目次

第 1 章

線型推測論

線型推測論が統計モデルの線形代数である．

Anderson による標準的教科書"An Introduction to Multivariate Statistical Analysis" (1958) が多変量解析の，George Box と Jenkins による Time Series Analysis, Forecasting and Control" (1976) が時系列解析の最初の金字塔であるが，これらはいずれも線形モデルを扱っている．大規模計算機時代を迎えて以降，非線形モデルが発展の中心となったが，応用における一次近似と，その後の非線形解析の方向性を見るための試金石としての意味で，線形モデルの意義は不動である．例えば，数学的に整理された解析力学を差し置いて，工学の場面では Newton 的力学がもっとも重要になることに等しい．実際，X のデザインを非線形にすることが可能であるから，線形性の仮定は応用上は強い制約ではない．

また，正規近似の手法で用いられる，正規分布から派生する確率分布（カイ 2 乗分布，t 分布，F 分布）についてまとめる．

1.1 射影行列と逆行列

1.1.1 射影

Euclid 空間 $\mathbb{R}^n$ では，任意の自己作用素は有界でである．よって，冪等律を満たす行列を射影行列といい，さらに自己共役であるものを直交射影という．

補題 1.1.1 (射影の特徴付け)**.** 線型写像 $P : \mathbb{R}^n \to \mathbb{R}^n$ について，

(1) $P \circ P = P$.
(2) $\mathbb{R}^n = \mathrm{Im}\, P \oplus \mathrm{Ker}\, P$ と表せる．

また，次の 2 条件とも同値．

(3) $Q \in \mathrm{GL}_n(\mathbb{R})$ が存在して，$Q^{-1}PQ = \mathrm{diag}[1, \cdots, 1, 0, \cdots, 0]$.
(4) $\mathrm{rank}P + \mathrm{rank}(I_n - P) = n$.

なお，このとき $\mathrm{rank}P = \mathrm{Tr}P = r$ である．

補題 1.1.2. 射影 $P \in M_n(\mathbb{R})$ について，

(1) P は直交射影である．
(2) $P^\top = P$.

1.1.2 一般化逆行列

記法 1.1.3.

(1) 行列 $A \in M_{mn}(\mathbb{R})$ の列ベクトルではられる線型空間を $L[A] \subset \mathbb{R}^n$ と表す．
(2) $X \in M_{np}(\mathbb{R})$ に対して，$L[X]$ が定める直交射影を P_X で表す．

(3) $H \in M_{pq}(\mathbb{R})$ は $L[H] \subset L[X^\top]$ かつ $\mathrm{rank}H = q$ を満たすとする．

$$L[X|H] := X((\mathrm{Im}\ H)^\perp) = \left\{X\gamma \in \mathbb{R}^n \mid \gamma \in \mathrm{Ker}\ H^\top = (\mathrm{Im}\ H)^\perp\right\}$$

への直交射影を $P_{X|H}$ で表す．$r := \mathrm{rank}X$ とする．なお，$\forall_{T\in B(H)}\ \mathrm{Ker}\ T^* = (\mathrm{Im}\ T)^\perp$ なので，$\mathrm{Ker}\ H^\top = (\mathrm{Im}\ H)^\perp$.

定義 1.1.4 (generalized inverse matrix)**.** 行列 $A \in M_{mn}(\mathbb{R})$ に対して，$AA^-A = A$ を満足する $A^- \in M_{nm}(\mathbb{R})$ を，A の**一般化逆行列**という．

補題 1.1.5. 行列 $A \in M_{mn}(\mathbb{R})$ について，

(1) 一般逆 A^- は存在する．
(2) 一意とは限らない．

［証明］.

(1) ある $B \in \mathrm{GL}_m(\mathbb{R}), C \in \mathrm{GL}_n(\mathbb{R})$ について，$D := BAC = \mathrm{diag}[1,\cdots,1,0,\cdots,0]$．$A^- := CD^\top B$ とおくと，$A = B^{-1}DC^{-1}$ より，$AA^-A = ACD^\top BA = B^{-1}DC^{-1}CD^\top BB^{-1}DC^{-1} = B^{-1}DD^\top DC^{-1} = B^{-1}DC^{-1} = A$.
(2) $A := (1\ 0; 0\ 0)$ について，A 自身も，$B := (1\ 1; 0\ 0)$ も一般逆である．

■

定理 1.1.6. $X \in M_{np}(\mathbb{R})$ とする．$X^\top X$ の任意の一般化逆行列 $(X^\top X)^-$ に対して，$P_X = X(X^\top X)^- X^\top$.

定理 1.1.7. $X \in M_{np}(\mathbb{R})$ とする．$X^\top X$ の任意の一般化逆行列 $(X^\top X)^-$ に対して，$G := (X^\top X)^- H$ と定める．このとき，$G^\top X^\top XG$ は正則であり，

$$P_{X|H} = P_X - XG(G^\top X^\top XG)^{-1}G^\top X^\top.$$

また，$\mathrm{rank}P_{X|H} = r - q$ である．

1.1.3　ブロック行列の逆

記法 1.1.8. $A \in M_p(\mathbb{R}), B \in M_{pq}(\mathbb{R}), C \in M_{qp}(\mathbb{R}), D \in M_q(\mathbb{R})$ について，$M := [A\ B; C\ D] \in M_{p+q}(\mathbb{R})$ とする．$D \in \mathrm{GL}_q(\mathbb{R})$ のとき $F := A - BD^{-1}C$，$A \in \mathrm{GL}_p(\mathbb{R})$ のとき $G := D - CA^{-1}B$ とおく．

命題 1.1.9. (1) D が正則であるとき，$|M| = |F||D|$．特に，D, M が正則であることと，D, F が正則であることとは同値．
(2) D, M が正則であるとき，

$$M^{-1} = \begin{bmatrix} F^{-1} & -F^{-1}BD^{-1} \\ -D^{-1}CF^{-1} & D^{-1} + D^{-1}CF^{-1}BD^{-1} \end{bmatrix}.$$

(3) A が正則であるとき，$|M| = |A||G|$．特に，A, M が正則であることと，A, G が正則であることとは同値．
(4) A, M が正則であるとき，

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BG^{-1}CA^{-1} & -A^{-1}BG^{-1} \\ -G^{-1}CA^{-1} & G^{-1} \end{bmatrix}.$$

1.2　カイ 2 乗分布

記法 1.2.1. Gamma 分布 $G(\alpha, \nu) : \mathbb{R}^2_{>0} \to P(\mathbb{R}, \mathcal{B}_1)$ の確率密度関数は

$$g(x;\alpha,\nu) = \frac{1}{\Gamma(\nu)}\alpha^\nu x^{\nu-1}e^{-\alpha x}1_{x>0}, \qquad \Gamma(\nu) = \int_0^\infty t^{\nu-1}e^{-t}dt.$$

と表せる．

定義 1.2.2 (noncentral chi-square distribution with k degree of freedom and noncentrality parameter δ)**.** $(X_j)_{j\in[k]}$ を $X_1 \sim N(\nu, 1), X_j \sim N(0,1)$ $(j = 2,\cdots,k)$ を満たす独立な実確率変数列とする．このとき，二乗和で定まる確率変数 $Y := \sum_{j=1}^k X_j^2$ が定める分布を**自由度 k，非心率 $\delta = \mu^2$ の非心カイ 2 乗分布**といい，$\chi^2(k, \delta)$ で表す．

注 1.2.3. $\delta = 0$ のとき，$\chi^2(k) = \chi^2(k,0)$ をカイ 2 乗分布という．自由度 $k = 0$ の非心カイ 2 乗分布は $x = 0$ 上のデルタ測度とし，非心率は $\delta = 0$ とする．

命題 1.2.4. $\chi^2(k, \delta = \mu^2)$ について，

(1) 確率密度関数は $f(x) := e^{-\frac{\mu^2}{2}} \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{\mu^2}{2}\right)^r g\left(x; \frac{1}{2}, r + \frac{k}{2}\right)$.

(2) 特性関数は

$$\varphi(u) = e^{-\frac{\mu^2}{2}} \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{\mu^2}{2}\right)^r (1-2iu)^{-r-\frac{k}{2}} = (1-2iu)^{-\frac{k}{2}} \exp\left(\frac{\mu^2 iu}{1-2iu}\right)$$

系 1.2.5. $X_j \sim N(\mu_j, 1)$ $(j \in [k])$ を独立な実確率変数列とする．このとき，$\delta := \sum_{j=1}^{k} \mu_j^2$ とすると，$\sum_{j=1}^{k} X_j^2 \sim \chi^2(k, \delta)$.

系 1.2.6. (X_j) $(j \in [k])$ を，$\chi^2(k_j, \delta_j)$ に従う独立同分布確率変数列とする．このとき，$\sum_{j=1}^{k} X_j \sim \chi^2\left(\sum_{j=1}^{k} k_j, \sum_{j=1}^{k} \delta_j\right)$.

1.3　フィッシャー・コクランの定理

命題 1.3.1. $A_i \in M_n(\mathbb{R})$ $(i \in [k])$ は $\sum_{i=1}^{k} A_i = I_n$ を満たすとする．このとき，次の 4 条件は同値．

(1) $\forall_{i \in [k]}\ A_i^2 = A_i$.

(2) $\forall_{i,j \in [k]} i \neq j \Rightarrow A_i A_j = O$.

(3) $\mathbb{R}^n = \bigoplus_{i=1}^{k} \mathrm{Im}\, A_i$ と直和分解される．

(4) $\sum_{i=1}^{k} \mathrm{rank} A_i = n$.

記法 1.3.2. $\mu := (\mu_1, \cdots, \mu_n)^\top \in \mathbb{R}^n$，$Y = (Y_1, \cdots, Y_n)^\top \sim N_n(v, I_n)$ とする．

定理 1.3.3 (Fisher-Cochran)**.** $A_i \in M_n(\mathbb{R})$ $(i \in [k])$ は $\sum_{i=1}^{k} A_i = I_n$ を満たす対称行列の列とする．$Q_i := Y^\top A i Y, \mathrm{rank} A_i =: n_i$ とおく．このとき，次の 5 条件は同値．

(1) $\exists_{\delta_i \geqslant 0}\ Q_i \sim \chi^2(n_i, \delta_i)$ かつ $Q_1, \cdots, Q_k$ は独立．

(2) $\sum_{j=1}^{k} n_i = n$.

(3) $\forall_{i \in [k]}\ A_i^2 = A_i$.

(4) $\forall_{i,j \in [k]}\ i \neq j \Rightarrow A_i A_j = O$.

(5) $\mathbb{R}^n = \bigoplus_{i \in [k]} \mathrm{Im}\, A_i$.

さらにこのとき，$\delta_i = \mu^\top A_i \mu$ と表わせ，特に $\sum_{j=1}^{n} \mu_j^2 = \sum_{i=1}^{k} \delta_i$.

系 1.3.4. $A \in M_n(\mathbb{R})$ を対称行列とする．次の 2 条件は同値．

(1) $\exists_{\delta \geqslant 0}\ \exists_{k \in \mathbb{Z}_+}\ Q := Y^\top A Y \sim \chi^2(k, \delta)$.

(2) $A^2 = A$.

このとき，$k = \mathrm{rank} A, \delta = \mu^\top A \mu$ が成り立つ．

1.3.1　独立性

系 1.3.5. $A_i \in M_n(\mathbb{R})$ $(i = 1, 2)$ を対称行列とし，$Q_i := Y^\top A_i Y$ は非心カイ 2 乗分布に従うとする．このとき，次の 2 条件は同値．

(1) $Q_1 \perp\!\!\!\perp Q_2$.
(2) $A_1 A_2 = O$.

補題 1.3.6. $A, B \in M_n(\mathbb{R})$ を対称な射影行列とする：$A^2 = A, B^2 = B$. 差は非負値行列とする：$A - B \geqslant 0$. このとき，

(1) $AB = BA = B$.
(2) $(A-B)^2 = A - B$.

定理 1.3.7. $A_i \in M_n(\mathbb{R})$ $(i = 0,1,2)$ を $A_0 = A_1 + A_2$ を満たす対称行列とし，$Q_i := Y^\top A_i Y$ は $Q_j \sim \chi^2(n_j, \delta_j)$ $(j = 0,1), Q_2 \geqslant 0$ a.s. とする．このとき，$n_2 := n_0 - n_1, \delta_2 = \delta_0 - \delta_1$ について

(1) $Q_2 \sim \chi^2(n_2, \delta_2)$.
(2) $Q_1 \perp\!\!\!\perp Q_2$.

1.4　t 分布と F 分布

1.4.1　t 分布

定義 1.4.1 (noncentral t-distribution). $Y \sim \chi^2(n), X \sim N(0,1)$ は独立であるとする．このとき，$X := \dfrac{Z}{\sqrt{\frac{Y}{n}}}$ が定める分布を，**自由度 n，非心率 δ の非心 t 分布**といい，$t(n,\delta)$ で表す．特に $\delta = 0$ のとき $t(n) := t(n,0)$ を**自由度 n の t 分布**という．

命題 1.4.2 (確率密度関数).

1.4.2　F 分布

定義 1.4.3 (noncentral F-distribution). $Y_1 \sim \chi^2(m,\delta), Y_2 \sim \chi^2(n)$ は独立であるとする．このとき，$X := \dfrac{Y_1/m}{Y_2/m}$ が定める分布を，**自由度 m,n，非心率 δ の非心 F 分布**といい，$F(m,n,\delta)$ で表す．特に $\delta = 0$ のとき $F(m,n) := F(m,n,0)$ を**自由度 m,n の F 分布**という．

命題 1.4.4 (確率密度関数).

1.5　ガウス・マルコフモデル

$Y = X\beta + \epsilon$ なる形で表せるモデルを線型回帰モデルといい，$X \in \mathbb{R}^p$ の次元が $p \geqslant 2$ を満たすとき，重回帰という．この誤差項 ϵ について，$E[\epsilon] = 0$ かつ $\mathrm{Var}[\epsilon] = \sigma^2 I_n$ なる仮定 (Gauss-Markov assumption) をおいた線型回帰モデルを **Gauss-Markov モデル**という．

記法 1.5.1 (Gauss-Markov model). n 個の実確率変数 Y_i $(i \in [n])$ が，定数関数 $x_i, \beta : \Omega \to \mathbb{R}^p$ と，期待値 0 かつ互いに無相関で等分散を持つ $E[\epsilon_i] = 0, E[\epsilon_i \epsilon_j] = \sigma^2 \delta_{ij}$ を満たす確率変数列 $\epsilon_i : \Omega \to \mathbb{R}$[†1] について

$$Y_i = x_i^\top \beta + \epsilon_i$$

なる関係を満たすとする．添字 $i \in [n]$ は観測の単位を表し，x_i を，$E[Y_i]$ を説明する p 個の要因を並べた変数，σ^2 は観測の分散を表すパラメータである．

$$Y := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \epsilon := \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad X := \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix}$$

[†1] 独立ならば無相関である．

をそれぞれ n 次元ベクトル値確率変数と $n \times p$ 次元定行列とすると，モデルは

$$Y = X\beta + \epsilon, \quad E[\epsilon] = 0, \quad \mathrm{Var}[\epsilon] = \sigma^2 I_n.$$

と表せる．すなわち，

$$\mathcal{P} := \left\{ \nu \in P(\mathbb{R}^n, \mathcal{B}_n) \mid \exists_{\beta \in \mathbb{R}^p} E[\nu] = X\beta, \exists_{\sigma \in (0,\infty)} \mathrm{Var}[\nu] = \sigma^2 I_n \right\}$$

なるモデルである（正規分布よりも遥かに広いクラスである）．

1.5.1　正規方程式

結局，線型汎関数 $\beta \mapsto c^\top \beta$ が推定可能であるとき，正規方程式の解から最良線形不偏推定量が構成できる．これが M-推定量の雛形となる．

定義 1.5.2 (sum of squared residuals, normal equation)**.** 未知パラメータ $\beta \in \mathbb{R}^p$ の推定を考える．

$$Q(\beta) = \mathrm{SSR}(\beta) =: (Y - X\beta)^\top (Y - X\beta) = \sum_{i=1}^{n} (Y_i - x_i^\top \beta)^2$$

について，$\widehat{\beta} := \arg\min Q(\beta)$ とすればよい．特にこれは，**正規方程式**

$$\frac{dQ}{d\beta} = 0 \Leftrightarrow X^\top X\beta = X^\top Y$$

の解である．この解は $L[X^\top X] = L[X^\top]$ より，この解は常に存在し，$X^\top X$ が正則であるとき一意的に存在する．

定理 1.5.3. $Y \in \mathbb{R}^n$ とする．正規方程式の任意の解 $\widehat{\beta}$ に対して，

(1) $X\widehat{\beta} = P_X Y$.
(2) $|Y - X\beta\beta| = \min_{\beta \in \mathbb{R}^p} |Y - X\beta|$.

1.5.2　最良線型不偏推定量

未知パラメータ β, σ^2 の関数（主に線型汎関数）$g(\beta, \sigma^2)$ の推定を考える．どの $(\beta, \sigma^2) \in \Pi \subset \mathbb{R}^p \times (0, \infty)$ が真の値であろうと，良い結果がほしいとなると，これを不偏性という．

定義 1.5.4 (unbiased estimator)**.** 統計量 $\delta(Y) : \mathrm{Meas}(\mathcal{X}, \mathbb{R}) \to \mathbb{R}$ について，

$$\forall_{(\beta,\sigma^2) \in \Pi} \quad E_{\beta,\sigma^2}[\delta(Y)] = g(\beta, \sigma^2)$$

が成り立つとき，これを**不偏推定量**であるという．

定義 1.5.5 (estimable)**.** $c \in \mathbb{R}^p$ について，$c \in L[X^\top]$ のとき，線型汎関数 $\mathbb{R}^p \ni \beta \mapsto c^\top \beta \in \mathbb{R}$ を**推定可能**であるという．

定理 1.5.6. $a \in \mathbb{R}^n, c \in \mathbb{R}^p, \beta \in \mathbb{R}^p$ について，

(1) 次の 2 条件は同値．
  (a) 確率変数 Y の線型関数 $a^\top Y$ が β の線型関数 $c^\top \beta$ の不偏推定量である．
  (b) $X^\top a = c$.
(2) 次の 2 条件は同値．
  (a) 線型関数 $c^\top \beta$ の線形不偏推定量が存在する．
  (b) $c^\top \beta$ は推定可能である．

定理 1.5.7. 線型関数 $c^\top \beta$ は推定可能であるとする．このとき，

(1) $c^\top$ と $c^\top\hat{\beta}$ は，ある $b \in \mathbb{R}^p$ を用いて $c = X^\top Xb, c^\top\hat{\beta} = b^\top X^\top Y$ と表される．
(2) $c^\top\hat{\beta}$ は正規方程式の解 $\hat{\beta}$ の取り方に依らず，$c^\top\beta$ の線形不偏推定量である．

定理 1.5.8 (BLUE: best linear unbiased estimator)**.** β の線型関数 $c^\top\beta$ は推定可能であるとする．このとき，正規方程式の任意の解 $\hat{\beta}$ に対して，$c^\top\hat{\beta}$ は $c^\top\hat{\beta}$ の**最良線型不偏推定量**である：

$$\forall_{(\beta,\sigma^2)\in\Pi} \quad \mathrm{Var}[c^\top\hat{\beta}] = \min\left\{\mathrm{Var}[a^\top Y] \geqslant 0 \mid a^\top Y \text{ は } c^\top\beta\text{の線形不偏推定量}\right\}.$$

ただし，Var とは，真値が (β,σ^2) であるときの分散を表す．

定理 1.5.9 (BLUE の特徴付け)**.** 次の 3 条件は同値である．

(1) 線型統計量 $a^\top Y$ は線型関数 $c^\top\beta$ の最良線型不偏推定量である．
(2) $c^\top\beta$ は推定可能で，任意の真値 $(\beta,\sigma^2) \in \Pi$ に対して，$c^\top\hat{\beta} = a^\top Y$ a.s.
(3) $X^\top a = c$ かつ $a \in L[X]$.

1.5.3　分散共分散行列

1.5.4　分散の推定

1.6　仮説検定

ある部分集合 $\Theta_0 \subset \Theta := \mathbb{R}^p$ の部分集合について，問題

$$H_0 : H^\top\beta = d \quad \text{vs.} \quad H_1 : H^\top\beta \neq d$$

を検定することを考える．Gauss-Markov モデルに対して，さらに誤差項の分布の正規性 $\epsilon \sim N_n(0,\sigma^2 I_n)$ を仮定する．この仮定は，誤差分布を楕円形分布に拡張することや，ロバスト推測論によって乗り越えられる．

1.6.1　理論

記法 1.6.1. $R_0^2 := \min_{\beta\in\mathbb{R}^p}|Y - X\beta|^2$ とし，$r := \mathrm{rank}X < n$ を仮定する．

定理 1.6.2. 任意の真値 $(\beta,\sigma^2) \in \Pi$ について，$\dfrac{R_0^2}{\sigma^2} \sim \chi^2(n-r)$.

定理 1.6.3. 次を仮定する．

(A1) $H \in M_{p,q}(\mathbb{R})$ $(q < p)$ は $L[H] \subset L[X^\top]$ かつ $\mathrm{rank}H = q$ を満たすとする．
(A2) $d \in \mathbb{R}^q$ について，$\Theta_0 = \left\{\beta \in \mathbb{R}^p \mid H^\top\beta = d\right\}$ と表せるとする．

このとき，$R_1^2 := \{\beta \in \Theta_0\}|Y - X\beta|^2$ について，

(1) 任意の真値 $(\beta,\sigma^2) \in \Pi$ について，R_0^2 と $R_1^2 - R_0^2$ は独立であり，$\dfrac{R_0^2}{\sigma^2} \sim \chi^2(n-r)$，かつ，ある $\sigma \in \mathbb{R}_+$ に対して，$\dfrac{R_1^2 - R_0^2}{\sigma^2} \sim \chi^2(q,\delta)$．特に，
$$\frac{(R_1^2 - R_0^2)/q}{R_0^2/(n-r)} \sim F(q, n-r, \delta).$$
(2) 真値 β が $\beta \in \Theta_0$ を満たすならば，$\dfrac{R_1^2 - R_0^2}{\sigma^2} \sim \chi^2(q)$ であり，特に
$$\frac{(R_1^2 - R_0^2)/q}{R_0^2/(n-r)} \sim F(q, n-r).$$

命題 1.6.4.

1.6.2　応用

このように，F 分布に基づく検定を F-検定という．このように，データのなすベクトルの 2 次形式による，データの変動の分解を基に検定を構成する手法を**分散分析** (ANOVA: ANalysis Of VAriance) と呼ぶ．

1.7　平均の検定

パラメトリックモデル $N(\mu, \sigma^2)$ を仮定し，ここからの無作為標本 Y_j $(j \in [n])$ から平均 μ に関する仮説 $\mu = \mu_0$ の検定は両側 t-検定となる．

1.8　重回帰分析

重回帰分析においては，**切片項 β_0 を含む Gauss-Markov モデル**が用いられる．

議論 1.8.1.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

なる関係で説明したい．n 個の観測 Y_i $(i \in [n])$ を説明したい場合，ベクトル表記で $Y = X\beta + \epsilon$ とまとめられるが，$i \in [n]$ が個体番号を表すとき，誤差項の独立性は成り立つであろう．こうして，Gauss-Markov モデルにたどり着く．

1.9　一元配置

one-way ANOVA

1.10　二元配置

第 2 章

統計的決定理論

（意志）決定理論の立場から推定と検定の小標本理論を考察する．統計的な推定を行うことは，特定の損失関数を採用して，これについて統計的決定問題を解くことに等しい．

統計的決定問題が設定されたとき，決定問題の下で起こる確率現象を解明するのが数理統計学の課題であって，特定の決定関数を無条件に是とするものではない．

2.1 統計推論の枠組み

決定すべきは関数＝射である．その文脈は 8-組で表現できる．

2.1.1 Wald の定義

定義 2.1.1 (statistical decision problem, sample space, decision / action space, loss function, nonrandomized, randomized decision function (Abraham Wald))**.** 次の 8-組 $(\mathcal{X}, \mathcal{A}, \mathcal{P}, \Theta, \mathcal{D}, \mathcal{B}, W, \Delta)$ を**統計的決定問題**という．

(1) $(\mathcal{X}, \mathcal{A})$ は可測空間で，**標本空間**という．
(2) Θ はパラメータの空間で，$\mathcal{P} := (P_\theta)_{\theta\in\Theta}$ は $(\mathcal{X}, \mathcal{A})$ の確率分布の族．
(3) $(\mathcal{D}, \mathcal{B})$ は可測空間で，**決定空間**または**行動空間**という．
(4) 第二変数について可測な関数 $W : \Theta \times \mathcal{D} \to \mathbb{R}_+$ を**損失関数**という．
(5) Δ は決定関数の族とする．
  (a) 可測写像 $\delta : \mathcal{X} \to \mathcal{D}$ を**非確率的決定関数**という．
  (b) 第一変数についての currying $\mathcal{X} \to \mathrm{Map}(\mathcal{B}, [0,1])$ は $(\mathcal{D}, \mathcal{B})$ 上の確率測度を与え，第二変数についての currying $\mathcal{B} \to \mathrm{Map}(\mathcal{X}, [0,1])$ は $\mathcal{X}$ 上の $\mathcal{A}$-可測関数となる関数 $\delta : \mathcal{B} \times \mathcal{X} \to [0,1]$ を**確率的決定関数**という．

注 2.1.2. 非確率決定関数 $\delta : \mathcal{X} \to \mathcal{D}$ が与えられたとき，任意の $B \in \mathcal{B}$ に対して $\tilde{\delta}(B|x) := 1_{\{\delta(x)\in B\}}$ で $\tilde{\delta} : \mathcal{B} \times \mathcal{X} \to [0,1]$ が定まるから，確率的決定関数の方がより一般的な設定となる．

例 2.1.3 (平均値の推定量の決定)**.** $(\mathcal{X}, \mathcal{A}) := (\mathbb{R}^n, \mathcal{B}_n)$, $\mathcal{P}_1 := \left\{ P = P_*^n \in \mathcal{M}(\mathbb{R}^n) \mid P_* \text{は } \mathbb{R} \text{ 上の確率分布で} \int_{\mathbb{R}} x^2 P_*(dx) < \infty \right\}$, $\Theta := \mathcal{P}_1$ とする．決定空間は平均値の全体としたいから $(\mathcal{D}, \mathcal{B}) := (\mathbb{R}, \mathcal{B}_1)$．推定量 $\Delta := \{T_1, T_2, T_3\}$ はそれぞれ $T_1 := X_1, T_2 := \sum_{j=1}^n \frac{X_j}{n}, T_3 := X_n$ という非確率的決定関数の族で，損失関数は $W(P, a) := (a - \mu_1(P))^2$ とする．

すると，損失関数 W から定まる危険関数は

$$\begin{aligned} R(P, T_i) &= \int W(P, T_i(x_1, \cdots, x_n)) P(dx_1, \cdots, dx_n) \quad (P \in \mathcal{P}_1) \\ &= \mathrm{Var}_P[T_i] \end{aligned}$$

となる．こうして，$R(P, T_1) = R(P, T_3) = \mathrm{Var}_P[X_1], R(P, T_2) = \dfrac{\mathrm{Var}_P[X_1]}{n}$ となり，T_2 が危険関数 R を最小にするとわかる．

2.1.2　危険関数

定義 2.1.4 (risk function, admissible)**.** **危険関数** $R:\Theta\times\Delta\to\mathbb{R}$ を次のように定める．

(1) 確率的な決定関数 $\delta:\mathcal{B}\times\mathcal{X}\to[0,1]$ については，$R(\theta,\delta):=\int_{\mathcal{X}}\int_{\mathcal{D}}W(\theta,a)\delta(da|x)P_\theta(dx)$ と定める．

(2) 非確率的な決定関数 $\delta:\mathcal{X}\to\mathcal{D}$ については，$R(\theta,\delta):=\int_{\mathcal{X}}W(\theta,\delta(x))P_\theta(dx)$ と定める．

2 つの決定関数 $\delta_1,\delta_2\in\Delta$ について，

(1) $\forall_{\theta\in\Theta}\ R(\theta,\delta_1)\leqslant R(\theta,\delta_2)$ が成り立つとき，δ_1 と δ_2 は**同程度に良い**決定関数であるという．

(2) 同程度に良い決定関数がさらに $\exists_{\theta_0\in\Theta}\ R(\theta_0,\delta_1)<R(\theta_0,\delta_2)$ を満たすとき，δ_1 は δ_2 より**一様に良い**決定関数であるという．

(3)「一様に良い」という関係は順序を定める．Δ に最大元が存在せず，極大元 δ_* のみが存在するとき，**許容的**であるという．

2.2　統計量の十分性と完備性

統計量とは可測関数 $T:(\mathcal{X},\mathcal{A})\to(\mathcal{T},\mathcal{B})$ をいう．考えている確率分布族 $\{P_\theta\}_{\theta\in\Theta}\subset P(\mathcal{X},\mathcal{A})$ に応じて，T がその構造を保つかどうかが変わる．が，一般的に好ましい性質が定義できる．

2.2.1　十分統計量

θ の十分統計量 T とは，θ の推定に関する限り，データから得られる情報を漏らさず含んでいることを表している．すなわち，T で条件付けると，その条件づき確率がもはや θ に依らなくなるような統計量 $T:\mathcal{X}\to\mathbb{T}$ をいう．

記法 2.2.1. 可測関数 $T:(\mathcal{X},\mathcal{A})\to(\mathcal{T},\mathcal{B})$ を考える．確率分布族 $\{P_\theta\}_{\theta\in\Theta}\subset P(\mathcal{X},\mathcal{A})$ についての条件付き確率 $P_\theta[A|T=t]$ が，母数 $\theta\in\Theta$ に依存しないとき，T を得てしまえば，これはデータの持つ情報のすべてを持っていると言える．

定義 2.2.2 (sufficient statistic)**.** $T:\mathcal{X}\to\mathcal{T}$ が $\{P_\theta\}$ に関して**十分**であるとは，次が成り立つことをいう：

任意の $A\in\mathcal{A}$ に対して，$\mathcal{B}$-可測関数 $q(A|-):\mathcal{B}\to\mathbb{R}$ が存在して，

$$\forall_{B\in\mathcal{B}}\ \forall_{\theta\in\Theta}\quad \int_B q(A|t)P_\theta^\top(dt)=P_\theta[A\cap T^{-1}(B)].$$

例 2.2.3 (order statistic)**.** 確率分布族

$$\mathcal{P}:=\{P\in P(\mathbb{R}^n)\mid\forall_{\sigma\in S_n}\ P^\sigma=P\}$$

を考えているとき，$x=(x_1,\cdots,x_n)$ の各成分を小さい順に並べて出来る列 $t(x)=(x_{(1)},\cdots,x_{(n)})$ へ写す可測関数 $T:\mathbb{R}^n\to\mathbb{R}^n;x\mapsto t(x)$ を**順序統計量**という．

2.2.2　Fisher の因子分解定理

十分統計量を構成する十分条件を探す．

2.2.3　Rao-Blackwell の定理

決定関数＝「推定量の候補」としては，十分統計量の関数のみを考えれば十分であることを定立する．あるいは，適当な推定量 δ に対して，$\delta_0 := E[\delta|T]$ とすれば，これは δ よりも同程度に良い決定関数となる．

定理 2.2.4 (Rao-Blackwell)**.** T をモデル $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$ に対する十分統計量であるとする．このとき，非確率的決定関数 δ は $\forall_{\theta\in\Theta}\ E_\theta[|\delta|] < \infty$ を満たすとする．これに対して，$\delta_0 := E[\delta|T]$ と定めると，これは同程度に良い決定関数である：

$$\forall_{\theta\in\Theta}\quad R(\theta,\delta) \leqslant R(\theta,\delta_0)$$

2.2.4　完備性

定義 2.2.5 (complete, boundedly complete)**.**

(1) 分布族 $\mathcal{P} \subset P(\mathcal{X},\mathcal{A})$ が [有界的] **完備**であるとは，任意の [有界] 可測関数 $f : \mathcal{X} \to \mathbb{R}$ に関して $\forall_{\theta\in\Theta}\ E_\theta[f] = 0 \Rightarrow f = 0\ \mathcal{P}$- a.s. が成り立つことという．ただし，$\mathcal{P}$- a.s. とは，任意の $P \in \mathcal{P}$ について P- a.s. であることをいう．

(2) 可測関数 $T : \mathcal{X} \to \mathcal{T}$ が [有界的] **完備**であるとは，$(\mathcal{T},\mathcal{B})$ 上に引き起こされる分布族 $(P_\theta^T)_{\theta\in\Theta}$ が [有界的] 完備であることをいう．

2.3　指数型分布族

2.4　統計的推定

2.4.1　不偏推定

定義 2.4.1 (unbiased estimator, U-estimable)**.** 確率分布族 (P_θ) に対して，

(1) $\forall_{\theta\in\Theta}\ E_\theta[\delta] = \theta$ を満たす推定量 δ を**不偏推定量**という．

(2) パラメータの関数 $g : \Theta \to \Theta$ に対して不偏推定量が存在するとき，g を U-**推定可能**という．

2.4.2　正則な統計的実験

- 確率空間 $(\mathcal{X},\mathcal{A},P_\theta)$ のエントロピー＝平均情報量とは，対数尤度の平均 $-E[\log L(\theta|x)|\theta]$ であり，系の乱雑さを表す．一様分布で，どの事象も起こり得て予測が難しいとき，値が大きくなる．
- 対数尤度の θ に関する微分 $V(x;\theta)$ の平均は $E[V(x;\theta)|\theta] = 0$ を満たすが（θ がいくら変わろうと確率分布である限り総重量は 1 なので），この分散に興味がある．もし至る所で変わるなら，分散は極めて大きくなるはずである．Fisher 情報量とはこれであり，分布 P_θ 自身が母数 θ に関してもつ情報の量を表す．

定義 2.4.2 (Fisher's finite information)**.** $L_2(\mathcal{X},\nu)$ について，

$$I(\theta) := 4\int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta}p^{1/2}(x;\theta)\right)^2 d\nu = 4\left\|\frac{\partial}{\partial\theta}p^{1/2}\right\|_\nu^2$$

を Fisher 情報，

$$I(\theta) = 4\int_{\mathcal{X}} \frac{\partial}{\partial\theta}p^{1/2}(x;\theta)\left(\frac{\partial}{\partial\theta}p^{1/2}(x;\theta)\right)^\top d\nu$$

を Fisher 情報行列という．ここで，関数 $p^{1/2}(-;u)$ が $u = \theta$ で微分可能で導関数が $L_2(\nu)$ 級であることは必ずしも仮定しない．

$$\exists_{\psi\in L^2(\mathcal{X},\mathrm{Meas}(\Theta,\mathbb{R}^K))}\ \int_{\mathcal{X}} |g(x;\theta+h) - g(x;\theta) - (\psi(x;\theta),h)|^2\, d\nu = o(|h|^2)\quad (h\to 0)$$

が満たされるとき，統計的実験 E は $\theta \in \Theta$ で**有限な Fisher 情報を持つ**という．$\dfrac{\partial p}{\partial \theta}(x;\theta)$ はこの L^2-微分の意味で表している．

要諦 2.4.3. 対数尤度関数 $\log L(\theta|x)$ の θ による微分をスコア関数という．

$$V(x;\theta) = \frac{\partial}{\partial \theta} \log L(\theta|x) = \frac{1}{L}\frac{\partial L}{\partial \theta}.$$

θ を増やしたら尤度が上がる場合，スコア関数は正で，さらに尤度が低ければ低いほど $1/L$ により拡大される．これはたしかに情報量の考え方である．実は $E[V(x;\theta)|\theta] = 0$ が成り立つ．そこで，スコア関数の 2 次のモーメント $I(\theta) := E[V(x;\theta)^2|\theta]$ を Fisher 情報量というと，これはスコア関数の分散である：$I(\theta) = \mathrm{Var}[V(x;\theta)]$．なお，2 つの定義の等価性

$$\begin{aligned}\left(\frac{\partial}{\partial\theta}p^{1/2}(x;\theta)\right)^2 &= \left(\frac{1}{2}p^{-1/2}\frac{\partial}{\partial\theta}p(x;\theta)\right)^2 \\ &= \frac{1}{4}p^{-1}(x;\theta)\left(\frac{\partial p(x;\theta)}{\partial\theta}\right)^2\end{aligned}$$

による．

定義 2.4.4 (regular statistical experiment)**.** 組 $E := (\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta}, \Theta \overset{\text{open}}{\subset} \mathbb{R}^K)$ が，P_θ が ν-絶対連続であるとき，統計的実験という．E が Θ 上で**正則**であるとは，次の 3 条件を満たすことをいう：

(1) Radon-Nikodym 微分の族 $p(x,-) : \Theta \to [0,1]$ は ν-a.e.x に関して連続である．
(2) 任意の点 $\theta \in \Theta$ において有限な Fisher 情報が存在する．
(3) 関数 $\psi(-,\theta) \in L^2(\nu)$ は連続．

2.4.3　最適性・有効性

一般に不偏推定量は複数ある．そこで，設定した損失関数に応じて，「最適」な推定量を選ぶ必要がある．一般には二乗損失を考えるため，分散を最適性の指標とし，分散が最も小さい推定量を**有効 (efficient)** であるという．

記法 2.4.5. 統計的決定理論の枠組みで不偏推定問題を考えると，次の通りになる．

(1) 標本空間 $(\mathcal{X}, \mathcal{A})$ 上の確率分布族 $(P_\theta)_{\theta\in\Theta}$ を考える．
(2) 決定空間は，$\mathcal{D} \subset \mathbb{R}^p$ を凸ボレル集合，$\mathcal{B}$ はその上のボレル集合族とする．
(3) 損失関数 W は凸とする．
(4) $\Delta := \Delta_g$ は p 次元関数 $g(\theta)$ の不偏推定量を非確率論的決定関数 $\delta : \mathcal{X} \to \mathcal{D}$ と観て，その全体

$\Delta_g \neq \varnothing$，すなわち，$g : \Theta \to \Theta^p$ は U-推定可能とする．

定理 2.4.6 (Lehmann-Scheffe)**.** T を分布族 $\mathcal{P}$ に対する完備十分統計量とする．任意の不偏推定量 $\delta \in \Delta_g$ に対して，$\delta_0 := E[\delta|T]$ とすると，δ_0 は $g(\theta)$ の最良の不偏推定量である：$\forall_{g'\in\Delta_g}\ \forall_{\theta\in\Theta}\ R(\theta,\delta_0) \leqslant R(\theta,\delta')$.

要諦 2.4.7. 一般には，$p = 1$ 次元とし，損失関数は二乗損失 $W(\theta,a) = [a - g(\theta)]^2$ を考えることが多い．このとき，危険関数は $R(\theta,\delta) = \mathrm{Var}_\theta[\delta]$ となる．したがって，分散を次の「最適性」の指標とする．

定義 2.4.8 (UMVUE: uniformly minumum variance unbiased estimator)**.** $g(\theta)$ の任意の不偏推定量 $\delta' \in \Delta_g$ に対して，$\forall_{\theta\in\Theta}\ \mathrm{Var}_\theta[\delta] \leqslant \mathrm{Var}_\theta[\delta']$ を満たす不偏推定量 δ を**一様最小分散不偏推定量**という．

2.4.4　Cramer-Rao の不等式

Fisher 情報量は，スコア関数の 2 次のモーメントと定義される．また，KL 情報量の二次の項としても登場する．これを用いて算出される，不偏推定量の「最適性」の情報理論的限界を，不偏推定量の分散行列の下界を与えることで示す定理を，Cramer-Rao の不等式という．したがって，Cramer-Rao の不等式の下界を達成する不偏推定量が存在するならば，それは UMV 不偏推定量である．

定義 2.4.9. パラメータ $\theta = (\theta_i) \in \mathbb{R}^q$ に依存する確率密度関数 $p_\theta(x)$ に対して，

$$I_{ij}(\theta) := \int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \log p_\theta(x) \cdot \frac{\partial}{\partial \theta_j} \log p_\theta(x) \cdot p_\theta(x) \mu(dx)$$

として定まる行列 $I(\theta)$ を **Fisher 情報行列**という．特に $q = 1$ の場合，Fisher 情報量という．

記法 2.4.10. パラメータの関数 $g : \Theta \to \Theta^p$ について，$J(\theta) := \dfrac{\partial g}{\partial \theta}$ を $p \times q$-行列とする．

定理 2.4.11 (Cramer-Rao)**.** $U \overset{\text{open}}{\subset} \mathbb{R}^q$ で添字付けられたパラメトリックモデル $(P_\theta)_{\theta \in U}$ と，偏微分可能な関数 $g : \mathbb{R}^q \to \mathbb{R}^p$ と，その $\theta \in U$ における不偏推定量 $\delta \in \mathcal{L}^2(\mathcal{X}^n; \mathbb{R}^p)$ とについて，次を仮定する：

(A0) 分布族 $(P_\theta)_{\theta \in U}$ はある σ-有限な参照測度 $\mu \in P(\mathcal{X})$ に関して絶対連続とし，その Radon-Nikodym 微分を $p_\theta : \mathcal{X} \to [0, 1]$ で表す．

(A1) $p_-(x) : U \to [0, 1]$ は P_θ-a.e. x に関して偏微分可能である．

(A2) $p_\theta(x) > 0$ を満たす点 (x, θ) の上で，スコア関数の第 i 成分 $\psi_i : \mathcal{X} \times U \twoheadrightarrow \mathbb{R}$ を $\psi_i(x, \theta) := \dfrac{\partial \log p_\theta(x)}{\partial \theta_i} = \dfrac{1}{p_\theta} \dfrac{\partial p_\theta}{\partial \theta_i}$ と定めると[†1]，二次の絶対積率が有限 $\psi_i(-, \theta) \in \mathcal{L}^2(P_\theta)$：

$$E_{P_\theta}[|\psi_i|^2] = \int_{\mathcal{X}} |\psi_i(x, \theta)|^2 p_\theta(x) \mu(dx) < \infty \qquad (\forall_{i \in [q]}).$$

(A3) 各 $\theta \in U$ における Fisher 情報行列 $I = (I_{ij})$ を，第 (i, j)-成分を

$$I_{ij}(\theta) := \mathrm{Cov}_{P_\theta}[\psi_i, \psi_j] = \int_{\mathcal{X}} \frac{\partial \log p_\theta(x)}{\partial \theta_i} \frac{\partial \log p_\theta(x)}{\partial \theta_j} \cdot p_\theta(x) \mu(dx) \qquad (\forall_{i,j \in [q]}).$$

とすることによって定めると，正定値な対称行列となる．

(A4) 不偏推定量 δ は $\psi := (\psi_1, \cdots, \psi_q)^\top$ が定める推定方程式の解であり：$E_{P_\theta}[\psi] = \displaystyle\int_{\mathcal{X}} \psi(x, \theta) p_\theta(x) \mu(dx) = 0 \in \mathbb{R}^q$，かつ，次の微分と積分の可換性が成り立つ：

$$\int_{\mathcal{X}} \delta(x) \psi_i(x, \theta) p_\theta(x) \mu(dx) \left(= \int_{\mathcal{X}} \delta(x) \frac{\partial p_\theta}{\partial \theta_i} \mu(dx) \right) = \frac{\partial}{\partial \theta_i} \int_{\mathcal{X}} \delta(x) p_\theta(x) \mu(dx) \qquad (i \in [q]).$$

このとき，関数 $g : \mathbb{R}^q \to \mathbb{R}^p$ の Jacobi 行列を $J(\theta) := \dfrac{\partial g}{\partial \theta} \in M_{p,q}(\mathbb{R})$ とすると，次の行列不等式が成り立つ：

$$\mathrm{Var}_\theta[\delta] \geqslant J(\theta) I(\theta)^{-1} J(\theta)^\top.$$

［**証明**］. 任意に $\theta \in U$ を取り，$I := I(\theta), J := J(\theta)$ と略記する．

共分散への翻訳　仮定 (A4) より，

$$J = \frac{\partial g(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} E_\theta[\delta] = E_\theta[\delta \psi^\top].$$

これと $E_\theta[\psi] = 0$ より，

$$\mathrm{Cov}_\theta[\psi, \delta] = E_\theta[\psi \delta^\top] = J^\top.$$

$$I = E_\theta[\psi \psi^\top] = \mathrm{Var}_\theta[\psi].$$

[†1] $\{x \in \mathcal{X} \mid p_\theta(x) = 0\}$ 上で $\psi_i(-, \theta)$ は定義されていないことに注意．

証明　すると，Cauchy-Schwartz の不等式同様，$\mathrm{Var}_\theta[\delta - JI^{-1}\psi] \geqslant O$ であることから，Cov の双線型性のみから従う．また，対称行列 S に対して，$S^{-1} = (S^{-1})^\top = (S^\top)^{-1}$ であることと Fisher 情報行列が対称であることに注意すると，

$$\begin{aligned} O &\leqslant \mathrm{Var}_\theta[\delta - JI^{-1}\psi] = \mathrm{Cov}[\delta - JI^{-1}\psi, \delta - JI^{-1}\psi] \\ &= \mathrm{Var}_\theta[\delta] - JI^{-1}\mathrm{Cov}_\theta[\psi, \delta] - \mathrm{Cov}_\theta[\delta, \psi]I^{-1}J^\top + JI^{-1}\mathrm{Var}[\psi]I^{-1}\psi^\top \\ &= \mathrm{Var}_\theta[\delta] - JI^{-1}J^\top. \end{aligned}$$

■

要諦 2.4.12. 本質的には Cauchy-Schwartz の不等式である．$\mathbb{R}^q$ の内積を $(-|-)$，$\mathcal{L}^2(P_\theta)$ の内積を $\langle -|- \rangle$ で表すと，$\|\langle \psi|\delta\rangle\|^2 \leqslant (\|\psi\|^2\|\delta\|^2)$.

2.4.5　ベイズ推定

ベイズ決定関数とミニマックス決定関数の関係を考える．

記法 2.4.13. 次の統計的決定問題を考える．

(1) 標本空間 $(\mathcal{X}, \mathcal{A})$ 上の確率分布族 $(P_\theta)_{\theta\in\Theta}$ について，Θ が可測であるとする．
(2) 損失関数 $W : \Theta \times \mathcal{D} \to \mathbb{R}_+$ は可測，危険関数 $R(-, \delta) : \Theta \to \mathbb{R}$ を可測とする．

定義 2.4.14 (prior distribution, Bayes risk, Bayes estimator, minimax decision function)**.**

(1) 先験的な確率分布 $\pi \in P(\Theta)$ を**事前分布**という．
(2) $R(\pi, \delta) := \int_\Theta R(\theta, \delta)\pi(d\theta)$ を，π に関する δ の**ベイズリスク**という．
(3) ベイズリスクを最小にする決定関数 δ を**ベイズ決定関数**という．これに対応するものが推定量であるとき，**ベイズ推定量**という．
(4) 一方で，リスク関数の上限 $\sup_{\theta\in\Theta} R(\theta, \delta)$ を最小にする決定関数を**ミニマックス決定関数**という．

2.4.6　非許容性

p 変量正規分布族 $(N_p(\theta, \sigma^2 I_p))_{\theta\in\mathbb{R}^p}$ の位置母数の推定量 $\bar{x} = n^{-1}\sum_{j=1}^n x_j$ は UMV 不偏推定量であるが，$p \geqslant 3$ のとき非許容である．

2.5　統計的仮説検定

2.5.1　仮説検定の枠組み

定義 2.5.1 (hypothesis testing, null hypothesis, alternative hypothesis, significance level, critical region)**.** 確率空間 $(\mathcal{X}, P)$ に対して，事前に与えられた設定 $\alpha \in (0, 1), H_0, H_1$ から定められる部分集合 $\mathcal{X}_1 \subset \mathcal{X}$ を検定という．

(1) 仮説検定において，仮説とは，母数または確率分布に関する（メタ）条件をいう．
(2) 疑いの目が向けられており，検定される仮説 H_0 を**帰無仮説**といい，それに対立する仮説 H_1 を**対立仮説**という．
(3) 文脈としては，H_0 を棄却し H_1 の成立を示すことで社会的な意味を見出そうという文脈に当てはまるように H_0, H_1 を選ぶ．
(4) 仮設 H_0 の下で計算した確率が「小さい」とはどういう意味かを形式化した指標を**有意水準**という．$\alpha = 0.05, 0.01$ などが用いられる．
(5) $\mathcal{X}_1 \subset \mathcal{X}$ を α と H_0, H_1 の意義から事前に定めて**棄却域**といい，観測値 x が $x \in \mathcal{X}_1$ のとき棄却し，$x \notin \mathcal{X}_1$ のとき採択する．$\mathcal{X}_1$ は，H_0 の下では非常に稀だが，H_1 の下ではそうでもないような観測値の集合である．

例 2.5.2. $\alpha = 0.05$ とする．20 回中事象 E が 13 回起こったとする．帰無仮説 H_0 を $P(E) = 1/2$ とし，対立仮説 H_1 を $P(E) > 1/2$ とする．すると，棄却域は $x_0 \in \mathbb{N}$ を用いて $\mathcal{C}_1 = \{x \in [20] \mid x \geqslant x_0\}$ と定めるのが良いと考えられる．いま，

$$\sum_{k=14}^{20} \binom{20}{k} \left(\frac{1}{2}\right)^{20} \approx 0.0577$$

より，$13 < x_0$ であって，観測値 13 は $\mathcal{C}_1$ には入らないから，棄却できない．「確率 0.05 程度の事象が偶々起こった」と解釈する方が選択される．

定義 2.5.3 (type I error, type II error)**.** H_0 が正しいのに H_0 を棄却してしまう誤りを**第一種の誤り**または**偽陽性**といい，H_1 が正しいのに H_0 を採択してしまう誤りを**第二種の誤り**または**偽陰性**という．

2.5.2　ランダム化検定

検定の構成法の例をあげる．検定＝棄却域の良さは，第一種の誤りの確率を有意水準 α で制限した後に，その範囲内で第二種の誤りの確率を最小化することで得られる．

例 2.5.4. 例 2.5.2 について，$x_0 = 15$ とすると，

$$\sum_{k=15}^{20} \binom{20}{k} \left(\frac{1}{2}\right)^{20} \approx 0.0207$$

より，棄却域のサイズは 5% を大きく切って 2% 程度となってしまう．そこで，$14 < x_0 < 15$ にあたる棄却域を構成したいが，$x_0 \in \mathbb{Z}$ の値は変えられない．そこで，**棄却域をランダム化**する．

$$\binom{20}{k} \left(\frac{1}{2}\right)^{20} \varphi + \sum_{k=15}^{20} \binom{20}{k} \left(\frac{1}{2}\right)^{20} = \alpha = 0.05$$

を満たす $\varphi \in (0,1)$ を用いて，観測値 $x = 14$ に対しては確率 φ で H_0 を棄却し，確率 $1-\varphi$ で H_0 を採択することとする．

2.5.3　形式化

定義 2.5.5 (test / critical (function), size, level-α-test, power function, uniformly most powerful (UMP) test)**.** 確率分布族 $\mathcal{P} = (P_\theta)_{\theta\in\Theta} \subset \mathcal{M}(\mathcal{X}, A)$ を考える．

(1) 棄却域 $\mathcal{C}_1 \subset \mathcal{X}$ の定めるパラメータ空間の分割 $\Theta = \Theta_0 + \Theta_1$ を考え，帰無仮説を $H_0 : \theta \in \Theta_0$ で表し，対立仮説を $H_1 : \theta \in \Theta_1$ と表す．

(2) $|\Theta_i| = 1$ の場合を単純仮説といい，そうでない場合を複合仮設という．[†2]

(3) 可測関数 $\varphi : \mathcal{X} \to [0,1]$ を**検定（関数）**という．$\exists_{A\in\mathcal{A}}\ \varphi = 1_A$ であるとき検定 φ は非確率的である．[†3]

(4) $\sup_{\theta\in\Theta_0} E_\theta[\varphi]$ を φ の**大きさ**という．大きさが α 以下の検定を**水準 α 検定**という．水準 α 検定の全体を $\Phi_\alpha \subset \mathrm{Meas}(\mathcal{X}, [0,1])$ で表す．

(5) $\beta_\varphi : \Theta_1 \to [0,1]; \theta \mapsto E_\theta[\varphi]$ を**検定出力関数**という．$1 - \beta_\varphi(\theta)$ は第二種の誤りの確率を表す．

(6) $\forall_{\varphi\in\Phi_\alpha}\ \forall_{\theta\in\Theta_1}\ \beta_{\varphi_0}(\theta) \geqslant \beta_\varphi(\theta)$ を満たす検定 $\varphi_0 \in \Phi_\alpha$ を**一様最強力検定**という．

要諦 2.5.6 (統計的決定理論の枠組みでの解釈)**.** 決定空間は，採択する仮説の添字からなる空間 $\mathcal{D} := 2 = \{0, 1\}$ である (0 が受容，1 が棄却)．損失関数は

$$W(\theta, a) = \begin{cases} 1_{\{1\}}(a), & \theta \in \Theta_0, \\ 1_{\{0\}}(a), & \theta \in \Theta_1. \end{cases}$$

で，検定関数 φ は確率的決定関数

$$\delta_\varphi(dz|x) = \varphi(x)\epsilon_1(dz) + (1 - \varphi(x))\epsilon_0(dz)$$

[†2] Θ が 1 次元で，$H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$ の形のときを**両側検定**，$H_0 : \theta \leqslant \theta_0, H_1 : \theta > \theta_0$ の形のときを**片側検定**という．

[†3] 単純に棄却域とそうでない域に分割するため．そうでない場合は，ランダム化検定の方法論を取り入れたことになる．

に対応する．ただし，ϵ_a はデルタ測度とした．したがって危険関数は

$$R(\theta,\delta_\varphi)=\int_{\mathcal{X}}\int_{\mathcal{D}} W(\theta,z)\delta_\varphi(dz|x)\epsilon_0(dz)$$
$$=\begin{cases}E_\theta[\varphi], & \theta\in\Theta_0,\\ 1-E_\theta[\varphi], & \theta\in\Theta_1.\end{cases}$$

となり，$\Phi_\alpha=\left\{\varphi\in\mathrm{Meas}(\mathcal{X},[0,1])\mid \sup_{\theta\in\Theta_0}R(\theta,\delta_\varphi)\leqslant\alpha\right\}$ と表せる．決定関数の全体は $\Delta\subset\Phi_\alpha$ となる．

Δ の中に入る一番自然な順序は一様順序で，これによる最大元が存在するときこれを UMP と呼ぶ．

定義 2.5.7 (nuisance parameter, goodness of fit test).

(1) Θ が 2 次元以上で，特定のパラメータにしか興味がないとき，分割 $\Theta_1+\Theta_2$ は商空間となる．この時の興味のない母数を**局外母数**または**撹乱母数**という．
(2) 帰無仮説を「確率変数が正規分布に従う」という命題にする場合，特に**分布の適合度検定**と呼ぶ．適合度検定や「確率分布が独立である」などの数学的な仮定が現実のデータと矛盾がないかチェックするための検定のクラスを**統計的モデルの診断**という．

2.5.4　Neyman-Pearson の補題

サイズ α の検定の中で，検出力 β が最も大きい検定の存在を保証し，尤度比検定によって達成されることを主張する補題．
「α を決めておき、その中で検出力が最も大きい検定法を選択する」という方針をネイマン・ピアソンの基準という。

2.5.5　単調尤度比と複合仮説の検定

2.5.6　一般化された Neyman-Pearson の補題

2.5.7　不偏検定

一様最強力検定は一般には存在しない．が，Neyman-Pearson の理論はある指数分布族モデルに対して一様最強力検定の構成を可能にする．

2.5.8　両側 t-検定

2.5.9　不変検定

2.6　区間推定

パラメータ $\theta\in\Theta$ に対して，区間 $S(x)$ を考えることを，区間推定という．

定義 2.6.1 (confidence coefficient, confidence region). 写像 $S:\mathcal{X}\to P(\Theta)$ は，$\alpha\in(0,1)$ を所与の定数として，次の 2 条件をみたすとする：

(1) $\forall_{\theta\in\Theta}\ \{x\in\mathcal{X}\mid\theta\in S(x)\}\in\mathcal{A}$.
(2) $\forall_{\theta\in\Theta}\ P_\theta[\theta\in S(x)]\geqslant 1-\alpha$

このとき，$1-\alpha$ を**信頼係数**，S を**信頼域**という．

第 3 章

大標本理論

非線形モデル・非ガウスモデルに対しては，統計量の分布が複雑なため，標準的な統計的決定理論を適用するのは困難であるが，これに替わって漸近的方法に基づく大標本理論が有力な解析手段となる．ここで，正規近似に基づく 1 次の漸近理論を扱う．

大標本理論の一番最初の応用先として汎関数の漸近分布を求めることがあるので，統計的検定問題を考える．漸近論が準備できたことにより，尤度比検定はより理解できるようになった．

これらの暗黙の枠組みは漸近決定理論であり，「統計的実験列の極限」というアイデアである．

(1) 一致性：大数の法則
(2) 一致性のオーダー：収束レート
(3) 漸近分布：中心極限定理とその精緻化

の 3 つの観点が肝要となる．

3.1 一致推定量

Ronald Fisher 1912-22 が開発．最尤推定量は一致性を満たすが，この現象はより広いクラスについて成り立つ．

3.1.1 最尤推定

統計推測においては，Bayes 法と最尤法とが双璧をなす．最尤推定量がごく自然な条件の下で一致性を持つのは，i.i.d. の場合は Wald，これを一般化して Huber が示した．また，正則な場合は有効な推定量にもなるが，これは「尤度最大」であることは関係がなく，「尤度方程式の根である」ことによる．そこで，Z-推定量なる対象に議論が移る．

問題 3.1.1. 可測空間 $(\mathcal{X}, \mathcal{A})$ 上の分布族 $(P_\theta)_{\theta\in\Theta}$ について，独立な観測値 $x_1, \cdots, x_n$ から θ を推定する問題を考える．

定義 3.1.2 (MLE: maximum likelihood estimator)**.** $(\mathcal{X}, \mathcal{A})$ 上の σ-有限測度 ν に関して P_θ が絶対連続であるとし，微分を $p(x,\theta) := dP_\theta/d\nu(x)$ と表す．

(1) $L_n(\theta) := \prod_{j=1}^{n} p(x_j, \theta)$ によって定まる関数 $\mathcal{X}^n \times \Theta \to \mathbb{R}$ を**尤度関数**という．これは，ある $\theta \in \Theta$ について，これが定める分布 P_θ から観測値 $x_1, \cdots, x_n$ が得られる条件付き確率に当たる．
(2) 各 $(x_1, \cdots, x_n) \in \mathcal{X}^n$ に対して，尤度関数を最大にする $\widetilde{\theta_n}(x_1, \cdots, x_n) \in \Theta$ を対応させる関数 $\widetilde{\theta_n} : \mathcal{X}^n \to \Theta$ を**最尤推定量**という．[†1]

注 3.1.3 (M-推定量としての最尤推定量)**.** 単調増加関数との合成を考えて，対数尤度関数 $l_n := \log \circ L_n$ の最大化を考えるにあたって，パラメータの空間 Θ が可微分構造を持つとき，ランダムな方程式 $\partial_\theta l_n(\theta) = 0$ の解としての最尤推定量を特徴付けることが出来る．これを**尤度方程式**という．このとき出現した対数尤度の微分を（標準化されていない）**スコア関数** $S(x,\theta) := \dfrac{\partial}{\partial\theta} l(\theta, x)$ と

[†1] ここでは推定量は可測という仮定は置かない．

いう．

$$L_n(\theta) := \prod_{j=1}^{n} p(x_j,\theta) \text{ の最大化} \Leftrightarrow l_n(\theta) := \log L_n(\theta) = \sum_{j=1}^{n} \log p(x_j,\theta) \text{ の最大化}$$
$$\Leftrightarrow S_n(x,\theta) := \frac{\partial}{\partial\theta} l_n(\theta) = \sum_{j=1}^{n} \frac{\partial}{\partial\theta} \log p(x_j,\theta) \text{ の零点を探す}$$

3.1.2　定義

要諦 3.1.4 (最尤推定量の一致性)**.** 大数尤度関数 $l_n(\theta) := \log L_n(\theta) = \sum_{j=1}^{n} \log p(x_j,\theta)$ は，独立同分布に従う確率変数の和である．よって，大数の法則により，真値 $\theta_0 \in \Theta$ について，

$$\forall_{\theta\in\Theta} \quad \frac{1}{n}[l_n(\theta) - l_n(\theta_0)] \to \int \log\left(\frac{p(x,\theta)}{p(x,\theta_0)}\right) P_{\theta_0}(dx) \quad P_{\theta_0}\text{- a.s.}$$

が成り立つ．$P : \Theta \to \{P_\theta\}$ が単射であるという識別可能性条件を仮定すると，相対エントロピーの非負性より，真値 θ_0 はこの右辺を最大にする唯一の点となる．そこで，左辺を最大にする推定量が $\widetilde{\theta_n}$ なのであったから，$n \to \infty$ のとき，$\widetilde{\theta}_n \to \theta_0$ が予想される．あとは，任意の $\theta \in \Theta$ に対する一様性を確認すれば良い．

補題 3.1.5. 可測空間 $(\Omega, \mathcal{F})$ 上の σ-有限測度 ν に関して絶対連続な確率測度 P, Q について，その微分を $p(\omega) := dP/d\nu(\omega), q(\omega) := dQ/d\nu(\omega)$ とおく．このとき，

$$\int_\Omega \log\left(\frac{p(\omega)}{q(\omega)}\right) P(d\omega) \leqslant 0$$

で，等号成立条件は $P = Q$ のとき．

定義 3.1.6 (consistency)**.** パラメータの推定量 $\theta_n : \mathcal{X}^n \to \Theta$ について，

(1) 任意の $\theta \in \Theta$ について，これが真値であったとき，$\widetilde{\theta_n} \to \theta$ a.s. $(n \to \infty)$ が成り立つとき，この推定量は**強一致性**を持つという．
(2) 推定量 $\widetilde{\theta_n}$ が可測で，任意の $\theta \in \Theta$ について，これが真値であったとき，$\widetilde{\theta_n} \xrightarrow{p} \theta$ $(n \to \infty)$ が成り立つとき，すなわち，$\widetilde{\theta}_n - \theta = o_p(1)$ が成り立つとき，この推定量は**弱一致性**を持つという．

たったこれだけの議論で，相対エントロピーなる確率測度間の距離概念と，実数の乗法群と加法群の間の準同型 $\mathbb{R}_{>0} \to \mathbb{R}$ としての対数関数とのいずれもが出てくるとは思わなかった．「尤度」なる意味論と「相対エントロピー」という意味論とが繋がることが一番の驚き．

補題 3.1.7 (確率収束の消息を統計の言葉で表現)**.**

(1) U_n が θ にある $r > 0$ について r-次平均収束するならば，弱一致推定量である．
(2) 特に，U_n が θ の不偏推定量かつ $\mathrm{Var}[U_n] \to 0$ が成り立つならば，θ の弱一致推定量である．

例 3.1.8 (一致推定量)**.**

(1) $U_n(X) := \overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ は母集団平均の強一致推定量である（大数の強法則による）．
(2) 全く同様のことは，一般の標本モーメント $\widehat{m}_r := \dfrac{1}{n}\sum_{j=1}^{n} X_j^r$ が強一致推定量であるということが言える．
(3) 一般の標本中心化モーメントも強一致推定量であるが，そのままでは不偏推定量にはならない．
(4) セミパラメトリックな線型回帰モデル $Y_j = \alpha + \beta X_j + \epsilon_j$ $(j \in [n])$ の OLS(ordinary least squares estimator) はいくらかの条件の下一致推定量である．
(5) 自己回帰モデル (AR:Autoregressive Model)$Y_j = \alpha + \beta Y_{j-1} + \epsilon_j$ の最小二乗推定量

$$\widehat{\alpha} = \overline{Y}_{1,n} - \widehat{\beta}\overline{Y}_{0,n-1}, \quad \widehat{\beta} = \beta + \frac{\sum_{r\in[n]} \epsilon_r (Y_{r-1} - \overline{Y}_{0,n-1})}{\sum_{r\in[n]} (Y_{r-1} - \overline{Y}_{0,n-1})^2}$$

は一致推定量である．ただし，$\overline{Y}_{i,j}$ を Y_i から Y_j までのデータを使った標本平均とした．

3.1.3　存在の必要条件

極限実験 $\{P_\theta\}$ において，δ_n が離散位相を定めていれば良い．

記法 3.1.9. $(\mathcal{X}_n, \mathcal{A}_n, \{P_{n,\theta}\}_{\theta\in\Omega})$ $(\Omega \overset{\text{open}}{\subset} \mathbb{R}^p)$ を統計的実験の列とする．

定義 3.1.10. 弱一致推定量 $(\widehat{\theta}_n)$ が**局所一様**であるとは，

$$\forall_{\theta\in\Omega}\ \forall_{\epsilon>0}\ \exists_{\delta>0}\ \lim_{n\to\infty} \sup_{\|\theta'-\theta\|<\delta} P_{n,\theta'}[\|\widehat{\theta_n} - \theta'\| > \epsilon] = 0$$

が成り立つことをいう．

命題 3.1.11.

(1) $\theta_1, \theta_2 \in \Omega$ に対して，$P_{n,\theta_1}, P_{n,\theta_2}$ を優越する σ-有限測度 μ_n が取れる．
(2) $L^1(\mu_n)$-ノルムが，$\{P_{n,\theta}\}$ 上に距離 δ_n を定める．
(3) この距離は全変動ノルムの 2 倍 $\delta_n(\theta_1,\theta_2) = 2\sup_{A\in\mathcal{A}_n} |P_{n,\theta_1}(A) - P_{n,\theta_2}(A)|$ と特徴付けられる．特に，$\mathrm{Im}\ \delta_n \subset [0,2]$.

定理 3.1.12. 統計的実験列 $(\mathcal{X}_n, \mathcal{A}_n, \{P_{n,\theta}\}_{\theta\in\Omega})$ $(\Omega \overset{\text{open}}{\subset} \mathbb{R}^p)$ について，

(1) θ の一致推定量が存在するならば，$\theta_1 \neq \theta_2 \Rightarrow \lim_{n\to\infty} \delta_n(\theta_1,\theta_2) = 2$.
(2) 局所一様一致推定量が存在するならば，$\forall_{\epsilon>0}\ \delta_{\delta>0}\ \forall_{\theta\in\Omega}\ \lim_{n\to\infty} \sup_{\epsilon<\|\theta'-\theta\|<\delta} \delta_n(\theta,\theta') = 2$.

［証明］.

(1)

■

3.2　一様大数の法則

真値 $\theta_0 \in \Theta$ はわからないため，あらゆる $\theta \in \Theta$ について一様な一致性を持ってほしい．最尤推定量はこの性質を持つ．

記法 3.2.1.

(1) $(\mathcal{X}, \mathcal{A}, P)$ の可算無限直積 $(\overline{\mathcal{X}}, \overline{\mathcal{A}}, \overline{P})$ 上に，$\pi_n := (\mathrm{pr}_1, \cdots, \mathrm{pr}_n) : \overline{\mathcal{X}} \to \mathcal{X}^n$ で $\mathcal{X}^n$ 上の可測関数を引き戻して議論する．
(2) $U : \mathcal{X}\times\Theta \to (-\infty,\infty]$ に対して，

$$\overline{U_n}(x,\theta) := \mathbb{P}_n U(x,\theta) = \frac{1}{n}\sum_{j=1}^n U(x_j,\theta)$$

と表す．ただし，$x = (x_1, \cdots, x_n) \in \mathcal{X}^n$ とした．

補題 3.2.2. (Θ, d) を距離空間，$\Theta_1 \subset \Theta$ をコンパクト集合，$U : \mathcal{X}\times\Theta \to [-\infty,\infty]$ を関数とし，以下の条件を仮定する：

(A1) $\forall_{x\in\mathcal{X}}\ \theta \mapsto U(x,\theta)$ は下半連続．
(A2) $\forall_{\theta\in\Theta}\ r$ が十分に小さければ，$x \mapsto \inf_{\theta'\in\Theta, d(\theta',\theta)<r} U(x,\theta')$ は可測．
(A3) ある可積分関数 $M : \mathcal{X} \to \mathbb{R}$ が存在して，$\forall_{(x,\theta)\in\mathcal{X}\times\Theta}\ U(x,\theta) \geqslant M(x)$.

このとき，$\overline{U}(\theta) := \int_{\mathcal{X}} U(x,\theta)P(dx) \in (-\infty,\infty]$ に対して，

(1)

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta_1} \overline{U_n}(\theta) \geqslant \inf_{\theta\in\Theta_1} \overline{U}(\theta) \quad \text{a.s.}$$

(2) $\overline{U}$ は下半連続.

定理 3.2.3 (一様大数の法則). (Θ, d) を距離空間，$\Theta_1 \subset \Theta$ をコンパクト集合，$U : \mathcal{X} \times \Theta \to [-\infty, \infty]$ を関数とし，以下の条件を仮定する：

(A1) $\forall_{x\in\mathcal{X}}\ U(x,-) : \Theta \to \mathbb{R}$ は連続.
(A2) $\forall_{\theta\in\Theta}\ U(-,\theta) : \mathcal{X} \to \mathbb{R}$ は可測.
(A3) ある可積分関数 $M \in \mathcal{L}^1(\mathcal{X})$ が存在して，$\forall_{(x,\theta)\in\mathcal{X}\times\Theta}\ |U(x,\theta)| \leqslant M(x)$.

このとき，$\overline{U}(\theta) := \int_{\mathcal{X}} U(x,\theta)P(dx) \in \mathbb{R}$ は連続で，

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_1} |\overline{U_n}(\theta) - \overline{U}(\theta)| = 0 \quad \text{a.s.}$$

3.3　一致推定量の収束の速度

定義 3.3.1. 一致推定量 $(\widehat{\theta}_n)$ が，無限大に発散する正数列 $\{c_n\} \subset l_\infty(\mathbb{N}; \mathbb{R}_+)$ に関して**オーダー** (c_n) **を持つ**とは，$c_n(\widehat{\theta}_n - \theta)$ が極限分布を持つこと，すなわち裾が無限大へ逃げて分布が潰れてしまうことがないことをいう：

$$\forall_{\theta_0\in\Omega}\ \exists_{\delta>0}\ \lim_{L\to\infty} \limsup_{n\to\infty} \sup_{\|\theta-\theta_0\|<\delta} P_{n,\theta}[c_n\|\widehat{\theta}_n - \theta\| > L] = 0.$$

3.4　最小コントラスト推定

パラメトリック法内での一般化

最尤推定の枠組みを一般化して，まずは一様大数の法則を準備した．最小コントラスト推定なる一般的な枠組みを用意して，一様大数の法則と一致性との間の十分条件を示す．U, U_n に関わる仮定は，ここまで削ぎ落とすことが出来る．基本的に，最尤推定のときのように，尤度関数 Φ_n を変形して $\overline{U}_n$ を作る．

3.4.1　定義と例

最尤推定量は，真の分布との対数尤度の差，すなわち KL-分離度を最小化するものとして得られた．経験分布の真の分布との KL-分離度は，$n \in \mathbb{N}$ に対応して関数を与える．これを $n \to \infty$ の極限で漸近的に最小にするクラスを最小コントラスト推定量という．

定義 3.4.1 (contrast function, minimum contrast estimator). (Θ, d) を距離空間，$(\overline{\mathcal{X}}, \overline{\mathcal{A}}, \overline{P})$ を確率空間とする.

(1) **コントラスト関数**とは，可積分関数 $\overline{\Psi} : \overline{\mathcal{X}} \times \Theta \to (-\infty, \infty]$ であって，次を満たすものをいう：

$$\forall_{\theta\in\Theta} \quad E_{\theta_0}[\overline{\Psi}(x,\theta)] \geqslant E_{\theta_0}[\overline{\Psi}(x,\theta_0)].$$

(2) これを最小にする写像 $\widetilde{\theta} : \overline{\mathcal{X}} \to \Theta$ を**最小コントラスト推定量**という.

以降，コントラスト関数の列 $(\overline{\Psi}_n)$ を漸近的に最小にする列 $(\widetilde{\theta}_n)$ を考える.

例 3.4.2 (最尤推定). 最尤推定量は，第一義的には，$\overline{\Psi}_n(\theta) := -\mathbb{P}_n \log p(x,\theta)$ を最小化した.
が，一致性の議論の中で，真の分布との対数尤度の差 $|l_n(\theta) - l_n(\widetilde{\theta})|$ の最小化を考えた．これは損失関数のようでより直感的である．すなわち，

$$\overline{U}_n(\theta) = \mathbb{P}_n \left(\frac{\log p(x,\theta_0)}{\log p(x,\theta)} \right)$$

と取ることも出来る．この考え方は，Z-推定量では出来ない，最小コントラスト推定量独自の議論である．

この大数の法則と定理とを併せた議論は完全に 3.1.4 の議論の厳密化となっている．

3.4.2　強一致性を持つための十分条件

定理 3.4.3 (最小コントラスト推定量が強一致性を持つための十分条件)**.** 関数列 $\overline{U}_n$，関数 $\overline{U}:\Theta\to(-\infty,\infty]$，空でない部分集合 $\Theta'\subset\Theta$，写像列 $\widehat{\theta}_n:\overline{\mathfrak{X}}\to\Theta$ は次の 3 条件を満たすとする．

(C1) 一様性：$\forall_{\epsilon>0}\ \liminf_{n\to\infty}\inf_{\theta\in\Theta:d(\theta,\Theta')\geqslant\epsilon}\overline{U}_n(\theta)\geqslant\inf_{\theta\in\Theta:d(\theta,\Theta')\geqslant\epsilon}\overline{U}(\theta)$ a.s.

(C2) 識別可能性：$\forall_{\epsilon>0}\ \inf_{\theta\in\Theta:d(\theta,\Theta')\geqslant\epsilon}\overline{U}(\theta)>\inf_{\theta\in\Theta'}\overline{U}(\theta)$.

(C3) 最小コントラスト推定量である：$\limsup_{n\to\infty}\left[\overline{U}_n(\widehat{\theta}_n)-\inf_{\theta\in\Theta'}\overline{U}_n(\theta)\right]\leqslant 0$ a.s.

このとき，一様大数の法則が成り立つ：$\limsup_{n\to\infty}\inf_{\theta\in\Theta'}\overline{U}_n(\theta)\leqslant\inf_{\theta\in\Theta'}\overline{U}(\theta)$ a.s. ならば，$\lim_{n\to\infty}d(\widehat{\theta}_n,\Theta')=0$ a.s.

3.4.3　一致性を持つ推定量の例

> 続いて，この枠組みを用いてより一般的なクラスについて一致性を得よう．この枠組みの強みは，最小コントラスト推定量が陽に（解析的に）計算できない場合でも，一致性を示すことが出来る点である．

例 3.4.4 (多項分布)**.** k 個の事象への分割 $\Omega=\sum_{i=1}^{k}E_i$ を考え，$P_\theta[E_i]=:\theta^i$ を k 個のパラメータとする．この試行を n 回繰り返し，E_i が観測された回数を y_i とすると，これは多項分布に従う確率変数である：$(y_1,\cdots,y_k)\sim M(n;\theta^1,\cdots,\theta^k)$．パラメータ空間を

$$\Theta:=\left\{(\theta^1,\cdots,\theta^k)\in[0,1]^k\mid\sum_{i=1}^{k}\theta^i=1\right\}$$

とする．コントラスト関数を

$$\overline{\Psi}_n(\theta):=-\sum_{i=1}^{k}\frac{y_i}{n}\log\theta^i$$

と定めると，これに対応する最小コントラスト推定量とは最尤推定量に他ならない．これに対して，関数を次のように定めれば，定理が用いることが出来て，一致性がわかる：

$$\overline{U}(\theta):=-\sum_{i=1}^{k}\theta_0^i[\log\theta^i-\log\theta_0^i]$$

$$\overline{U}_n(\theta):=-\sum_{i=1}^{k}\frac{y_i}{n}[\log\theta^i-\log\theta_0^i]$$

今回は実は最尤推定量 $\widehat{\theta_n}=(y_1/n,\cdots,y_k/n)$ は陽に計算できるが，一般に尤度方程式も，コントラスト関数の最小点も求めることが出来るとは限らない．この枠組みは，そのような場合でも通用するように出来ている．

3.4.4　独立観測の構造を仮定した場合

3.4.5　可測な最小コントラスト推定量の存在

3.5　M-推定量

最尤法のノンパラメトリック化

最尤推定量は対数尤度関数を最大化する推定量で，最小コントラスト推定量は真の分布との分離度を最小化する推定量であった．このように，損失関数または効用関数（2 つの概念を併せて目的関数）の極値点として与えられる推定量は，モデルが微分可能であるとき，これらは導関数の零点として得られる推定量として一般化できる．これを M-推定量という．[a] この一般化は，分布関数 f への言及を除去しており，ノンパラメトリック化と言える．

Huber が頑健推定の動機で 1964 年に最尤推定を一般化し，"maximum likelihood-type"の頭文字から M-推定量のクラスを定めた．[b] 実はこれは可微分構造を持つ正則なモデルに関して最適でもある．このセミパラメトリックな一般化を経験尤度法 (Owen 1988) という．

[a] この意味での推定量を Z-推定量とし，M-推定量は効用関数の最大化で得るものと狭義に解釈することもある．
[b] https://en.wikipedia.org/wiki/M-estimator

3.5.1　定義と例

M-推定量はロバスト推定の代表である．

定義 3.5.1 (estimating function / objective function)**.** $\Theta \subset \mathbb{R}^p$ とする．予め設定した関数 $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^p$ に対して，$x_1, \cdots, x_n$ が定める経験分布についての平均を

$$\overline{\psi_n}(x, \theta) := \mathbb{P}_n \psi_n(x, \theta) = \frac{1}{n} \sum_{j=1}^{n} \psi(x_j, \theta)$$

とおき，推定方程式 $\overline{\psi_n}(\theta) = 0$ の解として構成される推定量 $\widetilde{\theta_n}$ を **M-推定量** (または **Z-推定量**) という．ψ またはその標本平均 $\overline{\psi_n}$ を**目的関数**または**推定関数**という．

注 3.5.2. 狭義には，目的関数 $\psi(x, \theta)$ の標本平均 $\mathbb{P}\psi(x, \theta)$ を最大化する推定量 $\theta := \arg\max_{\theta \in \Theta} \mathbb{P}\psi(x, \theta)$ を M-推定量という．

例 3.5.3.

(1) 最尤推定量は，スコア関数を推定関数とする M-推定量であり，尤度方程式 $S_n(x, \theta) = \partial_\theta l(\theta, x) = 0$ の零点として特徴付けられる．

(2) 最小コントラスト推定量は，$\partial_\theta \Psi_n(\theta) = 0$ なる推定関数に対応する M-推定量である．

(3) 非線形最小二乗法は，最小二乗 $\rho(u) = u^2/2$ を損失関数として，これを最小化する M-推定量である．このときの推定関数は，この損失関数の微分なので，$\psi(x) = x$ である．なお，この ψ の代わりに，剪断する Huber の ψ（が定める損失関数）を用いたものを，頑健回帰という：

$$\psi(u_i) := \begin{cases} u_i, & |u_i| \leqslant H, \\ H, & u_i > H, \\ -H, & u_i < H. \end{cases}$$

このときの H を**調整定数** (tuning constant) という．

3.5.2　影響関数

Huber and Ronchetti (2009)[10] p.47. Hampel et al. (1986) p.85[9] M-推定量の影響関数は，推定関数 ψ の定数倍になる．特に最尤推定量について言えば，スコア関数の定数倍となる．また，$\mathrm{Var}[\mathrm{IF}^2]$ が漸近分散となる．最尤推定量について，これは Fisher 情報量である．

定理 3.5.4.

$$\mathrm{IF}(x;P,\theta) =$$

3.6　M-推定量の漸近正規性

一致性は大数の法則に当たるならば，その真値への収束の速度が同様に気になる．これが漸近正規性である．これを，M-推定量について議論する．

3.6.1　一致性を持つ M-推定量の漸近分布

記法 3.6.1. $\Theta \subset \mathbb{R}^p$ を可測集合，$\psi : \mathcal{X} \times \Theta \to \mathbb{R}^p$ を推定関数とし，$\psi(x,-)$ は連続とする．$\psi_n(\theta) := \sum_{j=1}^{n} \psi(x_j,\theta)$，$\overline{\psi}_n(\theta) := \mathbb{E}_n[\psi(\theta)] = \dfrac{1}{n}\psi_n(\theta)$ と定め，真の平均を $\overline{\psi}(\theta) := \displaystyle\int_{\mathcal{X}} \psi(x,\theta)P(dx)$ と定める．

定理 3.6.2 (漸近正規性の十分条件)**.** $\theta_0 \in \Theta^\circ$ とし，$B \subset \Theta$ を θ_0 の開近傍とする．次の 5 条件を仮定する．

(E1) 推定関数 $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^p$ は B 上 C^1 級.
(E2) $\psi(-,\theta)$ は可測.
(E3) $\psi(x,\theta_0) \in \mathcal{L}^2(P;\mathbb{R}^p)$ で，$P\psi(x,\theta_0) = 0$.
(E4) $\exists_{M \in \mathcal{L}^1(\mathcal{X},\mathbb{R})}\ \forall_{x\in\mathcal{X},\theta\in B}\ |\partial_\theta \psi(x,\theta)| \leqslant M(x)$．さらに，$\Gamma := -P\partial_\theta\psi(x,\theta_0)$ は正則.
(E5) 確率変数列 $\widetilde{\theta}_n : \mathcal{X}^n \to \Theta$ が $\widehat{\theta}_n \xrightarrow{p} \theta_0$ かつ $\psi_n(\widehat{\theta}_n) = o_p(\sqrt{n})$ を満足する.

このとき，

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) - \Gamma^{-1}\frac{1}{\sqrt{n}}\psi_n(\theta_0) = o_p(1) \quad (n \to \infty).$$

特に，$\Sigma := \Gamma^{-1}\Phi(\Gamma^\top)^{-1}$，$\Phi := \displaystyle\int_{\mathcal{X}} \phi\phi^\top(x,\theta_0)P(dx)$ について，

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N_p(0,\Sigma) \quad (n \to \infty).$$

3.6.2　収束レート

Θ 上の関数 $\theta \mapsto P(\Psi_\theta - \Psi_{\theta_0})$ と $\theta \mapsto E^*[|\mathbb{G}_n(\Psi_\theta - \Psi_{\theta_0})|]$ の θ_0 近くの Holder 指数によって，$d(\widehat{\theta}_n,\theta_0)$ の収束レートが定まる．

記法 3.6.3. コントラスト関数 $\Psi_\theta(x)$ について，目的関数 $\theta \mapsto \mathbb{P}_n\Psi_\theta(x)$ を（近似的に）最大化する推定量列を $(\widehat{\theta}_n)$ とする．

定理 3.6.4 (rate of convergence)**.** ある $C \in \mathbb{R}, \alpha > \beta \in \mathbb{R}$ と任意の $n \in \mathbb{N}$ と十分小さい任意の $\delta > 0$ について，次が成り立つとする：

$$\sup_{d(\theta,\theta_0)<\delta} P(\Psi_\theta - \Psi_{\theta_0}) \leqslant -C\delta^\alpha, \qquad E^*\left[\sup_{d(\theta,\theta_0)<\delta} |\mathbb{G}_n(\Psi_\theta - \Psi_{\theta_0})|\right] \leqslant C\delta^\beta.$$

$$\mathbb{P}_n \Psi_{\widehat{\theta}_n} \geqslant \mathbb{P}_n \Psi_{\theta_0} - O_P\left(n^{\frac{\alpha}{2\beta-2\alpha}}\right), \qquad \widehat{\theta}_n \xrightarrow{P*} \theta_0$$

このとき，$n^{\frac{1}{2\alpha-2\beta}} d(\widehat{\theta}_n, \theta_0) = O_P^*(1)$.

3.6.3　最尤推定量での例

$\psi(x,\theta) := \partial_\theta l(x;\theta)$ を推定関数とし，この零点を最尤推定量という．

定理 3.6.5 (最尤推定量の漸近正規性の十分条件).

3.6.4　モーメント法での例

良い条件の下でモーメント推定量は $\sqrt{n}$-一致性と漸近正規性を持つが，最尤推定量のものより漸近分散は大きい．これはデルタ法の結果である．

記法 3.6.6. $\{P_\theta\}_{\theta\in\Theta} \subset P(\mathcal{X})$ を分布族，$g : \mathcal{X} \to \mathbb{R}^p$ を所与の関数とする．$\psi(x,\theta) := g(x) - \int_{\mathcal{X}} g(y) P_\theta(dy)$ とすると，真値 $\theta_0 \in \Theta$ について $E[\psi(x,\theta_0)] = 0$ である．このとき，次の識別可能条件がなりたつとする

$$\theta_1 \neq \theta_2 \Rightarrow \int_{\mathcal{X}} g(y) P_{\theta_1}(dy) \neq \int_{\mathcal{X}} g(y) P_{\theta_2}(dy).$$

このとき，$\psi(x,\theta)$ の P_{θ_0}-平均 $E_{\theta_0}[\psi(x,\theta)]$ の零点は θ_0 に限る．たとえば $g(x) = x^n$ と取ると，この条件を満たす．

定義 3.6.7 (method of moments). この ψ を推定関数とした M-推定量 $\widehat{\theta}_n^\dagger$，すなわち次の θ についての方程式の解を g **に関するモーメント推定量**という：

$$\overline{\psi}_n(x,\theta) := \mathbb{P}_n[\psi(x,\theta)] = \frac{1}{n}\sum_{j=1}^n \psi(x_j,\theta) = 0 \quad \Leftrightarrow \quad \mathbb{P}_n g(x) = P_\theta g(x)$$

あるいは，あるノルム $\|-\|$ について，$\|\overline{\psi}_n(x,\theta)\|$ を最小にする $\widehat{\theta}$ を探す (これを近似的モーメント推定量 (approximate moment estimator) という)．このとき，推定関数 ψ を**モーメント関数**という．

定理 3.6.8. 次の条件を仮定する．

(M1) $e(\theta) : \theta \mapsto P_\theta f \in \mathbb{R}^k$ は $\Theta \overset{\text{open}}{\subset} \mathbb{R}^k$ 上の単射で，θ_0 上で連続微分可能で，非特異な Jacobi 行列 $e'_{\theta_0} \in \mathrm{GL}_k(\mathbb{R})$ を持つとする．
(M2) $P_{\theta_0}\|f\|^2 < \infty$.

このとき，モーメント推定量 $\widehat{\theta}_n$ は任意に 1 に近い確率で存在し，次を満たす：

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\theta_0} N\left(0, e'^{-1}_{\theta_0} P_{\theta_0} f f^\top (e'^{-1}_{\theta_0})^\top\right).$$

3.6.5　一般化モーメント法

Lars Hansen は Pearson 1894 によるモーメント法を拡張してファイナンスに応用し，2013 年のノーベル経済学賞を受賞した．モーメントに限らず，真の分布を特徴付ける別の母数 ψ で計算しやすいものがあったら，これに変換することで M-推定量を構成する．最尤法は一般化モーメント法の特殊な場合とみなせる．

3.7　LeCam のワンステップ更新による漸近有効推定量の構成

一致推定量さえ見つかれば，漸近分散は逆 Fisher 情報行列まで改善できる標準的算譜がある

M-推定量は推定関数の零点として定まるが，実際に方程式を解くことは容易でない場合も多い．また推定方程式の解が一致性を持つためには，推定関数が Θ の大域的性質を満たす必要がる．
一方で，有効でないが，一致性をもつ推定量は容易に見つかることがしばしばある．このとき，それを用いて，最尤推定量と同じ漸近分散を持つ推定量を構成できる．
推定量の漸近有効性は，数値解析のような Newton-Rhapson 反復法を採用すれば良いことは初等的に思いつくが，実はこの手続きは 1 回で Z-推定量 (最尤推定量) と同等にまで改善される．[a]

[a] https://www.ma.imperial.ac.uk/~das01/MyWeb/M3S3/Handouts/OneStep.pdf

定義 3.7.1 (one-step estimator)**.** Θ-値確率変数列 $(\widehat{\theta}_n^0)$ から，新たな推定量の系列 $(\widehat{\theta}_n)$ を

$$\widehat{\theta}_n := \widehat{\theta}_n^0 - (\partial_\theta \psi_n(\widehat{\theta}_n^0))^{-1} \psi_n(\widehat{\theta}_n^0)$$

で定めると，これは

$$\mathcal{X}_n^* := \left\{ x \in \mathcal{X}^n \mid \widehat{\theta}_n^0 \text{において} \psi_n \text{は微分可能で } \partial_\theta \psi_n(\widehat{\theta}_n^0) \text{ は正則で} \widehat{\theta}_n \in \Theta \right\}$$

上で定まるから，これを $\mathcal{X}$ 上に可測に延長する．延長の仕方は任意で良い．こうして得た $(\widehat{\theta}_n)$ を $(\widehat{\theta}_n^0)$ **を初期推定量とするワンステップ推定量**という．

要諦 3.7.2. 実は，推定関数の標本平均 $\overline{\psi}_n$ が，ある正則行列 $\dot{\psi}_0 \in \mathbb{R}^p$ について次を満たすとき，

$$\sup_{\sqrt{n}\|\theta-\theta_0\|<M} \left\| \sqrt{n}(\overline{\psi}_n(\theta) - \overline{\psi}_n(\theta_0)) - \dot{\psi}_0 \sqrt{n}(\theta - \theta_0) \right\| \xrightarrow{p} 0.$$

係数 $(\partial_\theta \psi_n(\widehat{\theta}_n^0))^{-1}$ は任意の推定量で良い．このとき，任意のランダムな行列 $\dot{\psi}_{n,0}$ はある $\dot{\psi}_0$ に収束するならば，

$$\widehat{\theta_n} := \widehat{\theta}_n^0 - \dot{\psi}_{n,0}^{-1} \overline{\psi_n}(\widehat{\theta}_n^0)$$

で定まる推定量も**ワンステップ推定量**といい，次の定理を満たす．

定理 3.7.3 (ワンステップ推定量が漸近正規であるための十分条件)**.** $\theta_0 \in \Theta^\circ$，$B \subset \Theta$ を θ_0 の開近傍とする．条件 (E1)〜(E4) を仮定し，次の条件を仮定する：

(E6) 初期推定量の列 $\widehat{\theta}_n^0 : \mathcal{X}^n \to \Theta$ が $\sqrt{n}$-一致性を持つ：$\widehat{\theta}_n^0 - \theta_0 = O_p(1/\sqrt{n})$.[†2]

このとき，ワンステップ推定量 $\widehat{\theta}_n$ に対して，定理 3.6.2 と同様の結果が成り立つ：

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\Gamma^{-1} \frac{1}{\sqrt{n}} \psi_n(\theta_0) + o_P(1).$$

特に，$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N_p(0, \Sigma)$.

3.8　Cramer の一致推定量の構成

一致推定量の漸近有効性を改善する算譜を得たので，次は一致推定量の構成算譜を考える．
これにより，Θ 全域で L^2-微分可能で非退化な Fisher 情報行列を持つ識別可能なパラメトリックモデル $(P_\theta)_{\theta\in\Theta}$ には正則で漸近最適な θ の統計量が存在することが示せる．

[†2] 条件 $\widehat{\theta}_n^0 - \theta_0 = O_p(1/\sqrt{n})$ は，弱一致性 $\widehat{\theta}_n^0 - \theta_0 = o_p(1)$ を含意することに注意．これに加えて，$n^{-1/2}$ 倍の範囲で真値 θ_0 を捕まえていることは要求する．

3.9　頑健推定

1964 年に Huber によって提唱された.

3.9.1　影響関数

影響関数は，有界線型作用素 $T : P(\mathcal{X}) \to \Theta$ の特別な微分とみなせる.

記法 3.9.1. 推定関数 $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^p$ ($\Theta \overset{\text{open}}{\subset} \mathbb{R}^p$) に対して，汎関数 $T : P(\mathcal{X}) \to \Theta$ は推定方程式 $\forall_{P \in \mathcal{P}} \int_{\mathcal{X}} \psi(x, T(P)) P(dx) = 0$ を満足すると仮定する．すると経験分布 $\mathbb{P}_n$ について，$\widetilde{\theta_n} := T(\mathbb{P}_n)$ は推定関数 ψ に関する M-推定量である．このようにして，経験分布の汎関数として得られる推定量のクラスを考える.

推定関数は θ について可微分とし，

$$\Gamma(P) := -\int_{\mathcal{X}} \partial_\theta \psi(y, T(P)) P(dy) \quad \in M_p(\mathbb{R})$$

とする.

注 3.9.2 (汎関数 $T : \mathcal{P} \to \mathbb{R}$ の描像は？). $\mathcal{P}$ の極点の凸包上での値は M-推定量を与え，極点の凸包は $\mathcal{P}$ 上弱稠密である．したがって $T(P)$ の分布は一意に定まる (?)

定義 3.9.3.

(1) 確率測度 $P \in P(\mathcal{X})$ に対して，$x \in \mathcal{X}$ が定めるデルタ測度との $\epsilon \in [0,1]$ が定める凸結合を $P_\epsilon^x := (1-\epsilon)P + \epsilon\delta_x \in P(\mathcal{X})$ $(x \in \mathcal{X}, \epsilon \in [0,1])$ と定める.

(2) T について，$P \in P(\mathcal{X})$ における $\epsilon \in [0,1]$ に関する微分係数 $\mathrm{IF}(x; T, P) := \dfrac{d}{d\epsilon} T(P_\epsilon^x)|_{\epsilon=0} (= \Gamma(P)^{-1} \psi(x, T(P)))$ を関数 $\mathcal{X} \to \Theta$ とみたとき，これを T の P における**影響関数**という．これは，各点 $x \in \mathcal{X}$ に観測を得た場合に，M-推定量を与える写像 T がどれほど影響を受けるかを表す指標とみなせる.

補題 3.9.4 (M-推定量の影響関数). T が前述条件を満たす M 推定量を与える汎関数であり，$\forall_{(x,\theta) \in \mathcal{X} \times \Theta} |\partial_\theta \psi|$ は可積分であるとする.

$$\mathrm{IF}(x; T, P) = \Gamma(P)^{-1} \psi(x, T(P)), \qquad \Gamma(P) := -\int_{\mathcal{X}} \partial_\theta \psi(y, T(P)) P(dy)$$

特に，$p = 1$ のとき，推定関数 $\psi(x, T(P)) \in \mathbb{R}^p$ の定数倍である.

［証明］. 推定方程式より

$$0 = \int_{\mathcal{X}} \psi(x, T(P + \epsilon\delta_x))(P + \epsilon\delta_x)(dx) = \int_{\mathcal{X}} \psi(x, T(P + \epsilon\delta_x)) P(dx) + \epsilon \psi(x, T(P + \epsilon\delta_x))$$

が成り立つ．この両辺を ϵ で微分して $\epsilon = 0$ を考えると，関数 $\forall_{(x,\theta) \in \mathcal{X} \times \Theta} |\partial_\theta \psi|$ は可積分であるとすると，微分と積分とは交換可能であるから，

$$0 = \left(\int_{\mathcal{X}} \partial_\theta \psi(x, T(P)) P(dx) \right) \cdot \frac{d}{d\epsilon} T(P + \epsilon\delta_x) \Big|_{\epsilon=0} + \psi(x, T(P)).$$

■

例 3.9.5. 正規母集団 $N(\theta, 1)$ の位置母数 θ に対する最尤推定量 $\widetilde{\theta_n}$ の影響関数は $\mathrm{IF}(x; T, N(\theta, 1)) = x - \theta$ となる．したがって，$x \in \mathbb{R}$ の絶対値が大きいほど，最尤推定量への影響が大きいため，ここに異常値が来ると，大きく推論に影響する危険がある.

3.9.2　バイアスロバスト推定量

頑健統計では，外れ値による影響がなるべく小さい M-推定量を構成することを考える．M-推定量の影響関数には推定関数 ψ が明示的に出現するため，頑健性の考察がしやすい．

定義 3.9.6 (bias robust estimation)**.** $(P_\theta)_{\theta\in\Theta}$ をモデルとする．ある定数 $M \in \mathbb{R}$ が存在して，$\Gamma_\theta := \Gamma(P_\theta)$ について $|\Gamma_\theta^{-1}\psi(x,\theta)| \leqslant M$ を満たす ψ が定める M-推定量を B-**ロバスト推定量**という．

例 3.9.7 (最尤推定)**.** 最尤推定量は，漸近分散が最小という意味では最適であるが，一般に影響関数は有界にはならない．特に，Y 方向の誤差について頑健であるが，X 方向の誤差については有界ではない．ここを補おうとするのが，**有界影響推定**または GM 推定である．

例 3.9.8. OLS によって係数を推定した回帰モデルは，崩壊点 (breakdown point) $1/n$ を持つ非頑健なモデルである．そこで，外れ値＝推定値に高い影響力を持つ影響点を検出する手法「回帰診断」があり，外れ値に対するウェイトを制御することで頑健なパラメータ推定を行うのが頑健回帰となる．

3.9.3　Fisher の一致性

Fisher の一致性を仮定すると，最適な不偏ロバスト推定量を構成できる．

定義 3.9.9 (Fisher consistency)**.** 写像 $T : P(\mathcal{X}) \to \Theta$ が，モデル (P_θ) について，$T(P_\theta) = \theta$ を満たすとき，**Fisher 一致性を満たす**という．

3.10　頑健回帰推定

Andersen *Modern Methods for Robust Regression* (2008)

3.11　尤度比検定

初等統計で扱われるカイ 2 乗検定の根拠は，この漸近理論によって明らかになる．

3.11.1　検定から見た漸近論

「密度推定は無理でも，汎関数の分布ならわかる」という技法である．

棄却域を定めるに当たって，決定理論から観て妥当な設計をするためには，検定統計量の帰無仮説の下での分布を知る必要がある．これを解析的に行なえない場合やモデルの misspecification が疑われる場合は漸近分布によって近似をすることとなる．

統計量が t-分布に従うと仮定するとき，χ^2-分布に従うと仮定するときの検定法を t-検定，χ^2-検定というのである．パラメトリックモデル $N(\mu,\sigma^2)$ を仮定したとき，S_n を標本標準偏差として $\dfrac{\sqrt{n}(\overline{X_n}-\mu)}{S_n}$ は τ-分布に従い，多変数の場合は二次形式 $Z^\top\Sigma^{-1}Z$ は χ^2-分布に従うことが判明している．また，Pearson の適合度検定や一部の尤度比検定は，検定統計量が漸近的に χ^2-分布に従う．だがこれらは toy example である．平均や二次形式を検定するために，よりよい検定統計量を構成できる．これらは大抵分布がわからないから，大標本理論＝漸近理論に頼ることとなる．

Neyman-Pearson の理論はある指数分布族モデルに対して一様最強力検定の構成を可能にする．Rao-Blackwell の理論は不偏推定量のクラスの中で分散が最小のものが存在することを明らかにする．一方でこのような理論が使えないときは，漸近最適理論

が呼ばれる．次に競われるのは検定出力関数 (power function) の近似や，漸近分散となる．例えば，滑らかなパラメトリックモデルに対しては最尤推定量が漸近最適である．

(1) 漸近的一致性を持つ．
(2) 真値への収束レートが最速である：$n^{-1/2}$.
(3) 極限分布の分散が最小である．実際，Cramer-Rao を漸近的に満たす．

3.11.2　尤度比検定

記法 3.11.1. $\Theta \subset \mathbb{R}^p, \theta_0 \in \Theta^\circ$ とする．$(\mathcal{X}, \mathcal{A})$ 上の σ-有限な確率分布族 $\{P_\theta\}_{\theta\in\Theta} \subset P(X)$ に対する検定

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_2 : \theta \neq \theta_0$$

を考える．$x = (x_1, \cdots, x_n) \in \mathcal{X}^n$ を無作為標本とする．

定義 3.11.2. 尤度関数 $L_n(\theta) := \prod_{j=1}^{n} p(x_j, \theta)$ の比

$$\Lambda_n := \frac{L_n(\theta)}{\sup_{\theta\in\Theta} L_n(\theta)} \in [0,1]$$

の値を考え，$\Lambda_n : \mathcal{X}^n \to [0,1]$ の値が小さい時に帰無仮説を棄却する検定を**尤度比検定**という．

注 3.11.3. $\widehat{\theta}_n$ が θ の最尤推定量であるとき，定義上これが尤度関数を最大にする点であるから，$\Lambda_n = \dfrac{L_n(\theta_0)}{L_n(\widehat{\theta}_n)}$ となる．

3.11.3　棄却域の定め方

では，棄却域はどう定めるのが「良い」か？ $\Lambda_n : \mathcal{X}^n \to [0,1]$ の分布は極めて複雑であるため，近似を用いる．ここで漸近論が登場し，漸近正規性が，ミニマックスの意味で「妥当」であることが言える．正しくは，漸近決定理論の枠組みで捉えるのが良い．

定義 3.11.4.

(1) θ に関する，規準化されたスコア関数を $Z_n(\theta) := \dfrac{1}{\sqrt{n}} \sum_{j=1}^{n} \partial'_\theta \log p(x_j, \theta)$ とする．
(2) $Z_n := Z_n(\theta_0)$ とする．

定理 3.11.5. $\widehat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$ を満たす最尤推定量が存在し，漸近正規性を持つための十分条件 [F1] [F4] を満たすとする 3.6.5．このとき，P_{θ_0} の下で

$$-2\log\Lambda_n \xrightarrow{d} \chi^2(p) \quad (n \to \infty)$$

3.11.4　複合仮説の検定

3.11.5　Rao 検定

複合仮説の検定に関しては，尤度比検定に変わって，ラオ検定（Lagrange 未定乗数法検定）を用いることが出来る．これは H_0 の下で尤度比検定と同じ漸近分布 $\chi^2(r)$ を持つ．

3.11.6　Wald 検定

Wald 検定も尤度比検定と漸近同等であって，特に H_0 の下で漸近的に分布 $\chi^2(r)$ に従う．

3.12　多項分布の検定

記法 3.12.1 (モデルの設定)**.**

$$\Delta := \left\{ p = (p_1, \cdots, p_k) \in \mathbb{R}_+^k \mid \sum_{i=1}^k p_i = 1 \right\}$$

とするとこれは多項分布の全体とみなせる．開集合 $\Theta \subset \mathbb{R}^l$ から $\mathrm{Int}(\Delta)$ への単射な C^1 級写像 $p :\hookrightarrow \Theta \to \mathrm{Int}(\Delta)$ で定まるパラメトリックモデル Θ を考える 3.4.4．さらに $U \overset{\text{open}}{\subset} \mathbb{R}^m$ $(1 \leqslant m < l \leqslant k-1)$ 上の C^1 級部分モデル $\theta : U \hookrightarrow \Theta$ を考える．

問題 3.12.2. 仮説検定問題

$$H_0 : \theta \in \theta(U) \quad \text{vs.} \quad H_1 : \theta \notin \theta(U)$$

を考える．スコア関数と漸近同等な θ, u の推定量 $\hat{\theta}_n, \hat{u}_n$ がモデル Θ, U のそれぞれに対して得られているとする．このとき，検定統計量 Q_n, Q_n^* を

$$Q_n = \sum_{i=1}^k \frac{[np_i(\hat{\theta}_n) - np_i \circ \theta(\hat{u}_n)]^2}{np_i \circ \theta(\hat{u}_n)}$$

$$Q_n^* = \sum_{i=1}^k \frac{[np_i(\hat{\theta}_n) - np_i \circ \theta(\hat{u}_n)]^2}{np_i(\hat{\theta}_n)}$$

とし，Q_n, Q_n^* の値が大きいときに H_0 を棄却することを考える．

定理 3.12.3. H_0 の下で，Q_n, Q_n^* の漸近分布はカイ 2 乗分布 $\chi^2(l-m)$ である．

要諦 3.12.4. 証明抽出により，尤度比検定と漸近同等であることがわかる．

例 3.12.5 (goodness-of-fit test)**.** 各事象 E_i が起こる確率 p_i は，与えられた値 p_i^* に等しい，という単純仮説の検定を**適合度検定**という．このとき，漸近分布は $\chi^2(k-1)$ である．

3.13　接触構造

P_n が帰無仮説，Q_n が対立仮説の法則という意味論がある．接触しているとは**漸近的絶対連続**であることをいい，統計量の極限において，測度変換を与える基礎になる．

3.13.1　尤度比と測度変換

測度変換の公式が成り立つためには，絶対連続性が必要である．

2 つの確率変数があったときに，一方のもう一方に対する絶対連続部分の Radon-Nikodym 微分を**尤度比確率変数**という．これは零集合上の差を除いて一意に定まる．「尤度比」という名前は実用上の重要性が先行して考察されている対象かと思いやすいが，実際は本質的に数学的に自然に重要な対象である．

絶対連続性 $Q \ll P \Leftrightarrow P(A) = 0 \Rightarrow Q(A) = 0$ は，P-a.e. 性質はそのまま Q-a.e. に成り立つことを意味し，$\mathrm{supp}\, q \subset \mathrm{supp}\, p$ だから P のグラフが上から覆いかぶさっている形になる．$q/p\, dP$ なる形式は Q^a を表す．実はこの Ω 全域での積分が 1 であるとき，$Q^\perp = 0$ である．実際，

$$\int f dQ \geqslant \int f \frac{dQ}{dP} dP \quad (f \geqslant 0)$$

が一般に成り立つが，等号が成り立って測度変換公式として使えるための必要十分条件が $Q \ll P$ である．
Lebesgue 測度は参照測度として優秀で，人類はみなこの測度に引き戻して計算を行う．

動機 3.13.1 (尤度比とは，数学的には絶対連続な確率測度組の Radon-Nikodym 微分のことに他ならない)**.**

(1) （参照測度 ν の下の）任意の確率測度 P, Q について，絶対連続部分と直交／特異部分とへの Lebesgue 分解 $Q = Q^a + Q^\perp$ が存在して，$Q^a \ll P$ かつ $Q^\perp \perp P$ かつ

$$\forall_{A\in\mathcal{F}} \quad Q^a(A) = \int_A \frac{q}{p} dP, \qquad \frac{dQ}{dP} = \frac{q}{p} \ P\text{-a.e.}$$

が成り立つ．

(2) こうして，Radon-Nikodym 微分という確率変数 $\dfrac{dQ}{dP} : \Omega \to \mathbb{R}_+$ を確率 P の下で考えるというテーマが定まる．これを**尤度比**という．

(3) 尤度比の平均値が 1 であることと，$Q^\perp = 0$ であることと，絶対連続 $Q \ll P$ であることとは同値．

補題 3.13.2. $P, Q \ll \mu$ を確率測度とし，対応する密度を p, q とする．

(1) 絶対連続部分 Q^a の参照測度 μ が介在しない表示 $Q^a(A) = \displaystyle\int_A \frac{q}{p} dP$ が成り立つ．

(2) 尤度比の P-平均値が 1 であること $\displaystyle\int \frac{q}{p} dP = 1$ と，$Q^\perp = 0 \Leftrightarrow Q[p = 0] = 0$ であることと，絶対連続 $Q \ll P$ であることとは同値．

(3) 一般に $\displaystyle\int f dQ \geqslant \int f \frac{dQ}{dP} dP$ が成り立ち，等号が成立するのは $Q \ll P$ のとき．

3.13.2　接触性

X の Q についての法則を，P の言葉で考えたいとき，q/p を積分核とすれば良いことがわかった．この測度変換の問題の漸近版を考える．いつ，極限実験 P を用いて，対象の実験列 (Q_n) の極限分布が計算可能になるか？

記法 3.13.3. 台となる測度空間 $(\Omega^n, \mathcal{A}^n)$ を共通とする統計的実験の列 $(P_n), (Q_n)$ を考える．そしてそこでの統計量の列 $X_n : \Omega^n \to \mathbb{R}^k$ との振る舞いを考察する．

定義 3.13.4 (contiguous)**.** モデル (Q_n) は (P_n) **下に接触している** $Q_n \lhd P_n$ とは，任意の可測集合の列 $(A_n), A_n \in \mathcal{A}^n$ に対して，$P_n(A_n) \to 0 \Rightarrow Q_n(A_n) \to 0$ が成り立つことをいう．

注 3.13.5. $\dfrac{dQ_n}{dP_n}, \dfrac{dP_n}{dQ_n}$ はそれぞれ P_n, Q_n の下で一様に緊密であるから，Prohorov の定理から，弱コンパクトである．すなわち，任意の部分列は，弱収束する部分列をもつ．以降，この意味で U に収束するとき，$\dfrac{dP_n}{dQ_n} \xrightarrow[Q_n]{d}{}^* U$ と表す．

補題 3.13.6 (接触性の特徴付け (LeCam 1))**.** 次の 4 条件は同値．

(1) $Q_n \lhd P_n$.

(2) $\dfrac{dP_n}{dQ_n} \xrightarrow[Q_n]{d}{}^* U$ ならば，$P[U > 0] = 1$ である．

(3) $\dfrac{dQ_n}{dP_n} \xrightarrow[P_n]{d}{}^* V$ ならば，$E[V] = 1$ である．

(4) 任意の統計量 $T_n : \Omega_n \to \mathbb{R}^k$ について，$T_n \xrightarrow{P_n} 0$ ならば $T_n \xrightarrow{Q_n} 0$ である．

要諦 3.13.7. 漸近的でない場合の消息でいうと，

$$Q_n\left(\frac{dP_n}{dQ_n} = 0\right) = 0 \quad \Leftrightarrow E_P\left[\frac{dQ}{dP}\right] = 1$$

は $Q_n \ll P_n$ に同値だが，$Q_n \lhd P_n$ と同値になるためには，上のような表現となる．

例 3.13.8 (漸近的対数正規性).

$$\frac{dP_n}{dQ_n} \xrightarrow[Q_n]{d} e^{N(\mu,\sigma^2)}$$

が成り立つとき，$Q_n \lhd P_n$ で，さらに互いに接触していることは $\mu = -\frac{1}{2}\sigma^2$ に同値.

定理 3.13.9 (漸近的測度変換 (LeCam 3)). $X_n : \Omega^n \to \mathbb{R}^k$ を確率変数列とする.

(1) $Q_n \lhd P_n$,
(2) $\left(X_n, \frac{dQ_n}{dP_n}\right) \xrightarrow[P_n]{d} (X, V)$

ならば，$L(B) := E[V, B] = E[1_B V]$ は確率測度を定め，これについて $X_n \xrightarrow[Q_n]{d} L$.

補題 3.13.10 (接触性の十分条件). P_n, Q_n はある σ-有限測度 ν_n に対して絶対連続であるとし，Radon-Nikodym 微分を p_n, q_n とおく.

$$\Lambda_n := \begin{cases} \log(q_n/p_n), & p_n, q_n > 0, \\ 0, & p_n = q_n = 0, \\ \infty, & p_n = 0, q_n > 0, \\ -\infty, & p_n > 0, q_n = 0. \end{cases}$$

ある確率変数 Λ が存在して P_n の下で $\Lambda_n \xrightarrow{d} \Lambda$ かつ $E[e^{\Lambda}] = 1$ であるならば，(Q_n) は (P_n) に接触している.

3.14　局所漸近正規性

パラメータのスケーリングの違いを除いて，局所漸近正規なモデルはあるガウスモデルに収束する.

議論 3.14.1. ある点 $\theta_0 \in \mathbb{R}^k$ に注目して，その周りの局所パラメータを $h := \sqrt{n}(\theta - \theta_0)$ とすると，元の統計的実験 $(\mathcal{X}^n, \mathcal{A}^n, (P_\theta^n)_{\theta\in\mathbb{R}^k})$ は，$(\mathcal{X}^n, \mathcal{A}^n, (P^n_{\theta_0+h/\sqrt{n}})_{h\in\mathbb{R}^k})$ と変換される．するとこの統計的実験は，元の統計的実験 $\theta \mapsto P_\theta$ が滑らかであるとき，$(\mathcal{X}^n, \mathcal{A}^n, (N(h, I_{\theta_0}^{-1}))_{h\in\mathbb{R}^k})$ に漸近する！

3.14.1　尤度の拡張

議論 3.14.2. $P_\theta \ll \mu$ とし，$p_\theta := \frac{dP_\theta}{d\mu}$ と表す．ここでは $\theta, h \in \mathbb{R}$ のような記法をするが，k 次元としても $\dot{l}(x)$ が k ベクトルになり，掛け算が内積または二次形式になるのみである．$l_\theta(x) := \log p_\theta(x) \in C^2(\mathcal{X})$ と仮定すると (この仮定は強すぎるが)，Taylor の定理より，

$$\log P_{\theta+h}(x) = l_\theta(x) + \dot{l}_\theta(h) + \frac{1}{2}\ddot{l}_\theta(x)h^2 + o_x(h^2). \quad h \in \mathbb{R}^k, h \to 0$$

$$\log \frac{p_{\theta+h}}{p_\theta}(x) = h\dot{l}_\theta(x) + \frac{1}{2}h^2\ddot{l}_\theta(x) + o_x(h^2). \quad h \in \mathbb{R}^k, h \to 0$$

ここで尤度比が現れる．特に x に観測値 $X_1, \cdots, X_n$ を代入し，h を $\frac{h}{\sqrt{n}}$ として $n \to \infty$ の極限を考えると，

$$\log \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) = \frac{h}{\sqrt{n}}\sum_{i=1}^n \dot{l}_\theta(X_i) + \frac{1}{2}\frac{h^2}{n}\sum_{i=1}^n \ddot{l}_\theta(X_i) + \sum_{i=1}^n o_x\left(\frac{h^2}{n}\right) \quad (n \to \infty)$$

ここで，$E_\theta[\dot{l}_\theta] = 0, -E_\theta[\ddot{l}_\theta] = E_\theta[\dot{l}_\theta^2] = I_\theta$ だから，

(1) $\Delta_{n,\theta} := \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{l}_\theta(X_i) \in \mathbb{R}$ とおくと $\Delta_{n,\theta} \xrightarrow{d} N(0, I_\theta)$（中心極限定理）.
(2) 大数の法則より $\frac{1}{n}\sum_{i=1}^n \ddot{l}_\theta(X_i) \xrightarrow{d} -I_\theta$.
(3) 3 項目はおそらく確率収束（仮定をうまく設定すれば良い）.

よって，

$$\log\prod_{i=1}^{n}\frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i)=h\Delta_{n,\theta}-\frac{1}{2}I_\theta h^2+o_{p_\theta}(1)$$

が成り立つ．実は，必要な過程は，L^2-微分の言葉で簡潔に書ける．

定理 3.14.3 (局所漸近正規性の十分条件)**.** モデル $(p_\theta)_{\theta\in\Theta\overset{\text{open}}{\subset}\mathbb{R}^k}$ は，密度関数の平方根 $\theta\mapsto\sqrt{p_\theta}$ が L^2-微分可能であるとする：

$$\exists_{\dot{l}\in\mathscr{L}^2(\mathfrak{X})}\quad\int\left(\sqrt{p_{\theta+h}}-\sqrt{p_\theta}-\frac{1}{2}h^\top\dot{l}_\theta\sqrt{p_\theta}\right)^2d\mu=o(\|h\|^2)\quad(h\to0).$$

なお，通常の意味で微分可能であった場合は，

$$\frac{1}{2}\dot{l}_\theta\sqrt{p_\theta}=\frac{\partial\sqrt{p_\theta}(x)}{\partial\theta}\quad\Leftrightarrow\quad\dot{l}_\theta=\frac{\partial}{\partial\theta}\log p_\theta(x)$$

が必要であることに注意．このとき，

(1) $E_\theta[\dot{l}]=0\in\mathbb{R}^k$ かつ $I_\theta:=E_\theta[\dot{l}_\theta\dot{l}_\theta^\top]\in M_k(\mathbb{R})$ が定まる．
(2) 任意の収束列 $\{h_n\}\subset\Theta, h_n\to h$ について，$\Delta_{n,\theta}:=\dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=1}^{n}\dot{l}_\theta(X_i)\in\mathbb{R}^k$ とすると，

$$\log\prod_{i=1}^{n}\frac{p_{\theta+h_n/\sqrt{n}}}{p_\theta}(X_i)=h^\top\Delta_{n,\theta}-h^\top\frac{1}{2}I_\theta h+o_{p_\theta}(1)$$

定義 3.14.4 (LAN: local asymptotic normal)**.** モデル $(P_{n,\theta}|\theta\in\Theta)_{n\in\mathbb{N}}$ が $\theta_0\in\Theta$ において**局所漸近正規**であるとは，

(1) 行列 $r_n, I_{\theta_0}\in M_k(\mathbb{R})$,
(2) P_{θ_0} の下で $N(0,I_{\theta_0})$ に分布収束する確率変数 $\Delta_{n,\theta}:\Omega\to\mathbb{R}^k$,

が存在して，任意の収束列 $\{h_n\}\subset\mathbb{R}^k, h_n\to h$ について，

$$\log\frac{dP_{n,\theta+r_n^{-1}h_n}}{dP_{n,\theta}}=h^\top\Delta_{n,\theta}-\frac{1}{2}h^\top I_\theta h+o_{P_{n,\theta}}(1)$$

と展開できることをいう．

例 3.14.5. L^2-微分可能なモデル $(P_\theta)_{\theta\in\Theta}$ が定める実験列 (P_θ^n) は，$r_n=\sqrt{n}I$ を行列として局所漸近正規である．

3.14.2　正規実験への収束

上の「局所漸近正規性」の意味を，統計的実験列の収束概念以前の状態で考えたい．

議論 3.14.6. 統計的実験が収束するとき，元の実験で弱収束する統計量は，極限実験での弱収束先に法則同等である．

定義 3.14.7. ランダム化された統計量 $T=T(X,U)$ とは，$U\sim U([0,1])$ にも依存する統計量をいう．

定理 3.14.8 (正規実験へ収束するための十分条件)**.** モデル $(p_\theta)_{\theta\in\Theta\overset{\text{open}}{\subset}\mathbb{R}^k}$ は，密度関数の平方根 $\theta\mapsto\sqrt{p_\theta}$ が L^2-微分可能であり，Fisher 情報行列 $I_\theta:=E_\theta[\dot{l}_\theta\dot{l}_\theta^\top]\in\mathrm{GL}_k(\mathbb{R})$ が可逆であるとする．実験 $(P_{\theta+h/\sqrt{n}})_{h\in\mathbb{R}^k}$ における統計量 $T_n:\mathfrak{X}^n\to\mathbb{R}^k$ が，任意の $P_{\theta+h/\sqrt{n}}$ $(h\in\mathbb{R}^k)$ について弱収束するとき，実験 $(N(h,I_\theta^{-1}))_{h\in\mathbb{R}^k}$ 上のランダム化された統計量 $T:=T(X,U):\mathfrak{X}\times[0,1]\to\mathbb{R}^k$ が存在して，任意の h について同じ収束先へ分布収束する．

3.15　漸近決定理論

$h_0 \in H$ 上の尤度比過程は，周りの h と h_0 との「分離度」を表す．実際対数尤度比の P_h についての平均が h, h_0 の間の KL-分離度である．そこで統計的実験が収束するとは，任意の尤度比過程がバージョンの違いを除いて収束することとすると，統計モデルがある種の"Riemann 多様体"として，各点において局所構造が等しいことと思える．
極限的実験を定めると，ここでの効率限界が，元の実験での下界を与える．

3.15.1　実験列の収束の定義

定義 3.15.1 (likelihood ratio process, limit experiment)**.** 統計的実験 $(\mathcal{X}, \mathcal{A}, (P_h)_{h\in H})$ について，

(1) $h_0 \in H$ 上の**尤度比過程**とは，$\left(\frac{dP_h}{dP_{h_0}}(X)\right)_{h\in H}$，または参照測度が存在するとき，$\left(\frac{p_h}{p_{h_0}}(X)\right)_{h\in H}$ をいう．これは，観測という確率変数 X に，モデルの密度関数族が定める積写像 $(p_h/p_{h_0})_{h\in H}$ を合成して得たものである．

(2) 実験列 $((\mathcal{X}_n, \mathcal{A}_n, P_{n,h}; h \in H))_{n\in\mathbb{N}}$ が**極限実験** $(\mathcal{X}, \mathcal{A}, P_h; h \in H)$ に収束するとは，それらの任意の $h_0 \in H$ 上の尤度比過程の任意の有限次元周辺分布が，h_0 が真のパラメータであるという過程の下で法則収束することをいう：

$$\forall_{h_0\in H}\ \forall_{I\subset H}\ |I| < \infty \Rightarrow \left(\frac{dP_{n,h}}{dP_{n,h_0}}(X_n)\right)_{h\in I} \xrightarrow{h_0} \left(\frac{dP_h}{dP_{h_0}}(X)\right)_{h\in I}.$$

3.15.2　漸近表現定理

法則収束する統計量の漸近的な振る舞いは，極限実験における統計量の考察に帰着する．実験内の最適統計量の考察と，実験の収束とは，完全に独立に議論できるのである！
実は，さらに仮定をおくと，最尤推定量は極限実験での最尤推定量に収束し，尤度比統計量の列は尤度比統計量に収束することなども言える．

記法 3.15.2. 実験 $\mathcal{E}_n = (P_{n,h}|h \in H)_{n\in\mathbb{N}}$ 上の統計量 $T_n : \mathcal{X}^n \to H$ を考える．任意の分布 P_h の下で，$T_n \xrightarrow{P_h} L_h \in P(H)$ に法則収束するとする．このとき，T_n の漸近的な振る舞いは極限法則 $\{L_h\}_{h\in H}$ に支配される．ここで，極限法則 $\{L_h\}_{h\in H}$ が，極限実験における（少し変換した）統計量の法則に対応する．

定義 3.15.3 (dominated, randomized statistic)**.**

(1) 実験 $\mathcal{E} = (P_h|h \in H)$ が**支配されている**とは，ある σ-有限測度 μ が存在して $\forall_{h\in H}\ P_h \ll \mu$ が成り立つことをいう．

(2) $(\mathcal{X}, \mathcal{A}, P_h|h \in H)$ 上の**ランダム化された統計量** T とは，可測写像 $T : \mathcal{X} \times [0,1] \to \mathbb{R}^k$ をいう．ただし，$[0,1]$ 上には一様分布を与える．

定理 3.15.4 (法則収束統計量の漸近表現定理)**.** 実験 $\mathcal{E}_n = (P_{n,h}|h \in H)_{n\in\mathbb{N}}$ は，μ に支配された実験 $\mathcal{E} = (P_h|h \in H)$ に収束するとする．$\mathcal{E}_n$ の統計量 T_n が，任意の $h \in H$ について法則収束するならば，$\mathcal{E}$ のランダム化された統計量 T が存在して，$\forall_{h\in H}\ T_n \xrightarrow{h} T$.

3.15.3　漸近正規性

$h \in H$ にて局所漸近正規なモデルは，正規分布モデル $N(Jh, J)$ へ収束する実験列を定める．

定理 3.15.5 (局所漸近正規ならば，正規実験に収束する)**.** $\mathcal{E}_n = (P_{n,h}|h \in H)$ を実験列で，パラメータ空間 $H \subset \mathbb{R}^d$ は $0 \in H$ を満たすとする．$h = 0$ 下で $N(0, J)$ $(J \in M_d(\mathbb{R}))$ に収束する確率変数列 $\Delta_n : \mathcal{X}^n \to \mathbb{R}^d$ を用いて

$$\log \frac{dP_{n,h}}{dP_{n,0}} = h^\top \Delta_n - \frac{1}{2} h^\top J h + o_{P_{n,0}}(1)$$

と展開できるとき，実験列 $\mathcal{E}_n$ は $(N(Jh,J))_{h\in H}$ に収束する．

3.15.4　一様分布

一様分布のモデル $(U([0,\theta]))_{\theta\in\mathbb{R}_+}$ は L^2-微分可能ではなく，正規な実験に収束しない．実は，指数実験に収束する．

3.15.5　Pareto 分布

Pareto 分布の極限実験は，正規実験と指数実験の組み合わせとなる．

3.15.6　漸近混合正規性

3.16　尤度比確率場の局所漸近構造

統計推測の最適性の問題は，尤度解析に帰着される．

3.16.1　漸近決定理論

漸近決定理論の方法は，その普遍性ゆえに，確率過程の統計推測論などの新しい分野と合流し発展を続けている．統計量の漸近有効性を議論するためには，統計的実験の列に対する漸近決定理論の立場で一般論を展開する方が好ましい．これは，小標本理論における十分性・不偏性・不変性などの精緻な理論を，非線形な場合にも拡張する企てである．

歴史 3.16.1. 漸近論への数学的アプローチで，初めて理論と言えるものは Wald 1943 と LeCam1960, 1972, 1979 である．そこで，統計的実験とは，多様体 $\{P_\theta\}_{\theta\in\Theta}\subset P(\mathbb{R}^n)$ とされた．いまでは無限次元の場合（セミ／ノンパラメトリック）も考えられている．実験を繰り返すにつれて，実験は局所的に簡単なモデルで近似可能になっていく，この現象を Gaussian shift という．よって，問題は元の実験の研究から，Gaussian shift の近似へと移り変わる（極限定理）．これは本質的には実験の滑らかさに起因する．

定義 3.16.2 (experiment, deficiency)**.** $\Theta\neq\varnothing$ に関する実験とは，3 つ組 $E=(\Omega,\mathcal{A},\mathcal{P})$ をいう．パラメータ空間 Θ に関する実験全体の集合を $\mathcal{E}(\Theta)$ で表す．$\delta(E,F):=\inf\left\{\epsilon>0\mid E\overset{\epsilon}{\subset}F\right\}$ $(E,F\in\mathcal{E}(T))$ に対して，$\Delta(E,F):=\max\{\delta(E,F),\delta(F,E)\}$ を**欠損**という．これは $\mathcal{E}(T)$ 上に擬距離を定める．

$$\Delta(E,F)=\sup\{\}$$

3.16.2　統計的実験

漸近正規性を定義する枠組みを整備する．Frechet 微分のような考え方をするが，L^2 ノルムにより条件が弱められている [6].

定義 3.16.3. 可測空間とその上の確率分布族 $(P_\theta)_{\theta\in\Theta}$ の組 $(\mathcal{X},\mathcal{A},\{P_\theta\}_{\theta\in\Theta})$ を**統計的実験**という．大標本理論では，直積による列 $(\mathcal{X}^n,\mathcal{A}^n,\{P_\theta^n\})_{n\in\mathbb{N}}$ を考える．

記法 3.16.4 (正則性の仮定)**.** $\Theta\overset{\text{open}}{\subset}\mathbb{R}^p$ とし，P_θ の真値 P_{θ_0} に対する Lebesgue 分解における，絶対連続部分の Radon-Nikodym 微分を dP_θ/dP_{θ_0}，特異部分を σ_{θ/θ_0} で表すと，

$$P_\theta[A]=\int_A\frac{dP_\theta}{dP_{\theta_0}}dP_{\theta_0}+\sigma_{\theta/\theta_0}(A)\quad(A\in\mathcal{A})$$

関数 $D_u : \mathcal{X} \to \mathbb{R}$ $(u \in \mathbb{R}^p)$ を

$$D_u(x) := \sqrt{\frac{dP_{\theta_0+u}}{dP_{\theta_0}}(x)} - 1$$

で定める．

(G1) 原点の近傍の $u \in \mathbb{R}^p$(十分小さな変位) に対して $D_u \in L^2(P_{\theta_0})$ であって，$u \mapsto D_u$ は $u = 0$ において $L^2(P_{\theta_0})$ の意味で微分可能である：

$$\exists_{\varphi=\varphi_{\theta_0}\in L^2(P_{\theta_0};\mathbb{R}^p)} \quad \int_{\mathcal{X}} (D_u(x) - u'\varphi(x))^2 P_{\theta_0}(dx) = o(|u|^2) \quad (|u| \to 0).$$

(G2) $\sigma_{\theta_0+u/\theta_0}(\mathcal{X}) = o(|u|^2)$ $(|u| \to 0)$.

補題 3.16.5 (L^2-微分の表示)**.** G1,G2 を仮定し，P_θ は確率密度関数族 $p(x,\theta)$ を持ち，対応 $\theta \mapsto p(x,\theta)$ は θ_0 において微分可能とする．このとき，

$$\varphi(x) = \frac{1}{2p(x,\theta_0)} \frac{\partial p}{\partial \theta}(x,\theta_0).$$

また，Fisher 乗法行列に当たる作用素 $\Theta \to \mathbb{R}$ を

$$I(\theta_0) := 4\int_{\mathcal{X}} \varphi(x)\varphi(x)^\top P_{\theta_0}(dx)$$

とおく．

補題 3.16.6.

3.16.3　局所漸近正規性

定理の $Z_n(u;\theta_0)$ の展開が成り立つとき，すなわち，G1,G2 が成り立つとき，統計的実験 $(\mathcal{X}^n, \mathcal{A}^n, \{P_\theta^n\}_{\theta\in\Theta})$ は θ_0 において**局所漸近正規**であるという．
対応 $\mathbb{R}^p \ni u \mapsto Z_n(u;\theta_0) \in C(\mathbb{R}^p)$ を尤度比確率場という．Ibragimov HasMinskii は，この関数の $C(\mathbb{R}^p)$ 上での弱収束を示した：$Z_n(-;\theta_0) \xrightarrow{d} Z(-;\theta_0)$.

記法 3.16.7 (normalized likelihood ratio)**.** 正規化された尤度比を

$$Z_n(u;\theta_0) := \prod_{j=1}^n \frac{dP_{\theta_0+n^{-1/2}u}}{dP_{\theta_0}}(x_j), \quad (u \in \mathbb{R}^p, x \in \mathcal{X}^n)$$

とおくと，Z_n は十分大きな $n \in \mathbb{N}$ に対して定義されている．

定理 3.16.8 (LeCam)**.** 正規な統計的実験の列 E_i について，$I(\theta_0)$ $(\theta_0 \in \Theta)$ は非退化とする．[G1],[G2] を仮定し φ を D_u の L^2-微分とする．このとき，ある p 次元と 1 次元確率変数 $\Delta_n(\theta_0), \rho_n(u,\theta_0)$ が存在して，正規化された尤度比は次の表現を持つ：

$$Z_n(u;\theta_0) := \exp\left(u^\top \Delta_n(\theta_0) - \frac{1}{2}u^\top I(\theta_0)u + \rho_n(u,\theta_0)\right).$$

また，

$$L\left[\Delta_n(\theta_0)|P_{\theta_0}^n\right] \to N_p(0, I(\theta_0)), \quad \rho_n(u,\theta_0) \xrightarrow{P_{\theta_0}^n} 0.$$

［証明］.

$$\Delta_n(\theta_0) := \frac{2}{\sqrt{n}}\sum_{j=1}^n \varphi(x_j), \quad \rho_n(u,\theta) := \log Z_n(u;\theta_0) - u^\top\Delta_n(\theta_0) + \frac{1}{2}u^\top I(\theta_0)u.$$

とすれば良い．■

要諦 3.16.9. この定理が保証する尤度比の性質は，推定量の性質を調べるに当たって強力な道具となる．特に，尤度比の対数 $\log Z_n$ は $N\left(-\frac{1}{2}u^\top I(\theta_0)u, u^\top I(\theta_0)u\right)$ に分布収束する．

定義 3.16.10 (LAN: Locally Asymptotically Normal (LeCam 1960))**.** 定理の $Z_n(u;\theta_0)$ の展開が成り立つとき，すなわち，G1,G2 が成り立つとき，統計的実験 $(\mathcal{X}^n, \mathcal{A}^n, \{P_\theta^n\}_{\theta\in\Theta})$ は θ_0 において**局所漸近正規**であるという．

要諦 3.16.11. LAN のとき，$P^n_{\theta_0+n^{-1/2}u}$ は $P^n_{\theta_0}$ に接触している．

3.16.4　局所漸近正規な例

> 統計モデルの列が正規モデルで近似可能であるという性質が，局所漸近正規性である．正則なパラメトリックモデルの独立同分布モデルは局所漸近正規で，これについての LeCam の定理は中心極限定理に他ならない．

定理 3.16.12 (Hajek の十分条件)**.** $\Theta\subset\mathbb{R}^1$ で添字付けられた実験 E_i の分布族 $f(x,\theta)=\dfrac{dP_\theta}{d\nu}$ が次の 3 条件を満たすならば，族 P_θ^n は LAN 条件を $\theta=t$ について満たす．

(1) 任意の $x\in\mathcal{X}$ について $f(x,-):\Theta\to[0,1]$ は $\theta=t$ の近傍 $U\overset{\text{open}}{\subset}\Theta$ で絶対連続である．
(2) ν-a.e.$x\in\mathcal{X}$ について，微分係数 $\dfrac{\partial f(x,\theta)}{\partial\theta}$ が $\theta\in U$ 上存在する．
(3) $I(\theta)$ $(\theta=t)$ は連続で半正定値である．

3.16.5　漸近有効性

> sup によって，パラメータの 1 点でのみ良いパフォーマンスを示す病的な推定量（超有効推定量）が競争から排除されている．これが漸近ミニマックスリスク関数である．

定理 3.16.13 (Hajek の不等式)**.** 統計的実験は θ_0 にて局所漸近正規であるとする．θ の任意の推定量系列 T_n について，

$$\forall_{\gamma,\delta>0}\quad \liminf_{n\to\infty}\sup_{\theta:|\theta-\theta_0|<\delta} E_\theta\left[|\sqrt{n}(T_n-\theta)|^\gamma\right]\geqslant E[|I(\theta_0)^{-1}\Delta(\theta_0)|^\gamma].$$

3.17　漸近有効性

> 滑らかなパラメトリックモデルの漸近下界は $N(0,I_\theta^{-1})$ が与える．

記法 3.17.1. $\sqrt{n}(T_n-\psi(\theta))$ は全ての $\theta\in\Theta$ について分布収束すると仮定して (この仮定を正則性という)，その効率性を比較する．

3.17.1　実験の下界

> 漸近有効性の議論は，正規な位置母数モデル (Gaussian shift モデル) での有効推定量の問題に帰着する．

定義 3.17.2 (regular)**.** 統計量の列 $(T_n:\mathcal{X}^n\to\Theta)$ が θ において**正則**または**漸近的法則同等**であるとは，ある極限分布 L_θ が存在して，

$$\forall_{h\in\Theta}\ \sqrt{n}\left(T_n-\psi\left(\theta+\frac{h}{\sqrt{n}}\right)\right)\xrightarrow[\theta+h/\sqrt{n}]{d}L_\theta$$

が成り立つことをいう．極限分布 L_θ が h に依ることを許すとき，**前正則**であるとする．前正則な統計量が正則であるとは，局所一様に取れる消息を表す．

定理 3.17.3.

(E1) 実験 $(P_\theta)_{\theta\in\Theta}$ は点 $\theta\in\Theta\overset{\text{open}}{\subset}\mathbb{R}^k$ で $\theta\mapsto dP^{1/2}$ が L^2-微分可能で，可逆な Fisher 行列 $I_\theta\in\mathrm{GL}_k(\mathbb{R})$ を持つとする．

(E2) $\psi : \Theta \to \mathbb{R}$ は θ で微分可能であるとする．
(E3) 実験 $(P_{\theta+h/\sqrt{n}})_{h\in\mathbb{R}^k}$ の統計量 T_n は前正則であるとする．

このとき，ランダム化された統計量 T が実験 $(N(h, I_\theta^{-1}))_{h\in\mathbb{R}^k}$ 上に存在して，$T - \dot{\psi}_\theta h \sim L_{\theta,h}$．

3.17.2　Gauss モデルの平均の推定

3.17.3　畳み込み定理

正則な統計量の中で，最適なものは解析的に定まる．

定理 3.17.4.

(E1) 実験 $(P_\theta)_{\theta\in\Theta}$ は点 $\theta \in \Theta \overset{\text{open}}{\subset} \mathbb{R}^k$ で $\theta \mapsto dP^{1/2}$ が L^2-微分可能で，可逆な Fisher 行列 $I_\theta \in \mathrm{GL}_k(\mathbb{R})$ を持つとする．
(E2) $\psi : \Theta \to \mathbb{R}$ は θ で微分可能であるとする．
(E3) 実験 $(P_\theta^n)_{\theta\in\Theta}$ の統計量 T_n は正則で極限分布 L_θ を持つとする．

このとき，ある確率測度 $M_\theta \in P(\mathcal{X})$ が存在して，$L_\theta = N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top) * M_\theta$ と表せる．

3.17.4　局所漸近 minimax 定理

全く違うアプローチで，最適な統計量が正規分布 $N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top)$ に漸近的に従うことを示す方法を考える．

3.17.5　下界を達成する推定量

漸近分布 $N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top)$ を持つ推定量が最適とわかったから，このようなものの必要十分条件を探す．これは実は漸近線型性である．線形性が正規性に対応するのである！

補題 3.17.5 (漸近線型性:最適性の特徴づけ)**.**

(E1) 実験 $(P_\theta)_{\theta\in\Theta}$ は点 $\theta \in \Theta \overset{\text{open}}{\subset} \mathbb{R}^k$ で $\theta \mapsto dP^{1/2}$ が L^2-微分可能で，可逆な Fisher 行列 $I_\theta \in \mathrm{GL}_k(\mathbb{R})$ を持つとする．
(E2) $\psi : \Theta \to \mathbb{R}$ は θ で微分可能であるとする．

このとき，実験 $(P_\theta^n)_{\theta\in\mathbb{R}^k}$ の統計量列 (T_n) について，次の 2 条件は同値．

(1) 漸近線型である：$\sqrt{n}(T_n - \psi(\theta)) = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^n \dot{\psi}_\theta I_\theta^{-1} \dot{l}_\theta^\top(X_i) + o_P(1)$.

(2) T_n は $\psi(\theta)$ に対する正則な統計量の中で最適である：$\forall_{h\in\Theta}\ \sqrt{n}\left(T_n - \psi\left(\theta + \dfrac{h}{\sqrt{n}}\right)\right) \xrightarrow[\theta+h/\sqrt{n}]{d} N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top)$.

要諦 3.17.6. $\Delta_{n,\theta} := \dfrac{1}{\sqrt{n}} \sum \dot{l}_\theta(X_i)$ の部分が $N(0, I_\theta)$ に収束する．

例 3.17.7. 適切な正則条件の下で，最尤推定量は漸近線型である．したがってデルタ法によって，任意の汎関数 $\psi(\theta)$ に対して，plug-in 推定量は漸近線型である．

3.18　射影

統計量の漸近分布を求める方法の 1 つに，既に漸近分布が知れている確率変数と漸近同等であることを示す方法がある．このとき，Slutsky の補題の簡単な系から，極限法則が一致することが従う．
すると，何で近似するかという問題が生じるが，射影は L^2-近似であることを考えると，この言葉を使うのが良かろう．

3.18.1　Hajek の射影

Hajek の射影は，ある確率変数 $X_1,\cdots,X_n$ と，それぞれの一変数関数 g_i との合成 $g_1(X_1),\cdots,g_n(X_n)$ の形で T を最もよく近似する元である．

記法 3.18.1.

$$S := \left\{ \sum_{i=1}^n g_i(X_i) \in L^2(\Omega) \;\middle|\; \forall_{i\in[n]}\ g_i(X_i) \in L^2(\Omega) \right\}$$

とおく．

補題 3.18.2 (Hajek projection)**.** $X_1,\cdots,X_n$ は独立とする．このとき，任意の $T \in L^2(\Omega)$ の S への射影は

$$\mathrm{pr}_S(T) = \sum_{i=1}^n E[T|X_i] - (n-1)E[T]$$

と表せる．

要諦 3.18.3. 特に $X_1,\cdots,X_n$ が同分布に従うとし，$T = T(X_1,\cdots,X_n) : \Omega \to \mathcal{X}^n \to \mathbb{R}$ を対称な可測関数とする．このとき，条件付き期待値は

$$E[T|X_i = x] = E[T(x, X_2, \cdots, X_n)]$$

と簡略化されるから，

$$S' := \left\{ \sum_{i=1}^n g(X_i) \in \mathcal{L}(\Omega) \;\middle|\; g \in \mathcal{L}(\mathcal{X}) \right\}$$

上への射影と一致する：$\mathrm{pr}_S = \mathrm{pr}_{S'}$.

3.18.2　Hoeffding の分解

一変数関数 $g_i : \mathcal{X} \to \mathbb{R}$ を用いるのではなく，より変数を多くして，$X_1,\cdots,X_n$ の間の複雑な依存関係も表現できるようにすることで，近似の精度を上げることを考える．これは，初めは $(H_i)_{i\in[n]}$ の線型結合で表せる元の空間だったものを，どんどん互いに直交する部分空間を追加することで大きくしていく算譜である．

記法 3.18.4. 任意の $A \in P([n])$ に対して，

$$H_A := \left\{ g_A(X_i : i \in A) \in L^2(\Omega) \mid g_A \in \mathcal{L}(\mathcal{X}^{|A|}),\ \forall_{B\in P([n])}\ |B| < |A| \Rightarrow E[g_A(X_i : i \in A)|X_j : j \in B] = 0 \right\}$$

と定める．

補題 3.18.5. 各 $\{H_A\}_{A\in P(n)} \subset L^2(\Omega)$ は組ごとに直交している．

定理 3.18.6. $X_1,\cdots,X_n$ を独立な確率変数，$T \in L^2(\Omega)$ とする．

(1) $\mathrm{pr}_{H_A}(T) = \sum_{B\in P(A)} (-1)^{|A|-|B|} E[T|X_i : i \in B]$.

(2) $\forall_{B\in\mathcal{P}(A)}\ T\perp H_B$ ならば，$E[T|X_i:i\in A]=0$.

(3) $\bigoplus_{B\in\mathcal{P}(A)} H_B$ は，$(X_i)_{i\in A}$ の関数で 2 乗可積分であるものの全体を含む部分空間となる．

要諦 3.18.7. 特に $X_1,\cdots,X_n$ が同分布に従うとし，$T=T(X_1,\cdots,X_n):\Omega\to\mathcal{X}^n\to\mathbb{R}$ を対称な可測関数とする．このとき，Hoeffding 分解は

$$T=\sum_{r=0}^{n}\sum_{|A|=r}g_r(X_i:i\in A),\qquad g_r(x_1,\cdots,x_r)=\sum_{B\in\mathcal{P}[r]}(-1)^{r-|B|}E[T(x_i\in B,X_i\notin B)]$$

となり，各項 $\sum_{|A|=r}g_r(X_i:i\in A)$ は退化した核を持った r 次の U-統計量で，全ての項は互いに直交だから，分散は容易に計算できる：

$$\mathrm{Var}[T]=\sum_{r=1}^{n}\binom{n}{r}E[g^2(X_1,\cdots,X_r)].$$

3.19　U-統計量

U-統計量の理論は Hoeffding 1948 による．U-統計量は最小分散不偏推定量の計算中に自然に現れる．
$r=1$ のとき，核 h を用いた標本平均の一般化となっている．すなわち，$U_{1,n}=\mathbb{P}h(X)$．これをさらに $r>1$ とすると，基本的には resampling 法の一種で，標本から大きさ r の再抽出をして平均を取る．これは標本分割に繋がるのではないか？

記法 3.19.1. $I_n=[n],(\mathcal{X},\mathcal{A})$ を可測空間，$X_j\ (j\in I_n)$ を独立同分布に従う $\mathcal{X}$-値確率変数，$h:\mathcal{X}^r\to\mathbb{R}$ を対称な可測関数とする．部分集合 $A:=\{i_1,\cdots,i_r\}\subset I_n$ に対して，$X_A=(X_{i_1},\cdots,X_{i_r})$ と略記すると，対称性から $h(X_A):=h(X_{i_1},\cdots,X_{i_r})$ と表して良い．すなわち，関数 $h(X_\_,\cdots,X_\_)$ は $I_n^r\to\mathbb{R}$ として定めたが，本質的には $[I_n]^r\to\mathbb{R}$ である．

議論 3.19.2. モデル $\mathcal{P}$ の母数 γ が**推定可能**であるとは，自然数 $r\in\mathbb{N}$ と r 変数関数 h が存在して，

$$\forall_{P\in\mathcal{P}}\ X_1,\cdots,X_r\overset{\text{i.i.d.}}{\sim}P\Rightarrow E_P[h(X_1,\cdots,X_r)]=\gamma$$

を満たすことをいう．このときの最小の r を**次数**という．条件を満たす h は必ず対称に取れる．

3.19.1　対称な関数

対称な関数の取り扱いと表示は仰々しいが，要は対称関数 $h:\mathcal{X}^r\to\mathbb{R}$ は集合 $[\mathcal{X}]^r$ 上の関数 $f:[\mathcal{X}]^r\to\mathbb{R}$ と一対一対応する．そして $[\mathcal{X}]^r$ なる対象は，$\mathcal{X}$ の点から r 個の抽出として自然に出現する．

定義 3.19.3. $E[|h(X_A)|]<\infty$ とする．部分集合 $B\subset A$ に対して，

$$h(X_A)_B:=\sum_{C:C\subset B}(-1)^{|B|-|C|}E[h(X_A)|X_C]$$

とする．$E[h(X_A)|X_\varnothing]:=E[h(X_A)]$ とした．

補題 3.19.4 (反転公式)**.** $h(X_A)=\sum_{B:B\subset A}h(X_A)_B$.

3.19.2　定義と例

標本 n 個のうち r 個を選んで，その r 個の部分集合 $A,|A|=r$ に対応する値 $h(X_A)$ の標本平均として得られる統計量のクラスを，U-統計量という．
$\theta:=E[h(X_1,\cdots,X_r)]$ の推定を考えると，当然 h 自身は θ の不偏推定量であるが，h を核とする U-統計量 $U_{r,n}$ は再び不偏推定量で，より小さい漸近分散を持つ (また漸近正規である)．実際，$U_{r,n}$ は h の射影 (条件付き期待値) となっており，

射影はノルム減少的であるから分散を減らすのである.

定義 3.19.5 (U-statistic with kernel h)**.** 対称関数 $h : \mathcal{X}^r \to \mathbb{R}$ を**核**とする r 次の U-**統計量**とは,

$$U_{r,n} := \binom{n}{r}^{-1} \sum_{A \in [I_n]^r} h(X_A) = \frac{1}{\binom{n}{r}} \sum_{\sigma : [r] \hookrightarrow [n]} h(X_{\sigma(1)}, \cdots, X_{\sigma(r)})$$

をいう.

要諦 3.19.6 (標本平均の拡張)**.** 単に 1 つ 1 つの観測点 X_i に等しく荷重平均を取る $\mathbb{P}$ を一般化して,

標本 n 個からの r 個のすべての選び方についての測度 $\overline{\mathbb{P}} \subset P([I_n]^r)$ について, ある対応 $h : [\mathcal{X}]^r \to \mathbb{R}$ に関しての値の平均 $U_{r,n} = \overline{\mathbb{P}}h(X, \cdots, X)$ をいう.

このとき, 標本平均の場合と違って, 各項は互いに独立ではない！

定義 3.19.7. U-統計量 $U_{r,n}$ が**退化次元 $i-1$ を持つ**とは, $2 \leqslant i \leqslant r$ に関して,

$$E[h(X_1, \cdots, X_r) \mid X_1, \cdots, X_{i-1}] = 0, \quad \wedge \quad E[h(X_1, \cdots, X_r) | X_1, \cdots, X_i] \neq 0$$

を満たすことをいう. また, $E[h(X_1, \cdots, X_r)|X_1] \neq 0$ であるとき, **非退化**であるという.

例 3.19.8 (rank correlation coefficient)**.** 明らかに標本平均, 不偏分散は, それぞれ $f = \mathrm{id}_{\mathcal{X}}, f(x_1, x_2) = \dfrac{(x_1 - x_2)^2}{2}$ に関する U-統計量である. 歪度も $k_{3,n}(x) = \sum \dfrac{(x_i - \overline{x}_n)^3 n}{(n-1)(n-2)}$ が定める U-統計量で, とくに 3 次の k-統計量である.

また, 2 次元の無作為標本 (X_j, Y_j) $(j = 1, \cdots, n)$ に関して, 順位の間の相関係数として

$$\tau := \frac{4}{n(n-1)} \sum_{i<j} 1_{\{(X_i - X_j)(Y_i - Y_j) > 0\}} - 1$$

を **Kendall の順位相関係数**という. また,

$$K := \left| \left\{ \{i, j\} \in [[n]]^2 \mid (x_i > x_j \wedge y_i > y_j) \vee (x_i < x_j \wedge y_i < y_j) \right\} \right|,$$
$$L := \left| \left\{ \{i, j\} \in [[n]]^2 \mid \neg(x_i > x_j \wedge y_i > y_j) \wedge \neg(x_i < x_j \wedge y_i < y_j) \right\} \right|$$

として $\tau = \binom{n}{2}^{-1} (K - L)$ とも表せる.

例 3.19.9 (K-statistic)**.** キュムラントの最小分散不偏推定量を Fisher の K-統計量と言い, John Tukey はこれから一般化 K-統計量 (polykay) を導いたが, これは斉次多項式が定める U-統計量に他ならない.

例 3.19.10 (V-statistic)**.** 対称な関数 h に関して,

$$V_n := \frac{1}{n^r} \sum_{(i_1, \cdots, i_r) \in I_n^r} h(X_{i_1}, \cdots, X_{i_r})$$

を V-統計量という. U-統計量は大きさ r の部分標本の抽出について平均を取ったが, V-統計量は大きさ r の部分標本の, 抽出の順序も区別し, また重複も許す. $[I_n]^r$ と I_n^r の違いである. V-統計量は Hoeffding の一年前である 1947 年に Richard von Mises が導入した.[†3] $v_1 > 0$ のとき, U-統計量と漸近同等である.

$h(x, y) = \dfrac{(x-y)^2}{2}$ が定める 2 次の V-統計量は

$$V_{2,n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

[†3] von Mises, R. (1947). "On the asymptotic distribution of differentiable statistical functions". Annals of Mathematical Statistics. 18 (2): 309 - 348. doi:10.1214/aoms/1177730385. JSTOR 2235734.

で，これは分散の最尤推定量である．しかし同じ核に対する U-統計量は

$$U_{2,n} = s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

となる．漸近同等ではあるが，不偏推定量となるのは後者である！

例 3.19.11 (Wilcoxon 統計量)**.** Wilcoxon 統計量が漸近正規であることも，U-統計量に関係する．

3.19.3　漸近正規性

構成の仕方から明らかに，大数の法則から U_n は母数 $\theta := E[h(X_1,\cdots,X_r)]$ の不偏推定量となっている．（また，これが逆マルチンゲールになっていることからも従う）．そこで，$\sqrt{n}(U_n-\theta)$ の漸近分布を考える．
そのための道具が Hoeffding の分解で，U-統計量は，より退化度の高い U-統計量の和として表せる．これは確率変数の分解に似ている．

記法 3.19.12. $E[|h(X_1,\cdots,X_r)|] < \infty$ とする．

$$h_k(x_1,\cdots,x_k) := E[h(x_1,\cdots,x_k,X_{k+1},\cdots,X_r)] \quad l \leqslant r$$

とし，$h_0 := E[h(X_1,\cdots,X_r)]$ とする．

補題 3.19.13. $\forall_{C\subset A}\ E[h(X_A)|X_C] = h_{|C|}(X_C)$．ただし，$h_0(X_\varnothing) = h_0$ とする．

定義 3.19.14. $\widehat{h_k} := h_k - \theta$ $(k = 0,1,\cdots,r)$ に関して

$$g_k(x_{[k]}) := \sum_{C:C\subset[k]} (-1)^{k-|C|}\widehat{h}_{|C|}(x_C)$$

と定める．このとき，$\widehat{h}_r(X_A)_B = g_{|B|}(X_B)$.

補題 3.19.15 (Hoeffding の分解)**.**

$$U_n - \theta = \sum_{k=1}^{r}\binom{n}{k}^{-1}\binom{r}{k}S_k(n)$$

ただし，$S_k(n) := \sum_{B:B\subset I_n,|B|=k} g_k(X_B)$ とした．

補題 3.19.16 (漸近分布の平均)**.** $h \in L^1(\mathfrak{X}^r)$ とする．$B \subset I_n, |B| = k > 0, b \in B$ について，

$$E[g_k(X_B)|X_{I_n\setminus\{b\}}] = 0.$$

補題 3.19.17 (漸近分布の分散)**.** $h \in L^2(\mathfrak{X}^r)$ とし，$v_k := E[g_k(X_1,\cdots,X_k)^2]$ とおく．$B, C \subset I_n$ について，

$$E[g_{|B|}(X_B)g_{|C|}(X_C)] = \delta_{B,C}v_{|B|}.$$

定理 3.19.18 (中心極限定理)**.** $h \in L^2(\mathfrak{X}^r), v_1 > 0$ とする．このとき，

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, r^2v_1). \quad (n\to\infty)$$

こうして，統計量を汎関数 $\mathcal{P} \to \mathbb{R}$ と捉え，この漸近分布を考える（あわよくば中心極限定理を目指す）発想が出てくる．
これが Hoeffding が von Mises から受け継いだものである．
そう，標本平均とは U-統計量の一例に過ぎないのであった．そして独立確率変数列の和からマルチンゲールの概念が生まれたのならば，U-統計量がマルチンゲールになるのは当然である．
標本平均も，これに核 h を重ねたものも，線型汎関数であることは変わらないから，漸近分布は正規である．

3.19.4　2 標本 U-統計量

3.19.5　退化 U-統計量

定義 3.19.19 (degenerated)**.** U-統計量の列 (U_n)(の核 (h_n)) が**退化している**とは，漸近分散が $r^2 v_1 = 0$ であることをいう．

3.20　情報量規準

3.20.1　枠組みと KL 距離

「真の分布からの KL-距離」自体も一つの情報量規準と考えられる．この情報量規準に従ってモデル選択をすることを「最尤法」と呼ぶことになるのだ（赤池 3.20.14）．

記法 3.20.1. σ-有限測度 ν に支配されている確率分布族 $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta} \subset P(\mathcal{X}, \mathcal{A})$ を考える．真の分布 Q の確率密度関数も存在して $q(x) := dQ/d\nu(x)$ と表され，$\mathcal{P}$ に属しているかは不明とする．n 個の無作為標本 $x = (x_1, \cdots, x_n)$ に基づく未知パラメータ θ の推定量を $\widehat{\theta}_n(x)$ とする．新たな観測に対する予測分布として，plug-in 分布 $P_{\widehat{\theta}_n}$ で Q を近似するのは自然である．

定義 3.20.2 (statistical manifold)**.** σ-有限測度 ν に支配されている確率分布族 $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ が**正則統計モデル**であるとは，

(1) パラメータ空間 Θ はある Euclid 空間 $\mathbb{R}^p$ の開集合と同相．
(2) $\theta \mapsto P_\theta$ は微分可能．
(3) 非特異な Fisher 情報行列を持つ．

このとき，$\mathcal{P}$ は p-次元多様体をなす．

定義 3.20.3 (divergence)**.** $(M, (\xi_x))$ を多様体とする．関数 $D : M \times M \to \mathbb{R}_+$ が**分離度**であるとは，

(1) 非退化性：$D[P, Q] = 0 \Leftrightarrow P = Q$.
(2) 十分近い 2 点 $\xi, \xi + d\xi$ について，$D[\xi, \xi + d\xi] = \dfrac{1}{2} g_{ij}(\xi) d\xi_i d\xi_j$ が定める行列 $G(\xi) = (g_{ij}(\xi))$ は正定値対称である．
(3) 点 $P \in M$ の r-近傍 $N(P, r) := \{Q \in M \mid D[P, Q] < r\}$ は r について単調に増大する．

要諦 3.20.4. 対称性も三角不等式も成り立たず，距離の概念とは違い，むしろ距離の二乗のような，分散のような概念である．そして，微小な 2 点間の「Riemann 距離」を与えるが，非対称性は受け継がれ，これが双対性という方向を持った新たな構造を授ける．

定義 3.20.5 (Kullback-Leibler information / divergence / relative entropy)**.** $P, Q \in \mathcal{P}$ について，

$$I[Q, P] := \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \nu(dx)$$

によって定まる関数 $I : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ を **Kullback-Leibler 情報量**という．

例 3.20.6 (正規分布族の KL 分離度)**.** $M_n(\mathbb{R})$ の中の半正定値行列のなす部分多様体 $\mathrm{GL}_n(\mathbb{R})_+$ は $n(n+1)/2$ 次元である．この多様体に

$$D[P, Q] = \mathrm{Tr}(PQ^{-1}) - \log|PQ^{-1}| - n$$

なるダイバージェンスが定まる．これは，P, Q を分散共分散行列とする平均 0 の正規分布の間の KL ダイバージェンスになっている．

3.20.2　バイアス補正

真の分布と，そのモデルがはじき出す予測分布との分離度を，そのモデルの misspecification の度合いを測る尺度に用いるという発想である．すると，「**モデル f の平均対数尤度が大きいほど良いモデル**」ということになる．しかし，分離度を経験分布から求めても，そのバイアスの推定の仕事が残っている．これはバイアスの漸近値 (の一致推定量 plug-in) を用いるということである．**このバイアスは，同じデータ X を，パラメータの推定と，推定されたモデルの平均対数尤度の推定とに二度用いたことによって生じる**．こうして得られる種々の分離度の推定量を，情報量規準という．

議論 3.20.7 (KL 情報量の最小化と平均対数尤度の最大化は同値)**.** 真の分布 Q と予測分布 $P_{\hat{\theta}_n}$ の分離度は

$$I[Q, P_{\hat{\theta}_n}] = \int_{\mathfrak{X}} q(z) \log q(z) \nu(dz) - \int_{\mathfrak{X}} q(z) \log p(z, \hat{\theta}_n) \nu(dz).$$

ここで，モデルの選択に依るのは第二項の**平均対数尤度**

$$h(Q, p, \hat{\theta}_n(x)) := \int_{\mathfrak{X}} \log p(z, \hat{\theta}_n(x)) Q(dz)$$

である．

要諦 3.20.8. Q の経験分布 P_n について，$h(P_n, p, \hat{\theta}_n(x))$ で推定することになるが，これが最大になるときの $\hat{\theta}_n$ を最尤推定量というのであった．これが，情報理論的な観点から再発見される最尤法の意味の一つである．

議論 3.20.9 (バイアスの推定)**.** 第二項 h を経験分布から求めることになるが，そのときのバイアスはどうなるか．

$$b_n := \int_{\mathfrak{X}^n} \left[h(P_n, p, \hat{\theta}_n(x)) - h(Q, p, \hat{\theta}_n(x)) \right] Q^n(dx).$$

このとき，次が成り立つ．

定理 3.20.10. ある $\theta_* \in \Theta$ と $\varphi_Q \in L^2(\mathfrak{X}, Q; \mathbb{R}^k)$ が存在して，$\int_{\mathfrak{X}} \varphi_Q(x) Q(dx) = 0$ かつ

$$R := \sqrt{n}(\hat{\theta}_n - \theta_*) - \zeta_n \xrightarrow{Q_n} 0, \quad \zeta_n := n^{-1/2} \sum_{j=1}^{n} \varphi_Q(x_j)$$

が成り立つと仮定する．[†4]

スコアを $s(x, \theta) = \partial_\theta \log p(x, \theta)$，$C := \mathrm{Cov}_Q[s(-, \theta_*)^\top, \varphi_Q]$ とすると，$nb_n \xrightarrow{n \to \infty} \mathrm{Tr} C$.

定義 3.20.11 (information criterion)**.** 尤度関数にペナルティ項 p を付けた

$$IC = -2 \left[\sum_{j=1}^{n} \log p(x_j, \hat{\theta}_n(x)) - p(n) \right]$$

を，推定量 $\hat{\theta}_n$ に対する**情報量規準**という．

3.20.3　AIC

$(-2) \log_e$(最大尤度) の第 1 項は変わらず，第 2 項のペナルティ項をいろいろ変えることで，情報量規準が作れる．ペナルティ項をモデルの次元としたものが AIC で，仮定が少ないものを採用する「オッカムの剃刀」の方針に則ったものと考えられる．

例 3.20.12.

[†4] 最小コントラスト推定量はこの仮定を満たす．

(1) バイアス修正を考えるならば，$p(n) = \mathrm{Tr}C$ と取るとよいが，これは Q に依存するので，その一致推定量を取るとよい．

(2) 推定量 $\hat{\theta}_n$ が Ψ が定める最小コントラスト推定量であるときの形は，小西貞則-北川源四郎による**一般化情報量規準** (GIC: Generalized Information Criterion) の特殊な場合である．これは一般の統計的汎関数 $T : \mathcal{P} \to \mathbb{R}^m$ に関する漸近バイアスの研究による．特に最尤推定量であるとき，補正項は $p(n) = \mathrm{Tr}(J_Q^{-1} I_Q)$ となり (Huber 1976, 竹内 1976)，この一致推定量で置き換えた情報量規準を**竹内の情報量規準** (TIC) という．I は Fisher 情報量行列で，J は Hesse 行列．

(3) $\hat{\theta}_n$ が最尤推定量である上で特にモデルが真の分布 Q を含むとき，$I = J$ が成り立ち，漸近バイアスはパラメータベクトル $\theta \in \Theta$ の次元となる．すなわち，パラメータ数を k として $p(n) = k$ となる．この場合を**赤池情報量規準** (AIC) という．

議論 3.20.13 (独立性・有意性検定は最小 AIC 推定で置き換えられる（あるいは近似的には同じ行為である）)**.** 二次関数に正規誤差 $N(0, \sigma^2)$ を持たせて発生させたデータに，1 次，2 次，5 次で fitting すると，実は 5 次の方が最小二乗誤差は小さいが，真のデータ生成規則と異なる．これを AIC が解決することが最初の例として挙げられている (1976[14])．これは「雑音で乱された信号の復元にも AIC が使えることを表している」と論じている．

続いて，独立性の検定に応用する．独立性を仮定した場合と仮定しなかった場合のどちらのモデルを採用するかを AIC で判定できるという．Fisher の χ^2-検定と結論が同じになると論じている．さらに χ^2-検定に対する有意性として「大量データの統計解析の自動化の可能性があることを示している」と提示している．

歴史 3.20.14 (AIC の誕生 (1971) [14] 参照)**.** AIC は予測問題と密接に関係があることが，従来の検定・推定の立場と最も異なる点であるという．

> 従来の推定論は，単に未知パラメータの推定という形で展開され，推定結果の利用目的が明確でないために，推定精度の評価基準，例えば 2 乗平均誤差などの選定上の指導理念を与えることができなかった．

(1) 定常時系列 (y_n) を，過去 k 段階前までの観測値の線型結合で予測することを考える．それぞれの線型結合の係数は，最小二乗誤差で捉え，最終予測誤差 (FPE) を

$$FPE := E[y_n - \widehat{a_{k0}} - \widehat{a_{k1}} y_{n-1} \cdots - \widehat{a_{kk}} y_{n-k}]$$

で定めて，最良の次数 $k \in \mathbb{N}$ を決めたい．なお，ここに $z_n := y_n - \widehat{a_{k0}} - \widehat{a_{k1}} y_{n-1} \cdots - \widehat{a_{kk}} y_{n-k}$ が互いに独立であると仮定できるとき，これをこれを自己回帰モデルという．この次数 $k \in \mathbb{N}$ の決定は従来検定を用いることが考えられていたが，実用的なものは知られていなかった．しかし FPE の最小化と考えると，推定論の問題である．FPE を推定して，その最小化を考えて k を選択すれば良い，とした (Akaike 1969,70).

> この最小 FPE 法と呼ばれる方法の実用上の有効性は著しく，現在まで多くの成功的な適用例，特に各種研究分野におけるスペクトル解析への応用例，が報告されている．この方法を $\{y_n\}$ がベクトル過程である場合に拡張したものは，セメント焼成工程の計算機制御の実現に用いられ，またそのほかの複雑な統計的システムの解析に利用されつつある．

(2) しかし実は，$\{y_n\}$ がベクトル過程である場合への拡張には困難がある．$X \in \mathbb{R}^q$ を共変量として，$y \in \mathbb{R}^p$ に対する回帰 $Y = FX + Z$ を**因子分析モデル**という．$F \in M_{pq}, Z$ の各成分は X の各成分と独立であるとする．これは，Y の成分間の相関関係を，低次元因子ベクトル X で説明しようとする．Z はノイズである．従来の決定法は，Z にパラメトリックな仮定を置いて，$q \in \mathbb{N}$ を増加させながら最尤法によって「q に伴う最大尤度の増加分の検定結果」が有意でなくなるような最初の q を**因子数**として決定するものであった．するとこれは，一体何を基準に考えているのかが明らかでない．FPE のように，「何を最小化しようとしているのか？」が明らかでない．

(3) 役者が揃った．因子分析モデルから最尤法が採用され，AR モデルから FPE という「統計量にまとめること」が採用されることになる．

> MLE は Fisher によって初めて詳細に論じられてから極めて広く用いられているが，その有効性の根拠が何にあるのかは明らかではなく，その合理性の説明は通常直感に訴えてなされてきた．前項で論じた因子分析法には最尤法が用いられている．そこで，最尤法が何を最適化しようとしているのかが明らかになりさえすれば，先の問題は解決されることになる．

この後すぐに，漸近論的な立場から，M-推定量のような発想「最尤法は $E[\log_e f(Y|\theta)]$ を最大にする θ の推定量である」さらに言い換えると，「KL-情報量を規準として，真の分布を最もよく近似する $f(y|\theta)$ を求めようとしている」と捉え直している．

最後にこの観点から，FPE の場合の平均二乗誤差の代わりに，$\mathbb{P}_n[I(g; f(|\hat{\theta}))]$ を取ることを考える．この値の推定は，第 1 項を無視して，$\mathbb{P}_n[E[\log_e f(Y|\hat{\theta})]]$ の推定と等価．そして実は $N \cdot \mathbb{P}_n[E[\log_e f(Y|\hat{\theta})]]$ の推定値は $l(\hat{\theta}) - k$ となる．KL 情

報量または FPE の場合に合致するように符号を反転させると，これが AIC である．

要諦 3.20.15. 赤池さんはすごく統計的決定理論の霊性を持っていることがわかる．因子分析モデルの最尤法はどのような意味で最適化を明らかにした．さらに，FPE は将来の値の予測であったが，AIC は将来の値の分布の予測，という質的な進化も経ている．この結果として，AIC はほぼ無際限の広い適用分野を見出すことが可能になった．そして赤池さん多分 Fisher 大好きだ．そしてパラメトリックな仮定というのは「当たれば最強」という点において統計学者の腕の見せ所であり，人類の希望でもあるのだろうな．我々はセミパラメトリックな Fisher になろう．

> 我々はここに適当なモデルの系列によってデータの外の情報あるいは知識（事前情報）を適切に表現することにより，しばしば飛躍的な予測精度の向上をもたらす高度に実用的な推定法が実現されているのを見ることができる．機械的盲目的な推定・検定の分類は，まさに逆立ちした有意性検定の虚像に基づくものといえよう．Fisher によって書かれた書物が現在に至るまで多くの応用分野の研究者によって利用されているのは，そこに与えられている各種のモデルの系列の持つ有効性の魅力によるものといえる．有意性検定を支えているものは，その一見客観的に見える論理構成ではなく，Fisher 自身実際のデータの解析を通じてその有効性を信ずるに至った，極めて経験的なその手続きの実用性である．この特性はある程度そのまま AIC に受け継がれ，それが AIC の特徴と同時にまた限界を形成している．

議論 3.20.16 (最小 AIC 推定の哲学から見た有意性検定)**.** Fisher の有意性検定は，例えば「接種と発病とが独立である」という仮説を，データによって反駁することを目的とするものとされてきた．しかし赤池 [14] はこう論じる．

> しかしこうした見方が論理的矛盾を含むものであることは明らかである．非常に厳密に見れば，接種は必ず大なり小なりの影響を人体に与えるはずで，その結果は発病と完全に無関係ではあり得ない．あるサイコロの正しさを検定するという問題も全く同様で，現実のサイコロで完全に対称なものが存在し得ないことは明らかである．このように，仮説は常に否定される立場にあり（帰無仮説），データによる検定結果を待つまでもなく結論は見えている．この論理的矛盾に加え，有意性検定における判定基準選定上の曖昧さは著しく，いわゆる神聖な 5 パーセント水準の信仰の発生など議論の絶え間がなく，遂に有意性検定の有用性を疑う人まで現れるようになった．

これは，認識論と Bayes 学派とを折衷する立場じゃないのか？？**帰無仮説を棄却できるというのは，「実は無効であるという制約された人工モデルの方が予測精度が良い」という我々の無知が生んだ奇異な状況を脱却できるかの試練である．**

有意性検定を，「帰無仮説は AIC を最小化するか？」という問題を解いている行為とほとんど等しいとみなせることを思い出すと，有効モデル＝対立仮説は常に何らかの意味では正しいが知識不足であり（いわばノンパラメトリックモデル？），無向モデル＝帰無仮説は人工的に制約を加えたモデルであると考えられる．実験データが十分に大きいならば，機械学習なり，$n(i,j)/N$ を $p(i,j)$ の推定量として使えば良い．しかし，データは限られてるので，無効モデルの方が有効な近似を与える可能性があるし，何より推定すべきパラメータは減少するので我々にも feasable である．この時，制約を加えている無効モデルは現実の構造からあまりにも隔たっているかもしれない．こうして，トレードオフの関係にある．

(1) 有意性検定における無効モデルは，相関の非存在を主張するから，考慮すべきパラメータ数は減少する．従って，人類の計算能力からすると予測精度は上がる．
(2) しかし，現実とは異なる仮定をおいているため，これを原因としたバイアスがあるはずである．

この 2 つのどちらのモデルが良いかを，「無効モデルの方が正しいと想定して比較検討」して，どちらかのより予測精度が高いモデルを採用してこれを「科学的知識」とするのが有意性検定である．

歴史 3.20.17 (「バイアス補正法」として拡張の歴史を辿った)**.** モデルの評価基準を「その予測能力としよう」として初めて提案したのは赤池 (1973,74)．−2 の係数はこのときの名残である．モデルが真の分布を含まない場合にも拡張したのが竹内啓 (76) である．小西・北川 (96) は，バイアス補正の方法を一般の統計量に拡張した．石黒 (97) はブートストラップ法によってバイアス補正を行うことを提案し，拡張情報量規準 EIC(Extended IC) と呼ばれる．

注 3.20.18. AIC の発明のときから取っている基本的な考え方として次の 3 点がある．

(1) モデルの良さはその予測能力でみる．
(2) 予測は分布に対して行う．
(3) 分布の近さは (二乗誤差などではなく)KL 情報量で測る．

要諦 3.20.19. 本懐はバイアスの補正であるとするなら，ロバスト統計と考えていることの方向性は同じではないか？

3.20.4　尤度とは何か

たぶん，セミパラメトリック化以外の精神は全て赤池先生がやってくれているような気がする．

議論 3.20.20 (尤度の客観性・主観性の二面性)**.** AIC の導入は，平均対数尤度 $l(\theta)/N = \mathbb{P}_n[\log f(Y|\theta)]$ が，$E[\log_e f(Y|\theta)]$ の推定値であるという着想の下に成立した．この見方に従えば，$\{f(y|\theta)\}_{\theta\in\Theta}$ の族が Y の真の分布を与える $g(y)$ を含まない場合においても，最尤法が統計的モデルのパラメータ決定に有効なものでありうることが容易にわかる．直交射影ではないが，集合 $\{f(y|\theta)\}_{\theta\in\Theta}$ の中で g に最も KL-距離の意味で近い点を選出する算譜なのだろう．つまり，赤池さんの言葉で言えば，最尤推定量 $\widehat{\theta}$ はエントロピーに関して g を最もよく近似する $f(y|\theta)$ を与える θ の，観測値 $y_1, \cdots, y_N$ に基づく推定値を与える．ただし，適用条件は，平均対数尤度が $E[\log_e f(Y|\theta)]$ の推定値として有効な限りであり，またモデル $\{f(Y|\theta)\}$ も人間が勝手に自分の責任において取捨選択するものである．「Fisher が明らかにし得なかった尤度の本質的に実験的帰納的な性格と，統計理論における尤度利用の必然性とが，ここに客観的に描き出されている」.

歴史 3.20.21 (稲垣宣生)**.** 赤池 [14] は $g(y)$ がある θ_0 によって $g = f(-|\theta_0)$ と与えられる状態を想定することは一つの**主観的な行動原理**であるという．つまり，「想定したモデルが正しいという仮定の下に，エントロピーの期待値を評価する」という行動指針はどういう意味を持つのか？という疑問を呈している．そして有意性検定も，AIC も，この行動原理に貫かれている．そして，この主観性を明確に表現する損失関数の下での最小 AIC 推定量の漸近最適性を証明したのが稲垣氏の仕事である．

3.20.5　エントロピー最大化原理

AIC（対数尤度を用いたモデル評価）を用いることは，次の指針に従って統計的推測を行うことと等価である [14].「予測分布」を作りたい．これを，観測データの関数として表すためのモデル $\{f_\theta\}$ を適切に定義し，エントロピーを評価基準としてその最適化を図り，モデル $\{f_\theta\}$ を選択し (AIC)，パラメータ θ を選択する (MLE)．なお，エントロピーとは KL-情報量 $-I(g;f_\theta)$ である．

要諦 3.20.22 (Boltzman のエントロピーの確率論的解釈)**.** エントロピーが増大するとは乱雑になるということであるが，これは確率論から翻訳することもできる．エントロピー $B(g;f_\theta) := -I(g;f_\theta)$ は，f_θ が定める確率分布に従う確率変数を独立に繰り返し観測した時に得られる標本分布として，g が得られる確率の大数に比例する尺度である [16].

3.21　密度推定

ここからはノンパラメトリック解析の手法を概観する．

記法 3.21.1. (X_j) を密度関数 f の定める独立同分布に従う実確率変数列とし，ここから f の推定を考える．

3.21.1　カーネル密度推定

枠組みとしては非常に自然であり，核 K の選択はたしかに正規性の仮定が自然だろう．

定義 3.21.2. 積分核の中でも統計分野で**カーネル**というと，次の 2 条件を満たすものをいう：

(1) $\int_{\mathbb{R}} K(u)du = 1.$
(2) 偶関数：$K(-u) = K(u)$ a.s..

なお，時系列解析では**窓関数**という．

定義 3.21.3 (KDE: Kernel density estimator, Parzen-Rosenblatt window)**.** ある可測関数 $K : \mathbb{R} \to \mathbb{R}$ とバンド幅 $h > 0$ について，$K_h(x) := \dfrac{1}{h}K(x/h)$ として平滑化し，

$$\widehat{f}(x) := \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right) = \frac{1}{n}\sum_{j=1}^{n} K_h(x - x_i)$$

として定まる推定量 $\widehat{f}$ を**核型推定量**または**パルツェン窓**という．

例 3.21.4. 標準正規分布の確率密度関数であるガウス関数

$$K(x) = \frac{1}{2\pi}e^{-x^2/2}$$

をカーネル関数として採用することが多い．

要諦 3.21.5. バンド幅は平滑化のためのパラメータである．いずれにしろ可測関数 $K_h : \mathbb{R} \to \mathbb{R}$ を用いて，$x - x_j$ の重み $K_h(x)$ 付き標本平均を $f(x)$ の推定値とするのである．

これは，$K_h(x)$ なる波形を各観測点 x_j において合成波を取ったものと見れる．実際，熱核を各 x_j において熱の総量を求める操作に等しい．なお，ヒストグラムもカーネル推定の例と見れる．カーネル推定は合成波の手法で，基本波形を求めようとする逆問題と思える．

3.21.2　誤差評価

誤差の評価法は様々だが，ここでは MISE を規準とする．これは L_2 危険関数とも呼ばれる．

定義 3.21.6 (mean integrated square error)**.** $\mathrm{MISE}(\widehat{f}) := \displaystyle\int_{\mathbb{R}} E[(\widehat{f}(x) - f(x))^2]dx$ を**積分 2 乗誤差**という．

定義 3.21.7 (super kernel)**.** 関数 $K : \mathbb{R} \to \mathbb{R}$ が $s \in \mathbb{N}_{\geqslant 2}$ について次の 2 条件を満たすとき，**クラス s の核**であるという：

(1) $x^s K(x) \in L^1(\mathbb{R})$ すなわち $\displaystyle\int_{\mathbb{R}} |x|^s |K(x)|dx < \infty$.

(2) $\forall_{1 \leqslant i < s} \displaystyle\int_{\mathbb{R}} x^i K(x)dx = 0$.

$\forall_{s \in \mathbb{N}_{\geqslant 2}} \displaystyle\int_{\mathbb{R}} x^s K(x)dx = 0$ を満たすとき，**超核**であるという．

定理 3.21.8. $f \in C^s(\mathbb{R})$ かつ $f^{(s)} \in L^2(\mathbb{R})$ とする．クラス s の核 $K \in L^2(\mathbb{R})$ について，ある定数 $C \in \mathbb{R}$ が存在して，

$$\forall_{n \in \mathbb{N}}\ \forall_{h > 0} \quad \mathrm{MISE}(\widehat{f}) \leqslant C\left(\frac{1}{nh} + h^{2s}\right).$$

要諦 3.21.9. 誤差を最小にする h のレートは $n^{-1/(2s+1)}$ で，そのときの積分 2 乗誤差のオーダーは $O(n^{-2s/(2s+1)})$ となる．これは必ずパラメトリックモデルの n^{-1} よりも遅くなる．

第 4 章

漸近展開とその応用

中心極限定理は，正規近似に基づく，分布の一次の漸近理論であった．Edgeworth による展開はその近似をより精密にしたものである．標本数がそれほど大きくない実際的な状況では，分布のより精密な近似を基礎として統計量を構成することが必要になる．

今日，漸近展開法も確率過程にまでその領域を広げ，新しい確率統計学が発展しつつあるが，このような新領域を理解する上でも，独立確率変数列における現象を理解することが重要である．

キュムラント関数をよく使う．その関係で，テンソルがよく出る．また，最尤推定量の漸近展開に出てくる係数の一部は接続係数にほかならない．

4.1 漸近展開

平均 0，分散共分散行列 $\Sigma > O$ の d 次元確率変数列 $\{Z_j\}_{j\in\mathbb{N}}$ は独立同分布に従うとする．中心極限定理によると，

$$S_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Z_j \xrightarrow{d} N_d(0, \Sigma) \quad (n \to \infty)$$

であるが，この正規近似はあまり精度は良くないため，標本数 n が小さい時にはより良い近似が必要になる．
数学的には，これは X の特性関数 $\varphi_X(u) = E[e^{iu\cdot X}]$ $(u \in \mathbb{R}^d)$ の漸近展開理論に等価である．

定義 4.1.1. 確率変数 $X : \mathcal{C} \to \mathbb{R}^d$ に対して，

$$\chi_r(u; X) := i^r \kappa_r[u \cdot X] = (\partial_\epsilon)_0^r \log \varphi_{u\cdot X}(\epsilon)$$

を**キュムラント関数**という．ただし，$(\partial_\epsilon)_0^r$ とは，ϵ で r 回偏微分をして $\epsilon = 0$ を代入したものをいう．これはキュムラント母関数の r 次項係数（を u の関数と観たもの）である．

4.2 平滑化補題

分布の差と特性関数の差の関係を与える補題を準備する．

4.3　特性関数の展開

4.4　漸近展開の正当性の証明

4.5　漸近展開の変換

4.6　最尤推定量の漸近展開

4.7　漸近展開と情報幾何

4.8　ブートストラップ法

ブートストラップ法では，データから作った分布 $\widehat{P}_n$ から，さらに標本を取る．そこから，リサンプリング法とも言われる．パラメトリックモデルではプラグイン推定量 $P_{\hat{\theta}_n}$，ノンパラメトリックモデルでは経験分布などを取る．

記法 4.8.1. (X_j) を確率分布 P に従う独立な実確率変数列とし，$R_n(\mathbf{X}_n, P) : \mathcal{X} \to \mathbb{R}$ を確率変数とする．$R_n(\mathbf{X}_n, P)$ の分布関数を

$$H_n(x; P) := \int_{\mathbb{R}^n} 1_{R_n(\mathbf{x},P)\leqslant x} P_n(d\mathbf{x})$$

と定め，$H_n(c_\alpha; P) = \alpha \in (0,1)$ を満たす点 c_α を求める問題を考える．

例 4.8.2. P が未知で，$\theta = \theta(P)$ の信頼区間を構成したいとき，推定量 $\widetilde{\theta_n}(\mathbf{X}_n)$ から定まる確率変数 $R_n(\mathbf{X}_n, P) := \sqrt{n}(\widetilde{\theta_n}(\mathbf{X}_n) - \theta(P))$ の分位点 c_α を求めたい．

定義 4.8.3. $\widehat{P}_n$ を P の推定量とする．

(1) 真の分布関数 $H_n(x; P)$ に対する近似 $H_n(x; \widehat{P}_n) := \int_{\mathbb{R}^n} 1_{R_n(\mathbf{x},\widehat{P}_n)\leqslant x} (\widehat{P}_n)^n(d\mathbf{x})$ を，**ブートストラップ分布**という．

(2) 推定分布 $\widehat{P}_n$ に従う大きさ n の無作為標本 $\mathbf{X}^* = (X_1^*, \cdots, X_n^*)$ を**ブートストラップ標本**という．

注 4.8.4 (computer-intensive)**.** 推定分布 $\widetilde{P}_n$ が複雑であるとき，積分 $H_n(x; \widehat{P}_n)$ はモンテカルロ法で近似する．これが，ブートストラップ法が計算機集約的と言われる所以である．

注 4.8.5 (studentization)**.** $R_n(\mathbf{X}_n, P)$ をステューデント化された確率変数で置き換えることで，ブートストラップ分布の近似精度の改善が可能である．

第 5 章

参考文献

参考文献

[1] 吉田朋広『数理統計学』(朝倉書店，2006)

[2] 竹村彰道『現代数理統計学』（学術図書，2020）

[3] 久保川達也『現代数理統計学の基礎』（共立出版，2017）

[4] 西山陽一『マルチンゲール理論による統計解析』（近代科学社，2011）

[5] Rabi Bhattacharya - Course in Mathematical Statistics and Large Sample Theory

[6] Ibragimov and Has'minskii - Statistical Estimation

[7] van der Vaart - Asymptotic Statistics

[8] Helmut Strasser "Mathematical Theory of Statistics"

[9] Hampel - Robust Statistics - 2005

[10] Huber and Ronchetti - Robust Statistics - 2009

[11] 稲垣宣生『数理統計学』

[12] E. L. Lehmann - Nonparametrics: Statistical Methods Based on Randks

[13] 竹内啓『統計的推定の漸近理論』

[14] 赤池広次 (1976) 「情報量規準 AIC とは何か」数理科学 153 号，5-11 ページ.

[15] 赤池広次 (1979) 「統計的検定の新しい考え方」数理科学 198 号，51-57 ページ.

[16] Sanov, I. N. (1961). IMS and AMS Selected Translations in Mathematical Statistics and Probability, Vol. 1, 213-244.