



SASユーザー総会2024論文集

会場：東京大学・弥生講堂

会期：2024年9月18日(水)～19日(木)

主催：SASユーザー会 世話人会

目次

論文部門

GAM を用いた非直線的な共変量の影響を調整した特定変数の効果の可視化

古川敏仁(株式会社バイオスタティスティカル リサーチ)

RTF に使える DDE

関根暁史(藤本製薬株式会社)

SQL の一貫性制約

関根暁史(藤本製薬株式会社)

Cochran-Mantel-Haenszel 検定の概要と SAS での実装方法

石本りさ(イーピーエス株式会社)

**オープンデータを活用した加古川市における見守りカメラ導入による狭域での犯罪事案抑止効果の検証
— 自転車盗の事例 —**

川向肇, 橋間花音, 中根滉大(兵庫県立大学)

**SAS がデータ解析の覇者になる最善の方策 —52 年間で完成した全判別データに対応できる世界初の
LSD の判別理論—**

新村秀一(成蹊大学), 新村秀樹(イノベストメント株)

SAS による経時的な 2 値データの解析方法

伊庭克拓, 浅野豊, 黒見真央, 鈴木聡, 松嶋優貴, 毛利誠(大塚製薬株式会社)

RTF Output Tips -SAS 初学者も SAS 熟練者もたのしめるものを目指して-

麻生美由紀(株式会社インテリム)

交互作用を含む共分散分析における最小 2 乗平均の活用

高橋行雄(BioStat 研究所株式会社)

古くてあたらしい共変量調整

森田祐介, 太田裕二, 浜田泉(ノーベルファーマ株式会社)

サンプリングのための標本設計と標本配分

高田浩成(イーピーエス株式会社)

ODS RTF での出力環境の整備 ~RTF スタイルと SAS マクロの運用・管理手順~

宮辻淳, 麻生美由紀, 平野勝也(株式会社インテリム)

3 つ以上の治療法を比較する場合の傾向スコア逆確率重み付け解析

中倉章祥, 岸本淳司(九州大学)

SAS でゲームを作る! ~PowerShell を橋渡し役とした SAS と Web ブラウザの相互連携~

伊住寿雄(ソーバル (株))

医薬・疫学研究向け SAS 用グラフパッケージ「SAS Plotter」

筒香孝輔

デザイン行列

中川雄貴(イーピーエス株式会社)

臨床試験データ解析における SAS と R の挙動差異

後藤駿輔, 松永友貴(ノバルティス ファーマ株式会社)

生成 AI を用いたデータセットの CDISC レビュー方法の模索

大山暁史(イーピーエス株式会社)

大規模言語モデルの新篇章：マルチモーダル AI の進化と展望

小野潔(コムチュア株式会社)

スクラムと CRISP-ML(Q)を併用した分析プロジェクトの進め方

吉野祥(株式会社クレスコ)

時間依存性 ROC 曲線に対する推定方法の性能比較

久貝航太郎, 魚住龍史(東京工業大学)

データステップを用いた統計検定の再現

折村奈美(イーピーエス株式会社)

Proc MI による欠測値補完の精度評価

小林邦世(イーピーエス株式会社)

コピュラモデルによる相関のある生存時間解析：SAS プログラムの開発

中園孝輔(統計数理研究所), 魚住龍史(東京工業大学), 江村剛志(統計数理研究所)

自主防災組織の ICT による支援の現状と課題

有馬昌宏, 川向肇(兵庫県立大学)

JAFROC(Jackknife alternative free-response receiver operating characteristic)解析の SAS による実装

別府汐(イーピーエス株式会社)

生存時間解析用 ADaM データセット (ADTTE) のソースコード詳解

田中真史(イーピーエス株式会社)

support vector machine を用いた格付推計モデルの構築—hpsvm procedure を用いたパラメータ推定—

坂巻英一(宇都宮大学)

寄席の「顔」が柳家から春風亭に変わるまで

坂部裕美子(公益財団法人 統計情報研究開発センター)

競合リスクを伴う生存時間解析について

樋口史(イーピーエス株式会社)

文字コードに関する SAS の対応

井上貴博(ノバルティス ファーマ株式会社)

抄録----プレゼンテーション部門

SAS の CALL ルーチンを可能な限り全て説明

森岡裕(イーピーエス株式会社)

Can SAS Monolinguals who make full use of Programming Code Translator Survive?

仲家諒(武田薬品工業)

A modelling framework for regression with collinearity

刈屋武昭(名古屋商科大学)

Dataset-JSON の SAS での実装と SAS 拡張属性の利用について

森岡裕(イーピーエス株式会社)

YAML ファイルの SAS での作成方法について

森岡裕(イーピーエス株式会社)

SAS によって描くグラフ軸目盛の自動調整

魚田孝志(イーピーエス株式会社)

SAS による個別データの可視化

徳田芳稀, 杉本真理(エイツーヘルスケア株式会社)

**SAS SAMPLE LIBRARY で公開されている Kaplan-Meier プロットの
カスタマイズマクロ%ProvideSurvivalMacros を知っていますか？**

—Graph Template Language (GTL) への第一歩にも—

太田裕二, 森田祐介, 浜田泉(ノーベルファーマ株式会社)

データステップ内で次の obs の値を参照する

浅井友紀(エイツーヘルスケア株式会社)

生存時間解析における 3 次スプラインを用いた事例の紹介

室永遼太郎, 浦狩保則(シミック株式会社)

視覚障害者を支える SAS プログラミング-Accessible Output と SAS Graphics Accelerator

森岡裕(イーピーエス株式会社)

**MCMC Procedure による混合モデルの実装とクラスタリング：1 次元混合正規モデルから多変量時系列
モデルまで**

福島綾介(イーピーエス株式会社)

AI-SAS for RWE の開発

米田卓司, 惟高裕一, 木口亮, 北西由武(塩野義製薬株式会社)

R プログラムを生成系 AI を用いて SAS に変換してみた事例について

三木悠吾, 森口泰孝(グラクソ・スミスクライン株式会社)

SAS Extension for Visual Studio Code の紹介

山添滉弥(エイツーヘルスケア株式会社)

打ち切りをもつ連続値評価項目の解析～意外と知られていない解析法たち～

馬場崇充, 渡辺秀章, 長谷川貴大(塩野義製薬株式会社)

SAS ハッシュオブジェクト基礎編

松永友貴(ノバルティス ファーマ株式会社)

解析結果 RTF ファイル検証手法の検討と報告

山野辺浩己(マルホ株式会社)

SAS で作成した画像データの外部への出力と TIPS

山野辺浩己(マルホ株式会社)

文字はなぜ切れる？変数長の決まり方

田中真史(イーピーエス株式会社)

傾向スコアマッチング前後の群間バランスの定量的な評価指標に対する考察

上野航太郎, 魚住龍史(東京工業大学)

データサイエンスを支えるデータエンジニアリングのチカラ —SAS Viya を中核とする SHIONOGI 統合解析環境—

坂井聡, 北西由武(塩野義製薬株式会社)

2 値型応答における意外と知られていない因果効果の推定方法 ~なんとなくオッズ比からの脱却~

堀江悠生, 馬場崇充, 渡辺秀章, 長谷川貴大(塩野義製薬株式会社)

SAS 新製品概要~LLM ガバナンス SAS Viya Workbench、Copilot、そして SAS Data Maker

土生敏明, WilliamKuan(SAS Institute Japan)

疫学研究で2群を比較する際に用いる確認手法について

矢田徹(イーピーエス株式会社)

Kappa 係数の意外な落とし穴とその対処法の紹介

江口凌平, 菊池宏和(シミック株式会社)

大規模言語モデルから動かす SAS

中松建

共分散分析の基礎

藤島直隆(株式会社分析屋)

~チュートリアル: ~SGPLOT で広がる Kaplan-Meier プロットの世界

魚住龍史(東京工業大学)

アウトカムを説明するモデル（非直線的な関係も含むモデル）への特定変数の調整後の寄与や直線性を可視化する-

SAS GAM を用いて-

古川敏仁

株式会社バイオスタティスティカル リサーチ

要旨

応答変数と説明変数間の関係が、一般化線形モデルでモデル化できるほど直線的ではない場合、共変量の影響を取り除く方法や、調整後の関係や寄与の大きさを可視化する GAM を用いた方法を解説する。

Big Data, GAM(generalized additive model), Adjusting, Visualization

1.はじめに

1つのアウトカムを説明する複数の変数が存在するとき、最も効率的な項目の組み合わせは何か？ その組み合わせの中でアウトカムを説明する特定の変数の寄与はどれくらいか？あるいは、他の変数の影響を取り除いた特定の変数とアウトカムとの関係はどのような形状を持っているのか？などは、多変量回帰モデル（線形、あるいは非線形においても）、常に検討課題となる重要な事項である。しかし、残念ながら、これらを明確に示す指標は見当たらない。今回は GAM を用いて視覚的にこれらを確認し、線形（一般化）モデルなどの解析結果と相補する手法を紹介する。

今回取り上げる例は 8749 例の非心臓手術中後の急性腎障害（acute kidney disease; AKI）に関するデータである（提供：東京慈恵会医科大学 坪川 恒久 麻酔科教授、エドワーズライフサイエンス協賛）。本データの研究主題は手術中低血圧時間（65, 60, 55mmHg 未満の総時間）が AKI とどのように関連するかであった。そのため、術中血圧時間に関しては 75 ~40mmHg まで 5mmHg 間隔で収集されたデータが存在した。

このデータから以下の 2つの点が着目される。

手術中低血圧時間以外のデータはアウトカム AKI に影響しないのだろうか？ 影響がとすれば、手術中低血圧時間以外のデータの影響を取り除いたとき、手術中低血圧時間以 AKI とまだ関連するのだろうか？

例えば、手術中血圧時間以外にも数多くの項目が測定され、例えばその中の 1つに手術時間

がある。手術中低血圧時間はどんなに長くなっても手術時間以上にはならないはずであるし、長い手術時間の人は手術中低血圧時間も長くなる可能性は高くなるはずである。つまり、手術時間などの項目は手術中低血圧時間と関係を持ち、お互いの評価にはそれぞれの影響を考慮した解析が必要となる。

術中血圧時間に関して、75~40mmHg まで 5mmHg 間隔で 8 項目あるが、どの血圧値が最もアウトカムと関係するのだろうか？ お互いの関係は密接にしているのは手術時間と血圧時間の関係のように明白であり、お互いの関係を考慮しなければ、やはり、どの血液項目が一番関係しているかはわからないであろう。

2 調整(Adjusting)

応答変数に対する当該項目の影響を、①当該以外の項目の影響を取り除いたり、②当該項目値と当該項目以外の項目値の線形結合による新たな情報を付加したりすることを調整と定義するが、ここからは、①の当該項目とアウトカムの関係から他の項目の影響を取り除くことに主眼を当てて論議する。アウトカムを Y 、関係を調査する項目を X 、 Y, X 間の関係から影響を取り除きたい項目の共変量セットを $X' = \{X_1, X_2, \dots, X_k\}$ としたとき、最も確実な方法は、共変量セット 1 組の値を固定 ($x' = \{x_1, x_2, \dots, x_k\}$) し、共変量に x' の値を持つ Y と X の値を収集し X と Y の関係を調べることである。しかしながら、この方法には以下の問題がある。

問題 1： いかにか **BIG DATA** といえども、 k 個の変数が同じ値を持つオブザベーションを十分な数だけ収集するのは容易ではない (特に k が大きいときは)。

問題 2： 説明変数 X と共変量 X' との間に関連がある場合、共変量 X' が特定の固定値を持つ場合、 X の範囲が限定され、 X と Y との関係を調査するには X の範囲が狭くなりすぎる可能性がある。

問題 3： そもそも、説明変数 X に影響を与える共変量セット X' の値はどうするかという問題がある。当然、共変量セット X' の値が違えば X と Y の関係も違ってくる可能性が高い。 X' の取りうる領域すべてにわたって、固定値 X' のもとで十分な数の X と Y の値の組み合わせを求めることは不可能である。

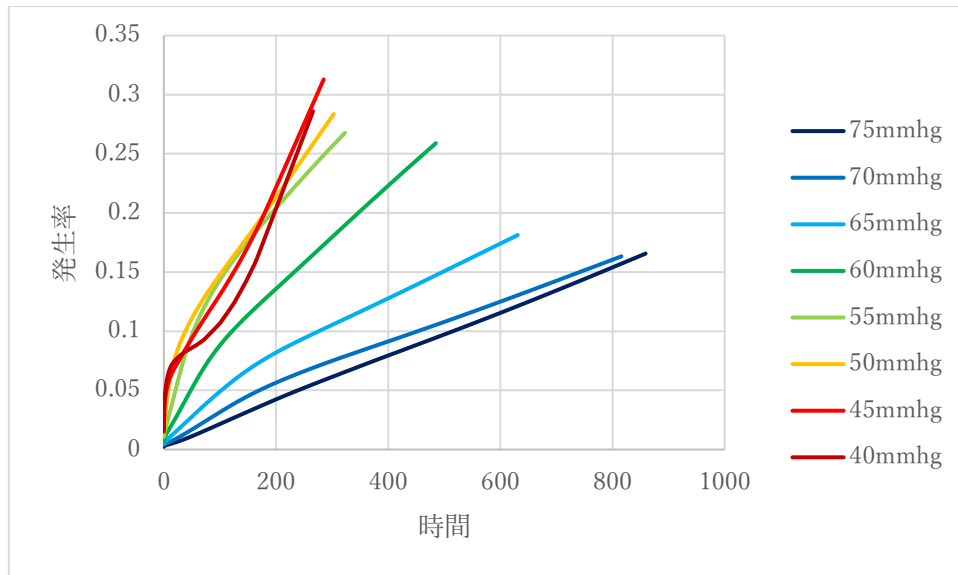
そこで、 X と Y との関係を共変量セット X' で調整するには何らかの数学的モデルが必要になる。ただ、線形モデル、線形混合効果モデル、一般化線形モデル等は良く用いられるが、いずれも、 X ($+X'$) と Y の間には線形を仮定しており、近年の **BigData** 解析結果からは、この線形の仮定が明らかに成り立たない事例もある。**GAM**(generalized additive model)は、1 変数 X' に対する Y の平滑化曲線を用いることにより、非直線の関係に「対応することができ、 Y と複数の X' の関係も、1 変数 X' に対する Y の平滑化曲線を、複数の X' に関して加法的に取り扱うことにより、 Y と X' の多変数の関係も、直線性の仮定なしにモデル化できる。つまり、線形モデル類よりは、より、真の Y と X との関係に迫ることができる可能性

を持ったモデルである。

3 GAMによる実例

8749 例の AKI データ (AKI 発生数 165 例 1.89%) で実際の GAM の結果について述べてみよう。このデータの目的は「術中平均血圧が一定以下になったトータル時間が AKI 発生と関連するか? また、関連するとすれば、どの血圧が最も AKI と関連するか?」を確認することである。血圧は 75mmHg から 40mmHg まで 5mmHg 間隔で 8 項目測定されている。AKI 発生率をそれら項目値ごとの SPLINE 平滑化曲線を求めると Fig1 になる。

Fig1 術中平均血圧が XXmmHg 以下になった時間と AKI 発生率の単独の関係を GAM による可視化



発生率を求める SAS プログラムは以下である。

```
/*発生率推移を求める SAS プログラム例*/;
```

```
PROC GAM DATA=DATASET ;
```

```
MODEL AKI(EVENT='1') = SPLINE(MMHG, DF=3)
```

```
/DIST=BINOMINAL;
```

```

OUTPUT OUT=OUT ALL ; RUN;
PROC SORT DATA=OUT;BY MMHG ;RUN;
FILENAME EXL DDE "EXCEL |FIG1!R5C3.:R500C4";
DATA _NULL_ ;
  SET OUT ;
  FILE EXL NOTAB; BY MMHG ;
  T='09'X; IF FIRST.MMHG;
  PUT MMHG T P_AKI ;
RUN;

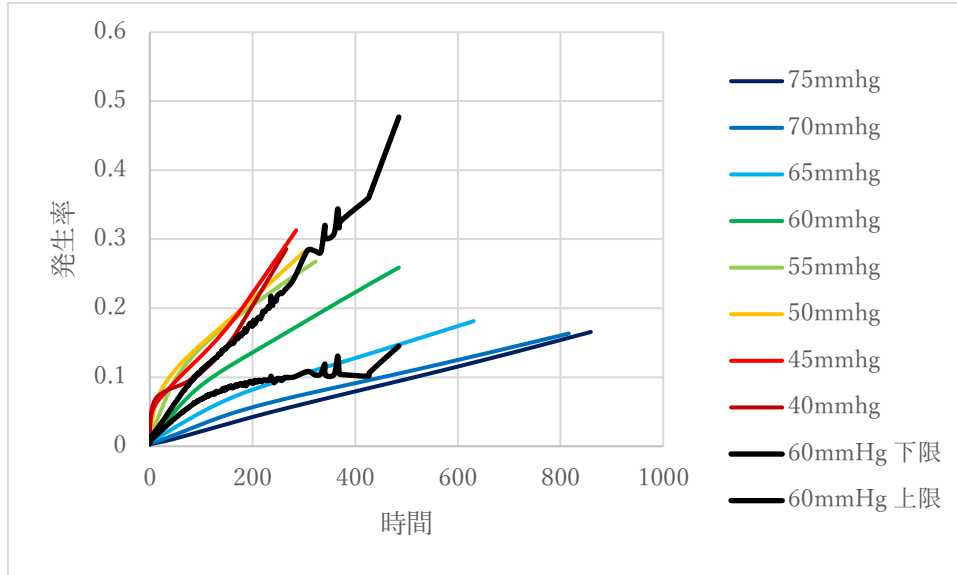
```

この例では、AKI: (0,1)応答の1に対応する発生率を血圧値 MMHG の自由度3のスプライン曲線で GAM を用いて算出し、対応する共変量（この場合はMMHG）に対する予測 AKI 発生率を OUTPUT ステートメントにて、OUT という SAS データセットに出力している。

グラフを見てみると、75mmHg~40 mmHg 以下になった時間に対する AKI 発生率は、時間0近辺ではほぼ0%であるのに対し、いずれも時間が大きくなればなるほど AKI 発生率が増加していることがわかる。また、各項目の AKI 予測確率最大値付近では、血圧が 60mmHg 以下の場合、AKI 発生率は25%以上高くなることがわかる。つまり、術中血圧が低下し低下時間が大きくなれば AKI 発生リスクも大きくなることが示されている。

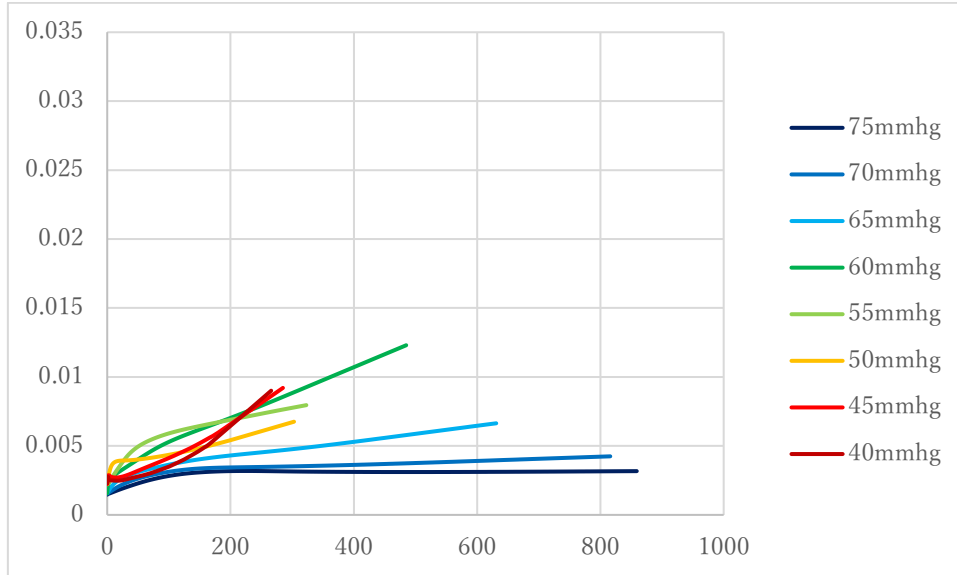
この曲線間に差はあるのだろうか？試しに 60mmHg の95%信頼区間曲線を boot-trap 法で求めてみると (Fig2)、75,70mm 以下時間よりは低血圧の方がより AKI と関連するだろうということはわかるが、信頼区間は広くそれ以上のことはこれ以上わからない。ただ、いずれの血圧時間においても曲線は単調増加であり、ある程度、Logistic モデルのようなモデルと適合することが Fig からは示唆される。

Fig2 術中平均血圧が XXmmHg 以下になった時間と AKI 発生率の単独の関係の GAM による可視化 60mmHg 以下時間の 95%信頼区間とともに



術中 XXmmHg 以下時間の特徴は、手術時間などに大きく影響を受ける。つまり、手術時間以上には術中 XXmmHg 以下時間は長くなることもないし、また、その他の手術環境、患者の背景因子にも大きく影響を受けるはずである。つまり、正味の術中 XXmmHg 以下時間と AKI 発生率の関係を見るためには、手術時間に代表される手術情報や患者背景情報の影響を除外しなければ、正味の術中 XXmmHg 以下時間と AKI 発生率の関係は観察できないことになる。種々の検討の結果、術中 XXmmHg 以下時間以外では、ASA PS (4 区分)、手術時間等 10 項目が AKI 発生率と関連することがわかっている。そこでこれら項目の影響を除去した術中 XXmmHg 以下時間と AKI の関係を GAM で可視化した。

Fig3 術中平均血圧が XXmmHg 以下になった時間と AKI 発生率の関係 他の共変量で調整した場合



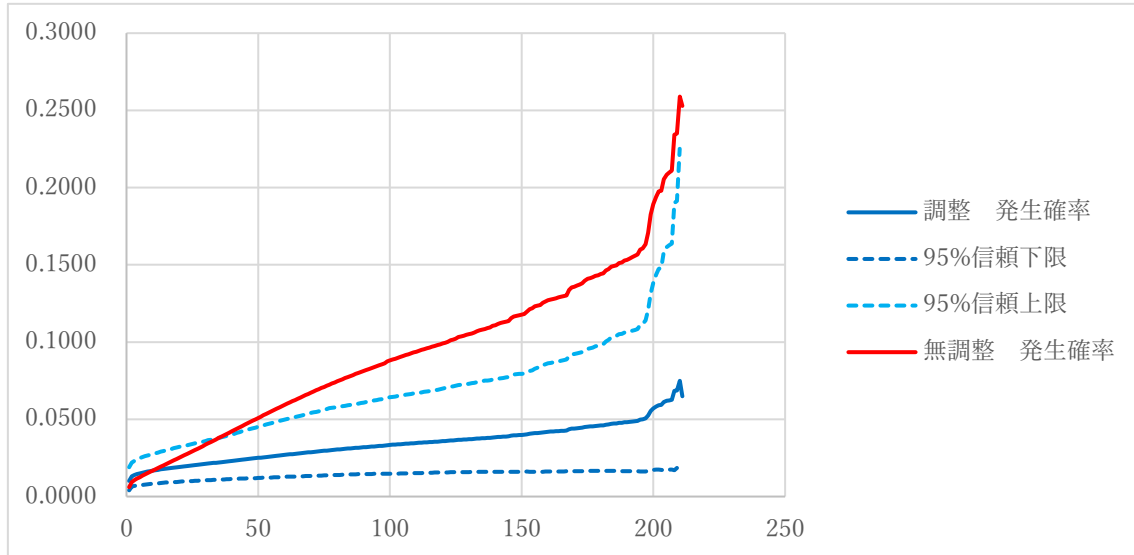
これは、AKIに影響する共変量の影響を除外した XXmmHg 以下血圧と AKI の関係である。この描画のための SAS プログラムは以下である。

```
DATA FRM;Set MMHG ;
  VAR1=A1;VAR2=A2;ASA_PS=1; . . . . . ;
OUTPUT;RUN;

PROC GAM DATA=DATASET ;
CLASS ASA_PS(REF='1');
MODEL AKI(EVENT='1') = SPLINE(MMHG, DF=3) PARAM(ASA_PS )
SPLINE(VAR1, DF=3) SPLINE(VAR2, DF=3) . . . . . ;
/DIST=BINOMINAL;
SCORE DATA=FRM OUT=outw ; RUN;
l; RUN;
```

GAM モデルで求めた特定の共変量セットの値は SCORE ステートメントによって求めることができる。FRM データセットに MMHG 項目の値と、特定の共変量セットの値を格納している。今回（一般的には）は、評価集団を代表する値を格納する（特に調べたい共変量セットの値でもよい、本文最後に、複数の共変量セットを使った例を表示した）出力された OUTW データセットに特定の共変量セットに対する期待値が格納される。

Fig5 術中平均血圧が 60mmHg 以下時間特異 AKI 発生率 調整 (95%信頼区間)、無調整



AKI 発生予測能に関して、Fig3 からは 0 分からの最大値までの変化は 60mmHg 以下時間が最も大きく、他の以下時間の規模はこれの半分以下である。そこで、10 共変量を含む Logistic モデルに、それぞれの XXmmHg 以下時間を入れてみると、p 値、AIC も 60mmHg が最も AKI 予測能が大きくなることがわかる (Table1)。また、共変量 10 個と XXmmHg 以下時間 1 項目を入れたモデルに他の XXmmHg 以下時間項目を入れたモデルでは 60mmHg 以下時間項目は必ず $p < 0.001$ 以下項目として残るが、逆に共変量 10 個と 60mmHg 以下時間 1 項目を入れたモデルに他の XXmmHg 以下時間項目を入れたモデルではいずれの他の XXmmHg 以下時間の残存 AKI 説明量はほぼ 0 であることが GAM (Fig6) でも Logistic モデル (Table2) でも示された。

以上のことから AKI を予測する性能という点で、60mmHg 以下時間項目が、他の時間項目よりも優れていることが示された。また、GAM による多変量モデルへの可視化は、多変量モデルにおける、各項目の寄与を可視化する優れた特性を持つことが示された。

ここまでで、AKI を予測するのに効果のある情報を持っているのは、一定血圧の時間の中では 60mmHg 以下時間項目であることは明確になった。では、60mmHg 以下時間を制御するとどの程度の AKI 発生を制御できるのであろうか？という根本的な疑問に到達する。残念ながら、その答えは、このデータからは完全には得ることができない。なぜなら、今回得られたデータはレトロスペクティブであり、プロスペクティブ介入試験ではないなら、因果関係のある仮説は証明できないからである。そこで、もし、60mmHg 以下時間と AKI との関係が、60mmHg 以下時間を制御する関係にそのまま保たれるという大きな仮定をした場合について簡単に以下の記述してみる。

まず、集団全体としての 60mmHg 以下時間の変化による AKI 発生率の変化は、Fig5 (他の因子の影響を取り除いた場合) に見られるごとく、最小 0 最大 5% 程度である。医学的な知見はないが、非常に小さな変化であると思われる。これを、手術時の患者状態を表す ASAPS 別に調べてみると Fig7 のようになる。GAM の特徴として、加法性の仮定があるから、ASA PS が違ってても、各 ASA PS 水準内では、60mmHg 以下時間の変化によって AKI 発生率は同じような倍率で増加することを示すことになる。(本、GAM は交互作用項を含まないためこのような変化になる。これが正しいかは、交互作用を入れたさらなる検討が必要になるが、データ数の関係で今回は難しいので、今回は、交互作用を含まないモデルでの記述をする)。

おおよそ、どの水準でも 60mmHg 以下時間が 0 から 30 分増えると AKI 発生リスクが 2 倍になる。全体 8737 例の 32.9% を占める ASA PS=1 集団 (60mmHg 以下時間 : 例数、最小値、下側四分位点、中央値、上側四分位点、最大値 ; 2876、0、0、3、11、428) では Fig7 青実線にあるように、0 分時の Baseline リスクが、非常に小さいため、全体を通して 10 倍ほどリスクが高まっても、AKI リスクは非常に小さい。また、全体の 60.4% を占める ASAPS=2 (5281、0、1、4、14、426) でも AKI リスクはそれほど大きくなく、また、0 分からのリスクが 2 倍となる前の 30 分以内の症例はこの水準内の 80% 占めている。つまり、全症例中の 90% は 60mmHg 以下時間による AKI 発生リスクは非常に小さいといえる。ある程度の AKI 発生リスクが、60mmHg 以下時間に影響を受ける ASAPS=3 (573、0、0、4、18、485) でも 30 分以上の変化は全集団の 1.5% 程度であり。また、ハイリスクな ASAPS=4 (7、0、3、18、112、120) においては、わずか 1,2 例である。臨床的な知識はないからこれ以上のコメントはできないが、これらは、手術時の血圧管理をどうするかという問題においては、重要な情報になると思われる。

以上まとめると、SAS GAM モデルを用いることにより、特定項目と応答変数との関係を直線性の仮定抜きに明示することができ、さらに、特定の共変量の影響を除いた、関係も明示できる。また、特定項目の変化に応じた応答変数期待値の変化を確認することにより、特定項目の応答変数の説明能力が確認でき、多変量のための影響を除いた特定項目と応答変数の関係を項目間で比較すれば、複数項目内の相対的な応答変数説明能力を知ることができる。GAM にはこのような優れたデータの可視化能力がある。

Table1. 共変量 10 項目とともに XXmmHg 項目を 1 つ AKI 予測モデルに組み込んだ時の p 値、AIC

		95%信頼区間			p値	AIC	最大値調整係数 RSQ
		Odds比	下限	上限			
75mmHg以下時間	平方根・平方根2	1.287	0.949	1.745	0.1047	1166	0.327
70mmHg以下時間	平方根・平方根2	1.313	1.023	1.685	0.0325	1164	0.329
65mmHg以下時間	平方根・平方根2	1.396	1.123	1.735	0.0027	1159	0.332
60mmHg以下時間	平方根・平方根2	1.518	1.239	1.861	<.0001	1152	0.336
55mmHg以下時間	平方根・平方根2	1.312	1.074	1.604	0.0079	1162	0.330
50mmHg以下時間	平方根・平方根2	1.289	1.032	1.611	0.0253	1164	0.329
45mmHg以下時間	平方根・平方根2	1.206	0.919	1.582	0.1760	1167	0.327
40mmHg以下時間	平方根・平方根2	1.116	0.802	1.553	0.5136	1168	0.326

Fig6 共変量 10 項目と術中平均血圧が 60mmHg 以下時間を含む AKI 予測 GAM モデルにおける XXmmHg 以下時間の残存説明能

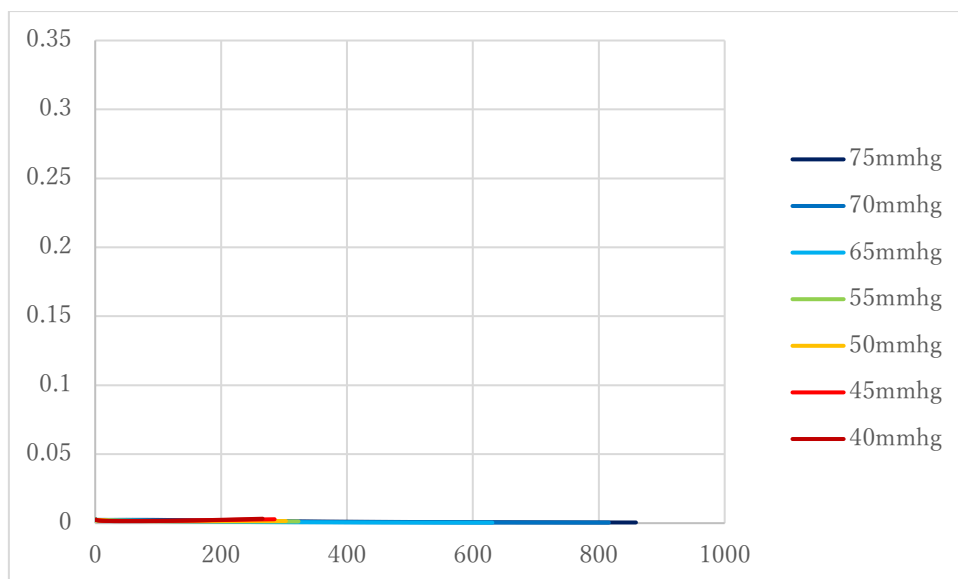
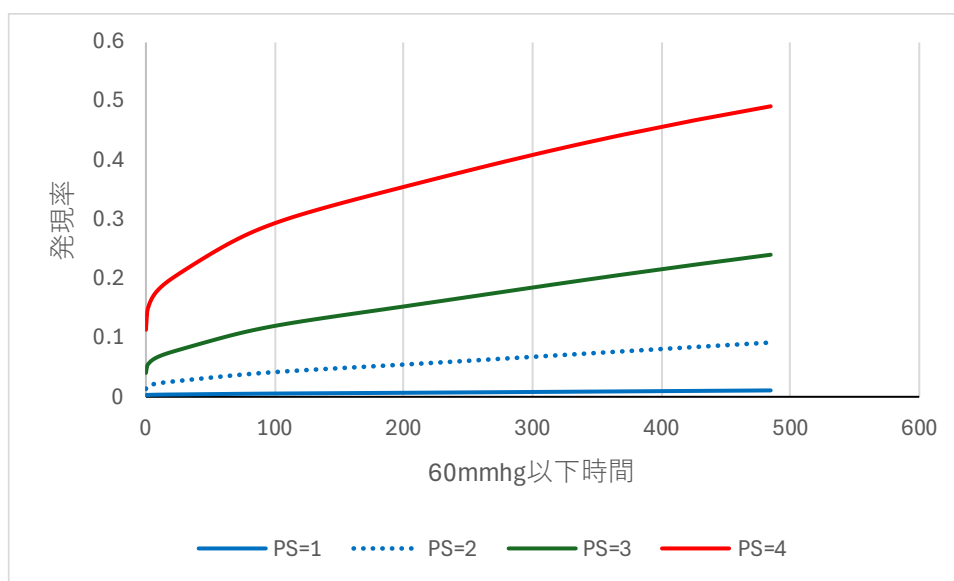


Table2. 共変量 10 項目と 60mmHg 項目を入れた Logistic モデルに他の XXmmHg 項目を 1 ついれ時の p 値、AIC

		95%信頼区間			p値	AIC	最大値調整
		Odds比	下限	上限			
75mmHg以下時間	平方根・平方根2	0.806	0.554	1.173	0.2598	1153	0.337
70mmHg以下時間	平方根・平方根2	0.774	0.534	1.122	0.1767	1152	0.337
65mmHg以下時間	平方根・平方根2	0.753	0.471	1.206	0.2380	1153	0.337
55mmHg以下時間	平方根・平方根2	0.791	0.564	1.109	0.1735	1152	0.337
50mmHg以下時間	平方根・平方根2	0.958	0.730	1.258	0.7597	1154	0.336
45mmHg以下時間	平方根・平方根2	0.949	0.709	1.272	0.7275	1154	0.336
40mmHg以下時間	平方根・平方根2	0.924	0.659	1.294	0.6446	1154	0.336

Fig7 ASA-PS 水準（共変量は各水準における平均値）ごとの、術中平均血圧が 60mmHg 以下時間特異 AKI 発生率 調整 GAM モデル



参考文献 (1) ‘How to use the GAMs for Big Data’,古川敏仁, SUJI-J2023

RTFに使えるDDE

○関根 暁史
(藤本製薬株式会社)

DDE programs to process RTF files.

Satoshi Sekine
Fujimoto Pharmaceutical Corp.

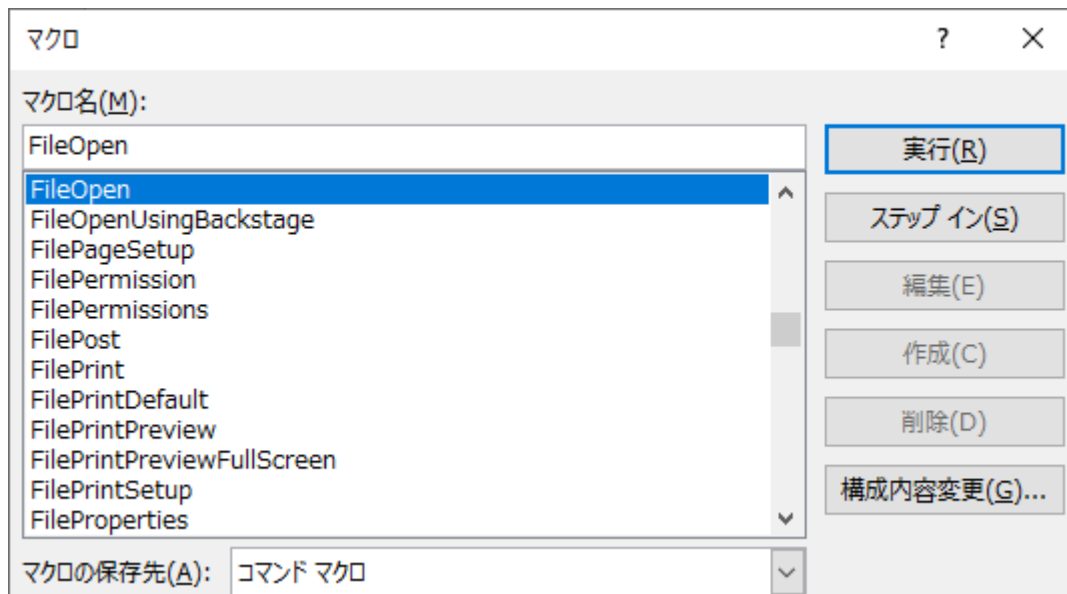
要旨

SAS の DDE は Excel だけでなく RTF(Word)にも適用できる。DDE を使用しての Word ファイルの体裁変更法等を紹介する。

キーワード： DDE(Dynamic Data Exchange), RTF, Word, SAS

1. はじめに

DDE は Excel にしか使用できないように見えるが、RTF(Word)にも適用することができる。Excel の DDE は Excel 4.0 マクロがベースになっていたが²⁾、Word の DDE はコマンドマクロがベースになっていると思われる。コマンドマクロに引数を持たせると DDE が動作することが多いようだ。



Word の DDE 技術を使用して、RTF の拡張子の変更、RTF ファイルの結合、RTF の余白変更、RTF のフォント変換、RTF の Excel へのコピー、文字列変換を利用した症例構成図の作成、RTF ファイルへ

のグラフ貼付、RTF ファイルへの文字列の書き込み、RTF ファイルへのページ番号の挿入、Word マクロを操作しての RTF の変更履歴を書き出す方法等を紹介する。

DDE への接続は filename ステートメントの使用で可能となるが、DDE 使用前には X ステートメントや call system を使用して事前に Word を立ち上げておく必要がある。起動した Word に触れるには sleep 関数等を使用して、適度な待機時間を置かねばならない。

```
data _null_;  
  call system("C:¥Program Files (x86)¥Microsoft Office¥root¥Office16¥WINWORD.EXE");  
  rc=sleep(3);*3秒待機;  
run;
```

2. RTF の拡張子の変更

一旦開いた RTF ファイルは、保存して閉じる際に拡張子を変更することができる。以下では RTF ファイルを Word ファイル(docx)に変換している。docx にしたい時の引数は 14 である。RTF を Word 化することによってファイルサイズを圧縮できることが多い。

```
filename cmd dde "winword|system";  
%let kakko=[;  
proc stream outfile=cmd resetdelim="br"; begin%nrstr(&kakko.)  
On Error Resume Next] br newline;  
%macro Convert(_infile=,_outfile=);  
[FileOpen("&inpath.¥&_infile.")] br newline;  
[FileSaveAs("&inpath.¥&_outfile.",14)] br newline;*docxに変更;  
[FileClose] br newline;  
%mend Convert;  
%Convert(_infile=Class.rtf,_outfile=Class.docx)  
  
[FileExit] br newline;  
:::
```

FileSaveAs の引数と拡張子の関係を以下の表にまとめた。直接 PDF 化することもできる。

引数	拡張子
2	.txt
4	.rtf
14	.docx
17	.pdf

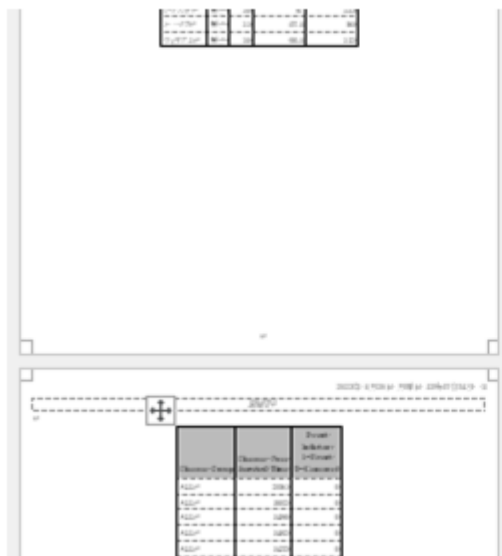
3. RTF ファイル結合

吉崎ら(2010)³⁾は、VBScript によってセクション区切りを入れながら複数の RTF ファイルの結合を試みていたが、DDE でも同様のことができる。VBSscript の Movedown は [EndOfDocument] に、InsertBreak 2 は [InsertBreak(2)] に対応している。以下は Class.rtf と BMT.rtf を連結して RTF 結合.rtf の名前で保存している例である。

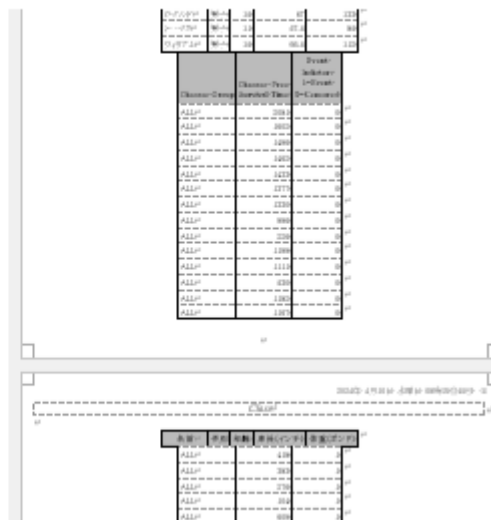
ただし残念ながら DDE では印刷の向きを考慮しながらの結合はできない。1 ページ目のファイルの印刷の向きが優先されることとなる。

```
[FileOpen("&first.¥Class.rtf")] br newline;*先頭ファイル;  
  
%macro Add(_file=);  
[EditSelectAll] br newline;  
[EndOfDocument] br newline;*下に移動;  
[InsertBreak(2)] br newline;*セクション区切りを入れたい場合;  
[InsertFile("&inpath.¥&_file.")] br newline;*追加ファイル;  
%mend Add;  
%Add(_file=BMT.rtf)  
  
[FileSaveAS("&outpath.¥RTF 結合.rtf"),4] br newline;
```

セクション区切りを入れた場合



セクション区切りを入れない場合



4. RTF ファイルの余白変更

Word の余白を、左 2.5cm、それ以外を 2.0cm 空けたいとする。DDE では単位がインチとなっているので、0.0353 で割れば余白がセンチの単位で空くことになる。options や proc template で余白を設定して RTF 出力できるため当該機能は不要のように見えるが、既存の余白設定がなされたファイルの再設定や新規 Word ファイルの設定に有用である。

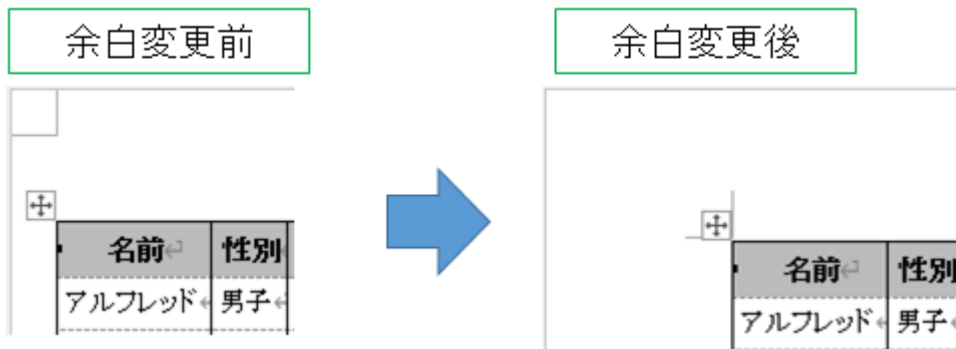
```
[FileOpen("&inpath. %&_infile. ") ] br newline;  
[PageSetupLayout. RightMargin=(2.0 /0.0353) ] br newline; *右2.0cm;  
[PageSetupLayout. LeftMargin=(2.5 /0.0353) ] br newline; *左2.5cm;  
[PageSetupLayout. TopMargin=(2.0 /0.0353) ] br newline; *上2.0cm;  
[PageSetupLayout. BottomMargin=(2.0 /0.0353) ] br newline; *下2.0cm;  
[FileSaveAs("&outpath. %&_outfile. ",4) ] br newline;
```

ページ設定

文字数と行数 余白 用紙 その他

余白

上(I): 20 mm 下(B): 20 mm
左(L): 25 mm 右(R): 20 mm



5. RTF ファイルのフォント変換

eCTD では、英数字フォントは Times New Roman を、日本語フォントはMS 明朝を使用することがお決まりとなっている。DDE を使用し、全体を選択して「MS 明朝」+「Times New Roman」+「9pt」に変換することとする。

```
[FileOpen("&inpath. %&_infile. ") ] br newline;  
[EditSelectAll] br newline; *全体を選択;
```

```
[Font("MS 明朝")] br newline;
[Font("Times New Roman")] br newline;
[FontSize(9)] br newline;
[FileSaveAs("&outpath. %&_outfile.",4)] br newline;
```

フォント変更前

名前	性別	年齢	身長(インチ)
アルフレッド	男子	14	69
アリス	女子	13	56.5
パーバラ	女子	13	65.3



フォント変更後

名前	性別	年齢	身長(インチ)
アルフレッド	男子	14	69
アリス	女子	13	56.5
パーバラ	女子	13	65.3

6. RTF から Excel へのコピー

DDE ならば RTF ファイルの内容をコピーして Excel に貼り付けることができる。

```
[FileOpen("&inpath. %&_infile.")] br newline; *Wordを開く;
[EditSelectAll] br newline; *全体を選択;
[EditCopy] br newline; *Wordの内容をコピー;

[new(1)] br newline; *新規Excelファイルを開いて;
[paste] br newline; *ペースト;
[save.as("&outpath. %&_outfile.")] br newline; *Excelとして保存;
```

Word の情報は直接 SAS で読み込むことはできないので、Excel に変換できたならば SAS でインポートして活用もできることになる。これにより RTF のコンペアを SAS することも可能となろう。

Before.rtf

Disease Group	Disease-Free Survival Time	Event Indicator: 1=Event 0=Censored
ALL	2081	0
ALL	1602	0
ALL	1496	0



After.xlsx

	A	B	C
	Disease Group	Disease-Free Survival Time	Event Indicator: 1=Event 0=Censored
1			
2	ALL	2081	0
3	ALL	1602	0
4	ALL	1496	0

7. Excel から RTF へのコピー

もちろん Excel ファイルの内容をコピーして RTF にペーストすることもできる。

```
[open("&inpath. %&infile. ")] br newline; *Excelを開く;  
[workbook.select("Sheet1")] br newline; *シート選択;  
[select("r1c1:r20c5")] br newline; *レンジ 選択;  
[copy] br newline; *Excelの内容をコピー;  
  
[FileNew] br newline; *新規Wordファイルを開いて;  
[EditPaste] br newline; *ペースト;  
[FileSaveAs("&outpath. %&outfile. ", 4)] br newline; *RTF として保存;
```

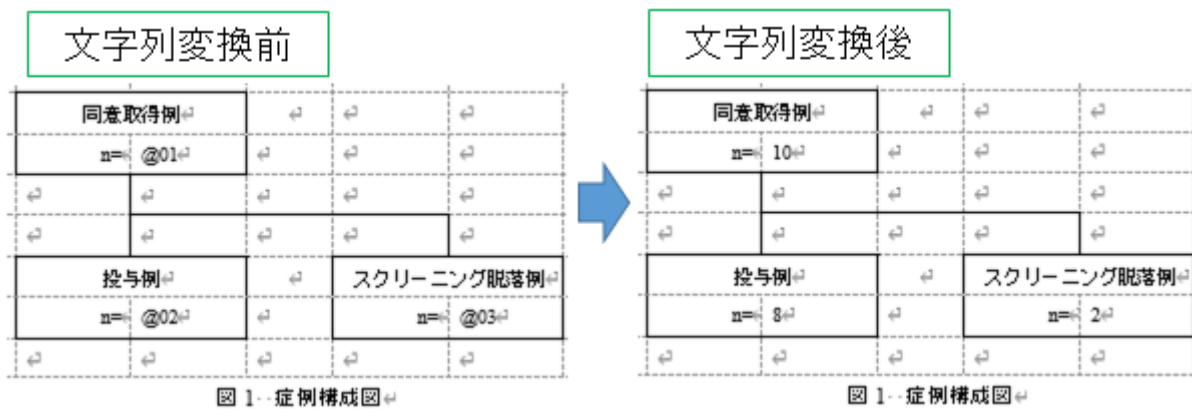
Excel で作った帳票を CSR や CTD に貼付するのに有用であろう。

Before.xlsx		After.rtf			
		名前	性別	年齢	身長(インチ)
1	アルフレッド	男子	14	69	
2	アリス	女子	13	56.5	
3	バーバラ	女子	13	65.3	
4	キャロル	女子	14	62.8	
5	ヘンリー	男子	14	63.5	
6	ミシャエル	男子	14	57.2	

8. 症例構成図の作成

例数が入るべきところに予め@01 などと打ち込まれているテンプレートを用意しておく。DDE の文字列置換機能を使って、その@01 を 10(例)などと置き換えてしまえば、簡単に症例構成図を作ることができる。

```
[FileOpen("&_inpath. %&infile. ")] br newline;  
[EditFind("@01")] br newline;  
[TableFormula("=10")] br newline;  
[EditFind("@02")] br newline;  
[TableFormula("=8")] br newline;  
[EditFind("@03")] br newline;  
[TableFormula("=2")] br newline;  
[FileSaveAs("&_outpath. %&outfile. ", 4)] br newline;
```



9. RTF ファイルへのグラフ貼付

複数枚のグラフを1つのパネルにすることは容易なことではない。しかし SG PLOT で描いた1つ1つの画像ファイルを RTF の1ページに貼付していく方法は比較的容易である。このような方法ならば個々例の臨床検査値図や薬物濃度プロファイルの作成が可能である。以下は RTF に6枚の画像ファイルを挿入する DDE プログラムである。EMF ファイルであっても構わない。この方法であれば画像の解像度の調整も効きやすい。

```
[FileOpen("&outpath. %&_outfile. ") ] br newline;
[InsertPicture("&inpic. %F1. emf") ] br newline;
[InsertPicture("&inpic. %F2. emf") ] br newline;
[InsertPicture("&inpic. %F3. emf") ] br newline;
[InsertPicture("&inpic. %F4. emf") ] br newline;
[InsertPicture("&inpic. %F5. emf") ] br newline;
[InsertPicture("&inpic. %F6. emf") ] br newline;
[FileSave] br newline;
```

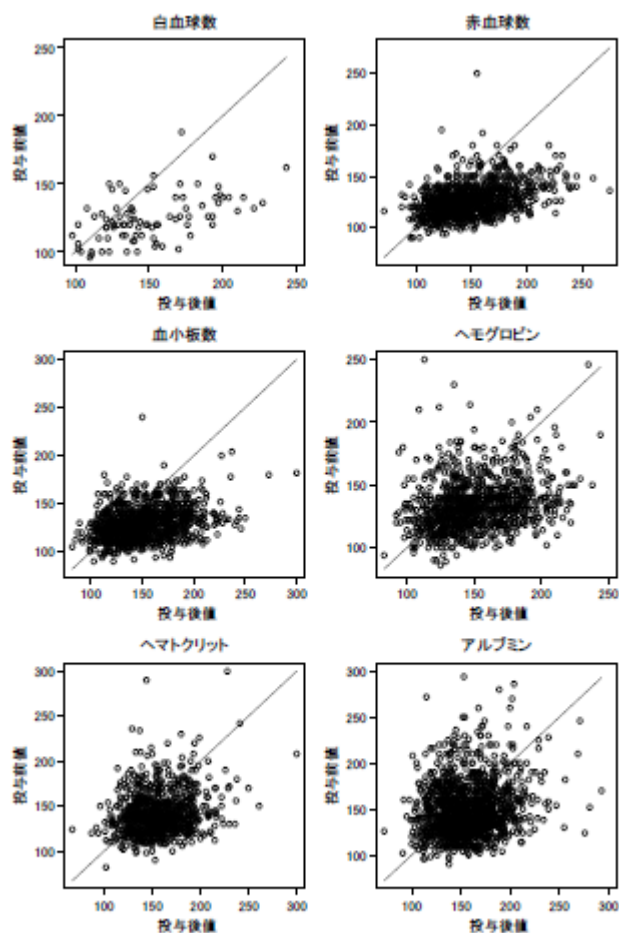


図14.3.1 臨床検査値の投与前後比較

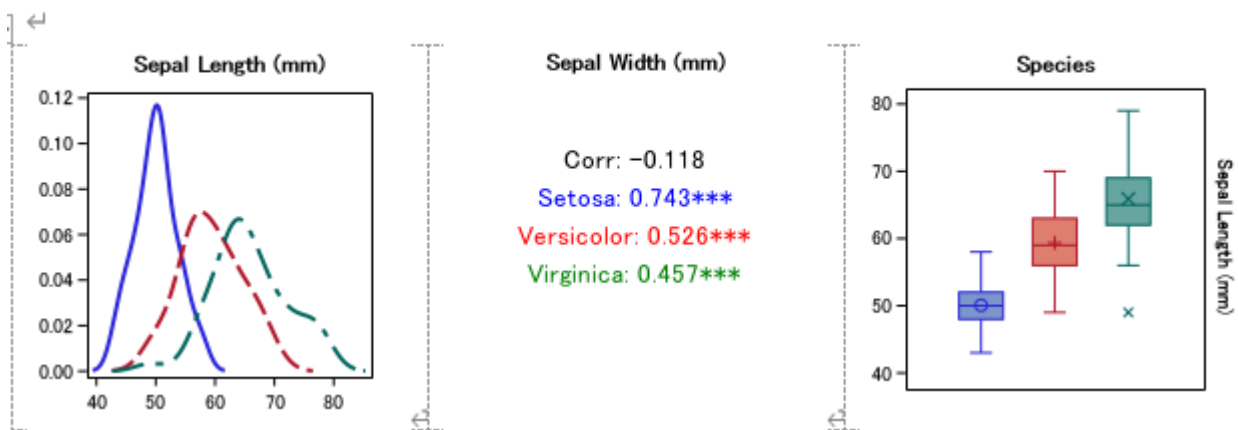
10. RTF ファイルへの文字列の書き込み

散布図行列⁵⁾の上に表題の文字列を書き込むこととする。「Figure1. Scatter Plot Matrix」を Arial 12pt で、図表ではなく本文領域に書き込んでいる。もちろん日本語での記入も可能である。

```
*--図表タイトル;
%let _title=Figure1. Scatter Plot Matrix;

[FileNew] br newline;
[Font("Arial")] br newline;
[FontSize(12)] br newline;
[Insert("&_title.")] br newline;*図表タイトル書き込み;
...
[FileSaveAs("&outpath.¥&_outfile.",4)] br newline;
```

Figure1. Scatter Plot Matrix



11. ページ番号挿入

DDE で RTF フッター部へ直接ページ番号を挿入する。ページ番号 / 総ページ数の表示とする。もちろんページ番号の字体の調整等は効く。

```
[FileOpen("&inpath.¥&_infile.")] br newline;
[ViewFooter] br newline;
[InsertPageField] br newline;
[Insert("/")] br newline;
[InsertNumberOfPages] br newline;
[EditSelectAll] br newline;
[Font("Times New Roman")] br newline;
[FontSize(10.5)] br newline;
[FileSaveAs("&outpath.¥&_outfile.",4)] br newline;
```

ページ番号挿入前

ページ番号挿入後

←



1./4←

12. すべてを承諾してコメントを除去

Word のすべての変更を反映し、変更の記録を停止し、すべてのコメントを削除する DDE プログラムです。

```
[FileOpen("&inpath.¥&_infile.")] br newline;
[AcceptAllChangesInDocAndStopTracking] br newline;*すべてを承諾;
[DeleteAllCommentsInDoc] br newline;*すべてのコメントを削除;
[FileSaveAs("&outpath.¥&_outfile.",4)] br newline;
```

コメント削除前

Name	Sex	Age	Height	Weight
アルフレッド	男子 女子	14	69.0	112.5
アリス	女子	13 2	56.5	84.0
バーバラ	女子	13	65.30	98.0



コメント削除後

Name	Sex	Age	Height	Weight
アルフレッド	女子	14	69.0	112.5
アリス	女子	12	56.5	84.0
バーバラ	女子	13	65.0	98.0

13. RTF の変更履歴のリストを作るプログラム

複雑な工程を DDE で踏むことはできないが、DDE は Word マクロのスイッチを押すことはできる。2つの RTF ファイルを Word 比較した変更履歴ファイルがあったとして、その変更履歴を別の Word ファイルにリストとして書き出す Word マクロ(Revision)を作ったとする。Word に事前登録しておいた Revison()マクロを DDE で叩くことにする。

```

Sub Revision()
Set wd1 = Application.ActiveDocument
cont = wd1.Revisions.Count
With Application.FileDialog(msoFileDialogFolderPicker)
MsgBox "リストファイル保存先を指定して下さい。"
If .Show = True Then
End If
End With
pt = Application.FileDialog(msoFileDialogFolderPicker).SelectedItems(1) & "¥"
sabun = cont
If cont > 10 Then
cont = 10
SaveName = pt & "_差分" & sabun & "個(Index=" & cont & ").doc"
Else
SaveName = pt & "_差分" & sabun & "個.doc"
End If
Set wd2 = Application.Documents.Add
Set objRange = wd2.Range()
wd2.Tables.Add objRange, cont + 1, 5
Set objTable = wd2.Tables(1)
objTable.Cell(1, 1).Range.Text = "Index"
objTable.Cell(1, 2).Range.Text = "Start"
objTable.Cell(1, 3).Range.Text = "Page"

```

```

objTable.Cell(1, 4).Range.Text = "Type"
objTable.Cell(1, 5).Range.Text = "Text"
For i = 1 To cont Step 1
  Index = wd1.Revisions(i).Index
  Start = wd1.Revisions(i).Range.Start
  Page = wd1.Revisions(i).Range.Information(3)
  typ = wd1.Revisions(i).Type
  Text = wd1.Revisions(i).Range.Text
  objTable.Cell(i + 1, 1).Range.Text = Index
  objTable.Cell(i + 1, 2).Range.Text = Start
  objTable.Cell(i + 1, 3).Range.Text = Page
  objTable.Cell(i + 1, 4).Range.Text = typ
  objTable.Cell(i + 1, 5).Range.Text = Text
Next
Application.ActiveDocument.SaveAs SaveName
End Sub

```

上記 Word マクロ Revision()は以下の方法で SAS から押すことができる。

```

[FileOpen("&inpath.¥&_infile.")] br newline;
[Revision()] br newline;*ワードマクロ適用;
[FileClose] br newline;

```

変更履歴

6 個の変更箇所

挿入された箇所: 3

削除された箇所: 3

移動された箇所: 0

書式が変更された箇所: 0

コメント: 0

Author 削除

男子

Author 挿入

女子

Author 削除

➔

Class.rtf

Name	Sex	Age	Height	Weight
アルフレッド	男子 女子	14	69.0	112.5
アリス	女子	13 2	56.5	84.0
バーバラ	女子	13	65.30	98.0
キャロル	女子	14	62.8	102.5
ヘンリー	男子	14	63.5	102.5
ジェームズ	男子	12	57.3	83.0
ジェーン	女子	12	59.8	84.5

Page	Type	Text
1	2	男子
1	1	女子
1	2	13
1	1	12
1	2	3
1	1	0

14. まとめ

コマンドマクロが全て有効とは限らないので、トライ&エラーではあるが、RTF に適用できる DDE を探していくことは可能であった。ODS RTF 等を使って SAS から出力したファイルに、DDE で体裁

変更の補填をするのに有益であると考えられた。DDE で不可能なことであっても Word マクロで実現可能なことであれば、DDE でその Word マクロを操作することもできた。

参考文献

- 1) Jay Zhou (2009). Importing Data from Microsoft Word into SAS, PharmaSUG 2009, Paper CC18
- 2) 森岡 裕 (2013). ライブラリ参照と名前定義を利用して EXCEL ファイルへの柔軟な入出力を実現する方法と応用例の提案 —解析結果のレポートニングからセルオートマトンまで—, SAS ユーザー総会 2013 論文集
- 3) 吉崎 正浩 (2010). RTF を用いた総括報告書の解析結果作成業務の効率化, SAS ユーザー総会 2010 論文集
- 4) [SAS Help Center: DDE Examples](#)
- 5) 徳田 芳稀 (2023). SAS による散布図行列の実装, SAS ユーザー総会 2023 論文集

SQLの一貫性制約

○関根 暁史
(藤本製薬株式会社)

Integrity Constraints in SQL

Satoshi Sekine
Fujimoto Pharmaceutical Corp.

要旨

proc SQL を使用することで、SAS データセットに何らかの規則(制約条件)を組み込むことができる。その制約条件は、データマネジメントにおけるロジカルチェック的機能にもなるであろう。

キーワード： proc SQL, 一貫性制約(整合性制約)

1. はじめに

データベースに入力できる値は、何らかの規則や基準に従ったものである必要がある。そのような制約条件をテーブルの仕様として組み込める仕組みを一貫性制約(整合性制約)と言う。SAS にももちろん proc SQL として一貫性制約の機能が備わっている。SQL の一貫性制約には、非ナル制約・一意性制約・検査制約・主キー制約・外部キー制約(参照制約とも言う)の 5 種類あり、これらを組み合わせることにより、データマネジメントにおけるロジカルチェックの代わりになるであろう。本論文ではこの 5 種類の一貫性制約の詳細を紹介する。

2. SAS における一貫性制約の使い方

2-1. 空のデータセットに一貫性制約を設ける

一貫性制約の付いた空のデータセットを用意したとする。その空のデータセットに proc SQL で何らかの値を入力したとき、制約条件に違反していたならばその値は発生できないこととなる(ERROR となる)。proc SQL による値の発生だけではなく、proc APPEND による行の積み上げでも同様である。制約は、空のデータセットを定義するときの変数の後ろに付与することができる。

```
proc sql;  
  create table AE(label=' Adverse Events')  
(AESTDTC Char(10) label=' Start Date/Time of Adverse Event' not null);
```

```
quit;
```

あるいは変数を宣言し終わったあとに、最後にまとめて一貫性制約をかけることもできる。以下では制約に `_NM0001_` という名前を付けている。

```
proc sql;
  create table AE(label='Adverse Events')
  (AESTDTC Char(10) label='Start Date/Time of Adverse Event',
  constraint _NM0001_ not null(AESTDTC) );
quit;
```

制約の情報を見たいときは `DESCRIBE` を用いればよい。

```
proc sql;
  describe table AE;
quit;
```

```
----- 一貫性制約の昇順リスト -----
#   一貫性制約   タイプ   変数
-----
1   NM0001      Not Null  AESTDTC
```

2-2. 既存のデータセットに一貫性制約を設ける

既存のデータセットに一貫性制約を設けようとしたとき、データセットに既に何らかの違反があったならば、一貫性制約はかからないこととなる(`ERROR` が出力される)。

以下では既存のデータセットに一貫性制約を追加している。

```
proc sql;
  alter table AE add constraint _NM0001_ not null(AESTDTC);
quit;
```

`proc SQL` でなくとも `proc DATASETS` で制約をかけることはできる。

```
proc datasets lib=work memtype=data nolist;
  modify AE;
  ic create _NM0001_=not null(AESTDTC);
quit;
```

2-3. 一貫性制約の解除

以下のような構文で制約を削除することができる。

なお `XPT` ファイルにはそもそも一貫性制約は設けられないのでご注意ください。

```
proc sql;
  alter table AE drop constraint _NM0001_;
quit;
```

3. 非ナル制約

非ナル制約とは、指定した列に必ず値を設定しなければならない制約である。例えば有害事象発現日に非ナル制約をかけてみる。エラーメッセージも入力してみる。

```
proc sql;
  create table AE(label=' Adverse Events')
  (AESTDTC Char(10) label=' Start Date/Time of Adverse Event' not null message="非ナル
  制約に抵触");
quit;
```

上記の空データセットからナル値を持つレコードを発生させようとする、制約に適合しないため ERROR が出てレコードの追加はなされない。すなわち AE を有害事象発現日が欠測のデータは入力できないデータセットとすることができる。

```
proc sql;
  insert into AE(AESTDTC) values ("");
quit;
```

```
6  proc sql;
7  insert into AE(AESTDTC) values ("");
ERROR: 非ナル制約に抵触 データセットWORK.AEへの追加/更新に失敗しました。
NOTE: データをVALUES句1からデータセットに追加するときに、この挿入は失敗।
NOTE: 一貫性のある状態にテーブルをリストアするために、前述のエラーが発生
```

VIEWTABLE: Work.Ae (Adverse Events)		
	AEREL	AESTDTC
1	NOT RELATED	2004-01-06
2	NOT RELATED	2004-01-27
3	NOT RELATED	×

NULL 値の入力を
禁止できる

4. 一意性制約

一意性制約とは、指定した列(又は列の組)の中にデータの重複を許さない制約である。以下ではバイタルサインのビジットの重複を禁じている。

```
proc sql;
  create table VS(label=' Vital Signs')
  (USUBJID Char(10) label=' Unique Subject Identifier',
  VSTEST Char(10) label=' Vital Signs Test Name',
  VISIT Char(10) label=' Visit Name',
  constraint _UN0001_ unique(USUBJID, VSTEST, VISIT) );
quit;
```

VIEWTABLE: Work.Vs (Vital Signs)		
	VSTEST	VISIT
1	Pulse Rate	BASELINE
2	Pulse Rate	WEEK 2
3	Pulse Rate	WEEK 2

重複レコードの入力を
禁止できる

5. 検査制約

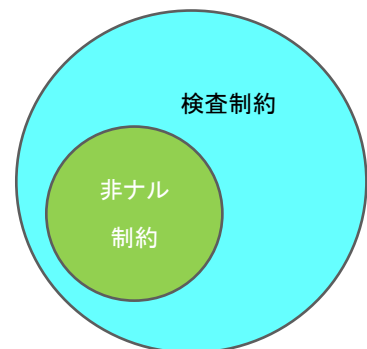
検査制約とは、条件を満たさないデータの入力を禁止する制約である。例えば、数値が特定の範囲内にあることを保証するために用いる。以下の例では、臨床検査値で0以下となる値の入力を禁止している。

```
proc sql;
  create table LB(label='Laboratory Test Results')
(LBCAT Char(10) label='Category for Lab Test',
  LBORRES Char(10) label='Result or Finding in Original Units',
  constraint _CK0001_ check((LBCAT="URINALYSIS") or (LBCAT<>"URINALYSIS" &
input(LBORRES, best.)>0) ) );
quit;
```

VIEWTABLE: Work.Lb (Laboratory Tests Results)		
	LBCAT	LBORRES
1	HEMATOLOGY	36.2
2	HEMATOLOGY	32.9
3	HEMATOLOGY	0.0

0以下の値の入力を
禁止

検査制約の条件式は何でも定義ができる。「ナル値でない」という条件であってもよいため、非ナル制約は検査制約に包含された方法と言える。



次に ADLB.AVALC に不等号("<"又は">")を含む値の入力を禁じてみる。

SQL では LIKE 述語やワイルドカードが使えるので条件指定が便利である。

```
proc sql;
  create table ADLB(label='Laboratory Test Results Analysis')
(AVAL Num(8) label='Analysis Value',
  AVALC Char(10) label='Analysis Value (C)',
  constraint _CK0002_ check((AVALC not like '%<') & (AVALC not like '%>')) );
quit;
```

VIEWTABLE: Work.AdIb (Laboratory Test Results Analysis)		
	AVAL	AVALC
1	0.15	<0.15
2	80	80>

不等号を含む語の入力を
禁止

6. 主キー制約

主キー制約とは、指定した列(又は列の組)にデータの重複およびナル値を許さない制約である。つまり主キー制約 = 非ナル制約 + 一意性制約とも言える。以下では症例 ID, 検査項目名およびビジットの組み合わせを主キーとして設定している。

```
proc sql;
  create table EG(label='ECG Test Results')
  (USUBJID Char(10) label='Unique Subject Identifier',
  EGTESTCD Char(8) label='ECG Test or Examination Short Name',
  VISIT Char(10) label='Visit Name',
  constraint _PK0001_ primary key (USUBJID, EGTESTCD, VISIT) );
quit;
```

VIEWTABLE: Work.Eg (ECG Test Results)			
	USUBJID	EGTESTCD	VISIT
1	CDISC01	INTP	SCREEN
2	CDISC01	INTP	WEEK 24
3	CDISC01	INTP	WEEK 24
4	CDISC01	INTP	

重複レコードおよび NULL 値
の入力を禁止



7. 外部キー制約

子テーブルの列に外部キー制約を設定すると、参照先の親テーブルの指定された列に記録された値しか書き込むことができないようになる。参照先の列に対しても、子テーブルから参照されている値は削除することができない制約が設けられる。外部キー制約を結ぶ際に、親テーブルには主キー制約がかけられていなければならない。

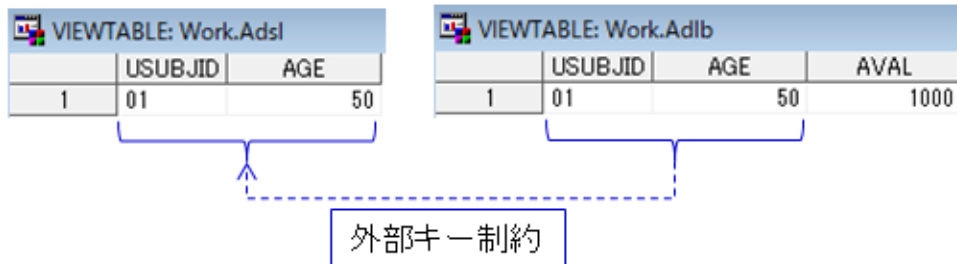
まずは ADSL(親テーブル)の USUBJID と AGE の組に主キー制約をかけたとする。

```
proc sql;
  create table ADSL
  (USUBJID Char(2), AGE Num, constraint _PK0001_ primary key (USUBJID, AGE) );
  insert into ADSL (USUBJID, AGE) values ("01", 50);
```

```
quit;
```

次に ADLB(子テーブル)の USUBJID,AGE から ADSL(親テーブル)の USUBJID,AGE に対して外部キー制約をかける。

```
proc sql;  
  create table ADLB  
  (USUBJID Char (2), AGE Num, AVAL Num, constraint _FK0001_ foreign key (USUBJID, AGE)  
  references ADSL);  
  insert into ADLB (USUBJID, AGE, AVAL) values ("01", 50, 1000);  
quit;
```



7-1. ON UPDATE RESTRICT

上記の条件で ADSL の年齢を 50 から 52 歳に更新しようとする、ERROR が出て更新することができない。ADSL と ADLB は関連付けられているので、どちらか一方の情報のみ更新することができないのである。つまりオプションを書かなくともデフォルトで ON UPDATE RESTRICT のオプションが付いている。

7-2. ON DELETE RESTRICT

同様に ADSL の年齢を 50 からナル値に変更しようとする、ERROR が出て更新することができない。何も書かなくともデフォルトで ON DELETE RESTRICT のオプションが付けられているからである。

7-3. ON UPDATE CASCADE

外部キー制約を結ぶ際に、ON UPDATE CASCADE のオプションを付けたとする。

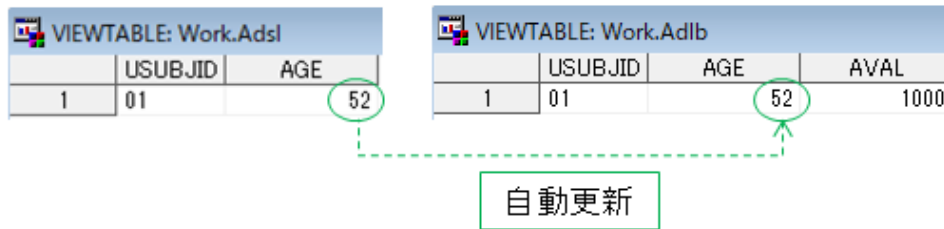
```
proc sql;  
  create table ADLB  
  (USUBJID Char (2), AGE Num, AVAL Num, constraint _FK0001_ foreign key (USUBJID, AGE)  
  references ADSL ON UPDATE CASCADE);  
  insert into ADLB (USUBJID, AGE, AVAL) values ("01", 50, 1000);  
quit;
```

この条件で親テーブル ADSL の年齢を 50 から 52 歳に更新してみる。

```
proc sql undopolicy=none;  
  update ADSL set AGE=52 where USUBJID="01";
```

```
quit;
```

すると子テーブル ADLB の年齢もつられて 52 歳に自動更新される。



ON UPDATE CASCADE とは、親テーブルの情報が子テーブルに自動的に伝播されるオプションである。

7-4. ON DELETE CASCADE

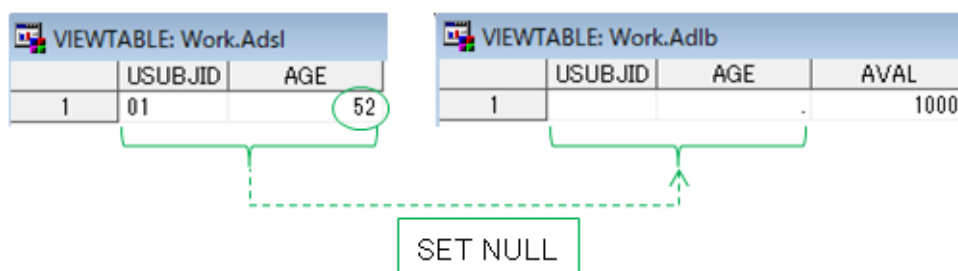
ON DELETE CASCADE の機能は SAS には実装されていない。

7-5. ON UPDATE SET NULL

外部キー制約を結ぶ際に、ON UPDATE SET NULL 句を付けたとする。

```
proc sql;  
  create table ADLB  
  (USUBJID Char (2), AGE Num, AVAL Num, constraint _FK0001_ foreign key (USUBJID, AGE)  
  references ADSL ON UPDATE SET NULL);  
  insert into ADLB (USUBJID, AGE, AVAL) values ("01", 50, 1000);  
quit;
```

上記の条件で ADSL の年齢を 50 から 52 歳に更新しようとする、親テーブルの更新された ID に対応する子テーブルの ID にナル値がセットされる。



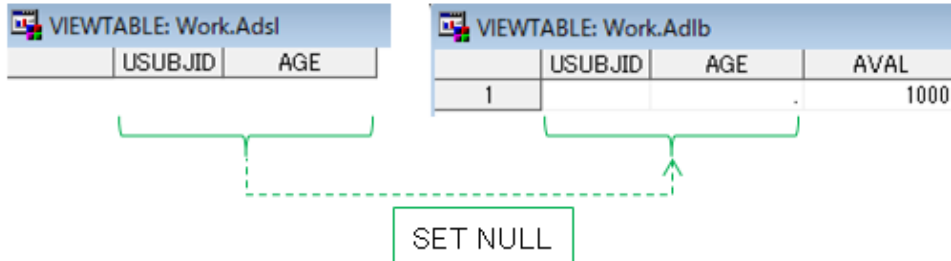
7-6. ON DELETE SET NULL

外部キー制約を結ぶ際に、ON DELETE SET NULL 句を付けたとする。

```
proc sql;  
  create table ADLB  
  (USUBJID Char (2), AGE Num, AVAL Num, constraint _FK0001_ foreign key (USUBJID, AGE)  
  references ADSL ON UPDATE SET NULL ON DELETE SET NULL);
```

```
insert into ADLB(USUBJID, AGE, AVAL) values ("01", 50, 1000);
quit;
```

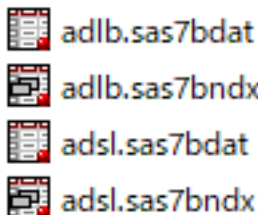
上記の条件で ADL の USUBJID="01" の行を削除しようとする、親テーブルの削除されたレコードに対応する子テーブルの ID にナル値がセットされる。



なお ON UPDATE SET NULL と ON DELETE SET NULL は共存できる。

8. sas7bndx ファイルについて

永久 SAS データセットについて、一意性制約、主キー制約又は外部キー制約を設けようとする、拡張子.sas7bndx のファイル(インデックスファイル)が同時に出来上がる。このインデックスファイルを誤って処分してしまったとしても、describe table を実行すれば復活するようだ。上記の制約をかけると、create index と同様の処理がかかるように思われる。



9. まとめ

5つの一貫性制約を駆使することにより、データベース内のデータ品質を維持し、データエラーや不整合性を減少させ、データベースの信頼性と保守性を向上させると考えられた。特に検査制約は、データマネジメントにおけるロジカルチェックの代替となり得る機能であった。データベースの設計時にこれらの制約を適切に使用することは、データベースプランニングとデザインの鍵となると思われる。

参考文献

- 1) SAS 9.4 SQL プロシジャユーザーガイド, 第4版, SAS ドキュメント 2023/04/25
- 2) 瀬戸美月 (2023). 徹底攻略データベーススペシャリスト教科書 令和5年度, インプレス
- 3) 三好康之 (2023). うかる! データベーススペシャリスト情報処理技術者試験学習書 2023年版, 翔泳社

Cochran-Mantel-Haenszel検定の概要とSASでの実装方法

○石本りさ

(イーピーエス株式会社)

Overview of the Cochran-Mantel-Haenszel Test and Its Implementation in SAS

Risa Ishimoto

EPS Corporation

要旨

Cochran-Mantel-Haenszel (CMH) 検定は複数の独立した分割表の層別解析を行うために多くの研究で広く用いられているが、初学者に向けて CMH 検定の解説を行っている文献は少ない。そのため、本稿では初学者が初めて CMH 検定を行う際の理解を助けることを目的とし、CMH 統計量の導出式について記述し、SAS の FREQ Procedure 及び LOGISTIC Procedure を用いた CMH 統計量の算出方法について紹介する。

キーワード : Cochran-Mantel-Haenszel 検定, proc freq, proc logistic

1. 緒言

臨床試験において収集されるデータにはカテゴリカルデータも多く、そのデータを解析する機会も多くある。カテゴリカルデータの解析において基本的な手法の1つには Pearson のカイ二乗検定があり、この検定においてクロス集計表における行変数及び列変数の独立性を調べることができる。しかし、Pearson のカイ二乗検定単独では交絡因子による影響を考慮した評価をすることはできない。CMH 検定では一つのクロス集計表を交絡因子で層別化することにより複数のクロス集計表を作成し、交絡因子の影響を調整した上で行変数及び列変数の連関性を調べることができる。広く用いられている検定手法の一つであるものの、初学者向けに解説を行っている文献は少ない。

そのため、本稿では初学者が初めて CMH 検定を行う際の理解を助けることを目的とし、CMH 統計量の導出式について記述し、SAS の FREQ Procedure 及び LOGISTIC Procedure を用いた CMH 統計量の算出方法について紹介する。さらに CMH 統計量の導出式への理解を深めるために、データステップのみを用いた CMH 統計量の導出についても解説する。

2. CMH 統計量の導出方法

前述したとおり、CMH 検定では一つのクロス集計表を交絡因子で層別化することにより複数のクロス集計表を作成し、層別解析を行う。イメージが捉えやすくなるよう、層化変数に性別および年齢を設定し、CMH 検定を行った際の概略図を Fig2.1 に示した。

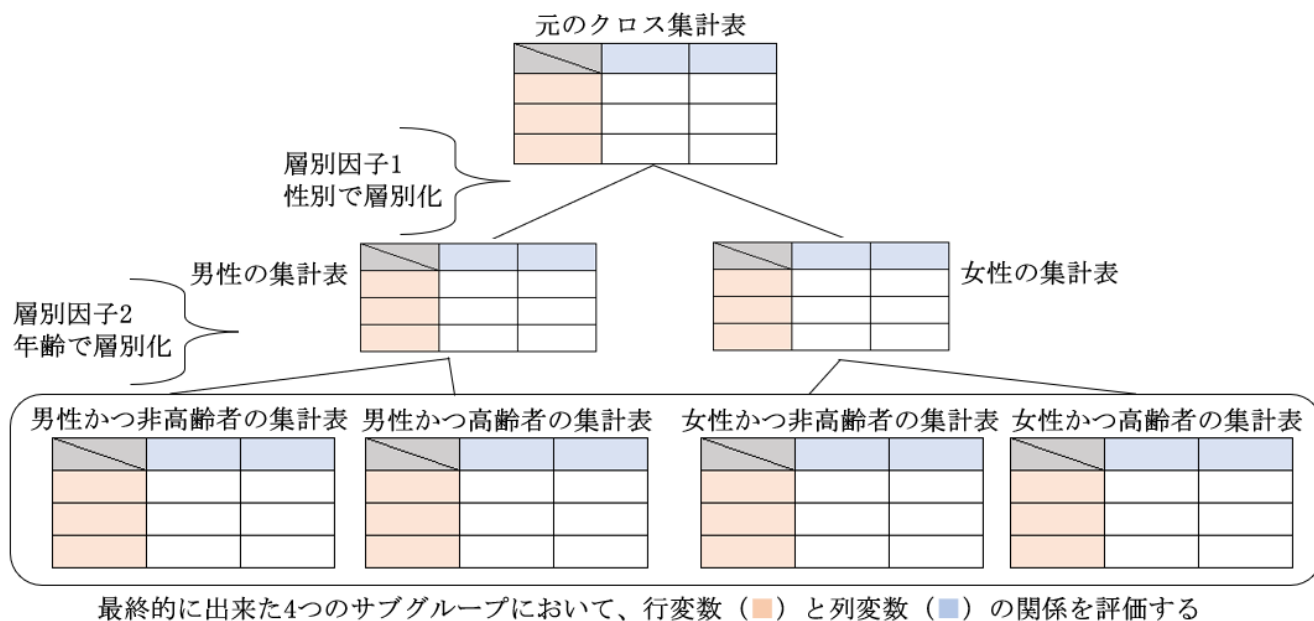


Fig 2.1 性別（2層）・年齢カテゴリ（2層）で CMH 検定を行った際の概略図

今回は 2×2 の分割表を想定し、Table 2.1 を用いて CMH 統計量の導出について最初に解説する。Table 2.1 は一つの分割表に対し層が k 個あると仮定した際の、 i 層目の分割表を示している。

Table 2.1 クロス集計表のサンプル

		Response		Total
		Better	Same	
Treatment	Active	a_i	b_i	m_{1i}
	Placebo	c_i	d_i	m_{2i}
Total		n_{1i}	n_{2i}	N_i

まず、CMH 検定ではオッズ比は全ての層で等質であるという仮定を置く。また、帰無仮説は「共通オッズ比は 1 である」とし、対立仮説は「共通オッズ比は 1 ではない」とする。オッズ比はある群における事象の起こりやすさに対する別の群における事象の起こりやすさであるため、オッズ比が 1 であれば、群と事象の起こりやすさに関連はないと判断することができる。つまり、CMH 検定における帰無仮説は「いずれの層においても行変数及び列変数の間に連関性はない」と言い換えることができる。

この帰無仮説の下で Table 2.1 における a_i は超幾何分布に従うことが知られている。超幾何分布とは、二種

類の性質で構成される集団から、無作為かつ非復元抽出を行ったときに、特定の性質のものがいくつあるかという確率を表す離散型確率分布である。例えば、母集団のサイズを N 、その中で特定の性質の要素数を M 、無作為に抽出される標本サイズを n 、標本中における特定の性質の要素数を a とする。その際、確率 $P(X = a)$ は以下の式で表される。

$$P(X = a) = \frac{{}^M C_k \times {}^{N-M} C_{n-a}}{{}^N C_n}$$

$$\max\{0, n - (N - M)\} \leq a \leq \min\{n, M\}$$

また、この確率変数 X における期待値と分散は下記になる。

$$E(X) = n \times \frac{M}{N} = \frac{nM}{N}$$

$$V(X) = n \times \frac{M(N - M)}{N^2} \times \frac{N - n}{N - 1} = \frac{nM(N - M)(N - n)}{N^2(N - 1)}$$

そのため、Table 2.1 における a_i の期待値と分散は下式にて表すことができる。

$$E(a_i) = \frac{n_{1i}m_{1i}}{N_i}$$

$$V(a_i) = \frac{n_{1i}m_{1i}n_{2i}m_{2i}}{N_i^2(N_i - 1)}$$

CMH 検定統計量は Z 統計量を二乗した値、つまりカイ二乗統計量 χ^2_{MH} で表される。すべての周辺度数を固定した条件のもとで、上記の期待値と分散を用いると算出式は以下となる。

$$\chi^2_{MH} = Z^2 = \frac{[\sum_{i=1}^k (a_i - E_i)]^2}{\sum_{i=1}^k V_i} = \frac{[\sum_{i=1}^k (a_i - n_{1i}m_{1i}/N_i)]^2}{\sum_{i=1}^k \frac{n_{1i}m_{1i}n_{2i}m_{2i}}{N_i^2(N_i - 1)}}$$

こうして導いた上記の検定統計量が自由度 1 のカイ二乗分布に従うことを利用して、CMH 検定では仮説検定を行う。導出式の複雑さから難しく見えるが、カイ二乗統計量の算出式に超幾何分布より導いた期待値と分散を代入したのみと考えるとシンプルである。また、資料によっては次式として表されていることもあるが、分子の表し方が異なるだけで内容としては上記式と同一である。

$$\chi^2_{MH} = \frac{[\sum_{i=1}^k (a_i d_i - c_i b_i)]^2}{\sum_{i=1}^k \frac{n_{1i}m_{1i}n_{2i}m_{2i}}{N_i^2(N_i - 1)}}$$

また、サンプルサイズが小さい場合、過度に有意な結果が生じるのを防ぐためにイエーツの補正を用いて

連続修正を行う場合がある。連続修正を行った場合の CMH 検定統計量の算出式は、分子から 0.5 を引いた下式となる。この補正をかけると CMH 検定統計量は小さくなるため、算出される p 値は大きくなり、統計的な有意差は生じにくくなる。

$$\chi^2_{MH} = Z^2 = \frac{[\sum_{i=1}^k (a_i - E_i)]^2}{\sum_{i=1}^k V_i} = \frac{[\sum_{i=1}^k (a_i - \frac{n_{1i}m_{1i}}{N_i} - 0.5)]^2}{\sum_{i=1}^k \frac{n_{1i}m_{1i}n_{2i}m_{2i}}{N_i^2(N_i - 1)}}$$

3. SAS による CMH 検定の実行

3.1 FREQ Procedure を用いた CMH 検定の実行方法

最初に FREQ Procedure を用いた CMH 検定の実行方法について解説する。サンプルデータとして、Table 3.1.1 のデータを用いて解説を行う。今回のサンプルデータは FREQ Procedure における CMH 検定の実行方法を解説した SAS Procedure Guide より引用した。^[1]

こちらのサンプルデータによる解析は、Treatment の Response に対する影響を調べることを目的としており、Gender で層別を行っている。

```
/* CODE 3.1 */
data Migraine;
input Gender $ Treatment $ Response $ Count @@;
datalines;
female Active Better 16  female Active Same 11
female Placebo Better 5   female Placebo Same 20
male Active Better 12   male Active Same 16
male Placebo Better 7   male Placebo Same 19
;
run;
```

Table 3.1.1 サンプルデータ

	Gender	Treatment	Response	Count
1	female	Active	Better	16
2	female	Active	Same	11
3	female	Placebo	Better	5
4	female	Placebo	Same	20
5	male	Active	Better	12
6	male	Active	Same	16
7	male	Placebo	Better	7
8	male	Placebo	Same	19

```

/* CODE 3.2 */
ods output cmh = cmh;
Proc freq data = <source dataset>;
    Tables gender * treatment * response / relrisk plots (only) = oddsratioplot (stats) cmh;
    Weight count;
run;

```

上記のように FREQ Procedure において、Tables ステートメント内に行変数、列変数及び層化変数を指定した上で CMH オプションを記述し、Weight ステートメントにて分析変数の数値の重みを設定することで CMH 検定を実行することができる。この時、左端変数が層化を指定し、二つの右端変数が表の行と列を指定することとなるため、変数を設定する際にはその点を注意する必要がある。

上記コードを実行した際、アウトプットに出力される結果が Table 3.1.1～Table 3.1.4 及び Fig 3.1.1 である。まず RELRISK オプションを設定したことにより、Table 3.1.1 内において、層別の分割表におけるオッズ比、相対リスクの値及び Wald 法で求められた 95%信頼限界が計算されている。CMH オプションにより共通オッズ比及び相対リスク推定値も計算されており、それらの結果を示しているのが Table 3.1.2 である。

また PLOTS オプションを設定しているため、gender の各水準及び全体のオッズ比とその 95%信頼限界がプロットされたグラフも出力されており、Fig 3.1.1 がその図を示している。このとき、“plots=oddsratioplot”ではなく“plots(only)=oddsratioplot”と記述することで、オッズ比プロットのみを特定してプロットを出力している。もし前者のコードで実行した場合には、オッズ比プロットに加えて相対リスクプロットも出力される。さらに、“oddsratioplot(stats)”において stats と指定していることで、プロット図内に統計量を表している。プロットをアウトプットに出力するためには、予め ods graphics オプションを on にする必要がある。

Table 3.1.1 層別変数毎のオッズ比及び相対リスク

表 1 : Treatment * Response の統計量
層別変数 : Gender=female

オッズ比と相対リスク			
統計量	値	95% 信頼限界	
オッズ比	5.8182	1.6755	20.2034
相対リスク (列 1)	2.9630	1.2740	6.8913
相対リスク (列 2)	0.5093	0.3103	0.8357

表 2 : Treatment * Response の統計量
層別変数 : Gender=male

オッズ比と相対リスク			
統計量	値	95% 信頼限界	
オッズ比	2.0357	0.6478	6.3975
相対リスク (列 1)	1.5918	0.7413	3.4180
相対リスク (列 2)	0.7820	0.5259	1.1626

Table 3.1.2 共通オッズ比及び相対リスクの推定値

共通オッズ比と相対リスク				
統計量	手法	値	95% 信頼限界	
オッズ比	Mantel-Haenszel	3.3132	1.4456	7.5934
	ロジット	3.2941	1.4182	7.6515
相対リスク (列 1)	Mantel-Haenszel	2.1636	1.2336	3.7948
	ロジット	2.1059	1.1951	3.7108
相対リスク (列 2)	Mantel-Haenszel	0.6420	0.4705	0.8761
	ロジット	0.6613	0.4852	0.9013

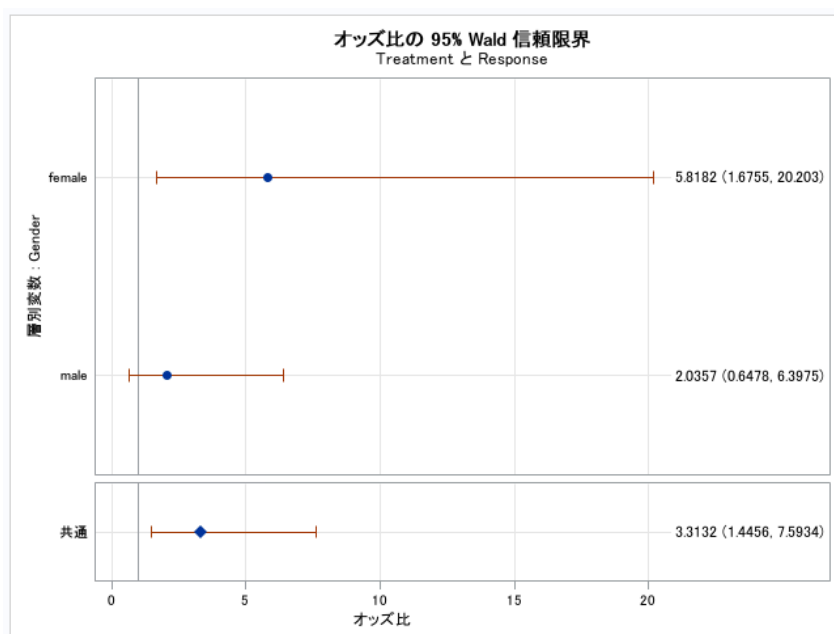


Fig 3.1.1 層別変数毎のオッズ比及び共通オッズ比のプロット図

Table 3.1.3 Cochran-Mantel-Haenszel 統計量の出力画面

Cochran-Mantel-Haenszel 統計量 (テーブルスコアに基づ)				
統計量	対立仮説	自由度	値	p 値
1	相関統計量	1	8.3052	0.0040
2	ANOVA 統計量	1	8.3052	0.0040
3	一般関連統計量	1	8.3052	0.0040

また、Table3.1.3 に示しているのが CMH 統計量の出力画面である。CMH 統計量においては、相関統計量、ANOVA 統計量及び一般関連統計量の三つの結果が表示される。相関統計量は行変数及び列変数が順序尺度である場合の CMH 統計量、ANOVA 統計量は列変数が順序尺度である場合の CMH 統計量、そして一般関連統計量は行変数及び列変数が順序尺度でない場合に用いられる CMH 統計量である。これらの統計量についていずれを使用するかについては、解析する分割表内の行や列が順序を持つかどうかで異なるが、今回のような 2×2 分割表の場合において順序関係は問題にならず、これらの三つの統計量は全て同じ値になる。¹¹⁾

Table 3.1.4 オッズ比等質性に対する Breslow-Day 検定の出力画面

オッズ比等質性に対する Breslow-Day 検定	
カイ 2 乗値	1.4929
自由度	1
Pr > ChiSq	0.2218

ods output 出力画面に最下段に表示される結果 (Table 3.1.4) がオッズ比等質性に対する Breslow-Day 検定の結果である。CMH 検定では 2 章で解説した通りに、オッズ比は全ての層で等質である、という仮定を置いていた。この仮定に対する検定が Breslow-Day 検定であり、帰無仮説は「各層におけるオッズ比は等しい」というものである。今回の結果では p 値が 0.2218 であり、帰無仮説は棄却されないため、オッズ比は全ての層で等質ではないとはいえないと判断できる。

3.2 LOGISTIC Procedure を用いた正確な CMH 検定の実行方法

ロジスティック回帰分析では 2 値の目的変数における特定の事象が発生する確率を、説明変数を用いて構築されたロジスティック回帰モデルより予測する。カテゴリカルデータの分析に対して用いられ、2 値アウトカムの評価ができるという点は CMH 検定と同様である。このロジスティック回帰分析において、 k 個の独立した 2×2 表の層別解析における周辺度数を固定した場合のスコア検定の結果は、CMH 検定の結果と一致することが報告されている。^[2] これに関する理論的背景については十分なロジスティック回帰分析におけるスコア検定への知識を要するため、詳細に関しては関連する参考文献を参照されたい。^{[3][4]}

今回はまず実際に 3.1 節と同様のサンプルデータを用いて LOGISTIC Procedure でロジスティック回帰分析を行い、CMH 検定の結果と一致することを確認した。CODE 3.3 が該当する LOGISTIC Procedure の実行コードである。着目している "gender" という層別変数を strata ステートメントで指定しプロシジャを実行すると、スコア検定のカイ二乗統計量は CMH 統計量と同様の "8.3052" と一致していることがわかる。ただし、LOGISTIC Procedure で出力された共通オッズ比については FREQ Procedure で出力した共通オッズ比とは一致しないため、一致するのはカイ二乗統計量と p 値のみであるという点は注意が必要である。

```

/* CODE 3.3 */
Proc logistic data=Migraine;
  Class treatment / param=ref;
  Freq count;
  Model response(Event='Better')=treatment;
  Strata gender;
  Exact treatment;
run;

```

Table 3.2.1 ロジスティック回帰モデルにおけるスコア検定統計量及び CMH 検定統計量の一致

包括的帰無仮説: BETA=0 の検定				Cochran-Mantel-Haenszel 統計量 (テーブルスコアに基づく)				
検定	カイ 2 乗値	自由度	Pr > ChiSq	統計量	対立仮説	自由度	値	p 値
尤度比	8.4831	1	0.0036	1	相関統計量	1	8.3052	0.0040
スコア	8.3052	1	0.0040	2	ANOVA 統計量	1	8.3052	0.0040
Wald	7.9947	1	0.0047	3	一般関連統計量	1	8.3052	0.0040

また、一般的に CMH 検定を行うにあたっては各層における各セルの期待度数が 5 以上となるような大規模なサンプルサイズを必要とする。このような場合において、LOGISTIC Procedure を用いて正確な検定を実行すれば、少数のサンプルサイズでも層別解析を行うことが可能である。^[5]

具体的にはロジスティック回帰分析におけるスコア検定をデータへと適用し、結果として CMH 検定の結果を導けることを利用すればよい。CODE 3.3 に記載しているように exact ステートメントを使用することで、Table 3.2.2 に示した正確な検定結果が出力される。

Table 3.2.2 exact ステートメントを指定した際の出力結果

条件付き正確確率検定				
効果	検定	統計量	p 値	
			正確	中間
Treatment	スコア	8.3052	0.0051	0.0039
	確率	0.00242	0.0051	0.0039

4. Procedure を用いない CMH 統計量の導出方法

今回は CMH 統計量導出式の理解を深めるためにも、FREQ Procedure や LOGISTIC Procedure を用いずにデータステップの中で CMH 統計量を導出することも試みた。用いるサンプルデータは同様の Table 3.1.1 である。また、SORT Procedure 及び TRANSPOSE Procedure については導出までのデータセットの変形のために用いている。

```
/* CODE 4.1 */
data Migraine;
  input Gender $ Treatment $ Response $ Count @@;
  datalines;
  female Active Better 16   female Active Same 11
  female Placebo Better 5   female Placebo Same 20
  male Active Better 12    male Active Same 16
  male Placebo Better 7    male Placebo Same 19
  ;
run;

proc sort data = Migraine;
  by Gender;
run;

proc transpose data = Migraine out = wk1 delimiter = _;
  var Count;
  by Gender;
  id treatment response;
run;
```

Table 4.1 CODE 4.1 の出力結果

	Gender	_NAME_	Active_Better	Active_Same	Placebo_Better	Placebo_Same
1	female	Count	16	11	5	20
2	male	Count	12	16	7	19

まずサンプルデータを作成し、後における計算のし易さを考慮し TRANSPOSE Procedure で数値データを横並びにしたものが Table 4.1 のデータとなる。(CODE 4.1)

その次に各群における CMH 統計量導出式内の分母と分子の値を算出する。導出式への当てはめが理解しやすくなるよう、それぞれのデータを Table 2.1 における $a_i \sim d_i$ までの記号で表した。その後に行行列の合計値と、全データの合計値を算出している。ここまでの整えば各層における導出式内の分母と分子が計算でき、Denom 変数が分母、Numer 変数が分子を表している。(CODE 4.2)

Table 4.2 CODE 4.2 の出力結果

	Gender	a	b	c	d	N
1	female	18	11	5	20	52
2	male	12	16	7	19	54

m1	m2	n1	n2	Denom	Numer
27	25	21	31	3.1864557953	5.0961538462
28	26	19	35	3.1324895825	2.1481481481

```

/* CODE 4.2 */
data wk2;
  set wk1;

  a = Active_Better; b = Active_same;
  c = Placebo_Better; d = Placebo_same;

  N = sum(of Active_ : Placebo_);
  m1 = sum(of a b); m2 = sum(of c d);
  n1 = sum(of a c); n2 = sum(of b d);

  Denom = divide(n1*m1*n2*m2, (N**3-N**2));
  Numer = a - (divide(n1*m1, N));
run;

```

次に各群において算出した分母と分子の総和を導くために、再度 TRANSPOSE Procedure を用いてデータの形を整えていく。Total 変数は female 群と male 群の分母と分子の各総和を表している。(CODE 4.3)

```

/* CODE 4.3 */
proc transpose data = wk2 out = wk3;
  var Numer Denom;
  id Gender;
run;

data wk4;
  set wk3;
  Total = female + male;
run;

proc transpose data = wk4 out = wk5;
  var Total;
run;

```

```

/* CODE 4.4 */
data wk6;
  set wk5;

  CMH = divide(Numer**2, Denom);
  prob = 1 - probchi(CMH, 1);
run;

```

VIEWTABLE: Work.Wk6					
	NAME	Denom	Numer	CMH	prob
1	Total	6.3189453779	7.2443019943	8.3051693355	0.0039532396



Cochran-Mantel-Haenszel 統計量 (テーブルスコアに基づく)				
統計量	対立仮説	自由度	値	p 値
1	相関統計量	1	8.3052	0.0040
2	ANOVA 統計量	1	8.3052	0.0040
3	一般関連統計量	1	8.3052	0.0040

Fig 4.1 データステップで導出した CMH 統計量と FREQ Procedure における出力画面

最終的に二乗した Numer 変数を Denom 変数で割ることにより、CMH 検定統計量を導いた。(CODE 4.4) CMH 統計量は 8.30516...であり、p 値は 0.003952...と計算されている。この結果は Fig 4.1 で示した通りに FREQ Procedure で求めた結果と一致しており、データステップでの再現ができた。

5. まとめ

今回は初学者向けに CMH 検定統計量の導出方法及び SAS での実行方法について解説した。CMH 検定統計量の導出式は一見複雑ではあるが、超幾何分布における期待値と分散の導出方法及びカイ二乗統計量の導出方法を理解していれば、データステップでの再現も可能な比較的シンプルな検定統計量である。本稿の内容を通じて CMH 検定の理解が深まり、実際のデータ解析に役立てれば幸いである。

また今回は 2×2 の分割表での解説のみ行ったが、 $r \times c$ ($r > 2, c > 2$) の分割表においては Table 3.1.3 に示した統計量のうち、順序尺度の有無を考慮し、どれが適切かを検討する必要がある。今後の著者の課題として、多元分割表における CMH 検定についての理解を深めていきたい。

6. 参考文献

- [1] SAS Institute Inc., Cary, NC. Base SAS® 9.4 Procedures Guide: Statistical Procedures, Fourth Edition (2015)
- [2] Day, N.E. and Byar, D.P. Testing hypotheses in case-control studies-equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics*, 35, 620-630. (1979)
- [3] 丹後 俊郎, 山岡 和枝, 高木 晴良. 新版 ロジスティック回帰分析 -SAS を利用した統計解析の実際-. 朝倉書店. (2013)
- [4] David W. Hosmer, Jr, Stanley Lemeshow, Rodney X. Sturdivant (著), 宮岡 悦良, 早川 有, 川崎 洋平, 下川 朝有 (訳). データ解析のためのロジスティック回帰モデル. 共立出版. (2017)
- [5] SAS Support, "Usage Note 32711: Computing an exact CMH test", <https://support.sas.com/kb/32/711.html>, Accessed July 28, 2024
- [6] John M. Lachin (著), 宮岡 悦良, 遠藤 輝, 黒沢 健, 下川 朝有, 寒水 孝司 (訳). 医薬データのための統計解析 原著第2版. 共立出版. (2020)
- [7] 折村 奈美. データステップを用いた統計検定の再現. SAS ユーザー総会 2024. (2024)

オープンデータを活用した加古川市における 見守りカメラ導入による狭域での犯罪抑止効果の検証 —自転車盗の事例—

○川向 肇¹, 橋間 花音², 中根 滉大²

(¹兵庫県立大学 大学院 情報科学研究科, ²兵庫県立大学 社会情報科学部)

A Trial of Verification of Crime Deterrent Effects in a Narrow Area by Introducing Street Monitoring Cameras in
Kakogawa City Utilizing Open Data
An Case Study on Bicycle Thefts
Hajime Kawamukai, Kanon Hashima, Kodai Nakane
Graduate School of Information Science, School of Social Information Science, University of Hyogo

要旨

加古川市においては市民生活の安全性向上を目的として、見守りカメラの導入が進められてきた。本報告では、この見守りカメラの防犯効果についての自転車盗を対象とする実証的分析を行った結果を示す。具体的には、オープンデータとして加古川市が公開している見守りカメラ設置位置の座標情報と兵庫県が公開している 2018 年から 2023 年の期間における自転車盗のデータから事案発生町丁目名を利用して、標準地域メッシュを空間的集計単位とする空間的データベースを整備した。この空間的データベースに国勢調査および経済センサスの統計データを結合し、JMP17.3 を用いて自転車盗の発生を促進する要因と抑止する要因についての検証を行った結果を示す。本報告では、見守りカメラの防犯効果を定量的に評価し、地域での犯罪抑止効果の有効性を検証した事例を紹介する。

キーワード : Open Data, JMP, Spatial Analysis, Crime Deterrent Effects, Street Monitoring Cameras, Kakogawa City, Bicycle Thefts

研究の出発点

加古川市における見守りカメラの導入

兵庫県加古川市においては、「子育て世代からシニア世代まで誰もが健康で安心して暮らせるまち」、「生涯にわたって活躍できるまちの実現を目指す姿」を掲げ、2016（平成 28）年から ICT を活用した安心・安全なまちづくり推進事業を実施している。三和（2019）で紹介されているとおり、2017 年から

全市の各小学校区に約 50 台、全市で約 1,500 台の見守りカメラにより道路上の監視を可能とすると同時に、みまもりサービスと呼ばれる BLE タグの所有者が付近を通過した際にその所有者の通過を検知する検知機能を付与した見守りカメラを全国有数の設置数及び設置密度で導入している。

この見守りカメラの導入以降、加古川市（2023）で提示されているように、加古川市内の刑法犯認知件数の年次間比較した場合、見守りカメラ導入以前の平成 28 年（2016 年）12 月末段階での年間 2,926 件から令和 4 年 12 月末段階での年間事案数 1,699 件（42.0%減、なお同期間の兵庫県全体の減少率は 35.0%減）と、大幅に減少したことが示されており、市域全体での刑法犯認知件数に関しては一定の効果が確認されている。

しかしながら、加古川市（2022）において注意喚起がなされているように、見守りカメラ設置後も加古川市内における自転車盗は相当数発生しており、市民にとっての体感的な日常生活における防犯上の安心感に大きく影響する可能性がある課題の一つとなっている。

そこで、筆者らは総務省統計局、兵庫県、加古川市から提供されている多様なオープンデータを利用し、これらのデータについて、GIS アプリケーションを用い、総務省統計局（2024）で紹介されている標準地域メッシュ（第 3 次地域区画）と呼ばれる 1 km×1 km の四角形の空間領域に関するデータベースとして再編したうえで、JMP を利用して分析した結果を紹介する。

標準地域メッシュ利用のメリット

地域に関するデータ分析をする際の対象地域に関する統計データの編成方法として、①歴史的空間分割による領域（自治体境界や町丁目などの小地域など）を対象とする方法、②標準地域メッシュを対象とする方法、の 2 種類に大別される。

従来、地域分析において利用されてきた歴史的な地域区分に基づく町丁目などの領域の場合、地域住民や地域に関する土地勘がある人々にとっては特段マップ表示などを伴わずとも、空間的事象に関する直感的な理解を与えることが容易である反面、都市化していない地域では、町丁目の領域に山林・湖沼・農業用地などが含まれる場合があり、町丁目領域内に非都市的土地利用が含まれる場合、その町丁目の領域が広域となる傾向となる。この場合、見守りカメラの密度などを含む都市施設密度の計算を行うため、山林等の非都市利用領域部分の面積を除く作業を実施する必要がある。これに対し、標準地域メッシュを利用する場合、非都市的土地利用のみからなる標準地域メッシュの場合、領域内に居住する住民が存在しない場合には、標準地域メッシュを分析対象から容易に排除可能である。さらに、単一の市区町村を分析対象として標準地域メッシュを利用して分析する場合、①各標準地域メッシュの面積がほぼ等しい、②時系列的に空間分割が安定、というメリットを有していることから、当該標準地域メッシュの指標値そのものが、地理空間に関する狭義の正規化された指標値としても利用可能となるというメリットがある。これに加え、本報告で分析するような社会的課題の分析にあたっては、特定地名や地域名称と結びついた分析結果が独り歩きし、特定の事象が多発している地域として人々が広く認識する場合、特定地名の地域に対する好ましくない印象が一般に拡散しかねないという課題も回避可能となる。以上の理由から、本報告での空間的集計単位として標準地域メッシュを利用することとした。

なお、統計データとしては、標準地域メッシュをほぼ 4 等分した 2 分の 1 地域メッシュ（4 次メッシュ）、標準地域メッシュをほぼ 16 等分した 4 分の 1 地域メッシュ（5 次メッシュ）も利用されることがあるが、後述するように自転車盗の発生位置情報の空間的特定可能性の精度が低いことから、標準地域メッシュを分析単位とすることとした。

先行研究

日本国内における自転車盗に関する研究としては管見する限り松本（1982）によるものが最も早く、その後、福岡市内の天神地区と呼ばれる都市内の狭い領域を対象にした発生密度に関する研究である杉野・有馬（2014）があり、橋本・矢田他（2022）のように岡山市および倉敷市の両市域を対象に発生密度の高い地点をカーネル密度法による推計を行う手法について検討し、4分の1地域メッシュ（5次メッシュ）を分析空間の解像度としつつ公的統計データ以外にも独自に取得したデータについて4分の1地域メッシュ単位で集計したうえで標準化回帰モデルの推計を行い自転車盗の発生を予測する空間的特性によりモデル化した事例、杉浦・樋野ほか（2022）ではより詳細な土地利用モデルや建物構造などの物理的側面に関するデータを利用し、駅周辺での自転車盗の発生要因を負の二項回帰モデルで推計した研究などがある。

利用したデータと標準地域メッシュによる空間的データベースの構築法

本報告で紹介する研究を実施する際に利用したデータソースとしては、以下に示す加古川市が提供するオープンデータ、兵庫県が提供しているオープンデータ、総務省統計局が提供しているオープンデータの3種類のデータである。以下データの入手先別に各データの概略を示し、分析に際しての加工方法などを示す。

加古川市のオープンデータ

加古川市が提供するオープンデータとしては、加古川市が設置している見守りカメラの緯度経度による設置位置の座標データを用い、オープンソースのGISソフトウェアであるQGIS3.34.5によりGISで利用可能な形式の地点のポイントデータに変換し、標準地域メッシュ単位で見守りカメラの設置数を集計した。

兵庫県のオープンデータ

兵庫県がオープンデータカタログページにおいて、平成30年から令和4年までの5年分の兵庫県下で発生した自転車盗に関して、発生地点の町丁目名、発生の状況、施錠の有無、被害者の年齢や所属種別、発生場所（駐輪場、住宅内など）などの属性が付与されたCSV形式のファイルデータを利用した。このデータにおける発生地点の地点情報は〇〇町〇丁目レベルであり、正確な位置座標を特定できない可能性が高いという問題があった。この制約から、2分の1地域メッシュや4分の1地域メッシュなどのより狭い範囲の領域を集計単位とした場合、自転車盗事案数に関する集計バイアスが大きくなるため、標準地域メッシュによる集計を行うこととした。

平成30年から令和4年の各年次別に提供されている自転車盗事案についてExcel形式のデータをJMPで読み込み、テーブルの連結機能を利用し、5年分のデータを一括で扱えるテーブルとして編成を行った。さらにJMPで一変量の分析を用い、加古川市内での発生時案についてのサブセットとして抽出・保存し、発生地点の町丁目の文字事案発生地情報を含むCSVファイルとして保存した。

事案発生地の文字情報を含むCSVファイルを用意すれば、東京大学CSIS CSV Geocoding Serviceを利用して町丁目の代表点の緯度経度による座標情報付きCSVファイルを構成することが可能となる。そこで、文字情報として記録された自転車盗の発生地点情報からGeocoding Serviceを利用して緯度経度のデータに変換を行い、GISで利用可能な地点のポイントデータとして構築した。

島田・大山（2019）において、施錠の有無が自転車盗事案の被害阻止につながるとの指摘があることから、自転車盗の被害発生ポイントデータについて（1）施錠された自転車の盗難事案のポイントデータと、（2）無施錠での自転車の盗難事案についてのポイントデータという二種類のポイントデータを作成し、見守りカメラの集計と同様の方法で標準地域メッシュを利用した施錠状態別の発生事案数の集計を行った。

総務省統計局のオープンデータ

総務省統計局(2024)では、指定統計である国勢調査・経済センサス基礎調査及び活動調査・農林業センサスの小地域及び標準地域メッシュなどに対する統計データと標準地域メッシュなどに関する領域に関する空間的統計データが比較的容易に利用可能な形式でオープンデータとして提供されている。そこで、標準地域メッシュデータの領域データのうち、加古川市の領域を含む第1次地域区画のコードが5234（東経134度から135度 北緯34度40分から35度20分）に対応する第1次地域区画（1次メッシュ領域）で加古川市全域が包含されるため、この第1次地域区画に含まれる標準地域メッシュの領域データと同領域に対応する2020年の国勢調査の人口・世帯及び人口移動、就業状態等及び従業地・通学地、2016年の経済センサス活動調査の産業大分類別事業所数及び従業者数の統計データをダウンロードし、自転車盗数と見守りカメラ数が集計されたデータとこれらの統計データをJMPにより結合した。

国土交通省のオープンデータ

国土交通省（2024）では、鉄道・道路や防災関連、地価などの国土に関する多様な特性に関するデータが公開されているが、本報告では杉浦・樋野他（2022）に倣い、鉄道駅の乗降客数のデータをダウンロードし、QGISで鉄道駅の重心を求めたうえで、その重心座標値をもとにGISで利用可能なポイントデータとして構築した。さらにこのデータから標準地域メッシュ内に含まれる鉄道乗降客数をQGIを利用し集計したうえで、自転車盗事案数、見守りカメラ数、国勢調査、経済センサスの調査項目の集計値の含まれるデータベースとJMPにより結合して分析用データベースを構築した。

分析結果

以下では、JMPを利用して各種の分析を行った結果を紹介する。

加古川市内の自転車盗の発生事案数などに関する記述統計分析

まず、加古川市内全域での自転車盗に関する年次的変動の状況を確認した結果（図1）を見る限り、加古川市内における自転車盗数は、経年的に緩やかな減少傾向を示しており、特にCOVID-19の流行期においては政府や自治体からの外出自粛要請に伴い、企業等に対しては感染症予防のための在宅勤務が推奨され、学校等の教育機関については、オンライン授業をはじめとする自宅での学習の推進がなされたため、2020年及び2021年は自転車盗の事案数が相当減少している。なおCOVID-19対策が緩和され、外出規制が緩和された2022年度においては、兵庫県下の全刑法犯事案数と同様、増加傾向が確認された。

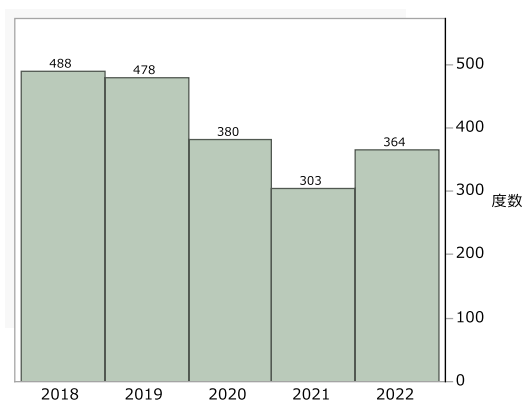


図1 加古川市内の年次別自転車盗事案数

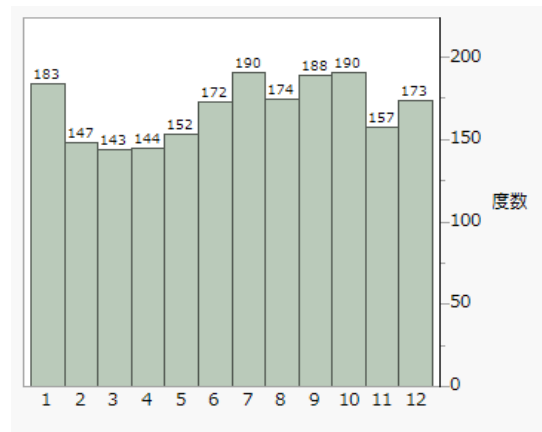


図2 加古川市内の月次別自転車盗事案数

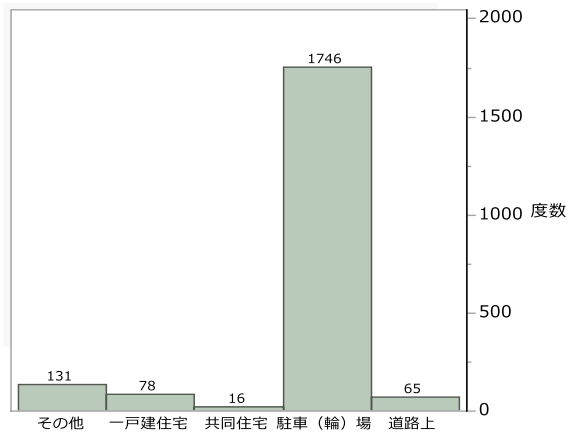


図3 発生場所別の自転車盗事案数

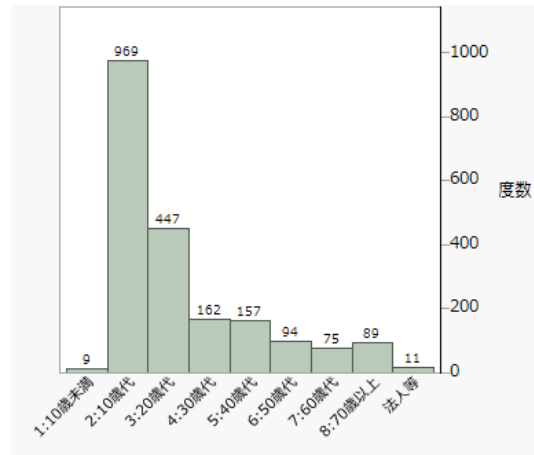


図4 被害者年齢階層別の自転車盗事案数

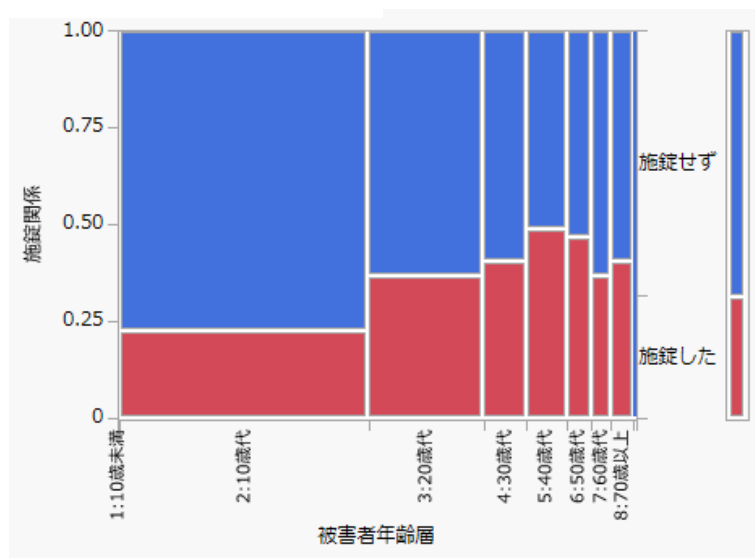


図5 自転車盗被害者年齢層の施錠の有無別に関するモザイク図

次に月次別の発生事案数（図2）を検証してみると、盛夏期にあたる8月を除く初夏から秋、年末年始の人々の移動数そのものが増加する時期、交通手段として自転車利用が快適となる秋の時期での自転車盗の増加傾向が確認された。

続いて、自転車盗の発生場所の種別（図3）について確認してみると、駐車場ないし駐輪場における事案数が圧倒的であり、路上の放置自転車が自転車の盗難被害の大半を占めているわけではない。加古川市内における自転車盗事案の発生場所としては圧倒的に駐車（輪）場が多い。図4に示す自転車盗の被害者属性に関する分析結果についてであるが、島田・大山（2019）でも示された通り、20歳代以下の若年層の被害者率が高く、特に加古川市内では10歳代の被害者の比率が約半数にあたる47.9%を占めている。

図5に示す被害者の年齢別施錠の有無別の分析結果からは、若年層での無施錠自転車の被害の比率が68.4%と高く、無施錠の自転車が盗難のターゲットとなっており、特に10歳代の自転車盗の被害者のうち77.4%が無施錠であった。

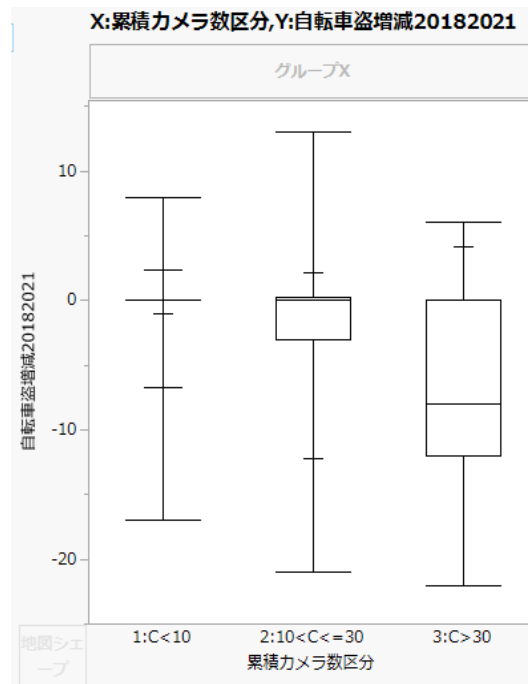


図6 標準地域メッシュ内の見守りカメラ数（2021年段階）
区分と自転車盗の変化に関する分位点箱ひげ図

これらの結果からは、小学校高学年の児童から中学校、高校の生徒への施錠の推奨、とりわけ、2ロックと呼ばれる駐輪時の二重施錠が自転車盗の事案の抑止に有効であるとされていることから、これらの自転車盗の予防対策についての集中的な一層の働きかけが有効であると考えられる。特に駐輪場等での被害が一定数存在することを考えると、駐輪場などでのこれらの若年層に特化したポスター等の掲示・駐輪時の二重施錠を呼びかけるキーホルダーや自転車用のハンドル部に貼付するステッカー等による施錠の意識づけなど、ワインバーグ（1990）などでも指摘されている危険回避行為を誘導するある種のナッジ理論などに基づく対策が有効かもしれない。

加古川市内の標準地域メッシュを集計単位とした自転車盗の発生事案数とその時間的推移

加古川市内における見守りカメラ設置直後であり加古川市内での自転車盗の事案数が最も多い2018年と最も少ない2021年における加古川市内の各標準地域メッシュ内（全標準地域メッシュ数157）における事案数の差を取り、見守りカメラの設置密度と自転車盗の事案数との関係を検証した（図6）。

この結果からは見守りカメラが少ない標準地域メッシュ領域（標準地域メッシュ内にカメラが10（メッシュ当たりの設置数の中央値）以下）では、自転車盗の発生事案数が限られることもあり、大きな変化は確認されない。

これに対し、見守りカメラの設置密度がやや高密度となる標準地域メッシュ領域（標準地域メッシュ内の設置カメラ数が11から20以下）では、自転車盗事案数の中央値による比較では設置直後の2018年から大きな差が確認されないという結果となったが、約半数のメッシュに関しては減少傾向にあるものの、一部自転車盗事案数の大きな増加が目立つ標準地域メッシュも存在する。とはいえ、標準地域メッシュ内の見守りカメラの設置数が31以上の標準地域メッシュでは、自転車盗事案数の差の75%分位点が0であることから、見守りカメラの設置数が多いこれらの標準地域メッシュのうち75%の標準地域メッシュで自転車盗事案数が減少していることを意味し、高密度での設置が一定の効果を上げていることが指摘されよう。

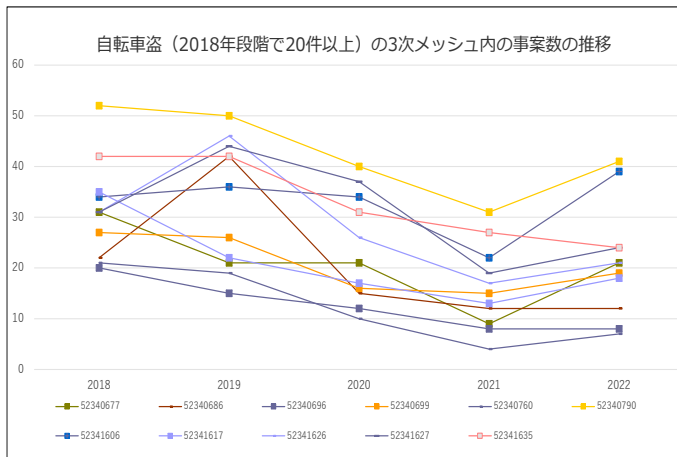


図 7 加古川市内での 2018 年の年間自転車盗 事案数が 20 以上の標準地域メッシュの 自転車盗の年次推移

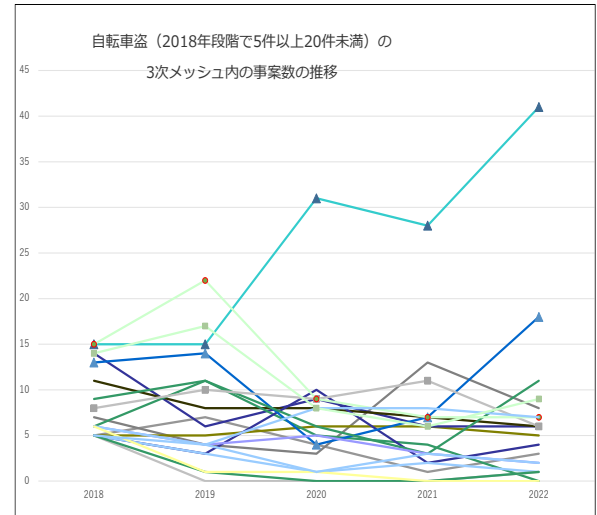


図 8 加古川市内での 2018 年の年間自転車 盗事案数が 5 件以上 20 件未満の標準地 域メッシュの自転車盗の年次推移

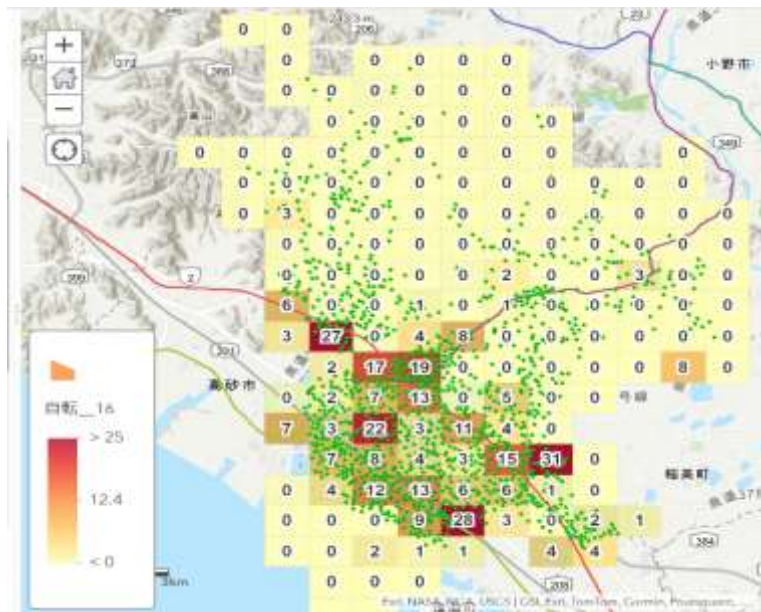


図 9 2023 年 1 年間における加古川市内の自転車盗事案の 発生状況 (詳細図 : <https://arcg.is/rfWGa> 参照)

これらの結果を踏まえ、2018 年以降、自転車盗が全体として増加傾向にあるメッシュを特定するため、個別の標準地域メッシュ単位の自転車盗事案数の増減を示すグラフ (図 7・図 8) を作成したところ、図 7 に示した 2018 年段階での自転車盗の発生事案数が比較的多い標準地域メッシュ (2018 年時点で 20 以上) では、おおむね減傾向、あるいは振動しながらも低減傾向を示すものの、自転車盗事案数が 2018 年時点での事案数が 15 程度のメッシュの一部で、増加傾向が確認された。特に急増が目立つ標準地域メッシュは、加古川市東南部の近年改装され 2022 年 3 月にオープンした大規模商業施設及び山陽電鉄別府駅を内包する標準地域メッシュであった。

加古川市内での 2023 年の自転車盗の発生状況に関する主題図表現 (図 9) を行ったマップからは、市内の中心部の加古川駅付近の都心型の商業集積地区、同市西部の工場跡地開発に伴い設置された大型商

業施設，同市東部の商業施設が密集するエリアで自転車盗事案が多数発生していることが確認された。加古川市内の標準地域メッシュを集計単位とした自転車盗に関する回帰分析結果

これらの事前分析を行ったうえで，各標準地域メッシュの領域内で発生した施錠・非施錠別の自転車盗数を説明変数とし，2020年度国勢調査の人口・世帯統計，人口移動・就業状態等及び従業地・通学地統計，2016年度の経済センサス活動調査の産業大分類別就業者数・従業者数を説明として利用した変数減少法による重回帰分析を行った。

施錠・非施錠自転車の両者の自転車盗の事案数に関する重回帰分析において，パラメータ推計値が正值となるであろうと想定，つまり自転車盗の増加要因であろうと想定される変数は，被害者実数が多い年齢層に対応する居住者である①高校在学者総数，②0歳～14歳人口総数，転居や世帯人員の増加に伴い新規の自転車購入がなされ，比較的使用期間の短く状態良い自転車が自転車盗の対象となりやすいと考えられるため，③居住期間5年未満の人口群，④世帯員人員数が5人以下の比較的規模の小さい一般世帯数群，利用する際に気のゆるみが生じやすいと考えられる⑤飲食サービス事業所数及び娯楽を含む宿泊業・飲食サービス，生活サービス業・娯楽業の事業所数，通院時の自転車の利用が一般的であると考えられる⑥病院・医院等の医療機関数，歯科医院数，⑦通勤通学に自転車を利用する人口の7グループの変数である。

パラメータ推計値が負値となるであろうと想定，つまり自転車盗の減少要因であろうと考えられる変数としては，①雇用総数及び生産年齢人口，犯罪抑止効果を持つと想定した②見守りカメラ数，③通勤通学時に鉄道・バス等の公共交通機関を利用する人口に関する変数群及び自家用車を利用する人口の3グループの変数である。

表1は，施錠自転車の自転車盗事案数に関する重回帰分析結果，表2は非施錠の自転車盗事案数に関する重回帰分析結果を示したものである。表1及び表2では紙幅の関係から，係数の有意水準が1%以下のものに限り表示することとし，薄赤色に着色したセルについては，係数の有意水準が1%以下で負，薄青色に着色したセルについては，係数の有意水準が1%以下で正值であることを示す。

施錠自転車の自転車盗数において，当初想定通りに有意に正值として符号条件を満たしたものは，居住期間1～5年の人口及び世帯員人員数が2～5人の比較的規模の小さい一般世帯数のみであり，他の変数は，符号条件を満たさず，またそれぞれの係数の正負が明確な結果とはならなかった。符号条件が当初の想定に反し，有意に負値となった変数としては，高校在学者数，0歳～14歳人口総数であった。この背景としては，自転車盗の被害防止のための施錠が学校等の教育機関において定期的実施され，施錠が推奨されていることが反映された結果かもしれない。歯科医院数が当初の想定とは異なり，有意に負となった結果に関しては，その理由を検討中である。

次に，非施錠の自転車盗の事案数に関する推計値である表2についても，有意に係数値が負値であると判断される変数群については薄赤色に着色し，係数値が有意に正值であると判断される変数群については薄青色で着色している。非施錠の場合の自転車盗の事案数に関する推計結果として，有意に正值となったものは雇用者数，宿泊・飲食サービス事業所数のみであった。逆に，当初の想定に反し通勤・通学に自転車を利用する人口と一人世帯のパラメータ推計値は，有意に負値となった。通勤通学に自転車を利用する人口の係数値が負となった背景としては，通勤通学に自転車を利用する人口と高校在学者数と15歳未満の人口などがかなり強い正の相関があるため，学校教育機関等での自転車への施錠の推奨により通勤通学に自転車を利用する人々の代理変数となった結果であると考えられる。

表1 施錠自転車に関する重回帰分析結果

パラメータ	推定値	自由度	平方和	"F値"	"p値"
サンプル数	157				
RMSE	2.327				
決定係数	0.451				
自由度調整済決定係数	0.403				
AICc	636.642				
BIC	671.642				
切片	-0.043	1	0.000	0.000	1.000
在学者 うち 高校 総数	-0.072	1	61.978	11.443	>0.001
居住期間 1～5年未満 総数	0.017	1	121.330	22.401	>0.001
居住期間 20年以上 総数	-0.019	1	226.231	41.768	>0.001
0～14歳人口 総数	-0.051	1	169.391	31.274	>0.001
15～64歳人口 総数	-0.010	1	102.594	18.941	>0.001
2人世帯数 一般世帯数	0.053	1	215.028	39.700	>0.001
3人世帯数 一般世帯数	0.037	1	30.458	5.623	>0.001
4人世帯数 一般世帯数	0.115	1	156.419	28.879	>0.001
5人世帯数 一般世帯数	0.190	1	141.188	26.067	>0.001
見守りカメラ総数	-0.090	1	47.907	8.845	>0.001
歯科診療所	-0.677	1	56.506	10.433	>0.001

表2 非施錠自転車に関する重回帰分析結果

パラメータ	推定値	自由度	平方和	"F値"	"p値"
サンプル数	157				
RMSE	3.554				
決定係数	0.530				
自由度調整済決定係数	0.496				
AICc	749.998				
BIC	780.006				
切片	-0.315	1	0.000	0.000	1.00
雇用者(役員を含む)(15歳以上) 総数	0.045	1	343.272	27.172	>0.001
15歳以上自宅外就業者・通学者 鉄道・電車	-0.022	1	109.023	8.630	>0.001
15歳以上就業者・通学者 乗合バス	-0.068	1	174.401	13.805	>0.001
15歳以上自宅外就業者・通学者 自家用車	-0.039	1	378.698	29.976	>0.001
15歳以上自宅外就業者・通学者 自転車	-0.027	1	148.571	11.760	>0.001
M宿泊業, 飲食サービス業	0.247	1	377.310	29.866	>0.001
N生活関連サービス業, 娯楽業	-0.281	1	177.346	14.038	>0.001
1人世帯数 一般世帯数	-0.023	1	231.853	18.352	>0.001
見守りカメラ総数	0.186	1	238.745	18.898	>0.001

なお、一人世帯のパラメータ推計値が有意に負値となる背景についても、現在検討中である。これら以外の変数については、有意に正負のいずれかであるとは言い難い結果が得られた。

次に、非施錠自転車盗の事案数を被説明変数としたモデルにおいては、係数値が正值となると想定した変数のうち、通勤・通学時にバス・鉄道等の公共交通機関を利用する人口と自家用車の利用する人口については、当初想定通り有意に正值、飲食・サービス事業所数についてもパラメータ推計値は正值であったが、生活サービス・娯楽業については、当初の想定とは異なり、パラメータ推計値は有意に負値であった。このような結果となった背景としては、生活サービス・娯楽業に含まれる生活サービス業の事業所数が多数であることにより影響を受けている可能性がある。

なお、表2で記述されていない変数のパラメータ推計値についても、有意にその正負について言及しがたいという結果が得られている。

これらの結果からは、施錠された自転車の盗難事案については見守りカメラは有効に機能していると考えられる反面、そもそも施錠すらなされなかった自転車については、見守りカメラが仮に十分存在・機能していたとしても自転車盗の抑止にはつながらないという状況があるものと言えよう。

今後の研究の展開

これらの分析結果は、大阪・神戸への郊外住宅地としての側面が強い加古川市域固有の結果である可能性があり、加古川市と同様に、近年、見守りカメラを多数設置しようとしている神戸市内における自転車盗の発生状況などの推計値などとの比較を行う必要があるものと考えている。

また、本報告で紹介した分析方法は、重回帰分析の手法を適用した結果であるが、自転車盗の事案数が正の整数値としてのみ定義されることを考えると、負の二項回帰分析など、より適切な方法での推計を試みる必要がある。

また、兵庫県のオープンデータサイトでは、自転車盗以外にも、ひったくり、オートバイ盗、自動販売機ねらい、自動車盗、車上狙い、部品狙いの6罪種についても、同様の分析を進める準備中である。

参考文献およびデータ入手サイト

加古川市(2023),「自転車盗難被害の注意喚起」, <https://www.city.kakogawa.lg.jp/soshikikarasagasu/shiminbu/shiminseikatsuanshinka/bouhann/17106.html>, (最終アクセス日:2024年3月26日).

加古川市(2023),「見守りカメラについて」, <https://www.city.kakogawa.lg.jp/soshikikarasagasu/shiminbu/shiminseikatsuanshinka/ICT/mimamori.html>, (最終アクセス日:2024年3月26日).

国土交通省(2024), 国土数値情報ダウンロードサイト, <https://nlftp.mlit.go.jp/ksj/>, (最終アクセス日:2024年8月6日).

杉浦 完征, 樋野 公宏, 浅見 泰司, 山田 育穂(2022), 鉄道駅周辺における環境要因と自転車盗発生との関係,「日本建築学会計画系論文集」, 87 (791), pp.123-132.

杉野弘明・有馬隆文(2014), 脆弱性と抵抗性に着目した自転車盗難に対する防犯力の発見,「都市・建築学研究」, 九州大学大学院人間環境学研究院紀要, 25, pp.25-33.

総務省統計局(2024),「統計地理情報システム」, <https://www.e-stat.go.jp/gis>, (最終アクセス日:2024年8月15日).

橋本 成仁, 矢田 篤史, 工藤 弘誉, 海野 遥香, 樋口 輝久(2022), 自転車盗多発区域推定モデルの構築に関する研究,「交通工学論文集」, 9 (4), pp.18-26.

兵庫県(2024),「オープンデータカタログページ」, https://web.pref.hyogo.lg.jp/opendata/index.php?p=57_n&desc=license&displayedresults=10, (最終アクセス日:2024年3月26日).

松本 巖(1982), 自転車盗の被疑者及び被害者の行動特性,「科学警察研究所報告 防犯少年編」, 23(1), pp.52-38.

三和宏幸(2019), 国内最大級の「見守りカメラ」日本初の官民連携「見守りサービス」,「新都市」, 73(2), pp.30-36.

村中 大輝・雨宮 護・大山 敏也(2016), 地方自治体による公共空間への防犯カメラ大規模設置事業の取り組み実態と評価,「都市計画論文集」, 51(3), pp.357-364.

ワインバーグ・G.M.(1990),「コンサルタントの秘密」(木村泉訳), pp.96-106.

SASがデータ解析の覇者になる最善の方策

—52年間で完成した全判別データに対応できる世界初のLSD（線形分離可能なデータ）の判別理論—

○新村秀一¹、新村秀樹²

(¹成蹊大学、²イノベストメント株)

The best way for SAS to become the leader in data analysis

-The world's first LSD discriminant theory that can handle all discriminant data completed in 52 years-

○Shuichi Shinmura¹, Hideki Shinmura²

(1: Seikei University, 2: Innvestment Corporation)

要旨

SAS と JMP を統計の個人家庭教師としてデータ解析の研究を行い、SCS（株）のブラブラ部長とスタッフの市川さんの2人でSAS/VAXを32の製薬企業と東洋信託銀行に販売した。また高機能で安価な数理計画法(MP)ソフトのLINDO製品を論文と書籍で紹介し販売した。統計とMPソフトでデータと数式で表される問題解決ができることを主張したが理解されなかった。しかしオバマ大統領がデータサイエンティストの重要性を示すと日本でも追従する統計学者が現れた。しかし最高品質の統計とMPソフトで、実証研究し成果を出すのが真のデータサイエンティストである。

1995年に米国で動物の遺伝子の出す蛋白量を計測するMicroarrayが商用化され、医師が癌と正常の約100例の5万個程度の発現量で「癌の遺伝子データ解析 (Theory2)」の研究を行い、使用データを公開した。多くの統計、機械学習(ML)、AI等の理工学研究者が論文を出したが全く人類の癌研究に役に立たない。筆者は169のMicroarrayが、全て最小誤分類数のMNM (Minimum Number of Misclassifications)が0 (LSD)であり、多変量の癌遺伝子を含むことを示した。さらにMNMを見つける最適線形判別関数のRIP (Revised IP Optimal LDF)で判別するだけで、n=100個以下のLSDのSM (Small Matryoshka)と、山中4因子に類似したほぼ5変数以下の最小次元のLSDのvital BGS (Basic Gene Set)を含むMatryoshka人形のデータ構造(Fact3)を初めて示した。

2023年に3冊目の本を執筆し、Theory2の技術で2015年までに分析した普通のデータによるTheory1 [1]で用いたデータを再分析し、驚くLSD判別の事実と効果を確認し、LSDの判別理論(Theory3)をSpringer [3]から出版し、来年に材料学会の国際会議と、先端的医薬の展覧会の特別講演を依頼された。

キーワード：LSDの判別理論、ChatGPTでできない組み合わせ最適化、癌の遺伝子データ解析、各種格づけ、試験の合否判定、スイス銀行紙幣の真札と偽札、日本車の小型車と普通車に見る産業製品、Irisデータに見る動植物の種の分類、2例の誤分類のある医学診断、間違ったIPソフトを検証する6データ

1. 世界が驚愕したLSDの判別理論

2024年4月13日に3冊目の判別理論 [3] を、世界最大の理工学出版社のSpringerから出版した。28,742円と高いが、Amazonは発売1週間で在庫切れになり、5月から2回目の在庫で受け付けたが6月

中旬の配送表示になっていて売れているようだ。日本の出版社の印税は通常 10% であるが、Springer は一律出版時に 750 ユーロである。しかし日本の出版社は、Fisher が確立した判別理論を完全否定した理論等、理解する努力もしないで出版してくれない。そこで英語が弱い私が、目のトラブルに悩まされながら 2023 年に日本語の下書き無しで執筆した。そのお陰で、来年 3 月にロンドンの「**第 7 回先端材料科学国際会議**」と 4 月にアムステルダムの「**第 3 回医薬品の発見、開発、デリバリーのイノベーションに関する博覧会**」の特別講演者の招待状が来た。

米国で 1995 年に動物の遺伝子の出す蛋白量（**発現量**）が Microarray で計測できるようになった。遺伝子の発現量は山中 4 遺伝子が示すように、生命現象を正確に制御する人類が出会った最高品質の計測値である。医師が、癌と正常を分ける多変量の発癌遺伝子の解明に取り組んだ。癌研究は不正研究を避けるため、高額な計測費用の Microarray データが無償公開されている。そして、統計、機械工学 (ML)、AI やバイオ工学等の研究者が、データ解析で癌の遺伝子診断（**Theory2**）にとり組んで多くの研究成果を出したが全て間違っている。

筆者は 169 の Microarray の 2 群が LSD である事を示した。恐らく品質が向上した RNA でもっと良い結果になるが、年齢の関係で私は研究できないので、SAS ユーザーの誰かが継承してほしい。この成功は、発現量が癌と正常を正しく 2 群に分ける多変量の癌遺伝子を含むことを示す。全研究が LSD を示していないので間違いである。そして例えば各 50 例の正常と癌患者の 1 万個の発現量は、変数 p が症例数 n より異常に大きいので、「高次元遺伝子データ ($n \ll p$) 解析は困難」というバズワードが論文で流行した。これは単に自分の研究の失敗の情けない言い訳である。彼らは Theory3 を習得し汚名を晴らすか、意味のない研究をやめるかの選択肢しかない。

山中先生が 3 万個以上と言われる人の遺伝子から万能細胞を作る少数の遺伝子を見つける目標を立てた。そして Rat の ES 細胞で活性化する 24 遺伝子を遺伝子 DB で特定した。これで 3 万以上の変数から 24 変数を選ぶ変数選択法 (FS) に成功した。間違った研究者は山中先生を見習うべきである。

筆者は 27 歳頃に、統計がデータで表される問題解決を行うので SAS を、数理計画法が数式で表される問題解決を行うのでシカゴ大学ビジネス・スクール Linus 教授の作った MP ソフトの LINDO 製品を自分の学習を兼ねて日本に紹介してきた。そして統計と数理計画法の共通テーマとして、判別データの MNM 見つける **RIP** を LINGO の Program1 で作成した。

2015 年 10 月 28 日に RIP で Harvard 大学医学部の Shipp らのデータ (77*7129) を判別すると、32 個の判別係数だけが非零で、残り 7097 個が 0 になり、簡単に 1 分で判別と変数選択が同時にできた。この 32 個の遺伝子の組みは LSD で、これを **SM** (Small Matryoshka) と命名した。Matryoshka は LSD と同義語である。残りの 7129 個を判別すると 2 番目の SM2 が求まり、RIP の判別を繰り返し SM に分割する Program3 を開発した。判別結果の **ER** (Error Rate) を評価する L00 を改良した k 重 CV が用いられているが間違いである。分析データを k 回コピーし疑似母集団の検証標本とする。各ケースに乱数を与えてソートし、 k 分割し学習標本とする。これが「母集団と標本の統計原則」を満たす正しい k 重 CV (Method1) であり Program2 を作成した。そして BGS が、検証標本の k 個の ER の平均値の **M2** を最小とする Best モデルになる事が分かった。BGS がオッカムの剃刀 (ケチの原理) を満たす最適モデルである。

Program3 で何と 7129 個の遺伝子が **222** 組みの LSD (SM) になり、残り 44 個の遺伝子が MNM=1 の雑音になった。これらも他の新しい遺伝子との組み合わせで SM になる可能性がある。既に医学的に正常細胞が癌化すると、「エピジェネティック変化」という異なった遺伝子の発現が起こる。この変化で 2 群が LSD の多くの多変量の発癌遺伝子の候補が見つかった。そしてこの多くの候補から、真の多変量の発癌

遺伝子を特定する4つの検証法も考えた。医師が Validation4 で多変量の癌遺伝子の特徴を、既に見つかっている400個以上の「遺産的な1変量の発癌遺伝子」の特徴で検討できる。

遺伝子データは横長のLSD ($n \ll p$)で、それらが n 個以下のSMから最小次元のLSD (BGS) 迄を含む「マトリョースカ人形の構造を持つという事実 (Fact3)」は、「世界初のLSDの判別理論の学際的な成果」で、全てのDNAや解析していないRNAを含む全判別データに適用できる。医療は新しい計測器が次々に開発されるので、**分析は永遠に繰り返す**必要がある。

222組みのSMに含まれる遺伝子数は[18, 62]で、必ず $n=77$ 症例以下である。一気に7129個から77以下のSMが見つかることは不思議かもしれない。これはRIPの定義域が n 次元以下の凸領域であるからである。これは山中教授が遺伝子DBの利用で解決したことと似ている。但し、 2^{77} 個の凸領域固有のNMの中からMNMの集合である最適凸体という一つの組み合わせ最適解を見つける必要がある。これはChatGPTでも統計でもアプローチできない。整数計画法 (IP) の分枝限定 (B&B) 法だけで実現できる。しかし「企業の要員計画や病院の看護師のスケジューリング計画」をIPで商品化している商用のIPソフトでも、今回3章で提案する6データで正しい組み合わせ最適解を求めないソフトがあり驚いた。これ迄IPソフトの正しさを検証する方法がなく野放しの状態だった。

SASは統計ソフトの中で唯一MPソフトをサポートしている。AIに対応することも重要だが、IPで筆者のTheory3を実現し、統計ソフトの覇者になることを強く切望する。

「高次元遺伝子データ解析は困難」という言い訳を論文に記す理工学研究者は、公的な研究費の無駄遣いである。余談であるが筆者は、不幸にして日本で評価されず公的研究費は貰っていない。それ以上に問題なのは、数学好きな多くの研究者が、高校数学で習う「連立方程式の解の条件」を自分の研究に生かせなかった点である。高次元データ ($n \ll p$) は、単にSall博士が2015年11月10日のJMPユーザー会で示した「横長データ ($n \ll p$)」と考えるべきである。 p の大きさは全く関係しない。彼の基調講演は、特異値分解 (SVD) を用いた高次元LDFとPCAの紹介で、Microarrayを分析していたがLSDでない。1ヶ月無償で借りて6データが**2%から17%**のERで、正しくLSDを識別しないことを報告した。残念ながら、JMPの折角の高次元LDFとPCAは日の目を見なくなった。しかしこの事実を知らない統計研究者が、研究をしている場合は忠告しているが無視されている。それと同じくLASSOでTheory2の研究をする統計研究家もいる。この2つの統計研究は、全くLSDには役に立たない判別研究であった。

JMPのロジスティック回帰や重回帰分析等の線形モデルで横長データを分析すると、 $DF=n$ で p 変数から n 変数を選んで普通のデータ解析として処理する。この事実を知らないことが、統計ソフトの使用を見下す頭でっかちの統計研究者の欠点である。しかし ${}_p C_n$ 組みの多くのモデルを検討する必要がある。恐らく以上の処理を正しく行っていない統計やMPソフトも多いと考える。そこで3章の6データで事前に確認する事が重要だ。またSASがRIPを提供すれば、多くのSASユーザーが、世界最先端の判別理論による成果を出せる。恐らく多くの企業の正しい評価システムにも役立つ。

また判別結果を表す誤分類数のNMは、0から $(n-1)$ 迄の n 個の整数値しかないので、「組み合わせ理論」で考えるべきだ。この主張に対して生半可な統計家は、データに固有なNMは信用できず正規分布に従うべきという。彼らの間違いは次の4点である。

- 1) 判別理論は、クラスを表す情報を「**外的基準**」と呼ぶ。明らかに外的基準は、統計で分析する計測値の上位概念である。これが理解できない人は、判別理論に関係しない方がよい。
- 2) 世界で一番判別分析の実証研究をしている筆者は、3変数以上のFisherの仮説を満たす2群の正規分布に出会ったことがない。まして1万以上のMicroarrayを考えるのは異常である。さらに2変数

以上の正規性の検定は研究されているが、広く認められた検定はないと考える。

- 3) 重回帰分析の予測値が 1 変量の正規分布と考えるのは、Gauss が 2 地点間の繰り返し測定で見つけた釣り鐘状の Gauss 分布の前提と合致する。また Fisher は実験計画法や分散分析に関係したが、これらも問題はない。しかし判別分析に適した Fisher の仮説を満たす実際の現象はない。
- 4) Fisher は、開発した LDF を Iris データで評価した。後世の統計家は彼を統計の神に祭り上げるため「Fisher の Iris データ」と呼んだ。Fisher を侮辱する行為であることを理解していない。後で紹介するように Iris データは LSD 判別に一番適したデータで、後世の判別理論の研究者に託した「Fisher の遺言データ」と考えるべきだ。筆者が彼の願いを実践した、判別理論の後継者である。

以上を纏めれば、癌の高次元遺伝子データは n 個以下の多くの LSD である SM に分割できる。更に山中 4 因子と同じ最小次元の LSD である BGS に分割できる。Microarray のような横長の LSD は、2 章で示すように p 次元の LSD から複数個の BGS の一つを最後に含む Matryoshka 人形の入れ子状の構造 (Fact3) をもつ。これは高次元の LSD でなくても、試験の合否判定、動植物の種や産業製品の分類等の LSD の他、誤分類のある 2 群判別データ全てを RIP で特定した誤分類を省いて LSD にできるので、全判別データに適用できる汎用的なデータ構造である。

180 例の自然分娩患者と 60 例の帝王切開患者の 19 変数の計測値で、出産前に自然分娩か帝王切開にするかの診断モデルの研究を、日本医科大学の鈴木教授から依頼された。2016 年に出版した Springer1 [1] で Sall 博士の SAS Regression Application と Goodnight 博士の「掃き出し演算子」の SAS のテクニカルレポートを翻訳した「SAS による回帰分析の実践 [4]」の多重共線性の知識で分析した。2024 年に Springer3 を執筆しながら、RIP で判別すると 2 名の患者がデータに一意に誤判別される。これを省いた 238 例は LSD になる事の重大さに気づいた。そこで Program4 で 54 万 ($=2^{19}-1$) 以上のモデルの中から 14 変数の BGS を見つけた。これは最小次元の LSD である。残りの 5 変数で 32 ($=2^5$) 個のモデルができ、5 変数に BGS を加えた 32 モデルだけが LSD である。ほぼ 54 万以上の他のモデルは判別に値しないので統計検定無しで棄却できる。LSD という分かり易い基準を使えば、統計検定も無意味になる。

そして間違っって普及している k 重 CV に代わって、「統計の母集団と標本の間隔を満たす k 重 CV」で、BGS が k 組の検証標本の平均 ER の M2 が最小モデルを Best モデルとして選んだ。このような見事な「オッカムの剃刀」を満たすモデルは、今まで誰も発見できなかった。そして省いた 2 例の誤分類患者を詳しく研究することが、判別分析本来の目的である。これが世界初の「**ケースの選別法**」で、全判別データの福音になる。試験の合否判定、各種格づけ等の信頼性が向上する。そして産業機械や動植物の計測値の品質が高いので、これまで以上の新発見ができるであろう。これが 3 月の国際材料会議の特別講演のテーマにしようとする。3 群以上の判別データは、2 群の組み合わせで考えるべきだ。

重回帰は、説明変数で目的変数の計測値を予測する手法である。両方は同じ計測値である。しかし判別理論は、例えば医師が経験と知識で決めた分類情報 (外的基準) の 2 群を、計測値で識別し外的基準を詳しく分析する手法である。この場合、分類情報は説明変数の上位概念である。そして n 例の 2 群があれば、NM (あるいは ER) は 0 から $(n-1)$ までの n 個の整数値しかとらない。これを無限の正規分布と考える Fisher 以降の判別理論は誤りで、「組み合わせ最適化」で考えるべきである。CPD の 54 万以上のモデルから最小の最適組み合わせモデルを見つけることが数分でできる。遺伝子診断に用いる多変量の発癌遺伝子を見つける最先端の研究は、ChatGPT や統計でこの組み合わせ最適解は得られない。不幸にして 1995 年以降の全理工学研究がこの研究に失敗した。米国の医学研究は統計研究者を雇い分析するが、当然失敗した。理由は、彼らは Book Smart で筆者のような統計と数理計画法ソフトで実証研究する Street

Smart な真のデータサイエンティストでないからである。理論統計家は、自分の研究が何の解明に役立つかを考えるべきである。頼りない記述統計を、Fisher らの世代の統計研究者は推測統計学で数学に近づけることに成功した。次の世代は、医学や工学のように人類に役立つことに舵を切るべきだ。数学に恋い焦がれる理工学研究者は反省すべきだ。

2004 年迄の第 1 世代の Microarray の医学プロジェクトは失敗した。そのために 2004 年以降のより精度の高い第 2 世代の Microarray が開発されたが、医学研究は過去の顕微鏡と生物的知識で見つけた 1 変量の 400 個以上の発癌遺伝子や抑制遺伝子を組み合わせる多変量の癌遺伝子の候補とする後向きの研究をしている。筆者の Theory3 で正しい vital BGS (M2=0 で 5 個程度以下の遺伝子を含む) を癌研究に用いるべきである。医学以外の判別データの中には、計測値の精度が落ちる分野もある。この場合、誤分類数が多いかもしれない。しかし誤分類を省いた LSD の結果で、その誤分類例を後で詳しく研究すれば新しい進歩につながる。各種格付け等の多くの社会に役立つテーマに利用できる。

筆者の開発した MP ソフトの LINGO の 4 プログラムの使い方を知れば、個人が消費電力の少ない PC で十分に最先端の研究に対応できる。また重要な誤分類のある医学診断は、CPD と同じ方法で診断すれば良い。さらに統計ソフトに実装して、統計手法の一つとして利用できる方が良い。

しかし Springer3 を出版後に、統計ソフトは SAS と JMP を用いれば良いが、多くの商用 IP ソフトが正しい組み合わせ最適解を求めないことに気づいた。そこで、SAS と JMP が 3 章で紹介する 5 例以下の 2 変数の 6 データを IP で解く RIP を提供し、統計ソフトの覇者になることを希望する。すでに 2017 年頃、MP ソフトを持たない JMP に、LINDO の c ライブラリーで開発することを提案したが断られた。SAS は唯一 MP を提供しているので筆者の親切な提案を受け入れる余地はある。ただし多くの商用 IP ソフトと同じく、正しい解は見つけないと考える。その場合は、c ライブラリーを購入し製品に実装すれば良い。間違いなく SAS や JMP の医療システムにも役立つだろう。筆者が、その開発とマニュアルの作成に協力できる。

2. Fisher の Iris データで LSD の Fact3 の説明

Fisher の Iris データは、**図 1** の赤のセトサ(G1)と青のバージニカ(G2)と緑のバーシクル(G3)の各 50 例で 4 変数の 3 群データである。動植物の種は、形態や臓器や種々の観点から決まる外的基準である。それを単にがく片、がく片幅、花びら、花びらの幅という 4 計測値で判別できれば、実用上役に立つわけだ。

図 1 の行列散布図から分かるように G1 と G2 の G1&2 と、G1&3 と G1&2&3 の 3 群が LSD で、G2&3 が LSD でない。2 変数以上の公認の正規性検定がないので Fisher の仮定を満たすか否かを判断できず、また Fisher の仮説を満たす Iris 以上のより良い実際のデータを見つけることができない。

表 1 は LSD でない G2&3 の Var 列で示す 15 モデルの RIP

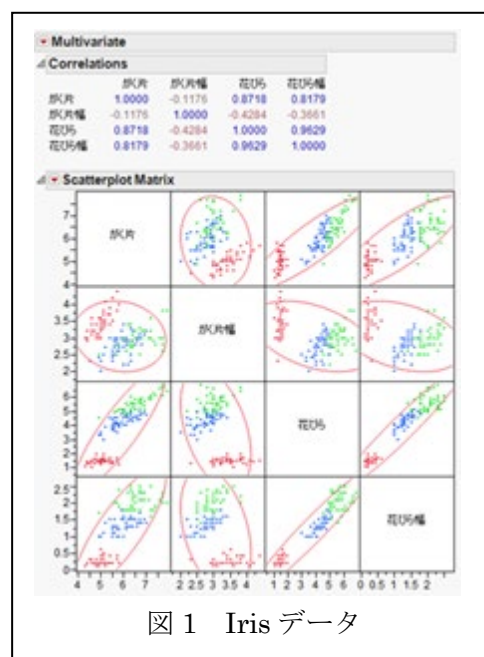


図 1 Iris データ

の MNM と、ロジスティック回帰、LDF、QDF と RDA（正則化判別関数）の NM から MNM を引いた差を表す。ロジスティック回帰は、(X1, X2, X3) で 2、(X3, X4) で 1 だけ大きく、残りの 13 モデルで MNM と同じである。2 次判別曲線で判別する QDF が LDF より良いが、線形超平面の LDF の ER の方が 2 次曲線の ER より信頼性が高い。RDA は Harvard 大学の教授との共同研究の手法であるが、2 組のパラメータの最適な組み合わせを選ぶ必要がある。JMP のお薦めのパラメータ値で種々検討したが結果は悪い。ソフトマージン最大化の SVM(S-SVM) も 1 個のパラメータを 0 から 1 の範囲でチューニングが必要である。筆者はそれを 10 等分して検証した。そして全ての組み合わせから分析データに最適な判別関数を選べることを OR 誌[8]で報告した。しかしチューニング法で悪戦苦闘するより、RIP という判別分析のスーパースターを SAS が提供し、ユーザーが世界最先端の癌の遺伝子解析や各種格づけや合否判定等ができる方が良い。

2024 年に執筆中、図 1 を見て 3 つの BGS の指摘が重要なことに気づいた。X3 と X4 は 2 組みの 1 変量の BGS で 3LSD を識別できる。さらに (X1, X2) が 1 組の 2 変量の BGS で G1&23 が識別できる。そして次の 14

体の Matryoshka 人形は、4 変数の LSD から必ず 3 組の何れかの BGS を含む入れ子構造になる。山中 4 因子と発現量は、巨大な Matryoshka 人形である事が分かる。これで遺伝子の出す発現量が示す、LSD の Fact3 を分かり易く説明できた。

人形 1: (X1, X2, X3, X4) ⊃ (X2, X3, X4) ⊃ (X3, X4) ⊃ (X3), 人形 2: (X1, X2, X3, X4) ⊃ (X2, X3, X4) ⊃ (X3, X4) ⊃ (X4),
 人形 3: (X1, X2, X3, X4) ⊃ (X2, X3, X4) ⊃ (X2, X4) ⊃ (X4), 人形 4: (X1, X2, X3, X4) ⊃ (X2, X3, X4) ⊃ (X2, X3) ⊃ (X3),
 人形 5: (X1, X2, X3, X4) ⊃ (X1, X3, X4) ⊃ (X3, X4) ⊃ (X4), 人形 6: (X1, X2, X3, X4) ⊃ (X1, X3, X4) ⊃ (X3, X4) ⊃ (X3),
 人形 7: (X1, X2, X3, X4) ⊃ (X1, X3, X4) ⊃ (X1, X4) ⊃ (X4), 人形 8: (X1, X2, X3, X4) ⊃ (X1, X3, X4) ⊃ (X1, X3) ⊃ (X3),
 人形 9: (X1, X2, X3, X4) ⊃ (X1, X2, X4) ⊃ (X2, X4) ⊃ (X4), 人形 10: (X1, X2, X3, X4) ⊃ (X1, X2, X4) ⊃ (X1, X4) ⊃ (X4),
 人形 11: (X1, X2, X3, X4) ⊃ (X1, X2, X4) ⊃ (X1, X2), 人形 12: (X1, X2, X3, X4) ⊃ (X1, X2, X3) ⊃ (X2, X3) ⊃ (X3),
 人形 13: (X1, X2, X3, X4) ⊃ (X1, X2, X3) ⊃ (X1, X3) ⊃ (X3), 人形 14: (X1, X2, X3, X4) ⊃ (X1, X2, X3) ⊃ (X1, X2).

3. 6 データで IP と統計ソフトが利用できるか否かの検証

3.1 52 年の研究の概略

筆者は 1971 年、京大の数学科の大学院に落ち、10 月頃に大卒の入社試験は終わっていたが NEC と SCS を特例で受け採用された。現 SCSK の一期生の 10 月迄の新人教育を受けず、6 月頃に NEC に派遣され、大阪成人病センターの野村医長預かりとなった。彼と NEC の開発する ECG 診断の千例程の正常と 10 以上の異常所見の 100 変数以上の計測値で、Fisher の LDF と QDF で診断論理の研究を 4 年間行った。しかし野村医師の開発した「枝別れ論理」より劣っていた。役に立つ判別理論の完成が、一生の研究テーマになった。2015 年にスイス紙幣データ (200*6) の真札と偽札で世界初の LSD 研究を LINGO と SAS や JMP

表 1 5 判別関数の比較

SN	Var	RIP	Logi	LDF	QDF	RDA
1	1, 2, 3, 4	1	0	2	2	2
2	2, 3, 4	2	0	2	2	1
3	1, 3, 4	2	0	1	1	2
5	1, 2, 3	2	2	5	6	4
4	1, 2, 4	4	0	1	2	1
7	3, 4	3	1	3	0	2
8	1, 3	4	0	2	2	2
9	1, 4	5	0	1	0	0
6	2, 4	5	0	0	2	2
10	2, 3	6	0	1	1	1
11	1, 2	25	0	0	4	4
12	4	6	0	0	0	0
13	3	7	0	1	0	0
14	1	27	0	0	3	3
15	2	37	0	5	5	5

で行い、Fisher を超える判別理論を完成した[1]。

そして2016年から無謀と思える正常と癌の約100例の1万個程度の6種のMicroarrayで癌と正常を分ける研究を行い、世界初の成功例を出版した[2]。2019年からは、163の2世代のMicroarrayを4種のLINGOプログラムを作り、数多くの誰も見つけていない成果を国際会議で発表し、その後にIEEEやSpringer等の研究叢書に14の研究論文として採択された。この方法は、自費で研究している筆者の論文を安く出す近道である。また研究者のSNSのResearch Gate (<https://www.researchgate.net/profile/Shuichi-Shinmura>)の利用を薦める。

2022年に日本癌学会の出す学会誌の論文掲載料が30万以下と他の医学誌に比べて安いので投稿したが不採用になった。編集長から論文に採用する重要研究テーマでないので採択できないというメールが届いた。そこで2023年にSpringerから52年間の研究をまとめた世界初の「LSDの判別理論のTheory3」を出版することにした[3]。27の画期的な発見、考え方や技術があるが、MNMを見つけるRIPによる4種のProgramを用いれば、全判別データが解決できる。多くのユーザーが、世界最先端の「癌の遺伝子データ解析」や、各種格づけ、試験の合否判定、動植物の種の分類、各種医学診断の研究を継承してほしい。しかし多くのMPや統計ソフトは正しい結果を出さない恐れがあり、それを確かめる6データで事前に自分のソフトの検証を行ってほしい。6データの判別結果を教えていただければ、癌診断と画期的な判別理論に関して手助けする意思はある。

3.2 MNM を見つける画期的な方法

図2はRIPをLDFの判別係数空間で見つける説明である。次の3例*2変数でできる7凸体からMNM=0のLSDになる最適凸体を見つける問題は、IP研究者の想定外の問題である。多くのIPソルバーは正しい組み合わせ最適解を見つけられない可能性がある。 x_1 はG1に属し、クラスを表す識別子 y_i を1とし、 x_2 と x_3 はG2に属し、 y_i を-1とする。G1 ($y_i = 1$): $x_1 = (-1/18, -1/12)$,

$$G2 (y_i = -1): x_2 = (-1, 1/2), x_3 = (1/9, -1/3)$$

H_i は、LDFの係数空間を次の3判別超平面で分割する。

$$H1 : -b_1/18 - b_2/12 = -1;$$

$$H2 : -b_1 + b_2/2 = 1; H3 : b_1/9 - b_2/3 = 1 \dots \dots (1)$$

H_i は、係数空間を判別スコア(${}^t\mathbf{x}_i * \mathbf{b} + 1$)の正と負の半平面に分割し、7つのCP(凸多面体)になる。RIPが x_1 に対応するH1の+半平面内の全ての点に対応するLDFは x_1 を正しく判別し、-半平面の点は x_1 を誤判別するLDFになる。そして各凸体の数字はNMで内点を囲む負の半平面の個数である。三角形がMNM=0の最適凸体である。この定式化は分かりにくいので、3.7節で2群データを連立不等式で解く方法を説明し、多くの方が理解しやすいように工夫した。

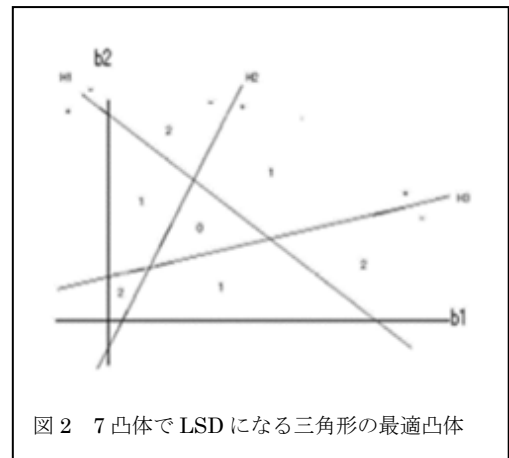


図2 7凸体でLSDになる三角形の最適凸体

3.3 4種のLSDと2種のMNM=1のデータ

表2は、4種のLSDと2種のMNM=1の6データである。 x_1 と x_2 の2変数が説明変数で、 $y_i=1$ が群1を、 $y_i=-1$ が群2を表す識別子の外的基準である。RIP行は6データのRIPの判別係数である。MNM=0の4データはLSDで、MNM=1は1個の誤分類を示す。2変数をプロットすれば、この結果は確認できる。そして使用するIPと統計ソフトでRIPと同じ結果が出れば、安心して世界最先端の研究に参画できる。恐らく多くのIPソフトが正しい解を出さないと考える。

表 2 2変数の6判別データ

		x1	x2	y _i	SCORE		x1	x2	y _i	SCORE	
Data	RIP1	4	1	-6	MNM=1	Data	RIP4	-1	-2	7	MNM=0
A	1	1	3	1	1	D	1	0	3	1	1
	2	2	2	1	4		2	1	1	1	4
	3	2	-1	1	1		3	4	1	1	1
	4	1	1	-1	1		4	3	3	-1	2
	5	2	1	-1	-3		5	4	2	-1	1
	RIP2	2	2	-7	MNM=0		RIP5	-4	-1	0	MNM=0
B	1	1	3	1	1	E	1	-1	-1	1	5
	2	2	2	1	1		2	0	1	-1	1
	3	3	1	1	1		3	1	-3	-1	1
	4	1	1	-1	3	RIP6	-1	0.5	1	MNM=1	
	5	2	1	-1	1	F	1	-1	-1	1	1
	RIP3	1	1	-1	MNM=0		2	1	1	1	1
C	1	1	1	1	1		3	0	1	-1	-1.5
	2	1	-1	-1	1	4	1	-3	-1	1	
	3	-1	1	-1	1						

3.4 新村の仮説と提案

MP と統計ソフトを利用している人は、是非この6データで検証し結果を教えてください。

表 3 は、表 2 の6データの判別結果で、RIP だけが6データを正しく識別する。恐らく次の理由で、IP ソフトの多くは6データで正しい解を出さないと考える。

- 1) MP の基本の LP (線形計画法) は高校数学の「領域の最大/最小問題」が示す通り、1 個の凸体を定義域とし、その制約条件の範囲で線形の目的関数の最大/最小問題を解く。LP で問題のある商用ソフトは考えられない。
- 2) しかし多くの学会の非線形最適化の発表で、追試できる例を後で検証すると間違っ了解を出す有名な商用ソフトがある事が分かった。大域的最適解を求める技術力がない企業のソフトと考える。小さなモデルは、LINGO の無償の評価版で確認できる。

	A	B	C	D	E	F
RIP	1	0	0	0	0	1
LDF	2	0	0	0	1	2
Logistic	1	0	0	0	1	1

- 3) IP を用いた RIP は、n 例の値を係数とする線形超平面 H_i で多数の凸体に分割した中から、MNM をもつ最適凸体を見つける。これは、これまでの IP 研究が考えていない問題である。恐らく商用の IP で要員計画等を商品化している開発者は、複数の凸体から最適凸体を探す事を考慮していない。
- 4) この問題を解決できるのは、高速な B&B 法を開発した LINDO 製品と、それと同等の少数の IP ソフトしかないと考える。多くの商用ソフトは対応できないと考える。

3.5 IP が統計や ChatGPT に負けない研究と応用

日本に SAS と JMP や数理計画法ソフトの LINDO 製品を本や研究論文で紹介し普及に貢献してきた。そして全判別データを LSD にして、CPD データ [3] で ChatGPT や統計でできない膨大な組み合わせ最適化問題

で癌の遺伝子データや誤分類のある医学診断等を解決した。MP は、LP の配合問題、IP の要員スケジューリングや政策決定、QP (2 次計画法) のポートフォリオ分析、NLP (非線形計画法) の DEA [6] 等が社会に貢献している。そして RIP が癌の遺伝子診断、誤分類のある医学診断、製品や企業評価等の各種格づけ、試験の可否判定、動植物や産業製品の判別などに貢献できる道を切り開いた。医学では、DNA から現時点では RNA の新しい機器が開発され、最新研究は最新の RNA の分析が要求される。筆者がそれに対応できないので、後進のやる気のある人に対応してほしい。多くの若手研究家が、私の国際的に評価の高い成果を踏まえて、この分野で世界最先端の研究成果に挑戦できる。

そのために、事前に分析に用いるソフトが正しいか否かを簡単に確認できる 6 データを作成した。結果を共有したいので、判別結果を教えてほしい。協力していただいた方には、筆者の研究成果等の資料を無償提供する事も考えたい [5, 6]。また次回の発表で協力者として紹介したい。

ChatGPT が世界中で注目され、統計や MP 等の多くの既存の学問が飲み込まれ衰退の危機にある。その中で MP は、私が開発した「LSD の判別理論」が誤分類のある医学診断や 169 の高次元遺伝子診断の最先端の医学研究で、「ChatGPT や統計でも対応できない組み合わせ最適解」が、個人の少しの努力と少ない消費電力の PC で、B&B 法の IP アルゴリズムで簡単に解決できた。即ち、IP による世界初の LSD の判別理論を理解すれば、最先端の研究成果を世界に発信できる。

3.6 1971 年から 52 年間の判別理論の 3 研究

以下の 3 期で、3 種の判別理論を完成した。

1) Fisher を超えた判別理論 (Theory1; ~2015) :

1971 年から判別理論を研究し、判別分析の 4 問題を Fact1 (LDF と NM の関係) と Fact2 ($MNM_k \geq MNM_{(k+1)}$) で解決した。判別関数の NM が信頼ができず、式(2)で MNM を見つける RIP を定式化した。

$$\text{MIN} = \sum e_i; y_i * (\mathbf{x} * \mathbf{b} + 1) \geq 1 - 1000 * e_i; \quad (2)$$

そして真札と偽札各 100 枚の 6 変数データ (200 * 6) を判別すると、(X4, X6) が最小の 2 変数の BGS になった。BGS を含む 16 (=2⁴) モデルが LSD で、残り 47 モデルは MNM が 1 以上の役に立たないモデルである事が Fact2 から分かる。筆者は、外的基準の 2 値に 1/-1 の整数を与えて重回帰分析として扱い、モデル選択を「全ての組み合わせ或いは逐次変数増加/減少法」で行い、AIC や BIC や Cp 等の統計量で検定した。このような「重回帰を用いるまやかしの検定」もいない。14 個の LSD と 47 個の MNM が 1 以上のモデルに分けることができる。その上で、正しい k 重 CV で BGS が最小 M2 の Best モデルを選ぶ事ができる。これ程強力なモデル検定と FS はない。LSD 研究の重要な成果である。

2) 169 の Microarray の解析 (Theory2; ~2023) :

1995 年から医学研究が、Microarray の発現量で癌の遺伝子データ解析 (Theory2) を研究した。研究不正を避けるためにデータが無償公開され、多くの理工学研究者も研究した。しかし全研究が以下の事実から間違いと断定できる。RIP で Shipp (77 * 7129) を判別すると 1 秒で LSD が分かる。即ち 2⁷¹²⁹ 組の凸体から MNM になる最適凸体を見つけた研究は皆無で、全理工学研究は間違いと断定できた。さらに 32 個の遺伝子の係数だけが非零で、これを SM1 と呼ぶ。残りの 7097 個が 0 であり、これを再度判別すると SM2 が求まる。そこで LINGO で SM に分割する RIP の Program3 を開発した。169 データで、LSD である多くの Type-1 の SM と、ほぼ 1 組の MNM > 0 である Type-2 の SM に分割できた。残念ながら医学研究もこの事実をまだ知らない。Program4 で BGS 分割できる。最尤推定法を用いた JMP のロジスティック回帰は、Alon (62 * 2000) の 2000 個を判別すると DF=62 で残りの 1938 個の遺伝子をモデルから省き、DF 分割できた。以上から Microarray のような横長の LSD は、その中に DF、SM という LSD を含

み、最後に BGS を含むマトリョシユカ人形の構造を持つ (Fact3)。169 の Microarray は全て LSD で Fact3 である。

開発した LINGO の 4 プログラムと判別結果の 4 検証法を組み合わせた、簡単で使いやすい 2 段階スクリーニング法 (Method3) を開発した。そして、10 重 CV の 10 個の検証標本の平均 ER の M2 が 0 で、遺伝子数が 5 個以下の vital BGS を見つけることを提案した。この vital BGS を Validation4 で、既に医学研究で見つけた 400 個以上の 1 変量の遺伝的発癌遺伝子が含まれているかを調べる事を提案した。vital BGS が含む遺伝子の特徴で、多変量の発癌遺伝子の特徴が分かり、新しい癌診断の世界が開かれた。

3) 世界初の LSD の判別理論 (Theory3 ; 2024)

2019 年以降の第 2 世代の Microarray の研究は論文でしか報告していない。そこで 2024 年に本で紹介しようと考えて執筆を始めた。そして Theory2 の技術と知識で、Theory1 で分析したデータを見直して驚く成果を得た。これで Theory1 と Theory2 を総合化した全ての判別に対応した Theory3 が完成した。

一番の収穫は、19 変数の CPD データで、術前に手術法を決定する問題である。Program1 の RIP で判別して MNM=2 が分かる。この 2 例を省いた 238 例は LSD になる事に気づいた。そこで Program4 で 19 変数の 54 万 ($=2^{19}-1$) 以上のモデルから 14 変数の BGS1 が見つかった。この BGS1 を含む 32 ($=2^5$) モデルだけが LSD で、残りの 54 万以上のモデルは雑音である。また $X_{12}=X_{13}-X_{14}$ という等式制約で新しい X_{12} が作られていて多重共線性の問題がある [4]。Springer1 では半年以上かけてこれを解消した。しかし BGS1 は最小次元の LSD であり、 X_{12} を含むと冗長な X_{13} と X_{14} は含まず多重共線性が自動的に解消された。また山中 4 因子に発癌性のある C-Myc が含まれるが、発癌性の少ない L-Myc に置き換えた別の山中 4 因子もある。同様に、CPD も X_{12} だけを含む BGS1 の他、 X_{13} と X_{14} だけを含む 3 種の BGS がある事が分った。これは iPS 研究と Theory2 が発現量で正確無比に制御されている現象であるからと考える。

4) 正しく LSD を判別する 3 つの LDF

RIP とハードマージン最大化の H-SVM が、理論的に LSD を正しく $SV_1 \geq 1$ であれば群 1, $SV_2 \leq -1$ であれば群 2、そして区間 $(-1, 1)$ にケースがないことで LSD を定義できる。しかし H-SVM は、誤分類のあるデータで計算エラーになり、また RIP のように多くの係数を 0 にする FS の機能もない。

最尤推定法を用いた JMP のロジスティック回帰は、6 種の Microarray の全ての Type-1 の SM が $NM=0$ で経験的に LSD になる。結局この 3 種の LDF だけが LSD を認識できる。これで、各統計ソフトのロジスティック回帰が LSD で $NM=0$ になるか否かを調べる必要がある。ロジスティック回帰はほぼ $DF=n$ の変数を選び、SM や BGS より変数が多いので使い道に限られる。例えば医師が遺伝子遺伝子を組み合わせで作った多変量の発癌遺伝子の候補が正しい組み合わせか否かは、ロジスティック回帰の NM が 0 か否かで確認出来る。

合格点が 50 点で、大問 2 問で合否判定を行う場合、式(3)の自明な LDF が得られ自明な LSD になる。

$$LDF = T1 + T2 - 49.5 \quad (3)$$

ただし試験毎に合格点は変わるので、式(3)の定数項の 49.5 を調整して、 $LDF > 0$ で合格、 $LDF < 0$ で不合格にすれば良い。しかし 0 判定は収束計算の判定が難しいので $LDF \geq 1$ で合格、 $LDF \leq -1$ で不合格にする。これが SVM の SV の考え方である。しかし LSD でない場合、誤分類されるケース i の e_i を 1、正しく分類されるケースの e_i を 0 とする 2 値の整数値で、1 の代わりに $(1-1000*e_i)$ にする。これが式(2)の右辺定数項を LSD であれば e_i を 0、誤分類されれば 1 にするこれ迄にない IP 問題の定式化で、誤分類データに対応できる。5 例の場合、32 ($=2^5$) 個の連立不等式 (32 の凸体) から MNM になる最適凸体を見つ

ける新しい問題になる。即ち多くの IP ソフトが、32 個の凸体の NM ($\sum e_i$) の中から MNM ($\text{MIN} = \sum e_i$) を見つける事ができないと確信する。

3.7 RIP を複数の連立不等式で考える

表 2 のデータ A は、SN=1, 2, 3 が群 1 ($y_i=1$) の計測値で、群 2 ($y_i=-1$) は SN=4, 5 の 2 例である。そして RIP でデータ A を判別すると RIP 列に $f(\mathbf{b})=4x_1+x_2-6$ という LDF と MNM=1 即ち SN=5 の判別得点が -3 の負の 1 例が誤判別される。この 5 点を平面にプロットとすれば簡単に確認できる。式 (2) の RIP の定式化は分かりにくい、判別を連立不等式で解くことは理解しやすい。

- 1) $LDF=b_1*x_1+b_2*x_2+b_3$ は、5 個の 2 変数 (X_1, X_2) の値から、 $LDF > 0$ と $LDF < 0$ で 2 群を分ける判別係数 (b_1, b_2, b_3) を LDF の判別係数空間で見つける。
- 2) 係数空間で式 (2) は次の連立不等式 (4) になる。5 個の 2 値変数 e_i で $32 (=2^5)$ 組みの連立不等式の凸体から MNM になる 1 個の最適凸体を選ぶ。これが従来の IP で考えていない問題であり、多くの IP ソフトが対応できないと考える。6 データは、図で結果を確認でき、簡単にソフトで確認出来る。

$$\begin{aligned}
 1*b_1+3*b_2+1*b_3 &>= 1 - 1000* e_1; \\
 2*b_1+2*b_2+1*b_3 &>= 1 - 1000* e_2; \\
 2*b_1-1*b_2+1*b_3 &>= 1 - 1000* e_3; \quad \dots \dots \dots (4) \\
 -(1*b_1 + 1*b_2 - 1*b_3) &>= 1 - 1000* e_4; \\
 -(2*b_1 + 1*b_2 - 1*b_3) &>= 1 - 1000* e_5;
 \end{aligned}$$

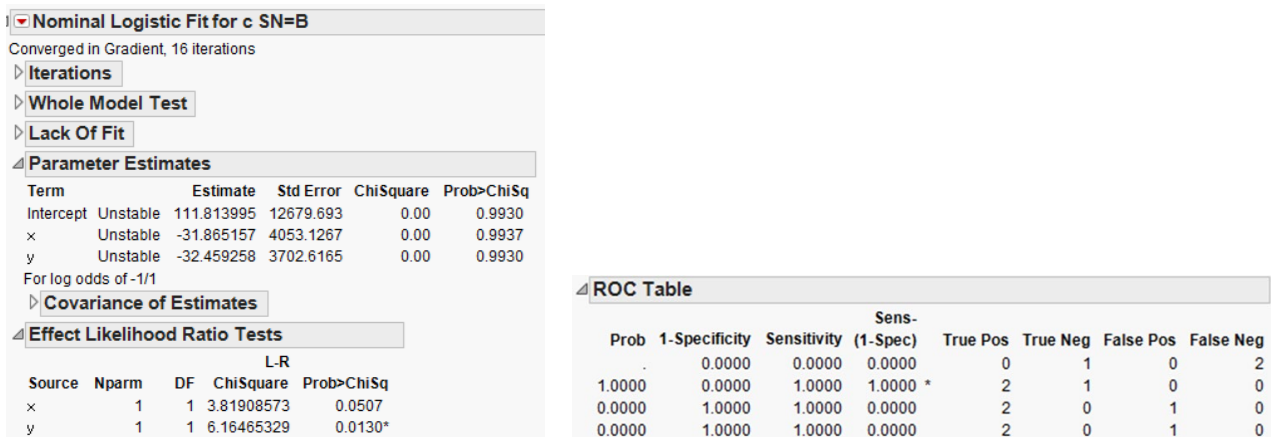


図 3 ロジスティック回帰

次のデータ B は、SN=3 を修正した LSD である。5 点をプロットし、 $RIP=2*X_1+3*x_2-7$ で LSD になることを確認してほしい。LSD を識別できないソフトは無用である。

但しデータ B の LSD の場合、図 3 のロジスティック回帰で確認する点を説明する。左のパラメータの推定値の SE は 3700 以上と大きくなり Unstable の表示が出る。これは Firth[7] が指摘している点である。このようなモデルは、統計的には棄却すべきモデルである。しかし筆者は千を超える SM で NM=0 になることを確認した。右の ROC 表で偽陽性と偽陰性の和が NM=0 を表す。以上から Fisher も関係した最尤推定法を用いて、実際の 2 群データがロジスティック曲線で分離でき、ROC 上の 2*2 分割表で偽陽性と偽陰性の和が 0 で LSD を確認することが必要である。フリーソフトなど大学の研究者が使う場合、サポートされていない可能性がある。論文の掲載料に 50 万円程度を払うのと、JMP の年間サブスクリプションを払うのと、どちらが重要か考える必要がある。勿論所属先が SAS や JMP を導入していれば幸いである。筆者は SCS では SAS の受託計算を行い自分の研究に SAS と IBM を自由に使った。成蹊大学では当

初 SAS で統計実習していたが、他のソフトを入れたいという他の教員の圧力で SAS を解約し JMP を入れた。退職後は JMP のサブスクリプションも契約したが、自費で研究費をまかなうのに耐えきれず解約し、古い永久ライセンスの JMP10 を利用している。よくサポートしていると感謝している。

4 終わりに

1971 年に大学院に落ちてきたばかりの SCS に 1 期入社した。幸いにも住友財閥の 2 代目総理事の孫の津田さんが NEC から出向して専務であった。住商で SCS の立ち上げに関与した中川企画課長に NEC に連れて行かれ、数日後に NEC の課長に大阪成人病センターの野村医長のところに連れて行かれ、「先生が自由に小僧のように使ってください」という専務の言葉を伝えた。今でいえば 2 重派遣である。しかし 2 年後から、以下の医師などと共同研究を行い、数学者にならずに本当に良かったと考える。

- 1) 野村医師らとの ECG、並びに心筋梗塞の予後判定の Medinfo77 で初の国際会議。開原東大講師と ROC を医学診断に利用した Lusted 博士が座長のセッションで発表。帰国後に開原先生の部屋のセミナーで、フラミングハム研究で用いられたロジスティック回帰等の輪読。小学校から東大まで同級生の三宅日本医大数学科教授と知り合い、「母集団と標本の ER と標本数の関係」と、「ヒューリスティック OLDF」の研究。1999 年にこれを IP で定式化した IP-OLDF で、岡山大学で論文博士。
- 2) 鈴木先生らと大阪府の癌の疫学データの研究。ROC で判別手法と医師診断を評価。
- 3) 三宅教授から、厚生省への SSM 認可の予備資料として、丸山先生の 32 万件の SSM の投与患者の予後評価を依頼され「医療情報学」に論文が採択された。その後、開原教授の下の大橋助教授が、SSM 認可の医学プロジェクトを立ち上げ特定の癌で認可された。
- 4) 2022 年に大学入試センターより、3 年間 13 教科の本試験と予備試験問題を関東圏の国立大学の現役 1 年生に実施した研究用のデータを貸与され 1 年がかりで分析した。社民党の菅政権の時代、某私立大学の問題配布ミスが社会問題になり、朝日新聞の「センター改組」のリーク記事で 3 本の投稿論文の発表取り下げ依頼があった。試験の合否判定の専門家にならずに本当に良かった考える。
- 5) 1995 年に東京都の病院の副院長から、介護保険の 1 分間タイム Study の Big データの分析手法の相談を受けて分類木を薦めた。彼は 3 ヶ月ほどで分析し、それを C コードでシステムを完成させた。国会の決めた実施時期に遅れていたため、分析に用いていないデータで検証しないで、かつ在宅患者まで実施し社会問題になった。この件と SSM の治験続行に関して、開原先生の貢献は口頭で説明する。

参考文献

- [1] Shinmura(2016) New Theory of Discriminant Analysis After R. Fisher. Springer.
- [2] Shinmura (2019) High-dimensional Microarray Data Analysis. Springer.
- [3] Shinmura(2024) The First Discriminant Theory of LSD. Springer.
- [4] Sall & Goodnight 著 (新村秀一訳) (1986) 「回帰分析の実践」朝倉書店
- [5] Linus 著(新村訳) : LINGO の解説上・中巻等
- [6] 新村秀一著 DEA による問題の発見と解決、Kindle 版の契約解除。
- [7] Firth (1993) Bias reduction of maximum likelihood estimates. Biometrika 80:27-39.
- [8] 新村秀一(2006) 改訂 IP-OLDF による SVM のアルゴリズム研究、オペレーションズ・リサーチ、51/11, 702-707.

SASによる経時的な2値データの解析方法

○伊庭 克拓、浅野 豊、黒見 真央、鈴木 聡、松嶋 優貴、毛利 誠

(大塚製薬株式会社 新薬開発本部 バイオメトリックス部 統計解析室)

Analysis Methods for Longitudinal Binary Data using SAS

Katsuhiko Iba, Yutaka Asano, Mao Kuromi, So Suzuki, Yuki Matsushima, Makoto Mouri

Office of Biostatistics, Department of Biometrics, Headquarters of Clinical Development, Otsuka Pharmaceutical Co., Ltd.

要旨

臨床試験では、治療の中止などによる欠測値を伴う経時測定データが得られることが多く、評価時点間の相関および欠測値を考慮した解析が必要とされることがある。また、臨床試験の評価項目には、治療に対する反応の有無などの2値変数がよく含まれる。SASでは、一般化推定方程式 (generalized estimating equation : GEE) , 重み付き GEE (weighted GEE : wGEE) , 多重補完法 (multiple imputation : MI) と組み合わせた GEE (MI-GEE) , pseudo 尤度に基づく一般化線形混合効果モデル (generalized linear mixed model : GLMM) など、経時的な2値データに対する周辺モデルによる解析を行うことができる。加えて、経時的な2値データに対して、マルコフ連鎖モンテカルロ (MCMC) 法で欠測値を補完し、ベイズ流の解析を行うことも可能である。本稿では、抗うつ薬の臨床試験データを事例として、SASでこれらの解析を行う方法を説明する。

キーワード：経時測定データ、2値データ、周辺モデル、GEE、GLMM、多重補完法

1. はじめに

治療法の有効性及び安全性を評価するために行われる臨床試験では、治療開始前（ベースライン）および治療開始後の複数の時点で経時的に評価を行うことが一般的であり、経時測定データが得られることが多い。経時測定データの特徴は、同一の被験者から得られた複数時点のデータに時点間の相関が生じることである。また、ほとんどの臨床試験では、治療の中止などで計画されたすべての時点での評価が行われないことによる欠測値が生じることから、解析では欠測値への対処も必要となる。

臨床試験の評価項目として、症状評価尺度のベースラインからの変化量などの連続変数や治療に対する反応の有無などの2値変数がよく用いられている。解析の主な目的は、各時点の各群の平均値や反応率（反応した被験者の割合）およびその群間差を推定することであり、特に計画された治療終了時点の群間差に関心が持たれる。各時点の各群の平均値や反応率を推定するために、経時測定データの周辺平均および周辺共分

散行列をモデル化する周辺モデル (marginal model) が用いられる。経時的な連続データの場合、周辺モデルに基づく mixed model for repeated measures (MMRM) がよく用いられており、SAS では MIXED プロシジャで実行することができる (五所・丸尾, 2017; 日本製薬工業協会, 2016)。SAS/STAT 15.3 (SAS 9.4 TSIM8) では、一般化推定方程式 (generalized estimating equation : GEE), 重み付き GEE (weighted GEE : wGEE), 多重補完法 (multiple imputation : MI) と組み合わせた GEE (MI-GEE), pseudo 尤度に基づく一般化線形混合効果モデル (generalized linear mixed model : GLMM) などで経時的な 2 値データに対する周辺モデルに基づいた解析を行うことができる。加えて、経時的な 2 値データに対して、マルコフ連鎖モンテカルロ (MCMC) 法で欠測値を補完し、ベイズ流の解析を行うことも可能である。本稿では、抗うつ薬の臨床試験データを事例として、経時的な連続データを 2 値化することによって得られた経時的な 2 値データに対して、SAS でこれらの解析を行う方法を説明する。なお、上記の解析方法は、もともと 2 値のデータに対しても用いることができる。

本稿では、Drug Information Association (DIA) Scientific Working Group on Estimands and Missing Data のウェブサイト (<https://www.lshtm.ac.uk/research/centres-projects-groups/missing-data#dia-missing-data>) で公開されている事例データ (Chapter15_example.sas7bdat) を用いた。この事例データは、抗うつ薬の有効性を評価した実際の臨床試験 (Goldstein et al., 2004) のデータに基づいており、実薬およびプラセボの 2 群に対し、ベースライン、Week 1, 2, 4, 6 の 5 時点でうつ病の重症度を評価するハミルトンうつ病評価尺度の 17 項目 (HAM-D17) 合計スコアが測定されている。本稿では、PATIENT (被験者番号), VISIT (時点 : 4, 5, 6, 7), THERAPY (群 : DRUG, PLACEBO), BASVAL (ベースライン時点の HAM-D17 合計スコア), HAMDTL17 (HAM-D17 合計スコア) の変数を解析に用いる。抗うつ薬の臨床評価方法に関するガイドライン (厚生労働省医薬食品局審査管理課長, 2010) では、主要な有効性評価に用いた症状評価尺度の経時的な合計スコアの推移に加え、副次評価項目として反応率などを評価することが推奨されている。本稿では、ガイドラインを参考に HAM-D17 合計スコアのベースラインから 50 %以上の減少を反応ありと定義し、ベースライン後の各時点の反応の有無を経時的な 2 値データとして解析する。各時点の被験者数と反応率 (反応例数/被験者数) を表 1 に示す。

以下の節の解析で用いるデータセットを準備するための SAS コードを以下に示す。導出した反応の有無の変数 RESP (1 : 反応あり, 0 : 反応なし) を含むデータセット (EXAMPLE) の一部を表 1 に示す。また、3 節、4 節および 6 節では、TRANSPOSE プロシジャを用いて同一被験者のすべての時点の HAM-D17 合計スコアが 1 オブザベーションに含まれるデータ構造に変換したデータセット (EXAMPLE_T) を用いる (Y1-Y5 : ベースライン, VISIT4, VISIT5, VISIT6, VISIT7 の HAM-D17 合計スコア)。

```
data EXAMPLE;
  set ADS.chapter15_example;
  if . < (HAMDTL17 - BASVAL) / BASVAL <= -0.5 then RESP = 1;
  if -0.5 < (HAMDTL17 - BASVAL) / BASVAL then RESP = 0;
run;

proc sort data=EXAMPLE;
  by PATIENT THERAPY BASVAL;
```

```

run;

proc transpose data=EXAMPLE out=EXAMPLE_T(rename=(BASVAL=Y1 _4=Y2 _5=Y3 _6=Y4 _7=Y5));
  var HAMDTL17;
  id VISIT;
  by PATIENT THERAPY BASVAL;
run;

```

表 1 各時点の被験者数と反応率

群	DRUG			PLACEBO		
	被験者数	反応例数	反応率(%)	被験者数	反応例数	反応率(%)
Week 1	84	7	8.3	88	5	5.7
Week 2	77	20	26.0	81	17	21.0
Week 4	73	25	34.2	76	20	26.3
Week 6	64	29	45.3	65	20	30.8

表 2 本稿で使用する事例データ (抜粋)

PATIENT	VISIT	THERAPY	BASVAL	HAMDTL17	RESP
1804	4	PLACEBO	20	14	0
1804	5	PLACEBO	20	7	1
1804	6	PLACEBO	20	7	1
1811	4	DRUG	29	18	0
1811	5	DRUG	29	12	1
1811	6	DRUG	29	6	1
1811	7	DRUG	29	3	1

2. 一般化推定方程式 (GEE)

応答変数が 2 値データの場合の回帰モデルとして、ロジスティック回帰モデルがよく用いられる (丹後ら, 2013)。ロジスティック回帰モデルは、応答変数が正規分布、二項分布、ポアソン分布などの指数型分布族に従うことを仮定した一般化線形モデル (generalized linear model: GLM) の特別な場合である (田中ら, 2008)。GLM は応答変数の各データが独立であることを仮定しており、時点間の相関を考慮することができない。GEE は経時測定データなどの相関のあるデータに対応できるように、GLM を拡張したモデルである (田中ら, 2008)。

被験者 $i (i = 1, \dots, N)$ の各時点 $1, \dots, n_i$ の応答変数を $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^T$ としたとき、GEE は周辺期待値 (本稿では各時点の反応率) $E[\mathbf{Y}_i] = \boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})^T$ を以下のようにモデル化する。

$$g(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\beta}$$

ここで、 \mathbf{X}_i は被験者 i の固定効果に対する計画行列、 $\boldsymbol{\beta}$ は固定効果パラメータベクトルである。 $g(\cdot)$ は GLM で

用いられるリンク関数である。また、周辺共分散行列は $Var[\mathbf{Y}_i] = \mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2}$ でモデル化される。 \mathbf{A}_i は応答変数の分散を期待値の関数として表す分散関数（丹後, 2019）を対角成分とする対角行列である。ベルヌーイ分布（試行回数 1 回の二項分布）の場合、分散関数は $V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$ である。 \mathbf{R}_i は \mathbf{y}_i の作業相関係数行列、 ϕ はスケールパラメータである（ベルヌーイ分布の場合は $\phi \equiv 1$ ）。 β の推定は以下の推定方程式に基づく。

$$S(\beta) = \sum_{i=1}^N \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mu_i(\beta)) = \mathbf{0}$$

ここで、 $\mathbf{D}_i = \partial \mu_i / \partial \beta$ である。この推定方程式は、GLM の対数尤度関数を β の各要素で偏微分することによって得られるスコア方程式を、 \mathbf{V}_i に相関構造を含めることで相関のあるデータに拡張したものである。 \mathbf{y}_i の分布を完全に特定しておらず、その 1 次のモーメント（平均）と 2 次のモーメント（分散）のみを特定しており、quasi 尤度法（疑似尤度と訳されることがあるが、5 節の pseudo 尤度と区別するため英語で記載する）に分類されるセミパラメトリックな方法である。GEE は作業相関構造 \mathbf{R}_i を誤特定しても、 β の推定量に一致性があるという特徴を持つ。ただし、GEE は尤度に基づく方法ではないため、欠測メカニズムが missing completely at random (MCAR) の場合のみ妥当であり、missing at random (MAR) の場合では推定値にバイアスが生じる（Mallinckrodt et al., 2020）。 \mathbf{R}_i に含まれるパラメータの推定値はモーメント法に基づいてピアソン残差から算出される（Dmitrienko and Koch, 2017; SAS Institute Inc., 2023）。 β の標準誤差 (SE) には、相関構造の誤特定に対してロバストな経験（サンドウィッチ）分散を用いるのが一般的であり、GEE プロシジャではデフォルトとなっている（SAS Institute Inc., 2023）。

HAM-D17 合計スコアに基づく反応の有無を応答変数、群、時点、ベースライン値、群と時点の交互作用、ベースライン値と時点の交互作用を固定効果とし、作業相関構造は unstructured（無構造）を仮定した。各時点の各群の最小二乗平均値（リンク関数で変換された線形予測子のスケール）をリンク関数の逆関数で変換（逆ロジット変換）し、各時点の各群の反応率とその SE を求めた。また、各時点の最小二乗平均値の群間差とその両側 95%信頼区間 (CI) を指数変換し、各時点のオッズ比とその両側 95%CI を求めた。それぞれの結果を図 1（反応率および SE）および図 2（オッズ比とその両側 95%CI）に示した。事例データを GEE で解析するための GEE プロシジャの SAS コードを以下に示す。

```
proc gee data=EXAMPLE;
  class PATIENT THERAPY VISIT;
  model RESP(event='1') = THERAPY VISIT THERAPY*VISIT BASVAL BASVAL*VISIT / dist=bin link=logit;
  repeated subject=PATIENT / type=un within=VISIT;
  lsmeans THERAPY*VISIT / ilink oddsratio diff cl;
run;
```

MODEL 文は、応答変数とモデルに含める固定効果の変数を指定し、応答変数が従う確率分布を DIST オプションで、リンク関数を LINK オプションで指定する。事例データでは応答変数がベルヌーイ分布に従うと仮定されるため、それぞれ二項分布 (DIST=BIN) およびロジット関数 (LINK=LOGIT) を指定している（確率分布ごとにデフォルトのリンク関数が決められており、二項分布の場合はロジット関数がデフォルトである）。REPEATED 文では、相関を持つ単位（クラスター）となる変数を SUBJECT オプションで指定する。

WITHIN オプションには、SUBJECT で指定した対象内での測定時点を示す変数を指定する。LSMEANS 文で群と時点の交互作用を指定することにより、各群の各時点の最小二乗平均値が計算される。ILINK オプションは、MODEL 文の LINK オプションで指定したリンク関数の逆関数で最小二乗平均値を変換した推定値を算出する。変換した推定値の SE の算出にはデルタ法（丹後, 2019）が用いられる。ODSSRATIO オプションは、ロジスティック回帰の場合に最小二乗平均値の差を指数変換し、オッズ比を算出する。なお、SAS では GENMOD プロシジャでも GEE を実行できるが、次節の wGEE には対応しておらず、デフォルトでは経験 SE が用いられないことに注意が必要である（SAS Institute Inc., 2023）。

3. 重み付き一般化推定方程式（wGEE）

wGEE は、欠測メカニズムが MAR の場合でも対応できるように、inverse probability weighting（IPW）法を用いて GEE を拡張した方法である（Mallinckrodt et al., 2020）。GEE プロシジャでは、各被験者の各時点（1 オブザベーション）に対して 1 つの重みを持つ observation specific（OS）と各被験者に対して 1 つの重みを持つ subject specific（SS）の 2 種類の重み付けを行うことができるが、OS の方がより有効な推定値を与えると報告されており（Preisser et al., 2002）、本稿では OS を用いた。wGEE は以下の推定方程式を用いる。

$$S(\boldsymbol{\beta}) = \sum_{i=1}^N \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}$$

ここで、 \mathbf{W}_i は y_{ij} が観測される確率の逆数である重み $w_{ij} = \Pr(r_{ij} | \mathbf{y}_i, \mathbf{X}_i)^{-1}$ を対角成分に持つ対角行列である。 r_{ij} は、 y_{ij} が観測された場合に 1、 y_{ij} が欠測となった場合に 0 の値を取る 2 値変数である。観測確率は、

$$\Pr(r_{ij} | \mathbf{y}_i, \mathbf{X}_i) = \lambda_{i1} \times \lambda_{i2} \times \dots \times \lambda_{ij}$$

と計算される。 $\lambda_{ij} = \Pr(r_{ij} = 1 | r_{ij-1} = 1, \mathbf{y}_i, \mathbf{X}_i)$ は、時点 $j-1$ で観測された下で、時点 j で観測される条件付き確率であり、一般的にロジスティック回帰モデルでモデル化される。wGEE は、この観測確率モデルが正しい場合に、欠測メカニズムが MAR の下で妥当である。なお、観測確率モデルの説明変数として、 $\mathbf{y}_i, \mathbf{X}_i$ 以外の変数を加えることも可能である。観測確率モデルの説明変数には、1 時点前の測定値がよく含まれている（Dmitrienko and Koch, 2017; Mallinckrodt et al., 2020; SAS Institute Inc., 2023; 駒寄ら, 2015）。

応答変数のモデルは、2 節の GEE と同じモデルを用いた。観測確率モデルには、群、時点、1 時点前の HAM-D17 合計スコア、ベースライン時点の HAM-D17 合計スコア、群と 1 時点前の HAM-D17 合計スコアの交互作用を固定効果として含んだ。最小二乗平均値に基づいて、各時点の各群の反応率および各時点のオッズ比を求め、図 1 および図 2 に示した。wGEE で推定した各時点の各群の反応率は、DRUG 群では Week 4 以降、PLACEBO 群では Week 2 以降において、GEE で推定した反応率よりも若干小さかった。

GEE プロシジャの仕様上、wGEE を実行するためには、データセットが①すべての時点の応答変数が欠測値の被験者が存在しないこと、②各被験者に応答変数が欠測値の時点も含むすべての時点のオブザベーションが存在すること、③ある時点で欠測値となった場合、それ以降のすべて時点が欠測値となる単調な欠測パターンである（非単調な欠測パターンではない）ことである（駒寄ら, 2015）。そのため、必要に応じて上記の条件を満たすようにデータセットの加工が必要となる。これらが満たされない場合に実行すると、「入力データの問題のためモデリングが中止されました」、「非単調 missingness は許容されません」といったエラーが出て実行できない。本稿の解析では、観測確率モデルに 1 時点前の HAM-D17 合計スコアを含めるための加工も必要となる。wGEE で解析できるようにするために、事例データのデータセットを加工するための SAS コードを以下に示す（データ加工の方法は例であり、他の方法も可能である）。1 節で準備した同一被験者の

すべての時点の HAM-D17 合計スコア (Y1-Y5) が 1 オブザベーションに含まれるデータセットを用いる。

```
data EXAMPLE_WGEE;
  set EXAMPLE_T;

  if nmiss(of Y2-Y5) = 0 then PATTERN = 0;
  if nmiss(of Y2-Y5) = 1 & nmiss( Y5) = 1 then PATTERN = 1;
  if nmiss(of Y2-Y5) = 2 & nmiss(of Y4-Y5) = 2 then PATTERN = 2;
  if nmiss(of Y2-Y5) = 3 & nmiss(of Y3-Y5) = 3 then PATTERN = 3;
  if PATTERN^in(0:3) then delete;

  basval = Y1;
  array YA{4} Y2-Y5;
  do VISIT = 4 to 7;
    HAMDTL17 = YA{VISIT - 3};
    if HAMDTL17 = . then RESP = .;
    if . < (HAMDTL17 - BASVAL) / BASVAL <= -0.5 then RESP = 1;
    if -0.5 < (HAMDTL17 - BASVAL) / BASVAL then RESP = 0;
    if VISIT = 4 then HAMD_PRE = 0;
    else HAMD_PRE = YA{VISIT - 4};

    output;
  end;
  drop Y1-Y5 _NAME_ _LABEL_ PATTERN;
run;
```

事例データでは、1 例 (PATIENT= 3618) だけ 1 時点で間欠な欠測値が生じており、非単調な欠測パターンとなっている。そのため、間欠な欠測値を補完するなど、何らかの対処が必要となる。本稿はプロシジャの実行方法の説明を目的としており、1 例のみと影響が小さいと考えられることから、非単調な欠測パターンの被験者を除外する (非単調な欠測パターンの被験者を除外して解析することを推奨している訳ではない)。NMISS 関数は引数の変数のうち、欠損値を持つ変数の数を返す。例えば、NMISS(of Y2-Y5)=0 は、Y2-Y5 の 4 変数のうち、欠損値が 1 つもないことを意味する。単調な欠測パターンでは、4 変数のうち欠損値を持つ変数の数を M とすると、4 変数の後ろから M 個の変数がすべて欠測値となっている。上記の SAS コードでは、それ以外のパターンとなる被験者のオブザベーションを削除する。そして、VISIT ごとに反応の有無 (RESP) および 1 時点前の HAM-D17 合計スコア (HAMD_PRE) を導出し、OUTPUT 文で各 VISIT のオブザベーションを発生させ、TRANSPOSE 前のデータ構造に戻す。なお、最初の時点 (VISIT=4) の HAMD_PRE は任意の数値 (ここでは 0) を格納することでよく、どのような値でも解析結果に影響しない (SAS Institute Inc., 2023; 駒寄ら, 2015)。加工後のデータセットの抜粋 (表 2 と同じ被験者) を表 3 に示す。加工後の事例データを wGEE で解析するための GEE プロシジャの SAS コードを以下に示す。

```

proc gee data=EXAMPLE_WGEE;
  class PATIENT THERAPY VISIT;
  missmodel THERAPY VISIT BASVAL HAMD_PRE VISIT*HAMD_PRE / type=obslevel;
  model RESP(event='1') = THERAPY VISIT THERAPY*VISIT BASVAL BASVAL*VISIT / dist=bin link=logit;
  repeated subject=PATIENT / type=un within=VISIT;
  lsmeans THERAPY*VISIT / ilink oddsratio diff cl;
run;

```

wGEE では、MODEL 文の応答変数のモデルの指定に加え、MISSMODEL 文で観測確率モデルに含める変数を指定する。MISSMODEL 文のオプションで TYPE=OBSLEVEL を指定することで、OS の重みを使用することができる。MISSMODEL 文以外の指定は 2 節と同じである。GEE プロシジャの出力には、応答変数のモデルのパラメータ推定値に加え、観測確率モデルのパラメータ推定値も表示される。 β の SE として、観測確率モデルの推定も考慮した経験 SE が算出される (Preisser et al., 2002; SAS Institute Inc., 2023)。

表 3 wGEE 用に加工したデータ (抜粋)

PATIENT	VISIT	THERAPY	BASVAL	HAMDTL17	RESP	HAMD_PRE
1804	4	PLACEBO	20	14	0	0 ^a
1804	5	PLACEBO	20	7	1	14
1804	6	PLACEBO	20	7	1	7
1804	7	PLACEBO	20	.	.	7
1811	4	DRUG	29	18	0	0 ^a
1811	5	DRUG	29	12	1	18
1811	6	DRUG	29	6	1	12
1811	7	DRUG	29	3	1	6

a 任意の数値 (ここでは 0) を格納することでよく、どのような値でも解析結果に影響しない。

4. 多重補完法と組み合わせた一般化推定方程式 (MI-GEE)

MI は、欠測値を複数回補完した疑似完全データをそれぞれ解析し、得られたパラメータ推定値を統合して最終的な推定値を得ることによって、欠測値の補完に対する不確実性を考慮する方法である (野間, 2017)。欠測メカニズムが MAR の場合、MAR を仮定した MI に基づく解析は妥当である。ただし、補完モデルを誤特定した場合、推定値にバイアスが生じることがある (Liu et al., 2011)。本稿の事例データのような、経時的な連続データを 2 値化することによって得られた経時的な 2 値データの場合、連続データを用いて補完を行ってから 2 値化の方が、より多くの情報を補完に用いることができる (Dmitrienko and Koch, 2017; Li et al., 2022)。そのため、本稿では HAM-D17 合計スコアに対して補完を行った後、反応の有無に 2 値化して解析を行った。なお、最初から 2 値データの場合でも、2 値変数に対する補完モデルを用いて解析することが可能である (野間, 2017)。MI での補完方法として、単調な欠測パターンのデータに対して、補完対象の変数を応答変数、補完に用いる変数を説明変数とした回帰モデルで欠測値を順次補完する単調回帰

法を用いた（日本製薬工業協会, 2016）。 y_{ij} を被験者*i*の*j*時点目（ $j = 1, \dots, 5$ ）のHAM-D17 合計スコアの測定値とする。ベースラインには欠測値がないものとし、それ以降の各時点（ $j = 2, \dots, 5$ ）の欠測値の補完に、補完対象の時点の測定値を応答変数、それ以前の時点の測定値を説明変数とした以下の回帰モデルを用いて欠測値を補完する。

$$y_{ij} \sim N(\mu_{ij}, \sigma_j^2), \quad \mu_{ij} = \beta_{j0} + \beta_{j1}y_{i1} + \dots + \beta_{j(j-1)}y_{i(j-1)}$$

ここで、 β の1つ目の添え字は応答変数（補完対象）の時点、2つ目の添え字（切片以外）は説明変数の時点を表している。非単調な欠測パターンのデータの場合、MCMC法で間欠な欠測値を補完し、単調な欠測パターンにしてから単調回帰法を行うことや、fully conditional specification（FCS）法を用いることができる（Dmitrienko and Koch, 2017）。本稿では、DRUG群とPLACEBO群で別々に、MCMC法で間欠な欠測値を補完してから単調回帰法を行った。MIプロシジャで欠測値を補完したデータセットを作成するSASコードを以下に示す。3節と同様に、1節で準備した同一被験者のすべての時点のHAM-D17合計スコア（Y1-Y5）が1オブザベーションに含まれるデータセットを用いる。

```
proc sort data=EXAMPLE_T;
  by THERAPY;
run;

proc mi data=EXAMPLE_T out=EXAMPLE_MIGEE1 nimpute=100 seed=12345;
  var Y1-Y5;
  mcmc impute=monotone;
  by THERAPY;
run;

proc sort data=EXAMPLE_MIGEE1;
  by _Imputation_ THERAPY;
run;

proc mi data=EXAMPLE_MIGEE1 out=EXAMPLE_MIGEE2 nimpute=1 seed=12345;
  var Y1-Y5;
  monotone reg(Y2-Y5 /details);
  by _Imputation_ THERAPY;
run;

data EXAMPLE_MIGEE3;
  set EXAMPLE_MIGEE2;

  basval = Y1;
  array YA{4} Y2-Y5;
```

```

do VISIT = 4 to 7;
  HAMDTL17 = YA{VISIT - 3};
  if . < (HAMDTL17 - basval) / basval <= -0.5 then RESP = 1;
  if -0.5 < (HAMDTL17 - basval) / basval then RESP = 0;
  output;
end;
drop Y1-Y5 _NAME_ _LABEL_;
run;

```

群別に補完を行うために、データセットを事前に群 (THERAPY) でソートする。PROC MI 文の NIMPUTE オプションで補完回数を指定することができ、本稿では 100 回とした。また、SEED オプションで疑似乱数のシード値を指定することができる。VAR 文で解析に用いる変数を指定する。MCMC 法で間欠な欠測値を補完する場合、MCMC 文のオプションで IMPUTE=MONOTONE を指定する。MCMC 法で間欠な欠測値を補完して得られたデータセットを、補完回数 (_Imputation_) および群 (THERAPY) でソートし、補完回数および群ごとに単調回帰法を行う。1 つ目の MI プロシジャで 100 個のデータセットが作成されているため、2 回目の MI プロシジャでは補完回数は 1 回で良い (NIMPUTE=1)。単調回帰法を行う場合、MONOTONE 文で REG を指定する。REG の後の括弧内に補完対象の変数を指定する。DETAILS オプションを指定することで、補完モデルの詳細が得られる。欠測値を補完した後、3 節のデータ加工と同様に、VISIT ごとに反応の有無の変数 (RESP) を導出し、OUTPUT 文で各 VISIT のオブザベーションを発生させ、TRANSPOSE 前のデータ構造に戻す。

GEE による解析は 2 節と同じモデルを用いた。各補完したデータセットの解析によって得られた最小二乗平均値および最小二乗平均値の群間差を Rubin のルール (野間, 2017) で統合し、統合した最小二乗平均値および最小二乗平均値の群間差に基づいて、各時点の各群の反応率 (図 1) および各時点のオッズ比 (図 2) を求めた。MI-GEE で推定した各時点の各群の反応率は、wGEE で推定した反応率に近かった。欠測値を補完したデータセットを GEE プロシジャで解析し、得られた推定値を MIANALYZE プロシジャで統合する SAS コードを以下に示す。

```

proc sort data=EXAMPLE_MIGEE3;
  by _Imputation_;
run;

proc gee data=EXAMPLE_MIGEE3;
  class PATIENT THERAPY VISIT;
  model RESP(event='1') = THERAPY VISIT THERAPY*VISIT BASVAL BASVAL*VISIT / dist=bin link=logit;
  repeated subject=PATIENT / type=un within=VISIT;
  lsmeans THERAPY*VISIT / ilink oddsratio diff cl;
  ods output lsmeans=MIGEE_LSMEANS diffs=MIGEE_DIFFS(where=(VISIT=_VISIT and THERAPY='DRUG'));
  by _Imputation_;

```

```

run;

proc sort data=MIGEE_LSMEANS;
  by VISIT THERAPY;
run;

proc sort data=MIGEE_DIFFS;
  by VISIT;
run;

proc mianalyze data=MIGEE_LSMEANS;
  modeleffects Estimate;
  stderr StdErr;
  ods output ParameterEstimates=MIGEE_LSMEANS_POOL;
  by VISIT THERAPY;
run;

data MIGEE_LSMEANS_POOL;
  set MIGEE_LSMEANS_POOL;
  RESP_POOL = 1 / (1 + exp(-Estimate));
  SE_RESP_POOL = exp(-Estimate) * StdErr / (1 + exp(-Estimate))**2;
run;

proc print data=MIGEE_LSMEANS_POOL;
  var VISIT THERAPY RESP_POOL SE_RESP_POOL;
run;

proc mianalyze data=MIGEE_DIFFS;
  modeleffects Estimate;
  stderr StdErr;
  ods output ParameterEstimates=MIGEE_DIFFS_POOL;
  by VISIT;
run;

data MIGEE_DIFFS_POOL;
  set MIGEE_DIFFS_POOL;
  OR = exp(Estimate);
  LowerOR = exp(LCLMean);
  UpperOR = exp(UCLMean);

```

```
run;

proc print data=MIGEE_DIFFS_POOL;
  var VISIT OR LowerOR UpperOR;
run;
```

補完回数ごとに解析を行うため、データセットを事前に補完回数の変数（_Imputation_）でソートする。BY 文で補完回数（_Imputation_）を指定し、補完回数ごとに GEE による解析を行って推定値を得る。ODS OUTPUT 文で、各群の各時点の最小二乗平均値が格納される LSMEANS、各時点での群間差の推定値が格納される DIFFS を指定し、結果をデータセットに出力する。LSMEANS のデータセットを時点（VISIT）および群（THERAPY）で、DIFFS のデータセットを時点（VISIT）でソートする。MIANALYZE プロシジャでは、MODELEFFECTS 文で推定値の変数、STDERR 文で SE の変数を指定する。統合した推定値から、反応率やオッズ比を計算するために、ODS OUTPUT 文で PARAMETERESTIMATES を指定し、結果をデータセットに出力する。最後に、データステップで統合された最小二乗平均値の推定値を逆ロジット変換し、反応率の推定値（RESP_POOL）を算出する。反応率の SE（SE_RESP_POOL）は、GEE プロシジャの ILINK オプションと同様にデルタ法で算出できる。また、統合された群間差の推定値およびその両側 95%CI を指数変換することで、オッズ比（OR）とその両側 95%CI（LowerOR および UpperOR）を算出する。

5. Pseudo 尤度に基づく一般化線形混合効果モデル（GLMM）

GLMM は、被験者 i の変量効果に対する計画行列を \mathbf{Z}_i 、変量効果パラメータベクトルを $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$ としたとき、条件付き期待値を以下のようにモデル化する。

$$g(E[\mathbf{Y}_i | \mathbf{b}_i]) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i$$

また、条件付き分散は、 $\text{Var}[\mathbf{Y}_i | \mathbf{b}_i] = \mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2}$ でモデル化される。 \mathbf{A}_i は分散関数を対角成分とする対角行列、 \mathbf{R}_i は共分散行列である。GLMM は GLM の拡張であり、リンク関数を用いることで、応答変数の分布が指数型分布族に含まれる分布に従うことを仮定できる混合効果モデルである (Dmitrienko and Koch, 2017)。モデルに含まれるパラメータは、変量効果を周辺化した周辺尤度を最大化することにより推定される。PROC GLIMMIX 文の METHOD オプションで指定可能な推定方法として、テイラー展開によるモデルの線形近似に基づく pseudo 尤度法 (RSPL, RMPL, MSPL, MMPL)、積分近似に基づく適応型ガウス-エルミート求積法 (QUAD) や Laplace 法 (LAPLACE) が実装されている (SAS institute Inc., 2023)。GLIMMIX プロシジャのデフォルトの推定方法である RSPL では、GLM の反復重み付き最小二乗法 (田中ら, 2008) と同様に、応答変数の期待値 $\boldsymbol{\mu}_i$ のテイラー展開による線形近似を行うことによって、pseudo データ \mathbf{P}_i に対する線形混合モデル

$$\mathbf{P}_i = \frac{\partial g(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i} (\mathbf{y}_i - \boldsymbol{\mu}_i) + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i \approx \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

を考え、制限付き最尤法によってパラメータを推定する。さらに、MMRM と同様に変量効果の項 $\mathbf{Z}_i \mathbf{b}_i$ を明示的にモデル化せず、誤差項 $\boldsymbol{\varepsilon}_i$ の共分散構造を規定することによって、周辺モデルによる解析 (pseudo データに対する MMRM 解析) を行うことができる。パラメータ推定では、pseudo データの更新とモデルの当ては

めによる二重の反復が必要となる。Pseudo 尤度に基づく GLMM は、GEE と同様に欠測メカニズムに MCAR を仮定した方法であるが、MAR の下でも適切な共分散構造が用いられた場合には、GEE よりも β の推定値のバイアスが小さいと報告されている (O’Kelly and Ratitch, 2014; Liu et al., 2011; Li et al., 2022)。

これまでの節の GEE による解析と同様に、HAM-D17 合計スコアに基づく反応の有無を応答変数、群、時点、ベースライン値、群と時点の交互作用、ベースライン値と時点の交互作用を固定効果とし、誤差共分散構造は unstructured を仮定した。また、最小二乗平均値とその群間差に基づいて、反応率 (図 1) とオッズ比 (図 2) を求めた。その際に用いる β の SE と t 統計量の自由度は Kenward-Roger 法 (五所・丸尾, 2017) で求めた。なお、PROC GLIMMIX 文の EMPIRICAL オプションを指定することで、各種の経験 SE を用いることも可能である (SAS institute Inc., 2023)。GLMM で推定した各時点の各群の反応率は、GEE で推定した反応率に近かった。事例データを pseudo 尤度に基づく GLMM で解析するための SAS コードを以下に示す。

```
proc glimmix data=EXAMPLE maxopt=100;
  class PATIENT THERAPY VISIT;
  model RESP(event='1') = THERAPY VISIT THERAPY*VISIT BASVAL BASVAL*VISIT / dist=binary link=logit ddfm=kr;
  random VISIT / type=un subject=PATIENT rside;
  lsmeans THERAPY*VISIT / ilink oddsratio diff cl;
run ;
```

GLIMMIX プロシジャの構文は GEE プロシジャと類似している。PROC GLIMMIX 文の MAXOPT オプションで反復計算の最大回数を指定できる。MODEL 文では、GEE プロシジャと同様に応答変数の分布 (DIST) とリンク関数 (LINK) を指定する。DDFM オプションで、検定や区間推定に用いる自由度の計算方法を指定する。Kenward-Roger 法の場合は、DDFM=KR と指定する。GLIMMIX プロシジャには REPEATED 文がなく、変量効果と誤差項に関する指定は両方とも RANDOM 文を用いる。MIXED プロシジャで MMRM 解析を行う際の REPEATED 文と同等の指定を行うには、RSIDE オプションを指定する。LSMEANS 文とそのオプションに関しては、GEE プロシジャと同じである。

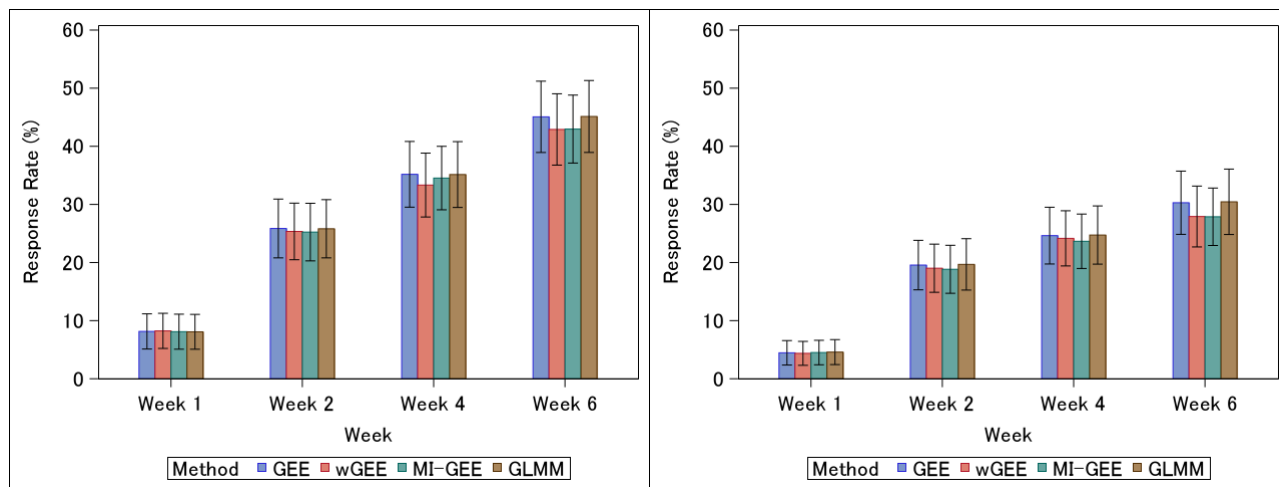


図 1 各解析方法の各時点の反応率および SE (左 : DRUG, 右 : PLACEBO)

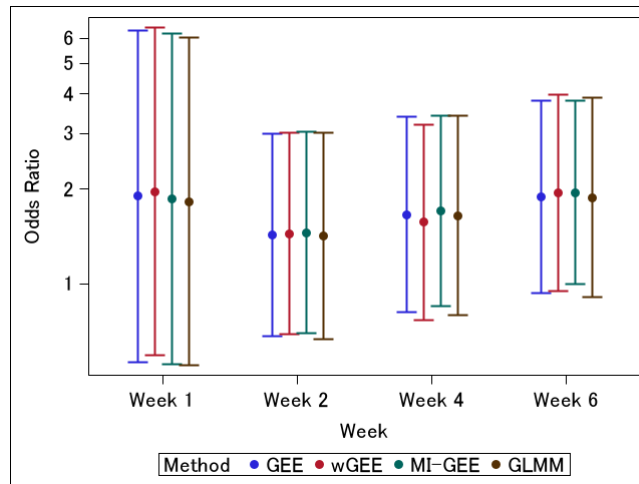


図2 各解析方法の各時点のオッズ比 (DRUG/ PLACEBO) および両側 95%CI

6. ベイズ流の欠測値補完による解析

本節では、MCMC プロシジャを用いて、4 節の MI で用いた単調回帰法と同様の回帰モデルに基づいて、HAM-D17 合計スコアの欠測値を補完した後に、主な関心がある治療終了時点 (Week 6) の反応の有無を導出し、ベイズ流のロジスティック回帰を行う方法を説明する (他の時点についても解析可能である)。なお、MCMC プロシジャで HAM-D17 合計スコアの欠測値を補完する部分は、伊庭ら (2023) の 6 節「回帰モデルに基づく欠測値の補完」と同じであり、本節の内容は、それを 2 値データの解析に応用したものである。MCMC プロシジャの概要や SAS コードの説明は、伊庭ら (2023) も参考にされたい。

4 節で示した回帰モデルに基づいて、群ごとに各時点の HAM-D17 合計スコアの欠測値を補完した (ただし、本節では時点 j の応答変数の分散 σ_j^2 は群で共通とした)。欠測値を補完した後、主な関心がある治療終了時点 (Week 6) の反応の有無 $resp_i$ に対して、以下のロジスティック回帰モデルで解析を行った。

$$resp_i \sim \text{Bernoulli}(p_i), \quad p_i = 1 / \exp(-(\alpha + \delta trt_i + \varphi y_{i1}))$$

ここで、 $trt_i = \{0: \text{PLACEBO}, 1: \text{DRUG}\}$ は群の指示変数、 α は切片、 δ は群、 φ は HAM-D17 合計スコアのベースライン値 y_{i1} に関するパラメータであり、これらのパラメータの事前分布には無情報の正規分布 $N(0, 10^4)$ を用いた。また、欠測値の補完に用いる回帰モデルに含まれる各 β の事前分布には無情報の正規分布 $N(0, 10^4)$ 、各 σ^2 の事前分布には形状パラメータおよびスケールパラメータを 0.01 とした無情報の逆ガンマ分布 $IG(0.01, 0.01)$ を用いた。欠測値の補完に用いる回帰モデルおよび反応率を推定するためのロジスティック回帰モデルにおけるパラメータおよび y_{ij} の欠測値 (パラメータとみなされる) は、単一の MCMC によってサンプリングされる。

SAS コードを以下に示す。回帰モデルに基づいて欠測値を補完する部分については、伊庭ら (2023) と同じコードである。本節でも、1 節で準備した転置した HAM-D17 合計スコアのデータセットを用いる。なお、もともと 2 値変数の場合でも、補完に用いるモデルを変更すれば同様に解析できる (SAS Institute Inc., 2023)。ロジスティック回帰モデルの応答変数は 2 値データのため、対数尤度関数を指定する部分では、ベルヌーイ分布の確率関数の対数変換値を返す LPDFBERN(X, P) 関数を用いている。HAM-D17 合計スコアのベースライン値が DRUG 群と PLACEBO 群を併せた全被験者の平均値 (17.9) の場合の最小二乗平均値に相当するパラメータを算出し、それを逆ロジット変換することで、各群の反応率を求めている。MCMC 法に

基づくベイズ流の解析では、これまでの節のようにデルタ法を用いる必要はなく、反応率の事後分布から事後標準偏差が直接得られる。また、 δ のMCMCサンプルを指数変換することで、オッズ比の事後分布を得ることができ、事後分布に基づいて95%最高事後密度（HPD）信用区間を求めることも容易である。

```
proc mcmc data=EXAMPLE_T seed=123456 nbi=100000 nmc=500000 thin=100 outpost=POST monitor=(_parms_ lsm1
lsm2 p1 p2 OR) diag=all plot(smooth)=all;
  parms beta_p20 beta_p21 beta_t20 beta_t21 sigma2;
  parms beta_p30 beta_p31 beta_p32 beta_t30 beta_t31 beta_t32 sigma3;
  parms beta_p40 beta_p41 beta_p42 beta_p43 beta_t40 beta_t41 beta_t42 beta_t43 sigma4;
  parms beta_p50 beta_p51 beta_p52 beta_p53 beta_p54 beta_t50 beta_t51 beta_t52 beta_t53 beta_t54 sigma5;
  parms alpha delta phi;

  prior beta: alpha delta phi ~ normal(0, var=10000);
  prior sigma: ~ igamma(shape=0.01, scale=0.01);

  if THERAPY='PLACEBO' then mu2 = beta_p20 + beta_p21 * Y1;
                        else mu2 = beta_t20 + beta_t21 * Y1;
  if THERAPY='PLACEBO' then mu3 = beta_p30 + beta_p31 * Y1 + beta_p32 * Y2;
                        else mu3 = beta_t30 + beta_t31 * Y1 + beta_t32 * Y2;
  if THERAPY='PLACEBO' then mu4 = beta_p40 + beta_p41 * Y1 + beta_p42 * Y2 + beta_p43 * Y3;
                        else mu4 = beta_t40 + beta_t41 * Y1 + beta_t42 * Y2 + beta_t43 * Y3;
  if THERAPY='PLACEBO' then mu5 = beta_p50 + beta_p51 * Y1 + beta_p52 * Y2 + beta_p53 * Y3 + beta_p54 * Y4;
                        else mu5 = beta_t50 + beta_t51 * Y1 + beta_t52 * Y2 + beta_t53 * Y3 + beta_t54 * Y4;
  if THERAPY='PLACEBO' then mu = alpha + phi * Y1;
                        else mu = alpha + delta + phi * Y1;

  p = 1 / (1 + exp(-mu));

beginnodata;
  lsm1 = alpha + phi * 17.9;
  lsm2 = alpha + delta + phi * 17.9;
  p1 = 1 / (1 + exp(-lsm1));
  p2 = 1 / (1 + exp(-lsm2));
  OR = exp(delta);
endnodata;

model Y2 ~ normal(mu2, var=sigma2);
model Y3 ~ normal(mu3, var=sigma3);
model Y4 ~ normal(mu4, var=sigma4);
```

```

model Y5 ~ normal(mu5, var=sigma5);

change = Y5 - Y1;
if . < change / Y1 <= -0.5 then RESP = 1;
if -0.5 < change / Y1 then RESP = 0;

ll = lpdfbern(RESP, p);
model general(ll);
run;

```

ロジスティック回帰モデルで得られた各群の反応率の調整済み事後平均およびオッズ比を表4に示す。本節の解析でも、これまでの節と一貫した結果が得られた。なお、ベイズ流の解析では事後確率に基づいたオッズ比以外の比較指標も柔軟に用いることができ、例えば、DRUG群の反応率がPLACEBO群の反応率を上回る事後確率は98.2%であった。また、本稿では無情報事前分布を用いたが、ベイズ流アプローチの特徴として、治療効果などに関する既存情報から事前分布を構成し、解析で考慮することも可能である。

表4 欠測値を補完したロジスティック回帰分析の反応率 (%) およびオッズ比

群	反応率 (DRUG)		反応率 (PLACEBO)		オッズ比	95%HPD 区間	
	調整済み事後平均	事後標準偏差	調整済み事後平均	事後標準偏差		下限	上限
時点 Week 6	42.0	5.8	25.6	4.9	2.27	0.86	3.92

7. まとめ

本稿では、SASで実行できる経時的な2値データの解析方法として、GEE, wGEE, MI-GEE および pseudo 尤度に基づく GLMM による解析を行う方法を説明した。加えて、経時的な2値データに対して、MCMC法で欠測値を補完し、ベイズ流の解析を行う方法も説明した。これらの解析方法は、いずれも経時的な2値データの反応率を推定するために用いることができるが、欠測メカニズムの仮定、解析可能な欠測パターン、観測確率モデルや補完モデルといった追加のモデリングの必要性、計算コスト、事前情報を考慮できるかなどの違いがある。実践においては、解析の目的、欠測値の発生状況などの解析対象となるデータの特徴を踏まえて、どの解析方法を用いるか検討する必要がある。なお、本稿で取り上げた解析方法は、欠測メカニズムが missing not at random (MNAR) の場合にはいずれも妥当ではない。欠測メカニズムをはじめとしたモデルの仮定からのずれに対する頑健性を、感度分析で検討することも重要である。

本稿が SAS で経時的な2値データを解析する際の参考になると幸いである。

参考文献

Drug Information Association Scientific Working Group on Estimands and Missing Data.
<https://www.lshtm.ac.uk/research/centres-projects-groups/missing-data#dia-missing-data> [2024/8/13 アクセス]

- Dmitrienko A, Koch G (2017). *Analysis of Clinical Trials Using SAS (Second Edition)*. SAS Press: Cary, NC, USA.
- Goldstein DJ, Lu Y, Detke MJ, Wiltse C, Mallinckrodt C, Demitrack MA (2004). Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine. *J Clin Psychopharmacol*. 24(4):389-399.
- Li Y, Feng D, Sui Y, Li H, Song Y, Zhan T, Cicconetti G, Jin M, Wang H, Chan I, Wang X (2022). Analyzing longitudinal binary data in clinical studies. *Contemporary Clinical Trials* 2022 Apr:115:106717.
- Liu GF, Zhan, X (2011). Comparisons of methods for analysis of repeated binary responses with missing data. *Journal of Biopharmaceutical Statistics* 21:371-392.
- Mallinckrodt C, Molenberghs G, Lipkovich I, Ratitch B (2020). *Estimands, Estimators and Sensitivity Analysis in Clinical Trials*. Chapman and Hall/CRC.
- O’Kelly M, Ratitch B (2014). *Clinical Trials with Missing Data -A Guide for practitioners-*. John Wiley & Sons.
- Preisser JS, Lohman KK, Rathouz PJ (2002). Performance of Weighted Estimating Equations for Longitudinal Binary Data with Drop-Outs Missing at Random. *Statistics in Medicine*, 21:3035-3054.
- SAS Institute Inc. (2023). *SAS/STAT 15.3 User’s Guide*. SAS Institute Inc., Cary, NC, USA.
- 伊庭 克拓, 浅野 豊, 松嶋 優貴, 毛利 誠 (2023). BGLIMM プロシジャおよび MCMC プロシジャによるベイズ流経時測定データ解析. SAS ユーザー総会論文集 2023.
- 厚生労働省医薬食品局審査管理課長 (2010). 抗うつ薬の臨床評価方法に関するガイドライン. <https://www.pmda.go.jp/files/000208191.pdf> [2024/8/13 アクセス]
- 五所 正彦, 丸尾 和司 (2017). 経時測定データ解析における mixed-effects models for repeated measures (MMRM) の利用. *応用統計学* 46(2):53-65.
- 駒寄 弘, 土居 正明, 横山 雄一, 鵜飼 裕之, 藤原 正和 (2015). 【企画セッション】欠測のあるデータにおける主解析の検討(3)Proc GEEによるwGEE法を用いた連続量経時データの解析. SASユーザー総会論文集 2015.
- 田中 豊, 森川 敏彦, 山中 竹春, 富田 誠 (訳), Dobson AJ (著) (2008). *一般化線形モデル入門*. 共立出版.
- 丹後 俊郎 (2019). *新版 統計モデル入門*. 朝倉書店.
- 丹後 俊郎, 山岡 和枝, 高木 晴良 (2013). *新版 ロジスティック回帰分析*. 朝倉書店.
- 日本製薬工業協会 (2016). 欠測のある連続量経時データに対する統計手法について Ver2.0. <https://www.jpma.or.jp/information/evaluation/results/allotment/lofurc0000007qqq-att/statistics01.pdf> [2024/8/13 アクセス]
- 野間 久史 (2017). 連鎖方程式による多重代入法. *応用統計学* 46(2):67-86

RTF Output Tips

- SAS初学者もSAS熟練者もたのしめるものを目指して -

○麻生 美由紀¹

(¹株式会社インテリム)

Practical Tips for RTF Output

- For Having RTF Output for ALL SAS Users to be Enjoyable -

Miyuki Aso

intellim Corporation

要旨

本論文では、SAS 初学者や RTF 出力を改めて学び直そうと考えている方に向けて、SAS での RTF 出力に関する学びのきっかけを掴んでもらうことを目的とし、基礎編・応用編に分けて RTF 出力に関する Tips を紹介する。まず、基礎編 (Basic Tips) として、RTF 出力に必要な要素として厳選した内容を実際の SAS プログラムを用いて解説する。具体的には、TEMPLATE プロシジャでのスタイルテンプレートの作成例、RTF 出力に関するオプションの指定例、ODS RTF ステートメントの使用例、REPORT プロシジャ及び ODS TABLE プロシジャの使用例を解説する。さらに、応用編 (Advanced Tips) として、日本語と英語が混在する RTF の帳票作成時に役立つ全角・半角フォントの制御方法について、筆者が SAS ユーザー総会 2018 で発表した内容の最新情報を紹介する。また、ODS RTF ステートメントとは異なる性質を持つ ODS TAGSETS.RTF ステートメントの使用例や RTF タグの活用事例を紹介する。本論文が SAS プログラマ間のさらなる意見交換や技術発展の一助になれば幸いである。

キーワード : RTF 出力, RTF Output, TEMPLATE プロシジャ, Style Template, スタイルテンプレート, ODS RTF ステートメント, ODS TAGSETS.RTF ステートメント, REPORT プロシジャ, ODS TABLE プロシジャ, フォント制御

1 はじめに

1.1 本論文の目的

本論文は SAS 初学者や RTF 出力の経験がない・少ない方或いは改めて学び直そうと考えている方に向けて、SAS での RTF 出力について学びのきっかけを掴んでもらうことを目的とし、基礎編・応用編に分けて RTF 出力に関する Tips を紹介する。

基礎編 (Basic Tips) では、実際に運用可能な SAS プログラムの例 (添付資料: ODS_RTF_sample_basic.sas) を通じて、RTF 出力に必要な要素を厳選して紹介する。

応用編 (Advanced Tips) では、SAS プログラムの例 (添付資料: ODS_RTF_sample_advanced.sas など) を通じて、RTF 出力での全角・半角フォントの制御方法、ODS TAGSETS.RTF ステートメントの使用例、RTF タグの活用事例など、RTF 出力に関する最新の技術や Tips を紹介する。

RTF 出力に関してより一層深い知識を得たい・プロシジャやステートメントの機能をもっと使ってみたいという方は、過去の SAS ユーザー総会の資料や SAS 社のマニュアルなどを活用いただきたい。

1.2 本論文で参照するプログラムと使用する SAS のバージョン

基礎編 (Basic Tips) では、ODS_RTF_sample_basic.sas を上から順に解説する。

応用編 (Advanced Tips) では、ODS_RTF_sample_advanced.sas 及び ODS_TAGSETS.RTF_sample.sas を参照する。RTF タグ活用事例は紹介に留める。

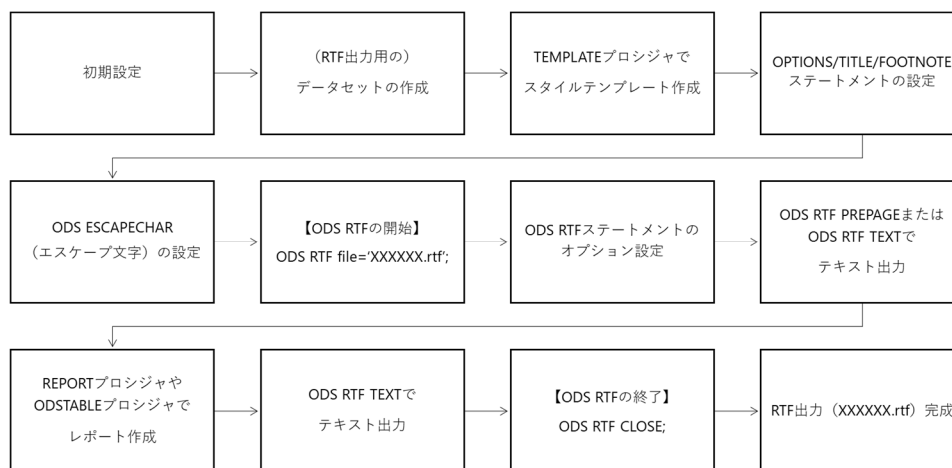
また、本論文で使った SAS のバージョンは SAS 9.4M7 (TS1M7) である。

2 Basic Tips

本章では、RTF 出力の基礎を ODS_RTF_sample_basic.sas の SAS プログラムコードを通じて紹介する。ここでは RTF 出力に必要な要素の説明に留め、各ステートメントや各プロシジャの詳細な説明は割愛する。基礎編のプログラムを参考に、自らプログラム作成する、プログラムの挙動を確認する、分からない部分は SAS 社のマニュアル [1] や論文を検索・参照するなどして、各ステートメントや各プロシジャの機能を理解していただきたい。

ここで、SAS における RTF 出力までの基本的な流れを図 2-1 に示す。

図 2-1 SAS における RTF 出力までの基本的な流れ



RTF 出力の基本的な流れを一通り経験する目的で、RTF 出力までの全工程を 1 つのプログラムに含めたものを紹介する。なお、ODS_RTF_sample_basic.sas は学習用のサンプルプログラムであり、実業務での運用は想定していない。このプログラムはあくまで参考として、各自の所属の会社・組織の方針に従い、各自で適切なプログラムを作成していただきたい。

2.1 TEMPLATE プロシジャを用いたスタイルテンプレートの作成例

TEMPLATE プロシジャで作成したスタイルテンプレートを ODS RTF ステートメントの STYLE オプショ

ンにおいて指定することで、フォントの設定や文字色の設定、罫線の設定、余白の設定などの RTF 出力の見た目を制御することができる [2]。ODS RTF ステートメントの STYLE オプションを指定しなくても RTF 出力は可能だが、大抵の場合、見た目の微調整が必要である。スタイルテンプレートを作成しておくことで、出力の全体的な見た目を制御できるため、各レポート作成用のプログラムにおいてスタイルを細かく指定する必要がなくなり、大変便利である。

まず、ODS PATH ステートメントでのスタイルテンプレートの保存先・参照先の変更について解説する。

CODE 1: ODS PATH Statement

```
ods path reset;  
ods path(remove) sasuser.templat;  
ods path(prepend) work.templat(update);
```

通常、新しく作成するテンプレートは SASUSER.TEMPLAT に保存される。CODE 1 では、ODS PATH ステートメントによってスタイルテンプレートの保存先を WORK.TEMPLAT (WORK ライブラリ内のスタイルテンプレートの保存先) に変更している。現在の ODS PATH は、[ods path show;]のコードを実行すると確認可能で、SAS のログ画面に図 2-2 のように表示される。

図 2-2 CODE 1 を実行した際の ODS PATH (左) とデフォルトの ODS PATH (右) を示すログ画面

<pre>ログ - (無題) 323 ods path show; 現在の ODS PATH リスト : 1. WORK.TEMPLAT(UPDATE) 2. SASHELP.TMPLMST(READ)</pre>	<pre>ログ - (無題) 327 ods path reset; 328 ods path show; 現在の ODS PATH リスト : 1. SASUSER.TEMPLAT(UPDATE) 2. SASHELP.TMPLMST(READ)</pre>
---	--

次に、ODS_RTF_sample_basic.sas における TEMPLATE プロシジャの使用例を解説する。

CODE 2: TEMPLATE Procedure

```
proc template;  
  define style Myrtftemp;  
    parent=styles.RTF;  
    style systemtitle/  
      font_face = "Times New Roman"  
      font_size = 9pt  
      font_weight = medium  
      font_style = roman  
      foreground = black  
      background= white;  
  (中略)  
  end;  
run;
```

CODE2 では、TEMPLATE プロシジャの DEFINE STYLE ステートメントで、新たにスタイルテンプレート

Myrtftemp を作成している。作成した Myrtftemp は、上述の CODE 1 の ODS PATH の設定によって、WORK.TEMPLAT 内に保存される。

スタイルテンプレートを新規に定義するのは難しいため、ここでは PARENT=ステートメントで、SASHELP.TMPLMST 内に保存されている styles.RTF を親スタイルテンプレートに指定して Myrtftemp を作成している。DEFINE STYLE ステートメントから end; までの DEFINE STYLE ステートメントブロック内において、STYLE ステートメントまたは CLASS ステートメントでスタイル要素を加工することで、スタイルテンプレートをカスタマイズしている。従前より REPLACE ステートメントがよく使われているが、現在、REPLACE ステートメントはサポートされておらず、TEMPLATE プロシジャのマニュアルでは STYLE/CLASS ステートメントの利用が推奨されている [3]。

ここで、スタイル要素 (style elements) とスタイル属性 (style attributes) について説明する。

スタイル要素は、ヘッダー/フッター、表ヘッダー、表フッター、表中の値、などの文書スタイルを構成する要素を指し、それぞれのスタイル要素に属するものとして、スタイル属性が存在する。CODE 2 に示す例では、STYLE ステートメントで指定している systemtitle がスタイル要素であり、font_face=などがスタイル属性に該当する。どのようなスタイル要素・スタイル属性が指定可能かを確認する方法は、SAS ヘルプセンター内のマニュアル[4][5][6]を参照する方法や過去の論文などを参照する方法が有効である。これらの方法は、特に SAS 初学者にとっては検索のハードルが高いと思われるため、ODS_RTF_sample_basic.sas の TEMPLATE プロシジャ内に、比較的頻繁に指定するスタイル要素とスタイル属性を例示している。なお、スタイルテンプレートの中身を知りたい場合、[proc template; source Myrtftemp; run;] を実行すると、ログ上にスタイルテンプレートの詳細な内容が表示される。また、SASHELP.TMPLMST や WORK.TEMPLAT 内に保存されているスタイルテンプレートを確認したい場合は、[proc template; list styles / store=sashelp.tmplmst; run;] を実行すると、アウトプット画面にスタイルテンプレート一覧が表示される。

2.2 OPTIONS/TITLE/FOOTNOTE ステートメントの使用例

CODE 3: OPTIONS Statement

```
options nodate nonumber;  
options orientation="portrait";
```

まず、RTF 出力における OPTIONS ステートメントの役割を紹介する。CODE 3 の options nodate nonumber; は、SAS Output に表示されるシステム日付及びページ番号を抑制するために指定している。また、options orientation=では、ページの縦 (Portrait)・横 (Landscape)を指定している。RTF 出力ではページ縦表示か横表示かで、出力可能な列数や出力可能な列の幅の合計が異なってくるので、レイアウト作成段階に、ページの縦・横をあらかじめ決めておくことが肝要である。

CODE 4: TITLE Statement and FOOTNOTE Statement

```
title1 j=right "Basic Tips";  
title2 j=right "ODS RTF Sample (proc report)";  
footnote1 j=center "Page ~{thispage} of ~{lastpage}";
```

次に、RTF 出力における TITLE ステートメント及び FOOTNOTE ステートメントについて紹介する。TITLE/FOOTNOTE ステートメントに指定した文字列は、RTF 出力のヘッダー/フッター部分に表示される

(ODS RTF ステートメントで BODYTITLE オプションを指定しない場合)。また、どちらのステートメントにも文字列の表示位置を指定する justify=(j=)オプションが存在する。CODE 4 では、タイトルは右寄せ (right)、フットノートは中央揃え (center) で配置されるように指定している。TITLE/FOOTNOTE ステートメントには、他にも文字色、太字、フォント等を指定するオプションが存在するが、ODS_RTf_sample_basic.sas では、既にスタイルテンプレートにおいて、ヘッダー/フッターのスタイルを定義しているため (2.1, 2.4 節を参照)、TITLE/FOOTNOTE ステートメントでのオプション指定は不要である。なお、TITLE と FOOTNOTE は最大 10 個まで指定できる [7]。

また、CODE 4 の footnote1 に示す通り、エスケープ文字 (2.3 節を参照) を利用して、[エスケープ文字(~)+{thispage}]で現在のページ番号、[エスケープ文字(~)+{lastpage}]で総ページ数を表示することができる。~{thispage}や~{lastpage}は、ODS RTF ステートメントのオプションで BODYTITLE オプションを利用しない場合のみ有効であり、ODS RTF ステートメントのオプション BODYTITLE を指定した場合には、ページ数が表示されないので注意が必要である。

2.3 ODS ESCAPECHAR ステートメントの指定例

ODS ESCAPECHAR ステートメントは、インラインフォーマット関数を利用するためのエスケープ文字を指定するステートメントである。CODE 5 の ods escapechar=では、エスケープ文字として「~」(チルダ) を指定している。RTF 出力の場合は、エスケープ文字+インラインフォーマット関数を文字列の先頭に指定することで、文字列のスタイルを変更することができる。また、セル内改行、上/下付き文字、ページ番号などを RTF 出力内に入れる際にも、エスケープ文字は用いられる。3.1 節で紹介する RTF 出力での全角・半角フォントの制御方法においても、エスケープ文字の指定が重要になる。エスケープ文字に使用可能な文字やインラインフォーマット関数の詳細については、SAS ヘルプセンター内のマニュアルを参照されたい [8]。

CODE 5: ODS ESCAPECHAR Statement

```
ods escapechar="~";
```

2.4 ODS RTF ステートメントの使用例

ODS RTF ステートメントは、Microsoft Word で使用する RTF 形式の出力を制御するためのステートメントである。また、ODS RTF ステートメントには CODE 6 で紹介する FILE=/ STYLE=/ STARTPAGE=/ BODYTITLE/ TEXT=/ PREPAGE=オプションの他にも多数のオプションが用意されているため、関心のある方は確認されたい [9]。

CODE 6: ODS RTF Statement

```
ods rtf file="&sasprgpath.¥Basic_procreport.rtf" style=Myrtftemp startpage=yes ;
```

ODS RTF ステートメントの FILE=オプションによって、RTF 出力先 (書き込み先) の RTF ファイルが開かれ、[ods rtf close;] または[ods _all_ close;]の操作を行うまで開いたままになる。STYLE=オプションにて、2.1 節の CODE 2 にて作成したスタイルテンプレート Myrtftemp を指定し、RTF 出力に定義済みのスタイルを適用する。STARTPAGE=オプションは改ページを制御するオプションで、YES の場合はプロシジャごとに改ページされる。STARTPAGE=YES がデフォルトの設定となっている。プロシジャごとに改ページしたくない場合は STARTPAGE=NO を指定し、改ページしたい場所に[ods rtf startpage=now;]を記述する。

なお、CODE 6 には記載していないが、ODS RTF ステートメントには BODYTITLE というオプションが存在する。BODYTITLE を指定すると、TITLE/FOOTNOTE ステートメントで指定したテキストが RTF 出力の本文中に出力される。一方、CODE 6 では BODYTITLE を指定していないため、TITLE/FOOTNOTE ステートメントで指定したテキストはそれぞれ RTF 出力のヘッダー/フッターに出力される。

CODE 7: ODS RTF PREPAGE= / TEXT= Option

```
ods rtf prepage="Listing16.2.1 SASHELP.CLASSFIT";  
[PROC REPORT の出力1]  
ods rtf text="This is a text like footnote.";  
ods rtf prepage="Listing16.2.2 SASHELP.CLASSFIT";  
[PROC REPORT の出力2]  
ods rtf text="This is a text like footnote.";
```

次に、任意のテキストを RTF 出力に表示するためのオプションの使用例を示す。ODS RTF PREPAGE=は RTF 出力の各プロシジャで作成される表の前に挿入される文字列を指定するために用いられる。一方、ODS RTF TEXT= (ODS TEXT=)は RTF 出力の任意の場所に文字列を挿入する際に用いられる。文字列を挿入するという点では似ているが、厳密には微妙に機能が異なる。ODS RTF ステートメントで STARTPAGE=YES を指定した場合、ODS RTF PREPAGE を [PROC REPORT の出力 1]、 [PROC REPORT の出力 2] の間に記述したとき、ODS RTF PREPAGE で指定したテキストは [PROC REPORT の出力 2] のタイトルのように出力される。一方、ODS RTF TEXT を [PROC REPORT の出力 1]、 [PROC REPORT の出力 2] の間に記述したとき、ODS RTF TEXT で指定したテキストは [PROC REPORT の出力 1] の脚注のように出力される。これらの文字列を挿入するオプションは、TITLE/FOOTNOTE ステートメントの代替として、表タイトルや表脚注を挿入する際に用いられる。ただし、ODS RTF STARTPAGE=NO か YES の違いで、表とテキストの間の空白行の有無が異なるため、注意が必要である。

2.5 REPORT プロシジャと ODS RTF プロシジャの使用例

本節では、RTF 出力の表作成に使用される代表的なプロシジャである、REPORT プロシジャ・ODS RTF プロシジャの使用例を紹介する。ODS RTF_sample_basic.sas においては、2.1 節の CODE 2 にて作成したスタイルテンプレートを利用するため、プロシジャ内のオプションやインラインフォーマット関数などを用いたスタイルの設定は最小限に留めている。CODE 8 及び CODE 9 は、SASHELP.CLASSFIT の一部の変数を計算・加工して作成した OUT1 から OUT5 までの文字値変数を持つデータセットを出力するための REPORT プロシジャ・ODS RTF プロシジャのコードである。どちらのコードも同様の見た目の出力を作成できる。

CODE 8: REPORT Procedure

```
%let COMMON_BTMBORDER=%str(borderbottomwidth=1pt borderbottomcolor=black borderbottomstyle=solid);  
  
proc report data=_classfit missing split=#;  
  column OUT1-OUT3 ("S={&COMMON_BTMBORDER.}Physique" OUT4 OUT5);  
  define OUT1/ display "Name" style(header)=[just=left] style(column)=[cellwidth=40mm just=left];  
(中略 ★下に続く)
```

(中略 ★上の続き)

```
compute after _page_/style=[just=left vjust=top borderbottomstyle=none];  
line "Customize footer by yourself."  
endcomp;  
run;
```

REPORT プロシジャには、表作成機能 (PRINT プロシジャとほぼ同様の機能) の他にも様々な機能が搭載されており、オプションもそれぞれのステートメントに対して多く用意されているが、本論文で紹介する機能は、表作成機能+ α に絞っている。なぜなら、特に統計解析業務においては、REPORT プロシジャを用いて要約・計算・データ加工などをすることはほとんどなく、出力用 (解析結果表示用) のデータセットを予め手前の解析プロシジャ及びデータステップで作成しておくことが一般的であり、実際に使用する機能は限られているからである。REPORT プロシジャのその他の機能については、SAS ヘルプセンター内のマニュアルや過去の論文等を参照されたい [10] [11]。

CODE 8 では、COLUMN ステートメント内で、複数の変数を括弧で囲み、OUT4 と OUT5 に共通の表ヘッダーを追加している。この表ヘッダーのテキストには、インラインフォーマット関数(CODE 8 の%let COMMON_BTMBORDER= 及び ~S={&COMMON_BTMBORDER.})の部分(参照)で下罫線を引く指定をしている。このため、図 2-3 のような表示になる。

図 2-3 CODE 8 の COLUMN ステートメントで表ヘッダーを追加したときの RTF 出力画面

			Physique	
Name	Sex	Age (years)	Height (cm)	Weight (kg)
Joyce	F	11	130.3	22.9
Louise	F	12	143.0	34.9

また、COLUMN ステートメントで宣言した変数について、DEFINE ステートメントで出力用に作成されたデータセットの変数をそのまま表示 (DISPLAY オプションを指定) している。DEFINE ステートメントでは、データセット内の変数ごとに、変数の役割や表ヘッダーとなるテキスト、スタイルを指定することができる。CODE 8 の DEFINE COL1/ の部分では、変数の役割を DISPLAY とし、STYLE(HEADER)=オプションでは表ヘッダーのスタイル属性 (just=left) を指定、STYLE(COLUMN)=オプションでは、表中の値とセル幅などのスタイル属性 (cellwidth=40mm just=left) を指定している。変数ごとに、文字の配置、文字色や背景色などを変更したい場合は、これらの変数ごとの個別のオプションが有効である。

また、compute after _page_ から endcomp; までの COMPUTE 構文の中で LINE ステートメントを用いることで、図 2-4 に示す通り表の脚注として文字列を追加することができる。さらに、COMPUTE ステートメントを用いれば、データセットの変数の値ごとにスタイルを変更 (例えば、特定の位置で下罫線を引くなど) が可能である。

図 2-4 CODE 8 の COMPUTE ステートメントで表の脚注を追加したときの RTF 出力画面

Alfred	M	14	175.3	51.0
Philip	M	16	182.9	68.0
Customize footer by yourself.				

CODE 9: ODSTABLE Procedure

```
proc odstable data=_classfit;
  column OUT1-OUT5;
  define header hd45; text "Physique"; start=OUT4; end=OUT5; style={&COMMON_BTMBORDER.}; end;
  define OUT1;
    define header hd1; text "Name"; style={just=left}; end; style={cellwidth=40mm just=left}; header=hd1;
  end;
  (中略)
  define footer btm1; text "Customize footer by yourself."; start=OUT1; end=OUT5; style={just=left vjust=top}; end;
run;
```

ODSTABLE プロシジャは、TEMPLATE プロシジャの DEFINE TABLE ステートメントと同様の機能を持ち、表形式の出力テンプレートを作成することに特化したプロシジャである。表を含む RTF 出力を作成する場合に多く利用されている。

CODE 9 では、COLUMN ステートメントで宣言した変数について、DEFINE ステートメントにおいてそれぞれ定義をしている。また、DEFINE HEADER、DEFINE FOOTER では、それぞれ表ヘッダーと表フッターの定義を行っている。表ヘッダーと表フッターも DEFINE ステートメントで定義できることが ODSTABLE プロシジャの特徴である。CODE 9 では、define header hd45; の部分で hd45 という表ヘッダーを定義し、start=OUT4; end=OUT5; とすることで、OUT4 と OUT5 に共通の表ヘッダーを追加している。これによって、REPORT プロシジャでの、COLUMN ステートメント内で複数の変数を括弧で囲む方法と同じ結果となる (図 2-3 と同様の結果となる)。define footer btm1;の部分では、btm1 という表フッターを定義し、start=OUT1; end=OUT5; とすることで、表フッターを追加している。これによって、REPORT プロシジャでの、compute after _page_ から endcomp;までの COMPUTE 構文の中で LINE ステートメントを用いて表の脚注のような文字列を追加する方法と同じ結果となる (図 2-4 と同様の結果となる)。なお、ODSTABLE プロシジャでのスタイルの定義の仕方は、REPORT プロシジャと類似しているため、詳細はプログラム内のコードを参照されたい。

RTF 出力での表作成の場合、各社・各組織の実業務で運用する上では、REPORT プロシジャと ODSTABLE プロシジャのどちらを使うかをプログラマ間で統一することを勧めたい。なぜなら、どちらのプロシジャを用いても工夫次第で同一の表現が可能とはいえ、そもそもの仕組みが異なるプロシジャであり、プログラマが各々別のプロシジャで表を作成すると、インラインフォーマット関数の指定箇所や表ヘッダー/表フッターの構成などの違いによって RTF タグの構成が異なり、後工程で RTF ファイルを比較することが困難になるからである。どちらのプロシジャが、或いは他のプロシジャが、より自分たちが表現したい内容を適切に表現できるかを十分に検討したうえで、プロシジャの使い分けとプログラマ間での使用プロシジャの統一を推奨する。

3 Advanced Tips

本章では、応用編として RTF 出力において有用な技術・Tips を紹介する。3.1 節で紹介するフォント制御の他にも応用的な技術を取り入れたプログラム ODS_RTF_sample_advanced.sas を用意した。基礎編同様、詳細な説明は省くが、実際に挙動を確かめることで、ある程度理解は可能である。本章で紹介する内容が、RTF

出力の可能性を広げる一助となれば幸いである。

3.1 RTF 出力での全角・半角フォントの制御方法

日本語と英語が混在する RTF の帳票作成時に役立つ全角・半角フォントの制御方法について、筆者が SAS ユーザー総会 2018 にて発表した方法の最新版を紹介する。

まず、全角・半角フォントの制御方法を開発するに至った背景を説明する。日本語と英語が混在している帳票において、「日本語：MS 明朝、英数字：Times New Roman」「日本語：MS ゴシック、英数字：Arial」などの組み合わせを依頼された場合、現状 (2024 年現在も)、TEMPLATE プロシジャで作成するスタイルテンプレートや REPORT プロシジャの STYLE オプションなどに、全角文字及び半角文字表示用のフォントを個別に指定できるオプションは存在しない。そこで筆者は、RTF ファイルを Word で開き直してフォントをかけ直すといった処理などをせずに、1 回の出力で、なおかつ RTF タグの加工もせずに、RTF 出力の全角・半角文字について任意のフォントを制御できる方法を開発し、SAS ユーザー総会 2018 にて発表した [13]。本論文では、SAS ユーザー総会 2018 で紹介したコードよりも簡単なコードを提案する。

全角・半角フォントの制御方法を利用する上での前提条件は、ODS RTF ステートメントを使用して RTF 出力を作成することと、2.3 節で紹介したエスケープ文字 (例示では「~」チルダ) を定義しておくこと、以上 2 つである。スタイルテンプレートにて定義するフォントは何でも良いが、全角文字が現れるときにだけ全角・半角フォントの制御をすれば良いので、半角文字表示用のフォント (例えば、Times New Roman) を指定しておくことを推奨する。

まず、全角・半角フォントの制御方法の基本の形を紹介する。インラインフォーマット関数でのフォントの指定を重ねて 2 回行う (~S={font_face='MS 明朝'}~S={font_face='Times New Roman'}) ことである。全角文字表示用フォント→半角文字表示用フォントの順で指定することがポイントである。

CODE A1: 全角・半角フォントの制御方法の基本の形

```
ods rtf prepage="~S={font_face='MS 明朝'}~S={font_face='Times New Roman'}Listing16.2.2 SASHELP.CLASS (全角半角フォント制御あり)";
```

次に、REPORT プロシジャでの全角・半角フォント制御方法を紹介する。

CODE A2: REPORT プロシジャでの全角・半角フォント制御の例

```
proc report data=_class missing split=#'
  style(header)={pretext="~S={font_face='MS 明朝'}~S={font_face='Times New Roman'}"}
  style(column)={pretext="~S={font_face='MS 明朝'}~S={font_face='Times New Roman'}"}
  style(lines)={pretext="~S={font_face='MS 明朝'}~S={font_face='Times New Roman'}"};
  column OUT1-OUT3 ("~S={COMMON_BTMBORDER.}体格" OUT4 OUT5);
  define OUT1/ display "名前" style(header)={just=left} style(column)={cellwidth=40mm just=left};
(中略)
run;
```

CODE A2 では、REPORT プロシジャの PROC REPORT ステートメントの STYLE オプションにて、各 location (header, column, lines) に対して {pretext="~S={font_face='MS 明朝'}~S={font_face='Times New Roman'}"} を指定している。このように pretext= を用いる方法によって、SAS ユーザー総会 2018 にて紹介した方法 (REPORT

プロシジャの COMPUTE ステートメントを用いて内部的にデータにインラインフォーマット関数を付与し、スタイルをデータセット内の文字列に対して直接制御する方法) よりも比較的簡単に全角・半角フォント制御が可能になった。ODSTABLE プロシジャでも同様に `pretext=`を用いる方法が有効である。ODSTABLE プロシジャでの全角・半角フォント制御例は `ODS_RTF_sample_advanced.sas` 内のコードを参照されたい。

3.2 ODS TAGSETS.RTF ステートメントの使用例

ODS TAGSETS.RTF ステートメントとは、2.4 節で紹介した ODS RTF ステートメントと同様に、RTF 形式の出力を制御するためのステートメントである [14][15]。ただし、ODS RTF ステートメントとは異なり、「Microsoft Word で使用する」という前提はなく、ODS TAGSETS.RTF ステートメントで作成する RTF 出力は、Microsoft Word に依存しない RTF 出力 (Measured RTF) となるのが特徴である。また、ODS RTF ステートメントとは種類の違うステートメントであり [16]、ODS RTF を ODS TAGSETS.RTF に置き換えただけでは上手く機能しない場合がある。加えて、ODS RTF と比べて RTF タグの構成も異なっている。

また、ODS TAGSETS.RTF ステートメントで生成される Measured RTF の特徴を以下に挙げる。

- Microsoft Word 以外の RTF リーダーで読み取ることができる
- 非常に大きな表の改ページを (メモリ消費量を節約しながら) 制御できる
- タイトル・フットノートを本文中に表として配置する

Measured RTF では、Microsoft Word 特有のヘッダー/フッターを使用しないため、ページ番号(`~{thispage}`や`~{lastpage}`) が使えないという難点があるが、RTF 出力自身が Microsoft Word に依存せずに、改ページやタイトル・フットノートの配置を決められるという利点がある (例えば、非常に大きな表を出力する場合、ODS RTF で作成される従来の RTF は表全体を読み込んでから Microsoft Word によって改ページの処理をしており、ページ間で情報を維持するために各ページの出力の際は多くのメモリを必要とするのに対して、Measured RTF は SAS のタグセットによってあらかじめ改ページ位置を決めているので各ページの出力の際は 1 ページ分のメモリのみで済む)。

筆者は、ODS TAGSETS.RTF ステートメントが、今後 ODS RTF ステートメントに取って代わるのか、どのような点が実業務上で役に立つかなどを検証するために、ODS TAGSETS.RTF ステートメントを用いて RTF 出力を作成するプログラム `ODS_TAGSETS.RTF_sample.sas` を作成した。以下に、ODS TAGSETS.RTF ステートメントの使用例と ODS RTF ステートメントとの違いなどを紹介する。

CODE A3: ODS TAGSETS.RTF Statement

```
ods tagsets.rtf file="&sasprgpath.¥Advanced_odstagsetsrtf_procreport.rtf"  
  
    style=Myrtftemp  
  
    startpage=yes  
  
    options(vspace='off');  
  
ods tagsets.rtf options(tables_off='prepage');  
  
ods tagsets.rtf prepage="Listing16.2.1 SASHELP.CLASSFIT";
```

まず、ODS TAGSETS.RTF ステートメント固有の機能を紹介する。CODE A3 の `OPTIONS(VSPACE='OFF')` オプションは、タイトルと表の間の空白及び表とフットノートの間の空白を取り除くためのオプションである。一方、`VSPACE='ON'` とすれば、ODS RTF ステートメントの出力と同様に空白を加えることができる(図 3-1 参照)。 `OPTIONS(TABLES_OFF='PREPAGE')` オプションは、引用符の中にスタイル要素名を指定するこ

とで、そのスタイル要素の表出力をオフにする (本文中に直接書き込む) ことが可能になる。

その他にも ODS TAGSETS.RTF ステートメント固有のオプションがあるので、詳細は SAS ヘルプセンター内のマニュアルを参照されたい。ODS TAGSETS.RTF には ODS RTF と共通しているオプションも存在するが、ODS RTF では使用可能だが ODS TAGSETS.RTF では使用不可の場合 (あるいはその逆) やコードの記載の仕方が違う場合があるので、使用の際はよく確認してほしい。

図 3-1 ODS TAGSETS.RTF OPTIONS(VSPACE='OFF') (上) と OPTIONS(VSPACE='ON') (下)の違い

Advanced Tips ODS TAGSETS.RTF Sample (proc report)				
Listing16.2.1 SASHELP.CLASSFIT				
			Physique	
Name	Sex	Age (years)	Height (cm)	Weight (kg)
Joyce	F	11	130.3	22.9

次に、ODS RTF ステートメントとの違いを紹介する。特に大きな違いが発生するのは、TITLE/FOOTNOTE ステートメントの挙動である。TITLE/FOOTNOTE ステートメントで指定したテキストは、ODS RTF (BODYTITLE 指定なし) の場合はヘッダー/フッター中のテキストとして出力され、各ページにまたがって表示されるが、ODS TAGSETS.RTF の場合は本文中のタイトル・フットノートとして出力され、各ページにまたがって表示される。また、ODS RTF (BODYTITLE または BODYTITLE_AUX 指定あり) の場合、本文中に出力されるが、表が 1 ページに収まらない場合に、表の一番上と一番下に表示されるのみである。一方、ODS TAGSETS.RTF の場合は、表が 1 ページに収まらない場合に、各ページに分割された表の上と下に、常にタイトル・フットノートとして表示される。

また、ODS TAGSETS.RTF での改ページの仕方、PREPAGE/TEXT ステートメントの出力位置についても、ODS RTF とは違う挙動を示す。表作成に使用するプロシジャが REPORT プロシジャか ODSSTABLE プロシジャかでも違いが生じる場合がある。出力位置や出力方法の違いについて詳細を確認したい場合は、それぞれのステートメントで作成した出力結果を見比べて確認することを推奨する。

以下は筆者の所感であるが、ほとんどの RTF 出力作成時において、ODS RTF ステートメントの代わりに ODS TAGSETS.RTF ステートメントをわざわざ使う必要はないと考える。ただし、非常に大きな表で各ページに出力する行数を決めたい場合 (TABLEROWS=オプションを使う) や表がページをまたぐ場合に継続タグをつけたい場合 (OPTIONS(CONTINUE_TAG='ON')オプションを使う) など、ODS TAGSETS.RTF ステートメントならではの機能を十分に理解できれば、効果的に使用可能と考える。実業務での実際の運用に至るかどうか、検討する余地は残されている。今後の研究課題として本論文で紹介した内容が参考になれば幸いである。

3.3 RTF タグの活用事例

RTF ファイルをメモ帳で開くと図 3-2 のように表示でき、RTF タグを確認できる。

図 3-2 SAS で出力した RTF ファイルをメモ帳で開いた時の様子

```

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
{#rtf1#ansi#ansicpg932#uc1#deff0#deflang1041#deflangfe1041
{#fonttbl
{#f1#fmodern#fprq1#fcharset128#cpg932 Courier;}
{#f2#froman#fprq2#fcharset128#cpg932 MS PMincho;}
{#f3#froman#fprq2#fcharset128#cpg932 Times New Roman;}
}{#colortbl;
#red0#green0#blue0;
#red0#green0#blue255;
#red0#green255#blue255;
#red0#green255#blue0;
#red255#green0#blue255;
#red255#green0#blue0;
#red255#green255#blue0;
}

```

SAS で出力した RTF ファイルの RTF タグは比較的シンプルな構成をしているため、RTF ファイルを SAS データセットとして再読み込みすることで、RTF タグの加工や RTF ファイル同士の結合、RTF タグを利用した RTF ファイルのコンペアなどが可能になる。ただし、SAS で出力した RTF ファイルを Microsoft Word で開き再保存した場合は、RTF タグに余分な情報が付加され、非常に複雑な RTF タグ構成を持つ RTF ファイルとなるので注意が必要である。さらに、RTF タグの構成を理解しないまま、RTF タグ加工や RTF ファイル結合などをすると、RTF ファイルを開くことができなくなる (RTF ファイルが破損する) 恐れがあるので注意が必要である。また、一度出力した結果に対して加工することになるため、実業務での運用は結果の信頼性の観点から憚られるかもしれない。ここでは、業務効率化の一環として RTF タグを活用した事例を 2 つ紹介する。ただし、実業務における運用事例ではなく、開発中の内容を含み、あくまで活用の可能性を検討しているものをご承知いただきたい。

【ODS RTF WORDSTYLE=オプションを利用したオリジナルの見出し 1 の作成】

ODS RTF WORDSTYLE=オプションを用いると Microsoft Word 上のみで使用可能な見出しの情報を RTF ファイルに付与することができる。使用例を CODE A4 に示す。CODE A4 では引用符の中に{¥s1 heading 1;}とあるが、これによって、¥s1 (RTF のマークアップ用のタグ) を「見出し 1」と認識させることができる。

¥s1 はエスケープ文字と併せて使用し、見出し 1 と認識させたいテキストの先頭に付加する。その後、一旦出力した¥s1 のタグを含んだ RTF ファイルを SAS データセットで再び読み込み、¥s1 によって「見出し 1」と認識されたテキストに対して本文中に直接書き込むように RTF タグの加工を行うと、図 3-3 のように見出し 1 のテキストがナビゲーションウィンドウに表示されるようになる。この方法は、オリジナルの目次 (Table of Contents) を作成する際にも利用することができる。

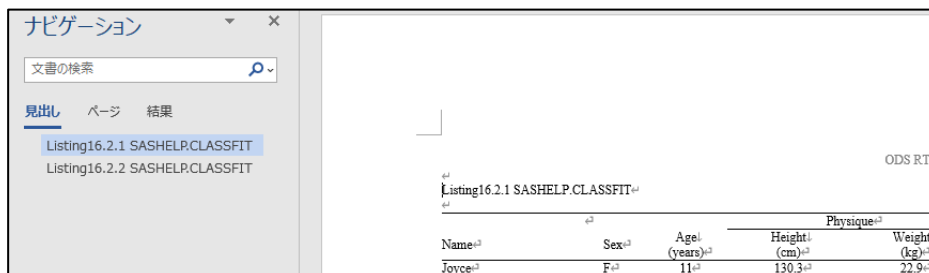
CODE A4: ODS RTF WORDSTYLE= Option and ODS RTF PREPAGE= Option

```

ods rtf file="xxxxxxx.rtf"
      wordstyle={¥s1 heading 1.};
ods rtf prepage=" ~¥s1 Listing16.2.1 SASHELP.CLASSFIT";
(中略)
ods rtf prepage=" ~¥s1 Listing16.2.2 SASHELP.CLASSFIT";

```

図 3-3 ODS RTF PREPAGE=で指定したテキストを見出し 1 として表示したときの Word 上の画面



【RTF ファイルの結合】

個別に出力した RTF ファイル同士の結合を行うための SAS プログラムを開発した。プログラムの流れは以下の通りである。

- 結合元の RTF ファイル (No.1) を SAS データセットとして読み込んで末尾の RTF タグを加工する。
- 結合する RTF ファイル (No.2) を SAS データセットとして読み込んで結合に必要な部分だけを抽出し、RTF タグを加工する。
- No.1 の加工済 SAS データセットと No.2 の加工済 SAS データセットを縦結合する。
- 上述の工程を結合する RTF ファイルの数だけ繰り返し、最終的に出来上がった SAS データセットを RTF ファイルとして再出力する。

ただし、同じスタイル定義・同じ環境で作成した RTF ファイルの結合に限られる、という制限がある。RTF ファイルの結合についてはその他にも方法が報告されているが、この方法は、Word や VBA などの機能を使用せずに、SAS の機能のみで実現しているという点で画期的である。

4 最後に

本論文の Basic Tips を参考にすれば基本的な RTF 出力が一通り実現可能になると考えている。ただ、ここで紹介した技術や Tips が決して RTF 出力の最適解ではなく、まだまだ改良の余地があると思う。プロシージャに搭載された機能をより効果的に使えば、より多様な表現ができるかもしれない。サブタイトルに記載しているように、紹介した技術や Tips が、全 SAS ユーザーにとってプログラム開発の手助けとなるとともに、プログラム開発を”たのしむ”為のツールとなることを期待している。各プログラマが試行錯誤しながらツールを引用・改良することで、自分が納得できるプログラムを開発することを、是非たのしんでいただきたい。また、そのようにプログラム開発する中で発見されるより良い技術や Tips の共有が、今後も活発になされていくことを期待する。

5 参考文献

- [1] SAS® Output Delivery System: Procedures Guide. SAS® Help Center (Programming Documentation for SAS® 9.4 and SAS® Viya® 3.5).
https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/odsproc/titlepage.htm
- [2] TEMPLATE Procedure: Creating a Style Template - Example 3: Modifying the Default Style with the CLASS Statement. SAS® Help Center.
https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/odsproc/p1trs6h1rlpxgjn17yofb92v31xz.htm
- [3] TEMPLATE Procedure: Creating a Style Template - REPLACE Statement. SAS® Help Center.
https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/odsproc/p126rcnu6y4s8cn1sg6i4sdq9g6i.htm
- [4] ODS Style Elements. SAS® Help Center.
https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/odsproc/p0pm4ysu0fb68dn1d8uj78weie6o.htm
- [5] Style Attributes Tables. SAS® Help Center.
https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/odsproc/n0otdo2g12obp3n0zmnghcn7p4vu.htm
- [6] Style Attributes Detailed Information. SAS® Help Center.

- https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/odsproc/p1pt77toue3iyun0z4I9gth5as9f.htm
- [7] TITLE Statement. SAS® Help Center.
https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/lestmtsglobal/p10gcmmf83iaxn1ilrx4pra969n.htm
- [8] ODS ESCAPECHAR ステートメント. SAS® Help Center.
https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/odsug/p11xia2ltavr8ln17srq8vn4rnqc.htm
- [9] ODS RTF ステートメント. SAS® Help Center.
https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/odsug/p1vsv8ucnjzjnn1wq5wrlp74mdb.htm
- [10] REPORT プロシジャ. SAS® Help Center.
https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/proc/p0bqogcics9o4xn17yvt2qjbgdpi.htm
- [11] 太田裕二. 2017. ODS 機能と REPORT プロシジャを用いた 解析帳票の作成. SAS ユーザー総会 2017.
https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups2017-b-11-02.pdf
- [12] The ODSTABLE Procedure. SAS® Help Center.
https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/odsproc/n1olg2oeqi9r5m1racah27f1ow8.htm
- [13] 麻生美由紀. 2018. ods rtf+proc report における解析帳票の日本語/英数字フォントの明示的制御. SAS ユーザー総会 2018.
https://www.sas.com/ja_jp/events/18/users-group-2018/sessions.html#m=4-2
- [14] ODS TAGSETS.RTF Statement. SAS® Help Center.
https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/odsug/p19rpsb989jyadn1kgunyh4t8g0e.htm
- [15] Xiangxiang Meng. 2010. ODS TAGSETS.RTF: Tips and Tricks. MWSUG-2010-67.
<https://www.mwsug.org/proceedings/2010/dataviz/MWSUG-2010-67.pdf>
- [16] ODS の仕組みの概要. SAS® Help Center.
https://documentation.sas.com/doc/ja/pgmsascdc/9.4_3.5/odsug/p027b3wsp0yg9yn1tfqekatdtn85.htm

交互作用を含む共分散分析における最小 2 乗平均の活用

高橋 行雄
BioStat 研究所(株)

Using LSMEANS in Analysis of Covariance Including Interactions

Yukio Takahashi
BioStat Research Co.,Ltd.

要旨: 共分散分析の適用に際し、共変量と群の交互作用が無いことが前提とされている。交絡変数を共変量と見なすような場合に、交互作用がある事例にしばしば遭遇する。なぜ交互作用が無いことが、前提になったのか、歴史的な背景を紐解くことにより、その理由を理解することができる。他方、線形モデルに交互作用項を含めることを前提にした解析方法の考え方が未整備であることも確かである。精度の高い実験の場合には、群間にわずかな傾きの差があっても、統計的には交互作用が有意な差となりやすく、精度の低い実験の場合には、多少の傾きの差があっても統計的には有意な差となりにくい。このようなジレンマは、予備検定の結果に判断を委ねたために起きる必然の結果である。そこで、本質的に有意な交互作用が出やすい事例を通じて統計的な有意差検定による交互作用が有意か否かの結果にかかわらず、交互作用項を含めた共分散分析における解析法を提示する。

キーワード: 線形モデル, 交互作用, 共分散分析, 交絡変数, 前後差, 前値, 抗うつ剤, 評価尺度

1. はじめに

量的な説明変数と反応の関係について、直線をあてはめる回帰分析がよく知られている。反応に影響を与える質的な説明変数も考慮したい場合には、それぞれの水準(カテゴリ)ごとに直線をあてはめ、反応に対する関係を仔細に検討したい。複数の直線が平行とみなせる場合、傾きが異なるが直線が互いに交わらない場合、直線がある一点を通るように見える場合、直線が互いに交差する場合など、千差万別である。このような場合について、どのような解析を行なったら良いのであろうか。

説明変数に質的な変数と量的な変数の両方を含み、反応が量的変数として得られた場合の解析法とし、共分散分析が連想される。しかしながら、共分散分析について解説されている成書は、なかなか見出しにくい。共分散分析について記述がある成書の多くは、共変量がある 2 群比較を主体にし、解析方法も計算手段が乏しい時代に定式化された単回帰分析を重層的に用いる煩雑な解析手順が示されている。共変量は、反応に影響を与える量的な変数であり、実験の開始直前に測定値を得ることはできるが、制御因子としてコントロールできない変数として定義されている。共分散分析は、共変量の影響を統計的に取り除き、制御因子としての質的な変数についての精度を高めた群間比較をするための方法として定式化されている。そのために、質的な変数 A と共変量 X との交互作用 $A \times X$ が無いことが大前提とされている。これは、交互作用が有ると共変量の影響を統計的に除いた群間比較ができなくなるためである。

量的な説明変数と質的な説明変数が混在するようなデータに対し、共分散分析を適用すれば、と思った瞬間に、伝統的な共分散分析の枠組みに絡め取られることになる。その呪縛は、“共変量”にある。共分散分析の目的は、質

的な変数の水準(カテゴリ)間の比較であり、邪魔になる“共変量”の影響を統計的に取り除きたいのである。量的な反応に対する回帰分析を行ない、質的な変数で層別(層別因子)したときに、どのような変化が回帰直線に起きるかを検討したいとの思いは、“共分散分析”を適用すればいいのではないかと思った瞬間に「平行でなければならない」という前提によって打ち碎かれる。

2. 共分散分析の伝統的な解析方法

SAS の一般線形モデル GLM プロシジャを用いた共分散分析に慣れ親しんでいる人達にとって、共分散分析の伝統的な解析方法には、目が眩むかもしれない。多くの成書で解説されている共分散分析の解析手順は、偏差平方和を主体に、単回帰分析を繰り返し適用した方法であり、GLM プロシジャの解析方法とはまったく異なる。どのような解析手順であるかを理解することにより、質的な変数 A と共変量 X との交互作用 A×X が無いことが大前提となっていることを理解することができる [奥野ら(1981), 宮原ら(1995), 朝香ら(1998), アーミテージら(2001), 青木(2004), フライス(2004), 浜田(2009), 芳賀(2014), 中澤(2018)]。

取り上げる事例は、奥野ら(1981)「多変量解析法[改訂版]」, 第 II 部 重回帰分析, 第 7 章 層別因子を含む回帰分析」の事例 7.1 である。原著に示されている解析法は、伝統的な共分散分析の解析手順であり、共分散分析を取り上げている成書などで共通の解析方法として定着している。ここでは、原著で示されている解析方法を Excel の力を借りつつ忠実に再現することから始める[高橋(2024a)]。

表 1 に示すデータは、亜硫酸ガス回収塔の洗浄用水の温度 x と、回収液の濃度 y との関係調べたが、測定回数が少ないので、長期間にわたって収集されたデータを使

わなければならない。その間に季節の変化ばかりでなく、操業条件も変わっているため、データを 4 つの季節に分けた結果が示されている。この事例で、洗浄水の温度 x は、季節 A の影響を受ける交絡変数であり、厳密な意味での“共変量”ではないが、共分散分析の事例として扱われている。

伝統的な共分散分析は、1 因子実験モデルでの解析に際し、反応 y に影響を与えることがはっきりしている量的変数があるが、実験に際し一定に保つことがどうしてもできない。この量的変数を「共変量」とし、統計的に共変量の影響を除去し、季節 A の水準間の比較をしたい。そのために、共変量の分布が季節 A に対し均一でなければ共変量として扱えないと認識されている。季節 A により洗浄水の温度 x の分布が異なるので共変量としての前提が満たされないことを無視したとしても、更なる適用条件とし、季節 A と洗浄水の温度 x の間に交互作用がないことも確認しなければならない。

(手順1) 全データに対する回帰分析

まず、 x と y に関する偏差平方和 ($S_{xx}^{(全体)}$, $S_{xy}^{(全体)}$, $S_{yy}^{(全体)}$) を計算し、回帰パラメータ ($\hat{\beta}_1^{(全体)}$, $\hat{\beta}_0^{(全体)}$) を求める。

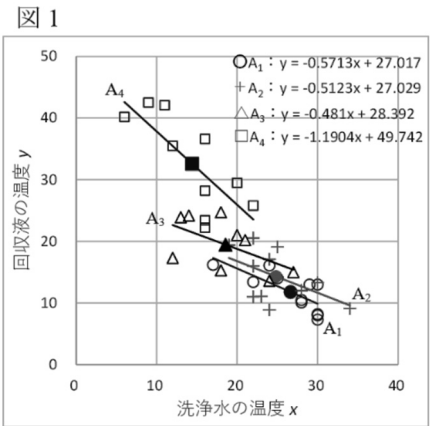
$$S_{xx}^{(全体)} = \sum_{i=1}^4 \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 = 1788.48$$

$$S_{xy}^{(全体)} = \sum_{i=1}^4 \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..}) = -2124.24$$

$$S_{yy}^{(全体)} = S_T^{(全体)} = \sum_{i=1}^4 \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = 3654.82$$

表1 季節ごとの洗浄用水の温度 x と回収液の濃度 y

No.	A ₁ (7月・8月)		A ₂ (9月・10月)		A ₃ (11月・12月)		A ₄ (1月・2月)	
	x	y	x	y	x	y	x	y
1	30	8.0	34	9.1	27	15.0	16	23.4
2	28	10.5	19	19.4	20	21.0	16	28.2
3	30	8.2	22	20.5	14	24.2	20	29.5
4	29	13.0	25	14.2	18	15.3	16	22.2
5	28	10.1	22	11.0	12	17.3	6	40.2
6	17	16.2	25	19.1	18	24.7	16	36.6
7	24	16.1	22	16.0	13	23.9	12	35.5
8	22	13.4	23	11.1	21	20.2	11	42.1
9	30	13.0	28	12.0	24	13.6	9	42.5
10	30	7.3	24	17.1			22	25.8
11	25	14.0	30	13.3				
12			24	8.9				
平均	26.64	11.80	24.83	14.31	18.56	19.47	14.40	32.60
		総平均	21.48	19.11	N=42			



次いで分散分析に必要な回帰の平方和 $S_R^{(全体)}$ および残差平方和 $S_e^{(全体)}$ を推定する. なお, $S_e^{(全体)}$ には, 季節 A の平方和 S_A および交互作用の平方和 $S_{R \times A}$ が含まれている.

$$S_R^{(全体)} = \frac{(S_{xy}^{(全体)})^2}{S_{xx}^{(全体)}} = \frac{(-2124.24)^2}{1788.48} = 2523.03 \quad \left| \quad S_e^{(全体)} = S_T^{(全体)} - S_R^{(全体)} = 3654.82 - 2523.03 = 1131.79$$

(手順 2) 季節 A の水準ごとの回帰分析の併合

表 2 に示す計算シート上で, 季節 A の水準ごとの偏差平方和 ($S_{xy}^{(i)}$, $S_{xx}^{(i)}$, $S_{yy}^{(i)}$) を計算し, i に関する合計を

$$S_{xx}^{(和)} = \sum_{i=1}^4 S_{xx}^{(i)} = 178.55 + 183.67 + 204.22 + 216.40 = 782.83$$

$$S_{yy}^{(和)} = \sum_{i=1}^4 S_{yy}^{(i)} = (-102.00) + (-94.08) + (-98.23) + (-257.60) = -551.92$$

$$S_{yy}^{(和)} = \sum_{i=1}^4 S_{yy}^{(i)} = 100.36 + 181.85 + 148.76 + 538.44 = 969.41$$

	A ₁	A ₂	A ₃	A ₄	和
n_i	11	12	9	10	42
\bar{x}_i	26.64	24.83	18.56	14.40	—
\bar{y}_i	11.80	14.31	19.47	32.60	—
$S_{xx}^{(i)}$	178.55	183.67	204.22	216.40	782.83
$S_{yy}^{(i)}$	-102.00	-94.08	-98.23	-257.60	-551.92
$\hat{\beta}_1^{(i)}$	-0.57	-0.51	-0.48	-1.19	—
$\hat{\beta}_0^{(i)}$	27.02	27.03	28.39	49.74	—
$S_{yy}^{(和)}$	100.36	181.85	148.76	538.44	969.41
$S_R^{(和)}$	58.27	48.19	47.25	306.64	460.36
$S_e^{(和)}$	42.09	133.65	101.51	231.80	509.05

として求め ($S_{xy}^{(和)}$, $S_{xx}^{(和)}$, $S_{yy}^{(和)}$) とする. また, 季節 A の水準ごとの回帰係数を

$$\hat{\beta}_1^{(i)} = S_{xy}^{(i)} / S_{xx}^{(i)}, \quad \hat{\beta}_0^{(i)} = \bar{y}_i - \hat{\beta}_1^{(i)} \bar{x}_i$$

で求め, 図 1 に示された回帰係数に一致することが確認できる. 水準ごとの回帰の平方和 $S_R^{(i)}$ を求め, それらの和を計算する. $S_R^{(和)}$ の和を $S_R^{(和)}$, $S_{yy}^{(和)}$ の和を $S_{yy}^{(和)}$ とし, それらの差から残差平方和 $S_e^{(和)}$ を求める.

$$S_R^{(和)} = \sum_{i=1}^4 (S_{xy}^{(i)})^2 / S_{xx}^{(i)} = 58.27 + 48.19 + 47.25 + 306.64 = 460.36$$

$$S_e^{(和)} = S_T^{(和)} - S_R^{(和)} = 969.41 - 460.36 = 509.05$$

要因	df	平方和	平均平方	F	p
x	1	2523.03	2523.03	168.52	0.0000
A	3	551.49	183.83	12.28	0.0000
x × A	3	71.25	23.75	1.59	0.2107
e	34	509.05	14.97	1.00	
全体	41	3654.82			

(手順 3) 季節 A について共通の傾き $\hat{\beta}_1^{(共通)}$ の推定

手順 2 で求めた $S_{xx}^{(和)} = 782.83$, $S_{xy}^{(和)} = -551.92$ から共通の傾き $\hat{\beta}_1^{(共通)} = -0.7050$ を算出する. 共通の傾きでの回帰の平方和 $S_R^{(共通)} = 389.11$ を推定し, $S_T^{(和)} = S_{yy}^{(和)} = 969.41$ との差から交互作用 $S_{x \times A}^{(共通)}$ を含む $S_e^{(共通)} = 580.29$ を算出する.

$$\hat{\beta}_1^{(共通)} = \frac{S_{xy}^{(和)}}{S_{xx}^{(和)}} = \frac{-551.92}{782.83} = -0.7050 \quad \left| \quad S_R^{(共通)} = \frac{(S_{xy}^{(和)})^2}{S_{xx}^{(和)}} = \frac{(-551.92)^2}{782.83} = 389.11$$

$$S_e^{(共通)} = S_T^{(和)} - S_R^{(共通)} = 969.41 - 389.11 = 580.29$$

(手順 4) 手順 1~3 で求めた平方和から分散分析表の作成

手順 1 の $S_{yy}^{(全体)}$ が, 総平方和 $S_T = S_{yy}^{(全体)} = 3654.82$, $S_R^{(全体)}$ が共変量 x の平方和 $S_x = S_R^{(全体)} = 2523.03$ となる. 手順 1 の $S_e^{(全体)} = 1131.79$ から手順 2 で示した $S_e^{(共通)} = 580.29$ との差が, 季節 A の平方和 $S_A = 551.49$ となる. 手順 3 の $S_e^{(共通)} = 580.29$ から手順 2 で示した $S_e^{(和)} = 509.05$ との差が, 交互作用 $S_{x \times A} = 71.25$ となる.

$$S_T = S_{yy}^{(全体)} = 3654.82$$

$$S_x = S_R^{(全体)} = 2523.03$$

$$S_A = S_e^{(全体)} - S_e^{(共通)} = 1131.79 - 580.29 = 551.49 \quad \left| \quad S_{x \times A} = S_e^{(共通)} - S_e^{(和)} = 580.29 - 509.05 = 71.25$$

$$S_e = S_e^{(和)} = 509.05$$

これらの平方和を用いて表 3 に示す分散分析表を作成し, 交互作用の有無について統計的判断を行なう. 表 3 に示すように交互作用の p 値は, $p = 0.2107$ と有意ではないので, 平行な直線があてはめられるとみなす.

(手順 5) 交互作用の平方和 S_{xxA} を残差平方和へプーリング
 交互作用 ($x \times A$) が有意ではないので、交互作用の平方
 和 $S_{xxA} = 71.25$ を残差平方和 $S_e = 509.05$ にプーリングし
 て表 4 示すように新たな残差平方和

$$S'_e = S_e^{(共通)} = 509.05 + 71.25 = 580.29$$

とし、分散分析表を作り直す。

(手順 6) 平行線をあてはめた場合の季節 A の水準ごとの切片の推定

手順 3 で求めた季節 A についての共通の傾き $\hat{\beta}_1^{(共通)} = -0.7050$ とす
 る場合の切片 $\hat{\beta}_0^{(i)}$ を

$$\hat{\beta}_0^{(i)} = \bar{y}_i - \hat{\beta}_1^{(共通)} \bar{x}_i$$

で推定すると表 5 に示すように $\hat{\beta}_0^{(1)} = 30.58$, $\hat{\beta}_0^{(2)} = 31.82$,
 $\hat{\beta}_0^{(3)} = 32.55$, $\hat{\beta}_0^{(4)} = 42.75$ が得られる。これらを用いて図 2 に示すよ
 うに共通の傾きを持つ平行な回帰直線を重ね書きする。

(手順 7) 季節 A の第 4 水準を基準とする水準間の差

図 2 に示したように、季節 A ごとの平行線に着目したときに、季節 A の
 A_4 水準を基準とし、それぞれの A_i 水準との平行線間の差と 95%信頼区間
 を推定し、図 3 に示すようにひげ付き線グラフを作成したい。

平行な回帰直線の任意の x に対し、基準となる水準を i' とすると、Y 軸
 方向の差 $d^{(i-i')}$ は、

$$d^{(i-i')} = (\hat{\beta}_0^{(i)} + \hat{\beta}_1^{(共通)} x) - (\hat{\beta}_0^{(i')} + \hat{\beta}_1^{(共通)} x) = \hat{\beta}_0^{(i)} - \hat{\beta}_0^{(i')}$$

のように、切片の差に帰着する。切片の差の分散を求めるために、 $\bar{y}^{(i)} = \hat{\beta}_0^{(i)} + \hat{\beta}_1^{(共通)} \bar{x}^{(i)}$ を $\hat{\beta}_0^{(i)}$ について解いた式、

および、 $\bar{y}^{(i')} = \hat{\beta}_0^{(i')} + \hat{\beta}_1^{(共通)} \bar{x}^{(i')}$ を $\hat{\beta}_0^{(i')}$ について解いた式を代入し、

$$\begin{aligned} \hat{\beta}_0^{(i)} - \hat{\beta}_0^{(i')} &= (\bar{y}^{(i)} - \hat{\beta}_1^{(共通)} \bar{x}^{(i)}) - (\bar{y}^{(i')} - \hat{\beta}_1^{(共通)} \bar{x}^{(i')}) \\ &= \bar{y}^{(i)} - \bar{y}^{(i')} - (\bar{x}^{(i)} - \bar{x}^{(i')}) \hat{\beta}_1^{(共通)} \end{aligned}$$

を得る。

(手順 8) 水準間の差の 95%信頼区間

回帰直線間の差 $d^{(i-i')}$ の分散 $Var(d^{(i-i')})$ は、平均値 $\bar{y}^{(i)}$, $\bar{y}^{(i')}$,
 $(\bar{x}^{(i)} - \bar{x}^{(i')}) \hat{\beta}_1^{(共通)}$ で表すことができ、これらは互いに独立で、 $(\bar{x}^{(i)} - \bar{x}^{(i')})$
 は誤差を持たない定数であり、 $\hat{\beta}_1^{(共通)}$ の分散は、 $Var(\hat{\beta}_1^{(共通)}) = \hat{\sigma}^2 / S_{xx}^{(共通)}$
 を用いて

$$\begin{aligned} Var(d^{(i-i')}) &= Var\left[\bar{y}^{(i)} - \bar{y}^{(i')} - (\bar{x}^{(i)} - \bar{x}^{(i')}) \hat{\beta}_1^{(共通)}\right] \\ &= Var(\bar{y}^{(i)}) + Var(\bar{y}^{(i')}) + Var\left[(\bar{x}^{(i)} - \bar{x}^{(i')}) \hat{\beta}_1^{(共通)}\right] \\ &= \frac{\hat{\sigma}^2}{n^{(i)}} + \frac{\hat{\sigma}^2}{n^{(i')}} + (\bar{x}^{(i)} - \bar{x}^{(i')})^2 \frac{\hat{\sigma}^2}{S_{xx}^{(共通)}} \\ &= \left(\frac{1}{n^{(i)}} + \frac{1}{n^{(i')}} + \frac{(\bar{x}^{(i)} - \bar{x}^{(i')})^2}{S_{xx}^{(共通)}}\right) \hat{\sigma}^2 \end{aligned}$$

となる。ここで、誤差分散の推定値 $\hat{\sigma}^2$ は、表 4 に示した交互作用をプーリングした残差の平均平方の 15.6836 を用い
 る。表 6 に示すように、季節 A_4 の切片 $\hat{\beta}_0^{(4)}$ 基準とした季節 A_i の切片 $\hat{\beta}_0^{(i)}$ の差について差の分散を計算し、次いで、
 95%信頼区間を求める。図 3 は、この結果を用いて Excel の線グラフを用いて作図した結果である。図から明らか
 のように、他の季節との間に、有意な差があることが示されている。

要因	df	平方和	平均平方	F	p
x	1	2523.03	2523.03	160.87	0.0000
A	3	551.49	183.83	11.72	0.0000
e	37	580.29	15.68	1.00	
全体	41	3654.82			

表5 傾きを共通とする場合の切片の推定

A	\bar{x}_i	\bar{y}_i	切片 $\hat{\beta}_0^{(i)}$	$\hat{\beta}_1^{(共通)}$
A ₁	26.64	11.80	30.58	-0.705
A ₂	24.83	14.31	31.82	-0.705
A ₃	18.56	19.47	32.55	-0.705
A ₄	14.40	32.60	42.75	-0.705

図 2

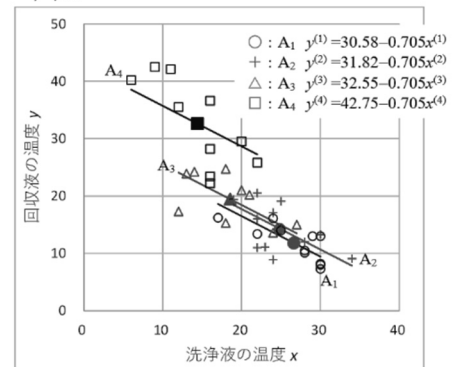
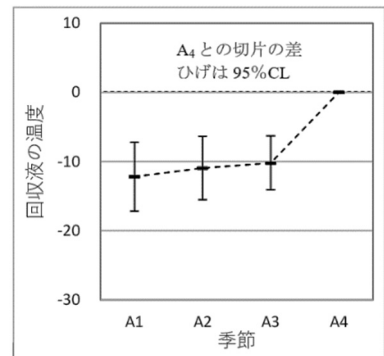


図 3



伝統的な共分散分析で定式化されているのは、ここまでである。

更なる結果の吟味をするために、図 4 に示すように、推定された回帰直線の 95%信頼区間を示し、季節 A ごとの洗浄水の温度 x が共通となる 20 度における回収液の推定温度 \hat{y}_i での 95%信頼区間付きの

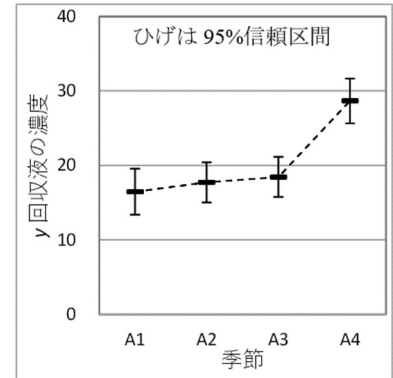
季節	切片	平均	差 ($\beta_0^{(i)}$)	$Var(\beta_0^{(i)})$	幅	差の95%CL
A_i	n_i	$\beta_0^{(i)}$	\bar{x}_i	$-\beta_0^{(i')}$	$t_{0.05} \times SE$	L 95% U 95%
規準 A ₄	10	42.75	14.40	—	—	—
A ₁	11	30.58	26.64	-12.17	5.9939	4.9606 -17.13 -7.21
A ₂	12	31.82	24.83	-10.94	5.0562	4.5561 -15.49 -6.38
A ₃	9	32.55	18.56	-10.20	3.6570	3.8747 -14.08 -6.33
A ₄	10	42.75	14.40	0.00		
$\sum_i S_{xx}^{(i)} = S_{xx}^{(共通)} =$			782.83		$\sigma^2 =$	15.6836
			$df =$	37	$t(0.05, 37) =$	2.0262

折れ線グラフも欲しいが、どうしたら良いのであろうか。図4に示すのは、Excel の行列計算、および、SAS の GLM プロシジャの lsmeans ステートメントの at オプションを用いて得た推定値を Excel の線グラフで作図した結果である。

伝統的な共分散分析は難解

奥野ら(1981)に示されている共分散分析の手順に忠実に Excel の力を借りて再現したのであるが、なんとも技巧的であり気が遠くなりそうである。私も実務で共分散分析を頻繁に使ってきたのであるが、ダミー変数を用いた線形モデルによる共分散分析を主体にしてきた。伝統的な解析手順については、面倒なことをしていると流し読みしていた。あらためて、共分散分析について記載されている成書を紐解いてみると、伝統的な解析法のオンパレードである。他方、Web 上での解説記事では、ダミー変数を用いた回帰分析を用いた解析方法がメジャーであるが、表層的な説明で終わっている。

図 4 洗浄水の温度が 20 度の場合



3. 前値を共変量とする交互作用がある抗うつ剤の評価

フライス著, KR(関西臨床データ解析)研究会誌(2004),「臨床試験のデザインと解析, 第 7 章 共分散分析と変化量の検討」にランダム化ブロック試験の 30 組の患者のうち、表 7 に示すように、初めの 10 組について治療前のうつ症状の Hamilton の評価尺度によるスコア X 、治療後のスコア Y が示されている。この試験データは、薬剤 A と前値 X に交互作用があるような場合について“共分散分析”による解析の考え方が丁寧に示されている。なお、Hamilton の評価尺度スコアは、大きいほど

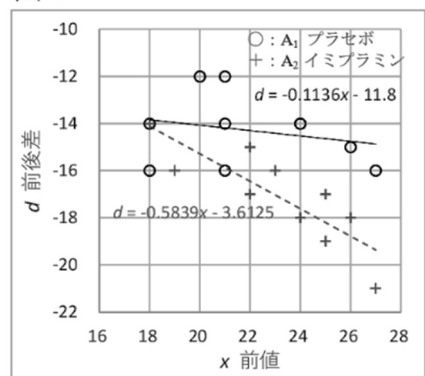
No	A ₁ : プラセボ			A ₂ : イミプラミン		
	前値 x	後値 y	差 d	前値 x	後値 y	差 d
1	18	4	-14	27	6	-21
2	21	7	-14	18	4	-14
3	24	10	-14	25	6	-19
4	24	10	-14	23	7	-16
5	18	2	-16	22	5	-17
6	27	11	-16	24	6	-18
7	21	9	-12	25	8	-17
8	21	5	-16	22	7	-15
9	26	11	-15	26	8	-18
10	20	8	-12	19	3	-16

うつ症状は重くなる。ここでは、簡単のためペアを考慮した解析ではなく、前値 X を共変量、薬剤を A、前後差(変化量) D を反応変数とする解析を行う。

前後差 D に対して薬剤間の比較をする際に、前値 X を共変量とするのは、「平均への回帰」と言われている現象の影響を取り除いて薬剤間の比較を行うためである。後値 Y を反応とする場合でも、同様に前値 X を共変量として含める必要がある。前後差 D による対応のない t 検定を適用したくなるのであるが、慎まなければならない。[高橋(2024b)]

図 5 に示すのは、前値 x と前後差 d の散布図に Excel の「近似曲線の追加」による線形近似(回帰直線)を重ね書きした結果である。前値が小さい場合に薬剤間の前後差 d は小さく、前値が大きい場合に薬剤間の前後差 d が広がっている。このような現象は、評価スコアが 0 以下にならないために、スコアの減少量に下限ができてしまうためである。

図 5



この図から A₁ プラセボ群は、回帰直線の傾きが小さいので、前値 x の大きさに関わらず前後差 d にはっきりとした違いが見いだせない。他方、A₂ イミプラミン群は、回帰直線の傾きが大きく、前値 x が大きければ前後差 d も大きく、前値 x が小さくなると前後差 d は、A₁ プラセボ群に近い前後差 d となることが、読み取れる。前値 x を共変量とする共分散分析を適用する際に、薬剤 A と前値 x に交互作用 $A \times x$ がないことが必須であるとの思い込みが蔓延している。交互作用 $A \times x$ を統計的に評価するため、薬剤 A を (1, -1) 対比型ダミー変数を a_1 として、交互作用モデルの回帰式、主効果モデルの回帰式

$$\text{交互作用モデル: } d_i^{(\text{交})} = \theta_0 + \theta_1 a_{1,i} + \theta_2 x_i + \theta_3 (a_{1,i} x_i) + \varepsilon_i$$

$$\text{主効果モデル: } d_i^{(\text{主})} = \theta_0 + \theta_1 a_{1,i} + \theta_2 x_i + \varepsilon_i$$

を用いる。薬剤 A が 2 水準なので、交互作用の検討は、分散分析表によらず $a_{1,i} x_i$ のパラメータ θ_3 に関する p 値で行なうことができる。

表 8 に示すのは、交互作用モデルに対応する (20 行×4 列) のデザイン行列 X である。デザイン行列 X は、(20 行×4 列) であるが、「定数に 0 を使用」をオフ(デフォルト)とした場合に、選択する変数は、切片 x_0 を除いた ($a_1, x, a_1 x$) である。表 9 上段に Excel の回帰分析を適用した交互作用モデルの結果を示す。表 9 下段に示す主効果モデルの場合は、交互作用のための変数 $a_1 x$ を除いた結果である。

		デザイン行列 X					
		交互作用モデル					
		主効果モデル					
No	A	x_0	a_1	x	$a_1 x$	y	d
1	A ₁	1	1	18	18	4	-14
2		1	1	21	21	7	-14
:							
10		1	1	20	20	8	-12
11	A ₂	1	-1	27	-27	6	-21
12		1	-1	18	-18	4	-14
:							
20		1	-1	19	-19	3	-16
平均				22.550	6.850		-15.7

表 9 上段のパラメータの推定値から交互作用を含む回帰式は、

$$\hat{d}^{(\text{交})} = \hat{\theta}_0 + \hat{\theta}_1 a_1 + \hat{\theta}_2 x + \hat{\theta}_3 a_1 x$$

$$= -7.7062 - 4.0938 a_1 - 0.3488 x + 0.2351 a_1 x$$

となる。(A₁, A₂)それぞれの回帰式は、

$$\hat{d}^{(\text{交}, A_1)} = -7.7062 - 4.0938 \times 1 - 0.3488 x + 0.2351 \times 1 \times x$$

$$= -11.8000 - 0.1136 x$$

$$\hat{d}^{(\text{交}, A_2)} = -7.7062 - 4.0938 \times (-1) - 0.3488 x + 0.2351 \times (-1) \times x$$

$$= -3.6125 - 0.5839 x$$

として求められ、これらは、図 5 の回帰式に一致する。

交互作用の検討は、変数 $a_1 x$ のパラメータ $\hat{\theta}_3 = 0.2351$ 、標準誤差 $SE(\hat{\theta}_3) = 0.1062$ を用い、 t 値と p 値は、

$$t = \frac{\hat{\theta}_3}{SE(\hat{\theta}_3)} = \frac{0.2351}{0.1062} = 2.2130$$

$$p = T.\text{dist.}2T(2.2130, 16) = 0.0418$$

となり、 $p < 0.05$ なので、有意な交互作用があると判断される。

交互作用の検討は、交互作用モデルの分散分析表と主効果モデルの分散分析表を合成することで行うことができる。表 9 下段に示されているパラメータから主効果モデルの回帰式は、

$$\hat{d}^{(\text{主})} = \hat{\theta}'_0 + \hat{\theta}'_1 a_1 + \hat{\theta}'_2 x$$

$$= -8.1924 + 1.2169 a_1 - 0.3329 x$$

交互作用モデル				
分散分析表 「定数に 0 を使用」 off				
	自由度	変動	分散	分散比
回帰	3	66.5524	22.1841	11.9721
残差	16	29.6476	1.8530	$=\sigma^2$
合計	19	96.2000		
	自由度	変動	分散	分散比
回帰	2	57.4778	28.7389	12.6171
残差	17	38.7222	2.2778	$=\sigma^2$
合計	19	96.2000		
	自由度	変動	分散	分散比
回帰	3	66.5524	22.1841	11.9721
残差	16	29.6476	1.8530	1.0000
合計	19	96.2000		

主効果モデル分散分析表「定数に 0 を使用」 off				
	自由度	変動	分散	分散比
回帰	2	57.4778	28.7389	12.6171
残差	17	38.7222	2.2778	$=\sigma^2$
合計	19	96.2000		
	自由度	変動	分散	分散比
回帰	2	57.4778	28.7389	12.6171
残差	17	38.7222	2.2778	$=\sigma^2$
合計	19	96.2000		

	係数	標準誤差	t	P -値
θ^{\wedge}_0 切片 x_0	-7.7062	2.4197	-3.1848	0.0058
θ^{\wedge}_1 a_1	-4.0938	2.4197	-1.6918	0.1101
θ^{\wedge}_2 x	-0.3488	0.1062	-3.2825	0.0047
θ^{\wedge}_3 $a_1 x$	0.2351	0.1062	2.2130	0.0418

	係数	標準誤差	t	P -値
θ^{\wedge}_0 切片 x_0	-8.1924	2.6717	-3.0664	0.0070
θ^{\wedge}_1 a_1	1.2169	0.3436	3.5415	0.0025
θ^{\wedge}_2 x	-0.3329	0.1175	-2.8327	0.0115

	df	平方和	平均平方	F	p
回帰($a_1 + x + a_1 x$)	3	66.5524	22.1841	11.9721	0.0002
($a_1 + x$) 主効果	2	57.4778	28.7389	15.5096	0.0002
($a_1 x$) 交互作用	1	9.0746	9.0746	4.8973	0.0418
残差	16	29.6476	1.8530	1.0000	
合計	19	96.2000			

となり, 共通の傾きは, $\hat{\theta}'_2 = -0.3329$ であり, それぞれの切片は, $\hat{\theta}'_0 \pm \hat{\theta}'_1 = (-6.9756, -9.4093)$ となる.

表 9 上段の交互作用モデルの分散分析表から, 交互作用を含む残差平方和 $S_{残差}^{(交)}$, 表 9 右下段の主効果モデルの分散分析表から, 交互作用を含まない平方和 $S_{残差}^{(主)}$ が,

$$S_{残差}^{(交)} = \sum_i (d_i - \hat{d}_i^{(交)})^2 = 29.6476 \quad | \quad S_{残差}^{(主)} = \sum_i (d_i - \hat{d}_i^{(主)})^2 = 38.7222$$

求められている. 表 10 に示すように, 交互作用モデルの偏差平方和 $S_{残差}^{(交)}$ と主効果モデルの偏差平方和 $S_{残差}^{(主)}$ のとの差が交互作用の平方和 $S_{A \times x}$

$$S_{A \times x} = S_{残差}^{(主)} - S_{残差}^{(交)} = 38.7222 - 29.6476 = 9.0746$$

となる. これは, 交互作用モデルの残差の平方和 $S_{残差}^{(交)}$ には, 交互作用の平方和 $S_{A \times x}$ が含まれていないために, 交互作用が含まれている主効果モデルの残差平方和 $S_{残差}^{(主)}$ との差を取ることで交互作用の平方和 $S_{A \times x}$ を取り出すことができる.

残差の平方和ではなく回帰の平方和の差からも交互作用の平方和を取り出すことができる. 表 9 上段から交互作用モデルの回帰の平方和は, $S_{回帰}^{(交)} = 66.5524$ であり, 主効果モデルの回帰の平方和は, $S_{回帰}^{(主)} = 57.4778$ である. この差からも交互作用の平方和 $S_{A \times x}$ が

$$S_{A \times x} = S_{回帰}^{(交)} - S_{回帰}^{(主)} = 66.5524 - 57.4778 = 9.0746$$

となり, 残差の平方和の差に一致する. 表 10 に示すように, 交互作用の F 値から p 値を求めると 0.0418 となり, 回帰パラメータに対する p 値に一致する.

交互作用が有意な差ではなく, 平行な 2 本の直線があてはめられれば, A_2 インプラミン群の A_1 プラセボ群に対する薬効は, 2 本の直線間の差が, 切片の差と同じになる. さて, 交互作用が有意となった場合には, 前値 x の大きさにより 2 本の直線間の差は異なり, どのような要約統計量によって薬効を論ずるかが課題となる. 恣意性が入り込まないように, 交互作用が有意となった場合は, 表 8 下段に示した前値 x の総平均 $\bar{x}_\cdot = 22.550$ における回帰直線の推定値の差で評価するなど, 前もって定めておく必要がある.

交互作用がない場合の平行な 2 本の直線間の差は, 前値 x が変化しても同じであり, そのために, 前値 $x=0$ の切片のパラメータの差が群間差となり, その 95%信頼区間から群間差について論ずることができる. ただし, 交互作用がある場合は, 前値 x の変化によって差が変化するので, A_2 インプラミンと A_1 プラセボの 2 本の回帰直線の差 $(\hat{d}^{(交, A_2)} - \hat{d}^{(交, A_1)})$ について, 前値 x を変化させたときの分散が必要となる. 2 本の回帰直線の差は,

$$A_2 \text{の回帰式: } \hat{d}^{(交, A_2)} = (\hat{\theta}_0 - \hat{\theta}_1) + (\hat{\theta}_2 - \hat{\theta}_3)x = -3.6125 - 0.5839x$$

$$A_1 \text{の回帰式: } \hat{d}^{(交, A_1)} = (\hat{\theta}_0 + \hat{\theta}_1) + (\hat{\theta}_2 + \hat{\theta}_3)x = -11.8000 - 0.1136x$$

$$\text{差} \quad \left\{ \begin{array}{l} (\hat{d}^{(交, A_2)} - \hat{d}^{(交, A_1)}) = -2\hat{\theta}_1 - 2\hat{\theta}_3x \\ = -2 \times (-4.0938) - 2 \times 0.2351x \\ = 8.1875 - 0.4702x \end{array} \right.$$

となる. この回帰式の差の推定値の分散は, 合成分散の一般式から

$$Var(\hat{d}^{(交, A_2)} - \hat{d}^{(交, A_1)}) = Var(-2\hat{\theta}_1 - 2\hat{\theta}_3x) = 4Var(\hat{\theta}_1) + 8Cov(\hat{\theta}_1, \hat{\theta}_3)x + 4Var(\hat{\theta}_3)x^2$$

が得られる. これらのパラメータの分散および共分散は, 表 9 上段に示したデザイン行列 \mathbf{X} と交互作用モデルの分散分析表の誤差分散の推定値 $\hat{\sigma}_{(交互)}^2 = 1.8530$ を用いて, パラメータの共分散行列 $\Sigma(\hat{\theta})$ が, $\Sigma(\hat{\theta}) = (\mathbf{X}^T \mathbf{X})^{-1} \hat{\sigma}_{(交互)}^2$ として得られ, $Var(\hat{\theta}_1) = 5.8550$, $Cov(\hat{\theta}_1, \hat{\theta}_3) = -0.2550$, $Var(\hat{\theta}_3) = 0.0113$ となるので, x を変化させ, 分散を計算することができる.

$\Sigma(\hat{\theta}) = (\mathbf{X}^T \mathbf{X})^{-1} \hat{\sigma}_{交互}^2$					
$\Sigma(\hat{\theta}) =$	5.8550	-0.6666	-0.2550	0.0233	$\hat{\theta}^0$
	-0.6666	5.8550	0.0233	-0.2550	$\hat{\theta}^1$
	-0.2550	0.0233	0.0113	-0.0008	$\hat{\theta}^2$
	0.0233	-0.2550	-0.0008	0.0113	$\hat{\theta}^3$
	x_0	a_1	x	$a_1 x$	

表12に示すのは、図5に示したように、前値 x を16から2刻みで30まで増加させた場合の各種の推定値である。ちなみに、 $x=22$ における $(\hat{d}_{A_2, x=22}^{(交)} - \hat{d}_{A_1, x=22}^{(交)})$ の p 値が、 $p = 0.0036$ と有意な差となっている。

図6に示すのは、表11で求められた2本の回帰直線間の差の推定値および差の95%信頼区間である。

前値が総平均 $\bar{x}_{..} = 22.550$ の場合についての予測値は、

$$\hat{d}_{x=22.55}^{(交, A_1)} = -11.8000 - 0.1136 \times 22.550 = -14.36$$

$$\hat{d}_{x=22.55}^{(交, A_2)} = -3.6125 - 0.5839 \times 22.550 = -16.78$$

$$\hat{d}_{x=22.55}^{(交, A_2)} - \hat{d}_{x=22.55}^{(交, A_1)} = -2.42$$

となる。分散は、

$$\begin{aligned} Var(\hat{d}_{x=22.55}^{(交, A_2)} - \hat{d}_{x=22.55}^{(交, A_1)}) &= 4Var(\hat{\theta}_1) + 8Cov(\hat{\theta}_1, \hat{\theta}_3)x + 4Var(\hat{\theta}_3)x^2 \\ &= 4 \times 5.8550 + 8 \times (-0.2550) \times 22.550 + 4 \times 0.0113 \times 22.550^2 = 0.3843 \end{aligned}$$

となり、 t 値および p 値は、

$$t_{x=22.55} = \frac{\hat{d}_{x=22.55}^{(交, A_2)} - \hat{d}_{x=22.55}^{(交, A_1)}}{\sqrt{Var(\hat{d}_{x=22.55}^{(交, A_2)} - \hat{d}_{x=22.55}^{(交, A_1)})}} = \frac{-2.42}{\sqrt{0.3843}} = -3.8981 \quad \left| \quad \begin{aligned} p &= T.\text{dist}.2T(|t|, 20-4) \\ &= T.\text{dist}.2T(|-3.8981|, 16) = 0.0013 \end{aligned} \right.$$

として計算されている。したがって、有意な差があると判定される。前値が $x=22$ の場合も同様に $p = 0.0036$ が得られ、有意な差と判定される。 $x=22$ 以上の場合も同様に有意な差となっている。これらの結果から、 A_2 インプラミン

は、 A_1 プラセボに対し、前値が $x=22$ 以上の場合に有意な差があると判定される。

前値の総平均 $\bar{x}_{..} = 22.550$ における p 値による判定だけでなく、2本の回帰直線の差の95%信頼区間の上限が0となる前値

x_0 が推定できれば、「 x_0 以上であれば、統計的に有意な差があると言える」ので、歯切れがよい。95%信頼区間の上限の式は、

$$\hat{y}_0 = (2\hat{\theta}_1 + 2\hat{\theta}_3 x_0) + t_{0.05}(20-4) \sqrt{4Var(\hat{\theta}_1) + 8Cov(\hat{\theta}_1, \hat{\theta}_3)x_0 + 4Var(\hat{\theta}_3)x_0^2}$$

で与えられるので、 $\hat{y}_0 = 0$ となる x_0 を求めるためには、

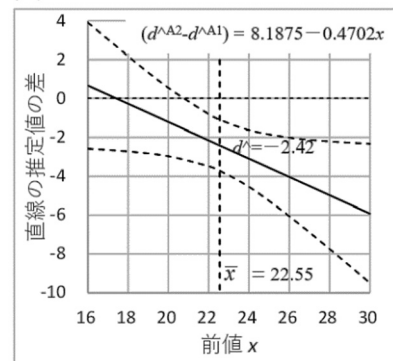
$$(-2\hat{\theta}_1 - 2\hat{\theta}_3 x_0)^2 = t_{0.05}^2(16) \cdot [4Var(\hat{\theta}_1) + 8Cov(\hat{\theta}_1, \hat{\theta}_3)x_0 + 4Var(\hat{\theta}_3)x_0^2]$$

を整理すると x_0 に関して2次式となるので、2次式の解の公式を用いて解を求めればよいが、煩雑な式となる。救いは、Excelのソルバーである。表13に示すのは、 $x=21.0000$ の場合の上限が -0.1916 であり、Excelのソルバーで上限が 0.0000 となるように x を探索的に変化させた結果とし、

$x=20.7260$ が得られた結果である。この結果として、 $p = 0.0500$ と有意な差となる x の推定値が得られたことになる。

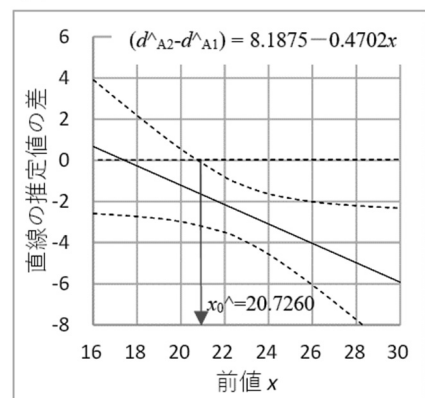
x	$\hat{d}_{A_1}^{(交)}$	$\hat{d}_{A_2}^{(交)}$	差	$Var(\text{差})$	$SE(\text{差})$	$L95\%$	$U95\%$	t 値	p 値
0	-11.80	-3.61	8.19	23.420	4.839	-2.072	18.447	1.692	0.1101
16	-13.62	-12.95	0.66	2.343	1.531	-2.581	3.909	0.434	0.6704
18	-13.85	-14.12	-0.28	1.334	1.155	-2.725	2.172	-0.240	0.8137
20	-14.07	-15.29	-1.22	0.686	0.828	-2.974	0.539	-1.469	0.1611
22	-14.30	-16.46	-2.16	0.400	0.632	-3.498	-0.817	-3.413	0.0036 *
22.550	-14.36	-16.78	-2.42	0.384	0.620	-3.730	-1.102	-3.898	0.0013 *
24	-14.53	-17.63	-3.10	0.474	0.689	-4.558	-1.638	-4.498	0.0004 *
26	-14.75	-18.79	-4.04	0.910	0.954	-6.061	-2.016	-4.233	0.0006 *
28	-14.98	-19.96	-4.98	1.707	1.307	-7.749	-2.209	-3.811	0.0015 *
30	-15.21	-21.13	-5.92	2.865	1.693	-9.508	-2.331	-3.497	0.0030 *

図6



x	\hat{d}_{A_1}	\hat{d}_{A_2}	$\hat{d}_{A_2} - \hat{d}_{A_1}$	$Var(\text{差})$	$SE(\text{差})$	$L95\%$	$U95\%$	t 値	p 値
20	-14.07	-15.29	-1.22	0.686	0.828	-2.974	0.539	-1.469	0.1611
21.0000	-14.19	-15.87	-1.69	0.498	0.706	-3.183	-0.192	-2.391	0.0294
22	-14.30	-16.46	-2.16	0.400	0.632	-3.498	-0.817	-3.413	0.0036 *
20	-14.07	-15.29	-1.22	0.686	0.828	-2.974	0.539	-1.469	0.1611
20.7260	-14.16	-15.71	-1.56	0.541	0.735	-3.117	0.000	-2.120	0.0500 *
22	-14.30	-16.46	-2.16	0.400	0.632	-3.498	-0.817	-3.413	0.0036 *
ソルバー解					$t_{0.05}(20-4) = 2.1199$				*: $p < 0.05$

図5



4. SAS の GLM プロシジャによる検証

多くの統計解析に関連する成書・文献・解説書では、共分散分析を含めて伝統的な解析方法を示し、統計ソフトの使い方と出力結果の見方が示されるのが常である。統計ソフトが、伝統的な解析方法に準じていれば問題はないのであるが、SAS の GLM プロシジャ、JMP の「モデルのあてはめ」、R の `lm()` 関数、および、`lsmeans` パッケージなど先進的な統計ソフトでは、デザイン行列 X を用いた線形モデルによる解析が行われ、最小 2 乗平均 (LSMEANS) のように、新たな統計量が提示されたときに、その扱いに苦慮することになる。

最小 2 乗平均をきちっと説明するためには、デザイン行列 X を用いた線形モデルによる行列計算が欠かせない [高橋ら (1989)]。ただし、GLM プロシジャが内部で生成するデザイン行列 X は、質的変数に対し最後の水準を基準とする (1, 0) オフセット型ダミー変数を採用しているので説明しにくい。そこで、JMP の「モデルのあてはめ」の内部で生成している (1, -1) 対比型ダミー変数を用いて、交互作用のある抗うつ剤の Hamilton 評価尺度の前後差のデータに対し、表 12 に示した様に Excel の行列関数を用いて最小 2 乗平均に対する群間比較の方法を示した。

表 8 に示したと同様の SAS データセットを作成し、交互作用ありの共分散分析を行ない、`lsmeans` ステートメントで薬剤群 A に対する最小 2 乗平均と 2 群間の出力させる。さらに、`at` オプションを使い、前値 x の総平均 $\bar{x}.. = 22.5500$ を明示的に示した場合の最小 2 乗平均、群間差の 95% 信頼区間の上限が 0.0 となる $\bar{x}.. = 20.7260$ となる場合も設定する。

表 14 に示すのは、GLM プロシジャによるタイプ I の平方和を用いた分散分析表であり、Excel による表 10 の分散分析表に対応し、交互作用が $p = 0.0418$ となることが再現されている。このように、交互作用の存在下でも `lsmeans` ステートメントで最小 2 乗平均を求めることができる。

表 15 に示すように、薬剤 A でのデフォルトの最小 2 乗平均は、(-14.3625, -16.7789) であり、表 12 に示した $x = 22.55$ の $\hat{d}_{A1}^{(交)}$ および $\hat{d}_{A2}^{(交)}$ に一致する。もちろん薬剤 A の群間差についても符号が逆であるが一致する。このことから、交互作用を含めた場合でも最小 2 乗平均は、前値 x の総平均における 2 本の回帰直線上の推定値であることが確認される。念のために `at` オプションで $x = 22.5500$

```
/* SAS プログラム 交互作用あり共分散分析 */
proc glm data=d04 ; /* 交互作用モデル */
  class A ;
  model d = A x A*x / solution SS1 XpX inverse ;
  lsmeans A /          stderr cl tdiff pdiff ;
  lsmeans A / at x=22.5500 stderr cl tdiff pdiff ;
  lsmeans A / at x=20.7260 stderr cl tdiff pdiff ;
  output out=o01 p=d_hat stdp=SE lcl=L95 ucl=U95 ;
run ;

proc glm data=d04 ; /* 主効果モデル */
  class A ;
  model d = A x / solution SS1 ;
  lsmeans A /          stderr cl tdiff pdiff ;
  lsmeans A / at x=22.5500 stderr cl tdiff pdiff ;
run ;
```

表14 GLMプロシジャによる分散分析表(元の出力を整形)

要因	自由度	平方和	平均平方	F 値	Pr > F
Model	3	66.5524	22.1841	11.97	0.0002
Error	16	29.6476	1.8530		
C. Total	19	96.2000			
Type I					
要因	自由度	平方和	平均平方	F 値	Pr > F
A	1	39.2000	39.2000	21.16	0.0003
x	1	18.2778	18.2778	9.86	0.0063
x*A	1	9.0746	9.0746	4.90	0.0418
A+x	2	57.4778	28.7389		

表15 交互作用がある場合の各種の最小2乗平均(元の出力を整形)

最小 2 乗平均 : x: デフォルト						
A	最小2乗平均	標準誤差	95% 信頼限界		t 値	Pr > t
A1	-14.3625	0.4378	-15.2906	-13.4344	-	<.0001
A2	-16.7789	0.4388	-15.2906	-15.8486	-	<.0001
A1-A2	2.4164	-	1.1023	3.7305	3.9	0.0013
最小 2 乗平均 : x=22.55						
A	最小2乗平均	標準誤差	95% 信頼限界		t 値	Pr > t
A1	-14.3625	0.4378	-15.2906	-13.4344	-	<.0001
A2	-16.7789	0.4388	-17.7092	-15.8486	-	<.0001
A1-A2	2.4164	-	1.1023	3.7305	3.9	0.0013
最小 2 乗平均 : x=20.726						
A	最小2乗平均	標準誤差	95% 信頼限界		t 値	Pr > t
A1	-14.1552	0.4685	-15.1484	-13.1621	-	<.0001
A2	-15.7139	0.5667	-16.9151	-14.5126	-	<.0001
A1-A2	1.5587	-	0.0000	3.1173	2.12	0.05

を指定した場合に一致する. 表 13 に示した水準間の差が $p = 0.05$ となる前値 $x = 20.7260$ を at オプションで与えると結果が再現される.

表 16 に示すのは, 交互作用を除いた主効果モデルでの薬剤 A に対する lsmeans ステートメントの最小 2 乗平均であり, at オプションでの前値 x の指定がないので, デフォルトの総

最小 2 乗平均: x: デフォルト						
A	最小2乗平均	標準誤差	95% 信頼限界		t 値	Pr > t
A1	-14.4831	0.4816	-15.4992	-13.4670	-	<.0001
A2	-16.9169	0.4816	-17.9330	-15.9008	-	<.0001
A1-A2	2.4338	-	0.9839	3.8837	3.54	0.0025

平均 $x = 22.5500$ が用いられている. もちろん at オプションで $x = 22.5500$ と明示した場合に一致する. これは, 表 15 上段の交互作用を含めた場合とほぼ同様の推定値が得られている. これは, 交互作用をモデルに含めた場合に, 前値 x の総平均を重心とみなして傾きを変化させるような平行線を求めていると解される. 平行な直線に傾きが異なる直線の場合も同様に前値 x の総平均を重心とみなして傾きを変えていると理解される. さて, 薬剤 A の $x = 22.5500$ における p 値は, 交互作用を含めた場合に $p = 0.0013$ であるが, 交互作用を無視した主効果モデルでは, $p = 0.0025$ と僅かであるが大きくなっている. このことは, 無理して平行線をあてはめる必要性が乏しいことを意味している.

5. 考察

伝統的な共分散分析では, 平行線があてはまる場合に, 切片の差が群間の差とみなせることを前提に解析法が定着している. 平行でない場合には, 前値 x の大きさにより群間差が異なり比較ができないと割り切っている. SAS の GLM プロシジャが提示した最小 2 乗平均は, 前値 x の総平均における群間差をデフォルトとしていて, 交互作用を含めた場合であっても, 同様の考え方で最小 2 乗平均が計算されていることを明らかにした.

伝統的な共分散分析では, なぜ交互作用がある場合の解析方法が示されなかったのであろうか. 異なる切片と異なる傾きを持つ回帰直線のあてはめは容易であり, 前値 x の総平均における回帰直線上の“最小 2 乗平均”の推定は容易にできる. また, それらの“最小 2 乗平均”の分散を別々に求め, 互いに独立であることを前提に定分散の加法性より定式化すれば済むように思われる. しかしながら, 複数の回帰直線を同時にあてはめ, 共通の分散を求めることは, 手計算の時代では, 行列計算は不可能であり, また, 群ごとの“最小 2 乗平均”の差の分散を求めるためには, 表 11 に示したパラメータの共分散行列が必要となることもあり, 平行であることを大前提にせざるをえなかったのであろう.

参考文献

- 1) 奥野忠一, 久米均, 芳賀敏郎, 吉澤正 (1981), 多変量解析法《改訂版》, 112-23, 日科技連出版社.
- 2) 宮原英雄, 丹後敏郎 編 (1995), 医学統計学ハンドブック, 144-50, 朝倉書店.
- 3) 朝香鐵一, 石川馨, 山口襄 共同監修 (1988), 新版 品質管理便覧 第 2 版, 174-81, 917-19, 日本規格協会.
- 4) アーミテージら著, 椿ら 共訳 (2001), 医学研究のための統計的方法, 原著第 3 版, 267-72, サイエンティスト社.
- 5) 青木繁伸 (2004), R による統計処理, 共分散分析, <http://aoki2.si.gunma-u.ac.jp/LaTeX/cov/cov.pdf>
- 6) フライス著, KR(関西臨床データ解析)研究会 訳 (2004), 臨床試験のデザインと解析, 189-223. アーム.
- 7) 浜田知久馬 (2009), SAS による共分散分析, SAS ユーザー総論文集, 301-337, Web で検索できる.
- 8) 芳賀敏郎 (2014), 医薬品開発のための統計解析 第 2 部 実験計画法 改訂版, 133-51, サイエンティスト社.
- 9) 中澤港 (2018), 保健・医療研究の進め方入門-R と EZR を用いて-, 125-30, <https://minato.sip21c.org/ebhc-text.pdf>
- 10) 高橋行雄, 大橋靖雄, 芳賀敏郎 (1989), SAS による実験データの解析, 307-33, 東京大学出版会.
- 11) 高橋行雄 (2024a), 層別因子を含む探索的な回帰分析入門, 195-203, カクワークス社.

第 6 章 伝統的な共分散分析からの脱却 <https://www.yukms.com/biostat/takahasi2/rec/012-06.htm>

- 12) 高橋行雄 (2024b), 層別因子を含む探索的な回帰分析入門, 30-37, カクワークス社.

第 1 章 層別因子を含む各種の回帰分析の実際, <https://www.yukms.com/biostat/takahasi2/rec/012-01.htm>

古くてあたらしい共変量調整

～ 共変量調整の現在・過去・未来 at 2024 ～

○森田 祐介, 太田 裕二, 浜田 泉

(ノーベルファーマ株式会社 データサイエンス部)

Old and New Covariate Adjustment

- The Statistical Journey of Covariate Adjustment: Past, Present, and Future (2024) -

Yusuke Morita, Yuji Ohta, Izumi Hamada

Data Science Department, Nobelpharma Co., Ltd.

要旨

ランダム化比較臨床試験の解析では、共変量の不均衡に伴うバイアスを補正し、治療効果の推定精度を向上させるために共変量調整が行われてきた。2023年には共変量調整に関するFDAガイダンスが発出され、重要なポイントが示された。例えば、①オッズ比やハザード比といった要約指標がもつ Non-collapsibility への注意喚起、②Non-collapsibilityに関連して、共変量で条件つけた Conditional な治療効果と、共変量で条件づけない Unconditional な治療効果を区別する必要性、③共変量を考慮した Unconditional な治療効果の推定方法として、G-computation (Standardization) や IPTW 法の紹介、④層別割付などの割付手法を考慮した標準誤差の推定の推奨、⑤モデル誤特定に対してロバストな標準誤差の推定手法の推奨である。共変量調整に関する研究の進展とそれに対応する R パッケージの開発が進む中で、FDA ガイダンスにおいてもこれらの進展が反映されている側面がある。本稿では、SAS を利活用しつつ、共変量調整に関する規制文書や解析手法を紹介し、これまでの状況を整理するとともに、未来を展望する。

キーワード：共変量調整, covariate adjustment, non-linear model, non-collapsibility, unconditional effect, G-computation, ANCOVA, CAUSALTRT, AIPW, TMLE, PROCOVA

1. はじめに

本章では、共変量調整について概説し、正規分布に従う連続量のアウトカム（応答変数）を想定した共分散分析（ANCOVA）の事例を用いて、共変量調整の利点を確認する。

1.1. 共変量調整とは

共変量は、臨床試験や観察研究において治療効果や他の主要なアウトカムに影響を与える可能性がある質的または量的な変数を指す。共変量には年齢、性別、疾患の重症度、主要な変数のベースライン値などが含まれることが多い。ICH E9 ガイドラインでは、治療効果の正確な評価を目的として、統計解析において共変量を考慮することが推奨されている[1]。また、試験デザインの段階では、共変量を割付因子として考慮す

る場合もあり、これにより群間での共変量のバランスを保つことが期待される。このように共変量を適切に管理することで、治療効果の推定精度を向上させ、共変量の不均衡によるバイアスを減少させる効果が期待される。

1.2. Simulation による共変量調整の効果確認

本節では、架空の事例を用いて、共変量調整を行った場合と行わなかった場合の結果をシミュレーションにより比較し、共変量調整の利点について確認する。

まず、シミュレーションで想定するデータおよび試験デザインを下表に示す。なお、帰無仮説下で第一種の過誤確率が制御されていることを確認するため、両群の平均値が完全に等しい状況についても評価を行った。

想定するデータ

投与群	ベースライン	最終観察時点	被験者内相関係数
Active	140 ± 10	130 ± 10	0.5
Placebo	140 ± 10	140 ± 10	0.5

平均値±標準偏差

被験者内相関係数（Pearson）は測定値の標準偏差と変化量の標準偏差が等しくなるよう便宜的に 0.5 とした

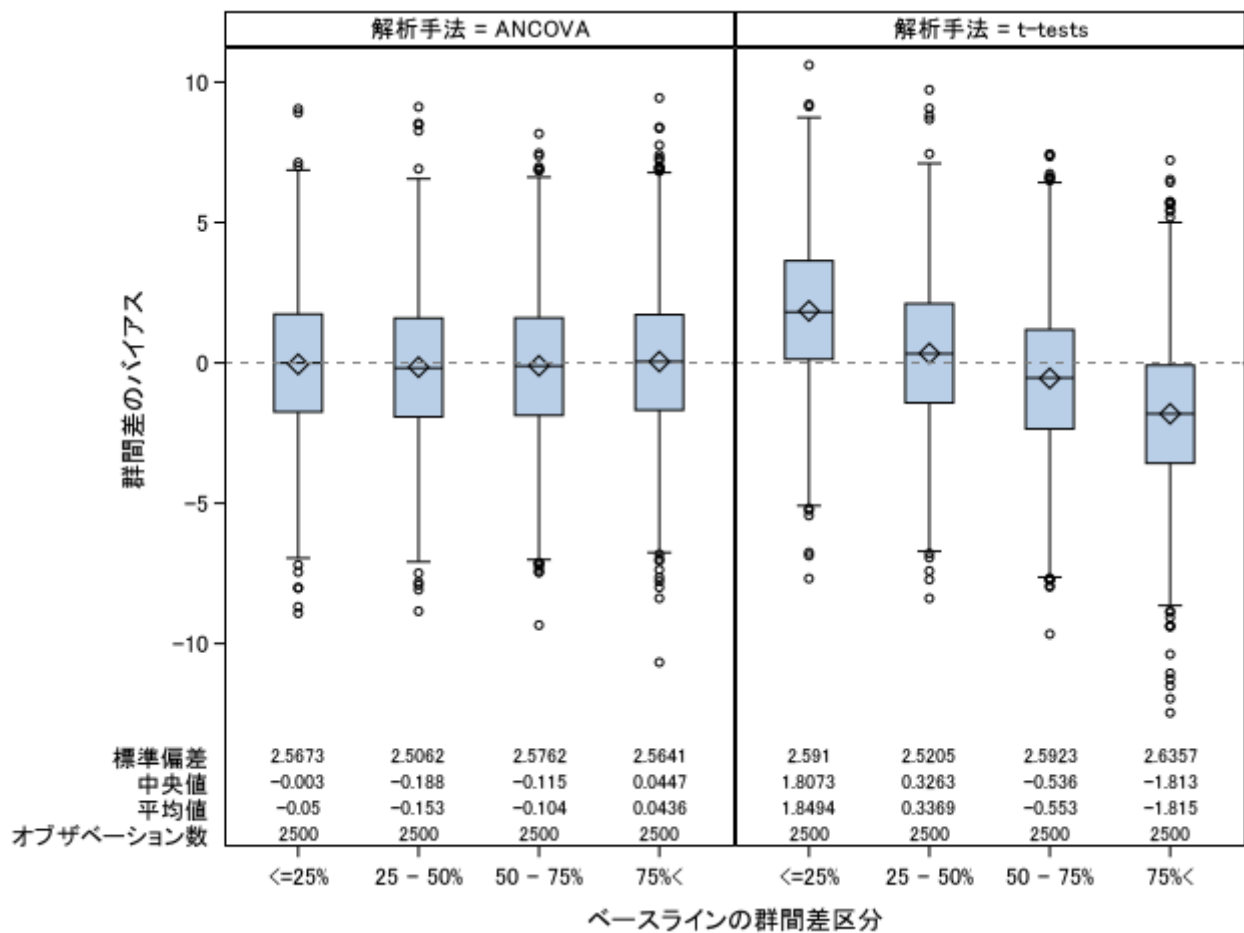
想定する試験デザイン／シミュレーション条件

主要評価項目	最終観察時点のベースラインからの変化量
要約指標	変化量の平均値の群間差（Active - Placebo）
割付比率	1 : 1
被験者数（検出力）	1 群 23 例（90%）
有意水準	両側 5%
繰り返し数	10,000 回
解析手法	① t 検定（Student の t 検定） ② ANCOVA（回帰モデル：変化量＝切片＋投与群＋ベースライン値）
その他	中止・脱落および欠測データは発生しないものとする
評価指標	検出力（第一種の過誤確率）、バイアス(推定値と真の群間差との差)、群間差のサンプル分散、95%信頼区間の被覆割合、群間差の標準誤差の平均、など

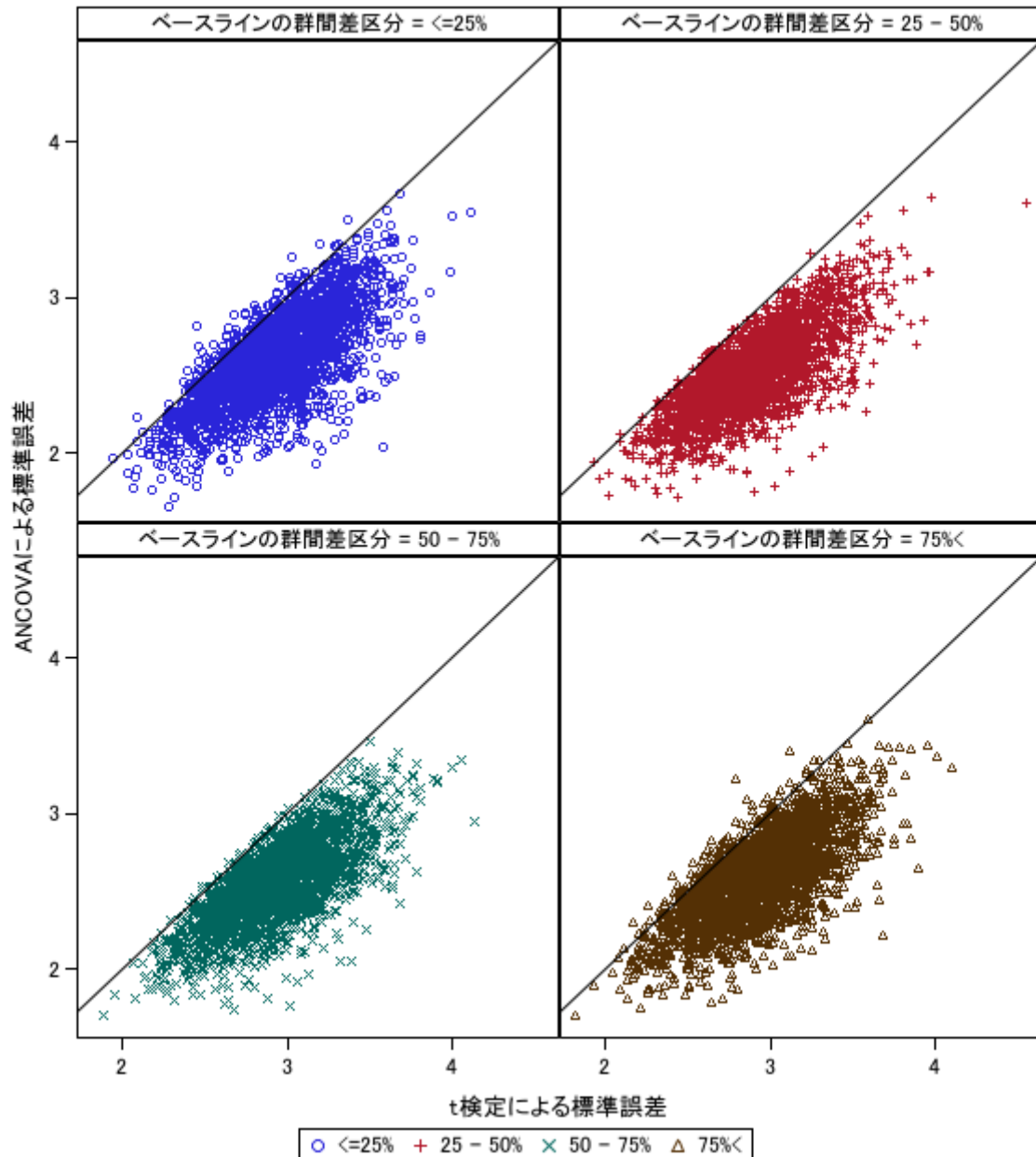
下表にシミュレーションの結果を示す。まず ANCOVA も t 検定も帰無仮説下で、第一種の過誤確率（検出力）は 5% 以下に制御されていた。対立仮説下では、ANCOVA の方が t 検定よりも検出力が上回っていた。これは、ANCOVA の方がベースラインの群間差に起因するバイアスが補正されることや誤差の一部をベースライン値で説明することで、t 検定よりも推定精度が向上するためと考えられる。

条件	解析手法	検出力	バイアス	群間差の サンプル分散	95%信頼区間 の被覆割合	群間差の 標準誤差の平均
帰無仮説下 (群間差=0)	ANCOVA	4.7%	-0.066	6.524	95.3%	2.568
	t検定	4.5%	-0.045	8.463	95.5%	2.929
対立仮説下 (群間差=10)	ANCOVA	97.0%	-0.066	6.524	95.3%	2.568
	t検定	92.0%	-0.045	8.463	95.5%	2.929

以下に、ベースライン値の群間差の4分位点に基づき区分したサブグループ別に群間差のバイアスを要約した箱ひげ図を示す。t検定では、ベースライン値の群間の不均衡が、主要評価項目である変化量の群間差にもバイアスをもたらすことが示されている。一方で、ANCOVAでは、ベースライン値を共変量にすることにより、変化量の群間差のバイアスが解消されている。



同様に、サブグループ別に、各試験の群間差の標準誤差の散布図（X軸：t検定の標準誤差，Y軸：ANCOVAの標準誤差）を示す。どのサブグループにおいても、t検定の標準誤差よりもANCOVAの標準誤差がおおよそ小さくなることが示されている。つまり、ベースライン値の群間差の大小によらず、ベースライン値を共変量として考慮することで群間差の推定精度が向上することが示されている。



以上、ANCOVAによる共変量調整によって、①ベースラインでの群間の共変量の不均衡が調整され、治療効果のバイアス補正が期待できること、②共変量のバランスの程度にかかわらず、治療効果の推定値の標準誤差が小さくなり、推定精度や検出力の向上が期待できることを確認した[2].

なお、本章では連続量アウトカムを例に説明したが、例えば二値アウトカムの場合にはロジスティック回帰モデル、生存時間アウトカムの場合にはCox比例ハザードモデル、発生頻度に関するアウトカムの場合にはポアソン回帰モデルや負の二項回帰モデルを用いた共変量調整が行われる。

2. 共変量調整に関する規制上の要求・推奨事項

本章では、医薬品開発における共変量調整に関する主要な規制ガイドライン・ガイダンスを紹介する。

2.1. 共変量調整に関する規制文書

本節では、共変量調整に関する主要な規制文書を紹介する。これらの文書は、共変量調整の理論的背景、重要な考慮事項、および具体的な方法論を提供しており、臨床試験の統計解析担当者が適切な手法を選択する際の指針となっている。

1. ICH E9 臨床試験における統計的原則 (1998 年) [1]

ICH E9 は、臨床試験における統計的原則を詳細に説明しており、特にランダム化比較試験の信頼性と妥当性を確保するための原則に焦点を当てている。この文書は、共変量調整を治療効果の推定精度向上のための重要な手段として推奨しており、共変量の選定や解析計画における取り扱いに関する指針を強調している。具体的には、共変量はランダム化後のバイアスを減少させ、結果の精度と効率を向上させるために利用されるが、ランダム化後に測定された共変量は、治療効果に影響される可能性があるため、調整に用いるべきではないと注意喚起されている。

2. EMA Guideline on Adjustment for Baseline Covariates in Clinical Trials (2015 年) [3]

欧州医薬品庁 (EMA: European Medicines Agency) のガイドラインは、臨床試験における共変量調整の重要性を強調しており、特に主要評価項目との相関が強い共変量を事前に特定して調整することで、解析の効率と精度を向上させることができるとしている。層別割付が行われた場合、その要因は主要評価項目のモデルに共変量として含めるべきであり、事後的な不均衡を理由に共変量を追加することは推奨されていない。また、モデルの仮定の確認や感度分析の重要性、調整された解析結果と未調整の解析結果を比較して整合性を確認する必要性、そしてプロトコルに沿った報告の重要性などが述べられている。

3. FDA Guidance for Industry: Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products (2023 年) [4]

アメリカ食品医薬品局 (FDA: Food and Drug Administration) によるこのガイダンスは、ランダム化比較臨床試験における共変量調整に関する具体的な解析指針を提供している。共変量調整の方法としては、線形回帰や非線形モデル (ロジスティック回帰モデル、Cox 回帰モデル) など、さまざまな統計モデルが言及されており、特に、非線形モデルでの共変量調整に関しても詳細に説明されている。また、モデル誤特定に対処するための標準誤差の計算方法についても言及されており、具体的に解析手法も提案されている。

いずれの規制文書も、共変量調整が統計解析の精度と妥当性を向上させることを認識し、共変量調整を推奨している。共変量は試験の計画段階で事前に特定し、プロトコルや解析計画書に明記する必要がある。ラ

ランダム化前に測定されるべきだとされている。ランダム化後に測定された変数は治療の影響を受ける可能性があるため、調整に使用すべきでないことが強調されている。

2.2. FDA ガイダンスのポイント

本節では、FDA ガイダンスの特筆すべきポイントについて概説する。FDA ガイダンスは本稿執筆時点で、共変量調整に関する最新の指針であり、既存の規制文書の中でも最も具体的かつ詳細に記載されている。

2.2.1. Non-collapsibility への注意喚起

Non-collapsibility（併合不可能性）とは、集団全体の効果指標が、共変量で規定されたサブグループの効果指標の重み付け平均とは異なる現象を指す。下表に示すように、FDA ガイダンスでは、バイオマーカー陽性・陰性のオッズ比（効果指標）がそれぞれ 8.0 であり、陽性と陰性の割合がそれぞれ 50% であっても、集団全体のオッズ比は 4.8 となる事例が提示されている（ $4.8 \neq 0.5 \times 8.0 + 0.5 \times 8.0$ ）。

一方、リスク差は集団全体で 33.8% であり、これは陽性と陰性のリスク差をサブグループの例数で重み付けした平均と一致する（ $33.8\% \approx 0.5 \times 46.7\% + 0.5 \times 21.0\%$ ）。同様に、リスク比も、各サブグループのプラセボ群の奏効例数の割合で重み付けすることにより、全体集団のリスク比は 2.81 となり、各サブグループのリスク比の重み付け平均と一致する（ $2.81 \approx 2.40 \times (100 \div 112) + 6.25 \times (12 \div 112)$ ）。このように、FDA ガイダンスは、臨床試験や観察研究において広く使用されるリスク差とリスク比は Collapsibility を持つ一方で、オッズ比やハザード比は Non-collapsibility を持つことを注意喚起している。なお、Wei らが、臨床試験で用いられる治療効果の要約指標について、Collapsibility の有無および関連する解析手法をまとめている [5]。

バイオマーカー	割合 N	奏効割合		オッズ比 (Non-collapsible)	リスク差 (Collapsible)	リスク比 (Collapsible)
		新薬	プラセボ			
陽性	50% N=600	80.0% 240/300	33.3% 100/300	8.0	46.7%	2.40
陰性	50% N=600	25.0% 75/300	4.0% 12/300	8.0	21.0%	6.25
合計	100% N=1200	52.5% 315/600	18.7% 112/600	4.8	33.8%	2.81

2.2.2. Conditional / Unconditional な治療効果の区別の必要性

FDA ガイダンスでは、臨床試験における治療効果の評価に際し、共変量で条件付けた「Conditional な治療効果」と、共変量で条件付けない「Unconditional な治療効果」を明確に区別することの重要性が指摘されている。Conditional な治療効果とは、共変量（例えば年齢、性別、基礎疾患など）によって規定される特定の患者群やサブグループにおける治療効果を指す。この効果は、特定の集団に適用されるため、その集団の特性に基づいて治療の有効性や安全性を評価する際に重要である。一方で、Unconditional な治療効果は、共

変量によって条件付けを行わず、対象とする集団の患者全体に対する平均的な治療効果を示すものであり、広範な患者集団における一般的な治療効果を評価する際に使用される。

ロジスティック回帰や Cox 回帰といった非線形モデルでは、Non-collapsibility に起因して、共変量で条件付けた場合と条件付けない場合で、治療効果の推定値および解釈自体が異なってくる。このため、FDA ガイダンスでは、試験計画を立案する際に、推定したい治療効果が条件付きか条件付きでないかを明確にし、それに対応する Estimand を定義することが推奨されている。

なお、FDA ガイダンスでは言及されていないが、変量効果（ランダム効果）の指定も治療効果の解釈に影響を与える。例えば、経時データ解析で被験者内の相関を考慮するため、被験者を変量効果として指定する場合、被験者で条件づけた“Subject-specific な治療効果”が推定される[6]。このように、解析を行う際には、共変量や変量効果の指定が解析結果に与える影響を認識し、慎重に検討する必要がある。

2.2.3. Unconditional な治療効果の推定手法としての G-computation の紹介

FDA ガイダンスでは、アウトカムが二値（例えば、奏効/失敗、治癒/治癒せず）である場合に、治療効果に影響を与える共変量を考慮しつつ、共変量で条件付けない Unconditional な治療効果を推定するための解析手法として、G-computation（Standardization）が詳述されている。G-computation の詳細な手順やその実装方法については、第 3 章で紹介する。

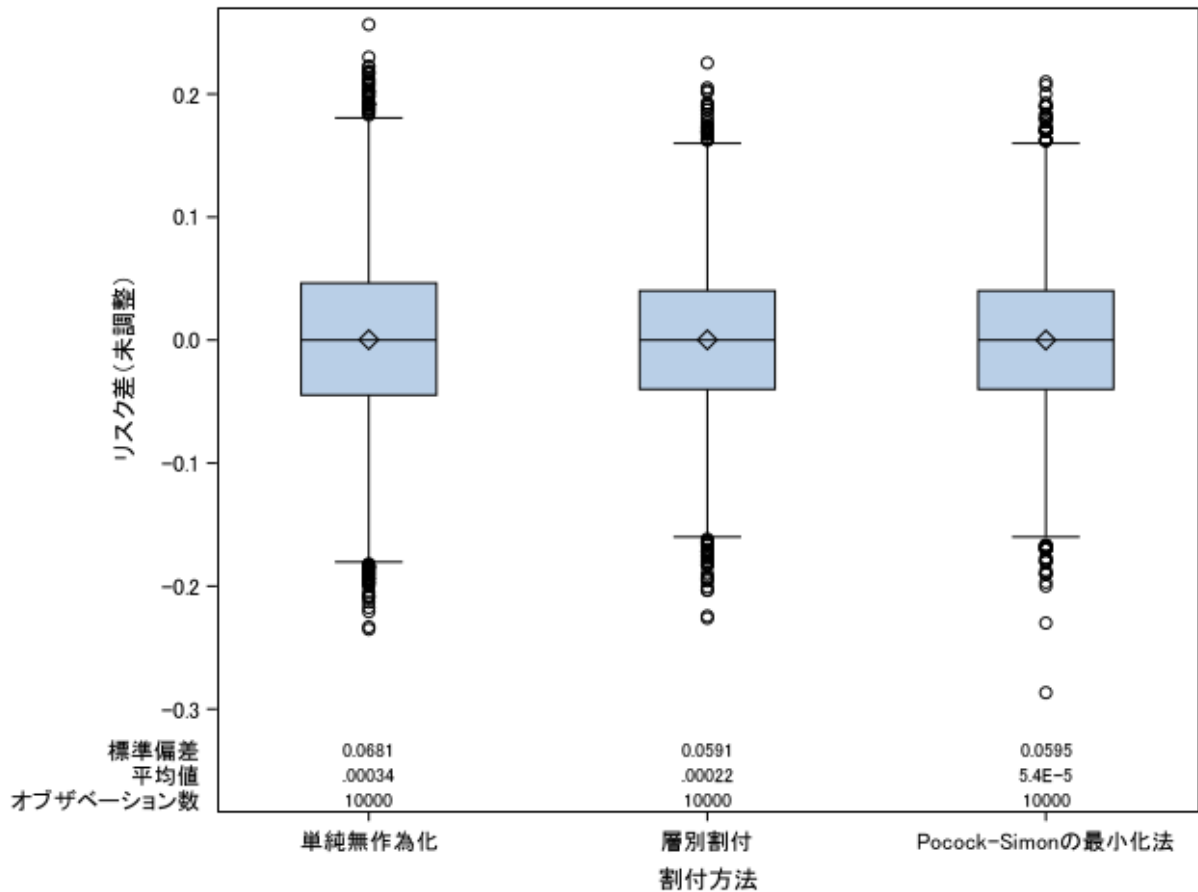
2.2.4. 割付手法を考慮した標準誤差の推定の推奨

臨床試験の解析段階ではなく計画段階で共変量を考慮するために、層別割付や最小化法による割付が行われる。これらの割付方法は、特定の共変量を考慮して被験者を各投与群に割り当てることで、その共変量の群間のバランスを取ることを目的としている。FDA ガイダンスでは、層別割付を考慮せずに解析した場合、標準誤差が過大推定され、Unconditional な治療効果の推定が過度に保守的になるリスクが指摘されている。このため、FDA ガイダンスは、標準誤差の計算において割付手法を考慮することを推奨している。

以下では、シミュレーションにより、特定の共変量を持つ集団に対して異なる割付手法を適用した際の未調整解析に基づくリスク差の分布を確認した。その結果、層別割付や Pocock-Simon の最小化法[7]を用いた場合、単純無作為化法と比較してリスク差のばらつきが小さいことが確認された。これは共変量を考慮した割付手法によって共変量の群間の不均衡が抑えられるためと考えられる。これらの結果から、標準誤差の推定に割付手法を考慮する必要性が示唆された。

シミュレーションの条件：

2 値のカテゴリカルな共変量 Z_1 と Z_2 を持つ被験者 200 名を発生させた。これらの共変量はそれぞれ独立にベルヌーイ分布 Bernoulli(0.5)に従うと仮定する。その後、①単純無作為化法、②層別割付（ Z_1 と Z_2 を層別因子として使用）、③Pocock 最小化法（ Z_1 と Z_2 を割付因子として使用）を用いてそれぞれ割付を行い、投与群 $A = \{0, 1\}$ を決定した。そして、データ発生モデルとして、 $\text{logit}(\Pr(Y=1|A, Z_1, Z_2)) = -1 + 1.5Z_1 + 2.0Z_2 + 0.4$ を用いて二値型応答変数 Y を設定し、未調整のリスク差を求めた。これを 10,000 回繰り返し、割付手法別にリスク差の分布を要約した。



なお、長谷川の総説では、割付手法に応じた適切な解析手法の選択が必要であることが解説されている。特に、Pocock-Simonの最小化法など予後因子を考慮した割付が行われる場合には、並べ替え検定を用いることが適切であり、標本が無作為抽出されていることを仮定した解析（t検定や共分散分析など）では対応できない課題に対処できることが指摘されている[8]。FDAガイダンスでも、共変量調整時に並べ替え検定を利用可能との言及がある。

2.2.5. モデル誤特定に対してロバストな標準誤差の推定手法の推奨

FDAガイダンスでは、回帰モデルが誤特定されている場合にも信頼性のある推定を行うため、ロバストな標準誤差の推定方法が推奨されている。特に、2群均等割り付け以外の割付手法では、Huber-Whiteのサンドイッチ分散が推奨されている。この手法は分散の均一性が成り立たない場合やモデルの誤特定があっても信頼性のある推定を行える。サンドイッチ分散はモデルの残差を利用して分散共分散行列を修正し、モデルの誤特定による影響を抑える役割を果たす。これにより、回帰モデルが正確でなくても第一種の過誤確率を漸近的に制御できることが報告されている[9]。さらに、FDAガイダンスでは、モデルと共変量の相互作用を考慮するロバスト推定法や非パラメトリックブートストラップ法も紹介されている。

3. G-computation による Unconditional なリスク差の推定

本章では、共変量を考慮したうえで Unconditional な治療効果を推定する解析手法の一つとして FDA ガイダンスでも紹介されている G-computation の手順と実装方法について、二値アウトカムの事例を用いて説明する。

3.1. 使用するデータ

FDA ガイダンス中の事例を用いる（再掲）。なお、表中の各セルの被験者数については著者らが割合に基づいて設定したものである。

バイオ マーカー	割合 N	奏効割合		オッズ比 (Non-collapsible)	リスク差 (Collapsible)	リスク比 (Collapsible)
		新薬	プラセボ			
陽性	50% N=600	80.0% 240/300	33.3% 100/300	8.0	46.7%	2.40
陰性	50% N=600	25.0% 75/300	4.0% 12/300	8.0	21.0%	6.25
合計	100% N=1200	52.5% 315/600	18.7% 112/600	4.8	33.8%	2.81

上記の表に基づいて、以下のような構造を持つ SAS データセットを作成した。

subjid	biomarker	trtpn	resp
597	1	9	0
598	1	9	0
599	1	9	0
600	1	9	0
601	2	1	1
602	2	1	1
603	2	1	1
604	2	1	1
605	2	1	1

変数名	説明
SUBJID	被験者番号 1~1200 の連番
BIOMARKER	バイオマーカーの結果, 1:陽性, 2:陰性
TRTPN	投与群, 1:新薬, 9:プラセボ
RESP	転帰(アウトカム), 1:奏効, 0:無効

3.2. Non-collapsibility の確認

ロジスティック回帰モデルを用いて、応答変数を奏効/無効の二値、曝露変数に投与群（新薬/プラセボ）、共変量にバイオマーカー（陽性/陰性）を指定した場合と指定しなかった場合の結果を比較し、Non-collapsibility を確認する。

共変量に、バイオマーカーを指定した場合、推定されるオッズ比は 8.0 となる。

```
*Conditional Odds Ratio;
proc logistic data = fda_guidance;
  class trtpn(ref='9') biomarker;
  model resp(event='1') = trtpn biomarker;
run;
```

オッズ比の推定

効果	点推定	95% Wald 信頼限界	
trtpn 1 vs 9	8.000	5.808	11.019
biomarker 1 vs 2	12.000	8.645	16.657

一方、共変量にバイオマーカーを指定しない場合、推定されるオッズ比は4.8となる。

```
*Unconditional (Marginal) Odds Ratio;  
proc logistic data = fda_guidance;  
  class trtpn(ref='9');  
  model resp(event='1') = trtpn ;  
run;
```

オッズ比の推定

効果	点推定	95% Wald 信頼限界	
trtpn 1 vs 9	4.816	3.711	6.249

以上、バイオマーカーの陽性・陰性が同じ割合で、各グループ内のオッズ比が8.0と等しいにも関わらず、バイオマーカーを共変量として指定した場合と指定しなかった場合で推定されるオッズ比に差が生じたことから、Non-collapsibilityが確認された。

3.3. G-computation の紹介

G-computation は共変量を考慮しながら、共変量で条件づけない Unconditional な治療効果を推定可能な手法として FDA ガイダンスでも紹介されている。G-computation は、各被験者が新薬による治療を受けた場合とプラセボによる治療を受けた場合に得られる奏効確率をそれぞれ推定し、それらを要約することで Unconditional な治療効果を導くものである[10,11]。

3.3.1. G-computation の手順

ステップ 1: 作業モデルの適用

まず、新薬群 (TRTPN=1) のデータを用いて、アウトカム RESP を事前に指定したベースライン共変量 BIOMARKER で回帰するロジスティック回帰モデル (切片を含む) を適用する。

ステップ 2: 予測

ステップ 1 のロジスティック回帰モデルに基づいて、全被験者 (プラセボ群も含む) のベースライン共変量 BIOMARKER に基づいて、新薬による治療が行われた場合のアウトカム (奏効確率) を予測する。

ステップ 3: 平均化

ステップ 2 で算出した各被験者に対する予測値 (奏効確率) の平均をとり、全被験者が新薬による治療を受けた場合の平均アウトカム (μ_1) の推定値を得る。

ステップ 1~3 をプラセボ群 (TRTPN = 9) についても同様に行い、プラセボ群の平均アウトカム (μ_0) の推定値を得る。最終的に、これらの推定値を用いて、リスク差など所望の要約指標を計算する。

リスク差 : $\mu_1 - \mu_0$

リスク比 : μ_1 / μ_0

オッズ比 : $\{\mu_1 / (1 - \mu_1)\} / \{\mu_0 / (1 - \mu_0)\}$

なお、FDA ガイダンスでは、ロジスティック回帰モデルを投与群別に適用せずに、投与群も含めたモデルを、全被験者を対象に適用したうえで、平均アウトカムを推定する手順が紹介されている。本節では、Van Lancker らの総説[12]にならって、投与群別に作業モデルを適用する方法を採用した。これにより、(もし存在するとすれば) 投与群と共変量間の交互作用を考慮することが可能となり、推定効率の向上につながることを期待される。

3.4. SAS による G-computation の実装方法

もっともシンプルな実装方法として CAUSALTRT プロシジャが利用できる[13]。本節では、G-computation の理解を促進するため、前節 3.3 の手順にならった LOGISTIC プロシジャによる実装例も併せて紹介する。

3.4.1. CAUSALTRT プロシジャによる実装

以下のプログラムで G-computation が実行できる。投与群別にロジスティック回帰モデルが適用される。

```
proc causaltrt data = fda_guidance method = regadj poutcomemod ;
  class trtpn(ref='9') biomarker;
  psmodel trtpn;
  model resp(event='1') = biomarker / dist = bin link = logit ;
  output out = out_causal potrt = p1 pocnt = p9;
  ods output causaleffects = ce_causaltrt;
run;
```

各群の奏効割合および奏効割合の差が推定される。なお、M推定[14]によるサンドイッチ分散推定に基づくロバスト標準誤差や信頼区間も出力される[15]。なお、ブートストラップ法に基づく標準誤差の推定も可能である。

CAUSALTRT プロシジャ							
因果効果の分析							
パラメータ	処置の水準	推定値	ロバスト標準誤差	Wald	95%信頼限界	Z	Pr > Z
POM	1	0.5250	0.0188	0.4882	0.5618	27.96	<.0001
POM	9	0.1867	0.0153	0.1566	0.2167	12.17	<.0001
ATE		0.3383	0.0228	0.2936	0.3830	14.83	<.0001

3.4.2. LOGISITC プロシジャによる実装

G-computation の理解を深めるため、前節 3.3 の手順にならった LOGISTIC プロシジャによる実装例を紹介する。なお、この実装方法は篠崎のスライド 67 の工夫を参考にした[16]。

まず、投与群別にロジスティック回帰モデルを適用してパラメータ推定を行いつつ、投与群を問わず全被験者の奏効確率の推定値を得るために前処理を行う。

```
data fda_guidance2;
  set fda_guidance;
  if trtpn = 1 then do;
    wt1 = 1;
    wt9 = 0;
  end;
  else do;
    wt1 = 0;
    wt9 = 1;
  end;
run;
```

つづいて、投与群別に、投与群を除いて共変量のみを指定したロジスティック回帰モデルを適用する。対象とする投与群のデータだけでパラメータ推定が行われるように、前処理で作成した重み付け変数をWEIGHT ステートメントに指定している点に留意されたい。

*新薬群に基づく推定値を利用して全被験者が新薬群の場合の奏効確率p1を推定;

```
proc logistic data = fda_guidance2;
  class biomarker;
  model resp(event='1') = biomarker;
  weight wt1; /* estimate only by new-drug group */
  output out = pred1 predicted = p1;
run;
```

*プラセボ群に基づく推定値を利用して全被験者がプラセボ群の場合の奏効確率p9を推定;

```
proc logistic data = fda_guidance2;
  class biomarker;
  model resp(event='1') = biomarker;
  weight wt9; /* estimate only by placebo group */
  output out = pred9 predicted = p9;
run;
```

続いて、各被験者について推定した各投与群の奏効確率およびその差の平均を算出して、各投与群の奏効割合およびその群間差の推定値を得る。

```
data pred;
  merge pred1 pred9;
  by subjid;
  p_dif = p1 - p9;
run;
```

subjid	biomarker	trtpn	resp	wt1	wt9	p1	p9	p_dif
597	1	9	0	0	1	0.7999961	0.333333	0.4666627169
598	1	9	0	0	1	0.7999961	0.333333	0.4666627169
599	1	9	0	0	1	0.7999961	0.333333	0.4666627169
600	1	9	0	0	1	0.7999961	0.333333	0.4666627169
601	2	1	1	1	0	0.2500002	0.040005	0.2099956126
602	2	1	1	1	0	0.2500002	0.040005	0.2099956126
603	2	1	1	1	0	0.2500002	0.040005	0.2099956126

*割合の差の推定;

```
proc means data = pred n mean;
  var p1 p9 p_dif;
  output out = out_logistic mean = u1 u9 riskdiff;
run;
```

CAUSALTRT プロシジャと同様のリスク差の推定値が得られる。

MEANS プロシジャ

変数	ラベル	N	平均
p1	Estimated Probability	1200	0.5249981
p9	Estimated Probability	1200	0.1866690
p dif		1200	0.3383292

この結果からリスク比
やオッズ比も算出可能

なお、このアプローチをとった場合の群間差の標準誤差の推定については、次節で紹介するが、その過程においてパラメータ推定値に対するサンドイッチ分散を推定したい場合は、GLIMMIX プロシジャ (EMPIRICAL ステートメントを指定) を利用できる[17]。(IML や Data ステップで頑張ってもよい。)

```
proc glimmix data = fda_guidance method = quad empirical = hc3;
  class trtpn(ref='9') biomarker;
  model resp(event='1') = trtpn biomarker /dist=bin link=logit solution covb;
run;
```

3.5. G-computation によるリスク差の標準誤差の推定

FDA ガイダンスでは、G-computation による Unconditional なリスク差の点推定方法については詳述されているが、その標準誤差の推定に関しては、ブートストラップ法や関連文献の紹介にとどまっている。一方、Magirr らは、二値アウトカムにおける G-computation 推定時に使用可能な標準誤差の推定方法として、Ge の方法および Ye の方法を紹介し、モデルの誤特定がある場合など、さまざまな状況下でシミュレーションを用いた性能比較を行っている[18]。本節では、まず条件付き分散公式を説明し、それに基づき、Ge の方法、Liu の方法、Ye の方法を概説し、SAS による計算例を示す。また、シミュレーションによる性能比較も行う。

3.5.1. 条件付き分散公式 : Law of Total Variance[19]

条件付き分散公式により、Unconditional なリスク差の分散 $\text{Var}(\widehat{RD})$ を次のように分解できる。

$$\text{Var}(\widehat{RD}) = E[\text{Var}(\widehat{RD} | X)] + \text{Var}(E[\widehat{RD} | X])$$

ここで、 \widehat{RD} はリスク差の推定量、 X は共変量を表す。また、式の第 1 項 $E[\text{Var}(\widehat{RD} | X)]$ は、共変量が固定された状態でリスク差の変動（分散）がどの程度生じるかを表している。つまり、特定の共変量を与えられた場合に、リスク差の推定量がどれだけばらつくかを示す。第 2 項 $\text{Var}(E[\widehat{RD} | X])$ は、共変量自体のばらつきがリスク差の期待値に与える影響を表しており、異なる共変量に応じてリスク差の期待値がどれだけばらつくかを示している。

3.5.2. Ge の方法[20]

Ge の方法は、条件付き分散公式の第 1 項の一部である $\text{Var}(\widehat{RD} | X)$ を推定するものである。まずロジスティック回帰モデルを用いてパラメータ推定値の条件付き分散 $\text{Var}(\hat{\beta} | X)$ を得る。この分散を基にデルタ法を適用して G-computation によるリスク差の標準誤差を推定する。この方法は、共変量を与えられた条件

下での治療効果のばらつきをモデルベースで評価するものであり、モデルが正しく指定されている場合には有効である。しかし、共変量で条件付けない分散 $\text{Var}(\widehat{RD})$ を推定したい場合は、条件付き分散公式の第2項 $\text{Var}(E[\widehat{RD} | X])$ による変動を考慮できていないため、分散が過小評価される可能性が指摘されている [18, 22].

なお、Ge の方法については、SAS IML を用いたマクロが文献[20]の Appendix に紹介されている。本稿では、このマクロを転記して利用した。ただし、マクロ中のロジスティック回帰モデルは、投与群別にモデルを適用するのではなく、投与群も含むモデルを適用する実装になっている点に留意されたい。

3.5.3. Liu の方法[21]

Ge の方法は、モデルベースの分散推定量を使用しており、モデル誤特定の際に分散を過小評価するリスクが存在する。この問題を解決するため、Liu らは、ロジスティック回帰モデルに基づくモデルベースの分散推定量の代わりにロバストなサンドイッチ分散推定量を採用することを提案している。さらに、条件付き分散公式の第2項 $\text{Var}(E[\widehat{RD} | X])$ を考慮し、各被験者のリスク差の推定値のサンプル分散を被験者数で除した項 $\widehat{\sigma}_{RD}^2/n$ を追加することで、共変量の変動に起因するリスク差の不確実性を捕捉している。この2つの工夫により、モデル誤特定の影響を軽減し、信頼区間が真の値を含む被覆確率（カバレッジ）を向上させることがシミュレーションで確認されている。

本稿では、Ge の方法のマクロを改変し、Liu の方法を実装したマクロを作成した。ただし、Ge のマクロと同様に、ロジスティック回帰モデルを投与群別に適用するものではない点に留意されたい。

3.5.4. Ye の方法[22]

Ye の方法は、モデルの誤特定や共変量の非線形性、複雑な分布が推定に与える影響を適切に反映したロバストな分散推定を提案している。この方法では、共変量がリスク差に与える影響を分析し、モデル残差によるばらつきとモデル予測値によるばらつきに分解し、それらを分散の加法性を用いて統合することで推定量を構成している。つまり、この方法は条件付き分散の公式の第1項： $E[\text{Var}(\widehat{RD} | X)]$ と第2項： $\text{Var}(E[\widehat{RD} | X])$ を考慮した推定量となっている。このアプローチにより、モデルが誤特定された場合や共変量の分布が複雑な場合でも、推定の精度が高く、信頼区間の被覆率が名目水準を維持することがシミュレーションで確認されている。

本稿では、Ye の方法が実装されている R パッケージの1つ `becca`[23]を参考にして、SAS へ移植した。

3.5.5. 標準誤差の計算例

以下に、3.1章のデータに対し3.5章で紹介したリスク差の標準誤差の推定方法を適用した結果を示す。

G-computation の 標準誤差の推定方法	リスク差		95%信頼区間	
	推定値	標準誤差	下限	上限
Ge の方法	0.3383	0.0225	0.2942	0.3825
Liu の方法	0.3383	0.0229	0.2935	0.3832
Ye の方法	0.3383	0.0228	0.2936	0.3831
M 推定 (CAUSALTRT)	0.3383	0.0228	0.2936	0.3830
参考：未調整 (正規近似)	0.3383	0.0259	0.2877	0.3890

Geの方法では、標準誤差が0.0225と他の方法より僅かに小さくなったが、ほぼ同じ結果であった。いずれの推定方法でも、未調整解析（正規近似）の場合より標準誤差が小さくなっており、G-computationによってリスク差の推定精度が向上していることがわかる。

なお、CAUSALTRTプロシジャで出力されるリスク差のロバスト標準誤差（3.4.1章）は、M推定によるサンドイッチ分散推定量に基づいているが[14]、本節で紹介した3つの推定方法との違いについては次節で評価する。

3.6. Simulation による標準誤差の推定方法の性能検討

Magirrらは、Geの方法およびYeの方法をシミュレーションで比較している[18]。本稿では、作成したプログラムの点検も兼ねて、Liuの方法、M推定も追加してMagirrらのシミュレーションを一部追試した。

検討したのは、以下の4つのデータ発生シナリオである。応答変数 $Y_i = \{0, 1\}$ 、投与群 $A_i = \{0, 1\}$ 、共変量 $X_i \sim N(0, 3^2)$ の正規乱数として、1試験あたり200例を、2群にランダムに割付け、各シナリオに基づき応答変数を設定した。各シナリオの繰り返し数は10,000回とし、いずれのシナリオでもG-computationの作業モデル $\text{logit}\{P(Y_i = 1|A_i, X_i)\} = \beta_0 + \beta_1 A_i + \beta_2 X_i$ を適用し、前節で紹介した標準誤差の推定方法を用いて計算を行い、95%信頼区間の被覆割合などを算出した。ただし、M推定はCAUSALTRTプロシジャを用いているため、投与群別に作業モデルを適用したG-computationとなっている。

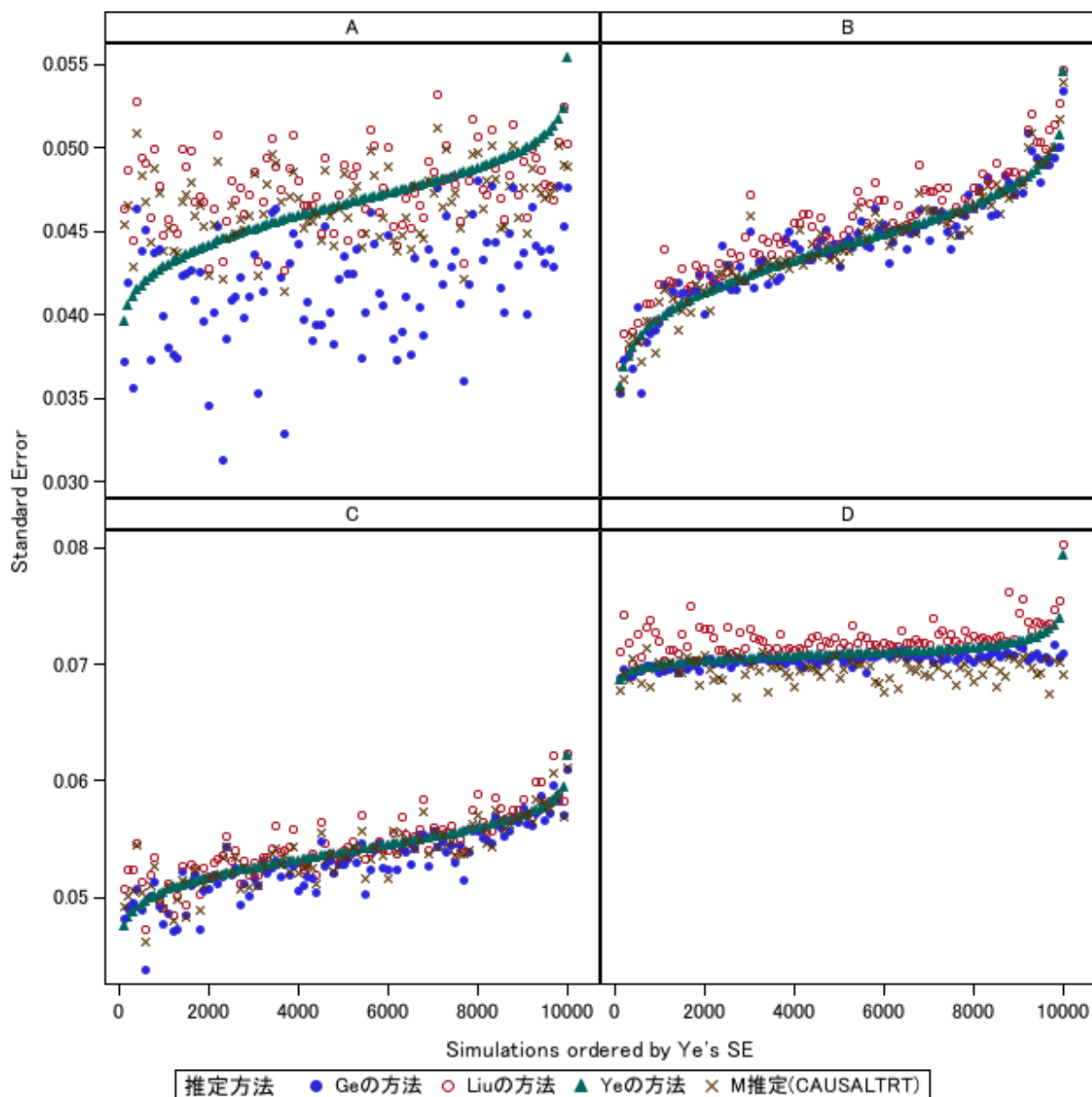
シナリオ	Simulation データの発生式[18]	説明
A	$\text{logit}\{P(Y_i = 1 A_i, X_i)\} = -2 + 5A_i + X_i$	治療効果が極大なシナリオ（オッズ比 ≈ 148 ）
B	$\text{logit}\{P(Y_i = 1 A_i, X_i)\} = -2 + 0A_i + X_i$	治療効果がないシナリオ（オッズ比 $= 1.0$ ）
C	$\text{logit}\{P(Y_i = 1 A_i, X_i)\} = -1.8 + 1.6A_i - 0.6X_i$	治療効果が現実的なシナリオ（オッズ比 ≈ 5.0 ）
D	$\text{logit}\{P(Y_i = 1 A_i, X_i)\} = -1 - 6A_i X_i + 3X_i$	解析モデルが誤特定となるシナリオ（交互作用）

以下にシミュレーション結果を示す。Magirrらの結果と同様に、シナリオAではGeの方法の95%信頼区間の被覆割合が91.58%と名目水準を下回ったが、それ以外のシナリオでは、4つの方法に大きな差異はなく、95%信頼区間の被覆割合も名目水準の95%に近い値であったが、Liuの方法では平均SEが若干大きく保守的な傾向があった。

シナリオ	n	リスク差		Geの方法		Liuの方法		Yeの方法		M推定 (CAUSALTRT)	
		平均	SD	平均 SE	CP(%)	平均 SE	CP(%)	平均 SE	CP(%)	平均 SE	CP(%)
A	200	0.5223	0.0465	0.0415	91.58	0.0474	95.47	0.0464	94.86	0.0463	94.92
B	200	-0.0004	0.0443	0.0437	94.59	0.0449	95.32	0.0439	94.83	0.0438	94.78
C	200	0.2336	0.0536	0.0529	94.27	0.0546	95.10	0.0538	94.75	0.0536	94.76
D	200	-0.0013	0.0710	0.0703	94.40	0.0722	95.06	0.0709	94.63	0.0695	94.65

n : 1試験あたりの例数, SD : 標準偏差, 平均 SE : 標準誤差の平均値, CP : 95%信頼区間の被覆割合

さらに、Magirr らの結果を参考に、シミュレーションの各試験番号を X 軸、リスク差の標準誤差を Y 軸にして、シナリオ別にプロットした結果を以下に示す。なお、表示の都合上、Ye の方法の標準誤差の昇順に試験番号を並び替え、100 番おきにプロットしている（つまり 100 試験分のプロットとなっている）。



シナリオ D で、M 推定（CAUSALTRT プロシジャ）の標準誤差が相対的にやや小さい理由として、データ発生モデルが共変量と投与群の交互作用項を含んでいることが考えられる。このため、投与群別に作業モデルを適用する CAUSALTRT プロシジャの方がデータに適合しやすくなり、標準誤差も小さくなったと推察される。

Magirr らは、Ge の方法で Conditional な治療効果を推定する場合は、95%信頼区間は名目水準の被覆割合を保つことを示し、推定方法の性能差は Estimand の違いに起因することを指摘している。また、Estimand に整合した標準誤差の推定方法を採用することが推奨されるとしている。さらに、Magirr らは条件付き分散の公式を引用し、Ge の方法と Ye の方法で標準誤差の推定値が近似する条件を考察し、治療効果や共変量の影響が現実的な範囲においては、どちらの方法も同等の性能を発揮し得ると述べている。

また Liu ら自身も、Ge の方法、Ye の方法および Liu の方法をさまざまな条件下でシミュレーションにより性能比較している[21]. Liu の方法は、サンプルサイズが小さい場合（30 例、60 例、90 例で評価）や、作業モデルが誤特定されている場合、または割付比が 2:1 など均等でない場合でも、他の方法に比べて信頼区間の被覆割合や第一種の過誤確率について一貫して優れた性能を示したと報告されている。

なお、Magirr らのシミュレーションでは、R の「becca パッケージ」を利用している[23]. becca は、4.1 章で後述する RobinCar パッケージをベースに、主に二値アウトカムのデータを扱うために軽量化されたものであり、G-computation を用いたリスク差などの推定と、その標準誤差の推定を簡便に行える。

4. 未来の展望

4.1. 解析手法の進展

Bannick らは、線形回帰、非線形回帰、機械学習モデル、G-computation などのさまざまな共変量調整法が、Augmented Inverse Probability Weighting (AIPW) のフレームワークの中で統一的に整理できると提案している。AIPW は、さまざまな回帰モデルや機械学習手法に対応し、これらを含む枠組みであり、二重頑健性 (Double Robustness) を持つ。この二重頑健性により、アウトカムモデルまたは傾向スコアモデルのどちらか一方が正しく指定されていれば、治療効果に対するバイアスのない推定が可能である。特に、無作為化試験では治療割り付けがランダムに行われるため、傾向スコアモデルが正確に推定されやすいとされている。また、さまざまな共変量適応的ランダム化 (CAR: Covariate-Adaptive Randomization) のもとで適用可能な新しい分散推定量も提案されており、これらの手法の有効性を理論的裏付けるとともに、シミュレーションでも確認している。さらに Joint Calibration という手法を導入することで、CAR のもとでもロバストな分散推定が可能のように工夫されている[24]. これらの手法は R パッケージ RobinCar (Robust Estimation and Inference in Covariate-Adaptive Randomization) に実装されており、実際の臨床試験データへの適用が容易となっている[25].

また、TMLE (Targeted Maximum Likelihood Estimation) を共変量調整にも活用することで、共変量を調整しない場合よりも漸近分散の小さい、より有効な推定量を構成することができる。TMLE は、因果推論や共変量調整の手法として注目されており、アウトカムの型にかかわらず、良好な性質を持つ推定量が得られる[26]. TMLE では、まずアウトカムモデルを使って治療効果の初期推定量を得た後に、傾向スコアモデルの情報を用いてこの推定を調整する。この際、「クレバー共変量 (Clever Covariates)」と呼ばれる特別な共変量を用いることで、調整が効率的に行われ、バイアスの少ない推定が得られる。また、TMLE はセミパラメトリックなアプローチを採用しており、データの特性に応じて柔軟に適応できるため、さまざまな状況で汎用的に利用できる点も特徴である。なお、臨床試験の共変量調整における TMLE の活用については、田栗らの総説の 6 章で詳しく解説されており、アウトカムモデルに正準リンクを用いた一般化線形モデルを活用した場合には、パラメータ推定のための反復アルゴリズムが不要であり、実用が簡単であることが述べられている[27].

一方、ヒストリカルデータを活用する PROCOVATM (Prognostic Covariate Adjustment) も提案されている[28]. PROCOVATM は、臨床試験における治療効果の推定精度を向上させるために開発された手法であり、主にフェーズ 2 およびフェーズ 3 の臨床試験で、連続量のアウトカムを解析する際に使用されることを意図

している。PROCOVA™は、ヒストリカルデータを基に機械学習モデルを用いて各被験者の予後を予測するスコアを算出し、そのスコアを共変量として調整することで、推定精度を高め、必要な被験者数の低減を図る。PROCOVA™は、ANCOVAの一形態と位置づけられており、バイアスのない治療効果の推定と適切な第一種の過誤確率の制御が可能である。EMAはPROCOVA™に関してDraft Qualification Opinionを発行しており、連続量のアウトカムを持つ臨床試験において、この手法が治療効果推定の精度や検出力を高める可能性があることを認めている[29]。また、FDAからはPROCOVA™はANCOVAの一形態であるとして、既存のFDAガイダンスとは矛盾しないというフィードバックがあったことが報告されている[30]。

4.2. その他の動き

近年の規制や解析手法の進展に伴う各種団体の動きにも注目したい。アメリカ統計協会（ASA）のOncology Estimand Working Group（WG）は、がん領域の臨床試験における治療効果の評価方法を標準化し、がん研究の進展と患者の治療成果の向上を目的とした専門家グループである[31]。このワーキンググループ内には、「Conditional vs. Marginal Effects」をテーマに掲げたサブチームが構成されており、Conditionalな治療効果とMarginal（Unconditionalな）治療効果に関する理解を推進し、臨床試験の設計・解析・解釈の質を向上させることを目指して活動している。その活動成果のひとつとして、実例を用いた各種治療効果の解説と実践上の課題点がまとめられている[5]。また、同じくASAのBiopharmaceutical Sectionには、臨床試験における共変量調整の統計手法の進展と普及を目的として「Covariate Adjustment Scientific WG」が新設された。このグループは、新しい共変量調整法の開発、知識の普及、標準化の推進を目指すとともに、ソフトウェアパッケージの開発、ジャーナルクラブやオンラインフォーラムを通じた情報共有を行っている[32]。

5. まとめ

本邦でも2024年6月に、ICH E9（R1）がステップ5となり[33]、FDAガイダンスでも推奨されるように、Estimandと整合した共変量調整法の選択がますます重要な状況となった。特に二値アウトカムや生存時間アウトカムにおいて、Non-collapsibilityに起因して、共変量で条件づけたConditionalな治療効果と共変量で条件付けないUnconditionalな治療効果のどちらかを推定するかを明確にすることが求められる。また、Estimandや割付手法を考慮した治療効果の標準誤差の推定も必要になってくる。臨床試験の統計解析担当者は、試験計画段階で関係者と協議し、これらの検討をリードする必要があるだろう。

医薬品開発の規制の進展と共に、共変量調整の理論や方法論の発展にも、引き続き注視していく必要がある。また、これらの発展がタイムリーかつユーザーフレンドリーにSASに実装される未来を期待したい。

連絡先：morita.yusuke@nobelpharma.co.jp

Appendix : 本稿に添付する SAS プログラム

プログラム名	内容
t_vs_ancova.sas	1.2 章の t 検定と ANCOVA の Simulation 比較に用いたプログラム
car_simulation.sas	2.2.4 章の割付手法による割合の差の分布確認用プログラム
fda_g-comp.sas	3.2～3.4 章の FDA ガイダンスの事例データに対する Non-collapsibility の確認と G-computation の実装例を示したプログラム
g-comp_riskdiff_se_ge.sas	3.5.2 章の G-computation によるリスク差の標準誤差を推定するプログラム (Ge の方法 : 文献[20]の Appendix より転記)
g-comp_riskdiff_se_liu.sas	3.5.3 章の G-computation によるリスク差の標準誤差を推定するプログラム (Liu の方法)
g-comp_riskdiff_se_je.sas	3.5.4 章の G-computation によるリスク差の標準誤差を推定するプログラム (Ye の方法)
g-comp_riskdiff_se_sim.sas	3.6 章の G-computation によるリスク差の標準誤差を比較する Simulation プログラム

なお、プログラムの動作確認は SAS9.4 TS1M7 で実施した。

引用文献・サイト

- [1] 厚生労働省. 臨床試験のための統計的原則. 厚生労働省; 1998.
- [2] 浜田知久馬. SAS による共分散分析. SAS ユーザー総会論文集 2009. 2009;301-337.
- [3] European Medicines Agency. Guideline on adjustment for baseline covariates in clinical trials. London: European Medicines Agency; 2015.
- [4] U.S. Food and Drug Administration. Guidance for industry: adjusting for covariates in randomized clinical trials for drugs and biological products. Silver Spring (MD): U.S. Food and Drug Administration; 2023.
- [5] Wei J, Xu J, Bornkamp B, Lin R, Tian H, Xi D, et al. Conditional and unconditional treatment effects in randomized clinical trials: estimands, estimation, and interpretation. Stat Biopharm Res. 2024;18(1):1-25. doi: <https://doi.org/10.1080/19466315.2023.2292774>.
- [6] Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. 2nd ed. Hoboken (NJ): Wiley; 2011.
- [7] Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics. 1975;31(1):103-15.
- [8] 長谷川貴大. 統計モデル : 母集団モデルと無作為化モデル. 保健医療科学. 2011 ; 60(1) : 38-43.
- [9] Rosenblum M, van der Laan MJ. Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. Biometrics. 2009;65(3):937-945.
- [10] 佐藤俊哉, 松山裕. 交絡という不思議な現象と交絡を取りのぞく解析: 標準化と周辺構造モデル. 計量生物学. 2011;32(Special Issue).
- [11] Hernán MA, Robins JM. Causal Inference: What If. Boca Raton (FL): Chapman & Hall/CRC; 2020.
- [12] Van Lancker K, Bretz F, Dukes O. Covariate adjustment in randomized controlled trials: General concepts and practical considerations. Clin Trials. 2024;21(3):312-325. doi: <https://doi.org/10.1177/17407745241251568>.
- [13] SAS Institute Inc. The CAUSALTRT procedure. SAS/STAT® 15.2 User's Guide. Cary (NC): SAS Institute Inc.; 2021.
- [14] 田中司朗. 医学のための因果推論 II —Rubin 因果モデル—. 朝倉書店; 2022.
- [15] SAS Institute Inc. The CAUSALTRT procedure: variance estimation for causal effect estimates [Internet]. SAS/STAT® 15.2 User's Guide. Cary (NC): SAS Institute Inc.; 2021 [cited 2024 Aug 31]. Available from: https://documentation.sas.com/doc/en/statug/15.2/statug_causaltrt_details08.htm.

- [16] 篠崎智大. 因果推論 II : モデルを用いた効果推定 [Internet]. Docswell; 2022 [cited 2024 Aug 31]. Available from: <https://www.docswell.com/s/she-knows-a-key/52639Z-2022-09-08-225555>.
- [17] SAS Institute Inc. The GLIMMIX procedure: empirical estimation [Internet]. SAS/STAT® 15.2 User's Guide. Cary (NC): SAS Institute Inc.; 2021 [cited 2024 Aug 31]. Available from: https://documentation.sas.com/doc/en/statug/15.2/statug_glimmix_syntax01.htm.
- [18] Magirr D, Baillie M, Wang C, Przybylski A. Estimating the variance of covariate-adjusted estimators of average treatment effects in clinical trials with binary endpoints [Internet]. 2024 Jun 17 [cited 2024 Aug 31]. doi: <https://doi.org/10.17605/OSF.IO/9MP58>.
- [19] 久保川達也. 現代数理統計学の基礎. 共立出版; 2017. p.66.
- [20] Ge M, Durham LK, Meyer RD, Xie W, Thomas N. Covariate-adjusted difference in proportions from clinical trials using logistic regression and weighted risk differences. *Drug Inf J*. 2011;45(4):481–93.
- [21] Liu J, Xi D. Covariate adjustment and estimation of difference in proportions in randomized clinical trials. *Pharm Stat*. 2024;1-22. doi: <https://doi.org/10.1002/pst.2397>.
- [22] Ye T, Bannick M, Yi Y, Shao J. Robust variance estimation for covariate-adjusted unconditional treatment effect in randomized clinical trials with binary outcomes. *Stat Theory Relat Fields*. 2023;7(2):159-63.
- [23] Przybylski A, Baillie M, Wang C, Magirr D. beeca: Binary Endpoint Estimation with Covariate Adjustment. R package version 1.0. 2024 [cited 2024 Aug 31]. Available from: <https://openpharma.github.io/beeca/>.
- [24] Bannick MS, Shao J, Liu J, Du Y, Yi Y, Ye T. A general form of covariate adjustment in randomized clinical trials. *arXiv* [Internet]. 2024 [cited 2024 Aug 31]. Available from: <https://doi.org/10.48550/arXiv.2306.10213>.
- [25] Bannick MS, Shao J, Liu J, Du Y, Yi Y, Ye T. RobinCar: Robust Inference for Covariate-Adaptive Randomization in Randomized Clinical Trials. R package version 1.0. 2024 [cited 2024 Aug 31]. Available from: <https://cran.r-project.org/web/packages/RobinCar/index.html>.
- [26] Van der Laan MJ, Rose S. Targeted learning in data science: causal inference for complex longitudinal studies. New York (NY): Springer; 2018.
- [27] 田栗正隆, 高橋邦彦, 小向翔, 伊藤ゆり, 服部聡, 船渡川伊久子, et al. 疫学分野での計量生物学の発展. *計量生物学*. 2024;44(2):129-200. doi: <https://doi.org/10.5691/jjb.44.129>.
- [28] Schuler A, Walsh D, Hall D, et al. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *Int J Biostat*. 2021;18(2):329-356.
- [29] European Medicines Agency Committee for Medicinal Products for Human Use (CHMP). Qualification opinion for Prognostic Covariate Adjustment (PROCOVA™). 2022 Sep.
- [30] Fisher CK. US FDA comments on Unlearn's PROCOVA methodology [Internet]. Unlearn Blog. [cited 2024 Aug 31]. Available from: <https://www.unlearn.ai/blog/us-fda-comments-on-unlearns-procova-methodology>.
- [31] ASA Biopharmaceutical Section Oncology Estimand Working Group. Oncology estimand working group webpage [Internet]. 2024 [cited 2024 Aug 31]. Available from: https://oncoestimand.github.io/oncowg_webpage/docs/.
- [32] ASA Biopharmaceutical Section Covariate Adjustment Scientific Working Group. Covariate Adjustment SWG homepage [Internet]. 2024 [cited 2024 Aug 31]. Available from: <https://carswg.github.io/>.
- [33] 厚生労働省. 臨床試験のための統計的原則 補遺 臨床試験における estimand と感度分析. 厚生労働省; 2024.

サンプリングのための標本設計と標本配分

○高田 浩成

(イーピーエス株式会社)

Sample design and sample allocation for sampling

Hiroshige Takata

EPS Corporation

要旨

社会や生態の調査を対象とする統計調査では、全数調査（センサス）を除き、時間や費用、その他の様々な事情から、標本抽出（サンプリング）により一部から全体を推定する。サンプル数が多いほど精度や正確度は高くなるのは、中心極限定理や大数の法則からも明らかであるが、目標精度や区間推定から基準を検討する。層化抽出法（層別サンプリング）は使用頻度の高い抽出方法で、さらに標本配分法として、層内の特定項目の分散を最小にするネイマン配分法がある。より少ないサンプル数による推定精度の向上を比例配分法等と比較して示す。SAS ユーザー総会 2023 においては、可能な限り多種の標本抽出法について、SAS での実装方法を紹介することに重きをおいたが、本稿では実用される頻度が多いと思われる手法にフィーチャーした。

キーワード：層別抽出法、多段抽出法、標本誤差、比例配分法、ネイマン配分法

1. 背景

EBPM (Evidence Based Policy Making) という「証拠に基づく政策立案」の取り組みが推進されている。政策目的を明確化したうえで、その目的のために効果が上がる行政手段を合理的根拠（エビデンス）に基づき企画立案することである。統計データは、エビデンスのうちでも重要なものであり、EBPM の推進に統計データの利活用が注目されている。

また、PPDAC サイクルは、問題解決のプロセスを Problem（問題）、Plan（計画）、Data（データ）、Analysis（分析）、Conclusion（結論）の 5 つのフェーズに分け、各フェーズの頭文字をつなげたものであり、統計調査のプロジェクトにおける手順にも対応していると言える。

これらの背景を踏まえ、サンプリング方法そのもの以外にも他の工程や状況に応じた条件について、確認する。本稿では、とくに標本設計と標本配分を対象とする。

2. 調査の分野

大きく2つの分野に分けて、サンプリングの考え方を述べる。

2.1 社会調査

政府・自治体・調査機関等により作成された統計情報を、調査対象の母集団情報を一覧できるサンプリング台帳として使用し、それに基づき、サンプリング対象をある程度絞るのが基本である。例えば、選挙予測の分野では、当確の速報を出すために、各報道機関が独自に調査を行うが、出口調査の場合、投票所を絞り込み、そのうち一部の投票者に直接アンケート調査を実施する。標本設計の上、多段抽出・層別抽出が行われるといえる。

このように、世の中の動向を捉えることを目的に、サンプリングの負担をできるだけ軽減するために、様々な方法が確立されている。また、個人情報保護等の制度・意識により社会調査が困難になる要因もある時代に、サンプリングを適切に行うためには、データの管理環境や協力者との信頼関係も忘れてはならない重要な要素である。

2.2 生態調査

生態調査は野外生物の社会調査とも言える。調査地域等以外は、人間のように、サンプリング台帳は存在せず、個体の識別が困難であるため、生息数の個体推定法として、様々な方法が存在する。

標識再捕獲法は、1度採集した個体に標識を付け、2回目まで同じ個体の採取を許容し、その中に標識個体の比率から全体個数を推定する（計算法はリンカーン・ペテルセン推定量が代表される）。コドラート法は、移動性の少ないものを対象に、生息域に一定の面積の方形枠を設定し、その内部の個体数を調べることで、全体の個体数や密度を推定する方法であるが、各区画の決定にはランダムサンプリングを行うことになる。

生物の個体数を知ることは資源の見積りや絶滅危惧種の保全において不可欠であり、調査するのが難しい対象にも工夫が検討され、科学的な推定を可能にしている。人間社会と自然環境の関わりは密接であり、今後も観測技術の発展により、サンプリングの考え方に基づいて改善が見込まれる分野でもある。

3. SAS によるサンプリングと精度の確認

3.1 SURVEYSELECT プロシジャ

本稿においては、SAS の SURVEYSELECT プロシジャによるランダムサンプリングを利用した。プロシジャ自身やそのオプション・ステートメントの機能を組み合わせることにより、統計調査に対応した様々なサンプリング方法やグループ分割を実装可能で、多くのデータを処理でき、乱数もシード値設定により再現可能である。

まず、その精度をサイコロの反復試行により確認した。サイコロは1から6までの目が出る確率が全て等しい一様分布とする。以下のコードで作成したサイコロのデータセットを用いて、サンプリングの観点から

本プロシジャにより、サイコロを投げたことを想定して、中心極限定理と大数の法則についてシミュレーションを行った。

```
/* サイコロの目 */
data DICE;
  do ROLL = 1 to 6;
    output;
  end;
run;
```

3.2 中心極限定理

中心極限定理(Central Limit Theorem, CLT)は、平均・分散が決まった母集団に対して、標本サイズが大きくなるほど、それに従った正規分布に近づくものである。例えば、サイコロ 1 回の目を対象とすると一様分布であるが、サイコロを 2 回投げて出た目の平均をとることで、平均・分散の特徴が生じる。

以下のコードの通り、SURVEYSELECT プロシジャでは、method=URS により復元抽出(Unrestricted Random Sampling)を実施し、outhits オプションにより重複して選択されるサンプル分もレコード生成しながら、reps=オプションにより反復施行した。

```
%macro M_CLT(REPS=);

/* サイコロ 2 回試行 (復元抽出) の繰り返し */
proc surveyselect data=DICE out=DICE2_REPS&REPS. seed=1234 method=URS sampsize=2 outhits reps=&REPS.;
run;

data DICE2_ME_REPS&REPS.;

  set DICE2_REPS&REPS.;
  by Replicate;

  retain ROLL_SUM 0;
  if first.Replicate then ROLL_SUM = 0;
  ROLL_SUM = ROLL_SUM + ROLL;

  if last.Replicate then ROLL_ME = ROLL_SUM / 2;
  else delete;

run;

/* 標本平均・中央値・分散を算出 */
proc summary data=DICE2_ME_REPS&REPS.;
  var ROLL_ME;
  output out=DICE2_STAT_REPS&REPS. n=N mean=ME median=MD var=VAR;
run;

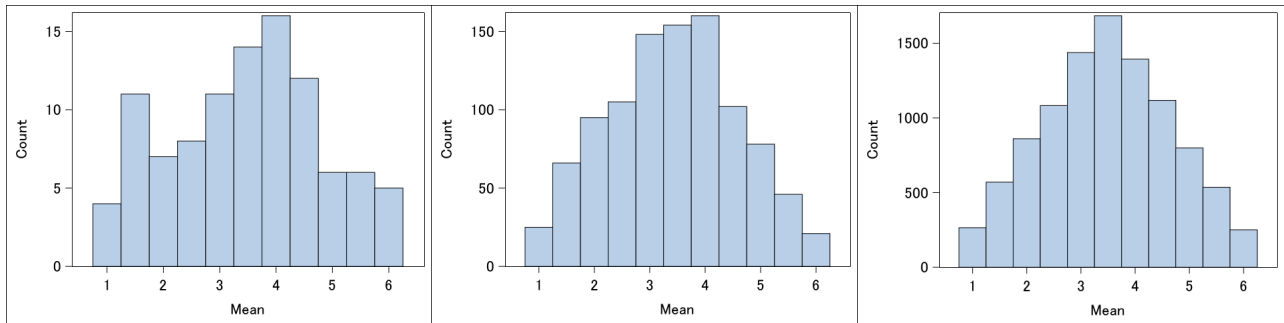
/* ヒストグラムを描画 */
proc sgplot data=DICE2_ME_REPS&REPS.;
  histogram ROLL_ME / scale=count binwidth=0.5;
  xaxis label='Mean' offsetmin=0.1 offsetmax=0.1 valueattrs=(size=16pt) labelattrs=(size=16pt) values=(1 to 6 by 1);
  yaxis label='Count' valueattrs=(size=16pt) labelattrs=(size=16pt);
run;

%mend M_CLT;

%M_CLT(REPS=100);
%M_CLT(REPS=1000);
```

```
%M_CLT(REPS=10000);
```

試行回数毎に、横軸を標本平均、縦軸を度数、ビン幅 0.5 としたヒストグラム(Fig 1)と標本の統計量のデータセット(Table 1)を以下に示す。ヒストグラムを比較すると、10,000 回試行で正規分布に近くなると言え、データセットにおいて、中央値は早い段階で 3.5 になり、平均は試行回数が増えるごとにそれに近づく。



(a)100 回試行

(b)1,000 回試行

(c)10,000 回試行

Fig 1 サイコロ 2 回分の平均算出を複数回試行した時の集計

Table 1 各標本の平均・中央値・分散

	N	ME	MD	VAR
1	100	3.475	3.5	1.8200757576
2	1000	3.4275	3.5	1.4048986486
3	10000	3.4817	3.5	1.4291080208

3.3 大数の法則

同一の試行を繰り返した場合、試行回数を増やすほど標本平均が 1 回の試行の期待値に近づく現象を大数の法則(Law of Large Numbers, LLN)と呼ぶ。サイコロを振って出た目で考え、以下のコードの通り、SURVEYSELECT プロシジャでは、標本サイズを 1 とし、reps=オプションにより試行回数をカウントしながらサンプリングを繰り返した。

```
%macro M_LLN(REPS=);

/* サイコロ 1 回試行の繰り返し */
proc surveyselect data=DICE out=DICE_REPS&REPS. seed=1234 sampsize=1 reps=&REPS.;
run;

/*現時点の平均を算出 */
data DICE_ME_REPS&REPS.;
set DICE_REPS&REPS.;
retain SUM 0;
SUM = SUM + ROLL;
ME = SUM / Replicate;
run;

/* 標本平均の変化を描画 */
ods graphics / antialiasmax=%eval(&REPS.);
proc sgplot data=DICE_ME_REPS&REPS.;
```

```

where &REPS. <= 10000 or mod(Replicate, 2) = 0 ;
step x=Replicate y=ME ;
xaxis label='Trial' valueattrs=(size=16pt) labelattrs=(size=16pt) ;
yaxis grid label='Mean' valueattrs=(size=16pt) labelattrs=(size=16pt) values=(3 to 4 by 0.5) ;
refline 3.5 / lineattrs=(color=black) ;
run ;

%mend M_LLN ;

%M_LLN(REPS=1000) ;
%M_LLN(REPS=10000) ;
%M_LLN(REPS=100000) ;

```

横軸を試行回数、縦軸を標本平均とした変化のグラフを以下に出力した(Fig 2)。100,000回で1~6の平均・中央値の3.5に収束していることが分かる。

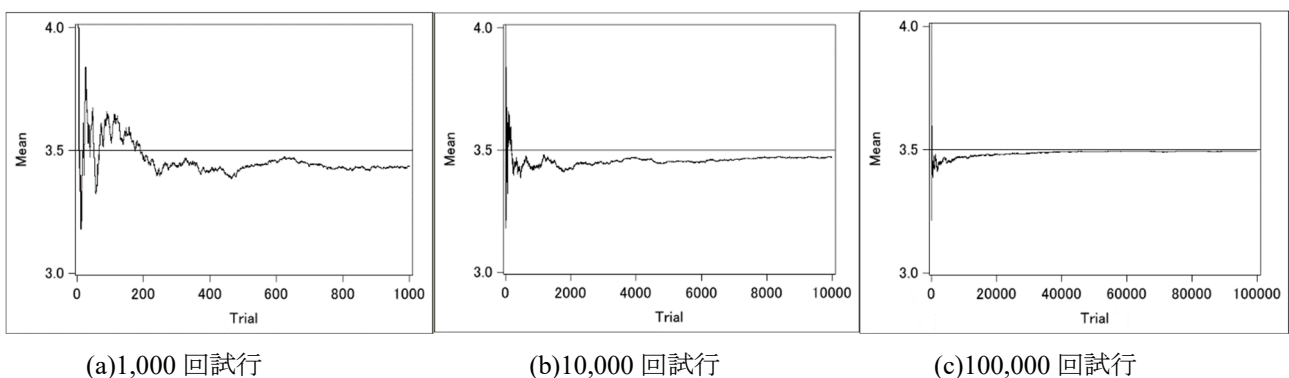


Fig 2 試行回数による標本平均の変化

4. 標本サイズと調整

4.1 標本設計

標本設計は母集団サイズから標本サイズを推定するが、確率的にサンプリングされたものには、標本誤差が生じることになる。N：母集団サイズ、n：標本サイズ、P：母集団比率、t：誤差係数、ε：標本誤差とすると、比率の標本誤差は次式で示される。

$$\varepsilon = t \sqrt{\frac{N-n}{N-1} \times \frac{P(1-P)}{n}}$$

t=1.96 は正規分布の95%信頼区間の幅を概ねの目安とする。Pは不明なことが多く、その場合は標本誤差が最大に見積もるようにP=0.5とする。以下のようにマクロでパラメーターを設定できるようにSASで実装すると、例えば、N=10000、n=1000の場合、ε=0.0294となり、±2.9%の標本誤差を見込まなければならない。

```

%macro M_SAMPLING_ERROR(T=1.96, N=, SN=, P=0.5) ;
data SAMPLING_ERROR ;
N = &N. ;

```

```

SN = &SN. ;
EPSILON = &T. * sqrt(((&N.-&SN.)/(&N.-1)) * (&P.*(1-&P.))/&SN.);
run ;
%mend M_SAMPLING_ERROR ;

%M_SAMPLING_ERROR(N=10000, SN=1000);

```

また、上記の標本誤差を考慮した、標本サイズは次式で示される。

$$n \geq \frac{N}{\left(\frac{\varepsilon}{t}\right)^2 \times \frac{N-1}{P(1-P)} + 1}$$

同様に以下の SAS マクロで実装すると、例えば、N=10000、ε=0.02 の場合、n=1936.3 となり、1937 以上の標本をサンプリングするのが望ましいと言える。なお、同じ ε=0.02 において N=100 であれば、n=96.04 となり、母集団サイズの大部分にあたる標本サイズを必要とすることになるため、母集団サイズが大きいほど効果を発揮するものである。また、多段抽出を行うと、概ね段数のルートに比例して標本誤差が大きくなり、ε を段数の平方根で除して調整する必要がある。その結果、より多くの標本が必要になる。

```

%macro M_SAMPLE_SIZE(N=, EPSILON=, T=1.96, P=0.5);
data SAMPLE_SIZE;
N = &N.;
EPSILON = &EPSILON.;
SN = &N. / ((&EPSILON./1.96)**2 * ((&N.-1)/(&P.*(1-&P.))) + 1);
run ;
%mend M_SAMPLE_SIZE ;

%M_SAMPLE_SIZE(N=10000, EPSILON=0.02);

```

4.2 層別抽出法と標本配分法

多くの場面でサンプリング対象の層が設定され、層別抽出が使用されることは多い。さらに、その機能として、Fig 3 のように、全体で決められた標本サイズを各層へ配分する標本配分法が存在する。標本配分法には種類がいくつかある。

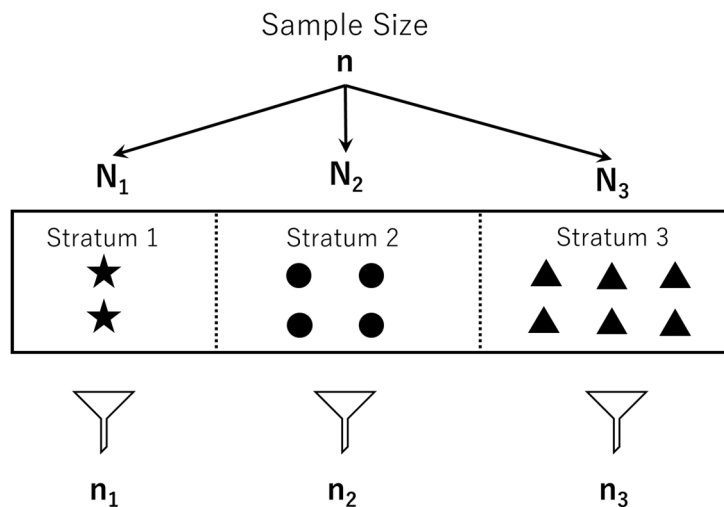


Fig 3 層別の標本配分のイメージ

比例配分法は、各層に、その母集団サイズに応じて平等に配分するものである。i を層、N を母集団サイズとすると、各層の標本サイズ n_i は次式で示される。

$$n_i = n \frac{N_i}{N}$$

ネイマン配分法は、予備調査等により各層の特定項目の値の分散が分かっている場合、それも考慮に入れて全体の分散を最小にするように標本配分を決定する。ただし、別の項目に関しては不適當な配分となり、比例配分法よりも分散が拡大する可能性もある。L を層の数、 S_i を各層の特定項目の標準偏差とすると、各層の標本サイズ n_i は次式で示される。 S_i が大きいほど多く配分されるようになる。

$$n_i = n \frac{N_i S_i}{\sum_{i=1}^L N_i S_i}$$

例として、3層で母集団サイズを全体 $N=12$ 、各層 $N_1=2, N_2=4, N_3=6$ 、標本サイズ $n=6$ の配分を想定する。比例配分法では、 $n_1=6*2/12=1, n_2=6*4/12=2, n_3=6*6/12=3$ となる。ネイマン配分法で、標準偏差を各層 $S_1=3, S_2=2, S_3=1$ と仮定した場合は、 $n_1=6*2*3/(2*3+4*2+6*1)=1.8, n_2=6*4*2/(2*3+4*2+6*1)=2.4, n_3=6*6*1/(2*3+4*2+6*1)=1.8$ により、 $n_1=2, n_2=2, n_3=2$ となり、各層の母集団サイズ以外にも層内のばらつきも考慮された結果になっている。

5. 標本設計と標本配分による多段サンプリング例

サンプリングの条件を複数組み合わせた例とその実装を示す。

5.1 調査方針

大企業と中小企業に層化された企業データを標本設計の上、一部の地域についてサンプリングする。業種によるが、日本において従業員数や資本金を基づいて定義された大企業・中小企業の割合はそれぞれ 0.3%、99.7%と言われる。各企業の製造業付加価値額（以下、付加価値額とする）が影響する調査を想定し、その額を考慮する。この場合、大企業と中小企業では重みが異なるが、効率良く企業標本を抽出したい。サンプリング手順としては、標本設計を行った上で、まずクラスター抽出で地域を決め、その地域全体を対象として、層を企業規模、特定項目を付加価値額とした層別抽出・標本配分を行った。全体像としては Fig 4 を想定した。

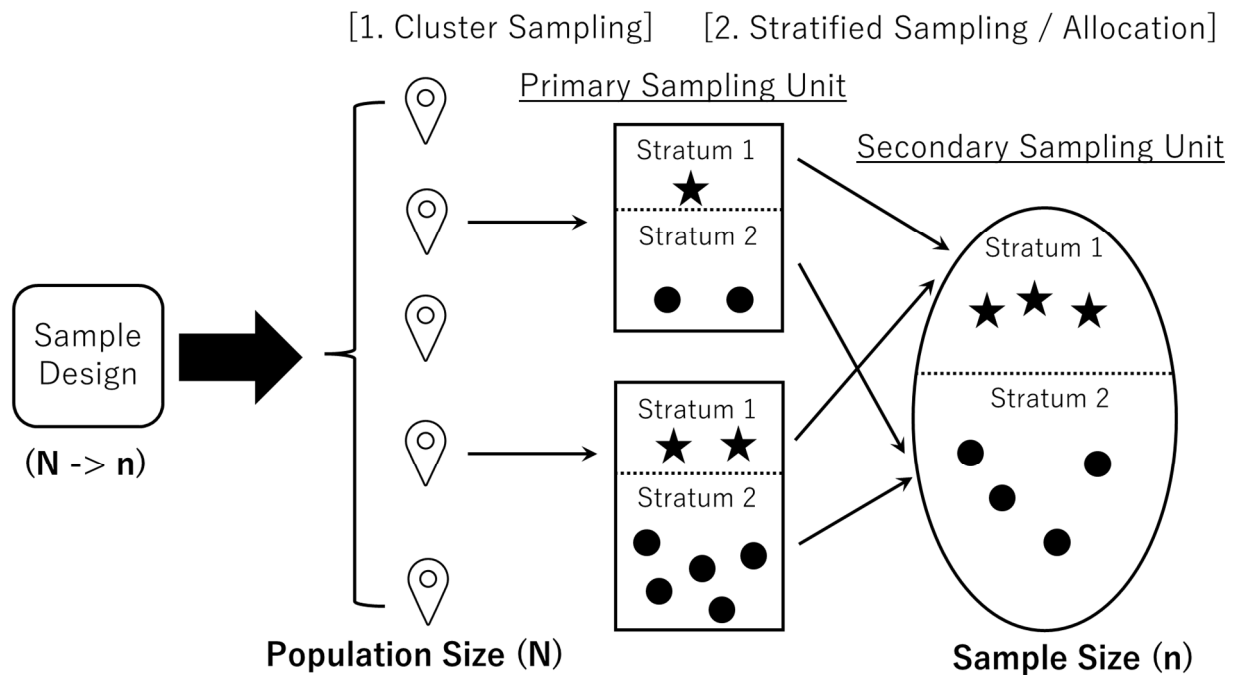


Fig 4 本実験における多段サンプリングの手順・構造

5.2 テストデータ

中小企業庁による 2006 年の情報を参考に、便宜上、日本国内の大企業を 1 万社、中小企業を 500 万社、付加価値額が大企業・中小企業の合計で各 50 兆円とした。また、母集団サイズを大きく考慮し、5 地域にそれぞれ大企業 1000 社、中小企業 10 万社を想定し、さらに、総額を企業数で割った平均が 50 億円、1000 万円で、これに基づき、ランダムに正規分布させたテストデータを作成した。付加価値額の単位は 100 万円とする。

```
data wk1 ;
/* 5 地域 */
do AREAN = 1 to 5 ;

  select(AREAN) ;
    when(1) AREA = "A" ;
    when(2) AREA = "B" ;
    when(3) AREA = "C" ;
    when(4) AREA = "D" ;
    when(5) AREA = "E" ;
  end ;

/* 大企業の平均値と母集団サイズ */
  TYPEN = 1 ;
  TYPE = 'Large' ;
  ME = 5000 ;
  COUNT = 1000 ;
  output ;
/* 中小企業の平均値と母集団サイズ */
  TYPEN = 2 ;
  TYPE = 'Small' ;
  ME = 10 ;
  COUNT = 100000 ;
```

```

output ;

end ;
run ;

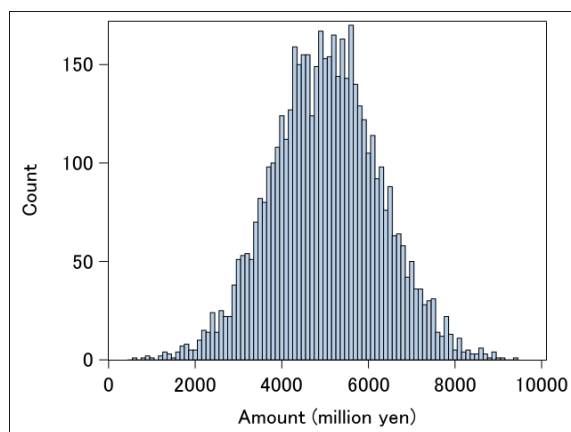
/* レコードを複製 */
data wk2 ;
set wk1 ;
do NO = 1 to COUNT ;
output ;
end ;
run ;

/* 正規分布データを作成 */
data wk3 ;
set wk2 ;
call streaminit(123) ;
/* 疑似乱数 (正規分布: 平均パラメーター・標準偏差パラメーター) を生成 */
if TYPEN = 1 then AMOUNT = ME * RAND('NORMAL', 1, 0.25) ;
else if TYPEN = 2 then AMOUNT = ME * RAND('NORMAL', 1, 0.2) ;
run ;

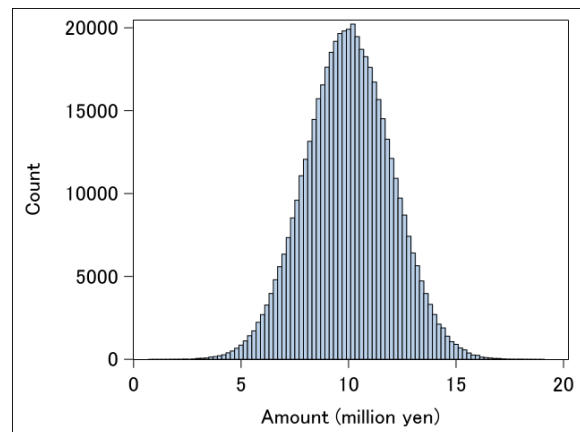
/* データ分布のヒストグラムを描画 */
/* 大企業 */
proc sgplot data=wk3 ;
where TYPEN = 1 ;
histogram AMOUNT / scale=count binwidth=100 ;
xaxis label='Amount (million yen)' valueattrs=(size=16pt) labelattrs=(size=16pt) ;
yaxis label='Count' valueattrs=(size=16pt) labelattrs=(size=16pt) ;
run ;
/* 中小企業 */
proc sgplot data=wk3 ;
where TYPEN = 2 ;
histogram AMOUNT / scale=count binwidth=0.2 ;
xaxis label='Amount (million yen)' valueattrs=(size=16pt) labelattrs=(size=16pt) ;
yaxis label='Count' valueattrs=(size=16pt) labelattrs=(size=16pt) ;
run ;

```

ビン幅を大企業 100、中小企業 0.2 として、データ分布を以下のようにヒストグラムを出力した。2つのヒストグラムの関係としては、母集団サイズの大きい Fig 4 (b)の中小企業の右側奥に、母集団サイズの小さい Fig 4 (a)の大企業が分布しているイメージである。



(a)大企業($N_1=5,000$)



(b)中小企業($N_2=500,000$)

Fig 4 付加価値額のテストデータの分布

5.3 各過程の実装

一式のコードは以下に示す。標本設計では、母集団サイズとなるデータセットのオブザベーション数から標本サイズを算出し、それをマクロ変数に格納し、後のサンプリングに引き渡すようにしている。標本誤差は $\varepsilon = 0.02$ (2%) として、二段抽出のため、 ε を $\sqrt{2}$ で除することで、厳格に見積もる。

第一次抽出単位としてのクラスター抽出では、SERVEYSELECT プロシジャの `samplingunit` ステートメントにより 2 地域の全体を抽出した。第二次抽出単位としての層別抽出では、`strata` ステートメントにより層を指定し、それに付随する `alloc`=オプションにより標本配分法、具体的に `alloc`=PROPORTIONAL により比例配分法、`alloc`=NEYMAN によりネイマン配分法を実装する。なお、ネイマン配分法では別途層別に特定項目の分散 (`_VAR_` 変数) を計算したデータセットを用意して指定する必要がある。

```
/*----- 標本設計 -----*/

%macro M_SAMPLE_SIZE_SERVEY(DS=, EPSILON=, P=0.5);

  data _NULL_ ;
    set &DS. nobs=N ;
    call symput('N', left(put(N,BEST.))) ;
  run ;

  %global SN ;
  data _NULL_ ;
    SN = &N. / ((&EPSILON./1.96)**2 * ((&N.-1)/(&P.*(1-&P.))) + 1) ; /* 標本サイズを算出 */
    call symput('SN', left(put(round(SN, 1), BEST.))) ;
  run ;

  %put population size = &N. ;
  %put sample size = &SN. ;

%mend M_SAMPLE_SIZE_SERVEY ;

/* パラメーター設定 */
%M_SAMPLE_SIZE_SERVEY(DS=wk3, EPSILON=0.02/sqrt(2)) ;

/*----- 第一次抽出単位：クラスター抽出 -----*/

proc surveyselect data=wk3 out=wk4 seed=1234 sampsize=2; /* 2 地域 */
  samplingunit AREAN ;
run ;

/*----- 第二次抽出単位：層別抽出・標本配分法 -----*/

/* 層別抽出前ソート */
proc sort data=wk4 ;
  by TYPEN ;
run ;

/** 比例配分法 **/
proc surveyselect data=wk4 out=wk5 seed=1234 sampsize=&SN. ; /* 標本設計による標本サイズ */
  strata TYPEN / alloc=PROPORTIONAL ;
run ;

/** ネイマン配分法 **/
```

```

/* 層別の分散を算出 */
proc summary data=wk4 ;
  by TYPEN TYPE ;
  var AMOUNT ;
  output out=wk6(drop=_TYPE_) var=_VAR_ ;
run ;

proc surveyselect data=wk4 out=wk7 seed=1234 sampsiz=&SN. ;
  strata TYPEN / alloc=NEYMAN var=wk6 ;
run ;

```

過程を含め、サンプリング結果を以下に示す。標本設計により、母集団サイズ 505,000 から標本サイズ 4,757 を見積もり、クラスター抽出により 2 地域に絞り (Table 2)、後の層別抽出に使用した。比例配分法では、各層の母集団サイズから比率 (SelectionProb, SamplingWeight 変数) を平等に、標本サイズを配分している。ネイマン配分法では、層別の分散を用いることで (Table 3)、大企業は母集団サイズが極端に小さいため、その効果が大きく、標本サイズ 2000 として全数抽出された。標本配分法別・層別にデータ分布のヒストグラムも出力した。標本データ (Fig 5) においても母集団データ (Fig 4) のように正規分布が保たれている。

抽出結果の出力データセットとして、実際は標本サイズ分 (今回の場合は 4,757 オブザベーション) の各値を持つレコードがサンプリングされるが、ここでは Table 5 のように省略した (サンプリング結果に付随する Total 変数から SamplingWeight 変数の値は同層の TYPEN では同じである)。

Table 2 クラスター抽出によるサンプリングのアウトプット出力と内訳データセット

SAS システム
SURVEYSELECT プロシジャ

選択の方法	Simple Random Sampling
抽出単位変数	AREAN

入力データセット	WK3
乱数シード	1234
標本サイズ	2
選択確率	0.4
サンプリングの重み	2.5
出力データセット	WK4

	AREAN	AREA	TYPEN	TYPE	COUNT
1	4	D	1	Large	1000
2	4	D	2	Small	100000
3	5	E	1	Large	1000
4	5	E	2	Small	100000

Table 3 ネイマン配分法で使用する層別の分散の算出結果

	TYPEN	TYPE	_FREQ_	_VAR_
1	1	Large	2000	1511153.7104
2	2	Small	200000	3.9854482381

Table 4 標本配分法によるサンプリングのアウトプット出力

(a)比例配分法

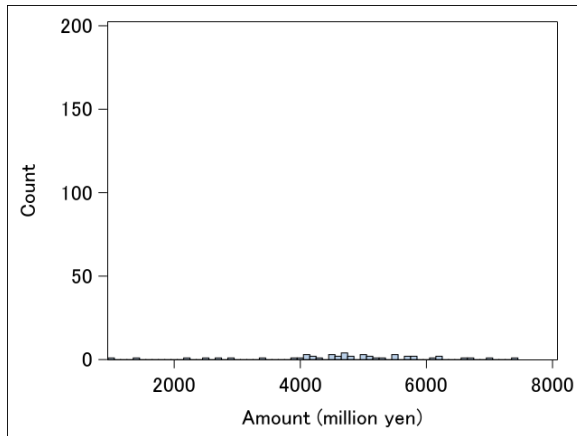
(b)ネイマン配分法

SAS システム	
SURVEYSELECT プロシジャ	
選択の方法	Simple Random Sampling
層変数	TYPEN
割付	Proportional

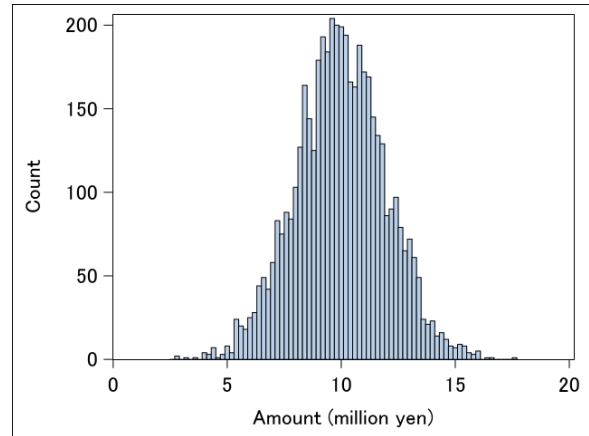
入力データセット	WK4
乱数シード	1234
層の数	2
総標本サイズ	4757
出力データセット	WK5

SAS システム	
SURVEYSELECT プロシジャ	
選択の方法	Simple Random Sampling
層変数	TYPEN
割付	Neyman

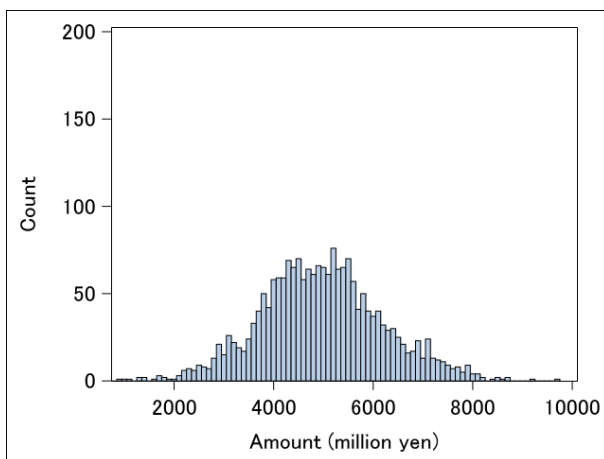
入力データセット	WK4
乱数シード	1234
Variance Input Data Set	WK6
層の数	2
総標本サイズ	4757
出力データセット	WK7



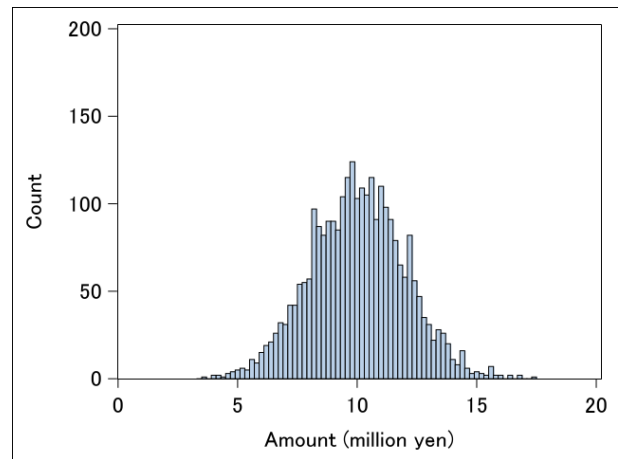
(a)比例配分法：大企業($n_1=47$)



(b)比例配分法：中小企業($n_2=4,710$)



(c)ネイマン配分法：大企業($n_1=2,000$)



(d)ネイマン配分法：中小企業($n_2=2,757$)

Fig 5 付加価値額の抽出データの分布の比較($n=4,757$)

Table 5 標本配分によるサンプリング結果のデータセット一部

(a)比例配分法

	TYPEN	TYPE	Total	AllocProportion	SampleSize	ActualProportion	SelectionProb	SamplingWeight
1	1	Large	2000	0.009901	47	0.00988	0.0235	42.55319
2	2	Small	200000	0.990099	4710	0.99012	0.02355	42.46285

(b)ネイマン配分法

	TYPEN	TYPE	Total	Variance	AllocProportion	SampleSize	ActualProportion	SelectionProb	SamplingWeight
1	1	Large	2000	1511153.71	0.860289544	2000	0.420433046	1	1
2	2	Small	200000	3.985448238	0.139710456	2757	0.579566954	0.013785	72.542618789

このように、指定の条件下で、標本設計から標本配分までの多段サンプリングが行えていることを確認できる。生産高等を推定するサンプリングを行う場合、比例配分法の場合は、層の規模の違いから、大企業を過小評価してしまう可能性があるが(Fig 5 a,b)、ネイマン配分法の場合は、全体からの視点では、バランスよく配分され、合わせると二峰性の分布と言える(Fig 5 c,d)。なお、一企業を等価として、特定項目に依存しない層別サンプリング、企業規模別に独立した調査がしたい場合、各層の母集団に対する割合が一定の比例配分法が適している。

6. まとめ

現実の統計調査において、調査対象数の確保の問題をはじめとして、様々な制約が掛かる場合が多い。限定された状況の中で、興味のある事柄について、できるだけ正確に知るためには、標本設計や標本配分への理解が重要になる。サンプルデータを用い、様々な標本抽出法を実装し、抽出された標本の性質を確かめておくことも、実際に標本抽出を計画する際に役立つと考える。

本稿では、標本誤差を考慮した層別多段サンプリングを SAS で実装し、その挙動を確認した。ビッグデータや AI を扱う時代においても、このような過程の必要性は残ると思われるため、今後も本領域に関心を持ちたい。SAS ユーザーが標本抽出について理解したり、実装したりする際の一助になれば幸いである。

参考文献

- [1] 廣瀬雅代・稲垣佑典・深谷肇一、サンプリングって何だろう：統計を使って全体を知る方法、岩波書店、2018 年
- [2] Bryan F. J. Manly (編)・Jorge A. Navarro Alberto (編)・深谷肇一 (訳)、生態学のための標本抽出法、共立出版、2023 年
- [3] 日本統計学会、調査の実施とデータの分析、東京図書、2023 年
- [4] SAS Institute Inc., SAS/STAT 15.2 User's Guide - The SURVEYSELECT Procedure, 2020
<https://documentation.sas.com/api/docsets/statug/15.2/content/surveyselect.pdf>
- [5] 総務省統計局、データ・スタート：EBPM 活用塾、2023 年
<https://www.stat.go.jp/dstart/point/>

- [6] 金森匡彦、地上調査の設計、日本森林技術協会、2014 年
https://redd.ffpri.affrc.go.jp/pub_db/course_materials/_img/2014_applieda/2014_applieda_chap01.pdf
- [7] 埼玉県、統計に関する Q&A：標本の抽出はどのように行えばよいのですか、2021 年
<https://www.pref.saitama.lg.jp/a0206/toukeifaq/q1-7.html>
- [8] 中小企業庁、日本の中小企業
<https://www.chusho.meti.go.jp/koukai/chousa/chushoKigyousentai9wari.pdf>
- [9] 株式会社社会情報サービス、BellCurve：統計 WEB：統計学の時間
<https://bellcurve.jp/statistics/course/>
- [10] 高田浩成、統計調査と SAS によるサンプリング方法、SAS ユーザー総会、2023 年
(各 Web サイトへの最終アクセスは 2024 年 8 月)

ODS RTFでの出力環境の整備

～RTFスタイルとSASマクロの運用・管理手順～

○宮辻 淳, 麻生 美由紀, 平野 勝也

(株式会社インテリム)

Preparation for ODS RTF programming

-Operation and management procedures for Style template and SAS macros-

Atsushi Miyatsuji, Miyuki Aso, Katsunari Hirano

intellim Corporation

要旨

Dynamic Data Exchange を用いて Excel 形式の解析図表を作成する場合、しばしば SAS システムと Excel との間でデータ交換する際に原因不明の I/O エラーが発生する問題がある。Excel を利用せずに SAS システムだけで解析図表を作成する方法として、ODS RTF を利用して Rich Text Format (RTF) 形式の解析図表を作成する方法が、日本では SAS Version 8.2 から利用できるようになった。この方法は、プログラム実行時の安定性が高く、何より SAS 社によりサポートされている機能であり、分からないことをいつでも質問できる体制がある、という点で安心して利用できる。本稿では、RTF 形式の解析図表を作成するために、検討・準備したこととして、RTF 出力時の体裁（フォントの種類と大きさ、四隅の余白、表の位置と罫線と塗りつぶし、等）を制御している Style テンプレートの準備と、効率的に業務を進めるために必要と考えた SAS マクロとその管理方針を紹介する。

キーワード：ODS RTF, RTF 出力, Style テンプレート, TEMPLATE プロシジャ, SAS マクロ

1 はじめに

近年、Rich Text Format (RTF) 形式の解析図表の納品をクライアントから要望されることが増えてきている。また、Dynamic Data Exchange (DDE) を用いて Excel 形式の解析図表を作成する際に問題（原因不明の I/O エラー、PDF 化時の文字切れ）が発生することがあるため、RTF 形式の解析図表を作成するための環境を整備しようと考えた。

RTF 形式で解析図表を作成する場合、SAS プログラムだけで完結する簡潔さと、プログラム実行時の安定性、SAS 社による Q&A 体制の充実が長所である。RTF 形式の解析図表の作成法に関しては、これまでに沢山のテクニカルな事例が報告されているため、本稿では、RTF 出力時の体裁（フォントの種類と大きさ、四隅の余白、表の位置と罫線と塗りつぶし等）を制御している Style テンプレートの作成・利用法を説明する。

そして、Excel DDE プログラミングと ODS RTF プログラミングの違いを振りかえり、効率的に ODS RTF プログラミングを進める上で必要と考えた SAS マクロと、その管理方針を紹介する。

2 Style テンプレート

本章では Style テンプレートの使用経験がない人に向けて、Style テンプレートの役割と Style テンプレートの参照・作成方法を説明して、作成した Style テンプレートの利用法を説明する。

2.1 Style テンプレートの役割

ODS RTF ステートメントを使用して RTF ファイルを作成する際に、Style=オプションを指定する場合と指定しない場合のプログラム (Program code 1) と出力結果 (Figure 1) を示す。

Program code 1

<pre>/*Style テンプレートの指定なし*/ ods rtf file="出力先の Path%OUT1.rtf" ; proc print data = sashelp.class noobs label ; run ; ods rtf close ;</pre>	<pre>/*Style テンプレートの指定あり*/ ods rtf file="出力先の Path%OUT1.rtf" style = Styles.roman9 ; proc print data = sashelp.class noobs label ; run ; ods rtf close ;</pre>
---	--

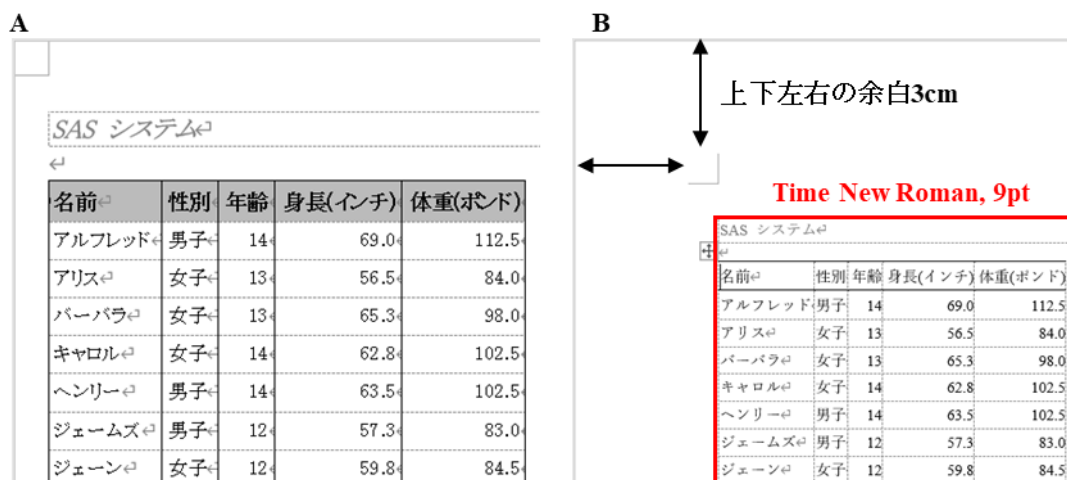


Figure 1. Style=オプションを指定する場合と指定しない場合の出力結果

Style=オプションで Style テンプレートを指定することで、出力の全体的な外観 (セルの塗りつぶし、フォント種とサイズ、余白、罫線など) が変わっていることに気付くと思う。実際、Style テンプレート (Styles.Roman9) では、セルの塗りつぶしをなくし、フォント種とサイズを Times New Roman と 9pt に、上下左右の余白を 3cm に、縦罫線をなくし横罫線を設定している。単純な設定であるが、Style テンプレートを利用することで、より解析図表らしい見た目の RTF ファイルを作成できることが理解できると思う。

Style テンプレートでは、半角/全角文字用のフォントを別々に定義できないため、「MS 明朝」に設定した Style テンプレート (Mincho9) と「Times New Roman」に設定した Style テンプレート (Roman9) を作成して、使い勝手を確認した。当たり前であるが、Style テンプレート (Mincho9) では数値を含むすべての出力が「MS

明朝」になってしまう。半角文字を「Times New Roman」で、全角文字を「MS 明朝」で表示したい場合には、Style テンプレート (Roman9) を利用して、全角文字を含む部分だけを、インラインフォーマットを利用して半角/全角フォントを指定するほうが効率的であったので報告する[1, 2]。

2.2 SAS 社が提供する Style テンプレートの参照方法

SAS 社が提供する Style テンプレートは SASHELP ライブラリに保存されているが、SAS データセットやカタログのようにエクスプローラでは参照できない。Style テンプレートを確認するための画面の遷移を Figure 2 に示す。まず、結果ウィンドウをアクティブ (A) な状態にして、「表示」メニューから「テンプレート」を指定 (B) することでテンプレートウィンドウ (C) を表示できる。Style テンプレートは、Sashelp.Tmplmst に保存されているため、それを選択することで主要な Style テンプレートを確認できる。テンプレートウィンドウで Styles 内の Rtf (RTF ファイルを作成する際のデフォルトの Style テンプレート) をダブルクリックすると、TEMPLATE プロシジャで記述された Style テンプレートのソースコードがテンプレートブラウザ (D) に表示される。

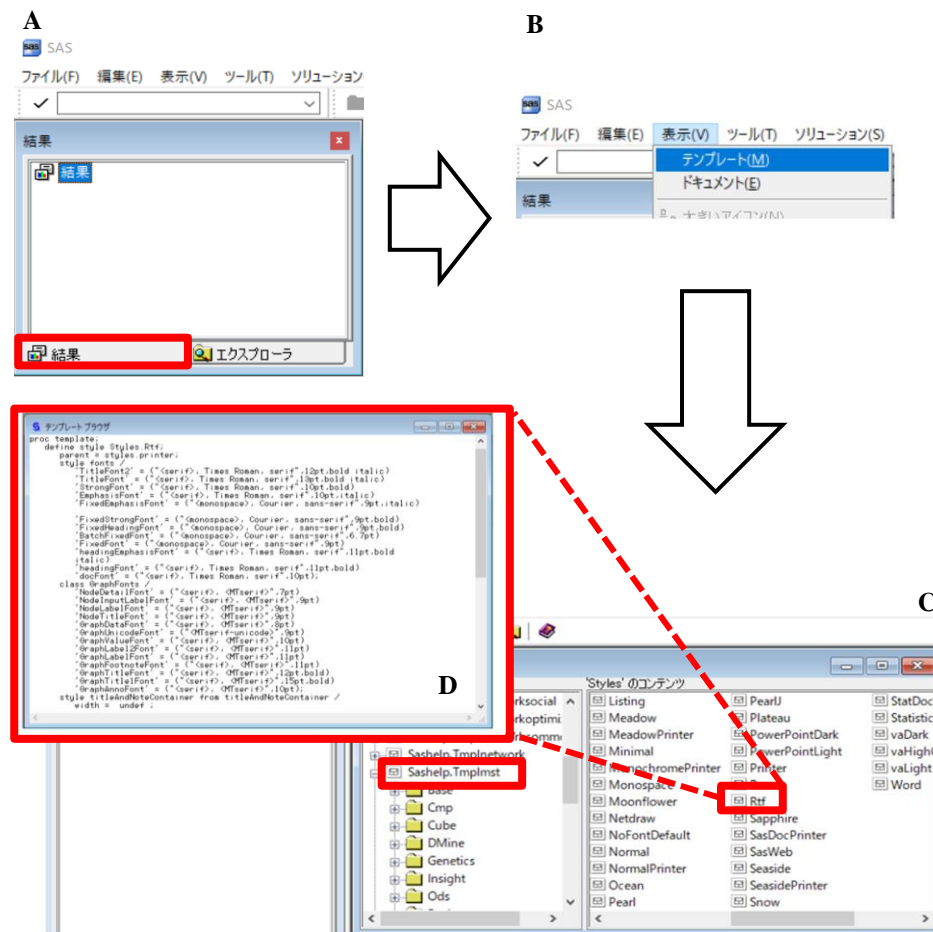


Figure 2. SAS 社が提供する Style テンプレートの参照方法

参照したい Style テンプレート (以下では、Styles.rtf) を特定できている場合には、Program code 2 を実行することで、ログ画面 (Figure 3) にソースコードを表示できる。

Program code 2

```
PROC TEMPLATE ;
    source styles.rtf ;
run ;
```



```
1 proc template ;
NOTE: HTML Bodyファイルの書き込み先: sashtml.htm
2 source styles.rtf ;
define style Styles.Rtf;
parent = styles.printer;
style fonts /
'TitleFont2' = ("<serif>, Times Roman, serif",12pt,bold italic)
'TitleFont' = ("<serif>, Times Roman, serif",13pt,bold italic)
'StrongFont' = ("<serif>, Times Roman, serif",10pt,bold)
'EmphasisFont' = ("<serif>, Times Roman, serif",10pt,italic)
'FixedEmphasisFont' = ("<monospace>, Courier, sans-serif",9pt,italic)
'FixedStrongFont' = ("<monospace>, Courier, sans-serif",9pt,bold)
'FixedHeadingFont' = ("<monospace>, Courier, sans-serif",9pt,bold)
'BatchFixedFont' = ("<monospace>, Courier, sans-serif",6.7pt)
'FixedFont' = ("<monospace>, Courier, sans-serif",9pt)
'headingEmphasisFont' = ("<serif>, Times Roman, serif",11pt,bold italic)
'headingFont' = ("<serif>, Times Roman, serif",11pt,bold)
'docFont' = ("<serif>, Times Roman, serif",10pt);
class GraphFonts /
NodeDetailFont' = ("<serif>, <MTserif>",7pt)
```

Figure 3. ログ画面

2.3 Style テンプレートの作成方法

クライアントの文書規定は様々であり、これらの文書規定に柔軟に対応するためには、最低限の設定を盛り込んだ Style テンプレートを準備しておくことが得策かもしれない。Program code 3 では、Style テンプレートを保存するフォルダを作成した後に、そのフォルダに対して SAS ライブラリ「mystyle」を割り当てている。次に ODS PATH ステートメントで書き込み権を設定したうえで、PROC TEMPLATE プロシジャで Style テンプレート (NewStyle) を作成している。

Program code 3

```
libname mystyle "Styleテンプレートを保存するフォルダ" ;
ODS PATH mystyle.template(WRITE) sashelp.tmplmst(read) ;
PROC TEMPLATE
define style NewStyle ;
parent = Styles.Rtf ;
class fonts /
'TitleFont' = ("Times New Roman",9pt)
'TitleFont2' = ("Times New Roman",9pt)
'StrongFont' = ("Times New Roman",9pt)
'EmphasisFont' = ("Times New Roman",9pt)
'headingEmphasisFont' = ("Times New Roman",9pt)
'headingFont' = ("Times New Roman",9pt)
'footFont' = ("Times New Roman",9pt)
'FixedEmphasisFont' = ("Times New Roman",9pt)
'FixedStrongFont' = ("Times New Roman",9pt)
'FixedHeadingFont' = ("Times New Roman",9pt)
```

```
'BatchFixedFont' = ("Times New Roman",9pt)
'FixedFont' = ("Times New Roman",9pt)
'docFont' = ("Times New Roman",9pt);

end;

run;
```

Style テンプレート (NewStyle) は、parent ステートメントを用いて、SAS 社が提供する Styles.Rtf をベースにして、フォントを Times New Roman、サイズを 9pt で設定している。作成される Style テンプレートは、事前に作成した Style テンプレートを保存するフォルダ内の「template.sas7bitm」に保存されている。

2.4 Style テンプレートの利用方法

2.3 節で作成した Style テンプレート「NewStyle」の利用例を Program code 4 に示す。利用例では、SAS ライブラリ「mystyle」を割り当て、次に ODS PATH ステートメントで Style テンプレートの参照順を設定している。そして、ODS RTF ステートメントの style=オプションで「NewStyle」を指定することで、作成する RTF ファイルの出力形式 (この事例では、すべてのフォントを Times New Roman、サイズを 9pt) を制御している。

Program code 4

```
libname mystyle "Styleテンプレートを保存したフォルダ";
ODS PATH mystyle.template(read) sashelp.tmplmst(read);

ODS RTF file ='xxx.rtf' style=NewStyle ;
proc print data= sashelp.class ;
run ;
ODS RTF close ;
```

3 Excel DDE プログラミングと ODS RTF プログラミングの違い

Excel DDE プログラミングで解析図表を作表する場合、図表計画書として作成した出力見本を出力先のテンプレートとして利用し、Figure 4 の赤枠の部分に集計結果を出力しているケースが多いと思う。つまり、テンプレートの一部、例えば、図表タイトルや項目名やカテゴリ、等を利用して解析図表を作成している。

①の検討時に、解析図表のタイトルと脚注を RTF ファイルのどの領域に出力するか、という問題に遭遇した。RTF ファイルを Word で開くと、Figure 5 のようにヘッダー部、本文、フッター部という 3 つの領域が存在する。ODS RTF プログラミングで、ODS RTF ステートメントに `bodytitle` オプションを指定しなければ、3 つの領域を利用できることが分かった。Title/Footnote ステートメントにテキスト情報を記述することでヘッダー部/フッター部に、ODS RTF `prepage=/text=` にテキスト情報を記述することで本文に、テキスト情報を出力できることを確認した。クライアントの要望が第一優先であるが、ODS RTF プログラミングの表現力を生かすためには、`bodytitle` オプションを利用しないほうがよいと我々は結論づけた。

4 SAS マクロを作成し運用するにあたって

4.1 SAS マクロの管理方針

業務で利用する SAS マクロは一元管理し、バージョンも管理したいと考えている。各プロジェクトの業務開始時に、最新版の SAS マクロをプロジェクトフォルダに配布する。プロジェクト進行中に新バージョンの SAS マクロがリリースされた場合には、各プロジェクトで導入の可否を判断いただくことを考えている。なお、一元管理を開始する前に、SAS マクロのベリフィケーションを実施し、その記録を残す予定である。

4.2 ベリフィケーション記録

SAS マクロごとに、Table1 に示したベリフィケーション記録を残す予定である。

Table 1. ベリフィケーション文書一覧

文書	作成者	承認の要否
ユーザ要求仕様書	利用者	要
指名書（プログラマー、ベリファイヤー）	承認者	—
ベリフィケーション計画書	プログラマー	要
マクロ仕様書	プログラマー	要
SAS マクロ	プログラマー	—
テスト計画書	ベリファイヤー	要
テスト報告書	ベリファイヤー	要
ベリフィケーション報告書	プログラマー	要

利用者からの新規の開発要求を受け付けて、SAS マクロを開発する否かを判断し、開発することを決定した後に、担当者をアサインする。そして、ベリフィケーション計画書とマクロ仕様書を作成してから、SAS マクロを作成し、テスト計画書に従いテストして、テスト結果をテスト報告書にまとめて、最後にベリフィケーション報告書を作成する、という一連の流れで記録を残す予定である。

5 最後に

Excel DDE プログラミングに慣れている人が、ODS RTF プログラミングに移行するのは簡単なことではない。ODS RTF での出力環境を整備することで、ODS RTF プログラミング未経験者に対して手助けとなることを期待している。ODS RTF プログラミングに移行することで、冒頭で述べたシステム連携時の不安定さや

DDE 機能のサポート期限の不透明さから解放されて、ODS 機能を活用する機会になればと思う。ODS 機能には RTF だけでなく EXCEL、WORD、PDF、HTML 等が存在し、これらを活用する契機となることも期待している。

SAS マクロの管理に関して少し記載したが、まだまだ手探り状態であるため、いろんな角度からアドバイスや情報交換いただけると有難い。

6 引用文献

- [1] 麻生美由紀. 2018. ods rtf+proc report における解析帳票の日本語/英数字フォントの明示的制御. SAS ユーザー総会 2018.
https://www.sas.com/ja_jp/events/18/users-group-2018/sessions.html#m=4-2
- [2] 太田裕二. 2017. ODS 機能と REPORT プロシジャを用いた 解析帳票の作成. SAS ユーザー総会 2017.
https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups2017-b-11-02.pdf

3つ以上の治療法を比較する 場合の傾向スコア逆確立重み付け解析

○中倉 章祥、岸本 淳司

(九州大学病院ARO次世代医療センター)

Propensity score method using inverse probability treatment weighting for comparing three or more treatments

Akiyoshi Nakakura, Junji Kishimoto

Center for Clinical and Translational Research, Kyushu University Hospital

要旨

治療による効果を比較する場合、治療法以外の患者背景は群間で等しくなっていることが望ましい。群間の共変量バランスを調整する方法として、傾向スコアを使用する方法がある。傾向スコアは共変量を持つ患者が治療群に割り当てられる確率と解釈することができ、ロジスティック回帰を使用することでも簡単に求めることができる。また、群間で共変量のバランスがとれているかを評価する方法として、共変量ごとに標準化差を計算することや傾向スコアの分布を確認する方法が一般的に使用される。SAS では傾向スコア解析のために PSMATCH プロシジャが準備されており、2 群間の比較においては傾向スコアの推定、傾向スコアを使った共変量の調整、共変量バランスの評価を簡単に実行することができる。しかし、3 群以上の傾向スコア解析に対応したプロシジャは今のところ存在しないため、解析のためには自身でプログラムを作成しこれらの手順を実施する必要がある。3 群以上で比較を行いたい場合に、一般化傾向スコアによる逆確率重み付けを使用して共変量を調整し、バランス評価を行う方法と SAS のプログラムを紹介する。

キーワード：傾向スコア 逆確率重み付け ATE ATT 標準化差 共変量調整 観察研究 バイオリプロット

1. はじめに

治療による効果を調べる臨床研究において、治療法の効果を正しく推定するためには群間で患者背景が揃っていることが望ましい。観察研究では特に治療は意図して決定されるため、治療法によって異なる患者背景を持つデータから治療法の比較を行う必要がある。解析時にベースライン共変量を調整することで、仮に患者背景が同等であった場合の妥当な比較を行う方法として、傾向スコア解析の需要はますます高まっている。傾向スコアは、観察されたベースライン共変量から計算された、治療法に割り当てられる確率と解釈できる。傾向スコアを正しく推定できた場合、傾向スコアで条件付けた共変量は群間でバランスが取れること

が知られている。傾向スコア解析は多くの場合 2 群間の妥当な比較方法として使用されるが、一般化傾向スコアとして拡張することで 3 群以上の治療法の比較においても提供可能である。

傾向スコアによる共変量調整はその簡便さから、広く使用されている。傾向スコア解析の手順は大きく①傾向スコアを推定する完全データを準備、②傾向スコアの推定、③傾向スコアによる共変量調整の実施、④共変量のバランス評価、⑤治療効果の推定、⑥感度分析、を順に実施することになる。[6] 傾向スコア解析の実施者は特に②③④の各段階で最適な方法を選択していく必要がある。SAS においても傾向スコア解析のために PSMATCH プロシジャが準備されており、傾向スコアの推定とマッチング及び重み付けによる共変量調整、共変量バランスの評価が簡単に実施可能である。しかし PSMATCH プロシジャは 2 群の傾向スコア解析にのみ対応しており、3 群以上の場合にも対応可能なプロシジャは現在のところ存在しない。一部の解析ソフトでは対応したプログラムも存在するが、SAS のみで解析が可能であればより望ましい。また、2 群の傾向スコアに比べて一般化傾向スコア実施のガイダンスは少ないことも使用を難しくしている。

本稿では、SAS で一般化傾向スコアを使用して共変量調整を行う方法を紹介する。2 章では 2 群の場合に使用される傾向スコアについて概説し、3 章では 3 群以上の場合に拡張した一般化傾向スコアを紹介する。4 章では作成したマクロを踏まえての SAS での実装方法を紹介する。

2. 2 群の傾向スコア

2 群の共変量を調整する場合、傾向スコアはロジスティック回帰モデルを使用して下記のように計算できる。

$$\hat{e}_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})}$$

傾向スコアはロジスティック回帰や Random forest、一般化ブースティング法などで推定可能である。SAS では PSMATCH プロシジャでの推定の他、LOGISTIC プロシジャを使っても簡単に計算できる。[1]

傾向スコアによる共変量調整は、マッチングや層別解析、逆確率重み付け解析を使って実施可能である。マッチングは傾向スコアが近いデータ同士をマッチングしていく方法である。層別解析は傾向スコアの値でいくつかの層を作成して解析を実施する。これらの方法についての詳細は多くのわかりやすい資料が存在するため、本稿では割愛する。[3] 逆確率重み付け解析 (IPW) は傾向スコアの逆数でデータを重み付けすることで共変量バランスをとる方法である。集団全体における治療の効果 (ATE) が知りたい場合と、治療群における治療の効果 (ATT) が知りたい場合で重みの付け方が異なる。

$$\text{(ATE の場合)} \quad w_{ATE_i} = \frac{z_i}{e_i} + \frac{1-z_i}{1-e_i}$$

$$\text{(ATT の場合)} \quad w_{ATT_i} = z_i + (1 - z_i) \frac{e_i}{1-e_i}$$

ただし、 z_i は i 番目の患者が治療群の場合に 1、対照群の場合に 0 となる指示変数である。

共変量バランスの評価では標準化平均差 (SMD) を計算する方法と、傾向スコアの分布を描画する方法がある。[2, 5, 9] 標準化平均差は各群の平均値と標準偏差を使って $SMD = (\bar{X}_1 - \bar{X}_0) / \sqrt{\hat{\sigma}_1 + \hat{\sigma}_0 / 2}$ で計算できる。[4, 8]

SAS では PSMATCH プロシジャを使用することで、ここまでの一連の流れを実施可能である。多くのプロシジャで実装されている WEIGHT ステートメントに計算した重みを使用することで、治療効果の IPTW 推定量を計算することができる。

3. 一般化傾向スコア

3.1. 一般化傾向スコアの推定

3群以上の共変量を調整する場合には、一般化傾向スコアを使用する。一般化ロジスティック（名義ロジスティック）回帰モデルを使用して、一般化傾向スコアを推定することができる。一般化傾向スコアでは、患者*i*が治療法*x*に割り付けられる確率が計算される。そのため、治療法*t* = (A, B, C)の3つの群がある場合、3つの傾向スコア*P_A*、*P_B*、*P_C*が計算される。ただし、*P_A* + *P_B* + *P_C* = 1である。

3.2. 逆確率重み付け

推定した傾向スコアをもとに、データに重みをつけることで群間の共変量を調整する。推定したい母集団によってATEの重みは下記の通り計算できる。[3]

$$w_{ATE_i} = \sum_{t=1}^T \frac{z_i[t]}{P_t}$$

ここで、*z_i[t]*は患者*i*が群*t*のとき1、そうでないとき0を示す指示変数である。また、群*t'*を参照群としたATTの重みは下記のとおりである。

$$w_{ATT_i}[t'] = \sum_{t=1}^T z_i[t] \frac{P_t}{P_{t'}}$$

3.3. 共変量バランスの確認

群間で共変量のバランスがとれているかを確認するためには、①共変量の標準化差を計算する、②傾向スコアの分布を描く、の2つの方法で確認することができる。

3.3.1. 標準化バイアス

SMDは2つの分布の距離の指標である。3群の場合にも全ての群間のSMDを計算することはできるが、群が多くなるほど煩雑なものになってしまう。McCaffreyらは標準化バイアス(SB)、または母集団からの標準化バイアス(PSB)を共変量のバランスの指標として使用することを提案した。[7] SMDからの拡張を考え、SB、PSBを再定義する。

SBは群1と群2の距離を標準化したものである。

$$SB = |\bar{X}_1 - \bar{X}_2| / \hat{\sigma}$$

ここで、 $\hat{\sigma}$ は対象とする集団の標準偏差である。ATE集団の場合、2つの群を併合した標準偏差は $\hat{\sigma} = \sqrt{\{\hat{\sigma}_1 * (n_1 - 1) + \hat{\sigma}_2 * (n_2 - 1)\} / (n_1 + n_2 - 2)}$ により計算できる。

特に、 $n_1 = n_2$ のとき、SBはSMDの絶対値に一致する。ATT集団の場合、 $\hat{\sigma}$ は基準とする群の標準偏差を使用する。

PSBは群*t*が想定される母集団からどれだけ離れているかを表す指標である。

$$PSB[t] = |\bar{X}_t - \bar{X}_p| / \hat{\sigma}_p$$

ここで、 \bar{X}_p と $\hat{\sigma}_p$ は対象とする集団の平均値と標準偏差である。ATEの場合、全ての群を併合した平均値と標準偏差は下記のように計算できる。

$$\bar{X}_p = \frac{\bar{X}_1 * n_1 + \bar{X}_2 * n_2 + \dots + \bar{X}_T * n_T}{n_1 + n_2 + \dots + n_T}$$

$$\hat{\sigma}_p = \sqrt{\frac{\hat{\sigma}_1 * (n_1 - 1) + \hat{\sigma}_2 * (n_2 - 1) + \dots + \hat{\sigma}_T * (n_T - 1)}{n_1 + n_2 + \dots + n_T - T}}$$

ATT の場合、 \bar{X}_p と $\hat{\sigma}_p$ は基準とする群の平均値と標準偏差を使用する。このように PSB を定義することで、群が 2 つの場合に PSB の合計は SB と一致する。すなわち、PSB の合計値 $Total\ PSB = \sum^t PSB[t]$ は SB の拡張ととらえることができる。

3.3.2. 傾向スコアの分布図

一般化傾向スコアでは、一人の患者につき治療法と同じ数の傾向スコアが推定される。そのため、それぞれの傾向スコアの分布を群ごとに描画する。3 つの治療法を比較するのであれば 3×3 の分布を描き、傾向スコア P_A 、 P_B 、 P_C それぞれで群間のバランスを評価する。どの傾向スコアを見ても群間で同じような分布となれば、共変量のバランスがとれているといえる。それぞれの傾向スコアの分布については、2 群の場合と同様にバイオリンプロットやヒストグラム、箱ひげ図などを使用して視覚的に判断すればよい。

4. SAS による実装

4.1. 解析データセット

本章で使用する解析データセット構造を表 1 に示す。データは PSMATCH プロシジャの Help で使用しているダミーデータ[10]をもとに、3 群のデータセットを作成している。

4.2. 傾向スコア推定と重みの算出

3 群の場合の傾向スコアは、PROC LOGISTIC プロシジャを使って簡単に求めることができる。応答変数が 3 水準以上の場合、LOGISTIC プロシジャはデフォルトで順序ロジスティック回帰が実行される。MODEL ステートメントにて link=glogit を指定することで名義ロジスティック回帰の値を計算できる。OUTPUT ステートメントで指定した名前のデータセットを作成する。オプションで predprobs=individual を指定することで、症例ごとに各群に割り当てられる確率（傾向スコア）が取得できる。

```
proc logistic data=drugs noprint;
  class Gender;
  model Drug=Age BMI Gender / link=glogit;
  output out=ps00 predprobs=individual;
run;
```

表 1. 本章で想定する解析データセット

PatientID	Drug	Gender	Age	BMI
1	Drug_A	Female	36	25.53
2	Drug_A	Male	28	24.69
3	Drug_B	Male	43	27.56
4	Drug_C	Female	45	22.48
5	Drug_A	Male	37	25.99
6	Drug_C	Male	40	26.23
⋮	⋮	⋮	⋮	⋮
485	Drug_B	Male	35	23.37
486	Drug_A	Male	34	19.24

PatientID : 患者 ID、Drug : 治療群 (DrugA/DrugB/DrugC)、Gender : 性別 (Female/Male)、Age : 年齢

ATE と ATT の重みは、データステップで計算する必要がある。ATE の重みが欲しい場合には実際の治療群に対応した傾向スコアの逆数、ATT の重みが欲しい場合には実際の治療群に対応した傾向スコアを基準群の傾向スコアで割った値を重みの列に格納する。

```

data ps01;
  set ps00;
  *ATE weightの場合;
  if Drug='Drug_A' then ATEw=1/IP_Drug_A;
  else if Drug='Drug_B' then ATEw=1/IP_Drug_B;
  else if Drug='Drug_C' then ATEw=1/IP_Drug_C;
  *ATT weightの場合 (Drug_Aを基準群とする) ;
  if Drug='Drug_A' then ATTw=1;
  else if Drug='Drug_B' then ATTw=IP_Drug_B/IP_Drug_A;
  else if Drug='Drug_C' then ATTw=IP_Drug_C/IP_Drug_A;
run;

```

4.3. 標準化バイアスの確認

ATE weight でデータを重み付けする前後の標準化バイアスを確認する。標準化バイアスを含む共変量の要約表を出力するマクロ (CalculatePSB.sas) を作成した。マクロファイルを読み込んだ後、下記のプログラムを実行することで共変量のバランスを評価するための表を得ることができる。

表 2. 傾向スコアと重みを格納したデータセット

PatientID	Drug	Gender	Age	BMI	IP_Drug_A	IP_Drug_B	IP_Drug_C	ATEw	ATTw
1	Drug_A	Female	36	25.53	0.32089	0.38795	0.29116	3.11635	1.0000
2	Drug_A	Male	28	24.69	0.85963	0.11874	0.02162	1.16329	1.0000
3	Drug_B	Male	43	27.56	0.50174	0.31896	0.17930	3.13523	0.6357
4	Drug_C	Female	45	22.48	0.03179	0.31310	0.65511	1.52647	20.6087
5	Drug_A	Male	37	25.99	0.65584	0.25133	0.09283	1.52477	1.0000
6	Drug_C	Male	40	26.23	0.54563	0.30822	0.14615	6.84213	0.2679
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
485	Drug_B	Male	35	23.37	0.57216	0.31132	0.11652	3.2121	0.544
486	Drug_A	Male	34	19.24	0.34399	0.44838	0.20763	2.9071	1.000

PatientID : 患者 ID、Drug : 治療群 (Drug_A / Drug_B / Drug_C)、Gender : 性別 (Female / Male)、Age : 年齢、IP_Drug_A : Drug_A の傾向スコア、IP_Drug_B : Drug_B の傾向スコア、IP_Drug_C : Drug_C の傾向スコア、ATEw : ATE weight、ATT : ATT weight

```
%CalculatePSB(data = /* 入力データセット */,
               out = /* 結果を出力するデータセット */,
               var_numeric = /* 連続量の共変量1 連続量の共変量2 ... */,
               var_category = /* カテゴリの共変量1 カテゴリの共変量2 ... */,
               group = /* 群を格納した変数名 */,
               w = /* 重みを格納した変数名 */
               );
```

data=には 4.2 項で作成した ps01 を入力する。Age と BMI は連続量であるため var_numeric=Age BMI とし、カテゴリ変数である Gender は var_category=に記載する。group=Drug で Drug を群に指定する。重み付け前と重み付け後の結果を比較するため、w=を指定しない場合と w=ATEw を指定した場合でそれぞれ実行する。

出力結果では、それぞれの共変量を、連続量においては平均と標準偏差、離散値においては割合で要約している (表 3 及び表 4)。2 群の傾向スコア解析で SMD>0.2 に印を付記する可能性があることに倣って、各群の標準化バイアスが 0.2 を超える場合には要約値の横に*印を出力する。Total PSB は 3 つの群の標準化バイアスの合計値である。2 群の場合には total PSB は群間差を標準化した値として解釈することが可能である。重み付け前は多くの項目で標準化バイアスが 0.2 以上であることから、群によって ATE 集団と異なる患者背景を持っていたことがわかる。重み付け後では標準化バイアスはいずれも 0.2 未満であり、どの群も ATE 集団から大きく外れていない。また標準化バイアスの合計値も重み付け前に比べて小さくなっており、共変量調整がうまくいっていることを示している。

表 3. 重み付け前の共変量の要約

変数名		Drug_A	Drug_B	Drug_C	total PSB
n		157	168	161	
Age	(Mean (SD))	35.68 (5.708) *	39.87 (6.134)	42.65 (6.034) *	1.241
BMI	(Mean (SD))	24.34 (1.997) *	23.77 (1.868)	23.65 (1.992)	0.429
Gender	Female (%)	29.3 *	46.4	60.2 *	0.662
	Male (%)	70.7 *	53.6	39.8 *	0.662

Total PSB : 全ての群の PSB の合計値。

群ごとに PSB が 0.2 を超える場合、要約値の横に*印を付記。

表 4. 重み付け後の共変量の要約

変数名		Drug_A	Drug_B	Drug_C	total PSB
n		490	485	458	
Age	(Mean (SD))	39.21 (6.275)	39.00 (6.762)	40.43 (5.895)	0.276
BMI	(Mean (SD))	23.76 (2.017)	23.86 (1.841)	23.96 (1.948)	0.105
Gender	Female (%)	44.9	43.2	48.2	0.110
	Male (%)	55.1	56.8	51.8	0.110

Total PSB : 全ての群の PSB の合計値。

群ごとに PSB が 0.2 を超える場合、要約値の横に*を付記。

以上の結果が `out=` で指定したデータセット名で作成される他に、2 つのデータセットが作成される。「`_sb_table`」には計算結果の詳細な値が保存されている。連続量の場合は例数、(重み付けした場合は) 重みの合計値、平均、標準偏差、最小値、最大値が要約値として格納されている。離散値の場合は要約値として度数と割合、水準の度数を合計した値が格納されている。各群の標準化バイアスの他に、対象とする集団の平均値 \bar{X}_p と標準偏差 $\hat{\sigma}_p$ は `P_MEAN` と `P_STD` の列に格納されている。また、`_sb_table` のデータセット中では全ての群の組み合わせの `SMD` も計算しているため、必要に応じて使用可能である。群の名前は内部的に `Group1`、`Group2`、...、と変換して処理されているため、実際の群との対応は「`_group`」データセットを確認するとよい。

4.4 傾向スコアの分布の確認

3 群以上の場合に適した傾向スコアのバイオリンプロットを描画するマクロ (`violin.sas`) を作成した。マクロファイルを読み込んだ後、下記のプログラムを実行することでバイオリンプロットを描画する。

```

%violin(data = /* 入力データセット */,
         out = /* 出力する画像の名前 */,
         var = /* 傾向スコアが格納された変数名 */,
         group = /* 群を格納した変数名 */,
         w = /* 重みを格納した変数名 */
        );

```

data=には4.2項で作成した ps01 を入力する。out=で指定したファイル名の png ファイルが作成される。

Drug_A、Drug_B、Drug_Cそれぞれの傾向スコアを描画するため、var=IP_Drug_A IP_Drug_B IP_Drug_C とスペースを空けて3つの変数名を入力する。group=Drug でDrugを群に指定する。重み付け前と重み付け後の結果を比較するため、w=を指定しない場合と w=ATEwを指定した場合でそれぞれ実行する。

調整前後のバイオリンプロットを図1、2に示す。上段はDrug_Aの傾向スコアの確率密度を実際に治療された群ごとに描いている。傾向スコアのため値の範囲は0~1であり、1に近いほどDrug_A群に割り当てられる確率が高い患者背景を持つ。同様に中段、下段ではDrug_B、Drug_Cの傾向スコアの分布が描かれている。Drug_Aの傾向スコアに着目してみると、調整前では3つの群で分布の形状が異なる。Drug_Aの群は傾向スコアの高い症例が多く、Drug_BやDrug_Cでは傾向スコアの低い症例が多くを占めている。一方、調整後ではやや凹凸はあるものの概ね3つの分布は近い形をしている。このことから、Drug_Aの傾向スコアに対して3つの群でバランスが調整されたといえる。Drug_BやDrug_Cの傾向スコアも重み付けにより群間で近い形状となっており、全体を通して患者背景のバランスが改善したことがうかがえる。

図1. 調整前のバイオリンプロット

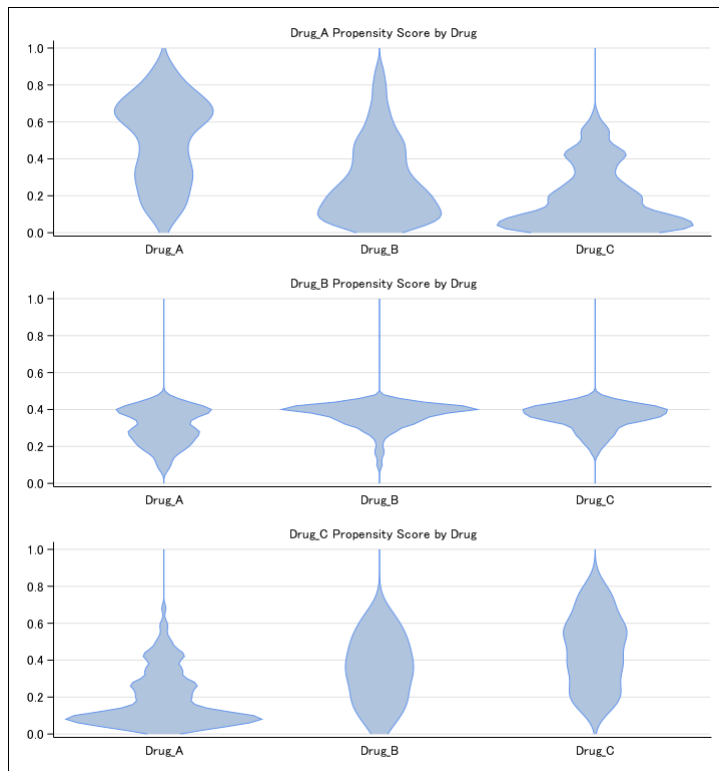
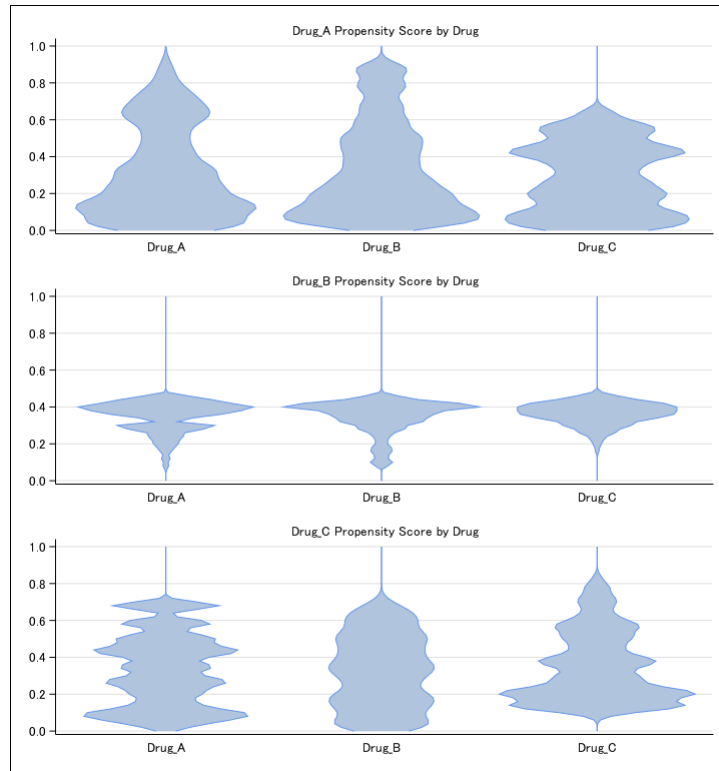


図 2. 調整後のバイオリンプロット



なお、2群で傾向スコア解析を実施する場合には群2の傾向スコア = 1 - 群1の傾向スコアとなるため、傾向スコアはどちらか一方のみを確認すれば十分である。その場合は var= に傾向スコアの変数を1つだけ記載することで不要な出力を削減できる。

5. まとめ

3つ以上の治療法を比較する場合に一般化傾向スコアを使用して共変量を調整する方法を紹介した。傾向スコア推定は実装の簡便さを優先して名義ロジスティック回帰モデルで推定した。IPTWの重みはデータステップにてATE、ATTの値を計算した。共変量バランスの評価は標準化バイアスと傾向スコアの分布によって実施することとし、それぞれ実装のためのマクロを作成した。標準化バイアスは、2群の傾向スコア解析で一般的に使用される標準化平均差を3群以上に拡張した値になるように計算式を再定義し、3群以上の場合にも表が煩雑にならないような表示方法を提案した。先行研究を基に標準化バイアスが0.2未満であることを共変量バランスがとれている基準として使用したが、標準化差と同様に一意な決まりがあるわけではない。また、標準化バイアスの合計値は2群のときに標準化差の絶対値と一致するが、3群以上の場合の値についての詳細な検討は実施できていない。本稿では傾向スコアの推定と共変量の調整、バランスの評価をそれぞれのステップで実施しているため、例えば別の方法で傾向スコアを推定した場合や共変量の調整にマッチングを使用した場合などに、共変量バランスの評価のみに紹介した手順の一部を使用することも可能である。

参考文献

- [1] 魚住龍史, 矢田真城, 山本倫生, 川口淳. SASによる傾向スコアマッチング. SAS ユーザー総会 論文集. 263-281, 2017.
- [2] 下川敏雄. 観察研究と傾向スコア分析, 医学統計セミナー, 2018.
- [3] 星野崇宏 (2009) 「調査観察データの統計科学—因果推論・選択バイアス・データ融合」, 岩波書店.
- [4] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009 Nov 10;28(25):3083-107.
- [5] Benedetto U, Head SJ, Angelini GD, Blackstone EH. Statistical primer: propensity score matching and its alternatives. Eur J Cardiothorac Surg. 2018 Jun 1;53(6):1112-1117.
- [6] Leite W. Practical propensity score methods using R. SAGE Publications; 2016.
- [7] McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med. 2013 Aug 30;32(19):3388-414.
- [8] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician. 1985;39:33–38.
- [9] Zhang Z, Kim HJ, Lonjon G, Zhu Y; written on behalf of AME Big-Data Clinical Trial Collaborative Group. Balance diagnostics after propensity score matching. Ann Transl Med. 2019 Jan;7(1):16.
- [10] SAS Institute Inc. SAS/STAT(R) User's Guide. SAS Institute Inc., Cary, NC, 2024.

付録. SAS プログラム

標準化差を計算するマクロ (calculatePSB.sas)、バイオリンプロットを描画するマクロ (violin.sas) をプログラム内で使用するため、ダウンロードしておく必要がある。

```
*-----;
*-- データセットを準備 -----;
*-----;
...省略

*-----;
*-- 傾向スコアを算出とATE weightを算出 -----;
*-----;

proc logistic data=drugs noprint;
  class Gender;
  model Drug=Age BMI Gender / link=glogit;
  output out=ps00 predprobs=individual;
run;
```

```

data ps01;
  set ps00;
  *ATE weightの場合;
  if Drug='Drug_A' then ATEw=1/IP_Drug_A;
  else if Drug='Drug_B' then ATEw=1/IP_Drug_B;
  else if Drug='Drug_C' then ATEw=1/IP_Drug_C;
  *ATT weightの場合 (Drug_Aを基準群とする);
  if Drug='Drug_A' then ATTw=1;
  else if Drug='Drug_B' then ATTw=IP_Drug_B/IP_Drug_A;
  else if Drug='Drug_C' then ATTw=IP_Drug_C/IP_Drug_A;

  label IP_Drug_A='Drug_A Propensity Score'
        IP_Drug_B='Drug_B Propensity Score'
        IP_Drug_C='Drug_C Propensity Score'
        ATEw='ATE weight'
        ATTw='ATT weight';
run;

proc print data=ps01 (obs=10) label;
  var PatientID Drug Gender Age BMI
      IP_Drug_A IP_Drug_B IP_Drug_C
      ATEw ATTw;
run;

*-----;
*-- 標準化差の確認 -----;
*-----;
** 標準化差を計算するマクロを使用 ;
%inc "C:\#####\CalculatePSB.sas";

*重み付け前のデータから計算される標準化バイアス;
%CalculatePSB(data = ps01,
              out = SMD_origin,
              var_numeric = Age BMI,
              var_category = Gender,
              group = Drug);

*ATE weightで重みつけた標準化バイアス;

```

```
%CalculatePSB(data = ps01,  
               out = SMD_IPTW,  
               var_numeric = Age BMI,  
               var_category = Gender,  
               group = Drug,  
               w = ATEw);  
  
*-----;  
*-- 傾向スコアの分布を確認 -----;  
*-----;  
** 傾向スコアの分布を描画するマクロ ;  
%inc "C:¥#####¥violin.sas";  
  
%violin(data = ps01,  
         out = violin plots_origin,  
         var = IP_Drug_A IP_Drug_B IP_Drug_C,  
         group = Drug);  
  
%violin(data = ps01,  
         out = violin plots_IPTW,  
         var = IP_Drug_A IP_Drug_B IP_Drug_C,  
         group = Drug,  
         w = ATEw);
```

SASでゲームを作る！

～Windows PowerShellを橋渡し役としたSASとWebブラウザの相互連携～

○伊住 寿雄

(ソーバル株式会社)

Making a game of SAS!

Linkage between SAS and Web Browsers Using Windows Powershell as a Bridging Tool

Izumi Toshio

Sobal corporation

要旨

SAS と Web ブラウザを相互連携させるための方法や、サンプルプログラムおよび応用として作成した簡単なゲームの紹介をする。

キーワード：

PC-SAS、 SAS Integrated Object Model、 Microsoft PowerShell、 Microsoft Forms、 Microsoft Edge WebView2、 HTML、 CSS、 JavaScript、 JSON

はじめに

SAS Integrated Object Model (以下 SAS IOM) を使用すると、Excel や Word などのようにバックグラウンドで実行できることはあまり知られていないと思われる。

SAS ユーザー総会 2015 において森田祐介氏が SAS IOM を用いて MS-Office の VBA から操作する方法を発表されていたが、VBA と似たような形の VBScript でも操作できると思われた。

しかし昨年 Microsoft 社が VBScript を非推奨とすることを発表したため、VBScript の代替として Windows PowerShell (以下 PowerShell) を用いた方法を検討し、さらに入力インターフェースとして Windows Forms (以下 Forms) 及び Microsoft Edge WebView2 (以下 WebView2) を用いて SAS-PowerShell- Web ブラウザの相互連携の仕組みを考慮してみた。

本稿では、上記の相互連携をさせるための方法や、サンプルプログラム及び応用として作成した簡単なゲームの紹介をする。

SAS Integrated Object Model とは

SAS を COM (Component Object Model) として動作させることができる SAS が持っている機能である。

COM として動作させることで、COM を扱えるプログラム言語からバックグラウンドで SAS を起動し、コードを実行することができるようになる。

ただし、SAS Integrated Object Model (以下 SAS IOM) を COM として動作させた場合、標準モード (シングルバイト、Latin1) での動作となる仕様のため、日本語等のダブルバイト文字が扱えない。

Windows PowerShell とは

米マイクロソフト (Microsoft) 社の Windows が標準で備えるコマンドライン方式のシェルおよびスクリプト実行環境の一つ。 .NET Framework 上で動作する。(IT 用語辞典 e-words より引用)。

非推奨となった VBScript の代替言語として今後 PowerShell がこれまでよりも多く使われる可能性がある。

Windows PowerShell(以下 PowerShell)と SAS IOM との連携

- PowerShell で SAS IOM を COM として呼び出す。
- PowerShell から SAS IOM へデータ送信

SAS.Workspace.LanguageService.Submit(\$code)を使用して SAS プログラムコードを送信

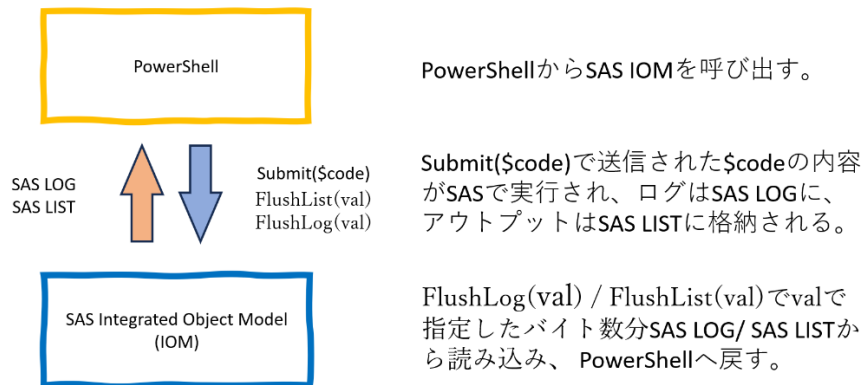
※\$code は SAS のプログラムコードを格納した変数

- SAS IOM から PowerShell へデータ受信

データ送信後、SAS.Workspace.LanguageService.FlushList(val)を使用して SAS LIST から出力データを読み出す。

なお、SAS.Workspace.LanguageService.FlushLog(val)を使用すると、SAS LOG からログを読み出すことができる。開発時このログを PowerShell のコンソールに出力させるようにしておくと、デバッグはやりやすくなる。

※val は数値。val で指定したバイト数分読み込む。val の省略はできないため、数値を多めにしておくか、読み込むデータがなくなるまで繰り返し読み込むような処理にする必要がある。



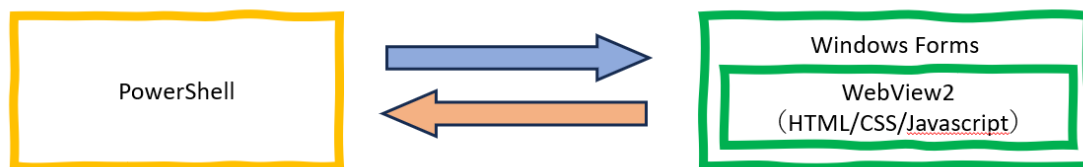
Microsoft Edge Webview2 とは

Microsoft .NET ベースの Windows アプリケーションに組み込むことができる部品化されたプログラム（コントロール）の一つで、Chromium ベースの Microsoft Edge の機能呼び出して Web ブラウザとして利用することができるもの。（IT 用語辞典 e-words より引用）

PowerShell と Webview2 の連携

- PowerShell で Windows Forms（以下 Forms）および Webview2 を呼び出し、Forms の表示領域に Webview2 をバインドさせることで PowerShell から Webview2 を操作できる状態にする。

なお、Webview2 を呼び出すには別途 Webview2 ランタイム（dll ファイル）が必要となり、Microsoft Teams（以下 Teams）がインストールされていれば、Teams のインストールフォルダに存在するので、それを使用する。Teams がインストールされていなければ、配布されている Webview2 ランタイムをインストールして使用する。



- PowerShell から Webview2 へデータ送信

PowerShell 側は Microsoft.Web.WebView2.WinForms.WebView2 インスタンスの

CoreWebView2.PostWebMessageAsString メソッドを使用してテキストデータを送信する。

```
$webview.CoreWebView2.PostWebMessageAsString($TextData)
```

※\$webview は Microsoft.Web.WebView2.WinForms.WebView2 のインスタンス

※\$TextData はテキストデータ

Webview2 側はブラウザ内で操作する JavaScript を用いてイベントリスナーを登録し、受信時の処理を行う。イベントリスナーの登録及び受信時の処理は JavaScript で以下のように記述する。

```
window.chrome.webview.addEventListener('message', function(event){
  実行時の処理をここに記述
})
```

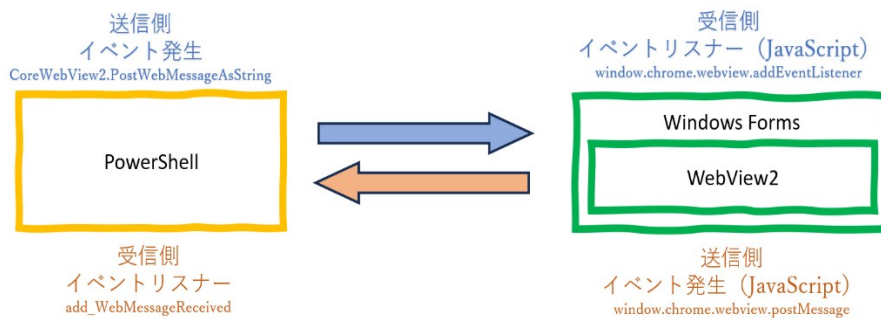
● Webview2 から PowerShell へデータ送信

Webview2 側はブラウザ内で操作する JavaScript を用いてデータを送信する。送信時の処理は JavaScript で以下のように記述する。

```
window.chrome.webview.postMessage(送信するデータ);
```

PowerShell 側は Microsoft.Web.WebView2.WinForms.WebView2 インスタンスの add_WebMessageReceived でイベントリスナーを登録し、受信時の処理を行う。

```
$webview_MessageReceived = {
  実行時の処理をここに記述
})
$webview.add_WebMessageReceived($webview_MessageReceived)
※$webview は Microsoft.Web.WebView2.WinForms.WebView2 のインスタンス
```



双方ともにテキストデータを送受信するが、複数のパラメータを送受信したりすることを考えると JSON 形式のテキストデータで行うのが良いと思われる。

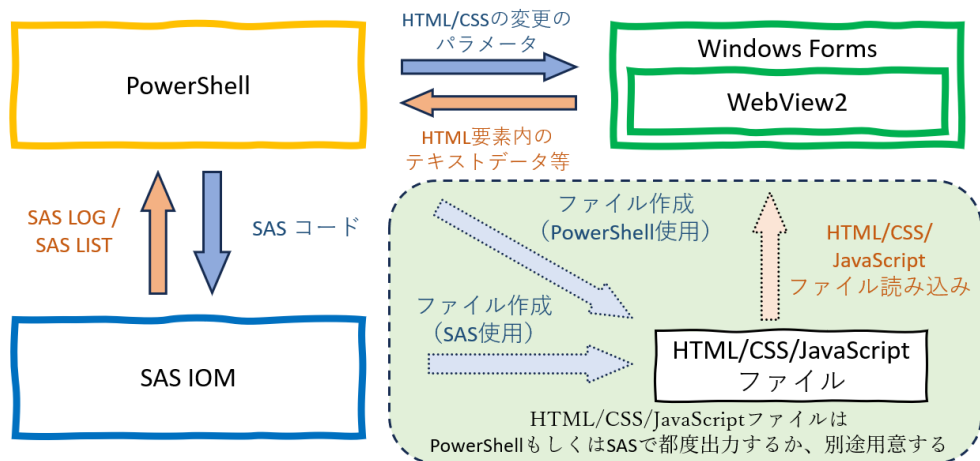
そして、受信側で JSON をパースすることにより、複数のパラメータが扱える。

PowerShell の受信側 : `$json = ($JsonData.WebMessageAsJson | ConvertFrom-Json)`

JavaScript の受信側 : `var jsondata = JSON.parse(event.data);`

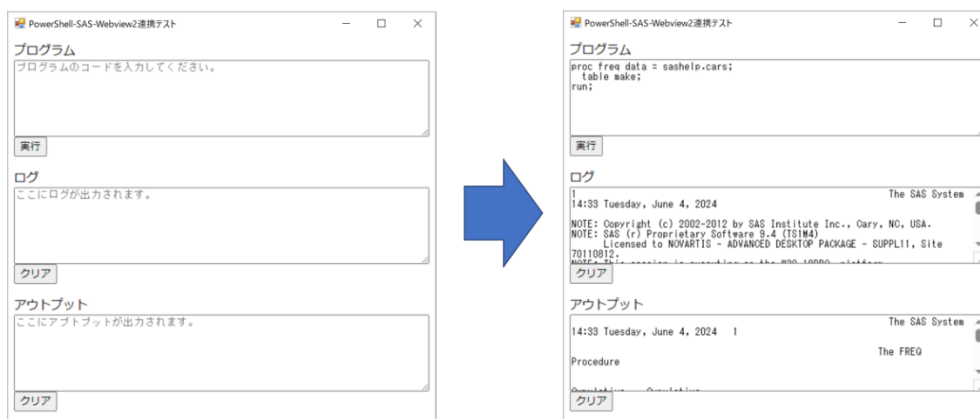
PowerShell と SAS IOM と Webview2 の連携

前述のように PowerShell を基準として、PowerShell-SAS IOM 間および PowerShell- Webview2 間の操作ができるようになり、以下のようなデータのやり取りが可能になる。



サンプルプログラム

サンプルプログラムとして、簡単なツールを紹介する。このツールは、ブラウザのプログラム入力欄に SAS のプログラムコードを入力し実行ボタンをクリックすると、SAS の実行ログ及びアウトプットがブラウザのログ欄及びアウトプット欄にそれぞれ表示される単純なものである。



サンプルプログラムのコードは別添の SAS_PowerShell_WebView2_test.ps1 を参照。

以下、プログラム内の各部分（番号が振ってあるコメント）について簡単に説明する。

```
<# 1. SAS ワークスペース・オブジェクトの取得 #>
$ObjWrkSp = New-Object -ComObject SAS.Workspace
```

1. SAS IOM を COM として起動し、オブジェクト変数 \$ObjWrkSp に格納する。以降 SAS とのデータのやり取りは、この変数を使って行う。

```
<# 2. SAS ワークフォルダのパス取得 #>
$code = @"
data _null ;
  file print noprint LINESIZE = 10000 PAGESIZE = 10000;
  workpath = pathname ( "work" );
  put @2 workpath;
```

```
run;
"@
$ObjWrkSp.LanguageService.Submit($code)
$WorkPath = [string]$ObjWrkSp.LanguageService.FlushList(100000).trim()
```

2. SAS のワークフォルダを使用するため、SAS との送受信を行い、SAS ワークフォルダのパスを取得する。
このフォルダは後述する表示用 HTML ファイルや Webview2 の一時ファイルの格納用として使用する。

```
<# 3. 表示用 HTML 作成 #>
$HTMLdata = @"
<html>
<head>
<style>
    textarea { width:100%; height:100px; }
</style>
<script>
//PowerShell からのデータ取得
window.chrome.webview.addEventListener('message', function(event){
    var jsondata = JSON.parse(event.data);
    //データの種別で出力先を変更
    if(jsondata.type == "log"){
        document.getElementById('log_strings').value = jsondata.value;
    }
    if(jsondata.type == "list"){
        document.getElementById('output_strings').value = jsondata.value;
    }
});
//プログラムコードを SAS に送信
function send_sas() {
window.chrome.webview.postMessage({"program_code":document.getElementById('program_code').value});
}
//ログのクリア
function clear_log() {
    document.getElementById('log_strings').value = "";
}
//アウトプットのクリア
function clear_output() {
    document.getElementById('output_strings').value = "";
}
</script>
</head>
<body>
<p>
    <label>プログラム</label>
    <textarea id="program_code" placeholder="プログラムのコードを入力してください。"></textarea>
    <button id="submit" onclick="send_sas()" type="button">実行</button>
</p>
<p>
    <label>ログ</label>
    <textarea id="log_strings" placeholder="ここにログが出力されます。"></textarea>
    <button id="submit" onclick="clear_log()" type="button">クリア</button>
</p>
<p>
    <label>アウトプット</label>
    <textarea id="output_strings" placeholder="ここにアウトプットが出力されます。"></textarea>
    <button id="submit" onclick="clear_output()" type="button">クリア</button>
</p>
</body>
```

```

</html>
"@
$HTMLPath = (Join-Path $WorkPath "SAS_PowerSell_WebView2_test.html")
#HTML ファイル書き出し
Write-Output $HTMLdata | Out-File $HTMLPath

```

3. 表示用の HTML ファイルを SAS ワークフォルダに書き出す。書き出す HTML ファイルには PowerShell とのデータを送受信するための JavaScript のコードも記述しておく。

```

<# 4. WebView2 の各種 DLL 取得および定義 #>
#WebView2 の各種 DLL があるフォルダパス
$WebView2Path = " (Webview2 がインストールされているパス) "
#WebView2 の各種 DLL 取得
[void][reflection.assembly]::LoadFile((Join-Path $WebView2Path "Microsoft.Web.WebView2.WinForms.dll"))
[void][reflection.assembly]::LoadFile((Join-Path $WebView2Path "Microsoft.Web.WebView2.Core.dll"))
#WebView2 の定義
$webview = [Microsoft.Web.WebView2.WinForms.WebView2]@{
    Location    = New-Object System.Drawing.Point(0, 0)
    Name        = 'webview'
    TabIndex    = 0
    ZoomFactor  = 1
    Dock        = [System.Windows.Forms.DockStyle]::Fill
    CreationProperties = New-Object 'Microsoft.Web.WebView2.WinForms.CoreWebView2CreationProperties'
}
#WebView2 が作成する一時フォルダの場所
$webview.CreationProperties.UserDataFolder = (Join-Path $WorkPath "data")

```

4. Webview2 を使用するための Webview2 ランタイムの DLL ファイル（なければ別途インストール）の読み込みや Webview2 の各種定義を行う。また、Webview2 が作成する一時ファイル類は SAS ワークフォルダ内の data フォルダを指定する。（フォルダが存在しない場合、実行時に新規に作成される）

```

<# 5. WebView2 (=JavaScript) からのメッセージ受信 #>
$webview_MessageReceived = {
    #引数取得
    param (
        $WebMessageKey,
        $JSONData
    )
    $ReceivedJSON = ($JSONData.WebMessageAsJson | ConvertFrom-Json)
    $Submitcode = $ReceivedJSON.program_code

    #SAS 実行
    $ObjWrkSp.LanguageService.Submit($Submitcode)

    #ログ取得
    $ReceivedLog = ""
    do {
        $PreReceivedLog = $ObjWrkSp.LanguageService.FlushLog(100000)
        $ReceivedLog += $PreReceivedLog
    }
    until ( [String]::IsNullOrEmpty( $PreReceivedLog ) )
    #ログを JSON データに変換
    $SubmitJSON = ( @{"type"="log"; "value"="$ReceivedLog"} | ConvertTo-Json)
    #ログを Webview2 に送信
    $webview.CoreWebView2.PostWebMessageAsString($SubmitJSON)
}

```

```

#アウトプット取得
$ReceivedList = ""
do {
    $PreReceivedList = $ObjWrkSp.LanguageService.FlushList(100000)
    $ReceivedList += $PreReceivedList
}
until ( [String]::IsNullOrEmpty( $PreReceivedList ) )
#アウトプットを JSON データに変換
$SubmitJSON = ( @{"type"="list"; "value"="$ReceivedList"} | ConvertTo-Json)
#アウトプットを Webview2 に送信
$webview.CoreWebView2.PostWebMessageAsString($SubmitJSON)
}
#イベントリスナ登録
$webview.add_WebMessageReceived($webview_MessageReceived)

```

5. WebView2 からのメッセージ受信時の処理を定義する。ブラウザ側でボタンがクリックされたとき、JavaScript によって JSON 形式でメッセージが送られると、ここで定義された処理が実行される。

```

<# 6. Windows Forms 取得および定義 #>
#アセンブリの読み込み
Add-Type -Assembly System.Windows.Forms
Add-Type -AssemblyName System.Drawing
#Forms のオブジェクト取得
$form = New-Object System.Windows.Forms.Form
#Forms の定義
$form.Text = "PowerShell-SAS-Webview2 連携テスト"
#Forms の読み込み時設定
$formLoaded = {
    $webview.Source = $HTMLPath
    $webview.Visible = $true
}
$form.add_Load($formLoaded)
#Forms の解放時設定
$formUnloaded = {
    $form.remove_Load($formLoaded)
    $form.remove_FormClosed($formUnloaded)
}
$form.add_FormClosed($formUnloaded)

```

6. WebView2 を使用するための Forms の読み込み及び各種定義を行う。また、Forms 読み込み時および Forms が閉じられる時の処理を設定する。

Forms 読み込み時:7.で Forms と Webview2 をバインドさせるが、Forms を読み込んで表示させる際の、Webview2 の表示非表示設定と読み込む HTML ファイルのパスを指定する。

Forms が閉じられる時 : Forms と Webview2 のバインドを解除する。

```

<# 7. Forms に WebView2 を追加して Forms を表示 #>
$form.SuspendLayout()
$form.AutoScaleMode = 'Font'
$form.ClientSize = New-Object System.Drawing.Size(800, 600)
$form.Controls.Add($webview)
$form.ResumeLayout()
#Forms を表示
[void]$form.ShowDialog()

```

7. Forms に Webview2 をバインドさせ、Forms を表示させる。

```
<# 8. 各オブジェクトの解放 #>
```

```
#webview2
```

```
$webview.Dispose()
```

```
$webview = $null
```

```
#form
```

```
$form.Dispose()
```

```
$form = $null
```

```
#3 秒処理を停止 (webview2 プロセスが終了するのを待つため。プロセスが終了していないと一時フォルダが削除できない)
```

```
Start-Sleep -Seconds 3
```

```
#一時作成のファイル/フォルダ類
```

```
remove-item (Join-Path $WorkPath "data") -Recurse
```

```
remove-item $HTMLPath
```

```
#SAS
```

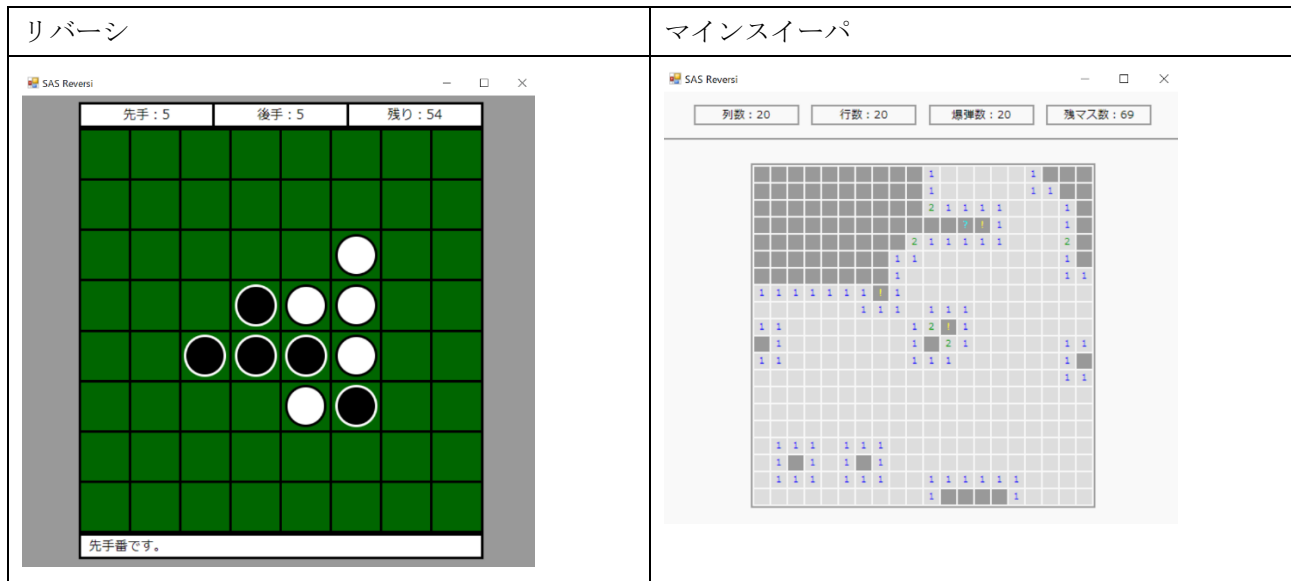
```
[void]$ObjWrkSp.Close
```

```
$ObjWrkSp = $null
```

8. Forms が閉じられた後、使用したオブジェクト類を解放する処理や一時的に作成されたファイル類を削除する。この処理を行わないと、使用したプロセスが残り、メモリーリークの原因となるため。

応用例

応用例として簡単なゲームを作成してみたので紹介する。



これらは主要なデータを SAS 側に持たせ、PowerShell を介してブラウザからの入力情報を SAS に送り、SAS 側では送られてきたデータを用いて処理を行った後、PowerShell を介して表示変更の情報などをブラウザに送る。

これらのゲームの使い方等は参考資料として別添のファイルに記載してあるので、そちらを参照していただきたい。

まとめ

PowerShell を介して SAS とブラウザの相互連携ができるので、この仕組みを応用すれば、SAS が苦手な GUI を使用したツールなどが開発できるようになると思われる。

また、SAS IOM ではシングルバイトしか扱えないため、直接日本語データを扱うことができないという制限はあるが、PowerShell で日本語テキストをコード値に変換後 SAS に送信し、SAS から帰ってきたデータを PowerShell で日本語テキストにデコードするなどの対応をすれば、手間はかかるが見た目上、日本語のデータを扱うことができそうではある。

参考資料

- 森田 祐介 (2015) SAS ユーザー総会 2015
SAS Integrated Object Model 活用によるユーザーエクスペリエンス向上への試み～MS-Excel を User Interface とした SAS Clinical Standards Toolkit の利用を事例として～

- Developing Windows Clients - SAS Help Center
<https://documentation.sas.com/doc/en/itechcdc/9.4/itechwcdg/p0xyn6hfw4e0an1dum74ovbcphw.htm>
(参照日 : 2024/08/21)

- 次期 HTA としての PowerShell+WebView2 の利用
<https://qiita.com/tkmry/items/dede9cadc0d2a2df6c1a>
(参照日 : 2024/08/21)

- WebView2 コントロールを使う | プログラムでネットサーフィン
https://web.biz-prog.net/readme/webview_new.html
(参照日 : 2024/08/21)

SAS用グラフパッケージ「SAS Plotter」

○筒香 孝輔¹

Modern graph package “SAS plotter”

¹Kosuke Tsutsugo

要旨

Python や R では少ない紙面で多くの情報を表現できるデータ可視化手法が提案され、簡単に作成ができるパッケージが公開されている。一方 SAS ユーザー間でそのような取り組みはほとんど実施されておらず、最近の SAS アップデートにおいてプロットステートメントの機能拡張は実施されていない。

「SAS Plotter」は SAS 環境下でより現代的なデータ可視化手法を手軽に利用できることを目指し開発された SAS マクロパッケージである。本発表では「SAS Plotter」の概要と利用例を紹介する。

キーワード：データビジュアライゼーション、ridgeline plot、raincloud plot、反転ヒストグラム、二次元 KDE プロット、サンキーダイアグラム、マルチヒストグラム

1.初めに

Python および R に代表されるオープンソースのデータ分析環境では、データ分析を効率的に実施できるパッケージがユーザー主体で開発されており、それらを無料で利用できる。

一方 SAS 言語ではユーザー間でコードを共有する文化自体が希薄であり、ユーザー主体で機能拡張の取り組みは活発とは言えない。ユーザーが作成したマクロを [GitHub](#) で共有されている事例は存在するが、公開されているマクロをユーザー自身で修正しなければ利用できないケースが多く、ユーザーが自由にマクロコードを変更できるためマクロコードの品質管理が困難であった。さらにオブジェクト指向が実装されておらずソースコードが公開されていない SAS 言語では、ユーザー主体の大規模開発が難しいのが実情である。主要な SAS 環境が SAS viya または SAS visual analytics となった現在では、従来 SAS 環境で利用されている `proc sgplot` や `GTL` の新規のプロットステートメント追加は過去数年にわたって実施されておらず、今後もデータ可視化の機能強化は期待できない。そのため残念なことにデータ可視化に関しては、従来の SAS 環境は他の開発環境と比較して劣っているといわざるを得ない。オープンソース言語への移行について業界として議論すべき話題ではあるだろう。

しかしながら SAS を前提とした事業プロセス、SAS 以外の言語を扱える人材不足、解析結果の再現性の観点からオープンソース言語への移行のハードルは依然として高い。そのため SAS 環境においても論文投稿にも活用できる洗練されたデータ可視化手法を少ない工数で作成する方法が求められていると思われる。

また Cdisc の取り組みによる臨床データの標準化が進んだ結果、データ可視化に必要な入力データの構造は調査に依らず一定になりつつある。これはデータ可視化の工程もある程度標準化が可能であり、汎用的な作図プログラムの開発は十分可能であると考えられる。

このような経緯から「データビジュアライゼーションに要するプログラミング工数の削減、および報告書作成の効率化」を目標に「SAS Plotter」は開発された。

「SAS Plotter」は SAS 9.4 単独で動作するグラフパッケージであり、適切なデータを用意すれば論文掲載にも活用できるグラフを簡単に作成できる。最新バージョン(Ver1.3)では7種類のグラフが作成できる。本論文では「SAS Plotter」の使用方法と実際の作図例を紹介する。

2. 「SAS Plotter」によるデータ可視化の実施例

2.1 セットアップ方法

「SAS Plotter」は version 1.3 から SAS パッケージフレームワークである「SAS PACKAGES」に正式に統合された。パッケージのインストール、ロード、ハッシュ値の確認はフレームワークを用いて実施する。パッケージのインストールは github 経由で実施可能であり Git のインストールは不要である。

「SAS PACKAGES」はマクロコードの品質管理問題の解決方法として有用であり、「ユーザー主体の機能拡張」という新しい取り組みを SAS 環境において可能にしたという点で画期的な成果だといえるだろう。

SAS PACKAGES の詳細は以下の web サイトから確認できる (Bartosz Jablonski, 2023)。

https://github.com/yabwon/SAS_PACKAGES

以下のコードを実行するだけで SAS Plotter のすべての作図マクロおよび専用カラーパレットが利用可能になる。

```
filename packages "<your path>";
filename SPFinit url
"https://raw.githubusercontent.com/yabwon/SAS_PACKAGES/main/SPF/SPFinit.sas";
%include SPFinit;

/* install sas package framework*/
%installPackage(SPFinit);

/*install SAS plotter*/
%installpackage(sasplotter);

%loadpackage(sasplotter);
```

インターネットに接続されていない環境の場合は、github から手動で各種ファイルをダウンロードする。

パッケージを格納するディレクトリを指定する必要がある。ディレクトリは `filename` ステートメントで「`packages`」という名前で指定すること。

作図事例はすべて SAS プログラムとして `github` 上で公開されている。例えば `Sankey` ダイアグラムの作図例は以下のコードを実行することで再現できる。

```
ods listing gpath="your output path";  
filename exam url  
"https://github.com/Supermanjp/SAS_Plotter/raw/main/example/sankey_example.sas" encoding='UTF-  
8';  
%include exam;
```

2.2 ridgeline plot

`ridgeline plot` は多群の応答変数の分布をカーネル密度推定(Kernel density estimation, KDE)で推定し、分布形状表示する表現手法である。x 軸は応答変数と y 軸はグループ化変数であり、密度の軸は存在しないことが特徴である。分布形状は平滑線として表現し、水平方向または垂直方向にずらして表示することにより分布の比較がしやすくなるよう工夫されている。密度そのものには興味がなく、分布形状の時間変化を可視化する場合に有用である。

「`SAS Plotter`」は応答変数とカテゴリ変数を指定するだけで自動的に座標を計算し `ridgeline plot` を簡単に作成することができる。さらに `group` パラメータを指定することでグループ変数別の応答変数の分布を同時に可視化することが可能である。

KDE による密度推定では、個別値の情報が失われてしまい要約統計量も出力されない欠点がある。「`SAS Plotter`」では四分位点ごとに平滑線を塗り分けることで4分位数を表現することができる。さらに `rug` パラメータを使用することで個別値をラグプロットとして同時表示することも可能である。

密度曲線の塗りつぶしに対応しており、群別および4分位別の塗りつぶしに対応している。

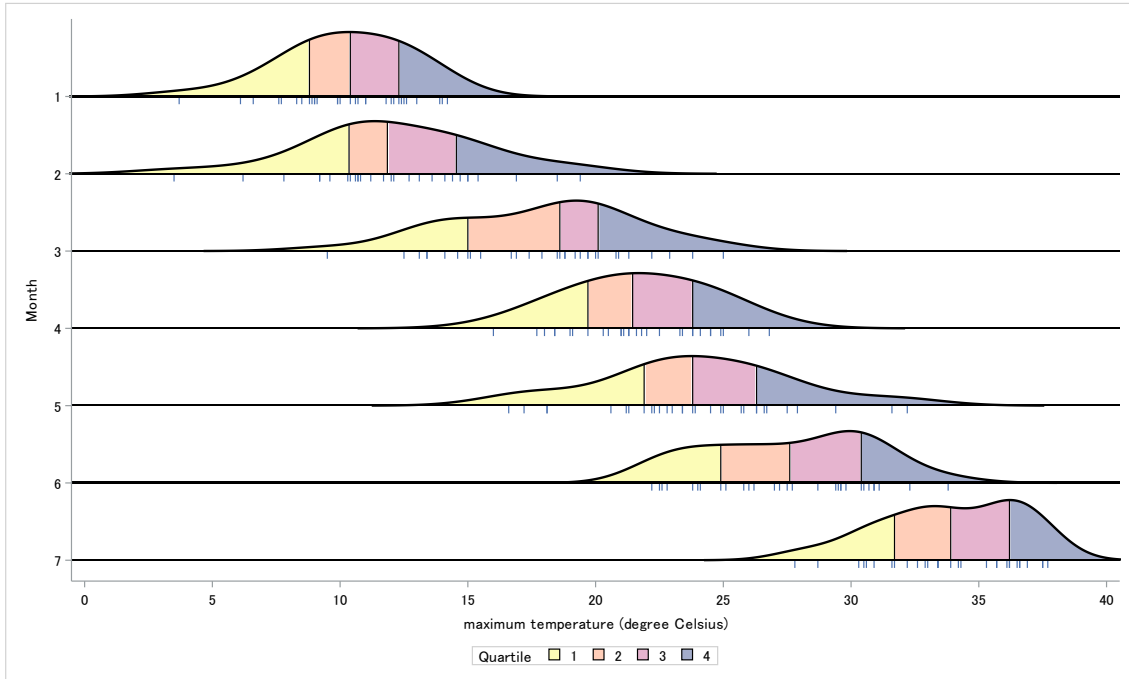


図 1 ridgeline plot の出力例（四分位ごとに塗り分け）

2.3 Raincloud plot

Raincloud plot は KDE で推定した分布形状、要約統計量、個別値をそれぞれ密度プロット、箱ひげ図、ストリッププロット（マーカー同士が重ならないようにマーカー位置を調整した 1 次元の散布図）を表示するプロットである (Micah Allen, 2021)。Ridgeline plot よりも要約統計量が見やすく群間比較がしやすい。Raincloud plot は各オブジェクトが重ならないように位置調整をする必要があるが、本パッケージではすべて自動で実施されさらに向きの切り替えや平均値の接続線の表示もパラメータの切り替えのみで実現可能である。

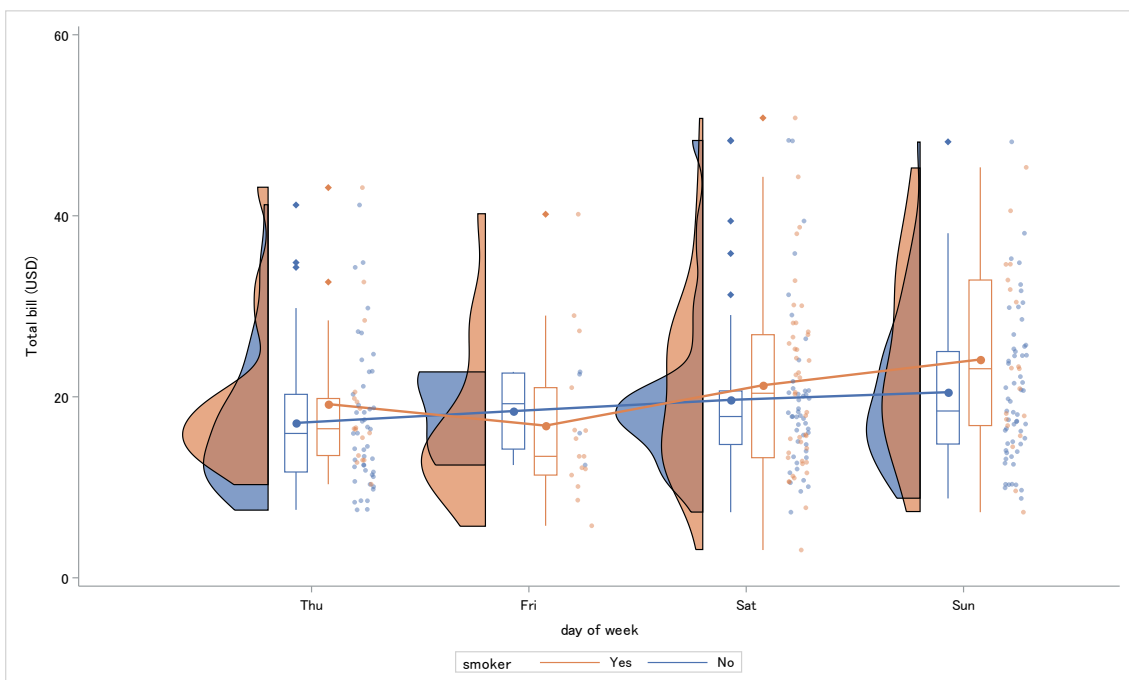


図 2 Raincloud plot の出力例

2.4 Paired Raincloud plot

Paired raincloud plot は対応のある 2 群の raincloud plot を作成することができる。2×2 クロスオーバー試験のように同一症例に対して異なる暴露を実施するデザインの試験結果の可視化に適している。

前項の raincloudplot に加えて対応のある個別値を線で結ぶ（スパゲッティプロット）機能が利用できる。

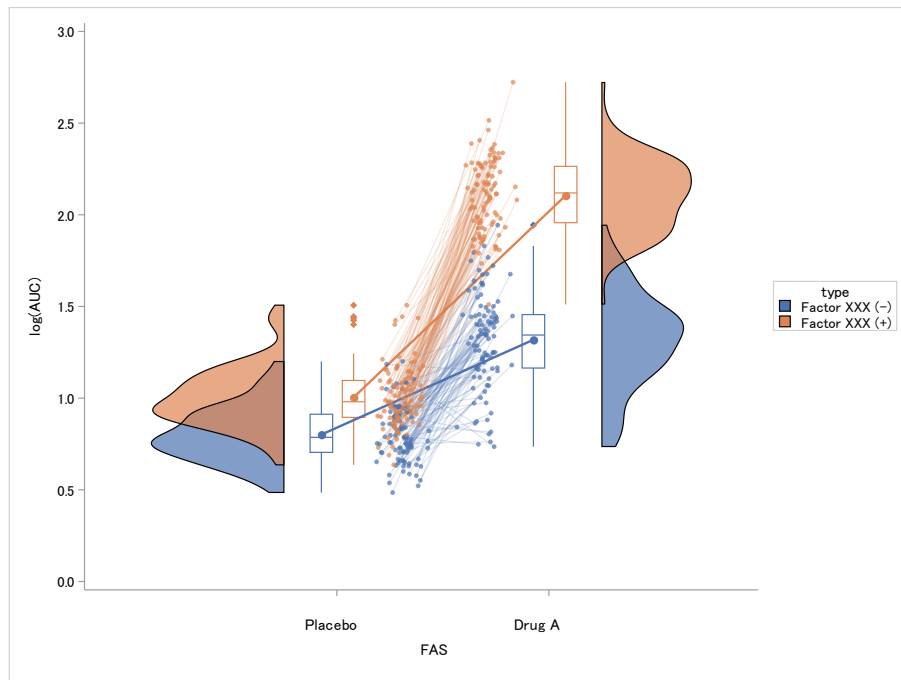


図 3 Paired raincloud plot の出力例

2.5 Mirrored histogram

Mirrored histogram（反転ヒストグラム）は 2 群の分布をヒストグラムとして表現することで分布を比較することができる表現手法である。一方の群のヒストグラムを上下または左右に反転させることでヒストグラム同士が重ならず 2 群の頻度分布を比較する場合に有用である。左右に反転させたものは「バタフライチャート」と呼ばれることもある。

Mirrored histogram は傾向スコアマッチングを用いた調査で良く用いられる。傾向スコアマッチングを実施する場合はマッチング前とマッチング後の傾向スコア分布を確認する必要があるが、Mirrored histogram を使用すると試験群と対照群の分布を比較しやすい。

「SAS Plotter」で作図する場合はあらかじめ集計済みのデータを用意する必要がある。

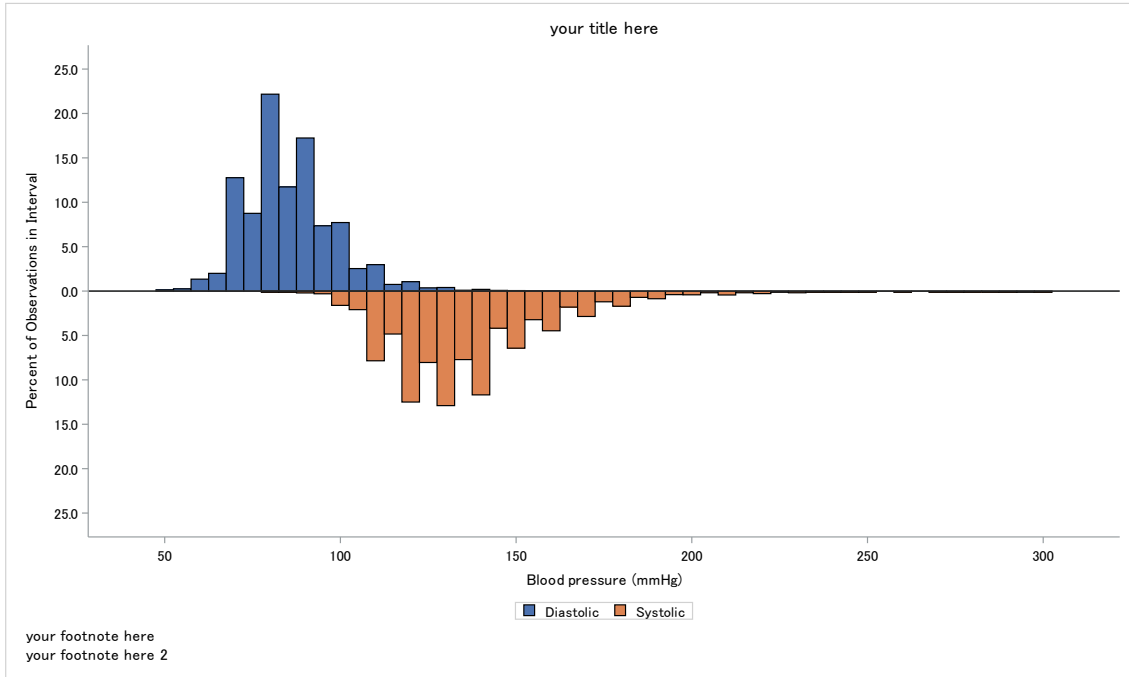


図 4 Mirrored Histogram の出力例

2.6 2D-KDE plot

2次元カーネル密度推定プロット(two-dimensional Kernel density estimation, 2D-KDE plot)は2つの連続変数間の分布を等高線として表現したグラフである。2変数間の相関関係を確認する場合は散布図を作成して確認することが一般的であるが、個別値だけではデータがどのように分布しているのかを把握しにくい欠点があった。2D-KDEを併用すれば個別値とデータの分布形状を同時に視覚化することができる。

「SAS Plotter」は散布図と等高線を同時に表示するだけでなく、平滑線とラグプロットを用いることで周辺密度と個別値を表示することも可能である。

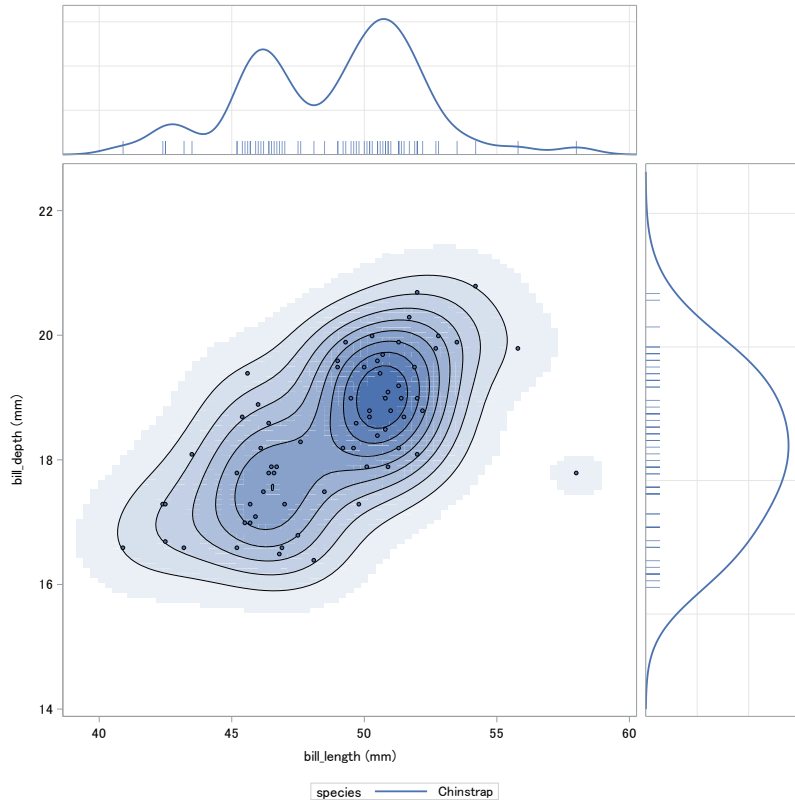


図 5 群別 2D-KDE plot の出力例 (個別値を同時表示)

2.7 Sankey Diagram

Sankey diagram (正確には Alluvial Diagram、沖積図) は患者の治療計画、疾病の進行度などの patient journey の可視化に用いられる (Nicolle M. Gatto, 2022)。近年では SAS visual analytics や tableau などの分析ソフトでも標準でサポートされている。SAS 環境でサンキーダイアグラムを作成した事例は過去にいくつか報告されているが (Chapel Hill, 2015) (Matange, 2015) (Jeffrey, 2024)、本グラフパッケージの Sankey diagram は、体裁変更、特定のリンクの強調表示および各種統計情報の表示を可能にした世界初の SAS パッケージである。

SAS visual analytics 等の他 SAS 製品や Python 等の他言語環境を用意しなくても SAS 単体で動作する。

本パッケージは臨床試験や疫学研究の利用を想定して開発されており、各ドメインのフォローアップ中症例数および割合を同時に表示することができる。執筆時点では臨床試験および疫学研究に特化したサンキーダイアグラム作成パッケージは R および python には存在しない。

Sankey diagram は隣接する domain 間の変化のみ可視化できる点に注意が必要である。例えば最初の domain で Node1 に属する患者が最終 domain でどの Node に属するかは diagram からは読み取れない。

入力データは 1 症例 1 レコードの形式で各 domain のカラムに node に対応するコード値を格納したデータを用意する。また domain および node の format を別途定義する必要がある。それ以外の前処理はパッケージで実施される。

表 1 入力データセット例

USUBJID	Day0	Day30	Day60
A001	Regimen A	Regimen B	Regimen C
B001	Regimen D	Regimen E	Regimen F

作図例を以下に示す。なお domain の間隔やテキストの体裁は format のコード値およびパラメータを修正することで変更可能である。

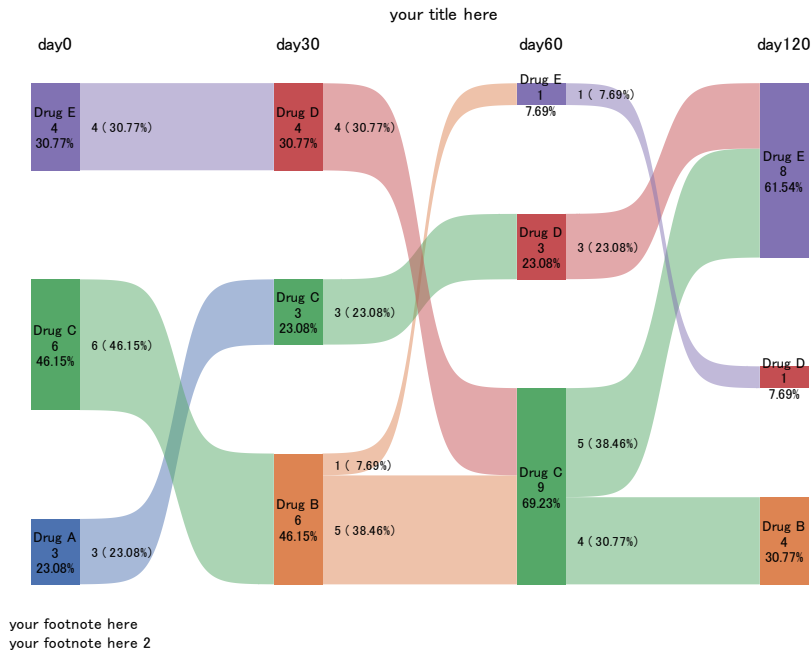


図 6 Sankey diagram の出力例

本パッケージでは疫学研究向けの独自機能を搭載している。Focus パラメータは特定の Link に着目したい場合に有用である。Focus パラメータで指定した条件式に合致する Link のみがカラー表示され、それ以外の Link はグレー表示となる。さらに条件式に合致する Link の統計量のみ表示される。

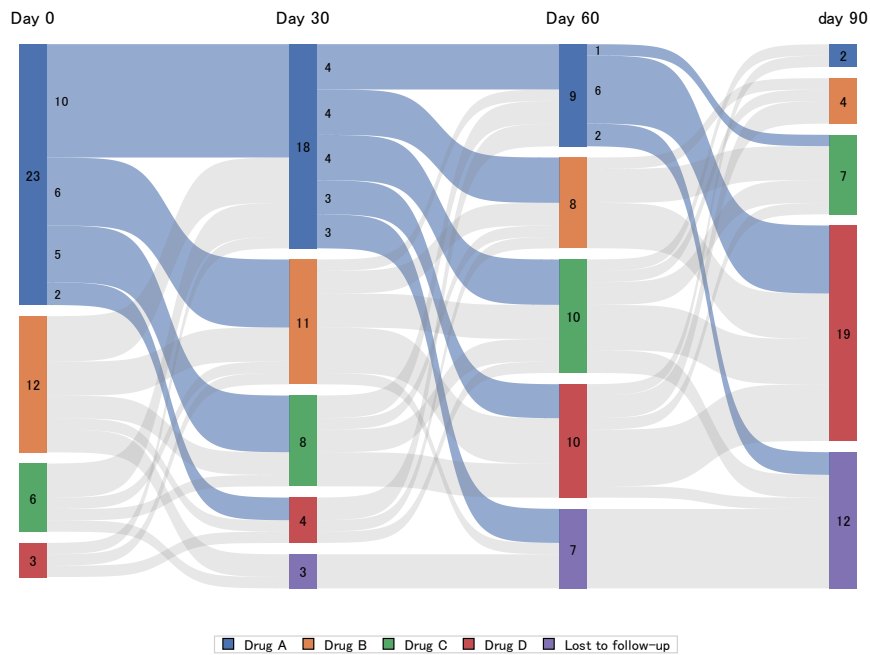


図 7 focus パラメータの活用例。Drug A を source node とする Link を強調表示した。

EndFollowup パラメータは何等かの理由でフォローアップが終了した患者集団に対応する Node を指定できる。指定するとフォローアップ終了に該当する Node はグレー表示となり、Sankey diagram 下部に各 Domain のフォローアップ中症例数を表示する。

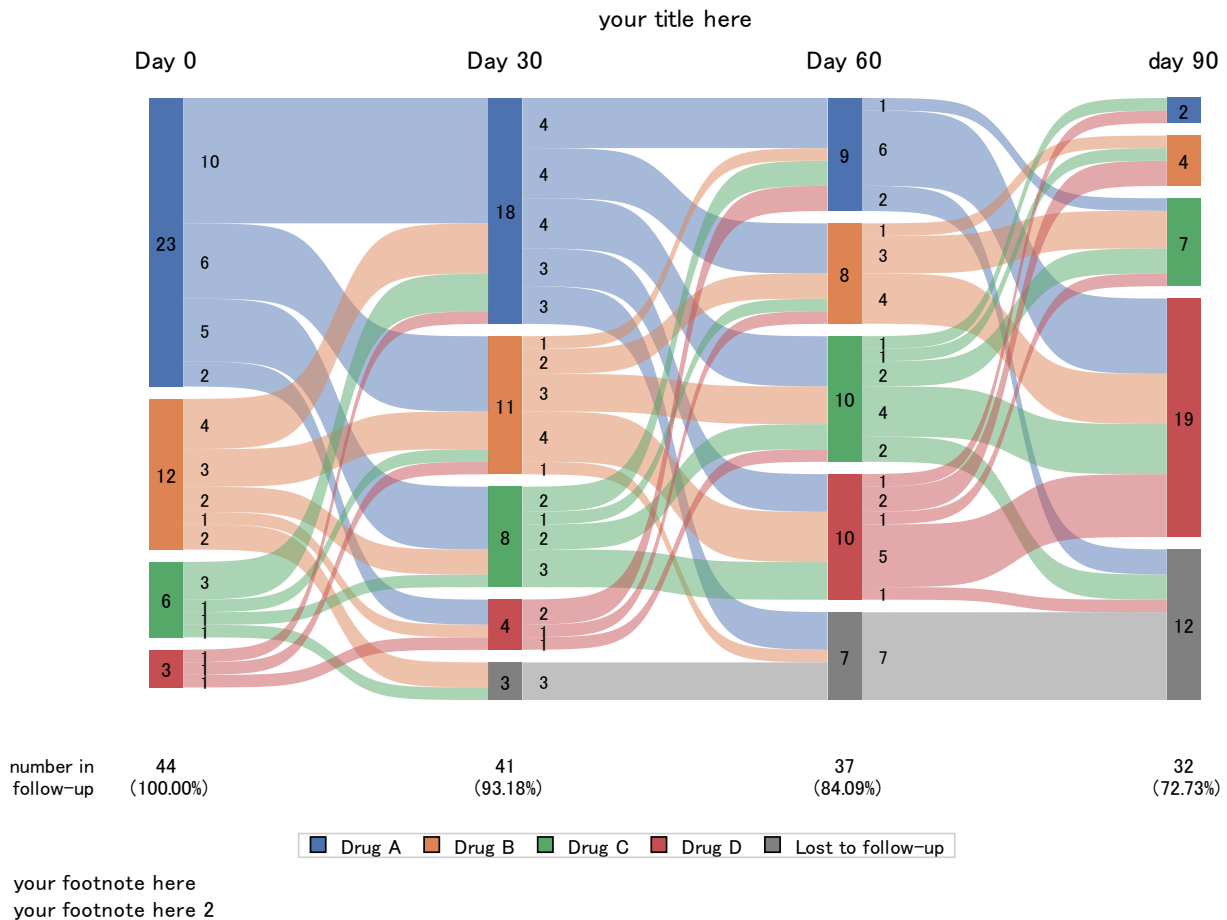


図 8 フォローアップ終了に相当する Node を指定した場合の表示例

2.8 Multihistogram

MultiHistogram は多群の離散変数の水準分布をヒストグラムとして可視化するのに有用である。

ヒストグラムは頻度を可視化するのに広く用いられる可視化手法であるが、多群のヒストグラムを比較するのが難しい欠点があった。複数のヒストグラムを重ね合わせる方法や、積み上げ棒グラフのように積み上げる方法があるが、可読性に難がある。マルチセルレイアウトを用いてヒストグラムを格子状に並べる方法が現時点では有効と思われるが、作図プログラムが複雑になりやすく必要な面積が多くなってしまふ欠点がある。

間隔尺度のスコアを連続変数と見なして箱ひげ図を作成する方法は簡単であるが、各水準の間隔が等しいことが妥当であるかどうかを検討する必要があるだけでなく、最頻値の情報が失われてしまう。

「SAS Plotter」の MultiHistogram は Madison らの報告例 (Madison R. Wierenga, 2020) を参考に、複数のヒストグラムを簡単に配置できるように設計されている。マルチセルレイアウトを利用するよりも少ない面積で簡単にヒストグラムを配置できる。また 2 群比較 (例えば被験薬と対照薬) を前提に設計されていることも特徴である。執筆時点では同種の作図が実施できるパッケージは python および R には用意されておらず、SAS 環境で先行して実装に成功した。

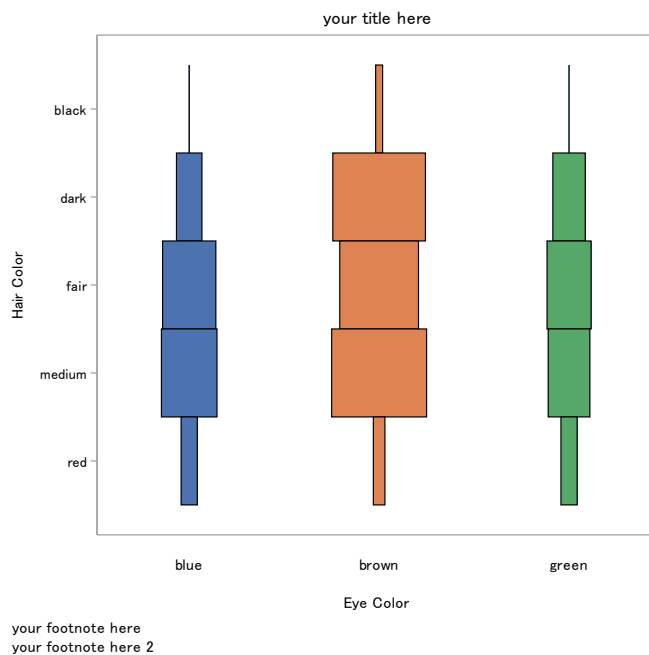


図 9 MultiHistogram の出力例

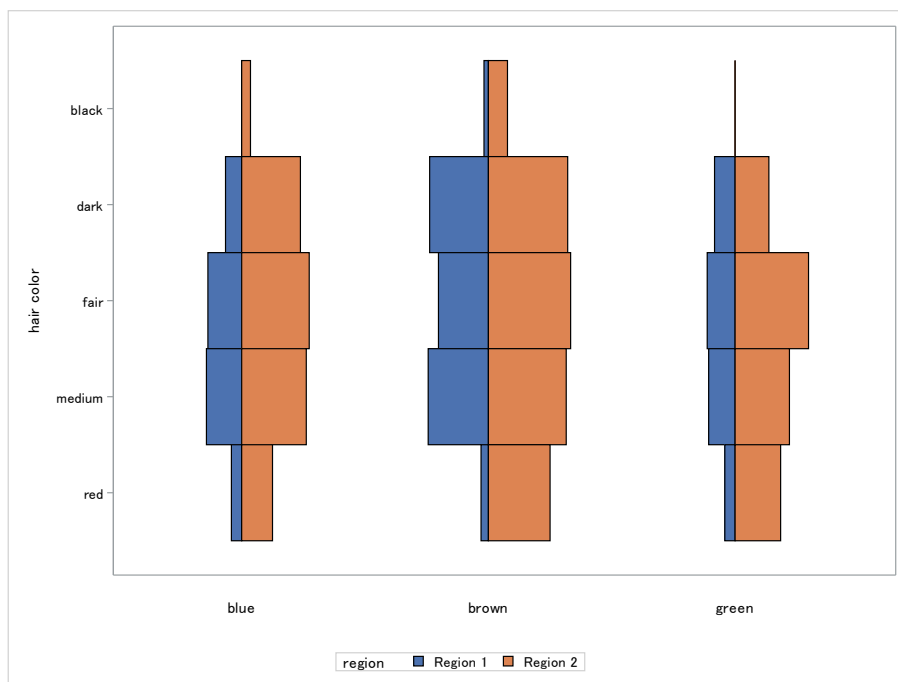


図 10 MultiHistogram の出力例（二値変数を有効にした場合、スプリットモード）

本マクロでは特に設定しない限り通常モードで作図する。これはヒストグラムのバーが応答変数に応じて左右または上下に伸びていく。一方スプリットモードを設定すると、2 値変数に基づいて **Mirrored histogram** と同様にヒストグラムを分割して表示できる。サブグループ別の 2 群ヒストグラムを表示したい場合、スプリットモードを使用するとより面積を節約できる。

MultiHistogram は応答変数軸が存在しないため、グラフから値を読み取ることが難しい。本マクロでは応答変数の値（頻度や割合など）をグラフ上に表示する機能を有する。この機能を利用すれば表の追加作成をしなくてもグラフから値を読み取れる。

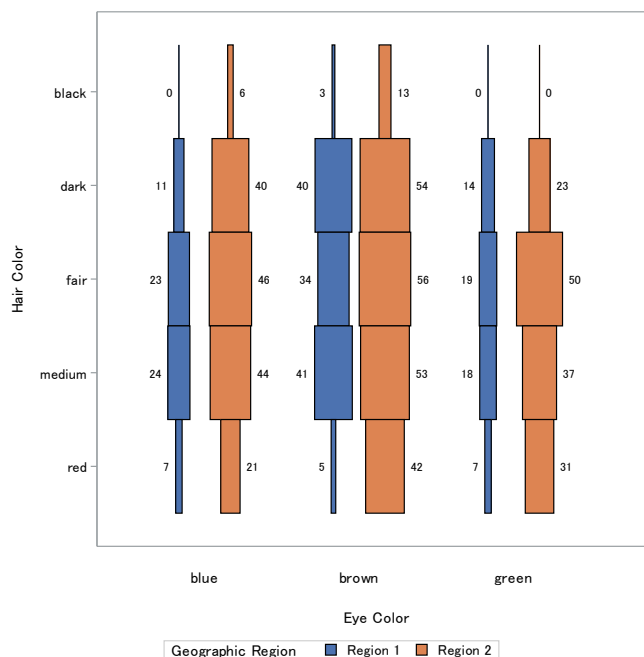


図 11 MultiHistogram の出力例（応答変数の同時表示）

3.各種ドキュメント

「SAS Plotter」のヘルプファイルは日本語版とグローバル版が用意されている。参考文献、作図例および詳細なパラメータの説明は以下のヘルプファイルを参照されたい。

グローバル版

https://superman-jp.github.io/SAS_Plotter/

日本語版

<https://picolabs.jp/sas-plotter-summary/>

日本語版については過去の紹介用スライドも公開されている。

4.謝辞

「SAS PACKAGES」への登録およびコードレビューを引き受けていただきました Bartosz Jablonski 氏、SAS community にて意見とアイデアをくださった Robert Allison 氏と海外 SAS ユーザー、および X(旧 twitter)にてネタ提供いただきました日本ユーザーの方々に感謝申し上げます。特に Bartosz 氏が開発した「SAS PACKAGES」により「SAS Plotter」の利便性と安全性は飛躍的に向上いたしました。

5.参考文献

- Bartosz Jablonski. (2023). Share your code with SAS Packages - a Hands-on-Workshop. *Western Users of SAS Software conference*, 208.
- Chapel Hill, S. R. (2015). Getting sankey with bar charts. *PharmaSUG2015* (p. DV07). PharmaSUG.
- Jeffrey, M. (2024). Combining Functions and the POLYGON Plot to Create Unavailable Graphs Including Sankey and Sunburst Charts. *PharmaSUG2024* (p. DV155). PharmaSUG.
- Madison R. Wierenga, C. R. (2020). Older us adults like sweetened colas, but not other chemesthetic beverages. *Journal of texture studies*, 722–732.
- Matange, S. (2015, 3 21). *Sankey diagrams*. Retrieved from Graphically speaking: <https://blogs.sas.com/content/graphicallyspeaking/2015/03/21/sankey-diagrams/>
- Micah Allen, D. P. (2021). Raincloud plots: a multi-platform tool for robust data visualization, version 2. *Wellcome open research*, 4, 63.
- Nicolle M. Gatto, S. V. (2022). Visualizations throughout pharmacoepidemiology study planning, implementation, and reporting. *Pharmacoepidemiology and Drug Safety*, 31(11), 1131-1217.

デザイン行列

○中川 雄貴

(イーピーエス株式会社)

Design Matrix

Yuki Nakagawa

EPS Corporation

要旨

デザイン行列は統計モデルで用いられる行列であり、定量的な説明変数の観測値やカテゴリカル変数の水準を表す値（ダミー変数）から構成される。デザイン行列を用いることで、重回帰分析や分散分析、共分散分析などの様々な統計モデルを類似した形で表現できる。このような基本的な統計モデルの理解だけでなく、SASなどのプログラムによる解析を理解する上でも非常に重要なものである。そのため、デザイン行列を早い段階で学んでおくことが良いと考えるが、デザイン行列を取り上げている初学者向けの書籍は現状多くはない。そこで本発表では、いくつかの例を用いて、デザイン行列を構成する際の考え方及び注意点に加えて、SASのプログラムや解析結果との対応関係についても紹介する。さらに、SASではデータから一般線形モデルのデザイン行列を構築できるGLMMODプロシジャが実装されている。自身のデザイン行列の学習にも役立つ経験から、GLMMODプロシジャの使用方法についても紹介する。

キーワード：デザイン行列、GLMMODプロシジャ

1. デザイン行列とは

デザイン行列（計画行列）^[1,2,3,4]は統計モデルで用いられる行列であり、重回帰分析や分散分析、共分散分析などの様々な統計モデルをシンプルな形で表現でき、基本的な統計モデルを理解する上で非常に重要なものである。デザイン行列は定量的な説明変数の観測値やカテゴリカル変数の水準を表す値（ダミー変数）、つまり既知の値から構成される。なお、ダミー変数が0と1だけからなる場合、指示変数と呼ばれる。p個の説明変数とN個の観測値がある時、目的変数 \mathbf{y} は $N \times 1$ のベクトル、 $\boldsymbol{\beta}$ は $p \times 1$ のパラメータのベクトル、デザイン行列 \mathbf{X} は $N \times p$ の行列となる。N × 1の誤差ベクトルを $\boldsymbol{\varepsilon}$ とすると、統計モデルは以下のように表現できる。

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1.1)$$

式(1.1)の $\mathbf{X}\boldsymbol{\beta}$ はモデルの線形成分である。なお、本稿では詳細は割愛するが、デザイン行列 \mathbf{X} がフルランクの場合、正規方程式を解くことにより、以下の式(1.2)からパラメータの推定値を求めることができる。ここで、 \mathbf{X}^t は行列 \mathbf{X} の転置行列、 $(\mathbf{X}^t\mathbf{X})^{-1}$ は行列 $(\mathbf{X}^t\mathbf{X})$ の逆行列を示す。

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \quad (1.2)$$

2. 重回帰モデルによる例

説明変数が連続量のみの場合、つまり重回帰モデルを例に取り上げる^[3]。目的変数が体重(Weight)、説明変数が身長(Height)、収縮期血圧(sBP)及び Hb1Ac からなるデータセット Data01 を例に考える。

データセット Data01

ID	Weight (kg)	Height (cm)	sBP (mmHg)	HbA1c (%)
1	60	165	120	5.0
2	75	170	130	6.0
3	80	175	135	5.5
4	95	175	150	6.5
5	85	180	145	5.5

各説明変数のパラメータの推定値を得るために、以下の式(2.1)の重回帰モデルを考える。 β_0 は切片のパラメータ、それ以外の β_x は対応する変数のパラメータとする。

$$\text{Weight} = \beta_0 + \beta_{\text{Height}} * \text{Height} + \beta_{\text{sBP}} * \text{sBP} + \beta_{\text{HbA1c}} * \text{HbA1c} + \varepsilon \quad (2.1)$$

まずは REG プロシジャにより、重回帰分析を実施する。なお、デフォルトで切片がモデルに含まれる^[5]点に注意が必要である。

```
proc reg data= Data01;
  model Weight = Height sBP HbA1c;
run;
quit;
```

パラメータの推定					
変数	自由度	パラメータ 推定値	標準誤差	t 値	Pr > t
Intercept	1	-125.20000	87.22362	-1.44	0.3874
Height	1	0.44000	0.69742	0.63	0.6417
sBP	1	0.64000	0.43635	1.47	0.3810
HbA1c	1	7.20000	4.91528	1.46	0.3813

このように、モデルがどのように構成されているかを意識しなくても、SAS では model ステートメントに変数を列記するだけで推定値を得ることができる。式(2.1)を基に、行列を用いて表現すると以下ようになる。

$$\begin{bmatrix} 60 \\ 75 \\ 80 \\ 95 \\ 85 \end{bmatrix} = \begin{bmatrix} 1 & 165 & 120 & 5.0 \\ 1 & 170 & 130 & 6.0 \\ 1 & 175 & 135 & 5.5 \\ 1 & 175 & 150 & 6.5 \\ 1 & 180 & 145 & 5.5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_{\text{Height}} \\ \beta_{\text{sBP}} \\ \beta_{\text{HbA1c}} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix} \quad (2.2)$$

デザイン行列は式(2.2)の網掛け部分であり、1列目が切片、2列目が身長、3列目が収縮期血圧、そして4列目が Hb1Ac に対応する。式(2.2)を行列を用いずに表現すると以下ようになる。

$$\begin{cases} 60 = \beta_0 + 165 \times \beta_{\text{Height}} + 120 \times \beta_{\text{sBP}} + 5.0 \times \beta_{\text{HbA1c}} + \varepsilon_1 \\ 75 = \beta_0 + 170 \times \beta_{\text{Height}} + 130 \times \beta_{\text{sBP}} + 6.0 \times \beta_{\text{HbA1c}} + \varepsilon_2 \\ 80 = \beta_0 + 175 \times \beta_{\text{Height}} + 135 \times \beta_{\text{sBP}} + 5.5 \times \beta_{\text{HbA1c}} + \varepsilon_3 \\ 95 = \beta_0 + 175 \times \beta_{\text{Height}} + 150 \times \beta_{\text{sBP}} + 6.5 \times \beta_{\text{HbA1c}} + \varepsilon_4 \\ 85 = \beta_0 + 180 \times \beta_{\text{Height}} + 145 \times \beta_{\text{sBP}} + 5.5 \times \beta_{\text{HbA1c}} + \varepsilon_5 \end{cases} \quad (2.3)$$

3. 2群のダミー変数の作成方法

重回帰モデルではデータが連続量のみであったが、カテゴリ変数が含まれる分散分析等の場合、各水準に対応するパラメータを考慮する必要がある。そこで、2群の平均の比較をするモデルを例にデザイン行列の定義の方法を紹介する^[4]。今回は群Aと群Bがあり、各群から2個ずつの測定値が得られたとする。

3.1. 各群の平均値を使用する場合

群Aの平均を β_A 、群Bの平均を β_B とする。この β_A と β_B を使用すると、モデルの線形成分 $X\beta$ は以下のように定義できる。

$$X\beta = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \end{bmatrix} \quad (3.1)$$

3.2. 総平均と総平均からの各群の差を使用する場合

2群の総平均を μ とし、 μ からの各群の平均値の差をそれぞれ α_A 、 α_B とする。この時、モデルの線形成分 $X\beta$ は以下のように定義される。

$$X\beta = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_A \\ \alpha_B \end{bmatrix} \quad (3.2.1)$$

このデザイン行列はランク落ちしており、パラメータを一意に推定できない。目的変数の値を y_{A1} 、 y_{A2} 、 y_{B1} 、 y_{B2} とし、行列を用いずに表現すると以下のようになり、パラメータを一意に推定できない（連立方程式の一意解が得られない）ことが分かる。

$$\begin{cases} y_{A1} = \mu + \alpha_A \\ y_{A2} = \mu + \alpha_A \\ y_{B1} = \mu + \alpha_B \\ y_{B2} = \mu + \alpha_B \end{cases} \quad (3.2.2)$$

3.3. 一方を参照群とし、他方の群の付加的効果を検討する場合

群Aを参照群とし、その平均を μ 、群Bの付加的効果を β とする時、群Bの平均は $\mu + \beta$ 表され、モデルの線形成分 $X\beta$ は以下のように定義できる。

$$X\beta = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} \quad (3.3)$$

このようにある群の効果を参照群（端点）からの差として定義するものを端点制約と呼ぶ。

3.4. 2群を対称に扱う場合

2群を対称に扱い、総平均を μ 、各群の μ からの差を γ とする。この時、群Aの平均は $\mu + \gamma$ 、群Bの平均は $\mu - \gamma$ となり、モデルの線形成分 $X\beta$ は以下のように定義できる。

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \gamma \end{bmatrix} \quad (3.4)$$

各群の平均から総平均を引いたものの和が 0 となるような制約を課しており、零和制約と呼ばれる。

4. 分散分析モデルによる例

3 章では 2 群の場合のデザイン行列の定義方法について記載したが、本章では分散分析モデルを例に 4 群の場合について取り上げる^[2]。今、薬剤 A～薬剤 D の 4 種類を一定期間使用した後のある検査項目の値を 2 サンプルから 1 回ずつ測定し、その変化量を格納したダミーデータセットを Data02 とする。

データセット Data02

ID	薬剤 A	薬剤 B	薬剤 C	薬剤 D
1	40	35	43	45
2	44	35	47	55
平均	42	35	45	50

各測定値を y_{ij} ($i=A, \dots, D, j=1, 2$)、全体平均を μ 、各薬剤のパラメータを β_i 、誤差項を ε_{ij} とし、3 つのモデルについて検討する。

4.1. 全体平均と各薬剤群の付加的な効果を導入したモデル

全体平均 μ と各薬剤群の付加的な効果 β_i を導入したモデルをデザイン行列により $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ の形で表現した際のモデルの線形成分 $\mathbf{X}\boldsymbol{\beta}$ は以下ようになる。

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta_A \\ \beta_B \\ \beta_C \\ \beta_D \end{bmatrix} \quad (4.1)$$

デザイン行列 \mathbf{X} の 1 列目が 2～5 列目の和で表現でき、ランク落ちしている。そのため、3.2 節と同様、パラメータを一意に推定できない。

4.2. 薬剤 D 群を参照群とするモデル

全体平均を使用せず、薬剤 D 群を参照群とし、それ以外の薬剤群の効果に着目するモデルを想定すると、モデルの線形成分 $\mathbf{X}\boldsymbol{\beta}$ とパラメータの推定値は以下ようになる。

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \\ \beta_D \end{bmatrix}, \quad \begin{bmatrix} \widehat{\beta}_A \\ \widehat{\beta}_B \\ \widehat{\beta}_C \\ \widehat{\beta}_D \end{bmatrix} = \begin{bmatrix} -8 \\ -15 \\ -5 \\ 50 \end{bmatrix} \quad (4.2)$$

β_D が薬剤 D 群の平均値 (50) となり、 $\beta_A \sim \beta_C$ が各薬剤群の付加的効果を表している。薬剤 A について注目すると、 $\beta_A + \beta_D$ ($-8 + 50 = 42$) となり、薬剤 A 群の平均値 42 が得られることが分かる。

4.3. 各群の平均を使用するモデル

全体平均を使用せず、それぞれの薬剤群の平均に着目するモデルを想定すると、モデルの線形成分 $\mathbf{X}\boldsymbol{\beta}$ とパラメータの推定値は以下ようになる。

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \\ \beta_D \end{bmatrix}, \quad \begin{bmatrix} \widehat{\beta}_A \\ \widehat{\beta}_B \\ \widehat{\beta}_C \\ \widehat{\beta}_D \end{bmatrix} = \begin{bmatrix} 42 \\ 35 \\ 45 \\ 50 \end{bmatrix} \quad (4.3)$$

得られたパラメータの推定値は各薬剤群の平均値となっていることが分かる。

5. 共分散分析モデルによる例

変数 **Score** が目的変数、薬剤、性別、年齢が説明変数であるデータセット **Data03** のデザイン行列を検討する。

データセット Data03

ID	Score	Drug	Sex	Age
1	134	Active	Male	24
2	128	Active	Male	37
3	116	Active	Male	43
4	129	Active	Female	48
5	150	Active	Female	52
6	134	Placebo	Male	28
7	144	Placebo	Male	39
8	120	Placebo	Male	40
9	132	Placebo	Female	41
10	140	Placebo	Female	58

各説明変数に加えて、薬剤と年齢の交互作用をモデルに含める場合、モデルは以下のように表すことができる。ここで、 β_0 は切片のパラメータ、それ以外の β_x は対応する変数のパラメータとし、薬剤は **Placebo**、性別は **Female** を参照群とする。

$$\mathbf{Score} = \beta_0 + \beta_{Drug} * \mathbf{Drug} + \beta_{Sex} * \mathbf{Sex} + \beta_{Age} * \mathbf{Age} + \beta_{Drug*Age} * (\mathbf{Drug} * \mathbf{Age}) + \varepsilon \quad (5.1)$$

この時、モデルの線形成分 $\mathbf{X}\boldsymbol{\beta}$ とパラメータの推定値は以下ようになる。交互作用のデザイン行列は説明変数のデザイン行列の列同士の積で表現できる。

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 1 & 24 & 1 * 24 \\ 1 & 1 & 1 & 37 & 1 * 37 \\ 1 & 1 & 1 & 43 & 1 * 43 \\ 1 & 1 & 0 & 48 & 1 * 48 \\ 1 & 1 & 0 & 52 & 1 * 52 \\ 1 & 0 & 1 & 28 & 0 * 28 \\ 1 & 0 & 1 & 39 & 0 * 39 \\ 1 & 0 & 1 & 40 & 0 * 40 \\ 1 & 0 & 0 & 41 & 0 * 41 \\ 1 & 0 & 0 & 58 & 0 * 58 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_{Drug} \\ \beta_{Sex} \\ \beta_{Age} \\ \beta_{Drug*Age} \end{bmatrix}, \quad \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_{Drug} \\ \widehat{\beta}_{Sex} \\ \widehat{\beta}_{Age} \\ \widehat{\beta}_{Drug*Age} \end{bmatrix} = \begin{bmatrix} 147.9816 \\ -3.84745 \\ -10.8422 \\ -0.18146 \\ 0.028796 \end{bmatrix} \quad (5.2)$$

6. GLMMOD プロシジャ

GLMMOD プロシジャは一般線形モデルのデザイン行列を構築するプロシジャである。REG プロシジャはカテゴリ変数を使用できないため、データセットの入力段階でダミー変数に置き換える必要がある。そこで、GLMMOD プロシジャを用いて、ダミー変数への置き換えを実装できる。

6.1. GLMMOD プロシジャの使用方法^[6]

基本的な構文は以下の通りである。

```
proc glmmod <options>;  
  BY variables;  
  CLASS variables;  
  FREQ variable;  
  MODEL dependents = independents </ options>;  
  WEIGHT variable;  
run;
```

続いて、主なステートメントとオプションについて記載する。

6.1.1. proc glmmod ステートメント

6.1.1.1. data=オプション

他の SAS プロシジャと同様、入力するデータセットを指定する。

6.1.1.2. outparm=オプション

モデル効果とデザイン行列の列との関係性をまとめたデータセットの出力名を指定する。

6.1.1.3. outdesign=オプション

デザイン行列のデータセットの出力名を指定する。

6.1.2. class ステートメント

モデルに含まれるカテゴリ変数を指定する。

6.1.3. model ステートメント

想定したモデルとなるよう目的変数及び説明変数を指定する。なお、REG プロシジャ等と同様、noint オプションにより、切片をモデルに含めないようにできる。

6.2. GLMMOD プロシジャの使用例

データセット Data03 を用いて、GLMMOD プロシジャの使用例を紹介する。

```
proc glmmod data=Data03 outdesign=Data03_design outparm=Data03_param;  
  class Drug Sex;  
  model Score = Drug Age Sex Drug * Age;  
run;
```

得られたデザイン行列 (outdesign=オプションからの出力) のデータセット Data03_design は以下となる。目的変数の変数名は変わらず Score、デザイン行列の各列の変数名は Col1~Col8 と連番形式で出力される。

Data03_design (outdesign=オプションからの出力)

Data03_param (outparm=オプションからの出力)

デザイン点									
オブザベーション 番号	Score	列の番号							
		1	2	3	4	5	6	7	8
1	134	1	1	0	24	0	1	24	0
2	128	1	1	0	37	0	1	37	0
3	116	1	1	0	43	0	1	43	0
4	129	1	1	0	48	1	0	48	0
5	150	1	1	0	52	1	0	52	0
6	134	1	0	1	28	0	1	0	28
7	144	1	0	1	39	0	1	0	39
8	120	1	0	1	40	0	1	0	40
9	132	1	0	1	41	1	0	0	41
10	140	1	0	1	58	1	0	0	58

パラメータの定義			
列の番号	関連する効果の名前	分類変数の値	
		Drug	Sex
1	Intercept		
2	Drug	"Active"	
3	Drug	"Placebo"	
4	Age		
5	Sex		"Female"
6	Sex		"Male"
7	Age*Drug	"Active"	
8	Age*Drug	"Placebo"	

このデザイン行列のデータセット Data03_design を用いて、REG プロシジャによる重回帰分析を行う

```
proc reg data=Data03_design;
  model Score = Col1 Col2 Col3 Col4 Col5 Col6 Col7 Col8;
run;
quit;
```

Note: モデルはフルランクではありません。パラメータに対する最小2乗解は一意ではありません。統計量によっては誤った解釈を招く恐れがあります。自由度が0またはBと表示されているパラメータについては、その推定値にバイアスが存在します。

Note: 次の変数は他の変数の一次結合であるため、パラメータは0に設定されます。

Col1 =	Intercept
Col3 =	Intercept - Col2
Col6 =	Intercept - Col5
Col8 =	Col4 - Col7

パラメータの推定						
変数	ラベル	自由度	パラメータ 推定値	標準誤差	t 値	Pr > t
Intercept	Intercept	B	137.13939	27.38019	5.01	0.0041
Col1	Intercept	0	0	.	.	.
Col2	Drug "Active"	B	-3.84745	33.90056	-0.11	0.9141
Col3	Drug "Placebo"	0	0	.	.	.
Col4	Age	B	-0.18146	0.71220	-0.25	0.8090
Col5	Sex "Female"	B	10.84224	11.77557	0.92	0.3994
Col6	Sex "Male"	0	0	.	.	.
Col7	Age*Drug "Active"	B	0.02880	0.80472	0.04	0.9728
Col8	Age*Drug "Placebo"	0	0	.	.	.

REG プロシジャによる結果は出力されるが、出力結果に Note が出ていたり、性別 (Sex) では Male が参照群になっていたり、想定したモデルとなっていない。また、出力されたデザイン行列のデータセット Data03_design を再度確認すると、カテゴリ変数の各水準を基準とした変数が作成され、ランク落ちした行列となっていることも分かる。したがって、フルランクかつ想定したモデルにするには、データセット Data03_design から適切に列を選択する必要がある。この際に参照するのが outparm=オプションで出力したデータセット Data03_param である。「列の番号」は Data03_design の「列の番号」と対応しており、「関連する効果の名前」から対応する説明変数を判別できる。さらに、カテゴリカル変数については、「分類変数の値」の列から、どの水準に対応するものかを確認できる。薬剤 (Drug) を例にすると、「列の番号 2」は Placebo 群を参照群とし、Active 群の付加的効果に注目するものであり、デザイン行列では Active 群に 1 が Placebo 群に 0 が割り当てられる。今回のモデルでは、式(5.1)で記載したように、薬剤 (Drug) では Placebo 群を、性別 (Sex) では Female 群を参照群とし、切片を含んだモデルとするため、Data03_design から Col1, Col2, Col4,

Col6, Col7 を選択し、再度 REG プロシジャによる重回帰分析を実行すると以下ようになる。また、REG プロシジャではデフォルトで切片が含まれるが、今回のデザイン行列では切片を考慮し、Col1 として含めており、切片の指定の重複を回避するため、model ステートメントに noint オプションを指定した。なお、Col1 を除き、noint オプションを指定しない場合でも同じ結果が得られる^[5]。

```
proc reg data=Data03_design;
  model Score = Col1 Col2 Col4 Col6 Col7 / noint;
run;
quit;
```

パラメータの推定						
変数	ラベル	自由度	パラメータ 推定値	標準誤差	t 値	Pr > t
Col1	Intercept	1	147.98162	34.45634	4.29	0.0078
Col2	Drug "Active"	1	-3.84745	33.90056	-0.11	0.9141
Col4	Age	1	-0.18146	0.71220	-0.25	0.8090
Col6	Sex "Male"	1	-10.84224	11.77557	-0.92	0.3994
Col7	Age*Drug "Active"	1	0.02880	0.80472	0.04	0.9728

なお、一般化逆行列を使用するなど、他の推定方法を用いることで、デザイン行列がフルランクではない場合でも推定が可能となるが、本稿の主題からは外れるため、取り上げないこととする。

7. デザイン行列と Estimate ステートメント

SAS には GLM プロシジャや GENMOD プロシジャ、LOGISTIC プロシジャなどの線形モデルによる解析を実行できる様々なプロシジャが実装されている。これらのプロシジャには特定のパラメータの水準における最小二乗平均の推定値を求めるための estimate ステートメントがあり、適切に使用するためにはデザイン行列と class ステートメントの理解が重要となる。GENMOD プロシジャをはじめ、いくつかのプロシジャの class ステートメントには param=オプションがあり、ダミー変数の生成方法を選択できる。線形モデルにおける class ステートメントによるダミー変数の生成が吉田 (2014)^[7]から、estimate ステートメントの使用方法が魚住 (2014)^[8]からそれぞれ紹介されており、詳細についてはこれらを参照されたい。

本稿では、GENMOD プロシジャにおける、glm 法 (GLMMOD プロシジャと同じ方法) と effect 法を取り上げ^[9]、デザイン行列と estimate ステートメントの関係について解説する。データセット Data03 を用い、glm 法と effect 法による特定の水準の最小二乗平均の推定値を得るための SAS コードとその結果は以下の通りとなる。なお、今回、例として分かりやすいものとするため、Placebo 群の推定値を推定対象とした。

```
proc genmod data=Data03 order=data;
  class Drug Sex / param=glm;
  model Score = Drug Sex Age Drug*Age;
  estimate "Placebo" intercept 1 Drug 0 1 Sex 0.5 0.5 Age 41 Drug*Age 0 41 / e;
run;
```

```
proc genmod data=Data03 order=data;
  class Drug Sex / param=effect;
  model Score = Drug Sex Age Drug*Age;
  estimate "Placebo" intercept 1 Drug -1 Sex 0 Age 41 Drug*Age -41 / e;
run;
```

glm 法の推定結果

パラメータ		自由度	推定値	標準誤差
Intercept		1	147.9816	24.3643
Drug	"Active"	1	-3.8475	23.9713
Drug	"Placebo"	0	0.0000	0.0000
Sex	"Male"	1	-10.8422	8.3266
Sex	"Female"	0	0.0000	0.0000
Age		1	-0.1815	0.5036
Age*Drug	"Active"	1	0.0288	0.5690
Age*Drug	"Placebo"	0	0.0000	0.0000

effect 法の推定結果

パラメータ		自由度	推定値	標準誤差
Intercept		1	140.6368	18.0586
Drug	"Active"	1	-1.9237	11.9857
Sex	"Male"	1	-5.4211	4.1633
Age		1	-0.1671	0.4201
Age*Drug	"Active"	1	0.0144	0.2845

また、それぞれのオプションで生成されるデザイン行列 X は次のようになる。デザイン行列 X の特定は、「分類変数の水準の情報」と「パラメータの情報」の出力を確認すると容易になる。なお、REG プロシジャで得られたパラメータの推定結果と異なる箇所があるが、これはプロシジャ間の推定方法が異なるためである。これについては本稿の主題から外れるため、詳細な解説は行わない。

glm 法の出力とデザイン行列

分類変数の水準の情報			パラメータの情報			
分類	水準	値	パラメータ	効果	Drug	Sex
Drug	2	"Active" "Placebo"	Prm1	Intercept		
Sex	2	"Male" "Female"	Prm2	Drug	"Active"	
			Prm3	Drug	"Placebo"	
			Prm4	Sex		"Male"
			Prm5	Sex		"Female"
			Prm6	Age		
			Prm7	Age*Drug	"Active"	
			Prm8	Age*Drug	"Placebo"	

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 24 & 24 & 0 \\ 1 & 1 & 0 & 1 & 0 & 37 & 37 & 0 \\ 1 & 1 & 0 & 1 & 0 & 43 & 43 & 0 \\ 1 & 1 & 0 & 0 & 1 & 48 & 48 & 0 \\ 1 & 1 & 0 & 0 & 1 & 52 & 52 & 0 \\ 1 & 0 & 1 & 1 & 0 & 28 & 0 & 28 \\ 1 & 0 & 1 & 1 & 0 & 39 & 0 & 39 \\ 1 & 0 & 1 & 1 & 0 & 40 & 0 & 40 \\ 1 & 0 & 1 & 0 & 1 & 41 & 0 & 41 \\ 1 & 0 & 1 & 0 & 1 & 58 & 0 & 58 \end{bmatrix}$$

effect 法の出力とデザイン行列

分類変数の水準の情報			パラメータの情報			
分類	値	デザイン変数	パラメータ	効果	Drug	Sex
Drug	"Active"	1	Prm1	Intercept		
	"Placebo"	-1	Prm2	Drug	"Active"	
Sex	"Male"	1	Prm3	Sex		"Male"
	"Female"	-1	Prm4	Age		
			Prm5	Age*Drug	"Active"	

$$X = \begin{bmatrix} 1 & 1 & 1 & 24 & 24 \\ 1 & 1 & 1 & 37 & 37 \\ 1 & 1 & 1 & 43 & 43 \\ 1 & 1 & -1 & 48 & 48 \\ 1 & 1 & -1 & 52 & 52 \\ 1 & -1 & 1 & 28 & -28 \\ 1 & -1 & 1 & 39 & -39 \\ 1 & -1 & 1 & 40 & -40 \\ 1 & -1 & -1 & 41 & -41 \\ 1 & -1 & -1 & 58 & -58 \end{bmatrix}$$

estimate ステートメントでは、ラベルを記載した後に、効果と係数を順に指定する。また、興味の対象外のカテゴリカル変数にはダミー変数の平均値を、連続量にはデータの平均値を指定するという点は、ダミー変数の生成法に依存しない共通の記載法である。glm 法では、カテゴリカル変数の各水準に対応する 0 と 1 からなるダミー変数が生成される。「パラメータの情報」の出力から、薬剤および年齢と薬剤の交互作用では Active, Placebo の順に、性別では Male, Female の順に係数を指定すれば良いことが分かる。したがって、estimate ステートメントでは、切片に『1』を、薬剤に『0 1』を、性別に『0.5 0.5』を、年齢には平均値の『41』を、交互作用項には『0 41』を指定することで、Placebo 群の最小二乗平均の推定値が得られる。一方で、effect 法では 2 水準のカテゴリカル変数に対して 1 と -1 からなるダミー変数が生成される。「分類変数の水準の情報」と「パラメータの情報」の出力から、薬剤 (Drug) で Active に着目したい場合は 1 を、Placebo に着目したい場合は -1 を指定すれば良いことが分かる。したがって、estimate ステートメントでは、切片に『1』を、薬剤に『-1』を、性別に『0』(Male の 1 と Female の -1 の平均) を、年齢には『41』を、交互作用項には『-41』を指定すれば良い。また、estimate ステートメントで e オプションを指定することで、各パラメータと estimate ステートメントで指定した係数との対応を確認できるため、想定した指定となっているかを確認できる。

glm 法での estimate ステートメントでの e オプションによる出力（上段）と推定結果（下段）

対比 Placebo の係数								
ラベル	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8
Placebo	1	0	1	0.5	0.5	41	0	41

対比推定の結果										
ラベル	平均推定値	平均		L'Beta 推定値	標準誤差	アルファ	L'Beta		カイ 2 乗	Pr > ChiSq
		信頼限界					信頼限界			
Placebo	135.1205	127.2755	142.9655	135.1205	4.0026	0.05	127.2755	142.9655	1139.6	<.0001

effect 法での estimate ステートメントでの e オプションによる出力（上段）と推定結果（下段）

対比 Placebo の係数					
ラベル	Prm1	Prm2	Prm3	Prm4	Prm5
Placebo	1	-1	0	41	-41

対比推定の結果										
ラベル	平均推定値	平均		L'Beta 推定値	標準誤差	アルファ	L'Beta		カイ 2 乗	Pr > ChiSq
		信頼限界					信頼限界			
Placebo	135.1205	127.2755	142.9655	135.1205	4.0026	0.05	127.2755	142.9655	1139.6	<.0001

このようにデザイン行列とダミー変数の生成法の理解が estimate ステートメントを適切に使用する上で、非常に重要であることが分かる。

8. 最後に

本稿ではデザイン行列の基本的な概要，GLMMOD プロシジャによるデザイン行列の生成法，そしてデザイン行列と estimate ステートメントの関係について解説した。SAS のプロシジャを利用して解析を行うとモデルへの意識が希薄になりがちである。しかしながら，estimate ステートメント等を使用して適切に推定を行うためには，デザイン行列や SAS によるダミー変数の生成方法を含め，プロシジャ内でどのようなモデルが作成されているかを十分に理解しておく必要がある。本稿がデザイン行列の理解の一助となり，統計モデルへの意識付けになれば幸いである。

9. 参考文献

- [1] Annette J. Dobson 著；田中豊 [ほか] 訳（2008），一般化線形モデル入門 原著第 2 版，共立出版
- [2] G.Verbeke, G.Molenberghs 編；松山裕，山口拓洋編訳（2001），医学統計のための線型混合モデル：SAS によるアプローチ，サイエンティスト社
- [3] 東京大学教養学部統計学教室編（1992），自然科学の統計学，東京大学出版会
- [4] 竹村彰通著（2020），現代数理統計学，学術図書出版社
- [5] SAS Institute Inc., SAS/STAT 15.3® User's Guide The REG Procedure
- [6] SAS Institute Inc., SAS/STAT 15.3® User's Guide The GLMMOD Procedure
- [7] 吉田早織, 魚住龍史, 線形モデルにおける CLASS ステートメントの機能, SAS ユーザー総会 論文集 2014
- [8] 魚住龍史, LS-Means 再考 - GLM と PLM によるモデル推定後のプロセス-, SAS ユーザー総会 論文集 2014
- [9] SAS Institute Inc., SAS/STAT 15.3® User's Guide The GENMOD Procedure
- [10] SAS Institute Inc., SAS/STAT 15.3® User's Guide The GLM Procedure

臨床試験データ解析におけるSASとRの挙動差異

○後藤 駿輔¹、松永 友貴¹

(¹ノバルティス ファーマ株式会社)

Discrepancy between SAS and R in clinical trial data analysis

Shunsuke Goto¹, Yuki Matsunaga¹

¹Novartis Pharma K.K.

要旨

近年製薬業界ではRの利用への関心が高まっている。SASとRには相違点が多く存在しており、SASユーザーがRの利用を開始する上での障壁になると考えられる。本論文ではSASとRの相違点についてまとめた。

キーワード：R、オープンソースソフトウェア

1. 初めに

製薬業界では近年、解析用のプログラミング言語としてRが注目され、その利活用の機会を増やしている。SASとRはいずれも統計解析に特化したプログラミング言語である一方で、言語の仕様や処理内容には多くの相違点が存在している。本論文ではSASとRの相違点についてまとめた。

2. SASとRの違い

2.1. General

2.1.1. 変数の命名方法

SASとRにおける変数の命名方法について、基本的には類似する点が多い。相違点について下記記載する。SASでは_（アンダースコア）で始まるマクロ変数名・データセット名が許容されているが、Rにおいては_で始まる変数名は許容されない。また、SASでは原則大文字と小文字が区別されないが、Rでは大文字と小文字が区別されるため、同じアルファベットでも大文字と小文字が異なっていれば、異なる変数とみなされる。

2.1.2. data型の種類

SASのdata stepで変数の型を指定する際、変数の型は主にNumeric型とCharacter型に大別される。一方、RではNumeric、Character、Date、Logical、Complexなどのデータ型が存在している。代表的

なデータ型を下記に記載する。

表 1 R の代表的なデータ型

データ型	説明
Numeric	数値を表すデータ型。Numeric 型の中でも、整数を表す Integer 型と倍精度浮動小数点型を表す Double 型に分類される。
Character	文字列で構成されるデータ型。
Complex	複素数を表すデータ型。
Logical	TRUE と FALSE で構成されるデータ型。
Factor	順序付けされたカテゴリを表すデータ型
Date	年月日を表すデータ型。基準日時からの経過日数が数値として格納される。
POSIXct/POSIXlt	年月日に加え、時間の情報も含むデータ型。POSIXct では基準時刻からの経過秒数が格納される一方で、POSIXlt では年月日・時刻がそれぞれ数値として格納される。

2.1.3. 日付・日時の基準

R における Date 型及びの基準日は、デフォルトでは 1970 年 1 月 1 日が設定されている。また、POSIXct/POSIXlt 型の基準時刻は、デフォルトではグリニッジ標準時間の 1970 年 1 月 1 日 0 時 0 分 0 秒が設定されている。一方で、SAS 日付値は 1960 年 1 月 1 日から指定した日付までの経過日数を表す値であり、SAS 日時値は 1960 年 1 月 1 日 0 時 0 分 0 秒 0 から指定した日付までの、経過秒数を表す値である。このように、SAS と R では基準とする日付・時点に 10 年の差分がある。SAS のデータセットを R で読み取る場合、また逆に R で作成した dataframe を SAS で読み取る場合には、適切な基準日やタイムゾーンの設定を行う必要がある。{xportr}や{haven}などのパッケージを用いてデータセットの出力・読み取りを行うことで、上記の差分を防ぐことが出来る。

2.1.4. 欠損値・非数値

SAS における欠損値は、数値型変数では、(1 個のピリオド) や特殊欠損値 (.A~.Z, ._)、文字型変数では''(ブランク)や' '(引用符で囲まれたスペース)として表現される。一方 R の Dataframe における欠損値は、Integer や Numeric 型の数値型変数では原則 NA、文字型変数では'' (ブランク) や NA として表現される。なお、厳密には Numeric や Character などの型ごとに NA が定義されている。

また、R には欠損値に加え、NULL、NaN、Inf などの非数値が定義されている。下記の表に非数値の説明を記載する。

表 2 R の代表的な非数値

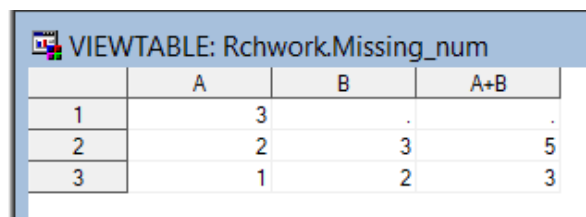
非数値	説明
NULL	データが存在しないことを表す。

NaN	Double 型の一つで、計算できない値を表す。例えば 0 を 0 で除した場合に、NaN が返される。
Inf	Double 型の一つで、無限大を表す。例えば 1 を 0 で除した場合に、Inf が返される。

2.2. 欠測値の扱い

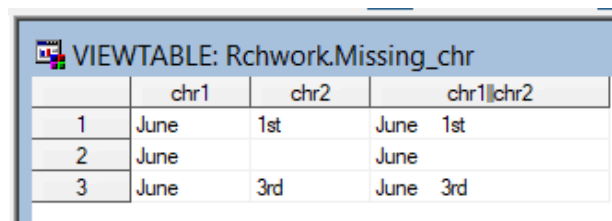
2.2.1. 欠損値を含む変数の処理

SAS では数値型の変数において、欠損値を含んだ計算は欠測値を返す。文字型変数においては、処理内容にも依存するが、単純な文字列の結合などであれば、非欠損値と同様の処理結果となる。



	A	B	A+B
1	3	.	.
2	2	3	5
3	1	2	3

図 1 欠測値を含む数値型変数の処理 (SAS)



	chr1	chr2	chr1 chr2
1	June	1st	June 1st
2	June		June
3	June	3rd	June 3rd

図 2 欠測値を含む文字型変数の処理 (SAS)

R において NA を含んだ処理の結果は、処理内容にも依存するが、データ型を問わず原則 NA を返す。一方、文字列型の変数において NA ではなく” (ブランク) が格納されている場合は、非欠損値と同様の処理結果となる。そのため、R の文字型変数では、欠損値として” (ブランク) が格納されているのか、NA が格納されているのかに応じて処理の結果が異なる。このことから、SAS と同様の処理を想定して R で文字列の結合等の操作を行った場合、意図と異なる計算結果が出力される場合がある。

	A	B	A+B
1	3	NA	NA
2	2	3	5
3	1	2	3

図 3 欠測値を含む数値型変数の処理 (R)

chr1	chr2_blank	join chr1&chr2_blank
1 June	1st	June1st
2 June		June
3 June	3rd	June3rd

chr1	chr2_na	join chr1&chr2_na
1 June	1st	June1st
2 June	NA	NA
3 June	3rd	June3rd

図 4 欠測値を含む文字型変数の処理 (R)

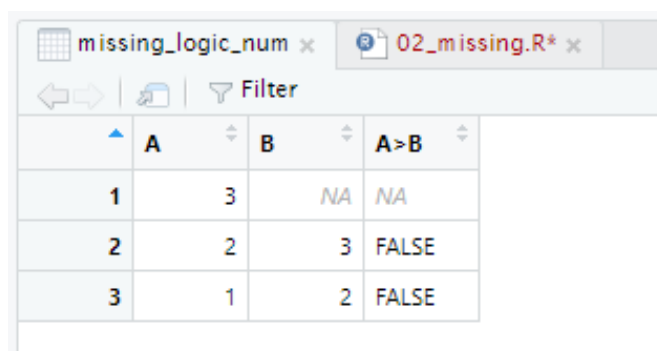
2.2.2. 欠損値を含んだ論理式

SAS の数値型変数の論理式において、欠損値は任意の数よりも小さい値とみなされる。そのため、 < -9999 のような論理式は TRUE を返す。

	A	B	A > B
1	3	.	1
2	2	3	0
3	1	2	0

図 5 欠損値を含んだ論理式 (SAS)

一方で、R の数値型変数の論理式においては、NA を含む場合は原則 NA を返す。このため、SAS と同様の論理式を想定してアルゴリズムを作成すると、差分が生じる原因の一つとなる。

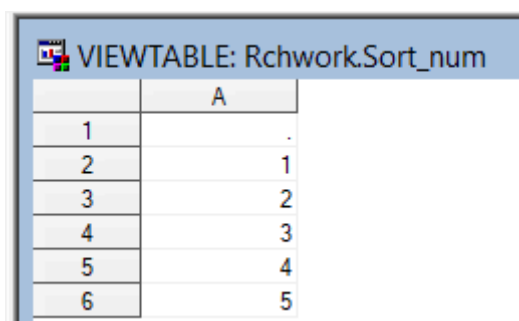


	A	B	A>B
1	3	NA	NA
2	2	3	FALSE
3	1	2	FALSE

図 6 欠損値を含んだ論理式 (R)

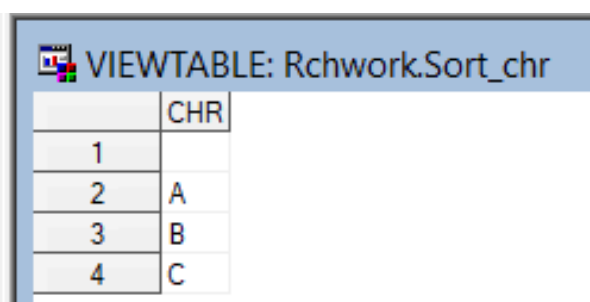
2.2.3. Sort 順

SAS と R では `ort` による欠損値の扱いが異なる。SAS の `sort procedure` においては、数値型変数・文字列型変数のいずれにおいても、欠損値は他のすべての数値・文字値よりも小さいものとみなされる。そのため、昇順の場合、欠損値は最初の行に `sort` される。



	A
1	.
2	1
3	2
4	3
5	4
6	5

図 7 欠損値を含む数値型変数の sort 順 (SAS)



	CHR
1	.
2	A
3	B
4	C

図 8 欠損値を含む文字型変数の sort 順 (SAS)

一方、Rのベクトルをsortする際に用いられるbase::sort()は、デフォルトではNAを削除してsortする。

```
R 4.3.1> num3 <- c(1,3,4,NA,5,2)
R 4.3.1> num3
[1] 1 3 4 NA 5 2
R 4.3.1> sort(num3)
[1] 1 2 3 4 5
```

図 9 欠損値を含む数値型変数の sort 順 (R)

これを防ぐためには、na.last という引数に TRUE または FALSE を指定する必要がある。TRUE を指定した場合、NA は最も後ろに、FALSE を指定した場合、NA は最も前に sort される。

```
R 4.3.1> sort(num3,na.last = TRUE)
[1] 1 2 3 4 5 NA
R 4.3.1> sort(num3,na.last = FALSE)
[1] NA 1 2 3 4 5
```

図 10 欠損値を含む数値型変数の sort 順 (R) na.last = TRUE と na.last = FALSE の場合

また、dataframe を sort するには base::order() や、dplyr::arrange() が用いられる。

base::order() はデフォルトでは sort key となる変数が NA となる行が、最も後ろに来るように sort する。

```
R 4.3.1> df_num[order(df_num$A),]
  A  B
1 1 001
6 2 003
2 3 002
3 4 001
5 5 002
4 NA 003
```

図 11 データフレームにおける欠損値を含む数値型変数の sort 順 (R)

base::order()

base::order() には base::sort() と同様に引数 na.last が存在しており、FALSE を指定することで、SAS と同様に欠損値を最も前に持ってくる事が出来る。

```
R 4.3.1> df_num[order(df_num$A,na.last = FALSE),]
  A  B
4 NA 003
1 1 001
6 2 003
2 3 002
3 4 001
5 5 002
```

図 12 データフレームにおける欠損値を含む数値型変数の sort 順 (R) `na.last = FALSE` の場合

`dplyr::arrange()`もまた `dataframe` の sort 時に良く用いられる関数だが、こちらは `na.last` のように NA の sort 順を制御するための引数はなく、NA は最も後ろに sort される。

```
R 4.3.1> dplyr::arrange(df_num,A)
  A  B
1 1 001
2 2 003
3 3 002
4 4 001
5 5 002
6 NA 003
```

図 13 データフレームにおける欠損値を含む数値型変数の sort 順 (R) `dplyr::arrange()`

2.3. 数値計算の処理の違い

2.3.1. Round

SAS と R では標準で実装されている Round 関数の処理内容が異なる。SAS の ROUND 関数では、「第 1 引数を第 2 引数の最も近い倍数(第 2 引数が省略された場合は最も近い整数)」に丸める。また、この時「第 1 引数が第 2 引数の最も近い 2 つの倍数のほぼ中間にある場合」には、絶対値が大きい方の倍数を返す仕様となっており、四捨五入が行われる。

一方 R の組み込み関数の `base::round()`では、最初の引数の値を指定された小数点以下の桁数に丸める。この `round()`は IEC 60559 に準拠しており、5 を丸める場合には偶数への丸めが行われる。SAS の ROUND 関数と同様の四捨五入を R で行う方法の一つとして、`{janitor}`の `round_half_up()`を利用することが挙げられる。

表 3 SAS の ROUND 関数と R の `base::round()`の比較

Original number	ROUND in SAS	round() in R	janitor::round_half_up() in R
-3	-3	-3	-3
-2.5	-3	-2	-3
-2	-2	-2	-2
-1.5	-2	-2	-2
-1	-1	-1	-1
-0.5	-1	0	-1
0	0	0	0
0.5	1	0	1
1	1	1	1
1.5	2	2	2
2	2	2	2

2.5	3	2	3
3	3	3	3

2.3.2. Percentile

SAS と R ではデフォルトの percentile の計算方法が異なる。例えば SAS の mean procedure や univariate procedure では、percentile の計算方法を指定するための引数 QNTLDEF が存在しており、全部で 5 つの計算方法が実装されている。デフォルトでは QNTLDEF=5 が指定されている。デフォルトの計算の場合、”指定した percentile に最も近い要素（複数ある場合は偶数番目の要素）”が返される。

```

data ds_quan;
  do num = 1 to 10;
    output;
  end;
run;

proc means data = ds_quan q1 q3;
  var num;
run;

```

The SAS System	
The MEANS Procedure	
Analysis Variable : NUM	
Lower Quartile	Upper Quartile
3.0000000	8.0000000

図 14 proc means における第一四分位点及び第三四分位点

R の stats::quantile() では、引数 type の値を変えることで、percentile の計算方法を変更することが可能である。stats::quantile() では全部で 9 つの percentile の計算方法が実装されている。それぞれの計算方法の詳細は Hyndman and Fan (1986) で説明されている。SAS の mean procedure や univariate procedure において QNTLDEF=5 を指定した場合の計算方法は、R の stats::quantile() において引数 type において type = 2 を指定した場合と同様の計算方法である。一方で、stats::quantile() ではデフォルトでは type = 7 が指定されており、SAS における percentile の計算方法とデフォルトでは異なる計算方法を使用している。

```

R 4.3.1> vec_quan <- seq(1,10)
R 4.3.1> vec_quan
[1] 1 2 3 4 5 6 7 8 9 10
R 4.3.1> stats::quantile(quan,c(0.25,0.75))
25% 75%
3.25 7.75

```

図 15 stats::quantile()における第一四分位点及び第三四分位点

表 4 stats::quantile()と mean/univariate procedure の計算方法の対応表

R stats::quantile()	SAS mean/univariate procedure
type = 1	QNTLDEF = 3
type = 2	QNTLDEF = 5
type = 3	QNTLDEF = 2
type = 4	QNTLDEF = 1
type = 6	QNTLDEF = 4

2.4. dataset/dataframe の属性

2.4.1. length

SAS で主に扱われる sas7bdat file や xpt file には変数に length の概念が存在しており、変数の最大長が設定される。一方、R の data frame には変数内における length が存在しないため、各変数は最大長による制限を受けない。data frame 自体には length は存在しないものの、{xportr}などの package を用いることで、xpt file 出力時に length を設定することが可能である。

2.4.2. label

SAS では各変数に label を付与することが可能である。R においても、{labelled}などの package を用いることで、dataframe の各列に対して label を付与することが可能である。また、dataframe から xpt file への出力時に label を付与する場合には、{xportr}や{haven}などの package を用いることで、label が付与された xpt file を出力することが可能である。

2.4.3. format

SAS では各変数について format と informat が設定されている。R の dataframe では、各列について SAS のように format といった属性が設定されているわけではない。しかしながら、{fmtr}などの package や base::format()を用いることで、特定の format に従うように各列の値を変換することが可能である。

3. まとめ

本論文では SAS と R の処理内容の差分とその解決方法についてまとめた。処理内容の原因としては、プログラミング言語としての仕様に違いや、処理・計算に用いられるアルゴリズムの違い、関数のデフォルトのオプションの違いがあげられる。公式のドキュメントなどを参照し、SAS と R の違いを把握することで、R への理解を深めることにつながると考えられる。

4. 参考文献

1. [SAS Help Center: Types of Data](#)
2. [Unix versus SAS Time – pharmaverse blog](#)
3. [ROUND 関数 \(sas.com\)](#)
4. [Round function - RDocumentation](#)
5. [Sample quantiles: A comparison of 9 definitions - The DO Loop \(sas.com\)](#)
6. [Sample Quantiles in Statistical Packages \(amherst.edu\)](#)

生成AIを用いたデータセットのCDISCレビュー方法の模索

○大山暁史

(イーピーエス株式会社)

CDISC Review with Generative AI

Akifumi Oyama

EPS Corporation

要旨

臨床試験の申請電子データ対応では、規制当局から CDISC に則った SAS データセット(SDTM/ADaM)の作成を求められているため、承認申請前の CDISC レビューは必須である。また、CRF の入力データ不備などを発見・修正できる可能性もあるため、開発段階での CDISC レビューも試験の質を向上させるために有用である。

しかし、CDISC レビューするためには高度な知識や経験が求められること、バリデーションツールが高額であること、日本語の公開資料が少なく CDISC の学習が容易ではないことから、CDISC レビューのハードルは高い状況であると考えられる。

近年、生成 AI による SAS プログラム生成など、臨床試験関連業務への応用が検討され始めているが[1]、生成 AI による CDISC レビューに関する調査は未だほとんど報告されていない状況である。本論文では主に DM ドメインに対して生成 AI を活用した CDISC レビュー方法を模索した結果について報告する。

キーワード : Generative AI, CDISC, Clinical Trial

注意事項

- ・本調査は 2024 年 7 月から 8 月の期間で実施した。
- ・今回は全て無償で利用できる範囲に絞った調査である。
- ・生成 AI の応答は必ずしも再現性があるわけではないため、本論文と同様の情報を入力しても結果が同じとは限らないことにご留意いただきたい。(ぜひ SAS ユーザーの皆様にもご自身で生成 AI を試していただきたい)
- ・本論文で紹介する生成 AI の出力は体裁調整のために著者が一部編集した。

調査方法

以下の流れで調査を実施した。

①生成 AI の選定・CDISC 知識確認

文献調査等により SAS データセットを CDISC レビューすることができる可能性の高い生成 AI を絞り込み、絞り込んだ生成 AI を対象に CDISC の情報を参照可能であるか確認した。

下記の質問項目に対する回答がレビュワーとして問題ない程度の精度であることを確認できた場合には、生成 AI がその項目の情報を参照可能であると判断した。

表 1 生成 AI への質問とその意図

質問番号	質問内容	質問の意図
Q1	CDISC とは何ですか。	CDISC の情報にアクセス可能か確認する。
Q2	CDISC 標準を導入するメリットは何ですか。	CDISC 化の意図となる情報にアクセス可能か確認する。
Q3	SDTM IG 3.3 において、DM ドメインの必須変数を教えてください。	SDTM IG v3.3[2]の内容にアクセス可能か確認する。
Q4	PMDA のバリデーションルール バージョン 5.0 において、RULE ID SD0002 はどのようなルールですか。	公知のバリデーションルール [3]の情報にアクセス可能か確認する。

②アップロードファイルのレビュー

CDISC に準拠した被験者 20 名分の DM ドメインのサンプルを様々なファイル形式で用意し、各生成 AI のアップロード可能なファイル形式とレビュー時の挙動を確認する。また、確認内容を基に効果的な使用方法を検討する。

③SAS 形式のレビューツール作成

各生成 AI で CDISC レビュー用の SAS プログラムの作成を試み、出力した SAS プログラムを評価する。

④SAS コード内での生成 AI API の利用

生成 AI の API を SAS コード内に組み込み、挙動を確認する。

調査結果・考察

① 生成 AI の選定・CDISC 知識確認

本調査で使用する生成 AI を選定するにあたり、多くの SAS ユーザーが手軽に使用できるように以下の基準を設定した。

- ・無料で使用できること
- ・日本語で入力・出力できること
- ・ファイルを読み込めること

上記基準や統計解析・申請電子データ関連での生成 AI を使用検討した文献 [4]を参照し、本調査では以下の生成 AI を選定した。

- ・ Chat-GPT 4o (Open AI 社) <https://openai.com/index/hello-gpt-4o/>
- ・ Gemini 1.5 Pro (Google 社) <https://gemini.google.com/app?hl=ja>
- ・ Copilot (Microsoft 社) <https://www.microsoft.com/ja-jp/microsoft-copilot>
- ・ Claude 3.5 Sonnet (Anthropic 社) <https://claude.ai/login?returnTo=%2F%3F>

各生成 AI に CDISC に関する質問を行った際の回答の概要を以下に示す。

Chat-GPT

Q1, Q2 の回答は問題なかった。

Q3 は必須変数を全て挙げたが、Exp 変数も挙げた。変数の説明やラベルのニュアンスが若干異なるものがあった。ARM, ACTARM, RFSTDTC, RFENDTC は重複して出力した。また、COUNTRY を「被験者が所属する国」を格納する変数であると、若干定義と異なる回答をした。

Q4 は尤もらしい回答をしたが、内容は全くの誤りであった。

表 2 Chat-GPT への質問と回答の概要

質問番号	質問内容	回答の概要
Q1	CDISC とは何ですか。	臨床試験データの標準化を推進する非営利団体。主に医薬品開発において、データの収集、管理、提出を効率化するための標準を策定している。SDTM や ADaM、CDASH、SEND などの標準がある。
Q2	CDISC 標準を導入するメリットは何ですか。	データの一貫性と品質の向上、効率的なデータ管理と解析、規制当局への提出が容易、データの互換性と再利用性の向上、データの透明性とトレーサビリティの確保、コスト削減とリソースの節約、規制対応の確保、技術とプロセスの進化への適応
Q3	SDTM IG 3.3 において、DM ドメインの必須変数を	STUDYID, DOMAIN, USUBJID, SUBJID, RFSTDTC, RFENDTC, RFXSTDTC, RFXENDTC, RFICDTC, ARMCD, ARM, ACTARMCD,

	教えてください。	ACTARM, COUNTRY, SITEID, AGE, AGEU, SEX, RACE, ARM, ACTARM, RFSTDTC, RFENDTC
Q4	PMDA のバリデーションルールバージョン 5.0 において、RULE ID SD0002 はどのようなルールですか。	すべての被験者識別子が一意であり、すべてのデータセットで一貫して使用されること

Gemini

Q1 の回答は、CDISC 自体の説明はできていたものの、「2018 年 4 月より厚生労働省が CDISC 準拠の臨床データの提出を推奨する」旨の根拠を著者の方で見つけられなかった。(PMDA は 2016/10/1 から受け入れ開始し、2020/3/31 に経過措置を終了している。)

Q2 の回答は問題なく、デメリットまで回答した。

Q3 では、必須変数が複数不足していることに加え、IG に存在しない変数を列挙した。SDTM IG に存在するはずのない、変数名が 8 文字以上の変数も出力していた。

Q4 については適切な回答を確認できた。

表 3 Gemini への質問と回答の概要

質問番号	質問内容	回答の概要
Q1	CDISC とは何ですか。	医療研究データとメタデータの取得、交換、申請、保存をサポートする標準の開発を推進する国際的な非営利団体。 厚生労働省も、CDISC 標準の採用を推進しており、2018 年 4 月からは、医薬品医療機器等法に基づく製造販売承認申請において、CDISC 準拠の臨床試験データの提出を推奨している。
Q2	CDISC 標準を導入するメリットは何ですか。	メリットは臨床試験の効率化、データ品質の向上、コスト削減、規制当局への対応、データの利活用、グローバルな展開 デメリットは導入コスト、人材育成、運用コスト、複雑性(習得に時間がかかる)
Q3	SDTM IG 3.3 において、DM ドメインの必須変数を教えてください。	STUDYID, DOMAIN, SDTCODE, SDTCDISCOVER, DTM, USUBJID, VISITNUM, VISITREF, FLAG, STUDYDAY, DYEAR, DMONTH, DDAY, DATECREATED, DATELASTMOD
Q4	PMDA のバリデーションルールバージョン 5.0 において、RULE ID SD0002 はどのようなルールですか。	必須項目に NULL 値 (空値) を格納してはならない。

Copilot

Q1, Q2 の回答は問題なかった。

Q3 は必須変数が不足しているだけでなく、Exp 変数や Perm 変数が含まれていた。また、SDTMIG に存在しない変数(RFTENDTC)を出力した。

Q4 では質問に対する回答が得られなかった。

表 4 Copilot への質問と回答の概要

質問番号	質問内容	回答の概要
Q1	CDISC とは何ですか。	臨床研究データとメタデータの獲得、交換、申請、および保存をサポートする標準の開発を行う非営利団体で SDTM や ADaM など様々なデータ標準を提供している。
Q2	CDISC 標準を導入するメリットは何ですか。	データの一貫性と信頼性、効率的なデータ管理、規制要件への適合、データ品質の向上
Q3	SDTM IG 3.3 において、DM ドメインの必須変数を教えてください。	STUDYID, DOMAIN, USUBJID, SUBJID, RFSTDTC, RFTENDTC, RFENDTC, DMDTC, DMDY, SITEID
Q4	PMDA のバリデーションルール バージョン 5.0 において、RULE ID SD0002 はどのようなルールですか。	(バリデーションルールについての簡単な説明はあったが、RULE ID SD0002 についての言及はなかった。)

Claude

Q1, Q2 の回答は問題なかった。

Q3 は必須変数を全て挙げられたが、Exp 変数が含まれていた。

Q4 では質問に対する回答が得られなかった。

表 5 Claude への質問と回答の概要

質問番号	質問内容	回答の概要
Q1	CDISC とは何ですか。	臨床試験データの電子的な収集、交換、提出、アーカイブに関する国際的な標準規格を開発・サポートする非営利組織で、主な目的は 1. 臨床研究データの品質と一貫性を向上させる 2. データ交換を効率化する 3. 規制当局への提出プロセスを合理化する 4. 臨床試験の効率と有効性を高める
Q2	CDISC 標準を導入するメリットは何ですか。	データの質と一貫性の向上、効率性の向上、規制対応の簡素化、グローバルな協力の促進、コスト削減、意思決定の迅速化、トレーサビリティの向上、新技術の導入が容易

Q3	SDTM IG 3.3 において、DM ドメインの必須変数を教えてください。	STUDYID, DOMAIN, USUBJID, SUBJID, RFSTDTC, RFENDTC, SITEID, ARM, ARMCD, ACTARM, ACTARMCD, COUNTRY, SEX
Q4	PMDA のバリデーションルールバージョン 5.0 において、RULE ID SD0002 はどのようなルールですか。	(PMDA の公式ウェブサイトやドキュメントを参照するようにと いう旨を出力した。)

各生成 AI の、レビューに必要な情報の参照可否をまとめると以下の通りである。

表 6 生成 AI の CDISC 情報参照可否

質問番号	Chat-GPT	Gemini	Copilot	Claude
Q1	可能	一部可能	可能	可能
Q2	可能	可能	可能	可能
Q3	一部可能	可能とは言えない	一部可能	一部可能
Q4	可能とは言えない	可能	可能とは言えない	可能とは言えない

全ての質問に適切に回答できた生成 AI はなかった。

Chat-GPT, Copilot, Claude は参照する情報が似た傾向がある可能性がある。一方で、Gemini は CDISC や IG を比較的うまく参照できていない可能性があるが、唯一 Q4(PMDA バリデーションルールの確認)を回答できた。バリデーションルールの遵守のみでは CDISC IG をカバーできているとは言えないため、実業務への応用を見据えるとバリデーションルールに規定されている項目以外の観点でもレビューすべきである。しかし、本調査の範囲では SDTM IG を参照でき(Q3)、バリデーションルールを参照できる(Q4)生成 AI が存在しなかった。

② アップロードファイルのレビュー

命令書は深津式プロンプト[5] を参考に質問の意図や前提情報を指定し、回答の範囲を限定した以下のプロンプトを用意した。

<p>#命令書：</p> <p>あなたは医療データ分析会社の CDISC の専門家です。</p> <p>以下の制約条件をもとに、添付するデータセットが CDISC 準拠しているかレビューしてください。</p> <p>また、精度の高い指摘をするために必要な情報があれば教えてください。</p> <p>#制約条件：</p> <ul style="list-style-type: none"> SDTM IG version 3.3 に準じているか確認し、準じていない箇所があれば箇条書きで挙げてください。 どのような観点でレビューしたのかを簡潔に示してください。
--

Chat-GPT

まず、Chat-GPT に SAS7BDAT 形式のサンプルデータセットをアップロードし、プロンプトを入力したところ、Chat-GPT は Python を用いて必須変数が全て存在することとデータ型の適切さを確認した。次に日付フォーマットや RFSTDTC と RFENDTC の整合性の確認を開始したが、無料版の Chat-GPT では一度にこれ以上の

処理ができないとのことで止まってしまった。後続の予定していた処理について Chat-GPT に問い合わせたところ、以下の旨の回答を得た。

手順 1: 日付フォーマットの確認

以下のカラムについて、各セルが YYYY-MM-DD 形式になっているか確認する：

RFSTDTC, RFENDTC, RFXSTDTC, RFXENDTC, RFICDTC, RFPENDTC

手順 2: 値の一貫性の確認

以下のカラムの値が指定されたリスト内に収まっているか確認する：

SEX: ["M", "F"]

RACE: ["WHITE", "BLACK OR AFRICAN AMERICAN", "ASIAN", "AMERICAN INDIAN OR ALASKA NATIVE", "NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER", "OTHER"]

ETHNIC: ["HISPANIC OR LATINO", "NOT HISPANIC OR LATINO", "UNKNOWN", "NOT REPORTED"]

手順 3: 論理的な一貫性の確認

RFSTDTC（開始日）が RFENDTC（終了日）以前になっていることを確認する。

Chat-GPT ではすべての変数を読み込めていない状況であり、レビュー箇所も CDISC レビューとしては十分ではないが、レビューの順序や Python 処理の方向性としては問題ないようであったため、プロンプトの工夫や有料版への切り替え次第で、SAS7BDAT をそのまま分析することができる可能性があると考えます。次に XPT 形式のデータセットファイルをアップロードしたところ、Chat-GPT が XPT 形式を読み込むための処理にリソースが使われてしまったようで、データの詳細の分析まで実施することができなかった。XLSX や CSV はアップロードできたが、DOMAIN や SEX の値が問題ないことと、列ラベルの確認を行うのみであった。

次に、Dataset-JSON (従来の XPT 形式の制約を克服することを目的とした、データ交換のために設計された最新のデータセット形式)のアップロードを試行することとした。(Dataset-JSON については参考文献[6][7]を参照されたい。)

Dataset-JSON に変換したデータセットを生成 AI に読み込ませたところ、以下のレビュー結果を出力した。

1. 必須項目の有無:
 - 必須項目は全て存在しています。
2. データ型の一致:
 - STUDYID: 文字列 (OK)
 - DOMAIN: 文字列 (OK)
 - USUBJID: 文字列 (OK)
 - SUBJID: 文字列 (OK)
 - RFSTDTC: 文字列 (日付) (OK)
 - RFENDTC: 文字列 (日付) (OK)
 - SITEID: 文字列 (OK)
 - BRTHDTC: 文字列 (日付) (OK)
 - SEX: 文字列 (OK)
 - RACE: 文字列 (OK)
 - COUNTRY: 文字列 (OK)
3. 値の妥当性:

- 日付形式がすべて「YYYY-MM-DD」で正しい (OK)
 - SEX の値が「M」または「F」である (OK)
 - COUNTRY が有効な国コードである (OK)
4. 項目の一貫性:
- STUDYID, DOMAIN, SITEID などの項目で一貫性が保たれている (OK)
5. 形式の正確性:
- 日付形式が正しい (OK)
 - 他の識別子も一貫している (OK)

必須項目(必須変数)の定義に一部誤りはあるが、レビューとして概ね問題なかった。なお、Dataset-XML も Dataset-JSON 同様に検証したが、概ね問題なかった。Dataset-JSON 形式では唯一出力に(OK)と記載しており、不適切な変数が存在した場合には、特定しやすいことが想定される。

Gemini

ファイル形式が JPG, JPEG, PNG, WEBP のみしか対応していなかったため、SAS7BDAT 形式だけでなく、SAS データセットの変換でよく用いられる XPT, XLSX, CSV, JSON, XML 形式はアップロードできなかった。データセットファイルを Gemini に読み込ませて、レビューするという方法は現段階では難しいことを確認した。

Copilot

SAS7BDAT 形式はサポートしていなかったが XLSX, CSV, JSON 形式に対応していたため、サンプルデータを変換し、Copilot にアップロードした。

しかし、Copilot 自体はいずれのファイル形式でもレビューを実施せず、レビュー方針についてざっくりと言及するのみであった。Copilot についても、データセットを読み込んでレビューするという方法は現段階では難しいことが示唆された。

Claude

Claude では SAS7BDAT, XPT, XLSX のアップロードはできなかった。CSV はアップロードできたが変数名のみしかレビューすることができなかった。JSON ファイル(Dataset-JSON)を読み込ませたところ、レビュー自体は実施できていたが、以下の通り全体的に誤りの多い指摘を出力した。

以下の観点でチェックを行いました。

- ・必須変数の存在
- ・変数名と定義の一致
- ・データ型の適切さ
- ・コード化された値の使用
- ・日付形式の一貫性

SDTM IG v3.3 に準じていない可能性がある箇所:

1. RFPENDTC (End of Participation Date/Time) が含まれていますが、これは SDTM IG v3.3 では定義されていません。
2. ARMNRS (Reason Arm and/or Actual Arm is Null) 変数が含まれていますが、これも SDTM IG v3.3 では定義されていません。

- 3. ACTARMUD (Description of Unplanned Actual Arm) 変数が含まれていますがこれも SDTM IG v3.3 では定義されていません。
- 4. 日付変数間で形式が一貫していません。
SDTM では通常、ISO 8601 形式を使用します。
- 5. AGEU の値が"YEARS"となっていますが、SDTM 標準のコード化された値では通常"YEARS"ではなく"Y"を使用します。

生成 AI のレビューの可能性についての調査結果を以下にまとめる。

表 7 各生成 AI のレビュー可能なファイル形式

ファイル形式	Chat-GPT	Gemini	Copilot	Claude
SAS7BDAT	可能	不可 (アップロード不可)	不可 (アップロード不可)	不可 (アップロード不可)
XPT	不可 (分析不可)	不可 (アップロード不可)	不可 (アップロード不可)	不可 (アップロード不可)
XLSX	不可とは言えない	不可 (アップロード不可)	不可	不可 (アップロード不可)
CSV	不可とは言えない	不可 (アップロード不可)	不可	不可
JSON	可能	不可 (アップロード不可)	不可	不可とは言えない
XML	可能	不可 (アップロード不可)	不可 (アップロード不可)	不可 (アップロード不可)

生成 AI にアップロードファイルをレビューしてもらうのであれば Chat-GPT 一択であるが、特に具体的な指示がなければ Chat-GPT は変数属性や値のコードについて確認するような挙動を示しており、レビューの観点が大きかった。プロンプトのニュアンス次第でパフォーマンスが上がるかもしれない(後述)が、申請に耐えうるバリデーションルールや IG の内容の網羅的な確認は実施できないようであった。

③ SAS 形式のレビューツール作成

各生成 AI を用いて CDISC レビューのための SAS プログラムの作成を試みた。

1. プロンプト入力での SAS プログラム作成

プログラムの作成能力を評価するため、著者がプロンプトで CDISC レビューの処理を指示した。

```
#命令書：
あなたは医療データ分析会社の SAS プログラマです。
以下の仕様に則って、DM ドメインが CDISC に準じているか確認をするための SAS プログラムを作成してください。
また、SAS プログラム作成において必要な情報があれば教えてください。
#仕様：
RFSTDTTC-BRTHDTC の年数が AGE と一致しない場合は、NGFL="Y"を出力する。
```

各生成 AI が出力した SAS プログラムについて以下に示す。

Chat-GPT

SAS 日付値が yymmdd10.形式で入っていれば、問題ない SAS コードを出力した。

```
data DM_check;
  set DM; /* DMドメインのデータセットを読み込みます */

  /* 日付のフォーマットを変更 */
  RFSTDTC_date = input(RFSTDTC, yymmdd10.);
  BRTHDTC_date = input(BRTHDTC, yymmdd10.);

  /* 年数の計算 */
  RFSTDTC_year = year(RFSTDTC_date);
  BRTHDTC_year = year(BRTHDTC_date);
  calculated_age = RFSTDTC_year - BRTHDTC_year;

  /* 年数が一致しない場合、NGFLを 'Y' に設定 */
  if calculated_age ne AGE then NGFL = "Y";
  else NGFL = "N";

  /* 必要ない変数をドロップ */
  drop RFSTDTC_date BRTHDTC_date RFSTDTC_year BRTHDTC_year calculated_age;
run;
```

Gemini

以下のように、Year 関数を文字型日付値に使用している以外は、問題ない SAS コードを出力した。

```
data check_cdisc;
  set your_dataset_name; /* 実際のデータセット名を指定 */
  age_calc = year(rfstctc) - year(brthdctc);
  ngfl = 'N';
  if age_calc ne age then ngfl = 'Y';
run;
```

必要な情報

- * データセット名: 分析対象のデータセットの正確な名前
- * 変数名: RFSTDTC, BRTHDTC, AGE の変数がデータセット内に存在し、正しいデータ型であることを確認
- * 日付フォーマット: RFSTDTC と BRTHDTC の日付フォーマットが SAS で認識できる形式であることを確認 (例: YYYY-MM-DD)
- * 出力形式: ngfl 変数の出力結果の扱い方 (レポート出力、CSV ファイルへの書き出しなど)

その他

- * 性能向上: データ量が多い場合は、where 文を用いて事前に条件を絞り込むことで処理時間を短縮できます。
- * エラー処理: データに欠損値や不正なデータが含まれている可能性がある場合は、エラー処理を追加することをおすすめします。

Copilot

以下の SAS コードを出力した。Copilot 自身が算出した AGE を日付の引き算の値と比較するという、仕様の意図を汲み取っていない処理であった。

```
/* 年齢の計算 */
data mydata;
  set mydata;
  AGE = year(RFSTDTC) - year(BRTHDTC);/* 年齢を計算 */
run;

/* 条件判定とエラーフラグの設定 */
data mydata;
  set mydata;
  if AGE ne RFSTDTC - BRTHDTC then NGFL = "Y";/* 年齢が一致しない場合、エラーフラグを設定 */
run;
```

Claude

変数 brthdct が brthdt と記載されており、このままでは実行できないコードであったがそれ以外は処理として問題なかった。ちなみに(恥ずかしながら)著者は intck 関数を知らなかったため、Claude の出力で SAS の理解を深めることができた。

```
/* RFSTDTC-BRTHDTCの年数を計算し、AGEと比較する */
data dm_check;
  set dm;
  /* RFSTDTC-BRTHDTCの年数を計算 */
  years = intck('year', input(brthdt, yymmdd10.), input(rfstdtc, yymmdd10.));

  /* 計算した年数とAGEを比較 */
  if years ne age then NGFL = 'Y';
  else NGFL = 'N';

  /* 結果を出力 */
  keep usbjid rfstdtc brthdt age years NGFL;
run;

RFSTDTC、BRTHDT の日付形式が`yymmdd10.`でない場合は、適切な形式に変更してください。
```

SAS プログラム自身がデータやコードを確認して、微調整する必要がある可能性はあるが、Chat-GPT, Gemini, Claude はプロンプトから簡単な CDISC レビューツールを概ね作成できた。

2. CORE を利用した SAS プログラム作成

CDISC の公開されたバリデーションシステムである CORE[8]のバリデーションルールを利用してチェックツールを作成することを試みた。Github[9]からルールファイルである rules.pkl を入手したが、Gemini, Copilot, Claude は pickle 形式をアップロードできなかった。Chat-GPT にはアップロード可能であったため、以下のプロンプトとともに rules.pkl をアップロードした。(なお、CG0666 は AGEU に値が入っていて AGE が null の場合にエラーを挙げるというルールである。)

#命令書：

あなたは医療データ分析会社の SAS プログラマです。

DM ドメインが、添付ファイルの CG0666 のルールに則っているか確認するための SAS プログラムを作成してください。

また、SAS プログラム作成において必要な情報があれば教えてください。

しかし、Chat-GPT が rules.pkl を読み込む途中でエラーが発生し、SAS プログラムを出力することができなかった。現時点では各種生成 AI に CORE のバリデーションルールファイルをそのまま input することは難しいと考えられる。

次に、rules.pkl を YAML に変換したファイルを Chat-GPT にアップロードしてみたところ、「このままでは処理できないが、特定のルールのみ絞った YAML ファイルを提供してもらえれば処理できる」旨の回答があった。そこで、rules.pkl の CG0666 部分だけを抽出し、YAML 化したファイルをアップロードしたところ、以下の通り適切な SAS 形式のチェックツールを出力した。

```
data dm_check;
  set dm;
  /* エラーが発生する条件 */
  if not missing(AGEU) and missing(AGE) then do;
    output;
  end;
run;

/* エラーメッセージを出力 */
proc print data=dm_check;
  var AGE AGEU;
  title "Errors found: AGE is missing when AGEU is provided.";
run;
```

同様に JSON 化したファイルも適切な SAS 形式のチェックツールを出力した。

Copilot や Claude についても YAML や JSON はアップロード可能であるため、Chat-GPT 同様に CG0666 部分だけを抽出したファイルを命令書とともに読み込ませたところ適切な SAS 形式のチェックツールを出力した。

ファイル形式を工夫し、情報量を抑えれば、Chat-GPT, Copilot, Claude で CORE のバリデーションルールを SAS プログラム化することは可能であることが示された。

④ SAS コード内での生成 AI API の利用

API(Application Programming Interface)とはシステム同士を連携させ、機能を拡張することができるインターフェースのことである。(API についての詳細な説明は Chat-GPT にお問い合わせいただければ幸いである。)各生成 AI について、API の利用可否と料金の発生有無を確認したところ、以下のように Gemini のみ無料で利用可能であった。そのため本調査では Gemini の API を利用することとした。

表 8 各種生成 AI の API サービス

	Chat-GPT	Gemini	Copilot	Claude
API 利用可否	可能	可能	不可	可能
料金	有料	無料	-	有料

まず、以下の方法で Gemini の API Key を取得した。[10]

- ① Build with the Gemini API (<https://ai.google.dev>)にアクセスし、Google アカウントでログインする。
- ② API キーの発行ページ (<https://aistudio.google.com/app/u/0/apikey>) にアクセスする。
- ③ プロジェクトの選択画面が表示されるので、新しい Google Cloud Project を作成し、API キーの発行ボタンを押す。

API キーを取得後、以下のようなマクロ(RunGEMINI)を定義し、生成 AI の API を SAS コード内で利用できるようにした[11]。なお、API Key が漏洩しないよう、共有環境の SAS プログラムには API Key を記載しないよう注意されたい。

```

%let api_key=/* ※ここに取得したAPI Keyをクォーテーションなしで張り付ける*/;
%let url = https://generativelanguage.googleapis.com/v1beta/models/gemini-1.5-pro-latest:generateContent;
%let url_query = &url.?key=&api_key.;
%put &url_query;

%macro RunGEMINI ;
/* POSTリクエスト本文 */
filename in temp;
data _null_;
file in;
put;
put "{";
put "contents": [{"parts":[{"text": "&prompt."}]}];
put "}";
run;

/* リクエストをPROC HTTPに入力する */
filename resp "%sysfunc(getoption(WORK))/echo.json";
proc http
method="POST" /*プロンプトの入力をするのでHTTPメソッドにPOSTを指定*/
url= "&url_query." /*生成AIのAPIのパスを指定 */
ct="application/json" /*コンテンツタイプを指定 */
in=in
out=resp;
run;

/* JSONを読み込み、表示可能な形式にする */
libname response JSON fileref=resp;

data outdata ;
set response.content_parts;

```

```

do row=1 to max(1,countw(text,'0A'x));
  outvar=scan(text,row,'0A'x);
  output;
end;
drop text;
run;

proc report data=outdata ;
column outvar;
define outvar / display "" style(column)=[cellwidth=6in fontsize=10pt asis=ON];
run;
%mend;

```

API を利用した SAS 内での CDISC レビューを 2 パターン試みた結果を以下に示す。

1. 必須変数のレビュー

DM ドメインの変数が必須変数であるか、以下コードを用いて確認した[1]。

```

/*プロンプトに含めるため、DMドメインの変数をマクロ変数var_listに格納*/
proc contents data = SDTM.DM
out = DM_VARS(keep = name label) noprint;
run;
proc sql noprint;
select name into: var_list separated by " " from DM_VARS;
quit;
%put &var_list.;

/*プロンプトを指定 */
%let prompt = %str(% "SDTM IG v3.3の必須変数がDMデータセットに存在するか確認してください。DMデータセットの変数は次の通りです: &var_list. %");

%RunGEMINI;

```

しかし Gemini は質問の意図を理解できない旨の回答を行った。そこでプロンプトを以下のような英語で入力してみたところ、(SITEID, SEX, COUNTRY を必須変数と認識せず、EFSTDTC, RFENDTC, RFICDTC, RFPENDTC を必須変数と認識するなど、) 認識している必須変数に過不足があったもののレビュー自体は実施することができた。

```

%let prompt = %str(% "check if all the REQUIRED variables according to SDTM IG v.3.3 are present in my DM dataset. The variables in my DM dataset are as follows: &var_list. %");

```

調査②(アップロードファイルのレビュー)では Gemini はファイルアップロードができなかったにも関わらず、SAS 中での API 利用ではレビューが実施できたため、SAS コード内での生成 AI の利用は意義があると考えられる。また、生成 AI に「適切でない回答を出力した場合には罰則を与える」などと脅すとパフォーマンスが上がる可能性を示唆している先行研究[12]もあり、本調査①(生成 AI の選定・CDISC 知識確認)の結果も合わせて鑑みると、レビュー可否や出力されているレビューコメントが、プロンプトの入力内容に大きく影響を受けていることが推測される。

2. 統制用語のレビュー

参考文献 [13] (pp. 22-26) を参考に、CDISC API key を取得の上、Race の統制用語を work.term_c74457_race.sas7bdat に格納した。次に、上述の「1. 必須変数のレビュー」確認時の &var_list. と同様のロジックで、DM.Race 内の値を &value_list_DM. に、term_c74457_race.sas7bdat 内のコードリストを &value_list_CT. に格納した上で以下プロンプトを Gemini に入力した。

```
%let prompt = %str("%&value_list_DM.の値が&value_list_CT.に存在しない値を出力する場合には指摘してください%");
```

しかし、人間を人種という括りでカテゴリ化するべきでないため回答できない旨を英語で出力した。英訳したプロンプトを入力しても同様の出力であったので、少なくとも現時点では、DM.RACE に格納されている値が統制用語に則っているかは Gemini では確認することは難しそうである。

倫理的に問題があると Gemini に懸念されない単語であればレビューできるか確認するため、DS.DSCAT を対象に同様の処理を実施した。しかし、日英ともにプロンプト内に展開される「PROTOCOL MILESTONE DISPOSITION EVENT DISPOSITION EVENT」などの DSCAT の値の羅列がランダムな文字列と認識され、統制用語との突き合わせを実施することができなかった。

まとめ

生成 AI での CDISC レビュー方法を様々な観点から模索した。無料版生成 AI はスポットであれば CDISC レビューに使える可能性があるというのが現段階の結論である。本調査で判明した、無料版生成 AI ででき得ること、現状では難しいこと・今後の課題、その他の気づいたことを以下に示す。

でき得ること

- ・簡単なデータセットレビュー(SAS7BDAT, Dataset-JSON, Dataset-XML) (Chat-GPT)
- ・プロンプト入力による簡単な CDISC レビューツールの作成 (Chat-GPT, Gemini, Claude)
- ・CORE のバリデーションルールを参照した CDISC レビューツールの作成 (Chat-GPT, Copilot, Claude)
- ・API を利用した SAS 環境内での必須変数の存在確認 (Gemini)

現状では難しいこと・今後の課題

- ・無料でできる範囲だと 1 日に実施できる処理に限られるため、網羅的な検証や複雑な分析が難しい。
- ・生成 AI は対話を重ねることで出力が洗練されていくものであることから、特定のプロンプトを 1 度入力すれば、求めるレビューを実施できるというのは現段階では難しいかもしれない。
- ・(人間と同様に)質問のニュアンスが変わると回答も変わりうるため、プロンプトの検討が必要。
- ・今回、被験者数が少ない(20 名)サンプルデータを用いたが、レビュー性能は変数やレコード数に依存する可能性がある。
- ・今回は単一データセットのみを調査したが、データセット間の整合性や申請書類とデータセットの整合性なども確認したい。

- ・生成 AI の性能を評価するための評価項目を業界として模索すべきと考える。(CDISC レビューの際は感度・特異度・正解率のどれに重きをおくのかなど)
- ・生成 AI を業務利用するには、責任の所在を明確にする必要がある。
- ・生成 AI は情報のソースが最新でない場合がある。IG はリリースされてから実務で使用するまでに数年のギャップがあるので大きな懸念はないかもしれないが、(COVID のガイドラインなど)リリースから仕様までのギャップが小さい情報などを取り扱う際は、生成 AI がどの時点の情報を参照しているのか、特に留意が必要と考えられる。

気づいたこと

- ・生成 AI の出力が正しいかどうかを判断するには CDISC 知識が必須である。そのため、生成 AI の性能が上がっても、SAS プログラマ自身も勉強が必要。
- ・生成 AI が出力したコードに著者が知らなかった SAS 関数が含まれていたため、学びになった。

最後に

生成 AI の発展は著しいため、すぐに情報の鮮度が落ちてしまう可能性があるが、未来の SAS ユーザーが 2024/7-2024/8 時点の情報を参照できるように調査結果を論文という形でまとめた。本論文が SAS ユーザーの議論のきっかけになり、SAS 界限での生成 AI の利用事例の共有がより一層活発になれば幸いである。

参考文献

各 Web サイトの最終アクセスは 2024 年 8 月

- [1] Illia Skliar (2023). Boosting SAS Programming Efficiency with ChatGPT: A Clinical Trials Perspective https://www.lexjansen.com/phuse/2023/cm/PAP_CM04.pdf
- [2] CDISC SDTMIG v3.3 <https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-v3-3>
- [3] 独立行政法人 医薬品医療機器総合機構 バリデーションルールバージョン 5.0 (2024 年 3 月 29 日) <https://www.pmda.go.jp/review-services/drug-reviews/about-reviews/p-drugs/0028.html>
- [4] Songgu Xie et al. (2024). Comprehensive Evaluation of Large Language Models (LLMs) Such as ChatGPT in Biostatistics and Statistical Programming <https://www.lexjansen.com/pharmasug/2024/AP/PharmaSUG-2024-AP-191.pdf>
- [5] Discovery AI (2023). 深津式プロンプト・システムで ChatGPT を最大限に活用する方法 https://note.com/discovery_ai/n/nb42d8fba539d
- [6] Yutaka Morioka, Yuki Nakagawa (2024) Creating Dataset-JSON Using proc JSON and Extended Attribute in SAS https://www.cdisc.org/sites/default/files/2024-06/2024_cdisc-japan_yutaka_morioka.pdf
- [7] 森岡裕 (2024) Dataset-JSON の SAS での実装と SAS 拡張属性に利用について SAS ユーザー総会 2024
- [8] CDISC CORE <https://www.cdisc.org/core>
- [9] GitHub cdisc-rules-engine <https://github.com/cdisc-org/cdisc-rules-engine/tree/main/resources/cache>
- [10] 中津川篤司 (2024). Google Gemini API を使って、Monaca アプリに AI を組み込む <https://press.monaca.io/atsushi/23517>
- [11] Deepanshu Bhalla. How to Integrate Google's Gemini AI Model Into SAS <https://www.listendata.com/2023/12/gemini-sas.html?m=1>
- [12] Sondos Mahmoud Bsharat, Aidar Myrzakhan, Zhiqiang Shen (2024) Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4 <https://arxiv.org/html/2312.16171v2>
- [13] 小澤義人 (2024). SAS and API <https://www.docswell.com/s/6484025/5GXDDJ-2024-03-26-061943>

大規模言語モデルの新章：マルチモーダルAIの進化と展望

小野 潔

コムチュア株式会社

The Next Frontier: How Multimodal AI is Transforming the Large Language Model

Kiyoshi Ono

Digital Innovation Dept., COMTURE Corporation

要旨

本論文では、ChatGPT-4o、Claude 3.0、Gemini、Llama といった主要な大規模言語モデル（LLM）を取り上げ、それぞれの技術的特徴や市場への影響を考察する。AI の進化に伴い、生成 AI エージェントの未実装の理由やセキュリティリスクが明らかとなり、今後の展望が議論される。また、Microsoft や Google が提供するクラウドベースの生成 AI 開発ツールの選択肢と、SAS Viya の生成 AI ツールの特徴や応用可能性についても分析し、AI 技術の将来における課題と可能性を提示する。

キーワード：大規模言語モデル（LLM） マルチモーダル AI 生成 AI エージェント クラウド AI 開発ツール
SAS Viya の生成 AI SAS Viya's Orchestration Layer SAS DATA MAKER
基盤モデル（Foundation Model） ドメインモデル（Domain Model）

1. 序論

本文は昨年 LLM 理論を中心に論じたが、今年には実務面から LLM を論じる。本章では、主要な LLM モデルとその市場への影響について解説する。

1.1 AI 市場を牽引する主要 LLM モデル

大規模言語モデル（LLM）の市場は、数多くのプレイヤーが競争を繰り広げる中で急速に拡大している。その中でも、特に注目されているのが ChatGPT-4o、Claude 3.0、Gemini、そして Llama の 4 つのモデルである。これらは、それぞれ異なる技術的特徴と強みを持ち、異なる開発元から提供されている。たとえば、ChatGPT-4o はエージェント機能の強化に力を入れており、複数のタスクを自律的に処理する能力が注目されている。一方、Claude 3.0 は安全性の確保に焦点を当て、データの誤用を防ぐ仕組みが強化されている。また、Google が開発した Gemini は、そのエコシステムとの深い統合が強みであり、Llama はオープンソースとしての柔軟性があり、カスタマイズ性の高さが評価されている。

これら 4 つのモデルは、それぞれの強みを活かしながら市場の需要に応じており、AI 技術のさらなる発展を牽引している。市場における各モデルの役割を理解することで、AI の進化がもたらす影響や未来の展望を把握することができる。

主要 LLM モデルの特徴と比較

モデル名	開発元	特徴と強み	応用分野	特記事項	月額料金	サービス開始月
ChatGPT-4o	OpenAI	・高速な応答とマルチモーダル処理 ・AIエージェントとしての機能拡充	ビジネス、教育、エンターテインメント	外部アプリとの連携が可能	ChatGPT Plus: 約20ドル	2024年5月
Claude 3.5 Sonnet	Anthropic	・高度な倫理性と安全性 ・誤情報やバイアスを回避	医療、金融、教育	AIの透明性と安全性を重視	約20ドル (Proプラン)	2024年6月
Gemini Pro	Google	・Google検索との統合 ・マルチモーダルデータ処理	ビジネス、検索エンジン、日常生活	高精度の情報提供、動画データ処理も対応	未公開 (Google Cloudと連携可能)	2024年4月
Llama	Meta	・オープンソース ・小規模デバイスでの利用可能	学術研究、非営利プロジェクト	カスタマイズ性が高く、オフラインでも利用可能	無料 (オープンソース)	2023年8月

1.2. AI の進化とマルチモーダルの夜明け

AI 技術の進化は急速に進んでおり、その影響は産業全体にわたって革新的な変化をもたらしている。その中心的な技術である生成 AI は、OpenAI が開発した ChatGPT シリーズによって広く認識され、今では日常生活からビジネスまで幅広い分野で利用されている。特に 2024 年に登場した ChatGPT-4o は、これまでの生成 AI の枠を超えて、さらなる機能向上を果たしている。

その中でも、ChatGPT-4o の大きな特徴として注目されているのがマルチモーダル AI の実装である。マルチモーダル AI とは、テキスト、画像、音声、さらには映像など、異なる種類のデータを統合的に処理し、相互作用させる技術を指す。この技術により、AI は単一の情報ソースに依存せず、複数のモーダル（形式）から得られる情報を組み合わせて、より正確かつ多次元的な洞察を提供することが可能になる。たとえば、マルチモーダル AI は、テキストと画像を同時に処理して状況を理解したり、音声や映像を組み合わせた高度な対話型 AI を構築するなど、これまでの単一モーダル AI にはできなかった複雑なタスクに対応できる。

特に 2024 年に登場した ChatGPT-4o は、AI 市場において重要な位置を占めている。ChatGPT-4o は、前バージョンである GPT-4 の機能を大幅に強化し、高速な応答とマルチモーダル処理の向上が実現された。日常生活では多様なデバイスとの連携やリアルタイムの意思決定を支援するなど、幅広いシナリオでの活用が期待される。

1.3. OpenAI の激動の歴史と現在の問題

OpenAI は 2015 年にイーロン・マスクとサム・アルトマンによって設立され、当初は非営利団体として汎用人工知能（AGI）の実現を目指していた。しかし、2018 年にイーロン・マスクが役員を辞任したことで内部での対立が表面化し、2019 年には営利部門が設立された。Microsoft からの 10 億ドルの出資を受け、技

術開発は加速したが、商業化に伴う内部の緊張も強まった。さらに、2020年には研究部門のアモディ兄妹が退社し、新たなAI企業 Anthropic を設立。これにより、OpenAI 内部の方針転換に対する反発が浮き彫りになった。

近年、OpenAI と Anthropic の LLM（大規模言語モデル）競争は激化しており、その背景には実は Microsoft（OpenAI 社の49%を保有する主株主）と Google の代理戦争があると言われている。Microsoft は OpenAI を支援し、AI 市場での優位性を強化する一方、Google は Anthropic を後押しし、自社の AI 技術を活かした市場拡大を図っている。この二社の競争は、単なる AI 開発企業同士の争いを越えた、大手テクノロジー企業の市場支配をかけた戦略的な動きとも言える。

OpenAI 挑戦の歴史

年号	内容
2015	非営利団体としてスタート サム・アルトマン、イーロン・マスク参加
2018	GPT-1 発表 イーロン・マスク役員辞任
2019	GPT-2 発表 営利部門と非営利部門分社 → アモディ兄妹退社
2020	GPT-3 発表 アンソロピック設立←アモディ兄妹
2022	ChatGPT 発表
2023	3月: GPT-4 発表 5月: アンソロピックにグーグル投資 11月: サム・アルトマン 解任騒動 12月: ニューヨーク・タイムズから批評
2024	2月: イーロン・マスク Open AI 提訴 3月: Claude 3.5 発表 5月: GPT-4o 発表 ササキヤス氏、ライカ氏 脱社 6月: Claude 3.5 Sonnet 発表 8月: アルトマンのWebにて、 ChatGPT Strawberry, Orion の開発示唆 9月: OpenAI 支社の日本社長が年内に新 LLM リリースを発表

2023年にはサム・アルトマン CEO が突然解任される事件が起こり、最終的に復職したが、OpenAI の内部ガバナンスに不安が残った。2023年にはニューヨーク・タイムズから記事の学習データ利用が疑われ、2024年にはイーロン・マスクが OpenAI を提訴し、非営利の理念を損なう商業化に対する批判が高まっている。また、外部からも生成 AI による「ハルシネーション」問題やデータ管理の不透明性が批判されており、技術的な信頼性に課題が残っている。

その一方で、ChatGPT シリーズは着実に技術的進化を遂げており、特に GPT-4 ではマルチモーダル処理が導入された。2024年に登場した ChatGPT-4o は、応答速度と処理能力が大幅に向上し、さらに広範な応用が期待されているが、誤情報の生成やデータ使用に関する課題が解決されていない点が依然として懸念されている。

1.3. 今後の LLM の進化と注目すべきポイント

本文今後、大規模言語モデル（LLM）は、より高度なマルチモーダル AI へと進化していくと予測される。その中で特に注目すべきは次の三点である。

① AIのエージェント化

AIが単なる対話型から進化し、ユーザーの指示に基づいて複数のアプリケーションやタスクを自律的に処理する「エージェント機能」の強化が進む。これにより、AIはより複雑な業務を自動化し、ユーザーの負担を大幅に軽減することが期待される。しかし、同時にセキュリティやプライバシーのリスクも増大するため、それに伴う安全対策が必須となる。

② 動画生成の可能性

今後、LLMは単にテキスト処理にとどまらず、動画生成にも対応する可能性が高い。これにより、映画、広告、教育など多くの分野での新しい応用が期待されるが、技術的な課題やデータ利用の倫理問題も存在する。技術の実現にはまだ時間がかかるが、将来的な可能性として注目される。

③ 学習データの安全性

AIの信頼性を担保するためには、学習に使用するデータソースの安全性が重要となる。特に、偏ったデータや誤情報による学習がAIの判断に悪影響を与えるリスクがある。また、データの著作権問題にも注意が必要であり、合法かつ透明性のあるデータ利用が求められる。

YouTube動画の無断使用については、LLM各社とも明言をさけている。

第2章：ChatGPT-4Oに生成AIエージェントが未実装の理由と展望

2.1 背景

2024年5月、GoogleとOpenAIは生成AI技術に関する新機能を発表した。しかし、「生成AIエージェント」のリリースはまだ行われていない。生成AIエージェントとは、ユーザーに代わって複数のアプリケーションを操作し、複雑なタスクを自動的に処理する機能であるが、技術的な課題やセキュリティリスクのため、実装が遅れている。本章では、生成AIエージェントの機能と未実装の理由を説明する。

2.2 生成AIエージェントの機能と応用例

生成AIエージェントは、単なるチャットボットを超えた高度な機能を持つ。ユーザーの指示に基づき、外部プログラムをAPIやUIを介して操作し、さまざまなアプリケーションを連携させて複雑なタスクを自動的に処理できる。例えば、引越しの際にエージェントが住所変更届を自動作成し、提出手続きを代行することができる。また、公共サービスの移転手続きを一括で処理するなど、ユーザーにとって煩雑な業務を効率的にこなすことが可能である。

さらに、現在の主流であるAIアシスタントは、ユーザーの指示に対して単に応答する「チャットボット」の機能に留まっており、より複雑な作業を自律的に処理するには限界がある。現行のAIアシスタントは、単純な情報検索やスケジュール管理といったタスクに特化しているが、将来的な生成AIエージェントの進化により、これを超えた高度な業務処理や意思決定が可能になると期待される。AIエージェントの本格的な登場により、AIは単なる補助的存在から、実際の業務や生活における実務を担う存在へと進化しようとしている。

2.3 未実装の理由

生成 AI エージェントがまだ実装されていない理由は、技術的な成熟度の不足とセキュリティリスクにある。エージェントはユーザーの権限を代行してシステムを操作するため、不正利用の危険性が大きい。たとえば、悪意ある攻撃者がエージェントを利用して、機密情報を操作したり、システムに侵入したりするリスクがある。また、生成 AI 特有の「ハルシネーション」（誤情報の生成）のリスクも深刻であり、これが誤った行動を引き起こす可能性がある。このような問題が技術的な障害となっている。

2.4 今後の展望

生成 AI エージェントが安全に実装されることで、ユーザーの生活や業務の効率化は劇的に進むことが期待されている。特に、複数のシステムを連携させた高度な自動化が可能となり、ビジネスや日常生活の様々な場面での負担が軽減されるだろう。しかし、その実現には、強力なセキュリティ対策やガイドラインの整備が欠かせない。AI アシスタントから生成 AI エージェントへの進化により、AI が社会に与える影響はさらに大きくなり、より多くの実務や意思決定を担うようになるだろう。

第3章: クラウドで動く生成 AI アプリの開発ツール

3.1 概要

2024 年 5 月から 6 月にかけて、Microsoft と Google はクラウドベースの生成 AI アプリ開発ツールを発表した。これにより、大規模言語モデル（LLM）を活用し、複雑な AI アプリケーションの迅速な開発が可能となっている。

クラウドLLMを利用する生成AIアプリの開発ツール			
	ツール名	位置付け	主な用途
マイクロソフト	Azure AI Studio	プロコードの開発ツール	様々なクラウドLLMを使う生成AIアプリの開発
マイクロソフト	Microsoft Copilot Studio	ローコードの開発ツール	マイクロソフトのAIアシスタント「Copilot」をカスタマイズして、エージェントなどを開発
グーグル	Vertex AI Studio	企業ユーザー用の開発ツール	様々なクラウドLLMを使う生成AIアプリの開発
グーグル	Google AI Studio	個人ユーザー用の開発ツール	グーグルのクラウドLLMを使う生成AIアプリの開発
グーグル	Firebase Genkit	スマートフォンアプリケーション向け開発ツール	クラウドLLMを使うスマートフォン用生成AIアプリの開発
グーグル	Vertex AI Agent Builder	ノーコードの開発ツール	RAG（検索拡張生成）を使う生成AIアプリなどの開発

クラウドLLMを利用する生成AIアプリを開発するためにマイクロソフトとグーグルが提供するツール

出典：日経コンピュータ 2024.06.20

3.2 クラウド AI 開発ツールの展望

Microsoft と Google が提供するクラウド AI 開発ツールは、生成 AI アプリケーションの開発を強力に支援する。開発者は、プロコード、ローコード、ノーコードといったツールをニーズに応じて選択し、最適な環境で効率的に AI アプリを構築できる。クラウド技術と LLM の進化により、今後さらに開発が加速し、柔軟で迅速なアプリケーション開発が期待される。

第 4 章: SAS Viya の生成 AI ツールの特長と活用領域

4.1 はじめに

SAS Viya は、企業のデータ分析と生成 AI（ジェネレーティブ AI）を組み合わせ、業務プロセスの効率化と自動化を実現する強力なプラットフォームである。本章では、SAS Viya が提供する生成 AI ツールの特長と応用領域について述べる。SAS Viya は、OpenAI GPT-4 や Google の Gemini Pro といった大規模言語モデル（LLM）と統合し、企業のニーズに合わせたカスタマイズ可能な AI モデルを提供している。SAS Viya は、基盤モデルとドメインモデルという二層のアプローチを採用し、柔軟かつ強力な AI ソリューションを提供している。

一方、昨年、ゴールドマンサックスは ChatGPT-Turbo 3.5 をリファインチューニング（再学習）し、金融のドメインモデルを開発したが、しばらくして発表された ChatGPT-4.0 の性能がリファインチューニングしたモデルを上回るという事実が明らかになった。この事例は、生成 AI 技術が急速に進化し、リファインチューニングを行うタイミングやそのモデルの寿命が大きく影響を受けることを示している。

4.2 ジェネレーティブ AI ツールの概要

SAS Viya が提供する生成 AI ツールは、AI モデルの開発から運用までのプロセスを効率化する機能を持つ。特にデータのプライバシー保護に配慮した合成データの生成が特徴である。たとえば、SAS Data Maker は、機密データを保持せずにトレーニングデータを作成できるため、データのセキュリティを強化する役割を果たす。また SAS Data Maker は Snow Flake のネイティブアプリケーション稼働も発表された。

SAS ViyaのジェネレーティブAIツールの特長と活用領域

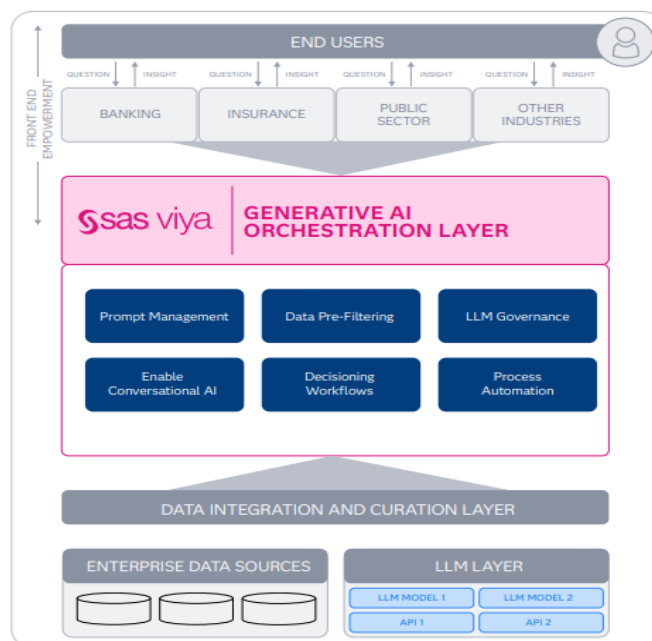
	Purpose	Key Benefit	Target User	Generative AI Mode
① GenAI オーケストレーション	高品質な合成データを生成し、AIモデル訓練に必要なデータを提供。プライバシー保護を確保。	業務効率を高め、複雑なプロセスの自動化を支援。	企業のデータサイエンティスト、業務プロセス担当者	OpenAIのGPT-4やGoogleのGemini Pro 1.5など、外部LLMを使用して業務プロセスに統合。これにより、特定の業務タスクにカスタマイズされたAIモデルを作成し、説明可能で信頼性の高い結果を提供。
② Viya Copilot	ジェネレーティブAIを利用して、コード生成、データクレンジングなどのタスクをサポートし、生産性を向上。	コード作成やデータ操作を効率化し、時間削減。	データサイエンティスト、開発者、ビジネスユーザー。	自然言語処理（NLP）をベースとした生成AIを採用し、コード生成やデータクレンジングの自動化を実現。これにより、ユーザーの作業を迅速化し、タスクを効率化する。
③ SAS Data Maker	高品質な合成データを生成し、	機密情報を守りつつ、データ不足の問題を解消	AIモデル開発者、企業のプライバシー担当者。	合成データ生成のために特殊な生成AIモデルを採用し、テーブル形式の高品質なデータを生成。この合成データはプライバシー保護に配慮し、既存の機密性の高いデータを使わずにAIモデルをトレーニングすることが可能。
④ 顧客エンゲージメント	マーケティング計画、オーディエンス構築など、カスタマージャーニーの最適化をサポート。	マーケターが迅速に顧客にアプローチできるよう支援。	マーケティング担当者、カスタマーエクスペリエンス担当者。	LLMを活用して、自然言語プロンプトに基づくオーディエンス構築やマーケティングコンテンツの生成を行う。また、オーディエンスデータからインサイトを抽出し、マーケターが効果的な意思決定を行えるようにする。

前頁の AI ツールの特長は、特にプライバシー保護を考慮しつつ、生成 AI による合成データの提供を実現している点にある。例えば、SAS Data Maker は、機密情報を保持せずに AI モデルのトレーニングデータを生成でき、データのセキュリティを確保する機能を持っている。

4.3 Generative AI Orchestration Layer

SAS Viya は Orchestration 機能を通じて、外部の生成 AI モデルを既存のビジネスプロセスに統合することを目指している。Orchestration Layer は、基盤モデルとドメインモデルを活用し、LLM（大規模言語モデル）を用いた業務プロセスの自動化を可能にする。このレイヤーは、複数のデータソースからの情報を統合し、AI による洞察提供の中心的な役割を果たすものである。特に、金融、保険、公共部門などの分野において、データの統合およびキュレーション機能を備えており、各ドメインに応じた高度な自動化を実現することができる。

Use SAS Viya to explain, govern and orchestrate LLMs to augment your existing use cases and processes



SAS Viya's Orchestration Layer Layer

名称	和名	役割
● Prompt Management	プロンプト管理	LLMに対する入力（プロンプト）の管理を行い、モデルが適切な応答を生成するよう調整する。
● Data Pre-Filtering	データ事前フィルタリング	モデルに入力するデータを事前に整理・フィルタリングし、生成AIが最適な結果を導き出すための準備する。
● LLM Governance	LLMのガバナンス	モデルの使用に対する管理と監視を行い、セキュリティやコンプライアンスに対する基準を維持する。
● Enable Conversational AI	対話型AIの有効化	ユーザーとの対話を支援するために、LLMを用いたAIチャットや質問応答の機能を提供する。
● Decisioning Workflows	意思決定ワークフロー	LLMが生成した洞察を基に、業務プロセスの意思決定を自動化する。
● Process Automation	プロセス自動化	モデルを活用し、繰り返し作業や定型業務を自動化することで、効率を向上させる。

4.4 活用領域の詳細

SAS Viya の生成 AI ツールは、多様な業界での業務プロセスに柔軟に対応する。基盤モデルが多様なデータを処理し、ドメインモデルがそれぞれの業界に特化したソリューションを提供することで、銀行業務では顧客の行動分析や信用リスクの予測、保険業界ではリスク評価や保険金請求の自動処理が進んでいる。また、公共部門では、大規模なデータ分析を通じた行政サービスの最適化が進められている。

4.5 結論

SAS Viya の生成 AI ツールは、データサイエンティストや業務担当者にとって、精度の高い予測と効率的な意思決定をサポートする信頼性の高いソリューションである。基盤モデルとドメインモデルの組み合わせにより、柔軟性と専門性を両立させ、さまざまな業界での業務プロセスの最適化を実現している。

5. 総括

本論文では、大規模言語モデル (LLM) と生成 AI の進化、そしてその将来の展望について検討しました。主要な LLM モデルである ChatGPT-4o、Claude 3.0、Gemini、Llama は、それぞれの強みを活かして AI 市場の成長に寄与しています。また、生成 AI エージェントの実装には技術的課題やセキュリティリスクがありながらも、将来的な技術革新に期待が持てます。Microsoft や Google のクラウドツールは、生成 AI アプリケーションの開発を効率化し、SAS Viya は業務の自動化とデータ保護に貢献しています。今後、AI 技術はマルチモーダル処理や動画生成、エージェント機能の進化を遂げ、技術的課題と倫理的問題に対応する必要があるでしょう。本論文は、AI の進化とその社会的影響についてまとめました。

6. 参考文献

- Open Ai GPT-4o
<https://platform.openai.com/docs/guides/chat-completions>
<https://platform.openai.com/docs/overview>
<https://openai.com/research/>
- ANTHROPIC
<https://docs.anthropic.com/en/docs/welcome>
- Microsoft
<https://www.stratoscloud.com/blog/microsofts-comprehensive-suite-of-ai-tools/>
- SAS Viya 生成 AI :
https://www.sas.com/en_sg/solutions/ai/generative-ai.html#:~:text=Achieve%20real%20value%20by%20using,and%20govern%20their%20life%20cycle.
- SAS Viya における生成 AI のトレンド
https://www.sas.com/ja_jp/news/press-releases/2024/april/jp-viya-uses-generative-ai.html
- SAS 生成 AI 概要と重要性
https://www.sas.com/ja_jp/insights/analytics/generative-ai.html
- 日経コンピュータ (2024 年 6 月 20 日),
クラウドで動く生成 AI アプリ、Microsoft と Google が合計 6 種類もの開発ツール。
- 日経コンピュータ (2024 年 5 月 17 日),
次の主役「生成 AI エージェント」、Google や OpenAI がまだ出さない怖い理由。
- 日経コンピュータ (2024 年 9 月 3 日),
AWS・MS・Google が生成 AI で開発支援を手厚く、コード生成からインフラ構築まで。
- 中田敦彦, (YouTube), OpenAI 激闘の歴史を知れば未来を予測できる!
https://www.youtube.com/watch?v=dHxFEqsGLsg&t=195s&ab_channel=%E4%B8%AD%E7%94%B0%E6%95%A6E5%BD%A6E3%81%AEYouTube%E5%A4%A7%E5%AD%A6-NAKATAUNIVERSITY.
- 野村総合研究所, (2024 年 2 月 28 日),
AI ナビゲーター 2024 年版: 生成 AI の進化がもたらす次世代ビジネス。

7. マルチモーダル AI のデモンストレーション

7.1. 前書き

本論文では、マルチモーダル AI のデモンストレーションを通じて、異なるデータ形式（テキスト、画像、音声など）を統合的に処理する AI の可能性を示す。マルチモーダル AI は、各データ形式の特徴を組み合わせることで、従来の単一モーダル AI に比べて、より高精度で直感的な意思決定を支援することが可能である。本デモでは、複数のデータソースから得られる情報をリアルタイムで解析し、ビジネスシナリオにどのように応用できるかを簡単な例を示す。

また、SAS Viya 上でも同様の処理が可能であるが、筆者は現時点ではその具体的な手法を詳細には把握していない。2024 年 4 月にラスベガスで開催された“SAS Innovate 2024”では、SAS Viya 上で ChatGPT を稼働させるハンズオンセッションが開催されており、AI と SAS の連携における新たな可能性が示唆された。

7.2. マルチモーダル AI のデモンストレーションの説明

下記を ChatGPT-4o に指定されたプロンプトを入力することで、ChatGPT が採用官の役割を担い、候補者の人事審査を行うことが可能である。ここで重要なのは、ChatGPT は当社のホームページに記載されている情報を初期入力として、不足情報を LLM で補うことで、トレーニングもせずに採用官の判断を行う点である。

さらに、Chrome の拡張機能である【ChatGPT 音声コントロール、Voice Control for ChatGPT】を使用すると、ChatGPT と音声で対話しながら、10 点満点で採用審査を行うことができる。ただし、現状では PC 上での ChatGPT-4o との音声対話は一部の米国ユーザーに限られている。スマートフォンでの会話応答は可能であるが、実は古いモデルが採用されており、モデルは ChatGPT-4o ではない。日本のユーザーは、Chrome 拡張機能を利用することで、PC 上で ChatGPT-4o との音声対話が可能となり、テキストを入力する機能も利用できる。

プロンプトエンジニアリングでは、ChatGPT API を活用することで、より複雑な情報をプログラムに入力し、高度な評価を行うことができる。マルチモーダル AI のデモンストレーションは、テキスト、画像、音声といった異なるデータ形式を統合的に処理する AI の能力を示し、これをビジネスシナリオに応用するための実践的なアプローチを提供している。

また、大量の情報（数百ページ）を処理する際、LLM は文書の最初と最後の部分はよく記憶するが、中間部分の記憶があいまいになる傾向がある。このような課題に対処するため、実務では RAG（Retrieval-Augmented Generation、検索拡張生成）プロセスを活用することで、より正確な応答を生成することができる。しかし、RAG のチューニングは難易度が高く、まだ十分に洗練されていない。このような RAG 技術を利用するアプリケーションは、外部知識を基に LLM に適切な回答を生成させることが可能であり、実務においてもその有効性が期待されている。

・次頁のプロンプトを ChatGPT-4o の入力フィールドに直接入力するか、ファイルとしてアップロードすることで ChatGPT-4o が採用官として、審査が実行される。音声会話をするには前もって Chrome の拡張機能【Voice Control for ChatGPT】のインストールが必要。

===== プロンプト =====

prompt = ""

依頼内容

次の{# 制約条件}に従い、{# 形式}のフォーマットに従い、**コムチュア株式会社の面接官が求める{# 質問}に対する候補者の{# 回答}を評価せよ。**

質問

[Question]

回答

[Transcript]

形式

・**10点満点で評価する。評価理由を簡潔に2, 3文で述べる。**

制約条件

- ・次の{# Overview}の企業・組織への採用面接である。
- ・候補者には{# Qualifications}が求められる。

Overview (コムチュア株式会社のホームページから取得した情報)

■コムチュア株式会社について

コムチュアは、1985年の創業以来、「お客様には“感動”を、社員には“夢”を」をモットーに掲げ、DXを成長領域と捉え、イノベーションの実現を目指しています。お客様の要望を形に変えることで、付加価値の高いサービス提供を続け、社会とともに発展していくことを目指しています。

■事業内容と主要な取引先

クラウドソリューション事業、デジタルソリューション事業、ビジネスソリューション事業、プラットフォーム・運用サービス事業、デジタルラーニング事業を展開。主要取引先には、金融・保険、製造・その他、流通・通信・サービス、官公庁・公共・団体等があります。

Qualifications

■基本条件:

- ・新しい技術に対する探究心を持ち、イノベーションに積極的に取り組むことができる
- ・コミュニケーション能力が高く、チームワークを重視できる
- ・お客様からの要望を形にできる問題解決能力
- ・社会と共に成長し、貢献する意欲を持つ

■歓迎条件:

- ・DXやクラウドサービス、デジタルトランスフォーメーションに関連する経験
- ・多様な背景を持つチームメンバーと協力して働く経験
- ・持続可能な成長に向けた新しいアイデアや提案を行える創造力

■求める人物像:

- ・顧客第一を心がけ、感動を提供する意欲がある
- ・自ら学び、成長する姿勢を持ち、高い目標に挑戦する
- ・素直でプラス思考の持ち主
- ・改善提案を積極的に行い、会社の成長に貢献する意欲がある

""

Question = "あなたの強みを教えてください。"

Answer = transcript.text

prompt = prompt.replace("[Question]", Question)

prompt = prompt.replace("[Transcript]", Answer)

=====

スクラムとCRISP-ML(Q)を併用した分析プロジェクトの進め方

○吉野 祥

(所属株式会社クレスコ)

How to proceed with an analysis project using Scrum and CRISP-ML(Q)

Sho YOSHINO

CRESCO LTD.

要旨

スクラムはアジャイル開発手法のひとつで、短いサイクルで計画と実装を繰り返すソフトウェア開発のフレームワークである。計画と実装を繰り返す手法は分析と親和性が高いため、スクラムを利用して分析プロジェクトを進めることもある。しかし、スクラムはソフトウェア開発の手法であるため、分析とは異なる点が存在する。また、スクラムだけでは分析プロジェクトを俯瞰的にどのように進めていくかを示すことができない。

データ分析プロジェクトのプロセスモデルとして、CRISP-DM が広く使われている。しかし、機械学習モデルを活用するプロジェクトにおいてはモデルを作成し利用し始めた後の品質を確認する観点がなかった。

CRISP-ML(Q)は、CRISP-DM から品質の観点を追加し発展させたプロセスモデルである。

CRISP-ML(Q)と併用することで、俯瞰的にプロジェクトを進めることができる。

本論では、スクラムを分析プロジェクトで適用する際のソフトウェア開発との違いを述べ、分析プロジェクトへと適用させ、CRISP-ML(Q)との併用を考察する。

キーワード：スクラム, アジャイル, CRISP-ML(Q)

はじめに

アジャイルは迅速、俊敏などの意味があり、IT 業界におけるソフトウェア開発の用語として使われている。近年では変化の激しい環境への対応が求められることから IT 業界以外でも使われている。

IPA が発行している DX 白書 2023⁹⁾における調査では、アジャイル開発を日本では 4.8%が全社的に活用しており、18.1%が事業部で活用している。また 18.6%がアジャイルの活用を検討している。米国では日本よりも活用しており、25.9%が全社的に活用しており、28.0%が事業部で活用している。19.4%が活用を検討している。アジャイルの原則とアプローチは、ソフトウェア開発以外でも利用されており、DX 白書 2023 に

おける調査では、2021 年度で IT 部門を含めた全ての部門において 5 割超が取り入れていなかったが、2022 年度では取り入れていない部門が 5 割を下回っており、検討または導入が進んでいる傾向にある。

アジャイルは多くのアジャイルフレームワークが存在しており、スクラムはその中でよく使われるフレームワークの一つである。

Digital.ai 社が毎年実施しているアジャイルに対する調査結果である The 17th State of Agile Report^[5]において、アジャイルフレームワークのいずれかを利用しているかという調査では、SAFe^[15]は 26%、スクラムは 19% 利用されていた。2020 年に行われた The 14th State of Agile Report^[8]まではスクラムが 1 位であったが、2021 年に行われた The 15th State of Agile Report から、2023 年の The 17th State of Agile Report^[5,6,7]ではスクラムは 2 位、SAFe は 1 位であった。調査結果からは SAFe がアジャイルフレームワークの主流といえるが、SAFe は組織全体の変革も含めた大規模アジャイルであるため、小規模のアジャイルとしてはまだスクラムが主流と考えられる。

スクラムは主にソフトウェア開発で活用され、経営やマーケティングなど、他の分野でも使われ始めている。スクラムの公式ガイドであるスクラムガイドでは 2020 年の改訂版において IT に関する記述を排除し、IT 以外にも適用しやすいように改訂されている^[1]。

変化の激しい分野において迅速に対応する手法としてスクラムは有用であるため、分析においても有用であると考えられる。特に分析は小規模で行われることから、アジャイルフレームワークの中ではスクラムが適用しやすい。Baijens らは、スクラムを分析に適用させ、CRISP-DM と組み合わせた手法 Scrum-DS を考案している^[4]。

CRISP-DM は分析のプロセスモデルの標準として広く使われている手法である。KDnuggets による 2014 年の調査では 200 人中 86 人が CRISP-DM を活用しており最も使われているプロセスモデルであった^[13]。Data Science Process Alliance による 2024 年の調査では 109 人の内、最も多い 49% が CRISP-DM を活用していた^[12]。また、同調査では 18% がスクラムを活用しており、CRISP-DM に次いで活用されていた。

CRISP-DM は広く使われているが、データマイニングを中心に考えられたものであり、CRISP-DM が生まれてから 20 年以上が経過し、当時とは状況が変化していることから、CRISP-DM を参考に多くのモデルが生まれている^[16]。例えば、Microsoft 社が提案した TDSP^[17]や SAS 社が提案した SEMMA^[18]がある。

CRISP-ML(Q)は、Schwaber らにより提案された CRISP-DM と互換性のあるプロセスモデルである^[3]。CRISP-ML(Q)は CRISP-DM の拡張で、機械学習における品質保証の手法を取り入れ、プロセスだけでなく、成果物への品質も考慮している。また、CRISP-DM とプロセスに大きな違いがないため、既に CRISP-DM を活用している組織にも適用しやすいのが特徴である。

本論では、スクラムを分析に適用し、品質保証の手法を取り入れた CRISP-ML(Q)との併用について考察する。

スクラム

スクラムは、スクラムの理論である「透明性」「検査」「適応」を実現するために、チーム構成であるスクラムチーム、スクラムで作成されるスクラムの作成物、スクラムで実施するスクラムイベントが定義されている。

スクラムチーム

スクラムチームは3人から10人で、複数人の**開発者**と**プロダクトオーナー**1人、**スクラムマスター**1人から構成されている。それぞれが兼任することはない。

開発者は、スプリントの計画であるスプリントバックログを作成し、スプリントゴールを達成するために日々バックログの内容を実施することになる。分析プロジェクトであれば専門家であるデータサイエンティストが担うことになる。組織や実施する内容によっては、データ分析基盤を開発・管理するデータエンジニアやアプリケーションを開発するソフトウェアエンジニア、アプリケーションの運用・保守するオペレーションエンジニアなど、データサイエンティスト以外が開発者として担うこともある。

プロダクトオーナーと**スクラムマスター**が担う責任はソフトウェア開発と大きな違いはない。

プロダクトオーナーは、プロダクトバックログの管理に責任を持つ。プロダクトゴールを策定しチームやステークホルダーに明示的に示す。プロダクトバックログを作成、状況に応じて適時並び替え、常に明示的に理解されるようにする。

スクラムマスターはスクラムチームがスクラムの有効性を理解してもらい、チームが生み出す価値に集中できるように支援する。プロダクトオーナーの支援やスクラムイベントを問題なく実施できるようにするだけでなく、進捗を妨げるようなスクラムチーム外からの影響も防ぐことも必要である。

例えば、ステークホルダーが要件の変更を開発者に直接伝えることを防ぎ必ずプロダクトオーナーを通すようにする。プロダクトゴールに向けて、なにをどのように実行すればよいのか、プロダクトバックログの管理はプロダクトオーナーが責任をもつため、その責任が移ることがないように防ぐ必要がある。

スクラムの作成物

スクラムの作成物は、**プロダクトバックログ**、**スプリントバックログ**、**インクリメント**から構成される。特にスクラムの作成物は分析とソフトウェア開発での大きな違いがある。

プロダクトバックログと**スプリントバックログ**はソフトウェア開発のものと同様に活用できる。プロダクトバックログは、ソフトウェア開発であれば、新機能の開発や機能の修正などになるが、分析ではデータから知ることができること、現状の把握や考察がバックログとなる。スプリントバックログでは、その把握をするためにはどのデータをどのように収集・加工・集計しなければならないなどタスクを分解していく。

バックログの作成においては、それぞれの優先度の設定や見積もりを行うことになる。分析では未知のデータを扱うこともあることから、初見での見積もりが非常に困難である。例えば、機械学習のモデルを学習するために、学習環境やデータ量と質、利用するアルゴリズムによっては、実行に日単位での時間がかかることもある。そのため、概念検証(PoC)として、一通りのタスクを先行して実施する、または繰り返しの初めてのタスクであれば仮の見積もりを行い、1回目の結果を用いて、繰り返しの2回目以降に見積もりの精度を上げるように努める。結果によっては、今後の計画にも大きく影響することもあるため、リスクとしてスクラムチームとステークホルダーで共有し話し合う必要がある。

インクリメントは、ソフトウェア開発ではスプリントで開発した動くアプリケーションになるが、分析においては必ずしもアプリケーションとは限らない。スプリント内で得られた分析結果から、その解釈と考察も成果物となる。考察をしやすくし今後の方針を決めやすくするために、ドキュメントを作ることもある。

アジャイルソフトウェア開発宣言^[17]では、「包括的なドキュメントよりも動くソフトウェアを」とあるが、これはドキュメントを作らなくてよいということではない。比較してどちらに価値を向けているか、ということであるため、価値があるならばドキュメントは作るべきである。スクラムチームとステークホルダ

ーはプロジェクトにおいて、スプリントごとにどのような作成物が作成されるか、スプリントゴールを事前に明確化しておくことで齟齬なく協力して進めやすくなるだろう。

スクラムイベント

スクラムイベントは、**スプリントプランニング**、**デイリースクラム**、**スプリントレビュー**、**スプリントレトロスペクティブ**の4つのイベントがあり、それらを内包するスプリントで構成される。チームはスプリントを何度も繰り返し、プロダクトゴールを目指す。

スプリントの長さは1ヶ月以内とされている。特に決まりはないが1週間単位で設定することが多い。スプリントごとに実施する内容で期限を決めず、期限を決めてからその期限内でできることを決めるタイムボックスという時間管理手法を用いる。もしスプリント内でスプリントゴールを満たさなくても決して延長してはならない。なぜそのような結果になったかを話し合い次に活用することが重要である。

スプリントの初めにスプリントプランニングを実施し、毎日デイリースクラムを実施、スプリントの終わりに、スプリントレビューとスプリントレトロスペクティブを実施する。

スプリントプランニングでは、スプリントの初めに実施するイベントである。プロダクトゴールに向けてプロダクトバックログを調整し、今回のスプリントでなにが実施できるかを話し合う。スプリントが1ヶ月であれば、最大8時間実施する。スクラムチームだけでなく、必要であるならばチーム外のメンバーも招集する。

分析では結果によって今後の計画が変わることがよくある。スプリントバックログに設定しても結果によっては不要になることもある。そのため、開発時間のすべてをバックログの消化に埋めるのではなく、ある程度のバッファを設けることも検討する。

デイリースクラムでは、毎日実施する15分のイベントである。開発者は進捗を検査し、スプリントバックログを調整する。スプリントバックログはデイリースクラムのときだけでなく、それ以外の時間でも調整してよい。特に分析では新たな発見により今後の方針が変化することがよくあるため、適時調整する。

スプリントレビューは最後から2番目に実施するイベントで、スプリントが1ヶ月の場合、最大4時間実施する。

ソフトウェア開発ではアプリケーションのデモをステークホルダーに見せることになるが、インクリメントで述べたように、分析ではアプリケーションを必ずしも見せることはできない。機械学習のモデルを作成し、アプリケーションに組み込むようなプロジェクトであれば、デモを見せることもあるが、モデルを作成する前やアプリケーションを作成しないプロジェクトでは動くアプリケーションを見せることはないだろう。

分析において動くアプリケーションを見せることはできないが、データの基礎統計や相関関係、モデルの評価指標結果などの結果を見せることはできる。しかし、これらの結果を見せるだけでは、プロダクトゴールに向けて、適切なフィードバックは得られない。提示した結果がプロダクトゴールに向けて、どのような変化をもたらすかが一見するだけでは判断が容易ではないからである。結果には、解釈と考察を添えることで、フィードバックが得られやすくなるだろう。場合によってはプロダクトバックログを調整し、今後の方針が変わることになる。

スプリントレビューでは品質の基準を満たしていない、リリースしていないアプリケーションはステークホルダーに見せないことになっている。分析では方針が何度も変わることも多いことから、まだ結果を示せる状態でなくても、今後の計画に影響を及ぼす可能性があるならば、示すことを検討すべきだろう。

スプリントレトロスペクティブでは、スプリントの最後に実施するイベントである。スプリントが1ヶ月

であれば最大3時間実施する。

スクラムチーム内で今回のスプリントを検査し、問題があれば、共有し、今後のスプリントで活用できるように話し合う。その問題が大きく時間がかかるようであれば、問題に関するスプリントバックログを追加する。

スクラムイベントではないが、リファインメントはスクラムイベントの中でも必要なタスクの一つである。

リファインメントは、プロダクトバックログの追加、見積もり、並び替えして最新の状態に保ち、次のスプリントの作業にすぐ移れるようにすることを指す。スクラムイベントではないため、いつ実施してもよい。スプリントプランニングの中で実施することもあれば、スクラムイベント外で個別に実施することもある。分析では、分析の結果によって要件が変化することがよくあるため、適時実施するべきである。場合によってはスプリントレビューでステークホルダーと話す機会を待たずして、ステークホルダーと話し合う機会を設けたほうがよいだろう。参加者は必ずしもスクラムチーム、ステークホルダーすべてではないため、内容に応じて柔軟に判断する。

CRISP-ML(Q)

CRISP-ML(Q)は下記6つのフェーズで構成されている。

1. ビジネスとデータの理解
2. データの準備
3. モデリング
4. モデル評価
5. モデルデプロイ
6. モニタリングとメンテナンス

CRISP-DM と違い、ビジネスとデータの理解が一つのフェーズに統合されている。ビジネス目標が、保有しているデータと密接に関係しており、それぞれに影響があるため、CRISP-DM と違い別フェーズではなく同じフェーズとなっている。

追加されたフェーズとして、モデルデプロイ後にモニタリングとメンテナンスがある。これは構築したモデルが劣化することを防ぐ目的がある。データはいつまでも同じようなデータ分布になることはなく、どんなデータも基本変化するものである。訓練時のデータにはないデータ分布に変化すると、モデルの精度が落ちるなどの問題が発生するため、デプロイ後も確認を継続しなければならない。

CRISP-ML(Q)のもう一つの特徴として、各フェーズに品質保証のためのプロセスが追加されている。各フェーズで確認すべきタスクが設けられており、その基準を満たさなければ品質を満たしていないとして、フェーズを繰り返すプロセスとなっている。機械学習では考慮していない事柄によっては、その後の大きく結果を変えてしまうことがある。また問題が起きても容易には気づかないこともある。そのような問題を回避するために品質保証のプロセスが追加されている。

フェーズのはじめに要件を定義し、実施するタスクを洗い出す。発生しうる品質上のリスクを特定し、もし品質上のリスクが発生しうるのであればリスクを軽減するようにフェーズを繰り返す。リスク軽減後、タスクをすべて消化するように繰り返し、すべて消化したとき次のフェーズに移る。

表1~6に筆者が要約した各フェーズで確認すべき項目を示す。

表 1. CRISP-ML(Q) 「ビジネスとデータの理解」フェーズにおける品質保証のための確認項目

1. ビジネスとデータの理解	
機械学習アプリケーションの範囲を定義	ビジネス面と機械学習面で満たすべき範囲を定義
成功基準を定義	ビジネス成功基準
	機械学習成功基準
	経済的成功基準
実現可能性の検討	機械学習技術の適用性; 概念検証 (PoC) で確認
	法的制約
	アプリケーションの要件; 堅牢性、スケーラビリティ、説明可能性、リソース需要など
データ収集	データ収集が可能か
	再現性と品質を確保するためデータをバージョン管理
データ品質の検証	データがどのように作られるか、どのような値であるか、ドメイン知識も含めて説明できるように調査、記述
	期待される特徴値、上下限値や分布、値のリスト、データ形式などであるかをデータの要件として記述
成果物文書のレビュー	上記をまとめて実現可能であるかをレビュー

表 2. CRISP-ML(Q) 「データの準備」フェーズにおける品質保証のための確認項目

2. データの準備	
データの選択	特徴量選択; フィルター法、ラッパー法、クロスバリデーションなど
	利用するデータの選択
	不均衡データへのサンプリング
データクリーニング	ノイズ除去
	データの補完; 欠損値補完
データ加工	特徴量エンジニアリング; ビニング、特徴量追加、クラスタリングやPCAなどの次元削減、オートエンコーダ、One-Hot エンコーディングなど
	データ拡張; 画像に回転、弾性変形、ガウスノイズを適用
データの標準化	データ加工、モデル学習できるように、ファイル形式を揃える
	正規化

表 3. CRISP-ML(Q) 「モデリング」 フェーズにおける品質保証のための確認項目

3. モデリング	
類似問題の文献調査	類似した出版物、特許、内部レポートなどのスクリーニング
モデルの品質尺度を定義する	パフォーマンス：未知のデータに対するモデルの性能評価
	頑健性：一貫性のない入力や実行環境の障害に対するモデルの耐性
	スケーラビリティ：本番システムにおける大量のデータに対するモデルの拡張能力
	説明可能性：モデルの直接的または事後的な説明可能性
	モデルの複雑さ：データの複雑さに見合ったモデルの能力
	リソース需要：モデルの展開に必要なリソース
モデルの選択	どのモデルを利用するか
モデルの学習	目的関数、最適化、正則化、クロスバリデーションなど
目的変数なしデータの活用	教師なし学習や半教師あり学習などを用いて学習に利用
モデルの圧縮	ディープラーニング：Pruning（枝刈り）、Quantize（量子化）、Distillation（蒸留）
アンサンブル法	アンサンブル法：ブースティング、バギング、混合エキスパート
再現性の確保	手法の再現性；アルゴリズム、データセット、ハイパーパラメータ、実行環境
	結果の再現性；平均性能を検証し、異なるランダムシードにて評価
	文書としてまとめる

表 4. CRISP-ML(Q) 「モデル評価」 フェーズにおける品質保証のための確認項目

4. モデル評価	
性能評価	テストデータを用いてモデルの性能を評価
ロバスト性の決定	
説明可能性の向上	
定義された成功基準と結果を比較する	結果を文書化し、成功基準と比較する

表 5. CRISP-ML(Q) 「モデルデプロイ」 フェーズにおける品質保証のための確認項目

5. モデルデプロイ	
推論ハードウェアを定義する	本番環境でモデルを学習するハードウェアを選定
本番環境下でのモデル評価	本番データと訓練データとの違いを確認
ユーザー受容性とユーザビリティを保証する	
不測のエラーのリスクを最小限に抑える	
デプロイ戦略	本番環境にデプロイする計画の策定

表 6. CRISP-ML(Q) 「モニタリングとメンテナンス」フェーズにおける品質保証のための確認項目

6. モニタリングとメンテナンス	
本番環境のモニタリング	非定常なデータ分布：訓練データから変化していないか
	アプリケーションの使用状況とパフォーマンスを追跡
	A/B テスト
モデル更新	新しいデータへの対応
本番環境のメンテナンス	ハードウェアの劣化監視
	システムのアップデートの影響監視

スクラムと CRISP-ML(Q)

スクラムと CRISP-ML(Q)は、それぞれの用語の違いはあるが、共通するプロセスがある。

CRISP-ML(Q)における要件の定義やタスクを洗い出す作業は、スプリントプランニングにて、スプリントゴールの設定やバックログの作成に当てはまる。各フェーズですべき品質保証のタスクをバックログやスプリントゴールとして利用できる。

CRISP-ML(Q)における品質を満たさなければフェーズを繰り返すプロセスは、スクラムではそのままスプリントとして適用できる。また、スクラムにおけるスプリントでは「品質を低下させない」ことが設けられており、考えとしても適合している。品質保証、スプリントゴールを満たさなければ、スプリント、そのときのフェーズを繰り返すことになる。CRISP-ML(Q)の品質保証は機械学習における抽象的な確認事項であるため、プロジェクトに応じてより詳細化してスプリントゴールを設定する必要はあるが、スプリントゴールを作るための参考となるだろう。

スプリントとしては、CRISP-ML(Q)におけるフェーズごとに区切るか、複数のフェーズをまとめてスプリントと設定することになる。

「ビジネスとデータの理解」フェーズは最初のスプリントとなる。成功基準、データ収集、データ品質が実現可能であるかを示した文書がスプリントで作成する成果物となる。この成果物は他のフェーズでの活用するものとなる。「ビジネスとデータの理解」フェーズはその後を決める大きなフェーズであるため、スプリントを何度も繰り返すことになるだろう。状況によってはプロジェクトの中止も判断する。

次は「データの準備」フェーズでスプリントとなる。Baijens らによる研究では「データの準備」フェーズに時間がかかることから初めての「データの準備」フェーズのスプリントではスプリントゼロとして、他のフェーズを交えないスプリントとしている^[4]。組織のルールやシステムの制約、データマネジメントの状況などによって、大きく影響するフェーズであるため、初回はスプリントを単独で実施するのが良いだろう。初回のスプリントの状況によって、「データの準備」フェーズを繰り返す、次のフェーズに移るか考慮する。

その後のフェーズである「モデリング」「モデル評価」フェーズは、状況に応じて、それぞれスプリントを用意するが、データが概ね揃い、実行までに時間がかからなくなれば、「データの準備」「モデリング」「モデル評価」フェーズを1つのスプリントとして設定する。1つのスプリント内で、何度も繰り返すこともあるだろう。再現性を保つため、手順や設定値は文書に残して成果物とする。

「モデル評価」フェーズでの結果は成果物として文書にまとめ、「ビジネスとデータ理解」フェーズにおける成功基準と比較し、基準を満たした場合、「モデルデプロイ」フェーズへ移行する。

本番環境にモデルがデプロイ後、「モニタリングとメンテナンス」フェーズに移行する。

その後、新たなデータによりモデルの劣化があれば、再び前のフェーズである「データの準備」フェーズに戻りフェーズを繰り返していくことになる。

スクラムチームの構成は CRISP-ML(Q)においても変更はないが、開発者は全てのフェーズで同じメンバーとは限らない。データ分析基盤を開発・保守するデータエンジニアと、データ分析を実施するデータサイエンティスト、アプリケーションを開発するソフトウェアエンジニア、アプリケーションの運用・保守を行うオペレーションエンジニアなど、組織によって多くのエンジニアが関わるだろう。例えば、データの収集は管理上データエンジニアしか行えず、モデリングからモデル評価まではデータサイエンティスト、デプロイはデータエンジニアとソフトウェアエンジニアが行い、モニタリングとメンテナンスはオペレーションエンジニアが行うなど分かれることが考えられる。フェーズごとに異なる開発者を想定して進めなければ、区切りごとにプロジェクトが滞ることになる。後半のフェーズで招集されると開発に問題が生じることもあるため、できるかぎり早い段階で、将来的に開発者になりうるメンバーは適時スクラムイベントに招集する、引き継ぐための文書をバックログとして設けるなど、影響が少ないように事前に調整しなければならない。

また、「モニタリングとメンテナンス」フェーズでは、運用保守がメインとなった場合、データ品質の監視は続けるが、組織によっては運用保守を扱うチームに引き継ぎスクラムチームを解散することもあるだろう。引き継ぐ先がスクラムチームでない可能性もある。フェーズごとに開発者が変化するときと違い、それまでスクラムチームにいなかったメンバーに引き継ぐことも考えられる。この場合も、もし可能ならばチームが解散する前に適時スクラムイベントに招集し、引き継ぐための準備やドキュメントの作成をバックログとして用意すべきである。

おわりに

スクラムと CRISP-ML(Q)の併用は、スクラムのチーム構成やイベント、作成物を崩すことなく適用でき、分析プロジェクトにおいて活用する方法を考察した。スクラムを使った分析プロジェクトのいち所となれば幸いである。

本論では、品質保証のために CRISP-ML(Q)を採用したが、挙げられている品質保証では品質が十分に満たされないこともあるだろう。CRISP-ML(Q)では機械学習の品質を考える上で基本的なポイントを抑えているが、利用するアルゴリズムや環境、規則、完成基準などによって品質基準は異なるため、プロジェクトごとにどこまで考えなければならないか決める必要がある。例えば、機械学習のより詳細な品質基準を決める際には、ML Test Score^[10]や機械学習品質マネジメントガイドライン^[11]が参考に、より品質を保つ手法を考えていく必要がある。

また、今回は1つのスクラムチームを想定した。データサイエンティスト、データエンジニア、ソフトウェアエンジニアなど複数の職種が交わると、1つのスクラムチームではなく、複数のスクラムチームでプロジェクトを進めることもありえる。そのため、LeSS^[14]やSAFe^[15]といった複数のチームを想定した大規模アジャイルフレームワークでの進め方も考える必要がある。

参考文献

[1] K. Schwaber ; J. Sutherland, “The Scrum Guide: The Definitive The Rules of the Game”, 2020.

- [2] Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R.,
“CRISP-DM 1.0 Step-by-step data mining guide. Technical report”, The CRISP-DM consortium, 2000.
- [3] Studer, S.; Bui, T.B.; Drescher, C.; Hanuschkin, A.; Winkler, L.; Peters, S.; Müller, K.-R.,
“Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology”, 2021.
- [4] Baijens, J.; Helms, R. W., Iren, D., “Applying Scrum in Data Science Projects”, 2020.
- [5] digital.ai, “The 17th State of Agile”, 2023.
<https://info.digital.ai/rs/981-LQX-968/images/RE-SA-17th-Annual-State-Of-Agile-Report.pdf>, (参照 2024-08-20)
- [6] digital.ai, “The 16th State of Agile”, State of Agile, 2022.
<https://info.digital.ai/rs/981-LQX-968/images/SOA16.pdf>, (参照 2024-08-20)
- [7] digital.ai, “The 15th State of Agile”, 2021.
<https://info.digital.ai/rs/981-LQX-968/images/SOA15.pdf>, (参照 2024-08-20)
- [8] digital.ai, “The 14th State of Agile”, 2020.
<https://info.digital.ai/rs/981-LQX-968/images/SOA14.pdf>, (参照 2024-08-20)
- [9] IPA 独立行政法人 情報処理推進機構, “DX 白書 2023”, 2023.
<https://www.ipa.go.jp/publish/wp-dx/dx-2023.html>, (参照 2024-08-20)
- [10] Eric Breck; Shanqing Cai; Eric Nielsen; Michael Salib; D. Sculley,
“The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction”, 2017.
- [11] 国立研究開発法人産業技術総合研究所 デジタルアーキテクチャ研究センター (DigiARC) ; サイバーフ
ィジカルセキュリティ研究センター (CPSEC) ; 人工知能研究センター(AIRC), “機械学習品質マネジメント
ガイドライン”, AIST デジタルアーキテクチャ研究センター, 2023.
<https://www.digiarc.aist.go.jp/publication/aiqm/AIQuality-requirements-rev4.2.0.0113-signed.pdf>, (参照 2024-08-20)
- [12] Jeff Saltz, “CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects”, Data Science
Process Alliance, 2024. <https://www.datascience-pm.com/crisp-dm-still-most-popular/>, (参照 2024-08-20)
- [13] Gregory Piatetsky, “CRISP-DM, still the top methodology for analytics, data mining, or data science projects”,
KDnuggets , 2014.
<https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>, (参
照 2024-08-20)
- [14] Larman, C. ; Vodde, B. , “Large-Scale Scrum More with LeSS”, Addison-Wesley, 2016.
- [15] Knaster, R.; Leffingwell, D. , “SAFe 5. 0 Distilled: Achieving Business Agility with the Scaled Agile Framework.”,
Addison-Wesley , 2020.
- [16] Martínez-Plumed, F.; Contreras-Ochando, L.; Ferri, C.; HernandezOrallo, J.; Kull, M.; Lachiche, N. J. A. H.;
Ramírez-Quintana, M. J.; Flach, P. A.,
“CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories.”, 2019.
- [17] Kent Beck; Mike Beedle; Arie van Bennekum; Alistair Cockburn; Ward Cunningham; Martin Fowler; James
Grenning; Jim Highsmith; Andrew Hunt; Ron Jeffries; Jon Kern; Brian Marick; Robert C. Martin; Steve Mellor; Ken
Schwaber; Jeff Sutherland; Dave Thomas,
“Manifesto for Agile Software Development”, 2001. <https://agilemanifesto.org/>, (参照 2024-08-20)

時間依存性ROC曲線に対する推定方法の性能比較

○久貝 航太郎, 魚住 龍史

東京工業大学工学院経営工学系

Performance comparison of estimating methods for time-dependent ROC curves

Kotaro Kugai and Ryuji Uozumi

Department of Industrial Engineering and Economics, Tokyo Institute of Technology

要旨

特定時点までのイベント発生の累積情報に基づいて感度および特異度を推定し、時点ごとの ROC 曲線を描出したものを時間依存性 ROC 曲線 (time-dependent receiver operating characteristic curve) という。時間依存性 ROC 曲線は、予後マーカーの性能の経時変化を評価する目的などに用いられ、その推定方法として、条件付 Kaplan-Meier 法、最近傍法、IPCW (inverse probability of censoring weighting) 法、再帰法の 4 つ方法が SAS に実装されているが、それらの推定の統計的性能は詳しく評価されていない。本稿は、これらの手法の推定精度の優劣が、生存時間データの設定を変えることによって変化するかどうかを調べ、各推定法の使い分けについて考察した。モンテカルロ・シミュレーションの結果、条件付 Kaplan-Meier 法、IPCW 法、再帰法の 3 手法は近い精度で AUC を推定した。一方、最近傍法はどの設定でも精度が悪く、打ち切りがマーカーと独立という条件の下では他手法を選択すべきだと考えられた。再帰法は追跡初期の精度が非常に優れていたが、リスク集合の小さくなる追跡後半で精度が大きく落ちる傾向が見られた。

キーワード：生存時間解析 PHREG Cox 比例ハザードモデル 時間依存性 ROC 曲線 時間依存性 AUC
IPCW 法 条件付 Kaplan-Meier 法 最近傍法 再帰法

1 はじめに

医学分野では、バイオマーカー（臨床検査値）を用いて対象者の疾患の有無を検査することや、死亡リスクを予測することで予後向上のための適切な治療を施すことが必要となる。時間依存性 ROC 曲線 (time-dependent receiver operating characteristic curve) は、2 値診断の評価に用いられる ROC 曲線を生存時間データに対して拡張したものであり、バイオマーカーが対象患者の予後を予測する予後マーカーとしての機能を有しているかを評価する手法として有用である。時間依存性 ROC 曲線を構成する時点ごとの感度と特異度を推定する手法として、条件付 Kaplan-Meier 法や最近傍法[1]、IPCW (inverse probability of censoring weighting) 法[2]、再帰法[3]などの現在代表的な手法が提案されており、これらの手法の推定式や特徴をまとめたレビュー記事も報告されている[4]。しかし、生存時間データの設定に応じて、どの推定方法を用いるのが適切か、統計的な性能比較は十分になされていない。本稿は、生存時間データの様々な状況を想定したもとの、時間

依存性 ROC 曲線の推定方法を数値シミュレーションによって統計的に性能比較する。

2 時間依存性 ROC 曲線

2.1 ROC 曲線の定義

被験者のマーカー値（臨床検査値）があるカットオフ値より高ければ陽性、そうでなければ陰性と診断するような検査を考える。このとき、疾患をもつ症例に対して検査で陽性となる確率を感度 (sensitivity)、疾患をもたない症例に対して検査で陰性となる確率を特異度 (specificity) と呼ぶ。対象者*i*の疾患の有無を D_i 、マーカー値を X_i 、カットオフ値を c とすると、感度と特異度は以下の式で表される。

$$\text{感度 (sensitivity)} : \text{Se}(c) = \Pr(X_i > c \mid D_i = 1)$$

$$\text{特異度 (specificity)} : \text{Sp}(c) = \Pr(X_i \leq c \mid D_i = 0)$$

感度と特異度はカットオフ値 c の関数であり、 c を変えることにより変化する。可能なすべてのカットオフ値に対し縦軸に（感度）、横軸に（1 - 特異度）としてひとつの曲線を描くことができ、これを ROC 曲線という [5]。

2.2 時間依存性 ROC 曲線における感度と特異度

時点 t における ROC 曲線を構成するためには、時点 t における感度と特異度を求めなければならない。いま、対象者*i*の生存時間 T_i およびマーカー値 X_i が得られているとする。カットオフ値を c としたときの時点 t における感度には 2 種類の形式があり、以下のように定義される。

$$\text{累積 (cumulative)} : \text{Se}^c(c, t) = \Pr(X_i > c \mid T_i \leq t) = \Pr(X_i > c \mid D_i(t) = 1)$$

$$\text{発現 (incident)} : \text{Se}^l(c, t) = \Pr(X_i > c \mid T_i = t) = \Pr(X_i > c \mid dD_i(t) = 1)$$

累積感度は、時間 t 以下で死亡する対象者*i*のマーカー値 X_i が c よりも大きな値をもつ確率、発現感度は時間 t で死亡する対象者*i*のマーカー値 X_i が c よりも大きな値をもつ確率である。ここで、 $D_i(t)$ は $D_i(t) = I\{T_i \leq t\}$ で定義される。 $I\{\cdot\}$ は指示関数であり、対象者*i*が時間 t 以内に死亡する場合は $D_i(t) = 1$ 、そうでない場合は 0 となる。これらの感度と同様に、時点 t における特異度にも 2 種類の形式がある。

$$\text{動的 (dynamic)} : \text{Sp}^D(c, t) = \Pr(X_i \leq c \mid T_i > t) = \Pr(X_i \leq c \mid D_i(t) = 0)$$

$$\text{静的 (static)} : \text{Sp}^S(c, t) = \Pr(X_i \leq c \mid T_i > t^*) = \Pr(X_i \leq c \mid D_i(t^*) = 0)$$

動的特異度は、時間 t よりも長く生存する対象者*i*のマーカー値 X_i が c 以下の値をもつ確率である。時点 t を動的に変更できる動的特異度に対し、静的特異度ではある任意の時間 t^* に固定される[6]。

本稿では、これらの形式のうち累積感度と動的特異度から構成される累積-動的(C/D)ROC 曲線に焦点を当てる。

2.3 SASによる時間依存性ROC曲線の推定

SASでは、PHREGプロシジヤを用いて時間依存性ROC曲線の推定を実行でき、以下の4種類の推定方法がサポートされている。推定方法は、オプションのROCOPTIONS内のMETHOD=で指定できる[5]。

1. IPCW法 (METHOD = IPCW)
2. 条件付 Kaplan-Meier法 (METHOD = KM)
3. 最近傍法 (METHOD = NNE)
4. 再帰法 (METHOD = RECURSIVE)

これらの推定方法は、累積感度と動的特異度をそれぞれ違う方法で推定している。具体的な数式は次節で解説するが、4つの推定方法の違いは打ち切りの扱いにあり、もし解析する生存時間データに打ち切りがない場合、IPCW法、条件付 Kaplan-Meier法、再帰法の3手法はまったく同じ値の感度と特異度を推定する。

表1 時間依存性ROC曲線の各推定方法の長所と短所

手法	長所	短所
IPCW法	・単調な感度と特異度を推定する	・打ち切りとマーカーが独立であるという仮定を必要とする
条件付 Kaplan-Meier法	・平易である	・非単調な感度と特異度を推定する場合がある ・打ち切りとマーカーが独立であるという仮定を必要とする
最近傍法	・単調な感度と特異度を推定する ・打ち切りとマーカーが独立でなくてもよい	・平滑化パラメータを要する
再帰法	・単調な感度を推定する	・非単調な特異度を推定する場合がある ・再帰的な計算を要する

表1は時間依存性ROC曲線の推定方法の性質として挙げられる「単調性」と「打ち切りとマーカーの独立への頑健性」についてまとめたものである[4]。通常のROC曲線は範囲[0.5, 1]内に必ず収まり、点(0,0)から点(1,1)にかけての階段状の形である。別の言い方をすると感度と特異度はトレードオフの関係にあり、カットオフ値 c が上がると感度が下がって特異度が上がり、 c が下がると感度が上がって特異度が下がる。つまり感度と特異度は c に対して単調である。しかし、時間依存性ROC曲線の推定の場合、推定法によっては c に対する単調性が保証されず、定義が「確率」であるはずの感度と特異度の推定値が[0,1]からはみ出ることがある。したがって「単調性」を満たさないことは明確な欠点である。また、治療後すぐに患者の様態が悪化し追跡を中止したような例を考える。この場合、この患者の脱落はマーカーと何らかの関連があったと考える方が自然であり、打ち切りとマーカーは独立でなかったといえる。このような打ち切り例を解析するときは注意を要するが、最近傍法などのいくつかの推定方法では、このような症例を含むデータに対しても時間依存性ROC曲線法を適応可能である。当然、「打ち切りとマーカーの独立への頑健性」があることは、解析できるデータの幅が広いということであり長所である。

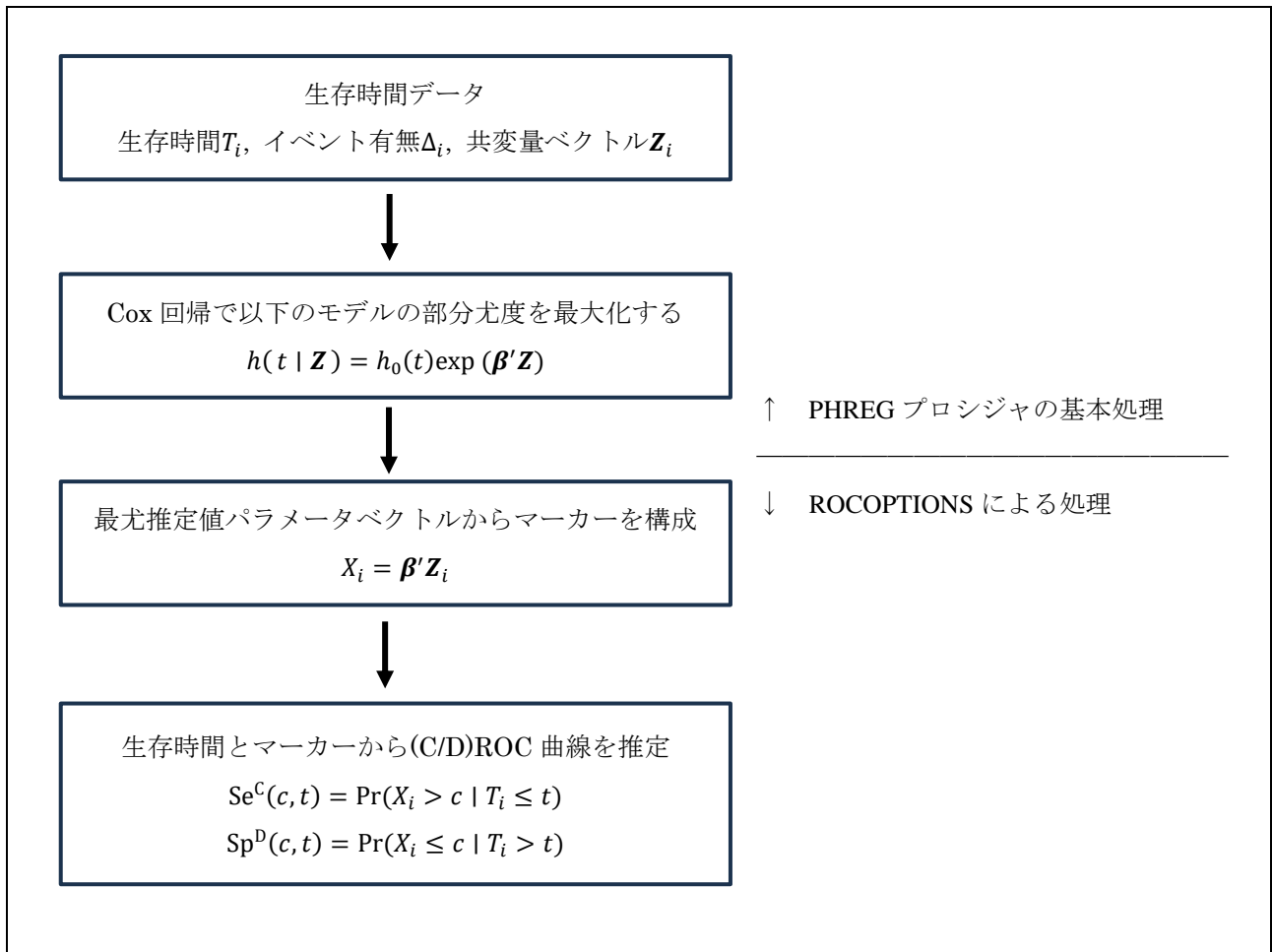


図 1 : PHREG プロシジャ内での時間依存性 ROC 曲線の推定の処理

時間依存性 ROC 曲線を推定する SAS プログラム[8]

```

PROC PHREG
  Data = SurvivalData /*解析対象のデータセットの指定*/
  ROCOptions(method = KM at = (1, 3, 5) auc) /*時間依存性 ROC 曲線の推定*/
  Model Time * Event(0) = Covariate; /*マーカーを構成する共変量*/
RUN;
  
```

また、前述した感度・特異度の定義からもわかるように、ROC 曲線のマーカーは実数値（曲線の性質上連続値である方が好ましい）が用いられる。一つの共変量のみを予後予測に用いる場合は値をそのまま X として用いて構わないが、複数の共変量を複合して使用したい場面では、複数の共変量を一つの実数値に変換する必要がある。その場合、線形予測子 X は k 次元共変量ベクトル Z を用いて $X = \beta' Z = \beta_1 Z_1 + \dots + \beta_k Z_k$ という線形結合で表される。

2.4 各推定方法の概要

本節では SAS に実装されている 4 つの推定手法の定義を概説する。各手法で共通する変数として、 n を症例数、 T_i を対象者 i ($i = 1, \dots, n$) の生存時間、 X_i を先述した方法で共変量から構成された対象者 i の線形予測子、 Δ_i を対象者 i のイベント観測の有無 (1 であれば死亡、0 であれば打ち切り)、 $I(\cdot)$ を指示関数 ($()$ 内の条件式が真ならば 1、そうでなければ 0 をとる関数) とする。

2.4.1 IPCW 法

$\hat{G}(t)$ を時点 t における打ち切り分布のカプラン・マイヤー推定量とする。打ち切り分布と生存時間分布が独立という仮定の下で、以下のように感度と特異度を推定する。

$$\widehat{\text{Se}}(c, t) = \frac{\sum_{i=1}^n \Delta_i I(X_i > c, T_i \leq t) / \hat{G}(T_i)}{\sum_{i=1}^n \Delta_i I(T_i \leq t) / \hat{G}(T_i)}$$

$$\widehat{\text{Sp}}(c, t) = \frac{\sum_{i=1}^n I(X_i \leq c, T_i > t)}{\sum_{i=1}^n I(T_i > t)}$$

IPCW 法は単調な感度と特異度を推定する。

2.4.2 条件付カプラン・マイヤー法

ベイズの定理を用いると、累積感度と動的特異度の定義は以下のように書き直すことができる。

$$\text{Se}^c(c, t) = \Pr(X_i > c \mid T_i \leq t) = \frac{[1 - S(t \mid X_i > c)] \Pr(X_i > c)}{1 - S(t)}$$

$$\text{Sp}^D(c, t) = \Pr(X_i \leq c \mid T_i > t) = \frac{S(t \mid X_i \leq c) \Pr(X_i \leq c)}{S(t)}$$

ここで、 $S(t)$ の推定量としてカプラン・マイヤー推定量 $S_{KM}(t)$ を使い、累積分布関数 $F_X(c) = \Pr(X \leq c)$ の推定量として経験分布関数 $\hat{F}_X(c) = \sum_i I(X_i \leq c) / n$ を用いる。

$$\widehat{\text{Se}}(c, t) = \frac{[1 - \hat{S}_{KM}(t \mid X_i > c)] (1 - \hat{F}_X(c))}{1 - \hat{S}_{KM}(t)}$$

$$\widehat{\text{Sp}}(c, t) = \frac{\hat{S}_{KM}(t \mid X_i \leq c) \hat{F}_X(c)}{\hat{S}_{KM}(t)}$$

非常に平易に計算できる一方で、非単調な感度と特異度を計算する。

2.4.3 最近傍法

最近傍法は、Akritas (1994) による 2 変量生存関数の最近傍推定量を組み込んだ手法である。2 変量生存関数 $S(c, t) = \Pr(X > c, T > t)$ を以下のように推定する。

$$\hat{S}_{b_n}(c, t) = \frac{1}{n} \sum_i \hat{S}_{b_n}(t | X = X_i) I(X_i > c)$$

ここで、条件付生存関数 $S_{b_n}(t | X = X_i)$ はカーネル関数 K_{b_n} で重み付けた Kaplan-Meier 推定量である。

$$\hat{S}_{b_n}(t | X = X_i) = \prod_{s \in \{X_i: i=1 \dots n, \Delta_i=1\}, s \leq t} \left[1 - \frac{\sum_j K_{b_n}(X_i, X_j) I(T_i = s) \Delta_i}{\sum_j K_{b_n}(X_i, X_j) I(T_i = s)} \right]$$

カーネル関数 K_{b_n} は $0 \leq 2b_n \leq 1$ であるバンド幅 b_n を用いて $K_{b_n}(X_i, X_j) = I(-b_n < \hat{F}_X(X_i) - \hat{F}_X(X_j) < b_n)$ と定義される。以上の 2 つの推定量および、線形予測子 X の経験分布関数 $\hat{F}_X(c) = \sum_i I(X_i \leq c)/n$ を用いて、感度と特異度を以下のように推定する。

$$\widehat{\text{Se}}(c, t) = \frac{1 - \hat{F}_X(c) - \hat{S}_{b_n}(c, t)}{1 - \hat{S}_{b_n}(t)}$$

$$\widehat{\text{Sp}}(c, t) = 1 - \frac{\hat{S}_{b_n}(c, t)}{\hat{S}_{b_n}(t)}$$

最近傍法は単調な感度と特異度を推定し、マーカーと打ち切りの独立に対する頑健性もある手法である。また、デフォルトのバンド幅は $b_n = 0.05$ であり、NNE オプションで `method=NNE (span=0.1)` のように 0 以上 0.5 以下の任意の値を指定できる。

2.4.4 再帰法

再帰法では、 $t_1 < t_2 < \dots < t_M$ をイベント発生時間とすると、時点 $t_m (1 \leq m \leq M)$ における感度と特異度は以下のように表現される [3]。

$$\widehat{\text{Se}}(c, t_m) = \frac{\sum_{k=1}^m I(X_k > c) \hat{S}_{KM}(t_{k-1}) / r_k}{1 - \hat{S}_{KM}(t_m)}$$

$$\widehat{\text{Sp}}(c, t_m) = \frac{\Pr(X_i < c) - \sum_{k=1}^m I(X_k \leq c) \hat{S}_{KM}(t_{k-1}) / r_k}{\hat{S}_{KM}(t_m)}$$

ここで、 X_k は時点 t_k でイベントが起きた被験者の線形予測子、 r_k は時点 t_k におけるリスク集合の大きさである。再帰法が推定する感度は単調であるが、特異度は非単調となる場合がある。

3 データと評価方法

被験者 $i = 1, \dots, n$ についてベースライン時点での予後マーカー値 X_i と観測生存時間 T_i^{obs} およびイベント観測の有無 $\Delta_i = \{1, 0\}$ が得られているという状況を考える. 本稿で用いるデータセットでは, 対象者 i の (真の) 生存時間 T_i^{true} を強度パラメータ $\lambda = -\frac{\log \pi}{5}$ の指数分布 $\exp(\lambda)$ から発生させた. ここで, λ は 5 年時点での対象者の生存率が π となるように生存時間を発生させるために, 生存関数 $S(t = 5) = \exp(-\lambda \times 5) = \pi$ により得られている. 対象者 i のベースライン時点での予後マーカー値 X_i は, 生存時間 T_i^{true} と独立同一分布から得られた指数乱数 $R_i \sim \exp(\lambda)$ と, 予後マーカーと生存時間の関連の強さ d ($0 \leq d \leq 1$) を用いて, $X_i = \sqrt{d} \times T_i^{true} + \sqrt{1-d} \times R_i$ とした [9].

π と d によって時間ごとの AUC の真値が決まるため, 表 2 に示したように π と d の組み合わせによって生存時間データのシナリオを考え, それぞれの時間ごとの AUC の真値を求めた. AUC の真値は, $n = 5000$ の生存時間データの AUC を 1000 回算出し, その平均値を真値とした (真値の一覧は付録にまとめている).

表 2 生存時間データの各シナリオの設定

シナリオ	I	II	III	IV	V	VI
π	0.2	0.5	0.8	0.2	0.5	0.8
d	0.25	0.25	0.25	0.50	0.50	0.50

条件付 Kaplan-Meier 法および IPCW 法は, マーカーと打ち切りが独立という仮定を必要とする推定方法のため, 対象者 i の打ち切り時間 C_i はマーカーと独立な以下の指数分布から生成した.

$$C_i \sim \text{Exp}\left(-\frac{\log 0.80}{5}\right)$$

表 3 生存時間データの各シナリオの時点ごとの死亡率と打ち切りの割合

設定		死亡率			打ち切りの割合		
シナリオ	π	$t = 1$	$t = 3$	$t = 5$	$t = 1$	$t = 3$	$t = 5$
I, IV	0.2	27.0%	58.6%	73.8%	3.73%	8.12%	10.2%
II, V	0.5	12.7%	32.0%	45.4%	4.08%	10.3%	14.6%
III, VI	0.8	4.27%	11.7%	18.0%	4.27%	11.7%	18.0%

対象者 i の観測生存時間 T_i^{obs} は $T_i^{obs} = \min(T_i^{true}, C_i)$ となる. このとき, 対象者の死亡が観測できるのは $T_i^{obs} = T_i^{true}$ のときであり, この対象者について $\Delta_i = 1$ となる. π と d の組み合わせ (シナリオ) および症例数を変えた場合に, 各推定方法が推定する AUC の精度がどのように変化するかを二乗平均平方根誤差 (root mean squared error, 以下 RMSE) を用いて調べる. RMSE は, 各シナリオでの AUC の真値を AUC^* とすると

$$\text{RMSE} = \left(\frac{1}{1000} \sum_{k=1}^{1000} (\overline{\text{AUC}} - AUC^*)^2 \right)^{\frac{1}{2}}$$

と計算される.

4 結果と考察

表 4 各推定方法における RMSE の比較 (シナリオ I)

設定		時間	推定方法ごとの RMSE			
シナリオ	n	T	IPCW	KM	NNE	RECURSIVE
I	50	1	<u>9.00</u>	9.02	9.06	9.01
$\pi = 0.2$		3	<u>7.12</u>	<u>7.12</u>	7.21	7.16
$d = 0.25$		5	7.14	<u>7.13</u>	7.38	7.57
	100	1	6.30	<u>6.28</u>	6.33	6.30
		3	<u>4.94</u>	4.96	5.10	4.96
		5	4.86	<u>4.81</u>	4.88	5.06
	200	1	<u>4.48</u>	<u>4.48</u>	4.55	<u>4.48</u>
		3	3.40	<u>3.39</u>	3.48	3.41
		5	3.38	3.37	<u>3.34</u>	3.53

表 5 各推定方法における RMSE の比較 (シナリオ III)

設定		時間	推定方法ごとの RMSE			
シナリオ	n	T	IPCW	KM	NNE	RECURSIVE
III	50	1	22.97	<u>22.62</u>	23.25	<u>22.62</u>
$\pi = 0.8$		3	13.72	13.70	13.83	<u>13.68</u>
$d = 0.25$		5	10.99	10.89	11.04	<u>10.87</u>
	100	1	16.38	<u>16.22</u>	16.75	16.23
		3	<u>9.15</u>	9.21	9.33	9.19
		5	<u>7.78</u>	7.82	7.87	7.79
	200	1	<u>10.91</u>	10.92	11.19	<u>10.91</u>
		3	6.39	6.43	<u>6.35</u>	6.39
		5	5.52	5.53	<u>5.47</u>	5.49

π : 5 年生存率, d : 生存時間とマーカーの関連強度, n : 症例数, t : 時点 (年)

本節には、紙幅の都合上、シナリオ I とシナリオ III の結果のみを示す。残りの 4 つのシナリオに対するシミュレーションの結果は、付録を参照されたい。

どのシナリオにおいても、IPCW 法、条件付 Kaplan-Meier 法、再帰法の 3 手法は精度が大きく変わらず、また推定精度の経時変化も似ている。一方で、最近傍法はどのシナリオでも精度が悪く安定しない。

再帰法は、すべてのシナリオにおいて追跡初期 ($t = 1$) での精度に優れるが、5 年生存率の低いシナリオでは追跡後期 ($t = 5$) で推定精度が大きく落ちた。この原因は再帰法による感度と特異度の推定がリスク集合の大きさを参照しているためである (2.4.4 節参照)。すなわち、リスク集合の大きさの逆数で重み付けをするため、患者が減った追跡後期ほど一度のイベント発生による推定値の変化が大きくなり、推定精度が損なわれている。イベント発生割合が高いデータでは、再帰法の使用は推奨されないと考えられる。

5 まとめ

本稿では、5年生存率とマーカーと生存時間の関連の強さの組み合わせを変えた複数のシナリオに対し、症例数と生存時間と独立な打ち切りの頻度を変えたケースで生存時間データを生成し、条件付 Kaplan-Meier 法、IPCW 法、最近傍法、再帰法の4つの推定方法による時点ごとの AUC の推定の RMSE を算出し、性能を評価した。まず最近傍法は、打ち切りがマーカーと独立であるデータにおいて、他の推定方法に精度で劣る。ただし、最近傍法は独立な打ち切りの仮定を必要としないため、独立な打ち切りを保証できないデータにおいては有効だと考えられる。条件付 Kaplan-Meier 法、IPCW 法、再帰法はほとんどのケースにおいて近い精度で AUC を推定する。再帰法はリスク集合の大きい初期の推定で非常に高い性能を発揮したが、死亡や打ち切りによってリスク集合が小さくなるほど精度が下がる傾向が見られた。

最後に時間依存性 ROC 曲線の推定方法の使い分けについて述べる。まず、打ち切りがマーカーと独立であることが保証できない場合、最近傍法のみ用いることができる。独立が保証できている場合、リスク集合の大きい追跡初期段階での推定には再帰法が最もふさわしい。事前にどの推定方法を使用するか決めておく必要がある場合、イベント発生割合が低いデータでは再帰法が最も適している。イベントや打ち切りの発生によりリスク集合が減少する後半の時点になると再帰法の精度は条件付 Kaplan-Meier 法および IPCW 法の精度を下回るため、イベント発生割合の高いデータでは、この2つの手法のどちらかから推定法を選ぶ必要があり、その場合は単調な感度と特異度を推定する IPCW 法を使用すべきだろう。

本稿が、時間依存性 ROC 曲線を使用したデータ解析の一助となれば幸いである。

参考文献

- [1] Heagerty PJ, Lumley T, Pepe MS. Time dependent ROC Curve for censored survival data and a diagnostic marker. *Biometrics*. 2000;56(2):337-44
- [2] Uno H, Cai TX, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*. 2007;102(478):527-37
- [3] Chambless LE, Diao G. Estimation of time - dependent area under the ROC curve for long - term risk prediction. *Statistic in Medicine*. 2006;25(20):3474-86
- [4] Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology* 2017;17:53
- [5] 大橋靖雄, 浜田知久馬, 魚住龍史. 『生存時間解析[第2版] -SASによる生物統計-』. 東京大学出版会. 2022.
- [6] 下川敏雄. 時間依存性 ROC 曲線とその応用. *日本計算機統計学会大会論文集*.2013;30:207-10
- [7] 魚住龍史, 矢田真城, 篠崎智大, 川口淳, 浜田知久馬. SASによる時間依存性ROC曲線とC統計量. *SAS ユーザー総会論文集*. 2018;6-24
- [8] Akritas MG. Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*. 1994;1299-1327
- [9] Beyene KM, El Ghouch A. Smoothed time-dependent receiver operating characteristic curve for right censored survival data. *Statistics in Medicine*. 2020;1-24

付録

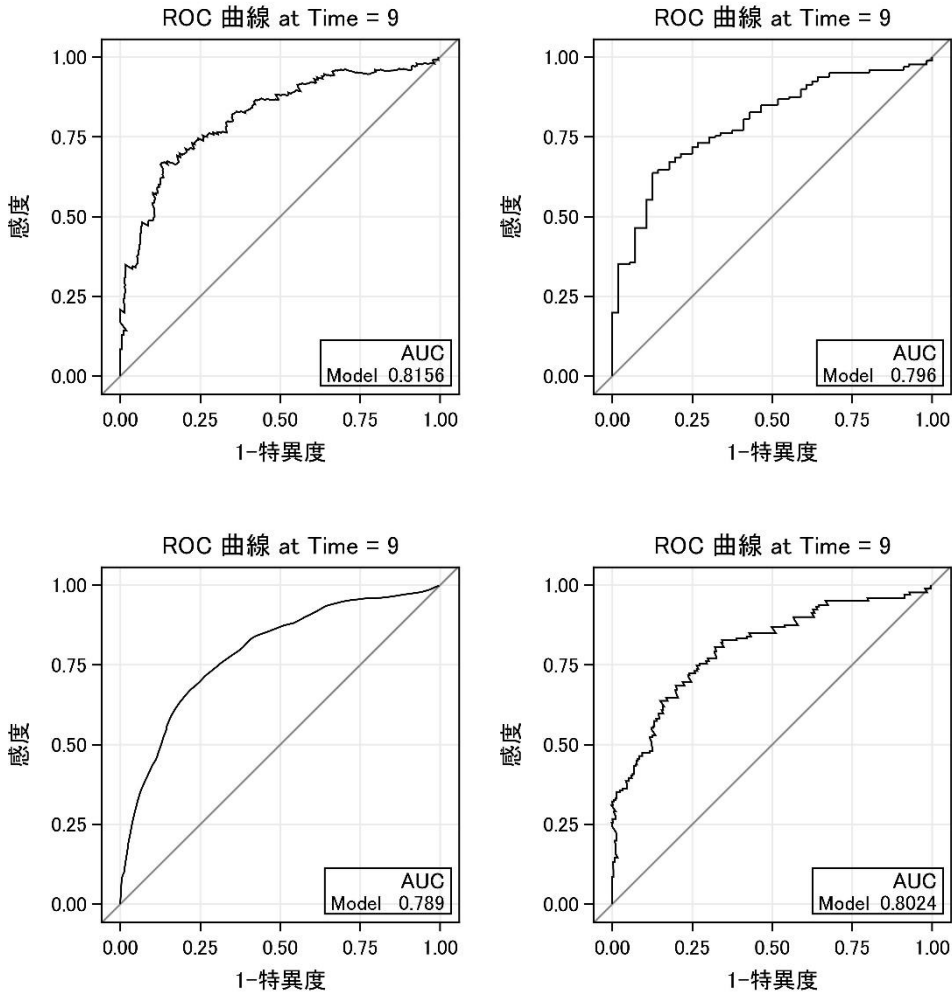


図 A 肝硬変データに対する各推定方法が推定する時間依存性 ROC 曲線
(左上 : KM, 右上 : IPCW, 左下 : NNE, 右下 : RECURSIVE)

図 A は肝硬変データに対して 9 年経過時の時間依存性 ROC 曲線を 4 手法で推定したものである。通常の ROC 曲線は範囲[0.5, 1]内に必ず収まり、点(0, 0) から点(1, 1)にかけての階段状の形である。別の言い方をすると感度と特異度はトレードオフの関係にあり、カットオフ値 c が上がると感度が下がって特異度が上がり、 c が下がると感度が上がって特異度が下がる。つまり感度と特異度は c に関して単調である。しかし時間依存性 ROC 曲線の推定の場合、推定法によってはこの感度と特異度の単調性が保証されない。条件付 Kaplan-Meier 法によって推定された感度と特異度は非単調であり、これは図 A 左上の歪な形状をみるとわかりやすい。定義に矛盾していることから、これは条件付 Kaplan-Meier 法の明確な欠点である。この欠点の改善のため、単調性を保つように感度と特異度を推定する手法として提案された最近傍法の構成する曲線（図 A 左下）をみると、他手法と異なる独特な、滑らかな形状をしつつも単調性を満たしていることがわかる。IPCW 法は単調な感度と特異度を推定し、これによって構成された時間依存性 ROC 曲線は通常の ROC 曲線と同様の階段状となる。再帰法は単調な感度を推定するが、特異度は単調ではない。つまり、カットオフ値 c が上がると感度は単調増加（非減少）するが、特異度も上がる場合がある。

表 A 生存時間データの各シナリオの時点ごとの AUC の真値 (有効数字 3 桁)

設定			AUC の真値		
シナリオ	π	d	$t = 1$	$t = 3$	$t = 5$
I	0.2	0.25	0.713	0.768	0.817
II	0.5	0.25	0.696	0.721	0.745
III	0.8	0.25	0.688	0.696	0.704
IV	0.2	0.50	0.789	0.852	0.899
V	0.5	0.50	0.767	0.799	0.827
VI	0.8	0.50	0.756	0.767	0.777

シナリオは 5 年生存率 π の高さおよびマーカーと生存時間の関連の強さ d の組み合わせによって 6 つに分けた。各シナリオにおいて、症例数 $n = 5000$ のデータを 1000 回生成し、それぞれのデータに対して 1,3,5 年経過時の時間依存性 ROC 曲線の AUC を計算、その平均を真値とした。表 A をみると、5 年生存率が低いほど (すなわちイベントが多く発生するほど)、マーカーと生存時間の関連が強くなるほど AUC が大きくなるのがわかる。時間が経過するほど AUC が大きくなるのは、累積のイベント発生数が大きくなるためと考えられる。

プログラム A AUC の真値計算用の生存時間データ生成コード

```
proc phreg data = surv plots(overlay=individual)=(roc) rocoptions(method = KM auc at=(1,3,5));
  model Time = marker;
run;
```

時間依存性 ROC 曲線の推定は、PHREG プロシジャの rocoptions() によって行い、() 内のサブオプションで推定に関する指定ができる。plots=auc とすれば、c 統計量 (曲線化面積) の時間変化をプロットできる。

表 B 各推定方法における RMSE の比較 (シナリオ II)

設定		時間	推定方法ごとの RMSE			
シナリオ	n	t	IPCW	KM	NNE	RECURSIVE
II	50	1	12.88	12.90	13.13	<u>12.86</u>
$\pi = 0.5$		3	8.44	8.41	<u>8.37</u>	8.40
$d = 0.25$		5	7.67	<u>7.65</u>	7.69	<u>7.65</u>
	100	1	<u>8.81</u>	8.83	8.94	<u>8.81</u>
		3	6.04	<u>6.03</u>	6.15	6.04
		5	<u>5.34</u>	<u>5.34</u>	5.54	5.37
	200	1	6.15	6.17	<u>6.13</u>	6.14
		3	4.28	<u>4.27</u>	4.39	4.28
		5	3.78	<u>3.77</u>	3.91	3.78

表 C 各推定方法における RMSE の比較 (シナリオ IV)

設定		時間	推定方法ごとの RMSE			
シナリオ	n	t	IPCW	KM	NNE	RECURSIVE
IV	50	1	<u>7.81</u>	7.83	7.83	<u>7.81</u>
$\pi = 0.2$		3	<u>5.71</u>	5.77	5.92	5.76
$d = 0.50$		5	<u>5.00</u>	5.15	5.39	5.59
	100	1	5.47	<u>5.45</u>	5.62	5.46
		3	<u>3.93</u>	3.99	4.09	3.96
		5	<u>3.45</u>	3.50	3.66	3.71
	200	1	<u>3.90</u>	3.91	4.06	<u>3.90</u>
		3	<u>2.73</u>	2.75	2.80	2.75
		5	<u>2.43</u>	2.48	2.47	2.64

表 D 各推定方法における RMSE の比較 (シナリオ V)

設定		時間	推定方法ごとの RMSE			
シナリオ	n	t	IPCW	KM	NNE	RECURSIVE
V	50	1	11.33	11.35	11.35	<u>11.31</u>
$\pi = 0.5$		3	7.23	7.26	7.46	<u>7.20</u>
$d = 0.50$		5	<u>6.38</u>	6.45	6.75	6.39
	100	1	7.83	7.85	7.89	<u>7.82</u>
		3	<u>5.17</u>	5.19	5.42	5.19
		5	<u>4.45</u>	4.51	4.71	4.47
	200	1	5.47	5.50	<u>5.43</u>	5.47
		3	<u>3.68</u>	<u>3.68</u>	3.88	3.70
		5	<u>3.18</u>	3.21	3.29	3.20

表 E 各推定方法における RMSE の比較 (シナリオ VI)

設定		時間	推定方法ごとの RMSE			
シナリオ	n	t	IPCW	KM	NNE	RECURSIVE
VI	50	1	20.32	<u>19.81</u>	20.45	19.82
$\pi = 0.8$		3	<u>12.03</u>	12.09	<u>12.03</u>	<u>12.03</u>
$d = 0.50$		5	9.53	9.54	<u>9.37</u>	9.44
	100	1	14.68	<u>14.67</u>	14.98	<u>14.67</u>
		3	<u>8.14</u>	8.20	8.22	8.22
		5	6.85	6.91	6.92	<u>6.84</u>
	200	1	9.72	9.73	9.87	<u>9.71</u>
		3	5.69	5.74	<u>5.62</u>	5.68
		5	4.86	4.90	<u>4.82</u>	4.84

π : 5 年生存率, d : 生存時間とマーカーの関連強度, n : 症例数, t : 時点 (年)

データステップを用いた統計検定の再現

○折村奈美

(イーピーエス株式会社)

Reproducing statistical tests using SAS data step

Nami Orimura

EPS Corporation

要旨

SAS プロシジャは検定を行ううえで非常に便利な機能である。複雑な検定を行う際も、SAS プロシジャを使用することで即時に結果を得ることができる。その反面、構文さえ理解すれば検定を実施することが可能なため、特に初めて行う検定は並行して学習を進めることが非常に重要である。検定の内容を理解しないままにすると、アウトプットの意味を正しく解釈できない可能性や、SAS プロシジャのみで結果を得ることができない解析に対応できなくなる可能性がある。そこで本稿では、データステップでの検定の再現過程を紹介し、特に初学者向けに SAS を用いた検定の学習の重要性を伝えることを目的とした。

キーワード: SAS プロシジャ, 検定

1. 緒言

SAS プロシジャは特定の処理を実現するための構文のまとまりで、機能ごとに大きく分けて解析、データの加工・変換、データセットの管理、その他の 4 つに分類される。特に解析の場面では、平均や標準偏差などの簡易な記述統計から、モデル解析や生存時間解析、多変量解析などの複雑な計算を必要とするものまで、さまざまな解析をシンプルなコードで実行することができる。また、各プロシジャにおいてオプション機能が豊富なため、多岐にわたる解析に対応できる特徴もある。解析業務でも頻繁に SAS プロシジャを使用するが、便利な反面、構文さえ理解すれば解析が実行できてしまうため、プログラマーが解析の本質を理解しないまま使い続けてしまうことも十分あり得る。このような状況になると、複雑な前処理を伴う解析や SAS プロシジャのオプションが未実装の解析に直面した際に、対応できない可能性が高くなる。そのため、特に初めて行う解析についてはプログラムだけでなく統計的な理論も併せて学習することが非常に重要である。

解析の学習にはさまざまな方法があり、本稿ではその一つとして SAS を用いたデータステップでの検定の再現を紹介する。

2. 検定の再現

本稿では原則的にプロシジャを用いずに統計処理を再現するが、以下のデータ操作系プロシジャは解析の流れの説明を平易にするため、例外的に使用した。

sort, transpose, summary, sql, rank

2.1. Student の t 検定

Student の t 検定は、対応のない 2 群間の平均値差を検定するパラメトリックな手法である。sashelp ライブラリに格納されている class データセットを使用して、Student の t 検定により性別間の体重を比較することとした。まず ttest プロシジャを用いた以下のコードを実行すると、Table. 1 の結果が得られた。Student の t 検定は等分散性を仮定した検定であるので、対応する出力結果である Pooled の行に着目すると自由度が 17, t 値が-1.93, p 値が 0.0702 である。この検定結果をデータステップで再現した。

```
proc ttest data = sashelp.class;
  class sex;
  var weight;
run;
```

Table. 1 ttest プロシジャの実行結果

手法	分散	自由度	t 値	Pr > t
Pooled	Equal	17	-1.93	0.0702
Satterthwaite	Unequal	16.962	-1.95	0.0680

等分散性				
手法	分子の自由度	分母の自由度	F 値	Pr > F
Folded F	9	8	1.37	0.6645

- ▶ 帰無仮説 H_0 : 2 群間の母平均は等しい
- ▶ 対立仮説 H_1 : 2 群間の母平均は等しくない
- ▶ 自由度: $n_1 + n_2 - 2$
- ▶ 公式

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

まず summary プロシジャを用いて、 S^2 (プールした分散) と t 値の計算に必要な性別ごとの n, 平均, 不偏分散を算出し、女子を”F”, 男子を”M”として _sex 変数を作成した (Fig. 1a). 得られたデータセット wk1 の形では公式に当てはめて計算ができないため, transpose プロシジャを用いて値を横並びにした wk2 を作成した (Fig. 1b).

```
proc summary data = sashelp.class nway;
  var weight;
  class sex;
  output out = _wk1 n = mean = var = / autoname;
run;

data wk1;
  set _wk1;
  if sex = “女子” then _sex = “F”;
  else if sex = “男子” then _sex = “M”;
run;
```

```

proc transpose data = wk1 out = n (drop = _) prefix = n_;
  var Weight_N;
  id _sex;
run;

proc transpose data = wk1 out = mean (drop = _) prefix = mean_;
  var Weight_Mean;
  id _sex;
run;

proc transpose data = wk1 out = var (drop = _) prefix = var_;
  var Weight_Var;
  id _sex;
run;

data wk2;
  set n;
  set mean;
  set var;
run;

```

(a)

VIEWTABLE: Work.Wk1						
	Sex	_TYPE_	_FREQ_	Weight_N	Weight_Mean	Weight_Var
1	女子	1	9	90.111111111	375.736111111	F
2	男子	1	10	108.95	516.525	M

(b)

VIEWTABLE: Work.Wk2						
	n_F	n_M	mean_F	mean_M	var_F	var_M
1	9	10	90.111111111	108.95	375.736111111	516.525

Fig. 1 Student の t 検定の再現過程

wk2 の値を公式に当てはめて計算すると、自由度が 17, t 値が-1.93...となった。さらに `probt` 関数を用いて自由度 17 の t 分布における t 値-1.93...以下となる確率を求め、両側検定であるため 2 倍して p 値を計算すると 0.0701...といった結果が得られた (Fig. 2)。これらの値は Table. 1 に示した `ttest` プロシジャを用いた検定結果と同様であることから、データステップで Student の t 検定を再現することができた。

```

data wk3;
  set wk2;

/*自由度*/
  df = n_f + n_m - 2;

/*S^2*/
  s2 = divide (((n_f - 1) * var_f + (n_m - 1) * var_m), df);

/*t 値*/
  t = divide ((mean_f - mean_m), sqrt (s2 * (1 / n_f + 1 / n_m)));

/*p 値*/
  if t < 0 then p = 2 * probt (t, df);
  else p = 2 * (1 - probt (t, df));
run;

```

VIEWTABLE: Work.Wk3				
	df	s2	t	p
1	17	450.27140523	-1.932246615	0.0701687815

Fig. 2 Student の t 検定の再現結果

2.2. カイ二乗検定（独立性の検定）

独立性の検定は、クロス集計表における 2 変数の集計結果の数値に関連があるか否かを検定する手法である。クロス集計表のテストデータ data1 を作成し、TREAT と RESP が独立であるか否かを検定した。freq プロシジャの tables ステートメントで CHISQ オプションを指定してカイ二乗検定を行うと、自由度が 1、カイ二乗値が 1.0256、p 値が 0.3112 といった結果が得られた (Table. 3)。この検定結果をデータステップで再現した。

```
data data1;
  TREAT = "N"; RESP = "N"; _FREQ_ = 5; output;
  TREAT = "N"; RESP = "Y"; _FREQ_ = 15; output;
  TREAT = "Y"; RESP = "N"; _FREQ_ = 8; output;
  TREAT = "Y"; RESP = "Y"; _FREQ_ = 12; output;
run;

proc freq data = data1;
  tables TREAT * RESP / CHISQ;
  weight _FREQ_ / zeros;
run;
```

Table. 2 テストデータ

度数 パーセント 行のパーセント 列のパーセント	表 : TREAT * RESP			
	TREAT	RESP		合計
		N	Y	
N	5 12.50 25.00 38.46	15 37.50 75.00 55.56	20 50.00	①
Y	8 20.00 40.00 61.54	12 30.00 60.00 44.44	20 50.00	
合計	13 32.50	27 67.50	40 100.00	② ③

Table. 3 freq プロシジャの実行結果
TREAT * RESP の統計量

統計量	自由度	値	p 値
カイ 2 乗値	1	1.0256	0.3112
尤度比カイ 2 乗値	1	1.0326	0.3095
連続性補正カイ 2 乗値	1	0.4558	0.4996
Mantel-Haenszel のカイ 2 乗値	1	1.0000	0.3173
ファイ係数		-0.1601	
一致係数		0.1581	
Cramer の V 統計量		-0.1601	

- ▶ 帰無仮説 H_0 : A と B は独立である
- ▶ 対立仮説 H_1 : A と B は独立ではない
- ▶ 公式

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(x_{ij} - E_{ij})^2}{E_{ij}}$$

自由度 $(m - 1) \times (n - 1)$ のカイ二乗分布に従う

x_{ij} : 観測度数, E_{ij} : 期待度数

m : 行数, n : 列数

まず Table. 2 の合計部分のデータを持つデータセット wk2 を作成した。summary プロシジャを用いて①, ②, ③で示した箇所をそれぞれ計算し、テストデータ data1 に結合して wk2 とした (Fig. 3a)。これで Table. 2 のクロス集計表の情報を持つデータセットを作成できたので、この wk2 を用いて以降の計算を行った。

次に、自由度を求めるためにクロス集計表の行数および列数を求めた。sort プロシジャと nodupkey オプションを用いて、wk2 の TREAT, RESP それぞれの文字列の重複レコードを除いたデータセット treat, resp を作成した。SQL を用いてこれらのデータセットのオブザベーション数をカウントして、マクロ変数 treat_n, resp_n に値を格納した (Fig. 3b)。これらの値がクロス集計表の行数および列数である。

①

```
proc summary data = data1 nway;  
  var _FREQ_ ;  
  class TREAT;  
  output out = wk1_treat (drop = _FREQ_) sum = total_treat;  
run;
```

②

```
proc summary data = data1 nway;  
  var _FREQ_ ;  
  class RESP;  
  output out = wk1_resp (drop = _FREQ_) sum = total_resp;  
run;
```

③

```
proc summary data = data1 nway;  
  var _FREQ_ ;  
  output out = _wk1_total (keep = total) sum = total;  
run;
```

```
data wk1_total;  
  set _wk1_total;  
  _TYPE_ = 1;  
run;
```

```
data wk2_treat;  
  merge data1 wk1_treat;  
  by TREAT;  
run;
```

```
proc sort data = wk2_treat;  
  by RESP;  
run;
```

```
data wk2_resp;  
  merge wk2_treat wk1_resp;  
  by RESP;  
run;
```

```
data wk2 (drop = _TYPE_);  
  merge wk2_resp wk1_total;  
  by _TYPE_;  
run;
```

```
proc sort data = wk2;  
  by TREAT RESP;  
run;
```

/*行数, 列数*/

```
proc sort data = wk2 out = treat nodupkey;  
  by TREAT;  
run;
```

```
proc sort data = wk2 out = resp nodupkey;  
  by RESP;  
run;
```

```
proc sql;
```

```

select count (*) into: treat_n trimmed from treat;
select count (*) into: resp_n trimmed from resp;
quit;
%put & = treat_n & = resp_n;

```

(a)

	TREAT	RESP	_FREQ_	total_treat	total_resp	total
1	N	N	5	20	13	40
2	N	Y	15	20	27	40
3	Y	N	8	20	13	40
4	Y	Y	12	20	27	40

(b)

```

692 %put &=treat_n &=resp_n;
TREAT_N=2 RESP_N=2
...

```

Fig.3 カイ二乗検定のテストデータ

次に自由度、期待度数、観測度数と期待度数との差を求めカイ二乗値を計算した。自由度 df はマクロ変数 $treat_n$, $resp_n$ を用いて計算すると 1 となった。期待度数 ef は TREAT の合計 ($total_treat$) \times RESP の合計 ($total_resp$) \div 全体の合計 ($total$)で、観測度数と期待度数との差 $diff$ は $_FREQ_ - ef$ で計算した。そして求めた ef を公式の E_{ij} に、 $diff$ を $x_{ij} - E_{ij}$ に代入して公式の $\frac{(x_{ij} - E_{ij})^2}{E_{ij}}$ 部分の計算を行うと、 $_x2$ の値が得られた (Fig. 4a)。カイ二乗値はこの $_x2$ の合計値であり、summary プロシジャを用いて算出すると 1.0256...となった。さらに、probchi 関数を用いて自由度 1 のカイ二乗分布におけるカイ二乗値 1.0256...以下となる確率を求め、1 から引いて p 値を計算すると 0.3111...といった結果が得られた (Fig. 4b)。これらの値は Table. 3 に示した freq プロシジャでの検定結果と同様であり、データステップでカイ二乗検定を再現することができた。

```

data wk3;
  set wk2;

/*自由度*/
  df = (&treat_n. - 1) * (&resp_n - 1);

/*期待度数*/
  ef = divide ((total_treat * total_resp), total);

/*観測度数と期待度数の差*/
  diff = _FREQ_ - ef;

/*χ2値*/
  _x2 = divide (diff ** 2, ef);
run;

proc summary data = wk3 nway;
  var _x2;
  output out = wk4 sum = x2;
run;

data wk5;
  set wk4;
  set wk3 (keep = df);

/*p 値*/
  prob = 1 - probchi (x2, df);
run;

```

(a)

	TREAT	RESP	_FREQ_	total_treat	total_resp	total	df	ef	diff	_x2
1	N	N	5	20	13	40	1	6.5	-1.5	0.34615
2	N	Y	15	20	27	40	1	13.5	1.5	0.16667
3	Y	N	8	20	13	40	1	6.5	1.5	0.34615
4	Y	Y	12	20	27	40	1	13.5	-1.5	0.16667

(b)

	_x2	df	prob
1	1.0256410256	1	0.3111846688

Fig.4 カイ二乗検定の再現結果

2.3. Wilcoxon の符号付順位検定

Wilcoxon の符号付順位検定は、対応のある 2 群間の中央値の差を検定するノンパラメトリック手法である。対応のあるテストデータ data2 を作成し、BEFORE と AFTER の値に差があるか否か Wilcoxon の符号付順位検定で確かめることとした。univariate プロシジャを用いて検定を行うと、統計量 S が-85、p 値が 0.0497 という結果が得られた (Table. 4)。以下に示した公式を使い、Wilcoxon の符号付順位検定の結果を再現した。

data data2;

length SUBJID \$200 BEFORE AFTER DIFF 8;

SUBJID = "A001"; BEFORE = 0.8; AFTER = 0.6; DIFF = -0.2; output;
 SUBJID = "A002"; BEFORE = 0.7; AFTER = 0.4; DIFF = -0.3; output;
 SUBJID = "A003"; BEFORE = 0.8; AFTER = 0.5; DIFF = -0.3; output;
 SUBJID = "A004"; BEFORE = 1.1; AFTER = 0.6; DIFF = -0.5; output;
 SUBJID = "A005"; BEFORE = 0.8; AFTER = 0.4; DIFF = -0.4; output;
 SUBJID = "A006"; BEFORE = 0.6; AFTER = 0.5; DIFF = -0.1; output;
 SUBJID = "A007"; BEFORE = 1.1; AFTER = 0.9; DIFF = -0.2; output;
 SUBJID = "A008"; BEFORE = 0.7; AFTER = 0.3; DIFF = -0.4; output;
 SUBJID = "A009"; BEFORE = 0.5; AFTER = 0.7; DIFF = 0.2; output;
 SUBJID = "A010"; BEFORE = 0.6; AFTER = 0.6; DIFF = 0; output;

SUBJID = "A011"; BEFORE = 0.7; AFTER = 0.5; DIFF = -0.2; output;
 SUBJID = "A012"; BEFORE = 0.8; AFTER = 0.2; DIFF = -0.6; output;
 SUBJID = "A013"; BEFORE = 0.4; AFTER = 0.7; DIFF = 0.3; output;
 SUBJID = "A014"; BEFORE = 0.5; AFTER = 0.3; DIFF = -0.2; output;
 SUBJID = "A015"; BEFORE = 0.7; AFTER = 0.6; DIFF = -0.1; output;
 SUBJID = "A016"; BEFORE = 1.2; AFTER = 0.7; DIFF = -0.5; output;
 SUBJID = "A017"; BEFORE = 0.3; AFTER = 0.5; DIFF = 0.2; output;
 SUBJID = "A018"; BEFORE = 0.8; AFTER = 0.5; DIFF = -0.3; output;
 SUBJID = "A019"; BEFORE = 0.9; AFTER = 0.8; DIFF = -0.1; output;
 SUBJID = "A020"; BEFORE = 0.6; AFTER = 0.9; DIFF = 0.3; output;

SUBJID = "A021"; BEFORE = 0.8; AFTER = 0.5; DIFF = -0.3; output;
 SUBJID = "A022"; BEFORE = 0.7; AFTER = 0.2; DIFF = -0.5; output;
 SUBJID = "A023"; BEFORE = 0.4; AFTER = 0.5; DIFF = 0.1; output;
 SUBJID = "A024"; BEFORE = 1.4; AFTER = 0.4; DIFF = -1.0; output;
 SUBJID = "A025"; BEFORE = 0.8; AFTER = 0.8; DIFF = 0; output;
 SUBJID = "A026"; BEFORE = 0.4; AFTER = 0.7; DIFF = 0.3; output;
 SUBJID = "A027"; BEFORE = 1.1; AFTER = 1.5; DIFF = 0.4; output;
 SUBJID = "A028"; BEFORE = 0.4; AFTER = 0.3; DIFF = -0.1; output;
 SUBJID = "A029"; BEFORE = 0.5; AFTER = 0.7; DIFF = 0.2; output;
 SUBJID = "A030"; BEFORE = 0.3; AFTER = 0.6; DIFF = 0.3; output;

run;

```
proc univariate data = data2;
  var DIFF;
run;
```

Table. 4 univariate プロシジャの実行結果

位置の検定 H0: Mu0=0				
検定	統計量			p 値
Student の t 検定	t	-2.23126	Pr > t	0.0336
符号検定	M	-5	Pr >= M	0.0872
符号付順位検定	S	-85	Pr >= S	0.0497

- ▶ 帰無仮説 H₀: 2 群の母集団に差がない
- ▶ 対立仮説 H₁: 2 群の母集団に差がある
- ▶ 公式

$$S = R^+ - \frac{n_t(n_t + 1)}{4}$$

R⁺: 正の符号付順位の合計, n_t: n

*DIFF=0 のデータは除く

まず差の符号付順位を求めた. DIFF の絶対値をとって DIFF_abs として, この値を用いて rank プロシジャで順位付けを行った. このとき DIFF=0 のレコードは使用しないので除いて, ties=mean オプションで DIFF の絶対値が同じ値のときに順位の平均値が振られるようにした. そして求めた順位に DIFF の符号を付けて signed_rank とした (Fig. 5).

```
data wk1;
  set data2;
  DIFF_abs = abs (DIFF);
run;

proc rank data = wk1 out = wk2 ties = mean;
  where DIFF_abs ^= 0;
  var DIFF_abs;
  RANKS rank;
run;

data wk3;
  set wk2;
  if 0 < DIFF then signed_rank = rank;
  else if DIFF < 0 then signed_rank = -rank;
run;
```

	SUBJID	BEFORE	AFTER	DIFF	DIFF_abs	rank	signed_rank
1	A001	0.8	0.6	-0.2	0.2	9	-9
2	A002	0.7	0.4	-0.3	0.3	16.5	-16.5
3	A003	0.8	0.5	-0.3	0.3	16.5	-16.5
4	A004	1.1	0.6	-0.5	0.5	25	-25
5	A005	0.8	0.4	-0.4	0.4	22	-22
6	A006	0.6	0.5	-0.1	0.1	3	-3
7	A007	1.1	0.9	-0.2	0.2	9	-9
8	A008	0.7	0.3	-0.4	0.4	22	-22
9	A009	0.5	0.7	0.2	0.2	9	9
10	A011	0.7	0.5	-0.2	0.2	9	-9
11	A012	0.8	0.2	-0.6	0.6	27	-27
12	A013	0.4	0.7	0.3	0.3	16.5	16.5
13	A014	0.5	0.3	-0.2	0.2	9	-9
14	A015	0.7	0.6	-0.1	0.1	3	-3
15	A016	1.2	0.7	-0.5	0.5	25	-25
16	A017	0.3	0.5	0.2	0.2	9	9
17	A018	0.8	0.5	-0.3	0.3	16.5	-16.5
18	A019	0.9	0.8	-0.1	0.1	3	-3
19	A020	0.6	0.9	0.3	0.3	16.5	16.5
20	A021	0.8	0.5	-0.3	0.3	16.5	-16.5
21	A022	0.7	0.2	-0.5	0.5	25	-25
22	A023	0.4	0.5	0.1	0.1	3	3
23	A024	1.4	0.4	-1	1	28	-28
24	A026	0.4	0.7	0.3	0.3	16.5	16.5
25	A027	1.1	1.5	0.4	0.4	22	22
26	A028	0.4	0.3	-0.1	0.1	3	-3
27	A029	0.5	0.7	0.2	0.2	9	9
28	A030	0.3	0.6	0.3	0.3	16.5	16.5

Fig. 5 Wilcoxon の符号付順位検定の再現過程

次に n および R^+ を計算し、公式に当てはめて S 値を算出した。summary プロシジャを用いて求めると n は 28 となり、 R^+ は where ステートメントで正の値を持つレコードのみを絞って合計を求めると 118 となった。これらを公式に当てはめて計算した結果 S 値が -85 となり、univariate プロシジャを用いて計算した結果と同様の値が得られた (Fig. 6)。

```

/*n*/
proc summary data = wk3 nway;
  var signed_rank;
  output out = wk4_n n = / autoname;
run;

/*R+*/
proc summary data = wk3 nway;
  where 0 < signed_rank;
  var signed_rank;
  output out = wk4_sum sum = / autoname;
run;

data wk4;
  set wk4_n;
  set wk4_sum;

/*S 値*/
S = signed_rank_Sum - divide (signed_rank_N * (signed_rank_N + 1), 4);
run;

```

	signed_rank_N	signed_rank_Sum	S
1	28	118	-85

Fig. 6 Wilcoxon 符号付順位和検定の再現結果

$N \geq 25$ のときに以下に示した式で計算された統計量 Z が標準正規分布に従うことを利用して、 p 値を求めた⁹⁾。まず式中の T を求めるために、 T_+ と T_- を計算した。それぞれ符号付順位が正の合計順位、負の合計順位であり、 T_+ が 118、 T_- が 288 となった。このうち値の小さいほうを T とするので、この場合 T は 118 である。

Z 値の公式の分母を v 、分子を e としてそれぞれ計算し Z 値を算出すると、-1.93... となった。標準正規分布の確率を返す probnorm 関数により p 値を求めると 0.0529... となり、univariate プロシジャを用いて求めた値と一致しなかった (Fig. 7)。そこで sashelp を参照すると、Wilcoxon の符号付順位検定の p 値の算出には標準正規分布でなく t 分布を用いると記載があったため、 t 分布での近似を用いて p 値の計算を行った。

```

/*T+*/
proc summary data = wk3 nway;
  where 0 < signed_rank;
  var rank;
  output out = wk5_plus sum = rank_plus;
run;

/*T-*/
proc summary data = wk3 nway;
  where signed_rank < 0;
  var rank;
  output out = wk5_minus sum = rank_minus;
run;

/*p 値*/

```

```

data wk6;
  set wk5_plus;
  set wk5_minus;
  set wk4 (keep = signed_rank_n);

  T = min (rank_plus, rank_minus);
  v = sqrt (divide (signed_rank_n * (signed_rank_n + 1) * (2 * signed_rank_n + 1), 24));
  e = T - divide (signed_rank_n * (signed_rank_n + 1), 4);
  z = divide (e, v);
  prob = probnorm (z) * 2;
run;

```

T_+ : 符号付順位が正の合計順位

T_- : 符号付順位が負の合計順位

$T = \min (T_+, T_-)$

右の式で標準化して p 値を求める

$$Z = \frac{T - \frac{n_t(n_t + 1)}{4}}{\sqrt{\frac{n_t(n_t + 1)(2n_t + 1)}{24}}}$$

	signed_rank_N	T	v	e	z	prob
1	28	118	43.314690025	-85	-1.935570989	0.0529202664

Fig. 7 Wilcoxon の符号付順位検定の再現結果

$n \geq 20$ のときに以下の式で示した t 値が自由度 $n-1$ の t 分布で近似されることを用いて p 値を算出した⁷⁾. $\text{Var}(S)$ (S 値の分散) を求めるために用いる n は $\text{DIFF}=0$ を含む n , d_0 は $\text{DIFF}=0$ の n , d_i は $\text{DIFF} \neq 0$ の n で、それぞれ $n=30$, $d_0=2$, $d_i=28$ である. これらの値を公式に代入して $\text{Var}(S)$ を求めると 1905.75 となり、この値を用いて t 値を計算すると -2.04... といった値が得られた. probt 関数を用いて自由度 29 の t 分布における t 値 -2.04... 以下の確率を求め、両側検定であるため 2 倍して p 値を計算すると 0.0496... となった (Fig. 8). この値は Table. 4 に示した univariate プロシジャでの検定結果と同様であり、データステップで Wilcoxon の符号付順位検定の結果を再現することができた.

Wilcoxon の符号付順位検定の p 値の算出方法を調べると、標準正規分布による近似を用いて計算されている例が多くみられたが、データステップで検定を再現してみることで SAS では t 分布による近似で計算されていることを知った.

```

/*n*/
proc summary data = data2 nway;
  var DIFF;
  output out = wk_n n = n;
run;

/*d0*/
proc summary data = data2 nway;
  where DIFF = 0;
  var DIFF;
  output out = wk_d0 n = d0;
run;

data wk7;
  set wk4; set wk_n; set wk_d0;
/*Var(S)*/
  var = divide (n * (n + 1) * (2 * n + 1) - d0 * (d0 + 1) * (2 * d0 + 1)
    - divide (signed_rank_n * (signed_rank_n + 1) * (signed_rank_n - 1), 2), 24);

```

```

/*t 値*/
_t = sqrt (divide (n * var - S**2, n - 1));
t = divide (S, _t);

/*自由度*/
df = n - 1;

/*p 値*/
prob = probt (t, df) * 2;
run;

```

$$t = \frac{S}{\sqrt{\frac{n \times \text{Var}(S) - S^2}{n - 1}}}$$

$$\text{Var}(S) = \frac{n(n + 1)(2n + 1) - d_0(d_0 + 1)(2d_0 + 1) - \frac{1}{2} \sum_{i>0} d_i(d_i + 1)(d_i - 1)}{24}$$

t が自由度 n-1 の t 分布で近似される

n: DIFF=0 を含む n
d₀: DIFF=0 の n
d_i: DIFF≠0 の n

	signed_rank_N	signed_rank_Sum	S	n	d0
1	28	118	-85	30	2

var	_t	t	df	prob
1905.75	41.500934763	-2.048146638	29	0.0496950674

Fig. 8 Wilcoxon の符号付順位検定の再現結果

2.4. 単回帰分析

単回帰分析は目的変数 y を説明変数 x によって予測する手法であり、回帰式は $y = ax + b$ で表される。sashelp ライブラリに格納されている baseball データセットを用いて、ヒット数から給料を予測することとした。ヒット数を説明変数、給料を目的変数として reg プロシジャを用いた以下のコードを実行すると、Table. 5 の結果が得られ、推定値を $y = ax + b$ の回帰式に当てはめると $y = 0.00995x + 4.84$ となった。さらに、データに対する回帰式の当てはまりの度合いを示す決定係数 R^2 は 0.2424 であった。以下に示した公式を使い、これらの結果をデータステップで再現した。

```

proc reg data = sashelp.baseball;
  model logSalary = nHits;
run;

```

Table. 5 reg プロシジャの実行結果

読み込んだオブザベーション数	322
使用されたオブザベーション数	263
欠損値を含むオブザベーション数	59

分散分析					
要因	自由度	平方和	平均平方	F 値	Pr > F
Model	1	50.21141	50.21141	83.50	<.0001
Error	261	156.94232	0.60131		
Corrected Total	262	207.15373			

Root MSE	0.77544	R2 乗	0.2424
従属変数の平均	5.92722	調整済み R2 乗	0.2395
変動係数	13.08274		

パラメータの推定						
変数	ラベル	自由度	パラメータ推定値	標準誤差	t 値	Pr > t
Intercept	Intercept	1	4.84098	0.12813	37.78	<.0001
nHits	Hits in 1986	1	0.00995	0.00109	9.14	<.0001

▶ 公式

最小二乗法

$y = ax + b$ の回帰式

$$a = \frac{S_{xy}}{S_x^2} \quad b = \bar{y} - \frac{S_{xy}}{S_x^2} \bar{x}$$

$$S_{xy} = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{n - 1}$$

\bar{x} : x の平均, \bar{y} : y の平均

S_x^2 : x の分散, S_{xy} : x と y の共分散

決定係数

$$R^2 = \frac{Q - S_e}{Q}$$

$$Q = \sum_{k=1}^n (y_k - \bar{y})^2 \quad S_e = \sum_{k=1}^n (y_k - \hat{y})^2$$

\hat{y} : 回帰式から求めた予測値

まず sashelp ライブラリの baseball データセットから logSalary が欠測値のレコードを除き, merge のためのキーとなる _TYPE_ 変数を作成し, データセット baseball_base とした. 次に x と y の n , 平均, 分散を算出し (Fig. 9a), 求めた値を baseball_base に merge して共分散の公式の $(x_k - \bar{x})(y_k - \bar{y})$ 部分を計算した (Fig. 9b). summary プロシジャを用いて _sxy の合計を求め, 共分散を計算すると 19.259... となった. これらの値を最小二乗法の公式に当てはめて計算すると $a=0.00995...$, $b=4.84...$ となり, reg プロシジャを用いて求めた回帰式と同じ $y = 0.00995x + 4.84$ の式が得られた (Fig. 9c).

```
data baseball_base;
  set sashelp.baseball;
  where ^missing (logSalary);
  _TYPE_ = 0;
keep Name Team logSalary nHits _TYPE_;
run;

/*x, yのn, 平均, 分散*/
proc summary data = baseball_base nway;
  var logSalary nHits;
  output out = wk1 n = mean = var = / autoname;
run;

data wk2;
  merge baseball_base wk1;
  by _TYPE_;
run;

data wk3;
  set wk2;
  diff_x = nHits - nHits_mean;
  diff_y = logSalary - logSalary_mean;
  _sxy = diff_x * diff_y;
run;
```

```

proc summary data = wk3 nway;
  var _sxy;
  output out = wk4 sum = sum_sxy;
run;

data wk5;
  set wk4;
  set wk1;

/*共分散*/
  sxy = divide (sum_sxy, (logSalary_n - 1));

/*a, b*/
  a = divide (sxy, nHits_var);
  b = logSalary_mean - a * nHits_mean;
run;

```

(a)

VIEWTABLE: Work.Wk1							
	TYPE	logSalary_N	nHits_N	logSalary_Mean	nHits_Mean	logSalary_Var	nHits_Var
1	0	263	263	5.9272215412	109.15969582	0.7906631036	1935.4095144

(b)

VIEWTABLE: Work.Wk3									
	Name	Team	nHits	logSalary	logSalary_Mean	nHits_Mean	diff_x	diff_y	sxy
1	Ashby, Alan	Houston	81	6.163314804	5.9272215412	109.15969582	-28.15969582	0.2360932628	-6.648314465
2	Davis, Alan	Seattle	130	6.1737861039	5.9272215412	109.15969582	20.840304183	0.2465645627	5.1384804869
3	Dawson, Andre	Montreal	141	6.2146080984	5.9272215412	109.15969582	31.840304183	0.2873885572	9.1504753992
4	Galarraza, Andres	Montreal	87	4.5163389723	5.9272215412	109.15969582	-22.15969582	-1.410882569	31.264728562
5	Griffin, Alfredo	Oakland	169	6.6200732065	5.9272215412	109.15969582	59.840304183	0.6928516653	41.460454405
6	Newman, Al	Montreal	37	4.248495242	5.9272215412	109.15969582	-72.15969582	-1.678726299	121.13637911
7	Salazar, Argenis	Kansas City	73	4.605170186	5.9272215412	109.15969582	-36.15969582	-1.322051355	47.80497486
8	Thomas, Andres	Atlanta	81	4.3174881135	5.9272215412	109.15969582	-28.15969582	-1.609733428	45.329603671
9	Thornton, Andre	Cleveland	92	7.0030654588	5.9272215412	109.15969582	-17.15969582	1.0758439176	-18.46115437
10	Trammell, Alan	Detroit	159	6.248319432	5.9272215412	109.15969582	49.840304183	0.3210978908	16.003616549

(c)

VIEWTABLE: Work.Wk5										
	sum_sxy	logSalary_N	nHits_N	logSalary_Mean	nHits_Mean	logSalary_Var	nHits_Var	sxy	a	b
1	5045.8961696	263	263	5.9272215412	109.15969582	0.7906631036	1935.4095144	19.259145686	0.0099509409	4.84097986

Fig. 9 単回帰分析の再現過程

データセットwk3に回帰式の a , b をmergeしてwk6を作成し、決定係数の公式の Q および Se を計算した。 Q の公式の $(y_k - \hat{y})^2$ 部分はwk3で求めたdiff_yを二乗した値であり、これを_qとした。 Se の公式の \hat{y} は $y=0.00995x+4.84$ の回帰式から求めた y であり、この値をy_regに格納して、公式の $(y_k - \hat{y})^2$ 部分を計算して_seとした (Fig. 10a)。summaryプロシジャを用いて_qと_seの合計を計算するとそれぞれ207.15..., 156.94...となり、これらの値を公式に当てはめて R^2 を計算すると0.2423...となった。これはregプロシジャで求めた R^2 と同様の値であり、以上からregプロシジャで行った単回帰分析をデータステップで再現することができた。

```

data wk6;
  merge wk3 wk5 (keep = a b _type_);
  by _TYPE_;
run;

data wk7;
  set wk6;

/*Q*/
  _q = diff_y **2;

/*Se*/

```

```

y_reg = nHits * a + b;
_se = (logSalary - y_reg) **2;
run;

proc summary data = wk7 nway;
  var _q _se;
  output out = wk8 sum = q se;
run;

data wk9;
  set wk8;

/*決定係数*/
  r2 = divide ((q-se), q);
run;

```

(a)

	Name	Team	nHits	logSalary	diff_y	_sxy	a	b	_q	y_reg	_se
1	Ashby, Alan	Houston	81	6.163314804	0.236032283	-6.6483145	0.01	4.841	0.056	5.647	0.2665747062
2	Davis, Alan	Seattle	130	6.1737861039	0.248564563	5.13848049	0.01	4.841	0.061	6.1346	0.0015353802
3	Dawson, Andre	Montreal	141	6.2146080364	0.287386557	9.1504754	0.01	4.841	0.083	6.2441	0.0008675633
4	Galarrraga, Andres	Montreal	87	4.5163389723	-1.41088257	31.2647286	0.01	4.841	1.391	5.7067	1.4163987235
5	Griffin, Alfredo	Oakland	169	6.6200732065	0.692851865	41.4604544	0.01	4.841	0.48	6.5227	0.0094837088
6	Newman, Al	Montreal	37	4.248495242	-1.6737263	121.136379	0.01	4.841	2.818	5.2092	0.3228857558
7	Salazar, Argenis	Kansas City	73	4.605170186	-1.32205136	47.8049749	0.01	4.841	1.748	5.5674	0.3258934156
8	Thomas, Andres	Atlanta	81	4.3174881135	-1.60973343	45.3296037	0.01	4.841	2.581	5.647	1.7676180033
9	Thornton, Andre	Cleveland	92	7.0030654588	1.075843918	-18.461154	0.01	4.841	1.157	5.7565	1.5540091576
10	Trammell, Alan	Detroit	159	6.248313432	0.321097891	16.0038165	0.01	4.841	0.103	6.4232	0.0305760302
11	Travis, Alan	Los Angeles	89	6.99998711	0.91987017	-17.629271	0.01	4.841	0.067	6.9264	0.766299636

(b)

	q	se	r2
1	207.15373314	156.94231859	0.2423872058

Fig. 10 単回帰分析の再現結果

3. まとめ

Student の t 検定, カイ二乗検定, Wilcoxon の符号付順位検定および単回帰分析について, SAS プロシジャで求めた結果をデータステップで再現する過程を紹介した. 検定についてじっくり考え, データステップで順を追って計算を行うことで理解が深まる. さらに, SAS プロシジャを実行する際に背景にある処理を意識してアウトプットを見るようになり, 結果を解釈する力を養ううえでも有効である. 検定を再現するうえで SAS を用いるメリットとしては, SAS プロシジャ未実装の検定を行う際に再現プログラムを活用できる点が大きい.

また, 本稿が活用された事例として, 本稿を事前に確認した別担当者が, half student's t-test の実装の際にデータステップでの処理をカスタマイズして解析を行ったことがあった. このことから, SAS プロシジャ未実装の検定に出会うことは決して珍しいことではなく, その際にスムーズに対応できるように日頃から基礎を理解して土台固めをしておくことが重要であるといえる.

4. 参考文献

- [1] SAS Institute, SAS technical News Summer 2006, 2006, p1-2
- [2] 大橋渉, 統計を知らない人のための SAS 入門, オーム社, 2010, p70-81
- [3] Social Survey Research Information Co., Ltd., 6-4. ノンパラメトリック検定-対応のある 2 標本の差の検定

| 統計学の時間 | 統計 WEB, <https://bellcurve.jp/statistics/course/26159.html>, Accessed Aug 6, 2024

- [4] 佐藤倫広, 統計解析入門者のための医療統計学の基礎と SAS の実践, ムイスリ出版, 2022, p49-51
- [5] GMO, カイ二乗検定とは? 検定手法を解説-GMO リサーチ, <https://gmo-research.jp/research-column/chi-square-test>, 2023, Accessed Aug 6, 2024
- [6] jmp, jmp Help, Wilcoxon の符号付順位検定の統計的詳細, <https://www.jmp.com/support/help/ja/17.2/index.shtml#page/jmp/statistical-details-for-the-wilcoxon-signed-rank-test.shtml>, 2023, Accessed Aug 6, 2024
- [7] SAS Institute, SAS Help Center: The UNIVARIATE Procedure: Tests for Location, https://go.documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/procstat/procstat_univariate_details17.htm, 2020, Accessed Aug 6, 2024
- [8] 株式会社アイスタット, ウィルコクソンの符号順位和検定 | ノンパラメトリック検定の一覧ページ | 統計的推定・検定の手法別開設, https://istat.co.jp/sk_commentary/non-parametric_test/Wilcoxon-signed-rank-test, Accessed Aug 8, 2024
- [9] Social Survey Research Information Co., Ltd., 27-1. 単回帰分析 | 統計学の時間 | 統計 WEB, <https://bellcurve.jp/statistics/course/9700.html>, Accessed Aug 8, 2024
- [10] 涌井良幸, ファーストブック 統計解析がわかる, 技術評論社, 2010, p172-184

Proc MIによる欠測値補完の精度評価

○小林 邦世

(イーピーエス株式会社)

Evaluating the accuracy of multiple imputation using the EMB algorithm with proc MI in SAS

Kuniyo Kobayashi

EPS Corporate

要旨

SAS の MI プロシジャを使用して Expectation-Maximization with Boot-strapping (以後 EMB と記載)アルゴリズムによる多重補完法の精度を評価した。EMB アルゴリズムは、期待値最大化法 (Expectation-Maximization, 以後 EM と記載) とブートストラップ法を組み合わせた新しい多重補完法であり、欠測データに対して高い精度で補完を行うことができるとされる。本論文では、シミュレーションデータを生成し、欠測データを発生させた後、EMB アルゴリズムを適用して補完データを作成した。また、R の Amelia パッケージに含まれる `overimpute` 関数に相当する補完値の評価方法を MI プロシジャを利用して SAS 上でも再現し、補完値の妥当性を検証した。結果として、EMB アルゴリズムを用いた多重補完は、欠測データに対して有効であり、補完値が真値を概ね正確に再現していることが確認された。これにより、SAS における EMB アルゴリズムの実用性と有効性が示された。

キーワード：多重補完, EMB アルゴリズム, MI プロシジャ, `overimpute`

1. 緒言

データに欠測が発生している場合、利用可能なデータの数が減少する可能性や、データの偏りが発生する可能性がある。

欠測値対処法の 1 つに多重補完法(Multiple Imputation)がある。多重補完法とは、欠測のあるデータについて複数回補完を行い、完全な (欠測値のない) データセットを複数作成したのち、各データセットに任意の統計手法を適用し、得られた複数の結果を統合する方法である[1]。

多重補完における欠測値補完には多くのアルゴリズムが提唱されている。現在、高橋ら(2013)[2]によって各ソフトウェアにおける各種アルゴリズムの性能比較は実施されているが、SAS における EMB アルゴリズムの性能比較はプロシジャとして未実装のため比較対象には含まれていなかった。近年、矢田・浜田(2017)[3]によって MI プロシジャを利用した SAS における EMB アルゴリズムによる多重補完法が実装可能となった。これを受け、今回、SAS における EMB アルゴリズムの性能評価を実施する。また、R の Amelia パッケージ内の、EMB アルゴリズムによる補完値を診断するための `overimpute` 関数と同様のメカニズムを SAS でも実装し、SAS 上で EMB アルゴリズムによって補完された値の診断も実施する。

2. Multiple Imputation の概要

多重補完法は Rubin, DB(1987)によって提案された方法であり, Multiple imputation(MI)法とよばれる. MI法の手法については, 欠測値を含むデータに対して,

1. 複数回の補完を実施.
2. 補完後のそれぞれの近似的な完全データに対する解析を行う.
3. 得られた結果を1つの最終結果に統合する.

過程がとられる.

補完の方法には様々な種類があり, 各補完方法を用いて複数回の補完を行うことで, 欠測値の補完に対しての不確実性を考慮することができるとされている. 欠測値を補完するための統計モデルのことを補完モデルとよび, 多重補完された完全データを用いて解析するための統計モデルを解析モデルとよぶ.

3. EMB アルゴリズム

EMB アルゴリズムはノンパラメトリック・ブートストラップ(bootstrap)と期待値最大化法から構成される比較的新しい多重補完法のアルゴリズムである[4].

EMB アルゴリズムは2つのステップに分けられる.

1. ブートストラップ法を利用して再標本を得る.
2. 1により得られた各標本に対して EM アルゴリズムを用いて欠測値を補完する.

2における EM アルゴリズムは E ステップと M ステップに分けられる. E ステップでは, 欠測値を含む不完全データ Y と, t 回目のパラメータ推定値 $\theta^{(t)}$ を条件として, 完全データの対数尤度の条件付き期待値を計算する.

$$Q(\theta|\theta^{(t)}) = \int l(\theta|Y) \Pr(Y_{miss}|Y_{obs}, \theta^{(t)}) dY_{miss} \quad (1)$$

続く M ステップで尤度の最大化計算を行う, 上記式(1)の $Q(\theta|\theta^{(t)})$ 関数を最大化することで, $t+1$ 回目の繰り返し時点におけるパラメータ $\theta^{(t+1)}$ を推定する.

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} Q(\theta|\theta^{(t)}) \quad (2)$$

これら E および M ステップを収束するまで繰り返すことで得られた値は最尤推定値であることが知られている[5].

SAS における MI プロシジャを用いた EMB アルゴリズムの実装は矢田ら(2017)[3]に詳しい.

4. overimpute の概要

overimpute は, EMB メカニズムでの多重補完において, 補完モデルの妥当性を判定するために Honaker ら(2011)[6]が R パッケージ Amelia 内の一関数として開発した手法である. 欠損データの補完メカニズムの性

質上、補完された値がその真値に近いかどうかを知ることは不可能である。しかしながら、補完値がどの程度正確なのか、という疑問は自然に発生する。その疑問を解消するため `overimpute` が開発された [6]。R パッケージ `Amelia` 内の `overimpute` 関数では、いわゆる `overimputation` 法を用いて補完値の評価を行う。

図 1 に `overimputation` 法の概略、出典に R の `Amelia` パッケージ内 `overimpute` 関数の出力結果を示す。補完データの任意の 1 変数に対して、観測値を対象とし、観測値を 1 つずつ順次人為的に欠測させる。ここで観測値とは補完する前のデータ(欠測データ)で欠測していない値を意味する。次に、任意の 1 値のみ欠測させた任意の 1 変数に対して、EMB アルゴリズムを用いて多重補完を実施して補完値を得る。1 つの 1 値欠測の補完データに対して 20 回ずつ補完を実施し、補完値のみ取り出す。そして各欠測値に対して、その補完値の平均値とその 90%信頼区間を構築して描画する。出典に描画の具体的な例示を示している。(補完回数と信頼区間の幅については関数内で固定。変更は想定されていない)

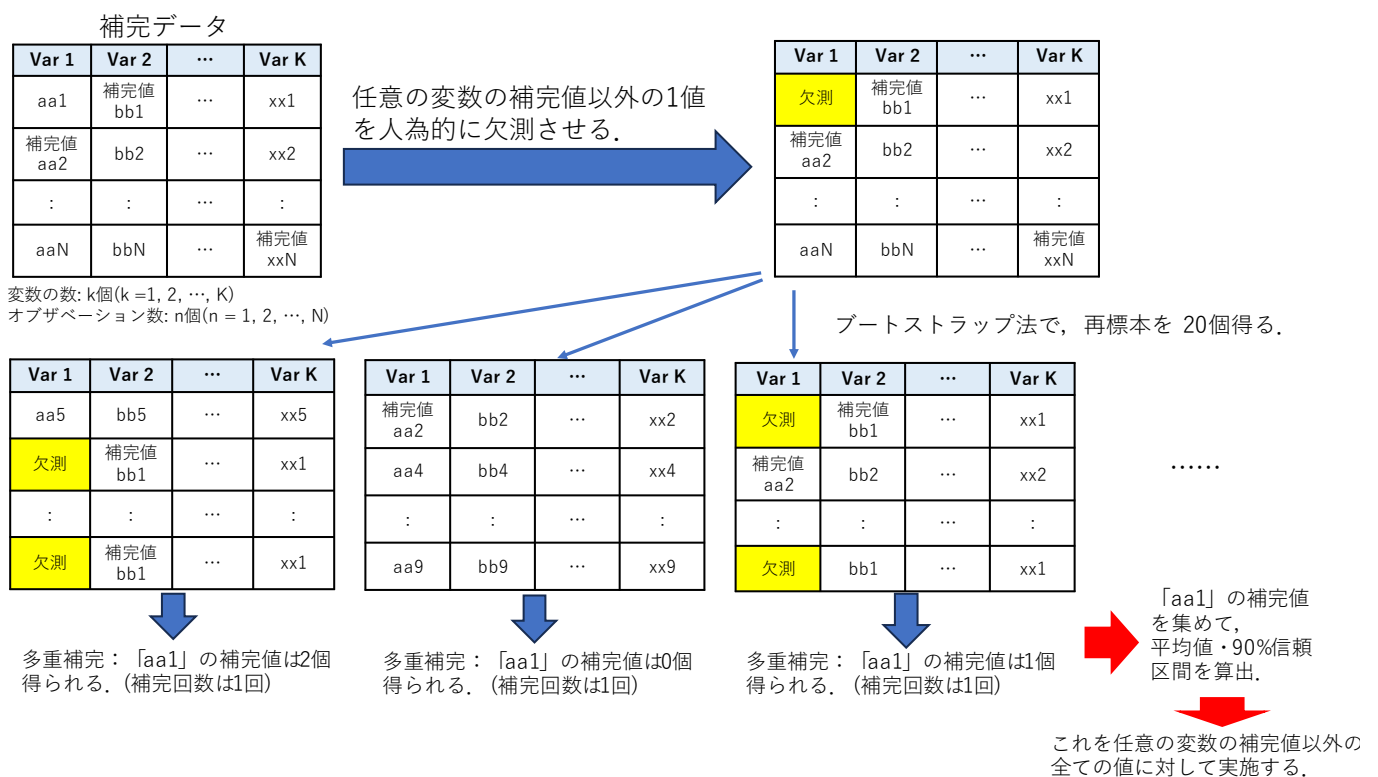
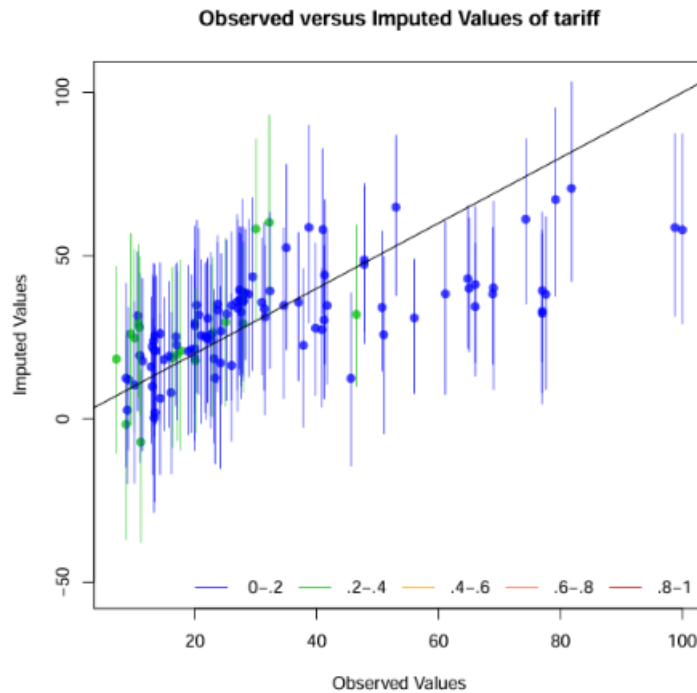


図 1: `overimputation` 法の概略図



出典：Honaker ら(2011)[6] Figure 8

R の "Amelia" パッケージ内 `overimpute` 関数の出力結果。横軸に人為的に発生させた真値、縦軸に真値に対する補完値をとる。図中のマーカーとそこから伸びる直線は、各真値に対する補完値の平均値とその 90% 信頼区間を示している。図内に引かれた 45 度線が、各真値に対する補完値の平均値の 90% 信頼区間内を通過することで、各真値に対する補完値が真値を正確に推定していると判断できる。

ここで、`overimpute` 関数の結果が直感的でない結果を示す場合があることを Honaker ら(2011)[6]は示唆している。極端な例で例えると、`overimpute` 関数にかける変数の観測値が、その変数内の唯一の観測値である場合には、その観測値が欠損しているように見せかけると、その変数の平均と分散以外に、その観測値を補完するために必要な情報が何も残らない。このようなケースでは、信頼区間が非常に大きくなることが予想される。

5. Overimputation の SAS での実装

R の Amelia パッケージ内、`Overimpute` 関数と同様の補完値の評価方法の SAS での実装を検討した。

まず、SAS で Overimputation を実装する。補完された各補完データセットの任意の変数に対して、補完される前のデータにおいて欠測していなかった値に対して 1 つずつ欠測を発生させ、MI プロシジャを利用して EMB アルゴリズムを用いて多重補完を実施する。任意の 1 値のみが欠測した各データセットに対して補完回数は R の `Overimpute` 関数を参考に 20 回とする。EMB アルゴリズムを用いて多重補完された補完データに対して、欠測箇所に補完された値のみ取り出し、真値に対する補完値の平均値とその 90% 信頼区間を作図する。

詳細なコードについては Appendix 3 を参照のこと。

6. 手法

6.1 完全データの発生方法

経時的測定データの完全データと、その完全データに対しさらにMARを仮定するシミュレーションデータを発生させた。シミュレーションデータ発生方法については横山(2016)[7]を参考にした。

シミュレーションデータは、大うつ病性障害患者を対象にベースラインを含んだ4時点でハミルトンうつ病評価尺度(HAM-D)を経時的に測定するものとする。例数は100で、各時点の平均と標準偏差については表1に従う。

表1 各時点の平均値と標準偏差の設定

	ベースライン(y0)	時点1 (y1)	時点2 (y2)	時点3 (y3)
平均値(標準偏差)	20.0 (4.0)	15.0 (5.0)	12.5 (6.0)	11.0(7.0)

6.2 欠測データの発生方法

6.1で発生させた完全データに対しさらにMCAR (MCAR, Missing Completely At Random)とMAR (MAR, Missing At Random)を仮定するシミュレーションデータを発生させる。欠測値発生方法については土井(2013)[8]を参考にした。

MCARの欠測メカニズムについては、ベースライン時点を除く各時点でそれぞれ10%の欠測を仮定した。また、MARの欠測メカニズムについては以下を(3)式を仮定した。MCAR、MARともに最終時点における欠測割合は全体の3割程度となるよう設定し、欠測パターンについては、簡単のため単調な欠測を仮定した。詳細なシミュレーションデータ発生コードについてはAppendix 1を参照のこと。

$$\text{logit}(p_t) = -0.9 + 0.09 * y_{t-1} \quad (3)$$

- t : 時点($t = 1, 2, 3$)
- y_t : 時点 t の測定値
- p_t : 時点 $t - 1$ で観察された症例が時点 t で欠測する確率

6.3 評価方法

高橋ら(2014)[2]を参考に、多重補完によって補完された補完値、補完データの評価を実施する。

まず、多重補完法によって補完された補完データに対して要約統計量を算出する。次に、時点ごとに補完値に対してのみ、真値との散布図を作成する。さらに、時点ごとに完全データと補完データの密度分布を作成する。

最後に overimputaion 法を利用し、EMB アルゴリズムによって補完された補完値の妥当性の評価も実施する。

7 結果

7.1 完全データの要約

シミュレーションで発生させた完全データの要約統計量を表 2 に示す。これにより事前設定したシミュレーションの設定を再現できていることが確認された。

表 2 シミュレーションにおける完全データの要約統計量

	ベースライン(y0)	時点 1 (y1)	時点 2 (y2)	時点 3 (y3)
平均値(標準偏差)	20.03 (4.38)	15.53 (5.04)	12.55 (6.30)	11.70 (7.35)

7.2 欠測データの要約統計量

シミュレーションで発生した完全データに対して単調な欠測を発生させて欠測データを作成した。欠測データの各時点の要約統計量についてはそれぞれ表 3 の結果になった。

表 3 シミュレーションにおける欠測データの要約統計量

	ベースライン(y0)	時点 1 (y1)	時点 2 (y2)	時点 3 (y3)
MCAR 下での欠測データ	20.03 (4.38)	15.44 (5.08)	12.17 (6.04)	12.00 (7.22)
MAR 下での欠測データ	20.03 (4.38)	15.89 (4.97)	11.86 (5.93)	12.84 (8.38)

なお、各時点の欠測データの割合についてはそれぞれ表 4 の結果となった。

表 4 シミュレーションにおける欠測データの各時点の欠測割合

	ベースライン(y0)	時点 1 (y1)	時点 2 (y2)	時点 3 (y3)
MCAR 下での欠測データ	0%	11.0%	23.0%	28.0%
MAR 下での欠測データ	0%	7.0%	19.0%	27.0%

7.3 EMB メカニズム下での補完データの要約

各時点の補完データの要約統計量についてはそれぞれ表 5 の結果になった。

MCAR メカニズム下での欠測データの補完データについては、概ね完全データの要約統計量を再現できている。MAR メカニズム下での欠測データの補完データについては、欠測割合の小さい時点 1 では完全データの要約統計量を再現できているものの、比較的欠測割合の大きい時点 2, 3 においては完全データの要約統計量との若干の乖離がみられた。

表5 EMBメカニズムで補完した補完データの要約統計量

平均値(標準偏差)	ベースライン(y0)	時点 1 (y1)	時点 2 (y2)	時点 3 (y3)
MCAR 下での欠測データ	20.03 (4.36)	15.36 (5.06)	12.26 (6.03)	11.97 (7.13)
MAR 下での欠測データ	20.03 (4.36)	15.62 (5.02)	12.00 (6.09)	12.35 (7.39)

7.4 EMBメカニズム下での補完データの完全データとの比較 (MCAR)

図2にEMBメカニズムによって補完された各時点の補完値と実際の値の散布図を示した。横軸に真値(Observed Values)を、縦軸には補完値(Imputed Values)をとった。散布図には欠測データセットにおける欠測値のみを採用し、欠測していない値については除外している。また散布図には参照線を引いた。また表6に補完値とその真値の相関係数をまとめた。

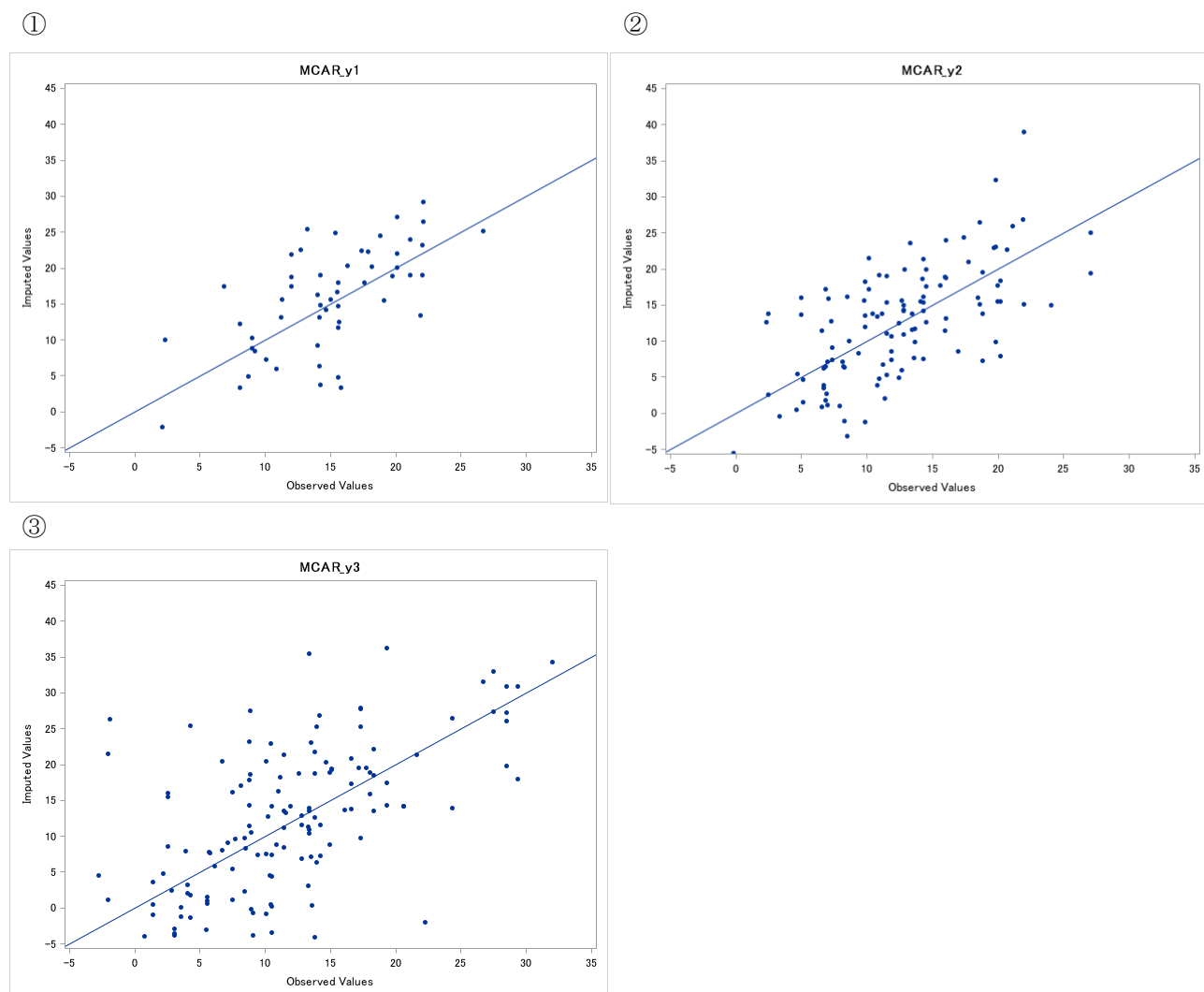


図2 EMBメカニズムによって補完された各時点の補完値と実際の値の散布図。①は時点1、②は時点2、③は時点3の結果をそれぞれ示している。

表6 各時点における補完値とその真値の相関係数

時点 1 (y1)	時点 2 (y2)	時点 3 (y3)
0.93681	0.75996	0.72078

MCAR メカニズム下での欠測データにおいて、各時点で、真値と補完値の間に大まかに直線の相関関係があることが分かった。時点 1, 2, 3 の順に散布図でのばらつきが大きくなることが明らかであることから、欠測割合が真値と補完値の相関関係に影響を及ぼすことが示唆された。

また、図3に完全データの密度分布と EMB メカニズムによって補完された補完データの密度分布を時点ごとに作成した。変数に含まれるすべての値を対象に密度分布を描画した。グラフ内の赤線は完全データ (True Value) の密度分布を、青線は補完データ (Multiple Imputed Value) の密度分布を示している。

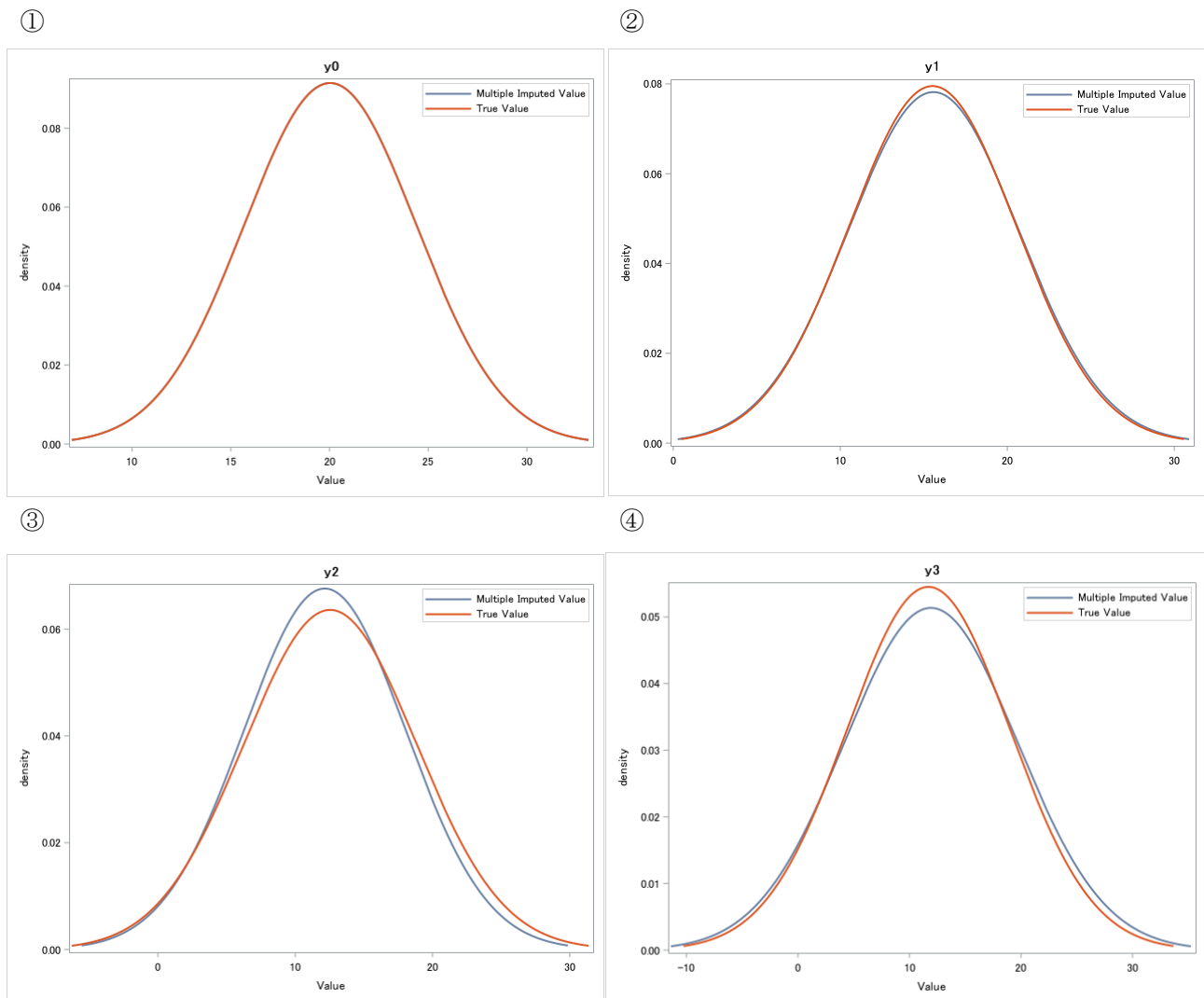


図3 EMB メカニズムによって補完された各時点の補完値と実際の値の散布図。①はベースライン時点、②は時点1、③は時点2、④は時点3の結果をそれぞれ示している。

データ内で比較的欠測割合の小さい時点1においては、完全データの密度分布と補完データの密度分布ほぼ一致したが、時点2、時点3と比較的欠測割合が大きい時点においては、完全データの密度分布と補完デ

ータの密度分布で若干の差異が生じる結果となった。しかしながらどの時点でも、補完データはある程度完全データを再現できていることが示唆された。

また、各時点で、Overimputation 法によって得られた補完値とその真値について散布図を作成した。結果は図 4 に示す。横軸に真値(Observed Values), 縦軸に overimputation 法による補完値(Imputed Values)をとった。また、参照線として 45 度線を採用している。

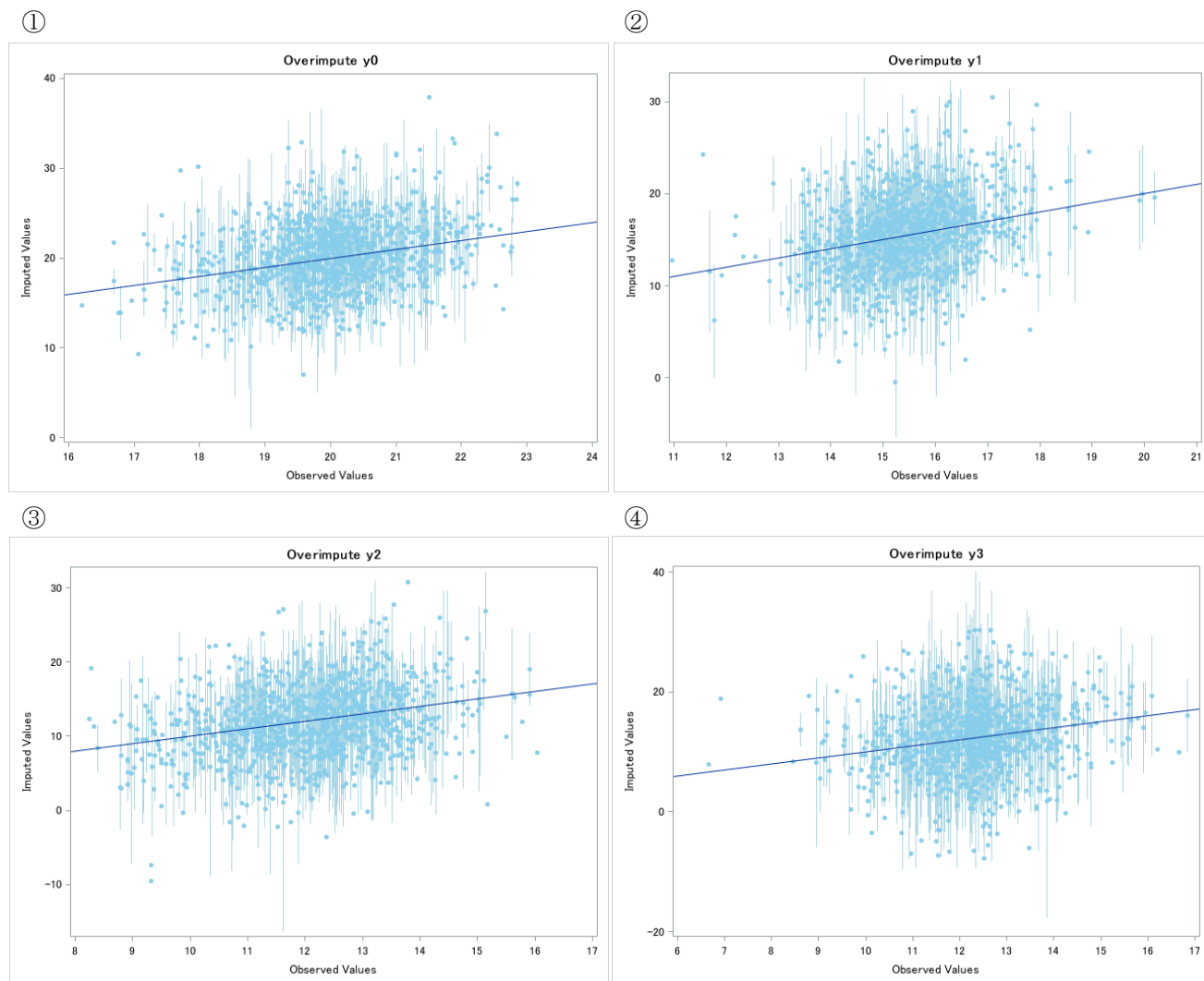


図 4 Overimputation 法によって得られた真値に対する補完値の平均値とその 90%信頼区間の図。①はベースライン時点, ②は時点 1, ③は時点 2, ④は時点 3 の結果をそれぞれ示している。

各時点において、45 度線が各真値の 90%信頼区間内に概ね収まっていることが確認された。したがって、MCAR メカニズム下での欠測データに対して補完された値は概ね妥当であると評価できる。

7.5 EMB メカニズム下での補完データの完全データとの比較 (MAR)

図 5 に EMB メカニズムによって補完された各時点の補完値と実際の値の散布図を示した。横軸に真値 (Observed Values) を、縦軸には補完値(imputed Values)をとった。散布図には欠測データセットにおける欠測値のみを採用し、欠測していない値については除外している。また散布図には参照線を引いた。ベースラインでは欠測値が発生していないため、散布図は描画していない。また表 7 に補完値とその真値の相関係数をまとめた。

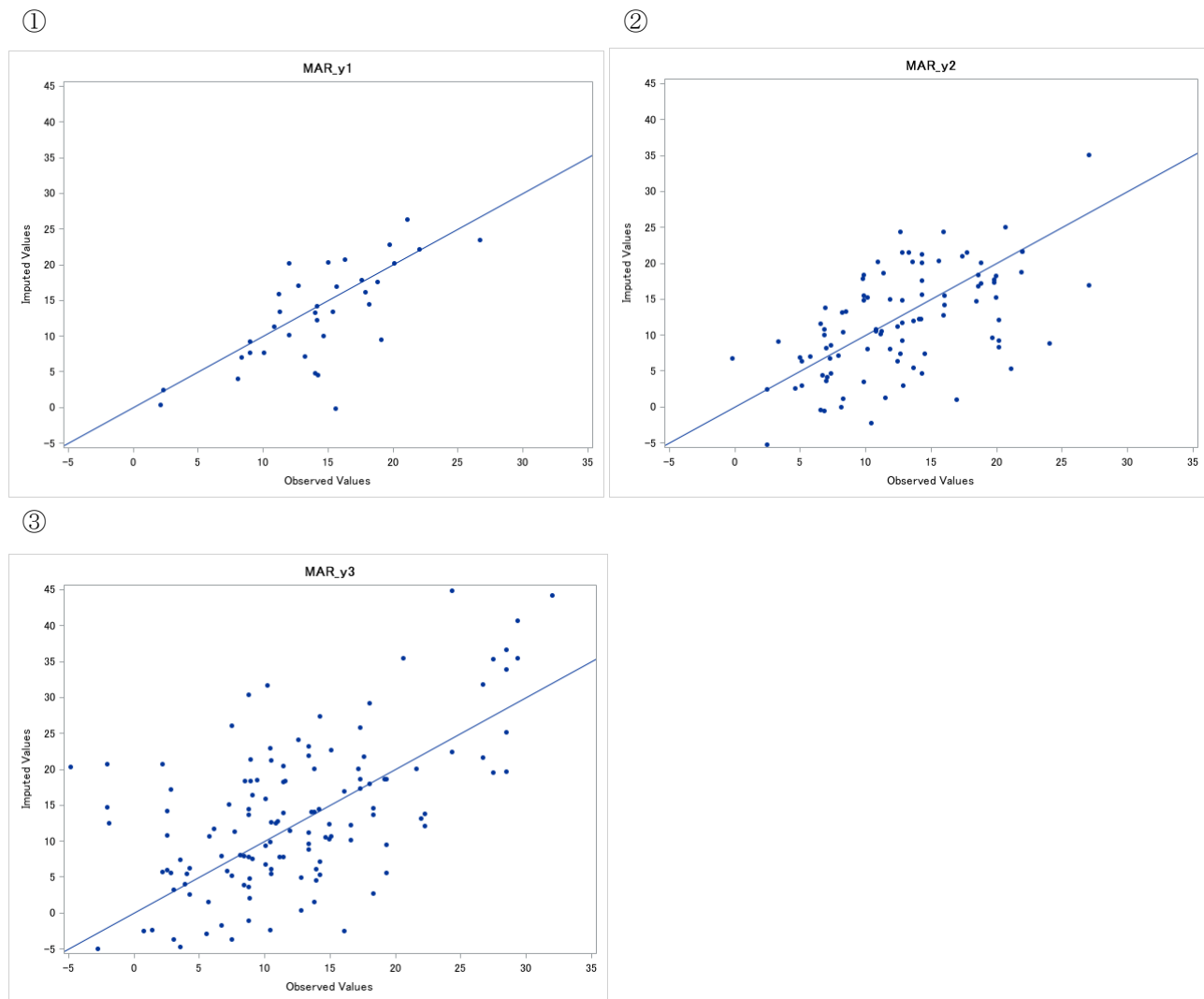


図 5 EMB メカニズムによって補完された各時点の補完値と実際の値の散布図。①はベースライン時点、②は時点 1、③は時点 2、④は時点 3 の結果をそれぞれ示している。

表 7 各時点における補完値とその真値の相関係数

時点 1 (y1)	時点 2 (y2)	時点 3 (y3)
0.92436	0.74892	0.73832

図 5, 表 7 より, MAR メカニズム下での欠測データにおいても各時点で, 真値と補完値の間に直線の相関関係があることが分かった。また時点を経るごとに相関係数がやや落ちることが確認されたが直線の相関

関係は保たれていることは確認された。時点 1, 2, 3 の順に散布図でのばらつきが大きくなった。特に時点 3 ではマーカーのばらつきが他の時点と比べてやや散見された。

また、図 6 に時点ごとに完全データの密度分布と EMB メカニズムによって補完された補完データの密度分布を作成した。変数に含まれるすべての値を対象に密度分布を描画した。グラフ内の赤線は完全データ (True Value) の密度分布を、青線は補完データ (Multiple Imputed Value) の密度分布を示している。

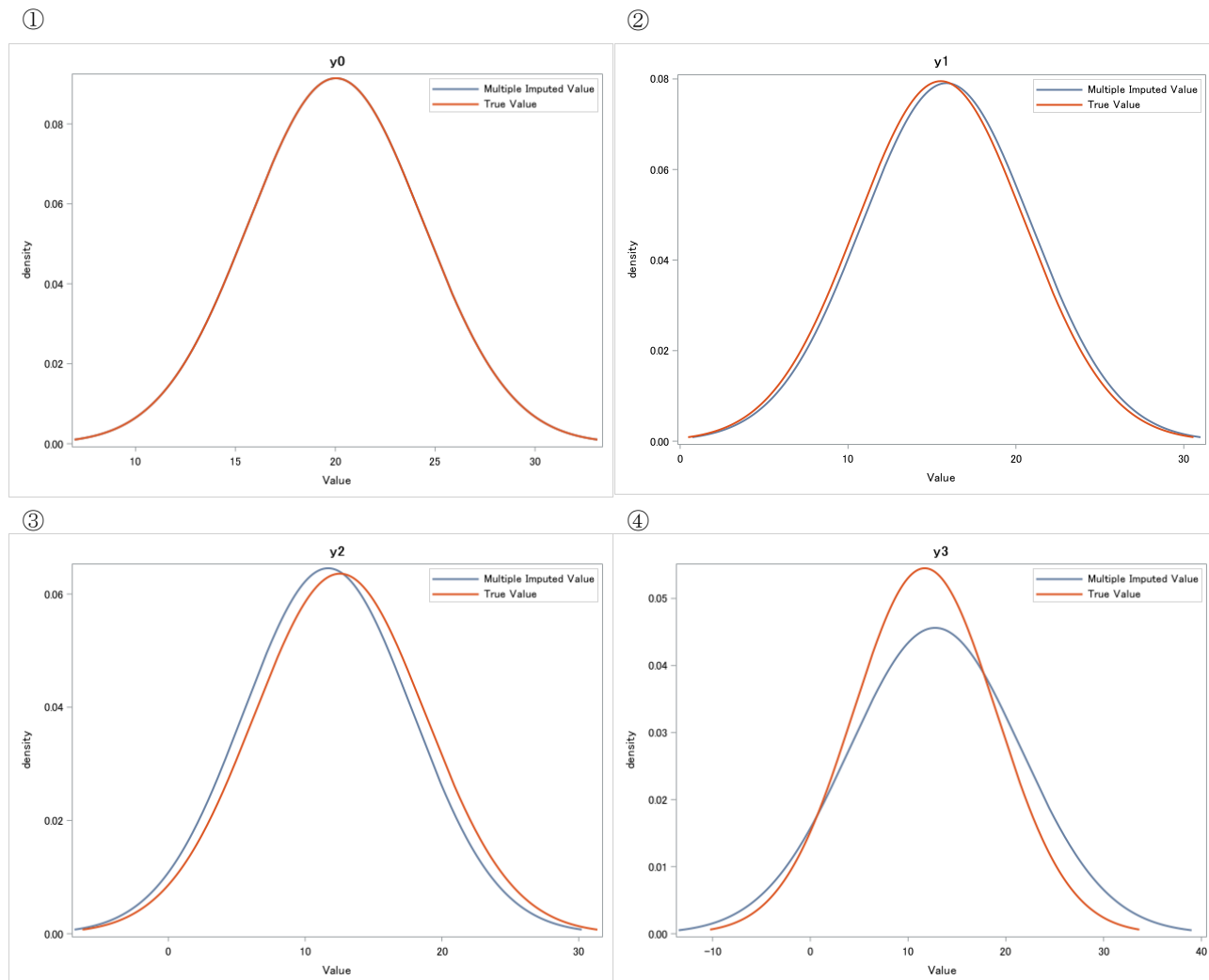


図 6 EMB メカニズムによって補完された各時点の補完値と実際の値の散布図。①はベースライン時点、②は時点 1、③は時点 2、④は時点 3 の結果をそれぞれ示している。

データ内で比較的欠測割合の小さい時点 1 においては、完全データの密度分布と補完データの密度分布ほぼ一致したが、比較的欠測割合が大きい時点 2, 3 においては、完全データの密度分布と補完データの密度分布で若干の差異が生じる結果となった。

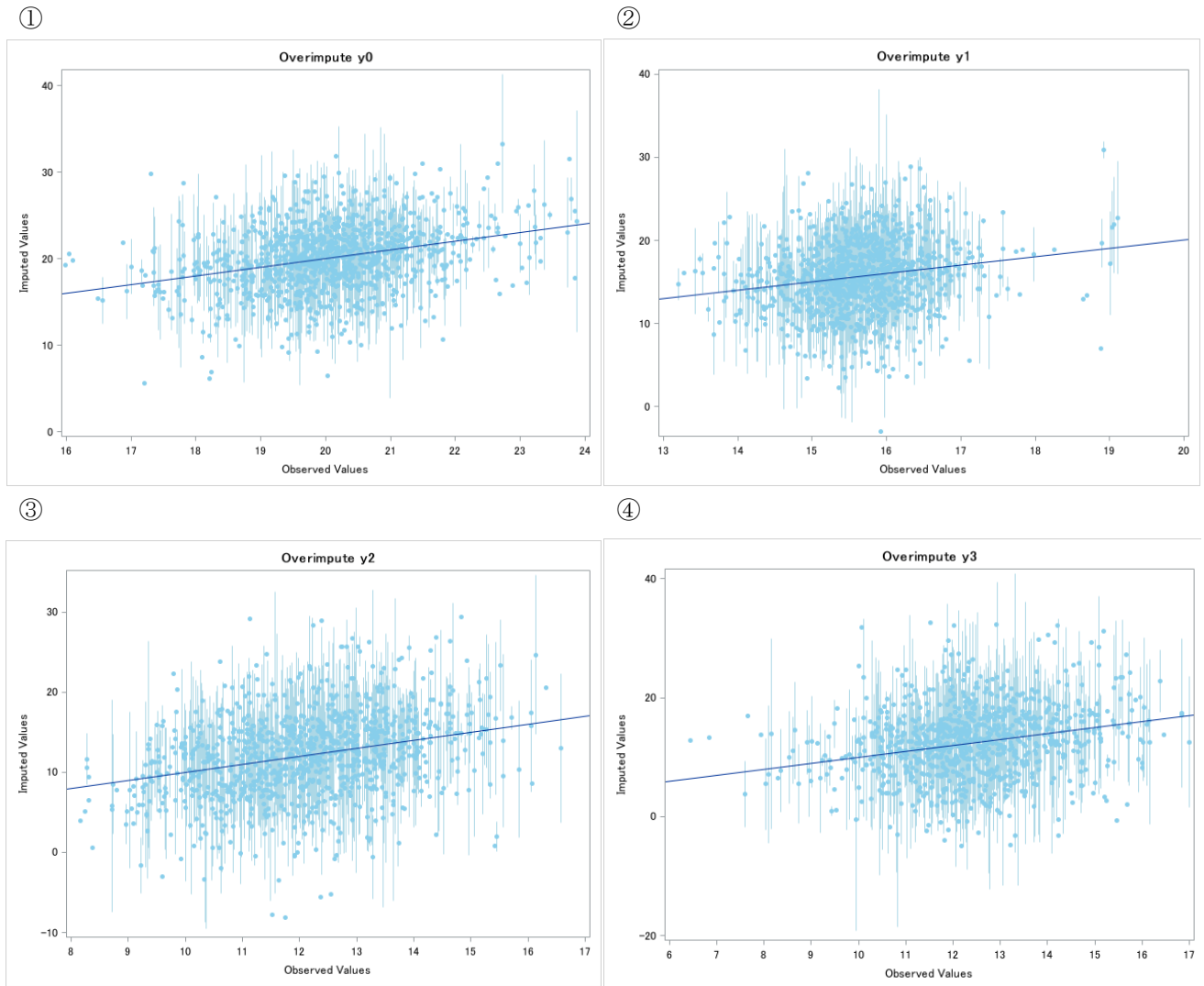


図7 EMBメカニズムによって補完された各時点の overimpute の結果。①はベースライン時点，②は時点1，③は時点2，④は時点3の結果をそれぞれ示している。

図7より，MCARメカニズム下での欠測データ同様，各時点において，45度線が各真値の90%信頼区間の中に概ね入っていることが確認された。したがって，MARメカニズム下での欠測データに対しても補完された値は概ね妥当であると評価できる。

8 考察

EMBアルゴリズムで補完された補完値について，散布図，密度関数，overimputation法の3視点から妥当性を評価した。散布図，密度関数からは直接的に，補完値が真値をほぼ再現していることが確認できた。また，SASで実装したoverimputation法によって，補完されたデータセットにおける補完値の妥当性を証明できた。MCAR, MARの両欠測メカニズム下での欠測データにおいて，SASにおけるEMBアルゴリズムを用いた多重補完による補完値，補完データについては概ね正確に推定を実施できていると考えられる。

9 まとめ

近年は R や Python を中心に、Open Source System の枠組みで、様々な統計解析手法についてのパッケージ開発が活発である。また SAS のような商用解析ソフトウェアにおいても機能拡張や新しい製品パッケージの開発が進んでいる。しかし、基本的には、手法の内容を理解し、各プログラム言語の機能を活用すれば、ソフトウェアが違うということのハードルは高いものではないと考える。SAS の場合においても、データステップと既存のプロシジャの組み合わせで、他ソフトウェアでのパッケージと同等の結果を得ることが可能である。本稿においても R の Amelia パッケージにおける Overimpute 関数の挙動を再現することができた。今後も研究を進め、利用するソフトウェアを問わずに万人がデータサイエンスに取り組めるように寄与できれば幸いである。

参考文献

- 1, Rubin, D.B, Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, 1987
- 2, 高橋将宜, 伊藤孝之, 様々な多重代入法アルゴリズムの比較 ~大規模経済系データを用いた分析~, 2014, 統計研究彙報 第 71 号 3(39~82)
- 3, 矢田真城, 浜田知久馬, EMB アルゴリズムを用いた Multiple Imputation, 2017, SAS ユーザー総会
https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups2017-d-12.pdf
(最終アクセス: 2024/08/22)
- 4, Honaker, James and Gary King, 2010, What to do About Missing Values in Time Series Cross-Section Data, American Journal of Political Science vol.54, no.2, pp.561-581
- 5, 高橋将宜, 渡辺美智子, 欠測データ処理 R による単一代入法と多重代入法, 2017
- 6, Honaker, James, Gary King, and Matthew Blackwell, 2011, Amelia II: A Program for Missing Data, Journal of Statistical Software vol.45, no.7, pp.1-47.
- 7, 横山雄一, 大浦智紀, 土井正明, MI プロシジャで実行可能な Pattern Mixture Model と Multiple Imputation に基づく解析, 2016, SAS ユーザー総会
https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups2016-a-02-06.pdf
(最終アクセス: 2024/08/22)
- 8, 土井正明, SAS による例数設計のシミュレーション, 2013, SAS ユーザー総会
https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups13-c-11.pdf
(最終アクセス: 2024/08/22)

Appendix. 1 シミュレーションデータ作成コード

```
/*完全データ作成*/  
data d0;  
    /*乱数固定*/  
    call streaminit(1);  
    do i = 1 to 100;  
        y0 = rand("normal", 20, 4);  
        y1 = rand("normal", 15, 5);  
        y2 = rand("normal", 12.5, 6);  
        y3 = rand("normal", 11, 7);  
    end;  
run;
```

```

                trt=1;
                output;
            end;
id = i;
run;

/*MCAR メカニズム下での欠測データ作成*/
data MCAR;
set d0;
call streaminit(123);
array y{4} y0-y3;
array p{3} p1-p3;
array m{3} m1-m3;
do t=1 to 3;
    p{t}=0.03;
    m{t}=rand('bernoulli',p{t});
end;
if m1=1 then do; y1=.; end;
if m2=1 then do; y2=.; end;
if m3=1 then do; y3=.; end;
run;

data _MCAR;
set MCAR;
array y{4} y0-y3;

    if y{1} = . then do; y{2} = .; y{3} = .; y{4} = .; end;
    if y{2} = . then do; y{3} = .; y{4} = .; end;
    if y{3} = . then do; y{4} = .; end;
run;

/*MAR メカニズム下での欠測データ作成*/
data MAR;
set d0;
call streaminit(123);
array y{4} y0-y3;
array p{3} p1-p3;
array m{3} m1-m3;
do t=1 to 3;
    p{t}=exp(- 0.9 - 0.09*y{t});
    m{t}=rand('bernoulli',p{t});
end;
if m1=1 then do; y1=.; end;
if m2=1 then do; y2=.; end;
if m3=1 then do; y3=.; end;
run;

data _MAR;
set MAR;
array y{4} y0-y3;

    if y{1} = . then do; y{2} = .; y{3} = .; y{4} = .; end;
    if y{2} = . then do; y{3} = .; y{4} = .; end;
    if y{3} = . then do; y{4} = .; end;
run;

```

上記サンプルコードで、MCAR メカニズム下での欠測データ”_mcar”と MAR メカニズム下での欠測データ”_mar”の2つのデータセットを作成可能。

Appendix. 2 EMB アルゴリズムによる欠測データ補完

```

/*補完対象の変数名をすべて指定する*/
%let var = y0 y1 y2 y3;

/*ブートストラップ法によって欠測データを増幅させる(今回の対象は”_mcar”データセット)*/
%macro analyze(data=, out=);
proc means noprint data=&data. vardef=n;
    output out=&out(drop=_freq_ _type_)

```

```

        var =&var. ;
        %bystmt;
run;
%mend;
%boot(data=_MCAR, samples = 5, random=1234, print=0, chart=0) /*data= で対象のデータセット名を指定 */
/*EMアルゴリズムによる推定*/
%macro m_miem(inds=, samples=, seed=, outem = , outmi=);
    data m_wkst1; run;
    data m_wkst2; run;
    %do irep = 1 %to &samples.;
        data m_wrk;
            set &inds.;
            if _sample_ = &irep. ;
        run;
        proc sort data = m_wrk; by _obs_;run;
        data cnt;
        set m_wrk; by _obs_;
        array aa(*) &var.;
        ct = 0;
        do i = 1 to dim(aa);
            if (aa(i))=. then ct +1;
        end;
        run;
        proc sort data= cnt; by ct; run;
        data _null_;
        set cnt; by ct;
        if last.ct then call symputx("ct", ct);
        run;
        %put &ct.;
        %if &ct. ne 0 %then %do;
            proc mi data = m_wrk seed = &seed. nimpute = 0 noprint;
                em outem = m_outem out=m_outmi initial = ac;
                var &var. ;
            run;
            data m_outem; set m_outem; _sample_ = &irep.; run;
            data m_outmi; set m_outmi; _sample_ = &irep.; run;
            data m_wkst1; set m_wkst1 m_outem; if &irep. ^=. ; run;
            data m_wkst2; set m_wkst2 m_outmi; if &irep. ^=. ; run;
        %end;
    %end;
    data &outem.; set m_wkst1; run;
    data &outmi.; set m_wkst2 ; run;
%mend m_miem;

%m_miem(inds= BOOTDATA, samples = 5, seed=1234, outem= outem, outmi=outmi);

data m_outmi; set OUTMI; keep _sample_ &var. _obs_; run;
data m_outmi; set m_OUTMI; by _sample_ ; if first._sample_ =1 then rep=0; rep+1; run;

data m_ind;
set BOOTDATA;
keep _sample_ M1 - M4;
array aa M1 - M4;
array bb &var.;

do over aa;
    if bb = . then aa = 0;
    else if bb ^= . then aa = 1;
end;
run;

data m_ind; set m_ind; by _sample_ ; if first._sample_ = 1 then rep = 0; rep +1 ;run;
data _m_var;
set OUTEM;
if _TYPE_ = 'COV';
run;

data m_var;

```

```

set _m_var;
keep _sample_ varno var;
  if _NAME_ = 'y0' then do ;
    varno = 1;
    var = y0;
  end;
  else if _NAME_ = 'y1' then do ;
    varno = 2;
    var = y1;
  end;
  else if _NAME_ = 'y2' then do ;
    varno = 3;
    var = y2;
  end;
  else if _NAME_ = 'y3' then do ;
    varno = 4;
    var = y3;
  end;
run;

proc transpose data = m_var out=s_var prefix=var;
var var;
by _sample_;
run;

data m_ind;
merge m_ind s_var(drop= _NAME_);
by _sample_;
run;

data m_outmi;
merge m_outmi m_ind;
by _sample_ rep;
array aa &var.;
array bb I_y0 I_y1 I_y2 I_y3;
array cc M1 - M4;
array dd var1 -var4;

do over aa;
  if cc= 1 then bb = aa;
  else if cc = 0 then bb = aa + sqrt(dd)*rand('normal');
end;
run;

data outemb;
set m_outmi;
if ^missing(_obs_);

run;

```

上記 EMB アルゴリズムで補完した補完データ”outemb”が得られる。

Outemb データには,

sample : ブートストラップ法によって作成された何個目の再標本か.

obs : 元のデータセットにおいて何番目のオブザベーションであったか.

I_y0 ~ I_y4 : 元のデータの y0 ~ y4 に対応する, 補完後のデータが格納された各変数.

M1 ~ M4 : 元のデータの y0 ~ y4 に対応して、同変数同_obs_のデータが欠測(0)か欠測でない(1)かを示す識別変数.

の 10 変数が含まれる.

sample	_obs_	m1	m2	m3	M4	I_y0	I_y1	I_y2	I_y3
1	1	1	1	1	1	20.079327333	20.060107213	13.402453154	13.464497937
1	1	1	1	1	1	20.079327333	20.060107213	13.402453154	13.464497937
3	1	1	1	1	1	20.079327333	20.060107213	13.402453154	13.464497937
4	1	1	1	1	1	20.079327333	20.060107213	13.402453154	13.464497937
5	1	1	1	1	1	20.079327333	20.060107213	13.402453154	13.464497937
2	2	1	1	1	1	23.130330022	14.887427042	14.289626925	13.813392828
3	2	1	1	1	1	23.130330022	14.887427042	14.289626925	13.813392828
4	2	1	1	1	1	23.130330022	14.887427042	14.289626925	13.813392828
4	2	1	1	1	1	23.130330022	14.887427042	14.289626925	13.813392828
4	2	1	1	1	1	23.130330022	14.887427042	14.289626925	13.813392828
1	3	1	1	1	1	23.037839117	24.403282795	13.856345204	8.8960515795
2	3	1	1	1	1	23.037839117	24.403282795	13.856345204	8.8960515795
2	3	1	1	1	1	23.037839117	24.403282795	13.856345204	8.8960515795

Appendix. 3 Appendix.2 にて EMB アルゴリズムで補完した補完データ”outemb”に対して、overimputation を実施するサンプルコード(対象変数: I_y0)

```

%let target_var = I_y1; /*overimpute法によって確認したい変数名を指定*/
%let non_target_var = I_y0 I_y2 I_y3; /*overimpute法によって確認したい変数以外の変数名を指定*
%let out_var = OI_y1 OI_y1 OI_y2 OI_y3; /*overimpute法の結果を出力するための変数名の指定*/
%let OI_data = outemb; /*overimpute法を実施するデータセット名を指定*/

/*---start prepare macro---*/
%macro analyze(data=, out=);
    proc means noprint data=&data vardef=n;
        output out=&out(drop=_freq_ _type_)
            var =&target_var.;
        %bystmt;
    run;
%mend;

%macro m_mioi(inds=, samples=, seed=, outem = , outmi=);
    data m_wkst1; run;
    data m_wkst2; run;
    %do irep = 1 %to &samples.;
        data m_wrk;
            set &inds.;
            if _sample_ = &irep. ;
        run;
        proc sql noprint;
            select count(*) into: ct trimmed from m_wrk where &target_var.=.;
        quit;
        %put &=ct.;
        %if &ct. ne 0 %then %do;
            proc mi data = m_wrk seed = &seed. nimpute = 0 noprint;
                em outem = m_outemoui out=m_outoui initial = ac;
                var &target_var. &non_target_var.;
            run;
            data m_outemoui; set m_outemoui; _sample_ = &irep.; run;
            data m_outoui; set m_outoui; _sample_ = &irep.; run;
            data m_wkst1; set m_wkst1 m_outemoui; if &irep. ^=. ; run;
            data m_wkst2; set m_wkst2 m_outoui; if &irep. ^=. ; run;
        %end;
    %end;
    data &outem.; set m_wkst1; run;
    data &outmi.; set m_wkst2; run;
%mend m_mioi;

%macro co_var();
data m_var;

```

```

set _m_var;
    %let target_count_var = &target_var. &non_target_var.;
    %let num_count = %sysfunc(countw(&target_count_var.));
    %do temp = 1 %to &num_count.;
    %let number = %scan(&target_count_var. , &temp.);
        if _NAME_ = "&number." then do;
            var = &number.;
            varno = &temp.;
            output;
        end;
    %end;
keep _sample_ varno var;
run;
%mend co_var;
/*---end prepare macro---*/

proc sql noprint; select count(*) into:num_obs trimmed from &OI_data.; quit;
%put &num_obs.;
proc sql noprint; select max(_obs_) into:max_obs trimmed from &OI_data.; quit;
%put &max_obs.;
data overimpute; run;

%macro overimpute();
%do obs = 1 %to &max_obs.;
    data wk1;
    set &OI_data.;
    if _OBS_ = &obs. and M1 = 1 then &target_var. =.;
    keep &target_var. &non_target_var. _obs_;
run;

    %boot(data=wk1, samples = 5, random=1234, print=0, chart=0)
    %m_mioi(ind= BOOTDATA, samples = 5, seed=1234, outem= outemioi, outmi=outioi);

    data m_outioi; set OUTIOI; keep _sample_ &target_var. &non_target_var. _obs_ ; run;
    data m_outioi; set m_OUTIOI; by _sample_ ; if first._sample_ = 1 then rep=0; rep+1; run;
    data m_ind;
    set BOOTDATA;
    keep _sample_ M1 - M4;
    array aa M1 - M4;
    array bb &target_var. &non_target_var.;

    do over aa;
        if bb = . then aa = 0;
        else if bb ^= . then aa = 1;
    end;
run;

    data m_ind; set m_ind; by _sample_ ; if first._sample_ = 1 then rep = 0; rep + 1 ;run;
    data _m_var;
    set OUTEMIOI;
    if _TYPE_ = 'COV';
run;

%co_var;

proc transpose data = m_var out=m_var prefix=var;
var var;
by _sample_;
run;

data m_ind;
merge m_ind m_var(drop= _NAME_);
by _sample_;
run;

data m_outioi;
merge m_outioi m_ind;
by _sample_ rep;
array aa &target_var. &non_target_var. ;

```

```

array bb &out_var.;
array cc M1 - M4;
array dd var1 -var4;

if _obs_ ^=.;
do over aa;
    if cc= 1 then bb = aa;
    else if cc = 0 then bb = aa + sqrt(dd)*rand('normal');
end;
run;

data overimpute;
set overimpute m_outoi;
if M1 = 0;
run;

%end;
%mend;

%overimpute;

/*overimpute描画*/
proc summary data=overimpute _mcar_y1 nway;
class I_y1;
var OI_y1;
output out=wk2 mean= std= /autoname;
run;
data wk3;
set wk2;
if n(OI_y1_Mean,OI_y1_StdDev)= 2 then do;
upper=OI_y1_Mean+OI_y1_StdDev;
lower=OI_y1_Mean-OI_y1_StdDev;
end;
run;
title "Overimpute y1";
proc sgplot data=wk3 noautolegend;
scatter x=I_y1 y=OI_y1_mean /
markerattrs=(size=5 color=skyblue symbol=circlefilled)
errorbarattrs=(color=lightblue thickness=0.1) errorcapscale=0.01
yerrorupper=upper
yerrorlower=lower;
lineparm x=0 y=0 slope=1;
xaxis values=(11 to 21) label="Observed Values";
yaxis label="Imputed Values";
run;

```

コピュラモデルによる相関のある生存時間解析： SAS プログラムの開発

○中藺 孝輔^{1,2}, 魚住 龍史², 江村 剛志^{1,3}

¹ 統計数理研究所医療健康データ科学研究センター, ² 東京工業大学工学院経営工学系,
³ 久留米大学バイオ統計センター

Copula-Based Survival Analysis with Dependent Survival Time Using SAS

Kosuke Nakazono^{1,2}, Ryuji Uozumi², Takeshi Emura^{1,3}

¹ *Research Center for Medical and Health Data Science, The Institute of Statistical Mathematics*

² *Department of Industrial Engineering and Economics, Tokyo Institute of Technology*

³ *Biostatistics Center, Kurume University*

要旨

ある治療が生存時間に与える効果は2群の生存時間が独立という仮定の下で、メディアン之差やEfron推定量などの古典的な方法で妥当な推定量を計算することができるが、2群の生存時間に依存関係がある場合、通常の推定量についての解釈が難しくなることが知られている。このような状況では、2群の生存時間の同時分布をモデリングし計算を修正する必要がある。そこで、同時生存時間分布のモデリングにコピュラモデルを用いることで相関のある生存時間データにおける治療効果の推定を行うことができる (Nakazono et al., 2024)。この推定には同時生存関数の偏微分の積分計算が必要であり、著者らはSAS/IMLによる積分計算を伴う数値計算プログラムを開発した。本発表では、コピュラモデルを用いた生存時間データにおけるMann-Whitney効果の推定法について、開発したSASプログラムの実行手順を報告する。

キーワード: 因果推論、接合関数、生存曲線、パラメトリックモデル、反事実モデル、治療効果、二群比較

1 はじめに

医学研究や信頼性工学の分野では、興味のあるイベントが発生するまでの時間を対象とした解析手法である生存時間解析がしばしば行われる。特に医学研究では死亡や興味のある疾患の再発、進行をイベントとして解析を行うことが多い。SAS/STATでは、生存時間解析を実行するプロシジャとして、LIFETESTやLIFEREG、PHREGなどといったプロシジャがサポートされており、生存時間解析の代表的な手法であるログランク検定やCox回帰を容易に実行できる (大橋ら, 2016, 2022)。

さて、二つの独立な母集団の生存時間 T_1, T_2 を比較する生存時間解析では、一方の生存時間がもう一方の生存時間よりも長い確率に興味があることがある。この確率は $T_1 > T_2$ の確率と $T_1 > T_2$ の確率の1/2の和であらわされ、Mann-Whitney効果と呼ばれる。Mann-Whitney効果 p はMann-Whitney検定やハザード比などの指標とも関連し、生存時間解析における治療効果の推定において重要な役割を果たしている。Mann-Whitney効果の推定については、Efron (1967) により基本的なアイデアが提案されており、この方法では比較する二つの母集団が独立で T_1, T_2 が連続である場合には二群の周辺生存関数 S_1, S_2 を用いた積分によって推定される。しかしながら、 T_1, T_2

が独立でない場合には2つの周辺生存関数を推定することにより、Efronの方法で推定した p と実際の確率が異なることがある。さらに一定の条件の下では周辺生存関数のみから推定した値と実際の値が $1/2$ を超えて異なるという、“Handのパラドックス”と呼ばれる現象が発生する (Hand, 1992; Greenland et al., 2020). それゆえ、周辺生存関数のみから得られる p の推定値は真の治療効果または因果効果として解釈が難しいという問題がある。Fay et al. (2018)はこの範囲に対してランダム化比較試験の場合の解釈を与えている。しかしながら、 T_1, T_2 の独立性が仮定できない状況で“一方の生存時間がもう一方の生存時間よりも長い確率”を計算するためには T_1, T_2 の同時分布をモデリングする必要がある。

生存時間などの非正規分布を持つ確率変数の同時分布のモデリングにおいて、コピュラモデル (Nelsen, 2006) が有用となる。コピュラの利点は、 T_1, T_2 それぞれの周辺生存関数と、これらの相関構造のモデリングを別々に行うことができることである。日本語でのコピュラの一般的な解説に塚原 (2003) などがあり、医学研究における生存時間解析へのコピュラの応用にも、江村・大庭 (2024) などがある。

本稿では二変量生存時間分布のモデリングにパラメトリックコピュラモデルを用いて、様々なパターンの従属関係がある生存時間データにおける p の計算方法を与えた提案法 (Nakazono, et al., 2024; 中園ら, 2024) を SAS で実行するプログラムについて解説する。また、ランダム化比較試験のデータセットに特定の確率分布を仮定した提案法を適用し、様々な従属構造の下での p の変化について観察する。開発した SAS マクロのプログラム全文は付録に添付した。

2 提案手法

本節では、Nakazono et al. (2024) で提案された手法を紹介する。はじめに生存時間変数のコピュラモデリング法について簡単に紹介する。

2.1 生存時間データにおけるコピュラモデル

コピュラとは閉区間 $[0,1]$ 上の一様分布を周辺分布を持つ同時分布関数である (Nelsen, 2006)。コピュラについて最も基本的な定理に Sklar の定理がある (Sklar, 1959)。

定理 2.1 Sklar の定理

任意の d 次元分布関数 F に対して、 F_1, \dots, F_d を周辺分布とすると、

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)),$$

を満たす関数 C が存在する。

通常、コピュラは、同時累積分布関数に $\mathbb{P}(T_1 \leq t_1, T_2 \leq t_2)$ に対して定義されるものであるが同時生存関数についても“生存コピュラ”と呼ばれる形で同じようにコピュラを定義することができる。ここで T_1, T_2 を、それぞれ S_1, S_2 を周辺生存関数とする連続な生存時間とする。Sklar の定理より、 T_1, T_2 の二変量生存時間関数はコピュラ C を用いて次のようにモデリングすることができる。

$$\mathbb{P}(T_1 > t_1, T_2 > t_2) = C(S_1(t_1), S_2(t_2)). \quad (1)$$

式 (1) のような表現は、周辺分布 S_1, S_2 と従属構造を示すコピュラ C が分離されているという点で有用である。コピュラには様々なものが提案されているが、Nakazono et al. (2024) ではクレイトン、グンベル、フランク、Farlie-Gumbel-Morgenstern (FGM)、グンベル-バーネット (GB) コピュラを用いた。本稿では COPULA プロシジャでも利用可能なクレイトン、グンベル、フランクコピュラを利用した。これらのコピュラの式を下記に示す。

1. クレイトン族:

$$C_\theta(u, v) = \max\left((u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, 0\right), \quad \theta \in [-1, \infty) \setminus \{0\}.$$

2. グンベル族:

$$C_\theta(u, v) = \exp\left\{-\left[(-\log u)^{\theta+1} + (-\log v)^{\theta+1}\right]^{\frac{1}{\theta+1}}\right\}, \quad \theta \in [0, \infty).$$

3. フランク族:

$$C_{\theta}(u, v) = -\frac{1}{\theta} \log \left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right), \quad \theta \in (-\infty, \infty) \setminus \{0\}.$$

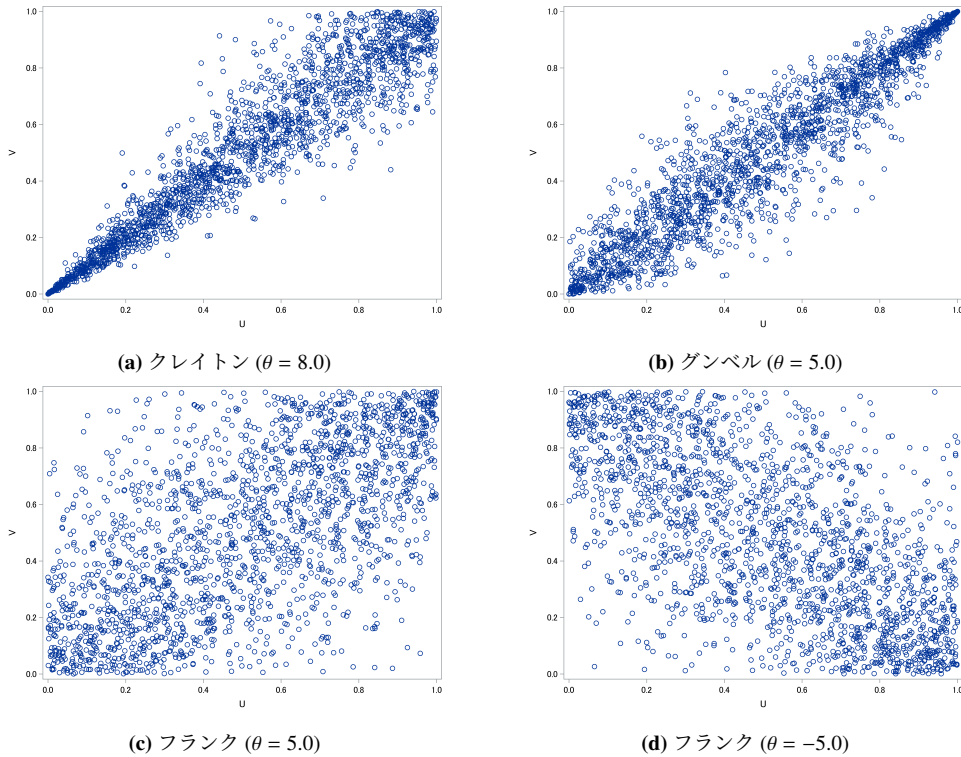


図 1: COPULA プロシジャを用いて発生させたコピュラを分布関数として持つ (U, V) の 2,000 個の乱数のプロット

ソースコード 1: クレイトンコピュラの疑似乱数を発生させる SAS プログラム

```
proc copula;
  var U V ;
  define COP clayton (theta=8) ;
  simulate COP/ ndraws=2000 seed=1111 outuniform=unifdata;
run;
```

これらの5つのコピュラは COPULA プロシジャを用いてそれぞれのコピュラに従う疑似乱数を生成することができる(矢田・浜田, 2014). ソースコード 1は COPULA プロシジャを用いて $\theta = 8.0$ のクレイトンコピュラに従う 2000 個の疑似乱数を発生させる SAS コードであり, 図 1は4つのコピュラから生成した疑似乱数の散布図である. 図からもわかるようにコピュラはそれぞれ異なる従属構造を示すことができ, クレイトンコピュラ(図 1 (a))は下側の裾に, グンベルコピュラ(図 1 (b))は上側の裾に強い従属があることがわかる. また, フランクコピュラ(図 1 (c),(d))は左右対称の従属構造となっている. これらのコピュラは生存時間解析やその他のデータ解析で広く利用されている. 例えば従属打ち切りのある生存時間解析や競合リスクの解析 (Schneider et al., 2023; Emura and Hsu, 2020), メタアナリシス (Yamaguchi and Maruo, 2019; Shih et al., 2022) での応用がある. このようにコピュラごとに異なる特性を持つことから, 従属構造のある生存時間データにおいては様々なコピュラを考慮することが重要である.

従属の強さを示すパラメータ θ はケンドールのタウへ 1 対 1 変換できる. ケンドールのタウは T_1 と T_2 の間の相関の強さを表す係数として知られ, 周辺分布に依存せずにコピュラのみから決定される.

$$\text{ケンドールのタウ} = 4 \int_0^1 \int_0^1 C_{\theta}(u, v) C_{\theta}(du, dv) - 1.$$

例えばクレイトンコピュラの場合, ケンドールのタウはコピュラパラメータ θ を用いて $\theta/(\theta + 2)$ と表される.

2.2 Mann-Whitney 効果 p の計算

Nakazono et al. (2024) では, Mann-Whitney 効果 p をコピュラモデル (1) の下で計算する手法を提案している. まず, コピュラの分布に従う一様分布の確率変数を $U = S_1(T_1), V = S_2(T_2)$ とする. このとき, $V = v$ を与えたときの U の条件付き分布は, C の v での偏微分によって与えられ, これを $C^{[0,1]}(u, v)$ とする. このとき条件付き分布を用いて p は次のように求まる.

$$\begin{aligned} p &= \mathbb{P}(T_1 > T_2) + \frac{1}{2}\mathbb{P}(T_1 = T_2) \\ &= \int_0^1 C^{[0,1]}(S_1(S_2^{-1}(v)), v)dv. \end{aligned} \quad (2)$$

また, T_1, T_2 に共有のフォローアップ期間 τ が設定されたもとでは, 時点 τ 以降にイベントが発生するデータはすべて打ち切りデータとして扱われる. そのため, τ までの確率 $\mathbb{P}(\min(T_1, \tau) > \min(T_2, \tau))$ と, τ 以降のデータをすべてタイデータとした $\frac{1}{2}\mathbb{P}(\min(T_1, \tau) = \min(T_2, \tau))$ を加える必要がある (Dobler and Pauly, 2018). よって, 式 (2) は以下のように修正される (Nakazono et al., 2024).

$$\begin{aligned} p_\tau &= \mathbb{P}(\min(T_1, \tau) > \min(T_2, \tau)) + \frac{1}{2}\mathbb{P}(\min(T_1, \tau) = \min(T_2, \tau)) \\ &= \int_{S_2(\tau)}^1 C^{[0,1]}(S_1(S_2^{-1}(v)), v)dv + \frac{1}{2}C(S_1(\tau), S_2(\tau)). \end{aligned} \quad (3)$$

式 (2), (3) より p, p_τ を計算するためには積分計算が必要だが, SAS では SAS/IML を用いてこれらの数値積分を実行することができる. 計算のための SAS プログラムは 3 節で解説する.

2.3 周辺生存関数

今回の SAS プログラムでは以下の 4 つの周辺生存関数を設定した. ここでは生存関数 $S_j(t)$ と Mann-Whitney 効果の計算に用いる $S_i(S_j^{-1}(v))$ を示す.

1. 指数分布:

$$\begin{aligned} S_j(t) &= \exp(-\lambda_j t), \quad \lambda_j > 0, \\ S_i(S_j^{-1}(v)) &= v^{\frac{\lambda_i}{\lambda_j}}, \end{aligned}$$

ここで λ_j は指数分布のパラメータ (rate parameter) である.

2. ワイブル分布:

$$\begin{aligned} S_j(t) &= \exp(-\lambda_j t^{k_j}), \quad \lambda_j > 0, k_j > 0, \\ S_i(S_j^{-1}(v)) &= \exp\left(-\lambda_i \left(-\frac{\log v}{\lambda_j}\right)^{\frac{k_i}{k_j}}\right), \end{aligned}$$

ここで λ_j は尺度パラメータであり, k_j は形状パラメータである.

3. 対数正規分布:

$$S_j(t) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \int_t^\infty \frac{1}{y} \exp\left\{-\frac{1}{2\sigma_j^2}(\log y - \mu_j)^2\right\} dy,$$

ここで μ_j は位置パラメータ, σ_j は尺度パラメータである. また, 対数正規分布については生存関数 $S_j(t)$ が閉形式の逆関数を持たないため, プログラム中では, SAS 関数 “quantile” を利用して $S_j^{-1}(v)$ を計算している.

4. Burr III 分布:

$$S_j(t) = 1 - (1 + t^{-c_j})^{-k_j}, \quad c_j > 0, k_j > 0,$$

$$S_i(S_j^{-1}(v)) = 1 - \left(1 + \left((1 - v)^{-\frac{1}{k_j}} - 1 \right)^{\frac{c_j}{k_j}} \right)^{-k_i},$$

ここで c_j, k_j は形状パラメータである.

上記の生存関数は医学や信頼性工学における生存時間解析で広く扱われており、疾患や故障の発生までの時間の解析や、競合リスクの解析など様々な適用例がある (武富ら, 2023). 図 2は上記の生存関数のプロットである. パラメータの設定によって幅広い生存関数を扱うことができることがわかる.

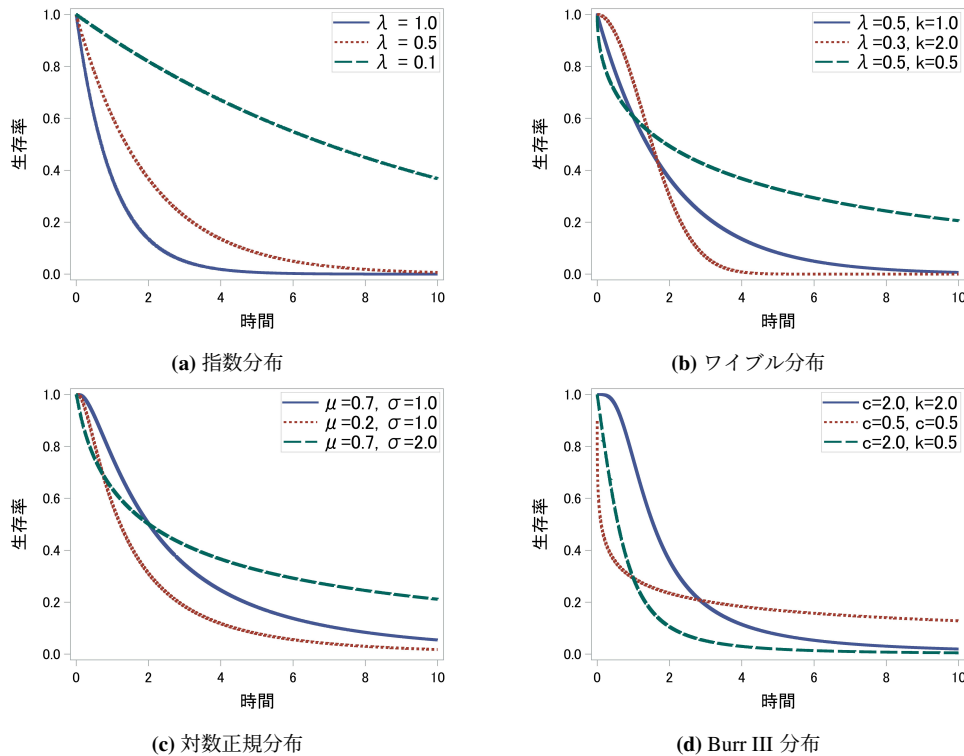


図 2: 本稿で扱う生存関数

3 SAS プログラムの開発

本節では 2 節で解説した Nakazono et al. (2024) による提案法を実行する SAS プログラムの解説を行う. 提案法ではコンピュータを用いてモデリングした同時生存関数の偏微分の積分をする必要があるが, SAS では SAS/IML の組み込み関数を用いることで容易に数値積分を行うことができる. ここでは付録 A に与えた SAS マクロコードの詳細について解説する.

作成したプログラムでは被積分関数を関数 HFunc として定義している. この関数 HFunc は 2.3 項で示した $S_i(S_j^{-1}(v))$ をそれぞれ,

- ワイブル分布

```
E=exp(-lambda1*((-log(x)/lambda2)**(k1/k2)));
```

- 対数正規分布

```
E=1-CDF('LOGNORMAL', quantile('LOGNORMAL', 1-x, lambda2, k2), lambda1, k1);
```

- Burr III 分布

```
E=1-(1+((1-x)**(-1/k2)-1)**(lambda1/lambda2))**(-k1);
```

とし, コピュラモデルの偏微分を,

- クレイトンコピュラ

```
v=x**(-theta-1)*(E**(-theta)+x**(-theta)-1)**(-1/theta-1);
```

- グンベルコピュラ

```
u=(-log(E))**(theta+1)+(-log(x))**(theta+1);
v=exp(-u**(1/(theta+1)))*u**(-theta/(1+theta))*(-log(x))**theta/x;
```

- フランクコピュラ

```
u=exp(-theta*E)-1;
v=(exp(-theta*x)*u)/(exp(-theta)-1+u*(exp(-theta*x)-1));
```

- FGM コピュラ

```
v=E+theta*E*(1-E)*(1-2*x);
```

- GB コピュラ

```
v=E*(1-theta*log(E))*x**(-theta*log(E));
```

とし, 被積分関数を変数 v として返す関数である. この関数を IML 関数 *quad* を用いて以下のように数値積分を行うことで式 (2) で示した Mann-Whitney 効果 p を求めることができる.

ソースコード 2: quad 関数による積分

```
eps=1E-8;
call quad(p, "MyFunc", {0 1}, eps);
p = round(p, 0.001);
```

プログラムの後半部分は式 (3) で示したフォローアップ期間を考慮した p_τ の計算に係る部分である. ここでは式3右辺第二項 $C(S_1(\tau), S_2(\tau))$ を計算し, 上記の関数で求めた右辺第一項と足し合わせている.

作成した SAS マクロのコードのアウトラインは以下のとおりである (全コードは付録 A を参照).

ソースコード 3: SAS コードのアウトライン

```
%macro mymacro(lambda1, k1, lambda2, k2, dist, theta, copula, tau);
proc iml;
  /** パラメータをセット **/
  lambda1 = &lambda1; lambda2 = &lambda2; k1 = &k1; k2 = &k2;
  theta = &theta; tau = &tau; copula = "&copula"; distribution = "&dist";

  /** 式(4), 式(5) の被積分関数を定義 **/
  start HFunc(x) global(lambda1, lambda2, k1, k2, theta);
    %if "&dist" = "weibull" %then %do;
      E = exp(-lambda1*((-log(x)/lambda2)**(k1/k2)));
    %end;
    %else %if "&dist" = .....

    %if "&copula" = "clayton" %then %do;
      v = x**(-theta-1)*(E**(-theta)+x**(-theta)-1)**(-1/theta-1);
    %end;
    %else %if "&copula" = .....
```

```

    return(v);
finish;

/** 積分を実行し p を計算 **/
eps=1E-8; call quad(p, "EFunc", {0 1}, eps);
p = round(p, 0.001); print p;

/** 式(5) 右辺第 2 項の計算 **/
%if "&dist" = "weibull" %then %do;
    S1 = exp(-lambda1*(tau**k1)); S2 = exp(-lambda2*(tau**k2));
%end;
%else %if "&dist" = .....
%end;

%if "&copula" = "clayton" %then %do;
    p1 = (S1**(-theta)+S2**(-theta)-1)**(-1/theta)/2;
%end;
%else %if "&copula" = .....
%end;

/** 式(5) 右辺第 1 項の計算 **/
call quad(p2, "EFunc", S2 || 1, eps);

/** p_tau の計算 **/
p_tau = round(p1+p2, 0.001);

/** 結果の出力 **/
print distribution lambda1 k1 lambda2 k2
      copula theta tau p p_tau;
quit;
%mend mymacro;

```

ここで2つの例を上記の SAS マクロで実行する。例1として、パラメータ $\lambda_1 = 0.01, k_1 = 1.1, \lambda_2 = 0.012, k_2 = 1.3$ のワイブル分布を周辺分布とし、コピュラに $\theta = 3.0$ のグンベルコピュラ、フォローアップ期間を $\tau = 10.0$ とする。また例2として、周辺生存関数を対数正規分布 ($\mu_1 = 0.7, \sigma_1 = 1.0, \mu_2 = 0.4, \sigma_2 = 1.5$)、コピュラをフランクコピュラ ($\theta = -1.0$)、フォローアップ期間を $\tau = 5.0$ とする。ソースコード4のようにこれらの例を上記の SAS マクロで実行することで表1のように、例1では $p = 0.949$ と $p_\tau = 0.580$ を、例2では $p = 0.560$ と $p_\tau = 0.563$ を得ることができる。

ソースコード 4: SAS マクロの実行例

```

/**例1**/
%mymacro(0.01,1.1,0.012,1.3,weibull,3.0,gumbel,10)

/**例2**/
%mymacro(0.7,1.0,0.4,1.5,lognormal,-1.0,frank,5)

```

表 1: SAS マクロによる p, p_τ の計算結果

	周辺分布	コピュラ	θ	τ	λ_1 or μ_1	k_1 or σ_1	λ_2 or μ_2	k_2 or σ_2	p	p_τ
例 1	ワイブル	グンベル	3.0	10	0.01	1.1	0.012	1.3	0.949	0.580
例 2	対数正規	フランク	-1.0	5	0.7	1.0	0.4	1.5	0.560	0.563

4 データ解析

本節では前節で解説したプログラムをデータセット VALung (Veterans Administration Lung cancer trial) に適用する。このデータセットは手術不能な進行性肺がんの標準治療群 (Therapy = 'standard') 69 例と試験治療群 (Therapy = 'test') 68 例からなるランダム化比較試験のデータであり、生存時間解析のテキストでも使用されている (大橋ら,

2016). 本稿では治療群を表す *Therapy*, 生存時間を表す *SurvTime*, そして打ち切りを表す *Censor* のみを利用した. はじめに周辺生存関数に指数分布とワイブル分布を仮定しそれぞれの分布のパラメータの最尤推定量 (MLE) を求める. ここで生存時間を $T_{ij}, j = 1, 2, i = 1, \dots, n_j$, 打ち切り時間を $C_{ij}, j = 1, 2, i = 1, \dots, n_j$, 観測する生存時間を $t_{ij} = \min(T_{ij}, C_{ij}), j = 1, 2$, 打ち切り指示変数を $\Delta_{ij} = \mathbb{1}(T_{ij} \leq C_{ij}), j = 1, 2$ とすると, 指数分布のパラメータの MLE は以下のように求まる.

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n_j} \Delta_{ij}}{\sum_{i=1}^{n_j} t_{ij}}, j = 1, 2.$$

また, ワイブル, 対数正規, Burr III 分布のパラメータの尤度関数は以下のとおりである.

- ワイブル分布

$$L_j(\lambda_j, k_j) = \prod_{i=1}^{n_j} \left(\lambda_j k_j t_{ij}^{k_j-1} \exp(-\lambda_j t_{ij}^{k_j}) \right)^{\Delta_{ij}} - \left(\exp(-\lambda_j t_{ij}^{k_j}) \right)^{1-\Delta_{ij}}, j = 1, 2.$$

- 対数正規分布

$$L_j(\mu_j, \sigma_j) = \prod_{i=1}^{n_j} \left(\frac{1}{t_{ij} \sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{1}{2\sigma_j^2} (\log t_{ij} - \mu_j)^2 \right\} \right)^{\Delta_{ij}} - \left(1 - \Phi \left\{ \frac{1}{\sigma_j} (\log t_{ij} - \mu_j) \right\} \right)^{1-\Delta_{ij}}, j = 1, 2.$$

ここでは対数正規分布の生存関数を変数変換によって標準正規分布の分布関数を用いて表現している.

- Burr III 分布

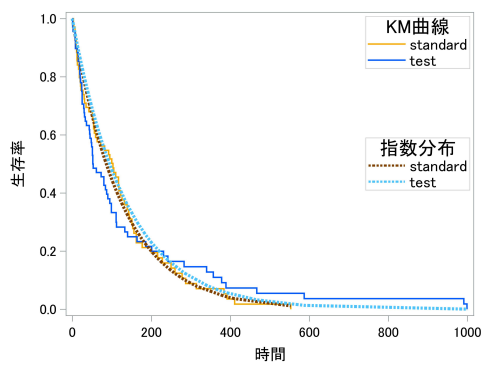
$$L_j(\lambda_j, k_j) = \prod_{i=1}^{n_j} \left(c_j k_j t_{ij}^{-c_j-1} \left(1 + t_{ij}^{-c_j} \right)^{-k_j-1} \right)^{\Delta_{ij}} - \left(1 - \left(1 + t_{ij}^{-c_j} \right)^{-k_j} \right)^{1-\Delta_{ij}}, j = 1, 2.$$

パラメータの MLE は上記の尤度関数から得た対数尤度関数を最大化することで得る. SAS では NL MIXED プロシジャを用いることで対数尤度関数を最大化する最尤推定量を求めることができる (大橋ら, 2022).

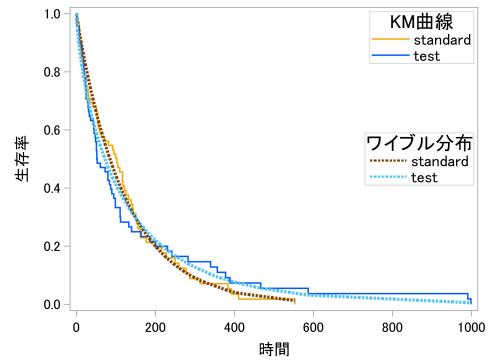
ソースコード 5: 試験治療群におけるワイブル分布のパラメータについての最尤推定

```
proc nlmixed data=VALung (where=(Therapy = 'test'));
  parms lambda=0.01 k=0.01;
  bounds lambda > 0, k > 0;
  if censor=1 then lh = lambda*k*SurvTime**(k-1)*exp(-lambda*SurvTime**k);
  else lh = exp(-lambda*SurvTime**k);
  ll = log(lh);
  model SurvTime ~ general(ll);
run;
```

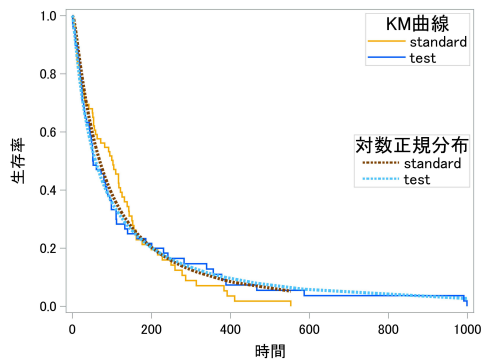
VALung における指数分布のハザードの MLE は $\hat{\lambda}_1 = 0.00734, \hat{\lambda}_2 = 0.00806$ であり, ワイブル分布のパラメータの MLE は $\hat{\lambda}_1 = 0.0258, \hat{k}_1 = 0.768, \hat{\lambda}_2 = 0.00868, \hat{k}_2 = 0.986$, 対数正規分布は $\hat{\mu}_1 = 4.07, \hat{\sigma}_1 = 1.47, \hat{\mu}_2 = 4.24, \hat{\sigma}_2 = 1.26$, Burr III 分布は $\hat{c}_1 = 0.667, \hat{k}_1 = 10.0, \hat{c}_2 = 0.769, \hat{k}_2 = 16.7$ であった. 図3は KM 曲線とそれぞれの分布関数を当てはめた生存曲線である. ここではフォローアップ期間は両群の最後のイベントのうち早く発生した時間 ($\tau = 553$) とした. 提案法のデータへの適用の結果は表 2 にまとめた. 得られた p, p_τ の範囲はそれぞれ, 指数分布: [0.516, 0.618], [0.5160, 0.618], ワイブル分布: [0.359, 0.485], [0.356, 0.479], 対数正規分布: [0.325, 0.476], [0.320, 0.476], Burr III 分布: [0.345, 0.482], [0.333, 0.482] であった. すべての確率分布で結果が帰無値である 0.5 をまたぐことはなかったが, 仮定する分布によって値は大きく異なることから, 同時生存関数のモデリングの仮定に結果が大きく依存することがわかる.



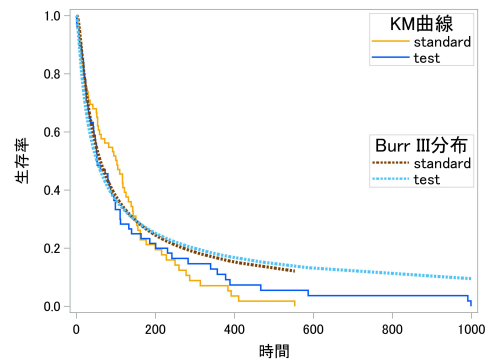
(a) 指数分布 ($\hat{\lambda}_1 = 0.00734, \hat{\lambda}_2 = 0.00806$)



(b) ワイブル分布 ($\hat{\lambda}_1 = 0.0258, \hat{k}_1 = 0.768, \hat{\lambda}_2 = 0.00868, \hat{k}_2 = 0.986$)



(c) 対数正規分布 ($\hat{\mu}_1 = 4.07, \hat{\sigma}_1 = 1.47, \hat{\mu}_2 = 4.24, \hat{\sigma}_2 = 1.26$)



(d) Burr III 分布 ($\hat{c}_1 = 0.667, \hat{k}_1 = 10.0, \hat{c}_2 = 0.769, \hat{k}_2 = 16.7$)

図 3: データセット VALung における KM 曲線と最尤法を用いた指数分布, ワイブル分布, 対数正規分布, Burr III 分布による生存曲線

表 2: データセット VALung における KM 曲線と最尤法を用いた指数分布 ($\hat{\lambda}_1 = 0.00734, \hat{\lambda}_2 = 0.00806$), ワイブル分布 ($\hat{\lambda}_1 = 0.0258, \hat{k}_1 = 0.768, \hat{\lambda}_2 = 0.00868, \hat{k}_2 = 0.986$), 対数正規分布 ($\hat{\mu}_1 = 4.07, \hat{\sigma}_1 = 1.47, \hat{\mu}_2 = 4.24, \hat{\sigma}_2 = 1.26$), Burr III 分布 ($\hat{c}_1 = 0.667, \hat{k}_1 = 10.0, \hat{c}_2 = 0.769, \hat{k}_2 = 16.7$) の下での p の推定結果

周辺分布	コピュラ	θ	ケンドールのタウ	\hat{p}	$\hat{p}_\tau(\tau = 553)$
指数	独立	-	0.00	0.523	0.523
		クレイトン	0.5	0.33	0.530
		2.0	0.50	0.559	0.556
		グンベル	1.0	0.50	0.546
		4.0	0.80	0.614	0.613
		フランク	-20.0	-0.82	0.516
	-5.0		-0.46	0.517	0.517
	5.0		0.46	0.542	0.542
	20.0		0.82	0.618	0.618
	ワイブル	独立	-	0.00	0.471
クレイトン			0.5	0.33	0.474
		2.0	0.50	0.485	0.479
		グンベル	1.0	0.50	0.431
		4.0	0.80	0.359	0.356
		フランク	-20.0	-0.82	0.476
-5.0			-0.46	0.477	0.477
5.0			0.46	0.453	0.452
20.0			0.82	0.375	0.374
対数正規		独立	-	0.00	0.466
	クレイトン		0.5	0.33	0.462
		2.0	0.50	0.455	0.447
		グンベル	1.0	0.50	0.417
		4.0	0.80	0.325	0.320
		フランク	-20.0	-0.82	0.476
	-5.0		-0.46	0.475	0.475
	5.0		0.46	0.439	0.438
	20.0		0.82	0.338	0.336
	Burr III	独立	-	0.00	0.470
クレイトン			0.5	0.33	0.467
		2.0	0.50	0.465	0.450
		グンベル	1.0	0.50	0.419
		4.0	0.80	0.345	0.333
		フランク	-20.0	-0.82	0.482
-5.0			-0.46	0.480	0.480
5.0			0.46	0.444	0.442
20.0			0.82	0.359	0.351

5 まとめ

ランダム化比較試験において, ある被験者の治療群でのアウトカムと対照群でのアウトカムが従属関係を持ちうるような状況は十分考えられる. しかしながら, 潜在アウトカムはいずれかの群でしか観測ができないため, 何らかの感度解析のフレームワークが必要となる. そこで本稿では, 従属関係がある生存時間データについて, パ

ラメトリックコピュラモデルを用いた Mann-Whitney 効果の計算方法 (Nakazono et al., 2024) の SAS プログラムを開発した。さらに実データに適用し、Mann-Whitney 効果の様々な相関構造の下での性能を考察した。提案法では同時生存関数の偏微分の積分が必要になるが、SAS/IML では組み込み関数を用いることで容易に数値積分を実行することができる。

付録 A: SAS コード

ソースコード3で紹介した SAS コードの全文を以下に示す。このプログラムを実行することで p, p_τ を計算する SAS マクロが利用できる。周辺分布として指数, ワイブル, 対数正規, Burr III 分布が, コピュラとしてクレイトン, グンベル, フランクコピュラが選択できる。

ソースコード A: SAS コード

```
/**
dist = {weibull lognormal burrIII}
copula = {clayton gumbel frank}
S1: lambda1, k1 S2: lambda2, k2
[exponential] scale parameter: lambda
                shape parameter: k=1 (dist = weibull)
[Weibull]       scale parameter: lambda
                shape parameter: k
[log-normal]   mu: lambda
                sigma: k
[Burr III]     c: lambda
                k: k
**/
%macro mymacro(lambda1, k1, lambda2, k2, dist, theta, copula, tau);
proc iml;
  /**set parameters**/
  lambda1 = &lambda1; lambda2 = &lambda2; k1 = &k1; k2 = &k2;
  theta = &theta; tau = &tau; copula = "&copula"; distribution = "&dist";

  start HFunc(x) global(lambda1, lambda2, k1, k2, theta);
  %if "&dist" = "weibull" %then %do;
    E = exp(-lambda1*((-log(x)/lambda2)**(k1/k2)));
  %end;
  %else %if "&dist" = "lognormal" %then %do;
    E = 1-CDF('LOGNORMAL',
              quantile('LOGNORMAL', 1-x, lambda2, k2), lambda1, k1);
  %end;
  %else %if "&dist" = "burrIII" %then %do;
    E = 1-(1+((1-x)**(-1/k2)-1)**(lambda1/lambda2))**(-k1);
  %end;

  %if "&copula" = "clayton" %then %do;
    v = x**(-theta-1)*(E**(-theta)+x**(-theta)-1)**(-1/theta-1);
  %end;
  %else %if "&copula" = "gumbel" %then %do;
```

```

    u = (-log(E))**(theta+1)+(-log(x))**(theta+1);
    v = exp(-u**(1/(theta+1)))*u**(-theta/(1+theta))
        *(-log(x))**theta/x;
%end;
%else %if "&copula" = "frank" %then %do;
    u = exp(-theta*E)-1;
    v = (exp(-theta*x)*u)/(exp(-theta)-1+u*(exp(-theta*x)-1));
%end;
    return(v);
finish;

/** integrate v on [a,b] */
eps=1E-8;
call quad(p, "HFunc", {0 1}, eps);
p = round(p, 0.001);

/**computing p with follow-up time tau**/
%if "&dist" = "weibull" %then %do;
    S1 = exp(-lambda1*(tau**k1)); S2 = exp(-lambda2*(tau**k2));
%end;
%else %if "&dist" = "lognormal" %then %do;
    S1 = 1-CDF('LOGNORMAL', tau, lambda1, k1);
    S2 = 1-CDF('LOGNORMAL', tau, lambda2, k2);
%end;
%else %if "&dist" = "burrIII" %then %do;
    S1 = 1-(1+(tau**(-lambda1)))**(-k1);
    S2 = 1-(1+(tau**(-lambda2)))**(-k2);
%end;

%if "&copula" = "clayton" %then %do;
    p1 = (S1**(-theta)+S2**(-theta)-1)**(-1/theta)/2;
%end;
%else %if "&copula" = "gumbel" %then %do;
    p1 = exp(-((-log(S1))**(theta+1)
        +(-log(S2))**(theta+1))**(1/(theta+1)))/2;
%end;
%else %if "&copula" = "frank" %then %do;
    p1 = -1/theta
        *log(1+(exp(-theta*S1)-1)*(exp(-theta*S2)-1)/(exp(-theta)-1))/2;
%end;
    call quad(p2, "HFunc", S2 || 1, eps);
    p_tau = round(p1+p2, 0.001);
    print distribution lambda1 k1 lambda2 k2
        copula theta tau p p_tau;
quit;
%mend mymacro;

```

参考文献

- Dobler, D. and M. Pauly (2018) “Bootstrap- and permutation-based inference for the Mann-Whitney effect for right-censored and tied data,” *TEST*, Vol. 27, No. 3, pp. 639–658.
- Efron, B. (1967) “The two sample problem with censored data,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 4, pp. 831–853.
- Emura, T. and J.-H. Hsu (2020) “Estimation of the Mann-Whitney effect in the two-sample problem under dependent censoring,” *Comput. Statist. Data Anal.*, Vol. 150, pp. 106990, 17.
- Fay, M. P., E. H. Brittain, J. H. Shih, D. A. Follmann, and E. E. Gabriel (2018) “Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments,” *Stat. Med.*, Vol. 37, No. 20, pp. 2923–2937.
- Greenland, S., M. P. Fay, E. H. Brittain, J. H. Shih, D. A. Follmann, E. E. Gabriel, and J. M. Robins (2020) “On causal inferences for personalized medicine: how hidden causal assumptions led to erroneous causal claims about the D -value,” *Amer. Statist.*, Vol. 74, No. 3, pp. 243–248.
- Hand, D. J. (1992) “On Comparing Two Treatments,” *Amer. Statist.*, Vol. 46, No. 3, pp. 190–192.
- Nakazono, K., Y.-C. Lin, G.-Y. Liao, R. Uozumi, and T. Emura (2024) “Computation of the Mann-Whitney Effect under Parametric Survival Copula Models,” *Mathematics*, Vol. 12, No. 10.
- Nelsen, R. B. (2006) *An introduction to copulas*, Springer Series in Statistics: Springer, New York, 2nd edition, pp.xiv+269.
- Schneider, S., R. C. P. dos Reis, M. M. F. Gottselig, P. Fisch, D. R. Knauth, and A. Vigo (2023) “Clayton copula for survival data with dependent censoring: an application to a tuberculosis treatment adherence data,” *Stat. Med.*, Vol. 42, No. 23, pp. 4057–4081.
- Shih, J.-H., Y. Konno, Y.-T. Chang, and T. Emura (2022) “Copula-Based Estimation Methods for a Common Mean Vector for Bivariate Meta-Analyses,” *Symmetry*, Vol. 14, No. 2.
- Sklar, M. (1959) “Fonctions de répartition à n dimensions et leurs marges,” *Publ. Inst. Statist. Univ. Paris*, Vol. 8, pp. 229–231.
- Yamaguchi, Y. and K. Maruo (2019) “Bivariate beta-binomial model using Gaussian copula for bivariate meta-analysis of two binary outcomes with low incidence,” *Jpn. J. Stat. Data Sci.*, Vol. 2, No. 2, pp. 347–373.
- 中藪孝輔・魚住龍史・江村剛志 (2024) 「パラメトリックコピュラモデルを用いた相関のある生存時間データにおける Mann-Whitney 効果の推定」, under review.
- 塚原英敦 (2003) 「接合分布関数とその応用」, 『応用統計学』, Vol. 32, No. 2, pp.77–88.
- 大橋靖雄・浜田知久馬・魚住龍史 (2016) 『生存時間解析 応用編 –SAS による生物統計–』, 東京大学出版会.
- 大橋靖雄・浜田知久馬・魚住龍史 (2022) 『生存時間解析 第2版 –SAS による生物統計–』, 東京大学出版会, 第2版.
- 武富奈菜美・山本和嬉 (2023) 「生存時間解析・信頼性解析のための統計モデル」, 『日本統計学会誌』, Vol. 52, No. 2, pp.69–112.
- 江村剛志・大庭幸治 (2024) 「生存時間変数に対する代替性評価 —メタアナリシスアプローチ—」, 『計量生物学』, Vol. 45, No. 1, pp.67–85.
- 矢田真城・浜田知久馬 (2014) 「SAS を用いたコピュラに従う擬似乱数の生成」, 『SAS ユーザー総会論文集 2014』, pp.643–656.

連絡先:

中藪孝輔

〒190-8562 東京都立川市緑町 10-3

統計数理研究所医療健康データ科学研究センター

Email: nakazono@ism.ac.jp

自主防災組織のICTによる支援の現状と課題

○有馬昌宏、川向肇

(兵庫県立大学)

Current State and Possibility of ICT Support for Voluntary Disaster Prevention Organizations

Masahiro Arima and Hajime Kawamukai

University of Hyogo

要旨

防災対策は自助、共助、公助に分類されるが、共助を担う地域組織として、自主防災組織が存在する。自主防災組織は、地域住民の隣保協同の精神に基づく自発的な防災組織であり、令和5年4月1日現在で166,923の自主防災組織が設置され、活動カバー率（全世帯数のうち、自主防災組織の活動範囲に含まれている地域の世帯数の割合）は85.4%に達しているとされている。しかし、過去の調査では、自主防災組織に加入して活動をしているとの認識を有している住民は少なく、自主防災組織を災害時に有効に機能させるには課題が山積であるとされている。本発表では、過去の調査結果を踏まえ、2023年8月から9月にかけて実施した「防災意識と地域の自主防災組織に関するウェブ調査」（有効サンプル3,027人）に基づき、回答者の災害リスクの認識状況、実施している防災対策、近隣との付き合いの状況などが自主防災組織への参加意識や活動状況にどのような影響を及ぼしているか、また自主防災組織を有効に機能させるために必要と考えられるスマートフォンなどのICTの活用状況と今後の課題について、JMPを使用して分析した結果を紹介する。

キーワード：自主防災組織、ICT利活用、全国ウェブ調査、組織機能化、組織活性化

1. はじめに

災害大国と自称される我が国では、1959年の伊勢湾台風による甚大な被害を契機として、1961年11月に成立した災害対策基本法において、市町村の責務として「住民の隣保協同の精神に基づく自発的な防災組織の充実を図り、市町村の有するすべての機能を十分に発揮するように努めなければならない」（第5条）とされ、加えて住民等の責務として「住民は、防災に寄与するように努めなければならない」（第7条）と規定されたことを出発点として組織化が始まった「自主防災組織」は、従来の自助と公助に加えて、共助を担う組織として、継続的に育成強化が図られてきている。令和5年版消防白書によれば、2023年4月1日時点では、全国1,741市区町村のうち、1,692市区町村で166,923の自主防災組織が組織され、その活動カバー率（全世帯数のうち、自主防災組織の活動範囲に含まれている地域の世帯数の割合）は85.4%に達している、

また、自主防災組織を育成するため、2022年度には、1,005市区町村において資機材の購入費及び運営費等に対する補助が、178市区町村において資機材等の現物支給が行われており、これらに要した経費は、2021年度の34億688万円から33億9,308万円へと増加している。

過去の消防白書で確認すると、昭和54年(1979年)版で自主防災組織の組織数(14,709)と活動カバー率(当時は組織率と定義しており19.6%)が示されるようになり、昭和56年(1981)版で自主防災組織の都道府県別結成状況の一覧表が掲載されるようになり、昭和57年(1982年)版では昭和56年度(1981年度)の資機材導入に対する支出が33億3315万円であることが示されている。

以上の経緯から、50年以上にわたって名目値の総額で1,500億円前後の助成金が自治体から支出されて自主防災組織の育成強化が図られてきていることが窺える(有馬・川向[1])。

しかし、この自主防災組織に関しては、令和4年(2022)版消防白書に「自主防災活動が効果的かつ組織的に行われるためには、地域ごとに自主防災組織を整備し」、「自主防災組織等の活動の活性化が地域防災力の更なる充実強化に当たり重要であるとの認識から」との記載があるように、自主防災組織の活性化と効果的かつ組織的な活動が求められているにも関わらず、自主防災組織が有効に機能した事例については令和4年版と令和5年版の消防白書のいずれにおいても、「自主防災組織が、地域住民の中心となってハザードマップの作成や避難訓練を実施するなど、日頃から地域防災力の向上に努めていた結果、地域住民の避難が適切に行われ、被害の軽減につながった事例もある」との具体的事例に欠けた紹介にとどまっている。成果や効果に関する具体的なエビデンスがないままでは、隣保協同の精神に基づく共助目的の自主防災組織の必要性は認められても、地域住民の積極的参加には繋がりにくい状況をもたらす要因の一つとなっているのではないかと考えられる。

また、令和4年版消防白書と令和5年版消防白書では、特集として「消防防災分野におけるDXの推進」と題する章が設けられているが、防災関連では、令和4年版では「消防教育訓練等におけるDXの推進」としてVRの消防学校での利用が提言され、令和5年版では「マイナンバーカードを活用した救急業務の迅速化・円滑化」として救急隊が搬送先医療機関の選定等を行う際に傷病者のマイナンバーカードを活用して過去の受診歴や服薬状況などの医療情報を取得して救急業務の迅速化や円滑化を図る取組が紹介されているだけで、しかもマイナンバーカードの読み取りなどでかえって現場滞在時間が長くなっている実証実験結果も示されており、DXの推進が軌道に乗っているとは言い難い状況である。加えて、「防災・危機管理eカレッジ」としてインターネット上で防災関連の動画が提供はされているものの、国立研究開発法人科学技術振興機構の助成金を得て有馬[2]で原型システムを提示したマイナンバーカードを活用する避難所の入退所管理や有馬他[3]で同じく原型システムを提示したスマートフォンを活用する安否確認のシステム、同じく有馬・川向[4]で提案したメタバースを利用する住民向けの仮想空間での体験型防災教育の検討などは全く考慮されていないのが現状である。

総務省[5]の令和6年版情報通信白書によれば、インターネットなどに接続するための端末について、2023年の情報通信機器の世帯保有率は、「モバイル端末全体」で97.4%であり、その内数である「スマートフォン」は90.6%に達している。このような状況のもと、かつての固定電話の普及率も低く、都市部では木造の住宅が密集する木造住宅密集地域(密集市街地)が拡がり、家庭用小型消火器もエアゾール式簡易消火具もカセットボンベ式ガス調理器なども存在しなかった昭和50年代の防災白書に記載されているようなバケツリレーによる初期消火、炊き出し、戸別訪問の情報伝達などの訓練を継続しては、自主防災組織への住民、特に若年層の参加は期待できず、自主防災組織の有事の際の機能化も難しく、現状が継続すれば、自主防災組織は名ばかりの組織になりかねないと危惧される状況である。

以上の認識のもと、本研究では、自主防災組織への参画が求められている住民へのウェブによる調査を実施して、その回答を分析することで、自主防災組織の現状と ICT を利活用する自主防災組織の DX の実現に向けての課題の明確化を試みる。

2. 自主防災組織の現状と課題に関する先行研究

自主防災組織の活性化と機能化に向けては、これまでも国や自治体による様々な調査が自主防災組織、自主防災組織の育成強化にあたる基礎自治体、および住民を対象に実施されてきている。

自主防災組織および基礎自治体に対する直近の全国規模の調査としては、消防庁の 2 つの検討会による調査を挙げることができる。

自主防災組織等の充実強化方策に関する検討会[6]の 2016 年の調査では、1,000 の自主防災組織等（女性防火クラブと少年消防クラブを含む、回答者は代表か役員）と 47 都道府県、および都道府県を介しての市町村と消防本部を対象に郵送調査で実施され、633 の自主防災組織等、47 都道府県、1,144 市町村、282 消防本部からの回答があり、集計結果をまとめているが、自主防災組織等からの回答からは、

- ①約 8 割が町内会単位で組織され、
- ②女性役員がない組織が約 2 割、
- ③代表の年齢が 60 歳以上の組織が約 8 割、
- ④避難勧告等（2021 年以降は避難指示）があった場合にとるべき行動を決めていない組織が約 5 割、
- ⑤構成員のうち高齢者の割合が 50%以上の組織が約 2 割、
- ⑥参加者やリーダーの不足の人材面が課題となっている組織が約 5 割、

といった結果が示されている。

また、自主防災組織等の地域防災の人材育成に関する検討会[7]の 2018 年の調査では、全国の都道府県・市区町村と 668 の自主防災組織等を対象にメールと郵送で実施され、412 組織からの回答に基づき、

- ①教育・訓練の構成員に占める参加者の割合（1 回あたり）が 10%未満の団体は 57.4%、
- ②教育・訓練への参加状況に関しては参加者、特に若年層の参加者が少ないことが課題、
- ③内容や運営に関しては、内容のマンネリ化（42.9%）、準備や当日の世話などの負担の大きさ（23.0%）、教育・訓練の効果の小ささ（21.5%）が課題、

などが集計結果および課題として示されている。

一方、住民を対象とした全国調査では、内閣府による 1999 年の「防災と情報に関する世論調査」（サンプルサイズ 3,000、有効回答者数 2,122（70.7%）、内閣府[8]）と 2003 年の「消防・救急に関する世論調査」（サンプルサイズ 3,000、有効回答者数 2,113（70.4%）、内閣府[9]）、および 2012 年の「消防に関する特別世論調査」（サンプルサイズ 3,000、有効回答者数 1,864（62.1%）、内閣府[10]）が挙げられる。

1999 年調査では、自主防災活動に関して、自主防災組織への帰属意識や参加状況は設問されていないが、町内会などの地域単位の活動への参加率は 17.2%、自主防災活動への非参加率は 70.4%であった。

自主防災組織への関心を設問した 2003 年調査では、「非常に関心を持っている」が 18.9%、「どちらかといえば関心を持っている」が 43.9%、「あまり関心を持っていない」が 31.2%、「全く関心を持っていない」が 3.0%という結果となっていたが、自主防災組織への帰属意識や活動状況は設問されていなかった。

また、2012 年調査では、消防団と自主防災組織に焦点を当てた設問がされており、自主防災組織に関しては、自主防災組織に対する関心度と自主防災組織への参加に対する意識が設問されている。自主防災組織に対する関心度では、「関心を持っている」が 45.8%、「どちらかといえば関心を持っている」が 35.%で「関心

を持っている」の小計が 81.7%, 「どちらかといえば関心を持っていない」が 11.3%, 「関心を持っていない」が 5.7%で, 「関心を持っていない」の小計が 17.0%で, 東日本大震災の 1 年後の調査ということも影響してか, 自主防災組織に対しての関心が高いという結果が示されている。

しかし, その後は, 直近で内閣府の世論調査として令和 4 年 (2022 年) 9 月に「防災に関する世論調査」が実施されているが, この調査も含めて内閣府の世論調査で自主防災組織に関する設問はなされていない。

このような状況で, 有馬他[11]は, 科学研究費補助金の助成を受けて 2011 年に自主防災組織の現状を住民の組織への帰属意識と参加状況の視点から把握するための応募型のウェブ調査を実施し, 7,133 人の有効回答から, 防災訓練などへの参加率は 45.0%であるにもかかわらず, 自主防災組織に加入していると認識している回答者の比率 (加入意識率) は 9.2%で, 74.4%の活動カバー率との間には大きな乖離があることを示している。

3. 自主防災組織に関するウェブ調査の概要

上述のように住民の自主防災組織に対する帰属意識や参加状況に関する全国規模の先行調査がほとんどないことから, 本研究では, 有馬他[11]の調査に引き続き, 自主防災組織の現状と課題について, 組織化の対象となる住民の帰属意識や参加意欲の現状を明らかにすることを目的に, 以下に示すウェブ調査を実施した。

調査名: 防災意識と地域の自主防災組織に関するウェブ調査

調査期間: 2023 年 8 月 20 日から 9 月 21 日の 33 日間。

調査委託機関: 株式会社データサービス (西宮市)。

調査対象: 日本に居住する 18 歳以上の住民。

調査方法: 自由応募型で, 目標サンプルサイズは 3,000 以上, 性別, 年齢層別, 居住都道府県別で必要最小サンプルサイズを条件として設定。

調査内容: 性別, 年齢, 居住市区町村, 職業, 家族構成, 住居形態・築後年数, 居住年数, 自宅の災害リスクの種別と判断根拠, 自治会の有無と自治会活動内容, 自主防災組織の認知と関心と加入意識, 加入の場合の対象災害種別や活動内容や参加頻度や ICT の利活用状況や活動の有効性評価など, 加入以外の場合の非加入理由や参加促進に有効な手段など, 防災訓練への参加状況, 地区防災計画の認知度, 防災アプリの利用状況や気象庁の「キキクル」および国土交通省の「浸水ナビ」の認知・利用状況など, 自宅で実施の防災対策, 過去の立退き避難経験, 近所との付き合いのレベル, インターネット上のサービスの利活用状況など。

有効回答者数: 33 日間の調査実施期間中の回答サイトへの訪問者数は延べで 19,380 人, 回答数は報償のインセンティブ目的の悪意のある複数回答者を除いて 3,382 人, 有効回答者数は 3,207 人。

回答者の属性: 男性 53.5%, 女性 45.5%, 30 歳未満 5.2%, 30 歳代 13.7%, 40 歳代 21.0%, 50 歳代 23.9%, 60 歳代 22.7%, 70 歳以上 13.4%。

図 1 に示すように, 40 歳代以下は女性の構成比率が高く, 50 歳代で男女の構成比率がほぼ等しくなり, 60 歳代以上では男性の構成比率が高く, 70 歳以上では圧倒的に男性の構成比率が高くなっている。

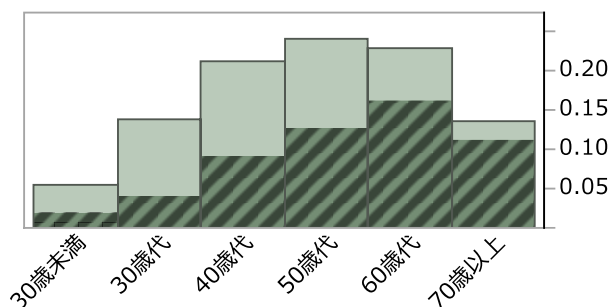


図 1 回答者の性別と年齢の構成 (網掛け部分が男性)

4. 住民から評価した自主防災組織の現状と課題

4.1 自主防災組織の知名度と関心度

自主防災組織が活性化し、有事に効果的かつ効率的に機能するためには、自主防災組織の存在そのものが認知されていなければならない。

自主防災組織の知名度は、回答者全体では「知っている」が52.5%、「知らない」が47.5%であったが、図2に示すように、年齢階級別の知名度は、30歳未満で41.7%、30歳代で44.2%、40歳代で49.6%、50歳代で49.5%、60歳代で57.3%、70歳以上で67.1%を示しており、高齢になるにつれて知名度が高くなる傾向が示されている。性別では、55.0%の男性に対して女性は49.9%で、知名度は男性の方が高い。

自主防災組織に対する関心度については、「非常に興味を持っている」が9.0%、「どちらかといえば興味を持っている」が37.5%、「あまり興味を持っていない」が42.1%、「全く興味を持っていない」が6.8%、「その他・わからない」が4.7%となっている。ただし、年齢階級別では、図4に示すように、「非常に興味を持っている」と「どちらかといえば興味を持っている」を併せると、30歳未満(50.0%)と30歳代(50.6%)の年代では50%を越えているが、40歳代(46.8%)、50歳代(41.1%)、60歳代(40.6%)と加齢が進むにつれて関心度は低下し、70歳以上で59.7%を示し、退職して年金生活となって地域での生活時間が長くなることで関心度が回復している状況が窺える。この結果は、自主防災組織の構成員の高齢化が進み、若年層の参加が課題であるという自主防災組織の代表者や役員の現状認識に対して、30歳代以下の若年層では、知名度は他の年

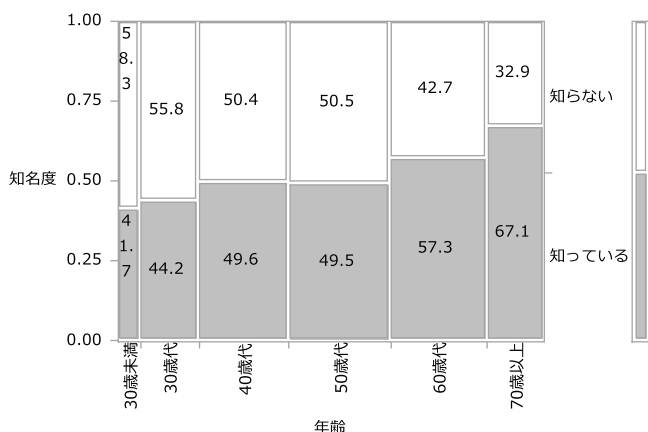


図2 知名度と年齢の関係

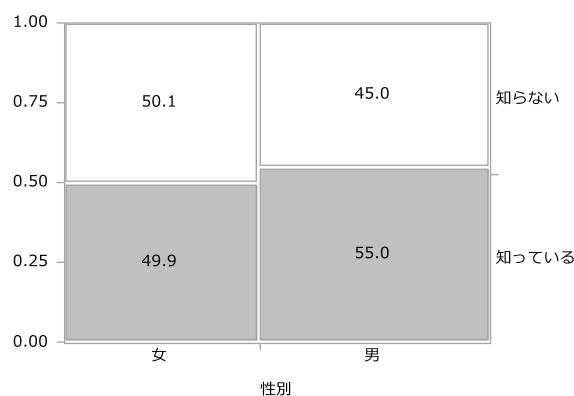


図3 知名度と性別の関係

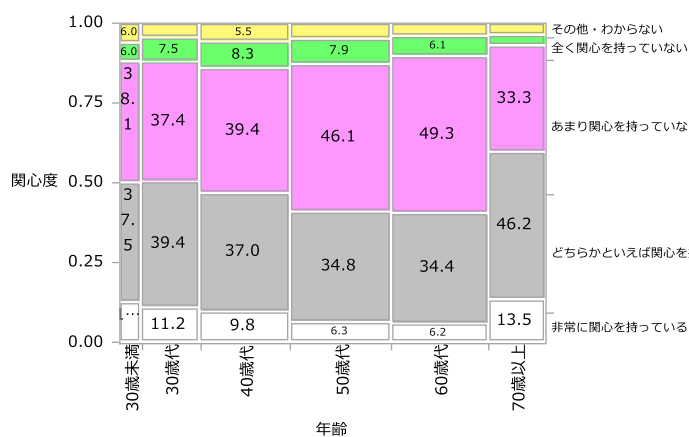


図4 関心度と年齢の関係

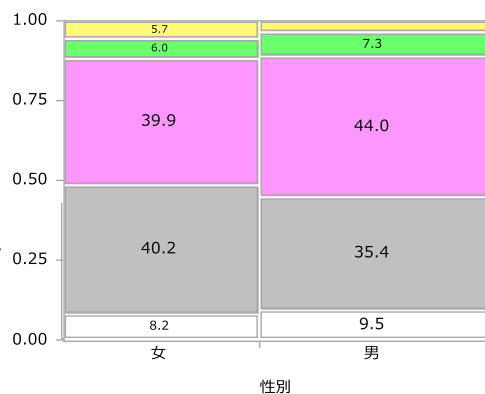


図5 関心度と性別の関係

年齢層よりも低いものの、関心度は高く、自主防災組織への参加のきっかけを設ければ、若年層の参加も見込めるのではないかと期待が持てる結果を示していると言えよう。また、「非常に関心を持っている」と「どちらかと言えば関心を持っている」を合わせた関心度の比率は、女性が 48.4%で男性の 44.9%を上回っており、女性の役員が少ないという状況は女性の関心度が低いためではないことが示唆されており、生理や授乳などで被災後の避難生活には女性の視点が特に重要で、男女共同参画社会の実現が自主防災組織では率先して実現されるべきであるのに旧来型の男性中心主義が継続されている現状の改革の可能性は示されていると言える。

4.2 自主防災組織の参加意識度と参加意識度を高めるための方策

自主防災組織の活性化と機能化に向けては、特に活動を積極的に担う若年層の自主防災組織への参加、すなわち若年層の自主防災組織への参加意識の高さが必要となる。

しかし、自主防災組織への加入意識率では、回答者全体では自主防災組織に「入っている」が 12.6%、「入っていない」が 48.0%、「(居住地域に) 自主防災組織はない」が 19.6%、「わからない」が 19.8%であり、自主防災組織の活動カバー率 85.4%であること、換言すると全国の世帯の 85.4%は自主防災組織が組織化されている地域に居住しているとされているという消防白書の報告と比較して、実際の住民の参加意識率は非常に低い状況であることが示されている。すなわち、自主防災組織が名目上でカバーしているとしている地域でも、その地域の居住者の多くは、自主防災組織に加入して平時はもとより、災害時に共助の活動を自主防災組織として行おうという意識は希薄であることが示されているのである。ただし、近年の消防白書では必ず言及されている「阪神・淡路大震災においては、地域住民が協力し合って、初期消火により延焼を防止した事例や、救助活動により人命を救った事例等が数多くみられた」とする説明と、付随の図としての阪神・淡路大震災における生き埋めや閉じ込められた際の救助（あるいは脱出）の状況の棒グラフでは、「自力で」が 34.9%と「家族に」が 31.9%で自助が 66.8%、「友人に・隣人に」が 28.1%と「通行人に」が 2.6%で共助が 30.7%、「救助隊に」という公助が 1.7%で、共助による救助の比率が高いことが示されているが、自主防災組織に参加していなくても、生き埋めや閉じ込めの現場に遭遇すれば救助しようとするのが人間であり、自主防災組織が組織されていて平時の救出訓練を行っているから救出率が高まるという因果関係は認められないのであり、因果関係があるとの説明もされていないのである。

しかし、平時に災害時の安否確認や閉じ込め事案の救出作業や避難所運営の訓練を行っておくことは、訓練を行っていない場合よりは被災者支援活動の効果は高まると考えられる。したがって、参加意識率を高めて、構成員に平時の自主防災組織の活動に参加してもらうことは重要である。

ここで、図 6 に示すように、自主防災組織に「入っている」との回答は、30 歳未満では 4.8%、30 歳代で 8.0%、40 歳代で 8.6%、50 歳代で 9.0%、60 歳代で 17.3%、70 歳以上で 24.6%を示しており、若年層での加入意識率が低く、自主防災組織を対象とする先行調査でも指摘されているように、自主防災組織の構成メンバーの高齢化ならびに参加者不足という課題が裏付けられている結果となっている。

また、性別では、図 7 に示すように、「入っている」との参加意識は、男性の 15.4%に対して女性は 9.4%で男性の方が 6 ポイント高く、参加しているかどうか「わからない」との回答は男性が 15.2%であるのに対して女性は 25.7%と 10 ポイントも高くなっており、女性の参加意識が低いという状況が窺える結果となっている。ただし、自主防災組織は世帯単位で加入しているとの認識があるとするれば、世帯主が活動に参加していれば、配偶者は「入っていない」との回答をする可能性があり、このような事情が男女間での参加意識度に差を生んでいる可能性は存在する。

なお、「(居住地域に) 自主防災組織はない」との 19.6%の回答率は、自主防災組織の活動非カバー率 (15.3%) とほぼ整合する結果となっている。

ところで、現状認識として、自主防災組織の若年層の参加率が低いという問題の存在を把握した上で、参加率が低いことを課題として単に嘆くのではなく、いかに若年層の参加率を高めていくかの課題解決のための方策を検討することが重要である。

図 8 と図 9 には、参加意識度に対して、知名度と関心度がどのように影響を及ぼしているかを検討するための知名度および関心度と参加意識度の関係を示すモザイク図を示している。

経営学のマーケティング分野の消費者購買行動モデルの代表的なものとして、マスメディア時代の AIDMA (Attention-Interest-Desire-Memory-Action), Web 時代の AISAS (Attention-Interest-Search-Action-Share), ソーシャルメディア時代の SIPS (Sympathize-Identify-Participate-Share&Spread) などがあるが、いずれのモデルでも、最初に商品・サービスの存在の認知が必要であり、我々の調査では、知名度が Attention を、関心度が Interest を、参加意識度が Action を測定する尺度となっている。知名度が低ければ関心度が低くなり、結果として参加意識度が低くなる。

実際、図 8 からは、「非常に関心がある」と「どちらかといえば関心を持っている」を合わせて、自主防災組織を知っている場合は 56.5%の関心度に対して、知らない場合の関心度は 35.4%で、大きな差が生じており、知名度が低ければ関心度が低くなることが示されている。また、図 9 からは、当然のことであるが、認知されていなければ興味・関心・共感を持ってもらう機会がなく、そのため活動に参加することはありえず、この過程の論理を反映して、自主防災組織を「知らない」回答者で「入っている」との回答者は 0.8%、自主

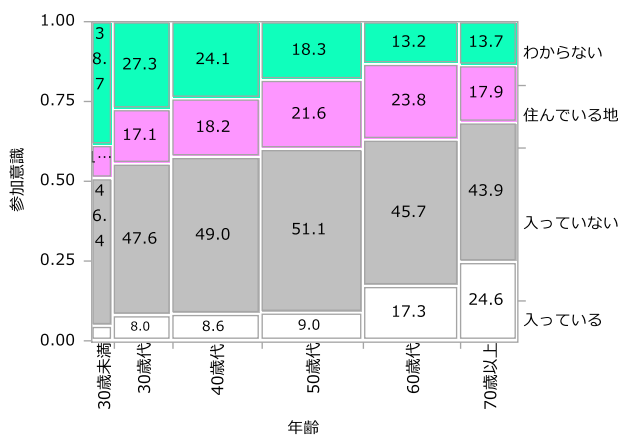


図 6 参加意識と年齢の関係

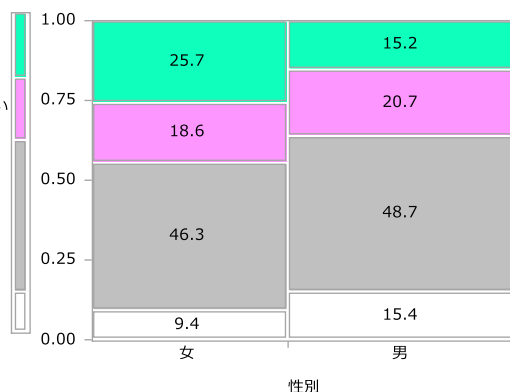


図 7 参加意識と性別の関係

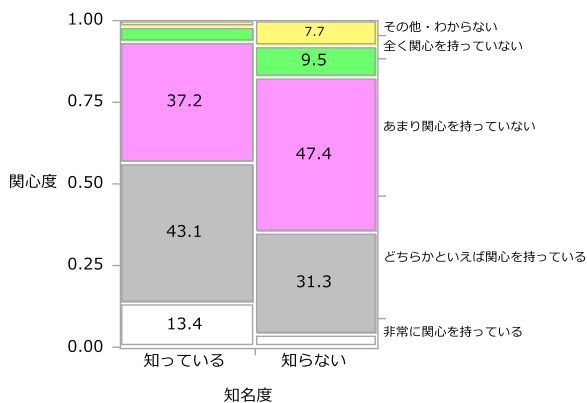


図 8 知名度と関心度との関係

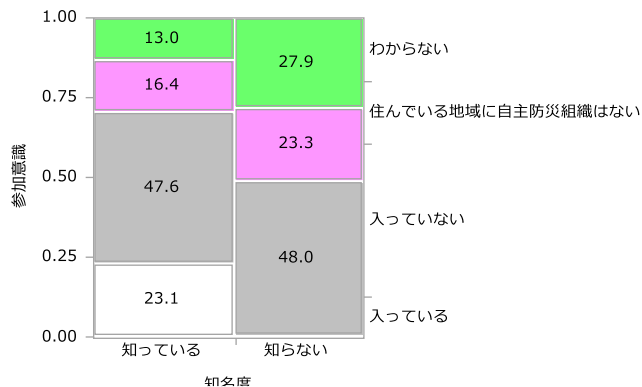


図 9 知名度と参加意識度との関係

防災組織を「知っている」との回答者で「入っている」との回答者は23.1%で、認知度が参加意識に大きな影響を及ぼしていることが分かる。

以上から、自主防災組織への参加を促すためには、まずは自主防災組織の存在を認知してもらい、さらに活動内容に関心・興味を抱いてもらうか、活動に共感してもらって活動の意義を理解してもらい、活動への参加と自主防災組織への帰属意識を持ってもらうというプロセスを機能させることが必要となる。

4.3 自主防災組織への加入意識のある住民の意識

今回のウェブ調査では、自主防災組織に「入っている」との回答者とそうでない回答者で異なる質問をしている。

自主防災組織に「入っている」との回答者（402人）には、

- 1) 加入している自主防災組織が対象とする災害種別（①地震 76.4%、②河川の氾濫 34.8%、③風害 23.8%、④土砂崩れ 14.5%、⑤津波 11.5%、⑥内水氾濫 10.3%、⑦不明 8.5%、母数は自主防災組織に「入っている」との回答者数、以下同様）
- 2) 活動内容（①消火訓練 57.0%、②食糧・飲料水等の備蓄 35.3%、③防災関連の講演会・講習会開催 33.3%、④発災後の安否確認 31.5%、⑤担架や AED 等を用いた救助・救命の実地の防災訓練 28.5%、⑥自治体による公的避難所開設支援 26.5%、⑦炊き出し 24.8%、⑧危険箇所の発見や防災地図作成 24.3%、⑨避難行動要支援者の避難所までの避難誘導 23.0%、⑩自治体の危機管理担当との連絡 20.8%、⑪自治体の公的避難所運営支援 19.3%、⑫危険接近時の避難情報の伝達 18.0%、⑬支援物資・食糧配布 17.5%、⑭自前の一時避難所開設・運営 16.8%、⑮災害ボランティアへの対応 8.0%、⑯わからない 6.3%、⑰防災士の育成・資格取得支援 6.0%、⑱DIG などゲーム感覚で学べる防災訓練 4.5%、⑲その他 1.3%）
- 3) 組織の活動の活発度（①活発である 11.4%、②まあまあ活動している 47.6%、③あまり活動していない 32.4%、④活動していない 3.3%、⑤わからない 5.3%）
- 4) 組織内での役割分担の認知度（①役割が決められていて知っている 37.3%、②決められているが具体的内容は知らない 17.5%、③以前は分担していたが今はしていない 27.8%、④わからない 17.3%）
- 5) 災害用機材保管場所の認知度（①知っている 77.1%、②知らない 20.1%、③資機材はない 2.8%）
- 6) 活動への参加度（①活動全てに参加した 14.0%、②殆ど参加した 23.3%、③まあまあ参加した 19.5%、④あまり参加していない 19.3%、⑤全く参加していない 24.0%）
- 7) あまり参加していないと全く参加していないの 43.3%の回答者の参加しない理由（複数回答、①仕事で時間がない 30.8%、②特に大きな災害の危険がない 20.3%、③体力がない 17.4%、④障害や持病等がある 13.4%、⑤家族の世話で時間がない 12.2%、⑥活動が大変そう 8.7%、⑦関心がない 7.6%）
- 8) 組織の活動の評価（①大いに評価できる 14.4%、②まあ評価できる 62.6%、③あまり評価できない 12.6%、④全く評価できない 2.5%、⑤わからない 7.8%）
- 9) 情報通信技術やサービスの利用（①利用している 33.7%、②利用していない 46.1%、③わからない 20.2%）
- 10) 情報通信技術やサービスを利用しているとの回答者での利用技術やサービスの内容（複数回答、①避難情報の連絡 59.8%、②日常の連絡 55.1%、③安否確認 43.3%、④避難開始の連絡 40.9%、⑤気象庁等の公的機関の災害情報収集 27.6%、⑥自治体の担当部署との連絡 24.4%、⑦避難所の入退所管理 20.5%、⑧河川等の監視カメラ確認 19.7%、⑨SNS 等による情報の受発信 17.3%、⑩避難行動要支援者の支援依頼 15.4%）
- ⑪発災時の組織の有用度評価で、避難誘導有用性（①役立つ 55.0%、②どちらとも言えない 30.1%、③役立

たない 7.5%, ③わからない 7.5%), 消火活動有用性 (①役立つ 50.9%, ②どちらとも言えない 33.7%, ③役立たない 8.2%, ④わからない 7.2%), 救命活動有用性 (①役立つ 45.0%, ②どちらとも言えない 37.0%, ③役立たない 10.1%, ④わからない 8.0%), 避難所運営有用性 (①役立つ 50.8%, ②どちらとも言えない 33.4%, ③役立たない 6.5%, ④わからない 9.3%),

を設問しており、上述の括弧内表記の単純集計結果からは、自主防災組織に加入しているという認識はあっても、役割の認識は不明確で、活動への参加は仕事や介護などの時間的制約や体力などが阻害要因となって活動頻度が低くなり、ICT や関連サービスが進展・普及しているにも関わらず、活動ではスマートフォンや QR コード等の ICT が利活用されていないなどの現状と課題が窺える結果が示されている。

4.4 自主防災組織への加入意識のない住民の意識

一方、自主防災組織に「入っている」という選択肢以外の回答者 (2,784 人) には、

- 1) 加入を求められた場合の対応 (①参加する 19.8%, ②参加しない 32.5%, ③わからない 47.7%),
- 2) 参加しない回答者への参加しない理由 (複数回答, ①活動が大変そう 38.7%, ②体力がない 37.3%, ③仕事で時間がない 35.7%, ④関心がない 26.9%, ⑤家族の世話で時間がない 15.4%, ⑥障害や持病がある 13.3%, ⑦特に大きな災害の危険がない 12.7%)

を設問しているが、報酬は自分や家族や隣人の命が守られることだけという地縁によるボランティアの活動に対して、時間的余裕と体力がないことに加えて、「活動が大変そうである」というイメージが、参加の認識のない居住者が自主防災組織の活動に参加するにあたっての阻害要因となっている状況が示唆されている。

4.5 自主防災組織の活性策と防災力向上策

自主防災組織の活動が活発ではないとされている状況のもと、自主防災組織の活動を活発にさせるための方策として市区町村に求める対策としては、複数回答で、①「自主防災組織の活動内容を自治体の広報紙などでもっと住民に PR する」が 52.4%, ②「活動内容を自治体のウェブサイトなどでもっと住民に PR する」が 32.4%, ③「消防署との連携を進める」が 29.7%, ④「自主防災組織に活動の場所や施設を提供する」が 26.2%, ⑤「自主防災組織にもっと財政援助を行う」が 23.5%, ⑥「消防団・水防団などとの連携を進める」が 22.0%, ⑦「活動内容を SNS でもっと住民に PR する」が 19.2%, ⑧「近隣の事業所の防火・防災組織との連携を進める」が 18.3%, ⑨「自主防災組織の指導者に対して教育を行う」が 17.3%, であり、自主防災組織を対象とした調査の結果と比較して、組織のリーダー養成に対する期待が低い結果となっている。

自主防災組織の活性化や機能化に加えて、地域の防災力を高めるには、行政による計画である「地域防災計画」(公助に相当)に加えて、地域住民が自発的に立案するコミュニティレベルでの防災計画である「地区防災計画」(自助と共助に相当)を策定することが有効であると考えられるが、この「地区防災計画」については、①「知らない・初めて聞いた」が 67.1%, ②「名前は聞いたことがあるが具体的内容は知らない」が 21.4%, ③「名前は聞いたことがあり具体的内容もある程度は知っている」が 6.9%, ④「名前は聞いたことがあり具体的内容もよく分かっている」が 2.5%, ⑤「居住地域で地区防災計画の策定を検討・策定中」が 1.2%, ⑥「地区防災計画を策定済み」が 1.8%であり、住民への周知が欠けている状況が明らかとなっている。

5. おわりに

地震、洪水、土砂災害などの自然災害が頻発する我が国では、隣保協同の精神に基づく共助での防災活動

は必要不可欠である。しかし、災害対策基本法に共助のための組織として自主防災組織が規定されて 60 年以上が経過した現在、自主防災組織の活動カバー率は 85.4%にまで高まったものの、消防庁で検討会が設置されて対応が議論されてきているように、構成員の高齢化やリーダー不足による自主防災組織が有事に効果的かつ効率的に機能しない可能性が危惧され、対策が検討されている。

しかし、その対策を検討するための現状把握のための調査は、国や自治体で実施されてはいるものの、自主防災組織の代表者や役員を対象とするもので、自主防災組織を構成して活動すべき地域の一般住民がどのように感じているかは明らかにできず、その結果として、有効な対策を講じることができていないのではないかと考えられる。

以上の問題意識をもって、本稿では、自主防災組織のリーダーからの視点ではなく、構成員として実働する住民の視点から自主防災組織の現状と課題を明らかにして自主防災組織を活性化させ、機能化させる策を見出すべく、自主防災組織に関する全国ウェブ調査の結果を紹介した。ただし、自主防災組織への参加の有無を目的変数とするロジスティック回帰分析や分類木によるディンジョンツリーを適用した分析などの結果は、調査で都市部であるか地方であるかなどの居住地の特性が不明であるために意味ある結果が得られていないので、本稿では割愛している。

しかし、単純集計やクロス集計を中心とした分析ではあっても、自主防災組織の組織構造や活動内容は、1960 年代から 1970 年代にかけての昭和の時代の情報通信技術と自然災害に対して脆弱であった時代の建築水準（旧耐震基準）や木造住宅密集地域（密集市街地）などの都市の空間構造を前提としたものが継続されてきているように観察され、建物の耐震化や地域の不燃化が進み、ICT が進展して普及している DX の時代にはそぐわない内容となっている状況は明確に示唆されているように思われる。

今後は、調査の回答データの詳細な分析を行い、DX 時代にふさわしい「有事に効果的に機能する自主防災組織」への転換が可能なような提言と自主防災組織の活動を支援する情報システムやアプリの開発を行っていきたい。

注

本稿は、参考文献に挙げた研究成果の[12]と[13]と[14]をもとに、加筆修正を加えたものである。また、本稿で参照している各年度版の消防白書の出典を列挙すると紙数を圧迫するので、参照した各年度の消防白書は参考文献には挙げていない。消防白書については、消防庁のサイト <https://www.fdma.go.jp/publication/#whitepaper> および国立国会図書館の「インターネット資料収集保存事業」のサイト <http://warp.da.ndl.go.jp/>で検索することで閲覧可能である。

謝辞

本研究は、科学研究費補助金（20K05031：「ソフト防災に資する防災情報の情報品質の向上と自主防災組織の活性化に関する研究」、研究代表者：有馬昌宏）の助成を受けている。

参考文献

- [1] 有馬昌宏・川向肇，ポストコロナ時代の自主防災組織の在り方に関する考察，日本情報経営学会第 85 回全国大会予稿集，pp. 49-52，2023.
- [2] 有馬昌宏，自主防災組織レベルでの災害時住民避難支援システムの構築の可能性，2010 年地域安全学会梗概集，No.26，pp.43-46，2010.

- [3] 有馬昌宏・田中宏明・陳斐龍・田中健一郎・亀井達樹・力宗幸男, 自主防災組織や学校などの地域組織の安否確認を支援するスマホアプリの開発, 2016年地域安全学会梗概集, No.39, pp.81-84, 2016.
- [4] 有馬昌宏・川向肇, 都市連動型メタバースの防災目的の利活用の可能性と課題, 日本情報経営学会 第84回全国大会予稿集, pp.93-96, 2022.
- [5] 総務省, 令和6年版情報通信白書, <https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/pdf/index.html>, 2024.
- [6] 自主防災組織等の充実強化方策に関する検討会, 自主防災組織等の充実強化方策に関する検討会報告書, https://www.fdma.go.jp/singi_kento/kento/items/kento189_27_houkokusyo.pdf, 2017.
- [7] 自主防災組織等の地域防災の人材育成に関する検討会, 自主防災組織等の地域防災の人材育成に関するアンケート調査結果(中間とりまとめ), https://www.fdma.go.jp/singi_kento/kento/items/jisyubousai_soshiki_iku sei23_sankou1.pdf, 2018.
- [8] 内閣府, 防災と情報に関する世論調査(平成11年6月調査), <https://survey.gov-online.go.jp/h11/bousai/>, 1999.
- [9] 内閣府, 消防・救急に関する世論調査(平成15年5月調査), <https://survey.gov-online.go.jp/h15/h15-shoubou/index.html>, 2003.
- [10] 内閣府, 消防に関する特別世論調査(平成24年8月調査), <https://survey.gov-online.go.jp/hutai/h24/h24-syoubou.html>, 2012.
- [11] 有馬昌宏・上野卓哉・有馬典孝, 全国ウェブ調査に基づく住民の自主防災組織の認知の現状と課題, 災害情報, No.10, pp.134-140, 2012.
- [12] 有馬昌宏・川向肇, 地縁によるボランティア組織である自主防災組織におけるICTの利活用状況と課題, 日本情報経営学会第86回全国大会予稿集, pp.65-68, 2023.
- [13] 有馬昌宏・川向肇, 自主防災組織の現状と課題, 地域安全学会梗概集, No.53, pp.159-162, 2023.
- [14] 有馬昌宏・川向肇, 自主防災組織の現状とICTによる支援の可能性, 第14回横幹連合コンファレンス予稿集, C-2-4, 2023.

JAFROC (Jackknife alternative free-response receiver operating characteristic) 解析のSASによる実装

○別府 汐¹

(¹イーピーエス株式会社 PMSセンター 統計解析部)

JAFROC (Jackknife alternative free-response receiver operating characteristic) analysis implementation in SAS

Ushio Beppu

Statistics Analysis Department, PMS center, EPS Corporation

要旨

JAFROC (Jackknife alternative free-response receiver operating characteristic) 解析は、近年発展目覚ましい分野である画像診断機器および AI 画像診断の性能を評価する際に頻繁に用いられる統計手法である^{[1][2]}。JAFROC 解析は専用ソフトウェアや R の RJafroc パッケージを用いることで実装することが出来るが、SAS においては専用のプロシジャは存在しない。本発表では JAFROC 解析の理論に基づき SAS のデータハンドリング手法および SAS/STAT の既存プロシジャを用いることで、SAS による JAFROC 解析の実装例を紹介する。実装例として 2 つの診断機器の病変検出能を比較する仮定の臨床試験をもとに、その評価値である FOM 値 (Figure of merit) を算出し、DBM (Dorfman-Berbaum-Metz) 法を用いて 2 つの診断機器間の比較を行う方法を紹介する。

キーワード : JAFROC, ROC 解析, GLM Procedure, 画像診断機器

1. 緒言

1.1 はじめに

画像診断機器の正確性を評価する際に使用される解析手法の 1 つとして ROC (Receiver Operating Characteristic) 解析が用いられており、診断機器による信号が付与された試料画像から読影者 (医師、放射線技師等) が病変の有無を判定し、その結果より得られる ROC 曲線の曲線下部分面積 (Area Under the ROC Curve: AUROC) の値によって性能を評価することが出来る^[3]。しかし通常の ROC 解析には以下の問題点が存在する。

- ・ 病変の位置については考慮されず、本来の病変と違う位置に信号が現れても正答とみなされる
- ・ 1 つの資料画像につき 1 個の病変の有無しか評価できない

この問題を解決するために、1 画像あたりに複数の信号が含まれた病変画像に対して、読影者が病変であると判定した箇所を自由にマークする FROC (Free-response ROC) 解析と呼ばれる手法が開発された^[4]。

JAFROC 解析は FROC 解析を発展させた手法であり、複数の読影者が参加することによる読影者毎の変動や試料画像内の病変数の多寡による影響を排除したうえで、2 つの診断機器の病変検出能を比較することが可能となる^[5]。

本稿では、JAFROC 解析の概要と SAS による JAFROC 解析の実装方法を報告するとともに、仮想の臨床試験例を用いて、実際に解析を行った例を報告する。

1.2 JAFROC 解析の理論的背景について

いま疾患群 N_1 人、非疾患群 N_2 人の計 N 人の被験者から、2 つの画像診断機器 i ($i = 1, 2$) でそれぞれ試料画像取得し、 J 人の読影者 ($J = 1, 2, \dots, j$) によって、画像診断機器間の病変検出能の差を比較する試験を考える。

JAFROC 解析では以下のような仮定を追加する。

- ・ 読影者は被験者の試料画像の中で病変であると判断した箇所を、何ヶ所でもマークすることが出来る
- ・ 読影者はマークした箇所に得点を与える (得点の与え方は試験によって異なる)
- ・ 被験者ごとの重みを均一にするため、疾患群の被験者については、それぞれの病変に対して臨床的重要度に応じて総和が 1 となるように重みが付けられる (この重みは読影者によらず一定)

以下に例を示す。診断機器 i によってイメージングされた被験者 s の試料画像に対して、読影者 j は病変と判断した箇所に 1~10 点の得点を与えるとすると、以下の Table 1.2.1 が作成される。

Table 1.2.1 病變得点のイメージ

診断機器	読影者	被験者	病変 ID	得点	重み
i	j	s	1	6	0.5
i	j	s	2	0 (病変と判定していない)	0.3
i	j	s	3	7	0.2

JAFROC 解析における診断機器の性能値 (Figure of Merit: FOM) は、以下の (1) 式で与えられる^[6]。

$$\theta_{ij} = \frac{1}{N_1 N_2} \sum_{s=1}^{N_1} \sum_{t=1}^{N_2} \sum_{m=1}^M W_{sm} \Psi(r_{ijsm}, \max(r_{ijt})) \quad (1)$$

$$\text{ただし } \Psi(r_{ijsm}, r_{ijt}) \begin{cases} 1 & r_{ijsm} > r_{ijt} \\ 0.5 & r_{ijsm} = r_{ijt} \\ 0 & r_{ijsm} < r_{ijt} \end{cases} \quad \text{および } \sum_{m=1}^M W_{sm} = 1$$

r_{ijsm} は診断機器 i 、読影者 j における疾患群の被験者 s の m 番目 ($1 \leq m \leq M$) の病変に与えられた得点、 $\max(r_{ijt})$ は診断機器 i 、読影者 j における非疾患群の被験者 t の得点の最大値を表す。なお、AUROC については疾患群と非疾患群の各被験者の試料画像に対して読影者 j が与えた得点について、Mann-whitney の U 検定

を行った際の U 統計量を N_1 と N_2 の積で除した値と一致することが知られており^[7]、(1) 式は更に各病変部位の重みを考慮した値であると解釈できる。

上記の FOM 値は診断機器、読影者ごとの値である。ここで被験者による影響を考慮するために、被験者 n ($n = 1, 2, \dots, N$) の全データを削除した上で (1) 式に基づき算出した $\theta_{ij(n)}$ を用いて、(2) 式に基づき被験者毎の疑似 FOM 値を算出する^[6]。

$$Y_{ijn} = N\theta_{ij} - (N - 1)\theta_{ij(n)} \quad (2)$$

ここまでで Jackknife 法により算出した疑似 FOM 値の算出方法について説明を行った。診断機器間の検出能を比較する際には診断機器、読影者、被験者およびこれらの交互作用を説明変数、疑似 FOM 値を目的変数とするモデル解析を行うが、その手法については 2 章以降に記載する。

1.3 JAFROC 解析のメリットとデメリット

通常の ROC 解析と比較した JAFROC 解析のメリットとして、ROC 解析では判別できない病変位置の誤答にも対応できるとともに、読影者間の変動や被験者による影響をモデルに組み込むことで 2 つの画像診断機器の検出能の差が有意であるかどうかを判定することが出来る点が挙げられる^[5]。

一方でデメリットとしては、疑似 FOM 値の算出方法が複雑かつ、モデル解析を用いることから解析を行うため、解析の実装方法が限られるという点が挙げられる。現在では R のパッケージ『RJafroc』等で解析を行うことが出来るが^[8]、SAS では専用のプロシジャは存在しない。

2. SAS による実装

2.1 仮想の臨床試験の設定とデータセットについて

本稿で用いる仮想の臨床試験のデータについては RJafroc に内蔵されている FROC 用のサンプルデータセット: dataset05 を用いる。試験の設定は以下のとおりである。

- ・ 2 種の診断機器の病変検出能を比較する
- ・ 読影者: 9 人
- ・ 被験者: 92 人 (疾患群 47 人、非疾患群 45 人)
- ・ 1 被験者につき、最大 3 箇所までマークされる

また、本データは以下の 3 つのデータセットから構成される。前処理として各症例について疾患群かどうかを判定するフラグ (DISFL) を持ったデータセット: Base を作成し、各データセットに結合しておく。

データセット: LL

変数	備考
ReaderID	読影者の ID (1 ~ 9)
ModalityID	診断機器の ID (1 ~ 2)
CaseID	被験者の ID (1 ~ 92)
LesionID	病変部位の ID
LLRating	病変部位の得点

データセット: NL

変数	備考
ReaderID	読影者の ID (1 ~ 9)
ModalityID	診断機器の ID (1 ~ 2)
CaseID	被験者の ID (1 ~ 92)
NLRating	偽病変部位 (実際には病変でない が信号が現れた箇所)の得点

データセット: TRUTH

変数	備考
CaseID	被験者の ID (1 ~ 92)
LesionID	病変部位の ID
Weight	病変部位の重み

2.2 Jackknife 法による疑似 FOM 値の算出

1.2 章にて AUROC は Mann-whitney の U 統計量と一致することを紹介した。SAS で Mann-whitney の U 統計量を算出する方法としては `npar1way` プロシジャ等が有名であるが、重みの項が含まれているため、通常の算出方法では算出できない。したがって本項では (1) 式に基づき、データハンドリングを用いて FOM 値を算出する。作成の方針は以下の通りである。

- ① 非疾患群について読影者、診断機器、被験者毎に最大の偽病変部位の得点を算出するとともに、マークが無い症例については得点を 0 で補完する。疾患群については読影者、診断機器、被験者、病変部位ごとにマークが無い部位については得点を 0 で補完する
- ② 読影者、診断機器ごとに r_{ijst} と $\max(r_{ijt})$ を比較するために、①で作成した疾患群のデータセット (LL_prejoin) と非疾患群のデータセット (NL_prejoin) 読影者と診断機器を Key として SQL の full join を用いて結合する

```

/*LL_prejoin: 疾患群の病変部位ごとの得点  NL_prejoin: 非疾患群の偽病変部位の得点*/
*READERID と Modality ID を key にして NLmax と LL_truth を fulljoin;
proc sql;
  create table wk12 as
  select A.*, B.ReaderID, B.ModalityID, B.CaseID as NLCaseID, B.NLRATING
  from LL_prejoin as A, NL_prejoin as B
  where A.ReaderID=B.ReaderID and A.ModalityID=B.ModalityID;
quit;
proc sort data=wk12; by  ReaderID ModalityID CaseID  lesionID NLCaseID; run;

```

- ③ (1) 式にしたがって、 θ_{ij} 値を算出する

```

data wk12_score;
  set wk12;
  by ReaderID ModalityID CaseID  lesionID NLCaseID;
*ポイント;
  if LLRATING > NLRATING      then point=1;
  else if LLRATING = NLRATING  then point=0.5;
  else if LLRATING < NLRATING  then point=0;
  Score=Weight*point;

```

```

*READERID と ModalityID 毎に加算していく;
    retain Jscore 0;
    if first.ModalityID                               then Jscore=Score;
    else                                               Jscore=Score + Jscore;
*FOM 値 (&obsd.: 疾患群の人数、&obsnd.: 非疾患群の人数);
    FOM=Jscore/(&obsd.*&obsnd.);
run;
data wk12_FOM;
    set wk12_score;
    by ReaderID ModalityID CaseID  lesionID NLCASEID;
    if last.ModalityID
run;

```

- ④ ②で作成したデータセット wk12 から n 番目の被験者 n を削除したデータセットを作成後、③と同様の処理を行い、その後 (1) 式にしたがって症例毎の疑似 FOM 値を算出する。

```

*Pseudo value の算出;
%macro pseudo();
    %do i=1 %to &CASE.;
        *症例の陽性陰性判定;
        data _null_;
            set BASE (where=(CASEID=&i));
            call symputx("DISFL",DISFL);
        run;

        data wk12_PS&i;
        set wk12_score (where=(CaseID^=&i and NLCASEID^=&i));
            by ReaderID ModalityID CaseID LesionID NLCASEID;
            PSID=&i;
        *READERID と Modality ID 毎の FOM 値を算出する;
            retain JSCORE_PS 0;
            if first.ModalityID then Jscore_PS=Score;
            else
                Jscore_PS=Score+Jscore_PS;
        /*          FOM_ps=Jscore/(&obsd.*&obsnd.);*/
            if &DISFL. = 0    then FOM_PS=Jscore_PS/(&obsd.*%eval(&obsnd.-1));
            if &DISFL. = 1    then FOM_PS=Jscore_PS/(%eval(&obsnd.-1)*&obsnd.);
            if last.ModalityID;
        run;
    %end;
%mpseudo;

```

```

%end;

data wk12_PSALL;
    set wk12_PS1 - wk12_PS&CASE.;
run;

%mend;
%pseudo;
*value と pseudo value をマージしてレスポンス値を出す;
proc sort data=wk12_FOM ; by ReaderID ModalityID; run;
proc sort data=wk12_PSALL; by ReaderID ModalityID; run;

data wk13;
    merge wk12_FOM (keep=ReaderID ModalityID FOM)
          wk12_PSALL (keep=ReaderID ModalityID FOM_PS PSID);
    by ReaderID ModalityID;
    *RESPONCE 値;
    RES=(&obsd.+&obsnd.)*FOM-%eval(&obsd.+&obsnd.-1)*FOM_PS;
run;

```

以下の Table 2.2.1 にこれまで述べた処理を行った結果の一部を示す。また、以下に RJafroc (R version 4.4.0, RJafroc_2.1.2 以降同様)のコードを示し、Table 2.2.2 に RJafroc に内蔵された関数を用いて算出した疑似 FOM 値を示す。この結果は SAS による算出結果と一致している。

RJafroc のコード

```

install.packages("RJafroc")
library(RJafroc)
#FOM 値出力
list1 <- UtilPseudoValues(dataset05, FOM = "wAFROC")$jkPseudoValues[1,1,1:15]
df1 <- data.frame(list1)

```

Table 2.2.1 SAS による疑似 FOM 値の算出

OBS	ModalityID	ReaderID	PSID	RES
1	1	1	1	0.35190
2	1	1	2	0.58292
3	1	1	3	0.27490
4	1	1	4	1.01563
5	1	1	5	1.01563
6	1	1	6	1.01563
7	1	1	7	1.01563
8	1	1	8	0.47291
9	1	1	9	1.01563
10	1	1	10	1.01563
11	1	1	11	0.58292
12	1	1	12	0.64893
13	1	1	13	0.58292
14	1	1	14	1.01563
15	1	1	15	1.01563

Table 2.2.2 R による実行結果

	list1
1	0.3519020
2	0.5829223
3	0.2748952
4	1.0156270
5	1.0156270
6	1.0156270
7	1.0156270
8	0.4729126
9	1.0156270
10	1.0156270
11	0.5829223
12	0.6489281
13	0.5829223
14	1.0156270
15	1.0156270

2.3 三元配置分散分析による 2 つの診断機器間の比較

これまでの章で述べた通り、JAFROC 解析では診断機器、読影者、被験者による変動を考慮した解析を行う。Dorfman らは診断機器を固定効果 (主効果)、読影者と被験者およびこれらの交互作用を変量効果とした三元配置分散分析を用いる Dorfman-Berbaum-Metz (DBM) 法を提唱した^[6]。

$$Y_{ijn} = \mu + \alpha_i + B_j + C_n + (\alpha B)_{ij} + (\alpha C)_{in} + (BC)_{jn} + \varepsilon_{ijn} \quad (3)$$

$$\sum_{i=1}^2 \alpha_i = 0$$

ここで μ は総平均、 α_i は診断機器 i の主効果、 B_j は読影者 j の変量効果、 C_n は被験者 n の変量効果、 $(\alpha B)_{ij}$ 、 $(\alpha C)_{in}$ 、 $(BC)_{jn}$ はそれぞれ診断機器 i と読影者 j 、診断機器 i と被験者 n 、読影者 j と被験者 n の交互作用、 ε_{ijn} は誤差項を表す。なお診断機器、読影者、被験者の 3 変数間の交互作用については誤差項に含まれる^[7]。Dorfman らは診断機器と読影者の交互作用の平均平方 $MS(\alpha B)$ 、診断機器と被験者の交互作用の平均平方 $MS(\alpha C)$ および誤差の平方平均 $MS(\varepsilon)$ について、 $MS(\alpha B) \leq MS(\varepsilon)$ かつ $MS(\alpha C) \leq MS(\varepsilon)$ が成り立つ場合、(3) 式から $(\alpha B)_{ij}$ および $(\alpha C)_{in}$ を除いたモデルを用いることを推奨し、2 つの診断機器に差がないとする帰無仮説に対する仮説検定 (F 検定) の F 統計量は(4) 式で表されると述べている^[6]。

$$F = \frac{MS(\alpha)}{MS(\varepsilon)} \sim F\{(I-1), ddf\} \quad (4)$$

$$\text{ただし } ddf = (I-1)(J-1)(N-1)$$

そして、 $(\alpha B)_{ij}$ および $(\alpha C)_{in}$ がモデルに残る場合（以下、特に断らない限り、この場合について述べる）、Dorfmanらは(5)式に示す Satterthwaite の近似 F 統計量を用いることを提唱している^[6]。このとき 2 つの診断機器に差がないとする帰無仮説に対する仮説検定 (F 検定)の F 統計量は (5) 式で表される。

$$F = \frac{MS(\alpha)}{MS(\alpha B) + MS(\alpha C) - MS(\epsilon)} \sim F\{(I-1), ddf_D\} \quad (5)$$

$$\text{ただし } ddf_D = \frac{\{MS(\alpha B) + MS(\alpha C) - MS(\epsilon)\}^2}{\frac{\{MS(\alpha B)\}^2}{(I-1)(J-1)} + \frac{\{MS(\alpha C)\}^2}{(I-1)(N-1)} + \frac{\{MS(\epsilon)\}^2}{(I-1)(J-1)(N-1)}}$$

なお $MS(\alpha)$ は診断機器の平均平方、 ddf および ddf_D はそれぞれ(4) 式、(5) 式に示した F 分布の自由度を表す。SAS では `random` ステートメントで変量効果に当たる変数および変量効果を含む交互作用項を指定し、さらに `test` オプションを使用することで、(5) 式の近似 F 統計量を得ることが出来る^[9]。コードを以下に示す。

```
*ModalityID:診断機器( $\alpha_i$ )、ReaderID:読影者( $B_j$ )、PSID:被験者( $C_n$ );
proc GLM DATA=wk13;
  class ModalityID ReaderID PSID;
  model RES =ModalityID ReaderID PSID ReaderID*ModalityID ModalityID*PSID ReaderID*PSID/ss3;
  random ReaderID PSID ReaderID*ModalityID ModalityID*PSID ReaderID*PSID / test;
  estimate 'ID1 - ID2' ModalityID 1 -1;
  ods OUTPUT ModelANOVA = ANOVA;
  ods OUTPUT RandomModelANOVA=RANOVA;
  ods OUTPUT Estimates=estimate;
  quit;
run;
```

以下の Table 2.3.1, Table 2.3.2Table 2.3.3 に実行結果を示す。なお、参考として、Table 2.3.4 には、上記のコードにおけるモデルの指定（`model` および `random` ステートメント）から、診断機器と読影者の交互作用および診断機器と被験者の交互作用を除いた場合における実行結果を示す。

Table 2.3.1 分散分析表
(ods OUTPUT ModelANOVA= で出力)

要因	自由度	Type III 平方和	平均平方	F 値	Pr > F
ModalityID	1	2.52054062	2.52054062	33.71	<.0001
ReaderID	8	13.49633191	1.68704149	22.56	<.0001
PSID	91	70.13689244	0.77073508	10.31	<.0001
ModalityID*ReaderID	8	0.68282675	0.08535334	1.14	0.3329
ModalityID*PSID	91	34.21044848	0.37593899	5.03	<.0001
ReaderID*PSID	728	91.67293479	0.12592436	1.68	<.0001

Table 2.3.2 2つの診断機器に差がないとする帰無仮説に対する仮説検定 (Satterthwaite の近似 F 統計量)
(ods OUTPUT RandomModelANOVA= で出力)

GLM プロシジャ
混合モデル分散分析に対する仮説検定
従属変数 : RES

要因	自由度	Type III 平方和	平均平方	F 値	Pr > F
ModalityID	1	2.520541	2.520541	6.52	0.0132
Error	60.449	23.364558	0.386516		
Error: MS(ModalityID*ReaderID) + MS(ModalityID*PSID) - MS(Error)					

Table 2.3.3 FOM 値の最小 2 乗平均値の差
(ods OUTPUT Estimate=で出力)

従属変数 : RES

パラメータ	推定値	標準誤差	t 値	Pr > t
ID1 - ID2	-0.07802732	0.01343947	-5.81	<.0001

Table 2.3.4 (3) 式から $(\alpha B)_{ij}$, $(\alpha C)_{in}$ を削除したモデルにおける 2 つの診断機器に
差がないとする帰無仮説に対する仮説検定

要因	自由度	Type III 平方和	平均平方	F 値	Pr > F
ModalityID	1	2.520541	2.520541	23.33	<.0001
ReaderID*PSID	728	91.672935	0.125924	1.17	0.0163
Error: MS(Error)	827	89.330531	0.108018		

Table 2.3.2 は近似 F 統計量を用いた F 検定の結果であり、誤差項として $MS(\alpha B) + MS(\alpha C) - MS(\varepsilon)$ が用いられているとともに、自由度は (5) 式に示した Satterthwaite の自由度となっている。参考として示した Table 2.3.4 の F 値は (4) 式に基づく。

$MS(\alpha B)$ および $MS(\alpha C)$ と $MS(\varepsilon)$ の大小関係については、Table 2.3.1 の F 値が 1 を上回っているかで判断することが出来る。いま診断機器と読影者の交互作用の F 値 = 1.14、診断機器と被験者の交互作用の F 値 = 5.03 であり、 $MS(\alpha B) > MS(\varepsilon)$ または $MS(\alpha C) > MS(\varepsilon)$ が満たされるため、Dorfman らの方法に従うと、Table 2.3.2 の結果が採用される。

Table 2.3.3 は 2 つの診断機器の FOM 値の最小 2 乗平均値の差を示している。なお最小 2 乗平均値の信頼区間については、以下の (6) 式で計算される [6]。

$$CI_{0.95} = (\bar{Y}_1 - \bar{Y}_2) \pm t_{0.025(df_D)} \sqrt{\frac{2}{JN} (MS(\alpha B) + MS(\alpha C) - MS(\varepsilon))} \quad (6)$$

\bar{Y}_1, \bar{Y}_2 は 2 つの診断機器の FOM 値の最小 2 乗平均値を表す。ただし、(6) 式の信頼区間については、SAS では直接算出することは出来ない。したがって以下に示したコードでデータステップを用いて算出する。

```

data Ranova_all;
    length ID $30;
    set Ranova (where=(SOURCE = "ModalityID") in=A)
        Ranova (where=(SOURCE = "ModalityID*ReaderID") in=B)
        Ranova (where=(SOURCE = "ModalityID*PSID") in=C)
        Ranova (where=(SOURCE = "Error: MS(Error)") in=D)
        estimate (in=E);
    *transpose の ID 用;
    if A=1 then ID="MS_Modality";
    if B=1 then ID="MS_Modality_Reader";
    if C=1 then ID="MS_Modality_Patient";
    if D=1 then ID="MS_Error";
    if E=1 then ID="Difference";
    *値;
    if whichn (1,A,B,C,D) > 0 then VAR=MS;
    if E=1 then VAR=Estimate;
run;
*proc transpose で横積みに;
proc transpose data=Ranova_all out=Ranova_col;
    var VAR;
    ID ID;
run;
data FIN_DBM;
    set Ranova_col;
    *分母の平均平方;
    MS=MS_Modality_Reader + MS_Modality_Patient - MS_Error;
    *自由度;
    DDF=(MS**2)/((MS_Modality_Reader**2)/(&READER.-1)+(MS_Modality_Patient**2)/(&CASE.-
1)+(MS_Error**2)/((&READER.-1)*(&CASE.-1)));
    *信頼区間;
    UCI2 = Difference + tinv(0.975,DDF)*SQRT(2/(&READER.*&CASE.)*MS);
    LCI2 = Difference - tinv(0.975,DDF)*SQRT(2/(&READER.*&CASE.)*MS);
    *F 値;
    FValue =MS_Modality/MS;
    *p 値;
    PValue = 1 - probf(FValue,1,DDF);
run;

```

以下の Table 2.3.5 に実行結果を示す。

Table 2.3.5 (6) 式に基づき算出した診断機器間の FOM 値の差の 95% 信頼区間

DBM法による信頼区間		
OBS	LCI2	UCI2
1	-0.13914	-0.016917

3. 補足事項 (Hillis による補正について)

2.3 章では診断機器、読影者、被験者およびこれらの交互作用をモデルに含んだ三元配置分散分析によって算出した近似 F 統計量を用いる DBM 法を紹介した。DBM 法は診断機器による効果を検定出来る一方で、診断機器と被験者の交互作用による平均平方項 $MS(\alpha C)$ に比べ、誤差による平均平方項 $MS(\epsilon)$ が大きい場合、自由度 ddf_D が 0 に近づくため、(6) 式に基づく信頼区間が非常に広くなることが言及されている^[10]。これに対して Hillis (2007) らは以下の (7) 式に基づく検定統計量と自由度を提案している^{[10][11]}。

$$F_H = \frac{MS(\alpha)}{MS(\alpha B) + \max\{MS(\alpha C) - MS(\epsilon), 0\}} \sim F\{(I - 1), ddf_H\} \quad (7)$$

$$\text{ただし } ddf_H = \frac{\{MS(\alpha B) + \max\{MS(\alpha C) - MS(\epsilon), 0\}\}^2}{\frac{\{MS(\alpha B)\}^2}{(I - 1)(J - 1)}}$$

SAS では F_H をプロシジャによって出力することは出来ないため、DBM 法に基づく信頼区間と同様にデータステップで算出を行う。以下にコードを示す。

```
*2.3 で作成したデータセット: Ranova_col を使用;
data FIN_Hill;
    set Ranova_col;
    *分母の平均平方;
    MS=MS_Modality_Reader + max(MS_Modality_Patient - MS_Error,0);
    *自由度;
    DDFH=(MS**2)/((MS_Modality_Reader**2)/(&READER.-1));
    *信頼区間;
    UCI2 = Difference + tinv(0.975,DDFH)*SQRT(2/(&READER.*&CASE.)*MS);
    LCI2 = Difference - tinv(0.975,DDFH)*SQRT(2/(&READER.*&CASE.)*MS);
    *F 値;
    FValue =MS_Modality/MS;
    *p 値;
    PValue = 1 - probf(FValue,1,DDFH);
```

```
run;
```

以下の Table 3.1 に実行結果を示すが、Table 2.3.5 と比較すると、 $MS(\alpha C) > MS(\epsilon)$ が満たされる状況では信頼区間の幅は、大きく異なることが確認できる。また RJafroc では p 値および信頼区間の算出方法について言及されていないが、上記の SAS の実行結果と一致することから、Hillis の補正が用いられていると推測できる(Table 3.2)。実際に Hillis の補正を利用する際には、参考文献^{[10][11][12]}をご確認いただきたい。

Table 3.1 Hillis による補正を行った解析結果 (DDFH:自由度 LCI2: 95%CI 下限 UCI2: 95%CI 上限)

Hillisの補正による検定結果

OBS	DDFH	FValue	PValue	LCI2	UCI2
1	164.053	6.52118	0.011571	-0.13836	-0.017695

RJafroc のコード

```
install.packages("RJafroc")
library(RJafroc)
StSignificanceTesting(
  dataset05,
  "wAFROC",
  method = "DBM",
  covEstMethod = "jackknife",
  analysisOption = "RRRC",
)
```

Table 3.2 RJafroc による解析結果

```

$RRRC$FTest
      DF      MS      Fstat      p
Treatment  1.00000  2.52054062  6.5211824  0.011570576
Error    164.05279  0.38651589      NA      NA

$RRRC$ciDiffTrt
      Estimate      StdErr      DF      t      PrGt      CILower      CIUpper
trt1-trt2 -0.078027318  0.030555086  164.05279 -2.5536606  0.011570576 -0.13835925 -0.017695389
```

4. 最後に

本稿では SAS によるデータハンドリング手法および SAS/STAT の既存プロシジャを用いることで、JAFROC 解析が実装出来ることを示した。JAFROC 解析のように特定の分野でのみ使用されるような解析については、SAS のプロシジャが存在しない場合があり、そのような場合専用ソフトウェアや幅広いパッケージを持つ R などに頼らざるを得ないことがある。しかし本稿の例で示したように、SAS には豊富なプロシジャが存在し、これらを適切に組み合わせることで、理論に基づいた解析を行って既存のソフトウェアと遜色ない解析を行うことが可能であるケースも存在する。

さらに、SAS を用いるメリットとして、データ構造が想定と異なってもデータハンドリングにより柔軟に対応できる点が挙げられる。また、モデル式に読影者の熟練度に対するパラメータ（読者経験年数、所属施設の方法等）を加えて結果に影響する要因を詳細に検討するなど、より詳細な解析を容易に実装できることも大きな利点である。これらのことから、SAS を用いた解析手法は実務的なメリットが非常に大きいと言える。加えて、理論に基づいて解析を実装することで、手法を深く理解せずに漫然と使用することを防ぐ効果も期待できる。

本発表が JAFROC 解析の理解と SAS の業務範囲の拡大に繋がれば幸いである。

5. 参考文献

- [1] Lee K., Lee S., Kwak J.S., Park H., Oh H., Koh J.C. Development and Validation of an Artificial Intelligence Model for Detecting Rib Fractures on Chest Radiographs. *J. Clin. Med.*, 13(13), 3850. (2024).
- [2] エルピクセル株式会社. 医用画像解析ソフトウェア EIRL Chest XR 添付文書第4版 (2024年5月改訂)
https://www.pmda.go.jp/PmdaSearch/kikiDetail/ResultDataSetPDF/171955_30400BZX00285000_A_00_04 (最終確認日 2024年8月26日)
- [3] Charles E. Metz. ROC解析の基礎. *日放技学誌*, 46(6), p.831 - 839, (1990).
- [4] 白石 順二, 小塚 和久, 畑川 政勝, 東田 満治, 渡辺 晋一. FROC解析におけるデータ収集簡略化の試み. *日放技学誌*, 47(4), p.620 - 626, (1991).
- [5] 白石 順二, 宇都宮 あかね. ROC解析における画像システム間の統計的有意差の検定方法 : Jackknife法とその適用. *日放技学誌*, 53(6), p.691 - 698, (1997).
- [6] Dorfman DD, Berbaum KS, Metz CE. ROC characteristic rating analysis: Generalization to the Population of Readers and Patients with the Jackknife method. *Invest.Radiol*, 27, p.723 - 731, (1992).
- [7] 工藤 健太郎, 舟尾 暢男 (2015). Wilcoxon-Mann-Whitney オッズ (WMWodds) と Wilcoxon 順位和検定について. https://www.sas.com/content/dam/SAS/ja_jp/doc/event/sas-user-groups/usergroups2015-b-03.pdf (最終確認日 2024年8月23日)
- [8] Chakraborty DP, Phillips P, Zai X (2022). Package ‘RJafroc’. RJafroc: Artificial Intelligence Systems and Observer Performance (r-project.org) (最終確認日 2024年8月23日)
- [9] SAS® Help Center. The GLM Procedure.
https://go.documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/statug/statug_glm_details55.htm (最終確認日 2024年8月23日)
- [10] Hillis SL. A Comparison of Denominator Degrees of Freedom Methods for Multiple Observer ROC Studies, *Statistics in Medicine*, 26, p.596 - 619, (2007)
- [11] Chakraborty DP, Zai X. Analysis of data acquired using ROC paradigm and its extension.
https://www.researchgate.net/publication/281069282_Analysis_of_Data_Acquired_Using_ROC_Paradigm_and_Its_Extensions/citations (最終確認日 2024年8月23日)
- [12] Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol.*, 15(5), p647 – 661, (2008).

生存時間解析用ADaMデータセット（ADTTE）のソースコード详解

田中 真史

(イーピーエス株式会社)

A Wordy Introduction to the Time-to-Event ADaM Programming

Masashi Tanaka

EPS Corporation

要旨

CDISC 準拠の臨床試験では、生存時間解析（Time-to-Event Analysis）に ADTTE データセットを用いる。SAS で ADTTE データセットを作成する際にどのような点が重要であるかを解説した。ADTTE データセットを作成する際のプログラム全体の構成について述べ、ソースコードの各行の意図を詳しく説明した。

キーワード：生存時間解析、データハンドリング、ADTTE、SAS、Time-to-Event Analysis

1. はじめに

CDISC 準拠の臨床試験で生存時間解析（Time-to-Event Analysis）を行う際に用いる ADaM データセット [1] [2]（ADTTE データセット）を SAS で作成する方法を詳しく解説する。本稿の対象は、ADTTE データセット（以下、「データセット」は省く）の各変数の意味は知っているが、その実装に不安がある解析担当者である。周辺の事柄から述べるので、ソースコードのみに興味がある場合は、3 章を眺めてほしい。SAS のバージョンは、9.4M7 を用いた。

ADTTE を作成する目的は、特定のイベントが起こるまでの時間と関連する情報をひとつのデータセットに整理して、生存時間解析の準備をすることである。臨床試験では、イベントの典型的な例が死亡であるため、また、生命表の分析から生存時間解析が発展してきた経緯があるため、「生存時間」解析と呼ぶことが多いが、イベントは死亡に限らないので、生存時間解析の応用範囲は広い [3] (1.2 章)。生存時間解析の特徴は、観察期間中にイベントが起こらなかった、又はイベントが起こる前にある条件を満たした場合のカテゴリとして、「打ち切り」(censor) を設けることである。

補足（あるいは蛇足）だが、辞書を引くと、censor（検閲、/sɛnsər/）には、censer（香炉）と sensor（感知装置）のふたつの同音異義語がある。綴りが似た単語に censure（非難、/sɛnʃər/）もある。単純なスペルチェックでは誤記を検出できないので、解析担当者は censor の綴りを正確に覚えておいた方がよい。感知装置の sensor と混同していると、イベントが起こった場合に、それを感知して変数 CNSR（Censor）を 1 以上にするのが自然であると錯覚する。しかしながら、CDISC の実装ガイド [1] (3.3.6 章) が勧告 (strongly recommend)

しているように、イベントを ADTTE に採用した場合は CNSR を 0 とし、打ち切りの場合は CNSR を 1 以上とするのが「検閲」の意味に忠実であろう。CDISC の実装ガイド [1] (3.3.6 章) は、打ち切りの理由に応じて CNSR の値を変えることを推奨 (recommend) しているが、同じく CDISC の実装ガイド [2] の Example 1 では、3 種類の打ち切りの全てについて、CNSR の値を 1 と定義している。

生存時間解析は、臨床試験の解析の中では、おそらく難しい部類に入る。ADTTE をログランク検定やコックス回帰といった手法で解析するが [3]、その詳細を検討するためには専門知識を有した担当者が解析作業に加わる必要がある。一方で、ADTTE の SAS による実装については、プログラム仕様が決まっていれば、生存時間解析の専門知識はおそらく必要ない。しかしながら、ADTTE の作成にはやや込み入った点があって、プログラム全体の構成を最初から考えることが重要になってくる。この点が SDTM からほとんど直接に作成できる ADaM データセットとは異なる。

2. ADTTE の作成方針

データセットを作成する際は、オブザベーション数の見積もりを最初に行う。CDISC の実装ガイド [2] の Example 1 で例示された基本的な ADTTE では、1 被験者 1 パラメータに対して 1 オブザベーションを発生させる。そのため、全体のオブザベーション数は被験者数×パラメータ数で計算できる (ADTTE に AVISIT を含めて、1 被験者 1 パラメータに複数のオブザベーションを発生させる場合もある)。作成の途中で、データセットのオブザベーション数がこの数を超えたり、1 被験者 1 パラメータに対して複数のオブザベーションが生じたりしていたら、作成方針を最初から考え直すべきである。データセットの重複を削除するなどして、オブザベーション数をその場しのぎで調整してはいけない。どのオブザベーションが削除されたかが不明瞭になり、プログラムの意図を追えなくなるからである。

ADTTE を作成する際、多数のデータセットを組み合わせるので、どのデータセットをベースにするかが最初に問題になる。簡単なのは、ADSL (Subject-Level Analysis Dataset) をベースにして 1 パラメータごとに作成する方法である。ADSL と他のデータセットを、被験者をキーにして 1 対 1 でマージして、必要な日付変数をすべてひとつのデータセットにまとめる。このとき、データセットの構造が 1 被験者 1 オブザベーションであることを確かめながら作業を進める。必要な日付変数を集約したら、それらの大小関係を比べて、どの日付を ADTTE.ADT (イベント又は打ち切りの日付) に採用するかを決める。

ADTTE.ADT を決定すると同時に、イベント又は打ち切りと、その内容をプログラム上で判定する。多くの場合、トレーサビリティを保つための変数として、SRCDOM、SRCVAR、SRCSEQ を作成するので、日付変数のソースデータの情報を作成途中のデータセットに保持しておく必要がある。SRCDOM、SRCVAR、SRCSEQ は、CDISC の実装ガイド [1][2] では「permissible」の変数なので作成は必須ではないが、実務上は最初から ADTTE に含めた方がよい。ADTTE のデータを検証する際に有用であるし、既存のソースコードに後から追加しようとする、ADT を決定してイベント又は打ち切りを判定するソースコードの核心部を大きく修正するか、核心部と重複する処理をプログラムの後半に追加するかのどちらかになり、影響が大きいからである。

ADTTE が作成できたら、LIFETEST プロシジャで Kaplan-Meier プロットを描いて、データを心行くまで眺めるとよい。

3. ADTTEの実装

3.1. 使用する疑似データセット

本稿では、ADSL (表 3.1.1、コード 3.1.4#1) と ADVS (体重やバイタルサインなどを含める ADaM データセット、表 3.1.2、コード 3.1.4#2) から表 3.1.3 の疑似的な ADTTE を作成する。ADTTE の設計は、CDISC の実装ガイド [2] の Example 1 を参考にした。

ADTTE には 6 人の被験者が含まれている。ADTTE のパラメータは体重又は腹囲の減少 (Weight/Waist Circumference Loss (days)) とした。そのパラメータに関連するイベントは、「10%を超える体重減少」と「5 cm を超える腹囲減少」の 2 種類とし、打ち切りは被験者の完了と中止の 2 種類とした。イベントのデータは ADVS に含まれており、「10%を超える体重減少」を CRIT1FL で表し、「5 cm を超える腹囲減少」を CRIT2FL で表した。イベント及び打ち切りのうち、投与開始日から発生までの日数が一番短いものを ADTTE に採用した。ここで、試験の中で発生したイベントの一部を、ADTTE に「採用」すると表現した。単に「イベント」といった場合、試験を通して発生したイベントを指しているのか、ADTTE に含まれているイベントのみを指しているのか区別する必要がある。複数のイベントや打ち切りが同日に起こった場合、それらを ADTTE に採用する優先順位を、「10%を超える体重減少」(最優先)、「5 cm を超える腹囲減少」、完了又は中止の順番とした (完了と中止はどちらか一方のみが起こる)。ADTTE を作成する際は、この優先順位に常に注意する。

AVAL は、STARTDT を起算日としたイベント又は打ち切りが発生した日付 (ADT) の日数である。ADT は、イベントについては ADVS.ADT から取得し、打ち切りについては ADSL.EOSDT から取得した。起算日 (STARTDT) は、治験薬の投与開始日 (ADSL.TRSTSDT) とした。CNSR=0 でイベントを、CNSR=1 で打ち切りを表し、イベント及び打ち切りの説明を EVNTDESC に格納した。EVNTDESC は「Event or Censoring Description」なので、打ち切りの理由も格納するのである [1][2] (「EVNTDESC」の変数名は良くないと思う)。

SRCDOM、SRCVAR、SRCSEQ にはソースデータの情報を格納した。ADTTE にイベントを採用した場合は、ADVS.ADT から日付を取得するので、SRCDOM="ADVS"、SRCVAR="ADT"、SRCSEQ=ADVS.ASEQ とした。打ち切りの場合は、SRCDOM="ADSL"、SRCVAR="EOSDT" とした (SRCSEQ は空白)。

ADTTE の各被験者について説明すると次のようになる。

- | | |
|-----------|---|
| 1001-0001 | 2007-05-30 に「10%を超える体重減少」と「5 cm を超える腹囲減少」が起こり、2007-06-30 に試験を完了した。優先順位の規約により、ADTTE には「10%を超える体重減少」をイベントとして採用した。 |
| 1001-0002 | イベントは起こらず、2007-06-19 に試験を完了した (打ち切り)。 |
| 1001-0003 | イベントは起こらず、2007-05-02 に試験を中止した (打ち切り)。 |
| 1001-0004 | イベントは起こらず、2007-06-26 に試験を完了した (打ち切り)。 |
| 1001-1005 | 2007-06-20 に「10%を超える体重減少」が起こり、同日に試験を完了した。優先順位の規約により、ADTTE には「10%を超える体重減少」をイベントとして採用した。 |
| 1001-1006 | 2007-04-20 に「5 cm を超える腹囲減少」が起こり、2007-05-20 に「10%を超える体重減少」が起こった。試験は 2007-07-20 に完了した。ADTTE には日付の早い「5 cm を超える腹囲減少」をイベントとして採用した。 |

表 3.1.1 ADSL

USUBJID Unique Subject Identifier
 TRTSDT Date of First Exposure to Treatment
 EOSSTT End of Study Status
 EOSDT End of Study Date

OBS	USUBJID	TRTSDT	EOSSTT	EOSDT
1	1001-0001	2007-01-01	COMPLETED	2007-06-30
2	1001-0002	2007-01-03	COMPLETED	2007-06-19
3	1001-0003	2007-01-03	DISCONTINUED	2007-05-02
4	1001-0004	2007-01-10	COMPLETED	2007-06-26
5	1001-1005	2007-01-11	COMPLETED	2007-06-20
6	1001-1006	2007-01-17	COMPLETED	2007-07-20

表 3.1.2 ADVS

USUBJID Unique Subject Identifier
 ASEQ Analysis Sequence Number
 PARAM Parameter
 PARAMCD Parameter Code
 ADT Analysis Date
 AVAL Analysis Value
 BASE Baseline Value
 CHG Change from Baseline
 PCHG Percent Change from Baseline

CRIT1 Analysis Criterion 1
 CRIT1FL Criterion 1 Evaluation Result Flag
 CRIT2 Analysis Criterion 2
 CRIT2FL Criterion 2 Evaluation Result Flag

OBS	USUBJID	ASEQ	PARAM	PARAMCD	ADT	AVAL	BASE	CHG	PCHG
1	1001-0001	100	Weight (kg)	WEIGHT	2007-05-30	70.1	78.2	-8.1	-10.36
2	1001-0001	121	Waist Circumference (cm)	WSTCIR	2007-05-30	83.0	91.0	-8.0	-8.79
3	1001-1005	138	Weight (kg)	WEIGHT	2007-06-20	62.5	70.5	-8.0	-11.35
4	1001-1006	65	Weight (kg)	WEIGHT	2007-05-20	66.5	75.0	-8.5	-11.33
5	1001-1006	102	Waist Circumference (cm)	WSTCIR	2007-04-20	80.0	87.0	-7.0	-8.05

OBS	CRIT1	CRIT1FL	CRIT2	CRIT2FL
1	Weight Loss of More than 10%	Y		
2			Waist Circumference Loss of More than 5cm	Y
3	Weight Loss of More than 10%	Y		
4	Weight Loss of More than 10%	Y		
5			Waist Circumference Loss of More than 5cm	Y

表 3.1.3 ADTTE

USUBJID Unique Subject Identifier
 PARAM Parameter
 PARAMCD Parameter Code
 AVAL Analysis Value
 STARTDT Time to Event Origin Date for Subject
 ADT Analysis Date
 CNSR Censor

SRCDOM Source Data
 SRCVAR Source Variable
 SRCSEQ Source Sequence Number
 EVNTDESC Event or Censoring Description

OBS	USUBJID	PARAM	PARAMCD	AVAL	STARTDT	ADT	CNSR
1	1001-0001	Weight/Waist Circumference Loss (days)	WTWCLOSS	150	2007-01-01	2007-05-30	0
2	1001-0002	Weight/Waist Circumference Loss (days)	WTWCLOSS	168	2007-01-03	2007-06-19	1
3	1001-0003	Weight/Waist Circumference Loss (days)	WTWCLOSS	120	2007-01-03	2007-05-02	1
4	1001-0004	Weight/Waist Circumference Loss (days)	WTWCLOSS	168	2007-01-10	2007-06-26	1
5	1001-1005	Weight/Waist Circumference Loss (days)	WTWCLOSS	161	2007-01-11	2007-06-20	0
6	1001-1006	Weight/Waist Circumference Loss (days)	WTWCLOSS	94	2007-01-17	2007-04-20	0

OBS	SRCDOM	SRCVAR	SRCSEQ	EVNTDESC
1	ADVS	ADT	100	WEIGHT LOSS
2	ADSL	EOSDT	.	COMPLETED
3	ADSL	EOSDT	.	DISCONTINUED
4	ADSL	EOSDT	.	COMPLETED
5	ADVS	ADT	138	WEIGHT LOSS
6	ADVS	ADT	102	WAIST CIRCUMFERENCE LOSS

コード 3.1.4 ADSL と ADVS

```
#
1 data adsl;
  length USUBJID $200 TRTSDT 8 EOSSTT $200 EOSDT 8;
  format TRTSDT e8601da. EOSDT e8601da.;
  infile cards dlm = '|' dsd missover;
  input USUBJID TRTSDT :e8601da. EOSSTT EOSDT :e8601da.;
  cards;
1001-0001|2007-01-01|COMPLETED|2007-06-30
1001-0002|2007-01-03|COMPLETED|2007-06-19
1001-0003|2007-01-03|DISCONTINUED|2007-05-02
1001-0004|2007-01-10|COMPLETED|2007-06-26
1001-1005|2007-01-11|COMPLETED|2007-06-20
1001-1006|2007-01-17|COMPLETED|2007-07-20
;
run;

2 data advs;
  length USUBJID $200 ASEQ 8 PARAM $200 PARAMCD $200 ADT 8 AVAL 8 BASE 8 CHG 8 PCHG 8 CRIT1 $200
  CRIT1FL $200 CRIT2 $200 CRIT2FL $200;
  format ADT e8601da.;
  infile cards dlm = '|' dsd missover;
  input USUBJID--ADT :e8601da. AVAL--CRIT2FL;
  cards;
1001-0001|100|Weight (kg)|WEIGHT|2007-05-30|70.1|78.2|-8.1|-10.36|Weight Loss of More than 10%|Y||
1001-0001|121|Waist Circumference (cm)|WSTCIR|2007-05-30|83|91|-8|-8.79|||Waist Circumference Loss of
More than 5cm|Y
1001-1005|138|Weight (kg)|WEIGHT|2007-06-20|62.5|70.5|-8|-11.35|Weight Loss of More than 10%|Y||
1001-1006|65|Weight (kg)|WEIGHT|2007-05-20|66.5|75|-8.5|-11.33|Weight Loss of More than 10%|Y||
1001-1006|102|Waist Circumference (cm)|WSTCIR|2007-04-20|80|87|-7|-8.05|||Waist Circumference Loss of
More than 5cm|Y
;
run;

/* EOF */
```

3.2. ADTTE のソースコードの説明

コード 3.2.1 で ADTTE を作成した。以下、コードの左側の番号 (#) を目印にして、順に説明を加える。wk01 データセットを作成する手順 (#7 から #12) がプログラムの中心である。「wk」は WORK ライブラリの略称のつもりである。安直ではあるが、データセット名や変数名は機械的に決めるのが簡単である。

ADVS から「10%を超える体重減少」のオブザベーション (CRIT1FL="Y") を抽出して、event1 データセットを作成した (#1、図 3.2.2)。ADTTE の ADT、CNSR、SRCDOM、SRCVAR、SRCSEQ、EVNTDESC の候補として、adt1、cnsr1、srcdom1、srcvar1、srcseq1、evntdesc1 を作成し (#2)、必要な変数だけを残した (#3)。同様に、「5 cm を超える腹囲減少」(CRIT2FL="Y") を対象にして event2 データセットを作成した (#4、図 3.2.2)。

打ち切りの情報は ADSL から取得して cnsr1 データセットを作成した (#5、図 3.2.2)。cnsr1 は打ち切りの候補をまとめたデータセットなので、ADTTE.CNSR の候補の cnsr11 の値を 1 にした (#6)。

#7 で、USUBJID をキーにして ADSL と event1、event2、cnsr1 をマージした。このとき、event1、event2、cnsr1 は 1 被験者 1 オブザベーションであるから、マージした結果のデータセットも 1 被験者 1 オブザベーションで、オブザベーション数は ADSL と同じになる。データ結合をする前に ADSL、event1、event2、cnsr1 を SORT プロシジャでソートする必要はない。ADaM データセットは USUBJID でソートされているものであり、所与のソート順を崩す処理をプログラムのどこにも入れていないからである。

event1、event2、cnsr1 に含まれる変数を ARRAY ステートメントで配列として参照できるようにした (#8)。#9 で、ADTTE のパラメータを定義した。

#10 と #11 が重要である。#10 の「if n(of dt[*])」の条件は、adt1、adt2、adt11 の全てが欠損値である状況を

考察から外すために加えた。「n」は N 関数で、引数のうちの非欠損値の数が戻り値である。次の行の i は、adt1、adt2、adt11 に番号 1、2、3 を与えたときに、最小の日付に付与された番号である（配列を使わずに書くと、「i=whichn(min(adt1, adt2, adt11), adt1, adt2, adt11)」）。MIN 関数は欠損値を無視するので、adt1、adt2、adt11 に欠損値が含まれる場合は、欠損値を除いた他の日付から番号が選択できる。adt1、adt2、adt11 に最小の日付が複数含まれる場合は、WHICHN 関数は前の方の引数から調べて戻り値を返すので、adt1（最優先）、adt2、adt11 の優先順位で i を定義できる。adt1 と adt2 が同日で adt11 よりも早い場合は、adt1 を優先して i は 1 になる。つまり、「10%を超える体重減少」と「5 cm を超える腹囲減少」が同日で打ち切りの日付よりも早い場合、i は「10%を超える体重減少」を示す 1 になる。このように ADTTE.ADT に採用する日付の由来を i の値で表すと、後が簡単である。「ADT=min(of dt[*])」と直接 ADT を求めに行くと、日付の由来を見失う。#11 で i を用いて、配列から ADT、CNSR、SRCDOM、SRCVAR、SRCSEQ、EVNTDESC を作成した。

#12 で STARTDT と AVAL を作成し、必要な変数だけを残した。AVAL の単位が月や年の場合は、日数から適当に換算する。

#13、#14 は ADaM データセットの体裁の調整であり、様々な方針があると思う。実務では、変数長の調整など、もう少し複雑な処理を加える。#13 の DATASETS プロシジャで変数ラベル、フォーマット、インフォーマットを一括で削除し、#14 で、変数ラベルとフォーマットを改めて定義して完成である。LIFETEST プロシジャで Kaplan-Meier プロットを描いてみよう（#15、図 3.2.3）。

コード 3.2.1 ADTTE の実装

```
#
1  data event1;
    set advs;
    where CRIT1FL = 'Y';
2  adt1      = ADT;
   cnsr1     = 0;
   srcdom1   = 'ADVS';
   srcvar1   = 'ADT';
   srcseq1   = ASEQ;
   evntdesc1 = 'WEIGHT LOSS';
3  format adt1 e8601da.;
   keep USUBJID adt1--evntdesc1;
run;

4  data event2;
    set advs;
    where CRIT2FL = 'Y';
   adt2      = ADT;
   cnsr2     = 0;
   srcdom2   = 'ADVS';
   srcvar2   = 'ADT';
   srcseq2   = ASEQ;
   evntdesc2 = 'WAIST CIRCUMFERENCE LOSS';
   format adt2 e8601da.;
   keep USUBJID adt2--evntdesc2;
run;

5  data cnsr1;
    set adsl;
6  adt11     = EOSDT;
   cnsr11    = 1;
   srcdom11  = 'ADSL';
   srcvar11  = 'EOSDT';
   srcseq11  = .;
   evntdesc11 = EOSSTT;
   format adt11 e8601da.;
   keep USUBJID adt11--evntdesc11;
run;
```

```

7 data wk01;
  merge adsl event1 event2 cnsr1;
  by USUBJID;
8 array dt[*] adt1-adt2          adt11;
  array cs[*] cnsr1-cnsr2        cnsr11;
  array s1[*] srcdom1-srcdom2     srcdom11;
  array s2[*] srcvar1-srcvar2     srcvar11;
  array s3[*] srcseq1-srcseq2     srcseq11;
  array ev[*] evntdesc1-evntdesc2 evntdesc11;
9 PARAM = 'Weight/Waist Circumference Loss (days)';
  PARAMCD = 'WTWCLOSS';
10 if n(of dt[*]) then do;
  i = whichn(min(of dt[*]), of dt[*]);
11 ADT      = dt[i];
  CNSR     = cs[i];
  SRCDOM   = s1[i];
  SRCVAR   = s2[i];
  SRCSEQ   = s3[i];
  EVNTDESC = ev[i];
  end;
12 STARTDT = TRTSDT;
  AVAL     = ADT - TRTSDT + 1;
  keep USUBJID PARAM PARAMCD AVAL STARTDT ADT CNSR SRCDOM SRCVAR SRCSEQ EVNTDESC;
run;

13 proc datasets nolist;
  modify wk01;
  attrib _all_ label = '';
  attrib _all_ format =;
  attrib _all_ informat =;
run;

14 data adtte;
  attrib
    USUBJID label = 'Unique Subject Identifier'
    PARAM   label = 'Parameter'
    PARAMCD label = 'Parameter Code'
    AVAL    label = 'Analysis Value'
    STARTDT label = 'Time to Event Origin Date for Subject' format = e8601da.
    ADT     label = 'Analysis Date' format = e8601da.
    CNSR    label = 'Censor'
    SRCDOM  label = 'Source Data'
    SRCVAR  label = 'Source Variable'
    SRCSEQ  label = 'Source Sequence Number'
    EVNTDESC label = 'Event or Censoring Description'
  ;
  set wk01;
run;

15 proc lifetest data = adtte;
  time AVAL*CNSR(1);
run;

/* EOF */

```

图 3.2.2 event1(上)、event2(中)、cnsr1(下)

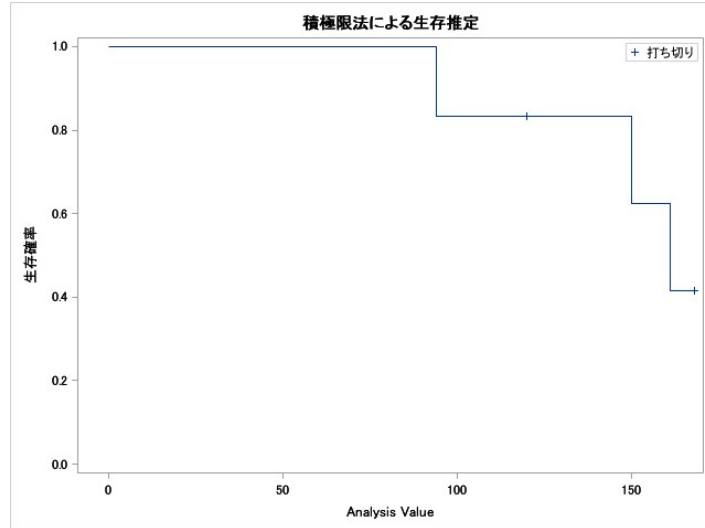
OBS	USUBJID	adt1	cnsr1	srcdom1	srcvar1	srcseq1	evntdesc1
1	1001-0001	2007-05-30	0	ADVS	ADT	100	WEIGHT LOSS
2	1001-1005	2007-06-20	0	ADVS	ADT	138	WEIGHT LOSS
3	1001-1006	2007-05-20	0	ADVS	ADT	65	WEIGHT LOSS

OBS	USUBJID	adt2	cnsr2	srcdom2	srcvar2	srcseq2	evntdesc2
1	1001-0001	2007-05-30	0	ADVS	ADT	121	WAIST CIRCUMFERENCE LOSS
2	1001-1006	2007-04-20	0	ADVS	ADT	102	WAIST CIRCUMFERENCE LOSS

OBS	USUBJID	adt11	cnsr11	srcdom11	srcvar11	srcseq11	evntdesc11
1	1001-0001	2007-06-30	1	ADSL	EOSDT	.	COMPLETED
2	1001-0002	2007-06-19	1	ADSL	EOSDT	.	COMPLETED
3	1001-0003	2007-05-02	1	ADSL	EOSDT	.	DISCONTINUED
4	1001-0004	2007-06-26	1	ADSL	EOSDT	.	COMPLETED
5	1001-1005	2007-06-20	1	ADSL	EOSDT	.	COMPLETED
6	1001-1006	2007-07-20	1	ADSL	EOSDT	.	COMPLETED

図 3.2.3 ADTTE データセットと Kaplan-Meier プロット

OBS	USUBJID	PARAM	PARAMCD	AVAL	STARTDT	ADT	CNSR	SRCDOM	SRCVAR	SRCSEQ	EVNTDESC
1	1001-0001	Weight/Waist Circumference Loss (days)	WTWCLOSS	150	2007-01-01	2007-05-30	0	ADVS	ADT	100	WEIGHT LOSS
2	1001-0002	Weight/Waist Circumference Loss (days)	WTWCLOSS	168	2007-01-03	2007-06-19	1	ADSL	EOSDT		COMPLETED
3	1001-0003	Weight/Waist Circumference Loss (days)	WTWCLOSS	120	2007-01-03	2007-05-02	1	ADSL	EOSDT		DISCONTINUED
4	1001-0004	Weight/Waist Circumference Loss (days)	WTWCLOSS	168	2007-01-10	2007-06-26	1	ADSL	EOSDT		COMPLETED
5	1001-1005	Weight/Waist Circumference Loss (days)	WTWCLOSS	161	2007-01-11	2007-06-20	0	ADVS	ADT	138	WEIGHT LOSS
6	1001-1006	Weight/Waist Circumference Loss (days)	WTWCLOSS	94	2007-01-17	2007-04-20	0	ADVS	ADT	102	WAIST CIRCUMFERENCE LOSS



4. おわりに

臨床試験におけるプログラミングの技術は、解析手法の知識とは別の重要性がある。実務では、説明可能なソースコードを書くべきであろう。意味と意図が明瞭なソースコードを運用できたら、作成したデータセットが正しいかどうかの検証が容易になり、統計解析の前提である ADaM データセット自体に誤りが含まれる可能性を下げられるはずである。そのような堅実な解析業務を期待して、基本的な事柄を長々と書いた。プログラミングの技術面に不安を残した状態で ADTTE を作成することになったとき、本稿がプログラミング作業の一助になれば幸いである。

引用文献

1. CDISC Analysis Data Model Team. Analysis Data Model Implementation Guide Version 1.1. (オンライン) (引用日: 2024年8月21日.) https://www.sastricks.com/cdisc/ADaMIG_v1.1.pdf.
2. CDISC Analysis Data Model (ADaM) Team. The ADaM Basic Data Structure for Time-to-Event Analyses. (オンライン) (引用日: 2024年8月21日.) https://www.pinnacle21.com/sites/default/files/2017-01/adam_tte_final_v1.pdf.
3. 大橋靖雄, 浜田知久馬, 魚住龍史. 生存時間解析: SASによる生物統計. 第2版. 目黒区: 東京大学出版会, 2022. 978-4-13-062322-3.

support vector machineを用いた格付推計モデルの構築

—hpsvm procedureを用いたパラメータ推定—

○坂巻英一¹

(¹国立大学法人宇都宮大学)

A introduction of credit ratings prediction model using support vector machine

-Parameter estimation using hpsvm procedure-

Yoshikazu, Sakamaki

School of Data Science and Management, Utsunomiya University

要旨

円安基調が続き物価のインフレ傾向が強まる中、老後の資産形成を目的とした投資が盛んに行われるようになり個人投資家による投資は増加傾向にある。株式投資をはじめとした金融商品には債券や株式を発行する企業などの財務状態が悪化することで、債務が回収できなくなるといったリスクも存在し信用リスク或いはデフォルトリスクと呼ばれる。

信用リスクの判断は金融のプロでも難しいのが現状である。そのため、客観的な視点からリスクを判断するために「信用格付」という指標が考え出された。一般に信用格付は債権や株式の発行体の信用リスクを評価した上でアルファベットや数字で表されることが多い。

先行研究を概観すると Altman(1968)以降リスクの大きさを定量的に計測するモデルはこれまで盛んに研究されてきたものの、アルファベットや数字で表された「信用格付」そのものを推定するモデルは実務的なニーズが少ないこともあり研究事例はあまり多くない。本稿では hpsvm procedure を用いたサポートベクターマシンによる格付推計モデルの構築方法を紹介する。併せて、実データを用いてモデルパラメータを推定した結果を報告する。

キーワード：格付推計モデル，サポートベクターマシン(SVM)，信用リスク，proc hpsvm

1 先行研究紹介

株価が上昇基調を続け NISA (小額投資非課税制度) による税制優遇が受けられる中、新たに投資を始める人々の数は近年増加傾向にある。こうした中、投資の初心者にとって如何にして投資先を決定するか、は重要な課題の一つとなっている。投資先の経営状態が悪化した場合、株価が下落するだけでなく、原資を回収することができなくなる可能性もある。投資先の企業が抱えるこうしたリスクは一般に信用リスクと呼ばれている。ところが信用リスクの判断は金融のプロでも難しいのが現実である。そのため客観的な視点からリスクを判断するために「信用格付」という指標が考え出された。

一般に信用格付けは債権や株式の発行体の信用リスクを評価した上で、アルファベットによって表されることが多い。例えば、大手格付け会社の Standard & Poors (S&P)は、「AAA」が最上位となり、「AA」「A」と続き最下位の「D」まで細かくランク付けされる(株式会社三井住友銀行, 2018)。

こうした格付は実務的には信用リスクモデルを基に定量的に計測されたリスク量をグループ分けした上で行われることが多い。信用リスクの計量化に関する研究は Altman(1968)まで遡ることができる。Altman は、

- k 企業番号 ($k=1,2,\dots,K$)
- m 変数番号 ($m=1,2,\dots,M$)
- v_m 財務指標 m に対するパラメータ
- x_{km} 企業 k における m 番目の財務指標

とした上で、財務指標の重み付け線形和により企業の経営状態 z を

$$z_k = \sum_{m=1}^M v_m x_{km} \quad (1)$$

により定式化している。その上で判別分析を用いてパラメータ v_m を推定している。

Altman がモデル構築の際に使用した財務指標は

- x_{k1} = working capital / total assets
- x_{k2} = retained earnings / total assets
- x_{k3} = earnings before interest and taxes / total assets
- x_{k4} = market value of equity / total liabilities
- x_{k5} = sales / total assets

の5つである。

このようにして算出された z-score に対して Altman(1968)は cut-off ポイントを設定した上で企業の経営状態を

$z > 2.99$	Safe Zone
$1.81 < z < 2.99$	Grey Zone
$z < 1.81$	Distress Zone

のように定義している。

当初、判別分析により推定されていたモデルパラメータは、その後、重回帰分析や2項ロジットモデル、順序ロジットモデルへと発展し、モデル内で使用される財務指標に改良が加えられながら、今日も金融機関をはじめとした信用リスク計測の分野で広く利用されている。

Altman(1968)以降、実務的には定量化されたリスク量をグループ分けした上で格付として使用することが多く、バーゼル規制以降、金融機関において行われている内部格付も多くがこの考え方を基礎として進められてきた。

そのため、財務指標を基に格付そのものを推定するモデルは実務的なニーズも少なく研究事例も少ないのが現状である。一方で、格付会社が公開している格付は、財務指標を基にアナリストの経験則等に基づく定性要因を加味した上で格付が決定されることがあり、定量情報だけを用いて推計された格付と実際に企業に対して付与される格付に差が生じることが考えられる。

こうした点を考慮した場合、財務指標から信用格付を直接推定する研究を行うことにもそれなりに意味があると考えられる。財務指標から信用格付を直接推定する手法として、坂巻(2004)(2006)は入れ子型ロジット

モデルを基礎とした格付推計モデルを提案している。また、田中・中川(2014)はサポートベクターマシン(SVM)を用いて、財務指標を基に信用格付を推定する方法を提案している。

今、記号を

i 格付番号 ($i=1,2,\dots,I$)

k 企業番号 ($k=1,2,\dots,K$)

β_m m 番目の財務指標に対する重みベクトル

x_{km} 企業 k における m 番目の財務指標

のように定義する。

田中らは企業に対する格付 i を数値により定義した上で、格付 i が j より小さいか大きいかを判別するための判別スコアを超平面

$$R_j(\mathbf{x}^h) = \beta_{j0} + \sum_{m=1}^M \beta_{jm} x_{km} \quad (2)$$

により定式化している。

この時、企業 h の財務指標と超平面の距離は

$$\frac{R_j(\mathbf{x}^h)}{\|\beta_j\|} \quad (3)$$

によって定式化される。

個々の企業が超平面によって完全に線形判別可能と仮定した場合、判別面は無数に書くことができるが、各企業と判別面の距離が最も離れた、いわば最も明確に分離できる判別面を書くのが望ましい。この観点から、企業の信用スコアと判別面との距離の最小値 $1/\|\beta_j\|$ (以下、最小幾何マージン) を可能な限り大きくした判別面を推定する。これを踏まえた上で記述された(4)式で示すモデルを中山らはハードマージン SVM と呼んでいる。

$$\begin{aligned} \text{Min } & \|\beta_j\|^2 \\ \text{s.t. } & \beta_{j0} + \sum_{m=1}^M \beta_{jm} x_{km} > 1, \quad h \in M_U^j \\ & \beta_{j0} + \sum_{m=1}^M \beta_{jm} x_{km} < -1, \quad h \in M_L^j \end{aligned} \quad (4)$$

一方で、実務的には与えられた判別面だけでは正しく判別を行うことができない企業が出てくる可能性があることを考慮したモデル構築を行う必要がある。田中らはスラック変数を導入した上でスラック変数の合計値の最小化を同時に考慮したモデルをソフトマージン SVM と呼び(5)式のように定式化している。

$$\begin{aligned} \text{Min } & \|\beta_j\|^2 + C \sum_{h=1}^H \xi_i^h \\ \text{s.t. } & \beta_{j0} + \sum_{m=1}^M \beta_{jm} x_{km} + \xi_i^h > 1, \quad h \in M_U^j \\ & \beta_{j0} + \sum_{m=1}^M \beta_{jm} x_{km} - \xi_i^h < -1, \quad h \in M_L^j \\ & \xi_i^h \geq 0 \end{aligned} \quad (5)$$

ところで、企業に付与された格付には通常格付間で偏りがあるため(5)式で定式化された最小化問題をそのまま解いた場合、パラメータ推定値に偏りが生じることが考えられる。

こうした影響を取り除くことを目的として、田中らは各格付 M_j に属する企業数 m_j に応じてスラック変数に重み H/m_j を乗じることにより比較的サンプル数が多い格付の影響を除去することを可能にしている。スラック変数に重みを乗じた状態を(6)式に示す。

$$\sum_{h=1}^H \xi_i^h \Rightarrow \frac{H}{M_U} \sum_{h \in M_U} \xi_i^h + \frac{H}{M_L} \sum_{h \in M_L} \xi_i^h \quad (6)$$

田中らは更に推定された格付間の交差が生じないように、

$$R_{i+1}(\mathbf{x}^h) - R_i(\mathbf{x}^h) \geq 1 \quad (7)$$

なる条件を付与した上でパラメータ推定を行っている。

本稿では `hpsvm procedure` の紹介を行うことを目的としているため(7)式で示す条件は考慮せず(5)式及び(6)式を基にソフトマージン SVM による格付推計を行った結果を報告する。

2 本稿における提案手法

田中らのモデルは今日機械学習におけるデータ分類アルゴリズムとして広く利用されている SVM を基礎とした手法を早い段階から採用している点で興味深い。一方で、提案手法を実データに適用する実証実験において図表 1 のように隣り合う格付をひとつのグループにまとめた上で、3つの格付区分(M_1 、 M_2 、 M_3)を対象とした検証実験となっている。

AAA, AA+, AA, AA-, A+	M_1
A, A-	M_2
BBB+, BBB, BBB-, BB+, BB	M_3

図表 1 田中らのモデルにおける格付分類 (出典：田中らの論文より抜粋)

隣り合う格付同士をグループにまとめた上で実証実験を行っている理由として、格付の中には AAA のようにもともとのデータ件数が少ないケースがあり、パラメータの推定結果を安定させるためにまとまったデータ件数を確保することがあるのではないかと考えられる。

このように分類先によって件数に大きな偏りがあるデータは不均衡データ (**imbalanced data**) と呼ばれており、回帰分析を始めとしたパラメータ推定を伴うモデルにおいてこうしたデータを使用した場合、パラメータの推定結果やモデルの分類精度が低下することが先行研究において指摘されてきた。

不均衡データをモデル内で扱う際、こうした課題を回避することを目的として、オーバーサンプリングやアンダーサンプリング、アンサンブル学習等、様々なアルゴリズムが研究されている。CNN を用いた画像分類を行うアルゴリズムにおいて、データ数が少ない分類先に属する画像について、画像を少しだけ回転、拡大・縮小するなどしてデータを複製し、データ数が多い分類先にデータ件数を併せた上でパラメータ推定を実施することがあり水増し法 (**augmentation**) と呼ばれている。水増し法はオーバーサンプリングのひとつといえることができる。

サポートベクターマシンを用いた格付推計モデルについて考えた場合、AAA や B、CCC といった格付を付与される企業は相対的に少ないのが現状であり、データ数に偏りが生じることになる。本稿ではデータ数が少ない格付に属する企業について、オーバーサンプリングによりデータを複製した上で分析に使用することとする。

3 実データを用いた実証実験

3.1 データ概要

本節では実データを基に第3節で示した方法により検証実験を行った結果を報告する。検証に使用したデータの概要は以下の通りである。

データ出典元 Kaggle

(データサイエンス学習者向けにオープンデータを提供するプラットフォームであり Google によって運営されている。)

データセット名 Corporate Credit Rating

URL https://www.kaggle.com/datasets/agewerc/corporate-credit-rating/corporate_rating.csv

財務指標数 16 個

分析ツール 本研究では分析ツールとして SAS OnDemand for Academics を使用する。SAS OnDemand for Academics とは、統計分析について教育および学習することを目的として SAS Institute 社が無償で提供する教育機関向けオンライン環境である。

データ概要 本データは米国の格付会社を実施した米国企業に対する格付データ並びに財務指標からなるデータセットである。このうち本稿では S&P による格付データを分析対象とし、2014 年度のデータをモデル構築用データ、2015 年度及び 2016 年度のデータを検証用データとして使用することとする。データセットの概要を図表 2 に示す。

Rating_Agency	rating_year							合計
	2010	2011	2012	2013	2014	2015	2016	
DBRS	1	2	3	2	4	5	9	26
Egan-Jones Ratings Company	0	0	0	575	1,129	1,119	3	2,826
Fitch Ratings	0	0	161	88	70	85	73	477
HR Ratings de Mexico S.A. de C.V.	0	0	1	1	1	1	1	5
Japan Credit Rating Agency, Ltd.	0	0	1	5	6	5	5	22
Moody's Investors Service	0	0	495	272	280	297	292	1,636
Standard & Poor's Ratings Services	40	323	410	444	454	550	592	2,813
合計	41	325	1,071	1,387	1,944	2,062	975	7,805

図表 2 検証実験で使用したデータの概要

また、S&P による格付データについて、格付毎の企業数は図表 3 の通りである。

rating_year	Rating(Rating)																			合計				
	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-	B+	B	B-	CCC+	CCC	CCC-		CC+	CC	C	D
2010	3	0	0	1	0	5	5	6	4	5	3	2	3	2	1	0	0	0	0	0	0	0	0	40
2011	10	0	10	3	10	40	21	36	40	36	22	16	23	9	23	6	9	6	1	0	1	1	0	323
2012	10	0	13	8	12	46	29	44	53	42	21	21	31	26	21	16	7	5	0	0	1	3	1	410
2013	9	0	10	9	11	51	29	39	67	44	22	30	32	30	28	19	8	4	0	0	2	0	0	444
2014	10	1	7	9	14	51	33	45	55	43	33	28	33	36	21	21	11	3	0	0	0	0	0	454
2015	16	4	28	7	11	75	23	45	56	55	48	51	34	47	19	20	10	1	0	0	0	0	0	550
2016	18	4	26	6	8	81	21	38	63	54	60	39	53	40	28	26	19	4	1	0	2	0	1	592
合計	76	9	94	43	66	349	161	253	338	279	209	187	209	190	141	108	64	23	2	0	6	4	2	2,813

図表 3 格付毎の企業数

また図表 5 に hpsvm プロシージャが出力したパラメータ推定値の例を示す。

```

コード   ログ   結果
126
127 ***** define model parameter w values *****;
128 array _w_{11} _temporary_ (
129     4.84418117966900E-03
130     7.24666210916470E-04
131     -3.01456991074630E-04
132     8.70457088828220E-05
133     -2.46374019248860E-04
134     3.39761757912550E-03
135     -4.18519041086270E-03
136     -2.80982995174130E-03
137     -5.46098489759730E-04
138     2.95572140561020E-03
139     -2.44789794373570E-03
140 );
141
142 if _missingflag_ = 1 then
143     substr(_WARN_, 1, 1) = 'M';
144 if _unknownflag_ = 1 then
145     substr(_WARN_, 2, 1) = 'U';
146
147 ***** compute the predictive value *****;
148 do _i=1 to _nw;
149     _PVAL_ = _PVAL_ + _x[_i]*_w[_i];
150 end;
151
152 ***** BETA *****;
153 _PVAL_ = _PVAL_ + 1.00275706168220E+00;
154 _P_ = _PVAL_;
155

```

図表 5 hpsvm プロシージャが出力したパラメータ推定値

図表 4 に示す SAS コードを実行すると「code file=ファイル名」により指定された SAS ファイルに図表 5 に示すコードが出力される。このファイルを図表 6 に示すような data step 内の %inc で指定することにより検証用データに対するスコアリングが可能となる。

```

data dataset;
set dataset (where=(Rating_Agency="Standard & Poor's Ratings Services" and YEAR=&year.));
if RATING <= &j. then y = 1;
else y = -1;
%inc "/home/uxxxxxxx/param_est.sas";
run;

```

図表 6 検証用データに対する格付推計

3.4 適合度検証

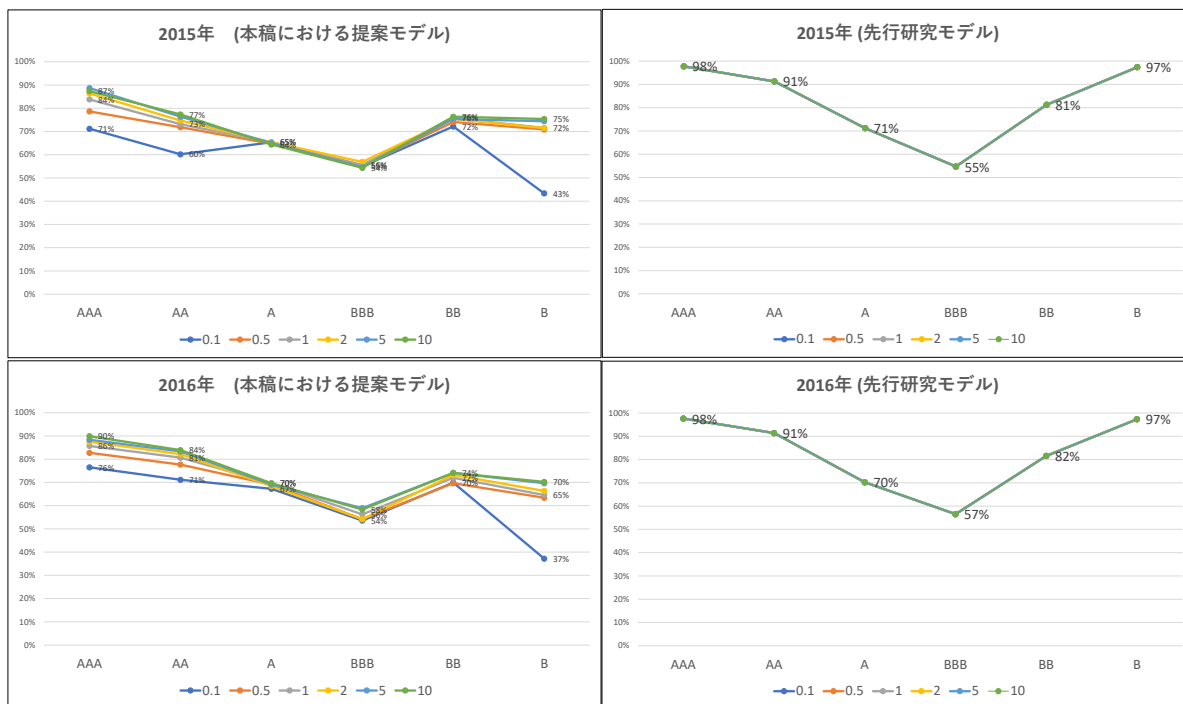
ここで田中らが行った先行研究における手法並びに本稿における提案手法により、混合行列を用いて適合度検証を行った結果を報告する。ここで混合行列とは実際の格付と予測される格付を基に作成された図表 7 のような行列を指す。本稿では(7)式を基に混合行列から算出される予測の的中率(Accuracy Ratio)を基にモデルの評価を行うこととする。

		予測	
		POSITIVE(正)	NEGATIVE(誤)
実際	POSITIVE(正)	TP(True Positive)	FN(False Positive)
	NEGATIVE(誤)	FP(False Positive)	TN(True Negative)

図表 7 混合行列の説明

$$\text{Accuracy Ratio} = \frac{TP + TN}{TP + FN + FP + TN} \tag{7}$$

図表 8 に予測の的中率を算出した結果を示す。分析を行った結果、予測の的中率に関しては先行研究モデルの方が本研究における提案手法と比較して、概ね高い値を示していることが確認できる。ここで、先行研究モデルのグラフが一本しかないのは、C (正則化パラメータ) の大きさによってパラメータの推定値に大きな変動がなく、全てのグラフが重なっているためである。



図表 8 予測的中率を算出した結果

これらの分析結果より予測的中率に関して見た場合、C（正則化パラメータ）の大きさに関わらず、総じて、先行研究モデルの方が本研究における提案モデルよりも高い的中率を示していることが分かる。このうち、本研究における提案モデルにおいては、格付構成企業数が相対的に多い A、BBB 及び BB に関しては C（正則化パラメータ）の値によって予測精度が大きく変わることはない反面、構成企業数が相対的に少ない AAA 及び AA に関しては、C（正則化パラメータ）が大きい程予測的中率が向上する傾向があることが分かる。

次に格付毎に予測的中率を集計した結果についてみる。

図表 9 は C（正則化パラメータ）を 10 に設定した上で格付毎の予測的中率を算出した結果である。

実際の格付	2015年 推計結果 (本稿における提案モデル)								
	AAA	AA	A	BBB	BB	B	CCC	合計	
AAA	5	0	2	0	0	0	0	7	
AA	5	1	4	7	3	0	0	20	
A	14	5	16	20	5	2	0	62	
BBB	8	11	16	30	14	2	0	81	
BB	7	4	14	19	34	3	0	81	
B	2	3	4	9	29	3	0	50	
CCC以下	1	1	2	1	1	2	0	8	
合計	42	25	55	169	1	17	0	309	

実際の格付	2015年 推計結果 (先行研究モデル)								
	AAA	AA	A	BBB	BB	B	CCC	合計	
AAA	0	0	0	7	0	0	0	7	
AA	0	0	0	17	3	0	0	20	
A	0	0	0	55	5	2	0	62	
BBB	0	0	0	65	14	2	0	81	
BB	0	0	0	44	34	3	0	81	
B	0	0	0	18	29	3	0	50	
CCC以下	0	0	0	5	1	2	0	8	
合計	0	0	0	211	86	12	0	309	

実際の格付	2015年 推計結果 (本稿における提案モデル)								
	AAA	AA	A	BBB	BB	B	CCC	合計	
AAA	8	0	0	0	0	0	0	8	
AA	8	4	4	4	1	0	0	21	
A	11	2	17	29	7	5	0	71	
BBB	7	11	13	30	19	0	0	80	
BB	5	2	9	27	44	7	0	94	
B	2	2	2	4	33	10	0	53	
CCC以下	0	1	0	0	4	3	0	8	
合計	42	22	40	200	1	31	0	336	

実際の格付	2015年 推計結果 (先行研究モデル)								
	AAA	AA	A	BBB	BB	B	CCC	合計	
AAA	0	0	0	8	0	0	0	8	
AA	0	0	0	20	1	0	0	21	
A	0	0	0	59	7	5	0	71	
BBB	0	0	0	61	19	0	0	80	
BB	0	0	0	43	44	7	0	94	
B	0	0	0	10	33	10	0	53	
CCC以下	0	0	0	1	4	3	0	8	
合計	0	0	0	203	108	25	0	336	

図表 9 格付毎の予測的中率を算出した結果

分析を行った結果、先行研究モデルではモデルから推計される格付が構成企業数の相対的に多い A、BBB 及び BB に集まる傾向がある反面、構成企業数が相対的に少ない AAA 及び AA についても実際の格付と推計される格付が一致する傾向があることが分かる。

4 結論

本稿では `hpsvm procedure` を用いて格付推計モデルを構築する方法について紹介を行った。本研究において参照した先行研究モデルでは複数の格付をひとつのグループに統合した上でモデル構築を実施していた。こうした格付の統合を行わなければならない背景に、構成企業数が少ない格付において推定されたパラメータを基に検証用データに対するクロスバリデーションを実施した場合、モデルの推計精度が低下してしまう、といった課題があるためではないかと考えられる。こうした課題を改善するために、本稿ではデータのオーバーサンプリングを実施し、ターゲットとなるクラスに属するデータ件数とターゲットとならないクラスに属するデータ件数を一致させる方法を提案した。先行研究モデル並びに本研究における提案モデルを **Kaggle** が公開している実データへ適用した上で予測的中率について検証を行った結果、先行研究モデルでは構成企業数が相対的に多い格付に推計結果が集まる傾向があった反面、本研究における提案モデルでは校正企業数が相対的に少ない AAA、AA といった格付についてもそれなりの精度で推計を実施することが可能であることが分かった。

最後に本研究の今後の展望について記述する。本稿では土のノッチを除外した上で 6 つの格付の中からひとつの分類先を選択する、という分類モデルを構築した。一方、実際の格付データには土のノッチが付与されていることが多いため、本研究で提案するような手法を実務で使用する場合、例えば、第一段階で格付を推計した上で第二段階でノッチを推計する、といったようなモデルを構築する必要があると考えられる。

`hpsvm procedure` はサポートベクターマシンを実行するためのプロシージャとして格付推計以外の分野においても広く利用されている。本稿における実証実験の結果がサポートベクターマシンによる機械学習に関する業務に従事する関係者の一助となれば幸いである。

以上

参考文献

Altman, E. (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, 23, 589-609.

株式会社三井住友銀行 (2018), わかると差が出る「信用リスクとは？」

URL <https://www.smbc.co.jp/kojin/money-viva/kihon-no-ki/0028/>

最終更新日 2018年5月23日

最終閲覧日 2024年7月31日

坂巻英一(2004), 入れ子型ロジットモデルを用いた企業格付モデル構築法の提案及び提案モデルの妥当性に関する検証

<https://www.sas.com/offices/asiapacific/japan/usergroups/wg/archive/041015saka.pdf>

最終閲覧日 2024年7月31日

坂巻英一(2006), 多段階多項企業格付予測のためのロジット・モデルに関する研究, *行動計量学*, 33(1), pp.69-85

田中克弘, 中川秀敏 (2014), 企業格付判別のための SVM 手法の提案および逐次ロジットモデルとの比較による有効性検証, *日本オペレーションズ・リサーチ学会和文論文誌*, vol.57, pp. 92-111

謝辞

本稿を執筆するに当たり多くの貴重なご助言を頂きました SAS Institute Japan 社テクニカルサポートの方々はこの場をお借り致しまして心より御礼申し上げます。

大学発StartUp

データのパワーで新たな可能性を開く
データサイエンスに特化した大学発コンサルティングファーム



2024 始動!

Data Mining and Knowledge Discovery
宇都宮大学 - 地域創生推進機構

寄席の「顔」が柳家から春風亭に変わるまで

○坂部 裕美子

(公益財団法人 統計情報研究開発センター)

The change in the head of the vaudeville theater from Yanagiya to Shunputei

Yumiko Sakabe

Statistical Information Institute for Consulting and Analysis

要旨

筆者は落語協会 HP 掲載の寄席番組データを SAS データ化して、各演者の公演プログラム登場回数を 1 年ごとに集計する、という作業を十数年にわたって行っているが、以前 SAS ユーザー会で報告した、寄席定席における一門別勢力分布の将来予想がほぼ当たっていた、という事実について報告する。併せて、最新の公演データから、さらに今後の勢力変化についても推測する。

キーワード：寄席定席、時系列比較、means プロシジャ、

---本文開始点---

1. 寄席定席プログラムについて

1.1 寄席定席とは

落語家は、師匠の下に弟子入りして修業し正規の協会員として認められると、都内にある 4 つの寄席（鈴木演芸場、新宿末廣亭、浅草演芸ホール、池袋演芸場）で前座として働き始めるのが最初の落語家としての仕事となる。寄席は落語家にとっての「ホームグラウンド」であり、二ツ目、真打と出世した後でも常に出演したい公演会場でもあるはずなのだが、特に落語協会では近年落語家の数が多くなりすぎて、ほとんどの落語家にとって、寄席出演そのものが難しい状況になっている。2023 年の落語協会主催の寄席興行出演者（ただし落語家に限る）について、プログラム登場回数（実際に高座に上がった回数ではなく、今後の出演予定者として公演プログラムに掲載された回数）のジニ係数を調べてみたところ、0.60 にも上っていた。

1.2 寄席定席プログラム登場回数集計

この「プログラム登場回数」の集計方法について説明する。寄席では先述の 4 つの演芸場で月に 3 つのプログラム（それぞれに昼の部と夜の部がある）の興行が行われているのだが、諸々の変遷を経て、現在落語協会主催興行はそのうち約 3 分の 2 になっている。プログラムには十数名ずつの出演者が配されるが、時折「交替枠」として複数の出演者が配される（それぞれの出演者内で 10 日間の出演枠を分け合う）ことがある。集計に際しては、交替枠は 1 を交替者数で割った値をそれぞれ登場回数とすることにした。また、初席（正

月の1～10日の興行)など、前半日程と後半日程で出演者が変わる場合は、それぞれの登場回数を0.5倍して合算することにした。

2. これまでの集計について

2.1 初めての集計 (2005年)

筆者が最初にSASユーザー会で発表を行ったのは2005年の「SASを用いた寄席定席興行の現状分析」だった。この時は4演芸場の平均的プログラム構成や色物の方が登場回数が多くなることなど、ごく基礎的な内容についての報告となったが、これは「どうやら今の寄席は柳家小さん一門の出演が多いようだ」という印象を得る最初の機会となった。

2.2 時系列での比較 (2009年)

最初の集計では、落語協会HPに掲載されていたプログラムを手元のExcelファイルに貼り付けて保存したものを使って集計していたが、その後研究者の方から戦後の公演プログラム(完全ではないがほぼ揃っている)のPDFを頂くことができたので、これをデータ化して一門別で回数比較をしたのが2009年の「落語家の勢力分布の変遷—師匠から弟子へ・親から子へ—」である。

タイトルにある通り、この時は「勢力(登場回数)の委譲はあるのか」ということに主たる関心を置いて集計を行ったのだが、昭和の頃のデータを集計してみると、1970～80年代から既に小さん一門は寄席に出まくっていたことが分かり、この一門の「岩盤」は崩れないのではないかとさえ思うようになっていた。しかし、時系列比較用のグラフをよく見ると、一門によって次世代への委譲がうまく進んでいるところとそうでないところがあるのに気がつき、「今後は彦六一門が伸びるのでは？」と予想していた。

そして現在、プログラム登場回数の1位、2位はこの一門の春風亭一朝・春風亭一之輔のいずれか、という年が大半なのである。

3. 今回の集計

3.1 一門のグループ分けと世代分け

今回も、2009年の集計と同様の「一門」分けをしようと思う。すなわち、誰の弟子か、もしくは大師匠(師匠の師匠)が誰になるかを考慮して、落語家全員を「小さん」(主に「柳家」)「志ん生」(主に「古今亭」)「彦六」(主に「春風亭」)「圓生」(「三遊亭」の一部)「金馬」(「三遊亭」の一部)「三平」(主に「林家」)「圓歌」(「三遊亭」の一部)「文楽」(主に「桂」)「圓蔵」(主に「橘家」)「その他」の10のいずれかの一門に分ける、ということである。

そして、世代については、前回から15年経っていることを考慮して区分を改め(前回は「1950年以前の生まれを第1世代としたが、その後かなりの方がお亡くなりになってしまった)、今回は当人の生年によって、1960年以前の生まれを第1、1970年以前の生まれを第2、1980年以前の生まれを第3、それ以降を第4とする4つの世代に分割する。

比較に使うデータは、前回の報告で2005年までのデータを使っているため、その続きということで2005、2010、2015、2019、2023年のデータとする。2020年はコロナ禍で寄席は長期休業となったため、データに連続性がないと判断し、コロナ前年から4年おきのデータとした。

3.2 一門別比較

まず、一門別×世代別の人数は以下の表1のとおりである。

表1 一年の世代別人数

	第1世代	第2世代	第3世代	第4世代	合計
文楽	7			1	8
圓歌	5	5	7	6	23
圓生	10	4	5	6	25
圓藏	8		2		10
彦六	9	6	8	19	42
金馬	3	2	4		9
小さん	30	13	25	31	99
三平	11	3	5	7	26
志ん生	25	8	15	15	63
その他	10	3	2	1	16

退潮が著しい（若い世代が入ってこない）一門もあるが、1980年以降生まれの大半が小さん・彦六・志ん生一門に行く中で、前回の報告では先行きが不透明だとした圓生・三平一門がさほど崩れていないのが感慨深い。

一門別にプログラム登場回数を合算したものの推移が、図1である。

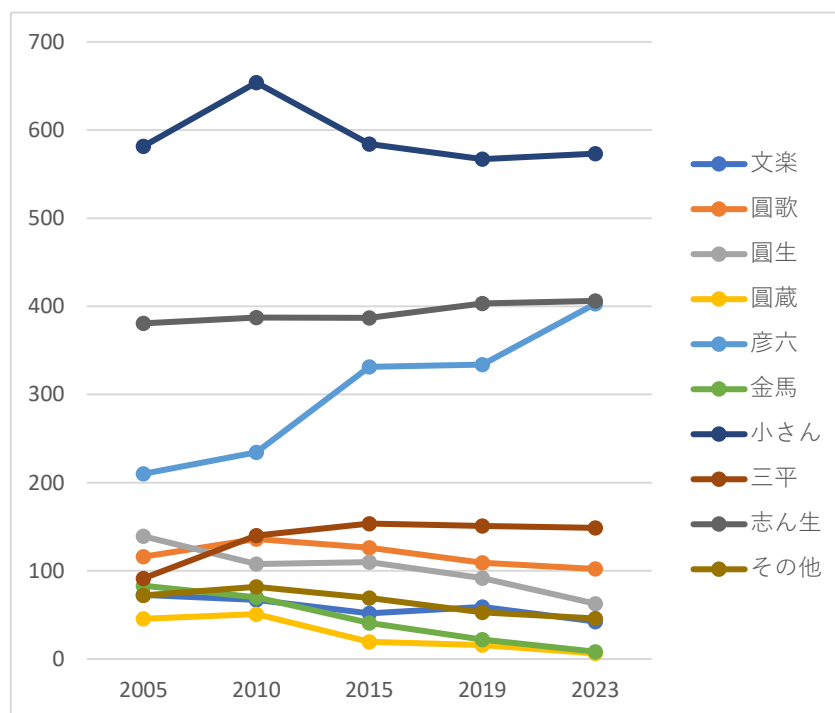


図1 一門別登場回数 (2005~2023年)

一門全員で見ると、小さん一門の圧勝状態は相変わらず続いている。が、目を見張るのが飛ぶ鳥を落とす勢いの彦六一門である。前回の報告では下位集団の一つに甘んじていたが、完全に抜け出して2位の志ん生き一門に、ついに2023年には追いついている。

3.2 一門別×世代別の比較

ここからは、特に注目される小さん一門・志ん生一門・彦六一門について、世代別の登場回数の推移を比較してみる。グラフは折れ線グラフが世代ごとの登場回数合計（第1軸）、棒グラフが世代ごとのその年の登場者数（第2軸）で、左から第1（青）、第2（黄）、第3（緑）、第4（エンジ）である。

まず、小さん一門を示す。

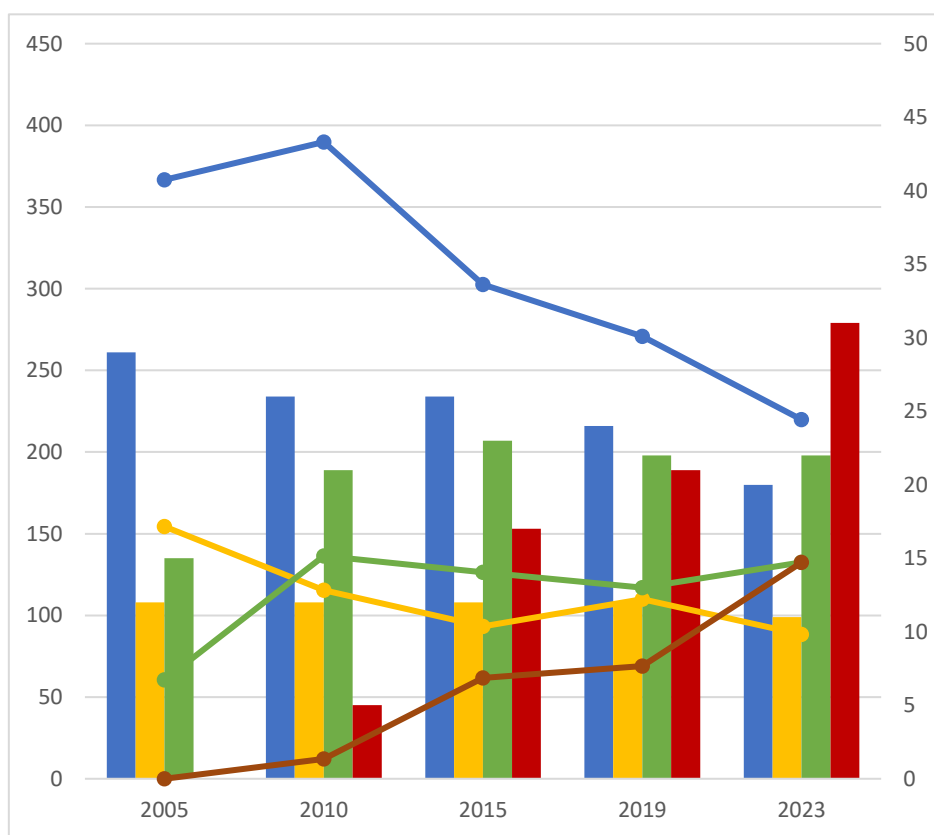


図 2-1 世代別人数・登場回数（柳家小さん一門）

第1世代（さん喬、権太楼、小三治、扇橋9、扇遊、小ゑんなど）が若干は回数を減らしながらもまだまだ安泰であるのに比して、第2（喬太郎、市馬、扇辰など）、第3（花緑、三三、こみちなど）世代があまり伸びてこない…などと見えてしまうのが小さん一門の凄いところで、2023年にはどの世代も合計100回くらいになっている。年ごとに変動はあるが、出演枠の合計はおおよそ1800前後なので、寄席の出番の4分の1程度は小さん一門が占めていることになる。

第4世代が人数ほどには回数を増やせていないのは、まだ二ツ目の若手が多いから、と考えることもできるが、ある師匠がトリを取る場合、弟子の二ツ目が全員、寄席の最初の出番に交替出演となるケースが多いため、1人当たりの回数が目減りした状態の若手が多くなってしまふ、というような事情はあるかも知れない。

続いて、志ん生一門を次ページ図 3-2 に示す。人数・回数合計では小さん一門に水をあけられているが、世代別回数で見ると、小さん一門以上に順調に若い世代が伸びてきていることが分かる。第 1 世代（雲助、志ん橋、志ん輔、馬生など）の回数が減少しているにもかかわらず、図 1 で全期間を通じてあまり合計の値が変わらないのは、その穴を見事に次世代が埋めていたから、ということになるだろう。

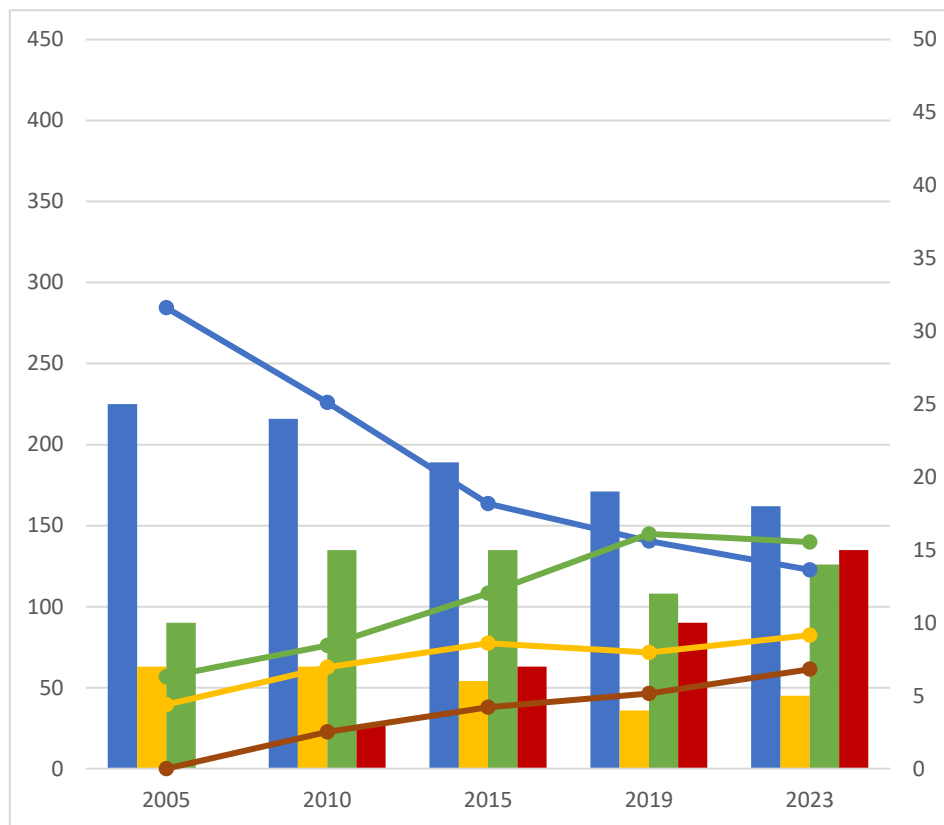


図 2-2 世代別人数・登場回数（古今亭志ん生一門）

志ん生一門の第 2 世代は人数が少なめだが、雲助一門の白酒、馬石がしっかり寄席に居場所を作っている。また、第 3 世代の菊之丞、文菊も今や回数上位の常連であり、いかにも世代交代がうまく進んだという印象である。

そして、彦六一門が図 2-3 である。小さん、志ん生一門と比べると人数も回数系も少ないのだが、よく見ると 2023 年は全世代が 100 回近くになっており、少数精鋭の総力戦という感じである。

第 1（一朝、木久扇、小朝、正雀など）世代が小さん、志ん生一門の中心に比べると若いので、まだ目立った回数の減少はない。その状況下で、もう第 2（玉の輔、園太郎、彦いち、文蔵など）、第 3（一之輔、三朝、木久蔵、きく麿など）世代が寄席の出番の一角を占め始めているということである。

実は、2009 年の集計の際には、一朝、玉の輔、園太郎といった、当時からある程度の回数出演していたメンバーがもっと回数を増やして彦六一門を引っ張るのではないかと、思っていた。しかし、この集計報告の後、この一門からスーパーstarが登場することになる。一之輔である。期待を受けての大型抜擢で真打に昇進して以降、その期待に違わぬ活躍を見せながらも、きちんと寄席にも出演し続けている。その姿勢が、日本テレビ「笑点」のレギュラーになっても変わらなかったことで、さらに驚かされた（詳述は省くが、「落語家は売れると寄席に出なくなる」というのがこれまでの常であった）。

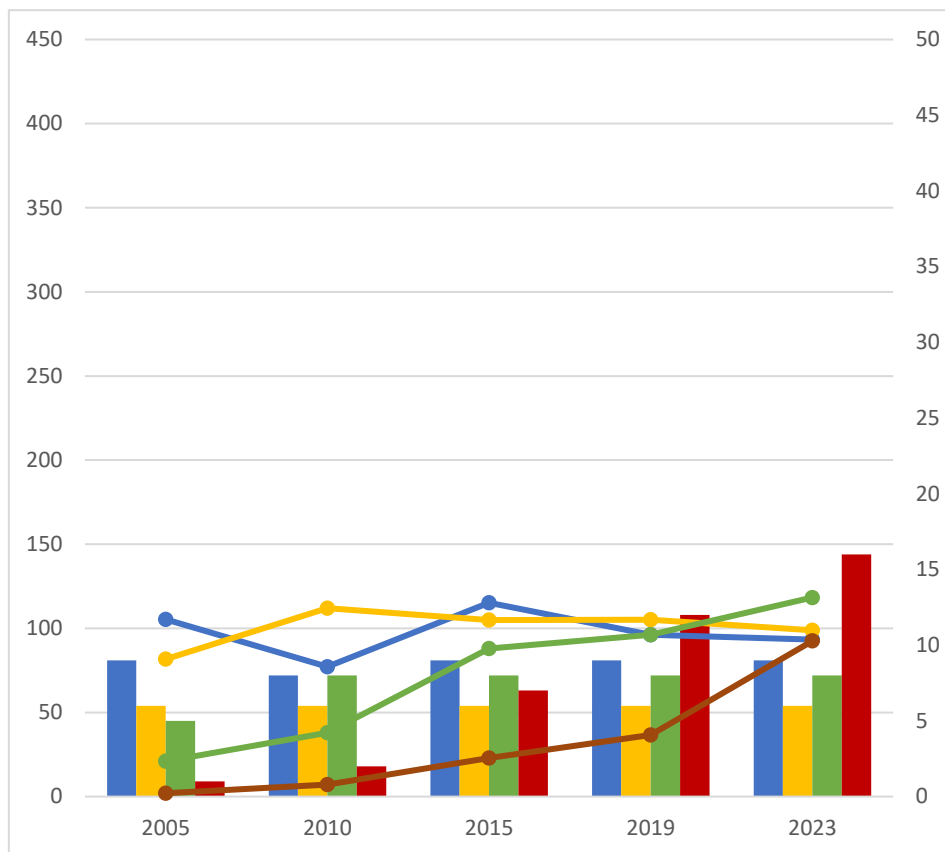


図 2-3 世代別人数・登場回数 (林家彦六一門)

4. 寄席の「顔」

ここまで見てくると、寄席出演枠の獲得勢力図としては、まだ小さん一門が圧倒的である、ということになりそうである。しかし、この結果にたどり着くまでには出演者を一門別に振り分け、それを合算するという手間がかかっている。実際に今回の作業では、2005 年で止まっていた落語家データを 2023 年時点まで更新する必要があったのだが、現在の落語家過多＝名前不足の状況もあってか、二ツ目の名前が次々に代替わりしており（この間に「小太郎」が別人で 3 人登場している）、作業は予想外に手間取ってしまった。

それよりも、分かりやすいのは単純な「年間登場回数ランキング」であろう。そこで、2005、2010、2015、2019、2023 年の登場回数上位 5 名を以下に挙げてみる。

表 2 登場回数年間上位 5 名

2005	2010	2015	2019	2023
柳家さん喬 (小さん一門)	林家正蔵 (三平一門)	春風亭一朝 (彦六一門)	春風亭一朝 (彦六一門)	春風亭一朝 (彦六一門)
柳亭市馬 (小さん一門)	入船亭扇遊 (小さん一門)	春風亭一之輔 (彦六一門)	林家正蔵 (三平一門)	古今亭文菊 (志ん生一門)
柳家権太楼 (小さん一門)	柳家権太楼 (小さん一門)	桃月庵白酒 (志ん生一門)	春風亭一之輔 (彦六一門)	春風亭一之輔 (彦六一門)

ん一門)	ん一門)	生一門)	六一門)	六一門)
春風亭一朝 (彦六一門)	柳家さん喬 (小さん一門)	林家正蔵 (三平一門)	古今亭文菊 (志ん生一門)	柳家さん喬 (小さん一門)
入船亭扇橋 (小さん一門)	柳亭市馬 (小さん一門)	三遊亭歌武蔵 (圓歌一門)	柳亭市馬 (小さん一門)	隅田川馬石 (志ん生一門)

先述のとおり、これ以前の時代も小さん一門が上位に君臨していたことを考えると、「回数トップ」つまりは寄席の「顔」が、2000年代に入って柳家から春風亭に変わった、と言ってもよいのではないだろうか。

参考文献

- 「SAS を用いた寄席定席興行の現状分析」坂部裕美子、2005年 SAS ユーザー会論文集 pp.273-282
- 「落語家の勢力分布の変遷－師匠から弟子へ・親から子へ－」坂部裕美子、2009年 SAS ユーザー会論文集 pp.255-264

競合リスクを伴う生存時間解析について

樋口史

(イーピーエス株式会社)

Survival analysis with competing risks

Fuhito Higuchi

EPS Corporation

要旨

競合リスクは、生存時間解析で対象とする主要なイベントの発生を妨げるイベントである。競合リスクを無情報打ち切りとして扱くと、Kaplan-Meier 推定量にはバイアスが入ることが知られており、競合リスクを考慮すべきとされている。従来では競合リスクの問題について考慮されていない解析が多かったが、近年では、競合リスクを考慮した解析が広まりつつある。競合リスクを考慮した解析手法としては、累積発生関数及び Gray 検定が用いられることが多く、SAS での実装も可能である。本発表では競合リスクの概要に加えてこれらの解析手法について紹介を行う。具体的には、生存時間解析の基礎の振り返りとして、一般的に用いられてきた Kaplan-Meier 推定量及び log-rank 検定の説明を行った上で、累積発生関数及び Gray 検定との数学的背景の違い、LIFETEST プロシジャの EVENTCODE オプションを使用した競合リスクを考慮した解析の実装方法、結果の解釈方法を紹介する。

キーワード：生存時間解析 競合リスク 累積発生関数 Gray 検定

1. はじめに

生存時間解析は、イベントが起きるまでの時間とイベントの関係に焦点を当てる解析方法であり、医学分野や工学分野において特に重要な解析方法である。医学分野においては疾病の再発や死亡を対象とした解析に用いられており、工学分野においては製品の故障を対象とした解析に用いられている。本稿では医学分野に焦点を当てるため、医学データに対する言い回しを用いる。

がん治療の試験を例に、死亡をイベントとする場合について考える。がんによる死亡を注目イベントとすると、事故や他の疾病による死亡など、他の原因によるイベントによって注目するイベントが観測できないことがあり、注目するイベント以外のイベントが起きるリスクを競合リスクという。競合リスクが原因で観測されたイベントを競合リスクイベントという。競合リスクイベントは、「注目イベントの発生を妨げる」という情報を持っているため、無情報打ち切りとして扱くと「生存時間と打ち切りが独立である」という生存時間解析を行う上での仮定を満たさなくなる。従来では、この競合リスクイベントを無情報打ち切りとして

扱う解析が多くあった。生存時間解析では、生存時間関数を推定する際に Kaplan-Meier 法がよく用いられるが、競合リスクが存在する場合、その推定量にバイアスが入ることが知られている。そのため、近年では競合リスクを考慮した解析が広まりつつある。そこで、本稿では、競合リスクを考慮する解析として多く用いられている累積発生関数及び Gray 検定について紹介する。また、それらの解析を SAS で実装する方法と結果の解釈方法も紹介する。

2. 生存関数とハザード関数

まずは生存時間解析で用いる関数の定義について紹介する。\$T\$ を生存時間を表す確率変数とすると、\$T\$ の分布関数は、確率密度関数 \$f(t)\$ を用いて式(2.1)で定義される。\$F(t)\$ は生存時間がある値 \$t\$ 以下となる確率やイベントが時点 \$t\$ までに発現する確率を表す。

$$F(t) = P(T \leq t) = \int_0^t f(u) du \quad (2.1)$$

生存時間関数 \$S(t)\$ は、生存時間が \$t\$ よりも大きくなる確率を表す。

$$S(t) = P(T > t) = 1 - F(t) \quad (2.2)$$

ハザード関数 \$h(t)\$ は、被験者に時点 \$t\$ までイベントが発生していないという条件のもとで、その次の瞬間にイベントが起きる確率を表す。

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (2.3)$$

式(2.2)の生存関数を推定する方法として、一般的に Kaplan-Meier 法が用いられる。時点 \$t_i\$ の直前までに生存している被験者数を \$n_i\$、時点 \$t_i\$ で発生したイベント数を \$d_i\$ とすると、Kaplan-Meier 推定量は式(2.4)で表される。\$n_i\$ はリスク集合と呼ばれる。イベントと打ち切りが同時に発生している場合は、打ち切りはイベントが起きた直後に起きているものとして扱う。

$$\hat{S}(t) = \prod_{t_i < t} \left(1 - \frac{d_i}{n_i}\right) \quad (2.4)$$

次に、群間の生存時間を比較するためによく用いられる log-rank 検定について紹介する。今回、2 群 (\$k = 1, 2\$) の場合を取り上げる。両群で \$r\$ 個のイベント発生時点 \$t_{(1)} < t_{(2)} < \dots < t_{(r)}\$ があり、時点 \$t_{(i)}\$ におけるイベント数を群 I (\$k = 1\$) で \$d_{1i}\$、群 II (\$k = 2\$) で \$d_{2i}\$ とし、その合計数を \$d_i\$ とする。さらに、時点 \$t_{(i)}\$ の直前において、群 I のリスク集合が \$n_{1i}\$、群 II のリスク集合が \$n_{2i}\$ とし、その合計数を \$n_i\$ とする。このとき、log-rank 検定の検定統計量は式(2.5)となる。

$$\frac{(\sum_{i=1}^r (d_{1i} - e_{1i}))^2}{\sum_{i=1}^r v_{1i}} \quad (2.5)$$

イベント数 \$d_{1i}\$ が超幾何分布に従い、\$e_{1i}\$ はイベント数 \$d_{1i}\$ の期待値 (\$e_{1i} = n_{1i} d_i / n_i\$)、\$v_{1i}\$ はイベント数 \$d_{1i}\$ の分散 (\$v_{1i} = n_{1i} n_{2i} d_i (n_i - d_i) / n_i^2 (n_i - 1)\$) を表す。各群の生存時間に差がないという帰無仮説の下、式(2.5)が \$\chi^2\$ 分布に従うことを用いて、検定を行う。

3. 競合リスク

競合リスクが存在する場合、式(2.4)の Kaplan-Meier 法を用いると各時点の生存率を過大評価してしまう。その問題点について、表 3.1 のデータを用いて示す。

表 3.1 サンプルデータ

被験者番号 <i>i</i>	生存時間 t_i	イベントの種類 δ_i
1	5	0
2	6	1
3	7	0
4	9	2
5	11	1
6	13	1
7	15	0
8	15	1
9	16	2
10	19	1

10 人の被験者*i*の生存時間を t_i とし、 δ_i はイベントの種類を表す。イベントの種類を表す δ_i は

$$\delta_i = \begin{cases} 0 & \text{(打ち切り)} \\ 1 & \text{(注目するイベント)} \\ 2 & \text{(競合イベント)} \end{cases}$$

と定義する。従来のように競合イベント($\delta_i = 2$)も打ち切り($\delta_i = 0$)と扱い、Kaplan-Meier 法による生存関数の推定値を求めると、表(3.2)のようになる。

表 3.2 $\delta_i = 1$ をイベント、 $\delta_i = 0, 2$ を打ち切りとした Kaplan-Meier 法による生存関数の推定値

時間 t_i	リスク集合 n_i	イベントが起きた数 d_i	$\hat{S}(t)$
5	10	0	1
6	9	1	$(1 - \frac{1}{9}) = 0.8889$
7	8	0	0.8889
9	7	0	0.8889
11	6	1	$0.8889 \times (1 - \frac{1}{6}) = 0.7407$
13	5	1	$0.7407 \times (1 - \frac{1}{5}) = 0.5926$
15	4	1	$0.5926 \times (1 - \frac{1}{4}) = 0.4444$
16	2	0	0.4444
19	1	1	$0.4444 \times (1 - \frac{1}{1}) = 0$

時間 $t_i = 9, 16$ では競合イベント($\delta_i = 2$)が起きているため、本来生存率は下がるはずだが、打ち切り($\delta_i = 0$)として扱っているため、過大に推定していることが分かる。このように、競合リスクが存在する場合にKaplan-Meier法を用いると、適切な解析結果を得られない。そこで、競合リスクが存在する場合、イベントの種類ごとに累積発生関数を推定することで、生存関数が過大に推定される問題を解決する。

3.1. 累積発生関数

時点 t まで生存したという条件のもとで、その次の瞬間に原因 $j(j = 1, 2, \dots, m)$ のイベントが起きる確率を表す原因別ハザード関数 $h_j(t)$ は式(3.1)で定義される。

$$h_j(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t, J = j | T \geq t)}{\Delta t} \quad (3.1)$$

時点 t まで生存したもしくは競合リスクイベントが起きた条件のもとで、その時間に原因 $j(j = 1, 2, \dots, m)$ のイベントが起きる確率を表す原因別部分分布ハザード関数 $\gamma_j(t)$ は式(3.2)で定義される。式(3.1)と異なる点は、分母に時点 t までに競合リスクイベントが起きた被験者が含まれることである。

$$\gamma_j(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t, J = j | T \geq t \cup (T \leq t \cap J \neq j))}{\Delta t} \quad (3.2)$$

原因 j のイベントが時点 t までに起きる確率を意味する累積発生関数 $F_j(t)$ は、原因別ハザード関数 $h_j(t)$ と生存関数 $S(t)$ を用いて式(3.3)で定義される。イベントの種類ごとの発現率を表しているため、式(2.1)とは異なり、 $F_j(\infty) < 1$ となる。そのため、式(3.3)は部分分布ともよばれる。

$$F_j(t) = P(T < t, J = j) = \int_0^t h_j(u) S(u) du \quad (3.3)$$

式(3.3)は式(2.1)と以下の関係式が成り立つ。このことから、式(2.1)で定義した分布関数 $F(t)$ は、時点 t までに原因 $j(j = 1, 2, \dots, m)$ のいずれかのイベントが発現する確率を表す。

$$F(t) = \sum_{j=1}^m F_j(t) \quad (3.4)$$

また、式(2.2)と式(3.4)から式(3.5)が成り立つ。そのため、生存関数 $S(t)$ は時点 t までに原因 $j(j = 1, 2, \dots, m)$ のいずれのイベントも発現していない確率を表すことから、無イベント生存率または無病生存率を意味している。

$$S(t) = 1 - \sum_{j=1}^m F_j(t) \quad (3.5)$$

n_{ji} を時点 t_{ji} の直前までいずれのイベントも発現せず、生存している被験者数(リスク集合)、 d_{ji} を時点 t_{ji} で原因 j のイベントが起きた数とする。 $\hat{S}(t-)$ は時点 t_{ji} の直前までに注目しているかどうかに関わらず、いずれのイベントも起きていない確率の推定値とする。これらを用いて、累積発生関数 $F_j(t)$ は式(3.6)で推定することができる。

$$\hat{F}_j(t) = \sum_{t_{ji} \leq t} \frac{d_{ji}}{n_{ji}} \hat{S}(t_{ji}-) \quad (3.6)$$

表 3.1 のデータを用いて、原因 $j = 1$ のイベントを注目するイベント($\delta_i = 1$)とし、 $\hat{F}_1(t)$ を求めた結果を表 3.3 に示す。ここで、 $\hat{S}(t)$ はいずれの原因によるものもイベントとして扱った場合の生存関数の推定値である。

表 3.3 原因 $j = 1$ のイベントを注目するイベント($\delta_i = 1$)とした累積発生関数の推定値

生存時間 t_{1i}	リスク集合 n_{1i}	イベントが 起きた数 d_{1i}	$\hat{S}(t)$	$\hat{F}_1(t)$
5	10	0	1	0
6	9	1	$\left(1 - \frac{1}{9}\right) = 0.8889$	$\frac{1}{9} \times 1 = 0.1111$
7	8	0	0.8889	0.1111
9	7	0	$0.8889 \times \left(1 - \frac{1}{7}\right) = 0.7619$	0.1111
11	6	1	$0.7619 \times \left(1 - \frac{1}{6}\right) = 0.6349$	$0.1111 + \frac{1}{6} \times 0.7619 = 0.2381$
13	5	1	$0.6349 \times \left(1 - \frac{1}{5}\right) = 0.5079$	$0.2381 + \frac{1}{5} \times 0.6349 = 0.3651$
15	4	1	$0.5079 \times \left(1 - \frac{1}{4}\right) = 0.3810$	$0.3651 + \frac{1}{4} \times 0.5079 = 0.4921$
16	2	0	$0.3810 \times \left(1 - \frac{1}{2}\right) = 0.1905$	0.4921
19	1	1	$0.1905 \times \left(1 - \frac{1}{1}\right) = 0$	$0.4921 + \frac{1}{1} \times 0.1905 = 0.6825$

同様に原因 $j = 2$ のイベントを注目するイベント($\delta_i = 2$)としたときの累積発生関数の推定値 $\hat{F}_2(t)$ を求めると、 $1 - \hat{S}(t) = 1 - (\hat{F}_1(t) + \hat{F}_2(t))$ から式(3.5)が成り立つことも確認できる。

3.2. Gray 検定

log-rank 検定は生存時間に差がないという帰無仮説の下で検定を行う一方で、Gray 検定は累積発生関数に差がないという帰無仮説の下で検定を行う。Gray 検定で用いる検定統計量は式(3.7)となる。

$$\int_0^t W_k(u) (\hat{y}_{jk}(u) - \hat{y}_j(u)) du \quad (3.7)$$

$W_k(t)$ は重み関数、 $\hat{y}_{jk}(u)$ は群 k の原因別部分分布ハザード関数の推定値である。詳細については LIFETEST プロシジャのマニュアル(参考文献[8])を参照されたい。

4. SAS による解析

LIFETEST プロシジャを用いて、累積発生関数の推定と Gray 検定を実装する方法について紹介する。(SAS9.4M3(SAS/STAT14.1)から可能)。

4.1. 累積発生関数の推定

表 3.1 のデータセットを用いる。データセット名を wk01 とし、ID 変数は被験者番号、VAL 変数は生存時間、status 変数はイベントの種類別に発生を示す変数とする。TIME ステートメントの VAL*status(0)で VAL 変数を時間変数、status=0 を打ち切りとして指定している。eventcode=1 で(failcode=1 でも可能)status=1 を注目イベントとして指定している。残った status=2 のデータは競合リスクによるイベントとなる。結果には累

積発生関数(Cumulative Incidence)の推定値に加え、その標準誤差と 95%信頼区間が出力される。デフォルトでは標準誤差が出力されるが、PROC LIFETEST ステートメントの CIFVAR オプションを用いることで、推定値の分散を出力することができる。標準誤差は Aalen 法と Delta 法の 2 通りの計算方法が実装されており、デフォルトは Aalen 法によって求められる。Delta 法を指定する場合は、PROC LIFETEST ステートメントの ERROR オプション(ERROR=DELTA)を用いる。95%信頼区間はデフォルトで設定されている二重対数変換に基づく方法で求められている。表(3.3)で求めた計算結果と SAS で出力された結果が同じであることが確認できる。また、ODS GRAPHICS を用いると、図(5.1)の累積発生関数の推定結果をグラフに出力することができる。グラフを見ると、競合イベントが発生していても注目イベントが発生していない区間では水平な線が引かれていることが分かる。

```
proc lifetest data=wk01;
  time VAL*status(0) / eventcode=1;
run;
```

Failed Event: status=1

Summary of Failure Outcomes			
Failed Events	Competing Events	Censored	Total
5	2	3	10

Cumulative Incidence Function Estimates				
VAL	Cumulative Incidence	Standard Error	95% Confidence Interval	
0	0	0	-	-
6	0.1111	0.1111	0.00469	0.4064
11	0.2381	0.1601	0.0274	0.5640
13	0.3651	0.1858	0.0664	0.6877
15	0.4921	0.1976	0.1162	0.7916
19	0.6825	0.3016	0.0250	0.9612

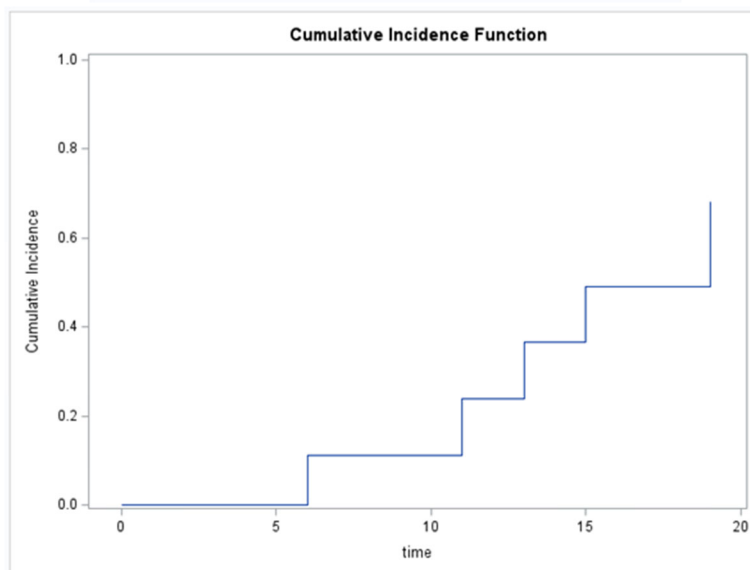


図 5.1 status=1 を注目イベントとした累積発生関数のグラフ

3章でも言及した，生存関数の推定値と累積発生関数の推定値との視覚的な比較を図5.2に示す．生存関数の推定値は(1-生存関数の推定値)でのプロットとした．垂直な灰色の線は，実線が $\delta = 1$ の注目するイベントが起きた時点，破線が $\delta = 2$ の競合リスクイベントが起きた時点を示す．時点 $t = 9$ で競合リスクが発生した後の注目イベントが発生した時点 $t = 11$ の推定結果について検討する．赤の破線は従来使用されることが多かった，競合リスクによるイベントを打ち切りとして扱ったKaplan-Meier法による生存関数の推定値のプロットである(図5.2内，CRE=censor)．一方で，緑の破線は，競合リスクによるイベントも注目するイベントとしたKaplan-Meier法による生存関数の推定値のプロットである(図5.2内，CRE=event)．時点 $t = 11$ のイベント発生率について，緑の破線は36.5%であるが，赤の破線では25.9%となっており，従来の方ではイベントの発生率が低く(生存率が高く)推定されていることが分かる．しかしながら，緑の破線では，イベントが注目したいものか否かが考慮されておらず，結果の解釈が難しいものとなる．このように従来用いられてきたKaplan-Meier法による推定では，イベントの考慮をできていなかったり，イベントの無発生割合を過大に推定したりといったバイアスが生じる．そこで，青の実線でプロットされた注目するイベント(今回は $\delta = 1$ のイベント)を指定した累積発生関数を用いることで，イベントが注目したいものかを考慮した上での発生率を推定することができる．さらに，注目するイベントの累積発生関数と橙色の破線でプロットされた競合リスクのイベントの累積発生関数の推定値の和が，緑の破線となることを見て取れ，式(3.5)の関係が成立することが視覚的にも分かる．

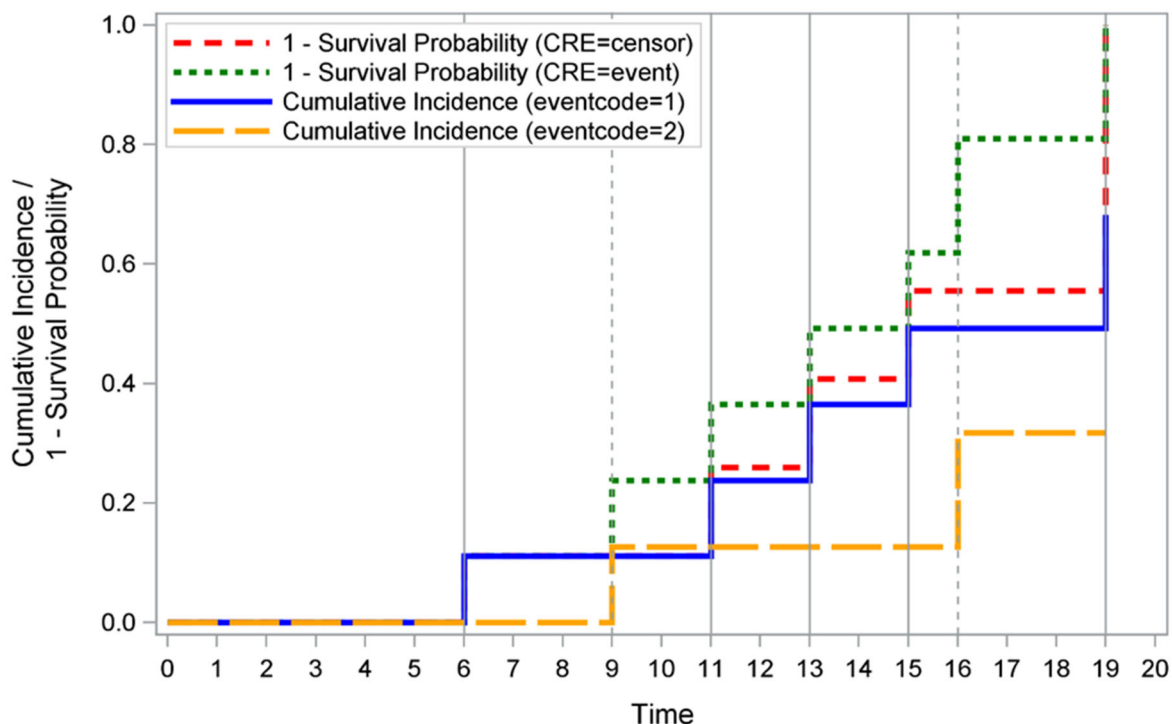


図5.2 イベントごとの累積発生関数，競合リスクをイベント，打ち切りとした場合の生存関数のグラフ

4.2. Gray 検定

Gray 検定の例として用いるデータは，造血幹細胞移植を受けた急性白血病患者 137 人(Bmt)のデータを使用する．Bmt データは参考文献[9]から入手可能である．ID 変数は被験者番号，Disease は疾患の種類，Dftime は生存時間，Status はイベントの種類，Gender は性別を示す変数である．Disease=1 は急性リンパ性白血病，Disease=2 は急性骨髄性白血病 低リスク，Disease=3 は急性骨髄性白血病 高リスクを示し，Status=0 は打ち切り，Status=1 は注目するイベント，Status=2 は競合イベントを示す．また，Gender=0 は女性，

Gender=1 は男性を示す。以下に、性別によって累積発生関数に差があるかを Gray 検定で行うプログラムと出力結果を示す。TIME ステートメントで Dftime*Status(0)で Dftime 変数を時間変数、Status=0 を打ち切りとして指定し、eventcode=1 で Status=1 を注目するイベントとして指定している。STRATA ステートメントに Gender を指定することで、性別間の累積発生関数が等しいという帰無仮説の下、性別によって注目するイベントの累積発生関数に差があるかを検定する。PLOTS オプションで CIF(TEST)と指定することで、累積発生関数のグラフ内に Gray 検定の p 値を出力することができる。また、TIMELIST オプションは指定した時点ごとの累積発生関数の推定値を出力する。今回、Gender にフォーマットを当てているため、STRATA ステートメントの ORDER オプションで INTERNAL と指定し、内部値順に累積発生関数の推定結果を出力するよう指定している。Gray 検定の結果には、 χ^2 値、自由度、p 値が出力される。2 群のため、検定統計量は自由度 1 の χ^2 分布に従う。p 値が 0.2958 であるため、有意水準 5%で帰無仮説は棄却されない。したがって、性別によって累積発生関数に差があるとは言えないという結果になる。

データセット Bmt				
ID	Disease	Dftime	Status	Gender
1	1	5.697467488	0	1
2	1	4.386036961	0	1
3	1	4.0958247775	0	1
4	1	4.0027378508	0	0
~				

```
proc lifetest data=Bmt plots=cif(test) timelist=0.5 1.0 1.5 2.0 4.0 6.0;
  time Dftime*Status(0) / eventcode=1;
  strata Gender / order=internal;
run;
```

Failed Event: Status=1

Summary of Failure Outcomes					
Stratum	Gender	Failed Events	Competing Events	Censored	Total
1	Female	20	16	21	57
2	Male	22	25	33	80
Total		42	41	54	137

Cumulative Incidence Function Estimates					
Stratum 1: Gender = Female					
Timelist	Dftime	Cumulative Incidence	Standard Error	95% Confidence Interval	
0.5	0.429843	0.2105	0.0546	0.1154	0.3248
1	0.744695	0.2648	0.0594	0.1573	0.3850
1.5	1.330595	0.3372	0.0640	0.2164	0.4620
2	1.711157	0.3553	0.0648	0.2318	0.4807
4	1.711157	0.3553	0.0648	0.2318	0.4807
6	1.711157	0.3553	0.0648	0.2318	0.4807

Cumulative Incidence Function Estimates					
Stratum 2: Gender = Male					
Timelist	Dftime	Cumulative Incidence	Standard Error	95% Confidence Interval	
0.5	0.298426	0.1000	0.0338	0.0464	0.1778
1	0.747433	0.1750	0.0428	0.1008	0.2661
1.5	1.24846	0.2250	0.0471	0.1403	0.3220
2	1.812457	0.2634	0.0498	0.1718	0.3641
4	2.047912	0.2763	0.0506	0.1826	0.3779
6	2.047912	0.2763	0.0506	0.1826	0.3779

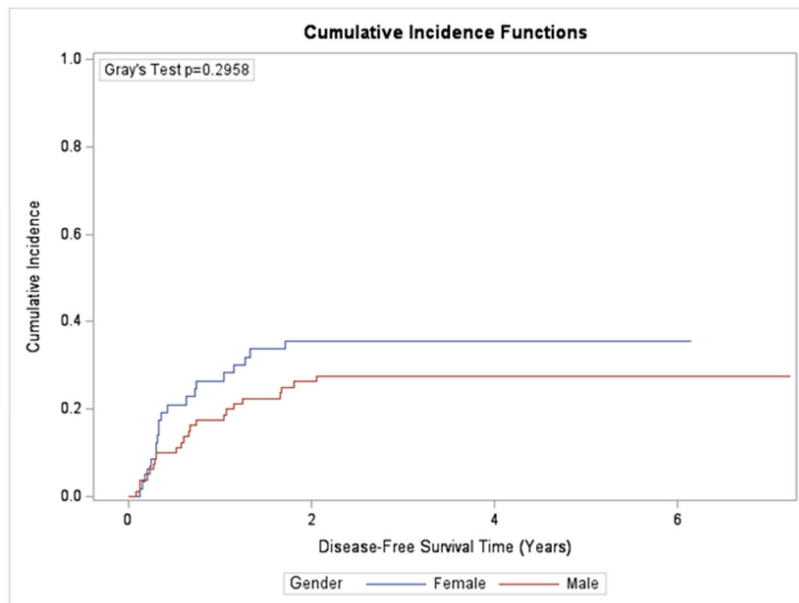


図 5.3 Status=1 を注目イベントとした性別ごとの累積発生関数のグラフ

Gray's Test for Equality of Cumulative Incidence Functions		
Chi-Square	DF	Pr > Chi-Square
1.0932	1	0.2958

4.3. 層別 Gray 検定

STRATA ステートメントの GROUP オプションに比較したい群を指定することで層別 Gray 検定も実装することができる。今回は、group=Disease と指定し、性別を考慮した上で疾患の種類ごとに累積発生関数に差があるかを検定した。3 群のため、検定統計量は自由度 2 の χ^2 分布に従う。p 値が 0.0028 であるため、有意水準 5% で帰無仮説を棄却する。したがって、性別を層とすると疾患の種類ごとの累積発生関数には差があると言えるという結果になる。

```
proc lifetest data=Bmt plots=cif(test) timelist=0.5 1.0 1.5 2.0 4.0 6.0;
  time Dftime*Status(0) / eventcode=1;
  strata Gender / group=Disease order=internal;
run;
```

Failed Event: Status=1						
Summary of Failure Outcomes						
Stratum	Gender	Disease	Failed Events	Competing Events	Censored	Total
1	Female	ALL	5	3	4	12
1	Female	AML-Low Risk	3	7	14	24
1	Female	AML-High Risk	12	6	3	21
Subtotal			20	16	21	57
2	Male	ALL	7	9	10	26
2	Male	AML-Low Risk	6	9	15	30
2	Male	AML-High Risk	9	7	8	24
Subtotal			22	25	33	80
Total			42	41	54	137

Cumulative Incidence Function Estimates						
Stratum 1: Gender = Female						
Disease	Timelist	Dftime	Cumulative Incidence	Standard Error	95% Confidence Interval	
ALL	0.5	0.353183	0.3333	0.1433	0.0938	0.6004
ALL	1	0.629706	0.4333	0.1591	0.1384	0.7022
ALL	1.5	0.629706	0.4333	0.1591	0.1384	0.7022
ALL	2	0.629706	0.4333	0.1591	0.1384	0.7022
ALL	4	0.629706	0.4333	0.1591	0.1384	0.7022
ALL	6
AML-Low Risk	0.5	0	0	0	.	.
AML-Low Risk	1	0.744695	0.0417	0.0419	0.00271	0.1810
AML-Low Risk	1.5	1.330595	0.1250	0.0695	0.0299	0.2918
AML-Low Risk	2	1.330595	0.1250	0.0695	0.0299	0.2918
AML-Low Risk	4	1.330595	0.1250	0.0695	0.0299	0.2918
AML-Low Risk	6	1.330595	0.1250	0.0695	0.0299	0.2918
AML-High Risk	0.5	0.429843	0.3810	0.1098	0.1768	0.5841
AML-High Risk	1	0.733744	0.4286	0.1125	0.2113	0.6302
AML-High Risk	1.5	1.278576	0.5238	0.1154	0.2834	0.7178
AML-High Risk	2	1.711157	0.5714	0.1155	0.3212	0.7590
AML-High Risk	4	1.711157	0.5714	0.1155	0.3212	0.7590
AML-High Risk	6

Cumulative Incidence Function Estimates						
Stratum 2: Gender = Male						
Disease	Timelist	Dftime	Cumulative Incidence	Standard Error	95% Confidence Interval	
ALL	0.5	0.298426	0.1154	0.0641	0.0280	0.2713
ALL	1	0.525667	0.1538	0.0725	0.0467	0.3188
ALL	1.5	1.048597	0.1923	0.0793	0.0678	0.3642
ALL	2	1.812457	0.2762	0.0922	0.1178	0.4613
ALL	4	1.812457	0.2762	0.0922	0.1178	0.4613
ALL	6
AML-Low Risk	0.5	0	0	0	.	.
AML-Low Risk	1	0.678987	0.1000	0.0558	0.0246	0.2389
AML-Low Risk	1.5	1.152635	0.1333	0.0633	0.0408	0.2810
AML-Low Risk	2	1.659138	0.1667	0.0695	0.0592	0.3213
AML-Low Risk	4	2.047912	0.2000	0.0747	0.0791	0.3602
AML-Low Risk	6	2.047912	0.2000	0.0747	0.0791	0.3602
AML-High Risk	0.5	0.273785	0.2083	0.0850	0.0734	0.3899
AML-High Risk	1	0.747433	0.2917	0.0956	0.1255	0.4812
AML-High Risk	1.5	1.24846	0.3750	0.1024	0.1840	0.5665
AML-High Risk	2	1.24846	0.3750	0.1024	0.1840	0.5665
AML-High Risk	4	1.24846	0.3750	0.1024	0.1840	0.5665
AML-High Risk	6	1.24846	0.3750	0.1024	0.1840	0.5665

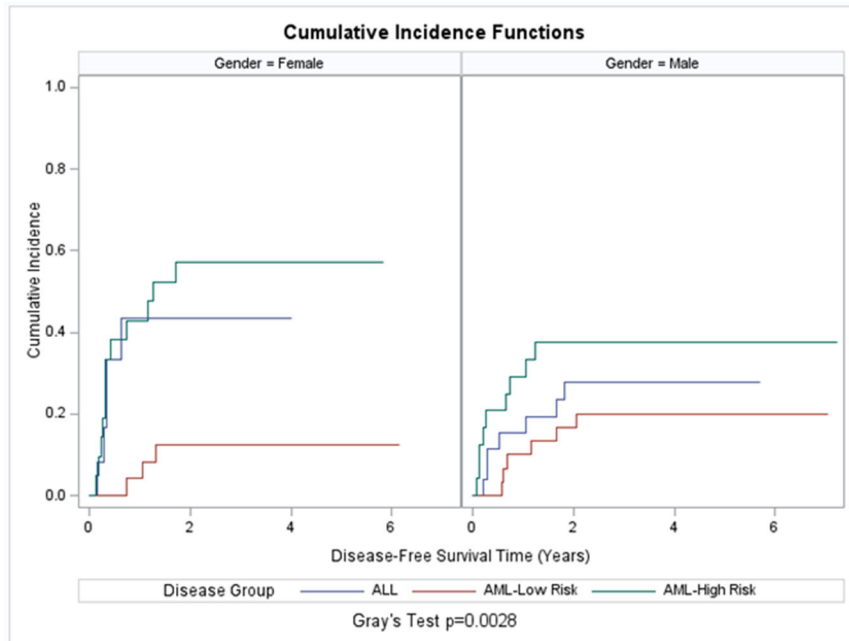


図 5.4 男女別の疾患の種類ごとの累積発生関数のグラフ

Gray's Test for Equality of Cumulative Incidence Functions		
Chi-Square	DF	Pr > Chi-Square
11.7625	2	0.0028

ここで、EVENTCODE オプションを複数指定する場合の挙動について解説する。EVENTCODE オプションで複数指定すると、指定した数値ごとに「異なる原因による」注目イベントとして扱い、それぞれの累積発生関数、Gray 検定の結果が出力される。TIME ステートメントの時間変数*イベント変数(打ち切り)で打ち切りを複数指定した場合、「区別することなく」指定した値をすべて同一の打ち切りとして扱う挙動と異なるため、注意が必要である。探索的な解析を行う際には、EVENTCODE オプションで複数指定することが有効である。

5. まとめ

本稿では、競合リスクを考慮した解析手法である累積発生関数と Gray 検定の概要、SAS での実装方法について解説をした。競合リスクを伴う場合、従来のように Kaplan-Meier 推定法を用いると、生存関数が過大に推定されること、累積発生関数を用いることでその問題が解決されることを数値例により示した。累積発生関数の推定、Gray 検定は LIFETEST プロシジャにより実装することが可能であり、EVENTCODE オプションや層別解析についても解説を行った。本稿が競合リスクを考慮した解析の基本を理解するための一助になれば幸いである。

6. 参考文献

- [1] Gray RJ. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. *Annals of Statistics* 1988; 16(3): 1141-1154.
- [2] 今泉敦・魚住龍史・浜田知久馬. 競合リスクが存在する下での生存時間分布の検定法の性能評価. SAS ユ

ーザー総会論文集 2018.

[3] 魚住龍史・浜田知久馬. SAS による競合リスクを伴う生存時間解析の理解. SAS ユーザー総会論文集 2018.

[4] 大橋靖雄・浜田知久馬・魚住龍史. 生存時間解析[第2版]-SASによる生物統計. 東京大学出版会, 2022.

[5] 西川正子. 生存時間解析における競合リスクモデル. 計量生物学 2008; 29(2): 141-170.

[6] 浜田知久馬. 生存時間解析の実務的課題の SAS での解決 競合リスク, 非 PH モデルの例数 設計, 最適症例数配分. SAS ユーザー総会 論文集 2017; 427.

[7] 宮岡悦良 監訳. 医薬統計のための生存時間データ解析[原著 第2版]. 共立出版株式会社, 2013

[8] SAS Institute Inc., SAS/STAT 15.3® User's Guide The LIFETEST Procedure

[9] SAS®9.4 and SAS® Viya® 3.5 Programming Documentation,

https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/statug/statug_code_liftex4.htm

(最終アクセス:2024/08/26)

文字コードに関するSASの対応

井上 貴博

(ノバルティス ファーマ株式会社)

SAS support for character encoding

Takahiro Inoue

Analytics and CDM Japan STATS PROG Cluster 1, Novartis Pharma K.K.

要旨

様々なデータが増え、文字化け等の予期せぬトラブルに遭遇することが多くなったが、SASにはオプション等が準備されておりトラブルを未然に防ぐことができる。本稿では文字コードを中心とした対処法を紹介する。

キーワード: Encoding, UTF-8, LATIN1, Shift_JIS, Shift-JIS, SJIS, SBCS, DSCS, MBCS, KPROPDATA

はじめに

長年にわたり、西洋諸国では英語および他の西洋言語の主要な文字コードとして LATIN1 および WLATIN1 が使用されてきました。一方、本邦では日本語を表すために Shift_JIS が Windows などの OS や SAS で広く使用されています。しかし、近年では Web 上で最も使用される文字コードとして UTF-8 が急速に普及しており、SAS Viya を含む多くのアプリケーションでも UTF-8 がデフォルトの文字コードとなりつつあります。

これまでは、各国で LATIN1、WLATIN1、Shift_JIS などの単一文字コードを使用して解析することが一般的でした。しかし、様々なデータを利用する機会が増えるにつれ、データの文字コードと SAS セッションのエンコードが異なるケースが増加し、文字化けなどの予期せぬトラブルに遭遇する可能性が高まっています。幸いにも、SASには様々なオプションが用意されており、適切に使用することでこれらのトラブルを回避することが可能です。本稿では、文字コードに関する適切な対処法を中心に紹介します。

なお、本稿は処理結果を再現できるよう SAS OnDemand for Academics を使用しています。

文字コード

コンピュータは電子機器であり、文字は番号(数字)の集まりとして保存される。文字を表す番号を文字符号、符号位置、コードポイントなどと呼び、文字と文字符号の関係を決めた対応表を「符号化文字集合(coded character set)」と呼び、一般的には「文字コード」と言われている。

以下に「SUJI-J」を例にして示す。

```

data _null_ ;
  text = 'SUJI-J' ;
  put text= / text=hex. / text=binary. ;
run ;

```

```

1      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69
70      data _null_ ;
71          text = 'SUJI-J' ;
72          put text= / text=hex. / text=binary. ;
73      run ;

text=SUJI-J
text=535547492D4A
text=01010011010101010001111010010010010110101001010
NOTE: DATA ステートメント処理(合計処理時間):

```

上記の実行結果(ログ)を表にまとめる。

String	S	U	G	I	-	J
hex.	53	55	47	49	2D	4A
binary	01010011	01010101	01000111	01001001	00101101	01001010

「SUJI-J」の 16 進数表記(hex.)は「535547492D4A」であることが分かる。「S」を表現する文字符号は 16 進数表記で「53」であり、符号化文字集合を使い「S」を表現していることが分かる。

	0	_1_	_2_	_3_	_4_	_5_	_6_	_7_	_8_	_9_	_A_	_B_	_C_	_D_	_E_	_F_
2_	(space)	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3_	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4_	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5_	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_

この「53」は 16 進数表記であり、2 進数表記では「01010011」となる。この「0」「1」一つが 1 ビットで 1 文字は 8 ビット、つまり 1 バイトで表現されている。

なお、通常の SAS データセットは保存領域がスペースで埋められてデータサイズが大きくなることが知られているが、この 16 進数表記(hex.)もしくは 2 進数表記(binary)では、見えないスペース(20/00100000)を確認することができる。

```

70      data _null_ ;
71          length text $6. ;
72          text = 'J' ;
73          put text= / text=hex. / text=binary. ;
74      run ;

text=J
text=4A2020202020
text=010010100010000000100000001000000010000000100000001000000
NOTE: DATA ステートメント処理(合計処理時間):

```

代表的な文字コード

SAS は様々な文字コードを選択することができるが、以下に代表的な文字コードを示す。

SBCS(Single-Byte Character Set): 1 バイト文字セット

- LATIN1(ISO 8859-1)

LATIN1、Latin1、ISO 8859-1 または Latin part1 と呼ばれる。拡張 ASCII コードの 1 つである。西ヨーロッパの文字データの一般的なエンコーディングであり、英語用の UNIX ホスト上の SAS セッションで使用される。

- LATIN9(ISO 8859-15)

LATIN1 と類似しており、混同されやすく、問題が生じる場合がある。8 つの異なる文字、特にユーロの文字「€」をサポートしている。LATIN9 は欧州の UNIX でよく使用されている。

- WLATIN1(Windows-1253)

Code page 1252 とも呼ばれる。LATIN1 のスーパーセットである。コード範囲[80-9F]に割り当てられた文字のみが異なり、追加文字が含まれている。これには、LATIN9 で追加された文字やいくつかの句読点、算術記号、引用符も含まれる。このコード範囲の引用符は「smart」とも呼ばれるが、ワープロソフトや一部のテキストエディタの自動修正機能でよく使用され、見た目も良いため ASCII 文字より好まれる傾向があるが、これらの文字は SAS ユーザーにとって悩みの種であり、予期しないトランスコーディングの問題につながる可能性がある。

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
8_	€		,	f	„	…	†	‡	^	%o	Š	<	Œ		Ž	
9_		‘	’	“	”	•	—	—	~	™	š	>	œ		ž	ÿ

DBCS(Double-Byte Character Set): 2 バイト文字セット

- Shift_JIS

Windows 及び UNIX ホスト上の日本語を表すために SAS が使用される。初期のマイクロソフト OS で採用されていたことから日本国内で広く使用されている。JIS X 0201(1 バイトのアルファベットとカタカナ)、JIS X 0208(かななど第二水準までの漢字コード)を組み合わせる可変長のエンコーディング方式である。

MBCS(Multi-Byte Character Set): 複数バイト文字セット

- UTF-8

Unicode は、世界中の文字を扱えるようにした ISO(国際標準化機構)規格の文字集合のこと。

UTF-8 は、Unicode の符号を 8 ビット単位で表す。複数言語をサポートしており、言語によって使用するバイト数が異なる。ASCII 文字との互換性であり、UTF-8 で使用できる最初の 128 文字は ASCII 文字と一致し、全く同じバイトで表現することができる。

Character length	Language
1 バイト	US_ASCII Characters
2 バイト	East and West European, Baltic, Greek, Turkish, Cyrillic, Hebrew, Arabic, and other supported character sets
3 バイト	Chinese, Japanese, Korean (CJK), Thai, Indic, and certain control characters
4 バイト	Emoji characters, less common CJK characters and various historic scripts

SAS セッションエンコーディング

通常のソフトウェアは、搭載するオペレーションシステムと同じエンコーディングのみを使用することが多いが、SAS はセッションエンコーディングとして、エンコーディングを指定することができる。セッションエンコーディングは、SAS 構文の SAS データセットを処理し、外部ファイルを読み書きするための環境で利用される。SAS セッションエンコーディングは、スタートアップ時ののみに設定可能で一般ユーザーは変更することは難しい。

まずは、自分が使用している SAS のセッションエンコーディングが何か確認することが必要であり、以下の構文にて確認することができる。

```
proc options option=encoding; run;
%put SESSION ENCODING is &sysencoding;
%put SESSION Encoding=%sysfunc(getoption(encoding));
```

実行結果(ログ)から使用している SAS セッションエンコーディングを知ることができ、以下に示す SAS OnDemand for Academics の場合、UTF-8 を使用していることが分かる。

```
69      proc options option=encoding; run;
      SAS (R) PROPRIETARY SOFTWARE RELEASE 9.4 TS1M7
ENCODING=UTF-8 SAS セッションのデフォルト文字セットエンコーディングを指定します。
NOTE: PROCEDURE OPTIONS処理(合計処理時間):
70      %put SESSION ENCODING is &sysencoding;
SESSION ENCODING is utf-8
71      %put SESSION Encoding=%sysfunc(getoption(encoding));
SESSION Encoding=UTF-8
72
```

UTF-8 での使用領域(length)

UTF-8 は、MBCS(Multi-Byte Character Set/複数バイト文字セット)で、言語によって使用するバイト数が異なる。

以下のテストデータを使用し、例を示す。

```
data test ;
      length name locale mean Frequency $30. ;
      input name $ locale $ mean $ Frequency $ ;
cards ;
Japanese_Japan ja_JP 平均 度数
Chinese_China zh_CN 均值 频数
Chinese_Taiwan zh_TW 平均值 次數
Korean_Korea ko_KR 평균 빈도
Polish_Poland pl_PL Średnia Liczebność
Thai_Thailand th_TH Mean Frequency
Thai google หมายถึง ความถี่
Hindi google अर्थ आवृत्ति
;
run ;
```

OBS	name	locale	mean	Frequency
1	Japanese_Japan	ja_JP	平均	度数
2	Chinese_China	zh_CN	均值	频数
3	Chinese_Taiwan	zh_TW	平均值	次數
4	Korean_Korea	ko_KR	평균	빈도
5	Polish_Poland	pl_PL	Średnia	Frequency
6	Thai_Thailand	th_TH	Mean	Frequency
7	Thai	google	หมายถึง	ความถี่
8	Hindi	google	अर्थ	आवृत्ति

前述の通り、SAS OnDemand for Academics はセッションエンコードで UTF-8 のため、この環境で確認していく。

なお、LOCALE オプションで表示言語を変更することができ、これを元に今回のサンプルデータは作成した。Thai と Hindi は SAS では英語表記のままであったため、翻訳サイトで翻訳したものも追加している。

options locale=ko_KR;		locale=ko_KR					options locale=th_TH;		locale=th_TH				
title "locale=ko_KR" ;		FREQ 프로시저					title "locale=th_TH" ;		The FREQ Procedure				
proc freq data = sashelp.class ;		Sex	빈도	백분율	누적 빈도	누적 백분율	proc freq data = sashelp.class ;		Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
table sex ;		F	9	47.37	9	47.37	table sex ;		F	9	47.37	9	47.37
run ;		M	10	52.63	19	100.00	run ;		M	10	52.63	19	100.00

以下の構文で、バイト数と文字数を確認する。

```
data leng ;
  set test ;
  l_mean = length(mean) ;      kl_mean = klength(mean) ;
  l_freq = length(Frequency) ; kl_freq = klength(Frequency) ;
run ;
proc print data = leng width = min ;
  var name locale mean l_mean kl_mean Frequency l_freq kl_freq ;
run ;
```

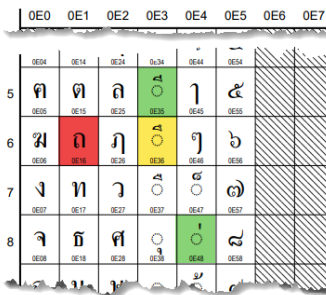
LENGTH 関数はバイト単位で長さ、KLENGTH 関数は引数内で最も右にある空白以外の文字の位置を表す整数を返すため、文字数と考えてよい。以下が実行結果(アウトプット)となる。

OBS	name	locale	mean	l_mean	kl_mean	Frequency	l_freq	kl_freq
1	Japanese_Japan	ja_JP	平均	6	2	度数	6	2
2	Chinese_China	zh_CN	均值	6	2	频数	6	2
3	Chinese_Taiwan	zh_TW	平均值	9	3	次數	6	2
4	Korean_Korea	ko_KR	평균	6	2	빈도	6	2
5	Polish_Poland	pl_PL	Średnia	8	7	Frequency	9	9
6	Thai_Thailand	th_TH	Mean	4	4	Frequency	9	9
7	Thai	google	หมายถึง	21	7	ความถี่	21	7
8	Hindi	google	अर्थ	12	4	आवृत्ति	21	7

「Frequency」は、ASCII characters であるため、9 バイト・9 文字で 1 文字あたり 1 バイトが使用されている。Polish の「Średnia」は、「Ś」が 1 文字あたり 2 バイトを使用しているため、8 バイト・7 文字となった。Japanese や Chinese 等の文字は、1 文字あたり 3 バイトが必要である。

なお、Thai の「ความถี่」は一見すると 5 文字だが「ถี่」は 16 進数表記で $\u0e16\u0e35\u0e48$ であり、1 つの子音に対して、2 つの母音が付与されており、3 文字分の領域を使用しているようだ。

```
70      data _null_ ;
71      length text $9. ;
72      text="ถี่" ; l=length(text) ; kl=klength(text) ;
73      uesc=kpropdata(text, "UESC", "UTF-8", "US-ASCII") ;
74      put text= l= kl= uesc= ;
75      text="ถี่" ; l=length(text) ; kl=klength(text) ;
76      uesc=kpropdata(text, "UESC", "UTF-8", "US-ASCII") ;
77      put text= l= kl= uesc= ;
78      run ;
text=ถี่ l=6 kl=2 uesc=\u0e16\u0e36
text=ถี่ l=9 kl=3 uesc=\u0e16\u0e35\u0e48
NOTE: DATA ステートメント処理(合計処理時間):
```



また、Hindi の「अर्थ」は英語の筆記体のようにつなげて表現するらしく、「अ」で 1 文字、「र्थ」は थ+र+् で 3 文字の領域を使用するため、見た目の文字数で使用領域を判断できない。

データセットのエンコード

SAS 9 および SAS Viya で作成されたデータセットは、ヘッダーにエンコーディング情報が格納されている。

データセットのエンコードは CONTENTS プロシジャで確認する。

```
proc contents data = sashelp.class ; run ;
```

右図の例では、SAS OnDemand for Academics の

SASHELP.CLASS のエンコードは、us-ascii ASCII (ANSI)であることが分かる。

CONTENTS プロシジャ			
データセット名	SASHELP.CLASS	オブザベーション数	19
メンバータイプ	DATA	変数の数	5
エンジン	V9	インデックス数	0
作成日時	2020/08/06 08:59:14	オブザベーションのバッファ長	40
更新日時	2020/08/06 08:59:14	前読済みオブザベーション数	0
保護		圧縮済み	NO
データセットタイプ		ソート済み	NO
ラベル	Student Data		
データ表種	SOLARIS_X86_64_LINUX_X86_64_ALPHA_TRU64_LINUX_A64		
エンコード	us-ascii ASCII (ANSI)		

デフォルトは、データセット作成に用いた SAS セッションのセッションエンコーディングが設定される。使用するデータと SAS セッションエンコーディングは一致することが推奨されるが、異なる場合、トランスコーディングが起きる。

以下の構文を例にして示す。

```
data test(encoding = 'SJIS'); text = 'Hello World!'; run ;
data wk010; set test; run ;
proc contents data = work.wk010 ;run ;
```

SJIS(Shift_JIS)で作成したデータセット「test」を元に「wk010」を作成したが、データセットはセッションエンコードの UTF-8 で作成される。ログにエンコードが一致せず、トランスコーディングで行われていることが示される。

データセット名	WORK.WK010	73 data wk010 ;
メンバータイプ	DATA	74 set test ;
ラベル		NOTE: データファイルWORK.TEST.DATAは別なホストにネイティブな形式が使用されているか、またはエンコーディングがセッションエンコーディングと一致していません。クロス環境データアクセスが使用されるため、追加のCPUリソースが必要となり、パフォーマンスが低下します。
データ表現	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64	75 run ;
エンコード	utf-8 Unicode (UTF-8)	NOTE: データセットWORK.TESTから1オブザベーションを読み込みました。
		NOTE: データセットWORK.WK010は1オブザベーション、1変数です。
		NOTE: DATA ステートメント処理(合計処理時間):

作成するデータセットのエンコードを SJIS(Shift_JIS)と指定した場合と同様である。以下の実行結果(ログ)が示すように、input 時にトランスコーディングされ、SAS セッションコーディングの UTF-8 で処理され、output 時に再度トランスコーディングされている。また、バイト数を確認する変数 L が 21 であることから、UTF-8 で処理していることが分かる。

```
75 data wk020(encoding = 'SJIS') ;
76 set test ;
NOTE: データファイルWORK.TEST.DATAは別なホストにネイティブな形式が使用されているか、またはエンコーディングがセッションエンコーディングと一致していません。クロス環境データアクセスが使用されるため、追加のCPUリソースが必要となり、パフォーマンスが低下します。 input
77 l = length(text) ;
78 put l = ;
79 run ;
NOTE: データファイルWORK.WK020.DATAは別なホストにネイティブな形式が使用されているか、またはエンコーディングがセッションエンコーディングと一致していません。クロス環境データアクセスが使用されるため、追加のCPUリソースが必要となり、パフォーマンスが低下します。 output
1-21
NOTE: データセットWORK.TESTから1オブザベーションを読み込みました。
NOTE: データセットWORK.WK020は1オブザベーション、2変数です。
NOTE: DATA ステートメント処理(合計処理時間):
```

なお、以下の実行結果(ログ/アウトプット)が示すように、エンコードを指定しデータセットを作成し、そのデータセットを上書きする場合、エンコード情報は上書きされず、以前のエンコードを保持するため、注意が必要である。

70 data wk010(encoding = 'SJIS') ;	CONTENTS プロシジャ		
71 a = 'あいえお' ;			
72 run ;	データセット名	WORK.WK010	オ
NOTE: データファイルWORK.WK010.DATAは別なホストにネイティブな形式が使用されているか、またはエンコーディングがセッションエンコーディングと一致していません。クロス環境データアクセスが使用されるため、追加のCPUリソースが必要となり、パフォーマンスが低下します。	メンバータイプ	DATA	変
NOTE: データセットWORK.WK010は1オブザベーション、1変数です。	エンジン	V9	イ
NOTE: DATA ステートメント処理(合計処理時間):	作成日時	2024/08/20 17:27:00	オ
処理時間: 00.01	更新日時	2024/08/20 17:27:00	削
73 data wk010 ;	保護		圧
74 a = 1 ;	データセットタイプ		ソ
76 run ;	ラベル		
NOTE: データファイルWORK.WK010.DATAは別なホストにネイティブな形式が使用されているか、またはエンコーディングがセッションエンコーディングと一致していません。クロス環境データアクセスが使用されるため、追加のCPUリソースが必要となり、パフォーマンスが低下します。	データ表現	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64	
NOTE: データセットWORK.WK010は1オブザベーション、1変数です。	エンコード	shift-jis Japanese (SJIS)	
NOTE: DATA ステートメント処理(合計処理時間):			

トランスコーディング

以下の構文を使用し説明する。

```
data _null_ ;
do text = 'ㄥ', 'ㄹ', 'ㄴ' ;
UTF8=kpropdata(text, "HEX", "UTF-8", "US-ASCII") ;
SJIS=kpropdata(kpropdata(text, "TRIM", "UTF-8", "SJIS"), "HEX", "SJIS", "US-ASCII") ;
put text= / UTF8= / SJIS= ;
end ;
run ;
```

一文字ずつ、UTF-8 と SJIS(Shift_JIS)の 16 進数表記で表示するもので、実行結果(ログ)は以下のようになる。

```

70     data _null_ ;
71         do text = '\', '□', 'ー' ;
72             UTF8=kpropdata(text, "HEX", "UTF-8", "US-ASCII") ;
73             SJIS=kpropdata(kpropdata(text, "TRIM", "UTF-8", "SJIS"), "HEX", "SJIS", "US-ASCII") ;
74             put text= / UTF8= / SJIS= ;
75         end ;
76     run ;

text=\\
UTF8=\xe3\x83\x8f
SJIS=\x83\x6e
text=□
UTF8=\xe3\x83\xad
SJIS=\x83\x8d
text=ー
UTF8=\xe3\x83\xbc
SJIS=\x81\x5b
NOTE: DATA ステートメント処理(合計処理時間):

```

「ハ」は、UTF-8 では「¥xe3¥x83¥x8f」、SJIS(Shift_JIS)で「¥x83¥x6e」となり、内部で保持する番号(文字符号)が異なることが分かる。このように「ハ」という文字表現は変更せずに、文字符号を変更することをトランスコーディングという。

なお、文字符号に対して、適切な文字コードをしないうち、以下の例のように文字化けが起きる。

```

70     data _null_ ;
71         do text = '\', '□', 'ー', 'ワ', 'ー', 'ル', 'ド' ;
72             SJIS_text=kpropdata(text, "TRIM", "UTF-8", "SJIS") ;
73             SJIS=kpropdata(SJIS_text, "HEX", "SJIS", "US-ASCII") ;
74             put text= / SJIS= SJIS_text= ;
75         end ;
76     run ;

text=\\
SJIS=\x83\x6e SJIS_text=◆n
text=□
SJIS=\x83\x8d SJIS_text=◆◆
text=ー
SJIS=\x81\x5b SJIS_text=◆[
text=ワ
SJIS=\x83\x8f SJIS_text=◆◆
text=ー
SJIS=\x81\x5b SJIS_text=◆[
text=ル
SJIS=\x83\x8b SJIS_text=◆◆
text=ド
SJIS=\x83\x68 SJIS_text=◆h
NOTE: DATA ステートメント処理(合計処理時間):

```

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
2_	(space)	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3_	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4_	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5_	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6_	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7_	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

上記は、SAS セッションエンコードが UTF-8 の SAS OnDemand for Academics で、SJIS(Shift_JIS)に変換したものを表示しようとした。この場合、「ハ」は「¥x83¥x6e」であり、「¥x6e」は符号化文字集合(coded character set)に該当する「n」が表示され、「¥x83」は該当する文字がなく「◆」が表示された。

トランスコーディング時の Length の拡張

SAS は、自動的にトランスコーディングするが、Length 等の保存領域は拡張されない。

以下のサンプルデータを用いて、例を示す。

データセット名	SJIS	
エンコード	Shift-jis Japanese(SJIS)	
変数	タイプ	長さ
text	文字	14

```

text
ハローワールド

```

このサンプルデータを PRINT プロシジャで出力すると実行結果(ログ/アウトプット)は、以下になる。

```
69      libname data '/home/u63934911/data' ;
NOTE: ライブラリ参照名DATAを次のように割り当てました。
      エンジン: V9
      物理名: /home/u63934911/data
70
71      proc print data = data.sjis ;
NOTE: データファイルDATA.SJIS.DATAは別なホストにネイティブな形式が使用されているか、またはエン
      コーディングがセッションエンコーディングと一致していません。クロス環境データアクセス
      が使用されるため、追加のCPUリソースが必要となり、パフォーマンスが低下します。
72      run ;

WARNING: データセットDATA.SJISのトランスコード時に文字データが一部損失しました。新しいエンコー
      ディングで表せない文字がデータに含まれていたか、またはトランスコード時に切り捨てが
      発生しました。
NOTE: データセットDATA.SJISから1オブザベーションを読み込みました。
NOTE: PROCEDURE PRINT処理(合計処理時間):
```

OBS	text
1	ハローワ

SJIS(Shift_JIS)の「ハローワールド」が自動的にトランスコーディングされ識読可能ではあるが、7文字に対してSJISの保存領域限界14バイトの変数であるため、UTF-8では領域不足で文字の切り捨てが生じる。

なお、この文字の切り捨ては、SAS OnDemand for Academicsでは以下の特徴があった。

```
72      if text = "ハローワールド" then fr1 = 1 ;
73      if text = substr("ハローワールド", 1, 14) then fr2 = 1 ;
74      if text = "ハローワ" then fr3 = 1 ;
75      put text= fr1= fr2= fr3= ;
76      run ;

WARNING: データセットDATA.SJISのトランスコード時に文字データが一部損失しました。新しいエンコー
      ディングで表せない文字がデータに含まれていたか、またはトランスコード時に切り捨てが
      発生しました。
text=ハローワ fr1=. fr2=. fr3=1
NOTE: データセットDATA.SJISから1オブザベーションを読み込みました。
NOTE: DATA ステートメント処理(合計処理時間):
```

14バイトで保存され、14バイトを超える文字符号が切り捨てられる場合、上記のfr2にフラグが立つことが予想される。しかし、上記の結果では、fr3にフラグが立ち、「ハローワ」という14バイト以内の識読可能な文字で切り捨てられていることが分かる。

Length 拡張には、LIBNAME ステートメントの CVP エンジンが有用である。

```
libname data cvp '/home/u63934911/data' inencoding=sjis ;
data utf8data ;
      set data.sjis ;
run ;
```

上記のように記載した場合、データセット「utf8data」の変数は、1.5倍に拡張される。

デフォルトは1.5倍に拡張されるため、Shift_JIS(DBCS)からUTF-8(MBCS)の変換は問題がないと思われるが、注意が必要な場合がある。

例えば、WLATIN1をUTF-8に変換した場合である。UTF-8で使用できる最初の128文字はASCII文字と一致し、全く同じバイトで表現することができるため誤解されやすいが、UTF-8はSBCSの全ての文字を1バイトで表現できるわけではない。

以下に WLATIN1 を UTF-8 に変換した場合に注意する例を示す。

```
69     data _null_ ;
70     do text = "€", " ", "α", " ", "€" ;
71         wlatin1=kpropdata(kpropdata(text, "TRIM", "UTF-8", "wlatin1"), "HEX", "wlatin1", "US-ASCII") ;
72         UTF8=kpropdata(text, "HEX", "UTF-8", "US-ASCII") ;
73         put text= / wlatin1= / UTF8= ;
74     end ;
75     run ;

text=
wlatin1=\x91
UTF8=\xe2\x80\x98
text=
wlatin1=\x92
UTF8=\xe2\x80\x99
text=
wlatin1=\x93
UTF8=\xe2\x80\x9c
text=
wlatin1=\x94
UTF8=\xe2\x80\x9d
text=€
wlatin1=\x80
UTF8=\xe2\x82\xac
NOTE: DATA ステートメント処理(合計処理時間):
```

上記の例のように「smart」と呼ばれる引用符や拡張された記号などは、3 バイトが必要となる。

切り捨てはログに表示されるため、ログを注視する。

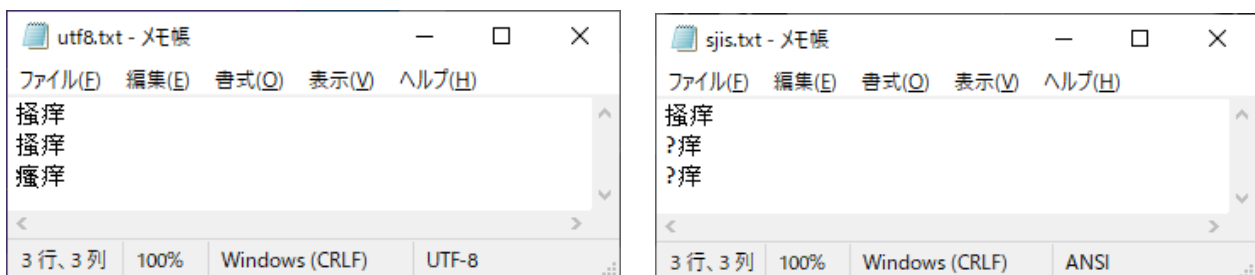
WARNING: データセットDATA.SJISのトランスコード時に文字データが一部損失しました。新しいエンコーディングで表せない文字がデータに含まれていたか、またはトランスコード時に切り捨てが発生しました。

LIBNAME ステートメントの CVP エンジンには、1.5 倍などの乗数の指定のほかに、加算する整数の指定方法のオプションが準備されているため、切り捨てが発生した場合は、オプションを利用して、適切に長さを拡張する必要がある。

文字コード間の対応している文字の違い

各文字コードのサポートしている文字が異なる点も、注意が必要である。

以下に「そう痒」の例を示す。「搔痒」「搔痒」「瘙痒」を UTF-8 で作成し(左図)、Shift_JIS を再度保存した(右図)。



左図の UTF-8 では全ての文字が識読可能であるが、右図の Shift_JIS では「?」となる文字が含まれている。

これは、「搔」と「瘙」は Shift-JIS では非対応で、代替文字の「?」に置き換えられているためである。なお、前述で示した「◆」は符号位置を保持しているため適切な文字コードで対応し識読可能になる場合はあるが、代替文字の「?」に置き換えられた後では、識読可能な状態に復元することはできない。そのため、使用するデータが、どの文字コードを使用しているか、また文字化け等により代替文字に置き換わっていないかの確認が重要となる。

SAS OnDemand for Academics のようにセッションエンコードが UTF-8 で使用するデータが文字化けしていないのであれば、Shift_JIS で非対応の文字は以下の構文で探すことができる。

```
if text ne kpropdata(kpropdata(text,'TRIM','UTF-8','SJIS'),'TRIM','SJIS','UTF-8') then put 'ER' 'ROR:' text= ;
```

```
70 data _null_ ;
71 do text = 'そう痒', '掻痒', '掻痒', '瘙痒' ;
72 if text ne kpropdata(kpropdata(text,'TRIM','UTF-8','SJIS'),'TRIM','SJIS','UTF-8') then put 'ER' 'ROR:' text= ;
73 end ;
74 run ;
```

ERROR: text=掻痒

ERROR: text=瘙痒

NOTE: DATA ステートメント処理(合計処理時間):

セッションエンコードが UTF-8 の環境で解析し、レポートなどを Shift_JIS で作成するなどの機会にご活用いただきたい。

おわりに

技術の進化に伴い、世界はかつてないほどつながりを見せています。データは世界中から集められ、Unicode とその UTF-8 は、これらの異なる言語を表現してくれます。SAS は、多言語をサポートしており、自動でトランスコーディングされる等の便利な一面はありますが、無意識にデータを扱って良いわけではありません。トラブルを未然に防ぐには、正しい理解とオプション等を適切に使用する必要があります。本稿は、様々な型式のデータを扱う SAS ユーザーの一助になれば幸いです。

参考文献

- [1] Base SAS Software - SAS 9.4 & SAS Viya 3.4
<https://support.sas.com/en/software/base-sas-support.html#documentation>
(参照: 2024-8-19)
- [2] Programming Documentation for SAS 9.4 and SAS Viya
https://go.documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=pgmsashome&docsetTarget=home.htm&locale=en
(参照: 2024-8-19)
- [3] SAS 9.4 と SAS Viya 3.4 プログラミングドキュメント
https://go.documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=pgmsashome&docsetTarget=home.htm&locale=ja
(参照: 2024-8-19)
- [4] The SAS Encoding Journey: A Byte at a Time
Mickaël Bouedo, SAS Institute
SAS Global forum, Paper SAS4561-2020
<https://support.sas.com/resources/papers/proceedings20/4561-2020.pdf>
(参照: 2024-8-19)
- [5] Have a Comprehensive understanding of SAS® K functions
Leo (Jiangtao) Liu, SAS Institute Inc.
Paper SAS1902-2018
<https://support.sas.com/resources/papers/proceedings18/1902-2018.pdf>
(参照: 2024-8-19)

SASのCALLルーチンを可能な限り全て説明

森岡 裕

(イーピーエス株式会社)

Showcase all SAS Call Routines as much as possible

Yutaka Morioka

(EPS Corporation)

SAS CALLルーチンは引数で指定した変数の値を変更したり、システム動作を起こすなどの機能を持っている。CALLルーチンは関数に似てるが、割り当てステートメントや式の中で使用できない点で関数とは異なる。すべてのSAS CALLルーチンは、CALLステートメントで呼び出される。

SAS CALLルーチンには様々な種類があり、状況によって使い分ければ、関数やプロシジャでの処理よりも有効なこともあるが、単純に存在や機能がユーザーに知られていないために、うまく活用されていないケースがあるようにも思われる。

本発表では以下の 57 個の CALL ルーチンについて説明する。以下はあくまで Base のリファレンスに紐づく CALL ルーチンであり、下記以外にも CALL ルーチンは多く存在している。

CALL ALLCOMB Routine	CALL POKELONG Routine	CALL SCAN Routine
CALL ALLCOMBI Routine	CALL PRXCHANGE Routine	CALL SET Routine
CALL ALLPERM Routine	CALL PRXDEBUG Routine	CALL SLEEP Routine
CALL CATS Routine	CALL PRXFREE Routine	CALL SOFTMAX Routine
CALL CATT Routine	CALL PRXNEXT Routine	CALL SORT Routine
CALL CATX Routine	CALL PRXPOSN Routine	CALL SORTC Routine
CALL COMPCOST Routine	CALL PRXSUBSTR Routine	CALL SORTN Routine
CALL EXECUTE Routine	CALL RANBIN Routine	CALL STDIZE Routine
CALL GRAYCODE Routine	CALL RANCAU Routine	CALL STREAM Routine
CALL IS8601_CONVERT Routine	CALL RANCOMB Routine	CALL STREAMINIT Routine
CALL LABEL Routine	CALL RANEXP Routine	CALL STREAMREWIND Routine
CALL LEXCOMB Routine	CALL RANGAM Routine	CALL SYMPUT Routine
CALL LEXCOMBI Routine	CALL RANNOR Routine	CALL SYMPUTX Routine
CALL LEXPERK Routine	CALL RANPERK Routine	CALL SYSTEM Routine
CALL LEXPERM Routine	CALL RANPERM Routine	CALL TANH Routine
CALL LOGISTIC Routine	CALL RANPOI Routine	CALL TSO Routine
CALL MISSING Routine	CALL RANTBL Routine	CALL VNAME Routine
CALL MODULE Routine	CALL RANTRI Routine	CALL VNEXT Routine
CALL POKE Routine	CALL RANUNI Routine	CALL WTO Routine

Can SAS Monolinguals

who make full use of Programming Code Translator

Survive?

仲家 諒

(武田薬品工業株式会社)

製薬業界で医薬品開発業務に従事する多くのSASプログラマーの最大の関心事は、RやPythonといったオープンソースソフトウェア（OSS）のリスクリングである。これらのOSS利用拡大は、昨今のAI/MLの大規模な流行やそれに伴うデータサイエンス領域の貢献度向上に基づいており、その利用拡大は幅広い分野で見られ、製薬業界の医薬品開発業務においても例外ではない。そのため、既存のSASプログラマーは近い将来のRやPythonの利用に備えるためこれらの学習を進めていく必要がある。

本発表では、昨今の発展が目覚ましい生成AIによるプログラミング言語の翻訳をフル活用すれば、もしかするとRやPythonを学習せずともSAS一本でこの先の将来も生きていくことができるのではないか？という仮説を立て検討した。検討結果とともに検討に際しての気づきも併せて共有する。なお、今回の検討ではSAS→Rのコード変換をGitHub Copilot Chatを用いて実施した。

A Modelling Framework for Regression with Collinearity

刈屋武昭³

(元名古屋商科大学、一橋大学名誉教授)

定数項を含む線形回帰モデルを

$$(1.1) \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad E(\mathbf{u}) = \mathbf{0}, \quad \text{Var}(\mathbf{u}) = \sigma^2 \mathbf{I}, \quad \mathbf{X}: N \times K$$

とする。このモデルにおいて、被説明変数 \mathbf{y} は説明変数行列 \mathbf{X} の重共線性を識別する能力がないので、次の伝統的な分析プロセス

(1.1) のモデルの指定— \mathbf{y} によるモデルの推定—変数モデル選択—事後診断

では重共線性の扱い方は正しくない。したがって、モデルが \mathbf{X} の列ベクトルからなす部分行列の集合全体になることから、

モデルの集合全体 $\{\mathbf{x}_\alpha\}$ —有効なサブモデル集合 \mathcal{H} の選択 (効率性・安定性条件を課す)
— \mathbf{y} を利用した AIC, 最尤法等による最適モデルの選択—有効性の確認

というモデル選択プロセスに変更する方法を提案した。最終的に求めるモデルが実証的に有効な回帰モデルであることを最初から、有効性の基準を満たさないモデルを排除しておいて、残りのモデルのクラス \mathcal{H} から、通常モデル推定プロセスにより最適モデルを選択する。したがって研究の基本的視点と制御対象とするモデル構造を次のように変更する。まず、有効性を毀損する K 個の説明変数行列 \mathbf{X} の構造にかかる 2 つのリスクとして、

1) 不安定性リスクとしてのマルチコ (Collinearity) リスク (C-リスク)

2) 最小 2 乗推定値 $\hat{\boldsymbol{\beta}}$ の非効率性 (Inefficiency) リスク (I-リスク)

を選択し、これらのリスク指標を制御する問題を設定した。ここで C-リスクは個別変数の VIF である。具体的には、各説明変数 k の N 個の平均予測分散

$$\text{IndPSV}_k \equiv \sum_{n=1}^N \text{Var}(x_{nk} \hat{\beta}_k) / N = \sigma^2 \frac{1}{N} [I_k \times C_k]$$

$$I_k \equiv \bar{x}_k^2 / s_k^2, \quad C_k \equiv \text{VIF}_k = 1 / (1 - \check{R}_k^2) \geq 0$$

において、与えられた (c, d) に対して $I_k \leq c$ 、 $C_k \leq d$ がすべての k について一様に成立するように変数の組のモデルの集合をアルゴリズムによって選択する。

一方、このモデル制御の目的を実行する方法として、1 節で述べたアルゴリズムと同様な考え方による変数増加的な方法と、上のモデル選択における制御問題を同時的に扱う、 \mathbf{X} の主成分分析による結果をモデル選択に利用する方法を発展させる。そこでは、第一主成分との相関係数が 0.9 (0.95) 以上となる変数を同一モデルに利用しない形でマルチコを避けたモデルの集合を導出。そのモデルに対して、AIC 等の最適化モデル選択基準で、最終的なモデルを選択する。応用例を示す。

Dataset-JSONのSASでの実装とSAS拡張属性の利用について

森岡 裕, 中川 雄貴

(イーピーエス株式会社)

Creating Dataset-JSON Using proc JSON and Extended Attribute in SAS

Yutaka Morioka, Yuki Nakagawa

(EPS Corporation)

SAS Version 5 (V5) transport file format, いわゆる version 5 の XPT フォーマットファイルについては仕様が公開されており, CDISC 形式での電子承認申請において, FDA をはじめとして規制当局が受け入れてることから, 事実上, CDISC 形式での電子承認申請における唯一の転送ファイル形式となっている.

しかし, XPT の規格策定からも, 受け入れ開始からも既に長い時間が経過しており, XPT V5 の持つ規格的な制限の問題が顕在化しつつある. そこで CDISC/FDA が中心となり, 新しいファイル転送の提言と検討がなされている. 新しい規格は, FDA 協力のもと受け入れのパイロットスタディが実施されており, そのフィードバック結果が良好であったことも含め, 近年, Dataset-JSON について注目度が増している.

CDISC の国際的なシンポジウムの日本版である 2024 CDISC Japan Interchange においても, FDA からパイロットスタディのフィードバックと今後の展望, また製薬メーカーおよび CRO から発表があり, Dataset-JSON をメインテーマとしたものが 3 演題あがった. 本発表はその一つの「Creating Dataset-JSON Using proc JSON and Extended Attribute in SAS」を SAS ユーザー総会用にカスタマイズしたものとなる.

SAS には Proc JSON という JSON ファイルを自由に作成可能なプロシジャがあり, また Libname JSON エンジンを使って, JSON ファイルを読み込むことも簡単にできる.

ただし, Dataset-JSON はファイル転送形式そのものに豊富なメタデータを格納する仕様となっているが, SAS データセットには, 型としては数値型・文字型の二つしかく, 設定できるメタデータも少ない. そのため SAS データセットと Dataset-JSON の変換の際に情報量の齟齬が発生する.

その問題に対するアプローチの一つとして SAS 拡張属性の利用を提案した. SAS 拡張属性は SAS v9.4 から実装された機能で, ユーザーが自由に属性を作成して, データセットレベルや変数レベルで付与することが可能である

本発表では, Dataset-JSON の存在自体の周知・必要性の啓蒙と, SAS での実装の技術的側面について紹介したい.

参考

- [1] Yutaka Morioka, Yuki Nakagawa, Satoru Orii, Li Zhang, and Hiroshige Takata (13 June 2024). “Creating Dataset-JSON Using proc JSON and Extended Attribute in SAS”, 2024 CDISC Japan Interchang.

SASによるYAMLファイルの作成

森岡 裕

(イーピーエス株式会社)

Generating YAML files with SAS

Yutaka Morioka

(EPS Corporation)

YAML (ヤメル, ヤムル) は、構造化データを人間にも見やすく、機械でも読みやすくしたファイルフォーマットである。

YAML は「**YAML Ain't a Markup Language**」と、再帰的な略称が主張するとおり、マークアップ言語ではない。スカラー (単一値)、シーケンス (リスト・配列) とマッピング (ハッシュ・辞書) という構造を組み合わせることで構造化データやオブジェクトを文字列としてシリアライズ (直列化) できるデータ形式である。要するに複雑な構造を1連の文字列だけで表現できる。

表現の機能で比較するとJSONが近く、JSON形式とYAML形式は相互変換性が高い。

ただし、YAMLの方が階層構造の表現がシンプルで、人間にも読みやすく、プログラム言語のようにコメントを内容に含めることができるため、まさにプログラム仕様のような手続きの記述や、人間も直接読み書きする何かの設定の記録をする場合にYAMLのほうが使い勝手がよいことも多い。

医薬領域でいうと、CDISCバリデータのPinnacle CommunityにおいてCDISC CORE (ここではオープンソースのバリデーションルールのようなものと単純化する) を利用する際の、各バリデーションルールの設定はYAML形式で作成されている。またCDISC ARS (Analysis Results Standard)における各Componentの記述はYAML形式になっている。

他にもSAS Viyaの設定ファイルはYAML形式であったりと、設定ファイルやAPI仕様がYAMLで記載されていることは比較的多い。

SASにはproc JSONはあるが、proc YAMLは存在しないため、他の機能を組み合わせてproc JSONに近い仕組みでYAMLを出力できないかを検討したのが本発表の根幹である。

プレーンテキストを作成することができるproc Streamをベースに、作成例を提示した。proc JSONにはSASデータセットの内容を出力できるexportステートメントが存在するが、Proc Streamにはそういった機能はなく、本来SASのプロシジャステップの中でデータステップを起こすことはできない。ただし、唯一の例外として、dosubl関数がある。dosubl関数はメモリ上に独自の別領域を構成し、そのなかで引数に指定した文字列 (サブライン) を実行することが可能である。別領域から現領域に、マクロを介して値を戻すことができるので、これを利用して疑似的にexportの仕様を実装した。

ただし、本発表の内容はあくまでお手本に過ぎない。SASに実装されていない、データ記述言語を読み書きする必要が生じた場合、こういったアプローチができるという指針を示したかったものであり、本発表がなにかのヒント・一助になれば幸いである。

SASによって描くグラフ軸目盛の自動調整

魚田 孝志

(イーピーエス株式会社)

Automatic adjustment of axis scales on a graph drawn by SAS

Takashi Uota

(EPS Corporation)

SGPLOTやSGPANELプロシ ज्याを始めとしたSGプロシ ज्याが利用できるようになって以降、SASにより簡単なコードでさまざまな高品質なグラフを描くことが可能になった。しかし、最低限のコードで散布図や推移図を作成すると、軸の最小目盛を下回る値や軸の最大目盛を上回る値がプロットされるケースが多い。そのため、X軸、Y軸ともにすべてのプロットを最小目盛から最大目盛内に含めるためには、XAXISやYAXISステートメントを記述する必要がある。現状では出力されたグラフを目で確認してからXAXISやYAXISステートメントでの指定を行っているのがないか。しかし、データの更新等で、作成したプログラムを繰り返し実行する場合には、XAXISやYAXISステートメントの指定を見直さなければ、測定結果がグラフの表示エリアから外れてしまう問題が生じる。各プロットがグラフ内に表示されているのかを確認しなければならないが、作成するグラフが多ければ多いほど確認に要する負担が大きくなり、確認漏れによる不備が生じる可能性も高くなる。

そこで、すべてのプロットがグラフの表示エリア内に収まるように最小目盛、最大目盛、目盛の間隔を導く方法を検討することにした。まずグラフの軸設定の条件を以下に設定した。

- (i) いずれのプロットも目盛の最小から最大の範囲に含まれている。
- (ii) 最小目盛と次の目盛の間、最大目盛と手前の目盛の間にはプロットが存在する。（限られた領域の中で推移や測定値の分布の違いを最大限明確になるように表現するため）
- (iii) 目盛の数は少なすぎず、多すぎない。（視覚的な情報を最適化し、かつ、プロットの値を特定しやすくするため）
- (iv) 各目盛は目盛間隔の倍数で設定される。

そして、この条件を満たしたグラフの作成は以下の手順で行った。

- (i) グラフに表示する目盛の数の目安を決める。
- (ii) 同一の目盛設定が適切と考えるグラフごとにプロットする最小値と最大値を算出する。
- (iii) (i), (ii)をもとに目盛間隔を決定する。
- (iv) グラフ内にプロットする値がすべて含まれるように、目盛の最小値と最大値を調整する。
- (v) SGプロシ ज्याを用いて、(iv)を元にXAXISステートメント、及び、YAXISステートメントに記述してグラフを描く。

本発表にて、そのプログラムコードを紹介する。

SASによる個別データの可視化

○徳田芳稀、杉本真理

(エイツーヘルスケア株式会社 データサイエンス第3部)

近年データの可視化について書籍や論文で多く取り上げられており、更にR/Python等でも可視化について多数のパッケージが利用可能である。SAS GTL (Graph Template Language)においても約40種類のPlotステートメントが利用可能である。¹ 臨床試験における可視化について、例えば、並行群間比較試験で興味のある評価項目について各群もしくは集団全体でどのような傾向があるかを確認したい場合、推移図や散布図等を用いることで視覚的に確認できる。一方でデータモニタリング委員会や症例検討等においては、安全性確認等の観点から集団全体ではなく特定の被験者データのみを確認したい場面が考えられる。しかしながら、集団全体から特定のデータを判別することは手間がかかり、特に症例数が多数の場合には困難を極める。そこで本発表では、個別データの可視化に焦点を当て、SAS GTLによる個別データの図示方法について検討した。SAS GTLでは、LAYOUTステートメント内に複数の同一もしくは異なるPlotステートメントを指定しできる。推移図を例に挙げると、強調したい特定のデータとそうでないデータについて、LAYOUT OVERLAYステートメント内で別々のSERIESPLOTステートメントに線の種類や色等を指定することで、特定の個別データを強調して可視化できる。本検討にあたって、Rのgghighlightパッケージを参照した。²

【参考資料】

1. SAS Help: SAS Graph Template Language, Plot Statements

https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/grstatgraph/p14nrue3d9gbcns17oofi9t8li76.htm

2. CRAN: Package gghighlight

<https://cran.r-project.org/web/packages/gghighlight/index.html>

SAS SAMPLE LIBRARY で公開されているKaplan-Meierプロットの カスタマイズマクロ%ProvideSurvivalMacrosを知っていますか？

—Graph Template Language (GTL) への第一歩にも—

○太田 裕二、森田 祐介、浜田 泉

(ノーベルファーマ株式会社 データサイエンス部)

Kaplan-Meierプロットは、打ち切りを考慮したイベントが発生するまでの期間データを可視化するために用いられる。

LIFETESTプロシジャとODS Graphics機能を組み合わせることで簡単に作成できるが、情報量や見栄えが不十分な状況がある。そのため、LIFETESTプロシジャで算出した統計データを加工し、GPLOTやSGPLOTなどのプロシジャを用いてカスタマイズする方法が取られることも多い。

SAS SAMPLE LIBRARYで提供されているマクロ%ProvideSurvivalMacrosは、LIFETESTプロシジャが作成するKaplan-Meierプロットをカスタマイズするための強力なツールである。

本発表では、%ProvideSurvivalMacrosの使用方法を解説し、さらにGraph Template Languageを交えながらマクロの仕組みについても紹介する。

データステップ内で次のobsの値を参照する

○浅井 友紀

(エイツーヘルスケア株式会社)

SASにおけるデータ処理にはデータセットを1行ずつ読み込みその結果を出力するという特徴がある。そのためデータステップでは同じobsの別変数の値を参照することは容易である。また前のobsの変数の値はlag関数やretainステートメントなどを使用すれば参照できるが、次のobsの変数の値を直接参照できる関数やステートメントはない。

データステップ内で次のobsの値を参照できれば、例えば睡眠のようにデバイスから多数のデータが取得され、前後のデータの取得時刻の間隔によって解析に使用するデータを特定する、といった処理が簡単に記述できるのではないか。

そこで本発表では次のobsの値を取得するプログラムの一例として、setステートメントを使用する方法とhashオブジェクトを使用する方法を紹介する。

生存時間解析における3次スプラインを用いた事例の紹介

○室永 遼太郎、浦狩 保則

(シミック株式会社)

自明の記法を用いて、比例ハザードモデルを対数で表現すると

$$\log(h(t)) = \log(h_0(t)) + \beta X + \alpha Z$$

である。 X と Z はともに共変量であるが、本発表では X について議論する。

比例ハザードモデルの解析では、年齢や白血球数などの連続量の説明変数(X とする)は通常、それらの線形成分がモデルに含まれる。しかし、このような説明変数の効果を線形成分のみで解釈することが好ましくないことがある。この問題への一つの対処法として、 X の非線形関数 $g(X)$ を用いて以下のようにモデル化し解析する。

$$\log(h(t)) = \log(h_0(t)) + \beta g(X) + \alpha Z$$

また、比例ハザードモデルの解析における重要な仮定は、ハザードの比例性である。これはハザード関数におけるある共変量の効果が時間上で一定であることを意味する。しかし実際には、共変量の効果が時間上で変動する場合がある。このように比例ハザード性を満たさない共変量(X とする)が見出された場合、以下のように X の回帰係数を時間の関数とするアプローチが主流となっている。

$$\log(h(t)) = \log(h_0(t)) + f(t)X + \alpha Z$$

$f(t)$: 時間の関数で表した回帰係数

最も単純な方法は、 $f(t)$ を時間の線形関数とすることであるが、そのようなモデル当てはめの柔軟性を高めるために $f(t)$ として時間の非線形関数を用いる方法が推奨されている。

本発表では、上記の $g(X)$ と $f(t)$ のそれぞれに3次スプラインを用いた場合を取り上げ、実際のデータを用いてSASでの実装方法を紹介する。

視覚障害者を支えるSASプログラミング

-Accessible OutputとSAS Graphics Accelerator

森岡 裕

(イーピーエス株式会社)

SAS Programming techniques for the Visually Impaired

-Accessible Output and the SAS Graphics Accelerator-

Yutaka Morioka

(EPS Corporation)

SASには視覚に障がいを持った方でも、表やグラフを音声として認識しデータサイエンスを行うため機能が備えられている。Accessible OutputとSAS Graphics Acceleratorを主に解説する。

私たちの周りにある表やグラフは実は、そのビジュアル的な表示位置に意味付けがされている。表の縦横の系列との関連性は、視覚がないと認識しにくく、色でグループ分けされたグラフは当然、色覚と視覚の健全性を前提に作られている

Accessibility (Access + Ability = Accessibility) という概念があり、簡単に言うと、どのような人であっても、Webページや、なんらかのアウトプットに対して、アクセスできるように、利用できるようにデザイン・作成するということである

例えばSASのODSのStyleではDAISY (Digital Accessible Information System)スタイルというものがあり、これを利用してアウトプットやグラフを作成すると、弱視や色覚異常の人でも視認性が増すように設計されたものである。

またSASにはAccessible Outputオプションが多く用意されており、これらを使ってアウトプットを作成することで、視覚障害者用の音声読み上げソフトが出力結果を解釈しやすいく、正しく音声読み上げしやすくなる

またSAS Graphics Acceleratorという無償で利用可能なGoogle Chrome用のアドインがある。これを使うとSASのグラフ出力を音声に変換(例えば、軸における値の大小が音階で表され、高い音だと高い値、低い音だと低い値といった例)することができる。

作成サンプルや、グラフを耳で聴くことになれるためのサンプルも多く提供されており、視覚に依存せずに、イメージを持つことが可能となっている。

本発表では必ずしもSASにこだわらず、PythonやRにおける、視覚障害者向けのパッケージについても紹介したい。

参考

[1] SAS® Graphics Accelerator

<https://support.sas.com/software/products/graphics-accelerator/index.html> (Accessed Aug 19, 2024)

MCMC Procedureによる混合モデルの実装とクラスタリング：

1次元混合正規モデルから多変量時系列モデルまで

○福島綾介

(イーピーエス株式会社)

本発表では臨床試験データを探索的に解析し、反応特性が異なる複数の被験者集団に分類可能な手法を紹介する。臨床試験では各被験者に対して複数の評価項目を設定し、医薬品等による介入後の変化を複数時点で観測することで多変量時系列データを得る。このような多変量時系列データに対してクラスタリングを行い、被験者集団全体をいくつかの集団に分類することを目指す。例えば、医薬品等の効果が強く現れやすい被験者集団と効果が現れにくい被験者集団に分類できる場合がある。このような分類ごとの試験開始前ベースライン値に違いがあるとき、医薬品等の効果を予測可能な被験者特徴が得られる可能性がある。このクラスタリングのために混合モデルと、時系列表現のために状態空間モデルを導入し、ベイズ統計に基づいた推定を行う。

SASではベイズ統計に基づいた解析のためにMCMC Procedureが利用可能である。MCMC Procedureはマルコフ連鎖モンテカルロ法 (Markov Chain Monte Carlo; MCMC) を実装可能であり、パラメータの事後分布を得られる。本発表では1次元混合正規モデルを例として混合モデルを実装するときに役立つ以下の4つのコツを紹介する。(A) No-U-Turn Sampler (NUTS)の利用。(B) LogSumExpの計算テクニック。(C) 提案分布にSlice Samplerの利用。(D) Label Switchingへの対処。さらに、時系列表現のために状態空間モデルを導入する。状態空間モデルは測定値として直接には観測できない潜在変数を取り扱うことが可能であり、測定値に欠測が含まれていても欠測値の補完なしで適用できる。また、順序選択モデルの適用によって離散変数の時系列にも適用できる。最後に、多変量時系列データへの応用例として、連続変数と離散変数の2変数が時系列として得られたときのクラスタリング例を示す。

本発表では状態空間モデルの混合モデルを用いてクラスタリングを行う。混合モデルの利用により被験者集団内における反応の不均一性を表現でき、反応の違いと関連がある集団特徴を理解できる可能性がある。この知見は医薬品等の効果予測を可能にするだけでなく、新たな医薬品の開発において新薬のターゲット集団の特定に貢献できる可能性もある。本発表で提供する方法が医薬品開発の推進のために、探索的データ解析の一つとしてSASで実行可能であることを紹介できれば幸いである。

AI-SAS for RWEの開発

○米田 卓司¹, 惟高 裕一¹, 木口 亮¹, 北西 由武¹

(¹塩野義製薬株式会社)

本文

塩野義製薬が開発したAI-SAS（略称. 正式名称：人工知能解析プログラマシステム, 通称：AI SAS Programmer）は、過去の臨床試験解析業務で蓄積された図表計画書, 解析用データセット等を学習データとしてモデルを組み上げ, 対象とする臨床試験の図表計画書を基に, 解析用プログラムを準自動で作成するシステムである。AI-SASにより塩野義製薬の臨床試験解析業務で約100時間（30%）の削減が確認できている。

一方で, 臨床試験以外のリアルワールドデータ（RWD）を用いた解析からのリアルワールドエビデンス（RWE）の創出が業界で活発化していることから, RWD解析の効率化も取り組むべき課題である。また, 臨床試験と異なりRWD研究では第三者による監視がほとんどないことから, 研究者にとって望ましい結果を得るために研究方法を調整した可能性を否定できず, プロセスの透明性が問題視されることがある¹⁾。そのため, 解析プロセスの透明性の確保による研究の質の担保も課題の一つである。これらの課題から, 透明性のあるフローでRWD解析を効率化することが, RWEの創出加速と質向上に必要であると考えている。

そこで, 我々はAI-SASをRWDに応用したAI-SAS for RWEの開発に取り組んでいる。データ形式・解析業務内容・解決課題範囲の違いから, 従来のAI-SASをそのままRWDに適用することではAI-SAS for RWEは実現しない。まず, 各RWDはローデータ形式が統一されていない。加えて, 膨大なレコードから傷病/薬剤コードと抽出期間を指定して各変数を作成する必要がある。解析用データセット作成にもリソースが大きく割かれる。またプロジェクトごとに, 各解析資料の作成形式や有無が異なっている。従って, 効率的に透明性を確保する観点から, 各解析資料の準自動作成と, 解析資料・プログラム・帳票の作成履歴を自動記録できる環境が必要となる。本発表では, これらの課題も踏まえたAI-SAS for RWE開発への取り組みについて発表する。

- 1) Dagenais S, Russo L, Madsen A, Webster J, Becnel L. Use of Real-World Evidence to Drive Drug Development Strategy and Inform Clinical Trial Design. Clin Pharmacol Ther. 2022 Jan;111(1):77-89. doi: 10.1002/cpt.2480. Epub 2021 Nov 28. PMID: 34839524; PMCID: PMC9299990.

Rプログラムを生成系AIを用いてSASに変換してみた事例について

○三木悠吾、森口泰孝

(グラクソ・スミスクライン株式会社)

近年、ADaMやTFL作成のためにSAS以外の他プログラミング言語、例えばRを用いて承認申請を行う事例が増えている。しかしながら、FDAやPMDA、NMPAなどeSubmissionが要求される規制当局では、SASのプログラムが要求される可能性がある。特にNMPAでは基本的にExecutableなプログラムが要求されるため、SAS以外の他プログラミング言語、特にオープンソースのプログラミング言語の提出では結果を再現するのが困難であることが想定される。そのため、規制当局からの要請に応じて、フレキシブルかつ迅速にSASプログラムを追加で準備することが求められる。R等の非SASプログラムを使用していない製薬会社であったとしても、導入品などにおいて他社で作成されたプログラムを受領する可能性はある。したがって、R等の非SASプログラムを使用していない製薬会社にとっても、R等のプログラムをSASに変換するニーズは非常に高いことが推察される。今回、生成系AIを用いて、RのプログラムをSASにうまく変換できるか実験を行ったため、その結果を紹介する。

SAS Extension for Visual Studio Codeの紹介

○山添 滉弥

(エイツーヘルスケア株式会社)

本発表ではBase SASに付属するコードエディタに代わるものとしてVisual Studio Code (VSCode)とそのSAS拡張機能について紹介する。Base SASにはコードエディタが付属しているが、エディタとしての機能は限られている。代替のサードパーティ製のエディタは設定次第でコードの記述は快適になる。しかし、SASに合わせてエディタの設定を行う敷居が高く、結局はSAS上で実行する必要がある。VSCodeおよびそのSAS拡張機能についてコードの記述および実行の面から紹介し、上記課題がある程度解決できることを共有する。

打ち切りをもつ連続値評価項目の解析

～意外と知られていない解析法たち～

○馬場 崇充, 渡辺 秀章, 長谷川 貴大

(塩野義製薬株式会社 解析センター)

データの測定法に起因し、打ち切りをもつ連続値評価項目は多くの状況で生じ得る。例えば、感染症領域でウイルス量を定量化する際には定量限界値 (LLoQ: lower limits of quantification) が存在し、定量限界未満では定量値が得られないという打ち切りが発生する。同様のことは、臨床検査値やPKデータでも発生し得る。そのような打ち切りをもつデータは制限従属変数と呼ばれるクラスに含まれる。感染症領域の試験でウイルス量が経時反復測定される場合、 $<LLoQ$ となる打ち切りデータに対して、 $LLoQ$ または $LLoQ/2$ の値を代入するという単一代入法を適用した上で、Mixed Effect Model for repeated measures (MMRM) または時点別の Analysis of Covariance (ANCOVA) によって因果効果 (群間差) を推定する方法がよく用いられてきた。しかし、単一代入法がそもそも適切であるかという点、また打ち切りを多く含むことで正規分布からの乖離が大きくなり、ANCOVAやMMRMにより推定される群間差が過大または過小評価される点から、より適切な方法を用いることが好ましい。

Moser et al. (2023) は最近のCOVID-19の治療薬の開発におけるウイルス量の解析における先述の課題を取り上げ、ウイルス量データに対する解析法として、単一代入法+ANCOVAまたはMMRMに加えて、パラメトリックな方法であるTobitモデルやノンパラメトリックな方法である打ち切りデータを伴う分位点回帰モデルの特徴を評価している。そして、結論としてTobitモデルまたは打ち切りデータを伴う分位点回帰モデルの利用を推奨している。

本発表では推奨される各手法の理論的背景を概説するとともに、SASで実装し、数値実験により各方法論の特徴を比較する。TobitモデルはPROC QLIM (SAS/ETS) またはPROC LIFEREG (SAS/STAT) で、打ち切りデータを伴う分位点回帰モデルはPROC QUANTLIFE (SAS/STAT) で実装が可能である。それぞれのプロシージャによる実装法を説明し、数値実験により見えてきた各方法の結果の差異と適用場面について議論する。

参考文献

1. Moser CB, Chew KW, Giganti MJ, et al. Statistical challenges when analyzing SARS-CoV-2 RNA measurements below the assay limit of quantification in COVID-19 clinical trials. *The Journal of Infectious Diseases* 2023; 228 (Supplement 2): S101-S110.

SASハッシュオブジェクト基礎編

○松永 友貴¹

(¹ノバルティス ファーマ株式会社)

ハッシュオブジェクトとは、キーと値の組み合わせでデータを表現するオブジェクトであり、迅速なデータの保存および取得が効率よく行えるメカニズムである。臨床試験データのような、構造化された大規模なデータとの相性が良く、処理時間の短縮や柔軟なコードの実装が可能である。SASにおいては、SAS9.1 - Base SAS Softwareより実装されている。

SASユーザー総会においては、2014年より森岡氏 (イーピーエス株式会社) によりSASハッシュオブジェクトに関する講演が行われている。高い有用性と専門性が話題を呼び、SASユーザー総会の名物講演と知られ、毎年盛況な演題となっている。一方で近年、講演内容の専門性の高さから、前提知識がない初心者にはハードルが高くなり始めている。本講演では、2014年の森岡氏の発表をベースとし、再度SASハッシュオブジェクトの基礎を振り返ることで、初心者の理解促進を目指す。

本公演では、SASハッシュオブジェクトの理解促進のために、ハッシュオブジェクト、MERGEステートメント、SQLプロシジャとの文法の比較について解説する。また、ハッシュオブジェクトを利用する最大のメリットである、プログラム実行時間の短縮について紹介する。

実際の使用例として、有害事象データと併用薬データおよび、臨床検査値と施設基準値を効率よくマッチングする方法を紹介する。また、解析業務に使用されたSASのプログラムコードをテキストとして読み込み、Perl正規表現とハッシュ反復子オブジェクトを使い、どのような処理が行われたかをレビューする方法を紹介し、テキスト処理におけるハッシュオブジェクトの特徴と利用メリットを明らかにする。

解析結果RTFファイルの検証手法の検討と報告

○山野辺 浩己

(マルホ株式会社臨床開発部)

臨床試験データの解析結果は、総括報告書作成等の理由からRTFファイルで作成されることが多い。

RTFファイルで作成された解析結果の検証は今まで様々な手法が試みられている。RTFタグセットを読みほどこきSASデータ化するなどといった手法から、一度別ファイルに変換してから読み込むなど多様な先行事例がある。発表者の解析環境による制約からPython等を使った一部の有力な手法は社内セキュリティ環境により選択出来ず、また社内検証担当者の既存の進め方に適合する検証方法を検討する必要がある。

これらの前提を踏まえ、RTFファイル同士の単純比較の効率化や、RTFファイルのSASを使ったデータセット化やエクセルを使った検証法等を検討し、以下の手法を実践した。

1. 比較用解析結果をRTFファイルで作成し、Microsoft Wordにおける比較機能を用いた方法
2. 比較用解析結果をSASデータセットで作成し、比較対象RTFファイルを出力する直前のSASデータセットとSASにおけるCOMPAREプロシジャを用いた方法
3. 比較用解析結果をエクセルファイルで作成し、比較対象RTFファイルをエクセルに変換した上で、エクセル関数を用いた方法

これらを実践したところ各手法に一長一短があり、それぞれに今後の検討すべき課題を得た。特別に秀でた手法はなく、準備、ユーザーとの親和性、調整コスト等を見極め、最良の方法を選ぶ必要がある。

SASで作成した画像データの外部への出力とTIPS

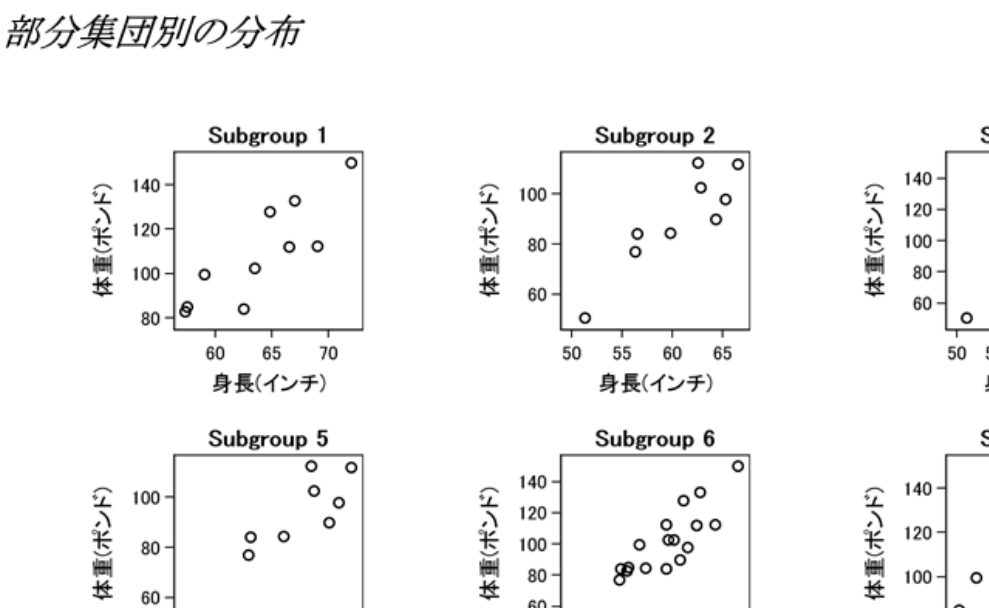
○山野辺 浩己

(マルホ株式会社臨床開発部)

近年SASユーザー総会ではグラフに関する発表が多く、データをただ解析するだけでなく、その可視化に関心が高まっている。グラフ作成にあたっての手法は様々な場で紹介されているが、本発表では作図手法ではなく、作図した画像データを出力するにあたってのTIPSをまとめた。必要に迫られ実践したものとして「複数のグラフをA4に並べる場合」、「パラメーターを変えて連続して主力する場合」、「密度が高くテキストが潰れてしまう場合」、また「複数ファイルのグラフを1点にまとめる場合」等を用いた手法とTIPSを紹介する。

【A4に複数の図を載せ、図ごとにTitleを値依存で出力する例】

部分集団別の分布



文字はなぜ切れる？変数長の決まり方

田中 真史

(イーピーエス株式会社)

SASの文字変数は固定長であるため、変数長がコードのどこでどのように決定されるかは、基本的な事柄であり、重要である。一方で、文字変数の挙動は数値変数に比べて非直感的であるため、数値変数の処理とは別物と捉えて勉強する必要がある。例えば

```
data z;  
  x = 'aaa';  
  y = x || 'bbb';  
run;
```

を実行すると、yは「aaabbb」になり文字切れしないが、他方

```
data z;  
  length x $20;  
  x = 'aaa';  
  x = x || 'bbb';  
run;
```

を実行すると、xは「aaa」になり文字切れする。本発表では、DATAステップにおける文字変数の変数長の挙動について、初学者にとって分かりにくいと思われる箇所を説明する。次の3点が要点である。変数長は

1. DATAステップで最初に変数が定義された場所で決まる。
2. DATAステップのシンタックスと読み込むデータセットのディスクリプタ部で決まる。
3. 読み込むデータセットの具体的なデータには依存しない。

SASが実用的なシステムであるならば、どれも当然のことであろう。変数長の挙動を理解すると、データの文字切れを起こしにくくなり、文字切れした場合も、その原因が特定できる。また、自信を持って不要なLENGTHステートメントを省くことができ、コードの可読性を上げられる。本発表で強調したいのは、SASの文字変数の挙動は理屈で予測する事柄であり、「慣れ」や「勘」でどうにかする事柄ではないことである。

傾向スコアマッチング前後の群間バランスの 定量的な評価指標に対する考察

○上野 航太郎, 魚住 龍史

東京工業大学工学院経営工学系

近年の観察研究では傾向スコアマッチングが用いられることが多く、マッチング前後の患者背景因子に対する群間バランスを評価する必要がある。群間バランスの評価として、どのような方法が挙げられるだろうか。視覚的な方法としては、群間の傾向スコアのミラーヒストグラムなど、多くの評価方法が挙げられる。また、PSMATCHプロシジャではクラウドプロットなども出力可能である。その一方、群間バランスの定量的な評価方法として、標準化平均差 (standardized mean difference, SMD) がよく使われる。PSMATCHプロシジャのASSESSステートメントでは、ODS GRAPHICSの機能により、指定した変数のSMDを出力できる。しかし、観察研究によっては、SMDでなくp値で群間バランスの評価を実施している事例も散見される。本発表では、2群の観察研究を想定し、モンテカルロシミュレーションによって、群間バランスの定量的な評価指標としてSMDとp値をそれぞれ用いる場合の性能比較を行った結果を報告する。

【参考文献】

魚住龍史, 矢田真城, 山本倫生, 川口淳. SASによる傾向スコアマッチング. SASユーザー総会論文集. 2017;263-281

魚住龍史, 矢田真城. SASによる傾向スコアの活用 : COVID-19患者に対する観察研究をひも解く. 日本SASユーザー会フォーラム. 2020

Geleris J, Sun Y, Platt J, et al. Observational Study of Hydroxychloroquine in Hospitalized Patients with Covid 19. New England Journal of Medicine. 2020;382 :2411 2418

Gum PA, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS. Aspirin Use and All-Cause Mortality Among Patients Being Evaluated for Known or Suspected Coronary Artery Disease: A Propensity Analysis. JAMA. 2001;286:1187-1194

データサイエンスを支えるデータエンジニアリングのチカラ

—SAS Viyaを中核とするSHIONOGI統合解析環境—

○坂井 聡、北西 由武

(塩野義製薬株式会社 DX推進本部 データサイエンス部)

塩野義製薬データサイエンス部は、RWD (Real World Data) 解析をはじめ、新規事業創出、社内DX支援など多岐にわたる業務を担っている。組織構成としても、データサイエンティストのみならずデータエンジニアも多数在籍している。

社会に目を移すと、ヘルスケアやビジネスのニーズは高度化・多様化の一途を辿っている。我々は、そのニーズにいち早く対応可能な解析環境を手に入れることが必要であった。そのような背景から、2023年度下期にSAS Viyaを中核とし、統合解析環境を刷新した。刷新前までの統合解析環境では、現在までの各種ニーズには対応可能だが、来るべき将来ニーズに対しても準備万全と自信を持って言える状況ではなかった。今回刷新した新しい統合解析環境のキーワードは、「スピード」、「柔軟性」、「コラボレーション」である。データの種類、量が爆発的に増加する昨今、社内外データをいかに上手くハンドリングし、スムーズな解析を後押しし、成果創出に繋げるかは、データエンジニアリングの一丁目一番地である。

本発表では、弊社が構築した統合解析環境がもたらした成果と期待感を事例交え紹介する。

2値型応答における意外と知られていない因果効果の推定方法

～なんとなくオッズ比からの脱却～

○堀江 悠生、馬場 崇充、渡辺 秀章、長谷川 貴大

(塩野義製薬株式会社 解析センター)

薬効評価のための無作為化比較試験で有効性の評価項目に2値型応答を用いる場合、共変量調整した有効性の群間差として回帰モデルに基づく調整済みリスク差、リスク比、及びオッズ比が用いられる。しかし、オッズ比はイベントが低頻度である場合のみリスク比の近似として解釈可能である。一方、より解釈が容易なリスク差とリスク比を用いることが推奨されている (Harrington *et al.*, 2019)。また、PROC GENMOD等の通常のプロシジャでは、用いるリンク関数に応じて、最尤法による回帰係数に基づく有効性の評価指標としてリスク差 (恒等リンク関数)、リスク比 (logリンク関数)、及びオッズ比 (ロジットリンク関数) が得られるが、それは共変量の平均値を条件付けたリスク差、リスク比、及びオッズ比として与えられる。恒等リンク関数が用いられることで、共変量の平均値を条件付けたリスク差は共変量調整した条件付きでないリスク差として解釈可能である。一方、恒等リンク関数が用いられない下で、共変量の平均値を条件付けたリスク比及びオッズ比は共変量調整した条件付きでない因果効果として解釈できない。共変量の条件付きの推定では、リスク差、リスク比、及びオッズ比の算出に必要な各群のリスクやオッズも出力されないことに注意したい。

一方で、リスク差とリスク比について、共変量に条件付きでない因果効果を推定するには、回帰標準化 (Hernán and Robins, 2020) と呼ばれる方法を用いることができる。回帰標準化に基づく共変量調整した条件付きでない推定では、各群のリスク、リスク差とリスク比、これらのデルタ法に基づく漸近分散を求めることができるため、より解釈しやすい形式で結果報告することが可能になるという利点も有する。SASでは、PROC CAUSALTRTプロシジャを用いることで共変量調整した条件付きでない各群のリスクとリスク差を算出できるが、当該プロシジャはリスク差にしか対応していないという課題がある。本発表では、PROC LOGISTICプロシジャを用いて回帰標準化に基づく共変量に条件付きでないリスク差、及びリスク比と、デルタ法に基づくそれぞれの漸近分散及びロバスト分散を算出するSASコードを開発し、実データに適応した結果を紹介する。

参考文献

Harrington D, D'Agostino RB Sr, Gatsonis C, et al. New guidelines for statistical reporting in the *journal*. *The New England journal of medicine*. 2019; 381(3):285-286.

Hernán MA, Robins JM. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2020.

SAS 新製品概要：LLM ガバナンス, SAS Viya Workbench, Copilot, そして SAS Data Maker

○土生 敏明¹、William Kuan¹

(¹SAS Institute Japan)

SAS は 2024 年に入り 3 つのソリューションを発表しました。

1. SAS Viya Workbench

SAS&OSS ユーザー向けに、クラウドベースの軽量なオンデマンドコンピューティング環境を提供するソリューションです。ユーザーが必要な時に必要なリソースを利用でき、効率的なデータ処理と分析を可能にします。

2. SAS Viya Copilot

SAS コードの自動生成ならびにプログラム作成支援を行うソリューションであり、利用することで SAS プログラム開発を大幅に簡略化できます。こちらを用いることで SAS プログラム初心者であっても、上級者と同様のプログラム作成が出来ます。また過去のプログラム資産に対し、本 Copilot 機能を用いることで、処理内容の可視化ができ、業務継続性の向上を図れます。

3. SAS Data Maker

参考になるデータをアップロードし、数クリックするだけで簡便に合成データの生成を行うソリューションです。本ソリューションを用いることで、データのプライバシーを保護しつつ、リアルなデータセットを作成することができます。

これらの新製品を今後リリース予定の機能も含め、デモを交えて紹介します。

疫学研究で2群を比較する際に用いる確認手法について

○矢田 徹

(イーピーエス株式会社)

要旨：

無作為割付を行わない疫学研究で群間比較する際に傾向スコアを用いたマッチングや重みを考慮した解析等を行うが、その際に用いるSASプロシジャや出力結果を紹介する。

キーワード：Epidemiology, Propensity Score, IPTW, Matching, Model fitting information

発表概略：

無作為割付を行わない疫学研究では、曝露の有無の影響を比較するために曝露群と非曝露群を設定しても、それらが均質な集団であることを保証できない。そのため、性別・年齢等の曝露やアウトカムに影響しうる変数から傾向スコアを算出してマッチングを行ったり、傾向スコアを用いた重みで調整したりして、できるだけ均質な集団になるように考慮する。SASには、傾向スコアを算出するプロシジャやそれを重みとして調整するプロシジャがあり、様々な種類のオプションが存在する。簡単なモデルを設定して構築したデータを用いて、群間の偏りを調整するプログラム例を紹介する。

LOGISTICプロシジャ、PSMATCHプロシジャを用いる。

【ページ設定】

用紙： A4・縦（1ページ）
余白： 上/下/左/右 25mm/25mm/20mm/20mm
行送り： 36行、18pt送り
文字間隔： 標準

■フォント

日本語： MS明朝
English： Times New Roman
数字： Times New Roman (0123456789)

■フォントサイズ・配置

演題名： 16ポイント・中央揃え
演者名： 11ポイント・中央揃え
※発表する演者の氏名の前に○をつける
※演者の所属が複数になる場合、各演者の氏名の後に所属を表す上付き数字を記載
演者所属名： 11ポイント・中央揃え ※全体をカッコで囲う
抄録本文： 10ポイント、両端揃え

■提出ファイル形式

- ・ Microsoft Office Word 2013 / 2016 / 2019 / 2021
- ・ PDF

提出する際は、こちらのテンプレートを使用して作成した **Word データ** と、保存形式を **PDF にして保存したファイル両方 (2 ファイル)** をご提出ください。

抄録は 1 ページ以内で作成してください。

ご不明点等ございましたら、ユーザー総会事務局までお問合せください。

【お問合せ先】

SASユーザー総会事務局

E-mail: info@sas-user2024.ywstat.jp

Kappa係数の意外な落とし穴とその対処法の紹介

○江口 凌平、菊池 宏和

(シミック株式会社)

背景：

カテゴリカルデータの一致度を評価する方法の一つにカッパ係数(κ 係数)がある。例えば、二人の観察者間の評価といった質的データ(カテゴリカルデータ)の一致度を、偶然による一致度を考慮した上で算出することが可能である。しかし、一致セル・不一致セルの割合に偏りが存在する場合には、一致率が同じでも κ 係数が異なる場合があり、 κ 係数の解釈が困難になる可能性がある。

検証：

SASのシミュレーションにより、 2×2 クロス表データの値の偏りを変動させた場合の κ 係数の動作特性を報告する。また、偶然による一致度を調整し、値の偏りに頑健な指標であるPABAK、AC1統計量を紹介し、 κ 係数同様に 2×2 クロス表データの値の偏りを変動させた場合の動作特性を報告する。

結果・まとめ：

κ 係数のシミュレーションの結果、一致セル・不一致セルの割合の不均衡が大きい場合、 κ 係数はその影響を受け、値が変動しやすい傾向にあることが分かった。したがって、 κ 係数の値を解釈する際には、一致セル・不一致セルの割合の偏り具合を確認する必要がある。また、PABAK、AC1統計量のシミュレーションの結果、一致セル・不一致セルの割合の不均衡が大きい場合でも、 κ 係数の場合と比較して値が変動しにくい傾向にあることが分かった。したがって、 κ 係数だけではなく、より偏りに頑健な統計指標であるPABAK、AC1統計量も併せて評価する必要があると考えられる。

大規模言語モデルから動かすSAS

○中松 建

(個人)

大規模言語モデル(LLM)をはじめとするAIの活用が多く組織で課題となっている中で、SASに関する有用な利用例や資料はまだ非常に少ない。LLMを活用するうえでは、正確性の向上のためにTool Callingと呼ばれる外部ツールを利用する言語モデルの機能と、機能を拡張できるライブラリをあわせて用いることが多く、ライブラリの代表的なもののひとつにlangchainがある。現在も非常に活発に更新されているライブラリであり、特に昨年時点では安定性がまだ不十分であったが、各種サービスやツールでもサポートされるようになり、ある程度安定性も高まってきていると考えられたことから、langchainをベースとしてLLMからのSASの利用について検討することとした。

本報告では、langchainおよび関連ライブラリのlanggraphによる、LLMからSASをツールとして動かすエージェントの実装例を通して、ライブラリを問わずLLMとSASの間のやり取りで必要となるような、セッション管理、オプション・設定、データ・メタデータの取得、出力形式の制御、出力結果へのアクセスの制限など、特にLLMからSASやデータを扱うのに関連するトピックや、そこでのsaspyの利用などについて報告を行う。

公開リポジトリ

https://github.com/k-nkmt/SAS_as_Tool_Examples

共分散分析の基礎

○藤島直隆

(株式会社分析屋)

臨床試験の統計解析において、共分散分析はベースライン時の測定値を共変量としてよく行われる解析手法である。

私は共分散分析を初めて学習した際に、基本的な統計学（ベイズの定理、単回帰分析、t検定など）に加えて交絡という概念も理解していないと、共分散分析を理解することが難しいと感じた。

しかし、`proc glm`や`proc mixed`を用いて共分散分析を行った際に、数行コードを書くことにより共分散分析を行うことができたため、実行自体は簡単だと感じた。

このことから共分散分析を理解できていないまま、SASのコードを実行してしまうと業務で自信を持って解析業務を行うことができないと感じたため、様々な資料を読み共分散分析の目的や利点などをSASで共分散分析を実行しながら理解することに努めた。その結果、自信を持ってSASによる実装を行えるようになった。

自分自身の経験を踏まえて、今回は共分散分析の基礎をテーマに業務で共分散分析を行うために最低限必要なことを理解することを目的とし、共分散分析の理論はもちろん、SASによる実装例と分析結果の解釈、さらにt検定や分散分析との違いも説明する。

キーワード：交絡、共変量、共分散分析、t検定、分散分析、`proc glm`、`proc mixed`

SGPLOTで広がるKaplan-Meierプロットの世界

○魚住 龍史

東京工業大学工学院経営工学系

Email: uozumi.r.ab@m.titech.ac.jp

SGPLOTプロシジャが利用可能となってから、SASによるグラフィックス機能は大幅に改善された。魚住・浜田 (2011) の報告では、Kaplan-Meierプロットを題材にした臨床開発担当者からのリクエスト対応として、SGPLOTプロシジャで対応可能な事例が紹介された。当時、SGPLOTプロシジャは新たに導入されたばかりであり、初期機能として限られたステートメントのみであった。Kaplan-Meierプロットにリスク集合を追加する方法を考えると、当時のSGPLOTプロシジャではリスク集合が横軸の上部に出力されてしまい、リスク集合を横軸の下部に出力するためには、代替手段としてテンプレートを用いてSGRENDERプロシジャを使う必要があった。しかし、SGPLOTプロシジャのXAXISTABLEステートメントが追加されたことで、リスク集合を横軸の下部に出力させることが可能となった (魚住ら, 2016)。LIFETESTプロシジャによる出力としても、OUTSIDEオプションを指定することで、リスク集合を横軸の下部に出力させることが可能である (大橋ら, 2016, 2022)。その一方、SGPLOTプロシジャを用いる場合、グラフに細かい指定を加えて出力することができる点が魅力である。川原田 (2019) の報告では、BANDステートメントで生存割合の95%信頼区間の付加など、さまざまな情報の追加方法を示している。堀ら (2021) の報告では、SGPLOTプロシジャのSTYLEATTRSステートメントの活用事例が示されている。

以上のように、これまでのSGPLOTプロシジャの機能拡張により、Kaplan-Meierプロットは高品質なグラフとして出力できるようになった。演者は10年以上SGPLOTプロシジャを用いて、Kaplan-Meierプロットを作成し続けている。しかし、演者が作成したKaplan-Meierプロットに対して、未だに細かいリクエストを受けることがある現状である。本講演では、Kaplan-Meierプロット作成に関連したSGPLOTプロシジャの機能をチュートリアル形式で解説し、演者が経験したKaplan-Meierプロットに対する細かいリクエストについて議論する。

参考文献

- [1] 魚住龍史・吉田早織・平井隆幸, 浜田知久馬. Kaplan-Meierプロット・Forestプロット作成の応用: グラフ出力範囲内・範囲外への数値出力. SAS ユーザー総会 論文集 2016, 216-232.
- [2] 魚住龍史・浜田知久馬. SG (Statistical Graphics) Procedures による Kaplan-Meierプロットの作成. SAS ユーザー総会 論文集 2011, 185-199.
- [3] 大橋靖雄・浜田知久馬・魚住龍史. 生存時間解析 応用編 -SASによる生物統計. 東京大学出版会, 2016.
- [4] 大橋靖雄・浜田知久馬・魚住龍史. 生存時間解析 第2版 -SASによる生物統計. 東京大学出版会, 2022.
- [5] 川原田貴広. 様々な付加情報を加えた Kaplan Meier Plot の作成事例. SAS ユーザー総会 論文集 2019, 102.
- [6] 浜田知久馬・魚住龍史. 生存時間解析 応用編 -SASによる生物統計. 東京大学出版会, 2016.
- [7] 堀隆正・日高優・魚住龍史. カプランマイヤープロット描画におけるスタイル要素の変更マクロ. SAS ユーザー総会 2020. Available at: https://www.sas.com/ja_jp/events/21/users-group-2021/sessions.html [2024/08/05]

主催：SAS ユーザー会 世話人会

代表世話人	伊藤 陽一	北海道大学
世話人(50音順)	井上 貴博	ノバルティス ファーマ株式会社
	上村 鋼平	東京大学
	岸本 淳司	九州大学 ARO 次世代医療センター
	菅波 秀規	興和株式会社
	竹村 尚大	SAS Institute Japan 株式会社
	林 行和	エイターヘルスケア株式会社
	森岡 裕	イーピーエス株式会社

SAS ユーザー総会 2024 についての問い合わせ先：info@sas-user2024.ywstat.jp

協賛(50音順)

イーピーエス株式会社
エイターヘルスケア株式会社
株式会社エス・シー・エー
ClinChoice 株式会社
SAS Institute Japan 株式会社
塩野義製薬株式会社
株式会社新日本科学 PPD
スタットコム株式会社
株式会社タクミインフォメーションテクノロジー
ノバルティスファーマ株式会社
株式会社日立製作所