^* は点推定可能でないときの方が良い．そのときでも，ψ^* の bounds が求まるときは多い．そこで，代わりに bounds を推定することを考えることが出来る．Manski など．

1.3　研究の種類

すべてが実験になる

独立変数への研究者による介入があるかどうかで，実験研究と観察研究を分けることが出来る．独立変数の群別への割り当てが無作為であるかどうかで，さらに細かく分類することが出来る．
因果効果は処置の有無が産む差と定義したのが潜在反応モデルであるが，因果推論の根本問題どころか，処置を施すこと自体が不可能な場合が多い．

[†2] 間接効果と直接効果を併せたものを総合効果といい，社会科学では多くこれを取り扱うことになる．

1.3.1　実験研究

実験研究または統制実験という．

定義 1.3.1 (experimental / treatment group, control group)**.** 無作為割当について，処置を行った群を**実験群**または**処置群**，行っていない群を**対照群**という．

1.3.2　調査観察研究

例 1.3.2.

(1) ランダム化比較試験などは，無作為割り当てを伴う．計量経済学では社会実験ともいう．EBM においてはメタアナリシス＝システマティックレビューの次に根拠が高いとされる．
(2) メタアナリシス＝システマティックレビューなど，１次研究の結果を統合するプロセスを **Research synthesis** という．
(3) 一般に**相関研究** (correlational study) や**観察研究** (observational study) は，無作為割当を伴わない研究を言う．計量経済学では，独立変数の値が共変量の値によって確率的に決定され，いわゆる独立変数に**内生性**がある場合の研究を言う．
(4) 政策評価では，自然実験なる概念もある (EBPM)．これは 2021 にノーベル経済学賞の主題であった．似た概念を，EBM ではコホート研究という．[†3]
(5) EBM でさらに一段階下の観察研究が，後ろ向きコホート研究とも呼ばれる，症例対照研究である．[†4]

1.4　RCMs：Rubin 因果モデル

因果推論の根本問題を，欠測データの枠組みで捉える．このとき用いることが出来る情報は，共変量情報である．このときに行う解析法は，パラメトリックだと非線形関係を取りこぼし，ノンパラメトリックでは共変量がそもそも多いために次元の呪いに陥る．そこで，特に関心のある部分はパラメトリックに仮定し，共変量が影響する関心のない部分はモデリングを避ける．これをセミパラメトリックモデルという．

1.5　SCMs：構造的因果モデル

数値への意味論の付与の仕方はただ一通りのみ許す定立を反事実モデルという．すなわち，因果効果の測定は，相関係数とは何の関係もなく[a]，（平均）因果効果と呼ばれる数量によって評価する．
次の問題として，構造的因果モデル M に数学的対象を付与する関手をあてがう．この方式は創造的行為であり，線型非線形を飛び交う議論になる．
構造的因果モデルの定義がやけに数学基礎論的に提示されたことが希望をくすぐる．これはこれから種々の数学手法を導入し，最終的には計算機の上に実装することが究極の祈りではなかろうか？
参考：[b]

[a] 相関関係と因果関係の乖離を一般に疑似相関という．
[b] 構造的因果モデルについて．黒木 学, 小林 史明

[†3] 特定の要因に曝露した集団と曝露していない集団を一定期間追跡し、研究対象となる疾病の発生率を比較することで、要因と疾病発生の関連を調べる観察研究の一種である。要因対照研究（factor-control study）とも呼ばれる。
[†4] 疾病に罹患した集団を対象に、曝露要因を観察調査する。次に、その対照として罹患していない集団についても同様に、特定の要因への曝露状況を調査する。

1.5.1　構造方程式モデルと因果グラフ

因果関係は 2 上の豊穣圏として定式化できる．この射をデータ生成過程と見て，確率分布の間の変換として定式化する数理モデルを，構造方程式モデルという．

記法 1.5.1 (数学基礎論的記号設定)**.**

(1) 大文字は確率変数，小文字はアルファベットとする．ここに目的言語とメタ言語の構造がある．
(2) これを用いて，代入 $=$ を $A=a$ と表し，介入という意味論を持つ．または $\mathrm{Do}(x=c)$ とも表す．モデルに対して $M_{x=c}$ という記法も，目的言語に於ける代入を意味する．
(3) 右上には個体名／添字を書く．反変ベクトルであるためである．

公理 1.5.2.

(1) (mean exchangability) $\forall_a\ E[Y_a|A=1]=E[Y_a|A_0]$．2 つに分けた集団が仮に逆であっても平均を変えない性質をいう．
(2) (consistency) $\forall_a\ E[Y_a|A=a]=E[Y|A=a]$.

例 1.5.3 (confounding, randomization)**.**

(1) 次の DAG が成立している場合（**交絡要因の存在**），平均交換律は成り立たない．
(2) 十分に大きな集団で無作為割賦を行うと，平均交換律を満たす．[†5][†6]

定義 1.5.4 (SEM: Structural Equation Model)**.** 4-組 $M=(v,u,f,p(u))$ を**構造方程式モデル**という．特にデータ生成過程 f をモデルに含む点が特徴的であり，f が $p(u)$ から，内生変数の確率分布 $p(v)$ を引き起こす点が特徴である．[†7]

- $v\in\mathbb{R}^p$ を内生変数 (endogenous variable) という．
- $u\in\mathbb{R}^q$ を外生変数 (exogenous variable) という．
- 関数 $f:\mathbb{R}^q\to\mathbb{R}^p$ をデータ生成過程という．
- $p(u):\mathbb{R}^q\to[0,1]$ は外生変数の確率分布を与える．

定義 1.5.5 (causal graph / path diagram / causal Bayesian network / DAG:directed acyclic graph)**.**

(1) データ生成過程のモデルを作る上での仮定を表現する図である．[†8]

1.5.2　構造的因果モデル

確率論に言語 Do を足したもの，と考えられる．こうして数学基礎論的にまとめているのは，伊藤清に加えた新たな純粋数学をどう作るかの気概を感じる．SCM は Judea Pearl (1995, 2009a) が導入．$f,p(u)$ によりデータ生成過程がモデルに入っていることが特徴である．

定義 1.5.6 (Structural Causal Models)**.** 構造方程式モデル $M=((x,y),e_y,f,p(e_y))$ に対して，介入 $M_{x=c}$ は次を構造方程式とするモデルとなる：

$$x=c,\qquad\qquad y=f_y(x,e_y).$$

(1) 集団において，x は y の原因になるとは，次が成り立つことをいう：$\exists_{c,d}\ p(y|\mathrm{Do}(x=c))\neq p(y|\mathrm{Do}(x=d))$.

[†5] ランダム化したにもかかわらず偶然で Exchangeability が成立しなかった場合には、共変量の調整によって後述する Conditional Exchangeability を目指していきます。[4]
[†6] 現実の世界では、割付された介入に従わない人が少なからずいます。例えば薬の効果を調べようとしてランダム化を行ったときに、薬を割付られたのにめんどくさがって薬をきちんと飲まない人がいるかもしれません。このような状況を割付への non-compliance と呼びます。[4]
[†7] * このデータ生成過程を記述しないのが多くの統計学や機械学習のモデルとなる。[1]
[†8] They can also be viewed as a blueprint of the algorithm by which Nature assigns values to the variables in the domain of interest.https://en.wikipedia.org/wiki/Causal_graph

(2) $E(y|\mathrm{Do}(x=d))-E(y|\mathrm{Do}(x=c))$ を平均因果効果と呼ぶ.

1.5.3　グラフィカルモデリング

宮川雅巳氏の著書が草分け.

議論 1.5.7 (工学分野が主流)**.** 例えば社会科学分野であまり使われていない理由は，中間変数間の部分的な独立性や変数の親子関係などの仮定が明確である場合が少ないからである．むしろそれらの確定が目的のひとつだったりする.

1.6　準実験デザインでの因果効果の推定

無作為割当が出来ない場合（観察研究）は，ロバスト性を目指して，共変量を利用することで，無作為割当の状態を推定し，因果効果の推定値を出すことになる．操作変数法と回帰分断デザインと中断時系列デザインは，観測されていない交絡因子も補正できる．差分の差分析は良くデザインすれば，ある程度ロバストである．そのほかは，全ての交絡因子が測定されているわけではない場合，非常に脆弱である.

定義 1.6.1 (marginal structural model)**.**

(1) 共変量 x の影響を除去した，潜在的な結果変数の周辺期待値を考えることを，**周辺構造モデル**という.
(2) 共変量の値に依存しない量を得るために，共変量の分布について平均を取ることを，**共変量についての周辺化**または**共変量調整**という.

1.6.1　目指すもの：強く無視できる割当条件

RCT のように比較が出来るためには,「共変量の全てが観測されている」=「十分な情報を持っている」ことが必要である．このことは，共変量 x が，これで条件づけると割り当て確率と潜在的結果変数とが独立になる，という条件に翻訳できる.

公理 1.6.2 (strongly ignorable treatment assignment (Rosenbaum & Rubin 1983), conditional effect, marginal effect)**.** 割当は，共変量のみに依存し，結果変数には依存しないとする仮定を**強く無視できる割当**という．これは，船内的な結果変数 $(y_1,\cdots,y_n)$ が，割当変数 z と，共変量 x の値についての条件付き確率について独立であること $(v_1,v_0)\perp\!\!\!\perp z|x$ と同値．また，これは欠測が RAM であることの十分条件である．このとき，共変量の値を所与とした結果変数の条件付き期待値の差 $E[y_1-y_0|x]$ を**条件付き効果**，共変量の影響を除去したあとの因果効果を**周辺効果**という.

補題 1.6.3 (mean independence)**.** 共変量が同じ対象者については，処置群と対照群で潜在的な結果変数の期待値は同じである：

1.6.2　操作変数法

無作為割り当ての理想的な状況からの乖離要因を考慮する．因果推論が叫ばれるずっと前から計量経済学の手法であった．傾向スコア解析で補正出来ないものは「観測されていない共変量」であり，これについては相変わらず操作変数法が用いられる.

定義 1.6.4 (instrumental variable)**.** 介入に影響を与える変数であって，介入を通じてのみしかアウトカムへの影響を持たない変数をいう．これを数学的に抽象化すると，回帰モデル $y=k(x)+e$ において，x,e は独立ではない（観測されていない共変量が存在する）とき，e と独立な確率変数 v を**操作変数**という．なお，条件 $E[e|X]=0$ を計量経済学では**外生性**といい，「内生性問題の解決に，操作変数法が用いられる」と表現される.

注 1.6.5. 数学的には単純なことであるが，$e \perp\!\!\!\perp v$ が実証的に証明できない場合が多いため，ある種の禁忌肢である．

例 1.6.6. 心臓カテーテル治療を介入とする実験において，自宅から病院までの距離は操作変数である．重症度が交絡因子となり，介入を受けられるかどうかを決定してしまっているためである．

公理 1.6.7. IV 法を用いるための前提条件は

(1) Exclusion restriction：IV が直接アウトカムに影響を与えず，介入因子のみを通じる．
(2) No instrument-outcome confounder：IV とアウトカムの両方に影響を与える交絡因子は存在しない．
(3) Instrumental relevance：介入因子を強く予測する．

例 1.6.8 (局所処置効果)**.** 不服従者を除いた「割り当て服従者の中での因果効果」を LATE という．このとき，割り当て自体が，実際に処置を受けるかどうかの操作変数になる，という理解をする．

1.6.3　回帰分断デザイン

自然実験がたまたまこの条件を満たすことがある．

歴史 1.6.9 (Regression discontinuity design)**.** Thistlethwaite and Campbell 60 による．

定義 1.6.10. 説明因子が $X \in 2$ のとき，ある連続変数 Z であって，$\exists_{c \in \mathbb{R}}\ Z > c \Rightarrow X = 1 \wedge Z < c \Rightarrow X = 0$ を満たすとき，c の近傍に Z の値を持つサブグループは，次の条件を満たすとき強く無視できる割り当て条件を満たし得る．

(1) カットオフ値 c の近傍において，アウトカムに影響を与えるそのほかの因子は大きく変化しない．
(2) カットオフ値 c の近傍において，潜在的アウトカムは連続である．

1.6.4　差分の差法

介入群と対照群のそれぞれについて，処置の前後の計 4-組のデータが得られたときに使える解析法である．

1.6.5　マッチング（均衡化）

共変量の値で同値類を作り，その中で処置群と対照群とに分ける．

当然，共変量の次元が多いと難しくなる（**次元問題**）．また，共変量の分布が大きくずれると，その部分についてはマッチング不可能（**サポート問題**）．このような場合，共変量を一元化する方法が**傾向スコア・マッチング**である．傾向スコアは，マッチングだけでなく，層別解析でも回帰分析でも利用される．

例 1.6.11.

(1) 共変量ベクトルが一致することを要請することを，完全マッチングという．
(2) Mahalanobis 距離によるマッチング (Rubin, 1980)．これは質的変数が共変量となることに弱い．
(3) 最近傍マッチング，caliper マッチング．

1.6.6　層別解析

交絡因子だと思われる変数に依って同値類を作って，その中で解析する．

1.6.7　回帰分析

例 1.6.12 (共分散分析)**.** 原因変数に加えて共変量も同時にモデルに投入すること．交絡因子を追加して重回帰モデルを立て，各変量による結果 y への影響の大きさを「分散分析」する．共変量が連続であるとき，これを離散化せずに利用できる点で層別解析よ

り優れている．

例 1.6.13 (局所多項式回帰 Stone 1977：ノンパラメトリック)**.** バンド幅をどう指定するか，次元の呪いなどの問題点が残る．突破口があるなら CS であるが，社会科学分野で解析手法として用いられることはまだあまりない．

1.6.8　逆傾向スコア荷重による一致推定量

定理 1.6.14 (IPW (Rubin 85))**.** 傾向スコアの逆数で積分した値は，平均処置効果の不偏推定量となり，傾向スコアの一致推定に成功すれば IPW 推定量も一致推定量である．

潜在的な結果変数と説明変数との間はパラメトリックな仮定を置くが，共変量 x とその他の変数の間には回帰関係を仮定せずに解析するとなると，これはセミパラメトリックモデルとなる．

すると，回帰モデルのパラメトリックな部分に脆弱性が生じる．平均処置効果の推定では，傾向スコアの推定もパラメトリックモデルにするなら，脆弱性は 2 つである．それぞれについて頑健にしたいが，さらに高い頑健性として，「二重の頑健性」が考え得る．

議論 1.6.15. そもそも逆確率による重みづけは，モンテカルロ法の importance sampling でも出現する．少しずれた確率分布から得られたサンプルから，真の確率分布の下での何らかの確率変数の期待値を計算する一般的手法である．

1.7　共変量選択

モデルに追加した説明変数 s は，X の上流にあり，かつ，その全体 S によって X, Y を共通に説明する全ての流れが遮断されているとき，バックドア基準を満たすという．

定義 1.7.1 (バックドア基準)**.** 因果グラフ G において，有向道 $X \to Y$ があるとする．このとき，頂点集合 S が次を満たすとき，(X, Y) についてバックドア基準を満たすという：

(1) $\forall_{s \in S}\ X \not\to s$.
(2) X を始点とする有向道を G から除いたグラフにおいて，S は X と Y を有向分離する．

1.8　因果探索の基本問題

経済学や社会科学や疫学や生活など，controlled experiment が不可能な場面は多く，non-experimental なデータから因果効果を推定する必要は各学問で肝要である．

1.8.1　因果探索の基本問題

違う因果関係が同じ相関関係を定め得るが，観測変数の分布には相違が現れる．これを足掛かりにして因果探索が行われる．

模型 1.8.1 (因果探索の基本問題)**.** 次の 3 つの構造的因果モデル $M = ((x, y), (z, e_x, e_y), f, p)$ を考える．ただし，$x, y \in \mathbb{R}^n$.

$$\text{(A)}\ \begin{cases} x = f_x(z, e_x) \\ y = f_y(x, z, e_y) \\ p(z, e_x, e_y) = p(z)p(e_x)p(e_y) \end{cases} \qquad \text{すなわち，} x \to y.$$

(B)
$$\begin{cases} x = f_x(y, z, e_x) \\ y = f_y(z, e_y) \\ p(z, e_x, e_y) = p(z)p(e_x)p(e_y) \end{cases}$$
すなわち，$x \leftarrow y$.

(C)
$$\begin{cases} x = f_x(z, e_x) \\ y = f_y(z, e_y) \\ p(z, e_x, e_y) = p(z)p(e_x)p(e_y) \end{cases}$$
すなわち，$x \quad y$.

外生変数 z, e_x, e_y を独立とした．これは，z 以外に未観測共通要因が無いことと同値．また，自律性 (autonomy) を仮定する，すなわち，u 介入を行っても，f, p に影響しないことを仮定する．また，因果関係が一方向である（f_x が y を引いたり，再帰的な構造がない）ことを仮定した．これを acyclic という．

この時，データ行列
$$X = \begin{pmatrix} x^1 & \cdots & x^n \\ y^1 & \cdots & y^n \end{pmatrix}$$
が与えられた時，これを生成したモデル M を決定する問題を，因果探索の基本問題という．変数を x, y 以外に追加した場合も，この問題に還元される．

例 1.8.2.

(1) x はチョコレートの消費量，y は一国のノーベル賞受賞者数，z は GDP と見れる．
(2) x は薬を飲むかどうか，y は病気に罹患しているか，z は病気の重症度とみれる．

1.8.2　3 つの手法

f, p について種々の仮定を置くことが考えられるが，数学的には，因果グラフが識別可能な仮定のクラスが重要となる．そのためには，p の非 Gauss 性が肝要になることが [1] の発見である．

1.8.2.1　non-parametric approach

何の仮定も置かないと，因果グラフは識別可能でない．理論的な限界点を明らかにする理論的な価値がある．

1.8.2.2　parametric-approach

実質科学からの事前知識や洞察を反映して，f と p に仮定をおいて 3 つのモデルを比較する営みである．特に数理的には，推定に必要な観測数が小さくなるなど，解析が容易になるなどが起こる．特に，数理のモデル進化の定番として，最初は f の線形性と p の Gauss 性を仮定することが多い．が，この場合も，観測変数の分布がいずれも同様な Gauss 分布となるので，因果探索の基本問題について，因果グラフは識別可能でない．これは Gauss 分布が 2 次元多様体をなすことに起因する．

注 1.8.3. 因果グラフの推測には，例え非線形な系に対しても線形性を仮定した方がうまくいくという報告が多い．その後にノンパラメトリックな方法で因果効果の大きさを定量化するという流れが考えられる．

1.8.2.3　semi-parametric approach

p の非 Gauss 性に注目すると，未観測の交絡要因 z が存在しようと（存在さえ未知だろうと），因果推論が可能になる．We have recently described how *non-Gaussianity* in the data can be exploited for estimating causal effects. In this paper we show that, with non-Gaussian data, causal inference is possible even in the presence of hidden variables (unobserved confounders), even when the existence of such variables is unknown a priori. Thus, we provide a comprehensive and complete framework for the estimation of causal effects between the observed variables in the linear, non-Gaussian domain. [5]

定義 1.8.4 (LiNGAM: Linear Non-Gaussian Model)**.** f は線型関数，p は非 Gauss な連続分布と仮定する手法をいう．

定理 1.8.5. LiNGAM の仮定を置いた場合，３つのモデルの因果グラフは識別可能である．[5]

第 2 章

欠測データの扱い

偏りのあるデータからの統計的な因果推論を考える．全てがデータになる。全てがアルゴリズムになる。そして、全てが実験になる。世界全体がデータと実験の塊になった世界。

[8] の話題は 3 つ．

(1) 実験が行われていない研究で得られるデータからの統計的な因果推論．

(2) 偏った抽出による標本・データを用いることで生じるバイアス（これを**選択バイアス**と呼ぶ）の統計的調整．労働経済学者の Heckman が草分け．実は，人工知能分野での「領域適応 (domain adaptation)」や「共変量シフト」とも関連が深い．インターネット調査などでは，レスポンスをする人自体に偏りがあるので，マーケティング分野で積極的に利用されるようになってきた．

(3) 複数の情報源から得られたデータを統計的に融合させて行うデータ融合．特に大きなプラットフォームを持つ主体が，消費者のカテゴリーを超えた購買行動やクロスメディアコミュニケーションを分析するために興隆した．

背景情報＝共変量を積極的に利用することで，欠測データの部分を予測できる可能性がある．(2),(3) の話題は，統計的因果推論と同型の問題構造を有しているので，(1) に集中すれば良い．

現実の構造を見抜いて，それに対応する数理構造を創る．これは想像と創造である．

さらに，セミパラメトリックモデルで置かれた仮定になるべく依存せずに関心のある量の推測を精度良く行う，ロバスト解析が重要になる．

2.1 歴史

歴史 2.1.1. Little and Rubin (87, 2002) にまとまっており，現場の暗黙知だった状況から，彼らが初めて数学的に話題にした．

2.2 欠測の分類

欠測メカニズムのモデル $p(m|\phi)$ をモデルに盛り込む．

定義 2.2.1 (monotone missingness)**.** 欠測パターンについて，(1),(2),(3) などで，「ある変数が欠測であれば，別の変数でも必ず欠測がある」という関係がすべての対象者で成立するとき**単調欠測**という．

例 2.2.2.

(1) 各変数レベルでの無返答 (item nonresponse).

(2) 打ち切り (censoring) と切断 (truncation)．後者は閾値を超えた観測数そのものも不明．または部分的なデータで全体を代表させる測定 (surrogate measurements).

(3) 脱落 (dropout) やパネルの摩耗 (attrition)．または noncommpliance.

(4) 調査全体への不参加 (unit nonresponse).

(5) rounding.

(6) 連続データの離散化.

狭義の欠測は (3) まで．(4) は選択バイアスを産む．(5),(6) を含めて不完全データともいう．

定義 2.2.3 (Rubin (76))**.** 欠測のメカニズム $p(m|y,\phi)$ について，欠測するかどうかは，

(1) MCAR: Missing Completely at Random：モデリングに用いている変数に依らない．
(2) MAR: Missing at Random：欠測値には依存せず，観測できた値に依存する．
(3) NMAR: Nonmissing at Random：欠測値そのものと，観測していない他の値にも依存する．

の 3 種類がある．$p(m|y,\phi)$ とは，(2) の場合を考えている．

例 2.2.4. 入試成績と，入学後の成績とのデータの組など，選抜効果は MAR を産む．

2.3　枠組み

欠測データに対する最尤推定は，「完全尤度」なる量の最大化を考えることとなる．

2.3.1　選択モデル

定義 2.3.1 (selection model, complete data likelihood, distinctness, ignorable missingness)**.** $y=(y_1^\perp,y_2^\perp)^\perp$ を完全データのベクトルとし，$p(y|\theta)$ を完全データのモデル，$p(m|y,\phi)$ を欠損の有無を表す確率変数 $m:\Omega\to 2$ に関するモデルとする．

(1) 2 つの同時分布が，直積測度 $p(y,m|\theta,\phi)=p(y|\theta)p(m|y,\phi)$ によって表せるとき，これを**選択モデル**という．結合確率 $p(y,m|\theta,\phi)$ を，**完全データの尤度**ともいう．
(2) このように，母数 θ,ϕ について変数分離可能である性質のことを，**分離性**という．
(3) ランダムな欠損 (MAR または MCAR) であり，かつ，分離性を満たす欠損を**無視できる欠測**という．
(4) 欠測データについての平均 $p(y_1,m|\theta,\phi)=\int p(y|\theta)p(m|y,\phi)dy_2$ を，**完全尤度**という．[†1]

注 2.3.2 (無視できる欠測について)**.**

(1) MAR のとき，完全尤度は $p(y_1,m|\theta,\phi)=p(y_1|\theta)p(m|y_1,\phi)$ をなる．これを**観測データの尤度**という．そこでこの場合は，$p(y_1|\theta)$ についての最尤推定のみを考えれば良いので，「無視できる」という．
(2) MCAR のときも同様に，$p(y_1,m|\theta,\phi)=p(y_1|\theta)p(m|\phi)$ が成り立つ．

2.3.2　パターン混合モデル

定義 2.3.3 (pattern mixture model)**.** 同時分布が $p(y,m|\xi,\omega)=p(y|m,\xi)p(m|\omega)$.

2.3.3　共有パラメータモデル

最近利用される．

定義 2.3.4 (shared parameter model)**.** 潜在変数として変量効果 β を仮定し，この下で条件付き独立となる $p(y,m|\xi,\omega,\beta)=p(y|\beta,\xi)p(m|\beta,\omega)$ とき，**共有パラメータモデル**という．

[†1] ここでの「完全」は完全データというときの「完全」とは違い，観測されたデータを完全に利用しているという意味である．

第 3 章

汎関数推定

Causal inference is over.

サンプル数を n とすると，$\mathcal{P}$ 上の汎関数 $\psi(P) \in \mathbb{R}$ を推定する関数 $\widehat{\psi} : \mathcal{X}^n \to \mathbb{R}$ を作る．一般に巨大なモデル $\mathcal{P}$ の推定は難しいが，その上の汎関数ならば，随分と豊かな理論が揃ってきた．

3.1 有効理論導入

定義 3.1.1 (influence function for estimator)**.** 推定量 $\widehat{\psi}$ が影響関数 ϕ を持つとは，$\sqrt{n}$-一致性と漸近正規性 (CAN と略す)

$$\sqrt{n}(\widehat{\psi} - \psi) = \sqrt{n}\mathbb{P}_n\phi + o_P(1) \Rightarrow N(0, \mathrm{Var}[\phi])$$

が成り立つことをいう．

注 3.1.2 (plug-in method)**.** ノンパラメトリックモデルでは，$\sqrt{n}$-CAN は高々 1 つしか存在しない．そこで，最も自然なアプローチは，分布の第一段階推定 $\widehat{P}$ を用いて，plug-in $\psi(\widehat{P})$ とすることである．しかしこれは，一般には $\sqrt{n}$-一致性を持たず，また信頼区間を定めない．$\widehat{P}$ が特定の推定量で，強い正則条件を持つ場合に良い振る舞いをする．

例 3.1.3 (第一段階によるバイアスが残る)**.** plug-in により汎関数 $\psi = \int p(x)^2 dx = E[p(X)]$ を推定することを考える．推定量を $\widehat{p}$ として，$\widehat{\psi} := \mathbb{P}_n[\widehat{p}(X)]$ とする．次が成り立つ：

$$\widehat{\psi} - \psi = (\mathbb{P}_n - P)p + P(\widehat{p} - p) + o_P(n^{-1/2}).$$

初項は中心極限定理より $O_P(n^{-1/2})$ である．第二項は Cauchy-Schwartz によると $|P(\widehat{p} - p)| \leqslant \|\widehat{p} - p\|$．ある Holder β に対して，$O_P(n^{-\beta(2\beta+d)})$ となる．

$\widehat{p}$ が一致性を持つ限り，plug-in 推定量も一致性を持つが，$\widehat{p}$ 由来のバイアスが残っており，収束は遅い．

しかし，$\widehat{\psi}$ として kernel estimator

$$\widehat{p}(t) := \mathbb{P}_n\left(\frac{1}{h}K\left(\frac{X - t}{h}\right)\right)$$

を選び，K は higher-order kernel で，$h \sim n^{-1/(\beta+d)}$ で，かつ $\beta > d$ であるならば，第二項は

$$P(\widehat{p} - p) = (\mathbb{P}_n - P)p + o_P(n^{-1/2}) = O_P(n^{-1/2})$$

と評価できる．

問題は，

(1) $\beta > d$ の範囲で，これが最適であるか？
(2) $\beta \leqslant d$ の場合どうすればよいのか？
(3) 滑らかさの構造以外に依れないか？
(4) 別の generic nuisance estimator を使いたい場合は？

これらの疑問が，nonparametric efficient theory と影響関数に手招いている．これが準備すれば，kernel estimator with careful tuning を用いなくて済み，より underlying structure に無頓着になることが出来，double machine learning などの柔軟な手法が取れる．

3.2　セミパラメトリックモデルの有効性

理論的に肝要なのは，パラメトリックモデルに対して Cramer-Rao の不等式が与えるような下界を，セミパラメトリックモデルに対して拡張することである．

3.2.1　セミパラメトリックモデル

定義 3.2.1. 有限次元の実パラメータで添字付けられない集合 $\mathcal{P} \subset P(\mathbb{R})$ を，セミパラメトリックモデルという．

例 3.2.2 (semiparametric model)**.**

(1) ノンパラメトリックモデル．はじめのセミパラメトリックモデルは，Begun et al. (83) により"parametric-nonparametric models"と呼ばれた．
(2) GEE(Generalized Method of Moment)，m-推定量，制限モーメントモデル．これは部分的に回帰関数を仮定するもので，$Y = \mu(X;\psi) + \epsilon$ とし，$\psi \in \mathbb{R}^d$ を興味のある母数，ϵ は $E[\epsilon|X] = 0$ のみを仮定する局外母数とする．すなわち，$E[Y|X] = \mu(X;\psi)$ のみが仮定となる．
(3) Cox モデル：T を生存時間，$Z = (X, T)$ とし，条件付き hazard の比のみを $\dfrac{\lambda(t|X=x)}{\lambda(t|X=0)} = \exp(x^\top \psi)$ と仮定する．

また，非 Euclid 空間上の汎関数に関するパラメトリックな仮定をおいた場合も，自然にセミパラメトリックモデルを生じる．$\gamma(v) := E[Y^1 - Y^0|V = v]$ とし，共変量 V を高次元あるいは連続な成分を持つとすると，$\gamma(v;\psi)$ $(\psi \in \mathbb{R}^q)$ として，パラメトリックな部分を抽出することは自然である．

注 3.2.3. 因果推論問題においては，観察研究とは違って，実験計画（傾向スコアなど）が既知の場合が多い．一方で，反応は全く未知のことが多いので，パラメトリックな仮定をおくことを避けたいのである．一方で，実験もしていない場合は，ノンパラメトリックモデルのみが，合理的なモデルとなる．この場合でも，影響関数・経験過程論・サンプル分割が必要になるが，機械学習を用いた場合でも，$\sqrt{n}$-収束や，有効な信頼区間を得ることが出来る．

3.2.2　古典的下界の一般化

Cramer-Rao の結果は，古典的なパラメトリックモデルに対する下界を与える結果である．

議論 3.2.4. $(P_\theta)_{\theta\in\Theta}$ が滑らかで，T は $\psi(\theta)$ の不偏推定量であるとする．このとき，

$$\mathrm{Var}[T] \geqslant \frac{\psi'(\theta)^2}{E[s_\theta^2]}.$$

ただし，$s_{\theta*} := \dfrac{\partial}{\partial\theta} \log p_\theta(z)|_{\theta=\theta*}$ はスコアである．

This is also a lower bound in **asymptotic minimax sense**:

$$\forall_{\hat{\psi}:\mathcal{X}^n\to\mathbb{R}} \quad \inf_{\delta>0} \liminf_{n\to\infty} \sup_{\|\theta'-\theta\|<\delta} E_{\theta'}\left[l\left(\sqrt{n}\left(\hat{\psi} - \psi(\theta')\right)\right)\right] \geqslant E\left[l\left(N\left(0, \frac{\psi'(\theta)^2}{E[s_\theta^2]}\right)\right)\right]$$

定義 3.2.5 (parametric submodel)**.** ノンパラメトリックモデル $\mathcal{P}$ の**パラメトリックな部分モデル** $\mathcal{P}_\epsilon := \{P_\epsilon\}_{\epsilon\in\mathbb{R}} \subset \mathcal{P}$ とは，真値を含む：$P = P_0$ ものをいう．

]

例 3.2.6. $E[h] = 0, \|h\|_\infty < M, \epsilon < 1/M$ について，真値 p に対して $p_\epsilon(z) := p(z)(1 + \epsilon h(z))$ とすると，$p_\epsilon(z) \geqslant 0$.

3.2.3　pathwise differentiablity

この pathwise differentiability が大事で，これを用いて plug-in 推定量のバイアスを修正する．

問題 3.2.7. 部分モデル $\mathcal{P}_\epsilon$ に対する下界 $\dfrac{\psi'(P_\epsilon)^2}{E[s_\epsilon^2]}$ を考える．ただし，

$$\psi'(P_\epsilon) = \frac{\partial}{\partial \epsilon}\psi(P_\epsilon)|_{\epsilon=0}, \quad s_\epsilon = s_\epsilon(z) = \frac{\partial}{\partial \epsilon}\log p_\epsilon(z)|_{\epsilon=0}.$$

議論 3.2.8. 仮に，1 次の von Mises-type 展開

$$\psi(Q) - \psi(P) = \int \phi(Q)d(Q-P) + R_2(Q,P), \qquad P[\phi] = 0, P[\phi^2] < \infty$$

が可能だとすると，道ごとの微分可能性

$$\frac{\partial}{\partial \epsilon}\psi(P_\epsilon)|_{\epsilon=0} = \int \phi(z;P)s_\epsilon(z)dP(z)$$

が従う．この時点で $\mathcal{P}$ をなんかの多様体だと思ってるよなあ．

定義 3.2.9 (influence function, influence curve, gradient (for a functional))**.** 上述の von Mises 展開を許す関数 ψ を，推定したい汎関数 ψ に関する**影響関数**または**影響曲線**または**勾配**という．

3.2.4　有効影響関数

畳み込み定理の系として得られる $\mathrm{Var}[\phi]$ は，Cramer-Rao の下界の，ノンパラメトリックな対応物である．このときの ϕ を，efficient influence function という．asymptotic minimax の意味で，$\sqrt{n}(\widehat{\psi} - \psi) \Rightarrow N(0, \mathrm{Var}[\phi])$ に優る推定量はない．なおこの式は，推定量 $\widehat{\psi}$ の影響関数の定義式 3.1.1 と同じ．

問題 3.2.10. ノンパラメトリックモデル $\mathcal{P}$ の下界を，asymptotic minimax の意味で与えたい．すなわち，すべての部分モデルの下界の中で最大のものを与えたい．

注 3.2.11. なお，asymptotic minimax sense とは，統計的決定理論の用語である（駒木先生など）．ある特定の点において驚くほど精度がよい推定量などが構成可能であるが，これを病的として見向きしないことが出来る枠組みが決定理論である．これが，「最適」という一般的用語が，暗黙のうちには特定の意味 (minimax sense) を持つ背景である．

議論 3.2.12. ψ が道ごとに微分可能ならば，Cauchy-Schwarz の不等式より，

$$\sup_{\mathcal{P}_\epsilon} \frac{\psi'(P_\epsilon)^2}{E[s_\epsilon^2]} = \sup_h \frac{P[\phi h]^2}{P[h^2]} \leqslant P(\phi^2) = \mathrm{Var}[\phi]$$

等号成立は，ϕ が有効な部分モデルのスコア (valid submodel score) である場合．ただし，ϕ が valid score であるとは，ϕ が接空間内にあることをいう．接空間とは，score space の閉包をいう．

定義 3.2.13 (efficient influence function)**.** 平均 0 で有限な分散を持つ $\phi : \mathcal{X}^n \to \mathbb{R}$ であって，任意の部分モデル $\{P_\epsilon\} \subset \mathcal{P}$ に対して

$$\frac{\partial}{\partial \epsilon}\psi(P_\epsilon)|_{\epsilon=0} = \int \phi(z;P)s_\epsilon(z)dP(z)$$

を満たすものを，**有効影響関数**という．

定理 3.2.14 (uniqueness of EIF)**.** $\mathcal{T}$ を接空間，ϕ を影響関数，Π を射影演算子とする．このとき，$\Pi(\phi|\mathcal{T})$ がただ一つの有効影響関数である．

3.2.5　影響関数の例

影響関数 ϕ の形がわかっているとき，弱い仮定を置くことで有効な推定量を構成できる．

議論 3.2.15 (影響関数の求め方)**.**

(1) pathwise derivative $\psi'(P_\epsilon)$ を計算し，ϕ について解く．
(2) データは離散だとして，point mass contamination の方向に対する Gateaux 微分を計算するほうがしばしば簡単になる（吉田先生の本にあった方法）．
(3) 単純な影響関数を既知として，連鎖律や Leibniz 則によって構成する．

例 3.2.16 (平均)**.** $\psi = E[Z] = \int z dP(z)$ の影響関数は，

$$\begin{aligned}\psi'(P_\epsilon) &= \frac{\partial}{\partial \epsilon}\int z dP_\epsilon(z)|_{\epsilon=0} = \int z \frac{\partial}{\partial \epsilon} dP_\epsilon(z)|_{\epsilon=0} \\ &= \int z \left(\frac{\partial}{\partial \epsilon}\log dP_\epsilon(z)\right) dP_\epsilon(z)|_{\epsilon=0} \\ &= \int (z-\psi)\left(\frac{\partial}{\partial \epsilon}\log dP_\epsilon(z)\right)|_{\epsilon=0} dP(z)\end{aligned}$$

より，$\psi(z;P) = z - \psi$ が有効影響関数で，von Mises 展開は $R_2 = 0$ について成り立つ．また，最適な推定量は $\widehat{\psi} = \mathbb{P}_n[Z]$ で，

$$\mathrm{Var}[\sqrt{n}(\widehat{\psi} - \psi)] = \mathrm{Var}[Z] = \mathrm{Var}[\phi].$$

3.3　ノンパラメトリック推定

続いて上界を考えたい．つまり，普通の機械学習によっても P が推定できるような推定量の構成法を作りたい．

記法 3.3.1. $\mathbb{P}[f] = \mathbb{P}[f(Z)] = \int f(z) d\mathbb{P}(z)$ とする．

3.3.1　導入

問題 3.3.2. von Mises 展開可能な ψ

$$\psi(\mathbb{Q}) - \psi(\mathbb{P}) = \int \phi(\mathbb{Q}) d(\mathbb{Q} - \mathbb{P}) + R_2(\mathbb{Q}, \mathbb{P})$$

を考える．このとき，plug-in 推定量は，$\widehat{\mathbb{P}}[\phi(\mathbb{P})] = 0$ より，

$$\psi(\widehat{\mathbb{P}}) - \psi(\mathbb{P}) = -\int \phi(\widehat{\mathbb{P}}) d\mathbb{P} + R_2(\widehat{\mathbb{P}}, \mathbb{P})$$

となり，1 次のバイアスを持つ．まずは，これを推定することを考えたい．

議論 3.3.3. $\widehat{\psi} := \psi(\widehat{\mathbb{P}}) + \mathbb{P}_n[\phi(\widehat{\mathbb{P}})]$ をワンステップ推定量とすると，次のように展開できる．

$$\begin{aligned}\widehat{\psi} - \psi &:= \left[\psi(\widehat{\mathbb{P}}) + \mathbb{P}_n[\phi(\widehat{\mathbb{P}})]\right] - \psi = (\mathbb{P}_n - \mathbb{P})\phi(\widehat{\mathbb{P}}) + R_2(\widehat{\mathbb{P}}, \mathbb{P}) \\ &= (\mathbb{P}_n - \mathbb{P})(\phi(\widehat{\mathbb{P}}) - \phi(\mathbb{P})) + (\mathbb{P}_n - \mathbb{P})\phi(\mathbb{P}) + R_2(\widehat{\mathbb{P}}, \mathbb{P}).\end{aligned}$$

(1) 分散が縮んでいくような項の標本平均．
(2) 固定された関数に対する標本平均 $\mathbb{P}_n[\phi(\mathbb{P})]$ に他ならない．これは CLT により収束する．
(3) NP 条件を与えられたとき，無視できる．

(1),(3) を処理できれば，最適な推定量が得られる．これらは，古典的にいう推定方程式とワンステップ推定 (one-step correction) に対応する．

3.3.2　経験過程と標本分割

標本分割法により，経験過程論の精緻な議論はある意味廃された．なんというか，精密なリサンプリングアルゴリズムで洗練されていくのは，すごくきれいな計算機科学的な仕事に思える．

問題 3.3.4. $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ としたとき，

(1) $R_1 := \mathbb{G}_n(\phi(\widehat{\mathbb{P}}) - \phi(\mathbb{P}))$.
(2) $R_2 := \sqrt{n}R_2(\widehat{\mathbb{P}}, \mathbb{P})$.

が 0 に $\mathbb{P}$-確率収束することを示したい．これが示せれば，$\sqrt{n}(\widehat{\psi} - \psi) \Rightarrow N(0, \mathrm{Var}[\phi])$ を得る．これは最適性を意味する．

議論 3.3.5. 結論から言うと，R_1 については，次のいずれかが成り立つ時，解決できる．

(1) $\phi(\mathbb{P})$ が複雑すぎない時（経験過程論による）
(2) $\widehat{\mathbb{P}}, \mathbb{P}_n$ を分割して overfittin を防ぐ（標本分割法による）

R_2 については，基本的にはあまり一般的な結論は引き出せず，ケースバイケースで対処する必要がある．しかし，今までの bias-corrected estimator でない方法とは違って，複雑なノンパラメトリックモデルにおいても，R_2 は結局無視できるほどに収束が早い．一般の plug-in などでは，この項は収束するのが非常に遅い．

補題 3.3.6 (19.24 van der Vaart (2000))**.** 次を仮定する．

(A1) $\widehat{f}, f \in \mathcal{F}$ を Donkser クラス $\mathcal{F}$ の元とする．
(A2) $\|\widehat{f} - f\|^2 = o_{\mathbb{P}}(1)$.

このとき，

$$\mathbb{G}_n\widehat{f} = \mathbb{G}_n f + o_{\mathbb{P}}(1)$$

補題 3.3.7 (sample splitting lemma (Kennedy et al.))**.** [†1] $\widehat{f}$ を標本 $Z^N = (Z_{n+1}, \cdots, Z_N)$ から定めた推定量とし，$\mathbb{P}_n$ で Z^N と独立に観測した標本 $(Z_1, \cdots, Z_n)$ が定める経験過程とする．このとき，$\sqrt{n}(\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) = O_P(\|\widehat{f} - f\|)$.

要諦 3.3.8. これにより，L_2-一致性を持ち，標本分割数 K が有限である限り，$R_1 = o_P(1)$ を得るし，モデルの複雑性に制限がないから，機械学習法も使える．

3.3.3　第二の剰余項

例 3.3.9. $\psi = E[Z]$ のとき，$R_2 = 0$．これは，$f = \dfrac{A}{\pi}(Y - g) + g$ などの，あらゆる $f(Z)$ に対して成り立つ．すなわち，IPW-スタイルの影響関数は，π が既知の場合の影響関数に他ならない．

例 3.3.10. $\psi = \displaystyle\int p(z)^2 dz$ の影響関数は $\phi(z|p) = 2(p(z) - \psi)$ である．一般の plug-in $\mathbb{P}_n(\widehat{p})$ は $\sqrt{n}$-一致性を持たない．が，$\widehat{p}$ を標本分割によって得て，IF-based bias-corrected 推定量

$$\begin{aligned}\widehat{\psi} &= \psi(\widehat{\mathbb{P}}) + \mathbb{P}_n[\phi(\widehat{\mathbb{P}})] \\ &= \int \widehat{p}^2 + \mathbb{P}_n(2\widehat{p} - \int \widehat{p}^2) = 2\mathbb{P}_n(\widehat{p}) - \int \widehat{p}^2\end{aligned}$$

を考えると，

$$\widehat{\psi} - \psi = 2(\mathbb{P}_n - \mathbb{P})(\widehat{p} - p) + 2(\mathbb{P}_n - \mathbb{P})p + R_2(\widehat{p}, p), \quad R_2(\widehat{p}, p) = -\int (\widehat{p} - p)^2.$$

あとは density estimation の問題になるが，従来の半分の smoothness しか要求しない．

[†1] "Sharp instruments for classifying compilers and generalizing causal effects"

例 3.3.11. $\psi = E[E[Y|X, A = 1]] =: E[\mu(X)]$ の影響関数は

$$\phi(Z; P) = \frac{A}{\pi(X)} (Y - \mu(X)) + \mu(X) - \psi.$$

3.4　open problems

(1) 新たな汎関数についてはどうする？
(2) $\sqrt{n}$-収束が達成不可能な場合はどうする？
(3) ψ が道ごとに微分可能でない場合はどうする？

第 4 章

適応的実験

適応的実験とバンディット問題と，因果推論の関係を考える．

4.1 離散時間マルチンゲール

定義 4.1.1 (filtration)**.**

⑴ 確率空間 $(\Omega, \mathcal{F}, P)$ 上の，$\mathcal{F}$ の部分 σ-加法族の増大列 $(\mathcal{F}_k)_{k\in\mathbb{N}}$ を**離散時間フィルトレーション**という．
⑵ 4つ組 $(\Omega, \mathcal{F}, (\mathcal{F}_k), P)$ を**離散時間確率基**という．
⑶ $(\xi_k)_{k=1,2,\cdots}$ が**離散時間マルチンゲール差分列**とは，$\forall_{k\in\mathbb{N}}\ \xi_k$ は $\mathcal{F}_k$ 可測・可積分で，$E[\xi_k|\mathcal{F}_{k-1}] = 0$ a.s. を満たすことをいう．
⑷ $(X_k)_{k\in\mathbb{N}}$ が**離散時間マルチンゲール列**とは，$\forall_{k\in\mathbb{N}}\ \xi_k$ は $\mathcal{F}_k$ 可測・可積分で，$E[\xi_k|\mathcal{F}_{k-1}] = X_{k-1}$ a.s. を満たすことをいう．

補題 4.1.2. 離散時間マルチンゲール差分列 $(\xi_k)_{k=1,2,\cdots}$ と，$\mathcal{F}_0$-可測な可積分確率変数 X_0 と，$\mathcal{F}_{k-1}$-可測確率変数 H_{k-1} であって $E[|H_{k-1}\xi_k|] < \infty$ であるようなものが与えられた場合，

$$X_k = X_0 + \sum_{j=1}^{k} H_{j-1}\xi_j$$

は離散時間マルチンゲール列になる．

4.2 停止時間と任意抽出定理

定義 4.2.1.

⑴ T が**停止時刻**であるとは，確率変数 $T : \Omega \to \mathbb{N} \cup \{\infty\}$ であって，$\forall_{k\in\mathbb{N}}\ \{\omega \in \Omega \mid T(\omega) = k\} \in \mathcal{F}_k$ を満たすことをいう．
⑵ 停止時刻が**有限**であるとは，$\forall_{\omega\in\Omega}\ T(\omega) < \infty$ を満たすことをいう．
⑶ 停止時刻が**有界**であるとは，$\exists_{c\in\mathbb{R}}\ \forall_{\omega\in\Omega}\ T(\omega) \leqslant c$ を満たすことをいう．

4.3 離散時間マルチンゲールに関する中心極限定理

定理 4.3.1. 離散時間確率基 $(\Omega^n, \mathcal{F}^n, (\mathcal{F}_k^n), P^n)$ 上の p-次元離散時間マルチンゲール差分列 $(\xi_k^n) = ((\xi_k^{n,1}, \cdots, \xi_k^{n,p})^T)$ と有限停止時刻の列 (T_n) を考える．

⑴ $\forall_{i,j\in[p]}\ \sum_{k=1}^{T_n} E^n[\xi_k^{n,i}\xi_k^{n,j}|\mathcal{F}_{k-1}^n] \xrightarrow{p} C^{(i,j)} \in \mathbb{R}$.
⑵ $\forall_{\epsilon>0}\ \sum_{k=1}^{T_n} E^n[\|\xi_k^n\|^2 1_{\{\|\xi_k^n\|>\epsilon\}}|\mathcal{F}_{k-1}^n] \xrightarrow{p} 0$.

が成り立つならば,

$$\sum_{k=1}^{T_n} \xi_k^n \xrightarrow{d} \mathbb{N}_p(0,\Sigma) \quad \left(\Sigma = (C^{(i,j)})_{i,j\in[p]}\right)$$

が成り立つ.

補題 4.3.2 (Lindeberg 条件の十分条件 (Lyapunov)). $\exists_{\delta>0} \sum_{k=1}^{k_n} E^n[\|\xi_k^n\|^{2+\delta}|\mathcal{F}_{k-1}^n] \xrightarrow{d} 0$ がなりたてば，(1),(2) が従う.

4.4　multiarm bandit について

arm に resorce を最適に allocate する問題をいう．最適制御／確率的スケジューリングの一形式でもあり，特に強化学習の一形式でもある（情報が増えていくので）．スロットマシンのことを one-armed bandit という，洒落た表現だ．ここから，ギャンブラーの気持ちになればよい．眼の前の arm の"exploitation"を選ぶか，"exploration"に時間を使うかの選択（これを policy という）を迫られるのである．製薬会社の経営方針などをモデルするのにも使われる．Robins が 1952 年にこの問題の重要性に注目した.

(1) 製薬会社の経営方針.
(2) 人道的な治験.
(3) インターネット広告配信，推薦システム.
(4) ゲーム木探索.
(5) 通信のルーティング.

4.5　傾向スコア

arm w を時刻 t に下げる確率 $e_t(w), e : T \times \mathcal{W} \to [0,1]$ を，傾向スコアという.

たとえば、喫煙の影響を知りたい場合を考える。人々を喫煙群に無作為に割り付けることは非倫理的であるため、観察研究が必要である。喫煙群と非喫煙群とを単純に比較することによって処置効果を推定すると、喫煙率に影響する要因（性別や年齢など）によるバイアスが生じる。PSM では、処置群とコントロール群の制御変数（この例では性別や年齢など）を同じくらいにすることによって、これらのバイアスを制御することを目指す。[†1]

定義 4.5.1 (propensity score).

(1) $Z_i \in 2$ は被験者 i が処置群に割り付けられたか，コントロール群に割り付けられたかを表す.
(2) バックグラウンド変数 X_i は被験者 i への割り当て前に観測された種々のデータとする.
(3) バックグラウンドで観測された共変量 X に対する，処置の条件付き確率を

$$e(x) := P[Z = 1|X = x]$$

と定め，これを**傾向スコア**という.

4.6　notation

Rosenbaum と Rubin を超えていく

adaptivity に由来するバイアスによって，不偏推定量がぐるぐる変わってしまう．片方の arm のみ sampled more なので，バイアスがかかる．これの代表的な解決が propensity score matching (83) であり，見事にバイアスは消えるが，今度は漸近分布の正規性が消える．したがって CLT が出来なかった．特に或るアームの引かれる確率 (assignment probability) が低い場合は裾の重い極限分布をもち，統計的推論が困難になる．あるいは実際の実験は付値確率が極限分布が収束しない

[†1] https://ja.wikipedia.org/wiki/傾向スコア・マッチング

状態で終わる．平均を整えても分散が爆発する．h_t が特定の条件を満たすように実験を設計することで，バイアスは掛かるが平均が良い推定量を得て，それは利用可能性の高いデータとなる．

定義 4.6.1. $(\Omega, \mathcal{F}, P)$ を確率空間とする．$\mathcal{W}, \mathcal{Y}$ を距離空間とする．

(1) $W_t : H_t \to \mathcal{W}$ は arm number，すなわち実現された処置 (realized treatment)．歴史に依存する先天的に定義された確率分布 (bandit algorithm) に従う．
(2) $Y_t : \mathcal{W} \to \mathcal{Y}; W_t \mapsto Y_t(W_t)$ は観測された結果．2 値関数に落とし込まれた場合は潜在結果 (potential outcome) で，t 番目の患者が w に assign された場合 1 となる．arm なら reward.
(3) $m : \mathcal{W} \to \mathbb{R}; \omega \to E[Y_t(\omega)]$ は Y_t に関する平均潜在結果 (mean potential outcome) を表す有界な可測関数で，$\hat{m}_t : \mathcal{W} \times H_{t-1} \to \mathbb{R}$ はその推定量．これは一致性を持たなくても良い．論文 [6] では Q という文字が用いられている．
(4) $\Delta(w, w') := E[Y_t(w)] - E[Y_t(w')]$ とする．
(5) $H_t := \{(Y_s, W_s) \in \mathcal{Y} \times \mathcal{W} \mid s \in t+1\} = \{(Y_s, W_s)\}_{s=0,1,\cdots,t}$ は歴史．$\mathbf{H} := P(\mathcal{Y} \times \mathcal{W})$ とする．[†2]
(6) $\mathcal{H} = (\mathcal{H}_t)_{t \in T+1}$ は $\mathcal{H}_t = \sigma[H_t]$ とするフィルトレーションである．
(7) $e_t(w) := P[W_t = w | H^{t-1}]$ は assignment probability，傾向スコアという．[†3]
(8) $P_{t-1} : H_t \times \mathcal{B}_{\mathcal{W}} \to \mathbb{R}$ は正則条件付き確率とする．
(9) 条件付き期待値 $E_{t-1} : \mathcal{Y} \to \mathbb{R}$ を，$E[f(W_t)]$ を $E[f]$ と略記する．
(10) $\psi_{h_{t-1}} : L^2(\mathcal{W}, \mathcal{B}_{\mathcal{W}}, P_{t-1}(h_{t-1}, \cdot)) \to \mathbb{R}$ は，H^{t-1}-条件付き自乗積分可能な関数（平均潜在結果）の空間上の有界線形汎関数であり，合成関数 $\psi(m)$ を推定することを考える．arm 毎の平均 $\psi := \mathrm{ev}_w$ など，推定したい統計量を表す．
(11) h_t は evaluation weight で，傾向スコアを打ち消すことで分散を収束させることを考える．

定理 4.6.2. ψ の一意的な Riez-representor $\gamma(-; H^{t-1}) \in L_2(P_{t-1})$

$$\forall_{f \in L_2(P_{t-1})}\ E[\gamma(W_t; H^{t-1}) f(W_t) | H^{t-1}] = \psi(f)$$

が存在する．

要諦 4.6.3. $\gamma(W_t)$ とは，AIPW の $\dfrac{1_{\{W_t = w\}}}{e_t(w)}$ に対応する．

例 4.6.4. arm 毎の平均 $\psi := \mathrm{ev}_w$ の Riez representer は $\gamma_t = \dfrac{1_{\{-=w\}}}{e_t(w)}$ である．

記法 4.6.5.

(1) $f(\omega) = E_{t-1}[\gamma_t(w_t) f(w_t)]$ を Riez-representer とする．
(2) 下付き文字 m_t, γ_t は条件付き $m(-|H^{t-1}), \gamma(-|H^{t-1})$ の略記である．同様に，$E[X|H^{t-1}] = E_{t-1}[X]$ と表す．

4.7　Malliavin calculus

Z のように，漸近分布から未知量を消すために標準偏差の推定量 $\hat{\sigma}$ で割って規格化することを，Studentization または self-normalized estimator という．t-分布を発見した William Gosset による．

[†2] [?] 第一稿では，歴史 H_t は $2t$-組として表現されている．論文 [6] では H^t としている．
[†3] time-varying and decided via some known algorithm, as it is the case with many popular bandit algorithms such as Thompson sampling[6]

4.8　adaptive weight の構成

適応的実験とは強化学習と実験計画の融合である．bandit algorithm で最適化される．bandit とは，one-armed bandit という別称を持つスロットの攻略法（どの台に賭けるか）の問題として 1950s に始まったため．スロット台のことを arm と呼ぶのか．そうして得た結果の最大活用を考える．
assignment probability が収束しない場合は，そのデータからの推論を困難にする．ここで，assignment probability が収束し，その極限分布に 3 つの仮定を課すと，頻度主義的な信頼区間が計算できることを提案 [6].

公理 4.8.1 (adaptive weight に関する公理)**.** evaluation weights h_t は

(1) (Infinite Sampling) $\dfrac{\left(\sum_{t=1}^T h_t\right)^2}{E\left[\sum_{t=1}^T h_t^2\gamma_t^2\right]} \xrightarrow[T\to\infty]{p} \infty$. [†4]

(2) (Variance Convergence) $\exists_{p>1}\ \dfrac{\sum_{t=1}^T h_t^2 E_{t-1}[\gamma_t^2]}{E\left[\sum_{t=1}^T h_t^2\gamma_t^2\right]} \xrightarrow[T\to\infty]{L_p} 1$. [†5]

(3) (Bounded Moments / Lyapunov condition) $\exists_{\delta>0}\ \dfrac{\sum_{t=1}^T h_t^{2+\delta} E_{t-1}[|\gamma_t|^{2+\delta}]}{E\left[\sum_{t=1}^T h_t^2\gamma_t^2\right]^{1+\delta/2}} \xrightarrow[T\to\infty]{p} 0$. [†6]

を満たす．ただし，γ_t は $\dfrac{1}{e_t}$ などの，荷重によって飼い慣らしたい量となる．

4.8.1　scoring rule

定義 4.8.2. $\hat{\Gamma}$ が $Q(w)$ に対する unbiased scoring rule であるとは，$\forall_{w\in\mathcal{W}}\ \forall_{t\in[T]}\ E\left[\tilde{\Gamma}_t(w)|H^{t-1}\right] = Q(w)$ が成り立つことをいう．

例 4.8.3.

(1) inverse propensity score weighted $\hat{\Gamma}_t^{IPW}(w) := \dfrac{1_{\{W_t=w\}}}{e_t(w)} Y_t$.

(2) augmented inverse propensity weighted は regression adjustment を加える (Robins 94).

$$\hat{\Gamma}_t^{AIPW}(w) := \frac{1_{\{W_t=w\}}}{e_t(w)} Y_t + \left(1 - \frac{1_{\{W_t=w\}}}{e_t(w)}\right) \hat{m}_t(w)$$

命題 4.8.4. 不偏スコア規則 $\hat{\Gamma}$ が定める量 $\hat{Q}_t(w) := \dfrac{1}{T}\sum_{t=1}^T \hat{\Gamma}_t(w)$ は，Q について不偏である：$E[\hat{Q}_T(w)] = Q(w)$．特に，$\hat{\Gamma}_t$ が $t \in T$ に相関する場合も成り立つ．

[証明]． 繰り返し期待値の法則による：

$$\begin{aligned} E\left[\hat{Q}_T(w)\right] &= E\left[\frac{1}{T}\sum_{t=1}^T \hat{\Gamma}_t(w)\right] \\ &= \frac{1}{T}\sum_{t=1}^T E\left[E\left[\hat{\Gamma}_t(w)|\mathcal{H}^t\right]\right] \\ &= \frac{1}{T}\sum_{t=1}^T E[Q(w)] = Q(w). \end{aligned}$$

■

[†4] Bandid algorithm はどの arm も無限回 assign する．
[†5] 条件付き期待値が，条件付きでない期待値に一致する．
[†6] $e_t = 1/\gamma_t$ の decay がいくら速くても，h_t も分散を収束させるくらいには十分速い．

4.8.2　漸近的正規な test statistics

Qualitatively, what we need for normality is for the variability of the estimator to be deterministic. だから,「逆数にしても発散しない，修正された傾向スコア」のようなものが必要になるのだ．それは，単に unbiased scoring rule を平均して推定量とするのではなく，evaluation weight $(h_t)_{t\in T+1}$ で荷重する．この (h_t) をうまく選ぶことで，assignment probability $e_t(w)$ を打ち消して挙動を漸近的正規にする．

> With such weights, the adaptively-weighted AIPW estimator (6), when normalized by an estimate of its standard deviation, has a centered and normal asymptotic distribution. Similar “self-normalization” schemes are often key to martingale central limit theorems (see e.g., de la Pen ☒ a et al., 2008).[6]

定義 4.8.5 (adaptively-weighted AIPW estimator)**.**

$$\widehat{Q}_T^h(w) := \frac{\sum_{t=1}^T h_t(w)\widehat{\Gamma}_t^{AIPW}(w)}{\sum_{t=1}^T h_t(w)}.$$

補題 4.8.6. (h_t) が $\sum_{t=0}^{T} h_t = 1$ を満たすならば，これが定める adaptively-weighted AIPW 推定量 $Q := \sum_{t=0}^{T} h_t(w)\widehat{\Gamma}(w)$ は不偏である：$E\left[h_t(w)\widehat{\Gamma}(w)|H^{t-1}\right] = h_t(w)Q(w)$.

4.9　General settings

傾向スコアによる AIPW を，ψ の Riez representer と捉える枠組みは誰が気づいたのか.

定義 4.9.1. $(\Omega, \mathcal{F}, P)$ を確率空間とする．$\mathcal{W}, \mathcal{Y}$ を距離空間とする．$\mathcal{W}$ は可算とする．$T \in \mathbb{N}$ とする.

(1) $H_t = ((Y_s, W_s); s \in [t])$ を歴史とする．これは $\mathbf{H}_t := (\mathcal{Y} \times \mathcal{W})^t$ の元である．各歴史が定める σ-加法族を $\mathcal{H}_t := \sigma[H_t]$ とし，$\mathcal{H} := (\mathcal{H}_t)_{t\in[T]}$ を filtration とする.

(2) この上の正則条件付き確率を $P_{t-1} : H_t \times \mathcal{B}_{\mathcal{W}} \to \mathbb{R}$ と定める.

(3) $W_t : H_t \to \mathcal{W}$ は arm number，すなわち実現された処置 (realized treatment) で，歴史に依存する先天的に定義された確率分布 (bandit algorithm) に従う．その確率 $e_t(w) := P[W_t = w|H^{t-1}]$ は assignment probability，または傾向スコアという.

(4) $Y_t : \mathcal{W} \to \mathcal{Y}; W_t \mapsto Y_t(W_t)$ は観測された結果．2 値関数に落とし込まれた場合は潜在結果 (potential outcome) で，t 番目の患者が w に assign された場合 1 となる．arm なら reward.

(5) $m : \mathcal{W} \to \mathbb{R}; \omega \to E[Y_t(\omega)]$ は Y_t に関する平均潜在結果 (mean potential outcome) を表す有界な可測関数で，$\hat{m}_t : \mathcal{W} \times H_{t-1} \to \mathbb{R}$ はその推定量．これは一致性を持たなくても良い．また，因果効果を $\Delta(w, w') := E[Y_t(w)] - E[Y_t(w')]$ とする.

(6) $\psi_{h_{t-1}} : L^2(\mathcal{W}, \mathcal{B}_{\mathcal{W}}, P_{t-1}(h_{t-1}, \cdot)) \to \mathbb{R}$ は，H^{t-1}-条件付き自乗積分可能な関数（平均潜在結果）の空間上の有界線形汎関数であり，合成関数 $\psi(m)$ を推定することを考える．arm 毎の平均 $\psi := \mathrm{ev}_w$ など，推定したい統計量を表す．論文 [6] では Q という文字が用いられている.

(7) Riesz representer $\gamma_{h_{t-1}} \in L^2(\mathcal{W}, \mathcal{B}_{\mathcal{W}}, P_{t-1}(h_{t-1}, -))$ を,

$$\forall_{f\in L^2(\mathcal{W},\mathcal{B}_{\mathcal{W}},P_{t-1}(h_{t-1},-))}\ \psi_{h_{t-1}}(f) = \int_{\mathcal{W}} \gamma_{h_{t-1}}(w)f(w)P_{t-1}(h_{t-1}, dw) = E[\gamma_{h_{t-1}}(W_t)f(W_t)|H_{t-1} = h_{t-1}]$$

を満たすものと定める.

(8) $(h_t)_{t\in[T]}$ は evaluation weight と呼ばれる実数列で，傾向スコアを打ち消すことで分散を収束させることを考える.

公理 4.9.2 (adaptive weight に関する公理)**.** evaluation weights h_t は

(a) (Infinite Sampling) $\dfrac{\left(\sum_{t=1}^T h_t\right)^2}{E\left[\sum_{t=1}^T h_t^2\gamma_t^2\right]} \xrightarrow[T\to\infty]{p} \infty$. [†7]

(b) (Variance Convergence) $\exists_{p>1}\ \dfrac{\sum_{t=1}^T h_t^2 E_{t-1}[\gamma_t^2]}{E\left[\sum_{t=1}^T h_t^2\gamma_t^2\right]} \xrightarrow[T\to\infty]{L_p} 1$. [†8]

(c) (Bounded Moments / Lyapunov condition) $\exists_{\delta>0}\ \dfrac{\sum_{t=1}^T h_t^{2+\delta} E_{t-1}[|\gamma_t|^{2+\delta}]}{E\left[\sum_{t=1}^T h_t^2\gamma_t^2\right]^{1+\delta/2}} \xrightarrow[T\to\infty]{p} 0$. [†9]

を満たす．ただし，γ_t は $\dfrac{1}{e_t}$ などの，荷重によって飼い慣らしたい量となる．

記法 4.9.3.

(1) 条件付き期待値 $E_{t-1}:\mathcal{Y}\to\mathbb{R}$ について，$E_{t-1}[f(W_t)]$ を $E_{t-1}[f]$ と略記する．
(2) $\gamma(W_t;H^{t-1})=\gamma_{H^{t-1}}(W_t)$ を $\gamma_t(W_t)$ と略記する．E_t, m_t なども同様．

定義 4.9.4 (一般化された不偏スコア規則)**.** 一般化された不偏スコア規則 $\widehat{\Gamma}_t:\mathcal{W}\to\mathbb{R}$ を，次のように定める：

$$\widehat{\Gamma}_t(W_t) := \psi(\widehat{m}) + \gamma(W_t;H^{t-1})(Y_t - \widehat{m}(W_t;H^{t-1})).$$

補題 4.9.5 (不偏スコアが不偏推定量となっている)**.**

(1) $E[\gamma_{H_{t-1}}(W_t)(Y_t - m(W_t))|\mathcal{H}_{t-1}] = 0$.
(2) $\psi_{H_{t-1}}(\widehat{m}_t(\cdot,H_{t-1})) - \psi_{H_{t-1}}(m) - \gamma_{H_{t-1}}(W_t)(\widehat{m}_t(W_t,H_{t-1}) - m(W_t))$ も martingale 差分列である．
(3) 2 つの和を $\dot{\xi}_t^T$ とおくと，$\dot{\xi}_t = \widehat{\Gamma}_t - \psi_{H_{t-1}}(m)$ と表せて，これも martingale 差分列である．

4.10　適応的実験のための荷重した統計量の中心極限定理

不偏スコア $\widehat{\gamma}_t$ を荷重 (h_t) で調整した統計量 $\widehat{\psi_T}$ についての中心的な極限定理を示す．

定理 4.10.1 (中心的極限定理)**.** 次を仮定する：

(1) 見本過程 $Y_t(w)$ の分散は上に有界で，0 でない (away from zero)．
(2) $\delta>0$ が存在して平均 $E[|Y_t(w)|^{2+\delta}]$ が $w\in\mathcal{W}$ について一様に上に有界．
(3) Riesz 表現子 γ_t は $\exists_{b>0}\ \forall_{t\in[T]}\ E_{t-1}[\gamma_t^2] > b$ を満たす．
(4) $\widehat{m}_t$ は $\exists_{m_\infty\in L^2(\mathcal{W},\mathcal{B}_\mathcal{W},P_{t-1}(h_{t-1},-))}\ \|\widehat{m}_t - m_\infty\|_{L_\infty(P_{t-1})} \xrightarrow[t\to\infty]{\text{a.s.}} 0$ を満たす一様有界な推定量の列．

荷重 $(h_t)_{t\in[T]}$ が次のいずれか：

(a) $\widehat{m}_t$ の一致性：$\|\widehat{m} - m\|_{L_\infty(P_{t-1})} \xrightarrow[t\to\infty]{\text{a.s.}} 0$.
(b) $E_{t-1}[\gamma_t^2] \xrightarrow[t\to\infty]{\text{a.s.}} \overline{\gamma}_\infty^2 \in (0,\infty]$

を満たすならば，

(1) (一致性) 推定量

$$\widehat{\psi}_T := \frac{\sum_{t=1}^T h_t\widehat{\Gamma}_t}{\sum_{t=1}^T h_t}$$

は $\psi(m)$ に確率収束し，

[†7] Bandid algorithm はどの arm も無限回 assign する．
[†8] 条件付き期待値が，条件付きでない期待値に一致する．
[†9] $e_t = 1/\gamma_t$ の decay がいくら速くても，h_t も分散を収束させるくらいには十分速い．

(2) (漸近的正規性) student 化した統計量は漸近的に正規である：

$$\frac{\widehat{\psi_T} - \psi(m)}{\widehat{V}_T^{1/2}} \xrightarrow[T\to\infty]{d} \mathcal{N}(0,1) \quad \widehat{V}_T := \frac{\sum_{t=1}^T h_t^2(\widehat{\Gamma}_t - \widehat{\psi}_T)^2}{\left(\sum_{t=1}^T h_t\right)^2}.$$

要諦 4.10.2. 分散は，「$\widehat{\psi}_T$ の 2 乗の，2 乗加重平均」と推定する．

定理 4.10.3 (荷重の構成)**.** Riesz 表現子の列 (γ_t) が

$$\exists_{\delta>0}\ \exists_{\delta\in\left[0,\frac{\delta}{2+\delta}\right)}\ \exists_{C,C'>0}\ \left[\left(\frac{E_{t-1}[|\gamma_t|^{2+\delta}]}{E_{t-1}[\gamma_t^2]^{2+\delta}} \leqslant C\right) \wedge \left(\forall_{t\in[T]}\ E_{t-1}[\gamma_t^2] \leqslant C' t^{\alpha}\right)\right]$$

を満たすとする．この時，

(1) $\forall_{t<T}\ \gamma_T < 1$,
(2) $\gamma_T = 1$,
(3) $\exists_{C''>0}\ \dfrac{1}{1+T-t} \leqslant \lambda_t \leqslant C'' \dfrac{E_{t-1}[\gamma_t^2]^{-1}}{t^{-\alpha} + T^{1-\alpha} - t^{1-\alpha}}$

を満たす付値率 (allocation rate) について，

$$h_t^2 E_{t-1}[\gamma_t^2] = \left(1 - \sum_{s=1}^{t-1} h_s^2 E_{s-1}[\gamma_s^2]\right)\lambda_t$$

によって帰納的に定義した荷重 (h_t) は 3 つの公理 4.9.2 を満たす．

4.11　連続 martingale の周りに展開される確率変数の漸近展開

もし推定量 $\widehat{\psi}_t$ が $M_n + r_n N_n$ の形を持ち，いくつかの仮定を満たす確率過程だと証明できたならば，吉田先生の Malliavin 解析を用いた論文の手法を適用することで漸近展開をすることができる．

定義 4.11.1. ある正な predictable な過程 $a^T = (a_t^T)$ と 0 に収束する正実数列 (r_n) について，

(1) $M_t := \sum_{t=1}^T a_t \dot{\xi}_t$.
(2) $Z_T := M_T + r_T N_T$.

と定める．

第 5 章

Targeted Learning

5.1 歴史

Pearl の構造因果モデルを理解するには，計量経済学からの系譜を知る必要があったのだ！

歴史 5.1.1 (Marschak と計量経済学のノンパラ化 + グラフィカルモデル)**.** Cowles 財団が構造方程式モデリングにお熱だったころ，Jacob Marschak (53) だけは，個々のパラメータ推定は大事ではなく，その全体で見ると良い（統計的汎関数の推定＝ノンパラへの転換）ことに気づき，Heckman(2000) が再発見した．これを"Marschak's Maxim"という：” the desired quantity can often be identified without ever specifying the functional or distributional forms of these economic models". 当時の統計学者からは，そもそも定義できない量を推定しようとするその姿勢は，形而上学的でお気楽に思えたのだろう．これは統計学者の観点からは出て来ないパラダイムシフトである，実際 Karl Pearson (11) が"what if"という言葉を追放した．実験的な研究者の政策問題 (policy question) には，因果の言葉が欠かせないのである．

個々で何より，Donald Rubin (74) が，Neyman (23) の潜在結果の概念を一般化し，統計学の中で地位を取り戻させた．これで Marschak's Maxim は化けの皮を被って転生した．何より，targeted question を，推定可能な量に落とし込むという意味で極めて強力であるが，このままでは，構造方程式モデルの表現力に劣る．実際，データ生成過程を理解する，という点では，まだまだ平均処置効果だけでは情報が少なすぎる．一方で，パラメトリックの頭になっていた計量経済学者が，成分をそもそも持たない定式化に考え方を移行する契機となった．いまでは，Pearl のグラフィカルモデルを用いて，政策問題と，推定すべき因果効果との関係に，アルゴリズム的な対応がついた．これが，最後のピースであった．

> Lacking the benefits of graphical models, nonparametric re- searchers have difficulties locating instrumental variables in a system of equations, recognizing the testable implications of such systems, deciding if two such systems are equivalent, if two counterfactuals are independent given another, whether a set of measurements will reduce bias, and, most importantly, reading the causal and counterfactual information that such systems convey (Pearl 2009, pp. 374 – 380).

参考文献

[1] 清水昌平『統計的因果推論』(講談社，機械学習プロフェッショナルシリーズ，2017).
[2] Anastasios A. Tsiatis "Semiparametric Theory and Missing Data" (Springer, 2006).
[3] Christopher Winship, Stephen L. Morgan "Counterfactuals and Causal Inference: Methods and Principles for Social Research" 2nd ed. (Cambridge University Press, 2014)
[4] 芝孝一郎さんのブログ記事データから因果関係をどう導く？：統計的因果推論の基本、「反事実モデル」をゼロから
[5] P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2), pp. 362-378, 2008.
[6] "Confidence Intervals for Policy Evaluation in Adaptive Experiments"
[7] "Malliavin calculus and asymptotic expansion for martingales" (1997).
[8] 星野崇宏『調査観察データの統計科学』