# SAS ユーザー総会 2017

# 論文集

SAS、SASを構成するプロダクト群は、SAS Institute Inc.の登録商標です。
その他、本論文集に記載されている会社名、製品名は、一般にそれぞれ各社の商標または登録商標です。
本論文集の一部または全部を無断転載することは、著作権法上の例外を除き、禁止されています。
本論文集の内容を実際に運用した結果の影響については、責任を負いかねます。

# 目次

## 基調講演

**医療における改正個人情報保護法と次世代医療基盤法** ―― 3
山本 隆一（一般財団法人医療情報システム開発センター）

## CDISC企画セッション

**ある実務担当者が考えるADaMダイバーシティと未来予測** ―― 7
渡辺 寿恵（小野薬品工業株式会社）

**PMDA申請電子データ提出経験談** ―― 8
栗栖 章（MSD株式会社）
鈴木 正人、福田 裕章

**多重性の調整が必要な試験を実施した際の承認申請時電子データ提出方法の検討** ―― 9
大内 喜海（協和発酵キリン株式会社、CJUG-ADaM）
福田 裕章（MSD株式会社、CJUG-ADaM）、大野 浩太（北海道大学病院 臨床研究開発センター、CJUG-ADaM）、
浅見 由美子（第一三共株式会社、CJUG-ADaM）

**ある実務担当者が考えるADaM開発アプローチと標準の紹介** ―― 10
坂上 拓（株式会社中外臨床研究センター）

**USでの申請承認プロセスとCDISC対応の事例** ―― 11
伊藤 雅憲（アステラス製薬株式会社）

**医薬品承認申請時電子データ提出の現状について** ―― 12
安藤 友紀（独立行政法人 医薬品医療機器総合機構）

## オープンデータ

**第5回Let's データ分析コンテストの規定課題のSASプログラム** ―― 15
周防 節雄（兵庫県立大学）

**Let's データ分析コンテストに用いる新擬似ミクロデータの概要** ―― 33
高橋 行雄（BioStat研究所株式会社）
周防 節雄（兵庫県立大学）、宮内 亨（独立行政法人統計センター）

## 金融・経済・経営関連

**自治体におけるオープンデータ化への対応の現状と課題** ―― 47
有馬 昌宏（兵庫県立大学）

**ID-POSだけで始める新しいRFM分析：データステップで実装するモンテカルロ法を用いて** ―― 53
片桐 智志（ネイチャーインサイト株式会社）

**最新AIの話題 ～AIの権利は誰のもの？～** ―― 63
小野 潔（株式会社インテック）

**One-Pass AUC Optimization手法の実務的拡張** ―― 72
木村 和央（株式会社金融工学研究所）

産業連関表を利用した地域経済の域外輸出の活性化要因の探索 ―― 83
栃澤 欣之(大塚製薬株式会社)
橋本 綾子、角元 慶二、菱垣 晴次

## SAS システム

プログラムバグを見つけるための注目すべきSASログメッセージ ―― 95
吉野 祥(株式会社クレスコ)

SAS Drug Development (SDD)実務の実際
ーデータマネジメント業務と解析業務はどう変わったかー ―― 107
松山 友映/高沼 正幸(協和発酵キリン株式会社)
兵頭 洋平、長谷川 優子、本山 佳代子、山下 昌洋

Jupyter Notebookを活用したプログラムライブラリ構築の検討 ―― 108
吹谷 芳博(あすか製薬株式会社)
藤澤 正樹

SAS/EGプロジェクト実行時のエラーハンドリングについて
ー任意の処理単位で分岐処理を実行する方法ー ―― 118
岩本 光司(武田PRA開発センター株式会社)

医薬品開発におけるSASプログラム
および関連ファイルのバージョン管理と共有について ―― 119
川上 貴弘(電助システムズ株式会社)

xmlv2エンジンを用いたannotated CRFの作成と検証方法の検討 ―― 123
三澤 喬(日本たばこ産業株式会社)

ODS Graphics Designerを用いたグラフの作成:SAS 9.4の機能を活用 ―― 130
平井 隆幸(日本化薬株式会社)
魚住 龍史(京都大学)

Report Writing Interfaceによる
データステップでのHTML形式レポート作成の検討 ―― 149
望戸 遼(イーピーエス株式会社)

SASと外部ソフトウェアの連携:Windows GUI構築とExcelデータの利用 ―― 161
高田 一行(株式会社タクミインフォメーションテクノロジー)
迫田 英之(株式会社タクミインフォメーションテクノロジー)、阿部 一也(株式会社タクミインフォメーションテクノロジー)、
村松 徳之(株式会社タクミインフォメーションテクノロジー)、宮原 秀夫(株式会社タクミインフォメーションテクノロジー)、
木村 範昭(専修大学)

初/中級者が陥りやすいプログラミングエラー ―― 162
井上 貴博(ノバルティス ファーマ株式会社)

Let's start STDRATE Procedure ―― 175
堀田 真一(ファイザー株式会社)

Pinnacle 21 CommunityのADaMチェック機能を補完する、
XML Mappingを使用したプログラムの紹介 ―― 176
西岡 宏(シミック株式会社)
中尾 親彰、上野 知子、山﨑 聖子、菊池 宏和

**LUAプロシジャを起点としたビッグデータ解析とチームコラボレーション ～LUAプロシジャと歩む，初めての共同作業～** ―― 177
木口 亮（塩野義製薬株式会社）
北西 由武

**ODS機能とREPORTプロシジャを用いた解析帳票の作成 －これから業務で利用を始める人のために－** ―― 186
太田 裕二（マルホ株式会社）

**EMBアルゴリズムを用いたMultiple Imputation** ―― 198
矢田 真城（エイツーヘルスケア株式会社）
浜田 知久馬（東京理科大学）

**Proc LuaによるSASマクロに依存しないプログラミングの提案** ―― 209
森岡 裕（イーピーエス株式会社）

**SAS以外の言語を利用した、SASで実現困難なロジックや機能の実装について** ―― 223
川上 貫弘（電助システムズ株式会社）

**SASプログラムの可視化 －SASプログラムステップフローチャート生成プログラムの紹介－** ―― 228
福田 裕章（MSD株式会社）

**Proc LUAを初めて使ってみた －SASでの処理を条件に応じて変える－** ―― 229
淺井 友紀（エイツーヘルスケア株式会社）

## SASシステム／その他関連分野

**SAS Tips. Top10 My Favorite SAS Functions.** ―― 233
松沢 享（株式会社 インターサイエンティフィックリサーチ）

## 生物統計・医薬関連

**PHREGプロシジャにおける共変量調整解析に関連したオプション機能** ―― 237
魚住 龍史（京都大学）
矢田 真城、浜田 知久馬（東京理科大学）

**区分指数モデル及びフレイルティモデルに対するベイズ流生存時間解析** ―― 249
矢田 真城（東京理科大学）
魚住 龍史（東京理科大学、京都大学）、浜田 知久馬（東京理科大学）

**申請電子データプロセスの効率化 ～メタデータ・社内ツールを中心に～ Metadata Driven Approach for eSubmission** ―― 261
飯島 康浩／梶原 正幹／小南 真由美（ノバルティス ファーマ株式会社）
中嶋 優一、北原 孝志

**SASによるベイズ流試験の成功確率の計算** ―― 262
張 方紅（ノバルティス ファーマ株式会社）
北原 孝志

**SASによる傾向スコアマッチング** ―― 263
魚住 龍史（京都大学）
矢田 真城（東京理科大学）、山本 倫生（岡山大学）、川口 淳（佐賀大学）

**Mantel-Haenszel法により調整したリスク差の信頼区間に関する一考察** ―――― 282
佐々木 英麿(武田薬品工業株式会社)
舟尾 暢男

**ベイズ流の単群第Ⅱ相試験デザインのSASマクロの作成** ―――― 301
橋本 大哉(国立病院機構名古屋医療センター)
嘉田 晃子

**公開医療ビッグデータを活用した医薬品マーケティングの可能性** ―――― 302
武藤 猛(MarkeTech Consulting)

**医薬品開発の意思決定におけるBayesian Posterior Probabilityの適用例 ～Random-Walk Metropolis vs. No-U-Turn Sampler～** 312
作井 将(武田薬品工業株式会社)
清水 康平、舟尾 暢男

**Rプログラムも使えるJMPの活用法 ～JMPに手法がない場合～** ―――― 328
原茂 恵美子(株式会社タクミインフォメーションテクノロジー)
佐藤 拓哉、篠津 和夫、佐藤 耕一

**第Ⅰ相臨床試験データの薬物濃度 －QTc解析** ―――― 329
塩境 一仁(第一三共株式会社)
渡橋 靖

**SASによる臨床試験の解析速報に用いる プレゼンテーションスライドの効率的な作成方法の提案** ―――― 339
吉田 直記(武田薬品工業株式会社)
山崎 文寛、舟尾 暢男、高浪 洋平

**レガシーデータ変換等におけるトランスコーディングの問題に関する考察と提案** 361
山崎 文寛(武田薬品工業株式会社)
高浪 洋平

**2値データのIntraclass Correlation Coefficientの推定マクロプログラム** 378
稲葉 洋介(国立国際医療研究センター)
田中 紀子

**GLMSELECTプロシジャにおけるLassoの有用性に関する検討** ―――― 392
川﨑 勝己(エイツーヘルスケア株式会社)
有光 導徳、新城 博子

**SASを利用したPinnacle21 Communityの セミオートメーション化による業務効率化** ―――― 393
片山 雅仁(塩野義製薬株式会社)
勘場 大、神谷 亜香里、江川 広明、角谷 伸一

**PSMATCH Procedure×ニューラルネットワークの奇跡のマッチング ～Real World Dataの利用可能性を付度する～** ―――― 394
馬場 崇充(塩野義製薬株式会社)
藤原 正和、北西 由武

**電子データ申請に向けた品質管理サイクルの最適化 ～プログラミング業務におけるdeja vuからの脱却～** ―――― 395
小林 千鶴(塩野義製薬株式会社)
平野 勝也、坂井 絵理、北西 由武

セントラルモニタリングにおいて
JMP Clinicalのレポートを効果的に選択する取り組みについて ―― 396
福升 悠一(エイツーヘルスケア株式会社)
藤田 侑子、小幡 智憲、畑山 知慶、近藤 秀宣

RTF出力のマクロプログラム化の工夫点の紹介について ―― 397
楊 敏(株式会社タクミインフォメーションテクノロジー)
宮川 元志、原茂 恵美子、篠津 和夫

Annotated CRF自動化に伴う業務効率化について ―― 398
藤原 由(株式会社タクミインフォメーションテクノロジー)
宮川 元志、楊 敏、篠津 和夫

一般化Wilcoxon検定を用いた中間解析の無効中止に関する研究 ―― 399
大久保 竣祐(東京理科大学)
浜田 知久馬、安藤 宗司

XSLプロシージャ(PROC XSL)を使ってみました。
--簡単に"define.xml"を読み込むために。-- ―― 400
小澤 義人(株式会社メディサイエンスプラニング)

多群における二重頑健推定量を用いた交絡の調整法に関する研究 ―― 401
江森 健人(東京理科大学)
浜田 知久馬、安藤 宗司

プラセボ効果を抑制し試験薬の効果を効率的に検出するための
エンリッチメントデザインに関する研究 ―― 402
舛田 優(東京理科大学)
浜田 知久馬、安藤 宗司

臨床試験における2段階デザインの性能評価に関する研究 ―― 403
井澤 慶信(東京理科大学)
浜田 知久馬、安藤 宗司

## 生物統計・医薬関連／SASシステム

proc STREAMによるAnalysis Results Metadataの作成 ―― 407
関根 暁史(藤本製薬株式会社)

帳票MockupからのRTF用テンプレートSASプログラム自動作成ツール
「Taiho TLF Automated Tool(TTAT)」の紹介 ―― 423
伊藤 衡気(株式会社タクミインフォメーションテクノロジー)
栗矢 芳之、銭本 敦(大鵬薬品工業株式会社)

信頼区間を塗りつぶす ～SGANNOを利用して～ ―― 424
関根 暁史(藤本製薬株式会社)

## チュートリアル

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モデルの被験者数設計, 最適被験者数配分 ―― 427
浜田 知久馬(東京理科大学)

**複数母集団のパラメータ推定 ー全国消費実態調査を用いてー** ―― 428
高橋 行雄（BioStat研究所株式会社）

## 統計理論

**GAMプロシジャ versus GAMPLプロシジャ** ―― 443
伊庭 克拓（大塚製薬株式会社）

## その他関連分野

**歌舞伎「仮名手本忠臣蔵」の上演パターン** ―― 459
坂部 裕美子（公益財団法人 統計情報研究開発センター）

**CDASHからSDTMを効率的に作成するためのTipsのご紹介** ―― 460
小松 邦岳（株式会社アスクレップ）
鈴木 淳一、村岡 了一、飛石 陵子

## Let's データ分析

参加カテゴリー：ニューカマー **SAフレッシャーズ**
自由課題レポート：**若者の消費傾向はライフスタイルが変化するにつれ、どのように変化するか？：横断研究** ―― 463
規定課題レポート ―― 470

参加カテゴリー：ニューカマー **タクミインフォメーションテクノロジー チームA**
自由課題レポート：**35歳未満の独身男女の消費傾向について** ―― 475
規定課題レポート ―― 482

参加カテゴリー：ニューカマー **タクミインフォメーションテクノロジー チームB**
自由課題レポート：
**居住地区の選択に関する探索的研究 ー人口の移動による過密化という社会問題に焦点を当てて－** 487
規定課題レポート ―― 493

参加カテゴリー：ニューカマー **タクミインフォメーションテクノロジー チームC**
自由課題レポート：**企業区分別にみる自動車保有世帯の傾向** ―― 499
規定課題レポート ―― 506

参加カテゴリー：ニューカマー **タクミインフォメーションテクノロジー チームD**
自由課題レポート：**高齢者世帯の消費動向** ―― 511
規定課題レポート ―― 516

参加カテゴリー：ニューカマー **野毛坂123**
自由課題レポート：**ストックの視点を含んだ世帯の消費行動の特徴抽出** ―― 522
規定課題レポート ―― 527

参加カテゴリー：学生 **おやつクラブ**
自由課題レポート：**経済格差を測る食事多様性指標の提案** ―― 533
規定課題レポート ―― 540

# 基調講演

# 医療における改正個人情報保護法と次世代医療基盤法

○山本　隆一
（一般財団法人医療情報システム開発センター）

2015年9月に改正個人情報保護法が可決され、2017年5月末に全面施行された。改正点は多岐にわたるが、その主なものは、それのみで個人を一意に識別可能な個人識別符号の導入とゲノムシークエンスが原則個人識別符号であることの明確化、匿名加工情報の導入、機微な情報として要配慮個人情報の概念が導入され、医療情報のほぼすべてが要配慮個人情報に指定されたこと、第三者提供における同意原則の強化、第三者提供におけるトレーサビリティの確保、罰則の強化、個人情報保護委員会の設立などである。この中でも医療情報の二次利用にもっとも大きな影響を与える点は、要配慮情報の概念の導入で、要配慮情報は明な同意なしに第三者提供することが原則禁止される。匿名加工情報であれば同意なしの第三者提供が可能であるが、匿名加工情報を作成する時点でその事実を公表しなければならず、また匿名加工の基準も高い。これまで簡単に匿名化と考えていた処理も匿名加工の基準さえ満たさないことがありうる。改正個人情報保護法では４章に個人情報を取得する事業者の責務が記載されているが、「学術研究」であれば４章は除外される。ただし、「学術研究」とはもっぱら学術研究を主な業務とする機関かあるいは学会等で行われる学術研究に限定され、別に定められた倫理指針に準拠することが求められる。この倫理指針も全体に見ればかなり厳しい指針になっており、学術研究であっても二次利用は相当困難になり、公益的であっても学術研究以外の二次利用は厳しく制限される。

この厳格化自体は機微な情報である医療情報の不用意な、あるいは本人が思いもよらないような二次利用を防止するという点では、IT化が進む医療や健康管理で国民の安心感を得ることに寄与することが期待され、意義が大きい。

その一方で医療・健康に係わる情報は当人の健康の維持回復だけに用い、医学研究や創薬、医療機器あるいは新し医療技術の開発に用いることができなければ、医学・医療の発展に重大な影響を与えることは明らかである。つまり少し長い目で見れば医学や創薬および医療機器の開発などに支障を来し、医療や健康管理の衰退を来しかねない。

この問題は改正個人情報保護法案の議論の時点から認知されており、個人の権利を確実に保護した上で、広い意味での公益的利用を促進する必要性が、主に内閣官房の健康医療戦略室を中心に議論され、2017年4月に「医療分野の研究開発に資するための匿名加工医療情報に関する法律」が成立した。広い意味での公益目的の二次利用を促進するために、しっかりとした匿名加工能力を持つ法人を認定し、この認定機関に医療機関等から情報を第三者提供する場合、明な同意ではなく、丁寧に説明した上でのオプトアウトで提供することを可とするもので、「次世代医療基盤法」と呼ばれている。広い意味での公益性が前提ではあるものの、患者等の権利を確実に保護し、二次利用を促進するという点では画期的な制度と言える。ただ実施は制定後１年以内とされており、まだこれから細部を詰めなければならない。またこの法律だけで十分かどうかも検討の余地がある。

本講演では改正個人情報保護法制の概略を示した上で、次世代医療基盤法を概説し、今後の問題点に触れたい。

# CDISC企画セッション

# ある実務担当者が考えるADaMダイバーシティと未来予測

○渡辺　寿恵
（小野薬品工業株式会社）

Clinical Data Interchange Standards Consortium (CDISC) Analysis Data Model (ADaM)リリースから約9年、ADaMへの理解の深まりやValidation Toolの普及によりCDISCに準拠したADaMデータ作成が容易になりつつある。また2020年から義務化されるPMDAへのCDISCデータ提出に向け、各企業、CROでも日常的にADaMを使用するようになってきている。このような時流の中、SDTMと比較して自由度の高いADaMは各企業内で独自の進化の道を歩み、多種多様な形態へと変貌していっている。本発表では、生命の実験室「カンブリア大爆発」の如く、不思議で奇妙な構造へと進化を遂げたADaMの事例や来るべきTechnological Singularityで消える？かもしれないADaMプログラマーの視点から標準開発アプローチやADaMの未来予測を下記の流れに沿って何となく紹介するものである。

第1章：増殖するADaM（枯渇するNaming Conventions）
　　大繁殖したADaM変数とその増殖原因、課題を考察する。

第2章：分裂するADaM（ADaM分化：TraceabilityとAnalysis readyの罠）
　　解析には使用しない中間データ，集計別データ。

第3章：結合するADaM（巨大化するADaM：ADSL，BDSの台頭）
　　5個のADaMでCSRのTLFは網羅できるのか？データの最適数について考察する。

第4章：変容するADaM（Therapeutic Area Standardへの拡張）
　　Therapeutic Area Standardに掲載されているADaM構造と標準開発について考察する。

第5章：ADaM進化予測（General MedicineからPrecision Medicine？）
　　多様性は淘汰されるべきか？ある実務担当者の考えるADaM 標準開発アプローチ。

以上

# PMDA申請電子データ提出経験談

○栗栖章[1]、鈴木正人[1]、福田裕章[1]
([1]MSD株式会社)

2016年10月より医薬品医療機器総合機構（PMDA）において医薬品承認申請時の臨床試験の電子データ提出がついに開始された．2014年6月に「承認申請時の電子データ提出に関する基本的考え方について」（基本的通知），2015年4月に「承認申請時の電子データ提出に関する実務的事項について」（実務的通知）及び「承認申請時の電子データ提出に関する技術的ガイドについて」（技術ガイド）が発出されて以来，業界では来たる提出開始までに様々な検討が行われ準備が進められてきた．また，通知発出後に寄せられた問い合わせ等を受け，通知に補足した方が良い情報は「申請電子データに関するFAQ」として公表され，提出すべきデータの詳細が明らかにされてきている．しかしながら申請者を取り巻く環境は様々で，既存の社内標準システムをどのように変えていくべきか，どのようなタイムラインで進めていくのか，どのように部署間で業務を分担し進めていくのかは，実際に経験してみないと分からないことも少なくない．また，PMDA申請だけでなく米国食品医薬品局（FDA）への申請を考慮に入れる場合は，規制要件の細かい違いを認識し，効率よく双方の電子データを準備する必要がある．

本発表では，FDAに提出したオンコロジー領域における国際共同治験の電子データをPMDAに提出した経験を踏まえて，その事例とそこから得られた学びを紹介する．

# 多重性の調整が必要な試験を実施した際の承認申請時電子データ提出方法の検討

○大内喜海[1]、福田裕章[2]、大野浩太[3]、浅見由美子[4]

（[1]協和発酵キリン株式会社及びCJUG-ADaM、[2] MSD株式会社、CJUG-ADaM、[3]北海道大学病院　臨床研究開発センター、CJUG-ADaM、[4]第一三共株式会社及びCJUG-ADaM）

「健康・医療戦略」における 3 つの基本理念（健康長寿社会の実現、経済成長への寄与、世界への貢献）を実現するための戦略の 1 つとして、独立行政法人医薬品療機器総合機構（以下、PMDA）の強化が挙げられており、規制当局である PMDA 自らが臨床データ等を活用した解析や研究を推進することが求められている。それを受けて、平成 26 年 6 月 20 日に厚生労働省医薬食品局審査管理課長より発出された「承認申請時の電子データ提出に関する基本的考え方について（薬食審査発 0620 第 6 号）」により、医薬品の承認申請時に臨床試験データの提出が求められることになった。同通知では、電子データは「Clinical Data Interchange Standards Consortium の規格（以下「CDISC 標準」という。）に準拠した形式で提出すること」、「解析データセットとして Analysis Data Model（ADaM）に基づくデータセット及びその定義ファイル（Define.XML 等）並びに ADaM データセット作成用のプログラムを提出すること」等が申請者（製薬企業、アカデミア等）に求められている。

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use（以下、ICH） E9 では、臨床試験において、主要変数が複数ある場合、試験治療間の多重比較、時間に伴う繰り返し評価（中間解析）等により検定の多重性が生じる場合、試験全体での第一種の過誤を 5%以下（片側仮説の場合は 2.5%）に抑えることを求めている。そのため、臨床試験のデザイン等に応じて様々な Multiple Comparison Procedures(以下、MCPs)を利用している。

近年、柔軟な仮説構造に対応できる Gatekeeping procedure 等が提案され、MCPs に関する活発な議論がなされてきたものの、それらの電子データ提出方法に関する議論は活発になされてこなかった。そこで CDISC、PMDA がそれぞれ公表している「ADaM Implementation Guide」、「承認申請時の電子データ提出に関する実務的事項について／技術的ガイドについて」等の関連資料と照らし合わせて、多重性の調整が必要な試験を実施した際の承認申請時電子データ提出方法を検討したところ、以下のような点で電子データ提出方法が明確ではないと考えられた。

・SAS のプロシジャが MCPs に対応していない場合
・前の検定結果に依存して、有意水準の配分が変わる場合
・MCPs を文章で表現することが困難な場合

本発表では、上記の点を踏まえて多重性の調整が必要な試験を実施した際の承認申請時電子データ提出方法について検討した結果を報告する。

# ある実務担当者が考えるADaM開発アプローチと標準の紹介

○坂上　拓

（株式会社 中外臨床研究センター バイオメトリクス部 データサイエンスグループ）

2020年からPMDAへのCDISCの標準データモデルでの申請電子データの提出義務化に伴い，各企業でのCDISCへの取り組みが始まり，実務担当者のナレッジの蓄積やバリデーションツール等の普及で，CDISC 標準データモデルに準拠したデータセットを開発できるようになってきた。

標準データモデルの一つで，解析を目的としたデータモデルであるADaMは，様々な解析要件に対応できるように，自由度が高めに規定されている。その高い自由度であるが故，特定のADaMを作成するにも多くのアプローチがあり，結果的に数多くのバリエーションのADaMが生まれることになる。これらは，いずれもADaMとして誤りではないが，ADaMとして最適なものであるとの確証も得られないのが現状である。最適なADaMを作成していくには，作成された様々なバリエーションのADaMを比較し，良い箇所を採用し，既存の標準モデルを改変しながら標準としてブラッシュアップして行くしかないが，実際，社内で発生したバリエーション以外を目にする機会もそれほどない。

本発表では，実務担当者として経験した様々な問題や疑問点を元に，ADaMの開発アプローチとして，作成するADaMの種類（ADSL，BDSやOCCDSといった標準構造のものから，ADaM OTHERやnon-ADaMと呼ばれるデータセットまで）や，ADaMでの新規変数の追加やIGで規定されている変数の運用，レコードの発生方法を実際のデータ構造と共に紹介する。

# USでの申請承認プロセスとCDISC対応の事例

○伊藤雅憲
(アステラス製薬株式会社)

USにおける感染症領域の第三相試験統計解析業務、FDAへの新薬承認申請関連業務、アドバイザリーコミッティ関連業務の経験を通して、申請承認プロセスとそれに伴うCDISC・電子データ申請業務のプロセスについて紹介する。本プロジェクトは2014年に申請された。申請パッケージに含まれる試験概要として、第三相試験が2本、第二相試験が2本、第一相試験がDDI、TQT、SAD、MADなど合計42本がファイリングされた。申請におけるCDISC標準の対応として、メタデータ／試験データのSDTMおよびADaMに関する記述であるData Definitions Document (Define.xml and PDF) が作成された。これはOpen CDISC (現pinnacle 21)実行でのデータの一致性の確認にも用いられた。また、有効性の複雑なデータハンドリングのロジックは、Supplemental data definition.pdfというものを作成し、ドメイン名、変数名、記述（ロジック）を整理した。また、CDISC基準に照らし合わせて、実際のSDTM及びADaMデータのすべてのドメインに関してWarningまたはErrorを検出したリストとしてOpen CDISC reportを作成した。本レポートは申請資料には含めないが、Reviewer’s Guideにて要約された。そのReviewer’s Guideはファイリングされた臨床試験、及び安全性併合解析それぞれについて作成されたFDA Reviewerのための取扱説明書であり、これらは申請資料に含まれた。事例ではReviewer’s Guideは統計担当者によって作成され、申請業務で多忙な時期にやや辛い負荷であったことを記憶している。
当日の発表では、CDISC関連以外にも、申請後のFDA照会事項対応、アドバイザリー会議の準備（Mock Meeting等）と当日の対応の経験談、外部レジストリデータベースを用いたマッチドペア解析による申請承認プロセス、パブリケーション業務と統計解析担当者の関わり、外部統計専門家とのコミュニケーションなどの経験談についても紹介する。

# 医薬品承認申請時電子データ提出の現状について

○安藤友紀
（独立行政法人医薬品医療機器総合機構）

医薬品医療機器総合機構（PMDA）では、承認申請データを一層活用した承認審査や相談の質の向上を図るため、医薬品承認申請時に添付される臨床試験データ利用の具体的な検討を行ってきた。電子データの受付が開始された2016年10月1日以降には、医薬品承認申請時に実際に臨床試験の電子データが添付される申請品目も出てきている。

申請時電子データ提出に関しては、これまで複数の通知や技術的ガイド、「申請電子データ提出に際して利用可能な規格一覧」、「バリデーションルール一覧」、及び「申請電子データに関するFAQ」により、提出すべきデータに関する方針が示されてきている。また、2015年より開始した申請電子データ提出確認相談の内容や、2016年10月1日以降の実際の電子データ受付の経験を踏まえて、2017年2月にはFAQの追加がなされ、今後の技術的ガイドの改定も予定されている。承認申請時に実際に提出された臨床試験の電子データについては、承認審査のプロセスにおいて、各品目の有効性及び安全性に関する理解や議論の補助等に実際に用いられてきている。

本発表では、これまでの申請時電子データ提出の状況を説明するとともに、受け入れの経験に基づく留意点等を紹介する。また、PMDAにおける電子データの使用状況について説明する。

# オープンデータ

# 第5回 Let's データ分析コンテストの規定課題の SAS プログラム

周防　節雄
兵庫県立大学・名誉教授

# Exemplary SAS Programs to Solve the Compulsory Exercises for the 5th Micro Data Competition in Japan SAS Users Forum 2017

Setsuo Suoh
Professor Emeritus of the University of Hyogo

## 要旨

SAS ユーザー総会 2017 における「Let's データ分析　第5回ミクロデータ分析コンテスト」の規定課題について、出題者の立場から、模範解答の SAS プログラムを例示する。

**キーワード**　全国消費実態調査、匿名データ、新擬似ミクロデータ、集計乗率、proc freq、proc tabulate、proc sgplot、proc univariate、SG プロシジャ、プログラミング・スタイル、ジニ係数、ローレンツ曲線、

## 1. はじめに

独立行政法人統計センターは、平成16年の全国消費実態調査のミクロデータから教育目的で教育用擬似ミクロデータを作成し、公的ミクロデータの利用実習用に提供してきた。SAS ユーザー総会では、2013 年から 2016 年まで、この擬似ミクロデータを使用してデータコンペを実施してきたが、2016 年度末でこのデータの提供が中止になった。そこで、今年に入って急遽「新擬似ミクロデータ」を SAS ユーザー会世話人有志が集まり、2004 年全国消費実態調査の匿名データを使って、今年のデータコンペ用の「新」擬似ミクロデータ[1]を作成した。

コンペでは、この新擬似ミクロデータを使った規定課題（付録3）と自由課題から成り、規定課題は参加者全員に課せられる。本稿では、この規定問題のためのSASプログラムを例示する。出力結果(付録2)にある図表のタイトルはプログラムの title 文と対応しているので、各プログラムの処理プロセスを追っていける。

## 2. 新擬似ミクロデータの概要

作成した新擬似ミクロデータと符号表（メタデータ）は、ウェブ上にアップロードしている[2]。

新擬似ミクロデータは 69,131 オブザベーション(世帯)から成る。変数としては、①世帯識別番号（変数名 No=1～69,131）、②14 個の世帯属性（変数名 X01～X14）と集計乗率（変数名 Weight）（表1）[3]、および、③203 個の収支項目（変数名 Y001～Y203）（表 2）がある。この両表の元ファイル

---

1 この新擬似ミクロデータの作成経緯とプロセスについては、本論文集に収録の高橋他(2017)を参照されたい。
2 SASデータセット http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroData.zip
CSV形式　http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroDataCSV.zip
3 この他に、調査年を意味する変数 Year があるが、現在の新擬似ミクロデータでは値は全て「2004」である。これは、将来、他の調査年次の擬似ミクロデータ化も予定しているからである。

(メタデータ)は、ウェブにアップロードした符号表「全消 2004 年新擬似ミクロデータ符号表.xlsx」に収録されている。

## 3. 規定課題 SAS プログラム模範解答プログラムの実行手順

規定課題のために、今回「作成」したSASプログラム(付録1)は、以下の①～⑩であるが、⑥のプログラムは③のSASプログラムが自動作成した。

① set_environment.sas 環境設定:SAS を立ち上げたときに★いつも★最初に実行する。

② setALL.sas → ③④⑤が%include を使って自動的に実行される。
③cr eateFor matNewZenshoGiji2004.sas→*自動作成*→⑥*procFormatNewZenshoGiji2004.sas*
④cr eatetLabelNewZenshoGiji2004.sas →*自動作成*→⑪*labelNewZenshoGiji2004.txt*
⑤setLABELweightsas→④の結果を使って、変数ラベルと 10 万世帯比 weight 変数を付加

⑥ procFormatNewZenshoGiji2004.sas → 変数 format が読み込まれる。

⑦ kitei2017.sas 規定課題の第1問～第4問を解答する。

★規定問題:第5問 ジニ係数計算とロレンツ曲線描画

⑧ make_weighted_dataX. sas
→ 集計乗率 weight 変数[4]を使って膨らませた SAS データセット"reconstruct"を作成

⑨ gini_lorenz2017 規定課題_ sas_forum_macro.sas → ジニ係数とロレンツ曲線を表示[5]

⑩ GiniLorenz2017.sas → SAS データセット"reconstruct"を家族分類別に分割した後で家族分類毎に、⑨を繰り返し自動実行して、ジニ係数とロレンツ曲線表示。

実行手順は以下の通り。

●初めて実行する時は、①②⑥⑦⑧⑩の順に実行する。
●同じ SAS のセッションで再実行の際は、⑦⑧⑩の順に実行する。
●SAS の新しいセッションで再実行する時は、①⑥⑦⑧⑩の順に実行する。

変数フォーマットの作成(プログラム⑥)と変数ラベル文(テキストファイル⑪)は、EXCEL ファイルの符号表をSASプログラムに読ませることで、全自動で処理をした。

## 4. SAS プログラミング・スタイル

本稿で示した解答例は一例であり、もちろん他のやり方もいろいろある。この解答例のプログラミング・スタイルは、ここ数年来、このユーザー総会でも一貫して採用している方式である。

SASのプログラムを書く際には、通常は人間のプログラマーが全てのSASコードを書いている。複雑なアルゴリズムをプログラム化するには、プログラマー自身でやらざるを得ないが、非常に単純だが量が多くて機械的なコード(いわゆる、"dirty work")の場合は、できれば「自動化」したい。例えば、大量の変数の procformat やlabel 文がそれに相当する。時間をかければもちろん出来るが、

4 集計乗率については、周防(2016)参照
5 ジニ係数とロレンツ曲線のSASプログラムについては、周防(2008)参照。そこに掲載の付録3のプログラムを規定問題用に一部書き換えて、プログラム⑨を作成した。

手作業でした場合、ケアレスミスが発生しやすい。今回の規定課題でも、proc format や label 文の作成は、エクセル形式の符号表ファイル(メタデータ)をSASプログラムに読ませることによって行った。

もう一つ重要なプログラミング・スタイルは、ひとつのプログラムで全ての処理をしないことである。つまり、全体の作業をいくつかのまとまった処理に分割してから、それぞれを個別のSASプログラムにする。SASプログラムには、メインプログラムとサブプログラムの概念が基本的にはない[6]。異なるプログラムの間は、SAS データセットで連携されている。第 3 節で示したプログラム間の連携はその例である。この方式のメリットは、バグが発生したときに、発生箇所を直ぐに発見できるので、容易に対処できることである。デメリットとしては、プログラムの数が増えて、混乱する可能性がある。その対策としては、第 3 節で示すように、**%include** を使って、いくつかの一連のプログラムが正しい順序で実行されるように、風呂敷に包むような形で、もうひとつ別の上位プログラムで包んでしまうのが便利である。第 3 節の②のプログラム **setALL.sas** がその上位プログラムに相当する。その上で、第 3 節の終わりにある三つの●で示すように、異なるシチュエーションに応じて、実行する順番をきちんと文書化しておく。この三つの実行手順を更に、それぞれ**%include** を使って、風呂敷で包むのも良い。

## 5. まとめ

今回のコンテストでも、集計乗率を使って、proc freq や proc tabulate をうまく使えるかが試された。集計用乗率という用語に余りなじみのないユーザーもいると思うが、公的ミクロデータの利用では必ず出てくる概念である。公的ミクロデータの分析では、初期段階で分布状況の把握のために、これらのプロシジャを使って度数分布を確認することがよく行われる。

グラフの出力も求められたが、こちらは、JMPを使う方が手軽だと感じた。筆者は、公的ミクロデータを主に利用する関係上、普段余りSASで特殊なグラフを描く必要性が少ないので、可能ならば、他のソフト、例えば EXCEL などでできる範囲内でお茶を濁すことが多い。昨年も今年も、規定課題のグラフ作成にはSGPLOTを使ってみたが、細かい微調整ができずにいて、正直なところ、自分でもまだ十分に使いこなせているとは言えない。SGPLOT は数年前から利用可能になった SG プロシジャ(Statistical Graphics Procedures)[7]に含まれるプロシジャであり、SAS/GRAPH とは別物である。上手に使えば、便利なツールになる。

## 参考文献


1. 周防節雄(2008)ジニ係数の計算とロレンツ曲線を描くSASプログラム、『SASユーザー総会2008論文集』、pp139-146
2. 高浪洋平(2011)SG プロシジャと GTL によるグラフの作成と ODS PDF による統合解析帳票の作成 ～TQT 試験における活用事例～、『SAS ユーザー総会2011論文集』、pp201-219
3. 周防節雄(2015)SAS プログラムで関数とサブルーティンを作成する方法、『SAS ユーザー総会2015論文集』、pp389-398
4. 周防節雄(2016)SAS ユーザー総会2016における「Let's データ分析第4回ミクロデータ分析コンテスト」の規定課題の SAS プログラム解説、『SAS ユーザー総会2016論文集』、pp330-347


---

6 数年前から SAS でも関数や副プログラムをユーザー自身が定義出来るようにはなった(周防(2015))。ただ、いわゆるメインプログラムという概念は SAS には馴染まない。

7 本総会でもこれまでに発表事例(高浪 2011)がいくつかある。

# 表１ 世帯属性と集計乗率

data=formatNewZenshoGiji2004 符号表

| OBS | F1 | comment | varName | V | meaning |
|---|---|---|---|---|---|
| 1 | | 世帯に関する項目の変数名及び符号 | | . | |
| 2 | | 項目名 | 変数名 | . | 符号内容 |
| 3 | | 調査年 | Year | . | 全国消費実態調査の調査年(西暦) |
| 4 | | レコード一連番号 | No | . | |
| 5 | | 大都市圏の別 | X01 | 1 | ３大都市圏 |
| 6 | | | | 0 | その他 |
| 7 | | 世帯区分 | X02 | 1 | 勤労者 |
| 8 | | | | 2 | 勤労者以外 |
| 9 | | | | 3 | 無職 |
| 10 | | 世帯人員 | X03 | 1 | １人 |
| 11 | | (住居と生計を共にしている世帯員数) | | 2 | ２人 |
| 12 | | | | 3 | ３人 |
| 13 | | | | 4 | ４人 |
| 14 | | | | 5 | ５人以上 |
| 15 | | 就業人員 | X04 | 0 | ０人 |
| 16 | | (就業している世帯員数) | | 1 | １人 |
| 17 | | | | 2 | ２人 |
| 18 | | | | 3 | ３人以上 |
| 19 | | 住居の構造 | X05 | 1 | 木造(防火木造含む) |
| 20 | | | | 2 | 木造(防火木造含む)以外 |
| 21 | | 住居の建て方 | X06 | 1 | 一戸建 |
| 22 | | | | 2 | 一戸建以外 |
| 23 | | 住居の所有関係 | X07 | 1 | 持ち家 |
| 24 | | | | 2 | 持ち家以外 |
| 25 | | 世帯主の性別 | X08 | 1 | 男 |
| 26 | | | | 2 | 女 |
| 27 | | 世帯主の年齢 | X09 | 5 | 24歳以下 |
| 28 | | | | 6 | 25～29歳 |
| 29 | | | | 7 | 30～34歳 |
| 30 | | | | 8 | 35～39歳 |
| 31 | | | | 9 | 40～44歳 |
| 32 | | | | 10 | 45～49歳 |
| 33 | | | | 11 | 50～54歳 |
| 34 | | | | 12 | 55～59歳 |
| 35 | | | | 13 | 60～64歳 |
| 36 | | | | 14 | 65～69歳 |
| 37 | | | | 15 | 70～74歳 |
| 38 | | | | 16 | 75歳以上 |
| 39 | | 企業区分･従業者規模 | X10 | 1 | 民営･自営1～4人 |
| 40 | | | | 2 | 民営･自営5～29人 |
| 41 | | | | 3 | 民営･自営30～499人 |
| 42 | | | | 4 | 民営･自営500人以上 |
| 43 | | | | 5 | 官公 |
| 44 | | | | 6 | 無職 |
| 45 | | 家族分類 | X11 | 1 | 単身世帯 |
| 46 | | | | 2 | 夫婦のみ世帯 |
| 47 | | | | 3 | 二世代世帯 |
| 48 | | | | 4 | 二世代(ひとり親)世帯 |
| 49 | | | | 5 | 三世代世帯 |
| 50 | | | | 6 | その他の世帯 |
| 51 | | 未就学児の有無 | X12 | 1 | 無 |
| 52 | | | | 2 | 有 |
| 53 | | 学校に通う世帯員の有無 | X13 | 1 | 無 |
| 54 | | | | 2 | 有 |
| 55 | | 65歳以上の世帯員数 | X14 | 0 | ０人 |
| 56 | | | | 1 | １人 |
| 57 | | | | 2 | ２人以上 |
| 58 | | 集計用乗率 | Weight | . | |

プログラム③createFormatNewZenshoGiji2004.sasで作成した

# 表 2 収支項目一覧

data=labelNewZenshoGiji2004 符号表

| OBS | F1 | varName | label | F4 |
|---|---|---|---|---|
| 1 | | 変数名一覧 | | |
| 2 | | 変数名 | 項目名 | |
| 3 | | ○世帯事項 | | |
| 4 | | Year | 調査年 | |
| 5 | | No | レコード連番号 | |
| 6 | | X01 | 大都市圏の別 | |
| 中略 | | | | |
| 13 | | X08 | 世帯主の性別 | |
| 14 | | X09 | 世帯主の年齢 | |
| 15 | | X10 | 企業区分・従業者規模 | |
| 16 | | X11 | 家族分類 | |
| 中略 | | | | |
| 20 | | Weight | 集計用乗率 | |
| 21 | | ○収支項目 | | |
| 22 | | Y001 | 年間収入 | (単位:万円) |
| 23 | | Y002 | 収入総額 | (単位:円)以下同様 |
| 24 | | Y003 | 実収入 | |
| 25 | | Y004 | 経常収入 | |
| 26 | | Y005 | 勤め先収入 | |
| 27 | | Y006 | 世帯主の勤め先収入 | |
| 28 | | Y007 | 世帯主の配偶者の勤め先収入 | |
| 29 | | Y008 | 他の世帯員の勤め先収入 | |
| 30 | | Y009 | 事業・内職収入 | |
| 31 | | Y010 | 農林漁業収入 | |
| 32 | | Y011 | 家賃収入 | |
| 33 | | Y012 | 他の事業収入 | |
| 34 | | Y013 | 内職収入 | |
| 35 | | Y014 | 本業以外の勤め先・事業・内職収入 | |
| 36 | | Y015 | 他の経常収入 | |
| 37 | | Y016 | 財産収入 | |
| 38 | | Y017 | 社会保障給付 | |
| 39 | | Y018 | 公的年金給付 | |
| 40 | | Y019 | 他の社会保障給付 | |
| 41 | | Y020 | 仕送り金(収入) | |
| 42 | | Y021 | 特別収入 | |
| 43 | | Y022 | 受贈金 | |
| 44 | | Y023 | その他の特別収入 | |
| 45 | | Y024 | 実収入以外の収入 | |
| 46 | | Y025 | 預貯金引出 | |
| 47 | | Y026 | 保険取金 | |
| 48 | | Y027 | 個人・企業年金保険取金 | |
| 49 | | Y028 | 他の保険取金 | |
| 50 | | Y029 | 有価証券売却 | |
| 51 | | Y030 | 株式売却 | |
| 52 | | Y031 | 他の有価証券売却 | |
| 53 | | Y032 | 土地家屋借入金 | |
| 54 | | Y033 | 他の借入金 | |
| 55 | | Y034 | 分割払・一括払購入借入金 | |
| 56 | | Y035 | 財産売却 | |
| 57 | | Y036 | その他の実収入以外の収入 | |
| 58 | | Y037 | 繰入金 | |
| 59 | | Y038 | 支出総額 | |
| 60 | | Y039 | 実支出 | |
| 61 | | Y040 | 消費支出 | |
| 62 | | Y041 | 食料 | |
| 63 | | Y042 | 穀類 | |
| 64 | | Y043 | 米 | |
| 65 | | Y044 | パン | |
| 66 | | Y045 | めん類 | |
| 67 | | Y046 | 他の穀類 | |
| 中略 | | | | |
| 224 | | Y203 | 繰越金 | |

プログラム④createLabelNewZenshoGiji2004.sasで作成した

太枠で囲った変数が、規定課題で使われる。

# 付録1　SASプログラム一覧

## ① set_environment.sas

```
/* set_environment.sas */
  %let drive=F; *★外付けHDのドライブ名を指定する;
*★全消2004年新擬似ミクロデータのSASデータセット、★SASプログラム、★符号表の保存先のパスを指定;
  %let path=¥全消¥擬似ミクロ作成プロジェクト¥★新情報¥最新情報¥全消2004年新擬似ミクロデータSAS版;
  %let pathSASprogram=¥★X60s(2014-8-9)¥SAS_Forum¥2017¥SASprogram;
  %let pathFugohyo=¥全消¥擬似ミクロ作成プロジェクト¥★新情報¥最新情報¥全消2004年新擬似ミクロデータSAS版;
  %let fugohyoEXCEL=全消2004年新擬似ミクロデータ符号表.xlsx;   *★符号表エクセルファイル名を指定する;
  libname newZ "&drive:&path"; *★この行はそのままにしておく;
```

## ② setALL.sas

```
/* setALL.sas */  *★3つのSASプログラムを実行;
%include "&drive:&pathSASprogram¥createFormatNewZenshoGiji2004.sas" *③;
%include "&drive:&pathSASprogram¥createLabelNewZenshoGiji2004.sas";   *④;
%include "&drive:&pathSASprogram¥setLABELweight.sas";                 *⑤;
```

## ③ createFormatNewZenshoGiji2004.sas

```
/* createformatNewZenshoGiji2004.sas */ *★新擬似ミクロデータ2004用proc formatを自動作成する;

filename out1 "&drive:¥★X60s(2014-8-9)¥SAS_Forum¥2017¥SASprogram¥procFormatNewZenshoGiji2004.sas";

proc import datafile="&drive:&pathFugohyo¥&fugohyoEXCEL"
  out=formatNewZenshoGiji2004 (rename=(f2=comment f3=varName f4=V f5=meaning))
  replace;
  sheet='世帯事項整形'; getnames=no;
run;

proc print data=formatNewZenshoGiji2004; title "data=formatNewZenshoGiji2004 符号表"; run;

data _NULL_; set formatNewZenshoGiji2004;
  file out1; *★proc formatの自動作成★;
  length buff $ 100;
  if _N_=1 then put '/* procFormatNewZenshoGiji2004.sas */';
  if _N_<=4 then return; *最初の4オブザベーションはスキップする;
  if varName="Weight" then do; put ';' / 'run;'; return; end; *最後の集計用乗率は無視する;
  if _N_=5 then put "proc format;";
  formatname=compress(varName || 'F'); *format名の末尾が数字は反則なので、Fを末尾に付す;
  if varName NE " " then put ';' / '*' comment ';' / 'value ' formatname;
  buff=V||'='''||V||':'||meaning||'''"; buff=kcompress(buff); put buff;
run;
```

## ④ createtLabelNewZenshoGiji2004.sas

```
/* setLABELweight.sas */ *★全消2004年新擬似ミクロデータ:データコンペ規定課題用★データセット★設定;
*★変数ラベル付加、及び、10万世帯比weight変数を作成して、新データセット★NZ2004★を新規作成;

libname newZ "&drive:&path";

data newZ.NZ2004; set newZ.zensho2004gijimicro  end=owari;
  %include "&drive:¥&pathSASprogram¥labelNewZenshoGiji2004.txt";*変数LABEL付与;
  cnt=1; *集計乗率利用時に使用する変数;
  sum_weight+weight; *★10万比の集計乗率を計算する準備;
  if owari then call symputx("sum_weight", sum_weight); *★集計乗率の合計値をmacro変数sum_weightに設定;
run;

%put "&sum_weight"; *★macro変数sum_weightの値の確認;

data newZ.NZ2004; set newZ.NZ2004 end=owari; drop total;
  weight100000=weight*100000/&sum_weight; *★10万比の集計乗率を計算;
  total+weight100000; if owari then put "検算:weight100000の合計＝" total;
  *★LOG画面出力結果→検算:weight100000の合計＝99999.999999;
  *★10万比で計算の際は、集計乗率変数weight100000を使えば簡単;
run; *★以後、規定課題用プログラム作成にはSASデータセット「newZ.NZ2004」を使用する;
```

## ⑤ setLABELweight.sas

```
/* setLABELweight.sas */ *★全消2004年新擬似ミクロデータ:データコンペ規定課題用★データセット★設定;
*★変数ラベル付加、及び、10万世帯比weight変数を作成して、新データセット★NZ2004★を新規作成;

libname
newZ "&drive:¥全消¥擬似ミクロ作成プロジェクト¥★新情報¥最新情報¥全消2004年新擬似ミクロデータSAS版";

data newZ.NZ2004; set newZ.zensho2004gijimicro  end=owari;
  %include "&drive:¥&pathSASprogram¥labelNewZenshoGiji2004.txt";*変数LABEL付与;
  cnt=1; *集計乗率利用時に使用する変数;
  sum_weight+weight; *★10万比の集計乗率を計算する準備;
  if owari then call symputx("sum_weight", sum_weight); *★集計乗率の合計値をmacro変数sum_weightに設定;
run;

%put "&sum_weight"; *★macro変数sum_weightの値の確認;

data newZ.NZ2004; set newZ.NZ2004 end=owari; drop total;
  weight100000=weight*100000/&sum_weight; *★10万比の集計乗率を計算;
  total+weight100000; if owari then put "検算 : weight100000の合計＝" total;
  *★LOG画面出力結果→検算 : weight100000の合計＝99999.999999;
  *★10万比で計算の際は、集計乗率変数weight100000を使えば簡単;
run; *★以後、規定課題用プログラム作成にはSASデータセット「newZ.NZ2004」を使用する;
```

## ⑥ procFormatNewZenshoGiji2004.sas　（自動作成された）

```
/* procFormatNewZenshoGiji2004.sas */
proc format;
;
*大都市圏の別 ;
value X01F
1="1:３大都市圏"
0="0:その他"
;
*世帯区分 ;
value X02F
1="1:勤労者"
2="2:勤労者以外"
3="3:無職"
;
*世帯人員 ;
value X03F
1="1:１人"
2="2:２人"
3="3:３人"
4="4:４人"
5="5:５人以上"
;
*就業人員 ;
value X04F
0="0:０人"
1="1:１人"
2="2:２人"
3="3:３人以上"
;
*住居の構造 ;
value X05F
1="1:木造(防火木造含む)"
2="2:木造(防火木造含む)以外"
;
*住居の建て方 ;
value X06F
1="1:一戸建"
2="2:一戸建以外"
;
*住居の所有関係 ;
value X07F
1="1:持ち家"
2="2:持ち家以外"
;
*世帯主の性別 ;
value X08F
1="1:男"
2="2:女"
;
*世帯主の年齢 ;
value X09F
5="5:24歳以下"
6="6:25～29歳"
7="7:30～34歳"
8="8:35～39歳"
9="9:40～44歳"
10="10:45～49歳"
11="11:50～54歳"
12="12:55～59歳"
13="13:60～64歳"
14="14:65～69歳"
15="15:70～74歳"
16="16:75歳以上"
;
*企業区分･従業者規模 ;
value X10F
1="1:民営･自営1～4人"
2="2:民営･自営5～29人"
3="3:民営･自営30～499人"
4="4:民営･自営500人以上"
5="5:官公"
6="6:無職"
;
*家族分類 ;
value X11F
1="1:単身世帯"
2="2:夫婦のみ世帯"
3="3:二世代世帯"
4="4:二世代(ひとり親)世帯"
5="5:三世代世帯"
6="6:その他の世帯"
;
*未就学児の有無 ;
value X12F
1="1:無"
2="2:有"
;
*学校に通う世帯員の有無 ;
value X13F
1="1:無"
2="2:有"
;
*65歳以上の世帯員数 ;
value X14F
0="0:０人"
1="1:１人"
2="2:２人以上"
;
run;
```

## ⑦ kitei2017.sas

規定課題の問1～問4用のプログラム

```
/* kitei2017.sas */ *★set_environment.sasを実行済みか確認せよ; options nocenter;
options MPRINT;

data kitei; set newZ.NZ2004; *作業用データセット「kitei」を作成;
  rename X02=HHkubun X08=sex  X09=HHage  X11=HHkind ;*使う変数を間違わないように;
        *世帯区分  世帯主性別 世帯主年齢  世帯分類;
run;

*★規定課題1★; *世帯主の年齢階層(X09)×世帯区分(X02);

proc freq data=kitei; tables HHage*HHkubun / norow nocol nopercent format=comma7.0; label;
  format HHage   X09F.;
  format HHkubun X02F.;
  title "★表1左半分：集計乗率なし"; run;

proc freq data=kitei; tables HHage*HHkubun / norow nocol nopercent format=comma7.0; label;
  format HHage   X09F.;
  format HHkubun X02F.;
  title "★表1右半分：集計乗率あり"; weight weight100000; *★10万比乗率使用; run;

*★規定課題2★; * 世帯分類(x11)×世帯区分(X02)×世帯主の性別(X08);

title "①表2：整形前";
proc tabulate data=kitei format=comma7.0;
  class HHkind HHkubun sex;
  var cnt / weight=weight100000; *★10万比乗率使用;
  table (HHkind ALL),
        (HHkubun ALL)*(sex ALL)*cnt; *★defaultでSUMが自動指定される;
  format HHkind  X11F.;
  format HHkubun X02F.;
  format sex     X08F.;
run;

title "★①表2：整形後";
proc tabulate data=kitei format=comma7.0;
  class HHkind HHkubun sex;
  var cnt / weight=weight100000; *★10万比乗率使用;
  table (HHkind ALL="計"),
        (HHkubun="" ALL="計")*(sex="" ALL="計")*cnt=""*SUM="";
  format HHkind  X11F.;
  format HHkubun X02F.;
  format sex     X08F.;
run;

title "★①表3 weight100000";
data TB3 (keep=Y044 log10Y044 weight100000); set kitei; log10Y044=log10(Y044+1); run;

proc univariate data=TB3;
  var Y044 log10Y044; *パン支出金額;
  weight weight100000; *四分位点だけ指定できる★？平均値だけ指定できる？？？;
  label    Y044="パン支出金額(円)";
  label log10Y044="1を加えた金額の常用対数(円)"; *format指定できる？？？小数点第3位まで;
run;

title "★設問３：箱ひげ図";
proc sgplot data=TB3;
  HBOX Y044 / weight=weight100000 BOXWIDTH=0.2 meanattrs=(color=black symbol=plus);
  label Y044="パン支出";
run;

title "★設問３：ヒストグラム";
data TB3; set TB3; log10Y044=log10(sum(Y044,1)); run;
proc sgplot data=TB3;
  histogram log10Y044 / weight=weight100000 ; run;
quit;
```

（次頁に続く）

```
title "★設問４：表4 weight100000";
%macro ratio(No); if Y042 NE 0 then r&No=Y&No/Y042*100; else r&No=.; ; %mend;
                                                        *★↑0値は欠損値に設定★;
data TB4(keep=Y042-Y046 weight weight100000 HHkubun HHkind diff r042-r046 cnt);
  set kitei;
  diff=Y042-sum(of Y043-Y046); label diff="穀類合計との差"; *データ確認,計算には不要;
  %ratio(042) %ratio(043) %ratio(044) %ratio(045) %ratio(046)
run;

proc print data=TB4 (obs=20); title "data=TB4 (obs=20)"; run;

title "★設問４：①表4：整形前";
proc tabulate data=TB4;
  var r042-r046 cnt / weight=weight100000;
  class HHkubun HHkind;
  table (HHkubun ALL) (HHkind ALL),
        cnt (r043-r046 r042)*MEAN;
run;

title "★設問４：①★表4：整形後";
proc tabulate data=TB4 out=table4;
  var r042-r046 cnt / weight=weight100000;
  class HHkubun HHkind;
  table (HHkubun ALL) (HHkind ALL),
        cnt="N"*SUM=""*F=comma7.0
       (r043="米%"*F=5.1
        r044="パン%"*F=5.1
        r045="めん類%"*F=5.1
        r046="他の麺%"*F=5.1
        r042="全体"*F=5.1)*MEAN="";
  keylabel ALL="全体";
  format HHkind  X11F.;
  format HHkubun X02F.;
run;

proc print data=table4 (obs=50); title "data=table4(1)"; run;

data table4; set table4; if HHkind NE .; run;

proc print data=table4; title "data=table4 "; run;

proc transpose data=table4 out=transTB4;
  var r043_Mean r044_Mean r045_Mean r046_Mean;
  by HHkind;
run;

proc print data=transTB4; title "data=transTB4 転置後"; run;

data transTB4; set transTB4; by HHkind;
  No+1; if No=5 then No=1; output;
run;

proc print data=transTB4; title "data=transTB4 変数No付加"; run;

proc format; *主食;
value SHUSHOKU
1="1:米"
2="2:パン"
3="3:めん類"
4="4:他の麺:"
; run;

title "★設問４：穀類の構成比の帯グラフ";
proc sgplot data=TRANSTB4;
   hbar HHkind / group=No response=COL1 stat=sum seglabel seglabelfitpolicy=NOCLIP;
   format No SHUSHOKU.;
   format COL1 F5.1;
   label COL1="穀類合計(100%)";
run;
quit;

*★設問５：ジニ係数とロレンツ曲線は別プログラムで処理する;
```

## ⑧ make_weighted_datasetX.sas

集計乗率を考慮してオブザベーションを膨らませ、かつ、必要な変数だけから成るデータセットを作成する。

```
/* make_weighted_dataset.sas */ options nocenter;
*Don't forget to run "set_environment.sas" for the 2nd macro specification below;
*Specify the following four macro variables;
*(1)dataset name;          %let dataset= NZ2004;
*(2)path for (1)dataset;  %let libname= &drive:&path; *Leave this as it is;
*(3)target variable;      %let var_name= Y001; *年間収入(単位:万円);
*(4)weight variable;      %let weight= weight;

libname gini "&libname";

proc summary data=gini.&dataset;
   output out=minmax(keep=_stat_ &weight); var &weight; run;
proc print data=minmax; title "min and max of &weight(1)"; run;
proc transpose data=minmax out=xminmax(keep=COL2 COL3 rename=(COL2=min COL3=max)); run;
proc print data=xminmax; title "min and max of &weight(2)"; run;

data gini;
  merge gini.&dataset xminmax;
  retain xmin;
  if _n_=1 then xmin=min;
  obs_no=int(&weight/xmin*10); *★膨らませるオブザベーションの数を計算;
  total_obs+obs_no;
run;

proc print data=gini(obs=100); title "gini(1)"; run;

data _NULL_; /*data total_obs;*/
  set gini end=final;
    if final then put total_obs=; *オブザベーション数の確認;
run;

data reconstruct;
  keep x11 &var_name repeat; *家族分類別、年間収入(単位:万円)、複製個数;
  set gini;
  do repeat=1 to obs_no; output; end;
run;

proc print data=reconstruct(obs=100);
title "data=reconstruct(obs=100)"; run;
```

## ⑨ gini_lorenz2017 規定課題_sas_forum_macro.sas

ジニ係数を計算し、ロレンツ曲線を作図する。

```
/* gini_lorenz2017規定課題_sas_forum_macro.sas */ options nocenter;
*-----------------------------------------------------------------------------;
*★"make_weighted_dataset.sas"の実行直後に、同じセッションで実行する;
*データセット名と変数名を指定して下さい。;
*SAS dataset名; %let dataset=&ds;
/**libname      ; %let libname=&drive:&path;*/
*変数名       ; %let var_name=Y001;
*-----------------------------------------------------------------------------;
*●アウトプット画面にgini係数だけを表示したい場合は、
   次のmacro変数commentに「*」を指定して下さい。
 ●計算経過も表示したい場合は半角ブランクを指定して下さい。;
%let comment= *;
*-----------------------------------------------------------------------------;
/*libname gini "&libname";*/

data original;
  keep &var_name;
  set work.&dataset;
  rename &var_name=income;
run;

/* Calculate gine coefficient */
proc sort data=original; by income; run;

&comment.proc print; title "(1)original"; &comment.run;

data income(keep=income cumm_income)
     cumm_income(keep=cumm_income rename=(cumm_income=total_income));
  set original end=last;

  if _n_=1 then do; save_income=income;
                         income=0; cumm_income=0;
                         output income;
                         income=save_income;
          end; /* Add "income=0" on top of dataset  "INCOME" */

  cumm_income+income;
  output income;
  if last then output cumm_income;
run;

&comment.proc print data=income;      title "(2)income"; &comment.run;
&comment.proc print data=cumm_income; title "(3)cumm_income"; &comment.run;

data standard_cumm_income;
  drop income;
  merge income cumm_income;
run;

&comment.proc print data=standard_cumm_income; title "(4)standard_cumm_income(1)";
&comment.run;

data standard_cumm_income;
  keep standard_cumm_income;
  set;
  retain xtotal_income;

  if _n_=1 then xtotal_income= total_income;
  standard_cumm_income=cumm_income/xtotal_income;
run;
```

注:周防(2008)の付録3のプログラムを使用した。

(次頁に続く)

```
&comment.proc print; title "(4)standard_cumm_income(2)"; &comment.run;

data trapezium;
  keep shorter standard_cumm_income; rename standard_cumm_income=longer;
  set;
  retain shorter 0;
  output;
  shorter=standard_cumm_income; /*To let "longer" to be "shorter" for next obs.*/
run;

&comment.proc print; title "(5)trapezium"; &comment.run;

proc datasets library=work; contents data=original out=no_of_obs(keep=nobs) noprint;
run; /* To count the number of observations in the dataset "original".*/

&comment.proc print data=no_of_obs; title "(6)no_of_obs"; &comment.run;

data gini;
  *keep shorter longer total_no under_bow total_under_bow bow gini;
  keep gini;
  merge  trapezium no_of_obs end=last;
  retain total_no;
  if _n_=1 then total_no=nobs;

  under_bow=(shorter+longer)*(1/total_no)/2;
  total_under_bow+under_bow;

  if last then do; bow=0.5-total_under_bow;
                        gini=bow*2;
                        output; * Make this OUTPUT valid, if you want only gini;
               end;
run;

proc print data=gini;
  title "(7)gini: variable 'gini' is the gini coefficient.";
  title2 "家族分類:&No";
  format gini F4.2;
run;

/* Draw Lorentz Curve */

data lorenz;
  keep standard_cumm_income diagonal n;
  retain total_no;
  rename standard_cumm_income=&var_name; /*income -> &var_name*/

  merge standard_cumm_income no_of_obs;

  if _n_=1 then total_no=nobs;
  n=(_n_-1)/total_no;
  diagonal=n; /* To draw a diagonal line */
run;

&comment.proc print; title "(8)lorenz"; &comment.run;

goptions reset=all htext=0.5cm vsize=8cm hsize=8cm;

symbol i=spline f=xswiss height=1;  /* i=join or spline*/

axis1 length=4.5cm label=(height=0.3cm f=simplex) value=(height=0.3cm f=simplex) offset=(0,0);
axis2 length=4.5cm label=(height=0.3cm f=simplex) value=(height=0.3cm f=simplex) offset=(0,0);

ods listing close;
ods rtf  file="Lorenz&No..rtf";
proc gplot; plot &var_name*n diagonal*n / overlay haxis=axis1 vaxis=axis2;
      title "Lorenz Curve:&No";
run; quit;
ods rtf close;
ods listing;
```

## ⑩ GiniLorenz2017.sas

```
/* GiniLorenz2017.sas */ *★gini係数の計算とLorenz曲線の作図;
*データセット"reconstruct"を家族分類毎のデータセット(D1～D6)に分割する;
%macro divide(No); data d&No ; set reconstruct; if X11=&No; run; %mend;
%macro repeat; %do No=1 %to 6; %divide(&No); %end; %mend;
%repeat;
*★家族分類1～6毎に、gini係数の計算とLorenz曲線の作図;
%macro GiniLorenz(No);
  %let ds=d&No;
  %include "&drive:¥&pathSASprogram¥gini_lorenz2017規定課題_sas_forum_macro.sas";
%mend;
%GiniLorenz(1)
%GiniLorenz(2)
%GiniLorenz(3)
%GiniLorenz(4)
%GiniLorenz(5)
%GiniLorenz(6)
*★全ての家族分類のgini係数の計算とLorenz曲線の作図;
%let No=全体; %let ds=reconstruct;
%include "&drive:¥&pathSASprogram¥gini_lorenz2017規定課題_sas_forum_macro.sas";
```

## ⑪ labelNewZenshoGiji2004.txt

プログラム④createtLabelNewZenshoGiji2004.sasが作成したラベル文。
プログラム⑤setLABELweight.sasの中でラベル文として挿入される。

```
label Year ="調査年";
label X01 ="大都市圏の別";
label X02 ="世帯区分";
label X03 ="世帯人員";
label X04 ="就業人員";
label X05 ="住居の構造";
label X06 ="住居の建て方";
label X07 ="住居の所有関係";
label X08 ="世帯主の性別";
label X09 ="世帯主の年齢";
label X10 ="企業区分・従業者規模";
label X11 ="家族分類";
label X12 ="未就学児の有無";
label X13 ="学校に通う世帯員の有無";
label X14 ="65歳以上の世帯員数";
label Y001 ="年間収入";
label Y002 ="収入総額";
label Y003 ="実収入";
label Y004 ="経常収入";
                中略
label Y013 ="内職収入";
label Y014 ="本業以外の勤め先・事業・内職収入";
label Y015 ="他の経常収入";
label Y016 ="財産収入";
                中略
label Y043 ="米";
label Y044 ="パン";
label Y045 ="めん類";
label Y046 ="他の穀類";
label Y047 ="魚介類";
label Y048 ="生鮮魚介";
                中略
label Y056 ="牛乳";
label Y057 ="乳製品";
label Y058 ="卵";
                中略
label Y132 ="(特掲)ガソリン";
label Y133 ="(特掲)自動車整備費";
label Y134 ="(特掲)自動車保険料";
label Y135 ="通信";
label Y136 ="(特掲)移動電話通信料";
label Y137 ="教育";
label Y138 ="授業料等";
label Y139 ="教科書・学習参考教材";
label Y140 ="補習教育";
label Y141 ="教養娯楽";
label Y142 ="教養娯楽用耐久財";
label Y143 ="教養娯楽用品";
label Y144 ="書籍・他の印刷物";
label Y145 ="教養娯楽サービス";
label Y146 ="宿泊料";
label Y147 ="パック旅行費";
label Y148 ="月謝類";
label Y149 ="他の教養娯楽サービス";
label Y150 ="(特掲)インターネット接続料";
                中略
label Y159 ="交際費";
label Y160 ="交際費(食料)";
label Y161 ="交際費(家具・家事用品)";
label Y162 ="交際費(被服及び履物)";
                中略
label Y189 ="他の非消費支出";
label Y190 ="実支出以外の支出";
label Y191 ="預貯金";
label Y192 ="保険掛金";
label Y193 ="個人・企業年金保険掛金";
label Y194 ="他の保険掛金";
                中略
label Y202 ="その他の実支出以外の支出";
label Y203 ="繰越金";
```

# 付録2　規定問題解答出力結果

## 問1

以下の各図表の上部にあるタイトルは、プログラムで指定したtilte文

★表1左半分：集計乗率なし

| 表：HHage * HHkubun | | | | |
|---|---|---|---|---|
| Hhage (世帯主の年齢) | HHkubun(世帯区分) | | | |
| | 1:勤労者 | 2:勤労者以外 | 3:無職 | 合計 |
| 5:24歳以下 | 765 | 18 | 15 | **798** |
| 6:25～29歳 | 2,289 | 157 | 66 | 2,512 |
| 7:30～34歳 | 4,223 | 511 | 107 | 4,841 |
| 8:35～39歳 | 5,217 | 728 | 193 | 6,138 |
| 9:40～44歳 | 5,492 | 1,076 | 205 | 6,773 |
| 10:45～49歳 | 5,702 | 1,305 | 277 | 7,284 |
| 11:50～54歳 | 6,194 | 1,797 | 392 | 8,383 |
| 12:55～59歳 | 5,682 | 2,221 | 624 | 8,527 |
| 13:60～64歳 | 3,237 | 2,226 | 2,534 | 7,997 |
| 14:65～69歳 | 1,343 | 1,743 | 3,590 | 6,676 |
| 15:70～74歳 | 469 | 1,197 | 3,214 | 4,880 |
| 16:75歳以上 | 166 | 940 | 3,216 | 4,322 |
| 合計 | **40,779** | 13,919 | 14,433 | **69,131** |

★表1右半分：集計乗率あり

| 表：HHage * HHkubun | | | | |
|---|---|---|---|---|
| Hhage (世帯主の年齢) | HHkubun(世帯区分) | | | |
| | 1:勤労者 | 2:勤労者以外 | 3:無職 | 合計 |
| 5:24歳以下 | **1,729** | 10 | 21 | **1,760** |
| 6:25～29歳 | 4,314 | 151 | 57 | 4,522 |
| 7:30～34歳 | 6,441 | 480 | 87 | 7,008 |
| 8:35～39歳 | 7,420 | 824 | 153 | 8,396 |
| 9:40～44歳 | 7,611 | 1,145 | 247 | 9,004 |
| 10:45～49歳 | 7,739 | 1,479 | 244 | 9,462 |
| 11:50～54歳 | 8,072 | 2,085 | 508 | 10,665 |
| 12:55～59歳 | 7,697 | 2,725 | 893 | 11,315 |
| 13:60～64歳 | 4,208 | 2,785 | 4,160 | 11,153 |
| 14:65～69歳 | 1,487 | 2,213 | 6,374 | 10,074 |
| 15:70～74歳 | 459 | 1,652 | 6,124 | 8,235 |
| 16:75歳以上 | 207 | 1,325 | 6,873 | 8,406 |
| 合計 | **57,386** | 16,874 | 25,741 | **100,000** |

## 問2

★①表2：整形後

| | 1:勤労者 | | | 2:勤労者以外 | | | 3:無職 | | | 計 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1:男 | 2:女 | 計 | 1:男 | 2:女 | 計 | 1:男 | 2:女 | 計 | 1:男 | 2:女 | 計 |
| 家族分類 | | | | | | | | | | | | |
| 1:単身世帯 | 6,503 | 4,207 | 10,710 | 1,030 | 867 | 1,897 | 2,463 | 5,688 | 8,151 | 9,995 | 10,763 | 20,758 |
| 2:夫婦のみ世帯 | 8,517 | 324 | 8,841 | 4,293 | 81 | 4,375 | 11,026 | 45 | 11,071 | 23,837 | 450 | 24,287 |
| 3:二世代世帯 | 27,267 | 378 | 27,645 | 6,190 | 51 | 6,242 | 3,264 | 32 | 3,295 | 36,721 | 461 | 37,182 |
| 4:二世代(ひとり親)世帯 | 1,543 | 2,421 | 3,963 | 934 | 430 | 1,364 | 671 | 1,303 | 1,974 | 3,148 | 4,153 | 7,302 |
| 5:三世代世帯 | 4,956 | 384 | 5,340 | 2,390 | 118 | 2,508 | 638 | 160 | 798 | 7,984 | 662 | 8,646 |
| 6:その他の世帯 | 654 | 234 | 887 | 413 | 75 | 488 | 278 | 174 | 452 | 1,344 | 482 | 1,826 |
| 計 | 49,439 | 7,947 | 57,386 | 15,250 | 1,623 | 16,874 | 18,340 | 7,401 | 25,741 | 83,029 | 16,971 | 100,000 |

## 問3

★①表3 weight100000

UNIVARIATE プロシジャ

変数：Y044 (パン支出金額(円))

重み変数：weight100000

白抜きの文字の箇所が解答に対応する。

| モーメント (重み付き) | | | |
|---|---|---|---|
| N | 69131 | 重み変数の合計 | 100000 |
| 平均 | 2284.02283 | 合計 | 228402283 |
| 標準偏差 | 2631.01273 | 分散 | 6922227.97 |
| 歪度 | 4.25528372 | 尖度 | 78.8350941 |
| 無修正 | 1.00E+12 | 修正済平方和 | 4.79E+11 |
| 変動係数 | 115.19205 | 平均の標準誤差 | 8.31999277 |

| 基本統計量 (重み付き) | | | |
|---|---|---|---|
| 位置 | | ばらつき | |
| 平均 | 2284.023 | 標準偏差 | 2631 |
| 中央値 | 1689 | 分散 | 6922228 |
| 最頻値 | 0 | 範囲 | 51776 |
| | | 四分位範囲 | 2181 |

| 分位点 (重み付き) | |
|---|---|
| レベル | 分位点 |
| 100% 最大値 | 51776 |
| 99% | 10300 |
| 95% | 6275 |
| 90% | 4835 |
| 75% Q3 | 3039 |
| 50% 中央値 | 1689 |
| 25% Q1 | 858 |
| 10% | 390 |
| 5% | 202 |
| 1% | 0 |
| 0% 最小値 | 0 |

## ★①表3 weight100000

UNIVARIATE プロシジャ

変数：log10Y044 (1を加えた金額の常用対数(円))

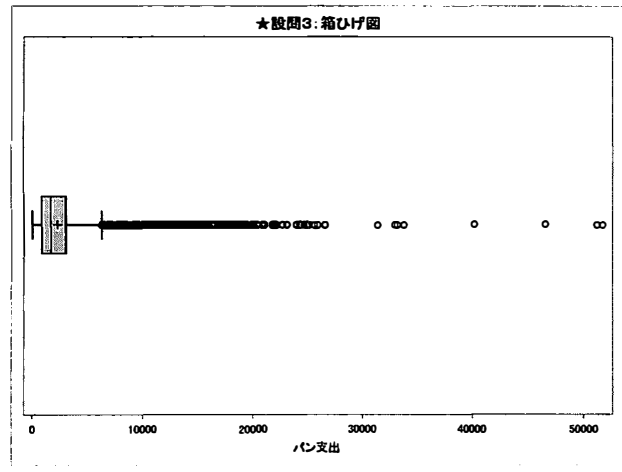
重み変数：weight100000

| モーメント (重み付き) | | | |
|---|---|---|---|
| N | 69131 | 重み変数の合計 | 100000 |
| 平均 | 3.13665089 | 合計 | 313665.089 |
| 標準偏差 | 0.7185186 | 分散 | 0.51626898 |
| 歪度 | -4.6970247 | 尖度 | 48.7932748 |
| 無修正平方和 | 1019547.56 | 修正済平方和 | 35689.6749 |
| 変動係数 | 22.9071908 | 平均の標準誤差 | 0.00227216 |

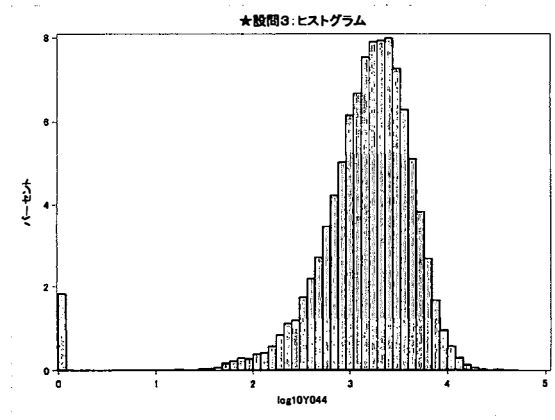
| 分位点 (重み付き) | |
|---|---|
| レベル | 分位点 |
| 100% 最大値 | 4.71414 |
| 99% | 4.01288 |
| 95% | 3.79768 |
| 90% | 3.68449 |
| 75% Q3 | 3.48287 |
| 50% 中央値 | 3.22789 |
| 25% Q1 | 2.93399 |
| 10% | 2.59218 |
| 5% | 2.3075 |
| 1% | 0 |
| 0% 最小値 | 0 |

| 基本統計量 (重み付き) | | | |
|---|---|---|---|
| 位置 | | ばらつき | |
| 平均 | 3.136651 | 標準偏差 | 0.71852 |
| 中央値 | 3.227887 | 分散 | 0.51627 |
| 最頻値 | 0 | 範囲 | 4.71414 |
| | | 四分位範囲 | 0.54888 |

問3 ②




問3 ③

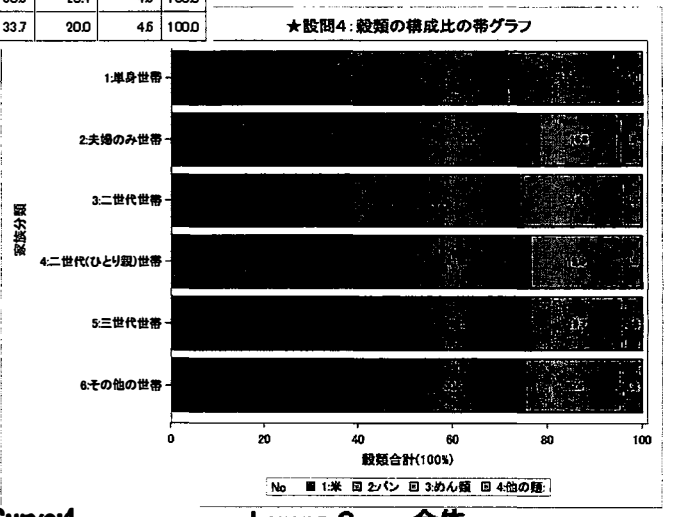


## 問4 ①

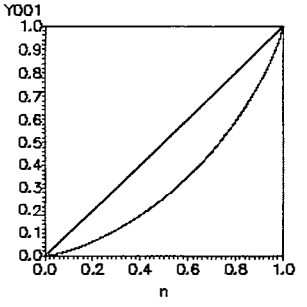
★設問4:①★表4:整形後

| | N | 米% | パン% | めん類% | 他の穀% | 全体 |
|---|---|---|---|---|---|---|
| 世帯区分 | | | | | | |
| 1勤労者 | 57,386 | 36.0 | 37.5 | 22.1 | 4.3 | 100.0 |
| 2勤労者以外 | 16,874 | 44.9 | 31.0 | 19.4 | 4.7 | 100.0 |
| 3無職 | 25,741 | 52.1 | 26.7 | 15.8 | 5.4 | 100.0 |
| 全体 | 100,000 | 41.5 | 33.7 | 20.0 | 4.5 | 100.0 |
| 家族分類 | | | | | | |
| 1単身世帯 | 20,758 | 33.8 | 37.7 | 23.0 | 5.5 | 100.0 |
| 2夫婦のみ世帯 | 24,287 | 49.5 | 28.9 | 16.8 | 4.8 | 100.0 |
| 3二世代世帯 | 37,182 | 39.0 | 35.9 | 20.8 | 4.3 | 100.0 |
| 4二世代(ひとり親)世帯 | 7,302 | 44.9 | 31.5 | 19.0 | 4.5 | 100.0 |
| 5三世代世帯 | 8,646 | 46.0 | 30.5 | 19.7 | 3.8 | 100.0 |
| 6その他の世帯 | 1,826 | 44.8 | 30.5 | 20.4 | 4.3 | 100.0 |
| 全体 | 100,000 | 41.5 | 33.7 | 20.0 | 4.5 | 100.0 |

## 問4 ②

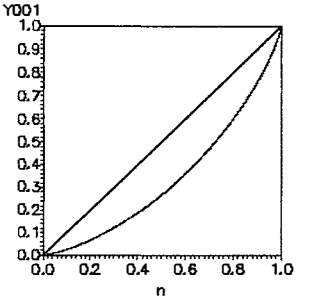



## 問5 ① Lorenz Curve:4




## Lorenz Curve:全体

参考




## 問

(7)gini: variable 'gini' is the gini coefficient.

家族分類:1

| OBS | gini |
|---|---|
| 1 | 0.34 |

家族分類:2

| OBS | gini |
|---|---|
| 1 | 0.31 |

家族分類:3

| OBS | gini |
|---|---|
| 1 | 0.26 |

家族分類:4

| OBS | gini |
|---|---|
| 1 | 0.35 |

家族分類:5

| OBS | gini |
|---|---|
| 1 | 0.27 |

家族分類:6

| OBS | gini |
|---|---|
| 1 | 0.33 |

家族分類:全体

| OBS | gini |
|---|---|
| 1 | 0.33 |

# 付録3 2017 年 第5回 SAS ミクロデータ分析コンテスト 規定課題

**1) 年齢階層別・世帯区分別 クロス表**

世帯主の年齢階層別・世帯区分別に集計用乗率なしのクロス集計表と、集計用乗率を使った 10 万世帯比のクロス集計表を左右に並べて、表 1 の様式で作成する。

表1

| 世帯主の年齢 | 世帯区分 勤労者 | 勤労者以外 | 無職 | 合計 | 世帯主の年齢 | 世帯区分 勤労者 | 勤労者以外 | 無職 | 合計 |
|---|---|---|---|---|---|---|---|---|---|
| 24歳以下 | 765 | | | 798 | 24歳以下 | 1,729 | | | 1,760 |
| 25～29歳 | | | | | 25～29歳 | | | | |
| 30～34歳 | | | | | 30～34歳 | | | | |
| 30～35歳 | | | | | 30～35歳 | | | | |
| 40～44歳 | | | | | 40～44歳 | | | | |
| 45～49歳 | | | | | 45～49歳 | | | | |
| 50～54歳 | | | | | 50～54歳 | | | | |
| 55～59歳 | | | | | 55～59歳 | | | | |
| 60～64歳 | | | | | 60～64歳 | | | | |
| 65～69歳 | | | | | 65～69歳 | | | | |
| 70～74歳 | | | | | 70～74歳 | | | | |
| 75歳以上 | | | | | 75歳以上 | | | | |
| 合計 | 40,779 | | | 69,131 | 合計 | 57,386 | | | 100,000 |

**2) 世帯分類別・世帯区分別 世帯主の性別 3 重クロス表**

世帯分類別・世帯区分別・世帯主の性別の 3 重クロス集計表を、集計用乗率を使って 10 万世帯比で表 2 の様式で作成する。

表 2

| | 勤労者世帯 | | | 勤労者以外の世帯 | | | 無職世帯 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 世帯分類 | 男 | 女 | 計 | 男 | 女 | 計 | 男 | 女 | 計 | 合計 |
| 単身世帯 | 6,503 | | | | | | | | | 20,758 |
| 夫婦のみ | | | | | | | | | | |
| 二世代 | | | | | | | | | | |
| 二世代(ひとり親) | | | | | | | | | | |
| 三世代 | | | | | | | | | | |
| その他 | | | | | | | | | | |
| 合計 | 49,439 | | | | | | | | | 100,000 |

**3) パン支出金額についての各種の統計グラフ**

① パン支出金額(変数名:Y044)及び、パン支出金額に 1 を加えた金額の常用対数変換 log10(Y044+1) を行い、集計用乗率を使って 10 万世帯比について、それぞれの四分位統計量および平均値を表 3 に示す様式で作成する。

② パン支出金額について箱ひげ図を作成する。

③ パン支出金額に 1 を加えた金額の常用対数変換値について、ヒストグラムを作成する。

表 3

| 統計量 | パンの支出金額(円) | 1を加えた金額の常用対数 |
|---|---|---|
| 最大値 | 51,776 | 4.714 |
| 第3四分位数 | | |
| 中央値 | | |
| 第1四分位数 | | |
| 最小値 | | |
| 平均値 | | |
| 件数 | 100,000 | 100,000 |

**4) 世帯区分別および世帯分類別の穀類(米、パン、めん類、他の穀類)の支出金額についての構成比**

「穀類」の支出金額(変数名:Y042)は、米(Y043)、パン(Y044)、めん類(Y045)、他の穀類(Y046)の合計金額である。

① これら 4 変数が穀類(Y042)に占める構成比(パーセント)を世帯ごとに計算した後、10 万世帯比による世帯区分別、及び、世帯分類別に構成比の平均値を、表 4 に示す様式で作成する。

表4

| | | | 構成比 (%) | | | | |
|---|---|---|---|---|---|---|---|
| 項目名 | 符号内容 | 件数 | 米 | パン | めん類 | 他の穀類 | 全体(%) |
| 世帯区分 | 勤労者世帯 | 57,386 | 36.0 | | | | 100.0 |
| | 勤労者以外の世帯 | | | | | | |
| | 無職世帯 | | | | | | |
| 世帯分類 | 単身世帯 | 20,758 | 33.8 | | | | 100.0 |
| | 夫婦のみ | | | | | | |
| | 二世代 | | | | | | |
| | 二世代(ひとり親) | | | | | | |
| | 三世代 | | | | | | |
| | その他 | | | | | | |
| | 全体 | 100,000 | 41.7 | | | | 100.0 |

② 世帯分類別に、穀類(米、パン、めん類、他の穀類)の構成比を帯グラフで表示する。

**5) 世帯分類別のローレンツ曲線およびジニ係数**

① 集計用乗率を考慮して、世帯分類「二世代(ひとり親)」世帯の年間収入(Y001)のローレンツ曲線を描く。

② 集計用乗率を考慮して、6つの世帯分類の年間収入(Y001)のジニ係数(小数点第2位まで)を計算して、表 5 の様式で表示する。

表 5

| 世帯分類 | ジニ係数 |
|---|---|
| 単身世帯 | |
| 夫婦のみ | |
| 二世代 | |
| 二世代(ひとり親) | |
| 三世代 | |
| その他 | |
| 全体 | 0. 33 |

# Let's データ分析コンテストに用いる新擬似ミクロデータの概要

○高橋 行雄[1], 周防 節雄[2], 宮内 亨[3]
[1]BioStat 研究所(株), [2]兵庫県立大学, [3](独)統計センター

Overview of new pseudo-micro data used in Let's data analysis contest

[1]Yukio Takahashi, [2]Setsuo Suoh, [3]Toru Miyauchi
[1]BioStat Research Co.,Ltd., [2]The Univ. of Hyogo, [2]National Statistics Center

**要旨：** (独)統計センターから提供されていた教育用擬似ミクロデータを用いて SAS ユーザー総会で「Let's データ分析コンテスト」を過去４回開催してきたが，2016 年度末で提供打切りとなった．そこで，2004 年全国消費実態調査の匿名データを用いて，ユーザー会世話人有志が新擬似ミクロデータを新規作成し，2017 年のコンテスト用に供した．新擬似ミクロデータを作成する際には，「匿名データから導いた統計表のみから作成する」ことが(独)統計センターから課せられた必須条件だった．そこで，世帯属性について 14 項目を厳選し，14,246 セルからなる多次元クロス表と集計乗率，収支に関する 203 項目のセル毎の対数平均値と標準偏差，年間収入 3 階級別の主要 21 項目間の対数相関行列を作成して Excel 形式でウェブ上に公開した．この公開情報だけを用いて新擬似ミクロデータを作成した．作成に際しての留意点は以下の通り．①年間収入 3 階級別の 69,131 世帯分の 21 次元正規乱数の作成．②主要 21 項目の 14,246 セル分の対数平均値と標準偏差に対して 21 次元正規乱数を適用し 69,131 世帯分のデータの作成．③主要 21 項目以外の収支 182 項目の 14,246 セルの対数平均値と標準偏差に正規乱数の適用し 69,131 世帯分のデータを作成．④収支金額が 0 円の世帯の割合を保つために一様乱数の使用．⑤下位の収支項目の合計がその上位の項目の収支金額となる様に「足し上げ構造」の保持．⑥正規乱数の適用に際し過剰な発散を防ぐための制約条件の設定．⑦作成された新擬似ミクロデータの収支項目の足し上げ構造の検証．⑧元の匿名データの各種の統計量と比較し，元の性質がどの程度保持されているかを検証．

**キーワード：** 全国消費実態調査，擬似ミクロデータ，匿名データ，SAS，JMP，デー分析コンテスト

## 1. はじめに

(独)統計センターから提供されていた教育用擬似ミクロデータを用いて SAS ユーザー総会で「Let's データ分析コンテスト」を過去４回開催してきたが，2016 年度末で提供打切りとなった．これに代わる擬似ミクロデータとして，平成 21 年全国消費実態調査(全消)による一般用ミクロデータが提供されているが，その内容は，教育用擬似ミクロデータの 197 項目に対して，20 項目と 1/10 と大幅に減少し，世帯属性では，例えば，世帯人員属性は「2 人」と「3 人以上」の 2 区分しかなく，利用しずらい．一般用ミクロデータを用いたデータ分析コンテストの実施可能性を検討してみたが，消費支出に関する項目が 10 大費項目に限定されており，公募方式のコンテストに用い

るミクロデータには適さないと判断した.

これまでの教育用擬似ミクロデータを用いた SAS/JMP データ分析コンテストには多数の応募があり，これがきっかけとなって匿名データを用いた実証研究を行う SAS/JMP ユーザーも徐々に増えてきた．統計センターから詳細な一般用ミクロデータの提供が早急に提供される見込みはなく，「Let's データ分析コンテスト」の継続が極めて難しい状況なった.

コンテスト実施の共同オーガナイザである周防が，統計センターの関係者に，匿名データを用いた擬似ミクロデータの作成の可否について問い合わせたところ賛同が得られたので, 2016 年から高橋・周防が使用許可を得ている全消の匿名データの使用目的に，「匿名データを用いた新擬似ミクロデータの作成のための統計表の作成」も追加して申請を行なった.

新擬似ミクロデータ作成は，匿名データ (47,797 世帯分) から導いた複数の統計表のみから作成することと，匿名化の観点から多次元クロス表でのセル度数が 3 以上となることが必須だったので，セル度数 1 及びセル度数 2 に該当する世帯のデータにノイズを入れたデータを加え 69,131 世帯分のデータとし，

① 世帯に関する情報 14 項目と集計乗率

② 14 次元クロス表のセル毎に収支に関する 203 項目の対数変換した平均値と標準偏差

③ 年間収入 3 階級別の主要 21 項目間の相関行列

を作成して，ウェブ上に公開した（高橋・周防）．この情報から新擬似ミクロデータを作成し，SAS データセットと CSV 形式の両方でウェブ上に公開した[1]．詳細手順については、紙幅の関係上、別の機会に詳解する．なお，今回の作業はすべて JMP で行ったが，SAS による作成も近いうちに開始したい.

## 2. 年間収入は対数正規分布に従うのか

匿名データの集計乗率を用いて，10 万世帯当たりの集計乗率を計算し，それを「度数」として使うと年間収入の平均と標準偏差は (609.7, 388.7) となり，変動係数は，63.8%と極めて大きい．その平均と標準偏差の正規分布に従う正規乱数を用いて，47,797 世帯分のデータを擬似的に生成すれば，平均と標準偏差は匿名データとほとんど同じになる．ただし，平均と標準偏差の関係から，実際にはありえない「年間収入がマイナス」の世帯が多数発生してしまう.

年間収入は，高いほう (右に裾を引く) 対数正規分布に従うと仮定できるのか．図 1a は，匿名データに対して正規分布をあてはめた結果で，左側に大きく裾を引いていることが観察される．図 1b は，正規乱数データに対する正規分布のあてはめでヒストグラム上にきれいに乗っている.

図 1a の匿名データは，右側のひげを超える点が全くない．これは，2500 万円を超える年間収入は 2500 万円とするトップコーディングが 268 世帯 (0.56%) に施されているためであり，99.5%分位点と最大値が 3.398 (2500 万円) と同じであることからも確認できる．図 1a の要約統計量の (平均 =2.698, 標準偏差=0.291) を用いた正規乱数で生成したデータでは，図 1b に示すようにきれいに左右対称となり，99.5%点は，3.469 (2944 万円) と若干大きくなる．2.5%点では log10 (年間収入) で 2.064 (116 万円)，乱数データで 2.134 (134 万円) と食い違いが生じている.

---

[1] SAS データセット ⇒ http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroData.zip
CSV 形式 ⇒ http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroDataCSV.zip

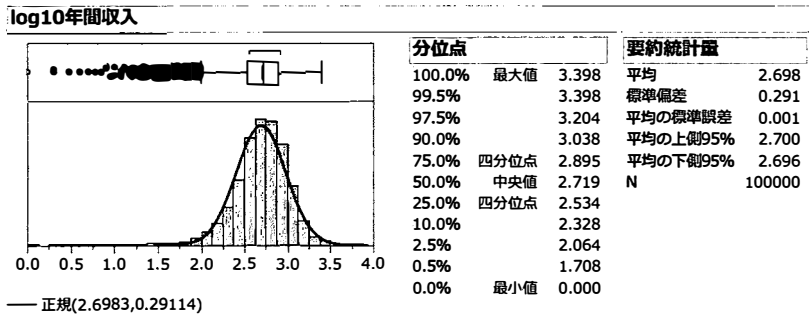



図 1a　常用対数変換した年間収入の分布の特徴

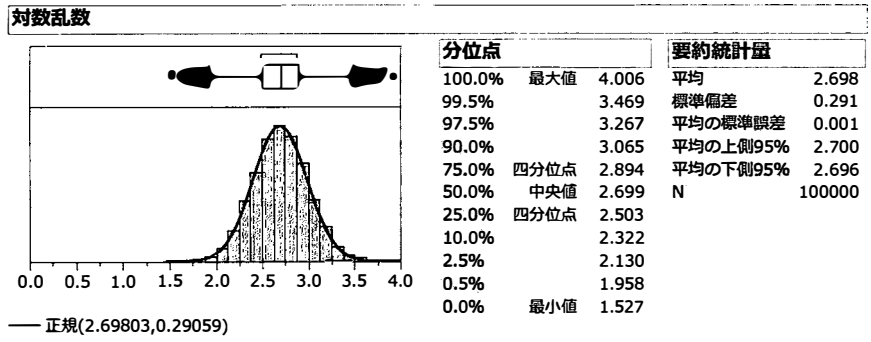



図 1b　正規乱数データ(平均=2.698, 標準偏差=0.291)の分布

データ分析コンテストでは，年間収入について各種の統計解析が行なわれることも想定されるので，このようなわずかな分布のゆがみも再現することが望ましいと考える．図 1a ヒストグラムから年間収入の分布は，視覚的にほぼ対数正規分布に従っていると見なせるかもしれないが，図 2 に示す正規分位点プロットを見れば，年間収入が低くなるにつれて，斜めの直線からの乖離が大きくなり，対数正規分布に従っているとみなすことは躊躇せざるを得ない．

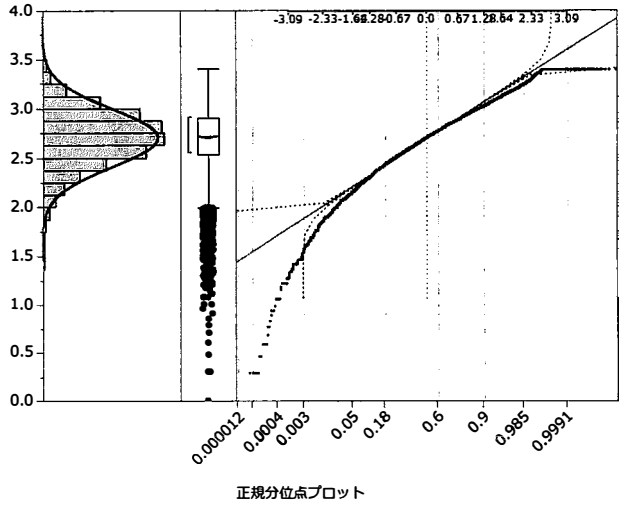



図 2 年間収入(対数)についての正規確率プロット

このような場合には，複数の母集団が内在していると仮定し，そこからランダムに 47,797 世帯がサンプリングされたとし，各母集団の(混合割合，平均，標準偏差)を最尤法によって推定することが可能である．

母集団が 2 つあると仮定した場合と，3 つあると仮定した場合の統計量を図 3 に示す．母集団が 2 つと仮定した場合に，「2 重正規分布のあてはめ」欄の割合 $\pi 1$ が 13.4%の小集団で，平均値は位置 $\mu 1$ =2.388(244 万円)であり，割合 $\pi 2$ が 86.6%を占める大集団で，平均値は 2.746(557 万円)である．ヒストグラム上の確率密度関数のあてはめも母集団が 1 つと仮定した場合に比べ改善している．さらに母集団が 3 つと仮定した場合には，「3 重正規分布のあてはめ」欄の割合(1.7%，8.0%，90.3%)の集団が分離されている．

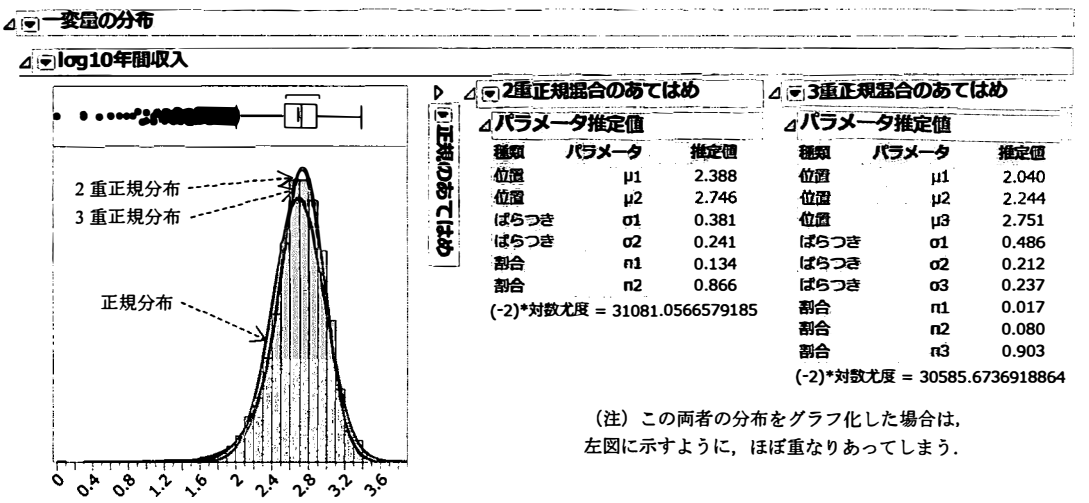



図 3 混合正規分布のあてはめ

統計的にどちらのあてはめが望ましいのかは，マイナス 2 倍の対数尤度を用いて評価する．表 1 に示すように，母集団=1 の場合は，(-2)*対数尤度=36995.97 で，母集団=2 の場合は，(-2)*対数尤度=31081.06 となり，5914.91 の減少となる．母集団=3 の場合は 30585.67 で，母集団=2 の場合に比べ 495.39 小さくなる，母集団=4 の場合は，母集団=3 の場合に比べ 45.82 小さくなる．母集団=5 の

表 1 仮定する母集団の数による対数尤度の差および最小母集団の統計量

| | | | 最小の母集団 | | |
|---|---|---|---|---|---|
| 母集団数 | (-2)対数尤度 | 対数尤度の差 | 構成比 | 対数母平均 | 母平均(万) |
| 1 | 36995.97 | - | 100.0% | 2.698 | 498.9 |
| 2 | 31081.06 | -5914.91 | 13.4% | 2.388 | 244.3 |
| 3 | 30585.67 | -495.39 | 1.7% | 2.040 | 109.6 |
| 4 | 30539.85 | -45.82 | 0.021% | 0.513 | 3.3 |
| 5 | 30539.84 | -0.01 | 0.025% | 0.587 | 3.9 |
| | $\chi^2(df=2, 5\%)$ | 5.99 | | | |

場合には変化が 0.01 である．統計的には，$\chi^2$=5.99(自由度 2，上側 5%点)以下ならば，母集団数の増加に意味がないので，母集団=4 が統計的に支持される．しかしながら，母集団=4 の場合，最小母集団の構成比は，0.021%と極めて小さい集団なので，母集団=3 とするのが現実的と思われる．その場合，年間収入の母平均は，それぞれ 110 万円，175 万円，563 万円となる．

年間収入以外の 202 項目についてもヒストグラムの視覚的な検討結果から，単一の対数正規分布とはみなすことができないと判断した．それ故に，世帯属性毎に，14 次元クロス表の 14,246 セルに対して 203 項目の(データ数，対数平均，標準偏差)の統計表を作成・公表した．

次節以降で，公表した統計表から新擬似ミクロデータを作成した過程を示す．

## 3. 正規乱数を用いた復元

公表した(14,246×3 レコード×収支 203 項目)の統計表から，(69,131 レコード×収支 203 項目のデータの作成が最初のステップである．表 2 に，統計表の一部を示す．

表 2　収支項目に関する統計表（抜粋）

| 14次元番号 | ラベル | Y001 | Y002 | Y003 | Y004 | Y005 | (中略) | Y038 | Y039 | Y040 | Y041 | (中略) | Y203 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Z01 年間収入 | Z02 収入総額 | Z03 実収入 | 経常収入 | 勤め先収入 | (収入項目) | Z06 支出総額 | Z07 実支出 | Z08 消費支出 | Z09 食料 | (支出項目) | Z21 繰越金 |
| 1 | 件数* | 3 | 3 | 3 | 3 | 3 | | 3 | 3 | 3 | 3 | | 3 |
| 1 | log10平均 | 2.192 | 5.226 | 5.112 | 5.119 | 5.104 | | 5.224 | 4.764 | 4.756 | 3.910 | | 4.719 |
| 1 | log10_SD | 0.005 | 0.015 | 0.010 | 0.034 | 0.009 | | 0.023 | 0.011 | 0.016 | 0.008 | | 0.023 |
| 2 | 件数 | 3 | 3 | 3 | 3 | 3 | | 3 | 3 | 3 | 3 | | 3 |
| 2 | log10平均 | 2.430 | 5.615 | 5.338 | 5.341 | 5.325 | | 5.627 | 5.303 | 5.207 | 4.557 | | 4.014 |
| 2 | log10_SD | 0.006 | 0.009 | 0.010 | 0.012 | 0.013 | | 0.010 | 0.010 | 0.014 | 0.011 | | 0.019 |
| | (中略) | | | | | | | | | | | | |
| 14246 | 件数 | 3 | 3 | 3 | 3 | 3 | | 3 | 3 | 3 | 3 | | 3 |
| 14246 | log10平均 | 3.112 | 6.16 | 5.921 | 5.912 | 5.771 | | 6.162 | 5.751 | 5.635 | 5.076 | | 5.052 |
| 14246 | log10_SD | 0.011 | 0.007 | 0.014 | 0.013 | 0.002 | | 0.011 | 0.013 | 0.027 | 0.004 | | 0.005 |

* 件数＝それぞれの収支項目の金額が0円であるケースを除いた出現頻度である．

① 14 次元表の 14,246 レコードのセル度数を用いて，該当するレコードをセル度数分複製し，順次縦方向に連結する．最終的には，セル度数の合計 69,131 レコードを持つファイルを作成する．

② 公表した統計表を転置して(14,246×203)×3 のファイルを作成する．

③ ①と②のファイルに対し，セル番号でマッチマージをして②の情報を①に付与し，(69,131×203)×3 のファイルを作成する．

④ ③のファイルに正規乱数を追加して，この正規乱数を使って「対数平均＋標準偏差×正規乱数」を計算する．ただし，標準偏差は±2 の範囲内に制限し，これに 0.5 を掛けて変動が±1 となるように制限を加えた．

⑤ ④のファイルを世帯毎に転置し，69,131×203 の擬似データのファイルを作成する．

⑥ 一様乱数を用いて欠測値処理をする．(これについては，第 6 節を参照)

## 4. 互いに相関を持つ主要 21 項目

主要 21 項目について，年間収入 3 階級別に 21×21 の相関係数行列を公表している．互いに相関を持つ正規乱数の作成は，相関係数行列をコレスキー分解した行列と正規乱数行列の積で求められ

る．手順を可視化するために JMP のスクリプトを用いて例示する．

まず，3 項目の 3×3 の相関係数行列を roh とする．Cholesky 関数でコレスキー分解を行なうと下三角行列 chol が得られる．

**roh(相関行列)**

| | | |
|---|---|---|
| 1 | 0.8 | 0.5 |
| 0.8 | 1 | 0.7 |
| 0.5 | 0.7 | 1 |

➔ `chol=Cholesky(roh);` ➔

**chol(コレスキー分解)**

| | | |
|---|---|---|
| 1 | 0 | 0 |
| 0.8 | 0.6 | 0 |
| 0.5 | 0.5 | 0.707 |

次に 3×5 の正規乱数行列 ysnorm を Random Normal 関数で生成する．

`ysnorm=J(3,5,Random Normal());` ➔

**ysnorm(正規乱数・相関 0)**

| | | | | |
|---|---|---|---|---|
| 0.274 | -0.224 | 0.219 | -0.420 | 0.246 |
| 0.908 | 1.485 | 1.663 | -1.682 | -0.761 |
| 0.629 | 1.432 | 0.360 | -0.023 | 0.467 |

下三角行列 chol と正規乱数行列 ysnorm の積を求め，互いに相関を持つ行列を計算し，転置(JMP のスクリプトでは「`」が転置記号)して．5×3 の 3 次元正規乱数行列 ymnorm を求める．

`ymnorm=(chol*ysnorm)` ; ➔

**ymnorm**

| | | |
|---|---|---|
| **0.274** | **0.764** | **1.036** |
| -0.224 | 0.712 | 1.643 |
| 0.219 | 1.173 | 1.196 |
| -0.420 | -1.345 | -1.067 |
| 0.246 | -0.260 | 0.073 |

最初に， chol の 1 行目 [1.0　0　0] と ysnorm の 1 列目 [0.274　0.908　0.629]$^T$ の積和が ysnorm の 1 行 1 列目に 0.272 と計算され，元の正規乱数のままとなる．次に，chol の 2 行目と ysnorm の 1 列目の積和なので，

$$0.8\times0.274+0.6\times0.908=0.764$$

ysnorm の 1 行 2 列目に 0.764 が得られる．さらに，chol の 3 行目 [0.5　0.5　0.707] に ysnorm の 1 列目 [0.274　0.908　0.629]$^T$ の積和によって，1.036 が得られる．

(相関を加味した正規乱数)

| | |
|---|---|
| **0.274** | **=1　x0.274 + 0　x0.908 + 0　x0.629** |
| **0.764** | **=0.8x0.274 + 0.6x0.908 + 0　x0.629** |
| **1.036** | **=0.5x0.274 + 0.5x0.908 + 0.0707x0.629** |

この操作を，ysnorm の 2 行目から 5 行目まで繰り返し，5×3 の 3 次元正規乱数行列 ymnorm が計算されている．実際に 21 次元の相関係数行列を 69,131 世帯分作る JMP スクリプトを次ページに示す．

このスクリプトを年齢 3 階級毎に，公表した相関係数行列をスクリプトに挿入し，69,131 行 21 列の互いに相関を持つ正規乱数行列を計算し，この行列をファイルに出力する．表 3 に第 1 階級のファイルの一部を示す．

```
//  相関を持つ正規乱数
nrows=69131 ;
seed=1234561 ;
Random Reset(seed);
roh=
[   1.000  0.462  ... 0.133,
    0.462  1.000  ... 0.217,
       :      :          :
    0.133  0.217  ...1.000]
;
ysnorm=J(Ncol(roh),Nrows, Random Normal());
chol=Cholesky(roh);
ymnorm=(chol*ysnorm)`;
dt = New TAble("ran") << Set Matrix(ymnorm);
```

表 3　下位 1/3 階級用の 69,131 行 21 列の互いに相関を持つ正規乱数

GIJI_年収区分_1低_21ran - JMP

ファイル(F)　編集(E)　テーブル(T)　行(R)　列(C)　実験計画 (DOE)(D)　分析(A)　グラフ(G)　ツール(O)　表示(V)　ウィンドウ(W)　ヘルプ(H)

GIJI_年収区分…
列(21/0)
行
すべての行　69,131
選択されている行　0

| | 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 | 列10 | 列11 | 列12 | 列13 | 列14 | 列15 | 列16 | 列17 | 列18 | 列19 | 列20 | 列21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.23 | 0.26 | 1.29 | -0.02 | -0.46 | 0.04 | 1.82 | 1.82 | -0.53 | 1.67 | -0.93 | -0.35 | 0.31 | 0.02 | 1.23 | 1.19 | 0.77 | 1.14 | 1.36 | -0.46 | -1.02 |
| 2 | -0.29 | 0.47 | 1.07 | 0.92 | -1.17 | 0.73 | 1.08 | 1.06 | 0.13 | 1.08 | 0.08 | 1.16 | 2.06 | 0.88 | 1.76 | 0.03 | 0.22 | 1.01 | -0.23 | 0.16 | -1.04 |
| 3 | -0.37 | 0.67 | 0.33 | 1.38 | 1.84 | 1.04 | 0.21 | 0.29 | 0.30 | -0.69 | -1.40 | 0.29 | -1.42 | 1.52 | 0.10 | -1.52 | 0.21 | 1.69 | -0.99 | -0.04 | 1.92 |
| 4 | -1.22 | -0.28 | -0.16 | -0.99 | -0.46 | -0.63 | 0.58 | 0.83 | 0.16 | -0.98 | -0.32 | -0.10 | 0.69 | -0.08 | 0.23 | 0.07 | -0.29 | 2.62 | -1.86 | -1.67 | 0.75 |
| 5 | 0.60 | 0.90 | 0.27 | 0.52 | 0.16 | 1.01 | 1.41 | 1.15 | 1.11 | 0.20 | -0.14 | 1.18 | 2.39 | 0.62 | 0.60 | -0.54 | 1.60 | 1.52 | 1.32 | 0.54 | -0.62 |

主要 21 項目に対し多次元正規乱数による誤差変動の与え方を以下に示す.

① 年間収入 3 階級の(69,131 世帯)×21 項目の多変量正規乱数ファイル毎に世帯毎に転置し(69,131×21)×1 とする.

② 年間収入 3 階級毎に転置されたデータを列方向に併合し, (69,131×21)×3 の乱数ファイルとする.

③ 世帯毎の 3 階級の符号に対応した乱数列を用いる.

④ 主要 21 項目は, 203 項目の一部なので 203 項目のファイルに上書きする.

## 5. 住居費および教育費など支出金額が 0 円となる世帯が多数ある場合

匿名データ 47,797 世帯を, 集計乗率を考慮して 10 万世帯とした場合に, 住居費が支出されている世帯は, 66,113 世帯あり, 支出金額が 0 円の世帯が 1/3 程度含まれている. 図 4a に示すように, 分布は左側に大きく裾を引いていて 3 重混合対数正規分布のあてはめが支持される. 母集団の月当たり住居費の母平均は,「3 重正規混合のあてはめ」欄の $\mu 1$, $\mu 2$, $\mu 3$ を用いて, それぞれ $10^{2.547}$=352, $10^{4.113}$=12,972, $10^{4.736}$=54,450 円/月と推定される.

図 4b に示すように, 教育費が支出されている世帯は, 10 万世帯中 28,540 世帯あり, 欠測値が 3/4 程度含まれている. 分布は左右ともに裾を引いていて, 3 重混合対数正規分布のあてはめが支持さ

れる．母集団の平均値は，住居費と同様に，それぞれ $10^{3.887}$=7,709，$10^{4.440}$=27,542，$10^{5.078}$=119,674 円/月と推定される．

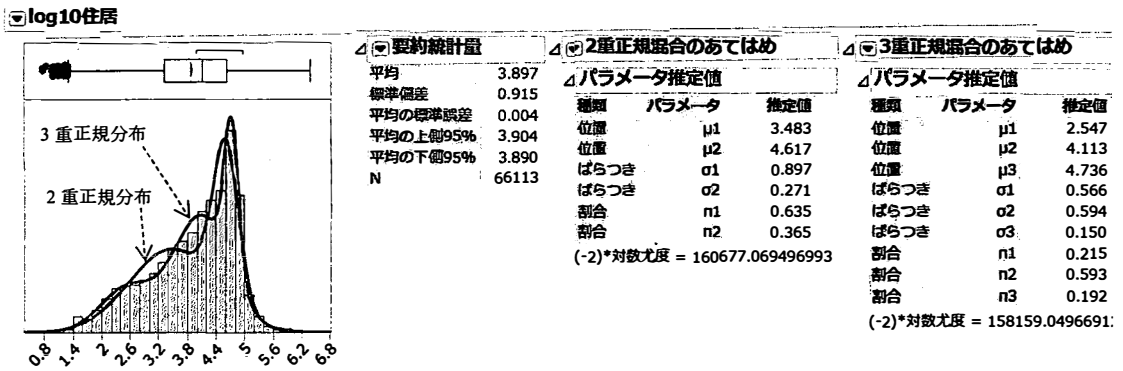



図 4a　匿名データ 47,797 世帯の 10 万世帯での対数住居費の分布

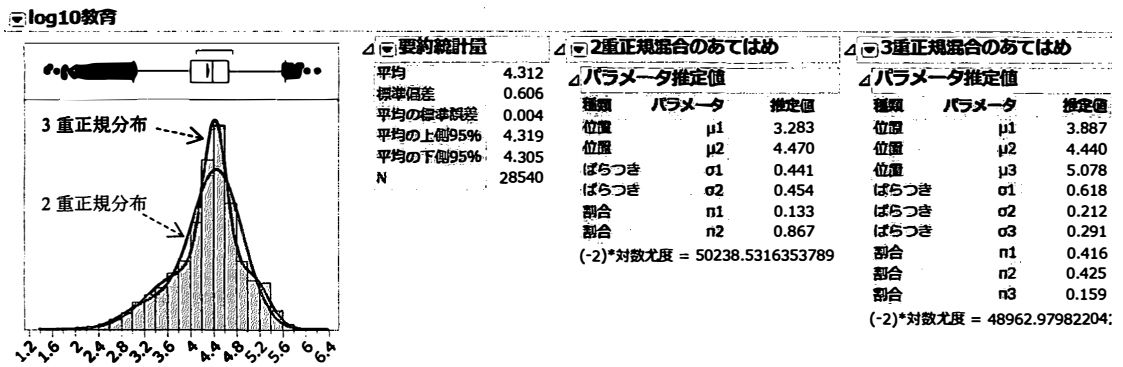



図 4b　匿名データ 47,797 世帯の 10 万世帯での対数教育費の分布

住居費には，家賃地代や修繕費などが含まれているが，持家でその年に修繕費が発生しなければ支出金額は 0 円となり，教育費は，対象になる家族がいなければ 0 円である．公表した統計量の算出に際して，常用対数変換を施したので 0 円は欠測値となり，(69,131 世帯×203 項目)×3 統計量のファイルに付加された世帯毎・項目毎の欠測値を含むデータ数を $N_i$ ，$i$=1,2,...,14246 セルとし，欠測値以外のデータ数を $n_{ij}$ ，$j$=1,2,..., $N_i$ ，$n_{i1} = n_{i2} = \cdots = n_{iN_i}$ として，欠測値の割合を保つ擬似データを，以下の基準で生成する．

① データ数が $n_{ij}$=0 の場合，対数平均値は欠測値なので，$N_i$ 個の擬似データ全てを欠測値．
② データ数が $n_{ij}$=1 の場合には，$N_i$ 個中，1 個は擬似データがあり，$N_i$ - 1 個は欠測値．
③ 0～1 の一様乱数を $u_{ij}$ とし，$u_{ij}$ が $n_{ij}/N_i$ の比より小さければ擬似データがあり，大きければ欠測値．
④ 表 3 に示すように $N_i$ が 5 で，データ数 $n_{ij}$ が 3 の場合，対数平均と標準偏差は，5 世帯が全て同じなので 5 世帯中 2 世帯は欠測値．
⑤ 一様乱数 $u_{ij}$ を発生させて，3/5=0.60 以下ならば擬似データがあり，0.60 以上ならば欠測値(表 4)．

表 4.　一様乱数による教育費が 0 円の世帯の割合の確保

| 世帯番号 | 14次元番号 | レコード数 $N_i$ | 繰返し $j$ | 項目名 | データ数 $n_{ij}$ | 対数平均 | 対数SD | 一様乱数 | $n_{ij}/N_i$ | 判定 | 扱い |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 9999 | 8888 | 5 | 1 | 教育費 | 3 | 4.00 | 0.10 | 0.70 | 0.60 | × | 欠測値 |
| 9999 | 8888 | 5 | 2 | 教育費 | 3 | 4.00 | 0.10 | 0.20 | 0.60 | ○ | 採用 |
| 9999 | 8888 | 5 | 3 | 教育費 | 3 | 4.00 | 0.10 | 0.80 | 0.60 | × | 欠測値 |
| 9999 | 8888 | 5 | 4 | 教育費 | 3 | 4.00 | 0.10 | 0.39 | 0.60 | ○ | 採用 |
| 9999 | 8888 | 5 | 5 | 教育費 | 3 | 4.00 | 0.10 | 0.50 | 0.60 | ○ | 採用 |

実際に作成された新擬似ミクロデータ 69,131 世帯の住居費と教育費について，10 万世帯での分布をそれぞれ図 5a と図 5b に示す．住居費が支出されている世帯数は，図 4a と比較して，66,113 世帯から 61,832 世帯になり，新擬似ミクロデータで 3 重正規混合分布の形状も保持されている．

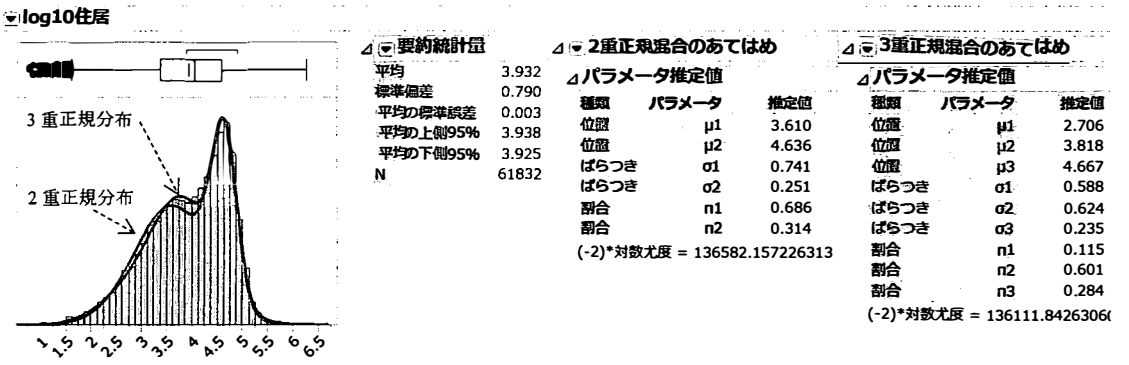



図 5a　新擬似ミクロデータの 10 万世帯での対数住居費

教育費が支出されている世帯数は，図 4b と比較すると 28,540 世帯から 29,066 世帯になり，ほぼ 0 円世帯の割合は保たれている．また，分布も 3 重正規混合分布の形状がやや滑らかになるが，(-2) 対数尤度の大きな減少もあり，新擬似ミクロデータで保持されている．

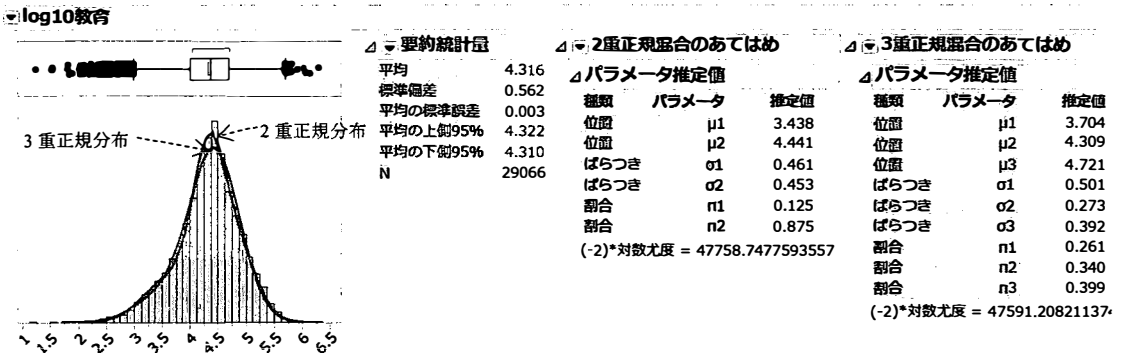



図 5b　新擬似ミクロデータの 10 万世帯での対数教育費

## 6. 足し上げ構造

表 4 に示す「消費支出」は，その下の項目「食料，住居，...，その他の消費支出」の 10 大消費項目の金額の合計が一致するように統制されており，これを「足し上げ構造」と呼ぶことにする．

この様な階層構造が他にも多くあり，表 5 に，主要 21 項目間の足し上げ構造について，「区分 $i$」と「足し上 $i$」を用いて表す．例えば「区分 2」では，21 と 22 の 2 種類の足し上げ構造があり，それぞれ，「足し上 2」の(0, 1)で親子関係が示されている．同じ区分内の親(0)は，子(1)の合計金額と一致する．「区分 3」の 33 では，「実支出」を親とし，子が「消費支出」と「非消費支出」の 2 変数から成る．この表に示した表記法は，新擬似ミクロデータに含まれる足し上げ構造を考慮して計算する際に使えるにように，今回考案した．

表 5　主要 21 項目足し上げ構造

| | | | レベル2 | | レベル3 | | レベル4 | | 世帯1 | | | 世帯2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Z | V | 区分 2 | 足上 2 | 区分 3 | 足上 3 | 区分 4 | 足上 4 | データ | 足し上げ | | データ | 足し上げ | |
| | | V0003_調査年 | | | | | | | 2004 | | | 2004 | | |
| 階層 | | V0004_No | | | | | | | 1 | | | 2 | | |
| 2 | Z01 | V0399_年間収入 | | | | | | | 76 | | | 738 | | |
| 2 | Z02 | V0400_収入総額 | 21 | 0 | | | | | 353,255 | 353,255 | o | 1,282,209 | 1,282,209 | o |
| 3 | Z03 | V0401_実収入 | 21 | 1 | | | | | 309,924 | | | 354,483 | | |
| 3 | Z04 | V0439_実収入以外の収入 | 21 | 1 | | | | | 3,715 | | | 672,663 | | |
| 3 | Z05 | V0452_繰入金 | 21 | 1 | | | | | 39,615 | | | 255,062 | | |
| 2 | Z06 | V0453_支出総額 | 22 | 0 | | | | | 353,255 | 353,255 | o | 1,282,209 | 1,282,209 | o |
| 3 | Z07 | V0454_実支出 | 22 | 1 | 33 | 0 | | | 256,381 | 256,381 | o | 485,670 | 485,670 | o |
| 4 | Z08 | V0455_消費支出 | | | 33 | 1 | 45 | 0 | 256,381 | 256,381 | o | 470,916 | 470,916 | o |
| 5 | Z09 | V0456_食料 | | | | | 45 | 1 | 68,774 | | | 90,732 | | |
| 5 | Z10 | V0498_住居 | | | | | 45 | 1 | 52,138 | | | 16,840 | | |
| 5 | Z11 | V0504_光熱・水道 | | | | | 45 | 1 | 17,125 | | | 22,710 | | |
| 5 | Z12 | V0509_家具・家事用品 | | | | | 45 | 1 | 11,387 | | | 12,172 | | |
| 5 | Z13 | V0519_被服及び履物 | | | | | 45 | 1 | 12,265 | | | 3,208 | | |
| 5 | Z14 | V0537_保健医療 | | | | | 45 | 1 | 2,132 | | | 17,123 | | |
| 5 | Z15 | V0542_交通・通信 | | | | | 45 | 1 | 43,016 | | | 12,012 | | |
| 5 | Z16 | V0553_教育 | | | | | 45 | 1 | 0 | | | 0 | | |
| 5 | Z17 | V0557_教養娯楽 | | | | | 45 | 1 | 12,614 | | | 16,505 | | |
| 5 | Z18 | V0567_その他の消費支出 | | | | | 45 | 1 | 36,930 | | | 279,615 | | |
| 4 | Z19 | V0598_非消費支出 | | | 33 | 1 | | | 0 | | | 14,754 | | |
| 3 | Z20 | V0609_実支出以外の支出 | 22 | 1 | | | | | 32,848 | | | 485,652 | | |
| 3 | Z21 | V0622_繰越金 | 22 | 1 | | | | | 64,026 | | | 310,887 | | |

表 6 に，消費支出の細目の足し上げ構造を示す．「穀類」は，「米，パン，めん類，他の穀類」の合計となっていることが確認できる．

足し上げ構造は，下位の区分から上位の区分に対して定義されているが，正規乱数を用いて変動を与えているので，足し上げ構造に若干の揺れが生じている．下位から上位への足し上げは，誤差変動の伝播により上位の項目の変動が大きくなりすぎる傾向がある．そこで，上位を先に固定してから下位方向に調整する「逆」足し上げ計算を行う．

子の金額を $z_{i1j}$（$i$ は各レベルの区分番号，1 は足し上げの子のコード，$j$ は項目番号），親の金額を $z_{i0}$（$i$ は各レベルの区分番号，0 は足し上げの親のコード），$N$ 個の子の項目の金額の合計を $z_{i1\cdot}$ とした時に，「逆足し上げ」により調整された子(項目番号 $j$)の金額 $z'_{i1j}$ を，

$$z'_{i1j} = \left(\frac{z_{i1j}}{z_{i1\cdot}}\right) z_{i0}$$

で定義する．こうすれば，

$$z_{i0} = \sum_{j=1}^{N} z'_{i1j}$$

表 6　食料における足し上げ構造（抜粋）

| | 階層 | 新擬似 Y名 | 匿名 V名 | レベル2 区分 | レベル2 足上 | レベル3 区分 | レベル3 足上 | レベル4 区分 | レベル4 足上 | レベル5 区分 | レベル5 足上 | レベル6 区分 | レベル6 足上 | 世帯1 データ | 世帯1 足し上げ | | 世帯2 データ | 世帯2 足し上げ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 39 | 3 | Y039 | V0454_実支出 | 22 | 1 | 33 | 0 | | | | | | | 256,381 | 256,381 | o | 485,670 | | |
| 40 | 4 | Y040 | V0455_消費支出 | | | 33 | 1 | 45 | 0 | | | | | 256,381 | 256,381 | o | 470,916 | 470,916 | o |
| 41 | 5 | Y041 | V0456_食料 | | | | | 45 | 1 | 503 | 0 | | | 68,774 | 68,774 | o | 90,732 | 90,732 | o |
| 42 | 6 | Y042 | V0457_穀類 | | | | | | | 503 | 1 | 602 | 0 | 11,957 | 11,957 | o | 10,362 | 10,362 | o |
| 43 | 7 | Y043 | V0458_米 | | | | | | | | | 602 | 1 | 6,359 | | | 4,627 | | |
| 44 | 7 | Y044 | V0459_パン | | | | | | | | | 602 | 1 | 3,719 | | | 3,209 | | |
| 45 | 7 | Y045 | V0460_めん類 | | | | | | | | | 602 | 1 | 1,712 | | | 1,962 | | |
| 46 | 7 | Y046 | V0461_他の穀類 | | | | | | | | | 602 | 1 | 168 | | | 565 | | |
| 47 | 6 | Y047 | V0462_魚介類 | | | | | | | 503 | 1 | 603 | 0 | 1,197 | 1,197 | o | 16,122 | 16,122 | o |
| 48 | 7 | Y048 | V0463_生鮮魚介 | | | | | | | | | 603 | 1 | 1,197 | | | 10,718 | | |
| 49 | 7 | Y049 | V0464_塩干魚介 | | | | | | | | | 603 | 1 | 0 | | | 2,504 | | |
| 50 | 7 | Y050 | V0465_魚肉練製品 | | | | | | | | | 603 | 1 | 0 | | | 1,142 | | |
| 51 | 7 | Y051 | V0466_他の魚介加工品 | | | | | | | | | 603 | 1 | 0 | | | 1,758 | | |

となり，子の項目の合計が，親の金額と一致する．実際の計算手順は以下の通り．

① 69,131 世帯毎に 203 項目を転置したファイルを作成する．

② 表 6 に示した階層データ構造を表現している「レベル 2 の区分と足し上,...，レベル 6 の区分と足し上」を変数名で①とマッチマージし（69,131 世帯×203 項目）レコード×「収支金額＋5 レベル×2（区分・足し上）」のファイルを作成する．

③ ②のファイルに対して「世帯×区分 $i$×足し上げ (0,1)」別に金額の合計を計算し，世帯毎に $z_{i0}$ と $z_{i1\cdot}$ を得る．

④ ②と③のファイルを，世帯番号，区分 $i$ でマッチマージすることによって，①のファイルに $z_{i1\cdot}$ を付加して，$z'_{i1j}$ を計算する．

⑤ すべての区分に対して手順③を繰り返す．

なお，一様乱数を用いて 0 円データを修正すると，足し上げ構造の親子関係に不具合が発生するケースがある．すなわち，親が欠測値なのに，子に値がある，逆に親に値があるが，子はすべて欠測値の場合が生ずる．このような場合には，逆足し上げ計算はできないので，親・子の一方が欠測値の場合には，事後処理として，両方とも 0 円に置き換える．

## 7. 新擬似ミクロデータと匿名データとの照合および足し上げ構造の検証

新擬似ミクロデータと匿名データとの照合するために，世帯属性の 14 項目の符号について構成比を算出し比較した結果，±5%程前後の相違があった．収支の 203 項目の対数平均の総平均は，匿名データで 3.811，新擬似ミクロデータで 3.765 と 1.2%小さかった．元の金額にすると 10.0%の差となる．元の金額で 50%を超える差異が 5 項目にあったが，ほぼ匿名データの特質は再現できたと思われる．

新擬似ミクロデータの 203 項目の金額について，すべて四捨五入により整数化を行ったので，足し上げ構造に若干のくるいが生じている．全足し上げ 3,387,419 箇所中，最大で±4 円は 13 箇所，±1 円差が 663,196 箇所 19.6%あった．今回は，急いで新擬似ミクロデータを提供しなければならない状況であったので，「足し上げには，四捨五入による不整合がある」との注を入れることに留めて，

以下の URL に新擬似ミクロデータを公開した(高橋・ 周防・ 宮内)将来的には，今の数字を使って，再度，通常の足し上げ計算をして，親の値を書き換えることも検討している．

SAS データセット http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroData.zip
CSV 形式 http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroDataCSV.zip

## 8. 今後の展望

新擬似ミクロデータは，誰でも何時でも自由にダウンロードして使えるが，約 7 万件のデータを用いた統計解析は，誰にでも容易にできるという訳ではない．どの様な解析がいかにしたらできるのかは，無料でダウンロードできる web 上の SAS ユーザー総会の論文集に収録されている「Let's データ分析コンテスト」のこれまでの優秀賞の論文が参考になる．今後，コンテストにチャレンジするユーザーが増えることを願っている．

今年の第 5 回目コンテストの実施に間に合うように，新擬似ミクロデータを提供できたことに安堵している．ただ，急造したために，一部にケアレスミスがあったことは，深くお詫びしたい．今回の 2004 年データに加え，1989 年，1994 年，1999 年の全国消費実態調査についても，匿名データを用いて擬似ミクロデータを作成し，来年以降の「Let's データ分析コンテスト」に供したい．


### 謝 辞

全国消費実態調査の匿名データの利用のために，必要な予算上の措置を賜った SAS Institute Japan (株)，ならびに，匿名データの利用に便宜を図って頂いた（独）統計センターのそれぞれの関係各位に深く感謝申し上げます．


### 参 考 文 献


1) 周防節雄(2015) 全国消費実態調査の匿名データとその符号表から自動的に SAS のデータセット，変数ラベルと変数フォーマットを作成する SAS プログラム，『SAS ユーザー総会 2015 論文集』，pp257-278.
2) 高橋行雄(2015) 統計センターの匿名データ 13 万件を用いた統計解析の実践，『SAS ユーザー総会 2015 論文集』，pp145-164.
3) 高橋行雄(2016) JMP による第4回 Let's データ分析の規定課題の解析，『SAS ユーザー総会 2015 論文集』，pp313-329.
4) 周防節雄(2016) SAS ユーザー総会 2016 における「Let's データ分析第4回ミクロデータ分析コンテスト」の規定課題の SAS プログラム解説，『SAS ユーザー総会 2016 論文集』，pp330-347.
5) 高橋行雄(2016) 統計センター提供の教育用擬似ミクロデータを用いた SAS/JMP によるデータ分析コンテスト，https://www.nstac.go.jp/services/pdf/161125_3-3.pdf.

# 金融・経済・経営関連

# 自治体におけるオープンデータ化への対応の現状と課題

有馬　昌宏

（兵庫県立大学応用情報科学研究科）

Current Status and Problems of Open Data Initiatives in Local Governments

Masahiro Arima

Graduate School of Applied Informatics, University of Hyogo

## 要旨

スマートフォンやソーシャルメディアの普及等を背景に，多種多様な情報を相互に連携させて新たな価値を生み出すことが期待され，国や自治体が保有する公共データが国民や企業が利用しやすい形で公開されること，すなわち，「オープンデータ」への取組が求められている．わが国では，2000 年に内閣に設置された高度情報通信ネットワーク社会推進戦略本部が 2011 年７月４日に「電子行政オープンデータ戦略」を策定し，オープンデータへの取り組みを本格化させた．しかし，政府はデータカタログサイト DATA.GO.JP を通じてオープンデータ化に先導的に取り組んではいるものの，自治体のオープンデータ化への対応は遅れており，地域や自治体規模によっても対応に差が生じてきている．そこで，本研究では，自治体におけるオープンデータ化の現状を自治体の公式ウェブサイトの目視調査で把握し，オープンデータ化への取組の違いを自治体の規模や財政力といった属性に加えて，電子自治体への組織的取組との関連で統計的に明らかにすることを試みた．

キーワード：オープンデータ，地方自治体，目視調査，名義ロジスティック回帰分析

## １．研究の背景と研究の目的

### １．１　オープンデータとは

スマートフォンやタブレット端末やソーシャルメディアの普及等を背景に，官民が有する多種多様な情報を相互に連携させて新たな価値を生み出し，経済を活性化するとともに透明性の高い政府を実現する手段の一つである「オープンデータ」が注目を集めている．オープンデータとは，総務省[1]によれば，「機械判読に適したデータ形式で、二次利用が可能な利用ルールで公開されたデータ」であり、「人手を多くかけずにデータの二次利用を可能とするもの」とされている．

オープンデータとして政府や自治体がデータを積極的に公開していく動きは，アメリカにおいて始まったと言える．アメリカ合衆国の第 44 代大統領に 2009 年１月 20 日に就任したバラク・オバマ氏が，就任式翌日

に行った行政機関の長への指令の中で，新政権はこれまでにないレベルのオープンな政府すると宣言したが，この考えはオープンガバメント・イニシアチブと呼ばれ，その具体的な取り組みがオープンデータであるとされている．

## １．２　わが国のオープンデータの取組

上述のアメリカの動きに触発され，日本でもオープンデータの動きが始まった．2011 年 7 月 4 日に，内閣府に設置された高度情報通信ネットワーク社会推進戦略本部（以降，IT 総合戦略本部と略記）は，電子行政オープンデータ戦略を策定している．この戦略によれば，オープンデータの活用を促進する目的は次の３つであるとされている（首相官邸[2]）．

①透明性・信頼性の向上：データの公開による透明性と信頼性の向上．機械判読が可能であり，公開のための手続きが不要であることから情報公開制度を上回る効果が期待できる．

②国民参加・官民協働の推進：官民協働の公共サービスや民間サービスが創出される．

③経済の活性化・行政の効率化：新ビジネスの創出や企業活動の効率化で経済が活性化される．また，公共データを用いて業務の効率化と高度化が図られる．

また，IT 総合戦略本部は 2013 年６月に「電子オープンデータ推進のためのロードマップ」を策定し，「二次利用の促進のための府省のデータ公開に関する基本的考え方」を策定し，そこでオープンデータの具体的な取組内容とその工程が示されている．その具体的取組の一つが政府のデータカタログサイト「DATA.GO.JP」である．DATA.GO.JP では府省の公開しているデータを横断検索し，府省の公開 URL からダウンロードできる機能やオープンデータの活用事例へのリンクなどの機能が設けられており，2017 年２月 22 日時点で 18,582 件のデータセットが公開されている．

地方公共団体においても，オープンデータを推進させるため，2015 年２月に内閣官房情報通信技術（IT）総合戦略室[3]により，「地方公共団体オープンデータ推進ガイドライン」が策定されている．その中で，「先進的にオープンデータに取り組んでいる自治体はあるものの，全体としては一部である」と記載されており，国としては地方自治体の取組み状況の低さを課題視していることがわかる．

## １．３　研究の目的

このような状況を受け，地方自治体でのオープンデータへの取組みの現状を把握し，地方自治体でオープンデータ化が進展しない原因を把握するために，オープンデータの利用者の視点で 1,718 の市町村と 23 の特別区の公式ウェブサイトを目視で調査することとした．

目視調査により，オープンデータに取り組んでいる基礎自治体をリストアップし，オープンデータに取り組んでいる自治体とそうでない自治体とではどのような違いがあるかを，オープンデータとして公開されている自治体のミクロ統計データ，具体的には総務省統計局の「統計でみる市区町村のすがた」で得られる人口規模と行政区分ならびに財政力指数のデータ，および電子自治体への取組状況を示す総務省の「地方自治情報管理概要（地方公共団体における行政情報化の推進状況調査）」のデータを用いて分析することを試みることとした．

電子自治体への取組状況を考慮に入れるのは，オープンデータの推進が，電子自治体推進に向けての直近の指針である 2014 年３月の「電子自治体の取組みを加速するための 10 の指針」（総務省[4]）において７番目の指針として挙げられており，電子自治体の推進体制とオープンデータの推進は密接な関係があると考えたからである．

## ２．先行研究の概要

わが国のオープンデータの推進の歴史は短く，そのために先行研究は多くなく，オープンデータの歴史や意義などの啓蒙的な解説が多くなっている．また，自治体でのオープンデータの推進の取組状況を把握する調査も数少ない．

このような中，総務省は 2013 年と 2014 年に全国地方自治体を対象に「地域における ICT 利活用の現状に関する調査」を実施しているが，オープンデータの取組については，2013 年は回答した 895 団体中で 6.8%が，2014 年は 733 団体中で 9.7%が「既に取組を推進している」と回答している（野村総合研究所[5]）．また，オープンデータの推進に向けての課題としては，「具体的な利用イメージやニーズの明確化」，「提供側の効果・メリットの具体化」，「個人情報等の機微情報の扱いに関する制度的な整理」，「提供に関わる費用や人的負担の軽減」，「データ形式・構造等の標準化，標準的なアプリケーション・プログラミング・インターフェイスの推進等，標準的な利用に必要なシステム面の各種標準化」，「提供情報の内容詳細・費用負担等の調整」，「知的財産等の権利処理の制度的な整理」などがあったとしている．

一方，オープンデータを推進している自治体とその数については，政府のデータカタログサイト DATA.GO.JP，日本のオープンデータ都市マップ（http://fukuno.jig.jp/2013/opendatamap），CityData（http://citydata.jp/），LinkData（http://ja.linkdata.org/home），BODIK（ビッグデータ&オープンデータ研究会 in 九州，http://www.bodik.jp/opendata/），自治体オープンデータ（https://www.open-governmentdata.org/），一般財団法人日本情報経済社会推進協会（オープンデータ・センサス　ステップアップガイド）などのサイトで確認することができるが，カウントの基準や政令市の行政区を含めるかどうかなどの対象自治体の定義の違いなどがあり，正確に把握するのは難しい状況である．

## ３．研究の方法と分析結果

### ３．１　自治体ウェブサイトの目視調査

先行研究のレビューを踏まえて，全国の 1,718 の市町村と 23 の特別区と 47 の都道府県に対して，公式ウェブサイトを訪問してオープンデータが公開されているかどうかの目視調査を行うことで，オープンデータに取り組んでいる自治体のリストを作成することとした．

具体的には，全自治体の公式ウェブサイトを調べた上で，2016 年 12 月から 2017 年 1 月にかけて，

①各自治体の公式ウェブサイトを訪問，

②サイト内検索で「オープンデータ」をキーワードとして検索，

③検索結果で何も表示されない場合はオープンデータに未対応と判定，

④検索結果が表示された場合は，オープンデータの公開サイトやカタログサイトが検索結果として表示されているかどうかを確認して，表示されていなければ未対応と判定，

⑤公開サイトあるいはカタログサイトがあれば，実際にアクセスして公開されているデータ形式を確認して，T. Berners-Lee のオープンデータの展開スキームであり評価指標の一つである「5 Star Open Data」を適用して，２つ星以上であれば対応と判定，１つ星以下は未対応と判定，

という手順で判定を行った．

目視調査の結果として，194 の基礎自治体がオープンデータに取り組んでいると判定され，基礎自治体のオープンデータ化率は 11.1%であることが判明した．都道府県別のオープンデータ化率は，福井県が 100%で

トップであり，42.4%の神奈川県，40.0%の静岡県が続く状況となっている．また，岩手県，山梨県，鳥取県，高知県，長崎県，大分県，宮崎県ではオープンデータ化率は0%となっている．

なお，この目視調査においては，先行研究で示したサイトは一切参考にせず，サイト内検索結果と検索結果のリンク先の確認による方法のみで行った．

## 3．2　オープンデータ化取組みの有無の分析

総務省の「地域における ICT 利活用の現状に関する調査」などの分析結果などによれば，オープンデータの推進の阻害要因としては，人員の不足や人的負担および費用負担の増加が挙げられている．そこで，事務権限の違いに加えて，人口規模が大きいと職員数も多くて人員不足にはなりにくく，財政状況がよければ費用負担に耐えられると考えられることを考慮して，人口規模別行政区分と財政力指数でオープンデータへの取組の有無の違いが説明できるのではないかとの作業仮説を立てた．また，オープンデータの推進が「電子自治体の取組みを加速するための 10 の指針」に挙げられていることから，電子自治体推進に向けての計画策定ならびに体制整備や基盤技術の導入がなされているかどうか，さらには職員の情報化研修に積極的な自治体ではオープンデータも推進されているのではないかとのもう一つの作業仮説も考えた．これらの作業仮説をまとめたものが図 1 である．

この作業仮説を検証するためにはデータが必要になるが，国のオープンデータ化政策推進の恩恵を受けて，人口や財政力指数のデータは e-Stat（http:// www. estat.go.jp/ SG1/estat/ eStatTopPortal.do）で「統計でみる市区町村のすがた」が excel 形式公開されており，電子自治体推進に関連する自治体の状況を調査した地方自治情報管理概要（地方公共団体における行政情報化の推進状況調査結果）の個票データは総務省のサイト（http:// www.soumu.go.jp/denshijiti/060213_02.html）から同じく excel 形式でダウンロード可能となっている．

目視調査結果と政府のサイトからダウンロードした市区町村別の統計データおよび調査の個票データからデータベースを作成し，統計解析ソフトウェアの JMP11.0 を利用して分析を行った結果をまとめたものが表 1 と表 2 である．

表 1 からは，

①人口規模別行政区分および財政力指数の高さに対応してオープンデータの推進が進展していること，

②自治体内部の判断で決定される電子自治体推進計画の策定状況や統合型 GIS の導入状況ともオープンデータの推進が密接に関係していること，

③職員研修もオープンデータの推進と関連があり，情報担当部門であるかどうかを問わず，広く職員に対して情報化に関する人材育成方針を策定している自治体ではオープンデータへの取組が進んでおり，民間企業等へ職員を研修派遣したり情報処理（技術）に関する資格取得を奨励している自治体でもオープンデータの推進が進んでいること，

などが読み取れる．すなわち，オープンデータの推進にあたっては，単に庁内にあるデータを「機械判読に適したデータ形式で、二次利用が可能な利用ルールで公開」していくことで達成されるものではなく，継続的に広範囲にわたってオープンデータを推進していくためには，体制の整備，職員のスキル向上，庁内の部局横断的な連携，情報システムの基盤，トップの強い

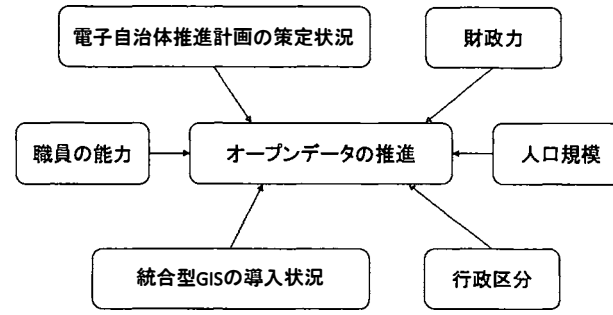



図 1　オープンデータ推進規定要因の分析枠組み

表 1　オープンデータ推進に影響を及ぼす要因およびその水準別にみた推進比率の違い

| 財政力指数 | 推進比率 | 自治体数 |
|---|---|---|
| 1.0.8以上 | 30.0 | 290 |
| 2.0.5以上0.8未満 | 15.1 | 451 |
| 3.0.3以上0.5未満 | 5.6 | 447 |
| 4.0.3未満 | 2.5 | 553 |
| 全市区町村 | 11.1 | 1,741 |

| 電子自治体推進計画 | 推進比率 | 自治体数 |
|---|---|---|
| 1.策定済み | 19.3 | 610 |
| 2.策定予定 | 15.1 | 53 |
| 3.総合計画等に含めて策定 | 7.2 | 611 |
| 4.策定予定なし | 5.1 | 467 |
| 全市区町村 | 11.1 | 1,741 |

| 統合型GIS | 推進比率 | 自治体数 |
|---|---|---|
| 1.導入済み | 14.4 | 854 |
| 2.整備中 | 11.8 | 51 |
| 3.調査中・導入検討中 | 9.7 | 330 |
| 4.導入予定なし | 6.5 | 506 |
| 全市区町村 | 11.1 | 1,741 |

| 人口規模別行政区分 | 推進比率 | 自治体数 |
|---|---|---|
| 1.政令市 | 95.0 | 20 |
| 2.中核市 | 49.5 | 91 |
| 3.特別区 | 43.5 | 23 |
| 4.人口10万人以上の市 | 26.3 | 156 |
| 5.人口5万人以上の市 | 15.9 | 270 |
| 6.人口5万人未満の市 | 7.1 | 253 |
| 7.町 | 2.3 | 745 |
| 8.村 | 0.6 | 183 |
| 全市区町村 | 11.1 | 1,741 |

| 職員研修 | 推進比率 | 自治体数 |
|---|---|---|
| 01.職員に対し情報化研修を実施 | 15.4 | 1,118 |
| 02.ITを活用した業務改善方法について職員研修を実施 | 24.6 | 114 |
| 03.部外の情報化研修に職員を派遣 | 17.1 | 865 |
| 04.情報処理(技術)に関する資格取得を奨励 | 32.4 | 68 |
| 05.業務担当部門と情報担当部門の人事交流を積極的に推進 | 18.2 | 203 |
| 06.民間企業等へ職員を研修派遣 | 33.3 | 21 |
| 07.情報化のために外部の専門人材を採用 | 22.0 | 118 |
| 08.情報担当部門に対しての情報化に関する人材育成方針を策定 | 43.4 | 53 |
| 09.一般職員に対して情報化に関する人材育成方針を策定 | 40.0 | 40 |
| 10.いずれも行っていない | 2.7 | 378 |

表 2　オープンデータ推進に影響を及ぼす要因の名義ロジスティック回帰分析結果

| | | | | |
|---|---|---|---|---|
| 疑似決定係数 | | 0.2716 | 0.1289 | 0.2931 |
| サンプル数 | | 1,741 | 1,741 | 1,741 |
| 切片 | | -5.204 *** | -3.651 *** | -5.946 *** |
| 人口規模 | 政令市(70万人以上) | 8.148 *** | | 6.868 *** |
| | 中核市(20万人以上) | 5.182 *** | | 3.922 *** |
| | 特別区 | 4.942 *** | | 5.683 *** |
| | 人口10万人以上の市 | 4.173 *** | | 2.999 *** |
| | 人口5万人以上の市 | 3.540 *** | | 2.537 ** |
| | 人口5万人未満の市 | 2.635 ** | | 2.106 ** |
| | 町 | 1.447 | | 1.027 |
| 財政力指数 | 0.8以上 | | 2.803 *** | 2.246 *** |
| | 0.5以上0.8未満 | | 1.922 *** | 1.697 *** |
| | 0.3以上0.5未満 | | 0.825 ** | 1.227 ** |

| | | | |
|---|---|---|---|
| 疑似決定係数 | | 0.0656 | 0.1342 |
| サンプル数 | | 1,741 | 1,741 |
| 切片 | | -3.288 *** | -4.227 *** |
| 推進計画 | 策定済み | 1.399 *** | 0.776 *** |
| | 今後策定予定 | 1.115 ** | 0.777 * |
| | 総合計画等に含む | 0.330 | 0.103 |
| 統合GIS | 導入済み | 0.666 *** | 0.357 |
| | 整備中 | 0.449 | 0.066 |
| | 調査・検討中 | 0.227 | 0.021 |
| 職員情報研修 | 01.職員に対し情報化研修を実施 | | 1.109 *** |
| | 02.IT活用の業務改善方法の職員研修実施 | | 0.345 |
| | 03.部外の情報化研修に職員を派遣 | | 0.768 *** |
| | 04.情報処理(技術)の資格取得を奨励 | | 0.447 |
| | 05.業務部門と情報部門の人事交流推進 | | 0.100 |
| | 06.民間企業等へ職員を研修派遣 | | 0.498 |
| | 07.情報化のために外部の専門人材を採用 | | 0.332 |
| | 08.情報部門の情報化人材育成方針を策定 | | 0.707 ** |
| | 09.一般職員の情報化人材育成方針を策定 | | 0.574 |
| | 10.いずれも行っていない | | 0.254 |

リーダーシップなど，組織全体としての一丸となったリーダーシップなど，組織全体としての一丸となった取組が必要となると考えられる．このような観点から考えると，有馬他[6]が提案している M. Hammer[7]の成熟度モデルである PEMM を自治体にも適用してオープンデータの推進，そして最終的には電子自治体の推進を試みることは，有用であるように思われる．

表 2 には，オープンデータの推進に影響を及ぼす諸要因を総合的に評価するために名義ロジスティック回帰分析を適用した結果をまとめて示している．これらの結果から，我々の仮説，すなわち，政令市および中核市とそれ以外の市町村の権限の相違，人口規模および財政力指数の大小，電子自治体推進計画の策定の有無，統合 GIS の導入状況，職員の情報化関連研修の有無がオープンデータの推進に影響を及ぼしているという仮説のすべてが支持されることが明らかとなった．なお，表 2 に要約した名義ロジスティック回帰分析では，被説明変数は（オープンデータに取り組んでいる）／（オープンデータに取り組んでいない）の対数オッズであり，表中の*，**，***はそれぞれ，10%，5%，1%の有意水準で有意であることを示している．

## 4．まとめと今後の課題

本研究では，自治体の公式ウェブサイトの目視調査で得られたオープンデータの推進状況のデータと政府がオープンデータとして公開している統計データおよび調査データを利用して，２つの作業仮説に基づき，オープンデータの推進の背後にある要因および阻害要因を明らかにすることを試みた．結果として，組織としてのケイパビリティが相応に高くなければオープンデータの推進に繋がらないのではないかとの感触を得ることができた．

ところで，本研究を進めている最中の2017年２月16日に，内閣官房情報通信技術（IT）総合戦略室から「オープンデータの取組状況に関するアンケート」の調査結果が公表された．この調査では，全自治体を対象にオープンデータ推進の現状と可能性や課題に関する設問がなされ，全自治体からの回答が得られている．全自治体のオープンデータの取組の状況はpdf形式で公表されているので，本研究での目視調査の結果と比較対照することで，さらなる知見が得られるのではないかと考える．また，筆者が所属する兵庫ニューメディア推進協議会の調査研究グループも2016年12月から１月にかけて全都道府県および全市区町村を対象に「災害ハザード関連空間データのオープンデータ化に関する調査」を実施しており，この調査の個票データを使った分析も残された課題となっている．可能であれば，これらの課題について対応した結果は，SASユーザー総会の本研究の発表時において紹介したい．

## 謝辞

本研究を行うにあたり，兵庫県立大学応用情報科学研究科博士前期課程２年次の鄧時韻氏には全国の自治体のウェブサイトを訪問しての目視調査を実施してもらった．ここに記して感謝します．なお，本論文は，筆者が指導した鄧時韻氏が経営情報学会関西支部第12回学生論文発表会ならびに経営工学会関西支部および経営システム学会関西支部主催の平成28年度学生論文発表会に提出した予稿に加筆修正したものである．

## 参考文献

[1] 総務省，「オープンデータ戦略の推進」，http://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/.

[2] 首相官邸，「電子行政オープンデータ戦略」，http://www.kantei.go.jp/jp/singi/it2/denshigyousei.html.

[3] 内閣官房情報通信技術(IT)総合戦略室，「地方公共団体オープンデータ推進ガイドライン」，2015.

[4] 総務省，「電子自治体の取組みを加速させるための10の指針」，2014.

[5] 野村総合研究所，『地域におけるICT利活用の現状等に関する調査研究報告書』，http://www.soumu.go.jp/johotsusintokei/linkdata/h26_07_houkoku.pdf，2014.

[6] 有馬昌宏・中土真輝・吉崎智信・山下綾子・島田達巳，「電子自治体実現に向けての成熟度モデルの構築の試み」，『経営情報学会2011年秋季全国研究発表大会予稿集』（ウェブ版），2011.

[7] Hammer, M.，“The Process Audit,” Harvard Busi-ness Review, April 2007, pp.111-123（マイケル・ハマー（有賀裕子訳），“PEMMでビジネスプロセスを改革する”，DIAMONDハーバード・ビジネス・レビュー，2007年９月号，pp.28-45，2007).

[8] 鄧時韻，『わが国自治体におけるオープンデータ化の現状と課題』，兵庫県立大学応用情報科学研究科修士論文，2017.

# ID-POS だけで始める新しい RFM 分析: データステップで実装するモンテカルロ法を用いて

片桐 智志[1]

[1]ネイチャーインサイト株式会社

# A New RFM Analysis Only with Scanner Data: Using Monte Carlo Simulation Implemented by a Data Step

KATAGIRI, Satoshi
Nature Insight Co., Ltd.

要旨

RFM 分析とは，最後に購買してからの経過日数(R，リセンシー)，購買頻度(F，フリクエンシー)，購買額(M，マネタリー)の 3 つの指標から，利用顧客のセグメンテーションや優良顧客かどうかの判定をおこなうフレームワークで，マーケティングでよく用いられる．従来の RFM 分析ではしばしば，リセンシーの大きさで顧客の離脱の有無が判断される. [阿部, 2011] は，この方法が理論的根拠に弱いことを指摘し，統計的なモデリングに基づいて, RFM の 3 指標を同時に推定する方法を提案している．この方法は，単純な ID-POS データだけでも実施できるという大きな利点がある．すなわち購買日とその購買金額のみを保存している購買履歴データと，履歴を購買した顧客と一意に紐づけられる ID データさえあれば可能となる．

この新しい RFM 分析の計算はモンテカルロ法によって行われるが，モデルの構造上, MCMC プロシジャで計算をすることが難しい．本稿では Base SAS の基本的な機能であるデータステップを用いて，モンテカルロ法による計算を実装した．データステップの特性上，この方法であれば，高性能なマシンでなくとも比較的高速で計算できることがわかった．

キーワード：RFM 分析 マーケティング POS データ MCMC モンテカルロ法 データステップ

## 初めに

RFM 分析は，最後の購買からの経過日数 (リセンシー, R)，購買頻度 (フリクエンシー, F)，購買額 (マネタリー, M) の 3 要素によって顧客ひとりひとりセグメンテーションを行い，適切な販売戦略を計画するという，マーケティングの技法の一種である．RFM の 3 指標は，ID-POS データや，オンラインショッピングサイトのトランザクションデータが保持する情報から簡単に得られるため，RFM 分析は様々な形態の小売業で幅広く利用できる．特に，会員制でないために顧客がいつ離脱したかわからない「非契約型」購買において，リセンシーは離脱を推測するために重要な指標である．

しかしながら，[阿部, 2004] [阿部, 2008] (Abe, 2009) では， RFM のうちマネタリーの要素を除いた RF 分析を挙げて，この方法が購買を続ける期間の長さという情報を捨てていることを指摘している．かりにある顧客のリセンシーの指標が良くても，顧客にとってそれが最後の購買となり，以後購入することがないかもしれないし，フリクエンシーが悪くても継続的に購入してくれるリピート客がいるかもしれない．そもそも，「離脱」は直接観測できないために，経験的にリセンシーがその代理変数として使われているにすぎない．そこで，前述の研究では，間接的に生存期間を推測できるように，顧客の生存期間がある分布にしたがうと仮定したモデルを提案している．

## 提案されるモデル

[阿部, 2011] は，(Abe, 2009) で提案した RF 分析に M の要素を加え，RFM の 3 指標すべてを同時に扱うモデルを提案した．これは次のような仮定をおいている．

I. 各顧客の離脱までの期間 (生存期間) はパラメータ $\mu$ の指数分布にしたがい，平均は $1/\mu$ である

II. 各顧客の，離脱するまでの期間(生存期間)の購買回数 $\lambda$ はパラメータポアソン分布にしたがう

III. 各顧客の購買額は，パラメータ $\ln\eta, \omega^2$ の対数正規分布にしたがい，平均は $\eta\exp(\omega^2/2)$ である．

IV. 各顧客の上記 3 つのパラメータ $\lambda, \mu, \eta$ の対数は以下のように多変量正規分布にしたがう．ただし，$i$ は顧客インデックスを表す．

$$\boldsymbol{\theta}_{\mathrm{i}} := \begin{bmatrix} \ln\lambda_i \\ \ln\mu_i \\ \ln\eta_i \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\theta} := \begin{bmatrix} \theta_\lambda \\ \theta_\mu \\ \theta_\eta \end{bmatrix}$$

以上の仮定は，「各顧客」とあるように，顧客ごとに異なる個別パラメータであり，$\mu, \lambda, \eta$ がそれぞれ R, F, M に対応している．これらを計算することで個人レベルの特徴を得ることができる．さらに，仮定 IV により，これらの個別パラメータは多変量正規分布の平均 $\boldsymbol{\theta}$ と分散 $\boldsymbol{\Sigma}$ という共通パラメータによって決まる．つまりこれは階層ベイズモデルである．[阿部, 2011] は，回帰係数を共通パラメータとして，顧客の年齢や性別といったデモグラフィック属性を個別パラメータで回帰させる，多変量回帰モデルの適用も提案している．しかし，本稿では，利用できるデータの制約から，多変量回帰モデルへと拡張せず，$\boldsymbol{\theta}$ と $\boldsymbol{\Sigma}$ の事前分布を，十分に分散のある正規逆ウィシャート分布とした．

従来の RFM 分析でも，リセンシーを用いることが問題であるというのは指摘したとおりである．つまり，各顧客の離脱時期に関する情報が潜在変数として存在するが，観測することができない，ということである．[阿部, 2011] のモデルに基づいて各パラメータを計算する場合もこの問題は同様に存在する．そこで，潜在変

数である，観測期間後に離脱したか否かを表す $z$ と，離脱時期を表す $y$ とを擬似的に生成する方法が提案している．これは，(Tanner & Wong, 1987) が提案した，潜在変数のあるモデルのパラメータを推定する方法，いわゆるデータ拡張 (data augmentation) 法を利用して，以下のような手順で計算できる．より詳細な説明は，(Abe, 2009) [阿部, 2011] でなされている．

1　以下 (1.a) – (1.d) を顧客ごとに行う．$t, T$ はそれぞれ，最初の購買日から，最後の購買日までの経過日数と，観測終了日までの経過日数で，$x$ はリピート購買回数 (フリクエンシーから 1 を引いた値) である．

1.a　$\lambda, \mu, t, T$ から，以下で与えられる確率をもとに $z$ の乱数を発生させる．

$$p(z = 1 \mid \lambda, \mu, t, T) := \frac{\lambda + \mu}{\lambda + \mu \exp[-(\lambda + \mu)(T - t)]}$$

1.b　もし，$z = 0$ なら，期間内に離脱したことを意味するため，平均 $1/(\lambda + \mu)$ の指数分布に従う $y$ の乱数を発生させる．ただし，$t < y < T$ の範囲にある切断分布でなければならない．

1.c　仮定 IV で定めた事前分布と以下の尤度関数をもとに，事後分布にしたがうパラメータ $\ln \lambda, \ln \mu$ の乱数をそれぞれ発生させる．今回は，正規分布を提案分布[1]として，ランダムウォーク・メトロポリス-ヘイスティングス (MH) 法により乱数を発生させる．

$$f(\lambda, \mu \mid x > 0, t, T, z, y) := \frac{x^{\lambda} t^{x-1}}{\Gamma(x)} \mu^{1-z} \exp[-(\lambda + \mu)(zT + (1 - z)y]$$

ただし，$x = 0$ の場合は，尤度は以下のようになる．

$$f(\lambda, \mu \mid x = 0, t, T, z, y) := \mu^{1-z} \exp[-(\lambda + \mu)(zT + (1 - z)y]$$

1.d　$\lambda, \mu$ が所与のもとでは，仮定 IV より，$\ln \eta$ の条件事前分布は正規分布となる．さらに仮定 3 より，購買額の対数が正規分布であり，かつ事前分布の共役性から，事後分布を陽に得られる．よって，ギブス・サンプラーによって事後分布にしたがう $\ln \eta$ の乱数を発生させられる．

2　顧客ごとの個別パラメータから，共通パラメータである平均 $\boldsymbol{\theta}_0$ と共分散 $\boldsymbol{\Sigma}_0$ を更新する．平均パラメータと共分散行列パラメータの事前分布をそれぞれ多変量正規分布と逆ウィシャート分布とすると，共役性から解析的に事後分布を得られるため，ここでもギブス・サンプラーを適用できる．

以上の手順を，サンプル系列が収束するまで行う．

## SAS を用いた実装

SAS にはマルコフ連鎖モンテカルロ法 (MCMC) を実行する MCMC プロシジャが存在するが，上記のような，多次元分布のパラメータを対数変換させる処理はさんだ計算を MCMC プロシジャで実行することはできない．SAS/IML で提供されている IML プロシジャは行列計算が可能であり，これを用いて基本的なモンテカルロ法の計算を行う方法が，すでに昨年の SAS Global Forum において，(Lofland, 2016) によって言及されている．しかし，今回は，SAS で最も基本的な機能のひとつである，データステップを用いて計算を行った．その理由は次のとおりである．我々が計算するモデルはパラメータが顧客数に比例して増大するため，メモリ容量を考慮する必要がある．データステップは，現在のオブザベーションだけをメインメモリに読み込み (プログラムデータベクトル, PDV) 逐次に処理する．よって，膨大なパラメータを膨大な回数繰り返し計算する必要があるが，一方で，1 つ前よりも過去の値をさかのぼって参照しない，という特徴のある MCMC のア

[1] 対数を正規乱数として生成するため，$\lambda, \mu$ の提案分布は対数正規分布となる．

ルゴリズムと相性が良い，と判断したからである．また，配列を行列に見立てて，行列どうしの積を求めたり，コレスキー分解を実行したり，逆行列を求めたりする CALL ルーチンが FCMP プロシジャ内でのみ利用可能である．これらを活用すれば比較的容易に今回のアルゴリズムを実装できる．

このプログラムによって出力される SAS データセットは，オブザベーションが各サンプルに対応し，すべてのパラメータを個別の変数 (列) として保持するワイド形式となる．そのソースコードがプログラム 1 である．ただし，行数の多さのため，解説するうえで重要でない，記述が冗長になる計算処理の箇所は一部省略している．以下，プログラム 1 の処理を解説する．

まず，膨大な数のパラメータを扱うため，パラメータを配列として宣言する．次に，入力データをハッシュオブジェクトとして読み込む (5-12 行目)．ハッシュオブジェクトとして読み込むことで，配列のインデックスなどを使わずにすむため，ソースコードの見通しが良くなり，また，ハッシュイテレータを用いることで反復操作が可能になる，という利点がある．読み込んでいる変数 `last, end, repeat, spend, omega` はそれぞれ，$t, T, x$，顧客ごとの 1 回あたりの購買額の平均値，$\omega^2$ である．

次に，あらかじめ指定したサンプリング回数だけ DO ループを実行し，ループ内の末尾で OUTPUT ステートメントを用いてデータセットへ出力する．このサンプリング回数ループの中では，先に示したアルゴリズムの (1)で書いたように，顧客の個別パラメータに対してそれぞれ乱数を発生させる必要があるため，ループ処理内でさらに顧客別のループ処理が必要となる．このループ処理は，ハッシュイテレータオブジェクトを利用する (16 行目)．このループ内で，まず(1-a) の式で与えた確率を計算する関数 `pz()`を FCMP プロシジャで登録しておいたものを呼び出し(18 行目)，この確率に従うベルヌーイ乱数を生成し $z$ とする．次に，(1-b) では，この $z$ が 1 である場合，さらに $y$ の乱数を生成する．乱数は切断指数分布にしたがうため，乱数を生成する関数 `randtexp()`をあらかじめ登録しておき呼び出した (21 行目)．次に，(1-c) と (1-d) でそれぞれ乱数を発生させる．ここでも尤度関数の計算に行列計算が必要なため，ソースコードを冗長にしないよう FCMP プロシジャの FUNCTION ステートメントで関数を登録している(24, 25 行目)．

26 行目でハッシュイテレータによるループが終了したら，(2) の共通パラメータの更新を行う．こちらはベクトルや行列を操作する処理のため，やはり FMCP プロシジャの SUBROUTINE ステートメントを用いて計算のためのルーチンを作成した．さほど複雑な処理ではないが，ループ処理のネストが多く，ソースコードが冗長になるので割愛した．

DO ループの処理の最後に，サンプリングが何回目かを判定する (28 行目)．すでにバーンイン期間を過ぎているなら，OUTPUT ステートメントでそのオブザベーションを出力する．これによって，データセットが (サンプルの回数) * (最低限必要なパラメータ) のサイズで出力されるようになる．さらに，ループの最後には，サンプリングの品質を確認するために採択率を計算して，別のデータセットで出力している (30 行目)．

## 実験と結果

### 用いたデータ

本稿ではオンラインで一般公開されている 2 種類のデータでそれぞれ推定を行う．これらのデータは RFM に必要な最低限の情報のみを保持しているため，先に述べたように多変量回帰モデルを利用できなかった．1 つは老舗オンライン楽曲 CD 販売サイトである CDNOW の販売履歴データで，Bruce Hardie の個人ページ[2]で公開されている．これは 1997 年から 1998 年 6 月までのデータで，詳細は (Fader & Hardie, 2001) に書かれて

[2] http://www.brucehardie.com/

いる．このデータセットはおよそ 23,000 人分のユニークな顧客 ID を含んでいるが，本稿では時間の制約上，ランダム抽出した 1,000 人分のデータに絞った．検証のため,データ全体の期間のおよそ半分にあたる, 39 週間分までを集計したものを訓練データとしてパラメータの計算に用いて，そこからその後の購買回数や金額の推定値を計算し，全体の購買履歴の集計データと比較検証する．このデータの記述統計が表 1，表 2 である(以下，これを CDNOW データと呼ぶ).

もう 1 つのデータとして, UCI Machine Learning Repository で公開されている, "Online Retail Data Set" を利用した[3]．これは (Chen, et al., 2012) で利用された, 2010 年 12 月から 2011 年 12 月にかけての，英国のあるオンライン通販サービスの購買履歴であり，データの詳細はこの上記の文献に書かれている．しかし，このデータには 2 点の問題がある. 1 つは，この履歴には返品のレコードも含まれており，かつ返品レコードは購買レコードと紐づけられない，という点である．そこで，返品の場合は金額がマイナスで計上されていることを利用して，商品カテゴリや商品 ID の一致でレコードをマッチングさせることで返品額と購入額を相殺した．それでも購入額や購買額がゼロ以下になるようなレコードは除去する，という処理を行った．もう 1 つは，履歴の取得期間に購買のあった利用客をすべて集計しているため，特定の期間に登録した利用客のみの購買データではなく，厳密な意味で RFM データではない，という点である．こちらは対処しようがないため，特に対処方法を設けなかった. 2010 年 12 月~5 月の 6 か月間の購買履歴だけを訓練データとした．このようにして作成した訓練データと検証データの記述統計表がそれぞれ表 3，表 4 である (以下，これを Online Retail データと呼ぶ).

**表 1 CDNOW 購買データの記述統計 (訓練データ)**

| ラベル | N | 平均 | 標準偏差 | 最小値 | 中央値 | 最大値 |
|---|---|---|---|---|---|---|
| リセンシー(日数) | 1000 | 180.48 | 77.10 | 1.00 | 206.00 | 272.00 |
| 購買期間 (日数) | 1000 | 48.51 | 75.90 | 0.00 | 0.00 | 269.00 |
| 観測期間 (日数) | 1000 | 228.98 | 24.09 | 34.00 | 229.00 | 272.00 |
| フリクエンシー (回数) | 1000 | 1.97 | 1.86 | 1.00 | 1.00 | 20.00 |
| 平均購買額 ($) | 1000 | 31.49 | 25.60 | 3.99 | 23.54 | 259.96 |

**表 2 CDNOW の購買データの記述統計 (検証データ)**

| | N | 平均 | 標準偏差 | 最小値 | 中央値 | 最大値 |
|---|---|---|---|---|---|---|
| リセンシー (日数) | 1000 | 363.43 | 180.70 | 0.00 | 468.00 | 544.00 |
| 購買期間 (日数) | 1000 | 137.82 | 181.33 | 0.00 | 1.50 | 540.00 |
| 観測期間 (日数) | 1000 | 501.25 | 23.36 | 462.00 | 501.50 | 545.00 |
| フリクエンシー (回数) | 1000 | 2.74 | 3.06 | 1.00 | 2.00 | 33.00 |
| 平均購買額 ($) | 1000 | 31.84 | 27.45 | 3.99 | 23.70 | 361.16 |

[3] ファイルは http://archive.ics.uci.edu/ml/datasets/Online+Retail で公開されている。

表 3 Online Retail データの記述統計 (訓練データ)

| | N | 平均 | 標準偏差 | 最小値 | 中央値 | 最大値 |
|---|---|---|---|---|---|---|
| リセンシー (日数) | 1000 | 64.51 | 51.65 | 1.00 | 52.00 | 181.00 |
| 購買期間 (日数) | 1000 | 46.18 | 58.50 | 0.00 | 0.00 | 179.00 |
| 観測期間 (日数) | 1000 | 110.69 | 53.76 | 1.00 | 116.00 | 181.00 |
| フリクエンシー (回数) | 1000 | 2.31 | 2.71 | 1.00 | 1.00 | 43.00 |
| 平均購買額 (£) | 1000 | 413.20 | 654.36 | 3.75 | 298.82 | 12864.20 |

表 4 Online Retail データの記述統計 (検証データ)

| | N | 平均 | 標準偏差 | 最小値 | 中央値 | 最大値 |
|---|---|---|---|---|---|---|
| リセンシー (日数) | 1000 | 363.43 | 180.70 | 0.00 | 468.00 | 544.00 |
| 購買期間 (日数) | 1000 | 137.82 | 181.33 | 0.00 | 1.50 | 540.00 |
| 観測期間 (日数) | 1000 | 501.25 | 23.36 | 462.00 | 501.50 | 545.00 |
| フリクエンシー (回数) | 1000 | 2.74 | 3.06 | 1.00 | 2.00 | 33.00 |
| 平均購買額 ($) | 1000 | 31.84 | 27.45 | 3.99 | 23.70 | 361.16 |

## 結果

サンプリングは 12,000 回行い，最後の 2,000 個のサンプルを利用した．受容率が 40 %程度になるよう，提案分布の分散パラメータを調整している．収束を判定するため，このサンプルを 4 系列生成し，(Gelman & Rubin, 1992) による $\hat{R}$ 統計量が 1.1 未満になることを確認した．

計算結果がどの程度妥当であるかを確認するため，事後パラメータの中央値をもとに推定した購買頻度と 1 回あたりの購買平均額，そして期間中の購買総額について，推定値と実現値の相関係数と平均二乗誤差 (MSE) を計算した．表 5 は CDNOW データでの推定結果で，購買総額における MSE の大きさを除けば，MSE，相関係数ともに良好である．一方で，表 6 は Online Retail データでの結果であるが，特に購買平均額・購買総額の MSE が非常に大きい．これは，Online Retail の購買履歴データから購買額を正確に集計することが不可能であることが一因であると考えられる．しかし，いずれのデータであっても，購買平均額，購買総額の MSE が大きいという傾向がある．これは，顧客に個別の分散を設定せず，購買履歴の全体から計算した分散値を適用していること，また，リピート購買のない，つまり期間中に 1 度しか買い物をしていない顧客が多く，それらの顧客の購買平均額の分散値の高精度な推定が難しいことが原因と考えられる．

表 5 CDNOW に対するモデルの当てはまり

| データ | 指標 | 購買回数 | 購買平均額 | 購買総額 |
|---|---|---|---|---|
| 訓練 | MSE | 1.85 | 217.32 | 7550.00 |
| 検証 | MSE | 5.16 | 335.47 | 11750.00 |
| 訓練 | 相関 | 0.97 | 0.94 | 0.96 |
| 検証 | 相関 | 0.87 | 0.84 | 0.91 |

表 6 Online Retail に対するモデルの当てはまり

| データ | 指標 | 購買回数 | 購買平均額 | 購買総額 |
|---|---|---|---|---|
| 訓練 | MSE | 6.43 | 102523.00 | 1.12E+07 |
| 検証 | MSE | 62.11 | 154002.00 | 1.93E+08 |
| 訓練 | 相関 | 0.99 | 0.92 | 0.94 |
| 検証 | 相関 | 0.96 | 0.87 | 0.88 |

また, [阿部, 2011] では, この方法はデータに含まれる顧客の人数による精度の低下はあまりない, と述べている. そこで, 人数ではなくデータの期間を短縮した場合についても検証した. Online Retail と CDNOW の 2 種類の訓練データの期間を 1 か月に短縮して同様に検証したところ, 両方のデータで, MSE と相関係数いずれにも悪化が見られた. 表 7, 表 8 がその結果である. 特に検証データでの MSE と相関係数の悪化が顕著である.

表 7 CDNOW 短縮データに対するモデルの当てはまり

| データ | 指標 | 購買回数 | 購買平均額 | 購買総額 |
|---|---|---|---|---|
| 訓練 | MSE | 1.08 | 517.54 | 2459.40 |
| 検証 | MSE | 13.51 | 459.15 | 39188.00 |
| 訓練 | 相関 | 0.91 | 0.94 | 0.54 |
| 検証 | 相関 | 0.29 | 0.79 | 0.37 |

表 8 Online Retail 短縮データに対するモデルの当てはまり

| データ | 指標 | 購買回数 | 購買平均額 | 購買総額 |
|---|---|---|---|---|
| 訓練 | MSE | 1.22 | 256187.00 | 1690000.00 |
| 検証 | MSE | 151.88 | 262485.00 | 319000000.00 |
| 訓練 | 相関 | 0.98 | 0.85 | 0.75 |
| 検証 | 相関 | 0.73 | 0.66 | 0.49 |

さらに, 計算時間を検証するため, 弊社の予備サーバ[4]を用いて計算した. 比較のため, 同等の処理を C++ で記述したプログラムの実行時間を併記する. 1,000 人分の顧客データに対し, C++, SAS プログラムそれぞれ, 12,000 回サンプリングし, うち 2,000 サンプルを保存する処理を 1 セットとし, 1 セットを 100 回繰り返した結果が表 9 である. データステップによる計算速度が, C++とさほど差のないことがわかる. なお, 計算アルゴリズムの性質から, 計算時間は顧客の増加におおむね比例して増える.

表 9 計算時間の比較[5] (単位: 秒)

| | N | 平均 | 標準偏差 | 最小値 | 最大値 |
|---|---|---|---|---|---|
| C++ | 100 | 257.026 | 2.234 | 251.63 | 261.89 |
| SAS DATA STEP | 100 | 325.039 | 2.66 | 317.551 | 331.411 |

[4] OS は 64bit Windows Server, クロック数は 3.1 GHz, 物理メモリも十分あるマシンを用いた. 補助記憶装置は HDD である.

[5]出力データセットのサイズが大きいため, 特に HDD では書き込みに時間がかかり, 処理時間に大きく影響する. 条件をなるべく近くするため, C++のプログラムには結果を csv 形式で出力する処理も含めた.

また，普及品水準の性能である弊社備品のノートPCや，SAS on Demand for Academics[6]でも同様に試したところ，やはりいずれも1セットあたりの計算が数分程度で終わることを確認した．

## 結論

以上の実験から，モンテカルロ法を用いて計算するRFM分析は，実際の顧客の特徴をよくとらえていることが実証的にも示された．[阿部, 2011] の方法をマーケティングの実務で応用できる見込みは大きい．

実行可能性の観点から言えば，SASのデータステップを用いた処理は高速であり，時間的なコストの点からも実用的であると結論できる．また，個別パラメータを計算するため，最低でも顧客数かける3の個別パラメータを出力することになり，サンプリング回数も併せて結果ファイルが膨大になりやすい．対象となる顧客が多数である場合は，高性能のマシンが要求されるが，近年のクラウド環境の普及を考慮すれば大きな問題とはならないと考える．

一方で，実用上の課題もいくつか存在する．

まず，今回使用したデータの推定結果はいずれも，購買額の推定値の精度が比較的よくないという点である．データに含まれている顧客は一度しか購買しないことが多いため，全ての購買額の分散を各顧客の標本分散の代用として推定したということが原因と考えられる．モデルにおける購買額の推定は，実質，単純な対数正規分布へのあてはめであるため，モデルの改善の余地があるといえる．

もう1つの課題は，モンテカルロ法の収束の遅さである．独立MH法やランダムウォークMH法ではサンプリングの効率が悪く，データしだいでは今回のように1万回程度のバーンインが必要になる．特に独立MH法は，提案分布が事後分布の形状から大きく乖離していた場合，局所解におちいる可能性がある．モンテカルロ法に一般的な性質として，分散を小さくしすぎると，局所解から抜け出にくくなる．しかし，今回のモデルでは，分散を大きくしすぎると $y$ の乱数を生成する過程で $y$ が無限大に発散しやすくなり，計算が安定しないという問題がある．このような状況では，うまく収束するように，何度も提案分布のパラメータを調整して計算しなおす必要があるため，実際の処理時間以上に時間がかかってしまう．安定的で，より収束のしやすい計算方法への改善は今後の課題である．

## 参考文献


Abe, M., 2009. “Counting Your Customers” One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model. *Marketing Science,* 28(3), pp. 541-553.

Chen, D., Sain, S. L. & Kun, G., 2012. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management,* 19(3), pp. 197-208.

Fader, P. S. & Hardie, B. G. S., 2001. Forecasting Repeat Sales. *Interfaces,* 31(3), pp. 94-107.

Gelman, A. & Rubin, D. B., 1992. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science,* 7(4), pp. 457-472.

Lofland, C. L., 2016. *MCMC in SAS: From Scratch or by PROC.* [Online]
Available at:
https://sasglobalforum2016.lanyonevents.com/connect/fileDownload/session/E8C6352BE6C09188E0DD97C7BB4893


[6] https://www.sas.com/ja_jp/industry/higher-education/on-demand-for-academics.html

32/9080_Lofland-FP_9080.pdf
[Accessed 21 6 2017].

Tanner, M. A. & Wong, W. H., 1987. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association,* 82(398), pp. 548-550.

阿部誠., 2004. CRM のデータ分析に理論とモデルを組み込む -消費者行動理論に基づいた RF 分析-. *流通情報,* Issue 426, pp. 10-17.

阿部誠., 2008. 消費者行動理論にもとづいた個人レベルの RF 分析: 階層ベイズによる Pareto/NBD モデルの改良. *日本統計学会誌,* 37[2], pp. 239-259.

阿部誠., 2011. RFM 指標と顧客生涯価値: 階層ベイズモデルを使った非契約型顧客関係管理における消費者行動の分析. *日本統計学会誌,* 41[1], pp. 51-81.

プログラム 1 MCMC を実行するソースコード

```
data &DS.(drop=accept_l accept_m)
    work.report&CHAIN.(keep=chain iteration accept_l accept_m) ;
  /* (中略) 初期値とハイパーパラメータの設定 */
  if _N_ = 1 then do ;
    length ID  $10 last end repeat spend omega 8 ;
    declare hash
      rfm(dataset: "&INDS.(keep=ID last end repeat spend omega)",
        ordered: 'yes') ;
    declare hiter rfmiter('rfm') ;
     rfm.definekey('ID') ;
     rfm.definedata('last', 'end', 'repeat', 'spend', 'omega') ;
     rfm.definedone() ;
   end ;
   do iteration=1 to &TOTAL_ITER. ; / *ここからMCサンプリング開始 */
     c = 1 ;
     do while(rfmiter.next() = 0) /* (1) 顧客ごとのパラメータ */;
       /* (1-a) 乱数 z の生成 */
       z_[c] = pz(exp(lnlambda_[c]), exp(lnlambda_[c]), last, end) >
                 rand('uniform') ;
       /* (1-b) 乱数 y の生成 */
       if z_[c] = 0 then y_[c] = randtexp(exp(lnlambda_[c]) + exp(lnmu_[c]),
         last, end) ;
       else y_[c] = end ;
       /* (1-c) (中略)lambda, muの乱数発生 */
       /* (1-d) (中略)etaの乱数発生 ---- */
     end ;
     /* (2) (中略) theta, Gamma の更新 */
     if iteration > &BURNIN. then output &DS. ; / *サンプルの保存 */
   end ;
   / * (中略)サンプル受容率の計算 */
  run ;
```

# 最新AIの話題
## ～AIの権利は誰のもの?～

○小野　潔
株式会社インテック
社会システムプラットフォーム開発部

# Topic of the Latest AI
## ～Who have the Right of AI ?～

Kiyoshi Ono
Social Systems Platform Development Dept, INTEC Inc.

## 要旨

AI の新しい理論や技術が提案され、今回の AI ブームを牽引している。これらは画像解析と IoT などのビックデータの業務分析に導入され、成果が出始めている。転移学習や蒸留は DeepLearning の最新技術であるが、AI の権利を守る法律が未整備のため、法律上のトラブルが懸念される。トラブルを避けるためには、AI 理論を理解したうえで、AI の権利を明示する当事者間の合意・契約が必要である。本稿の前半は最新 AI 理論とその動向を紹介し、後半は AI の権利を論じる。

キーワード: Deep Learning　転移学習　蒸留　強化学習　ベイジアンネットワーク
トピックモデル　ソストクラスタリング　AI の権利保護

## 1. はじめに

“DeepLearning”“強化学習”“トピックモデル”の新しい理論が、近年発展し続けている。またコンピュータの発達により“ベイジアンネットワーク”が見直され始めた。これら理論は画像解析や IoT のビックデータの業務分析に導入され、今回の第三次 AI ブームを牽引している。

昨年から AI 組込型の業務アプリケーションが広がっているが、その AI の権利を守る法律がまったくの未整備である。AI モデルのノウハウが主にベンチャー企業（以後、AI ベンダーと称する）を中心に有しているため、一般企業は AI ベンダーに AI モデルの開発を委託あるいは、共同で AI アプリケーションの開発を行なう。ところが一般企業の法務部も弁護士も急速な AI 理論の発展について行けないのが実情であり、今後は AI の権利トラブルが懸念される。最近は AI 専門家の中でも AI 発展の阻害要因として危惧されている。

AI の権利を考える上で、最新の AI 理論を理解することが重要である。本稿の前半は、AI の適応分野、AI 理論の動向、SAS の位置づけ、今話題の DeepLearning のネットワークのコピー技術を紹介する。後半は AI の権利について、AI ベンダーとユーザーの当事者間の合意・契約を論じる。

## 2. 今回のAIブームの根源は？

従来の統計モデルやAIモデルの構築は、分析者がリアルの対象物を観察し、特徴点を見出すことから、モデルの構築から始まる。モデル構築後の処理は自動化されるが、モデル化そのものには分析者が介在する必要があった。しかし最近(2012年以降)の機械学習やDeepLearningの特徴抽出は、分析者を介在せずに自動化されつつある。今回のAIブームの根源の一つは、次元圧縮手法を利用した自動的な特徴抽出である。DeepLearningは画像データの次元圧縮から特徴量を算出し、トピックモデルは文章や顧客行動の高次元データを次元圧縮で、低次元のクラスタリング(=セグメント)に射影する。そしてDeepLearningはコンピュータの眼となることで、従来のシステムがAI組込型の業務アプリケーションへ大きく変貌しようとしている。

AIは様々な業種で利用されつつある。金融業のAIの適応分野と用いられる技術を表1[1]に示す。

| 技 術 | 主な分析技術 | 適応分野 |
|---|---|---|
| パターン認識による分類 | DeepLearning（CNN,RNN,AutoEncoder)、OCR | 紙申込書の文字認識、チャート分析 |
| パターン認識による異常値検知 | DeepLearning、ベイジアンネットワーク、トッピクモデル | 不正検知（口座開設、不正取引等）、マネーロンダリング分析、不正取引、サイバーセキュリティ |
| テキストマイニング/音声認識/意味分析 | 構文解析、形態素分析、ベイジアンネットワーク、トピックモデル、DeepLearning（RNN, LSTM) | 業務マニュアル/規定書の検索、コールセンターAI、ビジネスチャット分析、接客の自動化、営業日誌分析 |
| マーケット/CRM分析 | データマイニング、強化学習、トピックモデル、ルールベース推論、ベイズ推論 | 金融商品のレコメンド、EBM（イベント・ベース・マーケティング）、MA（マーケティング・オートメーション） |
| スコアリング/予測分析 | データマイニング（決定木、ロジスティック回帰、アンサンブル学習) | 与信モデル（住宅ローン、無担保ローン、クレジット入会）、不動産/賃貸料の価格予測 |
| 投資戦略の構築 | データマイニング、強化学習、遺伝子アルゴリズム、アンサンブル学習 | ロボ・アドバイザー、アルゴリズム・トレード、高頻度取引、投信リコメンド/ポートフォリオ分析 |
| AI組込型の業務アプリケーション | データマイニング、ベイジアンネットワーク、DeepLearning、遺伝子アルゴリズム | 職場の最適配置、人事/採用・評価・配属のAI活用、事務処理の自動化 |

表 1 金融業のAI適応分野と技術

[1] データマイニングやDeepLeaning等は複数の分析手法やモデルの集合体であり、それぞれが得意とする分析対象や分析条件が違う。本稿では代表的モデルを示す。

## 3. 最新の AI の動向

筆者が注目する AI 理論の分野は、①画像のパターン認識、②シーケンシャル（時系列）データのパターン認識、③“より良い戦略”を選択する自己学習、④文章や行動の潜在意味解析、⑤因果関係を示すネットワークが挙げられる（表 2 参照）。

| 適用分野 | 分析モデル | 説　明 |
|---|---|---|
| 画像データのパターン認識<br>（画像からの特徴抽出） | DeepLearning:<br>・CNN (畳込みニューラルネットワーク)<br>・Auto Encoder | ニューラルネットワークは入力層、中間層、出力層から構成されるが、Deep Learning の中間層は 3～数 100 層にもなる。画像解析の CNN の中間層は畳み込み層とプーリング層で構成され、畳み込み層は画像の局所的な特徴を抽出し、プーリング層は局所的な特徴をまとめる。この機能により入力画像の特徴を維持しながら画像を次元圧縮できる。 |
| シーケンシャルデータのパターン認識<br>（動画/音声の特徴抽出、自然言語処理） | DeepLearning:<br>・RNN (再帰型ニューラルネットワーク)<br>・LSTM<br>・RBM (制限ボルツマンマシン) | RNN、LSTM は時系列データや音声データの分析、文脈の文法解析に優れている。与えられた画像から説明文を自動生成したり、文章を生成できる。CNN と異なるのは、データのフィードバックのために情報を蓄える記憶素子を有することである。 |
| 強化学習<br>（より良い行動を自己学習） | Reinforcement Learning | エージェント（行動主体）は行動を選択することで環境から報酬を得られ、エージェントはより良い行動の選択を学習する。つまり強化学習は一連の行動を通じて報酬が最も多く得られるような戦略を学習する。動的計画法の一つと位置づけられる。 |
| 潜在意味解析<br>（文章の意味解釈、ID-POS データ分析） | Topic Model（トピックモデル):<br>・LDA<br>・PLSA | 離散データに隠れた潜在的なトピックを推定するモデル。トピックとは、話題などの大ざっぱな意味である。単語の出現回数から文章のセグメント化でき意味解釈の分析が可能。文章の意味解釈だけでなくビックデータの行動解析が応用できる。 |
| 因果関係解析<br>(様々な因果関係から将来を予測) | Bayesian network<br>（ベイジアンネットワーク） | ベイジアンネットワークはグラフ構造を有するグラフィカルモデルである。最近は高速計算で大量データから知識を抽出できるようになり、再び注目されてきた。ベイジアンネットワークは消費者の属性と行動の因果関係を数量的推定できるので、広告などのマーケティングの分野にも使われる。 |

表 2 注目される AI 理論

(1) DeepLearning

DeepLearning は、多層構造(3 層以上の中間層)のニューラルネットワークを用いた機械学習である。画像データを入力すると、情報が第 1 層からより深くへ伝達されるうちに、各層で特徴量が自動計算され、学習が繰り返される。

画像データの解析は DeepLearning の独壇場であるが、多くの欠点を有する。一つは膨大なパラメーター・チューニングである。現在は自動パラメーター・チューニング機能を開発し、最適値を探索するが、分析対象により数カ月のシミュレーションを要する。さらに DeepLearning は脳を真似たはずなのに、脳とは程遠い複雑なネットワーク構造を形成する点が課題である。例えば DeepLearning による動画分析は動画を

分解し、静止画像ごとに複数ネットワークを学習させ、最後に全層を結合させる方法が用いる。色やサイズの分析も同様である。

DeepLearning は少し複雑なパターン認識をさせようとすると、複数のネットワークを構築することになるが、脳がそのような複雑な処理をしているとは思われない。脳医学の実験よれば、脳の信号の多くはフィードバックしていることが判明している。生物が少ない教師データで効率よく学習できるのは、脳がフィードバック型のネットワークを有するのが一因と言われる。画像データの特徴抽出によく用いられる CNN（畳込みニューラルネットワーク）はフィードフォワード型（一方通行型）のネットワーク構造である。対してシーケンシャル（時系列）データを学習する RNN（再帰型ニューラルネットワーク）はフィードバック型のネットワーク構造を有する。RNN,LSTM は収束条件等の克服すべき課題が多いが、世界中で技術競争を加速している。もし新技術が確立できれば、動画分析、音声分析、言語解析分析も格段に進む。

(2)強化学習(Reinforcement Learning)

強化学習は世界囲碁チャンピオンを破ったソフト（アルファ碁）に使われた戦略の学習法である。強化学習は試行錯誤を通じて環境状況に適応する学習制御する。教師あり学習とは異なり，状態入力に対する正しい行動を明示する教師がなく、代わりにポイントを報酬として与えることで学習させる。ロボット、自動運転自動車、ゲーミング、ナビゲーション、投資戦略の利用が進んでいる。

(3)トピックモデル（確率的潜在意味解析法)

トピックモデルはベクトル空間モデルを利用した自然言語処理の技法である。文書群とそこに含まれる単語群から、概念の集合(トピック)を生成することで、文書を生成される過程を確率的する。具体的には「文書」×「単語」の共起行列データから文書分類ができる。またトピックモデルはビックデータ（ID-POS、GPS、IoT 等）の様々な離散データの行動分析にも応用できる。例えば「顧客 ID」×「購入物」の共起行列データは ID-POS データ分析に利用されている。

トピックモデルは高次元のデータを低次元に変換する次元圧縮手法であり、同時にデータをクラスタリングする。このクラスタリングはソフトクラスタリングと呼ばれ、全ての変数が全てのクラスに所属し、その所属度合が確信度で算出される。また確率生成モデルであるため AIC が算出でき、これを利用すれば最適クラス数の近傍を自動算出できる。従来のハードクラスタリング[2]ではクラス数を計算で求められず、分析者の感性で決めること多い。トピックモデルで ID−POS データを分析するとクラス数が 50〜100 個にもなり、様々なマーケティング戦略を立案できる。実務では既に利用されている。

(4)ベイジアンネットワーク（Bayesian network)

ベイジアンネットワークは、因果関係を確率により記述するグラフィカルモデルの 1 つで、複雑な因果関係の推論を有向非巡回グラフ構造により表現し、個々の変数の関係を条件つき確率で表す確率推論のモデルである。コンピューターの能力の問題で、実務データ向きでなかったが、最近のコンピュータの向上と効率

---

[2]クラスタリング（k-means、ウォード法、階差層的クラスター分析等）はハードクラスタリングと呼ばれ、ある変数が 1 グループに所属すると他グループには所属できない。ハードクラスタリングは大量の変数やテラバイト級のビックデータの分析では全体的に距離が大きく離れ、解を得るのは容易でない。

の良いアルゴリズムの発見からビックデータにも対応できるようになった。ベイジアンネットワークはトピックモデルによるクラス分類、他属性、行動データ等を因果関係を結び付けることができるため、ビックデータの分析ツールとして研究されている。

## 4. 驚異の DeepLearning の汎用技術（転移学習と蒸留）

本稿後半の AI 権利保護では、DeepLearning のネットワークが大きな課題となる。ここでは DeepLearning の汎用技術である「転移学習」と「蒸留」を紹介する。

(1)転移学習

学習済みネットワークは汎用性が高く、強力な特徴抽出器として他タスクのパターン認識に転移できる。転移学習は学習済みネットワークを初期状態とし、他タスクの学習データで学習させることで、少数の学習データから他のパターンを学習させる技術である。最初からモデルを生成するよりも学習効率が向上する。

学習済みネットワークは入力層に近い層から出力層に近づくにつれて、徐々に低次の特徴（線、曲線等）から意味的な特徴（顔、物体等）に構造化される。低層の特徴をとると、高い識別構造が得られず、逆に高層の特徴を選ぶと、学習データに特化しすぎて性能が下がる。そこでファインチューニング(Fine-Tuning)の転移学習では、学習済みネットワークの手前の中間層を他タスクの出力層につけかえる。他部分は学習済みパラメータを初期値とし、他タスクの教師データを学習させる。

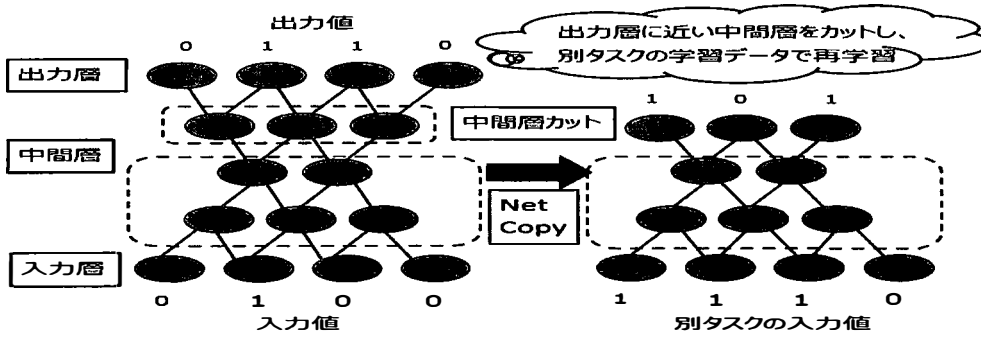



図 1　転移学習によるファインチューニングのイメージ

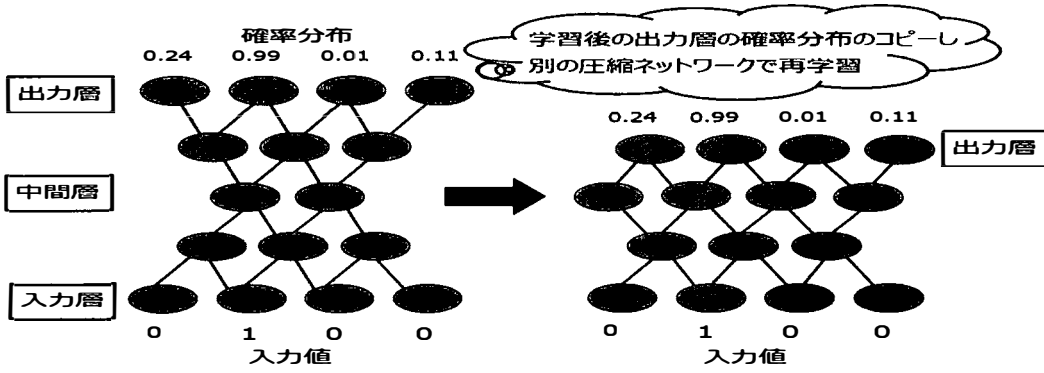



図 2　蒸留のネットワーク収縮のイメージ

(2)蒸留（Distillation）

蒸留は既に学習されたネットワークを他のネットワークで高速学習させたり、大規模なネットワークの学習結果を小さなネットワークに収縮する技術である。通常は与えられた画像が数字５の画像ならば、数字５に対応した出力ノードが１で、他の出力ノードが０となる学習を行う。それに対して蒸留は教師モデルの出力ノードの確率分布に合わせて学習させる。蒸留は複雑なネットワークとほぼ同じ性能をもたせながら、シンプルなネットワークに収縮できる。この技術を利用すれば、メモリが充分でなく処理能力が小さいスマートフォンやエッジ端末、さらにはIoTデバイスを搭載したチップでも、パターン認識が高速で実現できる。

## 5. SASのAIについて

SAS Viyaには様々な利点があるが、データ分析者にとって重要点は、SAS Viyaが公開REST API経由で一般的なプログラミング言語（Pyhon, Java, Lua）から利用できることである。他言語で開発したアプリケーションにSASの各種アルゴリズムを組込みやすくなる。今や、クラウド側では様々なDeepLearningやAIフレームワークが公開されているが、利用でできる統計関数はSAS関数に比べて機能が見劣りする。実は米国では統計ソフト（SAS、R、MatLab等）の優位性は認められているため、フレームワークの提供会社は必要最小限の統計関数しか搭載せず、不足分は統計ソフトを呼び出すことで、ユーザーニーズを満たしている。

SASのEnterpriseMinerver14.1ではベイジアンネットワークのプロシージャ（PROC HPBENET）が利用できる。また新規の特異値分解のSAS関数を利用すれば、トピックモデルを開発できる。

| | 概　略 |
|---|---|
| 教師なし学習 | 移動ウィンドウ主成分分析は、時間の経過と主成分の変化を評価 |
| | サポートベクトルデータ説明分析は、単一分類から成る機械学習技法 |
| | ロバスト主成分分析は、特徴抽出および異常検出のための行列分解を行う |
| 教師あり学習 | 部分最小二乗法回帰分析は、部分最小二乗法分析を行う |
| | 分位回帰分析は、モデル選択の自動化により分位回帰分析を予測モデルを適合 |
| テキストマイニング | テキスト構文解析およびトピック検出タスクから、ドキュメントを用語対ドキュメント行列に構文解析し、特異値分解により用語対ドキュメント行列からトピックを抽出 |
| | ブール値規則タスクは、ブール値規則をテキストデータから抽出 |

表 3　SASの新しいAI関数（SAS社Webから抜粋）

## 6. AI権利の予備知識

現状では法的にAIの権利は守れていない。そのためAIモデル開発ではAIの権利を明示した当事者間の合意・契約が重要になる。本稿では実務家の観点から、AIモデル開発に関連した契約関連に限定して論じる。

AIモデル開発は、データを提供したユーザーに権利が発生し、同時に学習データやモデルを開発したベンダー側にも権利が発生する。AIモデル開発の成果物はプログラム以外にも、例えば学習データ、パラメーターを含む学習済みモデル、モデルノウハウなど数多く存在する。通常のシステム開発ではその成果物であるプログラムは著作権法で保護される。同様にAIモデル開発のプログラムは著作権法で保護されるが、それ以外の成果物は著作権・特許権に該当しないため、AIの権利が完全には守られない。

この状態を放置すれば近将来、違法行為が発生する恐れがある。例えば、既にAIが音楽を作曲できるアプリケーションが存在するが、AIは人でないので作曲の著作権はない。しかし人による創作物とAI創作物を見分けることは困難であり、AIによる創作物を人による創作物と偽れば、法的保護が受けられる。そしてAIは人よりも高い生産性を有するので、AIの進歩がこのまま進めば、AIが作曲家、映像クリエイター、芸術家などの活躍の場を奪うことにもなる。また既に違法状態が続いているのが、撮影者の承諾なしに集めたインターネット画像の学習データへの利用である。これらの問題は関係者や業界団体が多く、解決の糸口さえも見出されていない。

## 7. AI権利の関係者

AIベンダーとユーザーとの合意・契約では、最初にAIの権利関係者洗い出す作業を行う。AI権利の主体候補は、①プログラム開発者、②AIの学習用データの提供者、③AIモデルの開発会社であるが、それ以外にも多くの権利関係者が存在する（表4参照）。

| 権利関係者 | 説　明 | 備　考 |
|---|---|---|
| 学習データ提供者 | インターネット上に写真を公開した撮影者 | 撮影人物と撮影者に著作権があるが、学習データ収集者が権利保有者に承諾を得ていないケースが多く、今後の課題に。 |
| 学習データの収集者 | インターネット上の写真を学習データとして収集した者 | DeepLearningでは該当タスクに適合した数万枚の写真データを収集する必要がある。 |
| 学習データの収集プログラムの開発者 | インターネット上の写真を収集プログラムの開発者 | 実際にインタネットから該当タスクに合う写真収集するプログラムが開発されている。 |
| 学習データの加工者 | 学習データを加工する熟練者 | AIの精度を向上（賢く）させるために、写真の一部カット、拡大、回転、反転等の加工したものを学習データに使う。 |
| 学習データの加工プログラムの開発者 | 加工プログラム開発者 | 上記のノウハウはプログラム開発されている。 |
| AIのプログラム開発者 | メインのAIのプログラム開発者 | プログラムには開発者の著作権があるが、一般にモデルのパラメーターには著作権はおよばない。 |
| AIのパラメーター算出者 | 該当タスクに適合するAIのパラメーターの算出者 | プログラム開発とパラメーター算出は別作業。モデルのパラメータをシミュレーション等で算出する。 |
| 学習を行った人 | AIに学習させた人 | 学習を行う人とパラメーター算出者は一致するケースが多い。 |
| AIコンテンツの流通者 | AIの創作した造形物・画像等を流通させた人 | 映画コンテンツの関係者はAI創作物を広めることで莫大な利益を得られるが、その権利保護を主張。 |
| 創作物をつくるAI | AIは既に初期の創作制作まで至っているがAIは人でないので権利は何も認められない | 事例：公立はこだて未来大学はAIで作成した小説を作家星新一氏のショートショート作品応募。スペインのマラガ大学は自動作曲「ラムス(Iamus)」を開発。 |

表 4　AIの権利関係者の例

## 8. 米国AI会社の戦略

AIのプログラム開発、学習によるパラメーター設定には、膨大な手間や時間、計算機資源を労する。それに対して転移学習、蒸留は、元の学習済みネットワークよりも優れた精度や計算速度の向上を有する新

しいネットワークが容易にできてしまう。しかも困ったことに、パラメーターやネットワーク構造が変わるため、転移学習、蒸留の立証は実質的にほぼ不可能である。

この手のAI権利を守る方法の一つは、クラウド側にネットワークを持たせ、ネットワーク自体を外部へダウンロードさせないことである。エッジ側では学習データの作成の前処理のみを行い、学習データをアップロードさせ、クラウド側でネットワークの学習をさせることである。

2015年頃から米国のAI会社では自社のクラウド側でDeepLearningやAIフレームワークを公開している。今や、手頃な使用料さえ払えば、誰でもがAIモデルを構築できる快適な環境を提供され、しかも学習済みネットワークのライブラリさえも提供されている。ここからはAI専門家の推測であるが、AIフレームワークの提供会社はクラウド側で元の学習済みネットワークを管理できるので、転移学習、蒸留の立証は容易である。もしAI権利の法律が制定されたら、AIフレームワークで構築したモデルの権利は、AIフレームワークの提供会社に帰属することになりかねない。

## 9. AIベンダーとユーザーとの合意・契約事項

AIベンダーとユーザーとの合意・契約では、①AIモデルを作成したプログラム、②学習データ、③パラメーターを含む学習済みモデル、④AIモデルを組込んだ業務プログラム、⑤ノウハウについての権利を明確にする必要がある。①～⑤を具体的に述べる。例えばAIベンダーが一般企業（ユーザー）が提供した学習データをSAS ViyaでAI組込型の業務アプリケーションを開発する場合を考える。AIモデルはSAS言語で開発、業務アプリケーションはJava言語で開発、そしてJavaからSAS ViyaのAIモデルを呼び出すシステムとする。

さて①AIモデル作成プログラムはAIベンダーが開発したSASプログラムを指す。②の学習データの作成にはAIベンダーのノウハウが必要になる。一般にユーザーの提供データをそのまま学習データに使用しない。賢いAIを創造するには、提供データのノイズ除去だけでなく、抽出対象物（文字、猫等）以外の部分をカット、あるいは対象物の縮小、拡大、回転、反転等の画像処理を行う。③はAIモデルの最適パラメーターを指す。DeepLearningのパラメーター数は膨大であり、出力しても意味が取れないが、成果物一つである。④はJavaで書かれた業務プログラムを指す。⑤のAIベンダーのノウハウは、学習データの作成方法、AIモデルの作成法、AIモデルのパラメーターの求めるためシミュレーション法、ネットワーク構造など多岐にわたる。AIの権利を明確にするとは、これらの事項の帰属を当事者間の合意・契約で定めることである。

さらに学習済みネットワークの開発には、特有な問題点が発生する。一つは派生モデルの取扱である。派生モデルは、学習済みモデルに別教師データを用いて転移学習を行い、生成した異なるモデルである。元モデルと派生モデルは全く違うパラメーターであるため、「派生モデルの生成の許諾権」を当事者間の合意・契約で明記する必要がある。

蒸留（Distillation）は、学習済みモデルのネットワーク構造をシンプルなネットワークに収縮させる技術である。収縮させたネットワークは、学習済みモデルとは全く別の構造のネットワークである。そのため転移学習とは別に、「蒸留行為の許諾権」を合意・契約書に明記する必要がある。また転移学習、蒸留の定義は定まっていないため、合意・契約書に転移学習と蒸留の許諾権を明記しただけ不十分であり、その定義まで記述しなければ効力を持たない。

AI モデルを不正競争防止法の「営業秘密」として保護させる条項を契約書に記載することも有効である。「営業秘密」の要件は純粋な法律問題なので、本稿では説明を省く。

## 10. おわりに

AI ベンダーと一般企業による AI モデルの共同開発は花盛りである。開発事業をスタートするにあたり、AI の権利保護の知る弁護士による契約書作成または契約書チェックは必要不可欠である。しかし AI の法律部門を有する法律事務所は少なく、また AI を知る法務の担当者でも DeepLeanring の転移学習、蒸留になると正しい理解が怪しい。彼らの勉強不足ではなく、AI 理論の急速の発展に法律や法務関係者が追い付けないためである。AI の権利を明確にするには、データサイエンティストを含めた当事者と弁護士のディスカッションから AI 権利とその関係者をすべて洗い出すことが、第一歩である。

今年中に新しい SAS Viya がリリースされる予定であり、特に DeepLearning の関数が強化される。まだ仕様が公開されていないので、本稿では SAS Viya と DeepLearning の転移学習・蒸留の言及ができなかった。

SAS システムには AI の関数やプロシージャーが数多くあり、これを利用した AI アプリケーションの開発が盛んになると予想される。しかし AI 関数を使いこなすエンジニアはまだ少ないので、一般企業は AI ベンダーとの共同開発が中心になろう。その際に本稿が役立て頂ければ幸いです。

なお、本報告書は個人的見解で書かれており、所属する株式会社インテックの意見をあらわすものではありません。

## 11. 参考文献


・麻生英樹、安田宗樹、前田新一、岡野原大輔、岡谷貴之、久保陽太郎、ボレガラ・ダヌシカ、“深層学習” 近代科学社(2015)
・岡谷貴之、“深層学習”、講談社(2015)
・C.M.ビショップ, “パターン認識と機械学習（上・下巻)”、丸善出版(2012)
・Willi Richert、“実践　機械学習システム”、オライリージャパン(2014)
・石垣司(東北大学大学院経済学研究科)、竹中毅(産業技術総合研究所サービス工学研究センター)、“日常購買行動に関する大規模データの融合による顧客行動予測システム”、人工知能学会、26 巻 6 号(2011)
・本村陽一、岩崎弘利、“ベイジアンネットワーク技術”、東京電機大学出版局(2006)
・TensorFlow is an Open Source Software Library for Machine Intelligence, TensorFlow, 2016/5, https://www.tensorflow.org
・“知的財産推進計画 2016”、知的財産戦略本部、2016/5 （首相官邸ホームページ内） http://www.kantei.go.jp/jp/singi/titeki2/kettei/chizaikeikaku20160509.pdf
・“次世代知財システム検討委員会報告書～デジタル・ネットワーク化に対応する次世代知財システム構築に向けて～”、知的財産戦略本部検証・評価・企画委員会、次世代知財システム検討委員会、平成 28 年 4 月 http://www.kantei.go.jp/jp/singi/titeki2/tyousakai/kensho_hyoka_kikaku/2016/dai5/sankou1.pdf

# One-Pass AUC Optimization手法の実務的拡張

木村和央[1]
株式会社金融工学研究所

A Practical Expansion of One-Pass AUC Optimiza tion

Kazuo Kimura
Financial Technology Research Institute Inc.

## 要旨

AUC，AR 値，Somers' D といった順序統計量は，教師ありクラス分類を目的としたモデルのパフォーマンス評価に用いられる極めて有用な指標である．しかし，伝統的な統計モデルのパラメータ推計には最尤法が用いられ，前述の順序統計量を直接最大化するような推計手法は少ない．このような状況下において，Gao et al. (2016) は，One-Pass AUC Optimization（OPAUC）という AUC を直接最大化するオンライン学習の手法を提案した．彼らは 2 値分類問題を扱ったが，本稿では順序のある多値分類問題へとその適用範囲を拡張した（拡張 OPAUC）．また，オンライン学習の範疇からは外れるが，サンプルを 2 度読み利用する Two-Pass AUC Optimization（拡張 TPAUC），さらに回数を増やした Multi-Pass AUC Optimization（拡張 MPAUC）を提案した．さらに，実務的なモデル構築時への対応も意識しつつ，各サンプルの重みを可変とし，パラメータへ制約条件を課すことも可能とした．これら拡張手法の定式化およびアルゴリズムを示した後，格付モデル構築を例に数値実験を行い，各手法の精度を比較した．残念ながら，提案手法は最尤法順序ロジットモデルの精度を上回ることはできなかったが，計算コストを考えれば十分実用に耐えうる結果が得られた．参考として，付録に SAS によるコーディング例を収録した．

キーワード：AUC，オンライン学習，最小二乗損失，格付推計，順序ロジットモデル，SAS/IML

## 1. はじめに

企業等が一定期間後にデフォルトとなる確率，あるいは格付に該当する確率を推計するために，ロジットモデルが金融機関実務の中で広く活用されている．モデル構築にあたっては，限られた時間コストの中で，人的スキルによらずに，最良でなくても構わないが一定精度を確保したモデルを得たいという実務上の要請がある．これに対し，木村（2015，2016）は，ソブリン（国）のデフォルト確率および単一ないし複数の格付該当確率モデルの構築を題材に，Wei ghtof Evi denceを用いた外れ値処理，変数変換，欠損値補完，同時推計といった手法の提案に取り組んできた．さまざまな工夫がなされているが，ここで用いられたロジットモデルないし順序ロジットモデルはいわば教師ありクラス分類を目的とした伝統的な線型の統計モデルである．

[1] 本稿の内容は筆者に属し，所属組織の見解ではない．

これらモデルのパフォーマンスを評価するための指標として，ROC 曲線の下側面積にあたる AUC（Area Under the Curve），それと単調の関係にあり金融分野でスタンダードな AR 値（AccuracyRatio），AR 値と同様の Somers' D といった順序統計量が重要視されてきた．一方で，当該モデル構築時におけるパラメータの推計は，一般には最尤法によりなされ，直接的に評価指標の最大化を目的とはしていない．つまり，推計手法と評価手法の間にギャップが存在している．

これに対し，山下・三浦（2011）は，AR 値の定義式にあらわれる階段関数をシグモイド関数で置換した近似 AR 値を直接最大化するという強力な推計手法を提案した．しかし，異なるクラスに所属するサンプル間の全組合せ $O(N^2d)$（$N$はサンプル数，$d$は説明変数の数）の記憶容量を保持した上で収束計算を行う必要があり，通常の最適化ルーチンでは大規模データへの適用は困難な状況にあった．筆者は，長らく大規模データに適用可能な手法がないか調査・探索していたが，最近になって，Gao et al.（2016）が，AUC の定義式にあらわれる階段関数を最小二乗型損失で置換した代理損失関数を最小化する One-Pass AUC Optimization（OPAUC）と呼ばれるオンライン学習の手法を提案していたことに気づいた[2]．オンライン学習とは，訓練データ（サンプル）が 1 件ごとに順次もたらされるという状況下で，モデルのパラメータをいかに推計するべきかというタスクである．OPAUC は，訓練データが逐次的に与えられるとき，学習を一からやり直すことなくこれまでの学習結果を更新することができるため，非常に計算効率が高い．また，OPAUC で必要とされる記憶容量はサンプル数に依らず，$O(d^2)$で済むため，大規模データへの適用も可能である．ただし，前述のバッチ学習と比較した際の精度低下の程度が気になるところである．

本稿では，Gao et al.（2016）の OPAUC をベースに，彼らが 2 値分類問題を扱ったところを，順序のある多値分類問題へとその適用範囲を拡張した（拡張 OPAUC）．また，オンライン学習の範疇からは外れるが，サンプルを 2 度読み利用する Two-Pass AUC Optimization（拡張 TPAUC），さらに回数を増やした Multi-Pass AUC Optimization（拡張 MPAUC）を提案した．さらに，実務的なモデル構築時への対応も意識しつつ，各サンプルの重みを可変とし，パラメータへ制約条件を課すことも可能とした．これら拡張手法の定式化およびアルゴリズムを示した後，格付モデル構築を例に数値実験を行い，提案手法のほか各手法の精度を比較した．参考として，付録に SAS によるコーディング例を収録した．

## 2. One-Pass AUC Optimization 手法とその拡張

### 2.1 問題設定

順序のある多値分類問題において AUC を最大化するような線型判別モデルのオンライン学習（追加学習）を考える[3]．基本的な枠組みは Gao et al.（2016）と同様であるので原論文も参照いただきたい．$t$番目に与えられたサンプルを$(\mathbf{x}_t, y_t)$，その重みを$\omega_t(>0)$とする[4]．ここで，$\mathbf{x}_t \in \mathbb{R}^d$は説明変数，$y_t \in \{0,1,\cdots,m\}$はクラス（被説明変数）である．また，$t$番目までのサンプルにより推計された線型モデルを$f_t(\mathbf{x}) = \mathbf{w}_t \cdot \mathbf{x}$とし，$\mathbf{w}_t \in \mathbb{R}^d$は係数パラメータである．

この設定の下で，AUC を定義する．各サンプルの重みを$\omega_i$としたデータセット$\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{0,1,\cdots,m\} | i \in [n]\}$, $[n] = \{1,2,\cdots,n\}$が与えられたとして，このデータセットをクラス$y_i$の値により分割する．$\mathcal{D}_0 = \{(\mathbf{x}_{i_0}^0, 0) | i_0 \in [n_0]\}$（各サンプルの重み$\omega_{i_0}$），$\mathcal{D}_1 = \{(\mathbf{x}_{i_1}^1, 1) | i_1 \in [n_1]\}$（同，重み$\omega_{i_1}$），$\cdots$，$\mathcal{D}_m =$

[2] 2016 年以前から OPAUC は提案されていたが，総括論文である Gao et al.（2016）のみ参考文献に掲載した．
[3] Ding et al. (2015) の記号および流れに沿った記載としているが，本稿では Adaptive Gradient 手法は扱わない．
[4] Gao et al. (2016) は，$y_t \in \{-1,+1\}$とした 2 値分類，各データの重みは$\omega_t = 1$とした問題を扱っている．

$\{(\mathbf{x}_{i_m}^m, m) | i_m \in [n_m]\}$ （同，重み$\omega_{i_m}$）．$\mathbf{w} \in \mathbb{R}^d$を係数パラメータとして，$\mathcal{D}$の AUC の定義は，

$$\mathrm{AUC}(\mathbf{w}) = \frac{\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}\left(\sum_{i_k=1}^{n_k}\sum_{i_l=1}^{n_l}\left(\mathbb{I}_{(\mathbf{w}\cdot\mathbf{x}_{i_l}^l > \mathbf{w}\cdot\mathbf{x}_{i_k}^k)} + \frac{1}{2}\mathbb{I}_{(\mathbf{w}\cdot\mathbf{x}_{i_l}^l = \mathbf{w}\cdot\mathbf{x}_{i_k}^k)}\right)\omega_{i_k}\omega_{i_l}\right)}{\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}\left(\sum_{i_k=1}^{n_k}\omega_{i_k}\right)\left(\sum_{i_l=1}^{n_l}\omega_{i_l}\right)}, \tag{1}$$

ここで，$\mathbb{I}_\pi$は指示関数であり，命題$\pi$が真なら 1，偽なら 0．なお，$\mathcal{D}$の AR 値（Somers' D）の定義は，

$$\mathrm{AR}(\mathbf{w}) = \frac{\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}\left(\sum_{i_k=1}^{n_k}\sum_{i_l=1}^{n_l}\left(\mathbb{I}_{(\mathbf{w}\cdot\mathbf{x}_{i_l}^l > \mathbf{w}\cdot\mathbf{x}_{i_k}^k)} - \mathbb{I}_{(\mathbf{w}\cdot\mathbf{x}_{i_l}^l < \mathbf{w}\cdot\mathbf{x}_{i_k}^k)}\right)\omega_{i_k}\omega_{i_l}\right)}{\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}\left(\sum_{i_k=1}^{n_k}\omega_{i_k}\right)\left(\sum_{i_l=1}^{n_l}\omega_{i_l}\right)}, \tag{2}$$

であり，$\mathbb{I}_{(\mathbf{w}\cdot\mathbf{x}_{i_l}^l > \mathbf{w}\cdot\mathbf{x}_{i_k}^k)} + \mathbb{I}_{(\mathbf{w}\cdot\mathbf{x}_{i_l}^l = \mathbf{w}\cdot\mathbf{x}_{i_k}^k)} + \mathbb{I}_{(\mathbf{w}\cdot\mathbf{x}_{i_l}^l < \mathbf{w}\cdot\mathbf{x}_{i_k}^k)} = 1$により，$\mathrm{AR}(\mathbf{w}) = 2 \times \mathrm{AUC}(\mathbf{w}) - 1$が確認される．

0 から 1 へ不連続に変化する指示関数を含むこれらの目的関数に対し，一般に連続関数を前提とした最適化手法を適用することは困難である．このため，最小 2 乗型の代理損失[5]

$$\ell(\mathbf{w}, \mathbf{x}_{i_l}^l - \mathbf{x}_{i_k}^k) = \left(1 - \mathbf{w}\cdot(\mathbf{x}_{i_l}^l - \mathbf{x}_{i_k}^k)\right)^2, \tag{3}$$

を用いて，AUC を最大化する代わりに，以下の損失関数を最小化することを考える．

$$\mathcal{L}(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}\left(\sum_{i_k=1}^{n_k}\sum_{i_l=1}^{n_l}\ell(\mathbf{w}, \mathbf{x}_{i_l}^l - \mathbf{x}_{i_k}^k)\,\omega_{i_k}\omega_{i_l}\right)}{2\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}\left(\sum_{i_k=1}^{n_k}\omega_{i_k}\right)\left(\sum_{i_l=1}^{n_l}\omega_{i_l}\right)}, \tag{4}$$

右辺第 1 項は L2 正則化項であり，モデルの複雑さを抑制する効果をもつ．

### 2.2 OPAUC 手法とその拡張

(4)の損失関数$\mathcal{L}(\mathbf{w})$の最小化をオンライン学習にて実現させるために，個々のサンプルに対する損失の和を用いて次のように捉えなおす．全サンプル数を$T$として，

$$\frac{1}{T_T}\sum_{t=1}^{T}\mathcal{L}_t(\mathbf{w}) \qquad \text{ここで，}$$

$$\begin{aligned}\mathcal{L}_t(\mathbf{w}) &= \omega_t\left[\frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{\substack{k=0,\\ k\neq y_t}}^{m}\left(\frac{\sum_{i=1}^{t-1}\mathbb{I}_{(y_i=k)}\omega_i}{\sum_{i=1}^{t-1}\mathbb{I}_{(y_i\neq y_t)}\omega_i} \times \frac{\sum_{i=1}^{t-1}\left(1 - \left(\mathbb{I}_{(y_t>y_i, y_i=k)} - \mathbb{I}_{(y_t<y_i, y_i=k)}\right)(\mathbf{x}_t - \mathbf{x}_i)^{\mathrm{T}}\mathbf{w}\right)^2\omega_i}{2\sum_{i=1}^{t-1}\mathbb{I}_{(y_i=k)}\omega_i}\right)\right] \\ &= \omega_t\left[\frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{\substack{k=0,\\ k\neq y_t}}^{m}\left(\frac{T_t^k}{T_t - T_t^{y_t}} \times \frac{\sum_{i:y_i=k}\left(1 - \left(\mathbb{I}_{(y_t>k)} - \mathbb{I}_{(y_t<k)}\right)(\mathbf{x}_t - \mathbf{x}_i)^{\mathrm{T}}\mathbf{w}\right)^2\omega_i}{2T_t^k}\right)\right].\end{aligned} \tag{5}$$

$(\cdot)^{\mathrm{T}}$は行列の転置，$T_t^k = \sum_{i=1}^{t}\mathbb{I}_{(y_i=k)}\omega_i$はクラス$k$に属するサンプル重みの和，$T_t = \sum_{k=1}^{m}T_t^k$はすべてのクラスのサンプル重みの和である．右辺の括弧内第 2 項は，$t$番目のサンプルのクラスである$y_t$とは異なるクラス$k(\neq y_t)$に属するこれまでに得られたデータとの代理損失の比較をすること，その際に各データの重み$\omega_i$を考慮した加重平均となるところが Gao et al. (2016)から自然に拡張された点である．なお，Gao et al. (2016)における設定，つまりサンプル重みが 1 の 2 値分類問題では，逐次得られるサンプルが i.i.d.であるならば，$\mathcal{L}_t(\mathbf{w})$は$\mathcal{L}(\mathbf{w})$の不偏推定量となることが証明されている．

次に，(5)の$\mathcal{L}_t(\mathbf{w})$の勾配（gradient）すなわち$\mathbf{w}$による微分を計算すると，

[5] $\mathbf{w}\cdot(\mathbf{x}_{i_l}^l - \mathbf{x}_{i_k}^k)$が 1 を超えるとき，指示関数は真であり 1 であるが，代理損失$\ell$は増加してしまうという問題がある．

$$\nabla\mathcal{L}_t(\mathbf{w}) = \omega_t \Bigg[ \lambda \mathbf{w} + \sum_{\substack{k=0,\\ k\neq y_t}}^{m} \frac{T_t^k}{T_t - T_t^{y_t}} \Bigg( \mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}\mathbf{w} - \left(\mathbb{I}_{(y_t>k)} - \mathbb{I}_{(y_t<k)}\right)\mathbf{x}_t + \frac{1}{T_t^k}\sum_{i:y_i=k}\left(\left(\mathbb{I}_{(y_t>k)} - \mathbb{I}_{(y_t<k)}\right)\mathbf{x}_i + \left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_t^{\mathrm{T}} - \mathbf{x}_t\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{w}\right)\omega_i \Bigg)\Bigg], \tag{6}$$

ここで，クラス$k$における説明変数の平均 $\mathbf{c}_t^k = \frac{1}{T_t^k}\sum_{i:y_i=k}^{t}\mathbf{x}_i\omega_i$，共分散行列 $S_t^k = \frac{1}{T_t^k}\sum_{i:y_i=k}^{t}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \mathbf{c}_t^k[\mathbf{c}_t^k]^{\mathrm{T}}\right)\omega_i$ を導入して上式を変形すると，

$$\nabla\mathcal{L}_t(\mathbf{w}) = \omega_t\left[\lambda\mathbf{w} + \sum_{\substack{k=0,\\ k\neq y_t}}^{m}\frac{T_t^k}{T_t - T_t^{y_t}}\left(-\left(\mathbb{I}_{(y_t>k)} - \mathbb{I}_{(y_t<k)}\right)(\mathbf{x}_t - \mathbf{c}_t^k) + (\mathbf{x}_t - \mathbf{c}_t^k)(\mathbf{x}_t - \mathbf{c}_t^k)^{\mathrm{T}}\mathbf{w} + S_t^k\mathbf{w}\right)\right], \tag{7}$$

と簡略化される．これにより，最適な係数パラメータ$\mathbf{w}$を得るために保持しなければならないデータは，各クラス（$k = 0,1,\cdots,m$）における説明変数の平均（$\mathbf{c}_t^k$）と共分散行列（$S_t^k$）のみであり，$O(d^2)$で済むことがわかる．このあとは，勾配降下法により$\mathbf{w}$を逐次更新していけばよい．ただし，本稿では，$\mathbf{w}$に対する制約条件も考慮する．$\mathbf{w}$の最大値を与えるベクトルを$\mathbf{b}_{\mathbf{max}}$，最小値を与えるベクトルを$\mathbf{b}_{\mathbf{min}}$とすると，

$$\mathbf{w}_t = \max(\min(\mathbf{w}_{t-1} - \eta\nabla\mathcal{L}_t(\mathbf{w}_t), \mathbf{b}_{\mathbf{max}}), \mathbf{b}_{\mathbf{min}}), \tag{8}$$

となる．ここで，$\eta$はステップ幅である．

なお，同時に平均（$\mathbf{c}_t^k$）と共分散行列（$S_t^k$）も$t$と$t-1$の間の漸化式にて更新していく必要がある．

$$\begin{aligned}
\mathbf{c}_t^k &= \frac{1}{T_t^k}\sum_{i:y_i=k}^{t}\mathbf{x}_i\omega_i = \frac{T_t^k - \mathbb{I}_{(y_t=k)}\omega_t}{T_t^k}\sum_{i:y_i=k}^{t-1}\frac{\mathbf{x}_i\omega_i}{T_t^k - \mathbb{I}_{(y_t=k)}\omega_t} + \frac{\mathbf{x}_t\mathbb{I}_{(y_t=k)}\omega_t}{T_t^k} = \frac{T_t^k - \mathbb{I}_{(y_t=k)}\omega_t}{T_t^k}\mathbf{c}_{t-1}^k + \frac{\mathbf{x}_t\mathbb{I}_{(y_t=k)}\omega_t}{T_t^k} \\
&= \mathbf{c}_{t-1}^k + \frac{\mathbb{I}_{(y_t=k)}\omega_t}{T_t^k}(\mathbf{x}_t - \mathbf{c}_{t-1}^k).
\end{aligned}$$

$$\begin{aligned}
S_t^k &= \frac{1}{T_t^k}\sum_{i:y_i=k}^{t}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \mathbf{c}_t^k[\mathbf{c}_t^k]^{\mathrm{T}}\right)\omega_i = \frac{1}{T_t^k}\sum_{i:y_i=k}^{t}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\omega_i - \mathbf{c}_t^k[\mathbf{c}_t^k]^{\mathrm{T}} \\
&= \frac{T_t^k - \mathbb{I}_{(y_t=k)}\omega_t}{T_t^k}\sum_{i:y_i=k}^{t-1}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\omega_i}{T_t^k - \mathbb{I}_{(y_t=k)}\omega_t} + \frac{\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}\mathbb{I}_{(y_t=k)}\omega_t}{T_t^k} - \mathbf{c}_t^k[\mathbf{c}_t^k]^{\mathrm{T}} \\
&= \frac{T_t^k - \mathbb{I}_{(y_t=k)}\omega_t}{T_t^k}\left(\sum_{i:y_i=k}^{t-1}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\omega_i}{T_t^k - \mathbb{I}_{(y_t=k)}\omega_t} - \mathbf{c}_{t-1}^k[\mathbf{c}_{t-1}^k]^{\mathrm{T}} + \mathbf{c}_{t-1}^k[\mathbf{c}_{t-1}^k]^{\mathrm{T}}\right) + \frac{\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}\mathbb{I}_{(y_t=k)}\omega_t}{T_t^k} - \mathbf{c}_t^k[\mathbf{c}_t^k]^{\mathrm{T}} \\
&= \frac{T_t^k - \mathbb{I}_{(y_t=k)}\omega_t}{T_t^k}(S_{t-1}^k + \mathbf{c}_{t-1}^k[\mathbf{c}_{t-1}^k]^{\mathrm{T}}) + \frac{\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}\mathbb{I}_{(y_t=k)}\omega_t}{T_t^k} - \mathbf{c}_t^k[\mathbf{c}_t^k]^{\mathrm{T}} \\
&= S_{t-1}^k - \frac{\mathbb{I}_{(y_t=k)}\omega_t}{T_t^k}(S_{t-1}^k + \mathbf{c}_{t-1}^k[\mathbf{c}_{t-1}^k]^{\mathrm{T}} - \mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}) + \mathbf{c}_{t-1}^k[\mathbf{c}_{t-1}^k]^{\mathrm{T}} - \mathbf{c}_t^k[\mathbf{c}_t^k]^{\mathrm{T}}.
\end{aligned}$$

$y_t \neq k$のとき，$\mathbb{I}_{(y_t=k)} = 0$であり，$\mathbf{c}_t^k = \mathbf{c}_{t-1}^k$と$S_t^k = S_{t-1}^k$であることが自然に導出される．

以上の定式化を**拡張 One-Pass AUC Optimization（拡張 OPAUC）**と呼称する．このターンで最終的に得られたサンプル重みの和，総和，平均，共分散行列は，$T_T^k$，$T_T$，$\mathbf{c}_T^k$，$S_T^k$（$k = 0,1,\cdots,m$）により固定し，$\mathbf{w}_T$を初期値として，再度，先頭のサンプルに戻って繰り返すことで，精度を向上させることが期待される．本手法を**拡張 Two-Pass AUC Optimization（拡張 TPAUC）**と呼称する．拡張 TPAUC はオンライン学習の範疇から外れるものの，計算負荷は拡張 OPAUC の 2 倍まではかからない．

## 2.3 拡張 OPAUC/TPAUC アルゴリズム

入　力：正則化パラメータ$\lambda(>0)$，ステップ幅$\eta(>0)$，
　　制約条件用最大値$\mathbf{b}_{\max}$，最小値$\mathbf{b}_{\min}$，
　　反復回数$U$（One-Pass は1，Two-Pass は2）.

初期化：パラメータ$\mathbf{w}_0=\mathbf{0}$，クラス$k=0,1,\cdots,m$として，　　パラメータおよび各種データを
　　平均$\mathbf{c}_0^k=\mathbf{0}$，共分散行列$S_0^k=[0]_{d\times d}$，　　ゼロクリア.
　　サンプル重みの和$T_0^k=0$，総和$T_0=0$.

処　理：**DO** $u=1$ **TO** $U$;　　開始 ループ：反復回数 $u$.
　**DO** $t=1$ **TO** $T$;　　開始 ループ：サンプル番号 $t$.
　《サンプルの読込》
　　説明変数$\mathbf{x}_t$; クラス$y_t$; サンプルの重み$\omega_t$;
　《各種データ更新》
　**IF** $u=1$ **THEN DO**;　　条件 初回ループのとき.
　　$T_t^{y_t}=T_{t-1}^{y_t}+\omega_t$; $T_t=T_{t-1}+\omega_t$;
　　$\mathbf{c}_t^{y_t}=\mathbf{c}_{t-1}^{y_t}+\omega_t\left(\mathbf{x}_t-\mathbf{c}_{t-1}^{y_t}\right)/T_t^{y_t}$;
　　$S_t^{y_t}=S_{t-1}^{y_t}+\omega_t\left(\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}-\mathbf{c}_{t-1}^{y_t}\left[\mathbf{c}_{t-1}^{y_t}\right]^{\mathrm{T}}-S_{t-1}^{y_t}\right)/T_t^{y_t}$
　　$+\mathbf{c}_{t-1}^{y_t}\left[\mathbf{c}_{t-1}^{y_t}\right]^{\mathrm{T}}-\mathbf{c}_t^{y_t}\left[\mathbf{c}_t^{y_t}\right]^{\mathrm{T}}$;
　　**DO** $k=0$ **TO** $m$;　　開始 ループ：クラス番号 $k$.
　　**IF** $k\neq y_t$ **THEN DO**;　　クラス番号$\neq y_t$のときは，
　　$T_t^k=T_{t-1}^k$; $\mathbf{c}_t^k=\mathbf{c}_{t-1}^k$; $S_t^k=S_{t-1}^k$;　　各種データを単純転送.
　　**END**;
　　**END**;　　終了 ループ：クラス番号 $k$.
　**END**;
　**IF** $u>1$ **THEN DO**;　　条件 2回目以降のループのとき.
　　**DO** $k=0$ **TO** $m$;　　開始 ループ：クラス番号 $k$.
　　$T_t^k=T_T^k$; $T_t=T_T$;$\mathbf{c}_t^k=\mathbf{c}_T^k$; $S_t^k=S_T^k$;　　1回目での最終確定値を単純転送.
　　**END**;　　終了 ループ：クラス番号 $k$.
　**END**;
　《損失関数の勾配》
　　$\nabla\mathcal{L}_t=\mathbf{0}$;　　損失関数の勾配$\nabla\mathcal{L}_t$をゼロクリア.
　**DO** $k=0$ **TO** $m$;　　開始 ループ：クラス番号 $k$.
　　**IF** $k\neq y_t$ **THEN DO**;　　クラス番号$\neq y_t$のときのみ計算.
　　**IF** $y_t>k$ **THEN** $\mathrm{Sgn}=+1$;　　クラス番号と$y_t$の大小比較で
　　**IF** $y_t<k$ **THEN** $\mathrm{Sgn}=-1$;　　下式で必要な符号を設定.
　　$\nabla\mathcal{L}_t=\nabla\mathcal{L}_t+T_t^k/\left(T_t-T_t^{y_t}\right)\omega_t$　　$\nabla\mathcal{L}_t$を計算.
　　$\times\left(\lambda\mathbf{w}_{t-1}-\mathrm{Sgn}\times\left(\mathbf{x}_t-\mathbf{c}_t^k\right)\right.$　　**なお，2回目以降のパラメータは前回**
　　$\left.+\left(\mathbf{x}_t-\mathbf{c}_t^k\right)\left(\mathbf{x}_t-\mathbf{c}_t^k\right)^{\mathrm{T}}\mathbf{w}_{t-1}+S_t^k\mathbf{w}_{t-1}\right)$;　　**の最終値が初期値となることに注意.**
　　**END**;
　**END**;　　終了 ループ：クラス番号 $k$.
　《パラメータ更新》
　　$\mathbf{w}_t=\max(\min(\mathbf{w}_{t-1}-\eta\nabla\mathcal{L}_t,\mathbf{b}_{\max}),\mathbf{b}_{\min})$;
　**END**;　　終了 ループ：サンプル番号 $t$.
**END**;　　終了 ループ：反復回数 $u$.

出　力：パラメータ$\mathbf{w}_T$

### 2.4 （参考）シグモイド関数を利用した近似 AR 値最大化

ここで，簡単に山下・三浦（2011）で提案された手法を説明しておく．(2)の AR 値の定義式に現れる指示関数（階段関数）を次のシグモイド関数にて置換する．

$$s_\sigma\left(\mathbf{w}, \mathbf{x}_{i_l}^l - \mathbf{x}_{i_k}^k\right) = 1/\left(1 + \exp\left(-\mathbf{w} \cdot \left(\mathbf{x}_{i_l}^l - \mathbf{x}_{i_k}^k\right)/\sigma\right)\right), \tag{9}$$

ここで，$\sigma$（$> 0$）は近似精度パラメータであり，小なるほど階段関数に近くなる．近似 AR 値$\mathrm{sAR}_\sigma(\mathbf{w})$は，

$$\begin{aligned}\mathrm{sAR}_\sigma(\mathbf{w}) &= 2 \times \frac{\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}\left(\sum_{i_k=1}^{n_k}\sum_{i_l=1}^{n_l} s_\sigma\left(\mathbf{w}\mathbf{x}_{i_l}^l - \mathbf{x}_{i_k}^k\right)\omega_{i_k}\omega_{i_l}\right)}{\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}\left(\sum_{i_k=1}^{n_k}\omega_{i_k}\right)\left(\sum_{i_l=1}^{n_l}\omega_{i_l}\right)} - 1 \\ &= 2 \times \frac{\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}\left(\sum_{i_k=1}^{n_k}\sum_{i_l=1}^{n_l} s_\sigma\left(\mathbf{w}, \mathbf{x}_{i_l}^l - \mathbf{x}_{i_k}^k\right)\omega_{i_k}\omega_{i_l}\right)}{\sum_{k=0}^{m-1}\sum_{l=k+1}^{m}(T_T^k T_T^l)} - 1\end{aligned} \tag{10}$$

となる．ここで，右辺の分母は$\mathbf{w}$によらない確定値であるので，分子を最大化すればよい．連続関数で近似されているため，理論上は最適化ルーチンにより最大化させることは可能であるが，異なるクラス間のサンプル同士のすべての組合せを保持しての最適化計算になるので，大規模データでは困難であることもわかる．

## 3. 数値実験

### 3.1 実験方法

格付モデル構築に，近似 AR 値最大化（山下・三浦，2011），最尤法による順序ロジットモデル，拡張 OPAUC，拡張 TPAUC，さらに反復回数$U$を 22 回とした拡張 Multi-Task Optimization MPAUC）の各手法を適用し，Somer'sD により精度を比較した．拡張 OPAUC/TPAUC/MPAUC については SAS/IML を用いて計算した（付録 A のコードを参照）．近似 AR 値最大化と最尤法による順序ロジットモデルについては SAS/STAT の NLMIXED プロシジャを用いて計算した（前者については付録 B のコードを参照）[6]．

実験には，金融・ノンバンクを除く上場企業のうち，格付投資情報センター（R&I）が長期発行体格付を付与した先を対象とし，本決算データから作成した財務指標を説明変数（$d = 76$）に，6 か月後の R&I 格付（AA+ 以上 / BB+ 以下を集約し，$m = 10$）をクラス（被説明変数）に用いた．$Y-2$，$Y-1$年の 2 か年のサンプル（$N$~800）を訓練データとしてモデルを推計し，$Y$年のサンプル（$N$~400）にてテストした[7]．各説明変数$x_{tj}$（$j = 1, \cdots, d$）は，訓練データにおける各変数の 5%タイル値$x_{.j,5}$，95%タイル値$x_{.j,95}$により，$x_{tj} \leftarrow \max\left(\min\left(\left(x_{tj} - x_{.j,5}\right)/\left(x_{.j,95} - x_{.j,5}\right), 1\right), -1\right)$，と正規化した．また，説明変数には符号条件を設定した．さらに，格付別サンプル数の偏りを補正するため逆数ウエイトにてサンプル重みを設定した．これは，クラス別サンプル数（個数），その総和を$N_T^k$，$N_T$（$k = 0,1,\cdots,m$）としたとき，サンプル重みを$\omega_t \leftarrow (N_T/(m+1))/N_T^{y_t}$と設定したということである．拡張 OPAUC/TPAUC/MPAUC のハイパーパラメータについては，正則化パラメータ$\lambda = 2^{[-10:-2]}$，ステップ幅$\eta = 2^{[-10:-2]}$を訓練データにて試行し，Somer'sD 最大時の係数パラメータを，テストデータへ適用した．その際に訓練データの並び順による影響を受けないよう，乱数でソートした 10 セットのデータにて試行し，その平均値で精度を比較した．

### 3.2 実験結果

図表 1 に$Y = 2007$~$2015$年の平均値として計算された Somers' D による手法間の精度比較を示した．FREQ プロシジャにて，TABLE ･･･/MEASURES;OUTPUT SMDRC;とすれば，Somers' D が得られる．

[6] 最尤法による順序ロジットも，制約条件を考慮する必要から，LOGISTIC ではなく，NLMIXED プロシジャを利用した．
[7] 翌年の格付は，当年の格付とは，それほど差があるわけではなく，テストで大きく精度低下することは考えにくい．

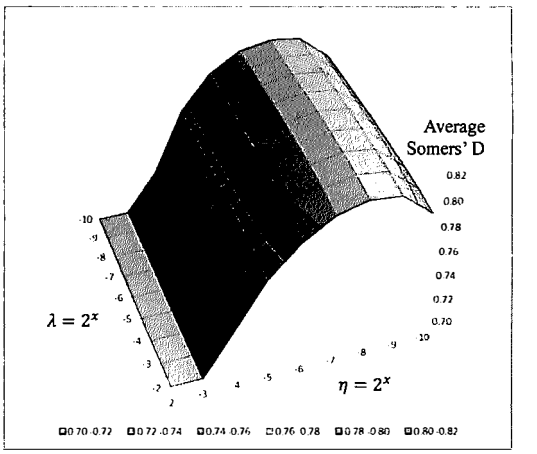
残念ながら，拡張 OPAUC の精度は最尤法による順序ロジットモデルにも劣後したが，計算コストとテストデータでの精度差を勘案すれば，実用に耐えうる結果である．また，2 回ないし複数回処理を繰り返す拡張 TPAUC / MPAUC で，精度が向上することも確認できた．

次に，図表 2 に，拡張 OPAUC 手法を例として，そのハイパーパラメータである$\lambda$と$\eta$の設定値が Somers' D へ与える影響を示した．本図によれば，ステップ幅は$\eta$~$2^{[-9:-6]}$程度に設定するのが望ましい．一方，正則化パラメータ$\lambda$が与える影響は小さいことがわかった．

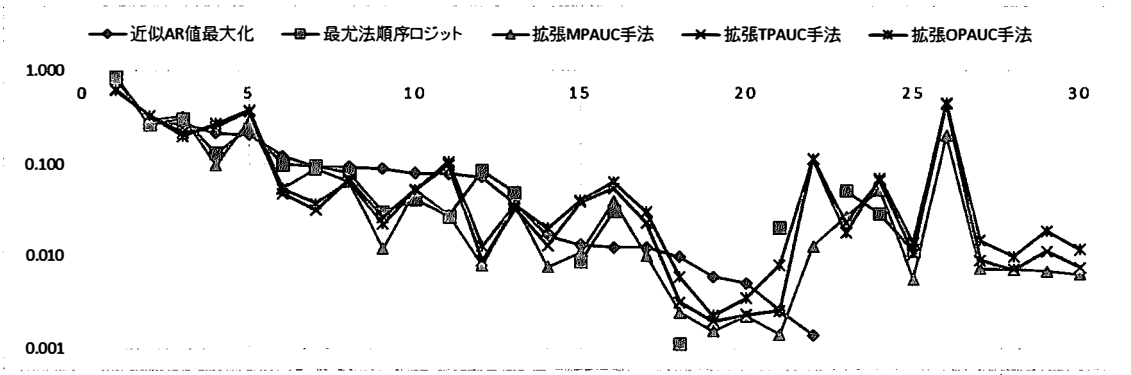
さらに，推計された係数パラメータが，各手法間でどの程度異なるかについて興味があるところであるが，図表 3 に$Y = 2015$年を例として示した．本図は，パラメータの 2 乗和すなわちノルムを 1 に調整してある．また，横軸の説明変数は近似 AR 値最大化のパラメータ降順にソートしてある．76 もの候補変数を投入したため，相関の高い変数も存在しており，各手法間でパラメータの値に乖離があってもモデル全体としては近い結果になっている可能性があることには注意する必要がある．係数パラメータの 1 番目は自己資本額（対数）であったが，近似 AR 値，最尤法，MPAUC が 0.8 以上であるのに対し，OPAUC と TPAUC が 0.7 未満と差があった．正則化項の存在の有無の影響によると考えられるが，繰り返し学習した MPAUC のパラメータは近似 AR 値，最尤法に近づいていた．近似 AR 値の 23 番目以降のパラメータと，入替りはあるが最尤法のパラメータはゼロになるのに対し，ほかの 3 手法のパラメータはゼロにはならなかった．ちなみに，22 番目は有利子負債比率，26 番目は EBITDA（対数）であった．

**図表 1：Somers' D による手法間の精度比較（2007～15 年の平均値±標準偏差）**

| モデル構築手法 | Somers' D 2007-15 年 | |
|---|---|---|
| | 訓練 | テスト |
| 近似 AR 値最大化 | .834±.011 | .816±.011 |
| 最尤法順序ロジットモデル | .832±.009 | .818±.009 |
| 拡張 OPAUC 手法 | .818±.007 | .807±.009 |
| 拡張 TPAUC 手法 | .821±.007 | .812±.008 |
| 拡張 MPAUC 手法（$U$=22） | .827±.007 | .814±.007 |




**図表 2：拡張 OPAUC ハイパーパラメータ設定値の影響（2007～15 年×10 セットの平均値）**




**図表 3：2015 年の係数パラメータ比較（ノルムは 1 に調整．近似 AR 値最大化におけるパラメータの降順でソート．）**

## 4. おわりに

本稿では，大規模データにおいて直接的に AUC 等の序列精度を最大化することが可能な OPAUC 手法を取り上げ，順序のある多値分類問題へ適用可能とし，さらに実務的に必要となる幾つかの点の拡張に貢献した（拡張 OPAUC）. また, オンライン学習の範疇からは外れるものの, One-Pass から Two-Pass あるいは Multi-Pass とすることで，大幅な精度向上を果たすことがわかった．ただし，いずれにせよ今回利用した格付モデル構築用の小サンプルデータにおいては, その精度は従来手法を越えることはできなかった. この点については, 別のサンプルにおいても試行してみる必要があるが，本稿で用いた損失関数はヒンジ型とは異なる最小二乗型損失であるがゆえに 2 次関数であるため, 一定値を超えてくると本来の損失はゼロであるにもかかわらず, 増加してしまうというデメリットがある．このあたりに問題はないのか，再考してみる必要もあろう.

今後の課題であるが，まずは Ding（2015）で提案されたステップ幅をサンプル番号 $t$ により変化させる手法（Adaptive Gradient Method）を導入したい．実際，Ding（2015）が提案した Adaptive Online AUC Maximization（AdaOAM）手法は，OPAUC 以上の精度が得られたと報告されている．さらに，ハイパーパラメータの探索においては，ベイズ推計的なアプローチを導入することで，さらなる計算速度の向上を図りたい.

## 参考文献


[1] 木村和央 (2015). ロジットモデル構築における Weight of Evidence を用いた変数変換と欠損値処理方法の提案. 2015SAS ユーザー総会論文集, 211-220.

[2] 木村和央 (2016). デフォルトと複数格付の同時推計によるロジットモデル構築手法の提案. 2016SAS ユーザー総会論文集, 85-99.

[3] 森内一朗, 木村和央 (2009). 格付モデルの構築と検証. 「金融工学とリスクマネジメント高度化」研究会公開資料, http://www.ffr-plus.jp/material/pdf/100913/kinkoken.pdf （2017 年 6 月 21 日）.

[4] 山下智志, 三浦翔 (2011). 信用リスクモデルの予測精度－AR 値と評価指標－. 朝倉書店.

[5] Ding, Y., Zhao, P., Hoi, S.C.H., Ong, Y-S. (2015). An Adaptive Gradient Method for Online AUC Maximization. The 29th AAAI Conference on Artificial Intelligence.

[6] Gao, W., Wang, L., Jin, R., Zhu, S., and Zhou, Z.-H. (2016). One-Pass AUC Optimization. Artificial Intelligence, 236, 1-29.

# 付録 A.　SAS/IML による拡張 OPAUC / TPAUC / MPAUC のコーディング

本稿の数値実験で用いた拡張 OPAUC / TPAUC / MPAUC のための SAS プログラムを掲載した．コーディング開発当初は，汎用性を重視して DATA ステップを用いていたが，行列演算速度に決定的な違いがあるように思われ，SAS/IML の利用に切り替えた．IML を利用するがゆえに，コーディングにおいては以下の点に留意する必要があり，場合によっては改善の余地もある．

まず，IML は 2 次元の配列（つまり行列）までしか扱えない点に苦心した．本稿では複数クラス分類問題への拡張を取り上げているが，たとえば，説明変数間の共分散行列をクラスごとに準備する必要がある．しかし，共分散行列それ自体が既に行列であり，クラス別の定義となると，通常はその変数名にクラス番号を付加して区別するという発想になる．このため，どのクラスの共分散行列に対してアクセスするか条件文を用いて制御することになり，そのためにプログラムが冗長になってしまったことは否めない．なお，横にクラスの数だけ共分散行列を連結した拡大行列にて処理するプログラムも試行してみたが，本稿掲載のものより計算時間を要した．拡大行列の一部にインデックスを用いてアクセスするよりも，条件文を多用しつつも行列自体を分割して扱うほうが有利ということが示唆された．

次に，メインである PROC IML の開始直後で，「read all var{…} into …」と SAS データセットから全レコードのデータをメモリに読み込んでいるが，大規模データセットにおいては，この方法ではメモリに収まらない可能性がある．これについては，もとのデータを保持する必要はないため，数回に分割して読み込み処理をすれば問題ない．ただし，分割しすぎると SAS データセットから IML 側への変換処理が多く発生し，それは計算時間に不利に働くと考えられるため，最適な分割を利用 PC の CPU，メモリ等の状況から見定める必要があろう[8]．

実際のプログラムは，事前処理として学習データをランダムソートして 10 パターン生成，事後処理として学習データ，テストデータ双方で Somers' D を計算するといったことが含まれるが割愛し，コアとなるデータを与えて係数パラメータを得る部分のみを掲載した．ただし，メインである PROC IML の前に，説明変数名リストおよびその頭に B を付加したパラメータ名リストの作成，マクロ変数への登録（CALL SYMPUT）に関するプログラムを後続の処理で呼び出しているため，参考までに掲載した．

**≪入力データ定義≫**

| | | |
|---|---|---|
| _in | 入力データセット名 | 以下のクラス変数データ，説明変数データ，サンプルの重み変数データを含む． |
| _y | クラス変数名 $y$ | 0, 1, ･･･, &_r の整数値をとるが，欠番があってもよい． |
| _xv | 説明変数名 $\mathbf{x}$ | 空白区切りで列挙（x1 x2 ･･･ x&_d），またはハイフン指定も可能（x1-x&_d）． |
| _wt | サンプルの重み変数名 $\omega$ | 0 以上の実数値を指定可能． |
| _r | クラス変数の最大値 | クラスは 0 から指定可能なので，クラス数は&_r + 1 個（欠番なしの場合）． |
| _d | 説明変数の数 | 説明変数の数を指定． |
| _con | 制約条件のある説明変数の数 | 説明変数の先頭から&_con 番目まで，制約条件を考慮． |
| _lambda | 正則化パラメータ $\lambda$ | ハイパーパラメータ．大きくすると，ペナルティがかかり，$\boldsymbol{w}$ の分散が減少． |
| _eta | ステップ幅 $\eta$ | ハイパーパラメータ．本稿では一定値． |
| _u | 反復回数 $U$ | 1 : One-Pass，2 : Two-Pass，3，… : Multi-Pass |

**≪出力データ定義≫**

| | | |
|---|---|---|
| _out | 出力データセット名 | 係数パラメータ $\boldsymbol{w}$ を含む．変数名は，説明変数名の頭に B を付加したもの． |

[8] SAS 社テクニカルサポート FAQ「部分行列を利用して行列演算を行なう方法」を参考にされたい．
https://www.sas.com/offices/asiapacific/japan/service/technical/faq/list/body/iml004.html

```
/******** OPAUC / TPAUC マクロ ********/

%macro opauc(_in, _out, _y, _xv, _wt, _r, _d, _con, _lambda, _eta, _u);

/****** 個々の説明　変数名称をマクロ変数へ格納 ******/
DATA _NULL_;
  SET &_in (OBS=1);

  /** 説明　変数パラメータリストの変数定義と初期化 **/
  ATTRIB VNAME LENGTH = $128.;            /* 一時変数 */
  ATTRIB VNAMS LENGTH = $4096.;           /* 説明　変数 */
  ATTRIB BNAMS LENGTH = $4096.;           /* パラメータ */
  VNAMS = ''; BNAMS = '';

  /** 説明　変数配列化 **/
  ARRAY _AX[&_d ] &_xv;

  /** 個々の説明　変数を取得後に連結してリスト化 **/
  DO _I = 1 TO &_d;
    CALL VNAME(_AX[_I],VNAME);
    VNAMS = TRIM(VNAMS)||' '||TRIM(VNAME);
    BNAMS = TRIM(BNAMS)||' '||'B'||TRIM(VNAME);
  END;

  /** 説明　変数パラメータリストをマクロ変数へ格納 **/
  CALL SYMPUT('_vxs',TRIM(VNAMS));
  CALL SYMPUT('_bxs',TRIM(BNAMS));

RUN;

PROC IML;

  /****** IML 配列へサンプルデータを格納 ******/
  use &_in;                       /* 入力データセット */
  read all var{&_y } into AY;     /* クラス */
  read all var{&_vxs } into AX;   /* 説明変数 */
  read all var{&_wt } into AU;    /* 各データの重み */

  /****** 制約条件判定用配列定義 ******/
  DX = {[&_con ] 0 [%eval(&_d - &_con)] 9999999};
          /* 頭から&_con 個の変数は 零 他は +∞* /

  /****** 初期化処理 ******/

  /**** パラメータ（w）配列をゼロクリアして定義 ****/
  BW = j(1,&_d,0);                /* &_d 列の零ベクトル */

  /**** 重み総和をゼロクリアして定義 ****/
  _TS = 0;

  /**** クラス別配列をゼロクリアして定義 ****/
  %macro a1;
    %do i = 0 %to &_r;

      /** クラス別重み合計（&_r+1 クラス分準備） **/
      BT&i = 0;                   /* スカラーの零 * /
      /** 平均ベクトル（&_r+1 クラス分準備）**/
      BC&i = j(1,&_d,0);          /* &_d 列の零ベクトル */
      /** 共分散行列（&_r+1 クラス×&_r+1 クラス分準備） **/
      BS&i = j(&_d,&_d ,0);       /* &_d×&_d の零行列 */

    %end;
  %mend a1; %a1;

  /****** メイン処理 ******/
  DO _U = 1 TO &_u;               /* 反復回数ループ */
  DO _T = 1 TO nrow(AY);          /* nrow(AY) = データ数 */

    /**** 1 回目かつ重みがゼロでないときのみ処理実行 ****/
    IF _U = 1 & AU[_T] > 0 THEN DO;

      /**** 各種データ更新 ****/
      %macro a2;
        %do i = 0 %to &_r;
          %if &i > 0 %then %do;
            ELSE                  /* 現データのクラスと　 */
          %end;                   /* 一致するときのみ計算 */
          IF AY[_T] = &i THEN DO;

            /** 重み総和 **/
            _TS = _TS + AU[_T];
            /** クラス別重み合計 **/
            BT&i = BT&i + AU[_T];
            /** 平均ベクトル**/
            AC = BC&i;
            BC&i = AC + (AX[_T,]-AC)/BT&i *AU[_T];
            /** 共分散行列 **/
            AS = BS&i;
            BS&i = AS + t(AC)*AC - t(BC&i )*BC&i
          + (t(AX[_T,])*AX[_T,] - AS - t(AC)*AC)/BT&i *AU[_T];

          END;
        %end;
      %mend a2; %a2;
    END;

    /****** 勾配計算用配列をゼロクリアして定義 ******/
    AG = j(1,&_d,0);              /* &_d 列の零ベクトル */

    /****** 損失関数の勾配計算 ******/
    %macro a3;
      %do i = 0 %to &_r;                  /*　 現クラス */
        IF AY[_T] ^= &i THEN DO;          /* ←以外計算 */
          IF BT&i > 0 THEN DO;            /* データが存在するクラス */
                                          /* のみを処理 */
            /** 現データのクラスとの比較 で符号をセット **/
            IF &i < AY[_T] THEN _SGN = 1;
            ELSE
            IF &i > AY[_T] THEN _SGN = -1;

            /** クラス別重み合計比率を計算 **/
            _WW = (BT&i /(_TS - BT&i )) * AU[_T];

            /** X－平均ベクトルの計算 **/
            CX = AX[_T,] - BC&i;

            /** クラス別の勾配計算 **/       /* ↓正則化項 */
            AG = AG + &_lambda * BW * _WW;
            AG = AG - _SGN * CX * _WW;      /* ←符号付項 */
            AG = AG + BW * (t(CX)*CX + BS&i ) * _WW;

          END;
        END;
      %end;
    %mend a3; %a3;

    /****** 制約条件を考慮した係数パラメータの更新 ******/
    BW = (BW - &_eta *AG) >< DX;
                    /* ><は，要素ごとに小なるほうを採用 */
  END;
  END;

  /****** 最終結果出力 ******/
  vname = {&_bxs };
  create &_out from BW[colname=vname];
  append from BW;
  close &_out;

QUIT;
%mend opauc;
```

# 付録 B.　NLMIXED プロシジャによる近似 AR 値最大化手法のコーディング

山下・三浦（2011）により提案されたシグモイド関数を用いて近似した AR 値を最大化するための SAS プログラムを掲載した．入出力変数は付録 A と同様である．「クラスの異なる全組合せデータの作成」が必要であり，組合せが 100 万件レコードを超えるとメモリに入りきらず，NLMIXED プロシジャが動かないといった問題が発生した．本稿の数値実験は約 40 万件であり，かなりの時間を要したが，計算結果は出力された．

```
/******** Sigmoid 近似 AR 値最大化マクロ ********/

%macro sigmoid(_in, _out, _y, _xv, _wt, _r, _d, _con);

  /****** 個々の説明変数名称をマクロ変数へ格納 ******/
  DATA _NULL_;
    SET &_in (OBS=1);

    /** 説明変数名称の一時保存用変数 **/
    ATTRIB VNAME LENGTH = $128.;

    /** 説明変数を配列化 **/
    ARRAY _AX[&_d ] &_xv;

    /** マクロ変数へ格納 **/
    DO _I = 1 TO &_d;
      CALL VNAME(_AX[_I],VNAME);
      CALL SYMPUT('_vx'||COMPRESS(PUT(_I,8.)),
                                  TRIM(VNAME));
    END;
  RUN;

  /****** クラスの異なる全組合せデータの作成 ******/
  %macro a5;
    PROC SQL;
      CREATE TABLE _TBL0 AS
      SELECT
        DA.&_y AS &_y._A,
        DB.&_y AS &_y._B,
        /** 2 つのサンプルの説明変数の差（dx）を計算 **/
        %do i = 1 %to &_d;
          (DA.&&_vx&i -DB.&&_vx&i ) AS &&_vx&i,
        %end;

        (DA.&_wt * DB.&_wt ) AS &_wt,          /* 重み */
        (DA.&_wt * DB.&_wt ) > 0 AS U          /* 分析対象F */
      FROM &_in AS DA, &_in AS DB
      WHERE &_y._A < &_y._B AND                /* 外積をとる */
        DA.&_wt > 0 AND DB.&_wt > 0            /* がクラスの */
      ORDER BY &_y._A, &_y._B;                 /* 大小関係を */
    QUIT;                                      /* 一方向限定 */
  %mend a5; %a5;

  /****** NLMIXED による Sigmoid 近似 AR 値最大化 ******/

  /**** ODS ファイル出力設定開始 ****/
  ODS OUTPUT AdditionalEstimates = _TBL1;

  /**** Sigmoid 近似 AR 値最大化ルーチン****/
  PROC NLMIXED DATA = _TBL0
    MAXITER = 10000 MAXFUNC = 10000;

    /** パラメータ定義と初期値ゼロを設定 **/
    PARMS
    %macro a6;
      %do i = 1 %to &_d;
        %if &i > 1 %then %do;      /* PARMS では          */
          ,                        /* 区切りに , が必要   */
        %end;
        B&&_vx&i = 0               /* パラメータの変数名は */
      %end;                        /* 説明変数名の頭に B  */
    %mend a6; %a6; ;

    /** 制約条件設定 **/
    BOUNDS
    %macro a7;
      %do i = 1 %to &_con;
        %if &i > 1 %then %do;      /* BOUNDS では         */
          ,                        /* 区切りに , が必要   */
        %end;
        B&&_vx&i >= 0              /* 説明変数の先頭から */
      %end;                        /* &_con の個数まで    */
    %mend a7; %a7; ;               /* 正値に制約          */

    /** 説明変数を配列化 **/
    ARRAY _AX[*] &_xv;

    /** パラメータを配列化 **/
    ARRAY _AB[*]
    %macro a8;
      %do i = 1 %to &_d;
        B&&_vx&i
      %end;
    %mend a8; %a8; ;

    /** Z スコア（Σw×dx）計算 **/
    _BX = 0;
    DO _J = 1 TO &_d;
      _BX = _BX + _AB[_J]*_AX[_J];
    END;

    Z = MAX(MIN(_BX,20),-20);      /* 次の Sigmoid-AR     */
                                   /* の発散を防止        */
    /** Sigmoid-AR 計算（sigma=0.02） **/
    SG = (2*1/(1+EXP(-Z/0.02))-1)*&_wt;

    /** Sigmoid-AR 最大化 **/
    MODEL U~GENERAL(SG);           /* U：分析対象F=1  */

    /** データ出力調整 **/
    %macro a9;
      %do i = 1 %to &_d;
        ESTIMATE "B&&_vx&i" B&&_vx&i;
      %end;
    %mend a9; %a9;

  RUN;

  /**** ODS ファイル出力設定終了 ****/
  ODS OUTPUT CLOSE;

  /****** パラメータの横展開 ******/
  PROC TRANSPOSE DATA = _TBL1
    OUT = &_out (DROP=_NAME_);
    VAR Estimate;
    ID Label;
  RUN;
%mend sigmoid;
```

# 産業連関表を利用した地域経済の域外輸出の活性化要因の探索

○栃澤欣之，橋本綾子，角元慶二，菱垣晴次

（大塚製薬株式会社）

# Discovery and estimate of the effects for activators of regional economy using interindustry table

Yoshiyuki Tochizawa, Ayako Hashimoto, Keiji Kakumoto, Haretsugu Hishigaki
Otsuka Pharmaceutical Co., Ltd.

**要旨**

地域経済およびその雇用の活性化においては、域外からの資金の流入が重要である（経済産業省）。そのため本研究では地域の域外輸出の増加を目的とした。まず、本研究では徳島県をモデルケースとし、産業連関表を利用して徳島県の移輸出産業を分析した。次に、徳島県の移輸出産業の移輸出を活性化させる中間投入を探索するため、移輸出と付随して変動する中間投入を中四国の産業連関表より解析した。また、このような移輸出の活性化と関連してその経済波及効果を推定するために産業連関分析を行った。産業連関分析は、経済波及効果に関する分析として、移輸出額を与えた場合の生産額の将来予測にも用いられている。最後に本研究の結果を、統計的因果推論(宮川, 2004)を参考にして考察した。

キーワード：地域経済、産業連関表、産業連関分析、域外輸出、経済波及効果、徳島県、中四国

## 1. 序論

地域経済は図 1 の模式図で表現できる(経済産業省)。①地域外を主な市場とする「域外市場産業（製造業、農業、観光）」と②地域内を主な市場とする「域内市場産業（日用品小売業、対個人サービス業)」に分けて考えることができる。資金の流れで両産業を模式的に捉えると、①域外市場産業が域外から資金を稼ぎ（第１フロー)、②雇用者所得として域内住民の消費活動元手となり(第２フロー)、③元手をもとに域内市場産業で消費され（第３フロー)、④域内市場産業が発展し、域内市場産業で働く住民の雇用。所得につながり、さらなる域内需要の拡大、域内市場の発展につながる（第４フロー)。これら４つの資金の流れを円滑に促すことが地域経済成長の鍵になる。域外から資金を流入させる域外市場産業は、地域経済の心臓部とも言え、域外から資金を稼いでくる産業の集積を促進し、競争力を強化することが重要と考えられている。

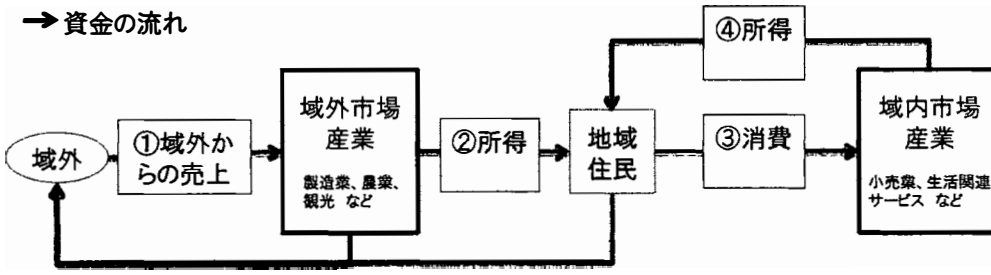



図 1　地域経済

地域経済にとって重要な域外から資金を稼いでくる中核産業の主な基準は以下の 5 つである。

1. 産業の規模が大きい。
2. 域外から資金を流入させる。
3. 産業集積していて労働生産性が高い。
4. 地域内への波及効果が大きい(地域外に流出しない)。
5. 設備投資を積極的である。

更に、地域の持続的成長を実現するには、以下の 5 つの視点で評価する。

1. 域外から資金を流入させる。
2. 域外市場産業で生まれた付加価値が域内に入る。
3. 再投資が域内で行われる。
4. 域外市場産業が持続的である。
5. 域内市場産業が所得を生み出している。

要約すると、域外市場産業が域外から資金を流入させ、その資金を域内市場産業が好循環させることが重要である。以上に基づき、域外市場産業を支援することは、地域経済およびその雇用の活性化において、有用と考えられている。そこで、本研究では移輸出を増大させる要因を探索するため、移輸出に付随し変動する中間投入の探索を行った。

本研究では、徳島県をモデルケースとし、産業連関表を利用して徳島県の移輸出産業を分析した。次に、中四国である香川県・高知県・岡山県・徳島県・鳥取県・山口県の産業連関表を利用して、徳島県の移輸出産業の移輸出を活性化させる中間投入を探索するため、移輸出と付随して変動する中間投入を解析した。また、このような移輸出の活性化と関連してその経済波及効果を推定するために産業連関分析を行った。産業連関分析は、経済波及効果に関する分析として、移輸出額を与えた場合の生産額の将来予測にも用いられている。最後に本研究の結果を、統計的因果推論(宮川, 2004)を参考にして考察した。

## 2. 解析内容

1　徳島県の移輸出産業を分析した。この時に最新年である 2011 年の 108 部門による産業連関表を利用した。

1.1　「移輸出計(各産業)/移輸出計(合計)」を「移輸出構成」として求めた。これにより、徳島県の移輸出を支える産業を示した。

1.2　「移輸出計(各産業)/県内生産額(各産業)」を「移輸出比率」として求めた。これにより、徳島県の移輸出への生産割合の高い産業を示した。

1.3　「移輸出計(各産業)×移輸出比率」を「成長効率」として求めた(表 1)。表 1 には、上位 20 位までを示した。これにより、徳島県の各産業を一定の割合で成長させた場合の移輸出の増大量を推定した。成長効率の上位 10 位の産業は、医薬品、電子デバイス、その他の電気機械、無機化学工業製品、食料品、電力、はん用機械、パルプ・紙・板紙・加工紙、耕種農業、飲料であった。

2　先の分析から得られた 10 産業の内、医薬品、無機化学工業製品、食料品、電力、はん用機械、耕種農業、飲料の 7 産業に関し、移輸出を活性化する中間投入の探索を行った。そのために、中四国の各県の産業連関表を利用した。また、以下では県内産業を総合した合計についても同様の解析を行った。

2.1　2000-2005-2011 年の産業連関表の集計上、電子デバイス、その他の電気機械、パルプ・紙・板紙・加工紙の 3 産業は 2005 年以前の産業連関表と対応していなかったため、残りの 7 産業に関し、移輸出を活性化する中間投入の探索を行った。その対応関係に関しては、表 2 とした。

2.2　中四国の各県の内、愛媛県、広島県、島根県は他の都道府県と産業連関表の区分が対応していなかったため、香川県・高知県・岡山県・徳島県・鳥取県・山口県の 6 県の産業連関表を利用した。利用した産業連関表は、2000 年の 104 部門、2005 年の 108 部門、2011 年の 108 部門によるものを利用した。この 6 県の 7 産業に関しても移輸出比率を確認した(表 3)。

2.3　各 7 産業に関し 2005-2011 年の移輸出の変動と、各 7 産業の 108 部門の中間投入および雇用者所得に関する変動との回帰分析を行った(表 4)。2005 年と 2011 年において、産業区分が対応していなかったものは除外して解析した。

2.4　同様に、各 7 産業に関し 2000-2005 年の移輸出の変動と、各 7 産業の 104 部門の中間投入および雇用者所得に関する変動との回帰分析を行った(表 4)。2000 年と 2005 年において、産業区分が対応していなかったものは除外して解析した。

2.5　上述の 2005-2011 年の相関係数の p 値(p1)を両側検定により算出し、2000-2005 年の相関係数の p 値(p2)を 2005-2011 年の相関係数と同方向の片側検定により算出した(表 4)。この時の有意水準は 0.05 とした。表 4 は、p1 と p2 が 0.05 以下となった中間投入について示した。

3　医薬品、無機化学工業製品、食料品、電力、はん用機械、耕種農業、飲料の 7 産業に関し、2011

年の徳島県の産業連関表に基づいた産業連関分析により移輸出の経済波及効果を推定した(表5)。移輸出の経済波及効果は逆行列係数と同値であるため、各産業の逆行列係数の列に各産業の移輸出の成長効率をかけることで経済波及効果の程度を推定できる。表 5 では、県内生産額が 1%増加した場合の移輸出による経済波及効果を推定した結果を示した。

4　徳島県の県外への移輸出は、輸出を含めた国内の需要と相関があることが考えられるため、国内生産額と徳島県の移輸出とで回帰分析を行った。この時、国内生産額を、国内の需要としてみなした。また、ほとんど移輸出を行っていない産業に関しては、県外移輸出と国内の需要との相関は見られないため、徳島県の産業の移輸出金額で上位 30 位の産業に関し回帰分析を実行した。この時の 30 産業は 2000- 2005- 2011 年の産業連関表で対応しているものに限定した中での上位 30 産業である。

4.1　上記の 30 産業に関し、2000 年から 2005 年の国内生産額の変動比率(国内生産額の変動額/2000 年の国内生産額)と、2000 年から 2005 年の徳島県の移輸出の変動比率(徳島県の移輸出の変動額/2000 年の徳島県の移輸出)とで回帰分析を行った(図 2)。

4.2　上記の 30 産業に関し、2005 年から 2011 年の国内生産額の変動比率(国内生産額の変動額/2005 年の国内生産額)と、2005 年から 2011 年の徳島県の移輸出の変動比率(徳島県の移輸出の変動額/2005 年の徳島県の移輸出)とで回帰分析を行った(図 2)。

5　最後にこれまでの結果を整理して考察した(図 3)。

## 3. 解析結果と考察

徳島県の移輸出産業を分析するため、「成長効率」で上位 30 産業に関し、「移輸出構成」、「移輸出比率」、「成長効率」を表 1 に示した。上位 10 産業で徳島県の移輸出の 65.5%を占めるため、これらの産業に注目し分析した。ここでは、「移輸出計(各産業)/移輸出計(合計)」を「移輸出構成」とし、「移輸出計(各産業)/県内生産額(各産業)」を「移輸出比率」とし、「移輸出計(各産業)×移輸出比率」を「成長効率」として求めた。

表 1　　徳島県の移輸出産業(単位は百万円)

| 産業 | 移輸出計 | 県内生産額 | 移輸出比率 | 移輸出構成 | 成長効率 |
|---|---|---|---|---|---|
| 医薬品 | 357614 | 362875 | 0.986 | 0.176 | 352429 |
| 電子デバイス | 199847 | 200703 | 0.996 | 0.098 | 198995 |
| その他の電気機械 | 131423 | 131471 | 1.000 | 0.065 | 131375 |
| 無機化学工業製品 | 120448 | 125901 | 0.957 | 0.059 | 115231 |
| 食料品 | 136984 | 177024 | 0.774 | 0.067 | 106000 |
| 電力 | 168143 | 271015 | 0.620 | 0.083 | 104319 |
| はん用機械 | 62158 | 75537 | 0.823 | 0.031 | 51149 |
| パルプ・紙・板紙・加工紙 | 64381 | 82857 | 0.777 | 0.032 | 50025 |
| 耕種農業 | 55375 | 74400 | 0.744 | 0.027 | 41215 |
| 飲料 | 33839 | 36069 | 0.938 | 0.017 | 31747 |
| プラスチック製品 | 37370 | 45285 | 0.825 | 0.018 | 30838 |
| 生産用機械 | 35718 | 44459 | 0.803 | 0.018 | 28696 |
| 紙加工品 | 35975 | 45375 | 0.793 | 0.018 | 28522 |
| 化学最終製品(医薬品を除く。) | 33883 | 41967 | 0.807 | 0.017 | 27356 |
| 銑鉄・粗鋼 | 26280 | 26280 | 1.000 | 0.013 | 26280 |
| 商業 | 86795 | 322290 | 0.269 | 0.043 | 23375 |
| 電子応用装置・電気計測器 | 18482 | 18503 | 0.999 | 0.009 | 18461 |
| 家具・装備品 | 19869 | 22165 | 0.896 | 0.010 | 17811 |
| その他の電子部品 | 18098 | 18518 | 0.977 | 0.009 | 17688 |
| 木材・木製品 | 22912 | 31312 | 0.732 | 0.011 | 16765 |
| その他の製造工業製品 | 17959 | 19501 | 0.921 | 0.009 | 16539 |
| 道路輸送(自家輸送を除く。) | 42057 | 107803 | 0.390 | 0.021 | 16408 |
| 衣服・その他の繊維既製品 | 16111 | 16756 | 0.962 | 0.008 | 15491 |
| 有機化学工業製品(石油化学基礎製品を除く。) | 16654 | 18634 | 0.894 | 0.008 | 14884 |
| 繊維工業製品 | 15857 | 18096 | 0.876 | 0.008 | 13895 |
| その他の金属製品 | 16528 | 23993 | 0.689 | 0.008 | 11386 |
| 運輸附帯サービス | 14505 | 19154 | 0.757 | 0.007 | 10984 |
| 宿泊業 | 16936 | 27703 | 0.611 | 0.008 | 10354 |
| 船舶・同修理 | 12516 | 15408 | 0.812 | 0.006 | 10167 |
| 水運 | 14476 | 22179 | 0.653 | 0.007 | 9448 |
| ゴム製品 | 9962 | 10630 | 0.937 | 0.005 | 9336 |
| 合成樹脂 | 9355 | 10066 | 0.929 | 0.005 | 8694 |
| 漁業 | 8111 | 10110 | 0.802 | 0.004 | 6507 |
| 建設・建築用金属製品 | 12314 | 24029 | 0.512 | 0.006 | 6310 |
| 林業 | 7889 | 14021 | 0.563 | 0.004 | 4439 |
| 畜産 | 9789 | 24645 | 0.397 | 0.005 | 3888 |
| セメント・セメント製品 | 7870 | 16094 | 0.489 | 0.004 | 3848 |
| 飼料・有機質肥料(別掲を除く。) | 4016 | 4538 | 0.885 | 0.002 | 3554 |
| 鋼材 | 2895 | 2898 | 0.999 | 0.001 | 2892 |
| 産業用電気機器 | 4960 | 8941 | 0.555 | 0.002 | 2752 |

しかし、2000-2005-2011 年の産業連関表では対応していない産業もあった。2000-2005-2011 年の産業連関表の対応を表 2 のような関係として以降の解析は行った。先述の徳島県の移輸出 10 産業の内 2000-2005-2011 年で産業連関表に対応のあった、医薬品、無機化学工業製品、食料品、電力、はん用機械、耕種農業、飲料の 7 産業に関し、移輸出を活性化する中間投入の探索分析を行った。

表 2　　産業連関表の分類の対応表

| 2005年版以前の分類名 | 2011年版の分類 |
|---|---|
| 飼料・有機質肥料（除別掲） | 飼料・有機質肥料（別掲を除く。） |
| 製材・木製品 | 木材・木製品 |
| 有機化学工業製品（除石油化学基礎製品） | 有機化学工業製品（石油化学基礎製品を除く。） |
| 化学最終製品（除医薬品） | 化学最終製品（医薬品を除く。） |
| 一般産業機械 | はん用機械 |
| 特殊産業機械 | 生産用機械 |
| 事務用・サービス用機器 | 業務用機械 |
| 電子計算機・同付属装置 | 電子計算機・同附属装置 |
| 道路輸送（除自家輸送） | 道路輸送（自家輸送を除く。） |
| 運輸付帯サービス | 運輸附帯サービス |
| インターネット付随サービス | インターネット附随サービス |
| 映像・文字情報制作 | 映像・音声・文字情報制作 |
| 社会保障 | 社会保険・社会福祉 |
| その他の公共サービス | その他の非営利団体サービス |
| 自動車・機械修理 | 自動車整備・機械修理 |
| 飲食店 | 飲食サービス |
| 家計外消費支出 | 家計外消費支出（行） |
| 間接税（除関税） | 間接税（関税・輸入品商品税を除く。） |
| 飼料･有機質肥料（除別掲） | 飼料・有機質肥料（別掲を除く。） |
| 出版・印刷 | 印刷・製版・製本 |
| 無機化学基礎製品 | 無機化学工業製品 |
| 有機化学基礎製品 | 石油化学基礎製品 |
| 有機化学製品 | 有機化学工業製品（石油化学基礎製品を除く。） |
| 化学最終製品□除医薬品） | 化学最終製品（医薬品を除く。） |
| その他の窯業・由石製品 | その他の窯業・土石製品 |
| 民生用電子・電気機器 | 民生用電気機器 |
| 通信機械 | 通信機械・同関連機器 |
| 電子応用装置・電気計測器 | 電子応用装置・電気計測器 |
| 再生資源回収・加工処理 | 再生資源回収・加工処理 |
| 道路輸送 | 道路輸送（自家輸送を除く。） |
| 貨物運送取扱 | 貨物利用運送 |
| 旅館・その他の宿泊所 | 宿泊業 |
| 家計外消費支出（行） | 家計外消費支出（行） |
| 資本減耗引当□社会資本等減耗分） | 資本減耗引当（社会資本等減耗分） |
| 間接税□除関税・輸入品商品税） | 間接税（関税・輸入品商品税を除く。） |

徳島県の移輸出産業でかつ 2000-2005-2011 年の産業連関表で対応のあった 7 産業に関し、今回集計した香川県・高知県・岡山県・徳島県・鳥取県・山口県の 6 県における移輸出比率に関して確認した(表 3)。表 3 より、医薬品、無機化学工業製品、食料品、電力、はん用機械、耕種農業、飲料の移輸出比率は 0.921, 0.839, 0.708, 0.304, 0.845, 0.599, 0.838 であり、全体の移輸出比率 0.408 と比較すると電力以外の 6 産業においては、中四国においても移輸出産業であることを確認した。逆に、徳島県では電力が移輸出産業となっていることが興味深い結果となった。

表 3　　各産業の移輸出比率(単位は百万円)

| 医薬品 | 香川 | 高知 | 岡山 | 徳島 | 鳥取 | 山口 | 計 |
|---|---|---|---|---|---|---|---|
| 移輸出 | 65137 | 0 | 5958.1 | 357614 | 0 | 311624 | 740333.1 |
| 県内生産額 | 65908 | 0 | 59413.9 | 362875 | 22 | 315538 | 803756.9 |
| 移輸出比率 | 0.988 | 0.000 | 0.100 | 0.986 | 0.000 | 0.988 | 0.921 |
| 無機化学工業 | 香川 | 高知 | 岡山 | 徳島 | 鳥取 | 山口 | 計 |
| 移輸出 | 27992 | 4085 | 75542.3 | 120448 | 918 | 137502 | 366487.3 |
| 県内生産額 | 32822 | 4085 | 97073.4 | 125901 | 1030 | 175969 | 436880.4 |
| 移輸出比率 | 0.853 | 1.000 | 0.778 | 0.957 | 0.891 | 0.781 | 0.839 |
| 食料品 | 香川 | 高知 | 岡山 | 徳島 | 鳥取 | 山口 | 計 |
| 移輸出 | 215573 | 63852 | 330657.6 | 136984 | 132435 | 142883 | 1022384.6 |
| 県内生産額 | 292153 | 111202 | 461409.7 | 177024 | 178276 | 223790 | 1443854.7 |
| 移輸出比率 | 0.738 | 0.574 | 0.717 | 0.774 | 0.743 | 0.638 | 0.708 |
| 電力 | 香川 | 高知 | 岡山 | 徳島 | 鳥取 | 山口 | 計 |
| 移輸出 | 10024 | 107 | 44153.6 | 168143 | 498 | 144626 | 367551.6 |
| 県内生産額 | 129341 | 64538 | 234254.8 | 271015 | 66650 | 442617 | 1208415.8 |
| 移輸出比率 | 0.078 | 0.002 | 0.188 | 0.620 | 0.007 | 0.327 | 0.304 |
| はん用機械 | 香川 | 高知 | 岡山 | 徳島 | 鳥取 | 山口 | 計 |
| 移輸出 | 56521 | 8324 | 174497.2 | 62158 | 2463 | 84758 | 388721.2 |
| 県内生産額 | 88465 | 8419 | 186049 | 75537 | 3032 | 98370 | 459872 |
| 移輸出比率 | 0.639 | 0.989 | 0.938 | 0.823 | 0.812 | 0.862 | 0.845 |
| 耕種農業 | 香川 | 高知 | 岡山 | 徳島 | 鳥取 | 山口 | 計 |
| 移輸出 | 34134 | 69058 | 32614.3 | 55375 | 26630 | 18451 | 236262.3 |
| 県内生産額 | 52482 | 86658 | 85485.4 | 74400 | 43158 | 52100 | 394283.4 |
| 移輸出比率 | 0.650 | 0.797 | 0.382 | 0.744 | 0.617 | 0.354 | 0.599 |
| 飲料 | 香川 | 高知 | 岡山 | 徳島 | 鳥取 | 山口 | 計 |
| 移輸出 | 1183 | 7629 | 125752.8 | 33839 | 1914 | 8686 | 179003.8 |
| 県内生産額 | 1901 | 10846 | 143813.1 | 36069 | 4123 | 16863 | 213615.1 |
| 移輸出比率 | 0.622 | 0.703 | 0.874 | 0.938 | 0.464 | 0.515 | 0.838 |
| 内生部門計 | 香川 | 高知 | 岡山 | 徳島 | 鳥取 | 山口 | 計 |
| 移輸出 | 2598419 | 674329 | 7820476.3 | 2031135 | 840687 | 6098997 | 20064043.3 |
| 県内生産額 | 7359204 | 3741965 | 16910660.8 | 5075741 | 3251973 | 12894409 | 49233952.8 |
| 移輸出比率 | 0.353 | 0.180 | 0.462 | 0.400 | 0.259 | 0.473 | 0.408 |

次に、各産業の移輸出と中間投入の相関関係を分析した(表 4)。表 4 は、2000-2005 年と 2005-2011 年に関し、移輸出の変動と各 7 産業の 108 部門の中間投入および雇用者所得に関する変動とで回帰分析を行った結果である。回帰係数で検定をしているため大体の項目で相関係数が 0.8 以上である。表 4 より、移輸出の変動に付随して金額が大きく変動している中間投入は、移輸出との変動金額の比率で、医薬品では研究で 14.4%、無機化学工業製品では雇用者所得で 9.4%、はん用機械では雇用者所得と鋼材で 28.4%と 26.8%、耕種農業では農業サービスで 19.2%、産業全体では鋼材で 10.2%と特に大きく、いずれも産業の構造的にも妥当な結果であった。食料品、電力、飲料に関しては相関のある中間投入が検出されなかった。実際に産業を推進する上では、各産業でこれらの中間投入が円滑に行われることが重要と考えられ、その各中間投入額の増減と移輸出額の増減は表 4 に示した比率で相関していることが示された。

表 4　各産業の移輸出と、重要と考えられる中間投入の相関関係(単位は百万円)

| 医薬品の中間投入産業 | 相関係数(2005-2011) | 相関係数(2000-2005) | p1 | p2 | 変動金額の標準偏差(2005-2011) | |
|---|---|---|---|---|---|---|
| | | | | | 中間投入 | 移輸出 |
| 研究 | 0.863 | 0.764 | 0.027 | 0.039 | 8167 | 56608 |
| その他の対事業所サービス | 0.812 | 0.992 | 0.050 | 0.000 | 903 | |
| 自動車整備・機械修理 | 0.869 | 0.864 | 0.025 | 0.013 | 343 | |
| 建設補修 | 0.878 | 0.948 | 0.021 | 0.002 | 162 | |
| 無機化学工業製品の中間投入産業 | 相関係数(2005-2011) | 相関係数(2000-2005) | p1 | p2 | 変動金額の標準偏差(2005-2011) | |
| | | | | | 中間投入 | 移輸出 |
| 雇用者所得 | 0.923 | 0.924 | 0.009 | 0.004 | 1885 | 20056 |
| 自動車整備・機械修理 | 0.939 | 0.840 | 0.005 | 0.018 | 229 | |
| 不動産仲介及び賃貸 | 0.906 | 0.849 | 0.013 | 0.016 | 40 | |
| 食料品の中間投入産業 | 相関係数(2005-2011) | 相関係数(2000-2005) | p1 | p2 | 変動金額の標準偏差(2005-2011) | |
| | | | | | 中間投入 | 移輸出 |
| 該当なし | - | - | - | - | - | 29452 |
| 電力の中間投入産業 | 相関係数(2005-2011) | 相関係数(2000-2005) | p1 | p2 | 変動金額の標準偏差(2005-2011) | |
| | | | | | 中間投入 | 移輸出 |
| 該当なし | - | - | - | - | - | 26796 |
| はん用機械の中間投入産業 | 相関係数(2005-2011) | 相関係数(2000-2005) | p1 | p2 | 変動金額の標準偏差(2005-2011) | |
| | | | | | 中間投入 | 移輸出 |
| 雇用者所得 | 0.923 | 0.987 | 0.009 | 0.000 | 7481 | 26366 |
| 鋼材 | 0.940 | 0.909 | 0.005 | 0.006 | 7066 | |
| 商業 | 0.990 | 0.925 | 0.000 | 0.004 | 1987 | |
| 自家輸送 | 0.880 | 0.952 | 0.021 | 0.002 | 198 | |
| 石油製品 | 0.887 | 0.987 | 0.018 | 0.000 | 156 | |
| その他の非営利団体サービス | 0.971 | 0.869 | 0.001 | 0.012 | 117 | |
| 無機化学工業製品 | 0.966 | 0.746 | 0.002 | 0.044 | 69 | |
| 倉庫 | 0.944 | 0.899 | 0.005 | 0.007 | 51 | |
| 建設・建築用金属製品 | 0.893 | 0.835 | 0.017 | 0.019 | 43 | |
| 貨物利用運送 | 0.913 | 0.900 | 0.011 | 0.007 | 22 | |
| 耕種農業の中間投入産業 | 相関係数(2005-2011) | 相関係数(2000-2005) | p1 | p2 | 変動金額の標準偏差(2005-2011) | |
| | | | | | 中間投入 | 移輸出 |
| 農業サービス | 1.000 | 0.998 | 0.000 | 0.000 | 21886 | 114227 |
| 耕種農業 | 0.995 | 0.994 | 0.000 | 0.000 | 5289 | |
| 畜産 | 0.999 | 0.999 | 0.000 | 0.000 | 1730 | |
| 化学最終製品（医薬品を除く。） | -0.991 | -0.851 | 0.000 | 0.016 | 1032 | |
| ガス・熱供給 | 0.998 | 0.996 | 0.000 | 0.000 | 0 | |
| 飲料の中間投入産業 | 相関係数(2005-2011) | 相関係数(2000-2005) | p1 | p2 | 変動金額の標準偏差(2005-2011) | |
| | | | | | 中間投入 | 移輸出 |
| 該当なし | - | - | - | - | - | 22006 |
| 内生部門計の中間投入産業 | 相関係数(2005-2011) | 相関係数(2000-2005) | p1 | p2 | 変動金額の標準偏差(2005-2011) | |
| | | | | | 中間投入 | 移輸出 |
| 鋼材 | 0.958 | 0.883 | 0.003 | 0.010 | 46506 | 453787 |

p1 は相関係数(2005-2011)の p 値、p2 は相関係数(2000-2005)の p 値を示した。

次に、表 4 の結果に基づいた中間投入により、各産業において一定の移輸出が向上されると想定して、徳島県の各産業の経済波及効果を産業連関分析により推定した(表 5)。表 5 では、県内生産額が 1%増加した場合の経済波及効果を推定している。また、表 5 は徳島県の 2011 年の産業連関表を利用し、2011 年の産業連関が現状を反映していると想定した。

表 5　　経済波及効果の推定(単位は百万円)

| | 耕種農業 | 食料品 | 飲料 | 無機化学工業製品 | 医薬品 | はん用機械 | 電力 |
|---|---|---|---|---|---|---|---|
| 列和 | 1.285 | 1.386 | 1.210 | 1.261 | 1.261 | 1.214 | 1.229 |
| 影響力係数 | 1.030 | 1.111 | 0.970 | 1.011 | 1.011 | 0.973 | 0.985 |
| 1%成長効率 | 412 | 1060 | 317 | 1152 | 3524 | 511 | 1043 |
| 経済波及効果(推定)<br>列和×成長効率 | 529 | 1470 | 384 | 1453 | 4446 | 621 | 1282 |

次に、徳島県の県外への移輸出は、輸出入を含めた国内需要と相関があることが考えられるため、国内生産額の変動比率と徳島県の移輸出の変動比率とで回帰分析を行った(図 1)。この時、国内生産額を、国内需要としてみなした。徳島県の県外への移輸出は、輸出入を含めた国内需要との間で、2000-2005 年と 2005-2011 年のどちらも相関係数の値が約 0.5 で、その p 値については $p<0.01$ であった。

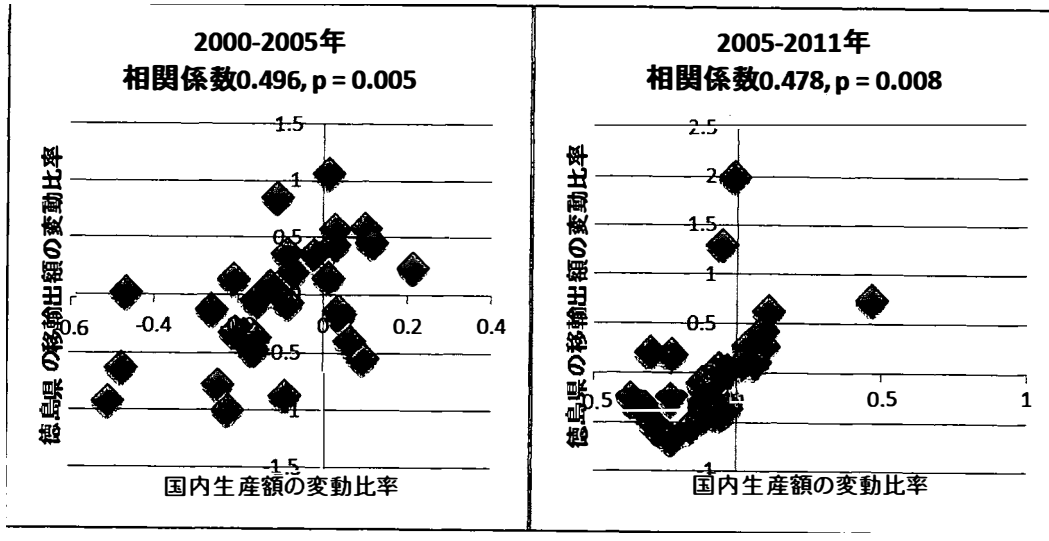



図 2　　国内生産額の変動比率と、徳島県の移輸出の変動比率との相関関係

## 4. 結論および考察

徳島県の主要な移輸出産業は、医薬品、電子デバイス、その他の電気機械、無機化学工業製品、食料品、電力、はん用機械、パルプ・紙・板紙・加工紙、耕種農業、飲料であった。医薬品では研究、無機化学工業製品では雇用者所得、はん用機械では雇用者所得と鋼材、耕種農業では農業サービス、産業全体では鋼材が、移輸出の拡大のために特に大きな金額を要する中間投入であると思われ、その投入金額のスケールは表 4 からおよそ推察することができた。このような中間投入の円滑化により、移輸出が 1%増加することを想定した時の経済波及効果の推定結果は表 5 に示された。また、このような徳島県の移輸出は国内生産との相関において相関係数が約 0.5 であり、地域経済の移輸出は国内需要と相関していることが示された。以上のことから、経済の円滑化のモデルは図 3 のように整理される。図 3 では、国内需要は徳島の移輸出に影響があるが、徳島の移輸出は国内需要にあまり影響がないとした。

図 3 では、徳島の未来の移輸出が、現在の国内需要の総合効果と中間投入の円滑化に依存することを示している。ここから、地域の移輸出の未来を予測および制御する判断材料として、本研究の結果からは国内需要と中間投入への介入の 2 点だけであると考えた。よって、国内の需要を考慮に入れながら中間投入の制御を行うことは、円滑化という間接的な方法しかないものの、地域の移輸出の活性化においては重要な役割を持つと考えられる (図 3 赤矢印)。その円滑化の方針を決定する際には、本研究の表 4 や産業連関表の逆行列係数表などの利用が有用であると考えられる。

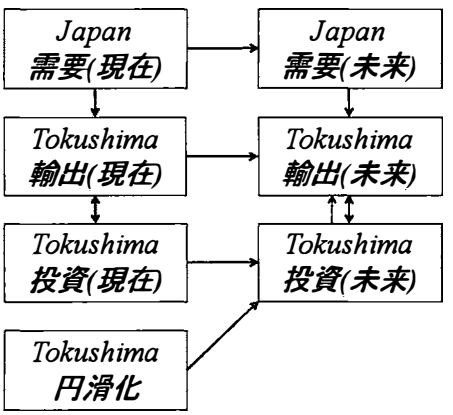



図 3　行政による地域経済円滑化のモデル

### Reference

[1] 経済産業省. 地域経済分析の考え方とポイント. Accessed at 20170531, Available at: <http://www.meti.go.jp/policy/local_economy/bunnseki/point/bunsekipoint.pdf>

[2] 総務省. 第 5 章　産業連関分析のための各種係数の内容と計算方法. Accessed at 20170411, Available at: < http://www.soumu.go.jp/main_content/000290893.pdf>

[3] 総務省. 第 6 章　産業連関分析の方法. Accessed at 20170411, Available at: < http://www.soumu.go.jp/main_content/000290979.pdf>

[4] 宮川 雅巳 (2004) 統計的因果推論—回帰分析の新しい枠組み. 朝倉書店

# SASシステム

# プログラムバグを見つけるための注目すべきSASログメッセージ

○吉野　祥

(株式会社クレスコ)

SAS Log message for remove bugs

Sho YOSHINO

CRESCO LTD.

## 要旨

SAS プログラムを実行したとき生成されるログには、SAS プログラムのソース内容やプログラムより作成されたデータセットの情報、各ステップの処理時間、使用したメモリ、発生した ERROR や WARNING などが出力される。ログに出力されたメッセージから、SAS プログラムがどのような動きをしたかを知ることができる。

ログを確認することは、デバッグ時にも有効である。ログからは、コーディングミスがないか、意図していない動作をしていないかを確認することができ、その原因を見つける手助けをしてくれる。ERROR や WARNING は強調して出力されるため、メッセージを見逃すことはないだろう。

しかし、ERROR や WARNING のような明確なメッセージを確認するだけでは、確認としては十分ではく、それ以外にも注目すべきメッセージは存在する。それらのメッセージを確認することで、ERROR や WARNING を見るだけでは気づかなかった動作にも気づくことができる。

また、注目すべきログメッセージを事前に把握しておくことは、SAS プログラムの品質向上や作成効率化に繋がると考える。

本稿では、その注目すべきログメッセージと注目すべき理由について紹介する。

キーワード：ログ、品質、デバッグ

## 注目すべき SAS ログメッセージ

WARNING や ERROR 以外の注目すべき SAS ログメッセージについて、実際の SAS ログを見ながら、メッセージをひとつひとつ紹介していく。

ログ 1 は、欠損値を含んだ四則演算が実施されたときに出力されるメッセージである。四則演算の中に 1 つの値でも欠損値を含んでいると結果が欠損値になる。

結果を欠損値にしないためには、例えば、プログラム 1 のような方法がある。変数が欠損値かどうかを判定する missing 関数を用いて、計算の前に欠損値を含んだ計算をしないようにする。coalesce 関数を用いて、欠損値の場合、違う値に変換してから計算するようにする。sum 関数のように、欠損値を含んでいても計算処理を行ってくれる関数を用いる方法である。

ログ 1 : 欠損値を含んだ計算により～

```
60          data TEST ;
61            A = . + 1 ;
62          run ;
NOTE: 欠損値を含んだ計算により、以下の箇所で
欠損値が生成されました。
      (回数)(行:カラム)
      61:9
```

プログラム 1 : 欠損値を考慮したプログラム例

```
data TEST ;
  set TEST ;
  if not missing(A) then B = A + 1 ;
  C = coalesce(A, 0) + 1 ;
  D = sum(A, 0) ;
run ;
```

ログ 2 : 変数～は初期化されていません

```
61          data TEST ;
62            A = B ;
63          run ;
NOTE: 変数 B は初期化されていません。
```

ログ 2 は、作成されていない変数を使用したときに出力されるメッセージである。例の場合、変数 B には欠損値が格納される。

このとき、誤った変数名を記載しているか、変数の初期化処理を記載し忘れている可能性がある。

ログ 3 は、分母が 0 の割り算(除算)を実施したときに出力されるメッセージである。行 61 と行 63 にあるように、計算式に変数があるかどうかでメッセージが異なる。分母に 0 が入る可能性がある場合は、事前に 0 かどうかを判定しておくとよい。

ログ 3 : 0 による除算が発生しました

```
60          data TEST ;
61            A = 1 / 0 ;
NOTE: コンパイルフェーズの途中の行 61 カラム 9 で、0 による除算が発生しました。
62            B = 1 ;
63            A = B / 0 ;
64          run ;
NOTE: 行 63 カラム 9 で、0 による除算が発生しました。
A=. B=1 _ERROR_=1 _N_=1
NOTE: 以下の箇所で演算式を計算できなかったため、結果を欠損値に設定しました。
      (回数)(行:カラム)
      1 63:9
```

ログ 4、ログ 5 は、関数の引数に無効な値を含んだときに出力されるメッセージである。ログ 4 は DATA ステップにて、ログ 5 は SQL プロシジャにて出力され、それぞれの関数の返却値は欠損値となる。同じ引数の無効という事象ではあるが、DATA ステップ、SQL プロシジャでメッセージが異なることに注意したい。

ログ 4 : 関数～の第～引数は無効です

```
60          data TEST ;
61            A = substr("TEST", 1, -1) ;
62          run ;
NOTE: 関数 SUBSTR(行 61 カラム 7)の第 3 引数は無効です。
A=TEST _ERROR_=1 _N_=1
```

ログ 5 : ～関数への引数が無効または欠落により／無効な引数～(関数～)です

```
66          proc sql ;
67            select substr(A, 1, -1)
68                 , year(.)
69              from TEST
```

```
70              ;
NOTE: YEAR 関数への引数が無効または欠落により、関数が欠損値を戻しました。
NOTE: 無効な引数 3 (関数 SUBSTR)です。欠損値が生成される可能性があります。
71          quit ;
```

ログ 6 の「数値を文字値に変換しました」は、文字処理に数値が含まれていた場合に出力される。逆に、「文字値を数値に変換しました」は、数値処理に文字が含まれていた場合に出力される。

**ログ 6 : 以下の箇所で～を～に変換しました**

```
61          data TEST ;
62            A = 1234567890123 ;
63            B = "TEST" || A ;
64            C = 123 + "456" ;
65            putlog A= B= C= ;
66          run ;
NOTE: 以下の箇所で数値を文字値に変換しました。
      (行:カラム)
      63:17
NOTE: 以下の箇所で文字値を数値に変換しました。
      (行:カラム)
      64:13
A=1.2345679E12 B=TEST1.2345679E12 C=579
```

ログ 6 のメッセージの問題は、値の変換が期待している値に変換されていない可能性があるという点である。特に数値は、best12. で変換されるため、13 桁を超えると、指数表記で変換されてしまう。ログ 6 の putlog ステートメントの結果のように、E 付きの数値で変換される。暗黙的に変換されるため、処理結果が期待された結果になっていることを確認しなくてはならないだろう。

**ログ 7 : 文字型変数は以下の個所において、デフォルトで 200 文字になりました**

```
60          options msglevel=i ;
61
62          data TEST ;
63            A = "ABC" ;
64            B = cat(A, "DEF") ;
65          run ;
INFO: 文字型変数は以下の個所において、デフォルトで 200 文字になりました。結果が切り捨てられる場合があります。
      (行:カラム)
      64:3      B
```

ログ 7 は、options msglevel=i を設定しているとき、事前に length ステートメントで長さを指定していない変数に対して、cat 関数などの文字関数の結果を格納した場合に出力されるメッセージである。長さが指定されていない場合、長さが 200 バイトになるため、200 バイトを超える文字を扱っていた場合に文字が切れてしまう。本メッセージを確認した場合、事前に length 関数を指定すべきである。

ログ 8 は、“ABCDE” EQ のようにダブルクォーテーションが閉じられた後、半角スペースを付けず、EQ などの演算子を記載したときに出力されるメッセージである。

ダブルクォーテーションの後に、識別子を付けるパターンとして、SAS 日付 (例. “01JAN2017” d)、SAS 時刻(例. “08:01:23” t)、SAS 日時(例. “01JAN2017 08:01:23” dt)がある。ログにあるように、今後のリリースで、上記以外のパターンが追加される可能性があるため、ダブルクォーテーションの後には半角スペースを挿入するべきである。また、コーディングルールとして、EQ などのニーモニック演算子を使用せず、= (イコール) のような記号演算子を採用することも検討するとよいだろう。

**ログ 8 : 引用符で囲まれた文字列の後の識別子の意味は～**

```
60          data TEST ;
61            A = "12345" ;
62            if "ABCDE"EQ A ;
```

49

```
NOTE 49-169: 引用符で囲まれた文字列の後の識別子の意味は、
             将来の SAS リリースで変わる可能性があります。
             引用符で囲まれた文字列と識別子の間にスペースを挿入することをお勧めします。
63         run ;
```

ログ 9 は、数値項目に対して、put 関数による文字列変換またはレポート出力(print プロシジャなど)する際に、小さい出力形式が指定されたときに出力されるメッセージである。例えば、3 桁の数値に対して、3 桁未満の出力形式が指定されたときである。

このメッセージが出力されているとき、実際の値とは異なる文字に変換またはレポート出力されている。記載した出力形式を更に大きな値にして修正すべきである。

**ログ 9 : 数値をプリントするには小さすぎる W.D 出力形式がありました**

```
60         data TEST ;
61           A = 123 ;
62           B = put(A, 2.) ;
63         run ;
NOTE: データセット WORK.TEST は 1 オブザベーション、2 変数です。
NOTE: 数値をプリントするには小さすぎる W.D 出力形式がありました。
      "BEST"出力形式によって小数点がシフトされる場合があります。
```

ログ 10 は、行末に、コロンを記載したとき出力されることがあるメッセージである。本メッセージは、goto 文のラベルを書いているにも関わらず、goto 文が記載されていないため出力される。

セミコロンとコロンの書き間違いはよく起こることであるが、この場合、構文エラーにならない。そのため、書き間違いを気付かず見逃しやすくなる。セミコロンの書き間違いではなく、goto 文の書き漏れである場合は、プログラム 2 のように載することで、本メッセージは出力されなくなる。

**ログ 10 : ～は未参照のラベル定義です**

```
60         data TEST ;
61           A = 1 ;
62           FORMAT :
63         run ;
NOTE: FORMAT は未参照のラベル定義です。
```

**プログラム 2 : goto 文の記載例**

```
data TEST ;
  A = 1 ;
  goto FORMAT ;
FORMAT :
  return ;
run ;
```

ログ 11、ログ 12 は、where ステートメントが必ず 0 または 1 となる書き方をしているときに出力されるメッセージである。where 句や DATA ステップそのものが意味のないコードとなる。テストコードの消し忘れの場合もあるため、確認すべきメッセージである。1 と 0 のみを格納したマクロ変数を用いた場合も発生するので、書き方を見直すと良いだろう。(例. where &TEST. → where &TEST. = 0)

**ログ 11 : WHERE 0 /* 明らかに FALSE な WHERE 句です。*/**

```
64         data TEST ;
65           set TEST ;
66           where 0 ;
67         run ;
NOTE: データセット WORK.TEST から 0 オブザベーションを読み込みました。
      WHERE 0 /* 明らかに FALSE な WHERE 句です。 */ ;
```

**ログ 12 : WHERE 1 /* 明らかに TRUE な WHERE 句です。*/**

```
69         data TEST ;
70           set TES T;
71           where 1 ;
72         run ;
NOTE: データセット WORK.TEST から 0 オブザベーションを読み込みました。
      WHERE 1 /* 明らかに TRUE な WHERE 句です。 */ ;
```

**ログ 13 : WHERE 式を置き換えました**

```
65          data TEST ;
66            set TEST ;
67            where A = 1 ;
68            where A = 2 ;
NOTE: WHERE 式を置き換えました。
69          run ;
```

ログ 13 は、where を 2 回以上記載したときに出力されるメッセージである。最後の where ステートメントが優先されるため、それ以前の where ステートメントは意味のない記述となる。

if ステートメントの場合は、複数記載していても、それぞれの条件が適用される。しかし、where ステートメントでは記載できないため、if ステートメントを where ステートメントに書き換えたときなどに発生することがある。複数の絞込条件が必要であるかを検討し、必要であるならば、if ステートメントに書き換え直すか、1 つの where ステートメントに and を用い書き直すべきである。

ログ 14 は、DATA ステップにて、無限ループが発生したときに出力メッセージである。SAS8 の場合、本メッセージは出力されず、終わらない無限ループが発生することになる。if 0 のように、条件を必ず満たさないサブセット化 if ステートメントの中に、set ステートメントや input ステートメントがある場合に発生することがある。

例のプログラムは、データセット TEST_1 の定義をコピーするためのプログラムである。run ステートメントの前に、stop ステートメントを記載し、処理を中断することで回避できる。

**ログ 14 : ループが発生したため、DATA ステップの実行を中止しました**

```
64          data TEST_2 ;
65            if 0 then set TEST_1 ;
66          run ;
NOTE: ループが発生したため、DATA ステップの実行を中止しました。
```

ログ 15 は、環境設定に指定された文字コード以外の文字コードで作成されたデータセットを扱ったときに出力されるメッセージである。メッセージにあるように、処理パフォーマンスが低下する。例えば、システム更改で文字コードを変更したときなどに発生する。

頻繁に扱うデータセットであるならば、文字コードを変えておくと良いだろう。文字コードの変換は、CVP= オプションを使うのが簡単であるが、文字の長さが自動的に変換される。適切な長さにならない場合もあるので、時間はかかるがデータセットに応じて、文字変数の長さを 1 つずつ指定し直すとよいだろう。

**ログ 15 : データファイル～に他のホストにネイティブな形式が使用されているか～**

```
60          data TEST(encoding="sjis") ;
61            A = 1 ;
62          run ;
INFO: データファイル WORK.TEST.DATA に他の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。
```

ログ 16 における「変数～はデータセットによって上書きされます」は、options msglevel=i を設定し、merge ステートメントによる結合時に、by ステートメントに記載していない変数をそれぞれのデータセットにある場合に出力されるメッセージである。同じ変数があった場合、merge ステートメントの後ろに書いたデータセットの値が格納される。本メッセージはどのデータセットの値が格納されたかを教えてくれる。期待していないデータセットの値が格納される場合を防ぐため、options msglevel=i は必ず設定しておくべ

きである。

ログ 16 における「MERGE ステートメントに BY 値を繰り返すデータセットが複数あります」は、merge ステートメントによる結合時に、n×n の結合がなされたとき出力されるメッセージである。

**ログ 16： 変数～はデータセット～によって上書きされます／MERGE ステートメントに BY 値を繰り返すデータセットが複数あります**

```
88          data TEST_3 ;
89            merge TEST_1 (in=SW1)
90                  TEST_2 (in=SW2)
91                  ;
92            by A ;
93            if SW1 and SW2 ;
94          run ;
INFO:  変数 B (データセット WORK.TEST_1) はデータセット WORK.TEST_2 によって上書きされます。
NOTE: MERGE ステートメントに BY 値を繰り返すデータセットが複数あります。
NOTE: データセット WORK.TEST_1 から 5 オブザベーションを読み込みました。
NOTE: データセット WORK.TEST_2 から 7 オブザベーションを読み込みました。
NOTE: データセット WORK.TEST_3 は 7 オブザベーション、2 変数です。
```

merge ステートメントによる統合は、SQL による結合とは処理結果が異なる。例えば、表 1、表 2 のようなデータセットを、ログ 16 のように変数 A で結合させた場合を考える。merge ステートメントによる結果は表 3 のようになる。対して、プログラム 3 のように SQL プロシジャで結合させた場合、表 4 のような結果となる。その結果の違いにおいて、特に注目すべきは件数である。大きく結果が異なるのが分かる。

本メッセージを確認した場合、作成されたデータセットの内容が期待した内容になっているかを確認すべきである。期待しない内容であるならば、n×n の結合とならないように事前加工をするか、SQL や FEDSQL プロシジャを用いることになるだろう。

**表 1 ： データセット TEST_1**

| A | B |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 2 | 3 |
| 3 | 4 |
| 3 | 5 |

**表 2 ： データセット TEST_2**

| A | B |
|---|---|
| 1 | 11 |
| 2 | 12 |
| 2 | 13 |
| 2 | 14 |
| 3 | 15 |

**表 3 ： merge により n×n の結合**

| A | B |
|---|---|
| 1 | 11 |
| 2 | 12 |
| 2 | 13 |
| 2 | 14 |
| 3 | 15 |
| 3 | 5 |

**表 4 ： SQL による n×n の結合**

| A1 | A2 | B1 | B2 |
|---|---|---|---|
| 1 | 1 | 1 | 11 |
| 2 | 2 | 3 | 14 |
| 2 | 2 | 2 | 14 |
| 2 | 2 | 3 | 13 |
| 2 | 2 | 3 | 12 |
| 2 | 2 | 2 | 12 |
| 2 | 2 | 2 | 13 |
| 3 | 3 | 5 | 15 |
| 3 | 3 | 4 | 15 |

**プログラム 3 ： SQL プロシジャによる n×n の結合**

```
proc sql ;
  create table TEST_3_SQL as
    select T1.A as A1
         , T2.A as A2
         , T1.B as B1
         , T2.B as B2
      from TEST_1 as T1
     inner join TEST_2 as T2
        on T1.A = T2.A
     order by T1.A
  ;
quit ;
```

ログ 17 は、libname ステートメントにて存在しない場所を指定したときのメッセージである。プログラムの初期開発時やライブラリの指定にマクロ変数を用いているときにはチェックすべきメッセージだろう。

**ログ 17 : ライブラリ～は存在しません**

```
61          libname TEST "/myfolder/wrong";
NOTE: ライブラリ TEST は存在しません。
```

ログ 18 は、データセットがないとき、datasets プロシジャにてデータセットを削除しようとしたときに出力されるメッセージである。タイピングミスで意図していないデータセット名を記載している可能性もあるため、開発当初は期待したデータセットが削除されているか確認しておくべきである。

対して、ログ 19 のように、delete プロシジャを用いた場合、NOTE メッセージではなく、WARNINGメッセージとして処理される。プログラムの用途に応じて、どのプロシジャを使うか選択するとよいだろう。

**ログ 18 : ファイル～は見つかりませんが、DELETE ステートメントに存在します**

```
61          proc datasets lib=work nolist ;
62            delete TEST ;
63          run ;
NOTE: ファイル WORK.TEST (memtype=DATA)は見つかりませんが、DELETE ステートメントに存在します。
```

**ログ 19 : データセットがない場合の delete プロシジャのログ**

```
60          proc delete data=TEST ;
61          run ;
WARNING: ファイル WORK.TEST.DATA は存在しません。
```

ログ 20、ログ 21 は、指定したデータセットにオブザベーションがないときに出力されるメッセージである。0 オブザベーションのまま処理が続くことが想定どおりであればよいが、処理の場所によって想定どおりでない場合もある。

オブザベーション数が重要である場合は、ログで判断せず、自動マクロ変数 SYSNOBS などで、オブザベーション数を取得し、プログラム上で判断させるとよいだろう。

**ログ 20 : 入力データセットは空です／データセット～から 0 オブザベーションを読み込みました**

```
66          proc sort data=TEST ;
67            by A ;
68            where A = . ;
69          run ;
NOTE: 入力データセットは空です。
NOTE: SAS ソートを使用しました。
NOTE: データセット WORK.TEST から 0 オブザベーションを読み込みました。
      WHERE A=.;
```

**ログ 21 : データセット～にオブザベーションがありません**

```
67          proc summary data=TEST nway missing ;
68            class A ;
69            output out=SUM ;
70          run ;
NOTE: データの要約には、複数の同時スレッドが使用されます。
NOTE: データセット WORK.TEST にオブザベーションがありません。
```

ログ 22 は、プログラム 4 を実行したときのログである。
における「～に対して、無効なデータが～にあります」は、変数 C が数値を指定しているにも関わらず、文字列を読み込んだときのメッセージである。
プログラム 4 の場合、文字列 klmn を読み込もうとしている。このメッセージが出力された場合、input ステートメントあるいは読込データに不備がある可能性がある。

ログ 22 における「INPUT ステートメントが行の終端に達したので、次の行を読み込みます」は、読み込んでいる行に input ステートメントで指定した変数分のデータがなく、不足分を次の行から読み込もうとしたときに出力されるメッセージである。変数 A , B はそれぞれ 1 行目の文字列 abcde , fghij を読み込み、変数 C は次の行、2 行目の文字列 klmn を読み込もうとしている。

1 行の項目数が一定ではない複雑なデータレイアウトの場合には無視してよいかもしれないが、基本的にはデータに不備がないかを確認すべきである。

ログ 23 は、ファイル 1 を読み込むプログラム 5 を実行したときのログである。ファイル 1 の長さは 10 バイトあるが、プログラム 5 にて指定している lrecl= の値は 5 バイトである。そのため、5 バイトまでしか読み取られず、6 バイト以降は切り捨てられたため出力されたメッセージである。本メッセージを確認した場合、データ、あるいは lrecl= の長さ、input ステートメントに指定した変数の長さを見直すべきである。

**プログラム 4 ： 無効なデータ**

```
data TEST ;
  infile cards ;
  input A $5. B $5. C ;
cards ;
abcde fghij
klmn
;
run ;
```

**ファイル 1**

```
1234567890
```

**プログラム 5 ： lrecl= が小さいプログラム**

```
data TEST ;
  infile IN lrecl=5 ;
  input A 8. ;
run ;
```

**ログ 22 ： ～に対して、無効なデータが～にあります／INPUT ステートメントが行の終端に達したので、次の行を読み込みます。**

```
NOTE: C に対して、無効なデータが行 64 カラム 11-11 にあります。
RULE:       ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8
64          abcde fghij
A=abcde B=fghi C=. _ERROR_=1 _N_=1
NOTE: LOST CARD.
66          ;
A=klmn B=  C=. _ERROR_=1 _N_=2
NOTE: INPUT ステートメントが行の終端に達したので、次の行を読み込みます。
```

**ログ 23 ： 切り捨てられた行があります**

```
NOTE: 1 レコードを入力ファイル IN から読み込みました。
      最小レコード長は 5 です。
      最大レコード長は 5 です。
      切り捨てられた行があります。
NOTE: INPUT ステートメントが行の終端に達したので、次の行を読み込みます。
```

ログ 24 は、SQL プロシジャにおいて、noexec オプションが記載されているとき出力されるメッセージで

ある。メッセージにあるとおり、SQL プロシジャは実行されない。noexec オプションは実行しないためのオプションで、SQL の構文チェックに利用できる。本メッセージを確認することで、開発時のオプション削除忘れが確認できる。

**ログ 24 : NOEXEC オプションが指定されているため、ステートメントは実行されません**

```
60          proc sql noexec ;
61            select a
62              from TEST
63              ;
NOTE: NOEXEC オプションが指定されているため、ステートメントは実行されません。
```

ログ 25 は、SQL プロシジャにおいて、case 式にて、else 句を記載しなかった場合に出力されるメッセージである。メッセージのように、when 句に記載した条件をすべて満たさなかった場合、欠損値が出力される。else 句も含めて条件分岐を明示的に記載することで、処理分岐の漏れを防ぐことができる。

**ログ 25 : CASE 式に ELSE 句がありません**

```
67          proc sql ;
68            select case when A = 2 then 1 end as A_1
69              from TEST
70              ;
NOTE: CASE 式に ELSE 句がありません。すべての WHEN 句の条件に合わなかった場合、
      CASE 式の結果は欠損値になります。
```

ログ 26 は、SQL プロシジャにおいて、group by 句に記載のない変数が、select 句の集約関数外にて記載されていた場合に出力されるメッセージである。例のプログラムであれば、変数 B が group by になく、select 句にある変数である。このとき、メッセージにあるように、再マージという処理が実行される。

**ログ 26 : クエリは元のデータに要約統計量の結果を再マージ**

```
74          proc sql ;
75            select A
76                 , B
77                 , sum(B) as B_SUM
78              from TEST
79             group by A
80             order by A
81             ;
NOTE: クエリは元のデータに要約統計量の結果を再マージします。
```

再マージは、集計処理が行われた後、集計結果を集計前のデータと結びつける処理である。表 1 のデータを変数 A で集計した結果が表 2 で、その集計結果を再マージした結果が表 3 である。

再マージは、SAS 独特の SQL の結果のため、特に SQL を慣れ親しんだユーザーにとっては期待していない結果の場合がある。再マージを許可したくない場合は、options nosqlremerge を設定することで、再マージを抑制し、ログ 27 のように ERROR にすることができる。

または、SAS9.4 以降であるならば、SQL プロシジャではなく、FED SQL プロシジャを用いることでも再マージを防ぐことができる。FED SQL プロシジャは、ANSI SQL99 に準拠しているため、SQL ユーザーにとっては、期待通りの結果になりやすくなるだろう。

表 5 : 集計前

| A | B |
|---|---|
| 10 | 100 |
| 20 | 200 |
| 20 | 300 |
| 30 | 400 |
| 30 | 500 |

表 6 : 集計後、再マージ前

| A | B |
|---|---|
| 10 | 100 |
| 20 | 500 |
| 30 | 900 |

表 7 : 再マージ後

| A | B | B_SUM |
|---|---|---|
| 10 | 100 | 100 |
| 20 | 200 | 500 |
| 20 | 300 | 500 |
| 30 | 400 | 900 |
| 30 | 500 | 900 |

ログ 27 : options nosqlremerge を設定して、再マージをしたときのメッセージ

```
ERROR: クエリは、要約統計量と元のデータとの再マージを要求しましたが、許可されません。
       NOREMERGE プロシジャオプションまたは NOSQLREMERGE システムオプションが指定されています。
```

ログ 28 は、SQL プロシジャにおいて、group by 句が記載されているが、select 句や having 句に集約関数(sum や max , count など)が記載されていない場合に出力されるメッセージである。group by が余分であるか、集約関数を追記すべきか見直す必要がある。なお、order by 句を記載しない場合、NOTE ではなく、ログ 29 のように WARNING が出力される。

ログ 28 : 関連するテーブルの SELECT 句も HAVING 句も要約関数を引用していないので～

```
60        proc sql ;
61          select A
62            from TEST
63           group by A
64           order by A
65           ;
NOTE: 関連するテーブルの SELECT 句も HAVING 句も要約関数を引用していないので、
      GROUP BY 句は無視されました。
66        quit ;
```

ログ 29 : 関連するテーブルの SELECT 句も HAVING 句も要約関数を引用していないので～(WARNING)

```
WARNING: 関連するテーブル式の SELECT 句も HAVING 句も要約関数を引用していないので、
         GROUP BY 句は ORDER BY 句に変換されました。
```

ログ 30 は、SQL プロシジャにおいて、select 句にない変数を order by 句に記載していた場合に出力されるメッセージである。order by 句に、誤った変数を記載していないか確認する必要がある。

ログ 30 : 指定したクエリには SELECT 句にない項目によるソートが含まれます

```
74        proc sql ;
75          select A
76            from TEST
77           order by B
78           ;
NOTE: 指定したクエリには SELECT 句にない項目によるソートが含まれます。
79        quit ;
```

## ログメッセージを確認する上で気をつけなければならないこと

紹介した SAS ログメッセージの有無を確認することで、バグの可能性がある箇所を発見しやすくなるだろう。しかし、ログを確認する上で気をつけなければならない点がある。

1 つ目は、**ログメッセージの内容はバージョンとともに変化する**ことである。

例えば、SAS9.4 と SAS9.3 で分母が 0 の割り算のときのメッセージは異なる。SAS8 のとき、where ステートメントが必ず 1 または 0 のときのメッセージは、日本語ではなく英語であった。

そのため、バージョンによっては、本稿で紹介したメッセージはそのまま使えない。それぞれのメッセージを出力するための SAS プログラムを実行し、どのようなメッセージが出力されるかを確認する必要がある。また、バージョンアップの都度、メッセージが変化していないかを確認すべきである。ツールなどで自動的に検索するときは、検索できなくなる。確認しない場合、意味のないメッセージの確認をすることになりかねない。

また、紹介したメッセージ以外にもバグになりうるメッセージはあるだろう。バージョンアップによって追加になるケースも十分にありうる。特に、新しい技術を使う場合は普段見かけないログがないかを注視しておくべきである。そのためには、普段出力されているログがどういうものかも慣れておくとより良いだろう。

2 つ目は、**同じ処理でも DATA ステップ、各プロシジャによって、メッセージの有無や内容が異なる**ことである。 紹介したメッセージは、主に DATA ステップや SQL プロシジャの処理中に出力されるメッセージである。しかし、他のプロシジャで同じ処理をしたとき、ERROR や WARNING になることもあれば、NOTE メッセージも出力されない場合がある。

例えば、プログラム 6 では、DS2 プロシジャにて初期化処理をしていない変数 B を使用している。しかし、DATA ステップとは異なり、特にログにメッセージは表示されない。

プログラム 7 は、SQL プロシジャにおける欠損値の含む四則演算で、特別なメッセージは表示されない。これは、FEDSQL プロシジャ、DS2 プロシジャにおいても同様である。結果が欠損値となることは同様である。

プログラム 8 は、lua プロシジャにおける分母が 0 の割り算で、特別なメッセージは表示されない。結果は欠損値となる。

プログラム 9 は、FEDSQL プロシジャにて、case 式に else 句がない場合も、SQL プロシジャと異なりメッセージは表示されない。

**プログラム 6 : DS2 プロシジャ 初期化していない変数の使用**

```
proc ds2 libs=work ;
   data TEST (overwrite=yes) ;
       dcl decimal(10, 0) A B ;
       method run();
           A = B ;
       end;
   enddata ;
   run ;
quit ;
```

**プログラム 7 : SQL プロシジャ 欠損値を含む四則演算**

```
proc sql ;
  select . + A as A
    from TEST
    ;
quit ;
```

プログラム 8 : lua プロシジャ 0 割の計算

```
proc lua ;
  submit ;
    local A = 1
    local B = 0
    A = A / 0
    print("A = " .. A)
    print("B = " .. B)
  endsubmit;
quit ;
```

プログラム 9 : FEDSQL プロシジャ else 句の記載なし

```
proc fedsql libs=work ;
  select case when A = 2 then 1 end as A_1
    from TEST
    ;
quit ;
```

紹介したメッセージは、あくまで紹介したステップやプロシジャの中でしか出力されないことを注意しなければならない。

3 つ目は、**紹介したメッセージが出力されている＝バグ、ではないということ**である。

例えば、ログ 22 の「INPUT ステートメントが行の終端に達したので、次の行を読み込みます」のように、扱うデータによって、必ず出力されてしまうことがある。

メッセージは確認すべきではあるが、ひとつひとつのプログラムに応じて、判断すべき場合がある。ただし、ルールに例外を多くつくると、確認に漏れが生じることになりかねない。あるプログラムでは確認すべきであるが、あるプログラムでは確認しない、となると確認しなくなるパターンが発生する可能性がある。各々の環境に応じて、例外は極力減らし、確認すべきメッセージを選定し、コーディングルールを設定するべきである。

最後に、**ログだけではあらゆるバグを見つけることはできない**ということである。ログの確認は、バグを見つけるための銀の弾丸ではない。2 つ目で述べた通り、紹介したメッセージは特定条件下でしか出力されない。したがって、ログだけでは、バグかもしれない事象をすべて確認することはできない。

紹介したメッセージからできることは、一部のバグかもしれない箇所を見つけやすくしてくれることである。ログだけに注目して、データを見ることを疎かにしてはならない。ログだけを確認するのではなく、作成されたデータの内容についても確認していかなければならない。

## まとめ

WARNING や ERROR 以外の注目すべき SAS ログメッセージについて、実際の SAS ログと共に紹介した。紹介したメッセージの有無をプログラムの実行後に確認することで、バグに気づきやすくなるだろう。作成されたデータから探すことも可能であるが、ログであれば、発生した箇所も特定できるので、デバッグも容易になる。ツールなどを用いて、ログを確認し、デバッグを容易にさせてくれるだろう。

しかし、先述したとおり、すべてのバグをログから見つけることはできない。ログも確認し、作成されたデータの内容も必ず確認することで、より良い品質のプログラムになるだろう。

本稿が、少しでも SAS プログラム開発の一助となれば幸いである。

# SAS Drug Development (SDD)　実務の実際
## - データマネジメント業務と解析業務はどう変わったか -

○松山友映[1]、○高沼正幸[1]、兵頭洋平[1]、長谷川優子[1]、本山佳代子[1]、山下昌洋[1]
([1]協和発酵キリン株式会社)

申請時電子データ提出に対応した医薬品データハンドリングプロセスを保証するには、一連のプロセスのトレーサビリティが重要となる。また情報資産管理の観点から、電子データの堅牢なセキュリティ対策も求められる。
一方で、申請用電子データを作成するには、データマネジメント部門、解析部門、臨床薬理部門、バイオマーカー解析部門、安全性管理部門など関連部門との連携が不可欠となる。さらに、昨今は国内外も含め外部リソースの活用も進み、地政学的な壁を乗り越えた業務連携も日常茶飯事となってきた。
このような環境変化の中では、PC-SAS(もしくはServer SAS)と社内のセキュアサーバの組み合わせといった従前の仕組みでは、データハンドリングプロセスの信頼性担保の脆弱性や情報セキュリティ管理の観点などから企業リスクが高まることが懸念される。さらに、セキュアかつシームレスなデータ共有・プロセスの共有は不可能である。そこで、これらの課題を解決するためのシステムの導入・運用が各社で始まっている。
当社では、2013年12月にSAS Drug Development(SDD)を本格稼働させ、現在全プロジェクトの電子データをSDD上で運用管理している。これまでの当社の経験を踏まえ、(1)プログラミング業務の効率化、(2)情報資産の有効活用と保護の両立、(3)シームレスな協業の促進　の3つの観点でSDD導入の有用性を紹介する。
(1)プログラミング業務の効率化（トレーサビリティの向上）
プロジェクト横断的にファイルの保管場所が統一され、情報資産へのアクセスが迅速かつ正確に行えるようになった。さらにファイルのバージョンやプログラム実行のトレーサビリティなどが自動的に記録されることによって、データセット作成プロセスの信頼性を保証するエビデンスを効率的に取得できるようになった。このトレーサビリティは実務面でも有用である。例えば、SDTMやADaMは、申請前のバリデーションチェックなどにより多少の「お化粧直し」が発生することが多い。その際に、変更対象となったデータセットの由来や使用されたプログラムなどが視覚的に特定できていることで、正確な修正対応が可能となる。
(2)情報資産の有効活用と保護の両立
臨床試験データは高度な営業秘密情報を含んでいるため、インサイダー情報になるリスクがある。そのため臨床試験データは、許可された者のみがアクセスでき、一方公開可能となった場合には情報の有効活用のため、適正な範囲で公開されなければならない。このきめ細かな情報資産のアクセス権設定を手動で行うことは煩雑であるため実務的には不可能であり、SDDのアクセス権設定機能が情報資産の有効活用と保護の調整弁として活躍している。
(3)シームレスな協業の促進
自社の海外拠点のみならず、国内外のCROとの協業はもはや製薬業界の常識である。特にSAS関連の業務では、所属、国境に拠らず作業環境を共有できるメリットは計り知れない。当社では海外拠点および国内外のCROとSDDを介した協業を推進しており、即実行可能なSASプログラムの共有、成果物の即時納品、作業進捗のリアルタイムな確認、標準マクロの共有など、その利点を享受している。

# Jupyter Notebookを活用したプログラムライブラリ構築の検討

○吹谷 芳博[1]，藤澤 正樹[1]

([1]あすか製薬株式会社)

# Examination of the program library construction using Jupyter Notebook

Yoshihiro Fukiya[1], Masaki Fujisawa[1]

[1]ASKA Pharmaceutical Co., Ltd.

## 要旨

プログラム開発とその周辺作業の効率化は，プログラム開発に欠かせない検討課題であり，惜しみない努力と時間が費やされている．吹谷（2016）は，R 言語のパッケージを用いて，SAS 言語のプログラムソースと文書を一括作成・管理する方法を検討した．しかし，パラメータ設定が煩雑であり，PC SAS の環境のみに制限されていたため，汎用性のあるツールとしての拡張性に欠ける．

本報告では，Jupyter Notebook でのプログラム実行環境下で，文書，プログラムソース及びプログラムの実行結果を一括管理できる方法を提案する．また，R 言語パッケージでの実装結果と比較し，その利点と欠点を浮き彫りにする．Jupyter Notebook の利便性を活かして SAS マクロなどのナレッジ管理ツールとしての使用事例を報告する予定である．その他の拡張ツールを導入することで，さらなる利便性の向上を図った内容も報告する予定である．

キーワード：Jupyter Notebook，Python，markdown，文芸的プログラミング

## 1. 序に代えて

プログラム開発とその周辺作業の効率化は，プログラム開発に欠かせない検討課題であり，惜しみない努力と時間が費やされている．とくにデータ解析で使用されるプログラムは汎用的に使用するために，社内標準化のルールや共通ルールに基づくシステムを構築しても往々にプログラムの共有化やメンテナンスで徒労に終わる．そこで効率の良いプログラムの標準化を目指すことを目的として，吹谷（2016）[6]は，R 言語のパッケージ（knitr）を用いて，SAS 言語のプログラムソースと文書を一括作成・管理する方法を検討した．しかし，パラメータ設定が煩雑であり，PC SAS の環境に限定されていたため，汎用性のあるツールとしての拡張性に欠けていた．

近年，Jupyter Notebook[1]が，データの分析作業を円滑に進めるツールとして脚光を浴びている．本ツールは，以前に IPython Notebook と呼ばれていたツールであり，Python でのプログラム開発に利用されていた．その利便性に着目した他言語のユーザーにより，さらにそれぞれの言語への移植が行われた．その結果，多言語に対応した Notebook 形式の Web アプリケーションとして開発され，対応する主要な言語（Julia + Python + R）から Jupyter と名付けられた．SAS は 2016 年 4 月に Unix SAS v9.4 で利用可能となり，同年の 7 月に SAS University Edition でも利用可能になった．

本報告では，Jupyter Notebook でのプログラム実行環境下で，文書（軽量マークアップ言語の markdown 形式で記載），プログラムソースおよびプログラムの実行結果を一括管理できる方法（Jupyter Notebook による文芸的プログラミング[4]）を提案する．また，R 言語パッケージでの実装結果と比較し，その利点と欠点を浮き彫りにする．Jupyter Notebook の利便性を活かした SAS マクロのナレッジ管理ツールとして，生物学的同等性試験の例数設計のプログラムへの適用事例を紹介する．さらに，その他の拡張ツールやプログラムを導入することで，一括実行によるナレッジ管理ツールの構築が可能となる．その適用事例にも触れる．

## 2. 方法

### 2.1. Jupyter Notebook の導入方法

Jupyter Notebook の導入には，最低限以下の 1 言語および 3 個のパッケージのインストールを要する．

- Python 3 系 : Jupyter Notebook を使う上で必須の言語（Python 2 系と Python 3 系があるが Python 3 系をインストールが望ましい）．
- Jupyter : Jupyter Notebook を動かすためのパッケージ.
- SAS_kernel[2] : Jupyter Notebook と SAS とをつなぐパッケージ.
- SASPy[ 3]: Python 上で SAS を呼び出すパッケージ（SAS_kernel と連携して Jupyter Nobebook に対して機能する）．

Jupyter Notebook を導入する際は Anaconda を推奨する．Anaconda は Continuum Analytics 社によって提供されている Python のディストリビューションである．Python 本体に加え，科学技術，数学，エンジニアリング，データ分析等，よく利用される Python パッケージが一括でインストール可能である．また，開発環境の構築が容易になり，Anaconda をインストールすることで Python 3 系と Jupyter を同時にインストールできる．

本稿では，Windows 7（32bit 版）OS および PC SAS（v9.4）の環境下で構築のための設定を試みた．インストール方法と設定方法の詳細は，SAS_kernel のドキュメント[2]に記載されているため，インストールから実装までのステップを簡潔に補遺に示した．

### 2.2. 動作確認

次の実行手順に従って動作確認を行った．動作結果を Figure 1 に示した．動作確認の結果，文書，プログラムソースおよびプログラムの実行結果が一つのファイルに表示されていた．

1. スタートから Jupter Notebook を選択する．
2. ブラウザ上で Jupter Notebook が立ち上がったら右上の「New」を選択し，「SAS」を選ぶ.
3. 以下の内容を入力し，実行する．

**title "Test";proc print data=SASHELP.CARS(OBS=5);run;**

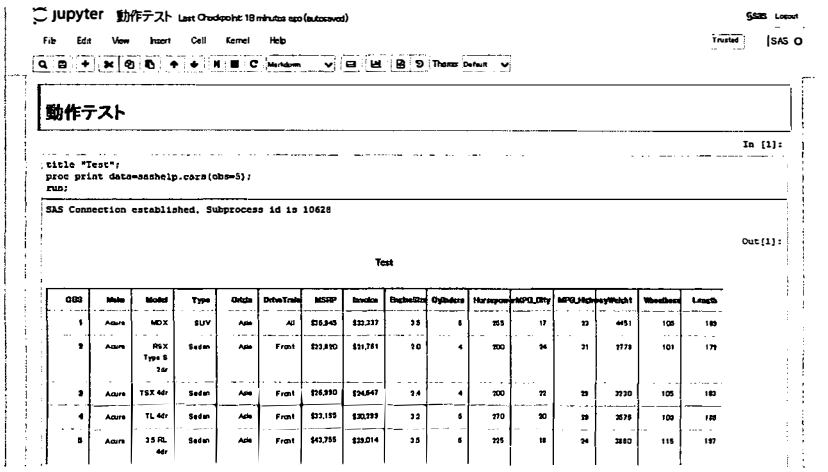



Figure 1. 動作結果

# 3. 実装結果

## 3.1. 実装した環境の性能評価

SAS_kernel のドキュメント[2]内の“Getting started”（入門）を参照して，PC SAS 環境下での出力結果を評価した．実装した環境の性能評価に利用したデータは，Kaggle （予測モデルの精度を競うデータ分析コンペティションが多数開催される分析コンペティションプラットフォームのサイト：https://www.kaggle.com/）から入手でき，高収入者および高スキル者に早期退職が多いことの原因を調査するために収集された 8 変数からなるデータセットである．本データは実データではなく，乱数により生成されたシミュレーション・データである[7]．レコード数は 14999 個であり，データに含まれる変数名は，以下のとおりであった．

·Last evaluation（最終学歴）

·Number of projects （担当しているプロジェクト数）

·Average monthly hours （月平均の勤務時間）

·Time spent at the company （会社で過ごす時間）

·Whether they have had a work accident （労働災害の有無）

·Whether they have had a promotion in the last 5 years （過去五年間の昇進の有無）

·Department （部署）

·Salary （収入）

出力結果を Figure 2 に示した．①は Markdown 形式で記載された“Getting started”の説明内容を示したものであり，②は CSV 形式のデータを取り込んだときのログの表示であり，③は取り込んだデータのコンテンツ情報をテーブル出力したものであり，④はヒートプロットのグラフ出力の結果である．②のようにプログラムの実行結果にグラフまたはテーブルが含まれなければ，ログの情報が出力される．そのため，Jupter Notebook の構成を考える段階では，プログラムソースを Jupter Notebook のどの箇所に含めるかを十分に検討し，余分なログの出力を制御することを要する．

“Getting started”を参照した実装結果を Table 1 に示した．問題点として，以下の 2 点を挙げる．

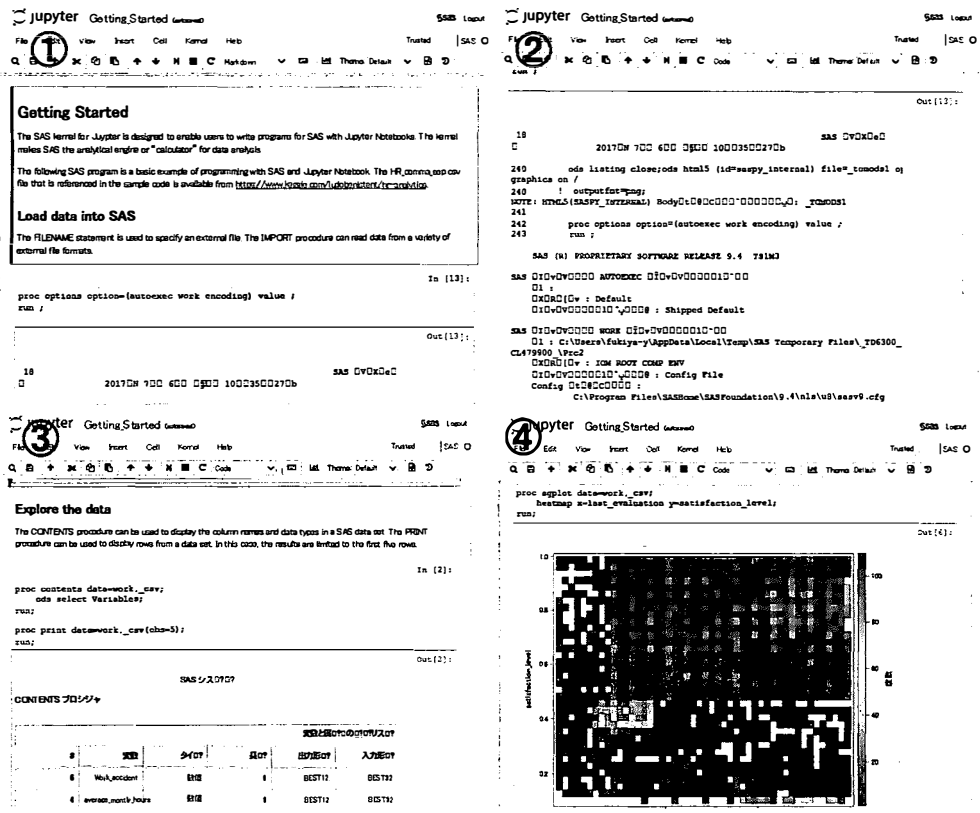



Figure 2. シミュレーション・データを用いた実装結果

Table1. 実装結果の評価

| | 評価結果 |
|---|---|
| **グラフ：画像** | オプションの指定なく綺麗に出力された. |
| **グラフ：タイトル** | 日本語を使用すると文字化けが起こった. |
| **テーブル** | 問題なく出力された．ただし，タイトルや変数ラベル及び文字列データに日本語が含まれる場合は文字化けを起こした. |
| **ログの出力** | 日本語が含まれる文字化けを起こすため, ログチェックは見読不能であった. |
| **処理速度** | 通常の SAS の GUI に比して処理速度が遅かった. |

1 点目は, 実行後のリストおよびログ関連の表記が設定されたエンコードの違いに起因して文字化けしていることである. SAS プログラムの実行はエンコードを UTF-8 に設定したが，正確に出力されなかった．SAS University Edition では，このような現象が起こらなかったため，SASPy の設定に由来する現象であるか否かの調査を要する．2 点目は，SAS の GUI よりも処理速度が遅くなることである（SAS プログラム約 8 秒に対して約 43 秒）．これは markdown のコンパル処理速度が含まれるため直接的な比較は困難であるが，それを加味しても遅く感じられた．このことは Web ブラウザ

を介して操作する SAS Studio でも同様の現象が起こり(約 23 秒), PC の CPU, メモリ及び HDD/SSD 等の処理に影響する構築環境と Web ブラウザの性能に依存する.

### 3.2. Jupyter Nobebook と R(kintr) との性能比較

Jupyter Notebook と R パッケージ(knitr) との性能評価の対比表を Table2 に示した. Jupyter Notebook の利点は, 操作方法および文章構成の柔軟性が R パッケージ(knitr) に比して高い. このため, Jupyter Notebook 環境が適切に設定されれば, 特殊な操作なく使用できる強靭なツールとなる. R パッケージ(knitr) を用いた場合の利点は, word 形式で解析結果を出力する場合に, Jupyter Notebook よりも有用となる.

今後の展開を考慮に入れると, Jupyter Notebook は拡張性が高く, 様々な追加パッケージの開発が活発であり, 有望である. 一方で, SAS University Edition は, Jupyter Notebook が同時にインストールされるため, URL による切り替えで Jupyter Notebook の使用可能となる. したがって, SAS University Edition では, このような使用方法の普及が見込まれる.

Table 2. 性能評価の対比表

| | **Jupyter Notebook** | **R(knitr)** |
|---|---|---|
| **操作方法** | ○ | △ |
| **文章構成の柔軟性** | ○ | △ |
| **出力形式 : html** | ◎ | ◎ |
| **出力形式 : word** | △ | ○ |
| **出力形式 : pdf*1** | △ | △ |
| **Unix SAS や SAS University Edition での利用** | ○ | × |
| **拡張性** | ◎ | ○ |

*1 : PDF 変換の際は Tex のツール群を要する.

## 4. 事例検討

### 4.1. ナレッジ管理の事例

Jupyter Notebook を使用したナレッジ管理ツールは, 1 マクロにつき, 1 ファイルに纏めることをルール化している. これは, マクロプログラムの更新と同時にドキュメントを更新するためである. また, 使用するデータは, 出典先および簡単な前処理のプログラムを含む, 実業務に則した形式とした. これは, 方法論と実例を含めることで, 蓄積したナレッジの積極的な活用を意図しているためである.

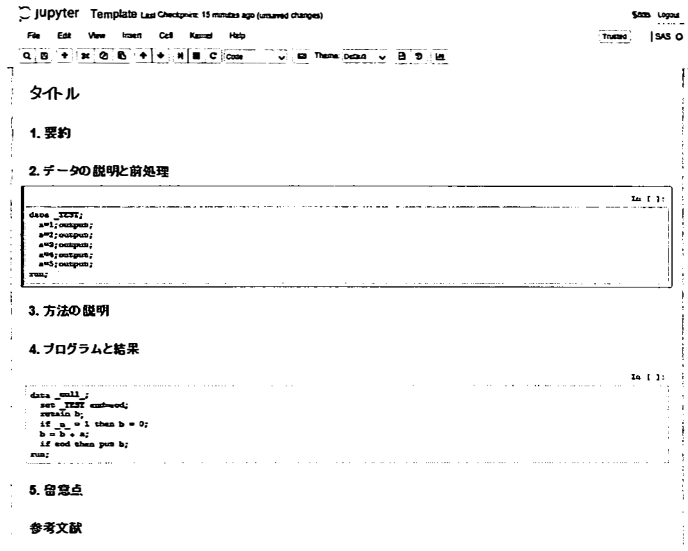



Figure 3. Jupyter Notebook をナレッジ管理として使うときの文書構成

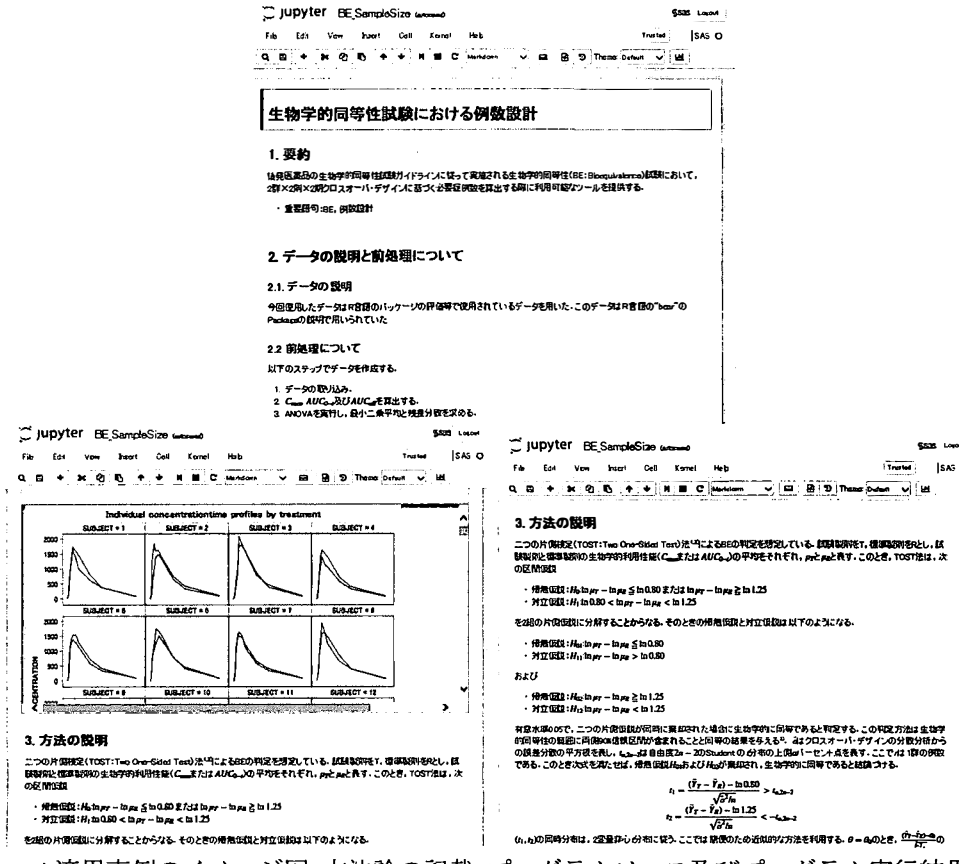



Figure 4.適用事例のイメージ図. 方法論の記載, プログラムソース及びプログラム実行結果を Jupyter Notebook で記載した例.

適用例は生物学的同等性の例数設定についてまとめたものを Figure 3 に示した．文書構成は，要約，データの説明と前処理，方法の説明，プログラムと結果，留意点の五つからなる．本事例の方法論及びプログラムソースの一部は，浜田[5]を参照した．本適用例は薬物濃度データから $C_{max}$ 及び $AUC_{0\text{-}t}$ を算出し，各パラメータについて分散分析結果を用いて，例数設計を行う一連の内容を記載した．Jupyter Notebook の長所は再度実行可能な状態で記録することが可能であるため，作成者以外のプログラマがプログラム構造を理解するために一部のプログラム実行により確認可能となる．また，一連の業務フローに対応する説明を Jupyter Notebook に含めることで，使用するマクロプログラムの位置づけも明確になる．

### 4. 2. 拡張ツールの事例

Jupyter Notebook は，html 形式で出力可能であることから情報共有用のドキュメントとして整備されている．ただし，html 形式単体では別途それぞれのリンクをつなぐフロントの html を構築しなければならないため，ドキュメンテーション・ツールとして既知である Sphinx の利用を検討した．とくに nbsphinx を利用して，Jupyter Nobebook を Sphinx のドキュメントに変換することを試みた．Jupyte Nobebook ファイル（. ipynb）を nbsphinx で変換した html 形式のドキュメントを Figure 4 に示した．本ドキュメントは，検索機能や章ごとのリンクが右側に表示される．

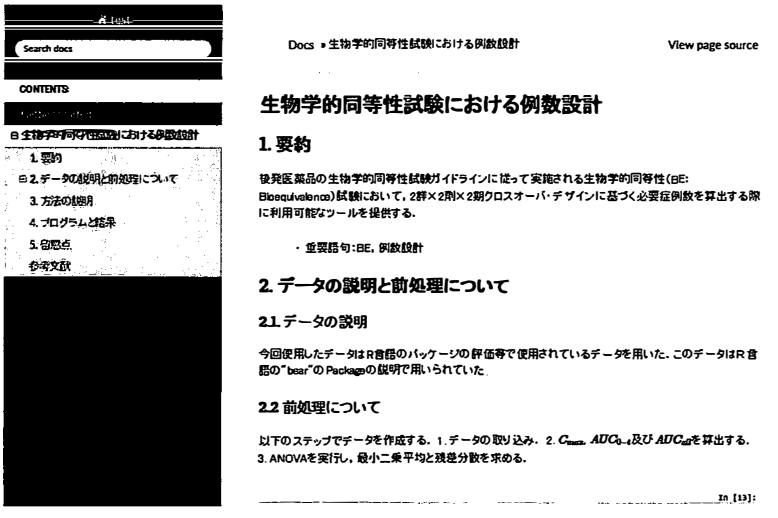



Figure 4.　Jupyte Nobebook ファイル（.ip ynb）を nbsphinx で変換した html 形式のドキュメント．

## 5. 結論

PC SAS 環境下において Jupyter Notebook を実装可能であることが確認された．ただし，日本語を含むデータセット，リスト及びログの表示が，エンコード設定に起因して文字化けしており，使用用途によっては注意を要する．PC SAS 環境下における Jupyter Notebook の実装の利点は，Jupyter の拡張が UNIX SAS や SAS University Edition に比して容易なことである．今後は Jupyter Nobebook で使うための日本語の表示の工夫を要する．

事例検討の結果から，プログラムソースと説明文書を一つのファイルで一元管理できることはマクロプログラムのナレッジ管理でも有効である．また，ファイル構造はJSONといわれる構造であり，プログラムソースと文書が保存されているため，運用の用途によって別々に抽出して管理することも可能である．Jupyter Notebookはデータサイエンス業界で活発に開発されているため，今後の拡張が期待される．

他の拡張ツールやプログラムを導入することで，さらなる利便性の向上を検討した結果，nbsphinxを利用することで閲覧しやすい形式でJupyter Notebookの結果を整形可能となった．また，この拡張ツールはとくにサーバーを立てることなく使用可能であるため，社内のナレッジ管理ツールとしての役割を担うこともできる．ただし，現状のエンコード設定の場合，日本語を検索できないため，search部分のカスタマイズを要する．

今後の課題は，SASPyの活用によりSASプログラムの資産を利用したインターラクティブなグラフの作成やDeep learningで使用されるTensorFlowのパッケージとの連携の検討である．SASをはじめとするデータ解析で利用されるプログラムが，シームレスに拡張されることを期待している．

## 参考文献


[1]:“Jupyter.” Jul 6, 2017. Available at <http://jupyter.org>.

[2]:“sas_kernel.” May 10, 2017. Available at <https://github.com/sassoftware/sas_kernel>.

[3]:“SASPy.” Jul 6, 2017. Available at < https://github.com/sassoftware/saspy>.

[4]: Knuth, Donald E. (1984). "Literate Programming". The Computer Journal. British Computer Society. 27 (2): 97–111.

[5]:浜田知久馬（2017）．新版 実用SAS生物統計ハンドブック．サイエンティスト社，382–387.

[6]:吹谷芳博（2016）．SASにおける文芸的プログラミングへの試み．SASユーザー総会2016.

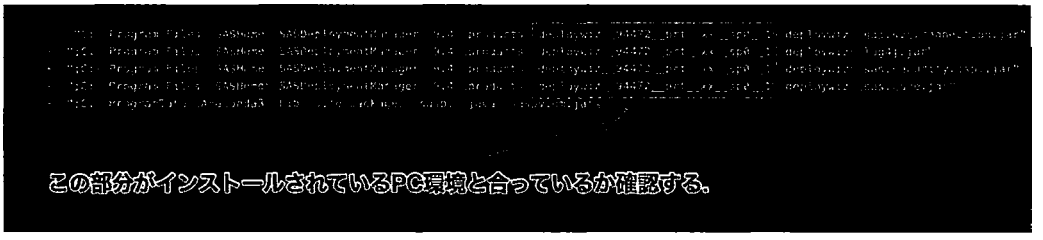
[7]: “Human Resources Analytics/Why are our best and most experienced employees leaving prematurely?”. Available at <https://www.kaggle.com/ludobenistant/hr-analytics>.

## 補遺

① Jupyter Notebook 構築手順

1. Anaconda をサイト（https://www.continuum.io/downloads）からダウンロードしてインストールを行う．構築時は Anaconda 4.3.1/Python 3.6.0 のものを使用．
2. コマンドプロンプトを立ち上げて以下のコマンドを入力し，SAS_kernel と SASPy をインストールする．
   A) pip install sas_kernel
   B) pip install saspy
3. 追加された SASPy のフォルダ内にある“sascfg.py”を以下の設定にする．

4. 次に以下の部分のフォルダ先が構築したい PC の設定と合ってるか確認する．




5. 環境変数の Path に以下を追加する．
   “C:¥Program Files¥SASHome¥SASFoundation¥9.4¥core¥sasext”
6. SAS 側の設定：C:¥.....¥SASHome¥SASFoundation¥9.4 直下の sasv9.cfg を以下のように修正し，UTF-8 環境で実行されるように設定した．（この変更をするとバッチサブミットなどに影響があるので変更する際は取扱いに充分注意を擦る．）
   - 修正前：C:¥...¥SASHome¥SASFoundation¥9.4¥nls¥**ja**¥sasv9.cfg
   - 修正後：C:¥...¥SASHome¥SASFoundation¥9.4¥nls¥**u8**¥sasv9.cfg

② 拡張ツール（nbsphinx）の設定と使用例（Read the docs のテーマ設定を含む）

1. コマンドプロンプトを立ち上げて以下のコマンドを入力し，nbsphinx と sphinx_rtd_theme をインストールする．
   A) pip install nbsphinx
   B) pip install sphinx_rtd_theme
2. コマンドプロンプトで sphinx ドキュメントを作成したいフォルダに移行し，以下のコマンドを入力する．その後、コマンドの指示通りにドキュメント名やヴァージン情報を入力する．
   A) python -m sphinx.quickstart

3. 実行後に作成された“conf.py”を以下の内容を追加し，保存する．

```
extensions = [
    'nbsphinx',
    'sphinx.ext.mathjax',
]

exclude_patterns = ['_build', '**.ipynb_checkpoints']
```

```
html_theme = "sphinx_rtd_theme"
```

4. .ipynb ファイルを“conf.py”があるフォルダに入れる．
5. index.rst に.ipynb ファイルのファイル名を記載する．
6. コマンドプロンプトで以下のコマンドを入力すると html 形式のドキュメントが作成される．

   A) python -m sphinx . _build/html

# SAS/EGプロジェクト実行時のエラーハンドリングについて
## －任意の処理単位で分岐処理を実行する方法－

岩本 光司

（武田ＰＲＡ開発センター株式会社 クリニカルデータマネジメント&テクノロジー）

SAS/EGで、複数の処理を組み合わせたSASプログラムを実行し、その途中でエラーが発生した場合に、対応した処理を行いたい場合がある。
プログラムの作成段階において、SASプログラマが、SAS/EGのログおよびログの要約を参照してエラーハンドリングしながら、SDTM変換プログラムをSAS/EG上でドメイン単位毎に作成する。
プログラムの実行段階において、データマネジャーが、SAS/EG上に配置されたプロジェクトを実行してSDTMを作成する。
SAS/EG上に配置されたプロジェクトには数十のデータセットを作成するプログラムが含まれており、また、それに付随する幾つかの処理が含まれる。この実行時に、何らかの原因により不正なRawデータ(データセットや変数の欠落、文字長が異なる、文字･数値変数の違い、Nullレコード等)が含まれることにより、処理実行途中でエラーが発生する場合がある。SAS/EGでは、エラーが発生しても最後の処理まで実行するか、または、途中で処理を停止するので、エラーが発生時に、後続の処理のスキップ、WORKとは異なる一時的に作成したデータの削除、エラー処理などの分岐処理を実施したい場合は、エラーハンドリング対応が必要となる。
これまでは、データステップやプロシジャ単位で、SYSERRをグローバルマクロ変数に引継いでエラー処理する方法を利用していた。しかし、この方法では、個々のプログラム作成段階で組み込む必要があり、検証プロセスの負担となっていた。また、SAS/DIを利用していたこともあって、プログラムの可読性にやや難があった。
今回は、SAS/EGのWorkspace Serverログを参照する方法について検討した。
SAS/EGのWorkspace Serverログは、SAS/EGを起動した際にユーザーIDおよび起動ジョブID毎に作成され、起動中に実行したプロセスのログを蓄積し、SAS/EGの終了時にクローズする。このログの中から、任意の処理単位（複数のデータステップまたはプロシジャを含む）で、エラーメッセージを検出して、エラーが発生した場合に指定した処理を行う方法である。上記で紹介したSDTM作成プログラム実行した際、各ドメイン作成するプログラムが終了する毎および各種処理終了後に、Workspace Serverログを参照して、エラー検出を行い、エラーが発生している場合は、以降の処理をスキップして指定した終了処理を実施するようなプログラムを挿入した。このWorkspace Serverログには、SAS/EGが生成するログとSAS/EG上で実行したSASプログラムのログがあるので、エラー発生の有無に係わらず、SASプログラムの実行ログを抽出して、保存できるようにした。
今回の検討の結果、メリットとして、Workspace Serverログを利用する方法では、SYSERRで挟み込まないのでプログラムがシンプルになり、可読性が向上する。また、個々のプログラムに組み込む必要がないので、個々のプログラムの検証プロセスに影響しないと考える。また、これらのことが品質向上に繋がると思われる。デメリットとして、Workspace Serverログは、SAS/EGの起動後にプロジェクトを複数回実行した場合、蓄積されてしまうので、１回のログを保存するには実行毎に再起動する必要がある。

# 医薬品開発におけるSASプログラムおよび関連ファイルのバージョン管理と共有について

○川上貴弘
（電助システムズ株式会社）

## 1　はじめに

医薬品開発におけるSASプログラムの実行においては、プログラムそのものだけでなく、実行ログやデータセット、結果を出力するためのテンプレート、出力帳票等、多くのファイルが密接に関連しており、バージョン管理やその保管は非常に気を使う作業である。

これらのファイルの管理には、1. フォルダやファイル名を日付やバージョンで分ける、2. バージョン管理システム（ Version Control System、以下VCSという）を利用する等の方法がある。

フォルダやファイル名の変更による管理手法はルールさえ決めれば比較的容易に誰にでも実現できるが、人の手による作業のため、ルール不遵守や不注意による誤操作でミスが発生したり、可能性は低いものの、悪意のある操作でファイルが失われてしまったりする可能性がある。当社でも重要なファイルが誤った操作により移動されてしまうというケースが何度か発生しており、フォルダへのアクセス権管理やファイル操作に関するルールの徹底はもちろんであるが、根本的な対策が必要であった。

そこで、プログラムの社内標準化に合わせて上記のファイルを安全に管理するために、フォルダやファイル名を日付で管理する手法から、オープンソースのVCS及び自社製SASプログラム管理ツール「SPITS」の効果的導入を検討した。

VCSではプログラムのバージョン管理を目的としているため、履歴参照や再現機能等、プログラムの管理に秀でている。2015年には「小さな組織がSAS Programのバージョン管理をするには」と題したセッションにてGitLabの紹介があったが、現在はクラウドをベースとしてさらに多くのサービスがリリースされており、手軽に始めることができる。しかしながら、当社では導入実績がなく、新しいシステムを導入することによる学習コストやクラウドならではの問題が想定された。

また、もう1つの選択肢であるSPITSは、一般的なVSCとは異なり、SASプログラムのバージョン管理のために開発された、SASプログラム、実行時ログ、アウトプット、データセット等を自動的に保存する機能を有するツールである。

社内標準化プロジェクトを機に、社内でより効果的に利用するための手段を合わせて検討した。

## 2　背景

当社では、基本的に社内サーバの共有フォルダにSASプログラムが保存されている。しかしながら、前述の通り、フォルダやファイル名による管理に不都合が生じてきている。これは、SASプログラムを使用（作成、実行）する社員が増えたことにより、ルールに対する意識レベルにバラツキが生じたこと、長期間の運用に伴って必要な修正が行われ、過去バージョンのファイルが増えたこと等が原因であると考えられる。

そこで、プログラムを含めたファイルの管理のためにVCSの導入を検討した。1つはgitおよびGitHub、もう1つは当社のSPITSである。

gitはオープンソースのVCSとして最も有名な分散型VCS（プログラムの履歴情報がサーバだけでなく、各ユーザーの手元にもコピーされるため、分散型と言われる）であり、GitHubはgitを使用す

るための最もユーザー数の多いコードホスティングサービス(コード管理を共有するためのwebサービス)である。

また、SPITSは多くのプログラムの管理が必要な特定のプロジェクトにおいて使用されているが、より多くの社員が共通利用するための手段を検討した。

## 3 gitを使用したバージョン管理環境の構築

gitを利用してバージョン管理を行うためには、まずgitのクライアントをインストールする。インストールはインストーラの手順に従っていけばよい。

gitはプログラムの履歴情報がすべて保管されたデータベース（リポジトリという）を管理することができるが、共有のためのリポジトリは別のサービスを利用して作成する必要がある。このために利用するのがGitHubである。

GitHubを利用するためにはまずwebからユーザー登録を行い、PCにインストールしたgitと連携するためのSSHキーの設定を行う。設定完了後、公開：無料／非公開：有料のいずれかを選択してリポジトリを登録する（リモートリポジトリという）。

リモートリポジトリを作成したら、まずはブラウザに管理対象のファイルをドラッグアンドドロップで登録した。これ以降の操作はGitHubを離れ、gitで行う。

なお、当初はすべてコマンドを入力して操作を行うCUI( Character User Interface )であるgit bashを利用したが、使用するコマンドが多く短期間では覚えきれないため、実際にはGUI版ツールであるSource Treeを導入して評価をした。

## 4 gitを使用したバージョン管理

主な流れは以下の通りである。

4.1 マスターとなるリモートリポジトリから自分のPCにリポジトリをコピー（クローンという）する。

4. 2 クローンされたPCのリポジトリに対し、ブランチと呼ばれる作業用コピーを作成する。

4. 3 ブランチのファイルに対して、変更を行う。メモ帳等のテキストエディタでも、SASでも構わない。

4. 4編集が完了したら編集を完了（コミット）する。コミットが完了した直後はPC上のリポジトリのブランチにのみ反映された状態である。

4. 5 ブランチに追加された情報をリモートリポジトリにも登録（プッシュ）する。プッシュした段階では、リモートリポジトリのブランチに変更が反映された状態である。

4. 6変更情報をマスターに登録するために、変更要求（プルリクエスト）を作成する。

4.7 変更をマスターに反映する場合は、取り込み（マージ）を行う。

4. 8作業を開始するときには、最新の情報を取り込んで（プル）から行う。

4. 9手順4. 2~4. 8を複数回行った。

## 5 gitを使用したバージョン管理の評価

5.1 課題

- SASのプログラムを実行するためには、プログラムの一式が必要であるが、プログラム内で指定しているexcelファイルやincludeするプログラムのパス等は、共有フォルダではなく、

git(Source Tree)が管理するPC上のフォルダとなるため、環境に合わせた修正が必要となる。

- 現在一般的にプログラムで使用されているファイルのエンコードはUTF-8であるが、SASのデフォルトはShift-JISであり、文字化けの原因となる。GUIアプリケーション(Source Tree)をShift-JISに切り替えることは可能であるが、GitHubはShift-JISに対応していないので、SASプログラムのエンコードを変更しない限り、文字化けの問題はついてまわる。
- GitHubは無料で提供されているが、履歴情報を保持するリポジトリがクラウド上に存在することになるため、解析結果などの機密情報を管理するには適さない。この点は社内のサーバやAWS, Azure等のプライベートクラウドを用いて、専用のリポジトリをインストール、構築することで解決できる。
- VCSはあくまでもプログラムのバージョン管理であり、SASプログラムを実行した結果を得ることはできない。具体的には、SASの実行結果と実行ログは、バッチ実行で生成されるものは管理対象として認識することができるが、SASのDMS( Display Manager System )上のものは認識できないので、プログラム中で保存先を指定する等の対応が必要となる。

5.2 評価できる点

- Source Treeを利用して管理を行うと、4.3～4.5の手順について、アプリケーションがヒントやアイコンで表示してくれるため、複雑なコマンドや手順を考えることなく、プログラムの作成と管理に集中ができた。
- 編集後には編集前後の比較（diff）が表示され、変更箇所を容易に確認できることから、不要な変更等を防ぐことができ、非常に有用であると感じた。
- 複数のユーザーが同時に操作しても、競合を検出する機能があるため、排他制御を気にする必要はない。
- 操作手順を理解してgitを利用すれば、PC上に不要な作業用フォルダやファイルを作成することなく、安全にプログラムの編集ができると思われる。

## 6 SPITSを使用したバージョン管理環境の構築

SPITSを利用してバージョン管理を行うためには、SPITSのインストールが必要であるが、標準ではローカル環境にすべてのファイルが構築される。その為、まずは共有フォルダを作成し、SPITSのデータベースを共有フォルダに設定してインストールした。

## 7 SPITSを使用したバージョン管理

主な流れは以下の通りである。

7.1 SAS上でSPITSを起動する。

7.2 プログラムを登録する。この時、SASのエディタで編集中のプログラムや、ファイルから登録する等、様々な方法で登録することが可能である。

7.3 プログラムを編集する。この時、SASのエディタだけでなく、登録した外部エディタ（デフォルトではメモ帳）やファイルから再登録することも可能である。

7.4 SASプログラムを実行する。SAS上で起動されているので、メニューから実行を選べばそのままプログラムが実行され、ログが出力され、結果が得られる。プログラムの実行が

完了すると、ログ、結果、およびその他の指定保存ファイルを自動的に認識し、保存することができる。

7.5 実行履歴を参照する。いつ誰が実行したかを記録しており、実行時の保存ファイルはいつでも再現が可能である。

7.6 関連ファイルの管理について。関連ファイル（例えば、テンプレートとして利用するExcelファイル等）を保存するだけのプログラムを登録した。SPITS上で実行する際には保存されたファイルを参照／再現して使用するマクロを使用することもできるため、関連ファイルをSPITSに保存してしまえば、共有フォルダ上のファイルを削除してしまうことも運用上可能である。

7.7 上記手順7.4～7.7を複数回行った。

## 8 SPITSを使用したバージョン管理の評価

8.1 課題

- プログラムの差分比較機能がない。
- gitと異なり、大規模な開発体制は想定していないので、レビュー/承認などの機能はない。
- すべてのファイルをバイナリとして保存するため、リポジトリの容量は大きくなる（大きなファイルの場合で、およそ元ファイルサイズの3倍程度）。
  - 現在でもcompress=binary オプションで2倍程度に抑えることは可能。

8.2 評価

- ファイルは全てSPITSのデータベース上に保存されているので、オリジナルのファイルがどこにあるかは関係ない（削除してしまっても問題ない）
- 使い慣れたSAS、日本語環境で学習コストがかからない

## 9 まとめ

オープンソースで次々に新しい機能が取り入れられるgitには機能面で劣る部分もあるが、SPITSの利点はSASに特化しているところであり、SASプログラムおよび関連ファイルを管理するには一日の長がある。さらに、差分比較、承認の機能を取り入れる等により、より安全なSASプログラム管理ができるようになると期待できる。

## 10 参考文献・サイト


1. 電助システムズ株式会社：SPITS使用の手引(121101A)
2. サルでもわかるGit入門：
   http://www.backlog.jp/git-guide/
3. git：About git
   https://git-scm.com/about
4. Qiita: git, GitHub, Source Treeに関するエントリ
   http://qiita.com/tags/git
   http://qiita.com/tags/github
   http://qiita.com/tags/sourcetree

# xmlv2 エンジンを用いたannotated CRF の作成と検証方法の検討

○三澤 喬

（日本たばこ産業株式会社　医薬事業部）

## 背景

承認申請時の電子データ提出に必要な資料の一つとして，注釈付き症例報告書（Annotated CRF）が存在する。これはCRFにより収集されるデータの各項目と，SDTMデータセットに含まれる変数を対応させて関係を示したものである。CDISCが発行したSDTM Metadata Submission Guideline（SDTM-MSG）に一定のルールの記載があるものの，作成ツールや作成方法はスポンサに委ねられている。

自社で作成プロセス検討の中で，PDFのAnnotation部分を FDF （Forms Data Format）やXFDF（XML Forms Data Format）として書き出すことが可能であることが分かった。さらに，SAS 9.3（TS1M2）よりMAPファイルを簡単に作成出来るようになり，XMLV2エンジンを用いることで，XMLファイルを容易にデータセット化出来ることが分かったため，工数削減及び品質向上を目的とし，手作業でAnnotationを行うのでなく，プログラムを用いて作成する方法を検討した。

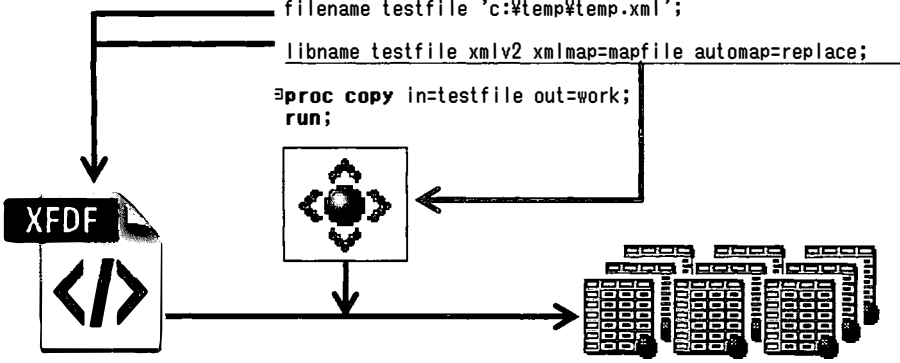
本資料では，XMLV2エンジンの活用方法として，PDFの注釈であるXFDF（XML Forms Data Format）形式のファイルを取り扱う事例を紹介する。実際の活用事例については，本発表のパワーポイント資料及びサンプルプログラムを参照していただきたい。

## XMLV2エンジン及びXFDFファイルの使用方法

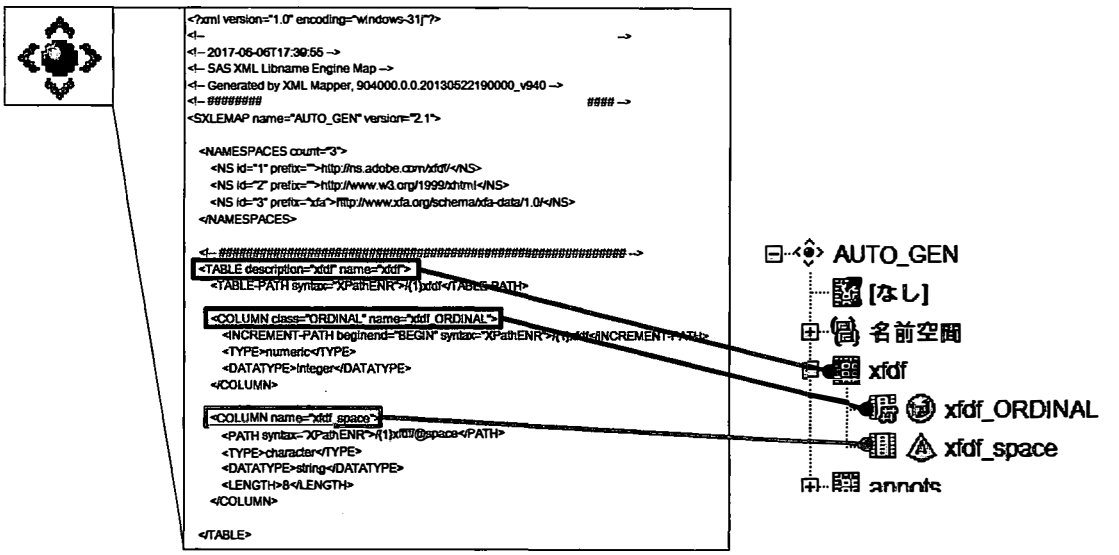
1.　XMLV2エンジンの使用方法

XML形式とは，タグを用いて表現されるテキストファイルである。製薬業界ではCDISCから公表されているdefine.xmlが有名である。プログラムはSASしか学んだことのない筆者は，SASデータセット以外の構造には疎く，以前はSASを用いてテキストとしてXML形式を取り扱っていた。しかしXMLV2エンジンの登場によりXML形式のファイルを，取り扱いやすいSASデータセットとして使用できるようになった。このエンジンは，XML形式のファイルの構造を指定したMAPファイルを用いることで，XML形式のファイルを取り扱いやすいSASデータセットに変換出来る非常に便利なエンジンである。SAS 9.3以前では，MAPファイル自体を作成することが非常に困難であったが，下図のようにSAS 9.3（TS1M2）よりMAPファイルを容易に作成出来るようになり，SASでXML形式のファイルを扱うことが簡単になった。

```
/* Create sample XML file to read in the next step */
libname testfile xmlv2 'c:¥temp¥temp.xml';

proc copy in=sashelp out=testfile;
   select class;
run;

/* Create the XMLMap file in temporary area and use this map to read data*/
filename mapfile  'c:¥temp¥test.map';
filename testfile 'c:¥temp¥temp.xml';

libname testfile xmlv2 xmlmap=mapfile automap=replace;

proc copy in=testfile out=work;
run;
```




MAPファイルとはXMLの構造をSASデータセットへ変換するためのファイルであり，下図のような構造をしている。




一から作成しなければいけなかったMAPファイルが容易に作成出来るようにはなったが，現状はダブルバイトやブランクが含まれる場合，Lengthが短くなることがあり，手作業で修正する必要があるため注意が必要である。

XML file

```
<p ↓
>SCORRES_when_SCTESTCD_=_SUBJINIT
</p>↓
```

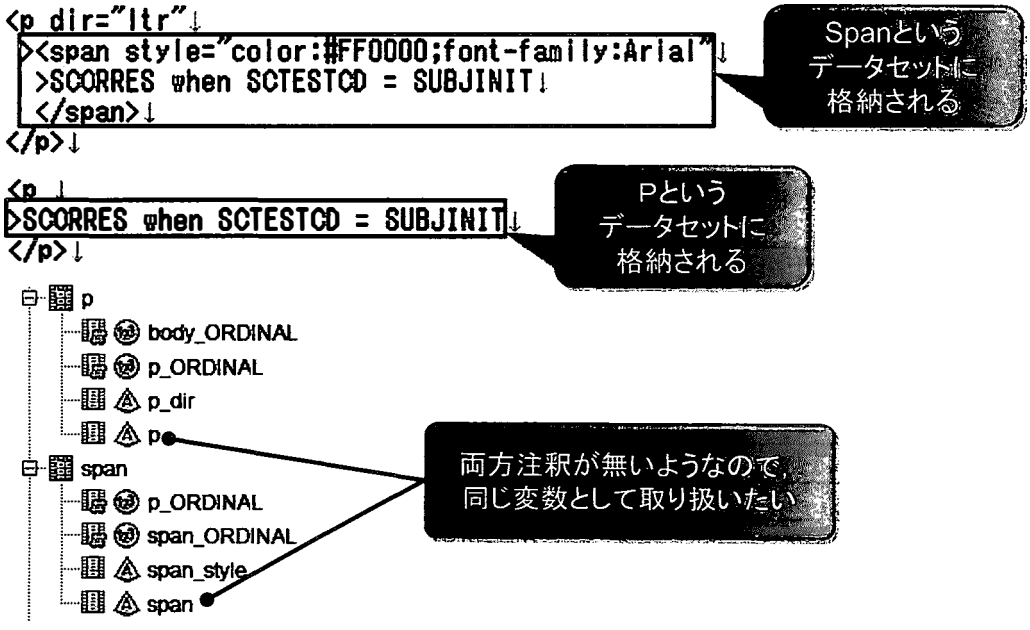
Map file

```
<COLUMN name="p">↓
    <PATH syntax="XPathENR">/{1}xfdf/{1}annots/{1}freetext/{1
    <TYPE>character</TYPE>↓
    <DATATYPE>string</DATATYPE>↓
    <LENGTH>122</LENGTH>↓
</COLUMN>↓
```

ただし，この自動作成したMAPファイルを用いてXMLファイルを読み込むと，ログ画面にMAPファイルとXMLファイルのLengthをWARNINGとして出力してくれるため，多くの時間をかけることなく修正が可能である。

**WARNING: 変数pにデータトランケーションが発生しました:列長=122 追加長=142**

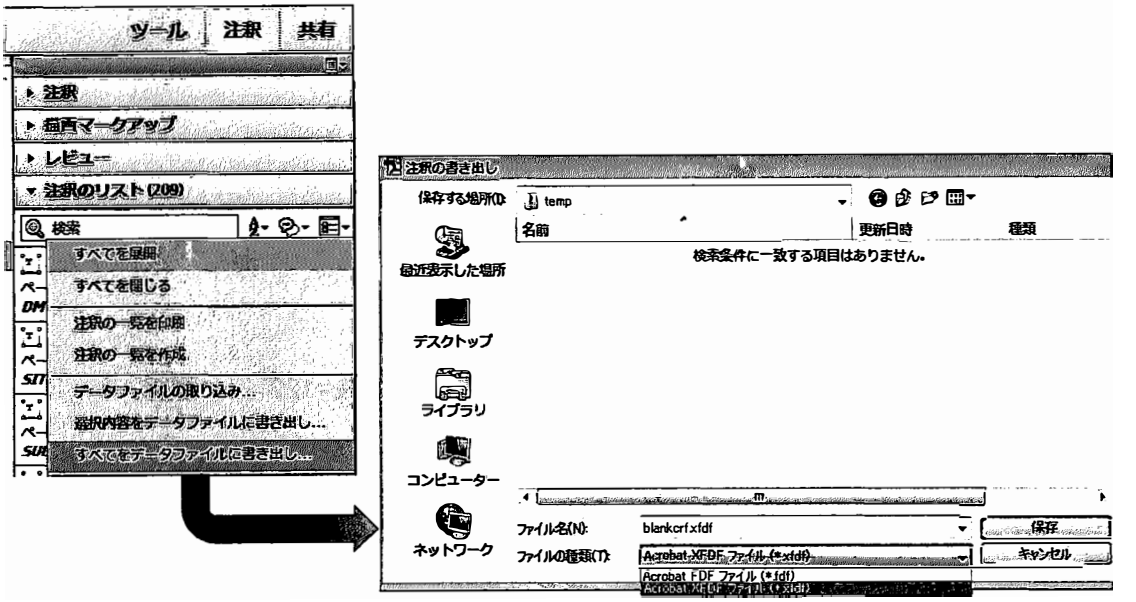
また自動で作成されるMAPファイルは，Element（要素）ごとにデータセットが作成されるため，必要に応じてMAPファイルを加工する必要がある。下図のようなデータをひとつのファイルとして認識させるためには，本発表のパワーポイント資料p. 46～47にある修正が必要である。



2. XFDFファイルの取り扱い

FDF又はXFDFファイルを用いることで，注釈の書き出しや取り込みを行うことが出来る。本資料では，XMLV2エンジンを用いてSASで扱えるように検討を進めたため，XFDFファイルの取り扱いを説明する。

XFDFファイルは，注釈が存在するPDFを開き下図のように[注釈]→[注釈のリスト]→[オプション]→[全てをデータファイルに書き出し]→[Acrobat XFDFファイル]を指定すると書き出しを行うことが可能である。

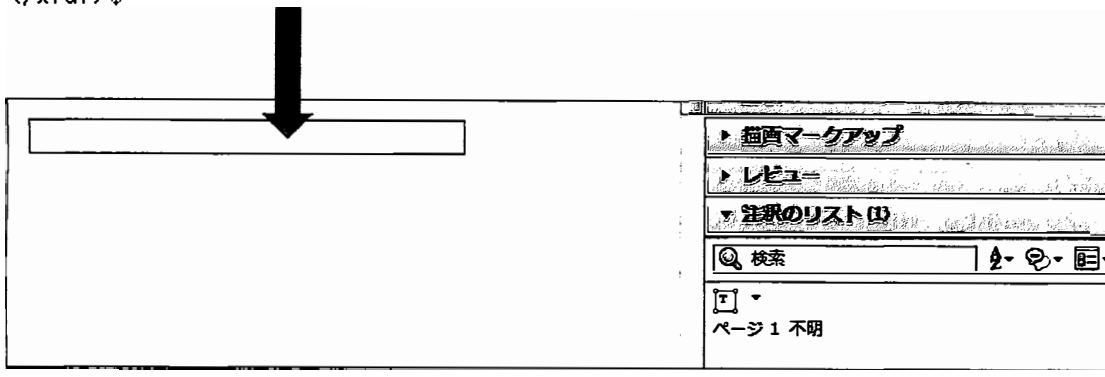



XML Forms Data Format Specification Ver.3.0によると，Annotationに必要な項目は，pageと言うページを指定するattribute(属性)と，rectという場所を指定する属性が最低限必要である。

| Attributes Element | | | |
|---|---|---|---|
| FDF annotation attributes | | Markup annotation attributes | |
| **page** | **Required** | creationdate | Optional |
| Common annotation attributes | | opacity | Optional |
| color | Optional | subject | Optional |
| date | Optional | Border style attributes | |
| flags | Optional | width | Optional |
| name | Optional | dashes | Optional |
| **rect** | **Required** | style | Optional |
| title | Optional | Freetext annotation attributes | |
| | | rotation | Optional |
| | | justification | Optional |
| | | intent | Optional |

しかし，この二つを定義しただけでは，枠のみが出来るだけで注釈が出来るわけではない。

```
<?xml version="1.0" encoding="UTF-8"?>↓
<xfdf xmlns="http://ns.adobe.com/xfdf/" xml:space="preserve">↓
  <annots>↓
    <freetext page="0" rect="50,740,300,760" >↓
    </freetext>↓
  </annots>↓
</xfdf>↓
```

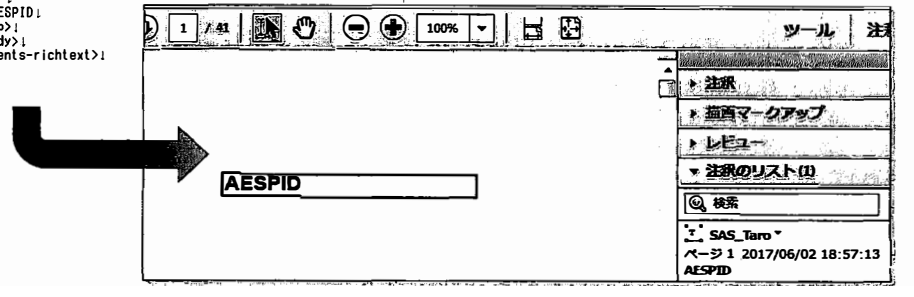



本検証では，注釈を作成するのに最低限必要な情報は下図の項目であると判断した。

なお，dateとtitleという属性は，注釈を確認する際に，作成日と作成者があると確認が容易になるため必要と判断した。

| Freetext Element | Body Element |
|---|---|
| page | style |
| color | P Element |
| date | <表示する文字列> |
| rect | |
| title | |

```
<?xml version="1.0" encoding="UTF-8"?>↓
<xfdf xmlns="http://ns.adobe.com/xfdf/" xml:space="preserve">↓
  <annots>↓
      <freetext color="#BFFFFF" ↓
                date="D:20170602185713+09'00'" ↓
                page="0              " ↓
                rect="50,685,200,700"↓
                title="SAS_Taro">↓
                  <contents-richtext>↓
                    <body style="font-size:12.0pt;text-align:left;color:#FF0000;font-weight:bold;font-style:normal;font-family:Arial;font-stretch:normal"
                     ><p ↓
                      >AESPID↓
                      </p>↓
                    </body>↓
                  </contents-richtext>↓
      </freetext>↓
  </annots>↓
</xfdf>↓
```



CDISCから発行されているdefine.xml v2.0に含まれるannotated CRFの注釈を見てみると，2つの方法で注釈が示されており，spanという要素で示される場合と，pという要素で示される場合が存在する。いずれの方法でも表現することが可能である。本発表のサンプルプログラムでは，pという要素を用いた表現方法を用いている。

```
><p dir="ltr"↓
  ><span style="color:#FF0000;font-family:Arial"↓
   >SCORRES when SCTESTCD = SUBJINIT↓
   </span>↓
 </p>↓
```

```
><p ↓
 >SCORRES when SCTESTCD = SUBJINIT↓
 </p>↓
```

これらのXML形式の取り込み方法について，SASのホームページや海外学会などの発表資料等を基に調査を行ったが，まだ事例が多くはなく，いくつか解決できない問題が発生した。しかしSASテクニカルサポートというSAS社のサービスを用いることで，多くの問題を解決することが出来た。多くの場合，質問をした翌営業日には回答が来るため，非常に有益なサービスである。本資料の調査で非常に丁寧な対応をして頂いた松井 優作様にはこの場を借りて感謝を申し上げたい。

## 活用事例

本発表では，SDTM変数のAnnotation自動作成とAnnotated CRFとdefine.xmlの整合性チェックについて記載をしている。活用事例の具体例はパワーポイント資料を参考にしていただきたい。

## 今後の展望

近年，製薬業界でもdefine-xmlやdataset-xmlなどXML形式のファイルを多く取り扱うようになって来たように感じる。筆者が調べたところ，現在のXMLV2エンジンでは，本資料にあるXFDFファイルのような複雑な構造のXMLファイルをimportすると複数のSASデータセットに分かれるが，これらの一部を修正した後に，元の形式通りの構造でexportすることは非常に困難であることが分かった。今後もXML形式のファイルを使用することが増えてくることが予想されるため，SASデータセットとXML形式ファイルを容易に変換できるようなプログラムを用意いただけると，よりSASが魅力的になると考えている。

また，今後本発表のようなSASを用いたXML形式ファイルの活用事例が多く発表され，XML形式のファイルがより身近になることを強く望む。

## 参照資料


Study Data Tabulation Model Metadata Submission Guidelines (SDTM-MSG) Ver.1.0

CDISC Define-XML Specification Version 2.0

XML Forms Data Format Specification Ver.3.0

Adobe® Acrobat® SDK 8.0 pdfmark Reference, Edition 1.0, November 2006

SAS® 9.4 XML LIBNAME Engine: User's Guide

Sample 47349: Create and use an XMLMap file in a SAS® session

承認申請時の電子データ提出等に関する技術的ガイドについて　平成２７年４月２７日
SDTM Annotations: Automation by implementing a standard process (PharmaSUG 2015) Geo Joy, Novartis
Merge multiple RTF files, convert to PDF and add bookmark (PharmaSUG China 2017) Xiang Wang

# ODS Graphics Designerを用いたグラフの作成：SAS 9.4の機能を活用

○平井 隆幸[1]、魚住 龍史[2]

([1]日本化薬株式会社解析チーム、[2]京都大学大学院医学研究科)

Creating Graphs Using the ODS Graphics Designer with Advanced Features: Application Based on SAS 9.4

Takayuki Hirai[1] and Ryuji Uozumi[2]

[1]Biostatistics Team, Nippon Kayaku Co.,Ltd, [2]Kyoto University Graduate School of Medicine

## 要旨

Output Delivery System (ODS) Graphics Designer とは， SAS データセットに対して，Graphical User Interface (GUI) に基づくグラフ作成を可能とするツールであり，SAS 9.2 (TS2M3) から正規版として利用ができるようになった．ODS Graphics Designer では，グラフギャラリーに様々なテンプレートが用意されており，作成したいグラフの種別を選択し，GUI に基づくカスタマイズを行うことができる．ODS Graphics Designer を用いれば，プログラミングを行う必要がないため，Graph Template Language (GTL) やテンプレートに関する知識がなくても迅速かつ容易にグラフ作成ができ，複雑なグラフ作成のためのカスタマイズも可能である．また，作成したグラフのコードについても，確認することができるため，GTL を用いてグラフ作成を行うプログラマーにとっても，GTL の学習に役立つツールとなる．さらに，SAS 9.4 (TS1M0) から自動チャート機能が拡張され，解析対象となるデータセットと変数を指定し，グラフの種類を選択すれば，事前に用意されたグラフテンプレートから該当するグラフを自動的に作成することも可能となった．本稿では，ODS Graphics Designer による様々な機能を用いたグラフ作成の活用例について報告する．

キーワード：ODS Graphics Designer，メインメニューバー，グラフギャラリー，プロットレイヤーパネル，挿入パネル，データの割り当てダイアログ，ポップアップメニュー，コードウィンドウ，自動チャート

## 1. はじめに

SAS を用いたグラフ作成報告は，これまで多くの報告が行われている．Output Delivery System (ODS) Graphics が評価版であった SAS 9.1.3 (SAS9.1 TS1M3)までは，一部評価版における報告 (吉澤ら, 2005; 長谷川,

2007) はされているものの，多くは GPLOT プロシジャを用いたグラフ作成報告 (林ら, 2003) や，GREPLAY プロシジャを用いた 1 ページに複数のグラフを表示する報告 (竹田, 2004) 等，SAS/GRAPH に基づく報告が行われている．2008 年 9 月に ODS Graphics が正規版となった SAS 9.2 (TS1M0) がリリースされた当初は，普及段階であったせいか，あるいは ODS Graphics を利用するには SAS/GRAPH のライセンスが必要であったせいか，ODS Graphics を用いたグラフ作成報告はほとんど行われていなかった． しかし，2011 年に SAS9.3 (TS1M0) がリリースされ (SAS Institute Inc., 2011), ODS Graphics は SAS/GRAPH から BASE/SAS の機能として組み込まれたため，BASE/SAS のみのライセンスで利用が可能となった．また同年に， SAS ハンドブック (宮岡・吉澤, 2011) 等において ODS Graphics の機能の解説も充実し始めたため，SAS 9.3 (TS1M0)リリース以降，ODS Graphics を用いたグラフ作成の報告が増加している．特に，GPLOT プロシジャに代わる新たなプロシジャとして，SGPLOT プロシジャによるデータの可視化の有用性について，多くの報告が行われている (魚住・浜田, 2011, 2012; 西本, 2013; 豊泉ら, 2014; 魚住ら, 2016). また医薬品開発における実務的な応用として，Statistical Graphics (SG) プロシジャや Graph Template Language (GTL) を用いたグラフ簡易作成ツールの開発事例についても報告されている (高浪, 2011, 2012). さらに山崎ら (2016) は，推移図や散布図等において，グラフで表現しきれなかった情報を付加した新たなグラフについて，SAS 9.4 の SG プロシジャを用いた作成方法を報告している．

このように時代とともに，SAS/GRAPH から ODS Graphics へ移行し，ODS Graphics の機能を利用したグラフ作成の報告が増えてきている．しかしながら，その多くはコードベースの報告である．SAS のプログラミングに馴染みのないユーザーにとっては，データや解析結果の可視化および解釈を行うにあたって，コードベースでグラフの作成方法を理解して作成することは難しく，時間を要するといった支障がある．

以上のような問題を抱えるユーザーに対し，もう一つのグラフ作成手段として，ODS Graphics Designer が提供されている．ODS Graphics Designer では，グラフギャラリーに様々なテンプレートが用意されており，作成したいグラフの種別を選択し, Graphical User Interface (GUI) に基づくカスタマイズを行うことができる．ODS Graphics Designer を用いれば，プログラミングを行う必要がないため，SAS のプログラミングに馴染みのないユーザーにとっても有用なツールとなる．また, SAS 9.4 (TS1M0) から自動チャート機能が拡張され，解析対象となるデータセットと変数を指定し，グラフの種類を選択すれば，事前に用意されたグラフテンプレートから該当するグラフを自動的に作成することも可能となった (SAS Institute Inc., 2015). SAS 9.2 がリリースされた頃から, ODS Graphics Designer は有用なツールとして宣伝されていたが (SAS Institute Inc., 2009), SAS ユーザー総会における報告は行われていない．そこで，本稿では，ODS Graphics Designer の利用方法について解説を行い，ODS Graphics Designer よる様々な機能を用いたグラフ作成の活用例について報告する．

## 2. ODS Graphics Designer の基本的な利用方法

### 2.1 ODS Graphics Designer の起動

通常，ODS Graphics Designer は，SAS プログラム・ウィンドウ環境における**メニューバーからツールをクリックし，プルダウンメニューから ODS Graphics Designer をクリック**して起動する．もしくは，プログラムエディタから**%SGDESIGN;**とコーディングし実行することで起動する ( SAS 9.2 (TS2M3)までは，起動方法はコーディングのみ). 他にも，SAS Enterprise Guide や SAS Studio を用いても起動できる．

SAS Enterprise Guide 7.1 : メニューバーのタスク ⇒ グラフ ⇒ ODS Graphics Designer を開く．

SAS Studio 3.5 : ボタン ⇒ ツール ⇒ ODS Graphics Designer で起動.

## 2.2 User Interface

ODS Graphics Designer を起動すると，図 1 に示したウィンドウが立ち上がる．主な操作方法を下記に示す．

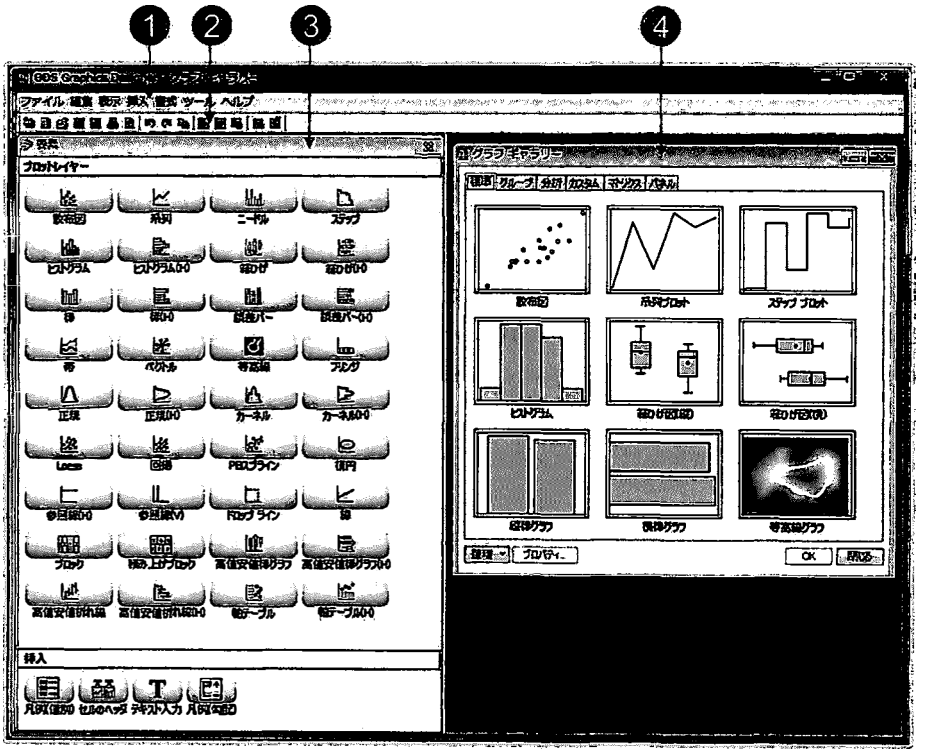



図 1 : ODS Graphics Designer の起動画面

① **メインメニューバー**

ODS Graphics Designer で行う様々なタスクを実行するメニューが用意されている．

**ファイル**：グラフを新規作成，開く，閉じる，編集，保存，グラフギャラリーに保存，印刷，終了．

**編集**：元に戻す，やり直し，すべて選択，コピー．

**表示**：グラフギャラリー，要素ペイン，グラフのコードを開く or 閉じる．

**挿入**：タイトル，フットノート，凡例 (全般) の挿入，行追加，列追加．

**書式**：グラフのスタイルの変更，各種プロパティからのカスタマイズ．

**ツール**：スタイルエディタツール，プリファレンスダイアログ，自動チャート機能の起動．

**ヘルプ**：ヘルプトピックスからオンラインヘルプ ODS Graphics Designer のページリンクへ移動．

② **ツールバー**

メインメニューバーの**ファイル**，**挿入**メニューのタスクを実行するアイコンが用意されている．

③ **要素ペイン**

グラフの中に，**プロットレイヤーパネル**から様々なプロットや線を，**挿入パネル**から凡例やヘッダー，フリーテキスト等をドラッグアンドドロップで挿入できる．

④ **グラフギャラリー**

事前に定義された一般的に使用されるグラフのギャラリーである．ギャラリーはグループ毎に分けられており，各グループは Tab で表示されている．デフォルトで図 2 に示す 6 グループが用意されている．

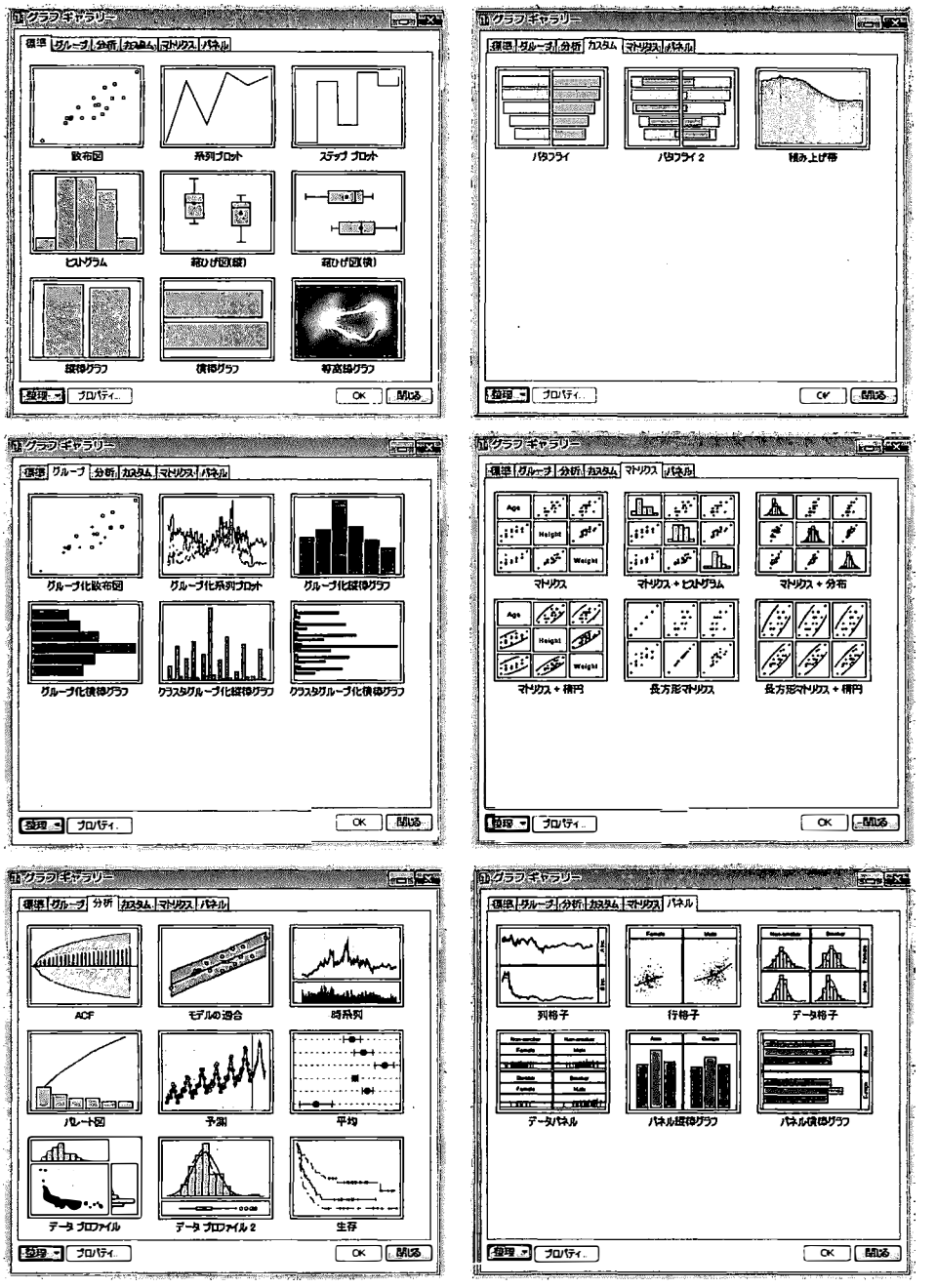



図 2：グラフギャラリーにおける 6 つのグループ

## 3. ODS Graphics Designer を用いたグラフの作成方法

新規にグラフを作成する際は，**①空のグラフに全ての要素を挿入していく方法**と，**②グラフギャラリーから事前に定義された一般的に使用されるグラフを選択しカスタマイズする方法**の 2 つがある．それぞれの作成方法を，例題を基に確認することで，ODS Graphics Designer を用いた作成方法の概要を説明する．

※③空の共有変数グラフに全ての要素を挿入していく方法もあるが，内容が多岐に渡るため割愛する．

## 3.1 性別の投与群別積み上げ棒グラフ

CDISC サンプルデータセット ADSL (Subject-Level Analy sisDataset: 被験者レベル解析データセット) を用いて，性別の投与群別積み上げ棒グラフを作成することで**①空のグラフに全ての要素を挿入していく方法の概要**を説明する．ADSL は，被験者識別コード，投与群，被験者背景情報等が含まれる 1 被験者 1 レコードのデータセットである．事前に SAS 社サポートページより SAS(R) Macros to support Dataset-XML v1. 0.をダウンロードし，WORK フォルダーにデータセット ADSL を取り込む．その後，解析対象集団を FAS に絞り，性別と投与群で並び替えてから利用した (付録 A).

1. **新規の空のグラフを作成する．**
   メインメニューバーから，**ファイル ⇒ 新規作成 ⇒ 空のグラフをクリック**すると，ワークスペースに空のグラフウィンドウが立ち上がる．
2. **空のグラフウィンドウ上に棒グラフアイコンを追加する．**
   要素ペインの**プロットレイヤーパネル**から，**棒グラフ**アイコンをドラッグアンドドロップする．

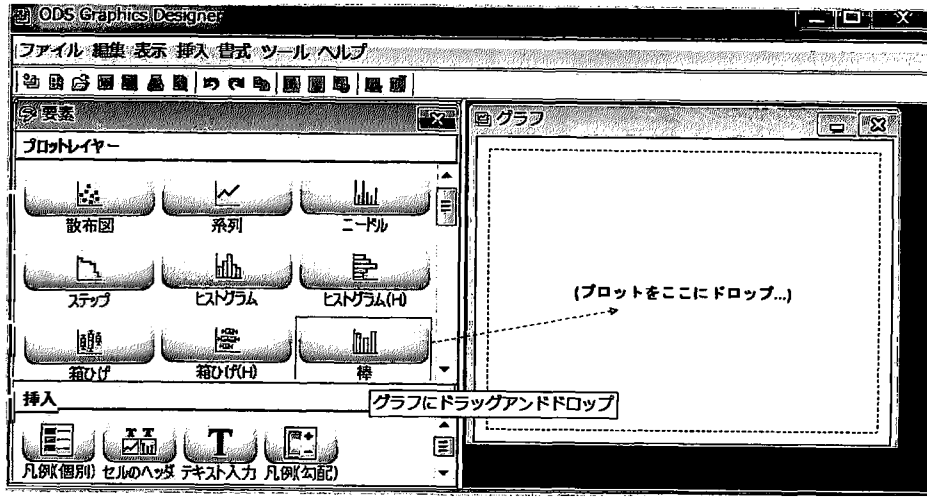



図 3：棒グラフ作成を選択する画面イメージ

3. **表示されるデータの割り当てダイアログに，ライブラリとデータ，各変数を割り当てる．**
   ライブラリ：**WORK**，データ：**ADSL**，変数カテゴリ：**SEX ⇒ OK**．性別の棒グラフ **(bar)** ができる．

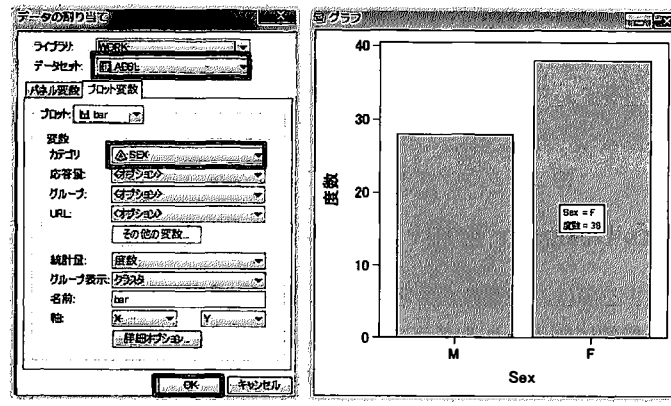



図 4：性別の棒グラフ

4. **グループ変数を割り当て，グループ表示を変更し，性別の投与群別積み上げ棒グラフに変更する．**
作成したグラフを右クリックし表示されるポップアップメニューから，**データの割り当て**をクリックする．ダイアログが再度立ち上がるため，**グループ変数 : ARM**，**グループ表示 : 積み上げ ⇒ OK** をクリックする．なお，グループ表示にはクラスタ (グループ値を隣接する個別の棒として表示) も用意されている．またデータが投与群別に並び替えられていない場合は，表示順が不揃いとなる．

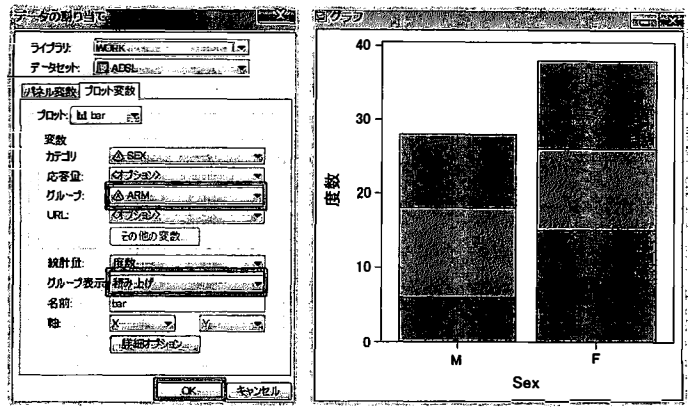



図 5：性別の投与群別積み上げ棒グラフの作成

5. **凡例の追加後，棒の幅を調整し性別の投与群別積み上げ棒グラフの完成**

i. メインメニューバーから，**挿入 ⇒ 凡例（全般）**をクリックする．凡例（全般）ダイアログの**プロットを選択**から，**bar にチェック**を入れ，OK をクリックする．グラフの下に**凡例が追加**される．

ii. グラフを右クリックし表示されるポップアップメニューから，**プロットのプロパティ**をクリックする．セルのプロパティダイアログが立ち上がるため，**プロット bar の棒 Tab** をクリックし，**棒の幅を 0.4** に変更し OK をクリックする．性別の投与群別積み上げ棒グラフが作成できる．

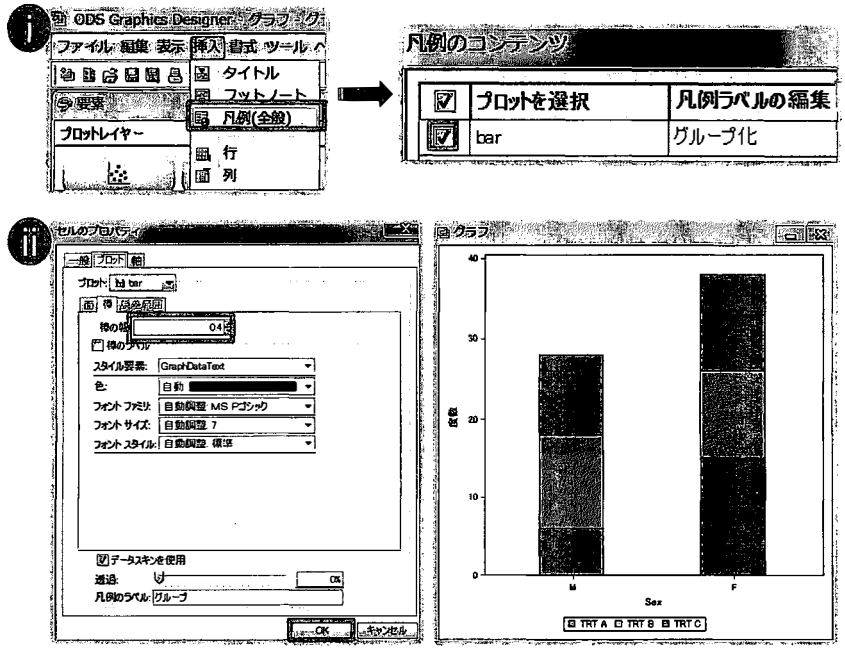



図 6：凡例表示，棒幅調整済，性別の投与群別積み上げ棒グラフの作成

### 3.2 経時データの推移図

次に，高浪 (2011)で用いられたデータセット_Sum_G を用いて，平均値推移図を作成することで**②グラフギャラリーからカスタマイズする方法の概要**を説明する．本稿では，網掛けを設定する代わりに，凡例を追加し描図する．_Sum_G には，ある臨床検査値 (変数 :VAL) について，VISIT : 検査時点，TREAT : 治療群，_MEAN : 平均，_STD : 標準偏差，VISIT2 : 系列線用検査時点，_STDU/_STDL : ±標準偏差が含まれる．公開されている SAS プログラムを実行し，WORK フォルダーにデータ_Sum_G を取り込み利用する．

1. **グラフギャラリーから，カスタマイズに用いる事前に定義されたグラフを選択する．**
   グラフギャラリーから **分析 Tab** を選択し，**平均** をダブルクリックする．デフォルトデータを用いて事前に定義された**平均のグラフ**と**データの割り当てダイアログ**が立ち上がる．この時点で上部に**タイトルプレースホルダー**，下部に**フットノートプレースホルダー**を持つ．プレースホルダーをダブルクリックすると，入力が可能となるため，タイトルとフットノートを必要に応じて入力する．

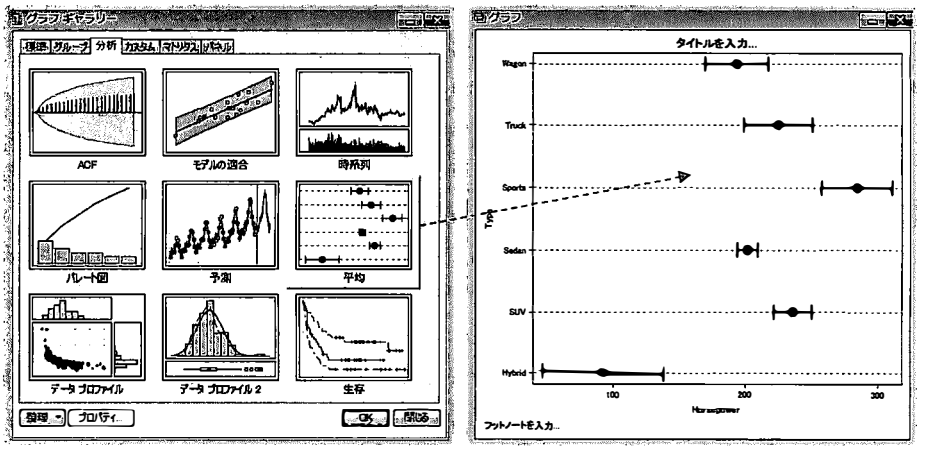



図 7 : グラフギャラリーから平均のグラフ作成画面

2. **表示されるデータの割り当てダイアログに，ライブラリとデータ，各変数を割り当てる．**
   **ライブラリ : WORK，データ : _Sum_G，変数 X : Visit，Y : _Mean，グループ : Treat** を選択する．さらに**その他の変数**をクリックし，**その他の変数ダイアログ**から変数 **Y エラー上部 : _STDU，Y エラー下部 : _STDL** を選択し，**OK** をクリックする．データの割り当てダイアログ下部より，OK をクリックする．平均値±標準偏差の推移図が表示される．

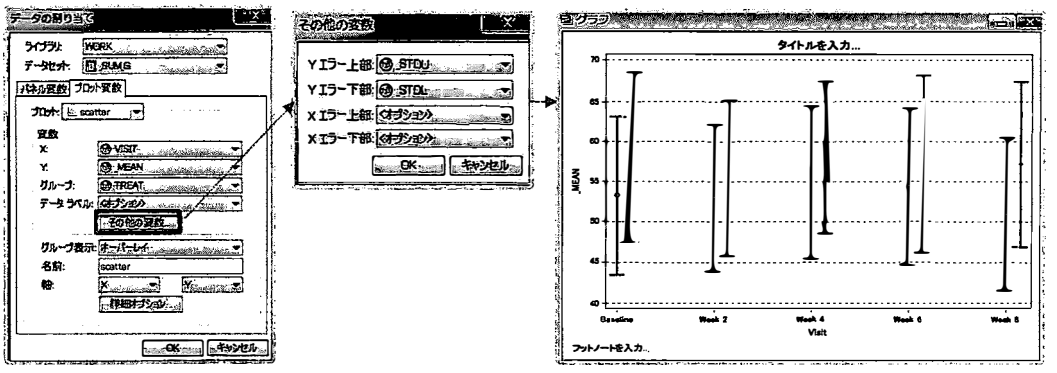



図 8 : 平均値±標準偏差の推移図作成画面

3. **系列線の追加**

作成した平均値±標準偏差の推移図において，平均値の Visit 間を系列線で結ぶことを考える．**プロットレイヤーから系列**を選択し，作成途中のグラフにドラッグアンドドロップする．

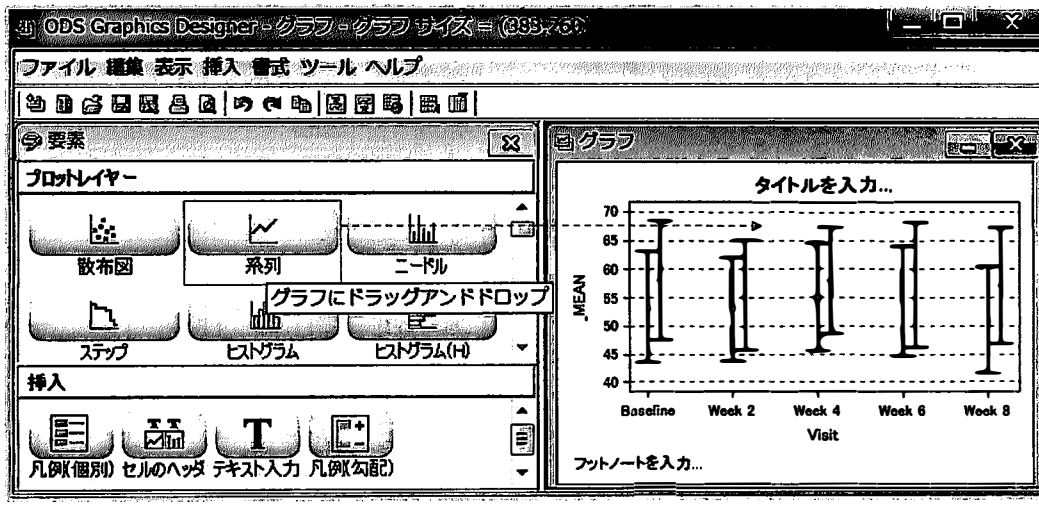



図 9：平均値±標準偏差の推移図に系列線を追加する画面

4. **系列線 (Series) のデータの割り当て**

表示される系列に対するデータの割り当てダイアログに，各変数を割り当てる．

**変数 X : Visit2，Y : _Mean，グループ : Treat** を選択し，**OK** をクリックする．

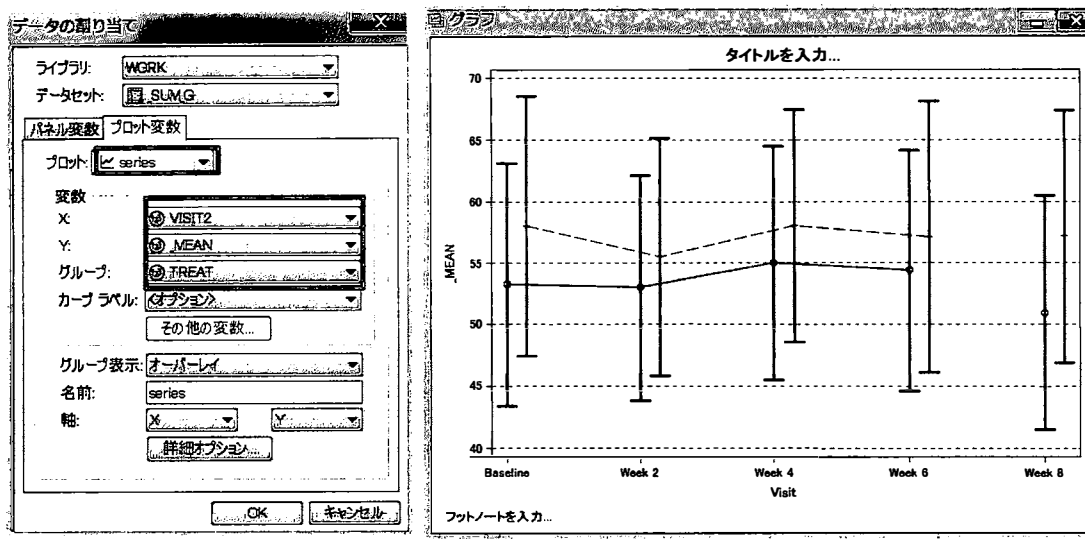



図 10：系列線のデータの割り当てと割り当て後のグラフ画面

5. **凡例の追加後，表示凡例を絞り，平均値±標準偏差の推移図の完成**

**i.** 3.1 節では，凡例をグラフセル領域の外側に挿入した．ここでは，内側に挿入したい．**挿入パネル**から**凡例(個別)**を挿入したい位置（下）にドラッグアンドドロップする．**凡例(個別)** が挿入される．

ii. マーカー記号と系列ともに表示されるため，凡例上で右クリックし表示されるポップアップメニューから，**凡例のコンテンツ**をクリックする．凡例のコンテンツダイアログが表示されるため，**Scatter** のチェックを外す．系列のみ凡例を表示させた平均値±標準偏差の推移図が作成される (図 11).

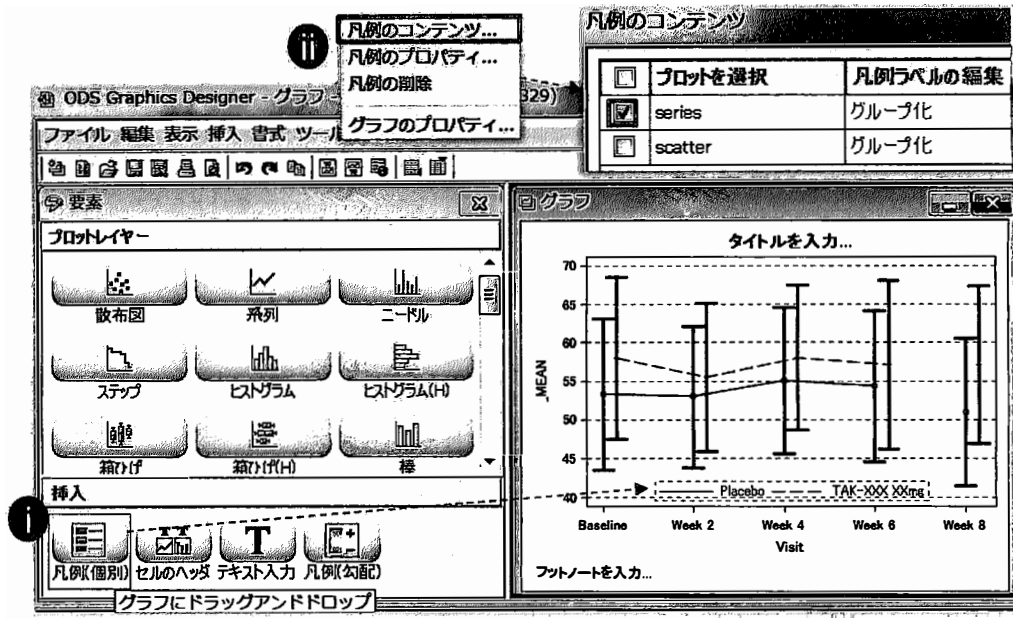



図 11：平均値±標準偏差の推移図に凡例を追加し表示凡例を絞る画面

## 4. GTL コードの確認

ODS Graphics Designer では，作成したグラフの GTL コードについても，確認することができる．メインメニューバーから，**表示 ⇒ コード** をクリックする．**コードウィンドウ**が立ち上がり，確認できる．またコードウィンドウを表示したまま，グラフに修正をかけると，修正した部分のコードが変更されるため，直感的に GTL を理解することも可能である．GTL を用いてグラフ作成を行うプログラマーにとっても，GTL の学習に役立つツールとなる．**コードウィンドウは読み取り専用**であるため，SAS のエディターに **Copy & Paste** し，SAS ファイルとして保存することで，コードベースの編集並びに実行が可能となる．

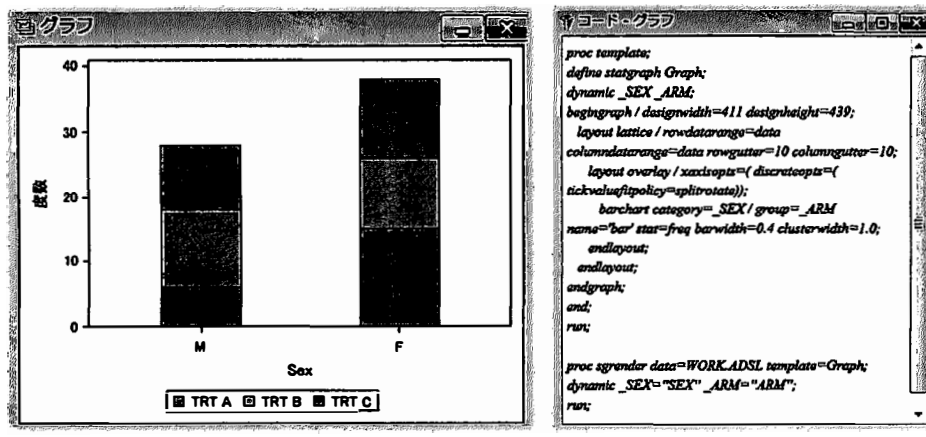



図 12：性別の投与群別積み上げ棒グラフの GTL コード

# 5. グラフの保存 (ファイル or グラフギャラリー)

## 5.1 SGD ファイルへの保存

ODS Graphics Designer では，メタファイルとなる SGD ファイルにグラフを保存する．SGD は，SAS ODS Graphics Designer 用の拡張子である．SGD ファイルは，グラフを再度編集できるように，GTL コード情報や関連するその他の情報を含んでいる．保存手順を下記に示す．

1. メインメニューバーから**ファイル ⇒ 保存**
   ※右の保存ダイアログが立ち上がる．
2. **保存**からファイルの保存先を選択する．
3. **ファイルの名**を入力する．
4. **ファイルのタイプ**から **SGD ファイル**をリスト選択し，保存をクリックする．

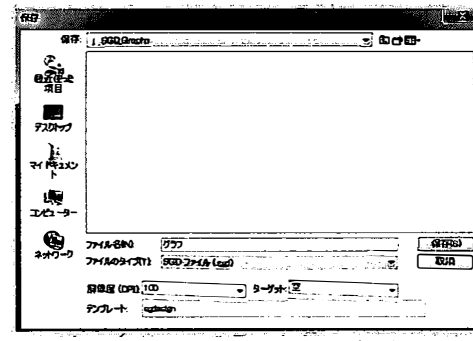



図 13 : SGD ファイル保存画面

## 5.2 Image ファイルへの保存

Image ファイルにグラフを保存することも可能である．ラスター形式として PNG (デフォルトフォーマット)，BMP，GIF，JPG，TIF，ベクター形式として EMF，SVG，PDF，PS といったフォーマットをサポートしている．PNG に関しては，保存前に **Dots per inch (DPI) の指定が可能**であり，高解像度の画像を得たい場合，DPI を大きい数値に変更して保存する．保存手順 1～3 は 5.1 節と同様であるため，手順 4 のみ下記に示す．

4. ファイルのタイプから希望の Image フォーマットをリスト選択し，保存をクリックする．図 14 は PNG ファイルを選択し，DPI を 300 として画像を得る設定である．

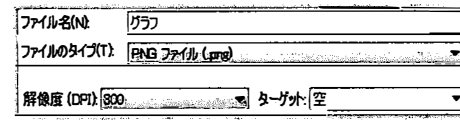



図 14 : PNG ファイル保存画面

## 5.3 グラフギャラリーへの保存

グラフギャラリーに保存も可能である．既存の 6 グループ Tab に追加は出来ないため，新規 Tab を追加し，グラフを保存する．新規 Tab への追加手順は，下記となる．

1. **メインメニューバー : ファイル ⇒ グラフギャラリー**に保存
   ※**グラフギャラリーに保存ダイアログ**が立ち上がる．
2. グラフを保存したいグループ Tab の**グループ名**を入力する．
   新規の場合は，次の手順に従う．
   a. をクリックする．
      ※新しいグループの作成ダイアログが立ち上がる．
   b. 新しいグループの名前を入力し，OK をクリックする．
      ※グラフギャラリーダイアログに名前が表示される．
3. グラフ名を入力する．

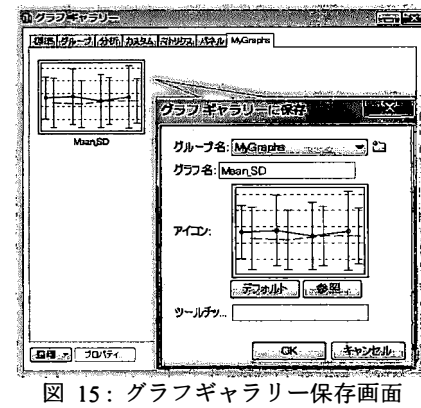



図 15 : グラフギャラリー保存画面

4. **アイコン**は，デフォルトでは作成したグラフのイメージとなるが，変更したい場合は変更可能である．参照をクリックし，変更したいイメージを選択し，開くをクリックする．
5. **ツールチップ**は，デフォルトでは作成したグラフの名前を表示するが，変更したい場合は変更可能である．ツールチップに表示したい内容を記入する．
6. OK をクリックすると，**新規 Tab が追加**され，コンテンツとして，グラフのアイコンが表示される．

# 6. グラフのカスタマイズ

グラフのカスタマイズは，グラフ上で右クリックし表示されるポップアップメニューから，**グラフのプロパティ**をクリックし表示される**グラフのプロパティダイアログ (全般 Tab** と**グループ属性 Tab)** から行う．主な操作方法を下記に示す．またセルのカスタマイズも，プロットのプロパティダイアログ **(一般 Tab，プロット Tab** と**軸 Tab)** から行うことができるが，本稿では説明を割愛する．

## 6.1 グラフ全般のカスタマイズ

グラフを掲載する際，カラー印刷を許容されない場合，モノクロ印刷されても識別しやすいよう，スタイルを変更したい場合がある．全般 Tab のスタイルから，モノクロの場合は Journal を選択し，OK をクリックすることで変更が可能である (図 16 左).

また棒グラフや円グラフの塗りつぶし部分の表示の見栄えをよくしたい場合，データスキンを適用できる．全般 Tab のデータスキンから，変更が可能である．同様に全般 Tab の背景色枠の色から背景色を変更できる．外枠はチェックを外すことで変更できる．サイズ (ピクセル) 枠の幅と高さから，サイズ変更ができる (デフォルトでは，幅 :640 ピクセル，高さ 480 ピクセル)．縦横比の維持はチェックを外すことで変更できる．サブピクセルは，より細かい表示が必要なグラフ等に用いると，細部を詳細に表示できる．例として，全般 Tab からデータスキン：光沢，背景色：灰色に変更した性別の投与群別積み上げ棒グラフを示す (図 16 右)．

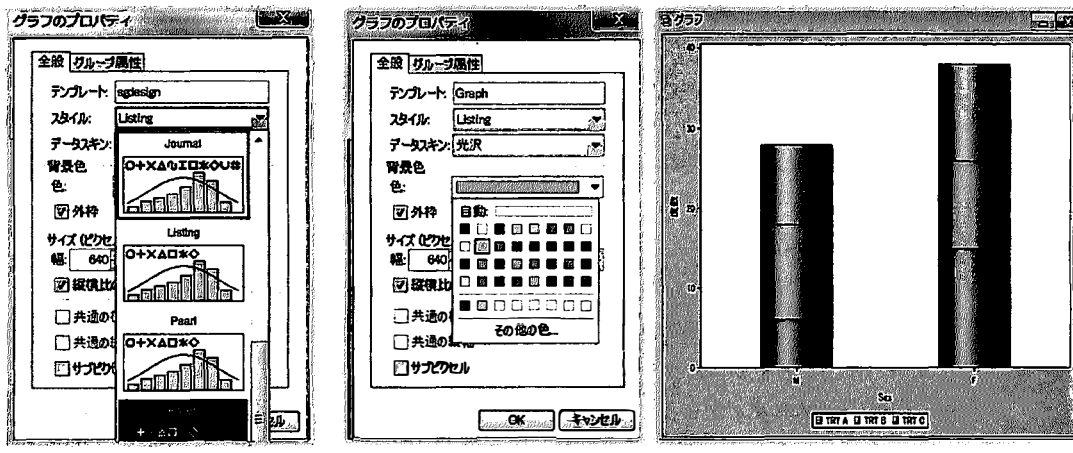



図 16：スタイル変更画面とデータスキンおよび背景色を変更した性別の投与群別積み上げ棒グラフ

## 6.2 グループ属性のカスタマイズ

コードベースにおいては，SAS9.4 より追加された **Styleattrs** ステートメントに **datacolors=(色 1 色 2 ...)**，**datacontrastcolors=(色 1 色 2 ...)**，**datasymbols=(マーカー記号 1 マーカー記号 2 ...)**，**datalinepatterns=(線のパターン 1 線のパターン 2 ...)** を指定することで，グループ属性をカスタマイズすることが可能となった．ODS Graphics Designer においても，GUI ベースにおいて，同様のカスタマイズを行うことが可能である．

1. グラフ上で右クリックし，グラフのプロパティをクリックする．**グループ属性 Tab** から，**種類：色，コントラストの色，マーカー記号，線のパターン**を選択し，開始スタイルをクリックするとデフォルトで指定されている色，コントラストの色，マーカー記号，線のパターンリストが表示される．

2. 上から順に設定されているため，1 番上の色，コントラストの色，マーカー記号，線のパターンのリストボックスをクリックする．色とコントラストの色の場合はパレットが表示され，表示したい色を選択し，OK をクリックすることで，変更可能である．マーカー記号と線のパターンの場合は，各種マーカー記号並びに線のパターンが表示される為，選択し変更可能である．さらに色とコントラストの色においては，その他の色オプションが用意されており，より多くの色を選択可能となる．

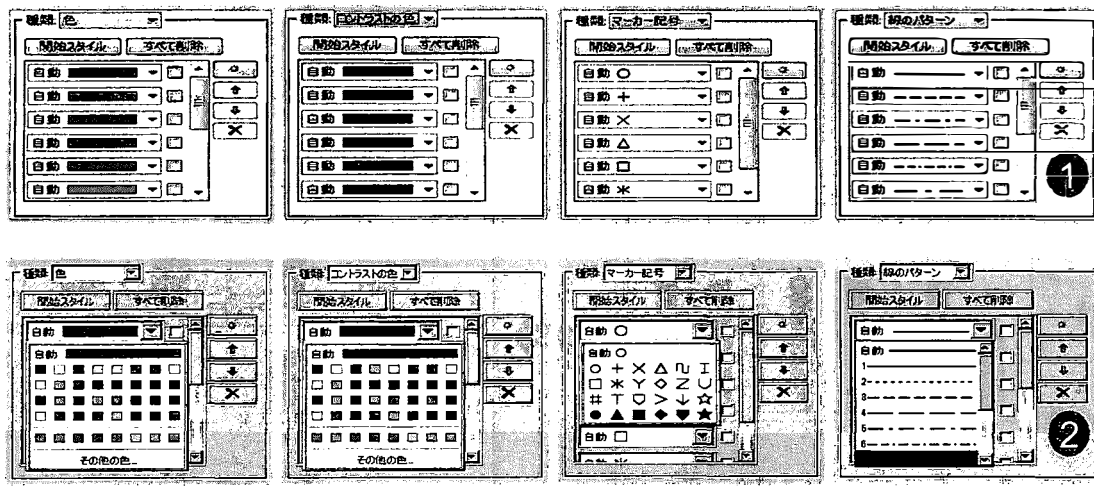


図 17：グループ属性のカスタマイズ画面手順 1 及び 2

# 7. 複数グラフの表示

## 7.1 行・列挿入に基づきセル分割し，複数のグラフ表示

セル (表示領域) に行列挿入を行い分割し，分割したセルにプロットレイヤーパネル等からグラフ要素を挿入し，複数グラフを表示することが可能である．メインメニューバーから挿入 ⇒ 行 or 列 でセルを分割し，以降は 3.1 節同様，分割した各セルに，プロットレイヤーパネル等から要素を追加しグラフ作成を行う．

## 7.2 パネル変数 (分類変数) の水準に基づきセル分割し，各水準に基づくグラフ表示

パネル変数（最大 2 変数まで）の水準に基づき，セル (表示領域) を分割し，分割したセルに各水準に基づくグラフを表示することが可能である．この機能は，SAS プログラムにおける SGPANEL プロシジャによるグラフ出力に相当する．レイアウトは**格子**あるいは**パネル**を選択することが可能である．表示例がグラフギャラリーのパネル Tab (図 2) に示されている．3.1 節で作成したグラフを性別の投与群別積み上げ棒グラフではなく，投与群の 3 水準に基づきセルを分割し性別の棒グラフを表示したい場合は，作成したグラフを右クリックし表示されるポップアップメニューから，**データの割り当て**をクリックする．データの割り当てダイアログが再度立ち上がるため，パネル変数 Tab をクリックし，**データ格子を選択**し，**列変数に ARM** を割り当てることで投与群の水準に基づく分割表示が可能である．パネル変数 Tab から，**データパネルを選択**し，**Class1 変数に ARM** を割り当てた場合は，パネルレイアウトとなる．しかしながらパネルレイアウトは，行列の水準数は同じと定義されているようで，必要のないセルも表示されてしまう欠点がある．

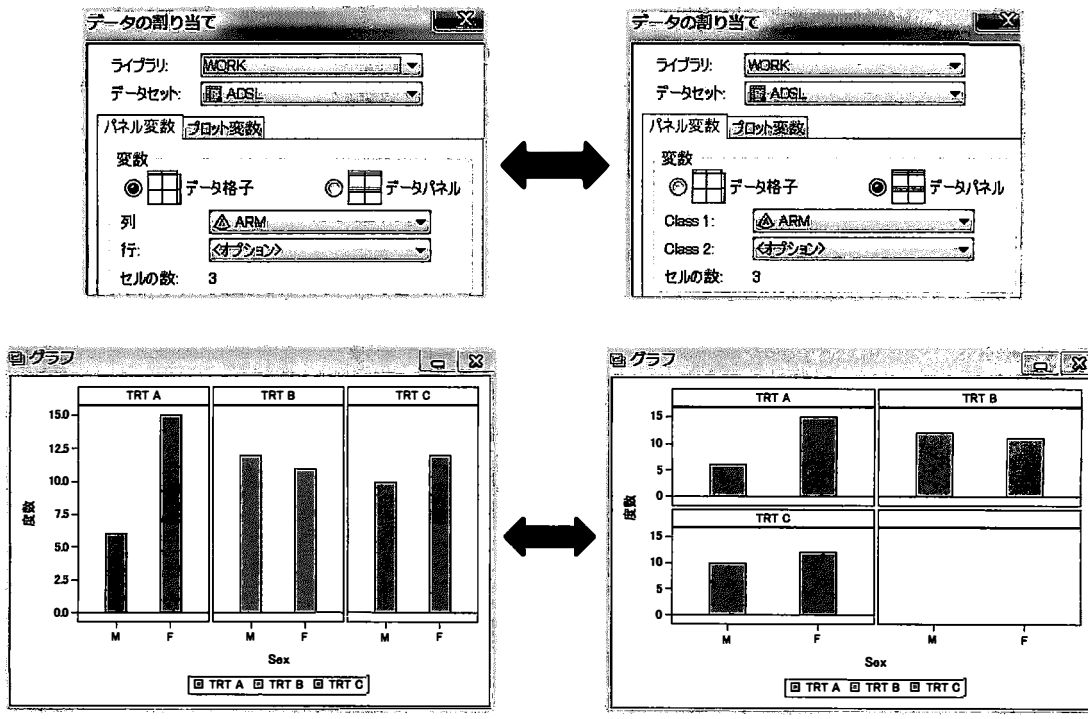



図 18：投与群のデータ格子 (左) or パネル(右) 表示による性別の棒グラフ

## 8. Batch モードでのグラフの実行

SGD ファイルとして保存したグラフは，SGDESIGN プロシジャを用いることによって，通常の SAS 環境やバッチでの実行も可能である．プロシジャの基本構文は，次の通りである．

```
/*SGDファイルの保存先とファイル名を指定*/
Proc Sgdesign sgd="C:¥SGD_Graph¥Bar.sgd";run;
```

同じ変数を持つ異なるデータに対して，SGD ファイルに保存されているグラフを作成したい場合，すなわち，ある特定集団にデータを絞り同様のグラフを作成したい場合には，Data=オプションを用いる．

```
/*SGDファイルの保存先とファイル名を指定，別のDataを指定*/
Data ADSL_C;set ADSL;where RACE="CAUCASIAN";run;
Proc Sgdesign sgd="C:¥SGD_Graph¥Bar.sgd" Data=ADSL_C;run;
```

さらに異なる変数を持つ異なるデータに対して，SGD ファイルに保存されているグラフを作成したい場合，すなわち，臨床検査値からバイタルサインに，データ並びに変数を変更し同様のグラフを作成したい場合には，共有変数グラフを用いる．共有変数は，GTL における DYNAMIC ステートメントで指定する DYNAMIC 変数のことで，Sgdesign プロシジャ実行時に値を与えることで，マクロ変数のように用いることが可能である．本稿では，共有変数グラフについての説明を割愛したため，こちらについても説明は割愛させて頂く．

# 9. 自動チャートの利用方法

SAS 9.4 (TS1M0) から自動チャート (Auto Chart) 機能が拡張され，解析対象となるデータセットと変数を指定し，グラフの種類を選択すれば，既存のテンプレートから該当するグラフを自動的に生成することが可能である．操作方法における重要な点を下記に示す．

## 9.1 自動チャートの起動

ODS Graphics Designer を起動後，メインメニューバーから，**ツール ⇒ 自動チャートをクリック**すると，自動チャートウィンドウが立ち上がる (図 19).

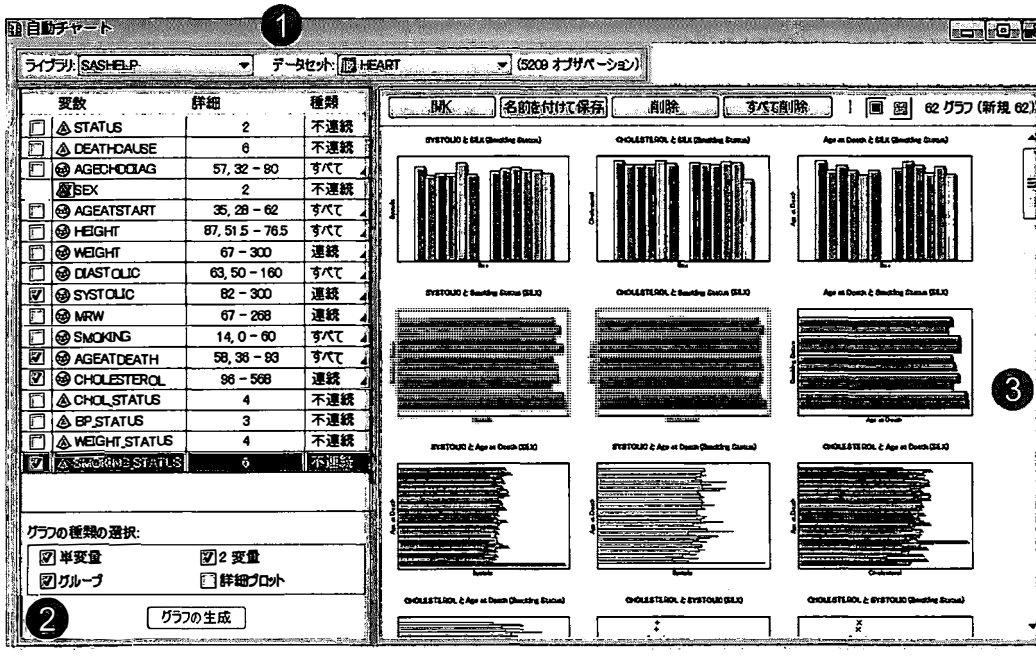



図 19：自動チャートウィンドウ

① 上部にあるライブラリとデータセットのリストボックスから，グラフ作成に用いるデータのライブラリとデータセット名をリスト選択する．

② 左部には，指定されたデータセットに含まれる変数の選択欄並びにグラフの種類の選択欄が用意されている．描図したい変数を 1 つ以上選択し，描図したいグラフの種類を選択する．選択後，グラフの生成をクリックすると，グラフの生成が自動的に開始される．

③ 右部には，自動チャート機能により，生成されたグラフのアイコンが表示される．グラフの中に，プロットレイヤーパネルから様々なプロットや線を，挿入パネルから凡例やヘッダー，フリーテキスト等をドラッグアンドドロップで挿入できる．

## 9.2 自動チャートの使い方

自動チャートを利用しグラフを作成する手順は，2 段階に別れる．はじめに，①自動チャート機能を用いてグラフを生成する．その後，②生成されたグラフの中から適切なグラフを選択し，詳細な部分をカスタマイズする．次の例題を用いて，概要を説明する．

### 9.2.1 自動チャートの使用例

例として，SASHELP ライブラリに保存されているデータセット HEART を用いる．心疾患の原因を探るために，フレーミンハムにおいて行った疫学研究データである．対象はフレーミンハムに住む健康な男女 5209 人 (30 - 62 歳) で 2 年に 1 度，身体検査，心臓に関する各種検査，アンケートが行われ，追跡中に心疾患を発症した人のデータをもとに心疾患の原因を探る研究が行われた．こちらのデータに自動チャートを用いて，喫煙状況別の男女の死亡年齢の平均値に対する横棒グラフを作成する．

**<1st step.> グラフ生成**

1. メインメニューバーから **ツール ⇒ 自動チャート** を選択し，自動チャートウィンドウを起動する．
2. 上部よりグラフ作成に用いるデータの**ライブラリ**と**データセット**をリスト選択する．(図 19 ①)
   a. **ライブラリ**から **SASHELP** 選択
   b. **データセット**から **HEART** 選択
3. 左部に指定されたデータセットに含まれる変数が表示されるため，描図したい変数 SEX，SYSTOLIC，AGEATDEATH，CHOLESTEROL，SMOKING_STATUS を選択する．(図 19 ②)
4. 左下部より，描図したい**グラフの種類**を選択する．**単変量，2 変量，グループ**を選択する．(図 19 ②) グラフの種類を選択すると，グラフの生成ボタンが利用できる．
5. **グラフの生成**をクリック．生成が開始され，右部に生成されたグラフアイコンが表示される．(図 19 ③)
6. 全てのアイコンは，単一のグラフであることがわかる．
7. 右上部の**削除**並びに**すべて削除**で必要のないグラフアイコンは削除できる．1 つのグラフアイコンをクリックし，**削除**をクリックで 1 つのグラフアイコンを削除できる．複数選択する場合は，Shift + クリックを用いて複数選択し，**削除**をクリックで複数のグラフアイコンを削除できる．**すべて削除**をクリックすると，生成されたすべてのグラフアイコンが削除される．
8. 左下部**グラフの種類**より，**詳細プロット**にもチェックを入れ，**グラフの生成**を再度クリックすると，パネル分類されたグラフや複数のセルを持つグラフ等より詳細なグラフを含むアイコンが追加される．

**<2nd step.> グラフのカスタマイズ**

生成されたグラフの中から適切なグラフを選択（複数選択可）し，詳細な部分をカスタマイズする．

9. 生成されたグラフアイコンの中からカスタマイズするものを選択する．ここでは，喫煙状況別の男女の死亡年齢に対する横棒グラフを選択する．
10. 選択されたグラフの作業領域が表示されるため，以降は 3.2 節同様，要素を追加しカスタマイズを行う．ここでは軸テーブルを追加し，グラフに平均値を追加している (図 20 右).

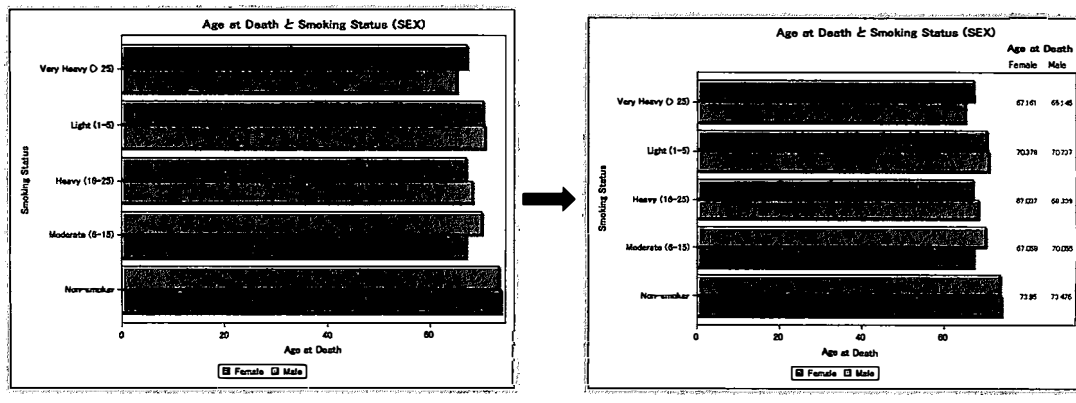



図 20：自動チャート機能による喫煙状況別の男女の死亡年齢に対する横棒グラフ

## 10. まとめ

本稿では，ODS Graphics Designer の操作方法を詳述し，例として，積み上げ棒グラフおよび経時データの推移図を作成することで ODS Graphics Designer におけるグラフ作成方法について，解説した．なお，より複雑なカスタマイズを要する例として，Kaplan-Meier および Forest プロットを作成する方法を付録に掲載した．ODS Graphics Designer は，グラフの作成，カスタマイズや管理に関して，GUI に基づき操作可能なため，SAS プログラミングに馴染みのないユーザーにとって，迅速かつ容易にデータを可視化および解釈するためのグラフ作成支援ツールとして，非常に有用なツールである．さらに，グラフのコードの確認やバッチモードでの実行も可能なため，GTL を用いてグラフ作成を行うプログラマーにとっても，GTL の学習ツールとして役立つといえる．SAS 9.4 では，自動チャート機能，コードベースにおける追加機能 (XAXISTABLE 並びに YAXISTABLE ステートメント等) に対応するプロットレイヤーの追加等，様々な拡張も行われている点も魅力的といえる．以上より，もう一つのグラフ作成手段として，ODS Graphics Designer は今後有用なツールになり得ると期待される．

## 参考文献


[1] Battiston C. SAS(R) ODS Graphics Designer – The Next Step in Amazing Data Visualization. Proceedings of the SAS Global Forum. SAS Institute Inc., Cary, NC, 2013. Available at http://support.sas.com/resources/papers/proceedings13/248-2013.pdf.

[2] Cheng W. Statistical Graphics for Clinical Research Using ODS Graphics Designer. Proceedings of the SAS Global Forum. SAS Institute Inc., Cary, NC, 2013. Available at https://support.sas.com/resources/papers/proceedings13/380-2013.pdf.

[3] Matange S. ODS Graphics Designer: An Interactive Tool for Creating Batchable Graphs. Proceedings of the SAS Global Forum. SAS Institute Inc., Cary, NC, 2009. Available at https://support.sas.com/resources/papers/proceedings09/331-2009.pdf.

[4] Matange S. Quick Results with SAS® ODS Graphics Designer. Proceedings of the SAS Global Forum. SAS Institute Inc., Cary, NC, 2012. Available at https://support.sas.com/resources/papers/proceedings12/153-2012.pdf.

[5] Matange S, Bottitta J. SAS ODS Graphics Designer by Example: A Visual Guide to Creating Graphs Interactively. SAS Institute Inc., Cary, NC, 2016.

[6] Muller RD. Charting Your Path to Using the "New" SAS(R) ODS and SG Graphics Successfully. Proceedings of the SAS Global Forum. SAS Institute Inc., Cary, NC, 2017. Available at http://support.sas.com/resources/papers/proceedings17/1369-2017-poster.pdf.

[7] Rebecca O, Leanne G. Getting Your Hands Reproducible Graphs. Proceedings of the SAS Global Forum. SAS Institute Inc., Cary, NC, 2015. Available at https://support.sas.com/resources/papers/proceedings15/3432-2015.pdf.

[8] SAS Institute Inc. SAS Technical News Summer 2009, 2009. Available at https://www.sas.com/offices/asiapacific/japan/periodicals/technews/pdf/09sum.pdf.

[9] SAS Institute Inc. SAS(R) 9.3 ODS Graphics: Procedures Guide (2nd edn.). SAS Institute Inc., Cary, NC, 2011.

[10] SAS Institute Inc. SAS(R) 9.4 ODS Graphics Designer: User's Guide (3rd edn.). SAS Institute Inc., Cary, NC, 2015.

[11] SAS(R) Macros to support Dataset-XML v1.0.0

[12] 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.

[13] 魚住龍史・浜田知久馬. がん臨床試験における腫瘍縮小効果の検討に有用なグラフの作成 -SGPLOT プロシジャの最新機能を活用-. SAS ユーザー総会 論文集 2012, 151-165.

[14] 高浪洋平. SG プロシジャと GTL によるグラフの作成と ODS PDF による統合解析帳票の作成 ～TQT 試験における活用事例～. SAS ユーザー総会 論文集 2011, 201-219.

[15] 高浪洋平. SAS と HTML アプリケーションによる CDISC ADaM 形式の解析用データセットを用いた有害事象の解析帳票・グラフ簡易作成ツールの開発事例. SAS ユーザー総会 論文集 2012, 185-205.

[16] 竹田眞. GREPLAY プロシジャによる 1 枚複数掲載図の作成について. SAS ユーザー総会 論文集 2004, 465-474.
[17] 豊泉樹一郎・財前政美・北西由武・都地昭夫. ODS GRAPHICS を用いた臨床試験データの可視化への挑戦. SAS ユーザー総会 論文集 2014, 307-323.
[18] 西本尚樹. データの可視化を加速する SAS/JMP のグラフ新機能－ODS Graphics と JMP グラフビルダー－. SAS ユーザー総会 2013.
[19] 長谷川要. ODS Statistical Graphics の臨床データへの適用事例. SAS ユーザー総会 論文集 2007, 377-380.
[20] 林行和・畑中雄介・小出起美雅・山口考一. SAS/GRAPH 入門 ~社内における教育研修事例~. SAS ユーザー総会 論文集 2003, 477-486.
[21] 平井隆幸・吉田早織・叶健・魚住龍史. ベクター形式を用いたグラフの作成と有用性. SAS ユーザー総会 論文集 2015, 303-310.
[22] 宮岡悦良・吉澤敦子. SAS ハンドブック. 共立出版, 2011.
[23] 山崎文寛・清水康平・高浪洋平. 臨床試験における SG プロシジャを用いた Graphic Design の考察. SAS ユーザー総会 論文集 2016, 233-249.
[24] 吉澤敦子・小林丈二・緑川修一・高原佳奈・宮岡悦良. ODS Statistical Graphics の紹介. SAS ユーザー総会 論文集 2005, 193-203.

# 連絡先

E-mail : *takayuki.hirai@nipponkayaku.co.jp*

# 付録

## 付録 A : 3.1 節 ADSL 加工 Pgm

```
data adsl; set adsl;if FASFL="Y";run;
proc sort data=adsl;by descending SEX ARMCD;run;
```

## 付録 B : Kaplan-Meier プロット

ODS Graphics Designer を用いた Kaplan-Meier プロットと Forest プロットの作成手順を紹介する. 魚住ら (2016) で, コードベースにおける作成方法が報告されている. 本報告で用いられている ODS Graphics による出力データセット SurvivalPlot, 並びに Forestdata を, ライブラリ WORK に取り込み利用する.

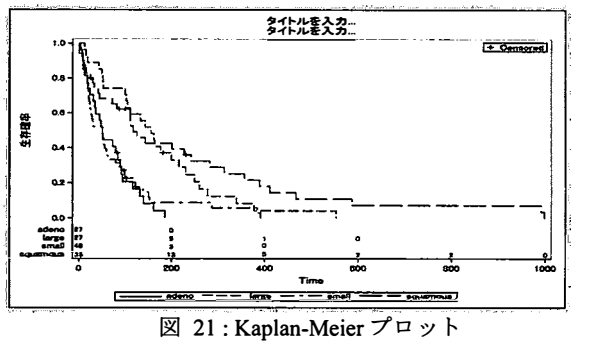



図 21 : Kaplan-Meier プロット

1. メインメニューバー : **ファイル** ⇒ **新規作成** ⇒ **空のグラフ**をクリック.
2. 空のグラフウィンドウ上に**ステッププロット**アイコンをドラッグアンドドロップする.
3. データの割り当てダイアログへの割り当て.
   ライブラリ : WORK, データ : SurvivalPlotData, 変数 X : Time, Y : Survival, グループ : Stratum, 名前を Survival とし, OK をクリック.
4. 凡例の追加
   メインメニューバー : **挿入** ⇒ **凡例 (全般)** をクリック. 凡例 (全般) ダイアログのプロットを選択から, Survival にチェックを入れ, OK をクリック.
5. タイトルの追加
   メインメニューバー : **挿入** ⇒ **タイトル**をクリック. グラフ上にタイトルが挿入される. 再度タイトルをクリックすると 2nd タイトルが挿入される. タイトルテキストの追加後 2nd タイトルのフォントを小さくしておく.

6. CENSORED の散布図を追加する．
プロットレイヤーから**散布図**アイコンを選択し，作成途中のグラフにドラッグアンドドロップする．

7. 散布図 (Scatter) のデータの割り当て
変数 X : Time，Y : Censored，名前 : Censor とし，OK をクリック．

8. マーカーの記号を＋に変更する．
Censor の散布図が挿入された．作成途中のグラフ上で右クリックし，**プロットのプロパティ**をクリック．セルのプロパティダイアログが表示されるため，プロットリストから，Censor を選択する．シンボルを○から＋に変更する．

9. 再度グループ毎の CENSORED の散布図を追加する．
プロットレイヤーから**散布図**アイコンを選択し，作成途中のグラフにドラッグアンドドロップする．

10. グループ毎の散布図 (Scatter) のデータの割り当て
変数 X : Time，Y : Censored，グループ : Stratum，名前 : CensorGroup とし，OK をクリックする．

11. マーカーの記号を＋に変更する．
CensorGroup の散布図が挿入された．作成途中のグラフ上で右クリックし，**プロットのプロパティ**をクリック．セルのプロパティダイアログが表示されるため，プロットリストから，CensorGroup を選択する．シンボルをグループから＋に変更する．

12. Censor プロットに対する凡例の追加
挿入パネル から凡例（個別）を挿入したい位置（今回は右上）にドラッグアンドドロップする．右上に凡例（個別）が挿入される．すべてのプロットの凡例が表示される．追加された凡例上で右クリックし表示されるポップアップメニューから，**凡例のコンテンツ**をクリックする．凡例のコンテンツダイアログが表示されるため，Censor 以外のプロットのチェックを外し，OK をクリックする．

13. リスク集合の追加
プロットレイヤーパネルから**軸テーブル(H)**を選択し，作成途中のグラフにドラッグアンドドロップする．

14. **軸テーブル(H)**のデータの割り当て
表示される軸テーブル(H) に対するデータの割り当てダイアログに，各変数を割り当てる．
X : TatRisk，値 : AtRisk，分類 : Stratum，グループ : Stratum，位置 : 下，名前 : AtRisk，OK クリック．

15. スタイルを Journal に変更し，Kaplan-Meier プロット完成
グラフ上で右クリックし表示されるポップアップメニューから，**グラフのプロパティ**をクリックする．グラフのプロパティダイアログの全般 Tab のスタイルから，Journal を選択し，OK をクリックする．

## 付録 C : Forest プロット

1. メインメニューバー : **ファイル ⇒ 新規作成 ⇒ 空のグラフ**をクリック．

2. 空のグラフウィンドウ上にプロットレイヤーパネルから，**散布図**アイコンをドラッグアンドドロップする．

3. データの割り当てダイアログへの割り当て．
ライブラリ : WORK，データ : Forestdata，変数 X : HAZARDRATIO，Y : SUBGRUOP，OK をクリック．その他の変数をクリック．その他の変数ダイアログから変数 X エラー上部 : HRUPPERCL，X エラー下部 : HRLPPERCL とし，OK をクリック．データの割り当てダイアログも，OK をクリック．Forest プロットが表示される．

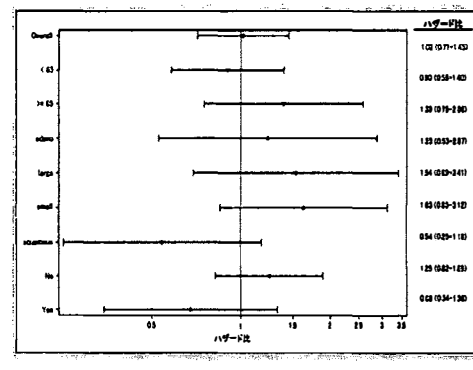


図 22 : Forest プロット

4. マーカーの記号を●に変更する．
作成途中のグラフ上で右クリックし，**プロットのプロパティ**をクリック．セルのプロパティダイアログが表示されるため，プロット Tab，プロット : Scatter のマーカーTab のシンボルを○から●に変更する．

5. X 軸を対数スケールに変更する．Y 軸を反転させる．
**軸 Tab** から軸 : X を選択し，種類を線形から Log10 に変更する．同様に軸 : Y を選択し，**ラベル**のチェックを外す．さらに **Y 軸 Tab** の**一般 SubTab** の反転にチェックを入れて，OK をクリックする．

6. X 軸の値 1 に**参照線(V)**を追加する．
プロットレイヤーから**参照線(V)**アイコンを選択し，作成途中のグラフにドラッグアンドドロップする．**参照線(V)**は，1 の位置にドロップすれば，適切に作成される．ずれた場合は，ドロップ後に表示されるデータの割り当て参照線ダイアログのパラメータ X を 1 に変更する．もし参照線の色や太さを変更した

い場合は，作成途中のグラフ上で右クリックし，**プロットのプロパティ**をクリック．セルのプロパティダイアログが表示されるため，プロット Tab でプロット：Vref の色と幅を希望のものに変更する．

7. HR と 95%信頼区間の値を追加
   グラフ上で右クリックし，**列の追加**をクリックする（通常右に追加される為，左に追加したい場合は，追加後再度作成途中のグラフ上で右クリックし，列の移動で移動する）．プロットレイヤーパネルから**積み上げブロック**を選択し，追加した領域にドラッグアンドドロップする．
8. **積み上げブロック**のデータの割り当て
   X：HAZARDRATIO，ブロック：HR，グループ：SUBGRUOP，軸：X2，OK をクリックする．
9. グラフ上で右クリックし，**プロットのプロパティ**をクリック．セルのプロパティダイアログが表示されるため，**プロット Tab** からプロット：group_block を選択し，**表示 SubTab** の**塗りつぶし，外枠，ラベル**のチェックを外す．**値 SubTab** の配置の横：を左から中央に変更する．一般 **Tab** の**塗りつぶし，外枠**のチェックを外す．軸 Tab の軸：X2 を選択し，値と目盛りのチェックを外す．さらに，**ラベル SubTab** のフォントスタイルを標準から太字に変更し OK をクリックする．
10. Forest プロットと積み上げブロックで作成した HR と 95%信頼区間の表の間にカーソルを合わせ，移動させることで，それぞれのグラフ領域での表示割合を変更し，完成．

# Report Writing Interfaceによるデータステップでの HTML形式レポート作成の検討

望戸　遼

（イーピーエス株式会社）

keyword：HTML、ODS、RWI

## 1.　要旨

SAS9.4よりOutput Derivery System（以下ODS）に追加されたReport Writing Interface（以下RWI）を使用して、HTML形式のレポート作成の実用性と可能性について検討する。

なお、HTML出力には、タグの綺麗さからODS HTML5ステートメントを使用する。

## 2.　RWI

### 2.1. RWI の基本的な記載

RWIはデータステップ内に記載する。RWIでのレポート出力の基本的なステップは以下の6ステップである。単純なテーブルであれば、これら6ステップで出力できる。

step 1.　RWIのODSオブジェクトを定義
step 2.　テーブルを開始
step 3.　行を開始
step 4.　セル（≒列）を出力
step 5.　行を終了
step 6.　テーブルを終了

step 1では、ODSオブジェクトを定義する。オブジェクト名は一文字目はアルファベットから始め32文字以内の英数字とする。

declare odsout *オブジェクト名*();
または、dcl odsout *オブジェクト名*();

step 2~6ではstep 1で定義したODSオブジェクトを利用し、RWIのオブジェクトメソッドを用いて、記載していく。

*オブジェクト名.メソッド名( オプション1, オプション2...)*;

以下に簡単な例を示した。

```
/*program01*/
data _null_;
   set sashelp.class end=end;
   if _n_=1 then do;
      dcl odsout r();                    /*...step 1*/
      r.table_start();                   /*...step 2*/
   end;
      r.row_start();                     /*...step 3*/
        r.format_cell(data:name);        /*...step 4*/
        r.format_cell(data:sex);
        r.format_cell(data:age);
        r.format_cell(data:height);
        r.format_cell(data:weight);
      r.row_end();                       /*...step 5*/
   if end then do;
      r.table_end();                     /*...step 6*/
   end;
run;
```

出力イメージ1

| | | | | |
|---|---|---|---|---|
| Alfred | M | 14 | 69 | 112.5 |
| Alice | F | 13 | 56.5 | 84 |
| Barbara | F | 13 | 65.3 | 98 |
| Carol | F | 14 | 62.8 | 102.5 |
| Henry | M | 14 | 63.5 | 102.5 |
| James | M | 12 | 57.3 | 83 |
| Jane | F | 12 | 59.8 | 84.5 |
| Janet | F | 15 | 62.5 | 112.5 |
| Jeffrey | M | 13 | 62.5 | 84 |
| John | M | 12 | 59 | 99.5 |
| Joyce | F | 11 | 51.3 | 50.5 |
| Judy | F | 14 | 64.3 | 90 |
| Louise | F | 12 | 56.3 | 77 |
| Mary | F | 15 | 66.5 | 112 |
| Philip | M | 16 | 72 | 150 |
| Robert | M | 12 | 64.8 | 128 |
| Ronald | M | 15 | 67 | 133 |
| Thomas | M | 11 | 57.5 | 85 |
| William | M | 15 | 66.5 | 112 |

program01にデータセットclassをRWIで出力するプログラムを、上述のstepと併せて示した。

step 1にてODSオブジェクトrを定義し、step 2にてtable_startメソッドを使いテーブルを開始する。ODSオブジェクトの定義とテーブルの開始は1回行えばよいため「if _n_=1 then do;」にて1obs目だけに処理を限定させる。

step 3にてrow_startメソッドを使い行を開始し、step 4にてformat_cellメソッドのdataオプションよりデータセットclassの各変数のデータを各セルへ出力し、step 5にてrow_endメソッドを使い業を終了する。step 3~5では各obsごとに出力したいため、特段処理を絞り込む条件は必要ない。

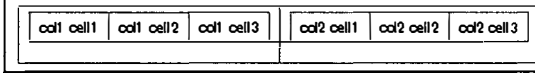
最後にstep 6にてtable_endメソッドを使いテーブルを終了する。テーブルの終了は開始と同様に、最後に1回行えばよいため「if end then do;」にて最終obsだけに処理を限定させる。

program01のままでは、データセットclassの変数の中身だけが出力され、変数名であったり変数ラベルが無いが、「r.table_start();」～「end;」の間に「r.row_start();」を一度かませ、「r.format_cell(data:"xxxxx...")」を変数分繰り返し記載すれば、変数名や変数ラベルでも、任意の列ラベルを出力することが出来る。

入れ子にすることも可能である。つまり、セルの中にテーブルを入れ込むことが可能である。

```
/*program02*/
data _null_;
  dcl odsout r();
  r.table_start();
  r.row_start();
    r.cell_start();
      r.table_start();
      r.row_start();
      r.format_cell(data:"col1 cell1");
      r.format_cell(data:"col1 cell2");
      r.format_cell(data:"col1 cell3");
      r.row_end();
      r.table_end();
    r.cell_end();
    r.cell_start();
      r.table_start();
      r.row_start();
      r.format_cell(data:"col2 cell1");
      r.format_cell(data:"col2 cell2");
      r.format_cell(data:"col2 cell3");
      r.row_end();
      r.table_end();
    r.cell_end();
  r.row_end();
  r.table_end();
run;
```

**出力イメージ2**




program02の例では、行1列2のテーブル（A）、行1列3のテーブル×2（B、C）の3つのテーブルがあり、Aのテーブルの各セルにテーブルBとCが入れ子になっている。

この時点で既存のプロシジャでは不可能であった出力が可能となった。

このようにRWIは、step 1のODSオブジェクトの定義後、step 2~6をデータセットの条件による枝分かれや入れ子を駆使して様々な形にハンドリングし、レポートを出力することが可能である。

## 2.2. RWI を使い HTML を出力する

次に、実用性をイメージした出力をRWIで行いHTMLファイルを出力していく。

なお、出力に用いたsasデータセットはrawdata.sasプログラムより作成したbaseデータセットを用いた。

以下、table_1のような測定値及び変化量の要約統計量を出力するレポートを考えていく。

table_1 要約統計量
full analysis set

| parameter | treatment | timing | value | | | | | | change value | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | n | mean | std | min | median | max | n | mean | std | min | median | max |
| xxxxx | xxxxx | xxxxx | 999 | 999 | 999 | 999 | 999 | 999 | | | | | | |
| | | xxxxx | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 |
| | | xxxxx | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 999 |
| | xxxxx | xxxxx | 999 | 999 | 999 | 999 | 999 | 999 | | | | | | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

table_1

| OBS | sort1 | sort2 | sort3 | out1 | out2 | out3 | out4 | out5 | out6 | out7 | out8 | out9 | out10 | out11 | out12 | out13 | out14 | out15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | systolic blood pressure (mmHg) | treatment a | visit 1 | 10 | 160.3 | 13.00 | 140 | 162.0 | 181 | | | | | | |
| 2 | 1 | 1 | 2 | | | visit 2 | 10 | 126.4 | 15.07 | 92 | 127.5 | 147 | 10 | -33.9 | 23.09 | -74 | -31.5 | 7 |
| 3 | 1 | 1 | 3 | | | visit 3 | 10 | 123.9 | 11.70 | 107 | 123.5 | 143 | 10 | -36.4 | 13.94 | -55 | -33.5 | -12 |
| 4 | 1 | 2 | 1 | | treatment b | visit 1 | 10 | 144.0 | 11.55 | 127 | 143.5 | 167 | | | | | | |
| 5 | 1 | 2 | 2 | | | visit 2 | 10 | 136.7 | 13.31 | 112 | 136.0 | 154 | 10 | -7.3 | 13.85 | -34 | -8.0 | 22 |
| 6 | 1 | 2 | 3 | | | visit 3 | 10 | 130.5 | 18.25 | 104 | 126.5 | 167 | 10 | -13.5 | 18.44 | -41 | -16.0 | 25 |
| 7 | 1 | 3 | 1 | | placebo | visit 1 | 10 | 149.6 | 12.86 | 123 | 150.5 | 170 | | | | | | |
| 8 | 1 | 3 | 2 | | | visit 2 | 10 | 150.8 | 15.84 | 133 | 150.5 | 173 | 10 | 1.2 | 22.74 | -21 | -5.5 | 38 |
| 9 | 1 | 3 | 3 | | | visit 3 | 10 | 155.8 | 15.80 | 134 | 151.5 | 181 | 10 | 6.2 | 22.38 | -29 | 2.0 | 58 |
| 10 | 2 | 1 | 1 | diastolic blood pressure (mmHg) | treatment a | visit 1 | 10 | 102.6 | 7.96 | 93 | 101.0 | 115 | | | | | | |
| 11 | 2 | 1 | 2 | | | visit 2 | 10 | 81.1 | 12.71 | 65 | 83.5 | 100 | 10 | -21.5 | 15.46 | -50 | -24.0 | 2 |
| 12 | 2 | 1 | 3 | | | visit 3 | 10 | 81.3 | 13.18 | 68 | 76.5 | 108 | 10 | -21.3 | 17.46 | -46 | -26.0 | 10 |
| 13 | 2 | 2 | 1 | | treatment b | visit 1 | 10 | 100.2 | 10.56 | 83 | 104.0 | 111 | | | | | | |
| 14 | 2 | 2 | 2 | | | visit 2 | 10 | 88.1 | 13.34 | 71 | 87.0 | 111 | 10 | -12.1 | 21.15 | -40 | -9.0 | 27 |
| 15 | 2 | 2 | 3 | | | visit 3 | 10 | 90.8 | 11.01 | 76 | 88.0 | 111 | 10 | -9.4 | 11.60 | -23 | -11.5 | 12 |
| 16 | 2 | 3 | 1 | | placebo | visit 1 | 10 | 98.6 | 12.40 | 79 | 97.0 | 117 | | | | | | |
| 17 | 2 | 3 | 2 | | | visit 2 | 10 | 103.4 | 11.93 | 90 | 100.0 | 133 | 10 | 4.8 | 17.57 | -21 | 4.0 | 26 |
| 18 | 2 | 3 | 3 | | | visit 3 | 10 | 100.5 | 8.28 | 85 | 100.5 | 112 | 10 | 1.9 | 15.23 | -19 | 0.0 | 29 |

上記に、出力したいsasデータセットtable_1を示した。これを元に出力していく。

```
/*program03*/
proc template;
    define style onlyfontstyle;
        class Container /
            fontfamily=""Times New Roman','MS Gothic'"
            fontsize=10pt
            fontweight=medium
        ;
    end;
run;
```

program03に、フォントを設定したテンプレート「onlyfontstyle」を示した。

「Container」Elementの設定が他のODSテーブルに使われる全てのElementへ引き継がれるので、ここでは簡単のため、文字フォント(fontfamily)、文字サイズ(fontsize)、文字スタイル(fontweight)を設定した。

```
/*program04*/
%let bb=borderbottomcolor=black borderbottomstyle=solid borderbottomwidth=1px;
```

program04に罫線のスタイル一式を指定するマクロ変数「&bb.」を示した。

罫線はStyle Attribute設定を用い、左、右、上、下のどれかを設定すればよい。上記テンプレートでは罫線の設定がゼロであるので、罫線色を黒、罫線スタイルを直線、罫線幅を1pxへ設定し、可視かさせている。

ここでは下線だけであるが、他罫線を引きたいのであれば、「bottom」を「left」「right」「top」にすればよいだけである。

なお、format_cellメソッドには罫線を指定するオプションがあるが、Style Attributeを用いた方が細かく設定できるので便利である。

```
/*program05*/
ods html5 file="&output./table_1.html" style=onlyfontstyle;
data _null_;
  set table_1 end=end;
  by sort1 sort2 sort3;
  if _n_=1 then do;
    dcl odsout r();
    r.table_start();
    r.head_start();
    r.row_start();
      r.format_cell(data:"table_1 要約統計量", just:"l", column_span:16);
    r.row_end();
    r.row_start();
      r.format_cell(data:"full analysis set", just:"l", column_span:16, style_attr:"&bb.");
    r.row_end();
    r.row_start();
      r.format_cell(data:"parameter", row_span:2, style_attr:"&bb.");
      r.format_cell(data:"treatment", row_span:2, style_attr:"&bb.");
      r.format_cell(data:"timing", row_span:2, style_attr:"&bb.");
      r.format_cell(data:"value", just:"c", column_span:6, style_attr:"&bb.");
      r.format_cell(data:"");
      r.format_cell(data:"change value", just:"c", column_span:6, style_attr:"&bb.");
    r.row_end();
    r.row_start();
    do i=1 to 2;
      r.format_cell(data:"n", just:"r", style_attr:"&bb.");
      r.format_cell(data:"mean", just:"r", style_attr:"&bb.");
      r.format_cell(data:"std", just:"r", style_attr:"&bb.");
      r.format_cell(data:"min", just:"r", style_attr:"&bb.");
      r.format_cell(data:"median", just:"r", style_attr:"&bb.");
      r.format_cell(data:"max", just:"r", style_attr:"&bb.");
      if i=1 then r.format_cell(data:"", style_attr:"&bb.");
    end;
    r.row_end();
    r.head_end();
    r.body_start();
  end;
    dmy="";
    array outs{*} out1-out9 dmy out10-out15;
    r.row_start();
    do i=1 to dim(outs);
      if i in(1,2) then do;
        if i=1 then span=9;
        if i=2 then span=3;
        if outs[i]^="" then
          r.format_cell(data:strip(outs[i]), row_span:span, style_attr:"&bb.");
      end;
      else do;
        if i=3 then do;
          if last. sort2 then r.format_cell(data:strip(outs[i]), style_attr:"&bb.");
                         else r.format_cell(data:strip(outs[i]));
        end;
        else do;
          if last. sort2 then r.format_cell(data:strip(outs[i]), just:"r", style_attr:"&bb.");
                         else r.format_cell(data:strip(outs[i]), just:"r");
        end;
      end;
    end;
    r.row_end();
  if end then do;
    r.body_end();
    r.table_end();
  end;
run;
ods html5 close;
```

program03、program04での設定を使い、フォント、罫線の設定を行ったprogram05を示した。

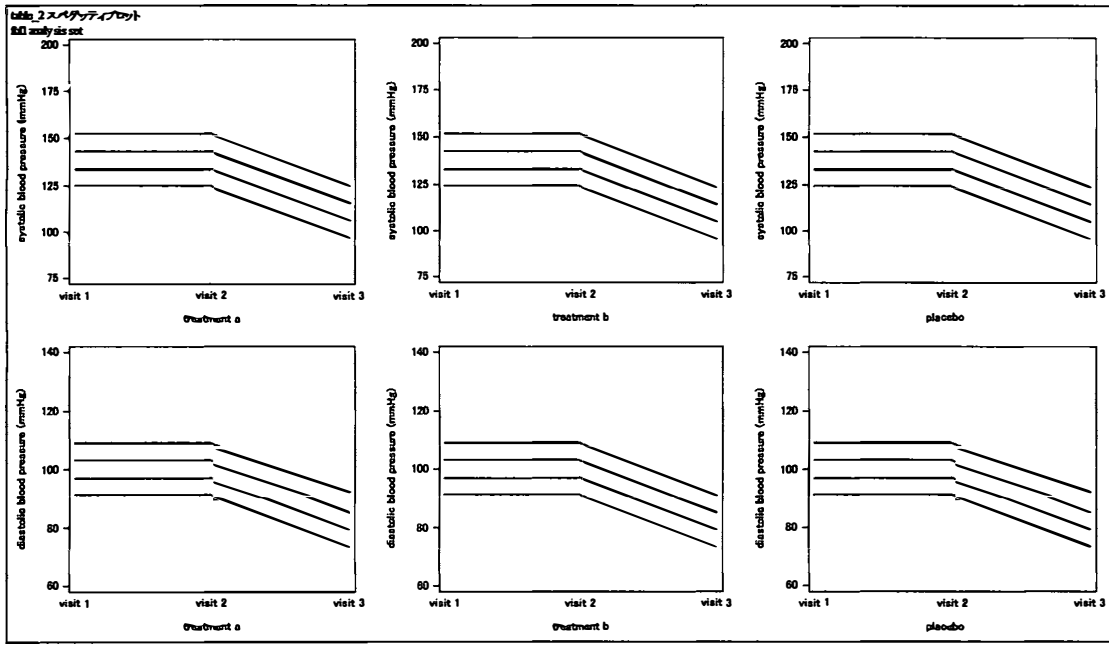
**青色太字**部分で、program03で作成した、「onlyfontstyle」を使ったHTMLファイルを作成するよう指定する。

**赤色太字**で示した format_cell.style_attr オプションにて、program04マクロ変数&bb.の下線設定を反映させる。

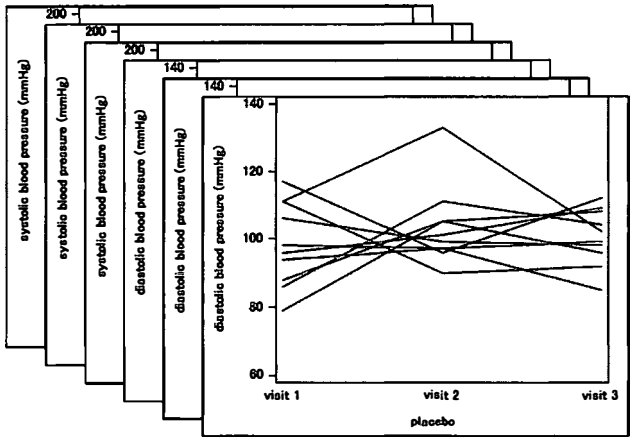
program05で出力されるHTMLファイルを以下に示す。

これで、目標とするtable_1イメージの出力が出来るようになった。

次に以下table_2のような、2パラメータ(systolic blood pressure, diastolic bloodpressure)、3群それぞれのグラフを出力するレポートを考えていく。




3列2行で各群各項目ごとにグラフを出力したいが、例えばGREPRAYプロシジャやGTLの各種レイアウトを使用し1ファイルとして出力することも可能であるが、HTMLファイルであれば1セルに1グラフを埋め込むことが可能なので、本稿では3列2行のセルに1つずつグラフを埋め込んでいく。




左記のように、1パラメータ1群ごとのグラフファイル(table_21.png～table_26.png)を用意した後、出力イメージに基づいて、table_1のときと同様に考えていく。

```
/*program06*/
ods html5 file="&output./table_2.html" style=onlyfontstyle;
data _null_;
    dcl odsout r();
    r.table_start();
        r.head_start();
            r.row_start();
                r.format_cell(data:"table_2 スパゲッティプロット", column_span:3, just:"l");
            r.row_end();
            r.row_start();
                r.format_cell(data:"full analysis set", column_span:3, just:"l");
            r.row_end();
        r.head_end();
        r.body_start();
            do p=1 to 2;
                r.row_start();
                    do a=1 to 3;
                        r.cell_start();
                        r.image(file:cats("table_2",(p-1)*3+a,".png"));
                        r.cell_end();
                    end;
                r.row_end();
            end;
        r.body_end();
    r.table_end();
run;
ods html5 close;
```

**青色太字**で示したimageメソッドが画像ファイルを貼り付けるメソッドとなり、HTMLファイルとしては<img>タグとなり、imageメソッドのfileオプションに指定したファイル名が<img>タグのsrc属性となる。

ここで注意すべき点として、画像ファイルの保存場所がどこかを考えることである。program06の例では画像ファイル名"だけ"を指定しているため、ここで出力されるHTMLファイルと同じディレクトリに存在するtable_21.png～table_26.pngを表示することになり、つまり相対パスで記載していることになる。もし、画像ファイルが必ず1箇所のディレクトリに存在する状況であれば、絶対パスで記載してもよいかもしれない。

program06で出力されるHTMLファイルを以下に示す。

併せて、グラフの画像ファイルを同じディレクトリに保存してtable_2.htmlファイルを表示すれば、目標とするtable_2イメージの出力が出来た。

最後に以下、table_3のような一覧表を出力するレポートを考えていく。

table_3 一覧表
full analysis set

| treatment | subject | parameter | timing | value | change from baseline |
|---|---|---|---|---|---|
| xxxx | xxxx | xxxx | xxxx | 99.9* | |
| | | | xxxx | 99.9 | 99.9 |
| | | | xxxx | 99.9 | 99.9 |
| | | xxxx | xxxx | 99.9* | |
| | | | xxxx | 99.9 | 99.9 |
| | | | xxxx | 99.9 | 99.9 |
| ... | ... | ... | ... | ... | ... |

*:baseline value.

table_3

| OBS | arm | subjid | visitnum | paramn | out1 | out2 | out3 | out4 | out5 | out6 | span1 | span2 | span3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 101 | 1 | 1 | arm:treatment a | subjid:101 | paramsystolic blood pressure (mmHg) | visitvisit 1 | aval:150~{super *} | chg: | 60 | 6 | 3 |
| 2 | 1 | 101 | 2 | 1 | arm: | subjid: | param | visitvisit 2 | aval:129 | chg:-21 | 60 | 6 | 3 |
| 3 | 1 | 101 | 3 | 1 | arm: | subjid: | param | visitvisit 3 | aval:138 | chg:-12 | 60 | 6 | 3 |
| 4 | 1 | 101 | 1 | 2 | arm: | subjid: | paramdiastolic blood pressure (mmHg) | visitvisit 1 | aval:111~{super *} | chg: | 60 | 6 | 3 |
| 5 | 1 | 101 | 2 | 2 | arm: | subjid: | param | visitvisit 2 | aval:84 | chg:-27 | 60 | 6 | 3 |
| 6 | 1 | 101 | 3 | 2 | arm: | subjid: | param | visitvisit 3 | aval:80 | chg:-31 | 60 | 6 | 3 |
| 7 | 1 | 102 | 1 | 1 | arm: | subjid:102 | paramsystolic blood pressure (mmHg) | visitvisit 1 | aval:154~{super *} | chg: | 60 | 6 | 3 |
| 8 | 1 | 102 | 2 | 1 | arm: | subjid: | param | visitvisit 2 | aval:137 | chg:-17 | 60 | 6 | 3 |
| 9 | 1 | 102 | 3 | 1 | arm: | subjid: | param | visitvisit 3 | aval:130 | chg:-24 | 60 | 6 | 3 |
| 10 | 1 | 102 | 1 | 2 | arm: | subjid: | paramdiastolic blood pressure (mmHg) | visitvisit 1 | aval:104~{super *} | chg: | 60 | 6 | 3 |
| 11 | 1 | 102 | 2 | 2 | arm: | subjid: | param | visitvisit 2 | aval:100 | chg:-4 | 60 | 6 | 3 |
| 170 | 3 | 309 | 2 | 1 | arm: | subjid: | param | visitvisit 2 | aval:164 | chg:19 | 60 | 6 | 3 |
| 171 | 3 | 309 | 3 | 1 | arm: | subjid: | param | visitvisit 3 | aval:148 | chg:3 | 60 | 6 | 3 |
| 172 | 3 | 309 | 1 | 2 | arm: | subjid: | paramdiastolic blood pressure (mmHg) | visitvisit 1 | aval:106~{super *} | chg: | 60 | 6 | 3 |
| 173 | 3 | 309 | 2 | 2 | arm: | subjid: | param | visitvisit 2 | aval:99 | chg:-7 | 60 | 6 | 3 |
| 174 | 3 | 309 | 3 | 2 | arm: | subjid: | param | visitvisit 3 | aval:98 | chg:-8 | 60 | 6 | 3 |
| 175 | 3 | 310 | 1 | 1 | arm: | subjid:310 | paramsystolic blood pressure (mmHg) | visitvisit 1 | aval:139~{super *} | chg: | 60 | 6 | 3 |
| 176 | 3 | 310 | 2 | 1 | arm: | subjid: | param | visitvisit 2 | aval:171 | chg:32 | 60 | 6 | 3 |
| 177 | 3 | 310 | 3 | 1 | arm: | subjid: | param | visitvisit 3 | aval:137 | chg:-2 | 60 | 6 | 3 |
| 178 | 3 | 310 | 1 | 2 | arm: | subjid: | paramdiastolic blood pressure (mmHg) | visitvisit 1 | aval:111~{super *} | chg: | 60 | 6 | 3 |
| 179 | 3 | 310 | 2 | 2 | arm: | subjid: | param | visitvisit 2 | aval:90 | chg:-21 | 60 | 6 | 3 |
| 180 | 3 | 310 | 3 | 2 | arm: | subjid: | param | visitvisit 3 | aval:92 | chg:-19 | 60 | 6 | 3 |

上記に、出力したいsasデータセットtable_3を示した。これを元に出力していく。

```
/*program07*/
ods escapechar="~";
ods html5 file="&output./table_3.html" style=onlyfontstyle;
data _null_;
  set table_3 end=end;
  by armn subjid paramn visitnum;
  if _n_=1 then do;
    dcl odsout r();
    r.table_start();
    r.head_start();
      r.row_start();
        r.format_cell(data:"table_3 一覧表", column_span:6, just:"l");
      r.row_end();
      r.row_start();
        r.format_cell(data:"full analysis set", column_span:6, just:"l",
                      style_attr:"&bb.");
      r.row_end();
      r.row_start();
        r.format_cell(data:"treatment", style_attr:"&bb.");
        r.format_cell(data:"subject", style_attr:"&bb.");
        r.format_cell(data:"parameter", style_attr:"&bb.");
        r.format_cell(data:"timing", style_attr:"&bb.");
        r.format_cell(data:"value", style_attr:"&bb.", just:"r");
        r.format_cell(data:"change|from|baseline",
                      style_attr:"&bb.",
                      split:"|", just:"r");
      r.row_end();
    r.head_end();
    r.body_start();
  end;
    r.row_start();
    array outs{*} out1-out6;
    array spans{*} span1-span3;
    do i=1 to dim(outs);
      if last. paramn then classstr=cats("class='",scan(outs[i],1,":"),"'")||" &bb.";
                      else classstr=cats("class='",scan(outs[i],1,":"),"'");
      if i in(5,6) then juststr="r";
      else if i=3   then juststr="l";
      else juststr="";
      if i in(1,2,3) then do;
        if scan(outs[i],2,":")^="" then
        r.format_cell(data:strip(scan(outs[i],2,":")),
                      row_span:spans[i],
                      style_attr:strip(classstr)||" &bb.");
      end;
      else do;
        r.format_cell(data:strip(scan(outs[i],2,":")),
                      style_attr:strip(classstr),
                      just:juststr);
      end;
    end;
    r.row_end();
  if end then do;
    r.body_end();
    r.foot_start();
      r.row_start();
      r.format_cell(data:"*:baseline value.", column_span:6, just:"l");
      r.row_end();
    r.foot_end();
    r.table_end();
  end;
run;
ods html5 close;
```

赤色太字で示したformat_cell.splitオプションにて、指定した文字を改行として扱うことが出来きる。HTMLファイルとしては<br/>タグとして出力される。

青色太字で示したformat_cell.style_attr オプションは、画面上に何か出力する設定ではない。「class」StyleAttributeを設定すると、HTMLファイルとして、該当セル(又は行、bodyそのもの)へメタデータを埋め込むことが可能である。本稿では出力に用いた元の変数名を格納している。

緑色太字で示したfoot_startメソッドは必須ではないが、HTMLに出力すると、<tfoot>タグになり、テーブルの構造をより明確に示すことが出来るため、使用することを勧める。

out5変数内に「~{super *}」と格納された箇所では、エスケープ文字と「super」インラインフォート関数を用いて「*」を上付きにして出力している。HTMLファイルとしては<sup>*</sup>となる。

program07で出力されるHTMLファイルを以下に示す。

3.　HTML であることを利用した出力

ここからはRWIを使ったノウハウではないが、HTMLファイルであることを利用して、より柔軟なレポート出力を考える。

これまで出力したHTMLファイル、table_1.html、table_2.html、table_3.htmlでは、それぞれ<table>～</table>の内容以外は共通であり、従って、<table>～</table>までの内容をまとめて、新たにHTMLファイルへ出力すことにより、一つのHTMLファイルへ統合することが可能である。

また、まとめる際にリンクを埋め込めばクリックすれば目的のレポートを参照できるし、インデックスのようなテーブルと実際のレポートとをブラウザ上で2画面表示させればインデックスとレポートを分割して表示も出来る。

以下に、これまで出力したtable_1～3.htmlをまとめ、さらに左側にインデックスを、右側にレポートを表示するようしたHTMLファイルを示した。これらは全てsas上でハンドリング可能であるため、HTMLファイルを直接手作業で修正する必要は無い。

## 4. まとめ

RWIと既存のプロシジャのメリット・デメリットについて、以下に一覧にまとめた。

| | RWI | 既存のプロシジャ |
|---|---|---|
| 出力の柔軟性 | ◎<br>dataステップを利用するため、基本的にほとんどの形のテーブルを網羅でき、sas関数を利用しつつ出力が可能。 | △<br>プロシジャの機能に制限されるため、限界がある。 |
| プログラムの可読性 | △<br>sasオブジェクト形式の記載となるため、慣れが必要かもしれない。また、目的の出力とするためにコーディング量が多くなる傾向がある。しかし、dataステップを利用するのでアルゴリズムは理解し易い。 | ○<br>プロシジャの記載ルールに則ればよいので、基本的には容易。 |
| RTFファイルへの出力 | ×<br>出力は出来るが、サポートされていない。使えないメソッドも存在する。 | ○<br>サポートされているため出力可能。 |

HTML形式ではあるがこれまでの出力結果から、RWIを利用した出力と既存のプロシジャを利用した出力とほとんど変わらない出力が可能であることを示せた。

プロシジャでは難しかったデータの加工をしつつ出力をするという点は非常に意義のあることと考える。

RWIはRTFファイルへの出力がサポートされていないため、実務上ではすぐに活躍することは無いのかもしれない。今後、RTFファイルへの出力がサポートされれば既存のプロシジャでの出力に取って代わるかもしれない。

## 5. 参考文献

・SAS® 9.4 Output Delivery SystemAdvanced Topics Second Edition

# SASと外部ソフトウェアの連携：Windows GUI構築とExcelデータの利用

○高田一行[1]、迫田英之[1]、阿部一也[1]、村松徳之[1]、宮原秀夫[1]、木村範昭[2]
（[1]株式会社タクミインフォメーションテクノロジー、[2]専修大学商学部）

情報通信技術やコンピュータ環境の進歩や革新により、業種を問わず人が行う業務も多様化している。人が行う日々の業務の効率化や標準化を図り、作業量を軽減するためには、利用しているソフトウェアやコンピュータ環境を有効利用する方法を考え、ユーザーインターフェースの工夫や作業手順の合理化など、様々な課題を解決する試みを続けなければならない。さらに課題は環境だけに留まらず分析対象のデータについても容量増大、品質保証、特別な形式のデータへの対応など、多くの課題に向き合う必要がある。SASは様々なコンピュータ環境やデータ形式、外部ソフトウェアとの連携など柔軟かつ豊富な機能を提供している。ここでは、業務を効率化するためのGUIメニューの構築、外部ソフトウェアとのタスクやデータの連携、様々な形式やレイアウトのデータ利用などの課題を、SASを利用して解決するための事例を紹介する。

複雑なレイアウトのExcel帳票の形式で保存されているデータ資産をデータベース化して、分析に利用したいというニーズは多い。しかし明確な行と列の関係を持つテーブル形式のデータの扱いは比較的容易だが、複雑に設計されたレイアウトのExcel帳票からデータを取り出すのは難しく、データの利用を諦めてしまうケースが散見される。

また様々な業務や機能に特化したソフトウェアを利用することは、専門性の高い業務を行うためには必要不可欠となっているため、目的の業務を実施する際に、いくつかの外部ソフトウェアのタスクを途中に入れ込みたい、または平行して外部ソフトウェアのタスクを実行したいというニーズも多い。

さらに使いやすいユーザーインターフェースを利用して業務の効率化を図ることも重要である。作業環境は特定できないため、ユーザーインターフェースはいくつかの環境での利用を想定する必要があり、簡単に短期間で構築することも重要な条件となる。

SAS Integration Technologiesは、これらの課題を解決するための機能を提供している。①SASと外部ソフトウェアとのシームレスな連携、②他言語で構築したイベントドリブンの使いやすいWindows GUIメニューとSASのタスクとの連携、さらにWindows GUIメニューはクライアントPC、サーバー環境、クラウド環境など環境を選ばず利用が可能である。

データの連携に関してもSASは多くのLIBNAMEエンジンをサポートしているため、様々なデータ形式のファイルを利用することができる。LIBNAMEエンジンを指定したデータファイルは、SASデータセットへの変換をすることなく、そのままSASの分析機能で利用することができる。

SAS、Windows GUIメニュー、外部ソフトウェアを連携するSAS Integration Technologiesを利用することにより、課題解決の道筋を見出すことができる。様々な環境やソフトウェアが持つ特徴を、最大限に引き出すことができ、利用を諦めていたデータ資産の活用が促進される。

# 初/中級者が陥りやすいプログラミングエラー

井上 貴博
（ノバルティス ファーマ株式会社）

Programming errors that easily occur on beginner/intermediate SAS programmers

Takahiro Inoue
Data Sciences & Scientific Operations Department, Novartis Pharma K.K.

## 要旨

初/中級者には理解し難く陥りやすいプログラミングエラーは多数ある。エラー発見・解消テクニックを紹介し、浮動小数点の丸め誤差、PDV など、理解していないと意図した結果が得られない事例を紹介する。

キーワード：プログラミング、テクニック、エラー、浮動小数点、丸め誤差、PDV、Program Data Vector

## 1. はじめに

SAS のパワーユーザーには当たり前になっているプログラミングエラーでも、SAS を構文を覚え慣れ始めた初/中級者に理解し難く、陥りやすいエラーは多数ある。しかし、そのテクニックや内部処理が体系的にまとめられた資料は少ない。そこで、本稿ではエラー発見・解消に役立つテクニックを紹介し、その後、浮動小数点の丸め誤差、Program Data Vector など、内部処理を理解していないと意図した結果が得られない事例を紹介する。なお、本稿は、SAS OnDemand for Academics を使用した。処理結果を再現できるようにプログラムも記載しているが、他の動作環境を使用する場合、結果が異なる場合がある。

## 2. ログの確認

SAS は、プログラムを実行するとログを生成する。このログには以下の情報が含まれている。

- 実行されたプログラム
- SAS データセットに関する情報(変数やオブザベーション数など)
- プログラム実行中に発生した警告・エラーメッセージなど
- プログラム実行に要した時間

プログラムが正常に実行されてない場合、ログに WARNING や ERROR が表示される。その原因やプログラム修正のためのヒントは、ログから見つけることができ、プログラムの実行後は、意図した通りに実行され結果が得られているかログを確認する必要がある。
詳細は後項に記載するが、自動マクロ変数で ERROR、WARNING の有無を確認できる。しかし、ERROR、WARNING ではない NOTE でも注意すべきメッセージが多数ある。ログに確認する際のキーワードの一例を以下に示す。

| | | | | |
|---|---|---|---|---|
| ERROR | WARNING | BY 値を繰り返すデータセット<br>(バージョンにより「BY 値」とスペースがないものもある) | | |
| 欠損値を含んだ | 初期化され | 変換 | 範囲外 | による割り算がありました |
| ゼロ割り | 無効な | LOST CARD | 処理を中止 | 文字を超えています |
| 実行を中止 | 切り捨て | 表示が制限 | ステートメントの余分な情報を無視します | |
| 既にライブラリ | 出力形式によって小数点がシフト | | | 変数がありません |
| 存在しません | 演算式を計算できなかった | | | |

## 3. エラーのタイプ

エラーが発生したプログラムを修正するには、その発生したエラーのタイプを知ることで、そのエラーのタイプに合わせたプログラム修正ができる。SAS 9.4 Language Reference [1] の Types of Errors in SAS の項では、以下のエラーのタイプが紹介されている。

- 構文エラー(Syntax errors)
- セマンティックエラー(Semantic errors)
- 実行時エラー(Execution-time errors)
- データエラー(Data errors)
- マクロ関連エラー(Macro-related Errors)※

  ※マクロコンパイル時および実行時に発生する。または、マクロから生成されたステートメント構文等のエラーだが、本稿での解説は割愛する。

また、Language Reference では紹介されていないが、論理エラー (Logic errors) と呼ばれるエラーがあり、以下にそれぞれの具体例を示す。

### 3.1. 構文エラー(Syntax errors)

SAS 構文エラーに沿っていない場合に発生するエラーで、以下のようなエラーがある。

- スペルミス
- 引用符(“ ”,’ ’)の不一致
- セミコロン(;)がない
- 無効なステートメントオプション
- 無効なプロシジャオプション

| サンプルプログラム | ログ |
|---|---|
| date temp ;<br>set a ;<br>run ; | 60 date temp ;<br>14<br>WARNING 14-169: dateはシンボルDATAのスペルミスであると判断しました。<br><br>61 set a ;<br>62 run ;<br><br>NOTE: データセットWORK.Aから1オブザベーションを読み込みました。<br>NOTE: データセットWORK.TEMPは1オブザベーション、1変数です。<br>NOTE: DATAステートメント処理(合計処理時間): |
| data を date とスペルミス。SAS では一部のスペルミスは自動的に変更し実行する。 | |
| data temp<br>set a ;<br>run ; | 60 data temp<br>61 set a ;<br>56<br>ERROR 56-185: オプションDATASTMTCHK=COREKEYWORDSの時、DATAステートメントにSETは使用できません。<br>DATAステートメントにセミコロンが抜けていないか、<br>またはDATASTMTCHK=NONEを使用してください。<br><br>62 run ;<br><br>NOTE: エラーが発生したため、このステップの処理を中止しました。<br>NOTE: DATAステートメント処理(合計処理時間): |
| data temp ;のセミコロン(;)が抜けていた。SAS のバージョンによっては、temp, set, a の 3 データセットが作成されることがあり、SET 文の参照データ a が上書きされる場合もある。 | |

### 3.2. セマンティックエラー(Semantic errors)

構文規則上は正しくても、使用意図として間違っているような場合に発生する。

| サンプルプログラム | ログ |
|---|---|
| data temp ;<br>array all(*) x1 - x5 ;<br>all = 3 ;<br>run ; | 61 data temp ;<br>62 array all(*) x1 - x5 ;<br>63 all = 3 ;<br>ERROR: 配列allへの参照が正しくありません。<br>64 run ;<br><br>NOTE: エラーが発生したため、このステップの処理を中止しました。<br>WARNING: データセットWORK.TEMPは未完成です。このステップは、0オブザベーション、<br>5変数で停止しました。<br>WARNING: このステップを中止したため、データセットWORK.TEMPを置き換えていません。<br>NOTE: DATAステートメント処理(合計処理時間): |
| array all(*) x1 - x5 ;および all = 3 ;を各行で見れば、構文規則上、問題はない。<br>しかし、all は配列として定義されているため、変数として all は使用できない。 | |

### 3.3. 実行時エラー(Execution-time errors)

コンパイル済みのプログラムを実行する際に発生する以下のようなエラーがある。

- 関数の無効な引数
- 無効な数学演算(0 除算)
- BY 変数が Sort されていない
- ARRAY 配列の引数が範囲外
- INFINE/FILE ステートメントで指定したファイルのオープンおよびクローズエラー
- INPUT 文での無効な引数(文字変数を数値変数として抽出)

### 3.4. データエラー(Data errors)

ステートメントは正しいが、データ値がステートメントに適さない場合に発生する。

| サンプルプログラム | ログ |
|---|---|
| data temp ;<br>dtc = '2017-02-31' ;<br>dt = input(dtc, yymmdd10.) ;<br>run ; | 61 data temp ;<br>62 dtc = '2017-02-31' ;<br>63 dt = input(dtc, yymmdd10.) ;<br>64 run ;<br><br>NOTE: 関数INPUT(行 63 カラム 7)の引数は無効です。<br>dtc=2017-02-31 dt=. _ERROR_=1 _N_=1<br>NOTE: 以下の箇所で演算式を計算できなかったため、結果を欠損値に設定しました。<br>(回数)(行:カラム)<br>1 63:7<br>NOTE: データセットWORK.TEMPは1オブザベーション、2変数です。 |
| 構文に問題はない。しかし、2017-02-31 は日付として存在しないため、引数が無効となり Null 値が設定される。ログには ERROR, WARNING ではなく NOTE として表示されないため、注意が必要である。 | |

なお、2050-02-01 のように日付として存在するが、未来の日付が入力された場合、SAS のログには表示されない。そのため、必要に応じて、入力データが適切であるかデータレビュー等を行う。

### 3.5. 論理エラー(Logic errors)

ここまで紹介したエラー以外に SAS のログからは発見できない論理エラーがある。例えば、以下のように年齢(age)を'10 代'、'20 代'、'30 代以上'と分類するプログラムを作成したとする。プログラム(誤)では比較演算子に「=」がなく、age = 10, 20, 30 は、agec が Null 値となってしまう。

| プログラム(誤) | プログラム(正) |
|---|---|
| if 10 < age < 20 then agec = '10 代' ;<br>else if 20 < age < 30 then agec = '20 代' ;<br>else if 30 < age then agec = '30 代以上' ; | if 10 <= age < 20 then agec = '10 代' ;<br>else if 20 <= age < 30 then agec = '20 代' ;<br>else if 30 <= age then agec = '30 代以上' ; |

また、数値変数の Null 値は最小値として扱われるため、IF 文の比較演算には注意が必要である。

| サンプルプログラム | ログ |
|---|---|
| data temp ;<br>input age ;<br>if age <= 1 then agec = 1 ;<br>else if 2 < age then agec = 2 ;<br>put age = agec = ;<br>cards;<br>.<br>1<br>2<br>;<br>run ; | 60 data temp ;<br>61 input age ;<br>62 if age <= 1 then agec = 1 ;<br>63 else if 2 < age then agec = 2 ;<br>64 put age = agec = ;<br>65 cards;<br><br>age=. agec=1<br>age=1 agec=1<br>age=2 agec=.<br>NOTE: データセットWORK.TEMPは3オブザベーション、2変数です。<br>NOTE: DATAステートメント処理(合計処理時間):<br>処理時間 0.00 秒<br>ユーザーCPU時間 0.00 秒 |
| age=. を agec=. とする場合 if age <= 1 then agec = 1 ;を if . < age <= 1 then agec = 1 ;とする。 | |

論理エラーとして紹介した 2 つの例では、ログに ERROR, WARNING は表示されず、通常の NOTE のみである。そのため、作成したプログラムおよびデータが意図した結果になっているか確認の必要がある。

## 4. エラー発見・解消のためのテクニック

SAS では、次のコンポーネントをデバッグに利用することができる。これを組み合わせることにより、検証時のデータ損失の予防や、エラー発生箇所の特定を容易にすることができる。

- オプション(システム/データセット)
- ステートメント
- 関数および CALL ルーチン
- 自動変数

以下にそれぞれの具体例を示す。

### 4.1. オプション(システム/データセット)を利用する

obs = 0 と noreplace オプションを同時に利用することで、データセットや外部データを読み込まず Data step を実行し、構文エラーの確認ができ、存在するデータセットへの上書きを防げる。設定をデフォルトに戻すには「options obs = max replace ;」と記述する。

サンプルログ(およびプログラム)

```
60          option obs = 0 noreplace ;
61          data temp.b ;
62              set a ;
63          run ;

NOTE: データセットWORK.Aから0オブザベーションを読み込みました。
NOTE: データセットTEMP.Bは0オブザベーション、1変数です。
WARNING: NOREPLACE オプションのため、データセットTEMP.Bを置き換えていません。
NOTE: DATAステートメント処理(合計処理時間):
```

大量の処理する場合、必然的に実行時間が長くなる。そのため、FISRTOBS=および OBS=データセットオプションを利用しデータセットをサブセット化することで、実行時間を短くしコードレビューを容易にすることができる。また、WHERE=データセットオプションを利用することにより、サブグループごとにサブセット化することも可能である。

サンプルログ(およびプログラム)

```
61          data b ;
62              set a(firstobs = 10 obs = 15) ;
63          run ;

NOTE: データセットWORK.Aから6オブザベーションを読み込みました。
NOTE: データセットWORK.Bは6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):

64          data b ;
65              set a(where = (10 < age <20)) ;
66          run ;

NOTE: データセットWORK.Aから9オブザベーションを読み込みました。
      WHERE (age>10 and age<20);
NOTE: データセットWORK.Bは9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):
```

### 4.2. ステートメントを利用する

前項の方法ではデータに偏りが発生する場合がある。そのため、ステートメントを利用したサブセット化も有用である。

サンプルプログラム

```
data temp ;
    do pickup=1 to totobs by 100;
        set a point=pickup nobs=totobs;
        output;
    end;
    stop;
run;
```

また、プログラムが正しくても、不正なデータにより処理に悪影響を受けることがある。このようなエラーを防ぐために Defensive coding が必要となる。例えば、以下のプログラムは、未来日が処理に悪影響があるとして ERROR を表示している。

サンプルプログラム

```
data temp ;
    length dtc $10. ; input dtc $ ;
    dt = input(dtc, yymmdd10.) ;
    if today() < dt then
      put 'ER' 'ROR: 未来日あり' dt = yymmdd10.;
cards;
2017-02-28
2050-02-01
;
run ;
```

ログ

```
60          data temp ;
61              length dtc $10. ; input dtc $ ;
62              dt = input(dtc, yymmdd10.) ;
63              if today() < dt then
64                put 'ER' 'ROR: 未来日あり' dt = yymmd
65          cards;

ERROR: 未来日ありdt=2050-02-01
NOTE: データセットWORK.TEMPは2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):
```

### 4.3. 関数を利用する

関数を利用して、データセットの存在を確認することができる。以下のプログラムは、データの存在を確認し ERROR が表示している。

サンプルログ(およびプログラム)

```
61          data _null_ ;
62              zaibuzai = exist('work.no') ;
63              if zaibuzai ne 1 then
64                put 'ER' 'ROR: データセットがありません' ;
65          run ;

ERROR: データセットがありません
NOTE: DATAステートメント処理(合計処理時間):
```

また、システムオプションの設定値によっては、プログラムが意図した結果にならない。この場合、GETOPTION 関数を利用して設定値を取得する。

サンプルログ(およびプログラム)

```
63          %LET p_size=%SYSFUNC(GETOPTION(PS));
64          %PUT 現在の PAGESIZE:  &p_size;
現在の PAGESIZE:  75
65          /* PAGESIZE が 100 未満なら 100 へ変更 */
66          %MACRO ps;
67             %IF &p_size < 100 %THEN %DO;
68
69                OPTIONS ps=100;
70             %END;
71          %MEND;
72          %ps;
73          %LET p_size=%SYSFUNC(GETOPTION(PS));
74          %PUT 現在の PAGESIZE:  &p_size;
現在の PAGESIZE:  100
```

### 4.4. 自動変数を利用する

SAS データセットでは、いくつかの自動変数が用意されている。以下は、文字変数が不正日付で数値に変換できなかった場合、_error_を使用して_n_(データ行)を表示するプログラムである。

サンプルプログラム

```
data date ;
    length dtc $10. ;
    input dtc $ ;
cards;
2017-02-01
2017-02-30
;
run ;

data temp ;
    set date ;
    dt = input(dtc, yymmdd10.) ;
    if _error_ = 1 then
      put 'ER' 'ROR: '
          _n_ '行目に不正日付あり' dtc = ;
run ;
```

ログ

```
 ～ 省略 ～

68          data temp ;
69              set date ;
70              dt = input(dtc, yymmdd10.) ;
71              if _error_ = 1 then
72                put 'ER' 'ROR: '
73                     _n_ '行目に不正日付あり' dtc = ;
74          run ;

NOTE: 関数INPUT(行 70 カラム 10)の引数は無効です。
ERROR: 2 行目に不正日付ありdtc=2017-02-30
dtc=2017-02-30 dt=. _ERROR_=1 _N_=2
NOTE: 以下の箇所で演算式を計算できなかったため、結果を欠損値に
```

引数の前に「?」または「??」をつけることにより、_error_およびログの出力等を制御することができ、複数の INPUT 文を使用する際などに利用し特定の変数のみを確認することができる。

サンプルログ(およびプログラム)

```
61          data temp ;
62              set date ;
63              dt = input(dtc,? yymmdd10.) ;
64              if _error_ = 1 then
65                put 'ER' 'ROR: ' _n_ '行目に不正日付あり' dtc = ;
66          run ;

ERROR: 2 行目に不正日付ありdtc=2017-02-30
dtc=2017-02-30 dt=. _ERROR_=1 _N_=2
NOTE: データセットWORK.DATEから2オブザベーションを読み込みました。
NOTE: データセットWORK.TEMPは2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):
[illegible]

68          data temp ;
69              set date ;
70              dt = input(dtc,?? yymmdd10.) ;
71              if _error_ = 1 then
72                put 'ER' 'ROR: ' _n_ '行目に不正日付あり' dtc = ;
73          run ;

NOTE: データセットWORK.DATEから2オブザベーションを読み込みました。
NOTE: データセットWORK.TEMPは2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):
[illegible]
```

ログに出力される ERROR、WARNING メッセージを格納する自動マクロ変数が用意されている。

サンプルログ(およびプログラム)

```
60          dat temp
            ___
            14
61              set date ;
                ___
                56
WARNING 14 -169: datはシンボルDATAのスペルミスであると判断しました。

ERROR 56-18 5: オプションDATASTMTCHK=COREKEYWORDSの時、DATAステートメントにSETは使用できません。
               DATAステートメントにセミコロンが抜けていないか、
               またはDATASTMTCHK=NONEを使用してください。

62          run;

NOTE: エラーが発生したため、このステップの処理を中止しました。
NOTE: DATAステートメント処理(合計処理時間):
      [illegible]  00 秒

63          %PUT &SYSWARNINGTEXT;
14-169: datはシンボルDATAのスペルミスであると判断しました。
64          %PUT &SYSERRORTEXT;
56-18 5: オプションDATASTMTCHK=COREKEYWORDSの時、DATAステートメントにSETは使用できません。
DATAステートメントにセミコロンが抜けていないか、          またはDATASTMTCHK=NONEを使用してください。
```

これにより、ログに ERROR、WARNING が存在するか確認することができる。

## 5. 浮動小数点の丸め誤差

### 5.1. 2 進算術演算と 10 進算術演算の差異

コンピュータは、有限精度の 2 進算術演算を使用している。通常、厳密な 2 進表現がない数値を扱う場合、コンピュータによって生成される結果は 10 進算術演算で生成される結果とわずかに異なる。例えば、10 進値の 0.1 および 0.3 には厳密な 2 進表現はない。10 進算術演算の 3*0.1 は厳密に 0.3 になるが、2 進算術演算ではこの等式は成り立たない。次の例に示すように、これらの 2 つの値を SAS に書き込むと、同じ値のように見える。ただし、差異を計算すると、値が異なることが分かる。

サンプルログ(およびプログラム)

```
61          data _null_;
62             point_three=0 .3;
63             three_times_point_one=3*0.1;
64             difference=point_three - three_times_point_one;
65             put point_three= ;
66             put three_times_point_one= ;
67             put difference= ;
68          run;

point_three=0 .3
three_times_point_one=0 .3
difference=-5 .55112E-17
NOTE: DATAステートメント処理(合計処理時間):
      処理時間              0.00 秒
      [illegible]  00 秒
```

SAS では、IEEE 準拠の浮動小数点表現 (倍精度浮動小数点表現) を用いて数値を取り扱っている。この浮動小数点表現は、効率よく実数を表現でき、また表現可能なレンジが広いという利点があるが、10 進数を 2 進数で格納するため、10 進数と 2 進数のマッピングに誤差が生じる場合がある。とくに 10 進数では有限小数点であっても、2 進数では無限小数点(循環小数)となる場合があり、コンピュータ内部では、物理的な格納つまり有限桁に変換され、精度が失われる。

### 5.2. 丸め誤差の及ぼす影響

前項では、3*0.1 が厳密に 0. 3 にならない例を示した。このような差異がある場合、IF 文で意図した結果にならない。例えば、以下のプログラムのように 0. 3 超のとき fr1 に1 を、0. 3 以下のとき fr2 に 1 を設定する IF 文を作成した場合、意図した結果が得られないことがある。

サンプルログ(およびプログラム)

```
62          data temp ;
63          a = 3 * 0.1 ;
64          if 0.3 < a then fr1 = 1 ;
65          if a <= 0.3 then fr2 = 1 ;
66          put a = fr1 = fr2 = ;
67          run ;

a=0.3 fr1=1 fr2=.
NOTE: データセットWORK.TEMPは1オブザベーション、3変数です。
NOTE: [illegible]ステートメント処理(合計処理時間): [illegible]
```

### 5.3. 回避策

SAS は、プログラムで"0. 3'を定数として示した場合、値は標準の入力形式で 3/10 として計算される。変数は前項で示した丸め誤差が生じるため、ROUND 関数を使用し丸め処理を行う必要がある。

サンプルログ(およびプログラム)

```
61          data temp ;
62          a = 3 * 0.1 ;
63          if 0.3 < round(a, 1e-8) then fr1 = 1 ;
64          if round(a, 1e-8) <= 0.3 then fr2 = 1 ;
65          put a = fr1 = fr2 = ;
66          run ;

a=0.3 fr1=. fr2=1
NOTE: データセットWORK.TEMPは1オブザベーション、3変数です。
[illegible]
```

本稿では 3*0.1 の算術式を例としたが、SAS が別の入力形式を使用して変数を読み込む場合や、SAS 以外のプログラムで変数を読み込む場合も、10 進値の厳密な値が生成されない可能性がある。また、厳密な 2 進表現ではない数値が計算に含まれている場合や、浮動小数点表現が異なる別の動作環境にデータセットを移植する場合も、正確な結果を得られない可能性がある。

### 5.4. ROUND 関数の注意点

通常、ROUND 関数は、結果の有効桁数が 9 以下で、次のどの条件にも該当しない場合、10 進算術演算で期待される結果を生成する。

- 丸めの単位が整数である。
- 丸め単位が 1e-15 以上の 10 のべき乗である。
  (丸め単位が 1 よりも小さい場合、ROUND は、丸め単位の逆数が 10 のべき乗との違いが最下位から第 3 または第 4 ビッドまでであれば、丸め単位を 10 のべき乗として扱う)
- 10 進算術演算で期待される結果が小数第 4 位以下である。

多くのパッケージソフトウェアや表計算ソフトウェアでは、内部的に暗黙の FUZZ 値を用いてこの問題を回避しているが、以前の SAS の ROUND 関数では引数の値をそのまま用いて処理を行うため、意図した結果が得ないことが多かった。このような問題は、FUZZ 関数を利用した構文を作成することにより、回避できた。[例 round(a, 1e-5)： round((a+(sign(a)*1e-10)),1e-5)]

## 6. Program Data Vector (PDV)

SAS は、Program Data Vector (以下 PDV) に一度新しいデータセットを作成する。どのように PDV で初期設定され、情報が保持されているかを理解することが正確なプログラムを書く上で必要不可欠となる。本項では、事例を使って PDV でどのような処理されているか紹介する。

### 6.1. 使用するデータセット

以下のデータを使用し、PDV が影響する事例を示す。

| Dataset: A | |
|---|---|
| id | var1 |
| a1 | 1 |
| a2 | 2 |

```
/* プログラム */
data A ;
    input id $ var1 ;
cards;
a1 1
a2 2
;
run ;
```

| Dataset: B | | |
|---|---|---|
| id | var1 | var2 |
| b1 | 1 | 101 |
| b2 | 2 | 202 |

```
/* プログラム */
data B ;
    input id $ var1 var2 ;
cards;
b1 1 101
b2 2 202
;
run ;
```

| Dataset: C | | |
|---|---|---|
| id | var1 | var2 |
| a1 | 1 | . |
| a2 | 2 | . |
| b1 | 1 | 101 |
| b2 | 2 | 202 |

```
/* プログラム */
data C ;
    set A B ;
run ;
```

### 6.2. PDV が影響する事例

Dataset: C に対して、if var1 = 1 then var2 = 0 ; という IF 文を使用し、Dataset: D を作成した例を示す。この例は意図通りの結果を得られている。

| Dataset: D | | | |
|---|---|---|---|
| id | var1 | var2 | var3 |
| a1 | 1 | 0 | 0 |
| a2 | 2 | . | . |
| b1 | 1 | 0 | 0 |
| b2 | 2 | 202 | . |

```
/* プログラム */
data D ;
    set C ;
    if var1 = 1 then var2 = 0 ;
    if var1 = 1 then var3 = 0 ;
run ;
```

次に Dataset: C を使わず、Dataset: A と B を SET 文で読み込み、同様の IF 文を使用した Dataset: E を作成した例を示す。

| Dataset: E | | | |
|---|---|---|---|
| id | var1 | var2 | var3 |
| a1 | 1 | 0 | 0 |
| a2 | 2 | 0 | . |
| b1 | 1 | 0 | 0 |
| b2 | 2 | 202 | . |

```
/* プログラム */
data E ;
    set A B ;
    if var1 = 1 then var2 = 0 ;
    if var1 = 1 then var3 = 0 ;
run ;
```

Dataset: E では、id = a2 の var2 が 0 になっていることに注目していただきたい。PDV の処理を理解されていない方にとっては、意図していない結果になっていると思う。

### 6.3. 処理プロセス・結果

PDV では、1 オブザベーション毎にデータが保持され処理される。次のオブザベーションの処理に移った際、新規変数はリセットされ(Null 値に変換)され、読み込み変数は読み込みデータで上書きされる。

前項で示した Dataset: E の場合、var2 は読み込み変数、var3 は新規変数である。新規変数である var3 は、オブザベーションが変わるたびに Null 値にリセットされる。その上で、IF 文の条件に合致した id=a1, a2 は var3 に 0 が代入され、条件が合致しない id=a2, b2 は Null 値が残る。

var2 は読み込み変数だが、Dataset: A に var2 がない。そのため、Dataset: B 由来の id = b1, b2 では変数が読み込みデータで上書きされるが、Dataset: A 由来の id = a1, a2 は読み込みされず上書きされない。そのため、id = a2 では上書きされず、前オブザベーションのデータが残り、0 となった。

### 6.4. PDV が影響した MERGE 文の事例

前項では SET 文の事例を示したが MERGE 文でも同様に PDV が影響し、意図しない結果が出る場合がある。

| Dataset: F | |
|---|---|
| id | var1 |
| 1 | XXXX |
| 1 | XXXX |
| 2 | YYYY |
| 2 | YYYY |

```
/* プログラム */
data F ;
    input id var1 $ ;
cards ;
1 XXXX
1 XXXX
2 YYYY
2 YYYY
;
run ;
```

| Dataset: G | | |
|---|---|---|
| id | var1 | var2 |
| 1 | ZZZZ | 7777 |
| 2 | ZZZZ | 9999 |

```
/* プログラム */
data G ;
    input id var1
$ var2 ;
cards ;
1 ZZZZ 7777
2 ZZZZ 9999
;
run ;
```

| Dataset: H | | |
|---|---|---|
| id | var1 | var2 |
| 1 | ZZZZ | 7777 |
| 1 | XXXX | 7777 |
| 2 | ZZZZ | 9999 |
| 2 | YYYY | 9999 |

```
/* プログラム */
data H ;
    merge F G ;
    by id ;
run ;
```

var2 は Dataset: F になく Dataset: G にのみ存在するため、全てのオブザベーションで読み込みが行われる。Dataset: F, G の両方に存在する var1 は、Datasets: F のデータが読み込まれた後、Dataset: G のデータで上書きされる。その際、Dataset: G のオブザベーションのみ上書きされるため、var1 は、各 ID の 1 行目が Dataset: G、2 行目が Dataset: F のデータとなる。

### 6.5. PDV に対する対策

通常、前項の事例で示したような IF 文で存在する変数の上書き、変数が重複する状態での MERGE は行わないほうが適切である。今回は、PDV の処理プロセスの説明のため、このようなプログラムを事例として紹介した。

もし、Dataset: E で示したように SET 文と IF 文を同一 Data step 内で処理する場合、以下のプログラムで Dataset: D と同様の結果を得られることができる。

```
/* プログラム */
data E_ ;
    set A(in = A) B ;
    if var1 = 1 then var2 = 0 ;
      else if A then var2 = . ;  /* var2 がない dataset:A は var2 を初期化 */
    if var1 = 1 then var3 = 0 ;
run ;
```

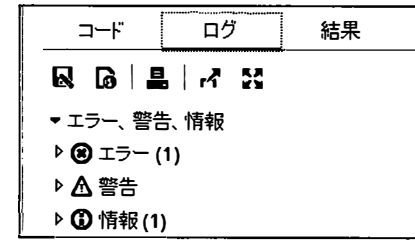
変数が重複する状態での MERGE に関しては、「options msglevel = i ;」を指定することにより、ログに重複した変数の情報を表示できる。設定をデフォルトに戻すには「options msglevel = N ;」と記述する。

サンプルログ(およびプログラム)

```
79          options msglevel=i;
80          data H ;
81          merge F G ;
82          by id ;
83          run ;

INFO:  変数 var1 (データセット WORK.F) はデータセットWORK.Gによって上書きされます。
NOTE: データセットWORK.Fから4オブザベーションを読み込みました。
NOTE: データセットWORK.Gから2オブザベーションを読み込みました。
[illegible]
```

## 7. おわりに

SAS のログでは、ERROR、WARNING が赤や緑などで目立つように表示される。また、SAS OnDemand for Academics などをはじめとする最新の SAS では、エラー、警告などの数をカテゴリで表示できるようになった。しかし、本稿で紹介した事例にようにログの NOTE でも注意すべきメッセージでも多くある。意図した結果を得るには、まずはログの確認が重要である。




また、どういったエラーがあるか内容を理解し、そのエラーに合わせたプログラミング、得られた結果が意図したものであるか、実際に入力されたデータ、出力されたデータを確認することが必要である。

また、SAS 初心者が SAS を学習する際、構文を学習することに重点を置くことが多い。しかし、本稿で紹介した事例のように、SAS がコンパイルと実行段階で、どのようにデータを処理するかを理解せずプログラミングした場合、意図した結果を得られないことがある。SAS の構文を覚えた初/中級者は、正しい結果を作成できるように本稿で紹介したような SAS 内部のデータ処理に対しても、理解を深めていただきたい。

## 8. 参考文献


[1] SAS 9.4 Language Reference: Concepts, Sixth Edition
< https://support.sas.com/documentation/cdl/en/lrcon/69852/HTML/default/viewer.htm#titlepage.htm>
(参照: 2017-6-19)

[2] SAS Technical News
< http://www.sas.com/offices/asiapacific/japan/periodicals/technews/>
(参照: 2017-6-19)

[3] SAS 9.4 関数と CALL ルーチン　リファレンス, 第 4 版
<http://support.sas.com/documentation/cdl_alternate/ja/lefunctionsref/67960/HTML/default/titlepage.htm>
(参照: 2017-6-19)

[4] SAS FAQ
<https://www.sas.com/ja_jp/support/technical/faq.html>
(参照: 2017-6-19)

[5] Essentials of the Program Data Vector (PDV): Directing the Aim to Understanding the DATA Step!
Arthur Xuejun Li, City of Hope National Medical Center, Duarte, CA
SAS Global Forum 2013, Paper 125-2013

[6] The Use and Abuse of the Program Data Vector
Jim Johnson, Ephicacy Corporation, North Wales, PA, USA
SAS Global Forum 2012, Paper 255-2012

# Let’s start STDRATE Procedure

○堀田　真一
（ファイザー株式会社）

SAS/STAT 12.1から利用可能になったSTDRATEプロシジャは、疫学で一般的に用いられる指標である率と割合を計算するツールである。STDRATEプロシジャは、率と割合の要因間の差や比、それらの信頼区間を計算することができる。また、層化や参照集団による標準化も可能である。STDRATEプロシジャの入力データセットの基本的な形式や、プログラム例、計算結果の見方を紹介する。さらに、使用上の注意点や、割合の計算についてFREQプロシジャとの比較も紹介する。

# Pinnacle 21 CommunityのADaMチェック機能を補完する、XML Mappingを使用したプログラムの紹介

○西岡　宏

（シミック株式会社　CDM事業本部　統計解析部）

要旨:
無償版Pinnacle 21では行われない、Define.xmlのValue-level MetadataのコードリストとADaMデータセットとの突合確認を行うSASプログラムを紹介する。

キーワード:
Define.xml, Value-level Metadata, XML Mapping, Pinnacle21

Pinnacle21は承認申請資料として提出されるADaMデータセットのCDISC準拠状況を検証するツールである。
これには有償版のPinnacle 21 Enterpriseと無償版のPinnacle 21 Communityがあり、機能に互換性があるとされてはいるものの、無償版(バージョン2.1.3および2.2)においては、Value-level MetadataにおけるコードリストとADaMデータセットのParameterごとの値との突合が行われない。
有償版の方は、コードリストとADaMデータセットに不整合があった場合、Rule ID: SD1228「Variable value not found in user-defined codelist when value-level condition occurs」のErrorを出力することができる。

そこで、以下のような流れでチェックするSASプログラムを作成した。

1. Define.xmlをSASデータセット化:
   SASのXMLをMappingする機能を用い、Value-level MetadataのあるADaMの変数情報および各変数に設定されているコードリスト情報をDefine.xmlから抽出しSASデータセットにする。

2. ADaMデータセットの変数内容を一覧化:
   各ADaMデータセットの変数のうち、Value-level Metadataを持つ変数の内容を一覧化する。

3. 突合:
   1.のSASデータセットと2.の一覧を突合し、Define.xmlに未定義の内容がADaMデータセットに含まれていないかチェックする。

承認申請時に申請者は有償版Pinnacleを用いたチェックを必須とされてはいないが、PMDAは有償版を使用して承認申請資料のチェックを行うため、無償版によるチェックのみでは、PMDAに承認申請資料を提出した後になってDefine.xmlとADaMデータセットの不整合が見つかるという状況が発生する可能性がある。
本発表で提示したような、無償版Pinnacleのチェック機能を補助する手段を用いることで、事前に対応が可能になると考えられる。

# LUA プロシジャを起点としたビッグデータ解析とチームコラボレーション
# ～LUA プロシジャと歩む，初めての共同作業～

○木口亮，北西由武

（塩野義製薬 解析センター）

# Big Data Analysis and Team Collaboration Starting from LUA Procedure
# ~ The First Team Programming Work with LUA Procedure ~

Ryo Kiguchi, Yoshitake Kitanishi
Shionogi & Co., Ltd

## 0. 要旨

Big Data という言葉がバズワードと呼ばれた時代も過ぎ，各企業，団体が本腰を入れて Big Data の利活用を推進している．Big Data 解析は，複数人で行われる場合が多く，関係者間で知識共有をすると同時に，ナレッジとして経験を蓄えていくことが重要である．これはプログラムにおいても同様である．Big Data 解析では，データクリーニング，欠測の扱いに関する検討 (Ex. 補完 / 削除) ，正規化，解析手法の選択，モデリング，解析，評価，考察，考察後の追加解析など，多くのステップを踏む．Big Data はデータ構造，種類などによって特徴が異なるため，解析のたびに各ステップの適切な手法等の選択はもちろん，ステップの組み合わせも検討する必要がある．そのため，過去に実施した解析プログラムをナレッジとして蓄え，アクセス可能な状態にすること，さらには現在進行している解析についてはそのプログラムを関係者間で可読性の高い状態で共有することで，解析フローを透明化し，集合知的な発想を促すことが重要とある．この一連の解析プロセスは，SAS ® 9.4 から正式に実装された Lua プロシジャを用いることで実現できる．Lua プロシジャは Big Data 解析における良いプログラムの条件の一つである可読性の高いプログラムを記述でき，かつ最大の特徴と考える Tables (プログラムメタデータ) という柔軟なデータ構造を有しており，このプロシジャを起点としたチームプログラミングを行うことで，チーム内で同一認識のもとで最適な解析プロセスを構築でき，同時に知識共有 / ナレッジの蓄積を行うことが可能である．本論文では，プログラム (Lua プロシジャ) を起点とした解析プロセスの最適化の提案を行う．

キーワード：Big Data 解析， Lua プロシジャ，解析プロセスの最適化，チームプログラミング

## 1. はじめに

Big Data という言葉が日本で流行しバズワードと呼ばれた時期も過ぎて，各企業，団体が本腰を入れて Big Data の利活用を推進している．Big Data と言っても，構造化データや非構造化データに代表されるように様々な構造を持ったデータが存在する．製薬企業が保有するデータもその例外ではなく，例えば，縦長データ（Ex. 医療ビッグデータ，副作用自発報告データ）や横長データ（ゲノムビッグデータ）の構造を持ったデータを保有している．また，データの種類に関しても，数値，テキスト，画像，音声，センサーなど様々である．データ構造やデータの種類によって，当然解析アプローチと解析手法

は異なる.加えて,これら以外にも,データクリーニング,欠測の扱いに関する検討 (Ex. 補完 / 削除),正規化,解析結果の評価,考察,考察後の追加解析など,Big Data 解析では多くの解析プロセスを検討する必要がある.それと同時に,各プロセスでの適切な手法等の選択はもちろん,ステップの組み合わせも検討しなければならない.このように多くの段階を踏んで解析が実施されることもあり,Big Data 解析は通常複数人,複数機能で実施される.そこで,重要になるのが「知識共有」と「ナレッジ蓄積」である.これはプログラムも同様であり,過去に実施した解析プログラムをナレッジとして蓄え,アクセス可能な状態にすること,さらには現在進行している解析についてはそのプログラムを関係者間で可読性の高い状態で共有することで,解析フローを透明化し,集合知的な発想を促すことが重要である.以上を踏まえると,Big Data 解析では,「解析プロセスの最適化」と「チームプログラミング」がキーワードになるだろう.

上記のキーワードを実現するために,SAS ® 9.4 から正式に実装された Lua プロシジャの活用することを考えた.本論文では,2 章で従来の SAS プログラムと Lua プロシジャのプログラムの比較を行ながら,Lua プロシジャの特徴をまとめ,3 章では Lua プロシジャを起点としたチームプログラミングの在り方と解析プロセスのパラダイムシフトについて議論し,最適な解析プロセスの構築を提案する.

## 2. Lua プロシジャ

Lua プロシジャは LUA 言語を基にしている.LUA 言語は C 言語をベースにしたスクリプト言語で幅広い業界で利用されている言語である.この LUA 言語が組み込まれたことで,SAS プログラミングがより簡便になった.加えて,LUA のデータ構造がサポートされたことで,これまでの SAS マクロの限界を突破し,より可読性の高いプログラムを構築できるようになった.Lua プロシジャの特徴は多々あるが,最も特徴的なのは Tables (プログラムメタデータ) と呼ばれる柔軟なインメモリーのデータ構造だろう.この Tables は,表 2 に示した Value タイプのデータを格納することが出来る (Anand, 2017).つまり,どのようなデータ形式でも格納可能なデータ構造である.

**表 2. Tables に格納可能な Value タイプ一覧 (Anand, 2017)**

| Type | Meaning | Example |
|---|---|---|
| Nil | Represents absence of a value | Nil (single value only) |
| Boolean | True or False | True / False |
| Number | Double precision floating point | 1.2356 |
| String | Sequence of characters | “Hello World” |
| Table | Associative array | x[“key”] = 100, x[2] = “value” |
| Function | Represents a method | type(), print() |
| Userdata | Arbitrary C data | lua_newuserdata() used |
| thread | Represents Threads of excecution | Coroutine |

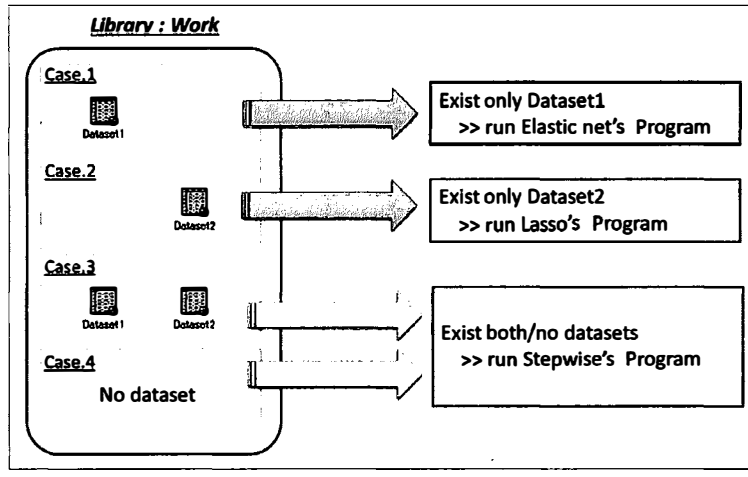
LUA のコーディングは非常にシンプルである.例えば ISO 国名コード USA, JPN, GBR, CAN, DEU を Tables に格納するプログラムを考える.ローカルで定義した変数を Countryname_L, グローバルで定義した変数を Countryname_G とすると,下記のように書ける.submit ステートメントと endsubmit ステートメントの間で LUA 言語を記述できる.LUA のデータ構造とシンプルなコーディングは非常に便利であり,可読性の向上に繋がっていることが分かる.

```
proc lua;
     submit;
          local Country_name_L = {'USA', 'JPN', 'GBR', ' CAN', 'DEU'}
                                   -- ローカルで定義. Country_name_L に文字値を格納
                Country_name_G = {'USA', 'JPN', 'GBR', ' CAN', 'DEU'}
                                   -- グローバルで定義. Country_name_G に文字値を格納
     endsubmit;
run;
```

次に，この Tables を活用し，SAS マクロ処理を簡略化したプログラムの例，及びデータ状況を踏まえるプログラム（可変的なプログラム）の例を示す. この例を通じて従来の SAS プログラムと比較し, Lua プロシジャの特徴を把握する.

## 2.1 SAS マクロ処理を簡略化したプログラム

表 2 で示したように，LUA の Tables は任意のデータ形式でデータ格納が可能なため，当然 SAS プログラムを変数として格納することが可能である. SAS プログラムをグローバル変数として予め格納しておき，sas.submit 関数でその変数を呼び出し実行することで，従来の SAS プログラムよりもシンプルなプログラムで記述できる．例えば，以下の処理を行うプログラムを記述してみる.




従来の SAS マクロを使用して記述した場合と Lua プロシジャで記述した場合のプログラムを下記に示す. 可読性という観点で比較すると, Lua プロシジャのプログラムの方が可読性に優れている. これは, 以下の 3 点が要因だろう.

1. SAS マクロ処理を簡略化できたこと
2. 実行したい機械学習（Elastic net, Lasso, Stepwise）のプログラムをメタデータとして整理した状態で，事前に準備したこと（下記プログラムの破線枠内）
3. 機械学習プログラムのメタデータ作成プログラム（下記プログラムの Step1）と，Dataset の有無を確認し各機械学習を実行するためのプログラム（下記プログラムの Step2）を切り分けたこと

■ 従来の SAS マクロを使用したプログラム

```
%MACRO ML_MACRO(lib,data1, data2);
  %IF (%SYSFUNC(EXIST(&lib..&data1.) )=1 and %SYSFUNC(EXIST(&lib..&data2.) )=0) %THEN
      %DO; proc glmselect data=work.Data plots=all ;
                model OUTCOME=COL1-COL10 / selection=elasticnet(steps=1000 L2=0.1  choose=AIC) ;
           run ;
      %END;
  %ELSE %IF( %SYSFUNC(EXIST(&lib..&data1.) )= 0 and %SYSFUNC(EXIST(&lib..&data2.) )= 1) %THEN
      %DO; proc glmselect data=work.Data plots=all ;
                model OUTCOME=COL1-COL10 / selection=lasso(steps=1000 choose=AIC) ;
           run ;
      %END;
  %ELSE  %DO; proc glmselect data=work.Data plots=all ;
               model OUTCOME=COL1-COL10 / selection=stepwise(select=SL choose=AIC) ;
           run ;
      %END;
%MEND;
%ML_MACRO (Work,  dataset1,  dataset2)
```

■ Lua プロシジャを用いたプログラム

```
proc lua ;
    submit;
      -- グローバル変数 MLcode を定義
      MLcode={}
      -- Step1. ML_code の中でリストとして各機械学習プログラムを事前に格納しておく
       -- Elastic net の実行プログラムを格納
        MLcode.elasticnet = [[
             proc glmselect data=work.Data plots=all ;
               model OUTCOME=COL1-COL10 / selection=elasticnet(steps=1000 L2=0.1  choose=AIC) ;
             run ; ]]
       -- Lasso の実行プログラムを格納
        MLcode.lasso      = [[
             proc glmselect data=work.Data plots=all ;
               model OUTCOME=COL1-COL10 / selection=lasso(steps=1000 choose=AIC) ;
             run ; ]]
       -- Stepwise の実行プログラムを格納
        MLcode.stepwise   = [[
             proc glmselect data=work.Data plots=all ;
               model OUTCOME=COL1-COL10 / selection=stepwise(select=SL choose=AIC) ;
             run ; ]]

      -- Step2. Dataset の有無を確認しプログラムを実行する
       if sas.exist("work.dataset1")~=0 and sas.exist("work.dataset2")==0 then
                                                                sas.submit(MLcode.elasticnet)
       elseif sas.exist("work.dataset1")==0 and sas.exist("work.dataset2")~=0 then
                                                                sas.submit(MLcode.lasso)
       else                                                     sas.submit(MLcode.stepwise)
       end
    endsubmit;
quit;
```

## 2.2 データ状況を踏まえるプログラム

データに依存したプログラムを記述することは効率的なことではなく，データに依存しないプログラムを記述することが望ましい．つまり，データ状況を踏まえる，可変的なプログラムで記述することが望ましい．この可変処理も Lua プロシジャで実現可能である．実装例を示す． 図3 に示す国/エリア別の 1US ドルあたりの通貨と ISO 通貨コードが格納されたデータがあるとする（2017/02/07 時点）．このデータを基に, 図 4 に示した各国/エリアの通貨を基準とした場合の他の通貨の値を算出するプログラムを記述する．従来の SAS プログラムと Lua プロシジャを用いたプログラムを下記に示す．Lua プロシジャで記述したプログラムの方が長いが，可変性の観点では優れている．例えば，新たに「Country/Erea = 'NewZealand', ISO CODE = 'NZD', Currency per USD = 1.38」というデータが追加された場合，従来の SAS マクロを利用したプログラムは%*NEW_VALURE*(NewZealand') を追記する必要があるが, Lua プロシジャは追記不要である．もちろん，For 文等で記述すれば追記が不要なプログラムを書けるが，America を除く処理を加味すると煩雑なプログラムになり，可読性の観点から見劣りするだろう．

| | Country/Erea | ISO CODE | Currency per USD |
|---|---|---|---|
| 1 | Japan | JPY | 112.38 |
| 2 | America | USD | 1 |
| 3 | Euro | EUR | 0.93 |
| 4 | Britain | GBP | 0.8 |
| 5 | Koria | KRW | 1133.02 |
| 6 | China | CNY | 6.85 |
| 7 | Australia | AUD | 1.31 |
| 8 | Canada | CAD | 1.3 |
| 9 | Switzerland | CHF | 0.99 |
| 10 | SouthAfrica | ZAR | 13.29 |

**図 3. 国/エリア別の 1US ドルあたりの通貨とその ISO 通貨コードのデータ**

| | Country/Erea | ISO CODE | Currency per USD | Currency per JPY | Currency per EUR | Currency per GBP | Currency per KRW | Currency per CNY | Currency per AUD | Currency per CAD | Currency per CHF | Currency per ZAR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Japan | JPY | 112.38 | 1 | 120.8387 | 140.475 | 0.0992 | 16.4058 | 85.7863 | 86.4462 | 113.5152 | 8.456 |
| 2 | America | USD | 1 | 0.0089 | 1.0753 | 1.25 | 0.0009 | 0.146 | 0.7634 | 0.7692 | 1.0101 | 0.0752 |
| 3 | Euro | EUR | 0.93 | 0.0083 | 1 | 1.1625 | 0.0008 | 0.1358 | 0.7099 | 0.7154 | 0.9394 | 0.07 |
| 4 | Britain | GBP | 0.8 | 0.0071 | 0.8602 | 1 | 0.0007 | 0.1168 | 0.6107 | 0.6154 | 0.8081 | 0.0602 |
| 5 | Koria | KRW | 1133.02 | 10.082 | 1218.3011 | 1416.275 | 1 | 165.4044 | 864.9008 | 871.5538 | 1144.4646 | 85.2536 |
| 6 | China | CNY | 6.85 | 0.061 | 7.3656 | 8.5625 | 0.006 | 1 | 5.229 | 5.2692 | 6.9192 | 0.5154 |
| 7 | Australia | AUD | 1.31 | 0.0117 | 1.4086 | 1.6375 | 0.0012 | 0.1912 | 1 | 1.0077 | 1.3232 | 0.0986 |
| 8 | Canada | CAD | 1.3 | 0.0116 | 1.3978 | 1.625 | 0.0011 | 0.1898 | 0.9924 | 1 | 1.3131 | 0.0978 |
| 9 | Switzerland | CHF | 0.99 | 0.0088 | 1.0645 | 1.2375 | 0.0009 | 0.1445 | 0.7557 | 0.7615 | 1 | 0.0745 |
| 10 | SouthAfrica | ZAR | 13.29 | 0.1183 | 14.2903 | 16.6125 | 0.0117 | 1.9401 | 10.145 | 10.2231 | 13.4242 | 1 |

**図 4. 各国/エリアの通貨を基準とした場合のデータを付加したデータ**

■ **従来の SAS マクロを利用したプログラム**

```
%MACRO NEW_VALURE(Country);
  data _null_;
    set work.meta(where = (Country="&Country.")  );
    call symputx("value", value);
    call symputx("unit", unit);
  run;
  data work.meta;
    set work.meta;
    &Country._value = round(value/&value., 10**-4);
    label &Country._value = "Currency per &unit.";
  run;
%mend NEW_VALURE;

%NEW_VALURE(Japan) %NEW_VALURE(Euro) %NEW_VALURE(Britain) %NEW_VALURE(Koria) %NEW_VALURE(China)
%NEW_VALURE(Australia) %NEW_VALURE(Canada) %NEW_VALURE(Switzerland) %NEW_VALURE(SouthAfrica)
```

■ **Lua プロシジャを用いたプログラム (プログラム解説なし．解説ありは Appendix に掲載)**

※より簡略化したプログラムも記述できるが，従来の SAS マクロを利用したプログラムと同様の処理を行うように記述している

```
proc lua;
   submit;
      local data = sas.open("work.meta")
      sas.where(data, "Country not in ('America')")
       while sas.next(data) do
         sas.submit_([[data work.meta_@Country@; ]],{Country= sas.get_value(data,"Country")})
         sas.submit_([[set work.meta;]])
         sas.submit_([[@Country@_value = round(value/@value@, 10**-4);;]],
                   {Country= sas.get_value(data,"Country"), value= sas.get_value(data,"value")})
         sas.submit_([[label @Country@_value = "Currency per @unit@";]],
                     {Country= sas.get_value(data,"Country"), unit= sas.get_value(data,"unit")})
         sas.submit([[run;]])
      end
      sas.submit([[data Result;  merge meta_:; run;]])
      sas.close(data)
   endsubmit;
run;
```

本章で述べた Lua プロシジャの特徴をまとめる．

✓ Tables (プログラムメタデータ)
- 任意のデータ形式でデータを格納することが可能である
- メタデータとして，整理された状態でプログラミング属性を保持できる

✓ 高い可読性
- LUA 言語でコーディングすることが可能で，SAS マクロ処理を簡略化できる
- 反復操作 (処理) が容易である
- プログラムを辞書形式で Tables に事前格納することで，実行プログラムを簡潔に記述できる

✓ データ駆動型のプログラミング
- 解析対象データの状況を事前に定義することで，変数やレコードなどが追加・削除されても，プログラムの修正なしで，データの状況を踏まえた解析が実装できる

上記の特徴を踏まえて，3 章では，Lua プロシジャを起点とした Big Data 解析のチームプログラミングの在り方と解析プロセスのパラダイムシフトを通じて，最適な解析プロセスの構築について議論する．

## 3. Lua プロシジャによるチームプログラミングと解析プロセスのパラダイムシフト

本章では，Lua プロシジャを起点とした最適な解析プロセスの構築について議論する．Big Data 解析は，1 章で示したように多くの解析プロセスを踏む背景もあって，通常複数人，つまりチームで実施される．そのため，知識共有とナレッジ蓄積が重要となる．これらをキーワードとして，まず 3.1 節で Lua プロシジャを活用したチームプログラミングの在り方と解析プロセスの構築の在り方について提案する．この提案を踏まえた上で，3.2 節では Lua プロシジャを起点とした従来の解析プロセスからのパラダイムシフトを述べる．なお，以降の議論は，チーム単位で Big Data 解析を実施することを前提としている．

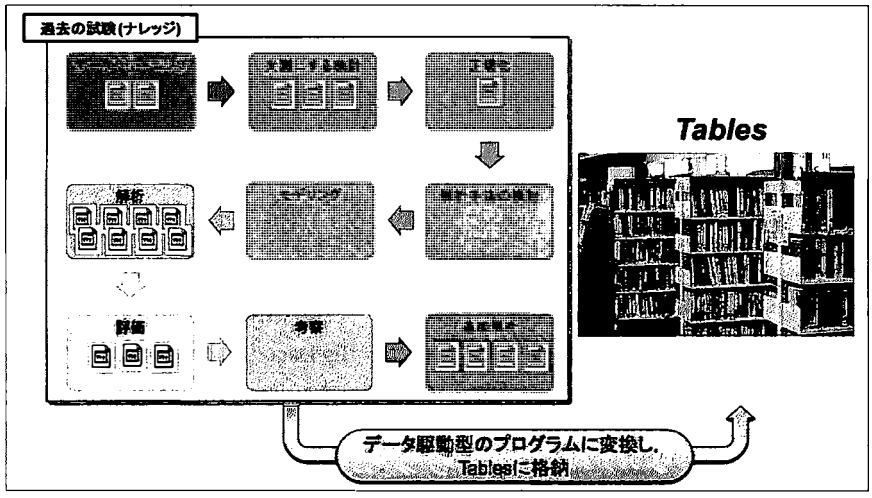
### 3.1 Lua プロシジャによるチームプログラミングと解析プロセスの在り方

Big Data 解析は，多くの解析プロセスを踏む背景があると同時に，探索解析の色合いが濃いため，解

析対象データに対して様々な解析手法を適用することになる．したがって，実際の Big Data 解析では，過去に実施した解析を適用したり，異なるバックグランド（専門分野など）を持った Data Manager, Informatician, Statistician, Data Scientist 等 がチームにアサインされることになる．これを勘案すると，下記の 2 つの観点が重要であり，これらは Lua プロシジャを活用することで実現が可能である．

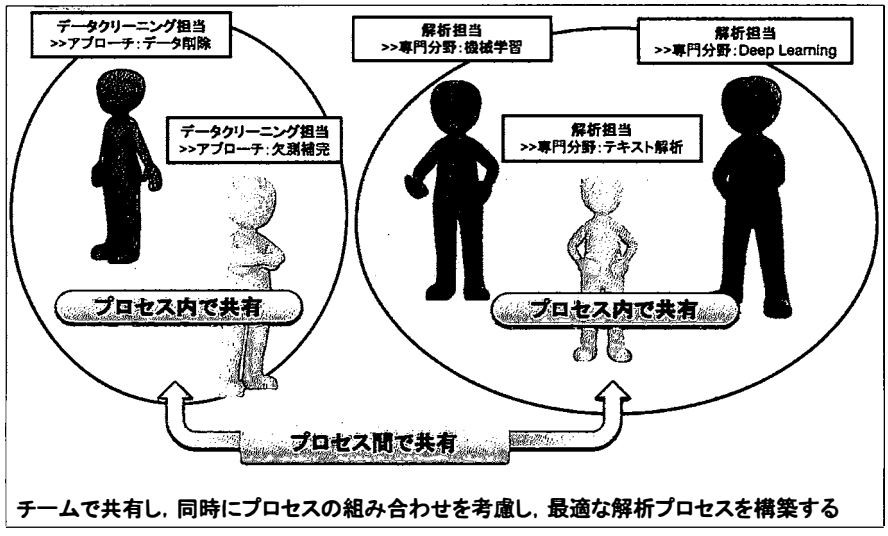
✓ 過去に実施した解析プログラムをナレッジとして蓄え，アクセス可能な状態にすること

- Tables を活用して，過去に実施した解析プログラムをリスト形式で格納することでナレッジを蓄積する (2.1 節の Lua プロシジャを用いたプログラム内の Step1 を参照)
- 格納する際，データ駆動型のプログラムで記載することで，いつでもアクセスが可能な状態にする




✓ プログラムをチーム内で可読性の高い状態で共有すること

- Lua プロシジャは潜在的に可読性が高い
- Tables に格納しておくことで，チームメンバーが作成したプログラムを横並びで比較，共有できる
- 事前に各プロセスのプログラムを格納しプロセス内 / 間で共有することで，最適な組み合わせの検証が容易に行うことができる (2.1 節の Lua プロシジャを用いたプログラム内の Step2 を参照)



過去の解析プログラムのナレッジがアクセス可能になることで，逐次利用可能となり，さらにチーム内で可読性の高い状態でプログラムを共有することで，チームメンバーが同一認識の下，リアルタイムでプロセスの組み合わせ等を検討及び探索し，最適な解析プロセスを効率的に構築することができる．

### 3.2 Lua プロシジャを起点とした従来の解析プロセスからのパラダイムシフト

3.1 節で述べたように，Big Dat a解析は探索解析の色合いが濃い．したがって，プログラムの観点で見ると，「データ設計書などに基づきプログラムを組む」よりも「実際のデータに基づきプログラムを組む」ことになるだろう．つまり，「演繹的プログラミング」よりも「帰納的プログラミング」を Big Dat a 解析では行うことになる．一方で，実際の Big Data 解析では，「演繹的解析プロセス」であることが多い．解析対象データの選定後，例えばデータクリーニングをあるチームメンバーが行い，次に別のチームメンバーが欠測の取り扱いの検討と処理を実施し，ようやくの解析担当のチームメンバーたちが探索解析を開始する，といった具合である．なぜ一足飛びに解析を実施できないのか？過去の解析プログラムやプロセスのナレッジを活用すれば，実施は容易に可能であるように思える．要因は様々あると思うが，プログラムがデータ駆動型でないことが，一因だろう．過去の解析プログラムやプロセスのナレッジは大抵の場合迅速なアクセスができないからだ．しかし，Lua プロシジャのプログラムでは，2.2 節及び 3.1 節で述べたようにデータ駆動型のプログラミングが可能であり迅速なアクセスを実現できる．Lua プロシジャを起点とすることで，データに基づく解析プロセスの構築が可能になり，解析を起点とした解析プロセスの最適化を検討できる．爆発的に増え続ける Big Dat aの解析を効率的に実施する上で「帰納的解析プロセス」へのパラダイムシフトは今後直面する課題の一つと思われるが，その一助として Lua プロシジャは強力なツールと言える．

## 4. おわりに

Lua プロシジャは，2 章で紹介したような従来の SAS プログラムにはない特徴を多く有している．本論文では，Lua プロシジャの特徴を最大限活用する事例として，Lua プロシジャを起点とした Big Dat a 解析のチームプログラミングの在り方と解析プロセスのパラダイムシフトを通じて，最適な解析プロセスの構築を挙げたが，2 点補足する．1 点目は，Lua プロシジャではデータ駆動型のプログラムを記述できるが，解析対象データの状況定義が不可欠である．2. 2 節で述べたように変数やレコードなどが追加・削除されてもデータ状況に合わせた可変的な処理が可能であるが，想定されるデータ状況をプログラム内で網羅的に定義しておく必要がある．例えば，変数が追加されたときの処理定義，レコードが増えたときの処理定義，複数のデータセットを Merge する場合の Key 変数が無くなったときの処理定義などがある．この定義は手間のかかる作業であるが，ある解析プログラムで一度定義すれば他の解析プログラムへの転用が可能であるため，継続的なタスクではない．2 点目は，解析プログラムのナレッジをアップデートしていくことである． 1 点目に挙げた定義の見直しや，新規の解析手法があればナレッジとして登録するなどを，定期的に行う必要がある．

1 人のプログラマー目線で SAS を眺めると，データハンドリングやプロシジャ活用による高度な解析など，プログラムの観点に目が行きがちであるが，視点を変えて SAS を眺めると別の知見が得られることがある．今回はチーム目線で Lua プロシジャを改めて眺めてみることで，Lua プロシジャを起点としたチームビルディングの構築という新しい発見があった．別の視点から SAS を眺めることで，既存の機能でも新たな発見があるかもしれない．

## 参考文献

Paul Tomas (2015). "Driving SAS® with Lua." SAS Global Forum 2015.

Anand Vijayaraghavan (2017). "A Comparison of the LUA Procedure and the SAS® Macro Facility" SAS Global Forum 2017.

Ryo Kiguchi (2017). "N-Stage Machine Learning Analysis with LUA Procedure Helps Solve Big Data Analysis Problems." SAS Global Forum 2017.

## Appendix　データ状況を踏まえたプログラム：Lua プロシジャを用いたプログラム (解説付き)

```
proc lua;
   submit;
      -- 解析対象のデータを関数sas.openで処理可能な状態にする．処理が終われば関数sas.closeを必ず実行する
      -- Tables (data) にmetaデータ (Figure.3のデータ) を格納する
      local data = sas.open("work.meta")
      -- Americaの変数付与は不要のため，処理対象から除く
      sas.where(data, "Country not in ('America')")
       -- 関数sas.nextでSASデータセット内の最初のレコードから最終レコードまで順番に見ていく．
       -- while文と組み合わせることで，最初から最終レコードまで1レコードずつ反復処理を行う宣言をする．
       while sas.next(data) do
         -- 関数sas.submitでdataステートメントを実行する．
         -- @Country@のCountryは引数．{}内で指定したデータをCountryに格納する．関数sas.get_valueを用いてデ
            ータを指定し利用可能にする．関数sas.get_valueでは，指定するデータが格納されているTables(data)，及
            び指定するデータの変数名 (今回はCountry)を引数として与える
         sas.submit_([[data work.meta_@Country@; ]],{Country= sas.get_value(data,"Country")})
         --解析対象データ (Figure.3)をsetする．
         sas.submit_([[set work.meta;]])
         -- 指定したCountryを基準とした場合の他の通を計算する
         sas.submit_([[@Country@_value = round(value/@value@, 10**-4);;]],
                   {Country= sas.get_value(data,"Country"), value= sas.get_value(data,"value")})
         --labelステートメントで作成した変数にラベルを付与する
         sas.submit_([[label @Country@_value = "Currency per @unit@";]],
                     {Country= sas.get_value(data,"Country"), unit= sas.get_value(data,"unit")})
         --runステートメントを実行する
         sas.submit([[run;]])
      -- while文のend
      end
      -- while文で作成した各国/エリアSASデータセットをmergeして処理終了
      sas.submit([[data Result;  merge meta_:; run;]])
      --全ての処理が終了したので，関数sas.closeで終了を宣言する
      sas.close(data)
   endsubmit;
run;
```

# ODS機能とREPORTプロシジャを用いた解析帳票の作成

## ―これから業務で利用を始める人のために―

太田　裕二

マルホ株式会社　データサイエンス部

Creation of Tables using ODS and REPORT Procedure

-For Users Begin to Work in Earnest-

Yuji Ohta

Data Science Dept., Maruho Co, Ltd.

## 要旨

SAS に ODS 機能が導入され RTF 形式のファイルに直接出力することができるようになって久しいが、弊社ではこれまで総括報告書用の解析帳票は Excel 形式で作成していた。社内の業務フロー検討の中で解析帳票を RTF 形式にする方針となり、RTF 形式対応の解析帳票テンプレート集を作成することとなった。

今野ら（2011）では、ODS RTF 及び REPORT プロシジャを用いて RTF 形式の一覧表を作成するプログラムが紹介されており、比較的簡単なプログラムのみで体裁がコントロール可能であることがわかる。しかしながら、これまでの私の経験では DDE 機能を用いた Excel 形式の解析帳票の作成が中心であり、ODS RTF 及び REPORT プロシジャに触れる機会が少なかった。そのため、体裁コントロールの概念や、期待する体裁にするためにどのようにスタイル定義を行うかを理解するために時間を費やした。また、多くの機能を備えた REPORT プロシジャの各ステートメントやオプションの役割を理解することにも苦渋した。

Ver.9.4 で ODS RTF 及び REPORT プロシジャを用いて解析帳票を作成する方法を整理した。私と同様にこれから業務で利用を開始する人を対象に紹介する。

キーワード：ODS RTF、REPORT プロシジャ、スタイル要素、スタイル属性、インラインフォーマット

## 1. 背景

承認申請時の CDISC 準拠電子データ提出の義務化を受けて、CDISC 準拠データを介した効率的な業務フローの検討を行っている。総括報告書用の解析帳票が英語データで構成される ADaM から作成されることを考慮して解析帳票の言語を英語にすることとなり、解析帳票テンプレート集を全面的に見直すこととなった。また、このタイミングで従来からの Word との親和性の高い RTF 形式で解析帳票を提供してほしいというメディカルライティング部門の要望にも対応することとした。これらの状況を踏まえ、RTF 形式対応の解析帳票テンプレート集を作成した。

解析帳票テンプレート集には、第 II 相以降の試験を想定した被験者背景、有効性、安全性の解析帳票及び症例一覧表を掲載した。有効性については連続値、カテゴリ、2 値、生存時間のデータに対応したそれぞれのテンプレートを掲載した。

事前にメディカルライティング部門と検討した解析帳票の要件を以下に示す。

| 要件 | 対応 |
|---|---|
| 書式の統一 | ✧ フォントは Couri eNew、8pt に統一<br>✧ 背景色は白、文字は黒に統一 |
| 総括報告書の本文中に編集なしで貼付できる | ✧ 用紙の向きは原則縦向き（発現時期別の有害事象発現状況などは除く）<br>✧ 余白は上 3cm、下 3cm、左 2.5cm、右 2.5cm<br>✧ 解析帳票の仕様（解析手法や値の導出方法など）等を説明する脚注は表の直下に表示（ただし、ページ番号等の情報は Word のフッター機能を用いて表示） |
| 1 ページの出力データを多くし、出力ページを少なくする | ✧ 複数の出力項目を同一列に表示する<br>✧ セル間やセル内の余分なスペースを省く<br>✧ 空白行ではなく、罫線で行間の区切りを示す |

## 2. ODS 機能

### 2.1 一般的な SAS プロシジャ

ODS 機能を使用することでプロシジャや DATA ステップからの結果を RTF、HTML、PDF などの形式の外部ファイルに出力することができる。大半の SAS プロシジャは図 1 に示したプロセスで結果を外部ファイルに出力する。

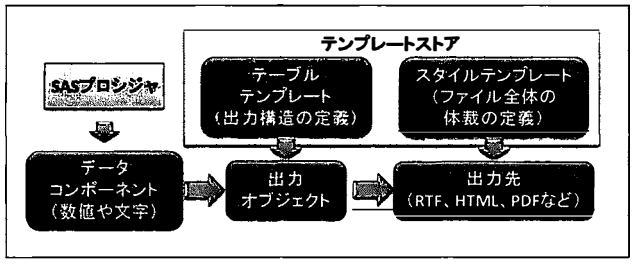



図 1 SAS プロシジャの出力プロセス（一部プロシジャを除く）

SAS プロシジャを実行すると、まず、プロシジャが導出したデータ（数値や文字）で構成されたデータコンポーネントが作成される。次に、データコンポーネントが出力構造を定義したテーブルテンプレートと結合して出力オブジェクトとなり、その後出力オブジェクトがスタイルテンプレートによって全体的な体裁（色、フォント、罫線、マージンなど）が定義されたファイルに出力される。

出力ファイルの体裁は、スタイルテンプレートでスタイル要素を定義することによってコントロールする。スタイル要素とは出力ファイルの特定の部分に適用するスタイル属性を集めたものであり、図 2 に示した「Sys temTil te」、「Head er」のような部分ごとに定義する。（全スタイル要素及び全スタイル属性は SAS 9.4 Output Deli verySys tem: User's Guid eの Cha pterl 1 及び Cha pterl 2 で参照可能）

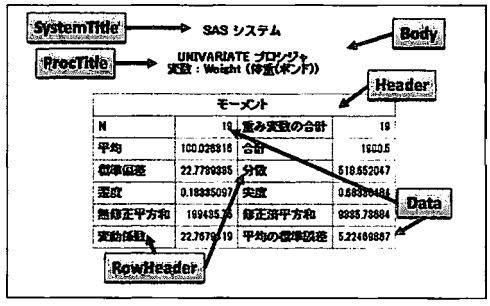



図 2 UNIVARIATE プロシジャの HTML 出力結果

スタイル要素の属性の数や種類は部分ごとに異なるため、体裁をコントロールするためには各部分のスタイル要素名や必要なスタイル属性を特定する必要がある。ODS MARKUP ステートメントを利用することでそのプロセスを簡略化することができる。オプション（TYPE =STYLE_POPUP）を指定した下記プログラムでは、ポインタをかざした部分のスタイル要素名が表示され、さらにダブルクリックするとスタイル要素の属性を表示したウインドウが開く HTML ファイルを作成することができる（図 3）。

```
ods markup file = 'ODS MARKUP.html'        /* 出力Bodyファイル名              */
           stylesheet = 'StyleSheet.css'  /* 出力スタイルシートファイル名 */
           type = style_popup             /* スタイル要素表示 オプション  */
           style = Styles.HTMLblue ;      /* スタイルテンプレート名        */
  ods select Moments ;
  proc univariate data = SASHELP.CLASS ;
    var WEIGHT ;
  run ;
ods markup close ;
```

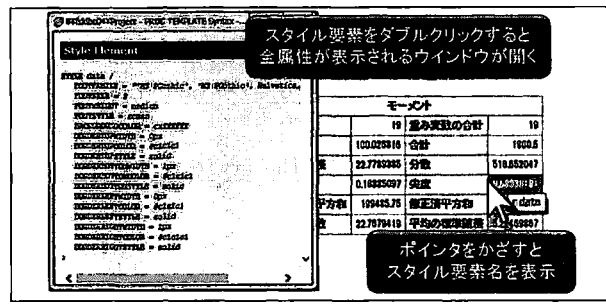



図 3　ODS MARKUP ステートメントで作成した出力ファイル

## 2.2 レポート作成プロシジャ

レポート作成プロシジャである REPORT プロシジャ、PRINT プロシジャ及び TABULATE プロシジャの出力プロセスは少し異なっている（図 4）。これらのプロシジャはテーブルテンプレートを持っておらず、結果の出力構造はプロシジャで定義される。

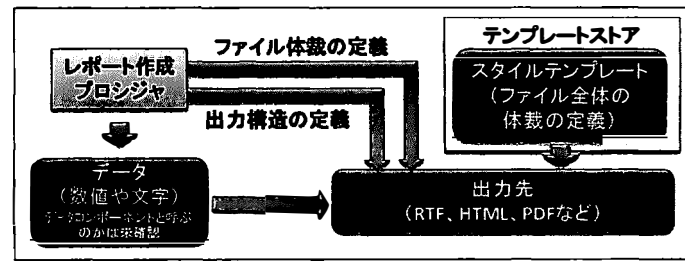



図 4　レポート作成プロシジャの出力プロセス

またプロシジャ内のステートメントで任意の出力構造に変更することもできる。出力ファイルの全体的な体裁は他のプロシジャと同様にスタイルテンプレートで定義されるが、プロシジャ内のステートメントでスタイル属性を変更することもできる。なお、スタイルテンプレートではスタイル要素名ごとにスタイル属性を指定するが、レポート作成プロシジャでは Location ごとあるいは特定のセルごとにスタイル属性を指定することができる。Location に関する詳細は 0 を参照。

# 3. 解析帳票作成プログラムの紹介

## 3.1 ODS RTF と REPORT プロシジャの主なステートメント

RTF 形式の解析帳票を作成するにはいくつかの方法があるが、ODS RTF 及び REPORT プロシジャを用いる方法を採用した。この方法で一覧表を作成するためのプログラムが今野ら（2011）で紹介されているが、比較的簡単なプログラムのみで体裁をコントロールした一覧表が作成できることがわかる。

本稿で紹介するステートメントの構文とその機能の概要を下に示した。ODS RTF 及び REPORT プロシジャにはこの他にも複数のステートメントやオプションがあるが、ここでは解析帳票テンプレート集の作成に用いたものに限定している。

```
ods rtf file = '出力ファイル名' style = スタイルテンプレート名
        notoc_data ;                                        /* 目次データの出力制御オプション */
  proc report data = 出力データ名 missing 他のオプション;    /* 欠測値を有効値として扱うオプション */
    column レポート項目1 レポート項目2 ・・・ ;                /* レポート対象となる変数 */
    define レポート項目 / オプション ;                        /* レポート項目の役割と表示方法 */
    compute 場所もしくはレポート項目 ;                        /* compute-endcomp間のSASステートメントを実行 */
      SASステートメント ;
    endcomp ;
    break タイミング ブレーク変数 / オプション ;              /* ブレーク時(ブレーク変数の値が変わった時)に */
                                                            /* オプションに指定したアクションを実行   */
  run ;
ods rtf close ;
```

注；ODS 機能を用いて解析帳票を作成するとき、事前に作成したユーザー定義スタイルテンプレートをすべての解析帳票に適用することで解析帳票を通じて書式を統一できるというメリットがある。したがって、解析帳票間で共通に使用するスタイルはスタイルテンプレートで定義し、個々の解析帳票の細かな体裁の調整に必要なスタイルはプロシジャ内で定義するという方法をとることができる。
弊社でもこの方法を採用する予定であるが、本稿ではそれぞれのステートメントが影響を与えた出力箇所を明確にするため、ユーザー定義スタイルテンプレートは用いずに REPORT プロシジャ及びグローバルステートメントのみでスタイル属性を指定する方法を紹介する。ユーザー定義スタイルテンプレートを用いて同体裁の解析帳票を作成するプログラムは付録で紹介している。

## 3.2 有効性の解析帳票を例として

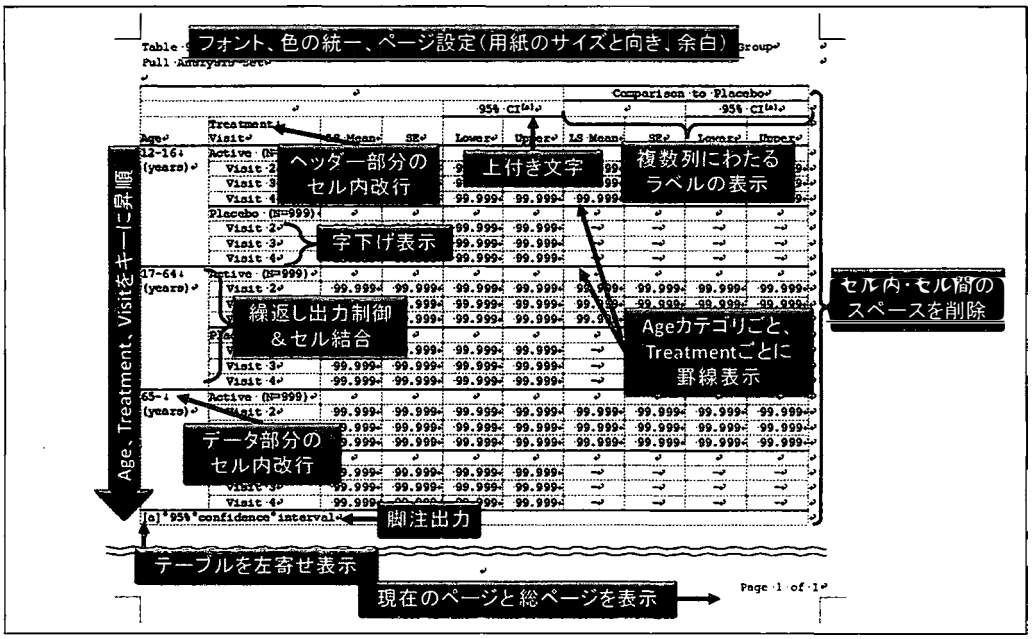
並行群間プラセボ対照試験における有効性の部分集団解析を例として、プログラムを紹介する。出力データセットの詳細と目標とする解析帳票を以下に示す。出力データセットにはすでに必要な統計量がすべて格納されていることを前提とし、REPORT プロシジャが有する要約統計量を算出する機能は使用していない。

出力データセットの詳細

| AGEGR1N | AGEGR1 | TRTPN | TRTP | AVISITN | AVISIT | TRT_VISITN | TRT_VISIT | LSMEAN | SE | LOWER | UPPER | C_LSME | C_SE | C_LOWE | C_UPPER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 12-16(years) | 1 | Active (N=999) | . | | 10 | Active (N=999) | | | | | | | | |
| 1 | 12-16(years) | 1 | Active (N=999) | 2 | Visit 2 | 12 | Visit 2 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 |
| 1 | 12-16(years) | 1 | Active (N=999) | 3 | Visit 3 | 13 | Visit 3 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 |
| 1 | 12-16(years) | 1 | Active (N=999) | 4 | Visit 4 | 14 | Visit 4 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 |
| 1 | 12-16(years) | 2 | Placebo (N=999) | . | | 20 | Placebo (N=999) | | | | | | | | |
| 1 | 12-16(years) | 2 | Placebo (N=999) | 2 | Visit 2 | 22 | Visit 2 | 99.999 | 99.999 | 99.999 | 99.999 | - | - | - | - |
| 1 | 12-16(years) | 2 | Placebo (N=999) | 3 | Visit 3 | 23 | Visit 3 | 99.999 | 99.999 | 99.999 | 99.999 | - | - | - | - |
| 1 | 12-16(years) | 2 | Placebo (N=999) | 4 | Visit 4 | 24 | Visit 4 | 99.999 | 99.999 | 99.999 | 99.999 | - | - | - | - |
| 2 | 17-64(years) | 1 | Active (N=999) | . | | 10 | Active (N=999) | | | | | | | | |
| 2 | 17-64(years) | 1 | Active (N=999) | 2 | Visit 2 | 12 | Visit 2 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 | 99.999 |
| 3 | 65-(years) | 2 | Placebo (N=999) | 3 | Visit 3 | 23 | Visit 3 | 99.999 | 99.999 | 99.999 | 99.999 | - | - | - | - |
| 3 | 65-(years) | 2 | Placebo (N=999) | 4 | Visit 4 | 24 | Visit 4 | 99.999 | 99.999 | 99.999 | 99.999 | - | - | - | - |

| 変数名 | 型 | 内容 | コードリスト |
|---|---|---|---|
| AGEGR1N | 数値 | 年齢カテゴリ（数値） | 1, 2, 3 |
| AGEGR1 | 文字 | 年齢カテゴリ | 12-16 (years), 17-64 (years), 65- (years) |
| TRTPN | 数値 | 投与群（数値） | 1, 2 |
| TRTP | 文字 | 投与群 | Active (N=999), Placebo (N=999) |
| AVISITN | 数値 | 解析時期（数値） | 2, 3, 4 |
| AVISIT | 文字 | 解析時期 | Visit 2, Visit 3, Visit 4 |
| TRT_VISITN | 数値 | 投与群と解析時期の合成変数（数値） | 10, 12, 13, 14, 20, 22, 23, 24 |
| TRT_VISIT | 文字 | 投与群と解析時期の合成変数 | Active (N=999), Visit 2, Visit 3, Visit 4, Placebo (N=999), Visit 2, Visit 3, Visit 4 |
| LSMEAN | 文字 | 投与群ごとの最小二乗平均 | — |
| SE | 文字 | 投与群ごとの標準誤差 | — |
| LOWER | 文字 | 投与群ごとの 95%信頼区間下限 | — |
| UPPER | 文字 | 投与群ごとの 95%信頼区間上限 | — |
| C_LSMEAN | 文字 | 投与群間の差の最小二乗平均 | — |
| C_SE | 文字 | 投与群間の差の標準誤差 | — |
| C_LOWER | 文字 | 投与群間の差の 95%信頼区間下限 | — |
| C_UPPER | 文字 | 投与群間の差の 95%信頼区間上限 | — |

目標とする解析帳票




### 3.2.1 出力順・変数の非表示・ヘッダーラベル

AGEGR1N と TRT_VISITN のように、データの出力順を指定するために使用するが解析帳票には表示させない変数がある。このとき、これらの変数を COLUMN ステートメントでレポート項目として変数名を記載する/*2*/。さらに、DEFINE ステートメントでは当該レポート項目の役割を ORDER として定義するとともに、NOPRINT オプションで非表示にする/*5*/。なお、AGEGR1 のように文字変数で役割を省略する場合のデフォルトの役割は DISPLAY である。数値変数のデフォルトの役割は ANALYSIS である。

DEFINE ステートメントのラベルオプションでは、ヘッダー部分に表示するラベルを指定する/*6*/。ラベルオプションを省略した場合には変数ラベル（変数ラベルが設定されていない場合には変数名）が表示される。ヘッダーラベルを任意の場所でセル内改行させて表示させるときは、/*1*/の SPLIT オプションで指定した区切り文字（デフォルトの区切り文字は'|'（スラッシュ））をテキスト内の改行させたい場所に挿入する/*7*/。

COLUMN ステートメントでラベルを追加することができる。/*3*/では LOWER と UPPER の 2 つのレポート項目にわたるラベルを表示させている。/*4*/では C_LOWER、C_UPPER にわたるラベルと、さらにその上に C_LSMEAN、C_SE、C_LOWER 及び C_UPPER の 4 つのレポート項目にわたるラベルを表示させている。

```
ods rtf file = "有効性の解析帳票_01.rtf" notoc_data style = Styles.Default ;
  proc report data = DATA missing split = '|' ;                          /* 1 */
    column AGEGR1N AGEGR1 TRT_VISITN TRT_VISIT                            /* 2 */
           LSMEAN SE ('95% CI' LOWER UPPER)                               /* 3 */
           ('Comparison to Placebo' C_LSMEAN C_SE ('95% CI' C_LOWER C_UPPER)) ; /* 4 */
    define AGEGR1N    / order noprint ;                                   /* 5 */
    define AGEGR1     / 'Age' ;                                           /* 6 */
    define TRT_VISITN / order noprint ;
    define TRT_VISIT  / 'Treatment|Visit' ;                               /* 7 */
    define LSMEAN     / 'LS Mean' ;
    （省略）
    define C_UPPER    / 'Upper';
  run ;
ods rtf close ;
```

### 3.2.2　セル結合・脚注

DEFINE ステートメントで役割が ORDER として定義されたレポート項目では、同じデータが連続している場合に行間で同じデータを繰り返して表示することが制御される/*2*/。また同時に PROC REPORT ステートメントで SPANROWS オプションを指定することで制御されたセルを結合することができる/*1*/。

/*3*/は、各出力ページの最後に COMPUTE と ENDCOMP との間に記述された SAS ステートメントを実行させる。LINE ステートメントで任意のテキストを出力しており、ここでは脚注を出力している/*4*/。

```
 : (省略)
  proc report data = DATA … (省略)  spanrows ;   /* 1 */
    : (省略)
    define AGEGR1      / order 'Age' ;            /* 2 */
    : (省略)
    compute after _page_ ;                        /* 3 */
      line '95% confidence interval' ;            /* 4 */
    endcomp ;
  run ;
 : (省略)
```

### 3.2.3　スタイル

これまでに紹介したプログラムで作成される結果を図 5 に示した。この結果の色やフォントは PROC REPORT ステートメントの STYLE オプションで指定されたスタイルテンプレート（STYLES.DEFAULT）に従っている。REPORT プロシジャでは Location ごとあるいは特定のセルごとにスタイル属性を指定することができる。（REPORT プロシジャで指定することができるスタイル属性は Base SAS 9.4 Procedures Guide の Table 55.9 で参照可能）

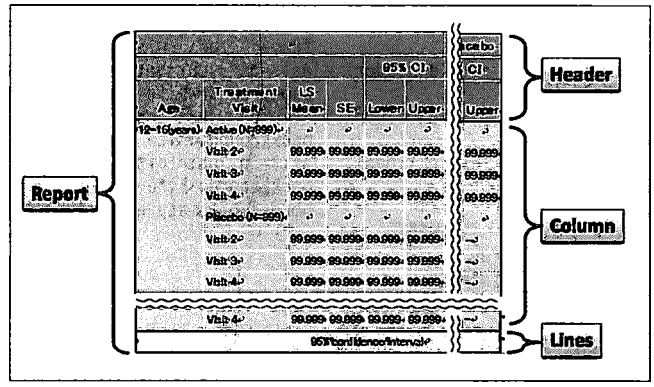



**図 5　Location**

本稿で取り扱う Location は図 5 に示した「Report」、「Header」、「Column」及び「Lines」の 4 種類である。スタイル定義の記述方法にはいくつかの方法があり、目的に応じて使い分けることができる。

<u>PROC REPORT ステートメントで記述</u>

PROC REPORT ステートメントでは、STYLE(*Location*)オプションを用いてテーブル全体に影響するスタイル属性とすべてのレポート項目に共通するスタイル属性を指定する。STYLE オプションの構文は以下の通り。

```
style(Location) = {スタイル属性 1 スタイル属性 2 ・・・}
```

/*1*/の STYLE(REPORT) オプションでは Location「Report」部分、つまりテーブル全体のスタイル属性を指定している。「just=left」ではテーブルの表示位置をファイルの中の左寄せにしている。「bordercolor=black」、「borderstyle=solid」及び「borderwidth=1pt」ではそれぞれテーブル中で共通して使用する罫線の色を黒、線種を実線、幅を 1pt と設定している。「frame=hsides」ではテーブル枠の罫線を指定しており、テーブルの上部と下部のみの表示にしている。「rules=groups」ではテーブル内の罫線を指定しており、「Header」と「Column」の間、「Column」と「Lines」の間のみの表示にしている。「cellpadding=0」と「cellspacing=0」ではそれぞれセル内の左右上下のスペースとセル間のスペースを削除している。

/*2*/、/*3*/、/*4*/では、それぞれ Location が「Header」、「Column」、「Lines」部分のスタイル属性を指定している。指定しているスタイル属性の詳細は以下の通り。

| 属性の種類 | 属性の値 | 内容 |
|---|---|---|
| asis | on | 先頭の空白と改行を無視せず表示する |
| vjust | bottom | テキストの縦方向の位置揃えを下揃えにする |
| | top | テキストの縦方向の位置揃えを上揃えにする |
| background | white | セルの背景色を白にする |
| foreground | black | セルの前景色(主にテキストの色)を黒にする |
| font | ('Courier New', 8pt) | フォントを'Courier New'、フォントサイズを 8ptにする |
| borderbottomstyle | none | セル下部の罫線を表示しない |
| just | left | テキストの横方向の位置揃えを左揃えにする |
| | right | テキストの横方向の位置揃えを右揃えにする |

```
: (省略)
  proc report data = DATA missing split = '|' spanrows
    style(report) ={just=left                                        /* 1 */
                    bordercolor=black borderstyle=solid borderwidth=1pt
                    frame=hsides rules=groups cellpadding=0 cellspacing=0}
    style(header) ={asis=on vjust=bottom                             /* 2 */
                    background=white foreground=black font=('Courier New', 8pt)}
    style(column) ={asis=on vjust=top                                /* 3 */
                    background=white foreground=black font=('Courier New', 8pt)}
    style(lines)  ={asis=on just=left vjust=top                      /* 4 */
                    background=white foreground=black font=('Courier New', 8pt)
                    borderbottomstyle=none} ;
    : (省略)
  run ;
: (省略)
```

DEFINE ステートメントで記述

PROC REPORT ステートメントで記述した場合はすべてのレポート項目に影響を与えるのに対し、DEFINE ステートメントで記述した場合は該当するレポート項目のみに影響を与える。両方のステートメントで重複してスタイル属性が指定されているレポート項目では DEFINE ステートメントでの指定が優先される。

/*1*/の STYLE(COLUMN)オプションでは、レポート項目 AGEGR1 の Location「Column」部分、つまりテーブルのデータ部分のスタイル属性を指定しており、テキストの表示位置を左揃え、セル幅を 1.6cm にしている。

/*2*/の STYLE(HEADER)オプションでは、レポート項目 AGEGR1 の Location「Header」部分のスタイル属性を指定しており、テキストの表示位置を左揃えにしている。

```
: (省略)
    define AGEGR1N    / order noprint ;
    define AGEGR1     / order
                        style(column)={just=left cellwidth=1.6cm}      /* 1 */
                        style(header)={just=left}                      /* 2 */
                        'Age' ;
    define TRT_VISITN / order  noprint ;
    define TRT_VISIT  / style(column)={just=left cellwidth=2.6cm}
                        style(header)={just=left} 'Treatment|Visit' ;
    : (省略)
  run ;
: (省略)
```

CALL DEFINE ステートメントに記述

COMPUTE と ENDCOMP との間で、IF ステートメントと CALL DEFINE ステートメントを組み合わせて使用することによって、特定のセルのスタイル属性を指定することができる。下のプログラムは年齢カテゴリが変わるときと投与群が変わるときに罫線を表示させ、かつ TRT_VISIT のデータのうち「Visit *n*」を表示す

るものにはインデント（字下げ）を施すプログラムである。なお、出力データセットにはあらかじめ OBS を特定するための数値変数_LINE_IND を追加している（図 6）。

スタイル定義を行う CALL DEFINE ステートメントの構文は以下の通りである。対象カラムに、レポート項目名あるいは_C*n*_（n は列数）を引用符で囲んで指定することにより列を特定でき、_row_と指定することですべての列を対象にすることができる。

```
call define(対象カラム, 'style', 'style=[スタイル属性 1 スタイル属性 2 ・・・] ')
```

_LINE_IND は解析帳票には表示しないが OBS を特定するために使用するため、/*1*/でレポート項目として変数名を記載している。/*2*/では NOPRINT オプションで非表示にしている。なお、_LINE_IND は数値変数であるため、役割をデフォルトの ANALYSIS から DISPLAY に変更している。

| AGEGR1N | AGEGR1 | TRTPN | TRTP | TRT_VISITN | TRT_VISIT | _LINE_IND |
|---|---|---|---|---|---|---|
| 1 | 12-16 (years) | 1 | Active (N=999) | 10 | Active (N=999) | 10 |
| 1 | 12-16 (years) | 1 | Active (N=999) | 12 | Visit 2 | 1 |
| 1 | 12-16 (years) | 1 | Active (N=999) | 13 | Visit 3 | 1 |
| 1 | 12-16 (years) | 1 | Active (N=999) | 14 | Visit 4 | 21 |
| 1 | 12-16 (years) | 2 | Placebo (N=999) | 20 | Placebo (N=999) | 0 |
| 1 | 12-16 (years) | 2 | Placebo (N=999) | 22 | Visit 2 | 1 |
| 1 | 12-16 (years) | 2 | Placebo (N=999) | 23 | Visit 3 | 1 |
| 1 | 12-16 (years) | 2 | Placebo (N=999) | 24 | Visit 4 | 21 |
| 2 | 17-64 (years) | 1 | Active (N=999) | 10 | Active (N=999) | 10 |
| 2 | 17-64 (years) | 1 | Active (N=999) | 12 | Visit 2 | 1 |
| 2 | 17-64 (years) | 1 | Active (N=999) | 13 | Visit 3 | 1 |
| 2 | 17-64 (years) | 1 | Active (N=999) | 14 | Visit 4 | 21 |
| 3 | 65- (years) | 2 | Placebo (N=999) | 24 | Visit 4 | 21 |

図 6　出力データセット（数値変数_LINE_IND を追加）

CALL DEFINE ステートメントを COMPUTE と ENDCOMP との間に記述する。COMPUTE ステートメントには必ず場所もしくはレポート項目を指定しなければならず、ここでは_LINE_IND を指定している/*3*/。

/*3*/、/*4*/は_LINE_IND=10 である OBS の AGEGR1 のセルに幅 1pt の下部罫線を表示している。ここで、罫線の色及び線種は指定していないが、PROC REPORT ステートメントの STYLE(REPORT)オプションで指定した属性に従っている。さらに、_LINE_IND の値に注目する。AGEGR1 のグループ内の最後の OBS（図 6 の 8OBS 目）ではなく最初の OBS（図 6 の 1OBS 目）に 10 が格納されていなければならない。これは、AGEGR1 は役割が ORDER として定義されており、1 行目から 8 行目までのセルが結合されていることが影響している。

/*5*/は_LINE_IND=1 である OBS の TRT_VISIT のセルにインデントを設定している。

/*6*/、/*7*/は_LINE_IND=21 である OBS の TRT_VISIT のセルにインデントを設定するとともに、TRT_VISIT 以降のすべてのセルに下部罫線を表示している。

```
: (省略)
  column AGEGR1N AGEGR1 TRT_VISITN TRT_VISIT … (省略) _LINE_IND ;                    /* 1 */
  define AGEGR1N    / order noprint ;
  define _LINE_IND  / display noprint ;                                              /* 2 */
  : (省略)
  compute _LINE_IND ;
    if _LINE_IND = 10 then                                                           /* 3 */
      call define('AGEGR1'   , 'style', 'style=[borderbottomwidth=1pt]') ;            /* 4 */
    if _LINE_IND =  1 then
      call define('TRT_VISIT', 'style', 'style=[marginleft=0.4cm]') ;                 /* 5 */
    if _LINE_IND = 21 then do ;
      call define('TRT_VISIT', 'style', 'style=[borderbottomwidth=1pt  marginleft=0.4cm]') ; /* 6 */
      call define('LSMEAN'   , 'style', 'style=[borderbottomwidth=1pt]') ;            /* 7 */
      : (省略)
      call define('C_UPPER'  , 'style', 'style=[borderbottomwidth=1pt]') ;            /* 7 */
    end ;
  endcomp;
  : (省略)
```

### 3.2.4　インラインフォーマット機能・テキストフィールドでのスタイル定義

エスケープ文字とインラインフォーマット関数を用いてテキストとセルのスタイルを変更することができる。インラインフォーマット関数の構文を以下に示す。エスケープ文字は ODS ESCAPECHAR ステートメントで任意の文字を指定することができる。

| *エスケープ文字{インラインフォーマット関数 引数}* |
|---|

主なインラインフォーマット関数は以下の通り。（この他の関数は SAS 9.4 Output Delivery System: User's Guide の Table 6.4 で参照可能）

| 関数名 | 引数 | 内容 | 例（エスケープ文字は^と表示） |
|---|---|---|---|
| SUPER | 上付きにする文字 | 文字を上付きに設定 | 'cm^{super(2)}' |
| NEWLINE | 改行したい行数 | セル内改行を挿入 | 'SOC^{newline 1}PT' |
| PAGEOF | なし | 現在のページと総ページ数を挿入 | 'Page ^{pageof}' |
| THISPAGE | なし | 現在のページを挿入 | 'Page ^{thispage} of ^{lastpage}' |
| LASTPAGE | なし | 総ページ数を挿入 | 'Page ^{thispage} of ^{lastpage}' |

また、以下の構文によりラベル、タイトル及びフットノートのようなテキストフィールドでスタイル定義が行える。

| *'エスケープ文字S={スタイル属性 1 スタイル属性 2 ・・・} テキスト'* |
|---|

/*1*/では、エスケープ文字に'~'（チルダ）を指定している。

/*3*/では、テキストフィールドでヘッダーセルのスタイル（横方向表示位置と下部罫線）を定義するとともに、SUPER 関数でテキストの一部を上付きにして「95%CI[a]」と表示させている。

/*2*/、/*4*/、/*5*/では年齢カテゴリ変数をこれまでの AGEGR1 から AGEGR1X に変更している。AGEGR1X は NEWLINE 関数を含むデータ（'12-16~{newline 1}(years)'、'17-64~{newline 1}(years)'、'65-~{newline 1}(years)'）で構成されていることから、年齢単位が改行された状態で表示される。

```
(省略)
ods escapechar = '~';                                                    /* 1 */
  (省略)
  column AGEGR1N AGEGR1X TRT_VISITN TRT_VISIT LSMEAN SE                  /* 2 */
         ('~S={just=center borderbottomwidth=1pt}95% CI~{super[a]}' LOWER UPPER)  /* 3 */
         ... (省略) _LINE_IND ;
  define AGEGR1N    / order noprint ;
  define AGEGR1X    / order                                              /* 4 */
                      style(column)={just=left   cellwidth=1.6cm}
                      style(header)={just=left  } 'Age' ;
  (省略)
  define _LINE_IND  / display noprint ;
  compute _LINE_IN ;
    if _LINE_IN = 1 then
      call define('AGEGR1X'  , 'style', "style=[borderbottomwidth=1pt]") ;  /* 5 */
    (省略)
  endcomp;
  (省略)
run ;
(省略)
```

### 3.2.5　ページ設定・タイトル・フットノート

OPTIONS ステートメントで SAS システムの日時とページを非表示にし、用紙のサイズ、向き及び余白を指定している。

タイトル及びフットノートは図 5 のいずれの Location にも属していない。そのため REPORT プロシジャ内ではなく TITLE、FOOTNOTE ステートメント内でスタイルを定義している。

```
/* ページ設定 */
options nodate nonumber                                  /* 日時非表示、ページ番号非表示 */
        papersize = "A4"                                                  /* 用紙サ イズA4 */
        orientation = portrait                          /* 用紙の向き;縦（横はlandscape） */
        topmargin = 3cm bottommargin = 3cm                                 /* 余白上下;3cm */
        leftmargin   = 2.5cm rightmargin  = 2.5cm ;                      /* 余白左右;2.5cm */

/* タイトル・フットノートの指定（位置、フォント、サイズ、テキスト）*/
title1
'~S={just=left background=white foreground=black font=("Courier New", 8pt)}
Table 9.9.9 Subgroup Analysis, MMRM for Percent Change from Baseline in VAS by Age Group' ;
title2
'~S={just=left background=white foreground=black font=("Courier New", 8pt)}Full Analysis Set' ;
footnote1
'~S={just=right background=white foreground=black font=("Courier New", 8pt)}
Page ~{pageof}' ; /* インラインフォーマット関数 */
:（省略）
```

## 3.3 複数ページにまたがる解析帳票を例に

### 3.3.1 改ページの挿入

各 OBS に表示したいページ番号が格納された変数（ブレーク変数）を追加し、BREAK ステートメントと PAGE オプションを組み合わせることで、ブレーク時（ブレーク変数の値の変更時）に改ページを挿入することができる。このとき、ブレーク変数を COLUMN ステートメントに記載するとともに、DEFINE ステートメントで役割を ORDER あるいは GROUP にする必要がある。

```
break after ブレーク変数 / page ;
```

### 3.3.2 ページごとにヘッダーを変更する

臨床検査の解析帳票では臨床検査項目ごとに改ページし、臨床検査項目名は列として表示するのではなく、ページのヘッダー部分に表示することがある。このとき、改ページは上の方法で挿入することができる。さらに、BEFORE オプションと_PAGE_オプションを指定した COMPUTE ステートメントと LINE ステートメントを使用することで各ページの最上部にテキストを表示することができる。下記プログラムでは臨床検査項目が格納された PARAMN の内容を表示することができる。なお、LINE ステートメントで変数のデータを出力する際には出力フォーマットを指定しなければならないことに注意が必要である。

```
proc format ;
  value F_PARAMN
  1 = 'Parameter Name 1'
  2 = 'Parameter Name 2'
  :（省略）
run ;
```

```
:（省略）
    compute before _page_ ;
      line PARAMN F_PARAMN. ;
    endcomp;
:（省略）
```

### 3.3.3 Word の表示機能を用いた改ページ

任意の行で改ページを行うためには各ページに出力する OBS 数を決めておく必要がある。症例一覧表ではレングスが長いデータが出力対象になることがあり、セル幅が十分に確保できない場合は自動的にセル内改行が挿入される。その結果、出力 OBS 数≠出力行数となり各ページに出力する OBS 数を決めることは難しくなる。しかし、1 症例の情報が 2 ページにわたることは可能な限り避けたい。このとき、プログラム上では改ページ処理は行わず、症例を識別するレポート項目の役割を ORDER としてかつ SPANROWS オプションを指定するのみとする。Word の「表のプロパティ」の「行の途中で改ページする」オプションを無効にし

た状態で出力ファイルを閲覧することで、1 ページに最大の症例数が表示され、かつ 1 症例の情報が 2 ページにわたって表示されることを避けることができる。

## 4. 最後に

ODS RTF 及び REPORT プロシジャを用いた体裁コントロールの概要と解析帳票を作成するためのプログラムを紹介した。弊社の解析帳票テンプレート集を作成するなかで使用したものを一通り網羅しており、これから解析帳票を RTF 形式で作成することにチャレンジしようと考えている方々には参考になるのではないかと思う。実際には今回紹介した内容以外にも多くのステートメントやオプションが使用可能であるため、より効率的な方法もあるかもしれないが、その点はご容赦いただきたい。

## 5. 参考資料


- ✧ 今野聡　富永一宏　高橋伸枝．ODS RTF と REPORT プロシジャを用いた一覧表の出力．SAS ユーザー総会 2011 論文集
- ✧ SAS 9.4 Output Delivery System: User's Guide, Fifth Edition
- ✧ Base SAS 9.4 Procedures Guide, Seventh Edition
- ✧ ODS によるレポート作成 1（SAS セミナーODS によるレポート作成コースドキュメント）
- ✧ ODS によるレポート作成 2（SAS セミナーODS によるレポート作成コースドキュメント）
- ✧ Lori S. Parsons. Enhancing RTF Output with RTF Control Words and In-Line Formatting. SAS Global Forum 2007 http://www2.sas.com/proceedings/forum2007/151-2007.pdf


## 6. 付録

```
/* テストデータ ***********************************************************************************/
%let SET_VALUE1=%str(LSMEAN  =''        ; SE  =''        ; LOWER  =''        ; UPPER  =''        ;
                     C_LSMEAN=''        ; C_SE=''        ; C_LOWER=''        ; C_UPPER=''        ;) ;
%let SET_VALUE2=%str(LSMEAN  =' 99.999' ; SE  =' 99.999' ; LOWER  =' 99.999' ; UPPER  =' 99.999' ;
                     C_LSMEAN=' 99.999' ; C_SE=' 99.999' ; C_LOWER=' 99.999' ; C_UPPER=' 99.999' ;) ;
%let SET_VALUE3=%str(LSMEAN  =' 99.999' ; SE  =' 99.999' ; LOWER  =' 99.999' ; UPPER  =' 99.999' ;
                     C_LSMEAN='-'       ; C_SE='-'       ; C_LOWER='-'       ; C_UPPER='-'       ;) ;
data DATA ;
  length AGEGR1 AGEGR1X TRTP AVISIT TRT_VISIT LSMEAN SE LOWER UPPER C_LSMEAN C_SE C_LOWER C_UPPER $200 ;
  do AGEGR1N=1 to 3 ;
    if AGEGR1N=1 then do ; AGEGR1='12-16 (years)' ; AGEGR1X='12-16~[newline 1](years)' ; end ;
    if AGEGR1N=2 then do ; AGEGR1='17-64 (years)' ; AGEGR1X='17-64~[newline 1](years)' ; end ;
    if AGEGR1N=3 then do ; AGEGR1='65- (years)'   ; AGEGR1X='65-~[newline 1](years)'   ; end ;
    TRTPN=1 ;  TRTP='Active (N=999)' ;
    AVISITN=. ; AVISIT=''        ; TRT_VISITN=TRTPN*10         ; TRT_VISIT=TRTP   ; &SET_VALUE1 ; _LINE_IND=10 ; output ;
    AVISITN=2 ; AVISIT='Visit 2' ; TRT_VISITN=TRTPN*10+AVISITN ; TRT_VISIT=AVISIT ; &SET_VALUE2 ; _LINE_IND= 1 ; output ;
    AVISITN=3 ; AVISIT='Visit 3' ; TRT_VISITN=TRTPN*10+AVISITN ; TRT_VISIT=AVISIT ; &SET_VALUE2 ; _LINE_IND= 1 ; output ;
    AVISITN=4 ; AVISIT='Visit 4' ; TRT_VISITN=TRTPN*10+AVISITN ; TRT_VISIT=AVISIT ; &SET_VALUE2 ; _LINE_IND=21 ; output ;
    TRTPN=2 ;  TRTP='Placebo (N=999)' ;
    AVISITN=. ; AVISIT=''        ; TRT_VISITN=TRTPN*10         ; TRT_VISIT=TRTP   ; &SET_VALUE1 ; _LINE_IND= 0 ; output ;
    AVISITN=2 ; AVISIT='Visit 2' ; TRT_VISITN=TRTPN*10+AVISITN ; TRT_VISIT=AVISIT ; &SET_VALUE3 ; _LINE_IND= 1 ; output ;
    AVISITN=3 ; AVISIT='Visit 3' ; TRT_VISITN=TRTPN*10+AVISITN ; TRT_VISIT=AVISIT ; &SET_VALUE3 ; _LINE_IND= 1 ; output ;
    AVISITN=4 ; AVISIT='Visit 4' ; TRT_VISITN=TRTPN*10+AVISITN ; TRT_VISIT=AVISIT ; &SET_VALUE3 ; _LINE_IND=21 ; output ;
  end ;
run ;
/* 前処理 *****************************************************************************************/
options nodate nonumber papersize = "A4" orientation = portrait topmargin = 3cm bottommargin = 3cm leftmargin = 2.5cm rightmargin= 2.5cm ;
ods escapechar = '~';
ods path work.templat (update) sasuser.templat (update) sashelp.tmplmst (read) ;
/* ユーザー定義スタイルテンプレート未使用 *********************************************************/
/* タイトル・フットノートの指定（位置、色、フォント、サイズ、テキスト）*/
title1 '~S={just=left background=white foreground=black font=("Courier New", 8pt)}Table 9.9.9 Subgroup Analysis, MMRM for Percent Change
from Baseline in VAS by Age Group' ;
```

```
title2 '~S={just=left background=white foreground=black font=("Courier New", 8pt)}Full Analysis Set' ;
footnote1 '~S={just=right background=white foreground=black font=("Courier New", 8pt)}Page ~{pageof}' ;

/* RTF出力開始 */
ods rtf file = "有効性の解析帳票.rtf" notoc_data style = Styles.Default ;
  /* REPORTプロシジャ */
  proc report data = DATA missing split = '|' spanrows
    style(report) ={just=left bordercolor=black borderstyle=solid borderwidth=1pt
                    background=white foreground=black frame=hsides rules=groups cellpadding=0 cellspacing=0}
    style(header) ={asis=on           vjust=bottom background=white foreground=black font=('Courier New', 8pt)}
    style(column) ={asis=on           vjust=top    background=white foreground=black font=('Courier New', 8pt)}
    style(lines)  ={asis=on just=left vjust=top    background=white foreground=black font=('Courier New', 8pt)
                    borderbottomstyle=none} ;
    column AGEGR1N AGEGR1X TRT_VISITN TRT_VISIT LSMEAN SE
           (""S={just=center borderbottomwidth=1pt}95% CI~{super[a]}" LOWER UPPER)
           (""S={just=center borderbottomwidth=1pt}Comparison to Placebo" C_LSMEAN C_SE
            (""S={just=center borderbottomwidth=1pt}95% CI~{super[a]}" C_LOWER C_UPPER)) _LINE_IND ;
    define AGEGR1N    / order noprint ;
    define AGEGR1X    / order style(column)={just=left cellwidth=1.6cm} style(header)={just=left} 'Age' ;
    define TRT_VISITN / order noprint ;
    define TRT_VISIT  / style(column)={just=left   cellwidth=2.6cm} style(header)={just=left  } 'Treatment|Visit' ;
    define LSMEAN     / style(column)={just=center cellwidth=1.4cm} style(header)={just=center} 'LS Mean' ;
    define SE         / style(column)={just=center cellwidth=1.4cm} style(header)={just=center} 'SE';
    define LOWER      / style(column)={just=center cellwidth=1.4cm} style(header)={just=center} 'Lower';
    define UPPER      / style(column)={just=center cellwidth=1.4cm} style(header)={just=center} 'Upper';
    define C_LSMEAN   / style(column)={just=center cellwidth=1.4cm} style(header)={just=center} 'LS Mean' ;
    define C_SE       / style(column)={just=center cellwidth=1.4cm} style(header)={just=center} 'SE';
    define C_LOWER    / style(column)={just=center cellwidth=1.4cm} style(header)={just=center} 'Lower';
    define C_UPPER    / style(column)={just=center cellwidth=1.4cm} style(header)={just=center} 'Upper';
    define _LINE_IND      / display noprint ;
    compute _LINE_IND ;
      if _LINE_IND = 10 then call define('AGEGR1X'  , 'style', 'style=[borderbottomwidth=1pt]') ;
      if _LINE_IND =  1 then call define('TRT_VISIT', 'style', 'style=[marginleft=0.4cm]') ;
      if _LINE_IND = 21 then do ;
       call define('TRT_VISIT', 'style', 'style=[borderbottomwidth=1pt marginleft=0.4cm]') ;
       call define('LSMEAN' , 'style', 'style=[borderbottomwidth=1pt]'); call define('SE'    , 'style', 'style=[borderbottomwidth=1pt]') ;
       call define('LOWER'  , 'style', 'style=[borderbottomwidth=1pt]'); call define('UPPER' , 'style', 'style=[borderbottomwidth=1pt]') ;
       call define('C_LSMEAN', 'style', 'style=[borderbottomwidth=1pt]'); call define('C_SE'  , 'style', 'style=[borderbottomwidth=1pt]') ;
       call define('C_LOWER', 'style', 'style=[borderbottomwidth=1pt]'); call define('C_UPPER', 'style', 'style=[borderbottomwidth=1pt]') ;
      end ;
    endcomp;
    compute after _page_ ;
      line '[a] 95% confidence interval' ;
    endcomp ;
  run ;
ods rtf close ;
```

```
/* ユーザー定義スタイルテンプレート使用 ***********************************************************************************/
/* ユーザー定義スタイルテンプレートの作成 */
proc template ;
  define style RTF_ORIGNAL ;  /* ユーザー定義スタイルテンプレートのテンプレート名 */
    parent = Styles.Default ; /* 継承もとのスタイルテンプレート名 */
    style fonts / 'docfont' = ('Courier New', 8pt) 'headingfont' = ('Courier New', 8pt)
                  'fixedfont' = ('Courier New', 8pt) 'batchfixedfont' = ('Courier New', 8pt)
                  'fixedstrongfont' = ('Courier New', 8pt) 'fixedemphasisfont' = ('Courier New', 8pt)
                  'emphasisfont' = ('Courier New', 8pt) 'strongfont' = ('Courier New', 8pt)
                  'titlefont' = ('Courier New', 8pt) 'titlefont2' = ('Courier New', 8pt);
    style color_list / 'fgA'  = black 'fgA1' = black 'fgA2' = black 'fgB1' = black 'fgB2' = black 'fgA4' = black
                       'bgA' = white 'bgA1' = white 'bgA2' = white 'bgA3' = white 'bgA4' = white   ;
    style output from container / borderwidth = 1 bordercolor=black borderstyle=solid frame = hsides rules = group
                                  cellspacing = 0 cellpadding = 0 ;
    end;
run;

/* タイトル・フットノートの指定（色、フォント、サイズはスタイルテンプレートで定義）*/
title1 '~S={just=left}Table 9.9.9 Subgroup Analysis, MMRM for Percent Change from Baseline in VAS by Age Group' ;
title2 '~S={just=left}Full Analysis Set' ;
footnote1 '~S={just=right}Page ~{pageof}' ;

/*RTF出力開始 */
ods rtf file = "有効性の解析帳票_STYLETEMPLATE.rtf" notoc_data
                                                  style = RTF_ORIGNAL ; /* ユーザー定義スタイルテンプレート指定 */
  /* REPORTプロシジャ */
  proc report data = DATA missing split = '|' spanrows
    /* コメントアウトされているスタイル属性はスタイルテンプレートで定義 */
    style(report) ={just=left /*bordercolor=red borderstyle=solid borderwidth=1pt background=white foreground=black*/
                    /*frame=hsides rules=groups cellpadding=0 cellspacing=0*/}
    style(header) ={asis=on vjust=bottom /*background=white foreground=black font=('Courier New', 8pt)*/}
    style(column) ={asis=on vjust=top /*background=white foreground=black font=('Courier New', 8pt)*/}
    style(lines)  ={asis=on just=left vjust=top /*background=white foreground=black font=('Courier New', 8pt)*/ borderbottomstyle=none} ;
    ：省略（COLUMNステートメント以降はユーザー定義スタイルテンプレート未使用バージョンと同様）
```

# EMB アルゴリズムを用いた Multiple Imputation

○矢田 真城[1]　浜田 知久馬[2]
[1]エイツーヘルスケア株式会社 開発戦略本部 生物統計部
[2]東京理科大学 工学部 情報工学科

Multiple Imputation with the EMB algorithm

Shinjo Yada[1] and Chikuma Hamada[2]
[1]*Biostatistics Department, Development Strategy Division, A2 Healthcare Corporation*
[2] *Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science*

## 要旨

多重補完法のひとつとして，MCMC 法に基づく方法が知られている．Honaker and King (2010)は，新しいアプローチとして，EMB (Expectation-Maximization with Bootstrapping)アルゴリズムによる多重補完法を示した．これは，MCMC 法に基づく方法において，事後分布からのモンテカルロ標本により欠測値を補完するプロセスを，ブートストラップ法を用いた EM アルゴリズムによる補完に置き換える方法である．本稿では，解析対象となるデータが全て連続変数である場合に対し，EMB アルゴリズムによる多重補完法を適用し，SAS を用いて実行するための具体的な方法を紹介する．

キーワード：Multiple Imputation，Expectation-Maximization with Bootstrapping，MI プロシジャ，%BOOT

## 1. はじめに

多重補完法 (multiple imputation) は，欠測値を，真の値に対する不確実性を反映させた値で置き換える方法である(Rubin, 1987)．多重補完法は，1)欠測値に対してなんらかの方法で生成した値で補完したデータを $M$ 組用意する($M$>1)，2)$M$ 組のデータセットに対し標準的な方法で解析する，3)得られた $M$ 組の解析結果を 1 つに統合する，という，3 つの手順で構成される．

MCMC (Markov Chain Monte Carlo)法は，欠測データの事後予測分布から生成した標本を用いて欠測値を補完する方法である．観測データを $\mathbf{Y}_{obs}$，欠測データを $\mathbf{Y}_{mis}$，興味のあるパラメータを $\boldsymbol{\theta}$ とし，$t$ 番目の反復過程における $\boldsymbol{\theta}$ の推定値を $\boldsymbol{\theta}^{(t)}$と表すとき，MCMC 法の具体的な手順は以下のようになる．Imputation step（以下 I step と表記）として，$\mathbf{Y}_{obs}$ と $\boldsymbol{\theta}^{(t)}$ 所与のもとでの $\mathbf{Y}_{mis}$ の条件付分布 $p(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\theta}^{(t)})$から欠測値 $\mathbf{Y}_{mis}^{(t)}$ を生成する．Posterior step（以下 P step と表記）として，$\mathbf{Y}_{obs}$ と I step において生成された $\mathbf{Y}_{mis}^{(t)}$ から一時的に構成される完全データ $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis}^{(t)})$に基づき，事後分布 $p(\boldsymbol{\theta}|\mathbf{y}_{obs}, \mathbf{y}_{mis}^{(t)})$から $\boldsymbol{\theta}^{(t+1)}$を生成する．このようにして形成

されるマルコフ連鎖( $\mathbf{Y}_{mis}^{(1)}, \boldsymbol{\theta}^{(1)}$), ( $\mathbf{Y}_{mis}^{(2)}, \boldsymbol{\theta}^{(2)}$),…が収束するまで I step と P step を繰り返し，反復過程の最後に得られた標本を用いて欠測値を補完する．

解析対象となるデータが全て連続変数の場合，それらが多変量正規分布に従うと想定し，興味のあるパラメータである平均ベクトル $\boldsymbol{\mu}$ と分散共分散行列 $\boldsymbol{\Sigma}$ に対して適当な事前分布を仮定したもとで，I step と P step をマルコフ連鎖が収束するまで繰り返すことになる．ただし，実務上の問題点として，取り扱う変数の数が増えるほど $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の事後分布からのサンプリングが困難になることがあげられる．$p$ 個変数がある場合，$\boldsymbol{\mu}$ で $p$ 個，$\boldsymbol{\Sigma}$ で $p(p+1)/2$ 個の合計 $p(p+3)/2$ のパラメータが存在するため，変数の個数が増えればそれだけ計算負荷が急速に増大する．

Honaker and King (2010)は，この問題を解決するための新しいアプローチとして，興味あるパラメータ $\boldsymbol{\theta}$ の事後分布からのサンプリングを，ブートストラップ法を用いた EM アルゴリズムで置き換えるという方法を提案した．以下，Honaker et al. (2011) に従い，これを EMB (expectation-maximization with bootstrapping) アルゴリズムと表記する．EMB アルゴリズムでは，欠測データを含むデータ($\mathbf{Y}_{obs}, \mathbf{Y}_{mis}$)からブートストラップ法を用いて $M$ 組のデータを生成し，生成された $M$ 組のデータごとに EM アルゴリズムにより $\boldsymbol{\theta}$ を推定し，推定された $\boldsymbol{\theta}$ と $\mathbf{Y}_{obs}$ から欠測値を補完する．Honaker and King (2010)は，この方法の利点として，扱えるデータの大きさは解析を行う PC のメモリーサイズのみに依存していることを挙げ，時系列横断データのような巨大なデータを扱う社会科学の分野にて適用することを推奨した．

本稿では，解析対象となるデータが全て連続変数である場合をとりあげ，EMB アルゴリズムによる多重補完法について示す．そして，SAS を用いて行うための方法を，具体的に説明する．

## 2. EMB アルゴリズムによる多重補完法とは

EMB アルゴリズムによる多重補完法は，補完回数を $M(>1)$として，以下の 4 つのステップから構成される．

Step1：欠測値を含んだ不完全データセット $D^{orign}$ から，$M$ 組のブートストラップ標本 $D_i^{resap}$ $(i = 1,2,\ldots,M)$を生成する．

Step2：$D_i^{resap}$ $(i = 1,2,\ldots,M)$に対し，EM アルゴリズムを用いて組ごとに $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を推定する．

Step3：組ごとに推定された $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を用いて，$D_i^{resap}$ の欠測値を補完したデータセット $D_i^{imput}$ を生成する．

Step4：$D_i^{imput}$ $(i = 1,2,\ldots,M)$に対し解析を実行し，これらの解析結果を統合する．

EMB アルゴリズムを用いた多重補完法では，まず，欠測値を含んだ不完全データセットから，ブートストラップ法を用いて，$M$ 組のブートストラップ標本を生成する．ブートストラップ法とは，観測されたデータからリサンプリングすることにより生成された擬似的なデータを用いて，パラメータ推定量の統計的誤差や統計量の分布を推定する方法のことである．$n$ 個のデータが観測されたとき，大きさ $n$ のブートストラップ標本とは，(0,1]上で発生させた $n$ 個の一様乱数 $u_1,u_2,\ldots,u_n$ を用いて，観測されたデータから，$[nu_1]+1$ 番目，$[nu_2]+1$ 番目,…, $[nu_n]+1$ 番目にあたるデータを抽出することで構成される（小西, 2008）．ここに[ ]はガウス記号である．生成されたブートストラップ標本を用いて，観測されたデータが従う確率分布に関するパラメータを推定し，ブートストラップ推定値を算出する．ブートストラップ推定値は，リサンプリング回数（ブートストラップ反復回数）を $M$ とするとき，大きさ $n$ のブートストラップ標本を $M$ 組生成させ，各ブートストラップ標本に基づき，モンテカルロ法によって数値的に近似することができる．

リサンプリング回数 $M$ が有限である以上，ブートストラップ推定値には数値近似による誤差が含まれる．釣り合い型ブートストラップは，数値近似による誤差を抑えるための方法のひとつである．一様リサンプリングを用いたブートストラップ法に比べ，同一のリサンプリング回数で推定量の分散を減少させることができる（汪・桜井, 2011）．一様リサンプリングでは，$n$ 個の観測データ $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ から大きさ $n$ のブートストラップ標本を生成する作業を，独立に $M$ 回繰り返す．このため，生成された $M \times n$ 個のブートストラップ標本において，観測データ $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ が全て $M$ 回ずつ登場するとは限らない．これに対して，釣り合い型ブートストラップ法では，$M \times n$ 個のデータにおいて $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ がいずれも $M$ 回ずつ登場するように，ブートストラップ標本が生成される．Step1 では，欠測値を含んだ $n$ 個のデータから，リサンプリング回数を $M$ として大きさ $n$ のブートストラップ標本を生成するのみであり，ブートストラップ推定値の算出は不要である．

Step2 では，Step1 で生成された $M$ 組のデータセットに対し，EM アルゴリズムを用いて興味のあるパラメータ $\boldsymbol{\theta}$ の推定値を算出する．EM アルゴリズムは，Expectation Step（以下 E Step と表記）と Maximization Step（以下 M Step と表記）の 2 つのステップを繰り返すことで，$\boldsymbol{\theta}$ の最尤推定量を導くためのアルゴリズムである．具体的には，$k$ 回目の反復過程で得られた $\boldsymbol{\theta}$ の推定値を $\boldsymbol{\theta}^{(k)}$ と表すとき，E step では，観測データ $\mathbf{Y}_{obs}$ と $\boldsymbol{\theta}^{(k)}$ が所与のもとで，完全データ $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ の対数尤度の条件付期待値 $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(k+1)})$ を計算する．続く M step では，E step にて求めた $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(k+1)})$ を最大化する $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k+1)}$ を求める．$\boldsymbol{\theta}^{(k+1)}$ を更新後のパラメータとして，$(k+1)$ 回目の E step を行う．このようにして，E step と M step を反復過程の収束条件が満たされるまで繰り返し，最後に得られたパラメータ値を，$\boldsymbol{\theta}$ の最尤推定値とする．

解析対象となるデータが全て連続変数の場合，これらが従う確率密度関数として多変量正規分布を想定すると，興味のあるパラメータ $\boldsymbol{\theta}$ は平均ベクトル $\boldsymbol{\mu}$ と分散共分散行列 $\boldsymbol{\Sigma}$ である．多変量正規分布は指数分布族であるため，E step は，観測データが与えられたもとでの十分統計量 $t(\mathbf{Y})$ の条件付き期待値 $E_{\theta^{(k)}}[t(\mathbf{Y}) | \mathbf{Y}_{obs}]$ を求めるステップと捉えることができる．また，M step は，方程式 $E_{\theta}[t(\mathbf{Y})] = E_{\theta^{(k)}}[t(\mathbf{Y}) | \mathbf{Y}_{obs}]$ を最大化する $\boldsymbol{\theta}$ を求めるステップとみなすことができる（小西, 2008）．欠測値を含む多変量正規確率変数に対する EM アルゴリズムの適用例は，渡辺（2008）に簡潔にまとめられている．

Step3 では，Step2 にて組ごとに推定された平均ベクトル $\boldsymbol{\mu}$ と分散共分散行列 $\boldsymbol{\Sigma}$ の推定値を用いて欠測値を補完する．$p$ 個の変数からなる $n$ 個のデータに対し，$i$ 番目のデータ $x_{i1}, x_{i2}, \ldots, x_{ip}$ を成分とする列ベクトル $\mathbf{x}_i$ $(p \times 1)$ が欠測に応じて以下のように分解されたとする．

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_i^0 \\ \mathbf{x}_i^1 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{i,mis} \\ \mathbf{x}_{i,obs} \end{bmatrix} \tag{1}$$

(1) の欠測パターンに応じて $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_0 \\ \boldsymbol{\mu}_1 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{01} \\ \boldsymbol{\Sigma}_{10} & \boldsymbol{\Sigma}_{11} \end{bmatrix} \tag{2}$$

と分解する（渡辺, 2008）．Step2 にて求められた $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値 $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ を

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \tilde{\boldsymbol{\mu}}_0 \\ \tilde{\boldsymbol{\mu}}_1 \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{00} & \tilde{\boldsymbol{\Sigma}}_{01} \\ \tilde{\boldsymbol{\Sigma}}_{10} & \tilde{\boldsymbol{\Sigma}}_{11} \end{bmatrix} \tag{3}$$

と表すとき，$\mathbf{x}_i$ にて欠測となっている $\mathbf{x}_{i,mis}$ の補完値 $\tilde{\mathbf{x}}_{i,mis}$ を

$$\tilde{\mathbf{x}}_{i,mis} = \tilde{\boldsymbol{\mu}}_0 + \tilde{\boldsymbol{\Sigma}}_{01}\tilde{\boldsymbol{\Sigma}}_{11}^{-1}(\mathbf{x}_{i,obs} - \tilde{\boldsymbol{\mu}}_1) + \tilde{\boldsymbol{\varepsilon}}_i \tag{4}$$

により算出する($i$=1,2,…,$n$). $p$ 個の変数に多変量正規分布を想定し，欠測である変数を目的変数，非欠測の変数を説明変数として線型回帰モデルをあてはめ，「予測値」として欠測値を補完する．線型回帰モデルにおいて，誤差項が互いに独立で平均 0，均一分散をもつ正規分布に従うとき，回帰係数の最尤推定値は，解析対象となるデータの平均値及び分散・共分散を用いて求めることができる（竹内, 2011）．この最尤解に Step2 にて求められた $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ を与えることで，(4)が導かれる．

(4)において，$\tilde{\boldsymbol{\varepsilon}}_i$ は根本的な不確実性（fundamental uncertainty）である（Honaker and King,2010;高橋・伊藤, 2013）．例えば，$\mathbf{x}_1$ では $x_{11}$ のみが欠測，$\mathbf{x}_2$ では $x_{21}$ のみが欠測であり，$\mathbf{x}_1$ と $\mathbf{x}_2$ では欠測以外全て同値である場合，(4)にて $\tilde{\boldsymbol{\varepsilon}}_i$ がなければ補完された値 $\tilde{x}_{11}$ と $\tilde{x}_{21}$ は確実に同値となる．このように，欠測以外の変数が全て同一であっても欠測である変数の補完値は不確実であることを考慮する必要がある．本稿では，Step2 にて求めた分散の推定値と標準正規分布に従う乱数をかけあわせた値を用いることとした．

以下，表 1 に示した不完全データを用いて，これら 4 つのステップについて説明する．表 1 のデータは，SAS/STAT(R) 14.1 User's Guide, The MI Procedure にて用いられているデータセット Fitness1 から作成したものである．データセット Fitness1 は，SAS/STAT(R) 14.1 User's Guide, The REG Procedure の Example 97.2 にあるデータセット Fitness をもとに作成されており，格納されている 3 つの変数 Oxygen，RunTime，RunPulse にはいずれも欠測値がある．Fitness1 は 31 個のデータがあったが，ここでは説明を簡単にするため，10 個のデータに絞った．表 1 で「NA」が欠測値を意味する．

表 1：Aerobic Fitness のデータ

| ID No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Oxygen | 44.609 | 45.313 | 54.297 | NA | 39.442 | 60.055 | 50.541 | 37.388 | 44.754 | 51.855 |
| RunTime | 11.37 | 10.07 | 8.65 | 11.95 | 13.08 | 8.63 | NA | 8.63 | 11.12 | 10.33 |
| RunPulse | 178 | 185 | 156 | 176 | 174 | 170 | NA | 170 | 176 | 166 |

Step1 では，欠測値を含んだ 10 個の不完全データに対し，大きさ 10 のブートストラップ標本を $M$ 組生成する．生成された $M$ 組のデータセットには，欠測データが含まれず補完する必要のない組も考えらえるが，欠測データが含まれている組が存在する可能性が高い．

Step2 では，Step1 で生成された $M$ 組のデータセットに対し，EM アルゴリズムを用いて $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を推定する．表 1 の Aerobic Fitness のデータにある Oxygen，RunTime，RunPulse の 3 変数に対する欠測値の補完を考えており，$M$ 組のデータセットごとに $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を推定するため，Step2 終了時には，3×1 の列ベクトル $\boldsymbol{\mu}$ と 3×3 の行列 $\boldsymbol{\Sigma}$ の推定値が $M$ 組算出される．

Step3 では，組ごとに Step2 で算出された $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を用い，組ごとに，欠測値を補完した 10 個のデータセットを生成することになる．例えば，Step2 で生成された $M$ 組のデータセットのうち，1 組目では，表 2 のように，10 個のデータのうち欠測データが，2 番目の Oxygen ($X_1$)，5 番目の RunTime ($X_2$)及び RunPulse ($X_3$) であったとする．

表 2：リサンプリングされた Aerobic Fitness のデータ

| ID No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Oxygen | 45.313 | NA | 37.388 | 45.313 | 50.541 | 44.754 | 45.313 | 44.609 | 45.313 | 54.297 |
| RunTime | 10.07 | 11.95 | 8.63 | 10.07 | NA | 11.12 | 10.07 | 11.37 | 10.07 | 8.65 |
| RunPulse | 185 | 176 | 170 | 185 | NA | 176 | 185 | 178 | 185 | 156 |

1 組目のデータに対して EM アルゴリズムを用いて得られた $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を

$$\tilde{\boldsymbol{\mu}} = \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{pmatrix}, \quad \tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \tilde{\sigma}_{11} & \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \\ \tilde{\sigma}_{21} & \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{31} & \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{pmatrix}$$

と表す．2 番目のデータでは欠測は Oxygen($X_1$)のみであるため，(3)に従い欠測パターンに応じて $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ を分解すると

$$\tilde{\boldsymbol{\mu}}_0 = \tilde{\mu}_1, \quad \tilde{\boldsymbol{\mu}}_1 = \begin{bmatrix} \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}}_{00} = \tilde{\sigma}_{11}, \quad \tilde{\boldsymbol{\Sigma}}_{01} = \begin{bmatrix} \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}}_{10} = \begin{bmatrix} \tilde{\sigma}_{21} \\ \tilde{\sigma}_{31} \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}}_{11} = \begin{bmatrix} \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{bmatrix}$$

となり，2 番目のデータでの欠測 $\mathbf{x}_{2,\mathrm{mis}}$ の補完値 $\tilde{\mathbf{x}}_{2,\mathrm{mis}}$ を

$$\tilde{\mathbf{x}}_{2,\mathrm{mis}} = \tilde{x}_{2,1} = \tilde{\mu}_1 + \begin{bmatrix} \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \end{bmatrix} \begin{bmatrix} \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{bmatrix}^{-1} \begin{bmatrix} x_{2,2} - \tilde{\mu}_2 \\ x_{2,3} - \tilde{\mu}_3 \end{bmatrix} + \tilde{\varepsilon}_{9,1}$$

$$= \tilde{\mu}_1 + \begin{bmatrix} \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \end{bmatrix} \begin{bmatrix} \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{bmatrix}^{-1} \begin{bmatrix} x_{9,2} - \tilde{\mu}_2 \\ x_{9,3} - \tilde{\mu}_3 \end{bmatrix} + \sqrt{\tilde{\sigma}_{11}} z_1$$

により算出する．ここに，$z_1$ は標準正規乱数である．5 番目のデータでは，欠測は RunTime ($X_2$)と RunPulse ($X_3$)であるから，欠測パターンに応じて $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ を分解すると

$$\tilde{\boldsymbol{\mu}}_0 = \begin{bmatrix} \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{bmatrix}, \quad \tilde{\boldsymbol{\mu}}_1 = \tilde{\mu}_1, \quad \tilde{\boldsymbol{\Sigma}}_{00} = \begin{bmatrix} \tilde{\sigma}_{22} & \tilde{\sigma}_{23} \\ \tilde{\sigma}_{32} & \tilde{\sigma}_{33} \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}}_{01} = \begin{bmatrix} \tilde{\sigma}_{21} \\ \tilde{\sigma}_{31} \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}}_{10} = \begin{bmatrix} \tilde{\sigma}_{12} & \tilde{\sigma}_{13} \end{bmatrix}, \quad \tilde{\boldsymbol{\Sigma}}_{11} = \tilde{\sigma}_{11}$$

となる．よって，5 番目のデータでの欠測 $\mathbf{x}_{5,\mathrm{mis}}$ の補完値 $\tilde{\mathbf{x}}_{5,\mathrm{mis}}$ を

$$\tilde{\mathbf{x}}_{5,\mathrm{mis}} = \begin{bmatrix} \tilde{x}_{5,2} \\ \tilde{x}_{5,3} \end{bmatrix} = \begin{bmatrix} \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{bmatrix} + \begin{bmatrix} \tilde{\sigma}_{21} \\ \tilde{\sigma}_{31} \end{bmatrix} \sigma_{11}^{-1} (x_{5,1} - \tilde{\mu}_1) + \begin{bmatrix} \tilde{\varepsilon}_{5,2} \\ \tilde{\varepsilon}_{5,3} \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{\mu}_2 \\ \tilde{\mu}_3 \end{bmatrix} + \begin{bmatrix} \tilde{\sigma}_{21} \\ \tilde{\sigma}_{31} \end{bmatrix} \sigma_{11}^{-1} (x_{5,1} - \tilde{\mu}_1) + \begin{bmatrix} \sqrt{\tilde{\sigma}_{22}} z_2 \\ \sqrt{\tilde{\sigma}_{33}} z_3 \end{bmatrix}$$

により算出する．ここに，$z_2, z_3$ は独立な標準正規乱数である．

このようにして欠測値が補完された $M$ 組のデータセットに対し，個別に解析を実行する．パラメータ推定値と標準誤差を算出し，最終的な解析結果としてそれらをひとつに統合する．

## 3. SAS による実装例

EMB アルゴリズムによる多重補完法は，2.に示したとおり，1)欠測値を含む不完全データからブートストラップ標本を $M$ 組生成する，2) $M$ 組のブートストラップ標本に対し EM アルゴリズムを用いて $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を求める，3) $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を用いて欠測値に対する補完を行う，4)補完されたデータセットごとに解析し得られた推定量を 1 つに統合する，という 4 つのステップから構成される．

ブートストラップ法に対応しているプロシジャは存在しないが，ブートストラップ法に対応するサンプルマクロ%BOOT が公開されており，SAS Institute Inc.の Technical Support よりダウンロードが可能である（http://support.sas.com/kb/22/220.html）．%BOOT を使用するにあたっては，まず，サンプルマクロダウンロード先にある%analyze を読みこんでおく必要がある．その上でブートストラップ法を行いたいデータセット名をマクロ引数 data =により指定すれば，BOOTDATA という名前でリサンプリングされたデータセットができあがる．デフォルトではリサンプリングの回数が 200 であるため，例えば，リサンプリングを 20 回行ないたい場合には samples = 20 と指定する．%BOOT では，特に指定しない限り，釣り合い型ブートストラップ法を行う（表 1）．

表 1: %BOOT のマクロ引数

| |
|---|
| data = 入力データセット名を指定． |
| samples = リサンプリング回数（ブートストラップ反復回数）を指定．デフォルトは200． |
| random = 擬似乱数を生成するときのシードを指定． |
| size= ブートストラップ標本の大きさを指定．デフォルトはdata =で指定される入力データセットのオブザベーション数として処理される． |
| balanced = デフォルトではsize =を指定しない限り1: 釣り合い型ブートストラップが適用される．size =を指定すると0: 一様リサンプリングが適用される． |
| biascorr = デフォルトは1: ブートストラップ偏り修正済み推定量 |
| alpha = ブートストラップ信頼区間の信頼係数を指定．デフォルトは0.05（95%信頼区間）． |
| print = デフォルトは1: ブートストラップ推定値を出力，出力しない場合には0と指定． |
| chart = デフォルトは1: ブートストラップ分布を出力，出力しない場合には0と指定． |

SAS プログラム 1 は，%BOOT を用いてブートストラップ法を行うプログラムの一例である．データセット Fitness1 に格納されている 3 つの変数 Oxygen，RunTime，RunPulse に対しブートストラップ法を行いたい場合，%analyze の MEANS プロシジャにてこれらの変数を VAR ステートメントで指定し，OUTPUT ステートメントで指定する変数もそれにあわせて更新しておく必要がある．Fitness1 には欠測データを含めて 31 個のデータがある．SAS プログラム 1 では，%BOOT において sample = 20 と指定し，size = は特に指定していないため，データセット BOOTDATA には，釣り合い型ブートストラップ法によりリサンプリングされた 31 個のデータが 20 組セットされる．変数_OBS_は，リサンプリングされたデータが Fitness1 のどのデータから得られたものかを表す．例えば_OBS_ = 24 ならそれは Fitness1 の 24 番目のデータに該当することを意味する．_OBS_の頻度集計を行うことで，リサンプリングされた 31×20 = 620 個のブートストラップ標本において，Fitness1 の 31 個のデータがそれぞれ 20 回ずつ抽出されていることが確認できる．

SAS プログラム 1: %BOOT によるブートストラップ法の実行

```
%macro analyze(data = , out = ) ;
    proc means noprint data = &data vardef = n ;
        output out = &out(drop = _freq_ _type_) var = Oxygen RunTime RunPulse ;
        var Oxygen RunTime RunPulse ;
        %bystmt ;
    run ;
%mend ;
%boot(data = Fitness1, samples = 20, random = 170723, print = 0, chart = 0);
```

EM アルゴリズムによる $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値の算出には，MI プロシジャを用いる．

MI プロシジャは，欠測値を補完した完全データを複数組作成するためのプロシジャである．どのような方法で欠測値を補完するかは，欠測パターンと補完したいデータのタイプに依存する．補完したいデータが連続量で欠測パターンが単調であれば，多変量正規性を仮定するパラメトリック回帰，あるいは傾向スコアを用いたノンパラメトリック法のいずれかが適用される．補完したいデータが 2 値で欠測パターンが単調であればロジスティック回帰が適用される．補完したいデータが連続量で任意の欠測パターンのときには，多変量正規性を仮定する MCMC 法が適用される．また Ver. 9. 3から，Fully Conditional Specification による多重補完も実行可能となった（SAS Institute Inc. ,2011 ; 多田, 2013）．これは，欠測パターンや補完したいデータのタイプによらず適用できる．

このように，MI プロシジャでは，様々なアルゴリズムを用いて欠測値を補完することができるが，EM ステートメントを用いることにより，多重補完は行わず EM アルゴリズムにより $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定を行うことも可能である．そこでこの機能を用いて，EMB アルゴリズムによる多重補完法を行う際の 2 つ目のステップ（$M$ 組のブートストラップ標本に対し EM アルゴリズムを用いて $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を求める）を実装する．

SAS プログラム 2 はその一例である. MI プロシジャにて, EM ステートメントにてオプション OUTEM = を指定することにより，EM アルゴリズムを用いた平均値，分散，共分散の推定値を保存することが可能となる．前述のとおり，多重補完なしで EM アルゴリズムを行いたいため，オプション nimpute = 0 と指定した．SAS プログラム 3.1 で作成されたデータセット BOOTDATA では，リサンプリングされた 31 個のデータが 20 組存在している．SAS プログラム 2 では，20 組のデータがデータセット BOOTDATA の変数_sample_により識別できることを利用して，20 組のデータそれぞれに対し $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の推定値を算出させている．MI プロシジャでは，$\boldsymbol{\mu}$ の初期値として入力データセットの変数ごとに算出された平均値を用い，$\boldsymbol{\Sigma}$ の初期値は，非対角成分を全て 0，対角成分を入力データセットの変数ごとに算出された分散として，E-step と M-step の反復計算が行われる．

SAS プログラム 2: 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 mi data = m_wrk seed = &seed. nimpute = 0 noprint ;
          em outem = m_outem out = m_outmi initial = ac ;
          var Oxygen RunTime RunPulse ;
        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 _sample_ ^= . ;run ;
        data m_wkst2 ;set m_wkst2 m_outmi ;if _sample_ ^= . ;run ;
    %end ;
    data &outem. ;set m_wkst1 ;run ; data &outmi. ;set m_wkst2 ;run ;
    proc datasets library = work nolist nowarn ;
      delete m_outem m_outmi m_wrk m_wkst1 m_wkst2 ;
    run ;quit ;
%mend m_miem ;
%m_miem(inds = BOOTDATA, samples = 20, seed = 170723, outds = outem, outmi = outmi) ;
```

MI プロシジャの EM ステートメントにてオプション OUT =を指定すると，欠測値が補完されたデータセットが生成され，EM アルゴリズムに基づき算出された期待値が出力される（SAS Institute Inc., 2015）．ただし，これは $\mathbf{x}_{i,mis}$ の補完値 $\tilde{\mathbf{x}}_{i,mis}$ を

$$\tilde{\mathbf{x}}_{i,mis} = \tilde{\boldsymbol{\mu}}_0 + \tilde{\boldsymbol{\Sigma}}_{01}\tilde{\boldsymbol{\Sigma}}_{11}^{-1}(\mathbf{x}_{i,obs} - \tilde{\boldsymbol{\mu}}_1) \tag{5}$$

により算出した場合に相当する．従って，(4)により欠測値を補完するためには，$\tilde{\boldsymbol{\varepsilon}}_i$ を加えるための SAS プログラムが別途必要となる．SAS プログラム 3 としてその一例を示す．

SAS プログラム 3: EMB アルゴリズムによる補完

```
data m_outmi ;set OUTMI ;keep _sample_ Oxygen RunTime RunPulse ;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-M3 ;
  array aa M1-M3 ; array bb Oxygen RunTime RunPulse ;
  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_ = "Oxygen" then do ;varno = 1 ;var = Oxygen ;end ;
  else if _NAME_ = "RunTime" then do ;varno = 2 ;var = RunTime ;end ;
```

```
  else if _NAME_ = "RunPulse" then do ;varno = 3 ;var = RunPulse ;end ;
run ;
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_outmi ;
  merge m_outmi m_ind ;
  by _sample_ rep ;
  array aa Oxygen RunTime RunPulse ;
  array bb I_Oxygen I_RunTime I_RunPulse ;
  array cc M1-M3 ;
  array dd var1-var3 ;
  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 ;
  keep _sample_ I_Oxygen I_RunTime I_RunPulse ;
run ;
```

SAS プログラム 2 まで実行した時点で，既にデータセット OUTMI に欠測値に対し(5)により補完された値が代入されている．このため，あとはデータセット OUTEM に保存されている $\tilde{\boldsymbol{\Sigma}}$ と標準正規乱数から $\tilde{\varepsilon}_i$ を算出し，データセット OUTMI に出力されている補完値に加えればよい．SAS プログラム 3 では，データセット BOOTDATA から，欠測なら 0，非欠測なら 1 をとる欠測識別変数をもたせたデータセット M_IND を用意し，この処理を行っている．

補完されたデータセットごとに解析し得られた解析結果を 1 つに統合するステップは，SAS プログラム 4 のとおり，UNIVARIATE プロシジャと MIANALYZE プロシジャを使って簡単に実行できる（SAS Institute Inc., 2015; 石田・斉藤,2014）．

SAS プログラム 4: パラメータ推定値の出力

```
proc univariate data = OUTEMB noprint ;
  var I_Oxygen I_RunTime I_RunPulse ;
  output out = out mean = Oxygen RunTime RunPulse stderr = SOxygen SRunTime SRunPulse ;
  by _sample_ ;
run ;
proc mianalyze data = out edf = 30 ;
  modeleffects Oxygen RunTime RunPulse ;
  stderr SOxygen SRunTime SRunPulse ;
run ;
```

## 4. おわりに

EMB アルゴリズムによる多重補完法を把握し，4 つのステップに分割した．分割したステップに対応するプロシジャあるいは SAS マクロを探し，SAS プログラム上でどのような処理をしているかを確認した．その上で，対応するプロシジャあるいは SAS マクロと分割したステップとを比較し，対応するプロシジャで必要なオプションを指定，足りない部分を実行するためのプログラムを用意した．以上により，%BOOT，MI プロシジャ，DATA ステップによる加工プログラムを連動させることで，EMB アルゴリズムによる欠測値補完の実装が可能となった．

代替案として SAS と R とを連動させる方法がある．R は，フリーで配布されている統計分析ソフトウェアであり，AmeliaII パッケージを用いて，EMB アルゴリズムによる多重補完法を行うことができる．そこで，SAS で用意した欠測値を含むデータセットを R にわたし，R に EMB アルゴリズムによる欠測値補完を行わせ，欠測値が補完されたデータセットを SAS に返し，SAS を用いて解析結果を統合することが考えられる．SAS 上で R を動かして解析結果を得る方法については，船尾(2009)が詳しく紹介している．

本稿では，多重補完法のひとつとして，EMB アルゴリズムによる多重補完法をとりあげた．この方法は，既存のプロシジャでは実行できない．しかし，対応するプロシジャがなくとも，SAS Institute Inc.より公開されている SAS マクロや関連するプロシジャを連動させることにより，簡単に実装することができた．

## 参考文献


[1] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

[2] Honaker, J., King, G. (2010). What to do About Missing Values in Time Series Cross-Section Data. *American Journal of Political Science*. **54**: 561–581.

[3] Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*. **45**:1–47.

[4]小西貞則，越智義道，大森裕浩 (2008). 計算統計学の方法 —ブートストラップ・EM アルゴリズム・MCMC. 朝倉書店.

[5]汪金芳，桜井裕仁 (2011). ブートストラップ入門. 共立出版.

[6]渡辺美智子 (2008). EM アルゴリズム. 21 世紀の統計科学. 日本統計学会 HP 版. http://park.itc.u-tokyo.ac.jp/atstat/jss75shunen/Vol3.pdf （閲覧日：2017 年 4 月 21 日).

[7]竹内啓（監修），市川伸一，大橋靖雄，岸本淳司，浜田知久馬,下川元継,田中佐智子（著）(2011). SAS によるデータ解析入門 第 3 版. 東京大学出版会.

[8]高橋将宜，伊藤孝之 (2013). 経済調査における売上高の欠測値補定方法について～多重代入法による精度の評価.～統計研究彙報 **70**(2):19–86.

[9]高橋将宜，伊藤孝之 (2014). 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～. 統計研究彙報 **71**(3):39–82.

[10]SAS Institute Inc. (2015). *SAS/STAT(R) 14.1 User's Guide*, Cary, NC, USA: SAS Institute Inc.

[11]SAS Institute Inc. (2011). *SAS/STAT(R) 9.3 User's Guide*, Cary, NC, USA: SAS Institute Inc.

[12]多田圭佑 (2013). MI Procedure による多重代入 SAS ver 9.3 における新機能の紹介. SAS ユーザー総会 論文集, 23-44.

[13]石田和也,斉藤和弘 (2014). PROC MIANALYZE を用いた、多重代入法による結果の統合. SAS ユーザー総会 論文集, 627-640.

[14]舟尾暢男 (2009). SAS でベイズ推定を行う方法 -proc MCMC vs R&WinBUGS-. SAS ユーザー総会論文集, 95-114.

## 連絡先

E-mail: *yada-s@a2healthcare.com*

# Proc LuaによるSASマクロに依存しないプログラミングの提案

森岡　裕

イーピーエス株式会社

Proc Lua Supersedes SAS Macro

Yutaka Morioka

Statistics Analysis Department1, EPS Corporation

## 要旨

SAS9.4M3 より追加された Proc Lua は、SAS の中で Lua 言語のコードを実行することが可能なプロシジャである。Lua 言語は、他言語に組み込まれることを前提として設計されているスクリプト言語であり、簡素な記述に柔軟な機能、高速な処理を特徴としている。

本稿では、Proc Lua の基本的な説明を行う。そして、従来 SAS マクロや call execute で行っていた処理について Proc Lua で実装することを提案する。SAS マクロや call execute の利用を必要最小限に限定することでコードの記述が簡潔かつ管理しやすいものとなり、他言語のプログラマや初心者にも理解しやすいプログラムが実現できる。

## 1. 具体的な使用例と既存機能との比較

Proc Lua について本稿では詳細の前に、具体的な使用例から説明する。

【例 1】データセットの存在による処理の分岐

```
/*CODE 1 */
%macro test;
%if %sysfunc(exist(TEST)) eq 1 %then %do;
  %put NOTE:WORK.TEST は既に存在します;
 %end;
 %else %do;
  %put NOTE: WORK.TEST を作成します;
   data TEST;
    set sashelp.CLASS;
   run;
  %end;
%mend test;
%test

%test
```

```
/*CODE 2 */
proc Lua;
 submit;
  if sas.exists("TEST") then
   print("NOTE: WORK.TEST は既に存在します")
  else
   print("NOTE: WORK.TEST を作成します")
   sas.submit([[data TEST;
                 set sashelp.CLASS ;
               run;]])
  end
 endsubmit;
quit;
```

指定した名前のデータセットが存在するかをチェックし、存在しない場合のみデータセットを作成するための SAS コードを実行するという処理について、SAS マクロで記述したのが CODE1、Proc Lua で記述したのが CODE2 となる。この例の場合、CODE1、CODE2 共に実行したい処理の記述部分は「data TEST;set sashelp.CLASS;run;」であり、それ以外は全て制御のためのコードとなる。

SAS マクロで記述した CODE1 の場合、制御部分(マクロ部分)は%が接頭についた %if、%else などを用い通常コード部分と区別する。また、いったん %macro test　～ %mend の部分でマクロを作成した後、%test で実行する必要

がある。例え、汎用的に使いまわす必要がない一回きりの処理であっても、一度 SAS マクロを作成してからそれを実行する必要がある。それは SAS マクロのデメリットのひとつといえる。

対して CODE2 の場合、SAS の実行コードは sas.submit という命令文の括弧の中に記述することで制御文と区別される。Lua でも汎用的な処理を関数化して使いまわすことができるが、そうでない場合は敢えて関数を定義してから実行するというステップを踏む必要がないため、コードが簡素になる。

【例 2】文字型データのループ

```
/*CODE 3 */
%let   DS1 = CLASS;
%let   DS2 = FISH;
%let   DS3 = IRIS;

%macro test;
 %do i = 1 %to 3;
  proc print data = SASHELP.&&DS&i noobs;
  run;
 %end;
%mend;

%test;
```

```
/*CODE 4 */
proc lua;
submit;
 local t = {"CLASS",
            "FISH",
            "IRIS"}
 for key, value in pairs(t) do
  sas.submit[[proc print data =SASHELP.@value@ noobs;
              run;
             ]]
  end
 endsubmit;
quit;
```

「CLASS」、「FISH」、「IRIS」という文字列を順番に処理部に渡すというループ処理について、SAS マクロは文字列をそのままループ値に指定できないため、工夫が必要になる。Proc Lua では、一時的にテーブルを作って文字列を格納し、テーブルの値を順に取り出して処理部に渡すことで実現ができる。

【例 3】データセットの内容から実行コードを生成する

| ITEM |
| --- |
| SEX |
| AGE |
| WEIGHT |

```
/*CODE 5 */
data _NULL_;
 set DS_1 end=eof;
 if _N_=1 then
  call execute("proc print data=
                SASHELP.CLASS; var");
 else do;
  call execute(ITEM || " ");
 end;
 if eof then call execute( ";run;");
run;
```

```
/*CODE 6*/
proc lua;
submit;
 sas.submit_([[proc print data=
                SASHELP.CLASS; var]])
 local dsid = sas.open("DS_1")
 while sas.next(dsid) do
  sas.submit_( sas.get_value(dsid,"ITEM"))
 end
 sas.close(dsid)
  sas.submit([[;run;]])
endsubmit;
run;
```

データセットが存在し、そこに格納されているデータからコードを生成して実行する場合、既存の方法では CODE5 のように call execute を利用していたが、同様の機能を CODE6 のように Proc Lua でも実現可能となった。

SAS では文字列はシングルまたはダブルのコーテーション、「' 」か「" 」で括って表現される。call execute で実行したいコード自体にコーテーションが含まれている場合などがあり、SAS マクロはシングルコーテーションの中では展開されないという仕様も加わり、文字列生成部分のクォート処理が複雑になりがちである。

Proc Lua の場合は SAS マクロのような独特な規則がないため、特に意識せずにコーティングが可能となる。

【例 4】SAS マクロの仕様に依存したコーディング

```
/*CODE 7 */
%macro test;
 %let mvar=0;
 %let mvar = A%eval(&mvar + 10);
 %if &mvar ne %then %do;
 %let mvar=%sysfunc(reverse(&mvar));
 %put mvar= &mvar;
%end;
%mend;

%test
```

```
/*CODE 8 */
proc lua;
submit;
 local lvar=0
 lvar = lvar+10
 if lvar ~= "" then
          lvar =string.reverse(lvar)
 end
 print(lvar)
endsubmit;
quit;
```

SAS マクロはテキスト展開の機能であるため、特別な規則が多く存在する。例えば、マクロ変数を整数値として計算する場合には%eval 関数が必要であり、特殊な文字のクォートには%str をはじめとした専用関数が必要である。

また SAS マクロステートメントの中で SAS 関数を使用する場合は%sysfunc 関数が必要となる。

SAS マクロを使ったプログラミングを行う場合は、常にそういった点に注意せねばならず、初心者にとっては大きなハードルとなる。

一方 Lua 言語にはそれほど特殊な規則がないため、処理をより自然に記述することができる

## 2. Proc Lua 及び Lua 言語の文法

### 2-1. コードとコメントの記述

Proc Lua は基本的に CODE9 のように proc lua;～quit;の間に記述する。

submit;～endsubmit;の間に Lua 言語でコーディングする。ただし外部ファイルとして既に存在する Lua 言語のファイルを読み込んで実行する場合は不要である。

```
/*CODE 9 */
proc lua <オプション>;
submit;

--Lua 言語による記述--

endsubmit;
quit;
```

```
/*CODE 10 */
proc lua ;
submit;
--1 行コメント
--[[複数行
--コメント]]
print("Hello World")
endsubmit;
quit;
```

CODE10 はコメントのつけ方の説明と、「Hello World」とログウインドウに出力するコードとなる。Lua ではステートメント区切りは改行になるため、末尾に「;」は必要ないが、つけることで明示的に区切ることができ、マルチステートメントも可能となる。文字列は「'」または「"」で括って指定する。また「 [[ 」で括ると改行やエスケープ文字「/ で改行など」を含む文字列もそのままの形で格納できる。

submit;～endsubmit;の間には Lua 言語の規則に反するものは記述できないため、通常の SAS コメントステートメント「/* */」などを入れるとエラーになるため注意が必要である。Print は Lua の標準関数で、指定した引数のものを出力する。Lua 言語にはいくつかの基本的な関数が用意されていて、Proc Lua でも利用できる。ただし、必要最低限なもので、それほど数は多くなく、Proc Lua では利用できないライブラリもある。

必要に応じて自作するか、ユーザーで公開されているものを取得して、読み込んで利用する、或いは Proc Lua の場合、「sas.関数名()」という記述で一部の SAS 関数をそのまま利用できるので先にそちらを試すとよい。

## 2-2. 変数のスコープ

Lua の変数は、ローカル変数とグローバル変数に分けられる。何も指定せずに値を代入すると、自動的にグローバル変数となる。変数の前に local とつけるとローカル変数になる。ローカル変数は宣言されたブロックの中でのみ参照可能であり、グローバル変数はどこからでも参照可能である。

CODE11 の結果のログウインドウを見てみると、最初の proc lua で[NOTE:Lua initialized.]とメッセージがでている。これは Lua の環境が新規に立ち上がったということを示す。一旦 proc Lua を quit で終了して、次に proc Lua を実行するときもその環境は引き継がれ、グローバル変数の値は保持されている。ログに「NOTE:Resuming Lua state from previous PROC LUA invocation.」というメッセージからもそのことが確認できる。

二番目の出力でグローバル変数 x のみ「1」と出力され y、z については「nil」と出力されている。nil は変数が未定義や欠損であることを示し、SAS の null と類似している。もし既に定義済みの変数を削除したい場合は「x=nil」と nil を代入すればよい。

```
/*CODE 11 */
options nosource;
proc lua ;
submit;
 x = 1 --グローバル変数
 local y = 2
 local z = x + y
 print(x,y,z)
 endsubmit;
quit;

proc lua ;
 submit;
 print(x,y,z)
 endsubmit;
quit;
options source;
```

```
1     /*CODE 11 */
2     options nosource;

NOTE: Lua initialized.
1     2     3
NOTE: PROCEDURE LUA処理(合計処理時間):
      処理時間              0.87 秒
      CPU時間               0.01 秒


NOTE: Resuming Lua state from previous PROC LUA invocation.
1     nil     nil
NOTE: PROCEDURE LUA処理(合計処理時間):
      処理時間              0.00 秒
      CPU時間               0.00 秒
```

## 2-3. Lua 環境の初期化

```
/*CODE 12 */
proc lua restart ;
submit;
--XXXXX
endsubmit;
quit;
```

```
/*CODE 13 */
proc lua terminate ;
submit;
--XXXXX
endsubmit;
quit;
```

CODE12、CODE13 は共に Lua 環境のリセットの方法を示している。restart は proc lua の実行開始前に環境を自動的にリセットし、terminate は実行中については既存の環境を適用し、実行終了後に自動的にリセットを行う。

## 2-4. Lua の演算子

Lua における演算子について以下に示す。

| 算術演算子 | |
|---|---|
| + | 加算 |
| - | 減算 |
| * | 積算 |
| / | 除算 |
| ^ | 累乗 |
| % | 剰余 |

| 比較演算子 | |
|---|---|
| < | 未満 |
| > | 超 |
| <= | 以上 |
| >= | 以下 |
| == | 等しい |
| ~= | 等しくない |

| 論理演算子 |
|---|
| and |
| or |
| not |

| 文字列連結演算子 |
|---|
| .. |

SAS では、割り当てステートメントも、if X=Y then といった値の比較演算も同じ「=」を使用するが、Lua では比較演算子の等価は「==」となっている点、不等は「^=」ではなく「~=」である点、及び SAS での文字連結「||」が「..」となっている点について SAS プログラマは特に間違えやすいので注意が必要である。

## 2-5. Lua の変数型

| Lua の基本変数型 |
|---|
| Numbers |
| Strings |
| Boolean |
| Tables |
| Functions |
| nil values |
| Userdata |

```
/*CODE 14 */
proc lua;
submit;
 local x="A"
 print(type(x))

 local x =1
 print(type(x))

 local x="1"
 local y=x + 1
 print(type(y)..":"..y)

endsubmit;
quit;
```

```
string
number
number:2
```

Lua は動的型付け言語であり、変数の型は格納されている値によって決定される。CODE14 は変数 x に値を入れた後、type 関数で型をチェックするコードとなる。x の値によって型が変わっていることが確認できる。

また、文字値"1"に+1 した場合に結果が 2 となっている。これは SAS の暗黙の型変換に似た機能で、数字とみなせる場合は文字値であっても数値で処理される。なお"A"+1 などとした場合は、エラーとなる。

SAS プログラマにとって注意が必要なのは、Boolean、つまり真偽値型についてである。Boolean 型は True と False で定義されている。SAS の場合、数値の 0 と「null」を「偽」、それ以外の数値を「真」とみなした処理が記述できるが、Lua においては「偽」は Boolean における「False」と「nil」のみであり、Boolean における「True」と「nil」以外のあらゆる値が「真」となる。

つまり、「0」も Lua では真とみなされるため、注意が必要である。

## 2-6. 多重代入・スワップ

```
/*CODE 15 */
proc lua ;
submit;
 local a , b , c = 1, 2, 3
 local a , b = b , a
endsubmit;
quit;
```

Lua では SAS の割り当てステートメントと異なり、CODE15 のような多重代入や変数の値入れ替え（スワップ）を書くことができる

## 2-7. Lua の制御コード

Lua における条件分岐とループ処理について CODE16、CODE17 に例示する。

SAS プログラマが間違えやすいのは、前述の比較演算子「==」を「=」としてしまうケースと if に必ず end が必要になる点である。また UNTIL 型の条件を書く箇所が下部にくることも SAS と異なるため間違いやすい。

```
/*CODE 16 */
proc lua;
submit;
local i; local x=1;
--条件分岐
if x == 1 then
   print("YES")
 end
--ループ(単純ループ)
for i = 1, 3, 1 do
      print( "FOR Loop"..i)
end
endsubmit;
quit;
```

```
/*CODE 17 */
proc lua;
submit;
local i
--ループ(WHILE 型)
 i = 1
 while i <= 3 do
    print( "WHILE Loop"..i)
    i = i + 1
end
--ループ(UNTIL 型)
 i = 1
 repeat
    print( "UNTIL Loop"..i)
    i = i + 1
 until ( i > 3 )
endsubmit;
quit;
```

## 2-10. Lua のテーブル

Lua のデータ構造は、唯一「テーブル」でのみ表現される。テーブルは配列と連想配列の両方の性質を持つ。SAS プログラマの感覚では、array 配列とハッシュオブジェクトの両方の使い方ができる構造と捉えるとわかりやすい。

```
/*CODE 18*/
proc lua;
submit;

local table1 = {1,"A",2,"B"}

--要素番号(キー)で取得
print(table1[1])
print(table1[2])
print(table1[3])
print(table1[4])

--イテレータ関数で順番に取得
for i, var in ipairs(table1) do
            print(i,var)
end

endsubmit;
quit;
```

```
/*CODE 19*/
proc lua;
submit;

local table1 = {a=1,b="A",c=2,d="B"}

--キーで取得
print(table1.a)
print(table1.b)
print(table1.c)
print(table1.d)

--イテレータ関数で順番に取得
for key, var in pairs(table1) do
            print(key,var)
end

endsubmit;
quit;
```

CODE18 はテーブルを配列のように使用している例である。テーブル名={値 1,値 2,…}で作成する。Lua のテーブルは、数値や文字型、或いはテーブルや関数型など、nil 以外のあらゆる型を値として格納することができる。

SAS の array 配列が格納順に 1,2,3…と要素番号を持つように、CODE18 のように値を格納した場合、1 からの連番が自動的にキーとなる。Table1.[1]と指定すると 1 をキーとする値 1 を指定したこととなる。

また数値のキーの場合 ipars 関数という特殊な関数(イテレータ関数)を使ってループを記述することでテーブル

内の値を順番に取得することができる。

CODE19 はテーブルをハッシュオブジェクトのように使用している例である。
テーブル名={キー1=値 1, キー2=値 2,…}で作成する。Table1.a とすると a をキーとしてもつ値を取得することになる。文字を含むキーの場合、pairs 関数でテーブル内のデータを順次取得できる。

## 2-8. Lua の関数

Lua におけるユーザー定義関数は function～end で定義する。括弧内に引数を定義し、return で戻り値を定義する。戻り値を持たない関数や、関数を戻り値とする関数も定義可能である。CODE20 では syntax1,2 として 2 種類の書き方を行っているが意味は同じとなる。 CODE21 のように複数の戻り値を設定することも可能である。

```
/*CODE 20*/
proc lua;
submit;

--syntax1
local function add(a,b)
     return a+b
end
print(add(2,1))

--syntax2
local subtract =function(a,b)
     return a-b
end
print(subtract (1,2))

endsubmit;
quit;
```

```
/*CODE 21*/
proc lua;
submit;

local function three(a)
            local   r1=a^2
            local   r2=a^3
            local   r3=a^4
     return r1,r2,r3
end

out1,out2,out3=three(2)
print(out1,out2,out3)

endsubmit;
quit;
```

Lua には標準ライブラリとして、基本的な関数群が用意されている。前述の print 関数や type 関数等は基本ライブラリに属する関数である。他に「ライブラリ.関数名」のルールで使用できる関数があり、数学関連の math ライブラリ、文字列操作の string ライブラリ、テーブル操作の table ライブラリなどがある。CODE22 ではその一部の使用例を示した。詳細は Lua のリファレンスを参照していただきたい。ただし明文化されていないが、Proc Lua では非対応のライブラリや関数も多い(例：os ライブラリなど)。

また Proc Lua には SAS ライブラリが実装されており、基本的に「SAS.SAS 関数名」の既述で、SAS 関数を用いることができる。CODE23 はその例になる。ただし、同じく明文化されてはいないが、Proc Lua では非対応または記述が通常の SAS コードと異なる関数もある。

```
/*CODE 22*/
proc lua ;
 submit;
 x=math.min(2,5,3,1)
 print(x) --1 が出力される
 y=string.reverse("ABC")
 print(y) --CBA が出力される
 z={5,3,1,2}
 table.sort(z)
 print(table.unpack(z))
 --1 2 3 5 が出力される
 endsub;
quit;
```

```
/*CODE 23*/
proc lua ;
 submit;
 x=sas.mean(2,5,3,2)
 print(x)
--3 が出力される
 y=sas.substr("ABCDE",2,3)
 print(y)
--BCD が出力される

 endsubmit;
quit;
```

2-9. クロージャ

```
/*CODE 24*/
%let total=0;

%macro cum(plus);
  %let total= %eval(&total + &plus);
  %put 合計=      &total;
%mend;

%cum(1)
%cum(2)
%cum(3)
```

```
/*CODE 25*/
/*関数定義部分*/
proc lua;
 submit;
  function L(ds)
   local nobs= 0
   return function(ds)
    local dsid = sas.open(ds)
    nobs = nobs+sas.nobs(dsid)
    sas.close(dsid)
    return nobs
   end
  end
endsubmit;
quit;

/*実行部分*/
proc lua;
 submit;
  L1=L()
  print ("合計=",L1("sashelp.class"))
  print ("合計=",L1("sashelp.iris"))
  print ("合計=",L1("sashelp.cars"))
 endsubmit;
quit;
```

SAS マクロに存在しない概念として、クロージャがあげられる。SAS の場合、マクロ変数はグローバル変数とローカル変数に分類されるが、ローカル変数はマクロ実行時のみ保持され、そのマクロ内でのみ保持参照されるが、実行が終わった時点で消滅する。そのため、マクロの実行をまたいで値を保持するためには、グローバルマクロ変数を使わざるを得ない。しかしグローバルマクロ変数はどこからでも変更可能なため、仕様上は予期せぬ干渉を防ぐことができなくなっている。例として CODE24 は、データセットを与えることで、そのデータセットの obs 数を累積カウントしていくマクロであるが、マクロ変数 tobs をグローバルマクロ変数とすることで処理が成立している。

これに対して、あるスコープの中でしか参照されずに、保持され続ける性質をもっているのがクロージャである。CODE25 は Lua のクロージャの例になる。ここでは local nobs=0 としているが、この nobs は function L の中で、実行時のみ存在、保持されるイメージとなり、呼び出されるたびに加算されていく。そのため、関数が呼び出されている箇所以外で print(nobs)としても結果は nil となり、存在しないことが確認できる。L1=L()のように関数を変数に代入しているが、例えば L2=L()として、L2 を使えば L1 と平行して別カウントをすることも可能である。このようにクロージャを使うことで名前の衝突を回避し、より簡易に処理の構造化が行える。

## 3. Proc Lua で SAS コードを実行する

### 3-1. sas. submit

Proc Lua で SAS コードを実行する場合、sas.submit を利用する。次頁の CODE26-CODE29 は全て「data OUT1;x=1;run;」という SAS コードを Proc Lua の中で実行するコードである。

CODE26 のように引数に直接コードを既述してもよいし、CODE27 のように Lua の変数に格納して引き渡してもよい。

実行コードに可変パラメータ部分を作り、実行時に与えて実行する場合、CODE28、CODE29 のようにコード

中に@@でパラメータ部分を括る。CODE28 のように予め Lua の変数として定義するか、或いは CODE29 のように sas.submit 関数の第 2 引数で{パラメータ変数=値,...}として実行する。

```
/*CODE 26*/
proc lua;
submit;
 sas.submit([[data OUT1;
              x=1;
             run;]])
endsubmit;
quit;
```

```
/*CODE 27*/
proc lua;
submit;
 code=[[data OUT1;
         x=1;
        run;]]
 sas.submit(code)
endsubmit;
quit;
```

```
/*CODE 28*/
proc lua;
submit;
 local ds="OUT1"
 local var="X"
 sas.submit(
[[data @ds@;
 @var@=1;
 run;]])
endsubmit;
quit;
```

```
/*CODE 29*/
proc lua;
submit;
 code=[[data @ds@;
 @var@=1;
 run;]]
 sas.submit(code,
{ds="OUT1",var="x"})
endsubmit;
quit;
```

3-2. SAS. submit_

sas.submit は与えられたコードを即時実行し、仮に実行コードに run ステートメントや quit ステートメントが存在しなかったとしても実行が完了状態になる。それに対して sas.submit_(末尾にアンダースコアがつく)は与えられたコードを即時に実行せず、sas.submit が実行されるまで保留させることができる。

これを用いれば CODE30 のように、実行コードを分解して実行し、途中で動的にコードを生成するといった処理を行うことができる。

```
/*CODE 30*/
proc lua;
submit;
 sas.submit_([[data OUT2;]])
 for i = 1,5 do
  sas.submit_([[x=@val@;output;]],{val=i})
 end
sas.submit([[run;]])
endsubmit;
quit;
```

```
data OUT2;
x=1;output;
x=2;output;
x=3;output;
x=4;output;
x=5;output;
run;
```

3-3. マクロ変数の取り扱い

SAS では、グローバルマクロ変数が定義されている場合、コード中のアンバサンド(&)のつく変数名は、プログラム実行時に展開される。

しかし、Proc Lua 内においては、このマクロ変数の展開が起きないようになっている。CODE31 を実行すると「&m1」は「AA」に展開されず、そのまま「&m1」と出力される。

Peoc Lua 内にマクロ変数に格納されている値を持ち込むには以下の 2 つの方法がある。

```
/*CODE 31*/
%let m1=AA;

proc lua ;
 submit;
  print("&m1")

  --&m1 と出力される

 endsubmit;
quit;
```

```
/*CODE 32*/
%let m1=AA;

proc lua ;
 submit;
   print(sas.symget("m1"))

  --AA と出力される

 endsubmit;
quit;
```

```
/*CODE 33*/
%let m1=AA;

proc lua ;
 submit"m1='&m1'";
  print(m1)
  --AA と出力される

 endsubmit;
quit;
```

```
/*CODE 34*/
proc lua ;
submit;

 local x=1;
 sas.symput("m2",x)

endsubmit;
quit;

%put &m2;
/*1 と出力される*/
```

一つ目は、sas.symget 関数を使用する方法(CODE32 参照)、もう一つは submit ステートメントに"Lua の変数名

='&マクロ変数' "を加える方法(CODE33 参照)となる。ただし submit を使用する場合、作成される Lua の変数は必ずグローバル変数になってしまうため、ローカル変数で持ち込みたい場合は local 宣言した変数に sas.symget で値を代入する必要がある。

Proc Lua 内でマクロ変数を作成する場合については、sas.symput 関数を利用する(CODE34 参照)。

## 3-4. データセットから値を読み込む

「1. 具体的な使用例と既存機能との比較]の「【例 3】データセットの内容から実行コードを生成する」でも例示したが、Proc Lua 内で SAS データセットから値を取得することができる。

```
/*CODE 3 5*/
proc lua;
submit;
local dsid = sas.open("sashelp.class")
  sas.where(dsid,"age<=13")
  while sas.next(dsid) do
    print(sas.get_value(dsid,"name")
          , as.get_value(dsid,"age"))
  end
 sas.close(dsid)
endsubmit;
quit;
```

```
アリス      13
バーバラ      13
ジェームズ      12
ジェーン      12
ジェフリー      13
ジョン      12
ジョイス      11
ルイーズ      12
ロバート      12
トーマス      11
```

CODE35 のように、sas.open でデータセットを指定し、そのデータセットを認識するための ID 値を設定する。その後、sas.where で抽出条件を設定し、sas.next で順次オブザベーションを移動し、get_value 関数に変数名を指定することで値を取得することができる。

```
/*CODE 3 6*/
proc lua;
submit;
  local dsid = sas.open("sashelp.class")
  for var in sas.vars(dsid) do
    print("変数名:",var.name)
    print("ラベル:",var.label)
    print("Length:",var.length)
  end
  sas.close(dsid)
endsubmit;
quit;
```

```
変数名:     Name
ラベル:     名前
Length:     12
変数名:     Sex
ラベル:     性別
Length:     4
変数名:     Age
ラベル:     年齢
Length:     8
変数名:     Height
ラベル:     身長(インチ)
Length:     8
変数名:     Weight
ラベル:     体重(ポンド)
Length:     8
```

CODE36 はデータセットの値ではなくラベルやフォーマットなどの属性情報を出力するコードとなる。

CODE35,36 のループを組み合わせればテーブルの中に変数ごとに子テーブルを持たせ、その中に値をいれるといったコードが実現できる。

```
/*CODE 3 7*/
data temp;
 set sashelp.class;
 where age=12;
run;
proc lua;
submit;
table_A=
sas.load_ds("temp")
print(table.tostring(table_A))
endsubmit;
quit;
```

```
table: 0DC5C4C0=
{
  [1]=table: 0DC5D490=
  {
    ["name"]="ジョイス"
    ["age"]=11
  }
  [2]=table: 0DC63610=
  {
    ["name"]="トーマス"
  } ["age"]=11
  [
  {"vars"]=table: 0DC68260=

    ["name"]=table: 0DC69E50=
    {
      ["fmt_dec"]=0
      ["type"]="C"
      ["label"]="名前"
      ["fmt_width"]=0
      ["informat"]=""
      ["name"]="Name"
      ["infmt_dec"]=0
      ["format"]=""
      ["infmt_width"]=0
      ["length"]=12
    }
    ["age"]=table: 0DC59CB0=
    {
      ["fmt_dec"]=0
      ["type"]="N"
      ["label"]="年齢"
      ["fmt_width"]=0
      ["informat"]=""
      ["name"]="Age"
      ["infmt_dec"]=0
      ["format"]=""
      ["infmt_width"]=0
      ["length"]=8
    }
  }
  ["name"]="A"
  ["nvars"]=2
}
```

また、別の方法として sas.load 関数を使うことで SAS データセットを決まった形で Lua のテーブルに取り込むことができる(CODE37)。この場合。obs ごとに子テーブルが作成される作りとな

り、obs 番号が要素番号となる。また変数ごとに属性情報がテーブル化(変数名がテーブルキー)となって同梱され、データセット名(name),変数数(nvars)も同様に作成される。

### 3-5. データセットに値を出力する

Proc Lua から SAS データセットを作成、或いは SAS データセットに値を追加することもできる。

```
/*CODE 38*/
proc lua;
submit;

--空のデータセットを作成
 sas.new_table("work.XX", {
 { name="X", type="C", length=20},
 { name="Y", type="N", length=8, format="yymmdds10."}
  })

 local A, B = 1 ,21035

 --値を追加してデータセットを更新
 local dsid = sas.open("work.XX", "u")
  sas.append(dsid)
  sas.put_value(dsid, "X", A)
  sas.put_value(dsid, "Y", B)
  sas.update(dsid)
 sas.close(dsid)

endsubmit;
quit;
```

```
/*CODE 39*/
proc lua;
submit;
 table_B={}
 table_B[1]={X="A",Y=1}
 table_B[2]={X="B",Y=2}
 table_B.vars={}
 table_B.vars.X={type="C",length=20}
 table_B.vars.Y={type="N",label="ラベル"}
 sas.write_ds(table_B,"YY")
endsubmit;
quit;
```

| X | Y |
|---|---|
| A | 2017/08/04 |

| ラベル | X |
|---|---|
| 1 | A |
| 2 | B |

CODE38 は sas.new_table で新規に SAS データセットを定義し、作成されたデータセットを、読み込みの時と同様に sas.open で掴み、sas.append で行を追加し、put_value で変数に値を挿入し、sas.update で更新している。sas.open の際に第二引数を u に指定し、編集モードで開くことが必要となる。

CODE39 は CODE37 の真逆の処理であり、sas.write_ds に sas.load_ds の際に作成される Lua テーブルと同じ構造のテーブルを与えることで Lua のテーブルを作成している。

## 4. Lua でオブジェクト指向

### 4-1. メタテーブル

Lua ではテーブル(値)にメタテーブルを付与することができる。メタテーブルは値に影響せずに情報を保持できるほかに、その値に対しての演算の挙動を既述することができる。

```
/*CODE 40*/
proc lua;
submit;
   tb={1,2,3}
   setmetatable(tb, {meta1="XXX", meta2="YYY"})
   mt = getmetatable(tb)
   print("tb の中身",table.tostring(tb))
   print("tb のメタテーブル中身",table.tostring(mt))
 endsubmit;
quit;
```

```
tbの中身     table: 0B41CB60=
{
  [1]=1
  [2]=2
  [3]=3
}
tbのメタテーブル中身     table: 0B41DE70=
{
  ["meta1"]="XXX"
  ["meta2"]="YYY"
}
```

CODE40 は setmetatable でテーブルにメタテーブルに値を設定し、getmetatable で取得する例となる。

```
/*CODE 41*/
proc lua;
submit;

--テーブル
tb1 = {a = 1, b = 2, c = 3}
tb2 = {a = 4, c = 5, d = 6}

--「+」テーブルマージを行う関数
function table_merge(a, b)
    local c = {}
    for key, val in pairs(a) do
        c[key] = val
    end
    for key, val in pairs(b) do
        c[key] = val
    end
    return c
end

--__add に関数を指定
setmetatable(tb1, {__add=table_merge})

--tb1 のメタテーブルにより、+演算子の挙動において関数が参照される
tb3 = tb1 + tb2

print("tb3 の中身",table.tostring(tb3))

endsubmit;
quit;
```

```
tb3の中身    table: 0A210CE0=
{
  ["a"]=4
  ["d"]=6
  ["c"]=5
  ["b"]=2
}
```

CODE41 では、メタテーブルを利用して「+」演算子の挙動を再定義し、SAS のマージに似せた機能を実装している例となる。

通常「+」演算子は、単一の値の加算を行う機能であり、テーブル型を指定するとエラーになる。

しかし setmetatable で__add=に挙動を定義した関数をしてやればテーブル同士の演算にメタテーブルが参照され、実現可能となる。__add のようにメタテーブルで演算子の挙動を定義する場合、アンダーバー2 つにキーワード(この場合 add)が決まった文法となる。

## 4-2. __index の利用

Lua のテーブルにおいて参照しようとするフィールドが存在しない場合、エラーとなる。しかし setmetatable で__index=にテーブルが指定されている場合、指定されたテーブルのフィールドが参照される。

その特性を利用して、オブジェクト指向言語における new 演算子と同様の機能をもつ関数を定義し、クラスからインスタンスを生成する挙動を実装することができる。(次頁 CODE42 参照)

```
/*CODE 42*/
proc lua ;
submit;
rand = {}

function rand:setobs(obs)
  self.obs = obs
end
function rand:setseed(obs)
  self.seed = obs
end

function rand:create_ds()
  print("seed=".. self.seed)
  sas.submit([[
          data out;
           call streaminit(@seed@);
            do i=1 to @obs@ ;
             x=rand("uniform");
             output;
             end;
           run;
     ]],{seed=self.seed,obs=self.obs})
end

function rand:new()
  local t = {}
  t.seed=1234
  t.obs = 10
  setmetatable( t, {__index = rand})
  return t
end

local rand1 = rand:new()
rand1:create_ds()

local rand2 = rand:new()
rand2:setobs( 20 )
rand2:setseed( 5678 )
rand2:create_ds()

endsubmit;
quit;
```

```
seed=1234
    data out;
     call streaminit(1234);
      do i=1 to 10 ;
       x=rand("uniform");
       output;
       end;
     run;

NOTE: データセットWORK.OUTは10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):
      処理時間            0.04 秒
      CPU時間             0.01 秒


seed=5678
    data out;
     call streaminit(5678);
      do i=1 to 20 ;
       x=rand("uniform");
       output;
       end;
     run;

NOTE: データセットWORK.OUTは20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):
      処理時間            0.00 秒
      CPU時間             0.01 秒


NOTE: PROCEDURE LUA処理(合計処理時間):
      処理時間            0.92 秒
      CPU時間             0.03 秒
```

## 5. まとめ

Lua は意図的に限りなくシンプルに設計されており、加えてテーブルが著しく応用の利く構造を持つため、SAS マクロに比べて処理を簡潔かつ柔軟に実装することができる。SAS マクロ特有のルールがなく、工夫を加えることで他言語の開発で一般的な概念を導入することができる。特にオブジェクト指向を取り入れることで、機能の体系的な構築と管理が可能となる。

学習は比較的短時間ですみ、他言語からの乗り入れは SAS マクロに比べて敷居が低いと思われる。

一方で、シンプルさの代償として提供されている機能は最低限となっているため、テーブルという構造をうまく利用してどのように仕組みを構築するかというところに頭を使う必要がある。

また SAS マクロは、SAS コード中のどこにでも挿入することができ、テキスト展開によって実行コードを生成

する強力な機能をもっているため、そのまま全く同じ形で Proc Lua に移植することは難しく、コードの構造自体を再考する必要があることが多い。

プログラム開発において、どのように Proc Lua と SAS マクロを取り入れれば、パフォーマンスを発揮するか、また既存の SAS マクロ資産との兼ね合いについては今後、集団での経験と知見の積み重ねが必要と思われる。

## 6. おわりに

SAS9.4 のリリース以降、Proc DS2 や Proc Lua が導入され、それまで SAS マクロのみで担っていたプログラム制御に選択肢が追加された。SAS マクロは強力で便利な機能ではあるが、プログラムの効率的な管理に向いているとは言いがたい。他言語の経験が無く、SAS に熟練したユーザーにとって、Proc Lua の導入と運用には多大な抵抗と苦痛を感じるであろうことが予想される。

しかし SAS マクロが 20～30 年以上、ほぼ同じ機能で留まっている間に、プログラミングの開発手法は大きく発展を遂げている。その中には学習コストを支払ってでも、キャッチアップすべき有用なものが多く存在すると考える。

今後、より実践的な例を研究、発表していきたい。

## 7. 参考文献

(1) Base SAS® 9.4 Procedures Guide, Seventh Edition

(2) Paul Tomas (2015), Driving SAS® with Lua, Paper SAS1561-2015

(3)Jiangtang Hu (2015), New Game in Town: SAS® Proc Lua with Applications, Paper AD-133

# SAS以外の言語を利用した、SASで実現困難なロジックや機能の実装について

○川上貴弘
（電助システムズ株式会社）

## 1　概要

SASを利用するにあたっては、多くの場合、「バリデーション」という作業を行う。計画を立てて期待された通りに動作することを検証することである。これによって、そのバージョンにおける動作が保証されるが、SASをバージョンアップする際、再バリデーションが必要となることから、バージョンアップを容易に実施できなくなるケースは少なくない。SASのバージョンアップの目的は、バグフィックスだけでなく、新しい機能の追加も含まれる。しかしながら、SAS9.4が最新である現在においても最新バージョンでのバリデーションが困難なため、SAS9.2、SAS9.3を使用せざるを得ず、そのため、最新バージョンで追加された便利な新機能を利用できずにいまだに複雑なコードで実現しているケースも散見される。

本発表では、さまざまな事情でSASの最新バージョンを使用できないユーザー向けに、旧バージョンのSASでは使用できない、もしくは複雑なコードで実現している方法を簡略化する方法の一例を紹介したい。

近年、Heartblood, poodle等、ベースとなる技術自体の脆弱性や技術の危殆化によるセキュリティインシデントが多く発生し、また、2017年2月にはGoogleがハッシュアルゴリズムSHA-1の対衝突性を破ることに成功したことで、暗号化方式は従来一般的であったSSLからTLSへ、ハッシュアルゴリズムはより安全であるSHA-256への移行が急速に進んでいる。しかしながら、SASにおいてはSAS9.4TS1M1でSHA-256ハッシュを生成する関数が追加されたものの、SAS9.3以前ではすでに容易に同一ハッシュ値が得られると指摘されているMD5しか使用できない状況であった（SAS9.3のSAS/SECUREを除く）。

また、ハッシュアルゴリズムは公開されているが、SASでSHA-256ハッシュを生成するプログラムを作成することは一般的なSASユーザーにとっては非常に困難であることは想像に難くない。加えて、Javaやruby、c++などで実現するコードは公開されているものの、そのためのインターネット接続環境やバージョン依存の実行環境を用意し、SASとの連携ができるよう修正することも容易ではない。そのため、より少ないコーディングで、なおかつSAS9.3以前のバージョンでもSHA-256ハッシュを生成する手段を実現できないかを調査、検討した。

## 2　調査と選定

調査の過程で、インターネット接続環境が必要な技術は除外した。また、度々脆弱性が見つかっているOpenSSLライブラリを使用する方法も除外した。その上で、ハッシュを生成して結果を返す仕組みがないか検討を行った結果、クロスプラットフォーム開発環境であるDelphiを採用することとした。

近年のDelphiはクロスプラットフォーム開発環境として、頻繁なアップデートによる機能追加が著しい。生成されるコードはdllや単体での実行が可能なexeファイルであり、アプリケーション間の連動に非常に都合がいいことも採用の理由として大きい。余談だが、行末がセミコロン「;」で終わることはSASユーザーにとって親近感を覚える仕様ではなかろうか。

今回は、2017年3月にリリースされたDelphi 10.2 Tokyoを使用して、SASでハッシュ値を生成するためのプログラムを作成した。

プログラムは以下の通り、わずか14行である。

## 3 プログラム

### 1. Delphiプログラム

ファイル名を「myHash.dpr」として以下を記述する。作成したファイルをDelphiで開き、Ctrl+F9を押せばコンパイルが完了し、myHash.dllが生成される。dllとは、それ自体では動作しないが、汎用的な機能がモジュール化され、アプリケーションプログラムからその機能を呼び出せるようにしたものである。

```
01 library myHash;
02 uses System.SysUtils, System.Classes, System.Hash;
03 {$R *.res}
04 function GetHash(s:PAnsiChar):PAnsiChar stdcall;
05 var
06   Hash: THashSHA2;
07 begin
08   Hash := THashSHA2.Create(SHA256);
09   Hash.Update(trim(String(s)));
10   Result:=PAnsiChar(AnsiString(Hash.HashAsString));
11 end;
12 exports GetHash;
13 begin
14 end.
```

プログラム各行は以下の通りである。

01行目:dllとして宣言する
02行目:Delphiのライブラリを指定する
03行目:
04行目: DelphiはUNICODEを扱うが、SAS(日本語)はANSIなので受け渡しの定義はANSIと明示する，GetHashは文字型「s」を受け取り、処理結果を返す関数(function)として定義している
05行目:関数内部で使用する変数を定義する
06行目:SHA-2ハッシュを実装するクラスを定義する
07行目:関数GetHashの開始を意味する
08行目:SHA-256を指定し、ハッシュを実装するオブジェクトを生成する(224～512も指定可能)
09行目:ANSI文字「s」を内部処理のためUNICODEに型変換してハッシュ値を生成する
10行目:UNICODEとして生成されたハッシュ値をANSIに型変換し、更に文字型に型変換する
11行目:関数GetHashの終わりを意味する
12行目:外部から呼び出される機能名(04行目のfunction名)を宣言する
13行目:プログラムの開始を意味する
14行目:プログラムの終わりを意味する

### 2. プログラム作成後の問題点

- SASでdllを読み込み、値の受け渡しをすることができなかった。そこで、SASのMODULEC関数にコントロール文字列 '*E' を加え、実行時エラーを逐一ログに出力することで状況を把握した。
  1) ERROR: Module myHash could not be loaded.
  2) ERROR: Routine GetHashX of module myHash could not be loaded.
  3) WARNING: No attribute table entry found for GetHashX.
  4) ERROR: Module GetHash was given over its maximum argument count of 1.

  結果としては、上記エラー1)から、dllをパスの通っているフォルダにコピーしておくこと、dllとSASのbit数が一致していなければならないことが判明した。
- dll読み込みの問題は解決したが、次は正しいハッシュが得られない結果となった。Delphi単体では正しいハッシュが得られるものの、SASで実行すると異なる結果が得られたのである。
  これは、Delphi側が扱う文字列はUNICODEであるのに対し、SASが扱う文字列はANSIであるためだった。Delphiは2009年にリリースされたバージョンからUNICODE化されているので、まずdllではANSI文字型としてSASから受取り、内部でUNICODE文字列に変換してからハッシュ値を生成し、再度ANSI文字型に変換してSASに返す、という処理を行うことで解決した。
- dllを利用するためのsascbtblの情報が少なく、マニュアルだけでなく、過去のSUGIの情報も参考にした。

### 3. SAS9.3における実行例

まず、コンパイルで生成されたmyHash.dllをパスの通ったフォルダにコピーする必要がある。最もわかりやすのはSAS起動用ショートカットの作業フォルダに指定されている%USERPROFILE%である。

コピーした上で以下のプログラムを実行する。

```
01 filename SASCBTBL 'tmp.txt';
02 data _null_;
03   file sascbtbl;
04   put 'routine GetHash';
05   put '  minarg=1';
06   put '  maxarg=1';
07   put '  MODULE=myHash';
08   put '  stackpop=called';
09   put '  returns =char64;';
10   put 'arg 1 char input byaddr format=$200.;';
11 run;
12
13 data _null_;
14   length hash $64;
15   source='ABC';
16   hash = modulec('GetHash', source);
17   put hash=;
18 run;
```

プログラム各行は以下の通りである。

```
01行目: 属性テーブルの作成。作成済ならばfilenameの割当のみ
02行目:
03行目: dllを使用するためにはsascbtblを使用する必要がある
04行目: Delphiコード中のexportで指定されたdllの機能名
05行目: パラメータは1個であるので、最小も1
06行目: パラメータは1個であるので、最大も1
07行目: dll名を指定する
08行目: GetHashにはstdcall属性を持たせているのでcalledとする
09行目: 返り値(SHA-256)は64バイトの文字列である
10行目: dllに渡す文字列の長さ(ここでは200字としているが、より長くすることも可能)
11行目:
12行目:
13行目:
14行目: 64バイトの変数長を確保する
15行目:
16行目: modulec関数でdllを呼び出し、文字列「ABC」のハッシュ値を生成する
17行目:
18行目:
```

ログには以下のように出力される。

```
hash=b5d4045c3f466fa91fe2cc6abe79232a1a57cdf104f7a26e716e0a1e2789df78
```

なお、一度実行済みで、tmp.txtがすでに存在するのであれば、以下だけでよい。

```
01 filename SASCBTBL 'tmp.txt';
02 data _null_;
03   length hash $64;
04   source='ABC';
05   hash = modulec('GetHash', source);
06   put hash=;
07 run;
```

### 4. SAS9.4における確認コード

以下を実行する。

```
01 data _null_;
02 length hash $64;
03   source='ABC';
04   hash = put(SHA256(source),hex64.);
05   put hash=;
06 run;
```

ログには以下のように出力される。

```
hash=B5D4045C3F466FA91FE2CC6ABE79232A1A57CDF104F7A26E716E0A1E2789DF78
```

大文字小文字の違いはあるが、UPPER/LOWERで容易に変更できるので、気にする必要はない。

## 4 まとめ

さまざまな事情でSASの最新バージョンを使用できないユーザー向けに、旧バージョンのSASでは使用できない、もしくは複雑なコードで実現している方法を簡略化する方法として、SHA-256ハッシュの生成を例に、別のアプリケーションと連携することで比較的容易に実現できることを示した。今回はdllを自作したが、他にも、WinAPIを使用するためのdllを利用したり、xコマンドを使用して外部アプリケーションと連動させたりすることもよく行われている。実際に弊社でも、dllやexeファイルをSASと連動させて複雑なバッチ処理を行う仕組みを構築している。

既存のソフトウェア資産等を利用し、SASをより有効に活用する方法が広まることへの一助となれば幸いである。

なお、掲載したコードはWindows版SAS9.3 TS1M2, SAS9.4 TS1M3, Delphi 10.2 Tokyoにて作成、実行させたものである。

## 5 参考文献


1. 日向俊二：Delphi XE2プログラミング入門, カットシステム (2012)
2. Embarcadero Technologies：Delphi online help docwiki(Delphi 10.2 Tokyo) http://docwiki.embarcadero.com/Libraries/Tokyo/ja/System.Hash.THashSHA2
3. SAS Institute Inc.：Windows版SAS® 9.4第3版 (2014)
4. SAS Institute Inc.：SAS® 9.4関数とCALLルーチン リファレンス第4版 (2015)
5. SAS Institute Inc.：SAS® 9.4 Companion for UNIX Environments, Sixth Edition (2016)
6. David H. Johnson.：SAS® with the Windows API (SUGI 30)
7. Edward Foster.：Using the WIN32 API from SAS (PhUSE 2006)


## 6 謝辞


1. SAS Technical Support (SAS Institute Japan Ltd.)

# SASプログラムの可視化
## － SASプログラムステップフローチャート生成プログラムの紹介 －

○福田裕章[1]

([1]MSD株式会社)

近年、SASのデータハンドリング機能は一層進化してきているが、未だデータステップやSQLプロシジャがその主流である。統計解析プロシジャを含めた、こうしたステップの繰り返しはSASの特徴の一つだと考えられるが、解析データセット作成プログラム等、多数のデータセットから多数の新規変数を導出するような場合、ステップ数が数百に達することもあり、そのプログラムをトレースしていくことでさえ多くの労力を費やしてしまうことになる。これを回避する一つの方策として、ステップフローの可視化、すなわちSASプログラムのステップをフローチャート化することを考えた。これまで、SASログからフローチャートを作成したような報告はあるが、ログは必ずしも利用可能とは限らず、また設定に依存して情報のアウトプットが制限されるため、SASプログラム自体からのステップフロー作成を実施することに注目した。SASプログラムを読み取り、どのコードがDATA/PROCステップで、インプットやアウトプットを指定しているのか判定するためには、SASシステムがそれらを判定するロジックと同様の処理をSASプログラム上で行えば良いと考えられるが、それには自由度の高い文字列処理が必要である。こうした柔軟な文字列処理は、データステップ中でPerl Regular Expression（PRX）及びハッシュオブジェクトを利用することで実現可能である。一方、フローチャートにおいては、SAS9.4より導入されたSGPLOTプロシジャにおけるPOLYGON並びにTEXTステートメントに着目した。これらのステートメントにより、Plot上に柔軟な図形とテキスト配置を実現することができる。本発表では、以上の機能を踏まえた、SASプログラムを可視化する手法の一つとしての、ステップフローチャートの生成プログラムを紹介する。

# Proc LUAを初めて使ってみた -SASでの処理を条件に応じて変える-

○浅井 友紀
(エイツーヘルスケア株式会社)

【はじめに】
これまで実行されるプログラムを分岐、繰り返すにはSASマクロが用いられてきた。
本発表ではSASマクロではなく、PROC LUAを用いてSASのコードの分岐や繰り返し処理を、EXCELの情報を変数名、変数ラベルとする0オブザベーションのデータセットを作成するプログラムを通して紹介する。

【方法】
本発表では簡便な例として、変数名等の定義情報をexcelファイルから読み込み、その定義情報の変数を持つ0オブザベーションのsasデータセットを作成するプログラムを取り上げる。
excelから読みこんだ変数名が8文字以上か変数ラベルが40字以上、もしくは変数名が重複しているならエラーをログに表示するようプログラムを分岐させた。
また複数のexcelシートを読みこみ、1シートにつきSASデータセットを1つ作成するようプログラムの繰り返し処理を行った。

【結果】
読みこむ予定のexcelファイルの例を図1に、図1をもとに作成したsasデータセットを図2に示す。

| Name1 | Name2 | Label | Type | Format |
|---|---|---|---|---|
| A | ID | ID | text | |
| A | A1 | label 1 | text | |
| A | A2 | label 2 | text | |
| A | A3 | label 3 | text | |
| A | A4 | label 4 | num | YYMMDD10. |
| A | A4_1 | label 41 | num | |
| A | A5 | label 5 | text | |
| A | A6 | label 6 | text | |
| A | A7 | label 7 | num | |
| A | A8 | label 8 | text | |

図1 読みこんだexcel(sheet3)

ライブラリ

| 名前 | Obs | Vars |
|---|---|---|
| SHEET1 | 0 | 10 |
| SHEET3 | 0 | 10 |

| # | 変数 | 種類 | 長さ | 出力形式 | 入力形式 | ラベル |
|---|---|---|---|---|---|---|
| 1 | ID | Character | 200 | | | ID |
| 2 | A1 | Character | 200 | | | label 1 |
| 3 | A2 | Character | 200 | | | label 2 |
| 4 | A3 | Character | 200 | | | label 3 |
| 5 | A4 | Numeric | 8 | YYMMDD10. | | label 4 |
| 6 | A4_1 | Numeric | 8 | | | label 41 |
| 7 | A5 | Character | 200 | | | label 5 |
| 8 | A6 | Character | 200 | | | label 6 |
| 9 | A7 | Numeric | 8 | | | label 7 |
| 10 | A8 | Character | 200 | | | label 8 |

図2 作成したSASデータセット(sheet3)

【総括】
PROC LUAを用いることで実行するコードの分岐、繰り返しを行うことが出来た。
本発表で取り上げた例は簡便な一例に過ぎない。様々なプログラムの書き方がPROC LUAによって実現できることを感じていただければ幸いである。

【参考文献】
Base SAS 9.4 プロシジャガイド 第4版
(http://www.sas.com/offices/asiapacific/japan/service/help/pdf/v94/proc.pdf)

# SASシステム／その他関連分野

# SAS Tips.
# Top10 My Favorite SAS Functions.

○松沢 享
(株式会社インターサイエンティフィックリサーチ)

SAS9から非常に便利な関数及びコールルーチンが追加された。本発表では、その中から特に便利なものを10個に厳選して、簡単な使用例と共に紹介する。

# 生物統計・医薬関連

# PHREGプロシジャにおける
# 共変量調整解析に関連したオプション機能

○魚住 龍史[1]　矢田 真城[2]　浜田 知久馬[2]
[1]京都大学大学院医学研究科 医学統計生物情報学
[2]東京理科大学大学院 工学研究科 経営 工学専攻

Investigating fascinating aspects associated with covariate-adjusted analysis using PHREG procedure

Ryuji Uozumi[1], Shinjo Yada[2], and Chikuma Hamada[2]
[1]*Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine*
[2]*Department of Management Science, Graduate School of Engineering, Tokyo University of Science*

## 要旨

PHREG プロシジャは Cox の比例ハザードモデルによる解析を実施するためのプロシジャであり，SAS/STAT のバージョンアップのたびに，構文数およびオプション数が増加している．本稿では，『生存時間解析 応用編 –SAS による生物統計』で取りあげられている PHREG プロシジャにおけるオプション機能のうち，BASELINE 文および LSMEANS 文に着目した．これらの有用なオプション機能として，BASELINE 文における DIRADJ オプションの機能および LSMEANS 文による Diffogram の作成に対する理解をより深めるための報告を行う．

キーワード：生存時間解析　比例ハザードモデル　PHREG　BASELINE　生存関数　Breslow 法　調整生存関数　DIRADJ　LSMEANS　Diffogram　Mean-Mean scatter plot　ODS GRAPHICS

## 1　はじめに

大橋・浜田 (1995) が刊行されてから 20 年もの歳月が流れ，2016 年 7 月に姉妹本という位置付けで，『生存時間解析 応用編 –SAS による生物統計』が刊行された (大橋ら, 2016)．現在，SAS 9.4 が最新版としてリリースされているが, LIFETEST プロシジャなどの生存時間解析に関連したプロシジャは SAS/STAT として提供されており，ほぼ毎年メンテナンス版がリリースされている．『生存時間解析 応用編 –SAS による生物統計』では，主に SAS/STAT 9.2 以降に追加された最新機能の使い方を幅広く解説しており，2016 年の SAS ユーザー総会の前夜祭の日に，SAS Learning Session として解説も行われた (浜田・魚住, 2016a)．また，『生存時間解析 応用編 –SAS による生物統計』で取りあげられていない生存時間解析に関連した最新機能も SAS

ユーザー総会で多く報告されており (浜田・魚住, 2016b; 魚住ら, 2016b; 矢田・浜田, 2016; 矢田ら, 2016, 2017), 今後も新たなプロシジャおよびオプション機能が追加されることが期待されている.

本稿では,『生存時間解析 応用編 –SAS による生物統計』の刊行後に, 読者から寄せられた質問のうち, PHREG プロシジャに関連した内容に焦点を当てることとする. PHREG プロシジャは Cox の比例ハザードモデル (Cox, 1972) による解析を実施するためのプロシジャであり, SAS/STAT のバージョンアップのたびに, ステートメントおよびオプションの数が増加している. 大橋・浜田 (1995) 刊行時は 10 のステートメント (文) が提供されているだけであったが, SAS/STAT 14.2 においては 24 のステートメントが提供されている (SAS Institute Inc., 2016). 著者らは, PHREG プロシジャにおけるオプション機能のうち, ベースライン生存関数を推定するための BASELINE 文, および固定効果に対する最小二乗平均を求める LSMEANS 文におけるオプション機能に着目した. BASELINE 文では, SAS/STAT 9.2 より DIRADJ オプションが追加されており (SAS Institute Inc., 2009), 平均生存関数の推定が可能である. また, ODS GRAPHICS ON の状態で, LSMEANS 文を指定して PHREG プロシジャを実行すると, 水準間の対比較の結果を表す Diffogram が作成される. 本稿では, BASELINE 文および LSMEANS 文における指定方法および有用なオプション機能の解説を行い, 指定方法によって異なる出力結果が得られる点に関して, 理解をより深めるための報告を行う.

## 2 BASELINE 文とオプション機能

PHREG プロシジャにおいて, BASELINE 文は大橋・浜田 (1995) が刊行された頃からサポートされている. BASELINE 文を指定することで, ベースラインハザード関数やベースライン生存関数などの推定を行うことが可能である.

本節では, BASELINE 文を用いてベースライン生存関数の推定を行うことを考える. 対象データとして, 肺癌のデータ (データセット名 : VALung) を用いる. データセット VALung (Veterans Administration Lung cancer trial) は, Kalbfleisch and Prentice (2002) で使用されたデータを一部抽出したものであり ($n = 137$), SAS/STAT PHREG プロシジャのマニュアル, および大橋ら (2016) 第 3 章においても用いられている. この研究の目的は, 男性の進行性肺癌患者を対象としたランダム化比較試験であり, 治療法 (変数名: Therapy) として, 試験治療 (Therapy = 'test') と標準治療 (Therapy = 'standard') を比較するために行われた. 評価項目は死亡までの時間 (日) (変数名: Time) で, 共変量の 1 つとして組織型 (変数名: Cell) が挙げられる. 組織型は 4 水準 (Cell = 'adeno' 「腺癌」, 'small' 「小細胞癌」, 'large' 「大細胞癌」, 'squamous' 「扁平上皮癌」) のカテゴリカル変数である. なお, PHREG プロシジャの処理の都合上, 文字変数である Therapy および Cell の代わりに, 治療法を表す数値変数 Therapy2 (Therapy2 = 1「試験治療」, 2「標準治療」), 組織型を表す数値変数 CellC (CellC = 1「腺癌」, 2「大細胞癌」, 3「小細胞癌」, 4「扁平上皮癌」) も用意した.

### 2.1 BASELINE 文による生存関数の推定

データセット VALung に対して, LIFETEST プロシジャおよび PHREG プロシジャで生存時間曲線を描くと, SGPANEL プロシジャでパネルを代えて出力すると図 1 のように示される.

なお, 本稿において示すプロットは, モノクロ印刷されても識別しやすいよう, ODS の出力としてジャーナルスタイル "STYLE = JOURNAL" を指定している. 例えば, 群別のグラフを示す場合, デフォルトでは群の違いが色で分けて出力されてしまう. ジャーナルスタイルの場合, モノクロ印刷で群の違いを識別できるよう, 実線と破線で分けて出力される.

また，図 1 のように，SAS/STAT のプロシジャの実行結果をデータセットとして作成し，SG (Statistical Graphics) プロシジャで出力する方法については，これまでの SAS ユーザー総会でも多くの報告が行われている (魚住・浜田, 2011; 魚住・浜田, 2012; 魚住ら, 2016a). SGPANEL プロシジャによる図 1 の詳細なプログラムは付録 A を参照されたい.

**SAS プログラム (詳細は付録 A を参照)**

```
ods graphics on;
ods output SurvivalPlot=lifetest00;
proc lifetest data=VALung plots=s method=breslow;
  time Time*Censor(1);
  strata Cell;
run;
data DummyCell; do CellC=1 to 4; output; end; run;
ods output SurvivalPlot=phreg00;
proc phreg data=VALung plots(overlay)=survival;
  class  CellC / param=glm;
  model Time*Censor(1) = CellC;
  baseline out=phregout00 survival=S covariates=DummyCell
   / nomean group=CellC;
run;
```

**出力結果**

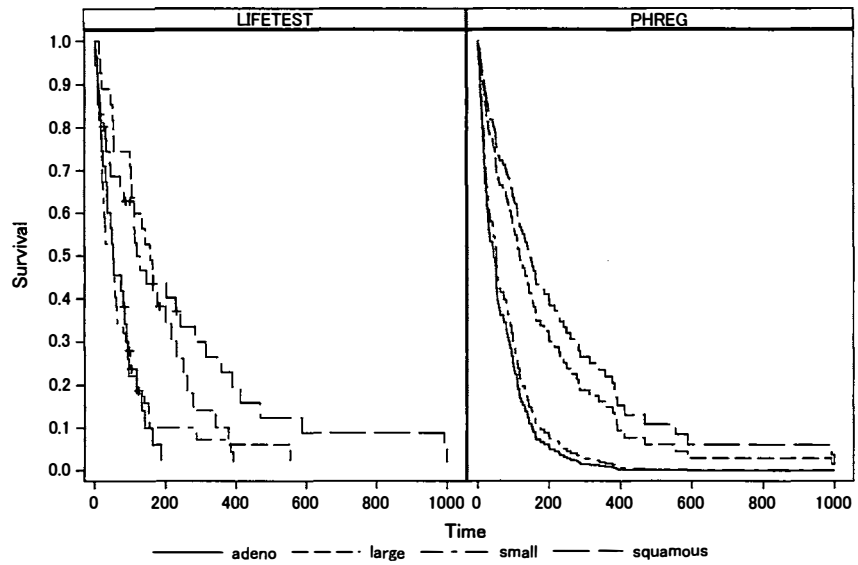



図 1 : LIFETEST プロシジャおよび PHREG プロシジャによる作成プラグラムと組織型別の生存時間曲線

図 1 における PHREG プロシジャのプログラムでは，BASELINE 文における OUT = オプション，あるいは ODS GRAPHICS の機能を用いた ODS OUTPUT SURVIVALPLOT = によって出力される．いずれのデータセットを用いても，同様の生存関数の推定結果が得られている．

図 1 の生存時間曲線は，LIFETEST および PHREG のいずれのプロシジャから作成しても，Br eslow 法による生存関数の推定を行っている．

ここで，$(T,D,X)$ を生存時間，イベントの有無，説明変数を表す確率変数として，$(t_i,d_i,x_i)$ はこれらの実現値を表し，$i=1,\ldots,n$ は互いに独立であることを仮定する．ただし，$d_i=1$ はイベント，$d_i=0$ は打ち切りを表す指示関数とする．さらに，$Y_i(t)=1(t_i\geq t)$ を個体 $i$ が時点 $t$ においてリスク集合に含まれることを表す指示関数，$N_i(t)=1(t_i\leq t,d_i=1)$ を個体 $i$ の時点 $t$ におけるイベントを表す指示関数とし，$\bar{Y}(t)=\sum_i Y_i(t)$ を時点 $t$ におけるリスク集合の大きさを表す指示関数，$\bar{N}(t)=\sum_i N_i(t)$ を時点 $t$ までの累積イベント数，$\Delta\bar{N}(t)$ を時点 $t$ におけるイベント数とする．

LIFETEST プロシジャでは，デフォルトの Kaplan-Meier 法による生存関数は

$$\hat{S}(t)=\prod_{t_i\leq t}\left(1-\frac{\Delta\bar{N}(t_i)}{\bar{Y}(t_i)}\right)$$

と推定される (Kaplan and Meier, 1958)．一方，Br eslow 法による生存関数は，Nelson-Aalen 法による累積ハザード関数 (Nelson, 1969; Aalen, 1978) を用いて，

$$\hat{S}(t)=\exp\left(-\int_0^t\frac{d\bar{N}(t_i)}{\bar{Y}(t_i)}\right)\approx\exp\left(-\sum_{t_i\leq t}\frac{\Delta\bar{N}(t_i)}{\bar{Y}(t_i)}\right)$$

と推定される．

一方，PHREG プロシジャを用いて Br eslow 法による推定を考える．説明変数ベクトル $X$ に対する比例ハザードモデル

$$\lambda(t\mid X)=\lambda_0(t)\exp(\beta^T X)$$

を考えると，ベースラインハザード関数が

$$\hat{\lambda}_0(t)=\frac{\Delta\bar{N}(t)}{\sum_i Y_i(t)\exp(\beta^T x_i)}$$

と推定される (Br eslow,1974)．このとき，累積ベースラインハザード関数は

$$\hat{\Lambda}_0(t)=\sum_{t_i\leq t}\hat{\lambda}_0(t_i)$$

となり，Br eslow 法による生存関数は

$$\hat{S}(t\mid x,\ \hat{\beta})=\exp(-\hat{\Lambda}_0(t)\exp(\hat{\beta}^T x))$$

と推定される．また，Kaplan-Meier 法による生存関数も

$$\hat{S}(t\mid x,\ \hat{\beta})=\prod_{t_i\leq t}\left(1-\exp(\hat{\beta}^T x)\hat{\lambda}_0(t_i)\right)$$

と同じように推定できる．

## 2.2 DIRADJ オプションを用いた生存関数の推定

図 1 の PHREG プロシジャでは，単変量の比例ハザードモデルに対して，BASELINE 文を用いて生存時間曲線を推定した．PROC PHREG 文において，PLOTS(OVERLAY) = SURVIVAL と指定しているものの，実際は 4 群が 1 つのプロットとして描かれない．しかし，BASELINE 文において DIRADJ オプションを指定すると，図 2 のように PHREG プロシジャの ODS GRAPHICS による機能として，4 群を 1 つのプロットに集約し

た生存時間曲線が出力される．なお，図 2 においては，単変量の比例ハザードモデルであることから，BASELINE 文において COVARIATES = を指定しなくても，図 1 と同様の出力が得られる．

BASELINE 文で DIRADJ オプションを指定して，図 2 のように 4 群を 1 つのプロットに集約した生存時間曲線が出力したが，DIRADJ では COVARIATES = で指定した変数で調整した生存関数を推定している．もし COVARIATES = を指定しないで実行した場合，GROUP = で指定した変数の各水準に対する調整生存関数として，GROUP = で指定していない変数の情報に基づいた重み付き生存関数が推定される．

**SAS プログラム**

```
ods output SurvivalPlot=phreg00diradj;
proc phreg data=VALung plots(overlay)=survival;
  class  Cell / param=glm;
  model Time*Censor(1) = Cell;
  baseline out=phregout00diradj survival=S
  / nomean diradj group=Cell;
run;
```

**出力結果**

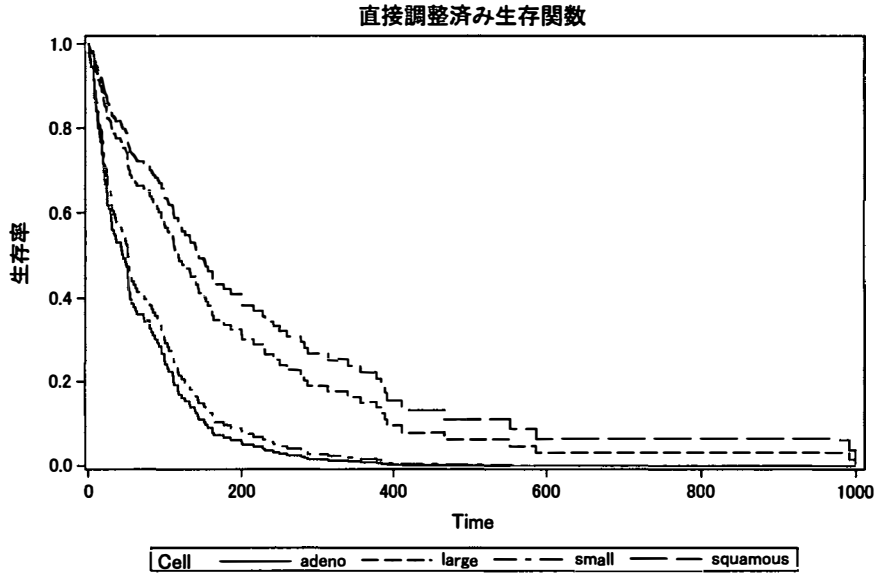



図 2 : DIRADJ オプションを用いた PHREG プロシジャによるプログラムと ODS GRAPHICS による出力

なお，もし図 2 のプログラムにおいて，BASELINE 文のオプションで METHOD = KM を指定しても，DIRADJ オプションは Breslow 法による生存関数の推定のみ対応しており，自動的に METHOD = BRESLOW と実行される．DIRADJ オプションは SAS/STAT 12.1 以降でサポートされている機能である (SAS Institute Inc., 2012).

ここで，わかりやすく比例ハザードモデルの説明変数として，組織型に加えて治療を表す 2 水準の変数も加えた場合で考え，各治療に対する調整生存関数を推定したいとする．このとき，DIRADJ および COVARIATES = をともに指定した場合は図 3 のプログラムとなる．

```
data Dummy; do CellC=1 to 4; do Therapy2=1 to 2; output; end; end; run;
proc phreg data=VALung plots(overlay)=survival;
  class CellC Therapy2 / param=glm;
  model Time*Censor(1) = CellC Therapy2;
  baseline out=out01 survival=S covariates=Dummy  / diradj nomean group=Therapy2;
run;
```

図 3 : DIRADJ および COVARIATES = オプションを用いた調整生存関数の推定プログラム

各治療 $i$ において組織型 $j$ 別に推定される (時点 $t$ における) 生存関数を $\hat{S}_{ij}$，例数を $n_{ij}$ とすると，表 1 のように表記できる．

表 1 : 各水準における推定生存関数

| | Cell ($j$) | | | |
|---|---|---|---|---|
| Treatment ($i$) | $\hat{S}_{11}$ | $\hat{S}_{12}$ | $\hat{S}_{13}$ | $\hat{S}_{14}$ |
| | $\hat{S}_{21}$ | $\hat{S}_{22}$ | $\hat{S}_{23}$ | $\hat{S}_{24}$ |

図 3 のプログラムでは，時点 $t$ における治療 $i$ に対する調整生存関数は

$$\bar{S}_{i\cdot} = \sum_j \hat{S}_{ij} / J$$

と推定されていることになる．ただし，$J$ は調整する組織型の水準数であり，今回のデータでは $J=4$ である．

```
proc phreg data=VALung plots(overlay)=survival;
  class CellC Therapy2/ param=glm;
  model Time*Censor(1) = CellC Therapy2;
  baseline out=out02 survival=S / diradj nomean group=Therapy2;
run;
```

図 4 : DIRADJ オプションを用いた重み付き調整生存関数の推定プログラム

一方，図 4 のプログラムでは，時点 $t$ における治療 $i$ に対する重み付き調整生存割合は

$$\tilde{S}_{i\cdot} = \sum_j n_{\cdot j} \hat{S}_{ij} / n_{\cdot\cdot}$$

と推定される．ただし，$n_{\cdot\cdot}$ はデータセット VALung の総例数，$n_{\cdot j}$ は組織型 $j$ における総例数である．同じ PHREG プロシジャ内の機能で例えるならば, LSMEANS 文における OM オプションに該当する内容といえる．データセット VALung の場合，組織型を表す変数の各水準における例数は adeno: $n_{\cdot 1} = 27$，large: $n_{\cdot 2} = 27$，small: $n_{\cdot 3} = 48$，squamous: $n_{\cdot 4} = 35$ となるため，時点 $t$ における治療 $i$ に対する重み付き調整生存割合は

$$\tilde{S}_{i\cdot} = \frac{27}{137}\hat{S}_{i1} + \frac{27}{137}\hat{S}_{i2} + \frac{48}{137}\hat{S}_{i3} + \frac{35}{137}\hat{S}_{i4}$$

となる．

# 3 LSMEANS 文とオプション機能

LSMEANS 文は GLM プロシジャや MIXED プロシジャにおいてサポートされていた機能で，PHREG プロシジャにおいては SAS/STAT 9.3 から新たに追加された (SAS Institute Inc., 2011). GLM プロシジャからの類推で LSMEANS「最小二乗平均」という言葉が使われているが，推定は最尤法で行われている (大橋ら, 2016). なお，LSMEANS 文ではすべての水準間の最小二乗平均の対比較が行われるため，対比較の数が多い場合は LSMESTIMATE 文や ESTIMATE 文で特定の線型仮説を明示的に指定することが有用である (浜田, 2013; 魚住, 2014).

## 3.1 LSMEANS 文による Diffogram の作成

ODS GRAPHICS ON の状態で LSMEANS 文を指定して PHREG プロシジャを実行すると，「最小二乗平均」の水準間の比較結果が出力される.『生存時間解析 応用編 –SAS による生物統計』の 3.2 節におけるプログラム 3.2.1 および出力結果を図 5 に示している．比例ハザードモデルの説明変数として，Kps，Cell，Prior，Therapy，Therapy と Prior の交互作用を含めている．なお，Kps はカルノフスキー (Karnofsky) の Performance Scale を表す連続変数，Prior は既往歴を表す 2 値カテゴリカル変数である.

**SAS プログラム**

```
ods graphics on;
ods select Diffs DiffPlot;
proc phreg data=VALung;
  class Prior Cell Therapy / param=glm;
  model Time*Censor(1) = Kps Cell Prior Therapy Prior*Therapy;
  lsmeans Cell / adjust=tukey cl exp;
run;
```

**出力結果**

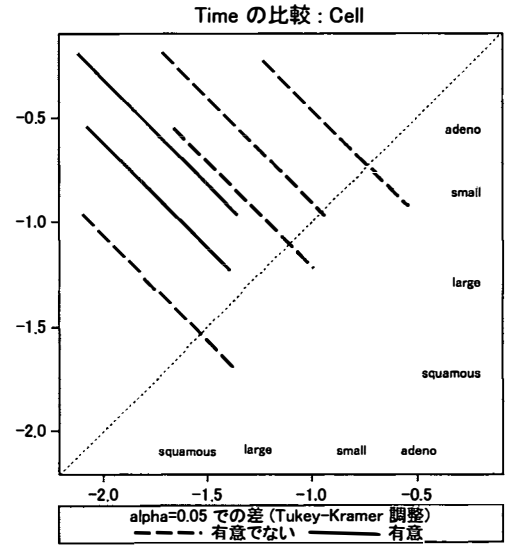



図 5 : LSMEANS 文を用いた Diffogram 作成プログラムとその出力

LSMEANS 文を指定して PHREG プロシジャを実行すると，水準間の対比較の結果が得られる．図 5 は ODS SELECT DIFFPLOT により出力の指定が行われており，Mean-Mean scatter plot あるいは Diffogram と呼ばれる (High, 2014)．LSMEANS 文で 4 水準の変数 Cell を指定したため，${}_4C_2$ = 6 通りの比較結果として，各群の最小二乗平均の 95%信頼区間として 6 本描かれている．対角線が帰無仮説を表しており，信頼区間と交わっているかで有意かどうか判定している．有意な対比較の信頼区間は実線，有意でない対比較の信頼区間は破線で示されている．

図 5 の Diffogram より，adeno-squamous「腺癌と扁平上皮癌」，small-squamous「小細胞癌と扁平上皮癌」の比較においては，Tukey-Kramer 法による多重性の調整を行って有意な結果が得られている (多重比較については，浜田 (2015) を参照されたい)．なお，この結果は比例ハザードモデルで組織型以外の変数で調整を行っているため，共変量および多重性の二重の調整を行っていることになる．

ここで，表 2 は図 5 のプログラムを実行して得られる出力結果で，ODS SELECT DIFFS の指定により出力された表である．しかし，「最小二乗平均の差」と出力されており，図 5 の Diffogram がどのように作成されたか疑問に思うのではないだろうか．

表 2 : LSMEANS 文による共変量を調整した推定結果

Cell の最小 2 乗平均の差

多重比較の調整 : Tukey-Kramer

| Cell | _Cell | 推定値 | 調整済下限値 | 調整済上限値 | 指数 | 調整済下限 Exp | 調整済上限 Exp |
|---|---|---|---|---|---|---|---|
| adeno | large | 0.7491 | -0.03358 | 1.5317 | 2.1150 | 0.9670 | 4.6261 |
| adeno | small | 0.3064 | -0.3906 | 1.0034 | 1.3586 | 0.6766 | 2.7277 |
| adeno | squamous | 1.1605 | 0.3898 | 1.9313 | 3.1916 | 1.4766 | 6.8983 |
| large | small | -0.4427 | -1.1149 | 0.2296 | 0.6423 | 0.3279 | 1.2581 |
| large | squamous | 0.4114 | -0.3158 | 1.1387 | 1.5090 | 0.7292 | 3.1227 |
| small | squamous | 0.8541 | 0.1632 | 1.545 | 2.3493 | 1.1773 | 4.6880 |

## 3.2 LSMEANS 文におけるオプション機能を用いた Diffogram の作成

図 6 は，LSMEANS 文において，図 5 よりも多くのオプションを指定したプログラム，およびその出力結果を示している．

まず，LSMEANS 文において MEANS オプションを指定することで，「最小二乗平均の差」の出力に加えて，表 3 の最小二乗平均の推定値を出力することができる．PHREG プロシジャの場合，対数ハザードが推定されていることを意味しており，「最小二乗平均の差」は対数ハザード比を推定していることに相当する．

次に，LSMEANS 文における PLOT = で ODS GRAPHICS の機能で出力させるグラフを明示的に指定している．PLOT = DIFF を指定しているので，図 5 に示したような Diffogram が作成される．PLOT = DIFF はデフォルトの指定である．さらに，PLOT = DIFF のオプションとして，NOABS および CENTER を指定している．図 5 の出力では，デフォルトである ABS オプションの機能により，最小二乗平均の推定値がすべて左上にプロットされるよう，縦軸・横軸のいずれを対照水準とするか対比較ごとに変動させて出力させていた．これ

に対して，NOABS を指定することで，縦軸・横軸のいずれを対照水準とするか固定して出力している．また，CENTER オプションを指定することで，最小二乗平均の推定値を○で特定して示している．

以上のように，MEANS オプションを指定することで，表 3 の最小二乗平均の推定値が出力されるため，図 5 および図 6 の出力で示した Diffogram がどのように作成されたか確認できるであろう．

なお，MEANS オプションも指定した下で，PLOT = DIFF の代わりに PLOT = MEANPLOT を指定すると，LSMEANS 文で指定した変数に対する各水準の最小二乗平均の推定値とその 95%信頼区間のグラフを作成できる (付録 B を参照されたい).

**SAS プログラム**

```
ods graphics on;
proc phreg data=VALung;
  class Prior Cell Therapy / param=glm;
  model Time*Censor(1) = Kps Cell Prior Therapy Prior*Therapy;
  lsmeans Cell / plot=diff(noabs center) means adjust=tukey cl exp;
run;
```

**出力結果**

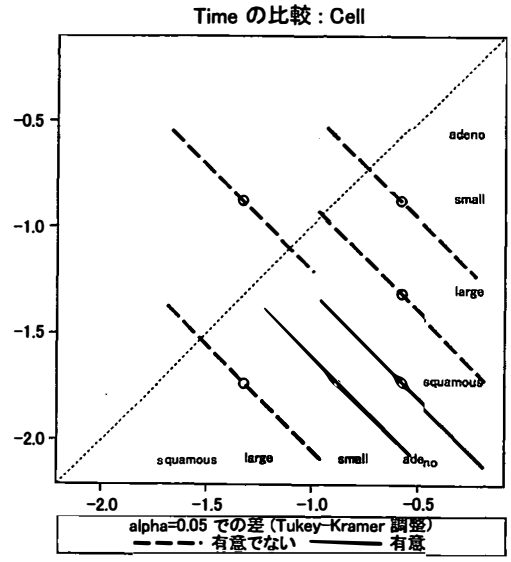



図 6 : LSMEANS 文のオプション機能を用いた Diffogram 作成プログラムとその出力

表 3 : LSMEANS 文による共変量を調整した推定結果

| Cell | 推定値 | 下限 | 上限 | 指数 | Exponentiated Lower | Exponentiated Upper |
|---|---|---|---|---|---|---|
| | Cell の最小 2 乗平均 | | | | | |
| adeno | -0.5778 | -1.5253 | 0.3697 | 0.5611 | 0.2176 | 1.4472 |
| large | -1.3269 | -2.2197 | -0.4341 | 0.2653 | 0.1086 | 0.6479 |
| small | -0.8843 | -1.8132 | 0.04472 | 0.413 | 0.1631 | 1.0457 |
| squamous | -1.7384 | -2.4919 | -0.9848 | 0.1758 | 0.08275 | 0.3735 |

## 3 まとめ

本稿では，比例ハザードモデルによる生存時間解析を行うためのプロシジャである PHREG プロシジャに関して，『生存時間解析 応用編 –SAS による生物統計』の刊行後に読者から寄せられた質問事項として，ベースライン生存関数を推定するための BASELINE 文，および固定効果に対する最小二乗平均を求める LSMEANS 文におけるオプション機能に着目した．BASELINE 文では，SAS/STAT 9.2 より DIRADJ オプションが追加されており，COVARIATE = オプションと併用することで調整生存関数の推定が可能である．COVARIATE = オプションを指定せずに DIRADJ オプションを指定すれば，調整する変数の各水準における例数で重み付けした調整生存関数の推定が行われていることに相当する．なお，ODS GRAPHICS ON の状態で，PROC PHREG 文において PLOTS(OVERLAY) = SURVIVAL および BASELINE 文で DIRADJ オプションを指定すれば，すべての群の生存関数の推定結果のグラフとして出力される．このような調整生存関数の推定は観察研究の論文報告でよく用いられており，例えば Yuan et al. (2017) において図として報告されている生存時間曲線が該当する．一方，ODS GRAPHICS ON の状態で，LSMEANS 文を指定して PHREG プロシジャを実行すると，水準間の対比較の結果を表す Diffogram が作成される．Diffogram の縦軸・横軸の値には，最小二乗平均の推定値 (PHREG プロシジャの場合は対数ハザードに相当する) が出力されており，LSMEANS 文において MEANS オプションを指定することで確認可能である．その他にも，Diffogram の解釈を役立てる機能として，LSMEANS 文における NOABS および CENTER の有用性を示した．

本稿で取りあげたすべての出力結果は，付録 A を含めた本稿に記述されている SAS プログラムで再現可能である．生存時間解析に従事する国内外の多くの実務家にとって，本稿の報告内容がお役に立てれば幸いである．

## 参考文献


[1] Aalen O. Nonparametric estimation of partial transition probabilities in multiple decrement models. *Annals of Statistics*. **6**:701–726, 1978.

[2] Breslow N. Covariance analysis of censored survival data. *Biometrics*. **30**:89–99, 1974.

[3] Cox DR. Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society, Series B*. **34**:187–220, 1972.

[4] High R. Plotting Differences among LSMEANS in Generalized Linear Models. *Proceedings of the SAS Global Forum*. SAS Institute Inc., Cary, NC, 2014. Available at http://support.sas.com/resources/papers/proceedings14/1902-2014.pdf.

[5] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. **53**:457–481, 1958.

[6] Klein JP, van Houwelingen HC, Ibrahim JG, Scheike TH. *Handbook of Survival Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 2013.

[7] Nelson W. Hazard plotting for incomplete failure data. *Journal of Quality Technology*. **1**:27–52, 1969.

[8] SAS Institute Inc. *SAS/STAT(R) 9.2 User's Guide* (2nd edn.). SAS Institute Inc., Cary, NC, 2009.

[9] SAS Institute Inc. *SAS/STAT(R) 9.3 User's Guide*. SAS Institute Inc., Cary, NC, 2011.

[10] SAS Institute Inc. *SAS/STAT(R) 12.1 User's Guide*. SAS Institute Inc., Cary, NC, 2012.

[11] SAS Institute Inc. *SAS/STAT(R) 14.2 User's Guide*. SAS Institute Inc., Cary, NC, 2016.

[12] Yuan C, Morales-Oyarvide V, Babic A, Clish CB, Kraft P, Bao Y, Qian ZR, Rubinson DA, Ng K, Giovannucci EL, Ogino S, Stampfer MJ, Gaziano JM, Sesso HD, Cochrane BB, Manson JE, Fuchs CS, Wolpin BM. Cigarette smoking and pancreatic cancer survival. *Journal of Clinical Oncology*. **35**:1822–1828, 2017.

[13] 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.

[14] 魚住龍史・浜田知久馬. がん臨床試験における腫瘍縮小効果の検討に有用なグラフの作成 –SGPLOT プロシジャの最新機能を活用–. SAS ユーザー総会 論文集 2012, 151–165.

[15] 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, 449-463.

[16] 魚住龍史・森田智視. 生存時間解析における三種の神器. 呼吸, **34**: 1083–1089, 2015.

[17] 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 2016a, 216-232.

[18] 魚住龍史・矢田真城・浜田知久馬. SAS プロシジャを用いた生存時間データに対する例数設計の変革. SAS ユーザー総会 論文集 2016b, 250-267.

[19] 魚住龍史. 有名論文から統計の基礎を学ぶ：COURAGE 試験における生存時間解析の側面. *Coronary Intervention*, **13**: 38–43, 2017.

[20] 大橋靖雄・浜田知久馬. 生存時間解析 –SAS による生物統計. 東京大学出版会, 1995.

[21] 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.

[22] 浜田知久馬. SAS 生存時間解析プロシジャの最新の機能拡張. SAS ユーザー総会 論文集 2013, 3-72.

[23] 浜田知久馬. SAS による多重比較「美女と野獣」の統計学. SAS ユーザー総会 論文集 2015, 367-388.

[24] 浜田知久馬・魚住龍史. SAS による生存時間解析の拡張機能. SAS Learning Session 2016a. 配付資料.

[25] 浜田知久馬・魚住龍史. SAS による生存時間分布の予測「Death Note の統計学」. SAS ユーザー総会論文集 2016b, 297.

[26] 矢田真城・浜田知久馬. ガンマ過程を用いた比例ハザードモデルによる解析. SAS ユーザー総会 論文集 2016, 100-109.

[27] 矢田真城・魚住龍史・浜田知久馬. 生存時間データに対するベイズ流例数設計. SAS ユーザー総会 論文集 2016, 268-278.

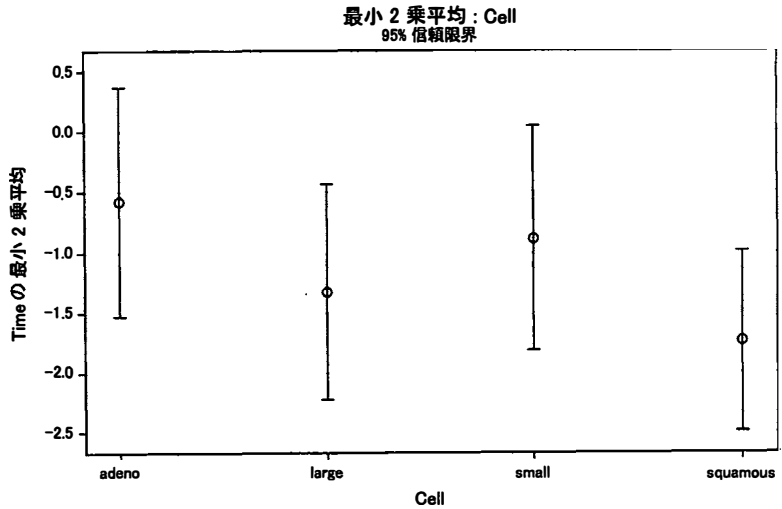
[28] 矢田真城・魚住龍史・浜田知久馬. 区分指数モデル及びフレイルティモデルに対するベイズ流生存時間解析. SAS ユーザー総会 論文集 2017.

## 連絡先

E-mail : uozumi@kuhp.kyoto-u.ac.jp

付録 A: 図 1 の SGPANEL プロシジャによる出力のためのプログラム詳細

```
*==================================================*;
* フォーマット ;
proc format;
  value Cellf 1='adeno' 2='large' 3='small' 4='squamous'; value procf 1='LIFETEST' 2='PHREG';
run;

*==================================================*;
* LIFETESTプロシジャによる出力結果 ;
*==================================================*;
ods graphics on;
ods listing close;
ods output SurvivalPlot=lifetest00;
proc lifetest data=VALung plots=s method=breslow;
  time Time*Censor(1); strata Cell;
run;

*==================================================*;
* PHREGプロシジャによる出力結果;
*==================================================*;
data DummyCell; do CellC=1 to 4; output; end; run;
ods output SurvivalPlot=phreg00;
proc phreg data=VALung plots(overlay)=survival;
  class  CellC / param=glm; model Time*Censor(1) = CellC;
  baseline out=phregout00 survival=S covariates=DummyCell / nomean group=CellC;
run;
ods listing;
ods graphics off;

*==================================================*;
* SGPANELプロシジャによる図1の出力 ;
*==================================================*;
data lifetest00;set lifetest00; proc=1; run;
data phreg00;set phreg00; proc=2; rename  Vector_=StratumNum; run;
data survival00; set lifetest00 phreg00; run;

ods html image dpi=400 style=journal;
ods graphics / reset noborder noscale width=600px height=400px imagefmt=png; ods listing close;
proc sgpanel data=survival00 noautolegend;
  panelby proc / novarname;
  step x=Time y=Survival / group=StratumNum name='Survival';
  scatter x=Time y=Censored / group=StratumNum markerattrs=(symbol=plus);
  rowaxis values=(0 to 1 by 0.1) label='Survival'; colaxis values=(0 to 1000 by 200) label='Time';
  keylegend 'Survival' / noborder position=bottom;  format StratumNum Cellf. proc procf.;
run;
ods listing; ods graphics off;
ods html close;
```




付録 B : LSMEANS 文のオプション機能を用いた最小二乗平均の推定結果のグラフ出力

# 区分指数モデル及びフレイルティモデルに対する ベイズ流生存時間解析

○矢田 真城[1]　魚住 龍史[1,2]　浜田 知久馬[1]
[1]東京理科大学大学院 工学研究科 経営 工学専攻
[2]京都大学大学院医学研究科 医学統計生物情報学

Bayesian survival analysis for piecewise exponential and frailty models

Shinjo Yada[1], Ryuji Uozumi[1,2], and Chikuma Hamada[1]
[1]*Department of Management Science, Graduate School of Engineering, Tokyo University of Science*
[2]*Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine*

## 要旨

SAS では，比例ハザードモデルによる生存時間解析を行うためのプロシジャとして，PHREG プロシジャが用意されている．Ver.9.2 より BAYES ステートメントが追加され，比例ハザードモデルによるベイズ流生存時間解析が可能となった．また，Ver.9.3 より RANDOM ステートメントが追加され，フレイルティモデルによる生存時間解析が可能となった．フレイルティモデルは，非独立な複数の生存時間データを解析するために，比例ハザードモデルを拡張したものである．比例ハザードモデルに対するベイズ流解析では，ハザードには事前分布を想定せず，部分尤度に基づいてパラメータ推定値を算出するアプローチと，ハザードにも事前情報を付与してパラメータ推定値を算出するアプローチがある．本稿では，区分指数モデルとフレイルティモデルに対するベイズ流生存時間解析についてまとめ，SAS による適用例を報告する．

キーワード：ベイズ流解析，生存時間解析，比例ハザードモデル，区分指数モデル，フレイルティモデル，PHREG プロシジャ，MCMC プロシジャ

## 1. はじめに

ある基準の時刻からある目的の反応（観測対象とする個体に対し一度だけ非再起的に起きる事象であるとする）が起きるまでの時間を解析対象とする解析手法の総称を，生存時間解析とよぶ (大橋・浜田, 1995; 大橋ら, 2016).

生存時間解析においてよく用いられる手法のひとつとして, Cox (1972)の比例ハザードモデルによる解析があげられる．比例ハザードモデルは，生存時間（基準ハザード関数）に特定の分布を仮定することなく共変量の効果を推定することから，セミパラメトリックなモデルとよばれる．SAS では，比例ハザードモデルを

行うためのプロシジャとして，PHREG プロシジャが用意されている．Ver.9.2 より，PHREG プロシジャに BAYES ステートメントが追加され(SAS Institute Inc., 2008)，比例ハザードモデルを用いて，ベイズ流に生存時間解析を行うことが可能となった．

比例ハザードモデルを用いてベイズ流に解析する場合，ハザード関数に特定の分布を仮定せず，回帰係数に事前分布を想定し，部分尤度に基づき共変量の効果を推定するアプローチがある(矢田ら, 2016)．これとは別に，ハザード関数を複数のパラメータでモデル化し，回帰係数及びハザード関数に対して事前情報を付与した上で，推論を行うアプローチがある．後者に属するアプローチのひとつとして，時間軸を幾つかの区間に分割し，各区間でハザードが一定であるとみなした，区分指数モデル(piecewise exponential model)による解析手法がある．

また，生存時間解析において，比例ハザードモデルのほかに用いられるセミパラメトリックなモデルとして，フレイルティモデル(frailty model)がある．フレイルティモデルは，非独立な複数の生存時間データを解析するために，比例ハザードモデルを拡張させたモデルである．SAS では，Ver.9.3 より，PHREG プロシジャに RANDOM ステートメントが追加され(SAS Institute Inc., 2011)，フレイルティモデルを用いた解析が実行可能となった．フレイルティモデルに対するベイズ流生存時間解析では，クラスターのフレイルティの分布を規定するパラメータに事前分布を想定し，部分尤度を用いて推論を行う．別のアプローチとして，ハザードをモデル化するパラメータにも事前分布を想定し，これらの同時分布を用いて推論を行う方法がある．

本稿では，区分指数比例ハザードモデルとフレイルティモデルに対するベイズ流生存時間解析をまとめ，SAS での適用例を紹介する．区分指数比例ハザードモデルについては，ハザード関数をモデル化するパラメータと回帰係数に事前分布を想定し，モデルパラメータ推定値を算出する方法を説明する．フレイルティモデルについては，回帰係数とフレイルティに事前分布を想定し，部分尤度に基づきモデルパラメータ推定値を算出する方法を説明する．最後に，フレイルティモデルについて，回帰係数とフレイルティに加えてハザードにも事前情報を与え，これらの同時確率分布を用いてモデルパラメータを推定する方法を示す．

## 2. 区分指数モデル

以下，解析対象となる時間を生存時間とよび，対象とする事象をイベントとよぶことにする．

$i$ 番目の被験者($i$=1,2,..,$n$)に対して，イベント発現時間あるいは打ち切り時間を $t_i$，打ち切りを表す変数を $v_i$（イベント発現のとき $v_i$=1，打ち切りのとき $v_i$=0)，$p$ 個の共変量を $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ と表し，$\mathbf{x}_i$ 所与のもとでのハザード関数を $\lambda_i(t \mid \mathbf{x}_i)$とおく．このとき，比例ハザードモデルでは，共変量 $\mathbf{x}_i$ をもつ被験者のハザード関数 $\lambda_i(t \mid \mathbf{x}_i)$に対し

$$\lambda_i(t \mid \mathbf{x}_i) = \lambda_0(t) \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) \tag{1}$$

を想定する．ここに $\lambda_0(t)$ は基準ハザード関数であり，$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ は共変量ベクトル $\mathbf{x}_i$ に対する回帰係数ベクトルである．また，共変量ベクトル $\mathbf{x}_i$ をもつ被験者の生存関数 $S(t \mid \mathbf{x}_i)$は，$\Lambda_0(t)$を累積基準ハザード関数として

$$S(t \mid \mathbf{x}_i) = \exp(-\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\, \Lambda_0(t)) \tag{2}$$

となる．

比例ハザードモデルに対するベイズ流アプローチとしては，回帰係数に対して事前分布を与え部分尤度を用いて推論を行う方法と，回帰係数のみならずハザード関数にも事前情報を与えて推論を行う方法とがある．前者の部分尤度を用いる方法については，黒田(2010)がその数理背景と PHREG プロシジャによる SAS プロ

グラムコードを紹介している．以下では，後者の方法として，時間軸を小さな区間に分け，区間ごとにハザードをモデル化することを考える．時間軸を，$0 = h_0 < h_1 < h_2 < \ldots < h_{J-1} < h_J$ と $J$ 個の区間$[0, h_1), [h_1, h_2), \ldots, [h_{J-1}, h_J)$に分割し，$t \in [h_{j-1}, h_j)$のとき $\lambda_0(t) = \lambda_j\ (j=1,2,\ldots,J)$であると考える．これは，時間軸を小さな区間に分割し，各区間内では基準ハザード関数が一定とみなすことを意味する．このとき，モデルパラメータは回帰係数 $\boldsymbol{\beta}$ と基準ハザード関数におけるパラメータ $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)^T$であり，累積基準ハザード関数 $\Lambda_0(t)$は

$$\Lambda_0(t) = \sum_{j=1}^{J} \lambda_j \Delta_j(t),$$

ここに

$$\Delta_j(t) = \begin{cases} 0 & (t < h_j) \\ t - h_j & (h_{j-1} \le t < h_j) \\ h_j - h_{j-1} & (t \ge h_j) \end{cases} \tag{3}$$

である．$n$ 人のデータ $D = \{t_i, v_i, \mathbf{x}_i\}_{i=1}^{n}$ が与えられたもとでの尤度は，ハザード関数 $\lambda_i(t \mid \mathbf{x}_i)$，生存関数 $S(t \mid \mathbf{x}_i)$に対応する確率密度関数を $f_i(t \mid \mathbf{x}_i)$として，

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid D) = \prod_{i=1}^{n} f_i(t_i \mid \mathbf{x}_i)^{v_i} S(t_i \mid \mathbf{x}_i)^{1-v_i} = \prod_{i=1}^{n} (\lambda_i(t_i \mid \mathbf{x}_i) S(t_i \mid \mathbf{x}_i))^{v_i} S(t_i \mid \mathbf{x}_i)^{1-v_i} = \prod_{i=1}^{n} \lambda_i(t_i \mid \mathbf{x}_i)^{v_i} S(t_i \mid \mathbf{x}_i)$$

と表すことができる．これに(1)から(3)まで代入し整理すれば

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid D) = \prod_{i=1}^{n} \prod_{j=1}^{J} (\lambda_i \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i))^{\delta_{ij} v_i} \exp\left[ -\delta_{ij} \{ \lambda_j (t_i - h_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (t_g - h_{g-1}) \} \right] \tag{4}$$

となる(Ibrahim et al., 2001)．ここに $\delta_{ij}$ は，$i$ 番目の被験者が $j$ 番目の区間$(h_{J-1}, h_J]$にてイベント発現あるいは打ち切りであれば 1，それ以外は 0 をとる関数である（$i = 1,2,.., n\,; j = 1,2,\ldots, J$）．(4)の形式で表現されるセミパラメトリックモデルは，区分指数モデルとよばれる(Ibrahim et al., 2001). $\boldsymbol{\beta}$ の事前分布として平均ベクトル $\boldsymbol{\mu}_0$，分散共分散行列 $\boldsymbol{\Sigma}_0$ の多変量正規分布を，$j$ 番目の区間における基準ハザード関数 $\lambda_j$ の事前分布として互いに独立に平均 $a_j / b_j$，分散 $a_j / b_j^2$ のガンマ分布（以下 $Ga(a_j, b_j)$と表記する）を想定すると，$\boldsymbol{\beta}$ と $\boldsymbol{\lambda}$ の同時事後分布 $p(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid D)$は

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid D) \propto L(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid D)\, p(\boldsymbol{\beta})\, p(\boldsymbol{\lambda})$$

で与えられる．ここに $p(\boldsymbol{\beta})$は $\boldsymbol{\beta}$ の事前分布

$$p(\boldsymbol{\beta}) \propto \exp\left( -(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_0) / 2 \right),$$

であり，$p(\boldsymbol{\lambda})$は $\boldsymbol{\lambda}$ の事前分布

$$p(\boldsymbol{\lambda}) \propto \prod_{j=1}^{J} \lambda_j^{a_j - 1} \exp(-b_j \lambda_j)$$

である．

PHREG プロシジャでは，BAYES ステートメントによりベイズ流生存時間解析を行うことができる．SAS プログラム 1 は，Gehan (1965)の急性白血病データ（以下「Gehan のデータ」とよぶ）に対し区間指数モデルをあてはめ，ベイズ流生存時間解析を行うための SAS プログラム例である．大橋ら(2016)に倣い，Gehan のデータから，再発または打ち切りを表す変数 Week，群を表す変数 Drug（プラセボ群であれば 0，6-MP 群であれば 1），打ち切りを表す変数 Remis（打ち切りであれば 0，イベント発生であれば 1）をもつデータセット Gehan を用意した上での実行を想定している．

SAS プログラム 1：区分指数モデルに対するベイズ流生存時間解析

```
data Covar ; do drug = 0,1 ; output ; end ; format drug drugf. ;run ;
proc phreg data = Gehan ;
  class drug ;
  model week*remiss(0) = drug ;
  hazardratio drug ;
  bayes seed = 1234 outpost = outsample nbi = 1000 nmc = 10000 thin = 2
    coeffprior = normal(var = 1e+3) statistics = (summary interval)
    piecewise = hazard(ninterval = 5 prior = gamma(shape = 1e-2 iscale =1e-2)) ;
    baseline covariates = covar out = out2 survival = Survival
      / nomean group = drug ;
  format drug drugf. ;
run ;
```

BAYES ステートメントにて，乱数のシードを 1234，事後分布からのモンテカルロ標本を保存する SAS データセット名を Outsample，burn-in 数を 1,000，生成させるモンテカルロ標本数を 10,000，モンテカルロ標本のサンプリング間隔を 2 と指定した．従って，5,000 個のモンテカルロ標本が Outsample に出力される．coeffprior = normal(var = 1e+3)と指定したことにより，回帰係数に対する事前分布として，平均 0，分散 1,000 の正規分布を与えたことになる．オプション PIECEWISE = が区間指数モデルをあてはめるためのキーワードである．piecewise = hazard と指定したことにより，各区間の基準ハザード関数に対して事前分布があてはめられる．PIECEWISE オプションにおいて，prior = gamma(shape =1e-2, iscale =1e-2))と指定したことにより，各区間の基準ハザード関数の事前分布として，平均 1，分散 100 のガンマ分布が用いられる．事後分布からのサンプリングアルゴリズムとして，適応的棄却メトロポリスサンプリング（adaptive rejection Metropolis sampling ; ARMS），ランダムウォークメトロポリスアルゴリズムのいずれかが使用可能であり，特に指定しない限り ARMS が適用される．ARMS は，Gilks et.al. (1995) により提案されたサンプリングアルゴリズムであり，メトロポリス - ヘイスティングアルゴリズムと適応的棄却サンプリング（Adaptive rejection sampling ; ARS）とを組み合わせた方法である．

出力結果 1 及び出力結果 2 は，SAS プログラム 1 により得られる出力結果から一部抜粋したものである．オプション PIECEWISE = において ninterval = 5 と指定したため，時間軸が 5 個の区間に分割され，基準ハザード関数は 5 個のパラメータによりモデル化される．出力結果 1 より，時間軸を[0,4.5), [4.5,6.5), [6.5,10.5), [10.5,15.5), [15.5,∞)の 5 つの区間に分割し，各区間の基準ハザード関数 $\lambda_j$ ($j$=1,2,…,5)に対して事後更新が行われたことが確認できる．区間数の代わりに分割したい区間を直接指定したい場合には，intervals = とすればよい． BAYES ステートメントにて statistics = (summary interval) と指定したことにより，事後分布からの要約統計量として，モデルパラメータごとに，モンテカルロ標本数，平均，標準偏差，パーセント点，95%信用区間，95%HPD（HPD : highest posterior density）信用区間が算出される（出力結果 2）．SAS プログラム 1 のように，BAYES ステートメントと HAZARDRATIO ステートメントとを同時に指定することで，ハザード比に関するベイズ流の推定値を出力することが可能である．OUTPUT 画面にて「ハザード比 drug」とタイト

ルされた出力結果から，対照群に対する 6-MP 群のハザード比の事後平均 0.1985, 95%信用区間(0.0764, 0.4049), 95%HPD 信用区間(0.0552, 0.3707)となることがわかる．

出力結果 1：区間指数モデルをあてはめたときの時間区切り

| ［下限, | 上限) | N | 事象 | ハザードパラメータ |
|---|---|---|---|---|
| 0 | 4.5 | 7 | 7 | Lambda1 |
| 4.5 | 6.5 | 6 | 5 | Lambda2 |
| 6.5 | 10.5 | 8 | 6 | Lambda3 |
| 10.5 | 15.5 | 7 | 6 | Lambda4 |
| 15.5 | Infty | 14 | 6 | Lambda5 |

出力結果 2：区分指数モデルに対するベイズ流生存時間解析 事後分布に関する要約統計量

| | | | パーセント点 | | | 信用区間 | | | |
|---|---|---|---|---|---|---|---|---|---|
| パラメータ | 平均 | 標準偏差 | 25% | 50% | 75% | 95% 信用区間 | | 95% HPD | |
| Lambda1 | 0.0712 | 0.0277 | 0.0514 | 0.0674 | 0.0870 | 0.0290 | 0.1366 | 0.0234 | 0.1256 |
| Lambda2 | 0.1545 | 0.0718 | 0.1029 | 0.1439 | 0.1940 | 0.0482 | 0.3283 | 0.0388 | 0.2983 |
| Lambda3 | 0.1209 | 0.0516 | 0.0841 | 0.1134 | 0.1502 | 0.0426 | 0.2408 | 0.0342 | 0.2217 |
| Lambda4 | 0.1737 | 0.0745 | 0.1202 | 0.1630 | 0.2169 | 0.0594 | 0.3471 | 0.0454 | 0.3201 |
| Lambda5 | 0.1752 | 0.0833 | 0.115 | 0.1619 | 0.2205 | 0.0556 | 0.3741 | 0.0457 | 0.3495 |
| drug6_MP | -1.7075 | 0.4307 | -1.9982 | -1.6939 | -1.4078 | -2.5724 | -0.9042 | -2.5516 | -0.8901 |

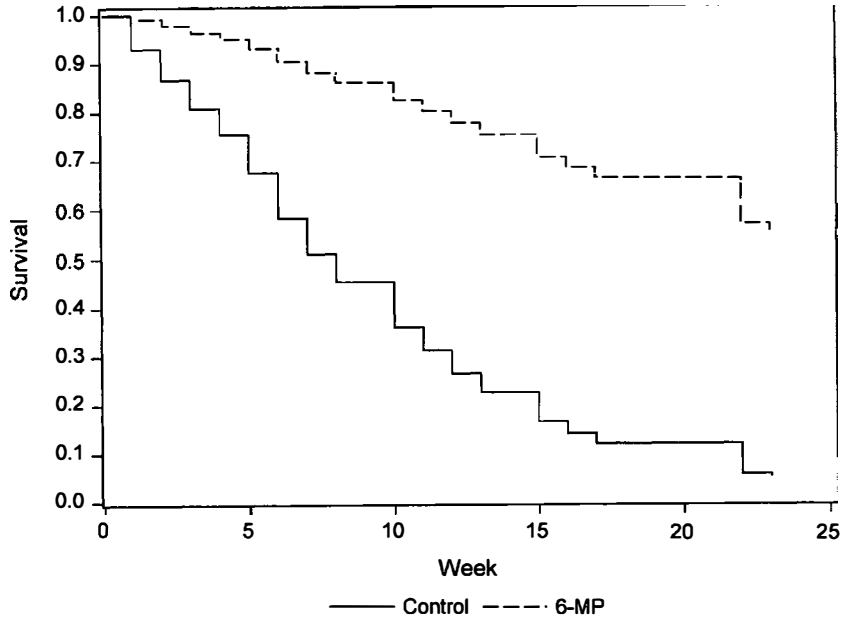



図 1：BASELINE ステートメントによる生存時間曲線

図 1 は，BASELINE ステートメントにより出力された予測生存時間曲線をプロットしたものである．データセット Out2 を用いて SGPLOT プロシジャにより作成した．BASELINE ステートメントにおいて，DIRADJ オプションを指定することで，GROUP = で指定した変数別に直接調整された生存時間曲線が可能である（大橋ら, 2016; 魚住ら, 2017）．しかし，ベイズ流生存時間解析を行う際には DIRADJ オプションが無効となるため, COVARIATES = では, 別途用意した群を表す共変量の値をもたせたデータセット Covar を指定している．

## 3. フレイルティモデル

フレイルティモデルは，比例ハザードモデルに変量効果をもたせたモデルである．例えば，眼に起こるある疾患をイベントとする生存時間データを解析したいとする．眼を個体とすれば，個体に生起するイベントは 1 つとなる．患者をクラスターと捉え，右眼と左眼でのイベントのおこりやすさは，クラスターの異質性に起因すると考える. クラスターの異質性を変量効果とみなし（生存時間解析ではフレイルティとよばれる），比例ハザードモデルに反映させる．

フレイルティモデルは，以下のように定式化される．

$i$ 番目のクラスターにおける $j$ 番目の個体に対して，$p$ 個の共変量を $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^{\mathrm{T}}$，ハザード関数を $\lambda_{ij}(t|\mathbf{x}_{ij})$とする($i = 1,2,\ldots, G$ ; $j=1,2,\ldots,n_i$)．　$i$ 番目のクラスターのフレイルティを $w_i$ として，基準ハザード関数 $\lambda_0(t)$を用いて

$$\lambda_{ij}(t|\mathbf{x}_{ij}) = \lambda_0(t)\, w_i \exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) \tag{5}$$

とモデル化する. (5)において $w_i$ が互いに独立に平均 1, 分散 $\theta$ のガンマ分布 $Ga(1/\theta,1/\theta)$を想定したモデルは，ガンマフレイルティモデルとよばれる．また，(5)において $\gamma_i = \log w_i$ とおけば

$$\lambda_{ij}(t|\, \mathbf{x}_{ij}) = \lambda_0(t) \exp(\gamma_i + \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) \tag{6}$$

となる．(6)において $\gamma_i$ が互いに独立に $N(0, \sigma^2)$に従うと想定したモデルを，対数正規フレイルティモデルとよぶ．$\boldsymbol{\beta}$ の事前分布を $p(\boldsymbol{\beta})$，$\boldsymbol{w} = (w_1, w_2,\ldots, w_G)^{\mathrm{T}}$ の分布を規定するパラメータ $\theta$ の事前分布を $p(\theta)$，パラメータ $\theta$ 所与のもとでの $\boldsymbol{w}$ の同時確率密度関数を $g(\boldsymbol{w}\,|\,\theta)$, 得られたデータ $D$ に対する部分尤度を $L_p(\boldsymbol{\beta}, \boldsymbol{w}\,|\,D)$と表すとき，モデルパラメータの事後分布 $p(\boldsymbol{\beta}, \boldsymbol{w}, \theta\,|\,D)$は

$$p(\boldsymbol{\beta}, \boldsymbol{w}, \theta\,|\,D) \propto L_p(\boldsymbol{\beta}, \boldsymbol{w}\,|\,D)\, g(\boldsymbol{w}\,|\,\theta)\, p(\boldsymbol{\beta})\, p(\theta) \tag{7}$$

で与えられる．

フレイルティモデルの適用例として，Lin (1994)の糖尿病性網膜症のデータをとりあげる．このデータ（以下「糖尿病性網膜症のデータ」とよぶ）は，糖尿病性網膜症患者に対するレーザー治療の効果を比較するため，失明までの時間を評価項目として，National Eye Institute で行われた試験の 197 名のデータであり，大橋ら (2016) の第 3 章や SAS/STAT(R) User's Guide, The PHREG Procedure においても用いられている．患者にはどちらかの眼にランダムにレーザー治療が割付けられ，もう片方の眼はレーザー治療なしの状態で，失明までの時間が観察された．本稿では，失明または打ち切りまでの時間を表す変数 Time，打ち切りを表す変数 Status (打ち切りであれば 0，イベント発生であれば 1)，レーザー治療あり・なしを表す変数 Treat (レーザー治療なしのとき 0，レーザー治療ありのとき 1)，基礎疾患の発症時期を表す変数 Type (発症時期が 20 歳未満なら 0，20 歳以上なら 1)をもつ SAS データセット Diab を用いる．

SAS プログラム 2：フレイルティモデルに対するベイズ流生存時間解析

```
proc phreg data = Diab ;
  class ID Treat Type ;
  model Time*Status(0) = Treat Type Treat*Type ;
  hazardratio Treat ;
  random ID / dist = lognormal ;
  bayes seed = 1234 outpost = outsample nbi = 1000 nmc = 10000 thin = 1
    coeffprior = normal(var = 1e+6) statistics = (summary interval)
    dispersionprior = igamma(shape = 1e-2, scale = 1e-2) statistics = (summary interval) ;
  format Type Typef. Treat Lazerf. ;
run ;
```

SAS プログラム 2 は，糖尿病性網膜症のデータに対し，治療法，基礎疾患の発症時期，治療法と基礎疾患の発症時期の交互作用を共変量とし，クラスター（被験者）のフレイルティを含めた，下記対数正規フレイルティモデルによるベイズ流生存時間解析を行うためのプログラム例である．

$\lambda_{ij}(t \mid x_{ij1}, x_{ij2}) = \lambda_0(t) \exp(\gamma_i + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij1} x_{ij2}) \quad (i = 1,2,\ldots,197; j = 1,2)$

$\beta_1, \beta_2, \beta_3 \sim N(0, 10^6)$

$\gamma_i \mid \theta \sim N(0, \theta)$

$\theta \sim IG(0.01, 0.01)$

ここに，$x_{ij1}, x_{ij2}$ は，$i$ 番目のクラスター（患者）の $j$ 番目の個体（右眼あるいは左眼）における治療法，基礎疾患の発症時期を表し，$\gamma_i$ が $i$ 番目のクラスター（患者）のフレイルティである．

RANDOM ステートメントにてクラスターを表す変数を指定し，オプション DIST = により分布を指定することで，クラスターがある特定の分布に従うものとして，モデルに反映することが可能となる．SAS プログラム 2 では，デフォルトである dist = lognormal と指定しており，(6)によって表される対数正規フレイルティモデルが設定される．なお，dist = gamma と指定すると，(5)によって表されるガンマフレイルティモデルが設定される．BAYES ステートメントにおいて，パラメータ $\theta$ の事前分布 $p(\theta)$を指定するオプションが DISPERSIONPRIOR = である．gamma(scale =$a$, iscale = $b$)とすれば $Gam(a,b)$を，igamma(scale = $a$, scale = $b$)とすれば平均 $b/(a\text{-}1)$，分散 $b^2/((a\text{-}1)^2(a\text{-}2))$の逆ガンマ分布（以下 $IG(a,b)$と表記する）を，それぞれ指定することができる．PHREG プロシジャでフレイルティモデルに対するベイズ流生存時間解析を行うと，事後分布からのサンプリングには，ランダムウォークメトロポリスアルゴリズムが適用される．

出力結果 3：フレイルティモデルに対するベイズ流生存時間解析 事後分布に関する要約統計量

| | | | パーセント点 | | | 信用区間 | | | |
|---|---|---|---|---|---|---|---|---|---|
| パラメータ | 平均 | 標準偏差 | 25% | 50% | 75% | 95% 信用区間 | | 95% HPD | |
| $\beta_1$ | -0.5056 | 0.2329 | -0.6637 | -0.5137 | -0.3444 | -0.9612 | -0.0571 | -0.9238 | -0.0332 |
| $\beta_2$ | 0.4437 | 0.2725 | 0.2786 | 0.4409 | 0.6145 | -0.1027 | 0.9984 | -0.1125 | 0.9708 |
| $\beta_3$ | -1.0518 | 0.3896 | -1.3135 | -1.0637 | -0.8009 | -1.7819 | -0.2413 | -1.7892 | -0.2686 |
| $\theta$ | 1.1576 | 0.4267 | 0.8682 | 1.1247 | 1.4134 | 0.2398 | 2.0801 | 0.1642 | 1.9070 |

出力結果 3 は，モデルパラメータの事後分布に関する要約統計量である．事後確率 $\Pr(\beta_l>0|D)$ ($l$=1,2,3)は，事後分布から生成された $R$ 個のモンテカルロ標本を $\beta_l^{(1)}, \beta_l^{(2)}, \ldots, \beta_l^{(R)}$ として

$$\Pr(\beta_l>0|D) = \frac{1}{R}\sum_{r=1}^{R} I(\beta_l^{(r)} > 0)$$

により求めることができる(Albert, 2009)．ここで，$I(\cdot)$はカッコの中が成り立てば 1 を，そうでなければ 0 をとる指示関数である．SAS プログラム 2 によって得られる，事後分布からのモンテカルロ標本を出力したデータセット Outsample を用いて算出すると，$\Pr(\beta_1>0|D) = 0.0069$, $\Pr(\beta_2>0|D) = 0.9408$, $\Pr(\beta_3>0|D) = 0.0004$ となった．OUTPUT 画面にて「Treat のハザード比」とタイトルされた出力結果から，レーザー治療なしに対するレーザー治療ありのハザード比は，20 歳以降に発症した場合，事後平均 0.2211, 95%信用区間(0.1163, 0.3893), 95%HPD 信用区間(0.1101, 0.3655)であるのに対し，20 歳未満で発症した場合，事後平均 0.6198，95%信用区間(0.3824, 0.9445)，95%HPD 信用区間(0.3658, 0.9066)であった．

上記は，ハザード関数には事前情報を付与せず，回帰係数にのみ事前情報を与えて，部分尤度に基づいて推論を行う方法である．比例ハザードモデルに対するアプローチと同様に，ハザード関数及び回帰係数に事前情報を付与し，これらの同時確率分布を用いて推論を行う方法がある．以降では，その一例として，ハザード関数に事前情報を付与するため，時間軸を小さな区間に分け，区間ごとにハザードをモデル化した，区分指数フレイルティモデル（piecewise exponential frailty model）について紹介する．

$i$ 番目のクラスターにおける $j$ 番目の個体に対して，イベントが観察されるまでの時間 $t_{ij}$，観察打ち切りまでの時間 $c_{ij}$ に対し，$u_{ij} = \min(t_{ij}, c_{ij})$, $\omega_{ij} = I(t_{ij} \leqq c_{ij})$として，$N_{ij}(t) = I(t_{ij} \leqq t, \omega_{ij}=1)$，$Y_{ij}(t) = I(t_{ij} \geqq t)$とおく．ここに $I(\cdot)$ はカッコの中が成り立てば 1 を，そうでなければ 0 をとる指示関数である．フレイルティモデルにおけるハザード関数 $\lambda_{ij}(t|\mathbf{x}_{ij})$を，強度関数 $I_{ij}(t)$でおきかえると

$$I_{ij}(t) = Y_{ij}(t)\, \lambda_0(t)\, w_i \exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$$

となる．打ち切りが non-informative であるとき，生存時間が互いに独立として尤度は

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i}\left[\prod_{t>0} I_{ij}(t)\right]^{dN_{ij}(t)} e^{-I_{ij}(t)dt} \tag{8}$$

となる．ここで，$dN_{ij}(t)$は微小時間$[t, t+dt)$に対する $N_{ij}$ の増分であり，$i$ 番目のクラスターにおける $j$ 番目の個体が $dt$ の間にイベントが生起すれば 1，そうでなければ 0 の，いずれかの値をとる．累積基準ハザード関数 $\Lambda_0(t)$の微小時間$[t, t+dt)$に対する増分を $d\Lambda_0(t)$と表すとき，微小時間$[t, t+dt)$に対する強度関数の増分 $I_{ij}(t)dt$ は

$$I_{ij}(t)dt = Y_{ij}(t)\, w_i \exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\, d\Lambda_0(t) \tag{9}$$

となり，尤度(8)は，(9)で表される $I_{ij}(t)dt$ を平均値とするポアソン分布の核 $[I_{ij}(t)dt]^{dNij(t)} \exp(-I_{ij}(t)dt)$ から構成されることがわかる．

時間軸を，$0 = h_0 < h_1 < h_2 < \ldots < h_{K-1} < h_K$ と $K$ 個の区間$[0, h_1), [h_1, h_2), \ldots, [h_{K-1}, h_K)$に分割し，$t \in [h_{k-1}, h_k)$のとき $\lambda_0(t) = \lambda_k$ ($k$=1,2,…, $K$)であると想定する．(8)，(9)の微小時間$[t, t+dt)$を区間$[h_{k-1}, h_k)$と置き換えることで，(8)により表される尤度を算出することができる．フレイルティモデルの場合，$\boldsymbol{\beta}$ の事前分布，基準ハザード $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)^T$ の事前分布のほか，$i$ 番目のクラスター($i$ = 1,2,…, $G$)のフレイルティにある特定の分布を想定し，その分布を規定するパラメータ $\theta$ の事前分布を与えることになる．

現時点では，PHREG プロシジャで区分指数フレイルティモデルを実行することはできない．ただし，DATA ステップを用いて入力データを加工することで，MCMC プロシジャによる実行が可能である．SAS/STAT(R) User's Guide, The MCMC Procedure では，糖尿病性網膜症のデータに対し，区分指数フレイルティモデル(10)

をあてはめてベイズ流生存時間解析を行うためのプログラムコード（SAS プログラム 3）が紹介されている．

$$\lambda_{ij}(t \mid x_{ij1}, x_{ij2}) = \lambda_0(t)\exp(\gamma_i + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij1}x_{ij2}) \quad (i = 1,2,\ldots,197;\ j = 1,2)$$

$$\lambda_k \sim Gam(0.01, 0.01) \quad (k = 1,2,\ldots,8)$$

$$\beta_1, \beta_2, \beta_3 \sim N(0, 10^6) \tag{10}$$

$$\gamma_i \mid \theta \sim N(0, \theta)$$

$$\theta \sim IG(0.01, 0.01)$$

SAS プログラム 3 では，時間軸を 0.1<6.545<13.95<26.47<38.8<45.88<54.35<62<80 と，8 つの区間[0.1, 6.545), [6.545, 13.95), [13.95, 26.47), [26.47, 38.8), [38.8, 45.88), [45.88, 54.35), [54.35, 62), [62, 80)に分割している．分割値（0.1, 6.545, 13.95, 26.47, 38.8, 45.88, 54.35, 62, 80）をもたせたデータセット Partition を用意する．データセット Partition と Diab から，197 個のクラスター（被験者）における 2 つの個体（右眼と左眼），合計 197×2＝394 個のデータに対し，各分割値でのリスクセットを変数 Y に，計数過程を変数 dN に，区間を識別する番号を変数 INT_INDEX に，それぞれ入力したデータセット Inputdata を用意する．その上で，MCMC プロシジャを用いてモデルパラメータの推論を行う．MCMC プロシジャでは，RANDOM ステートメントにより，クラスターのフレイルティ，基準ハザードに対して想定する分布を指定し，クラスターごと，区間ごとに，それぞれ互いに独立に分布をあてはめることを，オプション SUBJECT＝ を用いて設定している．

SAS プログラム 3：区分指数フレイルティモデルに対するベイズ流生存時間解析

```
data partition ;
input int_1-int_9 ;
datalines;
0.1 6.545 13.95 26.47 38.8 45.88 54.35 62 80
;
run ;
data _a ;set Diab ;
   if _n_ eq 1 then set partition ;
   array int[*] int_: ; array Y[8] ; array dN[8] ;
   do k = 1 to (dim(int)-1) ;
      Y[k] = (time - int[k] + 0.001 >= 0) ; dN[k] = Y[k] * ( int[k+1] - time - 0.001 >= 0) * status ;
   end ;
   output ;
   drop int_: k ;
run ;
data _b ;set _a ;
   array y[*] y: ; array dn[*] dn: ;
   do i = 1 to (dim(y)) ; y_val = y[i] ; dn_val = dn[i] ; int_index = i ;output ;end ;
   keep y_: dn_: type treat int_index id ;
run ;
data _b ;set _b ;rename y_val = Y dn_val = dN ;run ;
```

```
data inputdata ;set _b ;if Y > 0 ;run ;
proc mcmc data = inputdata nmc = 10000 outpost = postout seed = 12351 stats = summary diag = none ;
   parms beta1-beta3 0 s2 ;
   prior beta: ~ normal(0, var = 1e+6) ;
   prior s2 ~ igamma(0.01, scale = 0.01) ;
   random lambda ~ gamma(0.01, iscale = 0.01) subject = int_index ;
   random u ~ normal(0, var = s2) subject = id ;
   bZ = beta1*treat + beta2*type + beta3*treat*type + u ;
   idt = exp(bZ) * lambda ;
   model dN ~ poisson(idt) ;
run ;
```

出力結果 4：区分指数フレイルティモデルに対するベイズ流生存時間解析 事後分布に関する要約統計量

| | | | パーセント点 | | | 信用区間 | | | |
|---|---|---|---|---|---|---|---|---|---|
| パラメータ | 平均 | 標準偏差 | 25% | 50% | 75% | 95% 信用区間 | | 95% HPD | |
| $\beta_1$ | -0.4341 | 0.2290 | -0.5882 | -0.4274 | -0.2873 | -0.9034 | 0.0239 | -0.9658 | -0.0453 |
| $\beta_2$ | 0.3552 | 0.2194 | 0.2194 | 0.2069 | 0.5000 | -0.0529 | 0.8053 | -0.0536 | 0.8022 |
| $\beta_3$ | -0.3529 | 0.3477 | -1.0927 | -1.0927 | -0.6136 | -1.5162 | -0.1793 | -1.5600 | -0.2312 |
| $\theta$ | 0.3434 | 0.2063 | 0.1999 | 0.3137 | 0.4547 | 0.0422 | 0.8143 | 0.0267 | 0.7158 |

出力結果 4 は，SAS プログラム 3 を実行して得られる結果をまとめたものである（SAS プログラム 3 で，stats = summary を stats = all とすれば信用区間が得られる）．SAS プログラム 3 では，ハザード比は算出されないため，モンテカルロ標本から算出することになる．(10)では，レーザー治療なしのとき $x_{ij1} = 0$，レーザー治療なしのとき $x_{ij1} = 1$，発症時期 20 歳未満のとき $x_{ij2} = 0$，発症時期 20 歳以上のとき $x_{ij2} = 1$ とコーディングしているので，20 歳以降に発症した場合でのレーザー治療なしに対するレーザー治療ありのハザード比 $HR_1$ は

$$\frac{\exp(\beta_1 \times 1 + \beta_2 \times 1 + \beta_3 \times 1 \times 1)}{\exp(\beta_1 \times 0 + \beta_2 \times 1 + \beta_3 \times 0 \times 1)} = \exp(\beta_1 + \beta_3)$$

となる．従って，事後分布からの $R$ 個のモンテカルロ標本を $\beta_l^{(1)}, \beta_l^{(2)}, \ldots, \beta_l^{(R)}$ として($l = 1,2,3$)，ハザード比 $HR_1$ の事後平均は，$\exp(\beta_1^{(1)}+\beta_3^{(1)}), \exp(\beta_1^{(2)}+\beta_3^{(2)}), \ldots, \exp(\beta_1^{(R)}+\beta_3^{(R)})$の平均値として求めることができ，95%信用区間はこれら $R$ 個のモンテカルロ標本の 2.5%点と 97.5%点として求めることができる．一方，20 歳未満で発症した場合でのレーザー治療なしに対するレーザー治療ありのハザード比 $HR_2$ は

$$\frac{\exp(\beta_1 \times 1 + \beta_2 \times 0 + \beta_3 \times 1 \times 0)}{\exp(\beta_1 \times 0 + \beta_2 \times 0 + \beta_3 \times 0 \times 0)} = \exp(\beta_1)$$

となる．よってハザード比 $HR_2$ の事後平均は $\exp(\beta_1^{(1)}), \exp(\beta_1^{(2)}), \ldots, \exp(\beta_1^{(R)})$の平均値として，95%信用区間はこれら $R$ 個のモンテカルロ標本の 2.5%点と 97.5%点として，それぞれ求めることができる．SAS プログラム 3 により作成されるデータセット Outpost を用いて算出すると，20 歳以降に発症した場合でのレーザー治療なしに対するレーザー治療ありのハザード比 $HR_1$ の事後平均 0.2866，95%信用区間(0.1587,0.4645)，20 歳未満に発症した場合でのレーザー治療なしに対するレーザー治療ありのハザード比 $HR_2$ の事後平均 0.6650, 95%

信用区間(0.4052, 1.0242)であった.

## 4. まとめ

生存時間解析の手法として，比例ハザードモデルとフレイルティモデルをとりあげ，ベイズ流アプローチについて紹介した．比例ハザードモデルについては，時間軸をいくつかの区間に分割することでハザードを複数のパラメータで表現する，区分指数モデルをとりあげた．区分指数モデルは，PHREG プロシジャの BAYES ステートメントにおいて，PIECEWISE オプションを用いることで実行可能である．なお，区分指数モデルはパラメトリックに例数設計を行う際にも有用である (魚住ら, 2009). フレイルティモデルに対して，部分尤度に基づく推論を行いたい場合には，PHREG プロシジャの BAYES ステートメントにおいて，COEFFPRIOR オプションと DISPERSIONPRIOR オプションを用いることで実行可能である．一方，フレイルティモデルにおいて，ハザード関数にも事前情報を与えて解析を行うというアプローチも考えられる．その一例として，区分指数フレイルティモデルをとりあげた．区分指数フレイルティモデルは，計数過程に対応させた形式に加工した入力データを用いて，MCMC プロシジャの RANDOM ステートメントにより実行することができる.

ベイズ流生存時間解析には，これ以外にも多数の手法が存在する．本稿で説明した区分指数モデルでは，各区間での基準ハザード関数は互いに独立と仮定している．しかし，隣接する区間でのハザードは同じような傾向をもつと想定し, 基準ハザード関数に対して相関をもつ事前分布を適用することも考えられる. また, 累積ハザードに対して事前情報を付与する方法も考えられる．ガンマ過程 (Kalbfleisch,1978; Clayton,1991), 拡張ガンマ過程(Dykstra and Laud,1981), ベータ過程 (Hjort,1990)といった確率過程を用いた累積ハザード関数のモデル化が該当する．事前確率過程を用いる場合，確率密度関数を明示的に導出できないため，推論はシミュレーション実験に基づくものとなる．このため，各モデルに応じた事後確率過程からのサンプリングアルゴリズムが開発・研究されている．ガンマ過程を用いたベイズ流生存時間解析は，MCMC プロシジャを用いて実装が可能である (矢田・浜田, 2016).

本稿では，生存時間データに対するベイズ流解析のうち， PHREG プロシジャや MCMC プロシジャの既存の機能で実行可能な手法について説明した．PHREG プロシジャや MCMC プロシジャは，SAS/ STATとして提供されている．SAS/ STAT は毎年のようにバージョンアップされ，そのたびに新規プロシジャの追加，既存プロシジャでの新機能搭載がなされている (大橋ら, 2016; 浜田・魚住, 2016; 魚住ら, 2016). 今後，SAS/ STAT の機能拡張が行われていく中で，ベイズ流解析の手法も組み込まれ，ベイズ流アプローチがより身近になることを期待したい.

## 参考文献


[1]大橋靖雄・浜田知久馬 (1995). 生存時間解析 –SAS による生物統計. 東京大学出版会.

[2]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.

[3]Cox,D.R.(1972).Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B*. **34**:187–220.

[4]SAS Institute Inc.(2008). *SAS/STAT(R) 9.2 User's Guide*, Cary, NC, USA: SAS Institute Inc.

[5]矢田真城・魚住龍史・浜田知久馬 (2016). 生存時間データに対するベイズ流例数設計. SAS ユーザー総会 論文集, 268-278.

[6]SAS Institute Inc. (2011). *SAS/STAT(R) 9.3 User's Guide*, Cary, NC, USA: SAS Institute Inc.

[7]黒田普吾 (2010). Cox 回帰における Bayes 推定と PHREG プロシジャ.SAS ユーザー総会 論文集,23-44.

[8]SAS Institute Inc. (2016). *SAS/STAT(R) 14.2 User's Guide*, Cary, NC, USA: SAS Institute Inc.

[9]Ibrahim, J.G., Chen,M.H., Sinha,D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York.

[10]Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrary single-censored samples. *Biometrika*. **52**: 203–223.

[11]Gilks, W. R., N. G. Best, and K. K. C. Tan. (1995). Adaptive Rejection Metropolis Sampling with Gibbs Sampling. *Journal of the Royal Statistical Society*, **44**: 455-472.

[12]魚住龍史・矢田真城・浜田知久馬 (2017). PHREG プロシジャにおける共変量調整解析に関連したオプション機能. SAS ユーザー総会 論文集.

[13]Lin, D. Y. (1994). Cox Regression Analysis of Multivariate Failure Time Data: The Marginal Approach. *Statistics in Medicine*, **13**: 2233–2247.

[14]Albert,J. Bayesian Computation with R. Springer-Verlag, 2009. 石田基広・石田和枝(訳). R で学ぶベイズ統計学入門. 丸善出版, 2012.

[15]魚住龍史・水澤純基・浜田知久馬 (2009). 生存時間解析における Lakatos の症例数設計法の有用性の評価. SAS ユーザー総会 論文集, 143–152.

[16]Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society: Series B*. **40**:214–221.

[17]Clayton, D. G. (1991). A Monte Carlo Method for Bayesian Inference in Frailty Models. *Biometrics*, **47**:467–485.

[18]Dykstra,R.L. and Laud,P.W. (1981). A Bayesian nonparametric approach to reliability. *The Annals of Statistics*. **9**, 356-367.

[19]Hjort,N.L. (1990). Nonparametric Bayes estimatiors based on beta processes in models for life history data. *The Annals of Statistics*. **18**, 1259-1294.

[20]矢田真城・浜田知久馬 (2016). ガンマ過程を用いた比例ハザードモデルによる解析. SAS ユーザー総会 論文集, 100-109.

[21]浜田知久馬・魚住龍史 (2016). SAS による生存時間分布の予測「Death Note の統計学」. SAS ユーザー総会論文集, 297.

[22]魚住龍史・矢田真城・浜田知久馬 (2016). SAS プロシジャを用いた生存時間データに対する例数設計の変革. SAS ユーザー総会 論文集, 250-267.

## 連絡先

E-mail: *yada-s@a2healthcare.com*

ユーザー総会2017　抄録原稿

# 申請電子データプロセスの効率化～メタデータ・社内ツールを中心に

○飯島　康浩，○梶原　正幹，○小南 真由美，
中嶋 優一，北原 孝志
（ノバルティスファーマ株式会社）

■抄録本文

PMDA (Pharmaceuticals and Medical Devices Agency) は2016年10月より電子データ提出の受領を開始した。2016年11月から2品目の電子データを提出した経験より社内プロセスの標準化や社内ツールの有用性を実感した。本発表では，ノバルティスのデータ標準や申請電子データを準備する為のメタデータ， annotationを効率化するツールやDefine.xmlの確認ツールを紹介する。

Part1.

メタデータの導入は，CDISC標準を前提としたデータの更新・修正履歴の管理，試験間で運用するデータの整合性の確保，さらにはデータ作成プログラムのテンプレート化など様々なメリットが考えられる。ノバルティスでは2009年より，臨床試験データ管理を目的としたメタデータの利用し，症例報告書（Case Report Form）で取得されたデータから申請電子データまでをEnd to endに管理している。本Partではノバルティスのメタデータを紹介し，さらにはそれを管理するガバナンス体制を紹介する。

Part2.

注釈つき症例報告書（Annotated CRF）は，FDA (U.S. Food and Drug Administration)およびPMDAへの新薬の承認申請では提出が必須である。Annotated CRFには， CDISC標準のモデルの一つであるStudy Data Tabulation Model（SDTM）に格納される変数名および統制用語が表示され，これらの情報を手作業で付与する場合，Data Managerや統計解析担当者にとって大きな負担となることが考えられる。本Partでは，SASを用いて効率的に注釈付けを行う方法を紹介する。

Part3.

Define.xmlはSDTM，SEND，ADaM提出データセットのメタデータを記述するためのODMに基づくXMLスキーマ仕様である。多くの申請では各臨床試験及びそれらの併合データにもDefine.xmlが必要であるため，それら複数のDefine.xmlの整合性を確認することは多大な労力がかかる。本Partでは，これらをより簡便に確認できるようにSASマクロを使用したDefine.xml確認ツールを紹介する。

メタデータの活用，SASプログラムを用いたツールの利用は作業を包括的に標準化，効率化でき，解析プログラマーや解析担当者のリソースの負担を低減できる。標準化と簡便化は電子データ提出の経験から重要な要素の1つであると考える。

# SASによるベイズ流試験の成功確率の計算

○張　方紅　北原　孝志
（ノバルティスファーマ）

製薬業界で研究開発コストが高騰している中で，試験の成功確率を評価することが求められてきた．頻度論での古典的な検出力(Classical power)は，薬の効果を表すパラメータを未知の固定値とみなし，対立仮説としてある値を想定し，薬の効果がなしという帰無仮説を棄却して，効果がありと判断する確率となる．古典的な検出力はパラメータの特定な値に依存していることから，条件付検出力(conditional power)と言い換えている(Brutti, et. al., 2009). 条件付検出力は対立仮説で想定したパラメータの値に完全に依存することとパラメータの不確実性を考慮していないという指摘を受けている(Spiegelhalter et. al., 2004, Brutti, et. al., 2009).想定したパラメータの値が真であることを保障することができなく，その値が楽観過ぎるかもしれない．　一方，ベイズ接近法では，関心のあるパラメータを確率変数とみなし，その不確実性を事前分布p(θ)で表現する．古典的な条件付検出力に対して，事前分布p(θ)で積分して得られた無条件確率が提案され，平均検出力(expected power, average power)，または，Assuranceと名つげている(Spiegelhalter et. al., 2004；O'Hagan, et. al., 2005). Assuranceは薬の効果を表すパラメータの不確実性を考慮していることで，よりロバストな試験の成功確率を評価することができる．試験の成功確率として，Assuranceは海外で既に開発計画段階で利用され，日本では石田・川崎・土居は2017年の第4回データサイエンスラウンドテーブル会議で話題として提起されているが，まだ実例はあまり見られない．そこで，本発表は，2値データを評価項目と想定し，薬の効果があるかどうかの判定基準と成功確率の計算はそれぞれ頻度論とベイズ流があるということで，あわせて4つの場合，試験の成功確率を計算する方法を示す．また，実務者のため，SASプログラムを与える．

【参考文献】

Brutti1 P, Santis FD, Gubbiotti S. Mixtures of prior distributions for predictive Bayesian sample size calculations in clinical trials. Statist. Med. 2009; 28:2185–2201.

Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Wiley: Chichester, 2004.

O'Hagan A, Stevens J. and Campbell J. Assurance in clinical trial design. Pharmaceutical Statistics. 2005; 4: 187–201.

石田 弘輔, 川崎 洋平, 土居 正明 (2017) 第4回 データサイエンスラウンドテーブル会議 テーマ4 https://www.pmda.go.jp/files/000217616.pdf

# SASによる傾向スコアマッチング

○魚住 龍史[1] 矢田 真城[2] 山本 倫生[3] 川口 淳[4]
[1]京都大学大学院医学研究科 医学統計生物情報学
[2]東京理科大学大学院 工学研究科 経営工学専攻
[3]岡山大学大学院 環境生命科学研究科
[4]佐賀大学医学部附属 地域医療科学教育研究センター

Propensity score matching using SAS

Ryuji Uozumi[1], Shinjo Yada[2], Michio Yamamoto[3], and Atsushi Kawaguchi[4]
[1]*Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine*
[2]*Department of Management Science, Graduate School of Engineering, Tokyo University of Science*
[3]*Graduate School of Environmental and Life Science, Okayama University*
[4]*Faculty of Medicine, Saga University*

## 要旨

臨床研究では，ランダム化比較試験によって得られたデータの場合，群間比較を行う際の比較可能性が担保される．しかし，ランダム化比較試験でなく，観察研究によって得られたデータの場合，比較可能性が担保されず，群間比較としてはバイアスが混在した結果となってしまう場合がある．比較可能性を担保する方法のひとつとして，傾向スコアによるマッチングがよく行われている．これまで，SAS/STAT で提供されている LOGISTIC プロシジャで傾向スコアを算出して，マッチングに関しては DATA ステップでプログラムを構築しなければならなかった．しかし，SAS/STAT 14.2 (SAS 9.4 TS1M4) では，新たに PSMATCH プロシジャが追加され，傾向スコアによる分析を行うためのさまざまなツールが提供された．本稿では，PSMATCH プロシジャを用いて，傾向スコアによるマッチングを行う方法に関して解説を行い，JMP や R を用いて実施する場合との比較に関しても報告する．

キーワード：観察研究　傾向スコア　平均治療効果　ATE　ATT　マッチング　最近傍マッチング　PSMATCH　キャリパー　PSMODEL　MATCH　ASSESS　ODS GRAPHICS　箱ひげ図　度数分布　クラウドプロット

## 1　はじめに

臨床研究では，ランダム化比較試験によって得られたデータの場合，群間比較を行う際の比較可能性が担保される．しかし，ランダム化比較試験でなく，観察研究 (observational study) によって得られたデータの場

合，比較可能性が担保されず，群間比較としてはバイアスが混在した結果となってしまう．比較可能性を担保する方法として，マッチング，層別解析，および回帰モデルによる解析などが挙げられる (Faries et al., 2010).

近年，観察研究において傾向スコアによる解析がよく行われている．著者らも，臨床研究者から傾向スコアマッチングや重み付きの調整解析に関する統計相談に対応した経験を有している．本稿では，傾向スコアマッチング (propensity score matching) による調整解析に焦点を当てることとする．観察研究において，試験治療と対照治療の比較を行うと仮定すると，背景因子で条件付けたもと，試験治療を受ける確率が傾向スコアである．SAS/STAT では，LOGISTIC プロシジャにおける OUTPUT ステートメントのオプションを指定すれば，ロジスティックモデルによる予測確率が出力される．すなわち，応答変数として治療を表す変数を指定すれば，傾向スコアを出力させることが可能である．しかし，傾向スコアによるマッチングを行う場合，SAS/STAT のプロシジャで処理することはできず，DATA ステップでマッチングを行うためのプログラムを構築しなければならなかった．このため，傾向スコアによるマッチングを実施している臨床研究では，SAS 以外の統計解析ソフトとして，例えば R のパッケージ "Match" を用いた結果が報告されている (Hamatani et al., 2016; Hida et al., 2017). また，JMP のアドインを用いた傾向スコアによるマッチング解析の結果も報告されている (Yamashita et al., 2017). しかし，JMP のアドインを用いたマッチングでは，JMP スクリプト言語を用いない限りプログラムによる効率化ができない．加えて，JMP のアドインによる方法では，マッチングの条件指定の定義が特殊であることに留意しなければならない．したがって，SAS ユーザーであれば，SAS を用いて傾向スコアマッチングを行うことができれば，より効率的に実施できるといえる．

SAS/STAT 14.2 (SAS 9.4 TS1M4) では，新たに PSMATCH プロシジャが追加され，傾向スコアによる分析を行うためのさまざまなツールが提供された (SAS Institute Inc., 2016). 本稿では，PSMATCH プロシジャを用いて，傾向スコアによるマッチングを行う方法に関して解説を行い，JMP や R を用いて実施する場合との比較に関しても報告を行う．

## 2 傾向スコアとマッチング

本節では，ランダム化比較試験および観察研究における平均治療効果の推定について考えたもとで，傾向スコアおよびマッチングに関して解説する．

### 2.1 観察研究における平均治療効果の推定

2 値の指示変数 $Z$ として，$Z=1$ を治療群，$Z=0$ を対照群とする．さらに，潜在的結果変数 (potential outcome) として，治療群であった場合の結果変数 $Y(1)$，対照群であった場合の結果変数 $Y(0)$ のいずれかをとり，観察される結果変数は

$$Y = ZY(1) + (1-Z)Y(0) = \begin{cases} Y(0) & (Z=0) \\ Y(1) & (Z=1) \end{cases}$$

となる．このとき，

$$\mathrm{ATE} = E[Y(1) - Y(0)] \tag{1}$$

は平均治療効果 (average treatment effect; ATE) は呼ばれ，

$$\mathrm{ATT} = E[Y(1) - Y(0) \mid Z=1] \tag{2}$$

は治療群に対する平均治療効果 (average treatment effect on the treated; ATT) と呼ばれる．

ランダム化臨床試験のもとでは，$\{Y(1), Y(0)\} \perp Z$ であり，ATE は ATT に一致する．このとき，観測データから，ATE および ATT は $E[Y(1) \mid Z=1] - E[Y(0) \mid Z=0]$ と表せ，平均治療効果の不偏推定を可能とする (Lunceford and Davidian, 2004).

一方，観察研究のもとでは，$\{Y(1), Y(0)\} \perp Z$ が成り立たず，ATE は ATT と異なる．さらに，結果変数は直接的に治療効果を推定できない．ここで，ATT は，

$$\begin{aligned}\mathrm{ATT} &= E[Y(1) - Y(0) \mid Z=1] \\ &= E[Y(1) \mid Z=1] - E[Y(0) \mid Z=0] + E[Y(0) \mid Z=0] - E[Y(0) \mid Z=1]\end{aligned}$$

と分解することができ，$E[Y(1) \mid Z=1] - E[Y(0) \mid Z=0]$ は推定可能である一方，選択バイアスである $E[Y(0) \mid Z=0] - E[Y(0) \mid Z=1]$ は推定不能である．観察研究において，治療効果を正しく推定する方法として，傾向スコアに基づく調整解析が挙げられる．

## 2.2 傾向スコアとモデルによる推定

$X_i$ を被験者 $i \in \{1, 2, \ldots, N\}$ において観測される共変量ベクトル，$Z_i$ を被験者 $i$ の群を表す指示変数とすると，$X_i$ で条件付けたもとで治療群である確率 $e_i = \Pr(Z_i = 1 \mid X_i)$ が傾向スコアと定義される (Rosenbaum and Rubin, 1983)．2 群間で割付比が 1 : 1 のランダム化比較試験の場合，傾向スコアは $e_i = 1/2$ となる．また，傾向スコアはバランススコア (balancing score) とも呼ばれており，傾向スコアで条件付けたもとで，各群で観察される共変量の分布が似通っていることが望まれる ($Z_i \perp X_i \mid e_i$).

傾向スコアは，治療群を表す変数を応答変数としたロジスティックモデルから推定され，SAS/STAT では LOGISTIC プロシジャによって算出可能である．ここで，群を表す変数を Z (Z = 1 : 治療群，Z = 0 : 対照群)，調整する共変量として，性別，年齢，BMI を表す変数をそれぞれ Gender，Age，Bmi とする．このとき，プログラム 1 を実行させることで傾向スコア $e_i$ を推定することが可能であり，傾向スコアを表す変数 PS を含んだデータセットが出力される．傾向スコアを算出するためのロジスティックモデルは「傾向スコアモデル (propensity score model)」と呼ばれる．

プログラム 1 : LOGISTIC プロシジャによる傾向スコアの算出プログラム例

```
proc logistic data=<入力データセット名>;
  class Gender;
  model z(event='1')= Gender Age Bmi;
  output out=<出力データセット名> pred=ps;
run;
```

なお，傾向スコア推定の際のロジスティックモデルの説明変数として，どのような変数を指定するかについてはコンセンサスが取れていない．Ronsenbaum (2010) は，「傾向スコアによるマッチングでバランスを取りたい変数で考えるべきである．」と述べている．Austin et al. (2007) は，「ランダム化比較試験と同じように被験者背景の表を群別に作成し，アンバランスなベースラインの共変量を選択する．選択する共変量は重要な予後因子であるべきであり，この手順は傾向スコアによるマッチングでも用いると効果的である．」と述べている．

傾向スコア $e_i$ を用いて，マッチングや重み付きの調整解析を行うことが可能である．重み付きの調整解析に用いる「重み」として，傾向スコアの逆数で表される IPTW (inverse probability of treatment weighting) や ATT Weighting がよく用いられ，それぞれ

IPTW : $w_{\mathrm{IPTW},i} = \frac{Z_i}{e_i} + \frac{1-Z_i}{1-e_i}$ (3)

ATT Weighting : $w_{\mathrm{ATT},i} = Z_i + \frac{e_i(1-Z_i)}{1-e_i}$ (4)

として表される．なお，IPTW は ATE を推定するために，ATT Weighting は ATT を推定するために用いられる．SAS による傾向スコアを用いた重み付き解析として，IPTW 法による解析方法について，SAS ユーザー総会においても報告されている (古川, 2011).

## 2.3 傾向スコアによるマッチング

傾向スコアによるマッチングでは，治療群と対照群における被験者のうち，傾向スコアが似通った被験者をペアとして，新たな解析対象集団を構成する (“matched sample” と呼ばれる)．マッチングにより解析対象集団が構成された後，群間の治療効果を推定すれば，ATT の推定を可能とする．なお，本稿では，治療群よりも対照群の被験者数が多い状況を想定する．

傾向スコアによるマッチングを行うためには，被験者の傾向スコア間の差異を測るための尺度が必要となる．その理由は，傾向スコアモデルに連続変数を含めた場合，ある被験者と全く同一の傾向スコアを持つ被験者を見つけることは困難であるためである．被験者 $i$ と被験者 $j$ の傾向スコア間の差異を測る尺度として，傾向スコア自体の差 $|e_i - e_j|$ や，傾向スコアのロジット変換した値の差 $|\mathrm{logit}(e_i) - \mathrm{logit}(e_j)|$ が用いられる．

傾向スコアによるペアを構成するために，よく用いられる方法として，最近傍マッチング (greedy nearest neighbor matching) が挙げられる (Rosenbaum and Rubin, 1985)．最近傍マッチングでは，治療群の被験者に対して最も差異の小さい被験者を対照群から選択してペアを構成する．例えば，治療群 1 名に対して対照群から 1 名を選択する方法は 1 : 1 マッチングと呼ばれる．対照群から k 名を選択する k : 1 マッチングとして実施することも可能である．

最近傍マッチングによってペアを構成する場合，特定のキャリパー (caliper) を指定し，キャリパーの値に基づいて算出される幅以下であればペアの構成が行われる．傾向スコアマッチングの場合、キャリパーの値として，傾向スコアの推定値をロジット変換した値の標準偏差に 0.25 を掛けた値を推奨する報告がある (Rosenbaum and Rubin, 1985)．近年の報告としては，傾向スコアの推定値をロジット変換した値の標準偏差に 0.20 を掛けた値を用いることで，多くの状況下でバイアスが小さくなることが示されている (Austin, 2011)．キャリパーが傾向スコアの推定値をロジット変換した値の標準偏差に 0.20 を掛けた値で設定された事例は多く報告されており，例えば Bangalore et al. (2015) とその解説記事として田栗 (2017) を参照されたい．ただし，実際にマッチング実施後に，傾向スコアで調整した変数の分布に群間でバイアスがなくなったか確認することが重要であり，バイアスが生じたままである場合には，キャリパーの値を再考する必要がある (山本・森田, 2015).

SAS を用いて傾向スコアマッチングを行う場合，プロシジャでサポートされていない分析手法であるため，DATA ステップで最近傍マッチングのアルゴリズムを組まなければならなかった．DATA ステップで傾向スコアマッチングを行うための SAS マクロが報告されているが (Coca-Perraillon, 2007; Lanehart et al., 2012)，キャリパーの設定値が傾向スコアの推定値をロジット変換した値の標準偏差を掛けた値となっていない．そこで，3 節で解説するプロシジャの動作を理解する手助けとなるよう，過去に報告された SAS マクロを改良して，SAS の DATA ステップ上で，傾向スコアによる最近傍マッチングを行うための SAS マクロを開発した．詳細は，付録 A を参照されたい．

## 2.4 JMP を用いた傾向スコアによる 1 : 1 最近傍マッチング

SAS の代わりに JMP で分析を行う場合，次の URL からアドインをダウンロードすれば，傾向スコアによる 1 : 1 最近傍マッチングを行うことができる．ただし，JMP 11.0 以上のバージョンでなければならない．

アドイン入手 URL : http://www.jmp.com/content/dam/jmp/documents/jp/support/propensityaddin.zip

傾向スコアによる 1 : 1 最近傍マッチングを行うにあたっては，事前に傾向スコアを表す変数もデータセットに用意する必要がある．SAS を用いる場合は LOGISTIC プロシジャの OUTPUT ステートメント，JMP を用いる場合は「名義ロジスティックのあてはめ」のオプション「確率の計算式の保存」で対応できる．

以下の手順で 1 : 1 最近傍マッチングを行うことができる．

1. [アドイン] → [最近傍マッチング] → [最近傍マッチングの実行] を選択して，図 1 の入力画面を表示させる．
2. 「Y, 応答変数」 に適当な変数を指定する．
3. 「T, 処置変数 (2 水準)」に群を表す変数を指定する．
4. 「マッチングに使うスコア (数値)」に傾向スコアを表す変数を指定する．
5. 「乱数シード値」を設定する．
6. 「Caliper 係数 : a」を設定する．
7. 「元データにマッチング結果の列を追加する」にチェックする．
8. 「OK」をクリックして出力結果を示す．

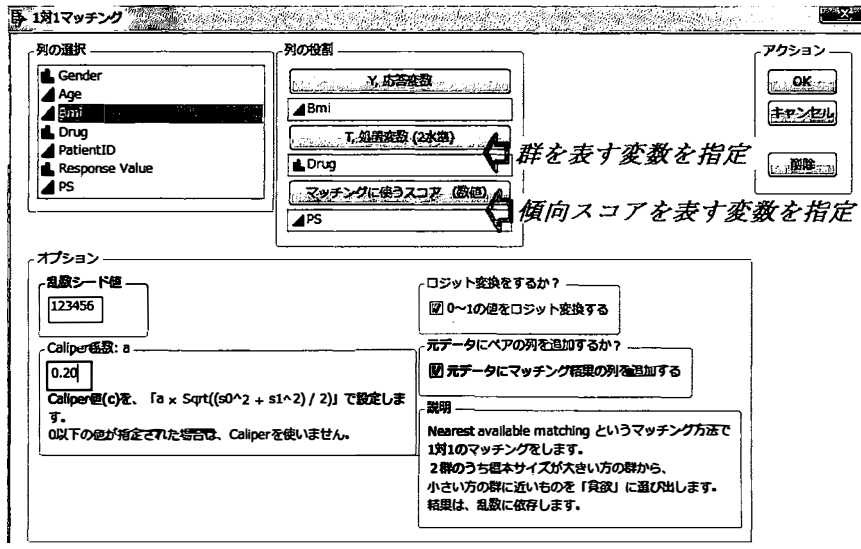



図 1 : JMP アドインによる最近傍マッチング実行画面

なお，手順 6 におけるキャリパーの値は 0.20 とすることが推奨されている (JMP ジャパン事業部, 2014). ただし，Austin (2011) は傾向スコアの推定値のロジット変換後の標準偏差を 0.20 倍した値をキャリパーとして利用することが推奨している一方，JMP アドインの実行画面で指定するキャリパーの値は群別に求めた不偏分散の平均値の平方根である．そのため，手順 6 のキャリパーの値を 0.20 と指定することは，Austin (2011) によって推奨されている 0.20 とは異なる．

以上の手順で，マッチングの結果として，マッチング前後におけるロジット変換後の傾向スコアの分布を確認でき，元データのすべてのオブザベーションに対して，「マッチした」あるいは「マッチなし」を表す変数が追加される．

SAS プログラムに関する素養がなく，JMP で統計解析を行っているユーザーであれば，上記の方法によって，傾向スコアマッチングを行う選択肢もある．ただし，JMP のアドインでは 1 : 1 マッチングのみで，k : 1 マッチングはサポートされていない．また，SAS に限らず，例えば R プログラムに関する素養があれば，R のパッケージ "Match" を用いて，傾向スコアマッチングを行うことが可能である．付録 B には，R のパッケージ "Match" を用いて，傾向スコアによる最近傍マッチングを行うための詳細を示している．しかし，SAS ユーザーであれば，SAS 上の DATA ステップあるいは SAS/STAT のプロシジャを用いて傾向スコアマッチングを行いたいと考えるであろう．

## 3 PSMATCH プロシジャを用いた傾向スコアによるマッチング

傾向スコアマッチングは，SAS/STAT のプロシジャでサポートされていない分析手法であるため，SAS で対応するためには，DATA ステップで最近傍マッチングのアルゴリズムを組まなければならなかった．しかし，SAS/STAT 14.2 (SAS 9.4 TS1M4) から PSMATCH プロシジャが追加され，傾向スコアマッチングができるようになった (SAS Institute Inc., 2016). PSMATCH プロシジャを用いれば，傾向スコア推定後の最近傍マッチングを行う場合に，前節で述べたような JMP あるいは R にデータセットを移動させて，マッチングを行う必要がなくなる．また，SAS の DATA ステップでマッチングのアルゴリズムを組むことなく，傾向スコアの推定からマッチングまで，SAS のプロシジャですべて完結できる．

本節では，PSMATCH プロシジャを用いて，傾向スコアによるマッチングをどのように実施するか解説し，出力結果の解釈を述べる．傾向スコアによるマッチングを行うための対象データとして，観察研究のデータ (データセット名：Drugs) を用いる ($n = 486$). データセット Drugsは，新規治療を行った「治療群」(変数：Drug= 'Drug_X') と標準治療を行った「対照群」(変数：Drug= 'Drug_A') を比較するために収集されたデータであるが，被験者あるいは主治医がどちらの治療を行うか選択できたため，観察研究のデータとして扱う．ベースラインの被験者背景として，年齢 (変数名：Age), 性別 (変数名：Gender ;Gender = 'Male'「男性」, 'Female'「女性」), BMI (変数名 Bmi) が挙げられる．

3.1 節から 3.3 節では，PSMATCH プロシジャの主なステートメントによる機能を詳述する．図 2 は，傾向スコアマッチングの手順における PSMATCH プロシジャの主なステートメントの役割を概念図として示している．

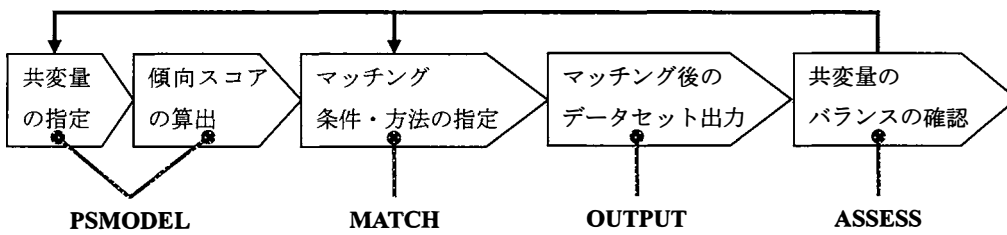



図 2：傾向スコアマッチングの手順における PSMATCH プロシジャの主なステートメントの役割

### 3.1 PSMODEL ステートメントによる傾向スコアモデル

LOGISTIC プロシジャでサポートされている MODEL ステートメントと同じように，PSMODEL ステートメントで傾向スコアモデルを指定すれば，傾向スコアの推定結果が出力される (プログラム 2).

プログラム 2 : PSMATCH プロシジャにおける PSMODEL ステートメントの利用例

```
proc psmatch data=drugs region=allobs;
   class z Gender;
   psmodel z(treated='1')= Gender Age Bmi;
   output out(obs=all)=psout00 ps=ps attwgt=attw atewgt=atew;
run;
```

- PROC PSMATCH ステートメント

PSMATCH プロシジャでは，REGION = で層別やマッチングを行う場合のオブザベーションのサポート範囲を指定できる．プログラム 2 では，簡便に REGION = ALLOBS と指定し，すべてのオブザベーションをマッチングのサポート範囲として取り扱っている．デフォルトは REGION = CS であり，治療群および対照群における傾向スコアについて，共通のサポート (common support) 範囲が選択される．

- PSMODEL ステートメント

傾向スコアモデルの指定を行っている．LOGISTIC プロシジャなどの MODEL ステートメントにおける指定と異なり，PSMODEL ステートメントの傾向スコアモデルで指定する応答変数も CLASS ステートメントに指定する必要がある．また，応答変数の水準指定には，TREATED = “治療群の水準” を指定する．なお，LOGISTIC プロシジャなどで予め傾向スコアを含んだデータセットを用いれば，PSMODEL ステートメントの代わりに PSDATA ステートメントを指定する方法もある．

- OUTPUT ステートメント

プログラム 2 では，OUTPUT ステートメントにおいて OUT = PSOUT00 と指定し，出力させるデータセット名を指定している．ここで，PSMATCH プロシジャでは，マッチング後の対象データを出力データセットとして保存することもできるため，OUT = のオプションとして OBS = が指定できる．OBS = ALL と指定すると，PROC PSMATCH ステートメントの DATA = で指定したデータセットすべてが出力される．

表 1 : プログラム 2 を実行させて得られたデータセット PSOUT00 (最初の 3 オブザベーション)

| Gender | Age | Bmi | Drug | PatientID | PS | ATEW | ATTW |
|---|---|---|---|---|---|---|---|
| Male | 29 | 22.02 | Drug_X | 284 | 0.36444 | 2.744 | 1 |
| Male | 45 | 26.68 | Drug_A | 201 | 0.22296 | 1.2869 | 0.28694 |
| Male | 42 | 21.84 | Drug_A | 147 | 0.11323 | 1.1277 | 0.12768 |

さらに，OUTPUT ステートメントにおいて，PS = を指定すれば，指定した変数名で傾向スコアの推定値が出力データセットに含まれる．また，ATEWGT = や ATTWGT = を指定すれば，式 (3) および (4) で示した IPTW および ATT Weigh tingが指定した変数名で出力データセットに含まれる (表 1).

## 3.2 MATCH ステートメントによる傾向スコアマッチング

PSMATCH プロシジャにおいて，MATCH ステートメントを指定すれば，傾向スコアによるマッチングができる．図 3 は，PSMATCH プロシジャでサポートされているマッチング方法である．デフォルトは METHOD = OPTIMAL である．入力データセットとキャリパーの値の指定によっては，マッチングができない場合もある．例えば，キャリパーの値が小さすぎると，ペアの構成数が少なくなってしまうため，治療群と同じ数の対照群をペアとして構成することが困難となる．このような場合に METHOD = OPTIMAL を指定して Fixed Ratio Matching を実行すると，"A feasible optimal fixed ratio matching that has the specified parameters does not exist." というエラーが出て，マッチングできない．

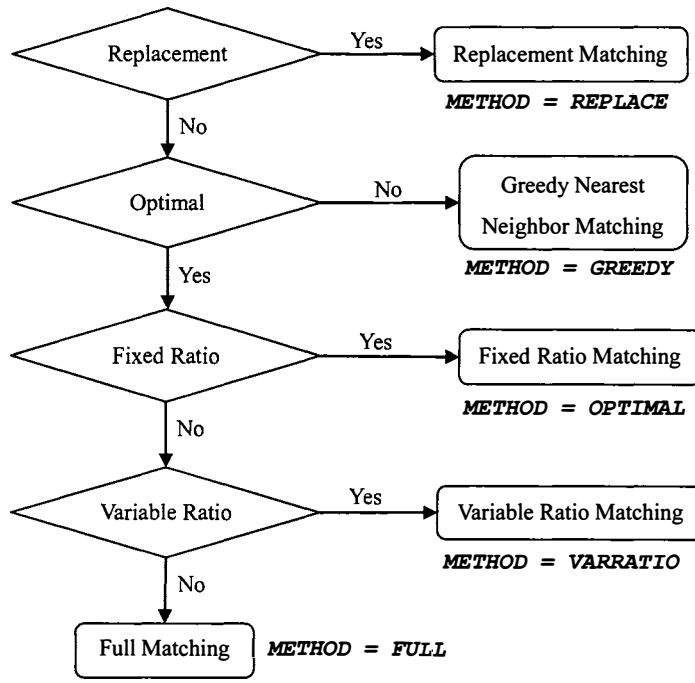



図 3 : MATCH ステートメントで指定可能なマッチング方法

データセット Drugs に対する傾向スコアマッチングとして，プログラム 3 は，前節で取りあげた最近傍マッチング "Greedy Nearest Neighbor Matching" を行うためのプログラムである．MATCH ステートメントにおいて METHOD = GREEDY を指定したプログラム 3 を実行すると，表 2 の出力データセット PSSORT00 が得られる．オプションで k = 1 を指定することで 1 : 1 マッチングを行っている．また，ORDER = DESCENDING を指定することで，治療群における傾向スコアが大きい順に対照群のペアを探している．ORDER = RANDOM(SEED = 乱数シード) と指定すれば，乱数シードに基づき治療群における傾向スコアのランダムな順で対照群のペアを探すことが可能である．なお，デフォルトは k = 1，ORDER = DESCENDING である．

- キャリパーの指定

プログラム 3 では，キャリパーとして CALIPER = 0.20 と指定している．STAT = LPS および CALIPER のオプションとして MULT = STDDEV と指定することで，キャリパーの値としては，推定された傾向スコアのロ

ジット変換した値の標準偏差に 0.20 を掛けた値を用いることを意味しており，Austin (2011) において推奨されている値に相当する．なお，STAT = LPS および MULT = STDDEV は指定しなくてもデフォルトとして実行される．

プログラム 3：PSMATCH プロシジャによる最近傍マッチング

```
proc psmatch data=Drugs region=allobs;
   class Drug Gender;
   psmodel Drug(treated='Drug_X')= Gender Age Bmi;
   match method=greedy(k=1 order=descending)
         stat=lps caliper(mult=stddev)=0.20;
   output out(obs=match)=psmatch00 ps=ps  lps=lps
          matchwgt=match matchid=matchsort;
run;
/* ペアの確認のために並び替え */
proc sort data=psmatch00 out=pssort00;
  by descending ps;
run;
```

- OUTPUT ステートメント

プログラム 3 のように，OBS = MATCH と指定することで，マッチングされたオブザベーションのみを含んだデータセットが出力される．なお，OBS = ALL と指定した場合，MATCHWGT = で指定した変数で，マッチングされたオブザベーションを識別することが可能である．プログラム 3 の場合，マッチングされたオブザベーションのみが変数 MATCH = 1 と出力される．

さらに，MATCHID = で指定した変数 MATCHSORT で，マッチングした順およびペアの構成を確認することができる．そのために，出力データセットを傾向スコア順に並び替えたデータセット PSSORT00 で確認すると，最初の 5 オブザベーションは表 2 のようになる．傾向スコアの大きい順にマッチングが行われ，被験者番号 (変数名：PatientID) として 87 と 123，200 と 349 のペアが構成されたことがわかる．

表 2：プログラム 3 を実行させて得られたデータセット PSSORT00 (最初の 5 オブザベーション)

| Gender | Age | Bmi | Drug | PatientID | PS | match | matchsort |
|---|---|---|---|---|---|---|---|
| Male | 27 | 25.87 | Drug_X | 87 | 0.64115 | 1 | 1 |
| Male | 27 | 25.76 | Drug_A | 123 | 0.63513 | 1 | 1 |
| Female | 30 | 27.75 | Drug_A | 200 | 0.61831 | 1 | 2 |
| Male | 33 | 28.06 | Drug_X | 349 | 0.60481 | 1 | 2 |
| Male | 27 | 25.14 | Drug_X | 114 | 0.60047 | 1 | 3 |

## 3.3 ASSESS ステートメントによる変数の群間差の評価

PSMATCH プロシジャの ASSESS ステートメントで，すべての解析対象データセット，あるいはマッチされた対象データセットに対して，指定した変数の群間差を評価できる．

プログラム 4 では，MATCH ステートメントに関して 3.2 節と同様の指定を行ったもとで，ASSESS ステートメントで LPS を指定して，傾向スコアをロジット変換した値の群間差を評価する．もし PS も指定すれば，傾向スコアの群間差も評価可能である．さらに，VAR = で指定した変数についても，群間差の評価を行う．また，VARINFO オプションを指定することで，各群に対する変数情報を出力させる (表 4)．もし WEIGHT = オプションで重みを指定すれば，式 (3)，(4) で示した IPTW，ATT Weighting を用いた重み付きで変数の評価を行う．プログラム 4 では，最後に，PLOTS = で ODS GRAPHICS による出力結果を得ている．

プログラム 4：PSMATCH プロシジャにおける ASSESS ステートメントの利用例

```
ods graphics on;
proc psmatch data=drugs region=allobs;
  class Drug Gender;
  psmodel Drug(treated='Drug_X')= Gender Age Bmi;
  match method=greedy(k=1 order=descending)
        stat=lps caliper(mult=stddev)=0.20;
  assess lps var=(Gender Age Bmi)
        / weight=none varinfo plots=(all);
run;
```

表 3：傾向スコアの情報

| Propensity Score Information | | | | | | |
|---|---|---|---|---|---|---|
| Observations | Treated (Drug = Drug_X) | | | Control (Drug = Drug_A) | | |
| | N | Mean | Std Dev | N | Mean | Std Dev |
| All | 113 | 0.310773 | 0.132467 | 373 | 0.208801 | 0.131969 |
| Region | 113 | 0.310773 | 0.132467 | 373 | 0.208801 | 0.131969 |
| Matched | 113 | 0.310773 | 0.132467 | 113 | 0.308755 | 0.130649 |

表 3 では，傾向スコアの情報を示しており，紙面スペースの関係上，一部の統計量に関しては出力を省略している．Observation = Region である出力結果も示されているが，PROC PSMATCH ステートメントで REGION = ALLOBS を指定しているため，Observation = All と同様の結果となっている．次に，Observation = All と Observation = Matched の結果を比較すれば，データセットのうち，どの程度のオブザベーションでペアが構成されたか確認できる．データセット Drugs の場合，治療群：113 名，対照群：373 名のうち，治療群：113 名，対照群：113 名がマッチされたことがわかる．3.2 節のプログラム 3 を実行させて出力したデータセット PSMATCH00 とオブザベーション数が一致していることを確認できるであろう．

表 4 は VARINFO オプションを指定して出力される内容で，ASSESS ステートメントで指定した変数情報の要約統計量が出力される．性別のようなカテゴリカルデータの場合，Mean の列には，平均の代わりに，最初

の順番のカテゴリにおける割合が出力される．表 4 における性別 Gender では，Female「女性」の割合が出力されている．

表 4：変数情報

| Variable Information | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Observations | Treated (Drug = Drug_X) | | | Control (Drug = Drug_A) | | |
| | | N | Mean | Std Dev | N | Mean | Std Dev |
| LPS | All | 113 | -0.880615 | 0.681761 | 373 | -1.520586 | 0.844486 |
| | Region | 113 | -0.880615 | 0.681761 | 373 | -1.520586 | 0.844486 |
| | Matched | 113 | -0.880615 | 0.681761 | 113 | -0.889200 | 0.674687 |
| Age | All | 113 | 36.309735 | 5.534114 | 373 | 40.404826 | 6.579103 |
| | Region | 113 | 36.309735 | 5.534114 | 373 | 40.404826 | 6.579103 |
| | Matched | 113 | 36.309735 | 5.534114 | 113 | 35.884956 | 5.468664 |
| Bmi | All | 113 | 24.492566 | 1.863797 | 373 | 23.753271 | 1.980778 |
| | Region | 113 | 24.492566 | 1.863797 | 373 | 23.753271 | 1.980778 |
| | Matched | 113 | 24.492566 | 1.863797 | 113 | 24.309027 | 1.882123 |
| Gender | All | 113 | 0.433628 | 0.495575 | 373 | 0.458445 | 0.498270 |
| | Region | 113 | 0.433628 | 0.495575 | 373 | 0.458445 | 0.498270 |
| | Matched | 113 | 0.433628 | 0.495575 | 113 | 0.495575 | 0.499980 |

表 5 では，ASSESS ステートメントで指定した変数に対して，標準化した群間差が出力される．Mean Difference は単純な群間の平均値の差であり，Standardized Mean Difference は Mean Difference / Divisor の計算結果である．Divisor 「除数」は各群の標準偏差をプールした標準偏差で構成される．例えば，年齢を表す変数 Age の All Obs「すべてのオブザベーション」を対象とした出力の場合，表 4 で求めた各群の標準偏差から，

$$\sqrt{\frac{(5.534114^2 + 6.579103^2)}{2}} = 6.079104$$

が Divisor となり，表 5 に出力された Mean Difference より，

$$\frac{\text{Mean Difference}}{\text{Divisor}} = \frac{-4.095091}{6.079104} = -0.673634$$

が Standardized Mean Difference として出力されている．

なお，ASSESS ステートメントで WEIGHT = NONE と指定したため，表 5 は重みを付けない計算結果が出力されている．例えば，WEIGHT = ATEWGT と指定すれば，式 (3) の IPTW で重み付けたした計算結果が出力される．また，ODS GRAPHICS ON の状態で実行しているので，表 5 で示した各変数に対する標準化した群間差に対応する図も出力される (付録 C)．付録 C では，±0.25 がカットオフ値として参照線が出力され，[–0.25, 0.25] であればバランスが取れているという目安として表している (SAS Institute Inc., 2016)．なお，カットオフ値として ±0.10 を推奨した報告も行われている (Normand et al., 2001)．

次に，ASSESS ステートメントで PLOTS = ALL を指定することで，ODS GRAPHICS による出力結果を得ることができる．ODS GRAPHICS による出力として，ASSESS ステートメントで指定した変数に対する群間の分布として，クラウドプロットが出力される (図 4)．

表 5：各変数に対する標準化した群間差

| Standardized Variable Differences (Treated - Control) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Mean Difference | | | Standardized Mean Difference | | | |
| | | | | Divisor | Mean Difference | | |
| | All Obs | Region Obs | Matched Obs | | All Obs | Region Obs | Matched Obs |
| LPS | 0.639971 | 0.639971 | 0.008584 | 0.767448 | 0.833894 | 0.833894 | 0.011185 |
| Age | -4.095091 | -4.095091 | 0.424779 | 6.079104 | -0.673634 | -0.673634 | 0.069875 |
| Bmi | 0.739296 | 0.739296 | 0.18354 | 1.923178 | 0.384414 | 0.384414 | 0.095436 |
| Gender | -0.024817 | -0.024817 | -0.061947 | 0.496925 | -0.049941 | -0.049941 | -0.124661 |

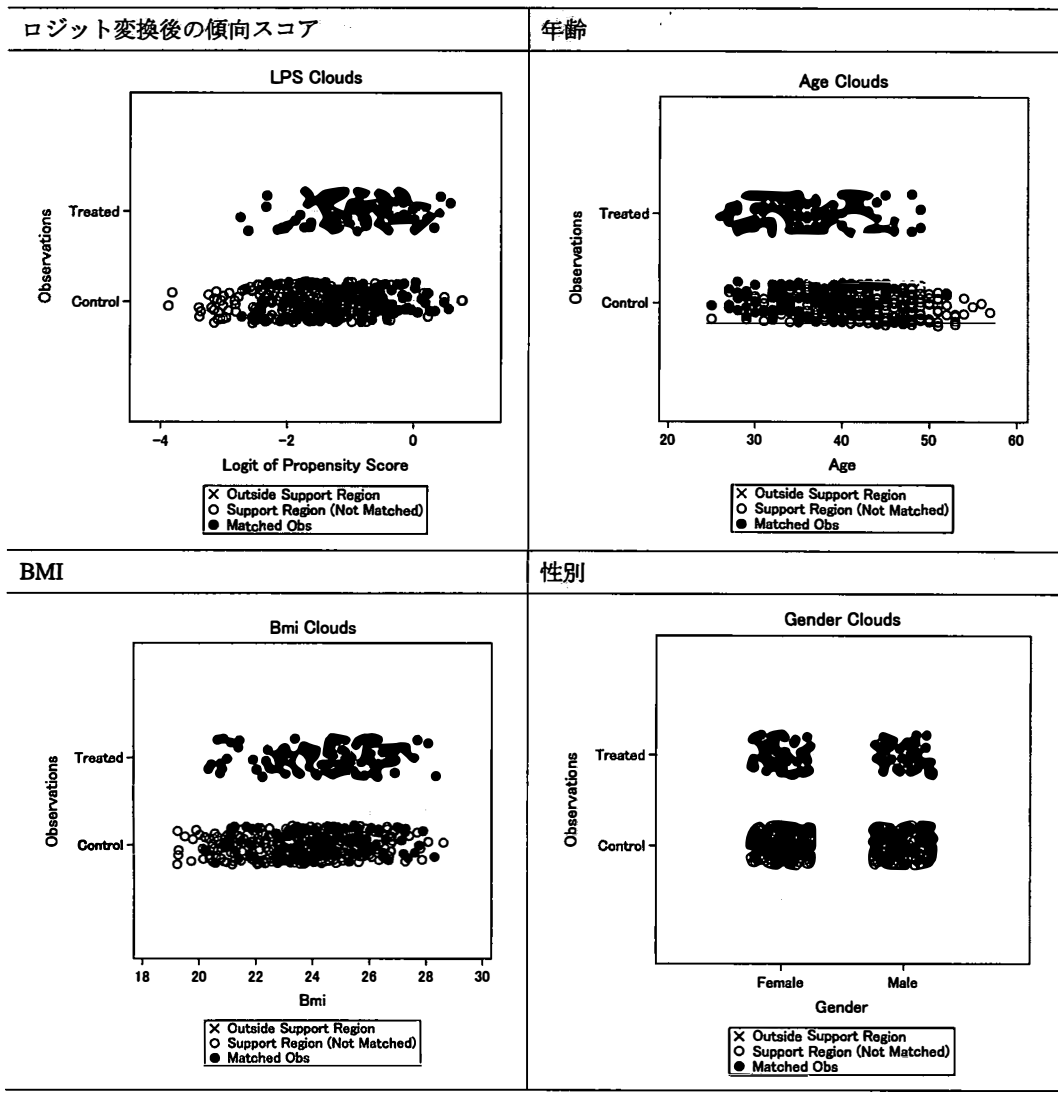



図 4：ODS GRAPHICS による各変数に対する群別のクラウドプロット

クラウドプロットでは，縦軸に群，横軸に対象となる変数の値をとって散布図を描いており，マッチされたデータとマッチされなかったデータの記号を変えて出力している．クラウドプロットをみれば，傾向スコアがどの程度の値のデータでペアが構成できなかったか，各変数がどの程度の値をとるデータでペアが構成できなかったか，視覚的に確認することができる．

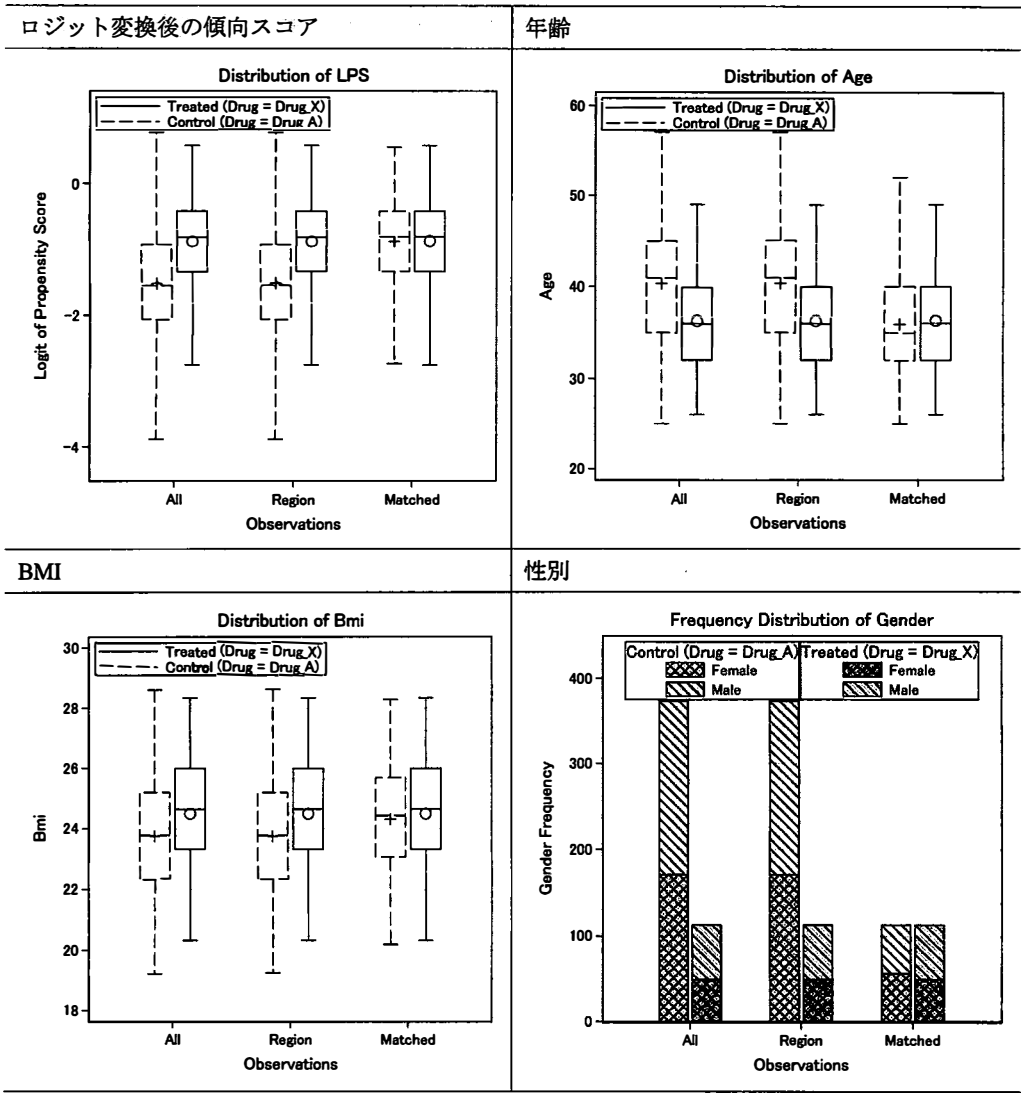



図 5 : ODS GRAPHICS による各変数に対する群別の箱ひげ図・度数分布

また，ODS GRAPHICS による出力として，連続変数に対しては箱ひげ図，カテゴリカル変数に対しては度数分布が出力される (図 5)．図 5 では，プログラム 4 において PLOTS = ALL と指定するだけでなく，PLOTS(ORIENT = VERTICAL) = ALL と指定し，縦軸に対象となる変数の値，横軸に群をとったグラフを描いている．なお，デフォルトは ORIENT=HORIZONTAL である．図 5 の箱ひげ図および度数分布より，各変数におけるマッチング前後で群間のバランスが視覚的に確認できる．なお，MATCH ステートメントにおいて，

EXACT = <カテゴリカル変数> と指定すると，指定したカテゴリカル変数は群間で完全にバランスが取れる．プログラム 4 の MATCH ステートメントにおいて，EXACT = Gender と指定すれば，"Some treated units have less than the specified K=1 matched controls because there are not enough available controls for these treated units." という警告が出て，マッチされた対象データは結果的に 113 名から 112 名に減少してしまう．しかし，群間の性別の度数分布は完全にバランスが取れていることが視覚的に確認できるであろう (付録 D).

なお，図 4 および図 5 で示した箱ひげ図やクラウドプロットの出力では，モノクロ印刷されても識別しやすいよう，ODS の出力としてジャーナルスタイル "STYLE = JOURNAL2" を指定した．特に，図 5 においては群別のグラフを示す場合，デフォルトでは群の違いが色で分けて出力されてしまうが．ジャーナルスタイルを指定したことで，モノクロ印刷に対応できるよう出力されている (魚住ら, 2016a, 2016b; 大橋ら, 2016).

## 4 まとめ

本稿では，観察研究データに対する統計解析手法として，傾向スコアによるマッチングに着目した．傾向スコアマッチングを SAS で実行する方法として，SAS/STAT 14.2 から追加された PSMATCH プロシジャによるプログラムについて解説し，実行結果の解釈に関して詳述した．PSMATCH プロシジャにおける PSMODEL ステートメントで，傾向スコアモデルにおける目的変数と説明変数を指定でき，MATCH プロシジャで様々なマッチングがサポートされている．3 節では，特に最近傍マッチングをについて取りあげ，PSMATCH プロシジャによるアウトプット画面および ODS GRAPHICS による出力結果について解説した．なお，2 節において，観察研究における平均治療効果の推定について整理したもとで，傾向スコアによる最近傍マッチングに関しても解説している．さらに，SAS のプロシジャで傾向スコアマッチングが実施できなかった状況下で，JMP や R で実行できることを述べた．そして，JMP で傾向スコアマッチングを行う場合の留意点を取りあげ，加えて R による最近傍マッチングの実行プログラムを付録 B に示した．また，PSMATCH プロシジャで実施している傾向スコアマッチングのアルゴリズムを理解する手助けとなるよう，DATA ステップで最近傍マッチングを実施するための SAS マクロを開発した (付録 A). この SAS マクロは，キャリパーの設定値として，臨床論文でよく報告されている，傾向スコアの推定値をロジット変換した値の標準偏差を掛けた値となっている点で有用である．

SAS ユーザー総会においても，これまで傾向スコアを用いた IPTW 法などに関して，SAS を用いた研究報告が行われている (古川・杉本, 2010; 古川, 2011). また，SAS/STAT 14.2 では PSMATCH プロシジャに加えて，平均因果効果を推定するためのプロシジャとして，新たに CAUSALTRT プロシジャも追加されている (SAS Institute Inc., 2016). PSMATCH プロシジャや CAUSALTRT プロシジャが新たに追加されたことによって，今後 SAS による傾向スコアを用いた統計解析が多く実施され，さまざまな分野におけるデータ解析に応用された結果が，論文・学会発表を通して報告されると予想される．傾向スコアによるマッチングを用いた調整解析を行う実務家にとって，本稿の報告内容がお役に立てれば幸いである．

## 参考文献

[1] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*. **26**:734–753, 2007.

[2] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*. **28**:3083–3107, 2009.

[3] Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*. **10**:150–161, 2011.

[4] Bangalore S, Guo Y, Samadashvili Z, Blecker S, Xu J, Hannan EL. Everolimus-eluting stents or bypass surgery for multivessel coronary disease. *New England Journal of Medicine*. **372**:1213–1222, 2015.

[5] Coca-Perraillon M. Local and global optimal propensity score matching. *Proceedings of the SAS Global Forum*. SAS Institute Inc., Cary, NC, 2007. Available at http://www2.sas.com/proceedings/forum2007/185-2007.pdf.

[6] Faries D, Leon AC, Haro JM, Obenchain RL. *Analysis of Observational Health Care Data Using SAS(R)*. SAS Institute Inc., Cary, NC, 2010.

[7] Hida K, Okamura R, Sakai Y, Konishi T, Akagi T, Yamaguchi T, Akiyoshi T, Fukuda M, Yamamoto S, Yamamoto M, Nishigori T, Kawada K, Hasegawa S, Morita S, Watanabe M. Open versus laparoscopic surgery for advanced low rectal cancer: a large multicenter propensity score matched cohort study in Japan. *Annals of Surgery*, 2017. DOI:10.1097/SLA.0000000000002329.

[8] Hamatani Y, Ogawa H, Uozumi R, Iguchi M, Yamashita Y, Esato M, Chun YH, Tsuji H, Wada H, Hasegawa K, Abe M, Morita S, Akao M. Low body weight is associated with the incidence of stroke in atrial fibrillation patients: Insight from the Fushimi AF Registry. *Circulation Journal*. **79**:1009–1017, 2015.

[9] JMP ジャパン事業部. JMP による傾向スコアを用いたマッチング、層別分析、回帰分析. SAS Institute Japan 株式会社, 2014. Available at https://www.jmp.com/content/dam/jmp/documents/jp/support/propensityscore201412.pdf.

[10] Lanehart RE, de Gil PR, Kim ES, Bellara AP, Kromrey JD, Lee RS. Propensity score analysis and assessment of propensity score approaches using SAS(R) procedures. *Proceedings of the SAS Global Forum*. SAS Institute Inc., Cary, NC, 2012. Available at http://support.sas.com/resources/papers/proceedings12/314-2012.pdf.

[11] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*. **23**:2937–2960, 2004.

[12] Normand ST, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology*. **54**:387–398, 2001.

[13] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. **70**:41–55, 1983.

[14] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. **39**:33–38, 1985.

[15] Rosenbaum PR. *Design of Observational Studies*. Springer-Verlag: New York, NY, 2010.

[16] SAS Institute Inc. *SAS/STAT(R) 14.2 User's Guide*. SAS Institute Inc., Cary, NC, 2016.

[17] Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical Science*. **25**:1–21, 2010.

[18] Yamashita Y, Uozumi R, Hamatani Y, Esato M, Chun YH, Tsuji H, Wada H, Hasegawa K, Ogawa H, Abe M,

Morita S, Akao M. Current status and outcomes of direct oral anticoagulant use in patients with atrial fibrillation in the real-world: The Fushimi AF Registry. *Circulation Journal*, 2017 DOI:10.1253/circj.CJ-16-1337.

[19] 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. 216–232, 2016a.

[20] 魚住龍史・矢田真城・浜田知久馬. SAS プロシジャを用いた生存時間データに対する例数設計の変革. SAS ユーザー総会 論文集. 250–267, 2016b.

[21] 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.

[22] 田栗正隆. 有名論文から統計の基礎を学ぶ: プロペンシティスコア. *Coronary Intervention*, **13**:63–69, 2017.

[23] 古川敏仁・杉本典子. Propensity Score 法によるバイアスの調整法に関する実務的な問題点. SAS ユーザー総会 論文集. 101–108, 2010.

[24] 古川敏仁. 観察研究において選択 bias を制御するために用いられる Propensity Score IPTW と層化調整法の、頑健性の観点からの使い分けについて. SAS ユーザー総会 論文集. 409–416, 2011.

[25] 山本倫生・森田智視. 傾向スコアによる調整解析. 呼吸. **34**:1187–1193, 2015.

## 連絡先

E-mail : *uozumi@kuhp.kyoto-u.ac.jp*

付録 A : DATA ステップによる最近傍マッチングを行うための SAS マクロ

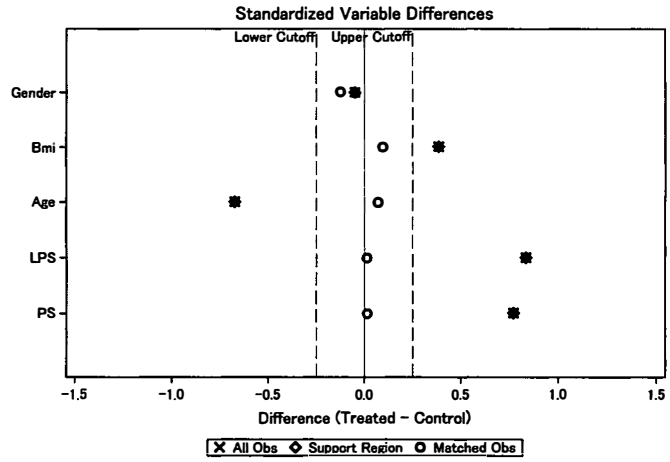
```
*==================================================*;
* 群を表す数値変数を準備 ;
data Drugs0;set Drugs;
  if Drug="Drug_X" then Group=1;else if Drug="Drug_A" then Group=0;
run;

*==================================================*;
* 傾向スコアを含んだデータセットを準備 ;
ods listing close;
proc logistic data=Drugs0;
  class Gender; model Group(event="1")=Gender Age Bmi;
  output out=ps00 pred=ps xbeta=ps_xb stdxbeta=ps_sdxb;
run;
ods listing;

*==================================================================================*;
* MACRO: PSMATCHING ;
*==================================================================================*;
* 傾向スコアによるマッチング後の被験者番号リストを作成する ;
* 傾向スコアによるマッチング後のデータセットを作成する ;
*==================================================================================*;
* 引数の説明 ;
*==================================================================================*;
* data: 入力対象データセット[傾向スコアなどを変数に含めたデータセットとする] ;
* id: 被験者番号を表すカテゴリ変数 ;
* grp: 群（治療群と対照群）を表すカテゴリ変数 ;
* tgrp_ind: 治療群を表す値 ;
* cgrp_ind: 対照群を表す値 ;
* ps_pred: 傾向スコアを表す変数 ;
* sort: マッチングを行う傾向スコアの順を指定 ;
* --- sort = descending   : 傾向スコアの大きい順 ;
* --- sort = ascending    : 傾向スコアの小さい順 ;
```

```
* --- sort = random        : ランダムな順 ;
* caliper: キャリパーの値(傾向スコアのロジット変換した値の標準偏差に掛ける値)を指定 ;
* seed: 乱数のシードを指定(sort = random) ;
* outid: マッチングされた被験者番号を群ごとに出力したデータセット ;
* outdata: 入力対象データセットからマッチングされなかった被験者番号を削除したデータセット ;
*==========================================================================*;

*-------------------------------------------------------------------------;
* マクロ開始 ;
%macro psmatching(data,id,grp,tgrp_ind,cgrp_ind,ps_pred,sort,caliper,seed,outid,outdata);

*-------------------------------------------------------------------------;
* 傾向スコアの変数名・傾向スコアのロジット変換した変数を準備 ;
data ps_p;set &data; ps_pred=&ps_pred; l_ps_pred=log(ps_pred/(1-ps_pred)); run;

*-------------------------------------------------------------------------;
* 各群のデータ数を読み込み ;
ods listing close; ods output summary=grp_N;
proc means data=ps_p n; class &grp; var ps_pred; run;
ods listing;
data _null_;set grp_N;
  where &grp=&tgrp_ind; %global tgrp_N; call symput("tgrp_N",ps_pred_N);
data _null_;set grp_N;
  where &grp=&cgrp_ind; %global cgrp_N; call symput("cgrp_N",ps_pred_N);
run;

*-------------------------------------------------------------------------;
* 傾向スコアのロジット変換した値の標準偏差を準備 ;
ods listing close; ods output summary=lps_p;
proc means data=ps_p stddev; var l_ps_pred; run;
ods listing;

*-------------------------------------------------------------------------;
* マッチングをランダムに行う場合の乱数を準備 ;
data one;set ps_p; seed=&seed; call streaminit(seed); ranvar=rand("uniform"); run;

*-------------------------------------------------------------------------;
* マッチングを行う傾向スコアの順を指定 ;
%if "&sort"="descending" %then %do;
proc sort data=one out=one; by &grp descending ps_pred;
proc transpose data=one out = data1; by &grp; run;
%end;

%else %if "&sort"="ascending" %then %do;
proc sort data=one out=one; by &grp ps_pred;
proc transpose data=one out = data1; by &grp; run;
%end;

%else %if "&sort"="random" %then %do;
proc sort data=one out=one; by &grp ranvar;
proc transpose data=one out = data1; by &grp; run;
%end;

*-------------------------------------------------------------------------;
* マッチングを行うために各群のデータ数だけ変数を準備 ;
data id_t (rename=(COL1-COL%eval(&tgrp_N)=tid1-tid%eval(&tgrp_N)));set data1;
  if &grp=&tgrp_ind and _NAME_="PatientID";
data ps_t (rename=(COL1-COL%eval(&tgrp_N)=tps1-tps%eval(&tgrp_N)));set data1;
  if &grp=&tgrp_ind and _NAME_="&ps_pred";
data id_c (rename=(COL1-COL%eval(&cgrp_N)=cid1-cid%eval(&cgrp_N)));set data1;
  if &grp=&cgrp_ind and _NAME_="PatientID";
data ps_c (rename=(COL1-COL%eval(&cgrp_N)=cps1-cps%eval(&cgrp_N)));set data1;
  if &grp=&cgrp_ind and _NAME_="&ps_pred";
run;

*-------------------------------------------------------------------------;
* キャリパーを指定して傾向スコアマッチングを実施 ;
data all;
```

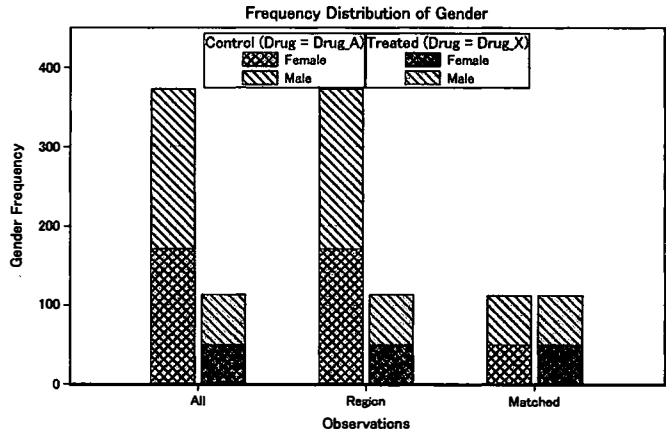
```
  merge id_t ps_t id_c ps_c lps_p;
  caliper=&caliper;
  array t_id {*} tid1-tid%eval(&tgrp_N); array c_id {*} cid1-cid%eval(&cgrp_N);
  array t_ps {*} tps1-tps%eval(&tgrp_N); array c_ps {*} cps1-cps%eval(&cgrp_N);
  array used_i {*} used1 - used%eval(&cgrp_N);
  array matched_t {*} m_tid1-m_tid%eval(&tgrp_N); array matched_c {*} m_cid1-m_cid%eval(&tgrp_N);
  array min_diff_i {*} min_diff1-min_diff%eval(&cgrp_N);
  match_N=0;
  do i=1 to &tgrp_N;
    min_diff=1; best_match=0;
    do j=1 to &cgrp_N;
      if used_i[j]=. then do;
        if abs(log(t_ps[i]/(1-t_ps[i]))-log(c_ps[j]/(1-c_ps[j])))
            < caliper*l_ps_pred_StdDev then do;
          if abs(log(t_ps[i]/(1-t_ps[i]))-log(c_ps[j]/(1-c_ps[j]))) < min_diff then do;
            min_diff=abs(log(t_ps[i]/(1-t_ps[i]))-log(c_ps[j]/(1-c_ps[j])));
            min_diff_i[j]=abs(log(t_ps[i]/ (1-t_ps[i]))-log(c_ps[j]/(1-c_ps[j])));
            best_match=j;
          end;
        end;
      end;
    end;
    if best_match > 0 then do;
      match_N=match_N+1; used_i[best_match]=1;
      matched_t[match_N]=t_id[i]; matched_c[match_N]=c_id[best_match];
    end;
  end;
run;
data all;set all;
  max_min_diff=max(of min_diff1-min_diff%eval(&cgrp_N));
  sum_min_diff=sum(of min_diff1-min_diff%eval(&cgrp_N));
run;

*-----------------------------------------------------------------------------;
* マッチングされた被験者番号を群ごとに出力したデータセット ;
data &outid;set all;
  array matched_t {*} m_tid1-m_tid%eval(&tgrp_N); array matched_c {*} m_cid1-m_cid%eval(&tgrp_N);
  do match = 1 to match_N; treat_id=matched_t[match]; control_id=matched_c[match]; output; end;
  keep match treat_id control_id;
run;

*-----------------------------------------------------------------------------;
* 群を表す変数を準備 ;
proc transpose data=&outid out=psmatch00; by match; var treat_id control_id;
data psmatch00;set psmatch00;
  if _NAME_="treat_id" then &grp=&tgrp_ind; else if _NAME_="control_id" then &grp=&cgrp_ind;
  &id=COL1; keep match &grp &id;
proc sort data=psmatch00 out=psmatch00; by &grp &id; run;
proc sort data=ps00 out=ps00sort; by &grp &id; run;

*-----------------------------------------------------------------------------;
* 入力対象データセットからマッチングされなかった被験者番号を削除したデータセット ;
data &outdata; merge ps00sort psmatch00(in=in1); by &grp &id; if in1=1 then output;
proc sort data=&outdata out=&outdata; by match &grp &id; run;

*-----------------------------------------------------------------------------;
* マクロ終了 ;
%mend psmatching;

*==============================================================================*;
* ランダムな順に傾向スコアマッチング ;
*==============================================================================*;
%psmatching(ps00,PatientID,Group,1,0,ps,random,0.20,123456,matchesr,psmatchdatar);

*==============================================================================*;
* 傾向スコアの大きい順にマッチング ;
*==============================================================================*;
%psmatching(ps00,PatientID,Group,1,0,ps,descending,0.20,123456,matchesd,psmatchdatad);
```

付録 B : R による最近傍マッチングの実行プログラム

```
### パッケージを使用 ###
library(Matching)

### 傾向スコアの推定結 果が含まれたCSVファイルPSOUT0をデータセットとして読み込み ###
data = read.csv("psout00.csv")

### 傾向スコアモデルの分子(治療群のカテゴリ)を指定 ###
Trt = data$Drug == "Drug_X"

### キャリパーの値を指定 ###
caliper0 = 0.20

### 傾向スコアのロジット変換 ###
data <- transform(data, ps.logit=log(ps / (1 - ps)))

### 乱数のシード ###
set.seed(123456)

### 1:1最近傍マッチングを非復元抽出で実行 ###
Matching = Match(Y=NULL, Tr=Trt, X=data$ps.logit, M=1, replace=FALSE, caliper = caliper0);
summary(Matching)
```




付録 C : 各変数に対する標準化した群間差 (ODS GRAPHICS)




付録 D : MATCH ステートメントで EXACT = Gender とした場合の性別の度数分布

# Mantel-Haenszel法により調整したリスク差の信頼区間に関する一考察

○佐々木 英麿、舟尾 暢男
武田薬品工業株式会社 日本開発センター 生物統計室

# A Consideration of the Confidence Intervals for Mantel-Haenszel Estimate of Common Risk Difference

Emma Sasaki, Nobuo Funao
Takeda Pharmaceutical Company, Ltd.

## 要旨

Mantel-Haenszel法により調整したリスク差に関する信頼区間の算出方法(Greenland信頼区間、Sato信頼区間、及びNewcombe信頼区間）を紹介し、各信頼区間の被覆確率をシミュレーションにより確認することで性能評価を行う。

キーワード：Mantel-Haenszel法、調整済みリスク差、信頼区間、被覆確率、FREQ プロシジャ、STDRATE プロシジャ

## 1. はじめに

2値応答変数を主要評価項目とした2群の並行群間比較試験では、主解析として割合の群間差（以下、リスク差）の評価を行い、感度分析として背景因子で調整した解析を実施するのが一般的である。また、治験薬の割付において層別割付を行った場合、主解析として層別因子で調整した解析を計画することがあるが、その際、Cochran-Mantel-Haenszel検定（以下、CMH検定）等を用いて群間比較を実施するとともに、Mantel-Haenszel法により調整したリスク差（以下、調整済みリスク差）に関する推定量を算出することが一般的である。いずれの場合においても、調整済みリスク差を算出する場合は、対応する信頼区間を合わせて算出することとなるが、信頼区間の算出方法についてはいくつか選択肢がある。

SASでは、リスク差に関する様々な信頼区間の算出が可能となっている。SAS V9.3で算出可能な信頼区間については飯塚、浜田（2013）が性能評価を行っており、SAS V9.4で新たに追加された信頼区間については飯塚、魚住、浜田（2014）が性能評価を行い、いずれもNewcombeスコアに基づく信頼区間を推奨しているが、調整済みリスク差に関する信頼区間の考察は、本ユーザー総会ではなされていない。

本稿では、FREQプロシジャに加え、SAS V9.3で新たに追加されたSTDRATEプロシジャを用い、調整済みリスク差に対する信頼区間の算出方法（Greenland信頼区間、Sato信頼区間、及びNewcombe信頼区間）を紹介し、これらのプロシジャを用いて調整済みリスク差の信頼区間を算出する際の留意点を示すとともに、それぞれの信頼区間の被覆確率をシミュレーションにより確認し、性能評価を行う。

## 2. Mantel-Haenszel法による調整済みリスク差の算出

まずは、表2.1の分割表を考える。

表 2.1　2×2 分割表

| 投与群 | イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | $x_1$ | $n_1 - x_1$ | $n_1$ |
| 群 2 | $x_2$ | $n_2 - x_2$ | $n_2$ |
| | | | $n_+$ |

このとき、群 $j$ の被験者のイベント発生割合（リスク）を $p_j$ とすると（ $j = 1, 2$ ）、各群のイベント発生数に関する確率変数 $X_j$ は二項分布 $B(n_j, p_j)$ に従うが、漸近的には正規分布 $N\left(n_j p_j, n_j p_j(1 - p_j)\right)$ に従う。よって、各群の割合（リスク）に関する確率変数 $Y_j$ は以下のように示すことができる。

$$Y_j = \frac{X_j}{n_j} \sim N\left(p_j, \frac{p_j(1 - p_j)}{n_j}\right) \tag{1}$$

2 群の並行群間比較試験において、群 1 と群 2 の割合（リスク）に関する確率変数 $Y_1$ と $Y_2$ はそれぞれ独立した正規分布に従うため、群 1 と群 2 のリスク差（$Y_1 - Y_2$）の分布は以下の通りとなる。

$$D = Y_1 - Y_2 \sim N\left(p_1 - p_2, \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}\right) \tag{2}$$

次に調整済みリスク差を考えるため、表 2.2 の分割表を考える。

表 2.2　層 $i$ の 2×2 分割表（ $i = 1, \cdots, k$ ）

| 投与群 | イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | $x_{i1}$ | $n_{i1} - x_{i1}$ | $n_{i1}$ |
| 群 2 | $x_{i2}$ | $n_{i2} - x_{i2}$ | $n_{i2}$ |
| | | | $n_{i+}$ |

$k$ 個の層を有する、ある層別項目を考えた場合、層 $i$ の群 $j$ における被験者のイベント発生割合（リスク）を $p_{ij}$ とすると（ $i = 1, \cdots, k;\ j = 1, 2$ ）、層 $i$ における各群のイベント発生数に関する確率変数 $X_{ij}$ は二項分布 $B(n_{ij}, p_{ij})$ に従い、漸近的には正規分布 $N\left(n_{ij} p_{ij}, n_{ij} p_{ij}(1 - p_{ij})\right)$ に従う。よって、層 $i$ における各群の割合（リスク）に関する確率変数 $Y_{ij}$ は以下の分布に従う。

$$Y_{ij} = \frac{X_{ij}}{n_{ij}} \sim N\left(p_{ij}, \frac{p_{ij}(1 - p_{ij})}{n_{ij}}\right) \tag{3}$$

よって、層 $i$ のリスク差（$Y_{i1} - Y_{i2}$）の分布は以下の通りとなる。

$$D_i = Y_{i1} - Y_{i2} \sim N\left(p_{i1} - p_{i2}, \frac{p_{i1}(1 - p_{i1})}{n_{i1}} + \frac{p_{i2}(1 - p_{i2})}{n_{i2}}\right) \tag{4}$$

式 (4) より、Mantel-Haenszel 型の重み $w_i = n_{i1}n_{i2}/(n_{i1}+n_{i2})$ を用いて重み付き推定量を求めると、Mantel-Haenszel 法により調整したリスク差（調整済みリスク差）が得られる。

$$\widehat{RD}_{MH} = \frac{\sum_{i=1}^{k} w_i * D_i}{\sum_{i=1}^{k} w_i} = \frac{\sum_{i=1}^{k}(x_{i1}n_{i2} - x_{i2}n_{i1})/n_{i+}}{\sum_{i=1}^{k} n_{i1}n_{i2}/n_{i+}} \tag{5}$$

次項では、調整済みリスク差の信頼区間を考える。

## 3. 調整済みリスク差の信頼区間の算出方法

### 3.1 Greenland 信頼区間

調整済みリスク差の信頼区間を算出するにあたって、まず、Greenland and Robins（1985）の方法に基づいた信頼区間（以下、Greenland 信頼区間）を考える。本方法では、式 (6) に従って調整済みリスク差の分散を計算し、この分散に基づいて信頼区間を算出する。

$$\widehat{\mathrm{var}}\left(\frac{\sum_{i=1}^{k} w_i * D_i}{\sum_{i=1}^{k} w_i}\right) = \frac{\sum_{i=1}^{k} w_i^2 \left(\frac{x_{i1}(n_{i1}-x_{i1})}{n_{i1}^3} + \frac{x_{i2}(n_{i2}-x_{i2})}{n_{i2}^3}\right)}{(\sum_{i=1}^{k} w_i)^2} \tag{6}$$

Sato（1989）では、Greenland 信頼区間はデータが疎な場合に一致性を持たないことを指摘している。これに対して Greenland & Robins は指摘を認めており、Greenland 信頼区間は層の例数が多い場合にのみ一致性を持つと訂正をしている。一方、Rothman, Lash, and Greenland（2012b）では、各層の各群の分母が全て 2 以上であれば、Greenland 信頼区間を用いることが可能だとしており、Greenland 信頼区間を使用する是非の判断が難しい状況である。今回の検討ではこの点も含めて確認を行う。

### 3.2 Sato 信頼区間

Sato（1989）では、Greenland and Robins（1985）の方法を改善し、層の数が多くデータが疎な場合でも妥当となるような調整済みリスク差の分散の推定方法を提案している。Sato（1989）の方法に基づいた信頼区間（以下、Sato 信頼区間）は式 (7) に従って調整済みリスク差の分散を計算し、この分散に基づいて信頼区間を算出する。この信頼区間は、すべての場合のデータにおいて妥当であることが Rothman, Lash, and Greenland（2012b）に記されている。

$$\widehat{var}^*(\widehat{RD}_{MH}) = \left(\widehat{RD}_{MH}\sum_{i=1}^{k} P_i + \sum_{i=1}^{k} Q_i\right)/(\sum_{i=1}^{k} w_i)^2 \tag{7}$$

このとき、$P_i$ 及び $Q_i$ はそれぞれ

$$\begin{aligned} P_i &= \left(n_{i1}^2 x_{i2} - n_{i2}^2 x_{i1} + \frac{n_{i1}n_{i2}(n_{i2}-n_{i1})}{2}\right)/(n_{i1}+n_{i2})^2, \\ Q_i &= \left(x_{i1}(n_{i2}-x_{i2}) - x_{i2}(n_{i1}-x_{i1})\right)/2(n_{i1}+n_{i2}) \end{aligned} \tag{8}$$

となる。

### 3.3 Newcombe 信頼区間

Yan and Su（2010）の方法で stratified Wilson 信頼区間に基づいて構築される stratified Newcombe 信頼区間（以下、Newcombe 信頼区間）では、調整済みリスク差（$\widehat{\mathrm{RD}}_{\mathrm{MH}}$）の $100(1-\alpha)\%$ 信頼下限と信頼上限 $(L, U)$ は以下の通りとなる。

$$L = \widehat{RD}_{MH} - z_{\alpha/2}\sqrt{\lambda_1 L_1(1-L_1) + \lambda_2 U_2(1-U_2)},$$
$$U = \widehat{RD}_{MH} + z_{\alpha/2}\sqrt{\lambda_1 U_1(1-U_1) + \lambda_2 L_2(1-L_2)} \tag{9}$$

このとき、$\lambda_1$ と $\lambda_2$ は

$$\lambda_1 = \sum_{i=1}^{k} w_i^2 / n_{i1} \ , \ \lambda_2 = \sum_{i=1}^{k} w_i^2 / n_{i2} \tag{10}$$

であり、$L_1, U_1$ は群 1 の割合（リスク）の stratified Wilson score 信頼上限と信頼下限を表し、$L_2, U_2$ は群 2 の割合（リスク）の str atifiedWilson score 信頼上限と信頼下限を表す。冒頭でも述べたが、層を考慮しないリスク差の信頼区間の性能評価の結果、飯塚、浜田 (2013) 及び飯塚、魚住、浜田 (2014) は、いずれも Newcombe スコアに基づく信頼区間を推奨している。

### 3.4 それぞれの信頼区間を算出する SAS プログラム

それぞれの信頼区間を算出する例を紹介するため、表 3.1 を考える。なお、下記サンプルデータは Rothman (2012a) から引用している。

表 3.1　サンプルデータ

| 投与群 | 層 1: イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | 8 | 98 | 106 |
| 群 2 | 5 | 115 | 120 |
| | | | 226 |

| 投与群 | 層 2: イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | 22 | 76 | 98 |
| 群 2 | 16 | 69 | 85 |
| | | | 183 |

まず、Greenland 信頼区間は STDRATE プロシジャにより算出が可能である。Greenland 信頼区間の算出プログラムをプログラム 3.1 に、実行結果を出力結果 3.1 に示す。

プログラム 3.1

```
data SAMPLEDATA1 ;
  input TREAT STRATA X TOTAL ;
  cards ;
    1 1  8 106
    1 2 22  98
    2 1  5 120
    2 2 16  85
  ;
run ;

ods graphics on;
proc stdrate data=SAMPLEDATA1
  method=mh
  stat  =risk
  effect=diff
  plots =all ;
  population group=TREAT event=X total=TOTAL ;
  strata STRATA / order=internal stats(cl=normal) effect ;
run;
ods graphics off;
```

出力結果 3.1（抜粋）

Risk Effect Estimates

| TREAT 1 | TREAT 2 | Risk Difference | 95% Normal Confidence Limits | | Standard Error | Z | Pr > \|Z\| |
|---|---|---|---|---|---|---|---|
| 0.1421 | 0.1072 | 0.0349 | -.027619 | 0.097419 | 0.0319 | 1.09 | 0.2739 |

次に、Sato 信頼区間及び Newcombe 信頼区間は FREQ プロシジャにより算出が可能である。Sato 信頼区間は riskdiff オプションへ「common column=1」を加えることによって算出が可能であり、Newcombe 信頼区間はさらに「cl=newcombe」を追記することによって算出可能となる。Sato 信頼区間及び Newcombe 信頼区間の算出プログラムをプログラム 3.2 に、実行結果を出力結果 3.2 に示す。なお、Greenland 信頼区間を STDRATE プロシジャで算出する際とは、データセットの形が異なることに留意されたい。

プログラム 3.2

```
data SAMPLEDATA2 ;
  input STRATA TREAT X N ;
  cards ;
  1 1 1 8
  1 1 2 98
  1 2 1 5
  1 2 2 115
  2 1 1 22
  2 1 2 76
  2 2 1 16
  2 2 2 69
  ;
run ;

proc freq data=SAMPLEDATA2 ;
  table  STRATA*TREAT*X / riskdiff(common column=1 cl=newcombe) nocol nopercent ;
  weight N ;
run ;
```

出力結果 3.2（抜粋）

共通比率（リスク）差

| 手法 | 値 | 標準誤差 | 95% 信頼限界 | |
|---|---|---|---|---|
| Mantel-Haenszel | 0.0349 | 0.0319 | -0.0276 | 0.0974 |
| Newcombe | 0.0349 | | -0.0302 | 0.1000 |
| 要約スコア | 0.0372 | 0.0299 | -0.0215 | 0.0959 |

列 1 (X = 1)

### 3.5 プログラムに関する注意点

前項では、調整済みリスク差とその信頼区間を STDRATE プロシジャ又は FREQ プロシジャを用いて算出する例を紹介したが、データの状態によっては、いくつかの信頼区間が算出されない場合があるため注意が必要である。

まず、表 3.2 のように、全ての層において、「全ての群が 0%」又は「全ての群が 100%」となった場合、STDRATE プロシジャ又は FREQ プロシジャを実施してもいずれの信頼区間も算出されない。すなわち、表 3.2 のような場合は全ての信頼区間が算出できない。

表 3.2　全ての信頼区間が算出できない場合

| 投与群 | 層 1: イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | 0 | 50 | 50 |
| 群 2 | 0 | 50 | 50 |
| | | | 100 |

| 投与群 | 層 2: イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | 0 | 40 | 40 |
| 群 2 | 0 | 60 | 60 |
| | | | 100 |

次に、表 3.3 のように、全ての層において、「全ての群が 0%又は 100%」となった場合、FREQ プロシジャによって算出される Sato 信頼区間のみ出力され、他の信頼区間は出力されない。

表 3.3　Sato 信頼区間のみ出力される場合

| 投与群 | 層 1: イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | 0 | 50 | 50 |
| 群 2 | 50 | 0 | 50 |
| | | | 100 |

| 投与群 | 層 2: イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | 40 | 0 | 40 |
| 群 2 | 0 | 60 | 60 |
| | | | 100 |

最後に、表 3.4 のように、いずれかの群において、「全ての層が 0%又は 100%」となった場合、FREQ プロシジャによって算出される Newcombe 信頼区間は出力されない。

表 3.4　Newcombe 信頼区間が出力されない場合

| 投与群 | 層 1: イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | 0 | 50 | 50 |
| 群 2 | 20 | 30 | 50 |
| | | | 100 |

| 投与群 | 層 2: イベントの有無 | | 合計 |
|---|---|---|---|
| | あり | なし | |
| 群 1 | 40 | 0 | 40 |
| 群 2 | 30 | 30 | 60 |
| | | | 100 |

次項のシミュレーションでは、いずれかの群で「全ての層が 0%」又は「全ての層が 100%」となった場合を除外して検討を行った。

## 4.　シミュレーションの設定

本項では、Greenland 信頼区間、Sato 信頼区間、Newcombe 信頼区間のそれぞれの被覆確率をシミュレーションにより算出し、性能評価を試みる。層別項目の層の数は 2 層、4 層、8 層の場合に分けて、各群の例数やイベント発生割合について以下のようなシミュレーションを設定した。なお、信頼区間を算出するための信頼係数は 95%、シミュレーション回数は 10000 回とした。

**2 層の場合のシミュレーション（表 5.1：No. 5.1.1.1～5.1.7.9）**

- 例数については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等例数**：各群において各層の例数が等しい場合として、20 例、50 例、100 例とした場合
  - ✧ **不等例数**：各群において（20 例、10 例）、（50 例、25 例）、（100 例、50 例）とした場合
- 割合については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等割合**：各群において各層の割合が等しい場合として、10%、30%、50%とした場合
  - ✧ **不等割合**：各群において、一方の層を 50%、他方の層を 5%、20%、40%とした場合

**2 層の場合のシミュレーション（表 5.1：No. 5.1.8.1～5.1.8.14）**

- 例数については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等例数**：各群において各層の例数が等しい場合として、10 例とした場合
  - ✧ **不等例数**：各群において（10 例、50 例）とした場合
- 割合については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等割合**：各群において各層の割合が等しい場合として、10%、50%とした場合
  - ✧ **不等割合**：各群において、一方の層を 50%、他方の層を 10%とした場合

**4 層の場合のシミュレーション（表 5.2：No. 5.2.1.1～5.2.7.9）**

- 例数については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等例数**：各群において各層の例数が等しい場合として、20 例、50 例、100 例とした場合
  - ✧ **不等例数**：各群において 3 つの層を 20 例、50 例、100 例、残りの層をその半分の例数とした場合
- 割合については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等割合**：各群において各層の割合が等しい場合として、10%、30%、50%とした場合
  - ✧ **不等割合**：各群において、3 つの層を 50%、残りの層を 5%、20%、40%とした場合

**4 層の場合のシミュレーション（表 5.2：No. 5.2.8.1～5.2.8.18）**

- 例数については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等例数**：各群において各層の例数が等しい場合として、10 例とした場合
  - ✧ **不等例数**：各群において半分の層を 20 例、残りの半分の層を 10 例とした場合
- 割合については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等割合**：各群において各層の割合が等しい場合として、10%、50%とした場合
  - ✧ **不等割合**：各群において、半分の層を 10%、残りの半分の層を 50%とした場合

**8 層の場合のシミュレーション（表 5.3：No. 5.3.1～5.3.18）**

- 例数については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等例数**：各群において各層の例数が等しい場合として、10 例、20 例とした場合
  - ✧ **不等例数**：各群において半分の層を 10 例、残りの半分の層を 20 例とした場合
- 割合については大きく分けて以下の 2 パターンを設定した。
  - ✧ **等割合**：各群において各層の割合が等しい場合として、10%、50%とした場合
  - ✧ **不等割合**：各群において、半分の層を 10%、残りの半分の層を 50%とした場合

**8 層の場合のシミュレーション（表 5.3：No. 5.3.19～5.3.22）**

- 例数及び割合について、以下のパターンを設定した。
  - ✧ **不等例数**：各群において半分の層を 10 例、残りの半分の層を 20 例、30 例、40 例とした場合
  - ✧ **不等割合**：
    - 各群において、半分の層を 10%、残りの半分の層を 40%、50%とした場合
    - 各群において、半分の層を 5%、残りの半分の層を 50%とした場合

## 5. 被覆確率の比較

各シミュレーションの設定に基づき出力した結果を、層ごとに表 5.1 から 5.3 にまとめた。表には、Greenland 信頼区間、Sato 信頼区間、Newcombe 信頼区間のそれぞれにおける被覆確率を示している。また、各層の群間差が全て 0 である設定（真の群間差が 0 になる設定）においては、CMH 検定の結果「有意差なし」となった割合（ $1-\alpha$ ）も併記した。なお、結果が 94.5%未満のものは格子掛け+太字（**XX.XX**）、97.5%を超えたものは網掛け+下線（XX.XX）にて表示した。

**2 層の場合のシミュレーション（表 5.1：No. 5.1.1.1～5.1.7.9）**

まず、Greenland 信頼区間においては、例数が小さい場合に被覆確率が 94.5%未満になることが多い傾向がみられたが、Newcombe 信頼区間では例数の影響はなかった。Sato 信頼区間においても、例数が小さい場合に被覆確率が 94.5%未満になることがあったが、Greenland 信頼区間よりも傾向は弱かった。

また、Sato 信頼区間は層内の割合の差が大きい場合（No. 5.1.3.1、5.1.3.4、5.1.3.7 等）に被覆確率が 97.5%を超える傾向がみられた。Newcombe 信頼区間では、層間の割合の差が大きい場合（No. 5.1.2.1、5.1.2.4、5.1.2.7、5.1.6.1、5.1.6.4、5.1.6.7 等）に被覆確率が 97.5%を超える傾向がみられた。

**2 層の場合のシミュレーション（表 5.1：No. 5.1.8.1～5.1.8.14）**

No. 5.1.1.1～5.1.7.9 のシミュレーション結果より、Greenland 信頼区間や Sato 信頼区間において例数が小さい場合に被覆確率が 94.5%未満になる傾向が示唆され、Sato 信頼区間や Newcombe 信頼区間においては割合の差が大きい場合に被覆確率が 97.5%を超える場合があったため、追加のシミュレーション No. 5.1.8.1～5.1.8.14 を行った。まず、Greenland 信頼区間においては、例数が小さい場合に被覆確率が 94.5%未満になることが多い傾向が明確になった。Sato 信頼区間も Greenland 信頼区間と同様の傾向がみられた一方、Newcombe 信頼区間では例数が小さい場合でも被覆確率が 94.5%未満になることは少なかった。

また、層内の割合の差が大きい場合のうち、No. 5.1.8.4、5.1.8.9 及び 5.1.8.14 等では、Sato 信頼区間にて被覆確率が大きくなる傾向がみられ、層間の割合の差が大きい場合のうち、No. 5.1.8.3、5.1.8.8 及び 5.1.8.12 等では Newcombe 信頼区間にて被覆確率が大きくなる傾向がみられたが、No. 5.1.8.13 の様に全ての信頼区間について被覆確率が 94.5%未満となる場合もみられることから、例数と割合のパターンにより結果が変わりうる点に注意する必要があることが分かる。

**4 層の場合のシミュレーション（表 5.2：No. 5.2.1.1～5.2.7.9）**

4 層の設定のシミュレーションの結果では、各群の各層にある程度の例数がある場合、割合の設定に関わらず、3 種類の信頼区間のいずれにおいても被覆確率は概ね 94.5%～97.5%の範囲内であったが、例数が少ない場合に Greenland 信頼区間のみ被覆確率が 94.5%未満となることが多かった。

**4 層の場合のシミュレーション（表 5.2：No. 5.2.8.1～5.2.8.18）**

シミュレーション No. 5.2.8.1～5.2.8.18 では各群の各層で例数が小さい場合に着目した。まず、No. 5.2.1.1～5.2.7.9 の結果と同様、例数が小さい場合に Greenland 信頼区間は被覆確率が 94.5%未満となることが多かった。次に、Sato 信頼区間は、2 層の際と同様、層内の割合の差が大きい場合（No. 5.2.8.4、5.2.8.9、5.2.8.11 等）に被覆確率が 97.5%を超える傾向があった。また、Newcombe 信頼区間は、2 層の際と同様、層間の割合の差が大きい場合に被覆確率が大きくなる傾向があることに加え、各群の各層の割合が小さい場合（No. 5.2.8.1、5.2.8.5、5.2.8.12 等）に被覆確率が 97.5%を超える傾向があった。なお、シミュレーション No. 5.2.8.2 のように、各群の各層で例数が小さく、割合が大きい場合には、どの信頼区間も 94.5%未満の被覆確率となった。ちなみに、CMH 検定の結果も 94.5%を下回った。

**8 層の場合のシミュレーション（表 5.3：No. 5.3.1～5.3.22）**

最後に、8 層の設定のシミュレーションでは、網羅的にシミュレーションを実施すると場合の数が膨大になるため、2 層と 4 層の場合のシミュレーション結果を踏まえ、例数が小さい場合に着目して検討を行った。まず、Greenland 信頼区間においては、2 層や 4 層の検討結果と同様、各群の各層の例数が小さい場合は、ほとんどのシミュレーション設定において被覆確率が 94.5%未満となったが、Sato 信頼区間及び Newcombe 信頼区間の被覆確率は、概ね 94.5%～97.5%の範囲内であった。また、Sato 信頼区間は、層内の割合の差が大きい場合（No. 5.3.4、5.3.8、5.3.13 等）に被覆確率が 97.5%を超える傾向があった。なお、層内の割合の差が縮まれば、被覆確率は名義水準（95%）に近づくことが No.5.3.19 より示唆される。さらに、No. 5.3.20 や 5.3.21 からは例数を増やすことにより、被覆確率が名義水準に近づく傾向であることが確認できる。Newcombe 信頼区間においては、2 層や 4 層の検討結果と同様、層間の割合の差が大きい場合（No. 5.3.18、5.3.22 等）に被覆確率が大きくなる傾向があった。

**シミュレーション結果のまとめ**

各群の各層の例数が小さい場合には、Greenland 信頼区間の被覆確率は名義水準を下回る可能性が高くなる一方、Newcombe 信頼区間は層の数や例数によらず被覆確率を名義水準以上に維持することが示唆された。Rothman, Lash, and Greenland（2012b）では、各層の各群の分母が全て 2 以上であれば、Greenland 信頼区間を用いることが可能だと述べているが、シミュレーションの結果からは必ずしもそうではなく、例数が少ない場合においては Greenland 信頼区間の適用は注意が必要であるといえる。また、Sato 信頼区間は、Greenland 信頼区間と比べて、層の例数が小さいことによる影響を受けにくいことが示唆されたが、層内の割合の差が大きい場合、名義水準を上回る傾向も見受けられた。さらに、Newcombe 信頼区間においても、層間の割合の差が大きい場合には、名義水準を上回る傾向であった。

ちなみに、本項のシミュレーションは SAS マクロ %MH() にて実施した（SAS プログラムは本稿の最後で紹介）。前項のシミュレーションで設定した状況以外の場合においても、本マクロを使用することで 3 種類の信頼区間の被覆確率を計算することが出来る。実際の臨床試験計画時等で活用いただければ幸いである。

表 5.1　2 層の場合の各設定における信頼区間の被覆確率

| | No. | 設定※ | | 被覆確率 | | | 1-α |
|---|---|---|---|---|---|---|---|
| | | 例数 | 割合 | Greenland | Sato | Newcombe | CMH 検定 |
| 等例数、等割合 | 5.1.1.1 | (20, 20)<br>(20, 20) | (10%, 10%)<br>(10%, 10%) | 95.67 | 95.77 | **98.36** | 96.62 |
| | 5.1.1.2 | (20, 20)<br>(20, 20) | (30%, 30%)<br>(30%, 30%) | 94.06 | 94.45 | 95.40 | 95.31 |
| | 5.1.1.3 | (20, 20)<br>(20, 20) | (50%, 50%)<br>(50%, 50%) | 94.60 | 94.64 | 94.64 | 94.64 |
| | 5.1.1.4 | (50, 50)<br>(50, 50) | (10%, 10%)<br>(10%, 10%) | 94.82 | 94.85 | 95.78 | 94.91 |
| | 5.1.1.5 | (50, 50)<br>(50, 50) | (30%, 30%)<br>(30%, 30%) | 94.80 | 94.89 | 95.09 | 95.19 |
| | 5.1.1.6 | (50, 50)<br>(50, 50) | (50%, 50%)<br>(50%, 50%) | 94.55 | 94.55 | 94.55 | 94.55 |
| | 5.1.1.7 | (100, 100)<br>(100, 100) | (10%, 10%)<br>(10%, 10%) | 95.03 | 95.12 | 95.49 | 95.46 |
| | 5.1.1.8 | (100, 100)<br>(100, 100) | (30%, 30%)<br>(30%, 30%) | 95.02 | 95.06 | 95.20 | 95.25 |
| | 5.1.1.9 | (100, 100)<br>(100, 100) | (50%, 50%)<br>(50%, 50%) | 95.20 | 95.20 | 95.20 | 95.20 |
| 等例数、層内で等割合、層間で不等割合 | 5.1.2.1 | (20, 20)<br>(20, 20) | (50%, 5%)<br>(50%, 5%) | 93.95 | 94.74 | **98.17** | 95.07 |
| | 5.1.2.2 | (20, 20)<br>(20, 20) | (50%, 20%)<br>(50%, 20%) | 94.18 | 94.55 | 96.49 | 95.46 |
| | 5.1.2.3 | (20, 20)<br>(20, 20) | (50%, 40%)<br>(50%, 40%) | 94.61 | 94.78 | 94.87 | 94.86 |
| | 5.1.2.4 | (50, 50)<br>(50, 50) | (50%, 5%)<br>(50%, 5%) | 94.52 | 94.83 | **97.91** | 95.08 |
| | 5.1.2.5 | (50, 50)<br>(50, 50) | (50%, 20%)<br>(50%, 20%) | 94.87 | 94.96 | 96.43 | 95.17 |
| | 5.1.2.6 | (50, 50)<br>(50, 50) | (50%, 40%)<br>(50%, 40%) | 94.93 | 94.94 | 94.95 | 94.95 |
| | 5.1.2.7 | (100, 100)<br>(100, 100) | (50%, 5%)<br>(50%, 5%) | 94.88 | 94.99 | **97.77** | 95.04 |
| | 5.1.2.8 | (100, 100)<br>(100, 100) | (50%, 20%)<br>(50%, 20%) | 94.93 | 95.01 | 96.28 | 95.21 |
| | 5.1.2.9 | (100, 100)<br>(100, 100) | (50%, 40%)<br>(50%, 40%) | 95.34 | 95.38 | 95.43 | 95.42 |
| 等例数、層内及び層間で不等割合 | 5.1.3.1 | (20, 20)<br>(20, 20) | (50%, 5%)<br>(5%, 50%) | 93.93 | **98.34** | **97.95** | - |
| | 5.1.3.2 | (20, 20)<br>(20, 20) | (50%, 20%)<br>(20%, 50%) | 93.91 | 96.57 | 96.47 | - |
| | 5.1.3.3 | (20, 20)<br>(20, 20) | (50%, 40%)<br>(40%, 50%) | 94.70 | 94.94 | 94.93 | - |
| | 5.1.3.4 | (50, 50)<br>(50, 50) | (50%, 5%)<br>(5%, 50%) | 94.59 | **98.67** | **97.68** | - |
| | 5.1.3.5 | (50, 50)<br>(50, 50) | (50%, 20%)<br>(20%, 50%) | 94.82 | 96.84 | 96.38 | - |
| | 5.1.3.6 | (50, 50)<br>(50, 50) | (50%, 40%)<br>(40%, 50%) | 94.79 | 94.87 | 94.80 | - |
| | 5.1.3.7 | (100, 100)<br>(100, 100) | (50%, 5%)<br>(5%, 50%) | 94.88 | **98.83** | **97.82** | - |
| | 5.1.3.8 | (100, 100)<br>(100, 100) | (50%, 20%)<br>(20%, 50%) | 94.94 | 96.98 | 96.18 | - |
| | 5.1.3.9 | (100, 100)<br>(100, 100) | (50%, 40%)<br>(40%, 50%) | 95.36 | 95.46 | 95.45 | - |

※ 各行の上段：（$n_{11}, n_{12}$）（$p_{11}, p_{12}$）、各行の下段：（$n_{21}, n_{22}$）（$p_{21}, p_{22}$）、α：Type I error rate

表 5.1　2 層の場合の各設定における信頼区間の被覆確率（続き）

| | No. | 設定※ | | 被覆確率 | | | 1-α |
|---|---|---|---|---|---|---|---|
| | | 例数 | 割合 | Greenland | Sato | Newcombe | CMH 検定 |
| 不等例数、等割合 | 5.1.4.1 | (20, 10)<br>(20, 10) | (10%, 10%)<br>(10%, 10%) | 97.19 | 97.26 | 98.78 | 97.33 |
| | 5.1.4.2 | (20, 10)<br>(20, 10) | (30%, 30%)<br>(30%, 30%) | 93.34 | 93.99 | 95.24 | 95.14 |
| | 5.1.4.3 | (20, 10)<br>(20, 10) | (50%, 50%)<br>(50%, 50%) | 94.57 | 94.98 | 95.15 | 95.12 |
| | 5.1.4.4 | (50, 25)<br>(50, 25) | (10%, 10%)<br>(10%, 10%) | 94.80 | 94.93 | 95.95 | 95.28 |
| | 5.1.4.5 | (50, 25)<br>(50, 25) | (30%, 30%)<br>(30%, 30%) | 94.60 | 94.84 | 95.34 | 95.31 |
| | 5.1.4.6 | (50, 25)<br>(50, 25) | (50%, 50%)<br>(50%, 50%) | 94.20 | 94.26 | 94.20 | 95.36 |
| | 5.1.4.7 | (100, 50)<br>(100, 50) | (10%, 10%)<br>(10%, 10%) | 94.95 | 95.01 | 95.66 | 95.15 |
| | 5.1.4.8 | (100, 50)<br>(100, 50) | (30%, 30%)<br>(30%, 30%) | 94.96 | 95.02 | 95.15 | 95.17 |
| | 5.1.4.9 | (100, 50)<br>(100, 50) | (50%, 50%)<br>(50%, 50%) | 94.41 | 94.45 | 94.41 | 94.97 |
| 不等例数、層内で等割合、層間で不等割合 | 5.1.5.1 | (20, 10)<br>(20, 10) | (50%, 5%)<br>(50%, 5%) | 93.77 | 95.04 | 97.77 | 95.52 |
| | 5.1.5.2 | (20, 10)<br>(20, 10) | (50%, 20%)<br>(50%, 20%) | 93.69 | 94.18 | 96.20 | 95.42 |
| | 5.1.5.3 | (20, 10)<br>(20, 10) | (50%, 40%)<br>(50%, 40%) | 94.29 | 94.85 | 95.21 | 95.17 |
| | 5.1.5.4 | (50, 25)<br>(50, 25) | (50%, 5%)<br>(50%, 5%) | 94.63 | 94.95 | 97.24 | 95.32 |
| | 5.1.5.5 | (50, 25)<br>(50, 25) | (50%, 20%)<br>(50%, 20%) | 94.69 | 94.91 | 95.71 | 95.27 |
| | 5.1.5.6 | (50, 25)<br>(50, 25) | (50%, 40%)<br>(50%, 40%) | 94.45 | 94.50 | 94.45 | 94.98 |
| | 5.1.5.7 | (100, 50)<br>(100, 50) | (50%, 5%)<br>(50%, 5%) | 94.86 | 94.99 | 97.47 | 95.07 |
| | 5.1.5.8 | (100, 50)<br>(100, 50) | (50%, 20%)<br>(50%, 20%) | 94.72 | 94.82 | 95.87 | 95.06 |
| | 5.1.5.9 | (100, 50)<br>(100, 50) | (50%, 40%)<br>(50%, 40%) | 94.67 | 94.70 | 94.67 | 94.80 |
| 不等例数、層内で等割合、層間で不等割合 | 5.1.6.1 | (10, 20)<br>(10, 20) | (50%, 5%)<br>(50%, 5%) | 93.01 | 95.03 | 98.56 | 95.37 |
| | 5.1.6.2 | (10, 20)<br>(10, 20) | (50%, 20%)<br>(50%, 20%) | 93.50 | 94.40 | 96.30 | 95.35 |
| | 5.1.6.3 | (10, 20)<br>(10, 20) | (50%, 40%)<br>(50%, 40%) | 94.13 | 94.78 | 95.27 | 95.21 |
| | 5.1.6.4 | (25, 50)<br>(25, 50) | (50%, 5%)<br>(50%, 5%) | 94.20 | 94.97 | 98.18 | 94.94 |
| | 5.1.6.5 | (25, 50)<br>(25, 50) | (50%, 20%)<br>(50%, 20%) | 94.54 | 94.73 | 96.15 | 95.16 |
| | 5.1.6.6 | (25, 50)<br>(25, 50) | (50%, 40%)<br>(50%, 40%) | 94.42 | 94.43 | 94.43 | 94.77 |
| | 5.1.6.7 | (50, 100)<br>(50, 100) | (50%, 5%)<br>(50%, 5%) | 94.75 | 95.10 | 97.99 | 95.12 |
| | 5.1.6.8 | (50, 100)<br>(50, 100) | (50%, 20%)<br>(50%, 20%) | 95.03 | 95.09 | 96.15 | 95.25 |
| | 5.1.6.9 | (50, 100)<br>(50, 100) | (50%, 40%)<br>(50%, 40%) | 94.85 | 94.85 | 94.85 | 94.86 |

※ 各行の上段：$(n_{11}, n_{12})$　$(p_{11}, p_{12})$、各行の下段：$(n_{21}, n_{22})$　$(p_{21}, p_{22})$、α：Type I error rate

表 5.1　2 層の場合の各設定における信頼区間の被覆確率（続き）

| | No. | 設定※ | | 被覆確率 | | | 1-α |
|---|---|---|---|---|---|---|---|
| | | 例数 | 割合 | Greenland | Sato | Newcombe | CMH 検定 |
| 不等例数、層内及び層間で不等割合 | 5.1.7.1 | (20, 10)<br>(20, 10) | (50%, 5%)<br>(5%, 50%) | 93.36 | 98.04 | 97.99 | - |
| | 5.1.7.2 | (20, 10)<br>(20, 10) | (50%, 20%)<br>(20%, 50%) | 93.54 | 96.14 | 96.36 | - |
| | 5.1.7.3 | (20, 10)<br>(20, 10) | (50%, 40%)<br>(40%, 50%) | 94.29 | 95.11 | 95.28 | - |
| | 5.1.7.4 | (50, 25)<br>(50, 25) | (50%, 5%)<br>(5%, 50%) | 94.37 | 98.38 | 97.72 | - |
| | 5.1.7.5 | (50, 25)<br>(50, 25) | (50%, 20%)<br>(20%, 50%) | 94.67 | 96.62 | 96.12 | - |
| | 5.1.7.6 | (50, 25)<br>(50, 25) | (50%, 40%)<br>(40%, 50%) | 94.93 | 95.35 | 95.44 | - |
| | 5.1.7.7 | (100, 50)<br>(100, 50) | (50%, 5%)<br>(5%, 50%) | 94.80 | 98.56 | 97.52 | - |
| | 5.1.7.8 | (100, 50)<br>(100, 50) | (50%, 20%)<br>(20%, 50%) | 94.70 | 96.80 | 96.13 | - |
| | 5.1.7.9 | (100, 50)<br>(100, 50) | (50%, 40%)<br>(40%, 50%) | 94.65 | 94.82 | 94.65 | - |
| 追加検討 | 5.1.8.1 | (10, 10)<br>(10, 10) | (10%, 10%)<br>(10%, 10%) | 98.98 | 99.12 | 99.78 | - |
| | 5.1.8.2 | (10, 10)<br>(10, 10) | (50%, 50%)<br>(50%, 50%) | 92.29 | 93.16 | 96.00 | - |
| | 5.1.8.3 | (10, 10)<br>(10, 10) | (50%, 10%)<br>(50%, 10%) | 92.77 | 94.15 | 97.31 | - |
| | 5.1.8.4 | (10, 10)<br>(10, 10) | (50%, 10%)<br>(10%, 50%) | 92.53 | 97.39 | 97.43 | - |
| | 5.1.8.5 | (50, 10)<br>(50, 10) | (10%, 10%)<br>(10%, 10%) | 94.37 | 94.57 | 96.94 | - |
| | 5.1.8.6 | (50, 10)<br>(50, 10) | (50%, 50%)<br>(50%, 50%) | 94.48 | 94.50 | 94.50 | - |
| | 5.1.8.7 | (50, 10)<br>(50, 10) | (50%, 10%)<br>(50%, 10%) | 94.48 | 94.62 | 95.96 | - |
| | 5.1.8.8 | (50, 10)<br>(50, 10) | (10%, 50%)<br>(10%, 50%) | 94.22 | 94.85 | 96.89 | - |
| | 5.1.8.9 | (50, 10)<br>(50, 10) | (50%, 10%)<br>(10%, 50%) | 94.42 | 97.02 | 96.50 | - |
| | 5.1.8.10 | (50, 10)<br>(10, 50) | (10%, 10%)<br>(10%, 10%) | 94.75 | 95.12 | 98.37 | - |
| | 5.1.8.11 | (50, 10)<br>(10, 50) | (50%, 50%)<br>(50%, 50%) | 92.66 | 93.43 | 94.58 | - |
| | 5.1.8.12 | (50, 10)<br>(10, 50) | (50%, 10%)<br>(50%, 10%) | 92.22 | 93.12 | 97.30 | - |
| | 5.1.8.13 | (50, 10)<br>(10, 50) | (10%, 50%)<br>(50%, 10%) | 92.34 | 94.41 | 94.26 | - |
| | 5.1.8.14 | (50, 10)<br>(10, 50) | (50%, 10%)<br>(10%, 50%) | 94.48 | 99.37 | 99.41 | - |

※ 各行の上段：（$n_{11}, n_{12}$）（$p_{11}, p_{12}$）、各行の下段：（$n_{21}, n_{22}$）（$p_{21}, p_{22}$）、α：Type I error rate

表 5.2　4 層の場合の各設定における信頼区間の被覆確率

| | No. | 設定※ | | 被覆確率 | | | 1-α |
|---|---|---|---|---|---|---|---|
| | | 例数 | 割合 | Greenland | Sato | Newcombe | CMH 検定 |
| 等例数、等割合 | 5.2.1.1 | (20,20,20,20)<br>(20,20,20,20) | (10%,10%,10%,10%)<br>(10%,10%,10%,10%) | 94.24 | 94.89 | 96.71 | 95.10 |
| | 5.2.1.2 | (20,20,20,20)<br>(20,20,20,20) | (30%,30%,30%,30%)<br>(30%,30%,30%,30%) | 94.11 | 94.53 | 95.00 | 94.97 |
| | 5.2.1.3 | (20,20,20,20)<br>(20,20,20,20) | (50%,50%,50%,50%)<br>(50%,50%,50%,50%) | 94.13 | 94.81 | 94.92 | 94.91 |
| | 5.2.1.4 | (50,50,50,50)<br>(50,50,50,50) | (10%,10%,10%,10%)<br>(10%,10%,10%,10%) | 94.65 | 94.78 | 95.04 | 94.97 |
| | 5.2.1.5 | (50,50,50,50)<br>(50,50,50,50) | (30%,30%,30%,30%)<br>(30%,30%,30%,30%) | 94.54 | 94.69 | 94.78 | 94.85 |
| | 5.2.1.6 | (50,50,50,50)<br>(50,50,50,50) | (50%,50%,50%,50%)<br>(50%,50%,50%,50%) | 94.83 | 94.85 | 94.85 | 94.85 |
| | 5.2.1.7 | (100,100,100,100)<br>(100,100,100,100) | (10%,10%,10%,10%)<br>(10%,10%,10%,10%) | 94.58 | 94.75 | 94.83 | 94.81 |
| | 5.2.1.8 | (100,100,100,100)<br>(100,100,100,100) | (30%,30%,30%,30%)<br>(30%,30%,30%,30%) | 94.87 | 94.95 | 94.96 | 94.98 |
| | 5.2.1.9 | (100,100,100,100)<br>(100,100,100,100) | (50%,50%,50%,50%)<br>(50%,50%,50%,50%) | 94.70 | 94.70 | 94.70 | 94.70 |
| 等例数、層内で等割合、層間で不等割合 | 5.2.2.1 | (20,20,20,20)<br>(20,20,20,20) | (50%,50%,50%,5%)<br>(50%,50%,50%,5%) | 94.50 | 94.99 | 97.11 | 95.49 |
| | 5.2.2.2 | (20,20,20,20)<br>(20,20,20,20) | (50%,50%,50%,20%)<br>(50%,50%,50%,20%) | 94.28 | 94.54 | 96.05 | 94.90 |
| | 5.2.2.3 | (20,20,20,20)<br>(20,20,20,20) | (50%,50%,50%,40%)<br>(50%,50%,50%,40%) | 93.98 | 94.87 | 95.09 | 95.06 |
| | 5.2.2.4 | (50,50,50,50)<br>(50,50,50,50) | (50%,50%,50%,5%)<br>(50%,50%,50%,5%) | 95.13 | 95.29 | 96.80 | 95.33 |
| | 5.2.2.5 | (50,50,50,50)<br>(50,50,50,50) | (50%,50%,50%,20%)<br>(50%,50%,50%,20%) | 94.81 | 94.88 | 95.93 | 95.02 |
| | 5.2.2.6 | (50,50,50,50)<br>(50,50,50,50) | (50%,50%,50%,40%)<br>(50%,50%,50%,40%) | 94.90 | 94.94 | 94.95 | 94.95 |
| | 5.2.2.7 | (100,100,100,100)<br>(100,100,100,100) | (50%,50%,50%,5%)<br>(50%,50%,50%,5%) | 94.96 | 95.05 | 96.84 | 95.17 |
| | 5.2.2.8 | (100,100,100,100)<br>(100,100,100,100) | (50%,50%,50%,20%)<br>(50%,50%,50%,20%) | 95.01 | 95.05 | 95.95 | 95.06 |
| | 5.2.2.9 | (100,100,100,100)<br>(100,100,100,100) | (50%,50%,50%,40%)<br>(50%,50%,50%,40%) | 94.89 | 94.89 | 94.89 | 94.89 |
| 等例数、層内及び層間で不等割合 | 5.2.3.1 | (20,20,20,20)<br>(20,20,20,20) | (50%,50%,50%,5%)<br>(5%,50%,50%,50%) | 94.10 | 96.86 | 96.77 | - |
| | 5.2.3.2 | (20,20,20,20)<br>(20,20,20,20) | (50%,50%,50%,20%)<br>(20%,50%,50%,50%) | 94.13 | 95.83 | 96.06 | - |
| | 5.2.3.3 | (20,20,20,20)<br>(20,20,20,20) | (50%,50%,50%,40%)<br>(40%,50%,50%,50%) | 93.94 | 94.84 | 95.01 | - |
| | 5.2.3.4 | (50,50,50,50)<br>(50,50,50,50) | (50%,50%,50%,5%)<br>(5%,50%,50%,50%) | 94.85 | 96.91 | 96.76 | - |
| | 5.2.3.5 | (50,50,50,50)<br>(50,50,50,50) | (50%,50%,50%,20%)<br>(20%,50%,50%,50%) | 94.67 | 95.80 | 95.77 | - |
| | 5.2.3.6 | (50,50,50,50)<br>(50,50,50,50) | (50%,50%,50%,40%)<br>(40%,50%,50%,50%) | 94.81 | 94.90 | 94.90 | - |
| | 5.2.3.7 | (100,100,100,100)<br>(100,100,100,100) | (50%,50%,50%,5%)<br>(5%,50%,50%,50%) | 94.70 | 96.98 | 96.65 | - |
| | 5.2.3.8 | (100,100,100,100)<br>(100,100,100,100) | (50%,50%,50%,20%)<br>(20%,50%,50%,50%) | 94.74 | 95.72 | 95.67 | - |
| | 5.2.3.9 | (100,100,100,100)<br>(100,100,100,100) | (50%,50%,50%,40%)<br>(40%,50%,50%,50%) | 94.87 | 94.87 | 94.87 | - |

※ 各行の上段：（$n_{11}, n_{12}, n_{13}, n_{14}$）（$p_{11}, p_{12}, p_{13}, p_{14}$）、各行の下段：（$n_{21}, n_{22}, n_{23}, n_{24}$）（$p_{21}, p_{22}, p_{23}, p_{24}$）、α：Type I error rate

表 5.2　4 層の場合の各設定における信頼区間の被覆確率（続き）

| | No. | 設定※ | | 被覆確率 | | | 1-α |
|---|---|---|---|---|---|---|---|
| | | 例数 | 割合 | Greenland | Sato | Newcombe | CMH 検定 |
| 不等例数、等割合 | 5.2.4.1 | (20,20,20,10)<br>(20,20,20,10) | (10%,10%,10%,10%)<br>(10%,10%,10%,10%) | 94.27 | 95.11 | 97.54 | 95.70 |
| | 5.2.4.2 | (20,20,20,10)<br>(20,20,20,10) | (30%,30%,30%,30%)<br>(30%,30%,30%,30%) | 94.39 | 94.73 | 94.94 | 95.09 |
| | 5.2.4.3 | (20,20,20,10)<br>(20,20,20,10) | (50%,50%,50%,50%)<br>(50%,50%,50%,50%) | 94.59 | 94.93 | 94.97 | 94.97 |
| | 5.2.4.4 | (50,50,50,25)<br>(50,50,50,25) | (10%,10%,10%,10%)<br>(10%,10%,10%,10%) | 94.78 | 94.92 | 95.47 | 95.15 |
| | 5.2.4.5 | (50,50,50,25)<br>(50,50,50,25) | (30%,30%,30%,30%)<br>(30%,30%,30%,30%) | 94.78 | 94.91 | 95.03 | 95.11 |
| | 5.2.4.6 | (50,50,50,25)<br>(50,50,50,25) | (50%,50%,50%,50%)<br>(50%,50%,50%,50%) | 94.94 | 95.21 | 95.27 | 95.27 |
| | 5.2.4.7 | (100,100,100,50)<br>(100,100,100,50) | (10%,10%,10%,10%)<br>(10%,10%,10%,10%) | 94.91 | 95.02 | 95.20 | 95.08 |
| | 5.2.4.8 | (100,100,100,50)<br>(100,100,100,50) | (30%,30%,30%,30%)<br>(30%,30%,30%,30%) | 94.91 | 95.00 | 95.08 | 95.10 |
| | 5.2.4.9 | (100,100,100,50)<br>(100,100,100,50) | (50%,50%,50%,50%)<br>(50%,50%,50%,50%) | 94.67 | 94.69 | 94.67 | 94.68 |
| 不等例数、層内で等割合、層間で不等割合 | 5.2.4.1 | (20,20,20,10)<br>(20,20,20,10) | (50%,50%,50%,5%)<br>(50%,50%,50%,5%) | 94.46 | 94.69 | 96.57 | 95.12 |
| | 5.2.5.2 | (20,20,20,10)<br>(20,20,20,10) | (50%,50%,50%,20%)<br>(50%,50%,50%,20%) | 94.26 | 94.98 | 95.56 | 95.44 |
| | 5.2.5.3 | (20,20,20,10)<br>(20,20,20,10) | (50%,50%,50%,40%)<br>(50%,50%,50%,40%) | 94.58 | 94.95 | 95.03 | 95.02 |
| | 5.2.5.4 | (50,50,50,25)<br>(50,50,50,25) | (50%,50%,50%,5%)<br>(50%,50%,50%,5%) | 94.84 | 95.21 | 96.54 | 95.54 |
| | 5.2.5.5 | (50,50,50,25)<br>(50,50,50,25) | (50%,50%,50%,20%)<br>(50%,50%,50%,20%) | 94.74 | 94.91 | 95.82 | 95.10 |
| | 5.2.5.6 | (50,50,50,25)<br>(50,50,50,25) | (50%,50%,50%,40%)<br>(50%,50%,50%,40%) | 94.77 | 95.20 | 95.32 | 95.29 |
| | 5.2.5.7 | (100,100,100,50)<br>(100,100,100,50) | (50%,50%,50%,5%)<br>(50%,50%,50%,5%) | 95.34 | 95.40 | 96.15 | 95.41 |
| | 5.2.5.8 | (100,100,100,50)<br>(100,100,100,50) | (50%,50%,50%,20%)<br>(50%,50%,50%,20%) | 94.97 | 95.16 | 95.34 | 95.23 |
| | 5.2.5.9 | (100,100,100,50)<br>(100,100,100,50) | (50%,50%,50%,40%)<br>(50%,50%,50%,40%) | 94.79 | 94.79 | 94.79 | 94.79 |
| 不等例数、層内で等割合、層間で不等割合 | 5.2.6.1 | (10,20,20,20)<br>(10,20,20,20) | (50%,50%,50%,5%)<br>(50%,50%,50%,5%) | 94.28 | 95.07 | 97.14 | 95.58 |
| | 5.2.6.2 | (10,20,20,20)<br>(10,20,20,20) | (50%,50%,50%,20%)<br>(50%,50%,50%,20%) | 94.15 | 94.69 | 96.06 | 95.16 |
| | 5.2.6.3 | (10,20,20,20)<br>(10,20,20,20) | (50%,50%,50%,40%)<br>(50%,50%,50%,40%) | 94.27 | 94.73 | 94.80 | 94.80 |
| | 5.2.6.4 | (25,50,50,50)<br>(25,50,50,50) | (50%,50%,50%,5%)<br>(50%,50%,50%,5%) | 94.74 | 95.10 | 96.89 | 95.31 |
| | 5.2.6.5 | (25,50,50,50)<br>(25,50,50,50) | (50%,50%,50%,20%)<br>(50%,50%,50%,20%) | 94.69 | 94.89 | 95.87 | 94.99 |
| | 5.2.6.6 | (25,50,50,50)<br>(25,50,50,50) | (50%,50%,50%,40%)<br>(50%,50%,50%,40%) | 94.43 | 94.91 | 95.19 | 95.12 |
| | 5.2.6.7 | (50,100,100,100)<br>(50,100,100,100) | (50%,50%,50%,5%)<br>(50%,50%,50%,5%) | 95.04 | 95.13 | 96.99 | 95.19 |
| | 5.2.6.8 | (50,100,100,100)<br>(50,100,100,100) | (50%,50%,50%,20%)<br>(50%,50%,50%,20%) | 94.84 | 94.88 | 95.75 | 94.95 |
| | 5.2.6.9 | (50,100,100,100)<br>(50,100,100,100) | (50%,50%,50%,40%)<br>(50%,50%,50%,40%) | 94.74 | 94.74 | 94.74 | 94.74 |

※ 各行の上段：（$n_{11}, n_{12}, n_{13}, n_{14}$）（$p_{11}, p_{12}, p_{13}, p_{14}$）、各行の下段：（$n_{21}, n_{22}, n_{23}, n_{24}$）（$p_{21}, p_{22}, p_{23}, p_{24}$）、α：Type I error rate

表 5.2　4 層の場合の各設定における信頼区間の被覆確率（続き）

| | No. | 設定※ | | 被覆確率 | | | 1-α |
|---|---|---|---|---|---|---|---|
| | | 例数 | 割合 | Greenland | Sato | Newcombe | CMH 検定 |
| 不等例数、層内及び層間で不等割合 | 5.2.7.1 | (20,20,20,10)<br>(20,20,20,10) | (50%,50%,50%,5%)<br>(5%,50%,50%,50%) | 94.10 | 96.34 | 96.23 | - |
| | 5.2.7.2 | (20,20,20,10)<br>(20,20,20,10) | (50%,50%,50%,20%)<br>(20%,50%,50%,50%) | 94.14 | 95.57 | 95.71 | - |
| | 5.2.7.3 | (20,20,20,10)<br>(20,20,20,10) | (50%,50%,50%,40%)<br>(40%,50%,50%,50%) | 94.50 | 95.02 | 95.05 | - |
| | 5.2.7.4 | (50,50,50,25)<br>(50,50,50,25) | (5%,50%,50%,50%) | 94.67 | 96.66 | 96.55 | - |
| | 5.2.7.5 | (50,50,50,25)<br>(50,50,50,25) | (50%,50%,50%,20%)<br>(20%,50%,50%,50%) | 94.73 | 95.70 | 95.62 | - |
| | 5.2.7.6 | (50,50,50,25)<br>(50,50,50,25) | (50%,50%,50%,40%)<br>(40%,50%,50%,50%) | 94.81 | 94.83 | 94.81 | - |
| | 5.2.7.7 | (100,100,100,50)<br>(100,100,100,50) | (50%,50%,50%,5%)<br>(5%,50%,50%,50%) | 94.71 | 96.64 | 96.37 | - |
| | 5.2.7.8 | (100,100,100,50)<br>(100,100,100,50) | (50%,50%,50%,20%)<br>(20%,50%,50%,50%) | 94.85 | 95.77 | 95.60 | - |
| | 5.2.7.9 | (100,100,100,50)<br>(100,100,100,50) | (50%,50%,50%,40%)<br>(4 ) | 94.77 | 94.82 | 94.77 | - |
| 追加検討 | 5.2.8.1 | (10,10,10,10)<br>(10,10,10,10) | (10%,10%,10%,10%)<br>(10%,10%,10%,10%) | 95.45 | 95.88 | 98.96 | 96.64 |
| | 5.2.8.2 | (10,10,10,10)<br>(10,10,10,10) | (50%,50%,50%,50%)<br>(50%,50%,50%,50%) | 93.66 | 94.28 | 94.30 | 94.30 |
| | 5.2.8.3 | (10,10,10,10)<br>(10,10,10,10) | (50%,50%,10%,10%)<br>(50%,50%,10%,1 ) | 93.55 | 94.86 | 97.48 | 95.38 |
| | 5.2.8.4 | (10,10,10,10)<br>(10,10,10,10) | (50%,50%,10%,10%)<br>(10%,10%,50%,50%) | 93.08 | 97.69 | 97.37 | - |
| | 5.2.8.5 | (20,20,10,10)<br>(20,20,10,10) | (10%,10%,10%,10%)<br>(10%,10%,10%,10%) | 94.38 | 94.84 | 97.82 | 95.59 |
| | 5.2.8.6 | (20,20,10,10)<br>(20,20,10,10) | (50%,50%,50%,50%) | 94.35 | 94.64 | 94.61 | 94.61 |
| | 5.2.8.7 | (20,20,10,10)<br>(20,20,10,10) | (50%,50%,10%,10%)<br>(50%,50%,10%,10%) | 94.39 | 94.85 | 96.99 | 95.20 |
| | 5.2.8.8 | (20,20,10,10)<br>(20,20,10,10) | (10%,10%,50%,50%)<br>(10%,10%,50%,50%) | 93.78 | 94.65 | 97.30 | 95.03 |
| | 5.2.8.9 | (20,20,10,10)<br>(20,20,10,10) | (50%,50%,10%,10%)<br>(10%,10%,50%,50%) | 94.07 | 97.72 | 97.02 | - |
| | 5.2.8.10 | (20,20,10,10)<br>(20,20,10,10) | (50%,10%,50%,10%)<br>(50%,10%,50%,10%) | 94.11 | 95.01 | 97.42 | 95.41 |
| | 5.2.8.11 | (20,20,10,10)<br>(20,20,10,10) | (50%,10%,50%,10%)<br>(10%,50%,10%,50%) | 93.76 | 97.88 | 97.08 | - |
| | 5.2.8.12 | (20,20,10,10)<br>(10,10,20,20) | (10%,10%,10%,10%)<br>( ) | 93.84 | 94.87 | 98.09 | 95.67 |
| | 5.2.8.13 | (20,20,10,10)<br>(10,10,20,20) | (50%,50%,50%,50%)<br>(50%,50%,50%,50%) | 94.06 | 94.30 | 95.26 | 95.20 |
| | 5.2.8.14 | (20,20,10,10)<br>(10,10,20,20) | (50%,50%,10%,10%)<br>(50%,50%,10%,1 ) | 93.67 | 94.57 | 97.39 | 95.17 |
| | 5.2.8.15 | (20,20,10,10)<br>(10,10,20,20) | (50%,50%,10%,10%)<br>( ) | 93.73 | 98.83 | 98.58 | - |
| | 5.2.8.16 | (20,20,10,10)<br>(10,10,20,20) | (10%,10%,50%,50%)<br>(50%,50%,10 ) | 93.19 | 96.49 | 95.73 | - |
| | 5.2.8.17 | (20,20,10,10)<br>(10,10,20,20) | (50%,10%,50%,10%)<br>(50%,10%,50%,10%) | 93.61 | 94.59 | 97.41 | 95.29 |
| | 5.2.8.18 | (20,20,10,10)<br>(10,10,20,20) | (50%,10%,50%,10%)<br>(10%,50%,10%,50%) | 93.37 | 97.84 | 97.21 | - |

表 5.3　8 層の場合の各設定における信頼区間の被覆確率

| No. | 設定※ | | 被覆確率 | | | 1-α |
|---|---|---|---|---|---|---|
| | 例数 | 割合 | Greenland | Sato | Newcombe | CMH 検定 |
| 5.3.1 | (10,…,10, 10,…,10)<br>(10,…,10, 10,…,10) | (10%,…,10%, 10%,…,10%)<br>(10%,…,10%, 10%,…,10%) | 94.05 | 95.55 | 98.16 | 95.90 |
| 5.3.2 | (10,…,10, 10,…,10)<br>(10,…,10, 10,…,10) | (50%,…,50%, 50%,…,50%)<br>(50%,…,50%, 50%,…,50%) | 92.93 | 94.90 | 95.25 | 95.23 |
| 5.3.3 | (10,…,10, 10,…,10)<br>(10,…,10, 10,…,10) | (50%,…,50%, 10%,…,10%)<br>(50%,…,50%, 10%,…,10%) | 93.44 | 94.73 | 97.29 | 95.21 |
| 5.3.4 | (10,…,10, 10,…,10)<br>(10,…,10, 10,…,10) | (50%,…,50%, 10%,…,10%)<br>(10%,…,10%, 50%,…,50%) | 93.84 | 98.27 | 97.41 | - |
| 5.3.5 | (20,…,20, 20,…,20)<br>(20,…,20, 20,…,20) | (10%,…,10%, 10%,…,10%)<br>(10%,…,10%, 10%,…,10%) | 94.49 | 95.06 | 95.94 | 95.20 |
| 5.3.6 | (20,…,20, 20,…,20)<br>(20,…,20, 20,…,20) | (50%,…,50%, 50%,…,50%)<br>(50%,…,50%, 50%,…,50%) | 94.48 | 95.11 | 95.11 | 95.11 |
| 5.3.7 | (20,…,20, 20,…,20)<br>(20,…,20, 20,…,20) | (50%,…,50%, 10%,…,10%)<br>(50%,…,50%, 10%,…,10%) | 94.67 | 95.18 | 97.17 | 95.31 |
| 5.3.8 | (20,…,20, 20,…,20)<br>(20,…,20, 20,…,20) | (50%,…,50%, 10%,…,10%)<br>(10%,…,10%, 50%,…,50%) | 94.62 | 98.36 | 97.47 | - |
| 5.3.9 | (20,…,20, 10,…,10)<br>(20,…,20, 10,…,10) | (10%,…,10%, 10%,…,10%)<br>(10%,…,10%, 10%,…,10%) | 94.17 | 94.91 | 96.60 | 95.31 |
| 5.3.10 | (20,…,20, 10,…,10)<br>(20,…,20, 10,…,10) | (50%,…,50%, 50%,…,50%)<br>(50%,…,50%, 50%,…,50%) | 94.05 | 94.99 | 95.66 | 95.51 |
| 5.3.11 | (20,…,20, 10,…,10)<br>(20,…,20, 10,…,10) | (50%,…,50%, 10%,…,10%)<br>(50%,…,50%, 10%,…,10%) | 94.31 | 95.01 | 96.62 | 95.14 |
| 5.3.12 | (20,…,20, 10,…,10)<br>(20,…,20, 10,…,10) | (10%,…,10%, 50%,…,50%)<br>(10%,…,10%, 50%,…,50%) | 93.69 | 94.85 | 97.42 | 95.10 |
| 5.3.13 | (20,…,20, 10,…,10)<br>(20,…,20, 10,…,10) | (50%,…,50%, 10%,…,10%)<br>(10%,…,10%, 50%,…,50%) | 94.43 | 98.08 | 97.14 | - |
| 5.3.14 | (20,…,20, 10,…,10)<br>(10,…,10, 20,…,20) | (10%,…,10%, 10%,…,10%)<br>(10%,…,10%, 10%,…,10%) | 93.85 | 95.01 | 96.84 | 95.14 |
| 5.3.15 | (20,…,20, 10,…,10)<br>(10,…,10, 20,…,20) | (50%,…,50%, 50%,…,50%)<br>(50%,…,50%, 50%,…,50%) | 93.64 | 94.68 | 94.99 | 94.96 |
| 5.3.16 | (20,…,20, 10,…,10)<br>(10,…,10, 20,…,20) | (50%,…,50%, 10%,…,10%)<br>(50%,…,50%, 10%,…,10%) | 93.43 | 94.49 | 97.30 | 94.84 |
| 5.3.17 | (20,…,20, 10,…,10)<br>(10,…,10, 20,…,20) | (10%,…,10%, 50%,…,50%)<br>(10%,…,10%, 50%,…,50%) | 93.61 | 94.68 | 97.24 | 95.12 |
| 5.3.18 | (20,…,20, 10,…,10)<br>(10,…,10, 20,…,20) | (50%,…,50%, 10%,…,10%)<br>(10%,…,10%, 50%,…,50%) | 94.40 | 99.08 | 98.62 | - |
| 5.3.19 | (20,…,20, 10,…,10)<br>(20,…,20, 10,…,10) | (40%,…,40%, 10%,…,10%)<br>(10%,…,10%, 40%,…,40%) | 94.35 | 97.16 | 96.70 | - |
| 5.3.20 | (30,…,30, 10,…,10)<br>(30,…,30, 10,…,10) | (50%,…,50%, 10%,…,10%)<br>(10%,…,10%, 50%,…,50%) | 94.74 | 97.96 | 97.00 | - |
| 5.3.21 | (40,…,40, 10,…,10)<br>(40,…,40, 10,…,10) | (50%,…,50%, 10%,…,10%)<br>(10%,…,10%, 50%,…,50%) | 94.67 | 97.60 | 96.68 | - |
| 5.3.22 | (20,…,20, 10,…,10)<br>(20,…,20, 10,…,10) | (50%,…,50%, 5%,…,5%)<br>(5%,…,5%, 50%,…,50%) | 94.13 | 98.84 | 97.93 | - |

※ 各行の上段：（$n_{11}, \ldots, n_{14}, n_{15}, \ldots, n_{18}$）（$p_{11}, \ldots, p_{14}, p_{15}, \ldots, p_{18}$）、各行の下段：（$n_{21}, \ldots, n_{24}, n_{25}, \ldots, n_{28}$）（$p_{21}, \ldots, p_{24}, p_{25}, \ldots, p_{28}$）

α：Type I error rate

## 6. まとめ

本稿では、調整済みリスク差に対する信頼区間の算出方法を3種類紹介した上で、それぞれの信頼区間の性能評価を行った。シミュレーション結果より、Greenland信頼区間、Sato信頼区間、Newcombe信頼区間のうち、層の数や例数によらず被覆確率が名義水準を維持したのはNewcombe信頼区間であった。各群の各層の例数が小さい場合、Greenland信頼区間は被覆確率が名義水準を下回る可能性が高いため、注意が必要である。Sato信頼区間やNewcombe信頼区間は、Greenland信頼区間ほど例数による影響を受けないが、層内や層間の割合の設定によっては、被覆確率が名義水準を上回ることがあるため、層内及び層間の割合の差を考慮し、適用する信頼区間を選択することが望ましい。

## 参考文献


- Greenland S. and Robins J. M. (1985), “Estimation of a common effect parameter from sparse follow-up data,” Biometrics, 41(1): 55-68.
- Sato T. (1989), “On variance estimator for the Mantel-Haenszel risk difference. (Letter),” Biometrics; 45: 1323-1324.
- Yan X. and Su X. G. (2010), "Stratified Wilson and Newcombe Confidence Intervals for Multiple Binomial Proportions," Statistics in Biopharmaceutical Research; Vol. 2, No. 3: 329-335.
- Rothman K. J. (2012a), “Epidemiology: An Introduction, 2nd Ed.,” Oxford University Press.
- Rothman K. J., Lash T. L., and Greenland S. (2012b), “Modern Epidemiology, 3rd Ed.,” Lippincott Williams & Wilki.
- SAS/STAT(R) 14.1 User's Guide.
- 飯塚 政人, 浜田 知久馬 (2013), "2群の割合の差における信頼区間の構成法の比較," SASユーザー総会.
- 飯塚 政人, 魚住 龍史, 浜田 知久馬 (2014), "FREQプロシジャによる割合の差の信頼区間," SASユーザー総会.


## 参考：5章のシミュレーションを実行するSASマクロ

```
* _N1 ：群1の各層の例数をスペース区切りで入力 ;
* _N2 ：群2の各層の例数をスペース区切りで入力 ;
* _P1 ：群1の各層の割合をスペース区切りで入力 ;
* _P2 ：群2の各層の割合をスペース区切りで入力 ;
* _ITE：シミュレーション回数                   ;
%macro MH(_N1=" 10, 20, 30", _N2=" 10, 20, 30", _P1="0.1, 0.2, 0.3", _P2="0.1, 0.2, 0.3", _ITE=10000) ;
  options nosource nonotes ;
  ods listing close ;

  data _NULL_ ;
    STR=0 ; FLAG=1 ;
    do while (FLAG=1) ;
      if (scan(&_N1., STR+1, ",")="") then FLAG=0 ;
      else STR=STR+1 ;
    end ;
    call symputx("_STR",STR) ;
  run ;

  data SAMPLE ;
    W=0 ; TRUE_D=0 ;
    array N1 {&_STR.} ; array N2 {&_STR.} ;
```

```
  array P1 {&_STR.} ; array P2 {&_STR.} ;
  array X1 {&_STR.} ; array X2 {&_STR.} ;
  do I=1 to &_STR. ;
    N1[I]=input(scan(&_N1., I, ","), best12.) ;
    N2[I]=input(scan(&_N2., I, ","), best12.) ;
    P1[I]=input(scan(&_P1., I, ","), best12.) ;
    P2[I]=input(scan(&_P2., I, ","), best12.) ;
    W    =W+(N1[I]*N2[I])/(N1[I]+N2[I]) ;
  end ;

  do I=1 to &_STR. ;
    TRUE_D=TRUE_D+(P1[I]-P2[I])*(N1[I]*N2[I])/(N1[I]+N2[I])/W ;
  end ;
  call symputx("_TRUE_D",TRUE_D) ;

  J=1 ;
  do while (J<=&_ITE.) ;
    do I=1 to &_STR. ;
      X1[I]=ranbin(777, N1[I], P1[I]) ;
      X2[I]=ranbin(777, N2[I], P2[I]) ;
    end ;
    if (sum(of X1[*]) ne 0 and sum(of X2[*]) ne 0 and
        sum(of X1[*]) ne sum(of N1[*])            and sum(of X2[*]) ne sum(of N2[*])) then do ;
      do I=1 to &_STR. ;
        NO=1 ; TREAT=1 ; X=X1[I] ; N=N1[I]       ; output ;
        NO=1 ; TREAT=2 ; X=X2[I] ; N=N2[I]       ; output ;
        NO=2 ; TREAT=1 ; X=1     ; N=X1[I]       ; output ;
        NO=2 ; TREAT=1 ; X=2     ; N=N1[I]-X1[I] ; output ;
        NO=2 ; TREAT=2 ; X=1     ; N=X2[I]       ; output ;
        NO=2 ; TREAT=2 ; X=2     ; N=N2[I]-X2[I] ; output ;
      end ;
      J=J+1 ;
    end ;
  end ;
  keep J NO TREAT I N X ;
run ;

ods output Effect=CI ;
proc stdrate data=SAMPLE
  method=mh
  stat  =risk
  effect=diff
  plots=none ;
  where NO=1 ;
  by J ;
  population group=TREAT event=X total=N ;
  strata I / order=internal stats(cl=normal) effect ;
run;
ods output close ;

data CP_STDRATE ;
  set CI ;
  if (LOWERCL <= &_TRUE_D. <= UPPERCL) then CP_WALD+(100/&_ITE.) ;
  if (_N_=&_ITE.) ;
  keep CP_WALD ;
run ;

ods output CommonPdiff=CI CMH=CMH ;
proc freq data=SAMPLE ;
  where NO=2 ;
```

```
    by J ;
    table  I*TREAT*X / riskdiff(common column=1 cl=newcombe) cmh nocol nopercent plots=none ;
    weight N ;
  run ;
  ods output close ;

  data CP_SATO ;
    set CI ;
    where (Method="Mantel-Haenszel") ;
    if (LOWERCL <= &_TRUE_D. <= UPPERCL) then CP_SATO+(100/&_ITE.) ;
    if (_N_=&_ITE.) ;
    keep CP_SATO ;
  run ;

  proc sql noprint ;
    select count(*) into:_ITE_NEW from CI where Method="Newcombe" ;
  quit ;

  data CP_NEW ;
    set CI ;
    where (Method="Newcombe") ;
    if (LOWERCL <= &_TRUE_D. <= UPPERCL) then CP_NEW+(100/&_ITE_NEW.) ;
    if (_N_=&_ITE_NEW.) ;
    keep CP_NEW ;
  run ;

  data CP_CMH ;
    set CMH ;
    where (Statistic=1) ;
    if (Prob>=0.05) then CP_CMH+(100/&_ITE.) ;
    if (_N_=&_ITE.) ;
    keep CP_CMH ;
  run ;

  options source notes ;
  ods listing ;
  data CP ;
    merge CP_STDRATE CP_SATO CP_NEW CP_CMH ;
  run ;

  title1 "N1 =" &_N1. "; N2 =" &_N2. ;
  title2 "P1 =" &_P1. "; P2 =" &_P2. ;
  title3 "True Difference=&_TRUE_D." ;
  proc print ; run ;
   title1 "" ; title2 "" ; title3 "" ;
%mend ;


* 2層の指定例 ;
%MH(_N1=" 20,  20", _N2=" 20,  20",
    _P1="0.1, 0.1", _P2="0.1, 0.1") ;

* 4層の指定例 ;
%MH(_N1=" 20, 20,  20,  20",  _N2=" 20,  20,  20,  20",
    _P1="0.1, 0.1, 0.1, 0.1", _P2="0.1, 0.1, 0.1, 0.1") ;

* 8層の指定例 ;
%MH(_N1=" 10,  10,  10,  10,  10,  10,  10,  10", _N2=" 10,  10,  10,  10,  10,  10,  10,  10",
    _P1="0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1", _P2="0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1") ;
```

以上

# ベイズ流の単群第II相試験デザインのSASマクロの作成

○橋本 大哉、嘉田 晃子

（国立病院機構名古屋医療センター 臨床研究センター）

Thall and Simon (1994, Biometrics)によって提案された事後確率によるベイズ流の単群第II相試験デザインは、主要評価項目である2値データを連続的にモニタリングし、試験治療と標準治療の反応率の事前分布および試験データから得られる尤度から、試験治療が標準治療に優る事後確率を求める。事後確率があらかじめ定めた上側の確率を超えれば試験治療が有効と宣言して試験を中止し（有効中止）、あらかじめ定めた下側の確率を下回れば試験治療が無効と宣言して試験を中止し（無効中止）、いずれにも当てはまらなければ試験を継続する。この手順は連続モニタリングする必要があるが、試験治療が標準治療と比べて極めて高い反応率を示す場合、あるいは低い反応率を示す場合に少数の評価で早期に試験を中止できるメリットがあり、希少疾患を対象とする単群の臨床試験を立案する際の選択肢の1つとなっている。一方、反応率の事前分布や中止確率及びモニタリングを開始する例数と終了する例数に関して、複数のシナリオを検討する必要があるため、マクロ化することで効率的な研究計画の立案ができる。

SAS でこのデザインに基づくプログラムの作成を検討したとき、事後確率（ベータ分布の確率密度関数と分布関数の積を積分した値で算出する）の計算が、試験治療の反応率の確率密度関数の形状が急峻な場合に、収束せずに誤った値を返すことが起こった。発表者らは、PROC IML の Submit ステートメントを用いて R 上で積分計算を実施することで、直接 R を起動することなくこの問題を回避し、SAS マクロの作成に成功したので報告する。

参考文献

[1] Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 1994; **50**: 337–49.

[2] 手良向聡，大門貴志訳．臨床試験デザイン―ベイズ流・頻度流の適応的方法．メディカル・パブリケーションズ．2014

# 公開医療ビッグデータを活用した
# 医薬品マーケティングの可能性

武藤　猛
MarkeTech Consulting 代表

Possibility of Pharmaceutical Marketing Using Open Medical Big Data

Takeshi Muto
President, MarkeTech Consulting

## 要旨

近年、医療の分野でもビッグデータとその活用の議論が盛んになっている。医療機関で発生した多種多様な医療データが関係機関に集約され匿名化の上で、医療費の適正化や医療の質向上など様々な目的で活用され始めている。一方、製薬企業という民間の立場から医療ビッグデータを眺めると、データのオープン性とデータ内容・粒度（医薬品別、医療機関別など）から判断して、医薬品マーケティングに活用できるものは、従来から公開されている DPC データや、昨年初めて公開された NDB オープンデータにほぼ限定される。本論文では、後者の NDB オープンデータをどのように医薬品マーケティングに活用できるかを検討した。

NDB オープンデータを活用した第一の事例は、医薬品市場構造の分析である。データが薬効領域別（大分類：8、中分類：33、小分類：137）であることに着目し、特に薬効領域（小分類）単位の市場構造を分析した。第二の事例は、都道府県単位の競合分析（エリアマーケティング）である。都道府県単位の競合状況を上位製品のシェアやシェア比率を算出し可視化することにより、今後打つべきエリア戦略が明確になる。

結論として、NDB オープンデータは、目的は限定されるが、医薬品マーケティングに十分活用可能である。ただし、NDB オープンデータは今後公開されるとしても年単位であること、およびデータ公開時期とデータ収集時期との時間差がかなり大きいことから、支店や営業所の業績評価には適さず、本社の中長期戦略立案のためのマーケティング向きと考えられる。

キーワード：医薬品マーケティング、医療ビッグデータ、NDB オープンデータ、市場構造分析、エリアマーケティング、公開データのメリットと限界

## 1. はじめに

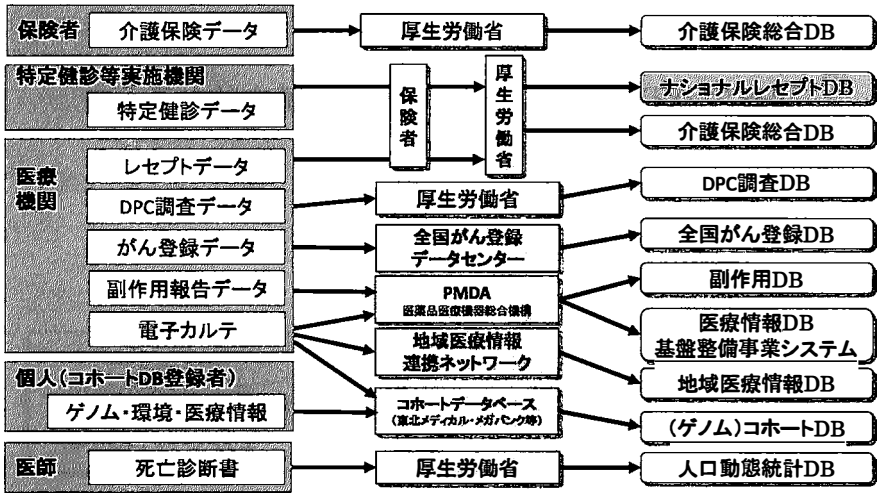
### 1.1 医療ビッグデータについて

最近、ビッグデータの話題が盛んである。医療分野も例外ではない。「ビッグデータ」の定義は必ずしも明

確ではないが、量（Volume）・迅速性（Velocity）・多様性（Variety）・正確性（Veracity）という「4V」を備えたものを「ビッグデータ」と称するのが一般的である[1]。ただし、医療分野ではこの定義がそのまま当てはまるとは限らない。例えば、電子カルテ情報やレセプト情報は、そのままでは一般公開されることは決してなく、プライバシーに対する厳格な配慮をした匿名化の上で、集計データのみが公開される。一方、最近リアルワールドデータ（RWD）という用語もよく用いられる。RWD は、RCT など臨床試験データと対比される、実診療行為に基づくデータベースであり、上記の医療ビッグデータと同義語と考えられる。

一般のビッグデータでは、インターネット上のデータや画像データ、センサーの計数値などを無差別に大量収集したものも含まれる。これに対して医療ビッグデータは、前述のように関係機関が一定ルールの下で開示する大規模医療データベースである。図表 1 は、大規模医療データベースの「発生元」、「集約機関」、および利用に供される「データベース」の流れを図示したものである[2]。

**図表1. 大規模医療データベース(DB)**




[出典]国会図書館:ライフサイエンスのフロンティア(第3章:医療データ等の利活用)、2015 などを参考にして作成

大規模医療データベースの活用は、医療保険業務、医療・介護の質向上、医療機関の経営改善、副作用の早期発見とその対策、地域医療改善など医療や介護本来の目的がその主なものである。製薬企業など民間企業の立場から医療ビッグデータを眺めると、データのオープン性とデータ内容・粒度（医薬品別、医療機関別など）から判断して、医薬品マーケティングに活用できるものは、従来から公開されている DPC データや、昨年初めて公開された NDB オープンデータに限定される。本論文では、後者の NDB オープンデータをどのように医薬品マーケティングに活用できるかを検討した。

### 1.2　本論文の目的

昨年公開された NDB オープンデータにおいて、医薬品別処方件数が公開されたことは朗報であり、このことにより医薬品マーケティングへの応用の道が開かれた。本論文の目的は、 NDB オープンデータを活用した医薬品マーケティングの可能性について、医薬品市場構造分析と特定薬効領域内競合分析という具体的な分析例を通じて検討することである。

## 2. 公開医療ビッグデータ活用の基本的な考え方

## 2.1　第 1 回 NDB オープンデータの公開内容

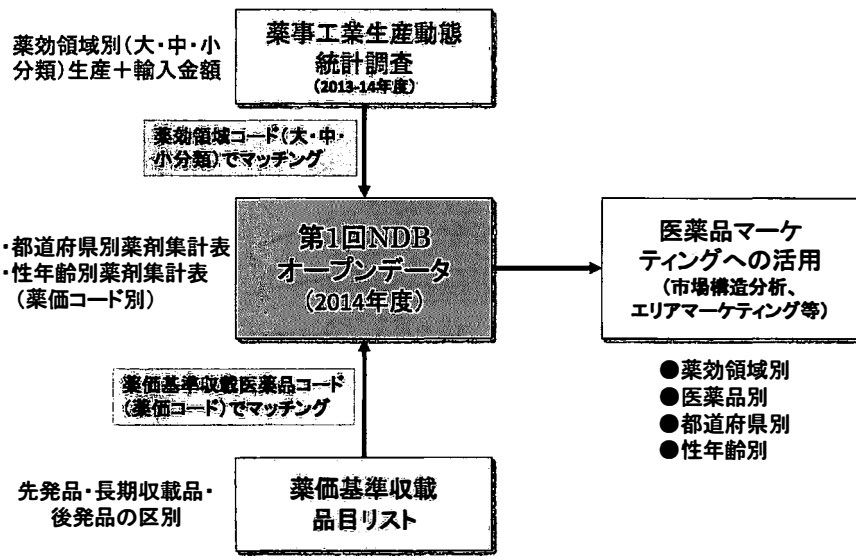
2016 年秋に公開された、第 1 回 NDB オープンデータの内容を紹介する[3]。NDB データベースとは、図表 1 に示されている「ナショナルレセプト DB」のことであり、レセプト（H21.4～27.1 診療分）約 92 億 5,000 万件、特定健診・特定保健指導（H20～25 年度分）約 1 億 4,200 万件から構成される。NDB オープンデータ作成の目的は、このレセプト情報・特定健診等情報データベース（NDB）の有効活用である。

NDB オープンデータの対象・公表形式（医薬品の場合）は、対象期間は 2014 年度（H26.4～H27.3）の 1 年間であり、内服・外用・注射それぞれにつき、外来院内・外来院外・入院別に集計され、薬価収載の基準単位別に薬効別上位処方数 30 位までの処方数が公開される。公開形式は、都道府県および性別・年齢別集計表である。患者のプライバシーへ配慮して、最小集計単位を 1000 としているので注意が必要である。

## 2.2　NDB オープンデータ活用の基本的考え方

第 1 回 NDB オープンデータは、多数の EXCEL シートとして公開されている。データ分析上重要なことは、データ薬価収載の医薬品名毎の処方件数が示され、また薬効領域（大分類：8、中分類：33、小分類：137）が付与されていることである。そこで、図表 2 に示す考え方で NDB オープンデータを活用することとする。

**図表2. NDBオープンデータ活用の基本的な考え方**




図表 2 に示すように、補足的に 2 種類のデータを NDB オープンデータにマッチングさせて分析用データベースを作成した。まず第一に、2014 年度薬事工業生産動態統計調査データと薬効領域コード（大・中・小分類）でマッチングさせ、薬効領域別の金額推移（成長率）を追加した。第二に、薬価基準主催品目リストと薬価基準収載医薬品コード（薬価コード）でマッチングさせ、先発品・長期収載品・後発品のフラグを取得した。ただし、この薬価基準主催品目リストには、先発品・長期収載品・後発品の区別が示されていない品目も多数あることに注意が必要である。上記 2 種類の補足データをマッチングさせた NDB オープンデータは、薬効領域別、医薬品別、都道府県別、性年齢別など多彩な分析が可能になる。なお、データ加工には SAS、データ分析には JMP を主に用いた。

## 3. 公開医療ビッグデータの活用①：医薬品市場構造の分析

### 3.1 NDB オープンデータの概要

完成した分析用データベースを用いて、NDB オープンデータの基本的内容を確認した。NDB オープンデータには、医薬品別の処方件数と薬価データが含まれる。これらを用いれば、医薬品別の金額（＝処方件数×薬価）が求められる。今回の NDB オープンデータから求められた 2014 年度の薬剤費合計は、7 兆 2529 億円となった。一方、2014 年度薬事工業生産動態統計調査によれば、「生産＋輸入」の医薬品金額は 9 兆 384 億円となり、NDB オープンデータの 2014 年度の薬剤費合計はこの金額の 80.2%に過ぎない。NDB オープンデータにおいては、前述のように最小集計単位を設定しているので、このようにデータ値が実際の金額より低くなったと考えられる。

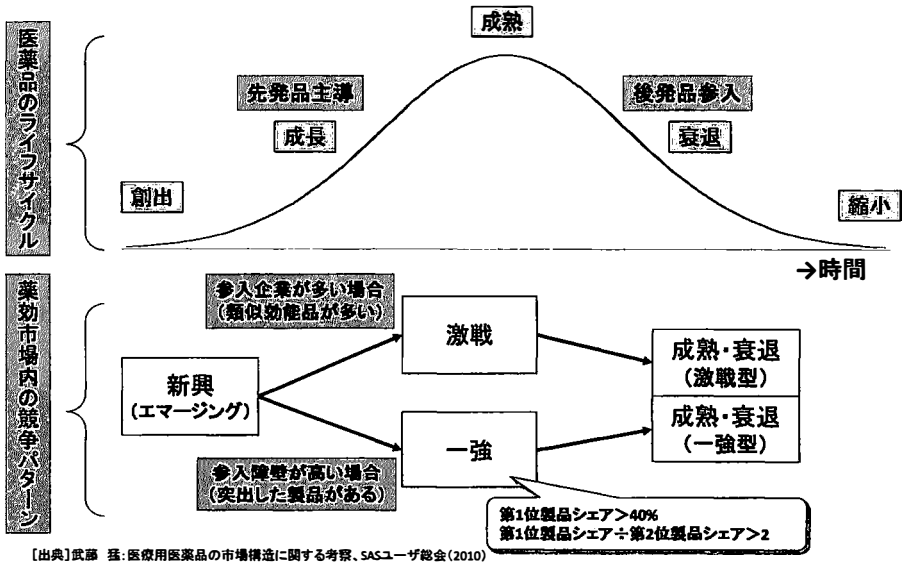
次に、先発品・長期収載品・後発品の数量比率と金額比率を求めたところ、各々45.4%・30.3%・24.3%、および 52.1%・40.2%・7.8%となった。ただしこの比率は、先発品・長期収載品・後発品の区別が明確なものだけを対象にしており、区別が不明な医薬品は含まない。

### 3.2 医薬品市場構造分析のアプローチ

医療用医薬品市場は、年間売上高が 10 兆円近くの巨大市場であるが、「競合」という視点からは、「薬効市場」に細分化された多数の競合市場の集合体である。NDB オープンデータには前述の通り、薬効領域（大・中・小分類）のフラグが付与されているが、「競合」（つまり医師により代替処方の可能性のある医薬品グループ）が小分類レベルで行われているとは限らず、場合によっては適応症や作用機序レベルに更に細分化する必要がある。ただし、このような細分化は現在のところ手作業によるしかない。

医薬品市場構造について、以前著者が発表した論文[4]を参照すると、図表 3 の仮説にまとめられる。

**図表3. 薬効市場のライフサイクルと市場盛衰のメカニズム（仮説）**




［出典］武藤 猛：医療用医薬品の市場構造に関する考察、SASユーザ総会（2010）

図表 3 に示すように、医薬品を含む多くの商品（消費財・生産財）について観察されるライフサイクルは、単独の医薬品だけでなく、薬効領域レベルでも観察される。薬効市場については、新興（エマージング）薬効領域は、領域の特性によって「激戦型」（参入企業・類似効能品が多い場合）または「一強型」（参入障壁

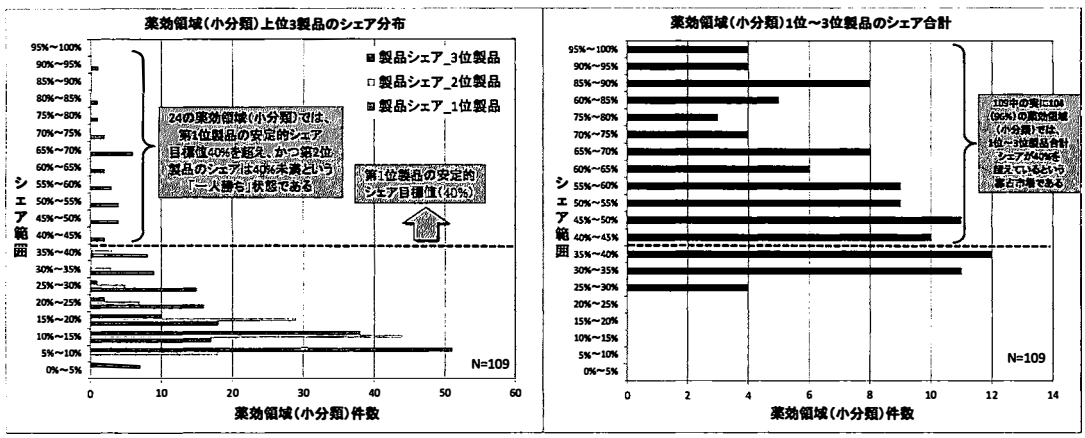
が高かったり効能上突出した製品がある場合）として成長し、市場の主力製品の特許切れに伴う後発品参入時期以降は成熟期および衰退期に入る。重要なことは、成熟期・衰退期にも「激戦型」または「一強型」の特性が持ち越されることである（少なくとも後発品のシェアが圧倒的になるまでは）。

### 3.3　医薬品市場構造分析の結果

医療用市場構図分析のいくつかの結果を紹介する。まず薬効領域の大分類・中分類・小分類の 3 レベルで、「市場規模対成長率」を分析した。いずれのレベルでも市場規模が大きくなると（有意に）成長率が下がることが分かった。次に薬効領域（小分類）別に、先発品・後発品比率と成長率との関係を分析した。その結果、薬効市場の拡大期は先発品が 60%以上で成長が見られ、一方、後発品が 10%程度以上を占めるようになると市場は縮小期に入るという興味深い事実が観察された。

次に上位製品のシェア分布と医薬品市場の特質との関係を検討する（図表 4）。

## 図表4. 上位製品のシェア分布と医薬品市場の特質




図表 4 の左側は、各薬効領域の上位 3 製品のシェア分布を示している。分析対象の 109 の薬効領域（小分類）中の 24 薬効領域では、第 1 位製品の安定的シェア目標値 40%（ランチェスター理論[5]による）を超え、かつ第 2 位製品のシェアは 40%未満という「一人勝ち」状態である。図表 4 の右側は、各薬効領域の 1 位～3 位製品のシェア合計の分布を示す。109 の薬効領域（小分類）中の実に 104（96%）の領域が、1 位～3 位製品合計シェアが 40%を超えているという寡占市場であることが分かる。

最後に、市場構造をより鮮明にするために、薬効領域（小分類）を対象に、因子分析とクラスター分析を行った。分析に用いた変数は、市場規模(億円)_2014、領域の薬剤数、成長率、金額比率 1_先発品、金額比率 2_長期収載品、金額比率 3_後発品、製品シェア_1 位製品、製品シェア_2 位製品、製品シェア_3 位製品、製品シェア_1 位～2 位製品、製品シェア_1 位～3 位製品、製品シェア比率_1 位÷2 位製品、製品シェア比率_1 位÷3 位製品、である。これらの変数データが揃っている 74 領域が分析対象である。因子分析およびクラスターの要約を図表 5 および図表 6 に示す。クラスターを可視化したものを図表 7 に示す。

図表 5 に示すように、因子分析の結果抽出された 2 つの因子を、「成熟因子」および「競合因子」と命名した。また、図表 6 に示すクラスター分析の結果、前述の「一強型」「激戦型」「成熟型」クラスターが鮮明となった。「一強型」「激戦型」「成熟型」クラスターに分類される薬効領域（小分類）の数（比率）は 4 (5.4%) : 23 (31%) : 47 (64%)と、「成熟型」が圧倒的に多く、「一強型」が非常に少ない。「一強型」「激戦型」だけに限れば両者の数（比率）は 4（15%）: 23（85%）となり、「激戦型」が圧倒的に多い。3 つのクラスターを図示

した図表７によれば、「成熟因子」×「競合因子」という２次元図上で、各クラスターはかなり鮮明に分離されていることが分かる。今回の分析だけでは十分でないが、多年度分のデータが集積されていけば、各クラスター内の薬効領域の動き（Ｙ字を横向きにした「一強型」「激戦型」各分枝から根元の「成熟型」に向かうと考えられる）も明らかになると予想される。

## 図表5. 市場構造の分析：因子分析

■医療用医薬品市場の因子構造

| 項目 | 因子1 | 因子2 |
|---|---|---|
| 固有値 | 4.768 | 2.735 |
| 寄与率 | 0.367 | 0.210 |
| 累積寄与率 | 0.367 | 0.577 |
| 因子名称 | 成熟因子 | 競合因子 |
| 市場規模(億円)_2014 | −0.498 | −0.319 |
| 領域の薬剤数 | −0.595 | −0.406 |
| 成長率 | −0.329 | 0.209 |
| 金額比率1_先発品 | −0.331 | 0.449 |
| 金額比率2_長期収載品 | 0.289 | −0.423 |
| 金額比率3_後発品 | 0.107 | −0.099 |
| 製品シェア_1位製品 | 0.954 | −0.103 |
| 製品シェア_2位製品 | 0.341 | 0.795 |
| 製品シェア_3位製品 | 0.090 | 0.814 |
| 製品シェア_1位～2位製品 | 0.950 | 0.191 |
| 製品シェア_1位～3位製品 | 0.910 | 0.341 |
| 製品シェア比率_1位÷2位製品 | 0.712 | −0.557 |
| 製品シェア比率_1位÷3位製品 | 0.762 | −0.515 |

因子パターン

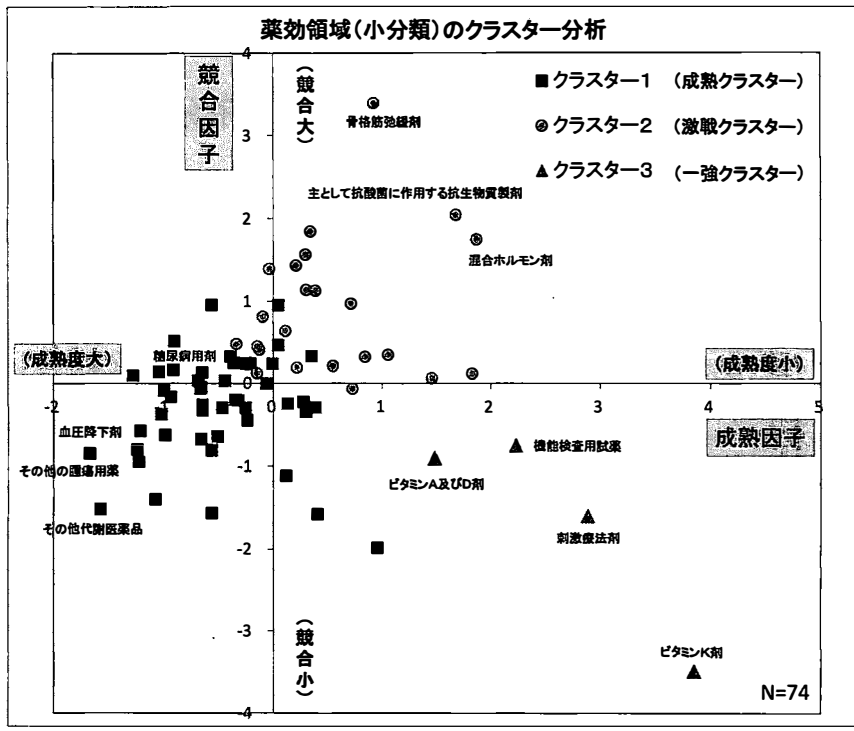
## 図表6. 市場構造の分析：クラスター分析

■医療用医薬品市場のクラスター構造（元の変数使用）

| クラスター番号 | クラスター1 | クラスター2 | クラスター3 |
|---|---|---|---|
| クラスター名称 | 成熟クラスター | 激戦クラスター | 一強クラスター |
| 薬効領域数 | 47 | 23 | 4 |
| 市場規模(億円)_2014 | 1297 | 501 | 183 |
| 領域の薬剤数 | 60 | 42 | 24 |
| 成長率 | −0.031 | −0.055 | −0.166 |
| 金額比率1_先発品 | 0.393 | 0.457 | 0.344 |
| 金額比率2_長期収載品 | 0.433 | 0.436 | 0.523 |
| 金額比率3_後発品 | 0.174 | 0.106 | 0.133 |
| 製品シェア_1位製品 | 0.194 | 0.325 | 0.667 |
| 製品シェア_2位製品 | 0.122 | 0.206 | 0.086 |
| 製品シェア_3位製品 | 0.096 | 0.133 | 0.060 |
| 製品シェア_1位～2位製品 | 0.317 | 0.531 | 0.753 |
| 製品シェア_1位～3位製品 | 0.412 | 0.664 | 0.813 |
| 製品シェア比率_1位÷2位製品 | 1.658 | 1.650 | 8.165 |
| 製品シェア比率_1位÷3位製品 | 2.132 | 2.659 | 12.441 |
| 具体的な薬効領域（小分類）の例 | 他に分類されない代謝性医薬品<br>その他の腫瘍用薬<br>血圧降下剤<br>糖尿病用剤<br>精神神経用剤<br>高脂血症用剤<br>眼科用剤<br>その他の血液・体液用薬<br>その他のアレルギー用薬<br>その他の中枢神経系用薬<br>その他のホルモン剤（抗ホルモン剤を含む）<br>抗ウイルス剤<br>血管拡張剤 | 消化性潰瘍用剤<br>鎮痛，鎮痒，収斂，消炎剤<br>解熱鎮痛消炎剤<br>合成抗菌剤<br>抗てんかん剤<br>脳下垂体ホルモン剤<br>気管支拡張剤<br>耳鼻科用剤<br>解毒剤<br>痛風治療剤<br>主としてカビに作用する抗生物質製剤<br>骨格筋弛緩剤 | ビタミンA及びD剤<br>機能検査用試薬<br>刺激療法剤<br>ビタミンK剤 |

クラスター分析に用いた元の変数の平均値

## 図表7. 市場クラスターの可視化




最後に、図表 6 から導かれる実用的な知見として、薬効領域の「一強型」または「激戦型」の簡便な判別法を紹介する。図表 3 にも記入済みではあるが、①「第 1 位製品のシェア」が 40%以上で、かつ②「第 1 位製品のシェア」÷「第 2 位製品のシェア」が 2 以上の場合、「一強型」と判断できる。このような薬効領域（小分類）では、第 1 位製品が独走状態にあり、第 2 位製品以下はキャッチアップが困難と考えられる。薬効領域（小分類）の圧倒的部分が「激戦型」である。製品力で圧勝している「一強型」に対して、「激戦型」においては、マーケティングおよび営業戦略によっては第 2 位以下の製品にも第 1 位を奪取する機会がある。

### 3.4 医薬品マーケティングへの示唆

市場規模と成長率との関係について、①薬効領域の大・中・小分類の各段階で、規模↑と成長率↓の関係が見出された。また、②単独製品および薬効市場は、ベル曲線型ライフサイクルを辿ると推測される。

薬効領域（小分類）内の競合状況について、①薬効市場の拡大期は先発品が 60%以上で成長が見られる。一方、②後発品が 10%程度以上を占めるようになると、市場は縮小期に入る。③薬効市場の多くは寡占状態であり、上位 3 製品の合計シェアが 40%を超える。④拡大期の市場は「激戦型」または「一強型」のいずれかに分かれる。⑤「激戦型」や「一強型」の市場構造は、ライフサイクルを通じて固定される（少なくとも後発品が圧倒するまで）。薬効市場の大部分は「激戦型」であり、医薬品マーケティングの役割が重要である。

## 4. 公開医療ビッグデータの活用②：特定薬効領域の競合分析

### 4.1 エリアマーケティングの基本的な考え方

医薬品のマーケティングにおいては、特定薬効領域における競合分析が主要な役割を果たす。近年、地域

医療のネットワーク化が進展するにともない、エリアマーケティングの役割が従来以上に重要になっている。今回公開された NDB オープンデータには、都道府県別集計データが含まれているので、エリアマーケティングの目的には好都合である。

前述のように、薬効市場は「激戦型」または「一強型」に分類される寡占市場である。薬効市場における製薬企業の立場は強者または弱者のいずれかであるが、自社の保有するすべての製品がトップシェアであることは考えられず、基本は「弱者」（つまり 2 位以下）の立場で競合戦略を立案することになる。医薬品市場における弱者の戦略とは、「ターゲティングを精緻化し、トップシェア獲得可能な細分化市場に注力する」ことと要約できる。著者の考案した具体的なエリア戦略立案手順については、以前に発表している[6]。

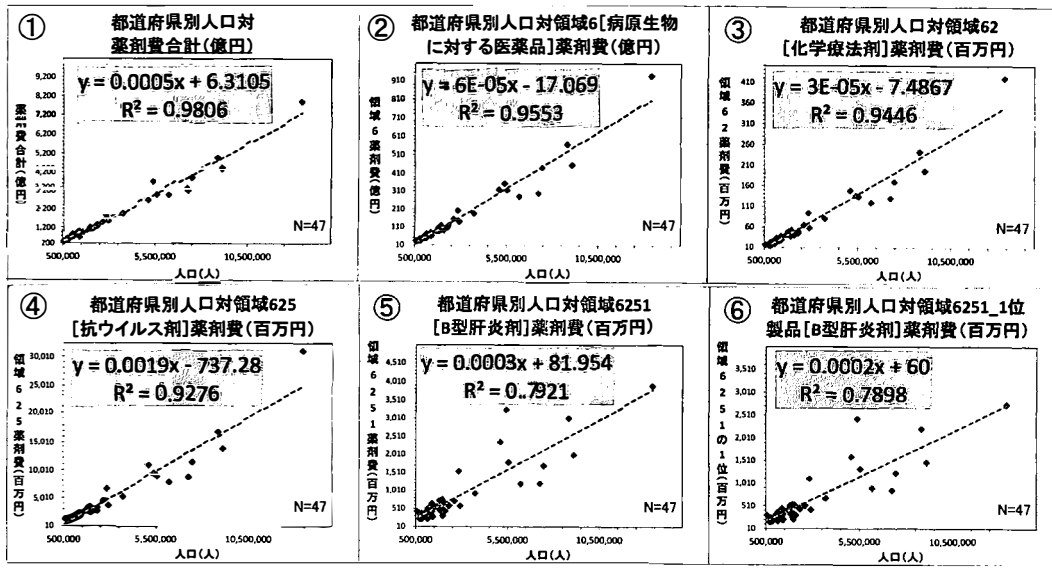
## 4.2 競合分析の対象（抗ウイルス剤）について

以下の事例では、薬効領域（小分類）の「625：抗ウイルス剤」を対象とする。細分類レベル（適応症）として、B 型肝炎、C 型肝炎、HIV、RS ウイルス、インフルエンザ、サイトメガロウイルス、ヘルペスが含まれる。「抗ウイルス剤」は「激戦型」であるが、細分類レベルでは「激戦型」と「一強型」の両方が含まれることにが興味深い。

## 4.3 抗ウイルス剤の競合分析結果

エリアマーケティングの基本前提は、「医薬品市場は人口（患者数）に比例する」ことである。抗ウイルス剤について、「都道府県別人口（患者数）対薬剤費」の相関関係を 6 段階で分析したのが図表 8 である。

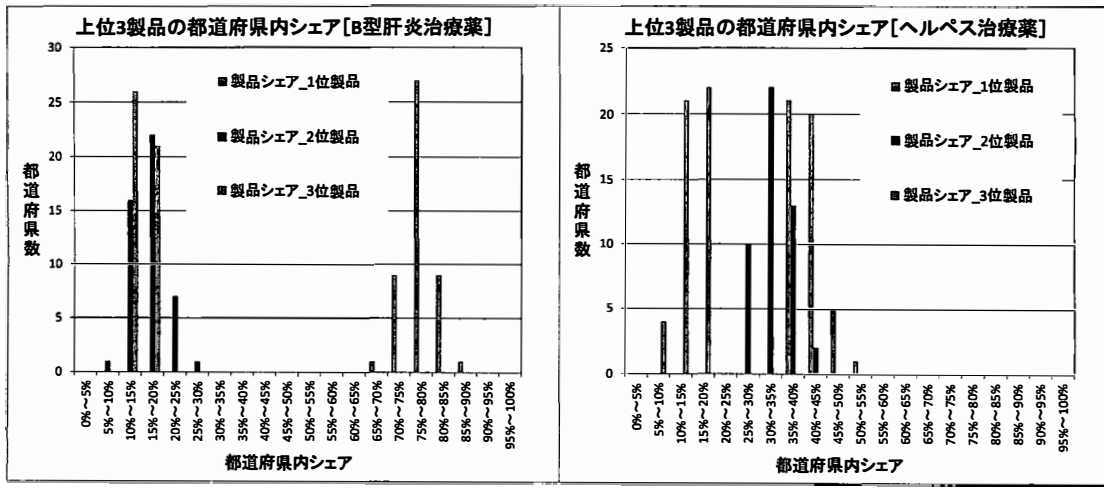
### 図表8. 医薬品市場は人口（患者数）に比例する



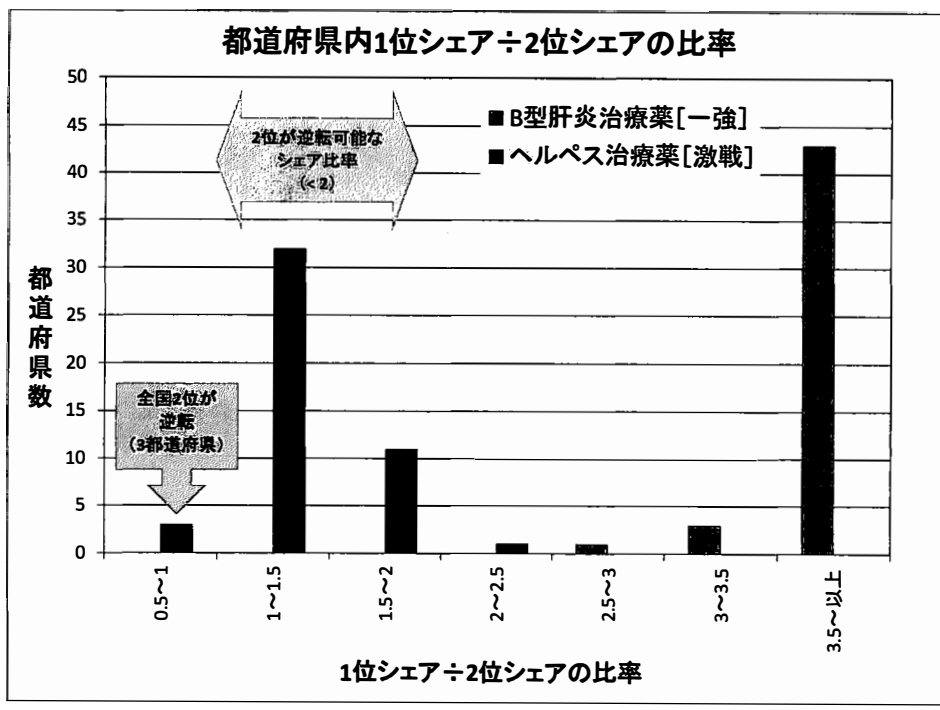

図表 8 によれば、④抗ウイルス剤のレベルまではばらつきが小さいのに、⑤および⑥の B 型肝炎治療薬のレベルになるとばらつきが急に大きくなる。これは、この段階ではエリア戦略の良否が重要であることを示唆している。

抗ウイルス剤の細分類レベルで、B 型肝炎治療薬は「一強型」であるのに対して、ヘルペス治療薬は「激戦型」である。都道府県レベルでの両者の競合状況を可視化するために、都道府県内シェアを比較したのが図表 9 である。また、「都道府県内 1 位シェア÷2 位シェア」の比率を図示したのが図表 10 である。

## 図表9. 都道府県内シェアの比較:「一強型」対「激戦型」




## 図表10.「一強型」対「激戦型」シェア比率の比較:比率の分布




図表 9 および図表 10 から、「一強型」または「激戦型」の競合状況が鮮明に読み取れる。さらに、製薬企業が各都道府県単位でどのようなエリア戦略を取るべきかについては、エリアバリューマトリックス (AVM) を作成し、地図上で表示することで明らかになる[6]。

### 4.4 医薬品マーケティングへの示唆

医薬品マーケティングへの適用可能性について、①エリア間の患者移動を考慮すると、「都道府県」とい

う集計単位は（患者フローがほぼ閉じているので）、エリアマーケティングに適している。ただし、②最小集計単位が 1000 であることから、処方数の少ない医薬品の分析は精度が悪くなる。③男女別・年齢別データは、競合医薬品間の患者プロファイルの比較に応用可能である（今回は分析しなかった）。

エリア単位の競合状況について、地図を用いて可視化することが有用である。都道府県単位で、「激戦型」や「一強型」の市場構造を可視化することは、エリア戦略立案や評価に役立つ[6]。

## 5. まとめ：公開医療ビッグデータ活用のメリットと限界

### 5.1　公開医療ビッグデータ活用のメリットと限界

公開医療ビッグデータ活用のメリットは次の通りである：(1)無償である、(2)このため自社の製品がなくても将来の進出予定領域など関心のある薬効領域について分析ができる、(3)医薬品データ分析の教育などに活用可能である。

一方、限界は次の通りである：(1)公開時期が年単位であること、(2)データ公開時期とデータ収集時期との時間差がかなり大きいこと、(3)薬剤データは一定条件を満たしたものに限られ悉皆データではないこと（特に後発品の品目はごく限られている）、(4)このため支店や営業所の業績評価（四半期・半期・年間）には適さず、(5)本社の中長期戦略立案のためのマーケティング向きである。

### 5.2　公開医療ビッグデータへの今後の期待

NDB オープンデータに関しては、今後毎年 1 回の定期的公開が期待される。時系列データが蓄積されることにより、医薬品マーケティングへの応用も格段に広がると考えられる。

## 参考文献


(1) 中山健夫：医療ビッグデータ時代の幕開け、週刊医学界新聞（2015 年 1 月 5 日）

(2) 国会図書館：ライフサイエンスのフロンティア（第 3 章：医療データ等の利活用）（2015）

(3) 厚生労働省ウェブサイト：第 1 回 NDB オープンデータ（2016 年 10 月）

(4) 武藤　猛：医療用医薬品の市場構造に関する考察、SAS ユーザ総会（2010）

(5) 福田秀人：ランチェスター思考－競争戦略の基礎、東京経済新報社（2008）

(6) 武藤　猛：医薬品エリアマーケティングへの新しいアプローチ、SAS ユーザ総会（2015）

# 医薬品開発の意思決定における

# Bayesian Posterior Probabilityの適用例

## ~ Random-Walk Metropolis vs. No-U-Turn Sampler ~

○作井 将、清水 康平、舟尾 暢男

武田薬品工業株式会社 日本開発センター 生物統計室

**Using Bayesian Posterior Probability for Go/No-Go Decision Making in Clinical Development**

Sho Sakui, Kohei Shimizu, Nobuo Funao

Takeda Pharmaceutical Company, Ltd.

## 要旨

2 つの薬剤効果を比較する並行群間/クロスオーバーデザインにて Bayesian Posterior Probability の算出と意思決定への適用例を紹介するとともに、mcmc プロシジャの random-walk Metropolis と No-U-Turn Sampler の結果を比較する。

キーワード：意思決定（Go/No-Go Decision）、Bayesian Posterior Probability、並行群間デザイン、クロスオーバーデザイン、mcmc プロシジャ、No-U-Turn Sampler（NUTS）

## 1 はじめに

医薬品開発では様々な場面で意思決定を行う機会があるが、重要なものとしては早期臨床試験終了後に次相以降の臨床試験実施の要否を検討することが挙げられる。特に、第 2 相試験や第 3 相試験（後期臨床試験）での失敗は、多くの患者さんに対して有用性が乏しい化合物を曝露したことへの倫理的問題が生じ、臨床試験に携わっていただいた医療機関の方々の多大な努力が無駄になり、医薬品開発コストの高騰や他の新薬候補の開発の遅れにも繋がることから、後期臨床試験への移行に関する意思決定（Go/No-Go Decision）が非常に重要となることは本稿で詳述するまでもない。この問題を解消する方法として、Go/No-Go Decision を行うためのデータを早期臨床試験にて取得することが挙げられる。そのために例えば、後期第 2 相試験の前に、有効性評価を主目的とした Proof-of-Concept（PoC）試験を実施することや、第 1 相試験において有効性評価項目に関するデータを収集することが行われる。ただ、これらを実施した場合でも、従来の頻度論にてデータ解析を行ってしまうと、被験者数が少ないことも遠因となり、当該試験のデータの記述に終始してしまうことが多く、Go/No-Go Decision に繋がる材料が得られにくい。一方、FDA（2010）や Walley et. al.（2015）では Bayesian Posterior Probability による意思決定方法が提案されており、後者では、PoC 試験終了後に有効性評価項目に関して「被験薬の効果がプラセ

ボに対して閾値以上の差がある」確率を Bayesian Posterior Probability として算出し、後期臨床試験への移行に関する Go/No-Go Decision の材料とすることを提案している。
本稿では、2 つの薬剤の有効性に関する効果を比較することを想定し、PoC 試験において良く用いられる並行群間デザインに加え、第 1 相試験においてしばしば適用されるクロスオーバーデザインについて Bayesian Posterior Probability を算出する例を紹介する。それぞれのデザインにおいて、まず解析的又は漸近的に Bayesian Posterior Probability を算出する SAS マクロを紹介した後、次に Markov Chain Monte Carlo（MCMC）にて同様の計算を行うための mcmc プロシジャの記述例を紹介する。mcmc プロシジャにて解析を行う際、random ステートメントによるコーディングを行うことがあるが、本ステートメントを用いると乱数列の収束やサンプリング効率が一般的に落ちる。そこで、比較的新しいサンプリング手法である No-U-Turn Sampler（NUTS）にて解析を行う例を紹介し、mcmc プロシジャにてデフォルトで用いられているアルゴリズム（random-walk Metropolis）との結果の違いも併せて考察することにする。

## 2 解析的/漸近的アプローチによる Bayesian Posterior Probability の算出例

早期臨床試験において 2 つの薬剤（1：被験薬、2：対照薬）の有効性に関する効果を比較することを考える。治験デザインは並行群間試験又は 2×2 クロスオーバー試験を想定し、有効性に関する応答に関して「対照薬に対する被験薬の平均値の差又は薬剤間の効果の差が 2 以上」となれば被験薬に関する臨床的な効果があると判定し、後期臨床試験へ移行することとする。実際の手順としては、例えば治験終了時に後期臨床試験への移行に関する Go/No-Go Decision のために Bayesian Posterior Probability を算出し、本確率が「あらかじめ設定した閾値（例えば 60%、70%、80%）」を上回った場合は後期臨床試験へ移行するものとする。頻度論に対するベイズ解析の利点として、事前情報を事前分布として設定し、解析に加味することが出来ることが挙げられるが、本稿ではこの点については議論せず、無情報事前分布を用いることとする。

### 並行群間デザイン

有効性に関する応答として、被験薬 $X_{1i} \sim N(\mu_1, \sigma_1^2)$、対照薬 $X_{2j} \sim N(\mu_2, \sigma_2^2)$, $i = 1, \cdots, n_1$, $j = 1, \cdots, n_2$、薬剤の平均値の差 $\delta = \mu_1 - \mu_2$ に興味があり、平均値の差 $\delta$ の事後分布（$\sigma_1^2 = \sigma_2^2 = \sigma^2$ だが $\sigma^2$ は未知）を求めることを考える。事前分布として無情報事前分布 $\sigma^{-(n_1+n_2+2)}$ を設定すると、$\delta = \mu_1 - \mu_2$ の事後分布は以下の Shifted & Scaled $t$ 分布に従う（計算過程は繁桝（1985）を参照）。

- $T = \frac{\delta - (\bar{x}_{1\cdot} - \bar{x}_{2\cdot})}{\sqrt{(n_1^{-1} + n_2^{-1})(s_1^2 + s_2^2)/\nu}}$ は自由度 $\nu = n_1 + n_2 - 2$ の $t$ 分布に従う
- $\bar{x}_{k\cdot} = n_k^{-1} \sum_{i=1}^{n_k} x_{ki}$, $s_k^2 = \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_{k\cdot})^2$, $k = 1, 2$

数値例として、各薬剤の例数は 20、応答に関する各薬剤の平均値はそれぞれ 3 と 0、標準偏差はそれぞれ 4 と 5、「平均値の差が 2 以上となる確率」を算出すると、Bayesian Posterior Probability は 75.54%となる。

```
%macro PROB1(_DELTA=2, _N1=20, _N2=20, _MU1=3, _MU2=0, _SD1=4, _SD2=5, _LOWER="Y") ;
  data PROB1 ;
    T=(&_DELTA.-(&_MU1.-&_MU2.))/
       sqrt( (1/&_N1.+1/&_N2.)*((&_N1.-1)*&_SD1.**2+(&_N2.-1)*&_SD2.**2)/(&_N1.+&_N2.-2) ) ;
    if (&_LOWER.="Y") then PROB=  cdf("t", T, &_N1.+&_N2.-2) ;
    else                   PROB=1-cdf("t", T, &_N1.+&_N2.-2) ;
```

```
  run ;
  proc print noobs ; run ;
%mend ;
%PROB1(_LOWER="N");   * 引数 _LOWER="Y" にて分布の下側確率を算出する ;
```

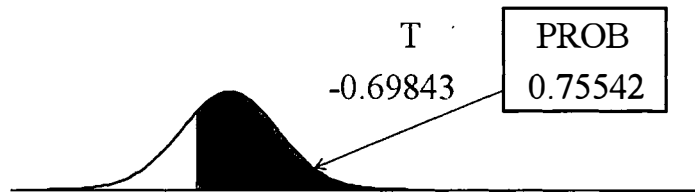



**図 2-1　%PROB1 の実行結果と Bayesian Posterior Probability のイメージ**

## 2×2 クロスオーバーデザイン

有効性に関する応答を $Y_{ijk}$ とし、以下の 2×2 クロスオーバーデザイン（AB/BA デザイン）を想定する。

**表 2-1　2×2 クロスオーバーデザインの例**

| | 時期 1 | 時期 2 |
|---|---|---|
| 順序 1 ($n_1 = 20$)<br>被験薬(1) → 対照薬(2) | $\bar{y}_{11\cdot} = 6$ | $\bar{y}_{12\cdot} = 3$ |
| 順序 2 ($n_2 = 20$)<br>対照薬(2) → 被験薬(1) | $\bar{y}_{21\cdot} = 2$ | $\bar{y}_{22\cdot} = 5$ |

解析するモデル式は $Y_{ijk} = \mu + \pi_j + \tau_{d[i,j]} + \lambda_{d[i,j-1]} + s_{ik} + \varepsilon_{ijk}$ なる分散分析モデルとする。また、Within-subjects residual SS を SSE、Between-subjects residual SS を SSP と略記する。

- $\mu$ :　切片項
- $\pi_j$ :　時期 $j$ の効果（$j = 1, 2$）
- $\tau_{d[i,j]}$ :　順序 $i$ 、時期 $j$ の薬剤効果（$i = 1, 2$ ;　薬剤 $d[i,j] = 1, 2$）
- $\lambda_{d[i,j-1]}$ :　薬剤 $d[i,j-1]$ の持ち越し効果（$d[i,0] = 0$ とする）
- $s_{ik}$ :　順序 $i$ の被験者 $k$ の効果（$i = 1,2;\ k = 1,\cdots,n_i$）、平均 0、分散 $\sigma_s^2$ の確率変数
- $\varepsilon_{ijk}$ :　誤差項（平均 0 、分散 $\sigma^2$）
- SSE :　$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n_i} y_{ijk}^2 - \sum_{i=1}^{2}\sum_{k=1}^{n_i} y_{i\cdot k}^2/2 - \sum_{i=1}^{2}\sum_{j=1}^{2} y_{ij\cdot}^2/n_i + \sum_{i=1}^{2} y_{i\cdot\cdot}^2/(2n_i)$
- SSP :　$\sum_{i=1}^{2}\sum_{k=1}^{n_i} y_{i\cdot k}^2/2 - \sum_{i=1}^{2} y_{i\cdot\cdot}^2/(2n_i)$
  $\left(y_{ij\cdot} = \sum_{k=1}^{n_i} y_{ijk}\ ,\ y_{i\cdot k} = \sum_{j=1}^{2} y_{ijk}\ ,\ y_{i\cdot\cdot} = \sum_{j=1}^{2} y_{ij\cdot}\ ,\ \bar{y}_{ij\cdot} = n_i^{-1}\sum_{k=1}^{n_i} y_{ijk}\right)$

ここで、$T = \tau_1 = -\tau_2,\ R = \lambda_1 = -\lambda_2$ なる制約をおくと（すなわち、$\tau_1 - \tau_2 = 2T, \lambda_1 - \lambda_2 = 2R$）、確率変数 $T$ と $R$ の推定値はそれぞれ $\hat{T} = (\bar{y}_{11\cdot} - \bar{y}_{12\cdot} - \bar{y}_{21\cdot} + \bar{y}_{22\cdot})/4,\ \hat{R} = (\bar{y}_{11\cdot} + \bar{y}_{12\cdot} - \bar{y}_{21\cdot} - \bar{y}_{22\cdot})/2$ となる。Bayesian Posterior Probability を算出するためには $T|y$ の分布に興味をもつことになるが、この計算は結構大変なので、Grieve (1994) の漸近式を用いる。$T$ と $R$ の無情報同時事前分布を $\left(\sigma^2(2\sigma_s^2 + \sigma^2)\right)^{-1}$ の定数倍とし、$m = (n_1 + n_2)/n_1 n_2$ とおくと、各事後分布は以下の分布に従う。

- $R|y$ :　$\dfrac{R - \hat{R}}{\sqrt{mSSP\ /\ 2(n_1 + n_2 - 2)}}$ は自由度 $n_1 + n_2 - 2$ の $t$ 分布に従う

- $T|R,y$ : $\frac{T-\hat{T}-R/2}{\sqrt{mSSE/8(n_1+n_2-2)}}$ は自由度 $n_1+n_2-2$ の $t$ 分布に従う

- $T|y$ : $\frac{T-\hat{T}-\hat{R}/2}{\sqrt{mb_0/8b_1}}$ は自由度 $b_1$ の $t$ 分布に従う $\left(b_1=\frac{(SSE+SSP)^2(n_1+n_2-6)}{SSE^2+SSP^2}+4,\ b_0=\frac{(b_1-2)(SSE+SSP)}{n_1+n_2-4}\right)$

数値例として、各順序の例数は 20、各順序・各時期の応答に関する平均値は上表、SSE と SSP はそれぞれ 250 と 480、「薬剤の平均値の差が 2 以上となる確率」を $T|R,y$ 及び $T|y$ の分布から算出すると、いずれも Bayesian Posterior Probability は 50%となる。ただし、実際の場面では、$T|R,y$ と $T|y$ のどちらの分布を使用するかは判断が難しく、結果も大きく異なる場合があるので注意が必要である。

```
%macro PROB2(_DELTA=2, _N1=20, _N2=20, _Y11=6, _Y12=3, _Y21=2, _Y22=5, _SSE=250, _SSP=480,
            _LOWER="Y") ;
  data PROB2 ;
    M =(&_N1.+&_N2.)/(&_N1.*&_N2.) ;
    R =(&_Y11.+&_Y12.-  &_Y21+&_Y22.)/2 ;
    T =(&_Y11.-&_Y12.-&_Y21.+&_Y22.)/4 ;
    T1=(&_DELTA.-R)/sqrt((M*&_SSP.)/(2*(&_N1.+&_N2.-2))) ;
    T2=(&_DELTA.-T-R/2)/sqrt((M*&_SSE.)/(8*(&_N1.+&_N2.-2))) ;
    B1=(&_N1.+&_N2.-6)*(&_SSE.+&_SSP.)**2/(&_SSE.**2+&_SSP.**2)+4 ;
    B0=(B1-2)*(&_SSE.+&_SSP.)/(&_N1.+&_N2.-4) ;
    T3=(&_DELTA.-T-R/2)/sqrt((M*B0)/(8*B1)) ;
    if (&_LOWER.="Y") then do ;
      PROB1=cdf("t", T1, &_N1.+&_N2.-2) ;
      PROB2=cdf("t", T2, &_N1.+&_N2.-2) ;
      PROB3=cdf("t", T3, B1) ;
    end ;
    else do ;
      PROB1=1-cdf("t", T1, &_N1.+&_N2.-2) ;
      PROB2=1-cdf("t", T2, &_N1.+&_N2.-2) ;
      PROB3=1-cdf("t", T3, B1) ;
    end ;
  run ;
  proc print noobs ; run ;
%mend ;
%PROB2(_LOWER="N");   * 引 数 _LOWER="Y" にて分布の下側確率を算出する ;
```

| M | R | T | T1 | T2 | B1 | B0 | T3 | PROB1 | PROB2 | PROB3 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 1 | 1.5 | 1.25831 | 0 | 65.8593 | 1294.93 | 0 | 0.10798 | 0.5 | 0.5 |

## 3 Markov Chain Monte Carlo による Bayesian Posterior Probability の算出例

前項では、並行群間デザインと 2×2 クロスオーバーデザインにおいて Bayesian Posterior Probability を算出する例を挙げたが、並行群間デザインの場合は共変量の追加や経時データへの拡張、クロスオーバーデザインの場合は高次のデザインへの拡張を行う等、デザインをより複雑にすると解析的/漸近的に事後分布を求めることが困難となる。また、2×2 クロスオーバーデザインにて Grieve (1994) の漸近式を用いた場合、$T|R,y$ と $T|y$ の分布の選択が難しい。そこで、先程と同様の設定で、今度は mc mc プロシジャによる Bayesian Posterior Probability の算出を試みる。なお、2×2 クロスオーバーデザインにおける mc mc プロシジャのモデルには持ち越し効果の代わりに順序効果を指定した。

## 並行群間デザイン

まず、各薬剤の例数は20、応答に関する各薬剤の平均値はそれぞれ3と0、標準偏差はそれぞれ4と5と想定し、乱数にてデータを生成した後、glm プロシジャにて頻度論の枠組みで解析を行い、各薬剤の平均値とその95%信頼区間及び平均値の差の点推定値とその95%信頼区間を算出する。

```
data MYDATA ;
  do GROUP=1 to 2 ;
    do ID=1 to 20 ;
      Y=3*(2-GROUP)+(3+GROUP)*normal(7777) ; output ;
    end ;
  end ;
run ;

proc glm data=MYDATA ;
  class   GROUP ;
  model   Y = GROUP / solution ss3 ;
  lsmeans GROUP     / cl pdiff ;
run ;
```

GLM プロシジャ

Y の最小 2 乗平均

| GROUP | | 95% 信頼限界 | |
|---|---|---|---|
| 1 | 2.725909 | 0.677707 | 4.774111 |
| 2 | -0.219506 | -2.267708 | 1.828696 |

効果 GROUP に対する最小 2 乗平均

| i | j | LSMean(i)-LSMean(j) の平均の差 | 95% 信頼限界 | |
|---|---|---|---|---|
| 1 | 2 | 2.945414 | 0.048819 | 5.842010 |

次に、mcmc プロシジャにより解析を行った後、「平均値の差が2以上となる確率」を算出すると、Bayesian Posterior Probability は 75.95%となる。各薬剤の事後平均とその95%確信区間及び平均値の差に関する事後平均とその95%確信区間は頻度論の結果と同様であるが、Autocorrelation Time や Efficiency の結果を見ると、サンプリングの効率が良くなく、自己相関が残っていることが伺える。

```
proc mcmc data=MYDATA outpost=OUT seed=777 nmc=10000 monitor=(_parms_ DIFF)
          statistics(alpha=0.05) ;
  parm mu1  0 mu2 0 ;
  parm sig2 1 ;
  prior mu:  ~ general(0) ;
  prior sig2 ~ general(-log(sig2), lower=0) ;
  diff = mu1 - mu2 ;
  if (GROUP=1) then mu = mu1 ;
  else              mu = mu2 ;
  model Y ~ normal(mu, var=sig2) ;
run ;

proc format ;
  value diff_f low -< 2   ="Diff <  2"
               2   <- high="Diff >= 2" ;
run ;

proc freq data=OUT ;
  tables diff / nocum ;
  format diff diff_f. ;
run ;
```

The MCMC Procedure
Posterior Summaries and Intervals

| Parameter | N | Mean | Standard Deviation | 95% HPD Interval | |
|---|---|---|---|---|---|
| mu1 | 10000 | 2.7678 | 0.9947 | 0.8838 | 4.7515 |
| mu2 | 10000 | -0.2749 | 1.0316 | -2.3385 | 1.7203 |
| sig2 | 10000 | 21.6602 | 5.1779 | 12.6395 | 31.5817 |
| diff | 10000 | 3.0428 | 1.4397 | 0.4024 | 6.0471 |

Effective Sample Sizes

| Parameter | ESS | Autocorrelation Time | Efficiency |
|---|---|---|---|
| mu1 | 1361.9 | 7.3427 | 0.1362 |
| mu2 | 1529.0 | 6.5402 | 0.1529 |
| sig2 | 1323.4 | 7.5563 | 0.1323 |
| diff | 1417.9 | 7.0528 | 0.1418 |

FREQ プロシジャ

| diff | 度数 | パーセント |
|---|---|---|
| Diff < 2 | 2405 | 24.05 |
| Diff >= 2 | 7595 | 75.95 |

そこで、mcmc プロシジャにおける Burn-in、thinning 及びサンプリングの数を 5 倍（nbi=5000、thin=5、nmc=50000）にして再実行する。結果として効率はある程度上がるが、計算時間が増えてしまう。

```
proc mcmc data=MYDATA outpost=OUT seed=777 nbi=5000 thin=5 nmc=50000 monitor=(_parms_ DIFF)
        statistics(alpha=0.05) ;
```

（以降は前プログラムと同一の命令）

Effective Sample Sizes

| Parameter | ESS | Autocorrelation Time | Efficiency |
|---|---|---|---|
| mu1 | 5572.2 | 1.7946 | 0.5572 |
| mu2 | 6005.7 | 1.6651 | 0.6006 |
| sig2 | 5909.9 | 1.6921 | 0.5910 |
| diff | 6085.1 | 1.6434 | 0.6085 |

参考までに、上記 2 つのプログラムの診断プロットを次頁に示す。

**【参考】**前掲のプログラムでは、薬剤を表すカテゴリ変数 GROUP の値によって条件分岐を行うことで、model ステートメントで使用するパラメータ mu の値を指定していた。切片項を表すパラメータと random ステートメントを併用することで、カテゴリ変数 GROUP の値による条件分岐を行う必要が無くなりコーディングが簡便となる（ただし、サンプリングの効率が落ちる）。次項の 2×2 クロスオーバーデザインにてこのコーディング方針を適用する。

```
proc mcmc data=MYDATA outpost=OUT seed=777 nbi=5000 thin=5 nmc=50000 statistics(alpha=0.05) ;
  parm   int  0 ;
  parm   sig2 1 ;
  prior  int  ~ general(0) ;
  prior  sig2 ~ general(-log(sig2), lower=0) ;
  random mu   ~ general(0) subject=GROUP monitor=(mu) init=0 zero=last ;

  model Y ~ normal(int+mu, var=sig2) ;
run ;
```

Effective Sample Sizes

| Parameter | ESS | Autocorrelation Time | Efficiency |
|---|---|---|---|
| int | 2988.7 | 3.3459 | 0.2989 |
| sig2 | 4802.6 | 2.0822 | 0.4803 |
| mu_1 | 2806.7 | 3.5629 | 0.2807 |

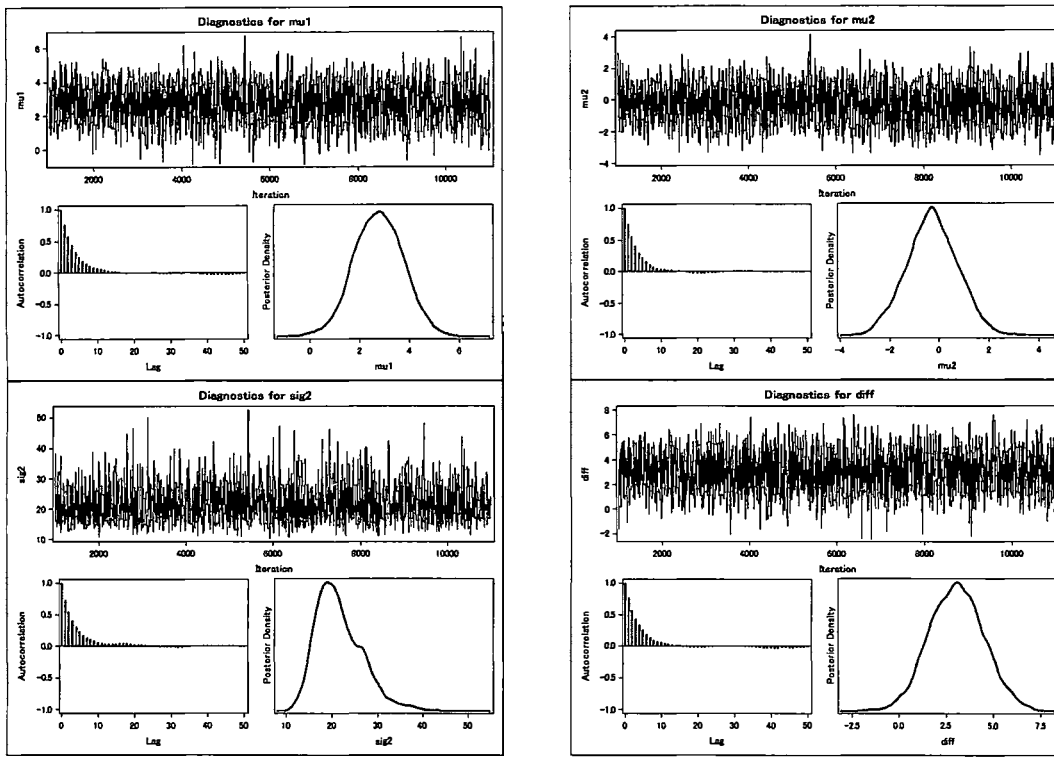



図 3-1　mcmc プロシジャの実行結果（nbi=1000、thin=1、nmc=10000）

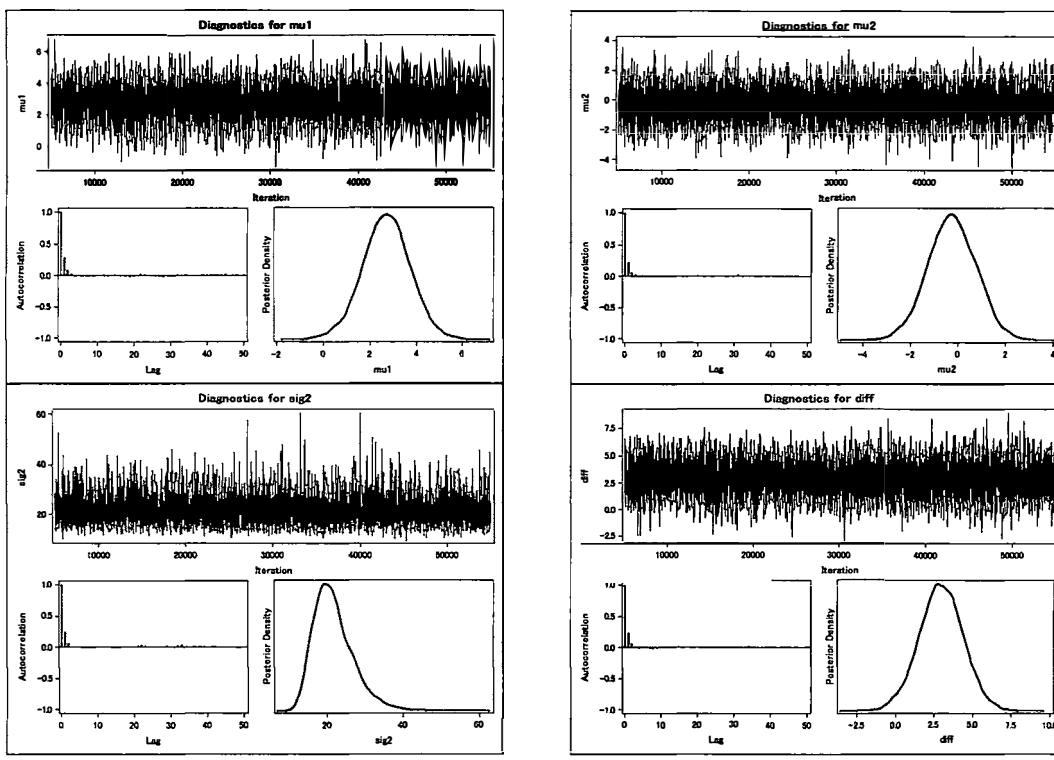



図 3-2　mcmc プロシジャの実行結果（nbi=5000、thin=5、nmc=50000）

## 2×2 クロスオーバーデザイン

まず、各順序の例数及び各セルの応答に関する平均値は表 2-1 と同様、各セルの応答に関する標準偏差は全て 3 と想定し、乱数にてデータを生成した後、glm プロシジャにて頻度論の枠組みで解析を行い、薬剤効果の値の差の点推定値とその 95%信頼区間を算出する。

```
data MYDATA ;
  GROUP=1 ;
  do ID=1 to 20 ;
    PERIOD=1; TREAT=1; Y=6+3*normal(7777) ; output ;
    PERIOD=2; TREAT=2; Y=3+3*normal(7777) ; output ;
  end ;
  GROUP=2 ;
  do ID=21 to 40 ;
    PERIOD=1; TREAT=2; Y=2+3*normal(7777) ; output ;
    PERIOD=2; TREAT=1; Y=5+3*normal(7777) ; output ;
  end ;
run ;

proc glm data=MYDATA ;
  class   ID GROUP PERIOD TREAT ;
  model   Y = GROUP ID(GROUP) PERIOD TREAT / solution ss3 ;
  test    h=GROUP e=ID(GROUP) ;
  lsmeans TREAT / cl pdiff ;
run ;
```

GLM プロシジャ

効果 TREAT に対する最小 2 乗平均

| i | j | 平均の差 | LSMean(i)-LSMean(j) の 95% 信頼限界 | |
|---|---|---|---|---|
| 1 | 2 | 2.834587 | 1.657083 | 4.012091 |

次に、mcmc プロシジャにより解析を行った後、「薬剤効果の差が 2 以上となる確率」を算出すると、Bayesian Posterior Probability は 91.16%となる。薬剤効果の値の差の点推定値とその 95%確信区間は頻度論の結果と同様である。しかし、mcmc プロシジャにおける Burn-in、thinning 及びサンプリングの数をそれぞれ 5000、5、50000 にして実行したにも関わらず、Autocorrelation Time や Efficiency の結果を見ると、サンプリングの効率が良くなく、自己相関が残っていることが伺える。この原因としては、モデルに複数の random ステートメントが含まれていることが挙げられる。

```
proc mcmc data=MYDATA outpost=OUT seed=777 nbi=5000 thin=5 nmc=50000 statistics(alpha=0.05) ;
   parms  residual 1 sig2 1 ;
   prior  residual ~ igamma(0.001, scale=0.001)  ;
   prior  sig2     ~ igamma(0.001, scale=0.001)  ;
   random seq      ~ normal(0, var=1000) subject=GROUP  monitor=(seq) zero=last ;
   random tau      ~ normal(0, var=1000) subject=TREAT  monitor=(tau) zero=last ;
   random pi       ~ normal(0, var=1000) subject=PERIOD monitor=(pi)  ;
   random s        ~ normal(0, var=sig2) subject=ID ;

   model y ~ normal(seq + tau + pi + s, var=residual) ;
run ;

proc freq data=OUT ;
  tables tau_1 / nocum ;
  format tau_1 diff_f. ;
run ;
```

The MCMC Procedure
Posterior Summaries and Intervals

| Parameter | N | Mean | Standard Deviation | 95% HPD Interval | |
|---|---|---|---|---|---|
| residual | 10000 | 8.0493 | 1.9739 | 4.5591 | 11.9088 |
| sig2 | 10000 | 2.0339 | 1.7767 | 0.00362 | 5.3782 |
| seq_1 | 10000 | 0.5742 | 0.7792 | -0.9330 | 2.0854 |
| tau_1 | 10000 | 2.8447 | 0.6341 | 1.6619 | 4.1611 |
| pi_1 | 10000 | 2.0109 | 0.7208 | 0.5419 | 3.3974 |
| pi_2 | 10000 | 2.6627 | 0.7194 | 1.1777 | 3.9717 |

Effective Sample Sizes

| Parameter | ESS | Autocorrelation Time | Efficiency |
|---|---|---|---|
| residual | 245.0 | 40.8182 | 0.0245 |
| sig2 | 153.8 | 65.0369 | 0.0154 |
| seq_1 | 1470.4 | 6.8010 | 0.1470 |
| tau_1 | 2486.9 | 4.0210 | 0.2487 |
| pi_1 | 1430.4 | 6.9912 | 0.1430 |
| pi_2 | 1482.3 | 6.7462 | 0.1482 |

FREQ プロシジャ

| diff | 度数 | パーセント |
|---|---|---|
| Diff < 2 | 884 | 8.84 |
| Diff >= 2 | 9116 | 91.16 |

## 4 No-U-Turn Sampler による Bayesian Posterior Probability の算出例

Hamiltonian Monte Carlo Method（HMC）は、解析力学等で用いられる Hamiltonian の概念を利用し、Leap-frog Method と呼ばれる手法で乱数を生成する。アルゴリズムの詳細は豊田（2015）が詳しいため本稿では詳述を避けるが、HMC は従来の手法と比較して収束が速く、乱数間の相関が低いため、通常の MCMC（例えば mcmc プロシジャでデフォルトとして用いられる random-walk Metropolis）よりもサンプリング数が数十分の一で済む場合があると言われている。また、No-U-Turn Sampler（NUTS）は HMC の変法で、HMC の計算で必要となる各パラメータを自動で調整しつつサンプリングを行うため、より実践的な手法である。NUTS は mcmc プロシジャの他、Stan という確率的プログラミング言語で実装されている。Stan の実装例としては例えば R のパッケージのひとつである RStan があるが、RStan では解析を実行する度に、R や Stan のコードを C++に変換してコンパイルを行った後、C++上でサンプリングを行うため、簡単なモデルでも実行時間がかかる。一方、mcmc プロシジャではアルゴリズムを random-walk Metropolis から NUTS に変更しても実行環境や動作は大きく変わらないため、実行時間はそれほど増えず、動作が軽い印象がある。本項では、先程紹介した mcmc プロシジャを NUTS で再実行し、効率等の観点から通常の MCMC（random-walk Metropolis）と比較を試みる。

以下、前項の mcmc プロシジャのプログラムについて、アルゴリズムを NUTS に変更して再実行した結果を紹介するが、アルゴリズムの変更方法は mcmc プロシジャの 1 行目に「alg=nuts」を追加するのみである。

```
proc mcmc data=MYDATA outpost=OUT seed=777 ...  statistics(alpha=0.05) alg=nuts ;
```

### 並行群間デザイン

前項の mcmc プロシジャのプログラムを再実行するが、その際、Burn-in 及び thinning の数をそれぞれ 5000 と 5、アルゴリズムを「デフォルト（random-walk Metropolis）」と「NUTS」の両方、NUTS の場合はサンプリング数を「50000」「5000」「500」の 3 通り、計 4 通りの実行結果を以下に示す。まず、事後分布に関する要約統計量はいずれも同様の結果となった。

| パラメータ | Posterior Summaries and Intervals | | | | | |
|---|---|---|---|---|---|---|
| alg=**normal**（デフォルト）<br>nbi=5000 thin=5 nmc=**50000** | Parameter | N | Mean | Standard Deviation | 95% HPD Interval | |
| | mu1 | 10000 | 2.7263 | 1.0324 | 0.7039 | 4.8101 |
| | mu2 | 10000 | -0.2373 | 1.0241 | -2.3522 | 1.6881 |
| | sig2 | 10000 | 21.7136 | 5.2970 | 12.9774 | 32.6929 |
| | diff | 10000 | 2.9636 | 1.4399 | 0.0918 | 5.7872 |
| alg=**nuts**<br>nbi=5000 thin=5 nmc=**50000** | Parameter | N | Mean | Standard Deviation | 95% HPD Interval | |
| | mu1 | 10000 | 2.7142 | 1.0440 | 0.7138 | 4.8019 |
| | mu2 | 10000 | -0.2074 | 1.0590 | -2.2917 | 1.7970 |
| | sig2 | 10000 | 21.6331 | 5.3229 | 12.2837 | 31.9301 |
| | diff | 10000 | 2.9215 | 1.4892 | -0.0224 | 5.8308 |
| alg=**nuts**<br>nbi=5000 thin=5 nmc=**5000** | Parameter | N | Mean | Standard Deviation | 95% HPD Interval | |
| | mu1 | 1000 | 2.7291 | 1.0202 | 0.9364 | 5.0202 |
| | mu2 | 1000 | -0.2457 | 1.0474 | -2.3428 | 1.7765 |
| | sig2 | 1000 | 21.7164 | 5.4621 | 12.4431 | 32.1279 |
| | diff | 1000 | 2.9748 | 1.4814 | 0.1426 | 6.0151 |
| alg=**nuts**<br>nbi=5000 thin=5 nmc=**500** | Parameter | N | Mean | Standard Deviation | 95% HPD Interval | |
| | mu1 | 100 | 2.6088 | 1.0192 | 0.4052 | 4.4017 |
| | mu2 | 100 | -0.2433 | 0.9066 | -1.7168 | 1.4576 |
| | sig2 | 100 | 22.3600 | 5.5752 | 13.8750 | 34.6725 |
| | diff | 100 | 2.8521 | 1.3400 | -0.1079 | 5.2955 |

次に、効率（Effective Sample Size）に関する結果は、アルゴリズムを NUTS に変更するだけでサンプリングの効率や自己相関が改善した。また、前表の結果も踏まえると、アルゴリズムを NUTS にした場合はサンプリング数を 1/10 程度に減らしても推定精度がある程度維持されていることが分かる。

| パラメータ | 効率（Effective Sample Size） | | | |
|---|---|---|---|---|
| alg=**normal**（デフォルト）<br>nbi=5000 thin=5 nmc=**50000** | Parameter | ESS | Autocorrelation Time | Efficiency |
| | mu1 | 5572.2 | 1.7946 | 0.5572 |
| | mu2 | 6005.7 | 1.6651 | 0.6006 |
| | sig2 | 5909.9 | 1.6921 | 0.5910 |
| | diff | 6085.1 | 1.6434 | 0.6085 |
| alg=**nuts**<br>nbi=5000 thin=5 nmc=**50000** | Parameter | ESS | Autocorrelation Time | Efficiency |
| | mu1 | 9641.5 | 1.0372 | 0.9642 |
| | mu2 | 10000.0 | 1.0000 | 1.0000 |
| | sig2 | 5277.5 | 1.8948 | 0.5278 |
| | diff | 10000.0 | 1.0000 | 1.0000 |
| alg=**nuts**<br>nbi=5000 thin=5 nmc=**5000** | Parameter | ESS | Autocorrelation Time | Efficiency |
| | mu1 | 936.5 | 1.0678 | 0.9365 |
| | mu2 | 1066.6 | 0.9375 | 1.0666 |
| | sig2 | 379.9 | 2.6320 | 0.3799 |
| | diff | 1000.0 | 1.0000 | 1.0000 |
| alg=**nuts**<br>nbi=5000 thin=5 nmc=**500** | Parameter | ESS | Autocorrelation Time | Efficiency |
| | mu1 | 91.9 | 1.0881 | 0.9191 |
| | mu2 | 78.0 | 1.2818 | 0.7801 |
| | sig2 | 33.5 | 2.9817 | 0.3354 |
| | diff | 107.0 | 0.9349 | 1.0696 |

参考までに、Geweke's Convergence Diagnostic の結果と、2〜3 番目のプログラムの診断プロットを以下に示す（なお、1 番目のプログラムの診断プロットは図 3-2 に掲載済み）。

| パラメータ | Geweke's Convergence Diagnostic | | |
|---|---|---|---|
| alg=**normal**（デフォルト）<br>nbi=5000 thin=5 nmc=**50000** | Parameter | z | Pr > \|z\| |
| | mu1 | 0.7724 | 0.4399 |
| | mu2 | -1.5570 | 0.1195 |
| | sig2 | 1.7239 | 0.0847 |
| | diff | 1.7669 | 0.0772 |
| alg=**nuts**<br>nbi=5000 thin=5 nmc=**50000** | Parameter | z | Pr > \|z\| |
| | mu1 | 0.6110 | 0.5412 |
| | mu2 | -0.7086 | 0.4786 |
| | sig2 | 0.2948 | 0.7682 |
| | diff | 0.9859 | 0.3242 |
| alg=**nuts**<br>nbi=5000 thin=5 nmc=**5000** | Parameter | z | Pr > \|z\| |
| | mu1 | -1.0353 | 0.3005 |
| | mu2 | -0.2225 | 0.8239 |
| | sig2 | 0.8094 | 0.4183 |
| | diff | -0.5526 | 0.5805 |
| alg=**nuts**<br>nbi=5000 thin=5 nmc=**500** | Parameter | z | Pr > \|z\| |
| | mu1 | 0.4566 | 0.6479 |
| | mu2 | 1.2622 | 0.2069 |
| | sig2 | -0.7638 | 0.4450 |
| | diff | -0.0709 | 0.9435 |

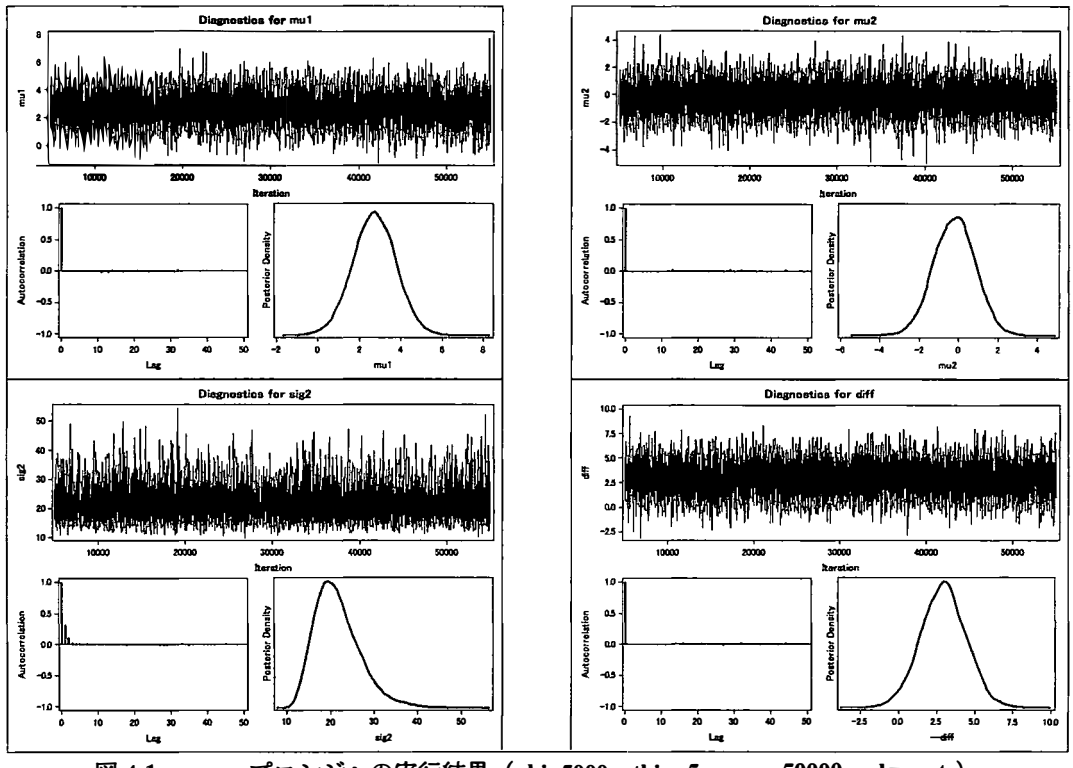



**図 4-1　mcmc プロシジャの実行結果（nbi=5000、thin=5、nmc=50000、alg=nuts）**

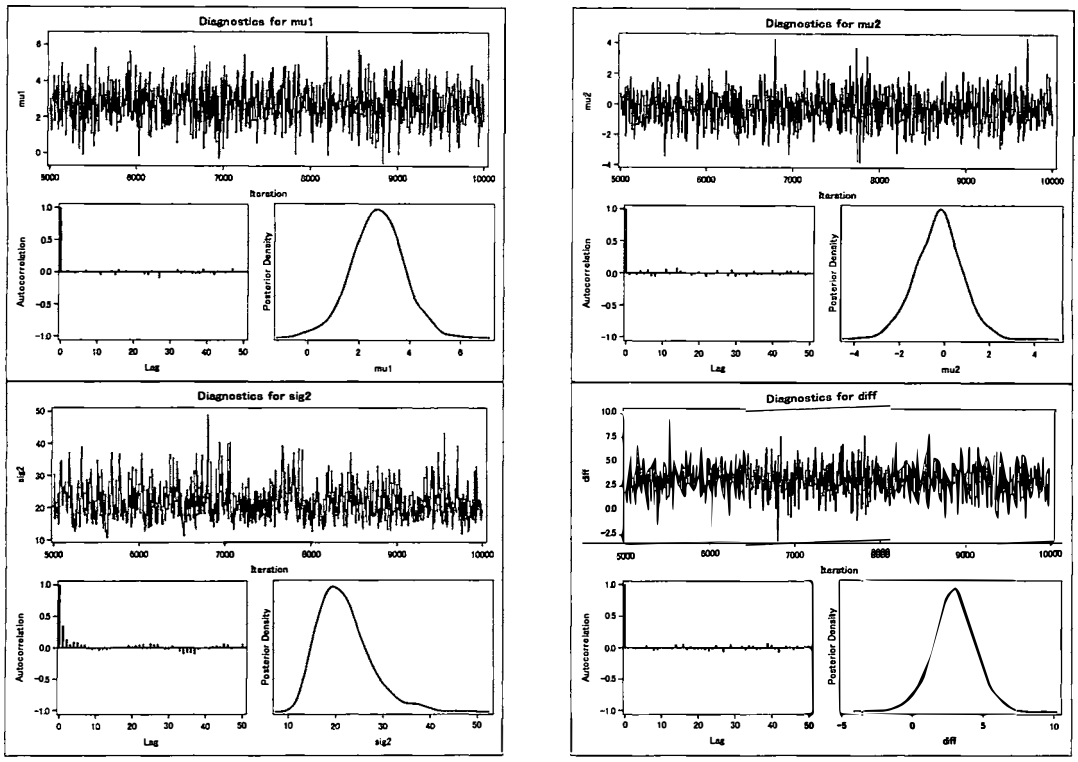



図 4-2　mcmc プロシジャの実行結果（nbi=5000、thin=5、nmc=<u>5000</u>、alg=<u>nuts</u>）

## 2×2 クロスオーバーデザイン

前項の mcmc プロシジャのプログラムを再実行するが、その際、Burn-in 及び thinning の数をそれぞれ 10000 と 10、アルゴリズムを「デフォルト（random-walk Metropolis）」と「NUTS」の両方、NUTS の場合はサンプリング数を「100000」「10000」の 2 通り、計 3 通りの実行結果を以下に示す。まず、事後分布に関する要約統計量はいずれも同様の結果となった。

| パラメータ | Posterior Summaries and Intervals | | | | | |
|---|---|---|---|---|---|---|
| alg=**normal（デフォルト）**<br>nbi=10000 thin=10 nmc=**100000** | Parameter | N | Mean | Standard Deviation | 95% HPD Interval | |
| | residual | 10000 | 8.0683 | 1.9779 | 4.5044 | 11.9148 |
| | sig2 | 10000 | 2.0432 | 1.8148 | 0.00182 | 5.4304 |
| | seq_1 | 10000 | 0.5564 | 0.7843 | -0.9888 | 2.0990 |
| | tau_1 | 10000 | 2.8352 | 0.6388 | 1.6269 | 4.1330 |
| | pi_1 | 10000 | 2.0284 | 0.7149 | 0.5868 | 3.4007 |
| | pi_2 | 10000 | 2.6823 | 0.7229 | 1.2121 | 4.0507 |
| alg=**nuts**<br>nbi=10000 thin=10 nmc=**100000** | Parameter | N | Mean | Standard Deviation | 95% HPD Interval | |
| | residual | 10000 | 7.9362 | 1.9161 | 4.4868 | 11.6854 |
| | sig2 | 10000 | 2.1671 | 1.7581 | 0.0305 | 5.4274 |
| | seq_1 | 10000 | 0.5342 | 0.7850 | -0.9584 | 2.1152 |
| | tau_1 | 10000 | 2.8416 | 0.6290 | 1.5967 | 4.0575 |
| | pi_1 | 10000 | 2.0477 | 0.7023 | 0.6432 | 3.3944 |
| | pi_2 | 10000 | 2.7046 | 0.7082 | 1.3497 | 4.1421 |

| パラメータ | Posterior Summaries and Intervals | | | | | |
|---|---|---|---|---|---|---|
| alg=**nuts**<br>nbi=10000 thin=10 nmc=**10000** | Parameter | N | Mean | Standard Deviation | 95% HPD Interval | |
| | residual | 1000 | 7.7681 | 1.8430 | 4.4485 | 11.1698 |
| | sig2 | 1000 | 2.3257 | 1.7355 | 0.0405 | 5.4617 |
| | seq_1 | 1000 | 0.5792 | 0.8020 | -0.8966 | 2.2719 |
| | tau_1 | 1000 | 2.8499 | 0.6397 | 1.5788 | 4.0941 |
| | pi_1 | 1000 | 2.0556 | 0.7144 | 0.5593 | 3.3252 |
| | pi_2 | 1000 | 2.7097 | 0.7136 | 1.3508 | 4.1274 |

次に、効率（Effective Sample Size）に関する結果は、アルゴリズムを NUTS に変更するだけでサンプリングの効率や自己相関が改善したが、並行群間デザインの場合と比較すると、改善の度合いは小さく、特にばらつきに関するパラメータについては満足出来る結果とはいえない。原因としては、やはりモデルに複数の random ステートメントが含まれていることが挙げられる。この状況下では、アルゴリズムを NUTS にしたとしてもサンプリング数の減少まで期待することは出来ないようである。

| パラメータ | 効率（Effective Sample Size） | | | |
|---|---|---|---|---|
| alg=**normal**（デフォルト）<br>nbi=10000 thin=10 nmc=**100000** | Parameter | ESS | Autocorrelation Time | Efficiency |
| | residual | 395.1 | 25.3128 | 0.0395 |
| | sig2 | 271.7 | 36.8032 | 0.0272 |
| | seq_1 | 2772.2 | 3.6073 | 0.2772 |
| | tau_1 | 4808.7 | 2.0796 | 0.4809 |
| | pi_1 | 2693.1 | 3.7132 | 0.2693 |
| | pi_2 | 2844.0 | 3.5162 | 0.2844 |
| alg=**nuts**<br>nbi=10000 thin=10 nmc=**100000** | Parameter | ESS | Autocorrelation Time | Efficiency |
| | residual | 1169.3 | 8.5525 | 0.1169 |
| | sig2 | 700.3 | 14.2789 | 0.0700 |
| | seq_1 | 9401.5 | 1.0637 | 0.9402 |
| | tau_1 | 8551.2 | 1.1694 | 0.8551 |
| | pi_1 | 9519.5 | 1.0505 | 0.9520 |
| | pi_2 | 9452.7 | 1.0579 | 0.9453 |
| alg=**nuts**<br>nbi=10000 thin=10 nmc=**10000** | Parameter | ESS | Autocorrelation Time | Efficiency |
| | residual | 160.1 | 6.2452 | 0.1601 |
| | sig2 | 82.4 | 12.1287 | 0.0824 |
| | seq_1 | 859.6 | 1.1634 | 0.8596 |
| | tau_1 | 704.1 | 1.4203 | 0.7041 |
| | pi_1 | 577.8 | 1.7307 | 0.5778 |
| | pi_2 | 659.0 | 1.5175 | 0.6590 |

参考までに、全ての Geweke's Convergence Diagnostic の結果と、1～2 番目のプログラムの診断プロットを以下に示す。

| パラメータ | Geweke's Convergence Diagnostic | | |
|---|---|---|---|
| alg=**normal**（デフォルト）<br>nbi=10000 thin=10 nmc=**100000** | Parameter | z | Pr > \|z\| |
| | residual | -0.6721 | 0.5015 |
| | sig2 | -0.0806 | 0.9358 |
| | seq_1 | 1.5806 | 0.1140 |
| | tau_1 | -0.1310 | 0.8957 |
| | pi_1 | -1.4010 | 0.1612 |
| | pi_2 | -1.2311 | 0.2183 |

| パラメータ | Geweke's Convergence Diagnostic |
|---|---|
| alg=**nuts**<br>nbi=10000 thin=10 nmc=**100000** | Parameter z Pr > \|z\|<br>residual -0.7958 0.4261<br>sig2 0.5213 0.6021<br>seq_1 1.3532 0.1760<br>tau_1 0.2820 0.7779<br>pi_1 0.6765 0.4987<br>pi_2 0.0095 0.9924 |
| alg=**nuts**<br>nbi=10000 thin=10 nmc=**10000** | Parameter z Pr > \|z\|<br>residual -0.3277 0.7431<br>sig2 0.9485 0.3429<br>seq_1 0.1327 0.8944<br>tau_1 -0.4700 0.6384<br>pi_1 0.6774 0.4981<br>pi_2 0.1971 0.8437 |

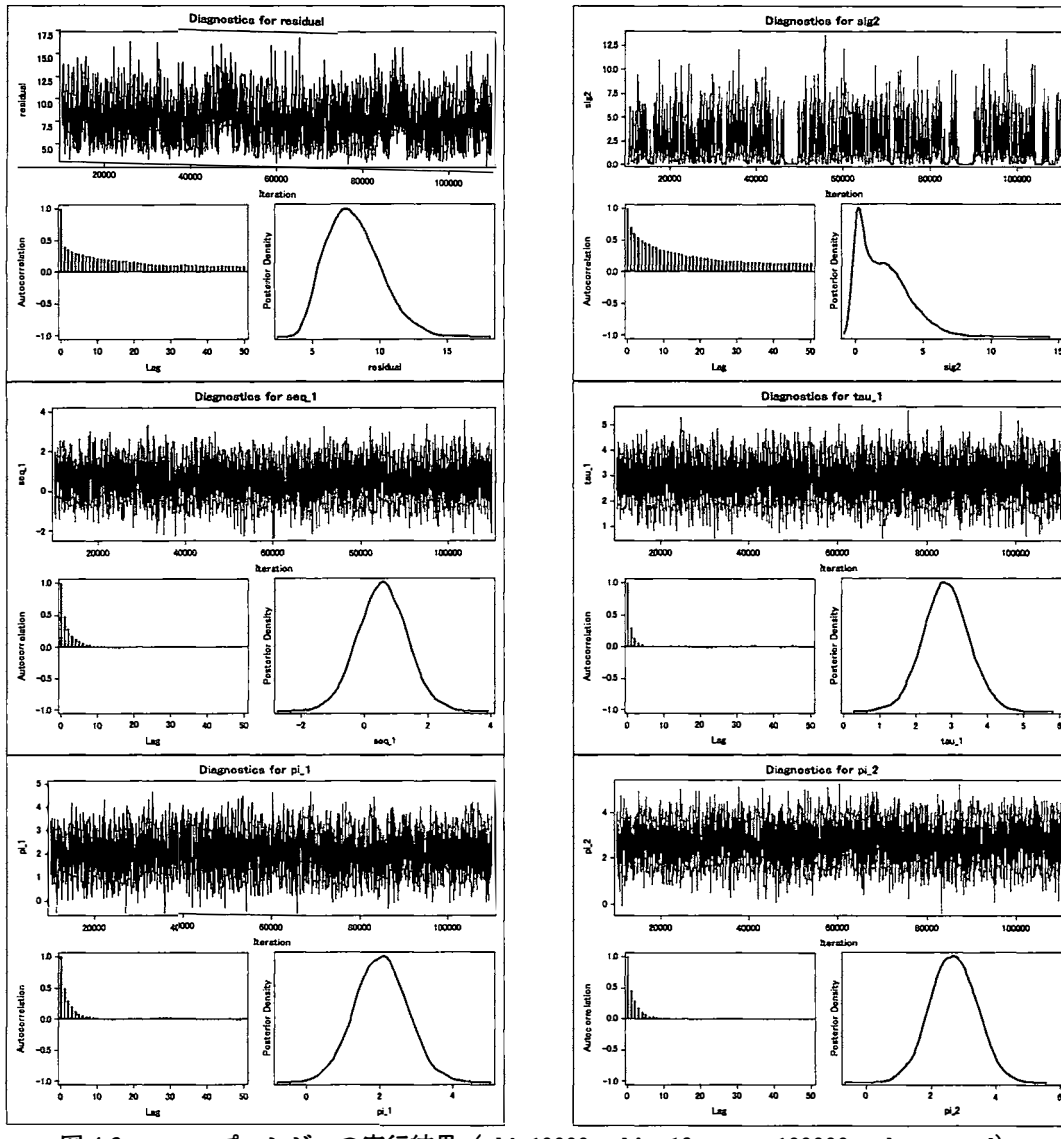



**図 4-3　mcmc プロシジャの実行結果（nbi=10000、thin=10、nmc=100000、alg=normal）**

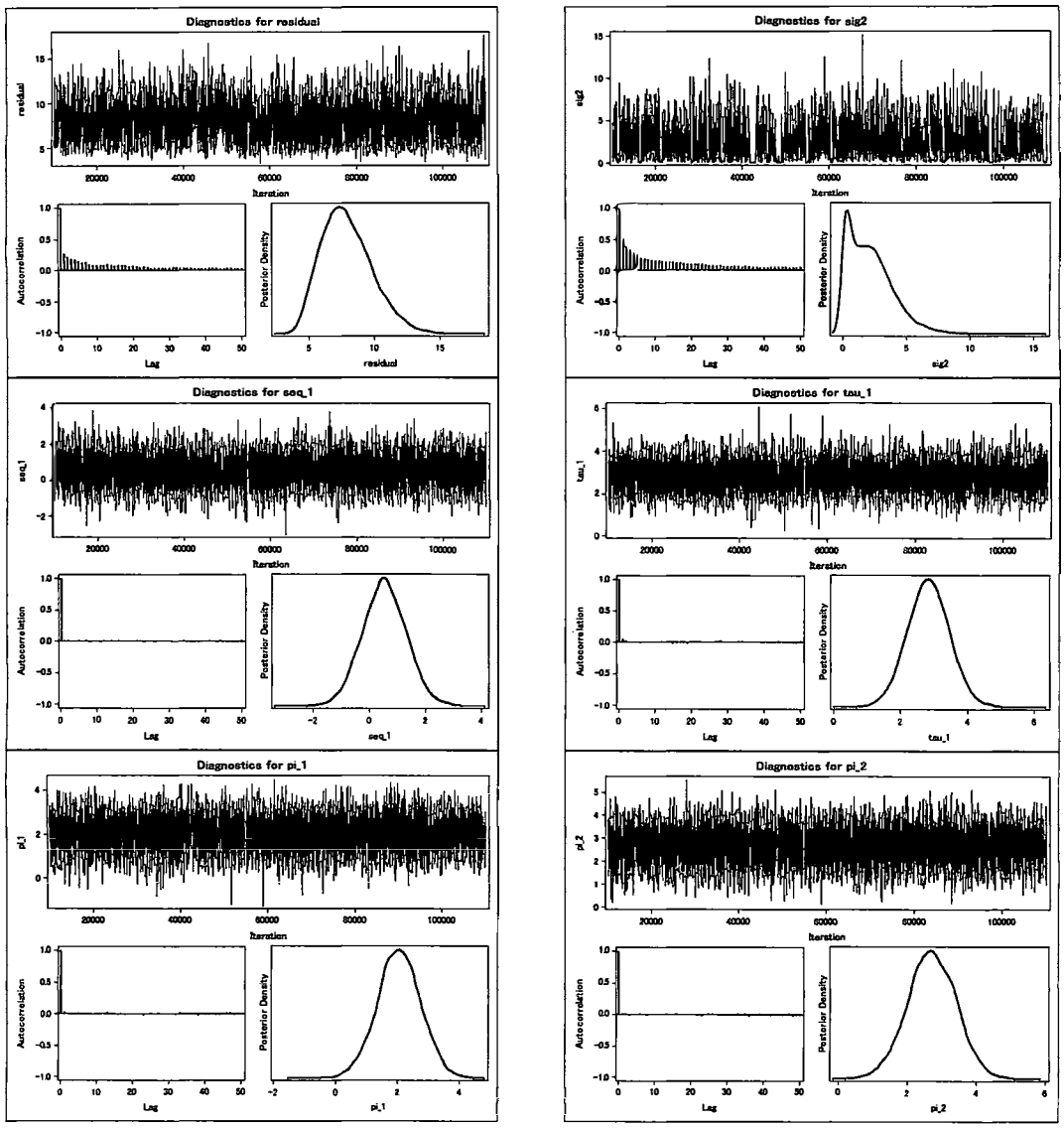



**図 4-4　mcmc プロシジャの実行結果（nbi=10000、thin=10、nmc=100000、alg=nuts）**

# 5　まとめ

臨床試験終了後に Bayesian Posterior Probability を算出することで、後期臨床試験への移行に関する意思決定（Go/No-Go Decision）のための材料が得られることを示した。早期臨床試験においては並行群間デザインに加え、クロスオーバーデザインもしばしば適用されるが、本稿ではいずれの場合においても本確率を計算することが出来る。また、mcmc プロシジャを適用して Markov Chain Monte Carlo（MCMC）による解析が可能となることも紹介したが、解析対象となるモデルや仮定が変わっても、mcmc プロシジャであれば柔軟に対応できることも申し添えておく。ただし、解析するモデルによっては乱数列の収束やサンプリング効率が良くない場合があり、特に random ステートメントを含むモデルを定義した場合は、乱数列の収束やサンプリング効率が一般的に落ちる。このような場合は NUTS によるサンプリングにより乱数列の収束状況や効率をある程度改善出来る場合があることを示した。

## 参考文献


- FDA (2010) "Guidance for the use of bayesian statistics in medical device clinical trials." https://www.fda.gov/MedicalDevices/ucm071072.htm
- Grieve AP (1994) "Bayesian analyses of two-treatment crossover studies," Statistical Methods in Medical Research; 3, 407-429.
- Jones B, Kenward MG (2014) "Design and Analysis of Cross-Over Trials, Third Edition," Chapman & Hall/CRC.
- SAS Institute Inc. (2015) "SAS/STAT(R) 14.1 User's Guide."
- Walley RJ. et. al. (2015) "Advantages of a wholly Bayesian approach to assessing efficacy in early drug development: a case study," Pharmaceutical Statistics, 14: 205-215.
- 繁桝 算男 (1985) "ベイズ統計入門" 東京大学出版会
- 豊田 秀樹 (2015) "基礎からのベイズ統計学" 朝倉書店
- 松浦 健太郎 他 (2016) "Stan と R でベイズ統計モデリング" 共立出版


以上

# Rプログラムも使えるJMPの活用法　～JMPに手法がない場合～

○原茂恵美子、佐藤拓哉、篠津和夫、佐藤耕一
(株式会社タクミインフォメーションテクノロジー)

医薬品が市販されるまでには、創薬研究から動物を対象とした非臨床試験、ヒトを対象とした臨床試験を経て、それらが審査されたのち市販されるという流れがある。統計ソフトは、非臨床試験ではSASやJMP、臨床試験ではSASが主に使われてきた。薬物動態、薬効・薬理、毒性試験といった非臨床試験では、近年、探索的なデータ分析を行う必要性から操作性やデータの可視化などにおいて優れた機能を持つJMPのニーズが高まっている。

JMPは探索的なデータ分析ができるソフトフェアで、プログラムを書くことなく、マウス操作によりさまざまな統計解析を行うことができる。しかし、非臨床試験に必要とされる統計手法の全てがJMPに搭載されているわけではない。JMPスクリプト言語（以下、JSL）を使用してプログラムを組むことによりカスタマイズすることができる。そのため、未搭載の解析手法を利用するための解決方法として

① JMP スクリプト言語の利用
② JMP から R 言語など他の言語のコードを実行

という方法が挙げられる。そこで今回、上記①および②の方法により、現在JMPに搭載されていないノンパラメトリックDunnett型多重比較のアドイン開発を行った。

ノンパラメトリックDunnett型多重比較は、要因が2群以上の場合に、対照群とその他の比較群の間に差があるかを、順位データに変換して比較する。本手法では、検定の初めに全群に対して元データの順位データへの変換を行う、ジョイントランキング法を採用した。

JSLにはDunnett型多重比較のp値および分位点を戻す関数が存在しているが、両側検定のみであった。そこで、両側検定にJSL、片側検定にR言語を使用し、アドイン「ノンパラメトリックDunnett型多重比較(ジョイントランキング法)」を開発し実装した。

値の検証は、SASのPROBMC関数で両側検定と片側検定を行い、それぞれJMPとRの出力結果と比較を行った。小数点以下5桁で比較した結果、p値および棄却限界値が全て一致した。以上から、開発したアドインでの解析結果に問題はなく、SASと同等の利用ができると判断した。

JMPは、Rプログラムを起動するJSLを実行することでR言語を利用することができる。ユーザーは、Rプログラムを利用していることを意識することなく、既存のJMPの手法同様に操作し、解析をおこなうことができる。JMPに搭載されていない解析手法は、JSLを使用してアドインを開発することによりJMPの操作性を損なうことなく利用が可能である。非臨床以外の他の分野においても、必要な手法がJMPに搭載されていない場合には、JSLや他言語を利用することで、JMPを効果的かつ有効に利用されることが示唆される。

参考文献

浜田・岸本 (1995) SASによるノンパラメトリック多重比較. 計算機統計学会, 8, 1, 83-89

Yoshida, M. (1988) Exact probablities assiociated with Tukey's and Dunnett's multiple comparisons procedures in imbalanced one-way ANOVA. Journal of the Japanese Society of Computional Statistics 1, 111-122

# 第 I 相臨床試験データの薬物濃度-QTc解析

渡橋 靖、○塩境 一仁

（第一三共株式会社 データサイエンス部）

Analysis of Concentration-QTc Models for Phase I Clinical Data

Yasushi Orihashi and Kazuhito Shiosakai

Biostatistics & Data Management Department, Daiichi Sankyo Co., Ltd.

## 要旨

ICH-E14 ガイドラインでは、一般的に「QT/QTc 評価試験 (thorough QT/QTc study)」（以下、TQT 試験）を実施し、薬物の催不整脈作用の潜在的可能性に関する臨床的評価を行うことが求められている。また、ICH E14 Q&A (R3) (2017) では、薬物濃度－QTc 解析が、TQT 試験の時点毎の解析を代替する可能性を述べている。TQT 試験では、QTc のベースラインからの変化量（ΔQTc）について、時点毎に被験薬とプラセボの差（ΔΔQTc）とその 95%片側信頼区間を計算し、全ての時点でその上限が 10 ms を下回れば、陰性と判定される。TQT 試験の代替として、第 I 相試験の QTc データを用いて薬物濃度－QTc 解析を行う場合、単に被験薬投与例のデータを用いて薬物濃度－QTc 解析を行うだけでは、QTc の日内変動を考慮していないため、薬物濃度データのないプラセボ投与例の QTc データを含めて日内変動で調整した薬物濃度と QTc の関係をモデル化し、TQT 試験のΔΔQTcに相当する評価を行う必要がある。薬物濃度－QTc 解析による陰性／陽性の判定は、臨床的に関連のある薬物への最大曝露量における QTc 延長を信頼区間で評価する。本発表では、最初に ICH-E14 ガイドラインと、その Q&A 作成の過程で行われた研究の概略をレビューする。次に、実例に基づく仮想データを用い、Garnett et al. (2016)のモデルを SAS で実行する方法、モデルや利用データによる結果の違い、および解析やデザイン上の課題や留意点などについて議論する。

キーワード：

ICH E14, QT 延長, concentration-QT model, proc mixed

## 1. ICH E14 ガイドライン

### 1.1 E14 ガイドライン

E14 ガイドラインでは、一般的に「QT/QTc 評価試験 (thorough QT/QTc study)」（以下 TQT 試験）を実施し、薬物の催不整脈作用の潜在的可能性に関する臨床的評価を行うことが求められている。TQT 試験の中心傾向の解析は、E14 ガイドラインでは「時間を一致させた被験薬群とプラセボ群の平均値の差（ベースライ

ン値による調整後）の収集の全期間を通じた最大値を用いて行われる」と記載されている。例えば、ベースライン値（並行群間比較試験では時間を一致させた投与前日の値（time-matched baseline）、クロスオーバー試験では投与日の投与直前の値(pre-dose baseline)）からの変化量（ΔQTc）について、時点毎に被験薬とプラセボの平均値の差（ΔΔQTc）とその信頼区間を計算し、95%片側信頼区間の上限が 10 ms を下回れば陰性と判定される。この解析方法は、薬物濃度情報を使わない機械的な評価であり、薬物濃度と QT 延長作用との間にヒステリシスの可能性がある場合や代謝物によって QT 延長作用の有無や程度が変わる場合などにも対応可能な頑健な方法である。しかし、複数時点のΔΔQTcの最大値を用いるため、例数、時点数、時点間の相関などに依存したバイアスが存在する保守的な評価法であることも知られている（日本製薬工業協会 2007）。E14 ガイドラインやその Q&A (R2)において、薬物濃度－QTc 間の評価は、用量選択のための情報、TQT 試験の解釈の明確化、QT 延長評価の信頼性向上など、総合的な評価の中で補助的な位置付けとされている。

### 1.2 Q&A (R3)

Q&A (R3)では、薬物濃度の QTc への影響を検討する薬物濃度－QTc 解析が、TQT 試験の時点ごとの解析 (intersection-union test) の代替となる可能性を述べている。但し、薬物濃度－QTc 解析だけで被験薬による QT 延長リスクを評価し、後期臨床試験における QTc 評価方法を判断するのではなく、非臨床データ、QT 延長の経時的変化、QT 延長の程度、外れ値のカテゴリカル解析、および潜在的な催不整脈作用のシグナルとなり得る特定の有害事象などを含む包括的な評価が必要である点も喚起されている。

薬物濃度－反応データのモデルには、異なる仮定に基づく、多種類があり、同一データに対して、モデル間で一致しない結果が得られる可能性があることが知られている。従って、モデル化の方法や仮定、モデル選択の基準、モデルの要素の根拠、および複数の試験で得られたデータを併合する可能性などを、解析に先立って事前規定し、バイアスを低減することが重要である。解析にあたっては、モデルの仮定、ヒステリシス、およびモデルの適合度などを検討し、その結果を文書化する必要がある。後期開発の患者対象の臨床試験において心電図を用いたより詳細な安全性評価が不要であると結論するためには、臨床的に意味のある最大の曝露状況において、薬物濃度－QTc 解析で推定された、薬物の QTc 間隔への作用の片側 95%信頼区間の上限が、10 ms 未満である必要がある。

## 2. 薬物濃度－QTc モデル

### 2.1 QT/QTc 評価試験での薬物濃度－QTc 解析

Garnett et al. (2008) によると、FDA が TQT 試験をレビューする際、薬物濃度－QTc 関係を検討するのがルーチンになっている。そのほとんどの場合において、ベースラインで調整し、時間を一致させた被験薬とプラセボの QTc の差、ΔΔQTcを反応変数とし、切片と薬物濃度に対する傾きを固定効果、および個人毎の切片と傾きを変量効果とする線形混合モデルが用いられている。Garnett et al. (2008) の Appendix B には薬物濃度と QTc の関係を検討するための SAS プログラムが載っており、変量効果の分散構造は unstructured が指定されている。

### 2.2 IQ-CSRC の研究

Innovation and Quality in Pharmaceutical Development (IQ) と Cardiac Safety Research Consortium (CSRC) の臨床薬理のメンバーが協力し、ヒトに初めて投与される単回投与試験で頑健で質の高い心電図の測定を行い、薬物濃度－QTc 解析を行うことで TQT 試験に置き換えることができるかどうかが研究された (Darpo et al. 2014)。QTc への影響がよく知られている 5 つの市販薬（偽陰性が重要なため、いずれとも規制当局が関心を

もつ基準値を超える QT 延長作用ありの薬（陽性の薬））と QT 延長作用なしの薬（陰性の薬）1 つを選び、20 名の健康成人に対し、プラセボ対照、ランダム化、3 期の不完備型デザインのクロスオーバー試験が行われ、6 つの実薬で各 9 名、プラセボ 6 名のデータが得られた。各期において Day 1 に低用量、Day 2 に高用量の投与が行われた。各実薬 9 名とプラセボ 6 名に対し、Fridericia 補正後 QTc の pre-dose baseline からの変化量ΔQTcFを反応変数とし、薬物濃度を連続量の共変量、投与群と投与後の計画された測定時点（プラセボで有意差なしの時点を統合）をカテゴリカルな固定効果とする線形混合モデルが当てはめられた。投与後の時点を固定効果とすることで、日内変動で調整された時点毎のプラセボとの差に近い推定値が得られることが過去の経験から分かっていた。線形混合モデルから QTc 延長の両側 90%信頼区間を求め、陽性の薬では、薬物濃度に対する傾きの信頼区間下限値が 0 より大きく、かつ、低用量投与での最高血中濃度（Cmax）幾何平均値での QTc 延長の信頼区間上限値が 10 ms を超えていること、陰性の薬では、高用量投与の Cmax 幾何平均値での QTc 延長の信頼区間上限値が 10 ms より下回ることを評価の基準とした。その結果、5 つ全ての陽性の薬で陽性、1 つの陰性の薬で陰性の判定結果となり、小規模な試験の薬物濃度－QTc 解析で、10 ms を超える QT 延長を検出でき、早期臨床試験の頑健で質の高い心電図データを用いた薬物濃度－QTc 解析が TQT 試験の代替となり得ることが分かった (Darpo et al. 2015)。

### 2.3 Garnett et al. (2016)のモデル

Garnett eta al. (2016)では以下の 4 つのモデルの検討を行っている。

Table. 2-1 線形 C-QTc モデルの推定と$\Delta\Delta QTc_{max}$の予測

| モデル名 | モデル | $\Delta\Delta QTc_{max}$の予測 |
|---|---|---|
| ΔΔQTc | $\Delta\Delta QTc_{Ikt} = \mu + W_k + (\theta + D_k)C_{Ikt} + \varepsilon_{Ikt}$ | $\Delta\Delta QTc_{max} = \mu + \theta\overline{C_{max}}$ |
| ΔQTc1 | $\Delta QTc_{Ikt} = \mu + W_k + (\theta + D_k)C_{Ikt} + \varepsilon_{Ikt}$ | $\Delta\Delta QTc_{max} = \theta\overline{C_{max}}$ |
| ΔQTc2 | $\Delta QTc_{Ikt} = \mu + W_k + p_t + (\theta + D_k)C_{Ikt} + \varepsilon_{Ikt}$ | $\Delta\Delta QTc_{max} = \theta\overline{C_{max}}$ |
| ΔQTc3 | $\Delta QTc_{Ikt} = \mu_I + W_k + p_t + (\theta + D_k)C_{Ikt} + \varepsilon_{Ikt}$ | $\Delta\Delta QTc_{max} = \mu_1 - \mu_0 + \theta\overline{C_{max}}$ |

処置$I = 0, 1$、被験者 k、時点tに対し、薬物濃度$C_{Ikt}$、固定効果：切片μ（共通または処置毎）、傾きθ、時点$p_t$、変量効果：切片$W_k$、傾き$D_k$、誤差項$\varepsilon_{Ikt}$（Garnett eta al. (2016) Table 1 を改変）

全てのモデルは個人毎の切片と傾きを変量効果として入れた線形混合モデルである。ΔQTc1モデルは時点効果を含めておらず、日内変動を無視した最も簡単なモデルであり、プラセボ効果は切片に含まれる。ΔQTc2モデルとΔQTc3モデルは日内変動のための時点効果が入っており、ΔQTc2モデルはプラセボと実薬で共通の切片、ΔQTc3モデルは別々の切片を用いている。ΔQTcを求める際、クロスオーバー試験では pre-dose baseline を、並行群間試験では time-matched baseline が用いられた。並行群間試験でΔΔQTcを計算する方法は記載されていないが、時点毎にプラセボのΔQTcの平均値を差し引くのが自然であると思われる。これらのモデルについて、QT 延長の有無やヒステリシスの有無などの 7 つの条件、被験者数が 12 例から 60 例などの様々な条件を変えてシミュレーションを行い、偽陰性の割合、検出力、バイアス、などの評価が行われた。その結果、ΔΔQTcモデルは並行群間試験で偽陰性が多くなり、3 つのΔQTcモデルの中ではΔQTc3モデルが最も好ましいという結果であった。

## 3. 薬物濃度－QTc 解析例

### 3.1 解析データ

本事例で用いる解析データは第一三共（株）内で実施したある治験を元に作成しているが、時点数、薬物濃度、及び QTc 等を加工している（患者に投与した時の最大曝露量の算術平均値での QTc が若干延長するよう加工）。なお、投与経路は経口投与であり、ヒステレシスは無いという前提である。

試験デザインは、健康成人を対象とした、単回投与用量漸増試験（被験者の重複なし）、コホートは計 7 コホート、および各コホートの被験者数は 8 名（実薬 6 名、プラセボ 2 名）である。投与前日、投与日ともに投与直前 0、投与後 0.5、1、2、3、4、6、12 時間に薬物濃度、及び 1 分程度の間隔で 3 回心電図を測定しその平均を取る triplicate ECG が得られている。

本事例で示す解析データセット「cqtc」を Table.3-1 に示す（代表として 2 被験者分）。各々の変数は「subjid：被験者番号」、「cohort：漸増コホート、1～7」、「trt：実薬/プラセボ、0 プラセボ、1 実薬」、「dosage：0 プラセボ、1（最小用量）～7（最大用量）」「time：時刻（hr）」、「conc：薬物濃度」、「base：QTc の投与前日の値（ベースライン値）」、「aval：QTc の投与後の値」、および「chg：QTc の投与前日から投与日の同時刻間の変化量（Time-matched baseline）」を表す。

Table.3-1 解析データ「cqtc」（2 被験者分）

| Subjid | cohort | Trt | dosage | time | conc | base | Aval | chg |
|---|---|---|---|---|---|---|---|---|
| 101 | 1 | 1 | 1 | 0 | 0.000 | 415.667 | 418.000 | 2.3333 |
| 101 | 1 | 1 | 1 | 0.5 | 1.980 | 412.000 | 403.667 | -8.3268 |
| … | … | … | … | … | …… | ……. | ……. | ……. |
| 102 | 1 | 0 | 0 | 0 | 0.000 | 411.000 | 417.333 | 6.3333 |
| 102 | 1 | 0 | 0 | 0.5 | 0.000 | 430.667 | 405.667 | -25.0000 |
| … | … | … | … | … | …… | ……. | ……. | ……. |

用量毎の QTc（算術平均値）と時刻の関係、及び QTc 変化量と薬物濃度の関係はそれぞれ、Fig.3-1、及び Fig.3-2 のとおりである。

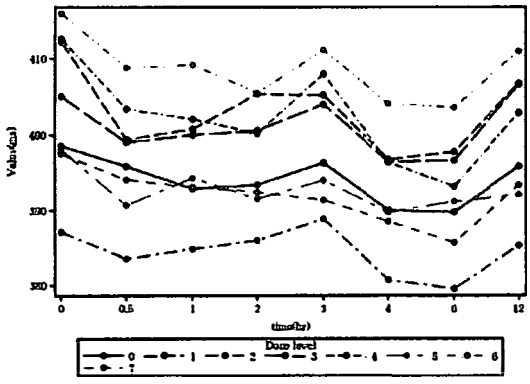



Fig.3-1 用量毎の QTc（算術平均値）と投与後時間の関係

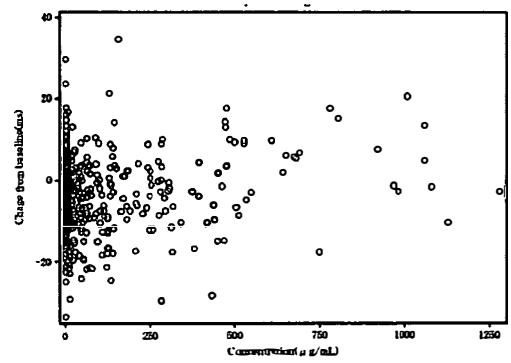


Fig.3-2 Δ QTc と薬物濃度の関係（散布図）

## 3.2 薬物濃度－QTc モデルと SAS プログラム

ΔQTcのモデルにおいてプラセボ群の QTc データを含めて解析するとき、プラセボ群の薬物濃度は測定されておらず、欠測である場合もあるため、例えば、ΔQTc2モデルのプラセボ群は $\Delta QTc_{lkt} = \mu + W_k + p_t + \varepsilon_{lkt}$

と記載すべきかもしれない。プラセボか実薬かでモデル式が異なる場合、PROC NLMIXED で if 文を用いた記述ができる。一方、プラセボ群の薬物濃度$C_{Ikt}$を 0 とする解析データセットを作成すれば、モデル式上は薬物濃度の項がなくなり、同じになる。このやり方は、固定効果や変量効果の推定に影響がないであろうか？そこで、(1) プラセボ群のモデル式から薬物濃度の項を削除する方法、および (2) プラセボ群も同じモデル式を用い薬物濃度を 0 とする方法、の 2 通りの方法を考える。線形混合モデルのモデル式 $y = X\beta + Z\gamma + \varepsilon$ において、固定効果の X は、方法 (1) についてもプラセボ群の薬物濃度を 0 とするしかなく、その影響については$\Delta$QTc2モデルと$\Delta$QTc3モデルの違いで論じるべきと思われる。一方、変量効果の項については、方法 (1) がプラセボ群の random slope がないモデル、方法 (2) がプラセボ群の random slope があるモデルとなり、Zと$\gamma$のコーディング方法が異なる。しかし、方法 (1) および(2)、それぞれにおいて、yの分散を計算すると、いずれも同じ分散を与えることが分かる。従って、対数尤度関数や固定効果の推定値は方法 (1) と (2) との間で一致する。変量効果については、方法 (1) ではそもそもプラセボ群の random slope がない。一方、方法 (2) ではプラセボ群の random slope が推定される。Random intercept と random slope の間に相関がなければ、プラセボ群の random slope の推定値は 0 となるが、相関がある場合はプラセボ群の random slope として 0 でない値も推定される。以上の考察から、本報告ではプラセボ群の薬物濃度を 0 とし、プラセボ群と実薬群を区別せず、PROC MIXED で同一 model ステートメント内で解析する。

PROC MIXED では、RANDOM Statement で変量効果の共分散構造を指定する。本報告で用いた解析データの場合、TYPE=UN (unstructured) とすると"Estimated G matrix is not positive definite"という NOTE が出たり、収束が不安定になったりする現象が見られ、random intercept と random slope の相関が高い傾向が認められた。従って、本報告では TYPE=VC (Variance Components) を用い、EMPIRICAL option で標準誤差を計算することとした。

## 3.3 解析結果

以下に SAS コード及び解析結果を示す。解析には SAS® System Release 9.2 を用いた。なお、本事例においては、別試験の患者対象試験の成績を参考に、患者に投与した時の最大曝露量の算術平均値を 900 μg/mL と想定している。

### 3.3.1 4 つのモデルでの最大曝露量での ΔΔQTc の違い

Garnett et al. (2016)の 4 つのモデルに対応する SAS コードを示す。

初めに$\Delta$QTc3から示す。

```
proc mixed data=cqtc empirical;
  class time subjid;
  model chg = trt time conc/solution cl alpha=0.1;
  random int conc/type=vc subject=subjid solution;
  estimate "ΔΔQTc at Cmax" trt 1 conc 900/cl alpha=0.1;
  title "ΔQTc3 model";
run;
```

引き続き$\Delta$QTc2であるが$\Delta$QTc3から以下の変更を行った SAS コードである。

```
model chg = trt time conc/solution cl alpha=0.1;
  →model chg = time conc/solution cl alpha=0.1;
estimate "ΔΔQTc at Cmax" trt 1 conc 900/cl alpha=0.1;
```

→ estimate "ΔΔQTc at Cmax" conc 900/cl alpha=0.1;

更にΔQTc1はΔQTc2から以下の変更を行った SAS コードである。

model chg = time conc/solution cl alpha=0.1;

→model chg = conc/solution cl alpha=0.1;

最後にΔΔQTcはΔQTc1から以下の変更を行った SAS コードである。ただし、chg2 は時刻毎にプラセボ群の算術平均値を引いた値を表す。

model chg = conc/solution cl alpha=0.1;

→□model chg2 = conc/solution cl alpha=0.1;

estimate "ΔΔQTc at Cmax" conc 900/cl alpha=0.1;

→ estimate "ΔΔQTc at Cmax" int 1 conc 900/cl alpha=0.1;

4 つのモデルでの conc=900 におけるΔΔQTcmaxの予測値は Table.3-2 のとおりである。Garnett et al. (2016) にて最も推奨されているΔQTc3との比較においてΔQTc1は 3 ms 以上の違いがある。また、90%両側信頼区間上限値より、ΔQTc1モデルのみが 10 ms の延長を見逃す可能性がある。

Table.3-2 ΔΔQTcmaxの予測値（90%両側信頼区間）

| モデル名 | $\Delta\Delta QTc_{max}$の予測値（ms） |
|---|---|
| ΔΔQTc | 5.2（-0.7, 11.1） |
| ΔQTc1 | 2.2（-3.8, 8.2） |
| ΔQTc2 | 5.7（-0.6, 12.1） |
| ΔQTc3 | 5.2（-1.2, 11.7） |

3.3.2 プラセボ群のデータを使わなかった場合

濃度データが無いため、解析にプラセボ群のデータを使わないという考え方もあろう。そこで、プラセボ群が無くとも解析可能なΔQTc1、及びΔQTc2に対してプラセボ群のデータを使わない解析を実施した。SAS コード上は「where trt=1;」を追加している。結果は Table.3-3 のとおりであり、ΔQTc1モデル、及びΔQTc2モデルともに、プラセボデータを使わない場合がプラセボデータを使う場合に比べ、点推定値で 1～2ms のズレが発生した。

Table.3-3 ΔΔQTcmaxの予測値（ms）（90%両側信頼区間）

| モデル名 | プラセボデータを使った場合 | プラセボデータを使わない場合 |
|---|---|---|
| ΔQTc1 | 2.2（-3.8, 8.2） | 3.7（-2.2, 9.5） |
| ΔQTc2 | 5.7（-0.6, 12.1） | 7.7（1.3, 14.0） |

3.3.3 投与前日の QTc を取得しない場合

単回投与用量漸増試験では投与前日の心電図データを取得しない場面が考えられる。そこで Garnett et al. (2016) にて最も良いとされたΔQTc3に対して、投与前日の time-matched baseline を使わず、pre-dose baseline を用いる場合の解析を実施した。解析ではΔQTc3から以下の変更を行った SAS コードを用いた。ただし、chg3 は時刻毎に投与日の時刻 0（time=0）の値を差し引いた値である。

model chg = trt time conc/solution cl alpha=0.1;

→model chg3 = trt time conc/solution cl alpha=0.1;

where time^=0;を追加

結果はTable.3-4のとおりであり、投与前日のtime-matched baselineの使用の有無で結果に大きな違いが出た。

Table.3-4 ΔΔQTcmaxの予測値（ms）（90%両側信頼区間）

| モデル名 | time-matched baselineを使った場合 | pre-dose baselineを使った場合 |
|---|---|---|
| ΔQTc3 | 5.2（-1.2, 11.7） | -4.2（-9.5, 1.2） |

## 3.4 影響度分析

Garnett et al. (2016) にて最も良いとされているΔQTc3に対して、各投与群の固定効果への影響を確認する目的で、被験者1名を除いたときの固定効果（被験者効果）を全被験者分算出し、全被験者データを用いた固定効果（全効果）との違いを分析した。以下にSASコードを示す。なお、Mixedプロシジャにおいて、本分析はmodelステートメントのinfluenceオプションで指定可能である。

```
/*フォーマットの指定：時刻0のデータをReferenceに持ってくるため*/
proc format;
  value ref 0="_0"
        ;
proc mixed data=cqtc empirical order=formatted;
  class time subjid;
  model chg = trt time conc/solution cl alpha=0.1 influence(effect=subjid iter=10 estimates);
  random int conc/type=vc subject=subjid solution;
  format time ref.;
run;
```

上記コーディングにより各被験者の被験者効果が得られる。Fig.3-3～3-7に被験者効果と全効果の差（効果の差）と用量の関係（箱ひげ図）、及びΔΔQTcmaxの推定値の差と用量の関係を示す。なお、参照線は差分0を表し、参照線から離れれば離れるほど当該被験者の情報が固定効果推定への影響が大きいことを示す。

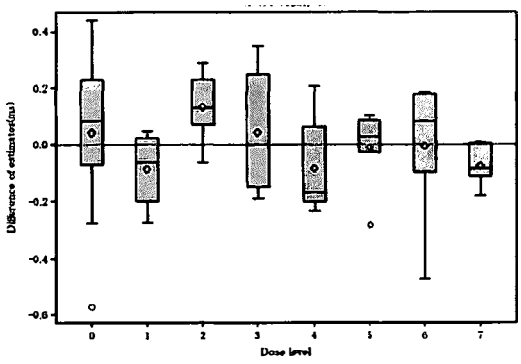



Fig. 3-3 切片の効果の差と用量の関係（箱ひげ図）

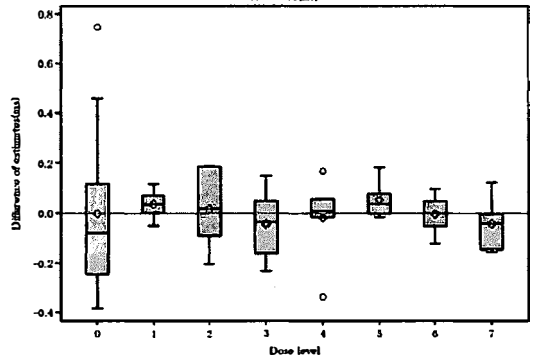



Fig. 3-4 実薬/プラセボの効果の差と用量の関係（箱ひげ図）

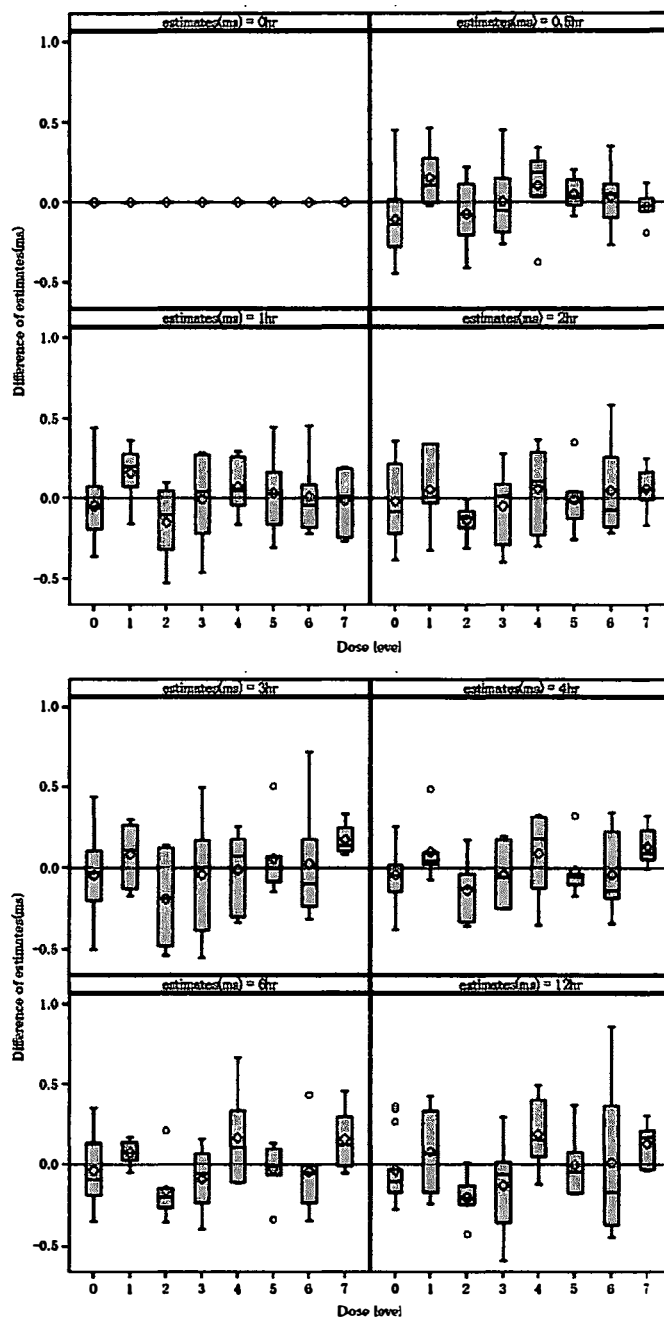



Fig. 3-5 各観測（投与後）時点における効果の差と用量との関係（箱ひげ図）

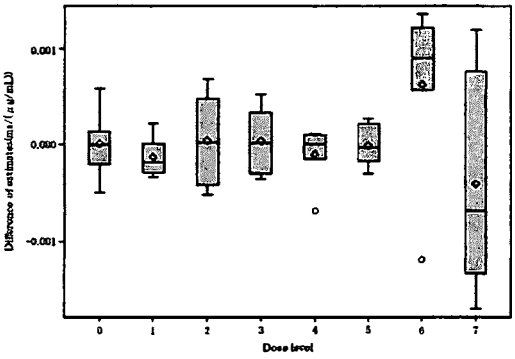



Fig. 3-6 薬物濃度の効果の差と用量の関係（箱ひげ図）

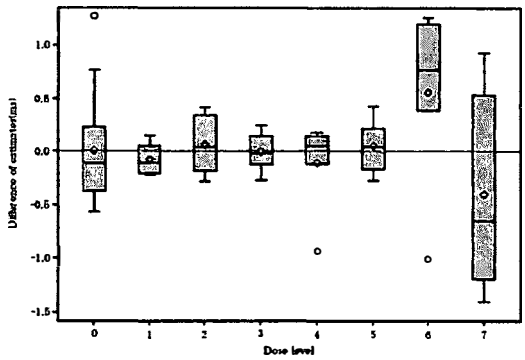



Fig. 3-7 ΔΔQTcmaxの推定値の差と用量の関係（箱ひげ図）

以上より、各固定効果、及びΔΔQTcmaxの推定にはプラセボ群の影響が大きいことが示唆された。また、薬物濃度に依存する効果の推定においては、高用量群のデータの寄与が大きいことが示唆された。

## 4. 考察

本論文では、既存の治験データを元に加工したデータに対して、Garnett et al. (2016)に記載された複数の解析を実施した。解析結果で示したように、どのモデルを用いるか、どのデータを用いるかによって結果が変わる。特に、時点効果を含めないΔQTc1モデルの結果が大きく異なり、起床時間、食事時間、活動性などの1日の時刻と関連する日内変動を無視することができないことが改めて認識された。また、薬物濃度に対する傾きの推定には高用量群のデータの影響が大きいが、日内変動を調整する時点効果の推定にはプラセボや低用量群のデータも影響し、プラセボデータを除くと結果が変わるため、プラセボを含む全てのデータを用いて薬物濃度-QTc モデルの解析を行うべきであろう。本論文で用いたデータでは、ΔΔQTc、ΔQTc2、およびΔQTc3、各モデルを用いた解析結果は点推定値、90%信頼上限ともに大きく変わらなかったが、どのモデルを用いた解析を主解析とするか、あるいは複数のモデルを当てはめて、どのような基準で使用するモデルを選択するかは、事前に規定する必要がある。変量効果の共分散構造については、本論文では Variance Components (VC) を用いたが、他のデータでは Unstructured (UN) を用いることができるかもしれないので、様々なデータで更なる検討が必要であろう。また、本論文では述べなかったが、主解析として用いるべき QT 間隔補正法（Fridericia 補正法、試験別補正法、または個体別補正法）、モデルの妥当性、特にヒステリシスがあるかどうかの視覚的な確認は必須である。

本論文で用いたデータは、1 用量あたり実薬 6 名、プラセボ 2 名の単回投与試験である。Darpo et al. (2014, 2015) の研究は、9 名のデータから 5 つの陽性の薬と 1 つの陰性の薬を正しく判定できることを示したが、本論文の例では、単回投与試験の最高用量と、臨床的に意味のある最大の曝露状況が近かったため、最高用量の 6 名だけでは、例数が不足であったかもしれない。その場合、最高用量だけ例数を増やすよう計画する、反復投与試験など他の試験とデータを併合した解析を計画するなどが必要であろう。

E14 Q&A では、TQT 試験が並行群間比較試験であれば time-matched baseline を用いるべきであり、クロスオーバー試験なら pre-dose baseline も適切であるとしている。第 I 相試験の多くは、プラセボや用量が並行群間のデザインで行われるので、薬物濃度-QTc モデルの解析を行うためには、投与前日に 1 日心電図測定を行

い、time-matched baseline を利用可能にすべきであろう。本論文で用いた例では、time-matched baseline を用いた薬物濃度-QTc モデルの解析結果と pre-dose baseline を用いた解析結果が大きく異なった。これは、time-matched baseline が適切なΔQTcモデルに pre-dose baseline のΔQTcを入れただけであるからとも考えられる。第 I 相試験の薬物濃度-QTc 解析に pre-dose baseline を用いることができれば、試験実施上のメリットが大きいので、pre-dose baseline が適した別のモデルを開発するなど、更なる研究が望まれる。

規制当局の一番の懸念は偽陰性である。TQT 試験では、偽陰性でないことを保証するために陽性対照が用いられる。第 I 相試験の場合、例えば、各コホートに QT 延長の陽性対照を 1 例ずつ設けることが考えられるが、その必要性については広く議論すべきであろう。E14 Q&A (R3) では、治療において予想される最大曝露量の相当倍の状況での反応を特徴づけるデータがあることが、陽性対照が不要となる必要条件とされている。今後多くの例で経験を重ね、十分議論するとともに、意思決定の頑健性の観点から薬物濃度-QTc 解析を考察することが必要である。

薬物濃度-QTc 解析は、薬物濃度解析担当者や pharmacometrician が担当している会社が多いのではないかと思われる。しかし、モデル構築や解析の段階で統計的に考察すべきことも少なくない。今後、様々な実例で検討を重ね、ベストなアプローチを探って行くためにも、より多くの統計解析担当者が参画し、広く議論すべきであろう。

## 参考文献


[1] ICH E14 「非抗不整脈薬における QT/QTc 間隔の延長と催不整脈作用の潜在的可能性に関する臨床的評価」 薬食審査発 1023 第 1 号 2009 年 10 月 23 日

[2] 日本製薬工業協会 医薬品評価委員会 統計・DM 部会 「QT 延長の統計解析に関する解説書」 医薬出版センター, 2007 年 5 月

[3] ICH E14 Q&A (R2) 「非抗不整脈薬における QT/QTc 間隔の延長と催不整脈作用の潜在的可能性に関する臨床的評価」に関するＱ＆Ａ 厚生労働省 医薬食品局審査管理課 事務連絡 2015 年 7 月 10 日

[4] ICH E14 Q&A (R3) 「非抗不整脈薬における QT/QTc 間隔の延長と催不整脈作用の潜在的可能性に関する臨床的評価」に関するＱ＆Ａ 厚生労働省 医薬・生活衛生局医薬品審査管理課 事務連絡 2017 年 5 月 23 日

[5] C. Garnett et al. “Concentration-QT Relationships Play a Key Role in the Evaluation of Proarrhythmic Risk During Regulatory Review” *J Clin Pharmacol* 2008;48:13-18

[6] B Darpo et al. ‘The IQ-CSRC Prospective Clinical Phase 1 Study: “Can Early QT Assessment Using Exposure Response Analysis Replace the Thorough QT Study?”’ *Ann Noninvasive Electrocardiol* 2014;19(1):70-81

[7] B Darpo et al. “Results From the IQ-CSRC Prospective Study Support Replacement of the Thorough QT Study by QT Assessment in the Early Clinical Phase” *Clinical Pharmacology & Therapeutics* 2015, 97(4), 326-335

[8] C Garnett et al. “Operational Characteristics of Linear Concentration-QT Models for Assessing QTc Interval in the Thorough QT and Phase I Clinical Studies” *Clinical Pharmacology & Therapeutics* 2016, 100(2), 170-178

# SASによる臨床試験の解析速報に用いる
# プレゼンテーションスライドの効率的な作成方法の提案

○吉田 直記、山崎 文寛、舟尾 暢男、高浪 洋平

（武田薬品工業株式会社）

# Increasing Efficiency in Creating Flash Report Presentation Slides in Clinical Trials using SAS

Naoki Yoshida, Fumihiro Yamasaki, Nobuo Funao, Yohei Takanami
Takeda Pharmaceutical Company Limited

## 要旨

臨床試験の結果を解析速報として報告する際の PowerPoint 形式のプレゼンテーションスライドを、SAS のマクロを用いて効率的に作成する方法を提案する。本マクロでは PowerPoint ファイルへの出力に ODS POWERPOINT、レイアウトの調整に ODS LAYOUT を用いる。なお、解析ソフト「R」も PowerPoint ファイルの作成が可能なため、その機能についても触れる。

キーワード：解析速報 (Flash Report), ODS POWERPOINT, ODS LAYOUT, R

## 1 はじめに

医薬品の臨床試験においては、試験終了後に作成される正式な解析報告書とは別に、当該医薬品の開発継続可否に関する意思決定を即座に行う目的で、解析速報の作成を求められることが多い。この解析速報は、当該臨床試験に携わった関係者に加えて、直接臨床試験に関わっていない第三者（医療関係者、各専門家、スポンサーの上層部等）に対して試験結果を報告するために作成されることから、内容には明瞭さが求められる。また、解析速報を基に意思決定を行う者の中には統計専門家以外の者も少なくないことから、試験結果の表示や解説は、統計の専門用語をふんだんに用いて説明された文章形式よりも、簡潔かつ見栄え良く示された図表形式の方が都合が良い。さらに、意思決定は可能な限り迅速に行う必要があることを鑑みると、解析速報は報告書形式よりも PowerPoint 形式で作成した方が、作成時間や閲覧時間を短縮することが出来、意思決定に関する会議の運営が容易となる。

さて、解析速報を作成する手順は、臨床試験に関するデータが出揃った時点で解析報告書のための図表を全て出力した後、必要となる図表を同定して解析速報（プレゼンテーションスライド）にマニュアルで転記・貼付を行う作業を繰り返すのが一般的であろうが、限られた時間の中で解析速報を作成する目的からは逆行している。すなわち、マニュアルでの図表の転記・貼付は決して効率的な方法とはいえず、作業中の間違いも起こりやすいため、作成後の解析速報を確認する作業が必須となり、さらなる時間の浪費に繋がる。

本稿では、胃潰瘍再発予防に関する架空の臨床試験を想定し、試験終了後に解析速報（プレゼンテーションスライド）を効率良く作成する方法を提案する。解析速報に出力する図表は、一般的な解析項目である被験者背景項目、主要評価項目、副次評価項目及び TEAE（Treatment Emergent Adverse Event：有害事象）に関するものとし、使用するサンプルデータは CDISC/ADaM 形式（簡単のため ADSL、ADTTE 及び ADAE のみ使用）にて作成しているため、本稿での提案方法を容易に実臨床試験に適用することが出来る。また、本稿で紹介する SAS マクロでは、PowerPoint ファイルへの出力は SAS 9. 4M1 から新たに追加された ODS POWERPOINT を用い、スライドのレイアウト調整には SAS 9. 4 M3から利用できる ODS LAYOUT を用いているが、SAS マクロの使用方法は本稿にて詳述しており、SAS マクロの使用時に読者が困ることが無いよう留意した。さらに、本稿の提案方法にて解析速報を作成すれば、マニュアルで作成する方法に比べて間違いが大幅に軽減され、試験終了後の作成時間の劇的な短縮が期待できる。なお、PowerPoint ファイルへの図表の出力は統計解析ソフト「R」でも実施可能である。本稿の後半では SAS と R による機能の違い、具体的には作業効率やスライドテンプレートの適用可否の観点から考察を行い、本稿の提案方法（すなわち、SAS の現行の機能）の限界と今後の展望を述べる。

## 2 ODS POWERPOINT、及び ODS LAYOUT の概要

### 2.1 ODS POWERPOINT

ODS POWERPOINT は、SAS で作成した表やグラフを簡単な構文で PowerPoint ファイルへ出力することを実現する。他の ODS ステートメントと組み合わせることで、テキストやリストをスライド上の任意の位置に出力することや、タイトルスライドからコンテンツスライドまで、プレゼンテーションスライドの大部分を SAS プログラムのみで作成することを可能にしている。また、ユーザー定義のスタイルテンプレートや style オプションを使用することで、背景・フォント・罫線等のスライドのデザインの詳細なカスタマイズが可能となる。ODS POWERPOINT を使用するには、図 2. 1. 1に示すように、PowerPoint ファイルへレポート出力するプロシジャの構文を ODS POWERPOINT ステートメントで囲んで記述する。

図 2. 1. 1 ODS POWERPOINT の基本構文

```
ods powerpoint file="example.pptx" option(s);
  <レポート出力するプロシジャの構文>
ods powerpoint close;
```

ODS POWERPOINT における主なオプションを表 2. 1. 1 に示す。

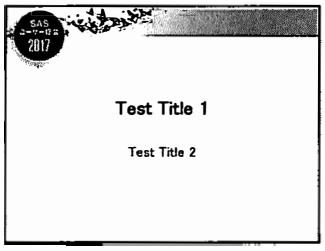
表 2. 1. 1 ODS POWERPOINT の主なオプション

| 指定方法 | 概要 |
|---|---|
| file="ファイル名" | 出力する PowerPoint ファイルを指定 |
| options(backgroundimage="ファイル名. pptx") | スライドの背景画像ファイルを指定<br>（GIF, JPEG 及び PNG 等が指定可能） |
| style=*style-element-name* \| *[style-attribute-name]* | スタイルテンプレート、スタイル属性を指定 |
| layout= | スライドのレイアウトを指定 |
| layout=TitleSlide | タイトルスライド |
| layout=TitleAndContent | タイトルとコンテンツ |
| layout=TwoContent | タイトルと 2 つのコンテンツ |

| 指定方法 | 概要 |
|---|---|
| gtitle / nogtitle | title ステートメントの内容をグラフに出力する/しないを指定（デフォルトは gtitle） |
| gfootnote / nogfootnote | footnote ステートメントの内容をグラフに出力する/しないを指定（デフォルトは gfootnote） |

PowerPoint ファイルの作成は、その用途に応じてレイアウト等がカスタマイズされたテンプレートを用いて作成することが多い。SAS 9.4 M3 では既存の PowerPoint ファイル（.pptx 又は.ppt）や PowerPoint テンプレート（.potx 又は.pot）を利用することはできないが、画像ファイルをスライドの背景として利用することで、同様の PowerPoint ファイルを作成することが可能である。具体的には、利用したいスライドをスクリーンショットによって画像化し、backgroundimage オプションで背景画像として指定すればよい。スライドの背景画像を指定する SAS プログラム例及び実行結果を図 2.1.2 に示す。

図 2.1.2　背景画像及び layout オプションを指定する SAS プログラム及び実行結果の例



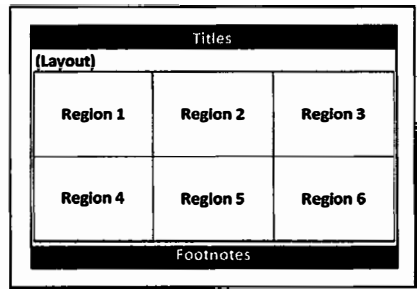

```
ods powerpoint file="Example1.pptx"
options(backgroundimage="C:¥Temp¥SUGJ2017_TitleImage.png")
layout=titleslide ;
proc odstext ;
   p "Test Title 1" / style=PresentationTitle ;
   p "Test Title 2" / style=PresentationTitle2 ;
run ;
ods powerpoint close ;
```

## 2.2 ODS LAYOUT GRIDDED、ODS REGION

PowerPoint ファイルにおけるスライドのレイアウトについて、より細かなカスタマイズが必要な場合は、ODS LAYOUT を用いることによってテキストや画像等のオブジェクトを任意の場所に出力することが可能となる。ODS LAYOUT には、ODS LAYOUT ABSOLUTE による絶対指定と、ODS LAYOUT GRIDDED によるグリッド指定の 2 種類の指定方法があるが、ODS POWERPOINT では ODS LAYOUT GRIDDED のみがサポートされている。ODS LAYOUT GRIDDED ではスライドが仮想的なグリッド（格子状）に分割され、ユーザーは ODS REGION を用いてオブジェクトを出力するグリッド上の番地（Region Container）を指定する。例えば、スライドのレイアウトを 6 分割して、それぞれにオブジェクトを配置することが可能である。ODS LAYOUT GRIDDED を使用するには図 2.2.1 に示すように、ODS REGION ステートメントとその番地に解析図表を出力するプロシジャの構文を、ODS LAYOUT GRIDDED ステートメントで囲んで記述する。

図 2.2.1　ODS LAYOUT GRIDDED の基本構文とレイアウトの分割イメージ（2 行×3 列）

```
ods layout gridded option(s) ;
   ods region option(s) ;
   <レポート出力するプロシジャの構文>
ods layout end ;
```



ODS LAYOUT における主なオプションを表 2.2.1 に示す。

表 2.2.1　ODS LAYOUT における主なオプション

| 指定方法 | 概要 |
|---|---|
| columns = *number* | グリッド全体を分割する列数を指定 |
| rows = *number* | グリッド全体を分割する行数を指定 |
| column_gutter = *number* | グリッドの列間の幅を指定 |
| row_gutter = *number* | グリッドの行間の幅を指定 |
| x = *dimension* | グリッド全体の横方向の開始位置を指定。x=0 は左端<br>（単位は mm、cm、in（インチ）、pct（%）等が利用可能） |
| y = *dimension* | グリッド全体の縦方向の開始位置を指定。y=0 は上端<br>（単位は mm、cm、in（インチ）、pct（%）等が利用可能） |
| order_type = row_major \|<br>column_ major | row_major　：行のすべての Regi on が埋まれば次の行へ移動<br>column_ major 列のすべての Regi on が埋まれば次の列へ移動<br>（ディフォルトは row_maj or） |
| advance = bygroup | bygroup 内のそれぞれについて、各 regi on に出力 |

ODS REGION の主なオプションを表 2.2.2 に示す。

表 2.2.2　ODS REGION における主なオプション

| 指定方法 | 概要 |
|---|---|
| column = *number* | グリッド上の列番号を指定 |
| row = *number* | グリッド上の行番号を指定 |

ODS POWERPOINT layout オプションで指定可能なレイアウトは、2.1 項で述べた 3 種類に限定されるが、ODS LAYOUT を使用すれば任意のレイアウトでスライドを作成することが可能である。いま、本 SAS ユーザー総会の発表者用 PowerPoi nt テンプレートと同様の外観となるような出力方法を検討する。ODS LAYOUT を用いて、タイトルスライドを作成する SAS プログラムを図 2.2.3、実行結果を図 2.2.2 に示す。図 2.2.3 に示した SAS プログラムでは、文字列の出力位置について、縦方向の出力開始位置を ODS LAYOUT GRIDDED の y オプションで指定し、横方向の出力位置を opti onsenter によって中央揃えにした。なお、ODS POWERPOINT の l ayoutオプションと ODS LAYOUT は、同一スライドにおいてはいずれか一方のみが使用可能である。

図 2.2.2　ODS LAYOUT によるタイトルスライド作成プログラムの実行結果とレイアウトイメージ

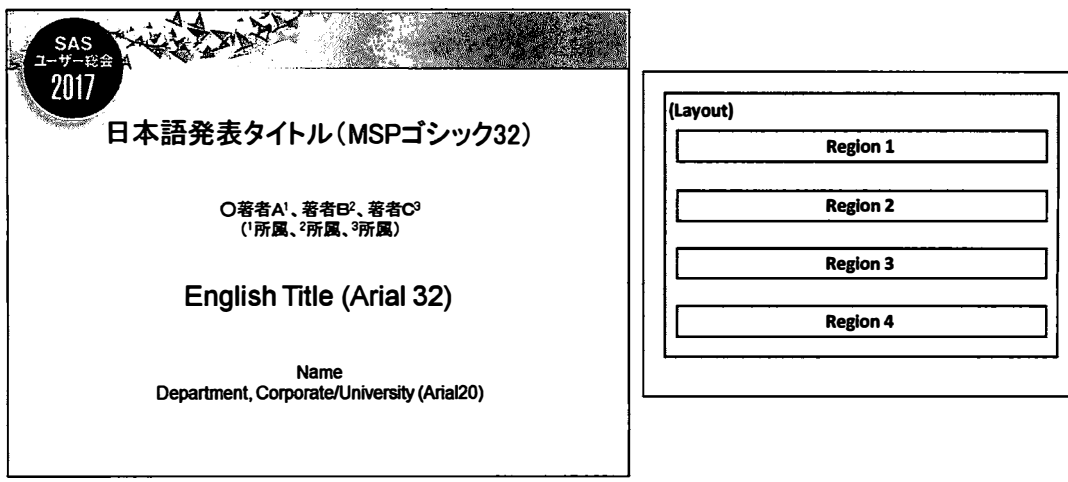
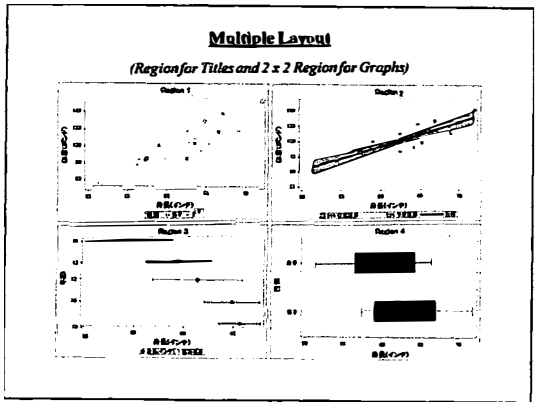


図 2.2.3　ODS LAYOUT によるタイトルスライド作成プログラム例

```
options center ; /* 出力を中央揃え */
ods escapechar="^" ; /* 上付き、改行などの特殊文字を出力 */
ods powerpoint file='Example2.pptx' ;
ods powerpoint options(backgroundimage="C:¥Temp¥SUGJ2017_TitleImage.png") ;
ods layout gridded y=20pct ; /* 出力位置を指定 */
ods region ;
proc odstext ;
  p "日本語発表タイトル（MSP ゴシック 32）" / style=[fontfamily="ＭＳ　Ｐゴシック" fontsize=32pt] ;
run ;
ods layout end ;
ods layout gridded y=40pct ;
ods region ;
proc odstext ;
  p "x 著者Ａ^{super 1}、著者Ｂ^{super 2}、著者Ｃ^{super 3}^{newline 1}（^{super 1}所属、^{super 2}
所属、^{super 3}所属）" / style=[fontfamily="ＭＳ　Ｐゴシック" fontsize=20pt just=center] ;
run ;
ods layout end ;
ods layout gridded y=60pct ;
ods region ;
proc odstext ;
  p "English Title (Arial 32)" / style=[fontfamily="Arial" fontsize=32pt] ;
run ;
ods layout end ;
ods layout gridded y=80pct ;
ods region ;
proc odstext ;
  p "Name^{newline 1}Department, Corporate/University (Arial20)" / style=[fontfamily="Arial"
fontsize=20pt just=center] ;
run ;
ods layout end ;
ods powerpoint close ;
```

次に、解析図表を出力するためのタイトルとコンテンツから構成されるスライドを作成する。SAS では同一スライド内に複数のグリッドを作成することで実現可能であり、そのプログラムを図 2.2.5、実行結果を図 2.2.4 に示す。図 2.2.5 に示した SAS プログラムでは、スライドにおけるタイトルの出力先として 1×1 のグリッドを、グラフの出力先として 2×2 のグリッドを作成した。

図 2.2.4　同一スライドに複数のグリッドを作成する
プログラムの実行結果及びレイアウトのイメージ



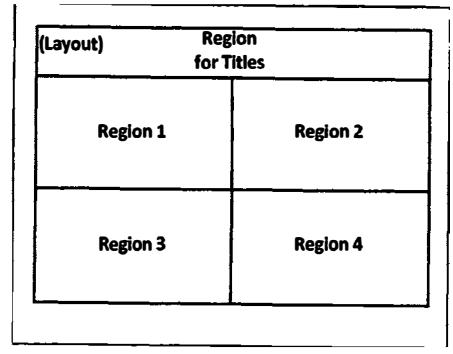



図 2.2.5　同一スライドに複数のグリッドを作成するプログラムの例

```
options center ; /* 出力を中央揃え */
ods powerpoint file='Example3.pptx' ;
ods layout gridded columns=1 rows=1 x=0pct ;
proc odstext ;
  p "Multiple Layout" / style=[fontfamily="Times New Roman" fontsize=24pt fontweight=bold
textdecoration=underline just=center] ;
  p "(Region for Titles and 2 x 2 Region for Graphs)" / style=[fontfamily="Times New Roman"
fontsize=20pt fontstyle=italic color=blue just=center] ;
run ;
ods layout end ;
ods layout gridded columns=2 rows=2 column_gutter=4pt row_gutter=1pt ;
ods graphics / reset width=4in height=2.6in ;
ods region ;
title1 "Region 1" ;
proc sgplot data=sashelp.class ;
  scatter x=height y=weight / group=sex ;
run ;
ods region ;
title1 "Region 2" ;
proc sgplot data=sashelp.class ;
  reg x=height y=weight / CLM CLI ;
run ;
ods region ;
title1 "Region 3" ;
proc sgplot data=sashelp.class(where=(.<age<16)) ;
  dot age / response=height stat=mean limitstat=stddev numstd=1 ;
run;
ods region ;
title1 "Region 4" ;
proc sgplot data=sashelp.class ;
  hbox height / category=sex ;
run ;
ods layout end ;
ods powerpoint close ;
```

## 2.3 PROC ODS TEXT

プレゼンテーションスライドでは、解析図表に加えてスライドのタイトルや図表に示す解析結果の解釈などを出力する必要がある。テキストの出力には、ODS TEXT プロシジャが使用できる。図 2.3.1 に基本構文を示す。

図 2.3.1 PROC ODSTEXT の基本構文

```
proc odstext ;
  p "output text" / option(s) ;
  list / option(s) ;
    item "list text" / options(s) ;
  end ;
run ;
```

| | |
|---|---|
| p | ：指定したテキストを出力 |
| list | ：箇条書き形式の開始位置を指定 |
| item | ：箇条書きで指定したテキストを出力<br>(list ～ end 内のみ指定可能) |

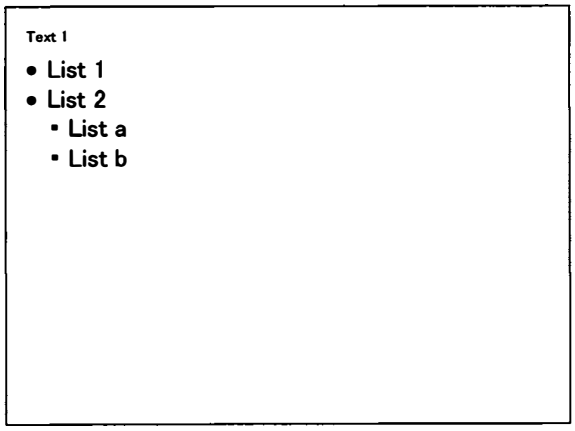
表 2.3.1 に主なオプションを示す。

表 2.3.1　PROC ODSTEXT で用いる主なオプション

| 指定方法 | 概要 |
|---|---|
| style=*style-element-name* \| *style-attribute-name]* | スタイルテンプレート、スタイル属性を指定 |
| style=[fontfamily="フォント名"] | フォントを指定 |
| style=[fontsize=*number* pt] | フォントサイズを指定 |
| style=[fontweight=bold] | 文字を bold（太字）に変更 |
| style=[fontstyle=italic] | 文字を italic（斜体）に変更 |
| style=[color=*color name*] | 文字の色を指定 |
| style=[backgroundcolor=*color name*] | 文字の背景色を指定 |
| style=[textdecoration=underline] | 下線 を指定 |
| style=[just=left \| center \| right] | テキストボックスの左揃え、中央揃え、右揃えを指定 |
| style=[bullet=disc \| circle \| square \| none] | 箇条書きの記号を指定（disc：黒丸） |

PROC ODSTEXT のプログラム例と実行結果を図 2.3.2 に示す。

図 2.3.2 PROC ODSTEXT プログラム例と実行結果




```
ods powerpoint file='Example4.pptx' ;
proc odstext ;
   p 'Text 1' ;
   list ;
      item 'List 1' ;
      item ;
         p 'List 2' ;
            list / style=[bullet=square] ;
               item 'List a' ;
               item 'List b' ;
            end ;
      end ;
   end ;
run ;
ods powerpoint close ;
```

## 2.4 出力図表のサイズ、スライドのサイズ

解析図表のサイズを調整せずに PowerPoint ファイルを作成すると、各スライドに出力可能な情報量を上回ることによって、意図しない出力結果となる可能性があるため、表の場合は出力する行数及びフォントサイズを、図の場合は出力のサイズを調整する必要がある。なお、図のサイズは図 2.4.1 に示すように ODS GRAPHICS ステートメントで指定可能である。

図 2.4.1 ODS GRAPHICS によるグラフのサイズ指定

```
ods graphics / reset width=5in height=2.5in ;
```

PowerPoint ファイルにおけるスライドのサイズによって、指定可能な図表の大きさも変化する。プロジェクターでの投影を行う場合は、多くの機器が対応している幅と高さが 4 : 3（標準）のスライドを作成するとよい。この場合、スライドのサイズは幅 10 inch：高さ 7.5 inch（幅 25.4 cm：高さ 19.05 cm）である。単位がインチ（inch）の場合にきりがよく、例えば半分や 1/3 といった調整が容易なため、図表サイズの指定もインチ単位での指定が容易と考える。

### 2.5 スライドの改ページ

ODS POWERPOINT では、各スライドに出力可能なスペースが存在する限り、同一スライドに引き続き出力され、意図した位置で改ページさせることができないが、各スライドの出力処理の末尾に図 2.5.1 に示すプログラムを記載することで、改ページさせることが可能となる。図 2.5.1 の SAS プログラムでは「dcl odsout 参照名 ()」により、指定した参照名でレポート作成インターフェイス（RWI：Report Writing Interface）を使えるようにする機能を用いて新たなスライドを作成することで、改ページを行う。RWI を使用すると、データステップでレポートの作成が可能となる。

図 2.5.1　改ページのプログラム

```
data _null_;
  dcl odsout obj();
run;
```

## 3 適用事例

### 3.1 仮想的な臨床試験

本稿で紹介した方法を適用する事例を紹介するため、胃潰瘍再発予防に関する架空の臨床試験を想定する。本治験は第 3 相試験として実施し、胃潰瘍既往のある患者さんに対して ABC-XXX 1mg 群又はプラセボを投与する。主要評価項目は胃潰瘍再発までの期間、副次評価項目は十二指腸潰瘍発症までの期間、その他の評価項目として胃潰瘍再発の割合、十二指腸潰瘍発症の割合及び TEAE（Treatment Emergent Adverse Event：有害事象）を設定する。また、背景項目として、年齢、性別、BMI、*H. pylori* 検査、カフェインの摂取、飲酒の習慣及び喫煙歴を収集する。解析するデータセットは ADaM 形式とし、簡単のため ADSL、ADTTE 及び ADAE のみ使用する。本治験の解析手法のうち主要評価項目に関する解析のみ詳述する。

<u>主解析</u>

最大の解析対象集団（以下、FAS）を対象として、Kaplan-Meier 法により、胃潰瘍再発に関する累積再発割合及びその両側 95%信頼区間を投与群別に算出するとともに logrank 検定を適用し、ABC-XXX 群とプラセボ群の群間比較を行う。また、Kaplan-Meier 法で算出した累積再発割合とその両側 95%信頼区間を Kaplan-Meier Plot にて図示すると共に、本グラフに logrank 検定結果を併記する。

<u>その他の解析</u>

- 感度分析：治験実施計画書に適合した対象集団（以下、PPS）にて主解析を実施。
- 調整解析：投与群及び各背景項目を独立変数とした Cox 回帰分析を実施。
- 部分集団解析：各背景項目の各カテゴリに絞り、Cox 回帰分析を実施。

### 3.2 解析するデータセットと本項の目的

本治験について、データベースの固定後、直ちに解析速報を作成し、解析結果を上層部に報告することを想定する。解析結果の形式は PowerPoint ファイルとし、ODS POWERPOINT にて自動作成する。その際、胃潰瘍再発及び十二指腸潰瘍発症に関する検定結果について、SAS 上で自動的に有意性の判定を行い、結果をテキストにて出力することとする。出力する解析結果、解析対象集団及び使用するデータセットは以下の通り。

- 背景項目に関する要約統計量（連続変数）又は頻度集計（カテゴリ変数）【無作為化された全例、ADSL】
- 主要評価項目に関する Kaplan-Meier Plot 及び要約表【FAS+PPS、ADTTE】
- 副次評価項目に関する Kaplan-Meier Plot 及び要約表【FAS、ADTTE】
- 主要評価項目に関する調整解析結果を Forest Plot にて表示【FAS、ADTTE】
- 主要評価項目に関する部分集団解析結果を Forest Plot にて表示【FAS、ADTTE】
- TEAE、因果関係ありの TEAE 及び SAE に関する頻度集計を Forest Plot にて表示【安全性データの解析対象集団、ADAE+ADSL】
- 程度別の TEAE に関する頻度集計を表にて表示【安全性データの解析対象集団、ADAE+ADSL】

表 3.2.1　データセット ADSL

| 変数名 | 型 | ラベル |
|---|---|---|
| STUDYID | 文字 | Study Identifier |
| USUBJID | 文字 | Unique Subject Identifier |
| SUBJID | 文字 | Subject Identifier for the Study |
| SITEID | 文字 | Study Site Identifier |
| AGE | 数値 | Age |
| AGEU | 文字 | Age Units |
| AGEGR1 | 文字 | Pooled Age Group 1 |
| AGEGR1N | 数値 | Pooled Age Group 1 (N) |
| SEX | 文字 | Sex |
| RACE | 文字 | Race |
| BMIBL | 数値 | Baseline Body Mass Index (kg/m2) |
| BMIGR1 | 文字 | Pooled Baseline BMI (kg/m2) Group 1 |
| BMIGR1N | 数値 | Pooled Baseline BMI (kg/m2) Group 1 (N) |
| ALCOHOL | 文字 | Consumption of Alcohol |
| CAFFEINE | 文字 | Consumption of Caffeine |
| HPYLORI | 文字 | Helicobacter Pylori Infection |
| TOBACCO | 文字 | Consumption of Tobacco |
| FASFL | 文字 | Full Analysis Set Population Flag |
| PPROTFL | 文字 | Per-Protocol Population Flag |
| SAFFL | 文字 | Safety Population Flag |
| COMPLFL | 文字 | Completers Population Flag |
| RANDFL | 文字 | Randomized Population Flag |
| ARM | 文字 | Description of Planned Arm |
| ACTARM | 文字 | Description of Actual Arm |
| TRT01P | 文字 | Planned Treatment for Period 01 |
| TRT01PN | 数値 | Planned Treatment for Period 01 (N) |
| TRT01A | 文字 | Actual Treatment for Period 01 |
| TRT01AN | 数値 | Actual Treatment for Period 01 (N) |

表 3.2.2　データセット ADTTE

| 変数名 | 型 | ラベル |
|---|---|---|
| STUDYID | 文字 | Study Identifier |
| USUBJID | 文字 | Unique Subject Identifier |
| SUBJID | 文字 | Subject Identifier for the Study |
| SITEID | 文字 | Study Site Identifier |
| AGE | 数値 | Age |
| AGEU | 文字 | Age Units |
| AGEGR1 | 文字 | Pooled Age Group 1 |
| AGEGR1N | 数値 | Pooled Age Group 1 (N) |
| SEX | 文字 | Sex |
| RACE | 文字 | Race |
| BMIBL | 数値 | Baseline Body Mass Index (kg/m2) |
| BMIGR1 | 文字 | Pooled Baseline BMI (kg/m2) Group 1 |
| BMIGR1N | 数値 | Pooled Baseline BMI (kg/m2) Group 1 (N) |
| ALCOHOL | 文字 | Consumption of Alcohol |
| CAFFEINE | 文字 | Consumption of Caffeine |
| HPYLORI | 文字 | Helicobacter Pylori Infection |
| TOBACCO | 文字 | Consumption of Tobacco |
| FASFL | 文字 | Full Analysis Set Population Flag |
| PPROTFL | 文字 | Per-Protocol Population Flag |
| SAFFL | 文字 | Safety Population Flag |
| COMPLFL | 文字 | Completers Population Flag |
| RANDFL | 文字 | Randomized Population Flag |
| ARM | 文字 | Description of Planned Arm |
| ACTARM | 文字 | Description of Actual Arm |
| TRTP | 文字 | Planned Treatment |
| TRTPN | 数値 | Planned Treatment (N) |
| TRTA | 文字 | Actual Treatment |
| TRTAN | 数値 | Actual Treatment (N) |
| ASEQ | 数値 | Analysis Sequence Number |
| ADT | 数値 | Analysis Date |
| PARAM | 文字 | Parameter |
| PARAMCD | 文字 | Parameter Code |
| PARAMN | 数値 | Parameter (N) |
| AVAL | 数値 | Analysis Value |
| STARTDT | 数値 | Time to Event Origin Date for Subject |
| CNSR | 数値 | Censor |

表 3.2.3　データセット ADAE

| 変数名 | 型 | ラベル |
|---|---|---|
| STUDYID | 文字 | Study Identifier |
| USUBJID | 文字 | Unique Subject Identifier |
| SUBJID | 文字 | Subject Identifier for the Study |
| SITEID | 文字 | Study Site Identifier |
| AGE | 数値 | Age |
| AGEU | 文字 | Age Units |
| AGEGR1 | 文字 | Pooled Age Group 1 |
| AGEGR1N | 数値 | Pooled Age Group 1 (N) |
| SEX | 文字 | Sex |
| RACE | 文字 | Race |
| BMIBL | 数値 | Baseline Body Mass Index (kg/m2) |
| BMIGR1 | 文字 | Pooled Baseline BMI (kg/m2) Group 1 |

| 変数名 | 型 | ラベル |
|---|---|---|
| BMIGR1N | 数値 | Pooled Baseline BMI (kg/m2) Group 1 (N) |
| ALCOHOL | 文字 | Consumption of Alcohol |
| CAFFEINE | 文字 | Consumption of Caffeine |
| HPYLORI | 文字 | Helicobacter Pylori Infection |
| TOBACCO | 文字 | Consumption of Tobacco |
| FASFL | 文字 | Full Analysis Set Population Flag |
| PPROTFL | 文字 | Per-Protocol Population Flag |
| SAFFL | 文字 | Safety Population Flag |
| COMPLFL | 文字 | Completers Population Flag |
| RANDFL | 文字 | Randomized Population Flag |
| TRTP | 文字 | Planned Treatment |
| TRTPN | 数値 | Planned Treatment (N) |
| TRTA | 文字 | Actual Treatment |
| TRTAN | 数値 | Actual Treatment (N) |
| AESEQ | 数値 | Sequence Number |
| AETERM | 文字 | Reported Term for the Adverse Event |
| AEDECOD | 文字 | Dictionary-Derived Term |
| AEBODSYS | 文字 | Body System or Organ Class |
| TRTEMFL | 文字 | Treatment Emergent Analysis Flag |
| AESTDTC | 数値 | Start Date/Time of Adverse Event |
| AESTDY | 数値 | Study Day of Start of Adverse Event |
| AEENDTC | 数値 | End Date/Time of Adverse Event |
| AESEV | 文字 | Severity/Intensity |
| ASEV | 文字 | Analysis Severity/Intensity |
| ASEVN | 数値 | Analysis Severity/Intensity (N) |
| AEBDSYCD | 数値 | Body System or Organ Class Code |
| AESER | 文字 | Serious Event |
| AEREL | 文字 | Causality |
| AREL | 文字 | Analysis Causality |
| ARELN | 数値 | Analysis Causality (N) |
| AEPTCD | 数値 | Preferred Term Code |
| AOCCFL | 文字 | 1st Occurrence within Subject Flag |
| AOCCPFL | 文字 | 1st Occurrence of Preferred Term Flag |
| AOCCSFL | 文字 | 1st Occurrence of SOC Flag |

### 3.3 解析速報の作成

2 章で紹介した方法を適用して、SAS 上で作成した解析結果を PowerPoint ファイルに直接出力する手順を紹介する。ここでは作成する SAS プログラムの役割を、解析結果を得るまでの解析パートとその解析結果を PowerPoint ファイルに出力する出力パートに分けて説明する。なお、解析パートの SAS プログラムは報告書形式の解析図表を作成するための SAS プログラムをそのまま利用することを想定しているため、報告書形式の解析図表と解析結果が一致することが保証される。出力パートの SAS プログラムは PowerPoint ファイルにおけるスライド上の解析図表とレイアウトが同一であれば、ODS POWERPOINT 等の出力先に関する ODS ステートメント及び REPORT プロシジャ等のレポート処理に関するプロシジャのみで構成することが可能である。解析速報としての PowerPoint ファイルは、複数の解析図表が一枚のスライドに含まれる場合があるため、本稿では、ODS EXCLUDE 及び ODS SELECT ステートメントを用いて出力を制御している。SAS プログラム全体の構造を図 3.3.1 に示す。

図 3.3.1 解析パート及び出力パートの構造

```
 1 ods powerpoint file = "<PowerPoint ファイルのパス及びファイル名>. pptx" ;
 2 ods exclude all ;
 3 <解析図表の 1 つ目の解析パート>
 4 ods select all ;
 5 <解析図表の 1 つ目のレポート表示に関する処理>
 6 ods exclude all ;
 7 <解析図表の 2 つ目の解析パート>
 8 ods select all ;
 9 <解析図表の 2 つ目のレポート表示に関する処理>
10 ・・・
11 ods powerpoint close ;
```

図 3.3.1 では 3 行目と 7 行目を解析パート、それ以外を出力パートとしている。解析パートでは、解析処理を直接記載するか、%INCLUDE ステートメントを用いて報告書形式の解析図表を作成する SAS プログラムのパス及びファイル名を記載する。このとき、各解析パートにおいて ODS EXCLUDE ALL 及び ODS SELECT ALL で囲むことで、PROC ステップの実行結果 (通常アウトプットウィンドウ等に表示される) の PowerPoint ファイルへの出力を抑制することが可能となる。出力パートは、まず ODS POWERPOINT ステートメントの FILE=オプションにおいて、出力する PowerPoint ファイルのパス及びファイル名を記載する。次に解析パートと出力パートのレポート表示に関する処理を解析図表の数だけ繰り返す。なお、本稿ではレポート表示に関する処理は REPORT プロシジャ、SG プロシジャ等を、タイトル表示に関する処理は ODSTEXT プロシジャを用いた。最後に ODS POWERPOINT ステートメントの CLOSE オプションを記載する。
ここで、3.1 仮想的な臨床試験で示した解析内容について、図 3.3.1 の解析パート及び出力パートのうちレポート表示に関する処理をまとめた SAS マクロプログラム群を紹介するが、各 SAS マクロプログラムは SAS ユーザー総会 2017 のウェブサイト上で公開される。これらの SAS マクロプログラムは、解析に用いるデータセット名、変数名及びレコードの抽出条件等を指定するだけで、報告書形式及び PowerPoint 形式の解析図表への出力直前の形に整えた SAS データセットを出力する。
なお、解析に用いるデータセット構造は、CDISC ADaM V2.1、ADaM IG V1. 1及び ADaM OCCDS V1. 0 に準拠することを前提としている。各解析内容における SAS マクロ名及び使用するキーワードパラメータの説明を以下に示す。

表 3.3.1 SAS マクロプログラムの説明

| 解析内容<br>SAS マクロ名<br>キーワードパラメータ | 説明及びパラメータの指定例 |
|---|---|
| 【各 SAS マクロを実行する前準備】解析対象集団の例数の算出 | |
| マクロ名：Calc_N_of_Subj<br>マクロパラメータ：<br>・TRTG =<br>・POP =<br>・TOTAL = | <br><br>投与群 (文字変数)<br>解析対象集団の変数 ("Y","N"が格納された変数)<br>投与群の合計の出力要否 ("Y","N"を指定) |
| 入力データセット及び変数 | データセット：ADSL |

| | |
|---|---|
| (固定値) | 変数：USUBJID (ユニークな被験者番号) |
| 出力データセット及び変数<br>(固定値) | データセット：DENOM<br>変数：TRT_N (投与群ごとの例数), TRT_LN (投与群ラベル＋改行用の記号"¥"＋例数), TRT_LN2 (投与群ラベル＋例数), TRT_L (投与群ラベル) |
| 出力マクロ変数 | TRT_GN (投与群の数), TRT_Nx1～TRT_Nx100 (投与群ごとの例数), TRT_LNx1～TRT_LNx100 (投与群ラベル＋改行用の記号"¥"＋例数), TRT_LN2x1～TRT_LN2x100 (投与群ラベル＋例数), TRT_Lx1～TRT_Lx100 (投与群ラベル), TRT_L2x1～TRT_L2x100 (投与群の数値変数値) |
| **背景項目に関する要約統計量（連続変数）【無作為化された全例】** | |
| マクロ名：<br>Calc_Summary_Demog_Chara<br>マクロパラメータ：<br>・TRTG =<br>・POP =<br>・TOTAL =<br>・IND =<br>・OUTD =<br>・AVAR =<br>・NDEC =<br>・ITEM1～ITEM10 =<br>・<br>・ITEMC1～ITEMC10 =<br>・TITLE = | <br><br><br>投与群(文字変数)<br>解析対象集団の変数 ("Y","N"が格納された変数)<br>投与群の合計列の出力要否 ("Y","N"を指定)<br>入力データセット名<br>出力データセット名 (レポート出力直前のデータセット)<br>解析変数 (複数指定する場合は半角スペース区切り)<br>小数点以下の桁数を指定<br>各解析項目のラベル (複数変数を指定していた場合は半角カンマ区切り)<br>各解析項目におけるカテゴリのラベル (半角カンマ区切り)<br>タイトル |
| 入力データセット及び変数<br>(固定値) | 変数：USUBJID (ユニークな被験者番号) |
| 出力データセット及び変数<br>(固定値) | データセット：OUTD=に指定した値<br>変数：解析結果の表示に必要な値 |
| **背景項目に関する頻度集計（カテゴリ変数）【無作為化された全例】** | |
| マクロ名：<br>Calc_Freq_Demog_Chara<br>マクロパラメータ：<br>・TRTG =<br>・POP =<br>・TOTAL =<br>・IND =<br>・OUTD =<br>・AVAR =<br>・ITEM1～ITEM10 =<br><br>・ITEMC1～ITEMC10 =<br>・TITLE | <br><br><br>投与群(文字変数)<br>解析対象集団の変数 ("Y","N"が格納された変数)<br>投与群の合計列の出力要否 ("Y","N"を指定)<br>入力データセット名<br>出力データセット名 (レポート出力直前のデータセット)<br>解析変数 (複数指定する場合は半角スペース区切り)<br>各解析項目のラベル (複数変数を指定していた場合は半角カンマ区切り)<br>各解析項目におけるカテゴリのラベル (半角カンマ区切り)<br>タイトル |
| 入力データセット及び変数<br>(固定値) | 変数：USUBJID (ユニークな被験者番号) |
| 出力データセット及び変数<br>(固定値) | データセット：OUTD=に指定した値<br>変数：解析結果の表示に必要な値 |
| **主要評価項目に関する Kaplan-Meier Plot【FAS+PPS、ADTTE】<br>副次評価項目に関する Kaplan-Meier Plot【FAS、ADTTE】** | |
| マクロ名： | |

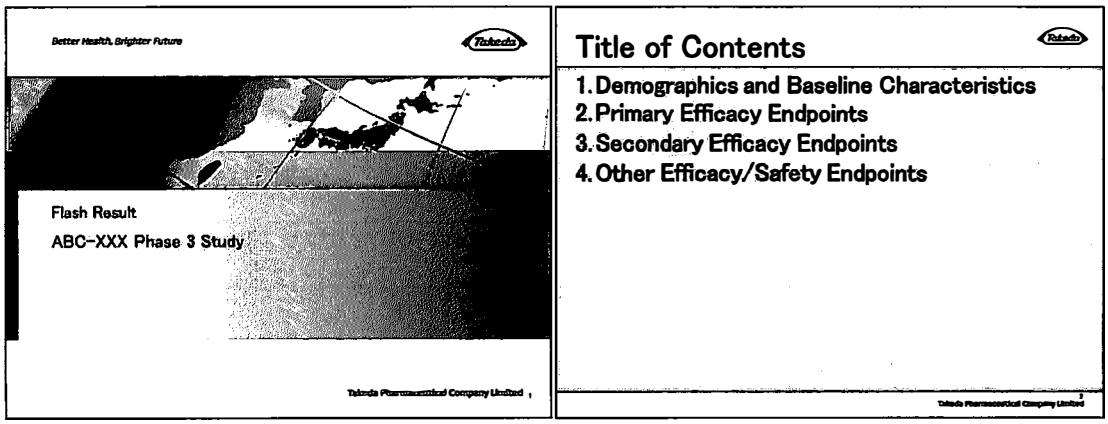
| | |
|---|---|
| Plot_Of_Cumulative_Incidence<br>マクロパラメータ：<br>・TRTG =<br>・POP =<br>・IND =<br>・OUTD =<br>・PARAMCD =<br>・CENSOR =<br>・INTERVAL =<br>・xLabel =<br>・yLabel =<br>・Text =<br>・Color =<br>・Symbol =<br>・TITLE = | <br><br>投与群(文字変数)<br>解析対象集団の変数 ("Y","N"が格納された変数)<br>入力データセット名<br>出力データセット名 (レポート出力直前のデータセット)<br>検査項目コード<br>打ち切りを示す値 (変数 CNSR の値を指定)<br>X 軸の時間間隔<br>X 軸のラベル<br>Y 軸のラベル<br>logrank 検定による生存関数の有意差の有無に関するコメント<br>プロットシンボル及び線の色<br>プロットシンボルの形<br>タイトル |
| 入力データセット及び変数(固定値) | 変数：AVAL (イベント発現又は打切りまでの期間), CNSR (打ち切り) |
| 出力データセット及び変数(固定値) | データセット：OUTD=に指定した値<br>変数：解析結果の表示に必要な値 |
| **主要評価項目に関する割合及び 95%信頼区間の要約表【FAS+PPS、ADTTE】<br>副次評価項目に関する割合及び 95% 信頼区間の要約表【FAS、ADTTE】** | |
| マクロ名：<br>Calc_Cumulative_Incidence<br>マクロパラメータ：<br>・TRTG =<br>・POP =<br>・IND =<br>・OUTD =<br>・PARAMCD =<br>・CENSOR =<br>・INTERVAL =<br>・TITLE = | % Calc_Cumulative_Incidence (TRTG=TRTP, POP=FASFL, IND=ADTTE, OUTD=TEST4, PARAMCD=GU, AVAR=AVAL, CENSORING_VAR=, CENSOR=, INTERVAL=) ;<br>投与群(文字変数)<br>解析対象集団の変数 ("Y","N"が格納された変数)<br>入力データセット名<br>出力データセット名 (レポート出力直前のデータセット)<br>検査項目コード<br>打ち切りを示す値 (CENSORING_VAR の値を指定)<br>X 軸の時間間隔<br>タイトル |
| 入力データセット及び変数(固定値) | 変数：AVAL (イベント発現又は打ち切りまでの期間), CNSR (打ち切り) |
| 出力データセット及び変数(固定値) | データセット：OUTD=に指定した値<br>変数：解析結果の表示に必要な値 |
| **主要評価項目に関する割合の投与群間差の点推定値及び 95%信頼区間【FAS+PPS、ADTTE】<br>副次評価項目に関する割合の投与群間差の点推定値及び 95%信頼区間【FAS、ADTTE】** | |
| マクロ名：<br>Proportion_Of_Analysis_Var<br>マクロパラメータ：<br>・TRTG =<br>・POP =<br>・IND =<br>・OUTD =<br>・PARAMCD1 =<br>・PARAMCD2 =<br>・CENSOR =<br>・CENSOR_LABEL = | <br><br>投与群(文字変数)<br>解析対象集団の変数 ("Y","N"が格納された変数)<br>入力データセット名<br>出力データセット名 (レポート出力直前のデータセット)<br>検査項目 1<br>検査項目 2<br>打ち切りを示す値 (変数 CNSR の値を指定)<br>CENSOR のラベル |

| | |
|---|---|
| ・EVENT_LABEL =<br>・INTERVAL =<br>・TITLE = | EVENT のラベル<br>X 軸の時間間隔<br>タイトル |
| 入力データセット及び変数<br>(固定値) | 変数：AVAL (イベント発現又は打ち切りまでの期間), CNSR (打ち切り) |
| 出力データセット及び変数<br>(固定値) | データセット：OUTD=に指定した値<br>変数：解析結果の表示に必要な値 |
| **主要評価項目に関する調整解析結果を Forest Plot にて表示【FAS、ADTTE】** | |
| マクロ名：<br>Adj_for_Covariates<br>マクロパラメータ：<br>・TRTG =<br>・POP =<br>・IND =<br>・OUTD =<br>・CENSOR =<br>・AVAR =<br>・ITEM1～ITEM10 =<br>・TITLE = | <br><br><br>投与群(文字変数)<br>解析対象集団の変数 ("Y","N"が格納された変数)<br>入力データセット名<br>出力データセット名 (レポート出力直前のデータセット)<br>打ち切りを示す値 (変数 CNSR の値を指定)<br>独立変数 (複数指定する場合は半角スペース区切り)<br>独立変数のラベル (複数変数を指定していた場合は半角カンマ区切り)<br>タイトル |
| 入力データセット及び変数<br>(固定値) | 変数：AVAL (イベント発現又は打切りまでの期間), CNSR (打ち切り) |
| 出力データセット及び変数<br>(固定値) | データセット：OUTD=に指定した値<br>変数：解析結果の表示に必要な値 |
| **主要評価項目に関する部分集団解析結果を Forest Plot にて表示【FAS、ADTTE】** | |
| マクロ名：<br>Subgroup_Analysis<br>マクロパラメータ：<br>・TRTG =<br>・POP =<br>・IND =<br>・OUTD =<br>・CENSOR =<br>・AVAR =<br>・ITEM1～ITEM10 =<br>・TITLE = | <br><br><br>投与群(文字変数)<br>解析対象集団の変数 ("Y","N"が格納された変数)<br>入力データセット名<br>出力データセット名 (レポート出力直前のデータセット)<br>打ち切りを示す値 (変数 CNSR の値を指定)<br>独立変数 (複数指定する場合は半角スペース区切り)<br>独立変数のラベル (複数変数を指定していた場合は半角カンマ区切り)<br>タイトル |
| 入力データセット及び変数<br>(固定値) | 変数：AVAL (イベント発現又は打切りまでの期間), CNSR (打ち切り) |
| 出力データセット及び変数<br>(固定値) | データセット：OUTD=に指定した値<br>変数：解析結果の表示に必要な値 |
| **TEAE、因果関係ありの TEAE 及び SAE に関する頻度集計を Forest Plot にて表示【安全性データの解析対象集団、ADAE+ADSL】** | |
| マクロ名：<br>All_TEAEs<br>マクロパラメータ：<br>・COND =<br>・GRP1 =<br>・GRP2 =<br>・GRP1_LABEL =<br>・GRP2_LABEL = | <br><br><br>集計対象とする有害事象の抽出条 件を指定<br>投与群 1<br>投与群 2<br>投与群 1 のラベル<br>投与群 2 のラベル |

| | |
|---|---|
| ・OUTD =<br>・TITLE = | 出力データセット名（レポート出力直前のデータセット）<br>タイトル |
| 入力データセット及び変数<br>(固定値) | 変数：USUBJID (ユニークな被験者番号), AEBODSYS (器官分類),<br>AEDECOD (基本語) |
| 出力データセット及び変数<br>(固定値) | データセット：OUTD=に指定した値<br>変数：解析結果の表示に必要な値 |
| **程度別の TEAE に関する頻度集計を表にて表示【安全性データの解析対象集団、ADAE+ADSL】** | |
| マクロ名：<br>Calc _Freq_AEINTE<br>マクロパラメータ：<br>・TRTG =<br>・POP =<br>・TOTAL =<br>・IND =<br>・OUTD =<br>・WHERE = | 指定例)<br>%Calc_Freq_AEINTE (TRTG=TRT01P, POP=RANDFL, TOTAL=Y,<br>IND=ADAE, OUTD=TEST2, WHERE=) ;<br>投与群（文字変数）<br>解析対象集団の変数 ("Y","N"が格納された変数)<br>投与群の合計列の出力要否 ("Y","N"を指定)<br>入力データセット名<br>出力データセット名（レポート出力直前のデータセット）<br>レコードの抽出 |
| 入力データセット及び変数<br>(固定値) | 変数：USUBJID (ユニークな被験者番号), AEBODSYS (器官分類),<br>AEDECOD (基本語), ASEV (程度) |
| 出力データセット及び変数<br>(固定値) | データセット：OUTD=に指定した値<br>変数：解析結果の表示に必要な値 |

表3 .3 .1 の SAS マクロプログラムについて、図3 .3 .1 に示した構造を適用して得られたスライドを図3 .3 .2 に示す。この出力結果では、各スライドの背景用画像ファイルを事前に作成し、ODS POWERPOINT の bac kgroundimage オプションに画像ファイルのパスを指定している。

図3 .3 .2 ODS POWER ポイントを用いて作成したプレゼンテーションスライド



## Demographics and Baseline Characteristics

| Variable | Treatment | N | Mean | S.D. | Min. | Q1 | Median | Q3 | Max. |
|---|---|---|---|---|---|---|---|---|---|
| Age (years) | ABC-XXX | 100 | 53.5 | 14.47 | 30 | 40.0 | 53.5 | 66.0 | 79 |
| | Placebo | 100 | 53.5 | 14.52 | 30 | 40.0 | 53.0 | 66.5 | 79 |
| | Total | 200 | 53.5 | 14.46 | 30 | 40.0 | 53.0 | 66.0 | 79 |
| BMI (kg/m^2) | ABC-XXX | 100 | 27.26 | 5.590 | 18.0 | 22.55 | 26.95 | 32.10 | 37.4 |
| | Placebo | 100 | 27.77 | 5.701 | 18.2 | 22.70 | 27.65 | 33.20 | 37.8 |
| | Total | 200 | 27.51 | 5.637 | 18.0 | 22.65 | 27.50 | 32.50 | 37.8 |



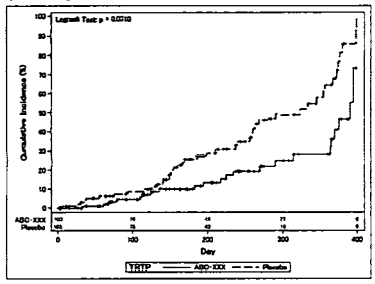
## Demographics and Baseline Characteristics

| Variable | Category | ABC-XXX | Placebo | Total |
|---|---|---|---|---|
| Age (years) | Min<= - <65 | 72 ( 72.0) | 70 ( 70.0) | 142 ( 71.0) |
| | 65<= - <=Max | 28 ( 28.0) | 30 ( 30.0) | 58 ( 29.0) |
| Alcohol Consumption | Yes | 31 ( 31.0) | 24 ( 24.0) | 55 ( 27.5) |
| | No | 69 ( 69.0) | 76 ( 76.0) | 145 ( 72.5) |
| BMI (kg/m^2) | Min<= - <25.0 | 39 ( 39.0) | 37 ( 37.0) | 76 ( 38.0) |
| | 25.0<= - <=Max | 61 ( 61.0) | 63 ( 63.0) | 124 ( 62.0) |
| Caffeine Consumption | Yes | 38 ( 38.0) | 32 ( 32.0) | 70 ( 35.0) |
| | No | 62 ( 62.0) | 68 ( 68.0) | 130 ( 65.0) |
| Gender | Male | 52 ( 52.0) | 61 ( 61.0) | 113 ( 56.5) |
| | Female | 48 ( 48.0) | 39 ( 39.0) | 87 ( 43.5) |
| H. pylori Infection | Positive | 46 ( 46.0) | 45 ( 45.0) | 91 ( 45.5) |
| | Negative | 54 ( 54.0) | 55 ( 55.0) | 109 ( 54.5) |
| Tobacco Consumption | Yes | 18 ( 18.0) | 21 ( 21.0) | 39 ( 19.5) |
| | No | 82 ( 82.0) | 79 ( 79.0) | 161 ( 80.5) |



## Cumulative Incidence of GU [FAS]

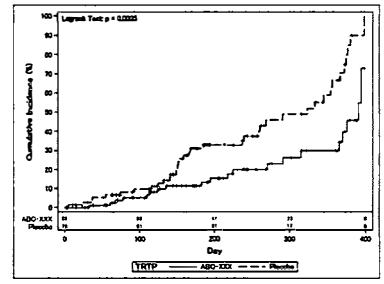
Superiority of ABC-XXX to Placebo was confirmed.






## Cumulative Incidence of GU [PPS]

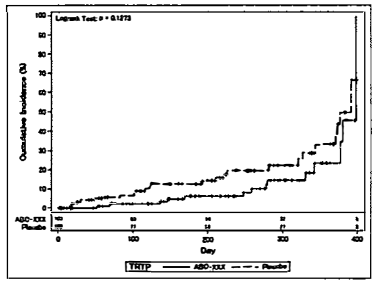
Superiority of ABC-XXX to Placebo was confirmed.






## Cumulative Incidence of DU [FAS]

Superiority of ABC-XXX to Placebo was not confirmed.






## Life Table of GU [FAS]

| Group | Time | At Risk | Event | Censor | % | 95% CI Lower | 95% CI Upper |
|---|---|---|---|---|---|---|---|
| ABC-XXX | 0 | 100 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| | 100 | 76 | 4 | 20 | 4.6 | 1.5 | 10.5 |
| | 200 | 49 | 10 | 41 | 13.4 | 6.7 | 22.4 |
| | 300 | 25 | 15 | 60 | 24.9 | 14.0 | 37.4 |
| | 400 | 0 | 23 | 77 | 73.2 | 38.2 | 90.4 |
| Placebo | 0 | 100 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| | 100 | 78 | 8 | 14 | 8.7 | 4.0 | 15.6 |
| | 200 | 40 | 23 | 37 | 29.0 | 19.2 | 39.5 |
| | 300 | 19 | 32 | 49 | 49.0 | 34.8 | 61.7 |
| | 400 | 0 | 44 | 56 | 100.0 | 100.0 | 100.0 |

## Life Table of GU [PPS]

| Group | Time | At Risk | Event | Censor | % | 95% CI Lower | 95% CI Upper |
|---|---|---|---|---|---|---|---|
| ABC-XXX | 0 | 88 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| | 100 | 66 | 4 | 18 | 5.3 | 1.7 | 12.0 |
| | 200 | 41 | 10 | 37 | 15.4 | 7.7 | 25.5 |
| | 300 | 22 | 14 | 52 | 26.4 | 14.9 | .9 |
| | 400 | 0 | 21 | 67 | 72.9 | 37.4 | 90.3 |
| Placebo | 0 | 78 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| | 100 | 61 | 7 | 10 | 9.7 | 4.2 | 17.9 |
| | 200 | 31 | 21 | 26 | 32.7 | 21.4 | 44.5 |
| | 300 | 17 | 27 | 34 | 48.9 | 33.6 | 62.5 |
| | 400 | 0 | 38 | 40 | 100.0 | 100.0 | 100.0 |

Takeda Pharmaceutical Company Limited

## Life Table of DU [FAS]

| Group | Time | At Risk | Event | Censor | % | 95% CI Lower | 95% CI Upper |
|---|---|---|---|---|---|---|---|
| ABC-XXX | 0 | 100 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| | 100 | 80 | 2 | 18 | 2.2 | 0.4 | 7.1 |
| | 200 | 56 | 5 | 39 | 6.3 | 2.3 | 13.3 |
| | 300 | 32 | 9 | 59 | 14.7 | 6.9 | 25.3 |
| | 400 | 0 | 14 | 86 | 100.0 | 100.0 | 100.0 |
| Placebo | 0 | 100 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| | 100 | 77 | 6 | 17 | 6.6 | 2.7 | 12.9 |
| | 200 | 56 | 12 | 32 | 14.4 | 7.8 | 22.9 |
| | 300 | 27 | 16 | 57 | 22.1 | 12.9 | 32.9 |
| | 400 | 0 | 23 | 77 | 66.4 | 21.9 | 89.6 |

Takeda Pharmaceutical Company Limited

## Proportion of GU/DU [FAS]

The difference in proportions of GU was statistically significant.

The difference in proportions of DU was not statistically significant.

| Test | Group | Yes (%) | No | Total | 95% CI Lower | 95% CI Upper |
|---|---|---|---|---|---|---|
| GU | ABC-XXX | 23 ( 23.0) | 77 | 100 | 15.2 | 32.5 |
| | Placebo | 44 ( 44.0) | 56 | 100 | 34.1 | 54.3 |
| DU | ABC-XXX | 14 ( 14.0) | 86 | 100 | 7.9 | 22.4 |
| | Placebo | 23 ( 23.0) | 77 | 100 | 15.2 | 32.5 |

| Test | Group | Diff. | 95% CI Lower | 95% CI Upper | Chi-square Test Chisq. | DF | p-value |
|---|---|---|---|---|---|---|---|
| GU | ABC-XXX - Placebo | -21.0 | -33.8 | -8.2 | 8.978 | 1 | 0.0027 |
| DU | ABC-XXX - Placebo | -9.0 | -19.7 | 1.7 | 2.122 | 1 | 0.1452 |

Takeda Pharmaceutical Company Limited

## Proportion of GU/DU [FAS]

The difference in proportions of GU was statistically significant.

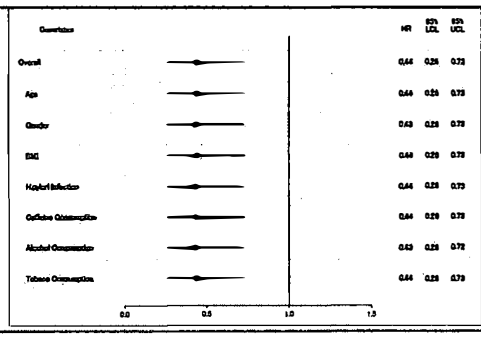
The difference in proportions of DU was not statistically significant.

| Test | Group | Yes (%) | No | Total | 95% CI Lower | 95% CI Upper |
|---|---|---|---|---|---|---|
| GU | ABC-XXX | 21 ( 23.9) | 67 | 88 | 15.4 | 34.1 |
| | Placebo | 38 ( 48.7) | 40 | 78 | 37.2 | 60.3 |
| DU | ABC-XXX | 13 ( 14.8) | 75 | 88 | 8.1 | 23.9 |
| | Placebo | 15 ( 19.2) | 63 | 78 | 11.2 | 29.7 |

| Test | Group | Diff. | 95% CI Lower | 95% CI Upper | Chi-square Test Chisq. | DF | p-value |
|---|---|---|---|---|---|---|---|
| GU | ABC-XXX - Placebo | -24.9 | -39.1 | -10.6 | 10.091 | 1 | 0.0015 |
| DU | ABC-XXX - Placebo | -4.5 | -15.9 | 7.0 | 0.311 | 1 | 0.5769 |

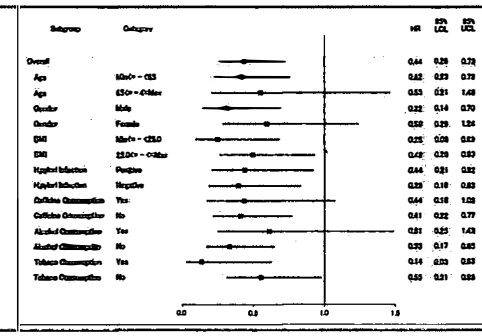
Takeda Pharmaceutical Company Limited

## Adjustments for Covariates

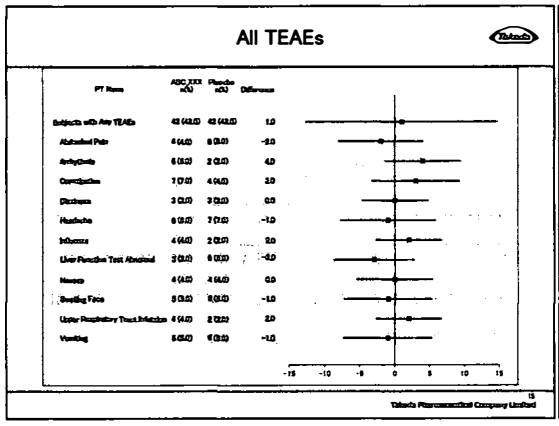


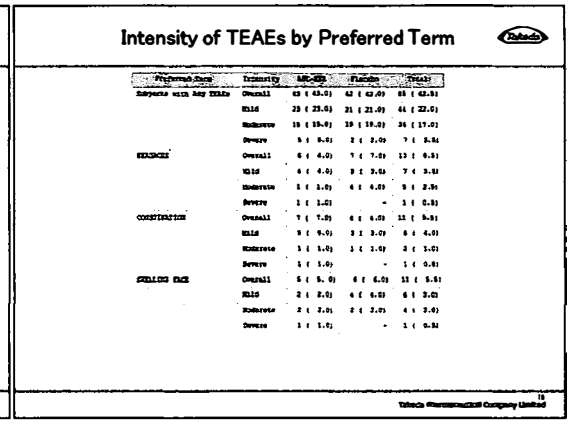

Takeda Pharmaceutical Company Limited

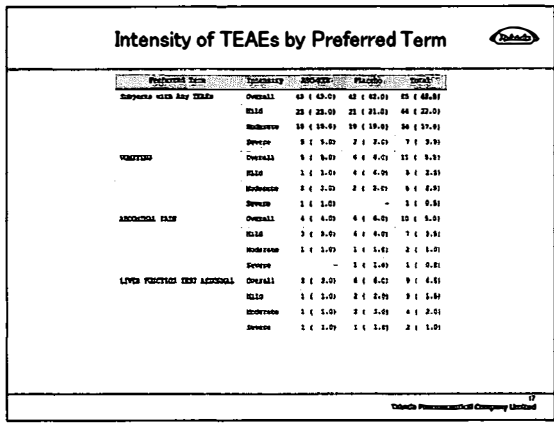
## Subgroup Analysis

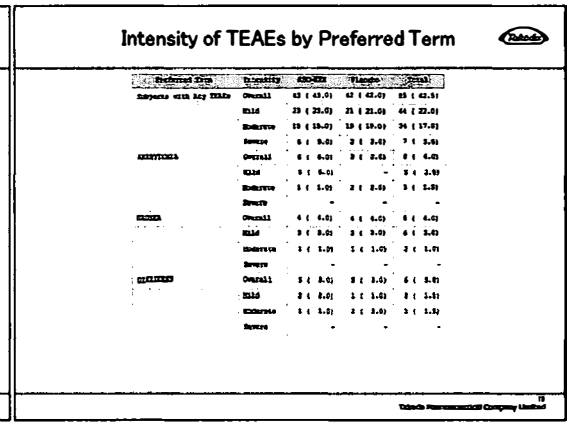


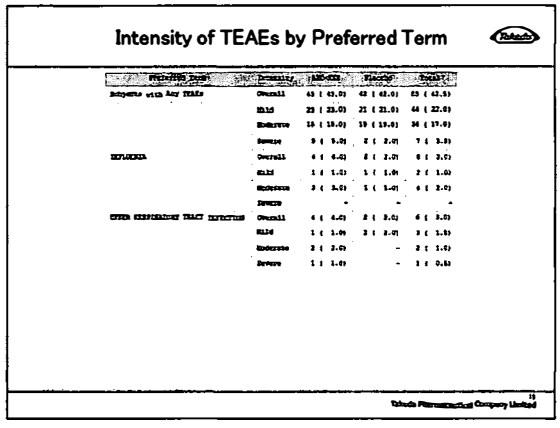

Takeda Pharmaceutical Company Limited



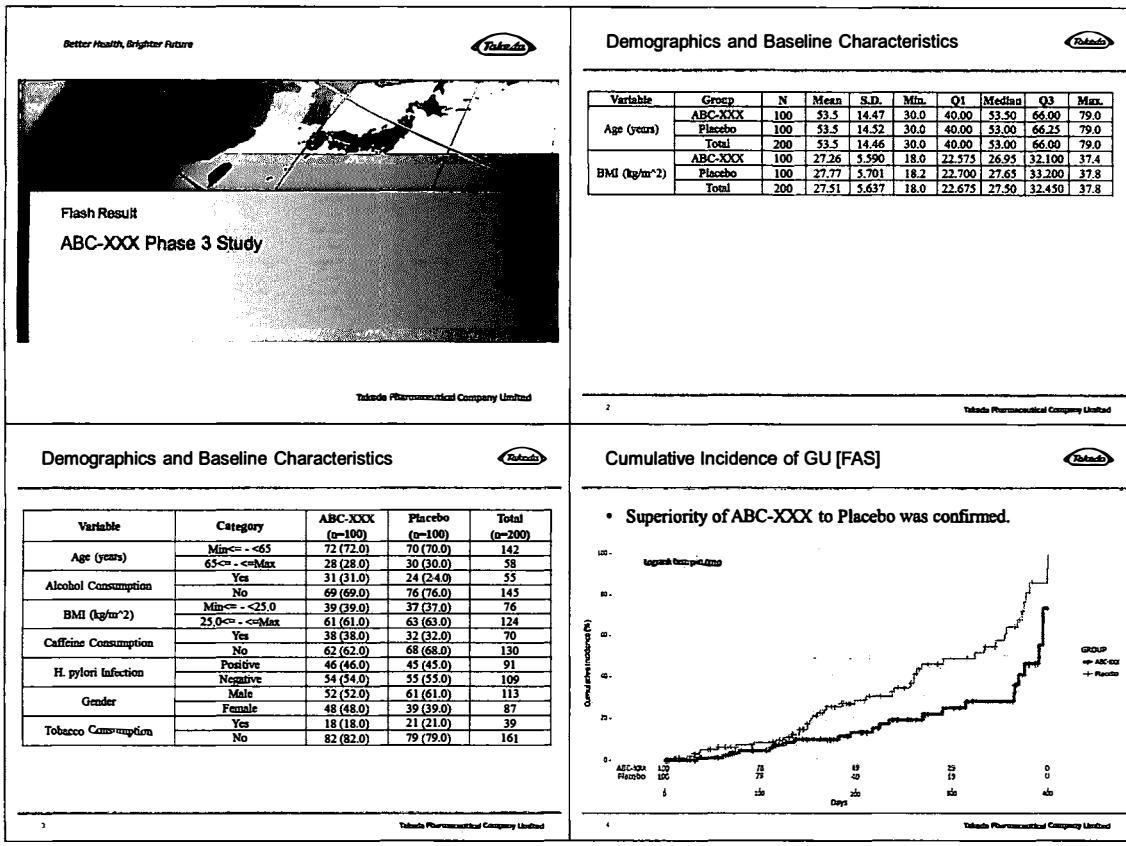

## 3.4 他のソフトウェアとの機能比較

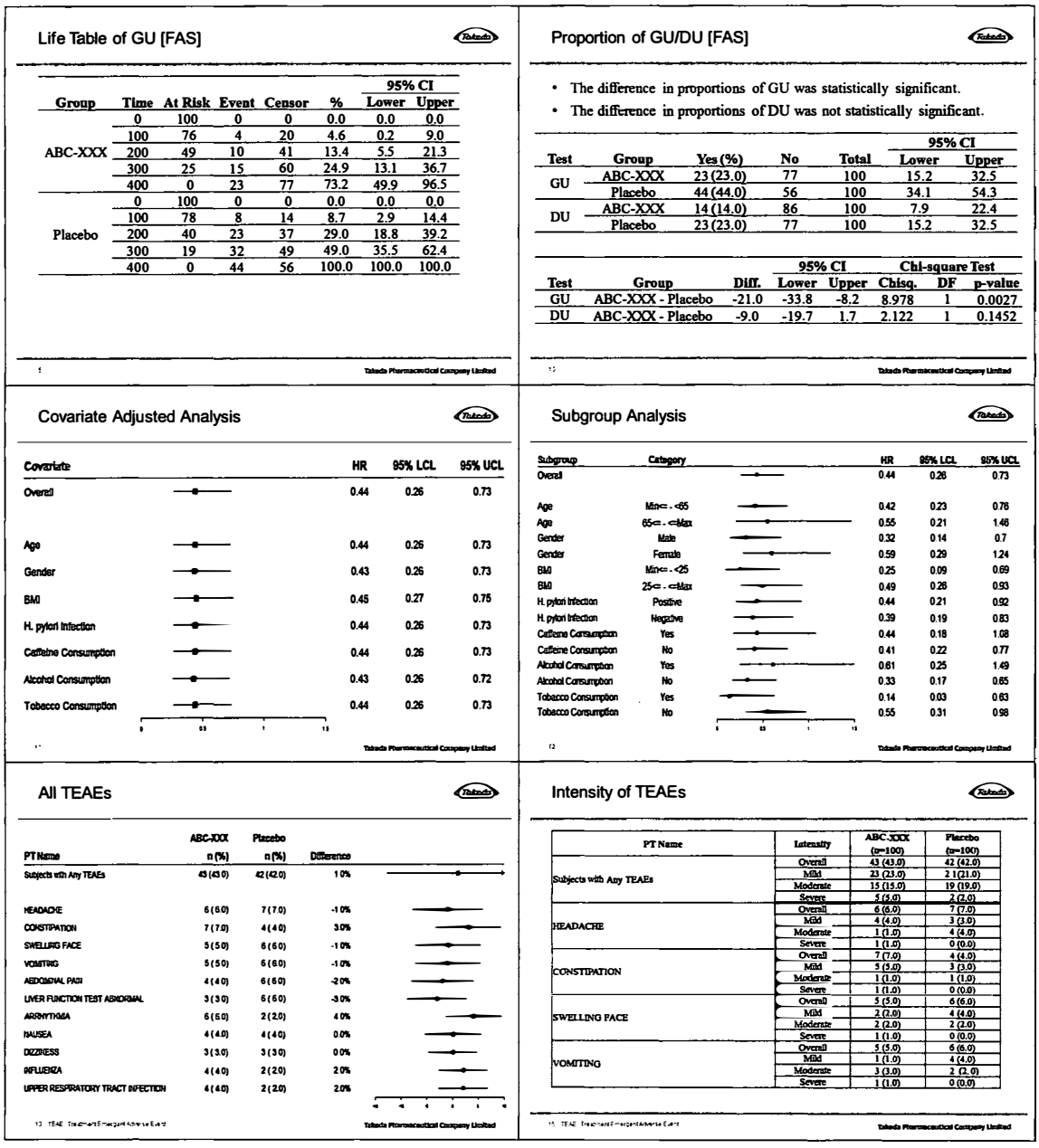
参考までに、同様の集計を統計ソフト R（主なパッケージ：ReporteRs）で実施した結果の抜粋を、SAS による方法との相違点と合わせて紹介する。

① SAS では解析結果をデータセットに出力する機能が豊富であるため、解析結果の二次利用が容易である。一方、R では解析結果の二次利用（特に解析結果をデータセット形式にて出力）に時間を要する場合が多い。この点は、解析結果を表形式で出力する際に顕著な作業効率の違いとなる。

② SAS の ODS POWERPOINT では、既存の PowerPoint ファイルや PowerPoint テンプレートの読込み・利用が出来ない。多くの製薬企業では PowerPoint スライドの形式が決まっている場合が多いため、スライドのスタイルを整えるのに困難を伴う。一方、R では既存の PowerPoint ファイルの利用が可能であるため、スタイルを整える必要がほとんどない上、既存の PowerPoint ファイルにあらかじめコンテンツを含めた上で（例：治験デザインや意思決定フロー）、解析結果を PowerPoint ファイルへ追記することが可能である。

図 3.4.1　統計ソフト「R」を用いて作成したプレゼンテーションスライド





Life Table of GU [FAS]

| Group | Time | At Risk | Event | Censor | % | 95% CI Lower | 95% CI Upper |
|---|---|---|---|---|---|---|---|
| ABC-XXX | 0 | 100 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| | 100 | 76 | 4 | 20 | 4.6 | 0.2 | 9.0 |
| | 200 | 49 | 10 | 41 | 13.4 | 5.5 | 21.3 |
| | 300 | 25 | 15 | 60 | 24.9 | 13.1 | 36.7 |
| | 400 | 0 | 23 | 77 | 73.2 | 49.9 | 96.5 |
| Placebo | 0 | 100 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| | 100 | 78 | 8 | 14 | 8.7 | 2.9 | 14.4 |
| | 200 | 40 | 23 | 37 | 29.0 | 18.8 | 39.2 |
| | 300 | 19 | 32 | 49 | 49.0 | 35.5 | 62.4 |
| | 400 | 0 | 44 | 56 | 100.0 | 100.0 | 100.0 |

Proportion of GU/DU [FAS]

- The difference in proportions of GU was statistically significant.
- The difference in proportions of DU was not statistically significant.

| Test | Group | Yes (%) | No | Total | 95% CI Lower | 95% CI Upper |
|---|---|---|---|---|---|---|
| GU | ABC-XXX | 23 (23.0) | 77 | 100 | 15.2 | 32.5 |
| | Placebo | 44 (44.0) | 56 | 100 | 34.1 | 54.3 |
| DU | ABC-XXX | 14 (14.0) | 86 | 100 | 7.9 | 22.4 |
| | Placebo | 23 (23.0) | 77 | 100 | 15.2 | 32.5 |

| Test | Group | Diff. | 95% CI Lower | 95% CI Upper | Chi-square Test Chisq. | DF | p-value |
|---|---|---|---|---|---|---|---|
| GU | ABC-XXX - Placebo | -21.0 | -33.8 | -8.2 | 8.978 | 1 | 0.0027 |
| DU | ABC-XXX - Placebo | -9.0 | -19.7 | 1.7 | 2.122 | 1 | 0.1452 |

Covariate Adjusted Analysis

| Covariate | HR | 95% LCL | 95% UCL |
|---|---|---|---|
| Overall | 0.44 | 0.26 | 0.73 |
| Age | 0.44 | 0.26 | 0.73 |
| Gender | 0.43 | 0.26 | 0.73 |
| BMI | 0.45 | 0.27 | 0.75 |
| H. pylori Infection | 0.44 | 0.26 | 0.73 |
| Caffeine Consumption | 0.44 | 0.26 | 0.73 |
| Alcohol Consumption | 0.43 | 0.26 | 0.72 |
| Tobacco Consumption | 0.44 | 0.26 | 0.73 |

Subgroup Analysis

| Subgroup | Category | HR | 95% LCL | 95% UCL |
|---|---|---|---|---|
| Overall | | 0.44 | 0.26 | 0.73 |
| Age | Min<= - <65 | 0.42 | 0.23 | 0.76 |
| Age | 65<= - <=Max | 0.55 | 0.21 | 1.46 |
| Gender | Male | 0.32 | 0.14 | 0.7 |
| Gender | Female | 0.59 | 0.29 | 1.24 |
| BMI | Min<= - <25 | 0.25 | 0.09 | 0.69 |
| BMI | 25<= - <=Max | 0.49 | 0.28 | 0.93 |
| H. pylori Infection | Positive | 0.44 | 0.21 | 0.92 |
| H. pylori Infection | Negative | 0.39 | 0.19 | 0.83 |
| Caffeine Consumption | Yes | 0.44 | 0.18 | 1.08 |
| Caffeine Consumption | No | 0.41 | 0.22 | 0.77 |
| Alcohol Consumption | Yes | 0.61 | 0.25 | 1.49 |
| Alcohol Consumption | No | 0.33 | 0.17 | 0.65 |
| Tobacco Consumption | Yes | 0.14 | 0.03 | 0.63 |
| Tobacco Consumption | No | 0.55 | 0.31 | 0.98 |

All TEAEs

| PT Name | ABC-XXX n (%) | Placebo n (%) | Difference |
|---|---|---|---|
| Subjects with Any TEAEs | 43 (43.0) | 42 (42.0) | 1.0% |
| HEADACHE | 6 (6.0) | 7 (7.0) | -1.0% |
| CONSTIPATION | 7 (7.0) | 4 (4.0) | 3.0% |
| SWELLING FACE | 5 (5.0) | 6 (6.0) | -1.0% |
| VOMITING | 5 (5.0) | 6 (6.0) | -1.0% |
| ABDOMINAL PAIN | 4 (4.0) | 6 (6.0) | -2.0% |
| LIVER FUNCTION TEST ABNORMAL | 3 (3.0) | 6 (6.0) | -3.0% |
| ARRHYTHMIA | 6 (6.0) | 2 (2.0) | 4.0% |
| NAUSEA | 4 (4.0) | 4 (4.0) | 0.0% |
| DIZZINESS | 3 (3.0) | 3 (3.0) | 0.0% |
| INFLUENZA | 4 (4.0) | 2 (2.0) | 2.0% |
| UPPER RESPIRATORY TRACT INFECTION | 4 (4.0) | 2 (2.0) | 2.0% |

Intensity of TEAEs

| PT Name | Intensity | ABC-XXX (n=100) | Placebo (n=100) |
|---|---|---|---|
| Subjects with Any TEAEs | Overall | 43 (43.0) | 42 (42.0) |
| | Mild | 23 (23.0) | 21 (21.0) |
| | Moderate | 15 (15.0) | 19 (19.0) |
| | Severe | 5 (5.0) | 2 (2.0) |
| HEADACHE | Overall | 6 (6.0) | 7 (7.0) |
| | Mild | 4 (4.0) | 3 (3.0) |
| | Moderate | 1 (1.0) | 4 (4.0) |
| | Severe | 1 (1.0) | 0 (0.0) |
| CONSTIPATION | Overall | 7 (7.0) | 4 (4.0) |
| | Mild | 5 (5.0) | 3 (3.0) |
| | Moderate | 1 (1.0) | 1 (1.0) |
| | Severe | 1 (1.0) | 0 (0.0) |
| SWELLING FACE | Overall | 5 (5.0) | 6 (6.0) |
| | Mild | 2 (2.0) | 4 (4.0) |
| | Moderate | 2 (2.0) | 2 (2.0) |
| | Severe | 1 (1.0) | 0 (0.0) |
| VOMITING | Overall | 5 (5.0) | 6 (6.0) |
| | Mild | 1 (1.0) | 4 (4.0) |
| | Moderate | 3 (3.0) | 2 (2.0) |
| | Severe | 1 (1.0) | 0 (0.0) |

## 4 おわりに

本稿では、SAS の ODS POWERPOINT 及び ODS LAYOUT を用いた PowerPoint ファイルの作成方法を紹介した後、臨床試験における結果を解析速報として報告する際のプレゼンテーションスライドを効率的に作成する方法として、SAS プログラムによる解析速報（プレゼンテーションスライド）の作成を提案した。その中で、報告書形式の解析図表を作成する SAS プログラムに、PowerPoint ファイルへの出力処理を追記するだけでよいことから、解析結果の一致性も保証されると述べた。

また、PowerPoint ファイルに解析図表を直接出力することによりマニュアルでの転記・貼付作業の回避が可能であるため効率的である。なお、統計ソフト「R」を用いて同様の PowerPoint ファイルを作成したが、既存の PowerPoint ファイルの利用が可能であるという利点がある一方、解析結果の二次利用に時間を要するため、SAS に一日の長がある。今後、ODS POWERPOINT で既存の PowerPoint ファイルの利用が可能となれば、事前に治験デザイン等の解析結果以外の内容を含む PowerPoint ファイルに解析結果を追加で出力することが可能となるため、さらなる機能追加が期待される。

連絡先
naoki.yoshida1@takeda.com

## 参考文献・引用文献

- Jane Eslinger. (2016). The Dynamic Duo: ODS Layout and the ODS Destination for PowerPoint; Available at http://support.sas.com/resources/papers/proceedings16/SAS5443-2016.pdf [Accessed 1 July 2017]
- Amber Carlson, Amelia Stein, Xiaobin Zhou. (2016). Proc Report, the Graph Template Language, and ODS Layouts: Used in Unison to Create Dynamic, Customer-Ready PowerPoints; Available at http://analytics.ncsu.edu/sesug/2016/RV-135_Final_PDF.pdf [Accessed 1 July 2017]
- Tim Hunter. (2016). A Second Look at the ODS Destination for PowerPoint; Available at http://support.sas.com/resources/papers/proceedings16/SAS3801-2016.pdf [Accessed 1 July 2017]
- Tim Hunter. (2013). A First Look at the ODS Destination for PowerPoint; Available at http://www.lexjansen.com/wuss/2013/143_Paper.pdf [Accessed 1 July 2017]
- 吉田早織, 平井隆幸, 叶健, 魚住龍史. (2015). ODS POWERPOINT の活用：SAS から Microsoft PowerPoint へのエクスポート; Available at http://www.sascom.jp/download/pdf/usergroups2015_D-04.pdf [Accessed 1 July 2017]
- 高浪洋平. (2011). SG プロシジャと GTL によるグラフの作成と ODS PDF による統合解析帳票の作成; Available at http://www.sascom.jp/download/pdf/usergroups11_B-19.pdf [Accessed 1 July 2017]

以上

# レガシーデータ変換等におけるトランスコーディングの問題に関する考察と提案

○山崎　文寛[1]、高浪　洋平[1]

[1]武田薬品工業株式会社　日本開発センター　生物統計室

Considerations and Proposals on Transcoding issues in Legacy Data Conversion

Fumihiro Yamasaki[1], Yohei Takanami[1]

[1]Takeda Pharmaceutical Company Limited

## 要旨

新医薬品の製造販売の承認申請時電子データ提出におけるレガシーデータ変換時のトランスコーディングに関する問題を整理し，医薬品医療機器総合機構 (PMDA) の規制要件に合致しないデータを検出・変換する方法を提案する．

キーワード：トランスコーディング，セッションエンコーディング，符号化方式 (Encoding)，文字セット (Character Set)，ASCII，Shift-JIS，Wlatin1，UTF-8，電子データ提出，統合解析，レガシーデータ変換

## 1　はじめに

日本では，2016 年 10 月 1 日に新薬の承認申請を行う品目について PMDA への電子データの提出が開始されたが，3 年半の経過措置期間を経て，2020 年 4 月 1 日より提出が義務化される予定である．その電子データの提出対象となる臨床試験及びその提出形式等は承認申請時の電子データ提出に関する実務的事項について (平成 27 年 4 月 27 日付け厚生労働省医薬食品局審査管理課長通知，以下「実務的通知」と記載) 等で述べられている．提出する電子データは CDISC 標準に準拠する必要があるが，国内外で実施される臨床試験の電子データ及び関連文書の作成に関わる SAS プログラマは，それらの作成環境にも目を向ける必要がある．

セッションエンコーディング (SAS 起動時に適用されるデフォルトの符号化方式) はプラットフォームに依存する．例えば Windows の日本語環境では Shift-JIS，Windows の英語環境では Wlatin1 であり，SAS プログラマが符号化方式を指定して SAS データセットを作成しない限り，セッションエンコーディングが，作成する SAS データセットにそのまま適用される．過去に国内外で実施した臨床試験の電子データ (レガシーデータ) を PMDA に提出する際，英語以外の言語が含まれる場合は，実務的通知等に従い英語に変換する必要がある．変換の際は，セッションエンコーディングが電子データの符号化方式と一致している SAS 環境を用いて言語のみ変換するか，SAS の Unicode サポート環境を用いて言語だけ

でなく符号化方式も UTF-8 等に変換 (トランスコーディング) する方法の 2 通りが考えられる．実務的通知に従い PMDA から統合解析 (ISS/ISE) に関する電子データが求められたが臨床試験ごとに符号化方式が異なる場合や電子データを使用する社内の解析環境 (例えばグローバルで統一された SAS の実行環境や統計解析システム等) が特定の符号化方式にしか対応していない場合には，後者の方法を用いる必要がある．UTF-8 は英語圏だけでなく多くの国や地域で使用される言語に対応している符号化方式であり，プラットフォームに依存しないため，業種を問わず幅広く利用されている．SAS の Unicode サポート環境を使用することで，レガシーデータを処理する際にその符号化方式に応じて適切な対応を行えば，英語以外の言語であったとしても文字化けや文字切れを避けることが可能である．

本稿では，文字セット及び符号化方式に関する基本的な情報を整理するとともに，主にレガシーデータ変換時に発生するトランスコーディングに関する問題について考察し，PMDA の規制要件に合致しないデータを検出・変換する方法を提案する．さらに，統合解析用データの作成等を目的として，Shift-JIS 及び Wlatin1 で作成された SAS データセットを効率的に UTF-8 にトランスコーディングする方法を提案する．

## 2 文字セットと符号化方式

コンピュータ上のデータはビット (bits, binary digits) とバイト (bytes) で構成される．ビットは情報の最小単位で，"0"と"1"の 2 進数である．バイトは 8 ビットを 1 単位としたものである．文字セットとは，実データ上で表現する 1 文字ごとにユニークな非負整数を割り当てたマッピング表である．例えば，「あ」の Unicode 値は「U+3042」である．このように文字と数字の個々の対応をコードポイントという．符号化方式は，個々のコードポイントをコンピュータに格納する方法を定義したものである．例えば，Unicode 値の「U+3042」を UTF-8 で符号化すると「0xE3 0x81 0x82」の 3 バイト列，UTF-16 で符号化すると「0x30 0x42」の 2 バイト列となる．このように，実データ上で表現される文字は，文字セットと符号化方式の組合せによって構成される．表 2.1 に，SAS データセット作成時に主に使用される文字セットと符号化方式の関係，符号化方式の SAS での指定名及び最大バイト数を示した．

| 文字セット | 符号化方式 | SAS で指定する名称 (短縮名) | SAS 上の最大バイト数 |
|---|---|---|---|
| JISX0208 | Shift-JIS | shift-jis (sjis) | 2 |
| | EUC | euc-jp (jeuc) | 4 |
| Unicode | UTF-8 | utf-8 (utf8) | 4 |
| | UTF-16BE | utf-16b | 2 |
| | UTF-16LE | utf-16l | 2 |
| Extended ASCII | Code Page 1252 (Windows Latin-1) | wlatin1 (wlt1) | 1 |

表 2.1 文字セットと符号化方式の関係

PMDA への電子データ提出に係る規制要件は，CDISC 標準に準拠することだけではない．文字セットに関する要件は，SAS プログラマがレガシーデータを CDISC 標準に準拠した形に変換する際に，予期せぬ不具合に遭遇することがある．本章では文字セットに係る PMDA の規制要件と，それに係る文字セット及び符号化方式に関する基本的事項を整理するとともに，実務におけるトランスコーディングに関する問題について考察する．

## 2.1 PMDA の規制要件

日本では，PMDA が承認申請時の電子データ提出等に関する技術的ガイド (平成 28 年 8 月 24 日付け薬機次発第 0824001 号 独立行政法人医薬品医療機器総合機構次世代審査等推進室長通知により改正，以下「技術的ガイド」と記載) において，英数字等の ASCII で規定されている文字セットのみで構成されたデータセットを求めており，PMDA Study Data Validation Rules (V1.0, 2015/11/18) に非 ASCII 文字及び非印字可能文字 (ASCII コードの 32 から 126 以外の値) が変数の値に含まれる場合は Warning となるルール (SDTM Rules V1.0, RULE ID: SD1029) を設けている．また米国でも，FDA が FDA Business Rules (Version 1.1, finalized March 2017) において，PMDA と同一のルール (Rule ID: FDAB041) を設けている．

PMDA，FDA ともに符号化方式の指定はなく，PMDA については技術的ガイドに従いデータガイドに電子データで使用した文字セット又は符号化方式の情報を記載するのみで問題ない．なお，PMDA の申請電子データに関する FAQ Q4-19 には申請時に提示すべき文字セット及び符号化方式の情報の具体例として表 2.1.1 に示す内容が記載されており，これらのいずれかを使用しておけば PMDA と申請者側とで特に議論が必要となることはないと考えられる．

| | |
|---|---|
| 【文字セット】JISX0208 | 【符号化方式】Shift-JIS |
| 【文字セット】JISX0208 | 【符号化方式】EUC-JP |
| 【文字セット】UNICODE (USC-2) | 【符号化方式】UTF-8 |

表 2.1.1 申請電子データに関する FAQ における文字セット及び符号化方式の情報の具体例

## 2.2 文字セット ASCII

PMDA の文字セットに関する規制要件である ASCII は，7 ビットで表現できる非負整数の 0 から 127 の数字に対して，大小のアルファベットや数字，英文でよく使用される記号等を割り当てた文字セットである．0 から 31 番目と 127 番目は制御文字と呼ばれ，改行・後退・タブなどの文字を含む．32 から 126 番目は印字可能文字と呼ばれ，32 番目は半角スペース，48 から 57 番目は数字，65 から 90 番目と 97 から 122 番目は大小アルファベット，これら以外は各種記号となっている．参考までに，表 2.2.1 に ASCII の 32 から 126 番目までの文字を示した．

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 |
| (space) | ! | " | # | $ | % | & | ' | ( | ) | * | + | , | - | . | / | 0 | 1 | 2 | 3 |
| 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 |
| 4 | 5 | 6 | 7 | 8 | 9 | : | ; | < | = | > | ? | @ | A | B | C | D | E | F | G |
| 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 |
| H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | [ |
| 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 | 111 |
| ¥ | ] | ^ | _ | ` | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o |
| 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 | 121 | 122 | 123 | 124 | 125 | 126 | | | | | |
| p | q | r | s | t | u | v | w | x | y | z | { | \| | } | ~ | | | | | |

表 2.2.1 ASCII 一覧

1972 年に ISO-646 として国際標準化された当時，既に様々な言語がコンピュータで使用されており，128 から 256 番目を追加した 8 ビットの拡張 ASCII が多数開発された．これらの拡張 ASCII に対応した符号化方式が ISO 8859 として標準化されている．ISO 8859 は地域ごとに ISO 8859-1～ISO 8859-16 (ISO 8859-12 は廃棄) の 15 種類が存在する．これまでの経験から，英語圏の国や地域で作成された電子データは符号化方式として ISO 8859-1 (Windows Latin-1) を採用していることが多いと思われるが，これらの

電子データを PMDA に提出する際は，日本語データは含まれないものの，非 ASCII 文字及び非印字可能文字 (ASCII コードの 32 から 126 以外の値) が含まれていないか確認した方がよいと思われる．

## 2.3 文字セット Unicode と符号化方式 UTF-8

レガシーデータの符号化方式には，それを作成した SAS 環境におけるデフォルトのセッションエンコーディングがそのまま適用されていることが多い．デフォルトのセッションエンコーディングはオペレーティングシステム (OS) 及びその言語環境 (Locale) の組合せによって異なる．表 2. 31 に デフォルトのセッションエンコーディングを抜粋して示した．

| Locale＼OS | Windows | UNIX | z/OS |
|---|---|---|---|
| English_US (en_US) | WLatin1 (wlt1) | Latin1 | open_ed- D47 |
| Japanese_Japan (jp_JP) | Shift-JIS (sjis) | Shift-JIS, EUC-JP 等 | ibm-939 |

表 2. 31 デフォルトのセッションエンコーディング

(SAS 9. 4NLS Reference Guide, Fifth Edition より抜粋)

例えば符号化方式が Shift-JIS や Wlatin1 であるレガシーデータについて，符号化方式をそのいずれかに統一することを考えた場合，どちらか一方でしか表現できない文字が存在する (例えば日本語は Wlatin1 では表現できない) ためどちらも選択することができない．

Unicode は，日本語を含む世界で使用されるほぼすべての言語における文字に対して，ユニークなコードを割り当てた文字セットである．Unicode の最初の 128 コードは ASCII と，最初の 256 コードは ISO8859- 1と同一である．また，日本語文字セットの JISX0208 をはじめ，一般的に使用されている多くの文字セットとの互換性がある．UTF-8 は，Unicode で使用できる 8 ビットの符号化方式であり，ASCII に対応する部分は 1 バイト，それ以外の部分は 2 から 4 バイトとなっており，Shift-JIS 及び Wlatin1 に対してバイト数の違いはあるものの互換性がある．一方，Shift-JIS と Wlatin1 の間には，ASCII 部分以外に互換性はない．表 2. 3. 2 に,UTF-8, Shift-JIS 及び Wlatin1 の各符号化方式の対応を示した．

| 符号化方式 | | UTF-8 | Shift-JIS | Wlatin1 |
|---|---|---|---|---|
| 文字セット | | Unicode | JISX0208 | Extended ASCII |
| 日本語表示 | | ○ | ○ | × |
| バイト数 | ASCII 部分<br>(英数字，記号) | 1 バイト | 1 バイト | 1 バイト |
| | ASCII 以外 | 2～4 バイト | 1～2 バイト | 1 バイト |
| | 平仮名，片仮名 | 3 バイト | 2 バイト | × |
| | 漢字 | 3 バイト<br>(一部 4 バイト) | 2 バイト | × |
| | 拡張 ASCII の 128 番目以降<br>(ü, Ä, ®など) | 2 バイト | × | 1 バイト |

表 2. 3. 2各符号化方式の比較

グローバル製薬企業は，特にレガシーデータにおいては，英語以外の言語を含む電子データを多く保有していると思われる．しかしながら，レガシーデータの符号化方式ごとに，それを取扱う SAS 環境を使い分けることは現実的ではない．SAS9. 4からは多くの言語における文字に対応した UTF-8 がセッションエンコーディングとなっている SAS 環境 (Unicode サポート) が利用できるようになり，符号化方式が Wlatin1 や Shift-JIS である SAS データセットを同時に取扱うことができるようになった．

Unicode サポートを使用すると，SAS のクロス環境データアクセス (CEDA: Cross-Environment Data Access) 機能によって，読み込んだ SAS データセットの符号化方式が UTF-8 以外の場合は，自動的にトランスコーディングが発生する．しかしながら，トランスコーディングによって 1 文字あたりのバイト数も変化するため，必要に応じて変数の長さを拡張する必要があるが，変数の長さが格納された値のバイト数に比べて不足する場合は，エラーが発生する．表 2.3.3 に SAS データセット読込時に各変数の長さを一律で調整する LIBNAME ステートメントのオプションを示した．このいずれかのオプションを用いることで，エラーの回避が可能である．

| 基本構文： | |
|---|---|
| LIBNAME ライブラリ名 <オプション>"ライブラリのパス"; | |
| オプション： | |
| CVPBYTES=数字 | ：トランスコーディング時に各変数で拡張するバイト数を指定する．例えば，「CVPBYTES=5」を指定すると，変数の長さが 5 の場合は 10 に，10 の場合は 15 になる |
| CVPMULTIPLIER=数字 | ：トランスコーディング時に各変数で拡張する乗数値を指定する．例えば，「CVPMULTIPLIER=1.5」を指定すると，変数の長さが 10 の場合は 15 になる |

表 2.3.3 トランスコーディング時に使用する LIBNAME ステートメントのオプション

## 2.4 非 ASCII 文字及び非印字可能文字の検出と変換

非 ASCII 文字及び非印字可能文字を検出するために，Srivastava (2017) を参考に検出結果を Excel ファイルに出力するマクロプログラムを作成した．本マクロプログラムの基本的な処理は，DATA ステップ関数の一つである BYTE 関数を用いて ASCII 文字のリストを作成し，そのリストに含まれない文字が各変数値に含まれる場合に非 ASCII 文字及び非印字可能文字として検出するようにしている．表 2.4.1 に本マクロプログラムの実行に必要な各マクロパラメータの引数を示した．なお，本マクロプログラムの詳細な処理内容は付録のソースコードを参照されたい．

| マクロ名<br>マクロパラメータ | ：%DetectNonPriASCII<br>(DataFolder=, KeyVariables=, CVPMult=) ; |
|---|---|
| DataFolder | ：検出対象の SAS データセットを格納したフルパス (必須) |
| KeyVariables | ：検出結果とともに表示させる被験者番号等のキー変数 (オプション)<br>(複数指定の場合は半角スペース区切り) |
| CVPMult | ：各変数の長さを一律拡張するための乗数値 (オプション) |

表 2.4.1 非 ASCII 文字及び非印字可能文字の検出マクロプログラム

本マクロプログラムは，セッションエンコーディングが SAS データセットの符号化方式と同一の SAS 環境又は UTF-8 環境で実行されることを想定している．UTF-8 環境で実行する場合は，トランスコーディング (符号化方式の変換) によって変数値のバイト数が増加する可能性がある．そこで，本マクロプログラムでは，検出結果の文字化け及び文字切れを防ぐため，すべての文字変数の長さを一律で拡張する機能を設けており，変数の長さに対する乗数値を CVPMult に指定することができる．例えば Shift-JIS 環境で 2 バイトの文字は，UTF-8 環境では最大で 3 又は 4 バイトとなるため「2」を指定しておけば文字切れが発生しない．

本マクロプログラムが出力する Excel ファイルは，データセットごとにシートが作成される．実行結果イメージについて，SAS データセットの符号化方式が Wlatin1 の場合を図 2.4.2，Shift-JIS の場合を図 2.4.3 に示した．

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | OBS | STUDYID | USUBJID | OBS | MEMNAME | NAME | ORGVALUE | DetectedString | CandidateValue |
| 2 | 1 | ABC-123 | ABC-123-1001-001 | 2 | ADAE | AETERM | NEUTROPENIA (0.86 MILLE/MM³) | ³ | |
| 3 | 2 | ABC-123 | ABC-123-1001-002 | 3 | ADAE | AETERM | BAKER´S CYST | ´ | |
| 4 | 3 | ABC-123 | ABC-123-1002-001 | 5 | ADAE | AETERM | ANÉMIA | É | |
| 5 | 4 | ABC-123 | ABC-123-1003-001 | 7 | ADAE | AETERM | INFECTION IN OUTER EAR ( PLUS DRANÆGE | Æ | |
| 6 | 5 | ABC-123 | ABC-123-1003-002 | 9 | ADAE | AETERM | DERMO-HYPODERMITIS ON LEFT WRIST | | |
| 7 | 6 | ABC-123 | ABC-123-1003-003 | 11 | ADAE | AETERM | ASTHÉNIA | É | |
| 8 | 7 | ABC-123 | ABC-123-1003-004 | 12 | ADAE | AETERM | ÖDEMA LOWER LIMBS | Ö | |
| 9 | 8 | ABC-123 | ABC-123-1003-004 | 13 | ADAE | AETERM | ÖDEMA RIGHT FOOT | Ö | |

図 2.4.2 SAS データセットの符号化方式が Wlatinl の場合の検出結果イメージ (Excel ファイル)

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | OBS | STUDYID | USUBJID | OBS | MEMNAME | NAME | ORGVALUE | DetectedString | CandidateValue |
| 2 | 1 | ABC-123 | ABC-123-1001-001 | 1 | ADAE | AETERM | 左足親指の傷感染 | 左足親指の傷感染 | |
| 3 | 2 | ABC-123 | ABC-123-1001-001 | 2 | ADAE | AETERM | 好中球減少 (0.86 MILLE/MM^3) | 好中球減少 | |
| 4 | 3 | ABC-123 | ABC-123-1001-002 | 3 | ADAE | AETERM | ベーカー嚢胞 | ベーカー嚢胞 | |
| 5 | 4 | ABC-123 | ABC-123-1002-001 | 4 | ADAE | AETERM | 下肢のうずき | 下肢のうずき | |
| 6 | 5 | ABC-123 | ABC-123-1002-001 | 5 | ADAE | AETERM | 貧血 | 貧血 | |
| 7 | 6 | ABC-123 | ABC-123-1003-001 | 6 | ADAE | AETERM | ステロイド誘発障害 | ステロイド誘発障害 | |
| 8 | 7 | ABC-123 | ABC-123-1003-001 | 7 | ADAE | AETERM | 外耳感染症 | 外耳感染症 | |
| 9 | 8 | ABC-123 | ABC-123-1003-002 | 8 | ADAE | AETERM | 右手の痙攣 | 右手の痙攣 | |
| 10 | 9 | ABC-123 | ABC-123-1003-002 | 9 | ADAE | AETERM | 左手首の皮膚皮下組織炎 | 左手首の皮膚皮下組織炎 | |
| 11 | 10 | ABC-123 | ABC-123-1003-003 | 10 | ADAE | AETERM | 便秘 | 便秘 | |
| 12 | 11 | ABC-123 | ABC-123-1003-003 | 11 | ADAE | AETERM | 衰弱 | 衰弱 | |
| 13 | 12 | ABC-123 | ABC-123-1003-004 | 12 | ADAE | AETERM | 下肢の浮腫 | 下肢の浮腫 | |
| 14 | 13 | ABC-123 | ABC-123-1003-004 | 13 | ADAE | AETERM | 右足の浮腫 | 右足の浮腫 | |

図 2.4.3 SAS データセットの符号化方式が Shift-JIS の場合の検出結果イメージ (Excel ファイル)

A 列から順に，検出結果の 1 からの連番 (A 列 : Obs)，検出値が含まれるキー変数 (B 列及び C 列 : KeyVariables として STUDYID 及び USUBJID を指定)，オブザベーション番号 (D 列 : OBS)，SAS データセット名 (E 列 : MEMNAME)，変数名 (F 列 : NAME)，検出値が含まれる変数の値全体 (G 列 : ORGVALUE)，検出値 (H 列 : DetectedString) となっている．図 2.4.2 では変数 AETERM において非 ASCII 文字がいくつか検出されている．また，図 2.4.3 は有害事象の医師記載名を日本語で取得していたため，変数 AETERM について全レコードで非 ASCII 文字として検出されている．なお，I 列の CandidateValue は後述する検出値の置換に使用する．DetectedString がブランクの場合は非印字可能文字が検出される．

続いて，マクロプログラム DetectNonPriASCII による検出結果について，該当する検出値が含まれる変数値全体を，ASCII に適合した値に変換するマクロプログラムを作成した なお，本マクロプログラムの実行前に，検出結果 (図 2.4.2 及び図 2.4.3 における I 列 : CandidateValue) に変換候補の値を入力しておく必要がある．図 2.4.3 のように日本語データが含まれる場合，PMDA への電子データ提出においては，実務的通知に従い「JAPANESE TEXT IN SOURCE DATABASE」等を CandidateValue に入力する．表 2.4.4 に本マクロプログラムの実行に必要な各マクロパラメータの引数を示した．なお，本マクロプログラムの詳細な処理内容は付録のソースコードを参照されたい．

| | |
|---|---|
| マクロ名<br>マクロパラメータ | : %ReplaceNonAscii<br>(InDataFolder=, OutDataFolder=, InExcelFile=,<br>InExcelSheet=, OutEncoding=, CVPMult=) ; |
| InDataFolder | : 変換対象の SAS データセットを格納したフルパス (必須) |
| OutDataFolder | : 変換後の SAS データセットを出力するフルパス (必須) |
| InExcelFile | : 変換候補の値を追記したファイルのフルパス (必須) |
| InExcelSheet | : 変換候補の値を追記したファイルにおけるシート名 (必須) |
| OutEncoding | : 変換後の符号化方式 (必須) |
| CVPMult | : 各変数の長さを一律拡張するための乗数値 (オプション) |

表 2.4.4 非 ASCII 文字及び非印字可能文字を ASCII 文字に変換するマクロプログラム

本マクロプログラムもまた，マクロプログラム DetectNonPriASCII 同様，セッションエンコーディン

グが変換対象の SAS データセットの符号化方式と同一の SAS 環境又は UTF-8 環境で実行されることを想定している．また，変換後の SAS データセットについては，Outencoding に指定した符号化方式が適用され，各変数の長さはデータ値の最大長に自動調整される．

図 2.4.5 に本マクロプログラムを実行した場合の出力結果のイメージを示したが，変数 AETERM の値が「JAPANESE TEXT IN SOURCE DATABASE」に置換されていることが分かる．

| | STUDYID | USUBJID | AESEQ | AETERM | AEDECOD |
|---|---|---|---|---|---|
| 1 | ABC-123 | ABC-123-1001-001 | 1 | JAPANESE TEXT IN SOURCE DATABASE | Wound infection |
| 2 | ABC-123 | ABC-123-1001-001 | 2 | JAPANESE TEXT IN SOURCE DATABASE | Neutropenia |
| 3 | ABC-123 | ABC-123-1001-002 | 1 | JAPANESE TEXT IN SOURCE DATABASE | Synovial cyst |
| 4 | ABC-123 | ABC-123-1002-001 | 1 | JAPANESE TEXT IN SOURCE DATABASE | Paraesthesia |
| 5 | ABC-123 | ABC-123-1002-001 | 2 | JAPANESE TEXT IN SOURCE DATABASE | Anaemia |

図 2.4.5 ASCII 文字への変換結果イメージ (SAS データセット)

## 2.5 トランスコーディングの必要性

PMDA への電子データ提出を目的としてレガシーデータ変換を行う際には，SAS データセットの構造を CDISC 標準に準拠するように変換し，ASCII の 32 から 126 番目に該当する文字のみを格納すれば規制要件を満たすため，必ずしもトランスコーディングは必要ではない．しかしながら，統合解析用の SAS データセットを作成する際や社内の統一されたシステム環境において解析を行う際にも，トランスコーディングが必要となる可能性があるため，複数の臨床試験データを PMDA に提出する場合，社内の業務手順やポリシーによっては，すべての電子データの符号化方式を一律 UTF-8 に統一しておくことも選択肢の一つとなる．2.3 項で述べたように，異なる符号化方式のデータの読込時には自動的にトランスコーディングが発生し，プログラム実行時のパフォーマンスが低下する可能性があることから，結果として SAS プログラマの作業時間の増大に繋がる可能性がある．上述した状況を想定し，符号化方式を UTF-8 にトランスコーディングするマクロプログラム (各変数の長さはデータの最大長に自動調整) を作成した．表 2.5.1 にマクロプログラムの使用方法を記載するが，その詳細は付録を参照されたい．

| | |
|---|---|
| マクロ名<br>マクロパラメータ | : %Transcoding2UTF8 (InDataFolder=, OutDataFolder=, Outencoding=) ; |
| InDataFolder | : 変換対象の SAS データセットを格納したフルパスを指定 (必須) |
| OutDataFolder | : 変換後の SAS データセットを出力するフルパス (必須) |
| Outencoding | : 変換後の符号化方式 (必須) |

表 2.5.1 トランスコーディングマクロプログラム

# 3 非 ASCII 文字及び非印字可能文字の検出・変換ツールの開発事例

本項では，2 項で紹介した非 ASCII 文字及び非印字可能文字の検出，並びにそれらの ASCII 文字への変換をサポートするツールの開発事例を紹介する．本稿では HTML アプリケーション (HTA) 及び SAS を用いてツールを開発した．HTA は Windows 上で動作可能なアプリケーションで，HTML 及び Java Script 等の機能を利用して手軽に GUI 画面を作成することができる．HTA の基本的な機能及び SAS との連携に関する詳細は舟尾・高浪 (2008)及び高浪 (2012) を参照されたい．

## 3.1 テストデータ

本項では，非 ASCII 文字及び非印字可能文字の検出・変換ツールのテストデータとして，米国及び日本でそれぞれ実施された 2 つの仮想の臨床試験から得られた，有害事象に関する解析用データセット ADAE を用いる．また，各 SAS データセットの符号化方式は，米国実施分は Wlati n1，日本実施分は Shi ft-JIS とした．なお，本テストデータは CDISC ADaM OCCDS 形式を参考にしたが，本ツールは CDISC 標準に準拠していなくても動作可能である．

| STUDYID | USUBJID | AESEQ | AETERM | ・・・ |
|---|---|---|---|---|
| ABC-123 | ABC-123-1001-001 | 1 | WOUND INFECTION LEFT BIG TOE | ・・・ |
| ABC-123 | ABC-123-1001-001 | 2 | NEUTROPENIA (0.86 MILLE/MM³) | ・・・ |
| ABC-123 | ABC-123-1001-002 | 1 | BAKER'S CYST | ・・・ |
| ABC-123 | ABC-123-1002-001 | 1 | TINGLE OF LOWER LIMBS | ・・・ |
| ABC-123 | ABC-123-1002-001 | 2 | ANÉMIA | ・・・ |
| ・・・ | ・・・ | ・・・ | ・・・ | ・・・ |

図 3.1.1 米国で実施された仮想の臨床試験で得られた有害事象に関する解析用データセット ADAE

| STUDYID | USUBJID | AESEQ | AETERM | ・・・ |
|---|---|---|---|---|
| ABC-123 | ABC-123-1001-001 | 1 | 左足親指の傷感染 | ・・・ |
| ABC-123 | ABC-123-1001-001 | 2 | 好中球減少 (0.86 MILLE/MM^3) | ・・・ |
| ABC-123 | ABC-123-1001-002 | 1 | ベーカー嚢胞 | ・・・ |
| ABC-123 | ABC-123-1002-001 | 1 | 下肢のうずき | ・・・ |
| ABC-123 | ABC-123-1002-001 | 2 | 貧血 | ・・・ |
| ・・・ | ・・・ | ・・・ | ・・・ | ・・・ |

図 3.1.2 日本で実施された仮想の臨床試験で得られた有害事象に関する解析用データセット ADAE

## 3.2 ツールの構成

ここでは，本ツールの画面・機能等の構成について説明する．本ツールの実行に必要なファイル及びフォルダは表 3.2.1 の構成で設置する．任意の場所に「実行ファイル.hta」及び files フォルダを設置する．また，file フォルダに設置する TranscodingMacro.sas は，表 2.4.1，表 2.4.4 及び表 2.5.1 で紹介した SAS マクロプログラムを 1 ファイルにまとめたものとなっている．検出・変換対象の SAS データセットを格納するフォルダ，及び検出・変換結果の SAS データセットを格納するフォルダは任意のフォルダを指定可能である．ツール実行時に，_EXE.log がログファイルとして，_EXE.sas 及び parameter.sas が実行時の一時ファイルとして，DetectNonPri ASCIIxlsx が検出結果ファイルとして，いずれも files フォルダに出力される．

| 実行ファイル.hta | ：本ファイルをダブルクリックするとツールが起動 | |
|---|---|---|
| files フォルダ | _EXE.log | ：実行時のログ<br>※実行時に自動作成される |
| | _EXE.sas | ：実行用の一時ファイル (各 SAS プログラムを%INCLUDE ステートメントで実行)<br>※実行時に自動作成される |
| | parameter. sas | ：実行用の一時ファイル (画面から与えた情報を SAS マクロ変数化)<br>※実行時に自動作成される |

| | | |
|---|---|---|
| | Transcoding Macro.sas | ：非 ASCII 文字及び非印字可能文字の検出・変換，符号化方式の変換を行う SAS マクロ．本稿で作成した以下の SAS マクロプログラムを使用<br>・DetectNonPriASCII<br>・ReplaceNonAscii<br>・Transcoding2UTF8 |
| | DetectNonPriASCII.xlsx | ：非 ASCII 文字及び非印字可能文字の検出結果ファイル，また変換候補文字を入力するファイル<br>※実行時に自動作成される |
| INDATA フォルダ | ：検出・変換対象の SAS データセットを格納するフォルダ<br>※任意の場所を指定可能 | |
| OUTDATA フォルダ | ：検出・変換結果の SAS データセットが格納されるフォルダ<br>※任意の場所を指定可能 | |

表 3. 2. 1 ツールの実行に必要なファイル及びフォルダ

続いて，本ツールの画面を図 3. 2. 2 に示す．

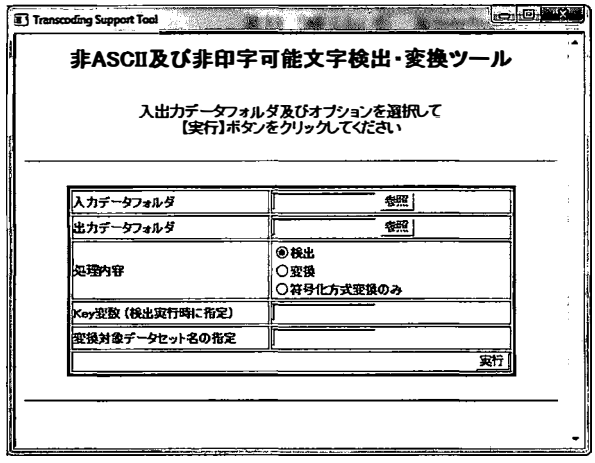



図 3. 2. 2ツールの入力画面

画面からの各入力パラメータとそれらの内容及び対応する SAS マクロ変数を表 3. 2. 3に示す．

| 入力パラメータ | 概要 | 対応する SAS マクロ変数 |
|---|---|---|
| 入力データフォルダ | 検出・変換対象の SAS データセットが格納されているフォルダ．「参照」ボタンをクリックしてフォルダを選択 | INDATA |
| 出力データフォルダ | 検出・変換結果の SAS データセットが格納されているフォルダ．「参照」ボタンをクリックしてフォルダを選択 | OUTDATA |
| 処理内容 | 非 ASCII 文字及び非印字可能文字の検出を行う場合は「検出」，非 ASCII 文字及び非印字可能文字から ASCII 文字への変換を行うとともに符号化方式を UTF-8 に変換する場合は「変換」，符号化方式を UTF-8 に変換する場合は「符号化方式変換のみ」を選択 | _Processing<br>1：検出<br>2：変換<br>3：符号化方式変換のみ |
| Key 変数 | 検出結果とともに表示する Key 変数を指定 (STUDYID, USUBJID 等) | _KeyVariables |
| 変換対象データセット名 | 変換対象の SAS データセットを指定 | _Datasets |

表 3.2.3 入力パラメータと SAS マクロ変数

本ツールの処理の流れを図 3.2.4 に示す.

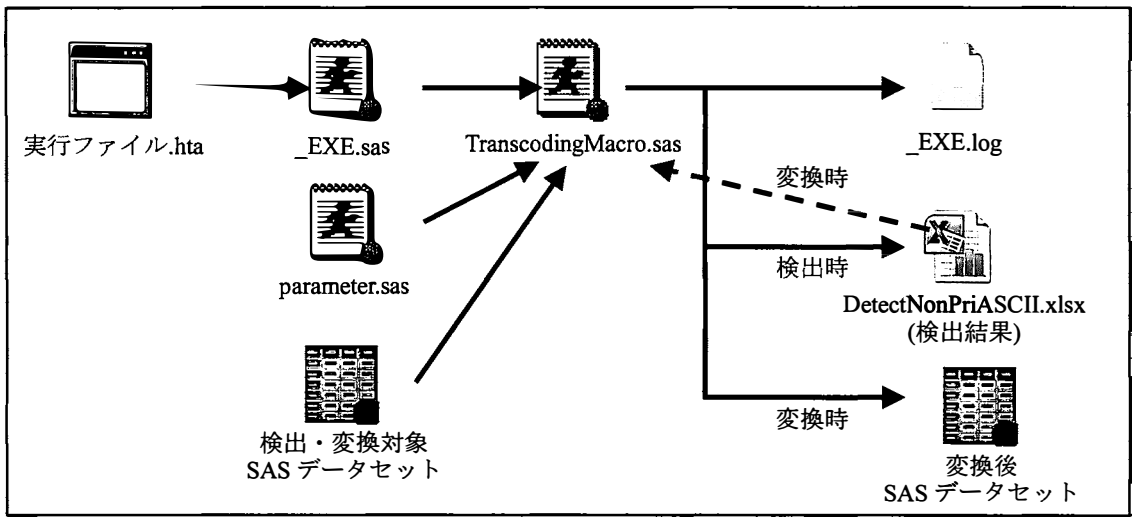



図 3. 2. 処理の流れ

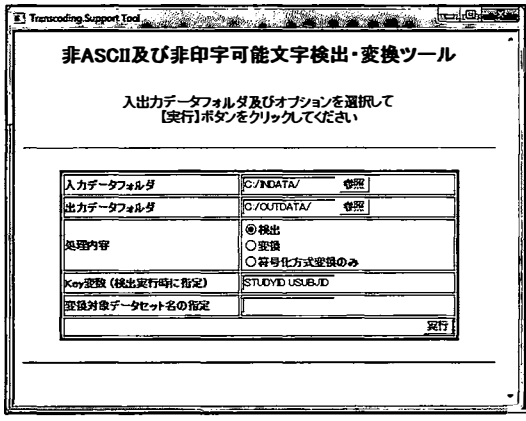
## 3.3 ツールの実行

本ツールはバッチモードで処理を実行する. 最初に「実行ファイル. hta」をテキストエディタ等で開き, 図 3.3.1 に示す HTA アプリケーションファイルにおける SAS の実行可能ファイル(.exe)及び環境設定ファイル(.cfg)のパスを SAS のインストール環境に合わせて変更する. なお, 環境設定ファイルは, SAS (Unicode サポート) が使用されるように必ず「u8」が含まれるパスを指定する.

```
// _EXE.sas の作成
  MyText = MyScript.CreateTextFile("./files/_EXE.sas", true) ;
  MyText.WriteLine("%inc '" + MyPath + "¥¥files¥¥parameter.sas' ;") ;
  MyText.WriteLine("%inc '" + MyPath + "¥¥files¥¥TranscodingMacro.sas' ;") ;
  MyText.Close() ;
// Shell 関連の操作を提供するオブジェクトを取得
  MyShell = new ActiveXObject("WScript.Shell") ;
// SAS の実行
MyShell.Exec('"C:/Program Files/SAS_9.4/SASFoundation/9.4/sas.exe"
              -CONFIG "C:/Program Files/SAS_9.4/SASFoundation/9.4/nls/u8/sasv9.cfg"
              -sysin "./files/_EXE.sas" -log "./files/_EXE.log" -nosplash -icon') ;
```

図 3.3.1 HTA アプリケーションファイルにおける SAS プログラムの実行部分

本ツールの画面で実際にパラメータを入力し, 検出・変換結果を出力する. ここでは図 3. 1. 2 に示した SAS データセット ADAE (日本で実施された仮想の臨床試験で得られた有害事象) をテストデータとして用いる. 最初に, 非 ASCII 及び非印字可能文字の検出を行う. 入力データフォルダ及び出力データフォルダをそれぞれ選択し, 処理内容として「検出」を選択する. また, Key 変数として STUDYID と USUBJID を入力する. これらの情報を選択・入力した画面及び生成された parameter.sas を図 3.3.2 に示す. なお, 本ツールは「C:¥Test」に設置し, TranscodingMacro.sas を「C:¥Test¥files」に設置した.




```
%let _CPATH = C:¥Test ;
%let _PPATH = C:¥Test¥files¥TranscodingMacro.sas ;
%let _Processing = 1 ;
%let _INDATA  = C:/INDATA/ ;
%let _OUTDATA = C:/OUTDATA/ ;
%let _KeyVariable = %str(STUDYID USUBJID) ;
%let _Datasets    = ;
```

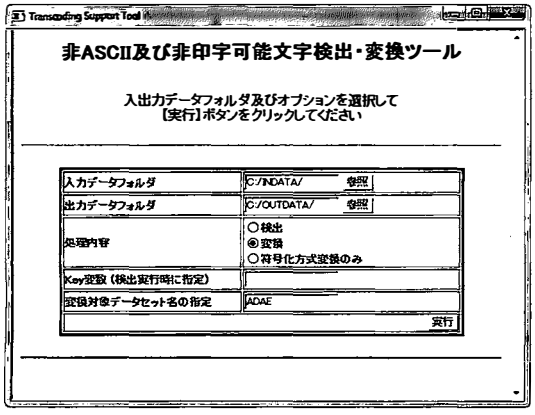
図 3.3.2 検出時のツールの画面及び生成された parameter.sas

各情報の選択・入力後，実行ボタンを押下することで図 3.2.4 に示した流れで処理が実行され，file フォルダに DetectNonPriASCII.xlsx が検出結果ファイルとして出力される．検出結果は図 2.4.3 と同一であるためここでは割愛する．次に，非 ASCII 及び非印字可能文字の変換を行う．図 2.4.3 に示した検出結果のファイルに変換候補の文字を入力したものを図 3.3.3 に示す．ここでは PMDA の実務的通知に従い，「JAPANESE TEXT IN SOURCE DATABASE」を入力した．

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | OBS | STUDYID | USUBJID | OBS | MEMNAME | NAME | ORGVALUE | DetectedString | CandidateValue |
| 2 | 1 | ABC-123 | ABC-123-1001-001 | 1 | ADAE_SJIS | AETERM | 左足親指の傷感染 | 左足親指の傷感染 | JAPANESE TEXT IN SOURCE DATABASE |
| 3 | 2 | ABC-123 | ABC-123-1001-001 | 2 | ADAE_SJIS | AETERM | 好中球減少 (0.86 MILLE/MM^3) | 好中球減少 | JAPANESE TEXT IN SOURCE DATABASE |
| 4 | 3 | ABC-123 | ABC-123-1001-002 | 3 | ADAE_SJIS | AETERM | ベーカー嚢胞 | ベーカー嚢胞 | JAPANESE TEXT IN SOURCE DATABASE |
| 5 | 4 | ABC-123 | ABC-123-1002-001 | 4 | ADAE_SJIS | AETERM | 下肢のうすき | 下肢のうすき | JAPANESE TEXT IN SOURCE DATABASE |
| 6 | 5 | ABC-123 | ABC-123-1002-001 | 5 | ADAE_SJIS | AETERM | 貧血 | 貧血 | JAPANESE TEXT IN SOURCE DATABASE |
| 7 | 6 | ABC-123 | ABC-123-1003-001 | 6 | ADAE_SJIS | AETERM | ステロイド誘発障害 | ステロイド誘発障害 | JAPANESE TEXT IN SOURCE DATABASE |
| 8 | 7 | ABC-123 | ABC-123-1003-001 | 7 | ADAE_SJIS | AETERM | 外耳感染症 | 外耳感染症 | JAPANESE TEXT IN SOURCE DATABASE |
| 9 | 8 | ABC-123 | ABC-123-1003-002 | 8 | ADAE_SJIS | AETERM | 右手の痙攣 | 右手の痙攣 | JAPANESE TEXT IN SOURCE DATABASE |
| 10 | 9 | ABC-123 | ABC-123-1003-002 | 9 | ADAE_SJIS | AETERM | 左手首の皮膚皮下組織炎 | 左手首の皮膚皮下組織炎 | JAPANESE TEXT IN SOURCE DATABASE |
| 11 | 10 | ABC-123 | ABC-123-1003-003 | 10 | ADAE_SJIS | AETERM | 便秘 | 便秘 | JAPANESE TEXT IN SOURCE DATABASE |
| 12 | 11 | ABC-123 | ABC-123-1003-003 | 11 | ADAE_SJIS | AETERM | 衰弱 | 衰弱 | JAPANESE TEXT IN SOURCE DATABASE |
| 13 | 12 | ABC-123 | ABC-123-1003-004 | 12 | ADAE_SJIS | AETERM | 下肢の浮腫 | 下肢の浮腫 | JAPANESE TEXT IN SOURCE DATABASE |
| 14 | 13 | ABC-123 | ABC-123-1003-004 | 13 | ADAE_SJIS | AETERM | 右足の浮腫 | 右足の浮腫 | JAPANESE TEXT IN SOURCE DATABASE |

図 3.3.3 検出結果に変換候補の文字を入力したファイル (Excel ファイル)

入力データフォルダ及び出力データフォルダをそれぞれ選択し，処理内容として「変換」を選択する．また，変換対象データセット名として ADAE を入力する．これらの情報を選択・入力した画面及び生成された parameter.sas を図 3.3.4 に示す．




```
%let _CPATH = C:¥Test ;
%let _PPATH = C:¥Test¥files¥TranscodingMacro.sas ;
%let _Processing = 2 ;
%let _INDATA  = C:/INDATA/ ;
%let _OUTDATA = C:/OUTDATA/ ;
%let _KeyVariable =  ;
%let _Datasets    = ADAE ;
```

図 3.3.4 変換時のツールの画面及び生成された parameter.sas

各情報の選択・入力後，実行ボタンを押下することで図 3.2.4 に示した流れで処理が実行され，出力

データフォルダに変換結果の SAS データセットが出力される．変換結果は図 2.4.5 と同一であるためここでは割愛する．図 3.3.5 に変換前後の SAS データセットの情報を示すが，SAS データセットの符号化方式が Shift-JIS から UTF-8 に変更されており，変数の長さはデータの最大長に調整されている．

| | |
|---|---|
| データセット名 | INDATA.ADAE |
| メンバータイプ | DATA |
| エンジン | V9 |
| 作成日時 | 2017/06/20 09:22:34 |
| 更新日時 | 2017/06/20 09:22:34 |
| 保護 | |
| データセットタイプ | |
| ラベル | |
| データ表現 | WINDOWS 32 |
| エンコード | shift-jis Japanese (SJIS) |

| # | 変数 | タイプ | 長さ | ラベル |
|---|---|---|---|---|
| 1 | STUDYID | 文字 | 7 | Study Identifier |
| 2 | USUBJID | 文字 | 16 | Unique Subject Identifier |
| 3 | AESEQ | 数値 | 8 | Sequence Number |
| 4 | AETERM | 文字 | 100 | Reported Term for the Adverse Event |
| 5 | AEDECOD | 文字 | 25 | Dictionary-Derived Term |
| 6 | AEBODSYS | 文字 | 52 | Body System or Organ Class |
| 7 | TRTEMFL | 文字 | 1 | Treatment Emergent Analysis Flag |
| 8 | AESTDTC | 文字 | 10 | Start Date/Time of Adverse Event |
| 9 | AEENDTC | 文字 | 10 | End Date/Time of Adverse Event |
| 10 | AESER | 文字 | 1 | Serious Event |
| 11 | AESEV | 文字 | 8 | Severity/Intensity |
| 12 | AEREL | 文字 | 11 | Causality |

| | |
|---|---|
| データセット名 | OUTDATA.ADAE |
| メンバータイプ | DATA |
| エンジン | V9 |
| 作成日時 | 2017/06/20 13:52:10 |
| 更新日時 | 2017/06/20 13:52:10 |
| 保護 | |
| データセットタイプ | |
| ラベル | |
| データ表現 | WINDOWS 32 |
| エンコード | utf-8 Unicode (UTF-8) |

| # | 変数 | タイプ | 長さ | ラベル |
|---|---|---|---|---|
| 1 | STUDYID | 文字 | 7 | Study Identifier |
| 2 | USUBJID | 文字 | 16 | Unique Subject Identifier |
| 3 | AESEQ | 数値 | 8 | Sequence Number |
| 4 | AETERM | 文字 | 32 | Reported Term for the Adverse Event |
| 5 | AEDECOD | 文字 | 25 | Dictionary-Derived Term |
| 6 | AEBODSYS | 文字 | 52 | Body System or Organ Class |
| 7 | TRTEMFL | 文字 | 1 | Treatment Emergent Analysis Flag |
| 8 | AESTDTC | 文字 | 10 | Start Date/Time of Adverse Event |
| 9 | AEENDTC | 文字 | 10 | End Date/Time of Adverse Event |
| 10 | AESER | 文字 | 1 | Serious Event |
| 11 | AESEV | 文字 | 8 | Severity/Intensity |
| 12 | AEREL | 文字 | 11 | Causality |

図 3.3.5 変換前後の SAS データセットの情報 (左：変換前，右：変換後)

# 4 まとめ

本稿では，文字セット及び符号化方式に関する基本的な情報を整理するとともに，主にレガシーデータ変換時に発生するトランスコーディングに関する問題について考察し，PMDA の規制要件に合致しないデータを検出・変換する方法を提案した．また，HTA を用いて，ラジオボタンや各種入力フィールド等から構成されるユーザーフレンドリーな画面から SAS プログラムを実行する方法を適用した，非 ASCII 文字及び非印字可能文字の検出・変換ツールの開発事例を紹介した．本ツールを使用することで，PMDA の規制要件に合致しないデータの検出・変換が容易に実施可能となった．PMDA への提出対象となった臨床試験における電子データにおいて，日本語データが含まれる場合は実務的通知に従って適切に処理されると考えられるが，海外で作成された電子データにも，PMDA の規制要件に合致しない非 ASCII 文字及び非印字可能文字が含まれる可能性があることに注意しておく必要がある．

連絡先

fumihiro.yamasaki@takeda.com

# 参考文献


- Hui Song et al. (2016). The Impact of Change from wlatin1 to UTF-8 Encoding in SAS Environment; PharmaSUG 2016; Paper BB15
- Jing Gao (2015). Multilingual data support in Dataset-XML with SAS® Clinical Data Integration; PharmaSUG China 2015; Paper 25
- Donna Dutton (2015). Data Encoding: All Characters for All Countries; PhUSE2015; Paper DH03
- Sridhar R Dodlapati et al. (2010). Non Printable & Special Characters: Problems and how to overcome them; NESUG 2010
- Abhinav Srivastva (2017). Reporting Non-Printable and Special Characters for Review in Excel; PharmaSUG2017; Paper BB04

- Manfred Kiefer (2012). SAS® Encoding: Understanding the Details; SAS Institute Inc.
- SAS Institute Inc. SAS 9.4 National Language Support (NLS): Reference Guide, Fifth Edition; Available at http://support.sas.com/documentation/onlinedoc/nls/
- SAS Institute Inc. Changing language during a SAS session https://support.sas.com/resources/papers/LocaleSwitching.pdf [Accessed 7 July 2017]
- Business Rules ; Available at https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm2005545.htm [Accessed 7 July 2017]
- PMDA 次世代審査・相談体制について (申請電子データ提出); Available at https://www.pmda.go.jp/review-services/drug-reviews/about-reviews/p-drugs/0003.html [Accessed 7 July 2017]
- 高浪洋平 (2012). SAS と HTML アプリケーションによる CDISC ADaM 形式の解析用データセットを用いた有害事象の解析帳票・グラフ簡易作成ツールの開発事例; SAS ユーザー総会 2012
- 舟尾暢男, 高浪洋平 (2008). HTMLアプリケーションを用いた簡易SASツールの開発事例 - 臨床試験における症例数設計用ツールの構築 -; SAS ユーザー総会 2008
- 高浪洋平, 舟尾暢男 (2015). 改訂版 統計解析ソフト「SAS」; カットシステム


# 付録

```
%macro DetectNonPriASCII(DataFolder=, KeyVariables=USUBJID, CVPMult=1) ;
   options nofmterr ;
   libname INDATA "&DataFolder." access=readonly cvpmult=&CVPMult. ; *入力データフォルダ ;
   ods output members=ListOfDatasets ; *データセットの一覧を取得 ;
   proc datasets lib=INDATA memtype=data ;
   quit ;
   data _NULL_ ; *データセットの数を取得 ;
      set ListOfDatasets ;
      call symput('DSNAME'||strip(put(_N_,best.)),NAME) ;
      call symputx('NofDS', _N_) ;
   run ;
   data _NULL_ ; *データセットごとの変数一覧を取得;
      set ListOfDatasets ;
      call execute('proc contents data=INDATA.'||NAME||'
                   out=ListOfVariables_'||NAME||
                   '(keep=MEMNAME NAME TYPE LENGTH VARNUM LABEL) noprint ; run ;') ;
      call execute('proc sort data=ListOfVariables_'||NAME||' ;') ;
      call execute('   where TYPE=2 ;') ;
      call execute('   by VARNUM ; run ;') ;
      call execute('data ListOfVariables_'||NAME||' ;') ;
      call execute('   set ListOfVariables_'||NAME||' ;') ;
      call execute('   by VARNUM ;') ;
      call execute('   retain MaximumLength 0;') ;
      call execute('   if MaximumLength < Length then MaximumLength = Length ;') ;
      call execute('   call symputx("MaxLengthOf'||strip(NAME)||'",MaximumLength) ; run ;') ;
   run ;
   data ListOfASCII ; *ASCII文字の一覧を変数ASCII_Listに格納 ;
      length ASCII_List $200. ;
      ASCII_List='' ;
      do i=32 to 126 ;
         ASCII_List=trim(ASCII_List)||byte(i) ;
```

```
      end ;
   run ;
   data _TEMP ; stop ; run ;
   %do i=1 %to &NofDS. %by 1 ; *データセットごとにASCII文字以外を検出 ;
      proc sql ;
         create table &&DSNAME&i. as
            select *
               from INDATA.&&DSNAME&i., ListOfASCII
      quit ;
      data _TEMP_&&DSNAME&i. ; stop ; run ;
      data _NULL_ ;
         set ListOfVariables_&&DSNAME&i. ;
         call execute('data _TEMP1 ;') ;
         call execute("   set &&DSNAME&i. ;") ;
         call execute("   length MEMNAME $32. NAME $32.
                                ORGVALUE $&&&&MaxLengthOf&&DSNAME&i..
                                DetectedString $&&&&MaxLengthOf&&DSNAME&i.. ;") ;
         call execute('   OBS=_N_ ;') ;
         call execute('   MEMNAME="'||strip(MEMNAME)||'"; ') ;
         call execute('   NAME="'||strip(NAME)||'"; ') ;
         call execute('   ORGVALUE=strip('||NAME||'); ') ;
         call execute('   DetectedString=kcompress('||NAME||',ASCII_List) ;') ;
         call execute('   CandidateValue="" ;') ;
         call execute('   if DetectedString^="" ;') ;
         call execute("   keep &KeyVariables. MEMNAME NAME ORGVALUE DetectedString OBS CandidateValue ; run ;") ;
         call execute("data _TEMP_&&DSNAME&i. ;") ;
         call execute("   set _TEMP_&&DSNAME&i. _TEMP1 ;") ;
         call execute("run ;") ;
         call execute('proc sql ; drop table _TEMP1 ; quit ;') ;
      run ;
   %end ;
   %do i=1 %to &NofDS. %by 1 ; *シートごとに検出結果を出力 ;
      %if &i.=1 %then %do ;
         ods excel file="&_CPATH.¥files¥DetectNonPriASCII.xlsx"
            options(sheet_interval='table'
                    sheet_name="&&DSNAME&i."
                    suppress_bylines='yes'
                    embedded_titles='no' ) ;
      %end ;
      %else %do ;
         ods excel
            options(sheet_interval='table'
                    sheet_name="&&DSNAME&i."
                    suppress_bylines='yes'
                    embedded_titles='no' ) ;
      %end ;
      proc print width=min data=_TEMP_&&DSNAME&i. ;
         var &KeyVariables. OBS MEMNAME NAME ORGVALUE DetectedString CandidateValue ;
      run ;
   %end ;
   ods excel close ;
   proc datasets lib=WORK kill memtype=data ;
   run ; quit ;
%mend DetectNonPriASCII ;

%macro ReplaceNonAscii(InDataFolder=, OutDataFolder=, InExcelFile=, InExcelSheet=, OutEncoding=, CVPMult=1) ;
   options nofmterr ;
   libname INDATA  "&InDataFolder." cvpmult=&CVPMult. access=readonly ; *入力データフォルダ ;
```

```
libname OUTDATA "&OutDataFolder." ; *出力データフォルダ ;
libname EXIN excel "&InExcelFile." *検出結果のオブザベーションごとに変換候補文字列を記載したファイル ;
   header=yes
   dbmax_text=2000
   mixed=yes ;
data _EXIN ;
   length CandidateValue $2000. ; *変換候補文字列の読込 ;
   set EXIN."&InExcelSheet.$"n ;
run ;
libname EXIN clear ;
proc sort data=_EXIN       out=_EX_DATA_VAR(keep=MEMNAME OBS1 NAME CandidateValue) ;
   by MEMNAME ;
run ;
data _CON_VARNUM ; stop ; run ;
data _WK_CON ; stop ; run ;
data _NULL_ ;
   set _EX_DATA_VAR ;
   by MEMNAME ;
   if first.MEMNAME then do ;
      call execute('proc contents data=INDATA.'||MEMNAME||'
                   out=_CON_VARNUM_'||MEMNAME||' noprint ; run ;') ;
      call execute('proc sort data=_CON_VARNUM_'||MEMNAME||' ;') ;
      call execute('   by VARNUM ; run ;') ;
      call execute('data _CON_VARNUM ; set _CON_VARNUM _CON_VARNUM_'||MEMNAME||' ; run ;') ;
      call execute("data _WK_"||MEMNAME||' ;') ;
      call execute('   set INDATA.'||MEMNAME||' ;') ;
   end ;
   *非ASCII及び非印字可能文字が含まれる変数値を変換候補文字列で置換 ;
   call execute('if _N_='||OBS1||' then '||NAME||'="'||strip(CandidateValue)||'" ;') ;
   if last.MEMNAME then do ; *各変数におけるデータの最大長を取得 ;
      call execute('run ;') ;
      call execute('data _WK_CON1_'||MEMNAME||'(keep=VARNAME VARLENGTH) ;') ;
      call execute('   set _WK_'||MEMNAME||' ;') ;
      call execute('   array CHARS _character_ ;') ;
      call execute('   length VARNAME $32. ;') ;
      call execute('   do over CHARS ;') ;
      call execute('      VARNAME=vname(CHARS) ;') ;
      call execute('      VARLENGTH=lengthn(CHARS) ;') ;
      call execute('      output ;') ;
      call execute('   end ; run ;') ;
      call execute('proc sql ; create table _WK_CON_'||MEMNAME||' as') ;
      call execute('   select VARNAME, max(VARLENGTH) as MAX_LENGTH, "'||strip(MEMNAME)||'" as MEMNAME') ;
      call execute('      from _WK_CON1_'||MEMNAME||' group by VARNAME ; quit ;') ;
      call execute('data _WK_CON ; set _WK_CON _WK_CON_'||MEMNAME||' ; run ;') ;
   end ;
run ;
proc sort data=_WK_CON ; by MEMNAME VARNAME ; run ;
data _NULL_ ; *各変数の長さをデータの最大長に変換 ;
   set _WK_CON ;
   by MEMNAME VARNAME ;
   MAX_LENGTH = ifn(MAX_LENGTH=0, 1, MAX_LENGTH) ;
   if      first.MEMNAME then do ;
      call execute('data _WK_'||MEMNAME||' ;') ;
      call execute('   length '||VARNAME||' $'||put(round(MAX_LENGTH,1.0),best.)||'. ') ;
   end ;
   else if last.MEMNAME then
      call execute(VARNAME||' $'||put(round(MAX_LENGTH,1.0),best.)||'. ; set _WK_'||MEMNAME||' ; run ;') ;
   else
```

```
      call execute(VARNAME||' $'||put(round(MAX_LENGTH,1.0),best.)||'. ') ;
   run ;
   proc sort data=_CON_VARNUM ; by MEMNAME VARNUM ; run ;
   data _NULL_ ;
      set _CON_VARNUM ;
      by MEMNAME VARNUM ;
      if      first.MEMNAME then do ;
         call execute("data OUTDATA."||MEMNAME||"(encoding=&OutEncoding.) ;") ; *符号化方式の変更 ;
         call execute('   format '||NAME) ;
      end ;
      else if last.MEMNAME then
         call execute(NAME||" ; set _WK_"||MEMNAME||' ; run ;') ;
      else
         call execute(NAME) ;
   run ;
   proc datasets lib=WORK kill memtype=data ;
   run ; quit ;
%mend ReplaceNonAscii ;

%macro Transcoding2UTF8(InDataFolder=, OutDataFolder=, OutEncoding=) ;
   options nofmterr ;
   libname INDATA "&InDataFolder." cvpmultiplier=4 ; *入力データフォルダ ;
   libname OUTDATA "&OutDataFolder." ; *出力データフォルダ ;
   ods listing close ;
   ods output Variables=_CON_VARNUM ;
   proc contents data=INDATA._ALL_ ; run ;
   ods listing ;
   data _CON_VARNUM ;
      set _CON_VARNUM ;
      length MEMNAME NAME $32. ;
      MEMNAME=compress(tranwrd(Member,'INDATA.','')) ;
      NAME   =strip(Variable) ;
      VARNUM =NUM ;
      keep MEMNAME NAME VARNUM ;
   proc sort data=_CON_VARNUM ; by MEMNAME VARNUM ; run ;
   data _WK_CON ; stop ; run ;
   data _NULL_ ; *各変数におけるデータの最大長を取得 ;
      set _CON_VARNUM ;
      by MEMNAME VARNUM ;
      if first.MEMNAME then do ;
         call execute('data _WK_CON1_'||MEMNAME||'(keep=VARNAME VARLENGTH) ;') ;
         call execute('   set INDATA.'||MEMNAME||' ;') ;
         call execute('   array CHARS _character_ ;') ;
         call execute('   length VARNAME $32. ;') ;
         call execute('   do over CHARS ;') ;
         call execute('      VARNAME=vname(CHARS) ;') ;
         call execute('      VARLENGTH=lengthn(CHARS) ;') ;
         call execute('      output ;') ;
         call execute('   end ; run ;') ;
         call execute('proc sql ; create table _WK_CON_'||MEMNAME||' as') ;
         call execute('   select VARNAME, max(VARLENGTH) as MAX_LENGTH, "'||strip(MEMNAME)||'" as MEMNAME') ;
         call execute('      from _WK_CON1_'||MEMNAME||' group by VARNAME ; quit ;') ;
         call execute('data _WK_CON ; set _WK_CON _WK_CON_'||MEMNAME||' ; run ;') ;
      end ;
   run ;
   proc sort data=_WK_CON ; by MEMNAME VARNAME ; run ;
   data _NULL_ ; *各変数の長さをデータの最大長に変換 ;
      set _WK_CON ;
```

```
      by MEMNAME VARNAME ;
      MAX_LENGTH = ifn(MAX_LENGTH=0, 1, MAX_LENGTH) ;
      if      first.MEMNAME then do ;
         call execute('data _WK_'||MEMNAME||' ;') ;
         call execute('   length '||VARNAME||' $'||put(round(MAX_LENGTH,1.0),best.)||'. ') ;
      end ;
      else if last.MEMNAME then
         call execute(VARNAME||' $'||put(round(MAX_LENGTH,1.0),best.)||
                     '. ; set INDATA.'||MEMNAME||' ; format _all_ ; informat _all_ ; run ;') ;
      else
         call execute(VARNAME||' $'||put(round(MAX_LENGTH,1.0),best.)||'. ') ;
   run ;
   data _NULL_ ;
      set _CON_VARNUM ;
      by MEMNAME VARNUM ;
      if      first.MEMNAME then do ;
         call execute("data OUTDATA."||MEMNAME||"(encoding=&OutEncoding.) ;") ; *符号化方式の変更 ;
         call execute('   format '||NAME) ;
      end ;
      else if last.MEMNAME then
         call execute(NAME||" ; set _WK_"||MEMNAME||' ; run ;') ;
      else
         call execute(NAME) ;
   run ;
   proc datasets lib=WORK kill memtype=data ;
   run ; quit ;
%mend Transcoding2UTF8 ;
```

# 2値データのIntraclass Correlation Coefficientの推定マクロプログラム

○稲葉　洋介[1]、田中　紀子[1]

([1]国立国際医療研究センター　データサイエンス部　生物統計研究室)

Macro program for calculating Intraclass Correlation Coefficient for binary data

Yosuke Inaba, Noriko Tanaka
Biostatistics Section, Department of Data Science, Center for Clinical Science
National Center for Global Health and Medicine

## 要旨

アウトカムが 2 値の場合の Intraclass Coefficient Correlation（級内相関係数：ICC）について，幾つかの異なる推定量を SAS で実装し，マクロプログラムを作成した.

キーワード：
Intraclass Coefficient Correlation, , ICC, Binary, Survival Cluster Randomized trial

## 1　はじめに

Intraclass Coefficient Correlation（級内相関係数：ICC）とは，同一クラスター（定義された群、あるいは繰り返し測定された個体）内のデータの相似性を示す定量的な指標である．Fisher(1925)によって当初同一個体で二回繰り返しデータに対して提案された ICC は、分散分析の枠組みで発展を遂げ、現在では“クラスター内の分散が全分散に占める割合”と一般的には定義される。これまでも、生物統計学分野においては、心理学や精神医学的測定や、病理、画像診断などに対する評価者間信頼性の指標としてしばしば用いられてきたが、近年では階層線形モデルの文脈で用いられることも多く、アウトカムの種類やサンプリングデザインに対応する様々な ICC の推定方法に関して研究が進められてきた（Donner( 1986), Donovan( 1986), Gibson( 1996)). SAS で ICC を計算する場合，アウトカムが連続量のブロックランダム化比較試験の場合は，PROC GLM や PROC MIXED，PROC NLMIXED 等のプロシージャからの出力から比較的容易に計算可能である．しかしアウトカムが 2 値の場合では，同じデザインから測定されたデータでも複数の定義があり(Ridout 1999)，推定量の計算方法も出版済みのもので 10 種類以上が知られている．その殆どは現状のプロシージャの出力をそのまま利用する事はできないため，個別にプログラミングが必要な状況である．そこで我々は，Ridout ( 1999)で比較検討されているアウトカムが 2 値の場合の 15 の ICC 推定量および一般化線形モデルによって推定される ICC を SAS により実装し，マクロプログラムを作成した．また Proc Survayselect を用いて 95%ブートストラップ信頼区間を計算する機能を付与した．ただしブートストラップ法はデータ量及び推定方法によっては膨

大な時間がかかるため，実行しない設定も可能である．

# 2　定義及び記号

本報告で用いる記号は基本的に Ridout(1999)に従う．クラスター数を $k$，$i$ 番目のクラスターのオブザベーション数を$n_i$とし，それぞれの要素の応答は 2 値確率変数$X_{ij}(i=1,\dots,k;\ j=1,\dots,n_i)$とする．$X_{ij}$の取りうる値は 1(成功)，0(失敗)で表し，$i$ 番目のクラスターの成功の総数を$Y_i=\sum X_{ij}$と表す．全てのオブザベーションにおいて成功確率は共通($Pr(X_{ij}=1):=\pi$ for $\forall i,j$)とし，異なるクラスター間のオブザベーションは全て独立と仮定する．同一クラスター内では，任意のオブザベーションの組$(X_{ij},X_{il})$の相関は共通とし，$\mathrm{Corr}(X_{ij},X_{il}):=\rho$ を ICC と定義する．一般化線形モデルの場合、

$$X_{ij}=\mu+\beta_i+e_{ij}$$

ただし

$$\boldsymbol{\beta i}\sim\boldsymbol{N}(\mathbf{0},\sigma_\beta):\text{クラスターの変量効果}$$

$$e_{ij}\sim\boldsymbol{N}(\mathbf{0},\sigma^2):\text{誤差}$$

とすると、

$$\boldsymbol{Cov}\,(X_{ij},X_{il})=\boldsymbol{E}(X_{ij},X_{il})-\boldsymbol{E}(X_{ij})E(X_{ij})=\boldsymbol{\mu^2}+\boldsymbol{E}(\boldsymbol{\beta_i}^2)-\boldsymbol{\mu}^2=\boldsymbol{\sigma^2_\beta}$$

$$\boldsymbol{Var}(X_{ij})=\boldsymbol{Var}(X_{ij})=\boldsymbol{\sigma^2_\beta}+\boldsymbol{\sigma^2}$$

ゆえに

$$\mathrm{Corr}(X_{ij},X_{il})=\frac{\sigma^2_\beta}{\sigma^2_\beta+\sigma^2}=\mathrm{ICC}$$

尚，各 15 推定量の推定方法の詳細は Ridout(1999)を参照されたい。

# 3　マクロプログラムの詳細

## 3.1 ソースコード

※Appendix 参照

## 3.2 引数

- Input: 入力するデータセット名を記載．ライブラリに格納されている場合は[library 名].[データセット名]と記載する．
  例：Library.raw
  入力するデータセットには，クラスターを格納する変数(Cluster に指定)及び応答変数(Response に指定)が必須となる．
- Cluster：Input に記載したデータセットの変数で，クラスター情報を格納している変数名を記載する．尚，type は問わない．
- Response：Input に記載したデータセットの変数で．応答を格納している変数名を記載する．尚，0 or 1 の 2 値数値変数に限る．
- Output：結果を出力するデータセット名を記載．ライブラリに格納する場合は[library 名].[データセット名]と記載する．データセットの詳細は項目「出力」参照の事。

- Method: $\hat{\rho}$の推定方法を指定する．指定できる引数を以下に示す．尚，複数の方法を半角スペース( )区切りで指定可能である．その場合は推定方法毎にレコードを分けて結果が出力される．
  - ◆ Ridout(1999)にて比較されている推定方法
  1. aov: $\hat{\rho}_{AOV}$ Analysis of Variance estimator.
  2. aovs: $\hat{\rho}^*_{AOV}$ modification of Analysis of Variance estimator
  3. keq : $\hat{\rho}_{KEQ}$ Moment estimator, $w_i = {}^{1}/_{k}$ for each $i$
  4. kpr : $\hat{\rho}_{KPR}$ Moment estimator, $w_i = {}^{n_i}/_{N}$ for each $i$
  5. keqs : $\hat{\rho}^*_{KEQ}$ modified Moment estimator, $w_i = {}^{1}/_{k}$ for each $i$
  6. kprs : $\hat{\rho}^*_{KPR}$ modified Moment estimator, $w_i = {}^{n_i}/_{N}$ for each $i$
  7. w : $\hat{\rho}_W$ Moment estimator, set $w_i$ by iterative scheme(反復回数は 10000 回，初期値は 0.02 を設定)
  8. ws : $\hat{\rho}^*_W$ modified Moment estimator, set $w_i$ by iterative scheme(反復回数は 10000 回，初期値は 0.02 を設定)
  9. stab : $\hat{\rho}_{STAB}$ stabilized Moment estimator($\kappa$は 0.45 を設定)
  10. ub : $\hat{\rho}_{UB}$ moment estimate of ICC from an unbiased estimating equation
  11. fc : $\hat{\rho}_{FC}$ Fleiss-Cuzick estimate of ICC
  12. mak : $\hat{\rho}_{MAK}$ Mak's estimate of ICC
  13. peq : $\hat{\rho}_{PEQ}$ weighted correlation estimate of ICC
  14. pgp : $\hat{\rho}_{PGP}$ weighted correlation estimate of ICC
  15. ppr : $\hat{\rho}_{PPR}$ weighted correlation estimate of ICC
  - ◆ 一般化線形モデルを用いた推定方法
  16. glmm（詳細は「2. 定義及び記号」を参照の事.）尚，Output が画面に出力される．また 95%ブートストラップ信頼区間は出力されない．
- CI:95%ブートストラップ信頼区間を Output のデータセットに計算する場合は CI=Y，出力しない場合は CI=N と記載する．省略可能．省略した場合はデフォルト値 CI=N がセットされる．
- Iter : 95%ブートストラップ信頼区間を作成する場合のリサンプリング回数を指定．Iter=xxxx の形式で記載する．省略可能．省略した場合はデフォルトで 50 が指定される。

## 3.3 出力

計算結果はデータセットに出力される。CI=Y を指定した場合とそれ以外の場合で出力される指標が異なる。

- CI=Y :
  - ◆ Method : 推定方法（Method の引数が表示される）
  - ◆ Rho_hat : $\hat{\rho}$の点推定値
  - ◆ Boot_lower : $\hat{\rho}$の推定値の 95%ブートストラップ信頼区間の下限値
  - ◆ Boot_upper : $\hat{\rho}$の推定値の 95%ブートストラップ信頼区間の上限値

  例 :

  CODE

```
%Binary_ICC(Input=library.raw2, Cluster=region, Response=y, Output=result,
Method=aov aovs peq, CI=Y,iter=100)
```

出力

| OBS | method | rho_hat | boot_lower | boot_upper |
|---|---|---|---|---|
| 1 | aov | .005851170 | -0.016595 | 0.037237 |
| 2 | aovs | .002868234 | -0.018023 | 0.029335 |
| 3 | peq | . | . | . |

■ CI=N

◆ Rho_hat：$\hat{\rho}$の点推定値

例：

CODE

```
%Binary_ICC(Input=sim, Cluster=region, Response=y, Output=result, Method=aov
kpr keqs w)
```

出力

| OBS | method | rho_hat |
|---|---|---|
| 1 | aov | 0.005851 |
| 2 | kpr | 0.006436 |
| 3 | keqs | 0.014250 |
| 4 | w | . |

### 3.4 注意点

■ マクロ中で，必要に応じて接頭語_mid_を付けた中間データセットを work ライブラリに生成しており，末尾で接頭語を指定して一括削除している．そのため work 内のデータセット名が_mid_から始まるデータセットが存在した場合，本マクロを実行すると削除される．

■ データと推定方法によっては ICC を推定できない場合がある．その場合は変数 rho_hat に欠損値が格納され，log に以下の Warning を出力する．

WARNING： ICC is not estimable by method [Method で指定した推定方法].

■ マクロ内で丸め処理は行っていない．

## 4 参考文献


FISHER, R. A. Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1925.

DONNER, Allan. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. International Statistical Review/Revue Internationale de Statistique, 1986, 67-82.

DONOVAN, A. M.; RIDOUT, M. S.; JAMES, D. J. Assessment of somaclonal variation in apple. II. Rooting ability and shoot proliferation in vitro. Journal of horticultural science, 1994, 69.1: 115-122.

GIBSON, G. J.; AUSTIN, E. J. Fitting and testing spatio- temporal stochastic models with application in plant epidemiology. Plant Pathology, 1996, 45.2: 172-184.

RIDOUT, Martin S.; DEMETRIO, Clarice GB; FIRTH, David. Estimating intraclass correlation for binary data. Biometrics, 1999, 55.1: 137-148.

# Appendix : Source Code of SAS Macro Binary_ICC

```
/*****************************************************************************
    NAME    : Binary_ICC.sas
    TITLE   : ICC calculaiton for binary data
    PRODUCT : SAS R9.4
    AUTHOR  : Yosuke Inaba
*****************************************************************************/
%Macro Binary_ICC(Input=, Cluster=, Response=, Output=, Method=, CI=N, iter=50);

%* Count methods to execute;
%let cnt=%eval(%sysfunc(count(&Method.,%str( ))) + 1);

%* calculate each method ;
%do J=1 %to &cnt.;
%LET CI2=Y;
%let _Method_=%scan(&Method.,&J.,%str( ));
%*-----------------------------------------------------------------------------;
%*     calculate ICC;
%*-----------------------------------------------------------------------------;
%if &_Method_.^=glmm %then %do;
proc iml;

start binary_ICC(cluster,response,method);
n=nrow(response);
u = unique(cluster);
k=ncol(u);
Y_i = j(1,k);
n_i = j(1,k);
do i = 1 to k;
    idx = loc(cluster=u[i]);
    Y_i[i] = sum(response[idx]);
    n_i[i] = nrow(response[idx]);
end;
n_0  = (1/(k-1))#(N - ((n_i##2)/N)[+]);
MS_b = (1/(k-1))#((Y_i##2/n_i)[+] - (1/N)#((Y_i[+])##2));
MS_w = (1/(N-k))#(Y_i[+] - ((Y_i##2)/n_i)[+]);
pi_i=Y_i/n_i;
```

```
%* The Analysis of Variance Estimator;
if method="aov" then do;
    rho_aov = (MS_b - MS_w)/(MS_b + (n_0 - 1)#MS_w);
    result = rho_aov;
end;

else if method="aovs" then do;
    MS_b = (1/(k))#((Y_i##2/n_i)[+] - (1/N)#((Y_i[+])##2));
    rho_aov = (MS_b - MS_w)/(MS_b + (n_0 - 1)#MS_w);
    result = rho_aov;
end;

%*Moment Estimator;
else if method="keq" then do;
    w_i =j(1,k,1/k);
    pi_w = (w_i#pi_i)[+];
    S_w = (w_i#((pi_i - pi_w)##2))[+];
    rho_keq=(S_w - (pi_w#(1-pi_w)#((w_i#(1 - w_i))/n_i)[+]))/(pi_w#(1 - pi_w)#((w_i#(1-w_i))[+]) -
(w_i#(1 - w_i)/n_i)[+]);
    result = rho_keq;
end;

else if method="kpr" then do;
    w_i =n_i/N;
    pi_w = (w_i#pi_i)[+];
    S_w = (w_i#((pi_i - pi_w)##2))[+];
    rho_keq=(S_w - (pi_w#(1-pi_w)#((w_i#(1 - w_i))/n_i)[+]))/(pi_w#(1 - pi_w)#((w_i#(1-w_i))[+]) -
(w_i#(1 - w_i)/n_i)[+]);
    result = rho_keq;
end;

else if method="keqs" then do;
    w_i =j(1,k,1/k);
    pi_w = (w_i#pi_i)[+];
    S_w = (w_i#((pi_i - pi_w)##2))[+];
    S_wn = (k - 1)#S_w/k;
    rho_keq=(S_wn - (pi_w#(1-pi_w)#((w_i#(1 - w_i))/n_i)[+]))/(pi_w#(1 - pi_w)#((w_i#(1-w_i))[+])
- (w_i#(1 - w_i)/n_i)[+]);
```

```
    result = rho_keq;
end;

else if method="kprs" then do;
    w_i =n_i/N;
    pi_w = (w_i#pi_i)[+];
    S_w = (w_i#((pi_i - pi_w)##2))[+];
    S_wn = (k - 1)#S_w/k;
    rho_keq=(S_wn - (pi_w#(1-pi_w)#((w_i#(1 - w_i))/n_i)[+]))/(pi_w#(1 - pi_w)#((w_i#(1-w_i))[+])
- (w_i#(1 - w_i)/n_i)[+]);
    result = rho_keq;
end;

else if method="w" then do;
    do j=1 to 10000;
    if j=1 then rho_w=0.02;
    else do;
        _w_i_sum=(n_i/(1 + rho_w#n_i))[+];
        w_i =(n_i/(1 + rho_w#n_i))/_w_i_sum;
         pi_w = (w_i#pi_i)[+];
        S_w = (w_i#((pi_i - pi_w)##2))[+];
        rho_w=(S_w - (pi_w#(1-pi_w)#((w_i#(1 - w_i))/n_i)[+]))/(pi_w#(1 - pi_w)#((w_i#(1-w_i))[+])
- (w_i#(1 - w_i)/n_i)[+]);
    end;
    end;
    result = rho_w;
end;

else if method="ws" then do;
    do j=1 to 10000;
        if j=1 then
            rho_ws=0.02;
        else do;
            _w_i_sum=(n_i/(1 + rho_ws#n_i))[+];
            w_i =(n_i/(1 + rho_ws#n_i))/_w_i_sum;
            pi_w = (w_i#pi_i)[+];
            S_w = (w_i#((pi_i - pi_w)##2))[+];
            S_wn = (k - 1)#S_w/k;
            rho_ws=(S_wn - (pi_w#(1-pi_w)#((w_i#(1 - w_i))/n_i)[+]))/(pi_w#(1 -
```

```
pi_w)#((w_i#(1-w_i))[+]) - (w_i#(1 - w_i)/n_i)[+]);
        end;
    end;
    result = rho_ws;
end;

else if method="stab" then do;
    kappa = 0;
    p = y_i[+]/n_i[+];
    w_i = n_i/N;
    pi_i = Y_i/n_i;
    pi_w = (w_i # pi_i)[+];
    s_w = (w_i # (pi_i - pi_w)##2)[+];
    rho_stab = (1/(n_0 - 1)) # ((N # s_w)/((k - 1) # p # (1 -  p)) + kappa - 1);
    result = rho_stab;
end;

else if method="ub" then do;
rho_ub = 1 - (N # n_0 # (k - 1) # Ms_w)/(Y_i[+] # (n_0 #
(k - 1) - y_i[+]) + (y_i##2)[+]);
result=rho_ub;
end;

else if method="fc" then do;
    pi_io=(Y_i[+])/N;
    rho_fc = 1 - (1/((N - k) # pi_io * (1 - pi_io)) # (Y_i # (n_i - Y_i)/n_i)[+]);
result=rho_fc;
end;

else if method="mak" then do;
    pi_io=(Y_i[+])/N;
    rho_mak = 1 - (k - 1) # ((y_i # (n_i - Y_i))/(n_i # (n_i - 1)))[+]
        /((Y_i##2/n_i##2)[+] + (Y_i/n_i)[+] # (k - 1 - (Y_i/n_i)[+]));
result=rho_mak;
end;
%* Estimators Based on Diret Calculation of Correlation Within Each Group;
else if method="peq" then do;
  mu_peq = ((n_i - 1) # Y_i)[+]/((n_i - 1) # n_i)[+];
  rho_peq = (1/(mu_peq # (1 - mu_peq))) # ((Y_i # (Y_i - 1))[+]/(n_i # (n_i - 1))[+] - mu_peq##2);
```

```
result=rho_peq;
end;

else if method="pgp" then do;
    mu_pgp = (Y_i/n_i)[+]/k;
    rho_pgp = (1/(mu_pgp # (1 - mu_pgp))) # (((Y_i # (Y_i - 1))/(n_i # (n_i - 1)))[+]/k - mu_pgp##2);
result=rho_pgp;
end;

else if method="ppr" then do;
    mu_ppr = Y_i[+]/N;
    rho_ppr = (1/(mu_ppr # (1 - mu_ppr))) # ((Y_i # (Y_i - 1)/(n_i - 1))[+]/N - mu_ppr##2);
result=rho_ppr;
end;

return(result);

finish binary_ICC;

%* Store modules to work library;
reset storage=Binary_ICC;
store module=_all_;

quit;

proc iml;

%*Read modules;
reset storage=Binary_ICC;
load module=_all_;

%*Convert input dataset to matrix;
use &Input.;
    read all into cluster var{&Cluster.};
    read all into response var{&response.};
close &Input.;

%*Calculate ICC;
result = binary_ICC(cluster,response,"&_Method_.");
```

```
%*Output ICC value to temporary dataset;
CREATE  _mid_result1 FROM result[colname='rho_hat'];
append from result;

quit;

%* Output Warning to log if ICC is not estimable;
data _mid_result1;
    set _mid_result1;
    if rho_hat<-1 or 1<rho_hat then do;
        call missing(rho_hat);
        a="WARN"||"ING :  ICC is not estimable by method &_Method_..";
        put a;
        call symputx("CI2","N");
    end;
    drop a;
run;
%end;

%else %do;
%* Calculate ICC by GLMM;
data _mid_glmm;
    set &Input.;
    subject=_N_;
run;

proc glimmix data=_mid_glmm ;
ods output CovParms=_mid_CovParms_rand_bin;
      class &Cluster.;
      model &response.=/ link=logit dist=binary;
      random &Cluster./ sub=subject ;
run;

data _mid_result1(keep=rho_hat);
    set _mid_CovParms_rand_bin(rename=(estimate=estimate_rand) keep=estimate);
    rho_hat=estimate_rand/(estimate_rand+3.289868134);
run;
%end;
```

```
%*-----------------------------------------------------------------------------;
%*     Construct 95% Confidential Interval by bootstraping;
%*-----------------------------------------------------------------------------;

%if &CI.=Y and &CI2.=Y  %then %do;
%*------  generate bootstrap sample using proc surveyselect ------;
proc surveyselect data=&Input. NOPRINT seed=1
     out=_mid_BootSSFreq(rename=(Replicate=SampleID))
     method=urs
     samprate=1
    reps=&iter.;
run;

%if &_Method_.^=glmm %then %do;
proc iml;
    %*Read modules;
    reset storage=Binary_ICC;
    load module=_all_;

    iter=&iter.;

    res=j(iter,1,0);

    do I=1 to iter;

        %*Convert dataset to matrix;
        use _mid_BootSSFreq;
            read all into cluster var{&cluster.} where(SampleID=I);
            read all into response var{&response.} where(SampleID=I);
        close _mid_BootSSFreq;

        %*Calculate ICC;
        res[I,1] = binary_ICC(cluster,response,"&_Method_.");

    end;

    %*Calculate 2.5 and 97.5 Percentiles;
    Prob = {2.5,  97.5} / 100;
```

```
    call qntl(Pctls, res, Prob);

    %* Convert matrix to dataset;
    Pctls=Pctls`;
    varNames = {"boot_lower" "boot_upper"};
    CREATE  _mid_result2 FROM Pctls[c=varNames];
    append from Pctls;
quit;
%end;

%else %do;
%let CI2=N;
%end;

%end;
%*------  Output result ------;
data _mid_mresult_&J.;
    length method $20;
    method=symget("_Method_");
    set _mid_result1;

    %* If set CI to Y and method is not GLMM then store 95%CI;
    %if &CI.=Y and &CI2.=Y %then %do;
        if _n_=1 then set _mid_result2;
    %end;

run;

%*------  delete temporary result datasets -------;
proc datasets nolist;
    delete _mid_result:;
quit;
%end;

%*------  Create output dataset by union result of each method ------;
data &Output.;
    set _mid_mresult_:;
run;
```

```
%*-----------------------------------------------------------------------------;
%*     Delete temporary datasets;
%*-----------------------------------------------------------------------------;
proc datasets nolist;
    delete _mid_:;
quit;

%Mend Binary_ICC;
%*-- Macro end ----------------------------------------------------------------;
```

# GLMSELECTプロシジャにおけるLassoの有用性に関する検討

○川﨑勝己、有光導徳、新城博子
(エイツーヘルスケア株式会社)

【背景・目的】

臨床薬理領域において、母集団解析を用いてモデル構築を行う際、共変量の影響を検討するが、そのための変数選択の方法としていくつかの種類をもつ。データは場合によって、リッチデータの場合もあれば、スパースサンプリングデータとなる場合もあり、データがもつ情報量が異なることも少なくない。最適なモデル構築のためには、より適切な変数選択の方法を用いたモデル探索は重要である。

一方、最近、機械学習を含めた広い分野でLassoを用いた変数選択の方法が注目を集めている。SASにおいても、最近のバージョンより、いくつかのプロシジャにおいて、変数選択のオプションとしてLassoが追加されている。

そこで、本発表ではGLMSELECTプロシジャを用いて、どのようなデータに対してLassoが有用となるか、シミュレーションにより検討を行った。

【方法】

シミュレーションにおける検討項目として、オブザベーション数、共変量候補変数の数および変数間の関連性等を変化させた様々なパターンを想定した。GLMSELECTプロシジャを用いて、Lassoを含むいくつかの変数選択の方法（Stepwise、Lasso、Adaptive Lasso等）による結果について比較を行い、傾向を探った。

シミュレーションの結果については、当日の発表で示す。

# SASを利用したPinnacle21 Communityのセミオートメーション化による業務効率化
# Semi-automation of Pinnacle21 Community using SAS

○片山 雅仁[1†], 勘場 大[1†], 神谷 亜香里[1], 江川 広明[1], 角谷 伸一[1]

([1]塩野義製薬株式会社 解析センター)

2016年10月よりPMDAへの申請電子データ (SDTM, ADaM, define.xml等) の提出が開始された. PMDAへの申請電子データはゲートウェイを介して提出され, バリデーションツールであるPinnacle21 Enterprise 3.0.5により, SDTM, ADaM, define.xmlのバリデーションが実施される. そのバリデーションでReject, Errorが発生しないように, 企業は事前にバリデーションツールを利用してSDTM, ADaM, define.xmlのバリデーションを実施し, バリデーション結果を確認する必要がある. バリデーションツールにて結果が許容できる内容になるまで繰り返すことになり, GUI (Graphical User Interface) で何度も実行することはしばしば手間に感じる. Pinacle21 Enterpriseには互換性のある無償版のPinnacle21 Communityというソフトウェアがあり, 利用している企業も少なくないと考えられる.

本発表では, SASプログラムでコマンドラインを実行し, バリデーションツールであるPinnacle21 Communityを効率的に動かすためのSASプログラムおよび手順を紹介する.

Sirichenko et al. (2015) はCLI (Command Line Interface) を利用し, SASプログラムでPinnacle21 Communityを実行し, データセットのバリデーションおよびData converterを実施する方法を紹介した. このSASプログラムを利用することにより, 予め必要なパラメータを指定しておくだけで, GUIで操作するよりも効率的にPinnacle21 Communityを動かすことができる.

しかしながら, そこに記載してあるSASプログラムを実行するためには, 工夫が必要であり我々はその原因を調査し, バリデーション結果が適切に出力させるプログラムの開発に成功した. また, そのプログラム加えて, Pinnacle21 CommunityのDefine.xml GeneratorをCLIで実行するプログラムも紹介する. 他方, Pinnacle21側の問題でdefine.xmlのバリデーションについては, CLIを利用することができないことがわかっていることも併せて報告する (Pinnacle 21 (2015), Pinnacle 21 (2016) 参照).

さらには, CLIを利用したSDTM及びADaMの効率的なバリデーションを紹介する. 具体的には, バリデーション結果の中には, 確認する必要性が薄いSeverityの低いRule IDも含まれているため, それらを社内ルールとして管理し, バリデーション結果ファイルに自動的に反映させることにより, 確認する必要性が高いRule IDを判別できるようにした. これにより, バリデーション結果確認を簡略化することができる.

[参考資料]

[1] Sirichenko, S., DiGiantomasso, M., and Collopy, T. (2015). Usage of OpenCDISC Community Toolset 2.0 for Clinical Programmers. Paper HT04. *PharmaSUG 2015*.

[2] Pinnacle 21. (2015). Issues with the CLI command for validating "define.xml" in 2.1 vs 2.0.2 (+ v2.2.2). [online]https://www.pinnacle21.com/forum/issues-cli-command-validating-definexml-21-vs-202. (2017-07-07参照)

[3] Pinnacle 21. (2016). CLI functionality, wrong version number reported, no define.XML support. [online]https://www.pinnacle21.com/forum/cli-functionality-wrong-version-number-reported-no-definexml-support. (2017-07-07参照)

†: Equal Contribution

# PSMATCH Procedure×ニューラルネットワークの奇跡のマッチング
## ～Real World Dataの利用可能性を忖度する～

○馬場崇充、藤原正和、北西由武
（塩野義製薬株式会社　解析センター）

医薬品開発において高いエビデンスレベルと高い効率性が求められる中，リアルワールドデータを医薬品開発に活用しようという機運が高まってきている．例の一つとしてクリニカルイノベーションネットワーク（CIN）が挙げられるが，このCINが目指すレジストリの目標の一つとして，希少疾患を対象とした臨床試験での対照群としての活用が挙げられる．しかしながら，CINをはじめとした多種多様なデータソースから得られたデータは観察データであるため，単純な比較による有効性や安全性の推測ができない．それは，例えば自社で行った臨床試験の実薬群と，Real World Dataなどの別データに由来する外部対照群という二つの群を考えた時に，被験者の背景因子等のエンドポイントに影響を及ぼす共変量の分布が群間で偏ることに起因する．観察データの活用に対して，より効率的な方法論の実装を模索していたところ，今回新たにSAS STAT14.2からPSMATCH Procedureによる傾向スコアマッチングが実装可能となった．これはRosembaum & Rubin (1983) の理論が基となっている．彼らは，強く無視できる割り当て条件のもと，共変量を条件付けたもとで被験者がどちらの群に属するかを表す確率である傾向スコアを定義し，傾向スコアの値が近い被験者を各群から抽出すると，抽出した被験者によりできた新たな群の間で共変量の分布が似通うことを示した．この理論を利用し，比較による推測を可能とする手法が傾向スコアマッチングである．このProcedureではpsmodelステートメントを用いることで，ロジスティック回帰モデルによって傾向スコアを容易に推定することができ，それを用いて傾向スコアマッチングを実装することができる．また，マッチング以外にも，傾向スコアによる層別解析，逆確率重み付きによる共変量のバランスを，出力されるグラフィックにより視覚的に確認できる等の機能も充実している．ただし，この傾向スコアを精度よく推定することが非常に重要であるが故に，psmodelステートメント（ロジスティック回帰モデル）による傾向スコアの推定が必ずしも最適とも限らない．一般に傾向スコア解析において考慮される①共変量は高次元であり，②共変量は互いに相関や傾向スコアのモデルの中で交互作用を持つことが予想される．また，③真の傾向スコアが必ずしも共変量の線形モデルであるとは限らない．このような場合には，しばしば傾向スコアのモデルを決定することは困難である．そこで，このような①～③の問題を扱うことを得意とした，機械学習のひとつであるニューラルネットワークによって傾向スコアを推定することでこれらの問題を克服することを考えた．ニューラルネットワークはIMSTAT Procedureのneuralステートメントを用いることで実装が可能であり，PSMATCH Procedureのpsdataステートメントによって，ニューラルネットワークで推定した傾向スコアをマッチングに利用できる．発表ではシミュレーションによってニューラルネットワークを用いた傾向スコアマッチングの性能を評価し，機械学習を用いた傾向スコア解析の有用性を説明するとともに，傾向スコアマッチングによるReal World Dataの活用について考える．

引用文献

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

# 電子データ申請に向けた品質管理サイクルの最適化
## ～プログラミング業務における*deja vu*からの脱却～

○小林 千鶴，平野 勝也，坂井 絵理，北西 由武
（塩野義製薬株式会社　解析センター）

2016年10月より，ついにPMDAへのCDISC標準に準拠した電子データの提出が開始（移行経過措置期間あり）され，2020年4月からは完全移行の予定である．一方，FDAでは2016 年12 月18 日以降に開始された試験において同様の対応が求められている．すなわち，申請者側もCDISC対応を急がなければならない．弊社においてもPMDA/FDAへの申請電子データ提出に適合する成果物を作成する必要がある．さらには今後，その標準化されたデータを蓄積・管理することで，試験横断的，疾患横断的な研究開発活動に利活用することも考えている．そこで，2016年度，部内において，申請データ作成プロセスの観点も含め，プログラミング業務手順を再検討した．

まず，手順検討に先立ち，Define.xml (Ver2.0) の作成にPinnacle21 Communityを利用することを決めた．しかし，Pinnacle21にインプットするためのメタデータは，多くの情報が多岐のシートにわたっていることから，その形式のままADaMデータ変換仕様書として利用した場合，仕様書作成者（Specifier）, Programmer, QC Checkerの負担が増えることが考えられた．さらに，作成したDefine.xmlが当局のバリデーションに適合するか否かを事前に確認することも必要である．もし，バリデーション結果としてRejectあるいは修正すべきError/Warningが出た場合，「（仕様書の修正 ⇒ ADaMデータ用プログラム修正 ⇒）define.xmlの修正 ⇒ define.xml再作成 ⇒ define.xmlバリデーション」のサイクルを繰り返す必要がある．このサイクルが増えるほど，これもまた担当者の負担になることは確実である．

以上のように，CDISC対応の観点を業務プロセスに組み込んだ場合，申請者側の事実上のタスクは膨大になる．そこで弊社では，この問題を緩和するためにSASマクロを利用したプログラミング業務手順を考案した．まず，仕様書とPinnacle21のためのメタデータを異なる成果物として切り分け，メタデータでシートをまたいで管理している関連情報を，仕様書では可能な限り同一シートで管理するようにした．そして，手作業で作成した仕様書からメタデータへの情報の自動マッピングおよび追加情報付与するためのSASマクロを開発した．さらには，Pinnacle21を用いたバリデーションで検知される一部のReject/Error/Warning, Pinnacle21では検出できない項目について検出するためのSASマクロも開発し，define.xml作成前に利用することとした．1つ目のSASマクロによって，仕様書からメタデータをオートメーションで作成することができ，仕様書修正が発生した場合，SASマクロを実行するだけで自動的にメタデータへの修正反映が可能となる．また，仕様書では情報を同一シートで管理することによって各担当者が作業をする際，シート間の行き来を最小限に抑え，容易に全ての情報を見ることが可能となった．また，2つ目のSASマクロを利用し事前確認を行うことによって，先に述べたサイクルの実施回数を最小限にすることが可能になった．この様にSpecifier，Programmer，QC Checkerの業務効率化を図っている．

本発表では，Pinnacle21 Communityを用いたDefine.xml Ver2.0作成に基づく，弊社における申請データ作成のプロセスを紹介する．そのプロセスにおいてはSASマクロを活用し，Specifier, Programmer, QC Checkerの業務効率化を図っている．その工夫点および効果についても紹介する．

# セントラルモニタリングにおいて
# JMP Clinicalのレポートを効果的に選択する取り組みについて

○福升悠一[1]、藤田侑子[1,2]、小幡智憲[1]、畑山知慶[1,3]、近藤秀宣[1]

([1]開発戦略本部セントラルモニタリング部、
[2]データマネジメント本部データマネジメント第2部、
[3]開発戦略本部生物統計部)

【背景】近年、効果的かつ効率的に臨床試験の品質を担保する方法としてRisk Based Monitoring(RBM)と呼ばれるモニタリング方法が注目を集めている。RBMにおいては、オンサイトモニタリング及びオフサイトモニタリングに加えて、臨床試験データを中央集約的にレビューすることで試験運用上のリスクを試験の実施中に検出するためにセントラルモニタリングが実施されており、これらのモニタリングを組み合わせたモニタリング方法にシフトしつつある。
セントラルモニタリングにおいては、臨床試験データを中央集約的にレビューする際に、データの可視化が行われ、可視化されたデータに基づいて臨床試験運用上のリスク検出のための検討が行われる。そのため、効果的にデータの可視化を行えるソフトの使用は不可欠である。JMP Clinicalはデータの可視化に優れたソフトであり、データレビューによって疑義事項が生じた場合においても、関連情報を速やかに確認することが可能である。そのため、セントラルモニタリングの効率的運用のためにはJMP Clinicalを導入は有用であると考えられる。
より効果的にセントラルモニタリングを実施するためにはJMP Clinicalの機能を適切に適用することが必要になるが、現状、JMP Clinicalで利用可能な約60種のレポートについて、どのようなリスクを検出するために用いるものであるのかレポートごとに整理されていない。また、それらをどのようなメンバー構成で評価することが適切であるのかといった指針もない。よって、JMP Clinicalをセントラルモニタリングに効果的に用いるためには、利用可能なレポートの詳細の整理とレビュー方法の検討が必要である。
【目的】セントラルモニタリングを行う際に必要となるJMP Clinicalのレポートを特定し、各レポートの適切なレビュー方法を検討する。
【方法】職種横断的・多角的な検討を行うためのセントラルモニタリングの実施者（CRA2名、DM2名、統計2名）によって、全61レポートを次の観点（「どのような分析が実施可能か？」、「どのようなリスクが検出可能か？」、「該当レポートを作成・操作・説明するのに相応しい役割は誰か？」）で評価する。
【結果】JMP Clinicalのレポートには2値の変数間における探索的な分析を行うレポートが含まれており、これらのレポートは開鍵後の分析や別レポート実行後の探索的なドリルダウンレポートとして有用であると評価した。しかし、2値変数としての比較対象群を含む“SDTM+”に加工することで、試験実施中において全体と各施設の比較も可能であると評価した。
また、JMP ClinicalはClinical Dataを対象としたレポートが中心であり、Data Integrityの確認を通して医療機関のプロセスを確認するのに有用であると評価した。その他のデータ（Secondary Data, Operational Data）は、Risk Based Monitoringレポートにより利用可能であると評価した。
【結論】本取り組みにより、各レポート機能とレビュー方法を一覧にまとめた評価者指針を作成することで、「誰が、いつ、何を」実施したら良いかが一目でわかるようになった。

# RTF出力のマクロプログラム化の工夫点の紹介について

○楊敏、宮川元志、原茂恵美子、篠津和夫
（株式会社タクミインフォメーションテクノロジー）

SASプログラムを利用した臨床試験等の図表作成において、書式設定、タイトル、脚注等の制御が煩雑になることが多くみられる。そこで、これらの煩雑さを解消し、効率よくプログラムを作成できるように、RTF出力のプログラムをマクロプログラム化した。これにより、RTFファイルの書式設定、帳票のタイトルや脚注の設定、改ページ制御等のプログラミング作業を共通化・簡素化することができた。

今回は、マクロプログラム化の過程で工夫した内容を紹介する。

① メタデータ情報を外部ファイルで管理

メタデータ情報を各プログラム内でハードコーディングすることはプログラムの汎用性を欠くことになり、メンテナンス時に、プログラムを修正しなければならない。また、設定した情報の確認は、プログラムを開いて確認する必要がある。これらの課題を解決するために、プログラム名、出力ファイル名、タイトル、脚注、データソース名等のメタデータ情報を外部ファイル（プログラム管理ファイル）に定義し、定義情報をマクロプログラムにより取得し、RTF出力のプログラムに反映させる仕組みを構築した。メタデータ情報を外部ファイルで一元管理し、プログラムを変更することなく出力情報を制御できるため、出力結果の共通化とプログラム作成の簡素化が実現できた。

② RTFコードを利用したページ番号の課題を解決するための改ページ制御マクロの作成

RTF コード（{thispage}や{lastpage}）を利用したページ番号の制御は、RTF ファイルが完全に読み込まれた後でないと、正しく表示されないため、複数の RTF ファイルを選択して一括印刷する場合や RTF ファイルが完全に読み込まれる前に印刷する場合は、ページ番号が正しく印刷されない事象が発生する。また、読み込みエラーが発生した場合なども、ページ番号が正しく表示されない。この課題を解決するために、改ページの位置情報（ページ番号及び総ページ数）を算出する改ページ制御マクロを作成した。改ページの位置情報は、出力データセットに付与され、REPORT プロシジャの PAGEBY（改ページ）及び FOOTNOTE で利用される。改ページのタイミングは、行単位、グループ単位等、柔軟に制御できるようにマクロの機能として実装し、改ページの際にグループ変数の値がページを跨ぐ場合には、次ページの先頭に'(続き)'のコメントを付与する機能も実現した。

③ SASプログラムで出力するRTFファイルの不要なコードを除去するマクロの作成

ODSステートメントを使用してRTFファイルを作成すると、SASで生成された不要なコードが付与され、帳票のレイアウトが崩れることがある。例えば、BODYTITLEオプションを利用したODS RTFステートメントでのRTF出力は、不要なセクション区切りが出力される。この課題を解決するため、作成済みのRTFファイルから不要なコードを取り除くマクロを作成した。

今回、マクロプログラム化した『RTF標準化マクロキット』は、上記で紹介した3種類のマクロを含めて、全8種類となる。RTF出力のプログラムをマクロプログラム化したことにより、最低限の指定で、効率よく同一スタイルのRTFの帳票を作成でき、図表作成プログラムをシンプルな構造にすることが可能となる。

# Annotated CRF自動化に伴う業務効率化について

○藤原由、宮川元志、楊敏、篠津和夫
（株式会社タクミインフォメーションテクノロジー）

2016年10月より国内の臨床試験の承認申請時にCDISC標準に準拠した電子データ提出が開始され、2020年4月の電子データ提出義務化に向け、製薬企業各社では対応が進められている。
そこで当社では、CDISC関連の電子データ提出物のうち、Annotated CRFに着目し、fdfファイル(pdfの注釈ファイル)作成の自動化を実現する3種類の方法について検討し、評価することとした。
define.xmlでのCRFページが記載される項目は、該当する変数名やValue Level MetadataでのCRFページの記載が考えられるためこちらの対応も検討事項とし、3種類の方法の自動化プログラム作成とメンテナンス作業の難易度も考慮に入れて検討した。

Annotated CRFのfdfファイル作成の自動化について、次の3種類の方法を検討した。

① 正規表現による実装

define.xmlの「<ItemDef」から「</ItemDef>」までの単語を1行として読み込み、SASでPerl正規表現を利用できるprxchange関数を利用して変数名及びページ番号を取得する実装方法を検討した。

② Pinnacle21 Community版を利用した実装

Pinnacle21 Community 版の「Create Spec」機能を利用して作成される Excel ファイルを利用して、変数名、ページ番号を取得する実装方法検討した。例として Excel ファイルの Variable シートを取得するプログラムを紹介する。この実装方法では、注釈を入れる位置について Excel で座標を入力し、自由に位置を指定できるカスタマイズ機能を実装した。

③ SAS XML Mapperを利用した実装

SAS社よりXML Engine用のXMLMapサポートツールとして提供されているSAS XML Mapperを利用した実装方法を検討した。さらに、XML Mapperで自動生成されるXML Map元に作成されるデータセットを利用して、変数名及びValue Level Metadataのページ番号を取得する実装方法も検討した。

上記３種類の実装方法の利点と欠点を考察する。

一つ目として、fdfファイルの作成に手作業なく自動作成ができるかという観点で比較すると、②の方法はdefine.xmlからExcelへの変換に手作業が必要なため、自動作成の実装には適さないと判断できる。

二つ目として、プログラムの可読性という観点で比較すると、①の方法で使用したprxchange関数はPerl正規表現の構文の理解が必要なため可読性が高いとは言えない。

三つ目として、Value Level Metadataの対応の観点で比較すると、①に比べて、②、③の方がValue Level Metadataの対応が容易に行える。

四つ目として、注釈位置・Not Submittedの対応の観点で比較すると、②の方法が容易である。

以上のことから、③の実装方法が最も自動化に適していると判断できるが、fdfファイル作成過程での手作業を許容できるのであれば②の実装方法も手作業が入る以外の欠点はない。

# 一般化 Wilcoxon 検定を用いた中間解析の無効中止に関する研究

○大久保竣祐、浜田知久馬、安藤宗司
（東京理科大学）

## 背景

医薬品を迅速に開発するためには、臨床試験を効率的に行う必要性がある。そのためには、臨床試験の途中の段階で、医薬品の効果を評価する必要がある。そのような背景から近年、アダプティブデザインは注目を集めている。

アダプティブデザインは臨床試験の途中で、中間解析を行うことにより、医薬品の有効性を評価し、試験の妥当性や被験者数などのデザインを再検討する。中間解析では、医薬品の無益性（最終的に効果がない）を判定するため条件付き検出力、予測検出力を評価する。連続量（正規分布）、二値（正規近似）、生存時間データ（Log-rank 検定）における上記の指標の算出方法が提案されており、SAS の SEQTEST プロシジャで計算可能である。しかしノンパラメトリックな生存時間の検定である一般化 Wilcoxon 検定については方法論の検討が不十分である。

## 研究目的

本研究の目的は、生存時間データ（一般化 Wilcoxon 検定）における条件付き検出力、予測検出力の算出方法を定式化し、その性能を評価することである。

## 提案方法

提案方法では、条件付き検出力関数$P_n(\delta)$の$\delta$を症例数設計の式から算出する。

$$P_n(\delta) = \Phi\left(\frac{\sqrt{n}Z_n + (N-n)\delta/\sqrt{2} - \sqrt{N}Z_{1-\alpha}}{\sqrt{N-n}}\right)$$

$$\boxed{\delta\sqrt{N} = Z_{1-\alpha} + Z_\beta \rightarrow \delta = \frac{Z_{1-\alpha} + Z_\beta}{\sqrt{N}}}$$

$N$: 最終解析時の被験者数(1 群)　　$n$: 中間解析時の被験者数(1 群

$Z_n$: 中間解析時の検定統計量　　$Z_{1-\alpha}$: 標準正規分布の 100（$1-\alpha$）パーセント点

## まとめ

帰無仮説の下では条件付き検出力（最尤推定）で評価すると最も無益な試験を中止しやすいことが分かった。また対立仮説の下では提案した方法で評価すると最も有益な試験を中止しにくいことが分かった。

## 参考文献

[1] Christopher J, Bruce WT. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC. 2000

[2] Huson LW. The Bayesian Bootstrap in a Predictive Power Analysis. Case Studies in Business, Industry and Government Statistics. Vol.3. No.1. pp.18-22. 2009

[3] Ware JH, Muller JE, Braunward E. The futility index: An approach to the cost-effective termination of randomized clinical trials. American Journal of Medicine. Vol.78. No.4. pp.635-643. 1985

# XSLプロシージャ（PROC XSL）を使ってみました。
# --簡単に"define.xml"を読み込むために。--

○小澤義人
株式会社メディサイエンスプラニング
医薬情報本部　データサイエンス部
EDC・CDI SCグループ

Why don't we try to use the XSL procedure ( PROCXSL)
We can read the "define.xml" using the SAS System easily

Yoshihito Ozawa
EDC･CDI SC Group
Data Science Dept.
Medical Information Div.
MEDISCIENCE PLANNING INC.

承認申請時の電子データ提出に際し、提出する電子データは、sas データセット(xpt)、define.xml およびそのスタイルシート(xsl)を提出することとなっている(1)。電子データに加え提出用の define.xml およびそのスタイルシートの作成は各社、種々の方法で作成している。電子データ提出が始まり、時間も経過してきたのでこれらのファイル作成方法などの知見は各社で蓄積してきているものと思われる。これらのファイルを作成する場合 excel の仕様書から作成することが多い。今後は define.xml をメタデータとして最初に作成される状況も揃ってきているのではないかと考える。define.xls ファイルを扱うには SAS XML Mapper を用い、map ファイルを作成するのが通常の方法であると考えられる(2)。SAS XML Mapper の使用方法自体も簡単ではなく、define.xml を扱うのは、excel からメタデータを取得するよりも面倒なものと考えられる。昨年の sas ユーザー総会においても、define.xml の利用例や SAS XML Mapper で作成した map ファイルの編集事例が紹介されている(3)。これを参考に、map ファイルを作成しようと試みたが、map ファイルの作成は困難に感じられた。一方、スタイルシートの作成も、map ファイル作成と同様困難なものと考えたが、スタイルシートに関しては提出用にすでに存在するスタイルシートがあるため、それを修正すればよいのではないかと思い修正変更を行ったところ、スタイルシートの修正変更は意外に簡単であった。その経験として、今回 define.xml とそのスタイルシートを用い、スタイルシートの変更により異なる csv ファイルを PROC XSL を使用し、作成することができる例示を行い、その csv ファイルを sas データセットに読み込むまでのプロセスおよびスタイルシートの変更の検討事例を報告する。

キーワード：PROC XSL CDI SC define.xml

参照:

(1) http://www.pmda.go.jp/files/000213558.pdf
平成 28 年 8 月 24 日 薬機次発第 0824001 号 「承認申請時の電子データ提出等に関する技術的ガイドについて」の一部改正について
(2) SAS Global Forum 2010: Paper 157-2010
Understanding the define.xml File and Converting It to a Relational Database
Lex Jansen
(3) SAS ユーザー総会 2016　C-05
SAS XML Mapper を用いた Define.xml のメタ情報の有効活用および業務効率化の取り組み
林 三男、山本繁

# 多群における二重頑健推定量を用いた交絡の調整法に関する研究

○江森 健人、浜田 知久馬、安藤 宗司
（東京理科大学）

観察研究から治療と効果の因果関係を検討する場合，交絡因子の影響を取り除いて効果を評価する必要がある．統計解析によって交絡因子の影響を調整する場合，数学モデルを用いた方法と層別解析が標準的な交絡調整法であり，臨床研究で頻繁に応用されている．

しかし，近年の臨床研究では上記の2つの標準的な交絡調整法に加え，傾向スコア法が盛んに応用されるようになった．傾向スコア法とは，ある治療を受ける確率のことである傾向スコアを用いて多数の交絡因子の影響を同時に調整する方法の総称である．実際には，各患者の傾向スコアは不明であることから，ロジスティック回帰モデルなどを用いて推定を行う必要がある．しかし，そのモデルの特定を誤った場合には，効果の推定に偏りが生じることが知られている．そこで新たに提案されているのが，二重頑健推定量である．二重頑健推定量とは，傾向スコアによる重み付き推定と共変量で治療効果を調整する回帰モデルを併用した推定量であり，傾向スコアと回帰モデルの少なくともどちらかが正しく特定されれば，効果を正しく推定できることが知られている．

また通常，傾向スコアは治療群と対照群といった2群の割付変数に対してモデル化し効果を推定する．それに対しImbensは，たとえ割付が3群以上でも，傾向スコアを利用できることを証明し，一般化傾向スコアとして提案した．この方法はLOGISTIC PROCEDUREでLINK=GLOGITオプションを指定することで適用できる．

本研究の目的は，シミュレーション実験により，一般化傾向スコアを用いて多群に拡張させた二重頑健推定量の性能を評価することである．評価指標としては，推定値と真値との差であるbiasと推定値と真値の差の二乗の平均値であるMSE（Mean Square Error）を用いた．

**参考文献**

[1]Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **64**:962–972.

[2]Imbens GW, The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**: 706-710.

[3]Peter C. Austin. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies, *Statistics in Medicine* 2010; **29**(20): 2137–2148

[4]Ping Feng, Xiao-Hua Zhou, Qing-Ming Zou, Ming-Yu Fan and Xiao-Song Li. Generalized propensity score for estimatingthe average treatment effect of multiple treatments, *Statistics in Medicine* 2012; **31**(7): 681–697

[5]Yosuke Fujii, Masayuki Henmi and Toshiharu Fujita. Evaluating the interaction between the therapy and the treatment in clinical trials by the propensity score weighting method, *Statistics in Medicine* 2012; **31**(3): 235–252.

# プラセボ効果を抑制し試験薬の効果を効率的に検出するための エンリッチメントデザインに関する研究

○舛田優、浜田知久馬、安藤宗司
（東京理科大学）

医薬品の効果は、医薬品そのものが持つ効果の他に、服用したという安心感や期待感によって生じる心理的効果が追加されることがあるが、これには個人差がある。このため、多くの臨床試験においては、プラセボを対照薬として用いることで、心理的効果を差し引いた医薬品本来の効果を正しく評価しようとしている。

一方で、特定の疾患領域（疼痛、精神・神経系領域等）における臨床試験では、プラセボ効果が大きく現れることにより、医薬品の効果が十分に示し難いということが問題となっている。このような問題点への対処として、従来では、Placebo lead-in designと呼ばれるエンリッチメントデザイン（医薬品の効果を見出し易い集団を選択するような試験デザイン）が多く用いられてきた。しかし、Placebo non-respoder（プラセボ効果があまり現れない被験者）を判定する際にバイアスが生じる可能性があることや、Placebo lead-in期の導入がランダム化後のプラセボの反応率に影響しないことを示唆するメタアナリシスの結果など問題点も存在する。

このような現状から、Sequential Parallel Comparison Design（SPCD）と呼ばれる、二段階のエンリッチメントデザインが提案されている。SPCDではStage1において全被験者集団での効果を、Stage2においてPlacebo non-responderの集団に対する効果を推定し、それらの重み付き平均に基づいて、全体としての医薬品の効果を推定する。近年ではTwo-way Enriched Design（TED）などのSPCDを拡張したようなデザインも提案されているが、それらの性能評価に関しては十分でないことから、本研究ではこれらのデザインの性能に関する検討を行う。

上記デザインでは応答に基づいて有効性の解析対象とする集団を選択することにより、潜在的に従来の試験デザインに比べて効率が向上することが期待される。先行研究[2]では、さらにSPCDにおいて中間解析を行い、条件付き検出力を用いたサンプルサイズの再推定等を行った場合の検討を行っている。本研究ではSPCDにおけるより詳細な検討と、TEDを想定した場合の性能に関して、シミュレーションにより検討した結果を報告する。

**参考文献**

[1] Maurizio Fava et al . The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics*. 2003; **72**:115-127

[2] Rachel K. Silverman, Anastasia Ivanova . Sample size re-estimation and other midcourse adjustments with sequential parallel comparison design. *Journal of Biopharmaceutical Statistics*. 2017; **27**(3):416-425

[3] Yeh-Fong Chen et al . Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemporary Clinical Trials*. 2011; **32**(4):592-604

# 臨床試験における2段階デザインの性能評価に関する研究

○井澤慶信、浜田知久馬、安藤宗司
（東京理科大学）

## 背景

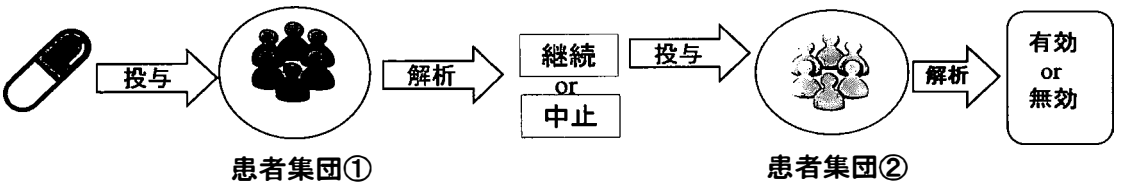
臨床試験における被験者数設計は、試験デザインの要素の中で最も重要なものの1つである。試験開始前に予め見積もった効果サイズに基づいて被験者数設計を決定する，固定デザインを用いるのが一般的である。しかし、この方法では見積もった効果サイズが、真の効果サイズから乖離すると、検出力が大きく変化してしまう。この問題を解消するために、試験の途中で被験者数再設計を伴う適応的デザイン[1]が提案されている。適応的デザインは効果サイズの不確実性に対する頑健性が高い一方で、盲検化の解除に伴う実施の煩雑さや、αエラーの上昇などの懸念点が挙げられている。

## 研究目的

本研究は適応的デザインが抱える懸念点を解消した新たな試験デザインの提案及び、その性能評価を行うことを目的としている。評価指標としては、一般的に広く使用されている検出力や期待被験者数に加え、過剰被験者数と不足検出力による2種の損失を統合したAPS(Average Performance Score)[2]を用いた。

## 提案方法

提案試験デザインは、各被験者集積時点の結果のみを用いて解析を実施する。従来の適応的デザインでは各解析時点の結果を統合するために、検定統計量の統合を行う必要がある。提案試験デザインは検定統計量の統合を行わないことで、実施の煩雑さを軽減することが狙いである。




## まとめ

シミュレーション実験により評価したところ、提案試験デザインはいずれの評価指標においても、適応的デザインとほぼ同等の性能を示した。今後は他の提案されている適応的デザインとの性能比較や、新たな評価指標を用いての評価を実施することを検討する。

## 参考文献

[1] Cui L，Hung HMJ，Wang S．Modification of sample size in group sequential clinical trials．*Biometrics*，55:853-857．1999

[2]Liu GF，Cui L．Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval．*Statistics in Medicine*，27：584-596．2008．

# 生物統計・医薬関連／SASシステム

# proc STREAMによるAnalysis Results Metadataの作成

関根　暁史
（藤本製薬株式会社）

proc STREAM: The Feasible Method for Creating define.xml Files.

Satoshi Sekine
Fujimoto Pharmaceutical Corp.

## 要旨

従来の put 文を用いる方法から変更して、define.xml の一部である Analysis Results Metadata を、STREAM プロシジャによる方法で作成した。

キーワード：STREAM プロシジャ，Analysis Results Metadata，define.xml，Pinnacle の Validator

## 1. はじめに

SAS を用いてテキストファイルを外部に吐き出す場合は、従来は put 文を用いていた。SAS9.4 より STREAM プロシジャが実装されている。put 文を用いる方法ではクオーテーションマークで文字列を挟み込む手間があるとともに、マクロ変数を仕込むためにシングル・ダブルクオーテーションの使い方を工夫せねばならなかったが、STREAM プロシジャの使用によりそれらが大幅に回避される。そこで STREAM プロシジャを使用することにより、define.xml の一部である Analysis Results Metadata(以下、解析結果メタデータ)を作成することとした。

本論文は、既存の解析結果メタデータ無しの define.xml に対し、SAS で生成した解析結果メタデータを挿入することをコンセプトとしている。出来上った define.xml を Pinnacle の Validator にかけても、極力 Warning は出さない仕様を目指した(Reject，Error はもちろん出さない)。

最後に STREAM プロシジャを用いた Pinnacle の ADaM データセットの Validation についても触れる。

STREAM プロシジャを使用して見たが、xml 文書などの作成に適していると考えられた。

## 2. put 文と STREAM プロシジャ

put 文においては、文書の書き始めをカラムポインタコントローラで設定しなければならなかった。またマクロ変数を仕込むために、クオーテーションマークを何重にも記入したり、時にはシングル・ダブルクオーテーションマークを反転させたりといった工夫が必要であった。そのためプログラム作成者本人にとっても可読性の良いプログラムとはならなかった。

```
%let stylesheet=define2-0-0.xsl;
/* put文 */
filename sample ".\sample.txt";
data _null_;
 file sample;
 put @1 '<?xml version="1.0" encoding="UTF-8" ?>'/
    @1 '<?xml-stylesheet type="text/xsl" href=" ' "&stylesheet." ' "?>' /
    @3 "<!-- Date: &sysdate9. -->" ;
run;
```

改行コード

マクロ変数を仕込むために、クオーテーションを何重にも記入していた

カラムポインタ

```
/* STREAMプロシジャ */
filename sample ".\sample.txt";
proc stream outfile=sample; begin
<?xml version="1.0" encoding="UTF-8" ?>&streamdelim. newline;
<?xml-stylesheet type="text/xsl" href="&stylesheet."?> &streamdelim. newline;
<!-- Date: &sysdate9. --> &streamdelim. newline;
;;;;
```

改行コード

空白文字は考慮される

beginで始まり最後はセミコロン4つで終わる。run; quit; ともに不要。

```
(sample.txt)   <?xml version="1.0" encoding="UTF-8" ?>
               <?xml-stylesheet type="text/xsl" href="define2-0-0.xsl"?>
                <!-- Date: 13APR2017 -->
```

実際の出力xml文書

一方 STREAM プロシジャでは、先頭の空白文字が考慮されるので、カラムポインタという概念は不要であったり、クオーティングを気にすることもなく、成果物そのものに近い可読性の良いプログラムを書くことが可能である。

## 3. 改行コードとコメントアウト

STREAM プロシジャでは、改行コード(&streamdelim. newline;)の存在が若干可読性を悪くしているが、resetdelim=オプションにより&streamdelim.はユーザーが自由に変更することが出来るので、これにより可読性が増す。当論文では以降、改行コードは mylabel newline;を使用するものとする。

```
proc stream outfile=sample resetdelim="mylabel"; begin
<?xml version="1.0" encoding="UTF-8" ?> mylabel newline; *改行コードの後ろ;
<?xml-stylesheet type="text/xsl" href="&stylesheet."?> &streamdelim. newline;
 <!-- Date: &sysdate9. --> /* resetdelimの指定 */ mylabel newline;
;;;;
```

/**/のコメントはどこでもコメントアウト出来る

元の改行コードも使用可

*; のコメントは改行コードの後ろであればコメントアウトされる

/**/で囲まれたコメントは通常の SAS の使用と同じで、どこに書いてもコメントアウトされる（外部ファイルに出力されることはない）。一方、*; で囲まれたコメントは改行コードの後ろであればコメントアウトされるが、改行コードの前に書いた場合は外部ファイルにその内容が出力される。

## 4. マクロ

STREAM プロシジャでは外部マクロの使用が可能である。以下は xml 文書としてのタイトルを出力するマクロである。マクロは最後にセミコロンを打たずとも展開する。

```
%macro TITLE(_title=);
 <!-- ********************************** --> mylabel newline;
 <!-- ***  &_title. *** --> mylabel newline;
 <!-- ********************************** --> mylabel newline;
%mend TITLE;

proc stream outfile=sample resetdelim="mylabel"; begin
%TITLE(_title=外部ファイルにリンクを貼る)
;;;;
```

セミコロンはなくとも展開する。逆にセミコロンが出力に残ってしまうと define.xml に不具合をもたらす。

```
 <!-- ********************************** -->
 <!-- ***  外部ファイルにリンクを貼る          *** -->
 <!-- ********************************** -->
```

xml文書におけるコメントアウト

## 5. 繰り返しの構造

put 文では保持したデータセットの内容をそのまま外に吐き出すことが出来た。しかし STREAM プロシジャはデータセットを保持したり、データを発生させたりすることが出来ないので、外側で作ったデータセットの内容をマクロ変数の形で STREAM の内部に引き渡すしかない。下に xml のタグが繰り返されている構造を STREAM に引き渡す方法を示す。

VIEWTABLE: Work.Link

| | ID | Title | Href |
|---|---|---|---|
| 1 | at14-3-01.sas | at14-3-01.sas | at14-3-01.sas |
| 2 | at14-5-02.sas | at14-5-02.sas | at14-5-02.sas |

```
<def:leaf ID="LF.at14-3-01.sas" xlink:href="at14-3-01.sas">
   <def:title>at14-3-01.sas</def:title>
</def:leaf>
<def:leaf ID="LF.at14-5-02.sas" xlink:href="at14-5-02.sas">
   <def:title>at14-5-02.sas</def:title>
</def:leaf>
```

SASプログラムにリンクを貼るxml文書

上記のようなデータセット link あったとして、<def:leaf> </def:leaf>で閉じられたタグが繰り返される出力を作って見る。先ずはデータセット link を加工して links を作成する。

```
data links; set link;
 text='<def:leaf ID="LF.'||strip(ID)||'" xlink:href="'||strip(href)||'"> mylabel newline;'; output;
 text=' <def:title>'||strip(Title)||'</def:title> mylabel newline;'; output;
 text='</def:leaf> mylabel newline;'; output;
run;
```

更にデータセット links の text の情報を、proc sql を用いてマクロ変数化し、マクロ変数を STREAM 内で展開すれば、上記の xml 文書の内容は実現する。

```
proc sql noprint;
 select text into : _repeat separated by '' from links;
quit;
proc stream outfile=sample resetdelim="mylabel";
begin
&_repeat.
;;;;
```

SQLによるマクロ変数化

マクロ変数の展開

先程のデータステップおよび proc sql の内容は、%let rc=%sysfunc(dosubl( ));で囲まれた中に記述することで、STREAM プロシジャ内に収めることは出来る。但しプログラムの可読性はよろしいとは言えないので、余りお薦め出来る方法ではない。

```
proc stream outfile=sample resetdelim="mylabel";
begin
%let rc=%sysfunc(dosubl(
data links; set link;
text='<def:leaf ID="LF.'||strip(ID)||'" xlink:href="'||strip(href)||'"> mylabel newline;'; output;
text=' <def:title>'||strip(Title)||'</def:title> mylabel newline;'; output;
text='</def:leaf> mylabel newline;'; output;
run;
proc sql noprint;
 select text into : _repeat separated by '' from links;
quit;
));
&_repeat.
;;;;
```

%let rc=%sysfunc(dosubl(
));
で閉じられた箇所に、データステップやプロシジャステップの記入は可能

## 6. DDE の操作

STREAM プロシジャは DDE も操作することが出来る。下記にエクセルを自動で閉じる工程を示している。書き始めの角括弧のみ特殊文字と捉えられてしまうのか、%NRSTR によるクオーティングをしておかないと STREAM プロシジャは作動しない。この一手間のみはかかるが、他は従前の put 文による方法の煩わしさからは解放されるであろう。このように xml 文書作成のみならず、従前 put 文で書いていた方法の大半は STREAM プロシジャによる方法に置き換え可能である。

```
* エクセルを自動で閉じる;
%let bracket=[;
filename xls dde 'excel|system';
proc stream outfile=xls; begin%NRSTR(&bracket.)
close(false)]
[quit()]
;;;;
```

書き始めの角括弧は特殊文字と捉えられてしまう

%NRSTRによるクオーティング

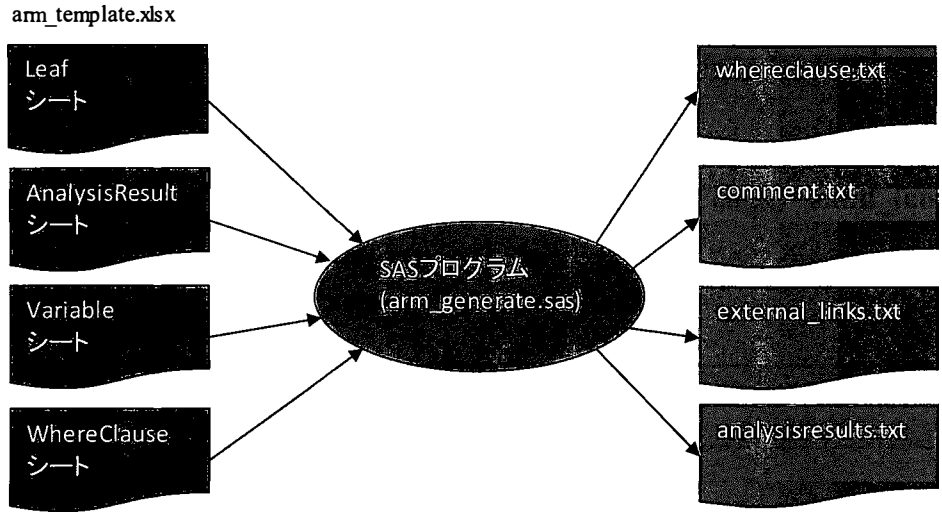
残りの角括弧のクオーティングは不要

以上のテクニックを用いて、いよいよ次章より解析結果メタデータを作成することとする。

4

## 7. 4つのエクセルシートと4つのテキストファイル

解析結果メタデータ作成に用いるエクセルテンプレート(arm_template.xlsx)を先ず用意した。arm_template.xlsx は、Leaf シート（リンクを貼りたい外部ファイルの定義）、AnalysisResult シート（解析結果メタデータ全般定義）、Variable シート（使用データセットおよび使用変数を定義）、WhereClause シート（データ抽出条件定義）の4つのシートから成り立っている。それらのシートを SAS プログラム(arm_generate.sas)が読み込み、情報を加工することで xml 言語で書かれた4つのテキストファイル、whereclause.txt（データ抽出条件）、comment.txt（複数のデータセットを使用する際の結合コメント）、external_links.txt（外部ファイルにリンクを貼る）、analysisresults.txt（解析結果メタデータ全般）をそれぞれ生成する。概念図は以下の通りである。




## 8. whereclause（データ抽出条件）

whereclause.txt はデータ抽出条件部分（Data References (incl. Selection Criteria)）を担っている。def:WhereClauseDef Element に PARAMCD が記入されていた場合、arm:AnalysisResult Element (analysisresults.txt の中にある）にも自動的に PARAMCD を書き入れる仕様とした。Data References (incl. Selection Criteria)の PARAMCD = “SYSBP”の内容は Analyis Parameter (s)にも連動して書き込まれることとなる。

| | |
|---|---|
| Analysis Parameter(s) | PARAMCD = "SYSBP" (Systolic Blood Pressure (mmHg)) |
| Analysis Variable(s) | CHG (Change from Baseline) |
| Analysis Reason | SPECIFIED IN SAP |
| Analysis Purpose | PRIMARY OUTCOME MEASURE |
| Data References (incl. Selection Criteria) | ADVS [PARAMCD = "SYSBP" and ANL01FL = "Y" and FASFL = "Y"] |

arm:AnalysisResult Element

def:WhereClauseDef Element

## 9. comment（複数のデータセットを使用する際の結合コメント）

例えば有害事象集計をする際に、有害事象発現例数は ADAE から、割合の分母は ADSL から引っ張って来ることになる。ADAE, ADSL という複数のデータセットの結合法をコメントとして入れられる。その際 COM.JOIN- ADAE- ADSL という CommentOID が作られるので、下欄の Comments にも同内容が記載されることになる。

| | |
|---|---|
| Data References (ind. Selection Criteria) | ADAE [TRTEMFL = 'Y' and AESER = 'Y']<br>ADSL [SAFFL = 'Y']<br>有害事象発現割合の分母はADSLから算出し、分子はADAEから算出した。ADAEとADSLは症例番号(USUBJID)で結合し、ADAE由来のレコードを分子とした。 |

同じ内容が連動する

**Comments**

| CommentOID | Description |
|---|---|
| COM.JOIN-ADAE-ADSL | 有害事象発現割合の分母はADSLから算出し、分子はADAEから算出した。ADAEとADSLは症例番号(USUBJID)で結合し、ADAE由来のレコードを分子とした。 |

## 10. external_links（外部ファイルにリンクを貼る）

具体的なプログラムは 5 章で詳しく述べさせて頂いたが、解析結果作成用 SAS プログラムまでリンクを貼るための xml である。解析結果作成用 SAS プログラム以外にも ADRG(データガイド)や ADaM 作成プログラムなどへもここからまとめてリンク定義してもよい。

| | |
|---|---|
| Programming Statements | [SAS version 9.4]<br>at14-5-02.sas |

## 11. analysisresults（解析結果メタデータ全般）

CSR の 1 つの表題の中で複数の解析を行っている場合、下記の R. 1, R. 2 のように複数の Table のブロックを発生することが出来る。

arm_template.xlsxのAnalysisResultシート

| Display Table | Display Title | Repeat Table | Repeat Title |
|---|---|---|---|
| Table 14-3.01 | 主要な解析：血圧値の投与前からの変化量 | R.1 | 収縮期血圧の投与前からの変化量の線形モデル解析 |
| Table 14-3.01 | 主要な解析：血圧値の投与前からの変化量 | R.2 | 拡張期血圧の投与前からの変化量の線形モデル解析 |

**Table 14-3.01**

| | |
|---|---|
| Display | Table 14-3.01 主要な解析：血圧値の投与前からの変化量 |
| Analysis Result | 収縮期血圧の投与前からの変化量の線形モデル解析 |
| Analysis Parameter(s) | PARAMCD = "SYSBP" (Systolic Blood Pressure (mmHg)) |
| Analysis Result | 拡張期血圧の投与前からの変化量の線形モデル解析 |
| Analysis Parameter(s) | PARAMCD = "DIABP" (Diastolic Blood Pressure (mmHg)) |

エクセルテンプレート(arm_template.xlsx)の Variable シートにおいて、AnalysisDataset にデータセット名を記載すれば Data References に反映されることになっている。Analyis Variable (s)で直接使用されない ADSL もここでは記入が必要となる。ADAE の変数である AEBODSYS, AEDECOD をカンマ区切りで記入することによって、Analyis Parameter (s)に分離されて表示されるようになっている（変数が 2 つ以上あってもカンマ区切りにされていれば反映可)。

arm_template.xlsxのVariableシート

| DisplayTable | Repeat Table | AnalysisDataset | AnalysisVariables |
|---|---|---|---|
| Table 14-5.02 | R.1 | ADAE | AEBODSYS,AEDECOD |
| Table 14-5.02 | R.1 | ADSL | |

| | |
|---|---|
| Analysis Variable(s) | AEBODSYS (Body System or Organ Class)<br>AEDECOD (Dictionary-Derived Term) |
| Analysis Reason | SPECIFIED IN SAP |
| Analysis Purpose | PRIMARY OUTCOME MEASURE |
| Data References (Ind. Selection Criteria) | ADAE [TRTEMFL = 'Y' and AESER = 'Y']<br>ADSL [SAFFL = 'Y']<br>有害事象発現割合の分母はADSLから算出し、分子は、由来のレコードを分子とした。 |

エクセルテンプレートの AnalysisResult シートにおいて、前もって空白文字と改行コード入りの ProgrammingStatements を記載しておけば、define. xml 上で纏に折り返す。このように xml 言語そのものの折り返しと、xml タグに組み込まれた文字列の折り返しの区別がないところが STREAM プロシジャの便利なところである。

arm_template.xlsxのAnalysisResultシート

ProgrammingStatements

```
proc mixed data = ADVS; mylabel newline;
  where PARAMCD="SYSBP" and ANL01FL='Y' and FASFL='Y'; mylabel newline;
  class TRTPN AVISITN USUBJID; mylabel newline;
  model CHG = TRTPN AVISITN BASE / ddfm = KR; mylabel newline;
  repeated AVISITN / type = TOEPH subject = USUBJID mylabel newline;
run;
```

改行コードは前もってエクセル内に仕込んでおく

Programming Statements

[SAS version 9.4]

```
proc mixed data = ADVS;
  where PARAMCD="SYSBP" and ANL01FL='Y' and FASFL='Y';
  class TRTPN AVISITN USUBJID;
  model CHG = TRTPN AVISITN BASE / ddfm = KR;
  repeated AVISITN / type = TOEPH subject = USUBJID;
run;
```

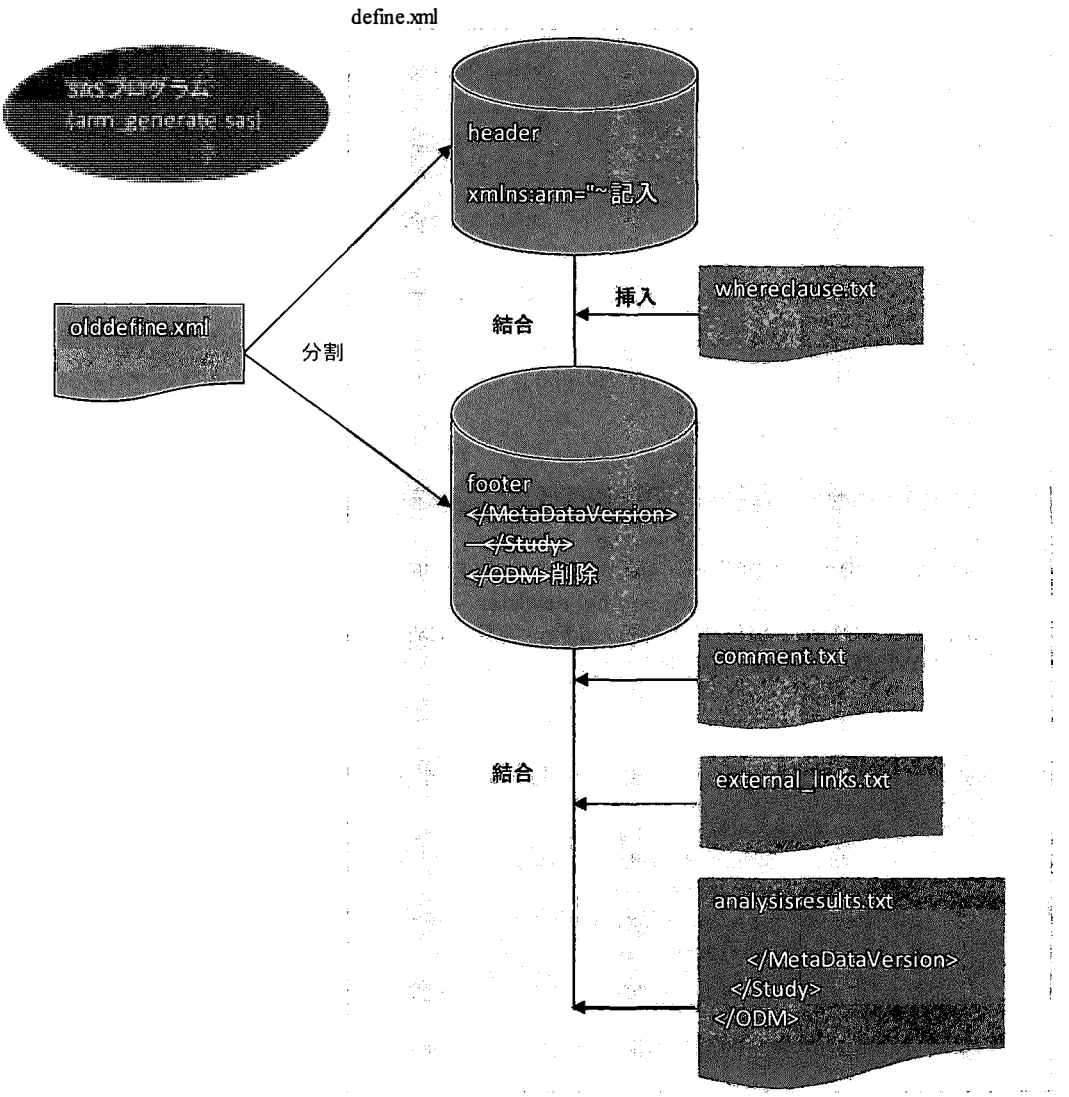
字下げは考慮される

改行

以上のように analysisresults.txt を作成するが、ファイルの最後に、</MetaDataVersion> </Stu dy></ODM>と追記しておく。

## 12. define.xml 作成

以上 4 つのテキストファイルが揃ったところで、既存の define.xml に解析結果メタデータを挿入することとする。既存の解析結果メタデータ無しの define.xml を「olddefine.xml」と名付けておく。olddefine.xml を SAS で読み込み 2 つのデータセット header および footer に分解する。”<ItemGroupDef OID”が出て来た後のパーツが footer であり、それより前のパーツが header である。なお header には「xmlns:arm="http://www.cdisc.org/ns/arm/v1.0"」を追記し、footer の「</MetaDataVersion> </Study> </ODM>」は削除しておく。更に 4 つのテキストファイル whereclause.txt、comment.txt、external_links.txt、analysisresults.txt の内容を再度読み込んで来て、下図の順番通りに情報を積み上げて合成し、内容を外部出力すれば define.xml は完成する。最後の出力 filename ステートメントにオプション encoding="UTF-8"を付けておけば、符号化方式に間違いがない。



以上、一見遠回りな作成方法を採っているように見えるが、この方法により「DD0007 Invalid content was found starting with element 'def:CommentDef'.」の Warning を回避することが出来ている。Pinnacle の Validator は v2.1.3 版を用いた。解析結果メタデータにおいて回避出来ない Warning とは「DD0078 Document 'LF.at14-5-02.sas' is not referenced」のようなリンクが貼られていないという Warning である。実際にはプログラムにリンクは貼られているのであるが、v2.1.3 版が解析結果メタデータ自体を監視していないことに由来するものであり、潜在的な問題である。

## 13. ADaM データセットの Validation

STREAM プロシジャは Pinnacle の ADaM データセットの Validation にも用いることが出来る。Pinnacle のコンポーネントには「validator-cli-2.1.3.jar」(cli とはコマンドラインインターフェース)が含まれており、これを SAS 側から制御することで Validation の自動化が可能となる。以下に ADaM データセットの Validation 法を概説する（ADaM の Validation に define.xml を添えることは可能である)。

```
/* プログラム: adam_validation.sas */
*=====Pinnacle21(Community版)の置き場所(絶対パスを指定);
%let _path= C:\pinnacle21-community-2.1.3\;
*=====ADaM, define.xml, 出力ファイルの置き場所(絶対パスを指定);
%let _abs= C:\pinnacle21-community-2.1.3\;

filename _temp "&_abs.\adam_validation.bat";
proc stream outfile=_temp resetdelim="mylabel"; begin
cd &_path. mylabel newline;
java -jar components\lib\validator-cli-2.1.3.jar
 -type=ADaM
 -source=&_abs.\*.xpt
 -source:type=SAS
 -config=\ADaM" "1.0" "(PMDA).xml
 -config:cdisc=2016-03-25
 -config:define=&_abs.\define.xml
 -report=&_abs.\adam_validation.xlsx
 -report:type=Excel
 -report:cutoff=1000
 -report:overwrite=yes
 >&_abs.\adam_validation.log mylabel newline;
exit
;;;;

options noxwait xsync;
x"&_abs.\adam_validation.bat";
```

カレントディレクトリをPinnacleの置き場所に合わせる。

batファイルに出力。

ADaMのxptファイルを全て指定している。

ADaM IGの指定。コマンドラインでは空白文字が許されないので、空白文字をダブルクオーテーションで囲っている。

NCI Terminologyの指定。

define.xml添付。この行はコメントアウトすることが出来る。

Validationレポートをエクセルファイルに出力。カットオフを1000としてファイルの上書き可としている。

Pinnacleの実行log取得。

SAS側からbatファイルを叩く。

(Pinnacleの実行log)

```
Beginning validation, please wait...
The validation has completed.
```

本来ならば define.xml そのものの Validation を SAS から制御したいところであったが、validator-cli-2.1.3.jar には define.xml の Validation 機能が落ちてしまっているようであった。

## 14. まとめ

STREAM プロシジャを使用することによりクオーティングを気にすることなく、define.xml の一部である解析結果メタデータを無理なく作成することが出来た。また define.xml の作成の工夫により Pinnacle の Validator が出す特定の Warninng を回避した。STREAM プロシジャは xml のようにタグで囲まれた文書の作成に向いていると思われるが、その他にも、従来から put 文で書かれている方法（DDE、コマンドライン、VBScript、R、Pinacle の Validator 等の操作）の大半は STREAM に置き換え可能と思われる。SAS プログラム等は SAS プレミアムラウンジ（Web）に置かせて頂いたので、ご参照頂きたい。

参考文献


1) Don Henderson (2014). PROC STREAM and SAS Server Pages: Generating Custom HTML Reports, SAS Global Forum 2014 Paper 1738-2014

2) Joseph Hinson (2015). Proc STREAM: The Perfect Tool For Creating Patient Narratives, PharmaSUG 2015 Paper AD03

3) Joseph Hinson (2016). The New STREAM Procedure as a Virtual Medical Writer, PharmaSUG 2016 Paper AD17

4) CDISC ADaM Metadata Sub-Team (2015). Analysis Results Metadata Specification Version 1.0 for Define-XML Version 2, 2015 Clinical Data Interchange Standards Consortium, Inc.

5) Chris Holland (2012). Implementing CDISC Using SAS An End-to-End Guide, SAS Institute Inc.

6) 高浪 洋平 (2014). SAS と Excel を用いた CDISC ADaM 標準における作業効率化の試み, SAS ユーザー総会 2014 論文集

7) 高浪 洋平 (2015). 解析結果メタデータの概要と SAS による作成方法の提案, SAS ユーザー総会 2015 論文集

8) Mark Crangle (2014). Running OpenCDISC from SAS, PhUSE 2014 Paper TS04

9) Sergiy Sirichenko (2016). Usage of Pinnacle 21 Community Toolset 2.1.1 for Clinical Programmers, PharmaSUG 2016 Paper HT04

```
/* プログラム: arm_generate.sas */

/* パス指定(全フ ァイルはここに出力される) */
%let path=.;*本プログラムと同じ場所に全て の読み込 みフ ァイルを置くことが前提;
%let metadata=arm_template.xlsx;
/* 既存の解析結果メタデータ無しのdefine.xml名 */
%let _file=olddefine.xml;*本プログラムと同じ場所に置く;

%let lt=%nrstr(&lt);
%let apos=%nrstr(&apos);
%let quot=%nrstr(&quot);

options noxsync;
x "'&path.¥&metadata.'";*エクセルを自動で開く;
/* 4つのエクセルシートの読み込 み*/
filename xls dde "EXCEL|Leaf!R2C1:R100C3";
data _link;
  infile xls dsd notab dlm='09'x missover;
  length ID Title Href $32767 ;
  input ID Title Href $ ;
  if ID="" then delete;
run;
filename xls dde "EXCEL|AnalysisResult!R2C1:R100C15";
data _result;
  infile xls dsd notab dlm='09'x missover ;
  length DisplayTable DisplayTitle CSR_ID CSR_Page RepeatTable RepeatTitle
         AnalysisReason AnalysisPurpose JoinComment AnalysisDocumentation SAP_ID
         SAP_Page SASversion ProgrammingStatements SASprogram $32767 ;
  input DisplayTable DisplayTitle CSR_ID CSR_Page RepeatTable RepeatTitle
        AnalysisReason AnalysisPurpose JoinComment AnalysisDocumentation SAP_ID
        SAP_Page SASversion ProgrammingStatements SASprogram $ ;
  if RepeatTitle="" then delete;
run;
filename xls dde "EXCEL|Variable!R2C1:R100C6";
data variable;
  infile xls dsd notab dlm='09'x missover ;
  length DisplayTable RepeatTable AnalysisDataset AnalysisVariables $32767 ;
  input DisplayTable RepeatTable AnalysisDataset AnalysisVariables $ ;
  if DisplayTable="" then delete;
run;
filename xls dde "EXCEL|WhereClause!R2C1:R100C6";
data whr;
  infile xls dsd notab dlm='09'x missover;
  length DisplayTable RepeatTable AnalysisDataset CheckVariable Comparator CheckValue $32767 ;
  input DisplayTable RepeatTable AnalysisDataset CheckVariable Comparator CheckValue $ ;
  if DisplayTable="" then delete;
run;

%macro TITLE(_title=);
 <!— ****** ********** *** *** *** *** *** *** *** ********* —> mylabel newline;
 <!— ***  &title. * ** —> mylabel newline;
 <!— *** ** *** *** *** *** *** ********** *** *** *** *** *** ** —> mylabel newline;
%mend TITLE;

proc sort data=whr; by DisplayTable RepeatTable; run;
proc sort data=_result out=result0; by DisplayTable RepeatTable; where DisplayTable^=""; run;
data result1;
  merge result0 whr(where=(CheckVariable in("PARAMCD","PARAM","PARAMN")) keep=DisplayTable RepeatTable AnalysisDataset
CheckVariable);
  by DisplayTable RepeatTable;
  if CheckVariable^="" then ParameterOID=strip(AnalysisDataset)||"."||strip(CheckVariable);
  obs1=_n_; drop AnalysisDataset CheckVariable;
run;
proc sort data=variable; by DisplayTable RepeatTable; run;
```

```
data variable0; set variable;
  by DisplayTable RepeatTable;
  length comjoin $32767;
  retain comjoin;
  comjoin=strip(comjoin)||"-"||strip(AnalysisDataset);
  if first.RepeatTable then comjoin=AnalysisDataset;
  if last.RepeatTable;
run;
proc sql;
  create table comjoin as select distinct DisplayTable,RepeatTable,comjoin from variable0 where comjoin like '%-%';
quit;

data variable1; merge variable comjoin; by DisplayTable RepeatTable;
  cont=countw(AnalysisVariables,",");
do i=1 to cont; length var $32767;
  var=scan(AnalysisVariables,i,",");output;
end;
run;
data variable2; set variable1;
  obs2=_n_;
run;
data result2; merge result1 variable2;
  by DisplayTable RepeatTable;
run;
proc sort data=result2; by obs1 obs2; run;
data result3;
  set result2;
  by notsorted DisplayTable notsorted RepeatTable notsorted AnalysisDataset;
  retain rep1 rep2 rep3 0;
  if first.DisplayTable then rep1=rep1+1;
  if first.RepeatTable then rep2=rep2+1;
  if first.AnalysisDataset then rep3=rep3+1;
run;

data comment;
  set result3(where=(JoinComment^=""));
  by rep1 rep2 rep3;
  length text $32767;
  if first.rep2 then do;
text='mylabel newline;                    <def:CommentDef OID="COM.JOIN-'||strip(comjoin)||'">'; output;
text='mylabel newline;                       <Description>'; output;
text='mylabel newline;                          <TranslatedText
xml:lang="en">'||strip(JoinComment)||'</TranslatedText>'; output;
text='mylabel newline;                       </Description>'; output;
text='mylabel newline;                    </def:CommentDef>'; output;
end;
run;
%let _repeat=;*initialize;
proc sql noprint;
  select text into : _repeat separated by ' ' from comment;
quit;
filename comment "&path.¥comment.txt" encoding="UTF-8" lrecl=32767 ;
proc stream outfile=comment quoting=both resetdelim="mylabel"; begin
%TITLE(_title=複数のデータセットを使用する際の結合コメント)
&_repeat.
;;;;
data arm;
  set result3;
  by rep1 rep2 rep3;
  length text $32767;
if first.rep1 then do;
text='mylabel newline;              <arm:ResultDisplay OID="RD.'||strip(DisplayTable)||'"
Name="'||strip(DisplayTable)||'">'; output;
text='mylabel newline;                 <Description>'; output;
text='mylabel newline;                    <TranslatedText xml:lang="en">'||strip(DisplayTitle)||'</TranslatedText>';
output;
text='mylabel newline;                 </Description>'; output;
```

```
if CSR_ID^="" then do;
text='mylabel newline;                    <def:DocumentRef leafID="LF.'||strip(CSR_ID)||'">'; output;
text='mylabel newline;                       <def:PDFPageRef PageRefs="'||strip(CSR_Page)||'" Type="PhysicalRef"/>';
output;*NamedDestination;
text='mylabel newline;                    </def:DocumentRef>'; output;
end;
end;

if first.rep2 then do;
text='mylabel newline;                    <arm:AnalysisResult'; output;
text='mylabel newline;                       OID="AR.'||strip(DisplayTable)||'.'||strip(RepeatTable)||'"'; output;
if ParameterOID^="" then do;
  text='mylabel newline;                     ParameterOID="IT.'||strip(ParameterOID)||'"'; output;
  end;
text='mylabel newline;                       AnalysisReason="'||strip(AnalysisReason)||'"'; output;
text='mylabel newline;                       AnalysisPurpose="'||strip(AnalysisPurpose)||'">'; output;
text='mylabel newline;                       <Description>'; output;
text='mylabel newline;                          <TranslatedText xml:lang="en">'||strip(RepeatTitle)||'</TranslatedText>';
output;
text='mylabel newline;                       </Description>'; output;
if comjoin^="" & JoinComment^="" then do;
text='mylabel newline;                       <arm:AnalysisDatasets def:CommentOID="COM.JOIN-'||strip(comjoin)||'">';
output;
end;
else do;
text='mylabel newline;                       <arm:AnalysisDatasets>'; output;
end;
end;

if first.rep3 then do;
text='mylabel newline;                          <arm:AnalysisDataset ItemGroupOID="IG.'||strip(AnalysisDataset)||'">';
output;
text='mylabel newline;                             <def:WhereClauseRef
WhereClauseOID="WC.'||strip(DisplayTable)||'.'||strip(RepeatTable)||'.'||strip(AnalysisDataset)||'"/>'; output;
end;

if var^="" then do;
text='mylabel newline;                             <arm:AnalysisVariable
ItemOID="IT.'||strip(AnalysisDataset)||'.'||strip(var)||'"/>'; output;
end;

if last.rep3 then do;
text='mylabel newline;                          </arm:AnalysisDataset>'; output;
end;

if last.rep2 then do;
text='mylabel newline;                    </arm:AnalysisDatasets>'; output;
text='mylabel newline;                    <arm:Documentation>'; output;
text='mylabel newline;                       <Description>'; output;
text='mylabel newline;                       <TranslatedText
xml:lang="en">'||strip(AnalysisDocumentation)||'</TranslatedText>'; output;
text='mylabel newline;                       </Description>'; output;
if SAP_ID^="" then do;
text='mylabel newline;                       <def:DocumentRef leafID="LF.'||strip(SAP_ID)||'">'; output;
text='mylabel newline;                          <def:PDFPageRef PageRefs="'||strip(SAP_Page)||'" Type="PhysicalRef"/>';
output;*NamedDestination;
text='mylabel newline;                       </def:DocumentRef>'; output;
end;
text='mylabel newline;                    </arm:Documentation>'; output;
text='mylabel newline;                    <arm:ProgrammingCode Context="'||strip(SASversion)||'">'; output;

if ProgrammingStatements^="" then do;
text='mylabel newline;                    <arm:Code>'; output;
text='mylabel newline;'||strip(ProgrammingStatements)||''; output;
text='mylabel newline;                    </arm:Code>'; output;
end;
```

13

```
if SASprogram^="" then do;/* 5.3.9 Child Elements*/
text='mylabel newline;                    <def:DocumentRef leafID="LF.'||strip(SASprogram)||'"/>'; output;
end;

text='mylabel newline;                       </arm:ProgrammingCode>'; output;
text='mylabel newline;                    </arm:AnalysisResult>'; output;

end;

if last.rep1 then do;
text='mylabel newline;                 </arm:ResultDisplay>'; output;
end;
run;
%let _repeat=;*initialize;
proc sql noprint;
  select text into : _repeat separated by ' ' from arm;
quit;
filename ar "&path.¥analysisresults.txt" encoding="UTF-8" lrecl=32767 ;
proc stream outfile=ar quoting=both resetdelim="mylabel"; begin
%TITLE(_title=解析結果メタデータ全般)
          <arm:AnalysisResultDisplays>
&_repeat.
mylabel newline;            </arm:AnalysisResultDisplays> mylabel newline;
      </MetaDataVersion> /*改行*/mylabel newline;/*改行*/
   </Study> mylabel newline; *改行;
</ODM>
;;;;

data whr_clause; set whr; by DisplayTable RepeatTable;
  cont=countw(CheckValue,",");
  do i=1 to cont; length whr $32767;
    whr=scan(CheckValue,i,",");output;
  end;
run;
data _clause;
  set whr_clause;
  by notsorted DisplayTable notsorted RepeatTable notsorted AnalysisDataset notsorted CheckVariable;
retain rep1 rep2 rep3 rep4 0;
if first.DisplayTable then rep1=rep1+1;
if first.RepeatTable then rep2=rep2+1;
if first.AnalysisDataset then rep3=rep3+1;
if first.CheckVariable then rep4=rep4+1;
run;
data clause; set _clause;
  length text $32767;
  by rep1 rep2 rep3 rep4;
if first.rep3 then do;
text='mylabel newline;         <def:WhereClauseDef
OID="WC.'||strip(DisplayTable)||'.'||strip(RepeatTable)||'.'||strip(AnalysisDataset)||'">'; output;
end;
if first.rep4 then do;
text='mylabel newline;            <RangeCheck Comparator="'||strip(Comparator)||'" SoftHard="Soft"
def:ItemOID="IT.'||strip(AnalysisDataset)||'.'||strip(CheckVariable)||'">'; output;
end;
text='mylabel newline;               <CheckValue>'||strip(Whr)||'</CheckValue>'; output;
if last.rep4 then do;
text='mylabel newline;            </RangeCheck>'; output;
end;
if last.rep3 then do;
text='mylabel newline;         </def:WhereClauseDef>'; output;
end;
run;
```

```
%let _repeat=;*initialize;
proc sql noprint;
  select text into : _repeat separated by ' ' from clause;
quit;
filename clause "&path.¥whereclause.txt" encoding="UTF-8" lrecl=32767 ;
proc stream outfile=clause quoting=both resetdelim="mylabel"; begin
%TITLE(_title=データ抽出条件)
&_repeat.
;;;;

proc sql;
  create table link as select distinct ID,Title,href from _link;
quit;
data links; set link;
  length text $32767;
text='mylabel newline;            <def:leaf ID="LF.'||strip(ID)||'" xlink:href="'||strip(href)||'">'; output;
text='mylabel newline;               <def:title>'||strip(Title)||'</def:title>'; output;
text='mylabel newline;            </def:leaf>'; output;
run;
%let _repeat=;*initialize;
proc sql noprint;
  select text into : _repeat separated by ' ' from links;
quit;
filename external "&path.¥external_links.txt" encoding="UTF-8" lrecl=32767 ;
proc stream outfile=external quoting=both resetdelim="mylabel"; begin
%TITLE(_title=外部ファイルにリンクを貼る)
&_repeat.
;;;;

/* 既存のdefineの読み込みと分割 */
filename olddef "&path.¥&_file." encoding="UTF-8" lrecl=32767;
data header footer;
  infile olddef firstobs=1 pad ;
  input @1 CNT $char32767.;
  ;;;;
  if index(CNT,"</MetaDataVersion>")>0 | index(CNT,"</Study>")>0 | index(CNT,"</ODM>")>0 then delete;
  CNT=tranwrd(CNT,'xmlns=','xmlns:arm="http://www.cdisc.org/ns/arm/v1.0" xmlns=');
  retain rep 0;
  if index(CNT,"<ItemGroupDef OID")>0 then rep=1;
  if rep=0 then output header;
  else if rep=1 then output footer;
run;
/* 4つのテキストファイルの読み込み */
data clause;
  infile clause firstobs=1 pad ;
  input @1 CNT $char32767.;
  ;;;;
run;
data comment;
  infile comment firstobs=1 pad ;
  input @1 CNT $char32767.;
  ;;;;
run;
data external;
  infile external firstobs=1 pad ;
  input @1 CNT $char32767.;
  ;;;;
run;
data analysisresults;
  infile ar firstobs=1 pad ;
  input @1 CNT $char32767.;
  ;;;;
run;
```

```
/* define.xml合成 */
filename newdef "&path.¥define.xml" encoding="UTF-8" lrecl=32767 ;
data _null_;
  set header clause footer comment external analysisresults;
  if CNT^="";
  file newdef;
  put +(kverify(CNT," ")-1) CNT;
run;

/* エクセルを自動で閉じる */
%let bracket=[;
filename xls dde 'excel|system' ;
proc stream outfile=xls; begin%NRSTR(&bracket.)
close(false)]
[quit()]
;;;;

/* define.xmlを自動で開く*/
x "&path.¥define.xml";
```

# 帳票MockupからのRTF用テンプレートSASプログラム自動作成ツール「Taiho TLF Automated Tool(TTAT)」の紹介

○伊藤衡気[1]、栗矢芳之[2]、銭本敦[2]

([1]株式会社タクミインフォメーションテクノロジー、[2]大鵬薬品工業株式会社)

## [Prologue]

臨床試験におけるTLFの主流がExcelからRTFに移行した今、解析結果などの本質的な部分ではなく、見た目を整えるテンプレートプログラムの作成に時間をとられるといった問題が生じている。
一度作成したものに、列の追加や罫線の有無、各列の幅の調整やラベルの修正など、細々した調整が入ることも少なくなく、手間がかかる作業となっていることは否めない。

## [Progress]

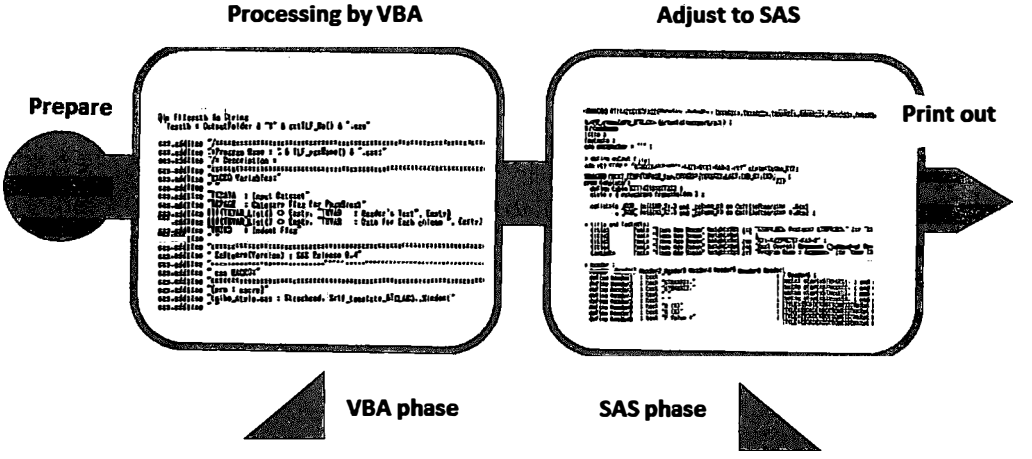
RTFで作成するMockupをExcelで作成している企業もあるが、今回我々はExcelでのレイアウト情報（罫線、文字列、インデント、列数等）をVBAで取得し、RTF出力用のSASマクロプログラムを自動生成するツール「Taiho TLF Automated Tool(以下 TTAT)」を開発した。TTATによって自動生成されたマクロプログラムのパラメータに、解析結果のデータセットの変数名を設定するだけで、Mockupの見た目通りのRTFが作成される。

## [Advantage]

TTATにより、テンプレートプログラム作成の手間自体が大きく省けることに加え、細々とした変更にも簡単に対応できる。さらに、Mockupの編集が、TTATから生成されるSASプログラムに反映されるため、MockupとRTFテンプレートの差異の予防にもなる。

## [Presentation]

プレゼンテーションでは、実際にはどのように動くのか、デモを交えて機能を紹介(及び公開)する。これにより、SASの可能性が、新しい技術の分野は言うまでも無く、本TTTATツールで用いたVBAなど、よく知られた既存の技術とのコラボレーションの中にもまだまだ広がっていることの一例となれたらと思う。



# 信頼区間を塗りつぶす ～SGANNOを利用して～

関根　暁史

（藤本製薬株式会社）

LIFETESTプロシジャはSAS9.2以降、信頼区間付きの高品質なODSグラフを出力するようになっている（図1）。確かにODSグラフは高品位ではあるのだが、LIFETESTプロシジャの出力そのものを解析報告や論文などに用いる方はほとんどいないと思われ、多くの方はLIFETESTプロシジャの出力データをSGPLOTプロシジャで受け止めてグラフを再作成されていると考えられる。しかし、SGPLOTプロシジャを使用してカプラン・マイヤー推定値の折れ線を描くのは可能であるが、95%信頼区間を塗るということまではデフォルトの機能だけでは難しく、信頼区間を線分で表現したとしてもLIFETESTのグラフよりも見劣りしてしまうだろう。そこで、SAS9.3より実装のSGANNO機能を用いて信頼バンドの情報のAnnotationデータセットを作成し、SGPLOTに提供することにより、LIFETESTに劣らぬ美しさを持つグラフの作成を実現した（図2）。SGANNO機能のうちPOLYGON関数を用いた。Annotationデータセットの作成がやや難解であったが、一度作成すれば他のものに流用可能であろう。

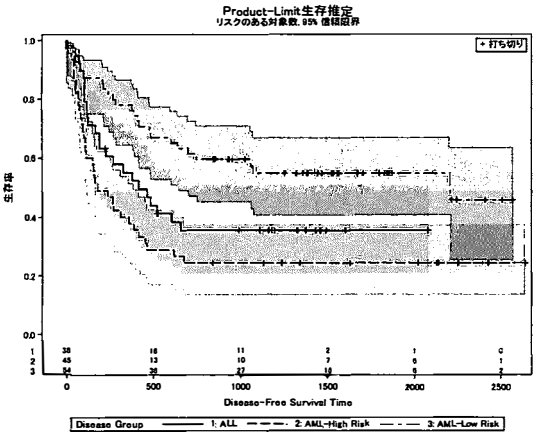



図1. デフォルトのLIFETEST出力

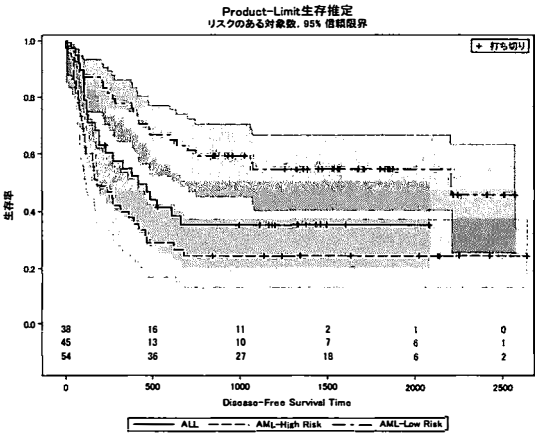



図2. SGANNOを挿入したSGPLOT

ポスター資料では、上記以外の関数の信頼区間やベン図の描画についても紹介する。また当日のポスター資料であるが、SASの機能（ODS PDF）のみを使用して作成した大判のものを掲出予定である。グラフ・ポスターともにSAS9.4を用いて作成したが、作成プログラムはSASプレミアムラウンジ（Web）に置かせて頂いた。

# チュートリアル

# 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モデルの被験者数設計，
# 最適被験者数配分

**○浜田知久馬**
**(東京理科大学)**

LIFETEST, PHREGの近年の拡張によって，生存時間解析における競合リスクの解析，cumulative incidenceの推定とGray検定，Cox回帰を拡張した，Fine and Gray(FG) modelの解析が可能になった． またがんの分子標的薬は，遺伝子型によって，反応の仕方が大きく異なり，生存時間曲線が比例ハザード型にならないことが多数報告されている．本発表ではチュートリアルとして，生存時間解析の実務的課題として競合リスク，非比例ハザードモデルの下での被験者数設計，最適被験者数配分等のSASでの解決法について解説する．

１）最適被験者数配分

計量データの場合，総被験者数が一定のとき，検出力が最大になる最適被験者数配分は，比較する2群に1:1の割り付けることである．これに対し生存時間解析では，情報量は被験者数ではなくイベント数であるため，対数ハザード比の分散は2群のイベント数の関数で表される．対数ハザード比の分散を最小にする最適被験者数配分について解説する．POWERプロシジャではgroupweight=オプションを利用して，非等例数の場合の被験者数設計を行うことができる．実際の生存時間解析では，登録期間の長さや，途中脱落等を考慮する必要があるが，これらに対してもPOWERプロシジャで対応することができる．

２）非比例ハザードモデルの下での被験者数設計

近年開発されているがんの免疫療法の免疫チェツクポイント阻害剤では，効果の発現するまでに一定の時間を要することが知られており，区分混合分布モデルを仮定して被験者数設計が行われている．また，Fleischerのモデルは，死亡までの経路として，増悪後の死亡と増悪を経ない死亡の2種類を想定する．区分混合モデル，Fleischerのモデルは比例ハザードモデルにしたがわない．このような場合でも，POWERプロシジャでcurve=オプションを指定して，モデルに基づいて各時点の生存率を指定することで被験者数設計を行うことができる．プログラム例を紹介するとともに比例ハザードモデルを前提に被験者数設計を行った場合と結果を比較する．

３）競合リスク(Competing Risks)の解析

複数の競合するイベントを解析するためにいくつかのアプローチが考えられる．イベントが2種類の場合で説明するとa)1，2の双方をイベント.b)1をイベント，2を打切り． c) 1をイベント，2を競合イベントとする．b)の解析のKM法，ログランク検定，Cox比例ハザードモデルに対応する，c)の競合リスクの解析法がcumulative incidence，Gray検定，Fine and Grayモデルであり，これらの解析がSASの最新バージョンで可能であり，プログラム例と解釈の仕方について解説する．

# 複数母集団のパラメータ推定 －全国消費実態調査を用いて－

高橋 行雄
BioStat 研究所（株）

Parameter estimation of multiple populations　－ Using Income and Expenditure Survey －

Yukio Takahashi
BioStat Research Co.,Ltd.

**要旨：** 得られたデータが，どのような分布に従うかを見極めることは，統計解析の基礎の基礎である．一般的には，「データの分布が正規分布に従うか否か」が主な興味であり，伝統的には，正規確率紙を用いた視覚的な判断であった．最近の JMP では，どのような分布のあてはめが適切かを判断するために，対数尤度および AICc の出力がされるようになった．さらに母集団の数を指定し，多重正規混合分布のパラメータ推定も容易に得られるようになっている．全国消費実態調査の匿名データから擬似ミクロデータを作成する際に，元の匿名データに含まれる年間収入など多くの収支項目についてヒストグラムを作成したところ，対数正規分布のあてはめがためらわれるような分布が多数存在することに気が付いた．新たに作成した擬似ミクロデータは，誰でも自由にダウンロードできるように公開したので，元の匿名データの分布の形状が新擬似ミクロデータでも再現しているか，との観点から比較検討を行った．解析の対象データの分布を同定するために JMP では，最尤法を用いているが，ほとんどの統計の教科書に取り上げられていないので，Excel を用いて多重正規混合分布の複数母集団のパラメータ推定を行い，JMP の結果と対比しつつ推定方法について解説する．

**キーワード：** 全国消費実態調査，匿名データ，多重正規混合分布，SAS，JMP，擬似ミクロデータ

## 1. はじめに

得られたデータが，どのような分布に従うかを見極めることは，統計解析の基礎の基礎である．一般的には，「データの分布が正規分布に従うか否か」が主な興味であり，伝統的には，正規確率紙を用いた視覚的な判断であった．最近の JMP では，どのような分布のあてはめが適切かを判断するために，対数尤度および AICc の出力がされるようになった．さらに母集団の数を指定し，多重正規混合分布のパラメータ推定も容易に得られるようになっている．

全国消費実態調査の匿名データから擬似ミクロデータを作成する際に，元の匿名データに含まれる年間収入など多くの収支項目についてヒストグラムを作成したところ，対数正規分布のあてはめがためらわれるような分布が多数存在することに気が付いた．新たに作成した擬似ミクロデータは，誰でも自由にダウンロードできるように公開したので，元の匿名データの分布の形状が新擬似ミクロデータでも再現しているか，との観点から比較検討を行った．

全国消費実態調査は，5年ごとに実施され，利用可能となっているのは，1989年，1994年，1999年，2004年の4回分であり，2017年のLet's データコンテストためには，2004年の匿名データに限定して新擬似ミクロデータを作成した．以下のURLに新擬似ミクロデータが公開されている．

SAS データセット　http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroData.zip
CSV 形式　http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroDataCSV.zip

解析の対象とするデータの分布を同定するために JMP では，最尤法を用いているが，ほとんどの統計の教科書に取り上げられていない．そこで，Excel を用いて最尤法による多重正規混合分布の複数母集団のパラメータ推定を行い，JMP の結果と対比しつつ最尤法による複数母集団のパラメータの推定方法について解説する．

## 2. 代表的な収支項目

2004年全国消費実態調査の匿名データは，47,797 レコード，1780 項目のフラットなファイルで提供されている．この中から，世帯属性に関する14項目および収支に関する203項目を抜き出し，さらに秘匿処理をしたファイルを作成し，この秘匿処理ファイルから，3種類の統計表を作成し公開した．

新擬似ミクロデータは，公表した3種類の統計表のみを用いて，各種の乱数を用いて作成した69,131 レコードのフラットなファイルであり，元の匿名データとの関連性は完全に切れている．新擬似ミクロデータを用いて各種の統計解析を行った場合に，計算された各種の統計量あるいは分布の形状が，元の匿名データの特性をどれだけ保持しているか検討した．代表的な収支項目として，年間収入，住居費，教育費を取り上げた．

## 3. 年間収入は対数正規分布に従うのか

図1は，年間収入についての匿名データの集計乗率を用いて，10万世帯当たりの集計乗率を計算し，それを「度数」として使ったヒストグラム，分位点，平均と標準偏差についての結果である．2500万円でトップコーディングがされているが，おおむね右に裾を引く対数正規分布と推測される．

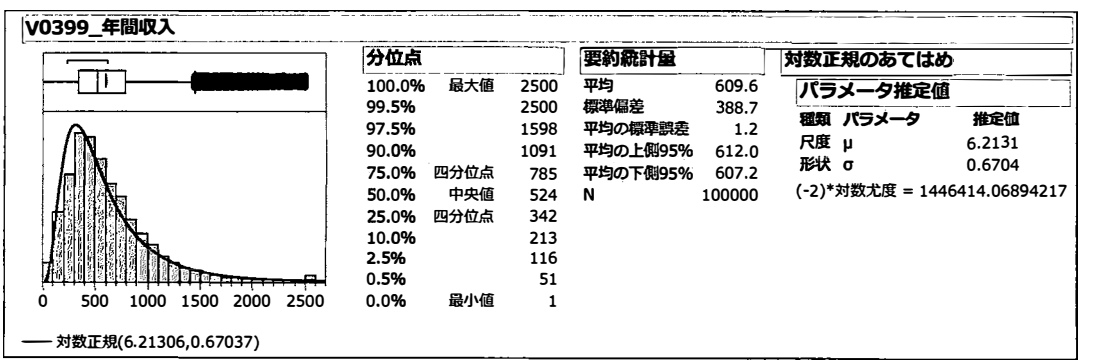



図1　年間収入の分布の特徴(匿名データ)

ヒストグラム上に対数正規分布の密度関数が上書きされている．その下に「対数正規（6.213, 0.670)」との出力があり，これらは図 1 の右の欄から（形状$\mu$，尺度$\sigma$）に対応していることが判読できる．その下に(-2)対数尤度=1446414.1 が出力されていることから，最尤法によって分布のパラメータ推定が行なわれていることが推測される．

図 2 は，年間所得を常用対数変換して正規分布をあてはめた場合である．ヒストグラム上に正規分布の確率密度関数をあてはめた場合，きれいにあてはまっていると思いたいのであるが，ボックスプロットでは，左側に多くの点があり，左右対称とはなっていないので，正規分布とは言い難い．

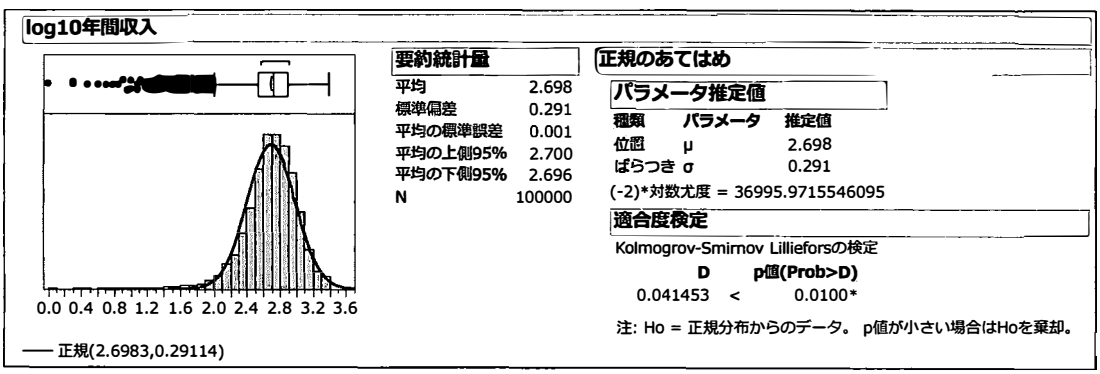



図 2　年間収入の常用対数変換値に正規乱数分布をあてはめた場合(匿名データ)

図 3 に示した正規確率プロットの直線からの乖離が，年間収入の低いほうで大きくなるので，複数の母集団が存在し，そこからのサンプリングデータとみなさざるを得ない．

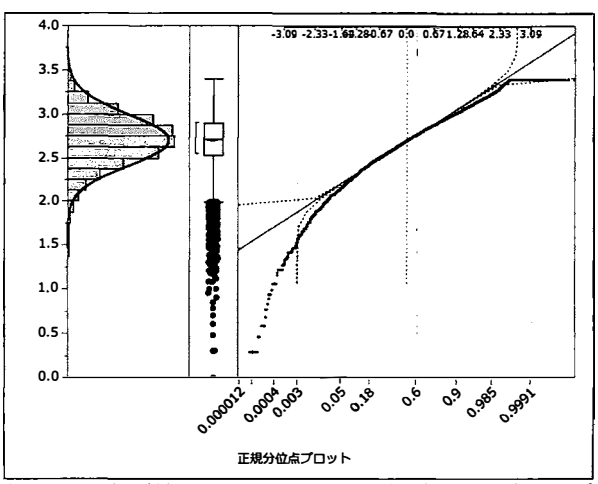
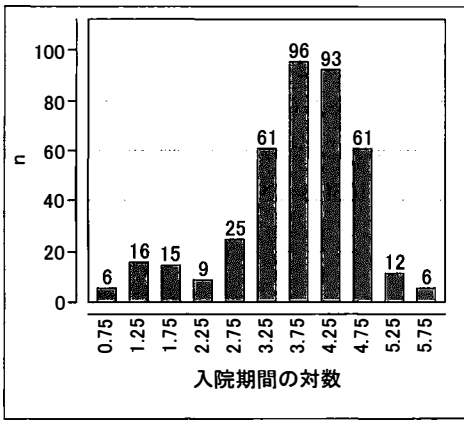



図 3 年間収入(対数)についての正規確率プロット(匿名データ)

## 4. JMP による多重正規混合分布のパラメータ推定

最尤法による多重正規混合分布のパラメータ推定については，吉村，大森，寒水(2009)の『医学・薬学・健康の統計学 理論の実用に向けて』「母数の推定法」(第 4 章)にモーメント法による推定と SAS/NLMIXED による推定が示されている．取り上げている例題は，「2 集団の混合分布」の節にあ

り，隔離が必要なある病気 T で，ある施設におけるある時期の患者 400 人の入院期間のヒストグラムが示されている．表は示されていないので，ヒストグラムの長さを計測して人数を推定した結果を図 4 に示す．




| $i$ | $y$ | $n_i$ |
|---|---|---|
| 1 | 0.75 | 6 |
| 2 | 1.25 | 16 |
| 3 | 1.75 | 15 |
| 4 | 2.25 | 9 |
| 5 | 2.75 | 25 |
| 6 | 3.25 | 61 |
| 7 | 3.75 | 96 |
| 8 | 4.25 | 93 |
| 9 | 4.75 | 61 |
| 10 | 5.25 | 12 |
| 11 | 5.75 | 6 |
| | $n =$ | 400 |

図 4　2 重正規混合分布らしきヒストグラム

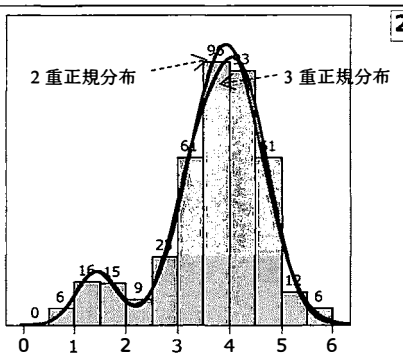
このヒストグラムは二山と感じられる．これを二つの正規分布 $N(\mu_1,\sigma_1^2)$，$N(\mu_2,\sigma_2^2)$ がある割合 $\pi : 1-\pi$ で混ざっているものと考える．母数 $\mu_1$，$\sigma_1^2$，$\mu_2$，$\sigma_2^2$，$\pi$ を最尤法で推定したい．

この例を一般化して，観測値が $n$ 個で，その値が $y_1, y_2, \ldots, y_n$ であるとしよう．これは確率密度関数 $f(y_i)$ が，

$$f(y_i) = \pi N(\mu_1,\sigma_1^2) + (1-\pi)N(\mu_2,\sigma_2^2)$$

で与えられる確率変数 $Y$ を，独立に $n$ 個観測した確率変数 $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ の実現値と考えられる．

JMP の「一変量のあてはめ」に，複数の正規分布の母集団パラメータ推定を行う機能があり，簡単にパラメータの推定が行える．あてはめた結果を図 5 に示す．




**2重正規混合のあてはめ**

**パラメータ推定値**

| 種類 | パラメータ | 推定値 |
|---|---|---|
| 位置 | μ1 | 1.425 |
| 位置 | μ2 | 3.946 |
| ばらつき | σ1 | 0.392 |
| ばらつき | σ2 | 0.700 |
| 割合 | π1 | 0.097 |
| 割合 | π2 | 0.903 |

(-2)*対数尤度 = 1050.25984656215

**3重正規混合のあてはめ**

**パラメータ推定値**

| 種類 | パラメータ | 推定値 |
|---|---|---|
| 位置 | μ1 | 1.454 |
| 位置 | μ2 | 3.273 |
| 位置 | μ3 | 4.205 |
| ばらつき | σ1 | 0.405 |
| ばらつき | σ2 | 0.471 |
| ばらつき | σ3 | 0.576 |
| 割合 | π1 | 0.102 |
| 割合 | π2 | 0.241 |
| 割合 | π3 | 0.657 |

(-2)*対数尤度 = 1051.20129941986

図 5　2 重正規混合分布

図 5 の「2 重正規混合のあてはめ」の欄から，第 1 母集団の割合 $\pi_1 = 0.097$，平均 $\mu_1 = 1.425$，標準偏差 $\sigma_1 = 0.392$，第 2 母集団の割合 $\pi_2 = 0.903$，平均 $\mu_1 = 3.946$，標準偏差 $\sigma_1 = 0.700$ と推定され，

ヒストグラム上に2重正規混合分布の密度関数が上書きされている．さらに，「3重正規混合のあてはめ」欄から，第1母集団の割合 $\pi_1 = 0.102$，平均 $\mu_1 = 1.454$，標準偏差 $\sigma_1 = 0.405$，第2母集団の割合 $\pi_2 = 0.241$，平均 $\mu_2 = 3.273$，標準偏差 $\sigma_2 = 0.471$ と推定され，第3母集団の割合 $\pi_3 = 0.657$，平均 $\mu_3 = 4.205$，標準偏差 $\sigma_3 = 0.576$ と推定さヒストグラム上に2重正規混合分布の密度関数が上書きされている．

ヒストグラム上の密度曲線の比較では，母集団数が2の場合に比べ母集団数が3の場合の違いは極わずかである．統計的には，母集団数が2の場合の(-2)対数尤度=1050.3に比べ，母集団数が3の場合の((-2)対数尤度=1051.2との差が，自由度2の $\chi^2$ 分布の上側5%点5.99よりも大きければ統計的には，母集団数が3が支持されるが，1051.2-1050.3=0.9 << 5.99と小さく母集団数が2の場合が採択される．

## 5. Excel による2重正規混合分布のパラメータ推定

Excel のソルバーを用いて最尤法により，未知母数 $\mu_1$，$\sigma_1^2$，$\mu_2$，$\sigma_2^2$，$\pi$ を推定したい．表1aに推定したいパラメータの初期値，母集団ごとに初期パラメータで計算された正規分布の確率密度が Normdist 関数で計算されている．それらの確率密度に対して混合割合を加味した対数尤度が-679.16と計算されている．対数尤度 $L$ は，

$$\ln L = \sum_{i=1}^{11} n_i f(y_i) = \sum_{i=1}^{11} n_i \ln\left[\pi N(\mu_1, \sigma_1^2) + (1-\pi) N(\mu_2, \sigma_2^2)\right]$$

なので，Excel シート中で逐次的に計算を積み上げる．

表1a　2重正規混合分布，初期値

正規混合分布の母数の推定

| $i$ | $y_i$ | $n_i$ | $L_{i1}$ | $L_{i2}$ | 対数 $L_i$ |
|---|---|---|---|---|---|
| 1 | 0.75 | 6 | 0.704 | 0.000 | -9.322 |
| 2 | 1.25 | 16 | 0.704 | 0.002 | -24.769 |
| 3 | 1.75 | 15 | 0.259 | 0.009 | -37.141 |
| 4 | 2.25 | 9 | 0.035 | 0.032 | -30.774 |
| 5 | 2.75 | 25 | 0.002 | 0.086 | -69.956 |
| 6 | 3.25 | 61 | 0.000 | 0.183 | -125.464 |
| 7 | 3.75 | 96 | 0.000 | 0.301 | -149.459 |
| 8 | 4.25 | 93 | 0.000 | 0.387 | -121.538 |
| 9 | 4.75 | 61 | 0.000 | 0.387 | -79.719 |
| 10 | 5.25 | 12 | 0.000 | 0.301 | -18.682 |
| 11 | 5.75 | 6 | 0.000 | 0.183 | -12.341 |

$n$ = 400　　対数 $L$ = -679.164　ここを最大化

変化させるセル5個　　$\pi$ = 0.500

$\mu_1$ = 1.000　　$\mu_2$ = 4.500

$\sigma_1$ = 0.500　　$\sigma_2$ = 1.000

$i$ = 1 の場合

$L_1$ =NORMDIST(C5,$E$21,$E$22,FALSE)

$L_2$ =NORMDIST(C5,$G$21,$G$22,FALSE)

対数 $L_i$ = D5 * LN( $F$19*E5 + (1-$F$19)*G5 )

対数 $L$ =SUM(H5:H15)

尤度$L_{i1}$の 1 行目は，$L_{1,1}$=0.704 と計算されている．これは，Excel の正規分布関数 `Normdist(0.75, 1.00,0.500,false)` で計算された正規分布の確率密度となっている．尤度$L_{i2}$の 3 行目は，$L_{3,2}$=0.009 と計算されている．これは，Excel の `Normdist(1.75,4.500,1.000,false)` で計算された正規分布の確率密度となっている．対数尤度$L_i$の 5 行目 対数尤度$L_5=-69.956$は，混合割合$\pi=0.3000$を用いて $25\times\ln(0.3\times0.02+(1-0.3)\times0.086)$ で計算されている．対数尤度$L=-679.164$は，対数尤度$L_i$，$i=1,2,...11$ の合計となっている．

表 1b には，ソルバーで対数尤度$L$を最大化するように 5 個のパラメータを探索的に動かした結果が示されている．推定されたパラメータは，JMP の結果と若干異なるが，収束精度など最適解を求めるアルゴリズムの違いに依存すると思われる．Excel の(-2)対数尤度は，2×524.95=1049.9 と JMP の 1051.2 より若干小さいので，Excel での計算結果が若干よい推定値と解釈される．

表 1b　2 重正規混合分布のパラメータの推定結果

正規混合分布の母数の推定

| $i$ | $y$ | $n_i$ | $L_1$ | $L_2$ | 対数$L_i$ |
|---|---|---|---|---|---|
| 1 | 0.75 | 6 | 0.250 | 0.000 | -22.161 |
| 2 | 1.25 | 16 | 0.838 | 0.000 | -39.725 |
| 3 | 1.75 | 15 | 0.722 | 0.004 | -38.762 |
| 4 | 2.25 | 9 | 0.161 | 0.030 | -28.317 |
| 5 | 2.75 | 25 | 0.009 | 0.131 | -53.152 |
| 6 | 3.25 | 61 | 0.000 | 0.346 | -71.178 |
| 7 | 3.75 | 96 | 0.000 | 0.546 | -68.054 |
| 8 | 4.25 | 93 | 0.000 | 0.519 | -70.658 |
| 9 | 4.75 | 61 | 0.000 | 0.297 | -80.490 |
| 10 | 5.25 | 12 | 0.000 | 0.102 | -28.657 |
| 11 | 5.75 | 6 | 0.000 | 0.021 | -23.794 |

$n$ = 400　　対数$L$ = -524.947　ここを最大化

変化させるセル5個　　$\pi$ = 0.099

$\mu_1$= 1.445　　$\mu_2$= 3.950

$\sigma_1$ = 0.429　　$\sigma_2$= 0.701

## 6. 年間収入についての多重正規混合分布のパラメータ推定

年間収入を常用対数変換して，正規分布をあてはめた結果を図 6 に示す．母集団＝2 と仮定した場合に，「2 重正規分布のあてはめ」欄の割合$\pi_1$が 13.4%の小集団で，平均値は位置$\mu_1$=2.388(244 万円)であり，割合$\pi_2$が 86.6%を占める大集団で，平均値は 2.746(557 万円)である．ヒストグラム上の確率密度関数のあてはめも母集団=1 と仮定した場合に比べ改善している．さらに母集団＝3 と仮定した場合には，「3 重正規分布のあてはめ」欄の割合(1.7%，8.0%，90.3%)の集団が分離されている．

統計的にどちらのあてはめが望ましいのかは，マイナス 2 倍の対数尤度を用いて評価する．表 2 に示すように，母集団=1 の場合は，(-2)*対数尤度=36995.97 で，母集団=2 の場合は，(-2)*対数尤

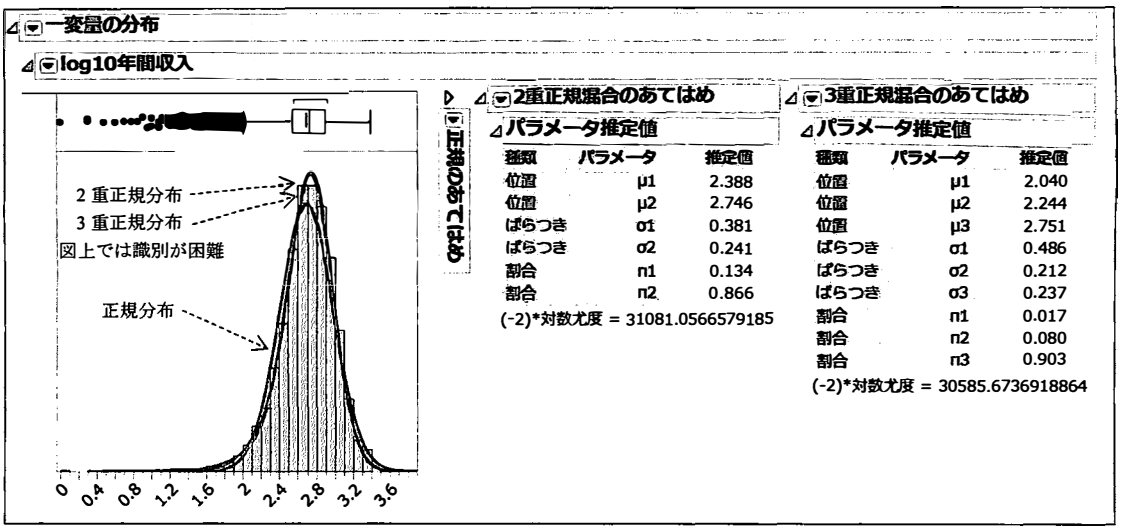



図 6 対数年間収入に対する正規混合分布のあてはめ(匿名データ)

度=31081.06 と 5914.91 の減少となる. 母集団=3 の場合は 30585.67 で, 母集団=2 の場合に比べ 495.39 小さくなる, 母集団=4 の場合は, 母集団=3 の場合に比べ 45.82 小さくなる. 母集団=5 の場合には変化が 0.01 である. 統計的には, $\chi^2$=5.99(自由度 2, 5%点)以下ならば, 母集団数の増加に意味がないので, 母集団=4 が統計的に支持される. しかし, 母集団=4 の場合, 最小母集団の構成は, 0.021%と極めて小さい集団なので, 母集団=3 とするのが現実的と思われる. その場合, 年間収入の母平均は, 110 万円, 175 万円, 563 万円となる.

表 2 仮定する母集団の数による対数尤度の差および最小母集団の統計量

| 母集団数 | (-2)対数尤度 | 対数尤度の差 | 最小の母集団 構成比 | 最小の母集団 対数母平均 | 最小の母集団 母平均(万) |
|---|---|---|---|---|---|
| 1 | 36995.97 | - | 100.0% | 2.698 | 498.9 |
| 2 | 31081.06 | -5914.91 | 13.4% | 2.388 | 244.3 |
| 3 | 30585.67 | -495.39 | 1.7% | 2.040 | 109.6 |
| 4 | 30539.85 | -45.82 | 0.021% | 0.513 | 3.3 |
| 5 | 30539.84 | -0.01 | 0.025% | 0.587 | 3.9 |
| | $\chi^2(df=1, 5\%)$ | 3.84 | | | |

## 7. 新擬似ミクロデータでの年間収入の分布

新擬似ミクロデータの年間所得の分布特性を, 匿名データと比較検討する. 図 7 に年間収入の 10 万世帯当たりの分布を示す. 匿名データは, 2500 万円でトップコーディングされていたが, 新擬似ミクロデータの場合は, 最大値が 6203 万円となっている. 平均値は, 609.9 と匿名データの 609.6 と, ほぼ等しい. 標準偏差は, 395.8 と匿名データの 388.7 よりも若干小さくなっている.

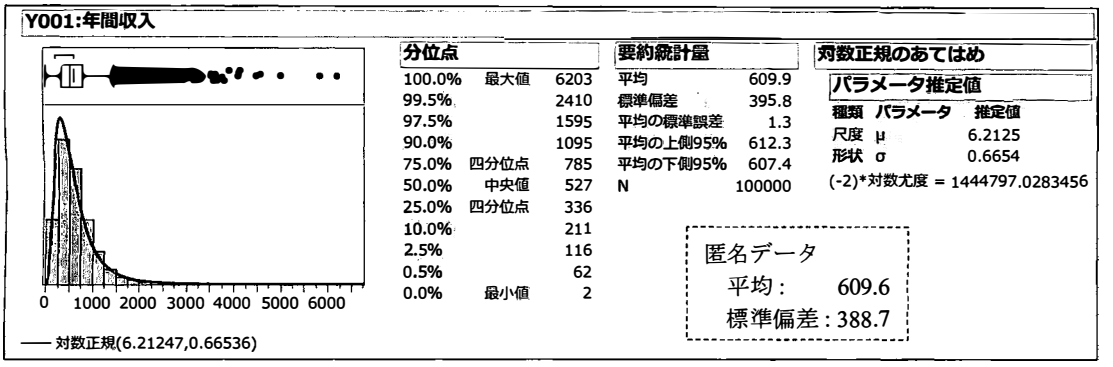



図 7　年間収入の分布(擬似ミクロデータ)

擬似ミクロデータの年間収入に，3 重正規混合分布をあてはめた結果を，図 8 に示す．匿名データの図 2 と比較して，平均は同じであるが，標準偏差は若干小さくなっている．

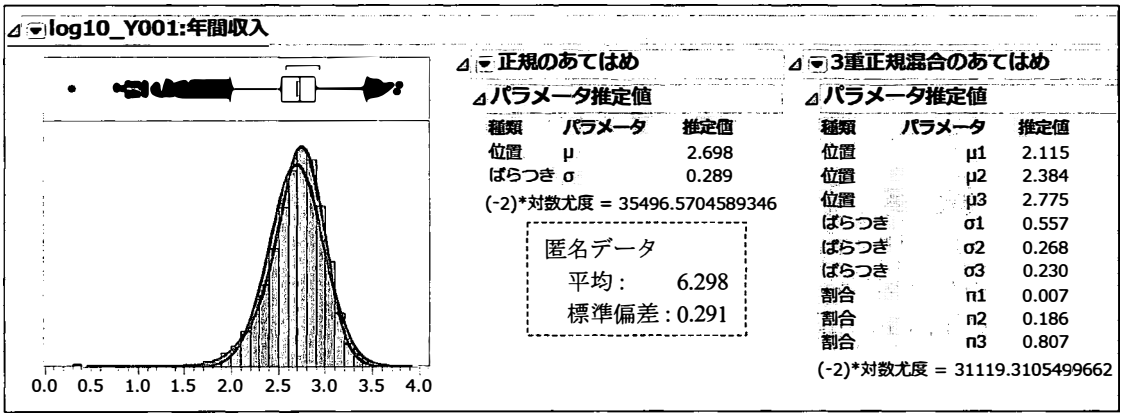



図 8　対数年間収入の分布(擬似ミクロデータ)

図 8 の擬似ミクロデータの 3 重正規混合分布のパラメータと図 3 の匿名データと比較した結果を表 3 に示す. 匿名データの第 1 母集団の位置 $\mu_1$ パラメータは 2.040 であり, 金額に直すと $10^{2.040}=110$ 万円となり，割合は $\pi_1=1.7\%$ である．新擬似ミクロデータでは，130 万円，0.7%である．第 2 母集団，第 3 母集団も，ほぼ同等と言い切れないが，何とか再現している．

表 3　3 重正規混合分布のパラメータ比較

| | 匿名データ | | | 新擬似ミクロデータ | | |
|---|---|---|---|---|---|---|
| 母集団 | 位置 $\mu$ | 万円 | 割合 $\pi$ | 位置 $\mu$ | 万円 | 割合 $\pi$ |
| 1 | 2.040 | 110 | 1.7% | 2.115 | 130 | 0.7% |
| 2 | 2.244 | 175 | 8.0% | 2.384 | 242 | 18.6% |
| 3 | 2.751 | 564 | 90.3% | 2.775 | 596 | 80.7% |

## 8. 住居費と教育費の分布の特徴

消費に関連する項目は，大きな分類項目から，中項目，小項目のような階層構造となっている．十大消費項目（食料, 住居, 光熱・水道, 家具・家事用品, 被服及び履物, 保健医療, 交通・通信, 教育, 教養娯楽, その他の消費支出）の合計は「消費支出」項目となる．この中で特徴的な住居費と教育費を取り上げる．

住居費と教育費には，多くの 0 円データが存在する．この 0 円データも含んだヒストグラムを図 9 に示す．両方とも右に長く裾を引くので，分布の全体像が見えにくい．

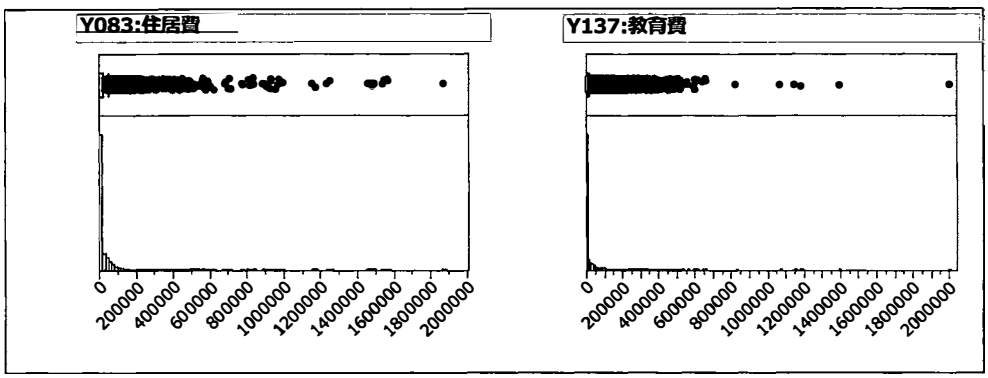



図 9　住居費と教育費の分布(擬似ミクロデータ)

常用対数変換すると 0 円データは，欠測値となってしまうので，1 円を加えて変換すると，$\log10(0+1)=0$ となり，金額が大きい場合には，1 円を加えても影響はわずかである．住居費も教育費もおおくの 0 円の世帯があることが図 10 に示されている．住居費の分布は，1/4 以上の 0 円世帯があり，$10^1$=10～$10^5$=100,000 円/月までは，ほぼ直線的に増加し，そこから急激に減少し $10^6$=1,000,000 円/月以上は，ごくわずかである．教育費は，半数以上の世帯で 0 円の世帯があり，$10^{4.5}$ =31,622 円/月あたりをピークにフラットな分布となっている．

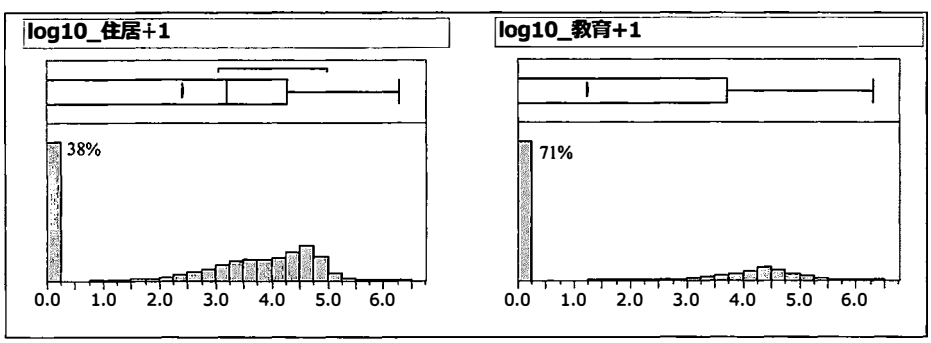



図 10　1 円を加えた対数住居費と教育費の分布(擬似ミクロデータ)

支出金額が 0 円が欠測値となる常用対数変換をして, 図 11 に示すように母集団の数を同定する．母集団の数を順次増やし，線の幅以上の変化がなくなった最大の数を母集団の数としたときに，住居費および教育費も同じ母集団の数は 3 となった．図 11 には，母集団の数を 5, 4, 3 と減らした時の確率密度曲線を重ね書きしたものであり，目視では識別が困難である．

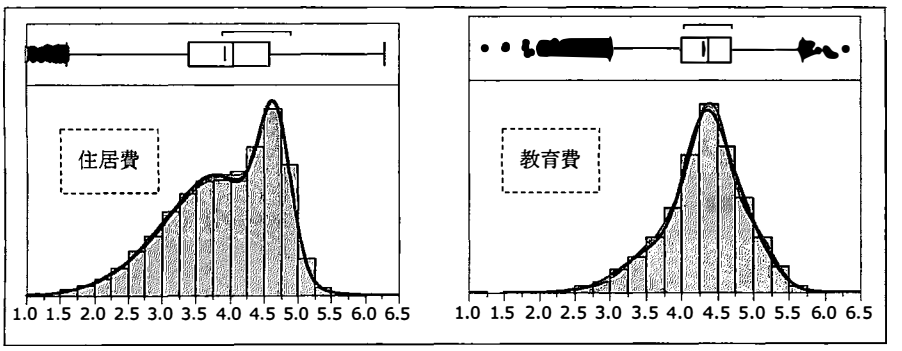



図 11　対数住居費と教育費に対する 3,4,5 重正規混合分布のあてはめ(擬似ミクロデータ)

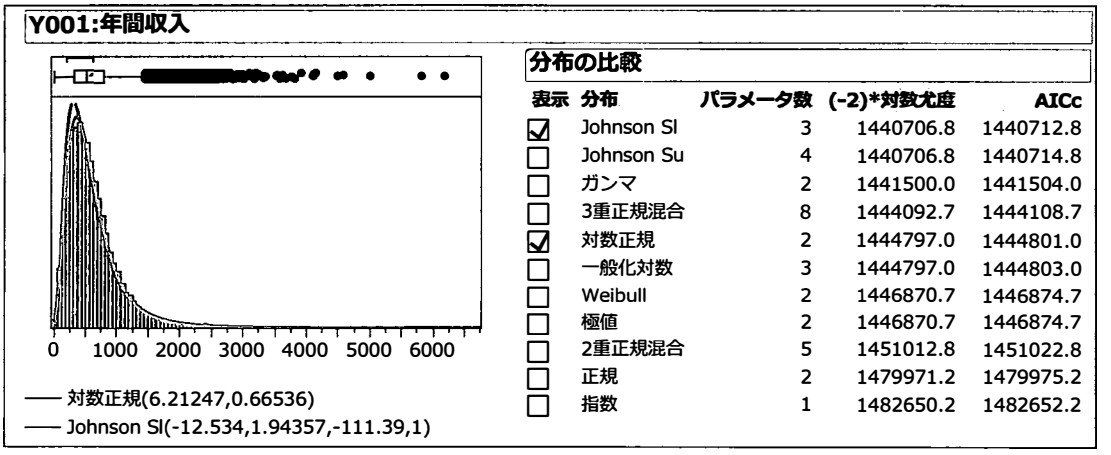
住居費の母集団の位置は，508 円，6,578，46,479 円/月となり，その割合は 11.5%，60.1%，28.8%と第 2 母集団が過半数を占めている．教育費の母集団の位置は，5,061, 20,380, 52,658 円/月となり，その割合は 26.1%，34.0%，39.9%とほぼ均等となっている．

表 4　3 重正規混合分布のパラメータ比較

| | 住居費 | | | 教育費 | | |
|---|---|---|---|---|---|---|
| 母集団 | 位置 $\mu$ | 円 | 割合 $\pi$ | 位置 $\mu$ | 円 | 割合 $\pi$ |
| 1 | 2.706 | 508 | 11.5% | 3. 704275 | 5,061 | 26.1% |
| 2 | 3.818 | 6,574 | 60.1% | 4.309195 | 20,380 | 34.0% |
| 3 | 4.667 | 46,479 | 28.4% | 4.721467 | 52,658 | 39.9% |

## 9. 年間収入の分布を 3 重正規混合分布と断定できるのか

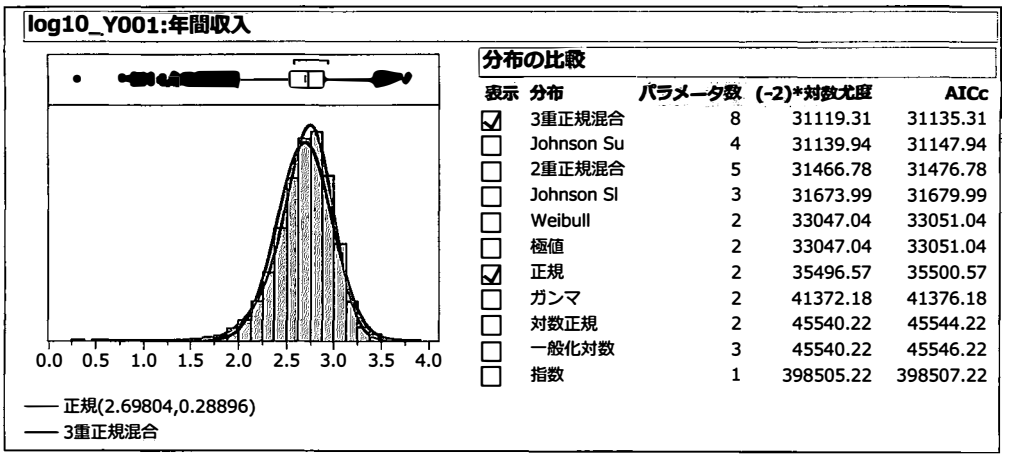
図 7 に年間収入に対して対数正規分布をあてはめを示したが，他の分布も含めて総合的に判断したい．JMP の「一変量のあてはめ」には，正規，対数正規，正規混合以外にも図 12 に示すように，多くの連続分布をあてはめた結果の(-2)対数尤度および AICc が出力されている．中ほどの対数正規分布のあてはめは，あまりよくないようである．




| 表示 | 分布 | パラメータ数 | (-2)*対数尤度 | AICc |
|---|---|---|---|---|
| ☑ | Johnson Sl | 3 | 1440706.8 | 1440712.8 |
| ☐ | Johnson Su | 4 | 1440706.8 | 1440714.8 |
| ☐ | ガンマ | 2 | 1441500.0 | 1441504.0 |
| ☐ | 3重正規混合 | 8 | 1444092.7 | 1444108.7 |
| ☑ | 対数正規 | 2 | 1444797.0 | 1444801.0 |
| ☐ | 一般化対数 | 3 | 1444797.0 | 1444803.0 |
| ☐ | Weibull | 2 | 1446870.7 | 1446874.7 |
| ☐ | 極値 | 2 | 1446870.7 | 1446874.7 |
| ☐ | 2重正規混合 | 5 | 1451012.8 | 1451022.8 |
| ☐ | 正規 | 2 | 1479971.2 | 1479975.2 |
| ☐ | 指数 | 1 | 1482650.2 | 1482652.2 |

図 12　年間収入に全ての連続変数をあてはめた結果(擬似ミクロデータ)

常用対数変換した年間収入に全ての連続変数をあてはめた結果を図 13 に示す．AICc による判定で 3 重正規混合分布(4 重以上は対象外)が最も小さいことがわかる．

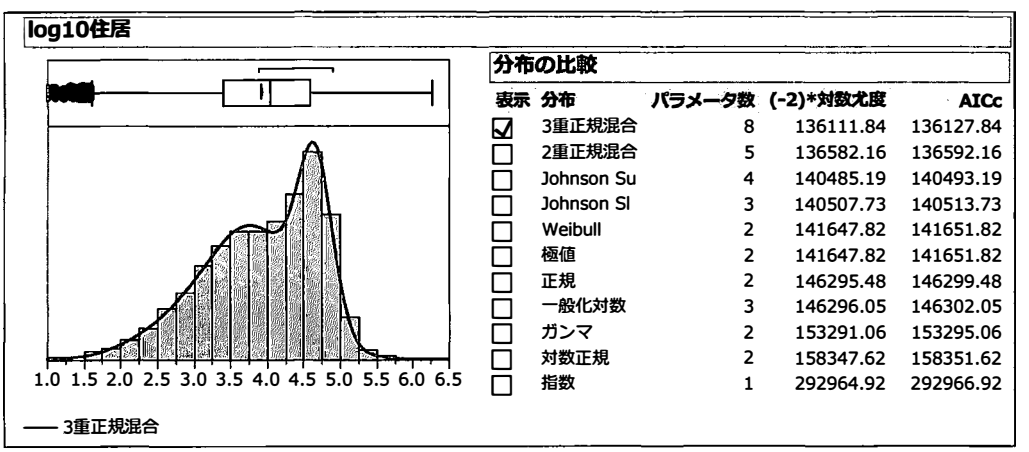



分布の比較

| 表示 | 分布 | パラメータ数 | (-2)*対数尤度 | AICc |
|---|---|---|---|---|
| ☑ | 3重正規混合 | 8 | 31119.31 | 31135.31 |
| ☐ | Johnson Su | 4 | 31139.94 | 31147.94 |
| ☐ | 2重正規混合 | 5 | 31466.78 | 31476.78 |
| ☐ | Johnson Sl | 3 | 31673.99 | 31679.99 |
| ☐ | Weibull | 2 | 33047.04 | 33051.04 |
| ☐ | 極値 | 2 | 33047.04 | 33051.04 |
| ☑ | 正規 | 2 | 35496.57 | 35500.57 |
| ☐ | ガンマ | 2 | 41372.18 | 41376.18 |
| ☐ | 対数正規 | 2 | 45540.22 | 45544.22 |
| ☐ | 一般化対数 | 3 | 45540.22 | 45546.22 |
| ☐ | 指数 | 1 | 398505.22 | 398507.22 |

図 13　対数年間収入に全ての連続変数をあてはめた結果(擬似ミクロデータ)

常用対数変換した住居費に全ての連続変数をあてはめた結果を図 14 に示す．AICc による判定で 3 重正規混合分布が最も小さい．この結果から一般化することはためらわれるが，家計に関連する収支項目について，常用対数変換した場合には，多重正規混合分布のあてはめが勧められる．




分布の比較

| 表示 | 分布 | パラメータ数 | (-2)*対数尤度 | AICc |
|---|---|---|---|---|
| ☑ | 3重正規混合 | 8 | 136111.84 | 136127.84 |
| ☐ | 2重正規混合 | 5 | 136582.16 | 136592.16 |
| ☐ | Johnson Su | 4 | 140485.19 | 140493.19 |
| ☐ | Johnson Sl | 3 | 140507.73 | 140513.73 |
| ☐ | Weibull | 2 | 141647.82 | 141651.82 |
| ☐ | 極値 | 2 | 141647.82 | 141651.82 |
| ☐ | 正規 | 2 | 146295.48 | 146299.48 |
| ☐ | 一般化対数 | 3 | 146296.05 | 146302.05 |
| ☐ | ガンマ | 2 | 153291.06 | 153295.06 |
| ☐ | 対数正規 | 2 | 158347.62 | 158351.62 |
| ☐ | 指数 | 1 | 292964.92 | 292966.92 |

図 14　対数住居費に全ての連続変数をあてはめた結果(擬似ミクロデータ)

## 10. 考察

全国消費実態調査の 2004 年の匿名データから作成された新擬似ミクロデータの年間収入，住居費，教育費について，どのような分布のあてはめが適切なのか検討した．常用対数変換した年間収入に対しては，3 重正規混合分布のあてははめが適していることが確認された．住居費あるいは教育費などでは，支出がなく 0 円の世帯がそれぞれ 10 万世帯比で 38%および 71%あり，これらを含めて分布のあてはめを行うことは適切でない．常用対数変換を行うことにより 0 円を欠測値にして

分布の検討を行った．

対数住居費は，一山型の分布とみなせないので，必然的に多重正規混合をあてはめることになる．母集団の数を順次増やして描かれた確率密度曲線の重なりから，最小の母集団数は3と推定できた．逆に母集団数を5とした場合から，減少させた場合からも母集団の数は3が適当と判断された．

対数教育費は，一山型の分布ではあるが，山の尖りが鋭く左右の裾が長い分布であり，3重正規混合が示唆される．他の常用対数変換した項目（米，パン，めん類，他の穀類）について分布のあてはめ結果でも，おおむね3重正規混合分布のあてはめが支持された．

このような分布の特性は，統計解析を行う際に十分考慮すべきであり，収支の金額(円)を用いて回帰分析を行うことは，全く推奨できない．常用対数変換を行っても誤差分布に正規分布を仮定するような統計解析も支持できない．収支金額が0円と世帯を除いた統計解析は，多くの世帯を無視することになり，バイアスに満ちた結果となってしまう．

このような特徴をもった全国消費実態調査データに対して，どのような統計的アプローチをしたら良いかは，これまでの Let's データ分析コンテストの規定課題に盛り込んできたので参考にしてもらいたい．

多重正規混合分布のあてはめについて，JMP の「一変量のあてはめ」は，画面操作により簡便に，計算スピードもデータ数が7万件でも数秒で行えることは，画期的である．文献的には EM アルゴリズムが提唱さてはいるものの，実際の適用例は見出すことができない．JMP では，ニュートン・ラフソン法による最尤法が使われているようであるが，多重正規混合分布の場合には，JMP の最尤解を初期値として Excel のソルバーを用いた最尤解を求めると，異なる結果となる．最大化する対数尤度を比較しても，JMP の方が小さく，適切な解とは言い難い．計算スピードを上げるために，何らかの計算の簡略化が行われているのかもしれない．

## 参 考 文 献


1) 周防節雄(2015) 全国消費実態調査の匿名データとその符号表から自動的に SAS のデータセット，変数ラベルと変数フォーマットを作成する SAS プログラム，『SAS ユーザー総会 2015 論文集』，pp257-278.
2) 高橋行雄(2015) 統計センターの匿名データ13万件を用いた統計解析の実践，『SAS ユーザー総会 2015 論文集』，pp145-164.
3) 高橋行雄(2016) JMP による第4回 Let 's データ分析の規定課題の解析，『SAS ユーザー総会 2015 論文集』，pp313-329.
4) 周防節雄(2016) SAS ユーザー総会 2016 における「Let's データ分析第4回ミクロデータ分析コンテスト」の規定課題の SAS プログラム解説，『SAS ユーザー総会 2016 論文集』，pp330-347.
5) 高橋行雄(2016) 統計センター提供の教育用擬似ミクロデータを用いた SAS/ JMP によるデータ分析コンテスト，ht ps: //wwwnst ac. go.jp/services/ pdf/161125_3-3.pdf.
6) 高橋行雄(2017) Let 's データ分析コンテストに用いる新擬似ミクロデータの概要，『SAS ユーザー総会 2017 論文集』，掲載予定.
7) 周防節雄(2017) 第5回 Let's データ分析コンテストの規定課題の SAS プログラム，『SAS ユーザー総会 2017 論文集』，掲載予定.

# 統計理論

# GAMプロシジャ versus GAMPLプロシジャ

○伊庭 克拓[1]

([1]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)

GAM Procedure versus GAMPL Procedure

Katsuhiro Iba
Otsuka Pharmaceutical Co. ,Ltd.

## 要旨

一般化加法モデル(Generalized Additive Models: GAM)は,一般化線形モデル(Generalized Linear Models: GLM)の線形予測子を平滑化スプライン等のノンパラメトリック回帰の加法モデルに拡張したモデルである。SASでは,従来のバージョンよりGAMプロシジャによって,またSAS/STAT 14.1より追加されたGAMPLプロシジャによってGAMによる解析を行うことができる。GAMプロシジャでは,Hastie & Tibshiraniの局所評点化法を用いてモデルの当てはめを行う。一方,GAMPLプロシジャはWood法を用いており,モデルの当てはめと共に平滑化スプラインにおける平滑化パラメータの最適化を行う等いくつかの変更点がある。本稿では,両プロシジャの実行方法の説明および実行結果の比較を行う。
キーワード:一般化加法モデル,GAM,GAMPL,平滑化スプライン

## 1.はじめに

一般化線形モデル(Generalized Linear Models: GLM)は,従来の線形モデルを応答変数の分布に正規分布,二項分布,ポアソン分布等を含んだ指数型分布族を仮定できるように拡張したモデルであり,様々な分野で広範に用いられている。正規線形モデル,ロジスティック回帰モデル,ポアソン回帰モデルはGLMの特別な場合である。SASではGENMODプロシジャで一般化線形モデルによる解析を実行できる。GLMは説明変数の効果に線形性を仮定しており,非線形な効果に対しては多項式や変数変換を用いるアプローチがあるが,適切な関数型を見つけることは簡単ではなく,またパラメトリックな関数では非線形な効果を十分に表現できない場合もある。

一般化加法モデル(Generalized Additive Models: GAM)は,GLMの線形性の仮定を緩め,GLMの線形予測子を説明変数の滑らかな関数の和に拡張したモデルである。滑らかな関数として平滑化スプラインやLOESS等のノンパラメトリック回帰を用いることができ,データ適応的に非線形な効果をモデル化する。

SASでは,従来のバージョンよりGAMプロシジャによって,またSAS/STAT 14.1より新たに追加されたGAMPLプロシジャによってGAMによる解析を実行できる。GAMプロシジャでは,Hastie & Tibshiraniの局

所評点化法を用いてモデルの当てはめを行う。一方，GAMPL プロシジャは Wood 法を用いており，平滑項を平滑化スプラインに限定することで，i) ペナルティ行列の低ランク近似による大規模データでの計算量の軽減，ii) 後退当てはめ法を用いずに全ての平滑化スプラインのパラメータを同時に推定，iii) ニュートンラフソン法等の数値最適化手法による平滑化パラメータの選択，といったいくつかの変更が行われている。

本稿では，最初に GAM と平滑化スプラインについて説明する。次に例題を用いて GAM プロシジャと GAMP L プロシジャの基本的な実行方法を説明する。続いて，GAM プロシジャと GAMP L プロシジャの構文をより詳しく紹介する。最後に SAS/STAT 14.1 User's Guide（以下，User's Guide）の Chapter 41 GAM プロシジャ及び Chapter 42 GAMP Lプロシジャの Getting Started及び Examplesの解析を双方のプロシジャで実行し，結果の比較を行う。

GLM については，丹後（2000）および田中ら（2008），GAM および平滑化スプラインについては辻谷 & 竹澤（2015），竹澤（2007）および Wood（2006）で詳しく説明されている。

## 2.一般化加法モデルと平滑化スプライン

### 2.1 一般化加法モデル（Generalized Additive Models: GAM）

GAM は，応答変数$y_i$の平均$\mathrm{E}(Y_i) = \mu_i, i = 1, \cdots, n$が説明変数$x_{i1}, \cdots, x_{ip}$によって，

$$g(\mu_i) = \eta_i = s_0 + s_1(x_{i1}) + \cdots + s_p(x_{ip})$$

と表されるモデルである。$g(\mu)$は GLM におけるリンク関数であり，$s_j(x_j), j = 1, \cdots, p$は平滑化関数である。各平滑項の解が一意になるように$\sum_{i=1}^{n} s_j(x_{ij}) = 0$の制約条件を設ける。$s_j(x_j) = \beta_j x_j$として線形項を含めることも可能である。

### 2.2 平滑化スプライン（Smoothing spline）

応答変数$y_i$は連続値とする。平滑化スプラインは，ペナルティ付き残差平方和

$$\sum_{i=1}^{n} (y_i - s(x_i))^2 + \lambda \int_{-\infty}^{\infty} \left( \frac{\partial^2 s(x)}{\partial x^2} \right)^2 dx \tag{1}$$

を最小化することで得られる。第 1 項は残差平方和であり，モデルのデータへの当てはまりが良いほど小さくなる。第 2 項は$s(x)$の凸凹に対するペナルティであり，$s(x)$が滑らかになるほど小さくなる。λ はモデルの当てはまりの良さと複雑さのトレードオフを調整する平滑化パラメータである。λ を小さくするとデータへの当てはまりが良くなり，最終的にデータに完全に当てはまる。一方，λ を大きくしていくと滑らかになり，最終的には 1 次式になる。

$$s(x) = \theta_0 + \theta_1 x + \frac{1}{12} \sum_{i=1}^{n} \delta_i |x - x_i|^3 \tag{2}$$

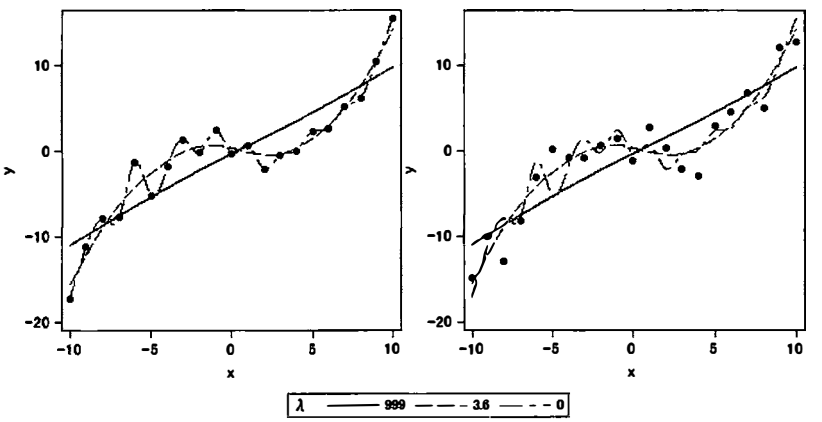
は 3 次の自然スプライン関数である。$\theta_0, \theta_1, \delta_i, i = 1, \cdots, n$が未知パラメータであり，自然スプラインには$\sum_{i=1}^{n} \delta_i = 0, \sum_{i=1}^{n} \delta_i x_i = 0$の制約が課される。$x_i, i = 1, \cdots, n$をノットと呼び，通常は説明変数の値と一致させるが，別の値にすることも可能である。

平滑化パラメータの選択について説明する。Figure 1 の左と右は同じ 3 次の多項式で誤差の値のみ異なっているシミュレーションデータである。左のデータのみを当てはめに用いて，平滑化パラメータの値を 999, 3.6, 0 と変化させたときの平滑化スプラインの予測値を描いた。当てはめに用いたデータ（左）に完全に当てはまるモデル（$\lambda = 0$）は，同じ現象に従う別データ（右）に対する当てはまりは必ずしも良いわけではな

いことが分かる。同じ現象に従う別データに対する予測精度を良くするという観点では，以下の Generali zed Cross Vali dati on(GCV）基準が最小になる平滑化パラメータを選択する。

$$\frac{n\sum_{i=1}^{n}(y_i - \hat{s}_\lambda(x_i))^2}{(n - \mathrm{tr}(\mathbf{H}_\lambda))^2}$$

GCV 基準は, L eave-one-outcross vali dati on 基準の近似である。$\mathbf{H}_\lambda$については 2.3 で説明する。Fi gurd の$\lambda = 3.6$は GCV 基準を用いて選択した結果であり，両方のデータにほど良く当てはまっている。




Fi gurd 平滑化スプラインの例

## 2.3 薄板平滑化スプライン（Thin-plate smoothing spline）

薄板平滑化スプラインは，平滑化スプラインを説明変数が 2 つ以上に拡張したもので，

$$\sum_{i=1}^{n}(y_i - s(\mathbf{x}_i))^2 + \lambda J_{m,d}(s) \tag{3}$$

を最小化することで得られる。$\lambda J_{m,d}(s)$はペナルティ項で，

$$J_{m,d}(s) = \int \cdots \int \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1! \cdots \alpha_d!}\left(\frac{\partial^m s}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}\right)^2 dx_1 \cdots dx_d$$

と表される。$s(\mathbf{x})$は薄板スプライン

$$s(\mathbf{x}) = \sum_{j=1}^{M}\theta_j\phi_j(\mathbf{x}) + \sum_{i=1}^{n}\delta_i\eta_{m,d}(\|\mathbf{x} - \mathbf{x}_i\|)$$

$$\eta_{m,d}(r) = \begin{cases} \dfrac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!\,(m-d/2)!}r^{2m-d}\log(r) & \text{if } d \text{ is even} \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!}r^{2m-d} & \text{if } d \text{ is odd} \end{cases}$$

である。第 1 項は$M = \binom{m+d-1}{d}$個のペナルティを受けない$m-1$次の多項式である。

$d = 1, m = 2$の場合は，(1)式および(2)式の平滑化スプラインに帰着する。$d = 2, m = 2$の場合，

$$J_{2,2}(s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left\{\left(\frac{\partial^2 s(x_1,x_2)}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 s(x_1,x_2)}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 s(x_1,x_2)}{\partial x_2^2}\right)^2\right\}dx_1 dx_2$$

および

$$s(x_1,x_2)=\theta_0+\theta_1x_1+\theta_2x_2+\sum_{i=1}^{n}\delta_i\frac{r_i^2}{8\pi}\log(r),\qquad r=\sqrt{(x_1-x_{i1})^2+(x_2-x_{i2})^2}$$

となる。平滑化スプラインと同様に$\sum_{i=1}^{n}\delta_i=0,\sum_{i=1}^{n}\delta_ix_{i1}=0,\sum_{i=1}^{n}\delta_ix_{i2}=0$の制約が課される。$x_1,x_2$それぞれの平滑化スプラインの加法モデル$s(x_1)+s(x_2)$では，$x_1,x_2$の交互作用は表現できないが，薄板平滑化スプライン$s(x_1,x_2)$ を用いると交互作用も表現できる。

$E_{ij}=\eta_{m,d}(\|\mathbf{x}_i-\mathbf{x}_j\|)$を要素とするペナルティ行列$\mathbf{E}$，$\phi_j(\mathbf{x})$を各行の要素とする行列$\mathbf{T}$を定義することで(3)式を以下の行列表現にできる。

$$\min\|\mathbf{y}-\mathbf{T}\boldsymbol{\theta}-\mathbf{E}\boldsymbol{\delta}\|^2+\lambda\boldsymbol{\delta}'\mathbf{E}\boldsymbol{\delta}\quad\text{subject to}\quad\mathbf{T}'\boldsymbol{\delta}=0\tag{4}$$

（薄板）平滑化スプラインでは，データ数が多くなるにしたがってパラメータ数も増加し，$\mathbf{E}$のサイズが膨大になることで(4)式を用いた計算が困難になる可能性がある。解決策の 1 つはノットの数を減らすことである。一方，Wood（2003）は$\mathbf{E}$の固有値分解による低ランク近似を行う方法を用いている。詳細は省くが，(4)式は最終的に次の最小化問題になる（User's Guide の p.2891）。

$$\min\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}\tag{5}$$

(5)式を最小にする$\widehat{\boldsymbol{\beta}}$は

$$\widehat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X}+\lambda\mathbf{S})^{-1}\mathbf{X}'\mathbf{y}$$

となる。よって，予測値は$\hat{\mathbf{y}}=\mathbf{X}\widehat{\boldsymbol{\beta}}=\mathbf{X}(\mathbf{X}'\mathbf{X}+\lambda\mathbf{S})^{-1}\mathbf{X}'\mathbf{y}=\mathbf{H}_\lambda\mathbf{y}$と表せる。$\mathbf{H}_\lambda$はハット行列と呼ばれ，$\mathrm{tr}(\mathbf{H}_\lambda)$は平滑化スプラインの有効自由度（実質的なパラメータ数）である。平滑化パラメータの値そのものは解釈が難しく，モデルの非線形の程度を表すのに有効自由度が用いられる。

## 2.4 一般化加法モデルの当てはめ

GAM プロシジャでは，応答変数が正規分布（加法モデル）の場合，以下の後退当てはめ法（backfitting algorithm）を用いる。

**後退当てはめ法（backfitting algorithm）**

1.初期化: $s_0=\bar{y},s_1^0=\cdots=s_p^0=0,m=0$

2.反復: $m=m+1$とし，$j=1,\cdots,p$について，偏残差$R_j$を応答変数とみなして$\hat{s}_j^{(m)}(x_j)$を求める。

$$R_j=y-s_0-\sum_{k=1}^{j-1}s_k^{(m)}-\sum_{k=j+1}^{p}s_k^{(m-1)}$$

3.$s_1^{(m)},\cdots,s_p^{(m)}$が変化しなくなるまで 2 を繰り返す。

応答変数が正規分布以外の指数型分布族の場合，以下の局所評点化法（local scoring algorithms）を用いる。

**局所評点化法（local scoring algorithms）**

1.初期化：$s_0=g(\bar{y}),s_1^0=\cdots=s_p^0=0,m=0$

2.反復：$m=m+1$

$\eta_i^{(m-1)}$および平均$\mu_i^{(m-1)}=g^{-1}(\eta_i^{(m-1)})$より，調整済み従属変数$z_i^{(m-1)}$および重み$w_i^{(m-1)}$を求める。

$$z_i^{(m-1)}=\eta_i^{(m-1)}+\left(y_i-\mu_i^{(m-1)}\right)\cdot\left(\frac{\partial\eta}{\partial\mu}\right)_i^{(m-1)},w_i^{(m-1)}=\left(V_i^{(m-1)}\right)^{-1}\cdot\left[\left(\frac{\partial\mu}{\partial\eta}\right)_i^{(m-1)}\right]^2$$

ここで，$V_i^{(m-1)}$は$\mu_i^{(m-1)}$における$Y_i$の分散である。

$z_i^{(m-1)}$を応答変数，各データの重みを$w_i^{(m-1)}$として，上述の後退当てはめ法を行う。

3.収束基準を満たすか逸脱度（deviance）$D(\boldsymbol{\mu})$が減少しなくなるまで 2 を繰り返す。

局所評点化法はGLMの当てはめに用いられる反復重み付き最小二乗法（Iteratively Reweighted Least Squares: IRLS）の各反復で重み付きの後退当てはめ法を行っているとみなせる。更に，各平滑項の平滑化パラメータを選択する必要がある。GAMプロシジャは平滑項ごとのGCV基準を用いて平滑化パラメータを自動選択するオプションがあるが，それを用いると局所評点化法が収束しない場合がある。

GAMPLプロシジャは，平滑項を平滑化スプラインに限定しており，その場合は後退当てはめ法は必要なく，$\mathbf{X} = [\mathbf{X}_0 : \mathbf{X}_1 : \cdots : \mathbf{X}_p]$，$\boldsymbol{\beta}' = [\boldsymbol{\beta}'_0 : \boldsymbol{\beta}'_1 : \cdots : \boldsymbol{\beta}'_p]$および

$$\mathbf{S}_\lambda = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_1 \mathbf{S}_1 & & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_p \mathbf{S}_p \end{bmatrix}$$

として，(5)式と同様の式を解くことで推定値が得られる。計画行列，パラメータベクトル，ペナルティ行列の添え字0は線形項，1~pは各平滑項に対応する。応答変数の分布が正規分布以外の場合，ペナルティ付き反復重み付き最小二乗法（Penalized Iteratively Reweighted Least Squares: P-IRLS）

$$\min \left\| \sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) \right\|^2 + \boldsymbol{\beta}' \mathbf{S}_\lambda \boldsymbol{\beta} \tag{6}$$

を用いる。$\mathbf{z}, \mathbf{W}$は局所評点化法で説明した調整従属変数と重みからなるベクトルと対角行列である。平滑化スプラインを用いたGAMの局所評点化法および(6)式は，ペナルティ付き対数尤度$\ell(\boldsymbol{\beta}) - 1/2\boldsymbol{\beta}'\mathbf{S}_\lambda\boldsymbol{\beta}$の最大化と等価である。

GAMPLプロシジャは平滑化パラメータの選択にWood（2004）のPerformance iteration法を用いる。

**Performance iteration 法**

1. 平滑化パラメータ$\boldsymbol{\lambda}$の初期値を与える。回帰パラメータ$\boldsymbol{\beta}$の初期値を与える。
2. 収束するまで以下を繰り返す。
   a. 調整従属変数$\mathbf{z}$，重み$\mathbf{W}$を求める。
   b. 平滑化パラメータ$\boldsymbol{\lambda}$の最適化を行う。
   c. 平滑化パラメータを使って線形予測子$\boldsymbol{\eta}$と予測値$\boldsymbol{\mu}$を得る。

2.bでは，モデル評価基準としてGCV基準もしくはunbiased risk estimator（UBRE）基準

$$\text{GCV}: \frac{n\|\mathbf{y} - \mathbf{H}_\lambda \mathbf{y}\|^2}{\mathrm{tr}(\mathbf{I} - \mathbf{H}_\lambda)^2}, \qquad \text{UBRE}: \frac{1}{n}\|\mathbf{y} - \mathbf{H}_\lambda \mathbf{y}\|^2 - \frac{2}{n}\sigma^2 \mathrm{tr}(\mathbf{I} - \mathbf{H}_\lambda)^2 + \sigma^2$$

を用いる。GAMPLプロシジャでは，スケールパラメータを持つ分布ではGCV基準が，そうでない場合はUBRE基準がデフォルトである。2.bでは，モデル評価基準の1次および2次導関数を求め，ニュートンラフソン法等の数値最適化法で平滑化パラメータを最適化する。Wood（2008）はP-IRLSの反復の外側で平滑化パラメータの最適化を行うOuter iteration法を提案しており，Performance iteration法より収束は遅いが安定していると言われている。SAS/STAT 14.1ではOuter iteration法は実装されているがExperimental（評価版）のため本稿では取り扱わない。

GAMでは，上記GCV基準の分子を逸脱度で置き換えたGCV逸脱度

$$\frac{nD(\boldsymbol{\mu})}{\mathrm{tr}(\mathbf{I} - \mathbf{H}_\lambda)^2} \tag{7}$$

も用いられることがある。本稿では，GAMプロシジャとGAMPLプロシジャを同じ基準で比較するために，実行結果の比較ではGCV逸脱度を用いる。

## 3.GAM プロシジャと GAMPL プロシジャの実行方法

GAM プロシジャと GAMPL プロシジャの実行例として，User's Guide の Chapter 41 Example 41.1 の脊柱後湾症データの解析を行う。椎弓切除術を受けた 83 例の小児患者について，応答変数として脊柱後湾症の有無（Kyphosis；1=あり，0=なし），説明変数として手術時の月齢（Age），何番目の脊椎から先を手術したか（StartVert）および手術した脊椎の個数（NumVert）が入力されている。GAM プロシジャと GAMPL プロシジャの実行例を Program 1 に示す。オプションを除けば，ほぼ同じプログラムで実行できる。

Program 1 GAM プロシジャと GAMPL プロシジャの実行例

```
ods graphics on;

proc gam data=kyphosis plots=components(additive clm);
model Kyphosis(event='1') =spline(Age) spline(StartVert) spline(NumVert) / dist=binomial method=gcv;
run;

proc gampl data=kyphosis plots;
model Kyphosis(event='1') =spline(Age) spline(StartVert) spline(NumVert) / dist=binomial;
run;
```

Output 1 GAM プロシジャの出力

Iteration Summary and Fit Statistics

| | |
|---|---|
| Number of local scoring iterations | 100 |
| Local scoring convergence criterion | 0.0000341057 |
| Final Number of Backfitting Iterations | 4 |
| Final Backfitting Criterion | 1.0915804E-9 |
| The Deviance of the Final Estimate | 51.209750057 |

WARNING: The local scoring algorithm did not converge.

Regression Model Analysis
Parameter Estimates

| Parameter | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | -1.66889 | 1.38043 | -1.21 | 0.2304 |
| Linear(Age) | 0.00787 | 0.00715 | 1.10 | 0.2746 |
| Linear(StartVert) | -0.19005 | 0.07180 | -2.65 | 0.0099 |
| Linear(NumVert) | 0.37633 | 0.18601 | 2.02 | 0.0465 |

Smoothing Model Analysis
Fit Summary for Smoothing Components

| Component | Smoothing Parameter | DF | GCV | Num Unique Obs |
|---|---|---|---|---|
| Spline(Age) | 0.999992 | 1.378752 | 60.494140 | 66 |
| Spline(StartVert) | 0.999605 | 0.954931 | 51.567921 | 16 |
| Spline(NumVert) | 1.000000 | 0.000000332 | 13.785016 | 10 |

Smoothing Model Analysis
Analysis of Deviance

| Source | DF | Sum of Squares | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Spline(Age) | 1.37875 | 10.547226 | 10.5472 | 0.0022 |
| Spline(StartVert) | 0.95493 | 3.921061 | 3.9211 | 0.0445 |
| Spline(NumVert) | 3.324874E-7 | . | . | . |

GAM プロシジャの出力を Output1 に示す。“Iteration Summary and Fit Statistics” はモデルの当てはめ結果の要約である。WARNING が出ているが，これは上述の GAM プロシジャで平滑化パラメータの自動選択を行う（method=gcv を指定）と局所評点化法が収束しない例である。しかしながら，反復計算過程を確認すると最初の数回の反復以降はモデルがほぼ変化しておらず，出力の説明が目的であるので，そのままの結果を示す。GAM プロシジャは，平滑項から線形項を分離する。例えば，$x_1, \cdots, x_3$に平滑化スプラインを仮定したモデル

$$\eta = \beta_0 + s_1(x_1) + s_2(x_2) + s_3(x_3)$$

は，GAM プロシジャでは以下のモデルに変換される。

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + s_1^*(x_1) + s_2^*(x_2) + s_3^*(x_3)$$

“Parameter Estimate” は線形項の推定値，“Fit Summary for Smoothing Components” は平滑項の当てはめ結果である。Smoothing Parameter の値は，λの代わりに$\lambda/(1+\lambda)$が示されている。また，表示されている自由度（DF）は分離した線形項の自由度 1 が除かれている。NumVert の自由度がほぼ 0 であるが，これは NumVert の影響が線形項でほぼ完全に説明でき，非線形効果が存在しないことを意味する。平滑項の効果は線形項のように単一のパラメータの値で要約できない。PROC GAM 文で plots オプションを指定すると平滑項の非線形効果をグラフィカルに表示できる（Figure2）。単に plots のみ指定した場合，$s^*(x)$が表示される。$s(x)$を表示するには components(additive)を追加する必要がある。“Analysis of Deviance” は非線形効果の有無（$H_0: s^*(x) = 0$）の検定結果である。

Output 2 GAMPL プロシジャの出力

適合度統計量

| | |
|---|---|
| 罰則付き対数尤度 | -26.94994 |
| 粗度ペナルティ | 2.68899 |
| 有効な自由度 | 6.12778 |
| 誤差の有効な自由度 | 75.87099 |
| AIC (小さいほどよい) | 63.46646 |
| AICC (小さいほどよい) | 64.61781 |
| BIC (小さいほどよい) | 78.28859 |
| UBRE (小さいほどよい) | -0.23419 |

パラメータ推定値

| パラメータ | 自由度 | 推定値 | 標準誤差 | カイ 2 乗値 | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -2.158016 | 0.474521 | 20.6822 | <.0001 |

平滑化成分の推定

| 成分 | 有効自由度 | 平滑化パラメータ | 粗度ペナルティ | パラメータ数 | ペナルティ行列のランク | ノット数 |
|---|---|---|---|---|---|---|
| スプライン(Age) | 2.11043 | 24744.0 | 1.6007 | 9 | 10 | 66 |
| スプライン(StartVert) | 2.01735 | 24.3226 | 1.0883 | 9 | 10 | 16 |
| スプライン(NumVert) | 1.00000 | 22215788 | 2.034E-7 | 9 | 10 | 10 |

平滑化成分の検定

| 成分 | 有効自由度 | 検定の有効自由度 | カイ 2 乗値 | Pr > ChiSq |
|---|---|---|---|---|
| スプライン(Age) | 2.11043 | 3 | 7.7823 | 0.0507 |
| スプライン(StartVert) | 2.01735 | 3 | 9.0739 | 0.0283 |
| スプライン(NumVert) | 1.00000 | 1 | 3.9377 | 0.0472 |

GAMPL プロシジャの出力を Output2 に示す。GAMPL プロシジャは線形項を分離しないため，線形項のパラメータが表示される“パラメータ推定値”には切片のみ表示されている。“平滑化成分の推定”は平滑項の当てはめ結果であり，自由度は GAM プロシジャの線形項と平滑項の自由度を足した値とほぼ同じになっている。PROC GAMPL 文で plots を指定すると，非線形効果をグラフィカルに表示できる（Figure3）。“平滑化成分の検定”は，平滑項のモデルへの寄与（$H_0: s(x) = 0$）の検定結果であり，GAM プロシジャと検定内容が異なるので結果の解釈する際に注意が必要である。

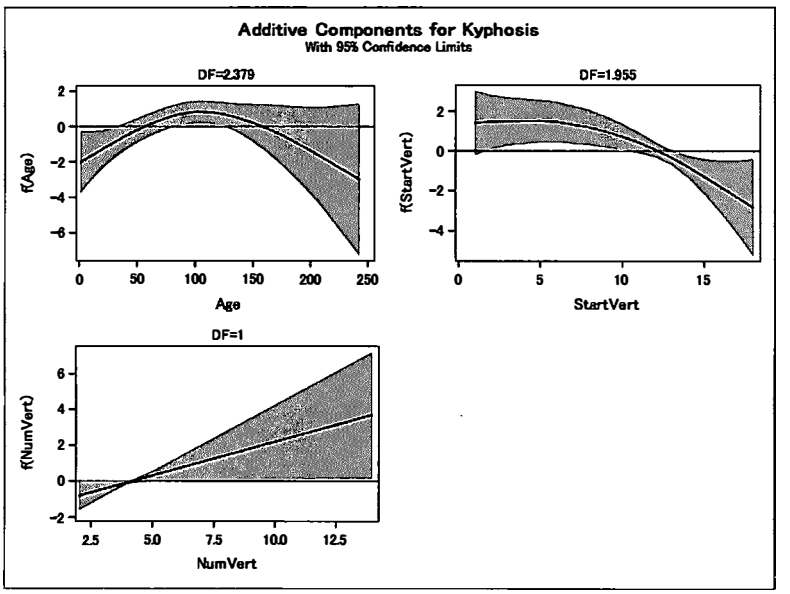



Figure 2 GAM プロシジャの各説明変数の効果

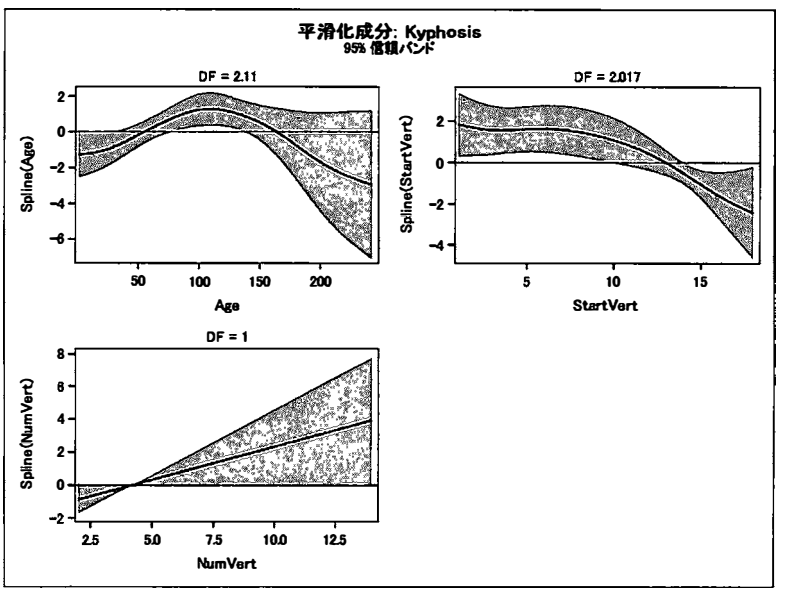



Figure 3 GAMPL プロシジャの各説明変数の効果

Figure2 および Figure3 の左上は手術時の月齢の非線形な効果を表しており，手術時の月齢とともに脊柱後湾症の発症率は増加していくが，約 100 カ月でピークを迎え，それを過ぎると発症率は減少に転ずる（小西 & 北川，2004）。

# 4.GAM プロシジャと GAMPL プロシジャの構文

## 4.1Syntax

GAM プロシジャと GAMPL プロシジャの Syntax は以下のとおりである（SAS，2015）。

```
PROC GAM < options > ;
CLASS variable < (options) > < variable < (options) > . . . > < / options > ;
MODEL dependent < (options) > = <PARAM(effects) > < smoothing-effects > < / options > ;
SCORE DATA= SAS-data-set OUT= SAS-data-set ;
OUTPUT OUT= SAS-data-set < keyword< =prefix > . . . keyword< =prefix > > ;
FREQ variable ;
```

```
PROC GAMPL < options > ;
CLASS variable < (options) >: : : < variable < (options) > > < / global-options > ;
MODEL response < (response-options) > = <PARAM(effects) > < spline-effects > < / model-options > ;
MODEL events / trials = <PARAM(effects) > < spline-effects > < / model-options > ;
OUTPUT <OUT=SAS-data-set > < keyword < =name > >. . . < keyword < =name > > < / options > ;
PERFORMANCE performance-options ;
FREQ variable ;
ID variables ;
WEIGHT variable ;
```

CLASS 文，FREQ 文および WEIGHT 文は，GENMOD プロシジャ等の回帰モデル系のプロシジャと同様の仕様であり，説明を省略する。GAMPL プロシジャの PERFORMANCE 文は SAS High-Performance Analytics に関する機能であり，本稿では省略する。

## 4.2MODEL 文

両プロシジャとも左辺に応答変数を指定する。応答変数が 2 値（ベルヌーイ分布）の場合に，追加オプションでどちらの値をイベントとするか指定できる。また，応答変数が二項分布の場合に，*event / trials* の 2 変数による指定も可能である。これは GENMOD プロシジャや LOGISTIC プロシジャと同様の仕様である。

右辺はモデル式を指定する。両プロシジャとも説明変数の効果に平滑化スプラインを仮定するには SPLINE () で囲う必要がある。また，線形項（CLASS 文で指定したダミー変数含む）を仮定する説明変数は PARAM() で囲う必要がある。両プロシャとも切片は自動で含まれる。指定できる効果を Table1 に示す。薄板平滑化スプラインの指定方法が異なるので注意が必要である。両プロシジャとも MODEL 文のオプション（/の後）で，

OFFSET=*variable* を指定することで，モデルにオフセット変数を加えることができる。

Table 1 MODEL 文で説明変数に仮定できる効果

| 効果 | GAM | GAMPL |
|---|---|---|
| パラメトリック項 | PARAM(variable1 variable2 ・・・) | PARAM(variable1 variable2 ・・・) |
| 平滑化スプライン | SPLINE(variable) | SPLINE(variable) |
| 薄板平滑化スプライン(2変数) | SPLINE2(variable1,variable2) | SPLINE(variable1 variable2) |
| 薄板平滑化スプライン(3変数以上) | × | SPLINE(variable1 variable2 ・・・) |
| LOESS | LOESS(variable) | × |

両プロシジャとも MODEL 文のオプションで DIST=*keyword* により応答変数の分布を指定できる。更に GAMPL プロシジャは LINK=*keyword* でリンク関数も指定できる。GAM プロシジャでは，リンク関数は正準リンク関数に限定されており，指定することが出来ない。指定できる分布とリンク関数を Table2 に示す（定義は User's Guide 参照）

Table2 指定できる分布とリンク関数

| 分布:keyword | リンク関数:keyword | GAM | GAMPL |
|---|---|---|---|
| 正規分布:NORMAL | Identity:IDENTITY | ○ | ○ |
| | Log:LOG | × | ○ |
| ベルヌーイ分布/2項分布:BINOMIAL | Logit:LOGIT | ○ | ○ |
| | Probit:PROBIT | × | ○ |
| | Complementary log-log:CLOGLOG | × | ○ |
| | Log-log:LOGLOG | × | ○ |
| ポアソン分布:POISSON | Log:LOG | ○ | ○ |
| ガンマ分布:GAMMA | Reciprocal:INV | × | ○ |
| | Negative reciprocal:非該当 | ○ | × |
| | Log:LOG | × | ○ |
| 負の二項分布:NEGATIVEBINOMIAL | Log:LOG | × | ○ |
| 逆ガウス分布:INVERSEGAUSSIAN | Reciprocal square:INV2 | ○ | ○ |
| | Log:LOG | × | ○ |

GAM プロシジャでは，MODEL 文のオプションで METHOD=GCV を指定することで，平滑項ごとの GCV 基準を用いた最適な平滑化パラメータの探索を行う。GAMPL プロシジャでは，METHOD=PERFORMANCE|OUTER で平滑化パラメータの最適化の方法を選択することが出来る。デフォルトは PERFORMANCE であり，OUTER は SAS/STAT 14.1 では Experimental（評価版）である。また，CRITERION=GCV| UBRE| GCAV で平滑化パラメータの最適化に用いる基準を指定できる。

GAM プロシジャは，SPLINE(変数,DF=*number*)で，各平滑化スプラインの自由度を任意の値に指定できる。GAMPL プロシジャは，SPLINE(変数/ Smoothing Options)で，平滑化スプラインをより詳細に調節できる。両プロシジャとも通常はユニークな説明変数の値をノットとする。GAMPL プロシジャは，ノット数が Smoothing Options の MAXKNOTS=*number* の値（デフォルトは 2000）を超える場合，ユニークな説明変数の値からランダムサンプリングで用いるノットを決める。サンプリングの結果を再現するには，GAMPL 文のオプションで SEED=*number* を指定する。Smoothing Options で MAXDF=*number* により最大の自由度（低ランク近似したペナルティ行列のランク）を指定できる。SMOOTH=*number* で平滑化パラメータを任意の値に指定できる。DF=*number* は指定した自由度になるよう平滑化パラメータの値が選択されるように見えるが，最大の自由度

が変更される（例えば，DF=3 および MAXDF=10 を指定するとペナルティ行列のランクが 3 になる）。他にも，平滑化パラメータの初期値や探索範囲の指定，ノットの調節なども可能である。詳細は User's Guide を参照されたい。

### 4.3OUTPUT 文

両プロシジャとも OUTPUT 文で予測値等を含んだデータセットを作成できる。その際，応答変数が欠測で説明変数が非欠測の observation の予測値も算出される。GAM プロシジャは MODEL 文で指定した説明変数および応答変数も作成されるデータセットに含まれるが，GAMPL プロシジャでは ID 文で指定した変数のみ作成されるデータセットに含まれる。出力できる項目の *Keyword* を Table3 に示す。*の項目は GAM プロシジャではデフォルトで平滑化スプラインごとの値が含まれるが，GAMPL プロシジャでは OUTPUT 文のオプションで COMPONENT を指定した場合のみ平滑化スプラインごとの値が含まれる。GAM プロシジャの SCORE 文では，モデルの当てはめに用いたデータセットとは別のデータセットを指定して，予測値を含んだデータセットを作成できる。

Table 3 OUTPUT文で出力できる項目

| | GAM | GAMPL | | GAM | GAMPL |
|---|---|---|---|---|---|
| 予測値 | PRED | PRED | 標準偏差(線形予測子)* | STD | STD |
| 線形予測子* | LINP | LINP | 下側信頼区間* | LCLM | LOWER |
| ピアソン残差 | RESID | PEARS | 上側信頼区間* | UCLM | UPPER |
| 残差 | × | RESID | ハット行列の対角要素* | ADIAG | × |

## 5.実行結果の比較

### 5.1 方法

User's Guide の Chapter 41 GAM プロシジャ及び Chapter 42 GAMPL プロシジャの Getting Started 及び Examples の解析を双方のプロシジャで実行し，結果を比較した。データおよびモデルの概要と実行結果を Table4 に示す（巻末）。モデルの当てはめ結果として，モデルの自由度（線形項と平滑項の総合計），逸脱度および(7)式の GCV 逸脱度を示した（両プロシジャの出力には含まれないが，予測値と自由度から容易に計算できる）。また，ログの実行時間（処理時間/CPU時間）も示した。

GAM プロシジャは METHOD=GCV で平滑化パラメータを選択した。GAM プロシジャはデフォルトでは，非線形効果の検定のために各平滑項を線形項に置き換えたモデルの再当てはめを行うが，処理時間が増加するので MODEL 文のオプションで ANNODEV=NOREFIT を指定し，複数モデルの当てはめを制限した。

GAMPL プロシジャは METHOD=PERFORMANCE で平滑化パラメータを選択した。GAMPL プロシジャでは，2 変量薄板平滑化スプラインの MAXDF はデフォルトでは 20 であるが，GAM プロシジャで自由度 20 を超えることがあったため，MAXDF=40 に変更した。また，SEED=12345 を指定した。

実行時間を評価する際，オプションは上記のみ最低限の指定としたが，両プロシジャはモデルの当てはめ以外の処理も行うため，実行時間は両プロシジャで採用されている手法の比較ではなく，あくまでプロシジャ間の比較である。また，PC のスペックおよび実行時の PC の状態にも依存するため厳密な比較ではない。

### 5.1 結果

実行時間の結果では，GAM プロシジャではデータ VOTE および MACK は 30 分くらい掛かっても結果が得られなかった。原因として GAM プロシジャはユニークなデータ値をノットにした計画行列とペナルティ行列をそのまま計算に用いるため，大規模なデータでは計算量が膨大になることが考えられる。ちなみに，両データとも，薄板平滑化スプラインを除くと計算は終了したことから，薄板平滑化スプラインの計算負荷が特に大きい可能性がある。KYPHO と PIMA では局所評点化法が収束せず，実行時間が局所評点化法の最大反復回数に依存するため，実行時間を表示していない。しかしながら，反復計算過程を見ると両データともに最初の数回の反復以降はモデルがほぼ変化していないことから，モデルの当てはめ結果を参考までに表示した。小規模なデータでは GAM プロシジャの方が速い傾向にあったが，ほとんどの場合 1 秒以内であり，体感できる差はなかった。

モデルの評価として，PIMA 以外のデータでは自由度は GAMPL プロシジャで小さい傾向であり，よって逸脱度は GAMPL プロシジャの方が若干大きかった。GCV 逸脱度は GAM プロシジャの方が小さいときと GAMPL プロシジャの方が小さいときがあり，値は大きく異ならないことから，モデルの良さに関して大きな違いは見られず，ほとんどの場合同じような解析結果が得られていた。PIMA では他のデータに比べて自由度の差が大きかった。GCV 逸脱度は GAM プロシジャの方が小さく，GAMPL プロシジャの平滑化パラメータを GAM プロシジャで得られた自由度に近くなるよう手動で調整すると元の GAMPL プロシジャの GCV 逸脱度より小さくなったことから，PIMA では GAMPL プロシジャで選択された平滑化パラメータは最適解に収束していないと考えられる。

## 6.まとめ

本稿では，GAM プロシジャと GAMPL プロシジャの特徴および実行方法を紹介し，いくつかのデータでの実行結果を比較した。

GAMPL プロシジャは，GAM プロシジャではサポートされていない負の二項分布を応答変数に仮定することができ，過分散（overdispersion）が生じているカウントデータの解析が可能である。User's Guide の Example 42.3 は負の二項分布を仮定した場合の解析も行っている。また，リンク関数の変更もでき，より広範なモデルを仮定することができる。一方，GAM プロシジャでは LOESS を用いた GAM も実行できる。

平滑化パラメータの選択において，GAM プロシジャで METHOD=GCV を指定すると局所評点化法が収束しない可能性がある。解決策としては，自由度の値を指定すれば収束するので，得られた自由度を参考に指定するか，自作のマクロで自由度を動かして最適な自由度を探索する等が考えられる。GAMPL プロシジャは Wood 法で平滑化パラメータの最適化が可能であり，P-IRLS は収束する。平滑化スプラインでは，平滑化パラメータの選択は重要な問題であり，多くの平滑化パラメータを持つモデルでは最適な平滑化パラメータの選択は複雑な問題になるので，アルゴリズムによって得られた平滑化パラメータが最適であるか確認することは有益であると考える。

解析対象のデータセットが大きくなると GAM プロシジャで計算が困難になる可能性があるが，そのような状況でも GAMPL プロシジャはペナルティ行列の低ランク近似を行うことで解析を実行できる可能性がある。しかしながら，ペナルティ行列の低ランク近似により，平滑化パラメータによる調整の前段階で自由度の上限が制限されるため，最大自由度（ペナルティ行列のランク）が十分かどうか注意を払う必要がある。なお，GAMPL プロシジャは，SAS High-Performance Analytics プロシジャに該当しており，本稿の解析は全て

シングルマシンモードで行っているが，環境によってはより大規模なデータの解析においても有用な可能がある。

本稿の結果は，User's Guide で取り上げられている限られたデータの範囲での検討に基づいており，実行した PC 環境も限定的であるため，別の状況において本稿と異なる傾向が見られるかも知れない。

以上，GAMPL プロシジャの登場により，SAS で GAM による解析を行う環境が拡充された。

## 謝辞

本稿について，貴重なご意見をいただきました大阪電気通信大学 名誉教授の辻谷将明先生に心より感謝申し上げます。

## 参考文献

小西貞則，北川源四郎 (2004). 情報量規準. 朝倉書店.

SAS Institute Inc. (2015). SAS/STAT(R) 14.1 User's Guide. Cary, NC, USA: SAS Institute Inc.

竹澤邦夫 (2007). みんなのためのノンパラメトリック回帰 第 3 版 上. 吉岡書店.

竹澤邦夫 (2007). みんなのためのノンパラメトリック回帰 第 3 版 下. 吉岡書店.

田中豊，森川敏彦，山中竹春，冨田誠 訳, Annette J. Dobson 著 (2008). 一般化線形モデル入門. 共立出版.

丹後俊郎 (2000). 統計モデル入門. 朝倉書店.

辻谷将明，竹澤邦夫 (2015). マシンラーニング 第 2 版. 共立出版.

Wood, S. (2003). Thin plate regression splines. Journal of the Royal Statistical Society, Series B 65:95–114.

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99:673–686.

Wood, S. (2006). Generalized Additive Models An Introduction with R, New York, Chapman & Hall.

Wood, S. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society, Series B 70:495–518.

## 補遺

GAMPL 文の SEED は 4.2 で説明したノットのランダムサンプリングの他にペナルティ行列の固有値分解の数値計算にも用いられるようである（User's Guide の p.2874）。前者は説明変数の値が十分に密であれば，モデルへの影響は少ないと思われる。後者は本質的にはモデルに影響しないと思われるが，数値計算が不安定になるデータでは影響するかも知れない。5 節の YIELD データを User's Guide の値のまま GAMPL プロシジャで SEED を指定せずに解析すると，実行ごとに結果が変わる現象が見られた。平滑化パラメータを固定しても（パラメータ推定値は一意に定まる），適合度統計量に表示される粗度ペナルティの値が実行ごとに変わったことから，低ランク近似後のペナルティ行列の値が変わっていると思われる。偶然だが説明変数 Catalyst を 1000 倍し，整数にしたところ改善された。

Table 4 GAM プロシジャと GAMPL プロシジャの実行結果の比較

| 用いたデータおよびモデル | | | | | 結果 | | |
|---|---|---|---|---|---|---|---|
| データ | 引用元 | Obs数 | モデル[a] | 分布 | | GAM | GAMPL |
| DIAB[b] | Getting Started: GAM Procedure | 43 | logCP = s(Age) + s(BaseDeficit)<br><br>logCP：C-ペプチド濃度の対数変 換，Age：年齢，BaseDeficit：不足塩基量 | 正規 | 自由度<br>逸脱度<br>GCV逸脱度<br>処理/CPU時間(秒) | 10.94<br>0.36<br>0.01523<br>0.27/0.01 | 9.70<br>0.38<br>0.01480<br>0.56/0.14 |
| KYPHO[b] | Example 41.1 | 83 | Kyphosis = s(Age) + s(StartVert) + s(NumVert)<br><br>Kyphosis：脊柱後湾症の有無，Age：手術　時の月齢StartVert：何番目の脊椎から先を手術したか，<br>NumVert：手術した脊椎の個数 | ベルヌーイ | 自由度<br>逸脱度<br>GCV逸脱度<br>処理/CPU時間(秒) | 6.33<br>51.21<br>0.7231<br>NA | 6.13<br>51.21<br>0.7193<br>0.12/0.14 |
| EQUIP[b] | Example 41.2 | 48 | removals = s(month)<br><br>removals：月あたりの撤去部品数，month：月 | ポアソン | 自由度<br>逸脱度<br>GCV逸脱度<br>処理/CPU 時間(秒 | 3.88<br>56.90<br>1.403<br>0.04/0.04 | 3.89<br>56.10<br>1.384<br>0.04/0.01 |
| YIELD[b] | Example 41.3 | 112 | Yield = s(Temperature,Catalyst)<br><br>Yield：化学反応の収率，Temperature：溶液の温度，Catalyst：加えた触媒の量 | 正規 | 自由度<br>逸脱度<br>GCV逸脱度<br>処理/CPU 時間秒) | 23.28<br>62.15<br>0.8844<br>0.07/0.06 | 21.84<br>64.73<br>0.8918<br>0.31/0.25 |
| VOTE[c] | Getting Started: GAMPL Procedure | 3107 | LogVoteRate = s(Pop) + s(Edu) + s(Houses) + s(Income) + s(Longitude,Latitude)<br><br>LogVoteRate：投票率の対数変 換，Pop：18歳以上の人口数，Edu：グレード12以上の学習を受けた人口数，<br>Houses：住戸数，Income：総所得，Longitude：経度，Latitude：緯度 | 正規 | 自由度<br>逸脱度<br>GCV逸脱度<br>処理/CPU 時間(秒) | NA<br>NA<br>NA<br>NA | 66.03<br>28.50<br>0.009576<br>4.22/10.99 |
| LIDAR[d] | Example 42.1 | 221 | LogRatio = s(RANGE)<br><br>LogRatio：2つのレーザー光源から受ける光量の比の対数変 換<br>RANGE：光が光源に反射される前に移動する距離 | 正規 | 自由度<br>逸脱度<br>GCV逸脱度<br>処理/CPU時間(秒) | 9.04<br>1.34<br>0.006596<br>0.04/0.04 | 9.00<br>1.33<br>0.006537<br>0.07/0.09 |
| PIMA[e] | Example 42.2 | 532 | Diabetes = s(Npreg) + s(Glucose) + s(Pressure) + s(Triceps) + s(BMI) + s(Pedigree) + s(Age)<br><br>Diabetes：糖尿病の有無，Npreg：妊娠回数，Glucose：ブドウ糖負荷試験の2時間血糖値，<br>Pressure：拡張期血圧，Triceps：上腕三頭筋皮下脂肪，BMI：BMI，Pedigree：糖尿病血統関数，Age：年齢 | ベルヌーイ | 自由度<br>逸脱度<br>GCV逸脱度<br>処理/CPU 時間秒) | 15.99<br>431.77<br>0.8627<br>NA | 22.91<br>424.17<br>0.8707<br>0.54/0.59 |
| MACK[f] | Example 42.3 | 634 | Egg_Count = s(Longitude,Latitude) + s(Depth) + s(Distance) + offset(Log_Net_Area)<br><br>Egg_Count：卵の数，Longitude：経度，Latitude：緯度，Depth：海底の深さ，<br>Distance：大陸棚からの距離，Log_Net_Area：収穫網の面積の対数変 換 | ポアソン | 自由度<br>逸脱度<br>GCV逸脱度<br>処理/CPU時間(秒) | NA<br>NA<br>NA<br>NA | 54.18<br>3848.33<br>7.257<br>2.63/2.20 |

a: s(変数名): 平滑化スプライン，s(変数名 1,変数名 2)：薄板平滑化スプライン，offset(変数名)：オフセット変数

b: User's Guide 中の SAS プログラムに記載されている。ただし，YIELD の Catalyst は元のデータを 1000 倍した値を用いた（補遺を参照）。

c: SASHELP ライブラリのデータセット Votel980 を使用した。

d: http:/ /www.stat.tamu.edu/~carroll/semiegbook/からダウンロードした。

e: http://archive.ics.uci.edu/ml/からダウンロードした。ただし，Glucose，Pressure，BMI および Triceps が 0 のデータは削除した（削除をすることで User's Guide 記載されているデータ数 532obs.に一致した）。Example 42.2 ではデータを training（モデル当てはめ用）と test（モデル評価用）に二分しているが，条件が分からないため，本稿では 532obs.をモデル当てはめに用いた。

f: フリーソフト R の gamair パッケージに含まれるデータを用いた。

# その他関連分野

# 歌舞伎「仮名手本忠臣蔵」の上演パターン

○坂部　裕美子
（公益財団法人　統計情報研究開発センター）

「仮名手本忠臣蔵」は、歌舞伎作品の中でも特に知名度が高く、上演回数も非常に多い。これは、もともと全11段に及ぶ長編物語である上に、全段の通し上演が多いことが大きく影響していると考えられるが、いずれかの部分を1～2幕抜き出して上演することも多くあり、全体からどの場面をどのように選んで上演するか、という個々の「上演パターン」は非常にバリエーションが大きい。この「忠臣蔵の上演パターン」を、SASを用いて集計した。

（1） 近年の上演について

日本俳優協会が作成した「歌舞伎公演データベース～戦後から現代まで」のデータを使用して1946年～2010年までの「忠臣蔵」上演パターンを調べた結果は、以下のとおりとなった。

| 上演された箇所 | 回数 | 上演された箇所 | 回数 |
|---|---|---|---|
| 落人（道行旅路の花聟）のみ | 40 | 八段目のみ | 14 |
| 七段目のみ | 22 | 大序・三・四・落人・五・六・七・八段目 | 5 |
| 九段目のみ | 19 | 大序・三・四・落人・五・六・七・九段目 | 4 |
| 大序・三・四・落人・五・六・七・十一段目 | 19 | 大序・二・三・四・落人・五・六・七・九・十一段目 | 4 |
| 五段目・六段目 | 15 | | |

最多は落人（道行旅路の花聟）の単独上演、次いで七段目の単独上演と、「上演パターン別」の頻度で見ると、1幕での上演が上位となる。

（2） 過去のすべての上演について

さらに、2016年に発売された「国立劇場上演資料集＜別冊＞　仮名手本忠臣蔵　上演年表　歌舞伎・文楽」というCDに収録されている、1748年の初演から2016年までの、現時点で把握できているすべての「忠臣蔵」上演データのうち、上演箇所が判明している分の記録を使用して同様の集計を行うと、以下のとおりとなった。

| 上演された箇所 | 回数 | 上演された箇所 | 回数 |
|---|---|---|---|
| 大序～十一段目（全段通し） | 389 | 五段目・六段目 | 54 |
| 大序～七段目 | 103 | 九段目のみ | 48 |
| 落人（道行旅路の花聟）のみ | 97 | 大序・三・四・五・六・七段目 | 40 |
| 大序～九段目 | 71 | 大序・三・四・落人・五・六・七・十一段目 | 32 |
| 七段目のみ | 64 | 八段目のみ | 28 |

初演当時は「全段通し」が基本上演形態だったこともあり、大半が通し上演となっているが、「七段目のみ」「九段目のみ」「五段目・六段目」というパターンはかなり以前から踏襲されていたようである。

# CDASHからSDTMを効率的に作成するためのTipsのご紹介

○小松 邦岳、鈴木 淳一、村岡 了一、飛石 陵子
(株式会社アスクレップ)

Rawデータからの SDTMの効率的な作成は、各社取り組んでいる大きな課題です。

CDASH標準を利用してCRFおよびRawデータを作成することで、SDTMの効率的な作成は可能ですが、注意すべきポイントがいくつかあります。また、EDCの仕様上の制約などにより、Rawデータが理想の定義にならないこともあるでしょう。

本発表では、EDC出力のRawデータからSDTMを効率的に作成するためのポイントと、そのために開発したツール（SASマクロ）のコンセプトを紹介いたします。

# Let's データ分析

# 若者の消費傾向はライフスタイルが変化するにつれ、どのように変化するか？：横断研究

参加カテゴリ：B）SAS または JMP の使用歴 3 年未満
チーム名：SA フレッシャーズ
谷　真宏、壁谷　勇佑、福澤　和輝
イーピーエス株式会社　統計解析 1 部

How do changes in lifestyle affect consumption trends in young people?
:a cross-sectional study.

Team name: SA Freshers
Mahiro Tani, Yuusuke Kabeya, Kazuki Fukuzawa
Statistics Analysis Department 1, EPS Corporation

**要約**：「さとり世代」の我々にとって、生活環境の違いが支出内訳に及ぼす影響を知ることは、今後のライフスタイルを検討する上で意義が大きい。そこで、全国消費実態調査に基づいた疑似ミクロデータを用いて、我々の同世代を、『単身世帯』、『夫婦のみ世帯』、『一子あり夫婦世帯』、『一子一人親世帯』の 4 群に分類し、消費実態を調査した。世帯数はそれぞれ 876 世帯、1,058 世帯、1,675 世帯、153 世帯であった。ライフサイクルが『一子あり夫婦世帯』へ進むにつれ、年齢が上がり、就業人数が増えることにより世帯収入は増大するが、一方で消費支出は主要耐久消費財やランニングコストが増加し、子供の保健医療や教育費が発生した。それに伴って、支出を抑えようと、娯楽支出を減らし、外食を抑えることで節約が行われていた。また『単身世帯』は他の世帯に比べ、飲料の消費が多かった。単身世帯のうちに外食費や飲料を節約する生活を送ることが、生活環境の変化に柔軟に対応するために大切であると分かった。

キーワード：疑似ミクロデータ、25～34 歳の世帯、さとり世代、格差社会

## 1.　はじめに

我々は入社一年目の会社員で、今年から社会人の仲間入りを果たした。我々の世代は、世間では「さとり世代」と呼ばれ、バブル崩壊後の不況の中で育ち、将来に漠然とした不安を抱え、浪費をしない傾向にあると指摘されている。実際に我々3 人も、「我々の生活環境はいつ変化するかわからない。自分の収入だけで、今後の生活を切り盛りできるだろうか。」という不安を抱いている。近年メディアでは、一人で子供を育て、経済的に苦しい暮らしをしている同世代や、定職に就けず、生活が困窮している同世代を度々目にする。また身近な友人では、既に結婚し、家庭を築いている同世代もいる。このように社会全体でみると、同世代の生活環境は様々であるといえる。

我々にとって、同世代の消費を上記のような世帯類型ごとに調査することは、生活環境の変化による消費の変化を知ることが出来、今後のライフスタイルを検討する上で意義が大きい。そこで、13 年前に我々のほぼ同年齢程度であった世代で、生活環境の違いが支出内訳にどう影響を及ぼすのかについて、平成 16 年　全国消費実態調査[1]（以下、全国消費実態調査）に基づく疑似ミクロデータ[2]を用いて調査した。全ての解析は SAS ver9.4（SAS Institute Inc., Cary, NC）で行った。

## 2. 方法

### 2.1. 対象集団

#### 2.1.1. 対象集団の抽出

本研究で用いる疑似ミクロデータ[2]は、全国消費実態調査に基づき統計データ利用推進及び教育用目的で疑似化されたデータである。分布特性は元の匿名データのそれに近似していることは確認されている。本データは元のデータの変数が一部削除されているため、対象集団を抽出するのに元のデータに比べ一定の制限がある。我々は、疑似ミクロデータにおける 69,131 世帯から、今回我々の調査したい若い世代を、生活環境の違いで分類するために、表 1 のような抽出条件を設けた。世帯人数に制限を加えたのは、世帯主または配偶者以外による収入の影響を除くためであり、また、子供の人数による影響を除くため、子供がいる世帯は一子とした。抽出した世帯を『単身世帯』、『夫婦のみ世帯』、『一子あり夫婦世帯』、『一子一人親世帯』とした。

**表 1**

| 共通条件 | | | |
|---|---|---|---|
| 就業人数 | 1～2人（無職を除外） | | |
| 世帯主の年齢 | 25～34歳 | | |
| | 世帯人数 | 家族分類 | 子供の有無※ |
| 単身世帯 | 1人 | 単身世帯 | なし |
| 夫婦のみ世帯 | 2人 | 夫婦のみ世帯 | なし |
| 一子あり夫婦世帯 | 3人 | 二世代の世帯 | あり |
| 一子一人親世帯 | 2人 | 二世代（一人親）の世帯 | あり |

※子供の有無・・・未就学児または学校に通う世帯員の有無

#### 2.1.2. 対象集団の世帯数

対象集団の各世帯数を図 1 に示す。

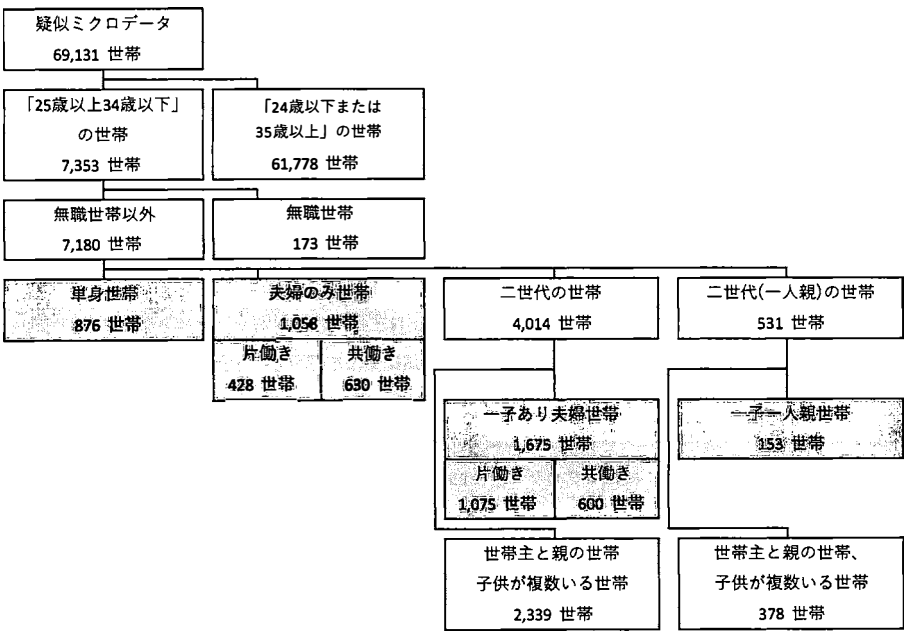



**図 1**

世帯数は『単身世帯』876 世帯、『夫婦のみ世帯』1,058 世帯、『一子あり夫婦世帯』1,675 世帯、『一子一人親世帯』153 世帯であった。また、『夫婦のみ世帯』1,058 世帯のうち、片働きの世帯及び共働きの世帯はそれぞれ 428 世帯と 630 世帯、『一子あり夫婦世帯』1,675 世帯では、それぞれ 1,075 世帯と 600 世帯であった。

#### 2.1.3. 疑似ミクロデータと全国消費実態調査及び国勢調査の比較

今回の対象集団のデータ抽出の妥当性と、結果の一般化可能性を検討するために、今回興味のある集団の世帯数の中で全国消費実態調査及び国勢調査の結果表と比較可能な世帯数を算出し、比較した。

結果を表 2 に示す。結果表は全国消費実態調査 世帯分布編の第 2 表[3]及び、平成 17 年国勢調査（以下、国勢調査）の第 3 表、第 6 表、第 7 表、第 10 表[4]を用いた。

表 2

| | 疑似ミクロデータ | | 全国消費実態調査 | |
|---|---|---|---|---|
| 25～34歳 | 対10万世帯比(※1) | | 対10万世帯比(※1) | |
| 総数 | 11,530 | | 12,320 | |
| | 疑似ミクロデータ | | 全国消費実態調査 | |
| 25～34歳 | 対10万世帯比(※2) | | 対10万世帯比(※2) | |
| 夫婦のみ | 1,915 | 37.1% | 1,922 | 35.5% |
| 一子あり夫婦 | 2,955 | 57.2% | 3,042 | 56.2% |
| 一人親 | 293 | 5.7% | 453 | 8.4% |
| 合計 | 5,163 | 100.0% | 5,417 | 100.0% |

| | 疑似ミクロデータ | 国勢調査 |
|---|---|---|
| 25～34歳 | 対10万世帯比(※1) | 対10万世帯比(※1) |
| | | |
| 総世帯数 | 11,530 | 13,612 |
| 夫婦のみ | 1,525 | 1,968 |
| 単身 | 4,052 | 6,055 |
| 一子あり夫婦 | 2,349 | 2,440 |
| 一人親 | 409 | 434 |
| 一子一人親 | 230 | 226 |

※1　総世帯を100,000世帯とした時の世帯数
※2　2人以上の世帯を100,000世帯とした時の世帯数

疑似ミクロデータと全国消費実態調査における各集団の世帯数の対 10 万世帯比は概ね一致した。

疑似ミクロデータと国勢調査における各集団の世帯数の対 10 万世帯比は、「単身世帯」以外の世帯でほぼ一致する結果となった。「単身世帯」については、疑似ミクロデータの世帯数の 10 万世帯比は国勢調査と比べて少なかったが、国勢調査は全例調査であるのに対し、全国消費実態調査は標本調査である。平成 12 年の国勢調査の結果をもとに、独自の手法[5]により算出した標本数をもとに調査している。「単身世帯」の標本数は少なく推定されていたものの、抽出の基準はほぼ同等である。そのため、10 年以上前の調査であるという点は考慮すべきであるものの、本研究の各世帯で得られた結果を日本の現在の同世代の各世帯の結果として一般化しても、大きなバイアスは生じないと考えられた。

### 2.2.　調査項目

対象集団の各世帯に対して以下の項目を調査した。調査にあたり、収入及び支出は各世帯人数分の経常収入及び消費支出とし、消費支出を 100 とした時の各支出の割合を評価した。加えて一子あり世帯については、子供の就学期分類ごとの世帯数を調査した（2.2.3.の項を参照）。

- 経常収入の分布
- 生活圏の内訳（世帯に関する項目の「大都市圏の別」を用いた。）
- 消費支出の内訳（2.2.1.の項を参照）
- 食費支出に対する『内食』・『中食』・『外食』の内訳（2.2.2.の項を参照）
- 『内食』支出に対する各食料費の内訳

#### 2.2.1.　消費支出

消費支出の内訳は、収支に関する項目のうち、「食料」、「住居」、「光熱・水道」、「家具・家事用品」、「被服及び履物」、「保健医療」、「教育」、「教養娯楽」、「交通・自動車等関係費」、「通信」、「その他の消費支出」の合計 10 個の変数を用いた。

#### 2.2.2.　『内食』、『中食』、『外食』の定義

『内食』、『中食』、『外食』の定義は先行研究[6]を参考に、収支に関する項目の「食料」のうち、「穀類」、「魚介類」、「肉類」、「乳卵類」、「野菜・海藻」、「果物」、「油脂・調味料」、「菓子類」、「飲料」、「酒類」の 10 項目を『内食』と定義した。また、「調理食品」を『中食』、「『外食』」を『外食』と定義した。

#### 2.2.3.　子供の就学期分類

子供の就学期を、世帯に関する項目の変数の「未就学児の有無」及び「学校に通う世帯員の有無」を用いて分類した。また、学校に通う世帯員有りの群を、収支に関する項目の「学校給食」の支払いがある群と無い群で分類し、以下の 3 群に分類した。

① 『未就学児群』
② 『給食費あり群』（小学生及び一部の中学生を想定）
③ 『給食費なし群』（一部の中学生及び高校生を想定）

2.2.4.　追加解析

調査項目の結果を確認した後、考察を行うために追加解析を行った。

但し、単身は1人分、
夫婦のみは1~2人分、
一子あり夫婦は1~2人分、
一子一人親は1人分、
の一ヶ月あたりの収入

3.　対象集団の収入及び特性に関する結果

3.1.　各世帯の経常収入の内訳

各世帯の経常収入の内訳を図 2 に示す

各世帯の経常収入の平均は、『単身世帯』で 26.9 万円、『夫婦のみ世帯』で 38.6 万円、『一子あり夫婦世帯』で 35.5 万円、『一子一人親世帯』で 19.0 万円であった。

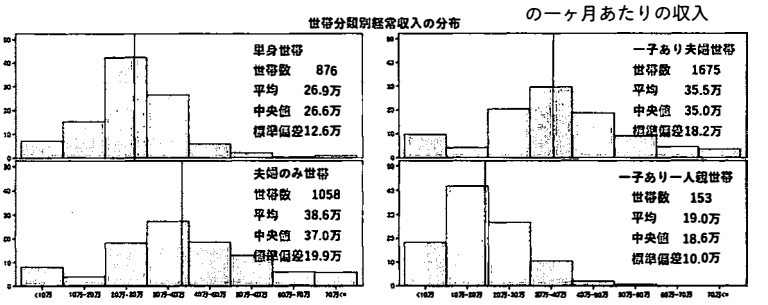



図 2

3.2.　各世帯の生活圏の内訳

各世帯の企業区分及び生活圏の内訳をそれぞれ表 3 に示す。

『一子一人親世帯』は、他の世帯と比較して「3 大都市圏」に住んでいる割合が高かったが、他の世帯では、「3 大都市圏」と「その他」に同程度の割合で住んでいた。

表 3

| | 抽出世帯 | |
|---|---|---|
| | 3大都市圏 | その他 |
| 単身世帯 | 431 (49.2%) | 445 (50.8%) |
| 夫婦のみの世帯 | 552 (52.2%) | 506 (47.8%) |
| 一子あり夫婦世帯 | 868 (51.8%) | 807 (48.2%) |
| 一子一人親世帯 | 96 (62.7%) | 57 (37.3%) |

3.3.　子供の就学期分類ごとの世帯数

一子あり世帯の子供の就学期分類ごとの世帯数を表 4 に示す。

表 4

| | 抽出世帯 | | | |
|---|---|---|---|---|
| | 未就学児 | 給食費あり | 給食費なし | 全体 |
| 一子あり夫婦世帯 | 1519 (90.7%) | 136 (8.1%) | 20 (1.2%) | 1675 (100.0%) |
| 一子一人親世帯 | 77 (50.3%) | 52 (34.0%) | 24 (15.7%) | 153 (100.0%) |

『一子あり夫婦世帯』では、子供が未就学児の世帯が 1,675 世帯中 1,519 世帯と 9 割以上を占めていた。それに対して、『一子一人親世帯』では、子供が学校に通っている世帯の割合が、『一子あり夫婦世帯』と比較して多く、約半数を占めた。

4.　対象集団の支出に関する結果と考察

4.1.　消費支出の内訳

消費費出の内訳を図 3 に示す。

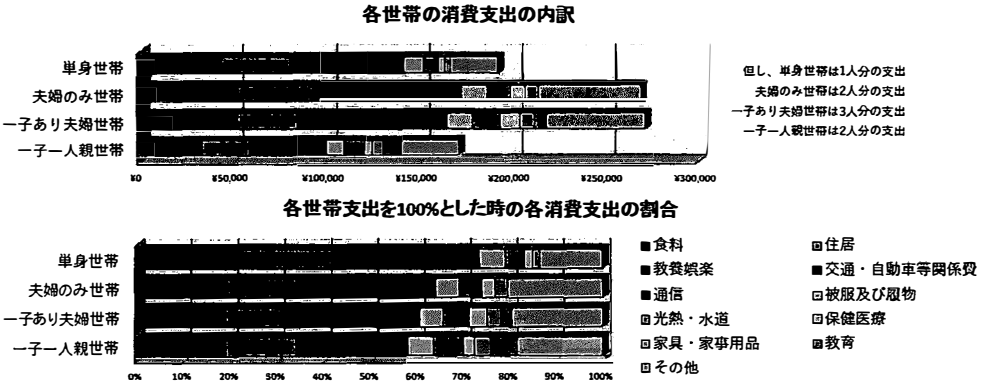



図 3

消費支出に占める割合は、「食料」、「被服及び履物」、「通信」は世帯間で大きな違いはなかった。2004 年はスマートフォンの普及前で、まだ通信支出を大きく占めていないことが要因であると考えられる。

また、「住居」、「家具・家事用品」、「保険医療」、「交通・自動車等関係費」、「教育費」、「その他支出」が各世帯の間で異なっていた。

「住居」は、割合でみると『単身世帯』で最も多くを占め、金額で見ても『単身世帯』が『一子あり夫婦世帯』を上回っていた。原因を調査するために追加で解析をしたところ、「住居」とは別に「土地家屋借金返済」の項目があり、『一子あり夫婦世帯』では 35.2%の世帯が支払っていた（表 5）。このことから、『一子あり夫婦世帯』は家屋を購入し、所持している世帯が比較的多いため、「住居」の内訳である「家賃地代」の支出が減少したと考えられた。また、経常収入が多いほど「土地家屋借金返済」の支出がある世帯が多かった。「家具・家事用品」の支出が単身世帯で少ないのも、家屋を所持していないことが原因であると考えられた。

「交通・自動車等関係費」について、追加解析として、「交通」と「自動車等関係費」に分け調査した。その結果『一子あり夫婦世帯』では、「交通」の支出が少なく、「自動車等関係費」の支出が多いことがわかった。自動車を所持している世帯を調べるために、「自動車保険料」の支払いがある世帯数を追加解析したところ、『一子あり夫婦世帯』で最も多く、56.9%の世帯で支払っていることが分かった（表 5）。

「保健医療」は『一子あり夫婦世帯』で多く、「教育」は一子あり世帯で多かった。どちらの支出も子供が出来ると増加するが、「保健医療」は未就学児で多く、「教育」は就学児で多く費用がかかることが原因であると考えた。

「教養娯楽」は、娯楽用品、書籍、月謝、旅行費、インターネット接続料などを含む。「教養娯楽」は特に単身世帯で多く、一子あり世帯で少なかった。内訳について、追加で解析をしたが、夫婦のみ世帯で旅行費が多いこと以外には各項目の比率に大きな違いは見られなかった。これらの費用は、子供が出来ることで時間を費やすことが出来ず、また、最初に節約される費用であることが原因だと考えられた。

「その他」は特に夫婦世帯で多い結果となった。「その他」について追加解析として、詳細に内訳を調べた結果を図 4 に示す。「その他」の内、こづかい（使途不明）が『夫婦のみ世帯』と『一子あり夫婦世帯』で約 2 万円を占めており他の世帯と異なっていた。世帯人員を考慮すると、交際費は単身世帯で比較的多く、一子あり世帯で少なかった。

### 4.2. 食費支出に対する『内食』・『中食』・『外食』の内訳

食費支出のうち、『内食』・『中食』・『外食』の内訳を図 5 に示す。但し、内食・中食・外食の分類基準は先行研究を参考として、「穀類」、「魚介類」、「肉類」、「乳卵類」、「野菜・海藻」、「果物」、「油脂・調味料」を内食、「調理食品」を中食、「外食」を外食とした[6]。

表 5

| | 各世帯数 | 家屋所持世帯※1 | | 車所持世帯※2 | |
|---|---|---|---|---|---|
| 単身 | 876 | 55 | 6.3% | 313 | 35.7% |
| 夫婦のみ | 1,058 | 278 | 26.3% | 459 | 43.4% |
| 一子あり夫婦 | 1,675 | 590 | 35.2% | 953 | 56.9% |
| 一子一人親 | 153 | 14 | 9.2% | 42 | 27.5% |

※1 土地家屋借金返済のある世帯
※2 自動車保険料の支払いのある世帯

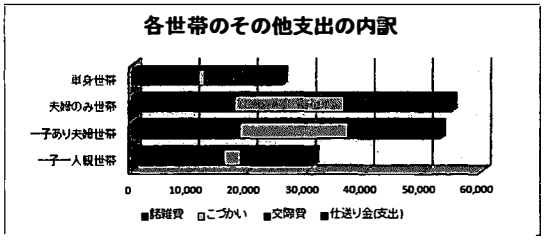



図 4

『単身世帯』では『外食』の割合が約 4 割と最も多くを占め、他の世帯と比較しても多かった。『中食』の割合はどの世帯においても大きな違いは無く、一方で『内食』の割合は『夫婦のみ世帯』、『一子あり夫婦世帯』及び『一子一人親世帯』で多かった。『一子あり夫婦世帯』では、片働き世帯が 64.2%であり、これらは育児（食育）または節約のため『内食』を増やす傾向があると考えられた。

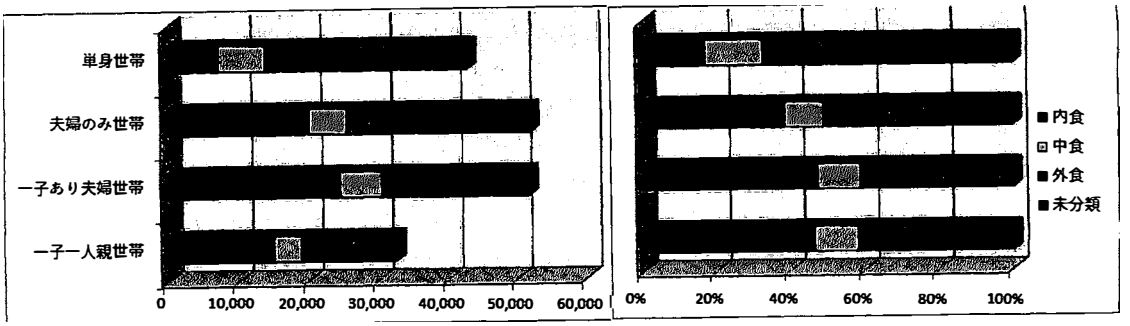



図 5

4.3.　『内食』支出に対する各食料費の内訳

『内食』支出のうち、各食料費の内訳を図 6,7 に示す。

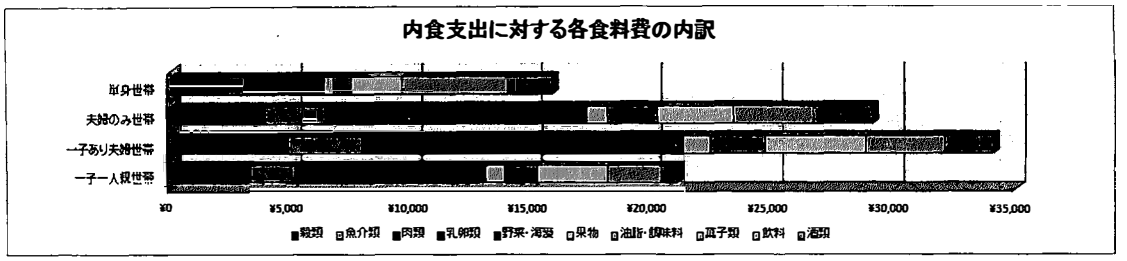



図 6

『単身世帯』では、特に『内食』支出に占める「飲料」の割合が他の世帯と比べて高く、金額も同様の結果が示された。また、「酒類」の割合が高く、「魚介類」、「肉類」、「野菜・海藻」の割合が低いことが分かった。これらのことから、『単身世帯』は『内食』においても調理が必要な肉や魚、野菜の消費が少なく、飲料や酒の消費が多いことが考えられた。

飲料は「茶類」、「コーヒー・ココア」、「他の飲料」で構成される。これらの内訳を調査したが、全ての飲料において、『単身世帯』での支出額が多かった。また『単身世帯』876 世帯中 61.4%の 538 世帯は男性で構成される。男女別でみると、男性の単身世帯での飲料の支出の平均は女性の 1.67 倍であった。このことから、男性の単身世帯では上記の傾向が顕著であると考えられた。

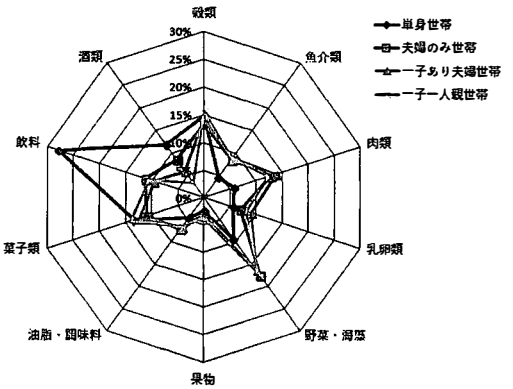



図 7

5.　まとめ

ライフサイクルは通常、『単身世帯』、『夫婦のみ世帯』、『一子あり夫婦世帯』(、場合によっては『一子一人親世帯』または『単身世帯』) という順に変化する。以上の結果から、ライフサイクルの変化に伴い、消費支出が次のような影響を及ぼすことが示唆された。

結婚後は配偶者の存在もあり世帯の収入は増加する。収入を共有することにより、使途不明の小遣い金が発生する場合がある。家屋を所持する世帯数が増えるので、家賃支出から土地家屋借金返済の支出に変化し、家具・家事用品の支出が増える。住居の変化に伴い、自動車を持つ世帯が増え交通費が増大する。子供を持つと、最初は保健医療費が、通学が始まれば教育費も発生する。

また、上記のような消費支出の増大に伴い、夫婦世帯や一子あり世帯では食費の節約を行い、消費支出の増大を抑えていると考えられる。一般的に外食は内食に比べて高額であるが、『単身世帯』は他と比べても『外食』の金額、割合がともに高かった。また、本研究で新たに男性の単身世帯で飲料の支出が多いことが分かった。このことから、『単身世帯』はライフサイクルが変化する前に、外食を控えること、また飲料を節約することで消費支出を抑えることが出来ると示唆された。しかし、世帯人数が増えれば清涼飲料やコーヒー等はパックで一度に大量に購入することができる。一方、単身では一度に大量に購入すると賞味期限までに消費しきれないため、支出がかさんでいる可能性がある。

本研究は横断研究であり、因果関係を判断することはできない。今後、複数年度の全国消費実態調査の結果を用いた研究を行うことが望まれる。疑似ミクロデータは、全国消費実態調査から品物の購入先や主要耐久消費財、貯蓄の現在高などの事項が削除されている。これらの変数を追加したデータを用いて研究を行うことが望まれる。本データは 2004 年のものと少々古く、晩婚化や未婚率の上昇、スマートフォンの普及、若者の車離れといった時代の変化が生じている。そのため本研究の結果から一概に、我々の同世代の生活環境の違いを示すことは難しい。以上より、本研究では幾つかの限界はあるものの、現代の新社会人の今後の生活に役立つ複数の知見が明らかとなった。我々3 人も、今後のライフスタイルを検討するうえで大いに参考としたい。

## 6. 参考文献


1. 総務省統計局. *平成 16 年度　全国消費実態調査*. 2004; Available at: http://www.stat.go.jp/data/zensho/2004/.（Accessed June 26,2017）
2. SAS 世話人会. *第 5 回　Let's データ分析　SAS/JMP によるミクロデータ分析コンテスト　新疑似ミクロデータ*. 2017; Available at: http://www.sascom.jp/lets/lets2017/.（Accessed June 26,2017）
3. 総務省統計局. *平成 16 年度　全国消費実態調査　全国　世帯分布編　報告書掲載表　第 2 表　世帯属性，都道府県別世帯分布*. 2004; Available at: http://www.e-stat.go.jp/SG1/estat/GL08020103.do?_toGL08020103_&tclassID=000000640073&cycleCode=0&requestSender=search. （Accessed June 26,2017）
4. 総務省統計局. *平成 17 年度国勢調査　新分類区分による遡及集計　全国結果*. 2011; Available at: http://www.e-stat.go.jp/SG1/estat/GL08020103.do?_toGL08020103_&tclassID=000001031076&cycleCode=0&requestSender=search. （Accessed June 26,2017）
5. 総務省統計局. *平成 16 年度　全国消費実態調査　用語の解説　付 7　調査世帯の選定方法と結果の推定式*. 2004; Available at: http://www.stat.go.jp/data/zensho/2004/kaisetsu.htm#furoku. （Accessed June 26,2017）
6. 3-girls from China 賈 慧玲、李 慧蛟、鄧 時韻. *Let's データ分析　ミクロデータ分析コンテスト　第 4 回　受賞作品　カテゴリーC　シングルマザー世帯の家計の実態－片働き世帯と共働き世帯との比較から－*. 2016; Available at: http://www.sascom.jp/lets/lets2016/.（Accessed June 26,2017）

**規定課題**

```
/*初期設定*/
%let path=フォルダパス名;
libname raw "&path.¥raw";
libname out "&path.¥out";
/*Weight の処理*/
ods output summary=weight;
proc means data=micro sum;
var weight;
run;
ods output close;
proc sql;
select weight_sum into: wei
from weight;
quit;
data micro1;
set micro;
calc_weight=weight*100000/&wei.;
run;
/*問題 1*/
data micro2;
set micro1;
output;
X02=4;
output;
run;
proc freq data=micro2;
table X09*X02/out=out1
  (drop=percent) ;
run;
proc freq data=micro2;
table X09*X02/out=out2
  (drop=percent) ;
   weight calc_weight;
run;
/*頻度 0 の対策：ダミーの作成*/
data dummy1;
do X09= 5 to 16;
do X02= 1 to 4;
count=0;output;
end;end;
run;
%macro outd1(num);
data out&num._1;
   merge    dummy1(in=x)
out&num.
   by X09 X02; if x;
run;
proc transpose data=
out&num._1 out=out&num._2;
   by X09;var count;
run;
data out&num._3;
   set out&num._2;
      nvar1 = round(col1,1);
      nvar2 = round(col2,1);
      nvar3 = round(col3,1);
      nvar4 = round(col4,1);
run;
%do i=1 %to 4;
proc sql;
   select sum(nvar&i.) into :
sum&i from out&num._3;
quit;
%end;
data data&num.;
   nvar1=&sum1;
   nvar2=&sum2;
   nvar3=&sum3;
   nvar4=&sum4;
   X09=100;
run;
data out&num._4
(keep=X09 var:);
   set out&num._3 data&num.;
      var%eval(6*(&num.-1)+1)        =
strip(put(nvar1,comma.));
      var%eval(6*(&num.-1)+2)        =
strip(put(nvar2,comma.));
      var%eval(6*(&num.-1)+3)        =
strip(put(nvar3,comma.));
      var%eval(6*(&num.-1)+4)        =
strip(put(nvar4,comma.));
run;
%mend;
%outd1(1);
%outd1(2);
data out1_5;
   length var1-var10 $200.;
   merge out1_4 out2_4;
   var5="";var6="";
   by X09;
run;
data out.res1;
   set out1_5;
   if X09=100 then var10=
"100,000";
run;
/*問題 2*/
data micro3;
   set micro1;output;
   X08=3;output;
run;
proc freq data=micro3;
   table X11*X08*X02/out=
out3(drop=percent) ;
   weight calc_weight;
run;
/*ダミー作成*/
data dummy3;
do X11=1 to 6;
do X08 = 1 to 3;
do X02=1 to 3;
count=0;output;end;end;end;
run;
data out3_1;
   merge dummy3(in=x) out3;
   by X11 X08 X02;
run;
proc sort data=out3_1;by X11
X02;run;
proc transpose data=out3_1 out=
out3_2;
   var count;by X11;
run;
proc sql;
create table data3 as
   select col1,col2,col3,col4,
            col5,col6,col7,col8,col9
   from
   (select   sum(col1)   as   col1   from
out3_2),
   (select   sum(col2)   as   col2   from
out3_2),
   (select   sum(col3)   as   col3   from
out3_2),
   (select   sum(col4)   as   col4   from
out3_2),
   (select   sum(col5)   as   col5   from
out3_2),
   (select   sum(col6)   as   col6   from
out3_2),
   (select   sum(col7)   as   col7   from
out3_2),
   (select   sum(col8)   as   col8   from
out3_2),
   (select   sum(col9)   as   col9   from
out3_2);
quit;
data out3_3;
   set out3_2 data3;
   if X11=. then X11=100;
col10=sum(col1,col2,col4,col5,col7,c
ol8);
run;
/*形式を整える*/
data out3_4(keep=X11 var:);
   set out3_3;
      var1    = strip(put(round
(col1,1),comma9.));
      var2    = strip(put(round
(col2,1),comma9.));
      var3    = strip(put(round
(col3, 1),comma9.));
      var4    = strip(put(round
(col4, 1),comma9.));
      var5    = strip(put(round
(col5, 1),comma9.));
      var6    = strip(put(round
(col6, 1),comma9.));
      var7    = strip(put(round
(col7, 1),comma9.));
      var8    = strip(put(round
(col8, 1),comma9.));
      var9    = strip(put(round
(col9, 1),comma9.));
      var10 = strip(put(round
(col10,1),comma9.));
run;
data out.res2;
   set out3_4;
run;
/*問題 3*/
/*①*/
data micro4;
   set micro1;
label Y044 = "パン支出金額（円";
logY044=log10(Y044+1);
run;
ods output summary=sum3_1;
proc means data=micro4 vardef=
wdf    max q3 median q1 min
mean;
var Y044;weight calc_weight;
run;
ods output close;
proc transpose data=sum3_1
out=out4_1;run;
ods output summary=sum3_2;
proc means data=micro4 vardef=
wdf    max q3 median q1 min
mean;
var logY044;weight calc_weight;
run;
ods output close;
proc transpose data=sum3_2
out=out4_2;run;
```

```
proc sort data=out4_1;by _label_;
run;
proc sort data=out4_2;by _label_;
run;
data out4;
   merge out4_1(drop=_name_
rename=(col1=pan))
            out4_2(drop=_name_
  rename=(col1=logpan));
   by _label_;
run;
data key;
  input key _label_ $40.;
cards;
1 最大値
2 上側四分位点
3 中央値
4 下側四分位点
5 最小値
6 平均
;
run;
proc sort data=key;by _label_;
run;
data out4_3;
   merge out4 key;
   by _label_;
run;
proc sort data=out4_3;by key;
run;
data out4_4(keep=out1 out2
  key);
   set out4_3;
   length out1-out2 $200.;
   out1=strip(put(pan,comma9.));
out2=strip(put(round(logpan,0.001),
5.3));
run;
/*件数追加*/
data kensu;
   length out1-out2 $200.;
   key=7;
   out1="100,000";
   out2="100,000";
run;
data out4_5;
   set out4_4 kensu;
run;
data out.res3_1;
   set out4_5;
run;
/*②*/
ods graphics / reset=index
imagename="res3_2"
imagefmt=bmp ;
ods html close;
ods listing gpath="&path.¥out";
ods graphics /ANTIALIASMAX
=3300;
proc sgplot data=micro4;
   vbox Y044;
   title "パン支出金額の箱ひげ図";
run;quit;
  ods listing close;
  ods graphics / reset=all;
/*  ods html; */ods listing;
/*③*/
proc univariate data=micro4 noprint;
   var logY044;
   histogram logY044;
  inset n="N" (8.0)
  mean="平均" (8.2)
  std="標準偏差" (8.3)
  / header='要約統計量' pos=NE;
  label logY044 ="log（パン支出金額
+1）（円）";
   title "パン支出金額+1 の常用対
数変換値のヒストグラム";
run;
quit;
/*問題 4*/
/*①*/
data micro5;
   length varname $200.;
   set micro1(where=(Y042^=0));
percent4=Y043/Y042*100;
varname="1:米";output;
percent4=Y044/Y042*100;
varname="2:パン";output;
percent4=Y045/Y042*100;
varname="3:めん類";output;
percent4=Y046/Y042*100;
varname="4:他の穀類";output;
run;
data micro5_1;
set micro5; output;
X11=7;output;
run;
/*dummy*/
data dummy_X02;
length out1-out6 $200.;
do X02=1 to 3;
out1="0";out2="-";out3="-";out4="-"
;out5="-";out6="-";output;end;
run;
data dummy_X11;
   length out1-out6 $200.;
   do X11=1 to 6;
out1="0";out2="-";out3="-";
out4="-";out5="-";out6="-";output;e
nd;
run;
%macro X02X11(var);
proc sort data=micro5_1;by &var.
varname;run;
ods output summary=out5_
&var.;
proc means data=micro5_1
mean;
   var percent4;
   by &var. varname;
   weight calc_weight;
run;
ods output close;
proc transpose data=out5_&var.
out=out5_1_&var. prefix=var;
   var percent4_Mean;
   by &var.;
run;
data out5_2_&var.;
   set out5_1_&var.;
   length out2-out6 $200.;
out2=strip(put(round(var1,0.1),5.1));
out3=strip(put(round(var2,0.1),5.1));
out4=strip(put(round(var3,0.1),5.1));
out5=strip(put(round(var4,0.1),5.1));
out6=strip(put(round(sum(var1,var2,
var3,var4),0.1),5.1));
run;
data out5_3_&var.;
   merge dummy_&var.
  out5_2_&var.;by &var.;
run;
%mend;
%X02X11(X02);
%X02X11(X11);
data out5_4;
   set out5_3_x02 out5_3_x11;
   keep out: sort;sort=_n_;
run;
data adj11;
    set micro1;output;
    x11 = 7;output;
run;
data adj02;
    set micro1;
run;
proc sort data=adj02 ;by x02;run;
proc means data=adj02 noprint;
    var calc_weight;
  by X02;output   out=tablex02
(drop=_freq_ _type_)   sum=var1
  / autoname;
run;
proc sort data=adj11 ;by x11;run;
proc means data=adj11 noprint;
var    calc_weight;by    X11;output
out=tablex11(drop=_freq_    _type_)
sum=var1 / autoname;
run;
data tablex02;
    length out1 $200.;
    set tablex02;
out1   = strip(put(round(var1,
  1),comma9.));
    sort=_n_;
run;
data tablex11;
    length out1 $200.;
    set tablex11;
    out1     =   strip(put(round(var1,
1),comma9.));
    sort=_n_ +3;
run;
data tablex00;
    set tablex02 tablex11;
run;
proc sort data=tablex00 out=
g01;by sort; run;
proc sort data=out5_4 out=
g02;by sort; run;
data out5_5;
    merge g02(in=inb) g01
(in=ina) ;
    by   sort;
run;
data out.res4_1(keep=out: sort);
   set out5_5;
run;
/*②*/
data out6;
set out5_5;
keep sort var:;
if sort>=4;
   var2=input(out2,5.1);
   var3=input(out3,5.1);
   var4=input(out4,5.1);
```

```
   var5=input(out5,5.1);
run;
proc transpose data=out6 out=
out6_1;
   var var2-var5;id sort;
run;
data out6_2;
   set out6_1;
   length varname $200.;
   label _name_     = "項目名"
     varname = "世帯(家族)分類"
       var        = "構成比（%）";
var=_4;varname="1";output;
var=_5;varname="2";output;
var=_6;varname="3";output;
var=_7;varname="4"; output;
var=_8;varname="5";output;
var=_9;varname="6";output;
var=_10;varname="7";output;
run;
proc format;
value $var "1" = "単身世帯"
              "2" = "夫婦のみ"
              "3" = "二世代"
         "4" = "二世代（ひとり親）"
             "5" = "三世代"
             "6" = "その他"
             "7" = "全体" ;
  value $name "var2" = "米"
                "var3" = "パン"
                "var4" = "めん類"
              "var5" = "他の穀類";
run;
proc sgplot data=out6_2;
   hbar   varname   /   response=var
group = _name_;
   format   varname   $var.   _name_
$name.;
run;
/*問題 5*/
/*①*/
%macro ginicalc(num);
%local sumweight;
%local sumY001;
%local obs;
data micro7_&num.;
set micro1;if X11=&num.;
run;
proc   sort   data=micro7_&num.;by
Y001;run;
data micro7_1_&num.;
   set micro7_&num.;
   retain cumweight 0;
   cumweight+calc_weight;
   retain cumY001 0;
   cumY001+Y001*calc_weight;
run;
ods output summary=
max1_&num.;
proc means data=micro7_1_
&num. max;
var cumweight cumY001;
run;
ods output close;
proc sql;
   select   cumweight_Max   into   :
sumweight from max1_&num.;
   select   cumY001_Max   into   :
sumY001 from max1_&num.;
   select count(*) into : obs from
micro7_1_&num.;
quit;
data micro7_2_&num.;
   set micro7_1_&num.;
cumweightp=cumweight/&sumweigh
t;
cumY001p=cumY001/&sumY001;xy=
_n_/&obs;
run;
%mend;
%ginicalc(4);
filename grafout "&Path.¥out¥
res5_1.bmp";
goptions device = bmp gsfname =
grafout gsfmode = replace;
ods graphics on;
symbol1 i=join v=dot c=red;
symbol2 i=join v=dot c=black;
proc gplot data=micro7_2_4;
plot cumY001p*cumweightp xy*
xy/ overlay;
label cumweightp = "人口の累積相
対度数"cumY001p     = "年間収入
の累積相対度数";title "「二世代（ひ
とり親）」の年間収入";
run;quit;
ods graphics off;
/*②ジニ係数の導出*/
%ginicalc(1);%ginicalc(2);
%ginicalc(3);%ginicalc(5);
%ginicalc(6);
%macro ginicalc2();
%do i=1 %to 6;
proc sql;
   select count(*) into : obs&i.
   from micro7_2_&i.;
quit;
data micro8_&i.;
   set micro7_2_&i.;
partsg=(xy-cumY001p)*1/&&obs&i.;
run;
ods output summary=gini_&i.;
proc means data=micro8_&i.
sum;var partsg;
run;
ods output close;
data gini__&i.;
   set gini_&i.;bunrui=&i;
run;
%end;%mend;
%ginicalc2;
data gini(drop=partsg_Sum);
   set gini__1-gini__6;
gini=strip(put(round(partsg_Sum,0.0
1),4.2));
run;
proc sort data=gini;by bunrui;
run;
data out.res5_2;
   set gini;
run;
/*エクセルへの出力*/
OPTIONS   NOXSYNC NOXWAIT ;
X   "&path.¥result¥Table_all.xls" ;
%macro   xlsout(in=,   var=,   sheet=,
range=);
    filename   excel   dde   "EXCEL   |
[Table_all.xls]&sheet.!&range.";
    data _null_;
        set &in.;
        file excel dlm = '09'x dsd
notab lrecl = 65488;
        put &var.;
    run;
    filename excel clear;
%mend;
%xlsout(in=out.res1,   var=var1-var4,
sheet=表 1,
range=R6C3:R18C6);
%xlsout(in=out.res1,
var=var7-var10,sheet=表 1,
range=R6C9:R18C12);
%xlsout(in=out.res2,  var=var1-var10,
sheet=表 2,
range=R6C3:R12C12);
%xlsout(in=out.res3_1,        var=out1
out2,sheet=表 3,range=R5C3:
R11C4);
%xlsout(in=out.res4_1,var=out1-out6
,sheet=表 4,range=R6C5:
R15C10);
%xlsout(in=out.res5_2,var=gini,sheet
=表 5,range=R5C3:R10C3);
%macro xlsr1c1(book=,sheet=);
    filename   sys   dde   'EXCEL   |
SYSTEM';
DATA _NULL_;
    FILE SYS;
    put          '[workbook.activate("'
"[Table_all.xls]&sheet." '")]';
    PUT '[SELECT("R1:R65536")]';
    PUT   '[FONT.PROPERTIES("ＭＳ
明朝")]';
    PUT '[SELECT("R1C1")]';
RUN;
    filename sys clear;
%mend;
%xlsr1c1(sheet=T5);
%xlsr1c1(sheet=T4);
%xlsr1c1(sheet=T3);
%xlsr1c1(sheet=T2);
%xlsr1c1(sheet=T1);
filename   excelsys   dde   'EXCEL   |
SYSTEM';
    data _null_;
    file excelsys lrecl = 65488;
put '[activate("' "Table_all.xls" '")]';
   put '[save()]';
    run;
filename excelsys clear;
```

**表1　年齢階層別・世帯区分別 クロス表**

| 世帯主の年齢 | 世帯区分 | | | 合計 |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | |
| 24歳以下 | 765 | 18 | 15 | 798 |
| 25～29歳 | 2,289 | 157 | 66 | 2,512 |
| 30～34歳 | 4,223 | 511 | 107 | 4,841 |
| 35～39歳 | 5,217 | 728 | 193 | 6,138 |
| 40～44歳 | 5,492 | 1,076 | 205 | 6,773 |
| 45～49歳 | 5,702 | 1,305 | 277 | 7,284 |
| 50～54歳 | 6,194 | 1,797 | 392 | 8,383 |
| 55～69歳 | 5,682 | 2,221 | 624 | 8,527 |
| 60～64歳 | 3,237 | 2,226 | 2,534 | 7,997 |
| 65～69歳 | 1,343 | 1,743 | 3,590 | 6,676 |
| 70～74歳 | 469 | 1,197 | 3,214 | 4,880 |
| 75歳以上 | 166 | 940 | 3,216 | 4,322 |
| 合計 | 40,779 | 13,919 | 14,433 | 69,131 |

| 世帯主の年齢 | 世帯区分 | | | 合計 |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | |
| 24歳以下 | 1,729 | 10 | 21 | 1,760 |
| 25～29歳 | 4,314 | 151 | 57 | 4,522 |
| 30～34歳 | 6,441 | 480 | 87 | 7,008 |
| 35～39歳 | 7,420 | 824 | 153 | 8,396 |
| 40～44歳 | 7,611 | 1,145 | 247 | 9,004 |
| 45～49歳 | 7,739 | 1,479 | 244 | 9,462 |
| 50～54歳 | 8,072 | 2,085 | 508 | 10,665 |
| 55～69歳 | 7,697 | 2,725 | 893 | 11,315 |
| 60～64歳 | 4,208 | 2,785 | 4,160 | 11,153 |
| 65～69歳 | 1,487 | 2,213 | 6,374 | 10,074 |
| 70～74歳 | 459 | 1,652 | 6,124 | 8,235 |
| 75歳以上 | 207 | 1,325 | 6,873 | 8,406 |
| 合計 | 57,384 | 16,874 | 25,741 | 100,000 |

**表2　世帯分類別・世帯区分別 世帯主の性別 3重クロス表**

| 世帯分類 | 勤労者世帯 | | | 勤労者以外の世帯 | | | 無職世帯 | | | 合計 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 男 | 女 | 計 | 男 | 女 | 計 | 男 | 女 | 計 | |
| 単身世帯 | 6,503 | 4,207 | 10,710 | 1,030 | 867 | 1,897 | 2,463 | 5,688 | 8,151 | 20,758 |
| 夫婦のみ | 8,517 | 324 | 8,841 | 4,293 | 81 | 4,375 | 11,026 | 45 | 11,071 | 24,287 |
| 二世代 | 27,267 | 378 | 27,645 | 6,190 | 51 | 6,242 | 3,264 | 32 | 3,295 | 37,182 |
| 二世代(ひとり親) | 1,543 | 2,421 | 3,963 | 934 | 430 | 1,364 | 671 | 1,303 | 1,974 | 7,302 |
| 三世代 | 4,956 | 384 | 5,340 | 2,390 | 118 | 2,508 | 638 | 160 | 798 | 8,646 |
| その他 | 654 | 234 | 887 | 413 | 75 | 488 | 278 | 174 | 452 | 1,826 |
| 合計 | 49,439 | 7,947 | 57,386 | 15,250 | 1,623 | 16,874 | 18,340 | 7,401 | 25,741 | 100,000 |

表3

| 統計量 | パンの支出金額（円） | 1を加えた金額の常用対数 |
|---|---|---|
| 最大値 | 51,776 | 4.714 |
| 第3四分位数 | 3,039 | 3.483 |
| 中央値 | 1,689 | 3.228 |
| 第1四分位数 | 858 | 2.934 |
| 最小値 | 0 | 0.000 |
| 平均値 | 2,284 | 3.137 |
| 件数 | 100,000 | 100,000 |

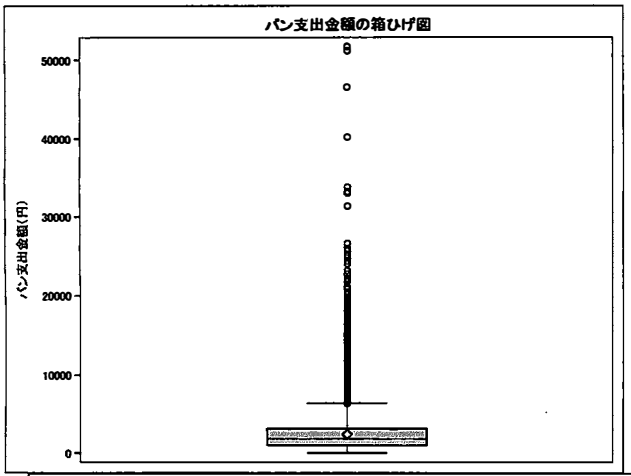
表5

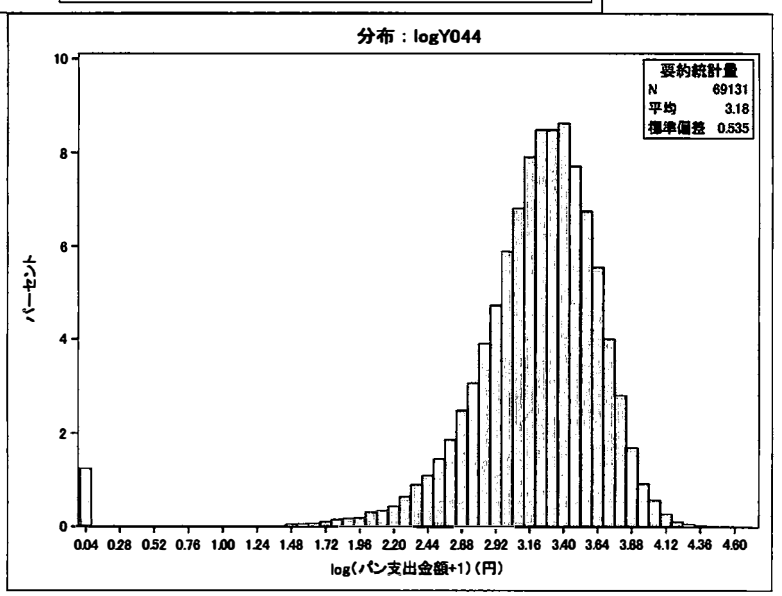
| 世帯分類 | ジニ係数 |
|---|---|
| 単身世帯 | 0.19 |
| 夫婦のみ | 0.15 |
| 二世代 | 0.15 |
| 二世代(ひとり親) | 0.18 |
| 三世代 | 0.15 |
| その他 | 0.17 |
| 全体 | 0.33 |

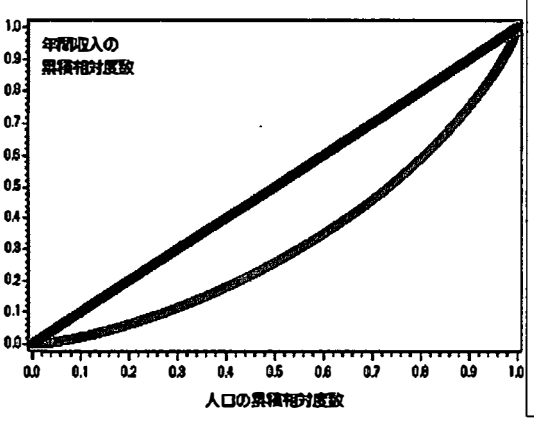
表4

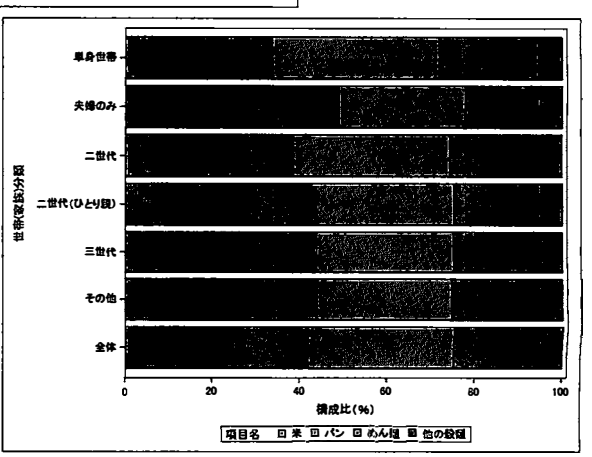
| 項目名 | 符号内容 | 件数 | 構成比(%) | | | | 全体(%) |
|---|---|---|---|---|---|---|---|
| | | | 米 | パン | めん類 | 他の穀類 | |
| 世帯区分 | 勤労者世帯 | 57,386 | 36.0 | 37.6 | 22.1 | 4.3 | 100.0 |
| | 勤労者以外の世帯 | 16,874 | 44.9 | 31.0 | 19.4 | 4.7 | 100.0 |
| | 無職世帯 | 25,741 | 52.1 | 26.7 | 15.8 | 5.4 | 100.0 |
| 世帯分類 | 単身世帯 | 20,758 | 33.8 | 37.7 | 23.0 | 5.5 | 100.0 |
| | 夫婦のみ | 24,287 | 49.5 | 28.9 | 16.8 | 4.8 | 100.0 |
| | 二世代 | 37,182 | 39.0 | 35.9 | 20.8 | 4.3 | 100.0 |
| | 二世代（ひとり親） | 7,302 | 44.9 | 31.6 | 19.0 | 4.5 | 100.0 |
| | 三世代 | 8,646 | 46.0 | 30.5 | 19.7 | 3.8 | 100.0 |
| | その他 | 1,826 | 44.8 | 30.5 | 20.4 | 4.3 | 100.0 |
| | 全体 | 100,000 | 41.6 | 33.7 | 20.0 | 4.6 | 100.0 |

下記、順に3②、3③、4②、5①の結果を添付する。


パン支出金額の箱ひげ図
パン支出金額(円)
50000
40000
30000
20000
10000
0


分布：logY044
要約統計量
N 69131
平均 3.18
標準偏差 0.535
パーセント
10
8
6
4
2
0
0.04 0.28 0.52 0.76 1.00 1.24 1.48 1.72 1.96 2.20 2.44 2.68 2.92 3.16 3.40 3.64 3.88 4.12 4.36 4.60
log(パン支出金額+1)(円)


年間収入の
累積相対度数
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
人口の累積相対度数


世帯(家族)分類
単身世帯
夫婦のみ
二世代
二世代(ひとり親)
三世代
その他
全体
0 20 40 60 80 100
構成比(%)
項目名 米 パン めん類 他の穀類

# 35 歳未満の独身男女の消費傾向について

参加カテゴリー「ニューカマー」
チーム名:タクミインフォメーションテクノロジー チーム A
チーム代表:芦澤宏治
チームメンバー:芦澤宏治、塚本洋平、米山隆樹
(株式会社タクミインフォメーションテクノロジー 医薬品開発事業 2 部 SAS コンサルティンググループ)

# About a consumption trend of the single man and woman younger than 35 years old

Koji Ashizawa , Yohei Tsukamoto , Ryuki Yoneyama
Corporate : Takumi Information Technology

自由課題
キーワード:疑似ミクロデータ、単身世帯、消費支出、平均値、Wilcoxon の順位和検定

## 要旨

本論文では、若年未婚者がどんなものを消費しているのかを探索的に研究することを目的とし、全国消費実態調査 2004 年新疑似ミクロデータを用いて男女間における消費内訳の比較及び費目の相関を調べた。また、自由課題と規定課題ともに SAS9.4 を用いて行った。

## はじめに

昨今、若者の消費離れが話題となることが多い。実際、1970 年代の大量生産・大量消費の時代と比較すると、消費は減少している[1]。一方で、年収階層によらず男性より女性の消費性向が高く、女性は男性より消費意欲が高い傾向がある[2]。そこで、本研究では、若年未婚者の具体的な消費費目の内訳を探索的に研究した。既婚者ではパートナーの収入も関係するため、今回は未婚者のみを対象として、横断的研究を行った。

## 謝辞

本論文の作成にあたり、データを提供してくださった SAS ユーザー会世話人の有志の皆さん、ありがとうございました。69131 件、219 変数もあるデータの分析をする機会ができ、心から感謝しています。

## §1.抽出条件と妥当性

### 抽出条件

「35 歳未満の独身男女」は、次をすべて満たすものと定義した。

条件 A:世帯区分は、勤労者世帯(X02 = 1)。　条件 B:世帯人員は、1 人(X03 = 1)。
条件 C:年齢は、35 歳未満(X09 = 5,6,7)。　条件 D:家族分類は、単身世帯(X11 = 1)。
条件 E:未就学児は、いない(X12 = 1)。　条件 F:学校に通う世帯員は、いない(X13 = 1)。
条件 G:65 歳以上の世帯人数は、0 人(X14 = 0)。
また、自由課題では集計用乗率は用いない。

妥当性

上で定義して得られた男女比と、国勢調査[3]による 35 歳未満の独身男女比は大きな差はない。

**図表 1.1**

| | 抽出世帯 | | 国勢調査 | |
|---|---|---|---|---|
| 性別 | 世帯数 | 割合 (%) | 世帯数 | 割合 (%) |
| 男 | 732 | 61.00 | 6717 | 54.40 |
| 女 | 468 | 39.00 | 5630 | 45.60 |

## §2.「35 歳未満独身男性」と「35 歳未満独身女性」の比較

まず、男女の消費支出について調べた。Shapiro-Wilk 検定の結果より正規性を仮定できなかったため(図表 2.1、2.2)、今回は Wilcoxon の順位和検定を用いた。その結果、男女間に有意差は検出されなかった(図表 2.4)。消費支出の内訳である 10 大費目に対しても Wilcoxon の順位和検定を用いると、差があることがわかった(図表 2.8)。男女別に消費支出及び 10 大費目の平均値を示し、各 10 大費目の消費支出に対する割合を求めた。教育は、女性 1 人のみのため有意差は出なかったが、他の費目は食料と同様に有意差があった。10 大費目における Shapiro-Wilk 検定及び Wilcoxon の順位和検定の結果は食料のみを載せた。

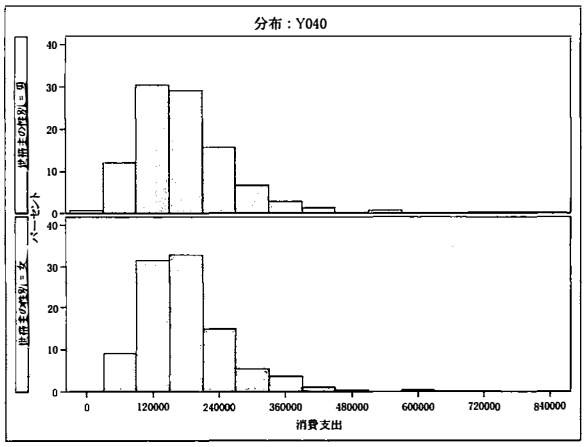
**図表 2.1(男性の消費支出の正規性検定)**

| 正規性の検定 | | | | |
|---|---|---|---|---|
| 検定 | 統計量 | | p 値 | |
| Shapiro-Wilk | W | 0.850239 | Pr < W | <0.0001 |

**図表 2.2(女性の消費支出の正規性検定)**

| 正規性の検定 | | | | |
|---|---|---|---|---|
| 検定 | 統計量 | | p 値 | |
| Shapiro-Wilk | W | 0.879892 | Pr < W | <0.0001 |

**図表 2.3(男女ごとの消費支出のヒストグラム)**




**図表 2.4(消費支出における男女間の差の比較)**

| Wilcoxon の順位和検定 (2 標本) | |
|---|---|
| 両側 Pr > \|Z\| | 0.3392 |

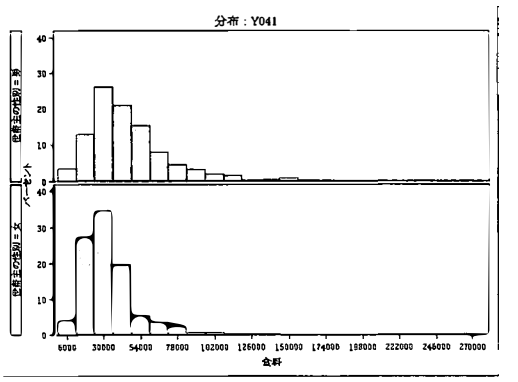
図表 2.5(男性の食料の正規性検定)

| 正規性の検定 | | | | |
|---|---|---|---|---|
| 検定 | 統計量 | | p 値 | |
| Shapiro-Wilk | W | 0.890857 | Pr < W | <0.0001 |

図表 2.6(女性の食料の正規性検定)

| 正規性の検定 | | | | |
|---|---|---|---|---|
| 検定 | 統計量 | | p 値 | |
| Shapiro-Wilk | W | 0.74767 | Pr < W | <0.0001 |

図表 2.7(男女ごとの食料のヒストグラム)




図表 2.8(食料における男女間の差の比較)

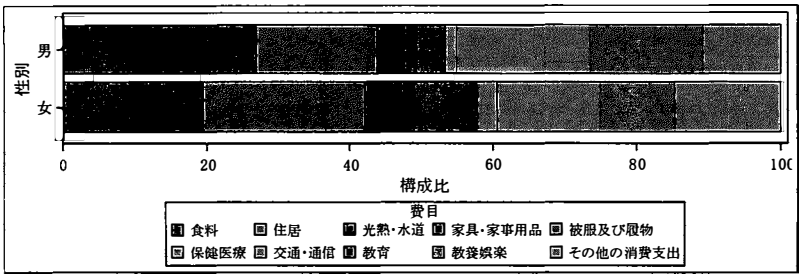
| Wilcoxon の順位和検定 (2 標本) | |
|---|---|
| 両側 Pr > \|Z\| | <.0001 |

消費支出の費目別の内訳をみると、男性では「食料」が最も多く、次に「交通・通信」、「教養娯楽」、となっていた。一方、女性では「住居」が最も多く、次に「食料」、「その他の消費支出(交際費など)」となっていた。生活必需性の高い食料と住居で、男女とも約半分を占めていた。生活必需性の高い、衣食住の割合が高いとされているが、被服及び履物は高くはなかった。購入頻度と単価の安さが起因となっているのではないだろうか。

図表 2.9(男女間における消費支出(月額)の差の比較)

| | 男性 | | 女性 | |
|---|---|---|---|---|
| 費目 | 平均 (円) | 割合 (%) | 平均 (円) | 割合 (%) |
| 食料 | 45,621 | 25.62 | 33,858 | 18.75 |
| 住居 | 29,476 | 16.55 | 39,318 | 21.78 |
| 光熱・水道 | 6,697 | 3.76 | 8,144 | 4.51 |
| 家具・家事用品 | 2,430 | 1.36 | 3,427 | 1.90 |
| 被服及び履物 | 7,477 | 4.20 | 17,458 | 9.67 |
| 保健医療 | 2,465 | 1.38 | 4,566 | 2.53 |
| 交通・通信 | 33,861 | 19.02 | 25,655 | 14.21 |
| 教育 | 0 | 0.00 | 49 | 0.03 |
| 教養娯楽 | 29,735 | 16.70 | 19,147 | 10.61 |
| その他の消費支出 | 20,301 | 11.40 | 28,920 | 16.02 |
| 消費支出(合計) | 178,063 | 100.00 | 180,541 | 100.00 |

**図表 2.10(費目の構成比)**

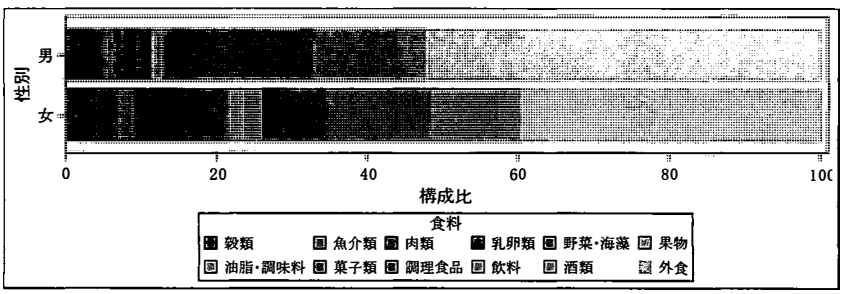



§3.「食料」「交通・通信」「教養娯楽」「住居」「その他の消費支出」について

男女それぞれの費目で、上位 3 位の費目の内訳を調べた。

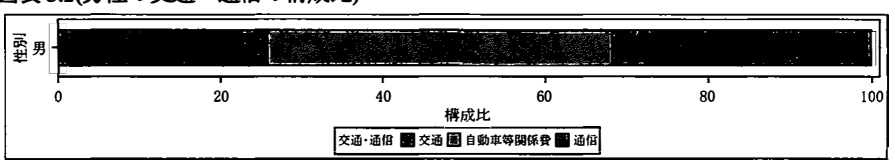
男女の食料の項目を調べると、外食及び調理食品は男性の方が高い傾向がみられたが、肉類及び乳卵類など素材に関する項目は、女性の方が高い傾向があった。これは[4][5]とほぼ同じ結果である。外食好きなら、料理をする機会が減るため素材を買うのも減ると推測された。

**図表 3.1(食料の構成比)**



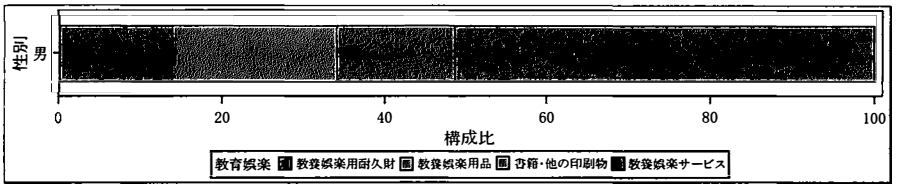

男性の交通・通信の項目を調べると、自動車等関係費と通信が割合を占めていた。通信が高くないのは、スマートフォンがまだ普及していない時代のデータであるためではないだろうか。最近のデータとも比較してみたい。

**図表 3.2(男性の交通・通信の構成比)**



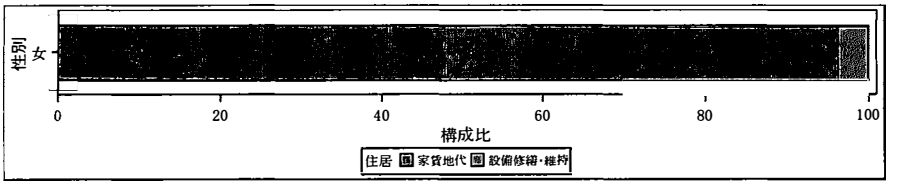
男性の教養娯楽の項目を調べると、教養娯楽サービスが構成比の約半分を占めていた。

**図表 3.3(男性の教育娯楽の構成比)**



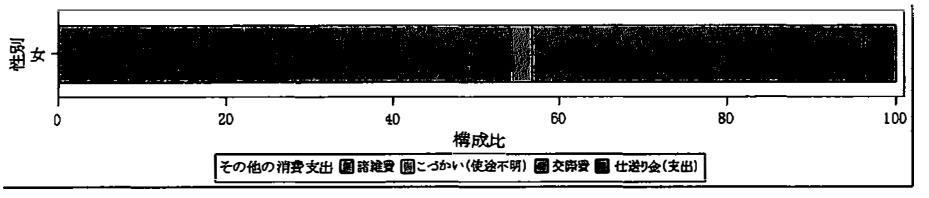

女性の住居の項目を調べると、家賃地代が構成比の約 96%を占めていた。独身女性の持ち家率が高くないことが要因だと推測された。

**図表 3.4(女性の住居の構成比)**




女性のその他の消費支出の項目を調べると、諸雑費(理美容サービスなど)と交際費への高い傾向がみられた。女性は消費意欲が旺盛で、特に美容に興味関心が高いということは世間でよく言われることだが、今回のデータにもあらわれた。

**図表 3.5(女性のその他の消費支出の構成比)**



## §4.上位 3 位と消費支出の相関

男女それぞれの費目の関連性を調べた。男女別に主要な費目間の関連を確認するために消費内訳上位 3 位までとした。正規性を仮定できなかったため、Spearman の順位相関係数を計算した。

**図表 4.1(男性)**

| Spearman の相関係数, N = 732<br>H0: Rho=0 に対する Prob > \|r\| | | | | |
|---|---|---|---|---|
| | Y040 | Y041 | Y126 | Y141 |
| Y041食料 | 0.6618 | 1.0000 | 0.3809 | 0.3514 |
| Y126交通・通信 | 0.6162 | 0.3809 | 1.0000 | 0.2630 |
| Y141教養娯楽 | 0.5873 | 0.3514 | 0.2630 | 1.0000 |

**図表 4.2(女性)**

| Spearman の相関係数, N = 468<br>H0: Rho=0 に対する Prob > \|r\| | | | | |
|---|---|---|---|---|
| | Y040 | Y041 | Y083 | Y151 |
| Y041食料 | 0.64504 | 1.00000 | 0.09727 | 0.42391 |
| Y083住居 | 0.43536 | 0.09727 | 1.00000 | 0.10205 |
| Y151その他の消費支出 | 0.67186 | 0.42391 | 0.10205 | 1.00000 |

男性は上位 3 位に関してそれぞれ正の相関があった。一方、女性は消費支出に対しては、それぞれ正の相関があったが、住居と食料、住居とその他の消費支出にはあまり相関がなかった。表は載せていないが残りの 7 つの費目に対しても、住居とはあまり相関がなかった。住居は他の消費項目とはあまり関係はない、または関係が弱いと考えられた。

## §5.まとめと展望

本研究では、35 歳未満の独身男女の消費内訳に焦点をあて、男女間の差を比較した。その結果、35 歳未満の独身男女の消費支出には差はあまりないが、10 大費目それぞれは差があることがわかった。また、男女それぞれにおける上位 3 位の費目に着目すると、男性はどの 2 つの組を選んでもすべて正の相関をもっていた。一方、女性の場合、住居は食料とその他の消費支出とはあまり相関がないことがわかった(図表 4.2)。男女とも生活必需性の高い費目への支出はもちろん多いが、生活必需性の低い費目(男性は教養娯楽、女性はその他の消費支出)にも多くの支出があった。

また、今回の独身男女(2 群)の比較の際には正規性と等分散性を仮定できなかった。本論文には載せていないが、対数変換しても正規性や等分散性を仮定できない費目がほとんどだった。正規性、等分散性を仮定しない検定なら Brunner-Munzel 検定が適切ではないだろうか[6][7]。Wilcoxon の順位和検定は等分散性を仮定しているため、本来は適切な方法ではない。ちなみにエクセル統計や R には Brunner-Munzel 検定が搭載されている(https://bellcurve.jp/ex/function/brunner.html 参照)。SAS9.4 では見つけられなかったため、本論文は Wilcoxon の順位和検定を用いた。ただし、今回はサンプルサイズが大きいため、Welch 検定でも Wilcoxon の順位和検定と同じ結果を得られた。ちなみに[7]を参考に Brunner-Munzel 統計量を計算すると、消費支出の p 値は約 0.1673 となり有意差は出ないという同じ結果が得られた。

今回は 2004 年のデータで比較をしたが、最新のデータと同じとは限らない。最近のデータをもとに、若者の消費傾向及び傾向の変化について比較したい。

§6.参考文献

[1]児美川孝一郎「若者の消費行動に見る日本社会の未来形」(2013)
< http://www.yhmf.jp/pdf/activity/adstudies/vol_43_01_02.pdf>
[2]「若年層の消費実態（１）ー収入が増えても、消費は抑える今の若者たち」
<http://www.nli-research.co.jp/report/detail/id=53061&pno=2?site=nli>
[3]総務省統計局「抽出速報集計結果」(2015)
<http://www.stat.go.jp/data/kokusei/2015/kekka.htm>
[4]総務省統計局「単身世帯の家計収支及び貯蓄・負債に関する結果」(2014)
< http://www.stat.go.jp/data/zensho/2014/pdf/gaiyo2.pdf>
[5]総務省統計局「男女・年齢階級別の家計」(2002)
<http://www.stat.go.jp/data/soutan/2002np/5.htm>
[6]名取真人「マン・ホイットニーの U 検定と不等分散時における代表値の検定法」霊長類研究(2014)173-185
<https://www.jstage.jst.go.jp/article/psj/30/1/30_30.006/_pdf >
[7]村上秀俊「ノンパラメトリック法」朝倉書店

# 付録 規定課題用SASプログラム

```
/*----------ライブラリの設定----------*/
libname zensho "¥¥tit-box¥koji_ashizawa¥2017ミクロデータ¥rawdata";
proc format;
  value X01F   1 = "3 大都市圏" 0 = "その他";
  value X02F   1 = "勤労者世帯" 2 = "勤労者以外の世帯" 3 = "無職世帯";
  value X03F   1 = "1人" 2 = "2人" 3 = "3人" 4 = "4人" 5 = "5人以上";
  value X04F   0 = "0人" 1 = "1人" 2 = "2人" 3 = "3人以上";
  value X05F   1 = "木造(防火木造含む)" 2 = "木造(防火木造含む)以外";
  value X06F   1 = "一戸建" 2 = "一戸建以外";
  value X07F   1 = "持ち家" 2 = "持ち家以外";
  value X08F   1 = "男" 2 = "女";
  value X09F   5 = "24歳以下" 6 = "25～29歳" 7 = "30～34歳"
               8 = "35～39歳" 9 = "40～44歳" 10 = "45～49歳"
               11 = "50～54歳" 12 = "55～59歳" 13 = "60～64歳"
               14 = "65～69歳" 15 = "70～74歳" 16 = "75歳以上";
  value X10F   1 = "民営・自営1～4人" 2 = "民営・自営5～29人"
               3 = "民営・自営30～499人" 4 = "民営・自営500人以上"
               5 = "官公" 6 = "無職";
  value X11F   1 = "単身世帯" 2 = "夫婦のみ" 3 = "二世代"
               4 = "二世代(ひとり親)" 5 = "三世代" 6 = "その他";
  value X12F   1 = "無" 2 = "有";
  value X13F   1 = "無" 2 = "有(専修学校及び各種学校を含む)";
  value X14F   0 = "0人" 1 = "1人" 2 = "2人以上";
run;
options fmtsearch = (zensho);
data work.zensho2004gijimicro;
  set zensho.Zensho2004gijimicro; * 規定課題用データ(HPの) *;
  format X01 X01F. X02 X02F. X03 X03F. X04 X04F. X05 X05F. X06 X06F.
X07 X07F. X08 X08F. X09 X09F. X10 X10F. X11 X11F. X12 X12F. X13 X13F.
X14 X14F.;
  label   Year = "調査年" No = "レコード連番号"
    X01 = "大都市圏の別" X02 = "世帯区分"
    X03 = "世帯人員" X04 = "就業人員"
    X05 = "住居の構造" X06 = "住居の建て方"
    X07 = "住居の所有関係" X08 = "世帯主の性別"
    X09 = "世帯主の年齢" X10 = "企業区分・従業者規模"
    X11 = "家族分類" X12 = "未就学児の有無"
    X13 = "学校に通う世帯員の有無" X14 = "65歳以上の世帯員数"
    weight = "集計用乗率" Y001 = "年間収入(単位千円)"
    Y002 = "収入総額" Y003 = "実収入" Y004 = "経常収入"
    Y005 = "勤め先収入" Y006 = "世帯主の勤め先収入"
    Y007 = "世帯主の配偶者の勤め先収入"
    Y008 = "他の世帯員の勤め先収入" Y009 = "事業・内職収入"
    Y010 = "農林漁業収入" Y011 = "家賃収入"
    Y012 = "他の事業収入" Y013 = "内職収入"
    Y014 = "本業以外の勤め先・事業・内職収入" Y015 = "他の経常収入"
    Y016 = "財産収入" Y017 = "社会保障給付"
    Y018 = "公的年金給付" Y019 = "他の社会保障給付"
    Y020 = "仕送り金(収入)" Y021 = "特別収入" Y022 = "受贈金"
    Y023 = "その他の特別収入" Y024 = "実収入以外の収入"
    Y025 = "預貯金引出" Y026 = "保険取金"
    Y027 = "個人・企業年金保険取金" Y028 = "他の保険取金"
    Y029 = "有価証券売却" Y030 = "株式売却" Y031 = "他の有価証券売却"
    Y032 = "土地家屋借入金" Y033 = "他の借入金"
    Y034 = "分割払・一括払購入借入金" Y035 = "財産売却"
    Y036 = "その他の実収入以外の収入" Y037 = "繰入金"
    Y038 = "支出総額" Y039 = "実支出" Y040 = "消費支出" Y041 = "食料"
    Y042 = "穀類" Y043 = "米" Y044 = "パン" Y045 = "めん類"
    Y046 = "他の穀類" Y047 = "魚介類" Y048 = "生鮮魚介"
    Y049 = "塩干魚介" Y050 = "魚肉練製品" Y051 = "他の魚介加工品"
    Y052 = "肉類" Y053 = "生鮮肉" Y054 = "加工肉" Y055 = "乳卵類"
    Y056 = "牛乳" Y057 = "乳製品" Y058 = "卵" Y059 = "野菜・海藻"
    Y060 = "生鮮野菜" Y061 = "乾物・海藻"
    Y062 = "大豆加工品" Y063 = "他の野菜・海藻加工品"
    Y064 = "果物" Y065 = "生鮮果物"
    Y066 = "果物加工品" Y067 = "油脂・調味料"
    Y068 = "油脂" Y069 = "調味料" Y070 = "菓子類" Y071 = "調理食品"
    Y072 = "主食的調理食品" Y073 = "他の調理食品"
    Y074 = "飲料" Y075 = "茶類" Y076 = "コーヒー・ココア"
    Y077 = "他の飲料" Y078 = "酒類" Y079 = "外食" Y080 = "一般外食"
    Y081 = "学校給食" Y082 = "賄い費" Y083 = "住居" Y084 = "家賃地代"
    Y085 = "設備修繕・維持" Y086 = "設備材料"
    Y087 = "工事その他のサービス" Y088 = "光熱・水道"
    Y089 = "電気代" Y090 = "ガス代" Y091 = "他の光熱"
    Y092 = "上下水道料" Y093 = "家具・家事用品" Y094 = "家庭用耐久財"
    Y095 = "家事用耐久財" Y096 = "冷暖房用器具" Y097 = "一般家具"
    Y098 = "室内装備・装飾品" Y099 = "寝具類" Y100 = "家事雑貨"
    Y101 = "家事用消耗品" Y102 = "家事サービス" Y103 = "被服及び履物"
    Y104 = "和服" Y105 = "洋服" Y106 = "男子用洋服"
    Y107 = "婦人用洋服" Y108 = "子供用洋服"
    Y109 = "シャツ・セーター類" Y110 = "男子用シャツ・セーター類"
    Y111 = "婦人用シャツ・セーター類"
    Y112 = "子供用シャツ・セーター類"
    Y113 = "下着類" Y114 = "男子用下着類"
    Y115 = "婦人用下着類" Y116 = "子供用下着類"
    Y117 = "生地・糸類" Y118 = "他の被服"
    Y119 = "履物類" Y120 = "被服関連サービス" Y121 = "保健医療"
    Y122 = "医薬品" Y123 = "健康保持用摂取品"
    Y124 = "保険医療用品・器具" Y125 = "保健医療サービス"
    Y126 = "交通・通信" Y127 = "交通" Y128 = "自動車等関係費"
    Y129 = "自動車等購入" Y130 = "自転車購入"
    Y131 = "自動車等維持" Y132 = "(特掲)ガソリン"
    Y133 = "(特掲)自動車整備費" Y134 = "(特掲)自動車保険料"
    Y135 = "通信" Y136 = "(特掲)移動電話通信料"
    Y137 = "教育" Y138 = "授業料等"
    Y139 = "教科書・学習参考教材" Y140 = "補習教育"
    Y141 = "教養娯楽" Y142 = "教養娯楽用耐久財"
    Y143 = "教養娯楽用品" Y144 = "書籍・他の印刷物"
    Y145 = "教養娯楽サービス" Y146 = "宿泊料"
    Y147 = "パック旅行費" Y148 = "月謝類"
    Y149 = "他の教養娯楽サービス"
    Y150 = "(特掲)インターネット接続料"
    Y151 = "その他の消費支出" Y152 = "諸雑費"
    Y153 = "理美容サービス" Y154 = "理美容用品"
    Y155 = "身の回り用品" Y156 = "たばこ"
    Y157 = "その他の諸雑費" Y158 = "こづかい(使途不明)"
    Y159 = "交際費" Y160 = "交際費(食料)"
    Y161 = "交際費(家具・家事用品)" Y162 = "交際費(被服及び履物)"
    Y163 = "交際費(教養娯楽)" Y164 = "他の物品サービス"
    Y165 = "贈与金" Y166 = "他の交際費"
    Y167 = "仕送り金(支出)" Y168 = "(再掲)教育関係費"
    Y169 = "(再掲)教養娯楽関係費" Y170 = "耐久財"
    Y171 = "読書" Y172 = "聴視・観覧" Y173 = "旅行" Y174 = "スポーツ"
    Y175 = "月謝" Y176 = "会費・つきあい費"
    Y177 = "他の教養娯楽" Y178 = "(再掲)情報通信関係費"
    Y179 = "非消費支出" Y180 = "直接税"
    Y181 = "勤労所得税" Y182 = "個人住民税"
    Y183 = "他の税" Y184 = "社会保険料"
    Y185 = "公的年金保険料" Y186 = "健康保険料"
    Y187 = "介護保険料" Y188 = "他の社会保険料"
    Y189 = "他の非消費支出" Y190 = "実支出以外の支出"
    Y191 = "預貯金" Y192 = "保険掛金"
    Y193 = "個人・企業年金保険掛金" Y194 = "他の保険掛金"
    Y195 = "有価証券購入" Y196 = "株式購入"
    Y197 = "他の有価証券購入" Y198 = "土地家屋借金返済"
    Y199 = "他の借金返済" Y200 = "分割払・一括払購入借入金返済"
    Y201 = "財産購入" Y202 = "その他の実支出以外の支出"
    Y203 = "繰越金";
run;
*10万人比集計用に合計WEIGHT算出;
proc means data = work.zensho2004gijimicro noprint;
  var Weight;
  output out = Weight(keep = Weight2) sum = Weight2;
run;

/*----------規定課題1----------*/
data Zensho2004gijimicro_1;
  set work.Zensho2004gijimicro;
  retain Multi;
  if _N_ = 1 then do;
    set Weight;
    Multi = 1e5 / Weight2;
  end;
  CountDummy = 1;
  TableType = 1;
  RedefinedWeight = 1; output;
  TableType = 2;
  RedefinedWeight = Weight * Multi; output;
  keep TableType X02 X09 CountDummy RedefinedWeight;
  format _all_;
run;
* 集計対象となるカテゴリの組み合わせを作成;
```

```
data ClassData1;
  do X09 = 5 to 16;
    do TableType = 1 to 2;
      do X02 = 1 to 3;
        output;
      end;
    end;
  end;
run ;
* 集計 ;
proc means data = Zensho2004gijimicro_1 noprint classdata = ClassData1;
  class TableType X09 X02;
  types   TableType * X09 * X02
    TableType * X02
    TableType * X09
    TableType ;
  var CountDummy;
  weight RedefinedWeight;
  output out = Means1 sum = SumCnt ;
run;
data Means2 ; set Means1;
  if missing( X09 ) then
    X09 = 17;
  if missing( X02 ) then
    X02 = 4;
  if missing( SumCnt ) then
    SumCnt = 0;
  length SumCntC $85;
  SumCntC = put( round( SumCnt, 1 ), comma12. -L );
    format _all_;
proc sort;
  by X09 TableType X02;
run;
proc format ;
  value X09Fn 5 = "24歳以下"   6 = "25～29歳"
        7 = "30～34歳" 8 = "30～35歳" 9 = "40～44歳"
        10 = "45～49歳" 11 = "50～54歳" 12 = "55～59歳"
        13 = "60～64歳" 14 = "65～69歳" 15 = "70～74歳"
        16 = "75歳以上" 17 = "合計" ;
run ;
proc transpose data = Means2 out = Means3 prefix = C;
  by X09 ; id TableType X02;
  var SumCntC;
run;
data Means4; set Means3;
 c25 = "" ;
run ;
proc report data = Means4 nowd ;
  column
    (X09 ("世帯区分" C11 C12 C13 C14)) ("09"x C25)
    (X09 ("世帯区分" C21 C22 C23 C24)) ;
  define X09 / "世帯主の年齢" order order=internal left;
  define C11 / "勤労者" right;
  define C12 / "勤労者以外" right;
  define C13 / "無職" right;
  define C14 / "合計" right;
  define C25 / "" ;
  define C21 / "勤労者" right;
  define C22 / "勤労者以外" right;
  define C23 / "無職" right;
  defineC24 / "合計" right;
  format X09 X09Fn. ;
run ;


/*-----------------規定課題2-----------------*/
data zensho2004gijimicro_2 ;
  set zensho2004gijimicro (keep=X02 X08 X09 X11 Y043 - Y046 weight) ;
  retain Multi ;
  if _N_ = 1 then do;
    set Weight;
    Multi = 100000 / Weight2;
  end ;
  CountDummy = 1 ;
  RedefWeight = Weight * Multi ;
  output ;
  format X02 X02F. X08 X08F. X11 X11F. ;
run ;
proc tabulate data = zensho2004gijimicro_2 ;
  class X02 X08 X11 ; var CountDummy ;
  table   (X11="" all="計") , (X02="" * (X08="" all="計") all="合計")
          * CountDummy="" * sum="" * f=comma8. /box="世帯分類" ;
  weight RedefWeight ;
  format
    X02 X02F. X08 X08F. X11 X11F. ;
run;


/*-----------------規定課題3-----------------*/
data Zensho2004gijimicro_3 ;
  set Zensho2004gijimicro ;
  retain Multi ;
  if _N_ = 1 then do;
    set Weight ;
    Multi = 1e5 / weight2 ;
  end ;
  Countdummy = 1 ;
  Countdummy2 = 1 ;
  RedefWeight = Weight * Multi ;
  Y044_1 = Y044 + 1 ;
  log2    = log10(Y044_1) ;
  run ;
***要約統計量データセット作成(①)***;
ods output Summary=mean ;
proc means data = Zensho2004gijimicro_3
            MAX Q3 MEDIAN Q1 MIN MEAN   STACKODS ;
  var y044 log2 ; weight RedefWeight ;
run ;
ods output close ;
data mean ; set mean ;
  if variable="Y044" then counttype=1 ;
  else if variable="log2" then counttype=2 ;
run ;
***"件数"をデータセットに出力(②)***;
ods output Summary=Total ;
proc means data = Zensho2004gijimicro_3 sum STACKODS ;
  var countdummy Countdummy2 ;
  weight RedefWeight ;
run ;
ods output close ;
data total ; set total ;
  rename variable = v2 ;
  if variable = "Countdummy" then counttype=1 ;
  else if variable = "Countdummy2" then counttype=2 ;
run ;
***①と②を結合(③)***;
data mean_total;
  merge mean total ; by counttype ;
run ;
***③の中身を表作成用に加工***;
data mean_total2; set mean_total;
  length MaxC Q3C MedianC Q1C MinC MeanC sumc IdLabel $100;
  if CountType=1 then do ;
    IdLabel = "パンの支出金額(円)";
    MaxC = put(round(Max, 1), comma12. -L);
    Q3C = put(round(Q3, 1e-1), comma12. -L);
    MedianC = put(round(Median, 1e-1), comma12. -L);
    Q1C = put(round(Q1, 1e-1), comma12. -L);
    MinC = put(round(Min, 1), comma12. -L);
    MeanC = put(round(Mean, 1e-1), comma12. -L);
    SUMC = put(round(Sum, 1), comma12. -L);
  end;
  else if counttype=2 then do ;
    IdLabel = "1を加えた金額の常用対数";
    MaxC = put(round(Max, 1e-3), comma12.3 -L);
    MinC = put(round(Min, 1e-3), comma12.3 -L);
    Q3C = put(round(Q3, 1e-4), comma12.3 -L);
    Q1C = put(round(Q1, 1e-4), comma12.3 -L);
    MedianC = put(round(Median, 1e-4), comma12.3 -L);
    MeanC = put(round( Mean, 1e-4), comma12.3 -L);
    SUMC = put(round( Sum, 1), comma12. -L);
  end ;
  label   MaxC="最大値" Q3C="第3四分位数"
    MedianC="中央値" Q1C="第1四分位数"
    MinC="最小値" MeanC="平均値"
    sumc="件数" ;
run ;
proc transpose data = mean_total2 out = mean_total3 prefix = C ;
  id counttype ;
  idlabel IdLabel ;
  var MaxC Q3C MedianC Q1C MinC MeanC sumC ;
run ;
***出力***;
proc report data = mean_total3(drop=_NA ME_) nowd ;
  column _LABEL_ c1 c2 ;
  define _LABEL_ / width=15 center "統計量" ;
  define c1 / width=15 right   "パンの支出金額（円）" ;
  define c2 / width=15 right   "1を加えた金額の常用対数";
run ;
```

```
ods graphics / height=5cm width=15cm;
proc sgplot data = Zensho2004gijimicro_3 ;
   vbox   Y044 ;
run ;
proc sgplot data = Zensho2004gijimicro_3 ;
   histogram log2 ;
   label log2 = "1を加えた金額の常用対数";
run ;
ods graphics / reset ;


/*----------------規定課題4----------------*/
proc format;
   value Item
      1 = "世帯区分" 2 = "世帯分類" 3 = "全体" ;
   value XF
      1 = "勤労者世帯" 2 = "勤労者以外の世帯"
      3 = "無職世帯" 101 = "単身世帯"
      102 = "夫婦のみ" 103 = "二世代"
      104 = "二世代(ひとり親)" 105 = "三世代"
      106 = "その他" 200 = " " ;
   value Grain
      0 = "件数" 1 = "米" 2 = "パン"
      3 = "めん類" 4 = "他の穀類" 5 = "全体(%)" ;
run ;
data Zensho2004gijimicro_4_1 ; set Zensho2004gijimicro;
   retain Multi;
   if _N_ = 1 then do;
      set Weight;
      Multi = 1e5 / Weight2 ;
   end;
   CountDummy = 1;
   RedefWeight = Weight * Multi ;
   array Variables(*) Y043 Y044 Y045 Y046 Y042 ;
   do I=1 to dim(Variables) ;
      Category = I;
      if Y042 ^= 0 then
         PR = 100 * Variables(I) / Y042 ;
      * 世帯区分 ;
      Item = 1 ; X = X02 ;
      output;
      * 世帯分類 ;
      Item = 2 ; X = X11 + 100 ;
      output;
      * 全体 ;
      Item = 3 ; X = 200 ;
      output ;
   end ;
   keep Item X Category PR CountDummy RedefWeight ;
   format _all_ ;
run ;
* 構成比 ;
proc means data = Zensho2004gijimicro_4_1 noprint nway;
   class Item X Category; var PR ;
   weight RedefWeight;
   output out = Zensho2004gijimicro_4_2 mean=MeanPR ;
run;
* 件数 ;
proc means data = Zensho2004gijimicro_4_1 noprint nway;
   where Category=5; class Item X ;
   var CountDummy ;
   weight RedefWeight;
   output out = Zensho2004gijimicro_4_3 sum = SumCnt ;
run;
data Zensho2004gijimicro_4_4 ;
   set Zensho2004gijimicro_4_2( in = In1 )
       Zensho2004gijimicro_4_3( in = In2 ) ;
   length ValueC XC $100;
   if In1 then do;
      ValueC = put( round( MeanPR, 1e-1 ), 12.1 -L );
   end;
   else if In2 then do;
      Category = 0;
      ValueC = put( round( SumCnt, 1 ), comma12. -L );
   end;
   IdLabel = put( Category, Grain. );
proc sort ; by Item X Category ; run ;
proc transpose data = Zensho2004gijimicro_4_4
   out = Zensho2004gijimicro_4_5 prefix=C ;
   by Item X ;
   id Category; idlabel IdLabel;
   var ValueC;
run ;
proc report data=Zensho2004gijimicro_4_5 ;
   column Item X C0 ( "構成比(%)" C1 - C4 ) C5 ;
   define Item / "項目名" order order=internal style(column)={just=left
cellwidth=2cm};
   define X / "符号内容" order   order=internal style(column)={just=left
cellwidth=3.5cm};
   define C0 - C5 / style(column)={just=right cellwidth=1.5cm};
   format Item Item. X XF.;
run;
ods graphics / height=5cm width=15cm;
proc sgplot data = Zensho2004gijimicro_4_2;
   where Item = 2 and Category ^= 5;
   hbar X / response=MeanPR group=Category ;
   format X XF. Category Grain.;
   label
      X = "世帯分類" MeanPR = "構成比" Category = "穀類分類" ;
run ;
ods graphics / reset ;


/*----------------規定課題5----------------*/
%macro kadai5(setai=) ;
data zensho2004gijimicro_5_&setai._0 ;
 set work.zensho2004gijimicro ;
   %if &setai ^= 7 %then %do ;
      where x11=&setai ;
   %end ;
   keep Y001 weight ;
run ;
proc sort data = zensho2004gijimicro_5_&setai._0
   out = zensho2004gijimicro_5_&setai._0 ;
   by Y001 ;
run ;
%* 累積世帯, 累積年収 *;
data zensho2004gijimicro_5_&setai._1
   Total_&setai._1(rename = (CumWeight=TotalWeight
                    CumIncome=TotalIncome)) ;
   set zensho2004gijimicro_5_&setai._0 end = EOF ;
   %* 累積世帯、累積年収の算出 *;
   CumWeight + Weight ;
   CumIncome + Y001*weight ;
   output zensho2004gijimicro_5_&setai._1 ;
   %* 総世帯数、総年収 *;
   if EOF = 1 then do ;
      output Total_&setai._1 ;
   end ;
run ;
data zensho2004gijimicro_5_&setai._2 ;
   set zensho2004gijimicro_5_&setai._1 end=EOF ;
   retain Integral PreRateIncome PreRateWeight 0 ;
   if _N_ = 1 then set Total_&setai._1 ;
   RateWeight = CumWeight / TotalWeight ;
   RateIncome = CumIncome / TotalIncome ;
   %* 均等分布線 *;
   EqualPlot = RateWeight ;
   %* 台形則により積分 *;
   Integral = Integral + (PreRateIncome + RateIncome)
            * (RateWeight - PreRateWeight) * 0.5 ;
   PreRateIncome = RateIncome ;
   PreRateWeight = RateWeight ;
   if EOF then do;
      Gini = 2 * ( 0.5 - Integral );
   end;
   label   RateWeight = "累積世帯数比率"
      RateIncome = "累積収入比率"
      EqualPlot = "均等分布線" ;
run ;
%mend ;
%kadai5(setai=4);
ods graphics / height=5cm width=15cm;
proc sgplot data = zensho2004gijimicro_5_4_2 noautolegend ;
   series x = RateWeight y = RateIncome /   lineattrs = (color = blue) ;
   series x = RateWeight y = EqualPlot /   lineattrs = (color = red) ;
run ;
ods graphics / reset;
%kadai5(setai=1) ; %kadai5(setai=2) ; %kadai5(setai=3) ; %kadai5(setai=5) ; %
kadai5(setai=6) ; %kadai5(setai=7) ;
proc format ;
   value NF 1 = "単身世帯" 2 = "夫婦のみ"
        3 = "二世代" 4 = "二世代(ひとり親)"
        5 = "三世代" 6 = "その他" 7 = "全体" ;
run ;
data gini ; format N NF. gini comma8.2 ;
   set   zensho2004gijimicro_5_1_2 zensho2004gijimicro_5_2_2
zensho2004gijimicro_5_3_2
```

```
        zensho2004gijimicro_5_4_2 zensho2004gijimicro_5_5_2
zensho2004gijimicro_5_6_2 zensho2004gijimicro_5_7_2 ;
  where gini ^= . ;
  N + 1 ;
  keep gini N ;
run ;
proc print data = gini label noobs;
  label N = "世帯分類"   gini = "ジニ係数" ;
run ;
```

# 付録 規定課題図表

## 1) 年齢階層別・世帯区分別 クロス表

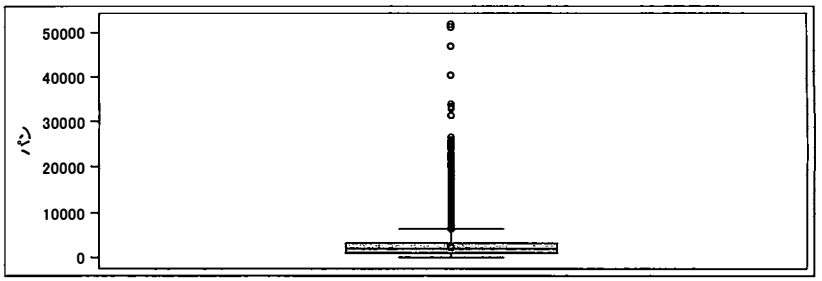
| | 世帯区分 | | | | | 世帯区分 | | | |
|---|---|---|---|---|---|---|---|---|---|
| 世帯主の年齢 | 勤労者 | 勤労者以外 | 無職 | 合計 | 世帯主の年齢 | 勤労者 | 勤労者以外 | 無職 | 合計 |
| 24歳以下 | 765 | 18 | 15 | 798 | 24歳以下 | 1,729 | 10 | 21 | 1,760 |
| 25～29歳 | 2,289 | 157 | 66 | 2,512 | 25～29歳 | 4,314 | 151 | 57 | 4,522 |
| 30～34歳 | 4,223 | 511 | 107 | 4,841 | 30～34歳 | 6,441 | 480 | 87 | 7,008 |
| 35～39歳 | 5,217 | 728 | 193 | 6,138 | 35～39歳 | 7,420 | 824 | 153 | 8,396 |
| 40～44歳 | 5,492 | 1,076 | 205 | 6,773 | 40～44歳 | 7,611 | 1,145 | 247 | 9,004 |
| 45～49歳 | 5,702 | 1,305 | 277 | 7,284 | 45～49歳 | 7,739 | 1,479 | 244 | 9,462 |
| 50～54歳 | 6,194 | 1,797 | 392 | 8,383 | 50～54歳 | 8,072 | 2,085 | 508 | 10,665 |
| 55～59歳 | 5,682 | 2,221 | 624 | 8,527 | 55～59歳 | 7,697 | 2,725 | 893 | 11,315 |
| 60～64歳 | 3,237 | 2,226 | 2,534 | 7,997 | 60～64歳 | 4,208 | 2,785 | 4,160 | 11,153 |
| 65～69歳 | 1,343 | 1,743 | 3,590 | 6,676 | 65～69歳 | 1,487 | 2,213 | 6,374 | 10,074 |
| 70～74歳 | 469 | 1,197 | 3,214 | 4,880 | 70～74歳 | 459 | 1,652 | 6,124 | 8,235 |
| 75歳以上 | 166 | 940 | 3,216 | 4,322 | 75歳以上 | 207 | 1,325 | 6,873 | 8,406 |
| 合計 | 40,779 | 13,919 | 14,433 | 69,131 | 合計 | 57,386 | 16,874 | 25,741 | 100,000 |

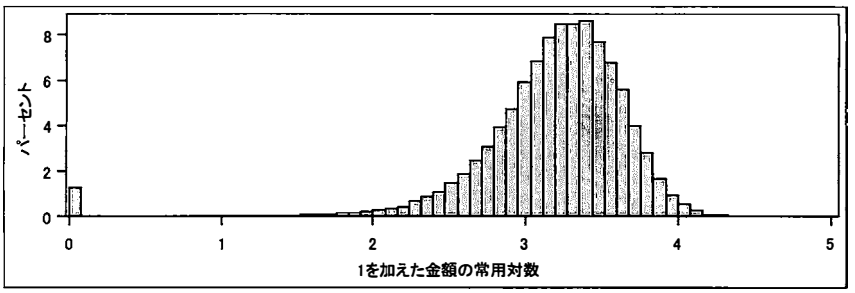
## 2) 世帯分類別・世帯区分別 世帯主の性別 3重クロス表

| 世帯分類 | 勤労者世帯 | | | 勤労者以外の世帯 | | | 無職世帯 | | | 合計 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 男 | 女 | 計 | 男 | 女 | 計 | 男 | 女 | 計 | |
| 単身世帯 | 6,503 | 4,207 | 10,710 | 1,030 | 867 | 1,897 | 2,463 | 5,688 | 8,151 | 20,758 |
| 夫婦のみの世帯 | 8,517 | 324 | 8,841 | 4,293 | 81 | 4,375 | 11,026 | 45 | 11,071 | 24,287 |
| 二世代の世帯 | 27,267 | 378 | 27,645 | 6,190 | 51 | 6,242 | 3,264 | 32 | 3,295 | 37,182 |
| 二世代（ひとり親）の世帯 | 1,543 | 2,421 | 3,963 | 934 | 430 | 1,364 | 671 | 1,303 | 1,974 | 7,302 |
| 三世代の世帯 | 4,956 | 384 | 5,340 | 2,390 | 118 | 2,508 | 638 | 160 | 798 | 8,646 |
| その他の世帯　（1～5以外の世帯） | 654 | 234 | 887 | 413 | 75 | 488 | 278 | 174 | 452 | 1,826 |
| 計 | 49,439 | 7,947 | 57,386 | 15,250 | 1,623 | 16,874 | 18,340 | 7,401 | 25,741 | 100,000 |

## 3) パン支出金額についての各種の統計グラフ

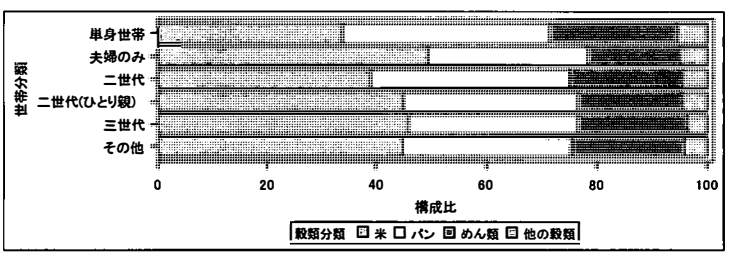
| 統計量 | パンの支出金額（円） | 1を加えた金額の常用対数 |
|---|---|---|
| 最大値 | 51,776 | 4.714 |
| 第3四分位数 | 3,039 | 3.483 |
| 中央値 | 1,689 | 3.228 |
| 第1四分位数 | 858 | 2.934 |
| 最小値 | 0 | 0.000 |
| 平均値 | 2,284 | 3.137 |
| 件数 | 100,000 | 100,000 |



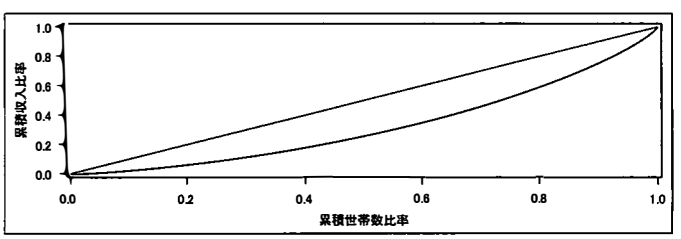




4)　世帯区分別および世帯分類別の穀類（米、パン、めん類、他の穀類）の支出金額についての構成比

| | | | 構成比(%) | | | | |
|---|---|---|---|---|---|---|---|
| 項目名 | 符号内容 | 件数 | 米 | パン | めん類 | 他の穀類 | 全体(%) |
| 世帯区分 | 勤労者世帯 | 57,386 | 36.0 | 37.6 | 22.1 | 4.3 | 100.0 |
| | 勤労者以外の世帯 | 16,874 | 44.9 | 31.0 | 19.4 | 4.7 | 100.0 |
| | 無職世帯 | 25,741 | 52.1 | 26.7 | 15.8 | 5.4 | 100.0 |
| 世帯分類 | 単身世帯 | 20,758 | 33.8 | 37.7 | 23.0 | 5.5 | 100.0 |
| | 夫婦のみ | 24,287 | 49.5 | 28.9 | 16.8 | 4.8 | 100.0 |
| | 二世代 | 37,182 | 39.0 | 35.9 | 20.8 | 4.3 | 100.0 |
| | 二世代(ひとり親) | 7,302 | 44.9 | 31.6 | 19.0 | 4.5 | 100.0 |
| | 三世代 | 8,646 | 46.0 | 30.5 | 19.7 | 3.8 | 100.0 |
| | その他 | 1,826 | 44.8 | 30.5 | 20.4 | 4.3 | 100.0 |
| 全体 | | 100,000 | 41.6 | 33.7 | 20.0 | 4.6 | 100.0 |




5)　世帯分類別のローレンツ曲線およびジニ係数




| 世帯分類 | ジニ係数 |
|---|---|
| 単身世帯 | 0.34 |
| 夫婦のみ | 0.31 |
| 二世代 | 0.26 |
| 二世代(ひとり親) | 0.35 |
| 三世代 | 0.27 |
| その他 | 0.33 |
| 全体 | 0.33 |

第 5 回ミクロデータ分析コンテスト：自由課題

# 居住地区の選択に関する探索的研究

## －人口の移動による過密化という社会問題に焦点を当てて－

株式会社 タクミインフォメーションテクノロジー(チーム B)
佐藤瑞羽　成田真実　大杉亮平　（ 参加カテゴリ：ニューカマー ）

**要旨**

過密化・過疎化は現代における大きな社会問題の一つである。本研究では、人口の移動による過密化という問題に焦点を当て、その背景要因について、「個人の諸属性」および「ライフスタイル」という 2 つの観点から探索的に横断的研究を行った。その結果、過密化には、若年層男性の都心部への人口移動や、所得水準、若年層のライフスタイルに対する多面的なニーズの変化等、様々な要因が作用している可能性が示唆された。

**問題**

昭和 30 年代後半から始まった高度経済成長過程において、地方から大都市圏への人口の流出が急激に生じ、大都市圏を中心とした過密化、およびそれに伴う農山村の過疎化が大きな問題となった(萩原,1993)。その後、過密化は一時的に緩和されたが、今後も三大都市圏(東京・大阪・名古屋)への人口過密と農山村の過疎化は並行して進む事が予想されている(総務省,2012)。過密化・過疎化における具体的な社会問題の一例として、以下のような問題が指摘されている(折田,1979;萩原,1993)。まず、農山村における過疎化については、人口の減少に伴う従来までの生活水準(医療・教育・福祉といったサービス)維持の困難さや、ひいては廃村化といった問題が指摘されている。一方、大都市圏における過密化については、交通の混雑や、住宅、道路、廃棄物問題、水・電力等のエネルギー等の逼迫、その他、オフィス需要の増大による地価高騰といった問題が指摘されている。

過密化・過疎化のプロセス、あるいは背景要因に関する研究は、既に多角的な観点からなされている。例えば、森本(1993)は大都市圏ほど民間賃貸住宅の比率が高い事を明らかにしている。森本(1993)の調査より、高額を支払い持ち家の購入を行わずとも、民間賃貸住宅を利用することで、住居を確保出来得るという点が、民間賃貸住宅が多い大都市圏に人口が集中する一要因であると考えられる。また、所得格差という観点からは、大都市圏と農山村では主要な産業が異なるため、農山村は大都市圏と比較して所得水準が低い事が明らかにされている(根本,1972;萩原,1984)。学歴という観点からは、清水(2010)は、大都市圏ほど高度な専門知識を必要とする職が多いため、高学歴者ほど大都市圏に集中する傾

向があると論じている。更に、年齢という観点からは、20 歳代および 30 歳代が大都市圏に集中する傾向がある事が明確化されている(伊豆田・中川・西川・佐藤,2016;折田,1979)。萩原(1984)や伊豆田ら(2016)は、上述の現象を、若年層のライフスタイルに対するニーズの変化に起因するものであると解釈している。加えて、伊豆田ら(2016)は、農山村の若年層が大都市圏へ移動した理由として、農山村には「娯楽」「買い物」「街並みの活気ぶり、賑やかさ」「交通手段の充実さ」等が欠如しているためであると論じている。その他、食傾向という観点からは、大都市圏と農山村では外食傾向に差が見られること、および家庭内における食事スタイルそのものに差が見られることが明らかにされている(内野,1992;吉田,1993)。上述の通り、人が居住地区の選択・決定を行う要因は無数に存在する事が明らかにされてきた。これらにより人口の移動が起こり、過疎化・過密化が発生していると考えられる。

## 目的

本研究では、問題の節において紹介した先行研究を参照しつつ、①「個人の諸属性」および、②「ライフスタイル」という 2 つの観点から、人が居住地区の選択・決定を行う上で影響力を持つ要因について探索的に横断的研究を行う。

## 方法

(1)応答変数の選択

応答変数には「大都市圏の別」という変数を用いた。「大都市圏の別」とは居住地区を示す変数であり、「三大都市圏」と「その他」という 2 値変数である。また、本変数を選択するにあたり、「その他」には問題の節で紹介した先行研究で過疎化地域として挙げられていた農山村等以外の地域も含まれていると考えられるため、本研究では過密化、すなわち三大都市圏への人口集中を主要な論点とする。また、本研究では、便宜的に「その他」を地方として扱った。

(2)説明変数の設定

まず、①「個人の諸属性」として、「住居の所有関係」、「世帯主の性別」、「世帯主の年齢区分」の 3 変数を選択した。なお、「世帯主の年齢区分」については、40 歳未満を「若年層」、40 歳以上を「中高年層」とした。

次に、②「ライフスタイル」を表す指標として採用した変数を記載する。なお、本研究では「生活様式。特に、趣味・交際などを含めた、その人の個性を表すような生き方」(新村,2008)をライフスタイルの定義として採用した。また、本定義は非常に包括的且つ抽象的であるため、ライフスタイルを表す指標として大まかに下記の 4 軸を設定し、変数の選択を行った。第 1 軸には「収入」を設定した。第 2 軸には「飲食」を設定した。第 3 軸には「公共料金と移動手段」を、第 4 軸には「趣味・娯楽」を設定した。各軸に含まれる変数は次の通りである。第 1 軸には「実収入」を、第 2 軸には「調理食品」、「飲料」、「酒

類」、「一般外食」、「菓子類」、「健康保持用摂取品」を、第3軸には「光熱・水道」、「交通」、「自動車等関係費」を、第4軸には「諸雑費」、「(特掲) 移動電話通信料」、「交際費」、「読書」、「聴視」、「スポーツ」を設定した。

(3)分析手法：ロジスティック回帰分析

「大都市圏の別」を目的変数とし(「3大都市圏」を基準とした)、(1)に記載した各説明変数を用いて、単変量のロジスティック回帰分析を行った。なお、勤労者以外の世帯については、収入を調査していないため、本研究からは除外した。また、本研究には重みづけを行わないデータを用いた。

## 結果

各説明変数のp値および、オッズ比、95%信頼区間におけるオッズ比の上限・下限をTable1として以下に掲載した。

Table1 各説明変数のオッズ比と95%信頼区間およびp値

| 説明変数 | 項目 | p値 | オッズ比 | 95%信頼区間 | |
|---|---|---|---|---|---|
| | | | | 下限 | 上限 |
| 住居の所有関係 | 持ち家以外vs持ち家 | <.0001 | 2.117 | 2.040 | 2.196 |
| 世帯主の性別 | 男vs女 | <.0001 | 1.176 | 1.125 | 1.230 |
| 世帯主の年齢 | 若年層vs中高年層 | <.0001 | 1.157 | 1.112 | 1.204 |
| 実収入 | 10000円 | <.0001 | 1.003 | 1.002 | 1.004 |
| 菓子類 | 10000円 | <.0001 | 1.251 | 1.190 | 1.315 |
| 調理食品 | 10000円 | <.0001 | 1.340 | 1.301 | 1.380 |
| 飲料 | 10000円 | <.0001 | 1.440 | 1.356 | 1.528 |
| 酒類 | 10000円 | 0.180 | 1.027 | 0.988 | 1.068 |
| 一般外食 | 10000円 | <.0001 | 1.422 | 1.396 | 1.449 |
| 健康保持用接種品 | 10000円 | 0.746 | 1.005 | 0.977 | 1.033 |
| 光熱・水道 | 10000円 | <.0001 | 0.942 | 0.926 | 0.958 |
| 交通 | 10000円 | <.0001 | 1.457 | 1.429 | 1.485 |
| 自動車等関係費 | 10000円 | <.0001 | 0.972 | 0.967 | 0.978 |
| (特掲)移動電話通信料 | 10000円 | <.0001 | 0.882 | 0.860 | 0.905 |
| 諸雑費 | 10000円 | 0.460 | 0.998 | 0.993 | 1.003 |
| 交際費 | 10000円 | <.0001 | 0.986 | 0.980 | 0.991 |
| 読書 | 10000円 | <.0001 | 1.967 | 1.869 | 2.069 |
| 聴視・観覧 | 10000円 | <.0001 | 1.599 | 1.511 | 1.693 |
| スポーツ | 10000円 | <.0001 | 1.225 | 1.179 | 1.273 |

なお、Table1における「ライフスタイル」を表す変数(実収入以降)のオッズ比とその95%信頼区間は、当該変数が10000円増加した時の値である。「酒類」、「健康保持用摂取品」、「諸雑費」の3変数のみ、「大都市圏の別」に対して有意な影響力が確認されなかっ

た。「個人の諸属性」については、「持ち家以外」、「男性」、「若年層」が「三大都市圏」に発現しやすい事が示された。また、「ライフスタイル」については、「実収入」、「調理食品」、「飲料」、「一般外食」、「菓子類」、「交通」、「読書」、「聴視」、「スポーツ」の支出が増加するほど、「三大都市圏」に発現しやすいという結果が得られた。一方で、「光熱・水道」、「自動車等関係費」、「(特掲) 移動電話通信料」、「交際費」の支出が増加するほど「三大都市圏」に発現しにくい、言い換えれば「その他」に発現しやすいという結果となった。

## 考察

本研究結果について、「個人の諸属性」と「ライフスタイル」それぞれの観点から考察する。第一に、「個人の諸属性」について論じる。まず、持ち家以外の住居に住んでいる場合には、都心部に住んでいる可能性が高いという知見は、森本(1993)の先行研究と合致するものとなった。さらに、本研究からは、伊豆田ら(2016)や折田(1979)の「若年層である場合に都心部に住んでいる可能性が高い」という指摘を支持する結果が得られた。また、女性と比較して男性の方が都心部に住んでいる可能性が高いという新たな知見が得られた。以上の結果から、過密化・過疎化には、若年層男性の都心部への人口移動が寄与していると考えられる。また、都心部における若年層男性の生活形態として、持ち家以外の住居に住んでいる可能性の高さから、民間賃貸住宅等の借家を住居として利用している可能性が示唆された。

第二に、「ライフスタイル」という観点から考察を論じる。本研究では、ライフスタイルを表す指標として大まかに、「収入」、「飲食」、「公共料金と移動手段」、「趣味・娯楽」の 4 軸を設定した。第 1 軸に設定した「収入」が増加するほど都心部に人口が集中するという結果は、根本(1972)や萩原(1984)の先行研究を支持するものとなった。これについては、上で論じた、若年層男性の都心部への人口移動と密接に関連していると推察される。若年層男性とは、すなわち主要な労働力人口帯であると推測される。また、年功序列型の日本の賃金形態から、一般的に若年層ほど賃金が安い事が予想される。故に、主要な労働力人口帯であるとともに低賃金層に該当する若年層男性ほど高所得を望み、所得水準の高い都心部へと移動する確率が高まると推察される。また、第 2 軸の「飲食」に関して、本研究では「健康保持用摂取品」と「酒類」は有意な影響力を持たないという結果が得られたが、それ以外の「飲食」に属する変数の影響力は有意であった。本研究結果は、吉田(1993)や内野(1992)の先行研究と整合性の取れた結果であると解釈される。また、食の外部化や中食利用の増加は、萩原(1984)や伊豆田ら(2016)の指摘する、若年層のライフスタイルに対するニーズの変化を一部反映していると推察される。そして、第 3 軸の「公共料金と移動手段」からは、公共交通機関への支出が増加するほど都心部に住んでいる可能性が高まる一方で、自動車関係の支出が増加するほど地方に住んでいる可能性が高くなる事が明らかにされた。上記の都心部と地方における交通手段の差異からは、伊豆田ら(2016)

の指摘する、若年層のニーズ変化による農山村から大都市圏への移動が読み取れる。また、「光熱・水道」の支出が多くなるほど地方に住む可能性が高まるという結果は、都心部と地方との家族形態の差異に起因するものであると考察される。本研究で用いたデータにおいて、地方に単身で住む世帯は 3141 世帯であり、一方、地方に単身以外で住む世帯は 36857 世帯であった。この比率から、地方は一世帯に住む人数が多いため、それだけ光熱費の支出も大きくなる傾向にある事が予測される。次に、第 4 軸の「趣味・娯楽」については、「諸雑費」のみ有意な影響力が認められなかったが、「読書」、「聴視・観覧」、「スポーツ」の支出が増加するほど、都心部に住んでいる可能性が高まる一方で「(特掲) 移動電話通信料」、「交際費」の支出が増加するほど地方に住んでいる可能性が高まるという結果が得られた。本結果については、清水(2010)の「高学歴者ほど都心部に集中しやすい」という知見から、知的な趣味(「読書」・「聴視・観覧」)への支出が増加するほど都心部に住む可能性が高まると解釈される。また、「スポーツ」への支出については、無料でスポーツの為に利用できる環境の有無に依拠した結果であると考えられる。例えば、都心部と比較して、地方ほど空き地や山林といった環境が多く存在すると予測され、地方におけるこうした環境が、スポーツへの支出の差異に繋がったと推察される。次に、「(特掲) 移動電話通信料」、「交際費」の支出が増加するほど地方に住んでいる可能性が高まるという結果について解釈を述べる。「(特掲) 移動電話通信料」については、地方ほど娯楽施設が不足している(伊豆田ら, 2016)ため、スマートフォンに代表されるような移動電話型のデバイスが提供するサービスへの支出が増加すると考えられる。「交際費」についても、若年層のライフスタイルに対するニーズの変化(萩原, 1984;伊豆田ら, 2016)が主要な背景要因であると推察されるが、「交際費」は多面的な性質を持つ支出項目であるため、その性質を要検討・吟味したうえで、解釈を述べるべきであろう。

## 総合考察

本研究結果から、まず、若年層男性ほど都心部に集中しやすい事が示された。さらに、都心部への人口集中には、所得水準の高さや、特に若年層のライフスタイルに対する多面的なニーズの変化等、様々な要因が作用している可能性が示唆された。また、これらの要因は個々が完全に独立している訳ではなく、段階的あるいは複合的に人が居住地区の選択・決定を行う要因として作用していると考察される。

## 今後の展望

本研究では、探索的に、各説明変数が居住地区の選択・決定を行う上での影響力の有無について検証を行った。今後は、「個人の諸属性」「ライフスタイル」という観点以外からも、居住地区の選択・決定を行う上で影響力を持つ要因を検討し、多重共線性や交絡といった問題に留意しながら、それらの要因間での影響力を比較する必要があろう。加えて、今後、過密化(あるいは過疎化の双方)についてその背景要因を探る際には、因果推論につ

いても十分検討を行った上で、変数選択を行う必要があろう。また、ロジスティック回帰分析以外のアプローチ(例えば判別分析等)を用いても、整合性の高い結果が得られる事の確認が望まれる。本研究で主要な論点としたものは過密化のみであったが、過密化・過疎化が現代における大きな社会問題であり、今後も改善の兆候が伺えない事から、様々な観点・アプローチを用いて、その背景要因を明らかにしていく事は重要な課題であろう。

## 引用文献


萩原清子(1984)「過疎問題の経済学的考察」 地域学研究 **15**,185-211.

萩原清子(1993)「過密・過疎問題―地域人口配分モデルによるアプローチ―」 総合都市研究 (50),93-106.

伊豆田義人・中川　恵・西川友子・佐藤佑香(2016)「都市部への移住に起因する地域過疎化と若者の価値観との関係のモデルについて」 山形県立米沢女子短期大学紀要 (52),65-78.

森本信明(1993)「大都市圏における民間賃貸住宅の位置と家賃問題」 都市住宅学(4),3-11.

根本和泰(1972)「過疎化過程の分析―過疎地域住民の残留と移動―」 地域学研究 **3**,89-113.

新村　出(2008)「広辞苑」［第六版］岩波書店.

折田仁典(1980)「過疎地域における人口移動モチベーションに関する調査」 秋田工業高等専門学校研究紀要 **15**,87-93.

清水昌人(2010)「近年における大都市圏の転入超過の分析」 人口問題研究 **66**(1),1-16.

総務省(2012) 「平成 24 年版情報通信白書　第 1 部　特集　ICT が導く震災復興・日本再生の道筋　第 2 節　グローバルに展開する ICT 市場　(3)三大都市圏への人口集中と過疎化の進展」
〈http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/html/nc112130.html〉(2017 年 6 月 5 日)

内野澄子(1992)「人口移動と食生活の関連」 日本栄養・食料学会誌 **45**(1),1-8.

吉田昌之(1993)「外食消費の要因分析」 千葉大学園芸学部学術報告 **47**,233-240.

# 付録：規定課題用 SAS プログラム・図表

SAS データセット「Zensho2004gijimicro」に対して符号表を参照して、フォーマット、ラベルを作成。それを割り当てた SAS データセット「SAS_MICRO」を作成して以下課題の入力データセットとして使用した。表、グラフはすべて RTF に出力した。また、本課題は SAS9.4 を用いて行った。

1. 年齢階層別・世帯区分別 クロス表

| 世帯主の年齢 | 世帯区分 | | | |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | 合計 |
| 24歳以下 | 765 | 18 | 15 | 798 |
| 25～29歳 | 2,289 | 157 | 66 | 2,512 |
| 30～34歳 | 4,223 | 511 | 107 | 4,841 |
| 35～39歳 | 5,217 | 728 | 193 | 6,138 |
| 40～44歳 | 5,492 | 1,076 | 205 | 6,773 |
| 45～49歳 | 5,702 | 1,305 | 277 | 7,284 |
| 50～54歳 | 6,194 | 1,797 | 392 | 8,383 |
| 55～59歳 | 5,682 | 2,221 | 624 | 8,527 |
| 60～64歳 | 3,237 | 2,226 | 2,534 | 7,997 |
| 65～69歳 | 1,343 | 1,743 | 3,590 | 6,676 |
| 70～74歳 | 469 | 1,197 | 3,214 | 4,880 |
| 75歳以上 | 166 | 940 | 3,216 | 4,322 |
| 合計 | 40,779 | 13,919 | 14,433 | 69,131 |

| 世帯主の年齢 | 世帯区分 | | | |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | 合計 |
| 24歳以下 | 1,729 | 10 | 21 | 1,760 |
| 25～29歳 | 4,314 | 151 | 57 | 4,522 |
| 30～34歳 | 6,441 | 480 | 87 | 7,008 |
| 35～39歳 | 7,420 | 824 | 153 | 8,396 |
| 40～44歳 | 7,611 | 1,145 | 247 | 9,004 |
| 45～49歳 | 7,739 | 1,479 | 244 | 9,462 |
| 50～54歳 | 8,072 | 2,085 | 508 | 10,665 |
| 55～59歳 | 7,697 | 2,725 | 893 | 11,315 |
| 60～64歳 | 4,208 | 2,785 | 4,160 | 11,153 |
| 65～69歳 | 1,487 | 2,213 | 6,374 | 10,074 |
| 70～74歳 | 459 | 1,652 | 6,124 | 8,235 |
| 75歳以上 | 207 | 1,325 | 6,873 | 8,406 |
| 合計 | 57,386 | 16,874 | 25,741 | 100,000 |

データセット作成時に 10 万世帯比の集計用乗率を計算し、tabulate プロシージャによって集計、出力を行った。左右に並べた出力をするために ODS ステートメントにて columns=2 のオプションを指定した。

```
* 重みなし ;
proc tabulate data = SAS_MICRO1;
  class S_age setaikubun ;
  var countdummy ;
  table
    ( S_age="" all="合計" )  ,
    ( setaikubun all="合計") * countdummy="" *
sum=""*f=comma8.
    / box = "世帯主の年齢"  misstext="0"  ;
  format setaikubun setaikubunf. ;
run ;
```

```
* 10 万世帯比の重みあり ;
proc tabulate data = SAS_MICRO1;
  class S_age setaikubun ;
  var countdummy ;
  table
    ( S_age="" all="合計" )  ,
    ( setaikubun all="合計") * countdummy="" *
sum=""*f=comma8.
    / box = "世帯主の年齢" misstext="0" ;
  weight redefinedweight ;
  format setaikubun setaikubunf. ;
run ;
```

2. 世帯分類別・世帯区分別 世帯主の性別3重クロス表

| 世帯分類 | 勤労者世帯 | | | 勤労者以外の世帯 | | | 無職世帯 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 男 | 女 | 計 | 男 | 女 | 計 | 男 | 女 | 計 | 合計 |
| 単身世帯 | 6,503 | 4,207 | 10,710 | 1,030 | 867 | 1,897 | 2,463 | 5,688 | 8,151 | 20,758 |
| 夫婦のみ | 8,517 | 324 | 8,841 | 4,293 | 81 | 4,375 | 11,026 | 45 | 11,071 | 24,287 |
| 二世代 | 27,267 | 378 | 27,645 | 6,190 | 51 | 6,242 | 3,264 | 32 | 3,295 | 37,182 |
| 二世代(ひとり親) | 1,543 | 2,421 | 3,963 | 934 | 430 | 1,364 | 671 | 1,303 | 1,974 | 7,302 |
| 三世代 | 4,956 | 384 | 5,340 | 2,390 | 118 | 2,508 | 638 | 160 | 798 | 8,646 |
| その他 | 654 | 234 | 887 | 413 | 75 | 488 | 278 | 174 | 452 | 1,826 |
| 合計 | 49,439 | 7,947 | 57,386 | 15,250 | 1,623 | 16,874 | 18,340 | 7,401 | 25,741 | 100,000 |

DATAステップで10万世帯比の集計用乗率を計算し、tabulateプロシージャにより各項目の集計、出力を行った。

```
proc tabulate data=SAS_MICRO;
  class x02 x11 x08 ;
  var CountDummy;
  table ( x11 ="" all="合計" ) ,
        x02=""   * (x08="" all="計" )
        *
        CountDummy=""
        *
        sum="" * f=comma8.   (CountDummy="" *sum="合計"*f=comma8. )
        / box="世帯分類"   ;
   weight RedefinedWeight;
   format x11 X11F.   x08 GenderN.   ;   run;
```

3. パン支出金額についての各種統計グラフ

① パン支出金額及びパン支出金額に 1 を加えた金額の常用対数変換値の統計量を算出した表

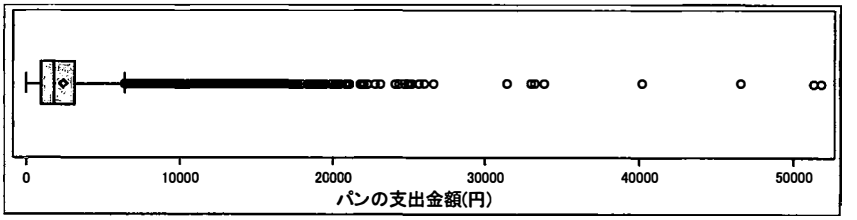
| 統計量 | パンの支出金額(円) | 1を加えた金額の常用対数 |
|---|---|---|
| 最大値 | 51,776 | 4.714 |
| 第3四分位数 | 3,039.0 | 3.4829 |
| 中央値 | 1,689.0 | 3.2279 |
| 第1四分位数 | 858.0 | 2.9340 |
| 最小値 | 0 | 0.000 |
| 平均値 | 2,284.0 | 3.1367 |
| 件数 | 100,000 | 100,000 |

summary プロシージャによって集計を行い、DATA ステップで出力に合わせた数値の計算・加工を行った。出力には report プロシージャを用いた。

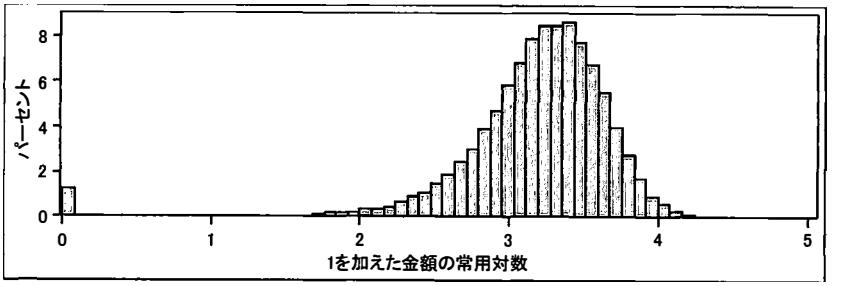
```
data micro2 ;
 length idlabel $30 ;
 set micro1 ;
 if transtype= 1 then do ;
   idlabel = "パンの支出@金額(円)" ;
   maxc     = put( round(max    , 1    ), comma12.   -L ) ;
   Q3c      = put( round(Q3     , 1e-1), comma12.1 -L ) ;
   medianc = put( round(median, 1e-1), comma12.1 -L ) ;
   Q1c      = put( round(Q1     , 1e-1), comma12.1 -L ) ;
   minc     = put( round(min    , 1    ), comma12.   -L ) ;
   meanc    = put( round(mean   , 1e-1), comma12.1 -L ) ;
   sumc     = put( round(sum    , 1    ), comma12.   -L ) ;
 end ;
 (右に続く)
```

```
else if transtype = 2 then do ;
   idlabel = "1 を加えた金額の@常用対数" ;
   maxc     = put( round(max    , 1e-3), comma12.3 -L ) ;
   Q3c      = put( round(Q3     , 1e-4), comma12.4 -L ) ;
   medianc = put( round(median, 1e-4), comma12.4 -L ) ;
   Q1c      = put( round(Q1     , 1e-4), comma12.4 -L ) ;
   minc     = put( round(min    , 1e-3), comma12.3 -L ) ;
   meanc    = put( round(mean   , 1e-4), comma12.4 -L ) ;
   sumc     = put( round(sum    , 1    ), comma12.   -L ) ;
 end ;
 label  maxc     = "最大値"    q3c     = "第 3 四分位数"
        medianc = "中央値"    q1c     = "第 1 四分位数"
        minc     = "最小値"    meanc   = "平均値"
        sumc     = "件数"   ;
run ;
```

② パン支出金額についての箱ひげ図




③ パン支出金額に 1 を加えた金額の常用対数変換値についてのヒストグラム




箱ひげ図、ヒストグラム共に sgplot プロシージャにて hbox ステートメント、histogram ステートメントの指定をして作成した。

4. 世帯区分別および世帯分類別の穀類（米、パン、めん類、他の穀類）の支出金額についての構成比

① 表4の再現

(表4-1)

| 項目名 | 符号内容 | 件数 | 構成比(%) 米 | パン | めん類 | 他の穀類 | 全体(%) |
|---|---|---|---|---|---|---|---|
| 世帯区分 | 勤労者世帯 | 57,386 | 36.0 | 37.6 | 22.1 | 4.3 | 100.0 |
| | 勤労者以外の世帯 | 16,874 | 44.9 | 31.0 | 19.4 | 4.7 | 100.0 |
| | 無職世帯 | 25,741 | 52.1 | 26.7 | 15.8 | 5.4 | 100.0 |
| 世帯分類 | 単身世帯 | 20,758 | 33.8 | 37.7 | 23.0 | 5.5 | 100.0 |
| | 夫婦のみ | 24,287 | 49.5 | 28.9 | 16.8 | 4.8 | 100.0 |
| | 二世代 | 37,182 | 39.0 | 35.9 | 20.8 | 4.3 | 100.0 |
| | 二世代(ひとり親) | 7,302 | 44.9 | 31.6 | 19.0 | 4.5 | 100.0 |
| | 三世代 | 8,646 | 46.0 | 30.5 | 19.7 | 3.8 | 100.0 |
| | その他 | 1,826 | 44.8 | 30.5 | 20.4 | 4.3 | 100.0 |
| 全体 | | 100,000 | 41.6 | 33.7 | 20.0 | 4.6 | 100.0 |

10 万世帯比の集計用乗率を計算し、means プロシージャにより各変数の構成比と世帯区分、世帯分類の件数を求めた。DATA ステップで数値の丸めを行い、出力用のデータセットを作成し、report プロシージャにより出力した。(表4-1) しかし、丸めを行う際に

```
put( round( MeanPR, 1e-1 ), 12.1 -L )
```

を記述したが、米の全体の値について問題で先に出力されていた値（41.7）と一致しなかった。

(表4-2)

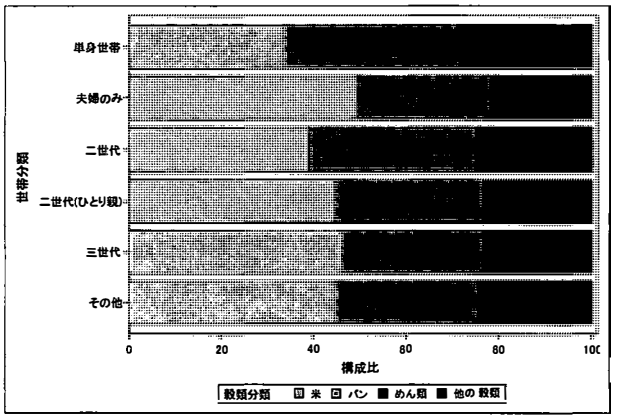
| 項目名 | 符号内容 | 件数 | 構成比(%) 米 | パン | めん類 | 他の穀類 | 全体(%) |
|---|---|---|---|---|---|---|---|
| 世帯区分 | 勤労者世帯 | 57,386 | 36.0 | 37.6 | 22.1 | 4.3 | 100.0 |
| | 勤労者以外の世帯 | 16,874 | 44.9 | 31.0 | 19.4 | 4.7 | 100.0 |
| | 無職世帯 | 25,741 | 52.1 | 26.7 | 15.8 | 5.4 | 100.0 |
| 世帯分類 | 単身世帯 | 20,758 | 33.8 | 37.7 | 23.0 | 5.5 | 100.0 |
| | 夫婦のみ | 24,287 | 49.5 | 28.9 | 16.8 | 4.9 | 100.0 |
| | 二世代 | 37,182 | 39.0 | 35.9 | 20.8 | 4.3 | 100.0 |
| | 二世代(ひとり親) | 7,302 | 44.9 | 31.6 | 19.0 | 4.5 | 100.0 |
| | 三世代 | 8,646 | 46.0 | 30.5 | 19.7 | 3.8 | 100.0 |
| | その他 | 1,826 | 44.8 | 30.5 | 20.4 | 4.3 | 100.0 |
| 全体 | | 100,000 | 41.7 | 33.7 | 20.0 | 4.7 | 100.0 |

ここで、問題の値通りにするために以下のプログラムで丸めを実施した。

```
put( round( MeanPR, 1e-2 ), 12.1 -L )
```

しかし、原則、2 回以上の丸めは禁止となるので、計算としてはふさわしくないが、今回はあくまで出力に合わせるために行った。(表4-2)

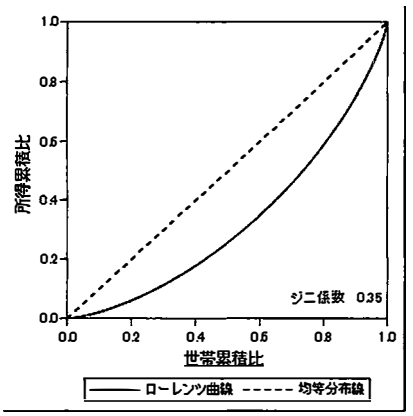
② 世帯分類別に穀類の構成比を帯びグラフで表示




sgplot プロシージャにて hbar ステートメントの指定をして作成した。

5. 世帯分類別のローレンツ曲線およびジニ係数

① 「二世代（ひとり親)」世帯の年間収入のローレンツ曲線（乗率使用）




② ジニ係数（表 5）

| 世帯分類 | ジニ係数 |
|---|---|
| 単身世帯 | 0.34 |
| 夫婦のみ | 0.31 |
| 二世代 | 0.26 |
| 二世代(ひとり親) | 0.35 |
| 三世代 | 0.27 |
| その他 | 0.33 |
| 全体 | 0.33 |

X 軸の値は「集計用乗率の累計/集計用乗率の合計」を算出して世帯累積比とした。Y 軸の値は「年収の累計/年収の合計」に重みづけをして所得累積比とした。ジニ係数はローレンツ曲線の下部の面積を台形の面積の和として算出後、1/2 から差し引くという計算を DATA ステップにて行った。グラフの出力は sgplot プロシージャにて実行、表 5 は report プロシージャにて出力した。

# 企業区分別にみる自動車保有世帯の傾向

参加カテゴリー：B）ニューカマー（SAS/JMP経験年数3年以内）
チーム名：株式会社タクミインフォメーションテクノロジー　チームC

坂本周一・齊藤祥子・髙田一行・野沢義夫

（株式会社タクミインフォメーションテクノロジー）

Trends in car ownership by households divided into company classification.

Shuichi Sakamoto, Akiko Saito, Ikkou Takata, Yoshio Nozawa
Takumi Information Technology Inc.

## 要旨

自動車産業の日本市場の低迷は 2005 年から続いており、その要因として、人口減少に伴う市場の縮小、景気の低迷、若者のクルマ離れなどが挙げられている。国土交通白書を代表とする各種文献では、世帯主の年齢や世帯形態別に自動車の保有割合が結果として示され、若者のクルマ離れといったライフスタイルの変化について言及されている。

本研究では新たな切り口として、企業の従業員規模で分類し、自動車の保有割合を比較することで、大手企業と中小に勤める世帯でライフスタイルの違いを示すことができると考えた。延いては、自動車産業の日本市場の低迷の一因を見出すことにつながると考え、研究を行った。

2004 年（平成 16 年）の全国消費実態調査データに基づく擬似ミクロデータを解析対象とし、自動車の保有世帯における「企業の従業員規模による違い」に焦点を当て、解析を行った。

結果として、「自動車の保有割合」と「年間収入」において、企業の従業員規模によって特徴があることが分かった。また、企業の従業員規模以外に、未就学児の有無や住居の所有関係の影響が強いことも示された。

キーワード：擬似ミクロデータ、自動車、企業規模、ライフスタイルの変化

## 1. 抽出条件

自動車の保有有無を判別するにあたり、当調査年度に自動車を購入した世帯及び、購入有無に関わらず、自動車に係る維持費が発生している世帯を保有世帯と設定した。

【疑似ミクロデータからの抽出条件】

・保有：自動車等購入（Y129）または、自動車等維持費（Y131）のいずれかが 1 円以上

・未保有：上記以外の世帯

全世帯での自動車の保有割合は以下の通りとなった。

| 自動車の保有 | 世帯数 | 割合 |
|---|---|---|
| 有 | 58,982 | 85.3% |
| 無 | 10,149 | 14.7% |

表１：自動車の保有割合

また、年間収入グループについては、疑似ミクロデータ内の全世帯を対象に変数 Y001（年間収入）を使用して順位付けをし、上位 20%から順に 20%ごとでグループ分けを行った。

| 分析変数 ：Y001 年間収入(単位万円)※ | | | | | |
|---|---|---|---|---|---|
| グループ | Obs 数 | 平均 | 中央値 | 最小値 | 最大値 |
| 上位 20% | 13,822 | 1,278 | 1,149 | 908 | 6,203 |
| 上位 20～40% | 13,783 | 771 | 766 | 659 | 907 |
| 40～60% | 13,811 | 569 | 568 | 489 | 658 |
| 下位 20～40% | 13,821 | 413 | 413 | 337 | 488 |
| 下位 20% | 13,894 | 230 | 243 | 2 | 336 |

表２：年間収入グループ

※　(Y001)年間収入に関して、擬似ミクロデータで定義されている単位を使用している。

## 2. 企業区分と年間収入における自動車の保有割合

はじめに、年間収入グループによって自動車の保有割合にどのような違いがあるのか、企業の従業員規模ごとに検証を行った。年間収入グループは、収入順に全体を五つに分けている。なお、今回の擬似ミクロデータでは、企業の従業員規模が大きくなるにつれ、年間収入が高くなる傾向が分かっている。図 2 の企業規模別では、データ全体の自動車の保有割合が高い傾向があり、各区分の差に大きな開きはなかったものの、図 1 より企業規模が大きくなるにつれ、自動車の保有割合が高くなる傾向が見られた。また、その中でも「上位 20～40%」に着目すると、自動車所有の割合は 95.4%（官公）、94.7%（民営・自営 500 人以上）となっており「民営・自営 1～4 人」の 86.1%とは大きく差があることが分かった。

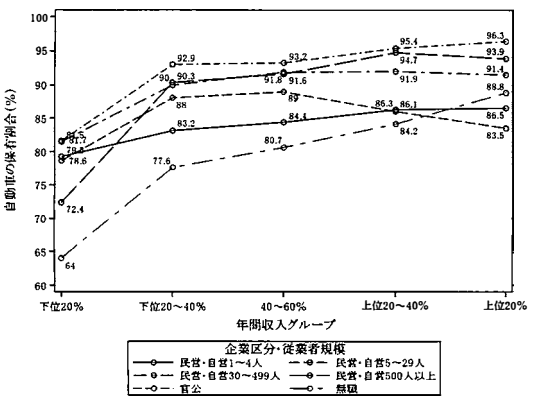



図１：年間収入グループと企業区分における自動車の保有割合

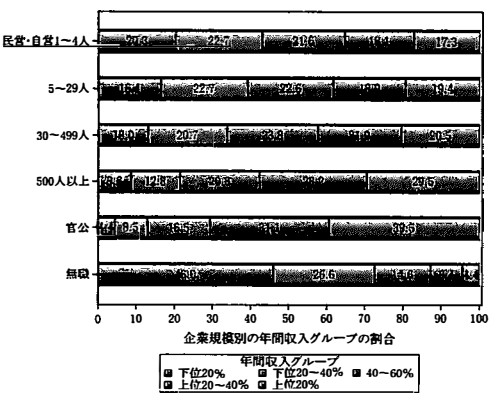



図２：企業別年間収入グループの割合

| 分析変数 ：Y001 年間収入(単位万円)※ | | | | | |
|---|---|---|---|---|---|
| 企業区分･従業者規模 | Obs 数 | 平均 | 最小値 | 中央値 | 最大値 |
| 民営･自営 1～4 人 | 12,895 | 631 | 20 | 537 | 4,591 |
| 5～29 人 | 10,615 | 673 | 23 | 562 | 6,203 |
| 30～499 人 | 12,881 | 687 | 15 | 599 | 5,017 |
| 500 人以上 | 9,865 | 775 | 16 | 723 | 3,238 |
| 官公 | 8,442 | 876 | 56 | 820 | 3,123 |
| 無職 | 14,433 | 408 | 2 | 356 | 3,102 |

表 3：企業区分・従業者規模における年間収入
※　(Y001)年間収入に関して、擬似ミクロデータで定義されている単位を使用している。

また、図 1 より「民営・自営 1～4 人」の自動車の保有割合は、年間収入グループの下位 20%とそれ以外で差が小さいことが読み取れる。これは地方農業従業者や個人事業主など、年間収入グループに関わらず、仕事上の理由で自動車を保有する必要があるのではないかと推測をしたが、自動車を保有している世帯の中で「3 大都市圏」(87.5%)と「その他」(82.3%)の世帯割合に大きな差が見られなかった。加えて農林漁業収入など関連する変数との相関を検証したが、こちらも相関は見られなかったため、この点については要因の特定にまで至らなかった。

## 3. 企業区分・年齢における自動車の保有割合

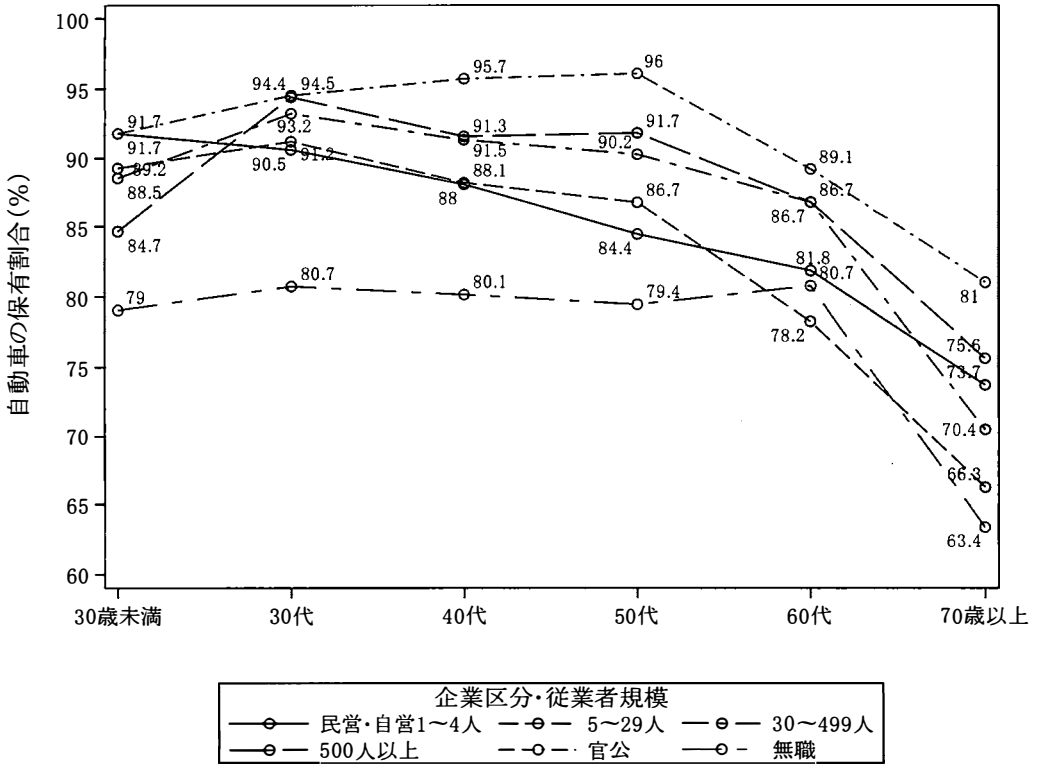



図 3：世帯主の年齢(世代)と企業区分における自動車の保有割合

| | 企業区分・従業者規模 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 民営・自営 1～4人 | | 5～29人 | | 30～499人 | | 500人以上 | | 官公 | | 無職 | |
| 世帯主の年齢 | 世帯数 | (%) | 世帯数 | (%) | 世帯数 | (%) | 世帯数 | (%) | 世帯数 | (%) | 世帯数 | (%) |
| 24歳以下 | 70 | 92.1 | 150 | 85.2 | 156 | 80.8 | 170 | 71.1 | 89 | 89.9 | 9 | 60.0 |
| 25～29歳 | 261 | 91.6 | 427 | 90.7 | 606 | 90.7 | 545 | 90.1 | 384 | 92.1 | 55 | 83.3 |
| 30～34歳 | 605 | 92.1 | 851 | 91.5 | 1,179 | 93.2 | 1,030 | 93.9 | 730 | 93.0 | 89 | 83.2 |
| 35～39歳 | 718 | 89.2 | 984 | 90.9 | 1,409 | 93.1 | 1,440 | 94.8 | 980 | 95.6 | 153 | 79.3 |
| 40～44歳 | 1,018 | 88.2 | 1,065 | 89.3 | 1,494 | 92.7 | 1,315 | 92.3 | 1,147 | 96.7 | 174 | 84.9 |
| 45～49歳 | 1,090 | 87.8 | 1,092 | 87.0 | 1,510 | 90.0 | 1,294 | 90.7 | 1,335 | 94.9 | 212 | 76.5 |
| 50～54歳 | 1,444 | 85.1 | 1,353 | 87.0 | 1,692 | 90.6 | 1,338 | 92.7 | 1,371 | 96.1 | 305 | 77.8 |
| 55～59歳 | 1,573 | 83.8 | 1,286 | 86.4 | 1,706 | 89.8 | 1,200 | 90.7 | 1,261 | 95.9 | 502 | 80.4 |
| 60～64歳 | 1,515 | 83.4 | 1,059 | 82.3 | 1,116 | 87.5 | 504 | 88.3 | 461 | 89.7 | 2,133 | 84.2 |
| 65～69歳 | 1,194 | 79.9 | 446 | 69.9 | 507 | 84.8 | 140 | 81.4 | 160 | 87.4 | 2,807 | 78.2 |
| 70～74歳 | 781 | 77.9 | 243 | 69.6 | 162 | 75.3 | 25 | 69.4 | 54 | 84.4 | 2,288 | 71.2 |
| 75歳以上 | 539 | 68.2 | 113 | 60.1 | 59 | 59.6 | 9 | 100.0 | 14 | 70.0 | 1,786 | 55.5 |

表4：企業区分と世帯主の年齢における自動車の保有割合

次に企業区分で分けた場合の年代ごとの自動車の保有の割合に差があるのか検証を行った。「民営・自営500人以上」、「官公」の自動車の保有割合の割合は35歳～44歳の間で最大値を取っており、それぞれ「官公」が「40～44歳」の時点で96.7%、「民営・自営500人以上」は「35～39歳」の時点で94.8%となっている。このことから生活が安定し、家族構成員が増加するなど生活的な要因の影響を受けて自動車保有の割合が増加していると考えられる。対して「民営・自営1～4人」、「民営・自営5～29人」では24歳～29歳までの間で最大値を取っており、その後は緩やかに下降傾向がある。ここに区分される世帯の詳細を分析したところ、就業人員が1人で未就学児有りの場合、一戸建てに住んでいる世帯割合(33.51%)より、一戸建て以外に住んでいる世帯割合(66.49%)が多く、このことから若い年齢で結婚し、未就学児のいる世帯が生活の足として自動車を所有していると推測される。

国土交通省の「全国都市交通特性調査」の「女性の子供の有無別の交通分担率」を参照すると、子供がいる就業者（平日）において全国の自動車の交通分担率は、どの年代においても子供のいない世帯より大きく、本データにおいてもこの傾向があるものと考えられる。

| | | 子供がいる就業者(平日)［%］ | | | | | | 子供がいない就業者(平日)［%］ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 年代 | 鉄道 | バス | 自動車 | 二輪車 | 徒歩 | | 鉄道 | バス | 自動車 | 二輪車 | 徒歩 |
| 全国 | 1987年 | 5 | 4 | 35 | 40 | 16 | | 2 | 2 | 28 | 37 | 30 |
| | 1999年 | 8 | 1 | 53 | 26 | 12 | | 1 | 1 | 45 | 26 | 26 |
| | 2010年 | 7 | 2 | 59 | 20 | 11 | | 2 | 1 | 52 | 20 | 24 |

表5：女性の子供の有無別の交通分担率（自動車のみ）［2013年国土交通白書より一部抜粋］

## 4. 企業区分と世帯人員における自動車の保有割合

企業区分・従業者規模ごとに、世帯人員で自動車の保有割合を比較した。世帯人員が1人の単身世帯の場合は、自動車の保有割合が低く、2人以上の世帯では、企業の従業者規模が大きくなるにつれ、自動車の保有割合も緩やかに上昇していた。また、世帯人員が3人以上の場合(3人、4人、5人以上)、企業の従業者規模と自動車の保有割合の関係は、近似していた。

単身世帯とそれ以外で差が見られたのは、収入の安定性、もしくは人員が増えることでの自動車利用

の必要性が影響を及ぼしていると推測される。自動車を保有（購入）する理由として、子どもの有無や高齢者の有無が影響していると考えた。

また、単身世帯の自動車の保有割合が低い理由として、単身世帯のうち、65 歳以上の女性が 47.08%と割合が高かった。これは女性の免許保有者のうち 65 歳以上の女性が 6.55%（「交通統計 平成 16 年」交通事故分析センター）しかいないことが大きな影響していると考えられる。

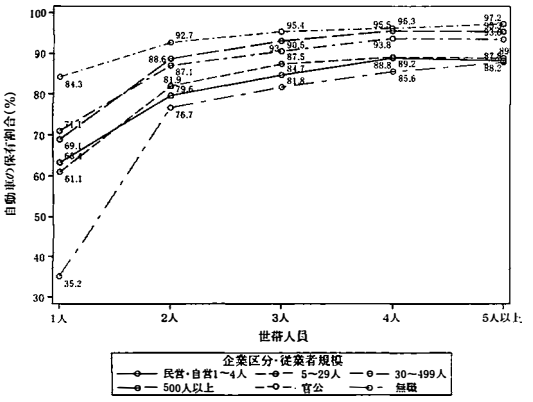



図 4：企業区分と世帯人員における自動車の保有割合

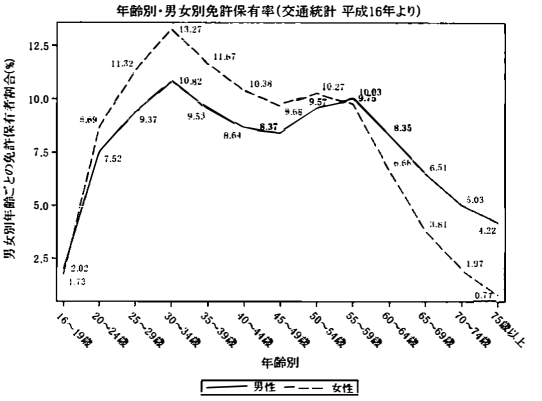



図 5：年齢別・男女別 免許保有者数「交通統計 平成 16 年より」

## 5. 企業区分と性別における自動車の保有割合

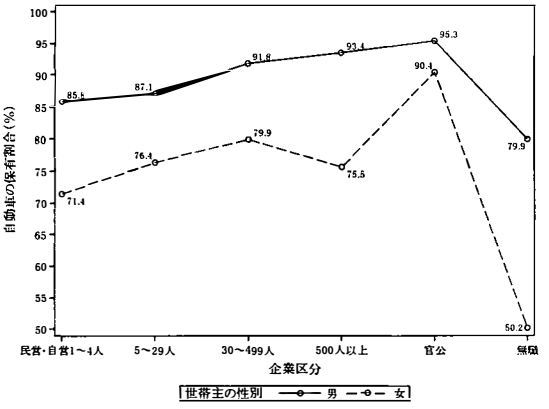



図 6：企業区分と性別における自動車の保有割合

| 分析変数：Y001 年間収入(単位千円) | | | | | |
|---|---|---|---|---|---|
| 企業区分・従業者規模 | Obs 数 | 平均 | 最小値 | 中央値 | 最大値 |
| 民営・自営 1～4 人 | 1,758 | 479 | 20 | 393 | 2,610 |
| 5～29 人 | 1,660 | 446 | 23 | 363 | 2,531 |
| 30～499 人 | 1,897 | 450 | 15 | 371 | 3,113 |
| 500 人以上 | 1,151 | 482 | 44 | 405 | 2,025 |
| 官公 | 1,253 | 703 | 56 | 667 | 2,504 |
| 無職 | 3,425 | 320 | 7 | 242 | 3,102 |

表 6：女性の世帯主と企業区分・従業員規模における年間収入

全体で世帯主の性別ごとの自動車の保有割合は、男性 88.3%、女性 69.6%と、男性の方が高いことが分かった。企業区分ごとに比較したところ、企業の従業者規模が大きくなるにつれ、男性の自動車保有の割合は緩やかに上昇している。

また、企業区分「官公」においては、男性と女性の自動車保有の割合の差が小さいことが読み取れる。これは、世帯主が女性の場合、企業区分「官公」の年収平均が、「民営・自営」の約 1.52 倍と大きく異なっていることが影響していると推測される。その他要因として考えられる、未就学児の有無に関して検討したが、自動車を保有している世帯主が女性の場合、未就学児の有無に関わらず「官公」（未就学児有：8.65%、無：91.35%）と「民営・自営」（未就学児有：9.58%、無：90.42%）と、ほとんど割合には違いが見られなかった。

## 6. 企業区分と戸建世帯における自動車の保有割合

次に企業規模ごとに「戸建世帯」と「戸建以外の世帯」で車の保有率に差があるかを検証した（図 7）。全体的に戸建世帯は、戸建以外の世帯よりも車保有率が高い傾向があった。要因として戸建世帯での駐車スペース、収入、家族構成など自動車を所有する因子が複合的に含まれているからではないかと推測する。駐車スペースについては、戸建購入の理由として「駐車スペースが無料」が 51.8%であることから、戸建における自動車の割合の高さは関係があるといえる。

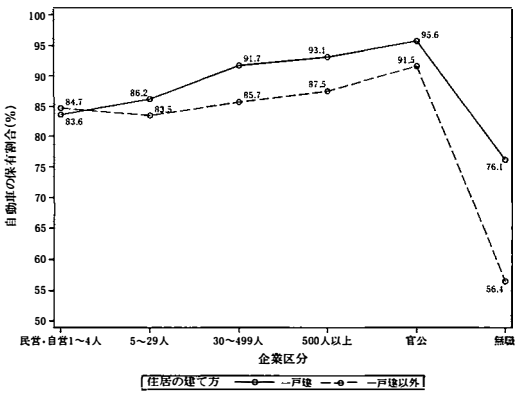



図 7：企業区分と住居の建て方における自動車の保有割合

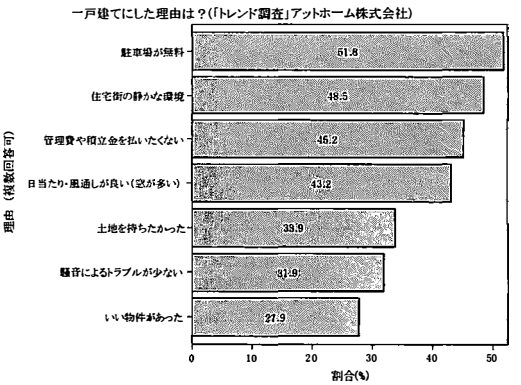



図 8：一戸建てにした理由　［トレンド調査より］

収入と戸建世帯については、図 9 で示す通り、世帯主の収入が高いほど、一戸建て世帯の割合が多くなる傾向にあることがいえる。加えて、先の図 2 より年間収入グループと企業の従業者規模が大きくなるにつれ、自動車保有の割合が上がる傾向があった。よって経済的な余裕から戸建世帯の方が戸建以外の世帯よりも自動車保有の割合が多いと推察する。世帯人員と車保有の相関については、図 4 より世帯人員が増えるにつれ、自動車保有の割合が上がる傾向があった。また、戸建世帯は世帯人員に単身以外の割合が高いことから、世帯人員が多い戸建世帯の方が戸建以外世帯よりも車保有の割合が大きくなると考える。

一戸建て世帯(図 7)を企業規模別でみた場合、世帯主の企業区分「民営・自営 1～4 人」では、戸建の車保有率が 83.6%、戸建以外の車保有率が 84.7%であり、その差は 0.9%である。よって「民営・自営 1～4 人」の区分での戸建以外で車保有率に大きな差は無いと考えられる。企業区分「民営・自営 5～29 人」「民営・自営 30 人～499 人」「民営・自営 500 人以上」の保有率について戸建世帯は戸建以外の世帯よりも割合が高い傾向がある。

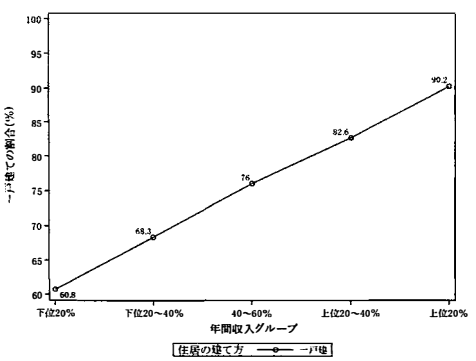



図 9：一戸建て世帯数と年間収入グループ

## 7. まとめ

本研究では自動車の保有率は年間所得や世帯人員の構成に大きな影響を受けることが分かった。企業の従業員の規模が大きくなるにつれ自動車の保有率が上昇することから、企業の従業員規模と自動車の保有率には相関があると考えられ、企業の従業員規模が大きくになるにつれ年間所得も大きくなることから、年間所得と自動車の保有率にも同じように相関があるといえる。「5. 企業区分と性別における自動車保有の割合」では官公に区分される女性の自動車の保有率（90.4%）は男性の割合（95.3%）と近似していた。女性の年間所得の平均が企業規模に関わらず近似している中で年間所得の平均が際立って高い官公区分の女性の自動車の保有率が高いということも大変興味深い結果であった。このことからも年間所得と自動車購入の保有率には関係があるといえるのではないだろうか。

一方で年齢や企業の従業員規模に関わらず自動車の保有をしている世帯には未就学児の有無も密接に関係しており、「3. 企業区分・年齢における自動車の保有割合」のデータから若い世代で企業の従業員規模が小さい場合でも、世帯人員の増加があった場合には自動車の保有率が高い水準であることが分かる。また、「6. 企業区分と戸建世帯における自動車の保有」のデータで一戸建てその世帯の生活状況によっても自動車購入の優先順位が高くなる傾向が見られた。

## 8. おわりに

今回使用した疑似ミクロデータは 2004 年時点のものであったが、それ以降の自動車の保有率関連の状況についても外部文献を利用して考察を行った。

国土交通白書（2013）によると 20-29 歳の運転免許保有率は 2001-2006 年の間で 85.5%と高い水準で推移しており、若者に関して自動車の保有はしていないが、将来的な必要性は感じていると考えられる。加えて、自動車の一時的な利用として近年ではレンタカーの他にカーシェアリングがビジネスとして展開されており、このサービスを利用した後の 10 代・20 代の自動車を保有したいという意向は 49.9%から 86.2%と大幅に上昇している。

こうしたサービスの展開は長期的な観点において、自動車の購入と保有率の上昇につながっていくのではないかと考える。

## 参考文献


- 「交通統計 平成 16 年」 交通事故分析センター
- 「2013 年国土交通白書」 国土交通省
  https://www.mlit.go.jp/hakusyo/mlit/h24/index.html
- 「トレンド調査」 アットホーム株式会社
  http://athome-inc.jp/wp-content/uploads/2015/07/12102301.pdf

付録　規定課題

*1)年齢階層・世帯区分別 クロス表*

*表1*

| 世帯主の年齢 | 世帯区分 勤労者 | 勤労者以外 | 無職 | 合計 | 世帯主の年齢 | 世帯区分 勤労者 | 勤労者以外 | 無職 | 合計 |
|---|---|---|---|---|---|---|---|---|---|
| 24 歳以下 | 765 | 18 | 15 | 798 | 24 歳以下 | 1,729 | 10 | 21 | 1,760 |
| 25～29 歳 | 2,289 | 157 | 66 | 2,512 | 25～29 歳 | 4,314 | 151 | 57 | 4,522 |
| 30～34 歳 | 4,223 | 511 | 107 | 4,841 | 30～34 歳 | 6,441 | 480 | 87 | 7,008 |
| 35～39 歳 | 5,217 | 728 | 193 | 6,138 | 35～39 歳 | 7,420 | 824 | 153 | 8,396 |
| 40～44 歳 | 5,492 | 1,076 | 205 | 6,773 | 40～44 歳 | 7,611 | 1,145 | 247 | 9,004 |
| 45～49 歳 | 5,702 | 1,305 | 277 | 7,284 | 45～49 歳 | 7,739 | 1,479 | 244 | 9,462 |
| 50～54 歳 | 6,194 | 1,797 | 392 | 8,383 | 50～54 歳 | 8,072 | 2,085 | 508 | 10,665 |
| 55～59 歳 | 5,682 | 2,221 | 624 | 8,527 | 55～59 歳 | 7,697 | 2,725 | 893 | 11,315 |
| 60～64 歳 | 3,237 | 2,226 | 2,534 | 7,997 | 60～64 歳 | 4,208 | 2,785 | 4,160 | 11,153 |
| 65～69 歳 | 1,343 | 1,743 | 3,590 | 6,676 | 65～69 歳 | 1,487 | 2,213 | 6,374 | 10,074 |
| 70～74 歳 | 469 | 1,197 | 3,214 | 4,880 | 70～74 歳 | 459 | 1,652 | 6,124 | 8,235 |
| 75 歳以上 | 166 | 940 | 3,216 | 4,322 | 75 歳以上 | 207 | 1,325 | 6,873 | 8,406 |
| 合計 | 40,779 | 13,919 | 14,433 | 69,131 | 合計 | 57,386 | 16,874 | 25,741 | 100,000 |

*2)世帯分類別・世帯区分別 世帯主の性別 3重クロス表*

*表2*

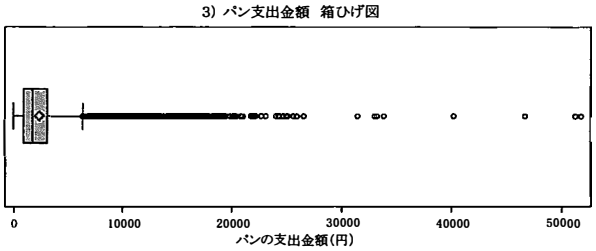
| 世帯分類 | 勤労者世帯 男 | 女 | 計 | 勤労者以外の世帯 男 | 女 | 計 | 無職世帯 男 | 女 | 計 | 合計 |
|---|---|---|---|---|---|---|---|---|---|---|
| 単身世帯 | 6,503 | 4,207 | 10,710 | 1,030 | 867 | 1,897 | 2,463 | 5,688 | 8,151 | 20,758 |
| 夫婦のみ | 8,517 | 324 | 8,841 | 4,293 | 81 | 4,375 | 11,026 | 45 | 11,071 | 24,287 |
| 二世代 | 27,267 | 378 | 27,645 | 6,190 | 51 | 6,242 | 3,264 | 32 | 3,295 | 37,182 |
| 二世代(ひとり親) | 1,543 | 2,421 | 3,963 | 934 | 430 | 1,364 | 671 | 1,303 | 1,974 | 7,302 |
| 三世代 | 4,956 | 384 | 5,340 | 2,390 | 118 | 2,508 | 638 | 160 | 798 | 8,646 |
| その他 | 654 | 234 | 887 | 413 | 75 | 488 | 278 | 174 | 452 | 1,826 |
| 合計 | 49,439 | 7,947 | 57,386 | 15,250 | 1,623 | 16,874 | 18,340 | 7,401 | 25,741 | 100,000 |

*3) パン支出金額についての各種統計グラフ*

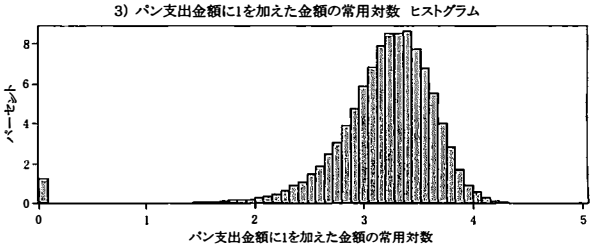
*表3*

| 統計量 | パンの支出金額(円) | 1を加えた金額の常用対数 |
|---|---|---|
| 最大値 | 51,776 | 4.714 |
| 第3四分位数 | 3,039.0 | 3.4829 |
| 中央値 | 1,689.0 | 3.2279 |
| 第1四分位数 | 858.0 | 2.9340 |
| 最小値 | 0 | 0.000 |
| 平均値 | 2,284.0 | 3.1367 |
| 件数 | 100,000 | 100,000 |

3) パン支出金額　箱ひげ図




3) パン支出金額に1を加えた金額の常用対数　ヒストグラム

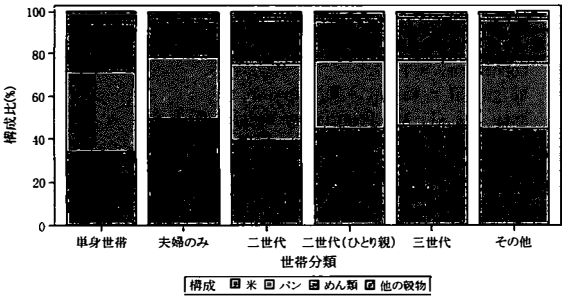


*4)世帯分類別の穀類(米、パン、めん類、他の穀類)の構成比*

*表4*

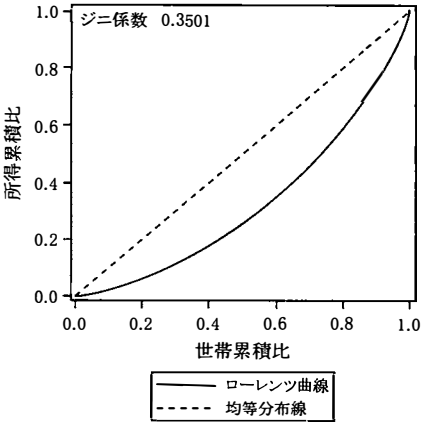
| | | | 構成比(%) | | | | |
|---|---|---|---|---|---|---|---|
| 項目名 | 符号内容 | 件数 | 米 | パン | めん類 | 他の穀物 | 全体(%) |
| 世帯区分 | 勤労者世帯 | 57,386 | 36.0 | 37.6 | 22.1 | 4.3 | 100.0 |
| | 勤労者以外の世帯 | 16,874 | 44.9 | 31.0 | 19.4 | 4.7 | 100.0 |
| | 無職世帯 | 25,741 | 52.1 | 26.7 | 15.8 | 5.4 | 100.0 |
| 世帯分類 | 単身世帯 | 20,758 | 33.8 | 37.7 | 23.0 | 5.5 | 100.0 |
| | 夫婦のみ | 24,287 | 49.5 | 28.9 | 16.8 | 4.9 | 100.0 |
| | 二世代 | 37,182 | 39.0 | 35.9 | 20.8 | 4.3 | 100.0 |
| | 二世代(ひとり親) | 7,302 | 44.9 | 31.6 | 19.0 | 4.5 | 100.0 |
| | 三世代 | 8,646 | 46.0 | 30.5 | 19.7 | 3.8 | 100.0 |
| | その他 | 1,826 | 44.8 | 30.5 | 20.4 | 4.3 | 100.0 |
| | 全体 | 100,000 | 41.7 | 33.7 | 20.0 | 4.7 | 100.0 |

4)世帯分類別の穀類(米、パン、めん類、他の穀類)の構成比 帯グラフ




*5)世帯分類別のローレンツ曲線およびジニ係数*

5) 世帯分類「二世代(ひとり親)」世帯の年間収入(Y001)のローレンツ曲線




*5) 6つの世帯分類の年間収入(Y001)のジニ係数*

*表5*

| 世帯分類 | ジニ係数 |
|---|---|
| 単身世帯 | 0.34 |
| 夫婦のみ | 0.31 |
| 二世代 | 0.26 |
| 二世代(ひとり親) | 0.35 |
| 三世代 | 0.27 |
| 全体 | 0.33 |

付録 規定課題のプログラム

*1)年齢階層・世帯区分別 クロス表*

```
* ADS ライブラリ定義: libname ADS "フォルダパス" ;
* 表1;
title "2017年 第5回 ＳＡＳミクロデータ分析コンテスト" ;
title2 "規定課題 1) 年齢階層・世帯区分別 クロス表" ;
title3 "表1" ;
* フォーマット作成 ;
proc format ;
  value X09FS        5 = "24歳以下"  6 = "25～29歳"
                     7 = "30～34歳"  8 = "35～39歳"
                     9 = "40～44歳" 10 = "45～49歳"
                    11 = "50～54歳" 12 = "55～59歳"
                    13 = "60～64歳" 14 = "65～69歳"
                    15 = "70～74歳" 16 = "75歳以上"
                 99 = "合計" ;
  value X02F         1 = "勤労者世帯" 2 = "勤労者以外の世帯"
                 3 = "無職世帯" ;
run ;
/* 10万人比集計用の WEIGHT 算出 */
proc summary data = ADS.Zensho2004gijimicro ;
  var Weight ;
  output out = work.Weight ( keep = TotalWeight )
         sum = TotalWeight ;
run ;
/* 表作成のデータ準備 */
data work.Kadai_1a ;
  set ADS.Zensho2004gijimicro ( keep = X02 X09 Weight );
  retain Multiplier ;
  if _N_ = 1 then do ;
    set Weight ;
    Multiplier = 1e5 / TotalWeight ;
  end ;
  CountDummy = 1 ;
  Tabletype = 1 ;
  RedefinedWeight = 1 ;
  output ;
  Tabletype = 2 ;
  RedefinedWeight = Weight * Multiplier ;
  output ;
  format _all_ ;
run ;
/* 集計対象となるカテゴリの組み合わせを作成 */
data work.ClassData1 ;
  do TableType=1 to 2 ; do X09 = 5 to 16 ; do X02 = 1 to 3 ;
  output;
  end; end; end;
run;
/* MEANS プロシジャによって集計データセット */
proc means data = work.Kadai_1a noprint classdata =
work.ClassData1 ;
  class TableType X09 X02 ;
  types TableType * X09 * X02
    TableType * X09  TableType * X02  TableType ;
  var CountDummy ;
```

*1)年齢階層・世帯区分別 クロス表【続き】*

```
  weight RedefinedWeight ;
  output out = work.Means1 sum = SumCnt ;
run ;
data work.Means2 ;
  set work.Means1 ;
  if missing( X09 ) then X09 = 99 ;
  if missing( X02 ) then X02 = 4 ;
  if missing( SumCnt ) then SumCnt = 0 ;
  length SumCntC $ 100 ;
  SumCntC = put( round ( SumCnt, 1 ), comma12. -L ) ;
run ;
proc sort data = work.Means2 ;
  by X09 TableType X02 ;
run ;
proc transpose data = work.Means2 out = work.Means3 prefix
= C ;
  by X09 ; id TableType X02 ; var SumCntC ;
run ;
/* 空欄を作成 */
data work.Means4 ;
 set work.Means3 ;
 C00 = "" ;
run ;
```

```
/* rtf 出力 */
ods rtf file = "rtf 出力先" style = listing bodytitle ;
proc report data = work.Means4 nowd ;
  column  ( X09 ( "世帯区分" C11 C12 C13 ) C14 )  ( "09"x C00 )
    ( X09 ( "世帯区分" C21 C22 C23 ) C24 ) ;
  define X09
  / "世帯主の/年齢" order order = internal center flow ;
  define C11 / "勤労者" right width = 8 ;
  define C12 / "勤労者/以外" right width = 8 ;
  define C13 / "無職" right width = 8 ;
  define C14 / "合計" right width = 8 ;
  define C00 / "" display ;
  define X09 / order order = internal flow width = 8 ;
  define C21 / "勤労者" right width = 8 ;
  define C22 / "勤労者/以外" right width = 8 ;
  define C23 / "無職" right width = 8 ;
  define C24 / "合計" right width = 8 ;
  format X09 X09FS. ;
run ;
ods rtf close ;
```

*2)世帯分類別・世帯区分別 世帯主の性別 3重クロス表*

```
* 表2;
proc format ;
 value X08F 1 = "男" 2 = "女" ;
 value X11F 1 = "単身世帯" 2 = "夫婦のみ"
            3 = "二世代"  4 = "二世代（ひとり親）"
            5 = "三世代"  6 = "その他"  ;
run ;
/*10 万世帯比の算出*/
data work.Micro2_1 ;
 set ADS.zensho2004gijimicro ;
  retain Multiplier ;
  if _N_ = 1 then do ;
   set Weight ;
   Multiplier = 1e5 / TotalWeight ;
  end;
  CountDummy = 1 ;
  Weight_HT = Weight * Multiplier ;
  keep X02 X11 X08 CountDummy Weight_HT ;
run ;
/*10 万世帯比の算出*/
data work.Micro2_1 ;
 set ADS.zensho2004gijimicro ;
  retain Multiplier ;
  if _N_ = 1 then do ;
   set Weight ;
   Multiplier = 1e5 / TotalWeight ;
  end;
```

```
  CountDummy = 1 ;
  Weight_HT = Weight * Multiplier ;
  keep X02 X11 X08 CountDummy Weight_HT ;
run ;
options orientation = portrait papersize = A4 nodate nonumber  ;
ods rtf file = "rtf 出力先"  bodytitle;
ods listing close;
title1 "2）世帯分類別・世帯区分別 世帯主の性別 3 重クロス表
";
title2 "表2";
proc tabulate data = work.Micro2_1 out = work.Micro2_1_F ;
  class X02 X08 X11 ;
  var CountDummy ;
  keylabel N = " " ;
  table  (X11 = "" ALL = "合計" ) , X02 = "" * (X08 = "" ALL =
"計") * ( CountDummy = ""* sum = "" * f = comma8.) (CountDummy
= "" * SUM = "合計" * f = comma8.)
          / box = "世帯分類";
  weight Weight_HT ;
  format X02 X02F. X08 X08F. X11 X11F. ;
run ;
ods rtf close ;
ods listing ;
title ;
```

*3) パン支出金額についての各種統計グラフ*

```
* 表3;
title1 " 3） パン支出金額についての各種統計グラフ" ;
title2 "表3" ;
/* パン支出金額に関するデータセット */
data work.Bread1 ;
  set ADS.zensho2004gijimicro ;
  retain Multiplier ;
if _N_ = 1 then do ;
    set Weight;
    Multiplier = 1e5 / TotalWeight ;
  end ;
  CountDummy = 1 ;
  RedefinedWeight = Weight * Multiplier ;
```

```
  TransType = 1 ;
  AVAL =Y044;
  output ;
  TransType = 2 ;
  AVAL=log10(Y044 + 1 ) ;
  output ;
  format _all_ ;
run ;
proc sort data = work.Bread1 ;
  by TransType ;
run ;
```

*3) パン支出金額についての各種統計グラフ【続き】*

```
/** means データを作成する*/
proc means data= work.Bread1 noprint nway;
  class TransType ;
  var AVAL CountDummy ;
  weight RedefinedWeight ;
  output out=BreadMeans1 max = Max q3 = Q3 median = Median
  q1 = Q1 min = Min mean = Mean sum = SumAval SumCnt ;
run ;
data work.BreadMeans2 ;
  set work.BreadMeans1;
  length MaxC Q3C MedianC Q1C MinC MeanC NC IdLabel $100 ;
  if TransType = 1 then do ;
    IdLabel = "パンの支出/金額（円）" ;
    MaxC = put( round( Max, 1 ), comma12.0 -L ) ;
    Q3C = put( round( Q3, 1e-1 ), comma12.1 -L ) ;
    MedianC = put( round( Median, 1e-1 ), comma12.1 -L ) ;
    Q1C = put( round( Q1, 1e-1 ), comma12.1 -L ) ;
    MinC = put( round( Min, 1 ), comma12. -L ) ;
    MeanC = put( round( Mean, 1e-1 ), comma12.1 -L ) ;
    NC = put( round( SumCnt, 1 ), comma12. -L ) ;
  end;
  else if TransType = 2 then do ;
    IdLabel = "1を加えた金額の/常用対数" ;
    NC = put( round( SumCnt, 1 ), comma12. -L ) ;
    MaxC = put( round( Max, 1e-3 ), comma12.3 -L ) ;
    MinC = put( round( Min, 1e-3 ), comma12.3 -L ) ;
    Q3C = put( round( Q3, 1e-4 ), comma12.4 -L ) ;
    Q1C = put( round( Q1, 1e-4 ), comma12.4 -L ) ;
    MedianC = put( round( Median, 1e-4 ), comma12.4 -L ) ;
    MeanC = put( round( Mean, 1e-4 ), comma12.4 -L ) ;
  end ;
  label MaxC = "最大値" Q3C = "第3四分位数" MedianC = "
中央値"
    Q1C = "第1四分位数" MinC = "最小値" MeanC = "平均値"
    NC = "件数" ;
run ;
proc transpose data = work.BreadMeans2 out =
work.BreadMeans3 prefix = C ;
id TransType ; idlabel IdLabel ;
  var MaxC Q3C MedianC Q1C MinC MeanC NC ;
run ;
data work.BreadMeans4 ;
  set work.BreadMeans3 ;
  if upcase( _NAME_ ) in( "MEANC" "NC" ) then StatType = 2 ;
  else StatType = 1 ;
run ;
ods listing close ;
ods rtf file = " rtf出力先" bodytitle ;
ods rtf startpage = no ;
proc report data = work.BreadMeans4 nowd split = "/" box ;
  column StatType _LABEL_ C1 C2;
  define StatType / order order = internal noprint;
  define _LABEL_ / "統計量" style (column) = {just=left
cellwidth=3cm};
  define C1 - C2 / style (column) = {just = right
cellwidth=3cm} ;
run;
/*3) ―②箱ひげ図を作成する*/
ods graphics / height = 6cm width = 14cm ;
title1 " 3） パン支出金額 箱ひげ図" ;
proc sgplot data = work.Bread1 ;
  where TransType = 1 ;
  hbox AVAL ;
  label AVAL = "パンの支出金額（円）" ;
run ;
/*3)―③ヒストグラムを作成する*/
ods graphics / height = 6cm width = 14cm ;
title1 " 3） パン支出金額に1を加えた金額の常用対数 ヒス
トグラム" ;
proc sgplot data = work.Bread1 ;
  where TransType = 2 ;
  histogram AVAL ;
  label AVAL = "パン支出金額に1を加えた金額の常用対数" ;
run ;
title ; ods graphics / reset ; ods rtf close startpage=yes ;
ods listing ; ods HTML ;
```

*4)世帯分類別の穀類(米、パン、めん類、他の穀類)の構成比*

```
* 表4 ;
/* 帳票出力用のフォーマットを作成 */
proc format ;
  value Coln 11    = "勤労者世帯" 12 = "勤労者以外の世帯"
    13 = "無職世帯" 101      = "単身世帯" 102 = "夫婦のみ"
    103  = "二世代" 104      = "二世代（ひとり親）"
    105  = "三世代" 106      = "その他" 200 = "全体" ;
  value TYPEF 0 = " "  1 = "世帯分類" 2 = "世帯区分" ;
  value $ _NAME_ "PerY043_Mean" = "米"
                 "PerY044_Mean" = "パン"
                 "PerY045_Mean" = "めん類"
                 "PerY046_Mean" = "他の穀物" ;
run ;
/*10万世帯比の分母となる乗率を算出し、マクロ変数に代入 */
data _null_ ;
  set ADS.zensho2004gijimicro(keep= Weight) end = EOF ;
  Weight2 + Weight ;
  if EOF = 1 then call symputx( "TotalWeight" , Weight2 ) ;
run ;
data work.kadai_4a ;
  set ADS.zensho2004gijimicro ;
/* 各穀類の構成比を世帯ごとに算出 */
  if Y042 ^= 0 then do ;
    PerY043 = Y043 / Y042 * 100 ;
    PerY044 = Y044 / Y042 * 100 ;
    PerY045 = Y045 / Y042 * 100 ;
    PerY046 = Y046 / Y042 * 100 ;
    PerAll  = sum(of Y043 - Y046 ) / Y042 * 100 ;
  end ;
/* 10万世帯比用の乗率を算出、カウント用のダミー変数を作成
*/
  Weight_10 = 100000 / &TotalWeight. * Weight ;
  Dummy = 1 ;
  keep X02 X11 Weight_10 PerY043 - PerY046 PerAll Dummy ;
run ;
/* 世帯区分、世帯分類ごとに統計量を算出 */
proc summary data = kadai_4a mean sum ;
  class X02 X11;
/* 組み合わせは全体、X02のみ、X11のみでアウトプット */
  types () X02 X11 ;
  var PerY043 - PerY046 PerAll Dummy ;
  weight Weight_10 ;
  output out = kadai_4b mean = sum = / autoname ;
run ;
data work.kadai_4c ;
  set work.kadai_4b ;
/* 世帯区分、世帯分類、全体のフラグを1つの変数にまとめる */
  if X02 then X02_X11 = 10 + X02 ;
  else if X11 then X02_X11 = 100 + X11 ;
  else X02_X11 = 200 ;
/* 各統計量の小数点を丸め */
  array XX (*) PerY043_Mean -- Dummy_Sum ;
  do i = 1 to dim(XX) ;
    XX(i) = round( XX(i) , 0.01 ) ;
```

*4)世帯分類別の穀類(米、パン、めん類、他の穀類)の構成比【続き】*

```
run ;
/ * 帯グラフ用データを作成 */
data work.kadai_4d ;
  set work.kadai_4b ;
  if X11 then X11 = 100 + X11 ;
run ;
proc sort data = work.kadai_4d ;
  by X11 ;
run ;
proc transpose data = work.kadai_4d out = work.kadai_4e ;
  by X11 ;
  var PerY043_Mean — PerY046_Mean ;
run ;
ods rtf file = "rtf出力先" bodytitle ;
title1 "4) 世帯分類別の穀類 (米、パン、めん類、他の穀類)
の構成比" ;
title2 "表4" ;
/ * 集計表を作成 */
proc report data = work.kadai_4c ;
  column _TYPE_ X02_X11 Dummy_Sum
        ("構成比(%)" PerY043_Mean  PerY044_Mean
                     PerY045_Mean  PerY046_Mean )
        PerAll_Mean ;
  define _TYPE_  / order order = internal descending
                   format = TYPEF. "項目名" ;
  define X02_X11 / order order = internal
                   format = Coln. "符号内容" ;
  define Dummy_Sum    / display format = comma8. "件数" ;
  define PerY043_Mean / display format = 5.1 "米" ;
  define PerY044_Mean / display format = 5.1 "パン" ;
  define PerY045_Mean / display format = 5.1 "めん類" ;
  define PerY046_Mean / display format = 5.1 "他の穀物" ;
  define PerAll_Mean  / display format = 5.1 "全体(%)" ;
run ;
/ * 帯グラフを作成 */
ods graphics / height = 8cm width = 14cm ;
title1 "4) 世帯分類別の穀類 (米、パン、めん類、他の穀類) の
構成比  帯グラフ" ;
proc sgplot data = work.kadai_4e ;
  vbar X11 / response = col1 group = _NAME_ ;
  format X11 Coln. _NAME_ $_NAME_. ;
  label X11 = "世帯分類" col1 = "構成比(%)" _NAME_ = "構成" ;
run ;
ods graphics / reset ;
ods rtf close ;
title ;
```

*5)世帯分類別のローレンツ曲線およびジニ係数*

```
* 表5 ;
/ *表出力用のフォーマットを作成 */
proc format ;
  value X11_2F 1 = "単身世帯" 2 = "夫婦のみ" 3 = "二世代"
    4 = "二世代(ひとり親)" 5 = "三世代"
    6 = "その他" 7 = "全体" ;
run ;
/ *合計算出用に Table_Type=2 とする */
data work.kadai_5a ;
  set ADS.zensho2004gijimicro (keep = X11 Y001 Weight) ;
  Table_Type = 1 ; Group= X11 ; output ;
  Table_Type = 2 ; Group = 7 ; output ;
run ;
proc sort data = work.kadai_5a ;
  by Table_Type Group Y001 ;
run ;
data work.kadai_5b ;
  set work.kadai_5a ;
  by Table_Type Group Y001 ;
/ * 各世帯分類ごとに合計する */
  if first.Group = 1 then do ;
    Sum_Weight = 0 ; Sum_Y001   = 0 ;
  end ;
  Sum_Weight + Weight ;
  Sum_Y001 + Y001 * Weight ;
/* Weight・Y001 の総合計をマクロ変数に格納 */
  if last.Group = 1 then do ;
    call     symputx(     cats("Total_Weight_",group)     ,
Sum_Weight ) ;
    call symputx( cats("Total_Y001_",group) , Sum_Y001 ) ;
  end ;
run ;
data work.kadai_5c work.kadai_5d ;
  set kadai_5b ;
  by Table_Type Group Y001 ;
  X_Weight            =            Sum_Weight            /
symgetn(cats('Total_Weight_',group)) ;
/ *X 軸の世帯累積比 */
  Y_Y001                          =          Sum_Y001        /
symgetn(cats('Total_Y001_',group)) ;/* Y 軸の所得累積比 */
  Lag_Y001 = lag1(Y_Y001) ;    / * 台形算出の上辺 */
  Dif_Weight = dif(X_Weight) ; /* 台形算出の高さ */
/ * 各世帯分類の10BS 目は三角形の面積、20BS 以降は台形の面積 */
  if first.Group = 1 then Integral = Y_Y001 * X_Weight * 0.5 ;
  else Integral + ( Lag_Y001 + Y_Y001 ) * Dif_Weight * 0.5 ;
  output work.kadai_5c ;
/ *各世帯分類の計算したジニ係数をアウトプット */
/ * 課題-1 用に世帯分類=4 のジニ係数をマクロ変数に格納 */
  if last.Group = 1 then do ;
    Gini = 2 * ( 0.5 - Integral ) ;
    if Group = 4 then call symputx("Gini4", put( round( Gini,
0.0001 ), 8.4)) ;
    output work.kadai_5d ;
  end ;
  label X_Weight = "世帯累積比" Y_Y001   = "所得累積比" ;
run ;
options nodate nonumber ;
ods rtf file = "rtf出力先" bodytitle ;
title1 "5) 世帯分類「二世代(ひとり親)」世帯の年間収入(Y001)
のローレンツ曲線" ;
ods graphics / height = 9cm width = 8cm antialiasmax = 8000 ;
proc sgplot data = kadai_5c ;
  where Group = 4 ;
  series x = X_Weight y = Y_Y001
   / lineattrs = (pattern = solid color = black) legendlabel
= "ローレンツ曲線";
  series x = X_Weight y = X_Weight
   / lineattrs = (pattern = shortdash color = black) legendlabel
= "均等分布線" ;
  inset ("ジニ係数" = "&Gini4.") ;
run ;
ods graphics / reset ;
title1 "5)  6つの世帯分類の年間収入 (Y001) のジニ係数" ;
title2 "表5" ;
proc print data = kadai_5d label noobs ;
  var Group Gini ;
  format Group X11_2F. Gini 8.2 ;
  label Group = "世帯分類" Gini  = "ジニ係数" ;
run ;
ods rtf close ;
title ;
```

# 高齢者世帯の消費動向

チーム名：タクミインフォメーションテクノロジー　チームD
〇中平 健

（株式会社タクミインフォメーションテクノロジー）

参加カテゴリー：ニューカマー（SASまたはJMPの使用歴3年未満）

Consumption trend of elderly households

Ken Nakadaira
Takumi Information Technology

## 要旨

今後も高齢化率の上昇が見込まれる我が国において、将来の消費動向は大勢を占める高齢者世帯にかかっているといっても過言ではない。そこで、「平成 16 年全国消費実態調査」に基づく新疑似ミクロデータで高齢者世帯の消費動向について分析を行った。その結果、高齢者世帯は消費意欲が旺盛であり、特に勤労者は 40 代前半と同程度の平均消費支出額であるとの結果を得た。

キーワード：疑似ミクロデータ、高齢化社会、消費動向、横断研究

## 1. はじめに

65 歳以上の高齢者が総人口に占める割合（高齢化率）は平成 16(2004)年 10 月 1 日時点において 19.5%であり[1]、平成 28(2016)年 10 月 1 日時点においては 27.3%となった[2]。平成 29 年版 高齢社会白書[2]によると、総人口が減少する中で高齢者が増加することにより高齢化率は上昇傾向にあり、2036 年に 33.3%で 3 人に 1 人となり、2065 年には 38.4%に達して、国民の約 2.6 人に 1 人が 65 歳以上の高齢者となる社会が到来すると推計している。このように、大きな割合を占める高齢者の消費動向がどのようなものであるか、「平成 16 年全国消費実態調査」に基づく新疑似ミクロデータを用いて分析する。

## 2. 分析方法

### 2.1 分析対象と定義

新疑似ミクロデータから「勤労者以外の世帯(X02 = 2)」を除く。これは、「勤労者以外の世帯」は収入が調査されておらず、実収入(Y003)が 0 円となっているためである。

以下、高齢者世帯を「世帯主の年齢が 60 歳以上(X09 >= 13)」と定義する。これは、高齢勤労者の収支状況を調べるためである。

なお、集計用乗率は用いない。

## 2. 2 分析手順

(1)世帯主の年齢区分別に可処分所得と消費支出の額を比較する。ここで、可処分所得とは

可処分所得 = 実収入(Y003) - 非消費支出(Y179)

で定義される。いわゆる手取り収入のことである[3]。

(2)世帯主の年齢区分別に消費支出の構成割合を比較する。

(3)高齢者を「勤労者世帯(X02 = 1)」と「無職世帯(X02 = 3)」に分け、(1)、(2)の比較を行う。

# 3. 分析結果と考察

## 3.1 年齢区分別の可処分所得と消費支出額

**表 1 年齢区分別の可処分所得と消費支出額**

| 世帯主年齢 | 変数 | ラベル | N | 平均 | 中央値 | 最小値 | 最大値 |
|---|---|---|---|---|---|---|---|
| 24 歳以下 | DI | 可処分所得 | 780 | 231749.94 | 213874.50 | 0 | 1862714.00 |
| | Y040 | 消費支出 | 780 | 191207.07 | 165919.00 | 33624.00 | 1081554.00 |
| 25～29 歳 | DI | 可処分所得 | 2355 | 302475.98 | 275735.00 | -26549.00 | 10037116.00 |
| | Y040 | 消費支出 | 2355 | 243804.57 | 219938.00 | 17702.00 | 1525319.00 |
| 30～34 歳 | DI | 可処分所得 | 4330 | 331261.78 | 320657.50 | -113257.00 | 1553094.00 |
| | Y040 | 消費支出 | 4330 | 263405.49 | 240228.00 | 15352.00 | 2035441.00 |
| 35～39 歳 | DI | 可処分所得 | 5410 | 372859.57 | 359619.00 | -54698.00 | 1944684.00 |
| | Y040 | 消費支出 | 5410 | 280641.22 | 260539.00 | 3157.00 | 1700795.00 |
| 40～44 歳 | DI | 可処分所得 | 5697 | 395982.11 | 379142.00 | -52497.00 | 3560369.00 |
| | Y040 | 消費支出 | 5697 | 308652.26 | 281802.00 | 12969.00 | 2991416.00 |
| 45～49 歳 | DI | 可処分所得 | 5979 | 431054.35 | 415242.00 | -64478.00 | 5201727.00 |
| | Y040 | 消費支出 | 5979 | 362257.53 | 323558.00 | 14373.00 | 4177899.00 |
| 50～54 歳 | DI | 可処分所得 | 6586 | 436556.37 | 411447.50 | -257165.00 | 2823127.00 |
| | Y040 | 消費支出 | 6586 | 370128.97 | 325904.50 | 4460.00 | 2365916.00 |
| 55～59 歳 | DI | 可処分所得 | 6306 | 421741.63 | 390974.00 | -133754.00 | 10309163.00 |
| | Y040 | 消費支出 | 6306 | 350289.68 | 304477.50 | 9216.00 | 3110781.00 |
| 60～64 歳 | DI | 可処分所得 | 5771 | 265452.81 | 221247.00 | -325611.00 | 2596767.00 |
| | Y040 | 消費支出 | 5771 | 309215.06 | 264943.00 | 5453.00 | 14970240.00 |
| 65～69 歳 | DI | 可処分所得 | 4933 | 219428.71 | 181185.00 | -128734.00 | 1938366.00 |
| | Y040 | 消費支出 | 4933 | 270092.60 | 238400.00 | 12701.00 | 1324595.00 |
| 70～74 歳 | DI | 可処分所得 | 3683 | 196111.06 | 155970.00 | -616545.00 | 1688032.00 |
| | Y040 | 消費支出 | 3683 | 243513.20 | 210901.00 | 16832.00 | 2438064.00 |
| 75 歳以上 | DI | 可処分所得 | 3382 | 169078.07 | 135042.50 | -162667.00 | 1041628.00 |
| | Y040 | 消費支出 | 3382 | 219893.26 | 187386.50 | 11181.00 | 1469101.00 |

**表 2 年齢区分別の勤労者割合**

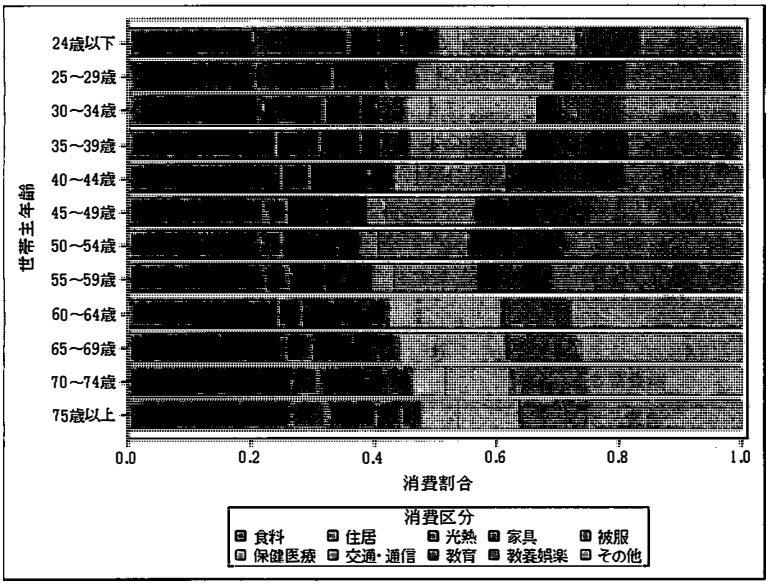
| 表：X02＊X09 | | | | | | |
|---|---|---|---|---|---|---|
| X02(世帯区分) | X09(世帯主年齢) | | | | | |
| 度数<br>列のパーセント | 非高齢者 | 60～64 歳 | 65～69 歳 | 70～74 歳 | 75 歳以上 | 合計 |
| 勤労者 | 35564<br>94.98 | 3237<br>56.09 | 1343<br>27.22 | 469<br>12.73 | 166<br>4.91 | 40779 |
| 無職 | 1879<br>5.02 | 2534<br>43.91 | 3590<br>72.78 | 3214<br>87.27 | 3216<br>95.09 | 14433 |
| 合計 | 37443 | 5771 | 4933 | 3683 | 3382 | 55212 |

ます、表 1 の平均可処分所得についてみてみると、50 代前半で最大となり、その後減少する。特に 60 歳以降大きく落ち込むが、これは 60 歳で退職する勤労者が多いため(表 2)である。

次に、表 1 の平均消費支出についてみる。50 代前半で最大となるのは可処分所得と同じであるが、60 歳以降の減少幅が小さいのは可処分所得と異なる点である。その結果、60 歳以降の年齢区分において平均消費支出金額が平均可処分所得を上回っている。ここから、金融資産の取り崩しなどで不足分を補っていることがうかがえる。

## 3.2 年齢区分別の消費支出構成割合

**図 1 年齢区分別の消費支出構成割合**




50 代前半からエンゲル係数は微増するが、教育費と交通・通信費の割合が減少することにより、教養娯楽費割合の減少が抑えられている。また、60 代から保健医療費の割合が大きく増加している。

## 3.3 高齢者の勤労者世帯と無職世帯の比較

### 3.3.1 可処分所得と消費支出額の比較

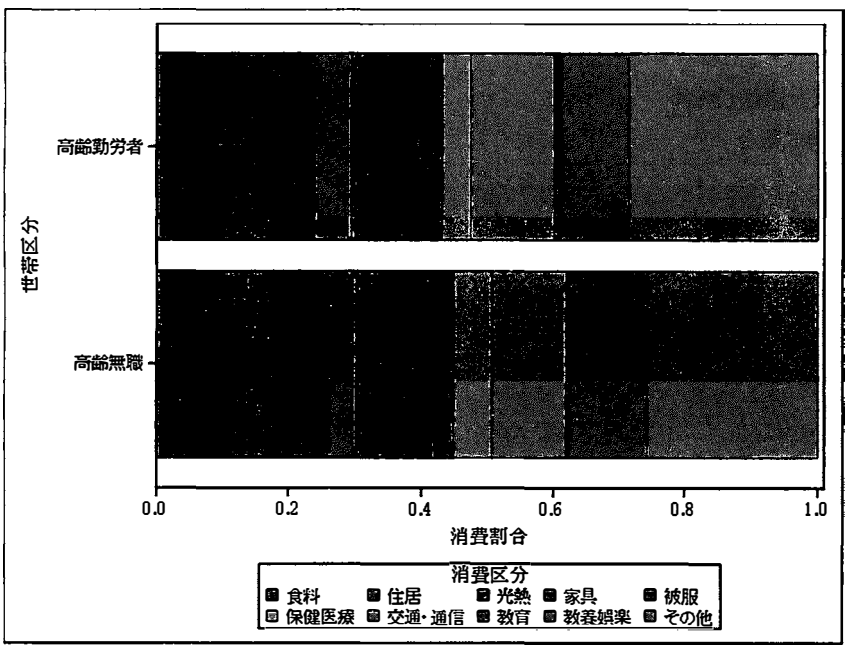
**表 3 可処分所得と消費支出額**

| 世帯区分 | 変数 | ラベル | N | 平均 | 中央値 | 最小値 | 最大値 |
|---|---|---|---|---|---|---|---|
| 高齢勤労者 | DI | 可処分所得 | 5215 | 348536.16 | 305343.00 | -48640.00 | 2596767.00 |
| | Y040 | 消費支出 | 5215 | 311135.54 | 269602.00 | 17237.00 | 5989880.00 |
| 高齢無職 | DI | 可処分所得 | 12554 | 166548.75 | 133691.50 | -616545.00 | 1688032.00 |
| | Y040 | 消費支出 | 12554 | 249706.33 | 215893.00 | 5453.00 | 14970240.00 |

まず、勤労高齢者世帯は表 1 の 30 代前半と同程度の平均可処分所得であり、平均消費支出は 40 代前半と同程度であることがわかる。次に、高齢無職世帯は勤労高齢者世帯と比較すると半分以下の平均可処分所得であるが、平均消費支出は可処分所得を上回り、金額は表 1 の 20 代後半と同程度である。

### 3.3.2 消費支出構成割合の比較

**図 2 消費支出構成割合**




消費支出の割合についてみると、高齢無職世帯の方が若干エンゲル係数は高くなっている。交通・通信費割合が少ない分教養娯楽費の割合は同程度である。高齢無職世帯の方が保健医療費割合が高くなっているが、これは消費支出額が低いためであり、金額ベースでは大きな差異は無いと推測できる。

## 4. おわりに

高齢化社会には大きな問題がある。まず 1 つは財源の問題である。平成 28 年度一般会計歳出のうち、一般歳出（基礎的財政収支対象経費から地方交付税交付金等を除いたもの）における社会保障関係費の割合は 55.0%にもなる。また、一般会計歳入のうち、4 割弱は将来世代の負担となる借金に依存している[4]。

もう 1 つは年金の世代間格差である。これは、公的年金において給付負担率(年金給付総額 / 保険料負担総額)が世代ごとに大きく異なるものであり、「不公平である」と考える若年世代も少なくない[5]。本稿は横断的研究であり、元データが時系列データではないため触れることができなかったが、これは今後の課題とする。

解決の一助となるかもしれない提言が日本老年医学会[6]からあった。これは、近年の平均寿命の延伸と「若返り」現象を考慮し、75 歳から 89 歳を高齢者、90 歳以上を超高齢者とする提案である。消費意欲が旺盛である高齢勤労者が増えれば、税収と消費支出の増加が見込める。

我が国は世界で最も高い高齢化率であり、世界に先駆け高齢化社会におけるビジネスモデルを構築する良い機会でもある。また、生産年齢人口の減少に対しては、AI・ロボットを活用し、労働生産性を向上させることも可能である。

問題は山積しているが、この社会の大きな変化をチャンスと捉え、立ち向かうことが肝要である。

## 参考文献

[1]平成 17 年版 高齢社会白書
< http://www8.cao.go.jp/kourei/whitepaper/w-2005/zenbun/17index.html>
[2]平成 29 年版 高齢社会白書
< http://www8.cao.go.jp/kourei/whitepaper/w-2017/zenbun/29pdf_index.html>
[3]平成 26 年 全国消費実態調査 用語の解説
<http://www.stat.go.jp/data/zensho/2014/kaisetsu.htm>
[4]財務省 平成 28 年度一般会計予算(平成 28 年 3 月 29 日成立)の概要
<https://www.mof.go.jp/tax_policy/summary/condition/002.htm>
[5]みずほ総合研究所 年金の世代間の給付と負担の差
<https://www.mizuho-ri.co.jp/publication/research/pdf/insight/pl151120.pdf>
[6]日本老年医学会 高齢者の定義と区分に関する、日本老年学会・日本老年医学会 高齢者に関する定義検討ワーキンググループからの提言(概要)
<https://www.jpn-geriat-soc.or.jp/proposal/pdf/definition_01.pdf>

## 付録 1. 規定課題用 SAS プログラム

```
* 規定課題共通 ;
%let Dir = .¥社内研修¥Temp¥SAS_09_データ ;
libname Data "&Dir.¥01_Raw" ;

proc format ;
  value X02_01F  1 = "勤労者"
                 2 = "勤労者以外"
                 3 = "無職" ;

  value X02_02F  1 = "勤労者"
                 2 = "勤労者以外"
                 3 = "無職世帯" ;

  value X08F  1 = "男"
              2 = "女" ;

  value X09F  5 = "24歳以下"
              6 = "25～29歳"
              7 = "30～34歳"
              8 = "35～39歳"
              9 = "40～44歳"
              10  = "45～49歳"
              11  = "50～54歳"
              12  = "55～59歳"
              13  = "60～64歳"
              14  = "65～69歳"
              15  = "70～74歳"
              16  = "75歳以上" ;

  value X11F  1 = "単身世帯"
              2 = "夫婦のみ"
              3 = "二世代"
              4 = "二世代(ひとり親)"
              5 = "三世代"
              6 = "その他"
              7 = "全体";

  value $StatF  Val_max = "最大値"
                Val_q3 = "第3四分位数"
                Val_med = "中央値"
                Val_q1 = "第1四分位数"
                Val_min = "最小値"
                Val_mean = "平均値"
                Weight_sum = "件数" ;

  value ItemF  1 = "世帯区分"
               2 = "世帯分類"
               3 = "全体" ;

  value XF  1 = "勤労者世帯"
            2 = "勤労者以外の世帯"
            3 = "無職世帯"
            101 = "単身世帯"
            102 = "夫婦のみ"
            103 = "二世代"
            104 = "二世代(ひとり親)"
            105 = "三世代"
            106 = "その他"
            200 = " " ;

  value GrainF  0 = "件数"
                1 = "米"
                2 = "パン"
                3 = "めん類"
                4 = "他の 穀類"
                5 = '全体(%)' ;
run;

*10万人比集計用のWEIGHT算出;
data work.WEIGHT (keep = weight1 weight2) ;
  set DATA.Zensho2004gijimicro (keep = weight) end
= EOF ;

  Weight_T + weight ;

  if EOF = 0 then delete ;
  else if EOF = 1 then do ;
    weight1 = 1 ;
    weight2 = 10 ** 5 / Weight_T ;
  end ;
run ;

* 規定課題01 ;
%let PgName = 規定課題_1 ;
filename Out "&Dir.¥04_OUTPUT¥&PgName..rtf" ;
options nofmterr mprint nodate pageno=1 ;
title1 "規定課題 1)年齢階層・世帯区分別 クロス表
" ;

data work.EX01_01 ;
  set DATA.Zensho2004gijimicro (keep = X02 X09
Weight);
  label X02 = "世帯区分" X09 = "世帯主の年齢" ;
  format X02 X02_01F. X09 X09F. ;
  if _N_ = 1 then set WORK.WEIGHT ;
run ;

ods listing close ;
ods rtf file = out columns = 2;
proc tabulate data = WORK.EX01_01 format =
comma7. ;
  class X02 X09 ;
  table (X09="" all="合計"), (X02 * n="") (all="合
計" * n="")
    / misstext = "0" box = "世帯主の年齢"  rts =
8;
run ;

proc tabulate data = WORK.EX01_01 format =
comma7. ;
  class X02 X09 ;
  var Weight ;
  table (X09="" all="合計"), (X02 * weight="" *
sum="")  (all="合計" * weight="" * sum="")
    / misstext = "0" box = "世帯主の年齢"  rts =
8 ;
  weight weight2 ;
run ;
ods rtf close ;
ods listing ;

* 規定課題02 ;
%let PgName = 規定課題_2 ;
filename Out "&Dir.¥04_OUTPUT¥&PgName..rtf" ;
options nofmterr mprint nodate pageno=1
orientation = landscape ;
title1 "規定課題 2) 世帯分類別・世帯区分別・世帯主
性別クロス表" ;

data WORK.EX02_01 (keep = X02 X08 X11 Weight
Weight1 Weight2) ;
  set DATA.Zensho2004gijimicro ;

  if _N_ = 1 then set WORK.Weight ;
  format X02 X02_02F. X08 X08F. X11 X11F. ;
run ;

ods listing close ;
ods rtf file = out ;
proc tabulate data = WORK.EX02_01 format =
comma7.;
  class X02 X08 X11 ;
  var Weight ;
  table (X11="" all="合計" ) * weight="" *
sum="" , X02="" * (X08="" all="計" ) all="合計"  /
misstext="0" box="世帯分類" ;
  weight weight2 ;
run ;
ods rtf close ;
ods listing ;

* 規定課題03 ;
%let PgName = 規定課題_3 ;
filename Out "&Dir.¥04_OUTPUT¥&PgName..rtf" ;
title "規定課題 3) パン支出金額についての各種の統
計グラフ" ;

data WORK.EX03_01 (keep = No Y044 Weight Weight1
```

```
Weight2 Count Log Val ) ;
  set DATA.Zensho2004gijimicro ;

  Count = 1 ;
  if _N_ = 1 then set WORK.Weight ;

  Log = 0 ; Val = Y044 ; output ;
  Log = 1 ; Val = log10(Y044 + 1) ; output ;
run ;

proc means data = WORK.EX03_01 noprint ;
  var Val Weight ;
  class log ;
  weight Weight2 ;
  output out = WORK.EX03_02 (drop = _TYPE_ _FREQ_
where = (log ^= . ))
    max = Val_max
    q3 = Val_q3
    median = Val_med
    q1 = Val_q1
    min = Val_min
    mean = Val_mean
    sum = Val_sum Weight_sum ;
run ;

proc transpose data = work.EX03_02 out =
work.EX03_03 (rename = (_NAME_ = STAT)
  where =(STAT ^= "Log" and STAT ^= "Val_sum")) ;
run ;

ods listing close ;
ods rtf file = out ;
title "規定課題 3)パン支出金額についての各種の統計
グラフ(1)" ;
proc print data = EX03_03 noobs label split= "#" ;
  label STAT = "統計量" COL1 = "パンの支出#金額
(円)" COL2 = "1を加えた金額の#常用対数" ;
  format STAT $STATf. COL1 comma7.0 COL2
comma6.3 ;
run ;

proc sgplot data = WORK.EX03_01 ;
  hbox Val ;
  where log = 0 ;
run ;

title "規定課題 3)パン支出金額についての各種の統計
グラフ(3)" ;
proc sgplot data = WORK.EX03_01 ;
  histogram Val ;
  where log = 1 ;
run ;
ods rtf close ;
ods listing ;

* 規定課題4 ;
%let PgName = 規定課題_4 ;
filename Out "&Dir.¥04_OUTPUT¥&PgName..rtf" ;
options nofmterr mprint nodate nonumber
orientation=portrait ls=155 ps=53;

proc summary data = Data.zensho2004gijimicro ;
  var Weight ;
  output out = Weight( keep = TotalWeight ) sum =
TotalWeight ;
run ;

data work.EX04_01 ;
  set Data.zensho2004gijimicro ;
  retain Multiplier;

  if _N_ = 1 then do;
    set Weight;
    Multiplier = 1e5 / TotalWeight;
  end;

  CountDummy = 1;
  RedefinedWeight = Weight * Multiplier;

  array Variables[*] Y043 Y044 Y045 Y046 Y042;

  do I=1 to dim( Variables );

    Category = I;

    if Y042 ^= 0 then
      PR = 100 * Variables[I] / Y042;

    * 世帯区分 ;
    Item = 1 ; X = X02 ; output ;

    * 世帯分類 ;
    Item = 2 ; X = X11 + 100 ; output ;

    * 全体 ;
    Item = 3 ; X = 200 ; output ;
  end ;

  keep Item X Category PR CountDummy
RedefinedWeight ;
  format _all_ ;
run ;

proc means data = work.EX04_01 noprint nway ;
  class Item X Category ;
  var PR ;
  weight RedefinedWeight ;
  output out = work.EX04_02
    mean=MeanPR ;
run ;

proc means data = work.EX04_01 noprint nway ;
  where Category = 5 ;
  class Item X ;
  var CountDummy ;
  weight RedefinedWeight ;
  output out = work.EX04_03
    sum=SumCnt ;
run;

data work.EX04_04 ;
  set
    work.EX04_02( in=In1 )
    work.EX04_03( in=In2 ) ;

  length ValueC XC $100 ;

  if In1 then do;
    ValueC = put( round( MeanPR, 1e-1 ), 12.1 -
L ) ;
  end;
  else if In2 then do ;
    Category = 0 ;
    ValueC = put( round( SumCnt, 1 ), comma12. -
L ) ;
  end;

  IdLabel = put( Category, GrainF. ) ;

proc sort ;
  by Item X Category ;
run ;

proc transpose data = work.EX04_04 out =
work.EX04_05 prefix=C ;
  by Item X ;
  id Category ;
  idlabel IdLabel ;
  var ValueC ;
run ;

ods listing close ;
ods rtf file = out bodytitle style=JournalCustom
nogtitle startpage=no ;
ods escapechar="^" ;

title "4)世帯区分別および世帯分類別の穀類の支出金
額についての構成比" ;

proc report data = work.EX04_05 nowd split=" " ;
  column
    Item X
    C0 ( '構成比(%)^R/RTF"¥brdrb¥brdrs¥brdrw1"' C1
- C4 ) C5 ;
  define Item / "項目名" order order=internal
```

```
style(column)={just=left cellwidth=2cm} ;
  define X / "符号内容" order  order=internal
style(column)={just=left cellwidth=3.5cm} ;
  define C0 - C5 /
style(column)={just=right cellwidth=1.5cm} ;

  compute before ITEM / style = { font_size = 0.1
cm } ;
  endcomp ;

  compute after / style = { font_size = 0.1 cm } ;
  endcomp ;

  format Item ItemF. X XF. ;
run;

ods graphics / height=5cm width=15cm ;
proc sgplot data = work.EX04_02 ;
  where Item = 2 and Category ^= 5 ;
  hbar X / response=MeanPR group=Category ;
  format X XF. Category Grain. ;
  label
    X = "世帯分類"
    MeanPR = "構成比"
    Category = "穀類分類" ;
run ;
ods graphics / reset ;
ods rtf close startpage = yes ;
ods listing ;

* 規定課題05 ;
%let PgName = 規定課題_5 ;
filename Out "&Dir.¥04_OUTPUT¥&PgName..rtf" ;
options nofmterr mprint nodate pageno = 1 ;
title1 "規定課題 5)世帯分類別のローレンツ曲線およ
びジニ係数(乗率使用)" ;

proc sort data = DATA.Zensho2004gijimicro (keep =
weight X11 Y001)
    out = work.EX05_01 ;
  by Y001 ;
run ;

%macro Lorenz1 ;
  %do i = 1 %to 7 ;
    data work.EX05_02 work.CUM
                      (keep = Weight_C Y001_C
                       rename = (Weight_C =
Weight_T Y001_C = Y001_T )) ;
      set work.EX05_01
        %if &i ^= 7 %then (where = (X11 = &i)) ;
        end = EOF ;

      Weight_C + weight ;
      Y001_C + (Y001 * Weight) ;

      output work.EX05_02 ;

      if EOF = 1 then do ;
        output work.CUM ;
      end ;
    run ;

    data work.EX05_03 work.Gini_&i (keep = X11
GiniC) ;
      set work.EX05_02 end = EOF ;
        if _N_ = 1 then set work.CUM ;

      retain Integral Y001_PR Weight_PR 0 ;

      Weight_R = Weight_C / Weight_T ;
      Y001_R = Y001_C / Y001_T ;

      Plot_E = Weight_R ;
      Integral + ((Y001_PR + Y001_R) * (Weight_R -
Weight_PR) * 0.5) ;

      Y001_PR = Y001_R ;
      Weight_PR = Weight_R ;

      output work.EX05_03 ;

      if EOF = 1 then do ;
        Gini = 2 * (0.5 - Integral) ;
        GiniC = put(round(Gini, 1e-2), 7.2 -L) ;
        %if &i = 7 %then X11 = 7 ; ;
        output work.Gini_&i ;
      end ;
    run ;

    %if &i = 4 %then %do ;
      ods listing close ;
      ods rtf file = out ;
      proc sgplot data = work.EX05_03 ;
        series x = Weight_R y = Y001_R /
        lineattrs = ( pattern=solid color = black)
        name="LORENZ" ;
        series x = Weight_R y = Plot_E /
        lineattrs = ( pattern=shortdash color =
black)
        name="EQUAL" ;
        keylegend "LORENZ" "EQUAL" ;
        xaxis min=0 max=1 offsetmin=0 offsetmax=0
label="世帯累積比" ;
        yaxis min=0 max=1 offsetmin=0 offsetmax=0
label="所得累積比" ;
      run ;
    %end ;
  %end ;
%mend ;

%macro Lorenz2 ;
  data work.GiniC ;
      set
      %do i = 1 %to 7 ;
        work.Gini_&i
      %end ;
      ;
  run ;
%mend ;

%Lorenz1 ;
%Lorenz2 ;

proc print data = work.GiniC noobs label ;
  label X11 = "世帯分類" GiniC = "ジニ係数" ;
  format X11 X11F. ;
run ;
ods rtf close ;
ods listing ;
```

## 付録 2.　規定課題図表

### 1)　年齢階層別・世帯区分別 クロス表

| 世帯主の年齢 | 世帯区分 | | | 合計 |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | |
| 24 歳以下 | 765 | 18 | 15 | 798 |
| 25～29 歳 | 2,289 | 157 | 66 | 2,512 |
| 30～34 歳 | 4,223 | 511 | 107 | 4,841 |
| 35～39 歳 | 5,217 | 728 | 193 | 6,138 |
| 40～44 歳 | 5,492 | 1,076 | 205 | 6,773 |
| 45～49 歳 | 5,702 | 1,305 | 277 | 7,284 |
| 50～54 歳 | 6,194 | 1,797 | 392 | 8,383 |
| 55～59 歳 | 5,682 | 2,221 | 624 | 8,527 |
| 60～64 歳 | 3,237 | 2,226 | 2,534 | 7,997 |
| 65～69 歳 | 1,343 | 1,743 | 3,590 | 6,676 |
| 70～74 歳 | 469 | 1,197 | 3,214 | 4,880 |
| 75 歳以上 | 166 | 940 | 3,216 | 4,322 |
| 合計 | 40,779 | 13,919 | 14,433 | 69,131 |

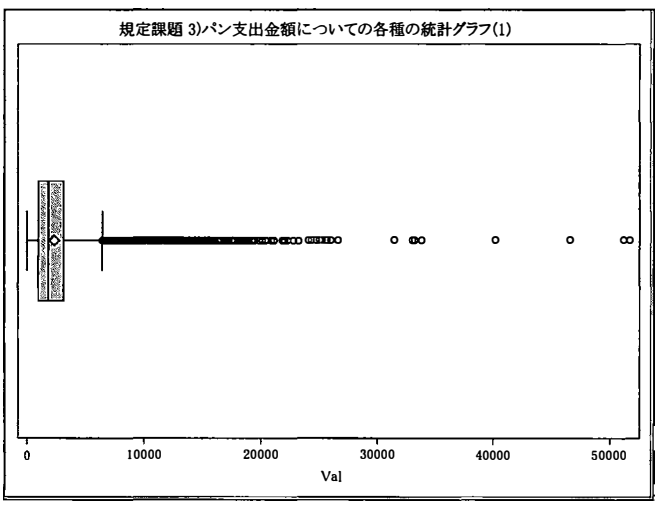
| 世帯主の年齢 | 世帯区分 | | | 合計 |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | |
| 24歳以下 | 1,729 | 10 | 21 | 1,760 |
| 25～29歳 | 4,314 | 151 | 57 | 4,522 |
| 30～34歳 | 6,441 | 480 | 87 | 7,008 |
| 35～39歳 | 7,420 | 824 | 153 | 8,396 |
| 40～44歳 | 7,611 | 1,145 | 247 | 9,004 |
| 45～49歳 | 7,739 | 1,479 | 244 | 9,462 |
| 50～54歳 | 8,072 | 2,085 | 508 | 10,665 |
| 55～59歳 | 7,697 | 2,725 | 893 | 11,315 |
| 60～64歳 | 4,208 | 2,785 | 4,160 | 11,153 |
| 65～69歳 | 1,487 | 2,213 | 6,374 | 10,074 |
| 70～74歳 | 459 | 1,652 | 6,124 | 8,235 |
| 75歳以上 | 207 | 1,325 | 6,873 | 8,406 |
| 合計 | 57,386 | 16,874 | 25,741 | 100,000 |

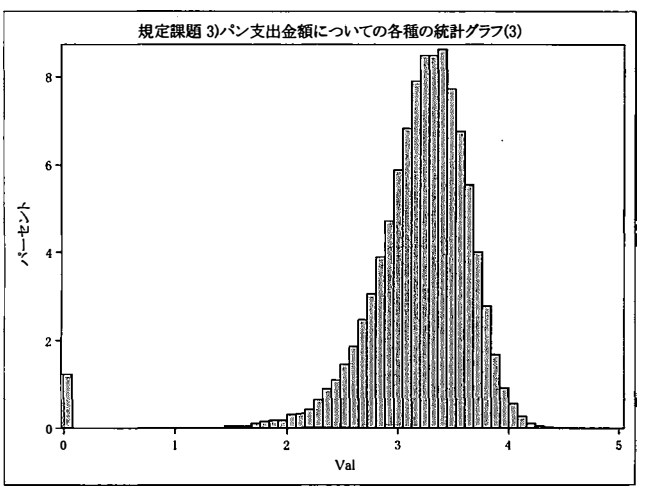
### 2)　世帯分類別・世帯区分別 世帯主の性別 3 重クロス表

| 世帯分類 | 勤労者 | | | 勤労者以外 | | | 無職世帯 | | | 合計 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 男 | 女 | 計 | 男 | 女 | 計 | 男 | 女 | 計 | |
| 単身世帯 | 6,503 | 4,207 | 10,710 | 1,030 | 867 | 1,897 | 2,463 | 5,688 | 8,151 | 20,758 |
| 夫婦のみ | 8,517 | 324 | 8,841 | 4,293 | 81 | 4,375 | 11,026 | 45 | 11,071 | 24,287 |
| 二世代 | 27,267 | 378 | 27,645 | 6,190 | 51 | 6,242 | 3,264 | 32 | 3,295 | 37,182 |
| 二世代(ひとり親) | 1,543 | 2,421 | 3,963 | 934 | 430 | 1,364 | 671 | 1,303 | 1,974 | 7,302 |
| 三世代 | 4,956 | 384 | 5,340 | 2,390 | 118 | 2,508 | 638 | 160 | 798 | 8,646 |
| その他 | 654 | 234 | 887 | 413 | 75 | 488 | 278 | 174 | 452 | 1,826 |
| 合計 | 49,439 | 7,947 | 57,386 | 15,250 | 1,623 | 16,874 | 18,340 | 7,401 | 25,741 | 100,000 |

3)　パン支出金額についての各種の統計グラフ

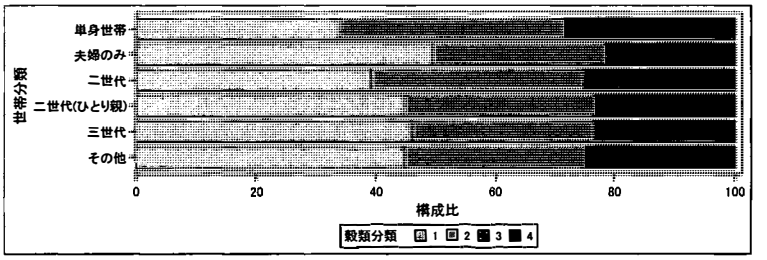
| 統計量 | パンの支出金額(円) | 1を加えた金額の常用対数 |
|---|---|---|
| 最大値 | 51,776 | 4.714 |
| 第3四分位数 | 3,135 | 3.496 |
| 中央値 | 1,811 | 3.258 |
| 第1四分位数 | 962 | 2.984 |
| 最小値 | 0 | 0.000 |
| 平均値 | 2,368 | 3.184 |
| 件数 | 100,000 | 100000 |



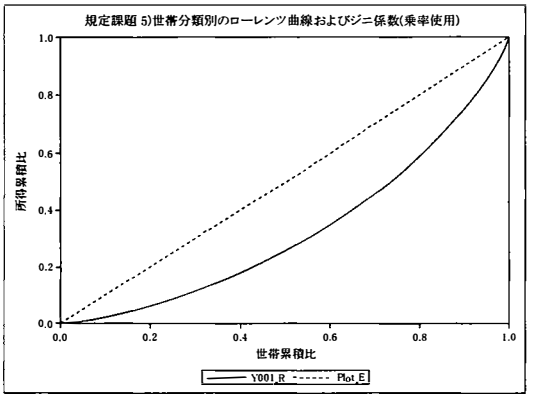



4)　世帯区分別および世帯分類別の穀類の支出金額についての構成比

| 項目名 | 符号内容 | 件数 | 構成比(%) 米 | パン | めん類 | 他の穀類 | 全体(%) |
|---|---|---|---|---|---|---|---|
| 世帯区分 | 勤労者世帯 | 57,386 | 36.0 | 37.6 | 22.1 | 4.3 | 100.0 |
| | 勤労者以外の世帯 | 16,874 | 44.9 | 31.0 | 19.4 | 4.7 | 100.0 |
| | 無職世帯 | 25,741 | 52.1 | 26.7 | 15.8 | 5.4 | 100.0 |
| 世帯分類 | 単身世帯 | 20,758 | 33.8 | 37.7 | 23.0 | 5.5 | 100.0 |
| | 夫婦のみ | 24,287 | 49.5 | 28.9 | 16.8 | 4.8 | 100.0 |
| | 二世代 | 37,182 | 39.0 | 35.9 | 20.8 | 4.3 | 100.0 |
| | 二世代(ひとり親) | 7,302 | 44.9 | 31.6 | 19.0 | 4.5 | 100.0 |
| | 三世代 | 8,646 | 46.0 | 30.5 | 19.7 | 3.8 | 100.0 |
| | その他 | 1,826 | 44.8 | 30.5 | 20.4 | 4.3 | 100.0 |
| 全体 | | 100,000 | 41.6 | 33.7 | 20.0 | 4.6 | 100.0 |




5)　世帯分類別のローレンツ曲線およびジニ係数




| 世帯分類 | ジニ係数 |
|---|---|
| 単身世帯 | 0.34 |
| 夫婦のみ | 0.31 |
| 二世代 | 0.26 |
| 二世代(ひとり親) | 0.35 |
| 三世代 | 0.27 |
| その他 | 0.33 |
| 全体 | 0.33 |

# ストックの視点を含んだ世帯の消費行動の特徴抽出

参加カテゴリー：ニューカマー
チーム名：野毛坂 123
メンバー：山川雄也、松本光右、三ツ井誠
株式会社　浜銀総合研究所 情報戦略コンサルティング部

## 要旨

投資行動の有無を基準とし、消費行動にどのような違いがあるかについて、「疑似ミクロデータ（平成 16 年全国消費実態調査）」を用いて調査・分析を行った。
その結果、
1.　投資無の世帯のほうが、投資有の世帯に比べて生活必需消費率が大きくなる
2.　収入が大きくなると、投資有無による生活必需消費率の差は小さくなる
という結論に至った。

キーワード：　疑似ミクロデータ、全国消費実態調査、投資行動、ストック、フロー

## 1　背景と目的

世界的に、貧困層と富裕層との格差については、問題視されて久しい。国家内での格差を計る指標にジニ係数がある。日本におけるジニ係数の推移は、年金などの国家の再分配機能を考慮した値は、横ばい傾向となっている一方、再分配機能を考慮しない、所得のみを考慮した値（"当初所得"による値）については、上昇傾向が続いている[1]。

通常、消費行動について考察する場合は、収入額・支出額といった、フローに着目することが多いが、今回われわれは、資本（“ストック”）の視点を取り入れ、資本の有無が消費行動にどのような違いをもたらすかの検証を目的として調査・分析を行う。

なお、分析で用いる疑似ミクロデータのもととなる、平成 16 年度全国消費実態調査は、ストックについての情報が乏しいため、その代替として投資行動の有無をストックの有無と置き換え、消費行動の特徴抽出を行う。

## 2　分析概要

本章では、背景と目的を踏まえ、世帯のストック有無が世帯の消費行動にどのような影響を与えるのかについて、擬似ミクロデータを用いた分析の概要を述べる。本分析で用いる疑似ミクロデータは、全国消費実態調査に基づく世帯単位の世帯のフローに関するデータであるため、世帯のストックに関する情報を得ることは難しい。そこで、以下に定義する投資行動の有無フラグを生成し、これを世帯のストックに関する情報の代替として扱うことで分析を行う。

**投資有無の定義**

- 投資フラグ＝1（投資行動あり）:「Y016：財産購入」「Y029：財産収入」「Y195：有価証券売却」「Y201：有価証券購入」のいずれかの金額が 0 円超
- 投資フラグ＝0（投資行動なし）:「Y016：財産購入」「Y029：財産収入」「Y195：有価証券売却」「Y201：有価証券購入」の全ての金額が 0 円

以下では、上記の投資有無が世帯の消費行動にどのような影響を与えているかを見ていく。具体的には、年間収入に対する生活必需消費率、投資有無によってどのように違うのかを分析する。ここで、生活必需消費率は、10 大費目の消費額合計に対する生活必需費目の消費額合計の割合と定義し、さらに 10 大費目および生活必需費目は、それぞれ以下で説明する。

**10 大費目および生活必需費目の説明**

- 10 大費目
  「食料」「住居」「光熱・水道」「家具・家事用品」「被服及び履物」「保険医療」「交通・通信」「教育」「教養娯楽」「その他の消費支出」
- 生活必需費目
  「食料」「住居」「光熱・水道」

ついで、分析対象について述べる。「X02：世帯区分」の「勤労者以外の世帯」および「無職世帯」は、投資の有無にかかわらず「勤労者世帯」の収入や消費行動と差異があると考えられるため、今回の分析では「勤労者世帯」のみを分析対象とする。

以下では、具体的な分析手順について説明する。まず擬似ミクロデータの中から、「X02：世帯区分」の「勤労者世帯」に該当するデータに絞り込み、先に述べた投資有無の定義に従ってフラグを立てる。次に、投資有無でデータを分割し、さらにそれぞれのデータを世帯の年間収入（変数名：Y001）に応じてカテゴリーを割り振る。世帯の年間収入（以下、収入と呼ぶ。）に関するカテゴリーの設定基準については表 1 の通りである。その後、収入カテゴリー、および投資有無別に生活必需消費率の平均を算出し、2 群の平均の検定を行う。これにより、世帯収入の変化に対する生活必需消費率の変化が、投資有無によってどのように違うのかを見る。

| カテゴリーNo. | カテゴリーの内容 | 投資有の件数 | 投資無の件数 |
|---|---|---|---|
| 1 | 収入が０～４００万円の世帯 | ４７１件 | ７６１２件 |
| 2 | 収入が４００～５００万円の世帯 | ４７０件 | ４８５６件 |
| 3 | 収入が５００～６００万円の世帯 | ６４６件 | ４８４６件 |
| 4 | 収入が６００～７００万円の世帯 | ７１４件 | ４１７５件 |
| 5 | 収入が７００～８００万円の世帯 | ６６５件 | ３６０４件 |
| 6 | 収入が８００～９００万円の世帯 | ５６９件 | ２９０４件 |
| 7 | 収入が９００～１０００万円の世帯 | ４６１件 | ２０４０件 |
| 8 | 収入が１０００万円以上の世帯 | １４４５件 | ５３０１件 |

表１：各カテゴリーの説明と件数について

## 3 分析結果

本章では、2章で述べた分析結果を提示する。まず、表1はカテゴリー別に投資有無ごとの平均生活必需消費率を算出した結果である。この結果、年収 700 万円未満のカテゴリーは投資有無で生活必需消費率に差があると考えられる。

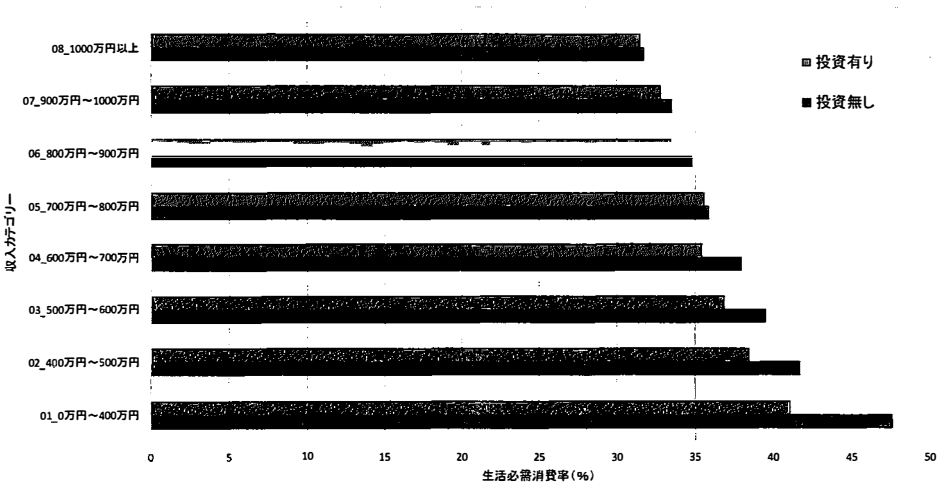



図1：各カテゴリーの投資フラグ有および無のデータにおける生活必需消費率の平均

続いて、収入カテゴリー別に、投資有無で生活必需消費率の平均が等しいかどうかについて、t検定により評価した結果を以下に示す。

| カテゴリー1(収入0～400万円)の結果 | | | カテゴリー2(収入400～500万円)の結果 | | |
|---|---|---|---|---|---|
| 投資フラグ | 平均 | 標準偏差 | 投資フラグ | 平均 | 標準偏差 |
| 有 | 41.0632 | 18.8244 | 有 | 38.4434 | 15.4383 |
| 無 | 47.5608 | 19.5521 | 無 | 41.7048 | 17.628 |
| | | | | | |
| 手法 | 分散 | p値 | 手法 | 分散 | p値 |
| Pooled | Equal | <.0001 | Pooled | Equal | <.0001 |
| Satterthwaite | Unequal | <.0001 | Satterthwaite | Unequal | <.0001 |
| | | | | | |
| 等分散性 | | | 等分散性 | | |
| 手法 | p値 | | 手法 | p値 | |
| Folded F | 0.2725 | | Folded F | 0.0002 | |

表2：カテゴリー1および2の検定結果

| カテゴリー3（収入500～600万円）の結果 | | | カテゴリー4（収入600～700万円）の結果 | | |
|---|---|---|---|---|---|
| 投資フラグ | 平均 | 標準偏差 | 投資フラグ | 平均 | 標準偏差 |
| 有 | 36.8873 | 14.9217 | 有 | 35.5075 | 15.1661 |
| 無 | 39.503 | 16.7877 | 無 | 37.9531 | 15.2249 |
| | | | | | |
| 手法 | 分散 | p値 | 手法 | 分散 | p値 |
| Pooled | Equal | <.0001 | Pooled | Equal | <.0001 |
| Satterthwaite | Unequal | <.0001 | Satterthwaite | Unequal | <.0001 |
| | | | | | |
| 等分散性 | | | 等分散性 | | |
| 手法 | p値 | | 手法 | p値 | |
| Folded F | 0.0001 | | Folded F | 0.903 | |

表３：カテゴリー３および４の検定結果

| カテゴリー5（収入700～800万円）の結果 | | | カテゴリー6（800～900万円）の結果 | | |
|---|---|---|---|---|---|
| 投資フラグ | 平均 | 標準偏差 | 投資フラグ | 平均 | 標準偏差 |
| 有 | 35.5187 | 16.2184 | 有 | 33.4348 | 14.5438 |
| 無 | 35.8307 | 14.5546 | 無 | 34.8154 | 14.2504 |
| | | | | | |
| 手法 | 分散 | p値 | 手法 | 分散 | p値 |
| Pooled | Equal | 0.5371 | Pooled | Equal | 0.0122 |
| Satterthwaite | Unequal | 0.5659 | Satterthwaite | Unequal | 0.0136 |
| | | | | | |
| 等分散性 | | | 等分散性 | | |
| 手法 | p値 | | 手法 | p値 | |
| Folded F | 0.0002 | | Folded F | 0.5188 | |

表４：カテゴリー５および６の検定結果

| カテゴリー7（900～1000万円）の結果 | | | カテゴリー8（収入1000万円以上）の結果 | | |
|---|---|---|---|---|---|
| 投資フラグ | 平均 | 標準偏差 | 投資フラグ | 平均 | 標準偏差 |
| 有 | 32.9353 | 13.2821 | 有 | 31.4879 | 13.5373 |
| 無 | 33.5424 | 14.1874 | 無 | 31.749 | 14.0381 |
| | | | | | |
| 手法 | 分散 | p値 | 手法 | 分散 | p値 |
| Pooled | Equal | 0.319 | Pooled | Equal | 0.4634 |
| Satterthwaite | Unequal | 0.2994 | Satterthwaite | Unequal | 0.4542 |
| | | | | | |
| 等分散性 | | | 等分散性 | | |
| 手法 | p値 | | 手法 | p値 | |
| Folded F | 0.0775 | | Folded F | 0.0877 | |

表５：カテゴリー７および８の検定結果

なお、上記の検定は、カテゴリー別の投資有無のデータがそれぞれ正規分布の仮定を満たしているかQ-Qプロットによって確認している。

さらに、収入カテゴリーごとに生活消費率の分布の乖離具合をカルバック・ライブラー情報量で確認する。２つの正規分布間のカルバック・ライブラー情報量は以下で表される。表６の結果より、年収０～４００万円のカテゴリー１の一番カルバック・ライブラー情報量が大きく、投資有無の影響が強い結果となった。

【カルバック・ライブラー情報量の定義】

$$\mathrm{KL}(\text{投資有} \mid \text{投資無}) = \frac{1}{2}\left(\frac{\sigma_{有}^2}{\sigma_{無}^2} + \frac{(\mu_{有} - \mu_{無})^2}{\sigma_{無}^2} - log\frac{\sigma_{有}^2}{\sigma_{無}^2} - 1\right)$$

$\mu_{有}$：投資有りの生活必需消費率の平均、$\sigma_{有}^2$：投資有りの生活必需消費率の分散（投資無しも同様）

| カテゴリー | KL情報量 |
|---|---|
| カテゴリー1（収入0～400万円） | 0.09456 |
| カテゴリー2（収入400～500万円） | 0.042964 |
| カテゴリー3（収入500～600万円） | 0.065704 |
| カテゴリー4（収入600～700万円） | 0.040486 |
| カテゴリー5（収入700～800万円） | 0.0070171 |
| カテゴリー6（収入800～900万円） | 0.017734 |
| カテゴリー7（収入900～1000万円） | 0.0094897 |
| カテゴリー8（収入1000万円以上） | 0.0009073 |

表６：ＫＬ情報量

## ４ 考察

図１からすべてのカテゴリーで投資有の生活必需消費率の平均は、投資無の平均と比べて低いという結果となった。しかし、年収が高くなるほど、投資の有無による生活必需消費率の違いに差がみられなくなる。一方、年収の低くなるほど、投資有無による生活必需消費率の差異が顕著に表れる。特に、年収７００万円未満で、投資フラグの影響が出ているように見受けられる。なお、統計的検定でも同様の結果が確認できた。また、年収カテゴリーごとに投資有無別の生活必需消費率の分布の乖離度をカルバック・ライブラー情報量で定量化し、年収が一番低いカテゴリーである０～４００万円で一番乖離しており、年収が低いところでより投資有無が影響を与えていることが分かった。

以上より、年収だけでなく、ストックを合わせてみると、低年収世帯のより苦しい生活状況が見えてくる結果になった。

## ５ 参考文献

[1] 平成 26 年 所得再分配調査報告書、厚生労働省政策統括官（総合政策担当）

## 付録：規定課題図表

### 1）年齢階層別・世帯区分別クロス表

集計の前に、用いた疑似ミクロデータの X02（世帯区分）、X08（性別）、X09（世帯主の年齢）、X11（世帯分類）、の各項目に対してフォーマットを設定する。

表 1 左側については、所定の様式の通りとなるよう、tabulate プロシジャを用いて集計を行った。表 1 右側についても、tabulate プロシジャを用いて集計を行ったが、10 万分比での集計を行うため、

"(weight / 全 weight の和) * 100,000"

を疑似ミクロデータの各オブザベーションに対して計算し、10 万分比集計用のウエイトとして使用した。結果は下記の通りである。

| | 世帯区分 | | | 合計 |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | |
| | N | N | N | N |
| 世帯主の年齢 | | | | |
| 01_24歳以下 | 765 | 18 | 15 | 798 |
| 02_25～29歳 | 2,289 | 157 | 66 | 2,512 |
| 03_30～34歳 | 4,223 | 511 | 107 | 4,841 |
| 04_35～39歳 | 5,217 | 728 | 193 | 6,138 |
| 05_40～44歳 | 5,492 | 1,076 | 205 | 6,773 |
| 06_45～49歳 | 5,702 | 1,305 | 277 | 7,284 |
| 07_50～54歳 | 6,194 | 1,797 | 392 | 8,383 |
| 08_55～59歳 | 5,682 | 2,221 | 624 | 8,527 |
| 05_60～64歳 | 3,237 | 2,226 | 2,534 | 7,997 |
| 06_65～69歳 | 1,343 | 1,743 | 3,590 | 6,676 |
| 07_70～74歳 | 469 | 1,197 | 3,214 | 4,880 |
| 08_75歳以上 | 166 | 940 | 3,216 | 4,322 |
| 合計 | 40,779 | 13,919 | 14,433 | 69,131 |

| | 世帯区分 | | | 合計 |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | |
| 世帯主の年齢 | | | | |
| 01_24歳以下 | 1,729 | 10 | 21 | 1,760 |
| 02_25～29歳 | 4,314 | 151 | 57 | 4,522 |
| 03_30～34歳 | 6,441 | 480 | 87 | 7,008 |
| 04_35～39歳 | 7,420 | 824 | 153 | 8,396 |
| 05_40～44歳 | 7,611 | 1,145 | 247 | 9,004 |
| 06_45～49歳 | 7,739 | 1,479 | 244 | 9,462 |
| 07_50～54歳 | 8,072 | 2,085 | 508 | 10,665 |
| 08_55～59歳 | 7,697 | 2,725 | 893 | 11,315 |
| 05_60～64歳 | 4,208 | 2,785 | 4,160 | 11,153 |
| 06_65～69歳 | 1,487 | 2,213 | 6,374 | 10,074 |
| 07_70～74歳 | 459 | 1,652 | 6,124 | 8,235 |
| 08_75歳以上 | 207 | 1,325 | 6,873 | 8,406 |
| 合計 | 57,386 | 16,874 | 25,741 | 100,000 |

表 1

### 2）世帯分類別・世帯区分別 世帯主の性別 3 重クロス表

所定の様式の通りとなるよう、tabulate プロシジャを用いて集計を行った。結果は下記の通りである。

| | 世帯区分 | | | | | | | | | 合計 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 勤労者 | | | 勤労者以外 | | | 無職 | | | |
| | 男 | 女 | 計 | 男 | 女 | 計 | 男 | 女 | 計 | |
| 世帯分類 | | | | | | | | | | |
| 単身世帯 | 6,503 | 4,207 | 10,710 | 1,030 | 867 | 1,897 | 2,463 | 5,688 | 8,151 | 20,758 |
| 夫婦のみ | 8,517 | 324 | 8,841 | 4,293 | 81 | 4,375 | 11,026 | 45 | 11,071 | 24,287 |
| 二世代 | 27,267 | 378 | 27,645 | 6,190 | 51 | 6,242 | 3,264 | 32 | 3,295 | 37,182 |
| 二世代(ひとり親) | 1,543 | 2,421 | 3,963 | 934 | 430 | 1,364 | 671 | 1,303 | 1,974 | 7,302 |
| 三世代 | 4,956 | 384 | 5,340 | 2,390 | 118 | 2,508 | 638 | 160 | 798 | 8,646 |
| その他 | 654 | 234 | 887 | 413 | 75 | 488 | 278 | 174 | 452 | 1,826 |
| 合計 | 49,439 | 7,947 | 57,386 | 15,250 | 1,623 | 16,874 | 18,340 | 7,401 | 25,741 | 100,000 |

表 2

3）パン支出金額についての各種の統計グラフ

①所定の様式となるよう、tabulate プロシジャを用いて集計を行った。結果は下記の通りである。

| | パンの支出金額(円) | 1を加えた金額の常用対数 |
|---|---|---|
| 最大値 | 51,776 | 4.714 |
| 第3四部位数 | 3,039 | 3.483 |
| 中央値 | 1,689 | 3.228 |
| 第1四分位数 | 858 | 2.934 |
| 最小値 | 0 | 0.000 |
| 平均値 | 2,284 | 3.137 |
| 件数 | 100,000 | 100,000.000 |

表 3

②sgplot プロシジャの vbox ステートメントを用いて箱ひげ図を描いた。結果は下記の通りである。

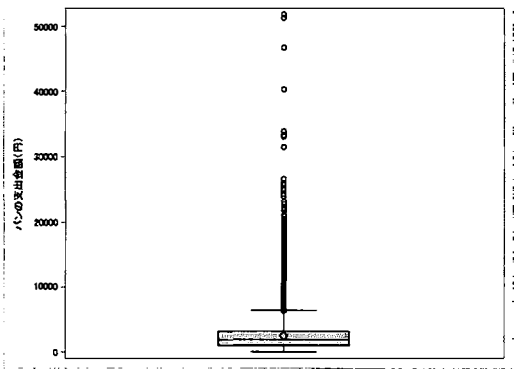



図 3.1

③univariate プロシジャの histogram ステートメントを用いてヒストグラムを描画した。結果は下記の通りである。

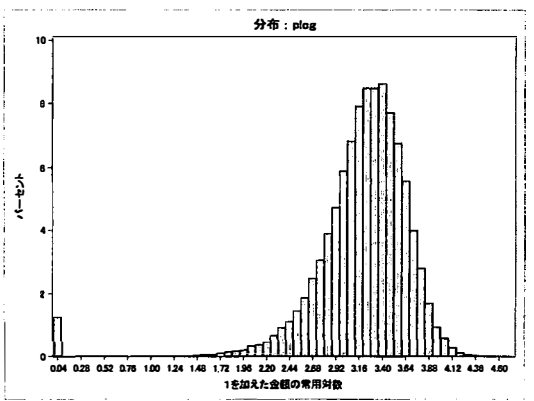



図 3.2

4）世帯区分別および世帯分類別の穀類（米、パン、めん類、他の穀類）の支出金額についての構成比

①所定の様式となるよう、tabulate プロシジャを用いて集計を行った。結果は下記の通りである。

| | 件数 | 米 | パン | めん類 | その他穀類 | 全体(%) |
|---|---|---|---|---|---|---|
| 世帯区分 | | | | | | |
| 勤労者 | 57,386 | 36.01 | 37.60 | 22.10 | 4.29 | 100.00 |
| 勤労者以外 | 16,874 | 44.91 | 30.97 | 19.43 | 4.69 | 100.00 |
| 無職 | 25,741 | 52.10 | 26.69 | 15.78 | 5.43 | 100.00 |
| 世帯分類 | | | | | | |
| 単身世帯 | 20,758 | 33.84 | 37.69 | 23.00 | 5.47 | 100.00 |
| 夫婦のみ | 24,287 | 49.49 | 28.87 | 16.79 | 4.85 | 100.00 |
| 二世代 | 37,182 | 39.01 | 35.92 | 20.76 | 4.31 | 100.00 |
| 二世代(ひとり親) | 7,302 | 44.94 | 31.57 | 19.01 | 4.49 | 100.00 |
| 三世代 | 8,646 | 46.01 | 30.52 | 19.67 | 3.80 | 100.00 |
| その他 | 1,826 | 44.83 | 30.54 | 20.36 | 4.27 | 100.00 |
| 全体 | 100,000 | 41.65 | 33.68 | 20.02 | 4.65 | 100.00 |

表 4

②sgplot プロシジャの vbar ステートメントを用いて棒グラフを描いた。結果は下記の通りである。

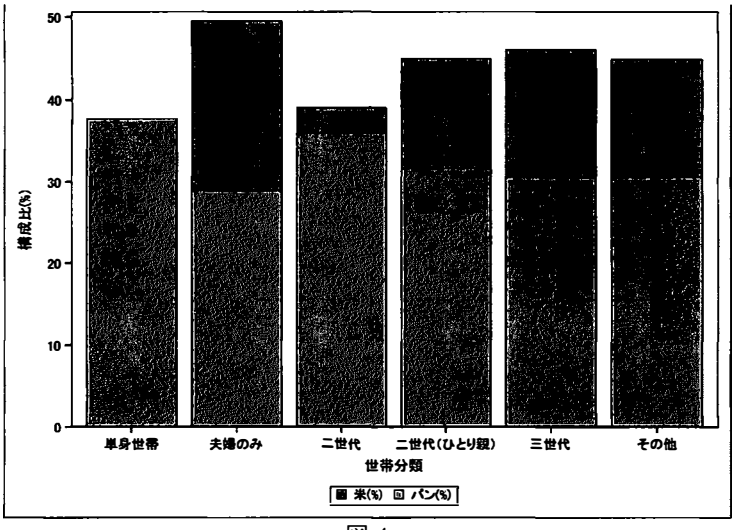



図 4

5）世帯分類別のローレンツ曲線およびジニ係数

①data、sql の各プロシジャを用いて、集計用乗率を考慮したローレンツ曲線を描くためのデータセットを作成した。このデータセットから sgplot プロシジャを用いてローレンツ曲線を描いた。結果は下記の通りである。

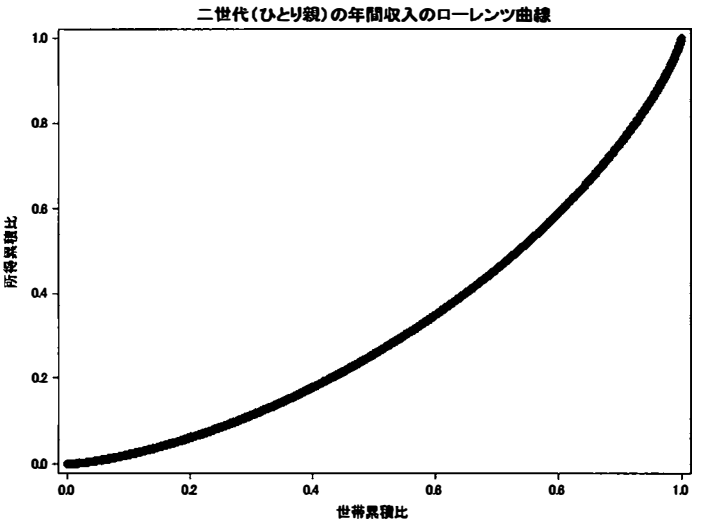



図 5

②data、sql の各プロシジャを用いて、集計用乗率を考慮したジニ係数を算出し、所定の様式となるようなデータセットを作成した。以下はそのデータセットを表示したものである。

| | 世帯分類 | ジニ係数 |
|---|---|---|
| 1 | 単身世帯 | 0.34 |
| 2 | 夫婦のみ | 0.31 |
| 3 | 二世代 | 0.26 |
| 4 | 二世代(ひとり親) | 0.35 |
| 5 | 三世代 | 0.27 |
| 6 | その他 | 0.33 |
| 7 | 全体 | 0.33 |

表 5

以上

# 付録：規定課題用SASプログラム

```
* フォーマットの作成 ;
proc format;
value kinrou 1  = "勤労者" 2  = "勤労者以外" 3  = "無職";
value age    5  = "01_24歳以下" 6  = "02_25～29歳"
             7  = "03_30～34歳" 8  = "04_35～39歳"
             9  = "05_40～44歳" 10 = "06_45～49歳"
             11 = "07_50～54歳" 12 = "08_55～59歳"
             13 = "05_60～64歳" 14 = "06_65～69歳"
             15 = "07_70～74歳" 16 = "08_75歳以上";
value gender 1  = "男" 2 = "女";
value bunrui 1  = "単身世帯" 2  = "夫婦のみ"
             3  = "二世代"  4  = "二世代（ひとり親）"
             5  = "三世代"  6  = "その他";
run;

/*規定問題（1）*/
/*集計乗率なしのクロス集計*/
ods html;
proc tabulate data=microdata missing;
  class  X09 X02;
  table  X09="世帯主の年齢" all="合計",
         ( X02="" all="合計" ) * f = comma8. / rts = 16;
  format X02 kinrou. ; format X09 age. ;
  title "表１";
run;
/*集計乗率ありのクロス集計*/
%global sum_weight;
proc sql noprint;
  select (sum(weight)) into : sum_weight
  from microdata;
quit;
proc sql;
  create table work.a as select
  X02, X09, weight,
  (100000*weight/&sum_weight) as wgt10man
  from microdata;
quit;
proc tabulate data=work.a order=internal;
  class  X09 X02;
  var wgt10man;
  table  X09="世帯主の年齢"*wgt10man="" all="合計"*wgt10man="",
   sum="" * ( X02 all="合計" ) * f = comma8. / rts = 16;
  format X02 kinrou. ; format X09 age. ;
  title "表１";
run;

/*規定問題（2）*/
proc sql;
  create table work.a2 as select
  X02, X08, X11, weight,
  (100000*weight/&sum_weight) as wgt10man
  from microdata;
quit;
proc tabulate data=work.a2 order=internal;
  class  X02 X08 X11;
  var wgt10man;
  table  X11="世帯分類"*wgt10man="" all="合計"*wgt10man="",
  sum="" * ( X02 * ( X08="" all="計" ) all="合計" ) * f = comma8. ;
  format X02 kinrou. ; format X08 gender. ; format X11 bunrui. ;
 title "表２";
run;

/*規定問題（3）*/
data pan;
 attrib price length=8. label="パンの支出金額（円）"
        plog  length=8. label="１を加えた金額の常用対数";
 format weight;
 set microdata;
 price = Y044; plog  = log10(Y044 + 1);
run;
proc sql;
  create table work.a3 as select
  price, plog,
  (100000*weight/&sum_weight) as wgt10man
  from pan;
quit;

/*----------表３の作成----------*/
proc tabulate data = work.a3 out=out_hako ;
  var price plog ;
  wgt wgt10man;
  table max = "最大値"        p75 = "第３四部位数"  median = "中央値"
        p25 = "第１四分位数" min = "最小値"        mean = "平均値"
        sumwgt = "件数",
        (price * f = comma8.  plog * f =comma15.3);
  title "表３";
run;

/*---------箱ひげ図の作成----------*/
proc sgplot data = work.a3;
  vbox price;
run;
/*----------ヒストグラムの作成----------*/
proc univariate data = work.a3;
  histogram plog;
run;

/*規定問題（4）*/
/*以下では、まず４変数（米、パン、めん類、他の穀類）が
穀類に占める構成比（パーセント）を世帯毎に計算する。*/
data micro_kokurui;
  attrib kome length=8. label="米"     pann   length=8. label="パン"
         menn length=8. label="めん類" sonota length=8. label="他の穀類"
         sum  length=8. label="全体(%)";
  set microdata;
  if Y042 ne 0 and Y043 ne . and Y044 ne . and Y045 ne . and Y046 ne . then do;
    kome = 100*Y043/Y042; pann   = 100*Y044/Y042;
    menn = 100*Y045/Y042; sonota = 100*Y046/Y042;
    sum  = kome + pann + menn + sonota;
  end;
  else if
   Y042 = 0 and Y043 ne . and Y044 ne . and Y045 ne . and Y046 ne . then do;
    kome = 0; pann = 0; menn = 0; sonota = 0; sum = 0;
  end;
  else do;
    kome = .; pann = .; menn = .; sonota = .; sum = .;
  end;
run;

/*以下では、ウェイトの計算をする。*/
proc sql;
  create table work.a4 as select
  weight, X02, X11, Y042, Y043, Y044, Y045, Y046,
  kome, pann, menn, sonota, sum,
  (100*kome/sum) as kome_p, (100*pann/sum)   as pann_p,
  (100*menn/sum) as menn_p, (100*sonota/sum) as sonota_p,
  (100 * kome/sum + 100 * pann/sum + 100 * menn/sum + 100 * sonota/sum) as sum_p,
  (100000*weight/&sum_weight) as wgt10man
  from micro_kokurui;
quit;

/*以下では、集計表を作る。*/
proc tabulate data=work.a4 out=outpct format=best12. missing;
  var kome_p pann_p menn_p sonota_p sum_p wgt10man;
  class X02 X11;
  table
    /*行指定*/
    X02 X11 all="全体",
    /*列指定 */
    wgt10man="件数"       * (sumwgt="" * f = comma8.0)
    kome_p="米"           * (mean=""   * f = 8.2)
    pann_p="パン"         * (mean=""   * f = 8.2)
    menn_p="めん類"       * (mean=""   * f = 8.2)
    sonota_p="その他穀類" * (mean=""   * f = 8.2)
    sum_p="全体(%)"       * (mean=""   * f = 8.2);
  weight wgt10man;
  format X02 kinrou. ; format X11 bunrui. ;
  title "表４";
run;

/*世帯分類別に、穀類（米、パン、めん類、他の穀類）の構成比を帯グラフで表示する。
*/
proc sgplot data=outpct;
  vbar X11 / response=kome_p_Mean legendlabel="米(%)" ;
  vbar X11 / response=pann_p_Mean legendlabel="パン(%)";
  yaxis label="構成比(%)";
run;

/*規定課題（5）*/
/*全体のローレンツ曲線とジニ係数の計算*/
proc sql;
  create table work.a5 as
  select
  X11, Y001, (100000*weight/&sum_weight) as wgt10man
  from micro_kokurui;
quit;
proc sort data=work.a5;
  by Y001;
run;
data a6;
  set work.a5;
  attrib a length=8. b length=8. ;
  retain a b 0;
  a = a + wgt10man; b = b + Y001*wgt10man;
run;
proc sql;
  create table work.a7 as select
   *, max(a) as max_a, max(b) as max_b
  from work.a6;
quit;
data a8;
```

```
  set a7;
  yoko = a / max_a; tate = b / max_b;
run;
data a9;
  set a8;
  attrib yoko_pre length=8. label="横軸" tate_pre length=8. label="縦軸";
  retain yoko_pre tate_pre;
  if _N_ = 1 then do;
    yoko_pre = 0; tate_pre = 0;
  end;
  output;
  yoko_pre = yoko; tate_pre = tate;
  drop X11 Y001;
run;
data count(keep=cnt);
  set a9 end=last;
  retain cnt 0;
  cnt + 1;
  if last;
run;

%global num;
proc sql noprint;
  select (cnt) into : num
  from count;
quit;
data LorenzGiniAll;
  set a9;
  attrib coef length=8.;
  do i=1 to &num;
    coef = 0.5 * (yoko - yoko_pre)*(yoko + yoko_pre - tate - tate_pre);
  end;
run;
%global sum_gini;
proc sql noprint;
  select (round(2*sum(coef),0.01)) into : sum_gini
  from LorenzGiniAll;
quit;
data GiniAll;
  attrib gini length=8.;
  gini = &sum_gini;
run;

/*世帯分類毎のローレンツ曲線とジニ係数の計算*/
%MACRO Lorenz(VAR);
proc sort data=work.a5;
  by X11;
run;
data b1;
  set work.a5;
  attrib a length=8. label="横軸"; attrib b length=8. label="縦軸";
  retain a b 0;
  by X11;
  if first.X11 = 1 then do;
    a = 0; b = 0;
  end;
  a = a + wgt10man; b = b + Y001*wgt10man;
run;
data subdata;
  set b1;
  if X11 = &VAR. then output;
run;
proc sql;
  create table work.subdata2 as select
   *, max(a) as max_a, max(b) as max_b
  from work.subdata;
quit;
data subdata3;
  set subdata2;
  yoko = a / max_a; tate = b / max_b;
  drop a b max_a max_b wgt10man;
run;
data subdata4;
  set subdata3;
  attrib yoko_pre length=8. label="横軸" tate_pre length=8. label="縦軸";
  retain yoko_pre tate_pre;
  if _N_ = 1 then do;
    yoko_pre = 0; tate_pre = 0;
  end;
  output;
  yoko_pre = yoko; tate_pre = tate;
  drop X11 Y001;
run;
data count(keep=cnt);
  set subdata4 end=last;
  retain cnt 0;
  cnt + 1;
  if last;
run;
%global num;
proc sql noprint;
  select (cnt) into : num
  from count;
quit;
data LorenzData0;
  set subdata4;
  rename yoko_pre = x_axis tate_pre = y_axis;
  drop yoko tate;
run;
data LorenzData1;
  attrib x_axis length=8. y_axis length=8.;
  x_axis = 1; y_axis = 1;
run;
data LorenzData;
  set LorenzData0 LorenzData1;
run;
roc sgplot data=LorenzData;
  where &VAR. = 4;
  scatter x = x_axis y = y_axis;
  title "二世代（ひとり親）の年間収入のローレンツ曲線";
  xaxis label="世帯累積比"; yaxis label="所得累積比";
run;
data LorenzGini;
  set subdata4;
  attrib coef length=8.;
  do i=1 to &num;
    coef = 0.5 * (yoko - yoko_pre)*(yoko + yoko_pre - tate - tate_pre);
  end;
run;
%MEND

%global sum_1 sum_2 sum_3 sum_4 sum_5 sum_6;
%Lorenz(VAR=1);
proc sql noprint;
  select (round(2*sum(coef),0.01)) into : sum_1
  from LorenzGini;
quit;
%Lorenz(VAR=2);
proc sql noprint;
  select (round(2*sum(coef),0.01)) into : sum_2
  from LorenzGini;
quit;
%Lorenz(VAR=3);
proc sql noprint;
  select (round(2*sum(coef),0.01)) into : sum_3
  from LorenzGini;
quit;
%Lorenz(VAR=4);
proc sql noprint;
  select (round(2*sum(coef),0.01)) into : sum_4
  from LorenzGini;
quit;
%Lorenz(VAR=5);
proc sql noprint;
  select (round(2*sum(coef),0.01)) into : sum_5
  from LorenzGini;
quit;
%Lorenz(VAR=6);
proc sql noprint;
  select (round(2*sum(coef),0.01)) into : sum_6
  from LorenzGini;
quit;
data Gini;
  gini_1 = &sum_1; gini_2 = &sum_2; gini_3  = &sum_3; gini_4 = &sum_4; gini_5
  = &sum_5; gini_6 = &sum_6;
  gini_all = &sum_gini;
  proc print;
run;
proc transpose data=Gini out=tp_Gini01;
run;
data tp_Gini02;
  input bunrui $25.;
  length bunrui $25.; length n_gini 8.;
  label bunrui = "世帯分類"; label n_gini = "ジニ係数";
  set tp_Gini01;
  n_gini = COL1;
  drop _NAME_ drop COL1;
  cards;
  単身世帯
  夫婦のみ
  二世代
  二世代（ひとり親）
  三世代
  その他
  全体
  ;
run;
```

以上

# 経済格差を測る食事多様性指標の提案

木下裕貴[1]、仕子優樹[1]、亀井健史[1]、堀口剛[1]、山本周平[2]、伊藤俊貴[2]

[1] 中央大学大学院理工学研究科都市人間環境学専攻
[2] 中央大学理工学部人間総合理工学科
カテゴリー：学生

Proposal of Dietary Diversity Index to Measure Economic Disparity

Yuki Kinoshita[1], Yuki Shiko[1], Takeshi Kamei[1], Go Horiguchi[1], Shuhei Yamamoto[2], Toshiki Ito[2]
[1]Civil, Human and Environmental Engineering, Chuo University
[2]Department of Integrated Science and Engineering for sustainable society, Chuo University

## 要旨

本研究の目的は、食生活の面から経済格差を測る食事多様性指標を提案し、その妥当性を検証することである。食事の多様性を測る指標として、生態学の分野で提唱された種の多様性指標（Shannon 指数）が代替可能であると考え、6 つの食品群の消費金額を用いる形で食事多様性得点を算出した。妥当性については、主成分分析の第一主成分がバランスの取れたパターンとなることを踏まえ、第一主成分得点との相関を見ることで検証した。その結果、1 人世帯では中程度の正の相関が見られたが、世帯人員が 2 人以上の集団では相関係数が低かった。この結果には多様性得点のバラツキの大きさが影響していると考えられるため、今回提案した食事多様性指標には妥当性があると言える。単価の調整等の課題はあるが、食事多様性指標の応用の幅は広い。

キーワード：食事多様性指標　全国消費実態調査　経済格差　妥当性　Shannon 指数

## 1. 背景・目的

全国消費実態調査は、家計の構造を「所得」、「消費」、「資産」の 3 つの側面から総合的に把握することを目的とし、5 年間隔で実施されている大規模調査である [1)]。この調査から得られるデータには、世帯別の費用、交際費や食費など多岐にわたる項目が含まれており、時系列的に他項目との関連を見ることができるという特徴を持つ。そのため、各府省の白書、大学や研究機関などで経済状況を把握することを目的に

広く利用されている。経済的な格差によってもたらされる格差社会は、経済活動の衰退や生活水準・食生活の悪化などの懸念を含むため[2]、こうした調査による経済格差の把握は重要である。

特に、食生活と経済格差の関連については、厚生労働省による調査、および小林（2010）が「所得の低い世帯では、所得の高い世帯と比較して、穀類の摂取量が多く野菜類や肉類の摂取量が少ない」[3]「高階層の人ほど、野菜と海藻をよく食べていた」[4]という結果を示している上、近年ではファストフードやコンビニ食の普及によって食の偏りが生まれることが指摘されている[5]。こうした背景を踏まえ、我々は食生活の面から経済格差を定量的に説明できるのではないかと考えた。

現在のところ、経済格差を示す定量的な指標としてはジニ係数やエンゲル係数がある。しかし、ジニ係数は所得のみから算出される係数であるため、食に関する情報を含んでいるわけではない。一方で、エンゲル係数については食費をもとに算出されるため、食の面から経済格差を測る指標ではあるが、食費をひとまとめにしているところに欠点がある。先で述べた通り、所得が偏食に影響していることが示唆されているため、食生活から経済格差をより詳細に測るためには、食のバランスを考慮した指標を考えるべきである。こうした指標は経済格差を測るばかりでなく、他の変数との関連が見られるため、疫学研究への応用等も可能となる。しかし、そのような食のバランスを定量的に評価する有効な指標はない。そこで、本研究では生態学の分野で提唱された種の多様性指標を食のバランスを測るために適用し、それを食事多様性の指標として利用することを提案する。最終的には、食事に特化した独自の多様性指標を開発することが目標ではあるが、本研究では種の多様性指標が食事多様性指標として代替可能かどうかを検討することを目的とする。

## 2. 方法

### 2.1 食事多様性指標の提案

生態学の分野では種の多様性の指標として Shannon 指数($H'$)がしばしば利用されており各種が均等に存在しているかを測る事ができる(1)[6]。この指数は多様性が高いほど値が大きくなり、多様性が低いほど値は小さくなるという特徴がある。我々はこの指標を食事の多様性に適用する事を考えた。そこで食事の多様性をある食品群のみの購買に偏らず様々な食材を購買している事であると定義した。その際、食品群としては穀類、魚介類、肉類、野菜・海藻類、乳卵類、果実類の 6 つの消費金額を使用し、(1)を用いて多様性得点を算出した。

$$H' = -\sum_{i=1}^{S} P_i \mathrm{In} P_i \qquad (1)$$

$$\left(P_i = \frac{i\text{番目の食品群の消費金額}}{\text{食品群の消費金額の合計}}、S\text{=食品群数}\right)$$

### 2.2 除外基準

農林業を営んでいる世帯は自給自足をしている可能性があり一般集団から外れた食事パターンになると考え、本研究の対象から除いた。また、穀類、魚介類、肉類、野菜・海藻類、乳卵類、果実類の 6 つの合計消費金額が 0 円の世帯についても除外対象とした。

### 2.3 解析方法

近年の疫学研究では、食習慣の評価に主成分分析や因子分析が用いられている。特に、主成分分析の第一主成分については、バランスの取れたパターン（大きさの指標）となることが多いため、第一主成分得点と多様性得点はどちらも各食品群に対する消費金額の均等度合を測る指標であると言える。そこで本研究では、多様性得点の妥当性を、主成分分析の第一主成分得点を用いスピアマンの順序相関分析を行うことで検証した。主成分分析には、多様性得点の算出に用いたものと同様の食品群を使用した。第一主成分得点は消費金額が多い程、得点が高くなってしまうため、解析に使用した食品群の総消費で除すことで補正した。この際、世帯人員が異なることによる交絡を考慮し、すべての解析は世帯人員別に行った。また、世帯人員ごとの食費支出の違いを把握するために、世帯人員ごとの食費の内訳を計算し、表にまとめた。

## 3. 結果

### 3.1 解析対象世帯

全国消費実態調査によって得られた 69,131 世帯のデータから、方法で示した除外基準に従い解析対象世帯を選定した。図 1 は対象世帯選択のフローチャートである。全調査世帯から農林漁業収入のある世帯（553 世帯）と穀類、魚介類、肉類、野菜・海藻類、乳卵類、果実類の 6 つの合計消費金額が 0 円であった世帯（16 世帯）を除いた 68,582 世帯が最終的に解析対象世帯となった。

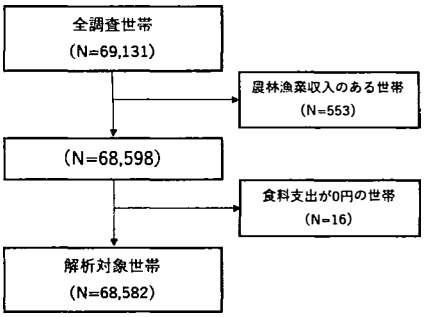



**図 1. 解析対象世帯選択フローチャート**

### 3.2 記述統計

表 1 に多様性得点（Shannon 指数）および主成分得点の平均と標準偏差を世帯人員ごとに示した。その結果、1 人世帯において多様性得点の平均が他の世帯人員よりも小さな値になっていた（1.46）。主成分得点の平均でも同様の傾向を示した（−0.22）。また多様性得点のバラツキを見るため、世帯人員ごとにヒストグラムを描いたところ、1 人世帯では他の世帯人員よりもバラツキが大きいことが確認できた（図 2）。

**表 1. 世帯人員ごとの多様性得点と第一主成分得点の平均および標準偏差**

| 世帯人員 | 世帯数 | 多様性得点 | 第一主成分得点* |
|---|---|---|---|
| 1 人 | 5608 | 1.46 ± 0.29 | -0.22 ± 1.41 |
| 2 人 | 20106 | 1.63 ± 0.10 | -0.09 ± 5.88 |
| 3 人 | 16488 | 1.63 ± 0.09 | -0.03 ± 0.09 |
| 4 人 | 14973 | 1.63 ± 0.08 | -0.03 ± 0.09 |
| 5 人以上 | 11407 | 1.63 ± 0.09 | -0.02 ± 0.05 |
| 全世帯 | 68582 | 1.62 ± 0.13 | -0.06 ± 3.21 |

*主成分得点は 1,000 倍した値を示している

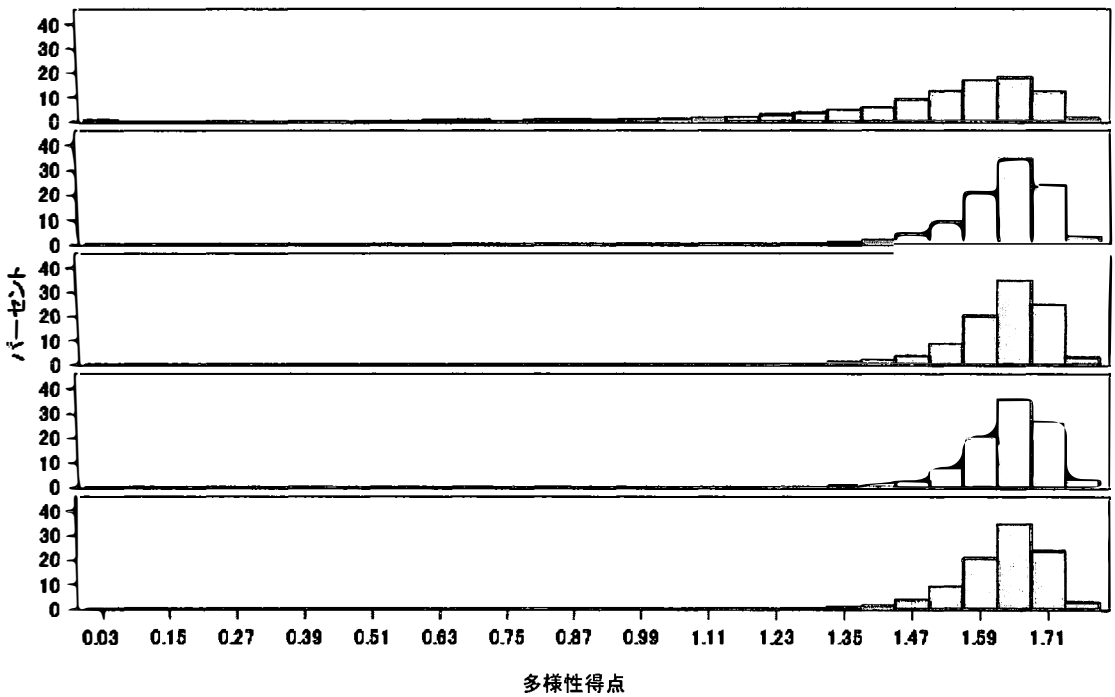



図 2. 世帯人員別の多様性得点のヒストグラム

*上から順に 1 人世帯、2 人世帯、3 人世帯、4 人世帯、5 人以上世帯、を示す

### 3.3 主成分分析の結果

各世帯人員における食パターンを確かめるため主成分分析を行った。表 2 に各食品群における固有ベクトルを世帯人員別に示した。いずれの固有ベクトルも正の値であったため、想定通り第一主成分はバランスの取れたパターンとなることが確認できた。また、世帯人員間に大きな差は見られなかった。

表 2. 各食品群の固有ベクトル：世帯人員別

| 世帯人員 | 穀類 | 魚介類 | 肉類 | 野菜・海藻類 | 乳卵類 | 果実類 |
|---|---|---|---|---|---|---|
| 1 人 | 0.38 | 0.42 | 0.39 | 0.46 | 0.37 | 0.42 |
| 2 人 | 0.39 | 0.43 | 0.38 | 0.45 | 0.39 | 0.40 |
| 3 人 | 0.38 | 0.44 | 0.40 | 0.47 | 0.36 | 0.40 |
| 4 人 | 0.37 | 0.44 | 0.40 | 0.47 | 0.37 | 0.39 |
| 5 人以上 | 0.33 | 0.45 | 0.43 | 0.48 | 0.36 | 0.39 |

### 3.4 世帯人員ごとの食費の内訳

世帯人員によって各食品群への支出に差があるのかを判断するため、食料支出に占める各食品群の割合の平均を求めた（表 3)。1 人世帯では、他の世帯人員と比較すると穀類への支出割合が高くなっており（0.263)、対して魚介類・肉類への支出割合が低くなっていた（0.162)。野菜・海藻類、乳卵類、果実類では各世帯人員で大きな差は見られなかった。

**表 3. 食料支出に占める各食品群割合の平均値：世帯人員別**

| 世帯人員 | 穀類 | 魚介類 | 肉類 | 野菜・海藻類 | 乳卵類 | 果実類 |
|---|---|---|---|---|---|---|
| 1 人 | 0.263 | 0.162 | 0.105 | 0.258 | 0.115 | 0.097 |
| 2 人 | 0.195 | 0.206 | 0.137 | 0.284 | 0.087 | 0.091 |
| 3 人 | 0.208 | 0.191 | 0.171 | 0.261 | 0.095 | 0.073 |
| 4 人 | 0.218 | 0.177 | 0.197 | 0.245 | 0.101 | 0.061 |
| 5 人以上 | 0.233 | 0.181 | 0.201 | 0.235 | 0.101 | 0.059 |

### 3.5 妥当性評価

多様性得点の妥当性を評価するため主成分得点とのスピアマンの順序相関分析を世帯人員別に行った（表 4)。その結果、1 人世帯では多様性得点と主成分得点間に中程度の相関が見られた（R=.468)。他の世帯人員では相関係数は低かった（R<.200)。

**表 4. 多様性得点と第一主成分得点の相関**

| 世帯人員 | 相関係数(R) |
|---|---|
| 1 人 | 0.468* |
| 2 人 | 0.197* |
| 3 人 | 0.145* |
| 4 人 | 0.121* |
| 5 人以上 | 0.135* |

*：p<0.001 世帯数が多いためすべて有意となった

## 4. 考察

Shannon 指数によって算出された多様性得点が、主成分分析によって得られる主成分得点と整合性があることを確認するために相関分析を行った結果、1 人世帯において中程度の正の相関が見られた。

多様性得点は食費が 6 つの食品群（穀類、魚介類、肉類、野菜・海藻類、乳卵類、果実類）に対して均等に分配されているほど高くなるため、バランス型の食事パターンを示した第一主成分の主成分得点と類似した傾向を持つ指標と考えられる。実際、1 人世帯においてはそのことを支持する結果が得られた。しかし、世帯人員が 2 人以上の場合においては、相関係数が低かった。そこで多様性得点の分布を見ると、世帯人員が 2 人以上ではバラツキが小さいことが分かる。つまり、集団全体として多様性に差がなかったこ

とが 2 つの指標間の相関が低い理由だと考えられるため、1 人世帯での結果が食事多様性指標の妥当性を示していると言える。こうしたことから、本研究の結果は、種の多様性指標である Shannon 指数は食の多様性を評価する場合にも使用可能であることを示唆する。

本研究の限界として、次の 2 点を挙げる。まず 1 つ目は食品によって単価が異なることによる影響である。主成分分析の場合は変数のバラツキのみを見ているため単価の影響はないが、多様性得点の場合は食費をそのまま利用するため、ある食品に対する費用がその食品の摂取量を表している訳ではないという意味で単価の影響を受けてしまう。そのため、より厳密に食の多様性を測りたい場合には食品の単価を考慮に入れる必要がある。2 つ目は多様性得点が食の良し悪しを評価するわけではない点である。特に今回提案した指標の場合、どの食品群にも均等にお金を使っていると得点が高くなるが、このことが良いかどうかはまた別の問題である。例えば、健康問題と関連付ける場合には理想的な食事バランスで重み付けを行うなどの調整をすることで、良し悪しの議論に落とし込めるのではないかと考える。

全国消費実態調査は 5 年間隔で行われているものであるため、時系列的な解析をするのに適していると言える。また、全国消費実態調査のデータには食費に関する細かい情報が含まれているが、そのデータを利用した研究はほとんどない。今回提案した多様性得点の主な特徴は、主成分得点とは違い集団に対して独立であること、食事に関する詳細な情報を反映した指標であることの 2 点である。そのため、年次比較や地域差の検討など集団間での比較が、食事に関して可能となる。つまり、この食事多様性指標の提案は、全国消費実態調査のデータと食事の研究を結び付け、さらには健康や疾病に関するアウトカムとの関連を見ることによって、疫学研究への応用も可能とする。また、ファストフードやコンビニ食などが今後さらに便利になっていくことによって食生活の多様化が進み、何をどのように食べるかの選択肢が増えると、嗜好の偏りを招く可能性があるため、食の多様性は減少していくと考えられる。したがって、今回提案した指標によるこうした傾向の把握は重要である。

本研究では、食事多様性指標の妥当性の検証までを行ったが、経済格差を食生活の面から測る指標の開発が最終的な目標であるため、経済格差との関連を見ることは必須である。そのためには、単価の影響を考慮に入れた解析をすべきであるため、それは今後の課題としたい。こうした課題はあるものの、本研究の成果は食生活から経済格差を測るための第一歩となるであろう。

## 5. 参考文献


1) 総務省統計局
http://www.stat.go.jp/data/zensho/2014/
2017 年 6 月 26 日アクセス可能

2) どうしてこうなった!?～不満だらけの日本の今～
https://booing.jp/discontent/gap.html
2017 年 6 月 26 日アクセス可能

3) 厚生労働省.国民健康栄養調査
http://www.mhlw.go.jp/stf/houdou/0000106405.html
2017 年 6 月 26 日アクセス可能

4) 小林 盾.社会階層と食生活 ー健康への影響の分析ー . 数理社会学会 機関紙「理論と方法」. 2010;25:81-93.

5) 永嶋久美子, 坂口早苗, 坂口武洋. 女子学生の偏食行動様式別食習慣および健康習慣の実態. 日本公衛誌. 2002;49: 447-455.

6) 多様度と類似度、分類学的新指標
http://www.mus-nh.city.osaka.jp/iso/argo/nl15/nl15-10-22.pdf
2017 年 6 月 26 日アクセス可能

# 付録　規定課題用 SAS プログラム

## 1) 年齢階層別・世帯区分別 クロス表

```
*フォーマットの作成;
proc format;
  value  household 1="勤労者"  2="勤労者以外"  3="無職";
  value  age 5 ="24歳以下"  6 = "25～29歳"  7 = "30～34歳"  8 = "35～39歳"  9 = "40～44歳" 10 = "45
          ～49歳"  11 = "50～54歳" 12= "55～59歳" 13 = "60～64歳" 14 = "65～69歳" 15 ="70～74歳"
          16 = "75歳以上";
  value family 1="単身世帯" 2="夫婦のみ" 3="二世代" 4="二世代(ひとり親)" 5="三世代" 6="その他" 7='合計
';
  value sex 1="男" 2="女";
  value kokurui 1="米" 2="パン" 3="めん類" 4="その他";
run;

*集計用乗率なしのクロス集計表;
proc tabulate data=micro.Zensho2004gijimicro;
  class x02/order=unformatted;class x09/order=unformatted;
  format x09 age. x02 household.;keylabel N=" " SUM="合計";
  table  x09=" " all="合計",(x02="世帯区分" all="合計")*(f=comma7.)
   /misstext="0" box="世帯主の年齢" row=float;
run;

*集計用乗率ありのクロス集計表;
proc tabulate data=micro.Zensho2004gijimicro;
  class x02/order=formatted;class x09/order=unformatted;
  format x09 age. x02 household.;var/WEIGHT=weight;
  keylabel N=" " SUM=" ";table  x09=" "*WEIGHT=" " all="合計"*WEIGHT=" ",x02="世帯区分" all="合計
  " /misstext="0" box="世帯主の年齢" row=float;
run;
data micro.Zensho2004gijimicro;
  set micro.Zensho2004gijimicro;
  weight2=weight/318.8730695;
run;

*集計用乗率を使った10万世帯比のクロス集計表;
proc tabulate data=micro.Zensho2004gijimicro;
  class x02/order=unformatted;class x09/order=unformatted;
  format x09 age. x02 household.;var/WEIGHT=weight2;
  keylabel N=" " SUM=" ";
  table  x09=" "*WEIGHT2=" " all="合計"*WEIGHT2=" ",
       (x02="世帯区分" all="合計")*(f=comma7.)/misstext="0" box="世帯主の年齢" row=float;
run;
```

## 2) 世帯分類別・世帯区分別　世帯主の性別 3 重クロス表

```
*集計用乗率なしのクロス集計表;
proc tabulate data=micro.Zensho2004gijimicro;
  class x02/order=formatted;class x08/order=unformatted;class x11/order=formatted;
  format x02 household. x08 sex. x11 family.;keylabel N=" " SUM=" ";
  table  x11=" " all="合計",x02=" "*(x08=" " all="計") all="合計"
   /misstext="0" box="世帯分類" row=float;
run;

*集計用乗率を使った10万世帯比のクロス集計表;
proc tabulate data=micro.Zensho2004gijimicro;
  class x02/order=unformatted;class x08/order=unformatted;class x11/order=unformatted;
  format x02 household. x08 sex. x11 family.;var/WEIGHT=weight2;
  keylabel N=" " SUM=" ";
  table  x11=" "*WEIGHT2=" " all="合計"*WEIGHT2=" ",
       (x02=" "*(x08=" " all="計") all="合計")*(f=comma7.) /misstext="0" box="世帯分類"
row=float;
run;
```

3) パン支出金額についての各種の統計グラフ

```
*パン支出金額の常用対数変換;
data micro.Zensho2004gijimicro;
  set micro.Zensho2004gijimicro;
  ly044=log10(y044+1);
run;
proc tabulate data=micro.Zensho2004gijimicro;
 var y044 ly044;weight weight2;
 table max="最大値" q3="第3四分位数" median="中央値" q1="第1四分位数"
 (min="最小値")*(f=comma7.)  mean="平均値" (sumwgt="件数")*(f=comma7.),
 (y044="パンの支出金額（円）")*(f=comma7.) (ly044="1を加えた金額の常用対数") *(f=comma7.3)
  /box="統計量" format_precedence=row;
run;

*箱ひげ図とヒストグラムの作成;
proc sgplot data=micro.Zensho2004gijimicro;
  vbox y044;
  label y044="パン支出金額";
run;
proc gchart data=micro.Zensho2004gijimicro;
  vbar ly044/ space=0 midpoints=0 to 5 by 0.25;
  label ly044="パン支出金額の常用対数変換値";
run;
quit;
```

4) 世帯区分別および世帯分類別の穀類の支出金額についての構成比

```
*構成比の計算;
data micro.percent;
  set micro.Zensho2004gijimicro;
  py043=y043/y042*100;
  py044=y044/y042*100;
  py045=y045/y042*100;
  py046=y046/y042*100;
  sum=py043+py044+py045+py046;
run;

*表の作成;
proc tabulate data=micro.percent;
  class x02/order=unformatted;
  class x11/order=unformatted;
  var py043 py044 py045 py046 sum/weight=weight2;
  format x02 household. x11 family.;
  keylabel N=" " SUM="件数";
  table  all="世帯区分"*x02=" " all="世帯分類"*x11=" " all="全体", (all=" "*weight2="
")*(f=comma7.)
       (all="構成比(%)"*(py043="米" py044="パン" py045="めん類" py046="他の穀類")*mean=" "
       sum="全体(%)"*mean=" ")*(f=5.1)/misstext="0" box="項目名";
run;

*データセットの変換;
data micro.trans;
 set table;
 if X11='.' then delete;
 drop _TYPE_--weight2_Sum sum_Mean X02;
 retain a 0;
 a=a+1;
run;
proc transpose data=micro.trans out=micro.bar;
var py043_Mean--py046_Mean;
by x11;
run;
```

```
data micro.bar2;
 set  micro.bar;
  if _name_='py043_Mean' then kokurui=1;
  if _name_='py044_Mean' then kokurui=2;
  if _name_='py045_Mean' then kokurui=3;
  if _name_='py046_Mean' then kokurui=4;
run;

*帯グラフの作成;
proc sgplot data=micro.bar2;
 vbar x11/ response=col1 group=kokurui groupdisplay=stack;
 label x11='世帯区分' kokurui='穀類品目' col1='構成比 [%]';
 format kokurui kokurui.;
run;
```

5) 世帯分類別のローレンツ曲線およびジニ係数

```
*ローレンツ;
%macro data(var);
proc freq data=micro.Zensho2004gijimicro;
where x11=&var;
tables y001/missing;
weight weight2;
ods output onewayfreqs=micro.table;run;
data micro.table;
set micro.table;
all_Nenkan_Syuunyuu=Y001*frequency;
num=1;
run;
proc means data=micro.table sum;
ods output summary=micro.table2;
var all_Nenkan_Syuunyuu;run;

data micro.table2;
set micro.table2;
num=1;
run;
data micro.table3;
merge micro.table(in=a) micro.table2;
by num;
if a;
sum+all_Nenkan_Syuunyuu;
menseki=(((sum+Lag(sum))/all_Nenkan_Syuunyuu_sum)*(Cumpercent-Lag(Cumpercent))/100)/2;
aaa=sum/all_Nenkan_Syuunyuu_sum;
cumpercent2=cumpercent/100;
label aaa="年間収入の累積相対度数"
cumpercent2="累積相対度数";
run;

proc means data=micro.table3 sum;
var menseki;
ods output summary=micro.table4_&var;run;

data micro.table4_&var;
set micro.table4_&var;
jini=(0.5-menseki_sum)*2;
x11=&var;
run;
```

```
data micro.table5;
function="move";
xsys="1";ysys="1";
x=0;y=0;
output;
function="draw";
xsys="1";ysys="1";
color="red";
x=100;
y=100;
output; run;
ods listing;
ods results;
title "年間収入のローレンツ曲線(二世代(ひとり親))";
proc gplot data=micro.table3;
plot aaa*Cumpercent2/
vref=0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.7,0.8,0.9,1.0
href=0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.7,0.8,0.9,1.0 anno=micro.table5;
run;
quit;
run;
%mend;
%data(1) %data(2) %data(3) %data(4) %data(5) %data(6) %data(all)
title;

*ジニ係数の出力;
data micro.total;
set  micro.table4_1 micro.table4_2 micro.table4_3
    micro.table4_4 micro.table4_5 micro.table4_6 micro.table4_all;
jini2=round(jini,0.01);
drop jini menseki_Sum all;
if x11='.' then x11=7;
label jini2='ジニ係数'
     x11='世帯分類';
run;
proc print noobs label;
format x11 family.;
run;
```

# 付録 規定課題図表

1) 年齢階層別・世帯区分別 クロス表

| 世帯主の年齢 | 世帯区分 | | | 合計 |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | |
| 24歳以下 | 765 | 18 | 15 | 798 |
| 25～29歳 | 2,289 | 157 | 66 | 2,512 |
| 30～34歳 | 4,223 | 511 | 107 | 4,841 |
| 35～39歳 | 5,217 | 728 | 193 | 6,138 |
| 40～44歳 | 5,492 | 1,076 | 205 | 6,773 |
| 45～49歳 | 5,702 | 1,305 | 277 | 7,284 |
| 50～54歳 | 6,194 | 1,797 | 392 | 8,383 |
| 55～59歳 | 5,682 | 2,221 | 624 | 8,527 |
| 60～64歳 | 3,237 | 2,226 | 2,534 | 7,997 |
| 65～69歳 | 1,343 | 1,743 | 3,590 | 6,676 |
| 70～74歳 | 469 | 1,197 | 3,214 | 4,880 |
| 75歳以上 | 166 | 940 | 3,216 | 4,322 |
| 合計 | 40,779 | 13,919 | 14,433 | 69,131 |

| 世帯主の年齢 | 世帯区分 | | | 合計 |
|---|---|---|---|---|
| | 勤労者 | 勤労者以外 | 無職 | |
| 24歳以下 | 1,729 | 10 | 21 | 1,760 |
| 25～29歳 | 4,314 | 151 | 57 | 4,522 |
| 30～34歳 | 6,441 | 480 | 87 | 7,008 |
| 35～39歳 | 7,420 | 824 | 153 | 8,396 |
| 40～44歳 | 7,611 | 1,145 | 247 | 9,004 |
| 45～49歳 | 7,739 | 1,479 | 244 | 9,462 |
| 50～54歳 | 8,072 | 2,085 | 508 | 10,665 |
| 55～59歳 | 7,697 | 2,725 | 893 | 11,315 |
| 60～64歳 | 4,208 | 2,785 | 4,160 | 11,153 |
| 65～69歳 | 1,487 | 2,213 | 6,374 | 10,074 |
| 70～74歳 | 459 | 1,652 | 6,124 | 8,235 |
| 75歳以上 | 207 | 1,325 | 6,873 | 8,406 |
| 合計 | 57,386 | 16,874 | 25,741 | 100,000 |

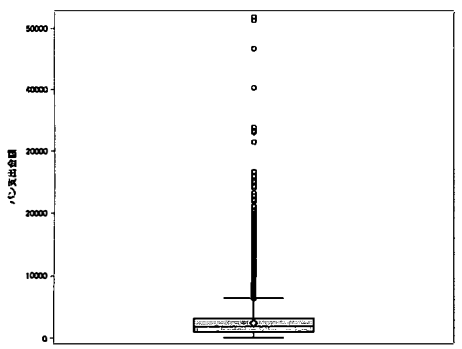
## 2) 世帯分類別・世帯区分別　世帯主の性別 3重クロス表

| 世帯分類 | 勤労者 | | | 勤労者以外 | | | 無職 | | | 合計 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 男 | 女 | 計 | 男 | 女 | 計 | 男 | 女 | 計 | |
| 単身世帯 | 6,503 | 4,207 | 10,710 | 1,030 | 867 | 1,897 | 2,483 | 5,668 | 8,151 | 20,758 |
| 夫婦のみ | 8,517 | 324 | 8,841 | 4,293 | 81 | 4,375 | 11,026 | 45 | 11,071 | 24,287 |
| 二世代 | 27,267 | 379 | 27,645 | 6,190 | 51 | 6,242 | 3,264 | 32 | 3,295 | 37,182 |
| 二世代(ひとり親) | 1,543 | 2,421 | 3,963 | 934 | 430 | 1,364 | 671 | 1,303 | 1,974 | 7,302 |
| 三世代 | 4,956 | 384 | 5,340 | 2,390 | 118 | 2,508 | 638 | 160 | 798 | 8,646 |
| その他 | 654 | 234 | 887 | 413 | 75 | 488 | 278 | 174 | 452 | 1,826 |
| 合計 | 49,439 | 7,947 | 57,386 | 15,250 | 1,623 | 16,874 | 18,340 | 7,401 | 25,741 | 100,000 |

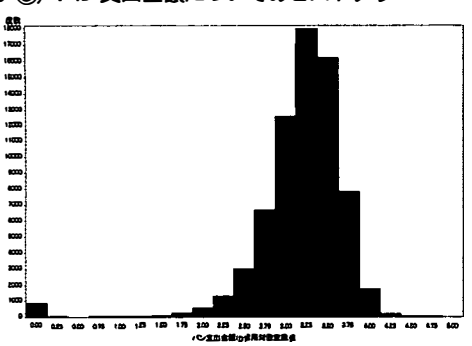
## 3-①) パン支出金額の四分位統計量と平均値

| 統計量 | パンの支出金額(円) | 1を加えた金額の常用対数 |
|---|---|---|
| 最大値 | 51,776 | 4.714 |
| 第3四分位数 | 3,039 | 3.483 |
| 中央値 | 1,689 | 3.228 |
| 第1四分位数 | 858 | 2.934 |
| 最小値 | 0 | 0 |
| 平均値 | 2,284 | 3.137 |
| 件数 | 100,000 | 100,000 |

## 3-②) パン支出金額についての箱ひげ図



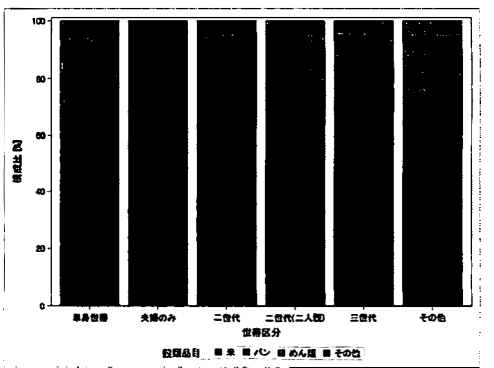

## 3-③) パン支出金額についてのヒストグラム



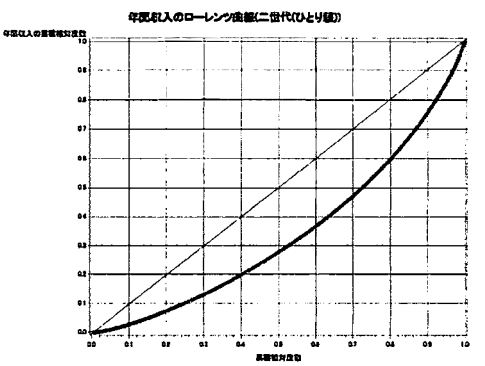
## 4-①) 世帯区分別及び世帯分類別の穀類支出金額に占める構成比の平均値

| 項目名 | | 件数 | 構成比(%) 米 | パン | めん類 | 他の穀類 | 全体(%) |
|---|---|---|---|---|---|---|---|
| 世帯区分 | 勤労者世帯 | 57,386 | 36.0 | 37.6 | 22.1 | 4.3 | 100.0 |
| | 勤労者以外の世帯 | 16,874 | 44.9 | 31.0 | 19.4 | 4.7 | 100.0 |
| | 無職世帯 | 25,741 | 52.1 | 26.7 | 15.8 | 5.4 | 100.0 |
| 世帯分類 | 単身世帯 | 20,758 | 33.8 | 37.7 | 23.0 | 5.5 | 100.0 |
| | 夫婦のみ | 24,287 | 49.5 | 28.9 | 16.8 | 4.8 | 100.0 |
| | 二世代 | 37,182 | 39.0 | 35.9 | 20.8 | 4.3 | 100.0 |
| | 二世代(ひとり親) | 7,302 | 44.9 | 31.6 | 19.0 | 4.5 | 100.0 |
| | 三世代 | 8,646 | 45.0 | 30.5 | 19.7 | 3.8 | 100.0 |
| | その他 | 1,826 | 44.8 | 30.5 | 20.4 | 4.3 | 100.0 |
| 全体 | | 100,000 | 41.6 | 33.7 | 20.0 | 4.6 | 100.0 |

## 4-②) 構成比の帯グラフ




## 5-①) 二世代（ひとり親）世帯における年間収入のローレンツ曲線




## 5-②) 各世帯分類の年間収入のジニ係数

| 世帯分類 | ジニ係数 |
|---|---|
| 単身世帯 | 0.34 |
| 夫婦のみ | 0.31 |
| 二世代 | 0.26 |
| 二世代(ひとり親) | 0.35 |
| 三世代 | 0.27 |
| その他 | 0.33 |
| 合計 | 0.33 |

## 主催：SASユーザー会　世話人会

| | | |
|---|---|---|
| **代表世話人** | **大橋 靖雄** | **中央大学　教授** |
| 世　話　人<br>（氏名50音順） | 伊藤 陽一 | 北海道大学　准教授 |
| | 小野 潔 | 株式会社インテック |
| | 岸本 淳司 | 九州大学ARO次世代医療センター　副センター長 准教授 |
| | 堺 伸也 | イーピーエス株式会社 |
| | 坂巻 英一 | |
| | 周防 節雄 | 兵庫県立大学　名誉教授 |
| | 菅波 秀規 | 興和株式会社 |
| | 高橋 行雄 | BioStat研究所株式会社 |
| | 野原 賢一 | 独立行政法人 統計センター |
| | 松岡 淨 | 日本メディア株式会社／順天堂大学 臨床研究支援センター |
| | 宮内 亨 | 独立行政法人 統計センター |
| | 八木 章 | 近畿大学　元教授 |
| | 山之内 直樹 | 第一三共株式会社 |

## 協賛

（掲載50音順）

イーピーエス株式会社
エイツーヘルスケア株式会社
SAS Institute Japan 株式会社 JMPジャパン事業部
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
日本メディア株式会社

## 協力

SAS Institute Japan株式会社

## SASユーザー総会事務局

〒160-0022　東京都　新宿区新宿6-27-56 新宿スクエア5F
TEL：03-5774-7018　受付時間：10:00～17:00（平日 12:00～13:00、土・日・祝日を除く）
E-mail：sasuser_entry@seminar.jp

SAS ユーザー総会 2017

**論文集**

2017年8月3日 初版第1刷発行
発行：SASユーザー会　SAS Institute Japan株式会社