目次

小標本データにおける最深回帰推定量の性能評価

大阪大学大学院基礎工学研究科　　藤木 美江

1　はじめに

最深回帰推定量 (Deepest Regression Estimator：DRE) は外れ値の影響を受けにくいロバストな推定量である．これは最小 2 乗法のような残差の距離を計算する方法とは異なり，データの個数計算を利用する回帰 depth (Rousseeuw and Hubert, 1999) という方法から導き出された推定量である．藤木・白旗 (2005) のシミュレーション実験より，最深回帰推定量はデータ数が大きい場合に精度が高いことがわかっている。

一方，分野によっては多くのデータが得られず，限られたデータの中で解析を行わなければならない．小標本データにおける最深回帰推定量に着目したのは，回帰 depth 法が線形モデルのみならず，より一般的なモデルへ適用可能で柔軟な性質を持つためである．最深回帰推定量が，どんなデータ数の場合においても安定しているならば，非常に有効で，かつ有用な推定量であると考えられる．

しかし，論文で提案されている近似アルゴリズムを利用したプログラム「Medsweep」は，定義に基づいた Exact アルゴリズムの計算結果と比較すると，特にデータ数が非常に小さい場合，直観的に健全だと思われない推定量となった．そのため，計算方法の見直しと新たな近似方法の提案が必要である．本報告では，これまで提案されている計算方法の条件設定の見直しを行い，新たな計算方法についての考察を示す.

2　Regression depth

単回帰における回帰 depth の概念を示すために不適合 (nonfit) の定義を与える．データ集合 $Z_n = \{(x_i, y_i) : i = 1, \cdots, n\} \subset R^2$ に対して, 回帰直線の候補を $\ell : y = \theta_1 x + \theta_2$ とする．$\boldsymbol{\theta} = (\theta_1, \theta_2)$ の θ_1 は傾き, θ_2 は切片とし，残差は $r_i(\boldsymbol{\theta}) = r_i = y_i - (\theta_1 x_i + \theta_2)$ とする.

定義 1 どの x_i とも一致しない実数 v が存在し, 次の (i) または (ii) が成り立つとき, データ集合 Z_n に対して, $\boldsymbol{\theta} = (\theta_1, \theta_2)$ は不適合という.

$$(i) \quad r_i(\boldsymbol{\theta}) < 0, \forall x_i < v \quad \text{かつ} \quad r_i(\boldsymbol{\theta}) > 0, \forall x_i > v$$
$$(ii) \quad r_i(\boldsymbol{\theta}) > 0, \forall x_i < v \quad \text{かつ} \quad r_i(\boldsymbol{\theta}) < 0, \forall x_i > v$$

定義 2 データ集合 Z_n に対して $\boldsymbol{\theta} = (\theta_1,\theta_2)$ の回帰 depth は $\boldsymbol{\theta}$ を不適合にするために取り除かれる必要のある観測値の最小個数である.

$$rdepth(\boldsymbol{\theta}, Z_n) = \min_v \{\sharp(r_i(\boldsymbol{\theta}) \leq 0 \ and \ x_i < v) + \sharp(r_i(\boldsymbol{\theta}) \geq 0 \ and \ x_i > v)\}. \tag{1}$$

ただし，$\boldsymbol{\theta} = (\theta_1, \theta_2) \in R^2$.

定義 1, 2 は x_i に同点がある場合も有効であり, 分布に関する仮定もしていない. Rousseeuw and Leory (1987, p.116) の定義に従うと，回帰 depth は scale invariant, regression invariant, そして affine invariant である．また, 重回帰の場合も同様に定義できる.

回帰 depth を計算するために, はじめに, 観測値を $x_1 \leq x_2 \leq \cdots \leq x_n$ と並べかえ,

$$rdepth(\boldsymbol{\theta}, Z_n) = \min_{1 \leq i \leq n} (\min(L^+(x_i) + R^-(x_i), L^-(x_i) + R^+(x_i)))$$

を使って計算する．ここで，回帰 depth が提案された Rousseeuw and Hubert (1999) において計算の条件設定に誤りがあることを発見し，(1) をもとにして条件の訂正を行なった．下記は訂正後のものである.

$$L^+(v) = \sharp\{i : r_i \geq 0, x_i < v\}, \quad R^-(v) = \sharp\{i : r_i \leq 0, x_i > v\},$$
$$R^+(v) = \sharp\{i : r_i \geq 0, x_i > v\}, \quad L^-(v) = \sharp\{i : r_i \leq 0, x_i < v\}.$$

3　最深回帰推定量

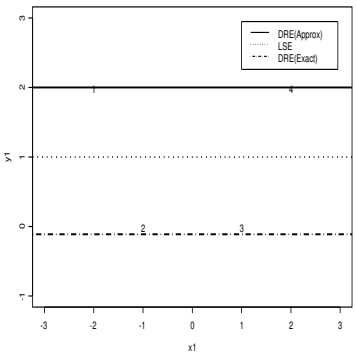
定義 3 p 次元における，最深回帰推定量 $DR(Z_n)$ は $rdepth(\boldsymbol{\theta}, Z_n)$ を最大にする $\boldsymbol{\theta}$ である．ただし，$rdepth(\boldsymbol{\theta}, Z_n)$ を最大にする $\boldsymbol{\theta}$ は，ただ一つとは限らない．

$$DR(Z_n) \quad = \quad \arg\max_{\boldsymbol{\theta}} \ rdepth(\boldsymbol{\theta}, Z_n)$$

定義 4 $\binom{n}{2}$ 個の 2 データ点を通る直線の中で，直線 $\theta_{1i}x + \theta_{2i}$ $(i = 1, \cdots, k)$ の $rdepth(\boldsymbol{\theta}, Z_n)$ を最大にするとき，

$$DR(Z_n) = (\overline{\theta}_{1\cdot}, \overline{\theta}_{2\cdot}).$$

右の図は手計算が可能な非常に少ないデータ数を用いて分析を行なった．これより定義 4 に基づいた Exact アルゴリズムの計算結果はデータより下に回帰直線がきてしまい，近似アルゴリズムでの推定量と大きく異なる結果になった．この例以外にも Exact アルゴリズムの場合，データに適合しない結果を生じた．そのため，最深回帰推定量の定義を下記のように提案する．



定義 5　$\binom{n}{2}$ 個の 2 データ点を通る直線の中で，直線 $\theta_{1i}x + \theta_{2i}$ の $rdepth(\boldsymbol{\theta}, Z_n)$ を最大にするとき $(i = 1, \cdots, k)$,

$$DR(Z_n) = (\text{med } \theta_{1\cdot}, \text{med } \theta_{2\cdot}).$$

4　シミュレーション

最深回帰推定量 (DRE) の性能評価をするために，平均 2 乗誤差 (MSE) を調べ，最小 2 乗推定量 (LSE)，Rank 推定量，LMS(Least Median of Squares) 推定量との比較をする．また，最深回帰推定量については計算方法の違いを調べるため，Exact アルゴリズム，近似アルゴリズム，新しく提案した方法の比較を行なう．モデルは $y = \theta_1 x + \theta_2 + \varepsilon$，ただし，$\varepsilon \sim N(0, \sigma^2)$，$x \sim U(-1, 1)$，$(\theta_1, \theta_2) = (0, 0)$．外れ値をデータ数の 10%と 30%を $\varepsilon \sim N(0, c^2\sigma^2)$ とした $(\sigma = 1, c = 3)$．また，データ数は小標本を対象としているので，$n = 10, 15, 20, 50$ の計算を行なった (繰り返しは 10,000 回)．

外れ値 10%の分析結果は，最深回帰推定量について提案した方法が近似アルゴリズムよりも MSE は小さくなり，Exact アルゴリズムと比較しても大差はなかった．推定量全体については，Rank 推定量の MSE が最小になった．表 1 は外れ値 30%の分析結果である．これより，最深回帰推定量については外れ値 10%と同様の結果が得られ，推定量全体については，データ数が $n = 10, 15$ のときに最深回帰推定量の MSE が最小になった．したがって，シミュレーション実験により，中央値を利用した方法は小標本データの場合において，従来の方法と大差がなく，MSE が最小となった．

表 1: MSE of Slope (Outlier 30%)

n	LSE	RANK	LMS	DRE(Exact)	DRE(New)	DRE(Approx.)
10	9.3802	2.8195	5.2417	2.7628	2.5403	3.8751
15	6.5220	1.5017	1.4294	1.2174	1.1853	1.4621
20	4.2252	0.7753	0.9041	0.6378	0.6403	0.8511
50	1.5868	0.2315	0.3283	0.2012	0.2043	0.2652

適応型計画における推定

アステラス製薬株式会社
開発本部 データサイエンス部
伊藤雅憲

1. はじめに

近年，臨床試験計画の方法の一つとして，試験途中での計画の変更を許容する柔軟な計画法が話題に上がっている．これは「アダプティブ・デザイン」の呼称で普及しており，新薬開発の迅速化・効率化が声高に叫ばれるなか，実地での適用の是非はともかくとして，様々な場面で議論の対象となっているようである．FDA (米国食品医薬品局) は”Innovation or Stagnation? -Challenge and Opportunity on the Critical Path to New Medical Products” のタイトルで 2004 年 3 月にレポートを発行しており，これに関連して 2006 年 3 月に発行された”Critical Path Opportunities List”では新薬開発の迅速化を図るための様々なツールが掲載されている．この中の一つとして適応型計画が俎上に上がっている．

何らかの仮説を検証することを目的として実施される臨床試験は，通常，試験計画段階において検定の有意水準，検出力，想定される有効サイズ及びそれらから導かれる必要症例数を「固定」し，目標症例数に達するまで計画を変更することなく遂行されるが，この方法では，次のような方策をとることができない：①試験薬の有効性が非常に優れていた場合に，早期に試験を終了する，②試験薬が無効であった場合に，早期に試験を中止し，試験に組み入れられた患者に対するリスクを低減する，③試験薬の安全性に重大な問題があった場合に，早期に試験を中止し，試験に組み入れられた患者に対するリスクを低減する，④想定していた試験薬の有効サイズに対する見込みが違っていたり，設定した試験薬の用量が有効な範囲から外れていた場合に，妥当な試験計画に変更する．①，②，③を可能とする方法として群逐次計画があるが，一般に群逐次計画は中間解析後の試験計画変更を許容しないため，④を達成することができない．

一方，適応型計画は，①，②，③，④をすべて可能にするものである．とくに，臨床試験において中間解析を 1 回だけ実施し，その前後をステージ 1，ステージ 2 と呼称し，ステージ 1 の結果に基づいてステージ 2 の試験計画を再構築し，ステージ 1 とステージ 2 の結果を総合的に検討する計画は 2 ステージ・デザインと呼ばれる．本報告では，この計画法において，関心のあるパラメータに対する統計的統合推定法に焦点をあてる．

2. 統合推定

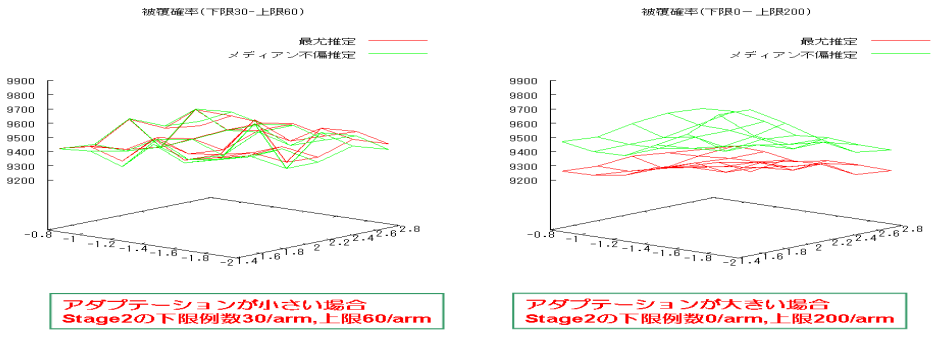
中間解析で症例数の再設定を実施する場合に，Brannath *et al.* (2006) では，関心のあるパラメータの全体平均（ステージ 1 とステージ 2 を併合した平均）について，最尤推定にバイアスが生じることを論じている．関心のあるパラメータ μ に対して，ステージ 1 とステージ 2 の平均値を $\bar{x}_1, \bar{x}_2$，事前に設定した症例数を n_1, n_2，ステージ 2 の再推定された症例数を $\tilde{n}_2$ とする．$\tilde{r} = \tilde{n}_2/n_1$ とするとき，単純な μ の最尤推定値は $\bar{x} = (n_1\bar{x}_1 + \tilde{n}_2\bar{x}_2)/(n_1 + \tilde{n}_2) = (\bar{x}_1 + \tilde{r}\bar{x}_2)/(1 + \tilde{r})$ となる．ここで $\bar{x} = \tilde{v}\bar{x}_1 + (1 - \tilde{v})\bar{x}_2 \ \ [\tilde{v} = 1/(1 + \tilde{r})]$ とおくと，$\tilde{n}_2$ を与えたもとでの $\bar{x}_2$ の条件つき平均は μ となるので，次の関係式 $\mathrm{E}_\mu(\bar{x}) - \mu = \mathrm{COV}_\mu(\tilde{v}, \bar{x}_1)$ が成り立つ．例えば，ステージ 1 の結果から，予想よりも平均値が増加し，ステージ 2 の必要症例数が減少した場合，共分散は正値をとり，した

がってバイアスが生じることになる．このバイアスをとり除く方策としては，推定量の構成において，事前に計画していたステージ 2 の症例数 n_2 を用いて不偏推定量を構成することが挙げられる．Brannath *et al.* (2006) は，情報量に基づく重み w_i を設定し [例えば，$w_i = \sqrt{n_i/(n_1+n_2)}$]，メジアン不偏推定量として $\hat{x}_m = \tilde{u}\bar{x}_1 + (1-\tilde{u})\bar{x}_2$ $[\tilde{u} = w_1\sqrt{n_1}/(w_1\sqrt{n_1} + w_2\sqrt{\tilde{n}_2})]$ を提案している．

3. 数値実験

上述した統合推定量の性能を検討するために，2 ステージデザインにおける簡単なシミュレーションを行った．解釈を簡単にするため，プラセボと試験薬の 2-arm 試験デザインとし，試験薬－プラセボのエフェクトサイズを 1.4，標準偏差 2.1 に想定した．また，それぞれ 1.2～2.0，1.8～2.7 までずらした場合も検討した．上記の仮定で，α =0.05，1- β =0.9 で固定デザインで設計するとおよそ 1 群あたり 60 例が必要となる．そこで，その半分を Stage1 に設定した．Stage1 で各群 30 例の正規乱数を生成し，得られたデータから統合検定で棄却限界値片側 0.0038（有意水準片側 0.025 に相当）より p 値が小さくなるための検出力を 90%として，症例数再算定を実施することとした．その際，Stage2 の症例数には次の二通りの制限を設けた：①下限 30 例，上限 60 例　②下限 0 例（Stage1 で終了有り），上限 200 例．このようなシミュレーション計画のもとで，最尤推定量とメジアン不偏推定量の性能を比較した．シミュレーション実験を 10,000 回反復し，真のエフェクトサイズからの平均平方誤差によって性能を評価した．さらに，各推定法において 95%信頼区間を構成し，真値の被覆確率を算出した（95%に近い方が妥当な信頼区間とみなした）．

以下，今回のシミュレーションによる限られた結果であるが，平均平方誤差の結果はほとんど変わらなかった．バイアスはメディアン不偏推定の方が小さかったが，分散のオーダーからするとほとんど影響しない．信頼区間については，アダプテーションが小さい（Stage2 の下限 30/arm，上限 60/arm）場合にはいずれも 9500 回付近で真値を被覆し，二つの推定法の結果に大きな差はなかった．しかし，アダプテーションが大きい（Stage2 の下限 0/arm，上限 200/arm）場合には最尤推定での通常の信頼区間の被覆確率はすべて 92%から 94%に分布し，95%に満たなかった．一方，メディアン不偏推定でのフレキシブル信頼区間では，概ね 95%を保っていた．



95%信頼区間の真値被覆確率の結果

構造方程式モデルを用いた因果推論

大分大学工学部知能情報システム工学科 和泉志津恵, 原 恭彦, 小畑経史

1. はじめに

構造方程式モデル(Structural Equation Model, SEM)とは，共分散構造モデル(Covariance Structure Model, CSM)とも呼ばれ，観測された変数間の分散・共分散の構造を分析する方法である。SEM では，直接観測できない潜在変数（構成概念）を導入し，潜在変数と観測変数（従属変数，独立変数）との間の因果関係を同定する。これまでのパス解析を含めた多重回帰分析と因子分析を拡張したものと捕らえることができる。

2. 構造方程式モデル

潜在変数を伴う構造方程式モデルにおいて，測定方程式（Measurement equation）は，i 番目の個体に対して，$y_i = \nu + \Lambda\eta_i + Kx_i + \varepsilon_i$ と表す。ここで，y_i は観測された連続型従属変数ベクトル，x_i は観測された独立変数ベクトル，η_i は観測されない潜在変数ベクトル，ε_i は誤差ベクトル （但し，$\varepsilon_i \sim N(0,\Theta)$），$\nu$ は切片ベクトル，Λ, K は回帰係数パラメータ行列とする。ある潜在変数が他の潜在変数によって説明されることを，構造方程式（Structure equation），$\eta_i = \alpha + B\eta_i + \Gamma x_i + \zeta_i$ で表す。ここで，η_i は観測されない潜在変数ベクトル，x_i は観測された独立変数ベクトル，ζ_i は誤差ベクトル（但し，$\zeta_i \sim N(0,\Psi)$），α は切片ベクトル，B, Γ は回帰係数パラメータ行列（但し，$diag[B]=0$）とする。このとき，期待値と分散共分散行列は，それぞれ

$$E(y_i \mid x_i) = \nu + \Lambda(I-B)^{-1}(\alpha + \Gamma x_i) + Kx_i,$$

$$V(y_i \mid x_i) = E\left[\left(y_i - E(y_i \mid x_i)\right)\left(y_i - E(y_i \mid x_i)\right)'\right] = \Lambda(I-B)^{-1}\Psi(I-B)^{-1'}\Lambda' + \Theta$$

となる。

観測された従属変数が 2 値の場合，i 番目の個体の j 番目の従属変数に対して $y_{ij} = 1\ (if\ y_{ij}^* > \tau_j), 0\ (otherwise)$ を満たす連続変数 y_{ij}^* の存在を仮定する。そして，y_{ij}^* の分布関数 $F(.)$ を用いて，$P(y_{ij}=1 \mid x_i) = P(y_{ij}^* \geq \tau_j \mid x_i) = 1 - F(\tau_j)$ となるプロビットモデルを考える。

従属変数が連続値をとる場合，最尤法によって母数の推定を行うが，2 値の場合には，重み付き最小二乗法によって母数を推定する。モデルの適合度の指標として，カイ二乗値，Tucker-Lewis 指標(TLI)，比較適合度指標(CFI)，RMSEA 等を用いる。SEM で取り扱う主なモデルとして，多重指標モデル(Multiple Indicator Multiple Cause Model)があげられる。これは，図 1 のように複数の観測変数によって潜在変数が説明され，その潜在変数が複数の観測変数の原因となるものである。

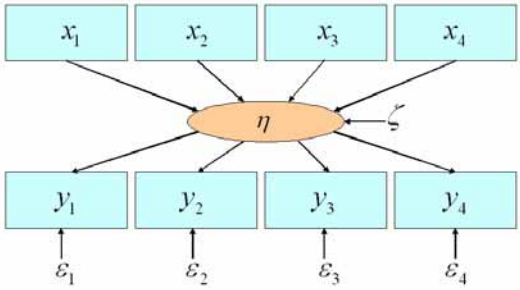


図１. MIMIC モデルの例

3. データ解析例

構造方程式モデルを用いて，２値を含むカテゴリカルな従属変数をもつ交通事故データの解析例を示す。2001～2006 年に大分県内で発生した人身事故において，第１当事者（交通事故に関与した者のうち，過失が重い方）となった運転者の約３万人の記録を用いた。過去の交通心理学的研究に基づいて，危険運転傾向と人為的エラーの MIMIC モデル(図２)を検討した。

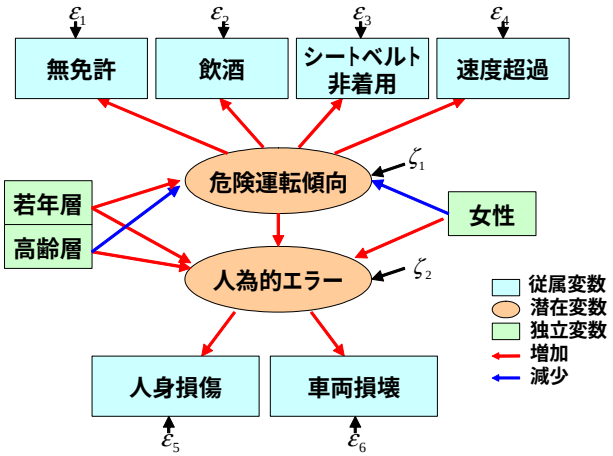


図２．運転者の危険運転傾向と人為的エラーのモデル

データ解析の結果，CFI，TLI，RMSEA 等のモデル適合度指標から上記のモデルへのデータのあてはまりが良いと示唆された。また，性別や年齢という運転者の背景因子や昼夜，天候，道路形状等の交通環境を説明変数に取り入れたモデルも検討した。その結果，速度超過やシートベルト非着用などの交通違反に顕著に表れる危険運転傾向は，男性や若年層の運転者が，夜間に単路やカーブを運転する際に強まる，また，事故（人身損傷，車両損壊）をもたらす人為的エラーは，女性や高齢層の運転者が，日中に交差点やカーブを運転する際に強まることが示唆された。

今後の課題として，交通事故のデータ中において今回検討されなかった背景因子を含めた総合的な検討が必要であり，漸近的分布非依存法（Asymptotically Distribution Free)による母数の推定値との比較やモデル適合度の検討も考えられる。

関数クラスター解析の背面形状分類への応用

（株）松下電工解析センター　藤原　ゆり
松下電工（株）　小林美佐世
山梨大学大学院医学工学総合研究部　下川　敏雄
特定非営利活動法人　医学統計研究会　後藤　昌司

１．背景と目的

現在、弊社のショウルームや弊社が経営するフィットネスクラブであるバランススタジオ RINTO で背面形状診断機を導入、運動機器ジョーバなどによる姿勢改善効果の検証に活用している。

現行の背面形状診断システムでは、視覚的分類である Staffel 分類を背面形状の凹凸部距離や角度に CHAID（カイ 2 乗統計量に基づく自動交互作用検出）を適用し、表現した定量的な分類アルゴリズムを用いている。本報告では、現行アルゴリズムでの距離、角度算出による関数曲線の情報量の損失を低減したより定量的で高精度な分類にするため、関数クラスター解析を背面形状データに応用し、その適用性を検討した。

２．アプローチ

関数クラスター解析の諸種の方法から最適な手法を選定するため、以下の項目について、1004 件の背面形状データを用いて比較検討した。

①背面形状データの曲線近似法

9 次関数近似、スプライン近似

②関数データにクラスター解析を適用する方策（個体間の Euclid 距離の算出法）

関数の積分値、関数式の係数

③クラスター数の最適化

視覚的な判断、統計指標（Euclid 距離総和、GAP 統計量、Silhouette 統計量）

３．結果

今回検討した関数クラスター解析の諸種の方法で、以下がよいと判断した。

①背面形状データの曲線近似法

分類した際のクラスターのまとまりがよく、背面形状の前後傾き、凹凸形状の特徴を分類可能という点で、9 次関数近似を用いる方がよい。

②関数データにクラスター解析を適用する方策（個体間の Euclid 距離の算出法）

分類した際の背面形状の前後傾き、凹凸形状の特徴を分類可能という点で、積分値を用いる方がよい。

③クラスター数の最適化

視覚的な判断、各統計指標での最適クラスター数は以下のようになり、総合すると最適クラスター数は 7 クラスターとなった。

視覚的な判断：7 または 8 クラスターが最適

Euclid 距離総和：6 クラスター以上がよい

GAP 統計量：意味のある考察はできず（最適クラスター数を導出できず）

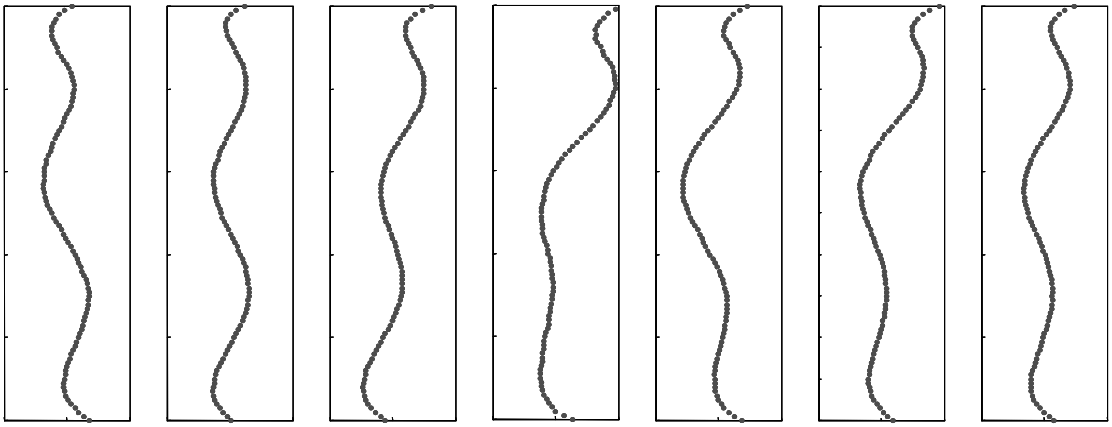
Silhouette 統計量： 7 または 10 クラスターが最適

上記①～③で決定した手法を用いて分類した 7 分類の背面形状の判別精度を、手法検討の際に用いたものと同じ 1004 件の背面形状データを用いて算出した。今回用いた判別方法は、 7 分類の各クラスターの代表とした中央値の形状との差の積分値に判別する方法とした。結果、判別精度は 96.5%と従来の CHAID を用いた分類法に比べ、大幅に向上した。

４．結論

関数クラスター解析を背面形状分類へ適用するにあたり、諸種の方法を検討した結果、以下の方法で適用すると、背面形状を図のような特徴的な形状に分類でき、判別精度は 96.5%と非常に高いことが分かった。

近似法：9 次関数近似、クラスター解析の適用の方策：積分値、クラスター数： 7



クラスタ1 後傾き S字　クラスタ2 S字　クラスタ3 出尻（凹背）　クラスタ4 高齢者 形状　クラスタ5 猫背（円背）　クラスタ6 高齢者 気味猫背　クラスタ7 前傾き S字

５．今後の課題

今回検討した 1004 件の背面形状データだけでなく、新たに計測されたデータにも対応できる精度のよい背面形状の判別アルゴリズムの作成が必要である。また、新たな背面形状分類法での背面形状診断のシステム化が早急に望まれる。

さらに、クラスター間の距離（遠さ近さ）を用いた背面形状診断カルテの MAP 化、背面形状の点数化により、診断システム利用者が診断結果を理解やすくなるようにしたい。あわせて、運動効果による姿勢改善の把握、腰痛や肩こりなどの不定愁訴と姿勢診断との関係の把握により、新たな運動機器開発、運動提案に活用したい。

関数データの回帰関数とその導関数の推定について

Dou Xiaoling　白旗慎吾　坂本　亘　　大阪大学　大学院基礎工学研究科

1. はじめに

関数データ解析においてはデータの回帰関数だけでなく, それらの導関数の推定もますます重視されてきている. B-Spline 基底関数で導関数を推定する方法は Ramsay & Silverman (2005) と Ferraty & Vieu (2006) によって提案されている. しかし, B-Spline 基底関数だけではなく, kernel 関数による導関数の推定方法 (Eubank, 1999) も広く応用されている. 本稿では, 私たちは B-Spline 基底関数法と kernel 法による回帰関数とその導関数の推定量について, それらの性能を調べ, 比べる.

2. 回帰関数と導関数の推定

n 個の観測値 $(t_1, y_1), ..., (t_n, y_n)$ が与えられ, モデル $y_i = x(t_i) + \varepsilon_i,\ \ i = 1, ..., n$ に従うと仮定する. ベクトルで $\boldsymbol{y} = x(\boldsymbol{t}) + \boldsymbol{\varepsilon}$ のように書ける. 誤差を表す $\varepsilon_i,\ i = 1, ..., n$ は平均ゼロ, 共通な分散 σ^2 を持つ独立な確率変数であり, $t_i,\ \ i = 1, .., n$ は非確率的なデザインポイントで, $a \le t_1 < t_2 < ... < t_n \le b$ と仮定する.

2.1 B-Spline 基底関数による推定

関数 $x(t)$ に対して, B-Spline 基底関数の展開を用いて $\hat{x}(t) = \sum_{k=1}^{K} \hat{c}_k \phi_k(t) = \hat{\boldsymbol{c}}^T \boldsymbol{\phi}(t) = \boldsymbol{\phi}^T(t)\hat{\boldsymbol{c}}$ を推定する. 長さ K のベクトル $\boldsymbol{c}$ は係数 c_k を含んである. $\boldsymbol{\phi}(t)$ は B-Spline 基底関数のベクトルである. Ramsay & Silverman (2005) では多くの場合においてはいい結果を導き, 特に導関数の推定には有効である「強力な」手法乱雑度ペナルティーアプローチ (または正則化アプローチ) を紹介してある. そこでは推定したい曲線の乱雑度の罰則, つまり, その関数の 2 階導関数の 2 乗の積分を考慮に入れて, 曲線を推定している. 定義されるペナルティ付残差 2 乗和

$$\begin{aligned}
\mathrm{PENSSE}_d(x) &= \sum_{i=1}^{n} \{y_i - x(t_i)\}^2 + \lambda_d \int (D^{d+2}x(t))^2 dt \\
&= \{\boldsymbol{y} - \boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{t})\}^T \{\boldsymbol{y} - \boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{t})\} + \lambda_d \int \left[D^{d+2}\boldsymbol{c}^T\boldsymbol{\phi}(t)\right]^2 dt \\
&= \{\boldsymbol{y} - \boldsymbol{\Phi c}\}^T \{\boldsymbol{y} - \boldsymbol{\Phi c}\} + \lambda_d \boldsymbol{c}^T \boldsymbol{R}_{d+2} \boldsymbol{c}
\end{aligned}$$

を最小化する係数ベクトルは $\hat{\boldsymbol{c}}_d = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda_d \boldsymbol{R}_{d+2})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$ と求められる. ここで, $D^m x(t)$ は関数 $x(t)$ の m 階導関数を表し, $\boldsymbol{R}_{d+2} = \int D^{d+2}\boldsymbol{\phi}(t) D^{d+2}\boldsymbol{\phi}^T(t) dt$ は $K \times K$ 行列で, $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_k(t_j)\}$ は $n \times K$ 行列である. d は求めようとする導関数の階数を表す, 本稿では $d = 0,\ 1,\ 2$ に注目する. λ_d は d 階導関数の平滑化パラメータである. このように, 罰則付き B-Spline 平滑化で得られる関数データの回帰曲線, 1 階導関数と 2 階導関数はそれぞれ

$$\hat{x}(t) = \boldsymbol{\phi}^T(t)\hat{\boldsymbol{c}}_0,\ \ D\hat{x}(t) = D\boldsymbol{\phi}^T(t)\hat{\boldsymbol{c}}_1,\ \ D^2\hat{x}(t) = D^2\boldsymbol{\phi}^T(t)\hat{\boldsymbol{c}}_2 \tag{1}$$

になる. よって, デザインポイントでデータの推定値, 変化率, 及び加速度の推定値ベクトルはそれぞれ, $d = 0, 1, 2$ のときに

$$\hat{x}^{(d)}(\boldsymbol{t}) = \boldsymbol{\Phi}^{(d)}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda_d \boldsymbol{R}_{d+2})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} = \boldsymbol{S}_{\phi,\lambda}\boldsymbol{y} \tag{2}$$

のように求められる.

平滑化パラメータの選択方法については一般化クロスバリデーション (GCV) の手法はよく用いられる. Ramsay & Silverman (2005) により GCV スコアは以下のように定義される.

$$\mathrm{GCV}(\lambda) = \frac{n^{-1}\sum_{i=1}^{n}\{y_i - \hat{x}(t_i)\}^2}{n^{-1}\mathrm{trace}(\boldsymbol{I} - \boldsymbol{S}_{\phi,\lambda})^2} = \frac{n\{\boldsymbol{y}^T[\boldsymbol{I} - \boldsymbol{S}_{\phi,\lambda}]^2\boldsymbol{y}\}}{\{\mathrm{trace}[\boldsymbol{I} - \boldsymbol{S}_{\phi,\lambda}]\}^2}. \tag{3}$$

2.2 kernel 関数による推定

本稿のシミュレーションに用いられたカーネル法は Gasser-Müller 推定量と Plug-in 法で選ばれた全体的な bandwidth で構成した手法である. Plug-in 法のアルゴリズムは以下のよう要約できる：あるとても小さな bandwidth の初期値 $\hat{h}_0$ を設定することからはじめ, $\hat{h}_{i-1}$ を求めたい回帰関数（または, 導関数）の 2 階導関数に代入して, bandwidth の式で「最適な」bandwidth $\hat{h}_i$ を求める. 何回かの反復で, $\hat{h}_i$ が収束し, $\hat{h}_{\mathrm{opt}}$ が得られる.

3. シミュレーションによる評価

本節では, B-Spline 基底関数法と kernel 法で回帰関数とそれらの導関数の推定を比較, 評価する. 前者は基底関数の次数 (degree) は 6 とする (order=7), 内部節点の位置を等間隔にしながら, 節点の個数を動かし, GCV Method で平滑化パラメータを求めた. 後者は使用する kernel 関数は多項式 kernel 関数 Gasser, Müller & Mammitzsch (1985) であり, Global Plug-in Method で全体的な bandwidth を求めた.

評価の基準については, まず, 推定された曲線全体の 2 乗誤差の平均

$$\mathrm{SSE}(\hat{x}^{(d)}) = \frac{1}{n}\sum_{i=1}^{n}(x^{(d)}(t_i) - \hat{x}^{(d)}(t_i))^2 \qquad (d = 0,\ 1,\ 2) \tag{4}$$

及び導関数の中央 70% の 2 乗誤差の平均, 例えば, $n = 100$ の場合

$$\mathrm{SSE}_{70}(\hat{x}^{(d)}) = \frac{1}{70}\sum_{i=16}^{85}(x^{(d)}(t_i) - \hat{x}^{(d)}(t_i))^2 \qquad (d = 1,\ 2) \tag{5}$$

を求める. N 本のシミュレーション曲線に対して回帰関数, 導関数を求めて, それらの SSE や SSE_{70} などの平均

$$\mathrm{MSSE} = \frac{1}{N}\sum_{j=1}^{N}\mathrm{SSE}(\hat{x}^{(d)}), \qquad\qquad \mathrm{MSSE}_{70} = \frac{1}{N}\sum_{j=1}^{N}\mathrm{SSE}_{70}(\hat{x}^{(d)}) \tag{6}$$

で推定手法の性能を評価する. また, 平滑化パラメータや bandwidth についても考察する.

シミュレーションに用いられたすべての関数を $y_i = x(t_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$ に従うとし, 等間隔なデザインポイント $t_i \ \ (i = 1, ..., n; n = 100)$ で観測される. また, 乱雑なデータの導関数を考えるため, 誤差の標準偏差を変えてみた.

シミュレーションによって, B-spline 法は関数定義域の中央 70% では, kernel 法よりうまく働くことが分かった. 境界領域では, 特に導関数の推定について, B-spline 法は非常に不安定な結果を与える. これに対して, kernel 法で求めた推定値はそれほど大きく変動しない.

4. まとめと今後の課題

本稿では, B-spline 基底関数法と kernel 法の回帰関数とそれらの導関数の推定をシミュレーションによって比較を行った。結果として, 等間隔に内部節点を設置する B-Spline 基底関数法は定義域の中央部分では kernel 法よりはよく当てはめているが, 端の近くでの当てはまりはあまりよくないことが分かった. B-spline 法の定義域の端で安定する値を得られるため, 節点の個数と位置を調節して回帰関数と導関数を推定する必要がある.

関数データ解析の数理・応用・展望

北海道大学　情報基盤センター
先端データ科学研究室　水田 正弘

１．はじめに

「データ」を解析し、そこから何らかの情報を得ることは、科学一般において普遍的な作業である。ここで、「データ」としてどのようなものを想定するかは、解析における目的や利用できる環境（特にコンピュータ環境）に依存する。コンピュータ環境が不十分な時代には、「データ」をいわゆる１次元の数値として扱うことしかできなかった。しかし、近年では、大量の超高次元の数値として扱うことが可能になってきた。

さらに、単なる多次元の数値ではないデータを想定する方法が開発されてきた。空間データ、時系列データ、シンボリックデータなどがある。また、Ramsay および Silverman などにより 1990 年ころから「関数データ解析法」が提案され、多くの研究者により推進されている。関数データ解析法に関する成書(Ramsay and Silverman, 1997, 2005)および応用をまとめた本（Ramsay and Silverman, 2002）が基本的な資料といえる。

本報告では、関数データ解析法の概要および今後の課題について紹介した。

２．関数データ解析の数理

通常のデータと関数データを比較検討する。簡単のために、通常のデータとして n 個の p 変量データ $\mathbf{x}_i \in \mathbf{R}^p \ i=1,2,\cdots,n$ 、関数データとして積分可能な n 個の一変数関数 $x_i(s)\ i=1,2,\cdots,n$ を考える。確率構造を入れなければ、通常のデータは、p 次元空間における n 個の点であり、関数データは無限次元空間における n 個の点ととらえることができる。また、p 変量データではベクトルのノルム、関数データでは L^2 ノルムを利用することで内積が定義でき、データ点間の距離を表現できる。

関数データが得られたとき、コンピュータなどを利用して解析するためには、関数データを有限個の数値で表現しなくてはいけない。関数データ解析における多くの研究では有限個の(正規直交)基底関数を利用して関数データを近似している。これにより関数データは、利用する基底関数の個数と同じ次元をもつ多変量データとなる。すなわち、従来からある多変量データに対する解析法がそのまま使える。

離散データの関数化および、関数データを有限個の数値パラメータで表現する方法の両者において、基底関数の選び方、利用する基底関数の個数の選び方が問題となる(荒木・小西, 2004 など)。

関数データに確率構造を入れる考え方はいくつかある。最も簡単なのは、関数データを基底関数などにより有限個の数値で表現することにより、通常の有限次元空間における確率構造と同様な議論ができる。別のアプローチとして、Random Function (Lifshits,

1995)の考え方を利用することができる。

４．関数データ解析の応用

関数データ解析法の応用例は、数多く報告されている。特に、Ramsay and Silverman (2002)では、骨の形状、人間の成長、手書き文字、など数多くの応用例が掲載されている。

また、本堂・南・白土・水田(2004)は、関数主成分分析を用いて動体追跡照射データの解析結果を報告している。動体追跡放射線照射とは、体内にある腫瘍に対し、その位置を追跡しながら適切なタイミングで放射線を照射する方法である。位置を知るために直径 2㎜程度の金マーカを利用している。金マーカの動きを、3 次元空間を値域とする関数データとみなして解析した。

５．関数データ解析の展望

関数データ解析に関する研究は、非常に活発に実施されており、COMPSTAT や ISI をはじめとする国際会議、論文、成書で多くの研究報告がなされている。詳しくは、http://www.psych.mcgill.ca/misc/fda/にある Bibliography および Conferences などを参考にされたい。

関数データ解析と(多分)独立に提案され、発展してきた新しいデータの捉え方として Diday(1987)らによるシンボリックデータがある。これは、従来のデータ構造の枠組みを一般化し、多様なタイプのデータを許容するシンボリックデータを定義し、それを解析する方法である。シンボリックデータ解析に関しては2つの成書、Billard and Diday (2006), Bock and Diday eds.(2000)および、電子ジャーナル
http://www.jsda.unina2.it/newjsda/index.htm が参考になる。

６．おわりに

本報告では、関数データ解析について、数理的な側面、応用面、今後の展望の 3 つの観点から考察した。データを関数として扱うことの長所・短所を考慮し、実際に役立つ解析法の開発が本質的だと思われる。また、シンボリックデータ解析法など新しいアプローチとの融合も重要な課題である。

一般化線形モデルで捉えたROC曲線の推測

大江基貴*・坂本 亘*・白旗慎吾*・後藤昌司†

1　序に代えて

臨床医学分野では，診断過程において検査結果から疾病の有無を的確に判断することが求められる．診断の確度は有病率に影響を受けない以下の 2 つの指標で与えられる．

- 感度　：対象疾患にかかっている患者を正しく診断・検出する能力.
- 特定度：対象疾患にかかっている患者を正しく診断・排除する能力.

一般にこれらの指標はトレード・オフの関係にある．これらを別々に評価することは誤った解釈を導く恐れがあり，同時に評価することが重要である．(感度，(1 − 特定度)) をプロットすることによって与えられる受信者動作特性 (Receiver Operating Characteristic:ROC) 曲線は，感度と特定度を同時に評価することができ，診断の確度を評価するのに用いられる．本稿で紹介する ROC-GLM では，ROC 曲線そのものを一般化線形モデル (Generalized Linear Model:GLM) の枠組みで捉え，モデルの適合を標準的な過程で接近する．この接近法では，パラメータの推定に関して 2 状態の分布を必要としないので，ROC 曲線の本質に沿った接近法であるといえる (Pepe, 2004).

2　診断とその確度

一般に診断は，以下の過程で行われる．

主訴・問診によって病歴を調査　→　診察による理学的な所見　→　鑑別診断

→　疑わしき疾患に対応した臨床検査　→　確定診断

ほとんどの疾患は病歴と理学的所見で診断がつくことが多いが，臨床検査による検査結果から最終的に診断をくだすことになる．この検査結果は連続的あるいは離散的な尺度で表されることもあるが，最終的には「しきい値」により陽性，すなわち「疾患あり」と陰性，すなわち「疾患なし」に 2 分される．

診断過程における検査結果 Y は連続的な尺度で与えられるとし，対象疾患にかかっている患者の検査結果を Y_D ，そうでない患者の検査結果を $Y_{\bar{D}}$ と表す．これらはそれぞれ生存関数 S_D ，$S_{\bar{D}}$ をもつ未知の分布に独立に従うとする．診断では，しきい値 c によって，$Y > c$ ならば陽性，そうでなければ陰性と判定される．ROC 曲線は偽陽性割合 FPF に対する真陽性割合 TPF のプロットを，可能なすべてのしきい値で求めた曲線である．ここに

$$\mathrm{TPF}(c) = S_{\bar{D}}(c), \tag{1}$$

$$\mathrm{FPF}(c) = S_D(c) \tag{2}$$

*大阪大学大学院 基礎工学研究科 システム創成専攻 数理化学領域
†特定非営利活動法人 医学統計研究会

である．このとき，ROC 曲線は以下のように書ける．

$$\mathrm{ROC} = \left\{ \left(S_{\bar{D}}(c), S_D(c) \right), -\infty < c < \infty \right\}. \tag{3}$$

ここで，$\mathrm{FPF}(c) = t$ とおき，(1) と (2) から c を消去すれば，(3) は以下の形に書くことができる．

$$ROC(t) = S_D\left(S_{\bar{D}}^{-1}(t) \right), \quad 0 \leq t \leq 1.$$

3　一般化線形モデルで捉えた ROC 曲線

GLM で捉えた ROC 曲線は以下のように書ける．

$$g(ROC(t)) = \sum_{k=1}^{K} \alpha_k h_k(t). \tag{4}$$

ここに，g は連結関数，$\mathbf{h}(t) = \{h_1(t), \ldots, h_K(t)\}$ は特定の基底関数ベクトル，$\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_K\}^{\mathrm{T}}$ はパラメータである．いま，疾患患者，非疾患患者に対する検査結果が，それぞれ n_D と $n_{\bar{D}}$ と与えられるとする．ここに，$\{Y_{D_i}, i = 1, \ldots, n_D\}$ と $\{Y_{\bar{D}_j}, j = 1, \ldots, n_{\bar{D}}\}$ は，それぞれ生存関数が S_D と $S_{\bar{D}}$ の未知母集団に独立に分布していると仮定する．いま，$B_{ij} = I[S_{\bar{D}}(Y_{D_i}) \leq t_j]$ を定義すると，これは 2 値データである．ここに，$t_j = S_{\bar{D}}(Y_{\bar{D}_j})$ である．その期待値は ROC 曲線の観測値として捉えることができる．

$$\mathrm{E}\{B_{ij}\} = \Pr[S_{\bar{D}}(Y_D) \leq t_j] = ROC(t_j).$$

これより，(4) において 2 値データ B_{ij} に対する GLM の適合の手順を利用することが可能である．推測に際して，$S_{\bar{D}}$ の代わりに経験生存関数 $\hat{S}_{\bar{D}}$ を用いる．このとき，$\hat{B}_{ij} = I[\hat{S}_{\bar{D}}(Y_{D_i}) \leq t_j]$ と書くことができる．

このとき，尤度方程式は以下のように与えられる．

$$\sum_{j=1}^{n_{\bar{D}}} w(t_j) \left(y_j - ROC(t_j) \right) \left(\frac{\partial \eta_j}{\partial \mu_j} \right) \mathbf{h}^{\mathrm{T}}(t_j) = \mathbf{0}. \tag{5}$$

ここに

$$y_j = \frac{1}{n_D} \sum_{i=1}^{n_D} \hat{B}_{ij}, \tag{6}$$

$$\eta_j = \sum_{k=1}^{K} \alpha_k h_k(t_j), \tag{7}$$

$$ROC(t_j) = g^{-1}(\eta_j) = \mu_j, \tag{8}$$

$$w(t_j) = \left(\frac{\partial \mu_j}{\partial \eta_j} \right)^2 \left\{ ROC(t_j)(1 - ROC(t_j)) \right\}^{-1} \tag{9}$$

である．(6) は ROC 曲線の観測値，(7) は線形予測子，(8) は線形予測子が与える系統的な y_j の平均，(9) は重み関数である．

(4) のパラメータの最尤推定量は，(5) をパラメータ $\boldsymbol{\alpha}$ について解く，すなわち Fisher のスコア法に基づく重みつき最小 2 乗法の反復過程により近似的に得られる．

疾病の地域集積性の検出について

岡山大学大学院環境学研究科　石岡文生
岡山大学大学院環境学研究科　栗原考次

1. はじめに

近年、環境リスク解析や環境保全のため、空間データ解析の必要性が高まっている。中でも、ある郡における病気の発生率などのように、領域毎に得られるデータに対して、有意に高いまたは低い値を示す地域（ホットスポット）の検出は、各種の空間データの大きな課題である。ホットスポット領域の検出手法として、これまでに様々な手法が提案されてきた。そんな中、我々は Echelon 解析（Myers et al., 1997）を利用することによって作成される位相的な階層構造に基づいてホットスポットを検出する手法を提唱している。本研究では、病気の発生率のような地域空間データに対して Echelon 解析と空間スキャン統計量に基づいたホットスポットを検出する方法について紹介し、さらに他のホットスポット検出法との結果の比較を行う。

2. 空間スキャン統計量

空間スキャン統計量は、ある領域内の地点に起きた現象が偶然によるものか否かを検定し、有意に高い地域群（ホットスポット）を検出するための尤度比検定統計量であり、以下の式で与えられる。

$$\lambda = \frac{(c(Z)/n(Z))^{c(Z)}\{(c(G)-c(Z))/(n(G)-n(Z)\}^{c(G)-c(Z)}}{(c(G)/n(G))^{c(G)}}$$

ここで、$n(G)$ をすべての領域 G での母集団の数、$n(Z)$ を領域 Z 内での母集団の数、$c(G)$ をすべての領域 G で属性を持つものの数、$c(Z)$ を領域 Z 内で属性を持つものの数を意味する。より尤度の高いホットスポット領域を検出するためには、λ が大きな値をとるような領域 Z の決め方（スキャン法）が重要になる。これまでに、Circular scan 法（Kulldorff, 1997）、Upper level set scan 法（Patil and Taillie, 2004）、Simulated annealing scan 法（Duczmal and Assunção, 2004）、Flexible scan 法（Tango and Takahashi, 2005）などのスキャン法が提唱されている。

3. Echelon 解析を利用したホットスポット検出

病気の発生率のような地域空間データは、対象とする地域が市や郡などいくつかの区画 D_i, i=1, 2, …, k に分割され、データは $h(D_i)$で与えられる。例として、アメリカ合衆国ノースカロライナ州の乳幼児突然死症候群（Sudden Infant Death Syndrome; SIDS）データ（Cressie and Chan, 1989）を用いる。データは、ノースカロライナ州の 100 郡において 1974 年 7 月から 1978 年 6 月の期間に観測されたデータである。この種の地域空間データのような場合、領域間の近隣情報 $NB(D_i)$を与えることにより、Echelon 解析によって、その位相的な構造を階層構造で表す事ができる。Echelon 解析により SIDS データの位相的な構造を求め、得られた階層構造の分類に基づき領域をスキャンし、空間スキャン統計量が有意に大きな値を示す領域 Z を検出する。その結果、1 番目のホットスポット（Most likely cluster）は、13 領域となり、そのときの対数尤度は 16.506、p 値は 0.001 となった。また、2 番目のホットスポット（Secondary cluster）として、6 領域が得られ、そのときの対数尤度は 15.303、p 値は 0.001 となった。これらのホットスポット領域を図 1 に示す。

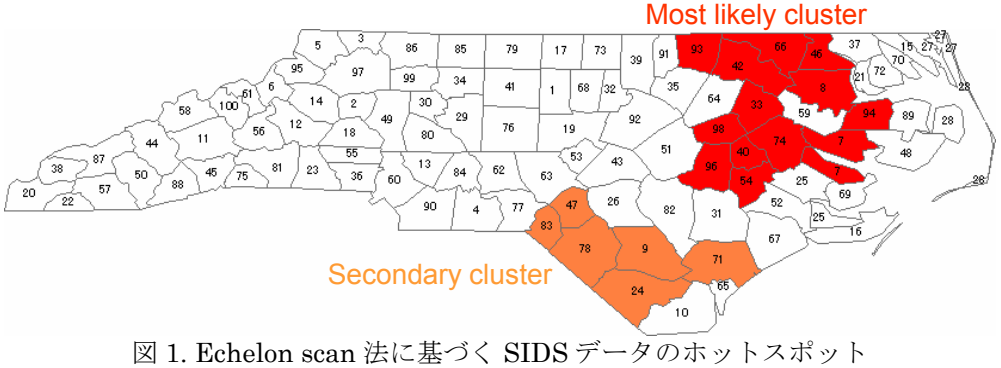


図 1. Echelon scan 法に基づく SIDS データのホットスポット

4. 他のホットスポット検出法との比較

SIDS データに対して、Circular scan 法 (Kulldorff, 1997)、Flexible scan 法 (Tango and Takahashi, 2005)によるホットスポット検出結果をそれぞれ図 2 に示す。

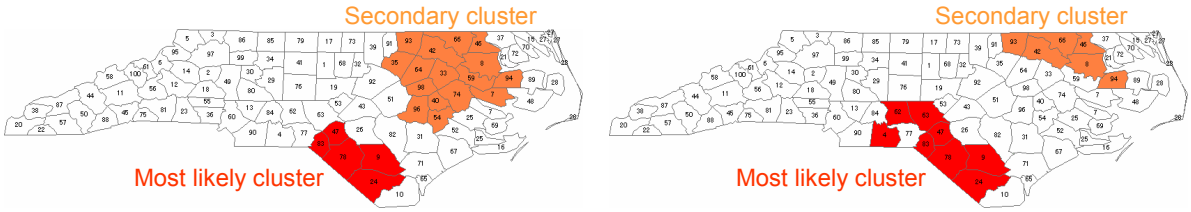


図 2. Circular scan 法（左図）と Flexible scan 法（右図）に基づく SIDS データのホットスポット

表 1 にこれらの手法と、我々の Echelon による手法の SIDS データのホットスポット検出結果を示す。いずれの手法の結果も、北部と南部にホットスポットが存在することが示唆された。北部のホットスポットでは、Echelon に基づく手法が、尤度の高いホットスポットを検出した。南部においては、Flexible scan による手法が尤度の高いホットスポットを検出した。

表 2 は、これら 3 つの手法の特性をまとめた結果を示している。Echelon に基づくホットスポット検出法は、(1)円形状に限らないホットスポット検出が可能。(2)空間データの構造に基づいたホットスポット検出が可能。(3)総当たりに領域をスキャンするのではなく、データの持つ階層構造のピークから優先的にスキャンするので効率がよい。(4)多次元空間データに対するホットスポット検出が容易になり、それと関連して時空間ホットスポットの検出が可能になった。これは、従来の手法では困難であった、時間推移とともに縮小するホットスポット、移動するホットスポット、分裂するホットスポットなどの連続する時間の中におけるホットスポット空間の推移を表現する事を可能にする。

表 1. 各スキャン法における SIDS データの北部と南部のホットスポット検出結果

北部	領域数	生誕数	SIDS 数	対数尤度	*p* 値
Echelon scan	13	36005	123	16.506	0.001
Circular scan	15	42006	131	12.585	0.001
Flexible scan	6	9763	49	15.968	0.001
南部	領域数	生誕数	SIDS 数	対数尤度	*p* 値
Echelon scan	6	17998	73	15.3025	0.001
Circular scan	5	16770	69	14.930	0.001
Flexible scan	8	22246	92	20.649	0.001

表 2. 空間スキャン統計量を用いた、スキャン別のホットスポット検出法の特性比較

	必要な空間情報	形状	高尤度	大量データ	計算コスト	時空間(多次元)	ソフト化
Echelon Scan	隣接情報	任意	○	◎	○	△	無
Circular Scan	距離情報	円形	△	○	○	△	有
Flexible Scan	距離情報かつ隣接情報	任意	◎(少数個)	×	△	○	有

ミニマックス推定量について

大阪大学大学院基礎工学研究科博士前期課程１年
中村将俊

ミニマックスとは，古くから知られている推定手法の一つである．過去には，多くの文献にその言葉が見受けられ，研究対象とされてきたものであるが，現代の統計学ではすでに周知の事実となっており，今の統計学を学ぶものにとっては教科書の隅に書かれた推定に関する『歴史』の一部と認識しているかもしれない．もちろん自分もその一人であったのだが，このような歴史の一部を深く学習するきっかけとなったのは，Lehmann,E.L.(1949).”*Lecture Notes on the Theory of Estimation* University of California”の訳出検討に，大学院に入学して始めて携わってきたことによる．ここでは，その内容の一部を紹介する．

1. 統計的決定問題

Wald(1939,1949) によって定式化された一般の統計的決定問題とは，$X = x$ が観測されるとき，決定 $\delta(x) \in \mathscr{D}$ をくだすという意味で，標本空間対決定空間の決定関数 $\delta(X)$ を選定する問題である．

そこで，より精確には，X が分布 F に従うような真の状況において，仮に決定 d をくだしたとすると，$W(F,d)$ の経済的損失が生じる．任意の特定な決定ルール $\delta(X)$ の確からしさは，その危険関数

$$R_\delta(F) = \mathrm{E}_F\left\{W[F,\delta(X)]\right\} = \int W[F,\delta(x)]dF(x)$$

で定義される．ここでの一般的な狙いは，あらゆる θ の値に対して $R_\delta(\theta)$ ができる限り小さくなるような $\delta(X)$ を選定することである．δ^* にかかわらず，すべての θ に対して

$$R_\delta(\theta) \leq R_{\delta^*}(\theta)$$

であることがわかっているとすると，$\delta(X)$ は一様に最良な推定量であり，ここで求めているものとなる．Wald (1949) によって開発された一般の統計的決定理論の出発点が，Gauss(1821) にあり，これらの理論のが Gauss に依ることを知るのは興味深い．

2. 提案された推定方法

ここにあげる方法は，Cramér(1946) の 33 章で議論されている．

(i) 最小二乗法. この方法の初期の業績に関する議論は，Plackett (1949) に与えられている．$\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ は確率変数である．その分布はパラメータ $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^{\mathrm{T}}$ に依存する．行列

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{ns} \end{bmatrix}$$

は $\mathrm{E}(\boldsymbol{X}) = \mathbf{A}\boldsymbol{\theta}$ を満たす行列とする．$s \leq n$ とし，$\mathbf{A}$ のランクは s であり，$\sigma_{X_iX_j} = \sigma^2\delta_{ij}$ とする．このとき，$\boldsymbol{\theta}$ の最小二乗推定量は

$$(\boldsymbol{x} - \mathbf{A}\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{x} - \mathbf{A}\boldsymbol{\theta}) = \sum_{i=1}^{n}\left\{x_i - (a_{i1}\theta_1 + a_{i1}\theta_2 + \cdots + a_{is}\theta_s)\right\}^2$$

を最小にする $\hat{\boldsymbol{\theta}}$ である．

(ii) モーメント法. この方法は，Karl Pearson によって紹介された．未知のパラメータの関数である母集団モーメントが，適切な標本モーメントの数に等しいとする．そのとき，結果として出てくる方程式の解が $\hat{\theta}$ である．

(iii) 最尤法．この推定法は Gauss によって紹介され，R. A. Fisher(1912,1925,1934) による一連の論文で開発された．X は確率密度関数 $p(\theta, x)$ をもち，θ は事前分布 R(a, b)，すなわち，ある範囲 (a, b) で短形であると仮定する．このとき，$X = x$ が与えられたもとでの，θ の事後分布は

$$p(\theta|x) = \frac{p(\theta, x)}{\displaystyle\int_a^b p(\theta, x)d\theta} = C(x)p(\theta, x)$$

となる．Gauss はこの事後分布の最頻値を，θ に対する最尤推定量 $\hat{\theta}$ として用いることを示唆している．

(iV) 最小カイ二乗法. 標本 $X_1, \ldots, X_n$ を r 個のグループに分け，$p_1(\theta), \ldots, p_r(\theta)$ は対応するグループに落ちる X の確率とする．θ の最小カイ二乗推定値は

$$\chi^2 = \sum_{i=1}^{r} \frac{[n_i - np_i(\theta)]^2}{np_i(\theta)}$$

を最小にする $\hat{\theta}$ である．

3. 大標本理論

最尤推定量の大標本理論として，(i) 一致性, (ii) 漸近正規性, (iii) 漸近効率性の性質は Fisher(1925), Hotelling(1930), Doob(1934) による一連の論文で，厳密さを増しながら証明されている [Wald(1949), Wolfowitz(1949) による証明も参照].

4.「最良」推定量の概念

(i) あるクラスの推定値への制約する方法としての不偏性の原理がある．
不偏推定量，すなわち $\mathrm{E}_\theta[\delta(X)] = \theta$ となる推定量 $\delta(X)$ が対象となる．

(ii) 全体としての危険関数の最適性を要求するミニマックス原理がある．

参考文献

1. Lehman,E.L.(1949).*Lecture Notes on the Theory of Estimation*, Associated Students Store, University of California, Berkley 4, September, 1949-50.
2. Lehman,E.L.& George Casella(1998).*Theory of Point Estimation (Second Edition)*.Springer.
3. Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Math. Stat.*, **10**, 299.
4. Wald, A. (1949). Statistical decision functions. *Annals of Math. Stat.*, **20**, 165.
5. Gauss, C. F. (1821). Theorie der den kleinsten Fehlern unterworfenen combination der Beobachtungen. *Abhandlungen zur Methode der Kleinsten Quadrate*, Berlin 1887.
6. Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
7. Plackott, R. L. (1949). A histrical note on the method of least squares. *Biometrika*, **36**, 458.
8. Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Mess. of Maths.*, **41**, 155.
9. Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Cambrigde Phil. Soc.*, **22**, 700.
10. Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc.*, **144**, 285.
11. Hotelling, H. (1930). The consistency and ultimate distribution of optimum statistics. *Trans. Am. Math. Soc.*, **32**, 847.
12. Doob, J. L. (1934). Probability and statistics. *Trans. Am. Math. Soc.*, **36**, 759.
13. Wolfowitz, J. (1949). On Wald's proof of the consistency of the maximum likelihood estimate. *Annals of Math. Stat.*, **19**, 40.

保健指導に関する評価

五十川直樹

大阪大学大学院　基礎工学研究科　システム創成専攻　数理科学領域

2004 年 4 月，大分県において，健康診断の成績の異常が疾病発生につながることから，疾患発生の予防を意図して健康診断が実施された．臨床検査項目は，身長，体重，BMI，収縮期血圧，拡張期血圧，TC，TG，HDL である．健康診断の結果をもとに放置群，指導群（生活改善の指導），治療群（薬物治療あり）に分け，その後，後者の 2 群については 2 年間にわたって，生活改善の指導あるいは治療が行われた．そして，臨床検査値に改善がみられるか否かが「追証」された．健康診断の対象者は 1141 例 (男性５４３名，女性５９８名) であった．治療群の規定は基準値（収縮期血圧:160 以上，拡張期血圧:100 以上，TC:90 未満 160 以上，TG:250 以上，HDL:25 以上）に基づいており，また指導群の規定は医師の判断によっている．

本研究では，データ適応型確率プロットを用いた各臨床検査値における外れ値の診断，データ適応型階層モデルによる臨床検査値の分布形状・変動要因の同定，プロフィール別参照範囲の設定を行ない，各臨床検査値の特性の把握を試みた．治療群と指導群の４ヵ月後の健康診断の成績を用いて，改善の効果を検討し，さらに，初めに行った健康診断の成績からデータ適応型解析・分類回帰樹木法 (CART) を用いて指導群と治療群の設定根拠を吟味した．

1　解析手法

1.1　データ適応型プロット (外れ値の診断)

臨床検査値の潜在基礎分布にデータ適応型分布である (多変量) ベキ正規分布を想定した．臨床検査値とモデルの変換曲線との乖離の度合いから，その検査値が外れ値であるか，あるいは単に分布の裾の部分であるかを検討することができる．

1.2　データ適応型階層モデル

データ適応型階層モデルを用いて，臨床検査値の分布形状・変動要因の同定を測る．次の階層仮説を考え，各仮説のもとでの AIC を比較することにより，最適モデルを選択する．H_N(ベキ正規性の仮説)，H_U(変換の一様性の仮説)，H_S(等尺度性の仮説)，H_L(位置パラメータに対する線形制約の仮説)[ここで，$H_N \subset H_U \subset H_S \subset H_{L1} \subset \cdots H_{Lm}$ である].

－線形仮説 H_{Lm}(本研究) －

H_{L11}：年代×性の交互作用なし　H_{L12}：年齢の影響なし　H_{L13}：性の影響なし

H_{L21}：年代×性の交互作用なし　H_{L22}：性の影響なし　H_{L23}：年齢の影響なし

1.3　データ適応型判別解析

健常者群，指導群・治療群の分類に寄与する臨床検査値の探索にデータ適応型判別解析を用いる．その判別関数は l 番目の t 変量ベキ正規母集団 $Q_l (l = 1, 2)$ から抽出された n_l 個の非負の t 変量観測値 $\{\boldsymbol{x}_{li}\}_{i=1}^{n_l}$ を用いて構成される．t 変量ベキ正規分布の確率密度関数を用いて，データ適応型判別解析の判別関数は

$$d(\boldsymbol{x}) = \log f_{MPN}(\boldsymbol{x} \mid \boldsymbol{\lambda}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_1) - \log f_{MPN}(\boldsymbol{x} \mid \boldsymbol{\lambda}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\mu}_2)$$

で与えられる．新たな観測値 $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$ から $d(\boldsymbol{x}) > \mu$ であれば，Q_1 に，$d(\boldsymbol{x}) < \mu$ であれば，Q_2 に判別する．ここで $\mu = \log\{q_2/q_1\}$ であり，q_1, q_2 は Q_1，Q_2 に対応する事前確率であり，$q_1 + q_2 = 1$ である．

t 個の説明変数の中の t_r 個の説明変数が判別に寄与する効果は ROC 曲線の AUC(曲線下面積) によって評価する．

2　データ解析例

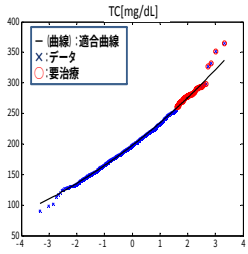


図 1: データ適応型分布
[外れ値なし]

TC

仮説	系列 1	2	3
HPN	11066		
Hu	11066		
Hs	11074		
HL1.	11076	11313	11311
HL2.	11076	11075	11311

図 2: AIC 結果
[性差なし・年齢差あり]

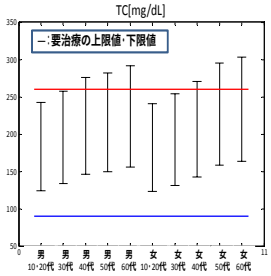


図 3: プロフィール別参照範囲

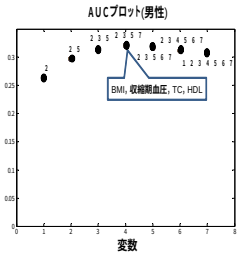


図 4: 放置群と指導群・治療群の分類における ROC 曲線の AUC(男性)．1：体重，2：BMI，3：収縮期血圧，4：拡張期血圧，5：TC，6：TG，7．HDL

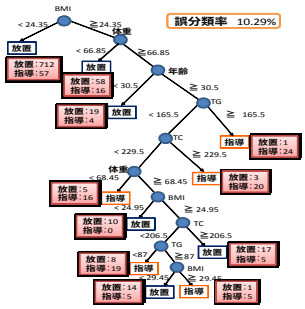


図 5: 指導群と治療群の分類基準 (CART)

3　今後の課題

TC は年齢差の影響，HDL は性別差の影響を考慮した基準値を検討する必要性が示唆された．ただし，背景因子が年齢・性別といった限られた状態での解析を行っているため，他の生活習慣等を考慮した上で解析を行うことも検討する．

適応型判別解析と CART の結果から放置群，指導群，治療群の分類に大きく寄与している臨床検査値が BMI，TC であることわかった．また，明確な規定のなかった放置群と指導群の分類には，BMI，体重，拡張期といった主に体型の特徴が効いていた．

今後にかけて，健康診断の成績に基づく医師の指導の有効性を引き続き評価するとともに定期的な健康診断を行う必要性について検討したい．

臨床検査値の変動の評価

興和株式会社
丸尾和司

医療の場では患者，医師，看護師などの共同参画による総合的な医療の諸事象の改善あるいはリスク回避の対策がはかられている．そこで，予防，診断，治療，予後の過程のすべての面において，その具体的なデータを提供するのは，広い意味で，臨床検査値である．そこに注目するとき，医療における「リスク・マネジメント」として，臨床検査値に潜む諸問題を考察する価値は高いと考えられる．

本稿では，臨床検査値の参照範囲を設計し，評価するために，臨床検査値の分布形状に着目し，背景因子に起因する臨床検査値の変動を包括的にとらえるモデルとしてベキ正規分布に基づく解析過程を提示した．このモデルのうえで臨床検査値の外れ値や健康群と疾病群の分類が一貫して柔軟にとらえることができることを示した．さらに，従来の参照範囲との比較・検討を行い，従来の参照範囲の推定法によってもたらされる情報損失を，参照限界値の推定値の漸近挙動に基づき評価した．これらのモデルを実際に，K クリニックにおいて，2003 年度に行われた 8815 例の人間ドック・データに適用し，参照範囲を導出し，評価した．その結果，殆どすべての臨床検査値の参照範囲が従来の参照範囲よりも広く推定された．このことは，従来の形式的に設定された「集団の特定できない」参照範囲を，実際の集団に適用することの危険性を示唆している．

本研究の遂行で留意した件は以下のとおりである．

- モデリングの重要性：データを統計的に扱うときに，その第一歩として注目するのがそこに切り口を入れるモデルである．統計的にモデルとは，そのデータの潜在基礎構造を支配する分布である．従来，臨床検査値の分布には，正規分布あるいは対数正規分布が仮定されて，定型的に用いられることが多い．しかし，現象に無関係に規定されるこれらの理論分布の適合で臨床検査値の分布(モデル)を記述することは，その背後にある現象の構造を説明することにつながらない．臨床検査値の参照範囲を設定する場合でも諸種の解析法の適用に先行して，臨床検査値の分布を探索することが先決である．
- テータに適応させて接近すること：上記のことと関連して，臨床検査値は一般に正値のみをとり，その分布は歪んでいることが多く，また，その歪みの程度も多岐にわたる．このような臨床検査値の特徴を捉えるためには，データに適応させて接近を測るのが自然である．いわば，「データに語らせる」姿勢である．本稿では，データ適応型分布族であるベキ正規分布(Goto *et al.,* 1979；Invited paper at the 10th International Biometric Conference)に焦点をあてた．ベキ正規分布は形状パラメー

タ(ベキ・パラメータ)を含み，きわめて多様な分布に適合する．このような包括的モデルについて考慮すべき事柄として，①解析全体の流れにおける論理の整合性，②モデルの柔軟さ，③適合性とモデル不適の影響の評価のしやすさ，④計算の容易さなどが考えられる．

- 誤ったモデルに起因する情報損失の評価：前述のように，医学論文では，正規分布や対数正規分布のような理論分布で臨床検査値の分布を記述することが多い．本稿では，データ適応型分布であるベキ正規分布を真のモデルと仮定したもとで，これらの「誤った」モデルを規定することに起因するデータの情報損失(偏り，ARE など)を，誤ったモデルのもとでの最尤推定値の漸近挙動に基づき評価した．その結果，正規分布や対数正規分布といった現象に無関係に規定される理論分布の適合で臨床検査値の分布を記述することによって，相当の情報損失が生じた．とくに，正規分布にもとづく推定値の偏りは全体的に，致命的といえるほどに大きかった．
- 臨床検査値の変動をプロフィールで捉えること：臨床現場では，一つの臨床検査値ごとに単一の参照範囲が設定され，適用されることが多い．しかし，このような集団の特定できない参照範囲を実際の患者に適用することは，誤った診断結果を導く惧れがある．この問題への一つの対処法は，臨床検査値の変動を，個体の日常生活様式に基づくプロフィールで捉えることである．すなわち，参照範囲の設定に先行して，性別，年代，生活習慣などの様々な背景因子でプロフィールを構成し，臨床検査値の変動に及ぼす影響の大きさを評価することが重要である．このことが，ひいては「集団」に基づく情報を「個」に還元することに繋がる．本稿では，背景因子の変動と，臨床検査値の分布形状を包括的に捉えるモデルを提示した．
- データ解析を「過程」として捉えること：臨床検査値の参照範囲には，上記以外にも様々な問題点がある．たとえば，外れ値の診断や，健康群と疾病群の識別などの問題である．臨床検査値の参照範囲の設定ガイドラインなどでは，これらの問題点を目標個別に捉えていることが多いが，このような捉え方では論理の不整合を招くことが考えられる．このような場面では，すべての解析を「断面」でなく「過程」として捉えることが必要である．前述のとおり，まずモデルを探索し，そのモデルを基盤にすることで解析過程に一貫性を付与することができる．本稿では，データ適応型分布(モデル)族のベキ正規分布のもとで，これらの解析過程を一貫して柔軟に捉えることができることを示した．実際に，臨床検査値の分布を同定し，外れ値を吟味した．すなわち，モデルを同定していることで「外れ値」を異常値として識別することができた．さらには，データ適応型判別解析法(下川・後藤，2004；計算機統計学)を適用することにより，健康群と疾病群の特徴を評価した．

ベキ正規分布に基づく ROC 曲線の構成

下川敏雄 * 後藤昌司 †

1. 序に代えて

医学分野において，診断テストから疾病の有無を予測することは，重要な要件の一つである．とくに，診断医学では，適切に疾病の有無を予測するための良好なバイオマーカの探索が研究主題となっている．そこでは，種々のバイオマーカーの評価だけでなく，最適なカットオフ値の選定も議論されている．このとき，有用な統計的方法の一つが ROC 曲線 (Receiver Operating Characteristic) 曲線である．そこでは，解釈の簡便さ，あるいはその後の統計的推測の観点から，正規分布に基づく ROC 曲線が広く用いられている．他方，実質科学では，このような仮定を満たすことは稀である．観測値の正規性の充足を意図するために，観測値にベキ変換を行い，変換値のデータに対して正規分布に基づく方法を適用する方法が提案されている (Zou *et al.*, 1998)．しかしながら,「変換」に基づく方法では，ROC 曲線に対する潜在分布の推定から曲線の解釈までの一貫性を保持できない．

そのため，本論文では，それぞれの群の潜在基礎分布にデータ適応型分布，とくにベキ正規分布 (Goto *et al.*,1979,1983 : Goto & Inoue,1980) を想定する，ベキ正規 ROC 曲線の方法を提案する．ここに，データ適応型分布とは，複数の理論分布を包括できる分布である．これにより，疾患群および健常者群が異なる分布形状を示す場合にも ROC 曲線を構成することができる．さらに，ベキ正規分布が正規分布を包括することから，既存の正規分布に基づく ROC 曲線の適切性の評価にベキ正規に基づく ROC 曲線を適用することができる．

2. ROC 曲線の構成

いま，健常者群および患者群に対する検査値 x が，それぞれ，累積分布 $G(x)$, $F(x)$ をもつ未知の分布に従うとする．任意の検査は，しきい値 u によって，$x > u$ ならば陽性，そうでなければ陰性と判定される．このとき，疾患を正しく陽性と判断することを真陽性 (TP : True Positive)，疾患にもかかわらず陰性と判断することを偽陰性 (FN : False Negative)，疾患でないことを正しく陰性と判断することを真陰性 (TN : True Negative)，疾患でないにも関わらず陽性と判断することを偽陽性 (TP ; True Positive) という．また，u における TP の確率 $TPR(u) = \Pr(x \geq u|D = 1)$ は感度と呼ばれ，FP の確率 $FPR(u) = \Pr(x \geq u|D = 0)$ は (1-特定度) と呼ばれる．これらは

$$FPR(u) = 1 - G(u), \quad TPR(u) = 1 - F(u), \quad (-\infty < u < \infty)$$

で定義される．

3. ベキ正規分布に基づく ROC 曲線

いま，ベキ正規母集団 PND($\lambda_x, \mu_x, \sigma_x$) から抽出された n_x 個の健常者群の検査値 $\{x_{i_x}\}_{i_x=1}^{n_x}$，および，ベキ正規母集団 PND($\lambda_y, \mu_y, \sigma_y$) から抽出された n_y 個の疾患群の検査値 $\{y_{i_y}\}_{i_y=1}^{n_y}$ を考える．λ_x,λ_y,μ_x,μ_y,σ_x,σ_y が既知のとき，$A(\lambda, \mu, \sigma) \approx 1$ を仮定したもとでの，しきい値 u に対する感度 $TPR_{\mathrm{PND}}(u)$ および 1-特定度 $FPR_{\mathrm{FNR}}(u)$ は

$$\begin{aligned} FPR_{\mathrm{PND}}(u) &= 1 - \Phi(z) \\ TPR_{\mathrm{PND}}(u) &= \begin{cases} 1 - \Phi\left(\dfrac{\exp\{\varsigma(z)\}^{\lambda_y/\lambda_x} - \lambda_y\mu_y - 1}{\lambda_y\sigma_y}\right) & , \lambda_x \neq 0, \lambda_y \neq 0 \\ 1 - \Phi\left(\dfrac{\exp\{\xi(z)\}^{\lambda_y} - \lambda_y\mu_y - 1}{\lambda_y\sigma_y}\right) & , \lambda_x = 0, \lambda_y \neq 0 \\ 1 - \Phi\left(\dfrac{\log\{\varsigma(z) - \mu_y\}}{\lambda_x\sigma_y}\right) & , \lambda_x \neq 0, \lambda_y = 0 \\ 1 - \Phi\left(\rho - \dfrac{1}{\sigma_1}(\mu_x - \mu_y)\right) & , \lambda_x = 0, \lambda_y = 0 \end{cases} \end{aligned} \tag{1}$$

で与えられる．ここに，$z = (u^{(\lambda_x)} - , \mu_x)/\sigma_x$, $\rho = \sigma_x/\sigma_y$, $\xi(z) = \sigma_x z + \mu_x$, $\varsigma(z) = \lambda_x\xi(z) + 1$ である．通常，λ_x,λ_y,μ_x,μ_y,σ_x,σ_y は未知であるため，これらのパラメータの最尤推定値 $\hat{\lambda_x}$,$\hat{\lambda_y}$,$\hat{\mu_x}$,$\hat{\mu_y}$,$\hat{\sigma_x}$,$\hat{\sigma_y}$ でおきかえることで,

*山梨大学 大学院医学工学総合研究部, e−mail:shimokawa@yamanashi.ac.jp
†特定非営利活動法人 医学統計研究会, e−mail:gotoo@bra.or.jp

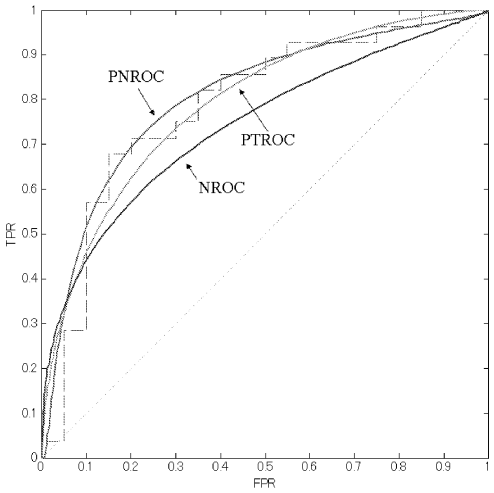


図 1: 卵巣癌データに対する ROC 曲線

(EROC 曲線:経験 ROC 曲線，PNROC 曲線:ベキ正規 ROC 曲線，NROC 曲線：正規 ROC 曲線，PTROC 曲線：ベキ変換に基づく ROC 曲線

$FPR_{\mathrm{PND}}(u)$ および $TPR_{\mathrm{PND}}(u)$ の推定値 $\widehat{FPR_{\mathrm{PND}}}(u)$,$\widehat{TPR_{\mathrm{PND}}}(u)$ を得る．したがって，PNROC 曲線は，しきい値 $u_j(j=1,2,\ldots)$ に対して，$(\widehat{FPR_{\mathrm{FNR}}}(u_j)$,$\widehat{TPR_{\mathrm{PND}}}(u_j))$ を座標軸上にプロットすることで与えられる．

ベキ正規分布が正規分布を包括しているため，PNROC 曲線は，NROC 曲線を特別な場合として表すことができる．すなわち，仮説 $\mathrm{H}_0:\lambda_x=\lambda_y=1.0$ が成りたつとき，NROC 曲線が適切であり，それ以外では，PNROC 曲線が適切である．これらの選定には，赤池の情報量規準 (AIC:Akaike's Information Criteria) を用いることができる．NROC 曲線と NROC 曲線の包含関係は，NROC 曲線の適切性の評価に PNROC 曲線を適用できることを意味している．

4. 事例検討

卵巣癌の有無を識別するために，卵巣癌患者 30 名および健常者 23 名の遺伝子データが観測されている (Pepe, 2003)．本解析の目標は，任意の遺伝子が卵巣癌の発現に寄与しているか否かを評価することにある．

図 1 は，卵巣癌データに対する ROC 曲線である．NROC 曲線は，EROC 曲線 (階段関数) から大きく外れていた．これに対して，PTROC 曲線および PNROC 曲線は EROC 曲線に適合した形状を示したが，PTROC 曲線の FPR は 0.1 から 0.2 付近であり，EROC 曲線から若干外れていた (因に，$\hat{\lambda}=0.911$ である)．これに対して，PNROC 曲線は，他の 2 個の ROC 曲線に比して最良な適合を示した (因に，$\hat{\lambda}_x=-1.750,\hat{\lambda}_y=-0.367$ である)．ROC 曲線のモデル (分布) の適切性を評価するために，潜在基礎分布に基づく AIC を計算した．ここで，PTROC 曲線は，ベキ変換に基づいているため，他の方法と AIC を比較することができない．そのため，PTROC 曲線の代替としてベキ正規分布の形状パラメータを $\lambda_x=\lambda_y=\lambda$ としたもとで，PNROC 曲線の AIC を計算した．NROC 曲線での AIC は，568.0 であり，他の 2 手法に比して極端に適合が悪かった．これに対して，$\lambda_x=\lambda_y=\lambda$ をしたときの PNROC 曲線の AIC は 543.2 であり，PNROC 曲線の AIC は，542.1 であった．すなわち，PNROC 曲線の AIC が最も小さく，最良の適合を示唆した．

5. 結びに代えて

本論文では，ROC 曲線を構成する 2 群 (健常者群，患者群) の潜在基礎分布にベキ正規分布を想定する，ベキ正規 ROC 曲線を提案した．ベキ正規 ROC 曲線は，データが歪んだ分布に従う場合にも，良好な適合結果を示した．さらに，ベキ正規分布が正規分布を包括していることから，正規分布に基づく ROC 曲線の推測の適切性を評価することができた．

参考文献

紙面の都合上，割愛する．

メディカルライターから見た統計家との関わり

～CSR 作成に関して～

ワイス株式会社
メディカルライティング室
内川 葉子

1.0 メディカルライターのタスク

会社によってタスクは異なると思われるが，以下のような文書作成の業務を担うことが多い。

1) 治験に関する文書の執筆（文書化も含む）
   (1) 治験薬概要書
   (2) 治験実施計画書
   (3) 同意説明文書
   (4) 治験総括報告書
2) 承認申請に関する文書の執筆
   (1) コモンテクニカルドキュメント
3) 投稿論文

2.0 治験総括報告書の構成

治験総括報告書は ICH E3 ガイドライン「治験の総括報告書の構成と内容に関するガイドラインについて（平成 8 年 5 月 1 日薬審第 335 号）」に沿って作成する。以下に，特に統計家が関わってほしい（私見）項目（メディカルライターに適切なアドバイスをしてもらったり，統計家がドラフティングしてほしい項目）をあげる。

2.1 統計家にアドバイスをお願いしたい項目

9.2　対象群の選択を含む治験デザインについての考察

- 選択された特定の対象や用いた治験デザインについて，必要に応じ考察すること。
- 治験デザインや対照群に関する問題点を対象疾患や治療法に照らして考察
- デザインその他の特徴について考察
- 用量及び投与間隔を選択した理由が明白でない場合には，その合理的な説明

9.5.2　測定項目の適切性

- 有効性又は安全性の評価法が標準的なものでなかった場合，その信頼性，正確性及び適切性について記述
- 検討したが使用しなかった他の評価尺度
- 代用エンドポイントが用いられた場合，正当性を説明

2.2 ガイドラインに適合するように計画段階で考慮してほしい項目

10. 治験対象患者
10.1 患者の内訳
- 無作為割付けした患者数，組み入れた患者数，治験を完了した患者数
- 無作為割付け後の全ての中止理由

10.2 治験実施計画書からの逸脱
12. 安全性の評価
12.1 治験薬が投与された症例数，期間及び用量

2.3 統計家にドラフティングしてほしい項目

11. 有効性の評価
11.4.2 統計・解析上の論点
11.4.2.1 共変量による調整
11.4.2.2 脱落又は欠測値の取扱い
11.4.2.3 中間解析及びデータモニタリング
11.4.2.4 多施設共同試験
11.4.2.5 多重比較・多重性
11.4.2.6 被験者の「有効性評価の部分集団」の使用
11.4.2.7 同等性を示すことを意図した実対照薬を用いた試験
11.4.2.8 部分集団の検討

ここに挙げた以外にも，統計家の関わる項目は積極的に作成に関与していただきたい。

3.0 帳票のレイアウト

解析計画を立てる段階で，当該帳票類が CTD に使われることを想定し，統一された帳票レイアウトにすることが望ましい。帳票類は読みやすい配列，ソートのかけやすい配列などがあげられる。

読み手にとってわかりやすい表現とは何か　ー留意すべきポイントー

アムジェン株式会社　薬事安全性本部
メディカルライティング室長　藤井久子

1. メディカルライティングの目的

読者が必要とする情報を正確に，簡潔に，わかりやすく伝えること。読み手と書き手が背景情報を共有しているとは限らない。背景から説明しなければならないこともある。メディカルライターは読み手が誰なのかを考えた上で，書き方を変える必要がある。

2. 文書の一部を分担して作成する際に気をつけるべきこと

2.1 ハウススタイルを守る

- 句読点は「、。」「，．（日本薬局方の規定）」「,。（昭和 27 年 4 月 4 日付け内閣通知　公用文作成の要領）」のどれでも構わないが，社内で統一する。
- カタカナ表記は全角表記する。表などのスペースが限られたところでは半角カタカナを使いたくなるが，文字コードがきちんと判別できない場合文字化けを起こすため，半角カタカナの使用は勧められない。
- 数字は全角か，半角か？（すべての数字を半角にする場合，一桁数字のみ全角とする場合が考えられる。いずれにしても社内で統一していればよい）
- 数値と単位の間にスペースを入れるか，入れないか？
- 一連の値すべてに測定単位をつけるかつけないか？
  【例】　2 mL，20 mL 及び 40 mL　　2，20 及び 40 mL
- 使用禁止文字を使っていないか？
  【例】 ㌢，㍻，㈱，①，Ⅳ　文字化けを防ぐため機種依存文字は使わないほうがよい。
- 送り仮名のつけ方，漢字と平仮名の使い分けのルールを守る
  【例】明かに／明らかに　　組合せ／組み合わせる　　さらに（接続詞）／更に（副詞）
- ひとつの文書の中で使う略語に複数の意味を持たせない。
  【例】SD　標準偏差（standard deviation）／安定（stable disease）／単回投与（single dose）

3. わかりやすい文章を書くためのヒント

3.1 主語・述語を明確にする（文章が長くなったら要注意）

【修正前】高血圧を合併している**患者が**本剤投与後軽度**上昇（最高血圧 148=>160）し**，医師の評価では軽微とされたが，経過観察で入院したため重篤と評価した症例である。（平成 13 年承認分申請資料概要から）

主語は患者，述語は上昇？？？

【修正案】高血圧を合併している患者の最高血圧は，本剤投与後，148 から 160 に上昇した。医師は血圧の上昇を軽微と評価したが，経過観察のために入院したことから，治験依頼者は重篤と評価した。

3.2 結論を最初に

【修正前】海外主要試験において安全性を評価した 107 例中 58 例（54.2%）のうち肝疾患を合併した被験者集団 58 例中 51 例（87.9%）に 420 件の有害事象が認められ，肝疾患の合併のない被験者集団においては 49 例中 48 例(98.0%)に 428 件の有害事象が認められた。（平成 18 年承認分申請資料概要から）

「有害事象のことを言いたい」というのが文章の初めでわかると読者は心構えができる。

【修正案】3.3 参照

3.3 修飾関係をはっきり

【修正前】海外主要試験において**安全性を評価した 107 例中 58 例**（54.2%）のうち肝疾患を合併した被験者集団 58 例中 51 例（87.9%）に 420 件の有害事象が認められ，肝疾患の合併のない被験者集団においては 49 例中 48 例(98.0%)に 428 件の有害事象が認められた。（平成 18 年承認分申請資料概要から）

安全性を評価したのは 107 例中 58 名のみ？？？

【修正案】海外主要試験で安全性を評価した 107 名のうち，58 名（54.2%）は肝疾患を合併していた。有害事象は，肝疾患を合併した被験者では 58 名中 51 名（87.9%）に 420 件，肝疾患の合併症のない被験者では 49 名中 48 名（98.0%）に 428 件発現した。

3.4 「～において」はなるべく使わない

「～において」を使うと主語が曖昧になる。

【修正前】当該被験者においては，有害事象に対する処置として鎮痛薬を投与した。

誰が，誰に，投与したのか？

【修正案】有害事象に対する処置として，医師は当該被験者に鎮痛薬を投与した。

3.5 動詞を名詞化しない

動詞を名詞化すると表現が冗長になる。科学的な文章には不適。

【修正前】重篤な有害事象の報告は行われなかった。

【修正案】重篤な有害事象は報告されなかった。

3.6 受動態を多用しない

受動態を用いると，文章が長くなり，行為者が曖昧になる。

【修正前】治験薬との関連はないと判断された。

誰が判断したのか？　医師？それとも治験依頼者？

【修正案】治験責任医師は，治験薬との関連はないと判断した。

ただし，行為を受ける側に重点があるときは受動態を用いる。

【例】19,185 例の患者が登録され，このうち 9,599 例（50.0%）が X 群に，9,586 例（50.0%）が Y 群に無作為に割り付けられた。（平成 18 年承認分申請資料概要から）

受動態を用いてもよい理由：誰が登録したかよりも何名が登録されたかが重要。誰が割り付けたかよりも何名が割り付けられたかが重要。

3.7 「それぞれ」の使い方

項目が 3 つ以上あるときに「それぞれ」を使うと読み返さないと対応がわからなくなる。

【修正前】本剤投与前の AST, ALT 及び総ビリルビン値はそれぞれ 28 U/L, 31 U/L 及び 8.0 mmol/L であり施設基準値範囲内であった。（平成 17 年承認分申請資料概要から）

【修正案】本剤投与前の AST 値は 28 U/L，ALT 値は 31 U/L，総ビリルビン値は 8.0 mmol/L でありいずれも施設基準値範囲内であった。

論文執筆の作法

長崎大学大学院医歯薬学総合研究科　　柴田義貞

１．はじめに

言葉は生き物である。新しく生まれる言葉もあれば、死んで行く言葉もある。「作法」は現在の日本ではほとんど死語かもしれない。広辞苑によれば、「作法」は次のように説明されている。

①物事を行う方法。「小説―」

②起居・動作の正しい法式。「礼儀―」

③きまり。しきたり。宇津保蔵開中「例の―に政事あらせてこそ候はせ給へ」

この小論では、筆者の独断と偏見で科学論文を執筆する上での基本事項を述べる。いわゆる個別科学の論文執筆を念頭に置いているが、統計的方法論に関する論文執筆にも少しは寄与できればと願っている。

２．論文とは

論文は小説などの文芸作品である。論文も小説も、広く他人に読んでもらって始めて存在価値を示す「作品」である。共にオリジナリティがなければならず、論文執筆は作家などの創作活動と酷似している。しかし、論文は、文芸作品ではない。論文の文章は文字通り論理的でなければならないが、論理でない文章の文芸作品もある。

ところで、文章の基本は起承転結であるが、これを論文に当てはめると、起は「序」承は「対象と方法」転は「結果」結は「考察と結語」ということになるであろう。さらに、この大きな起承転結のそれぞれが起承転結から構成されることになる。

３．論文執筆

まず、論文執筆の時期であるが、大きく分けて次の３通りがあり、後ほど良い論文を書くことができる。

（１）データを解析した後に執筆する

（２）データ収集後、データ解析の前に執筆する

（３）データ収集前に執筆する

（２）、（３）の場合、論文には本体が空白の図表を記載し、データ解析は論文の内容に沿って進めて行く。（３）の場合はさらに、データ収集も論文に沿って行われる。なお、いずれの場合においても、「考察と結語」は当然ながら、すべての「結果」が得られてからのみ執筆可能になる。

いずれの場合も、論文の目次を作成してから執筆を開始することが大切である。書き慣れていれば、目次自体が頭の中に入り、後は筆の進むままということになるかもしれないが、初めのうちは目次をしっかり構成しておく必要がある。

表題を定め、全体の構成を「序」「対象と方法」「結果」「考察と結語」の４章とした上で、それぞれの章について、節（必要ならば小節も）の見出しを決め、そこに記述すべき事柄を箇条書きにして行く。なお、「序」などは節の見出しは設けないが、パグラフごとに記述すべき内容を箇条書きにしておく。このようにして出来上がった目次全体に目を通すと、内容の過不足、記載順序の適不適などが分かってくる。目次を推敲し、確定したならば、後は箇条書きの箇所を肉付けして行けば論文は仕上がる。

良い論文を書くには、悪い論文を書かないようすればよい。そのためには、どのような論文が悪い論文であるかを理解しておかなければならない。完全ではないが、以下にその例を列挙しておく。

（ア）論旨の通りが悪い

これは主として接続詞の使い方が論理的でないことによる。順接と逆説の接続詞を適切に使い分けることができれば、論理的で退屈しない文章が出来上がる。広辞苑によれば、順接とは「甲乙二つの文または句を接続するしかたで、甲から当然に生ずる順当な事態として乙が成立することを表すもの。」とあり、挙げられた例文は『雨が降った。それで道が悪い』である。英文では、therefore, hence などである。同じく広辞苑によれば、逆接とは「前後する甲乙二つの文または句における接続のしかたの一で、甲で述べた事実と相反する事態またはそれから予想されるものとは違う事態が乙において成立することを表すもの。」とあり、『雨が降った。しかし出かけた』が例文に挙げられている。英文では、but, however, while などである。

（イ）図表の長短を考えない

図の長所は一目瞭然ということである。しかし、詳細な情報は表示できない。一方、表は詳細な情報は表示できるが、一目瞭然ではない。

（ウ）データの表示と統計手法が対応していない

典型的な例は、データの分布は平均±標準偏差として要約する一方で、２群の比較には Wilcoxon の順位和検定を用いるというものである。

（エ）「結果」において図表の内容を繰り返す

「結果」では、図表の内容への言及は必要最小限に抑え、図表に示しきれなかった重要な点を述べるべきである。

（オ）「考察」において結果の内容を繰り返す

「考察」においては、結果への言及は必要最小限に抑え、研究結果を過去の研究と比較しながら、その長短を論じるべきである。

（カ）主観的な言辞を多用する

「非常に重要な結果である」とか「得られた結果は非常に興味深い」などの記述は避けるべきである。それらは、読者の判断に委ねるべきである。

データ解析環境 R の多面的利用

東海大学理学部数学科　山本義郎

1. はじめに

統計解析環境 R は、フリーソフトウェアであるが、協力者の作成した機能を取り込みやすくしたパッケージという仕組みにより、利用できる機能は膨大なものとなり、今や商用のソフトウェアを超える機能を有するまでになったといっても過言ではない

パッケージにより追加される R の機能としては、単に多種多様な統計解析機能だけではなく、R の利用環境(GUI)やデータベースの接続を含む他言語との相互接続の機能などがある。本報告では、R を多面的に利用するための利用法のいくつかについて紹介した。

2. R におけるプログラミング

R ではプログラミングは可能であり、C や Fortran どうようのプログラミングも可能であるが、プログラミングには少々癖があり、効率のよいプログラミングをするためには R の癖を理解することが必要となる。最も有名なものとしては S 言語の特徴でもある「ループは使わず行列・配列に持ち込め」があるが、その他にも様々な癖がある

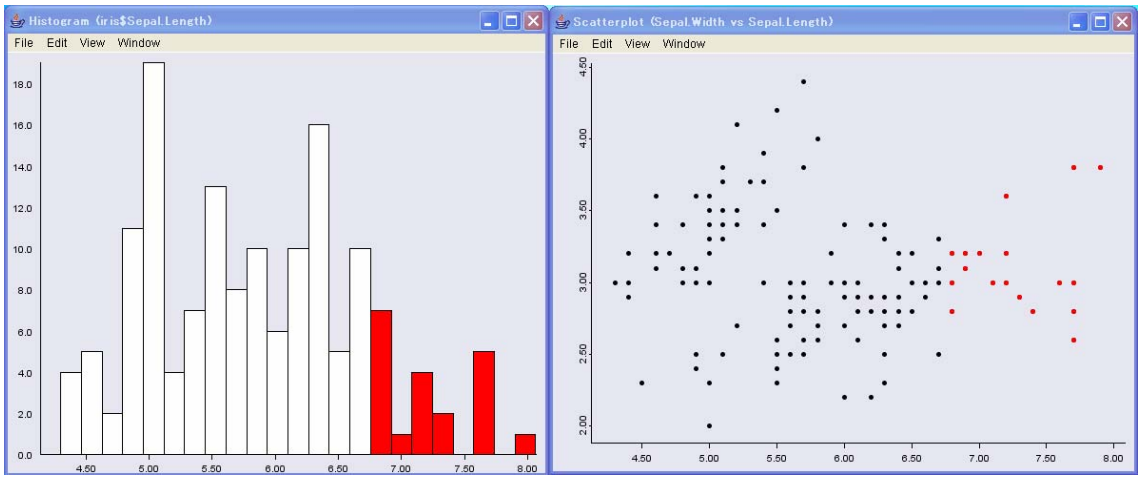
このように R には不得意な分野もあるため、R で効率の悪い計算となりそうな部分については、C や Fortran でプログラミングし、R の中で C や Fortran のプログラムを利用できる。同様に、C や Fortran で記述された各種計算ライブラリも利用できる。フリーの数値計算ライブラリとして定評のある GSL も例えば GSL の C 言語用ライブラリは R では `.C` 関数を利用することにより利用できる。R には GSL で提供されている関数と同様の計算を行う関数が用意されているが、その実装のされ方(引数の型)が不都合な場合や、関数を再帰的に利用する場合には、直接 GSL の関数を利用した関数を記述したり GSL を使って作成した C の独自関数を R から利用したい方が効率のよい場合がある。

3. R の実行環境

R は対話型のアプリケーションであり、プロンプトに命令を打ち込む旧式のインタフェースであるため、初心者や文系の講義などにおいては少々敷居が高い。そのような要求に応えるパッケージがいくつかある。その代表的なものとして、R コマンダー (R Commander , Rcmdr パッケージ)と JGR がある。R コマンダーは Tcl/tk による GUI で、メニューにより R を使うことができる。Rcmdr ライブラリをインストールして、読み込むことにより利用できる。R コマンダーを起動すると別窓として R コマンダーが現れ、その中でメニューを利用して解析が実行できるが、グラフは起動した R の中で表示されるため使い勝手が悪くなる。これを回避するためには、R の起動時オプションにとして `--sdi` をつけ SDI タイプ(各ウィンドウが独立して現れる)とするとよい。メニューから分析やグラフの指定などを選択し、ダイアログボックスで解析の詳細を指定することにより、R コマンダー内の出力ウィンドウに結果が表示される(図 3)。メニューとダイアログボックスで指定した命令はスク

リプトウィンドウに表示されるので、コマンドに慣れるまで利用し、ある程度理解したら、通常のコマンドプロンプトを利用できるようにスクリプトを読むようにするのがよい。

JGR (Java Gui for R) は Java によるインターフェースであり、コマンド入力において補完が利用できるなどメニューを利用した分析やコマンドの保管などが利用できる。Windows では、JGR アプリをインストールして利用する必要がある。JGR を利用すると、(図 4)。また、JGR のグラフ描画機能は、JGR パッケージをインストールすることにより R でインタラクティブグラフとして利用することができる。



JGR は Java による GUI で R を操作しつつ、グラフ作成で若干不便である R のグラフを用いず独自のグラフを利用するといった使い方をしている。このように、R は他のアプリケーションとの連携の機能も続々開発されており、RExcel では Excel でデータの編集を行い、高度な解析については R で実施できる。グラフについては Rggobi パッケージを利用することによりインタラクティブにグラフを操作可能な Rggobi が利用できる。他にも、Xlisp-stat や J などの解析ソフトとの相互連携も可能である。

4. Web の活用－Web アプリケーションの統計エンジンとして－

R の活用としては、さらに Web アプリケーションや Web サービスにおける統計エンジンとしての利用できることがあげられる。すでに紹介しているように、R では各種ソフトウェアとの連携が可能であり、その延長として Java から R へアクセス可能な JRI、PHP や Perl、Ruby などの Web 言語の API が用意されそれらで構築された Web アプリケーションでは、解析処理は R を利用することで煩わしいプログラミングから開放され、また安心して解析結果を利用することができる。図 6 は、CGI により作成されている Web 上での解析システムであるが、解析をしているフォームは Perl による CGI で記述し、指定された解析を、R を起動し解析するプログラムについても Perl で記述している。

このようなシステムでは、利用者に統計解析システムを利用する能力がなくとも、サーバに用意されたデータに対して、解析を行うことが可能となる。このような利用方法は今後ますます要求が高まると思われる。

樹木構造接近法における R

下川敏雄[1]・杉本知之[2]

データマイニングあるいは機械学習に関する研究の興隆に伴い, 統計ソフトウェアの自由化と利便性が急速に拡大している．その主体的な役割を担っているのが，データ解析環境 R である．R は, SAS などの有償ソフトウェアに代わる, 統計ソフトウェアのデファクトスタンダードになりつつある一方で, 提供されている幾つかのパッケージには不十分な点が多い. 例えば, 樹木構造接近法では, `tree`, `rpart`, `party` の 3 種類のパッケージが用意されているものの, 得られるオブジェクトの形式が異なるために, それぞれの結果に注意して解釈することが必要である．さらに, それぞれのパッケージの内容を知らずに適用することは, 統計的手法の誤用に繋がる.

自動交互作用検出法 (AID [Automatic Interaction Detection]: Morgan & Sonquist, 1963) に端を発する樹木構造接近法は, 情報技術の発展やニーズの拡大 (例えば, データマイニングやケモメトリックス) により, 急速にその版図を広げている．そのような理由として, 1) 樹木構造接近法の解釈の平易さとデータの型の制約がないこと, 2) 結果がグラフィカルに提示できることから解釈が容易である, 3) 応答と説明変数の非線形構造および交互作用を自動的に提えることができる, などが挙げられる．さらに, これらの手法は, 既に統計ソフトウェア R などに実装されており, 誰もが活用できる環境が整備されている.

本報告では, 樹木構造接近法に議論を限定して, R の適用の実際を議論し，それぞれのパッケージの特徴について言及する.

R には，分類回帰樹木法 (CART[Classification And Regression Trees]: Breiman *et al.*, 1984) あるいはその類似法のパッケージとして，`rpart` と `tree` の 2 種類が用意されている．前者は, CART 法の全ての過程を実装しているものの, 後者には交差確認法が含まれておらず, 部分樹木系列からの最適部分樹木の選択に，若干の主観性を伴う惧れがある．そのため，通常は，`rpart` が R における樹木構造接近法として推奨される．その他に，分岐基準の剪定を条件付き推測の枠組みで捉えることで，より包括的な樹木を提供する Hothorn *et al.*(2006) の方法は，`party` 関数により実装されている．また，Chipman *et al.* (1998) による Bayes 流樹木構造接近法は，`bayestree` 関数により提供されている.

ここでは，Forbes2000 のデータを用いて通常の CART 法の例示を与える．このデータは企業の市場価格 (`market value`), 売上高 (`sales`), 利益 (`profits`), 負債 (`assets`) が記載されたものである．解析の目標は，市場価格への影響の大きさを探索することである．例えば，以下のような R のソースにより，欠測値のない 1995 例に対して，CART 法の第 1 ステップが適用され，ある樹木図が描かれる (結果については割愛する)：

```
data("Forbes2000", package="HSAUR")
Fores2000 <- subset(Forbes2000,!is.na(profits))
layout(matrix(1:2,ncol=2))
library("rpart")
forbes_rpart <- rpart(profits~assets+marketvalue+sales, data=Forbes2000)
plot(forbes_rpart,margin=.10)
text(forbes_rpart,cex=0.7)
```

[1]山梨大学 大学院医工学総合研究部 社会システム工学系
[2]大阪大学 大学院医学系研究科

```
plot(forbes_rpart,compress=T,uniform=T,branch=0.4,margin=.10)
text(forbes_rpart,cex=0.8)
```

このもとで, 各樹木の複雑度コストの交差確認推定値を知るためには, printcp() 関数や plotcp() 関数を用い，plotcp(forbes_rpart) や printcp(forbes_rpart) のように書く.

さて，第 1 ステップでの樹木図は，統計的な最適性に対して何ら考慮された樹木ではない．そこで，交差確認推定値を用いて刈り込みを行い，ある最適樹木を構成するための R ソースを以下に示す．ここでは，結果の解釈を支援するために，終結ふしの下側にボックス・プロットを描き，箱の横幅を各ふしの標本サイズの平方根に対応させている (結果は割愛する).

```
opt <- which.min(forbes_rpart$cptable[,"xerror"])
cp <- forbes_rpart$cptable[opt,"CP"]
forbes_prune <- prune(forbes_rpart,cp=cp)
layout(matrix(1:2,ncol=1))
plot(forbes_prune, uniform=T,margin=0.1,branch=0.5,compress=T)
text(forbes_prune, cex=0.8)
rn <- rownames(forbes_prune$frame)
lev <- rn[sort(unique(forbes_prune$where))]
where <- factor(rn[forbes_prune$where],levels=lev)
boxplot(Fores2000$profit~where, varwidth=T,
        ylim=range(Fores2000$profit)*1.3, pars=list(axes=F)
        ,ylab="Profits in US dollars")
abline(h=0,lty=3)
axis(2)
```

このときの樹木図から，樹木は最初に市場価格 (markevalue) で分岐し，市場価格が 83.33 未満の企業は，サイド，市場価格で分岐した．他方，市場価格が 83.33 以上の企業は，売り上げ (sales) で分岐した．最終的に 5 個の終結ふしが得られ，終結ふし内の平均利益は左から右にいくほど上昇傾向が見られ，最適樹木の寄与率は 0.346 程度である.

CART 法の妥当性やより好ましい変数重要度測度を求めたいとき，より予測確度の高い樹木構造接近法の利用が考えられる．例えは，Boosting CART(Friedman, 2001) 法では，gbm パッケージを，MARS(Friedman, 1991) 法では，mda もしくは earth パッケージ用いることができる．その他のアンサンブル型樹木構造接近法において，Bagging(Breiman, 1996) 法は adabag パッケージ，Breiman(2001) のランダム・フォレスト法は RandomForest パッケージ, ランダム・フォレスト法の生存時間解析バージョンは SurvivalRandomForest(Ishwaran & Kogalur, 2006) に実装されている．こららの方法の特徴や適用結果，および通常の CART 法との適用比較などについて，当日の発表において報告する.

また，生存時間データにおける樹木構造接近法の使用においても，同様の R の関数により実行可能である．例えば，rpart パッケージでは，基線ハザードに指数分布を想定したもとで，全尤度およびその偏分を算出し，LeBranc & Crowley(1992) と同様に，CART アルゴリズムのもとで，樹木を構成する．他方，party パッケージでは，全ての終結ふしが異なる生存時間分布をもつように，全ふし対で多重比較を行い，全ての対で有意になるような分岐を探索する (Hothorn *et al.*, 2006)．我々はこれらの方法の適用例を，当日の発表において議論し，適用上の便宜をはかるため，有用なグラフィカル表示を与える関数や併合過程を行うための新たな関数を紹介する.

数理概念の具現化ツール R:離散データ解析への応用

大分大学工学部　越智義道

1．はじめに

統計的なデータ解析において重要な観点は，統計的推測法を適切に適用することに尽きるが，データの適切な収集から始まり，最終報告の記述にいたるまで，それには諸相の観点がある．なかでも，データの解析過程で適切な解析手法を選択し，さらに現実のデータに照らして，解析手法あるいは解析モデルをカスタマイズし解析を進めることが不可欠であり，これは統計家に課せられた重要なタスクである．いわゆる教科書に記載された標準的な解析モデル，解析手順では，現実のデータを十分に記述し説明することができないことが多い．このとき，統計家は数理的にモデルのカスタマイズを行い，場合によってはそのモデルにもとづく推測過程そのものに関わる改善を求められる．このプロセスにおいては，定型的なデータ解析にもとづく分析パッケージでは十分でなく，分析者の数理的なアイデアを効果的に具現化する機能を合わせ持つことが求められる．統計ソフトウエア R は基本的な分析ツールを備えると共に，それを修整・拡張し，新たな解析手法を構築する上で非常に柔軟な環境を提供する．本報告では，離散データ解析への応用例を通じて，データ解析場面での数理的なアイデアの具現化に対する R の可能性について検討する．

２．統計解析環境 R

R は Ihara & Gentleman(1993)らによって開発され，現在では国際的な開発チーム R Development Core Team によって活発に開発が行われている統計解析システムであるが，その特徴は，先行する商用システム S と使用感を等しくしたフリーな統計解析ソフトウエアである．S と同様 R も解析システムでありながらインタプリタ言語的色彩を強く持ち拡張性・柔軟性に優れている．また，GNU GPL ライセンスによってソースが公開されている．

R Development Core Team ではその特徴を“効果的なデータハンドリングとストーレッジ機能・配列，特に行列計算関連の一連の演算機能・データ解析のための広範で一貫性のある総合的な中間的なツール群・データ解析のための画面・印刷用のグラフィカル機能・条件分岐，ループ，ユーザ定義再帰関数，入出力機能を含む綿密に作りこまれた簡潔で効果的なプログラミング言語”にあると述べている(http://www.r-project.org/)．その目標とするところは S 言語と同様に，“アイデアを，速やかに，正確にソフトウエア化すること”(Chambers,1988)にあるといえる．

3．離散データ解析への応用

R には基本的な離散データ解析ツールは基本コマンドあるいはパッケージとして用意さ

れている．例えば分割表におけるカイ２乗検定(chisq.test)，Fisher の正確検定(fisher.test)，あるいは一般化線形モデルにもとづく解析法(glm)などはそのコマンド，あるいは適宜用意されたオプションを設定することによって分析を遂行することが可能である．あるいは，これらの関連の分析ツールを組み合わせて，一連の分析手続きを関数として取りまとめることも可能である．この際には，そのインタプリタとしての言語機能と履歴管理，関数編集機能，デバック機能などの強力なプログラミング支援機能を利用することができる．また，解析の計算上の中間ツールもユーザに公開され，基礎的な分布関連関数，組み合わせ関連関数(sample, choose, factorial)，最適化関数(optim, optimize)，非線型方程式の解法(uniroot)，数式処理(微分)関数(deriv, deriv3)などを利用することによって柔軟に解析を進めることが可能である．

例えば，分割表での条件付推測におけるオッズ比の信頼限界は，

$$p_{-}=\sum_{u=0}^{a}\Pr(A=u)=\sum_{u=0}^{a}\frac{\binom{n_{1\bullet}}{u}\binom{n_{2\bullet}}{n_{\bullet 1}-u}\phi^{u}}{\sum_{v}\binom{n_{1\bullet}}{v}\binom{n_{2\bullet}}{n_{\bullet 1}-v}\phi^{v}},\quad p_{+}=\sum_{u=a}^{n_{\bullet 1}}\Pr(A=u)\quad p_{-}(\phi)=\frac{\alpha}{2},\quad p_{+}(\phi)=\frac{\alpha}{2}$$

なるようなϕに関する多項式方程式の解を計算する必要があるが，

```
> hypergeo<-function(N,M,n,x,phi=1){a<-0:n;V<-sum(choose(N,a)*choose(M,n-a)*phi^a);
        choose(N,x)*choose(M,n-x)*phi^x/V}
> p_minus<-function(phi){ s=0; for(a in 0:4){ s=s + hypergeo(32,36,21,a,phi) }; (s-0.05)^2}
> optimize(p_minus,c(0.001,1))                    (a,n_1.,n_2.,n_.1)=(4,32,36,21),α=0.1 の場合
```

のように簡潔に表現することによって計算することができる．

また，P_iを反応確率，X_iを用量とするような量反応モデルにおいて，ベースライン用量モデル(Crump *et al*., 1976)

$$P_i=\frac{\exp\left(\beta_0+\beta_1\log(X_i+\exp\beta_2)\right)}{1+\exp\left(\beta_0+\beta_1\log(X_i+\exp\beta_2)\right)}$$

などは，一般化線形モデルの枠組みを超えるため，通常の glm などでは分析できないが，尤度寄与をパラメータの関数(li)として expression 関数によって定義しておくと，

```
> dli <- deriv3( li, c("b0","b1"), function( b0, b1, n, y, x ){})
> MLstage<-function(a,Ni,Yi,Xi){ addli<-adli(a[1],a[2],a[3],Ni,Yi,Xi);
  adll<-apply(attr(addli,"gradient"),2,sum); addll<-apply(attr(addli,"hessian"),c(2,3),sum);
  a<-a-solve(addll,adll); a}
```

のような形でニュートン・ラフソン法の１ステップでの計算を簡潔にかつその数理的表現と一貫性をもって記述することができる．本報告では，これら離散データ解析の数理的展開と R での実装化の特性について議論するとともに，Busvine(1938)の Grain beetle のエチレンオキシドへの反応データ，Hoekstra(1987)の aphid のニコチン反応データなど実データへの適用とその効果についても言及する．

臨床評価における欠測値のとり扱い

永久保太士 †
† アスビオファーマ株式会社

1. 序に代えて

新薬開発などの臨床試験では, 薬剤を投与した患者に対して時間による変化を比較することで治療効果を評価する. このように同一の対象を時間を追って観測して得られたデータは, 経時対応データと呼ばれる. 臨床試験では計画されたすべてのデータを測定できることはほとんどなく, データ解析を実施する際に欠測値の問題に対処する必要がある.

欠測値は治療効果を推定, 比較するときに偏りの起こる原因の一つであり, 理論的に興味深い問題である. 欠測値の問題は ICH E9 ガイドラインや欧州医薬品委員会 (Committee for Proprietary Medical Products) の Points to Consider on Missing Data(PtC) でも論じられている. 欠測値によって起こる偏りを除くことは, 困難もしくは不可能であるので, PtC では欠測値を回避できるデザインを事前に工夫することが重要であると述べられている. また, 欠測値のとり扱い方法によって結果が異なることが危惧されるため, 治験実施計画書の統計解析計画の項でとり扱いを予め定めることが重要であり, その方法の詳細とそれが適切と考えられる理由を含め, 正当化が必要であることも論じられている. 他に, 欠測値が生じる割合や時点に関する群間比較や感度解析の実施が推奨されている.

多数の欠測値が生じることで臨床試験の妥当性を崩す. 試験の計画と実施, 解析計画, 結果の報告と解釈において, 適切に欠測値の問題をとり扱うことが, 試験の結論を十分に支持するために必要である. 経時対応データにおいて欠測が生じた場合の統計解析手法としては, 欠測値を無視した単純な解析手法からデータが欠測する理由である欠測メカニズムを取り入れた複雑な解析手法までさまざまな手法が提案されている.

本報告では, 欠測値のとり扱い法を紹介し, 臨床試験においてどのように欠測値を考慮すべきかを Mallinckrodt *et al.* (2004) に基づいて報告したい. 2 節では, 欠測値の発生するメカニズムと無視可能性について説明する. 3 節では, 欠測値のとり扱い法について, それぞれの概要を説明する. 4 節では, 臨床試験において欠測値に対して考慮すべき点について述べ, 5 節でそれらをまとめる.

2. 欠測値の発生

2.1 欠測値の単調性

対象 $k(k = 1, \ldots, n)$ について時点 $t(t = 1, \ldots, T)$ で応答 Y_{kt} が測定されるとする. 各時点で各対象が測定されているかどうかを表す指示変数を R_{kt} とする. R_{kt} は Y_{kt} が測定されていれば 1, 測定されていなければ 0 をとる. 欠測値に関する指示変数 R_{kt} のパターンによって欠測値は単調と非単調に分類される. 単調な欠測値は, 欠測のある対象について, ある時点 t' まではすべての応答が測定されるが, 時点 $t' + 1$ 以降はすべての応答が欠測である場合であり, 脱落とも呼ばれる. 非単調な欠測は, それ以外の欠測パターンである. 経時対応データでは何らかの理由で, ある時点以降のデータが得られないことがあり, これは単調な欠測である.

2.2 欠測のメカニズム

欠測メカニズムは通常三つに分類される. 一つ目は欠測が, 観測された値と観測されなかった値のどちらにも依存しないとき, データは完全にランダムに発生する (Missing Completely At Random: MCAR). 二つ目は欠測が, 観測された値に依存し, 観測されなかった値に依存しないとき, データはランダムに発生する (Missing At Random: MAR). 三つ目は欠測が, 観測されなかった値に依存するとき, データはランダムでなく発生する (Missing Not At Random: MNAR). MCAR と MAR は無視可能な欠測である.

3. 欠測値のとり扱い

欠測のあるデータに対する解析には, 単純なものから複雑なものまでいくつかの方法がある. 欠測値への代表的なとり扱い法は, 完全例解析, 補完法, 得られたデータを用いた解析, モデル化などが

ある.

完全例解析はすべての測定が記録された対象のみを解析対象とする. この方法の明らかな利点は, 説明が非常に容易で理解しやすいことである. しかし, 欠測メカニズムが MCAR である必要がある. そこで, 欠測メカニズムが MAR の場合に, 妥当な推測を行う方法の一つとして Robins *et al.* (1995) により, 欠測するまでの対象ごとの履歴を考慮して欠測確率をモデル化する Inverse Probability of Censoring Weighted(IPCW) 法が提案されている.

補完法は欠測値にある値を補完して解析を行う方法である. この方法は容易であるという魅力をもち, 一度欠測値が補完されると完全データに対して利用可能であるすべての解析手法が適用できる. 従来から用いられている LOCF は補完法の一つであり, 最後に観測された値で欠測値を補完する方法である. しかし, 欠測に一つの値のみを補完する単一補完法では, 擬似的な完全データセットからの点推定値は求められるが, 欠測に伴う情報の損失が適切に評価できるとは言い難い. 多重補完法では, ある一つの欠測値に対して複数回の補完を行うことにより, この不確実性を考慮することができる.

得られたデータを用いた尤度に基づく解析は, 欠測メカニズムが MCAR, MAR の場合, 欠測を考慮せずに妥当な結果を導く. 欠測メカニズムが MNAR である場合, 偏りのない推定値を得るためには測定過程と欠測過程の同時分布に対するモデルを考える必要がある. その方法は選択モデルとパターン混合モデルがある.

4. 臨床試験における欠測値の考慮

すべての問題の最良な解決法は予防である. 欠測値を最小にする試験計画と実施によって欠測値を減らすことができる. 欠測の原因を記録するデータ収集によって問題をより特徴付けることができ, 適切な統計的方法を選ぶことができる.

異なる統計的方法はときに異なる結果を導く. そのために, 欠測値のとり扱いや妥当だと推測されるが主要な解析とは仮定の異なるいくつかの感度解析の選択について, 治験実施計画書で事前に規定することが必要不可欠である (とくに, 検証的第 III 相試験).

欠測値の不均衡に関連のある要因と試験結果, 欠測がある患者とない患者で異なる特徴をもつかどうかを調べる解析を行うべきである. 事前に規定された主要な解析と感度解析の結果を報告することに加えて, 欠測値の範囲とパターンが予測されたものから重大な乖離をもつならば, 追加の感度解析を実施し報告することが大切である. すべてを含めた結果が, 試験の結論についての頑健性の評価を可能にする.

5. 結びに代えて

本報告では, 欠測値のあるデータについていくつかのとり扱い法と臨床試験において考慮すべき点を述べた. 欠測値のとり扱いで, これらのどの方法を用いても妥当な結果が得られるわけではなく, それぞれの方法の仮定が成り立っている必要がある. しかし, これらの解析で必要となる欠測メカニズムの仮定の多くは, データから検証することができない. このような点も含めて, 臨床試験の計画段階において, これまで行われた試験から得られたデータを吟味し, 欠測がどのような理由で起きるのか, どの測定値と関連しているのかといった欠測に関する情報を把握することが重要である. 加えて, これら欠測に関する情報を記録できる計画を立てるべきである. 欠測メカニズムが MNAR である状況では妥当である欠測のとり扱い法が限られ, より複雑になる. 欠測に関する情報を記録することで, MNAR である場合をなるべく避け, MAR の仮定のもとで妥当な解析法を適用することが可能になる.

また, 他に計画時に重要なことは欠測を最小限に抑えることである. このためにも欠測に関する情報の把握は重要である. 欠測を最小限に抑えることおよび適切な解析法を用いることで, 欠測による偏りを最小限にする. そのための主要な解析および感度解析の適切な選択には, どのような程度, パターンの欠測ではどの解析法が適しているのかを実施のデータ解析をとおして検討していくことが重要になってくる.

An Approach to Rationalize Partitioning Sample Size into Individual Regions in a Multi-regional Trial

Norisuke Kawai
Statistics and Clinical Programming Group, Pfizer Global Research and Development, Pfizer Japan Inc., Tokyo, Japan

1. Introduction

In this paper, we will propose an approach to rationalize partitioning the total sample size in a multi-regional trial among the constituent regions. We will focus on trials where a new treatment is compared to a placebo. The primary objective is to demonstrate that the new treatment is superior to the placebo. Two principles are important to our discussion. First, no region can proclaim itself to be the region of interest and demand the treatment to show a statistically significant result at the usual significance level within the region. Second, the overall sample size is determined by the study's primary objective of demonstrating an overall treatment effect.

In our approach, we assume that the true (but unknown) mean treatment effect is uniform across regions. We define a consistent trend to have occurred if the point estimates for the treatment effect in different regions are all positive. Our approach is to find the minimal sample size for the smallest region so that there is a high probability (80% or 90%) of observing consistent trend in treatment effect across regions if the treatment effect is indeed positive and homogeneous across regions.

2. Probability of Observing a Consistent Trend across All Regions

We will formulate the probability of observing a consistent trend across regions. Consequently, for a fixed number of regions and a total sample size that is determined to achieve a desired power, the probability of observing consistent trend across regions depends only on the allocation of patients among regions.

3. Unconditional Probability

We will examine the probability of observing consistent trend across all regions numerically. For convenience, we will call this probability the *unconditional* probability to distinguish it from the concept of *conditional* probability we will introduce later.

As a result, for the case of three regions, if the sample size is determined to provide an 80% statistical power for the primary analysis, the smallest region should

contribute at least 21.3% of the patients so that the probability of observing consistently positive results across regions would be at least 80%. When the sample size is determined to provide a 90% overall power for the primary analysis, the proportion of patients coming from the smallest region can drop to 15.1% for a 80% chance to observe consistently positive results across regions. In the latter case, if we want the probability to be at least 90%, the smallest region should contribute at least 27.7% of the patients to the multi-regional trial.

4. Conditional Probability

So far, the probabilities discussed above are not conditional on whether or not the overall treatment effect is statistically significant. In practice, for a confirmatory trial, inference concerning regional results is relevant only if the overall treatment effect is statistically significant. Because of this, we will investigate the probability of observing a consistent trend across regions conditional on first concluding a significant overall treatment effect at the 0.05 significance level (two-sided test).

Based on our simulations for the case of three regions, the proportion of patients from the smallest region could be as low as 15% for us to have an 80% probability that the observed treatment effects are consistent across the three regions, under the assumption of a positive and uniform treatment effect across regions and conditioning on concluding a statistically significant overall treatment effect first.

5. Concluding Remarks

When the treatment effect is positive and uniform across regions, we could still observe a negative treatment effect in a region. What we propose to do in this paper is to have enough patients in each region so that the chance of observing a negative treatment effect in any region is controlled at a pre-specified level. In other words, our approach wants to control the chance for observing a *qualitative* region-by-treatment interaction when the treatment effect is positive and uniform across the regions. On the other hand, our approach is not concerned about the chance of observing a *quantitative* region-by-treatment interaction. From a regulatory perspective, interactions are often not equally important or concerning. They state that "qualitative as opposed to quantitative interactions are the most worrisome and difficult to interpret." We believe that our approach could serve as a starting point to discuss the scientific rationale for deciding the number of subjects for different regions in a multi-regional trial. The concept behind the approach is straightforward and is intuitively easy for people involved in planning the study to understand. The strengths of this approach lie in its mathematical simplicity and applicability to various types of endpoints.

医療に必要な科学的根拠とは何か

佐藤俊之
第一三共株式会社　データサイエンス部　統計解析グループ

1. 序に代えて

科学的根拠に基づく医療 (Evidence-based medicine, Sackett *et al.*, 1996) の実践が強調され浸透しつつある一方で (福原, 2006), 全人的医療の視点で医療と科学の関係を捉え直そうと主張する臨床家も少なくない (例えば丸橋, 2004: 米山, 2005: 渥美, 2007) ことから, 「患者のための薬」情報である「医療に必要な科学的根拠」とは何か, その生産に企業の統計家は何ができるのか, この機会に検討することとした.

企業の統計家が, 審査的視点である「疾患のための薬」としての価値を重視するあまり、患者の実像をを見ず「平均的な国際人」という虚像に処方する想定しかできない「科学的根拠」だけを提供していると危惧されるし，そのことがひいては医療者や患者に失望の因ともなりそうである. 医療に必要な科学的根拠, すなわち医薬品の真価把握のための情報を提供するという観点から, 定量科学者である企業の統計家は, 治療効果のプロフィールの推定, 適応疾患像の摘出などの「患者のための薬」情報の摘出のための統計的方法論の整備に取りかかる時期にきている.

2. 企業統計家が生産すべき医療に必要な科学的根拠とは何か : 審査に必要な「疾患のための薬」情報に加えて, 医療に必要な「患者のための薬」情報を考える必要性を共有する

歩くとすぐに息切れがするという主訴をもって, 56 歳の男性が医院を訪れた. 彼は NYHA(New York Heart Association) の心機能分類Ⅲの拡張型心筋症で, 左心室収縮不全 (左室駆出率: LVEF=31%) であり, 肺うっ血・肺水腫の所見が胸部 X 線から認められた. さらに，彼は高血圧の合併患者で ACE 阻害薬, ジギタリス, 利尿剤を半年前から服用している. 血清脳性 Na 利尿ペプチド (BNP) は 168pg/mL で, バイタルサインは, 心拍数が 110 拍/分, 血圧が 137/76 mmHg であった (参考 : 慢性心不全治療ガイドライン, 2005: 矢尾板・丸山, 2007: 蔦本・堀江, 2007). 現在の治療状況および頻脈であることから，β 遮断薬の追加投与が検討されるだろう. そうすると, 臨床論文に「新しい β 遮断薬が慢性心不全に対して, 標準薬に比し 1 年も生存時間の中央値が長かった」と記載されていて, この論文の適格性基準にこの患者が合致したときに (例えば Sackett *et al.*,1997), 喘息, 房室ブロック, 末梢循環障害のような禁忌の患者でないならば, この新しい β 遮断薬を用いようと判断されるかもしれない. あたかも臨床的根拠が患者を選んでいるかのようである.

患者が臨床的根拠を選べるようにするためには, 何に取り組むべきであろうか. 患者は年齢, 既往歴といった測定可能情報と人間性, 社会性といった測定不可能な情報を併せもつので, 企業の統計家は疾患の有無だけでなく測定可能なデータについても併せて探索的に吟味することで「患者のための薬」情報の摘出が目指すところ, 全人的治療・患者中心の医療へ近づく貢献ができると考

える. 医薬品の特徴をデータを通して評価するとき, その医薬品の総合的な解析, 例えば市販後を予測する「擬似市販後臨床評価」の解析などが欠落しているため, 企業の統計家は事後解析から洞察する姿勢,「経験に学ぶ」姿勢および知恵を熟成する過程を定式化する必要がある.

3. 擬似市販後解析の事例.

本稿では模擬データに多水準モデルを適合し, 母数空間に階層仮説を設定し, 情報量基準の意味で最適な仮説, すなわち最適なモデルを選定し, そのモデルが必要とした要因をもとに有効性および安全性の結果から, 患者プロフィールの類型化を行った. 今回のデータでは患者類型は 5 個にまとめることができた; Ⅰ) プロフィール 5 と 8(拡張型でかつ, NYHA がⅡ・心拍数が 100 拍/分未満または NYHA がⅢ-Ⅳ・心拍数が 100 拍/分以上), Ⅱ) プロフィール 2(虚血性でかつ, NYHA がⅡ・心拍数が 100 拍/分以上), Ⅲ) プロフィール 7(拡張型でかつ, NYHA がⅢ-Ⅳ・心拍数が 100 拍/分未満), Ⅳ) プロフィール 4 と 6(虚血性・NYHA がⅢ-Ⅳ・心拍数が 100 拍/分以上または拡張型・NYHA がⅡ・心拍数が 100 拍/分以上), Ⅴ) プロフィール 1 と 3(虚血性でかつ, 心拍数が 100 拍/分未満).

4. 事後解析に対する批判をどう捉えるか

今回のような事後解析結果の解釈にあたって, 事前に設定した仮説や多重性の留意がなされていないから使えないとの批判がある. しかし, 企業の統計家が新鮮なデータを用いて医薬品を「医療に必要な科学的根拠」という視点から評価するためには, 審査上で必要な管理値に囚われない洞察が必要になる. したがって, 解析の過程を明らかにし, 結果の利用の限界や品質を明確にしておくことで, 解析結果の利用可能性を評価することができると考える.

5. 結びに代えて

本稿では, 審査に必要な「疾患のための薬」情報の生産を重視するあまり, 医療に必要な「患者のための薬」情報の摘出に十分な議論ができていない現状に着目し, 第Ⅲ相臨床試験データが得られた時期を想定し, そのデータを用いた擬似市販後評価の事例を述べた. 連関要因図を始めとした, 適応患者像の摘出までの事後解析過程を明示した上で, その結果を解釈・運用することは有用であることを述べた. 今回の擬似市販後評価を複数の薬剤に対して実施すれば, 冒頭に紹介された 56 歳の心不全患者は次のような補完的考察が可能になる.「この患者は, 拡張型心筋症で NYHA がⅢで心拍 110 か. X 薬の適応患者像は … で、Y 薬の適応患者像は … か. 確率的な要素はあるが, X 薬が最も得意とする患者で, 他薬より, 効果がある可能性が高いな. 最大用量まで, 心不全, 徐脈等の副作用に注意しながら慎重に投与してみるか.」

遺伝子情報などを利用した相当な確度の予後因子が明らかになりつつあるが, 一方で今後にかけて少子高齢化, および, 病態の多様化が進み, 現象を把握するにはデータの多用な属性を考慮したより広範な推測過程の方法論を構築し, 精緻にし続けていく必要がある.